International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 10, Issue 5 (May 2014), PP.26-35

Can Modern Interconnects Improve the Performance of Hadoop Cluster? Performance evaluation of Hadoop on SSD and HDD with IPoIB

Piyush Saxena
M.Tech (Computer Science and Engineering), Amity School of Engineering and Technology, Amity University, Noida, India +91-9451427546

Abstract:- In today's world where Internet is most required and where pentabytes of data is produced per hour, there is a drastic need to speed up the performance and throughput of the cloud system. Traditional cloud systems were not able to give the performance that the storage devices like SSD and HDD were meant to deliver.
In the last paper we showed that the hadoop on SSD and HDD did not showed much difference in performance as these were traditionally connected to the processing system that acts as a hindrance to the system. Another reason that could be spotted with the pattern of data access was that there was less of Random Access Memory with low caching resources available. To these issues, another set of experiments were conducted using a highly improved connecting method than the conventional 10 GigE and by implementing Distributed shared memory that can make the access patterns much faster. The improved methods that were considered for the test purposes were IPoIB and RDMA-IB.
In this paper we will also present that Modern Interconnects used in Hadoop (MapReduce) with SSD can outperform the traditional Interconnecting technique like 10 GigE networks. In addition, we also demonstrate that the use of sockets or conventional TCP/IP applications can be still used with new technology and with improved throughput and less latency when IBoIP is used.

Keywords:- Hadoop, HDFS, SSD, HDD, HiBench, Benchmarking, 10 GigE, IPoIB, RDMA-IB.

I. INTRODUCTION TO HADOOP AND HDFS

In today's digital age, a big measure of data is been processed on the internet. Allotting optimal data processing with advantageous response duration acts the output to the requests by the consumer. There are frequent users that assay to enter the alike data above the web and it is a challenging task for the server to deliver optimal result. The large amount of data the internet has to deal with every day has made conventional solutions extremely uneconomical. There are difficulties like processing large documents split into many disaffiliated sub-tasks that are segmented with the available nodes, and processed in parallel. Due to this, MapReduce and Hadoop came into existence.

Hadoop is a free-of-cost, programming architecture that is java-based and supports the processing of large amounts of applications on systems that have thousands of nodes and involves multiple pentabytes of data. The Hadoop Distributed File System helps faster data transfer rates between the nodes and makes the cluster to persist functioning performances uninterrupted in case of node failure. This system actually lowers the risk of complete system failure even when a significant no. of nodes are in-operative.[2]
Hadoop was motivated by MapReduce (Fig.1) that was introduced by Google, a software framework in which an application is broken down into numerous small parts. Any of these parts (also called fragments or blocks) can be run on any node in the cluster.[3]

MapReduce-based studies have been actively carried out for the efficient processing of big data on hadoop. Hadoop runs on clusters of computers that can handle large amounts of data and support distributed applications.[4] In the last few years, lots of research has been carried out to improve the performance of hadoop. One of the hindrances is the performance issues of the storage device used as it is connected to the system by a slower connecting interface like Bus. Even the difference in the Devices used for storage creates the hindrance.[5]

The performance of the Hadoop system is also bound on the type of workload that we consider. This is why we consider HiBench as the standard model for testing Hadoop Distributed File System (HDFS). In this paper, we try to study and evaluate the performance of Hadoop Distributed File System on a Hadoop Cluster system that contains flash memory based SSD (Solid State Drive) and Hard Disk Drive by optimizing each parameter on HiBench.

Technology has advanced fast, and datasets have grown even faster as it is easier to generate and incarcerate data. The large Big Data, are a warehouse of information. The primary challenge in the investigation of Big Data is to conquer the I/O blockage present on modern systems.[6] Lethargic I/O systems overpower the very use of having high end processors. They cannot provide data fast adequate to utilize all of the accessible processing power. Outcome of this is wastage of power and increases in the price of in commission large clusters. An approach is the use of Modern interconnects like IPoIB and RDMA-IB in place of Traditional Interconnects like 10 GigE.

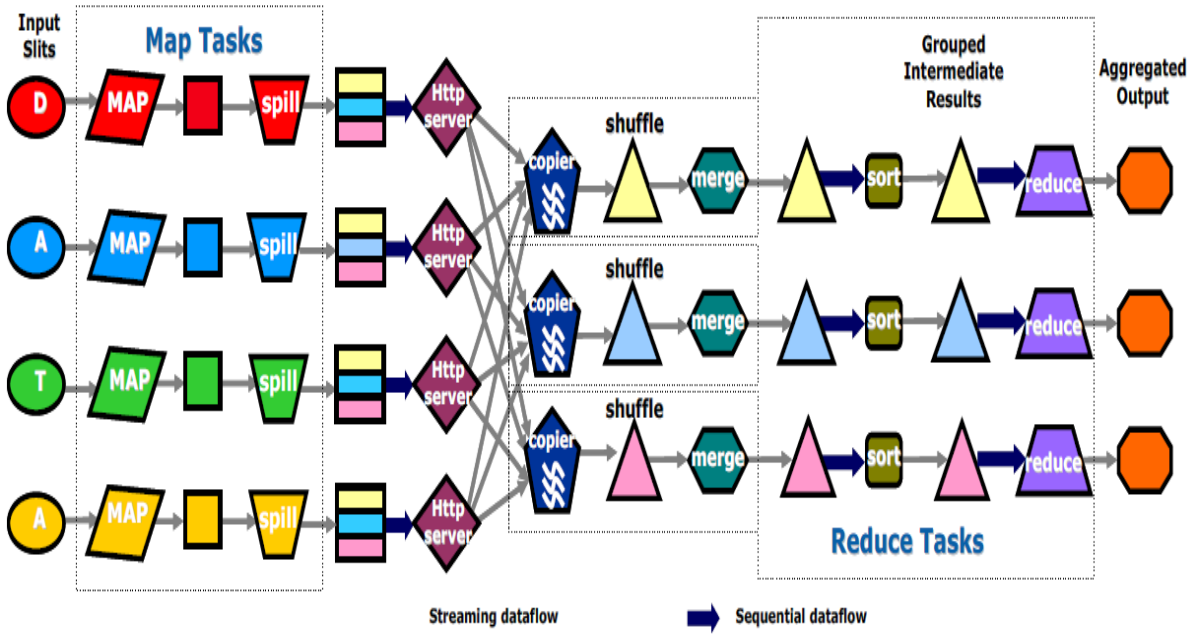

Fig 1.) Hadoop MapReduce Architecture

II. TRADITIONAL INTERCONNECT 10 GIGE NETWORK

10 gigabit Ethernet is a communication methodology that can give data transfer speeds up to 10 billion bits per second. 10 gigabit Ethernet is also known as 10GE, 10GbE or 10 GigE.

