Datacenter Fabric Workshop

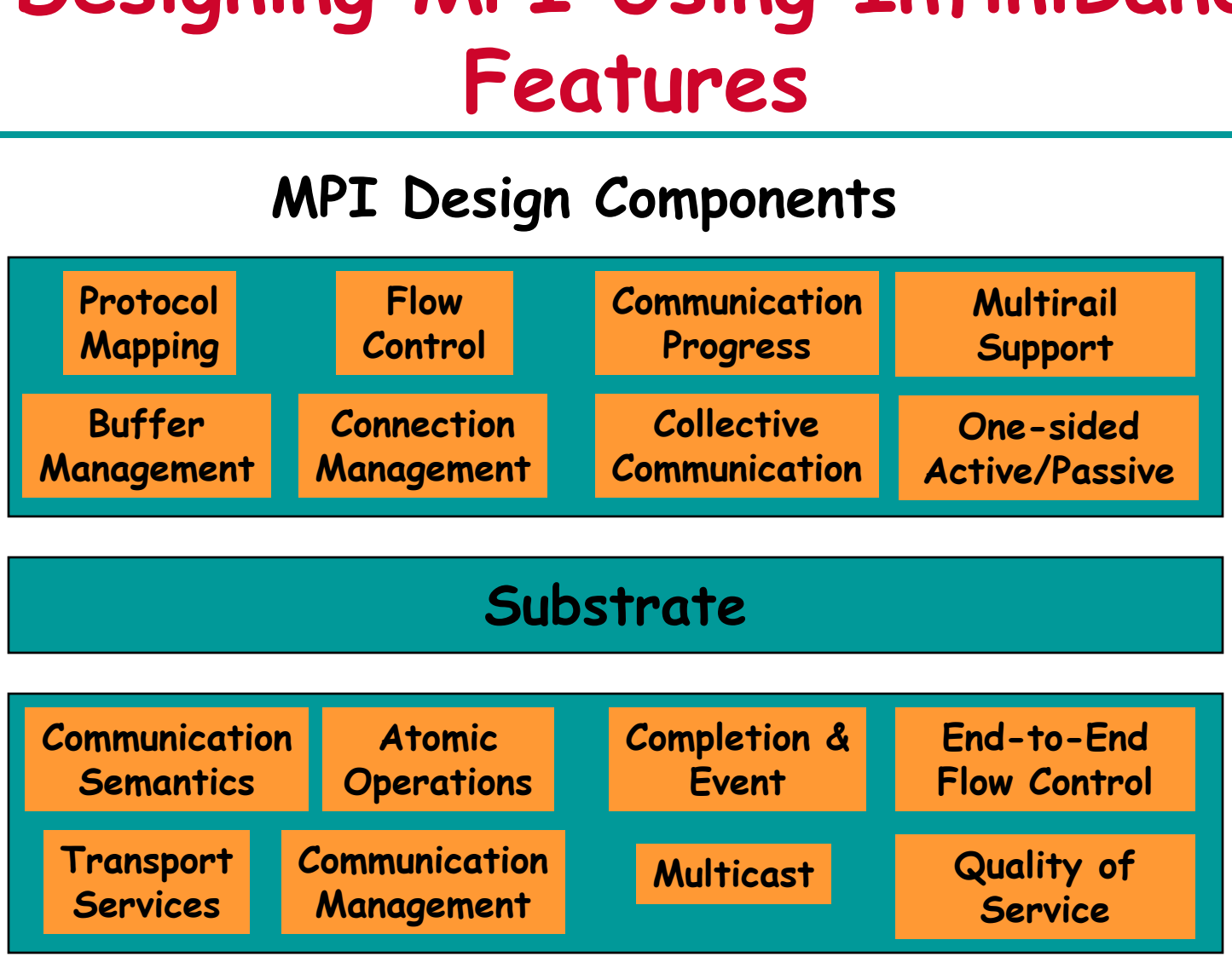
OSU MPI over InfiniBand (MVAPICH): Latest Status, Performance Numbers and Future Plans

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

- Overview of MVAPICH and MVAPICH2 Projects
  - Current Status
  - Features
- MVAPICH-Gen2 1.0
  - Features and Performance numbers
  - Comparison with other interconnects
- Upcoming New MVAPICH2 Design
  - Features and Sample Performance number
- Future Plans
  - Upcoming MVAPICH and MVAPICH2 releases
  - Scalability/Reduced memory usage
  - Fault Tolerance
- Conclusions

