



Building Multirail InfiniBand Clusters: MPI-Level Design and Performance Evaluation

J. Liu, A. Vishnu and D. K. Panda

Department of Computer Science and
Engineering

The Ohio State University

{liuj,vishnu,panda}@cse.ohio-state.edu



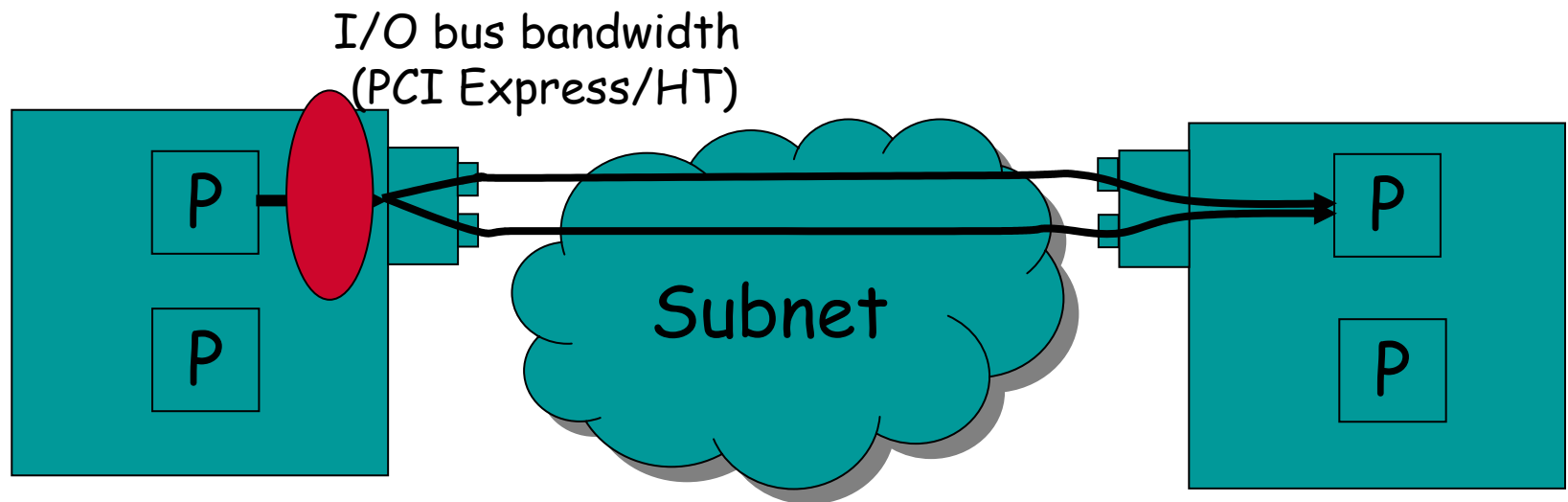
Presentation Outline

- Motivation
- Multirail MPI Design Challenges
- Detailed Design Issues
- Performance Evaluation
- Conclusions and Future Work

InfiniBand

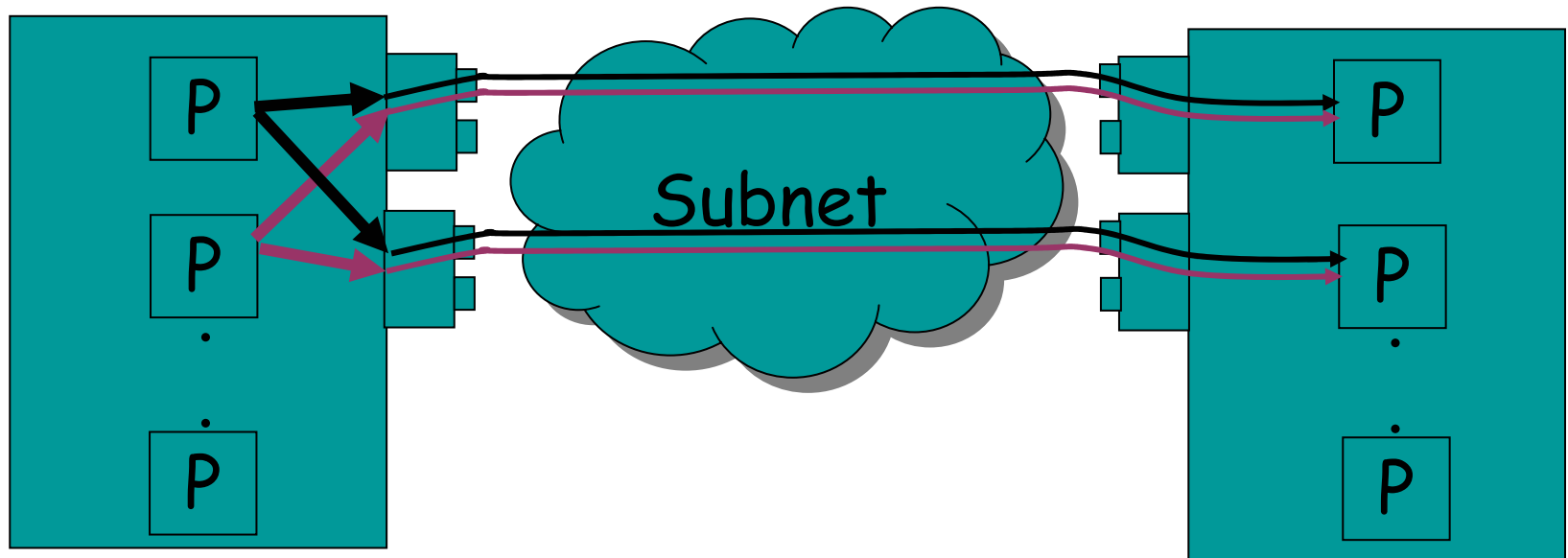
- An interconnect technology to connect I/O nodes and processing nodes
- Provides high performance
 - Low latency
 - 10 Gbps in each direction for 4X links
 - Emerging PCI-Express provides 8X/16X bandwidth
 - DDR mechanism can provide 8X (20.0 Gbps) bandwidth
- Supports many novel features
 - Send/Receive, RDMA, Atomic, Multicast, QoS ..
- Increasingly being used in large clusters

