IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 15, Issue 4 (Nov. - Dec. 2013), PP 81-87
www.iosrjournals.org

Performance Evaluation of Soft RoCE over 1 Gigabit Ethernet

Gurkirat Kaur, Manoj Kumar[1], Manju Bala
Department of Computer Science & Engineering, CTIEMT Jalandhar, Punjab, India
Department of Electronics and Communication Engineering, CTIEMT, Jalandhar, Punjab, India [1]

Abstract: *Ethernet is most influential & widely used technology in the world. With the growing demand of low latency & high throughput technologies like InfiniBand and RoCE have evolved with unique features viz. RDMA (Remote Direct Memory Access). RDMA is an effective technology, which is used for reducing system load & improves the performance. InfiniBand is a well known technology, which provides high-bandwidth and low-latency and makes optimal use of in-built features like RDMA. With the rapid evolution of InfiniBand technology and Ethernet lacking the RDMA and zero copy protocol, the Ethernet community has came out with a new enhancements that bridges the gap between InfiniBand and Ethernet. By adding the RDMA and zero copy protocol to the Ethernet a new networking technology is evolved called RDMA over Converged Ethernet (RoCE). RoCE is a standard released by the IBTA standardization body to define RDMA protocol over Ethernet. With the emergence of lossless Ethernet, RoCE uses InfiniBand efficient transport to provide the platform for deploying RDMA technology in mainstream data centres over 10GigE, 40GigE and beyond. RoCE provide all of the InfiniBand benefits transport benefits and well established RDMA ecosystem combined with converged Ethernet. 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 Intel's MPI Benchmarks. Our result shows that Soft RoCE is performing better than Ethernet in various performance metrics like bandwidth, latency & throughput.*

Keywords: *Ethernet, InfiniBand, MPI, RoCE, RDMA, Soft RoCE*

I. Introduction

There are various network interconnects, which provide low latency and high bandwidth. A few of them are Myrinet Quadrics, InfiniBand. In the high performance computing area, MPI is the de facto standard for writing parallel applications [1]. InfiniBand is scalable & high speed interconnects architecture that has very low lo latencies. InfiniBand provides high speed connectivity solutions between servers and switches that can be implemented alongside Ethernet & fibre channel in academic, scientific &financial High Performance Computing (HPC) environments [2]. RDMA over Converged Ethernet (RoCE) is a network protocol that allows Remote Direct Memory Access (RDMA) over an Ethernet network. RoCE is a link layer protocol and hence allows communication between any two hosts in the same Ethernet broadcast medium. Although the RoCE protocol benefits from the characteristics of a converged Ethernet network, the protocol can also be used on a traditional or non-converged Ethernet network [3] .Soft RoCE is the software implementation of RoCE. In this paper, we provide a performance comparison of MPI implementation over Soft RoCE and Gigabit Ethernet .The rest of the paper is organised as follows Section 1 gives the Introduction & overview of interconnects & MPI Implementation, InfiniBand and RoCE, RDMA protocol. Section 2 presents the Experimental setup & results of Soft RoCE and Gigabit Ethernet using IMB benchmark. Finally, Section 3 concludes the paper.

Interconnects & 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 & MPI Implementation.

1.1Ethernet Interconnect

In computer networking, Gigabit Ethernet (GbE or 1 GigE) is a term describing various technologies for transmitting Ethernet frames at a rate of a gigabit per second (1,000,000,000 bits per second), as defined by the IEEE 802.3-2008 standard [4]. TCP/IP is a communication protocol & can provide reliable, ordered, error checked delivery of a stream of bits between the programs running on the computers connected to a local area network. The TCP path support TCP message communication over the Ethernet using the socket interface to the operating system's network protocol stack. TCP utilizes events to manage the message communication. There are 3 types of events occurs: read event, write event, exception event. A read event occurs when there is an incoming message or data coming to the socket. A write event occurs when there is additional space available in the socket to send the data or message. An exception event occurs when any unexpected or exceptional condition occur in the socket [5].

