



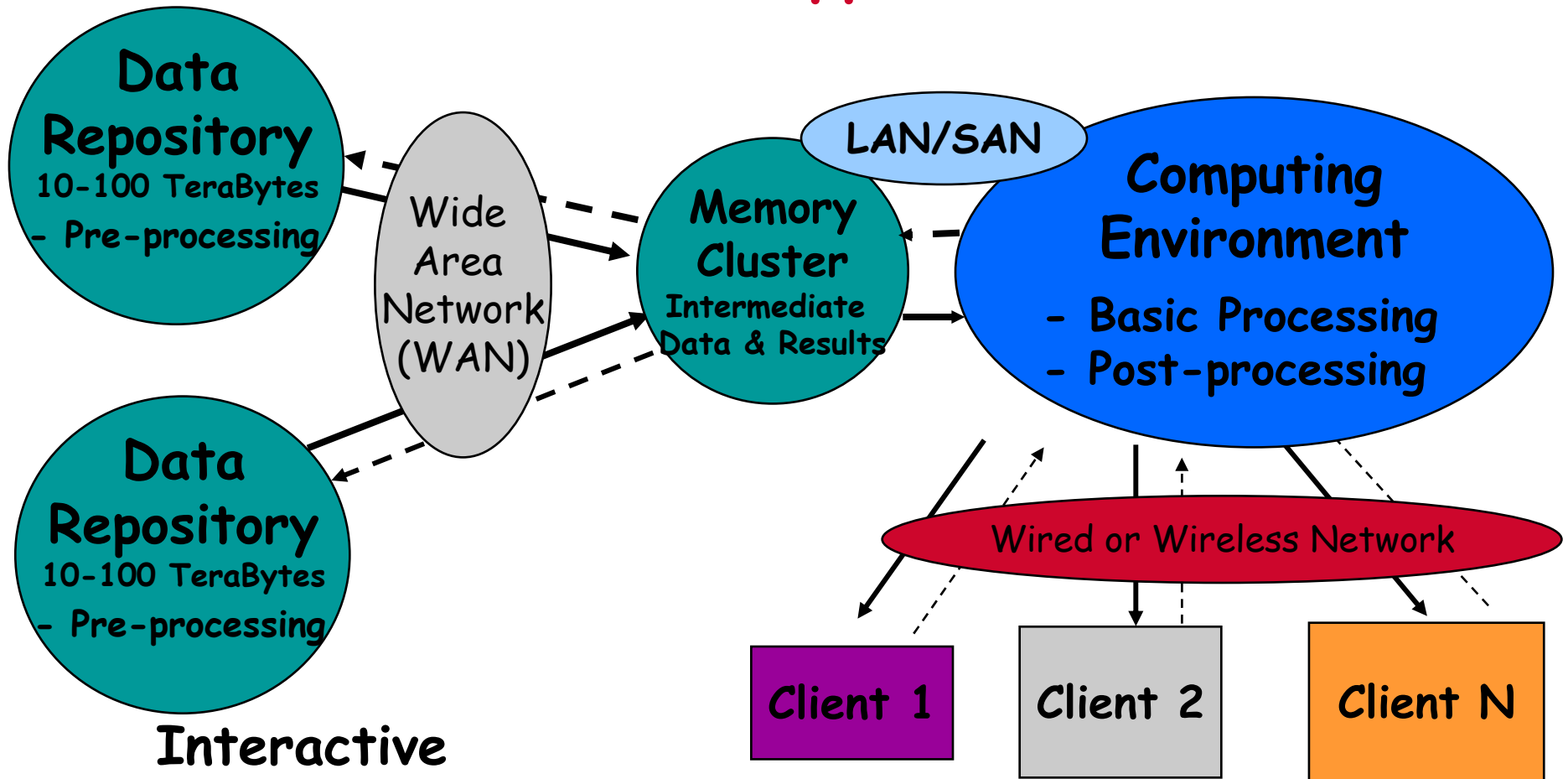
Presentation Overview



- Networking and I/O Requirements for Distributed Data-Intensive Applications
- Emerging Networking Technology and Protocols
 - InfiniBand
 - Sockets Direct Protocol
 - RDMA over IP
- Systems and Networking Research at OSU
 - High Performance MPI with InfiniBand for Clusters
 - Parallel File Systems
 - Multi-Tier Data Center
- Overview of NSF Research Infrastructure (RI) Grant
- Conclusions



Vision for the Next Generation Architecture to Handle Distributed Data- Intensive Applications



**Interactive
Collaborative**

End-to-end QoS



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
 - Parallel processing of image data
- Good Storage Area Networks high performance I/O
 - High performance file systems
- Good WAN connectivity and protocols
 - Distributed data access and operations
- Quality of Service (QoS) for supporting interactive applications
- With low cost and high performance





Presentation Overview



- Networking and I/O Requirements for Distributed Data-Intensive Applications
- Emerging Networking Technology and Protocols
 - InfiniBand
 - Sockets Direct Protocol
 - RDMA over IP
- Systems and Networking Research at OSU
 - High Performance MPI with InfiniBand for Clusters
 - Parallel File Systems
 - Multi-Tier Data Center
- Overview of NSF Research Infrastructure (RI) Grant
- 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.





Rich Set of features of InfiniBand



- 10.0-60.0 Gigabit Networking Technology
- Channel Semantic
 - Send and Recv
- Memory Semantic
 - RDMA read
 - RDMA write
 - RDMA atomic operations (e.g. Fetch & Add, Compare & swap)
- 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
- Many more features
- Low cost
 - Current network adapter is around \$600 (targeted to come down to \$100 by the end of this year)



⋮

Sockets Direct Protocol (SDP) on IBA

Existing applications written with Sockets semantics to get performance benefits

- 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

•
•

RDMA over IP

- Takes the memory semantics ideas from InfiniBand
 - RDMA read
 - RDMA write
- Targeted for WAN to deliver high-levels of communication performance
- Can work with Ethernet
 - Gigabit Ethernet
 - upcoming 10.0 Gigabit Ethernet



Presentation Overview



- Networking and I/O Requirements for Distributed Data-Intensive Applications
- Emerging Networking Technology and Protocols
 - InfiniBand
 - Sockets Direct Protocol
 - RDMA over IP
- **Systems and Networking Research at OSU**
 - **High Performance MPI with InfiniBand for Clusters**
 - Parallel File Systems
 - Multi-Tier Data Center
- Overview of NSF Research Infrastructure (RI) Grant
- Conclusions





High Performance MPI over IBA



- Facilitates scalable parallel processing on clusters connected with InfiniBand
- MPI (Message Passing Interface) is a common programming model for parallel processing
- Designed and developed a high performance MPI by taking advantage of InfiniBand features
 - MVAPICH - MPI-1 standard
 - MVAPICH2 - MPI-2 standard
- **Distributed as Open-source**
- 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 180 organizations and industry (over 25 countries)
- Available in the software stack distributions of IBA vendors (including IBGold CD)

