



Performance Evaluation of InfiniBand over PCI Express



Jiuxing Liu, Amith Mamidala,
Abhinav Vishnu, and Dhabaleswar K. Panda

Department of Computer Science and
Engineering

The Ohio State University



•
•
•

Presentation Outline

- Introduction
- Motivation and Methodology
- Experimental Results
- Conclusions



• • • • • • • • • •



InfiniBand



- A new interconnect technology that connect I/O nodes and processing nodes
 - Nodes connect to the fabric using Host Channel Adapters (HCAs)
- High performance
 - Very low latency
 - 10 Gbps bandwidth in each direction for 4x links
 - 8 Gbps or 1 GB/s available due to 8B/10B encoding
 - 2 GB/s total link bandwidth
- Many features
 - Send/receive, RDMA, atomics, multicast, QoS, ...





PCI



- The standard local I/O interface in the last 10 years
- Based on a parallel bus
- Load/store software interface
- PCI-X is an extension to PCI
 - 64 bit/133 MHz PCI-X supports around 1.0 GB/s peak bandwidth





Drawbacks of Using InfiniBand with PCI



- High latency
 - HCAs are connected through an additional I/O bridge
- Limited bandwidth
 - Parallel bus in PCI limits the bus frequency due to signal skew
 - Bandwidth shared by all devices connected to the same bus
 - Peak bandwidth of 1.0 GB/sec can not fully support the bidirectional transfer demand (2.0 GB/sec) of IBA 4X link



PCI Express



- The next generation local I/O interconnect
- Point-to-point, serial interface
- Scalable bandwidth
 - 8x PCI Express can support
 - Bidirectional demand of 2.0 GB/sec for IBA 4X link
 - Can also support two 4X ports/adapters (4.0 GB/sec total)
- Software backward compatibility with PCI
- Can also support the new Advanced Switching (AS) interface





Advantages of Using InfiniBand with PCI Express



- No bandwidth bottleneck
 - Different link speeds available
 - Links are dedicated
- Low latency
 - HCAs are connected directly to the memory controller through PCI Express links

•
•
•

Presentation Outline

- Introduction
- Motivation and Methodology
- Experimental Results
- Conclusions




• • • • • • • • • •



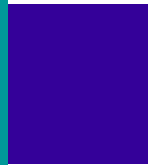
Motivation



- Conduct an initial performance comparison of InfiniBand HCAs with PCI-X and PCI Express interfaces
- Quantify the benefits of PCI Express in the context of inter-node communication with InfiniBand
- To examine whether both ports in the IBA HCAs can be activated to achieve much better performance with PCI Express



Use of Mellanox InfiniHost HCAs



- Support different I/O interfaces
 - PCI-X 64 bit/133 MHz (MT23108 HCAs)
 - 8x PCI Express (MT25208 HCAs)
- 2 physical 4X ports in each HCA
 - 4 GB/s total theoretical peak bandwidth (bidirectional)
- Was ideal for our evaluation





Methodology



- Micro-benchmarks
 - Implemented at different levels
 - Native InfiniBand VAPI interface
 - MPI interface
- MPI Applications
- Characterize different aspects of communication performance
 - One-port measurements
 - Two-port measurements
 - Different configurations to use both ports

VAPI-level Micro-Benchmarks

- Latency (single port)
 - Send/receive
 - RDMA write
 - RDMA read
 - Atomic operations
- Bandwidth (single port)
 - Uni-directional
 - Bi-directional
- Bandwidth (both ports)
 - One process per node
 - Two processes per node



MPI-level Evaluation





- Micro-benchmarks
 - Latency
 - Uni-directional bandwidth
 - Bi-directional bandwidth
- Collective communication (using Pallas)
 - Broadcast, all-to-all, reduce, and all-reduce
- Applications (some NAS benchmarks)



Experimental Testbed



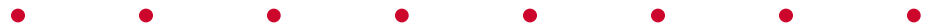
- Node configuration
 - Supports both PCI Express and PCI-X
 - 3.4 GHz Intel Xeon running in 64 bit mode (EM64T)
 - 512 MB main memory
 - Linux Redhat operating system
 - Intel compiler 8.1 (Beta Version)
 - InfiniBand HCAs
 - MT25208 HCAs with 8x PCI Express
 - MT23108 HCAs with 64 bit/133 MHz PCI-X
 - InfiniBand Switch
 - InfiniScale 24 port switch
- 
- 



Intel EM64T Technology



- Intel Extended Memory 64 Technology
- 64 bit extension to IA-32
- X86-64 instruction set
- Compatibility with 32 bit applications when running in 64 bit mode



⋮

Presentation Outline

- Introduction
- Motivation and Methodology
- Experimental Results
- Conclusions



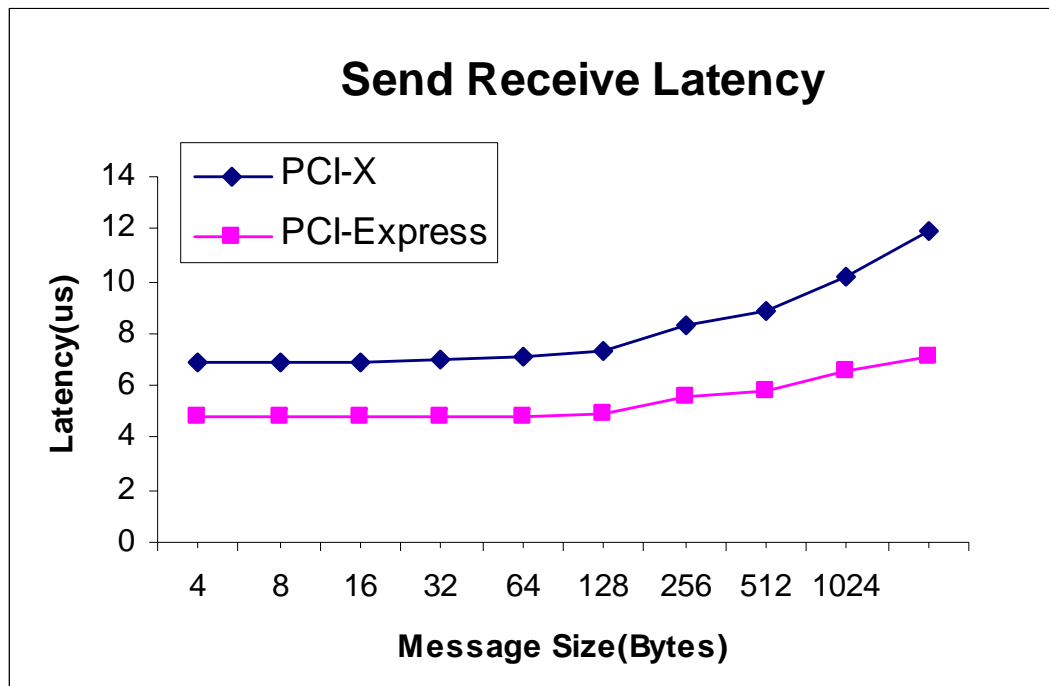


VAPI Level Latency (One Port)



- Test carried out in a ping-pong fashion
 - Timing a number of iterations after warm-up
 - Average time of half round trip
- Different InfiniBand Operations
 - Send/receive, RDMA write, RDMA read, atomic operations

