Research and Evaluation of RoCE in IHEP Data Center

Shan Zeng[1] , *Fazhi Qi*[1] , *Lei Han*[2], *Xiangyu Gong*[2], *Tao Wu*[2]

[1]Computing Center, Institute of High Energy Physics Chinese Academy of Science, Beijing 100049, China
[2]Huawei Nanjing R&D Center, Huawei Technologies Co. ,Ltd., Nanjing, 210012, China

Abstract. With more and more large-scale scientific facilities are built, more and more HPC requirements are needed in IHEP. RDMA is a technology that allows servers in a network to exchange data in main memory without involving the processor, cache or operating system of either server, which can provide high bandwidth and low latency. There are two RDMA technologies which were InfiniBand and a relative newcomer called RoCE – RDMA over Converged Ethernet. This paper introduces the RoCE technology, we research and compare the performance of both IB and RoCE in IHEP data center, and we also evaluate the application scenarios of RoCE which can support our future technology selection in HEPS. In the end, we present our future plan.

1 Introduction

Modern data centers are tasked with delivering intelligent multi-media responses to real-time human interactions. Massive amounts of data are being churned and sifted by highly parallel applications, such as online data intensive services (OLDI) and artificial intelligence (AI), which historically required high performance network.

Generalized cloud infrastructure is also being deployed in the data center of IHEP. The key to advancing cloud infrastructure to the next level is the elimination of loss in the network, which include not just packet loss, but also throughput loss and latency loss. There simply should be no loss in the data center network. Congestion is the primary source of loss and in the network, congestion leads to dramatic performance degradation. New advancements in high-speed distributed solid-state storage, coupled with remote direct memory access (RDMA) and new networking technologies to better manage congestion, are allowing these parallel environments to run atop more generalized next generation cloud infrastructure.

RDMA is a technology that allows servers in a network to exchange data in main memory without involving the processor, cache or operating system of either server, which can provide high bandwidth and low latency[1].

There are two RDMA technologies which were InfiniBand(IB) and a relative new comer called RDMA over Converged Ethernet (RoCE). We compare the two technologies in architecture and do some performance evaluation on the general MPI (Message Passing interface) scenarios.

1.1 InfiniBand

InfiniBand (IB) is a computer networking communication standard used in high-performance computing that features very high throughput and very low latency. It is used for data interconnect both among and within computers. IB can transfer data directly to and from a storage device on one machine to user space on another machine, bypassing and avoiding the overhead of a system call.
InfiniBand architecture (IBA) is designed around a point-to-point switched I/O fabric, whereby end node devices are interconnected by a cascaded switch device. IBA defines hardware transport protocols sufficient to support both reliable messaging and memory manipulation semantics without software intervention in the data movement path. IBA is divided into multiple layers where each layer operates independently of one another. As shown in Fig.1, IBA protocol layer is divided into the five layers: Physical, Link, Network, Transport, and Upper Layers[2].

