White Paper

NetEffect PCI-Express 10-Gigabit iWARP Ethernet: A Performance Study

Mohammad Javad Rashti Ahmad Afsahi
Parallel Processing Research Laboratory
Department of Electrical and Computer Engineering
Queen’s University, Kingston, ON, CANADA K7L 3N6
mohammad.rashti@ece.queensu.ca ahmad.afsahi@queensu.ca

Abstract

Recently, NetEffect Inc. has introduced the first and only iWARP–enabled 10-Gigabit Ethernet Channel Adapter. This white paper presents the performance of the NetEffect PCI Express 10-Gigabit iWARP Ethernet adapter at the NetEffect verbs layer and the MPI implementation. At the verbs layer, our evaluation includes the basic latency and bandwidth tests for both Send/Receive and RDMA write operations, as well as the cost of memory registration. At the MPI level, we present the basic latency and bandwidth, aggregate bandwidth, hotspot bandwidth, and the computation/communication overlap ability for the NetEffect MPI implementation.

1. Introduction

iWARP is a set of standards enabling *Remote Direct Memory Access* (RDMA) over Ethernet. The iWARP specification proposes a set of descriptive interfaces, called iWARP *verbs*, allowing direct access to the *RDMA enabled NIC* (RNIC). iWARP supporting RDMA and OS bypass, coupled with TCP/IP Offload Engines (TOE), can fully eliminate the host CPU involvement in an Ethernet environment. With the iWARP standard and the introduction of 10-Gigabit Ethernet, there is now an alternative path to the proprietary interconnects for high-performance computing, while maintaining compatibility with the existing Ethernet infrastructure and protocols.

Figure 1 presents a block diagram of the recently introduced NetEffect PCI-Express 10-Gigabit iWARP RNIC. The NetEffect RNIC consists of a Protocol Engine integrating iWARP, TOE, and NIC acceleration logic in hardware, a RAM based Transaction Switch operating on in-flight data, and a local Memory Controller for buffering non-RDMA connections. Memory registration, generating completions, and managing errors and exceptions are part of the Protocol Engine responsibility. The RNIC core hardware is connected to a local 64/133MHz PCI-X bus that is bridged over an x8 PCI-Express I/O interconnect.

The NetEffect RNIC supports IPv4 TOE with several on-board protocol resources and up to 32 independent IP addresses. The card has 256MB on-board DDR2 memory. It can be accessed using user-level and kernel-level libraries such as NetEffect verbs, standard sockets, SDP, uDAPL, and MPI.

