OSU MPI (MVAPICH and MVAPICH2): Latest Status, Performance Numbers and Future Plans

Presentation at OpenIB Sonoma Workshop (Feb '06)

by

Dhabaleswar K. (DK) Panda

Department of Computer Science and Engg.

The Ohio State University

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

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

Presentation Overview

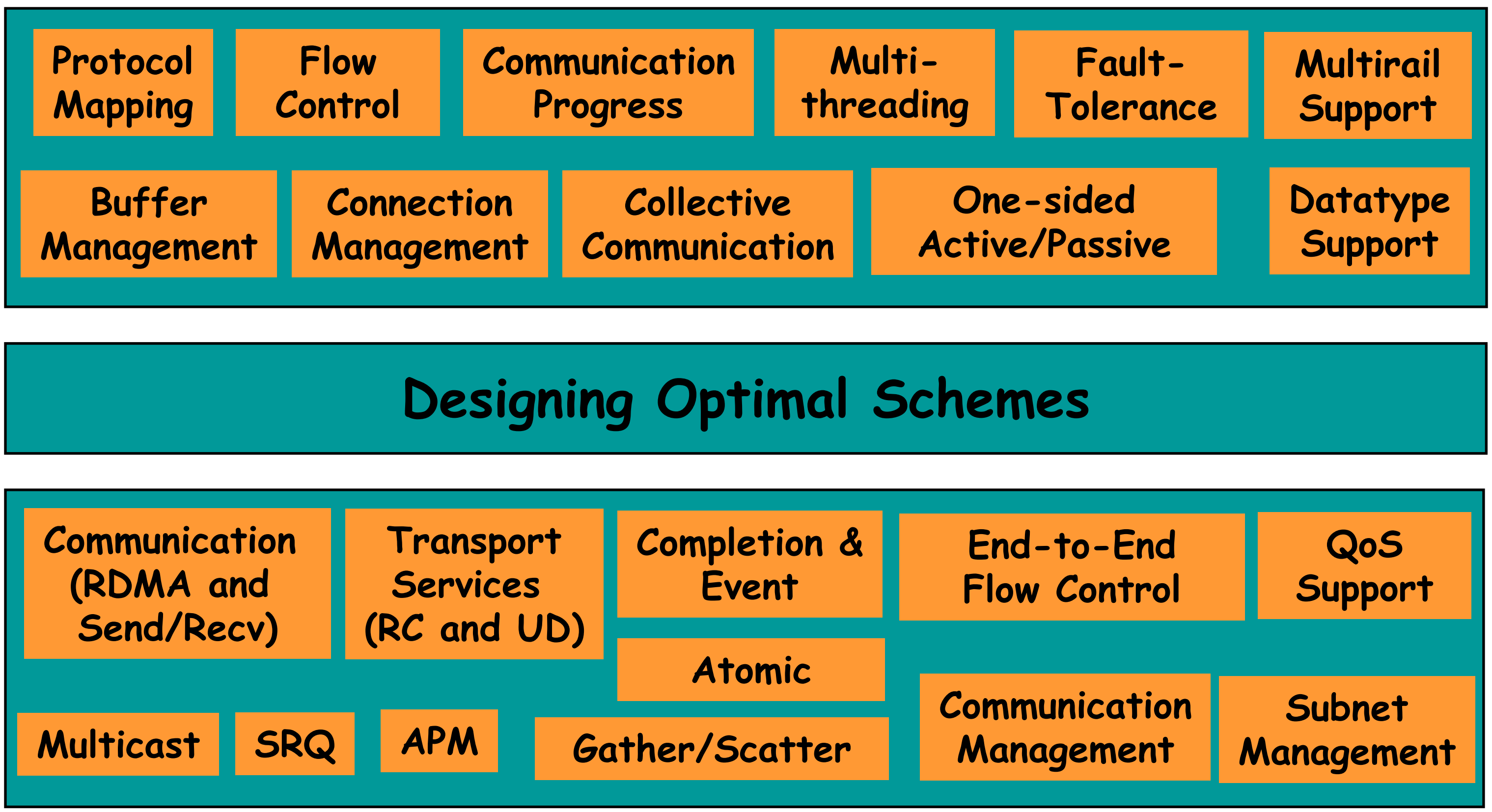
- Overview of MVAPICH and MVAPICH2 Projects
- MVAPICH 0.9.6 Features and Performance
  - Point-to-point
    - VAPI and Gen2
    - Mellanox and PathScale adapters
  - Adaptive RDMA Fast Path
  - RDMA Read
  - Collectives (Multicast, Barrier, All-to-All, All-gather)
  - Multi-rail support
  - Blocking support
  - uDAPL support
  - SDR/DDR comparison
- MVAPICH2 0.9.2 Features and Performance
  - Two-sided (VAPI and Gen2)
  - One-sided (VAPI and Gen2)
  - uDAPL support
  - Comparison of 0.9.6 with 0.9.2
- Upcoming MVAPICH 0.9.7 Features and Performance
  - Integrated VAPI, Gen2, UDAPL support
  - SRQ with Flow Control
  - Fault Tolerance
    - Memory-to-memory Reliability

Presentation Overview (Cont'd)

- Upcoming Features and Sample Performance
  - Fault Tolerance
    - Checkpoint-Restart
    - Automatic Path Migration (APM)
  - Multithreading
  - Multi-Network Support with uDAPL
  - Adaptive Connection Management
  - QoS Features and Routing
- Overview of Additional Projects
  - SDP
  - iWARP
  - Lustre, GFS, NFS over RDMA
  - Xen over IB
  - Multi-tier DataCenter
- Conclusions

Designing MPI Using Features of InfiniBand and RDMA Interconnects

MPI Design Components



Features of InfiniBand and RDMA Interconnects

MVAPICH/MVAPICH2 Software Distribution

- Focusing both
  - MPI-1 (MVAPICH)
  - MPI-2 (MVAPICH2)
- Open Source (BSD licensing)
- Started from IB 1X (2001)
- First high performance MPI over IB 4X was demonstrated at SC '02 (12-node blade server)
- Since then it has enabled a large number of production IB clusters all over the world to take advantage of IB
  - Largest being Sandia Thunderbird Cluster (4000 node with 8000 processors)
- Have been directly downloaded and used by more than 310 organizations worldwide (in 32 countries)
  - Time tested and stable code base with novel features
- Available in software stack distributions of many vendors
- Available in the OpenIB/gen2 stack

MVAPICH/MVAPICH2 Software Distribution

- Multiple Implementations on different low-level APIs
  - VAPI
    - MVAPICH 0.9.6 (MPI-1) and MVAPICH2 0.9.2 (MPI-2)
    - MVAPICH 0.9.5/0.9.6 is available with the software stack of many IBA and server vendors including Mellanox IBGD
  - OpenIB Gen2 stack
    - Two different versions are available at the OpenIB SVN
      - MVAPICH-Gen2 1.0
      - MVAPICH2 0.9.2
    - MVAPICH-Gen2 is also available with Mellanox IBG2
  - uDAPL
    - To achieve portability across different interconnects through uDAPL
    - Available for both MPI-1 (MVAPICH 0.9.6) and MPI-2 (MVAPICH2 0.9.2)
      - Tested with uDAPL-Solaris/IBA, uDAPL-OpenIBGen2/IBA and uDAPL-Myrinet/GM
  - TCP/IP
    - Based on MPICH and MPICH2
    - Can work with
      - IP over IB
      - Any other network supporting TCP/IP stack (such as Level5, Chelsio, …)

MVAPICH/MVAPICH2 Software Distribution (Cont'd)

- Available and Optimized for
  - Platforms
    - IA-32, IA-64, Opteron, EM64T and Apple G5
    - PPC/IBM support will be added in mvapich 0.9.7
  - Operating Systems
    - Linux, Solaris and Mac OSX
  - Compilers
    - GCC, Intel, PathScale and PGI
  - InfiniBand Adapters
    - PCI-X and PCI-Express (SDR and DDR with mem-full/mem-free cards)
- More details at
  http://nowlab.cse.ohio-state.edu/projects/mpi-iba/
- A set of microbenchmarks (two-sided, one-sided, broadcast)
  - Known in the IBA community as OSU benchmarks
- Sample Performance Numbers on Various Platforms

MVAPICH/MVAPICH2 Users: National Labs and Research Centers

Alabama Supercomputer Center
Argonne National Laboratory
Astrophysics Institute Potsdam (Germany)
AWI Polar and Marine Research Center (Germany)
CASPUR, Interuniversity Consortium (Italy)
CLUMEQ Supercomputer Center (Canada)
Cornell Theory Center
C-DAC, Center for Development of Advanced Computing (India)
Center for Computational Molecular Science and Technology, Georgia Tech
Center for High Performance Computing, Univ. of New Mexico
Center for Math. And Comp. Science (The Netherlands)
CCLRC Daresbury Laboratory (UK)
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)
CSC-Scientific Computing Ltd. (Finland)
CWI, The Netherlands Center for Mathematics and Computers Science (The Netherlands)
ECMWF, European Center for Medium-Range Weather Forecasts (UK)
ENEA, Casaccia Res. Center (Italy)
Fermi National Accelerator Laboratory
Fraunhofer-Inst. for High-Speed Dynamics (Germany)
Glushkov Inst. of Cybernatics (Ukraine)
High Performance Computing Center, Texas Tech Univ.
HPC and Mass Storage Institute, Catholic Univ. of Louvain (Belgium)
IFP, French National Oil and Gas Res. Center (France)
Inst. for Experimental Physics (Germany)
Inst. For Industrial Mathematics, ITWM (Germany)
Inst. for Program Structures and Data Org. (Germany)
INT, Institut National des Telecommunications (France)
Inst. of Astronomy, Czech Adademy of Sciences (Czech Republic)
Inst. of Computational Mathematics and Mathematical Geophysics (Russia)
Inst. of Physics, Chinese Academy of Sciences (China)
Inst. For Meteorological Research (Iceland)

MVAPICH/MVAPICH2 Users: National Labs and Research Centers

Inst. "Rudjer Boskovic" (Croatia)
IRSN (France)
Joint Institute for Computational Sciences, JIST
Kavli Inst. for Astrophysics and Space Research
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)
Max Planck Institute for Plasma Physics (Germany)
Michigan State University – HPC Center
NASA Ames Research Center
NCSA
National Center for High Performance Computing (NCHC, Taiwan)
National Center for Atmospheric Research
National Supercomputer Center in Linkoping (Sweden)
Oak Ridge National Laboratory NCCS Division
Ohio Supercomputer Center
Open Computing Centre "Strela" (Russia)
Pacific Northwest National Laboratory
Pittsburgh Supercomputing Center
Princeton Plasma Physics Laboratory
Ponzan Computing and Networking Center (Poland)
Rennaissance Computing Institute, Univ. of North Carolina, Chapel Hill
Research & Development Institute Kvant (Russia)
Sandia National Laboratory
SARA Dutch National Computer Center (The Netherlands)
Science Applications International Corporation
Stanford Center for Computational Earth and Environmental Science
Swiss Institute of Bioinformatics (Switzerland)
Texas Advanced Computing Center
Trinity Center for High Performance Computing (Ireland)
United Institute of Informatics Problems (Belarus)
University of Florida HPC Center
U.S. Army ERDC MSRC
U.S. Census Bureau
U.S. Geological Survey
Wegner Center for Climate and Global Change (Austria)
Woods Hole Oceanographic Inst.

MVAPICH/MVAPICH2 Users: Universities

- Aachen Univ. of Applied Sciences (Germany)
- Drexel University
- Engineers School of Geneva (Switzerland)
- Florida A&M University
- Georgia Tech
- Grdansk Univ. of Technology (Poland)
- Gwangju Inst. Of Science and Technology (Korea)
- Hardvard University
- Indiana University
- Indiana State University
- Johannes Kepler Univ. Linz (Austria)
- Johns Hopkins University
- Korea Univ. (Korea)
- Kyushu Univ. (Japan)
- Miami University
- Mississippi State University
- MIT Lincoln Lab
- Mount Sinai School of Medicine
- Moscow State University (Russia)
- Northeastern University
- Nankai University (China)
- Old Dominion University
- Oregon State University
- Penn State University
- Pohang Univ. of Science and Tech., POSTECH (Korea)
- Purdue State University
- Queen's University (Canada)
- Rostov State University (Russia)
- Russian Academy of Sciences (Russia)
- Seoul National University (Korea)
- Shandong Academy of Sciences (China)
- South Ural State University (Russia)
- Stanford University
- Technion (Israel)
- Technical Univ. of Berlin (Germany)
- Technical Univ. of Clausthal (Germany)
- Technical Univ. of Munchen (Germany)
- Technical Univ. of Chemnitz (Germany)
- Tokyo Univ. of Technology (Japan)
- Tsinghua Univ. (China)
- Univ. of Arizona
- Univ. of Berne (Switzerland)
- Univ. of Bielefeld (Germany)
- Univ. of California, Berkeley
- Univ. of California, Davis
- Univ. of California, Los Angeles
- Univ. of Chile (Chile)
- Univ. of Erlangen-Nuremberg (Germany)
- Univ. of Florida, Gainesville
- Univ. of Geneva (Switzerland)
- Univ. of Hannover (Germany)
- Univ. of Houston
- Univ. of Karlsruhe (Germany)
- Univ. of Lausanne (Switzerland)
- Univ. of Laval (Canada)
- Univ. of Luebeck (Germany)
- Univ. of Massachusetts Lowell
- Univ. of Milan (Italy)
- Univ. of Minnesota
- Univ. of Paderborn (Germany)
- Univ. of Pisa (Italy)
- Univ. of Pittspurgh
- Univ. of Politecnica of Valencia (Spain)
- Univ. of Potsdam (Germany)
- Univ. Du Quebec a Chicoutimi (Canada)
- Univ. of Rio Grande (Brazil)
- Univ. of Rostock (Germany)
- Univ. of Sherbrooke (Canada)
- Univ. of Siegen (Germany)
- Univ. of Surrey (UK)
- Univ. of Stuttgart (Germany)
- Univ. of Tennessee, Knoxville
- Univ. of Tokyo (Japan)
- Univ. of Toronto (Canada)
- Univ. of Twente (The Netherlands)
- Univ. of Vienna (Austria)
- Univ. of Westminster (UK)
- Univ. of Zagreb (Croatia)
- Vienna Univ. of Technology (Austria)
- Virginia Tech
- Wroclaw Univ. of Technology (Poland)

MVAPICH/MVAPICH2 Users: Industry (1)

Abba Technology
Adelie Linux (Canada)
Advanced Clustering Tech.
Agilent Technologies
AMD
AMD (Japan)
Alliance Technologies
Ammasso
Annapolis Micro Systems, Inc.
Apple Computer
Appro
Array Systems Comp. Inc. (Canada)
Ascender Technologies Ltd (Israel)
Ascensit (Italy)
Atipa Technologies
AWE PLC (UK)
BAE Systems
Barco Medical Imaging Systems
Best Systems Inc. (Japan)
Bluware
Broadcom
Bull S.A. (France)
CAE Elektronik GmbH (Germany)
California Digital Corporation
Caton Sistemas Alternativos (Spain)
Cisco Systems
Clustars Supercomputing Tech. Inc. (China)
Cluster Technology Ltd. (Hong Kong)
Clustervision (Netherlands)
Compusys (UK)
Cray Canada, Inc. (Canada)
CSS Laboratories, Inc.
Cyberlogic (Canada)
Dell
Delta Computer Products (Germany)
Diversified Technology, Inc.
DRS Technologies
Dynamics Technology, Inc.
Easy Mac (France)
Emplics (Germany)
ESI Group (France)
Exadron (Italy)
ExaNet (Israel)
Faster Technology
Fluent Inc.
Fluent Inc. (Europe)
Fujitsu Ltd. (Japan)
FMS-Computer and Komm. (Germany)
General Atomics
GraphStream, Inc
Gray Rock Professional
HP
HP (Asia Pacific)
HP (France)
HP Galway Limited (Ireland)
HP Solution Center (China)
High Performance Associates
IBM
IBM (China)
IBM (France)
IBM (Germany)
INTERSED (France)
IPS (Austria)
Incad Ltd. (Czech Republic)
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)
Level 5 Networks

MVAPICH/MVAPICH2 Users: Industry (2)

Linux Networx
Linvision (Netherlands)
Livermore Software Technology Corp.
Lumerical Solutions Inc. (Canada)
.Megaware (Germany)
Mercury Computer Systems
Mellanox Technologies
Meiosys (France)
Microsoft
Microway, Inc.
Motorola
NEC Europe, Ltd
NEC (Japan)
NEC Solutions, Inc.
NEC (Singapore)
NetEffect
NICEVT (Russia)
NovaGlobal Pte Ltd (Singapore)
OCF plc (United Kingdom)
OctigaBay
Open Technologies Inc. (Russia)
OptimaNumerics (UK)
Panasas
PANTA Systems
ParTec (Germany)
PathScale, Inc.
Platform Computing (UK)
Pultec (Japan)
Pyramid Computer (Germany)
Q Associated Ltd. (UK)
Qlusters (Israel)
Quadrics (UK)
Quant-X GmbH (Austria)
Rackable Systems, Inc.
Raytheon Inc.
Remcom Inc.
RJ Mears, LLC
RLX Technologies
Rocketcalc
Rosta Ltd. (Russia)
SBC Technologies, Inc.
Scyld Software
Scalable Informatics LLC
Scotland Electronics (Int'l) Ltd (UK)
SGI (Silicon Graphics, Inc.)
Siliquent
Silverstorm technologies
Simulation Technologies
SKY Computers
SmallTree communications
Societe Generale Investment Banking (France)
Solers Inc.
Space Exploration Technologies
STMicroelectronics
Streamline Computing (UK)
Sumisho Computer Systems Corp. (Japan)
SUN
Systran
Texh-X Corp.
Telcordia Applied Research
Telsima
Terra Soft Solutions
Thales Underwater Systems (UK)
Tomen
Topspin
Totally Hip Technologies (Canada)
Transtec (Germany)
T-Platforms (Russia)
T-Systems (Germany)
Unisys
Vector Computers (Poland)
Verari Systems Software
Virtual Iron Software, Inc.
Voltaire
Western Scientific
WorkstationsUK, Ltd. (UK)
Woven Systems, Inc.

Larger IBA Clusters using MVAPICH and Top500 Rankings (Nov. '05)

- 5th: 4000-node Dell PowerEdge 3.6 GHz (Thunderbird) cluster at Sandia National Laboratory
- 20th: 1100-node dual Apple Xserve 2.3 GHz cluster at Virginia Tech
- 51st: 576-node dual Intel Xeon EM64T 3.6 GHz cluster at Univ. of Sherbrooke (Canada)
- 226th: 356-node dual Opteron 2.4 GHz cluster at Trinity Center for High Performance Computing (TCHPC)
- 277th: 272-node dual Intel Xeon EM64T 3.4 GHz cluster at SARA (the Netherlands)
- 301st: 200-node dual Intel Xeon EM64T 3.2 GHz cluster at Texas Advanced Computing Center
- 305th: 315-node dual Opteron 2.2 GHz cluster at NERSC/LBNL
- More are there ….

Recent Releases and mvapich-discuss mailing list

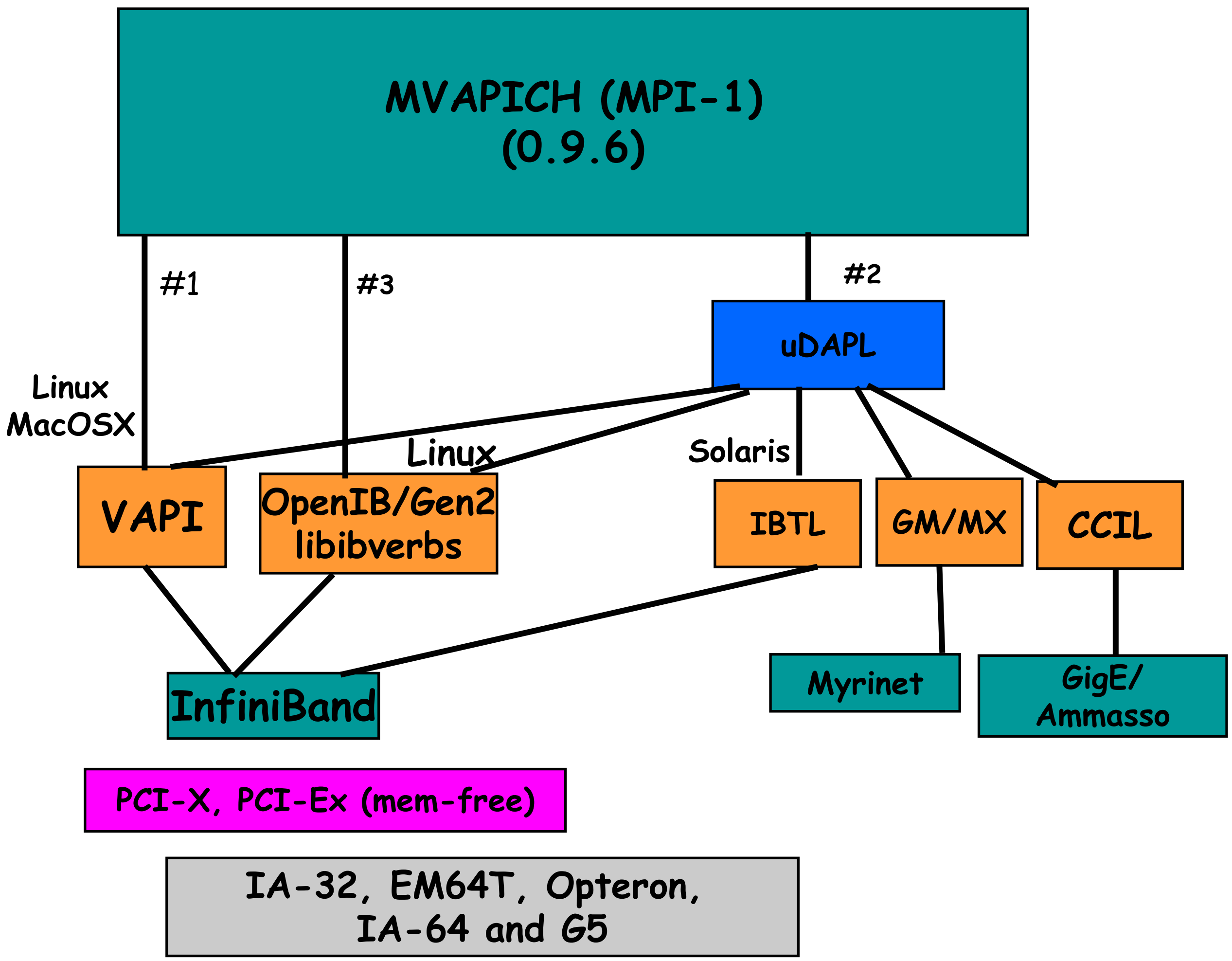
- Two releases have been made during the last two months
  - MVAPICH 0.9.6
  - MVAPICH2 0.9.2
- Established a new *mvapich-discuss* mailing list
  - Any mvapich/mvapich2 user, developer or vendor can subscribe
  - Post questions, comments and patches
  - Being actively used

http://www.cse.ohio-state.edu/mailman/listinfo/mvapich-discuss/

Presentation Overview

- Overview of MVAPICH and MVAPICH2 Projects
- MVAPICH 0.9.6 Features and Performance
  - Point-to-point
    - VAPI and Gen2
    - Mellanox and PathScale adapters
  - Adaptive RDMA Fast Path
  - RDMA Read
  - Collectives (Multicast, Barrier, All-to-All, All-gather)
  - Multi-rail support
  - Blocking support
  - uDAPL support
  - SDR/DDR comparison
- MVAPICH2 0.9.2 Features and Performance
  - Two-sided (VAPI and Gen2)
  - One-sided (VAPI and Gen2)
  - uDAPL support
  - Comparison of 0.9.6 with 0.9.2
- Upcoming MVAPICH 0.9.7 Features and Performance
  - SRQ with Flow Control
  - Fault Tolerance
    - Memory-to-memory Reliability

