

10-Gigabit iWARP Ethernet: Comparative Performance Analysis with InfiniBand and Myrinet-10G

Mohammad J. Rashti and Ahmad Afsahi

Department of Electrical and Computer Engineering

Queen's University

Kingston, ON, Canada

2007 Workshop on Communication Architectures for Clusters

March 26, 2007



- Introduction
- Overview of iWARP Ethernet
- Experimental Platform
- Performance Results
- Conclusions and Future Work

- More than 72% of the top500 computers in Nov. 2006 ranking are **clusters**.
- Clusters have become the predominant computing platforms providing high-performance mostly due to availability of :
 - Fast computational engines
 - High-speed interconnects
- **High-performance clusters** are extremely desirable to tackle challenging and emerging applications.
 - A new era in parallel processing is shaping with the emergence of multi-core SMP/NUMA nodes.
 - It is believed that research will be more focused at optimizing applications, parallel programming models, as well as **improving the communication subsystems**.

- High-performance clusters need high-performance networks and efficient communication system software. Such networks include:
 - Myrinet
 - Quadrics
 - InfiniBand
- Contemporary networks offer extremely **low latency** and **high bandwidth** using:
 - OS bypass
 - RDMA → true zero-copy data transfer and low host CPU overhead
 - Specialty protocols

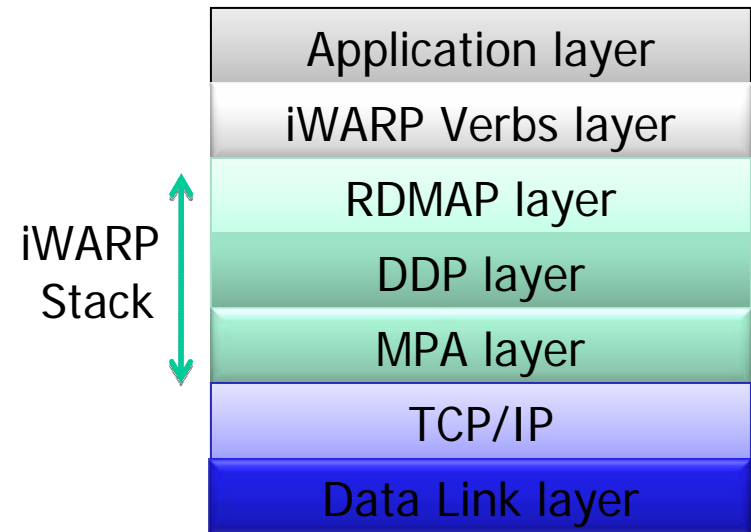
→ However, their main drawback has been their incompatibility with existing **Ethernet infrastructure** (42% of the top500 use Ethernet).

- There are currently two trends in high-performance networking community to bridge the performance, cost and compatibility gap between Ethernet and **Ethernor** networks:
 - Modern networks (such as Myrinet 10-G) have opted to support Ethernet by porting their messaging software over Ethernet.
 - Emerging high-performance 10-Gigabit Ethernet networks have aggressively pushed towards using **OS bypass** and **RDMA over Ethernet** in addition to **TOEs**. This has been facilitated with the standardization of the **iWARP Ethernet** by RDMA consortium.

It is important to the research community to have a systematic assessment of such 10-Gigabit iWARP Ethernet networks for high-performance computing in comparison to leading cluster interconnects such as Myrinet and InfiniBand.

- Introduction
- Overview of iWARP Ethernet
- Experimental Platform
- Performance Results
- Conclusions

- **iWARP**: is a set of standard extensions to TCP/IP and Ethernet.
 - **Verbs** layer is the user-level interface to the RDMA-enabled NIC.
 - ❖ Similar to IB, iWARP uses QPs, work descriptors etc.
 - **RDMAP** layer is responsible for RDMA operations (and Send/Recv), joint buffer management with DDP, bookkeeping, and error reporting to Verbs layer.
 - **DDP** layer is used for direct zero-copy data placement (tagged and untagged), as well as segmentation and reassembly.
 - **MPA** layer assigns boundaries to DDP messages, and takes care of CRC.



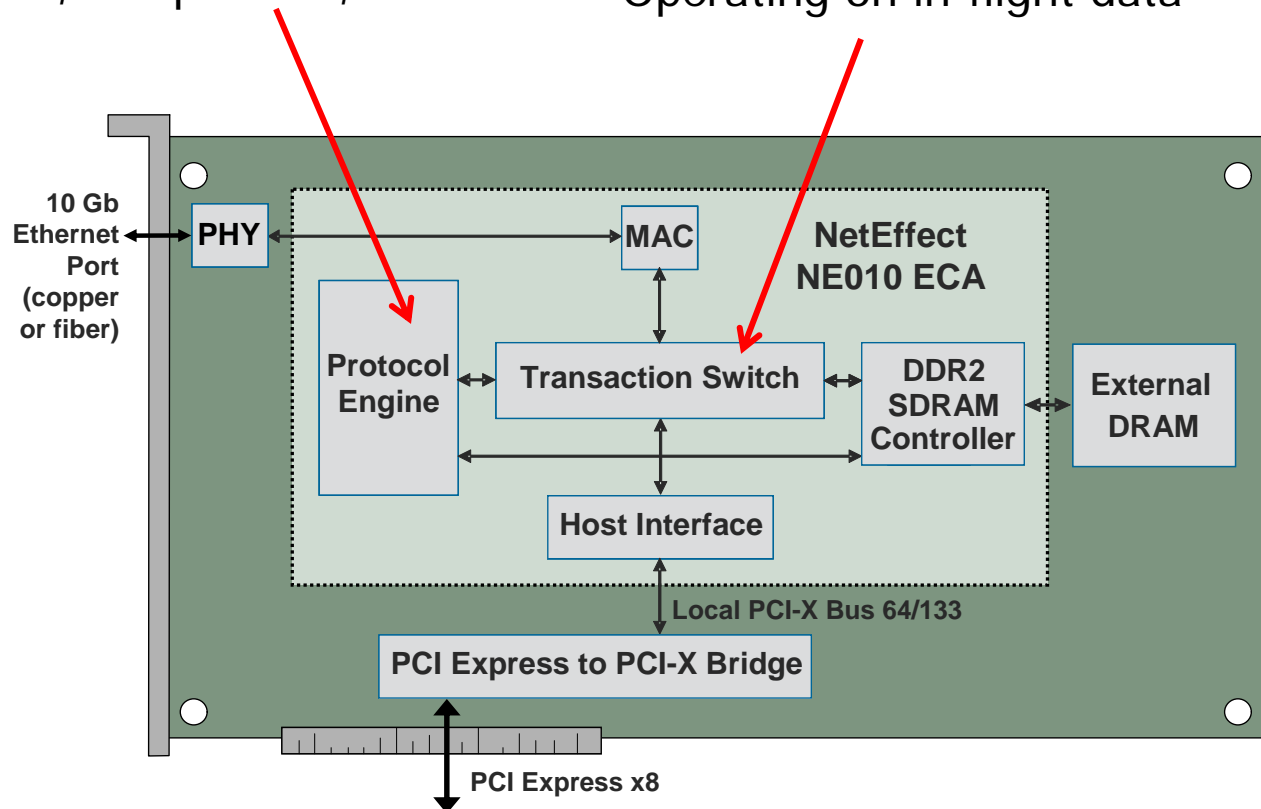
NetEffect 10-Gigabit iWARP Ethernet

IEEE CAC-2007

- NetEffect has recently introduced a 10-Gigabit Ethernet RNIC.