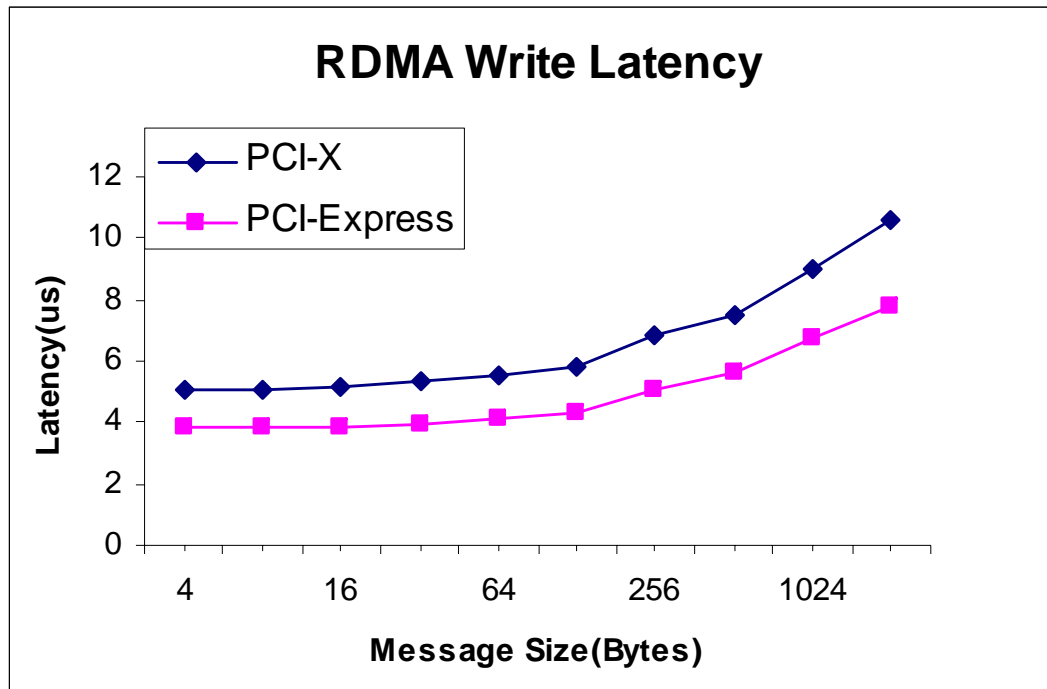
VAPI Send/Receive Latency



PCI-X	PCI-Express
7.1us	4.8us

- The VAPI level Send Receive latency improves by 24% for PCI-Express

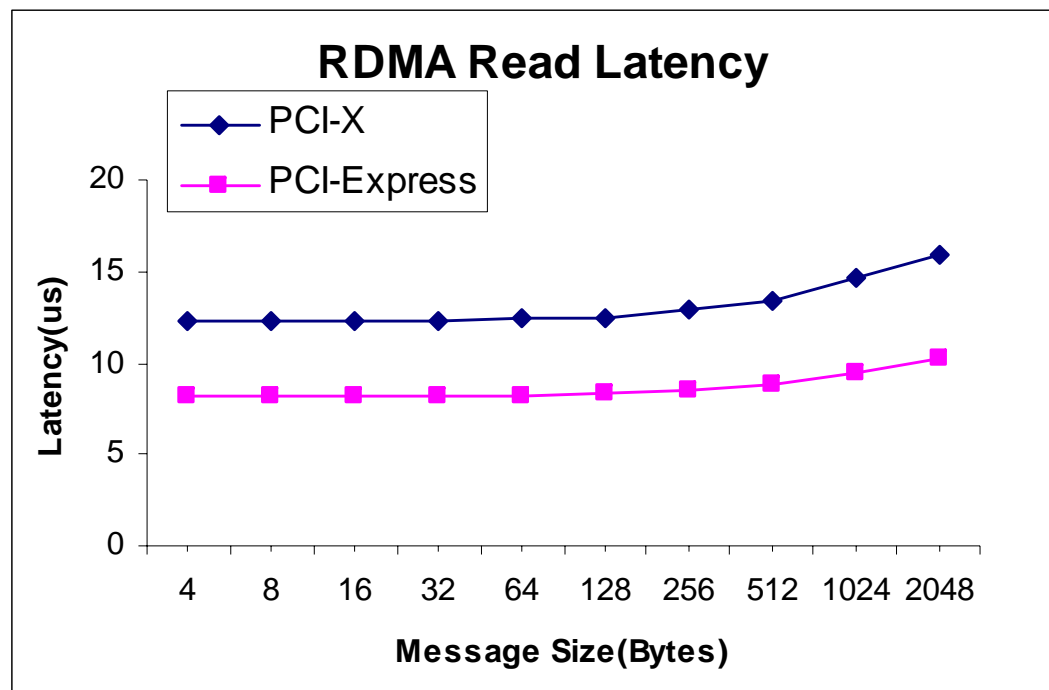
VAPI RDMA Write Latency



PCI-X	PCI-Express
5.0us	3.8us

- The VAPI level RDMA Write latency improves by 24% for PCI-Express

VAPI RDMA Read Latency



PCI-X	PCI-Express
12.4us	8.1us

- The VAPI level RDMA Read latency improves by 34% for PCI-Express



Atomic Operation Latency



Test Program	PCI-X	PCI-Express
Comp & Swap	12.4us	9.2us
Test & Set	12.3us	9.2us

- The Atomic Operation Latency improves by 25% for PCI-Express



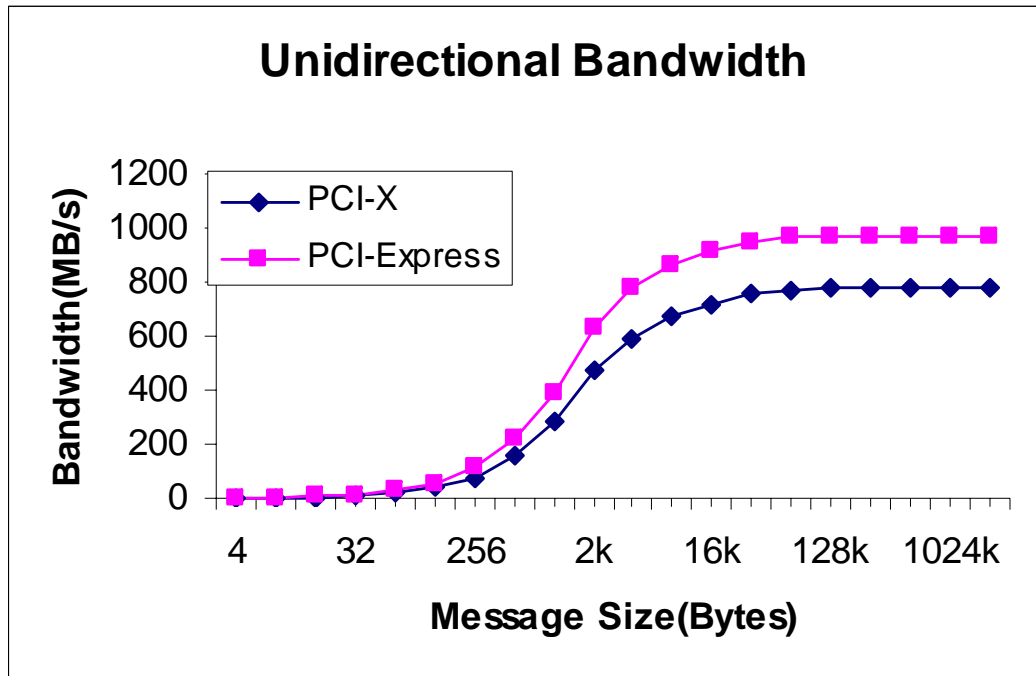
VAPI Level Bandwidth



- Determine the maximum sustained data rate
- Focus on RDMA Write
- Predefined Queue Size Q
- Procedure
 - Sender sends Q back-to-back messages
 - Waits for all messages to finish
 - Repeat for multiple iterations



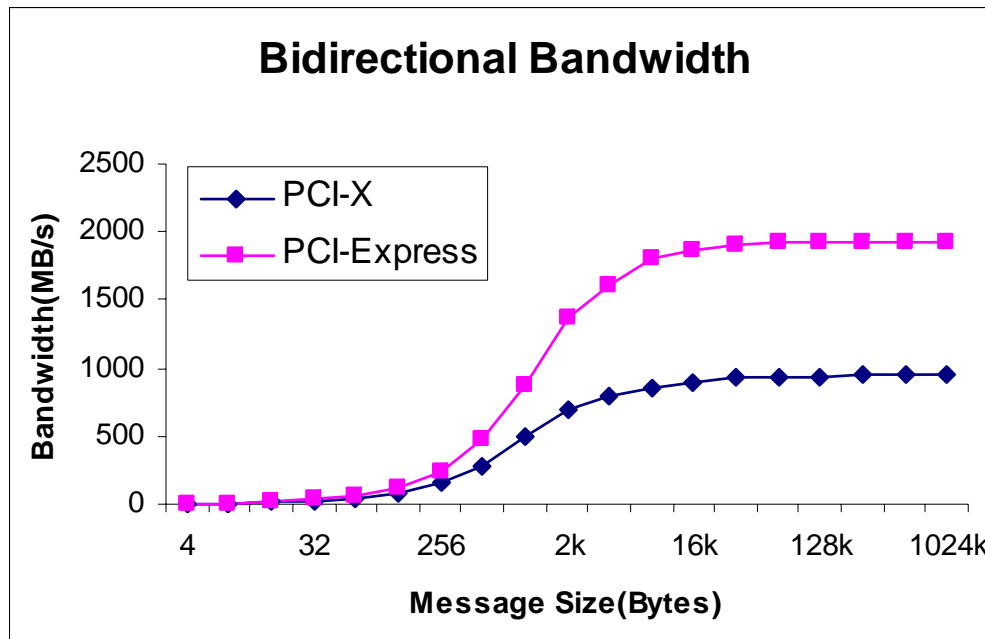
VAPI Uni-Directional Bandwidth (One Port)