It supports full duplex connections that can be connected by network switches and shared medium operation with CSMA/CD.[7] It can work properly with the existing protocols. Since the 10 GigE works in full-duplex method, it doesn't need Carrier Sense Multiple Access/Collision Detection protocols that is extremely important as this improves the efficiency and the speed of 10 Gb Ethernet as it can be easily deployed in the existing network, thus giving a cost-efficient methodology that support high-speed, low-latency requirements.[8]

10-Gigabit Ethernet offers distances between physical locations up to 40 kilometers over a single-mode fiber and multi-mode fiber systems.

Technically 10 Gb Ethernet is a Layer 1 and Layer 2 protocol that follows the Ethernet attributes like Media Access Control (MAC) protocol, the Ethernet frame format, & min and max frame size. This technology supports both LAN and WAN standards. (Fig 2.)

Issues faced in deploying 10 Gb Ethernet are due to the costs of fibre channels, but the benefits received are very large.

III. MODERN INTERCONNECT IPOIB NETWORK

Infini Band (IB) [9] is a uniform organization regional Network that is applicable in HPC and data centre environments Infini Band Technology has high speed data transfer at a very low latency time. To allow the legacy IP based applications over Internet Protocol based apps over InfiniBand in Data Centers. Internet Protocol over InfiniBand protocol uses an interface on top of Infini band 'Verbs' Layer that allows the applications running on sockets to use host based TCP/IP protocol stack that is converted into native InfiniBand Verbs that looks invisible to the application. Sockets Direct Protocol (SDP) is a development of the sockets based boundary interface, allows the process to bypass the TCP/IP protocol stack and translate socket based packets into the verbs layer RDMA operations, still maintaining TCP streaming socket symbolism. [10]

SDP has the benefits of trespassing software layer that is required in IPoIB. The results of this are SDP has better latency and performance than IPoIB.

The uses of the InfiniBand are in modern computing and high performance computing. The benefits of IBoIP is reducing communication latency as well as providing higher available bandwidth to clients in the local DCN.

The administration of network load is of concern in the new networking technologies. Quality of Service provisioning could be used to control the traffic for intra-network loads that can have main concern over the input data stream (IDS). The traffic loads and flexibility in fine tuning of the performance of the network is also a bottle neck for the system wide performance. Such technology would boost the performance of traditional data centers that still work on Ethernet Best Effort Service with low or no requirement for modifying the conventional socket applications.

It is important to evaluate the behavior of the H/W level QOS provisioning for InfiniBand network with applications on the optimized socket based protocols. This yields in a step to use of this new technology to harness high-speed interconnects for existing Internet applications.

In this paper, we will analyze and see the performance improvements in case of modern interconnects like IPoIB in comparison of the traditional Bus interconnects or 10 GigE hardware.

InfiniBand is a prominent cluster interconnecting technology with very low latency and very high performance. Native InfiniBand verbs is the lowest software layer of the InfiniBand network that allows direct user-level access to IB Host Channel Adapter (HCA) resources by omitting the Operating System. At the IB verbs level, a queue pairing form is used for message underneath both Send/Receive and RDMA semantics. InfiniBand needs the user to register the buffer before using it for communication.

InfiniBand HCAs has 2 ports that can operate as 4X InfiniBand or 10-GigE. The architecture of HCA includes a stateless offload engine for network interface card (NIC) based protocol processing.

Sockets Direct Protocol was designed originally for InfiniBand that has now been redefined as a transport –agnostic protocol for RDMA network based fabrics. It was made known to improve and progress the performance of sockets by using the RDMA protocol of the InfiniBand network. SDP is a byte-stream protocol that is built on TCP stream socket connotations. SDP uses a protocol switch inside the operating system kernel that clearly alternates between kernel TCP/IP stack above IB (IPoIB) along with the SDP above IB (which sidesteps the kernel TCP/IP stack) [11].

SDP acquires bi-form layouts of data interchange. In the buffered-copy arrangement, the socket data is duplicated in a preregistered buffer foregoing the network transfer. In the zero-copy arrangement, the consumer buffer is lucidly registered for broadcasting to bypass data reproduction. (Fig 2.)

IV. MODERN INTERCONNECT RDMA-IB

InfiniBand Host Channel Adapters (HCA) and further network equipments can be approached by the upper layer software using an interface called Verbs. The verbs interface is a low level communication interface that follows the Queue Pair (or communication end-points) model.

Queue pairs are required to establish a channel between the two communicating entities. Each queue pair has a certain number of work queue elements. Upper-level software places a work request on the queue pair that is then processed by the HCA. When a work element is completed, it is placed in the completion queue. Upper level software can detect completion by polling the completion queue. Verbs that are used to transfer data are completely OS-bypassed. (**Fig 2**.)

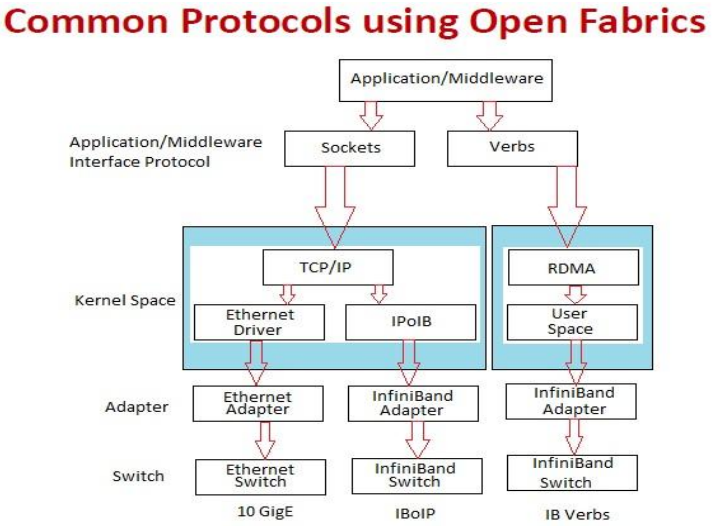

Fig 2.) Various Interconnect Technologies and architecture

V. TEST BED SYSTEM USED FOR THE ANALYSIS

4 node all 1U servers (Quanta Stack) with 2 Intel Xeon X5670 CPU's, each one has 6 cores that is equal to 12 physical cores with 96GB of Memory and Ethernet/ Infiniband as network. Storage Device 100 GB SSD and 2 TB HDD.

