16th IEEE Symposium on High Performance Interconnects

Performance Analysis and Evaluation of PCIe 2.0 and Quad-Data Rate InfiniBand [*]

Matthew J. Koop Wei Huang Karthik Gopalakrishnan Dhabaleswar K. Panda

Network-Based Computing Laboratory
Department of Computer Science and Engineering
The Ohio State University
{koop, huanwei, gopalakk, panda}@cse.ohio-state.edu

Abstract

High-performance systems are undergoing a major shift as commodity multi-core systems become increasingly prevalent. As the number of processes per compute node increase, the other parts of the system must also scale appropriately to maintain a balanced system. In the area of high-performance computing, one very important element of the overall system is the network interconnect that connects compute nodes in the system. InfiniBand is a popular interconnect for high-performance clusters. Unfortunately, due to limited bandwidth of the PCI-Express fabric, InfiniBand performance has remained limited.

PCI-Express (PCIe) 2.0 has recently become available and has doubled the transfer rates available. This additional I/O bandwidth balances the system and makes higher data rates for external interconnects such as InfiniBand feasible. As a result, InfiniBand Quad-Data Rate (QDR) mode has become available on the newest Mellanox InfiniBand Host Channel Adapter (HCA) with a 40 Gb/sec signaling rate. In this paper we perform an in-depth performance analysis of PCIe 2.0 and the effect of increased InfiniBand signaling rates. We show that even using the Double Data Rate (DDR) interface, PCIe 2.0 enables a 25% improvement in NAS Parallel Benchmark IS performance. Furthermore, we show that when using QDR on PCIe 2.0, network loopback can outperform a shared memory message passing implementation. We show that increased interconnection bandwidth significantly improves the overall system balance by lowering latency and increasing bandwidth.

1 Introduction

Recently higher-core counts have begun to replace the traditional increase in clock frequency as the use for the additional transistors that continue to be made available per Moore's Law [7]. Many high-performance clusters are made up of many high-volume "off-the-shelf" components to provide a good cost-performance ratio. As commodity processors have moved towards multicore, high-performance commodity clusters have also seen a significant increase in core counts per node.

One of the most important components of a high-performance cluster is the compute node interconnect. The performance of the interconnect can make a significant difference in overall application performance. The InfiniBand Architecture [8], a popular high-performance interconnect, is an industry standard technology which aims to provide low-latency and high-bandwidth communication.

As nodes have increased in processing power and core counts, the other system components must also scale to provide a balanced system architecture. Increased memory speeds and processor speeds are two components and the I/O subsystem is another. To maintain a proper balance the I/O must also be scalable.

As these I/O requirements have increased, PCI-Express [17] (PCIe) was introduced as a new switched serial point-to-point network. It replaced the traditional PCI shared bus architecture that was unable to scale to the number of devices and bandwidth required. However, in the years since the first version of the PCIe standard, I/O requirements have continued to increase. As a result, PCIe 2.0 [18] was released in late 2007, which

[*]This research is supported in part by Department of Energy grants #DE-FC02-06ER25749 and #DE-FC02-06ER25755; NSF grants #CNS-0403342 and #CCF-0702675; grants from Intel, Mellanox, Cisco, Sun Microsystems, and Linux Networx; Equipment donations from Intel, Mellanox, AMD, Advanced Clustering, IBM, Appro, Microway, PathScale, Silverstorm and Sun Microsystems.

1550-4794/08 $25.00 © 2008 IEEE
DOI 10.1109/HOTI.2008.26

doubled the maximum transfer rate per lane to 5 GigaTransfers/sec (GT/s).

InfiniBand has benefited significantly from this increase in the I/O performance. As pressure has increased on the interconnect due to increased node computing power, InfiniBand Host Channel Adapters (HCAs) have increased performance as well. Current adapters generally run at 4X Single Data Rate (SDR) (10Gb/sec signaling rate) or Double Data Rate (DDR) (20Gb/sec). Using PCIe 1.1, InfiniBand DDR has been limited in performance by the PCI interconnect. As a result, increased signaling rates for InfiniBand have not been able to be introduced until now. To maintain system balance InfiniBand Quad Data Rate (QDR) has been introduced at a 40Gb/sec signaling rate, taking advantage of the newly-introduced PCIe 2.0.

