Research and Evaluation of RoCE in IHEP Data Center

Shan Zeng, *Fazhi Qi* **,** *Lei Han* **,** *Xiangyu Gong* **,** *Tao Wu*

zengshan@ihep.ac.cn

5-19-2021

Outline

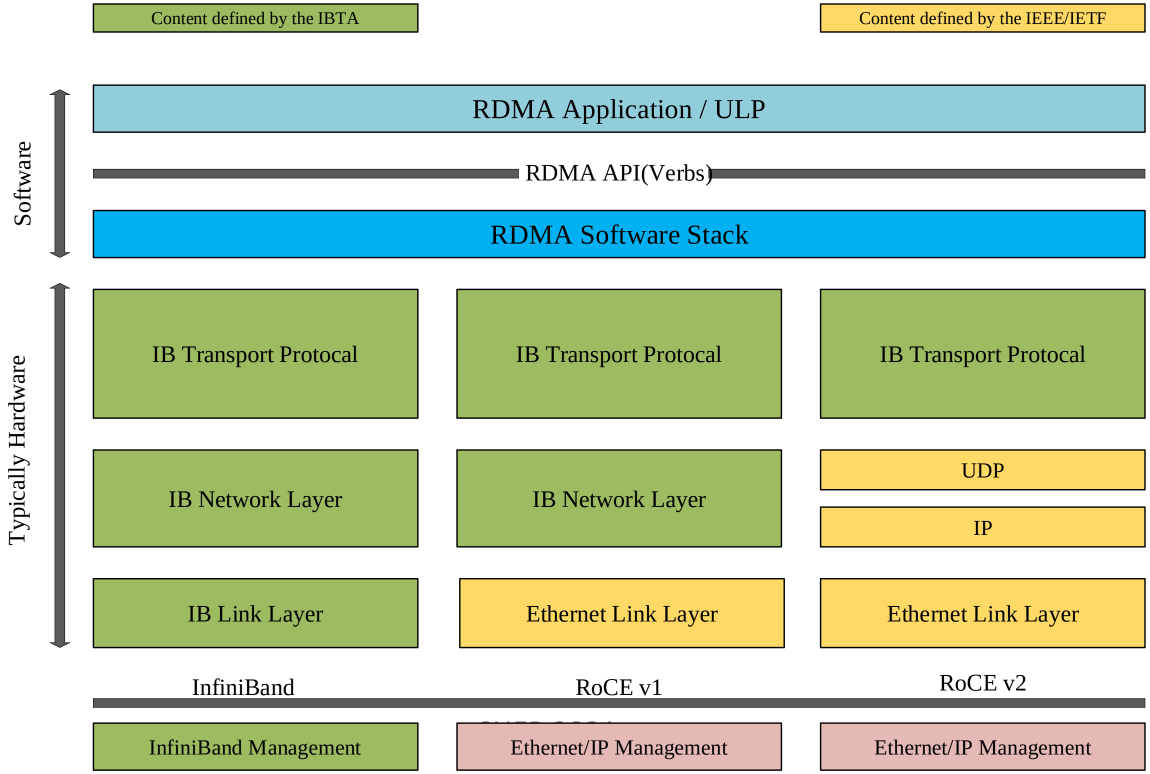
- **Motivation/Background**
- **Introduction of RDMA**
  - RoCE
  - RoCE vs IB
- **Testbed setup**
- **Performance evaluation of RoCE**
- **Summary**

Background

- **More and more large scientific facilities are being built or running**
- **Various types of applications and corresponding computing models are emerging**
  - LQCD
  - OLDI (online data intensive services)
  - …….
- **HPC requires high performance network**
- **More features are needed in high performance network**
  - High bandwidth
  - Low latency
  - Zero package loss
  - Stable
  - Scalable
  - Flexible
  - Manageable