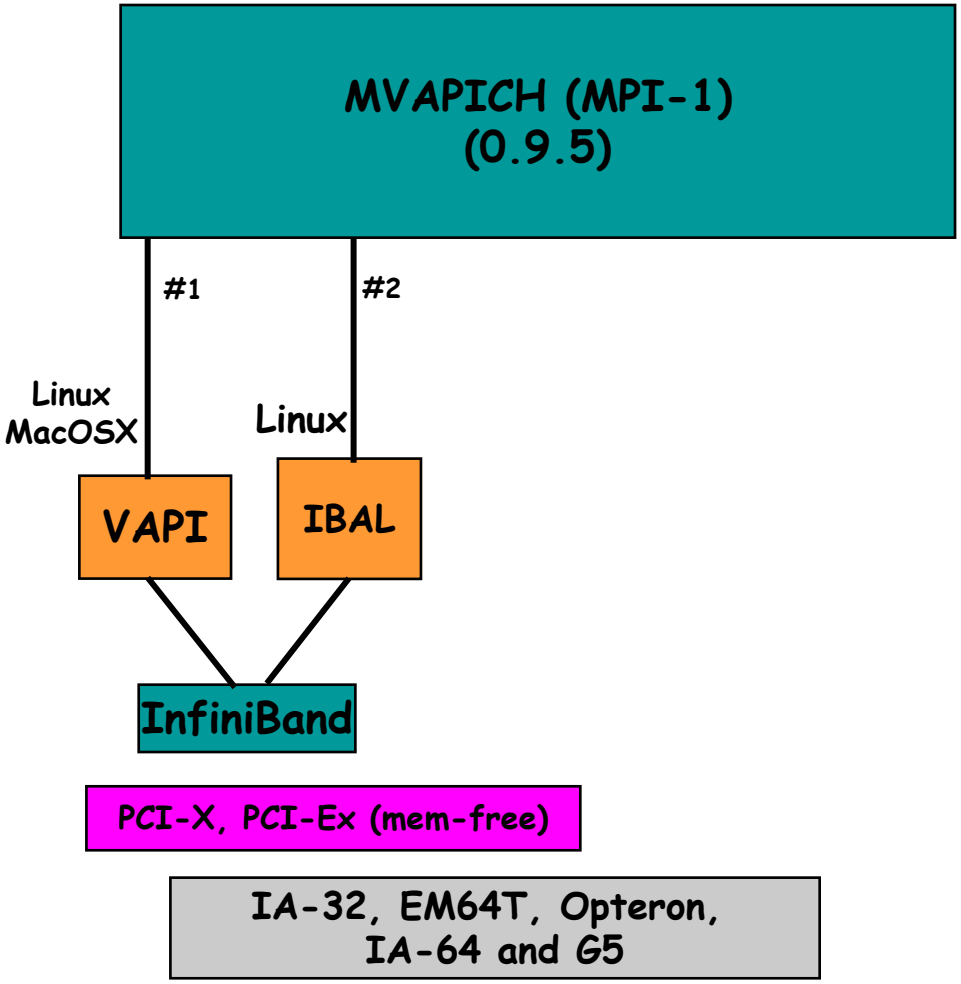
Designing MPI Using InfiniBand Features



High Performance MPI Implementation with IBA – Research Agenda at OSU

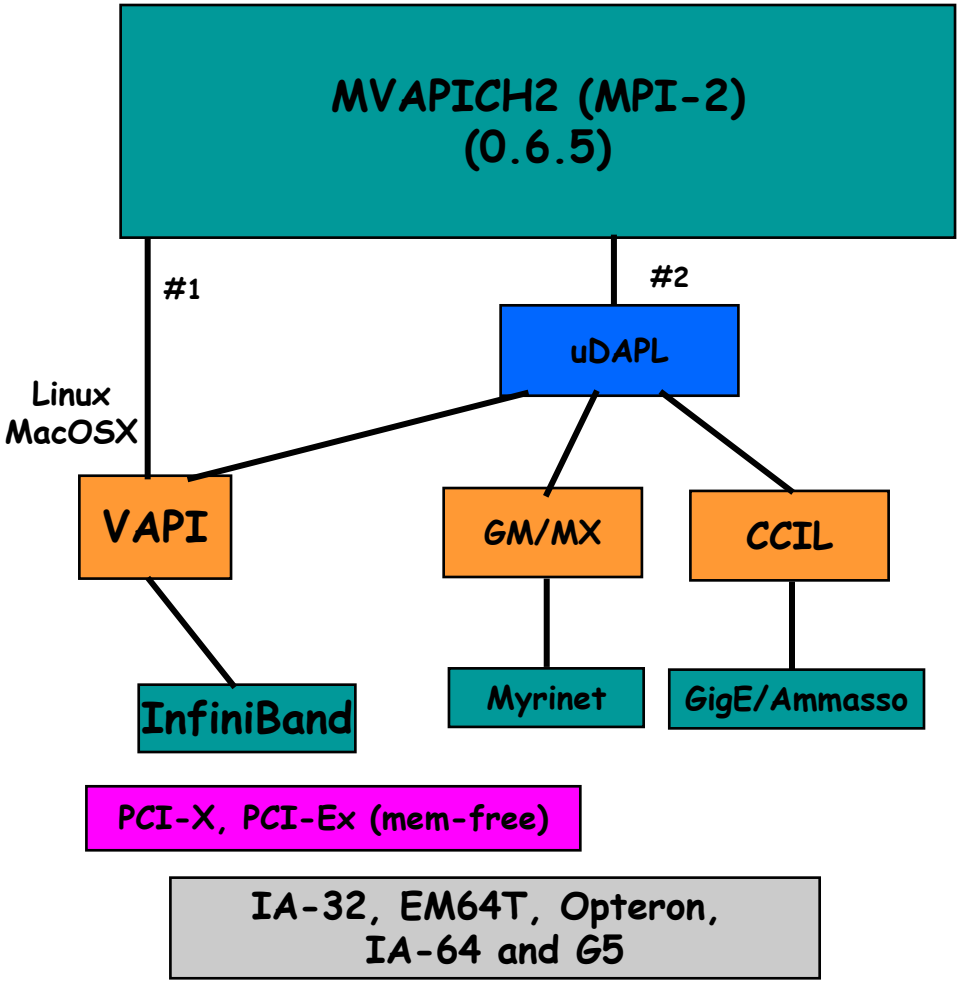
- Point-to-point communication
  - RDMA-based design for both small and large messages
- Collective communication
  - Taking advantage of IBA hardware multicast for Broadcast
  - RDMA-based designs for barrier, all-to-all
- Flow control
  - Static vs. dynamic
- Connection Management
  - Static vs. dynamic
- Multi-rail designs
  - Multiple ports/HCAs
  - Different schemes (striping, binding, adaptive)
- MPI Datatype Communication
  - Taking advantage of scatter/gather semantics of IBA
- MPI-2 One-sided communication and synchronization

MVAPICH Project: Past Developments and Current Status



- RDMA-based point-point and collectives
- Multi-rail support
  - Multiple ports/adapters
  - Multiple adapters
  - Multiple paths with LMC
- Enabling Broadcast support with IBA multicast for larger systems
- Optimized shared memory support
  - Bus-based architecture
  - NUMA architectures
- Optimized for scalability
  - Three different modes: small, medium, and large clusters
- Totalview Debugger (Etnus) support
- MPD Support

MVAPICH2 Project: Past Developments and Current Status



- RDMA channel designs
  - two-sided operations
  - one-sided operations
- Optimized one-sided operations
  - Get
  - Put
  - Accumulate
- Active target synchronization
- Optimized for scalability
  - Three different modes: small, medium, and large clusters
- MPD Support
- Portability across multiple interconnects

MVAPICH/MVAPICH2 Software Distribution

- Open Source (current versions are MVAPICH 0.9.5 and MVAPICH2 0.6.5)
- Have been directly downloaded by more than 250 organizations and industry (across 28 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)
CASPUR, Interuniversity Consortium (Italy)
Cornell Theory Center
C-DAC, Center for Development of Advanced Computing (India)
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)
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)
IFP, French National Oil and Gas Res. Center (France)
Inst. for Experimental Physics (Germany)
Inst. for Program Structures and Data Org. (Germany)
Inst. of Physics, Chinese Academy of Sciences (China)
Inst. "Rudjer Boskovic" (Croatia)
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)
Max Planck Institute for Plasma Physics (Germany)
NASA Ames Research Center
NCSA
National Center for High Performance Computing (Taiwan)
National Center for Atmospheric Research
National Supercomputer Center in Linkoping (Sweden)
Ohio Supercomputer Center
Open Computing Centre "Strela" (Russia)
Pacific Northwest National Laboratory
Pittsburgh Supercomputing Center
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
United Institute of Informatics Problems (Belarus)
U.S. Census Bureau
U.S. Geological Survey
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)
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
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)
Tsinghua Univ. (China)
Univ. of Arizona
Univ. of Berne (Switzerland)
Univ. of Bielefeld (Germany)
Univ. of California, Berkeley
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 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 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)
Virginia Tech
Wroclaw Univ. of Technology (Poland)