MVAPICH 0.9.6 Design



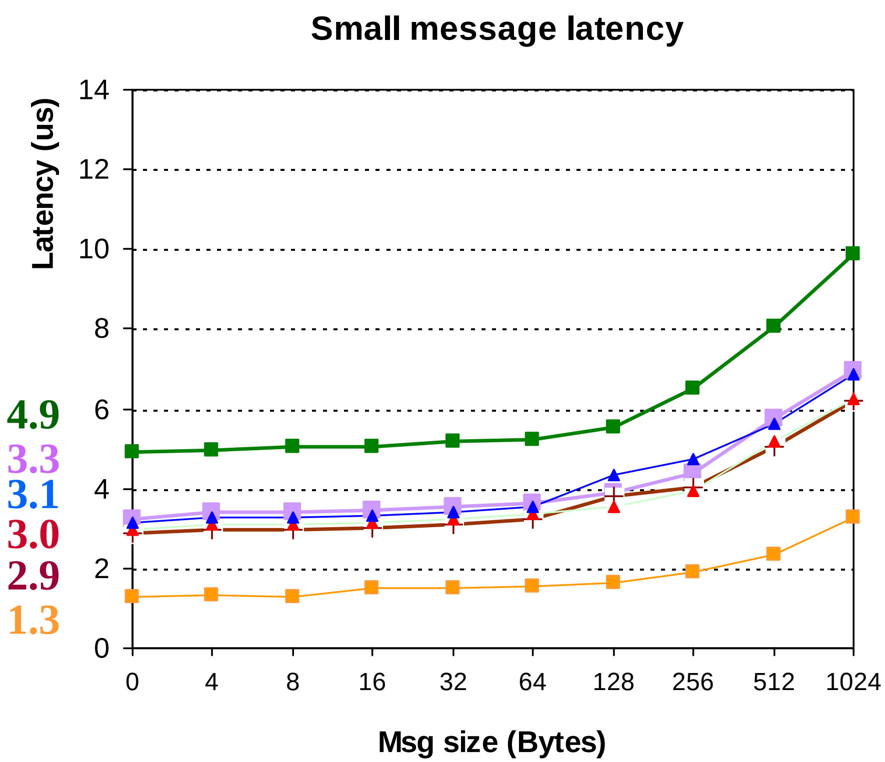
- Platforms
  - EM64T, Opteron, IA-32 and Mac G5
- Operating Systems
  - Linux, Solaris and Mac OSX
- Compilers
  - gcc, intel, pathscale and pgi
- InfiniBand Adapters
  - Mellanox adapters with PCI-X and PCI-Express (SDR and DDR with mem-full and mem-free cards)
- TCP/IP support also exists (through MPICH)
- Successive version will unify the codebase with OpenIB/Gen2 libibverbs version (MVAPICH-Gen2 1.0)

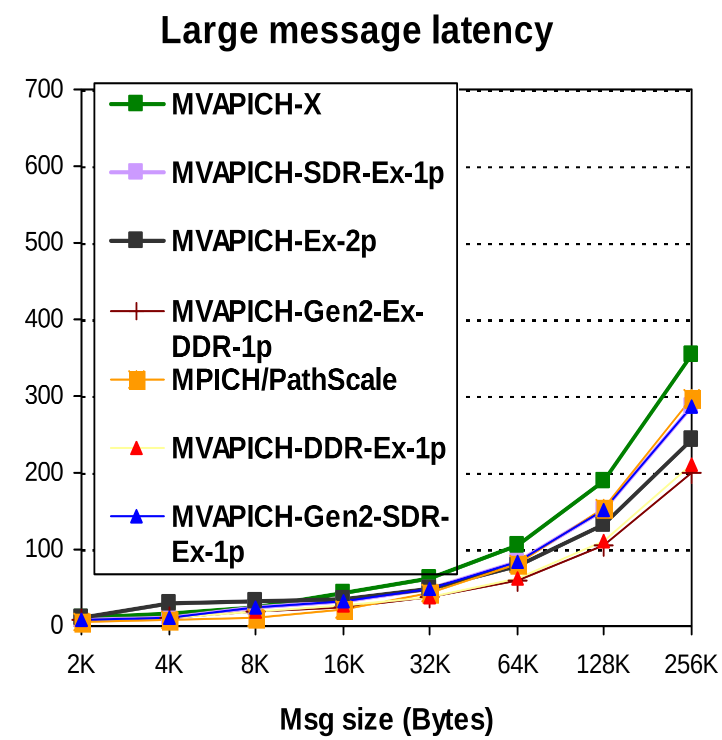
MVAPICH 0.9.6 Features

- RDMA-based point-point and collectives
- Multi-rail support
  - Multiple ports/adapters
  - Multiple adapters
  - Multiple paths with LMC
- Optimized Collectives
  - Broadcast support with IBA multicast
  - RDMA-based Barrier
  - RDMA-based All-to-all
- Optimized shared memory support
  - Bus-based architecture
  - NUMA architectures
- RDMA-based optimized collectives
  - Barrier
  - All-to-all
- Optimized for scalability
  - Three different modes: small, medium, and large clusters
- Totalview Debugger (Etnus) support
- MPD Support
- Shared Library support
- ROMIO support for MPI-IO

- Several New Features
- Adaptive Buffer Management and RDMA polling set
  - Significant reduction in memory usage and provide scalability
- RDMA Read support
  - Better overlap of computation and commn.
- Blocking communication support
- Enhanced RDMA-based collectives
  - All-gather
- uDAPL-based Portability
  - multiple interconnects and OS

MPI-level Latency (One-way): IBA (Mellanox and PathScale)

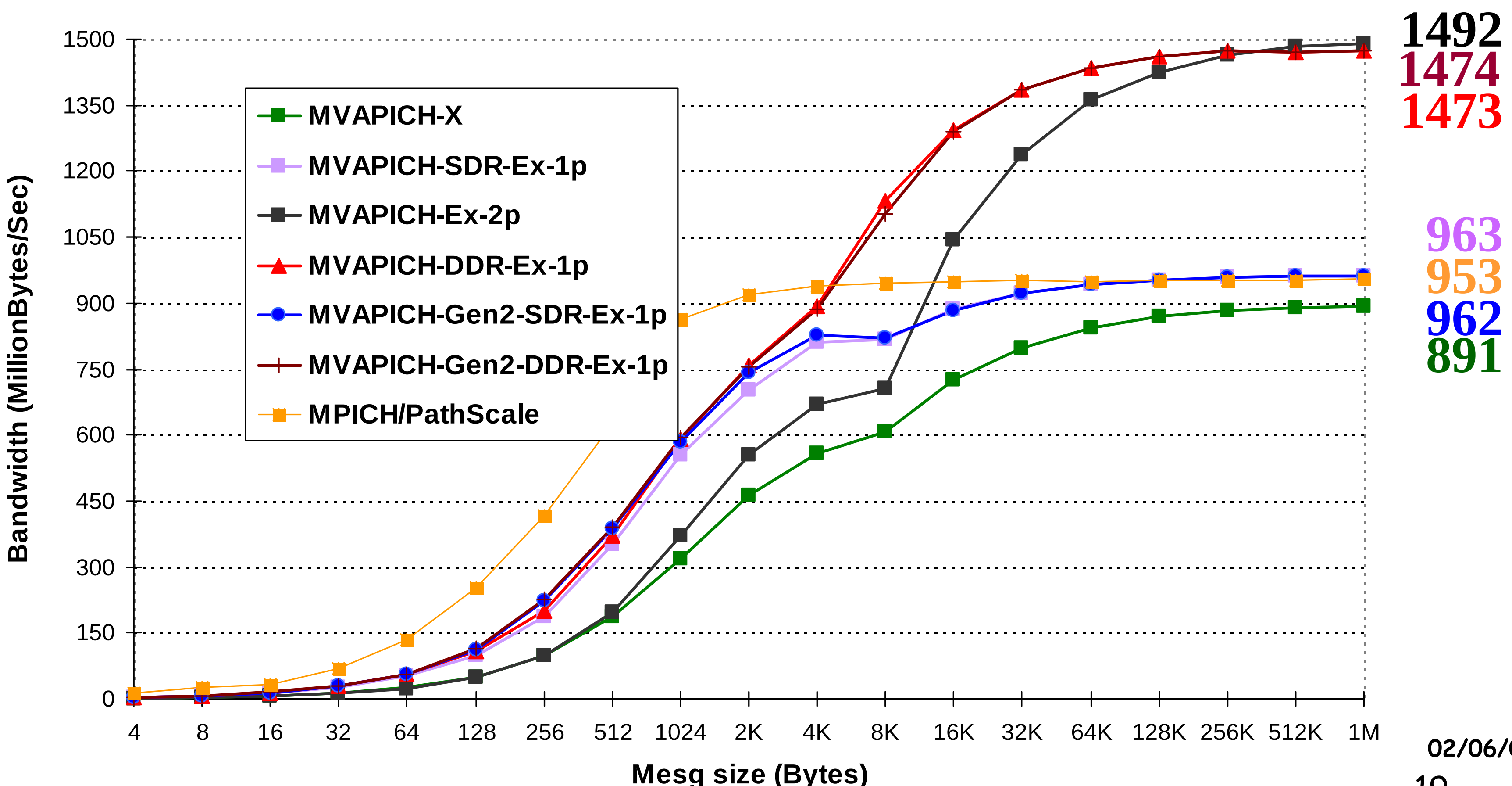




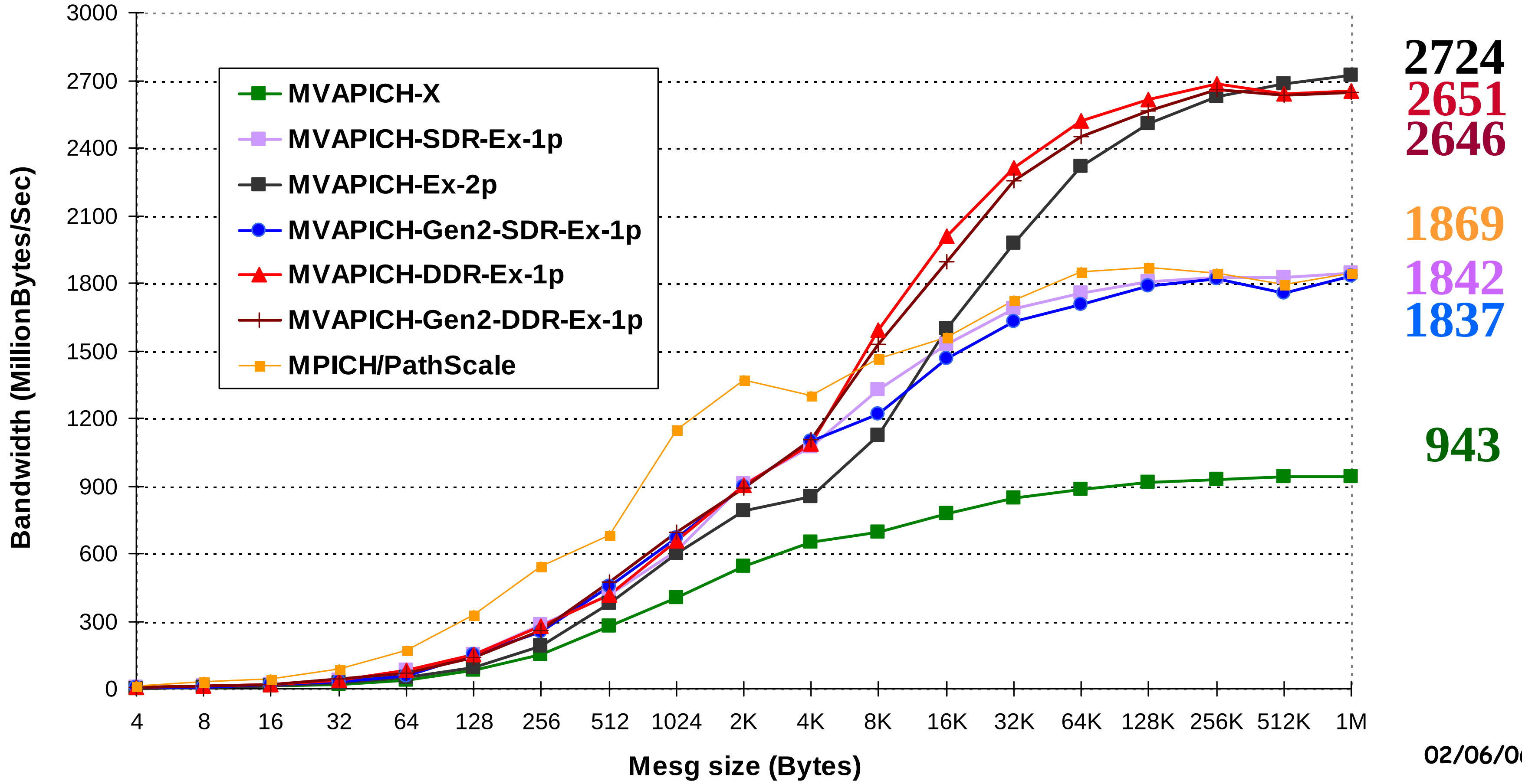
More detailed results are available with the following publications from the group

- Supercomputing ’03, Supercomputing ‘04
- Hot Interconnect ’04, Hot Interconnect ‘05
- IEEE Micro (Jan-Feb) ’04 and (Jan-Feb) ‘05, best papers from HotI ’04 and HotI ‘05

MPI-level Bandwidth (Uni-directional): IBA (Mellanox and PathScale)

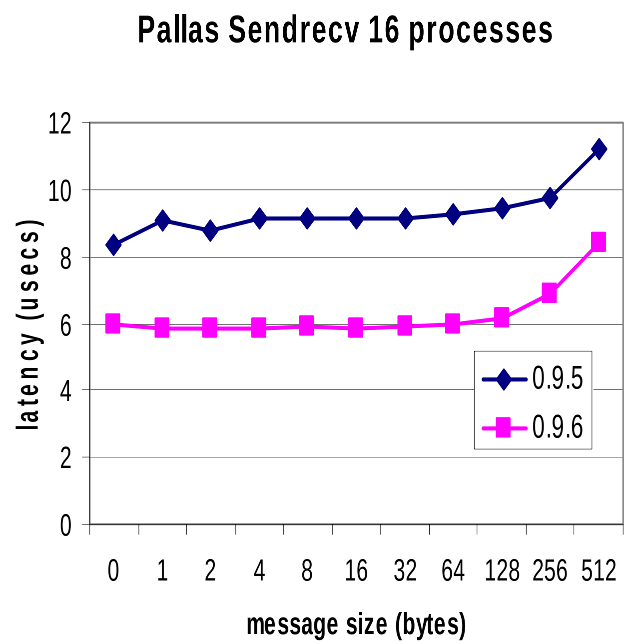


MPI-level Bandwidth (Bi-directional): IBA (Mellanox and PathScale)



MVAPICH 0.9.6 Feature: Adaptive RDMA Fast Path

- Connections Start with send/recv
- Switches to RDMA Fast Path if Communication frequency is higher
- Polls only on active connections



- 2 processes are involved in the involved in the data transfer with the rest waiting in Barrier

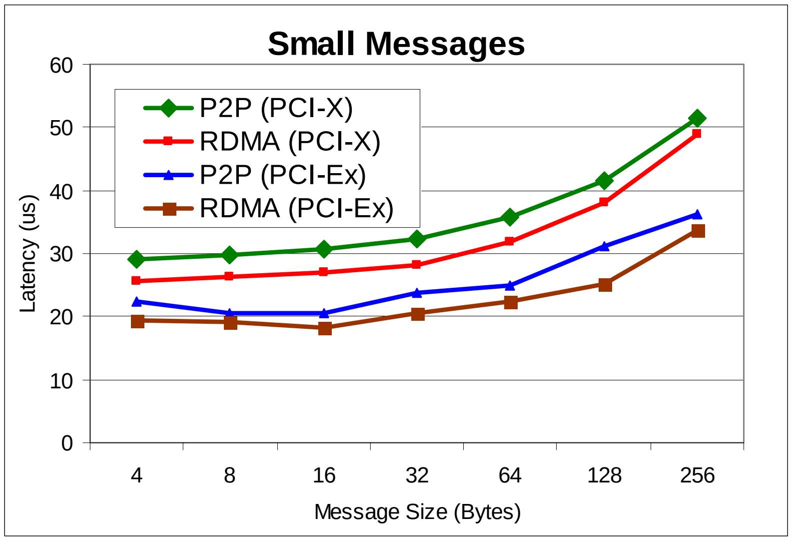
High Performance and Scalable Collectives

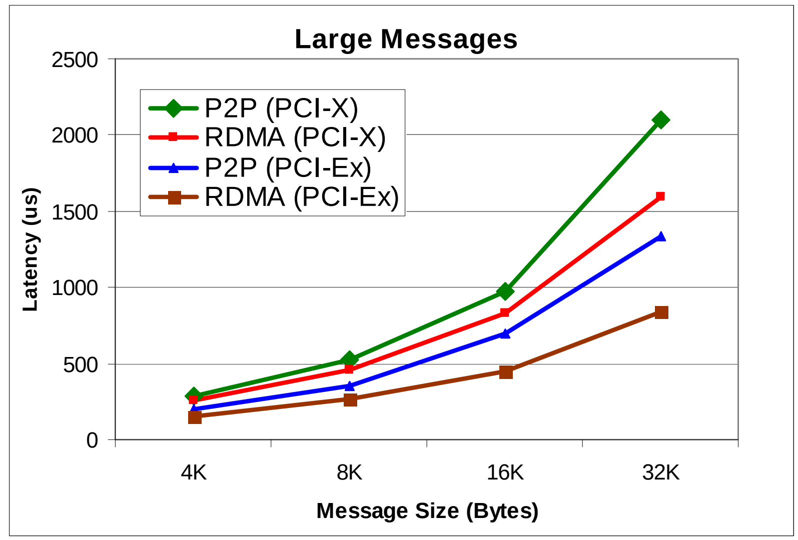
- Reliable MPI Broadcast using IB hardware multicast
  - Capability to support broadcast of 1K bytes message to 1024 nodes in less than 40 microsec
- RDMA-based designs for
  - MPI_Barrier
  - MPI_All_to_All

J. Liu, A. Mamidala and D. K. Panda, Fast and Scalable MPI-Level Broadcast using InfiniBand's Hardware Multicast Support, Int'l Parallel and Distributed Processing Symposium (IPDPS '04), April 2004

S. Sur and D. K. Panda, Efficient and Scalable All-to-all Exchange for InfiniBand-based Clusters, Int'l Conference on Parallel Processing (ICPP '04), Aug. 2004

MVAPICH 0.9.6 Features: RDMA-based MPI_Allgather

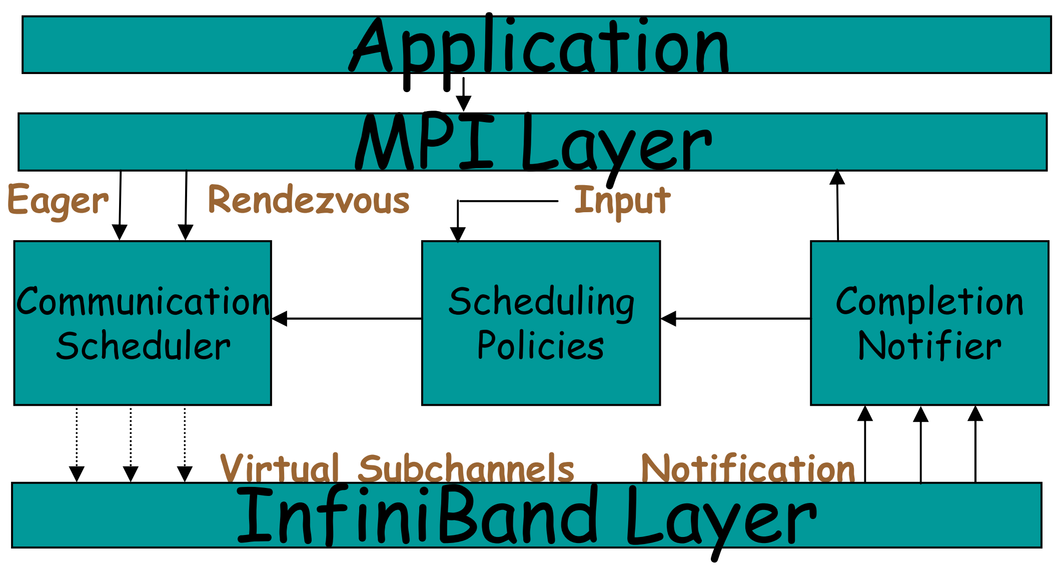




- RDMA based MPI_Allgather does 16.6% better for PCI-X and 13.6% better for PCI-Ex for small messages (4 bytes)
- For large messages (32KB), RDMA design does 30% better for PCI-X and 37% better for PCI-Ex

S. Sur, U. Bondhugula, A. Mamidala, H.-W. Jin and D. K. Panda, High Performance All-to-all broadcast for InfiniBand-based Clusters, Int'l Symposium on High Performance Computing (HiPC '05), Dec '05

Multi-Rail MPI Design for InfiniBand Clusters

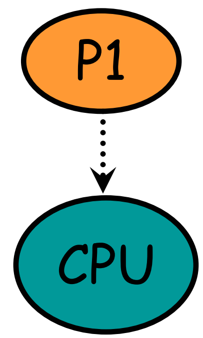


- Multiple ports/ adapters
- Multiple adapters
- Multiple paths with LMCs

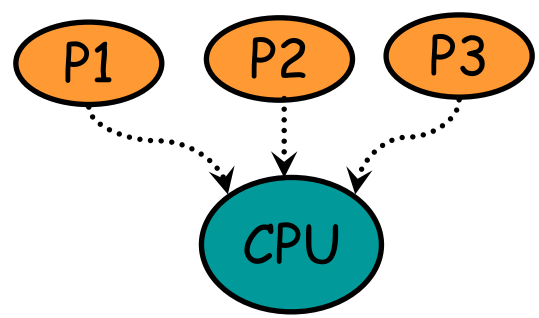
J. Liu, A. Vishnu and D. K. Panda. Building MultiRail InfiniBand Clusters: MPI Level Design and Performance Evaluation. Presented at Supercomputing '04, April, 2004

MVAPICH 0.9.6 Features: Blocking Mode Progress Engine

- Polling mode progress engine completely occupies CPU
- Blocking mode allows sharing CPU with other applications when MPI is idle-waiting
- Multiple processes can be mapped onto the same CPU when using blocking mode

