ISSN:0975-9646

Comparing Ethernet & Soft RoCE over 1 Gigabit Ethernet

Gurkirat Kaur, Manoj Kumar[1], Manju Bala[2]

[1]*Department of Computer Science & Engineering,*
CTIEMT Jalandhar, Punjab, India
[2]*Department of Electronics and Communication Engineering,*
CTIEMT, Jalandhar, Punjab, India

ABSTRACT: In recent years, we have witnessed a growing interest in optimizing the high performance computing (HPC) solutions using advanced CPU and Interconnect technologies. recent trends in HPC systems have shown that future increases in performance can only be achieved through increases in system scale using a larger number of components, such as multi-cores and faster interconnects. RDMA is an emerging technology, which is used for reducing system load & improves the performance. In this paper, we evaluate the heterogeneous Linux cluster, having multi nodes with fast interconnects i.e. gigabit Ethernet & Soft RoCE. This paper presents the heterogeneous Linux cluster configuration & evaluates its performance using MVAPICH platform & OSU Benchmarks. Our result shows that Soft RoCE is performing better or equal to Ethernet in various performance metrics like bandwidth, latency & throughput.

Keywords: HPC, MPI, RDMA, RoCE, Soft RoCE

1. INTRODUCTION

In recent years, we have witnessed a growing interest in optimizing the high performance computing solutions using advanced CPU and Interconnect technologies. This interest is motivated by the fact that single CPU-chips are reaching their physical limits in terms of heat dissipation and power consumption. Therefore recent trends in HPC systems have shown that future increases in performance can only be achieved through increases in system scale using a larger number of components, such as multi-core CPUs and ultra-fast interconnects. In terms of HPC interconnects, there are several network interconnects that provide ultra-low latency (less than 1 µsec) and high bandwidth (several Gbps). Some of these interconnects may provide flexibility by permitting user-level access to the network interface cards for performing communication, and also supporting access to remote processes' memory address spaces . Examples of these interconnects are Myrinet from Myricom, Quadrics and InfiniBand [1]. The main focus of this paper is on the RDMA over Converged Ethernet (RoCE) which is an InfiniBand (IB) protocol that can be used over the Ethernet infrastructures. RoCE provide all of the InfiniBand transport benefits and also provide well established RDMA ecosystem combined with converged Ethernet. It is also called link layer protocol which allows the communication between the two hosts on the same Ethernet broadcast domain. The main advantage of RoCE is that it can implement in hardware as well as in software. The software implementation of RoCE is called Soft RoCE. Our Objective of this paper is to evaluate the performance of a heterogeneous Linux cluster using the one of the most commonly used MPI Implementations. The rest of the paper is organised as section 1 gives the introduction, overview of MPI implementation, RoCE. Section 2 describes the Experimental Setup. Section 3 describes the Results and Discussions of Soft RoCE over1 gigabit Ethernet using OSU Micro Benchmark. Section 4 concludes the paper.

1. MPI Implementation

In this section, we briefly define the technologies used to benchmark the Linux cluster. These are the Gigabit Ethernet interconnect technology, the OFED's Soft RoCE Distribution MVAPICH-The MPI Implementation and OSU Micro benchmark.

1.1MVAPICH

Message Passing Interface (MPI) has been the most popular programming model for developing parallel applications. MVAPICH is maintained by the Department of Computer Science & Engineering of Ohio State University [1] .MVAPICH is software that delivers the best performance, scalability & fault tolerance for the high-end computers & servers using the InfiniBand interconnect & other RDMA - enabled interconnect network technologies. Few standards are available MVAPICH and MVAPICH2 and MVAPICH2-X. MVAPICH is based on MPI-1 & MVICH.MVAPICH2 is based on MPI-2 & MPICH2 & MPICH [2]. MVAPICH2-X is based on MVAPICH2 & supports all MPI-3 features. It includes all the MPI-1 & MPI-2 features. It includes several features like:

(1) RDMA fast path utilizing RDMA operations for efficient small messages.
(2) High performance and scalable support for one sided communication.
(3) High performance two-sided communication scalable to multi-thousand nodes.
(4)

1.2 Micro Benchmark- OSU MPI Benchmark

The Ohio Micro Benchmark suite is a collection of independent MPI message passing performance micro benchmarks developed and written at The Ohio State University. It includes traditional benchmarks and performance measures such as latency, bandwidth and host overhead and can be used for both traditional and GPU-enhanced nodes. It is a suite of micro-benchmarks for testing various MVAPICH2 MPI operations. The OSU

Micro Benchmark (OMB) suite has been the most widely used set of benchmarks to compare the performance of different MPI libraries on clusters. We will focus to measures the point to point MPI Benchmarks, Collective Benchmarks, and One-sided MPI Benchmarks using OSU Micro Benchmark.

2. RDMA OVER CONVERGED ETHERNET (ROCE)

RoCE is an InfiniBand Trade Association (IBTA) Standards.. RoCE utilizes the Open Fabrics Enterprise Distribution (OFED) verbs interface as the software interface between application layer and hardware. RoCE takes advantage of transport services support of various modes of communication, such as reliable connected services and datagram services. RoCE uses well defined verbs operations including kernel bypass, send/receive semantics, RDMA read/write, user-level multicast, user level I/O access, zero copy and atomic operations. It also takes advantage of Data Center Bridging (DCB) i.e. Priority Based Flow Control, Enhanced Transmission Selection, Congestion Control and Data Center Bridging Exchange (DCBX) Protocol.

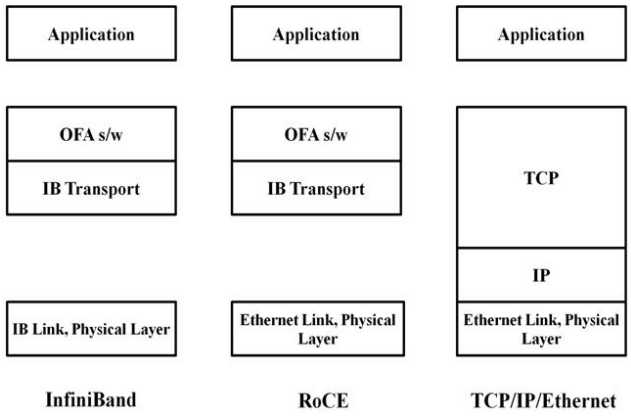

Figure 1: Three Different Interconnect Technologies

Conceptually, RoCE is simple enough, but there is a subtlety that is easy to overlook. Many of us, when we think of Ethernet, naturally envisage the complete IP architecture consisting of TCP, IP and Ethernet. But the truth is that RoCE bears no relationship to traditional TCP/IP/Ethernet, even though it uses an Ethernet layer. The Figure 1 also compares the two RDMA technologies to traditional TCP/IP/Ethernet. As the Figure makes clear, RoCE and InfiniBand are sibling technologies, but are only distant cousins to TCP/IP/Ethernet. Indeed, RoCE's heritage is found in the basic InfiniBand architecture and is fully supported by the open source software stacks provided by the Open Fabrics Alliance. Few key benefits and key features of RoCE are defined as follows.

2.1 Key Benefits

- RoCE-The new specification, pronounced "Rocky," provides the best of both worlds: "InfiniBand Efficiency and Ethernet Ubiquity"**.**
- It uses Ethernet switched Fabric instead of InfiniBand adapters & switches.
- It delivers the advantages of RDMA, such as lower latency or improved CPU utilization. Many latency-sensitive applications have been ported to run over RoCE and RoCE has been already deployed in mainstream data centers.
- RoCE end user benefits include improved application performance, efficiency, and cost and power savings.
- RoCE can be implemented in both hardware & software. So it can run anywhere.
- RoCE based network management is the same as that for any Ethernet, eliminating the need for IT managers to learn new technologies.