1.2 Micro Benchmark- Intel MPI

Intel MPI is a multi fabric message passing library that is based on message passing interface, v2 (MPI-2) specifications, Intel MPI library focus on making the application to perform better on the Intel Architecture based cluster. This MPI implementation enables the developers to upgrade or to change the processors & interconnects as new technology become available without doing changes in the software or the operating system environment. This benchmark provides an efficient way to measure the performance of a cluster, including node performance, network latency and throughput.IMB 3.2.4 is categorized into 3 parts: IMB-MPI1, IMB-EXT, and IMB-IO. We will focus on the IMB-MPI1 which is used in our evaluation. The IMB-MPI1 benchmarks are classified into 3 classes of benchmarks: Single Transfer Benchmark, Parallel Transfer Benchmark & Collective Transfer Benchmark [6].

InfiniBand, RoCE

The InfiniBand emerged in 1999 by joining the two competitive technologies i.e. Future I/O and Next Generation I/O. The resultant of the merger is the InfiniBand architecture, all rooted on the virtual Interface Architecture, VIA. The Virtual Interface Architecture is based on two collective concepts: direct access to a network interface straight from application space, and ability for applications to exchange data directly between their own virtual buffers across a network, all without involving the operating system. The idea behind InfiniBand is simple; it provides applications a - messaging service. This service is easy to use & help the applications to communicate with other applications or processors. InfiniBand Architecture gives every application a direct access to the messaging service. Direct access means that the applications need not to rely on the operating system for the transfer of data or messages [7].

As depicted in the Figure 1, the network protocols can broadly be divided into two categories: Socket based and Verbs based. Data transfer depends upon the type of interconnect being used. InfiniBand or High Speed Ethernet (HSE), both is further sub-divided into two. If the interconnect is IB the socket interface uses the IPoIB driver available with OFED stack and the verb interface will use the native IB verbs driver for the IB Host Channel Adapter (HCA) being used. If the interconnect is HSE or Gigabit Ethernet, the sockets interface uses the generic Ethernet driver and the verbs interface uses the RoCE driver available with OFED stack. With the initiation of Converged Enhanced Ethernet, a new option to use a non-IP-based transport option is available, which is called RDMA over Converged Ethernet (RoCE, pronounced as "rock-ie"). RoCE is a new protocol that allows performing native InfiniBand communication effortlessly over lossless Ethernet links. RoCE packets are encapsulated into standard Ethernet frames with an IEEE assigned Ethertype, a Global Routing Header (GRH), unmodified InfiniBand transport headers and payload [8]. RoCE provides all of the InfiniBand transport benefits and well-established RDMA ecosystem combined with all the CEE advantages i.e. Priority based Flow Control, Congestion Notification, Enhanced Transmission Selection (ETS) and Data Centre Bridging Protocol (DCBX). RoCE can be implemented in hardware as well as software. Soft RoCE is the software implementation of RoCE which combines the InfiniBand Efficiency & Ethernet's ubiquitous.