National Labs/Research Centers

Alabama Supercomputer Center
Argonne National Laboratory
AWI Polar and Marine Research Center (Germany)
Cornell Theory Center
Center for High Performance Computing,
Univ. of New Mexico
Center for Mathematics and
Computer Science (The Netherlands)
CEA (France)
CERN, European Organization for
Nuclear Research (Switzerland)
CINES, National Computer Center of Higher
Education (France)
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)
Sandia National Laboratory
SARA Dutch National Computer Center (The Netherlands)
Science Applications International Corporation
United Institute of Informatics Problems (Belarus)
U.S. Census Bureau

•
•

MVAPICH/MVAPICH2 Users: Universities

Engineers School of Geneva (Switzerland)
Georgia Tech
Gdansk 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 Erlangen-Nuremberg (Germany)
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 Tokyo (Japan)
Univ. of Toronto (Canada)
Univ. of Twente (The Netherlands)
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.
Remcom Inc.
RLX Technologies
Rosta Ltd. (Russia)
SBC Technologies, Inc.
Scyld Software
Scalable Informatics LLC
Scotland Electronics (Int'l) Ltd (UK)
SGI (Silicon Graphics, Inc.)
Siliquent
Simulation Technologies
SKY Computers
SmallTree communications
Streamline Computing (UK)
SUN
Sysran
Telcordia Applied Research
Telsima
Thales Underwater Systems (UK)
Tomen
Topspin
Totally Hip Technologies (Canada)
Transtec (Germany)
T-Platforms (Russia)
T-Systems (Germany)
Unisys
Vector Computers (Poland)
Voltaire
WorkstationsUK, Ltd. (UK)
Woven Systems, Inc.



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

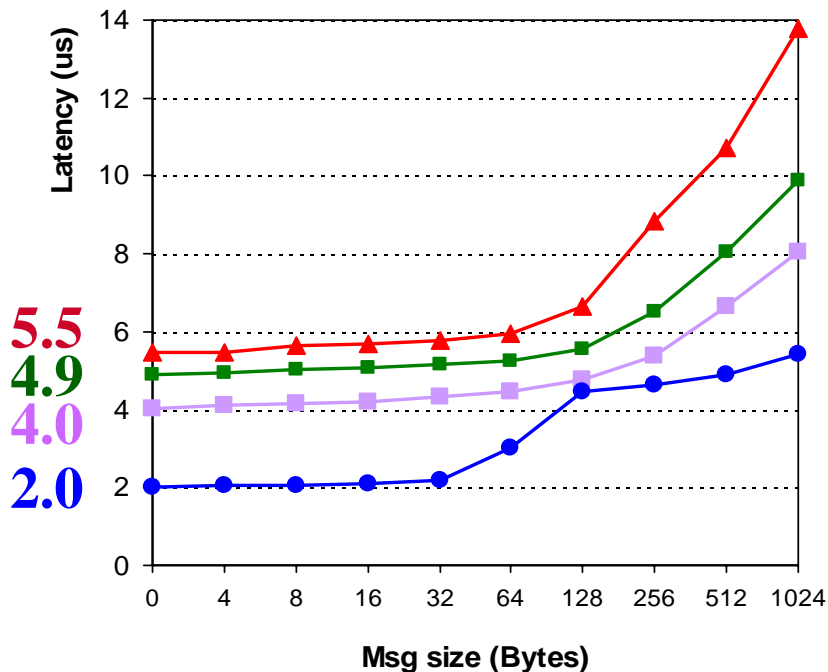


- 7th: 1100-node dual Apple Xserve 2.3 GHz cluster at Virginia Tech
- 98th: 288-node dual Opteron 2.2 GHz cluster at United Institute of Informatics Problems (Belarus)
- 211th: 192-node dual Xeon 3.06 GHz cluster at Mississippi State University
- 334th: 128-node dual Xeon 3.06 GHz cluster at Sandia/Livermore
- 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