VI. CLASSIFICATION OF MICRO BENCHMARK WORKLOADS

1.) **Sort:** It is a representation of a large subset of real world MapReduce jobs that is transforming data from one representation to another. Sort requires an Input Output bound system resource utilization with the data access patterns as equal quantities of data access. The input data is generated using the *RandomTextWriter* program contained in the Hadoop distribution. Time taken by Reduce stage is twice the time taken by Map stage. **(Fig.3.1) [12]**

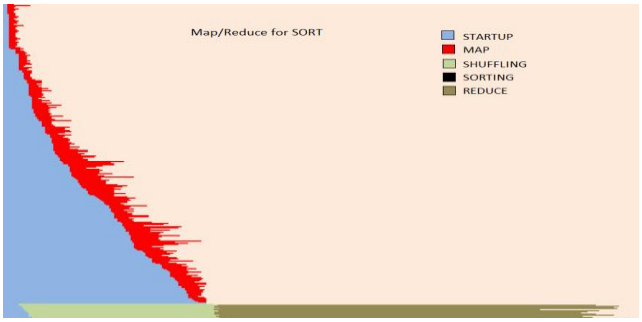

Fig 3.1) MapReduce for SORT workload

2.) **Word Count:** It is also a representation of a large subset of real world MapReduce jobs that is transforming data by extracting a small amount of interesting data from a large data set. Word Count requires a CPU bound system resource utilization with the data access patterns as reducing quantities of data access. The input data is generated using the *RandomTextWriter* program contained in the Hadoop distribution. Time taken by Reduce stage is nearly the same as the time taken by Map stage. **(Fig.3.2)**

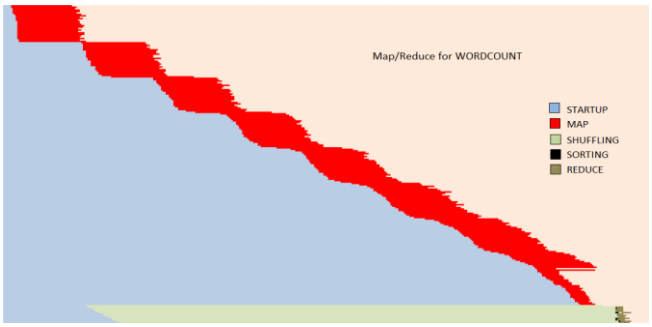

Fig 3.2) MapReduce for WORD COUNT Workload

3.) **TeraSort:** It sorts 10 billion 100-byte records generated by the *TeraGen* program contained in the Hadoop distribution. TeraSort requires CPU bound system resource utilization during Map stage and Input Output bound system resource utilization during Reduce stage with the data access patterns as reducing and then growing quantities of data access. Time taken by Reduce stage is 1.5 times the time taken by Map stage. **(Fig.3.3)**

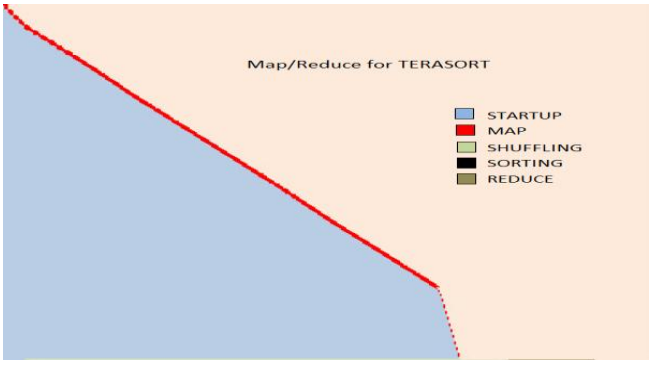

Fig 3.3) MapReduce for TERA SORT workload

VII. PERFORMANCE EVALUATION OF SSD AND HDD ON 10GIGE AND IPOIB

For the Performance evaluation and Analysis of the performance of SSD and HDD the considered work loads are Sort, Word Count and Tera Sort on two different workloads viz. 10 GigE and IBoIP. The size of data taken for all the workloads is 6550021992 bytes that is 6.1001GB of data. [13][14][15] (X-Axis -> Percentage ; Y-Axis -> Time in sec)

1) Sort Work Load: Since Sort has an Input Output bound resource utilization it is easily observed that SSD (Fig.4.1) buffers the data much earlier and at a faster rate than HDD (Fig.5.1) that tends to buffer at a constant speed. Due to this reason the SSD had an earlier chance to start off with the Reduce phase as compared to the HDD. It can also be inference from the graduated behavior of the graph that HDD works in a much stabilized manner as compared to the SDD. Over all SSD finishes off its job with the processors 39seconds earlier than the HDD. This proves that the SSD works much faster than HDD in the scenario of Sort Workload.

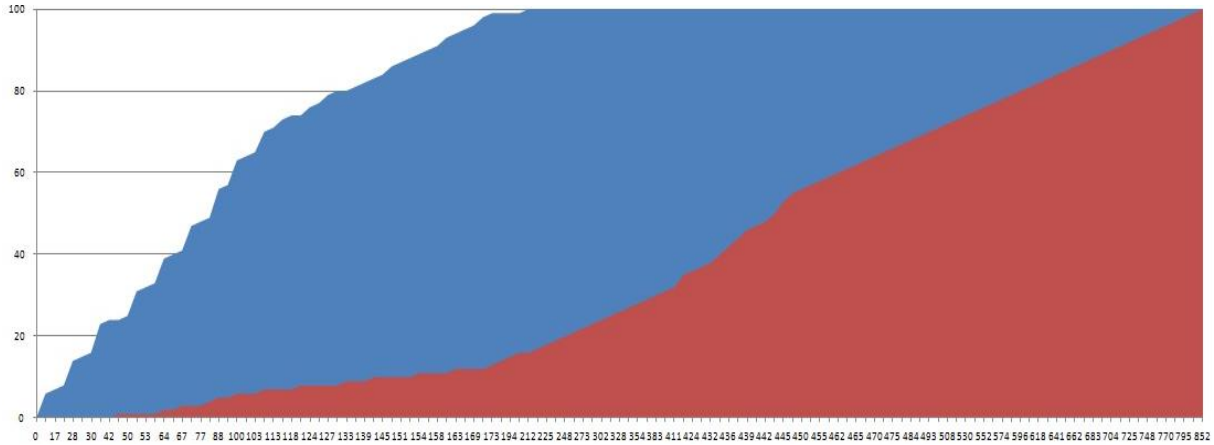

Fig 4.1.) SORT workload on Solid State Drive 10GigE (Blue -> Map Phase; Red -> Reduce Phase)