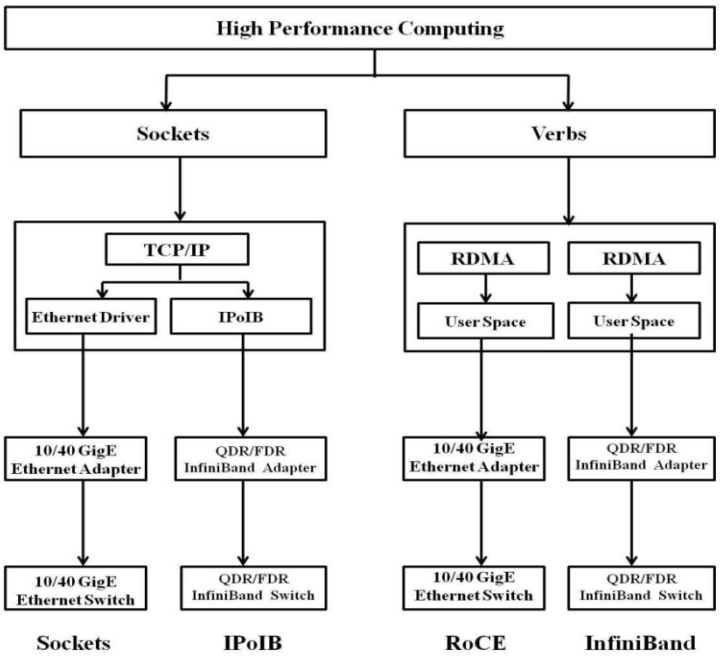

Figure 1: Network Protocol Stack

Remote Direct Memory Access

Remote Direct Memory Access (RDMA) is a network interface card (NIC) that lets one computer to place the data or information directly to the memory of another computer. This technology helps in reducing the latency by minimizing the bottleneck on bandwidth and processing overhead. Traditional architectures impose a considerable load on the server's CPU & memory because the data or information must be copied between the kernel & the application. Memory bottlenecks become severe when the connection speed increases the processing power & the memory bandwidth of servers. It supports zero copy networking with kernel bypass. Zero copy networking lets the NIC to transfer the data directly to and from the application memory, eliminating the need of copying the data between the kernel memory and the application memory [9]. RDMA also reduces the cost of data movement by eliminating the redundant copies through the network path & also improves overall resource utilization.

II. Experimental Setup

In this section, we have reported the performance comparison of Soft RoCE & Ethernet over 1 gigabit Ethernet network adapter using IMB 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 Intel's MPI Benchmark to run the various experiments. Secondly, a detailed performance evaluation of Soft RoCE & Ethernet over 1 gigabit Ethernet is done and then the comparisons between both are done using IMB Benchmark. To provide more close by look at the communication behaviour of the two MPI Implementations, we have used a set of micro benchmarks. They have included 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.

III. Result And Discussion

1. Intel MPI Benchmark

IMB 3.2.4 is the popular set of benchmarks which provides an efficient way to perform some of the important MPI functions. IMB benchmark consists of three types of functions: IMB-MPI1, IMB-EXT, and IMB-IO. We have focused on the IMB-MPI1 for our performance evaluation. The IMB-MPI1 benchmark introduces 3 classes of Benchmarks: Single Transfer Benchmark, Parallel Transfer Benchmark, and Collective Benchmark.

1.1 Single Transfer Benchmark

Here, we have focused on measuring the throughput on single message transferred between two processes. It involves two active processes between into communication. We have used two benchmarks in this category that is PingPong & PingPing for our performance evaluation.

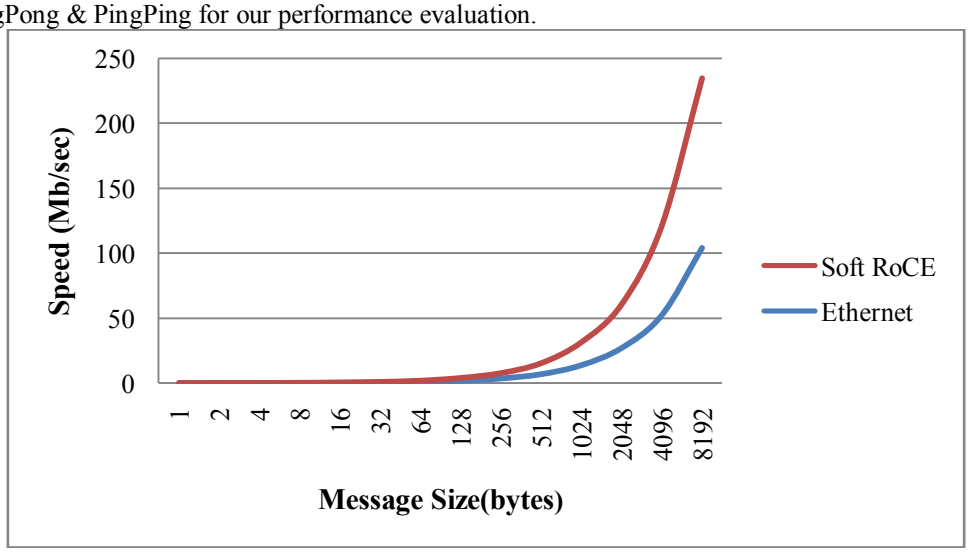

Figure 2: PingPong Throughput Test

In Figure 2, we have used IMB PingPong Test, which is a classic pattern for measuring the throughput of a single message sent between two active processes. In this test, we have compared the throughput of Ethernet interconnect & soft RoCE, which is the software implementation of RoCE Interface. They are almost same for small message sizes, but as the message size from 128 bytes, soft RoCE performance is better upto 8192 bytes.

