Benefits of High Speed Interconnects to Cluster File Systems: A Case Study with Lustre *

Weikuan Yu Ranjit Noronha Shuang Liang Dhabaleswar K. Panda

Network-Based Computing Lab
Dept. of Computer Sci. & Engineering
The Ohio State University
{yuw,noronha,liangs,panda}@cse.ohio-state.edu

Abstract

Cluster file systems and Storage Area Networks (SAN) make use of network IO to achieve higher IO bandwidth. Effective integration of networking mechanisms is important to their performance. In this paper, we perform an evaluation of a popular cluster file system, Lustre, over two of the leading high speed cluster interconnects: InfiniBand and Quadrics. Our evaluation is performed with both sequential IO and parallel IO benchmarks in order to explore the capacity of Lustre under different communication characteristics. Experimental results show that direct implementations of Lustre over both interconnects can improve its performance, compared to an IP emulation over InfiniBand (IPoIB). The performance of Lustre over Quadrics is comparable to that of Lustre over InfiniBand with the platforms we have. Latest InfiniBand products can embrace latest technologies, such as PCI-Express and DDR, and provide higher capacity. Our results show that over a Lustre file system with two Object Storage Servers (OSSs), InfiniBand with PCI-Express technology can improve Lustre write performance by 24%. Furthermore, our experimental results indicate that Lustre meta-data operations do not scale with an increasing number of OSSs, in spite of using high performance interconnects.

*This research is supported in part by DOE grant #DE-FC02-01ER25506, NSF Grants #CNS-0403342 and #CNS-0509452; grants from Intel, Mellanox, Sun MicroSystems, Cisco Systems, and Linux Networks; and equipment donations from Intel, Mellanox, AMD, Apple, IBM, Microway, PathScale, SilverStorm and Sun MicroSystems.

1. Introduction

While CPU clock cycle and memory bus speed are reaching the level of sub-nanoseconds and 10Gbytes/sec, disk access time and data transfer rate are still lingering around several milliseconds and 300Mbytes/sec, respectively. Since systems with ever-increasing speed are being deployed at the scale of thousands of nodes, IO speed needs to keep pace with the demand of high performance computing applications. On these systems, high speed interconnects are typically being utilized due to the very low latency and very high bandwidth they can achieve. Recently, utilizing high-end interconnect technologies to bridge the gap between CPU/memory speed and IO speed has been exploited in storage and file systems by striping IO accesses across multiple storage servers over the network. These cluster-based storage and file systems can combine the advantages of both high speed network accesses of the interconnects and large storage capacity available from the commodity nodes. The trend as such makes it very promising to build scalable petabyte storage area networks through commodity clusters. Many commercial [13, 15, 8, 7] and research projects [21, 11, 2] have been developed to provide parallel file systems for IO accesses using such architectures.

Currently, several high speed interconnects can provide very high bandwidth at the level of 10Gbps or even higher, including InfiniBand [14], Quadrics [27, 4], Myrinet [5], and 10Gigabit Ethernet [12]. Minimum achievable latency between two nodes already has gone below 2μs. It is important not only to the system designers and practitioners who build the systems, but also to the hardware vendors who produce these interconnects

to find out how effectively these interconnects' communication characteristics can be integrated into the file system.

A popular cluster file system, Lustre [7], has recently been designed and developed to address the needs of next generation systems using low cost Linux clusters with commodity computing nodes and high speed interconnects. Currently, it is available over several different interconnects, including Gigabit Ethernet, Quadrics [27] and InfiniBand [14] (in a prototype form). In order to gain more understanding of high speed interconnects and their performance impacts to storage and file systems, we intend to evaluate the performance of Lustre over different interconnects and study the following questions:

1. *Which file system operations of Lustre can benefit more from low latency and high performance of high speed interconnects?*
2. *Can the latest IO-bus technologies, such as PCI-Express [24], help Lustre performance of high speed interconnects?*
3. *What aspects of Lustre still need to be optimized even with high speed interconnects and latest IO-bus technologies?*