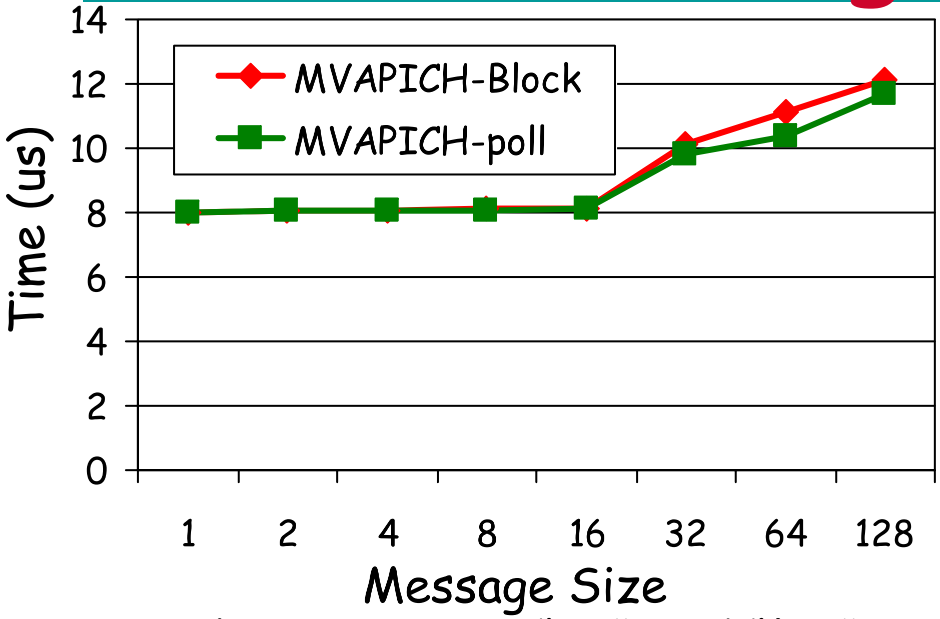


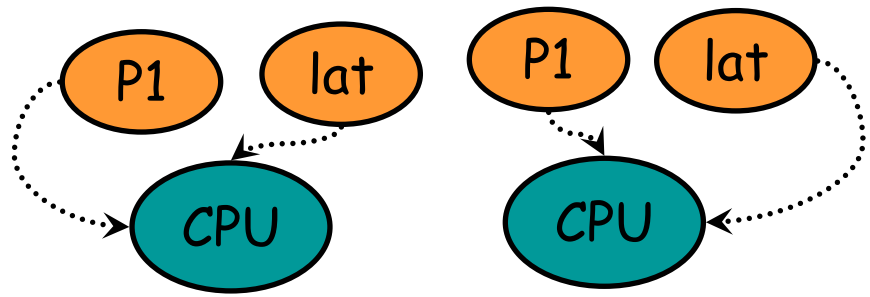
Polling Progress Mode



Blocking Progress Mode

Benefits of Blocking Mode Progress



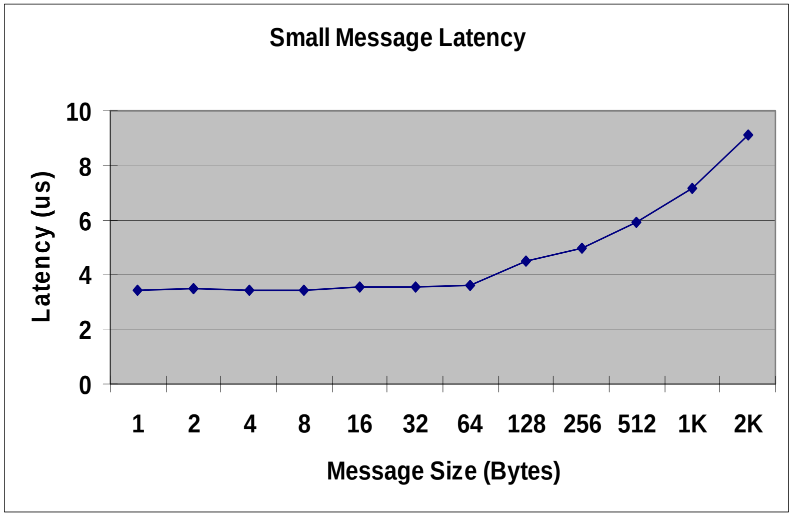


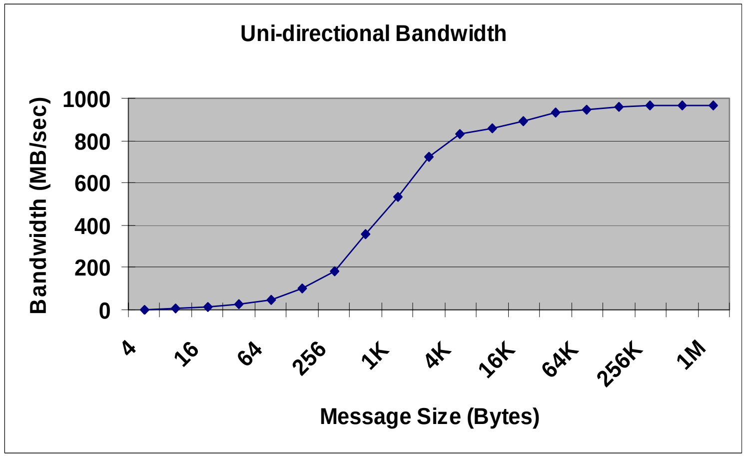
P1: MPI process waiting Polling in MPI_Wait
Lat: OSU Latency Test

- Both processes, "P1" and "lat" are mapped to the same CPU
- Latency is reported by "lat" which is OSU latency test
- MVAPICH-Poll represents the baseline performance if only "lat" is present
- If both "P1" and "lat" are present in Poll mode, the latency is in order of milliseconds

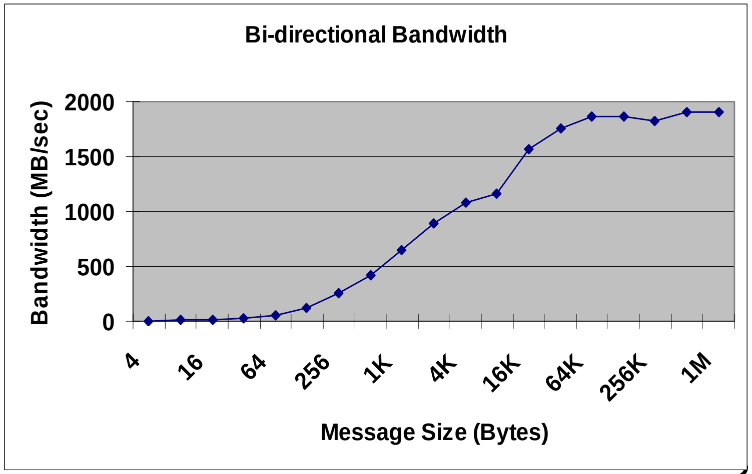
MVAPICH-0.9.6 uDAPL/Gen2 over InfiniBand: MPI-Level Performance

3.43





968



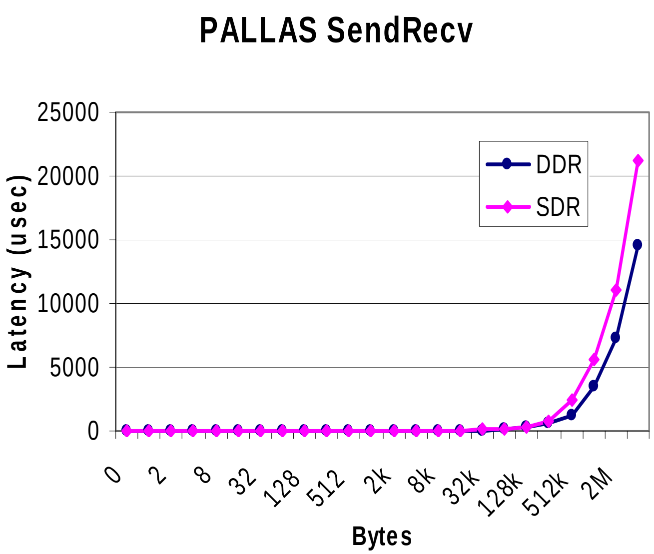
1912

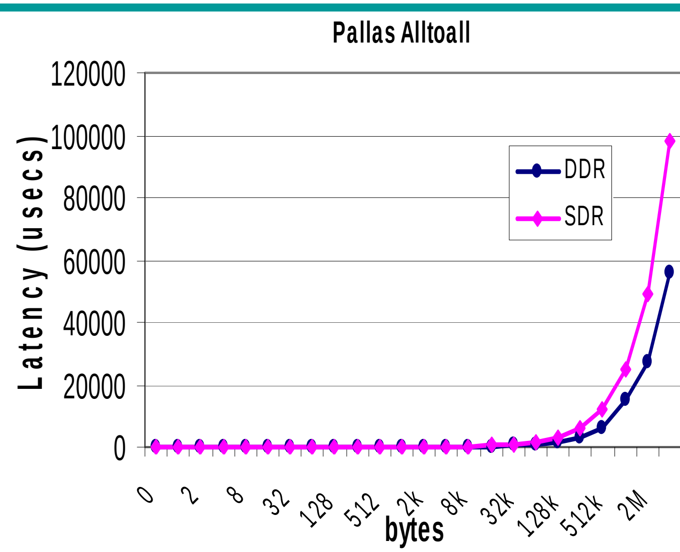
EM64T, PCI-Ex, SDR

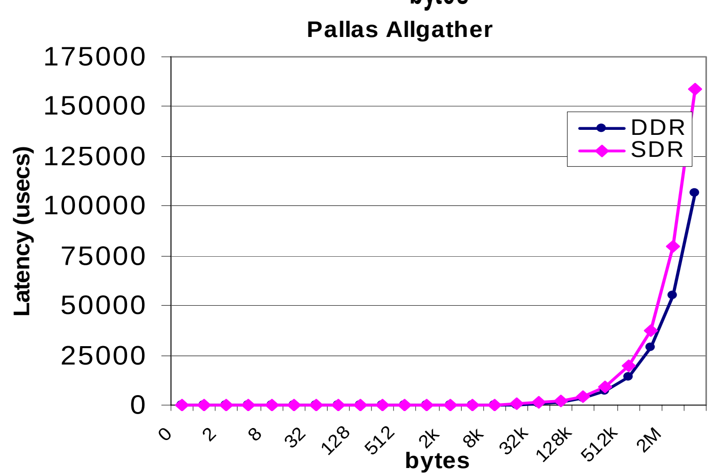
L. Chai, R. Noronha and D.K. Panda
MPI over uDAPL: Can High Performance and Portability Exist Across Architectures?
CCGrid'06, May 2006

IB support for Solaris is enabled through this uDAPL-based design

SDR/DDR Comparison for Micro-Benchmarks (Pallas)

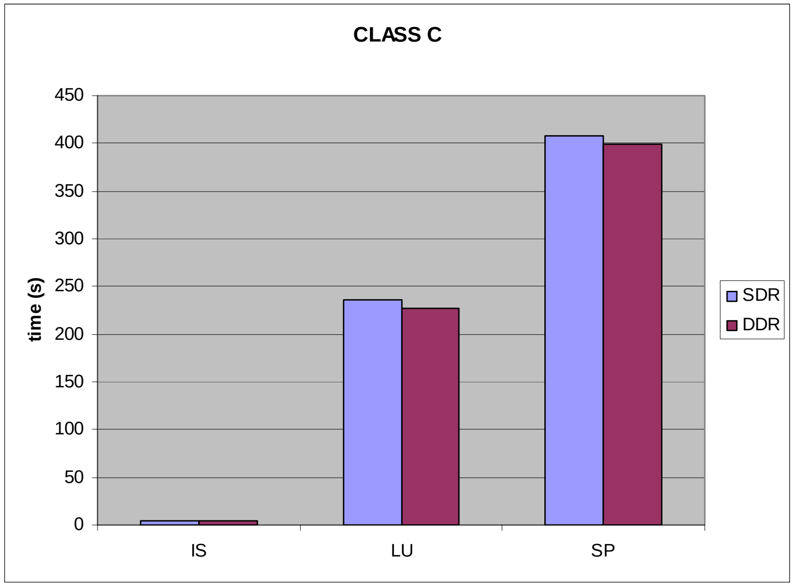


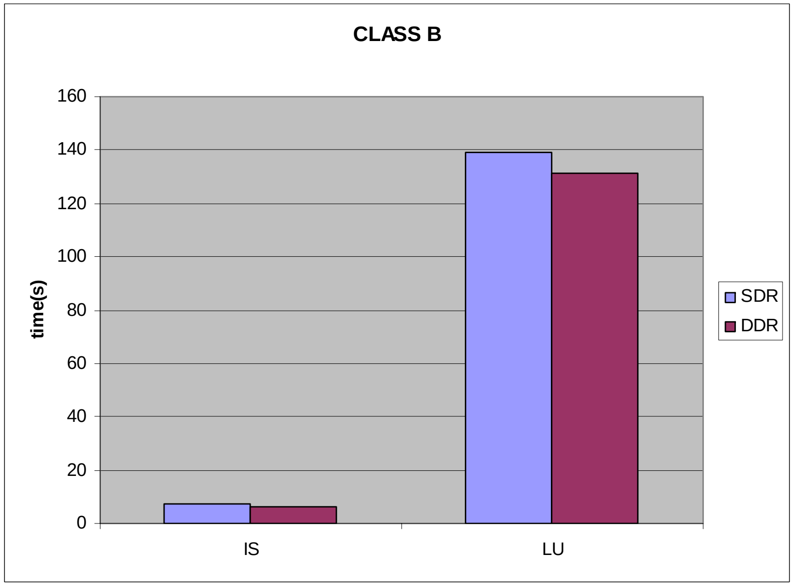




- MVAPICH 0.9.6
- EM64T, PCI-Ex
- 4x2 nodes

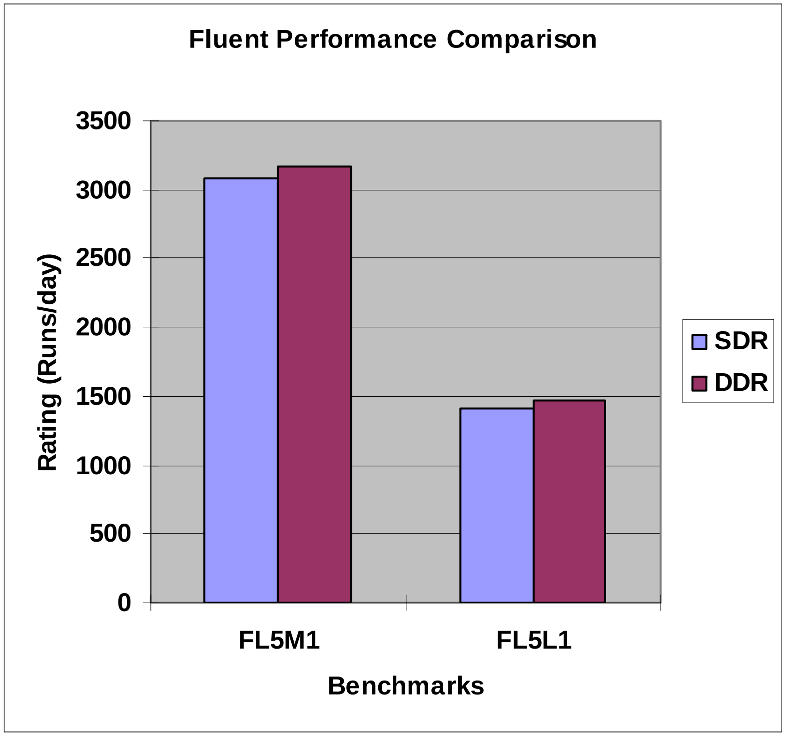
SDR/DDR Comparison for NAS Applications





- EM64T, 8x2 processes on VAPI
- DDR shows improvement for bandwidth sensitive applications

SDR/DDR Comparison for Fluent



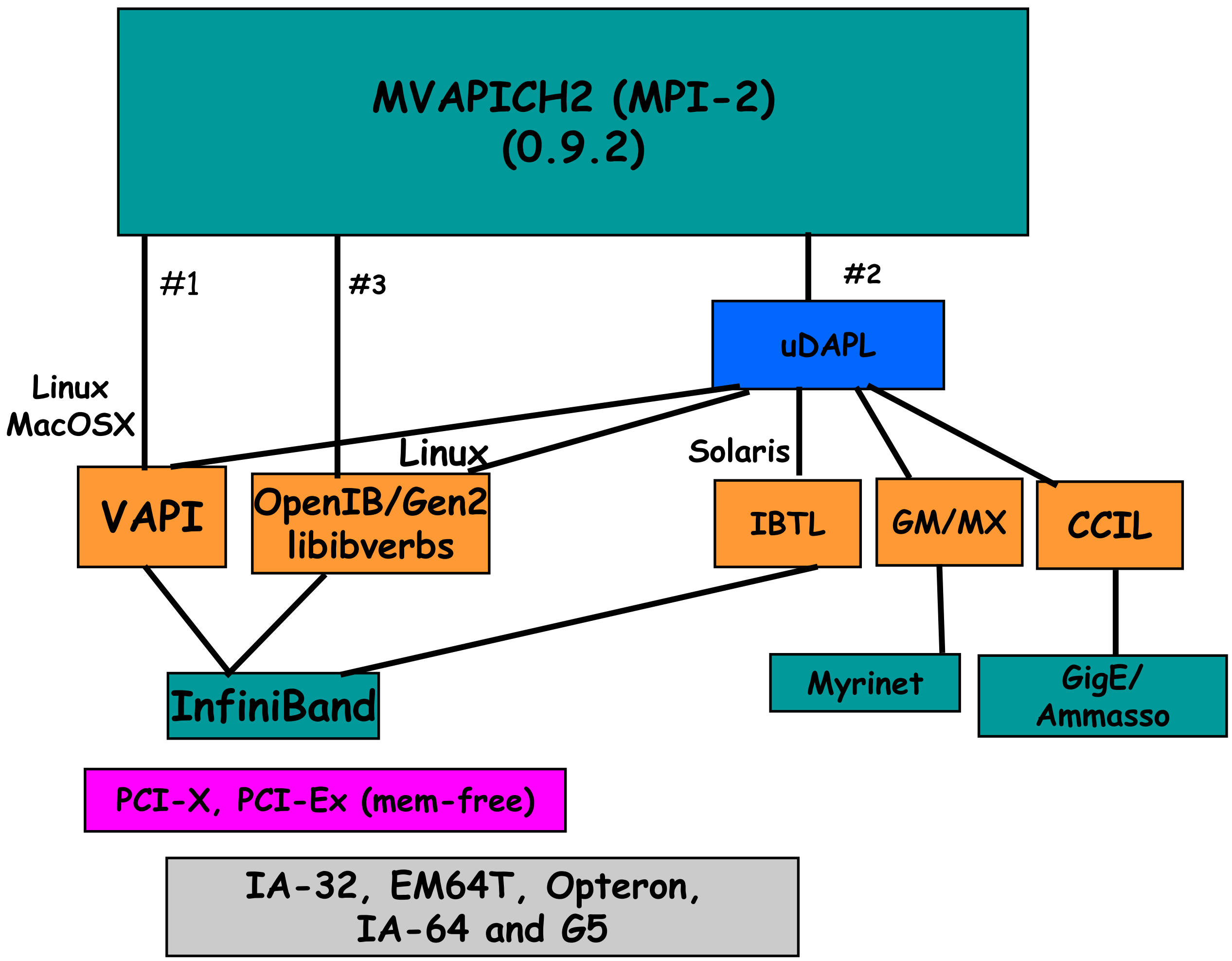
- MVAPICH 0.9.6 (VAPI)
- EM64T, PCI-Ex
- 8x2 processes

- Fluent is dominated by small messages
- For medium and large data sets we see some benefits by using DDR

Presentation Overview

- Overview of MVAPICH and MVAPICH2 Projects
- MVAPICH 0.9.6 Features and Performance
  - Point-to-point
    - VAPI and Gen2
    - Mellanox and PathScale adapters
  - Adaptive RDMA Fast Path
  - RDMA Read
  - Collectives (Multicast, Barrier, All-to-All, All-gather)
  - Multi-rail support
  - Blocking support
  - uDAPL support
  - SDR/DDR Comparison
- MVAPICH2 0.9.2 Features and Performance
  - Two-sided (VAPI and Gen2)
  - One-sided (VAPI and Gen2)
  - uDAPL support
  - Comparison of 0.9.6 with 0.9.2
- Upcoming MVAPICH 1.1 Features and Performance
  - SRQ with Flow Control
  - Fault Tolerance
    - Memory-to-memory Reliability

MVAPICH2 0.9.2 Design

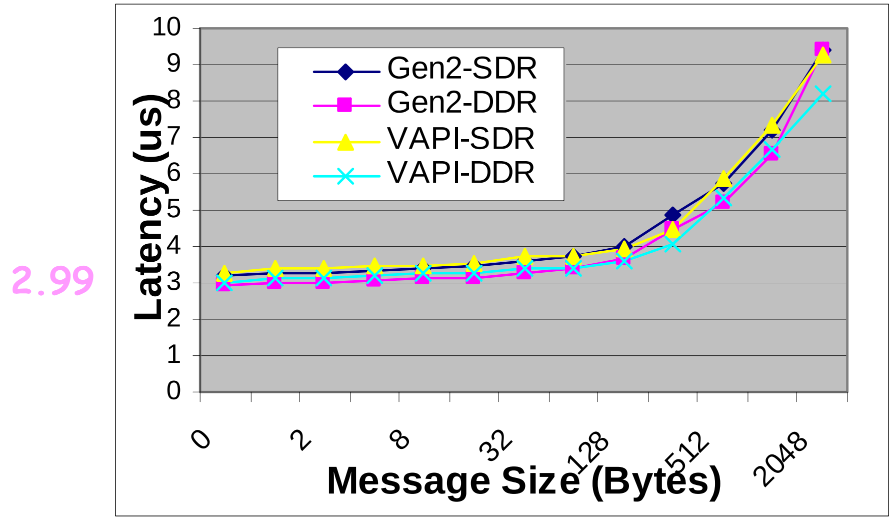


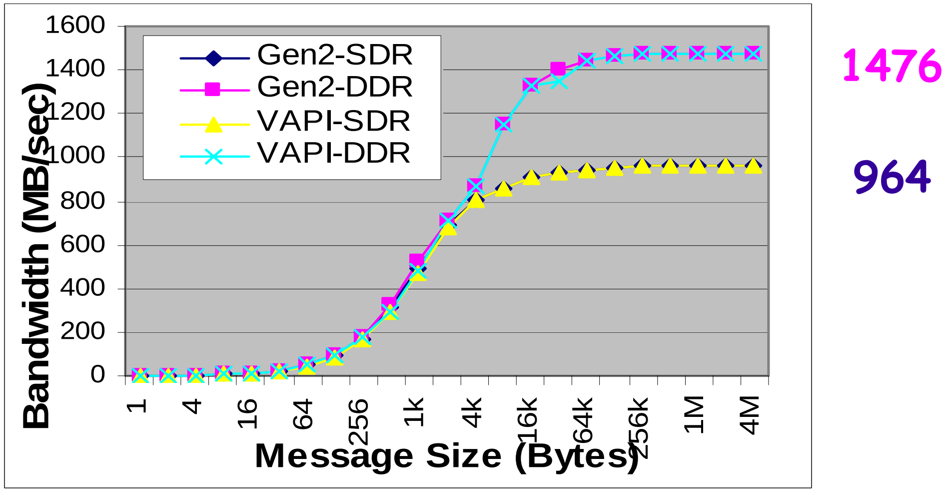
- Platforms
  - EM64T, Opteron, IA-32 and Mac G5
- Operating Systems
  - Linux, Solaris and Mac OSX
- Compilers
  - gcc, intel, pathscale and pgi
- InfiniBand Adapters
  - Mellanox adapters with PCI-X and PCI-Express (SDR and DDR with mem-full and mem-free cards)
- TCP/IP support also exists (based on MPICH2)

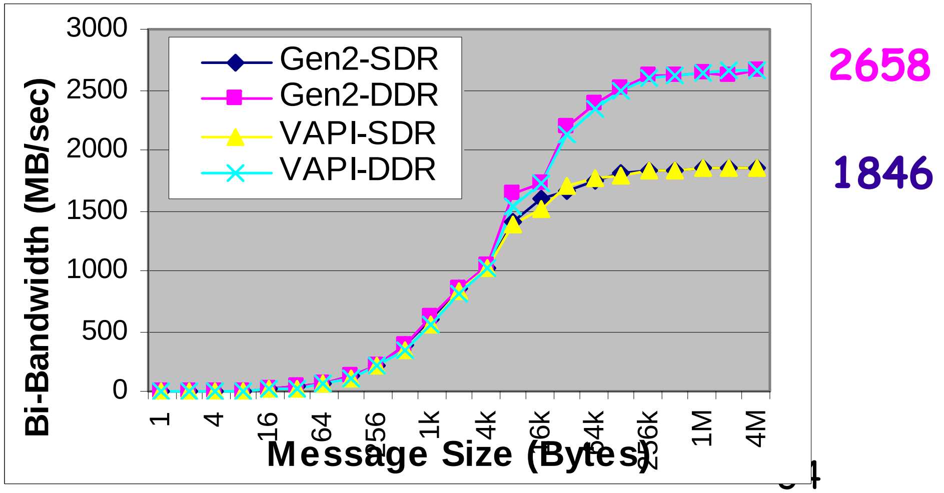
MVAPICH 0.9.2 Features

- **Released 01/11/06**
- High-Performance and Optimized Support for many MPI-2 functionalities
  - One-sided
  - Collectives
  - Datatype
- Support for other MPI-2 functionalities (as provided by MPICH2)
- High-Performance and Scalable ADI3-level design
- Optimized and scalable one-sided operations
  - Communication Calls
    - Get
    - Put
    - Accumulate
  - Synchronization Calls
    - Fence
    - General active target synchronization
    - Passive (lock and unlock)
- Optimized shared memory support
  - Bus-based architecture
  - NUMA architectures
- Portability across multiple interconnects through uDAPL
  - InfiniBand
    - uDAPL over Gen2 on Linux
    - uDAPL over VAPI (IBGD) on Linux
    - uDAPL over IBTL on Solaris
  - Myrinet (DAPL-GM Beta)
- Optimized for scalability
  - Three different modes: small, medium, and large clusters
- MPD Support
- Shared Library support
- ROMIO Support for MPI-IO
- All features and performance of MVAPICH + One-sided and Portability

