Experimental Analysis of InfiniBand Transport Services on WAN

Weikuan Yu, Nageswara Rao, Jeffrey Vetter
Computer Science and Mathematics
Oak Ridge National Laboratory
Oak Ridge, TN 37831-6173
{wyu,raons,vetter}@ornl.gov

Abstract

InfiniBand Architecture (IBA) has emerged as a standard system-area interconnect in industry for both data-center and high performance computing. While IBA continues to evolve with more capabilities, it has yet to permeate the field of grid and pervasive computing across wide area. This requires a software infrastructure to take the full advantage of IBA networking mechanisms without being dwarfed by the long distances. In this paper, we present a detailed analysis of InfiniBand transport services for their suitability to data transfer and message passing in the wide area. Three of the common InfiniBand transport services have been examined, including Reliable Connection (RC), Unreliable Connection (UC), and Unreliable Datagram (UD). Our analysis indicates that UC and UD are better suited to provide high bandwidth for MPI on wide-area networks (WAN). We have also demonstrated that the current existing MPI implementations, designed over RC and UD, can be tuned to provide improved MPI bandwidth on WAN. Furthermore, by developing a UC-based MPI implementation, we have shown that, at a distance of 8600 miles, MPI over UC can improve MPI bandwidth by as much as 100%.

1 Introduction

For the past few years, the computational power of commodity PCs has seen an astounding growth rate, especially with the development of multi-core chip technologies. This trend results in more and more large-scale computational systems being deployed around the world. A challenge arises due to the large data sets produced by these high-performance computing systems: they must be transported to storage, visualization and analysis systems, which are often remotely located. The enabling technologies for computing and data movement across wide-area networks (WAN) have traditionally been Ethernet and the legacy TCP/IP protocol and other tools atop, such as GridFTP [4] and bbcp [9]. For high-performance grid environments, these methods are not optimally matched, for several reasons. First, they scale poorly for high bandwidth operations that require 10Gbps throughputs over connections of thousands of miles. While these methods have been very effective for low bandwidth operations over shared IP networks, they require significant tuning and optimization for dedicated connections typically provisioned in high-performance wide-area environments. Second, these methods are characteristically different from interconnect technologies used for connecting processing nodes and/or storage systems, and hence require the users to be adept at multiple technologies that utilize disparate framing and flow control methods.

InfiniBand (IB) [11] has emerged as a frontier system-area interconnect technology in industry for both high performance data-centric and compute-centric environments. For the past few years, it has been able to keep up very closely with the speed of the processor technology by providing low latency and high bandwidth needed for both compute and data movement operations. A growing number of high performance computing platforms, such as the Ranger at Texas Advanced Computing Center [20], have been built by clustering the computational power of many commodity PCs together using IB. However, the current deployments of IB have been confined within typical connection lengths of a few meters using copper connectors, or hundreds of meters using optical connections.

Recently, a new class of devices is being produced, e.g. by Obsidian [16] Research and Network Equipment Technologies. These devices extend the reach of IB by transporting IB frames over wide-area networks. They convert network packets from IB network to SONET or Ethernet, and can be simply placed at both ends of a wide-area connection for services. Obsidian Research's Longbow XR switches have been demonstrated to support good throughput for RDMA data transmission across very long ranges [8], particularly with its low-level networking transport capabilities. Much remains to show on how the performance of IB over WAN can be affected by different network parameters, and on how effective an implementation of Message Passing Interface (MPI)[3] over IB can be on WAN.

IB offers a variety of data transport mechanisms, including the traditional send/receive, RDMA (Remote Direct Memory Access) write and read [18], as well as unreliable datagram; but questions about which ones perform effectively, under what condition, and how to tune and optimize them are largely open. In addition, MPI has been established as the *de facto* parallel programming standard. Thus, it is also desirable to investigate how well IB supports MPI over WAN. Three common IB transport services are analyzed on

978-0-7695-3187-8/08 $25.00 © 2008 IEEE
DOI 10.1109/NAS.2008.65

WAN, including Reliable Connection (RC), Unreliable Connection (UC), and Unreliable Datagram (UD). Our objectives are to answer the following imminent questions for IB on WAN, including (1) whether and how the performance characteristics of these transport services would differ on WAN, compared to the enterprise-level deployments; (2) which transport service over IB will provide the best performance for WAN; and (3) whether and how the upper level programming models, such as MPI, over IB should be customized to the long latency, and large latency-bandwidth products that are common on WAN.