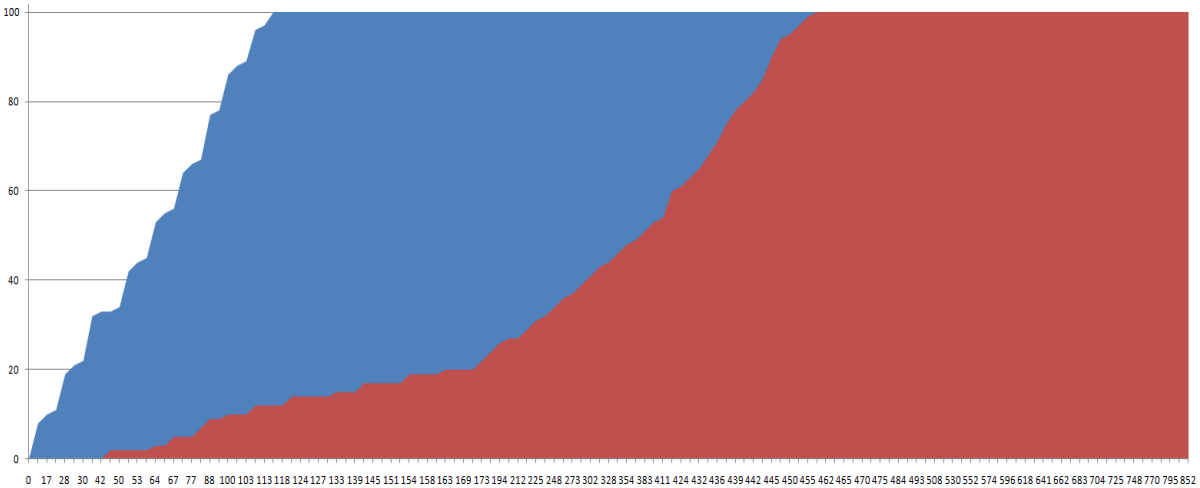

Fig 4.2.) SORT workload on Solid State Drive IPoIB (Blue -> Map Phase; Red -> Reduce Phase)

As of the performance change between the Modern Interconnect using IPoIB comapred to the Traditional Interconnect using 10 GigE can be analysed from the benchmarking results of SSD and HDD used on both types of interconnects. The analysis is as follows:

For SSD: the level of improvement is an average of 45% with precise improvement of 44% in Map Phase and 46% in Reduce Phase. The Map phase completed at 113sec in case of IBoIP as compared to 212sec in case of 10GigE. The Reduce phase completed at 455sec in IPoIB as compared to 852sec in case of 10GigE. The reduce phase started from 30% of map phase.

For HDD: the level of improvement is an average of 27% with precise improvement of 26% in Map Phase and 27% in Reduce Phase. The Map phase completed at 196sec in case of IBoIP as compared to 265sec in case of 10GigE. The Reduce phase completed at 699sec in IPoIB as compared to 953sec in case of 10GigE. The reduce phase started from 28% of map phase.

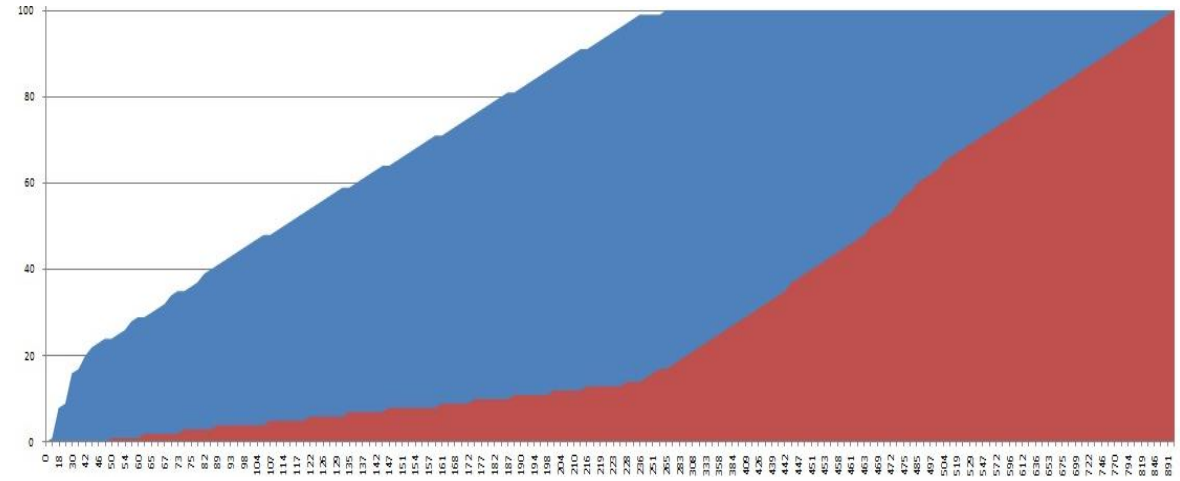

Fig 5.1) SORT workload on Hard Disk Drive 10GigE (Blue -> Map Phase; Red -> Reduce Phase)

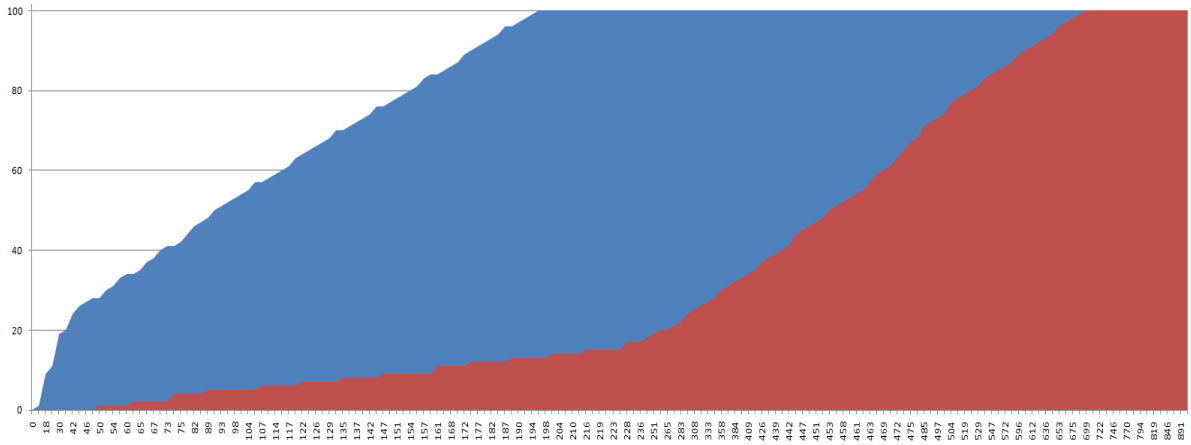

Fig 5.2) SORT workload on Hard Disk Drive IPoIB (Blue -> Map Phase; Red -> Reduce Phase)

