



Micro-Benchmark Level Performance Comparison of High- Speed Cluster Interconnects





J. Liu, B. Chandrasekaran, W. Yu, J. Wu,
D. Buntinas, S. Kini, P. Wyckoff[†] and D.K. Panda

Dept. of Computer and Info. Science Ohio Supercomputer Center [†]
The Ohio State University Columbus, OH 43212



Presentation Outline

- Trends in High-Speed Cluster Interconnects
 - Motivation behind the Micro-Benchmarks
 - Micro-Benchmarks and Results
 - Conclusions
- 
- 

Trends in High-Speed Cluster Interconnects

- Low latency (less than 10 μ s)
- High bandwidth (several GigaBytes per sec)
- Leading products
 - Myrinet
 - Quadrics
- A newcomer
 - InfiniBand

Myrinet

- High-speed interconnect based on technology in MPP
- M3F-PCIXD-2 network interface cards
 - PCI-X 64bit/133MHz interface
 - 2 Gbps link bandwidth
 - 225 MHz Lanai-XP processor
 - GM-2.0.1 software
- Myrinet 2000 switch

Quadrics

- Elan3 QM400 network interface cards
 - PCI 64bit/66MHz interface
 - MMU and 64 MB on-board SDRAM
 - 400 MB link bandwidth
 - libelan-1.4.3 software
- Elite-16 switch

InfiniBand

- A new interconnect that connect I/O nodes and processing nodes
- Mellanox InfiniHost 4X HCAs
 - PCI-X 64bit/133MHz interface
 - 10 Gbps link bandwidth
 - VAPI (software interface)
 - Reliable Connection (RC) service
- Mellanox InfiniScale switch

Common Characteristics of these Interconnects

- User-level access to network interfaces
- Remote DMA (RDMA) in addition to the standard Send/Receive
- Sophisticated network adapters to offload communication tasks from host CPU

Protocol Layers in Clusters

HPC
(MPI)

File Systems
(PVFS)

Data Centers
(Apache/Database)

User-Level Communication
Protocols

GM (Myrinet)
VAPI (InfiniBand)
ElanLib (Quadrics)

Cluster Interconnects
(Adapters + Switches)

How to Choose an Interconnect?

- Upper layers will have different communication characteristics
- Standard latency and bandwidth tests with
 - A single buffer at the sender and a single buffer at the receiver
 - A single connection
 - Polling for completion
- Results may be ideal
- Could be misleading

Our Objective

- Carryout a meaningful performance comparison among the three cluster-interconnects
- Taking into account of
 - communication characteristics of upper layers
 - common characteristics of lower-layers
 - User-mode network access
 - RDMA
- Explore different aspects of communication
 - Not just measure simple cases
- Helps
 - Upper layer designers to know about the strengths/limitations of the lower layers
 - Lower layer designers to optimize the implementation for a given upper layer

Methodology

- A suite of micro-benchmarks
 - Implemented with minimum software overhead
 - Characterize different aspects of communication performance
 - Traditional measurements
 - Latency, Bandwidth, ...
 - Other micro-benchmarks
 - Completion notification overhead, buffer reuse impact, hot spot, ...

Micro-Benchmarks Overview

- Latency and bandwidth
 - Send/receive and RDMA
 - Uni-directional and bi-directional
- Host communication overhead
- Completion notification overhead
 - Using polling
 - Using interrupt
- Buffer reuse impact
 - Different buffer reuse rate
- Host spot tests
 - Hot spot send
 - Host spot send and receive

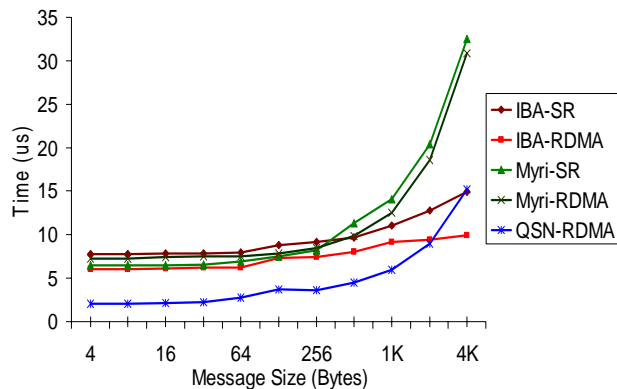
Experimental Testbed

- Eight SuperMicro SUPER P4DL6 nodes of Linux Clusters
 - ServerWorks GC chipsets
 - Dual Xeon 2.4 GHz processors
 - 512 MB memory
 - 400 MB/s FSB
- Connected with all three interconnects

Latency

- Frequently used to characterize interconnect performance
- Send/receive and RDMA write
 - RDMA only for Quadrics
- Test carried out in a ping-pong fashion
 - Multiple iterations
 - Average half round-trip time