2.2 Key Features

- Take advantage of **DCB** Ethernet. It is also called Converged Ethernet.
- It supports these IEEE standards.
  - ✓ **802.1Qbb-**Priority Flow Control
  - ✓ **802.1Qaz**- Enhanced Transmission Selection
  - ✓ **802.1Qau**- Congestion Negotiation
- Traffic classification at layer 2 improves network efficiency.
- Lowest latency of 1.3 microseconds on lossless Ethernet
- RoCE is basically the InfiniBand protocols made to work over Ethernet infrastructure.
- RoCE focuses on server-to-server and server-to-storage networks, delivering the lowest latency and jitter characteristics and enabling simpler software and hardware implementations
- RoCE supports the OFA verbs interface seamlessly. The OFA verbs used by RoCE have been proven in large-scale deployments and with multiple Independent Software Vendor (ISV) applications in High Performance Computing (HPC) and Enterprise Data Center (EDC) sectors. Such applications can now be seamlessly offered over RoCE without any porting effort required.

3. Soft RoCE

RoCE can also be implemented in software named Soft RoCE. It is a software-based RoCE Linux driver called Soft RoCE is provided by System Fabric Works (SFW). It is an open source IB transport and network layers in software over ordinary Ethernet. It interoperates with hardware RoCE at other end of wire. rxe_cfg is the configuration tool for the RXE software implementation of the RoCE protocol. The RXE software is presently available in a special OFED-1.5.2 Distribution from System Fabric Works (SFW).

2. EXPERIMENTAL SETUP

In this section, we have reported the performance comparison of Soft RoCE & Ethernet over 1 gigabit Ethernet network adapter using OSU Micro Benchmark. To perform the Benchmark evaluation, a setup required to be designed. This setup consists of a the heterogeneous Linux cluster design consists of 2 nodes having Intel's i3 core 2.13 GHz & Intel's i5 core 2.67 GHz processors. The Operating system running on both the Nodes are SUSE's Linux Operating System i.e. SLES 11 SP 1 with kernel

version 2.6.32.12-0.7 (x86_64). Each node is equipped with a Realtek PCIe network adapter with the connection speed of up to 1 gigabit. The MTU used for is 1500 bytes. OFED's Soft RoCE Distribution version 1.5.2 (System Fabrics Works (SFW) offers a new mechanism in its OFED release of supporting RDMA over Ethernet). We have used MVAPICH2 i.e. MPI platform for our experiments. We have used Ohio State's MPI Benchmark (OMB) to run the various experiments. Secondly, a detailed performance evaluation of Soft RoCE & Ethernet we use OMB benchmark for measuring the performance of Soft RoCE & Ethernet over 1gigabit network adapter. To provide more close by look at the communication behaviour of the two MPI Implementations, we have used a set of micro benchmarks. They include a basic set of performance metrics like latency, bandwidth, host overhead and throughput. The results are the average of the ten test runs for all cases. In addition we use OMB benchmarks to characterize the few aspects of the MPI implementation.

- Point to point Communication.
- Performance evaluation of collective communication.
- One sided Communication

3. Results and Discussion

In OSU Micro Benchmark, there are three categories of benchmark and these benchmarks are used to measure the performance of Soft RoCE over 1 Gigabit Ethernet network adapter. The three categories are Point to Point Benchmarks, Collective Benchmarks and One Sided Benchmarks. We have focused on the Point to point Benchmarks and Collective Benchmarks.

3.1 Point to Point MPI Benchmark

In point to point communication, one process sends a message and a second process receives it. A number of important MPI functions involve communication between two specific processes. A popular example is MPI_Send, which allows one specified process to send a message to a second specified process. In Point to Point MPI Benchmark category, we have used Bandwidth test, Latency test and Bidirectional latency test.