Our characterization and analysis of IB over WAN in this paper is carried out using the UltraScience Net (USN), which provides OC192 connections of various lengths up to 8,600 miles. In particular, we collect benchmark measurements on both RDMA data transfers as well as MPI operations with 4x (peak 8Gbps) IB links over OC192 connections of lengths 1,400 and 8,600 miles. In particular, we examine the performance characteristics of RC, UC and UD on long-range WAN, and the performance impact of the number of concurrent connections called Queue Pairs (QP). We found that UC and UD are better suited for WAN. In the same environment, we also analyze the performance of existing MPI implementations on top of RC and UD. We show how the bandwidth of MPI on WAN can be affected by different parameters. Furthermore, using an in-house prototyped MPI implementation over UC, we demonstrate that UC offers as much as 100% bandwidth improvements for MPI over WAN.

The rest of the paper is organized as follows. In the next section, we provide an overview of IB. A detail dissection of IB transport services is provided in Section 3. In Section 4, we describe our testing environment in detail. In Sections 5, we present the network-level performance analysis of IB transport services. In Section 6, we present the MPI-level experimental results. Section 7 presents related work. Finally, we conclude the paper in Section 8.

2 An Overview of InfiniBand

The InfiniBand Architecture (IBA) [11] is an open specification designed for interconnecting compute nodes, IO nodes and devices in a system area network. As shown in Figure 1, it defines a communication architecture from the switch-based network fabric to transport layer communication interface for inter-processor communication. Processing nodes and I/O nodes are connected as end-nodes to the fabric by two kinds of channel adapters: Host Channel Adapters (HCAs) and Target Channel Adapters (TCAs). Currently, only HCAs are fabricated from various companies like Mellanox and IBM. The collection of nodes and switches is referred to as an IB subnet, which is how the most, if not all, of the IB platforms have been configured. IBA specification stipulates that IB subnets can be connected into a global IB fabric through IB routers. However, the IB router implementation is at its very initial phase, and a production release is yet to be awaited. IBA specifies a wealth of transport services and protocols in its communication stack. We will dissect more into these transport services in Section 3.

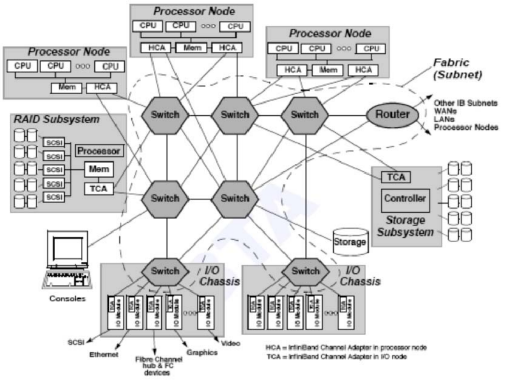


Figure 1 Diagram of IB Subnet (Courtesy of IBTA)

The OpenFabrics Enterprise Distribution (OFED) [1] supports IB. OFED is developed and maintained by the OpenFabrics Alliance [1], which is an organization that promotes the development of RDMA [18] interconnect technologies, including both IB [11] and iWARP [19]. OFED includes software packages that support a broad range of environments, including message passing, file system and storage. In this paper, we utilize the low level networking communication libraries and MVAPICH [15].

3 Dissections of IB Transport Services

IB provides five types of transport services: Reliable Connection (RC), Reliable Datagram (RD), Unreliable Connection (UC), Unreliable Datagram (UD), and Raw Transport. Among them, RC and UC are connection-oriented, a process has to establish a distinct connection composed of a pair of send and receive queues for every process with which it is communicating. The other three are connectionless, which means one QP can be used to communicate with any other peers. RD and Raw, though defined in the IBA specification, are not currently implemented. We will not discuss RD and Raw in the rest of the paper.

RC supports both type of communication semantics: send/receive and RDMA, along with network atomic operations. It is also the most commonly used service, due to its reliability, high performance, and its rich features. Only a single packet

can be communicated on top of UD, which means its operation can carry at most 2048 bytes (the largest packet size currently supported on IB). Still, UD is gaining more interest due to its advantages in supporting hardware multicast and scalable connection management [13, 22]. The availability of IB in the wide area raises an imminent set of questions on whether the performance of IB transport services would remain the same on WAN. Research on IB in the wide area is at its nascent phase [8]. The distinct difference on WAN is the physical distance between two sites. Roughly speaking, there will be a 20,000 microsecond round-trip delay for every 1000 miles. This is an overwhelming number compared to the microsecond latency that IB is capable of achieving within a local area network. It is necessary to take a close look at the level of network transactions in order to understand the implications of WAN distance to IB.