In this paper, we have used a set of micro-benchmarks and application benchmarks to evaluate the performance of Lustre over two high speed interconnects: InfiniBand and Quadrics. It is to be noted that there is also a prototype implementation of Lustre over Myrinet/GM [20]. However, in our experience, the current Lustre implementation over GM does not support large message communications, so it is not included in this work. Our benchmarks include not only the traditional sequential IO benchmarks, but also parallel IO benchmarks. Due to the limitation on available system size, we have conducted experiments on two clusters: one 8-node Xeon cluster with both interconnects, and another 4-node EM64T cluster with PCI-Express InfiniBand HCAs. Gigabit Ethernet is also available as the default network. However, its bandwidth is an order of magnitude less than InfiniBand and Quadrics. Thus we have not considered Gigabit Ethernet for a fair comparison. Instead, we have chosen to use an Ethernet implementation over InfiniBand hardware, IPoIB, to evaluate the performance of Lustre over TCP/IP-type networks.

Our experimental results indicate that Lustre file IO operations benefit more from high speed communication provided by both InfiniBand and Quadrics compared to an IP emulation over InfiniBand (IPoIB). In addition, the read and write bandwidth of an MPI-Tile-IO benchmark over Lustre over Quadrics is about 13% higher than the same over InfiniBand with the platforms we have. However, latest InfiniBand products integrate well with latest technologies, e.g., taking advantage of PCI-Express and DDR (Double Data Rate). Our results show that over a Lustre file system with two Object Storage Servers (OSSs), InfiniBand with PCI-Express technology can improve Lustre write performance by 24%. Our results also suggest that Lustre meta-data operations do not scale with an increasing number of Object Storage Servers, though it benefits slightly from high performance of latest interconnect technologies.

The rest of the paper is presented as follows. In the next section, we provide an overview of Lustre. Section 3 provides an overview of two interconnects: InfiniBand and Quadrics. Section 4 describes the details of our experimental testbed. Sections 5, 6 and 7 provide performance results from our evaluation. Section 8 gives a brief review of related work. Section 9 concludes the paper.

2. Overview of Lustre

Lustre [7] is a Posix-compliant, stateful, object-based parallel file system. It provides fine-grained parallel file services with its distributed lock management. Lustre separates essential file system activities into three components: clients, meta-data servers and storage servers. These three components are referred to as Object Storage Client (OSC), Meta-Data Server (MDS) and Object Storage Server (OSS), respectively.

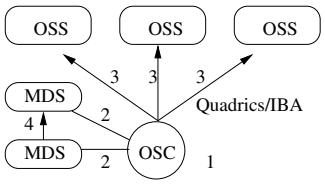


Fig. 1. Lustre System Architecture

Fig. 1 shows a diagram of Lustre system architecture. Meta-data operations are decoupled from file IO operations in Lustre. To access a file, a client first obtains from the primary MDS its meta-data, including file attributes, file permission and the layout of file objects. Subsequent file IO (storage) operations are done directly from the client to the OSS. By decoupling meta-data operations from IO operations, data IO can be carried out in a parallel fashion, which allows greater aggregated bandwidth. Lustre also provides a fail-over MDS, which is quiescent normally but can provide complete meta-data services in

case the primary MDS fails.

MPI-IO is the IO extension of MPI-2 [19] standard. It provides a high performance and portable parallel IO interface. An MPI-IO implementation over a file system requires a file system specific ADIO implementation. Currently, Lustre MPI-IO support is available through UFS-based ADIO implementation because its compatibility with the Unix IO interface.

3. Overview of High Speed Interconnects

In this section, we provide an overview of two interconnects: InfiniBand [14] and Quadrics [27].

3.1. InfiniBand

The InfiniBand Architecture (IBA) [14] is an open specification designed for interconnecting compute nodes, IO nodes and devices in a system area network. As shown in Fig. 2, it defines a communication architecture from the switch-based network fabric to transport layer communication interface for inter-process communication. In an InfiniBand network, compute nodes are connected to the fabric by Host Channel Adapters (HCA).