Latency Results



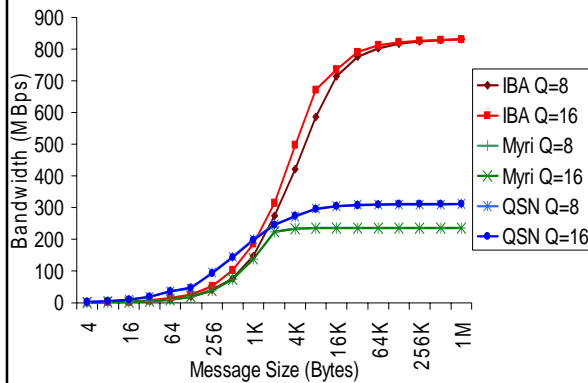
IBA SR	7.8 us
IBA RDMA	6.0 us
Myri SR	6.5 us
Myri RDMA	7.3 us
QSN RDMA	2.0 us

- Quadrics delivers best latency;
- IBA RDMA is better than Myrinet

Bandwidth

- Determine the maximum sustained data rate
- Predefine Queue Size Q
- Procedure
 - Sender sends Q back-to-back messages
 - Waits for $Q/2$ messages to finish
 - Sends another $Q/2$ messages
 - At least $Q/2$ and at most Q outstanding messages

Bandwidth Results



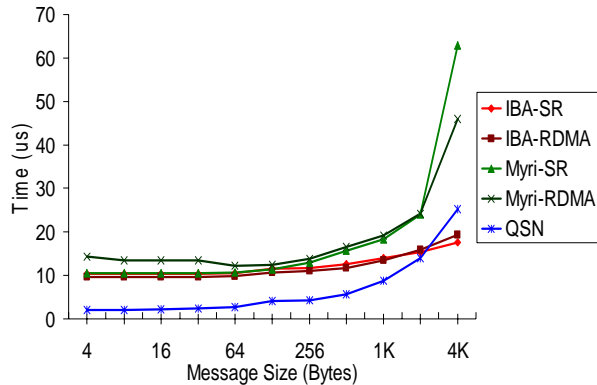
IBA Q=8	830 MBps
IBA Q=16	830 MBps
Myri Q=8	236 MBps
Myri Q=16	236 MBps
QSN Q=32	311 MBps
QSN Q=16	311 MBps

- Peak bandwidth reached with Q size 16 for all interconnects

Bi-Directional Latency and Bandwidth

- More stress on the communication layer
- Similar to uni-directional tests, but both sides send simultaneously

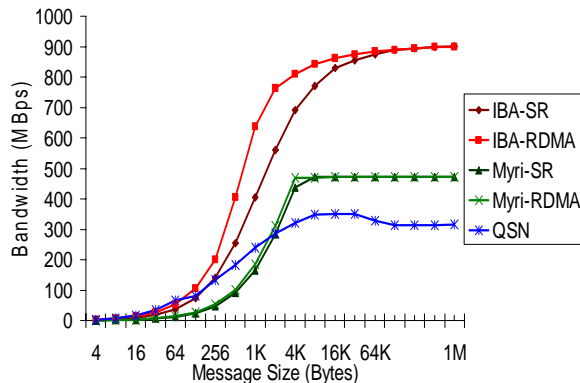
Bi-Directional Latency Results



IBA SR	10.3 us
IBA RDMA	9.7 us
Myri SR	10.5 us
Myri RDMA	14.4 us
QSN RDMA	2.1 us

- Bi-directional latency is worse than uni-directional latency
- Quadrics < IBA < Myrinet

Bi-Directional Bandwidth Results



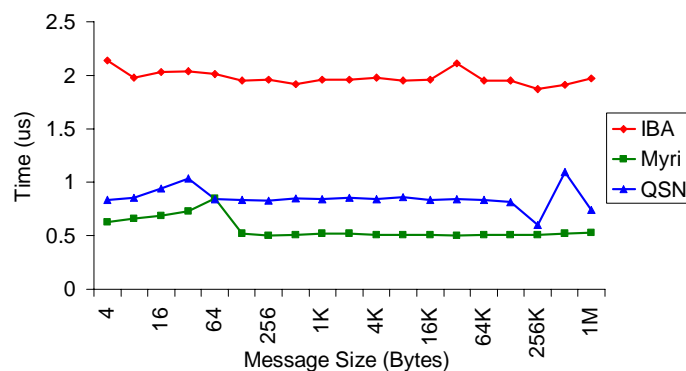
IBA SR	901 MBps
IBA RDMA	900 MBps
Myri SR	471 MBps
Myri RDMA	470 MBps
QSN	316 MBps

- PCI-X becomes the bottleneck for IBA
- Myrinet performs better than Quadrics
- Myrinet performance is limited by link bandwidth