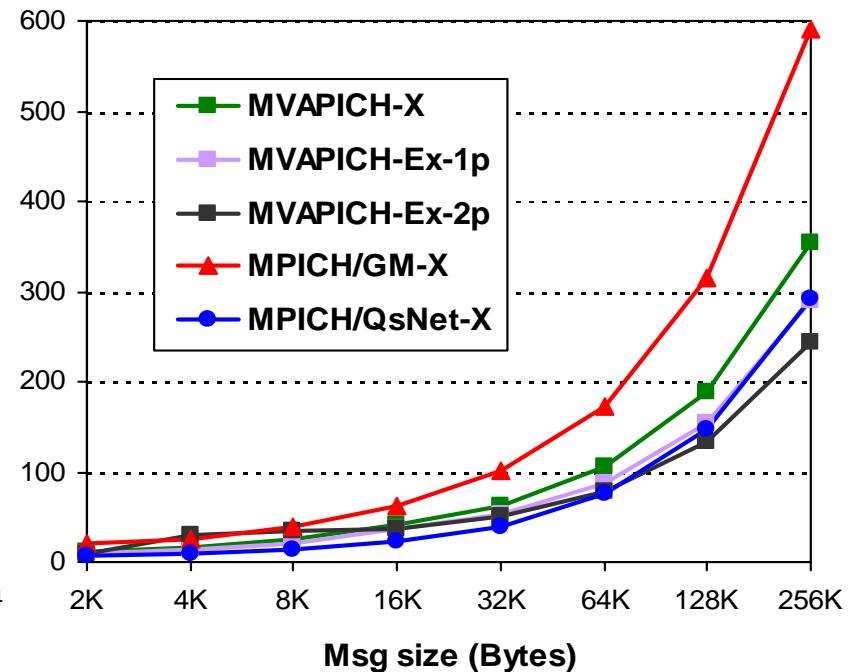


MPI-level Latency (One-way): IBA vs. Myrinet vs. Quadrics

Small message latency

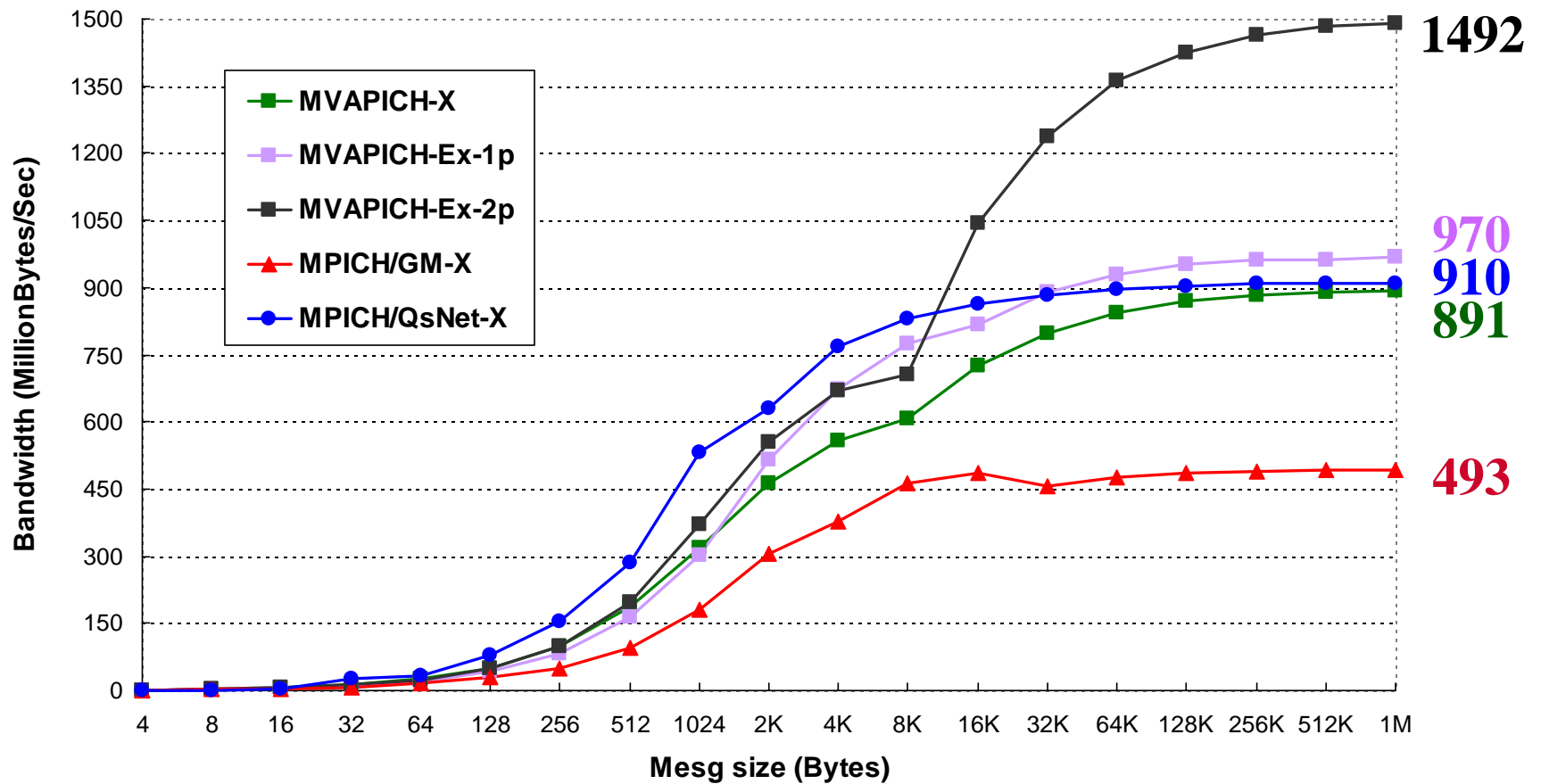


Large message latency

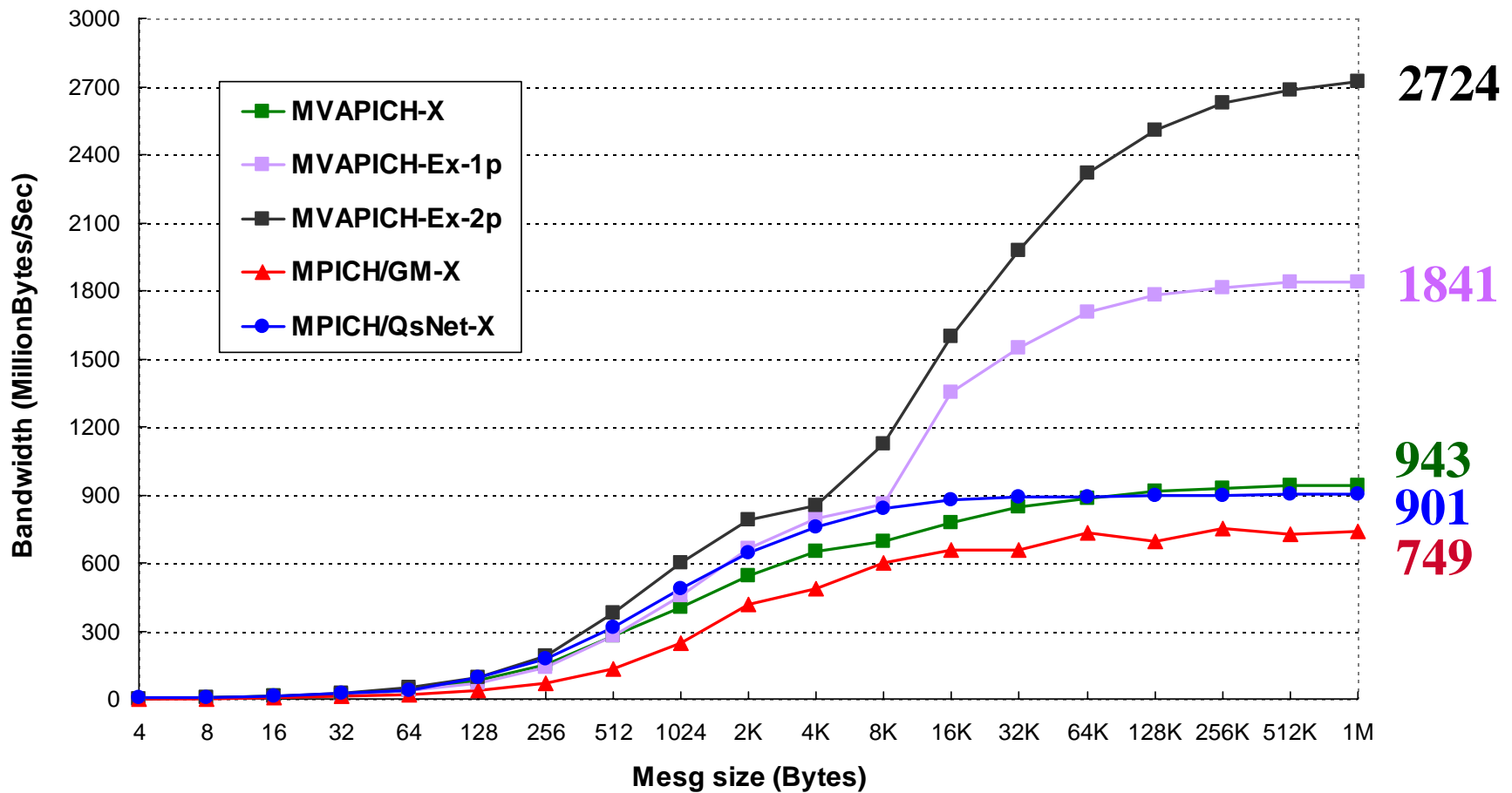


- SC '03
- Hot Interconnect '04
- IEEE Micro (Jan-Feb) '05, one of the best papers from HotI '04