PCI-X	PCI-Express
778MB/s	972MB/s

- The VAPI level Uni-Directional Bandwidth improves by 25% for PCI-Express

VAPI Bi-Directional Bandwidth (One Port)



PCI-X	PCI-Express
945MB/s	1932MB/s

- The VAPI level Bi-Directional Bandwidth improves by 104% for PCI-Express



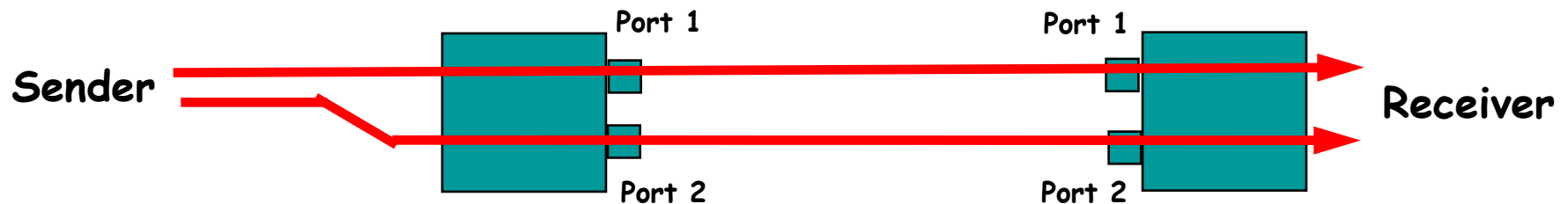
VAPI Bandwidth with Two Ports



- Both ports were activated during the tests
- Large messages can use both ports
 - Striping: Messages are divided into smaller chunks and sent out using both ports.
 - Binding: Message are never divided. But different processes can use different ports.

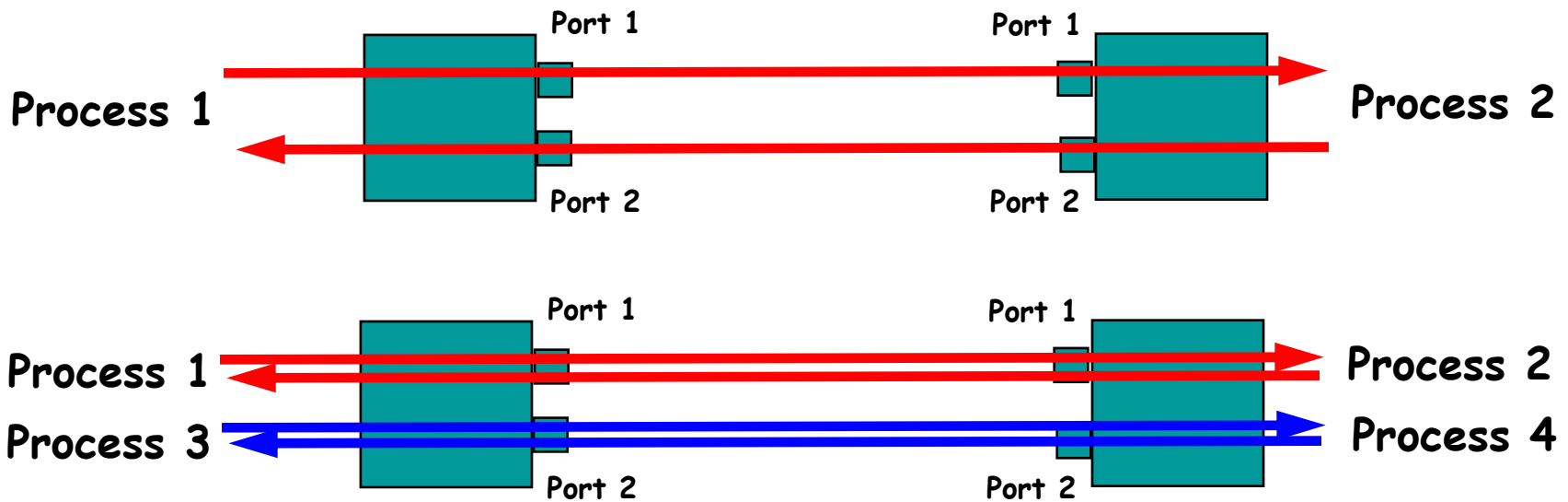


Striping



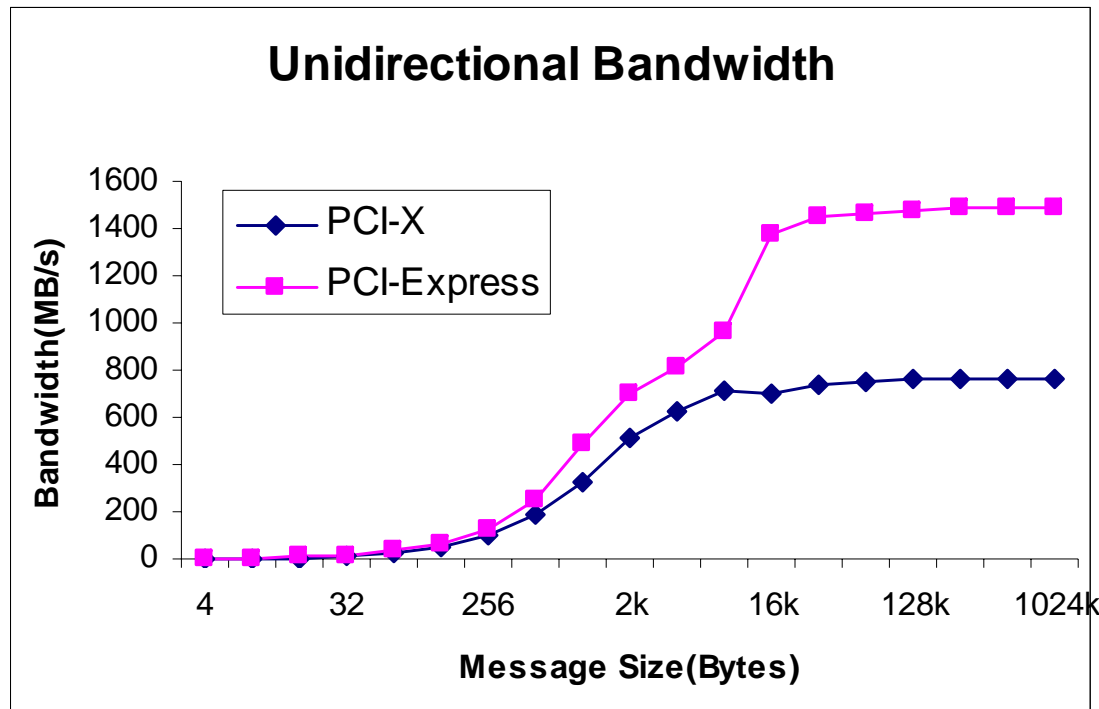
- Striping takes advantage of both ports in an HCA for
 - Both uni-directional and bi-directional traffic
 - Both one and two processes on a single node

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

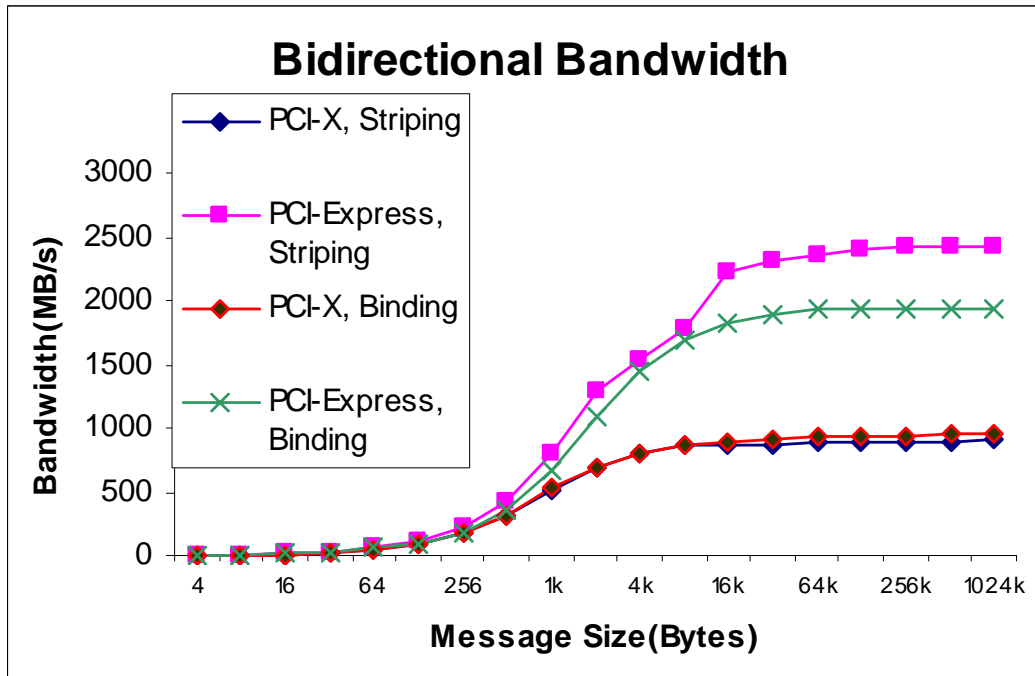
VAPI Uni-Directional Bandwidth (Two Ports)