Integrating iWARP, IPv4 TOE and NIC accelerator in hardware, also responsible for memory registration, completions, errors

Operating on in-flight data



Software:
NetEffect verbs
OpenFabrics
verbs, sockets,
SDP, uDAPL
and MPI

NetEffect NE010e Ethernet Channel Adapter architecture

- Introduction
- Overview of iWARP Ethernet
- **Experimental Platform**
- Performance Results
- Conclusions

Experimental Platform

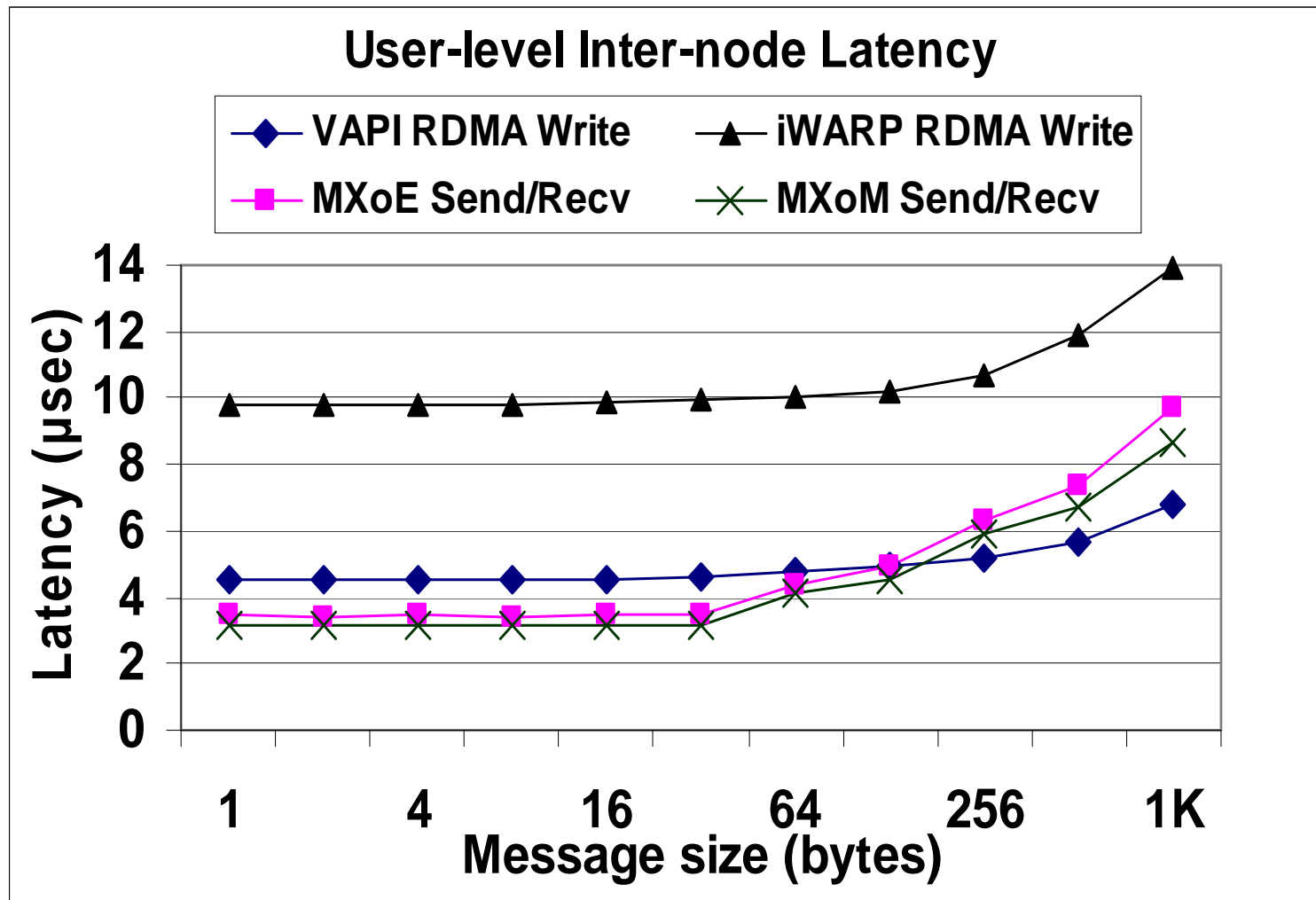
IEEE CAC-2007

	Specification
Nodes	Dell PowerEdge 2850s, each a dual-processor 2.8GHz Intel Xeon SMP with 1MB L2 cache/processor, 2GB memory and an x8 PCIe slot
Myrinet	Myricom single-port Myri-10G NICs with CX4 and PCIe x8 interface (were forced to work on x4 mode for effective performance). Myri-10G 16-port switch for MXoM and 12-port Fujitsu XG700-CX4 switch for MXOE. MPICH-MX based on MPICH 1.2.7..1.
InfiniBand	Mellanox dual-port 10GB/s (memfree) HCA cards with a PCIe x8 interface and a Mellanox 12-port 4X MTS2400-12T4 InfiniBand switch. MVAPICH2 based on MPICH2 1.0.3 over VAPI version 0.9.5. OpenFabrics/Gen2
iWARP Ethernet	NetEffect single-port NE010e 10-Gigabit ECAs with CX4 and PCIe x8 interface (local PCI-X 64/133MHz) and 12-port Fujitsu XG700-CX4 switch. MPICH2-iWARP based on MPICH2 version 1.0.3. NetEffect Verbs and OpenFabrics/Gen2
Kernel	Fedora Core 4 SMP for IA32, kernel version 2.6.11, and Fedora Core 5 SMP for x86-64, kernel version 2.6.17.7

- Introduction
- Overview of iWARP Ethernet
- Experimental Platform
- Performance Results
- Conclusions

