Comparison of High Performance Network Options:

EDR InfiniBand vs. 100Gb RDMA Capable Ethernet

LA-UR-16-25357

Kari Erickson
New Mexico Institute of Mining and Technology
Socorro, NM
kari.erickson@student.nmt.edu

Luke Kachelmeier
University of New Mexico
Albuquerque, NM
lkachelmeier@unm.edu

Faith Van Wig
Missouri Institute of Science and Technology
Rolla, MO
fvvnr5@mst.edu

I. Introduction

InfiniBand (IB) has long been the network of choice for high performance computing (HPC). However, advancements in both Ethernet and IB technology, as well as other high-performance networks, have made it necessary to analyze the performance of these network options in detail – specifically, we look at 100Gb Ethernet and IB. Advancements in Ethernet include upwards of 100Gb data rates and standardization of RDMA-over-Converged-Ethernet (Routable RoCE) [1]. Similarly, IB has introduced Enhanced Data Rate (EDR) hardware, which nearly doubles previous bandwidth, increasing it to 100Gb [2]. The goal of this study is to compare and contrast each option by looking at their respective bandwidth and latency performance, as well as message injection rates and general deployment effort.

II. Experimental Setup

Our cluster consists of six diskless compute nodes, each with a Mellanox ConnectX-4 EDR HCA. When running as an InfiniBand link layer, they communicate across a Mellanox MSB7700-ES2F EDR Mellanox switch. When running as an Ethernet link layer, they communicate across a 100Gb Juniper QFX5200 Data Center switch.

The cluster is then tested for performance using native IB, native Ethernet, and RoCE via OpenMPI 1.10.3 built with Mellanox OFED3.3's MXM libraries. OSU Micro Benchmarks test point-to-point latency, bandwidth, bidirectional bandwidth, and message injection rate performance. Intel Micro Benchmarks test collective network latency using the MPI_Alltoall and MPI_Alltoallv tests.

III. Preliminary Results

Preliminary point-to-point results show that the 8-byte latency across a switch using native IB and RoCE compare favorably – 1.1µs compared with 1.4µs. However, brief collective testing of our setup is currently showing the hop latency through the Juniper switch is significantly higher than that through the Mellanox switch – 300ns as opposed to 100ns. Asymptotic MPI bandwidth for both transports both were around 12GB/s; however, results from IMB all-to-all and all-to-all-v tests currently indicate that the Mellanox switch is delivering significantly better performance than the Juniper switch. For example, the time taken for an MPI_Alltoall operation at 4 MB transfers is about 50% greater across the Juniper switch than the Mellanox switch.

TABLE I. Directly Connected Latency

Message Size	Average Latency		
	Native IB (us)	*RoCE (us)*	*Native Ethernet (us)*
8	0.892	0.988	12.662
16	0.938	1.014	12.692
32	0.95	1.026	12.718
64	0.958	1.024	12.712
128	1.344	1.418	12.78
256	1.402	1.474	12.884
512	1.518	1.584	13.104
1024	1.752	1.814	13.55

FIGURE I. Directly Connected Latency

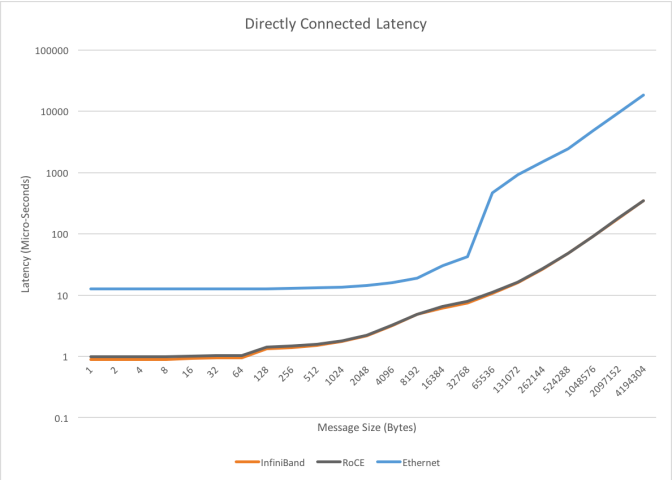
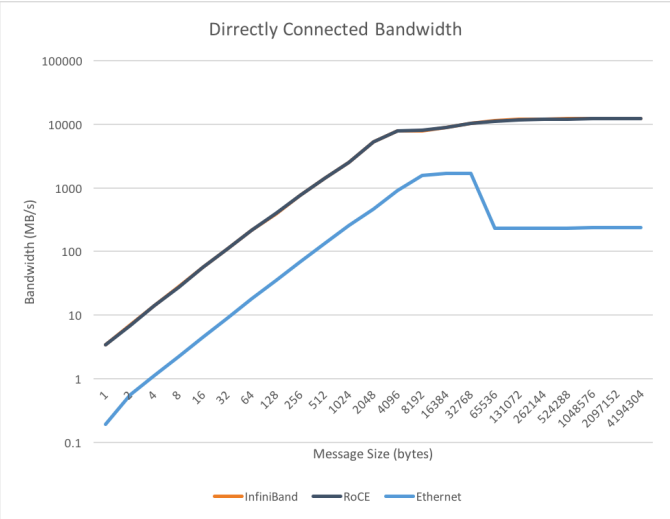


TABLE II. DIRECTLY CONNECTED BANDWIDTH

Message Size	Average Bandwidth		
	Native IB (MB/s)	*RoCE (MB/s)*	*Native Ethernet (MB/s)*
131,072	11,848	11,670	232
262,144	12,086	11,911	233
524,288	12,235	12,076	234
1,048,576	12,318	12,145	235
2,097,152	12,351	12,178	235
4,194,304	12,371	12,199	235

FIGURE II. DIRECTLY CONNECTED BANDWIDTH



IV. DISCUSSION

Although both switches advertise 100Gb/s bandwidth, EDR uses 64/66 bit encoding [3]. This means that the bandwidth for any of the network options has a theoretical maximum of 97Gb/s. RoCE and IB behave similarly in direct comparison. Both network choices are showing to be viable options for HPC.

Results for native Ethernet may reflect difficulties faced in optimizing Raw 100Gb Ethernet performance. While directly connected, iperf results peaked at 63Gb/s and OSU point-to-point latency benchmark runs peaked at 16Gb/s. This is potentially due to a lack of OpenMPI optimization for communicating with TCP over multi-lane EDR cables, but it is also possible that there is a deeper problem in our Ethernet configuration, which we will continue to investigate.

There is some potential in terms of optimizing the performance of the all-to-all Ethernet and RoCE tests. However, if the results remain as is, we can come to the conclusion that native IB performs significantly better for communications across larger networks.

Most of our current point-to-point results were generated with directly connected cable connections. We will replace these results with those running through a switch for an applicable comparison after both switches have been optimized.

V. FUTURE WORK

We would like to optimize the performance of native Ethernet in order to gain a more reliable comparison of networking options. Furthermore, we are working on testing the performance of both switches and hope to have confirmed and accurate results by August.

We would also like to test the performance of NFS across InfiniBand as well as IP over IB.

REFERENCES

[1] Berry, Frank. "QLogic Announces Wrol's 1st 100Gb Ethernet Adapter." *Worlds Top Data Centers*. Web. 20 July 2016.

[2] "InfiniBand Roadmap." *InfiniBand Roadmap: IBTA*. InfiniBand Trade Association. Web. 18 July 2016.

[3] "InfiniBand Overview." *High Performance Computing*. University of Iowa. Web. 20 July 2016.