PCI-X	PCI-Express
768MB/s	1486MB/s

- The VAPI level Uni-Directional Bandwidth improves by 94% for PCI-Express

VAPI Bi-Directional Bandwidth (Two Ports)

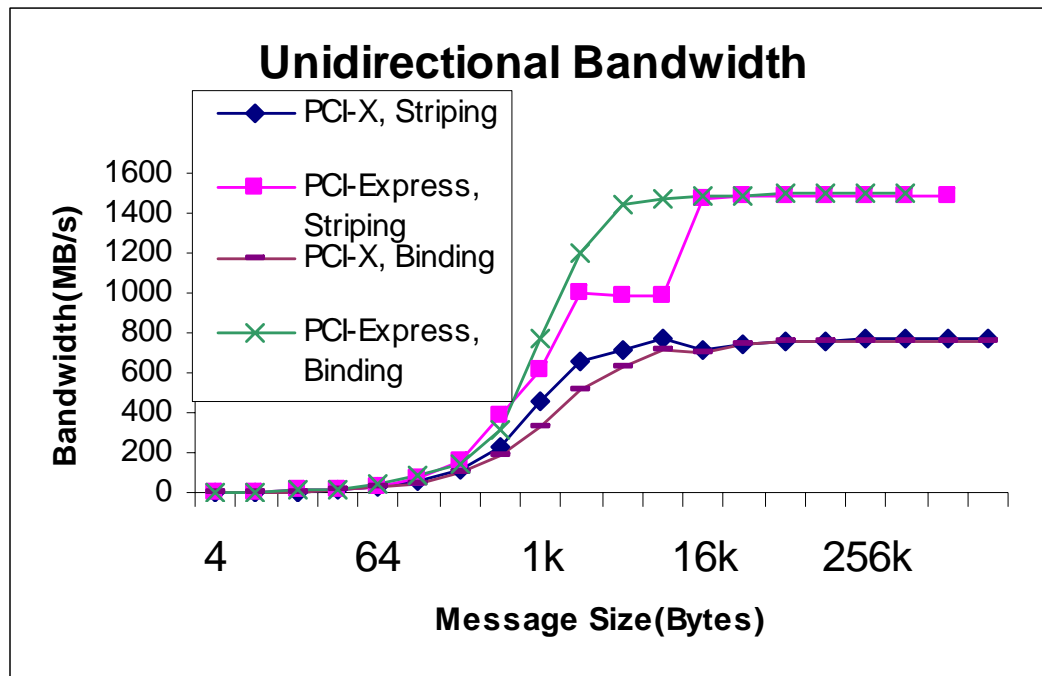


Policy	PCI-X	PCI-Express
Striping	946 MB/s	2483 MB/s
Binding	945 MB/s	1932 MB/s

- The VAPI level Uni-Directional Bandwidth improves by 189% for PCI-Express (Striping)



VAPI Uni-Directional Bandwidth (Two processes per Node with Two Ports)

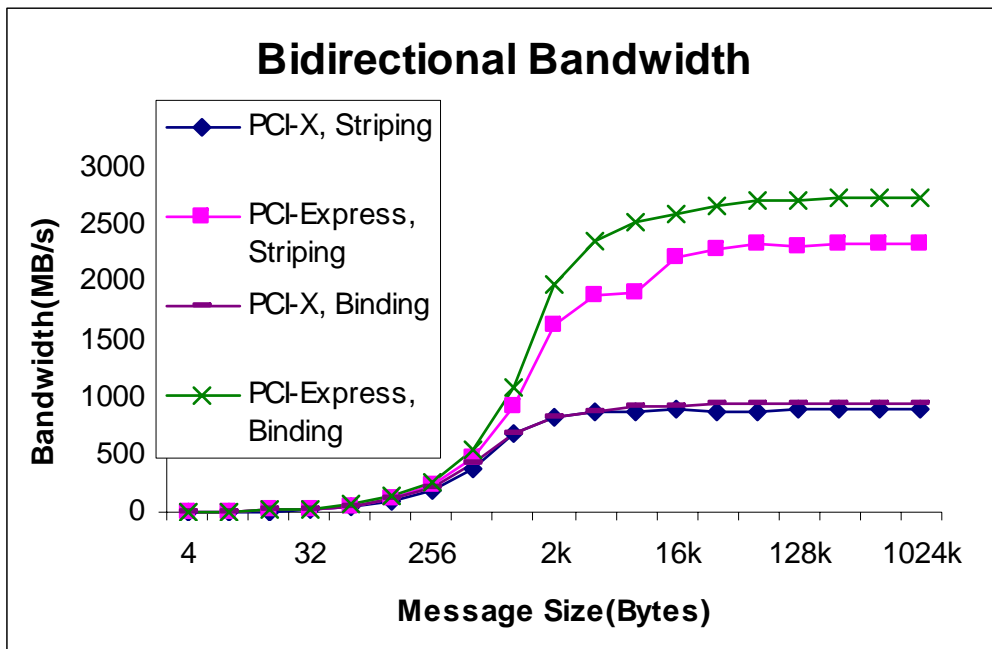


Policy	PCI-X	PCI-Express
Striping	761 MB/s	1486 MB/s
Binding	766 MB/s	1500 MB/s

- The VAPI level Uni-Directional Bandwidth improves by 96% for PCI-Express



VAPI Bi-Directional Bandwidth (Two processes per Node with Two Ports)



Policy	PCI-X	PCI-Express
Striping	945 MB/s	2457 MB/s
Binding	946 MB/s	2727 MB/s

- The VAPI level Bi-Directional Bandwidth improves by 189% for PCI-Express (Binding)



MPI Level Tests



- Single port tests using MVAPICH 0.9.4
 - Will be released this week
- Two port tests using a preliminary version of multi-rail MVAPICH
 - To be released with MVAPICH 0.9.5 in 3-4 weeks



MVAPICH Software Distribution

- Based on MPICH and MVICH
- Open Source (current version is 0.9.2)
- Have been directly downloaded by more than 115 organizations and industry
- Available in the software stack distributions of IBA vendors

National Labs/Research Centers

Argonne National Laboratory
Cornell Theory Center
Center for Mathematics and Computer Science
(The Netherlands)
Inst. for Experimental Physics (Germany)
Inst. for Program Structures and Data Organization
(Germany)
Lawrence Berkeley National Laboratory
Los Alamos National Laboratory
Max Planck Institute for Astronomy (Germany)
NASA Ames Research Center
NCSA
National Center for Atmospheric Research
Ohio Supercomputer Center
Pacific Northwest National Laboratory
Pittsburgh Supercomputing Center
Research & Development Institute Kvant (Russia)
Science Applications International Corporation
Sandia National Laboratory

Universities

Georgia Tech
Indiana University
Korea Univ. (Korea)
Korea Inst. Of Science and Tech. (Korea)
Kyushu Univ. (Japan)
Mississippi State University
Moscow State University (Russia)
Northeastern University
Penn State University
Russian Academy of Sciences (Russia)
Stanford University
Technion (Israel)
Technical Univ. of Munchen (Germany)
Technical Univ. of Chemnitz (Germany)
Univ. of Geneva (Switzerland)
Univ. of Houston
Univ. of Karlsruhe (Germany)
Univ. of Massachusetts Lowell
Univ. of Paderborn (Germany)
Univ. of Potsdam (Germany)
Univ. of Rio Grande (Brazil)
Univ. of Sherbrooke (Canada)
Univ. of Stuttgart (Germany)
Univ. of Toronto (Canada)

MVAPICH Users (Cont'd)