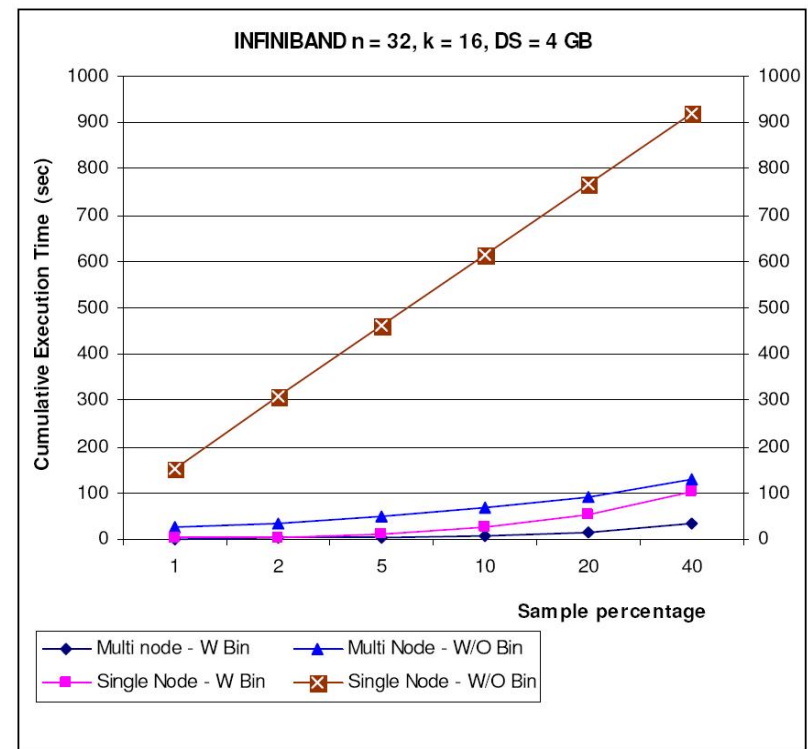
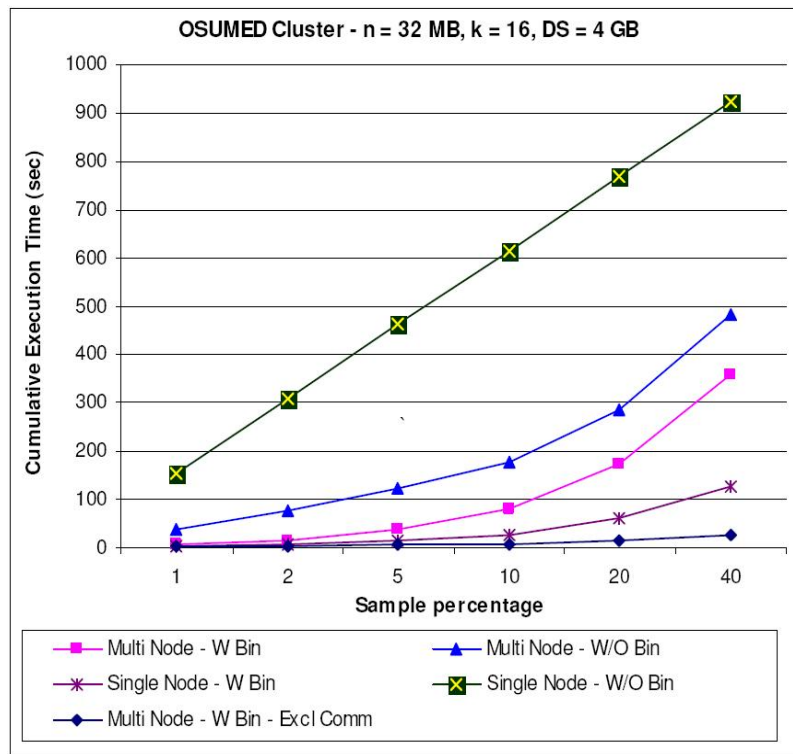
MPI-level Bandwidth (Uni-directional): IBA vs. Myrinet vs. Quadrics



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



Performance Benefits with a Data Mining Application (Courtesy Srini and his students)





Presentation Overview



- Networking and I/O Requirements for Distributed Data-Intensive Applications
- Emerging Networking Technology and Protocols
 - InfiniBand
 - Sockets Direct Protocol
 - RDMA over IP
- **Systems and Networking Research at OSU**
 - High Performance MPI with InfiniBand for Clusters
 - **Parallel File Systems**
 - Multi-Tier Data Center
- Overview of NSF Research Infrastructure (RI) Grant
- Conclusions





PVFS (PVFS-1 and PVFS-2) over InfiniBand



- A cluster file system for low-cost cluster systems
 - First version built upon TCP/IP
 - Second version is under development
 - <http://www.parl.clemson.edu/pvfs>
- To address issues to deploy IBA in cluster file systems
 - Design efficient transport layers
 - Contiguous and non-contiguous data movement
 - Communication buffer management
 - Memory registration/deregistration
- Jointly done with ANL (Rob Ross and group) and OSC (Pete)

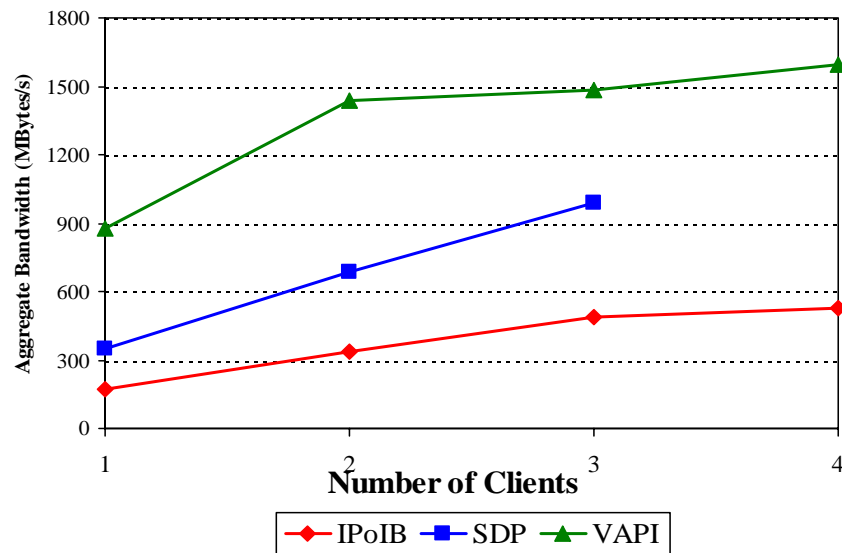
J. Wu, P. Wyckoff, and D. K. Panda, PVFS over InfiniBand: Design and Performance Evaluation, Int'l Conference on Parallel Processing (ICPP), Oct 2003.

J. Wu, P. Wyckoff, and D. K. Panda, Supporting Efficient Noncontiguous Access in PVFS over InfiniBand, Cluster Computing Conference, Dec. 2003.

