Design Alternatives and Performance Trade-offs for Implementing MPI-2 over InfiniBand

W. Huang, G. Santhanaraman,

H. -W. Jin, and D. K. Panda

Network Based Computing Laboratory

The Ohio State University

{huanwei,santhana,jinhy,panda@cse.ohio-state.edu)

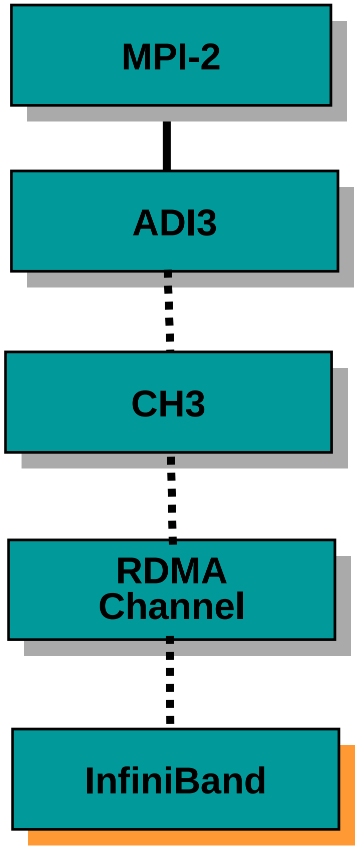
Outline

- Background
  - InfiniBand, MPICH2, MVAPICH2
- Motivation
- Design and Implementation
- Performance Evaluation
- Conclusion and Future work

Background-InfiniBand

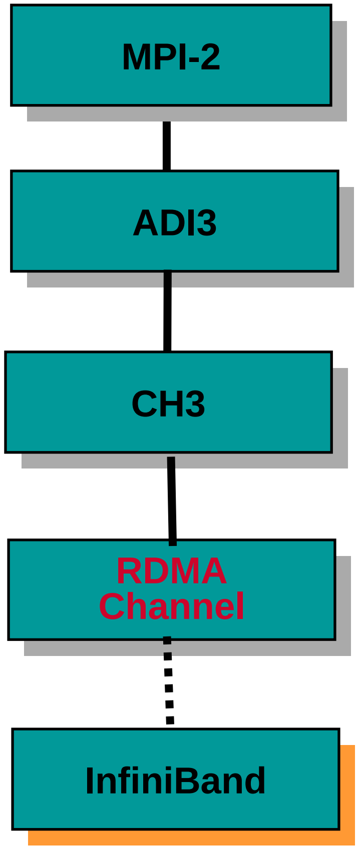
- The InfiniBand Architecture (IBA) is a new industry standard for high speed interconnect
- IBA supports channel semantics (send/recv) and RDMA semantics.
- User Level Verbs Interface:
  - VAPI: Mellanox implementation
  - Gen2: OpenIB implementation

Background-MPICH2



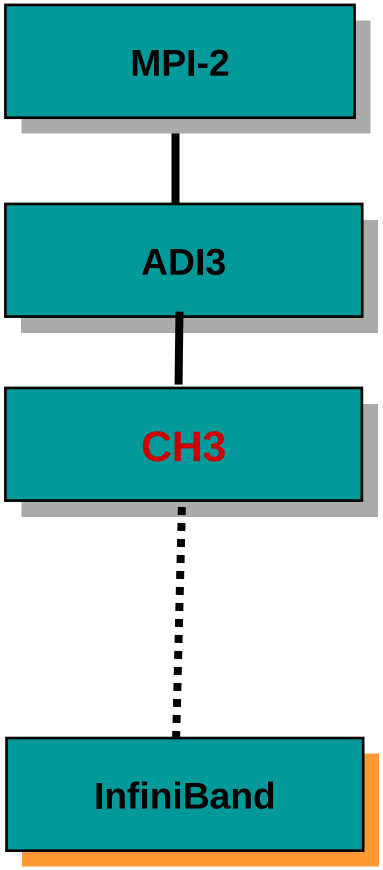
- Layered Design of MPICH2 leaves three choices for MPI2 over IBA:
  - RDMA
  - CH3
  - ADI3