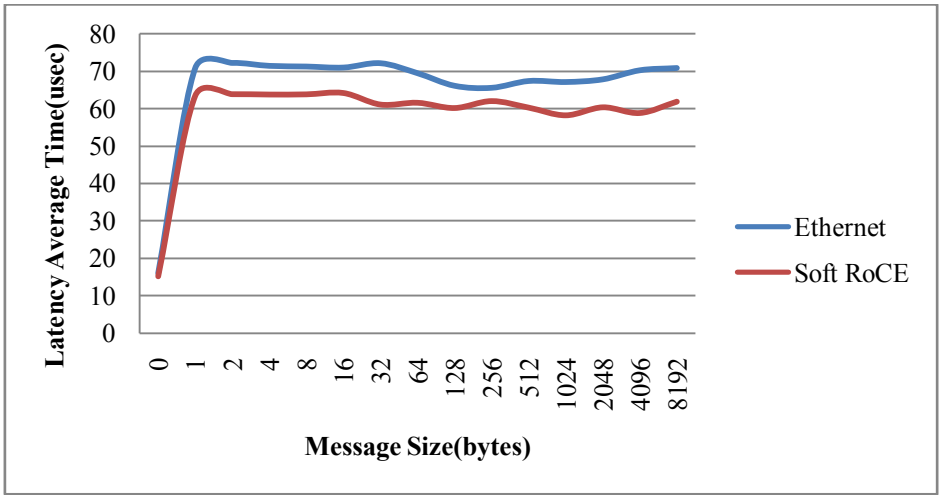

Figure 3: PingPong Latency Test

In Figure 3, we have used IMB PingPong Test, to measure the latency of Soft RoCE & Ethernet . Here in this comparison, we have seen that from the small message size Soft RoCE is performing better than the Ethernet and as the message size increases there is no turn down in the performance of Soft RoCE , which can be seen from 256 bytes.

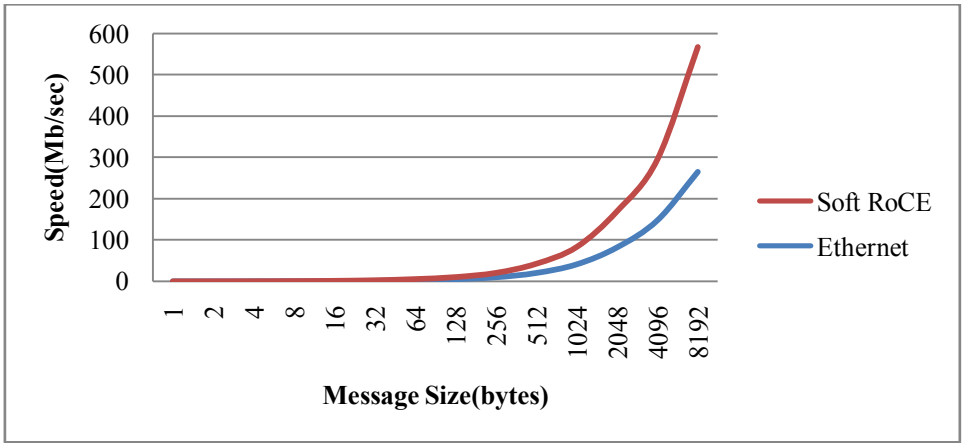

Figure 4: PingPing Throughput Test

In Figure 4, we have used IMB PingPing Test, it is same as PingPong test, and the only difference is that both the processes send the message to each other at same message size i.e. 256 bytes. In this test, we have evaluated the throughput of Ethernet interconnect & soft RoCE which is the software implementation of RoCE Interface. They are almost same upto 1k message size, but as the message size increases performance of the soft RoCE starts increasing.

1.2 Parallel Transfer Benchmark

Here, we have focused on calculating the throughput of simultaneous occurrence of the messages sent or received by a particular process in the periodic chain. We have used two benchmarks to evaluate of the performance of Sendrecv and Exchange benchmarks.