MVAPICH2-0.9.2 Performance with MPI-Level Two-Sided Communication





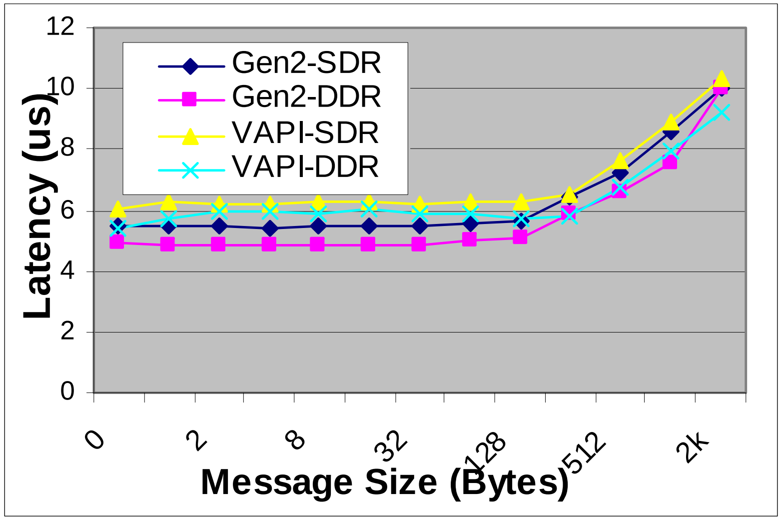


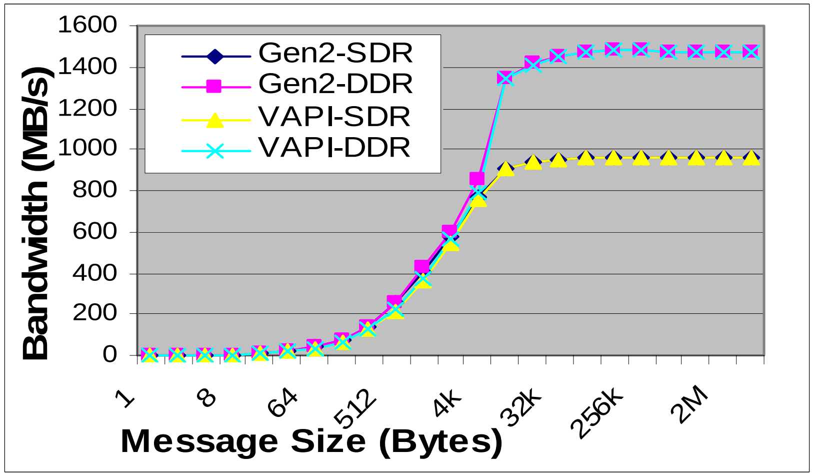
- Single port results only (EM64T, PCI-Ex)

Results for other platforms at
http://nowlab.cse.ohio-state.edu/projects/mpi-iba/

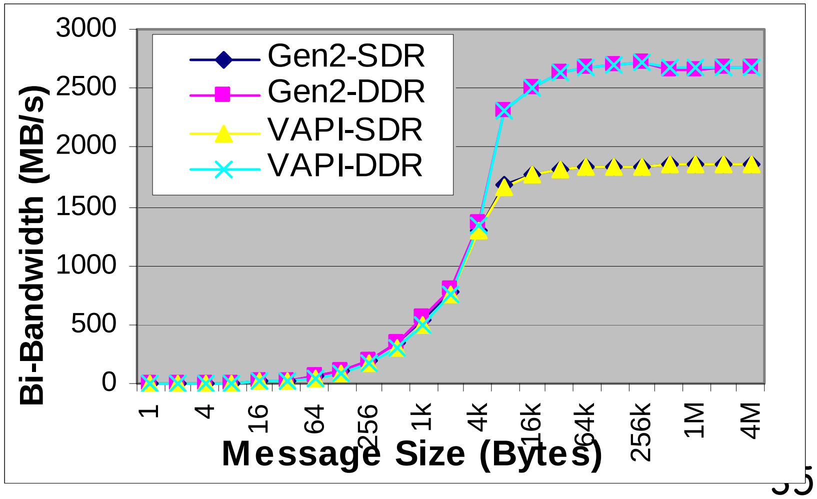
MVAPICH2-0.9.2 Performance with MPI One Sided Put (Active Target)

5.45
4.84





1476
964



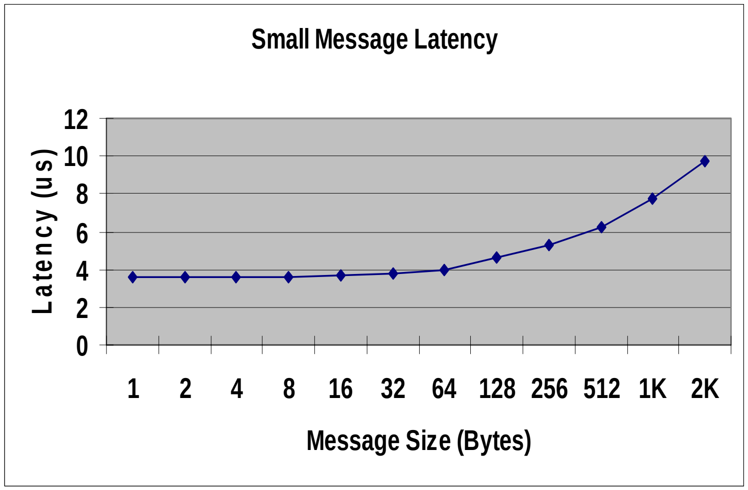
2667
1847

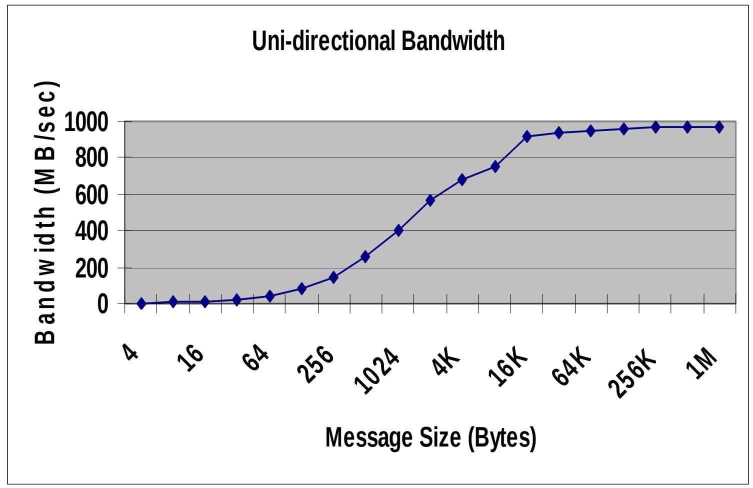
- Single port results only (EM64T, PCI-Ex)

Results for other platforms at
http://nowlab.cse.ohio-state.edu/projects/mpi-iba/

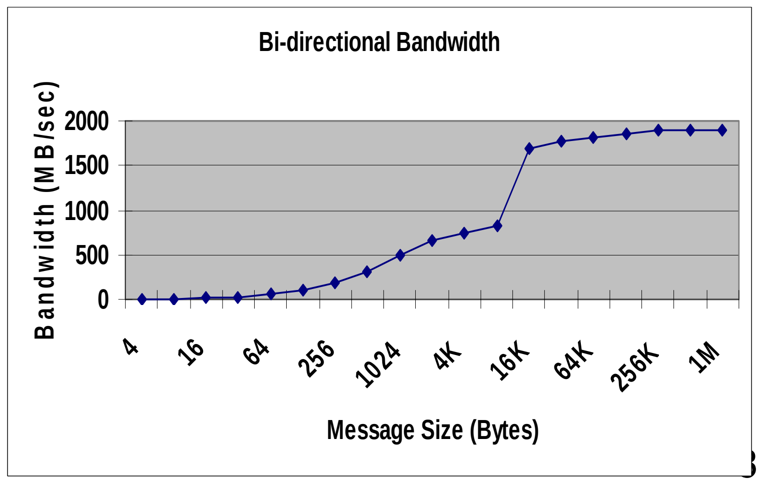
MVAPICH2-0.9.2 uDAPL/Gen2 over InfiniBand: MPI-Level Performance (Two-sided Operations)

3.57





968

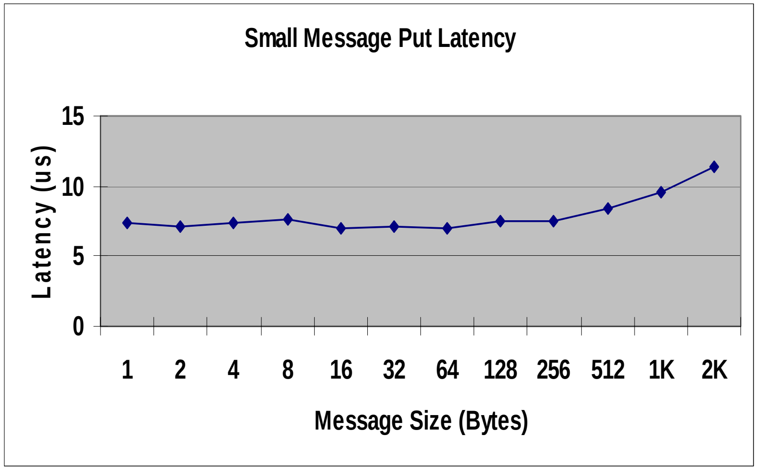


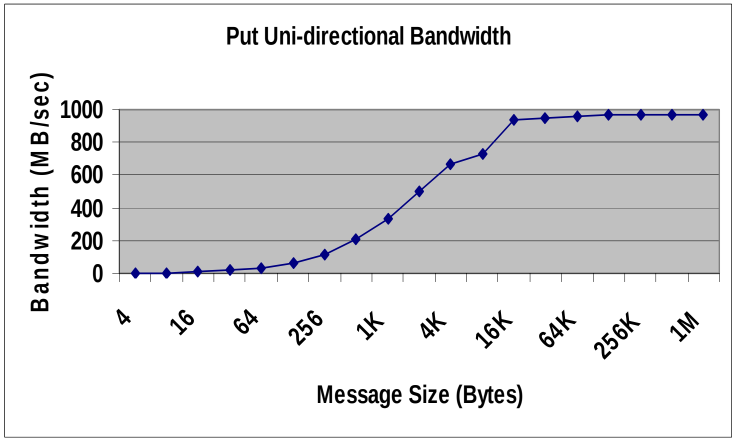
1914

EM64T, PCI-Ex, SDR

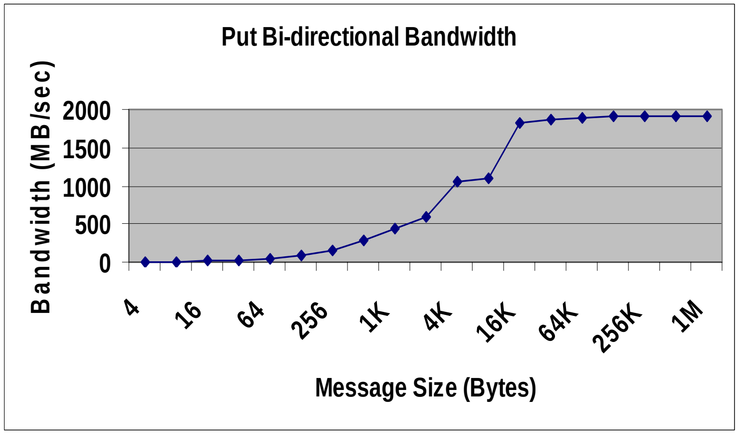
MVAPICH2-0.9.2 uDAPL/Gen2 over InfiniBand: MPI-Level Performance (One-sided Operations)

7.3





968

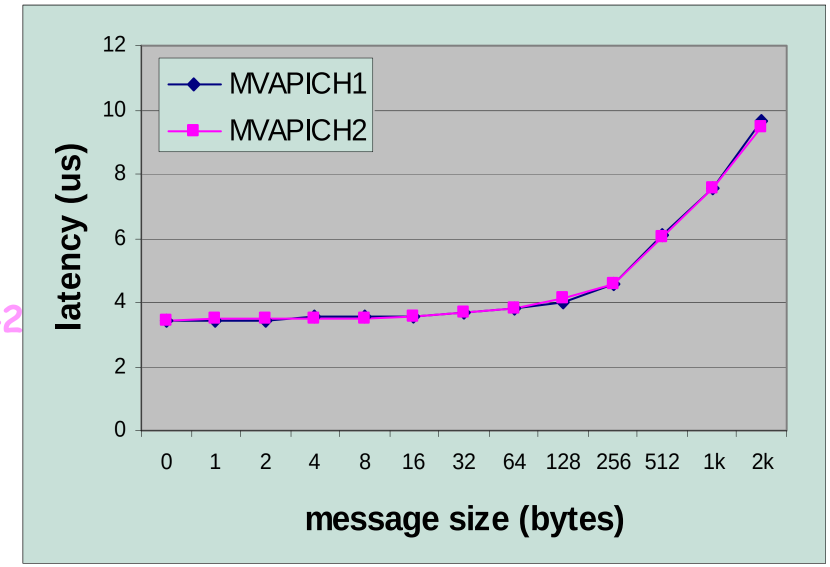


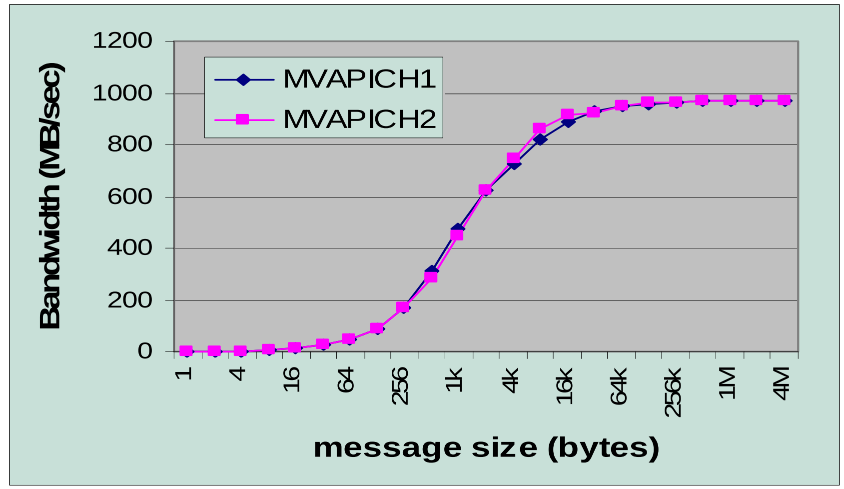
1914

EM64T, PCI-Ex, SDR

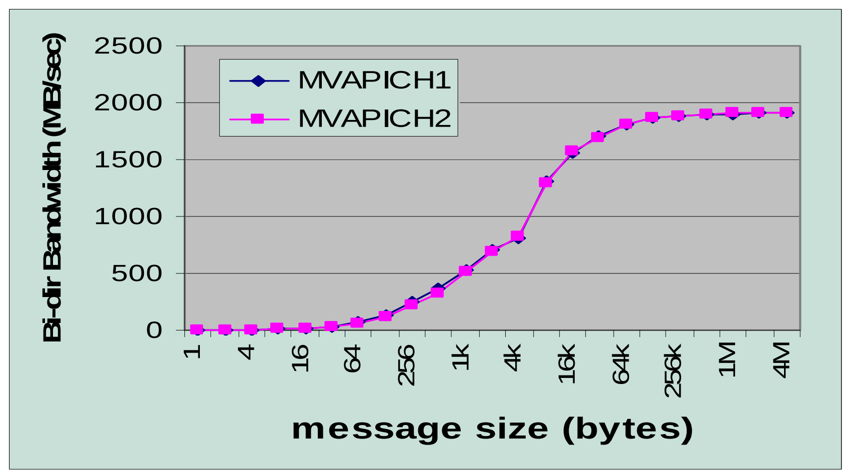
Performance Comparison of MAPICH-0.9.6 and MVAPICH2-0.9.2

3.42





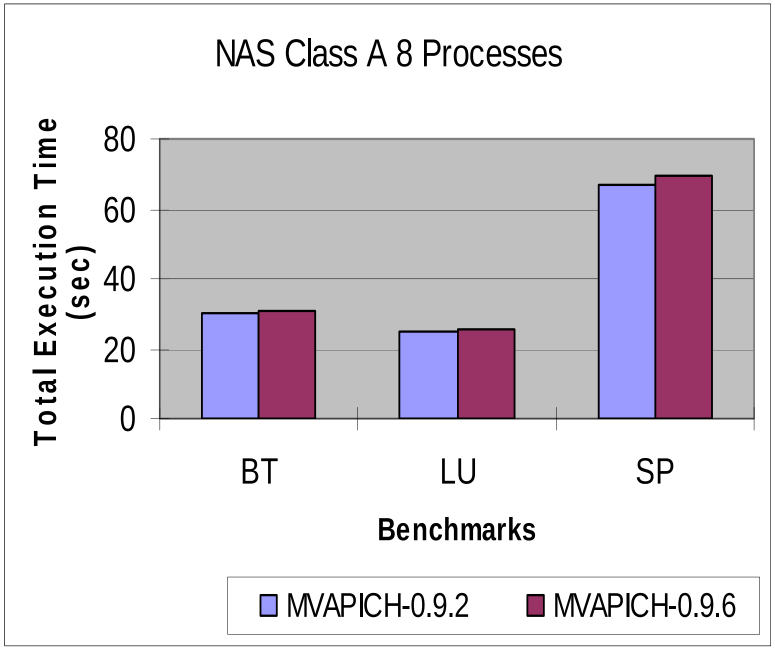
968

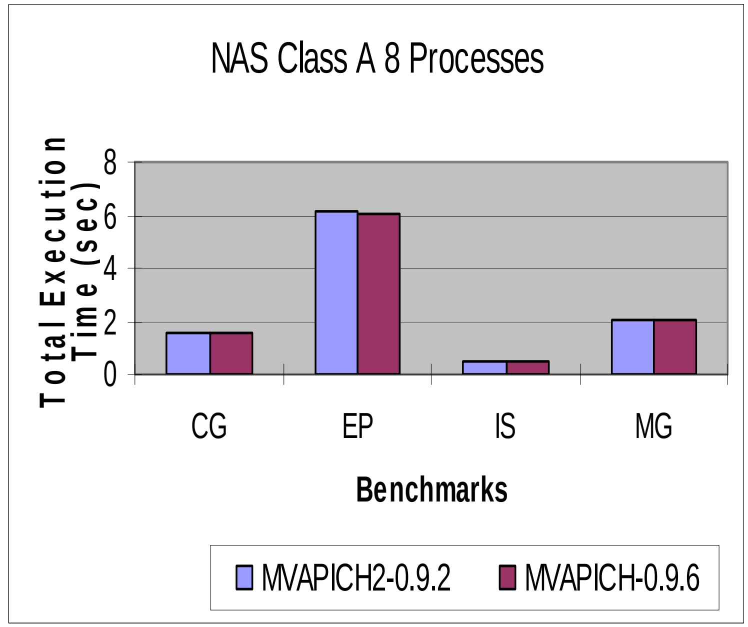


1914

- Single port results only (EM64T, PCI-Ex, SDR)

Performance Comparison: MVAPICH2-0.9.2 vs. MVAPICH-0.9.6





- MVAPICH2-0.9.2 performs very closely to MVAPICH-0.9.6
- Uses point-to-point and SMP
- MVAPICH 0.9.6 has added RDMA-based collectives which will be available with the next release of MVAPICH2
- MVAPICH2 0.9.2 has added one-sided communication

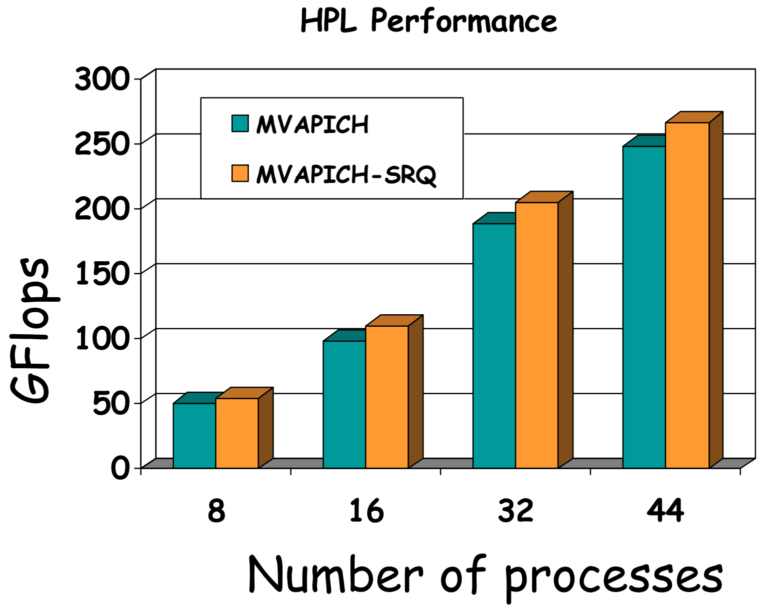
Presentation Overview

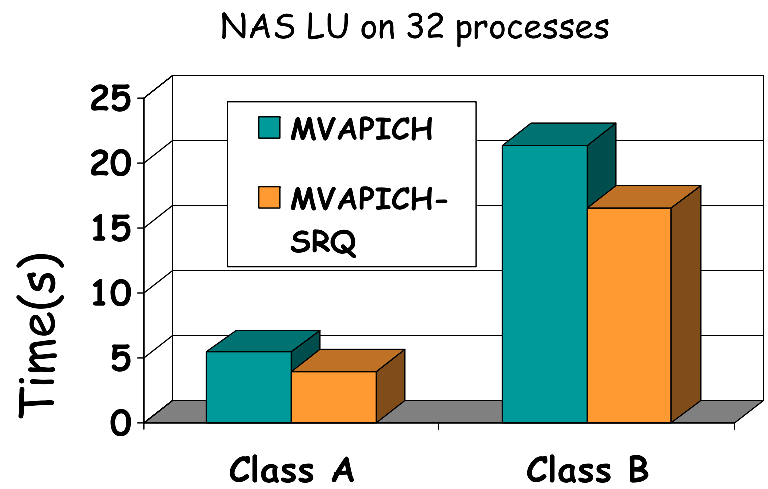
- Overview of MVAPICH and MVAPICH2 Projects
- MVAPICH 0.9.6 Features and Performance
  - Point-to-point
    - VAPI and Gen2
    - Mellanox and PathScale adapters
  - Adaptive RDMA Fast Path
  - RDMA Read
  - Collectives (Multicast, Barrier, All-to-All, All-gather)
  - Multi-rail support
  - Blocking support
  - uDAPL support
  - SDR/DDR Comparison
- MVAPICH2 0.9.2 Features and Performance
  - Two-sided (VAPI and Gen2)
  - One-sided (VAPI and Gen2)
  - uDAPL support
  - Comparison of 0.9.6 with 0.9.2
- Upcoming MVAPICH 0.9.7 Features and Performance
  - SRQ with Flow Control
  - Fault Tolerance
    - Memory-to-memory Reliability

MVAPICH 0.9.7

- Combines all features of MVAPICH together with Gen2
- Will be released in the next few weeks
- Additional features
  - SRQ with Flow Control
  - Fault Tolerance
    - Memory-to-memory Reliability
- High Performance and Scalable designs for
  - Point-to-point
  - Collectives
- Will support multiple interfaces
  - Gen2
  - VAPI
  - uDAPL
  - TCP/IP (based on MPICH)
- Can scale to multi-thousand nodes
- Additional architecture/platform to be supported
  - PPC/IBM