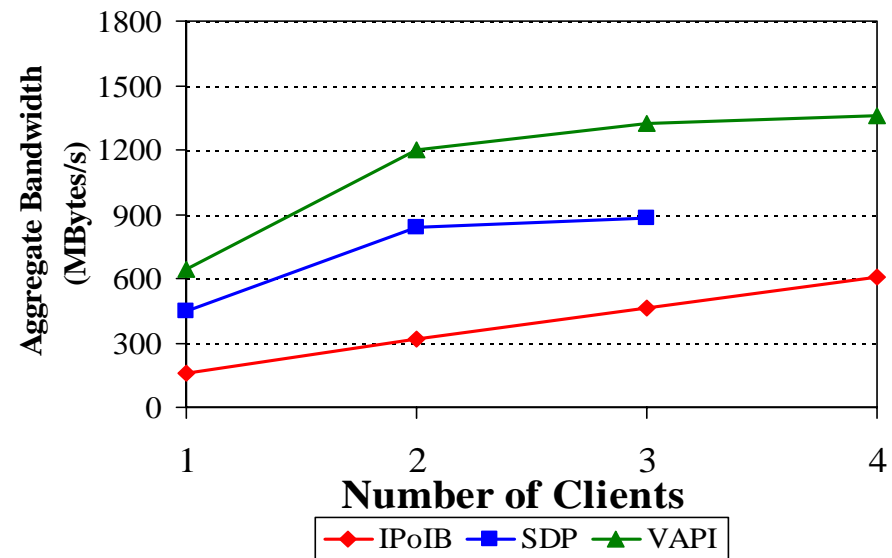


PVFS on *ramfs*

PVFS Read, 4 I/O Servers



PVFS Write, 4 I/O Servers



- PVFS over VAPI
 - 2.6 - 4.5 times faster than IPoIB
 - 1.5 - 2.5 times faster than SDP
- Memory copies in IPoIB and SDP
 - Reduced throughput and high host overhead
 - I/O servers are saturated



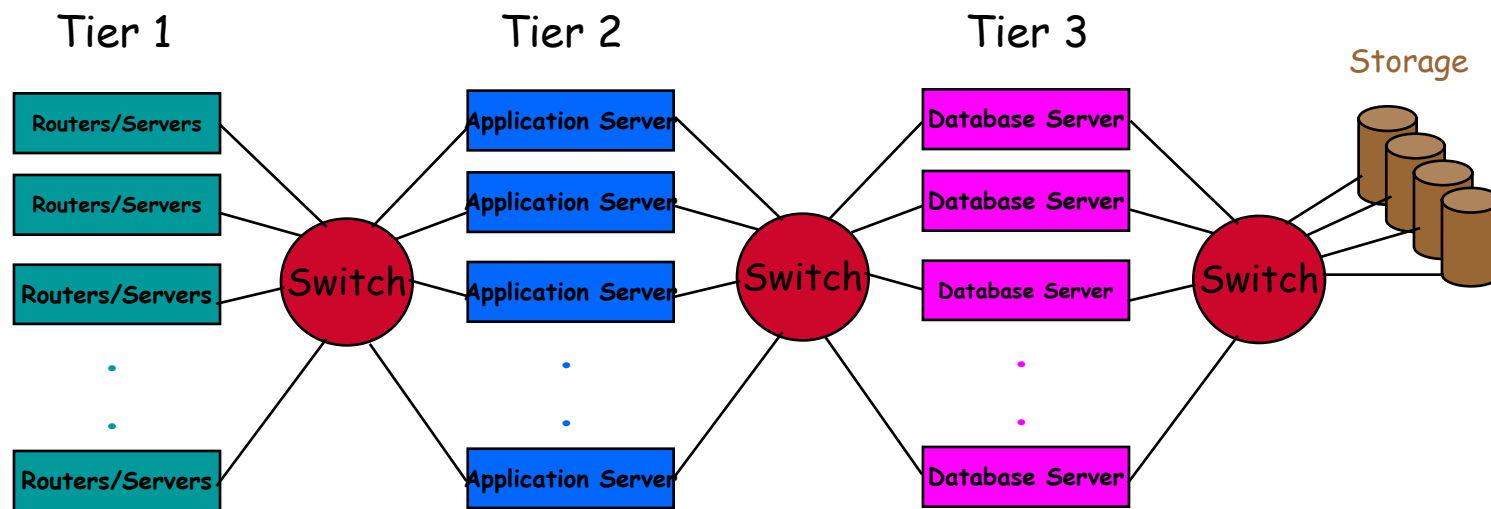
Presentation Overview



- Networking and I/O Requirements for Distributed Data-Intensive Applications
- Emerging Networking Technology and Protocols
 - InfiniBand
 - Sockets Direct Protocol
 - RDMA over IP
- **Systems and Networking Research at OSU**
 - High Performance MPI with InfiniBand for Clusters
 - Parallel File Systems
 - **Multi-Tier Data Center**
- Overview of NSF Research Infrastructure (RI) Grant
- 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