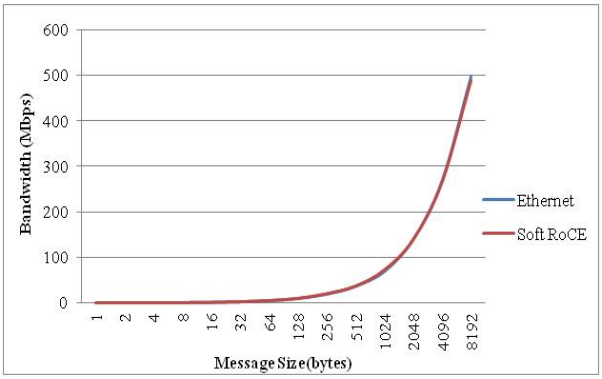

Figure 2: MPI_Bandwidth Test

In Figure 2, we have used OSU Bandwidth test and these tests were carried out by having the sender sending out a fixed number (equal to the window size) of back-to-back messages to the receiver and then waiting for a reply from the receiver. The receiver sends the reply only after receiving all these messages [4]. This process is repeated for several iterations and the bandwidth is calculated based on the elapsed time and the number of bytes sent by the sender. It is seen in the figure 2, that soft RoCE and Ethernet are performing almost same for message size 8192 bytes.

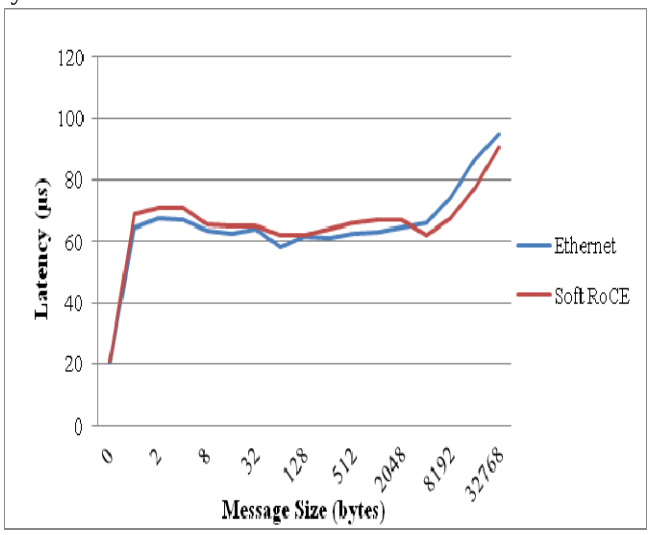

Figure 3: MPI_Latency Test

In Figure 3, we have used OSU latency tests are carried out in a ping-pong fashion. The sender sends a message with a certain data size to the receiver and waits for a reply from the receiver. The receiver receives the message from the sender and sends back a reply with the same data size. Many iterations of this test are carried out and average one-way latency numbers are obtained. Here, in this comparison Ethernet is performing better than Soft RoCE for message size upto 2 Kbytes but afterwards the performance of the Soft RoCE starts increasing from 4Kbytes to 32Kbytes.

