Designing Next Generation High Performance Clusters and Datacenters with InfiniBand

Dhabaleswar K. (DK) Panda
Department of Computer Science and Engineering
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

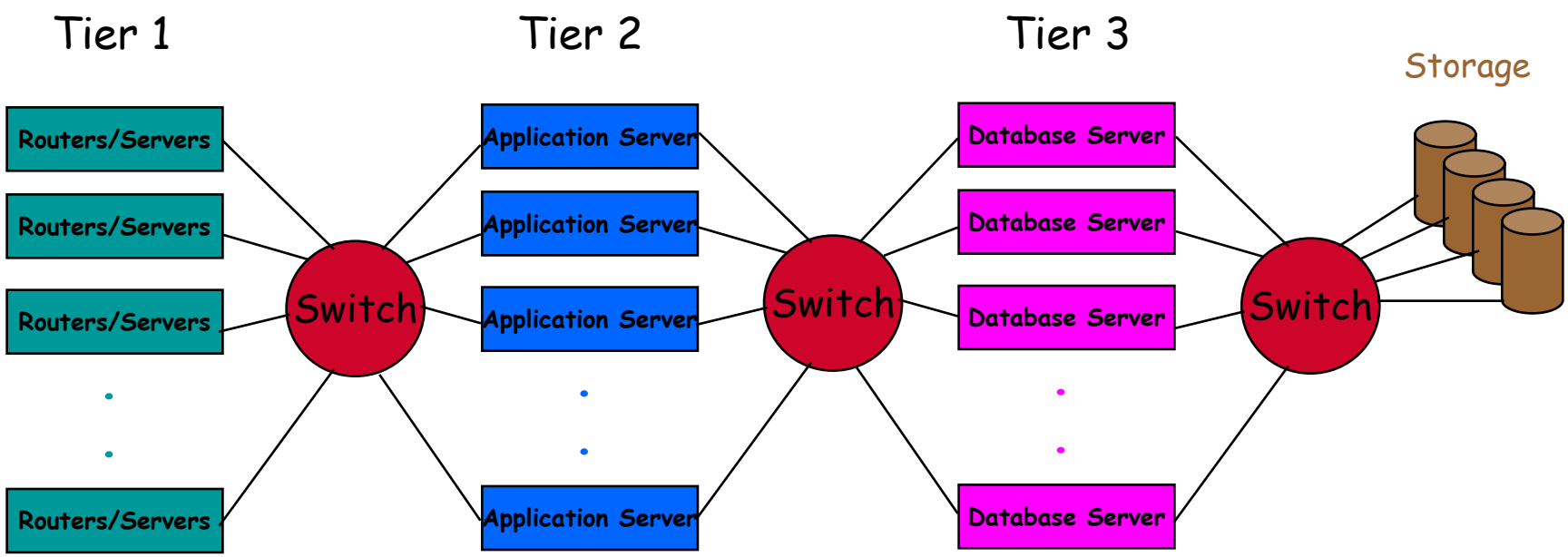
Presentation Overview

- Networking and I/O Requirements for HPC Clusters and Datacenters
- InfiniBand (IBA): Why, What, and Current State
- Designing High Performance MPI for IBA
  - MVAPICH and MVAPICH2
  - Latest performance numbers
  - Research challenges
- Designing Multi-Tier Datacenters with IBA, latest performance numbers, and challenges
- Conclusions

Trends for Clusters

- Top 500 list of Supercomputers (www.top500.org)
  - June 2001: 33/500 (6.6%)
  - Nov 2001: 43/500 (8.6%)
  - June 2002: 80/500 (16%)
  - Nov 2002: 93/500 (18.6%)
  - June 2003: 149/500 (29.8%)
  - Nov 2003: 208/500 (41.6%)
  - June 2004: 291/500 (58.2%)
  - Nov 2004: 294/500 (58.8%)

Increasing Use of Clusters for Multi-Tier Data Centers



- All major search engines and e-commerce companies are using clusters for multi-tier datacenters
  - Google, Amazon, Financial institutions, ...

Networking and I/O Requirements for HPC Clusters and Datacenters

- Good Systems Area Network with excellent performance (low latency and high bandwidth) for interprocessor communication (IPC) and I/O
- Good Storage Area Networks high performance I/O
- Good WAN connectivity in addition to intra-cluster SAN/LAN connectivity
- Quality of Service (QoS) for supporting interactive applications
- RAS (Reliability, Availability, and Serviceability)
- With low cost

Presentation Overview

- Networking and I/O Requirements for HPC Clusters and Datacenters
- InfiniBand (IBA): Why, What, and Current State
- Designing High Performance MPI for IBA
  - MVAPICH and MVAPICH2
  - Latest performance numbers
  - Research challenges
- Designing Multi-Tier Datacenters with IBA, latest performance numbers, and challenges
- Conclusions

Why InfiniBand?

- Traditionally HPC clusters have used
  - Myrinet or Quadrics
  - proprietary interconnects
- Datacenters have used
  - Ethernet (Gigabit Ethernet is common)
  - 10.0 Gigabit Ethernet is not yet available with low cost
  - QoS and RAS capabilities are not there in Ethernet
- Storage and File Systems have used
  - Ethernet with IP
  - Fibre Channel
  - Does not support high bandwidth and QoS, etc.

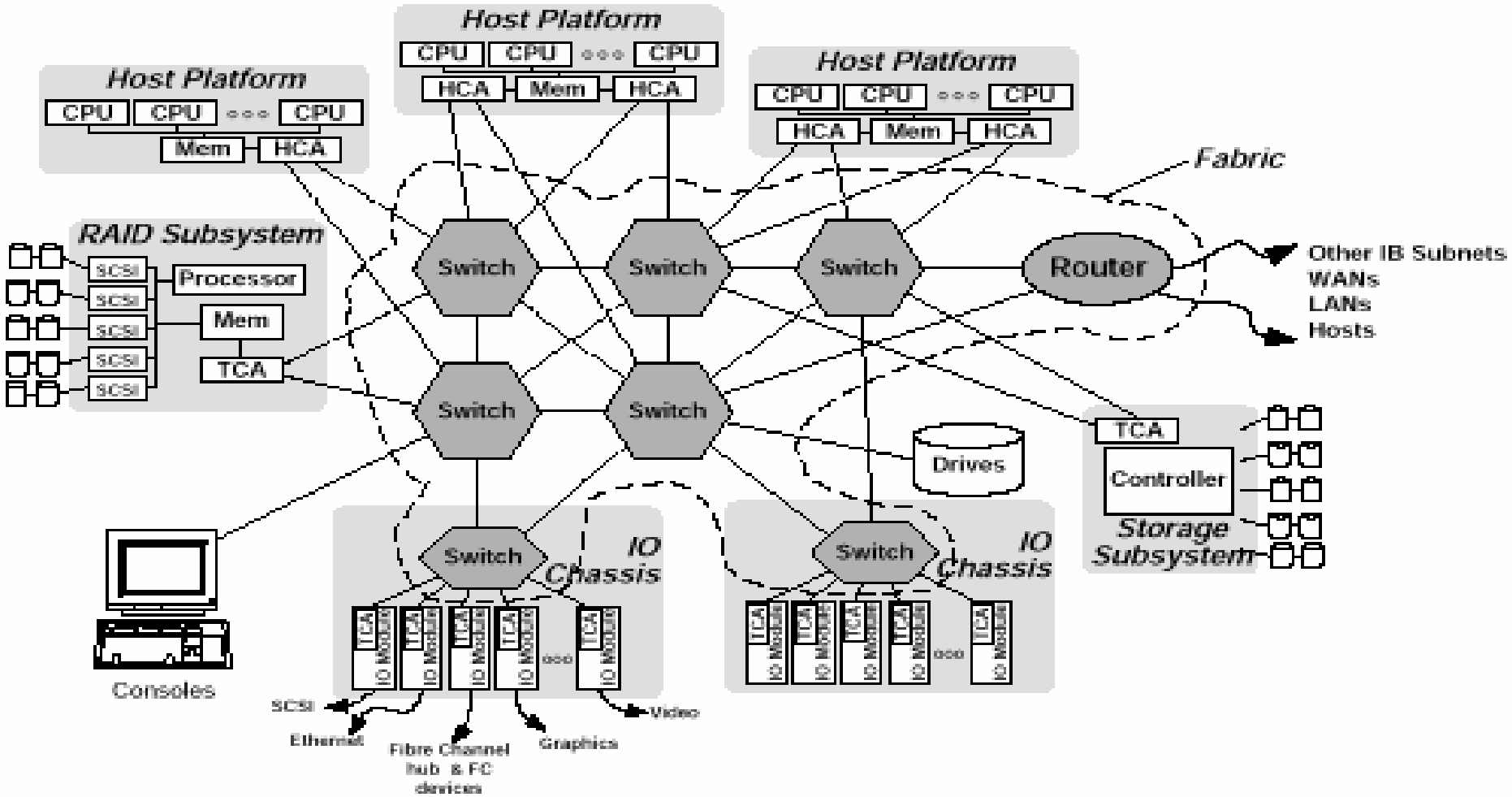
Need for A New Generation Computing, Communication & I/O Architecture

- Several initiatives were working to address high performance I/O issues
  - Next Generation I/O
  - Future I/O
- VIA consortium was working towards high performance interprocessor communication
- An attempt was made to consolidate all these efforts as an open standard
- The idea behind InfiniBand Architecture (IBA) was conceived ...
- The original target was for Data Centers

IBA Trade Organization

- IBA Trade Organization was formed with seven industry leaders (Compaq, Dell, HP, IBM, Intel, Microsoft, and Sun)
- Goal: To design a scalable and high performance communication and I/O architecture by taking an integrated view of computing, networking, and storage technologies
- Many other industry participated in the effort to define the IBA architecture specification
- InfiniBand Architecture (Volume 1, Version 1.0) was released to public on Oct 24, 2000
- www.infinibandta.org

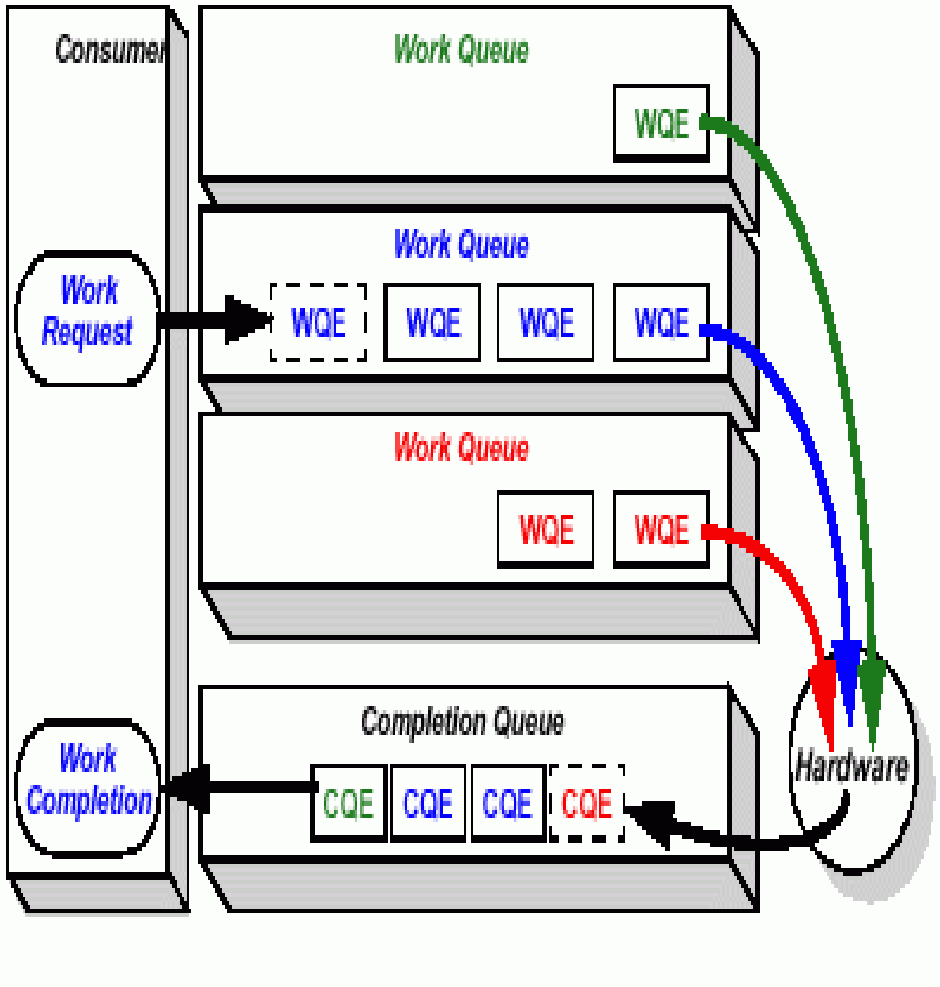
A Typical IBA SAN



Communication Operations

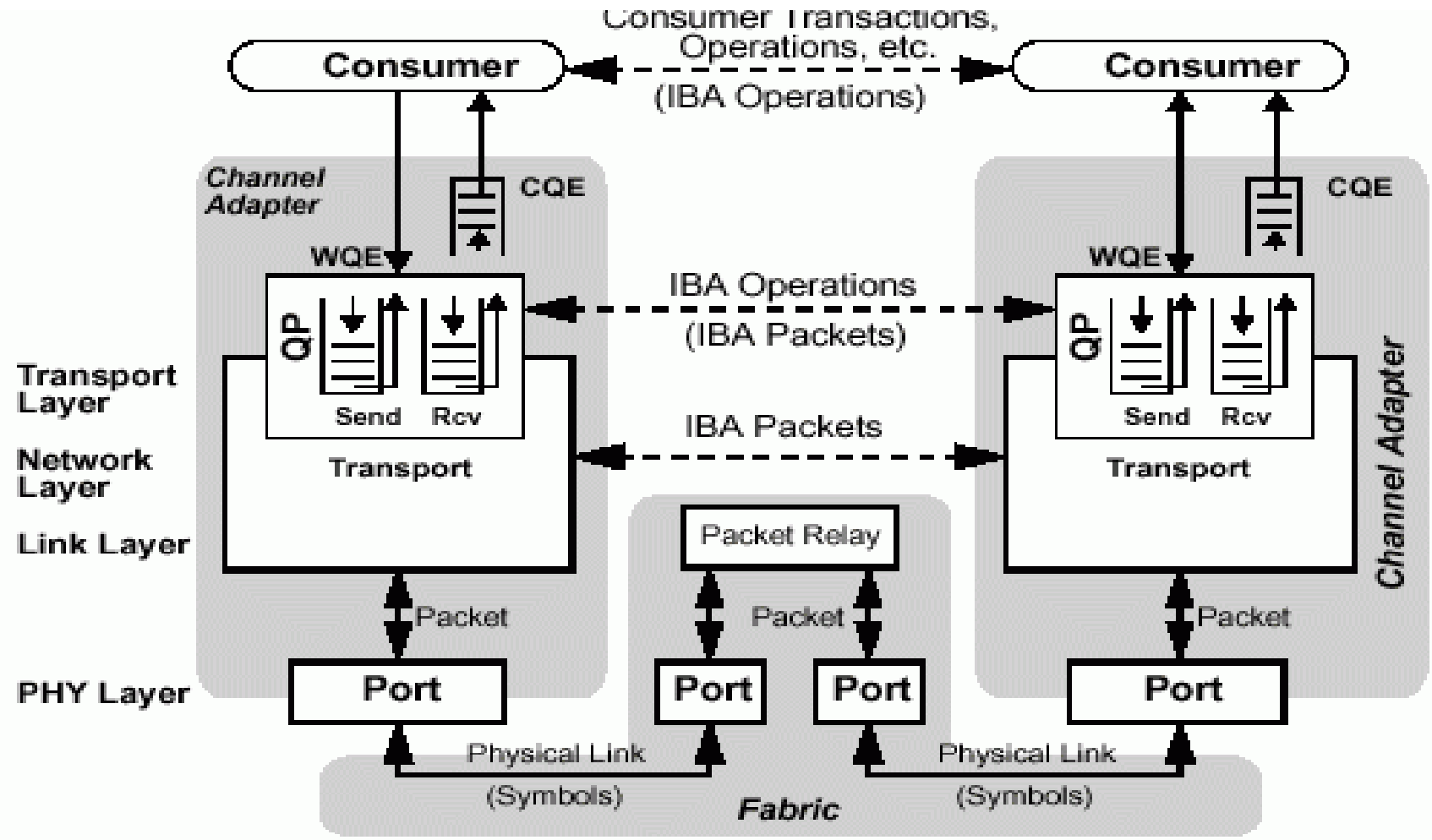
- Channel Semantic
  - Send and Recv
- Memory Semantic
  - RDMA read
  - RDMA write
  - RDMA atomic operations (e.g. Fetch & Add, Compare & swap)

Communication: Queuing Model on HCA

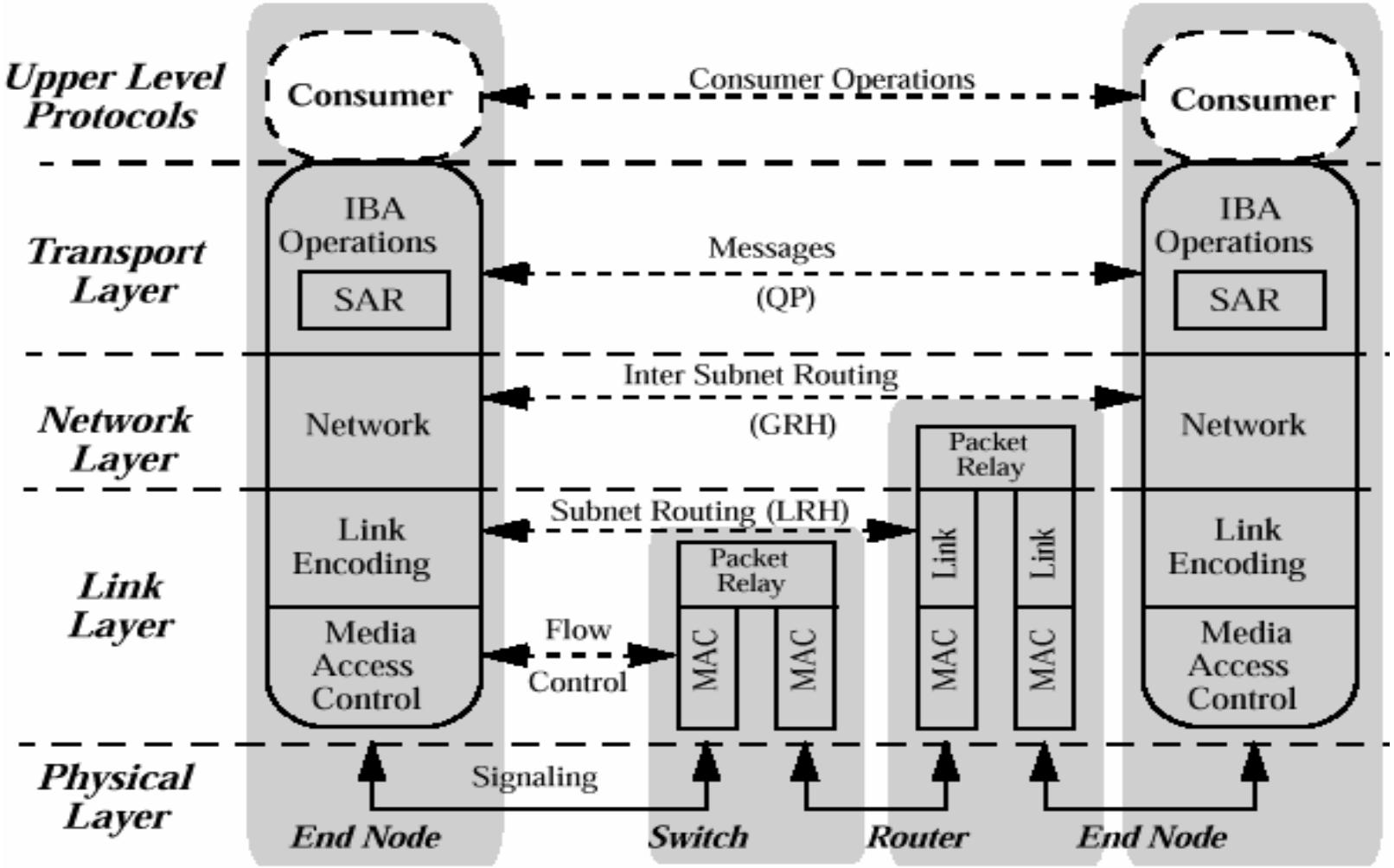


- Similar to VIA
- Work Request submitted => Work Queue Element (WQE) placed in a Work Queue
- HCA executes WQE
- A Completion Queue Entry (CQE) placed in associated CQ

Communications Stack



IBA Layered Architecture



Transport Services

Class of Service	Connected	Acknowledged	Transport
Reliable Connection	yes	Yes	IBA
Unreliable Connection	yes	no	IBA
Reliable Datagram	multiplexed	Yes	IBA
Unreliable Datagram	no	no	IBA
RAW Datagram	no	no	Raw

Reliability in Transport Services Types

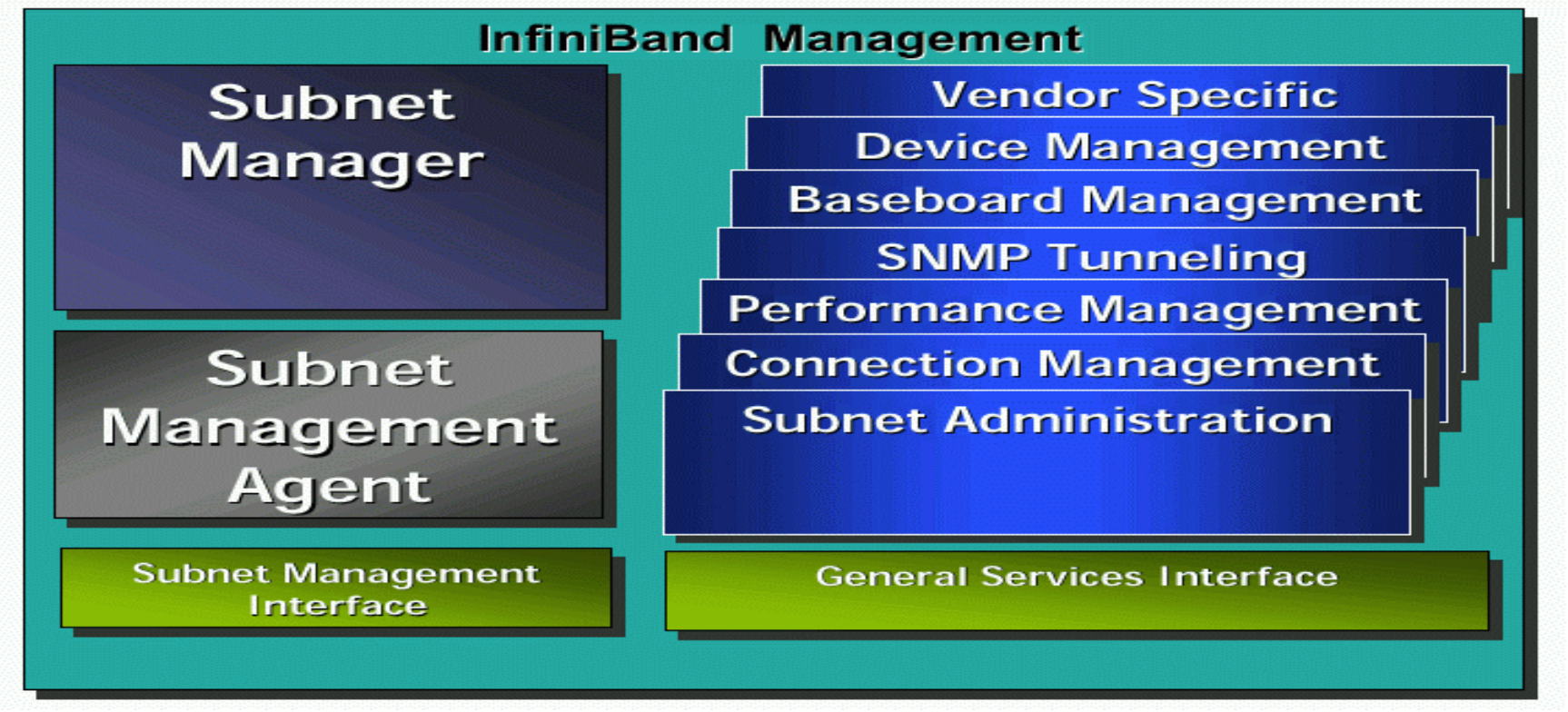
Comparison of IBA Transport Service Types

Attribute		Reliable Connection	Reliable Datagram	Unreliable Datagram	Unreliable Connection	Raw Datagram (both IPv6 & ethertype)
Scalability (M processes on N Processor nodes communicating with all processes on all nodes)		**M^2*N** QPs required on each processor node, per CA	**M** QPs required on each processor node, per CA.	**M** QPs required on each processor node, per CA.	**M^2*N** QPs required on each processor node, per CA.	**1** QP required on each end node, per CA.
Reliability	Corrupt data detected	Yes				
	Data delivery guarantee	Data delivered exactly once		No guarantees		
	Data order guaranteed	Yes, per connection	Yes, packets from any one source QP are ordered to multiple destination QPs.	No	Unordered and duplicate packets are detected.	No
	Data loss detected	Yes		No	Yes	No
	Error recovery	**Reliable**. Errors are detected at both the requestor and the responder. The requestor can transparently recover from errors (retransmission, alternate path, etc.) without any involvement of the client application. QP processing is halted only if the destination is inoperable or all fabric paths between the channel adapters have failed.		**Unreliable**. Packets with some types of errors may not be delivered. Neither source nor destination QPs are informed of dropped packets.	**Unreliable**. Packets with errors, including sequence errors, are detected and may be logged by the responder. The requestor is not informed.	**Unreliable**. Packets with errors are not delivered. The requestor and responder are not informed of dropped packets.