Host Communication Overhead

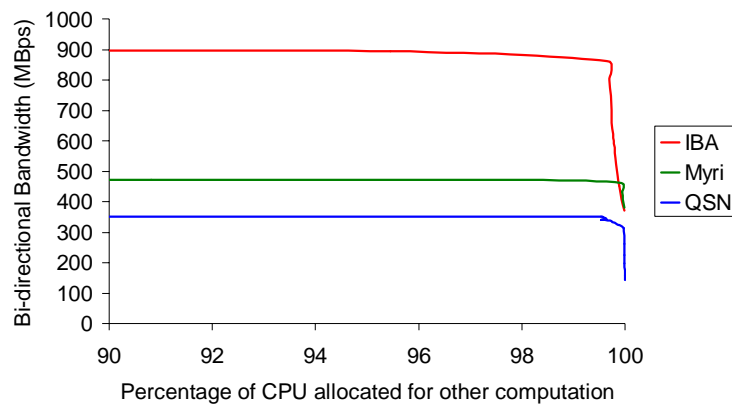
- The time CPU spent on communication
- A measure of the capability to overlap communication with computation
- Overhead in latency tests
 - Directly measure the time
- Overhead in bandwidth tests
 - Maximum computation inserted without decreasing bandwidth (CPU utilization)

Overhead in Latency Test



- All three interconnects show small overhead
- Myrinet < Quadrics < IBA

Overhead in Bandwidth Test

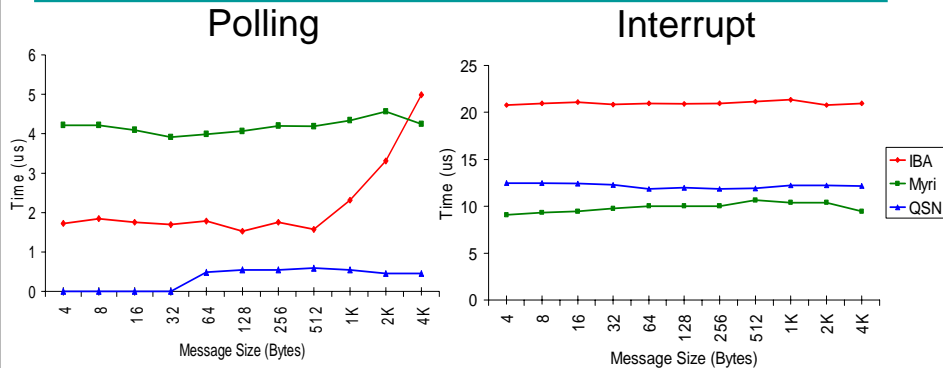


- All three interconnects allow a very large percentage of computation without decreasing performance

Completion Notification Overhead for RDMA

- RDMA operations are transparent to the remote side
- Explicit mechanism needed to signal message arrival
 - Polling on memory content may not be safe
- Checking the Notification
 - Polling
 - Interrupt
- Use micro-benchmark to measure the latency increase when notification is enabled

Overhead of Checking the Notification (Polling and Interrupt)



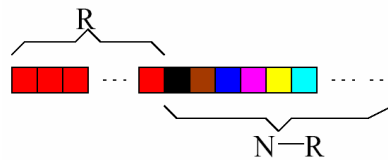
- Polling: Quadrics < IBA < Myrinet
- Interrupt: Myrinet < Quadrics < IBA

Buffer Reuse Pattern

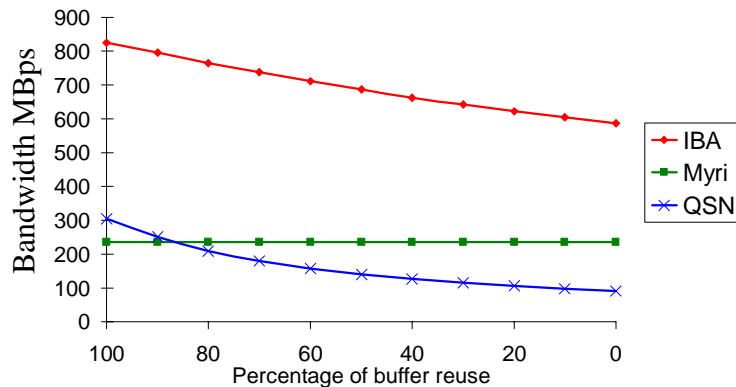
- Communication performance is sensitive to buffer reuse pattern
 - Result of address translation in network interfaces
- Cannot be characterized if only one buffer is used in the test

Buffer Reuse Tests

- Buffer Reuse Rate Test
 - N iterations in tests
 - First R iterations use the same buffer
 - (N-R) iterations use completely different buffer
 - Rate: R/N