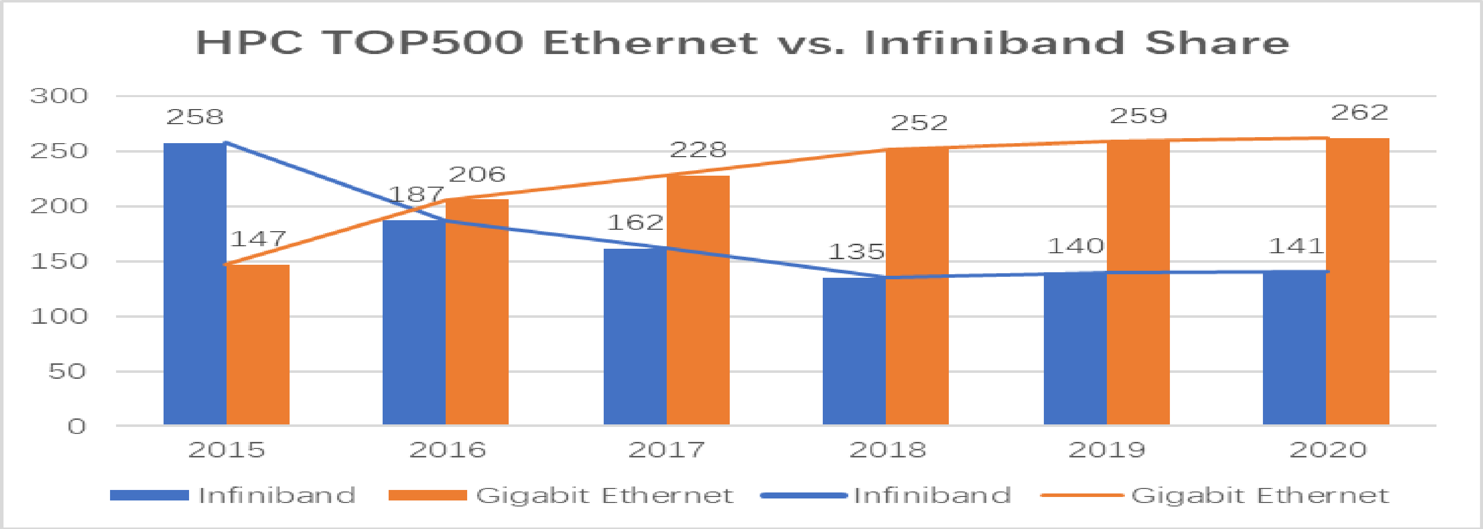
RDMA

- **RDMA: Remote Direct Memory Access**
- **Provide high bandwidth and low latency**
  - Allows servers in a network to exchange data in main memory without involving the processor, cache or operating system of either server
- **2 common flavors of practice in RDMA**
  - InfiniBand (IB)
  - RoCE (RDMA over Converged Ethernet)



IB vs RoCE

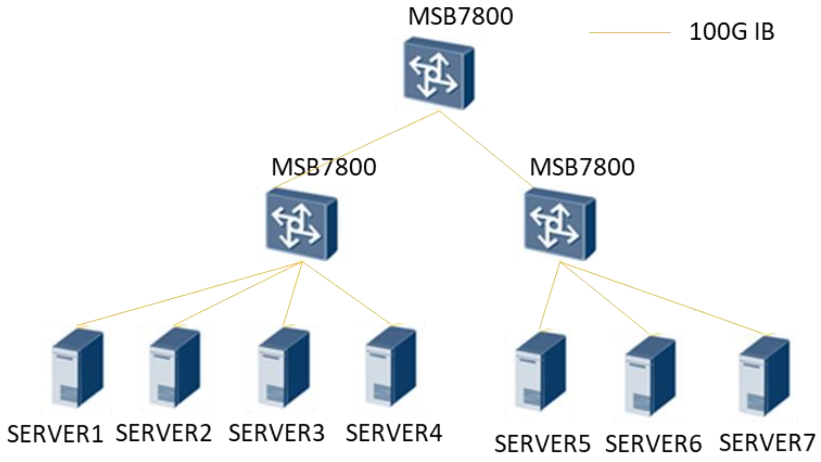
name	Underlying ISO Stacks	Ecosystem	Configuration	Cost
IB	IB link layer and network protocol	close	complicated	Higher
RoCEv2	Ethernet link layer and IP/UDP protocol	More open	easy	lower