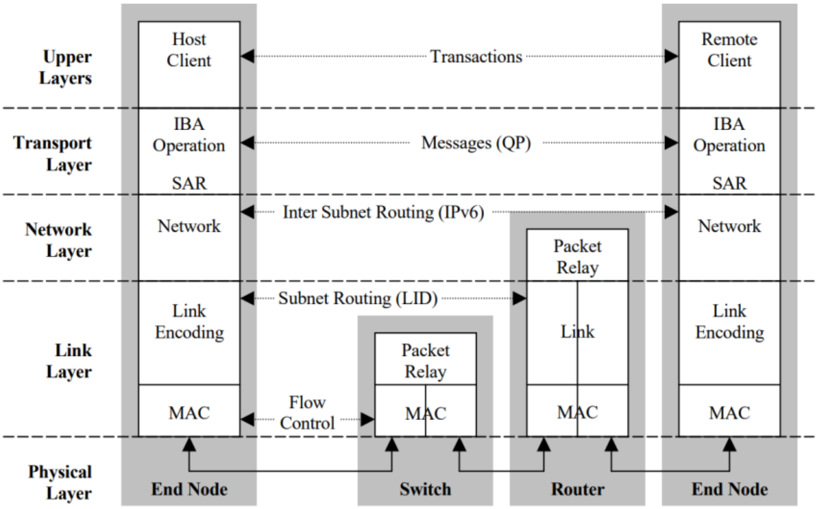

Fig.1. IBA protocol layers

Instead of sending data in parallel, which is what PCI does, InfiniBand sends data in serial and can carry multiple channels of data at the same time in a multiplexing signal. The principles of InfiniBand mirror those of mainframe computer systems that are inherently channel-based systems. InfiniBand channels are created by attaching host channel adapters (HCAs) and target channel adapters (TCAs) through InfiniBand switches. HCAs are I/O engines located within a server. TCAs enable remote storage and network connectivity into the InfiniBand interconnect infrastructure, called a fabric. InfiniBand architecture is capable of supporting tens of thousands of nodes in a single subnet.

1.2 RoCE

RoCE is a network protocol defined in the InfiniBand Trade Association (IBTA) standard, allowing RDMA over converged Ethernet network. Shortly, it can be regarded as the application of RDMA technology in hyper-converged data centers, cloud, storage, and virtualized environments. It possesses all the benefits of RDMA technology and the familiarity of Ethernet [3]. Generally, there are two RoCE versions: RoCE v1 and RoCE v2. It depends on the network adapter or card used. The underlying ISO stacks of IB, RoCE v1 and RoCE v2 can be shown in Fig.2.

RoCE v1: The RoCE v1 protocol is an ethernet link layer protocol allowing two hosts in the same ethernet broadcast domain (VLAN) to communicate. It uses ethertype 0x8915, which limits the frame length to 1500 bytes for a standard Ethernet frame and 9000 bytes for an Ethernet jumbo frame.
RoCE v2: The RoCE v2 protocol overcomes the limitation of version 1 being bounded to a single broadcast domain (VLAN). It operates at L3. By changing the packet encapsulation to include IP and UDP headers, RoCE v2 can now be used across both L2 and L3 networks. This enables L3 routing, which brings RDMA to network with multiple subnets for great scalability. Therefore, RoCE v2 is also regarded as Routable RoCE (RRoCE). Owing to the arrival of RoCE v2, the IP multicast is now also possible [4].
Both RoCE v1 and RoCE v2 require a lossless network configuration. RoCE v1 requires a lossless L2 network, and RoCE v2 requires that both L2 and L3 are configured for lossless operation.

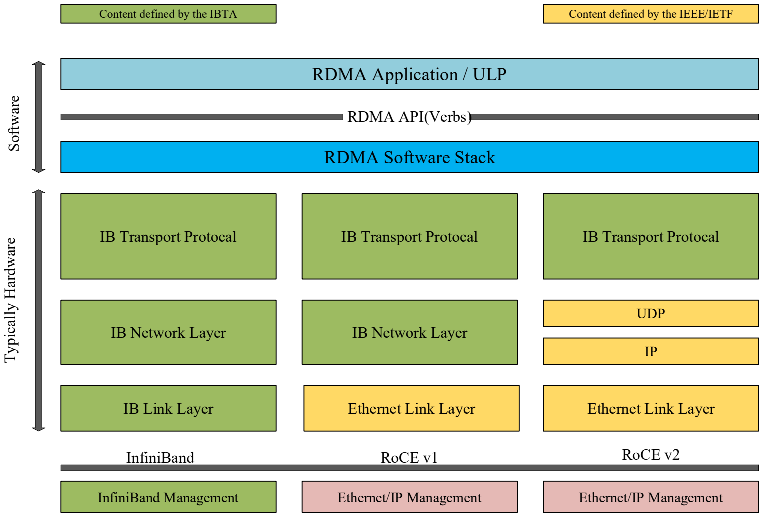

Fig.2. Underlying ISO Stacks of the Flavors of RDMA

1.3 IB vs RoCE

Based on the HPC TOP500[5], as shown in Fig,3. More and more companies in HPC top500 choose ethernet network as their network infrastructure. In the year of 2020, 262 companies in top500 choose ethernet network.