RDMA Channel



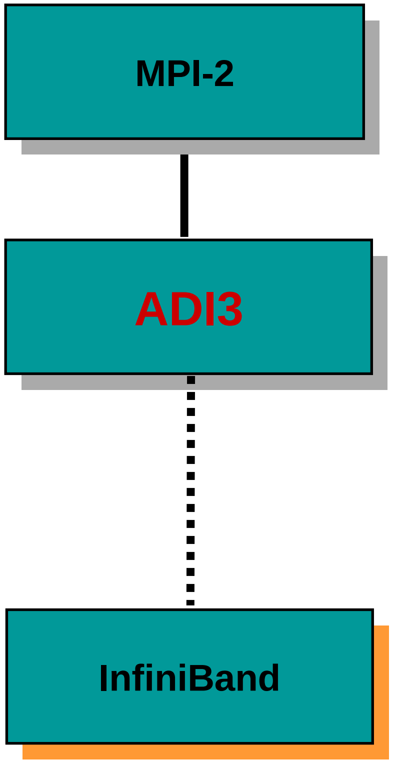
- Design at RDMA channel layer:
  - Bottom most position in the hierarchy, no need to implement progress engine
  - Simple: only 5 interfaces need to be implemented
  - Designed for RDMA capable network, a fit to IBA's RDMA semantics

CH3 layer



- A more complex channel device
  - Responsible to make communication progress
  - More flexible than RDMA channel layer, being capable to access more performance oriented features

ADI-3 layer



- Full featured Abstract Device Interface:
  - Highest portable layer in MPICH2
  - Most complex layer but flexibility for even more optimizations

MVAPICH2

- MVAPICH2 is an open-source MPI-2 implementation over InfiniBand at RDMA channel level
  - http://nowlab.cse.ohio-state.edu/projects/mpi-iba/index.html
  - Latest release is MVAPICH2-0.6.5
- A new release version based on ADI3 layer is coming out
- Together with MVAPICH, MVAPICH2 is being used by more than 260 organizations worldwide (across 29 countries)

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 260 organizations and industry (across 29 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.

Motivation

- How to design highly optimized MPI2 over InfiniBand in MPICH2 layered structure?
- Understanding the performance and complexity trade-offs in a quantitative manner
- Experiences can be applied to design efficient MPI2 over other interconnects

Outline

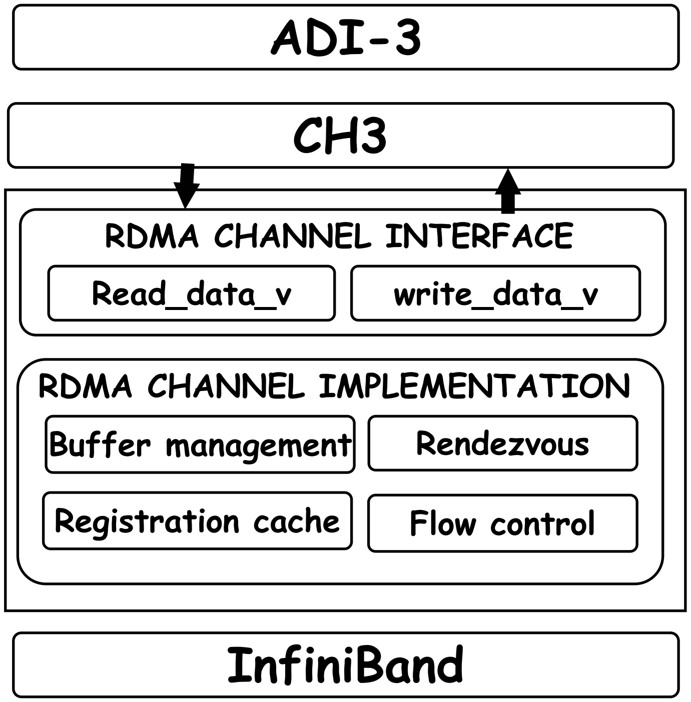
- Background
- Motivation
- Design and Implementation
- Performance Evaluation
- Conclusion and Future work

