Comparison of High Performance Network Options:
EDR InfiniBand vs. 100Gb RDMA Capable Ethernet

Luke Kachelmeier, Faith Van Wig, Kari Erickson

Mentors: Susan Coulter, Howard Pritchard

Affiliations: LANL CSCNSI, University of New Mexico, Missouri University of Science and Technology, New Mexico Institute of Mining and Technology

LA-UR-16-27754

Background Information

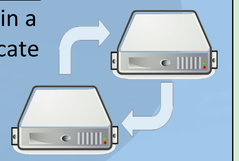
InfiniBand (IB) - A Layer-2 network solution that advertises high bandwidth and low latency.

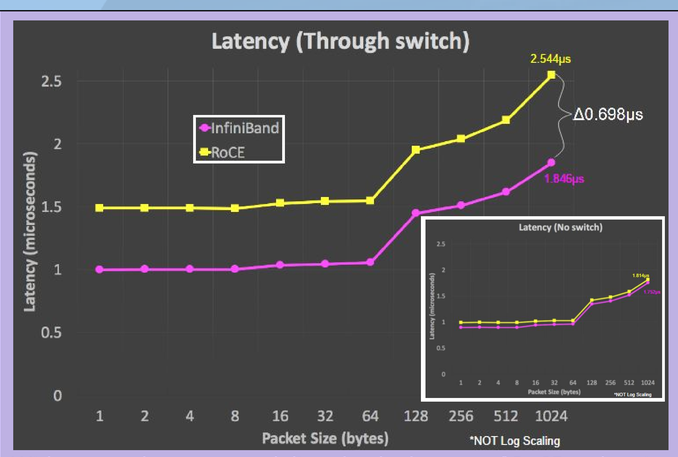
RDMA - Remote Direct Memory Access. Allows for nodes to access connected nodes' memory directly. IB uses this protocol natively.

RoCE - RDMA over Converged Ethernet.

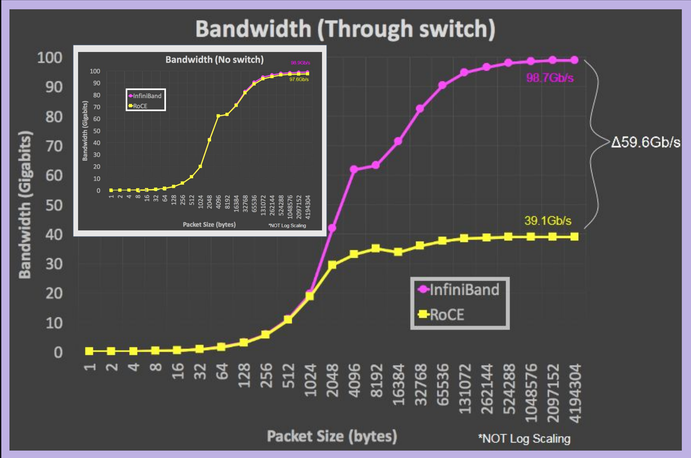
Single-Rank Performance

Point-to-Point tests check the performance of the cluster in a workload scenario that only two nodes need to communicate across a single rank. We use OSU point-to-point tests for latency and bandwidth, osu_latency and osu_bw. Native ethernet is not included in these graphs so differences between IB and RoCE are noticeable.





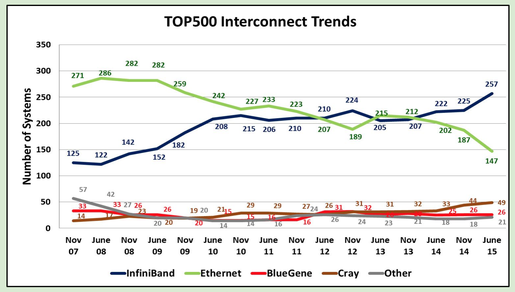
With two nodes connected directly, the latency of RoCE and native IB are very comparable. However, when connected through a switch, IB outperforms RoCE by a magnitude of 10. Ethernet latency performed between 12 and 14 microseconds.



With two nodes connected directly, the bandwidths for RoCE and IB are once again very similar. When going through the switch, however, the switch could only read RoCE as a 40Gb network instead of 100Gb, which explains the drop in RoCE performance with the switch.