In this paper we evaluate InfiniBand on a PCIe 2.0 system. We evaluate the benefits of PCIe 2.0 on both DDR and QDR data rates on the Mellanox ConnectX [11] HCA. We also investigate the general trend of additional interconnect bandwidth on application performance on multi-core machines. On QDR systems up to 2.5 GB/sec unidirectional and 5.09 GB/sec bidirectional bandwidth is observed. We show that even using the DDR InfiniBand PCIe 2.0 enables a 25% improvement in NAS Parallel Benchmark (NPB) IS performance. Furthermore, we show that when using QDR on PCIe 2.0, network loopback can outperform a shared memory message passing implementation in some cases. On the NAS kernel IS we show a 6% improvement using QDR loopback as compared to the shared memory implementation.

The remaining part of the paper is organized as follows: PCI Express is detailed in Section 2. In Section 3 we provide an overview of the InfiniBand Architecture. Section 4 gives an overview of our methodology and platform for evaluation. We present a performance evaluation of SDR, DDR and QDR rates with and without PCIe 2.0 in Section 5. In Section 6 we discuss related work in this area. Finally, we conclude the paper in Section 7.

2 Overview of PCI Express

PCI was the standard local I/O bus technology for many years. It was implemented as a parallel shared bus at the physical layer. Over time the bus frequency and bus width were increased. However, as demands increased the parallel bus implementation of PCI was found to be limiting [16].

In parallel bus architectures timing skew limits scalability. When board traces are of different lengths the signaling times to the various components on the bus differ. This skew prevents significantly increasing the bandwidth of the bus. Also, parallel buses can end up requiring a significant number of traces on the board. These issues are also much of the rationale for serial signaling in FB-DIMM memory technology, SATA, and other standards.

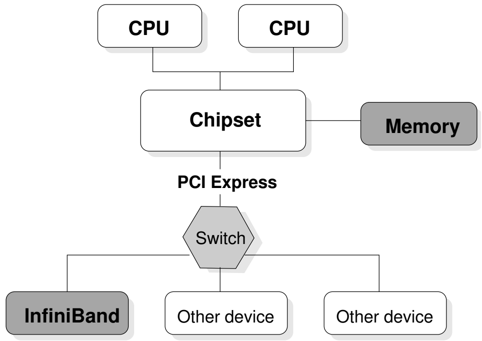


Figure 1. PCI Express Architecture

As a result of these limitations, PCI Express (PCIe) was introduced by Intel and others as the successor to PCI. Instead of the shared parallel bus of PCI, PCIe introduced a high-performance point-to-point serial interface [17]. Figure 1 shows the PCI Express architecture. These point-to-point links avoid the shared bus architecture of PCI. Additionally, PCIe allows a varied number of “lanes” to be be configured per device. Each lane in PCIe 1.1 supports 2.5 GigaTransfers/sec (GT/sec), 250 MB/sec, in each direction. As a result, an 8x PCIe 1.1 channel can achieve 2 GB/sec in each direction. It is important to note, however, that these are theoretical rates and other system components and implementation slow these rates.

PCIe 2.0 [18] doubles the transmission speed to 5.0 GT/sec, as many I/O devices were beginning to need additional bandwidth. While the PCIe 1.1 standard allowed up to a 32x slot, these were rarely used due to the extreme length. In practice, most server boards only implemented 8x slots, limiting theoretical bandwidth to 2GB/sec. PCIe 2.0 doubles the signaling rate, allowing an 8x PCIe 2.0 slot to provide 4GB/sec of bandwidth in each direction.

In addition to the doubled signaling speed, PCIe 2.0 also features backward compatibility with PCIe 1.1. Any PCIe 2.0 device will work in a PCIe 1.1 board and vice-versa. PCIe 2.0 can auto-negotiate link speeds and the number of lanes. This is expected to allow power savings when a only a slower bus bandwidth is required [18].

3 Overview of the InfiniBand Architecture

InfiniBand [8] was designed as a high-speed, general-purpose I/O interconnect. In recent years it has become a popular interconnect for high-performance computing to connect commodity machines in clusters from small to large scale.

3.1 InfiniBand Rates

The physical layer of InfiniBand is comprised of bidirectional links of 2.5Gb/sec. These links can be combined into 4X (10Gb/sec) and 12X (30Gb/sec) links. These speeds, however, are the gross data transfer speeds. The theoretical data transfer rate is 8/10ths of the gross due to an 8/10 encoding at the physical layer.

The InfiniBand specification also allows Double Data Rate (DDR) and Quad Data Rate (QDR) modes. In DDR operation, each InfiniBand lane is clocked at double the rate. In this allows the signaling rate per lane to double to 5Gb/sec. Similarly, QDR operation is clocked at quadruple the rate, allowing a 10Gb/sec signaling rate per lane. Thus, a 4X QDR InfiniBand link has a signaling rate of 40Gb/sec, or 32Gb/sec data rate.