Bottleneck: Link Bandwidth



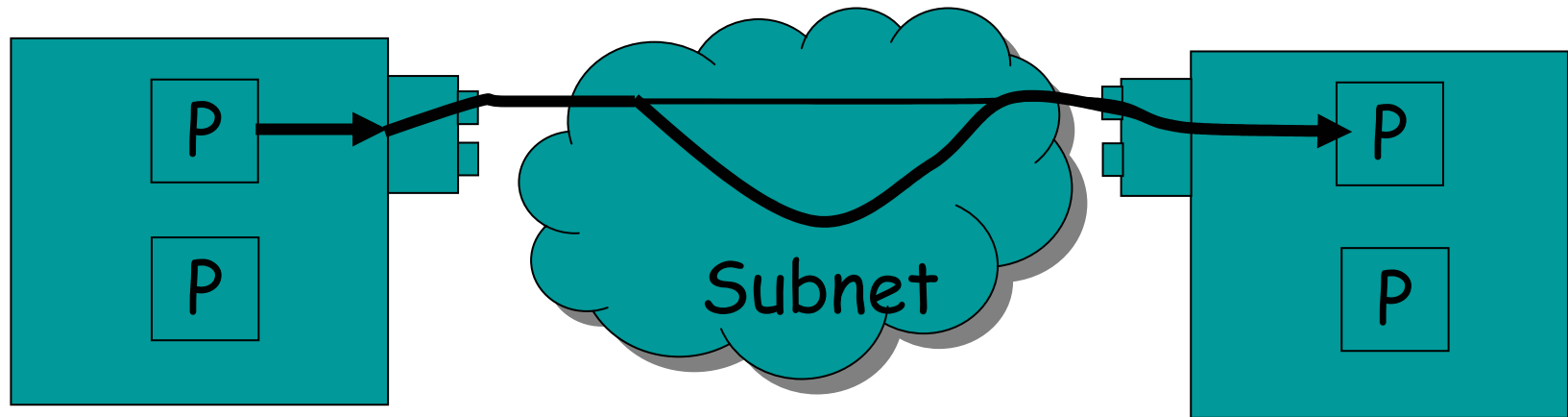
- Can Multiple Ports be used to alleviate bandwidth bottleneck (especially for PCI-Express)

Bottleneck: Communication between Multi-way SMP Systems



- Can Multiple Adapters with single ports be used for PCI-X systems to alleviate I/O bus bottleneck

Bottleneck: Network Congestion



- Can we use multiple paths to alleviate bandwidth bottleneck using LMC (both for PCI-X and PCI-Express Systems)

IBA Multirail Network Configurations

- Multiple Adapters
 - Can be used for SMP systems with I/O bus being the bottleneck
- Multiple Ports
 - Can be used with the systems, with link bandwidth as the bottleneck
- Multiple Paths with LMC
 - Can be used with above systems, when network congestion is the bottleneck

Problem Statement

- Can we design a unified MPI framework, with low overhead, flexibility, and adaptivity to support following multirail network configurations with InfiniBand:
 - Multiple Ports
 - Multiple Adapters on SMP systems
 - LMC
- What are the design challenges and issues
- How much performance benefits can be achieved with the new MPI framework on modern InfiniBand clusters

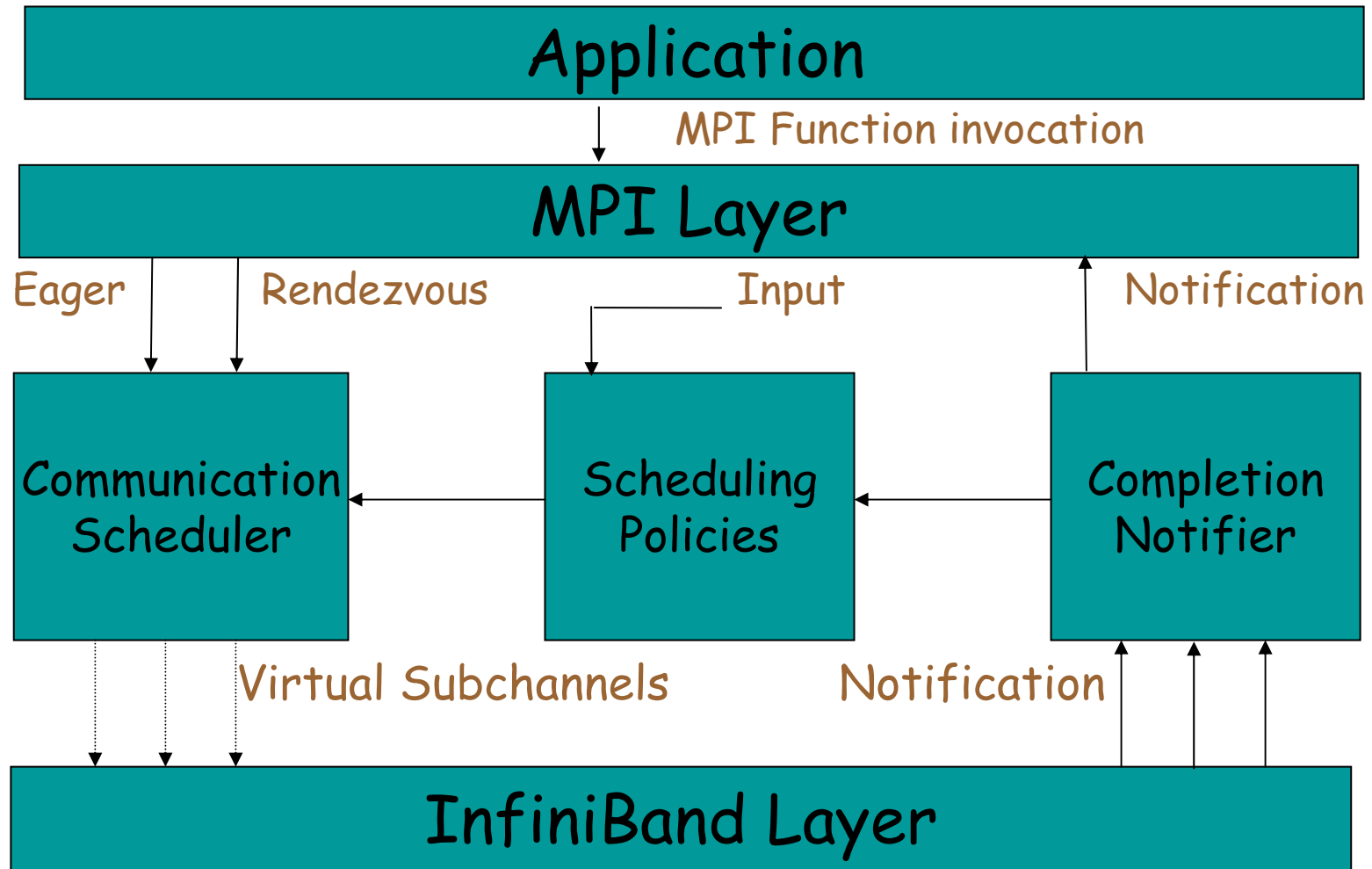
Presentation Outline

- Motivation
- *Multirail MPI Design Challenges*
- Detailed Design Issues
- Performance Evaluation
- Conclusions and Future Work