Performance impact of SRQ Flow control design

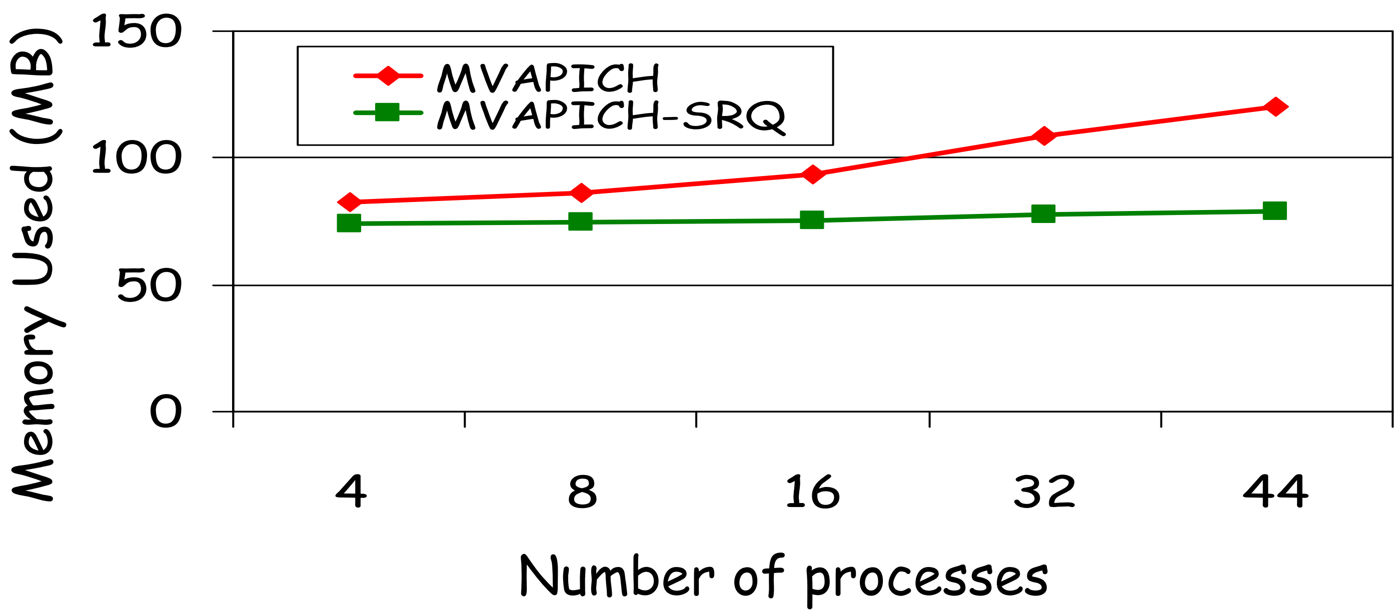




- MVAPICH-SRQ yields 7-8% benefit in overall HPL GFlops rating
- LU Class B performance is improved by 22% on 32 processes
- Benefits mainly stem from:
  - Reduced memory polling overhead
  - Enlarged window size due to unique flow control mechanism

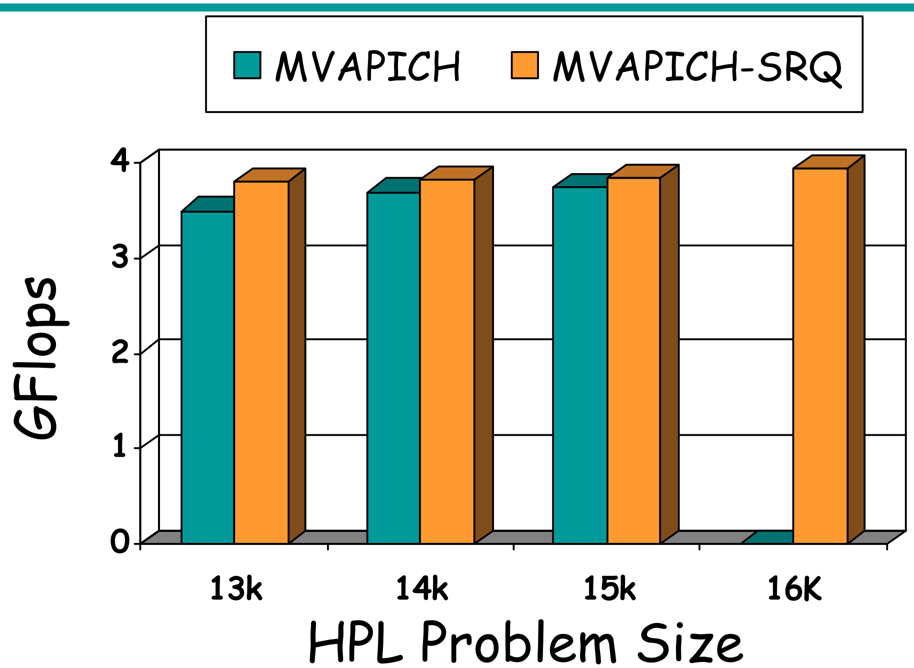
S. Sur, L. Chai, H.-W. Jin and D. K. Panda, *Shared Receive Queue Based MPI Design for InfiniBand Clusters*, Int'l Parallel and Distributed Processing Symposium (IPDPS), April 2006, to be presented

Memory Scalability with SRQ



- Startup memory utilization on for a real MPI process is reduced
- Analytical model predicts that MVAPICH-SRQ needs only 300MB of MPI internal buffers for a cluster of 16,000 nodes.
- Combined with Adaptive Connection Management, a MPI program on 16,000 nodes would require less than 500MB of registered memory at startup

Impact of Memory Scalability on HPL



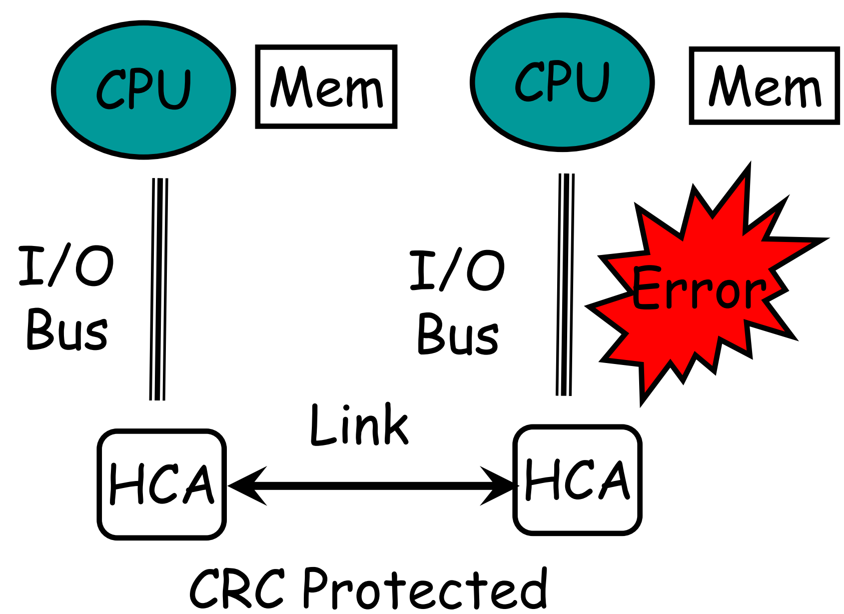
- HPL is run on 4 nodes with increasing problem size
- Increasing problem size increases memory consumption
- Larger problem sizes yield better GFlop ratings
- MVAPICH-SRQ is able to run larger problem sizes due to less consumption of memory by MPI library

Fault Tolerance

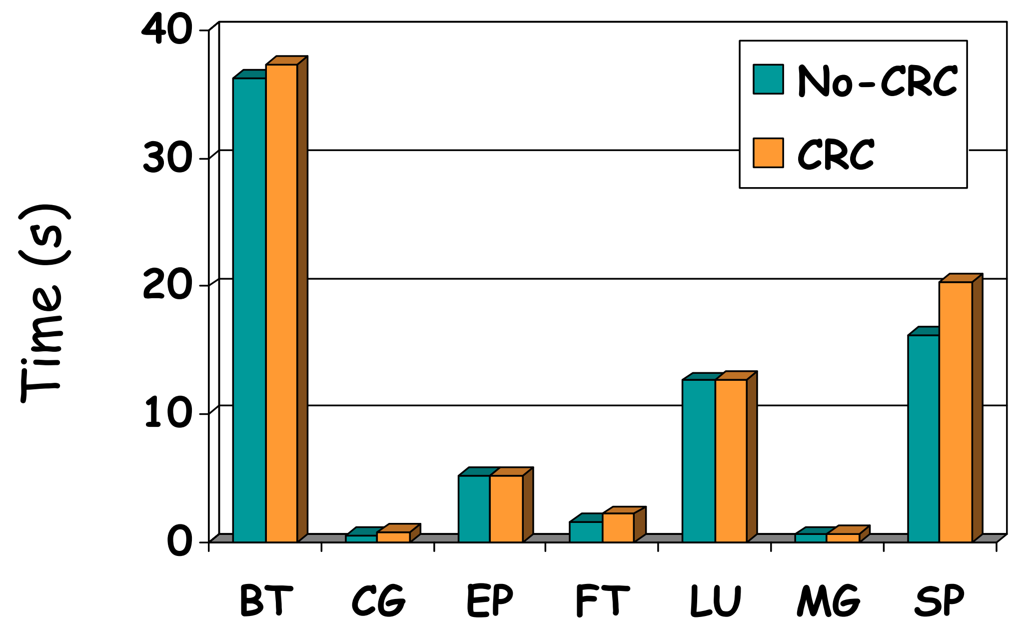
- Component failures are the norm in large-scale clusters
- Imposes need on reliability and fault tolerance
- Working along the following three angles
  - End-to-end Reliability with memory-to-memory CRC
    - Will be available with MVAPICH 0.9.7 and MVAPICH2
  - Reliable Networking with Automatic Path Migration (APM) utilizing Redundant Communication Paths
    - Will be available with MVAPICH2 0.9.3
  - Process Fault Tolerance with Efficient Checkpoint and Restart
    - Will be available with MVAPICH2 0.9.4

Memory-to-Memory Reliability

- InfiniBand enforces HCA to HCA reliability using CRC
- No check to see if data is transmitted reliably over I/O Bus
- In different situations (high-altitudes or in hotter climates), error rate increases sharply
- MVAPICH uses CRC-32 bit algorithm to ensure safe message delivery



Impact of Reliable mode on Performance



- NAS Benchmarks (Class A) are run in 8x2 mode
- Impact on end application performance is relatively small

Presentation Overview

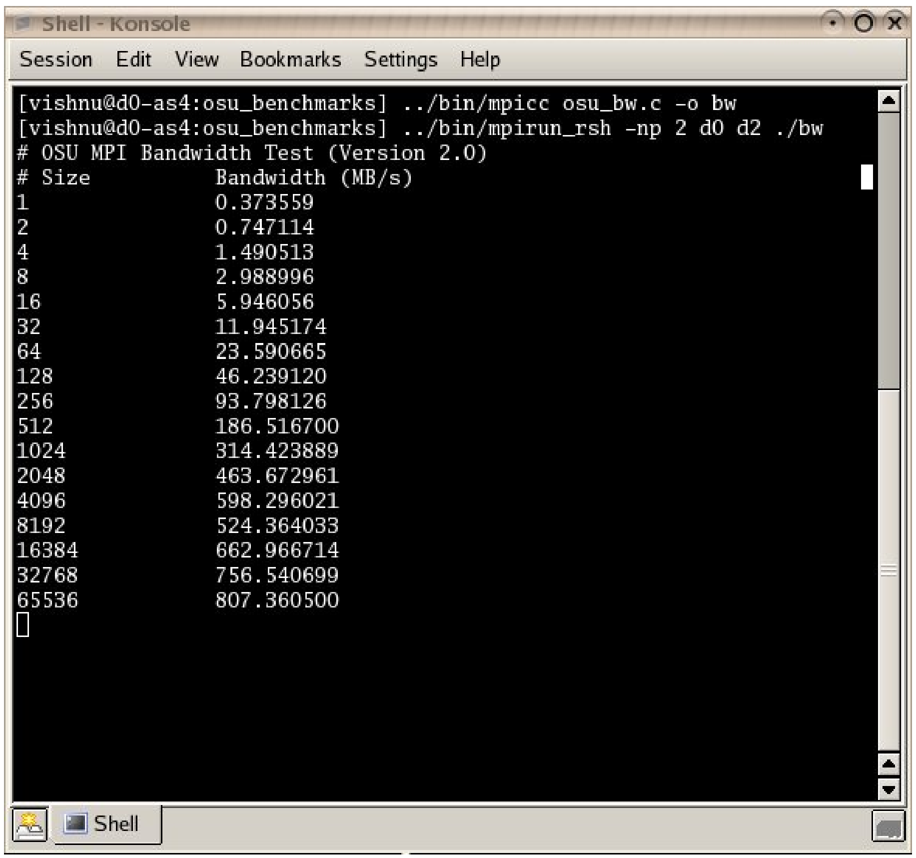
- Upcoming Features and Sample Performance
  - Fault Tolerance
    - Checkpoint-Restart
    - Automatic Path Migration (APM)
  - Multithreading
  - Multi-Network Support with uDAPL
  - Adaptive Connection Management
  - QoS Features and Routing
- Overview of Additional Projects
  - SDP
  - iWARP
  - Lustre, GFS, NFS over RDMA
  - Xen over IB
  - Multi-tier DataCenter
- Conclusions

Network-Level Fault Tolerance with APM

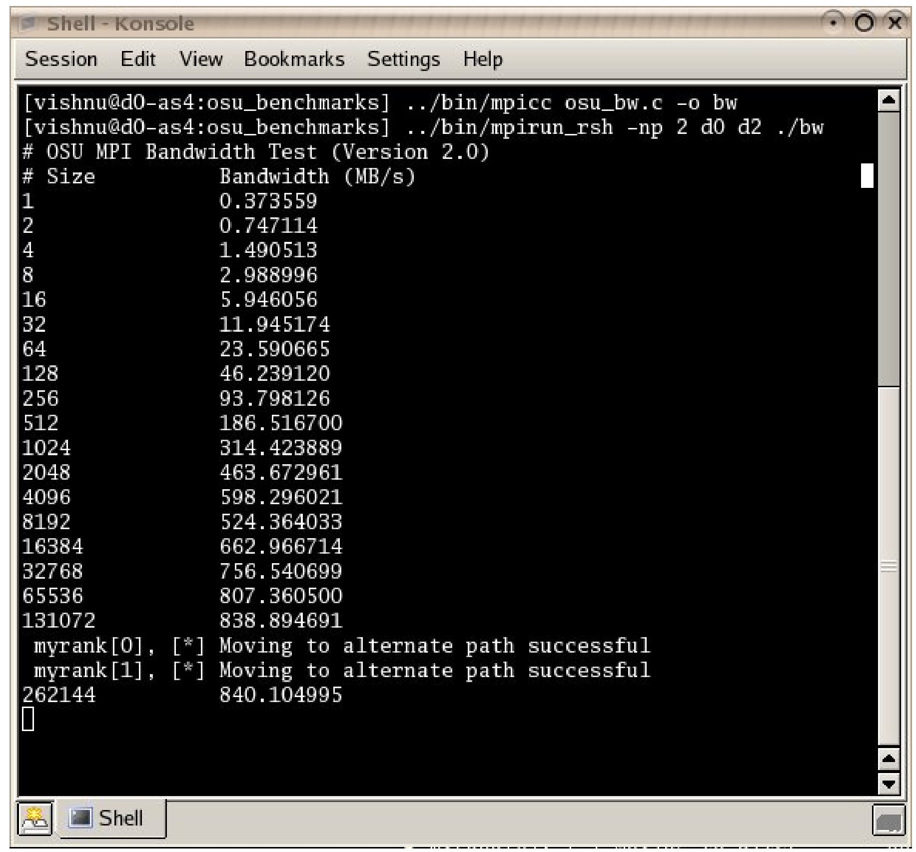
- Designed a solution using InfiniBand Automatic Path Migration (APM) Hardware mechanism
  - Utilizes Redundant Communication Paths
    - Multiple Ports
    - LMC
- Available for VAPI only because Gen2 does not support APM yet

Screenshots: APM with OSU Bandwidth test

Step #1: Bandwidth Test Running

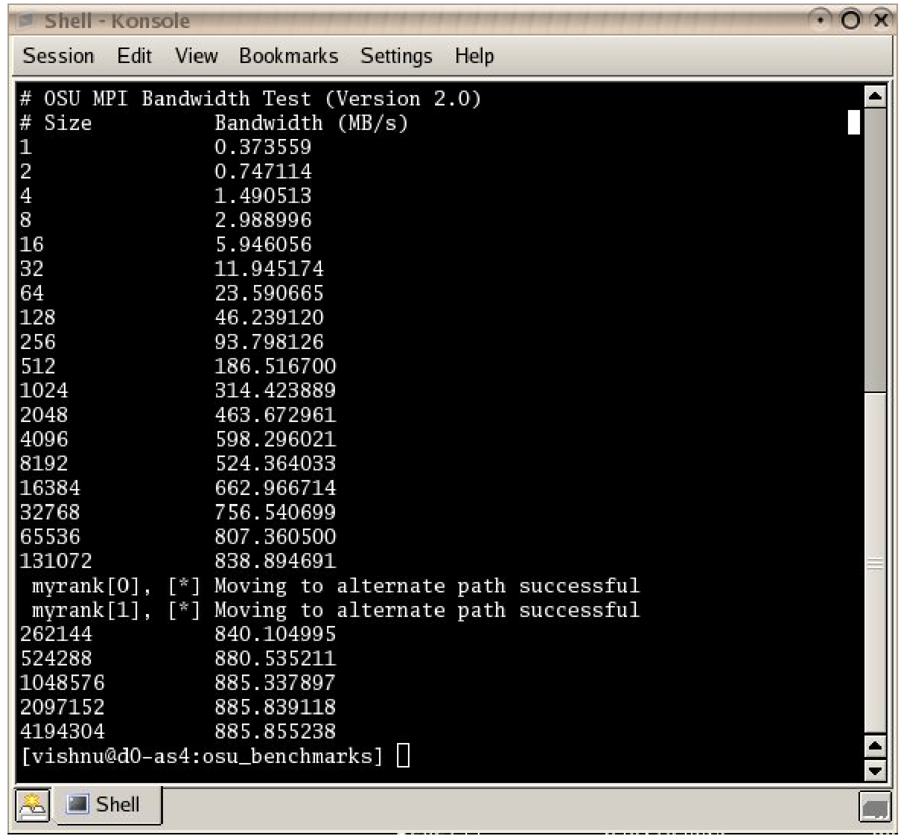


Step #2: Fault on Link, APM Triggered



Screenshots: APM with OSU Bandwidth test

Step #3:
Bandwidth Test Resumes and Finishes



Checkpoint/Restart Support for MVAPICH2

- Process-level Fault Tolerance
  - User-transparent, system-level checkpointing
  - Based on BLCR from LBNL to take coordinated checkpoints of entire program, including front end and individual processes
  - Designed novel schemes to
    - Coordinate all MPI processes to drain all in flight messages in IB connections
    - Store communication state and buffers, etc. while taking checkpoint
    - Restarting from the checkpoint

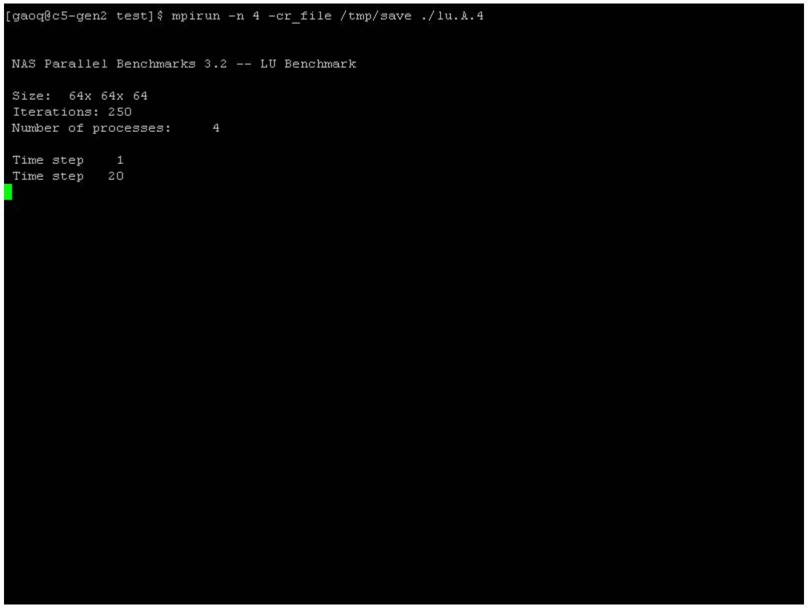
A Running Example

- Show how to checkpoint/restart LU from NAS benchmark
- There are two terminals:
  - Left one for normal run
  - Right one for checkpoint/restart

A Running Example (Cont.)

Terminal A:

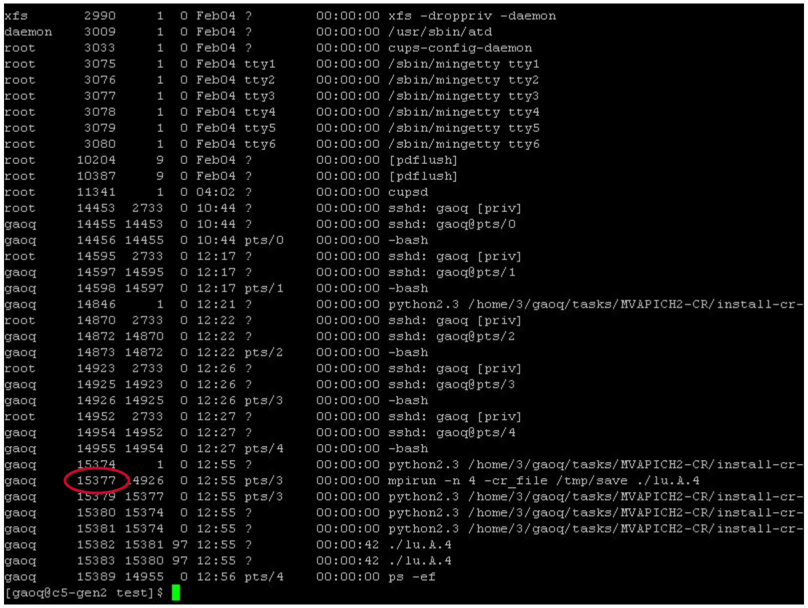
Start running LU



1

Terminal B:

Get its PID



2

A Running Example (Cont.)