3.2 Host Channel Adapters (HCAs)

There are a variety of InfiniBand Host Channel Adapters (HCAs) available on the market. These are currently designed by one of three companies: IBM, Mellanox, and QLogic. In this paper we will be focusing on the new ConnectX [11] card from Mellanox [1] as it has been designed to support QDR operation.

ConnectX is the fourth generation InfiniBand Host Channel Adapter (HCA) from Mellanox Technologies [1]. It is PCIe 2.0 capable. It provides two network ports to connect to the network fabric. Each port can be independently configured to be used either as 4X InfiniBand or 10 Gigabit Ethernet. Currently available versions of the cards do not allow mixed modes and allow only InfiniBand or 10GbE modes. In this paper we evaluate only the InfiniBand implementation of ConnectX.

4 Methodology and Platform

The goal of our evaluation is to provide a quantitative analysis of the benefits of both increased data rates from SDR to QDR, as well as the benefit of PCIe 2.0.

To facilitate this evaluation, we evaluate four different configurations on a single testbed. In this evaluation we will only change the network device, leaving all other factors the same. Recall that PCIe 2.0 is backwards compatible, so providing a PCIe 1.1 compliant device in a PCIe 2.0 environment will run as if it were PCIe 1.1.

The configurations are:

- *Single Data Rate, PCIe 1.1 (SDR-1.1)*: This combination is the lowest bandwidth configuration. As a result, PCIe 1.1 on an 8x lane allows full expected data rates.

- *Double Data Rate, PCIe 1.1 (DDR-1.1)*: This configuration is limited by the PCIe 1.1 8x bus. The theoretical bandwidth of PCIe 1.1 8x equals 2 GB/s which is at best equal to the 4x DDR rate (2.5 GB/s raw, 2 GB/s data). The overhead incurred by encapsulating InfiniBand packets in PCIe packets probably accounts for the limitation.

- *Double Data Rate, PCIe 2.0 (DDR-2.0)*: Using an 8x PCIe 2.0 lane, full InfiniBand bandwidth is possible.

- *Quad Data Rate, PCIe 2.0 (QDR-2.0)*: This is a prototype implementation of QDR on PCIe 2.0. Evaluation of QDR on PCIe 1.1 is not useful since DDR already exceeds what PCIe 1.1 can allow and thus is not evaluated here.

In all cases we use the same network device, a ConnectX HCA. This allows us to isolate performance differences precisely. We have HCAs with each of the different firmwares from our configuration available.

Our evaluation system consists of two compute nodes featuring the Intel "Harpertown" platform. Each node has dual Intel "Harpertown" 2.83GHz, quad-core processors. The nodes have 8 GB of main memory (FB-DIMMs). The platform is equipped with one PCIe 2.0 slot. We use RedHat Enterprise Edition 5 (RHEL5) with kernel 2.6.18-8.el5 as the operating system on these nodes. In addition, OpenFabrics [15] Enterprise Edition (OFED) 1.3 is used to provide the InfiniBand interface stack.

As of this writing, there is no QDR-capable switch widely available commercially. As such, QDR can only be evaluated in a back-to-back configuration with no switch present. To provide an equal platform for evaluation between all configurations, the back-to-back configuration is used in all cases. This is also the reason why all experiments only use two nodes.

Since the focus of our evaluation is on high-performance computing, we evaluate performance using benchmarks built on the Message Passing Interface (MPI) [12], which is the most prominent parallel programming model used in high-performance computing. For our evaluation we use MVAPICH [14], a derivative of MVICH [9] (an MPI over VIA [4]), from the Ohio State University that is optimized significantly for InfiniBand and is used by over 700 organizations worldwide.

5 Experimental Evaluation

In this section we present the results of our experimental evaluation. We start our evaluation with microbenchmark results, focusing on the impact of MPI level latency and bandwidth achieved on the PCI Express 2.0 systems. Than we use more comprehensive benchmarks, the NAS Parallel Benchmarks [2] to show the impact of this new architecture.

5.1 Microbenchmarks

In this section we investigate each of our configurations using various microbenchmarks to analyze the basic performance and scaling.

***Latency and Bandwidth**:* We start our performance evaluation with basic microbenchmark results. In Figure 2 we show the latency and bandwidth comparison of all the four configurations mentioned in the previous section.