Industry

Abba Technology
Advanced Clustering Tech.
AMD
Ammasso
Appro
Array Systems Comp. (Canada)
Atipa Technologies
Agilent Technologies
Clustars Supercomputing-
Technology Inc. (China)
Clustervision (Netherlands)
Compusys (UK)
CSS Laboratories, Inc.
Dell
Delta Computer (Germany)
Emplics (Germany)
Fluent Inc.
ExaNet (Israel)
GraphStream, Inc.
HP
HP (France)

IBM
IBM (France)
IBM (Germany)
INTERSED (France)
InfiniCon
Intel
Intel (China)
Intel (Germany)
Intel Solution Services (Hong Kong)
Intel Solution Services (Japan)
JNI
Kraftway (Russia)
Langchao (China)
Linux Networx
Linvision (Netherlands)
Megaware (Germany)
Mercury Computer Systems
Mellanox Technologies
Meiosys (France)
Microway, Inc.
NEC (Japan)
NEC Solutions, Inc.
NEC (Singapore)
NICEVT (Russia)
OCF plc (United Kingdom)

OctigaBay (Canada)
PANTA Systems
ParTec (Germany)
PathScale, Inc.
Pultec (Japan)
Pyramid Computer (Germany)
Qlusters (Israel)
Raytheon Inc.
RLX Technologies
Rosta Ltd. (Russia)
SBC Technologies, Inc.
Scyld Software
SGI (Silicon Graphics, Inc.)
SKY Computers
Streamline Computing (UK)
Systran
Tomen
Telcordia Applied Research
Thales Underwater Systems (UK)
Transtec (Germany)
T-Platforms (Russia)
Topspin
Unisys
Voltaire
WorkstationsUK, Ltd. (UK)
Woven Systems, Inc.



Larger IBA Clusters using MVAPICH and Top500 Rankings



- 1105-node cluster at Virginia Tech
 - 3rd in Nov. '03 ranking
- 192-node cluster at Mississippi State University
 - 150th in June '04 ranking
- 128-node cluster at Sandia/Livermore
 - 111th in Nov '03 ranking and 211th in June '04 ranking
- 256-node cluster at Los Alamos
 - 116th in Nov '03 ranking and 218th in June '04 ranking
- 128-node cluster at Ohio Supercomputer Center (OSC)
 - 272th in June '04 ranking
- More are getting installed



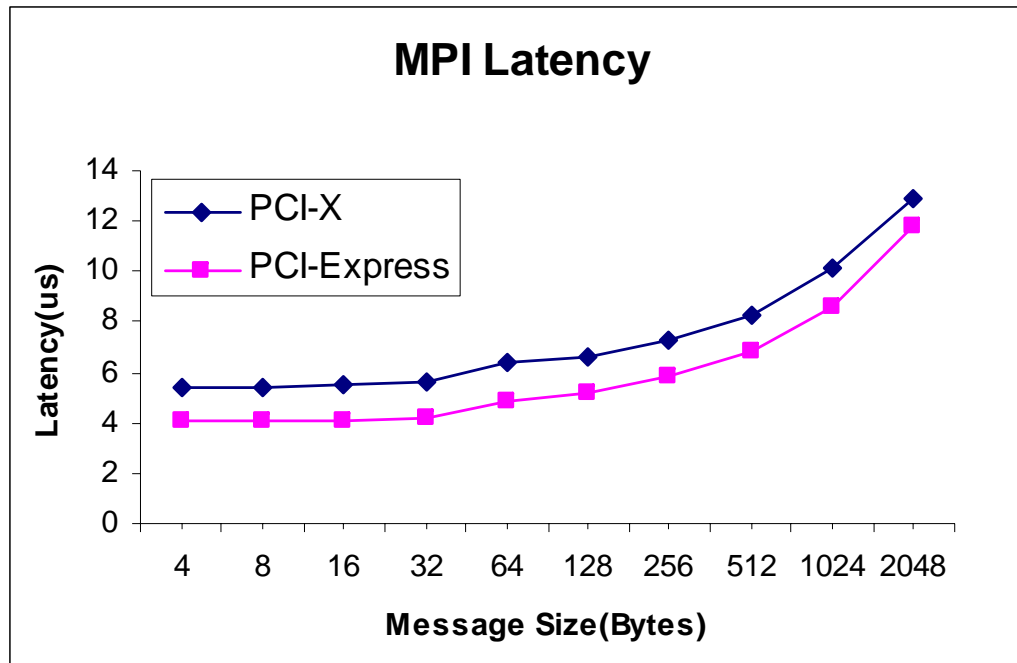


MPI-level Evaluation



- Micro-benchmarks
 - Latency
 - Uni-directional bandwidth (1-port and 2-ports)
 - Bi-directional bandwidth (1-port and 2-ports)
- Collective communication (using Pallas)
 - Broadcast, all-to-all, reduce, and all-reduce
 - Only 1-port implementation
- Applications
 - NAS benchmarks (IS and FT)
 - Only 1-port implementation

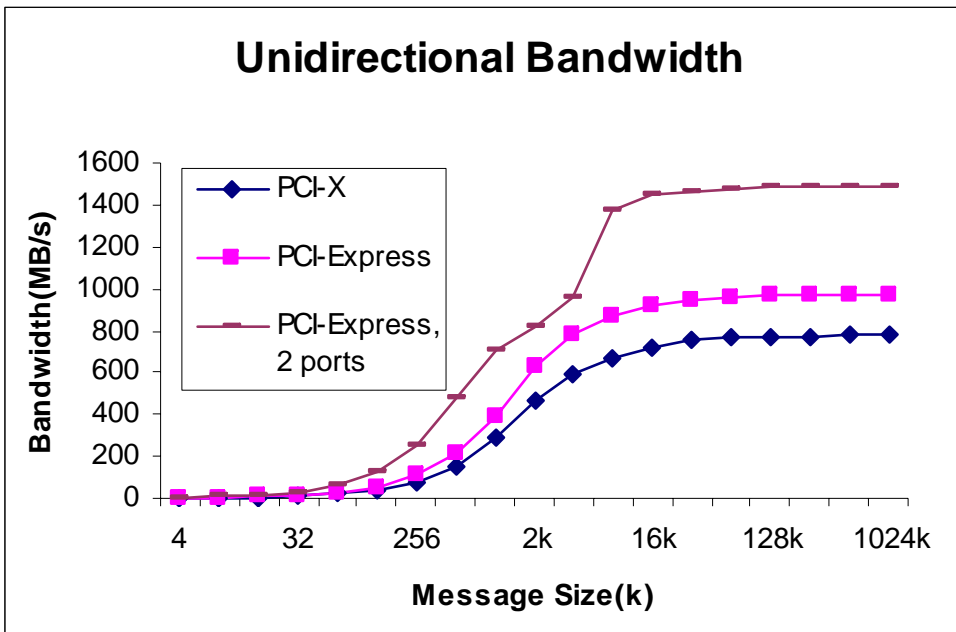
MPI Latency



PCI-X	PCI-Express
5.4us	4.0us

- The MPI latency improves by 23% for PCI-Express

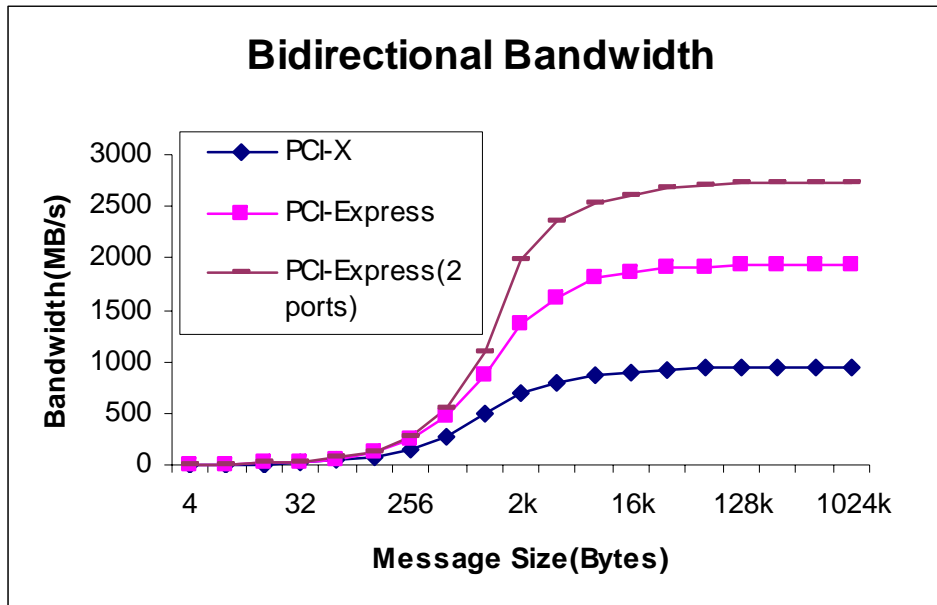
MPI Uni-Directional Bandwidth



PCI-X	PCI-Express (1 port)	PCI-Express (2 ports)
778MB/s	972 MB/s	1494MB/s

- The MPI level Uni-Directional Bandwidth improves by 93% for PCI-Express (2 ports)