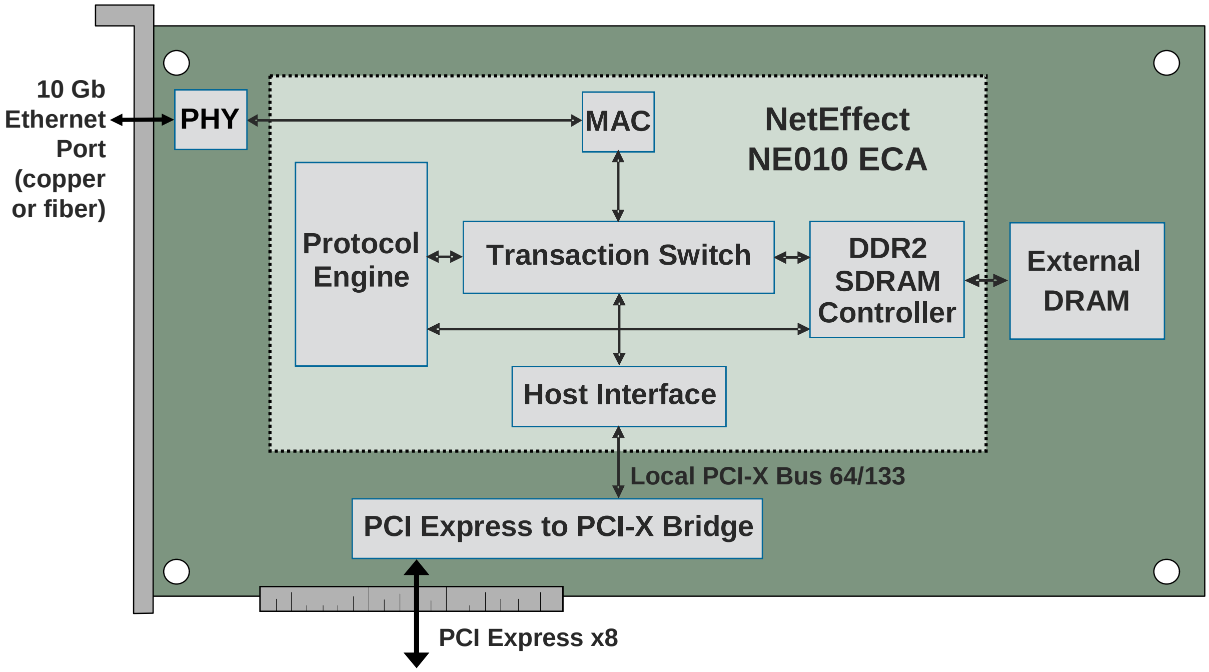


Figure 1. NetEffect NE010e Ethernet Channel Adapter architecture.

This white paper is organized as follows. The experimental framework is found in Section 2. Section 3 presents the NetEffect verb performance results. In Section 4, we analyze the NetEffect MPI implementation. Finally, Section 5 concludes this paper.

2. Experimental Framework

We have conducted our experiments using four, first generation, NetEffect NE010e 10-Gigabit Ethernet channel adapters, each with an x8 PCI-Express interface and CX-4 board connectivity. Each card is installed on an x8 PCI-Express bus of a Dell PowerEdge 2850 server. Each machine is a dual-processor Intel Xeon 2.8GHz SMP with 1MB L2-cache per processor. The total physical memory per machine is 2GB. A 12-port Fujitsu XG700-CX4 10-Gigabit Ethernet switch is used to connect the nodes together.

In terms of software, the machines run Linux Fedora Core 4 SMP for IA32 architecture with kernel version 2.6.11. The NetEffect MPI is based on MPICH2 version 1.0.3, and implemented on top of the NetEffect iWARP verbs.

3. iWARP Verbs Performance Results

This section presents the latency and bandwidth results for the NetEffect verbs layer communication in addition to memory registration/deregistration cost.

3.1. Ping-Pong Latency and Bandwidth

Figure 2 presents the latency of a typical ping-pong micro-benchmark running on two machines across the switch. Latency is defined as the time it takes for a message to travel from the sender address space to the receiver address space. The graph compares the case of using the Send/Receive verbs with the case of using RDMA write operation for sending the data. It is evident that RDMA write has a better performance than the Send/Receive operation.

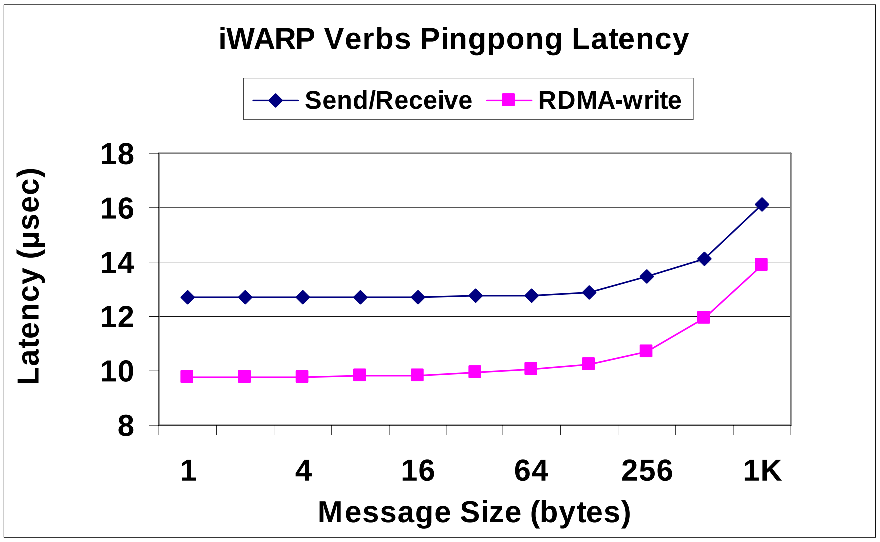


Figure 2. Verbs layer ping-pong latency.

Bandwidth results for the ping-pong benchmark are presented in Figure 3. Bandwidth is computed using the latency results. Nearly 900MB/s is achieved, which is almost 83% of the 1064MB/s available bandwidth on the internal PCI-X bus. The RDMA write has a slightly better bandwidth than that of the Send/Receive verbs.