From Figure 2(a) we can clearly observe the impact of reduced latency with the PCI-Express 2.0 architecture. DDR-2.0 and QDR-2.0 reduce the small message latency by approximately 0.2μsec. QDR-2.0 is able to achieve 1.06μsec for 1 byte message. DDR-2.0 performs comparably with QDR-2.0 for messages smaller than 64 bytes, but shows worse latency after that. This is because at 128 bytes MVAPICH stops using an InfiniBand feature that allows small amounts of data to be *inlined*. InfiniBand allows small messages to be "inlined" directly with the send request. When this is done only a single PIO operation to the HCA is required. When inlining is not used, above 128 bytes in this case, an additional DMA transaction is required. Note that this inline value can be increased at the cost of additional memory usage for InfiniBand userspace resources. Thus, the faster signaling speed of QDR allows that configuration lower latency.

Figure 2(b) and Figure 2(c) show similar results. Not surprisingly, both DDR signaling and the extra bandwidth with PCI Express 2.0 benefit message throughput. While the bandwidth and bi-directional bandwidth of PCI Express 1.0 peak at 1301 MB/s [1] and 2551 MB/s with DDR signaling, PCI-Express 2.0 is able to achieve a higher throughput at 1942 MB/s and 3873 MB/s, respectively. The QDR-2.0 configuration further improves the throughput to 2575 MB/s and 5023MB/s for uni- and bi-directional. It is to be noted that the QDR-2.0 configuration does not achieve double the throughput the DDR-2.0 configuration. This is because the throughput cap of PCI Express 2.0. Further revisions, including PCIe 3.0 are anticipated to include even higher signaling rates.

***Multi-pair Latency and Bandwidth**:* Next we look at multi-pair latency and bandwidth. These are important metrics since with multi-core architectures it is very common to host multiple computing processes per physical node. The multi-pair tests will reveal how quickly the HCA can handle the network traffic. We launch 2, 4, or 8 MPI processes on each physical node. Each MPI process on a node is paired with one process on the other node. We then run latency and bandwidth test simultaneously between these pairs of MPI processes.

In Figure 3, we observe the same trend as in the basic latency and bandwidth tests. Using the 4-pair performance as example, QDR-2.0 achieves the best latency at 1.07μsec for 1 byte messages. For throughput, the peak values (aggregated from all 4 pairs) reported are almost the same as the above basic tests since it is bounded by the bandwidth supported by the PCI Express bus. The major difference here is for medium-sized messages. For example, at 2K bytes, the throughput capability of DDR-2.0 configuration reaches 1872 MB/s as shown in Figure 3(e). While for QDR signaling rate, the bandwidth can be further increased to 2103 MB/s.

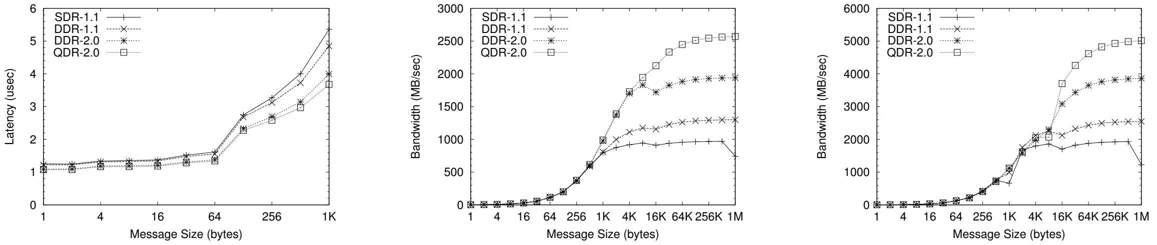
***Latency and Bandwidth Scaling Tests**:* As we have mentioned previously, the capability to support simultaneous network traffic from multiple MPI processes on a node is an important metric for multi-core computing platforms. We have presented comparison with different signaling rates and PCI Express speeds. Here we show direct comparisons of how the latency and bandwidth scale with multi-pair traffic under the same configuration.

Figure 4 shows the multi-pair latency under different configurations. As we can see, all configurations are able to maintain the same level of latency up to 4 pairs. Starting from 8 pairs, however, latency with all configurations almost doubles. This could be due to a limit on the maximum number of PCI transactions per second the hardware can handle. Note that messages smaller than 64 bytes still show the same level of latency due to the effect of inline messages. Recall that messages that are not inlined require an additional DMA operation, which must traverse the PCI Express channel.

The multi-pair aggregated bandwidth is shown in Figure 5. As we can observe from the figures, PCI Express 2.0 (Figures 5(c) and 5(d)) allows very good scaling for medium-sized messages (64 to 4K bytes) as compared with PCI Express 1.1 systems (Figures 5(a) and 5(b)) because of the increased signaling rate. The throughput can not be further enhanced with 4 or 8 pairs due to the PCI transaction limits.