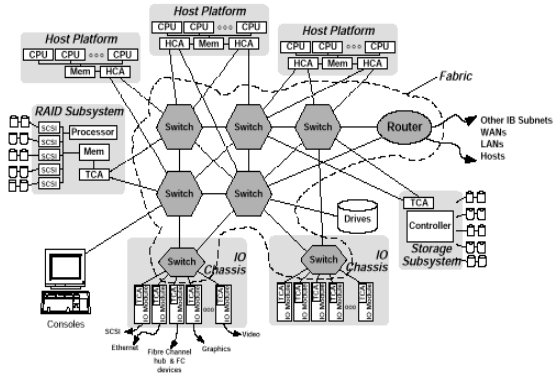


Fig. 2. The Switch Fabric of InfiniBand Architecture (Courtesy InfiniBand Trade Association)

Two communication semantics are supported in IBA: channel semantics with traditional send/receive operations and memory semantics with RDMA operations. RDMA operations allow one side of the communication parties to exchange information directly with the remote memory without the involvement of the remote host.

There are several different implementations of InfiniBand SDK available. We have chosen to evaluate an open-source implementation, OpenIB (Gen-1), for the evaluation of Lustre over InfiniBand. OpenIB also supports an emulated IP implementation, IPoIB. IP-based application can run directly over the same InfiniBand fabric. In this work, we use IPoIB as a substitute of Ethernet IP for the evaluation of Lustre over TCP/IP-like networks.

3.2. Quadrics

QsNet^{II} [4]. is the second generation network from Quadrics [27]. This release provides very low latency, high bandwidth communication with its two building blocks: a programmable Elan-4 network interface and the Elite-4 switch, which are interconnected in a fat-tree topology. Interprocess communication is supported by two different models: Queue-based model (QDMA) and Remote Directed Message Access (RDMA) model. QDMA allows a process to post messages (up to 2KB) to remote queues exposed from other processes; RDMA enables a process to write messages directly into remote memory exposed by other processes.

As shown in Fig. 3, Quadrics provides two communication libraries: `libelan` and `libelan4` user-level libraries and a kernel communication library, on top of its Elan4 network [27]. Lustre (CFS) [7] implementation is built upon the kernel communication library. It has been deployed in many large-scale clusters, such as Thunder from Lawrence Livermore National Laboratory. The implementation over the latest Quadrics Elan4 release is used in this evaluation.

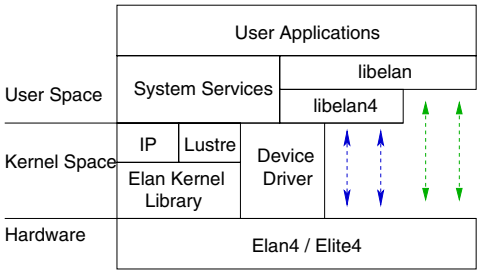


Fig. 3. Quadrics/Elan4 Communication Architecture

4. Experimental Testbed

In this section, we describe the experimental testbed used for the performance evaluation of Lustre [7] over InfiniBand [14] and Quadrics [27]. We have used two different clusters. The first cluster consists of eight SuperMicro SUPER X5DL8-GG nodes, each with dual Intel Xeon 3.0 GHz processors, 512 KB L2 cache, PCI-X

64-bit 133 MHz bus, 533MHz Front Side Bus (FSB) and a total of 2GB PC2100 DDR-SDRAM physical memory. These eight nodes are connected with both InfiniBand and Quadrics. The other is a four-node cluster, each node with dual 3.4GHz Intel EM64T Processors, 1024KB L2 cache and 1GB main memory. These nodes have both 8x PCI Express and 64 bit/133 MHz PCI-X interfaces, and they are connected with InfiniBand.

Quadrics: Quadrics^{II} network [27, 4] consists of a dimension one quaternary fat-tree [10] with a QS-8A switch and eight Elan4 QM-500 cards. Quadrics network is used in the PCI-X Xeon cluster. The operating system used for Quadrics experiments is RedHat AS-3 Linux kernel version 2.4.21-27.0.2.EL, with both Quadrics kernel patches and Lustre kernel patches. Quadrics 4.31qsnet Eagle release is used in our experiments. The peak achievable network bandwidth is 910 Mbytes/Sec.