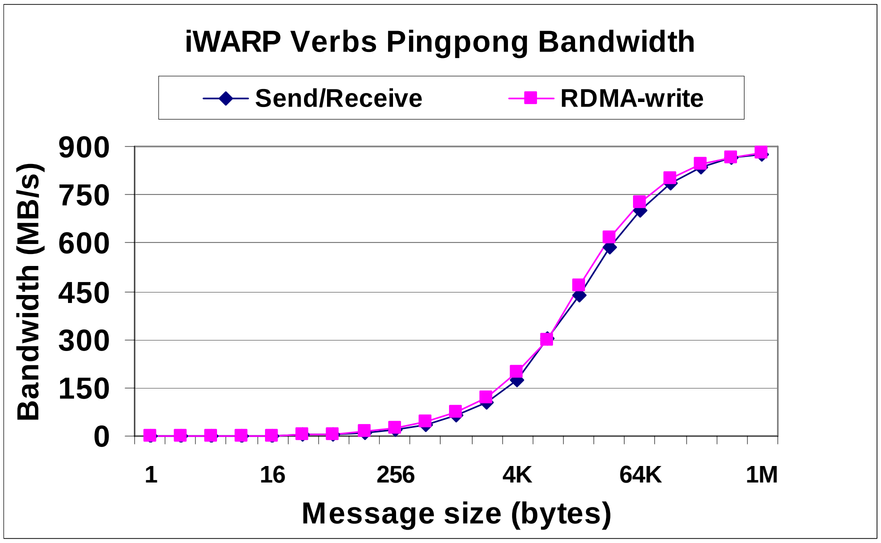
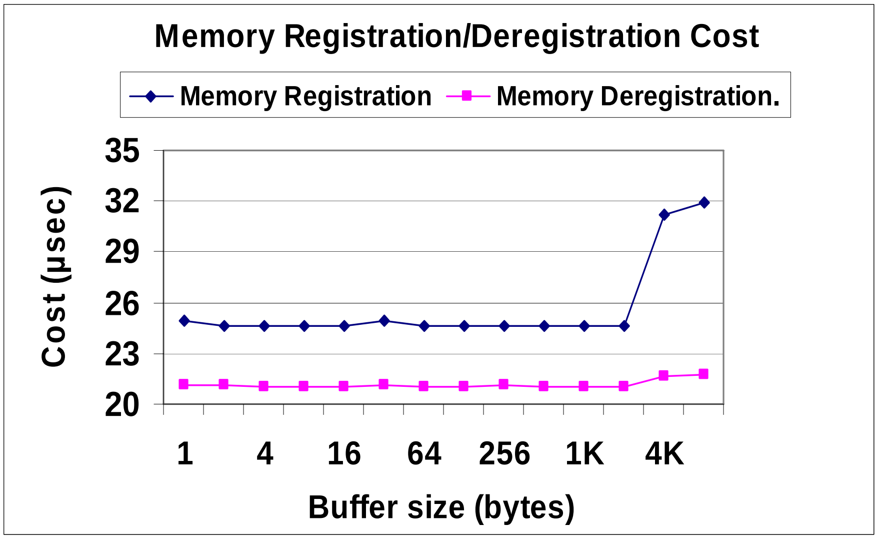


Figure 3. Verbs layer ping-pong bandwidth.

3.2. Memory Registration/Deregistration Cost

The iWARP standard requires message buffers to be registered (pinned-down) prior to being used for communication. This is to make sure the data pages remain in the main memory, and is accessible to the RNIC for the transfer. The results, shown in Figure 4, clearly indicate that memory registration is costlier than memory deregistration. Interestingly, memory deregistration cost is almost independent of the buffer size.