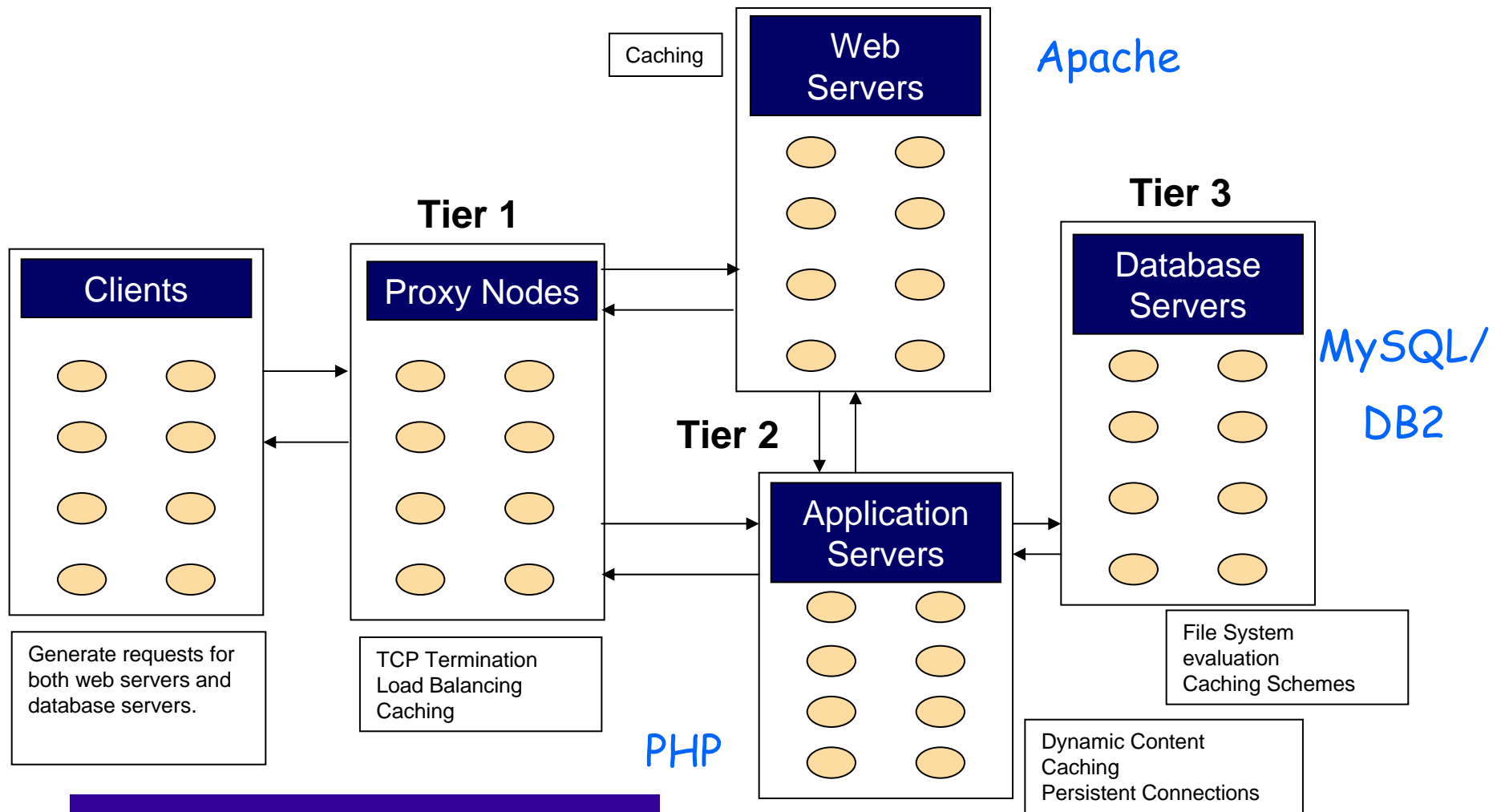
Our Objectives



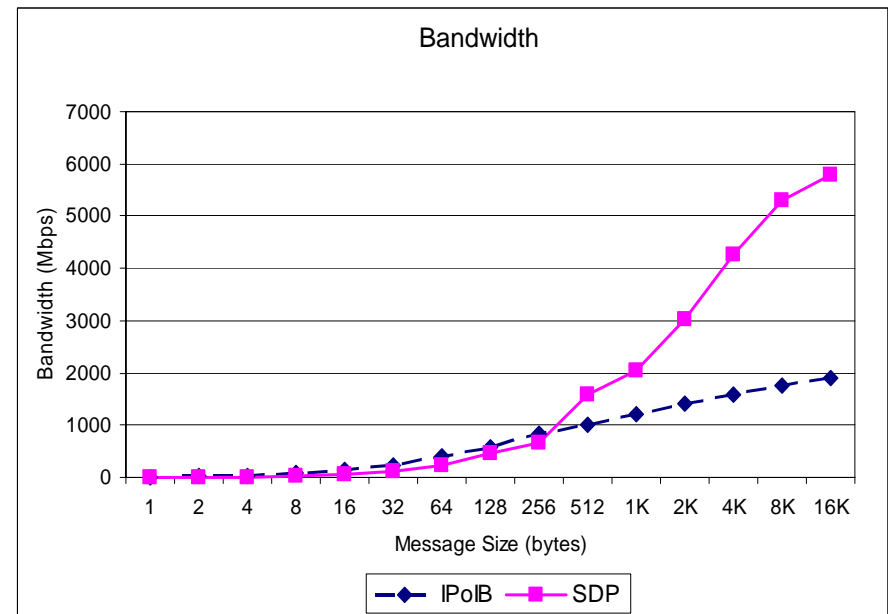
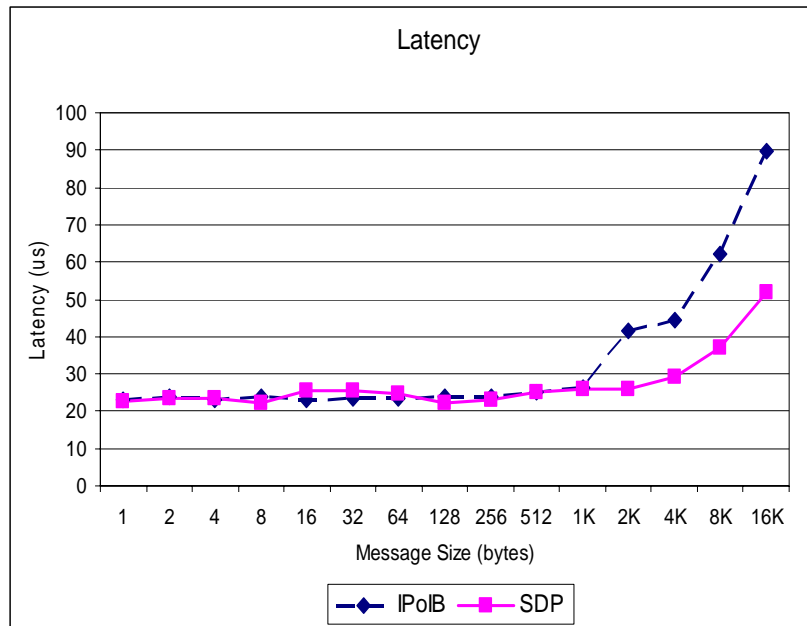
- To study the importance of the communication layer in the context of a multi-tier data center
 - Sockets Direct Protocol (SDP)
 - Explore whether InfiniBand mechanisms and features can help designing various components of datacenter efficiently
 - Web Caching/Coherency
 - Re-configurability and QoS
 - In memory databases
- 



3-Tier Datacenter Testbed at OSU

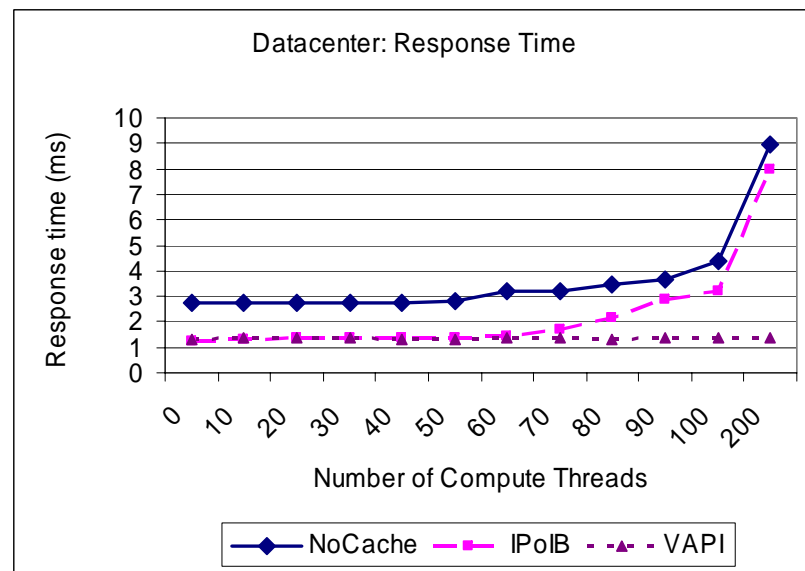
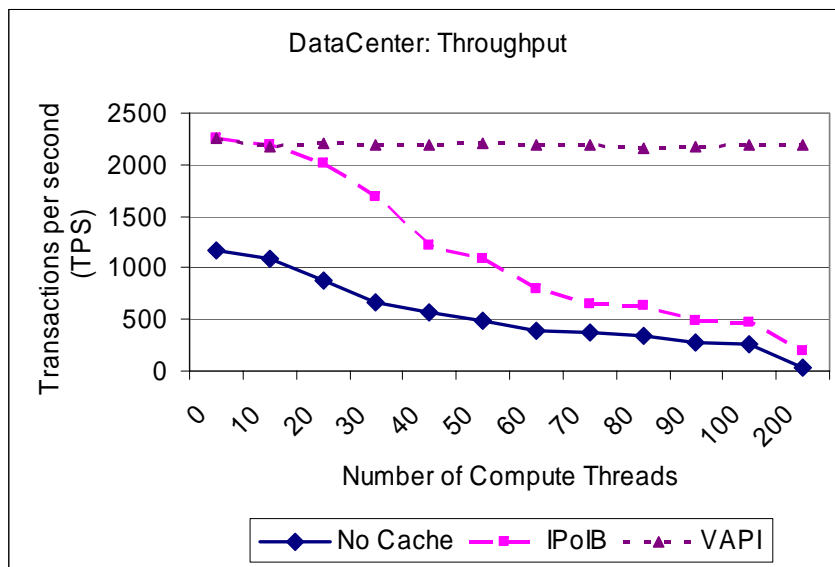