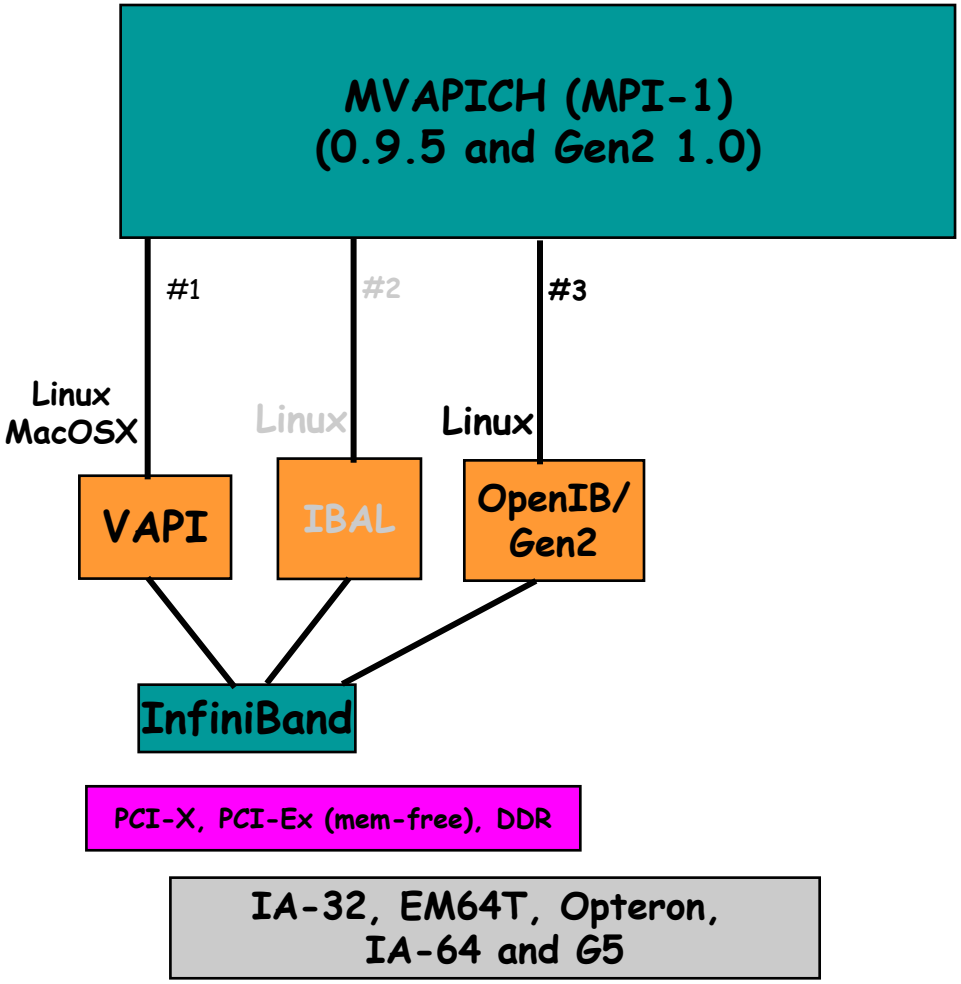
MVAPICH/MVAPICH2 Users: Industry

Abba Technology
Advanced Clustering Tech.
Agilent Technologies
AMD
Ammasso
Annapolis Micro Systems, Inc.
Apple Computer
Appro
Array Systems Comp. (Canada)
Ascender Technologies Ltd (Israel)
Ascensit (Italy)
Atipa Technologies
AWE PLC (UK)
BAE Systems
Barco Medical Imaging Systems
Best Systems Inc. (Japan)
Bluware
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.
Dynamics Technology, Inc.
Easy Mac (France)
Emplics (Germany)
ESI Group (France)
Exadron (Italy)
ExaNet (Israel)
Fluent Inc.
Fluent Inc. (Europe)
FMS-Computer and Komm. (Germany)
General Atomics
GraphStream, Inc
Gray Rock Professional
HP
HP (Asia Pacific)
HP (France)
HP Solution Center (China)
High Performance Associates
IBM
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)
Linux Networx
Linvision (Netherlands)
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)
PANTA Systems
ParTec (Germany)
PathScale, Inc.
Pultec (Japan)
Pyramid Computer (Germany)
Qlusters (Israel)
Quadrics (UK)
Quant-X GmbH (Austria)
Rackable Systems, Inc.
Raytheon Inc.
Remcom Inc.
RJ mears, LLC
RLX Technologies
Rosta Ltd. (Russia)
SBC Technologies, Inc.
Scyld Software
Scalable Informatics LLC
Scotland Electronics (Int'l) Ltd (UK)
Scotland Electronics Int'l Lrd. (UK)
SGI (Silicon Graphics, Inc.)
Siliquent
Silverstorm technologies
Simulation Technologies
SKY Computers
SmallTree communications
STMicroelectronics
Streamline Computing (UK)
SUN
Systran
Texh-X Corp.
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)
Verari Systems Software
Virtual Iron Software, Inc.
Voltaire
Western Scientific
WorkstationsUK, Ltd. (UK)
Woven Systems, Inc.

Presentation Overview

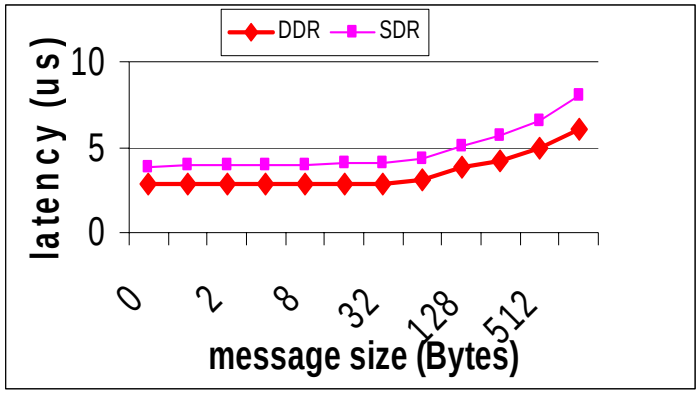
- Overview of MVAPICH and MVAPICH2 Projects
  - Current Status
  - Features
- MVAPICH-Gen2 1.0
  - Features and Performance numbers
  - Comparison with other interconnects
- Upcoming New MVAPICH2 Design
  - Features and Sample Performance number
- Future Plans
  - Upcoming MVAPICH and MVAPICH2 releases
  - Scalability/Reduced memory usage
  - Fault Tolerance
- Conclusions

MVAPICH-Gen2 1.0 Release

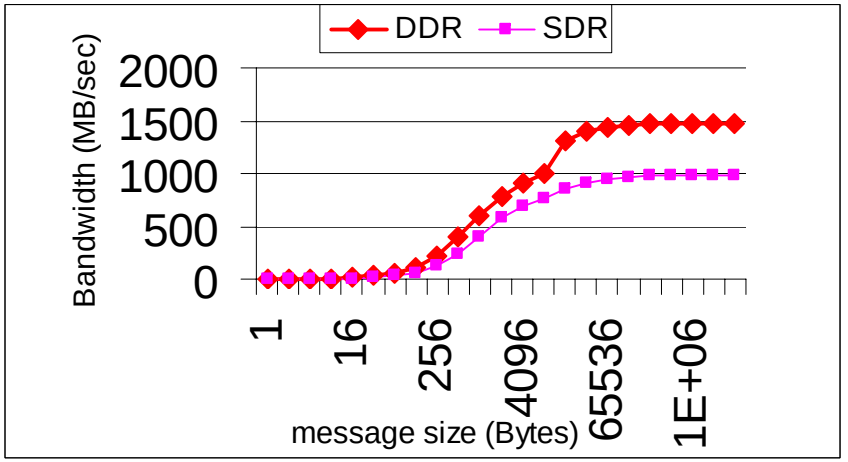


- Released on 08/21/05
- RDMA-based optimized designs for
  - point-to-point communication
- Collectives based on point-to-point
- Optimized shared memory support
  - Bus-based architecture
  - NUMA architectures
- Shared library support
- Additional features will be added in successive releases