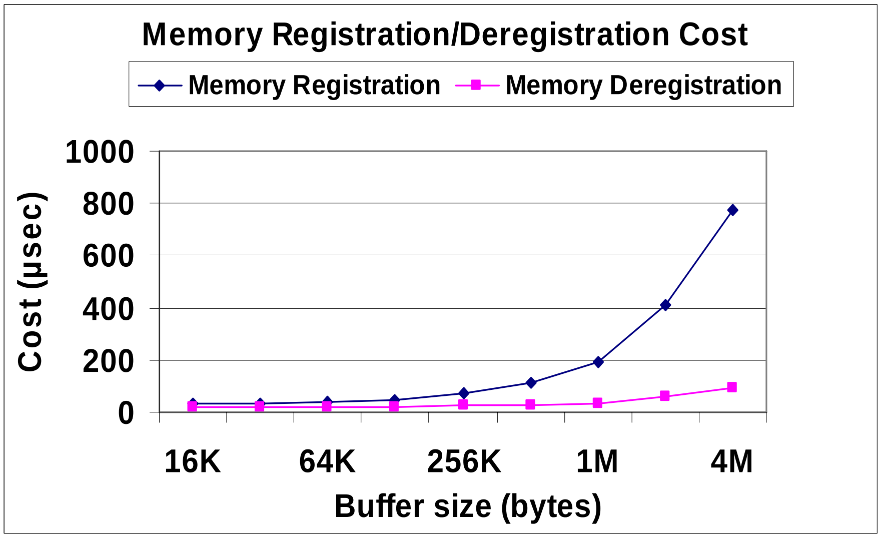


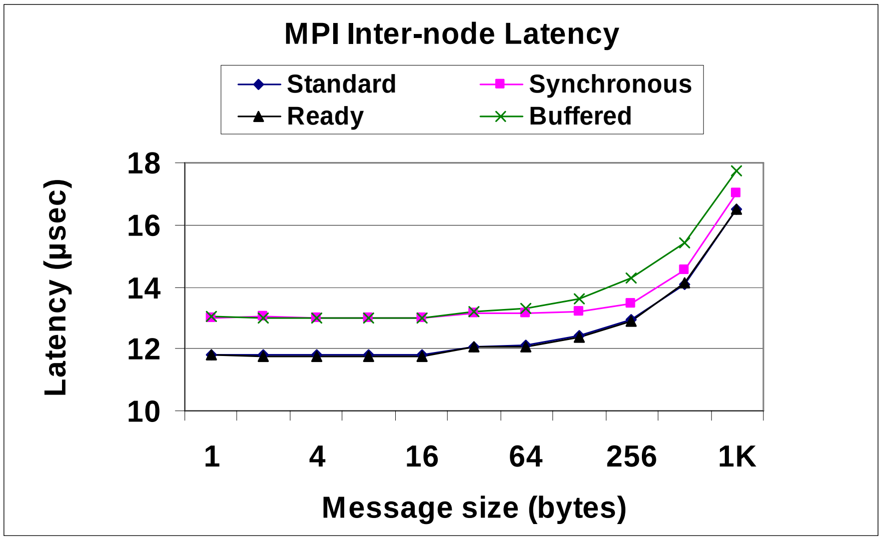
Figure 4. Memory registration/de-registration cost for small and large buffers.

4. MPI Performance Results

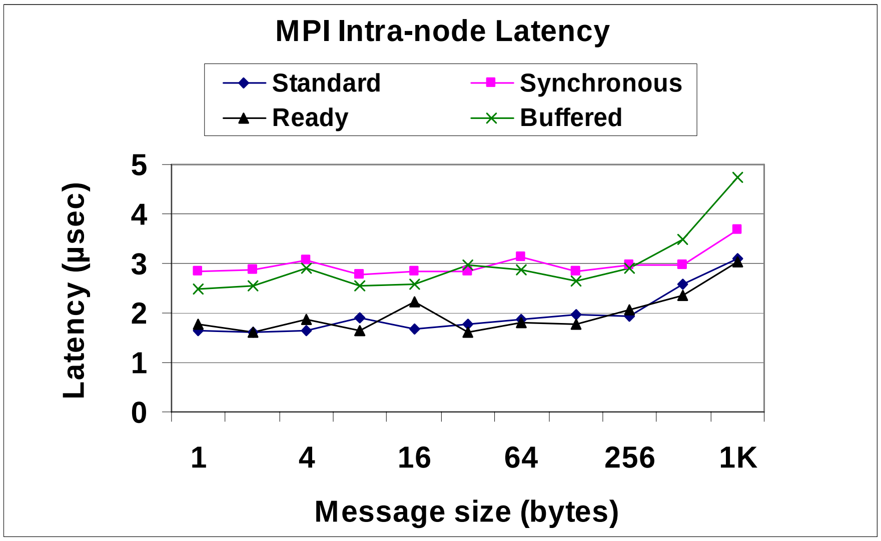
This section includes the results for several MPI micro-benchmarks to understand the performance and feasibility of the NetEffect iWARP RNIC for high-performance computing.

4.1. MPI Ping-Pong Latency

Figure 5 shows the ping-pong latency of different MPI communication modes for both inter-node and intra-node communications. The short message MPI latency (~11.5μs) is unprecedented for Ethernet networks, thanks to the RDMA over Ethernet, OS bypass, and TOE capabilities of the NetEffect iWARP RNIC.



(a) Inter-node ping-pong latency



(b) Intra-node ping-pong latency

Figure 5. MPI ping-pong latency.