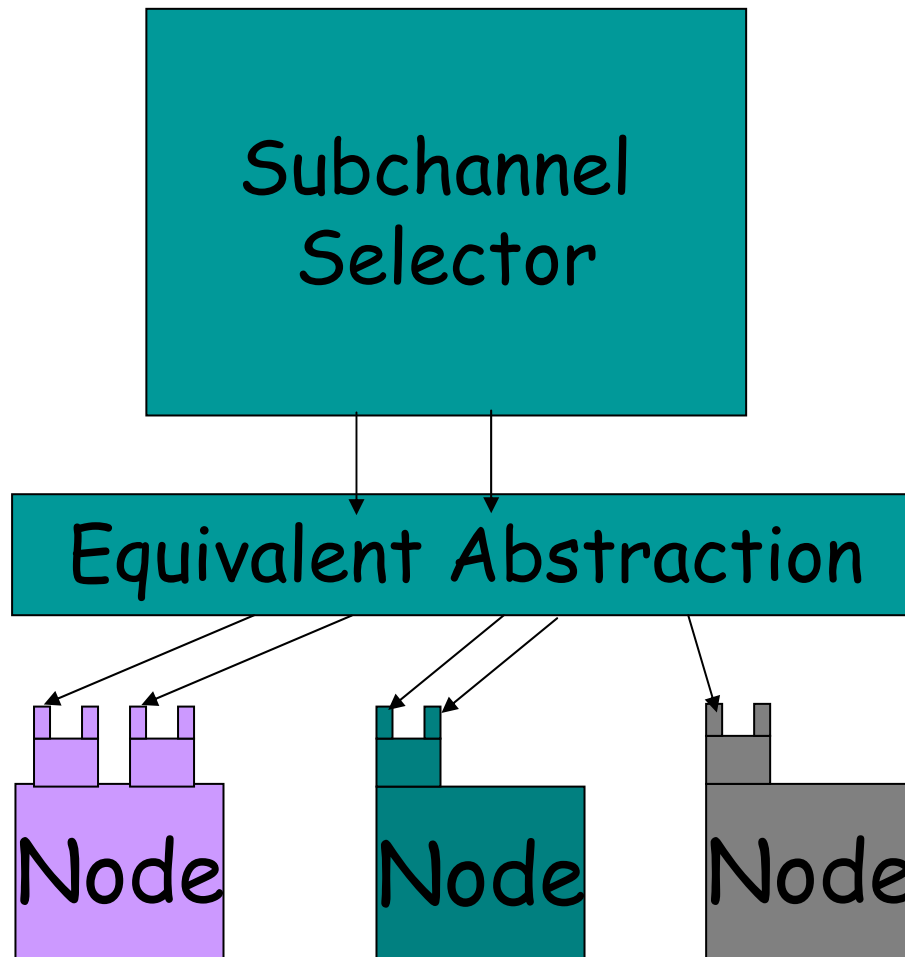
Multirail MPI Design Challenges

- How to abstract communication subchannels?
 - Channel of communication between end nodes through different ports/adapters/paths
- How to design communication scheduler and scheduling policies?
 - Schedules message transmission on a selected subchannel
- How to handle Completion Notification?