SDP vs. IPoIB: Latency and Bandwidth (3.4 GHz PCI-Express)



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

Strong Cache Coherency with RDMA Polling: Datacenter Performance with Dynamic Data



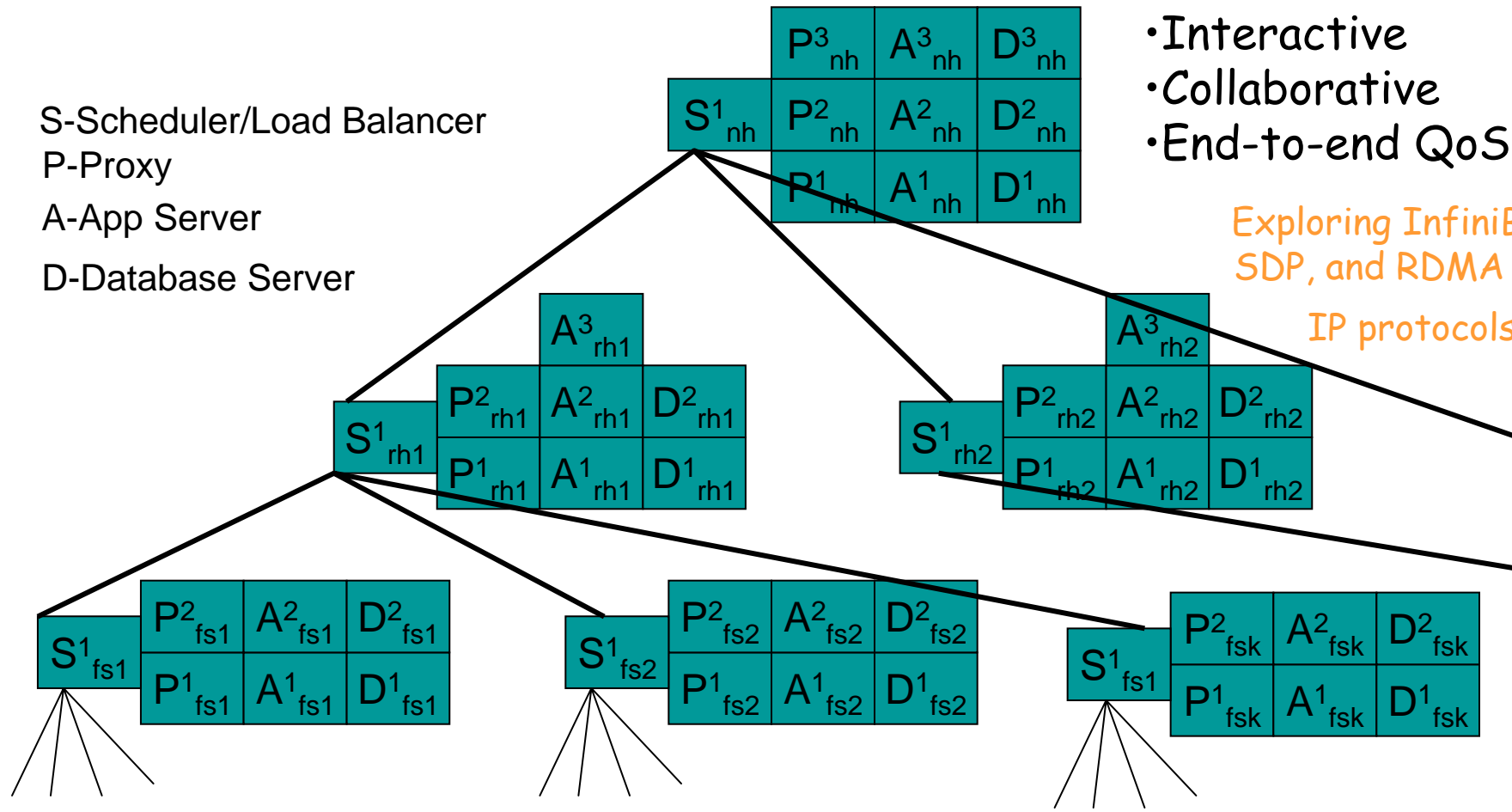
The VAPI module can sustain performance even with heavy load on the back-end servers

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

Distributed Datacenter Design

- Interactive
- Collaborative
- End-to-end QoS

Exploring InfiniBand,
SDP, and RDMA over
IP protocols





Presentation Overview



- Networking and I/O Requirements for Distributed Data-Intensive Applications
- Emerging Networking Technology and Protocols
 - InfiniBand
 - Sockets Direct Protocol
 - RDMA over IP
- Systems and Networking Research at OSU
 - High Performance MPI with InfiniBand for Clusters
 - Parallel File Systems
 - Multi-Tier Data Center
- Overview of NSF Research Infrastructure (RI) Grant
- Conclusions



High-End Computing and Networking Research Testbed for Next Generation Data Driven, Interactive Applications

PIs: D. K. Panda, G. Agrawal, P. Sadayappan, J. Saltz and H.-W. Shen

Other Investigators: S. Ahalt, U. Catalyurek, H. Ferhatosmanoglu,
H.-W. Jin, T. Kurc, M. Lauria, D. Lee, R. Machiraju, S. Parthasarathy,
P. Sinha, D. Stredney, A. E. Stutz, and P. Wyckoff

Dept. of Computer Science and Engineering, Dept. of Biomedical
Informatics, and Ohio Supercomputer Center