Design and Implementation

- Designed and implemented MPI-2 over InfiniBand based on
  - RDMA channel
  - CH3
  - ADI-3 layers

  for fair comparison
- For each layer, we present possible optimizations and performance trade-offs

RDMA Channel Layer

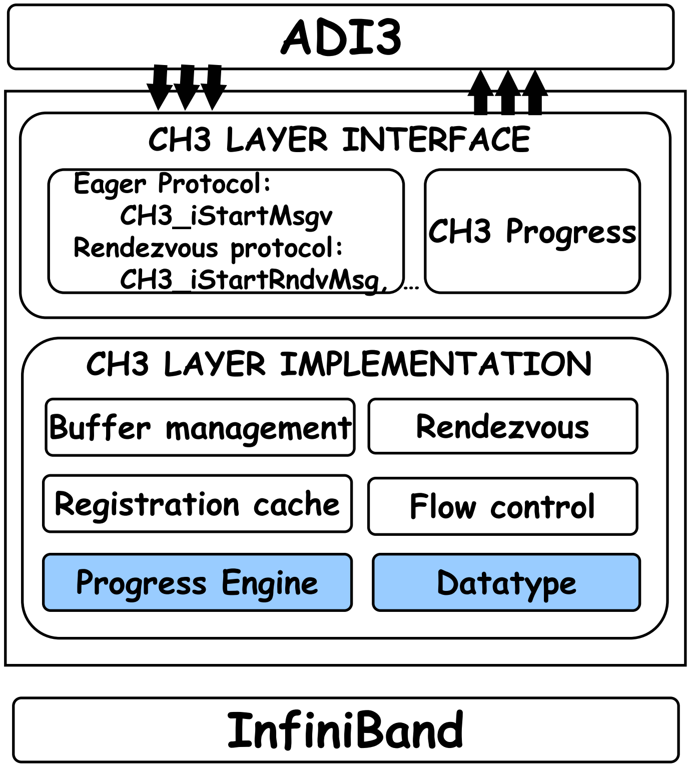


- A purely RDMA based design:
  - Provides stream semantics and progress is taken care of by CH3 layer
- Design Components:
  - Buffer management: eager protocol for short messages, copy and early completion
  - Rendezvous protocol
  - Registration cache
  - Flow control

RDMA Channel Layer

- Performance Trade-off:
  - CH3 only makes one outstanding request to RDMA channel
  - Progress engine is blocked for large messages since RDMA channel needs to keep the buffer for RDMA operations
  - Throughput is affected