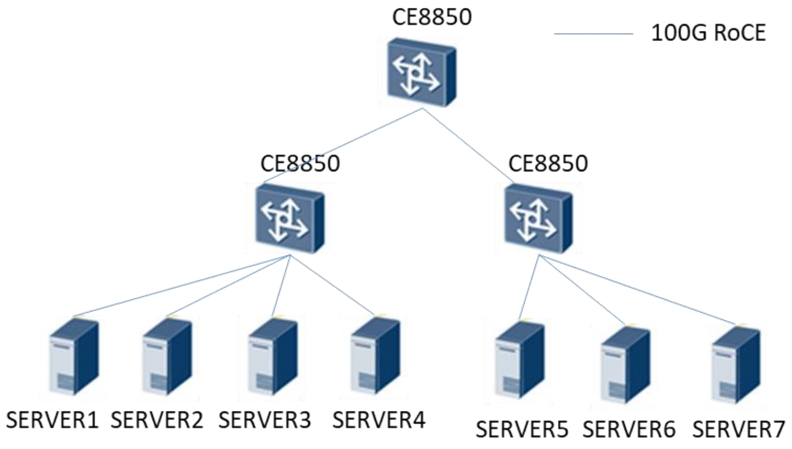


Experimental setup

Switch	Type	Vendor
Leaf-RoCEv2	CE8850-64CQ-EI	HUAWEI
Spine-RoCEv2	CE8850-64CQ-EI	HUAWEI
Leaf-IB	MSB7800	Mellanox
Spine-IB	MSB7800	Mellanox

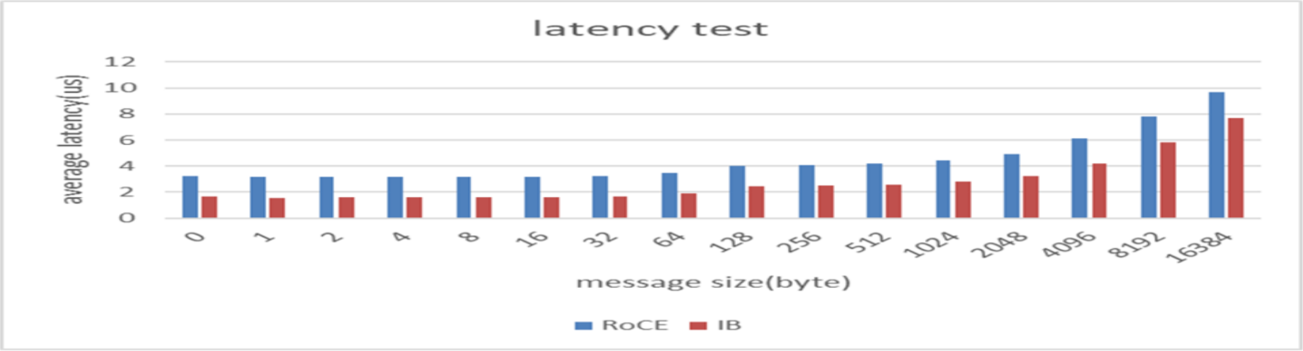
Server	OS	NIC/Driver Version	MPI Version	Benchmarks
DELL R640	Centos7.5	MCX556A/OFED4.7-3.2.9	HMPI, Version:b007	OSU Micro Benchmarks