4.2. MPI Overhead over Verbs

Figure 6 presents the overhead of the standard MPI latency over verbs RDMA write latency. The lower MPI intra-node latency indicates that the MPI implementation distinguishes between the inter-node and intra-node communications. In fact, for intra-node message passing, it uses local buffers instead of communication resources. Figure 6 also shows that for small messages, MPI imposes 20% overhead for inter-node communication. The overhead increases for 32KB messages but drops sharply for larger message sizes.

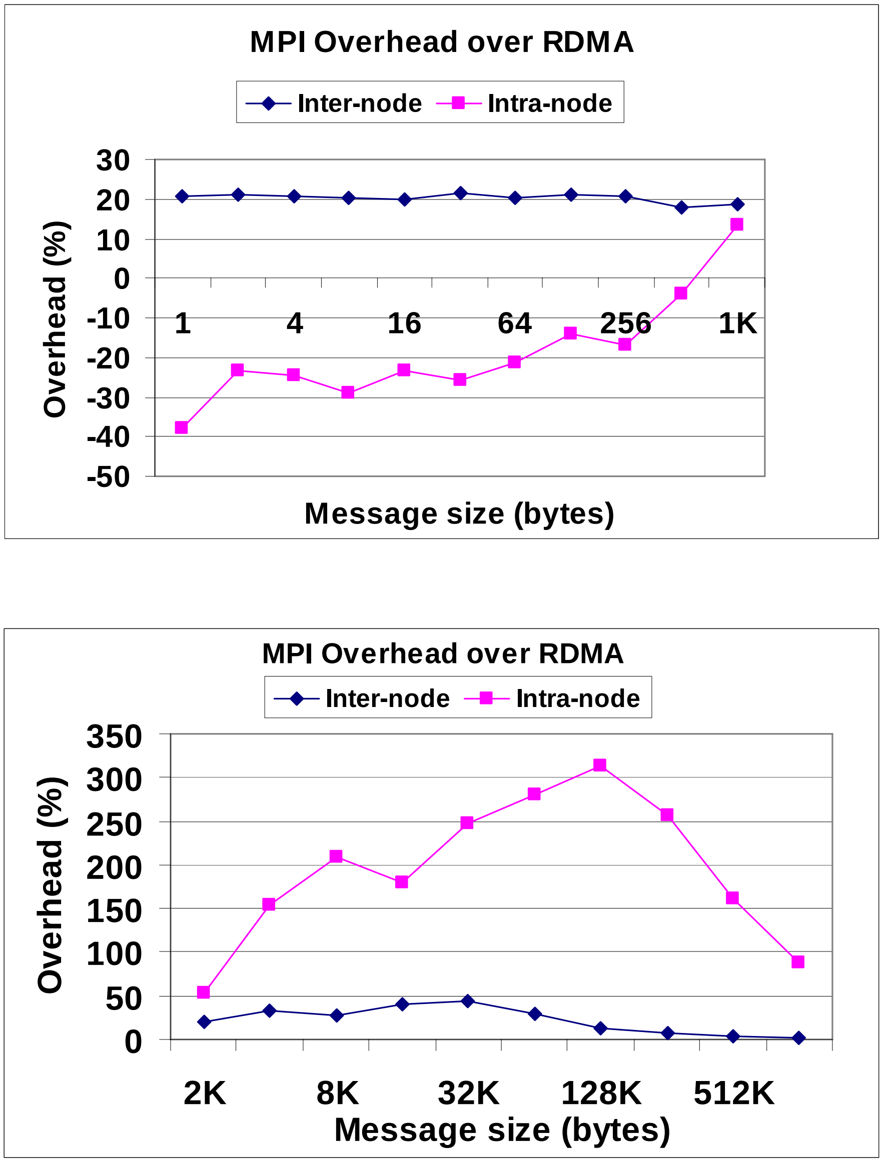


Figure 6. MPI latency overhead over verbs RDMA write.

4.3. MPI Communication Bandwidth

MPI inter-node bandwidths are presented in Figure 7. In the unidirectional case, the flow of messages is only from one side to the other side of the network. The bidirectional mode is in fact the ping-pong method. In the both-way benchmark, two communicating processes post windows of non-blocking send and receive calls and wait for their completion. This is done to saturate the network. The effective both-way bandwidth for 1MB messages is 925MB/s, which is 87% of the available bandwidth on the internal PCI-X bus. The bandwidth drop for the unidirectional case at 128KB is due to *Eager/Rendezvous* protocol switch. It is clear that the unidirectional communication can also well saturate the communication path for small and medium size messages. However, for large messages (when Rendezvous protocol is used), the one-way saturation is not possible and results are very close to the bidirectional case.