InfiniBand: The InfiniBand network consists of a Mellanox InfiniScale 144 port switch. The PCI-X InfiniBand cluster is running in RedHat AS-3 Linux kernel 2.4.21-27.ELsmp, patched with Lustre kernel patches. InfiniBand software stack IBGD-1.7.0 and HCA firmware version 3.3.2 for MT23108 PCI-X HCAs are used in this cluster. The peak achievable network bandwidth for this cluster is 888 Mbytes/Sec. The PCI-Express InfiniBand cluster is running in RedHat AS-4 Linux kernel 2.6.9-5.ELsmp with Lustre kernel patches. Software stacks used in this cluster are InfiniBand IBGD-1.8.0, firmware 3.3.3 for MT23108 PCI-X HCAs and firmware 5.1.0 for MT25208 PCI Express HCAs. The peak achievable network bandwidth for this cluster is 960 Mbytes/Sec. Since the Lustre OpenIB-Gen1 implementation is still in a prototype form and many file system activities are unstable, we have conducted best-effort experiments and presented the results that could be obtained consistently here.

5. Performance Evaluation with Sequential IO Benchmarks

In this section, we provide a microbenchmark evaluation of its Unix sequential IO performance, such as read and write bandwidth, as well as IO transaction throughput. Two sets of experiments are performed. The first one uses a popular sequential IO benchmark suite, IOzone [1] and its fileop benchmark, to measure read and write bandwidth and meta-data operation performance. The other uses the Postmark benchmark [3] from Network Appliances to measure the IO transaction rates for workloads typically seen on the Internet electronic mail servers. The eight-node PCI-X cluster is used and configured with a single client and various number of servers. One of the servers is dedicated as an MDS and the rest of them OSSs.

5.1. IOzone Benchmarks

IOzone – IOzone [1] performs a variety of file IO tests on a file system. We have evaluated the performance of Lustre with read and write tests using a file of 256MB. Fig. 4 shows the Read and Write performance of Lustre with up to six OSSs. The write performance initially increases with the number of OSSs. However, over any of the three network configurations, two servers can saturate the IO demand of a single IO client. In terms of read performance, because the entire file can be cached at the client side, we observe the same bandwidth across three different configurations. For files larger than 256MB, we were not able to conduct IOzone experiments over OpenIB due to the Lustre stability problem.

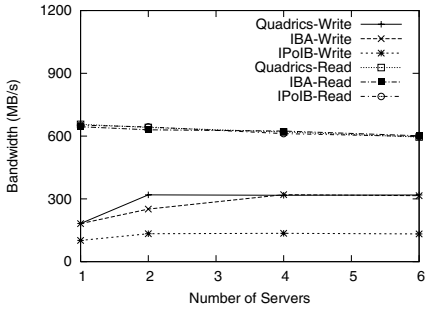


Fig. 4. Read and Write Performance of IOzone

Fileop – Another benchmark, fileop, is distributed along with IOzone. It tests the performance of a variety meta-data operations including create, stat, access, readdir, link, unlink and delete. We have observed a similar trend of performance results among these operations. Fig. 5 shows the performance for two of operations: create and stat. For both create and stat operations, the number of achievable transactions decreases with the increasing number of OSSs. This suggests that a single active meta-data server provided by Lustre can potentially be a performance and scalability bottleneck. This is because the meta-data has to allocate and manage meta-data before creating actual data objects for all the OSSs. Thus, the scalability issue can occur when the meta-data server has to handle the attribute access or update of many such dispersed objects. We plan to investigate this issue further by providing distributed storage for meta-data or journaling upon meta-data commitment.

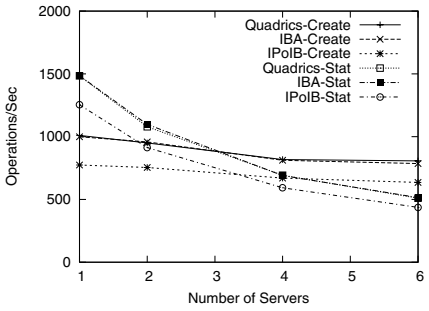


Fig. 5. Performance of File Create and Stat Operations

5.2. Postmark

Postmark [3] is a benchmark that measures file system performance on small short-lived files. These type of workloads are typically seen in computing systems which process e-mail and network groups and other communication intensive environments. It first creates a pool of text files and then performs two different sets of transactions on the pool, either create/delete a file or read/append a file. The transaction throughput is measured to approximate workloads on an Internet server such as electronic mail.