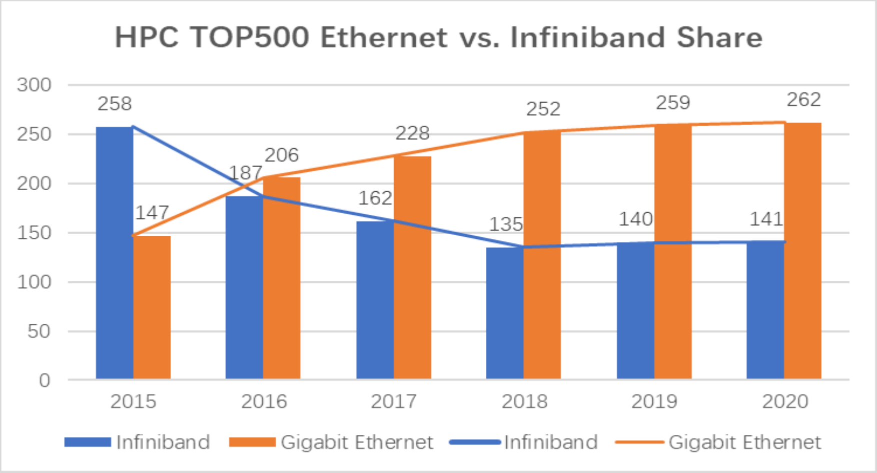

Fig. 3. HPC TOP500 Ethernet vs. Infiniband Share

2 Experimental Setup

Our cluster consists of seven DELL R640 computing nodes, each with a Mellanox ConnectX-5 EDR HCA. When running under InfiniBand infrastructure, they communicate across a Mellanox MSB7800 EDR Mellanox switch, the network topology is shown in Fig.4. When running under ethernet network infrastructure, they communicate across a 100Gb HUAWEI CE8850-64CQ-EI data center switch, the network topology is shown in Fig.5. We then compare the performance between the two network infrastructures, using OSU Micro Benchmarks [6] via HMPI b007 built with Mellanox OFED 4.7-3.2.9 [7]. We test point-to-point latency and bi-directional bandwidth using OSU_latency and OSU_bibw, and we also test collective communication performance using OUS_ALLREDUCE and OSU_ALLTOALL.