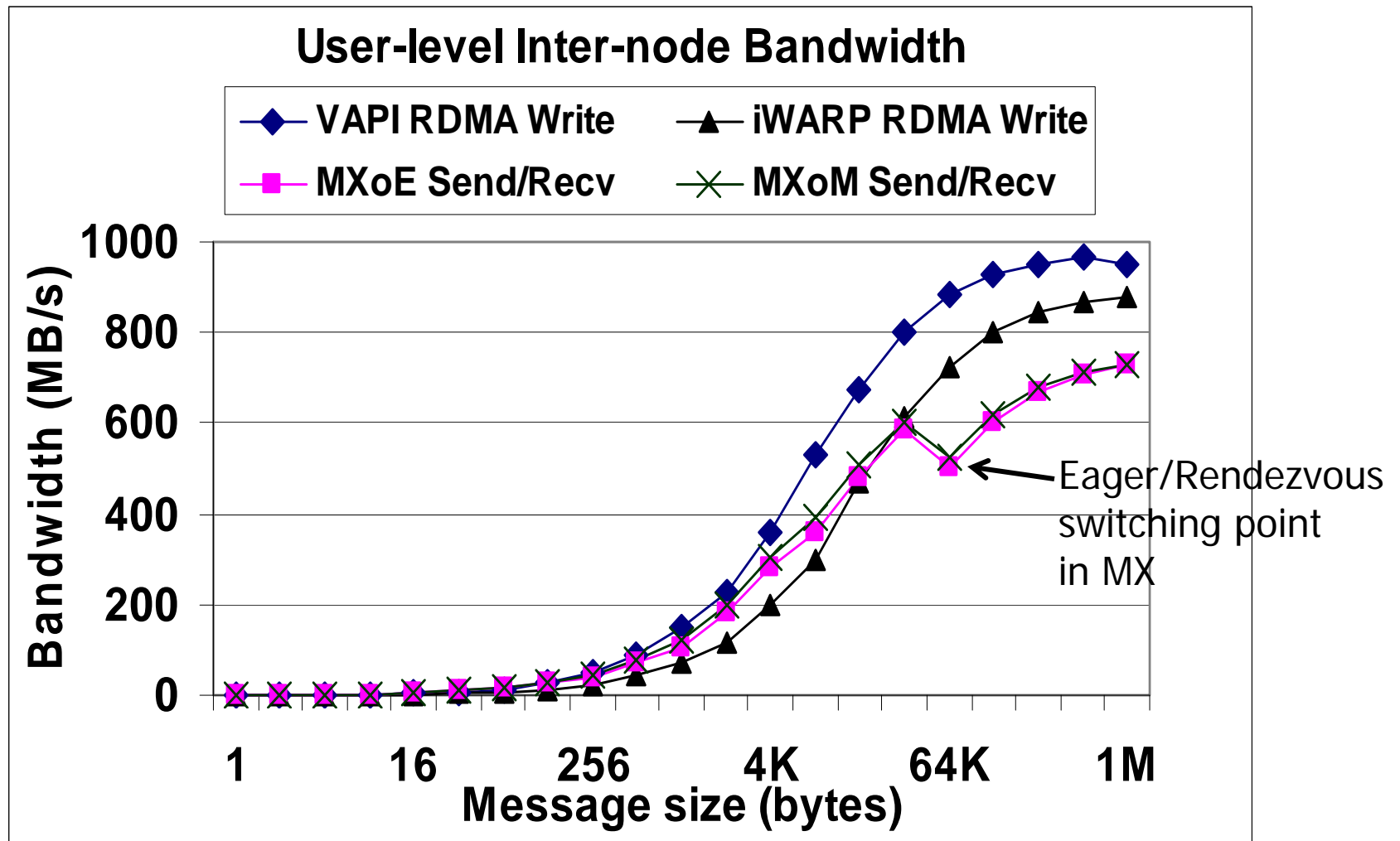
User-level Ping-Pong Performance

- Latency (using the main communication model)



User-level Ping-Pong Performance

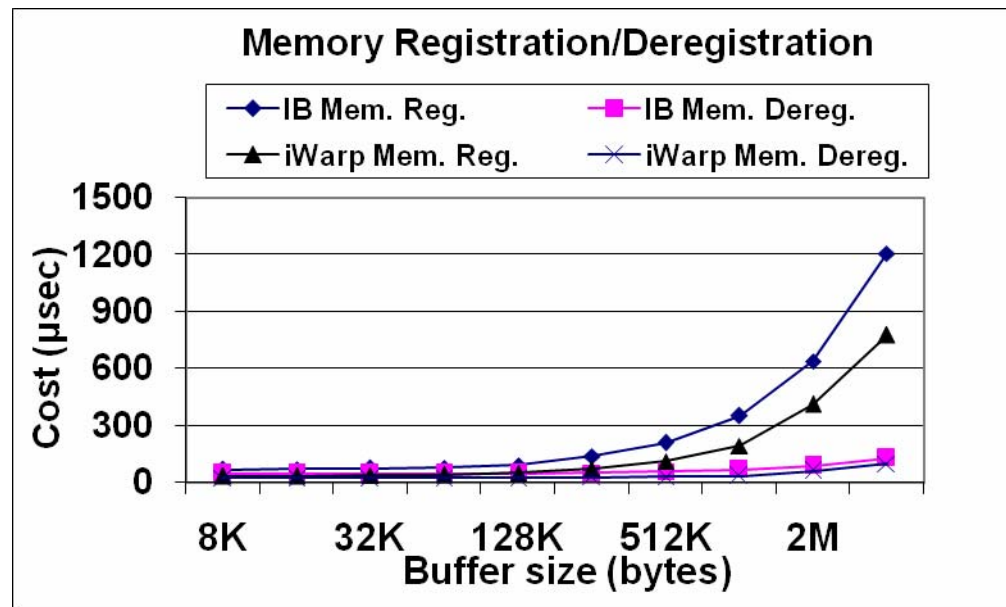
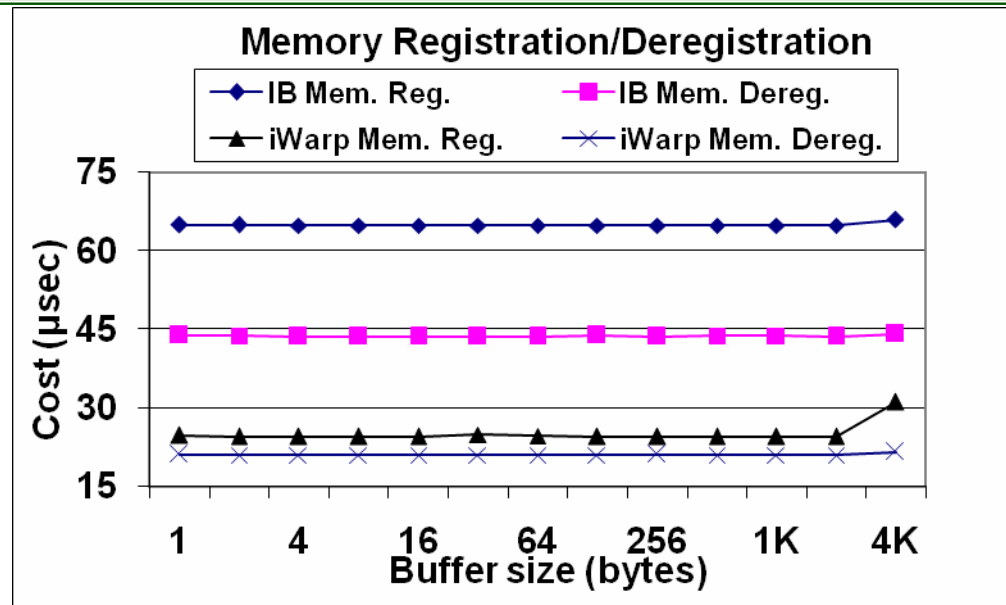
- Bandwidth (one-way)



Memory Registration/deregistration

- **iWARP vs. IB**

- MX does not have explicit memory registration.



• Why Multiple Connection Performance Matters?