Rich Set of QoS and RAS Features

- Range of Network Features and QoS Mechanisms
  - Service Levels (priorities)
  - Virtual lanes
  - Partitioning
  - Multicast
    - allows to design a new generation of scalable communication and I/O subsystem with QoS
- Protected Operations
  - Keys
  - Protection Domains
- Flexibility for supporting Reliability, Availability, and Serviceability (RAS)
  - Multiple CRC fields
    - error detection (per-hop, end-to-end)
  - Fail-over
    - unmanaged and managed
  - Path Migration
  - Built-in Management Services

Management Model



- QP0 provides Subnet Management Interface (services)
- QP1 provides General Services Interface (services)
- QP0 and QP1 are special QPs on each port

Available IBA Products - Hardware

- Adapters:
  - Dual port 4X (10.0 Gbps) with PCI-X 133 MHz interface
  - Dual-port 4X (20.0 Gbps) with 8X PCI-Express interface
- Switches:
  - Various sizes (8, 16, 24, 64, 96, 144, and 288 4X ports) using Mellanox chipset + value added services/features
  - Larger systems can be built with a set of these switches as building blocks
  - Some switch ports can support 12X (30 Gbps)
- Switch Routers with Gateways
  - IB-to-FC
  - IB-to-IP
- Customized adapters to work with IBA switches
  - Cray XD1 (formerly by Octigabay)
  - PathScale

Available IBA Products – Hardware (Cont'd)

- Single-Chip Adapter
  - Uses "MemFree" technology
  - No memory in HCA for storing information about connections each node has with other nodes
  - Uses main memory of the system, especially with PCI-Express
  - Suitable for `Landed on Motherboard (LOM)' design
  - Targets sub $100 adapters
  - More details on Mellanox Press Release (Sept. 07, 2004)
- DDR Technology
  - IBTA has released the specs for
    - DDR (Double Data Rate)
    - QDR (Quad Data Rate)
  - Mellanox demonstrated IBA adapters and switches with DDR at SC '04
    - 4X: 10.0 Gbps -> 20.0 Gbps
    - 12X: 30.0 Gbps -> 60.0 Gbps
  - More details on Mellanox Press Release (Nov. 02, 2004)

Available Products - Software Stack

- Lower-level
  - VAPI (Verbs-Level API) from Mellanox
  - Modified and customized VAPI from other vendors
  - IBAL (IB Access layer) with UVP (User Verb Provider) from Mellanox and Intel
    - open source (http://infiniband.sourceforge.net)
  - A new initiative of OpenIB for Linux+IBA community
    - http://www.openib.org

Available Products - Software Stack (Cont'd)

- Upper-level Software stack
  - MPI (Message Passing Interface)
  - SDP (Sockets Direct Protocol)
  - IPoIB (IP over IB)
  - SRP (SCSI RDMA Protocol)
  - uDAPL (user-level Direct Access Provider Library)
  - kDAPL (kernel-level Direct Access Provider Library)

Available Products – Subnet Manager

- Subnet Manager
  - From different vendors with different functionalities and features
    - MiniSM (original) from Mellanox
    - VFM (Voltaire Fabric Manager) from Voltaire
    - Fabric Manager from Topspin
    - InfiniView Fabric Manager from InfiniCon
    - OpenSM from Mellanox and Intel
      - Open-source initiative (http://infiniband.sourceforge.net)

Openib.org

- A new open source organization
- Focusing on effort for Open Source IB support for Linux community
- Design of complete software stack with `best of breed' components
  - Gen1
  - Gen2
- Users should be able to download the entire stack and run without any problem

IBA Installations

- Several small-scale clusters (up to 64 nodes) have already been installed at various places all over the world
  - North America, Europe
  - Asia, South America
- Larger cluster installations and TOP 500 rankings (Nov. '04)
  - 10,1600-processors cluster at NASA-Columbia (2nd)
  - 2200-processors at Virginia Tech (7th)
  - 576-processors at United Institute of Informatics Problems, Belarus (98th)
  - 936-processors at Univ. of Sherbrooke, Canada (127th)
  - 576-processors at Univ. of Sanghai (209th)
  - 384-processors at Mississippi State University (211th)
  - 256-processors at Sandia/Livermore (334th)
  - 512-processors at Los Alamos (346th)
  - 256-processors at Ohio Supercomputer Center (456th)
  - 288-processors at AMD Developer Center (461st)
  - 192-processors at Sandia (479th)
  - 2048-node cluster in Japan (RIKEN) (14th), partially InfiniBand
  - More are getting installed ….

Presentation Overview

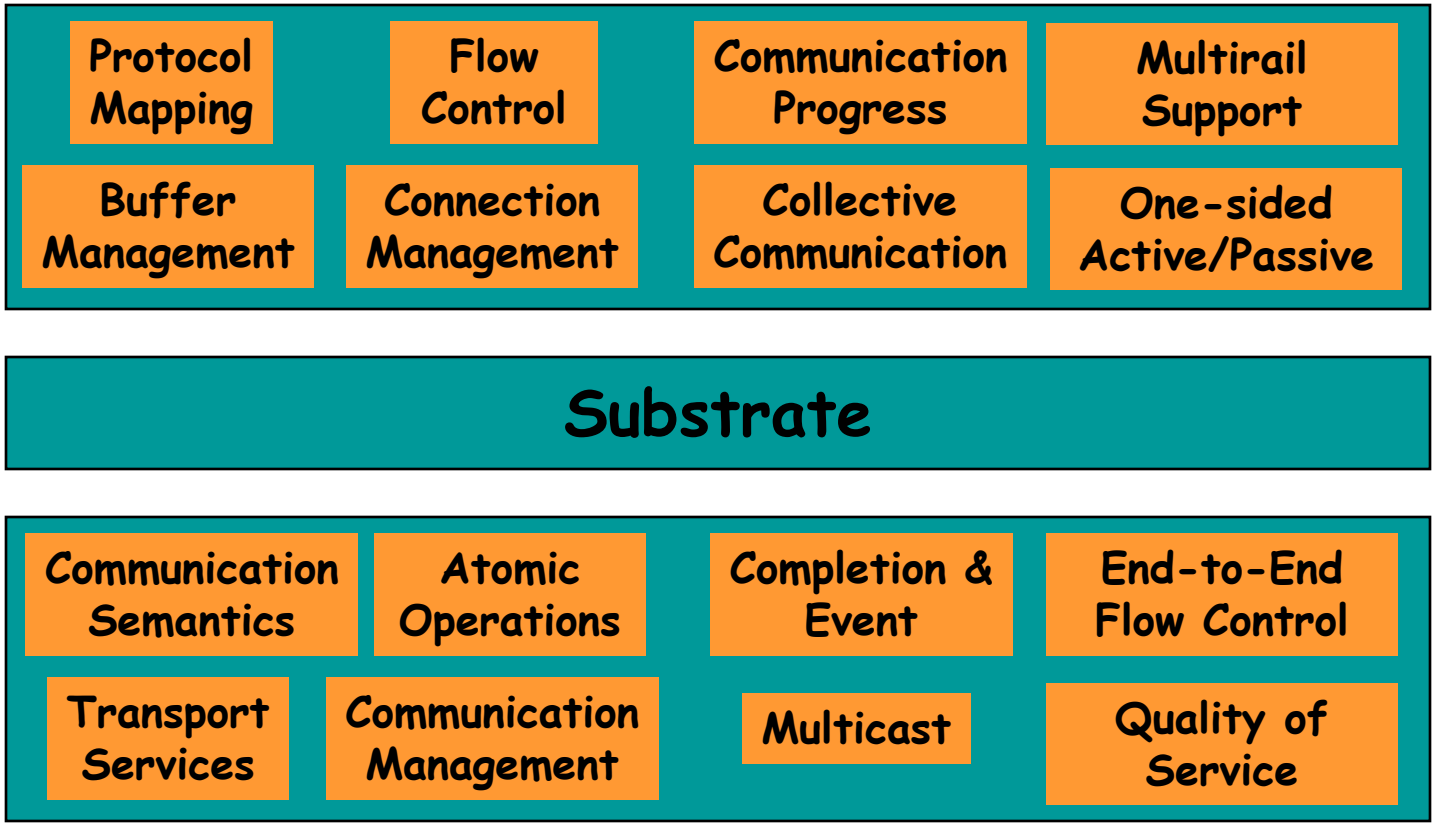
- Networking and I/O Requirements for HPC Clusters and Datacenters
- InfiniBand (IBA): Why, What, and Current State
- Designing High Performance MPI for IBA
  - MVAPICH and MVAPICH2
  - Latest performance numbers
  - Research challenges
- Designing Multi-Tier Datacenters with IBA, latest performance numbers, and challenges
- Conclusions

Designing High-End Systems with IBA – OSU Research Focus

- Clusters for High Performance Computing (Message Passing)
- Multi-tier Data Centers
- Cluster-based File Systems
- Storage Systems
- Distributed Shared Memory
  - Shared memory with coherency
  - Logical shared memory (ARMCI)

Designing MPI Using InfiniBand Features

MPI Design Components



InfiniBand Features

High Performance MPI Implementation with IBA – Research Agenda at OSU

- Point-to-point communication
  - RDMA-based design for both small and large messages
- Collective communication
  - Taking advantage of IBA hardware multicast for Broadcast
  - RDMA-based designs for barrier, all-to-all
- Flow control
  - Static vs. dynamic
- Connection Management
  - Static vs. dynamic
- Multi-rail designs
  - Multiple ports/HCAs
  - Different schemes (striping, binding, adaptive)

MPI Implementation with IBA – Research Agenda at OSU (Cont'd)

- Focusing both on MPI-1 and MPI-2
- Implementation on VAPI and IBAL (OpenIB)
- Optimization on IA-32, IA-64, Opteron, EM64T, and Apple G5 architectures
- Efficient Implementation of MPI Datatype Communication
- Performance Evaluation with other interconnects
- Applications-level performance evaluation

MVAPICH and MVAPICH2

- MVAPICH – MPICH over VAPI
  - Currently focuses on
    - point-to-point communication
    - collective communication
    - multi-rail support
- MVAPICH2 – MPICH2 over VAPI
  - Currently focuses on
    - point-to-point communication
    - one-sided communication
- Available for different architectures
  - EM64T, G5, IA-32, IA-64, and Opteron
  - PCI-X and PCI-Express
- More details at
  http://nowlab.cis.ohio-state.edu/projects/mpi-iba/

MVAPICH/MVAPICH2 Software Distribution

- Open Source (current versions are MVAPICH 0.9.4 and MVAPICH2 0.6.0)
- Have been directly downloaded by more than 160 organizations and industry
- Available in the software stack distributions of IBA vendors

National Labs/Research Centers

Alabama Supercomputer Center
Argonne National Laboratory
AWI Polar and Marine Research Center (Germany)
Cornell Theory Center
Center for Mathematics and Computer Science (The Netherlands)
CEA (France)
CERN, European Organization for Nuclear Research (Switzerland)
CLC, Center for Large-Scale Computation Chinese University (Hong Kong)
ECMWF, European Center for Medium-Range Weather Forecasts (UK)
Fermi National Accelerator Laboratory
Fraunhofer-Inst. for High-Speed Dynamics (Germany)
Inst. for Experimental Physics (Germany)
Inst. for Program Structures and Data Organization (Germany)
IRSN (France)
Korea Institute of Science and Technology (Korea)
Lawrence Berkeley National Laboratory
Los Alamos National Laboratory
Max Planck Institute for Astronomy (Germany)
Max Planck Institute for Gravitational Physics (Germany)
NASA Ames Research Center
NCSA
National Center for High Performance Computing (Taiwan)
National Center for Atmospheric Research
Ohio Supercomputer Center
Open Computing Centre "Strela" (Russia)
Pacific Northwest National Laboratory
Pittsburgh Supercomputing Center
Research & Development Institute Kvant (Russia)
Science Applications International Corporation
Sandia National Laboratory
United Institute of Informatics Problems (Belarus)

MVAPICH/MVAPICH2 Users: Universities

Engineers School of Geneva (Switzerland)
Georgia Tech
Grdansk Univ. of Technology (Poland)
Indiana University
Indiana State University
Korea Univ. (Korea)
Kyushu Univ. (Japan)
Mississippi State University
Moscow State University (Russia)
Northeastern University
Nankai University (China)
Oregon State University
Penn State University
Russian Academy of Sciences (Russia)
Stanford University
Technion (Israel)
Technical Univ. of Clausthal (Germany)
Technical Univ. of Munchen (Germany)
Technical Univ. of Chemnitz (Germany)
Tsinghua Univ. (China)
Univ. of Arizona
Univ. of Berne (Switzerland)
Univ. of Florida, Gainesville
Univ. of Geneva (Switzerland)
Univ. of Houston
Univ. of Karlsruhe (Germany)
Univ. of Massachusetts Lowell
Univ. of Milan (Italy)
Univ. of Paderborn (Germany)
Univ. of Pisa (Italy)
Univ. of Politecnica of Valencia (Spain)
Univ. of Potsdam (Germany)
Univ. of Rio Grande (Brazil)
Univ. of Sherbrooke (Canada)
Univ. of Stuttgart (Germany)
Univ. of Toronto (Canada)
Univ. of Westminster (UK)
Virginia Tech
Wroclaw Univ. of Technology (Poland)

MVAPICH/MVAPICH2 Users: Industry

Abba Technology
Advanced Clustering Tech.
AMD
Ammasso
Apple Computer
Appro
Array Systems Comp. (Canada)
Ascensit (Italy)
Atipa Technologies
Agilent Technologies
BAE Systems
Bull S.A. (France)
Clustars Supercomputing-
Technology Inc. (China)
Cluster Technology Ltd. (Hong Kong)
Clustervision (Netherlands)
Compusys (UK)
CSS Laboratories, Inc.
Dell
Delta Computer Products (Germany)
Emplics (Germany)
ESI Group (France)
Exadron (Italy)
ExaNet (Israel)
Fluent Inc.
FMS-Computer and Komm. (Germany)
GraphStream, Inc.
HP
HP (Asia Pacific)
HP (France)
HP Solution Center (China)
High Performance Associates
IBM
IBM (France)
IBM (Germany)
INTERSED (France)
InfiniCon
Intel
Intel (China)
Intel (Germany)
Intel Solution Services (Hong Kong)
Intel Solution Services (Japan)
InTouch NV (The Netherlands)
Invertix Corporation
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
OptimaNumerics (UK)
PANTA Systems
ParTec (Germany)
PathScale, Inc.
Pultec (Japan)
Pyramid Computer (Germany)
Qlusters (Israel)
Quant-X GmbH (Austria)
Raytheon Inc.
RLX Technologies
Rosta Ltd. (Russia)
SBC Technologies, Inc.
Scyld Software
SGI (Silicon Graphics, Inc.)
Siliquent
Simulation Technologies
SKY Computers
SmallTree communications
Streamline Computing (UK)
SUN
Systran
Telcordia Applied Research
Telsima
Thales Underwater Systems (UK)
Tomen
Topspin
Totally Hip Technologies (Canada)
Transtec (Germany)
T-Platforms (Russia)
Transtec (Germany)
Unisys
Voltaire
WorkstationsUK, Ltd. (UK)
Woven Systems, Inc.

Larger IBA Clusters using MVAPICH and Top500 Rankings (Nov. '04)

- 7th: 1100-node dual Apple Xserve 2.3 GHz cluster at Virginia Tech
- 98th: 288-node dual Opteron 2.2 GHz cluster at United Institute of Informatics Problems (Belarus)
- 211th: 192-node dual Xeon 3.06 GHz cluster at Mississippi State University
- 334th: 128-node dual Xeon 3.06 GHz cluster at Sandia/Livermore
- 346th: 256-node dual Opteron 1.6 GHz cluster at Los Alamos
- 456th: 128-node dual Xeon 2.4 GHz cluster at Ohio Supercomputer Center
- 461st: 144-node dual Opteron 2.0 GHz cluster at AMD Developer Center
- 479th: 96-node dual Xeon 3.06 cluster at Sandia/Albuquerque
- More are getting installed ….

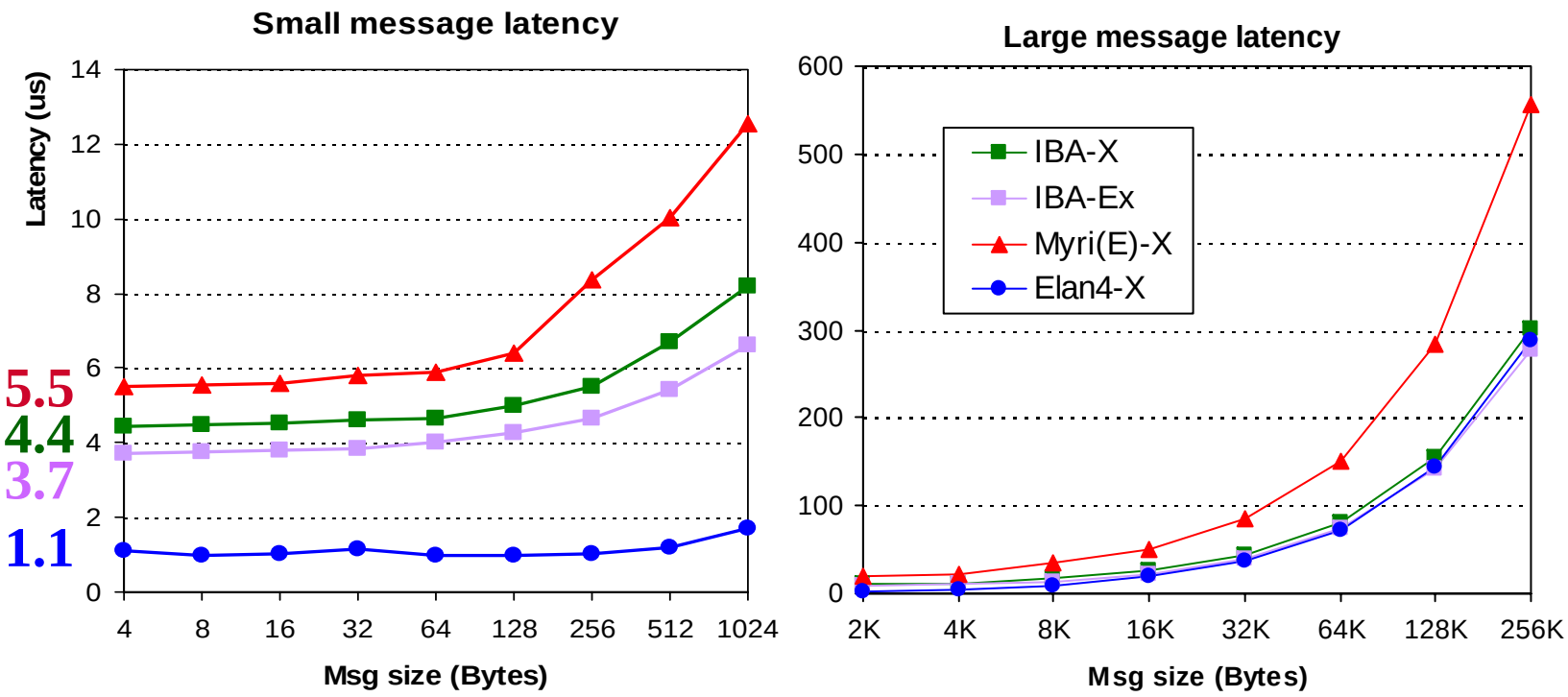
Latest Performance Results

- MVAPICH-1
  - Point-to-point communication
    - IBA (PCI-X and PCI-Express) vs. Myrinet vs. Quadrics/Elan4
  - Collective communication
    - Broadcast and all-to-all
  - Multi-rail systems
  - Support for MPI Datatype Communication
- MVAPICH2
  - Two-sided
  - One-sided
    - Active
    - Passive

IBA vs. Myrinet and Quadrics

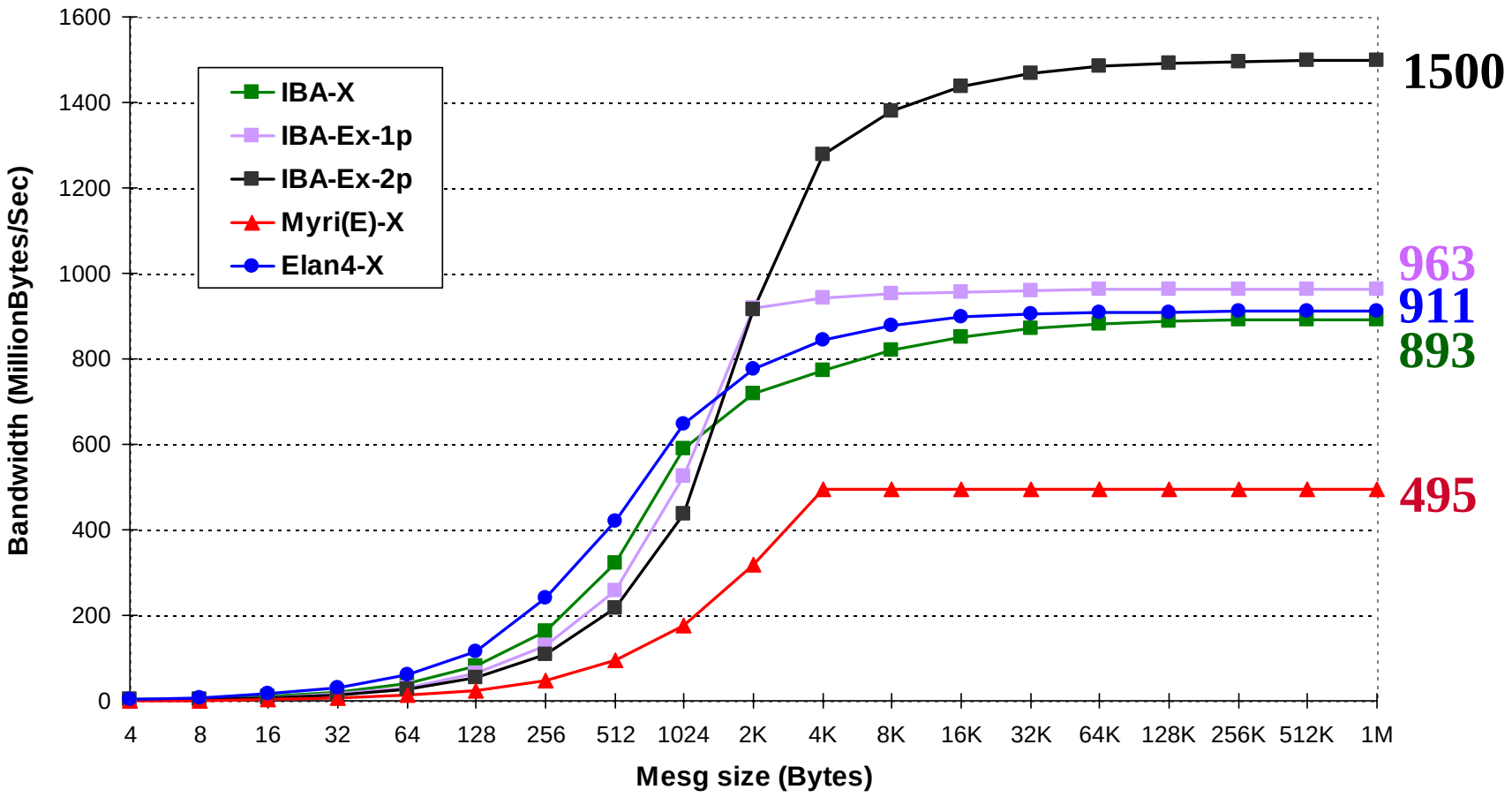
- System configurations and Interconnects
  - 3.4 GHz EM64T Xeon with 8X PCI-Express
    - IBA 4X (10.0 Gbps) adapters with PCI-Express interface
      - 1-port and 2-ports
  - 3.0 GHz Xeon with PCI-X 133 MHz interface
    - IBA 4X (10.0 Gbps) adapters with PCI-X
    - Myrinet LANai XP (E) cards with PCI-X
    - Quadrics Elan4 cards with PCI-X 133
  - All results are with respective switches being present
- Results
  - User-level communication layer
  - MPI-level
    - MVAPICH 0.9.4 for IBA (0.9.5 with dual-port support)
    - MPICH/GM for Myrinet
    - MPICH/QsNet for Quadrics