MVAPICH-Gen2 with InfiniBand 4X SDR and DDR: MPI-Level Performance

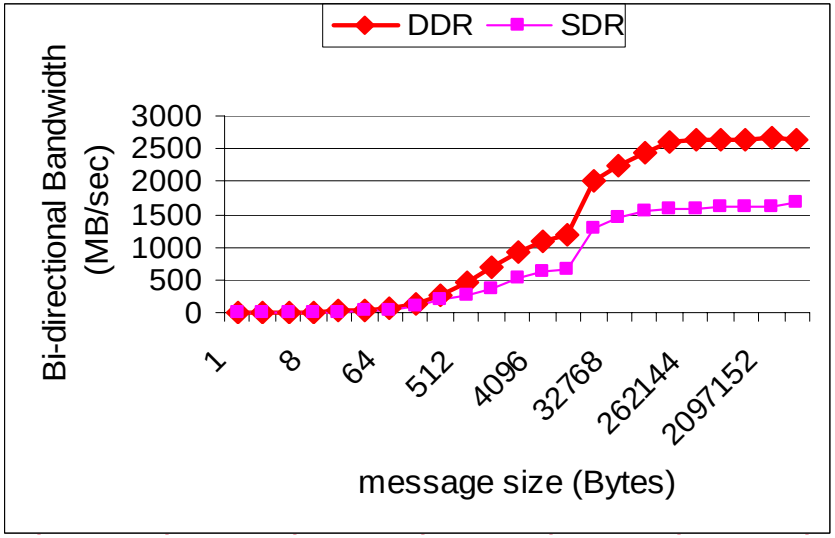


3.85
2.84



1458

981



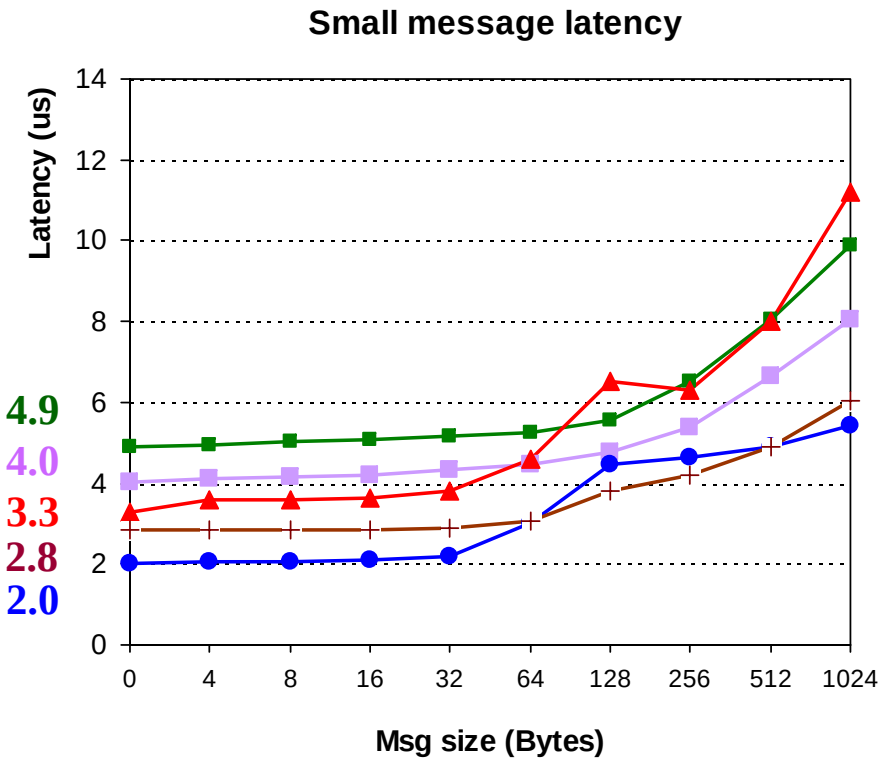
2646

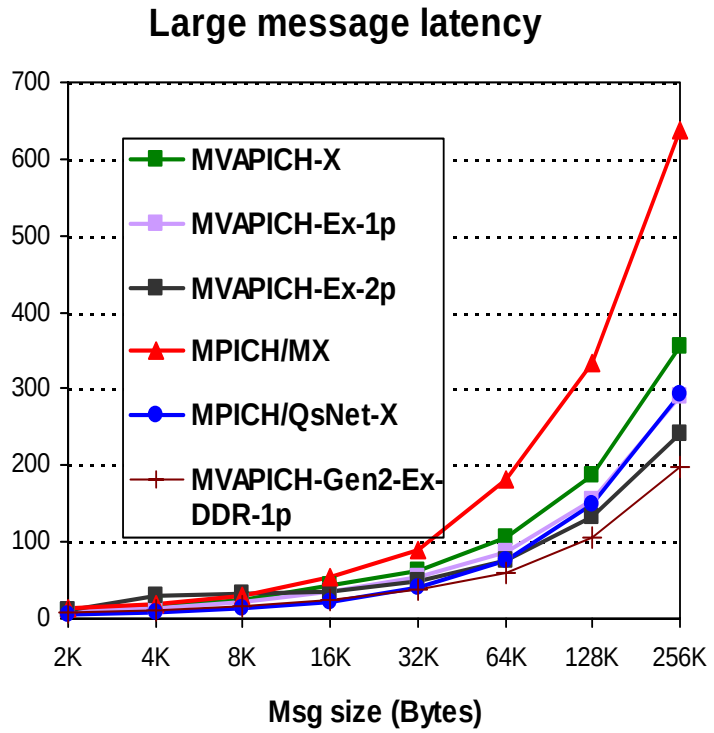
1670

- Single port results only

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

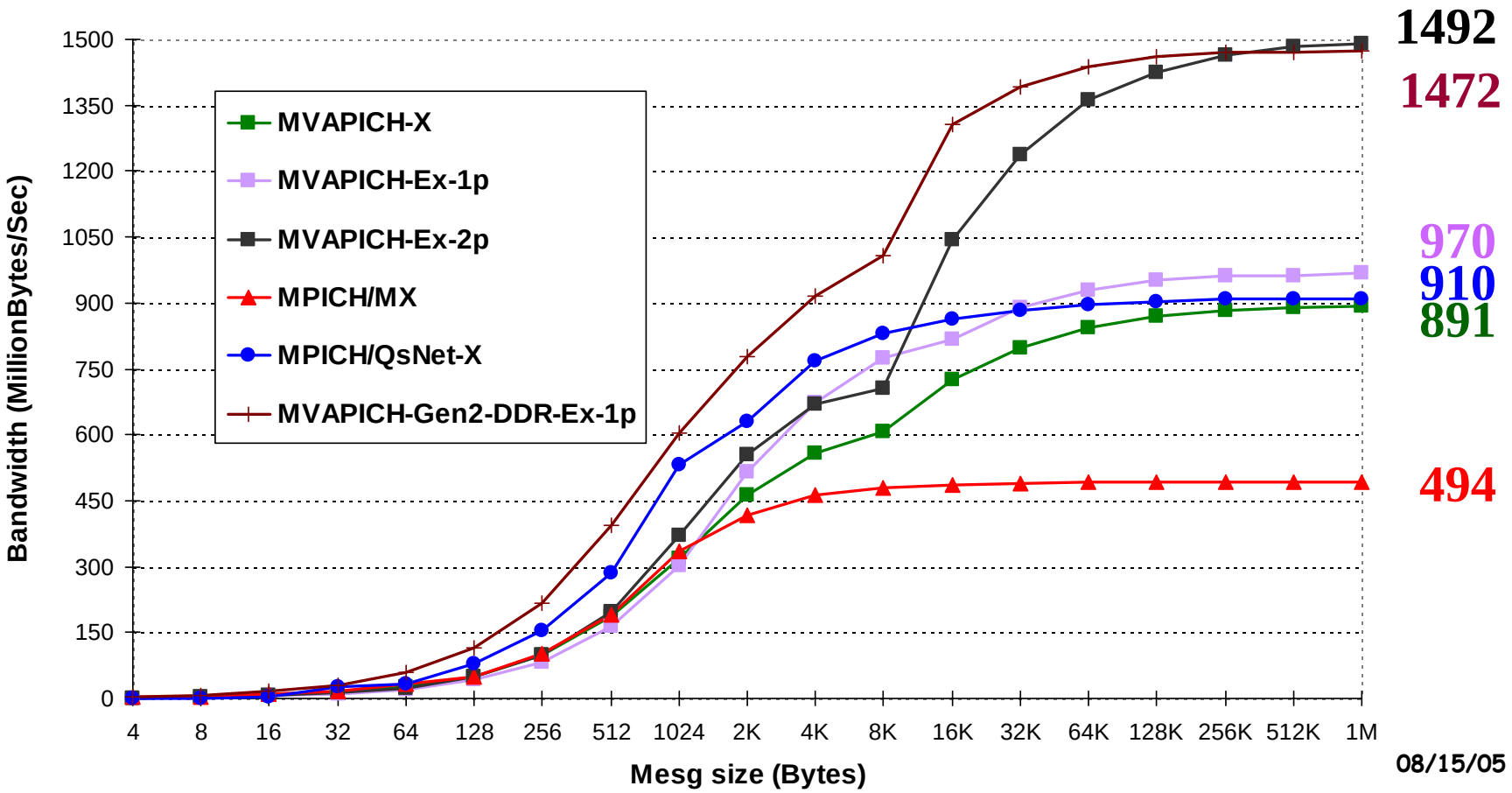
MPI-level Latency (One-way):
IBA vs. Myrinet vs. Quadrics