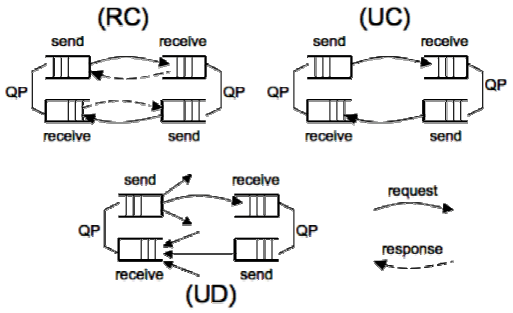


Figure 2 A Comparison of IB Transport Services

Figure 2 shows a comparison of IB transport services. The hallmark of RC is the reliable, in order and corruption-free delivery of messages. This reliability machinery is the combination of a packet sequencing protocol and the NAK/ACK protocol [11]. Every request packet is sent with a packet sequence number (PSN). Each one via RC is expected to have either an explicit or implicit response packet, serving as an acknowledgement. An acknowledgement for a packet of larger PSN implicitly acknowledges earlier packets. While this two-way reliability protocol works well within LAN, it presents a hindrance to the performance across wide area. The PSN protocol also enforces a limit on the window of outstanding, unacknowledged PSNs. This indirectly thralls the number of concurrent messages for a single QP. So it is questionable whether it would be wise to maintain the same reliability guarantees on WAN for individual IB packets, which we will dwell further through the experimental analysis in Section 5. In contrast, this is not the case for UC and UD. All packets will be injected into the IB network as soon as WQEs are processed by the HCA. A UD QP can also send data to and receive data from any other peer QPs. It is debatable whether the unreliable IB transport services can sustain a network or physical environment that can trigger a very high bit transmit error. However, the UC and UD services do provide its consumer the freedom of achieving the best bandwidth in a given environment. The needed reliability can be provided by the higher-level libraries or applications. Sections 5 and 6 provide a detailed experimental analysis on the performance of these transport services on WAN.

4 Configuration of UltraScience Net

UltraScience Net [2] (USN) is a wide-area experimental network testbed that supports the development of next-generation computational science applications. It spans more than four thousand miles from Oak Ridge (Tennessee) to Atlanta, Chicago, Seattle and Sunnyvale (California) using dual OC192 backbone connections as shown in Figure 4(L). USN provides dedicated OC192 (9.6Gbps) channels for large data transfers, and also high-resolution, high precision channels for fine control operations. Further details are available from the cited USN website [2]. In this study, we have created OC192 connections over connections of lengths 1400 and 8600 miles by creating loops from Oak Ridge to Chicago and Sunnyvale, respectively. These loop-back connections are realized using the SONET switching capability of Ciena CDCI devices. These connections are dynamically provisioned using automated scripts.

Two Longbow XR devices are configured for IB-over-SONET operation, and are connected to CDCIs on WAN side and to two IB switches at Oak Ridge. Figure 3 (R) shows the diagram of the configuration of our test environment. The first IB switch is a Flextronics DDR switch with 144-ports; the other is a Cisco SDR switch with 24 ports. Each cluster was running its own IB subnet manager. The Flextronics cluster was running OpenSM from OpenFabrics Stack on another node; the Cisco cluster was running an embedded subnet manager inside the switch. Two Intel WoodCrest nodes were used in this study, each connected to one of the IB switches. The Woodcrest nodes contained two Xeon 5150 dual-core processors and 8 GB of FB-DIMM memory. The processor clock rate was 2.66 GHz. These computer nodes are equipped with Mellanox 4x DDR InfiniHost HCAs, which are connected to the PCI-Express x8 Interface.

5 Performance Characterization of IB Transport Services on WAN

The RC transport service has been a primary choice for network I/O within local area networks due to its performance benefits and reliability. However, as dissected earlier, the reliability of different protocols may change this conventional wisdom on WAN dramatically. In this section, we provide a performance characterization of IB transport services at different