User-level Latency (One-way): IBA/VAPI vs. Myrinet/GM vs. Quadrics/Elan4

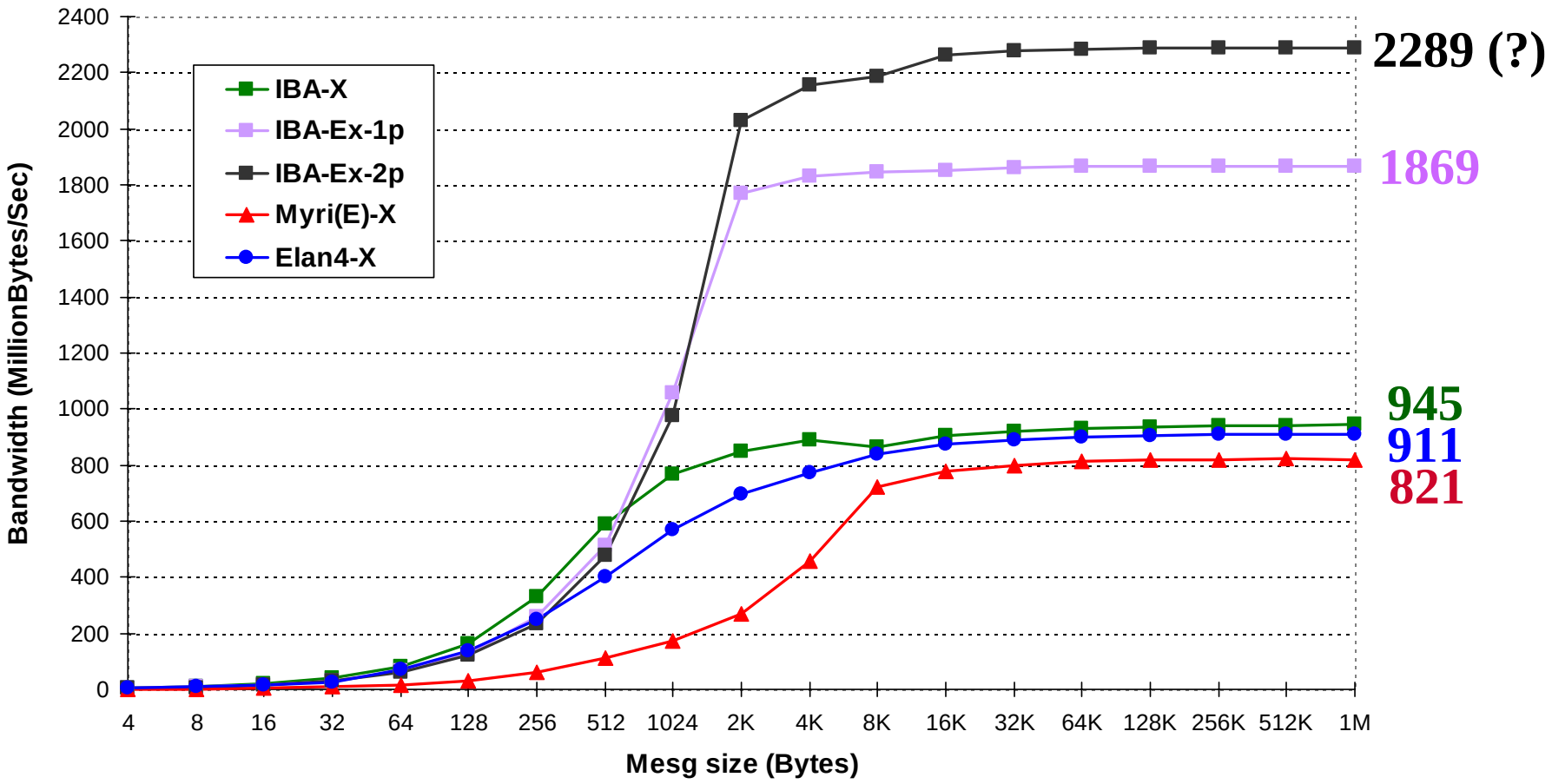


- Hot Interconnect '03
- IEEE Micro (Jan-Feb) '04 – one of the best papers from HotI '03

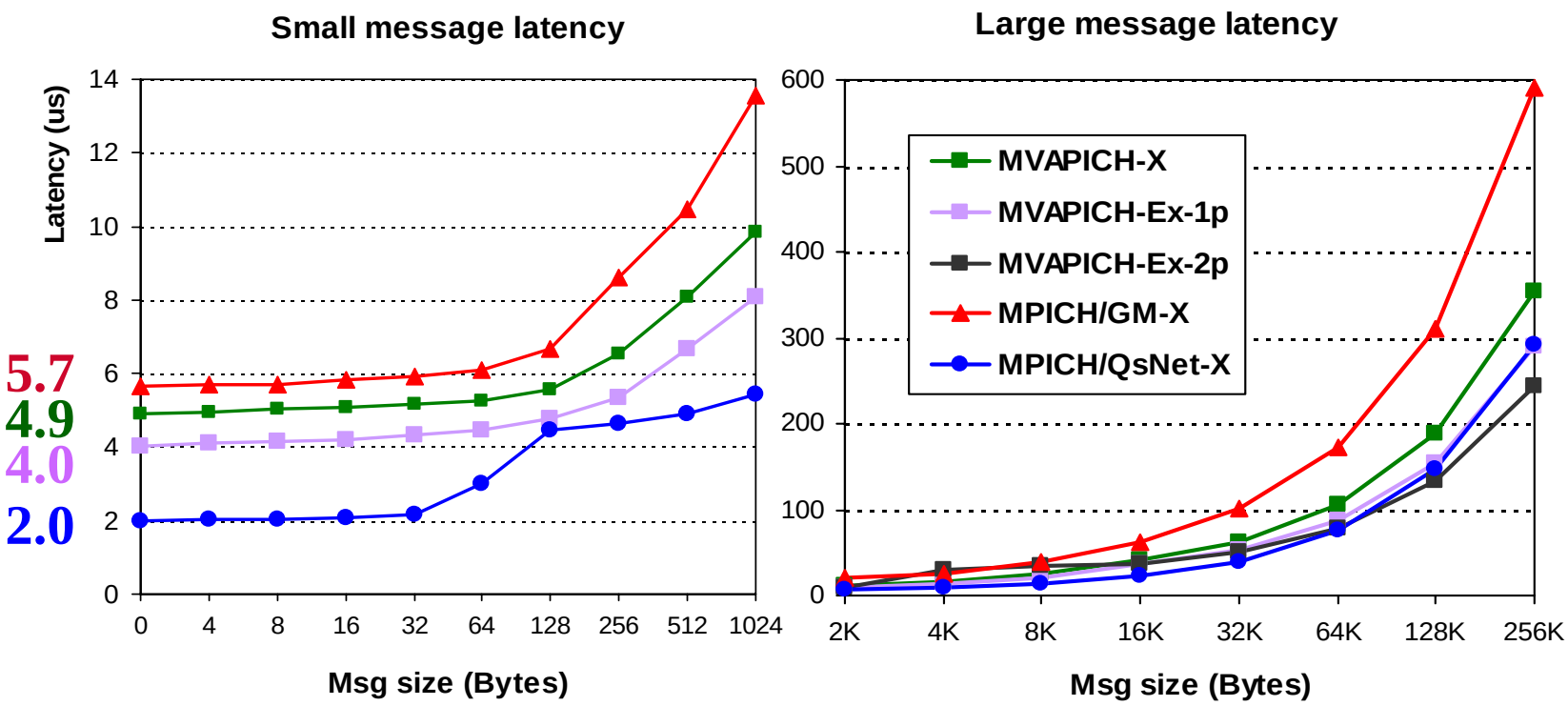
User-level Bandwidth (Uni-directional):
IBA/VAPI vs. Myrinet/GM vs. Quadrics/Elan4



User-level Bandwidth (Bi-directional):
IBA/VAPI vs. Myrinet/GM vs. Quadrics/Elan4

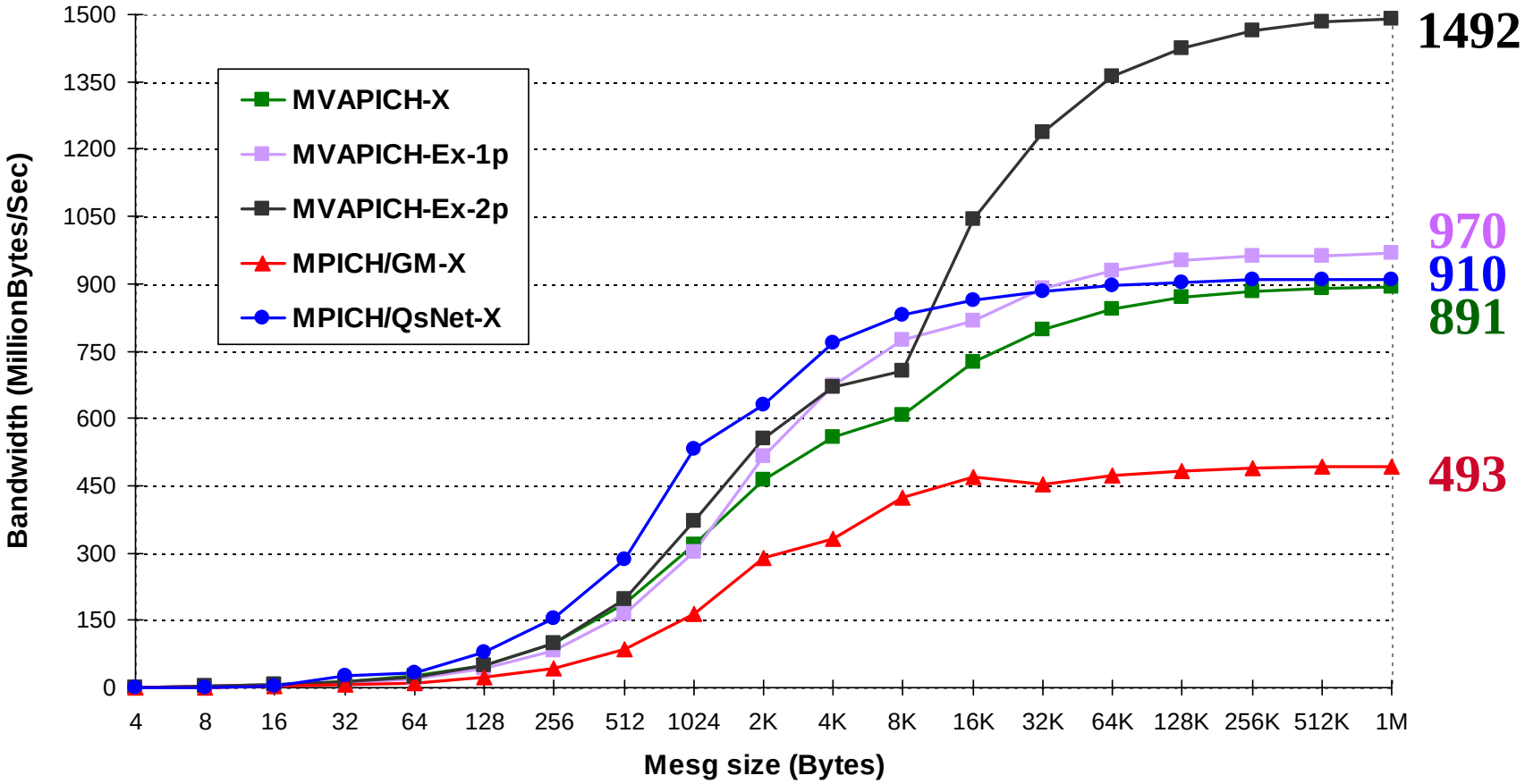


MPI-level Latency (One-way):
IBA vs. Myrinet vs. Quadrics

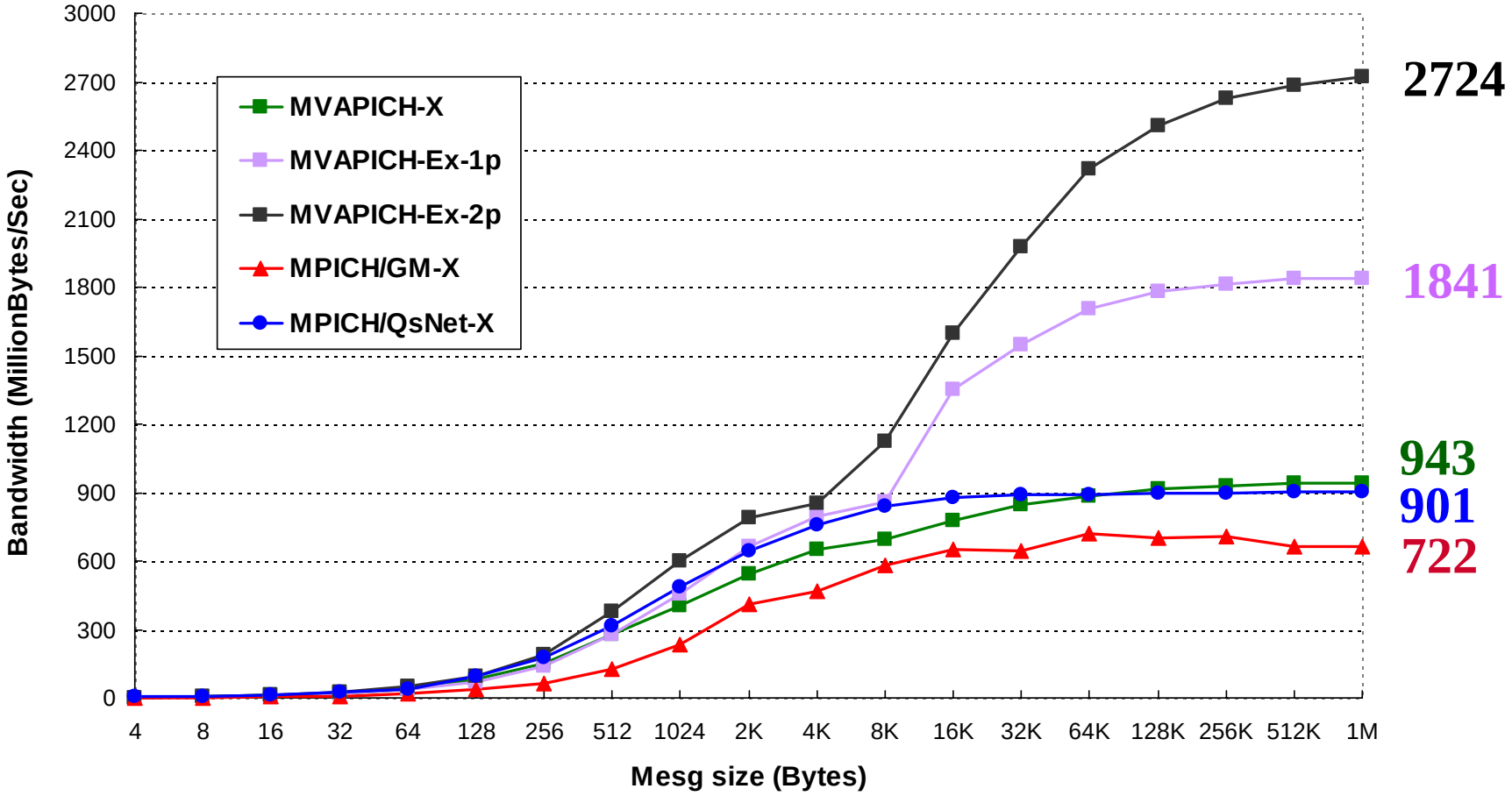


- SC '03
- Hot Interconnect '04
- IEEE Micro (Jan-Feb) '05, one of the best papers from HotI '04

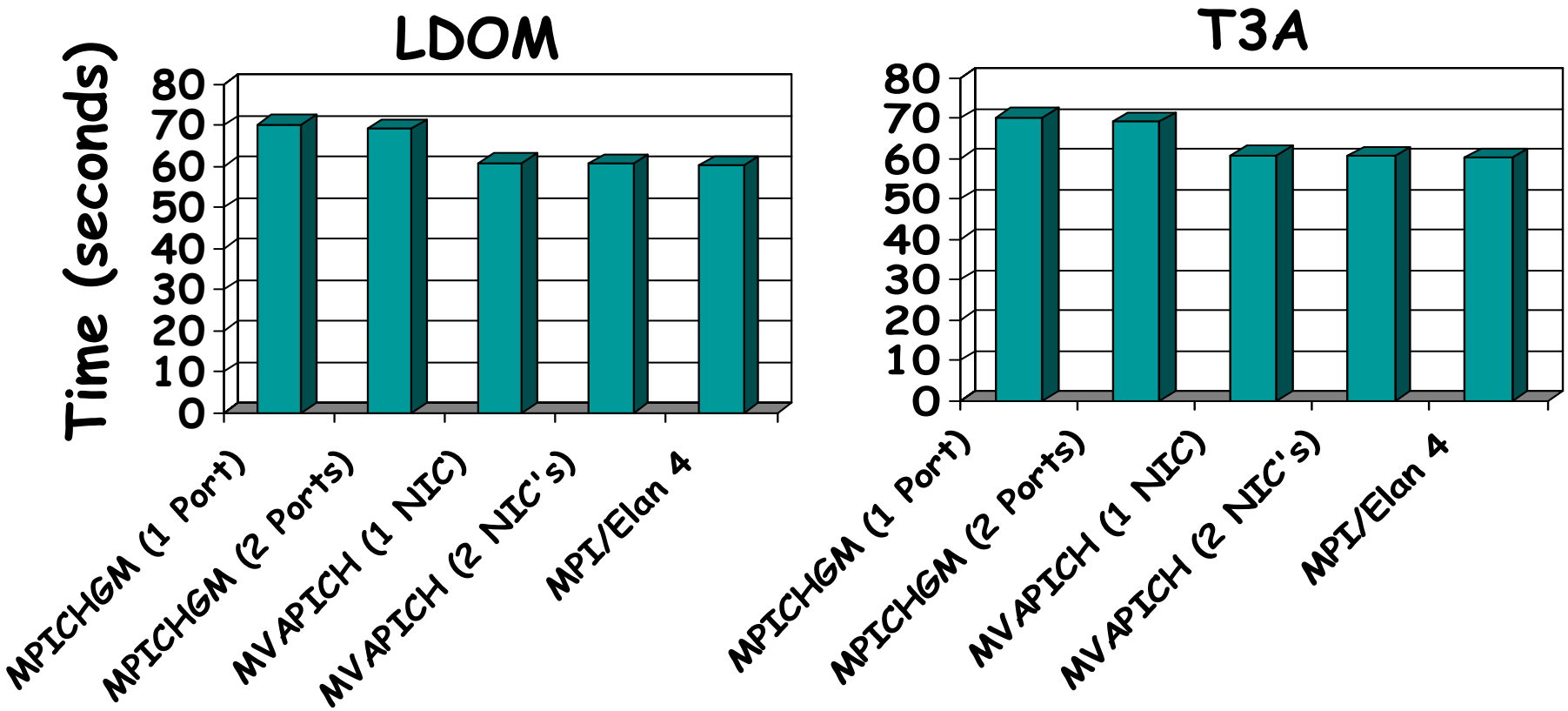
MPI-level Bandwidth (Uni-directional):
IBA vs. Myrinet vs. Quadrics



MPI-level Bandwidth (Bi-directional):
IBA vs. Myrinet vs. Quadrics

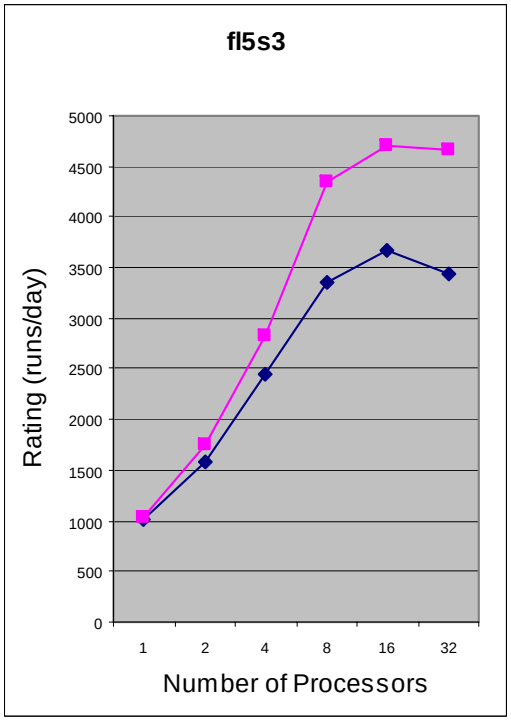


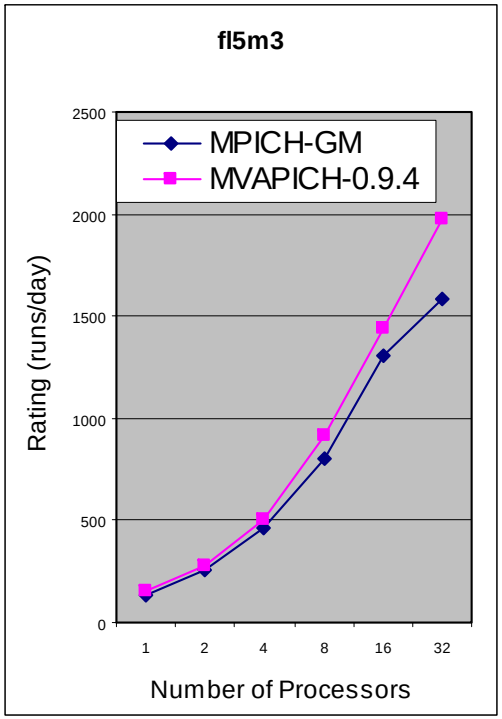
Performance of MM5: Network Comparison (8x2 processors)

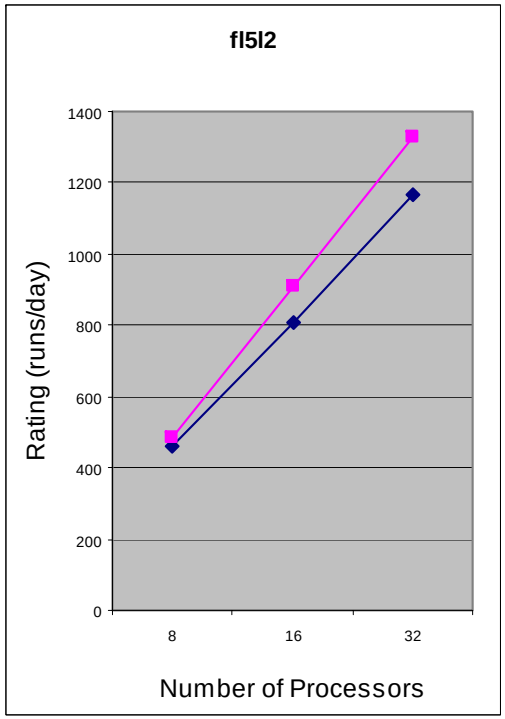


- Up to 12% improvement with MVAPICH compared with MPICH-GM

Performance Comparison for Fluent







- Ratings are better with MVAPICH/IBA than MPICH-GM/Myrinet

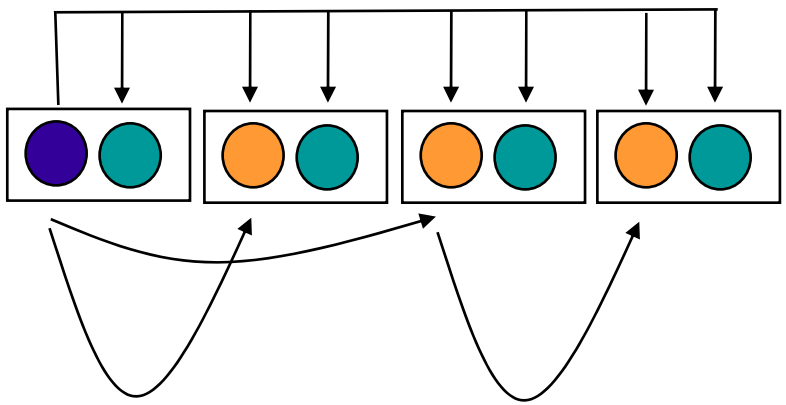
Latest Performance Results

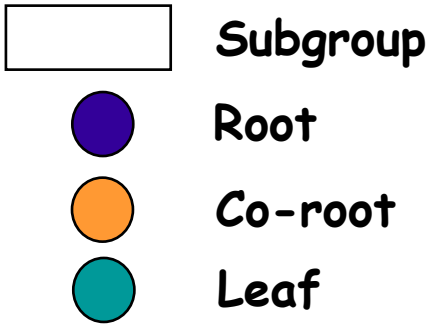
- MVAPICH-1
  - Point-to-point communication
    - IBA (PCI-X and PCI-Express) vs. Myrinet vs. Quadrics/Elan4
  - Collective communication
    - Broadcast and all-to-all
  - Multi-rail systems
  - Support for MPI Datatype Communication
- MVAPICH2
  - Two-sided
  - One-sided
    - Active
    - Passive