We have measured the performance of postmark with 100,000 transactions on 4096 files. Table 1 shows the average transaction rates out of 10 different executions for Lustre over Quadrics, InfiniBand and IPoIB. Lustre over Quadrics has the highest number transactions per second for the postmark workload. The transaction rate decreases with more number of OSSs. This confirms the earlier fileop results on the scalability bottleneck of Lustre meta-data server with increasing number of OSSs.

Table 1. Postmark Performance (Transactions per Second)

OSS	Quadrics	IBA	IPoIB
1	500	320	283
2	250	220	170
4	186	177	132
6	150	153	113

6. Performance Evaluation with Parallel IO Benchmarks

Parallel IO with concurrent read and write are typical IO patterns in scientific applications. We have evaluated the performance of management operations as well as data access operations. In the first experiment, a parallel benchmark is used to measure the rate of parallel management operations. Then two application benchmarks: MPI-Tile-IO [28] and BT-IO [29], are used to measure the performance of parallel data IO operations. The eight-node PCI-X cluster is used in these experiments. Four of eight nodes used are configured as server nodes (3 OSSs and 1 MDS), and the other four as client nodes.

6.1. Parallel Management Operations

Parallel applications can potentially generate a lot of management operations that do not involve massive data transfer. The performance of these management operations is important to the parallel applications. To evaluate the strength of different interconnects to the performance of these management operations, we have performed the following experiments using a microbenchmark program [16], *op_test*, in the PVFS2 [2] distribution.

Four of eight nodes used are configured as server nodes (3 OSSs and 1 MDS), and the other four as client nodes. The first experiment measures the average time to create a file using collective `MPI_File_open` with different numbers of clients. The second experiment measures the average time to perform a resize operation using collective `MPI_File_set_size` with different numbers of clients. Table 2 shows the performance of MPI-IO management operations over Lustre with different interconnects. These results suggest a scalability problem of Lustre for MPI-IO management operations as well. Quadrics provides very low latency communication. In contrary to intuition, the meta-data operations perform worse on Quadrics for two nodes, but seem to have a lower increasing trend as system size increases. Further investigation needs to confirm the trend with more number of clients and gain more insights into this.

Table 2. The Performance of Parallel IO Management Operations

No. of clients	Quadrics	IBA	IPoIB
Create (milliseconds)			
2	4.79	3.83	3.93
4	6.57	6.85	6.29
Resize (milliseconds)			
2	4.89	3.18	3.77
4	6.48	6.46	5.85

6.2. Performance of MPI-Tile-IO

MPI-Tile-IO [28] is a tile reading MPI-IO application. It tests the performance of tiled access to a two-dimensional dense dataset, simulating the type of workload that exists in some visualization and numerical applications. Four of eight nodes are used as server nodes and the other four as client nodes running MPI-Tile-IO processes. Each process renders a 2×2 array of displays, each with 1024×768 pixels. The size of each element is 32 bytes, leading to a file size of 96MB.

We have evaluated both the read and write performance of MPI-Tile-IO over Lustre. As shown in Fig. 6, Lustre over Quadrics achieves 13% higher read bandwidth and 14.2% higher write bandwidth compared to Lustre over InfiniBand, Lustre over IPoIB achieves less read and write bandwidth performance compared to the other two. This is consistent with the relative bandwidth difference between Quadrics/Elan4, 4x InfiniBand, and IPoIB over 4x InfiniBand. However, the current InfiniBand cards can take advantage of higher bandwidth of PCI-Express architecture, which provides higher peak bandwidth. In addition, the upcoming 4x DDR and 12x InfiniBand releases will provide cards with higher bandwidth. So InfiniBand is likely to provide better benefits to Lustre in the near future. In future as resources are available, we plan to experiment with a larger cluster with PCI-Express HCAs, and verify that MPI-Tile-IO can benefit from the Lustre performance improvement provided by PCI-Express technology.