CH3 Layer


ADI3

CH3 LAYER INTERFACE

Eager Protocol:
CH3_iStartMsgv
Rendezvous protocol:
CH3_iStartRndvMsg, …

CH3 Progress

CH3 LAYER IMPLEMENTATION

Buffer management

Rendezvous

Registration cache

Flow control

Progress Engine

Datatype

InfiniBand

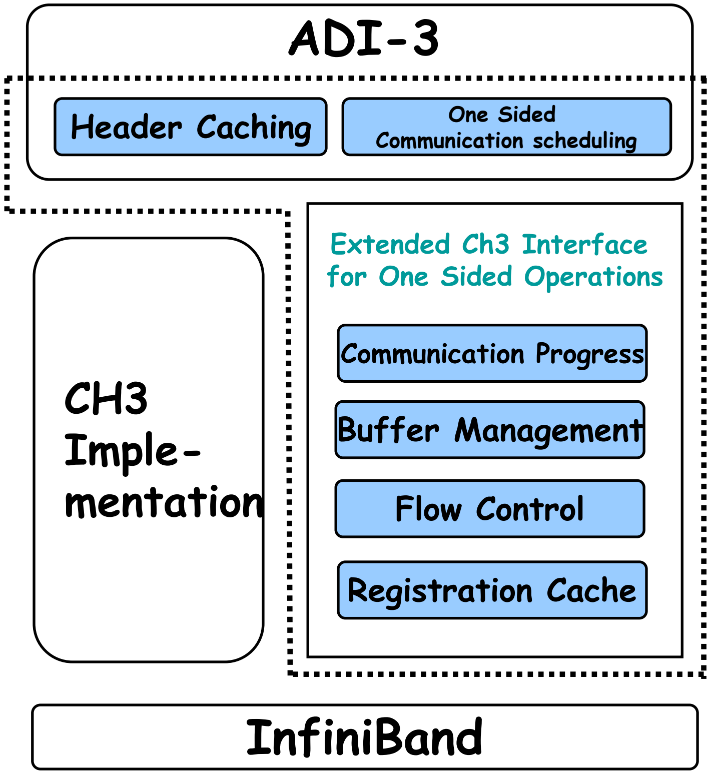
- Functionalities in RDMA channel layer are also needed here
- Extra Design Components:
  - Progress engine: all communication requests are handled at this layer
  - Datatype: ADI-3 flattens the datatype and provides CH3 layer the datatype information

Advantages of CH3 layer Design

- All communication requests are passed to CH3 layer by ADI3:
  - Multiple RDMA operations for large messages can start simultaneously. Removes restrictions in RDMA channel design
- CH3 has global picture of all datatype vectors for a MPI message
  - Optimization is possible at this layer*:

*G. Santhanaraman, J. Wu and D. K. Panda. Zero-copy MPI Derived Datatype communication over InfiniBand. EuroPVM/MPI '04

ADI-3 Layer



- A full ADI-3 layer implementation is extremely complex
- We extend CH3 implementation by ADI-3 level optimizations:
  - Header Caching
  - One Sided Communication

Optimizations at ADI-3

- Header Caching:
  - Cache header content at the receiver
  - Shrink header size if header content is the same as the last one
- One Sided Communication:
  - Direct one sided implementation [1]
  - One sided scheduling [2]

[1] J. Liu, W. Jiang, H. –W. Jin, D. K. Panda, W. Gropp, and R. Thakur. Higher Performance MPI-2 One-Sided Communication over InfiniBand. (CCGrid '04).

[2] W. Huang, G. Santhanaraman, H. –W. Jin, and D. K. Panda. Scheduling of MPI-2 One Sided Operations over InfiniBand. (CAC '05)