With many high-performance computing platforms now commonly reaching 8 and 16 cores per computing node, the fact that multi-pair latency and bandwidth

[1] MB/s stands for Million Bytes (10e+06) in this paper.



(a) Latency

(b) Bandwidth

(c) Bidir-bandwidth

Figure 2. Latency and bandwidth comparison

do not scale beyond more than 4 pairs of simultaneous communication suggests a critical communication bottleneck remains to be addressed. We believe this is an area that must be addressed for next-generation adapters and platforms.

5.2 NAS Parallel Benchmarks

In this section we show the performance results of NAS Parallel Benchmarks (NPB) [2]. NPB is developed at NASA and contains a set of benchmarks which are derived from the computing kernels common on Computational Fluid Dynamics (CFD) applications.

As we can observe from some of the communication sensitive benchmarks in Figure 6, including Integer Sort (IS), Fast Fourier Transform (FT), and Conjugate Gradient (CG), the increased performance coming with DDR signaling and PCI Express 2.0 is transformed to application-level performance improvement. NAS IS is very sensitive to the communication performance of large messages. Here we can observe that DDR-2.0 outperforms DDR-1.1 and SDR-1.1 configurations by 25.9% and 19.8%, respectively, when intranode communication is through network loopback (NoSMP case). And QDR is able to improve another 3.1% compared with DDR-2.0 case. Other benchmarks do not have as large communication portions as IS, but comparing with the SDR-1.1 configuration, QDR is still able to achieve an improvement of 9.2% for FT and 6.7% for CG.

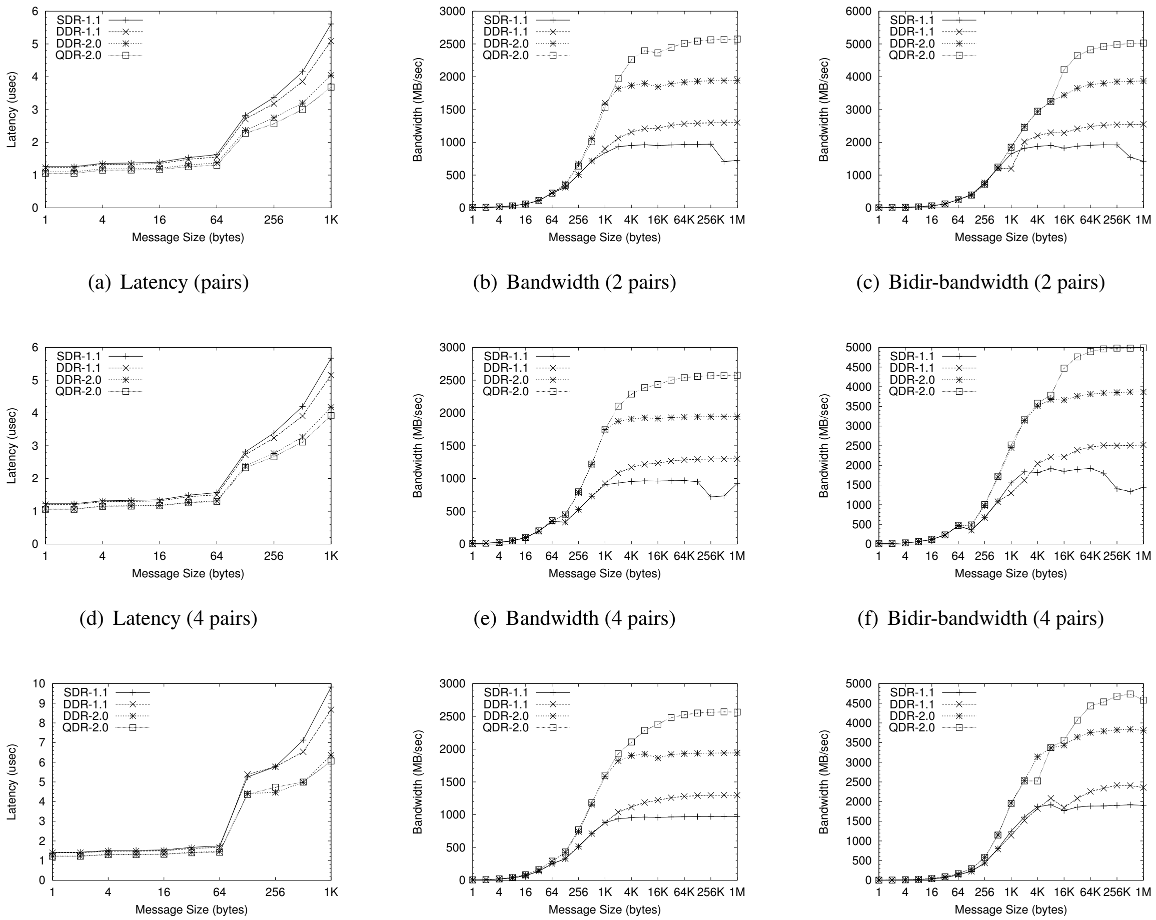
Another interesting observation is that with the increased performance brought by PCI Express and QDR leads to similar or better performance by using network loopback instead of shared memory for intra-node communication. Focusing on the QDR configuration, IS shows 5.9% better performance using network loopback instead of using shared for intra-node communication. For other benchmarks the performance difference is very similar and in the worst case it shows only 1.5% degradation for CG. In contrast, shared memory configuration significantly outperforms network loop back for the slower SDR-1.1 configuration, up to 7.0% for CG and 8.4% for IS. This suggests an interesting alternative for designing the communication library on the newer generation systems with fast signaling rates (QDR) and PCI Express 2.0. A simpler design using network for both intra- and inter- node communication may be able to achieve better performance than using a more complicated multi-communication method in some cases.