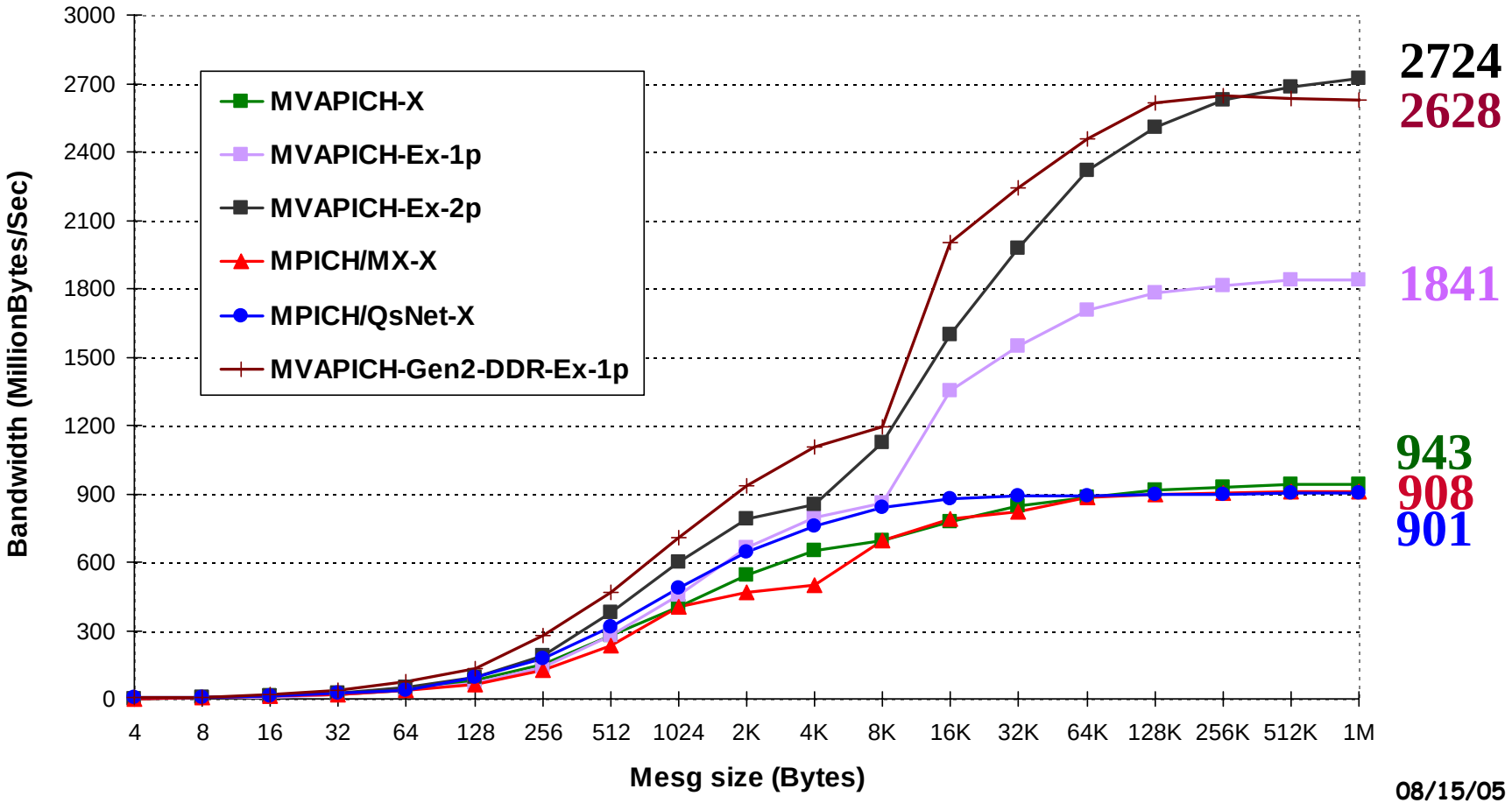


- SC '03
- Hot Interconnect '04
- IEEE Micro (Jan-Feb) '05, one of the best papers from HotI '04

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



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

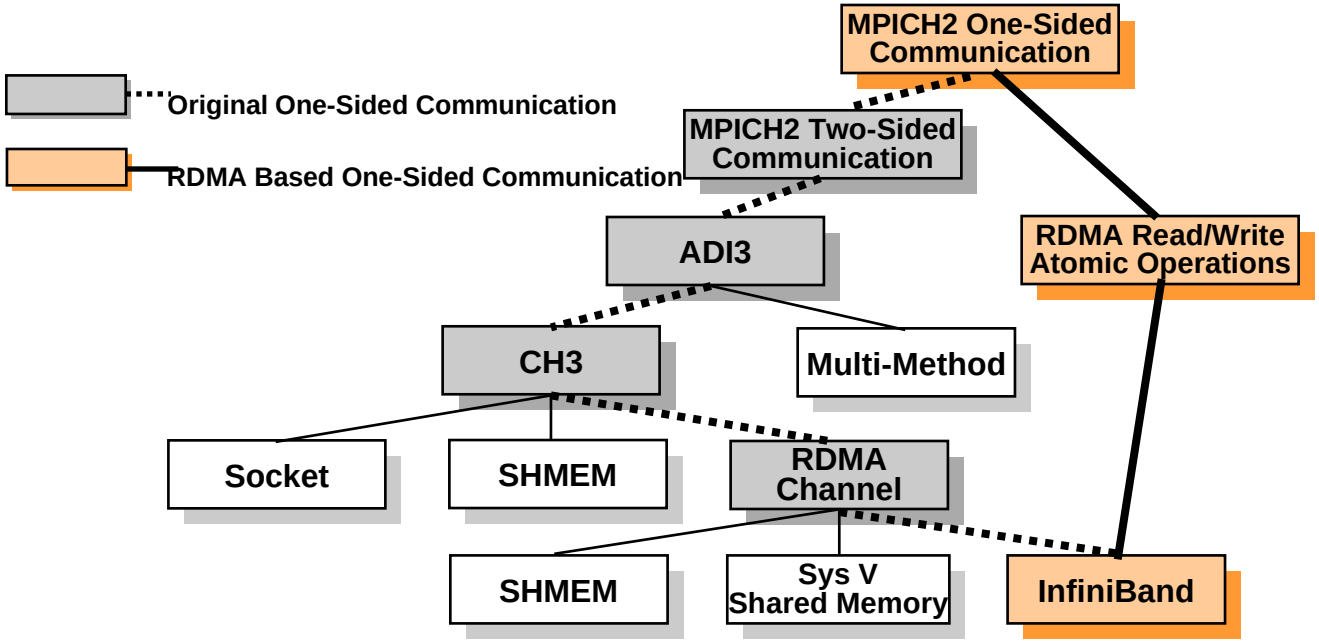


08/15/05

Presentation Overview

- Overview of MVAPICH and MVAPICH2 Projects
  - Current Status
  - Features
- MVAPICH-Gen2 1.0
  - Features and Performance numbers
  - Comparison with other interconnects
- Upcoming New MVAPICH2 Design
  - Features and Sample Performance number
- Future Plans
  - Upcoming MVAPICH and MVAPICH2 releases
  - Scalability/Reduced memory usage
  - Fault Tolerance
- Conclusions