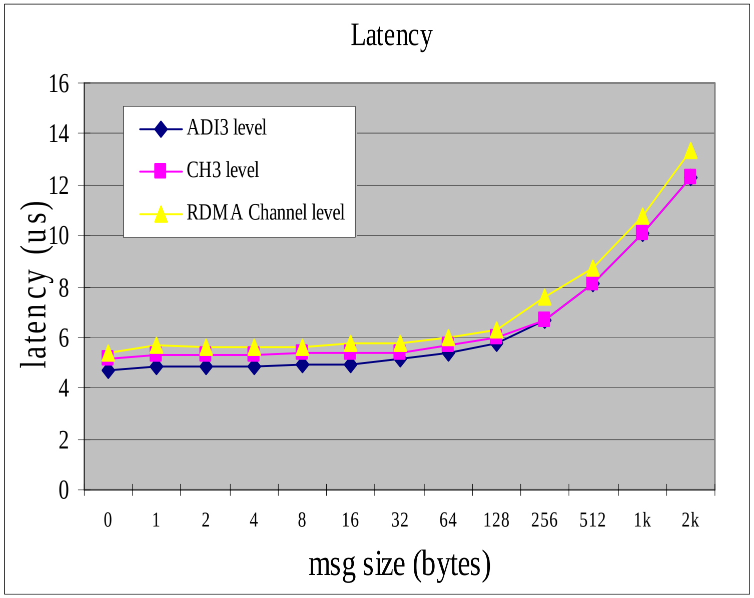
Outline

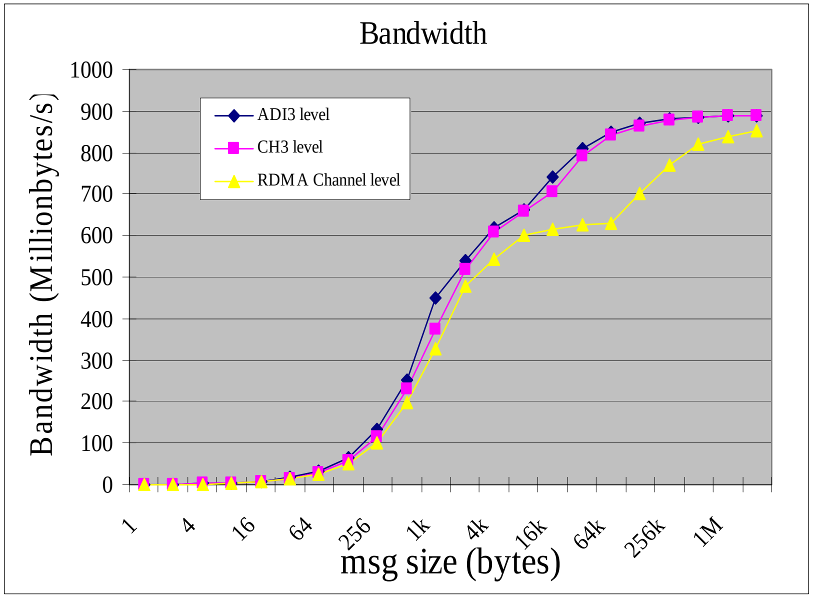
- Background
- Motivation
- Design and Implementation
- Performance Evaluation
  - Micro-Benchmarks
  - Application: NAS (IS), HPCC
- Conclusion and Future work

Experiment Setups

- Micro-benchmark:
  - Dual Intel Xeon 3.0 GHz (IA32), 2 GB memory, PCI-X HCA
  - Dual Intel Xeon 3.2 GHz (EM64T), 512 MB memory, PCI-Ex HCA
- Application:
  - Dual Intel Xeon 2.6 GHz, 2 GB memory, PCI-X HCA (16 nodes)