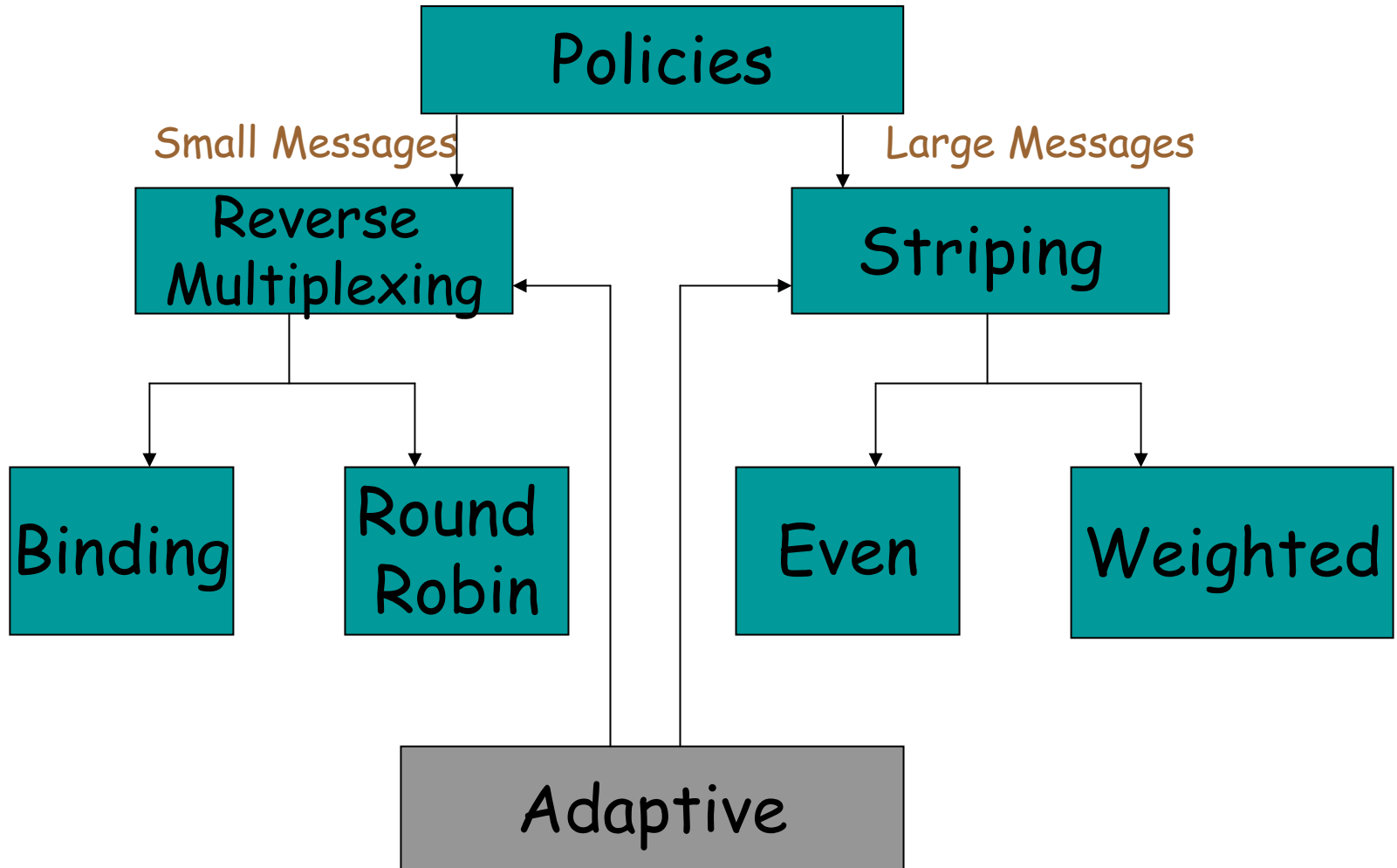
Proposed MPI Design Framework



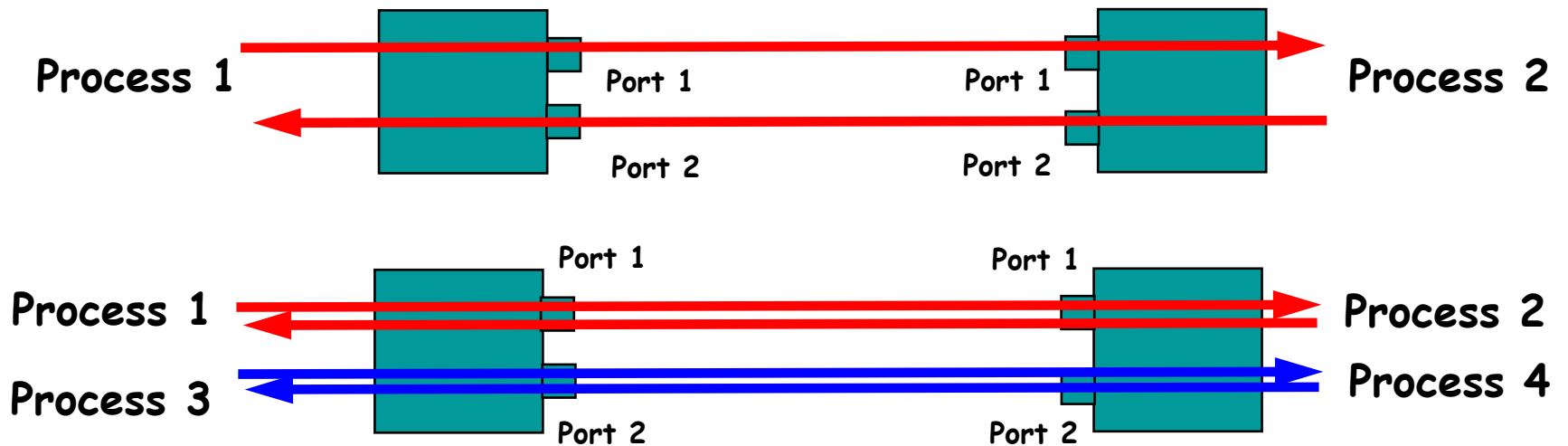
Communication Scheduler



Scheduling Policies

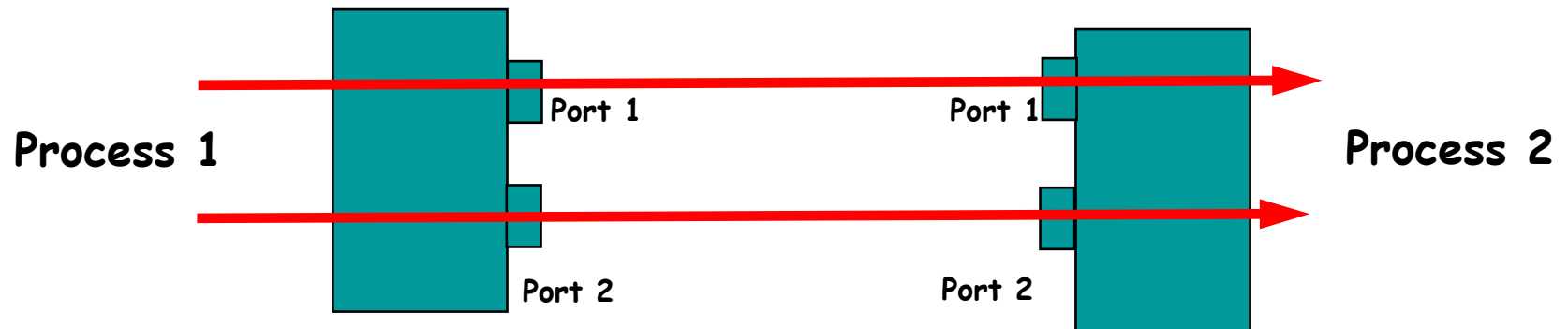


Process Binding



- Binding can take advantage of both ports in an HCA only when
 - Traffic is bi-directional, or
 - More than one process is on a single node

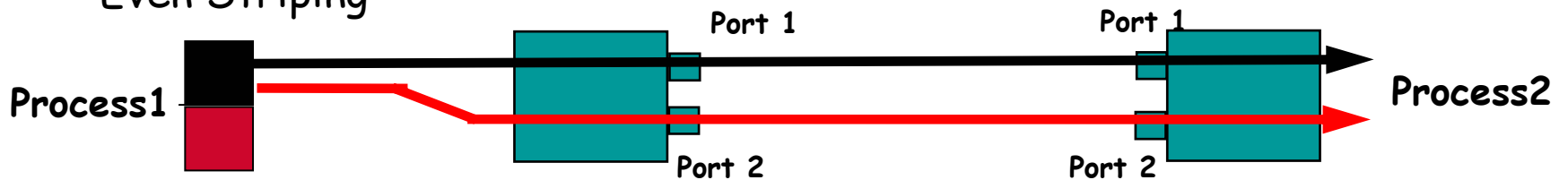
Round Robin



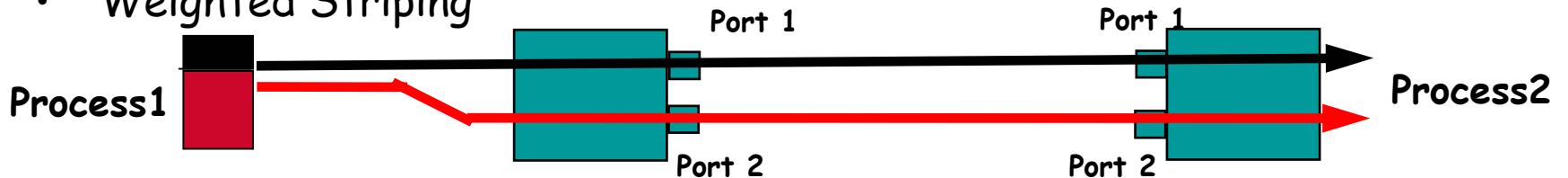
- A set of messages (window) are sent through subchannels, in round robin fashion
- Allows effective use of multiple subchannels for medium size messages

Striping

- Even Striping



- Weighted Striping



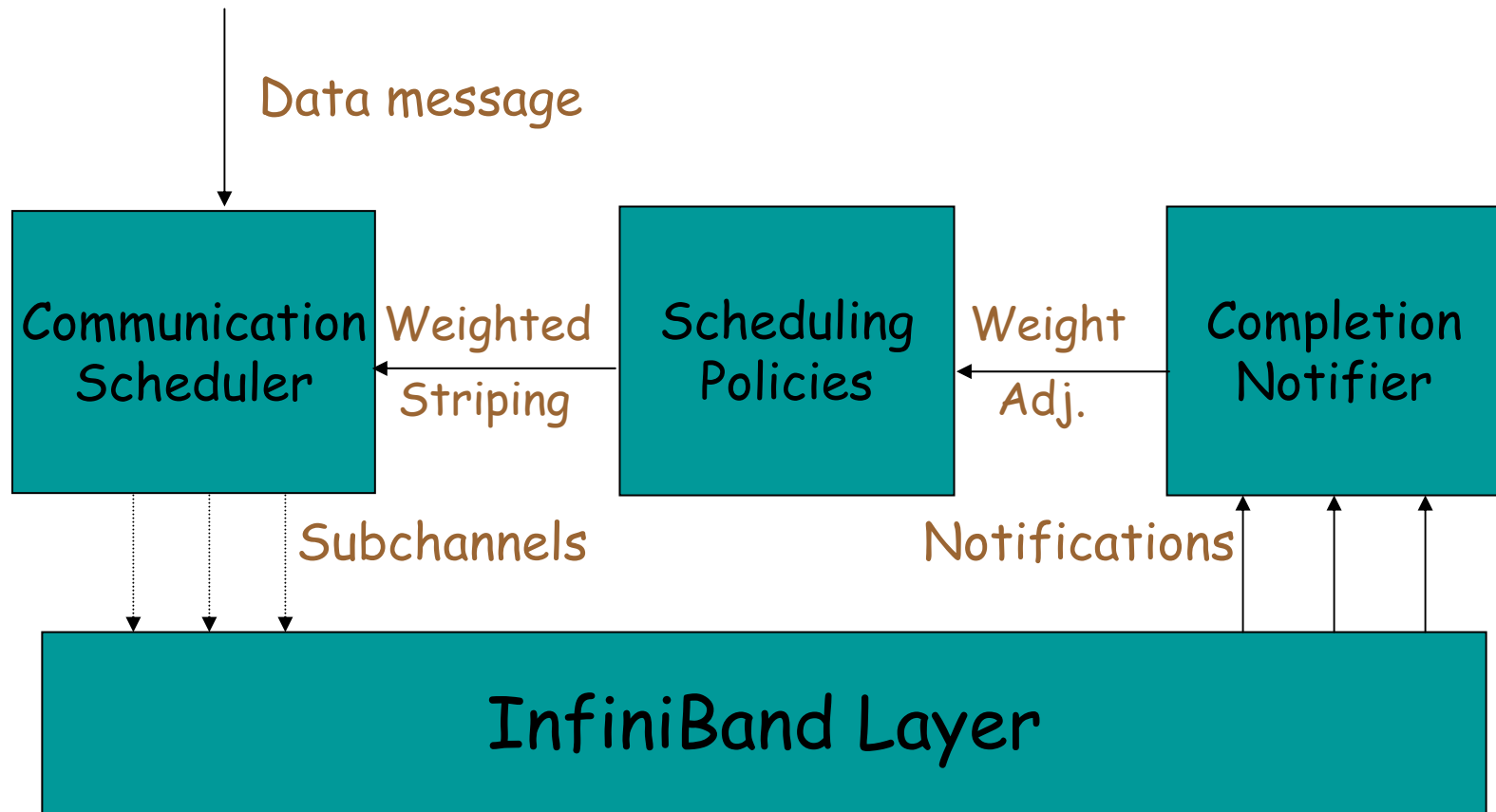
- Striping takes advantage of both ports in an HCA for
 - Both uni-directional and bi-directional traffic
 - Both one and two processes on a single node