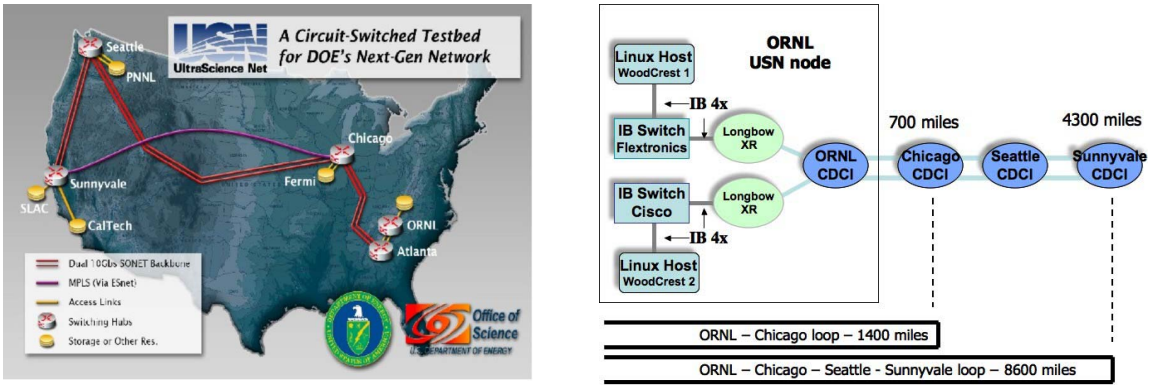


Figure 3 Configuration of Test Environment for InfiniBand on WAN: (L) USN; (R) Our Test Bed

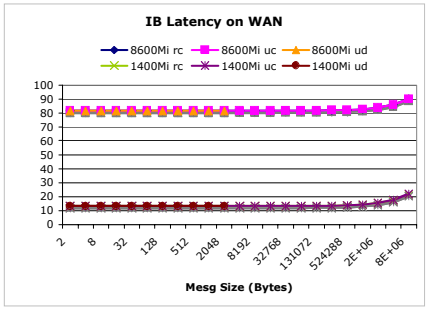


Figure 4 IB Send/Receive Latency on WAN

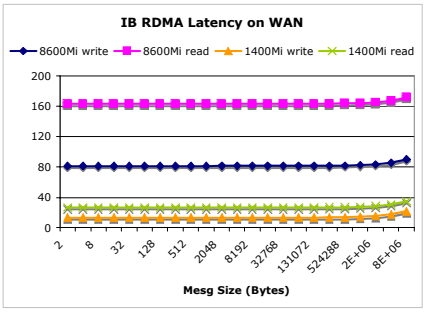


Figure 5 IB RDMA Latency on WAN

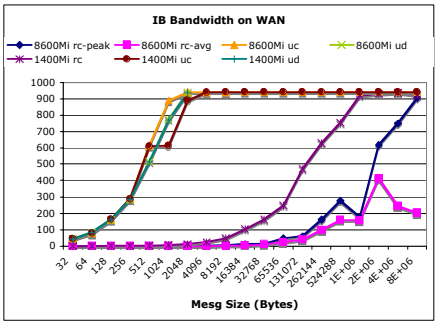


Figure 6 IB Send/Receive Bandwidth on WAN

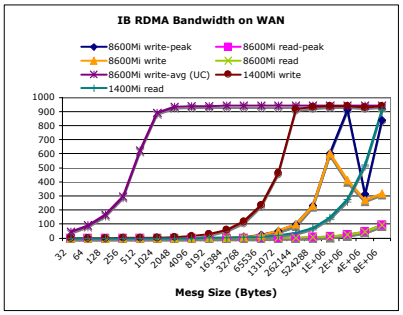


Figure 7 IB RDMA Bandwidth on WAN

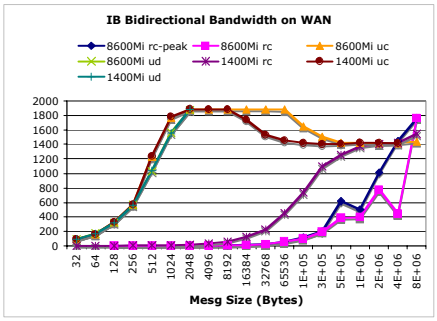


Figure 8 Bidirectional Send/Receive Bandwidth on WAN

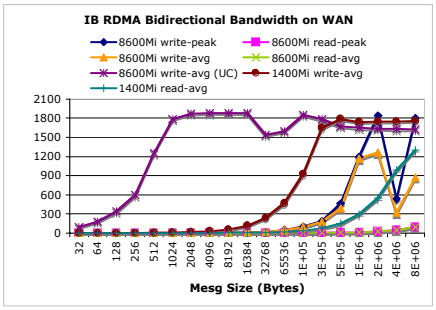


Figure 9 RDMA Bidirectional Bandwidth on WAN

distances on WAN. We used the OFED release version 1.2.5.4 for examining the network-level performance of IB transport services.

5.1 Network-Level Latency

We measured the latency of send/receive operations on top of three IB transport services: RC, UC, and UD. As shown in Figure 4, the latency using any of the three is 13 milliseconds at 1400 miles, 81.5 milliseconds at 8600 miles. The original latency differences (generally within 10usec) among these transport services are immaterial compared to the huge latency introduced by the physical distance. We also measured the latency of RDMA operations. As shown in Figure 5 the latency for RDMA is also determined by the physical distances for both reads and writes. RDMA read is twice as costly due to its communication nature of traveling through a round trip. Together, these results suggest that, as far as latency is concerned, there is no compelling reason to choose one transport service over another for designing IB-based communication libraries or storage protocols on WAN.