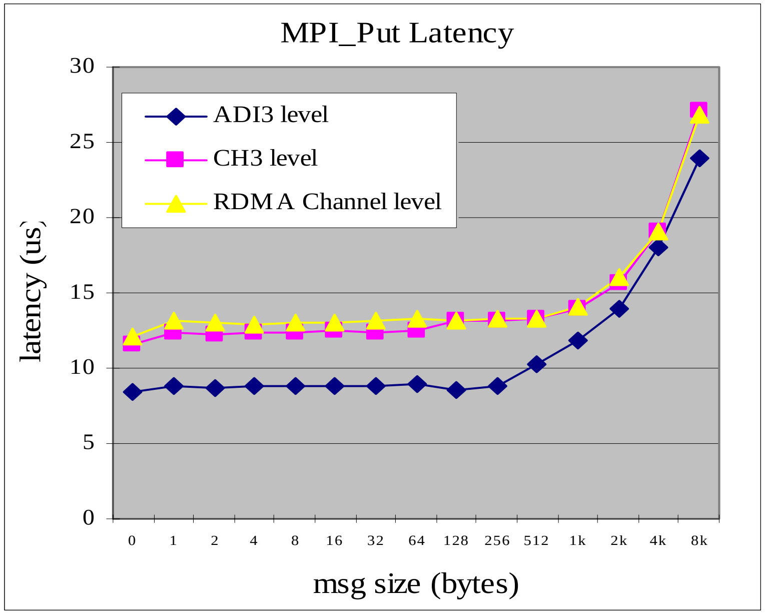
Point-to-Point Performance

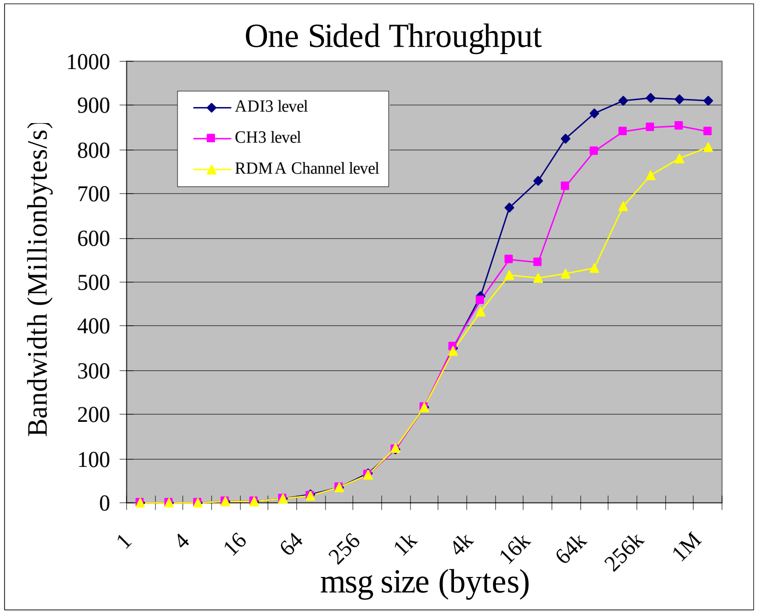




- Latency: 5.6us (RDMA Channel), 5.3us (Ch3), 4.9 us (ADI-3)
- Bandwidth: 28% improvement from RDMA channel to CH3 channel