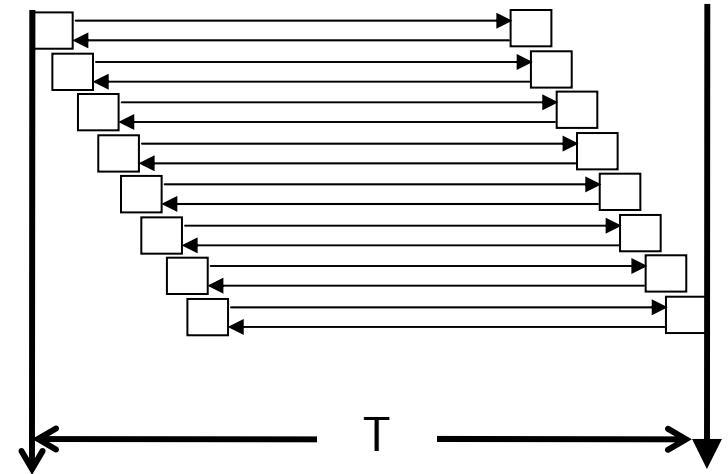
- With the emerging multi-core multiprocessor servers, there is now immense pressure on the NIC hardware and its communication system software to provide scalable performance with the increasing number of connections.
- Common off-the-shelf servers are 2-way or 4-way SMP nodes. Each processor is currently a dual-core or quad-core (soon it will be an 8- or 16-core). MPI runs typically a process on every core. Therefore, each core will have connections to several other processes.

Number of Servers	Processors per server	Cores per processor	Connections per NIC
4	2	2	48
8	2	2	112
64	2	4	4032
256	2	4	16320
1024	2	4	65472

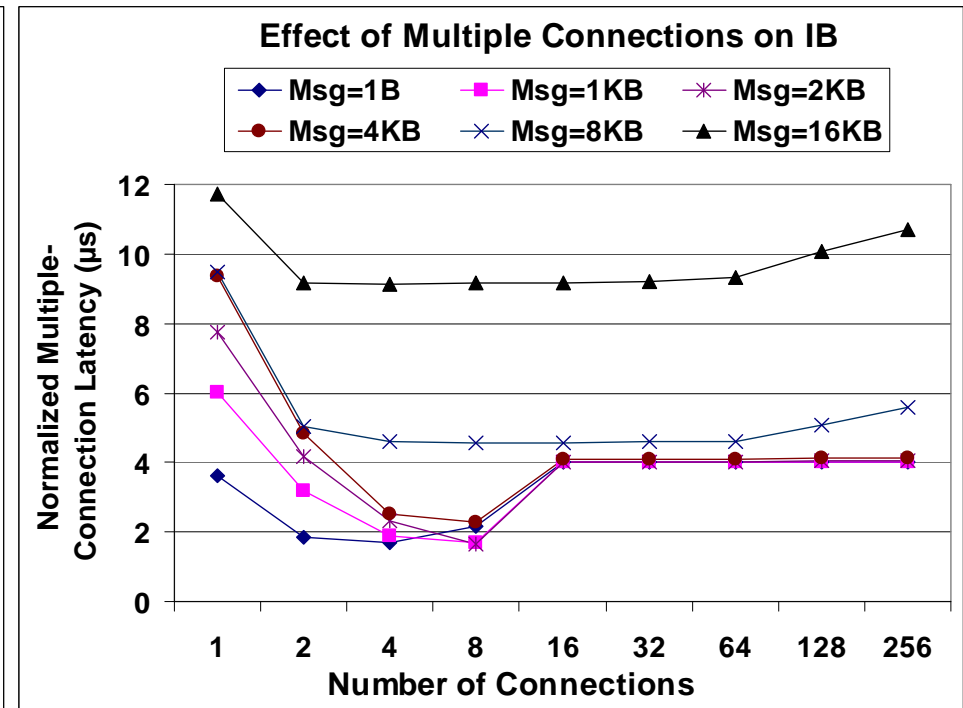
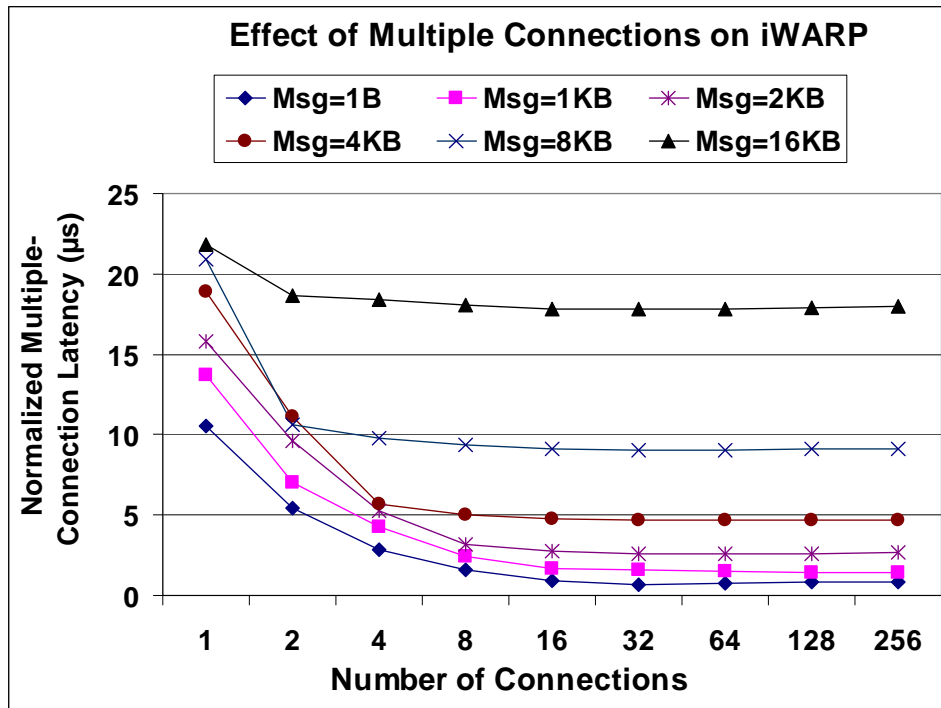
Multiple Connection Scalability

IEEE CAC-2007

- Both the iWARP and InfiniBand standards have a QP communication model and both support connection-oriented RDMA operations.
 - We use OpenFabrics/Gen2 as a common user-level interface to compare the hardware implementations of these standards in terms of multiple connection scalability.
- Pre-establish up to 256 connections between two processes on two nodes.
- Perform a ping-pong test using all of the connections in parallel.
- A fixed number of messages are sent and received over the connections in a round-robin fashion.
- We define the cumulative half way round trip time, T , divided by the number of messages and connections as the **normalized multiple-connection latency**. This shows how well communication over multiple connections can be performed in parallel.