Challenges in using IBA Hardware Multicast for MPI Broadcast

- IBA Hardware Multicast
  - UD-based
  - No reliability
  - Limited to maximum one MTU (1K/2K bytes)
- MPI_Bcast should have
  - Reliability
  - Message ordering
  - Support for any Message size
- Need to bridge this gap
  - Performance
  - Scalability

Co-Root Based Approach

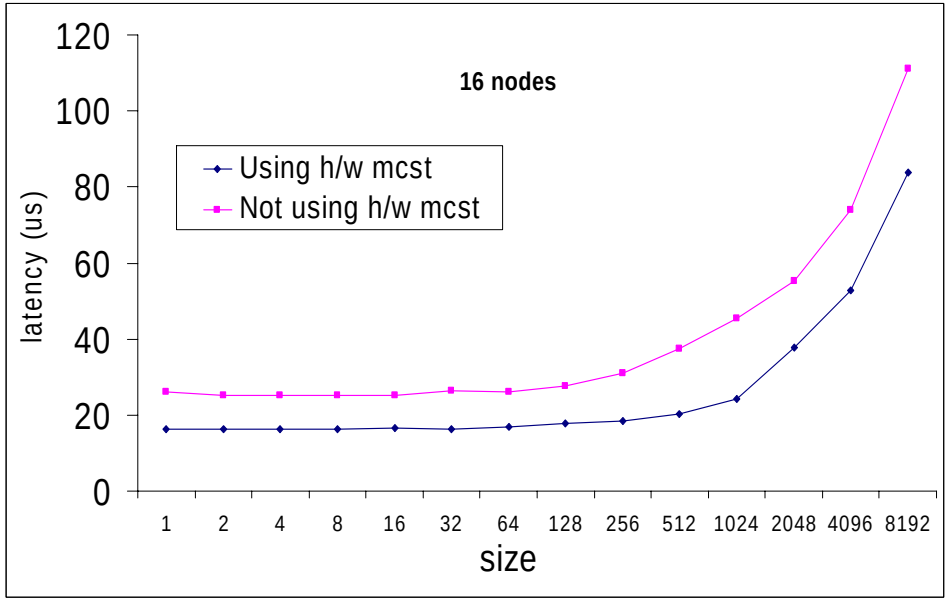


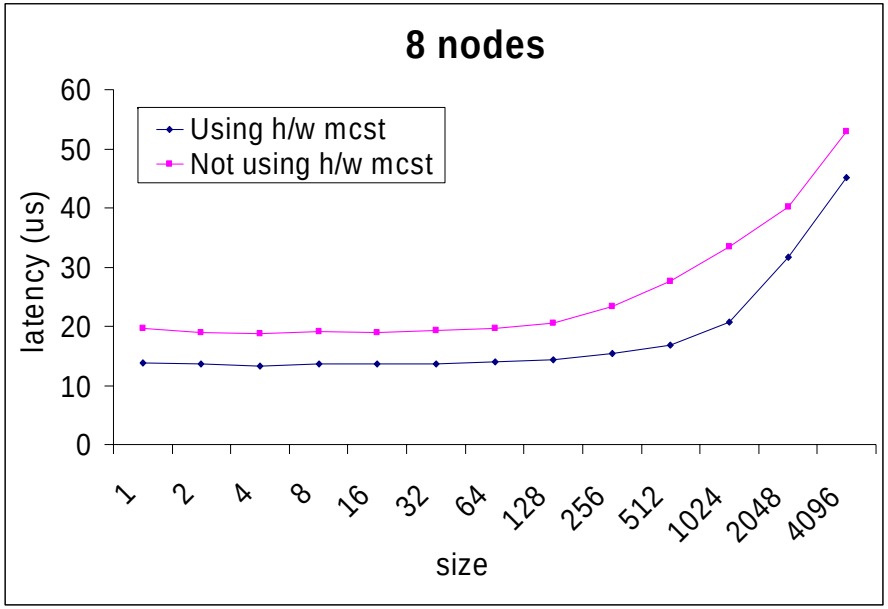


J. Liu, A. Mamidala and D. K. Panda, Fast and Scalable MPI-Level Broadcast using InfiniBand's Hardware Multicast Support, presented at IPDPS '04

- Two level structure
- Root does multicast
- Root does a broadcast to a set of co-roots (depending on system size)
- Root and co-roots responsible for ACK collection and retransmission in their respective subgroups
- Pipelining of packets
- Sliding window-based scheme for ack collection (not in critical path)

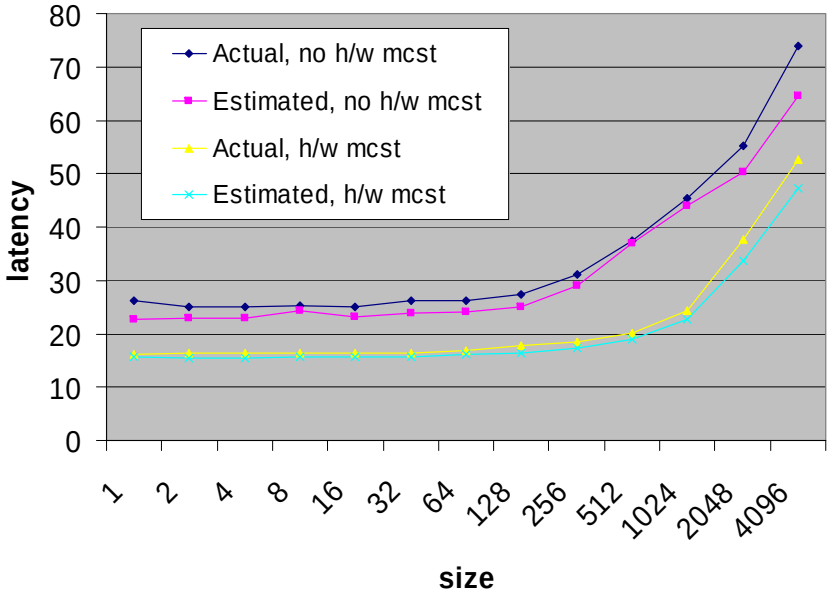
MPI_Bcast Latency

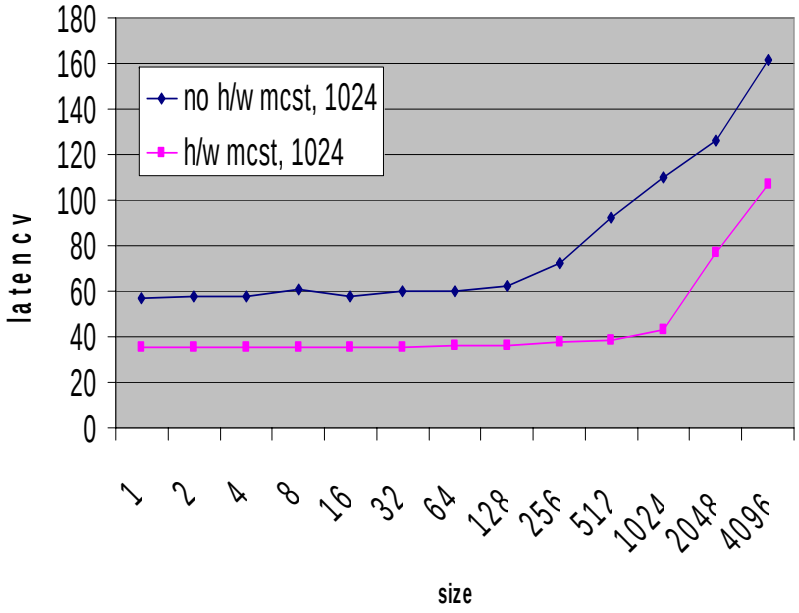




- Using hardware multicast improves the latency by 1.6 times for 16 nodes and 1.5 times for 8 nodes

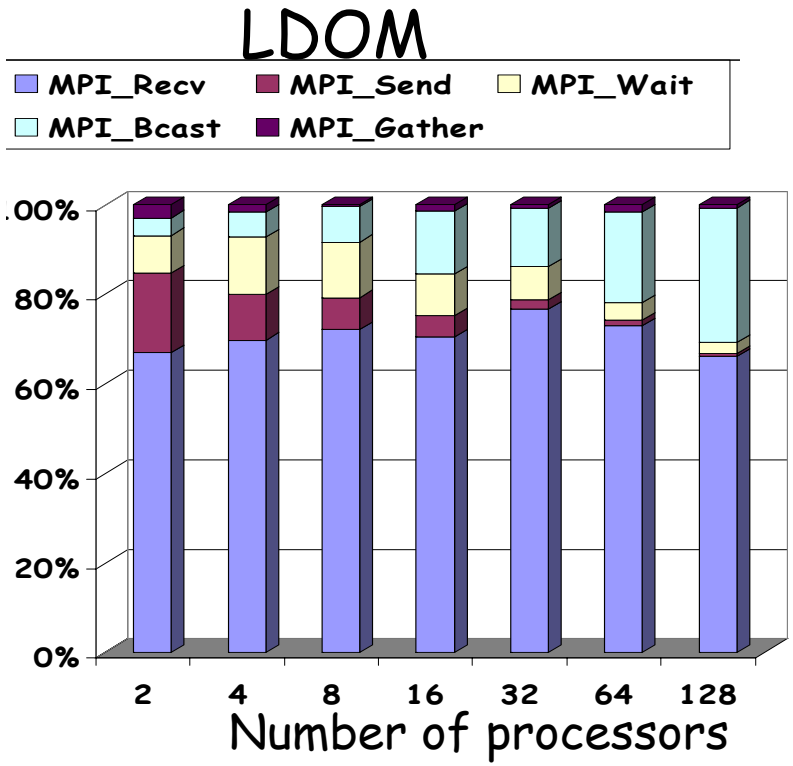
Analytical Model – Extrapolation for 1024 nodes

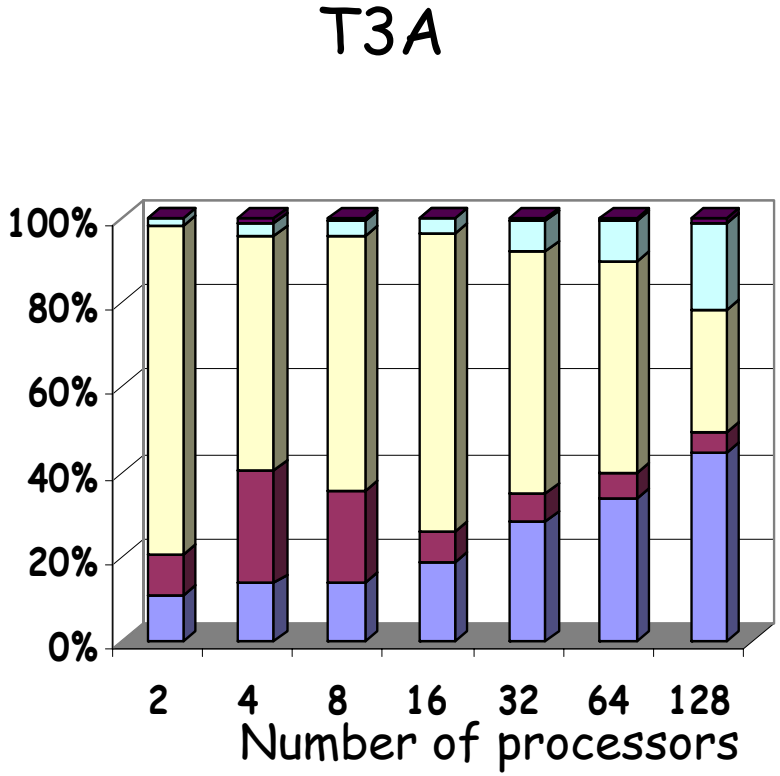




- Can broadcast a 1Kbyte message to 1024 nodes in 40.0 microsec

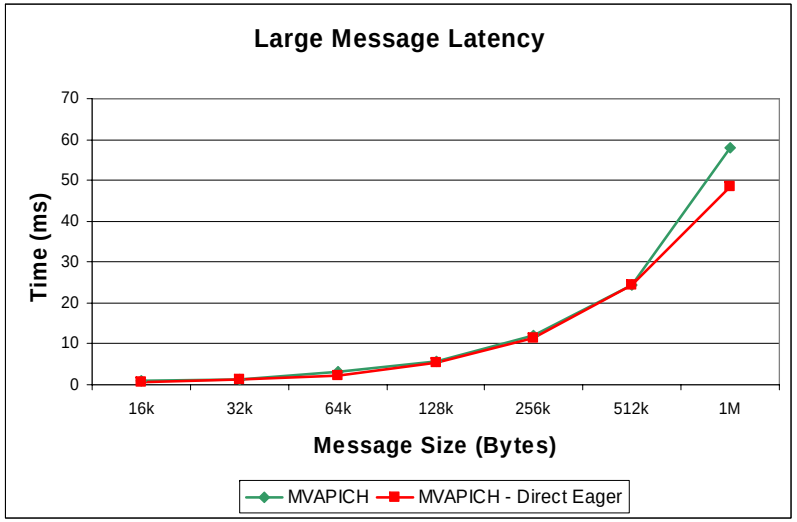
Impact of Hardware Broadcast on MM5

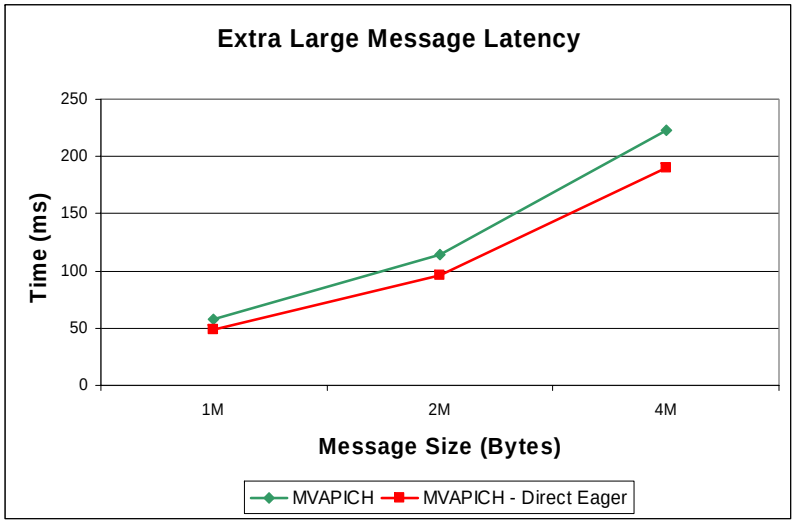




- MPI_Bcast up to 10% of total execution time (30% of communication time)
- Up to 5% improvement in execution time using hardware broadcast at processors (16 processors)

MPI All-to-all Latency on 16 Nodes





- All-to-all latency is reduced by factor of 3.07 for 32 bytes on 16 nodes.
- Analytical model suggests that the Direct Eager scheme will perform 64% better for 4k message size on 1024 nodes

S. Sur and D. K. Panda, Efficient and Scalable All-to-all Exchange for InfiniBand-based Clusters, presented at ICPP '04, Aug. '04

Latest Performance Results

- MVAPICH-1
  - Point-to-point communication
    - IBA (PCI-X and PCI-Express) vs. Myrinet vs. Quadrics/Elan4
  - Collective communication
    - Broadcast and all-to-all
  - Multi-rail systems
  - Support for MPI Datatype Communication
- MVAPICH2
  - Two-sided
  - One-sided
    - Active
    - Passive

Designing Multi-Rail IBA Systems

- Clusters use SMP nodes
  - 2-way and 4-way are typical
  - some with 8-way, 16-way, 32-way nodes
- Some nodes provide support for multiple independent PCI-X buses
  - Supermicro P4DL6 and X5DL8-GG
- Can these systems use multiple HCAs (with single/multiple ports each) to enhance communication performance on SMP-based clusters for
    - HPC
    - File systems
    - Storage systems

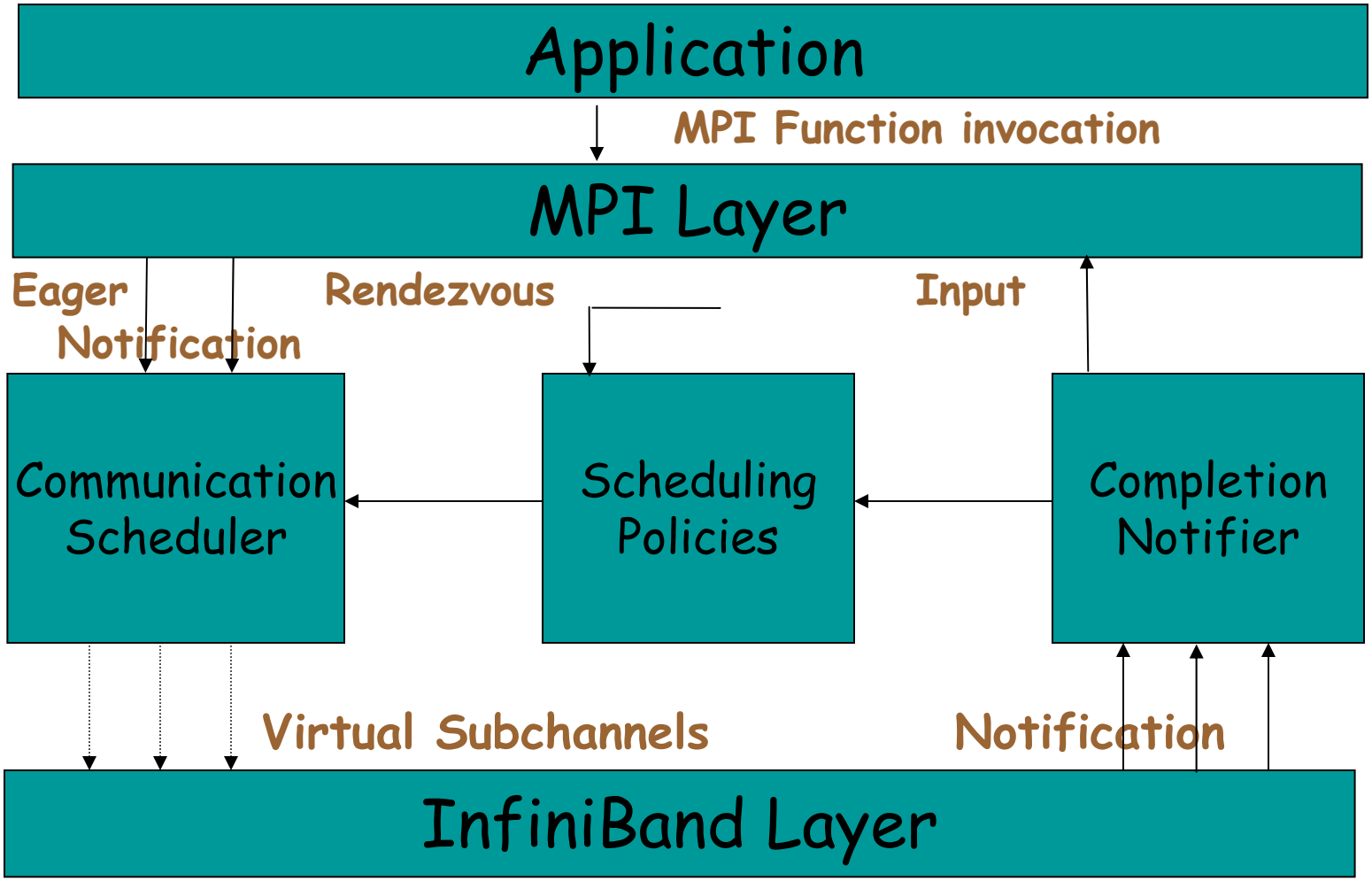
Using Multiple Ports/HCAs: Issues and Preliminary Studies

- Point-to-point communication
  - Registering memory of multiple NICs and making them available to a given process/processor
  - Handling out-of-order messages
  - Striping communication (long messages) across
    - Multiple ports/NICs
  - Load balancing communication (both long and short) across
    - Multiple ports/NICs
    - Static vs. Dynamic (based on network congestion status)
- Collective communication
  - Many interesting issues
  - Striping multicast data over multiple rails
  - Dedicated collective communication rail

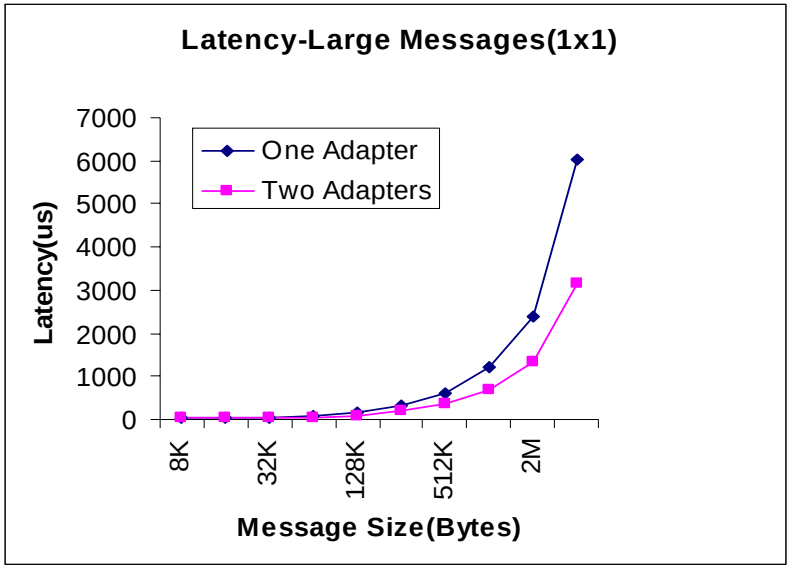
Unified Approach to Multirail Network Configurations

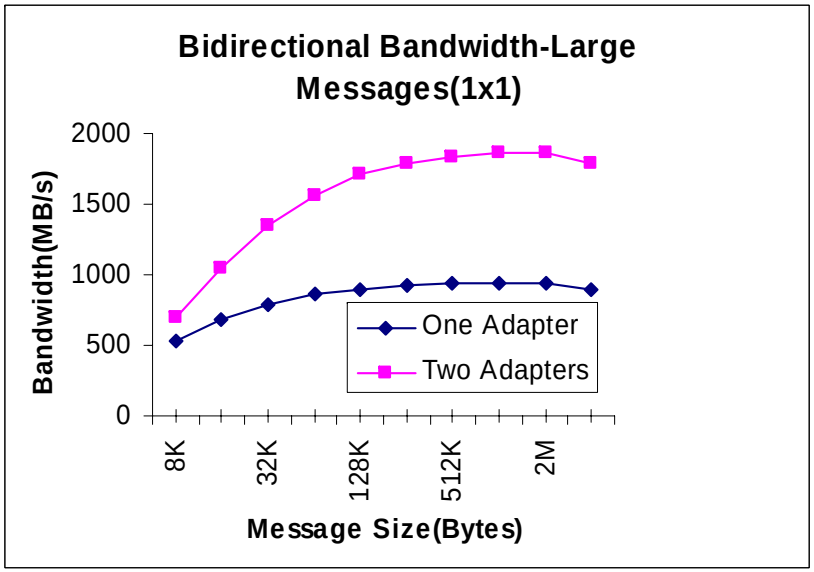
- Multiple Adapters
  - Can be used for SMP systems with I/O bus being the bottleneck
- Multiple Ports
  - Can be used with the systems, with link bandwidth as the bottleneck
- Multiple Paths with LMC
  - Can be used with above systems, when network congestion is the bottleneck

Proposed MPI Design Framework



Performance Evaluation of PCI-X (Two Adapters)





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

J. Liu, A. Vishnu, and D. K. Panda, Building Multirail InfiniBand Clusters: MPI-Level Design and Performance Evaluation, Presented at SC '04, Nov. 2004

Multi-rail support will be available in 0.9.5 (soon)