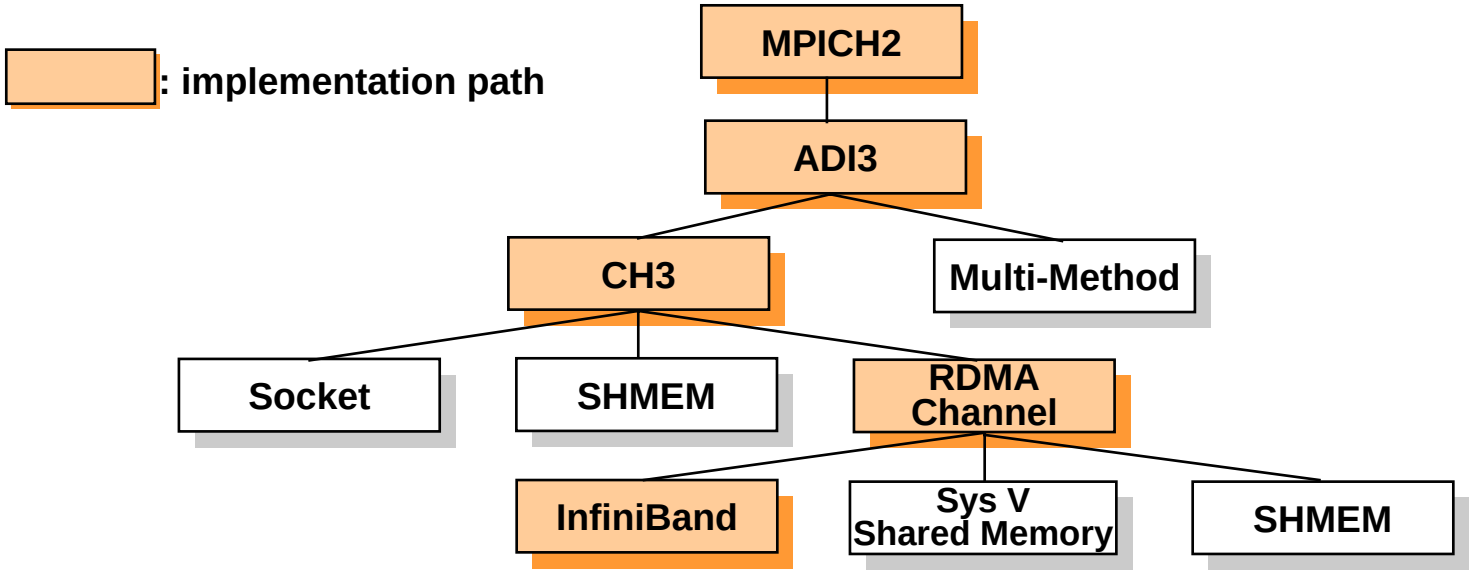
Two-sided and One-Sided Communication Implementation Structure in MVAPICH2



• 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 Int'l Parallel and Distributed Processing Symposium (IPDPS), April 2004

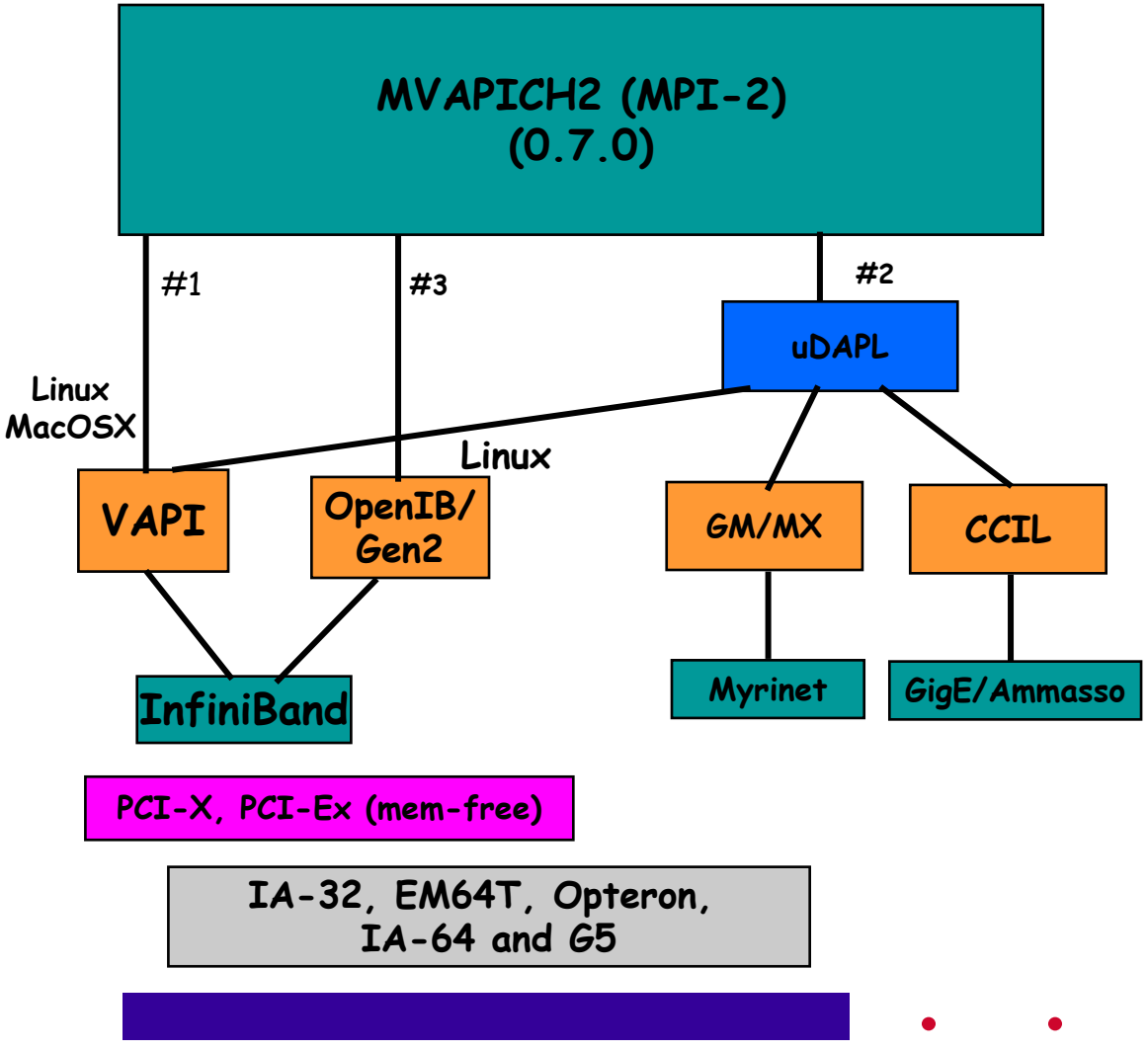
• W. Jiang, J. Liu, H. Jin, D.K. Panda , W. Gropp and R. Thakur, High Performance MPI-2 One-Sided Communication over InfiniBand, CCGrid, May 2004

New MVAPICH2 Design



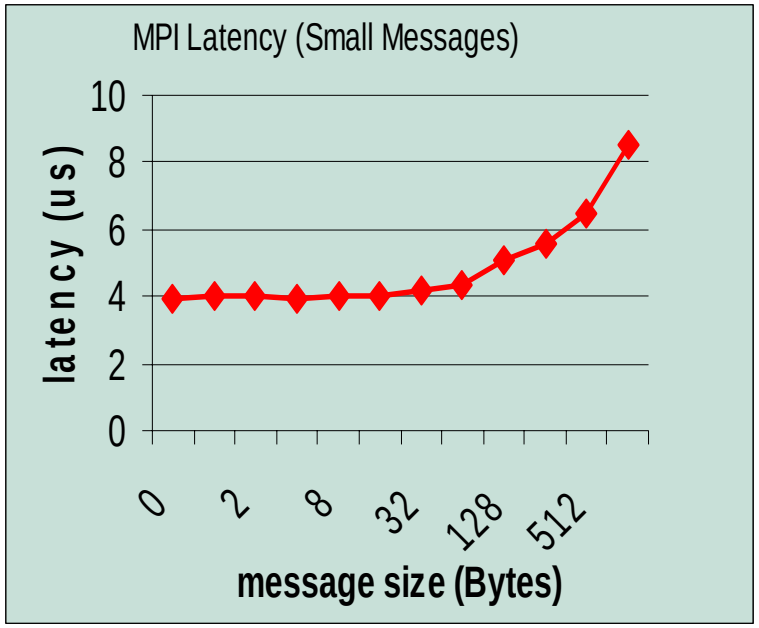
- RDMA channel provides limitations
  - communication overhead is higher
  - shared memory support is not there
  - multicast can not be used
- Moving our designs (two-sided and one-sided) to ADI3 layer and unify it with MVAPICH
- MVAPICH2 will have all benefits and performance as that of MVAPICH + One-sided