2) <u>Word Count Work Load:</u> Since Sort has a CPU bound resource utilization it is easily observed that SSD (Fig.6.1) and HDD (Fig.7.1) both buffers approximately at the same rate but with a little variation in the speed as SSD buffers about 3 seconds faster than HDD. Due to this reason the SSD had an earlier chance to start off with the Reduce phase at 47 seconds as compared to the HDD that starts at 49 seconds. It can also be inferred from the abrupt behavior of the graph that HDD takes a longer time in the reduce phase as compared to the SSd that takes less time. Over all SSD finishes off its job with the processors 6seconds earlier than the HDD that is not a very major time difference. But, still this proves that the SSD works faster than HDD in the scenario of Word Count Workload.

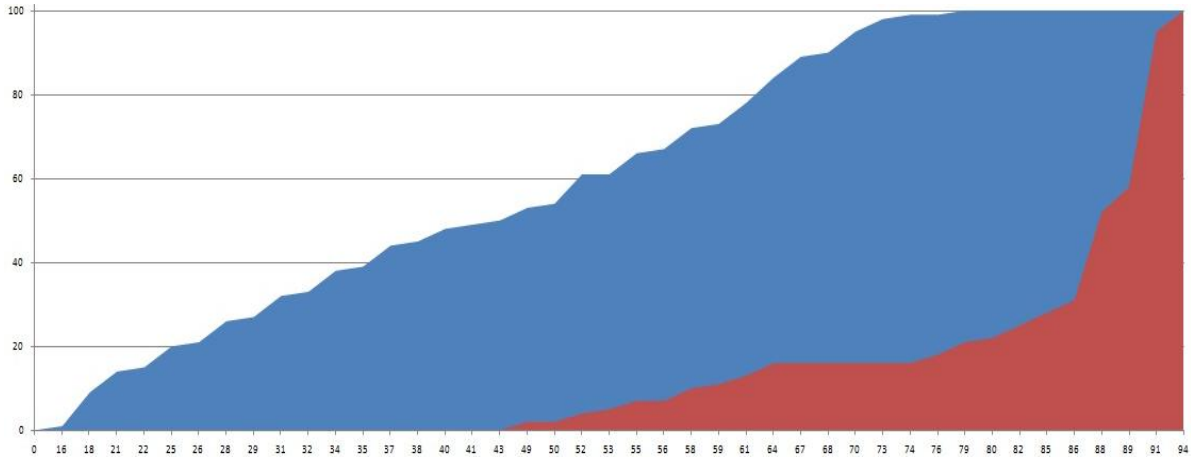

Fig 6.1) WORD COUNT Workload on Solid State Device 10GigE

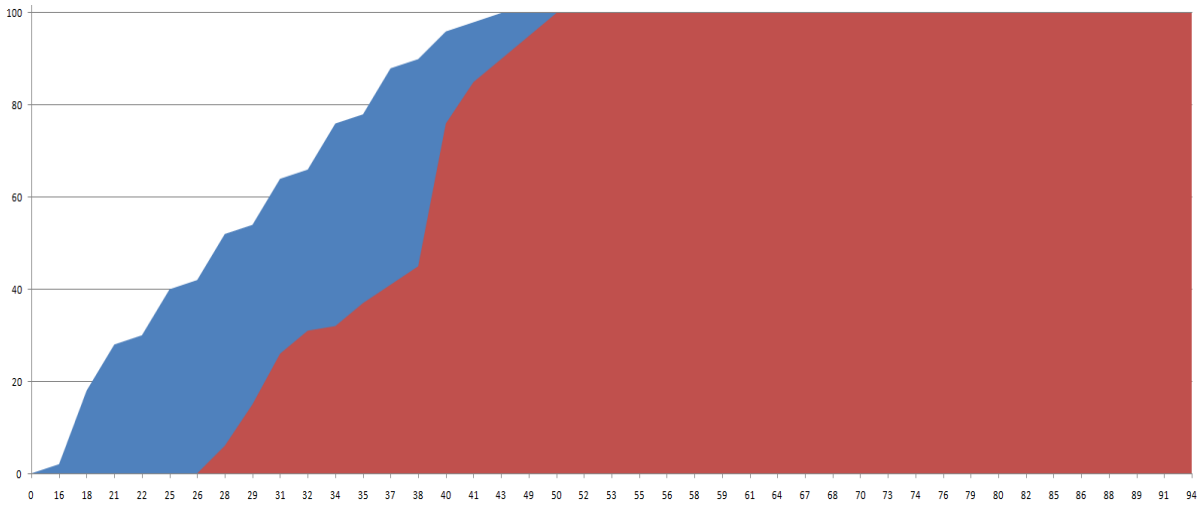

Fig 6.2) WORD COUNT Workload on Solid State Device IPoIB

As of the performance change between the Modern Interconnect using IPoIB comapred to the Traditional Interconnect using 10 GigE can be analysed from the benchmarking of SSD and HDD. The analysis is as follows:

For SSD: the level of improvement is an average of 46% with precise improvement of 45% in Map Phase and 47% in Reduce Phase. The Map phase completed at 43sec in case of IBoIP as compared to 79sec in case of 10GigE. The Reduce phase completed at 50sec in IPoIB as compared to 94sec in case of 10GigE. The reduce phase started from 44% of map phase.

For HDD: the level of improvement is an average of 29% with precise improvement of 26% in Map Phase and 33% in Reduce Phase. The Map phase completed at 64sec in case of IBoIP as compared to 86sec in case of 10GigE. The Reduce phase completed at 67sec in IPoIB as compared to 100sec in case of 10GigE. The reduce phase started from 65% of map phase.