Overview

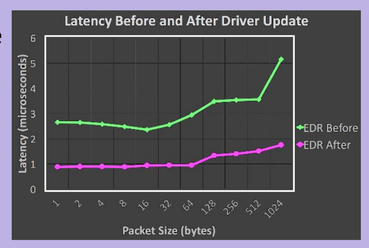
InfiniBand (IB) has long been the network of choice for high performance computing (HPC). New advances in technology have made Ethernet a potential network alternative. The goal of this study is to compare the deployment effort and performance of each technology.



http://blog.infinibandta.org/tag/top500/

Impact

We found an unexpected side effect of this project while optimizing native IB performance. Simply updating the cards' drivers from 2.2-1 to 3.3-1 made a significant impact on latency, as shown on the left. Los Alamos National Laboratory is now planning on updating their drivers on production systems where applicable.



Conclusions

The 100Gb Ethernet hardware from these two different companies are not compatible. This is not unexpected, given that 100Gb bandwidth hardware and software is very new and not yet well supported or standardized. The slight variations in hardware manufacture, a lack of testing for a support matrix, and a lack of availability in the open source community for these new technologies make even basic network setup challenging, particularly when using components from different vendors. This project is currently bringing these companies, Mellanox and Juniper, together to work on this incompatibility

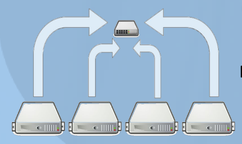
Contact Us

Luke Kachelmeier ----------lkachelmeier@unm.edu
Faith Van Wig ---------- fvvnr5@mst.edu
Kari Erickson----------kari.erickson@student.nmt.edu

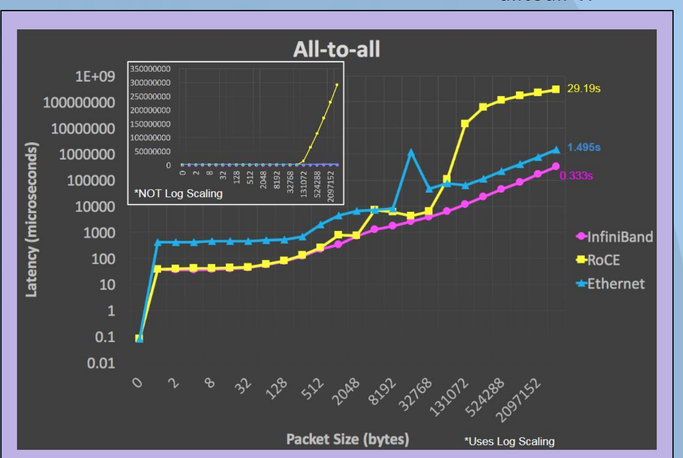
Experimental Setup

- CentOS 6.8 controlled via Warewulf 3.6-1
- Mellanox ConnectX4 EDR adapter cards, v.12.16.1020
- Mellanox 7700 EDR InfiniBand Switch
- Juniper QFX5200 100Gb Ethernet Switch
- MLNX OFED and related modules, v. 3.3-1
- OpenMPI 1.10.3 using MXM libraries

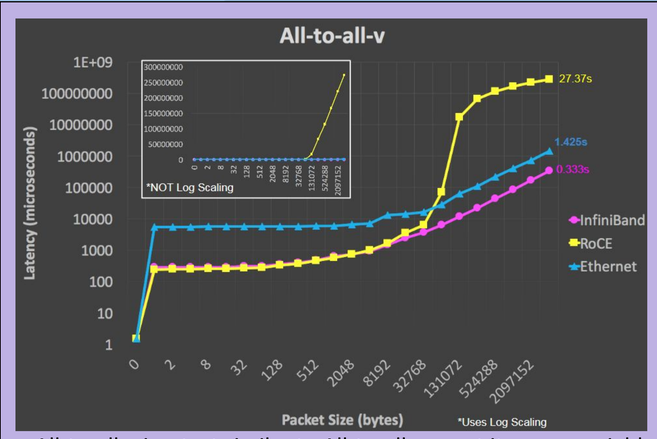
Multi-Rank Performance



Collective tests check the performance of the cluster with a workload scenario in which multiple ranks communicate. We use the IMB tests for collective latency, alltoall, and alltoall-v.



Our collective test all-to-all, tests latency of the network by sending packages between the nodes with a set package size. The smaller graph shows non-log scaling to demonstrate RoCE's drop in performace, while the larger graph compares how the different network options compare to each other.



All-to-all-v is a test similar to All-to-all, except it uses a variable size of messages as it passes them back and forth. The smaller graph shows non-log scaling, demonstrating how much RoCE drops in performance at around 120000 sized packets.