One Sided Operations

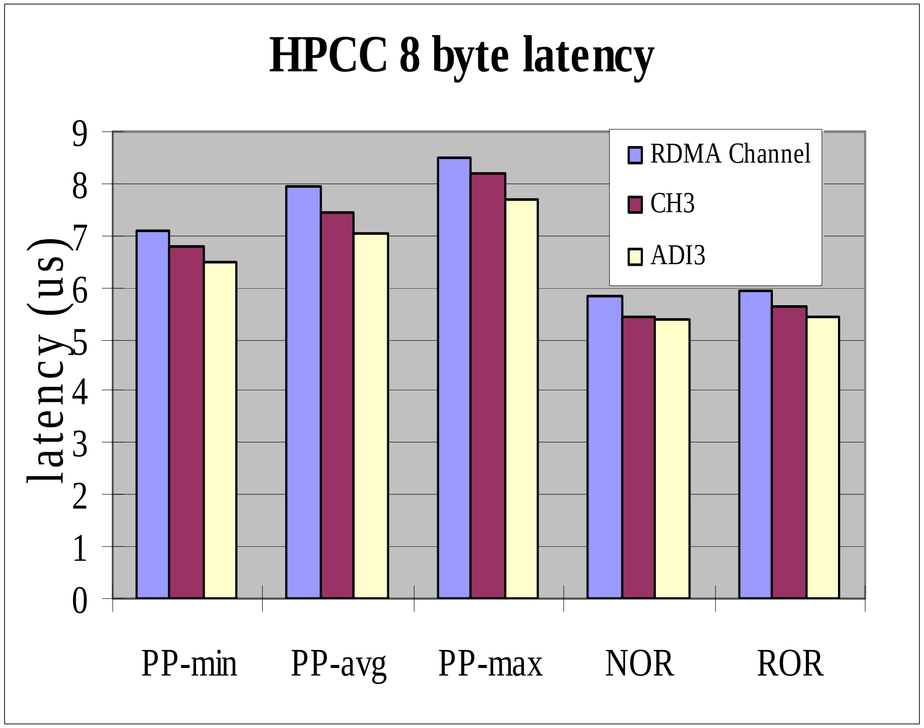




- Latency: 30 % improvement by optimization at ADI-3

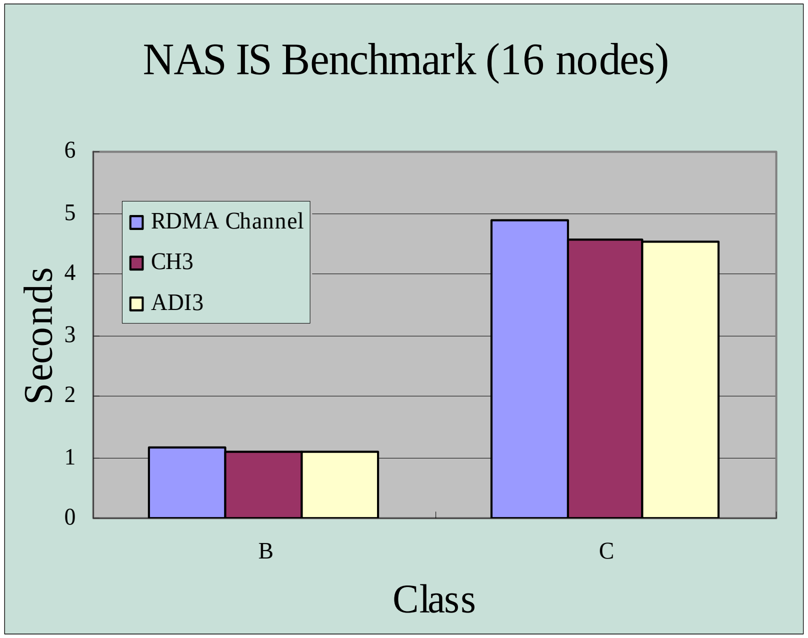
- Bandwidth: 28% improvement from RDMA channel to CH3 channel. Another 8.1 % by scheduling at ADI-3 (840 MB/s -> 910 MB/s)

HPCC Latency results



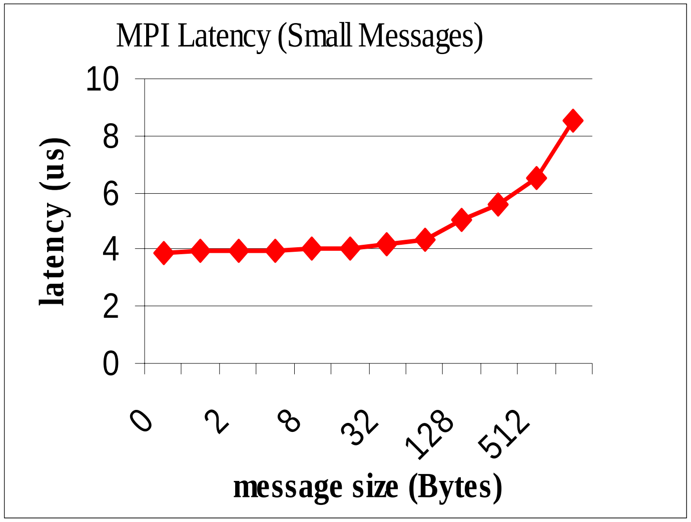
- HPCC suit 8 bytes latency results (16 nodes)
  - Minimum pingpong
  - Average pingpong
  - Maximum pingpong
  - Natural Ordered ring access
  - Random Ordered ring access
- Improvements:
  - CH3 over RDMA channel: 7%
  - ADI-3 over CH3: 6 %