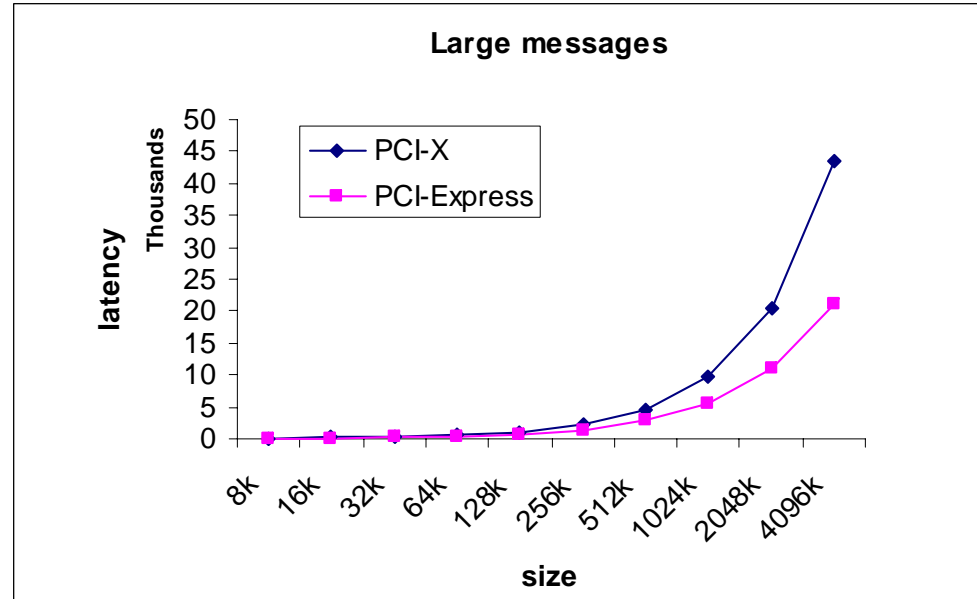
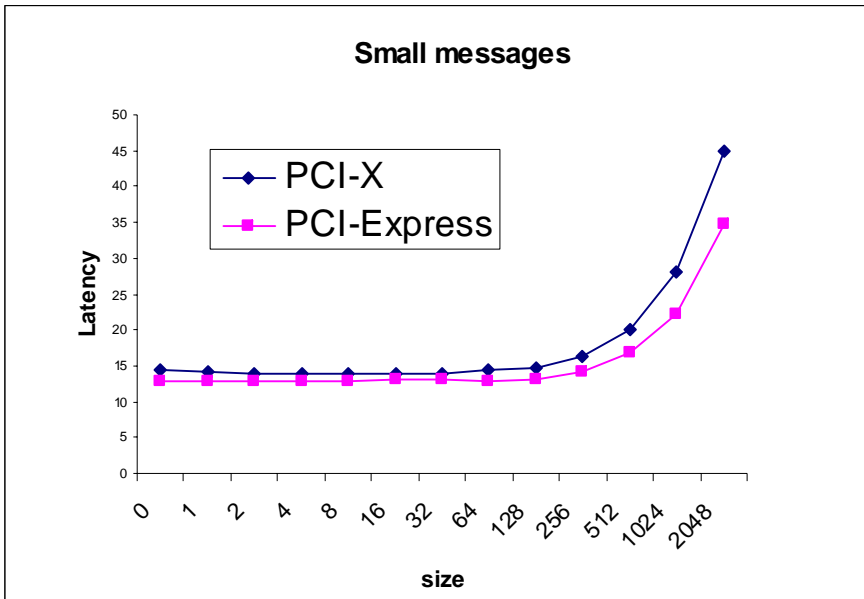
MPI Bi-Directional Bandwidth



PCI-X	PCI-Express (1 port)	PCI-Express (2 ports)
945MB/s	1942MB/s	2727MB/s

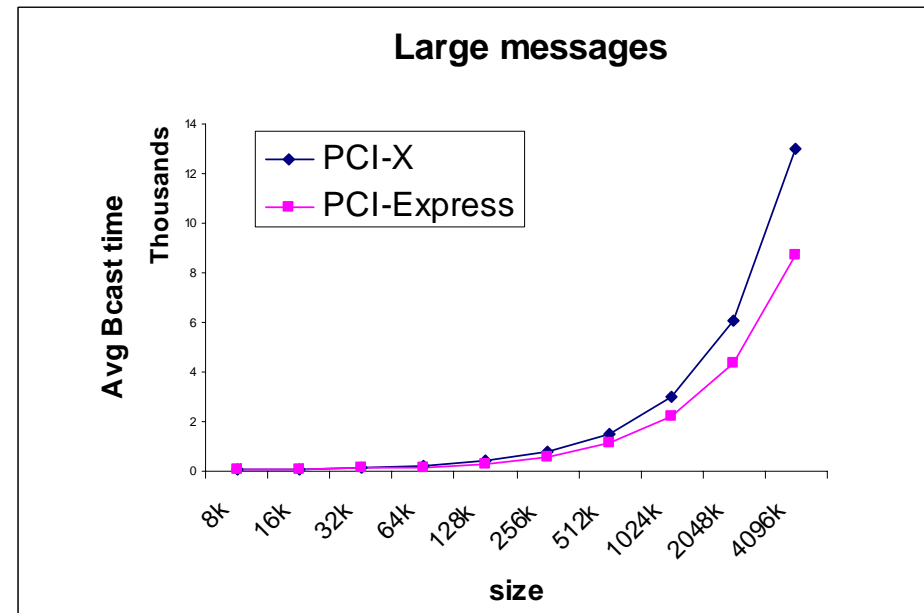
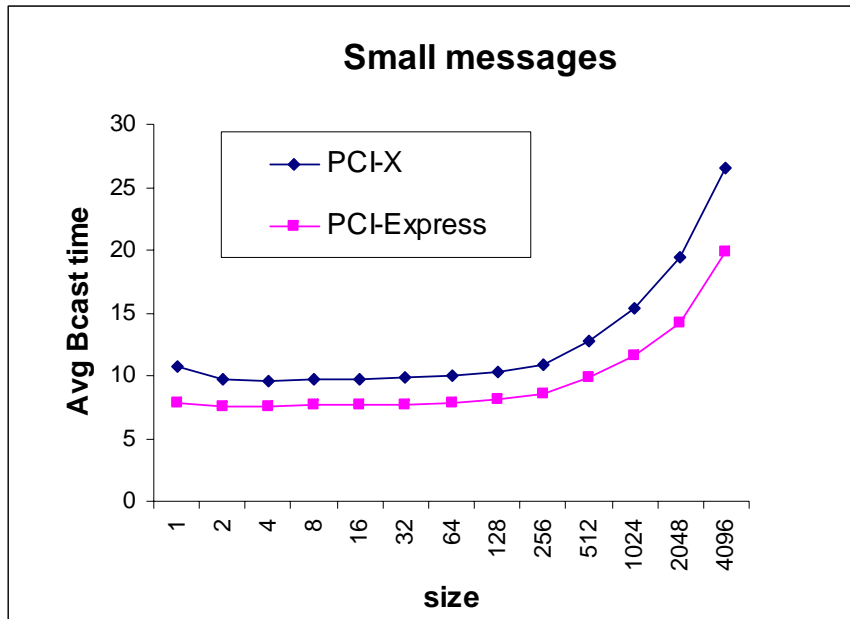
- The MPI level Bi-Directional Bandwidth improves by 188% for PCI-Express

All-to-All: Pallas (4 nodes)