New MVAPICH2 Design

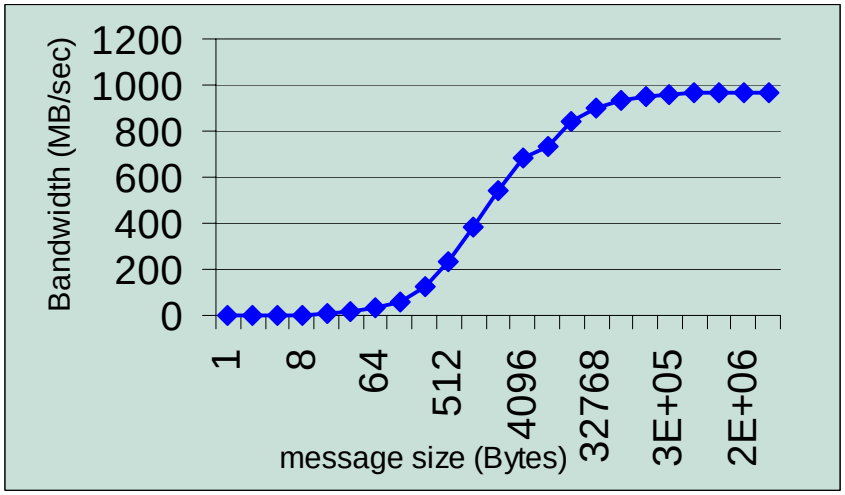


- ADI3 level design
  - two-sided operations
  - one-sided operations
- Optimized one-sided operations
  - Get
  - Put
  - Accumulate
- Shared memory support for Bus and NUMA-based systems
- Multicast support and RDMA-based collectives
- Optimized for scalability
  - Three different modes: small, medium, and large clusters
- MPD Support
- Totalview Support
- Both VAPI and Gen2 support
- Portability across multiple interconnects
- All features and performance of MVAPICH + One-sided and Portability

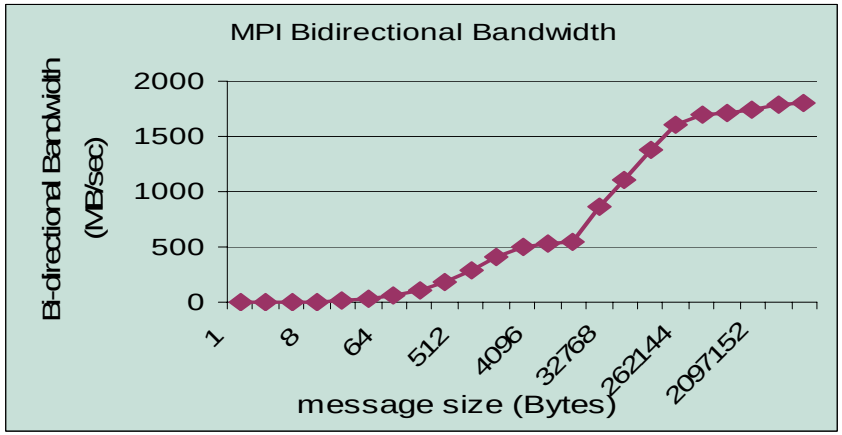
MVAPICH2-Gen2 with InfiniBand 4X SDR: MPI-Level Performance



3.91



968.2

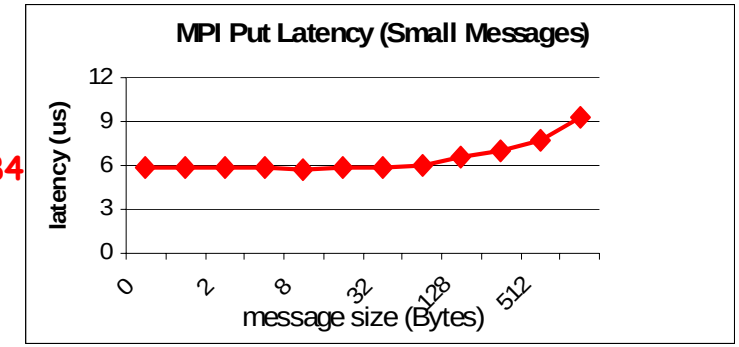


1801

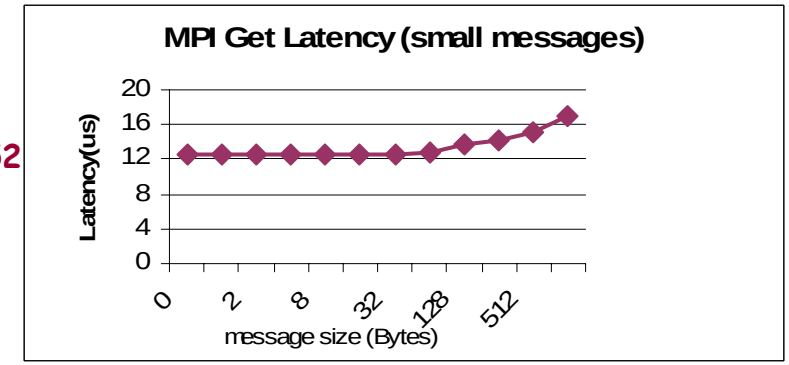
- Single port results only
- DDR results expected to be same as MVAPICH-Gen2 1.0

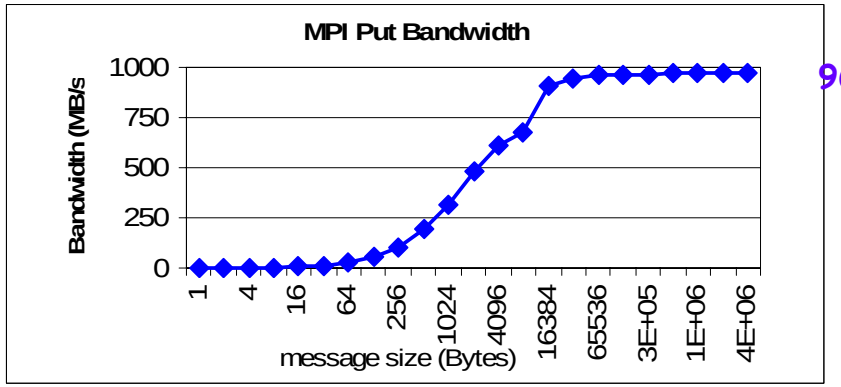
MVAPICH2-Gen2 with InfiniBand 4X SDR: MPI One Sided Performance