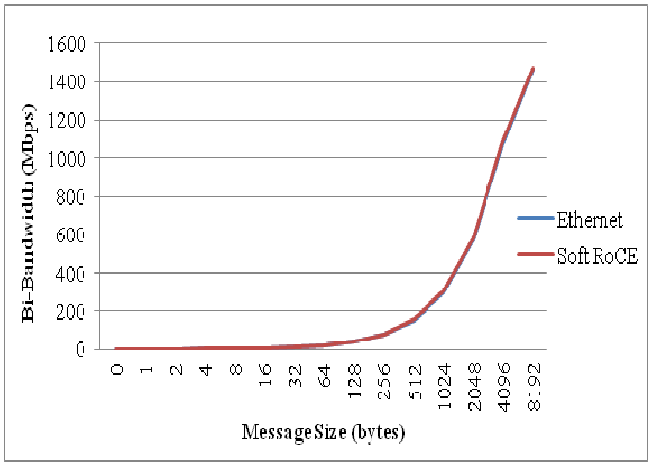

Figure 4: MPI_Bidirectional Bandwidth Test

In Figure 4, we have used the bidirectional bandwidth test which is similar to the bandwidth test, except that both the nodes involved send out a fixed number of back-to-back messages and wait for the reply [4]. This test measures the maximum sustainable aggregate bandwidth by two nodes. Here, in bidirectional bandwidth test Soft RoCE and Ethernet are performing almost same upto 8192 bytes.

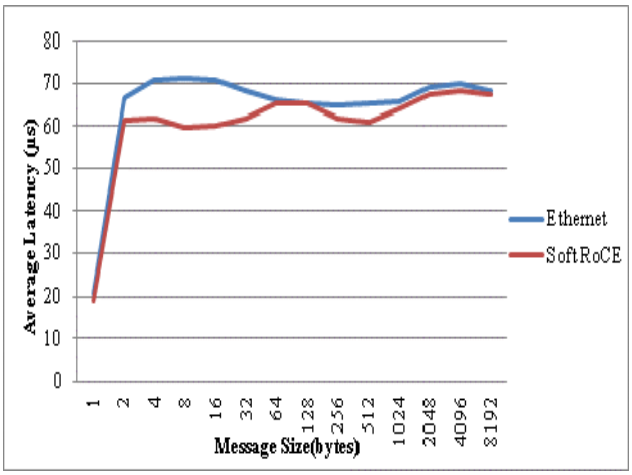

Figure 5:MPI_Multi_Latency Test

In Figure 5, we have used Multi Latency test and this test is very similar to the latency test. However, the only difference is that at the same instant multiple pairs are performing the same test simultaneously. In order to perform the test across just two nodes the hostnames must be specified in block fashion. Using the multi_latency test provide the complete communications latency output for message sizes for 1, 2, NCORE/2, and NCORE where NCORE equals the total number of assignable cores on the node.

3.2 Collective MPI Benchmark

The OSU collective benchmarks report the average communication latencies for a given collective operation, across various message lengths. These benchmarks report the average latency for each message length [3]. In this category we used broadcast test, allreduce test, gatherv test and reduce_scatter test. All tests are used to measure the average latency with different ways.

Figure 6: MPI_Broadcast Test

In Figure 6, we used Broadcast test it is a data movement operation. In this test root broadcasts or sends a message to all other processes in the group. Here, as shown in figure 5 the average time to deliver a message for soft RoCE is low as compared to Ethernet. At message size 512 bytes the performance decreases for Soft RoCE but again it starts improving at message size 1024 bytes.

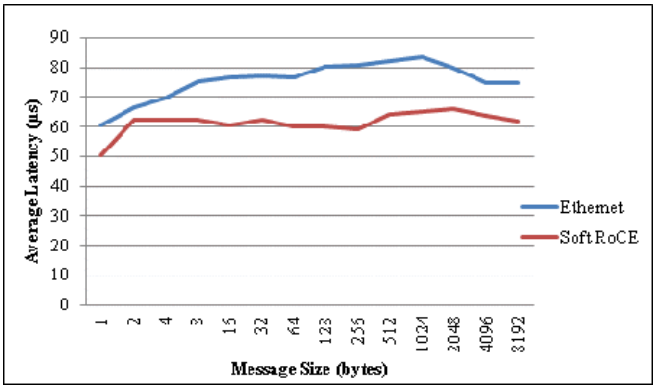

Figure 7: MPI_Gatherv Test

In Figure 7, we have used the Gatherv test and it gathers or collects varying amounts of data from all processes to the root process. Each process including the root process, sends a message to the root, and the root executes n receives. As shown in figure, the performance of Soft RoCE is better than the Ethernet. The performance difference (>=10µs) is maintained from the 4 bytes till the 8192 bytes message size.