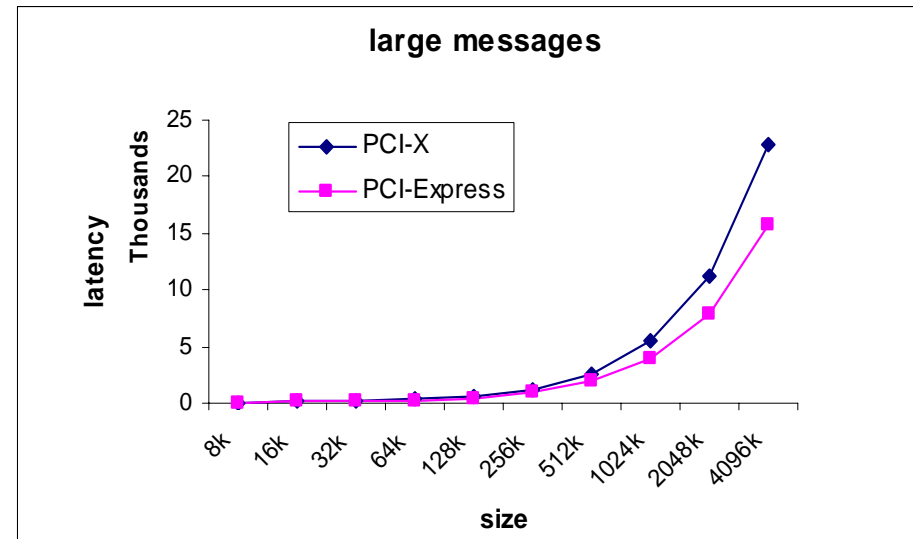
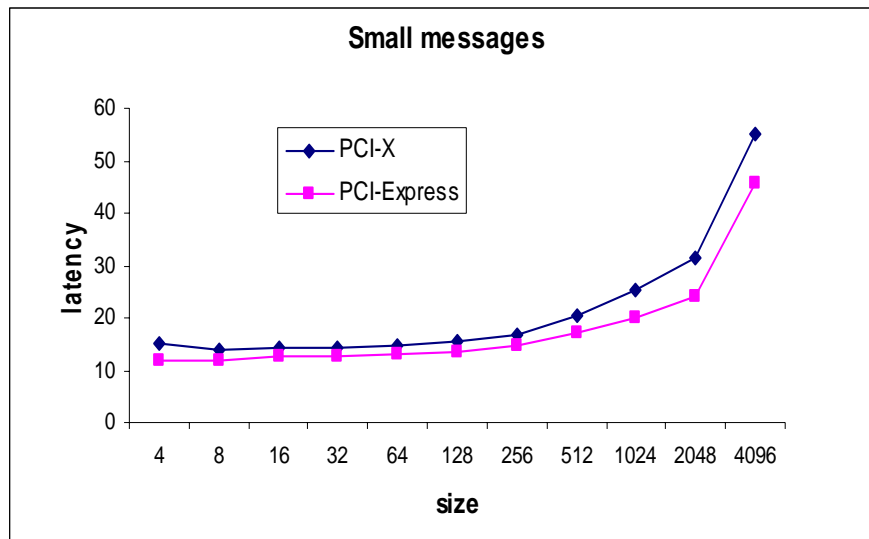
- The All-to-All latency improves by 2 times for PCI-Express for large messages

Bcast: Pallas (4 nodes)



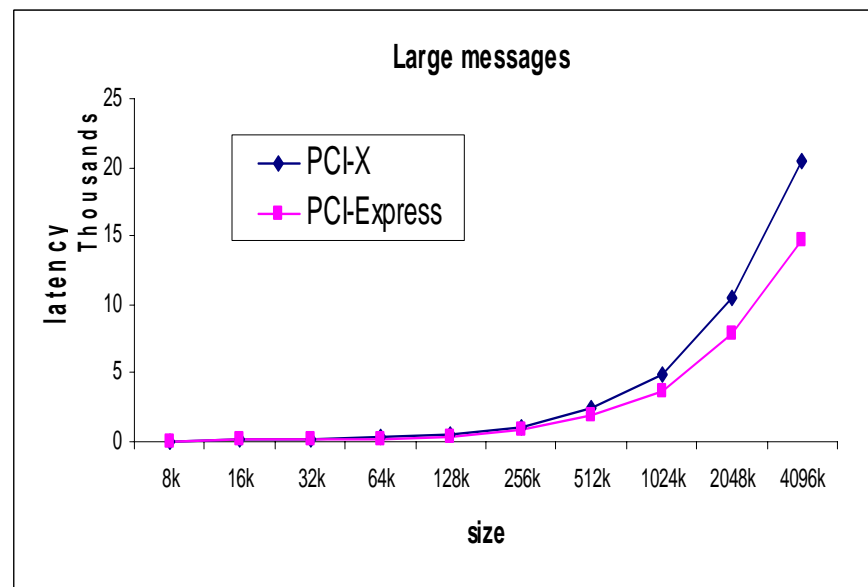
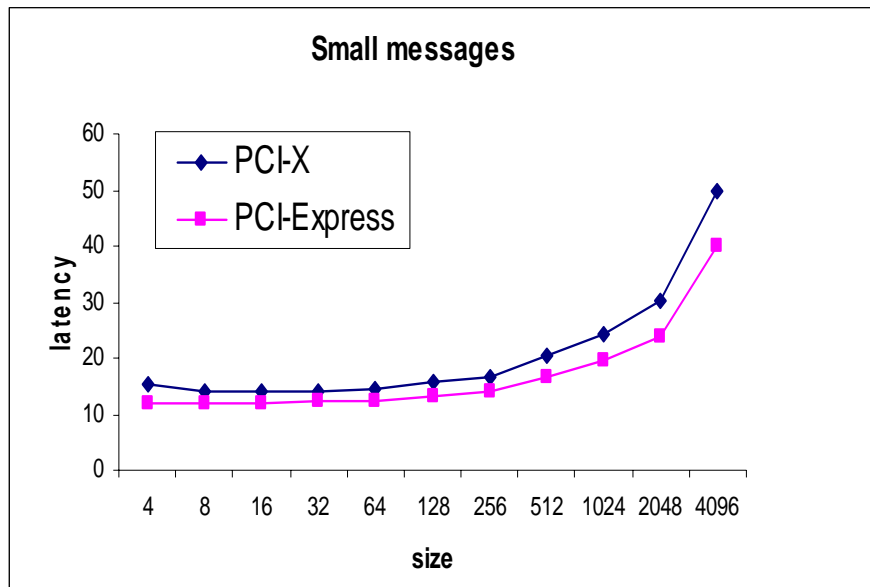
- The Average Bcast time improves by 1.5 times for PCI-Express for large messages

All-reduce: Pallas (4 nodes)



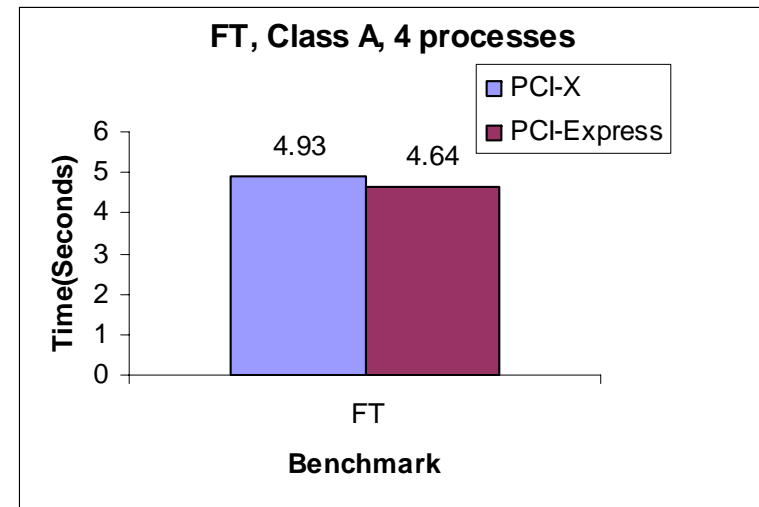
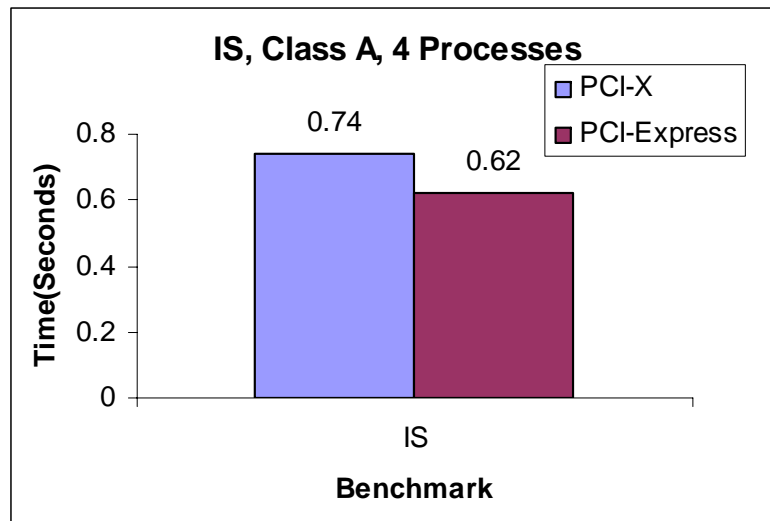
- The All-reduce latency improves by 1.45 times for PCI-Express for large messages