Terminal A:

LU is running

```
[gaoq@c5-gen2 test]$ mpirun -n 4 -cr_file /tmp/save ./lu.A.4


 NAS Parallel Benchmarks 3.2 -- LU Benchmark

 Size:  64x 64x 64
 Iterations: 250
 Number of processes:     4

 Time step    1
 Time step   20
 Time step   40
 Time step   60
 Time step   80
 Time step  100
 Time step  120
 Time step  140
```

3

Terminal B:

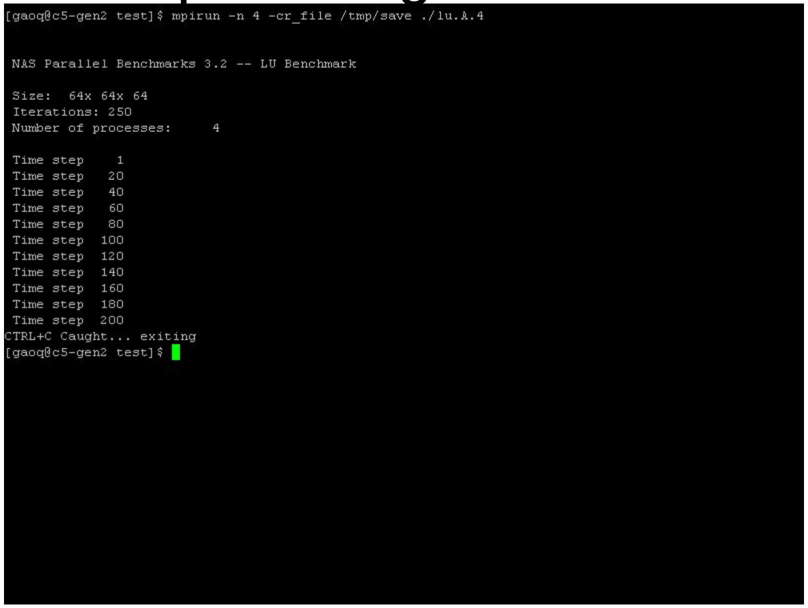
Now, Take checkpoint

```
root      3033     1  0 Feb04 ?        00:00:00 cups-config-daemon
root      3075     1  0 Feb04 tty1     00:00:00 /sbin/mingetty tty1
root      3076     1  0 Feb04 tty2     00:00:00 /sbin/mingetty tty2
root      3077     1  0 Feb04 tty3     00:00:00 /sbin/mingetty tty3
root      3078     1  0 Feb04 tty4     00:00:00 /sbin/mingetty tty4
root      3079     1  0 Feb04 tty5     00:00:00 /sbin/mingetty tty5
root      3080     1  0 Feb04 tty6     00:00:00 /sbin/mingetty tty6
root     10204     9  0 Feb04 ?        00:00:00 [pdflush]
root     10387     9  0 Feb04 ?        00:00:00 [pdflush]
root     11341     1  0 04:02 ?        00:00:00 cupsd
root     14453  2733  0 10:44 ?        00:00:00 sshd: gaoq [priv]
gaoq     14455 14453  0 10:44 ?        00:00:00 sshd: gaoq@pts/0
gaoq     14456 14455  0 10:44 pts/0    00:00:00 -bash
root     14595  2733  0 12:17 ?        00:00:00 sshd: gaoq [priv]
gaoq     14597 14595  0 12:17 ?        00:00:00 sshd: gaoq@pts/1
gaoq     14598 14597  0 12:17 pts/1    00:00:00 -bash
gaoq     14846     1  0 12:21 ?        00:00:00 python2.3 /home/3/gaoq/tasks/MVAPICH2-CR/install-cr-
root     14870  2733  0 12:22 ?        00:00:00 sshd: gaoq [priv]
gaoq     14872 14870  0 12:22 ?        00:00:00 sshd: gaoq@pts/2
gaoq     14873 14872  0 12:22 pts/2    00:00:00 -bash
root     14923  2733  0 12:26 ?        00:00:00 sshd: gaoq [priv]
gaoq     14925 14923  0 12:26 ?        00:00:00 sshd: gaoq@pts/3
gaoq     14926 14925  0 12:26 pts/3    00:00:00 -bash
root     14952  2733  0 12:27 ?        00:00:00 sshd: gaoq [priv]
gaoq     14954 14952  0 12:27 ?        00:00:00 sshd: gaoq@pts/4
gaoq     14955 14954  0 12:27 pts/4    00:00:00 -bash
gaoq     15374     1  0 12:55 ?        00:00:00 python2.3 /home/3/gaoq/tasks/MVAPICH2-CR/install-cr-
gaoq     15377 14926  0 12:55 pts/3    00:00:00 mpirun -n 4 -cr_file /tmp/save ./lu.A.4
gaoq     15379 15377  0 12:55 pts/3    00:00:00 python2.3 /home/3/gaoq/tasks/MVAPICH2-CR/install-cr-
gaoq     15380 15374  0 12:55 ?        00:00:00 python2.3 /home/3/gaoq/tasks/MVAPICH2-CR/install-cr-
gaoq     15381 15374  0 12:55 ?        00:00:00 python2.3 /home/3/gaoq/tasks/MVAPICH2-CR/install-cr-
gaoq     15382 15381 97 12:55 ?        00:00:42 ./lu.A.4
gaoq     15383 15380 97 12:55 ?        00:00:42 ./lu.A.4
gaoq     15389 14955  0 12:56 pts/4    00:00:00 ps -ef
[gaoq@c5-gen2 test]$ checkpoint 15377
Checkpoint Done
[gaoq@c5-gen2 test]$
```

4

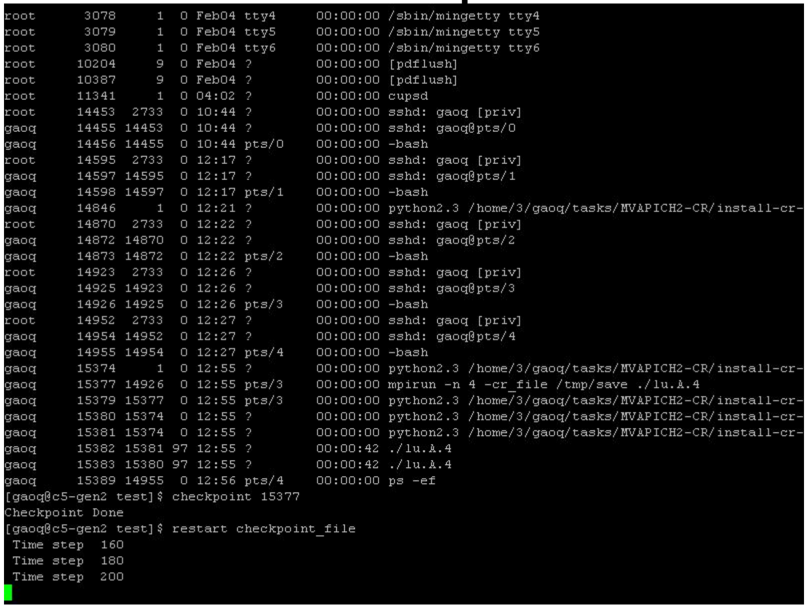
A Running Example (Cont.)

Terminal A:
LU is not affected.
Stop it using CTRL-C



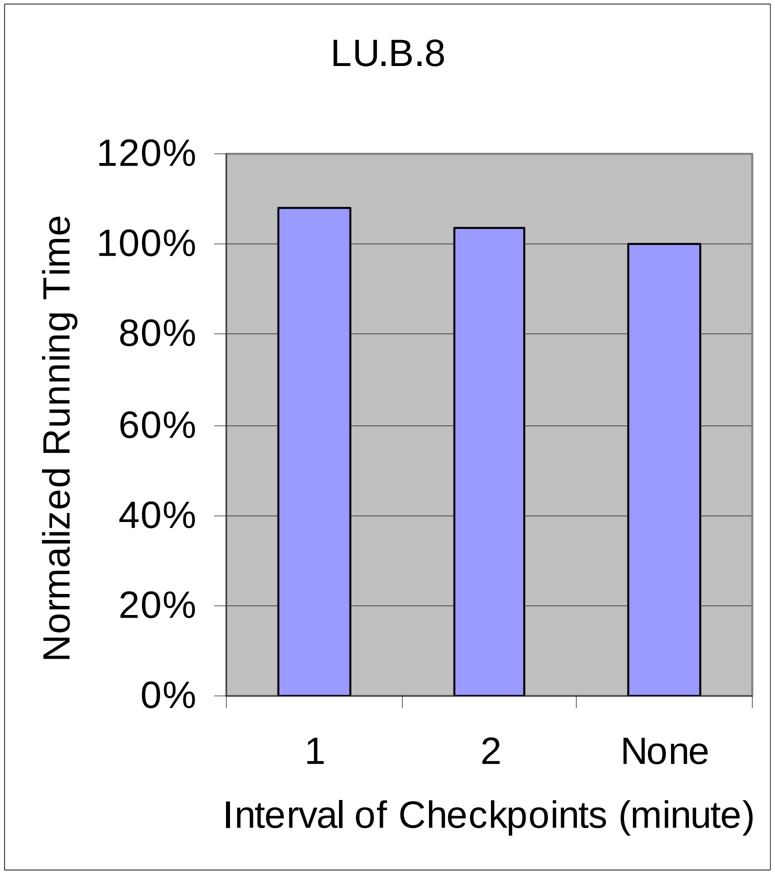
5

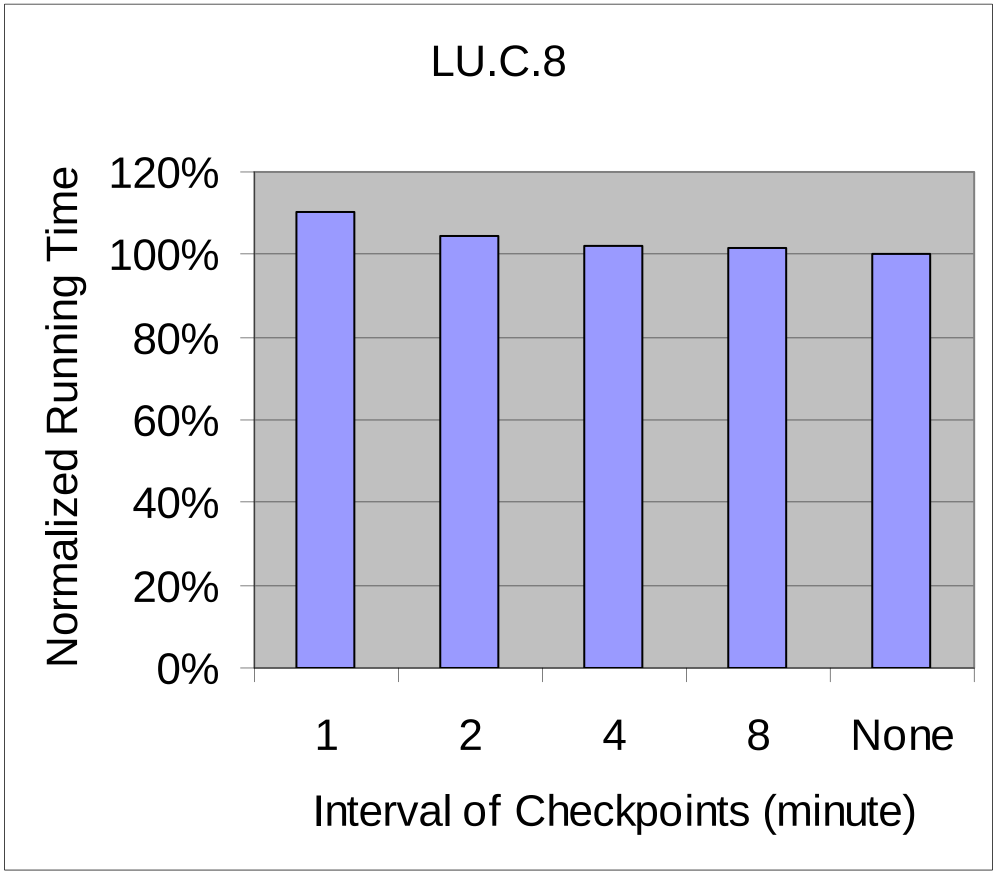
Terminal B:
Then, restart from
the checkpoint



6

Performance Impact for Checkpointing



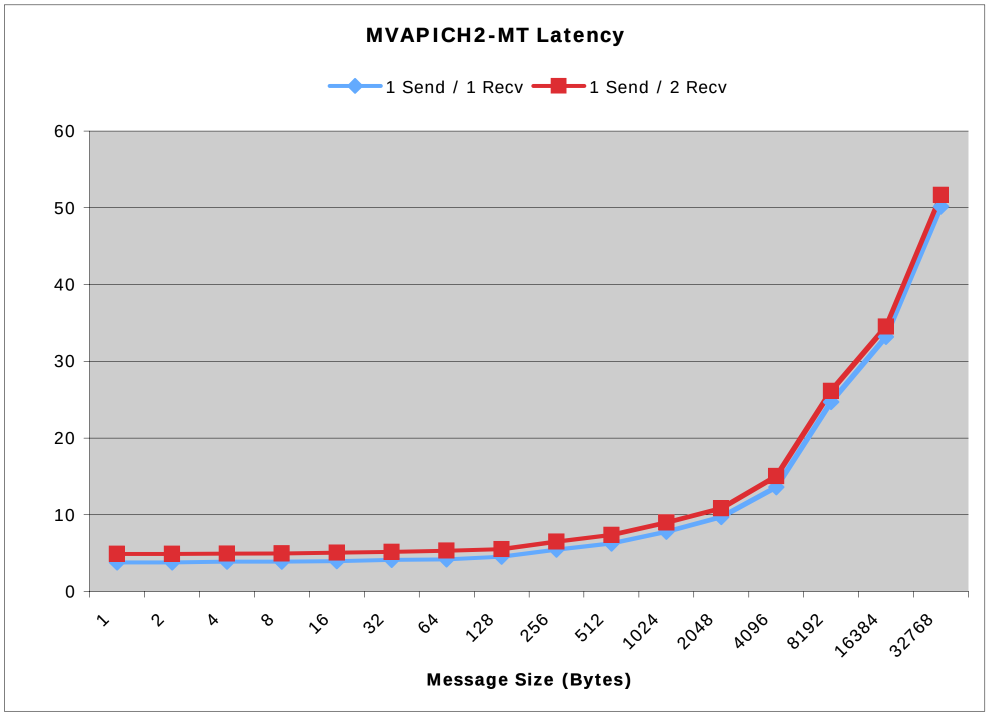


Very Little Overhead

Multithreading

- Emerging Multi-core architectures promise performance boost for Multi-threaded applications
- MVAPICH2 has a prototype design of Multi-threaded support to enable Multi-threaded applications
- Additional designs are being studied
- Will be released with MVAPICH2 0.9.3 in a few weeks

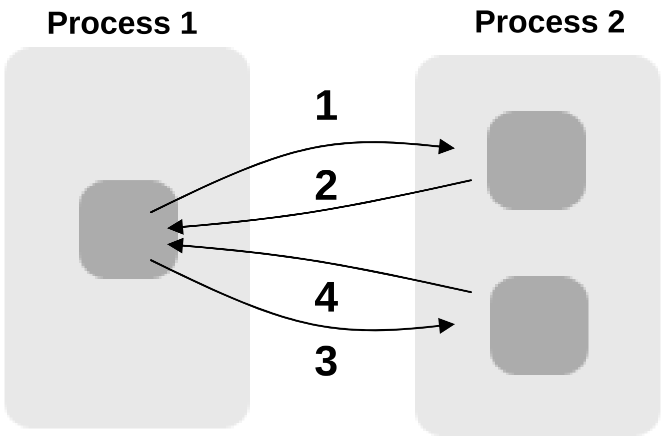
Multithreaded Latency Test



•Overhead is very small, ~1us even when there is severe contention.
• No performance impact when no contention

Ping-Pong Latency Test

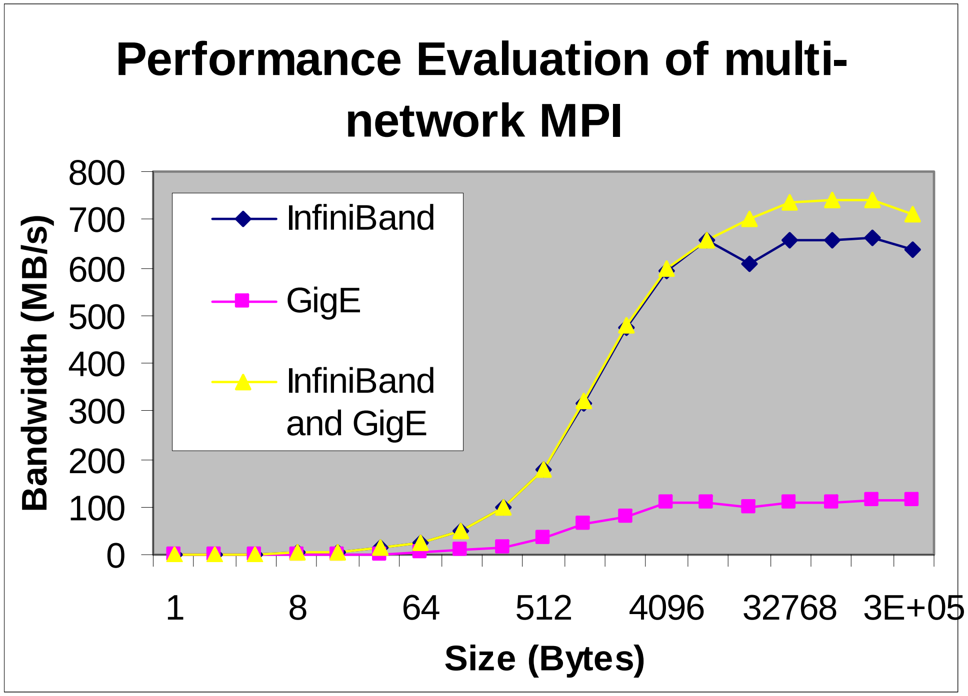
• Reference case: one thread per process.
•Multithreaded case: one thread on one process and two on the other



Multi-Network Support using uDAPL

- Clusters with different RDMA-enabled Interconnects are being deployed
- A combination of these interconnects can be used for performance/fault-tolerance
- Network-independent interfaces like uDAPL have become available
  - How do we design support for multi-network using network independent interfaces?

Performance Evaluation: Multi-Network MVAPICH/uDAPL



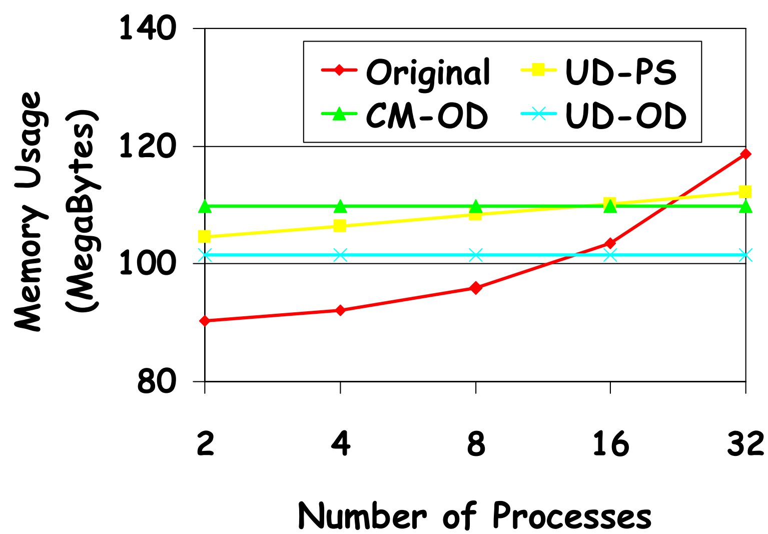
- Weighted Striping is used for scheduling between networks (10:1 for IB:GigE)
- Peak Bandwidth increases from 659 MB/s to 750MB/s for 128K message size

Adaptive Connection Management

- Problems with Static Connections
  - Prolonged startup time
  - Low connection utilization
  - Heavy resource usage
- Adaptive Connection Management
  - Establish connection for processes with frequent communication pattern
    - On-demand connection management:
      Establish a new RC connection only at the time it is needed.
    - Partially static:
      Start with 2*logN connections to match with binary search tree commonly used in collective algorithms, and establish additional RC connections as needed.
    - Using UD or IBCM for establishing new RC connections

W. Yu, Q. Gao and D.K. Panda, *Adaptive Connection Management for Scalable MPI over InfiniBand*. International Parallel and Distributed Processing Symposium, Rhodes Island, Greece, April 2006. To be presented.

Scalable Memory Usage



- Adaptive connection management can help achieve scalable memory usage
- Initial memory usage is reduced to logarithmic with partially static and a minimum constant with on-demand
- Working on optimal solutions together with SRQ + Flow Control for MVAPICH to scale to tens of thousands of nodes and higher

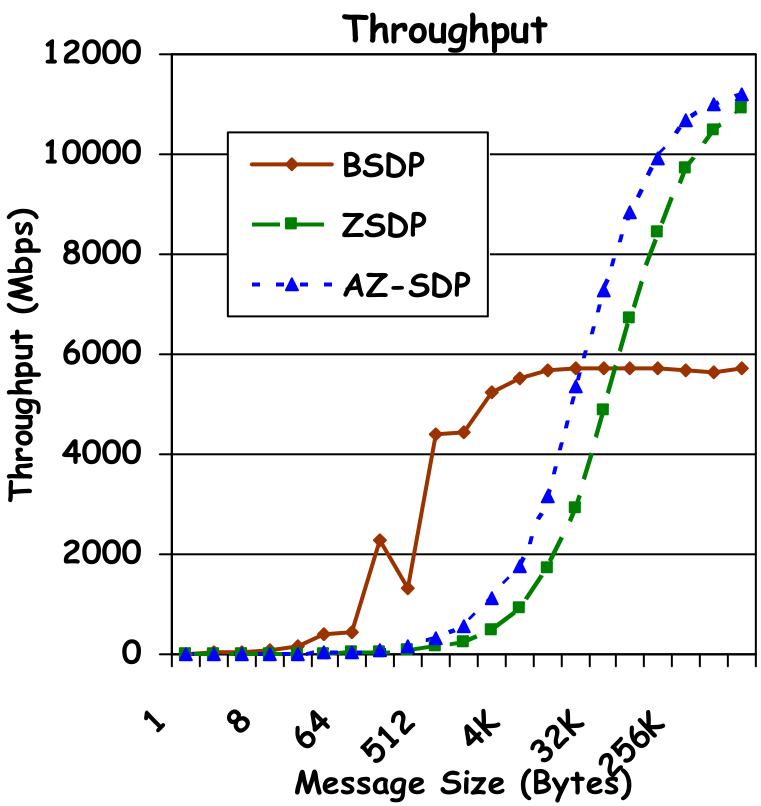
QoS Features, Routing and Added Features

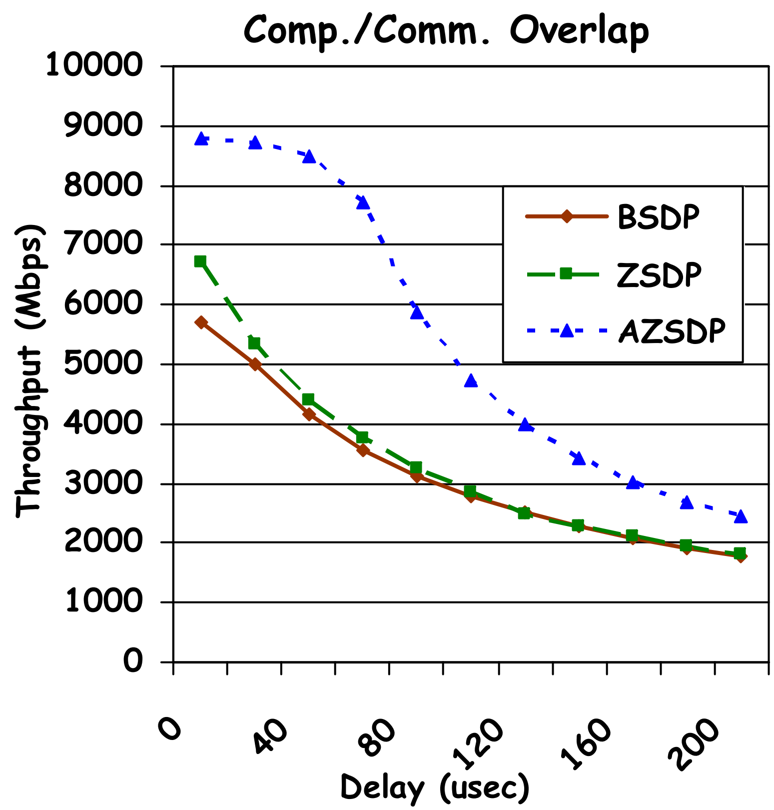
- As multi-thousand nodes with IB are deployed, many open challenges exist for
  - Usage of SL for traffic differentiation
    - Pt-to-pt and collective
  - Identifying optimal paths in the fabric
    - Support adaptive routing
  - Carrying out topology-aware collective operations
  - UD-based communication
  - Using kernel-based multicast support
- Requires support from SM and CM
  - Many of these mechanisms are not available yet …. gradually being available in SM and CM and core modules
- Carrying out research on these angles and solutions will be available soon

Presentation Overview

- Upcoming Features and Sample Performance
  - Fault Tolerance
    - Checkpoint-Restart
    - Automatic Path Migration (APM)
  - Multithreading
  - Multi-Network Support with uDAPL
  - Adaptive Connection Management
  - QoS Features and Routing
- Overview of Additional Projects
  - SDP
  - iWARP
  - Lustre, GFS, NFS over RDMA
  - Xen over IB
  - Multi-tier DataCenter
- Conclusions