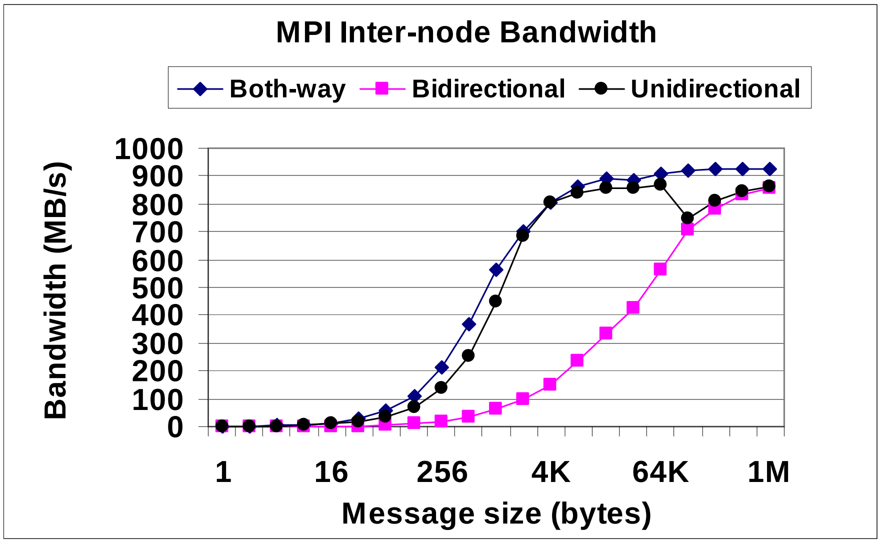


Figure 7. MPI inter-node bandwidth.

4.4. Aggregate Communication Bandwidth

The preceding bandwidth results in Section 4.3 are for the cases where there are only two communicating processes involved. Figure 8 shows the achieved aggregate bandwidth when multiple pairs are communicating at the same time across the switch. In this test, each pair performs the both-way bandwidth test. Then, the individual measured bandwidths are added up as the aggregate bandwidth. The top plot in Figure 8 shows that medium size messages can effectively achieve the sustained bandwidth at large messages over the RNIC local PCI-X bus, when multiple pairs are communicating on a pair of nodes. The aggregate achieved bandwidth when using all four nodes is close to 1.9GB/s.