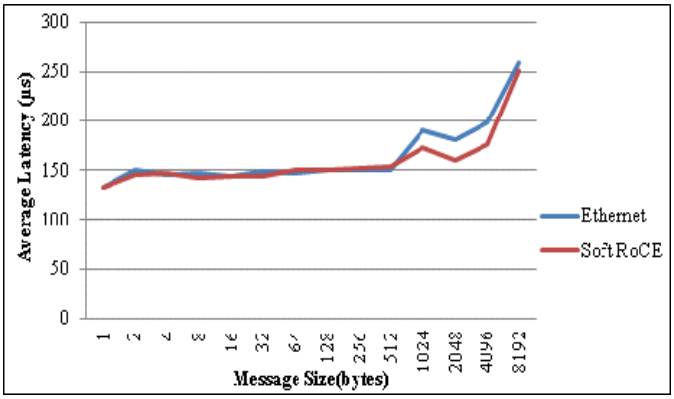

Figure 8:MPI_ Allreduce Test

In Figure 8, we have used Allreduce Test. It is a collective data movement routine. This test combines the values from all processes and distributes the results back to all the processes. This test is equivalent to an MPI_Reduce followed by an MPI_Bcast. Here also the Soft RoCE is performing better than Ethernet for small messages (1byte to 64 bytes) and again the performance gap starts increasing from 1024 bytes afterwards.

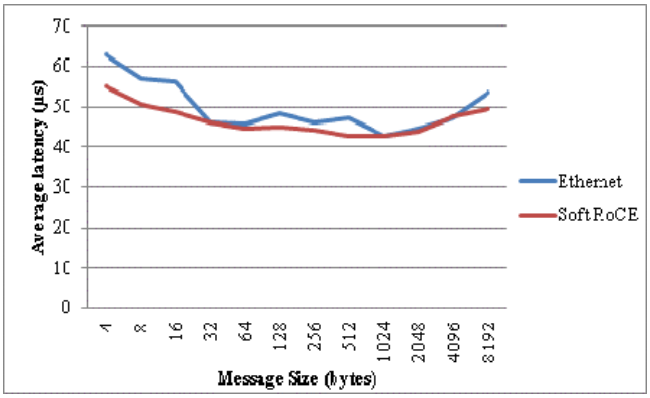

Figure 9: Reduce_Scatter Test

In Figure 9, we have used Reduce_Scatter Test and it is collective data movement. In this test each process sends a number to the root process and the total number is calculated by the root process then the root process sends a message to all processes. The size of the message equals to the chosen message size * number of processes. This test is equivalent to an MPI_Reduce followed by an MPI_Scatter operation. Here in the figure, it is seen that the latency for the small message size starts at higher side but as the message size increases the performance goes on increasing for both Soft RoCE and Ethernet. Besides this, Soft RoCE is performing better than the Ethernet in this test also.

CONCLUSION

This paper presents the Linux cluster configuration & evaluates its performance using Ohio State University (OSU) Micro Benchmark and we have evaluated the performance of Soft RoCE against the conventional Ethernet over the most commonly available 1 gigabit network adapter. In the meantime, it is publicized that the Soft RoCE showed varying performance gain in most of the cases over the conventional Ethernet.

REFERENCES

[1] Basem Madani, raed al-Shaikh, "Performance Benchmark and MPI Evaluation Using Westmere-based Infiniband HPC cluster", IJSSST, *Volume 12, Number 1, page no 20- 26, Feb. 2011*

[2] Pranabesh Kanti Thander1, Alpana Rajan, Anil Rawat, "Analysis of MPI Variants on Cluster with InfiniBand interconnect" in SACET-09, Oct., 28-29, 2009

[3] The InfiniBand website. [Online] available at blog.infinibandta.org

[4] The mvapich website. [Online] available at http://mvapich.cse.ohio-state.edu