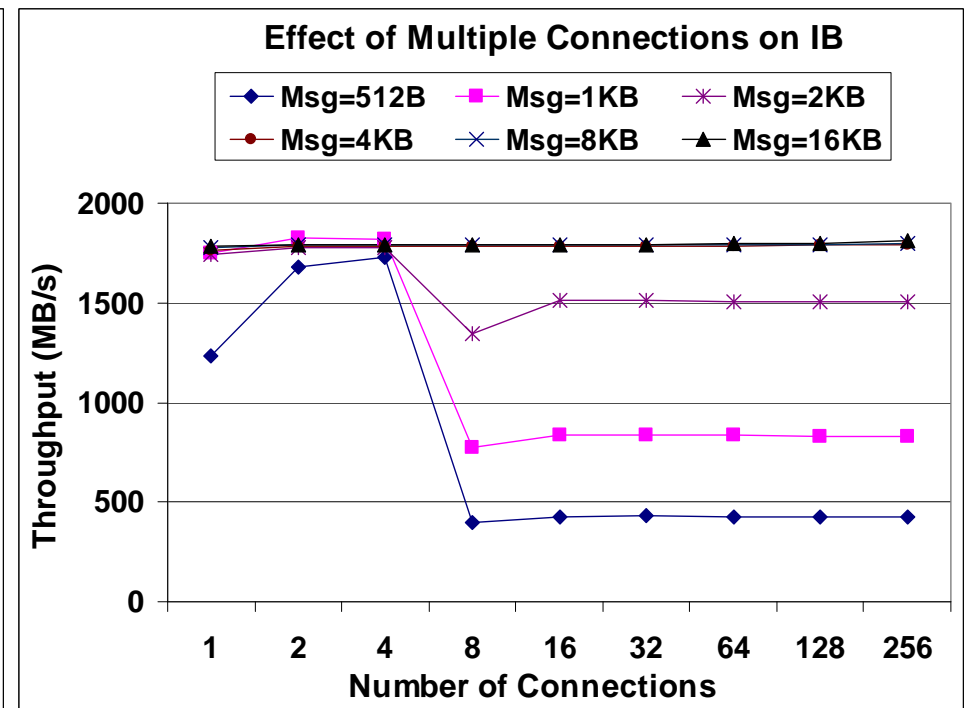
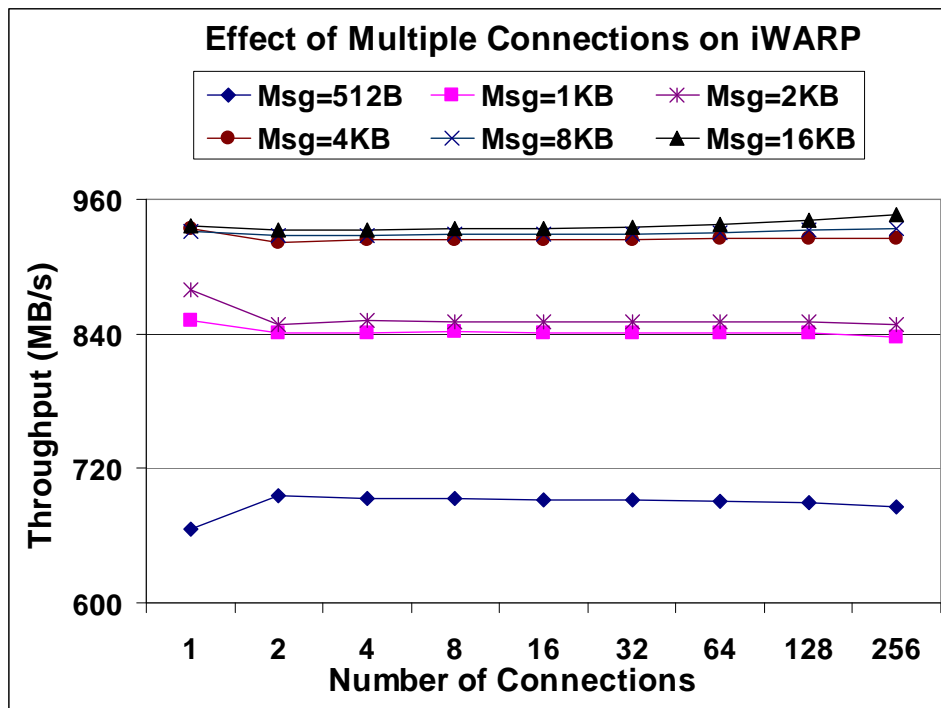


- Normalized multiple-connection latency of NetEffect iWARP and Mellanox InfiniBand

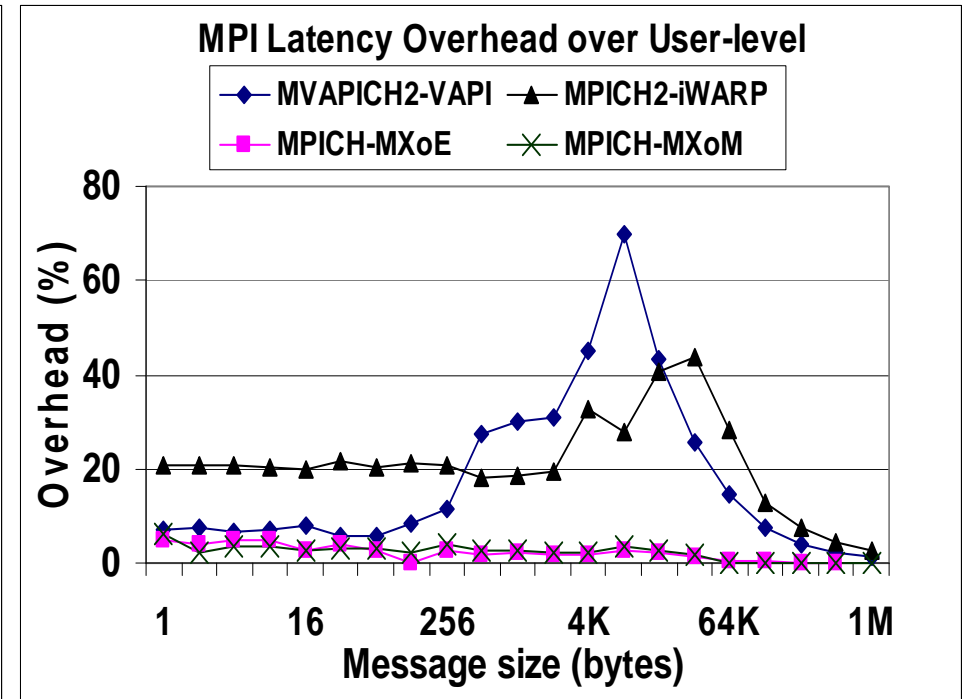
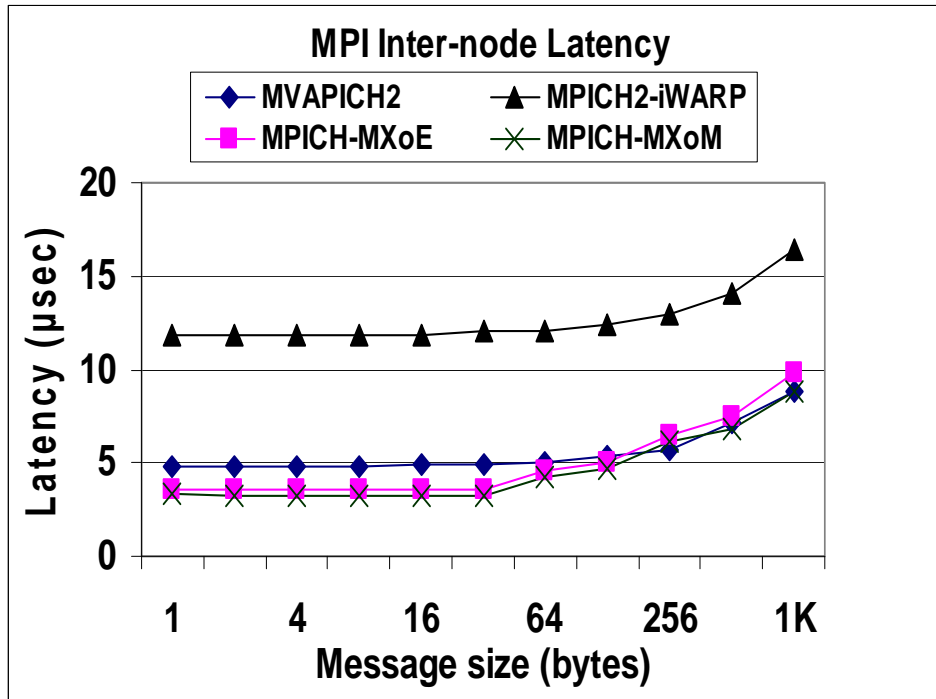