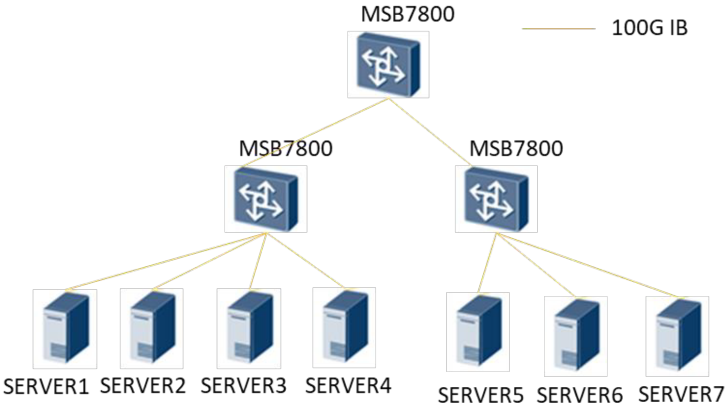

Fig. 4. IB Network topology of the cluster

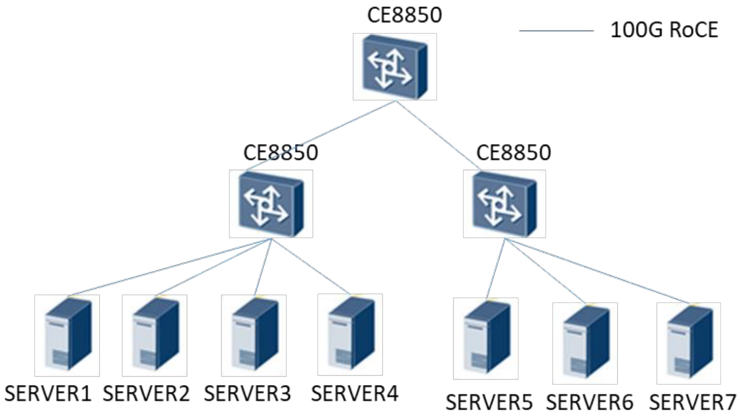

Fig. 5. RoCE Network topology of the cluster

3 Performance Evaluation Results

3.1 Throughput

RoCE network bidirectional point-to-point bandwidth is similar to IB network. The detailed test results can be seen in Fig.6.

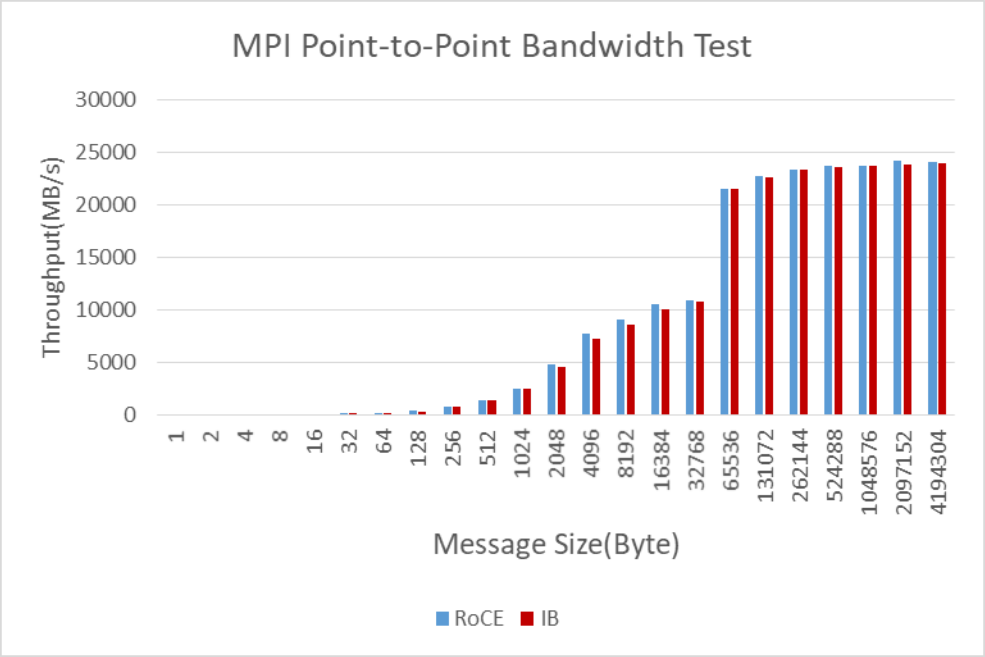

Fig. 6. MPI point-to-point bandwidth test result

3.2 Latency

RoCE network latency is from 1.5 to 1.6 us larger than IB network in a 3 hops spine-leaf topology, because of nearly 0.5 us switch latency gap between RoCE and IB switches per hop, which is related to the forwarding mechanism differences. The detailed test results can be seen in Fig.7.