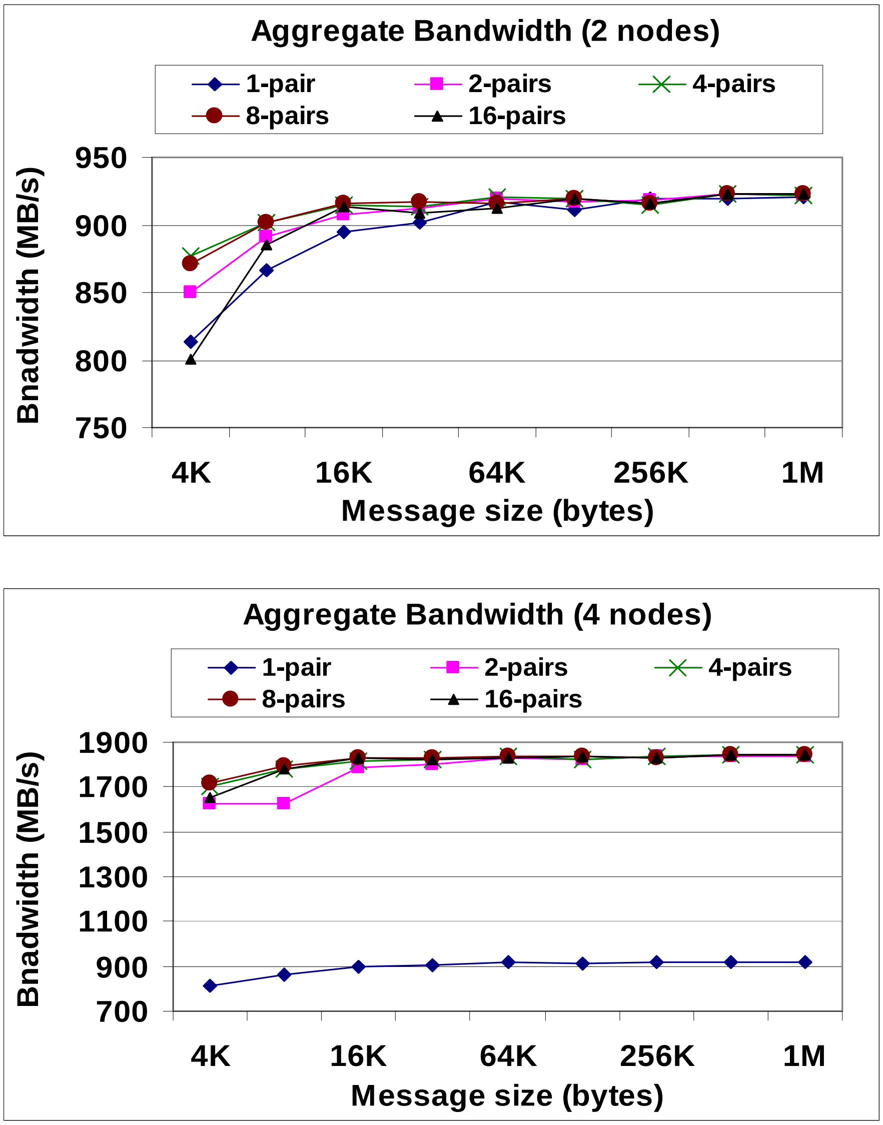
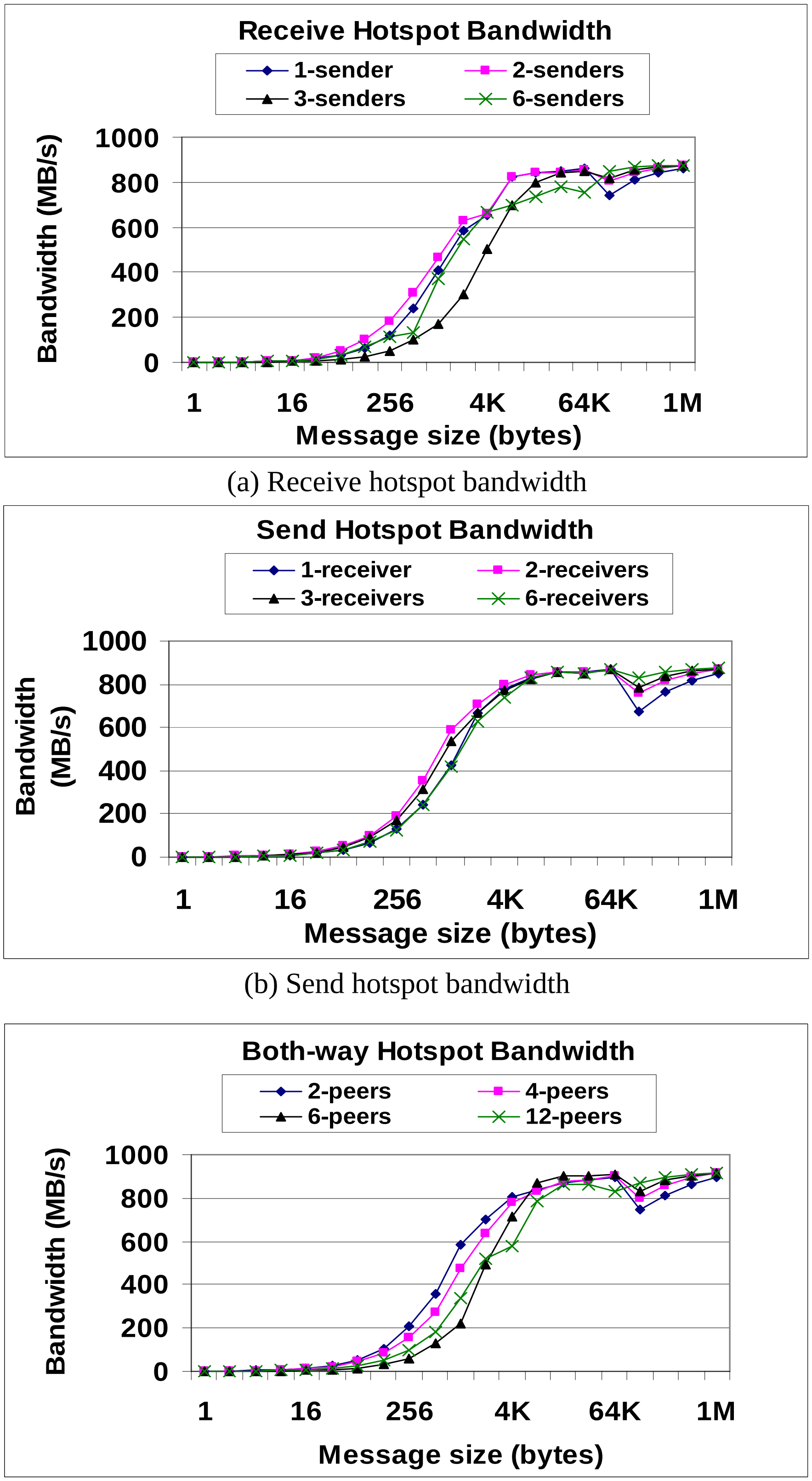


Figure 8. MPI Aggregate bandwidth.