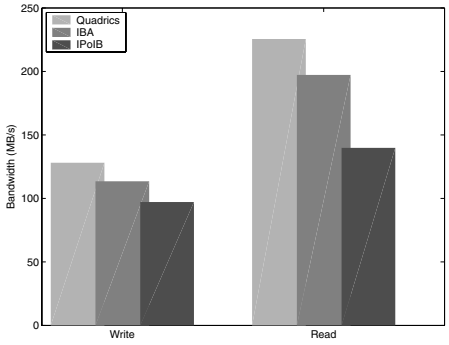


Fig. 6. Performance of MPI-Tile-IO

6.3. Performance of NAS BT-IO

The BT-IO benchmarks are developed at NASA Ames Research Center based on the Block-Tridiagonal problem of the NAS Parallel Benchmark suite. These benchmarks test the speed of parallel IO capability of high performance computing applications. The entire data set undergoes complex decomposition and partition, eventually distributed among many processes. These steps involve intensive IO accesses. More details are available in [29].

The BT-IO problem size class A is evaluated. Table 3 shows the BT-IO performance of Lustre is quite comparable between Quadrics and InfiniBand. Both of their performance is better than Lustre over IPoIB. However, in contrast to MPI-IO results, the BT-IO performance does not seem to differ much between Quadrics and InfiniBand. This is due to the difference between the communication characteristics of MPI-Tile-IO and BT-IO. MPI-Tile-IO appears to be more bandwidth-bound application.

Table 3. Performance of BT-IO Benchmark (seconds)

Type	Duration	IO Time
BT	61.34	–
BT/IO Quadrics	69.08	7.74
BT/IO IBA	69.11	7.77
BT/IO IPoIB	73.59	12.25

7. Benefits of PCI-Express to Lustre over InfiniBand

PCI-Express [24] is an emerging board-level interconnect technology that provides a high performance, point-to-point, full-duplex, and serial IO-bus interface. It provides much higher IO-bus bandwidth compared to the traditional parallel PCI [23] technology and its extension, PCI-X [23]. InfiniBand [14] is one of the leading interconnects that can embrace the benefits of PCI-Express. In this section, we provide an evaluation of PCI-Express benefits to Lustre over InfiniBand.

On the four-node cluster with both PCI-Express and PCI-X, we have configured Lustre with up to two OSSs, one MDS and a client. we have measured the IOzone performance of Lustre with a file of 256MB. Fig. 7 shows the Lustre performance over InfiniBand with PCI-Express and PCI-X HCAs. Again, we observe the comparable read bandwidth for both PCI-Express and PCI-X because the entire file of 256MB can be cached in Lustre client-side cache. The write bandwidth of Lustre over InfiniBand is improved by about 24% with PCI-Express for two OSSs. Note that the absolute write bandwidth is less than that of Section 5.1. This is because the local file system used in these experiments is *ldiskfs*. This disk file system was recently implemented by Lustre for Linux 2.6. Though *ldiskfs* does not provide equivalent performance compared to the heavily optimized version of ext3 used for Lustre over Linux 2.4, the performance

numbers obtained with it still provide a valid comparison between PCI-Express and PCI-X.

8. Related Work

Previous research have studied the benefits of high speed interconnects to parallel IO accesses in storage networks. Zhou et. al. [31] have studied the benefits of VIA networks in database storage. Wu et. al. [30] have described their work on InfiniBand over PVFS1 [22]. DeBergalis et. al. [9] have described a file system, DAFS, built on top of networks with VIA-like semantics. Yu et. al. [30] have described their work on Quadrics over PVFS2 [2]. These research have shown the benefits of utilizing high speed interconnects over the traditional TCP/IP networks.

Literature on the performance evaluation of different interconnects has largely focused on parallel computing applications. Petrini et. al. [26, 25] have presented detailed information about the Quadrics network architecture and performance of its Elan communication layer. Liu et. al. [18, 17] have presented performance evaluations of InfiniBand, Quadrics and Myrinet, including both micro-benchmark level evaluation of their user-level communication libraries and parallel application level evaluation of the MPI libraries on top of them. Brightwell et. al. [6] have compared the performance of InfiniBand and Quadrics Elan-4 Technology, also in the context of high performance parallel applications.

Our work compares the performance benefits of InfiniBand and Quadrics to the IO subsystem, using a cluster file system, Lustre [7]. The communication capabilities of these two interconnects are exposed to Lustre through their kernel level networking API. This work is complementary to the previous work on the evaluation of different interconnect technologies.