5.2 Network-Level Bandwidth and Bidirectional Bandwidth

We measured the bandwidth of send/receive operations on top of three IB transport services, RC, UC and UD. Figure 6 shows the bandwidth comparisons of send/receive operations at both 1400 miles and 8600 miles. The UC-based send/receive bandwidth can reach the peak rate of 941 MB/sec. With RC, the peak bandwidth is 935 MB/sec and 906 MB/sec, at 1400 miles and 8600 miles, respectively. However, at 8600 miles, RC can achieve good throughput only with message sizes 1MB or bigger. Also note that, at 8600 miles, there is a big variation of bandwidth results. Figure 6 includes both the peak and average numbers for RC bandwidth at 8600 miles. We also measured the latency of the RDMA operations on top of RC and UC. As shown in Figure 7, RDMA read has a very low performance. There are two factors involved for this. One is that an RDMA read operation takes a round-trip to complete; the other, performance-killing, factor is that there is a low limit on the number of concurrent RDMA read operations, typically set at 4 or other low numbers. These factors severely limit RDMA read from achieving a good throughput. Therefore RDMA read is not well qualified for network I/O on WAN. On top of RC, RDMA write has better performance than the send/receive operations at long distances. As is the case for send/receive operations, UC also supports much faster data transfer for RDMA write operations. Note that, the drastic drop of RC bandwidth only happens across long distances on WAN. As explained in Section 3, this is due to the thwarting effect of the reliability mechanisms in RC. In contrast, UC and RC have comparable performance within LAN, and RC is preferred because of its additional reliability.

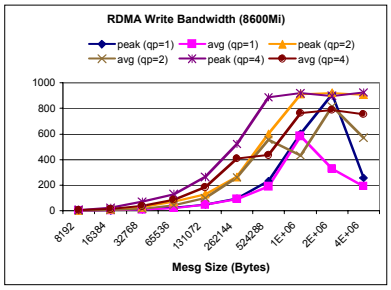
We have measured the bidirectional bandwidth of send/receive and RDMA operations on top of RC, UC and UD. Figure 8 and Figure 9 show the results that are collected in a similar way like the unidirectional case. The bidirectional bandwidth performance on top of RC, UC and UD exhibit similar characteristics like the unidirectional case. UC and UD can achieve the peak rate of 1887 MB/sec; while RC can achieve a peak rate of 1761 MB/sec. At 8600 miles, only measurements with large messages can reach its peak rate.

5.3 Tuning the Network-Level Bandwidth

In view of the low bandwidth for RC on WAN, we explored a number of different tuning options to improve it. The objective is to increase the number of concurrent messages as much as possible. There are two possible methods to achieve this. One is to increase the number of concurrent operations for an IB QP. The other is to increase the number of the concurrent QPs. The programs available in OFED provide options to examine the impacts of these tuning methods. The PSN protocol for RC reliability limits the number of outstanding IB packets for a given QP. So only the second method achieves the desired effects on WAN. Figure 10 shows the impact of using concurrent QPs for IB RDMA on WAN. In both unidirectional and bidirectional cases, the bandwidth of RDMA write has been improved, especially for lower or medium messages. By employing more concurrent QPs, more messages are successfully injected onto the wire for better throughput. Using 4 QPs and RC-based RDMA, at 8600 miles, we have measured the highest sustained unidirectional bandwidth as 934 MB/sec; the highest sustained bidirectional bandwidth as 1762 MB/sec.

6 MPI Using Different IB Transport Services on WAN

There are several popular MPI implementations over IB. Both RC and UD have been utilized to support data movement in MPI. In this section, we used the popular MPI implementations from the Ohio State University, MVAPICH-0.9.7 and MVAPICH-1.0-beta. The former is used for RC-based MPI evaluation; the latter for the UD-based MPI evaluation. They are referred as MVAPICH/RC and MVAPICH/UD, respectively. These choices are made solely based on conveniences and the availability of UD in MVAPICH-1.0-beta. Except the difference in the underlying transport services, these implementations have virtually no additional differences in terms of MPI-level