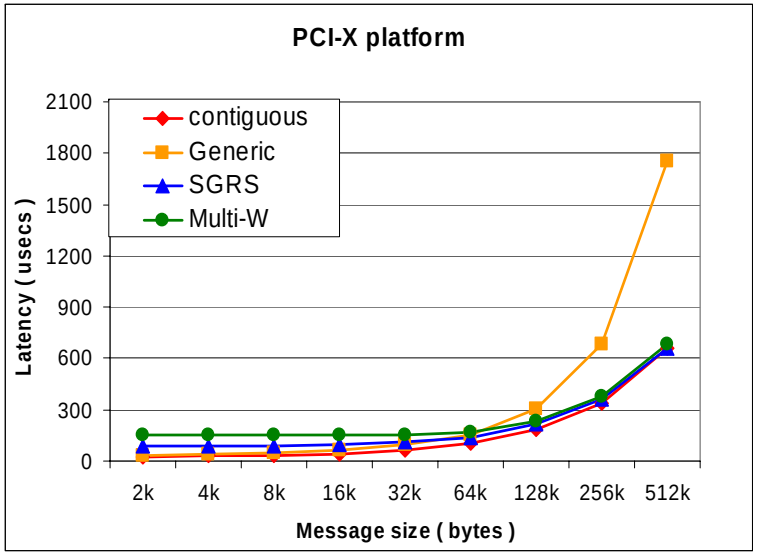
Latest Performance Results

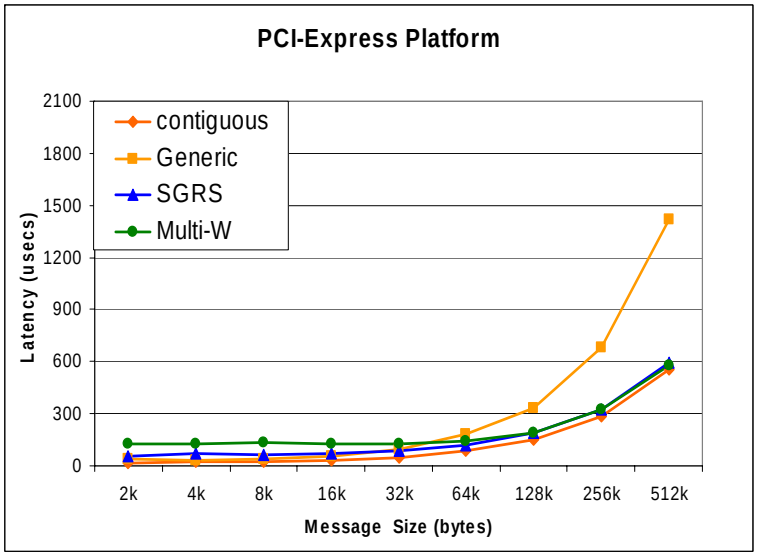
- MVAPICH-1
  - Point-to-point communication
    - IBA (PCI-X and PCI-Express) vs. Myrinet vs. Quadrics/Elan4
  - Collective communication
    - Broadcast and all-to-all
  - Multi-rail systems
  - Support for MPI Datatype Communication
  - PCI-X and PCI-Express performance comparisons
- MVAPICH2
  - Two-sided
  - One-sided
    - Active
    - Passive

Performance Issues in MPI Datatype Communication

- Many scientific applications use datatype processing
- Memory copies due to pack/unpack operations
- Memory registration and deregistration on dynamic pack and unpack buffers
- High startup costs due to a large number of communication operations
- Low network utilization due to small message sizes
- High cost for datatype processing

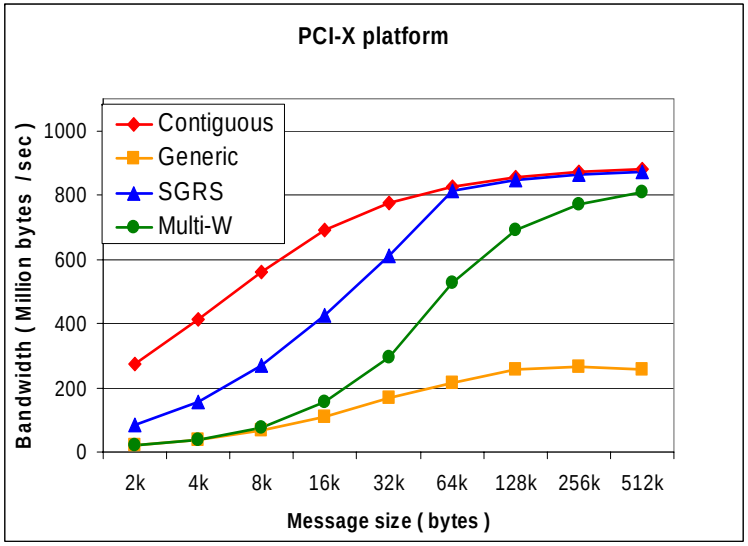
Ping-Pong Vector Latency

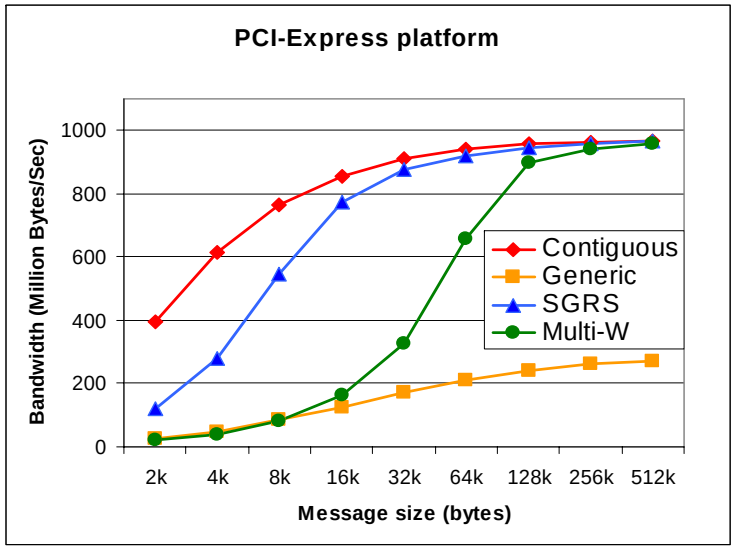




- **Vector**: variable numbers of columns in a 64*4096 integer array (64 non-contiguous segments)
- For small messages Generic scheme does better due to low copy cost
- PCI-X - SGRS scheme performs best beyond 64k
  reduces latency up to 62% compared to Multi-W
- PCI-Express - SGRS scheme performs best beyond 32k
  reduces latency up to 68% compared to Multi-W

Vector Unidirectional Bandwidth

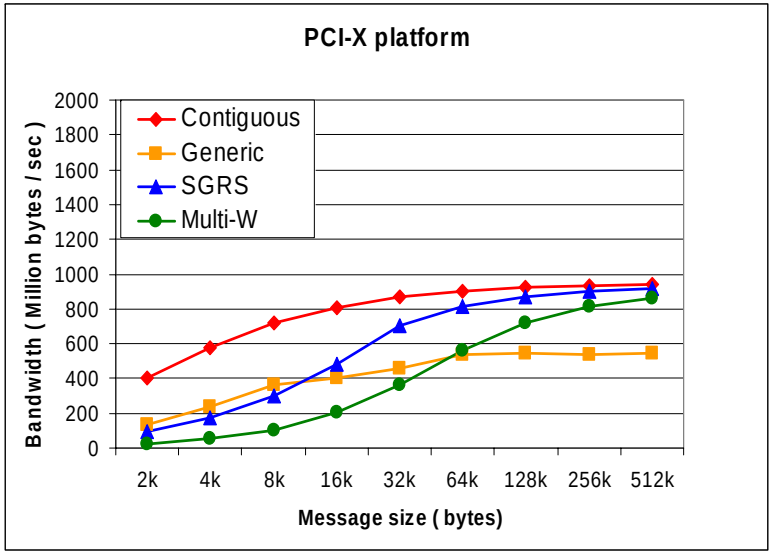


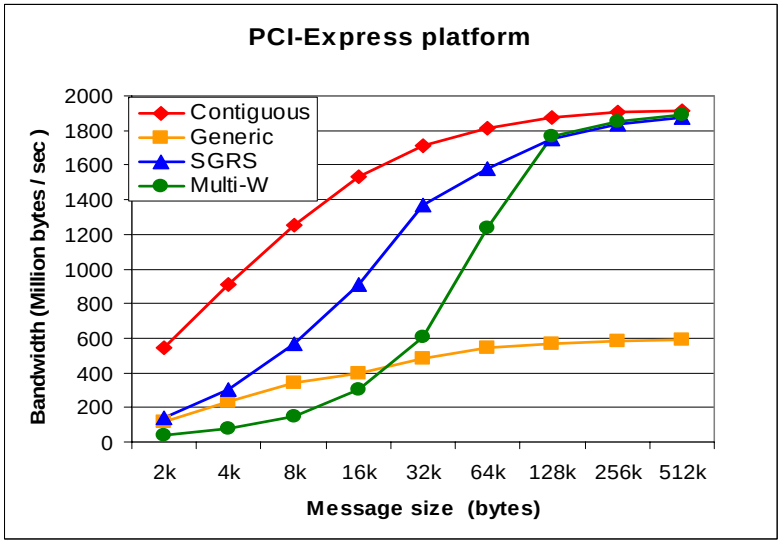


- SGRS scheme performs the best , bandwidth close to peak contiguous bandwidth
- On PCI-X
  - SGRS scheme does overall better than Multi-W
  - up to 4.0 times better than generic scheme
- On PCI-Express platform
  - up to 6.4 times better than generic scheme

- Euro/PVM '04
- Special issue of IJCNN (one of the best five papers)

Vector Bidirectional Bandwidth



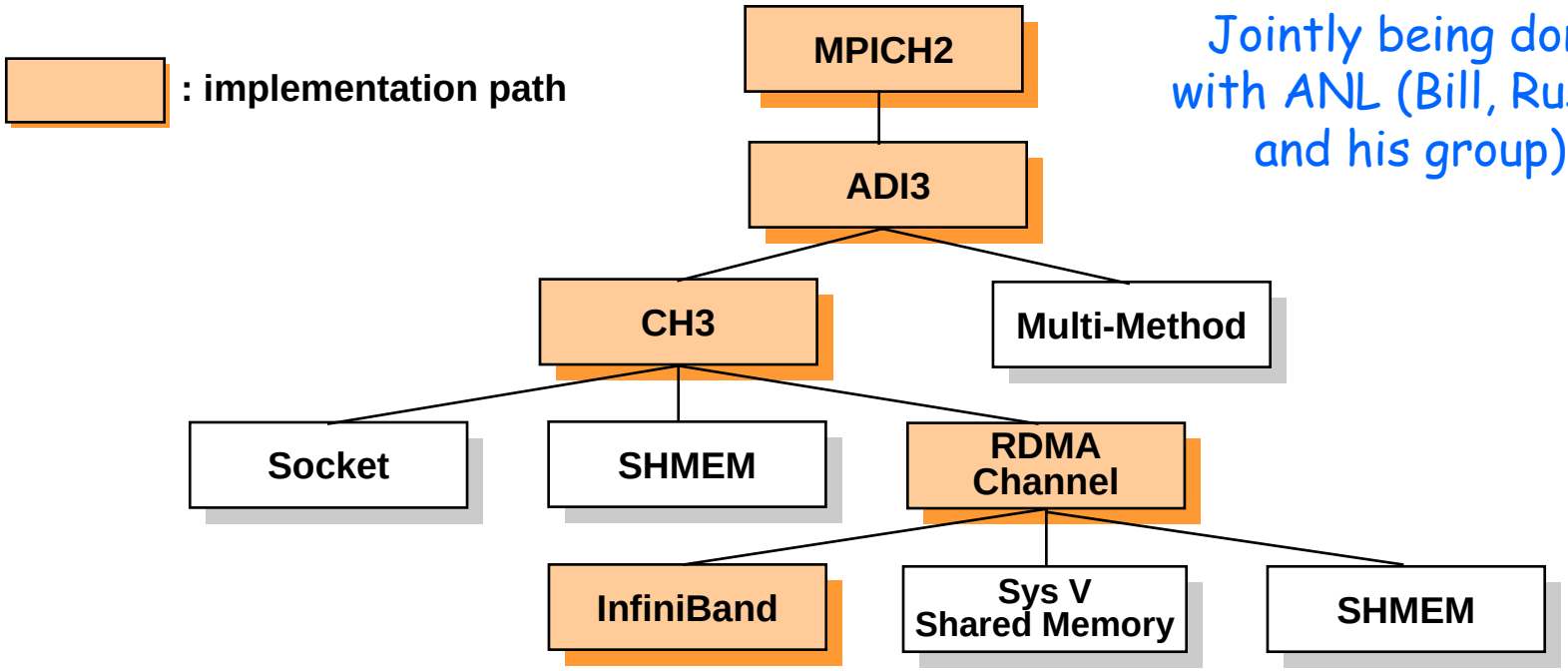


- On PCI-X
  - SGRS scheme performs overall better than Multi-W
  - SGRS scheme : up to 1.67 times improvement over Generic scheme
- On PCI-Express
  - Generic (copy based) scheme saturates around 600 MB/s
  - SGRS scheme : up to 3.15 times improvement over Generic scheme
  - SGRS scheme : peak bidirectional bandwidth of 1876 MB/sec

Latest Performance Results

- MVAPICH-1
  - Point-to-point communication
    - IBA (PCI-X and PCI-Express) vs. Myrinet vs. Quadrics/Elan4
  - Collective communication
    - Broadcast and all-to-all
  - Multi-rail systems
  - Support for MPI Datatype Communication
- MVAPICH2
  - Two-sided
  - One-sided
    - Active
    - Passive

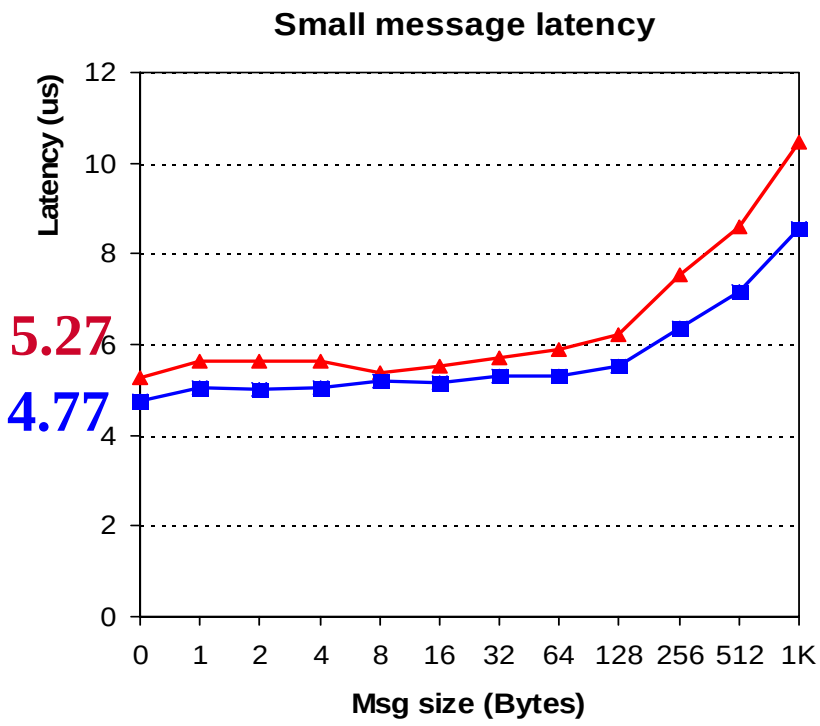
MVAPICH2 Implementation Structure

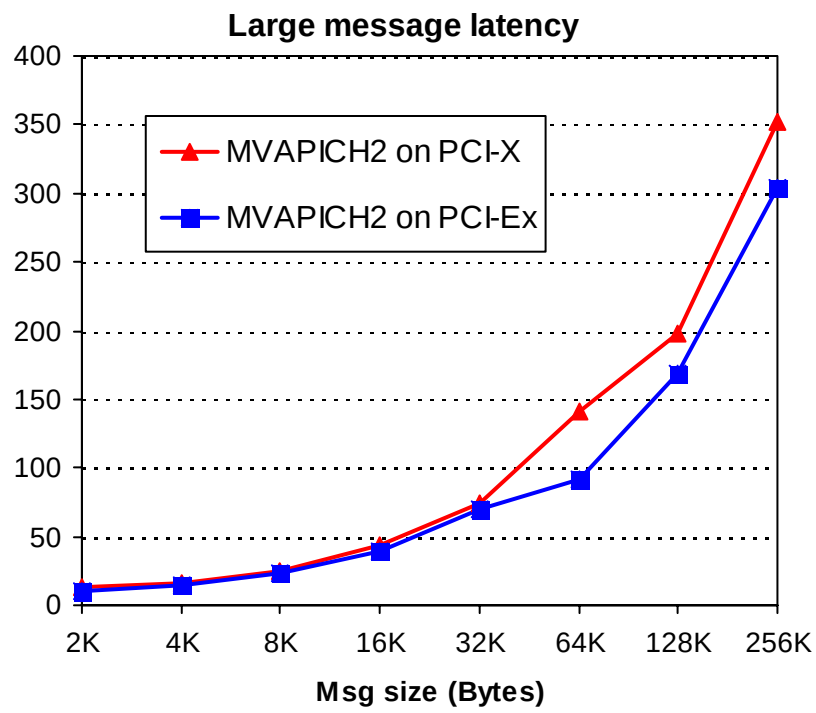


Jointly being done with ANL (Bill, Rusty, and his group)

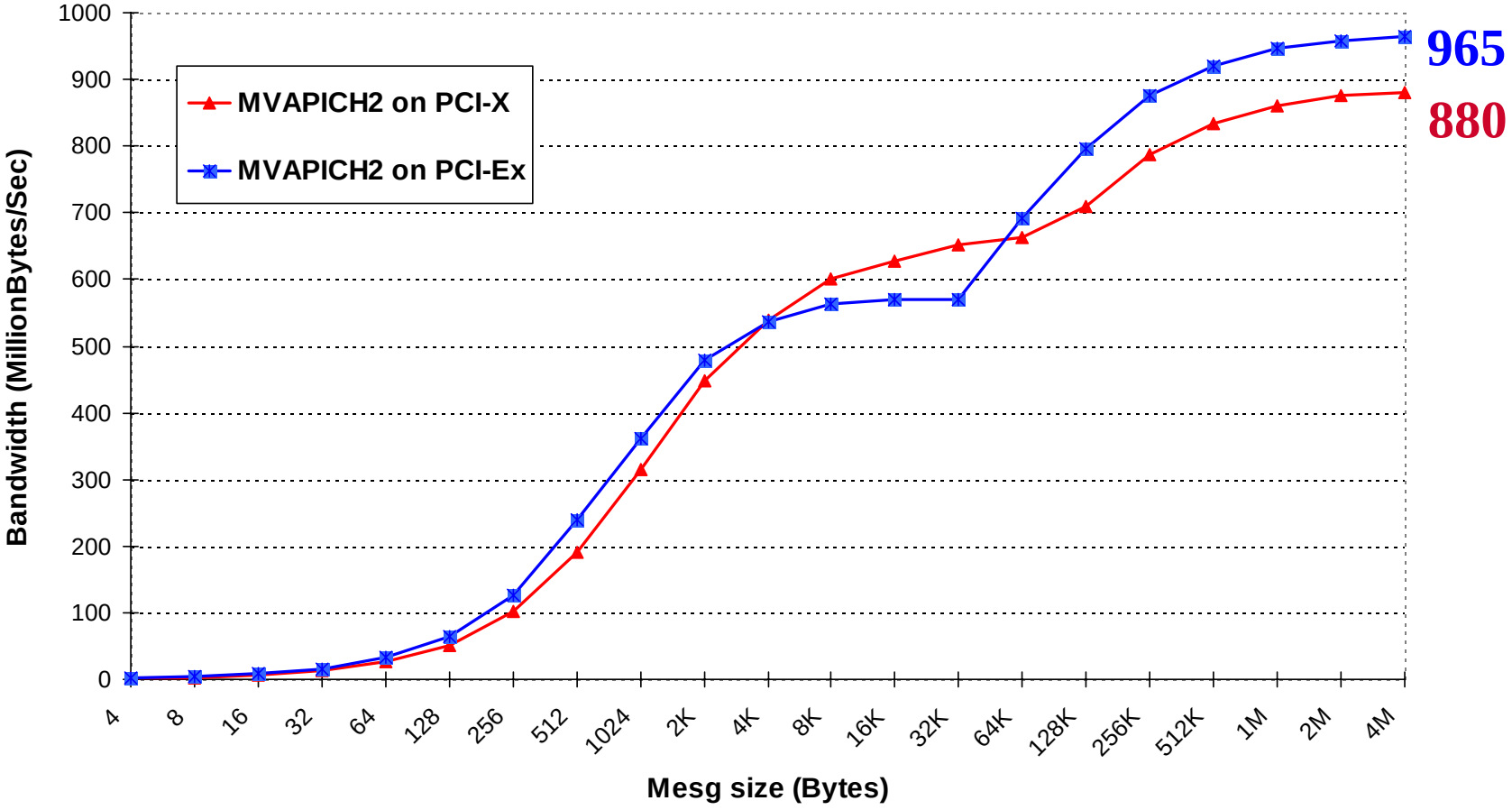
J. Liu, W. Jiang, P. Wyckoff, D. K. Panda, D. Ashton, D. Buntinas, W. Gropp, and B. Toonen, Design and Implementation of MPICH2 over InfiniBand with RDMA Support, IEEE International Parallel and Distributed Processing Symposium (IPDPS), April 2004

MVAPICH2 Send/Recv Latency on PCI-X and PCI-Express

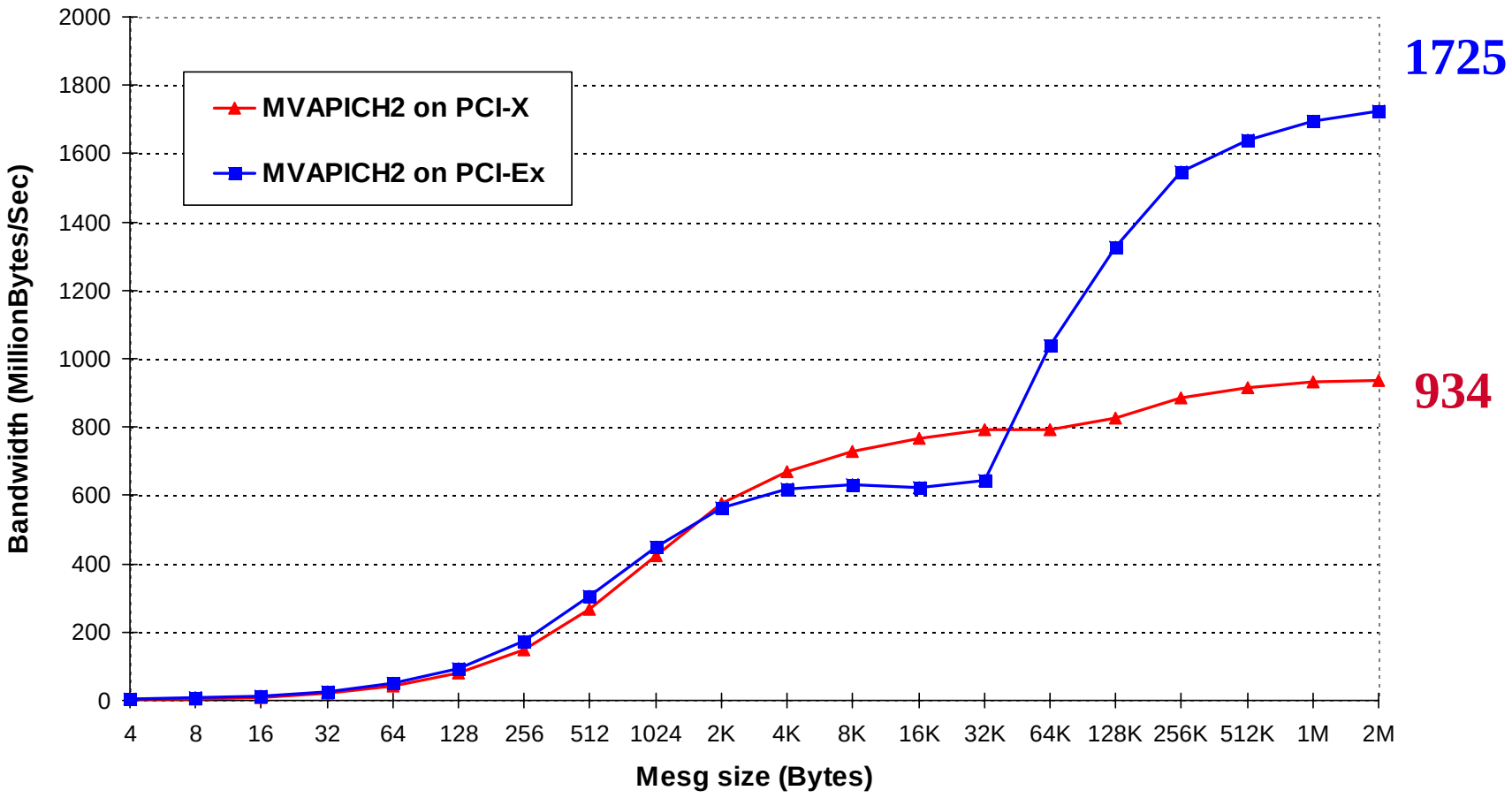




MVAPICH2 Send/Recv Bandwidth on PCI-X and PCI-Express



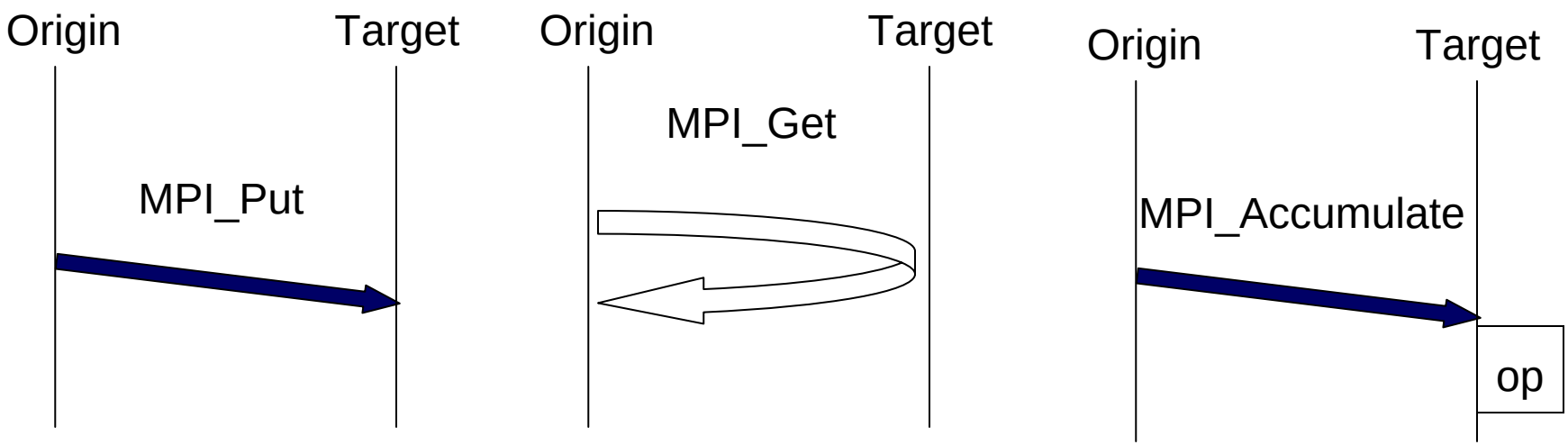
MVAPICH2 Send/Recv Bi-Directional Bandwidth on PCI-X and PCI-Express