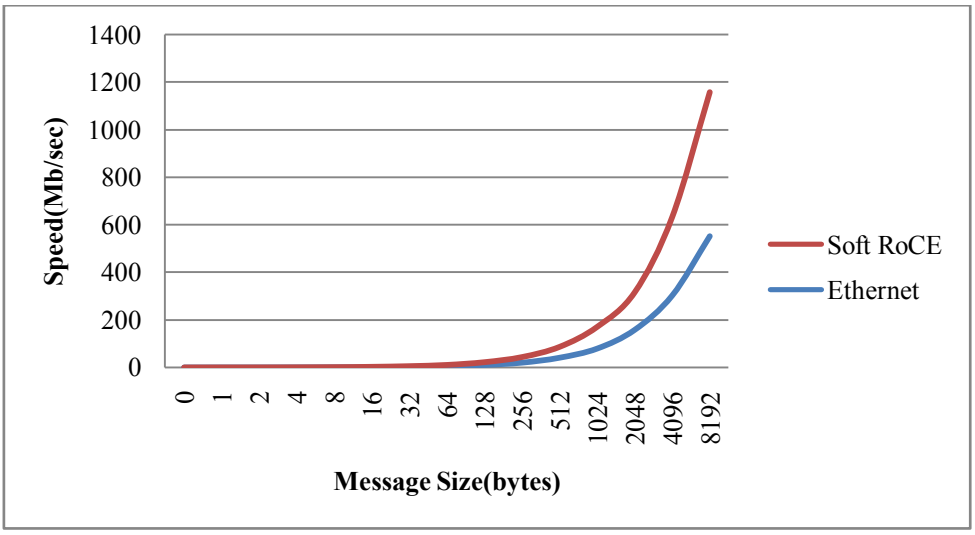

Figure 5: Sendrecv Test

In Figure 5, we have used IMB Sendrecv test each process send the message to its right & receives from its left neighbour in the periodic chain. The total turnover count is 2 messages per sample (1 in, 1 out). For small message size the performance is almost same but from message size 256 bytes onwards the Soft RoCE is performing faster than the Ethernet.

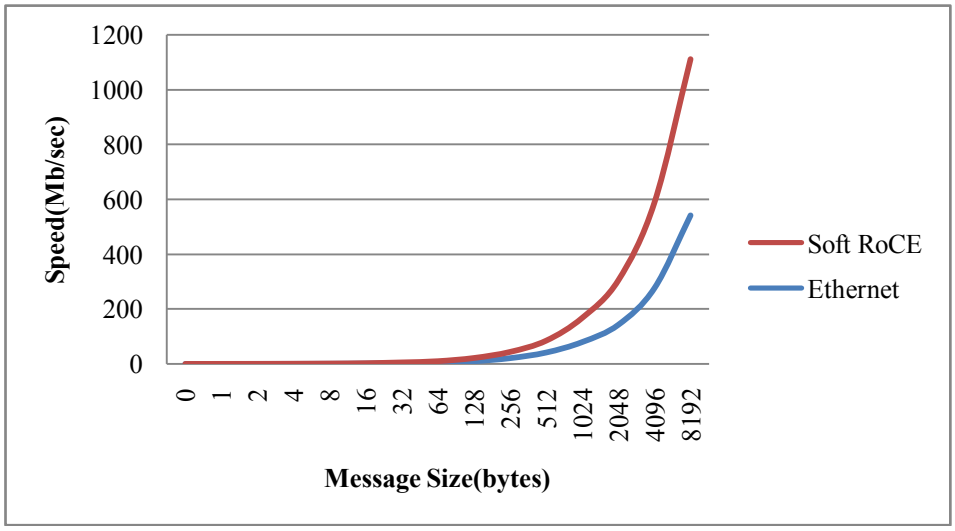

Figure 6: Exchange Test

In Figure 6, we have used IMB Exchange test in this test the processes are arranged in the periodic chain and each process exchange the message with both left and right neighbour in the periodic chain. The total turnover count is 4 messages per sample (2 in, 2 out). For small message size the performance is almost same but from message size 128 bytes onwards the performance of Soft RoCE is better than the Ethernet.

1.3 Collective Benchmark

Here, we have focused on MPI collective operations. This type of benchmark measure the time needed to communicate between a group of processes in different behaviour. There are several benchmarks come under this category. For our evaluation we have used Reduce, Gatherv and Scatter. In Reduce benchmark each process sends a number to the root & then total number will be calculated by the root.