Reduce: Pallas (4 nodes)



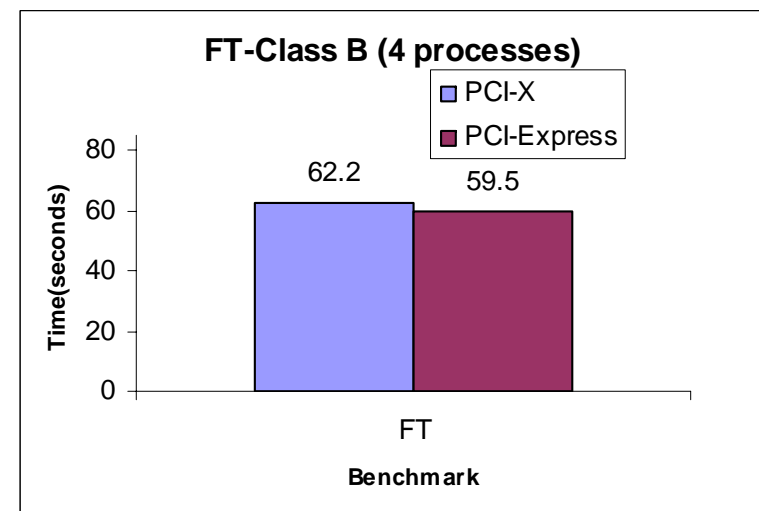
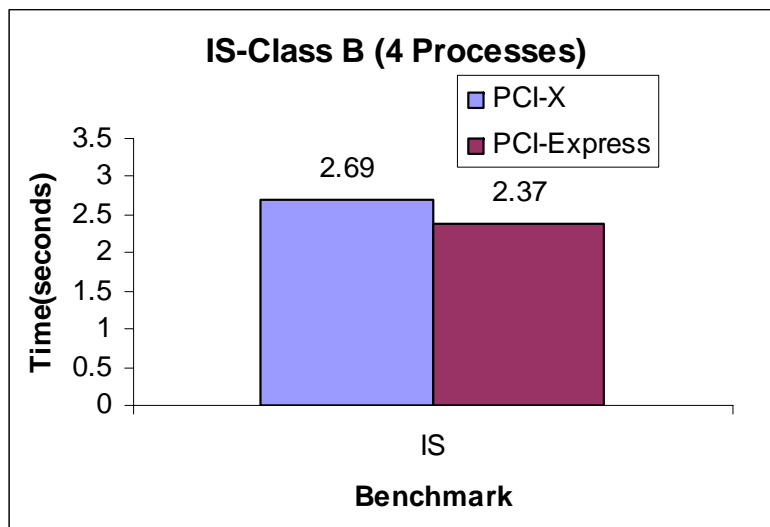
- The Reduce latency improves by 1.39 times for PCI-Express for large messages

NAS: 4 Nodes (4x1), Class A



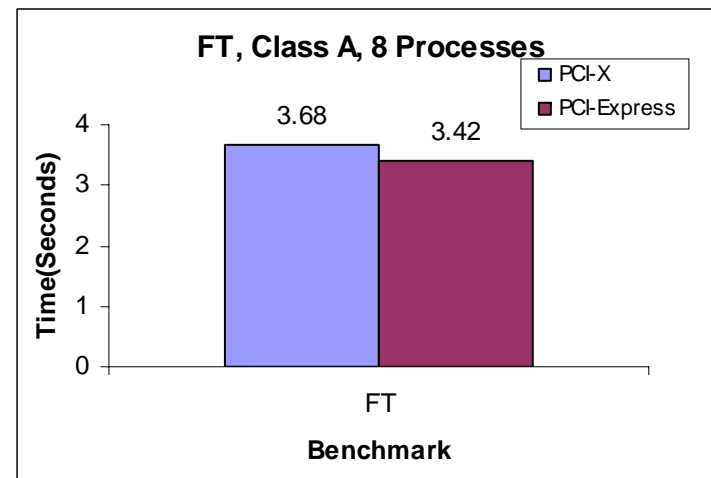
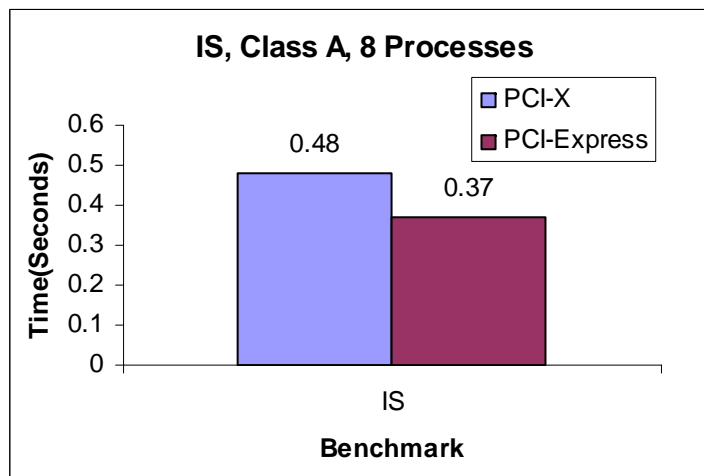
- The Execution time improves by 16% for IS and 7% for FT

NAS: 4 Nodes (4x1), Class B



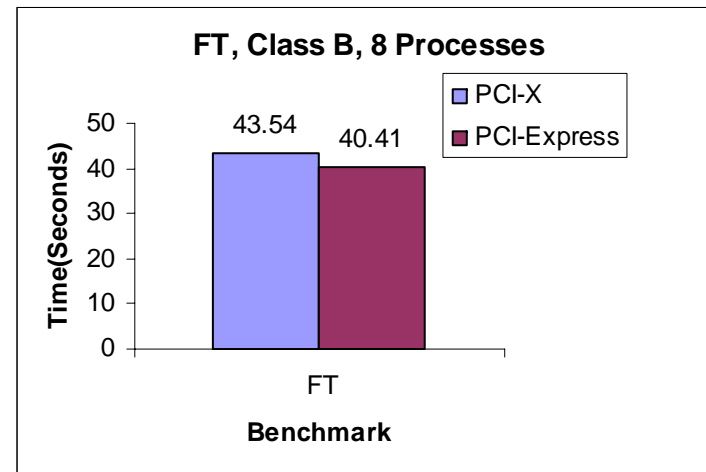
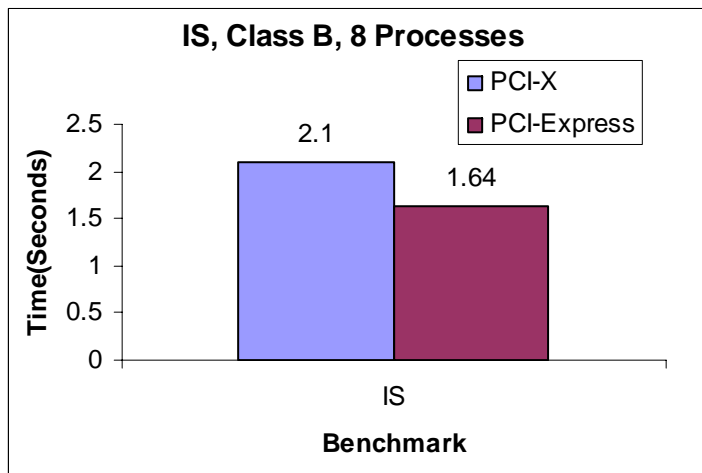
- The Execution time improves by 12% for IS and 5% for FT

NAS: 4 Nodes (4x2), Class A



- The Execution time improves by 24% for IS and 7% for FT

NAS: 4 Nodes (4x2), Class B





- The Execution time improves by 22% for IS and 8% for FT



Conclusions



- A comprehensive performance comparison of InfiniBand with PCI-X and PCI Express
 - Presented a set of micro-benchmarks at the VAPI- and MPI- level
 - Also MPI-level collective communication and applications
 - Results show that PCI Express can bring significant performance improvements
 - 24% reduction in small message latency
 - Up to 2.9 times factor of improvement in bandwidth
 - Also at the applications-level
- 
- 

•
•
•

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/>

An updated copy of the paper with the latest Performance numbers will be available on the web page soon.

• • • • • • • •