Adaptive Striping

- Motivation



- Link bandwidth not available directly to the MPI implementation
- Bottleneck in the switches or network may reduce the path bandwidth
- Path bandwidth can also be affected by other ongoing communication at the same node.

Feedback Loop in Adaptive Striping





Presentation Outline

- Motivation
 - Multirail MPI Design Challenges
 - Detailed Design Issues
 - Performance Evaluation
 - Conclusions and Future Work
- 
- 

Detailed Design Issues

- Handling Multiple Adapters
 - Completion Queue Polling
 - Multiple Completion Queues are polled for detecting individual completions
 - Buffer Registration
 - Application buffer registration is done with all HCAs
- Out of order Message Processing
 - Multiple subchannels generate out of order messages, processing delayed unless an in-order message is received
- Multiple RDMA completion notifications
 - Across different subchannels

Incorporation into MVAPICH (OSU MPI for InfiniBand)

- MVAPICH is based on MPICH and MVICH
- Open Source
- Current versions are
 - MVAPICH 0.9.4 (MPI-1)
 - MVAPICH2 0.6.0 (MPI-2)
- <http://nowlab.cis.ohio-state.edu/projects/mpi-iba/>
- Directly downloaded and being used by more than 150 organizations and industry worldwide
- Available in the software stack distributions of IBA vendors
- Empowers multiple InfiniBand clusters in the TOP 500 list
- The proposed design has been integrated in MVAPICH
 - Will be released as MVAPICH 0.9.5 version within a few weeks
 - Will be incorporated into MVAPICH2 framework shortly

Presentation Outline

- Motivation
- Multirail MPI Design Challenges
- Detailed Design Issues
- Performance Evaluation
- Conclusions and Future Work

Experimental Testbed

- Testbed1 (PCI-Express with Two Ports)
 - Four 3.4 GHz Intel Xeon Dual Processor, EM64T architecture
 - 1GB Main Memory
 - 4X PCI-E with MT25208 HCAs
 - InfiniScale MTS2400 switch

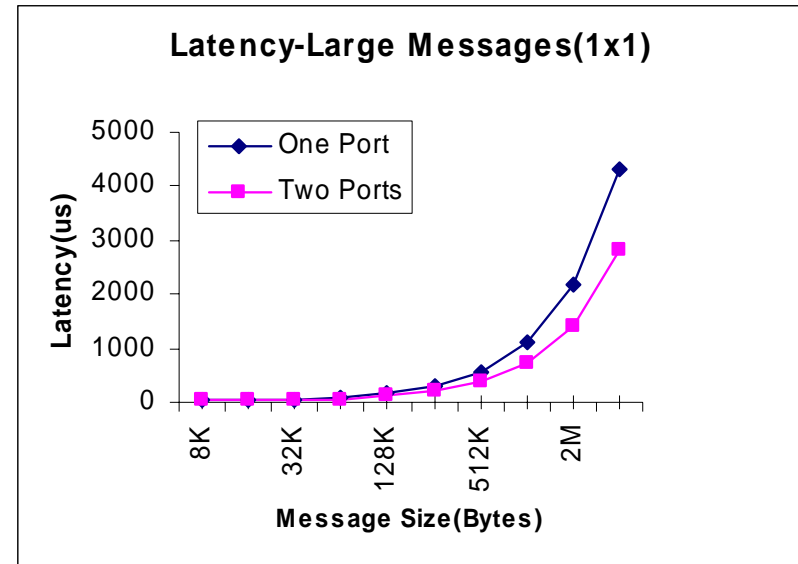
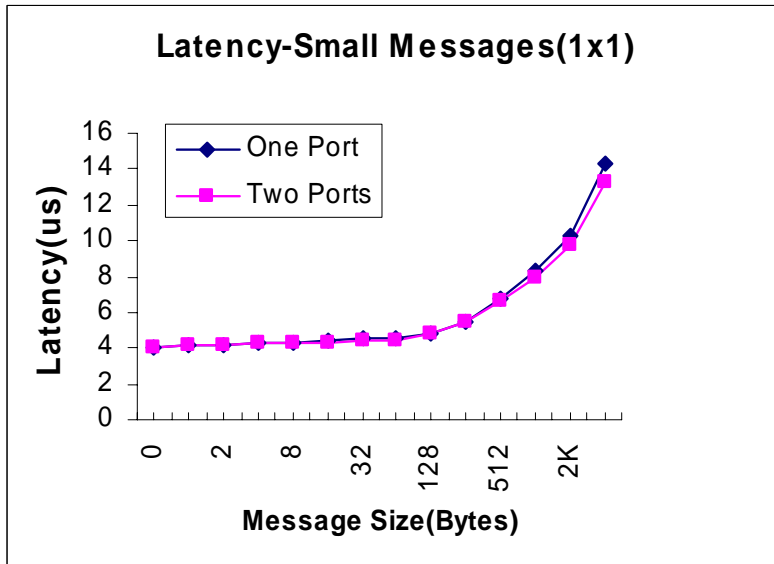
Experimental Testbed

- Testbed2 (PCI-X with Two Adapters)
 - Eight 3.0 GHz Intel Xeon Dual Processor, IA32 architecture
 - 2GB Main Memory
 - 4X PCI-X 64-bit 133MHz with MT23108 HCAs
 - InfiniScale MTS2400 switch

MPI-level Evaluation

- Micro-benchmarks
 - Latency
 - Uni-directional bandwidth
 - Bi-directional bandwidth
- Collective communication (using Pallas)
 - Broadcast
 - All-to-All
- Applications
 - NAS Parallel Benchmarks
 - IS
 - FT