Latest Performance Results

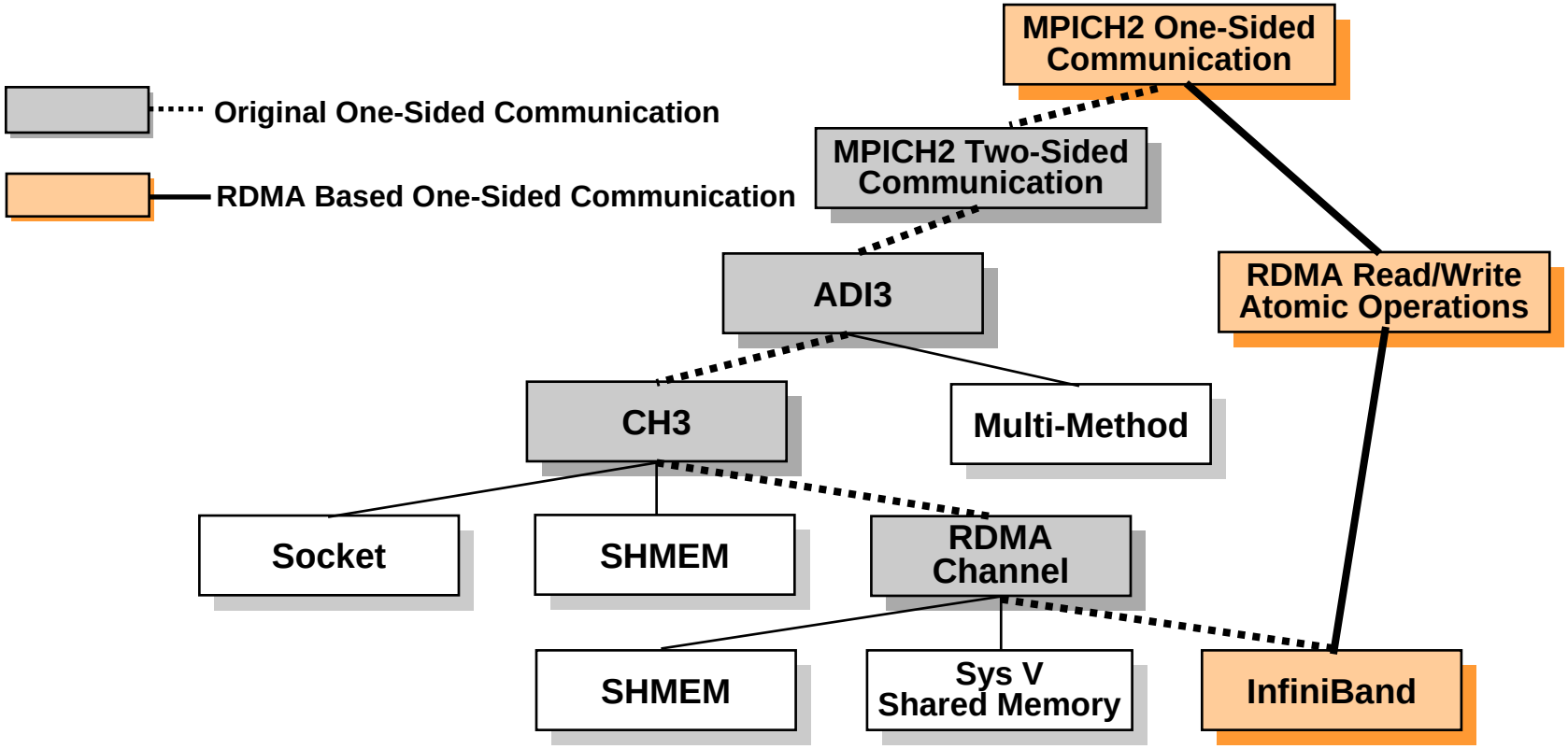
- MVAPICH-1
  - Point-to-point communication
    - IBA (PCI-X and PCI-Express) vs. Myrinet vs. Quadrics/Elan4
  - Collective communication
    - Broadcast and all-to-all
  - Multi-rail systems
  - Support for MPI Datatype Communication
- MVAPICH2
  - Two-sided
  - One-sided
    - Active
    - Passive

MPI-2 One-Sided Communication (Cont'd)



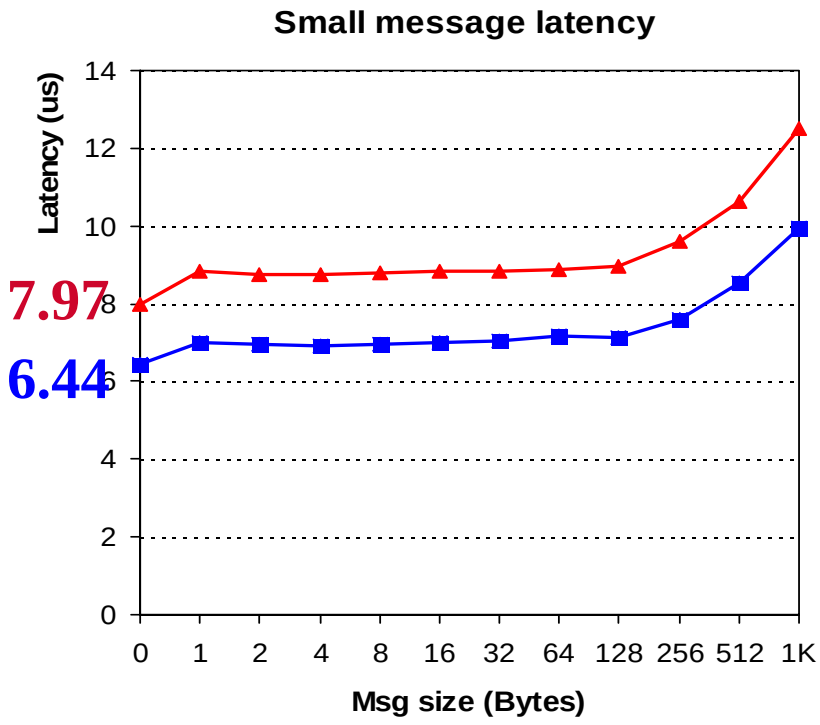
- Communication functions
  - MPI_Put
  - MPI_Get
  - MPI_Accumulate

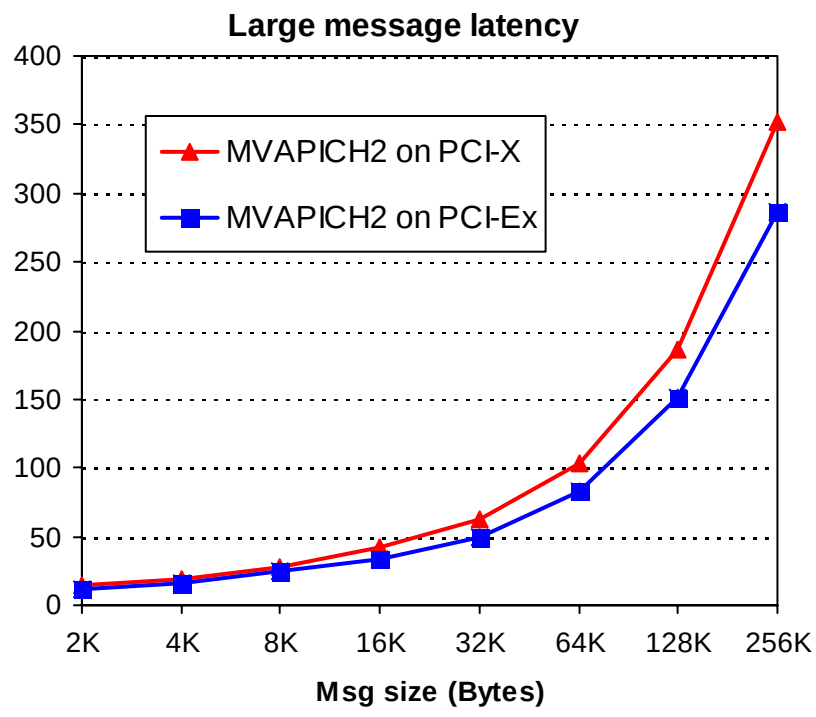
One-Sided Communication Implementation Structure



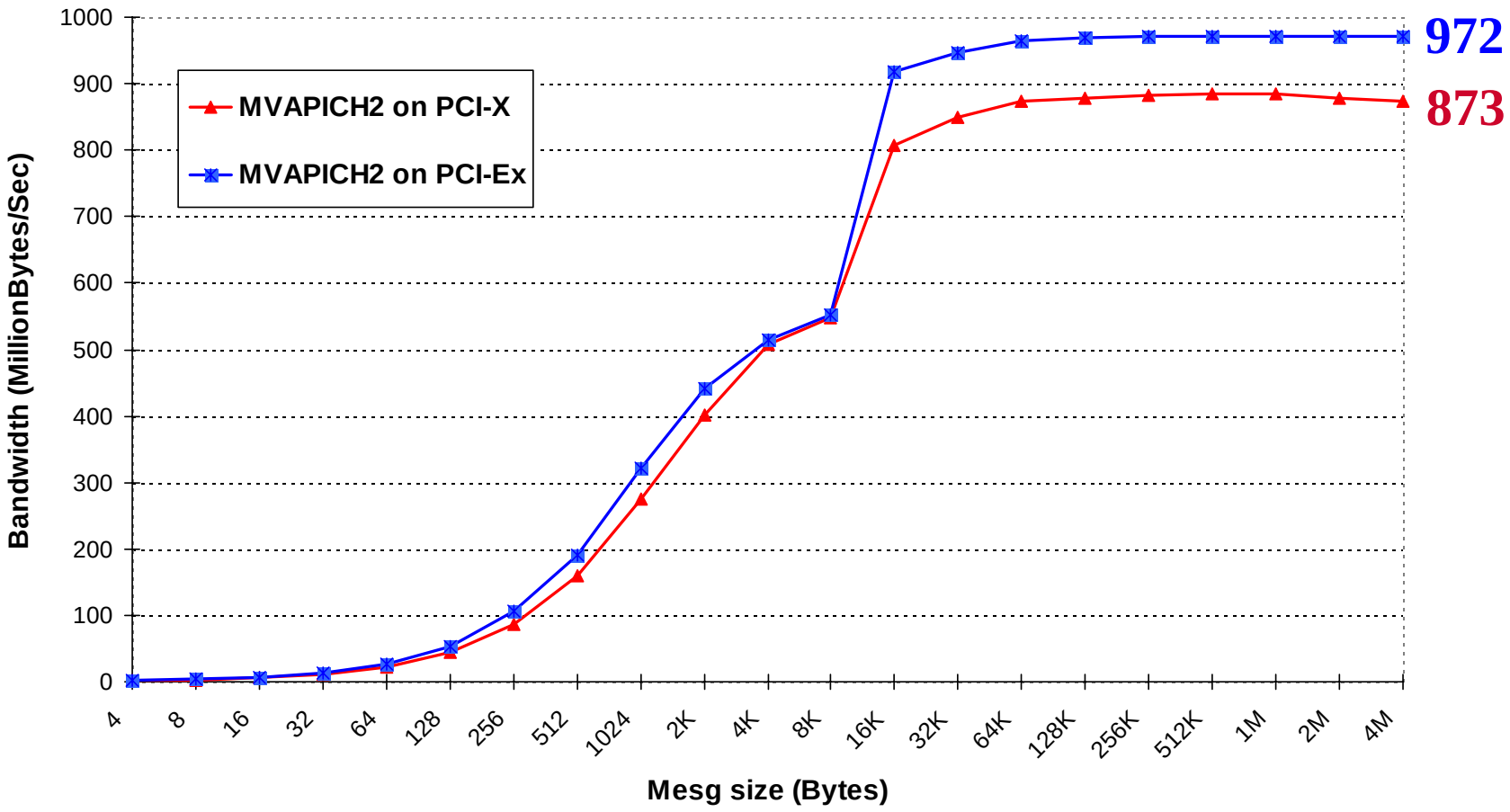
High Performance MPI-2 One-Sided Communication over InfiniBand, W. Jiang, J. Liu, H. Jin, D.K. Panda , W.Gropp and R. Thakur. IEEE International Symposium on Cluster Computing and the Grid (CCGrid), May 2004

MVAPICH2: One-Sided Put Latency with Active Target Implementation

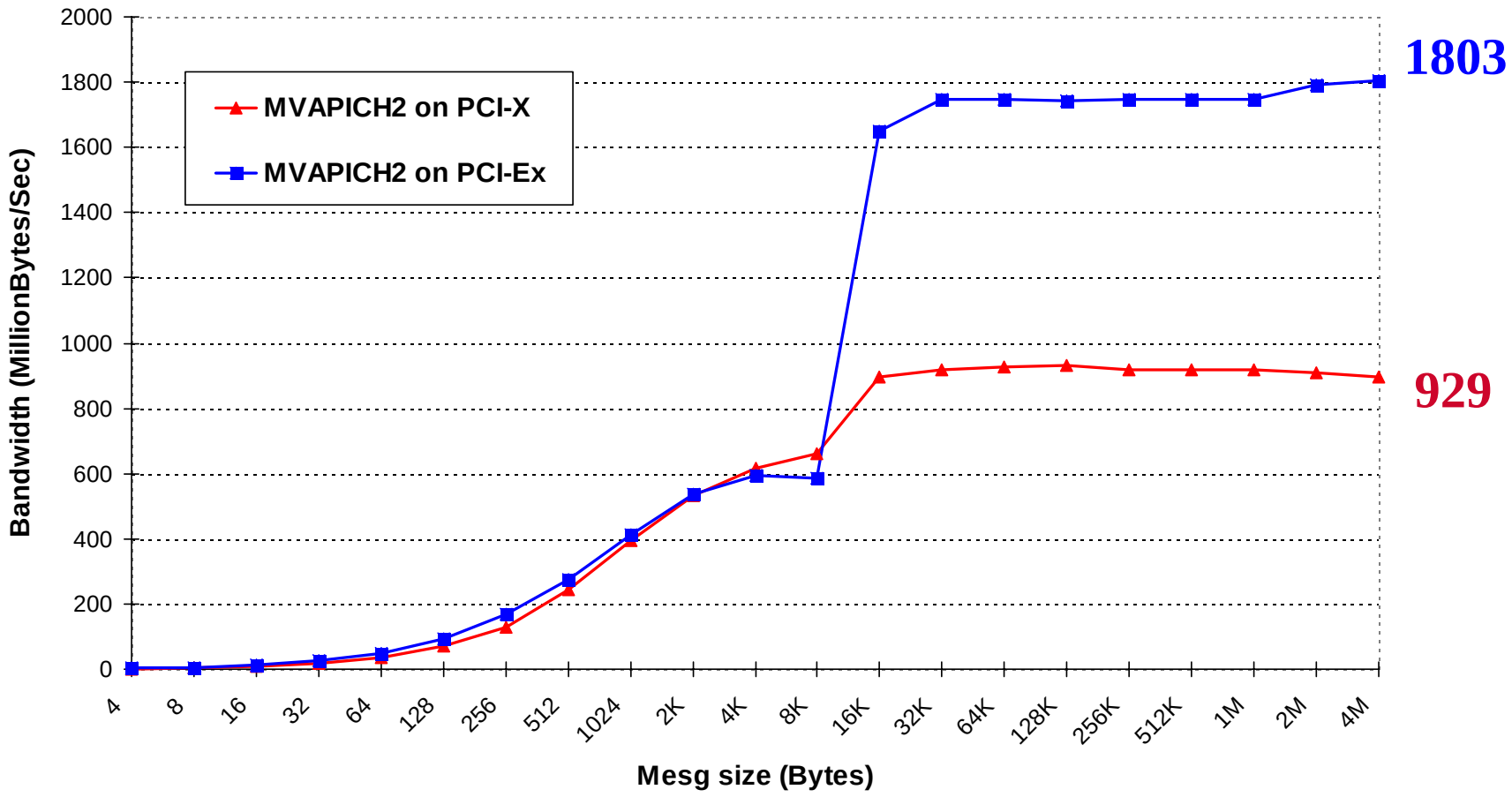




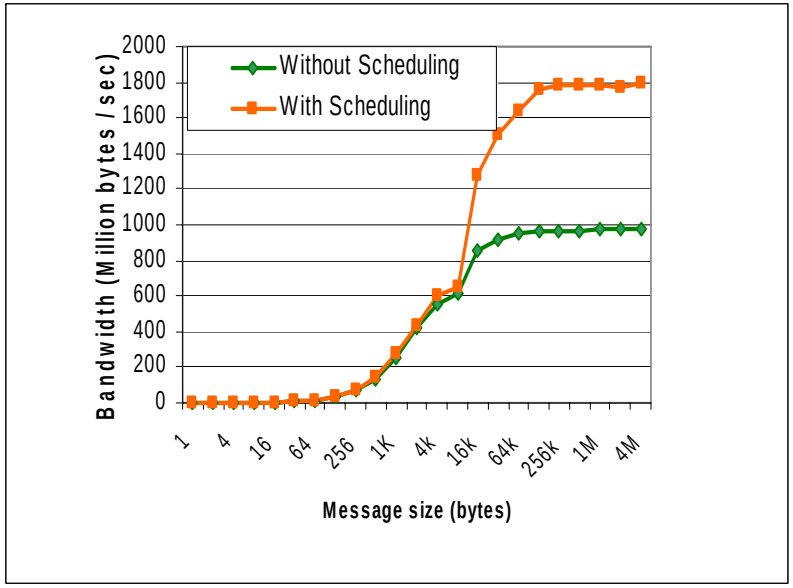
MVAPICH2: One-Sided Put Bandwidth with Active Target Implementation

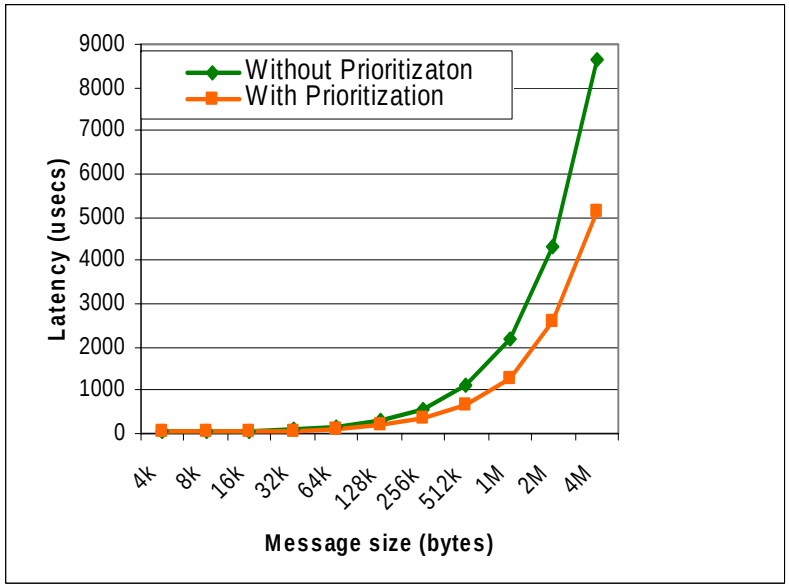


MVAPICH2: One-Sided Put Bidirectional Bandwidth with Active Target Implementation



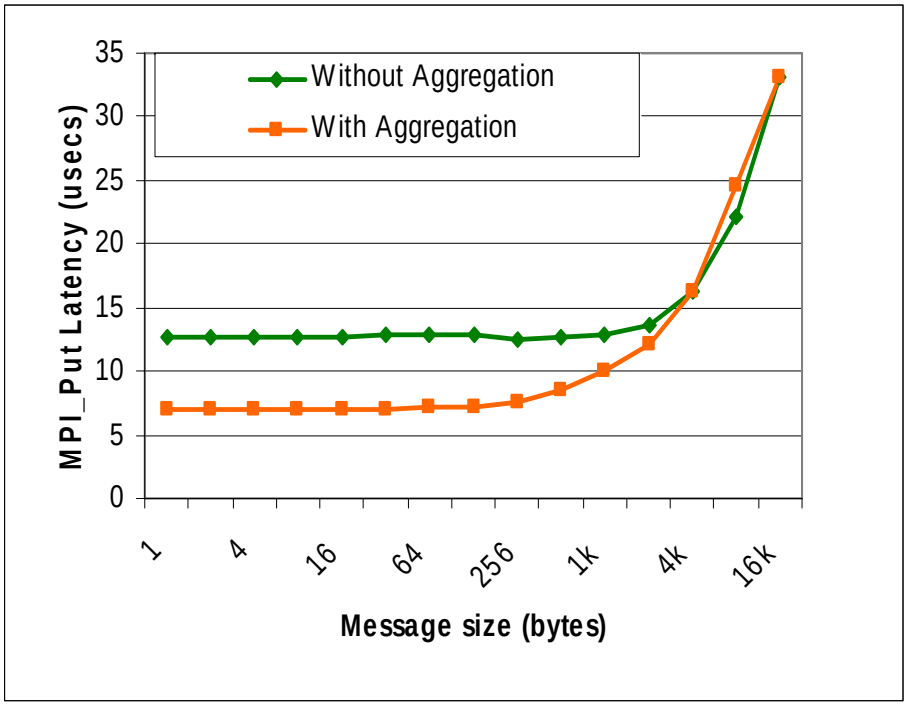
Optimized Scheduling of Gets and Puts - Benefits





- Scheduling by Interleaving MPI_Put and MPI_Get improves the throughput by up to 76%
- Scheduling by Prioritizing MPI_Get over MPI_Put reduces the latency by up to 40%

Aggregation of small RMA and Synchronization message - Benefits



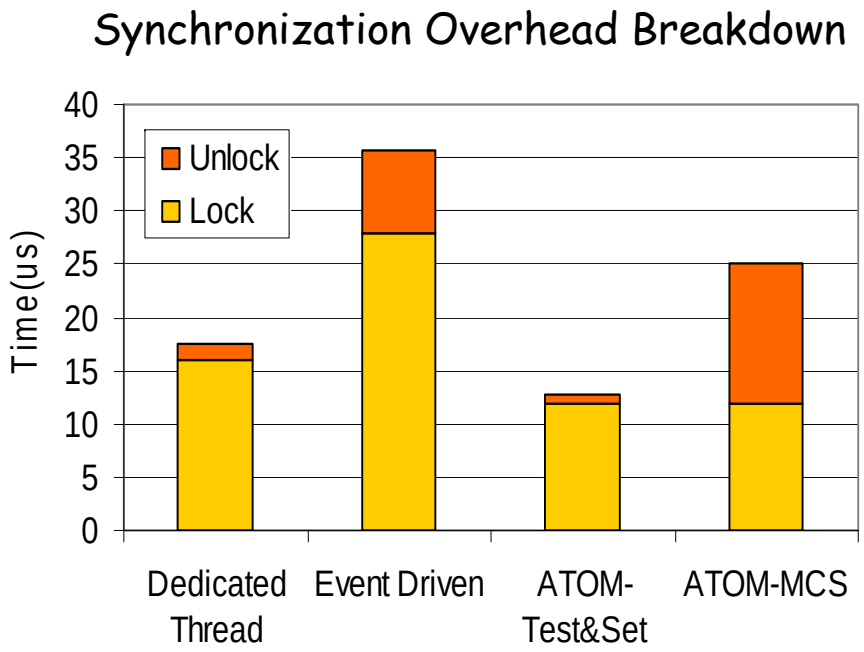
- Aggregates the synchronization message with one small RMA operation
- Reduces the latency of MPI_Put and MPI_Get for small messages up to 44% and 42% respectively