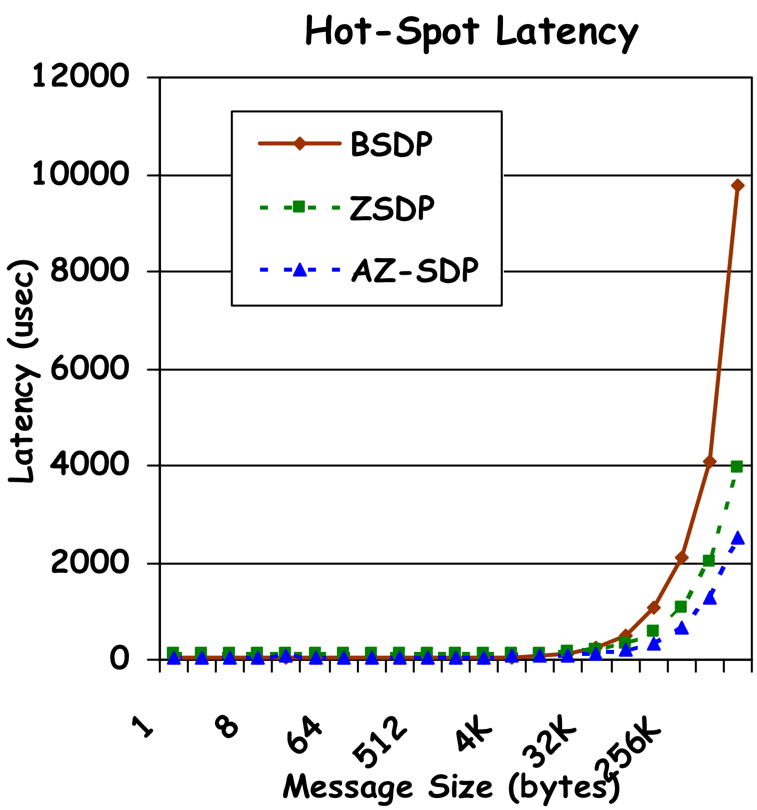
Different SDP Implementations (PCI-Express, DDR)

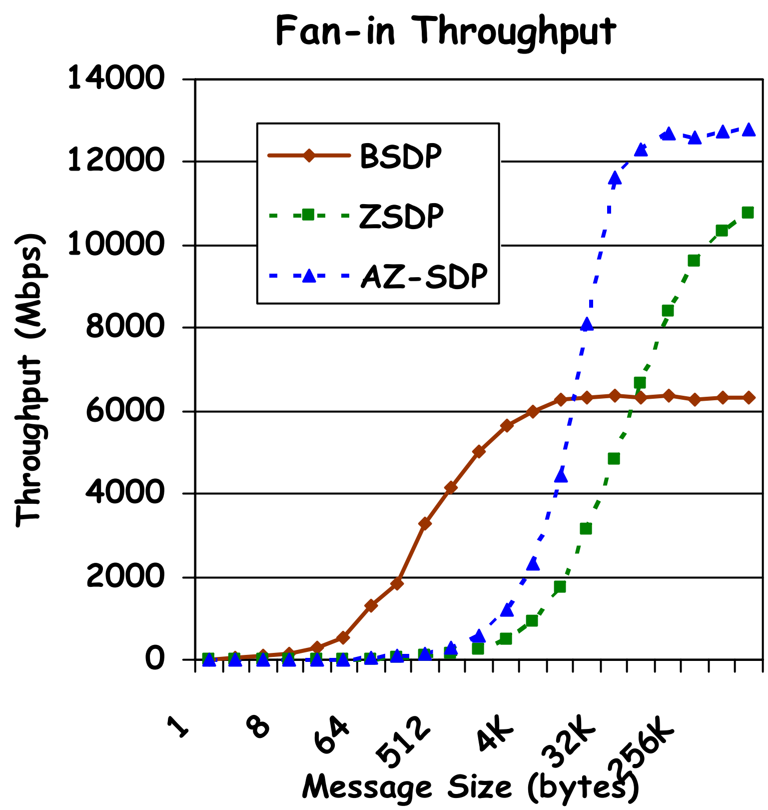




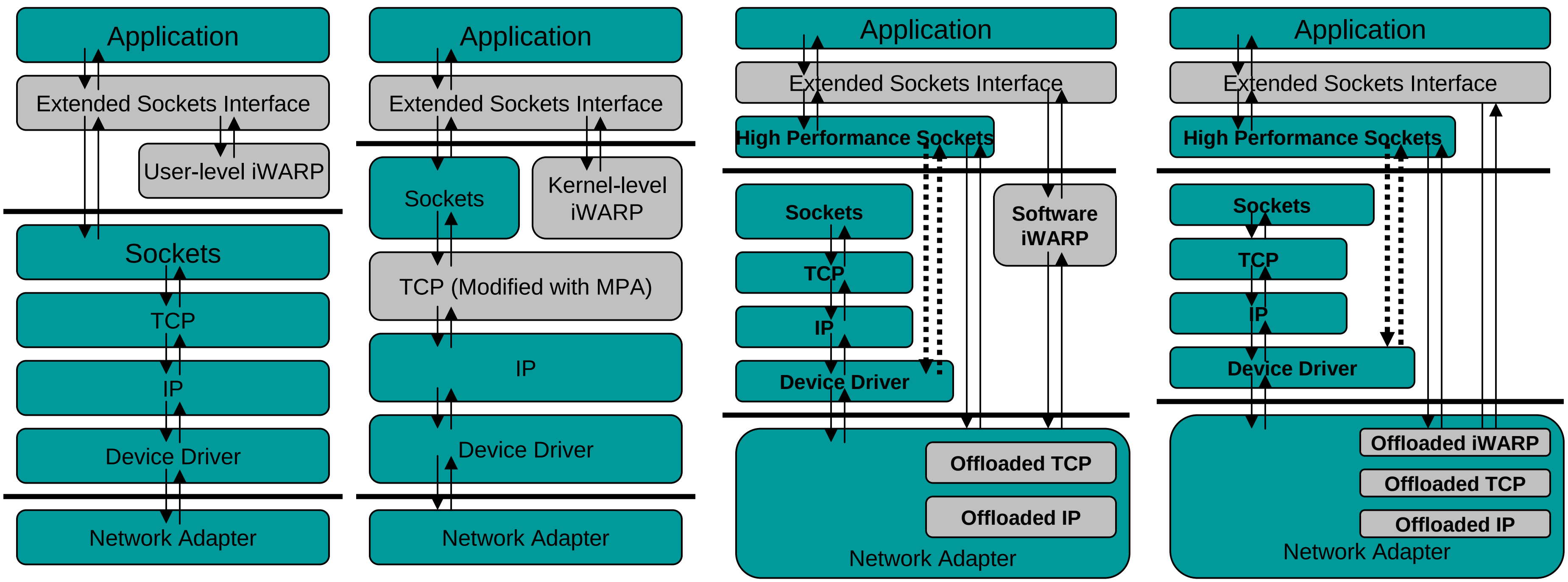
P. Balaji, S. Bhagvat, H. -W. Jin and D. K. Panda, Asynchronous Zero-copy Communication for Synchronous Sockets in the Sockets Direct Protocol (SDP) over InfiniBand, to be presented at CAC '06, in conjunction with IPDPS '06, April 2006

Multi-Connection Benchmarks





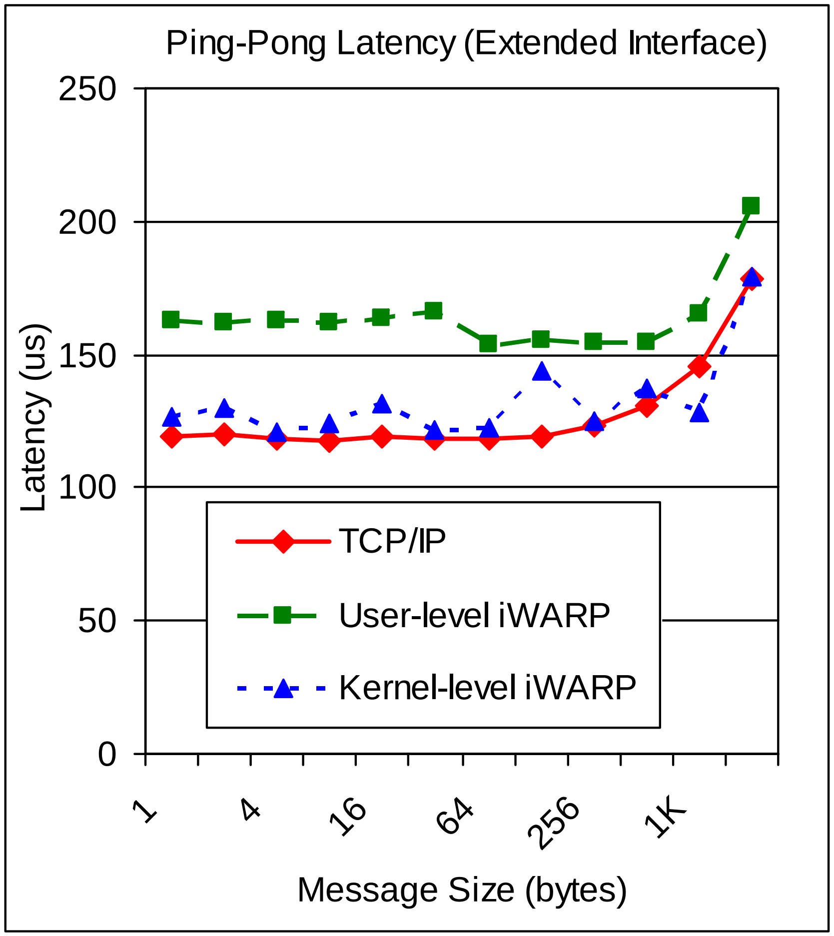
Software iWARP and Extended Sockets Interface

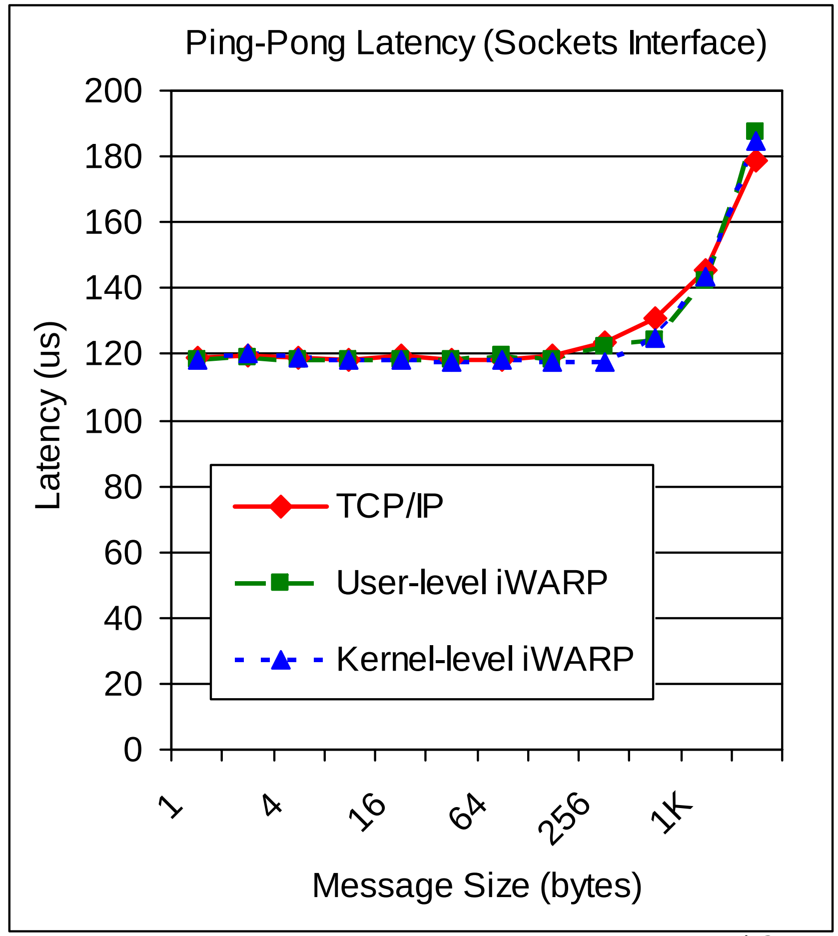


Regular Ethernet Adapters **TCP Offload Engines** **iWARP compliant Adapters**

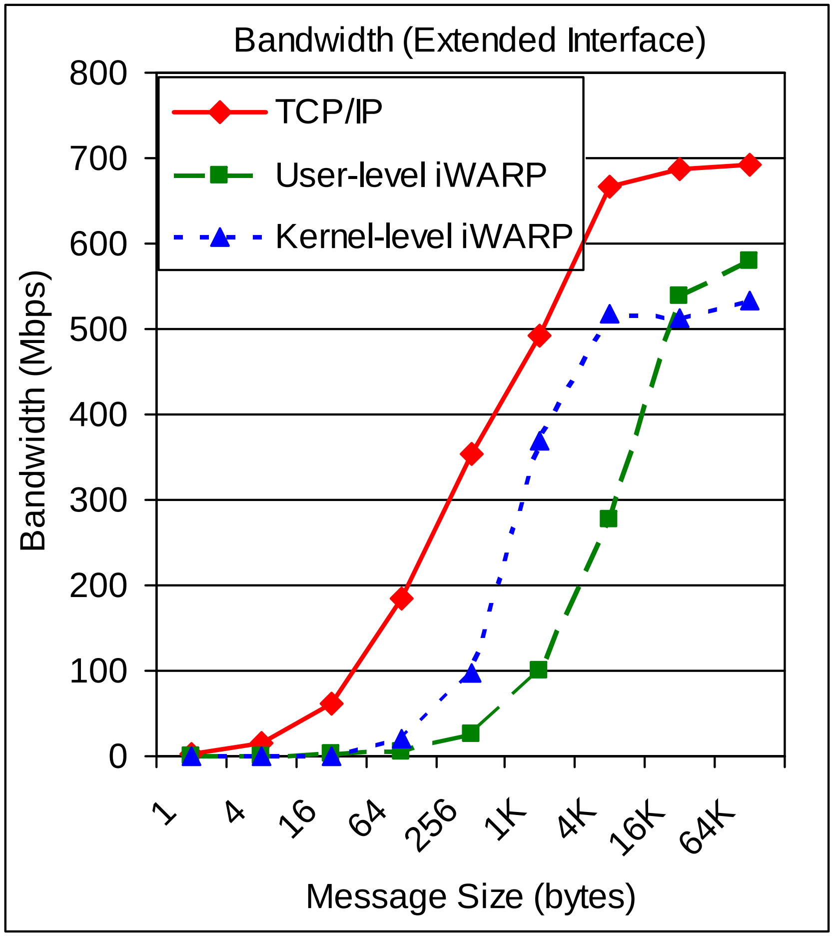
P. Balaji, H. -W. Jin, K. Vaidyanathan and D. K. Panda, Supporting iWARP compatibility and features, (RAIT 2005), Sept. 2005, in conjunction with the IEEE Cluster 2005.

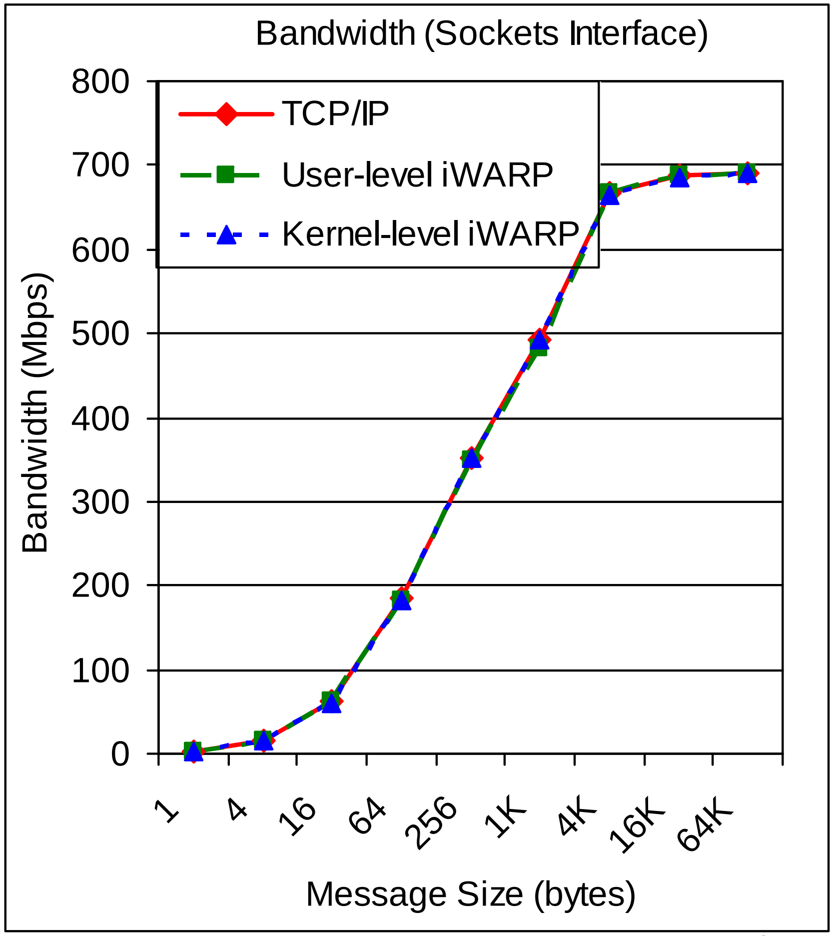
Ping-Pong Latency Test





Uni-directional Stream Bandwidth Test





PVFS (PVFS-1 and PVFS-2)

- 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

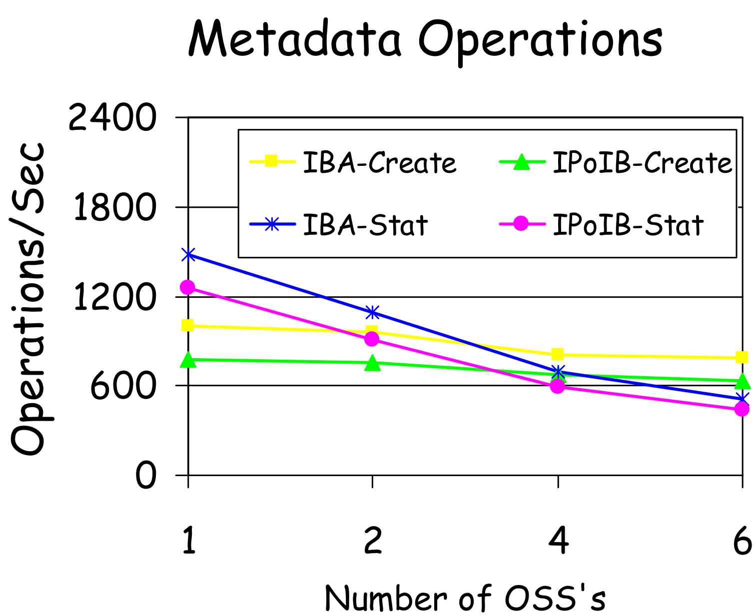
J. Wu, P. Wyckoff, and D. K. Panda, PVFS over InfiniBand: Design and Performance Evaluation, Int'l Conference on Parallel Processing (ICPP), Oct 2003.

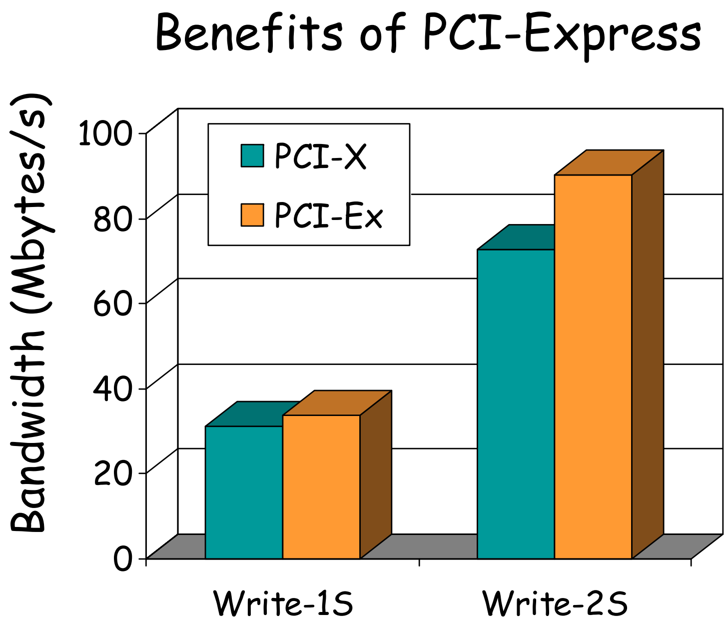
J. Wu, P. Wyckoff, and D. K. Panda, Supporting Efficient Noncontiguous Access in PVFS over InfiniBand, Cluster Computing Conference, Dec. 2003.

W. Yu, S. Liang and D. K. Panda, High Performance Support of PVFS2 over Quadrics. The 19th ACM International Conference on Supercomputing (ICS '05), June 2005

W. Yu and D. K. Panda, Benefits of Quadrics Scatter/Gather to PVFS2 Noncontiguous I/O, International Workshop on Storage Network Architecture and Parallel I/Os (SNAPI) 2005.