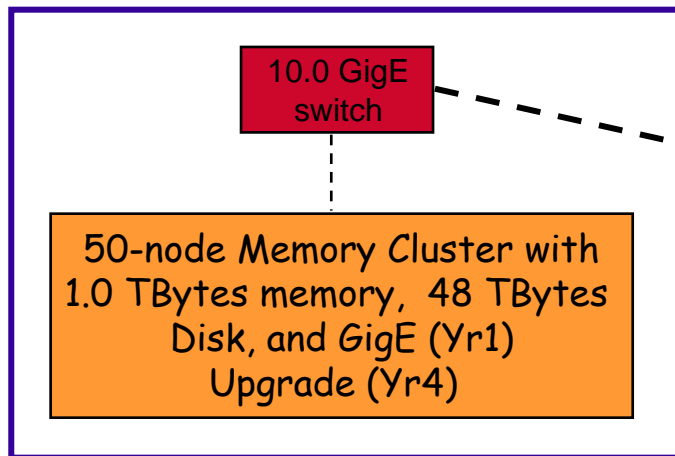
The Ohio State University

Total funding: \$3.01M (\$1.53M from NSF + \$1.48 from Ohio State
Board of Regents and Various units in OSU)

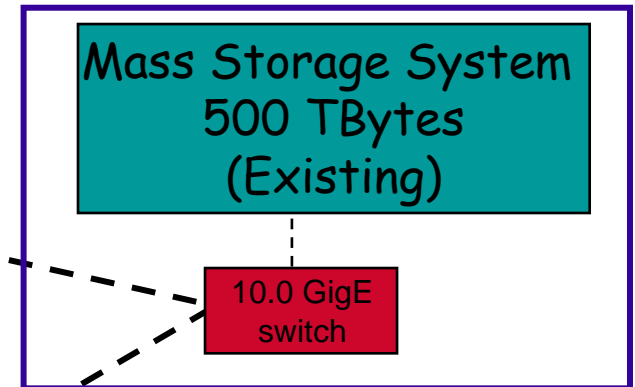


Proposed Experimental Testbed

BMI



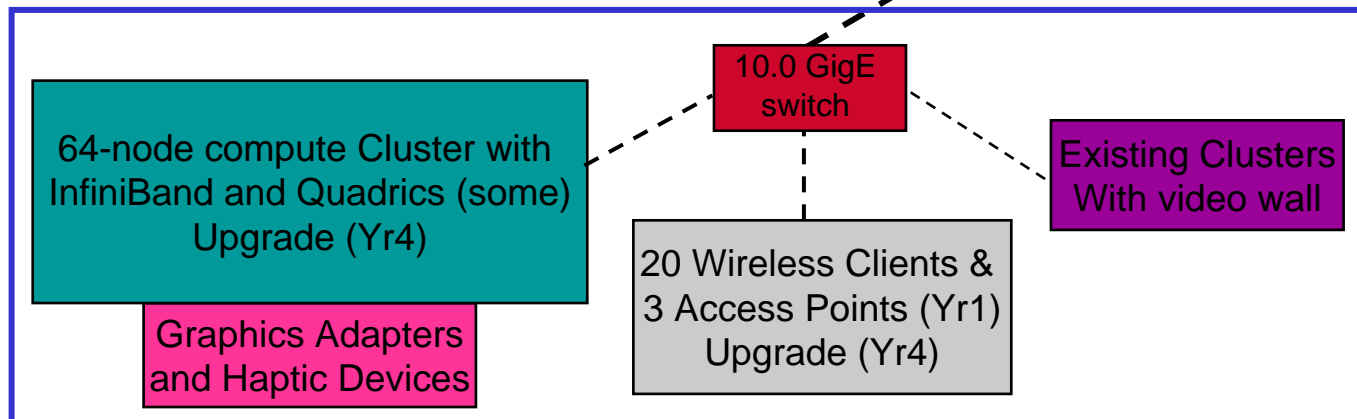
OSC



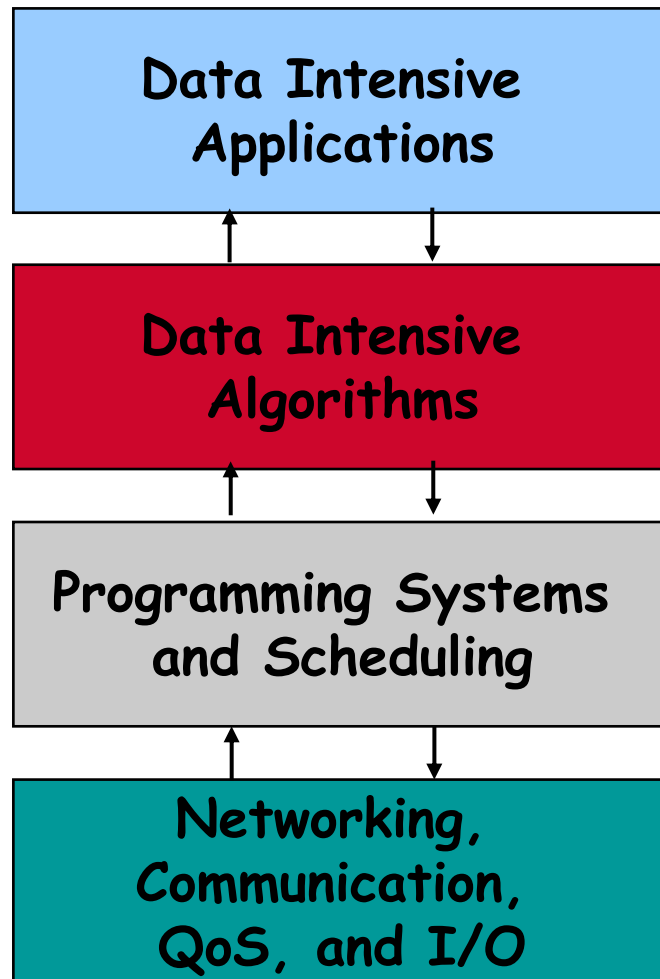
2x10=20 GigE (Yr1)
40 GigE (Yr4)

2x10=20 GigE (Yr1)
40 GigE (Yr4)

CSE



Collaboration among the Components and Investigators



Saltz, Stredney, Sadayappan
Machiraju, Parthasarathy,
Catalyurek, and Other OSU
collaborators

Shen, Agrawal, Machiraju,
and Parthasarathy,

Saltz, Agrawal, Sadayappan,
Kurc, Catalyurek, Ahalt, and
Hakan

Panda, Jin, Lee, Lauria,
Sinha, Wyckoff, and
Kurc



Presentation Overview



- Networking and I/O Requirements for Distributed Data-Intensive Applications
- Emerging Networking Technology and Protocols
 - InfiniBand
 - Sockets Direct Protocol
 - RDMA over IP
- Systems and Networking Research at OSU
 - High Performance MPI with InfiniBand for Clusters
 - Parallel File Systems
 - Multi-Tier Data Center
- Overview of NSF Research Infrastructure (RI) Grant
- **Conclusions**





Conclusions



- Distributed Data Intensive Applications (like Imaging) provides a lot of systems-level challenges
 - Computing
 - Networking
 - I/O and File Systems
- Interplay between capabilities of systems and what can be achieved at the upper layers including users
- Modern networking technologies and protocols are providing new ways to design such systems and deliver the capabilities to the upper layers
- The experimental testbed under the NSF RI grant will allow us to explore many of these issues



•
•
•

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

• • • • • • • •