5.84



12.52

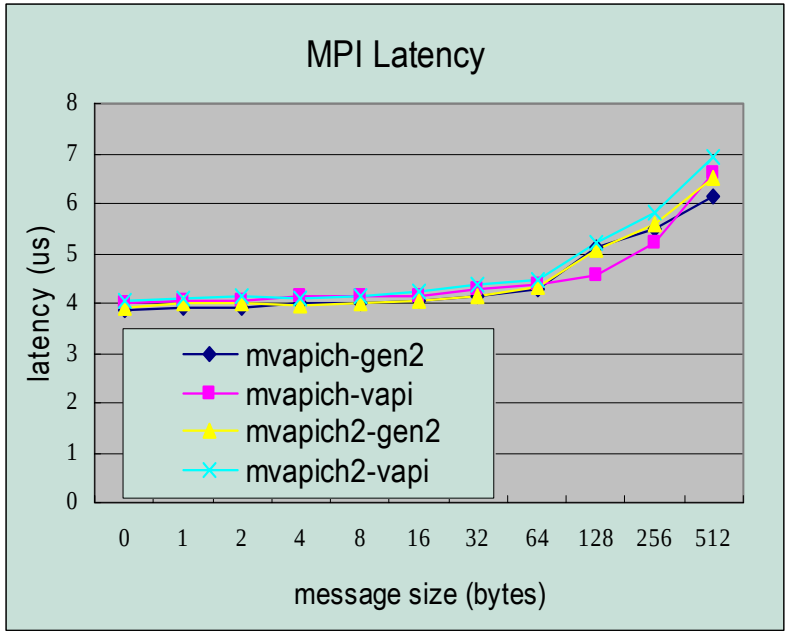


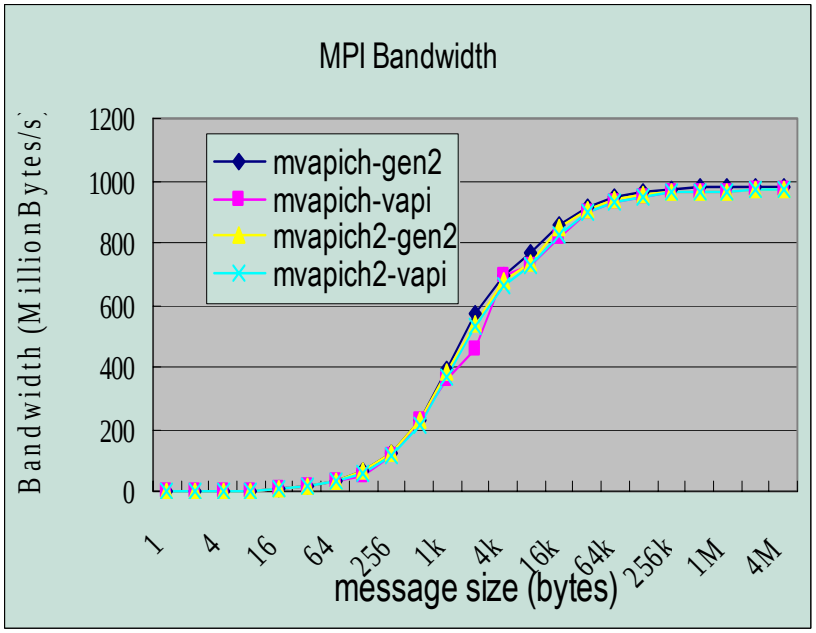


968

- Single port results only

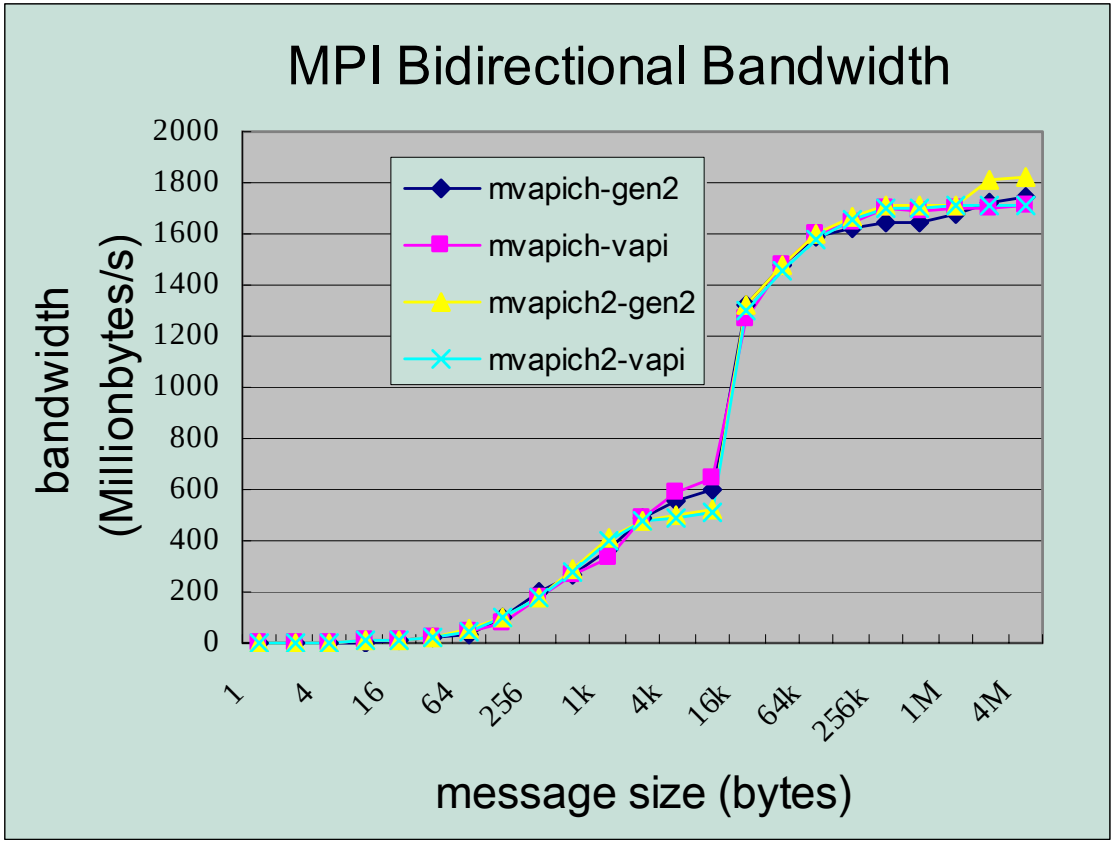
Performance Comparison of MVAPICH and MVAPICH2 (Two-sided Operations)





- Performance comparison for four versions of MVAPICH:
  - mvapich-0.9.5, mvapich-gen2 1.0;
  - mvapich2-0.7.0-vapi, mvapich2-0.7.0-gen2

Performance Comparison of MVAPICH and MVAPICH2 (Two-sided Operations)



MVAPICH2 new design

Will be available soon

Presentation Overview

- Overview of MVAPICH and MVAPICH2 Projects
  - Current Status
  - Features
- MVAPICH-Gen2 1.0
  - Features and Performance numbers
  - Comparison with other interconnects
- Upcoming New MVAPICH2 Design
  - Features and Sample Performance number
- Future Plans
  - Upcoming MVAPICH and MVAPICH2 releases
  - Scalability/Reduced memory usage
  - Fault Tolerance
- Conclusions

Upcoming MVAPICH and MVAPICH2 Releases

- MVAPICH 0.9.6 (in the next 2-3 months)
  - uDAPL support
    - Portability across many interconnects
  - Solaris support
    - Both uDAPL layer
    - Native IBTL layer
  - Additional Features
    - Blocking support
    - RDMA-Read based Rendezvous for better communication progress
    - Optimized registration cache (reduced memory requirement)
- Continued successive releases of MVAPICH Gen2
- MVAPICH2 0.7.0 (in the next 2-3 months)
  - With the new design
  - Gen2, uDAPL, and VAPI support
  - Solaris support

Scalability/Reduced Memory Usage

- Large-scale IB clusters (~4,000 nodes) are being deployed with MVAPICH
  - More are on the horizon
- MVAPICH project is gearing up to meet the requirements imposed by large clusters
- Working on several design changes to MVAPICH
  - Point-to-point Communication
    - Shared Recv Queue (SRQ)
    - Enhanced Flow Control
    - Advanced Buffer management strategy
    - Adaptive connection management
  - Collective Communication
    - Exploiting intra-node shared memory
    - Using hardware multicast for other collective operations
  - Resource Management Infrastructure
- Gil's presentation will show some initial numbers
- These solutions will be available in future releases of MVAPICH

Fault Tolerance

- Component failures are the norm in large-scale clusters
- Imposes need on reliability and fault tolerance
- Working along the following angles
  - Reliable Networking with APM utilizing Redundant Communication Paths
  - Process Fault Tolerance with Efficient Checkpoint and Restart
    - Exploiting RDMA for very low overhead checkpoints
  - End-to-end Reliability with memory-to-memory CRC

Conclusions

- Provided a brief overview of
  - Current status of MVAPICH and MVAPICH2 projects
  - Future Releases
  - Research Challenges
- Our RDMA-based MPI design is applicable to IBA as well as other emerging interconnects with RDMA support

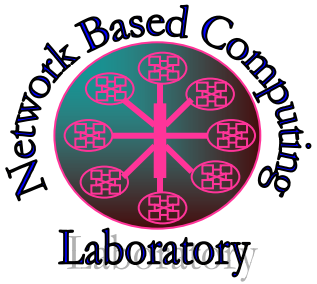
Acknowledgements

Our research is supported by the following organizations

- Current Funding support by

- Current Equipment donations by

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