6 Related Work

In this section we discuss related work in this area.

Liu, et al., provided a performance evaluation of PCI-X based InfiniBand adapters versus the first-generation PCI-Express InfiniBand adapters [10]. Each of these adapters only operated at SDR speeds. Our work shows the benefits of PCI-Express 2.0 and additionally explores the benefit of increased bandwidth available through DDR and QDR. Previous evaluation by Surs, et al., compared ConnectX and InfiniHost III HCAs from Mellanox and showed performance improvements for multi-core systems [19]. While our work also uses the ConnectX HCA, we are comparing data rates and PCIe technology instead of different HCAs.

There have been many other studies of interconnect technologies in the past as well. QLogic (previously PathScale) are the developers of InfiniPath, another InfiniBand adapter. InfiniPath was evaluated in [5] by Dickman et al. and in [3] by Brightwell et al. The InfiniPath adapter in these work works only at SDR speeds. Myrinet MPICH-MX [13] is an alternative to InfiniBand. Their unique approach of "partial-offload" results in very good latency and bandwidth. Applications must be written using an MPI-like interface (MX) which includes message tag matching (required for MPI). In [6], Doerfler et al. showed that MPICH-MX has low overhead for a posted send and receive.



(a) Latency (pairs)

(b) Bandwidth (2 pairs)

(c) Bidir-bandwidth (2 pairs)

(d) Latency (4 pairs)

(e) Bandwidth (4 pairs)

(f) Bidir-bandwidth (4 pairs)

(g) Latency (8 pairs)

(h) Bandwidth (8 pairs)

(i) Bidir-bandwidth (8 pairs)

Figure 3. Multi-Pair latency and bandwidth comparison

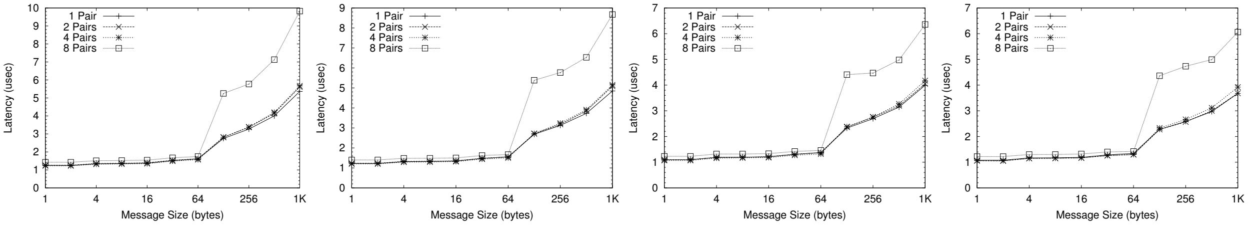
7 Conclusion

As core counts per node continue to increase all of the system components must also scale to maintain a balanced system. In a high-performance computing system the compute node interconnect plays a central role in overall performance. It is especially important for the interconnect to scale along with the rest of the system.

InfiniBand is a popular cluster interconnect for high-performance computing. InfiniBand bandwidth on commodity clusters, however, has been limited due to the 8x PCIe slots found on most platforms. The PCIe 2.0 specification, however, has doubled the data rate. This allows InfiniBand to scale the overall system bandwidth.