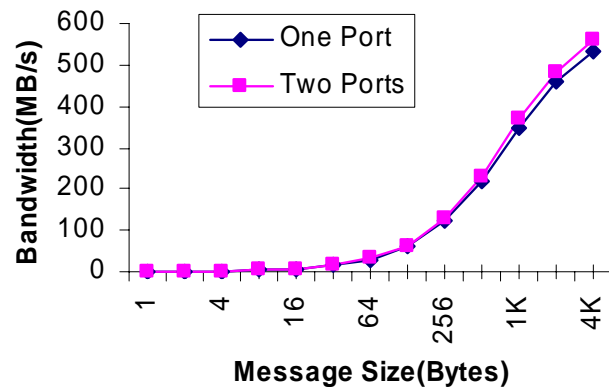
Performance Evaluation of PCI-Express (Two Ports)



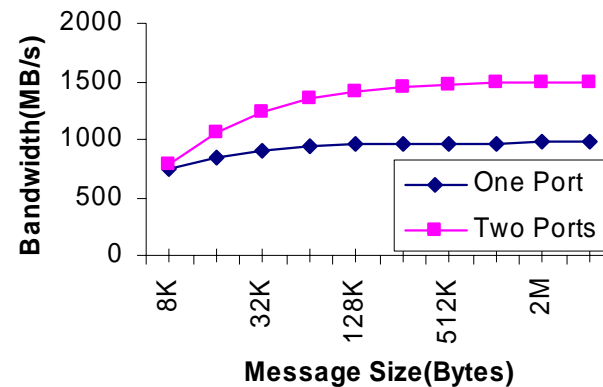
- For small messages, MPI incurs a very low overhead of less than 2%
- Latency improves by 34% for large messages for dual ports

Performance Evaluation of PCI-Express (Two Ports)

Unidirectional Bandwidth-Small Messages(1x1)

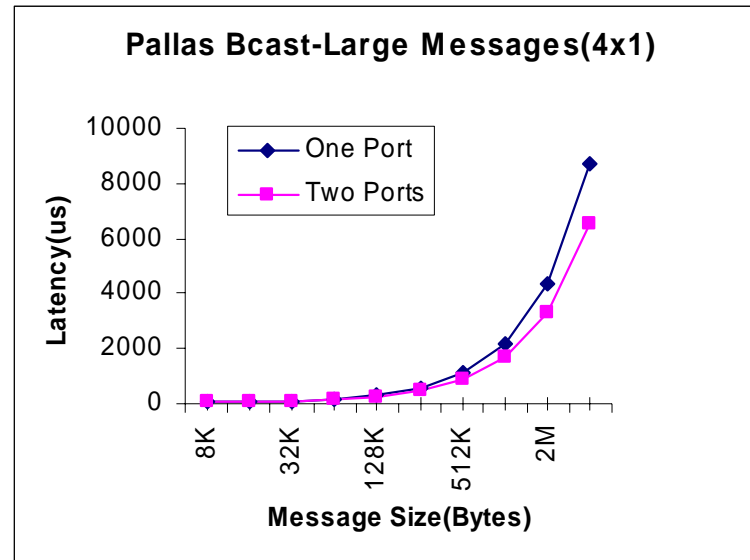


Unidirectional Bandwidth-Large Messages(1x1)



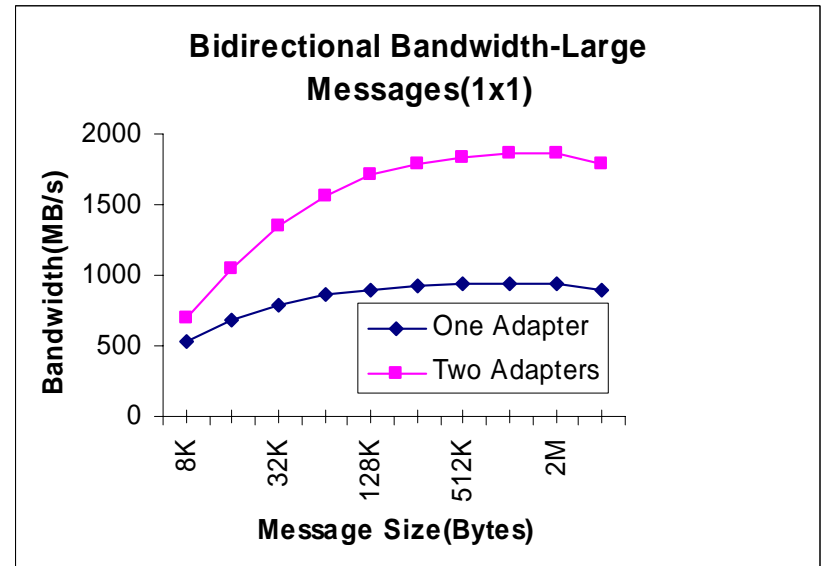
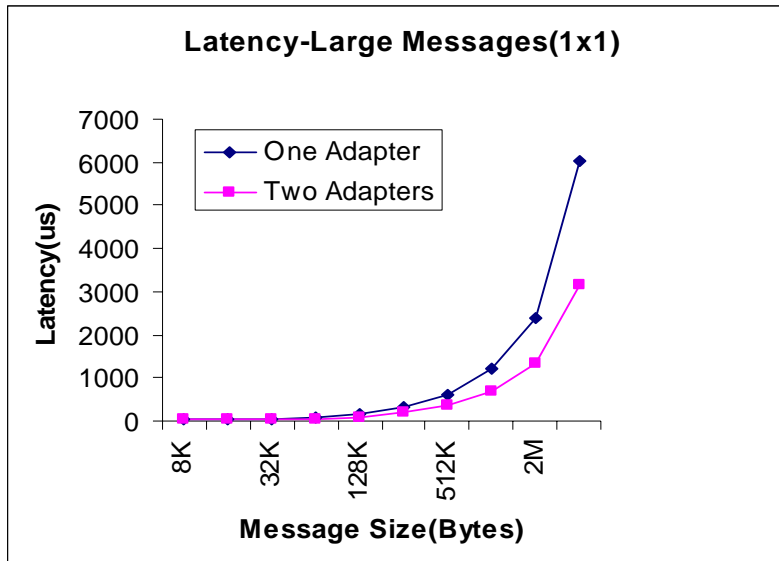
- Unidirectional Bandwidth
 - 1492 MB/sec (~1.5 GBytes/sec) with two ports
 - improves by 49% compared to one port
- Round Robin scheduling improves bandwidth by 6% for medium size messages for multiple ports
- Bidirectional bandwidth of ~2724 MBytes/sec (~2.7 GBytes/sec)

Performance Evaluation of PCI-Express (Two Ports)