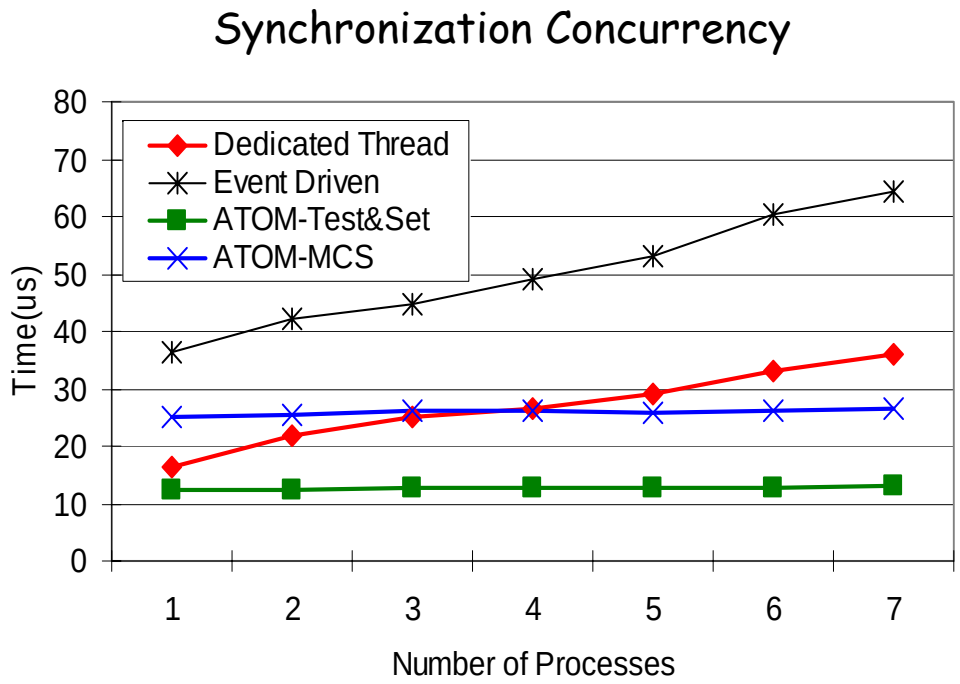
One-sided with Passive Target Synchronization

- Passive Target Synchronization
  - MPI_Win_lock, MPI_Win_unlock
  - No Involvement of Target even in Synchronization
- Design Alternatives
  - Thread Based Design
    - Dedicated Thread
    - Event Driven Thread
  - Atomic Operation Based Design
    - MCS
    - Test&Set

W. Jiang, J. Liu, H. –W. Jin, D. K. Panda, D. Buntinas, R. Thakur, and W. Gropp, Efficient Implementation of MPI-2 Passive One-Sided Communication over InfiniBand Clusters, presented at Euro PVM/MPI, Sept. 2004

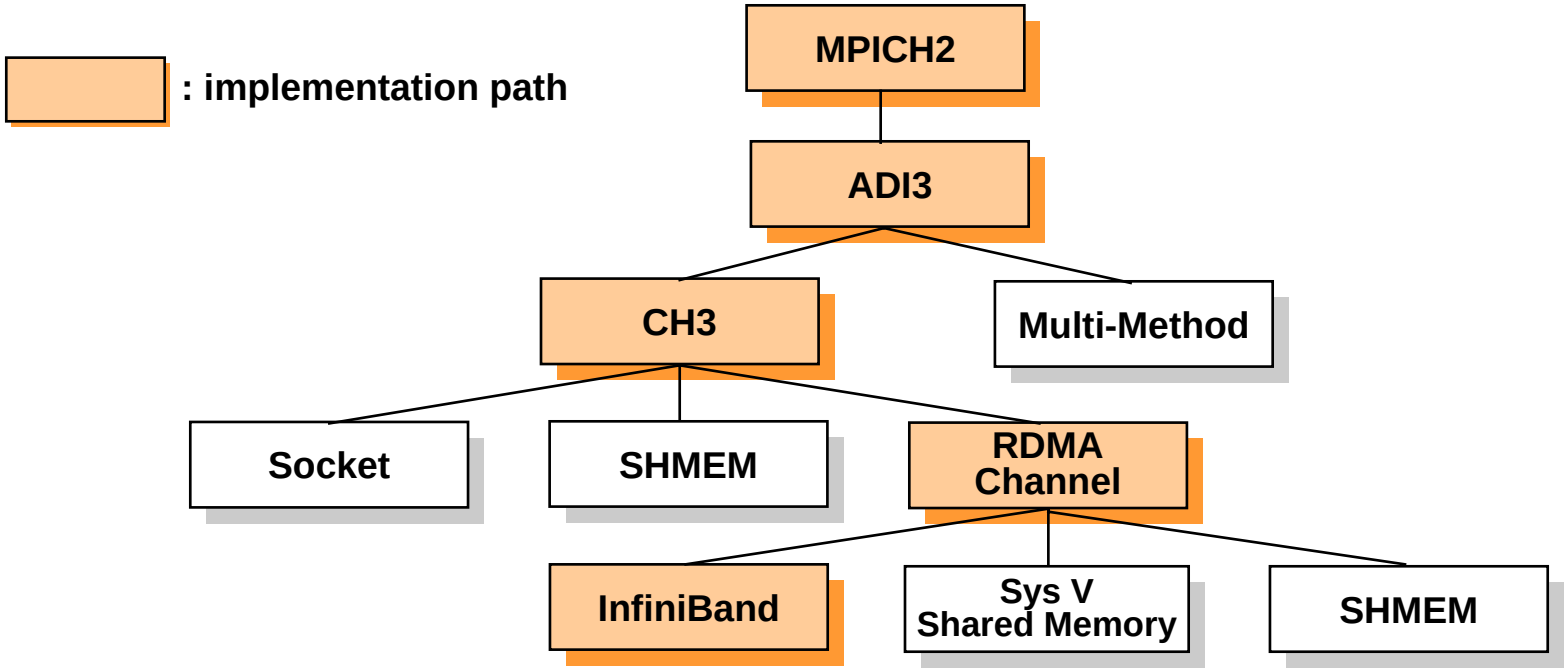
Synchronization Overhead Comparisons





Test&Set shows the least synchronization overhead and the best concurrency

Our Current Plan on MVAPICH2



- RDMA channel provides limitations
  - communication overhead is higher
  - shared memory support is not there
  - multicast can not be used
- Move our designs (two-sided and one-sided) to ADI3 layer and unify it with MVAPICH
- MVAPICH2 will have all benefits and performance as that of MVAPICH + One-sided

Presentation Overview

- Networking and I/O Requirements for HPC Clusters and Datacenters
- InfiniBand (IBA): Why, What, and Current State
- Designing High Performance MPI for IBA
  - MVAPICH and MVAPICH2
  - Latest performance numbers
  - Research challenges
- Designing Multi-Tier Datacenters with IBA, latest performance numbers, and challenges
- Conclusions

Research Challenges

- Scalability to 10,000 nodes and beyond for designing ultra-scale systems
- Efficient buffer management
  - IBA shared recv queue
  - IBA atomic operations
  - Efficient polling techniques
  - Mem-free adapters and its load on memory sub-system and PCI-Express/HT interfaces
- Connection management
  - RC vs. UD vs. RD
  - Static vs. dynamic

Research Challenges (Cont'd)

- Collectives
  - Efficient and reliable schemes to use multicast support
    - Hardware vs. software support for reliable multicast
  - Management of multicast groups
    - Dynamic creation
    - Interaction with Subnet and Connection Manager
  - Taking advantage of multicast and atomic support for
    - Barrier
    - All-reduce
  - Taking advantage of Service Level (SL) support for Collectives
  - Topology-aware collectives
  - Extended collectives (inter-communicator) in MPI-2

Research Challenges (Cont'd)

- Multi-rail Support for designing large-scale fault-tolerant systems
  - Study of Automatic Path Migration (APM) technique together with Subnet Manager (SM) and Connection Manager (CM)
  - Efficient usage of LMCs and adaptive routing in large fabric
  - Effective usage of Virtual Lanes, Service Levels and partitioning techniques and their integration with the MPI stack, connection manager, and scheduler for efficient resource allocation

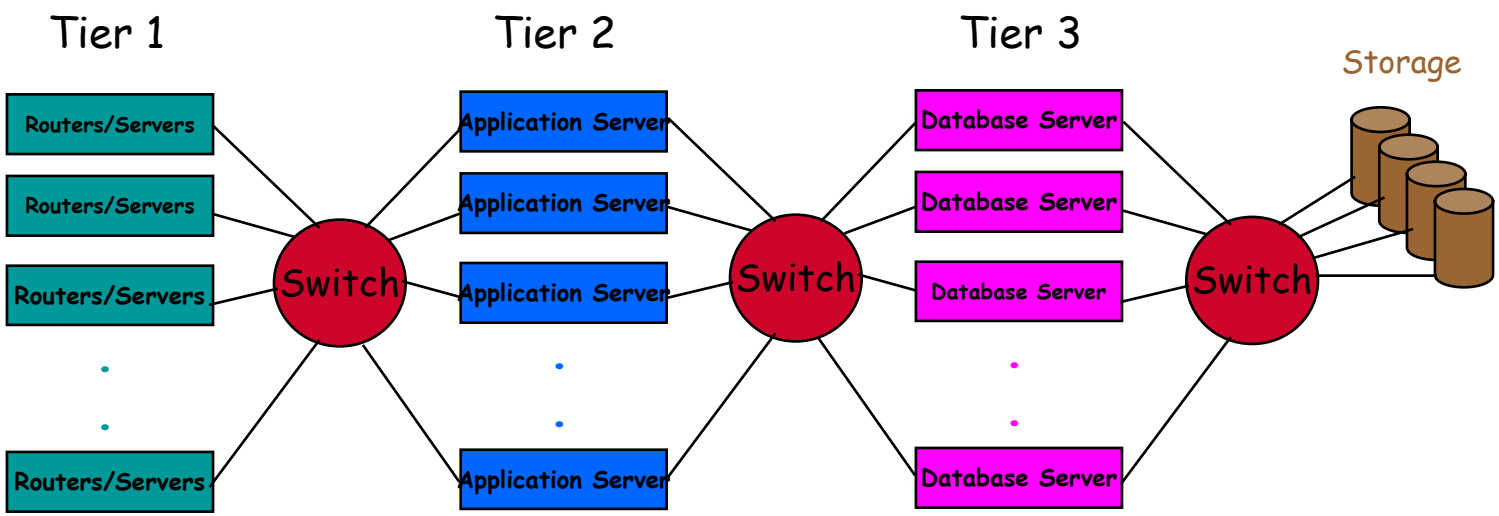
Research Challenges (Cont'd)

- MPI-2 One-sided Communication
  - Impact of active vs. passive target synchronization on applications
  - Supporting with shared-memory Intra-node communication
    - Kernel-level
    - NUMA-aware
- Collectives with one-sided communications
- Support for Datatype communication

Presentation Overview

- Networking and I/O Requirements for HPC Clusters and Datacenters
- InfiniBand (IBA): Why, What, and Current State
- Designing High Performance MPI for IBA
  - MVAPICH and MVAPICH2
  - Latest performance numbers
  - Research challenges
- Designing Multi-Tier Datacenters with IBA, latest performance numbers, and challenges
- Conclusions

Data Centers – Issues and Challenges

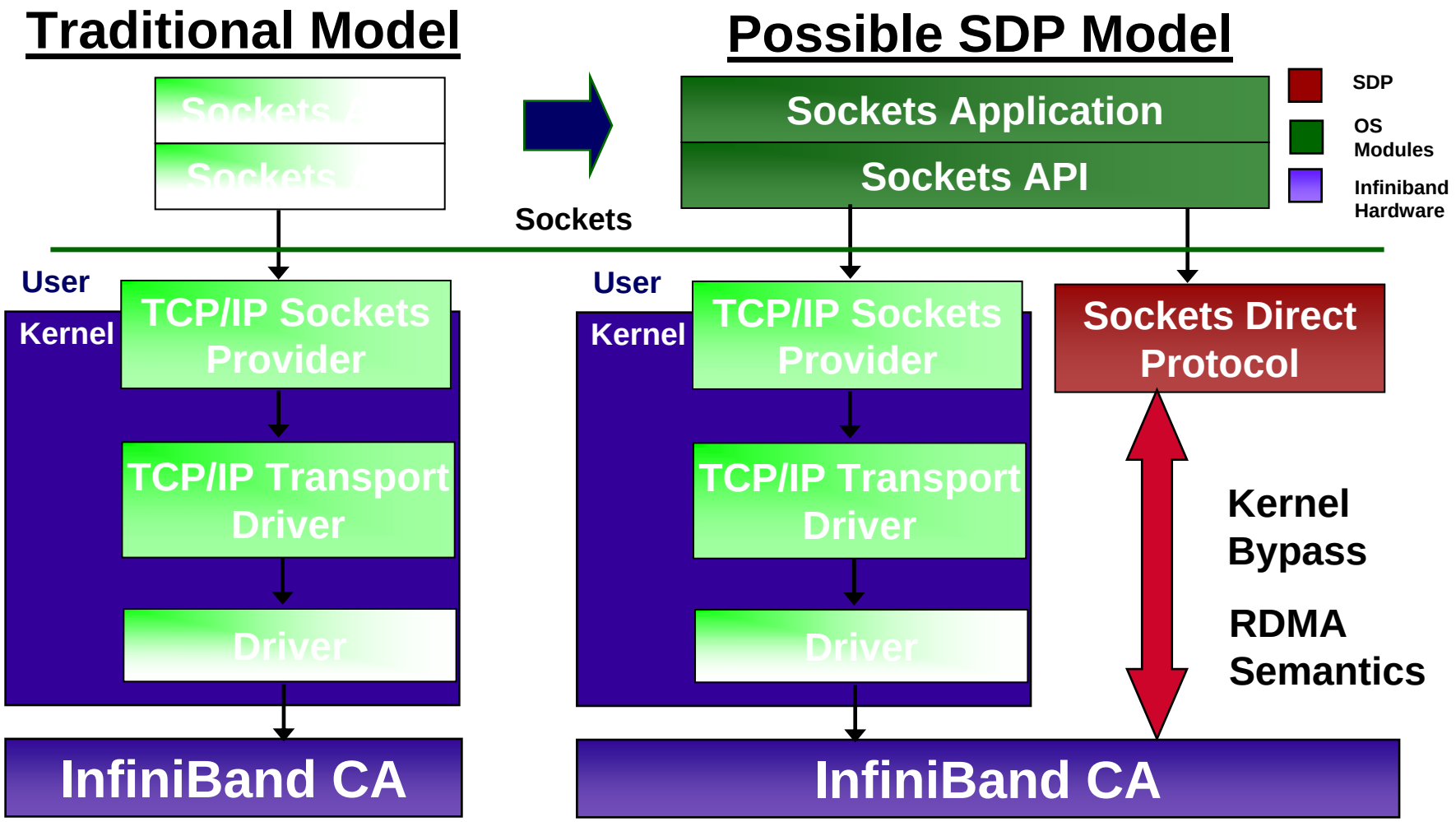


- Client requests come over TCP (WAN)
- Traditionally TCP requests have been forwarded through multi-tiers
- Higher response time and lower throughput
- Can performance be improved with IBA?
  - High Performance TCP-like communication over IBA
  - TCP Termination

Sockets Direct Protocol (SDP) on IBA

- SDP increases efficiency by providing
  - ✓ Protocol offload
  - ✓ Zero copy – use of RDMA reads/writes
  - ✓ OS/Kernel bypass
  - ✓ Interrupt avoidance
  - ✓ Reliable in-order delivery in hardware
  - ✓ Transparent to applications

SDP Architectural Model

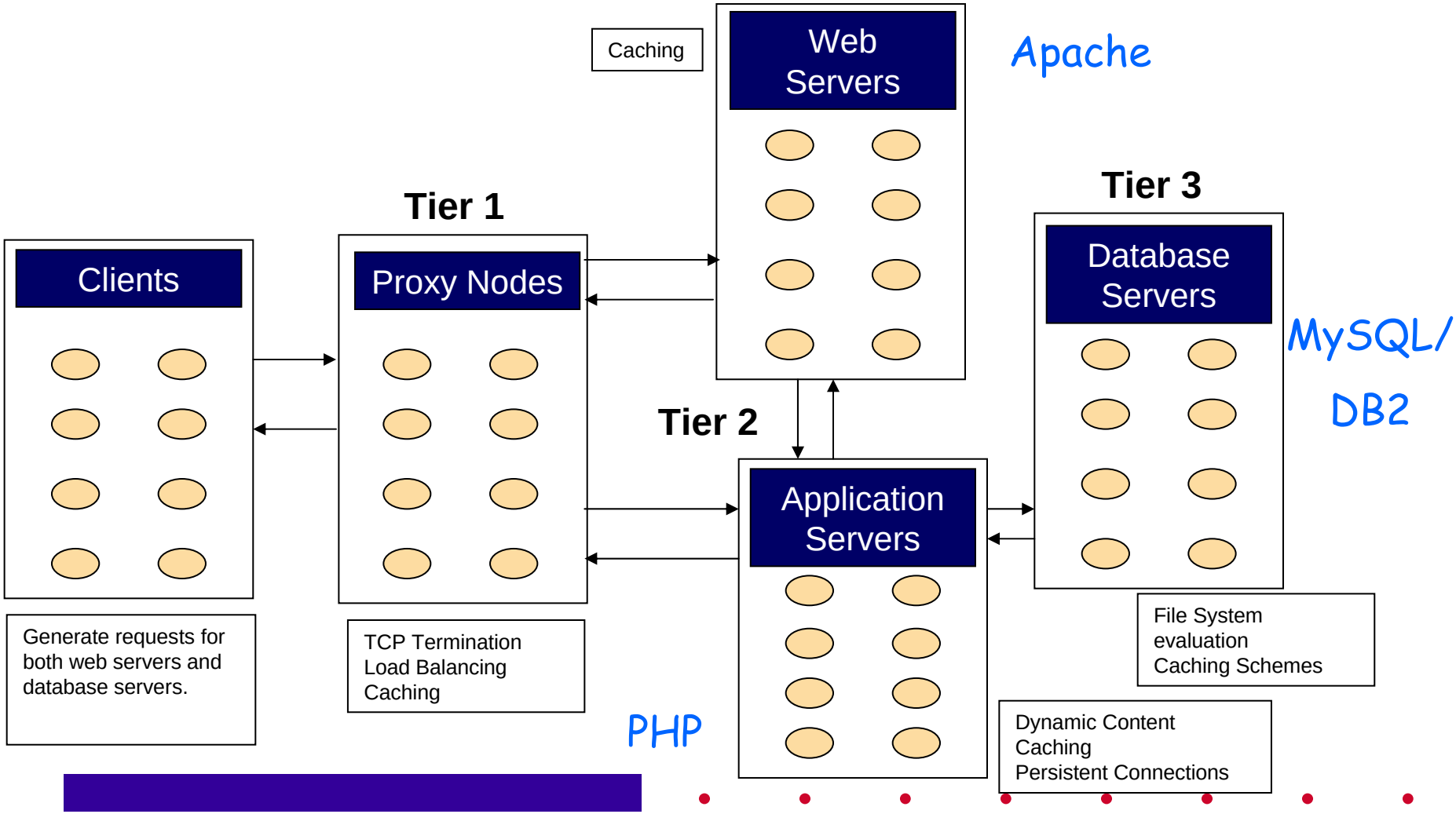


Source: Infiniband[SM] Trade Association 2002

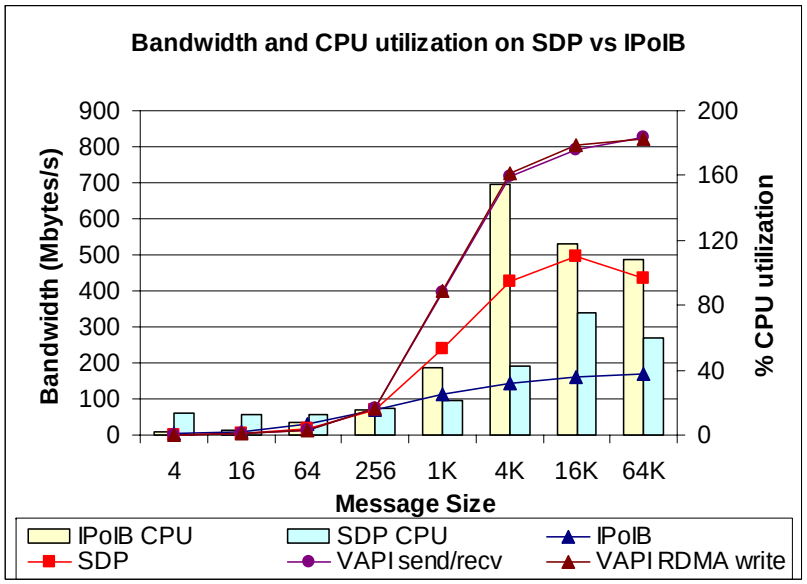
Our Objectives

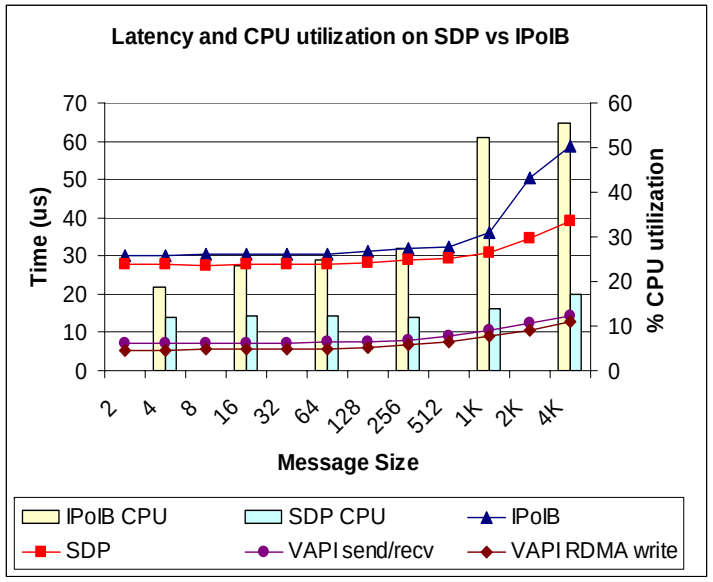
- To study the importance of the communication layer in the context of a multi-tier data center
  - Sockets Direct Protocol (SDP) vs. IPoIB
- Explore whether IBA mechanisms can help designing various components of datacenter efficiently
  - Web Caching/Coherency
  - Re-configurability and QoS
  - Load Balancing, I/O and File systems, etc.
- Studying workload characteristics
- In memory databases

3-Tier Datacenter Testbed at OSU



SDP vs. IPoIB: Latency, Bandwidth, and CPU Utilization (3.0 GHz PCI-X)

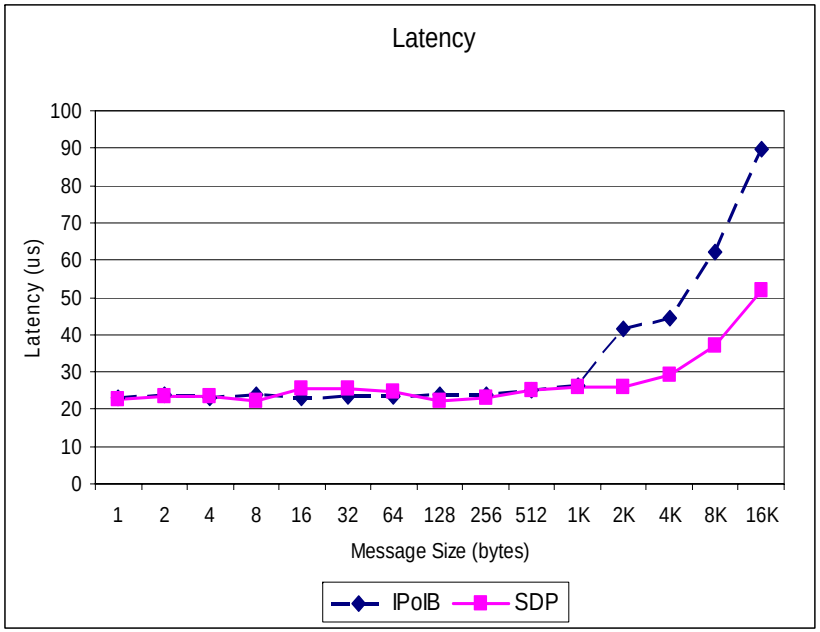


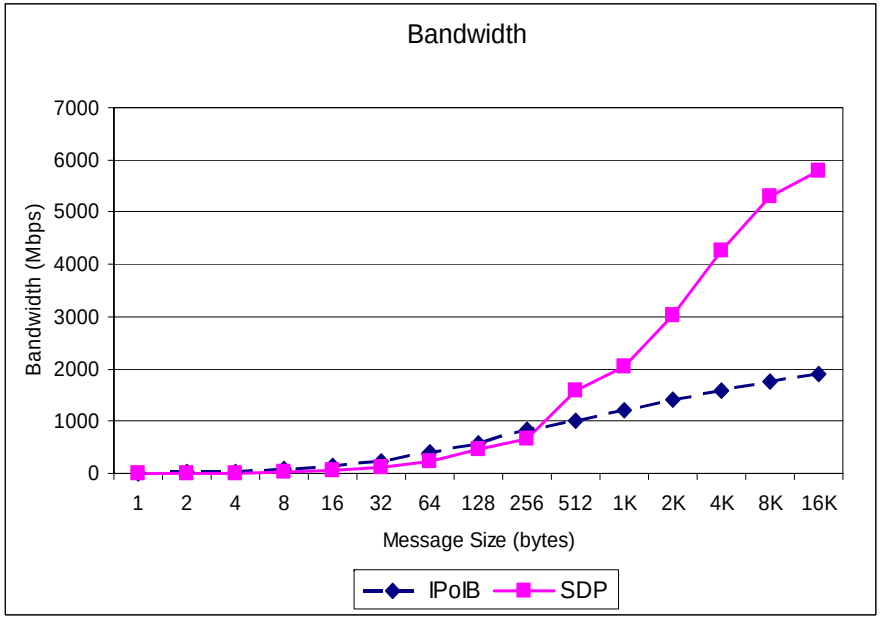


SDP enables high bandwidth (up to 500 MBytes/sec or 4000 Mbps), low latency (27 μs) message passing with low CPU utilization

P. Balaji, S. Narravul, K. Vaidyanathan, S. Krishnamoorthy, J. Wu, and D. K. Panda, Sockets Direct Protocol over InfiniBand: Is it Beneficial?, Presented at ISPASS'04, March 2004.