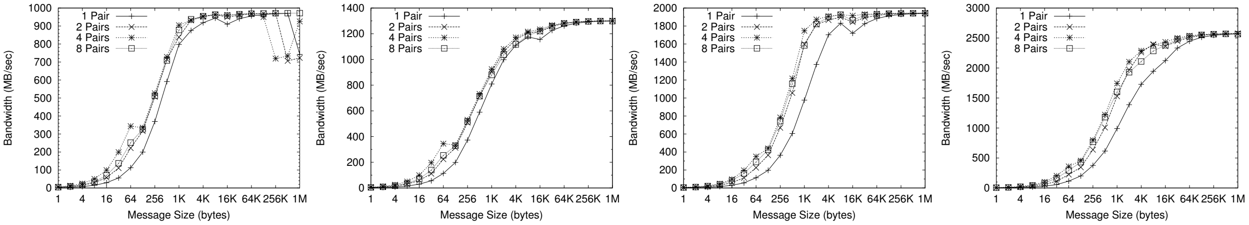
In this work we have evaluated the effect of increased interconnect bandwidth on various benchmarks. We additionally address the benefits of PCIe 2.0. We show an QDR performance reaching 2.5 GB/sec uni-directional and over 5.0 GB/sec bi-directional. We also observe near 1μsec one-way MPI latency when using QDR over PCIe 2.0. We show that even using the DDR interface, PCIe 2.0 enables a 25% improvement in NPB IS performance. Furthermore, we show that when using QDR on PCIe 2.0, network loopback can outperform a shared memory message passing implementation. On IS we show a 6% improvement using QDR loopback as compared to the shared memory implementation. This is in contrast to slower data rates, such as SDR, where the shared memory implementation is up to 8.4% slower than network-loopback.

In the future we wish to further explore the effect of interconnect bandwidth at larger scale on application performance. We also plan to to evaluate further QDR performance at scale when a QDR-enabled switch becomes available to us.



(a) SDR-1.1 (b) DDR-1.1 (c) DDR-2.0 (d) QDR-2.0

Figure 4. Latency scaling to multiple concurrent pairs



(a) SDR-1.1 (b) DDR-1.1 (c) DDR-2.0 (d) QDR-2.0

Figure 5. Bandwidth scaling to multiple concurrent pairs

References

[1] Mellanox Technologies. http://www.mellanox.com.

[2] D. H. Bailey, E. Barszcz, J. T. Barton, D. S. Browning, R. L. Carter, D. Dagum, R. A. Fatoohi, P. O. Frederickson, T. A. Lasinski, R. S. Schreiber, H. D. Simon, V. Venkatakrishnan, and S. K. Weeratunga. The NAS parallel benchmarks. volume 5, pages 63–73, Fall 1991.

[3] R. Brightwell, D. Doerfler, and K. D. Underwood. A Preliminary Analysis of the InfiniPath and XD1 Network Interfaces. In *Workshop on Communication Architecture for Clusters, held in conjunction with IPDPS*, 2006.

[4] Compaq, Intel, and Microsoft. VI Architecture Specification V1.0, December 1997.

[5] L. Dickman, G. Lindahl, D. Olson, J. Rubin, and J. Broughton. PathScale InfiniPath: A First Look. In *13th Symposium on High Performance Interconnects (HOTI)*, Palo Alto, CA, 2005. IEEE Computer Society Press.

[6] D. Doerfler and R. Brightwell. Measuring MPI Send and Receive Overhead and Application Availability in High Performance Network Interfaces. In *13th European PVM/MPI Users' Group Meeting*, Bonn, Germany, 2006.

[7] Gordon Moore. Moore's Law. http://www.intel.com/technology/mooreslaw/index.htm.

[8] InfiniBand Trade Association. InfiniBand Architecture Specification. http://www.infinibandta.com.

[9] Lawrence Berkeley National Laboratory. MVICH: MPI for Virtual Interface Architecture. http://www.nersc.gov/research/FTG/mvich/ index.html, August 2001.

[10] J. Liu, A. Mamidala, A. Vishnu, and D. K. Panda. Performance Evaluation of InfiniBand with PCI Express. In *Hot Interconnect 12 (HOTI 04)*, August 2004.

[11] Mellanox Technologies. ConnectX Architecture. http://www.mellanox.com/products/connectx_architecture.php.

[12] Message Passing Interface Forum. *MPI: A Message-Passing Interface Standard*, Mar 1994.

[13] Myricom Inc. MPICH-MX. http://www.myri.com/scs/download-mpichmx.html.