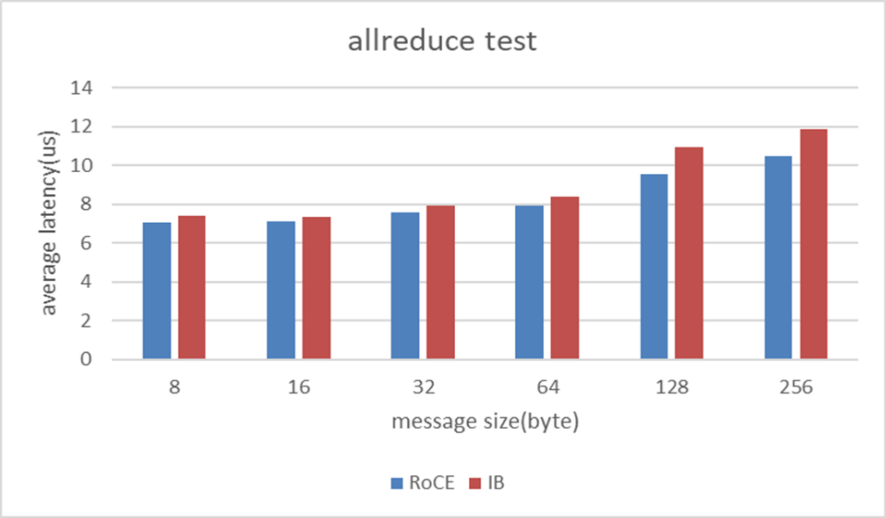
Evaluation Results(I)

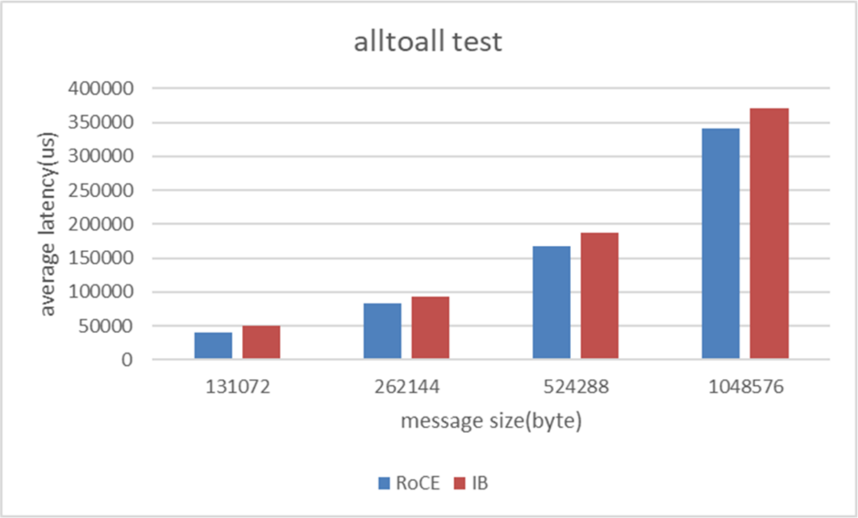
- **Network bidirectional bandwidth**
  - RoCE performs same level with IB
- **Network static latency**
  - RoCE is from 1.5 to 1.6 us larger than IB network in a 3 hops spine-leaf topology
  - Caused by the forwarding mechanism differences between RoCE and IB switches
  - nearly 0.5 us switch latency gap between RoCE and IB switches per hop



Evaluation Results(II):MPI

- **Resources**
  - CPU cores: 168
  - PPN (process per node) is set to 24
- **MPI allreduce**
  - RoCE performs a bit better than IB in allreduce average latency test
  - The improvement ranges from 4.5% to 13% when message size ranges from 8 bytes to 256 bytes
- **MPI alltoall**
  - RoCE performs a bit better than IB in alltoall average latency test
  - The improvement ranges from 7.9% to 17.2% when message size ranges from 131072 bytes to 1048576 bytes





Conclusion

- **IHEP started to research and evaluate RoCE in the end of last year**
  - We do some basic MPI benchmark test
  - RoCE performs slightly better than IB network in both point-to-point and collective tests except for the static latency test.
- **Future work**
  - More benchmarks should be tested to better evaluated RoCE, such as Linpack
  - more HEP applications will be tested in RoCE environment, such as Lustre and EOS
- **Cooperation will be needed and welcomed**

Thanks for your attention