9. Conclusions

In this paper, we have evaluated the performance of Lustre over two leading high speed interconnects: InfiniBand and Quadrics. We have employed both sequential Unix IO benchmarks and parallel IO benchmarks. Our experimental results indicate that Lustre benefits more from the high speed communication provided by both InfiniBand and Quadrics, compared to an IP emulation over InfiniBand (IPoIB). The performance of Lustre over Quadrics is slightly better, yet comparable with that over InfiniBand with the platforms we have. Furthermore, we have also shown that the emerging PCI-Express IO-bus technology can improve the write bandwidth of Lustre over InfiniBand by 24%. It is interesting to note that Lustre meta-data operations scale rather poorly with the increasing number of Object Storage Servers (OSSs).

In future, we plan to investigate further on how to optimize the performance of Lustre using the latest InfiniBand implementations. We also intend to study the performance and scalability of Lustre with large systems. Moreover, we intend to study the impact of the higher capacity provided by latest InfiniBand releases.

Additional Information – Additional information related to this research can be found on the following website: http://nowlab.cse.ohio-state.edu/

References

[1] IOzone Filesystem Benchmark.

[2] The Parallel Virtual File System, version 2. http://www.pvfs.org/pvfs2.

[3] PostMark: A New File System Benchmark. Tech. Rep. TR3022, october 1997.

[4] J. Beecroft, D. Addison, F. Petrini, and M. McLaren. QsNet-II: An Interconnect for Supercomputing Applications. In *the Proceedings of Hot Chips '03*, Stanford, CA, August 2003.

[5] N. J. Boden, D. Cohen, R. E. Felderman, A. E. Kulawik, C. L. Seitz, J. N. Seizovic, and W.-K. Su. Myrinet: A Gigabit-per-Second Local Area Network. *IEEE Micro*, 15(1):29–36, 1995.

[6] R. Brightwell, D. Dourfler, and K. D. Underwood. A Comparison of 4X InfiniBand and Quadrics Elan-4 Technology. In *Proceedings of Cluster Computing, '04*, San Diego, California, September 2004.

[7] Cluster File System, Inc. Lustre: A Scalable, High Performance File System. http://www.lustre.org/docs.html.

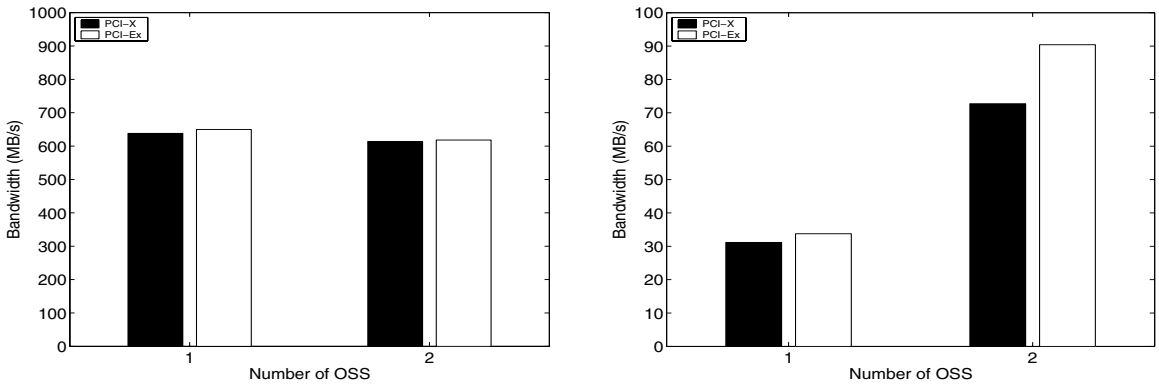
[8] A. M. David Nagle, Denis Serenyi. The Panasas ActiveScale Storage Cluster – Delivering Scalable High Bandwidth Storage. In *Proceedings of Supercomputing '04*, November 2004.

[9] M. DeBergalis, P. Corbett, S. Kleiman, A. Lent, D. Noveck, T. Talpey, and M. Wittle. The Direct Access File System. In *Proceedings of Second USENIX Conference on File and Storage Technologies (FAST '03)*, 2003.