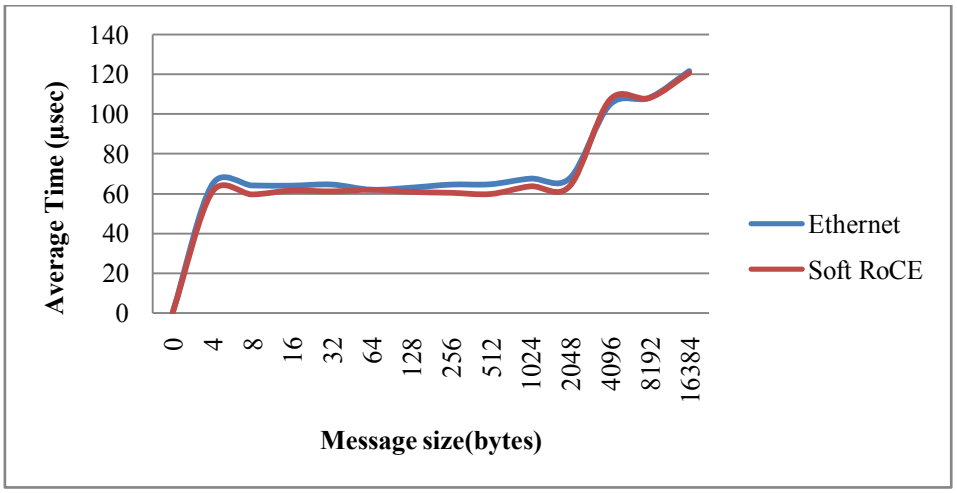

Figure 7: Reduce Test

In Figure 7, we have used Reduce Test, this is a collective MPI benchmarks which are used for collective operations. Reduce benchmark each process sends a number to the root & then total number will be calculated by the root. Soft RoCE is performing slightly better than Ethernet in message size < 4096 bytes. Afterwards both are performing slightly up and down.

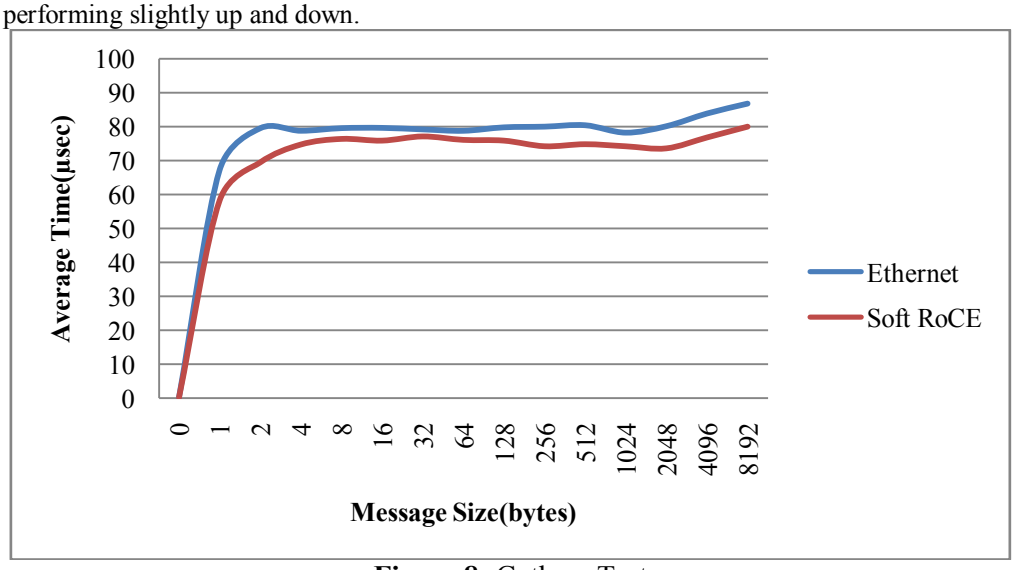

Figure 8: Gatherv Test

In Figure 8, we have used Gatherv test, in this test all process input *X* bytes & the root gather or collect *X* * *p* bytes where *p* is the number of processes. As shown in figure, Soft RoCE is performing faster than the Ethernet.