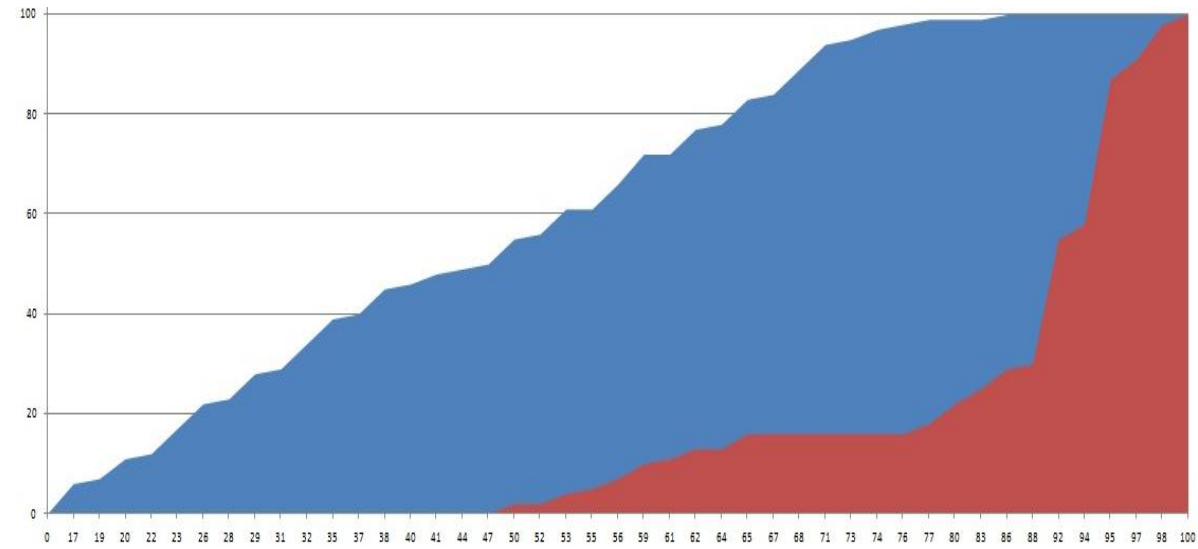

Fig 7.1) WORD COUNT Workload on Hard Disk Drive 10GigE

3) <u>TeraSort Work Load:</u> Since Sort has a CPU bound system resource utilization during Map stage and Input Output bound system resource utilization during Reduce stage it is easily observed that SSD (Fig.8.1) buffers the data much earlier 19Sec and at a faster rate than HDD (Fig.9.1) 21Sec that tends to buffer at an abrupt speed. Due to this reason the SSD had an earlier chance to start off with the Reduce at 23sec as compared to the HDD that starts at 24 second. It can also be observed that the reduce phase for SSD and HDD takes equal amount of time i.e. process is independent of SSD or HDD & dependent on processor. SSD finishes its job 1second earlier than the HDD that is not a negligible difference. But, still this proves that the SSD has lower latency than HDD in the scenario of Tera Sort Workload