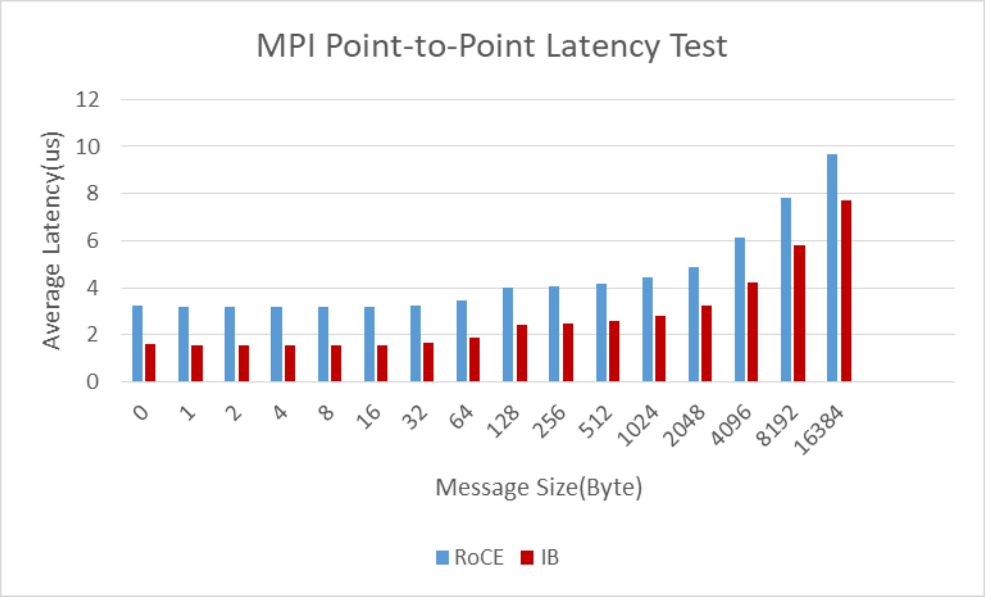

Fig. 7. MPI point-to-point latency test result

3.3 MPI ALLREDUCE

168 cores are used during the test and ppn(process per node) is set to 24. RoCE network allreduce average latency is slightly shorter than IB. The improvement ranges from 4.5% to 13% when message size ranges from 8 bytes to 256 bytes. The detailed test results can be seen in Fig.8.

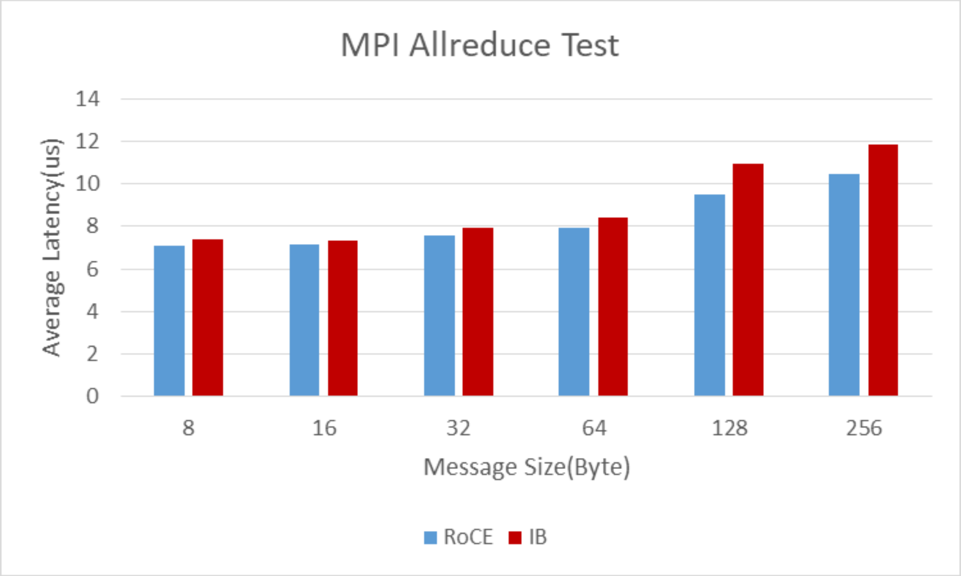

Fig. 8. MPI Allreduce test result

3.4 MPI ALLTOALL

168 cores are used during the test and ppn(process per node) is set to 24. RoCE network alltoall average latency is slightly shorter than IB. The improvement ranges from 7.9% to 17.2% when message size ranges from 131072 bytes to 1048576 bytes. The detailed test results can be seen in Fig.9.