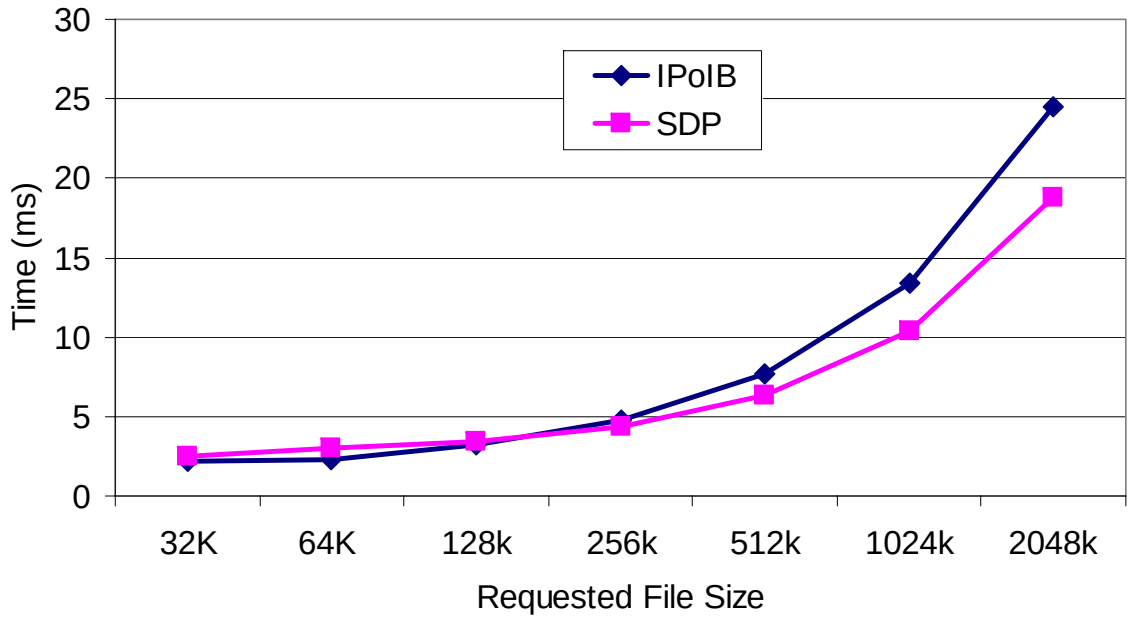
SDP vs. IPoIB: Latency and Bandwidth (3.4 GHz PCI-Express)





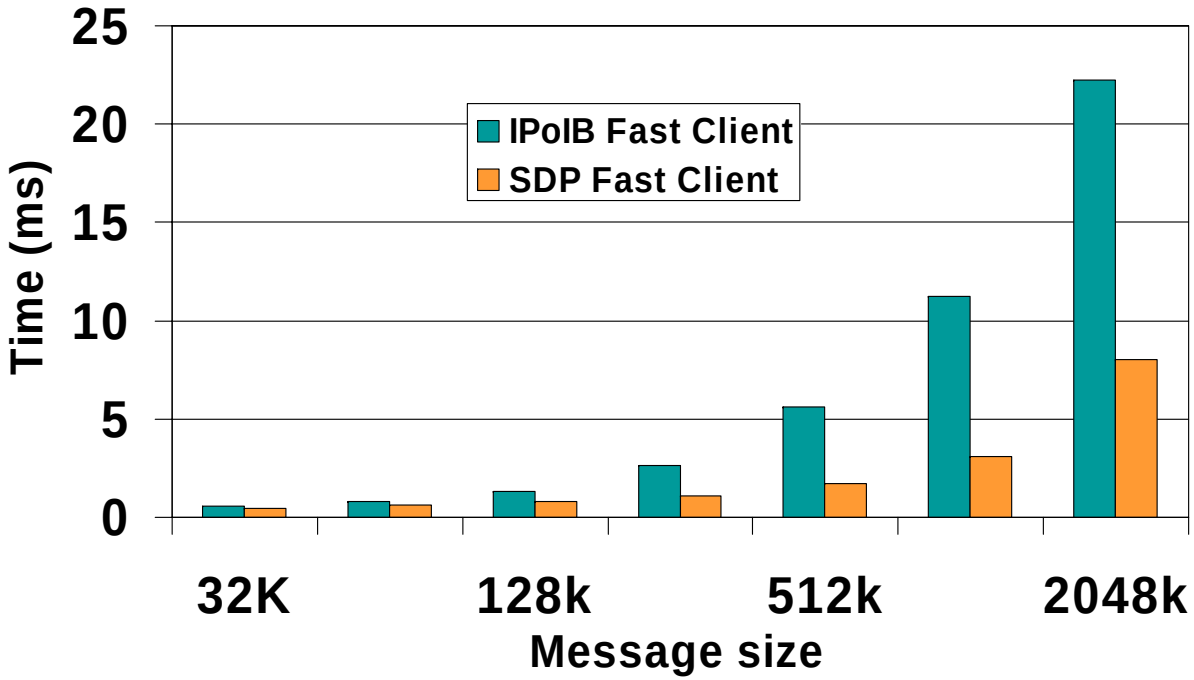
SDP enables high bandwidth (up to 750 MBytes/sec or 6000 Mbps), low latency (21 μs) message passing

Datacenter Response Time



- Response time at the client (client over IPoIB)
- SDP performs better for larger file sizes
- For small messages, dynamic connection set-up time becomes the bottleneck

Web Server response time



- Response time at the Web Server for Fast Clients
- SDP takes less than 50% time as compared to IPoIB

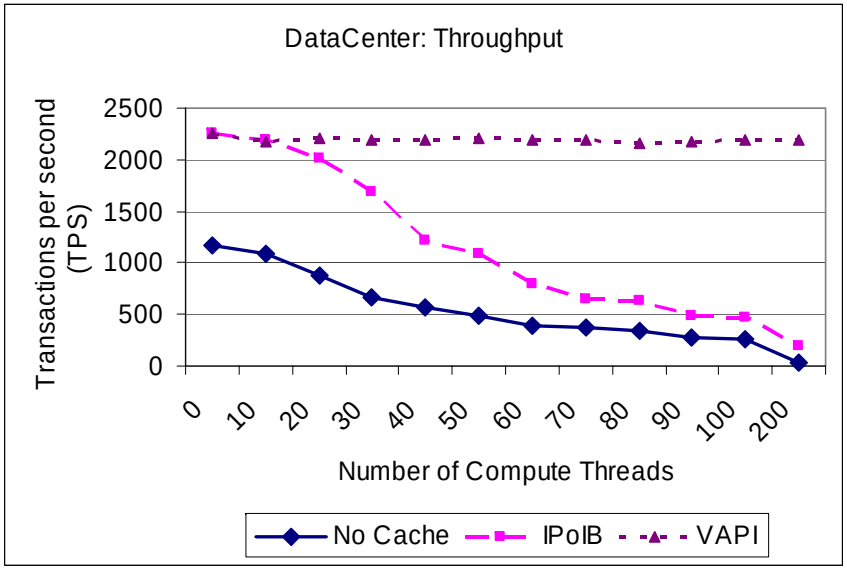
Strong Cache Coherency

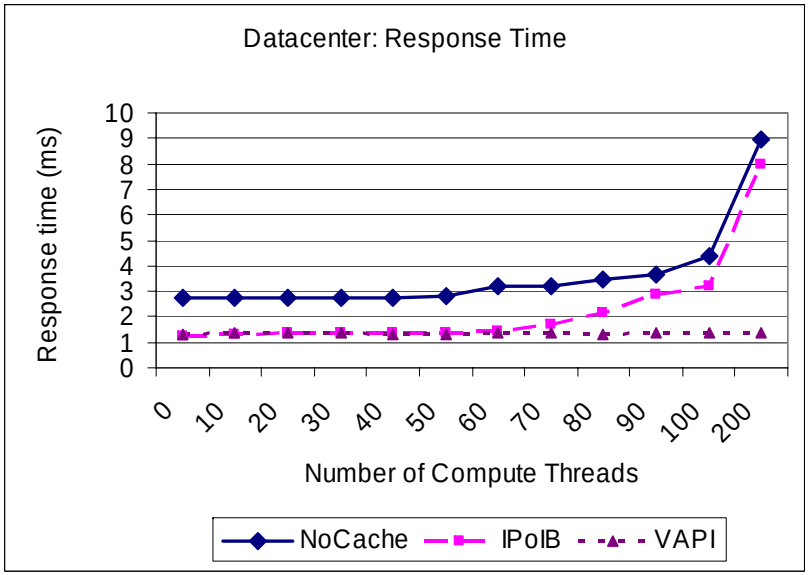
- Dynamic data caching – challenging!
- Cache Consistency and Coherence
  - Become more important than in static case
- Cache Consistency
  - Non-decreasing views of system state
  - Updates seen by all or none
- Cache Coherency
  - Refers to the average staleness of the document served from cache
  - Updates seen in a unique order at all nodes
  - Two models of coherence
    - Bounded staleness models (TTL, invalidation etc)
    - Strong or immediate coherence models
- Can we improve this with RDMA polling?

Strong Cache Coherency with Client Polling?

	Consistency	Coherency
No Caching	✓	✓
Client Polling	✓	✓✓
Invalidation	✓	✗
TTL	✗	✗

Strong Cache Coherency with RDMA Polling: Datacenter Performance





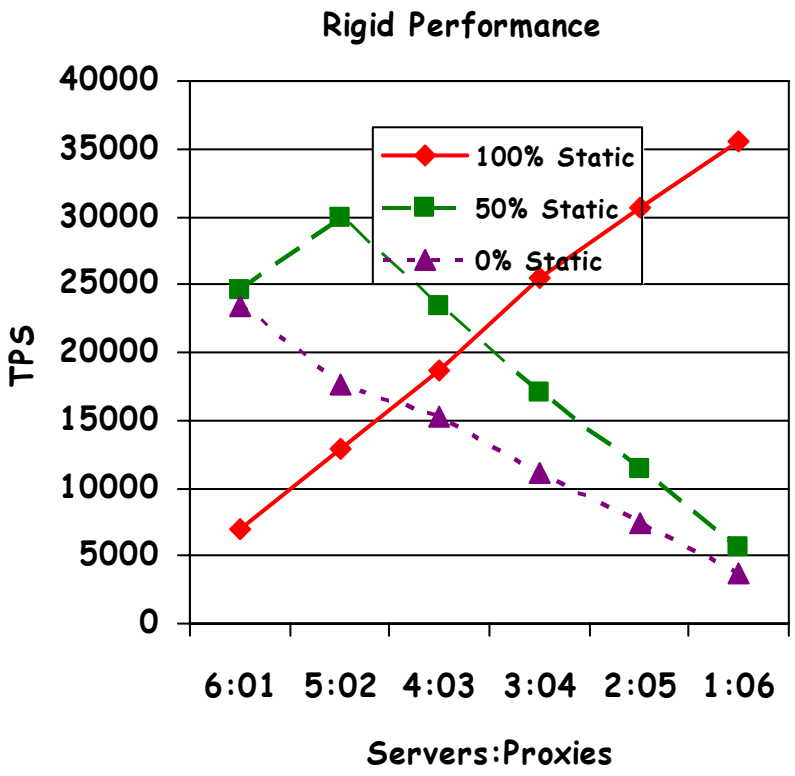
The VAPI module can sustain performance even with heavy load on the back-end servers

S. Narravul, P. Balaji, K. Vaidyanathan, S. Krishnamoorthy, J. Wu, and D. K. Panda, Supporting Strong Cache Coherency for Active Caches in Multi-Tier Data-Centers over InfiniBand, Presented at SAN'04, Feb 2004

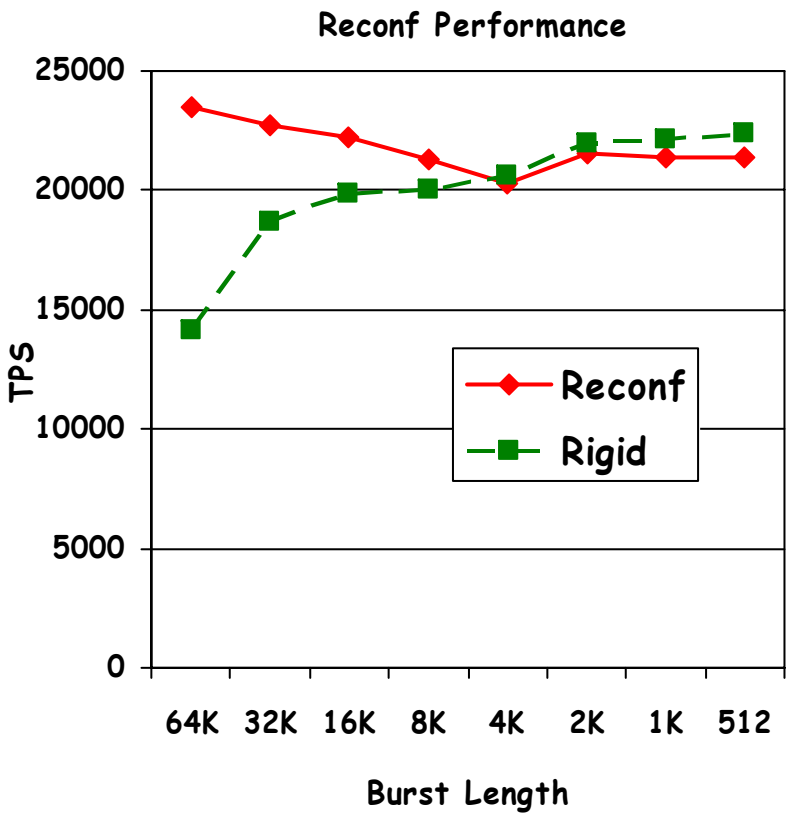
Dynamic Re-configurability and QoS

- More datacenters are using dynamic data
- How to decide the number of proxy nodes vs. application servers
- Current approach
  - Use a fixed distribution
  - Incorporate over-provisioning to handle dynamic data
- Can we design a dynamic re-configurability scheme with shared state using RDMA operations?
  - To allow a proxy node work like app server and vice versa as needed
  - Allocates resources as needed
  - Over-provisioning need not be used
- Scheme can be used for multi web-site hosting servers

Dynamic Re-configurability with Shared State using RDMA Operations

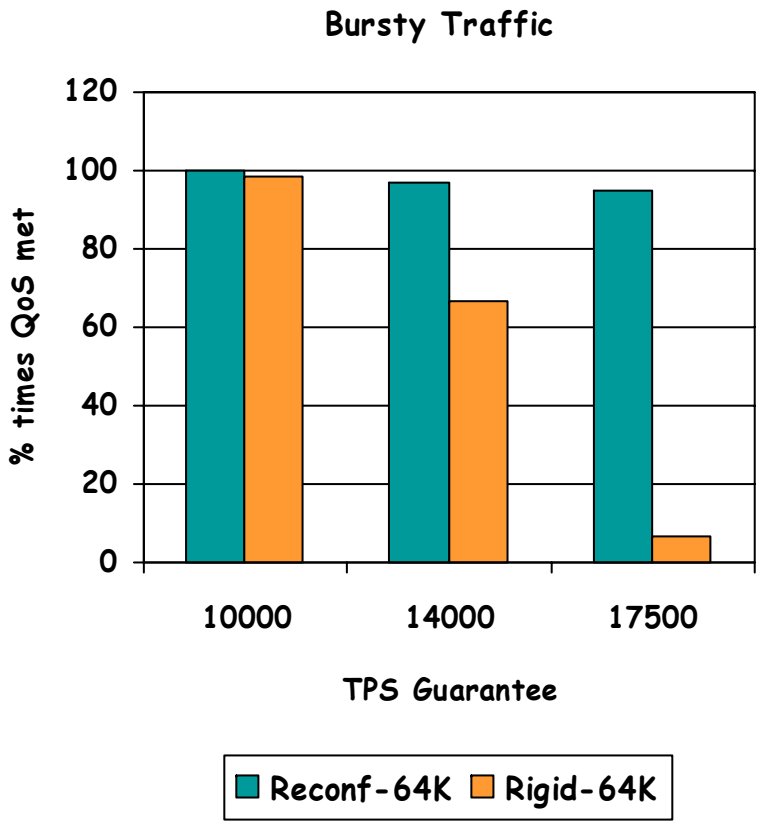


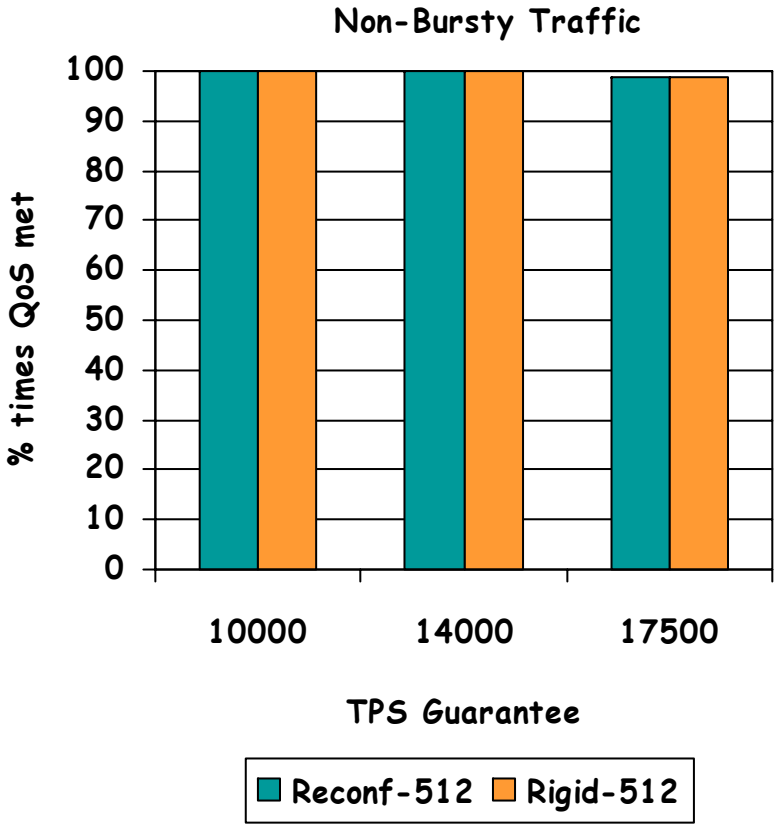
Different kinds of requests have different architectural requirements



Re-configurability dynamically allocates resources depending on the requests

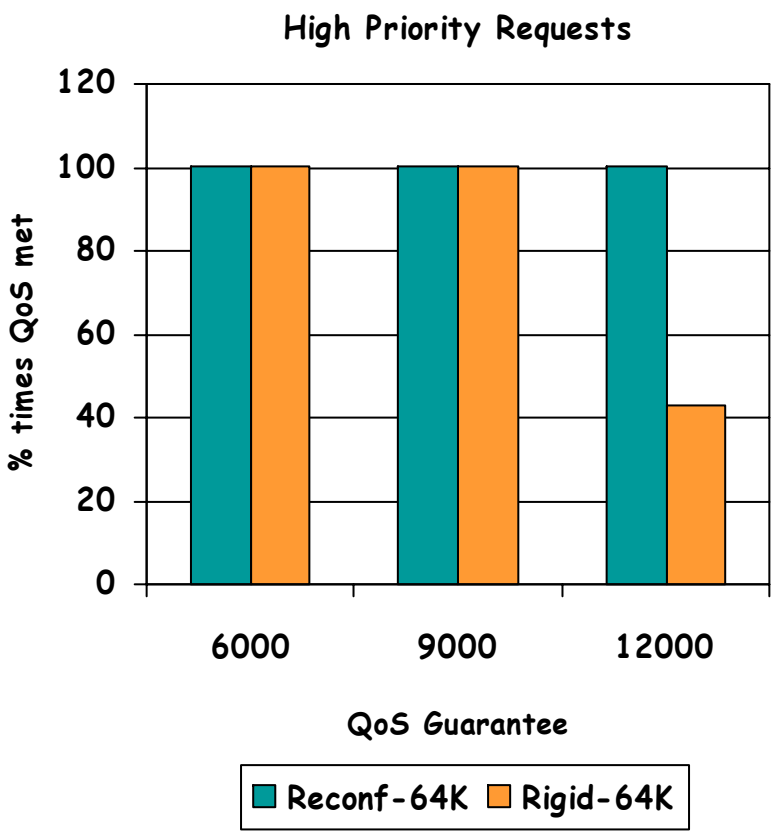
QoS meeting capabilities

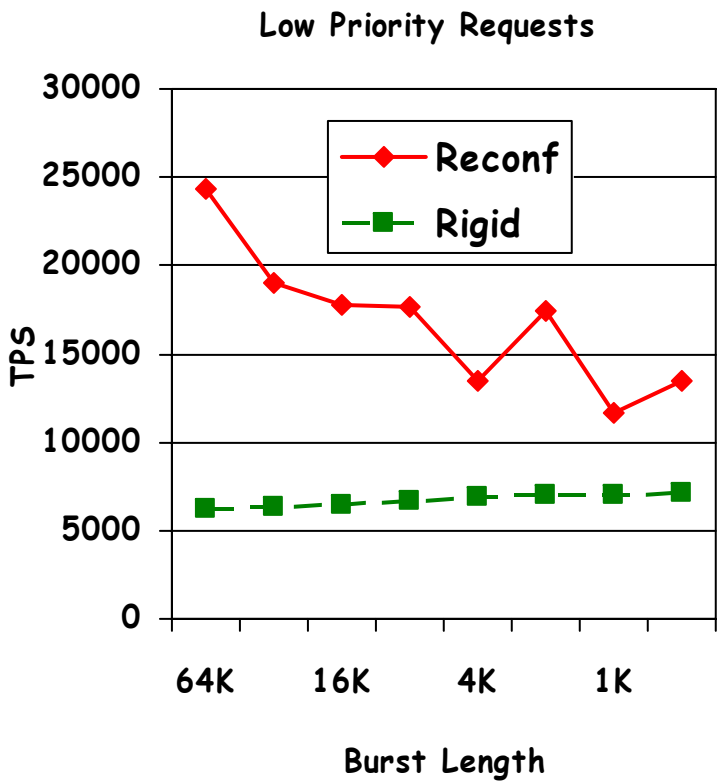




Re-configurability handles bursty traffic efficiently without affecting non-bursty traffic

Against Over-Provisioning for Multi-Website hosting servers





Better QoS guarantees for High priority requests; Better Performance for Low priority requests

Datacenter: Future Plans

- Study the impact with
  - Updated versions of SDP (Voltaire and/or OpenIB stack)
  - Detailed Evaluation on PCI-Express platforms
- Design efficient schemes for other components
  - Load Balancing
  - I/O and File systems, etc.
- Studying workload characteristics
- Designing in-memory databases with IBA features (RDMA and atomic)

Presentation Overview

- Networking and I/O Requirements for HPC Clusters and Datacenters
- InfiniBand (IBA): Why, What, and Current State
- Designing High Performance MPI for IBA
  - MVAPICH and MVAPICH2
  - Latest performance numbers
  - Research challenges
- Designing Multi-Tier Datacenters with IBA, latest performance numbers, and challenges
- Conclusions

Conclusions

- Provided a brief overview of InfiniBand
- Presented of the issues and solutions in designing HPC clusters and datacenters with IBA
- IBA (with PCI-Express and HT interfaces) is coming up as an entire new networking technology
- Will lead to design and deployment of a new generation of high-end computing systems, storage, servers, and datacenters
- Many open research issues in designing these systems by taking advantage of the IBA features …. will need challenging solutions

Web Pointers

NBC home page

http://www.cse.ohio-state.edu/~panda/
http://nowlab.cis.ohio-state.edu/

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