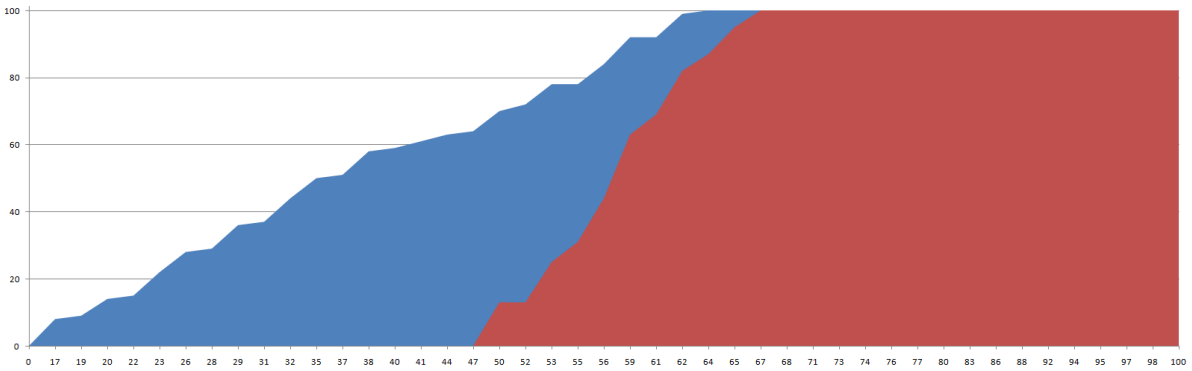

Fig 7.2) WORD COUNT Workload on Hard Disk Drive IPoIB

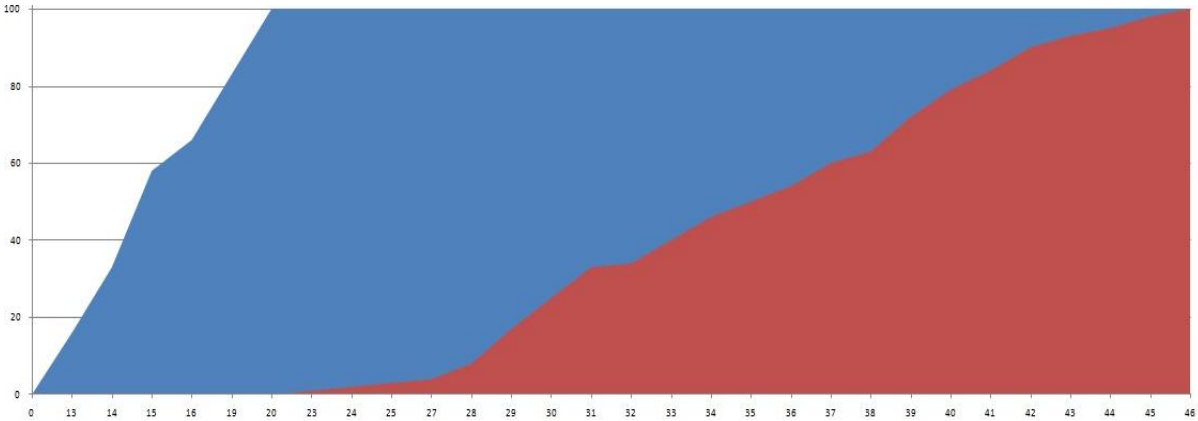

Fig 8.1) TERASORT workload on Solid State Drive 10GigE

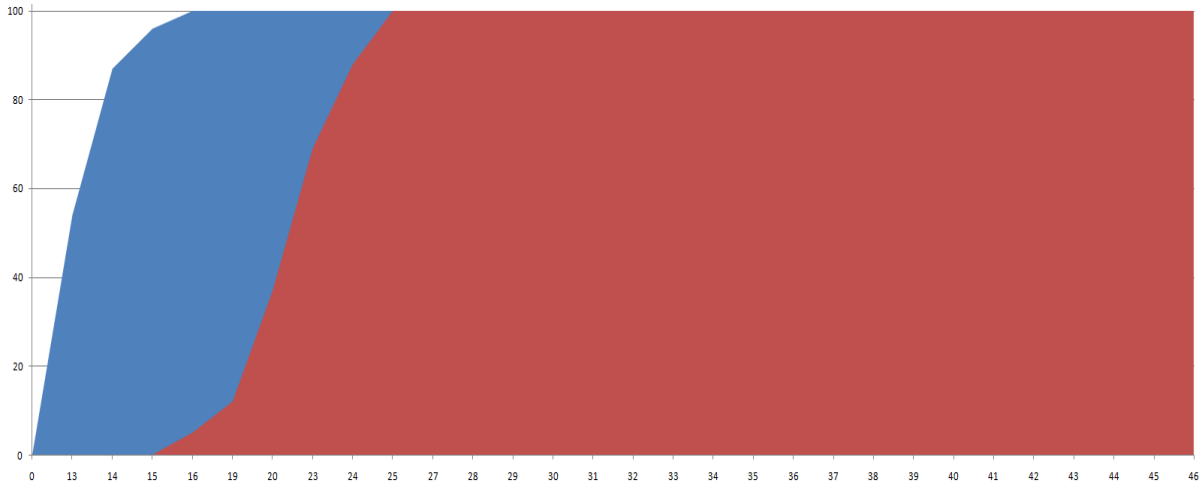

Fig 8.2) TERASORT workload on Solid State Drive IPoIB

As of the performance change between the Modern Interconnect using IPoIB comapred to the Traditional Interconnect using 10 GigE can be analysed from the benchmarking results of SSD and HDD used on both types of interconnects. The analysis is as follows:

For SSD: the level of improvement is an average of 44% with precise improvement of 41% in Map Phase and 46% in Reduce Phase. The Map phase completed at 14sec in case of IBoIP as compared to 28sec in case of 10GigE. The Reduce phase completed at 25sec in IPoIB as compared to 46sec in case of 10GigE. The reduce phase started from 98% of map phase.

For HDD: the level of improvement is an average of 25% with precise improvement of 24% in Map Phase and 26% in Reduce Phase. The Map phase completed at 16sec in case of IBoIP as compared to 21sec in

case of 10GigE. The Reduce phase completed at 34sec in IPoIB as compared to 46sec in case of 10GigE. The reduce phase started from 100% of map phase.

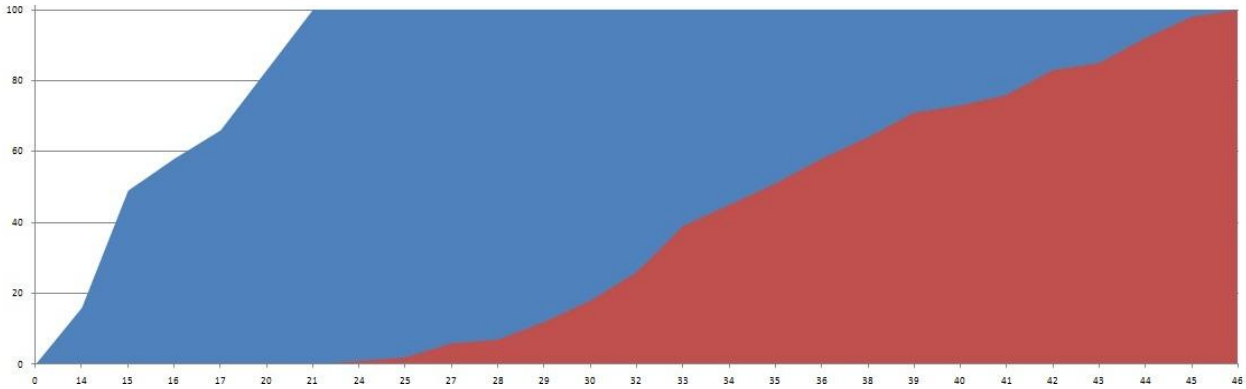

Fig 9.1) TERASORT workload on Hard Disk Drive 10GigE

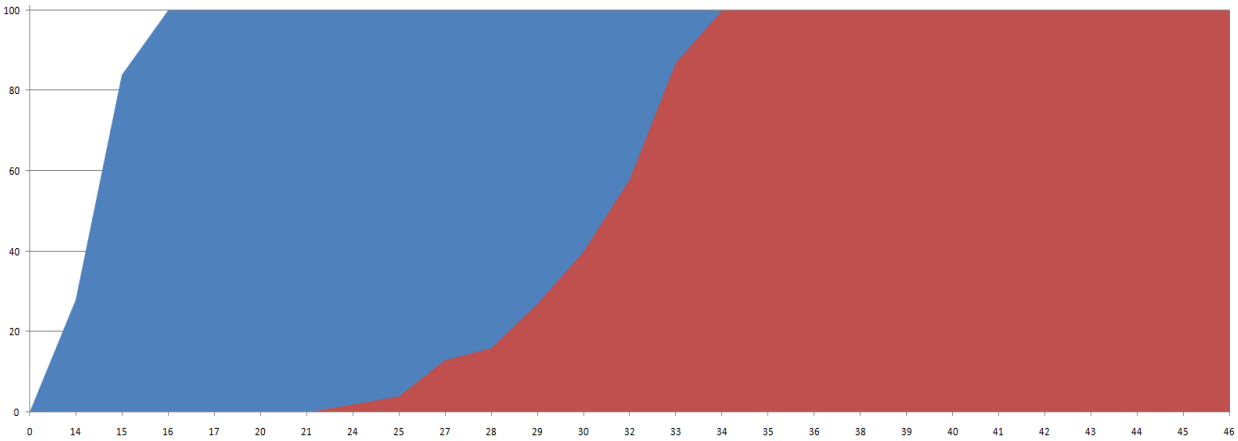

Fig 9.2) TERASORT workload on Hard Disk Drive IPoIB

VIII. CONCLUSION AND FUTURE SCOPE

From the above results and analysis the performance of SSD and HDD is nearly the same for the same Interconnect used, but positive results can be seen for better performance of SSD than HDD with use of IPoIB (Fig 10). Also the difference in the performance is very visible and drastic. So, an observation that can be monitored is that the Map phase in any of the workload is performing well until the random access memory is not consumed or the interconnect technology of the network used is of very high throughput and low latency. This concludes that there is a need to involve a Distributed Shared Memory (DSM) or the need to improve the Interconnect technology for networking from the traditional 10GigE to IBoIP to improve the performance of the SSD and HDD and get better significant results [16]. Another connection technique like InfiniBand using RDMA is used to connect then better performance in terms of latency, speed of access and fault tolerance can be achieved [10].