[10] J. Duato, S. Yalamanchili, and L. Ni. *Interconnection Networks: An Engineering Approach*. The IEEE Computer Society Press, 1997.

[11] J. Huber, C. L. Elford, D. A. Reed, A. A. Chien, and D. S. Blumenthal. PPFS: A High Performance Portable Parallel File System. In *Proceedings of the 9th ACM International Conference on Supercomputing*, pages 385–394, Barcelona, Spain, July 1995. ACM Press.

[12] J. Hurwitz and W. Feng. End-to-End Performance of 10-Gigabit Ethernet on Commodity Systems. *IEEE Micro '04*.



(a) Read (b) Write

Fig. 7. Benefits of PCI-Express to the Performance of Lustre over InfiniBand

[13] IBM Corp. IBM AIX Parallel I/O File System: Installation, Administration, and Use. Document Number SH34-6065-01, August 1995.

[14] Infiniband Trade Association. http://www.infinibandta.org.

[15] Intel Scalable Systems Division. Paragon System User's Guide, May 1995.

[16] R. Latham, R. Ross, and R. Thakur. The impact of file systems on mpi-io scalability. In *Proceedings of the 11th European PVM/MPI Users' Group Meeting (Euro PVM/MPI 2004)*, pages 87–96, September 2004.

[17] J. Liu, B. Chandrasekaran, J. Wu, W. Jiang, S. P. Kini, W. Yu, D. Buntinas, P. Wyckoff, and D. K. Panda. Performance Comparison of MPI implementations over Infiniband, Myrinet and Quadrics. In *Proceedings of Supercomputing '03*, November 2003.

[18] J. Liu, B. Chandrasekaran, W. Yu, J. Wu, D. Buntinas, S. P. Kini, P. Wyckoff, and D. K. Panda. Micro-Benchmark Performance Comparison of High-Speed Cluster Interconnects. *IEEE Micro*, 24(1):42–51, January-February 2004.

[19] Message Passing Interface Forum. *MPI-2: Extensions to the Message-Passing Interface*, Jul 1997.

[20] Myricom Corporations. The GM Message Passing Systems.

[21] N. Nieuwejaar and D. Kotz. The Galley Parallel File System. *Parallel Computing*, (4):447–476, June 1997.

[22] P. H. Carns and W. B. Ligon III and R. B. Ross and R. Thakur. PVFS: A Parallel File System For Linux Clusters. In *Proceedings of the 4th Annual Linux Showcase and Conference*, pages 317–327, Atlanta, GA, October 2000.

[23] PCI-SIG. PCI and PCI-X. http://www.pcisig.com.

[24] PCI-SIG. PCI Express Architecture. http://www.pcisig.com.

[25] F. Petrini, W.-C. Feng, A. Hoisie, S. Coll, and E. Frachtenberg. The Quadrics Network (QsNet): High-Performance Clustering Technology. In *Hot Interconnects 9*, Stanford University, Palo Alto, CA, August 2001.

[26] F. Petrini, A. Hoisie, W.-C. Feng, and R. Graham. Performance Evaluation of the Quadrics Interconnection Network. In *Workshop on Communication Architecture for Clusters 2001 '01*, April 2001.

[27] Quadrics, Inc. Quadrics Linux Cluster Documentation.

[28] R. B. Ross. Parallel i/o benchmarking consortium. http://www-unix.mcs.anl.gov/rross/pio-benchmark/html/.

[29] P. Wong and R. F. Van der Wijngaart. NAS Parallel Benchmarks I/O Version 2.4. Technical Report NAS-03-002, Computer Sciences Corporation, NASA Advanced Supercomputing (NAS) Division.

[30] J. Wu, P. Wyckoff, and D. K. Panda. PVFS over InfiniBand: Design and Performance Evaluation. In *Proceedings of the International Conference on Parallel Processing '03*, Kaohsiung, Taiwan, October 2003.

[31] Y. Zhou, A. Bilas, S. Jagannathan, C. Dubnicki, J. F. Philbin, and K. Li. Experiences with VI Communication for Database Storage. In *Proceedings of the 29th Annual International Symposium on Computer Architecture*, pages 257–268. IEEE Computer Society, 2002.