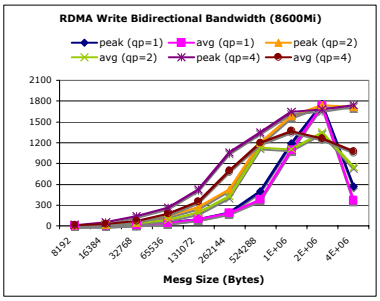
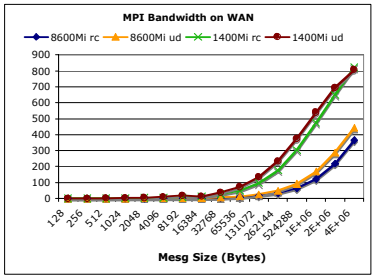


Figure 10 The Impact of Concurrent Queue Pairs



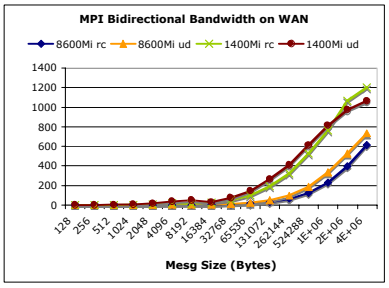
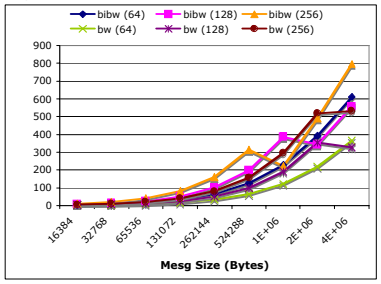


Figure 11 MPI Bandwidth over Different IB Transport Services: (L) Unidirectional; (R) Bidirectional



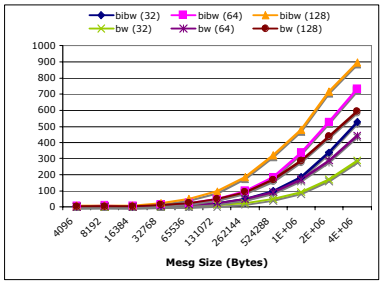
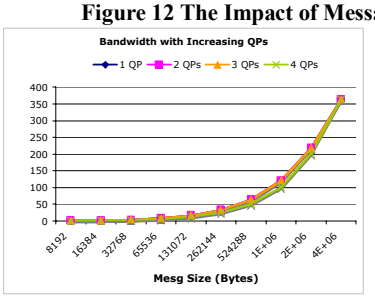


Figure 12 The Impact of Message Window Size to MPI Bandwidth: (L) RC; (R) UD



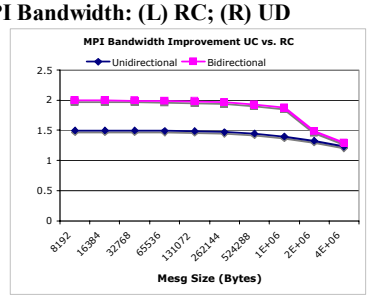
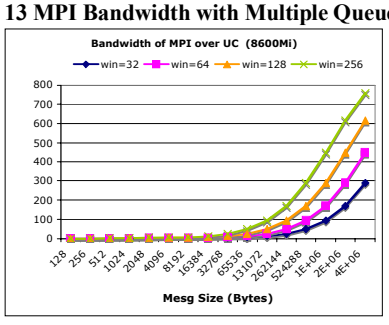


Figure 13 MPI Bandwidth with Multiple Queue Pairs

Figure 14 MPI Bandwidth using UC compared to RC



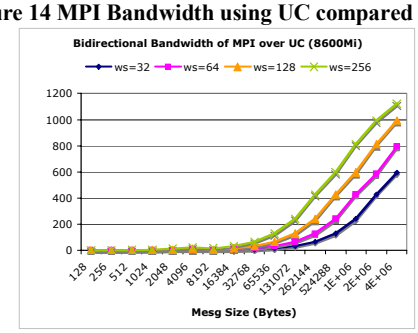


Figure 15 MPI Bandwidth over UC with Varying Message Window Sizes: (L) Unidirectional; (R) Bidirectional

implementation overhead. We used the MPI latency and bandwidth benchmarks distributed in these implementations. Similar to the observation made in Section 5, the original latency differences (generally within 10usec) among these transport services are immaterial for MPI on WAN. The physical distance determines the MPI-level latency, which is 13 milliseconds at 1400 miles and 81.6 milliseconds at 8600 miles. In addition, the large-message latency is tripled compared to that of the small ones because the three-way rendezvous protocol for large messages.