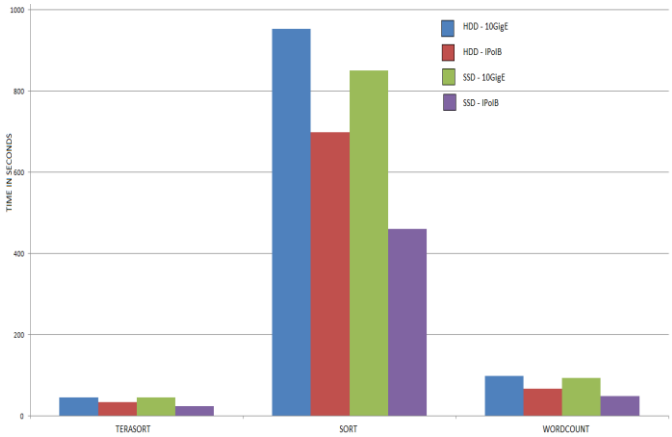

Fig 10) Comparison of performances of the Interconnect Technologies

In the future, a model to have DSM as a part should be used to be implemented with the use of InfiniBand on RDMA that supports the use of Verbs and with the technology of Optical Fibers to achieve faster performances and Remote Dynamic Memory Access. [17][18][19] The modern interconnects like IPoIB and RDMA-IB has got a lot of potential and their powers need to be researched and harnessed on in the future.[20]

REFERENCES

[1] Hadoop Home: http://hadoop.apache.org/

[2] Jacky Wu, "Hadoop HDFS & MapReduce", Help Guidelines LSA Lab, NTHU, Taiwan 2013.8.7.

[3] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung, "The Google File System", SOSP'03, October 19–22, 2003, Bolton Landing, New York, USA. Copyright 2003 ACM.

[4] Piyush Saxena, Satyajit Padhy, Praveen Kumar, "Optimizing Parallel Data Processing With Dynamic Resource Allocation", International Conference on Reliability, Infocom Technologies and Optimization.,pp. 735-739, Jan. 29-31, 2013.

[5] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", OSDI'04, 2004, Bolton Landing, New York, USA. Copyright 2004 ACM.

[6] "Can High-Performance Interconnects Benefit Hadoop Distributed File System? ", Lab Resources of Network-Based Computing Laboratory, Department of Computer Science and Engineering, The Ohio State University, USA.

[7] N. S. Islam, M. W. Rahman, J. Jose, R. Rajachandrasekar, H. Wang, H. Subramoni, C. Murthy, and D. K. Panda, "High Performance RDMA-based Design of HDFS over InfiniBand", Research Resources of Department of Computer Science and Engineering and IBM T.J Watson Research Center, The Ohio State University Yorktown Heights, NY. November 10-16, 2012, Salt Lake City, Utah, USA.

[8] Open Fabrics Enterprise Distribution, http://www.openfabrics.org/.

[9] X. Ding, S. Jiang, F. Chen, K. Davis, and X. Zhang. DiskSeen: Exploiting Disk Layout and Access History to Enhance I/O Prefetch. In Proceedings of USENIX07, 2007.

[10] InfiniBand Trade Association Home: http://www.infinibandta.org/

[11] C. Gniady, Y. C. Hu, and Y.-H. Lu. Program Counter Based Techniques for Dynamic Power Management. In Proceedings of the 10th International Symposium on High Performance Computer Architecture, HPCA '04, Washington, DC, USA, 2004. IEEE Computer Society.

[12] Shengsheng Huang, Jie Huang, Jinquan Dai, Tao Xie, and Bo Huang, "The HiBench Benchmark Suite: Characterization of the MapReduce-Based Data Analysis", ICDE Workshops'10, Oct. 2010, 2010 IEEE.

[13] Lan Yi, "Experience with HiBench: From Micro-Benchmarks toward End-to-End Pipelines", WBDB 2013 Workshop Presentation, Intel China Software Center, 2013.07.16.

[14] Dominique Heger, "Hadoop Performance Tuning - A Pragmatic & Iterative Approach", Research details by DHTechnologies - www.dhtusa.com, 2013.

[15] Jason Dai, "Toward Efficient Provisioning and Performance Tuning for Hadoop", Apache Asia Roadshow 2010, Intel China Software Center, June 2010.

[16] Remote Direct Memory Access : http://en.wikipedia.org/wiki/Remote_direct_memory_access

[17] Liang Ming , Dan Feng, Fang Wang, Qi Chen, Yang Li, Yong Wan, Jun Zhou, "A Performance Enhanced User-space Remote Procedure Call on InfiniBand*", Photonics and Optolectronics Meetings (POEM)., 2011.

[18] Fan Liang, Chen Feng, Xiaoyi Lu, Zhiwei Xu, "Performance Benefits of DataMPI:A Case Study with BigDataBench", ACM SOFT BPOE '14, Mar 1, 2014, Salt Lake City, Utah, USA.

[19] Xiaoyi Lu, Nusrat S. Islam, Md. Wasi-ur-Rahman, Jithin Jose, Hari Subramoni, Hao Wang, and Dhabaleswar K. (DK) Panda, "High-Performance Design of Hadoop RPC with RDMA over InfiniBand", National Science Foundation grants #OCI-0926691, #OCI-1148371 and #CCF-1213084, 2013 IEEE.

[20] K. Gupta, R. Jain, H. Pucha, P. Sarkar, and D. Subhraveti, "Scaling Highly-Parallel Data-Intensive Supercomputing Applications on a Parallel Clustered Filesystem," in The SC10 Storage Challenge.

Piyush Saxena Pursuing Master of Technology in Computer Science and Engineering from Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India, Area of Interest: Cloud Computing, Data Mining and Warehousing and Soft Computing.