- For the IB card, the normalized multiple-connection latency of messages smaller than 4KB (up to 8 connections) decreases, but after that the latency increases. We speculate this might be due to processor-based communication in IB NIC core (iWARP NIC uses a state-based approach).

- Throughput of NetEffect iWARP and Mellanox InfiniBand
 - We do a both-way communication test for a certain amount of time, where each process sends messages to its peer in a round-robin fashion over the established connections.

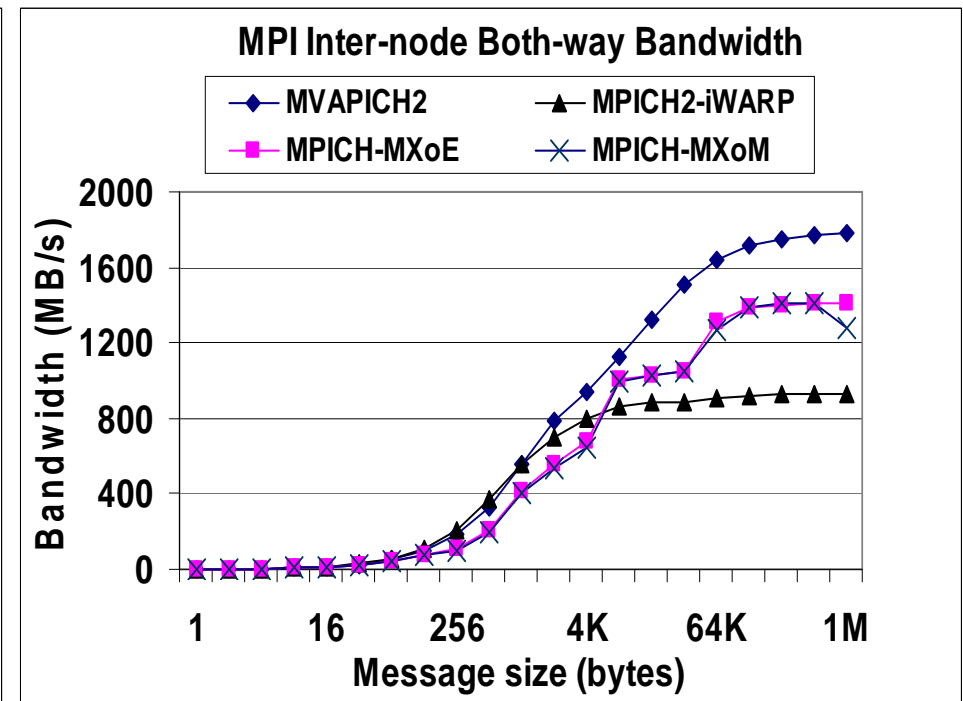
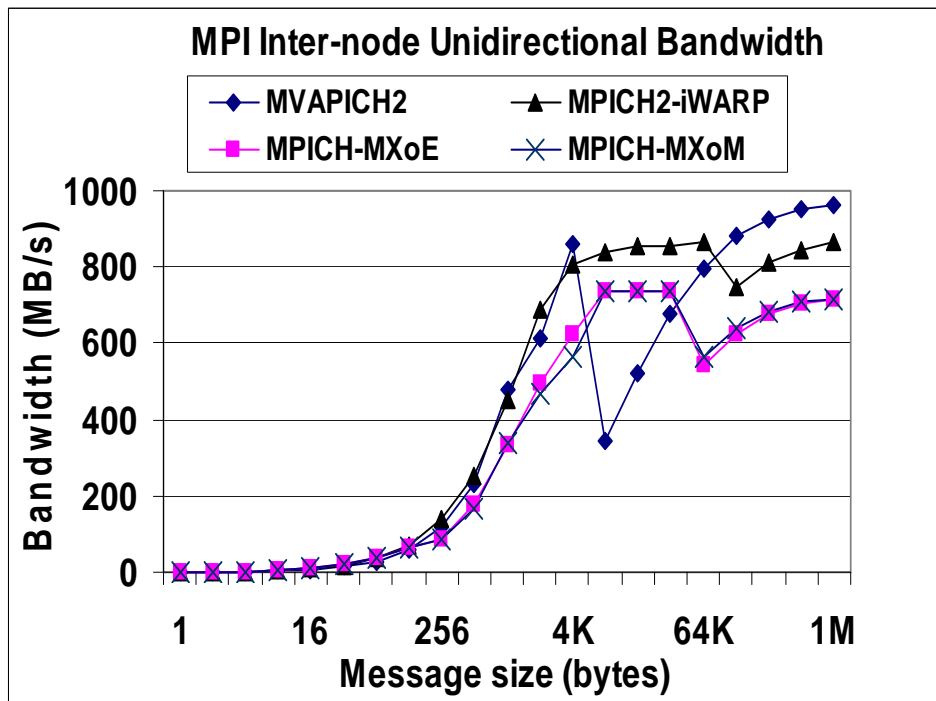


- MPI ping-pong latency and its overhead over user-level



Note MX-10G library has semantics close to MPI.