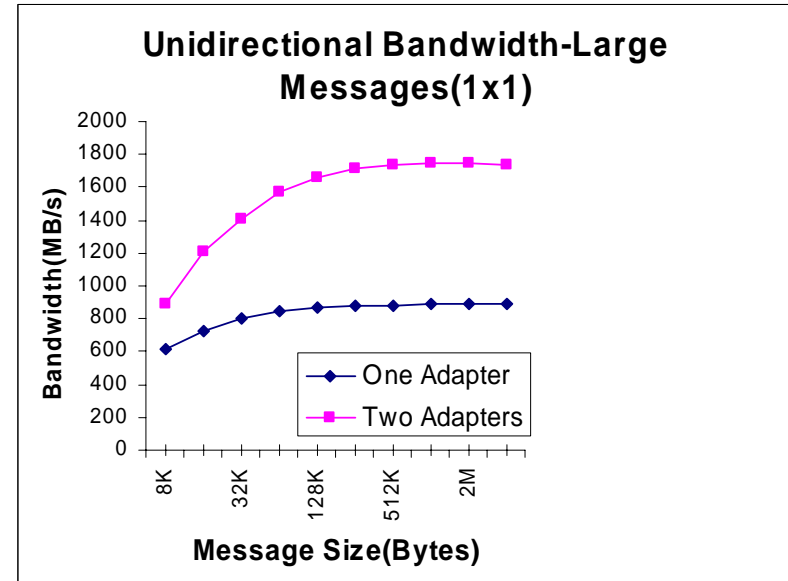
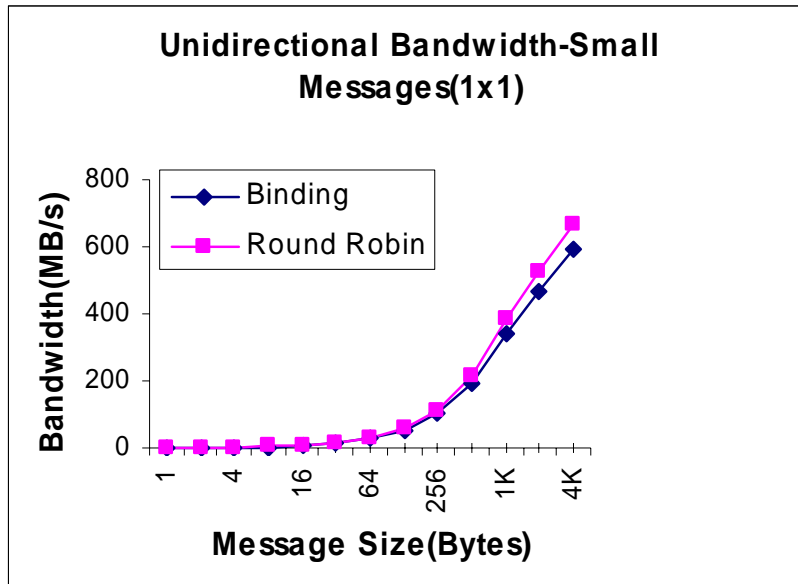
- Broadcast latency improves by 23% using dual ports on PCI-Express

Performance Evaluation of PCI-X (Two Adapters)



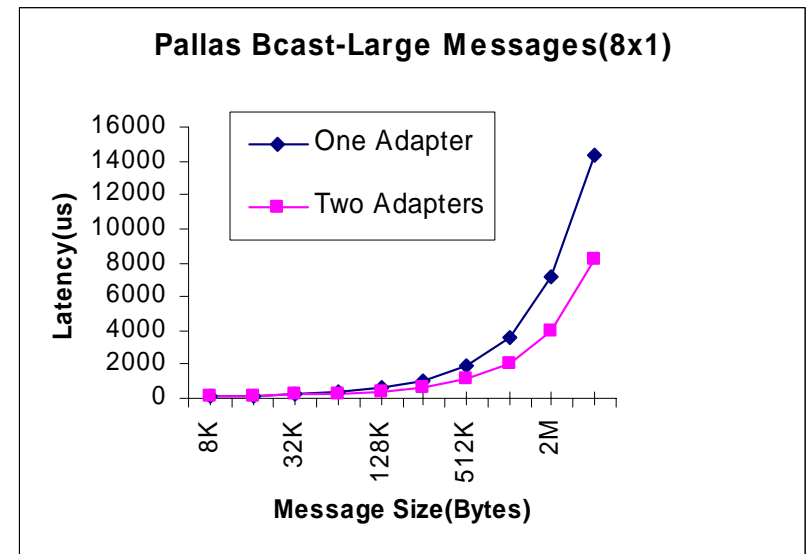
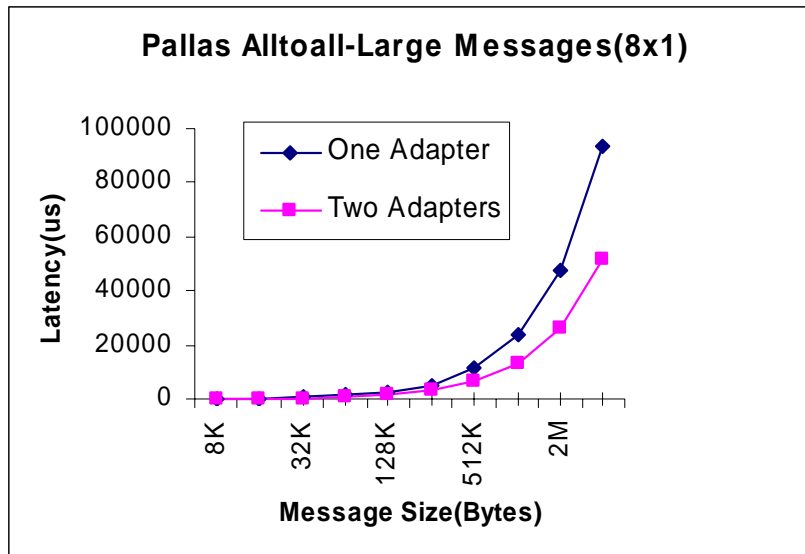
- Latency improves by 49%
- Bidirectional bandwidth
 - around 1800 MBytes/sec
 - improves by 98%

Performance Evaluation of PCI-X (Two Adapters)



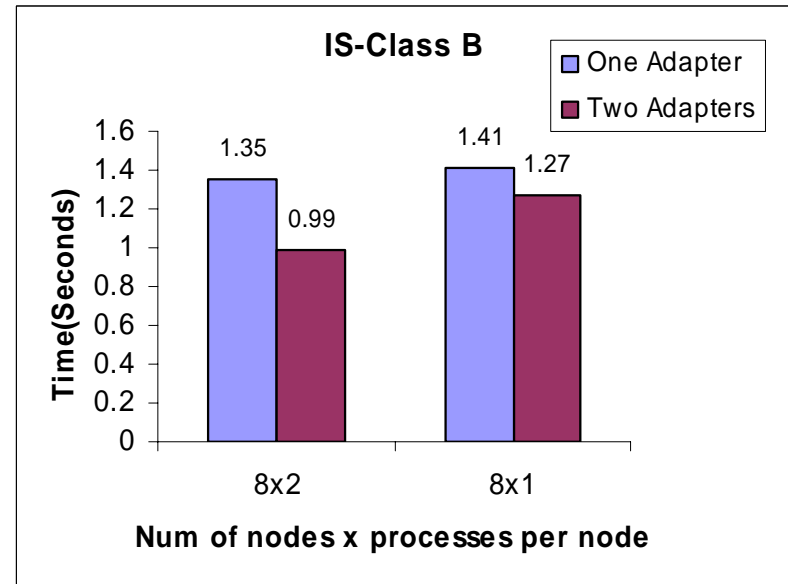
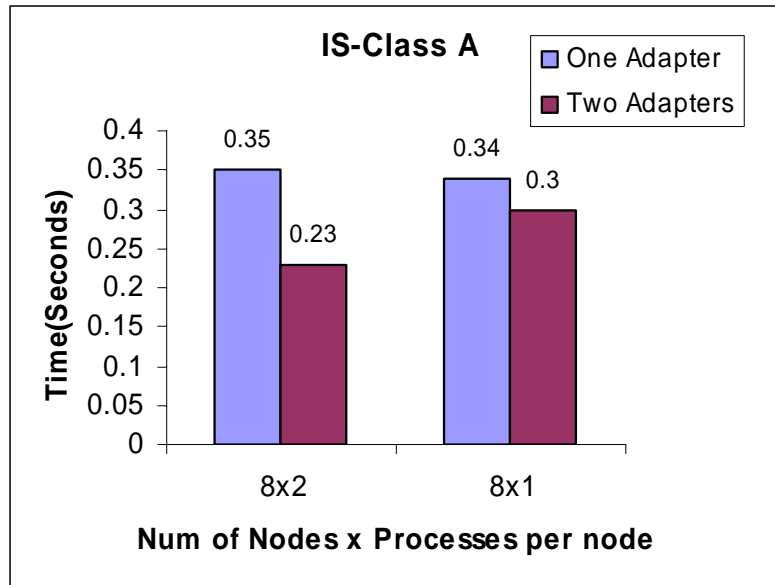
- Unidirectional Bandwidth almost doubles
- Round Robin scheduling for Medium Size Messages improves the bandwidth by 15%