6.1 MPI Unidirectional Bandwidth and Bidirectional Bandwidth

We measured the MPI bandwidth and bidirectional bandwidth on top of RC and UD. Figure 11 shows the comparisons at both 1400 miles and 8600 miles. At 1400 miles, MVAPICH/RC and MVAPICH/UD can achieve 822 and 808MB/sec, respectively, while at 8600 miles, the achieved bandwidths are 364 MB/sec and 443 MB/sec. At long distances, MVAPICH/UD attains better performance for both the unidirectional and bidirectional bandwidths. However, compared to the bandwidth improvement of UD to RC at the network-level, the MPI-level improvement is much reduced. This is a side effect of MPI processing, which has to enable message fragmentation and reliability in order to utilize UD for messages of arbitrary lengths.

Tuning the MPI Bandwidth: Similar to the network-level tuning, we examined several different methods for the MPI bandwidth tuning. These include: (1) increasing the number of concurrent MPI messages; (2) making use of concurrent QPs; and (3) adjusting the threshold of MPI rendezvous protocol. The benchmark test programs in MVAPICH, by default, post a window of 64 concurrent messages before waiting for the completion of the entire window. In experimenting method (1), we have varied the window size to be 64, 128 and 256. The second and third methods are available via environmental variables in MVAPICH. Figure 13 illustrates the case for method (2). Neither using multiple QPs nor adjusting the thresholds improves the bandwidths for MPI over UD and RC. In contrast, the number of concurrent messages does impact the bandwidth. Figure 12(L) shows the results of unidirectional and bidirectional bandwidths *(bw and bibw)* with MVAPICH/RC. At 8600 miles, both uni- and bi-directional bandwidths have been improved with more concurrent messages, reaching 533 MB/sec and 795 MB/sec, respectively. Figure 12(R) shows the impact of an increasing window size to the bandwidths of MVAPICH/UD at 8600 miles. Larger window sizes improve both uni- and bi-directional MPI bandwidths, reaching 593 MB/sec and 896 MB/sec, respectively. Note that, to prevent a bandwidth drop of MVAPICH/UD at large window sizes, we increased the environmental variable MV_UD_ZCOPY_QPS to 128. This is needed for the MVAPICH/UD buffer management to handle the burst pressure from many concurrent messages.

6.2 MPI over UC for WAN

In view of the high bandwidth of UC on WAN, we prototyped a UC-based MPI based on MVAPICH 0.9.7. The main purpose was to demonstrate the suitability of UC for message passing on WAN. We customized the connection establishment in MVAPICH to make use of UC for date transfer. In addition, to cope with a limited number of losses per message, an acknowledgment is required before a UC-based data message is marked for completion. Figure 14 shows the improvement factor for MPI over UC compared to RC. For most middle-size messages, MPI over UC improves the bandwidth by as much as 100% with a default window size of 64 in the bandwidth tests. Figure 15 shows the absolute bandwidth results (unidirectional and bidirectional) with MPI over UC when the window size varies between 32 and 256. With an increasing window size, the peak unidirectional and bidirectional bandwidths at 8600 miles are much improved, reaching 758 MB/sec and 1119 MB/sec, respectively. Together these results suggest that UC is an ideal candidate for high-bandwidth MPI on WAN.

7 Related Work

There is a rich set of literature on the analysis of network technologies both for the local area networks and wide-area networks. Many studies were carried out to study the performance of a spectrum of high performance networking technologies, including Myrinet [5], Quadrics [6, 17], IB [14], as well as 1/10Gigabit Ethernet [21]. The performance assessments in these studies were performed in a single cluster environment, focusing on intimate interactions among the processor architecture, the memory technologies, and the networking technologies. Studies on 1/10Gigabit Ethernet had also been performed both in a cluster environment and on the WAN [10].

Our previous work [8] was among the first to reveal the performance of IB on WAN. However, much of the study addressed the issues on deploying IB for a Cray XT3 system, and little analysis has been performed on how to tune the performance of IB on WAN. In addition, no attempt has been made to exploit the benefits of IB for higher-level application programming models such as MPI [3]. This work serves as a detailed study to reveal how the performance of IB on WAN can be affected by different network parameters. We also shed light on how the IB reliability machinery can impact its

performance of IB on WAN, while has little effect on the performance on LAN. Furthermore, we demonstrate how the MPI [3] programming model can take advantage of different IB transport services.

8 Conclusions

In this paper, to investigate the suitability of IB-based message passing in the wide area, we have carried out a detailed analysis of IB transport services at both the network-level and the MPI-level, using USN. The network-level analysis indicated that the differences in communication latency among the three IB transport services (RC, UC and UD) are immaterial, as is predominantly dominated by the physical distance. However, RC, UC and UC exhibit drastically different characteristics for the data throughput. UC and UD are better suited for fast data movement on WAN. On top of the same transport services (RC or UC), RDMA write exhibits the best bandwidth potential compared to Send/Receive and RDMA read. Particularly, we have documented that RDMA read is not well qualified for WAN-based data movement due to its round-trip nature and the low limit of concurrent operations. We have also evaluated several existing MPI implementations designed over RC and UD. We found that significant tuning is needed for these MPI implementations to offer better bandwidth on WAN. In view of the potential of UC, we have developed an in-house UC-based MPI implementation. At a distance of 8600 miles, we have shown that, MPI over UC can improve the MPI bandwidth by as much as 100%.