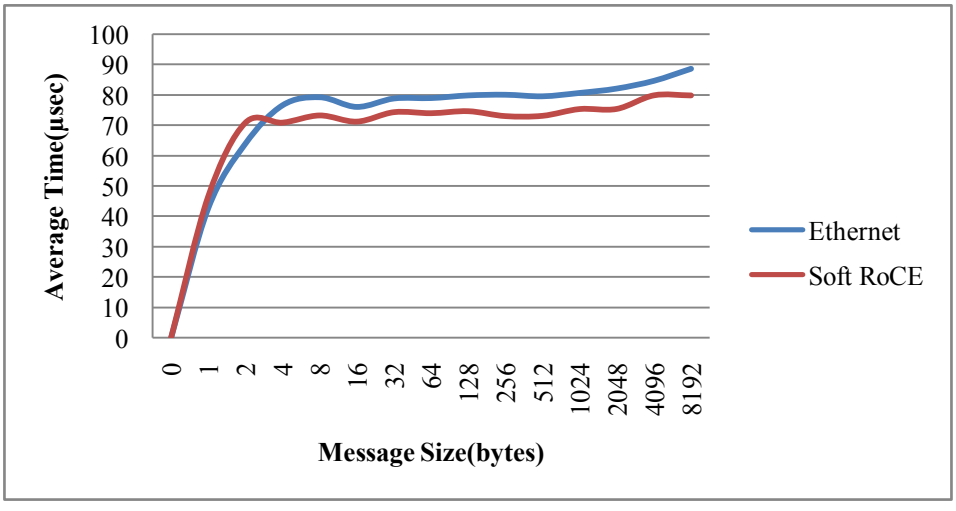

Figure 9: Scatter Test

In Figure 9, we have used Scatter benchmark, in this test the root process input the X * p bytes (X bytes for each process) & all process receives X bytes. Soft RoCE performs better when the message size exceeds 2 bytes.

IV. Conclusion & Future Scope

In recent trends High Performance Cluster (HPC) Systems has shown that in future the increase in the performance can only be achieved with the right combination of multi-cores processors &faster interconnects. This paper presents the Linux cluster configuration & evaluates its performance using Intel's MPI Benchmark and we have evaluated the performance of Soft RoCE against the conventional Ethernet over the most commonly available 1 gigabit network adapter. Meanwhile, it is shown that the Soft RoCE showed varying performance gain in every case over the conventional Ethernet. One day, one can foresee the RDMA capable Ethernet being provided by default on all the servers operating with a high data rates.

References

[1] Jiuxing Liu,Balasubramanian chandrasekaran, Jiesheng Wu, Weihang Jiang, Sushmitha Kini, Weikuan Yu,Darius Buntinas, Peter Wyckoff, DK Panda, "Performance Comparison of MPI Implementations over InfiniBand, Myrinet and Quadrics",*1-58113-695-1/03/0011 @ 2003 ACM*

[2] Panduit, "Introduction to InfiniBand", *2006@* InfiniBand Trade Association

[3] *(2013)*, The Wikipedia website.[Online] available at http://en.wikipedia.org/wiki/RDMA_over_Converged_Ethernet

[4] (*2013),* The Wikipedia website.[Online] available at http://en.wikipedia.org/wiki/Gigabit_Ethernet

[5] Supratik Majumder, Scott Rixner, "Comparing Ethernet and myrinet for mpi communication",in Proceedings of the 7th workshop on Workshop on languages, compilers, and run-time support for scalable systems , *10.1145/1066650.1066659 @ 2004 ACM*

[6] 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*

[7] Paul Grun, "Introduction to InfiniBand for End Users", *Copyright @ 2012* InfiniBand Trade Association

[8] Jerome Vienne, Jitong Chen,Md.Wasi-ur-Rahman,Nusrat S.Islam, Hari Subramoni, Dhabaleswar K.(DK) Panda, " Performance Analysis and Evaluation of InfiniBand FDR and 40GigE RoCE on HPC and Cloud Computing Systems", in 2012 IEEE 20th annual Symposium on High-Performance Interconnects, *978-0-7695-4831-9/12 @ 2012 IEEE*

[9] *(2003),* The networkworld Website(Online) available at http://www.networkworld.com/details/5221.html