4.5. MPI Hotspot Bandwidth

We have conducted these tests to assess the ability of a communicating process to keep up with more than one peer in communication. In the send and receive hotspot tests, a single hotspot process is either the source (send hotspot), or the sink (receive hotspot), of the messages to/from other processes. In the both-way case, the hotspot process is sending messages to a group of processes as well as receiving messages from another group. As shown in Figure 9, it is clear the hotspot process, in each case, is capable of coping with the communication pressure from its peers. The bandwidth is increasing, except with a drop at the protocol switching point.



(a) Receive hotspot bandwidth

(b) Send hotspot bandwidth

(c) Both-way hotspot bandwidth

Figure 9. MPI hotspot bandwidth.

4.6. Computation/Communication Overlap Ability

The ability of network to overlap useful computation with the pending communication is an important factor in achieving high performance. In Figure 10, we present the ability of the NetEffect iWARP RNIC and its MPI implementation in overlapping useful computation with a pending non-blocking send or receive operation.

We report the results in the percentage of the communication time that can be overlapped with computation without a significant increase. The results clearly show the excellent ability of the NetEffect RNIC in overlapping computation with the send operation. The receive operation overlap ability is also significant up to 64KB messages, but it drops sharply afterwards.

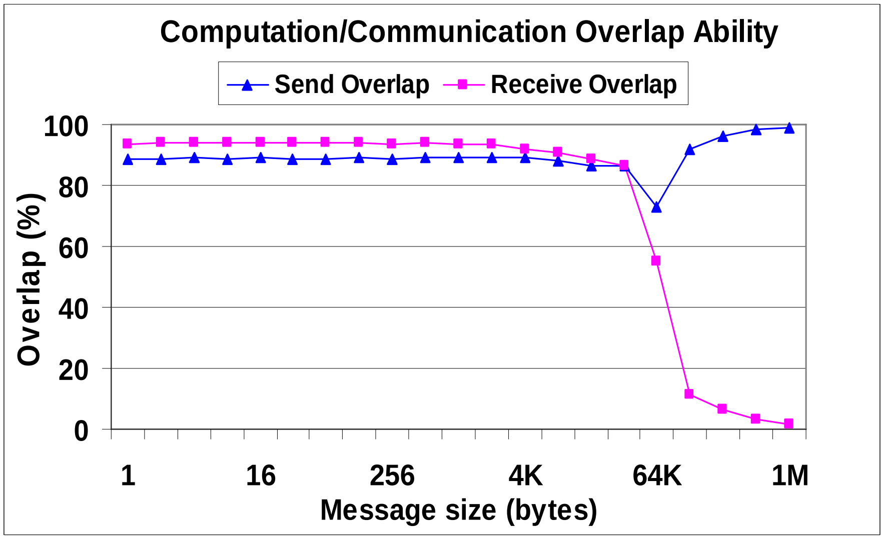


Figure 10. Ability to overlap computation with communication.

5. Conclusions

Recently, NetEffect Inc. has introduced an iWARP–enabled 10-Gigabit Ethernet Channel Adapter. We have experimented with the NetEffect RNIC at the verbs and MPI levels. In this report, we present the latency and bandwidth results as well as the cost of memory registration/deregistration at the verbs layer. MPI results are presented for basic latency and bandwidth, aggregate bandwidth, hotspot bandwidth, and the computation/communication overlap ability. The overall results show that the iWARP 10-Gigabit Ethernet will play a significant role in high-performance computing. In essence, the NetEffect RNIC achieves an unprecedented latency for Ethernet, reaches nearly 90% of available bandwidth, and is able to overlap up to 100% of the pending communication time with the computation.

Acknowledgments

The authors would like to thank the NetEffect for their tremendous support in providing the NetEffect iWARP RNIC, switch, and the necessary resources to conduct this experiment. We are thankful to Brian Hausauer, Gary Montry, Vadim Makhervaks, Terry Hulett, and Allan Torres of the NetEffect for their insightful comments during the development of this work. This research is also supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC), Canada Foundation for Innovation (CFI), Ontario Innovation Trust (OIT), and Queen’s University.