In the future, we plan to study how to utilize MPI over UC for wide-area parallel programming across multiple InfiniBand clusters. We also plan to investigate the performance potential of IB for wide-area storage protocols such as NFS over RDMA [7] and iSER [12]. Furthermore, we plan to evaluate new generations of IB technologies, such as connect-X HCAs, for wide-area data transfer. Tests of NX5010 devices from Network Equipment Technologies are currently being carried out, and their basic capabilities to reach distances of several thousand miles have been verified.

Acknowledgement

This research is sponsored by High Performance Networking Program of Department of Energy, by Department of Defense under contract to UT-Battelle LLC, and also by the Office of Advanced Scientific Computing Research; U.S. Department of Energy. The work was performed at the Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725.

References

[1] OpenFabrics Alliance, http://www.openfabrics.org/
[2] UltraScience Network, http://www.csm.ornl.gov/ultranet/
[3] MPI: A Message-Passing Interface Standard, 1994.
[4] T. G. Alliance, GridFTP 4.0, http://www.globus.org/toolkit/docs/4.0/data/gridftp/
[5] N. J. Boden, D. Cohen, and others, Myrinet: A Gigabit-per-Second Local Area Network, *IEEE Micro,* pp. 29-35, Feb 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, 2004.
[7] B. Callaghan, T. Lingutla-Raj, and A. Chiu, NFS over RDMA, in *ACM SIGCOMM 2003 Workshops*, 2003.
[8] S. Carter, M. Minich, and N. S. V. Rao, Experimental Evaluation of Infiniband Transport over Local and Wide-Area Networks, in *High Performance Computing Symposium (HPC'07)*, Norfolk, VA, 2007.
[9] A. Hanushevsky, bbcp, http://www.slac.stanford.edu/~abh/bbcp/
[10] J. Hurwitz and W.-c. Feng, Analyzing MPI Performance over 10-Gigabit Ethernet, *Journal of Parallel and Distributed Computing, Special Issue: Design and Performance of Networks for Super-, Cluster-, and Grid-Computing,* vol. 65, pp. 1253-1260, 2005.
[11] InfiniBand Trade Association, InfiniBand Architecture Specification, Release 1.2.
[12] M. Ko, M. Chadalapaka, U. Elzur, H. Shah, P. Thaler, and J. Hufferd, iSCSI Extensions for RDMA Specification, http://www.ietf.org/internet-drafts/draft-ietf-ips-iser-05.txt
[13] M. J. Koop, S. Sur, Q. Gao, and D. K. Panda, High Performance MPI Design using Unreliable Datagram for Ultra-Scale InfiniBand Clusters, in The 21st ACM International Conference on Supercomputing Seattle, WA: ACM, 2007.
[14] 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,* vol. 24, pp. 42-51, January-February 2004.
[15] Network-Based Computing Laboratory, MVAPICH: MPI for InfiniBand on VAPI Layer, http://mvapich.cse.ohio-state.edu/
[16] Obsidian Research Corporation, http://www.obsidianresearch.com/
[17] F. Petrini, W.-c. Feng, A. Hoisie, S. Coll, and E. Frachtenberg, The Quadrics network: High-performance clustering technology, *IEEE Micro,* vol. 22, pp. 46-57, January/February 2002.
[18] R. Recio, P. Culley, D. Garcia, and J. Hilland, An RDMA Protocol Specification (Version 1.0),
[19] A. Romanow and S. Bailey, An Overview of RDMA over IP, in *Proceedings of International Workshop on Protocols for Long-Distance Networks (PFLDnet2003)*, 2003.
[20] Texas Advanced Computing Center, http://www.tacc.utexas.edu/resources/hpcsystems/
[21] K. Voruganti and P. Sarkar, An Analysis of Three Gigabit Networking Protocols for Stroage Area Networks, in *Proceedings of International Conference on Performance, Computing, and Communications*, 2001.
[22] W. Yu, Q. Gao, and D. K. Panda, Adaptive Connection Management for Scalable MPI over InfiniBand, in *International Parallel and Distributed Processing Symposium (IPDPS '06)*, Rhodes Island, Greece, 2006.