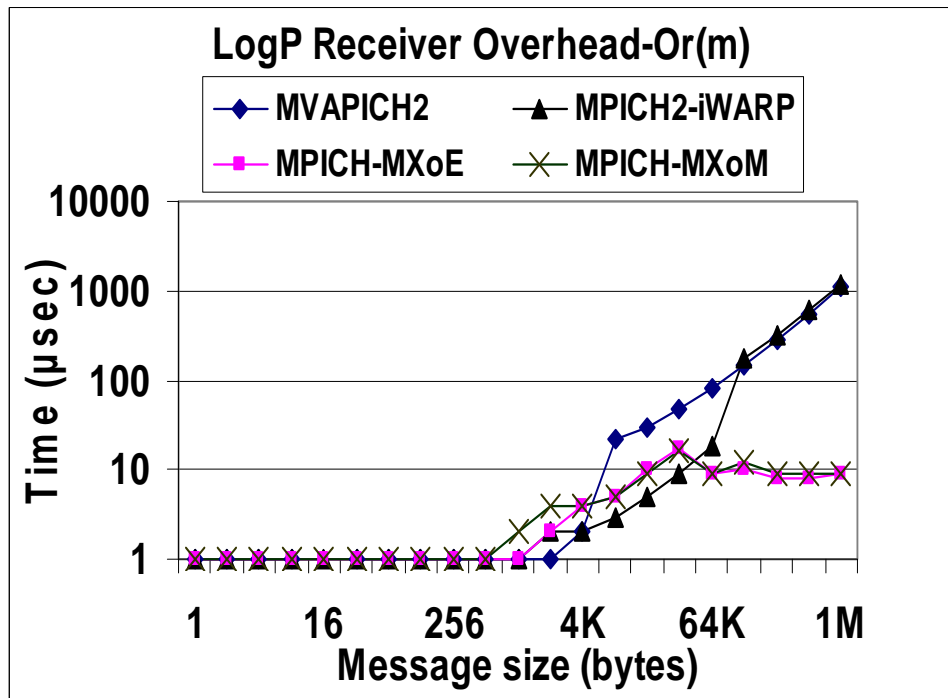
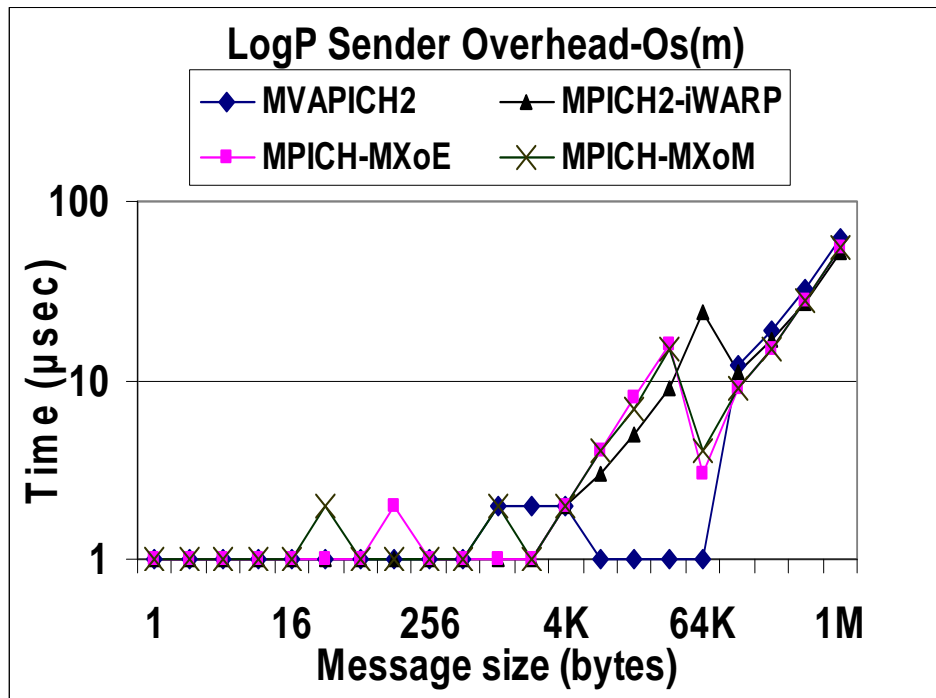
- In the **unidirectional** test, the sender repeatedly transmits windows of non-blocking messages to the receiver, waits for each window to be completed and then for the last message to be acknowledged.



- In the **both-way** test, both the sender and receiver post a window of non-blocking send operations, followed by a window of non-blocking receive calls.

Parameterized LogP Parameters

- The gap value is roughly the same and grows steadily with the message size for all messages.



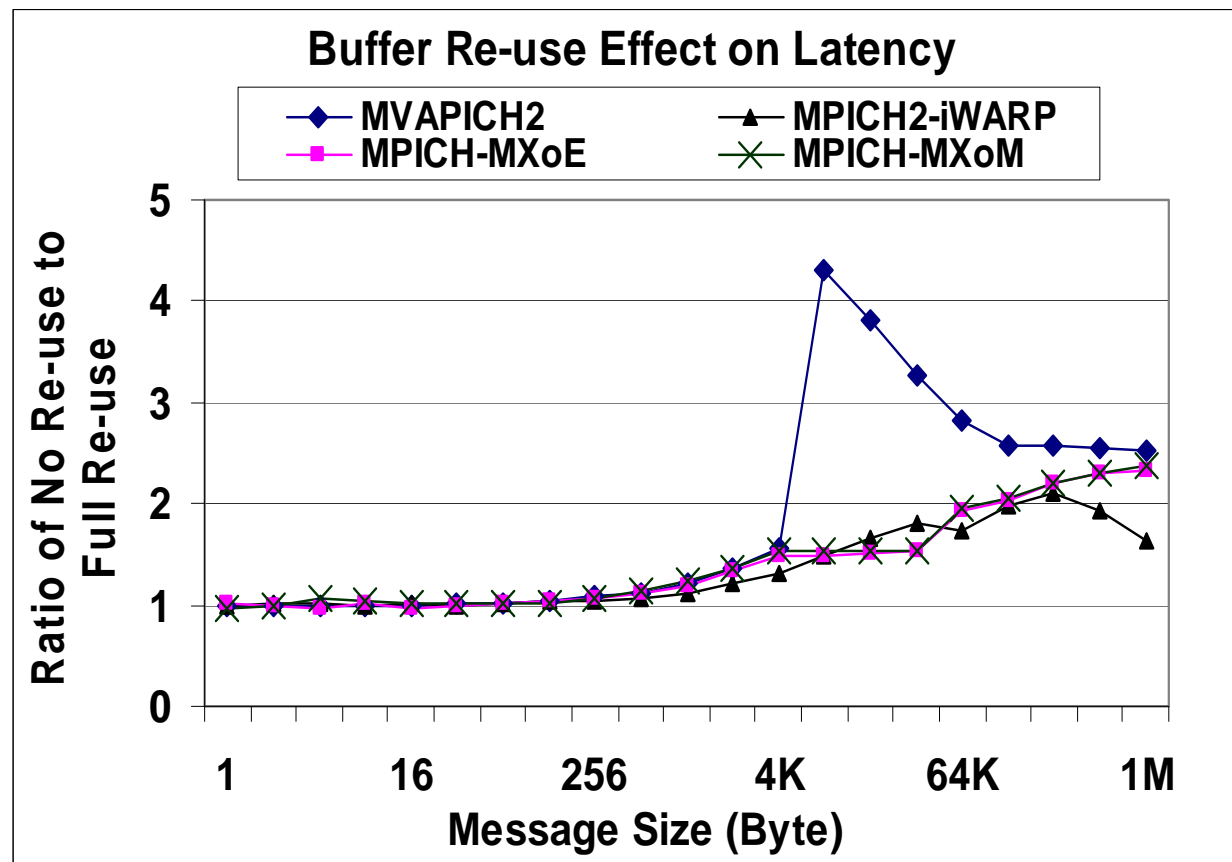
- Large receiver overhead for the Rendezvous-type messages for the iWARP and IB networks shows that mostly the receiving process is involved in the data transfer. Myrinet has a progression thread that is awakened for starting large message transfers.