Buffer Reuse Rate Test Results

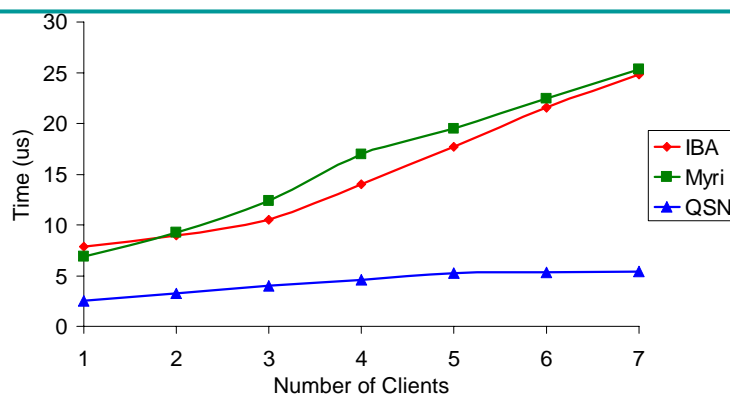


- InfiniBand and Quadrics are sensitive to reuse rate
- Myrinet is not sensitive (only drops slightly)

Hot Spot Tests

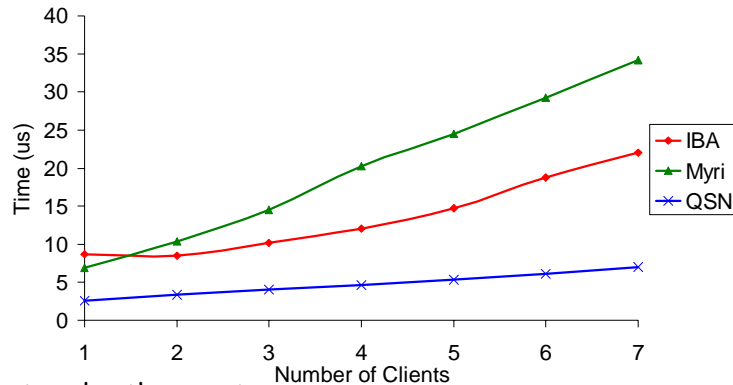
- Measure how interconnects handle unbalanced communication patterns
- Hot spot send
 - Master node sends data to a number of slave nodes, and receives an ack from one of them.
- Hot spot send and receive
 - Master node sends a note to slave nodes and receives data from each one of slave nodes

Hot Spot Send Results



- InfiniBand and Myrinet latencies increase with the number of slaves
- Quadrics scales very well

Hot Spot Send and Receive Results



- Myrinet scales the worst
- InfiniBand latency also increases with the number of slaves
- Quadrics scales very well

Conclusions

- Presented a suite of micro-benchmarks
 - Focus on user-level RDMA operations
 - Characterize different aspects of communication performance
- A comprehensive performance comparison of InfiniBand, Myrinet and Quadrics
- Different interconnects have different strengths and weaknesses
- Upper layer designers can use these tests to evaluate and select an interconnect
- Lower layer designers can use these tests to optimize their implementation for a given upper layer

Follow-up Work

- A complete suite of micro-benchmarks for InfiniBand (MIBA)
 - Performance Tools '03 Conference, to be presented, Sep. '03
- MPI level performance comparison of InfiniBand, Myrinet and Quadrics
 - Supercomputing '03 Conference, to be presented, Nov. '03

Web Pointers

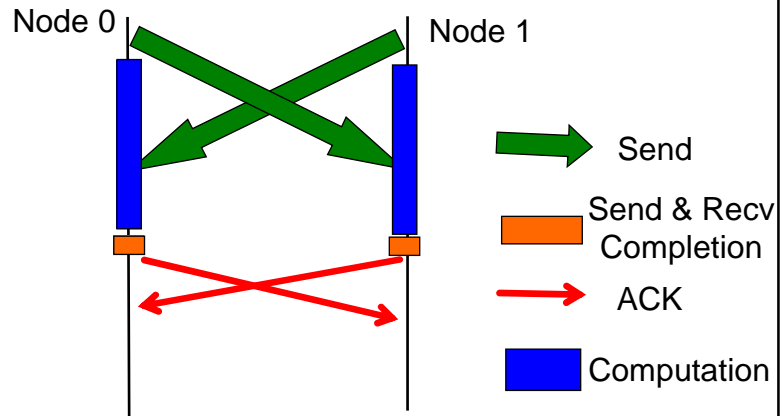
 home page

<http://www.cis.ohio-state.edu/~panda/>

<http://nowlab.cis.ohio-state.edu/>

E-mail: panda@cis.ohio-state.edu

Overhead in Bandwidth Test



- How much computation one can do without affecting bandwidth ?