Performance Evaluation of PCI-X (Two Adapters)



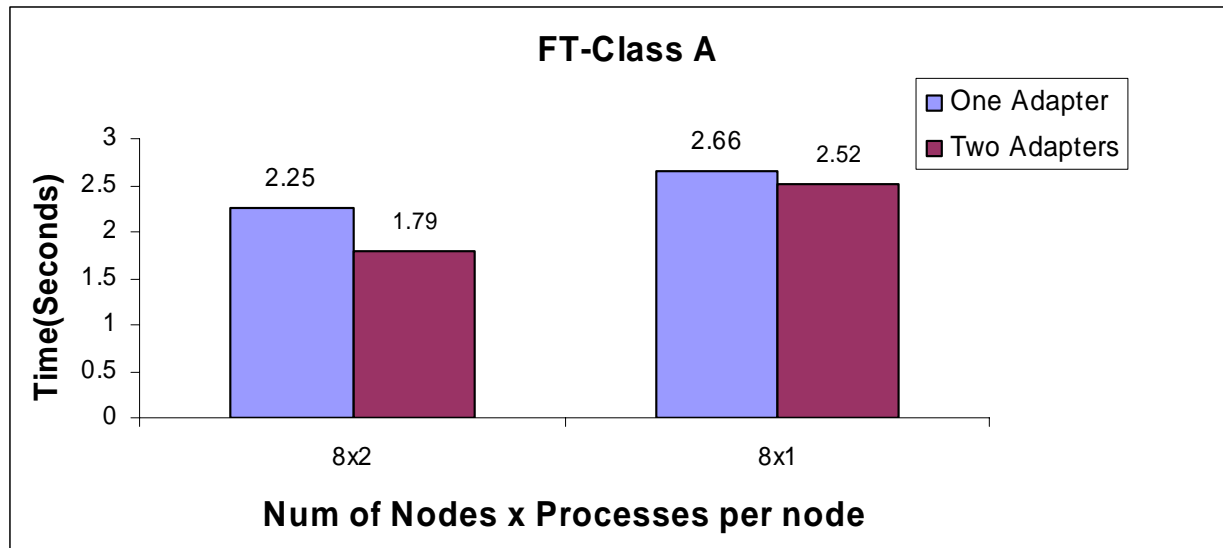
- AlltoAll latency reduces by 41% for large messages
- Broadcast latency reduces by 50% for large messages

Performance Evaluation of PCI-X (Two Adapters)



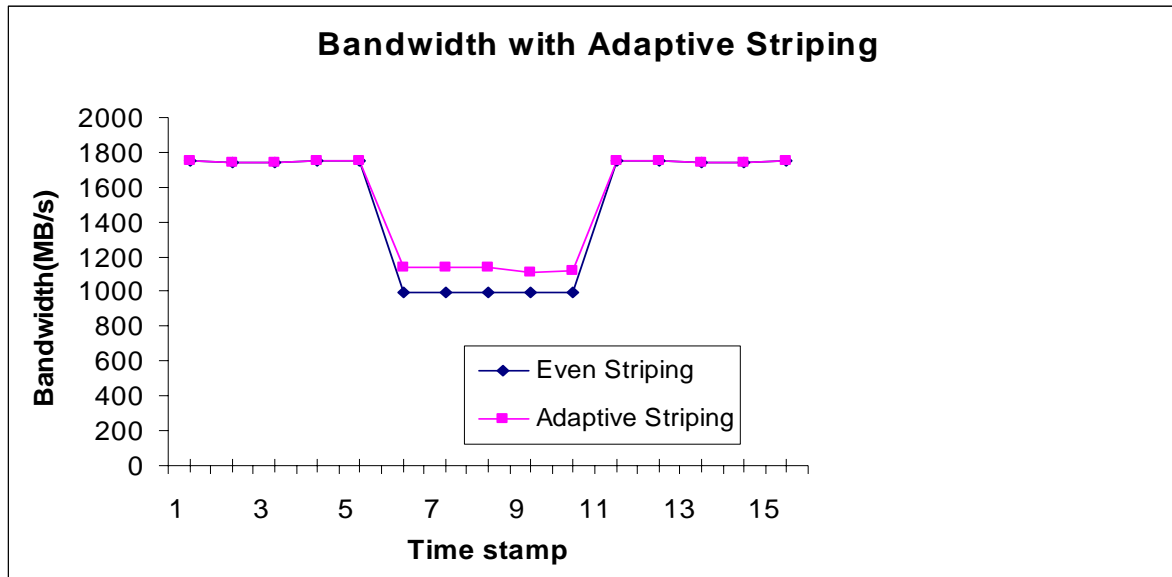
- For 8x2 configuration:
 - Reduction in execution time by 34% for Class A, and 28% for Class B
- For 8x1 configuration:
 - by 11% for Class A, and 10% for Class B

Performance Evaluation of PCI-X (Two Adapters)



- For 8x2 configuration:
 - reduction in execution time by 20%
- For 8x1 configuration:
 - reduction in execution time by 6%

Performance Evaluation of PCI-X (Two Adapters)



- Adaptive Striping outperforms Even Striping when ongoing communication is present

Presentation Outline

- Motivation
- Multirail MPI Design Challenges
- Detailed Design Issues
- Performance Evaluation
- Conclusions and Future Work

Conclusions

- Presented the need for different ways for connecting multirail configurations in emerging InfiniBand clusters
- A detailed MPI framework was presented
- Various Scheduling policies for message transmission were proposed (Striping, Binding, reverse multiplexing, etc.)
- Multirail configuration shows an improvement in execution time of up to 34% for NAS IS
- In case of network congestion, proposed adaptive scheme outperforms the static even striping scheme
- These solutions will help to reap the benefits of InfiniBand in designing clusters with multirail configurations



Continuing and Future Work

- Adding Collective Communication with Multirail support
- Evaluation of LMC mechanism for taking advantage of multiple paths in the subnet
- Evaluation of enhanced scheduling policies for multirail networks with different bandwidths

Web Pointers

NBC

home page

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

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

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