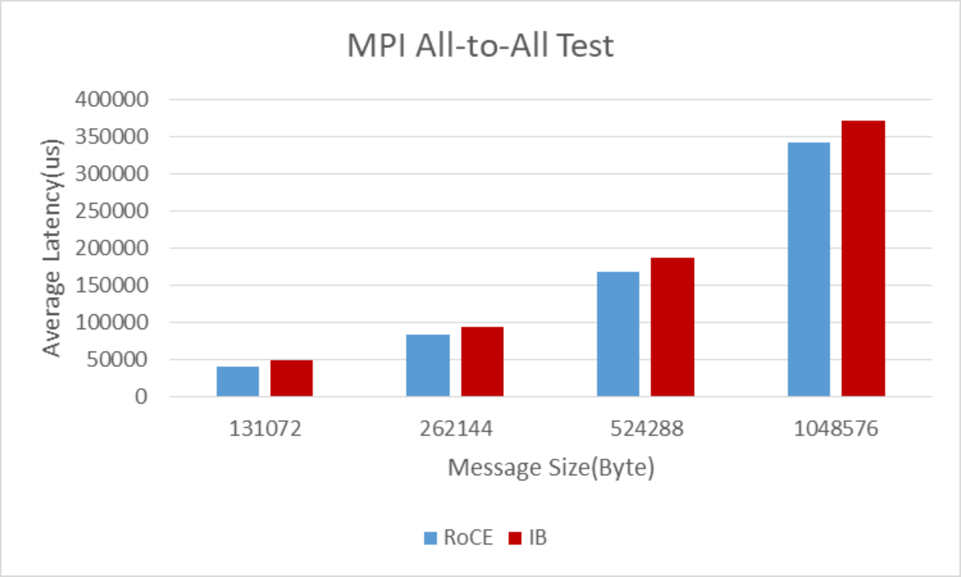

Fig. 9. MPI All-to-All test result

3.5 Conclusion

RoCE network performs slightly better than IB network in both point-to-point and collective tests, except for the latency test.

4 Future Work

We would like to optimize the performance of RoCE in order to gain a more reliable comparison of networking options. Furthermore, we will test more HEP applications on RoCE environment, such as the performance of the storage system(Lustre) which will be used in HEPS across InfiniBand as well as RoCE.

References

1. Liu J , Wu J , Kini S P , et al. High Performance RDMA-Based MPI Implementation over InfiniBand[C]// Proc Acm International Conference on Supercomputing. IEEE, 2003:295.
2. Mellanox Technologies. Introduction to InfiniBand. https://www.mellanox.com/pdf/whitepapers/IB_Intro_WP_190.pdf
3. Krawczyk R D , Colombo T , Neufeld N , et al. Feasibility tests of RoCE v2 for LHCb event building[J]. The European Physical Journal Conferences, 2020, 245:01011.
4. Schelten N , Steinert F , Schulte A , et al. A High-Throughput, Resource-Efficient Implementation of the RoCEv2 Remote DMA Protocol for Network-Attached Hardware Accelerators[C]// International Conference on Field-Programmable Technology (FPT'20). 2020.
5. TOP500. https://www.top500.org/
6. OSU Micro Benchmarks. http://mvapich.cse.ohio-state.edu/benchmarks/
7. Mellanox OFED 4.7-3.2.9. https://docs.mellanox.com/display/MLNXOFEDv473290/Installing+Mellanox+OFED