Effect of Message Buffer Re-use

- Pinning/unpinning is expensive and an application buffer re-use pattern may have a significant impact on performance.
 - We statically allocated 1024 separate buffers. Full re-use (100% re-use) always uses the same buffer.

For Eager size messages, buffer re-use impact ratio is 1.8 for iWARP, 1.55 for IB and 1.53 for Myrinet.

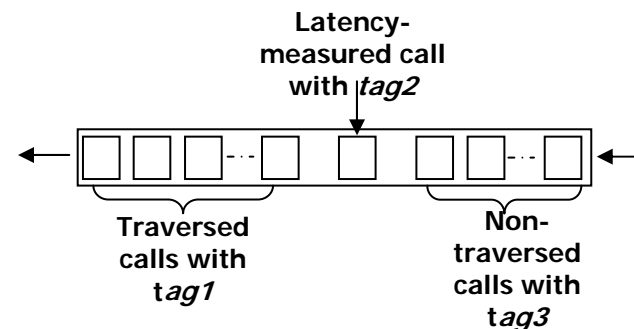
For Rendezvous size messages, it is up to 4.3 for IB, 2.1 for iWARP and 2,4 for Myrinet



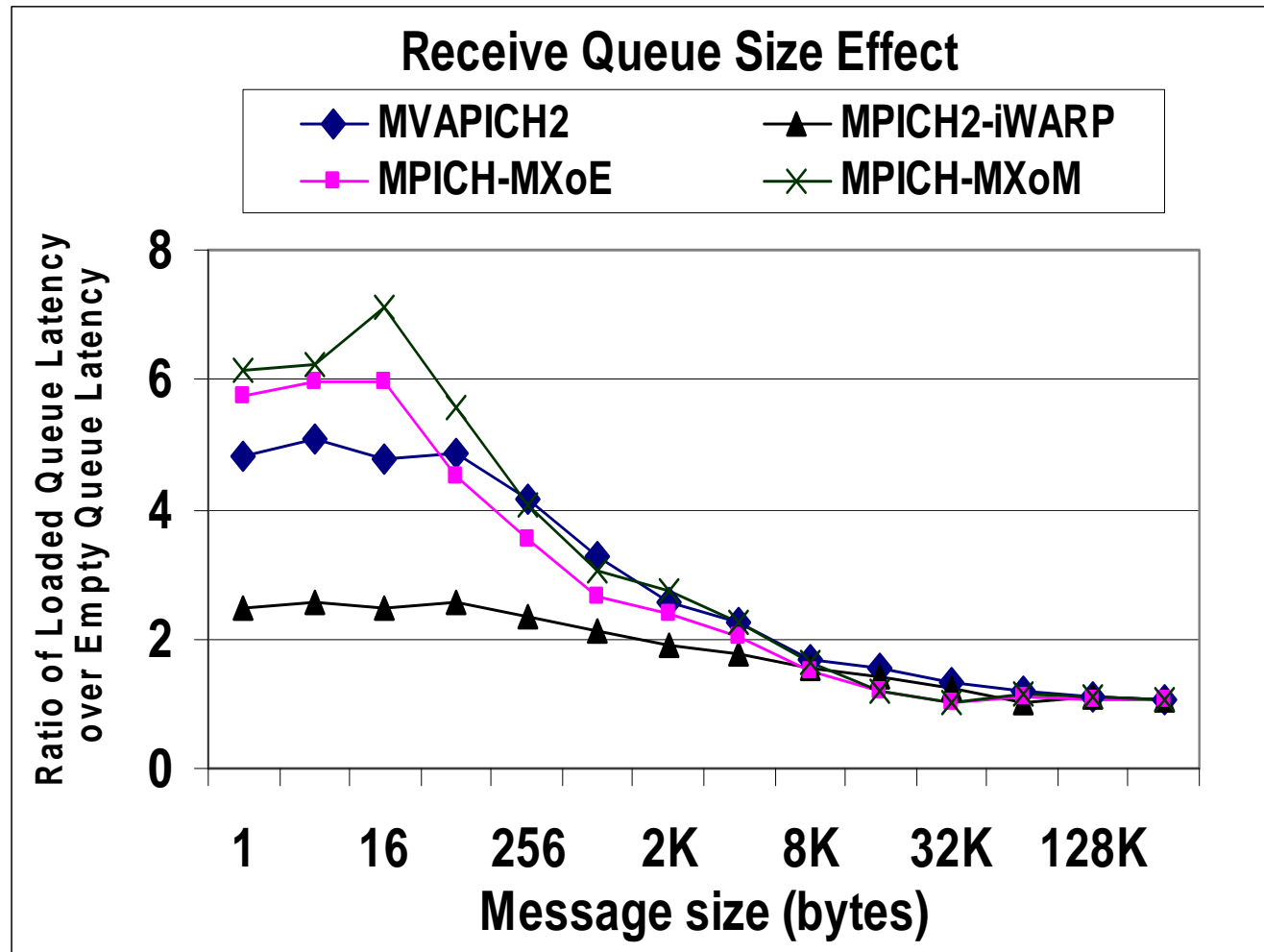
- **Unexpected Message Queue:** is used to save unexpected messages temporarily. When a receive call is posted, the unexpected message queue is traversed for a matching message.
 - Each process sends a certain number of small, unexpected messages to the other side. Then, the processes synchronize and start communicating in a ping-pong fashion. 1000 unexpected messages are posted.



- **Receive Call Queue:** is used to keep early posted receive calls. It is traversed upon reception of a message from the network to find a matching receive call.
 - Both sides pre-post a certain number of non-blocking receive calls with a certain tag (*tag1*). These calls sit at the beginning of the receive call queue, and are called traversed calls.
 - Then, the actual latency-measured non-blocking receive call is posted with a different tag (*tag2*). This call sits in the queue after the traversed calls.
 - At this time, both sides synchronize and start communicating. One side sends a message with tag2 and waits for a similar message in response.
 - Upon reception of a message at either side, the queue is traversed to find the matching receive with tag2.



- Receive Call Queue:



- Introduction
- Overview of iWARP Ethernet
- Experimental Platform
- Performance Results
- Conclusions

- TCP Offload Engines and RDMA help to achieve low host CPU utilization in Ethernet networks.
- We compared Myri-10G, Mellanox InfiniBand and NetEffect iWARP in terms of both single and multiple-connection user-level latency, MPI basic latency and bandwidth, LogP parameters, buffer re-use and queue usage.
- Myrinet is the best in latency, while InfiniBand is the best in bandwidth.
- For the iWARP:
 - Significant improvement in Ethernet latency is observed.
 - Better scalability for multiple-connection latency is noted.
 - A higher level of performance in MPI queue usage and buffer re-use is delivered than those of InfiniBand.
 - Could be a key player in future as technology matures.

- We would like to extend our study to computation/communication overlap and independent progress.
- We would like to put the networks to the test in a larger testbed.
- Extend our performance study to uDAPL, SDP and applications.
- Intend to enhance the MPI over iWARP RNIC.

- This work was supported by:
 - Natural Sciences and Engineering Research Council of Canada (NSERC)
 - Canada Foundation for Innovation (CFI)
 - Ontario Innovation Trust (OIT)
 - Queen's University

 - NetEffect, Inc.
 - Myricom, Inc.