[14] Network-Based Computing Laboratory. MVAPICH: MPI over InfiniBand and iWARP. http://mvapich.cse.ohio-state.edu.

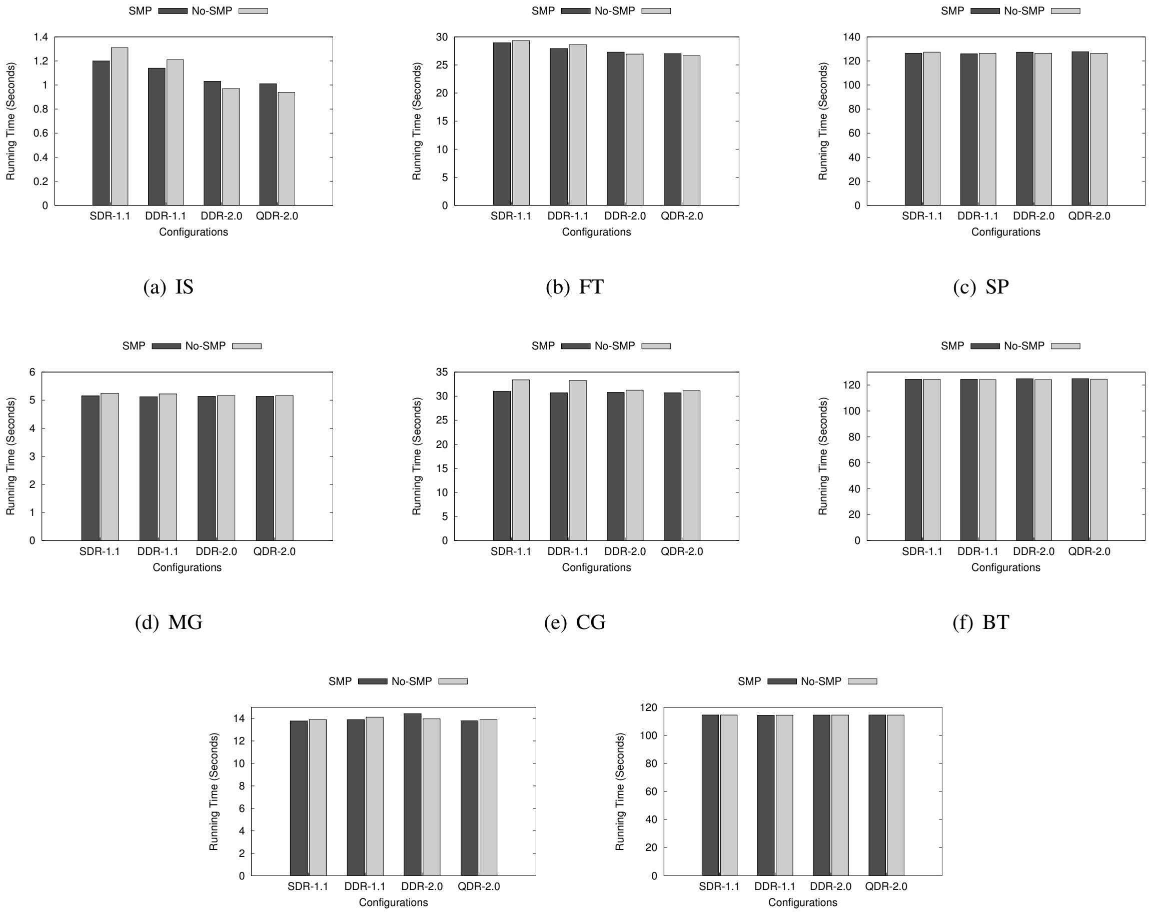
[15] OpenFabrics Alliance. OpenFabrics. http://www.openfabrics.org/, April 2006.

[16] PCI-SIG. Creating a PCI Express Interconnect (Whitepaper). http://www.pcisig.com/specifications/pciexpress/resources.

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

[18] PCI-SIG. PCI Express Base 2.0 Specification. http://www.pcisig.com/specifications/pciexpress/base2.

[19] S. Sur, M. Koop, L. Chai, and D. K. Panda. Performance Analysis and Evaluation of Mellanox ConnectX InfiniBand Architecture with Multi-Core Platforms. In *15th IEEE Int'l Symposium on Hot Interconnects (HotI15)*, Palo Alto, CA, August 2007.



(a) IS

(b) FT

(c) SP

(d) MG

(e) CG

(f) BT

(g) EP

(h) LU

Figure 6. NAS Parallel Benchmarks (Class B), 16 Processes on 2 nodes