Lustre Performance (VAPI)

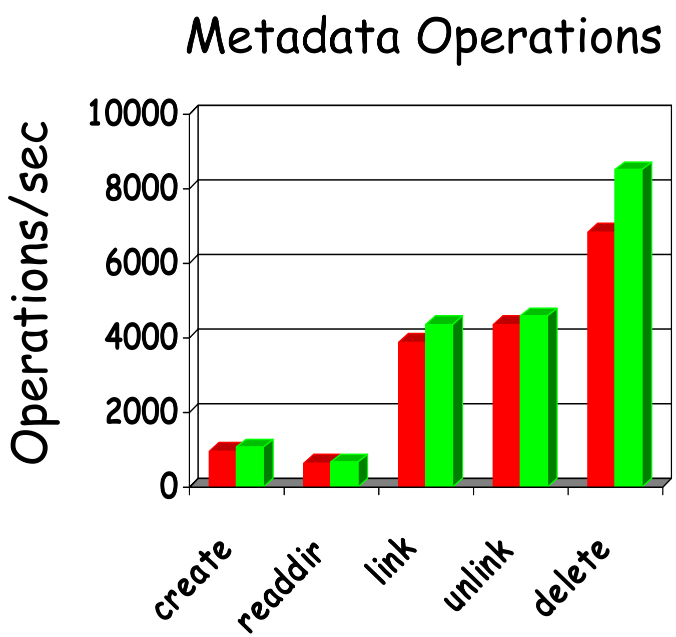


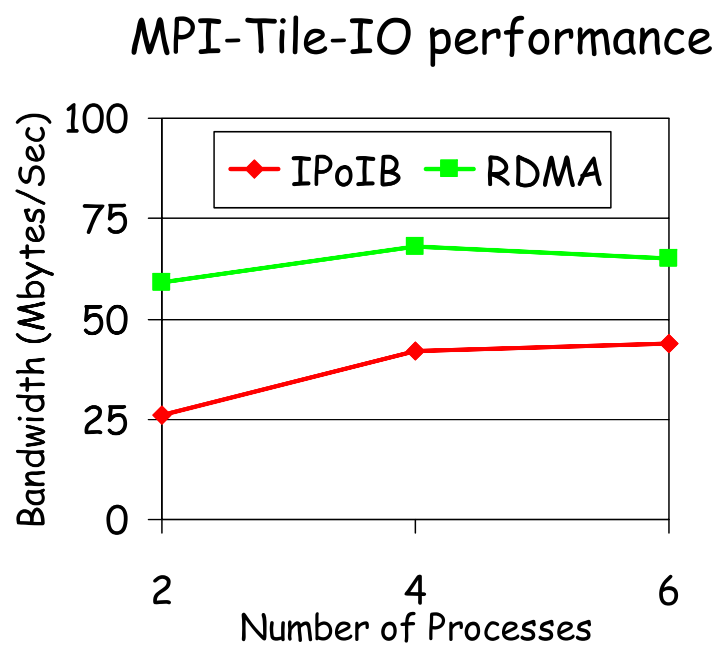


- Compared to IPoIB, IBA can improve the performance of Lustre meta-data operations, which does not scale with an increasing number of OSSs
- PCI-Express improves Lustre write bandwidth by 25% for 2 OSSs

W. Yu, R. Noronha, S. Liang and D. K. Panda, ***Benefits of High Speed Interconnects to Cluster File Systems: A Case Study with Lustre, To be presented at CAC 2006.***

Efficient NFS over RDMA for Solaris

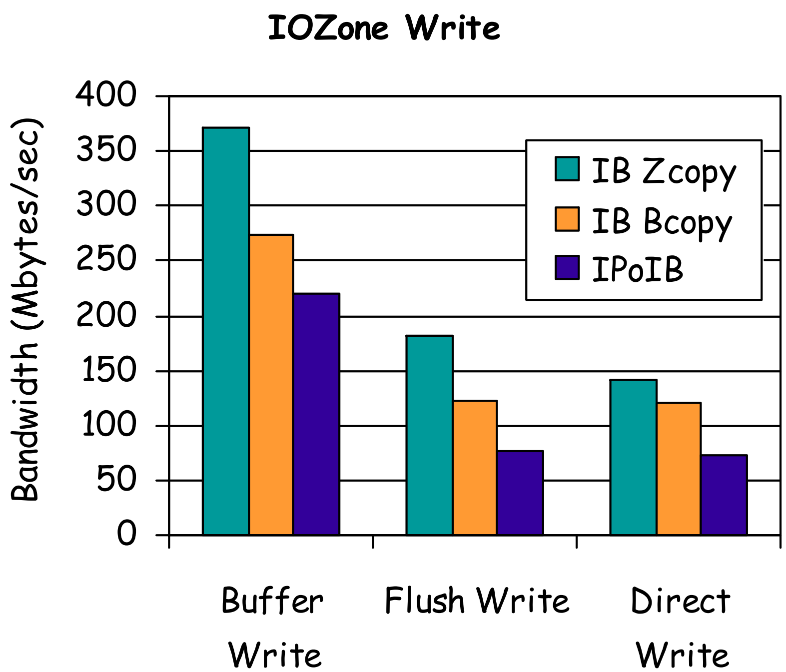


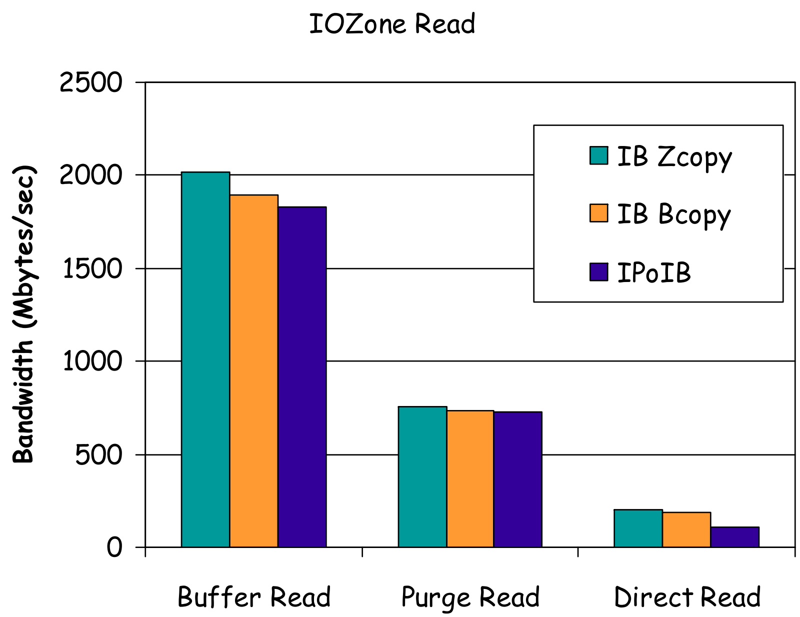


- RDMA is beneficial for operations which require communication
- Noncontiguous write bandwidth improved by 126%

Joint Project with SUN and NetApp

Global File System (Red Hat GFS) over OpenIB/Gen2





- Data copy overhead is significant for block I/O protocol, zero copy RDMA implementation improves performance up to 47% compared with copy based scheme and 136% compared with IPoIB

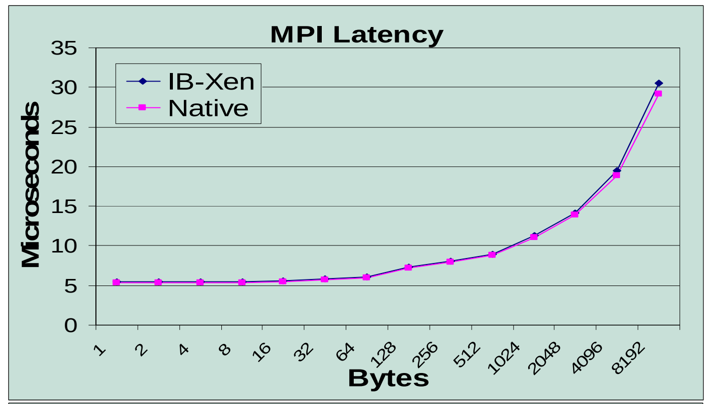
Xen-IB: Virtualizing InfiniBand in Xen

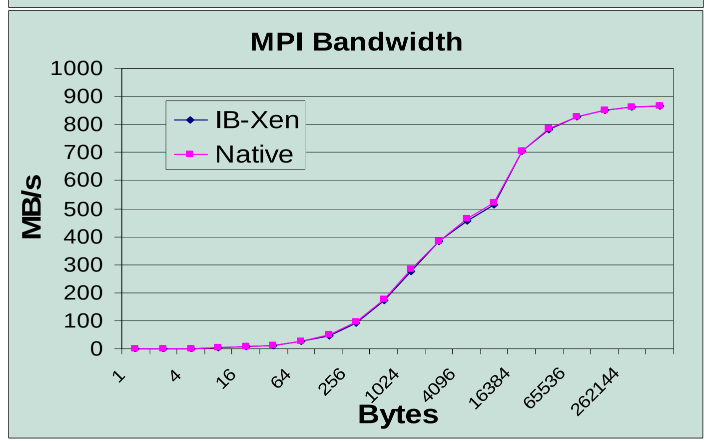
Design Overview:

- Follows Xen split driver model
  - Involving backend module for privileged operations
  - Bypassing dom0 (and VMM) for time critical operations
- Para-virtualization: Present virtual HCAs to guest domains
- Same IB-Gen2 Verbs Interface for applications in guest domains (domU)

Implementation:

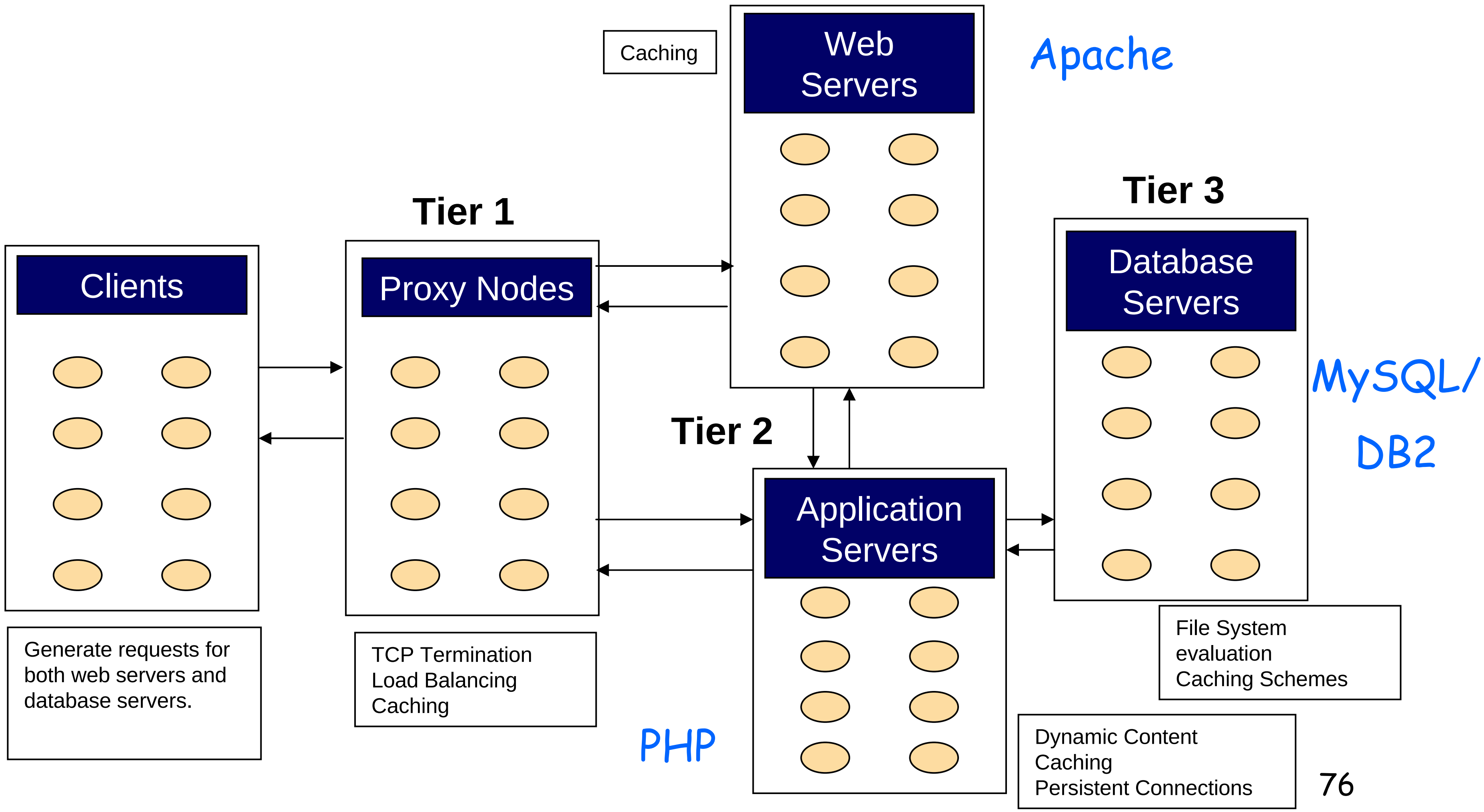
- Prototype based on Gen2 stack
- Close to native performance



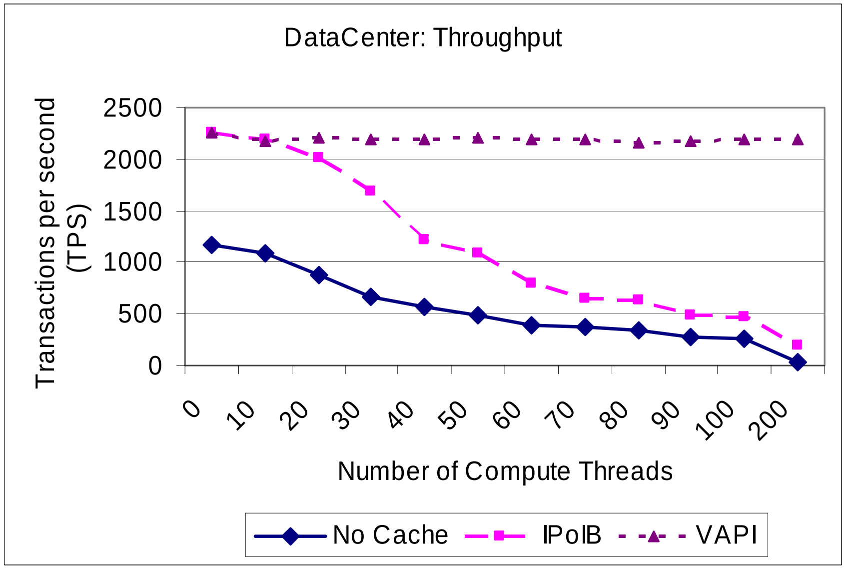


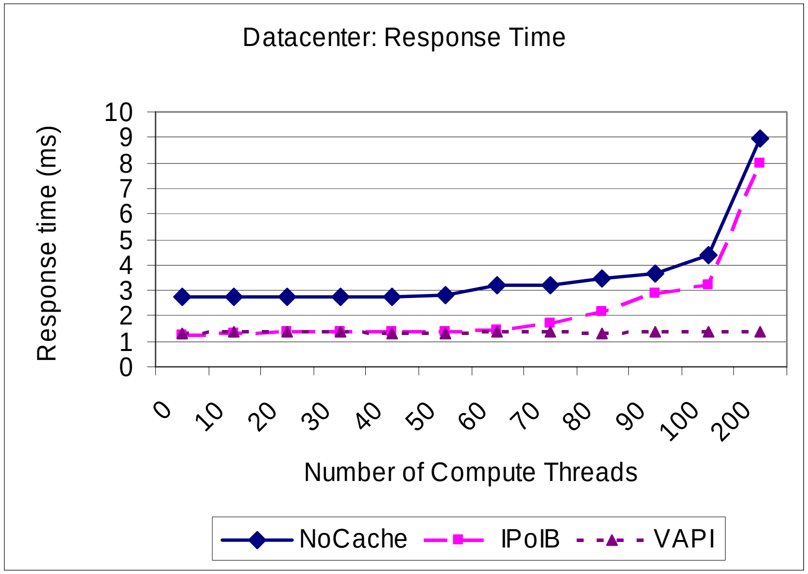
"Virtualizing InfiniBand in Xen – Prototype Design, Implementation and Performance". Presented at Xen Summit 2006

3-Tier Datacenter Testbed at OSU



Strong Cache Coherency with RDMA Polling: Datacenter Performance

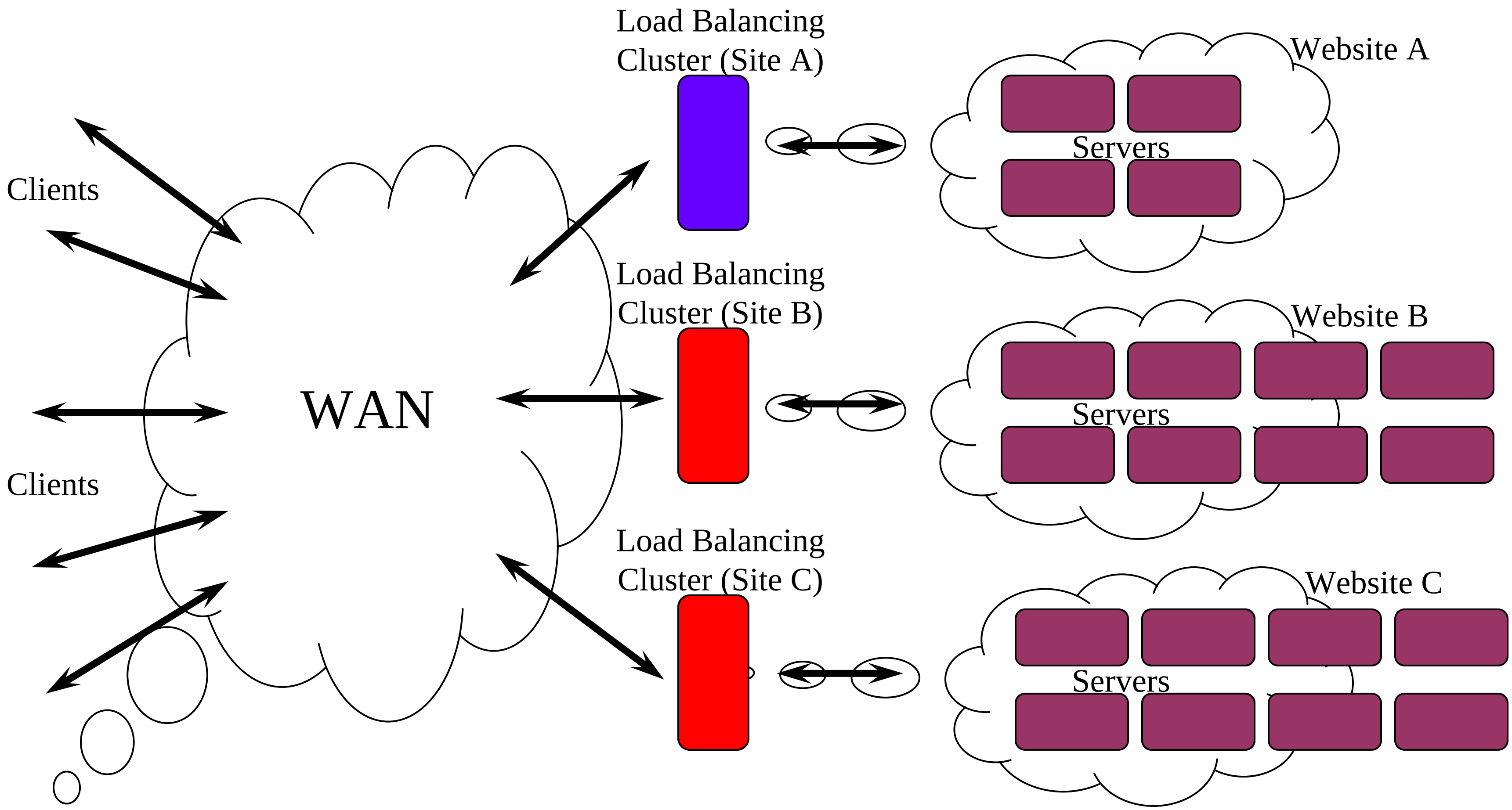




The VAPI module can sustain performance even with heavy load on the back-end servers

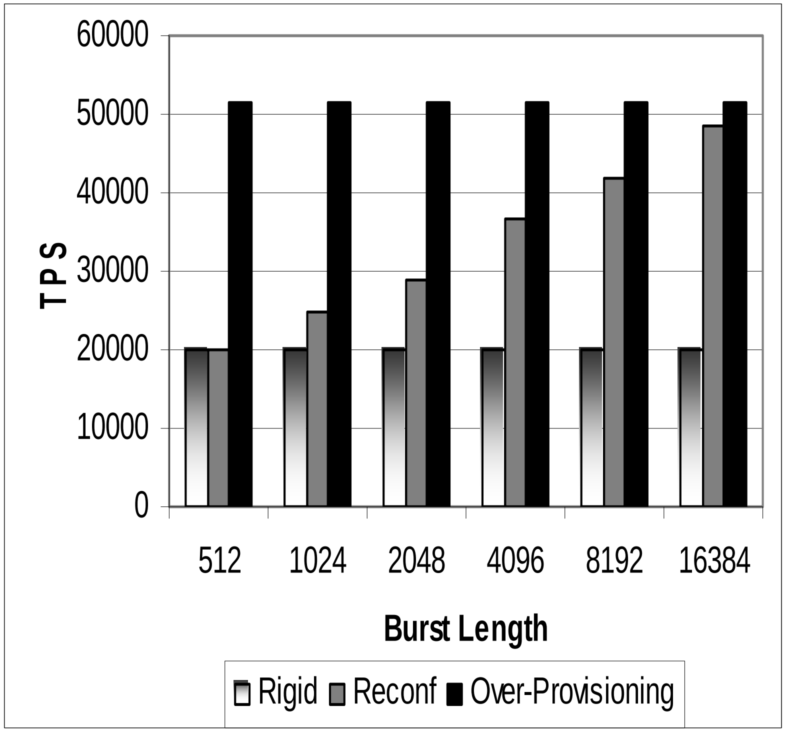
S. Narravul, P. Balaji, K. Vaidyanathan, S. Krishnamoorthy, J. Wu, and D. K. Panda, Supporting Strong Cache Coherency for Active Caches in Multi-Tier Data-Centers over InfiniBand, SAN'04, Feb 2004

Dynamic Reconfigurability in Shared Multi-tier Data-Centers

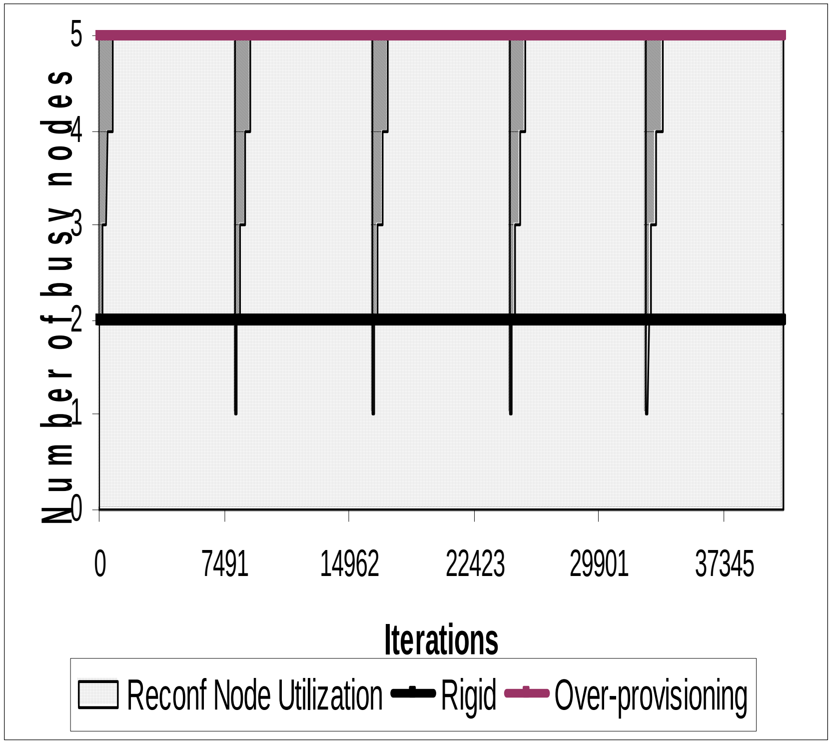


Nodes reconfigure themselves to highly loaded websites at run-time

Dynamic Re-configurability with Shared State using RDMA Operations



Performance of dynamic reconfiguration scheme largely depends on the burst length of requests



For large burst of requests, dynamic reconfiguration scheme utilizes all idle nodes in the system

P. Balaji, S. Narravula, K. Vaidyanathan, S. Narravula, H. -W. Jin, K. Savitha and D. K. Panda, Exploiting Remote Memory Operations to Design Efficient Reconfigurations for Shared Data-Centers over InfiniBand, RAIT '04

Conclusions

- MVAPICH and MVAPICH2 are being widely used in stable production IB clusters delivering best performance
- The user base stands at more than 310 organizations in 32 countries and is steadily growing
- Available with software stack distributions of many vendors
- Also available at the OpenIB/SVN
- New features for scalability, high performance and fault tolerance support are aimed to deploy large-scale IB clusters (20,000-50,000) nodes in the near future
- Besides MPI, many other open research issues in extracting performance of IB and iWARP in enterprise environments
  - SDP, File systems, Virtualization, Datacenters
- OSU is taking a lead in designing and developing novel solutions for these environments and integrated solutions will be available soon

Acknowledgements

Our research is supported by the following organizations

- Current Funding support by

- Current Equipment support by

Acknowledgements

- Current Students
  - Pavan Balaji (PhD)
  - Sitha Bhagvat (MS)
  - Lei Chai (PhD)
  - Qi Gao (PhD)
  - Prachi Gupta (PhD)
  - Wei Huang (PhD)
  - Matthew Koop (PhD)
  - Shaung Liang (PhD)
  - Amith Mamidala (PhD)
  - Sundeep Narravula (PhD)
  - Ranjit Noronha (PhD)
  - G. Santhanaraman (PhD)
  - Sayantan Sur (PhD)
  - K. Vaidyanathan (PhD)
  - Abhinav Vishnu (PhD)
  - Weikuan Yu (PhD)
- Current Post-Doc
  - Hyun-Wook Jin
- Current Programmer
  - Shaun Rowland
- Past Students
  - D. Buntinas (PhD)
  - B. Chandrasekharan (MS)
  - Weihang Jiang (MS)
  - Sushmita Kini (MS)
  - S. Krishnamoorthy (MS)
  - Jiuxing Liu (PhD)
  - Jiesheng Wu (PhD)

Web Pointers

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

http://nowlab.cse.ohio-state.edu/

MVAPICH Web Page

http://nowlab.cse.ohio-state.edu/projects/mpi-iba/

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