NAS-Integer Sort

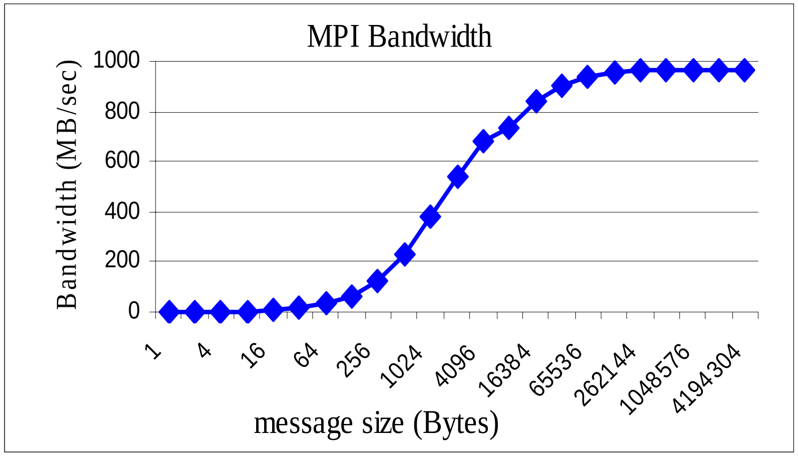


- NAS-IS: ADI-3 and CH3 designs show 7% benefit compared to RDMA channel design

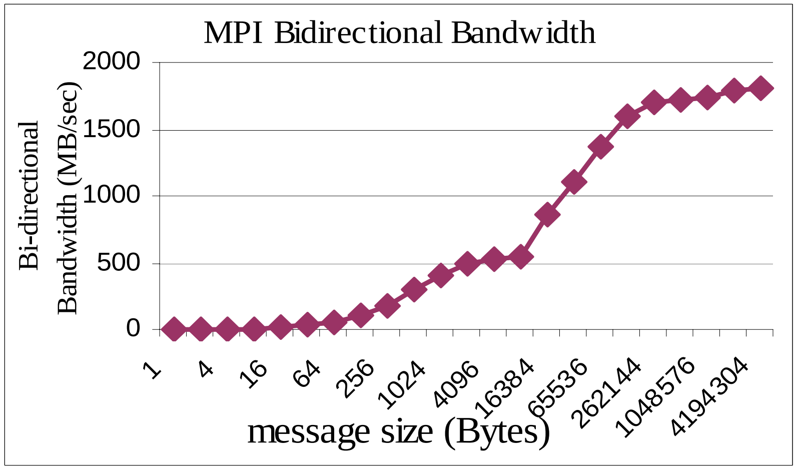
MVAPICH2-Gen2 with InfiniBand 4X SDR: MPI-Level Performance



3.91



968.2



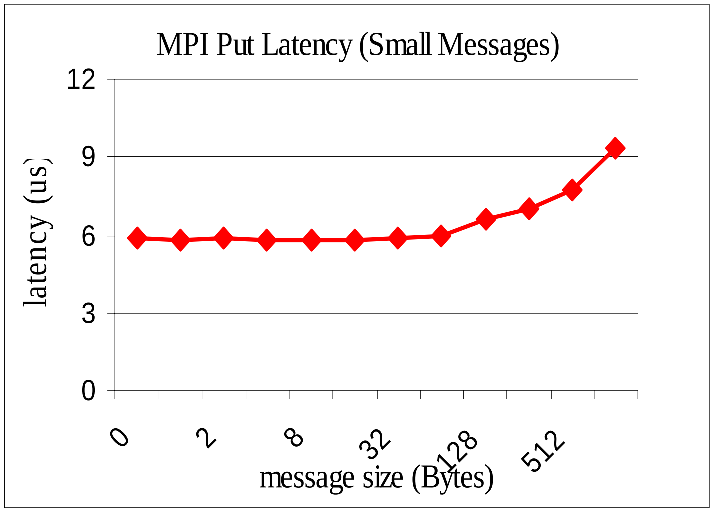
1801

- Single port results only
- DDR results expected to be same as MVAPICH-Gen2 1.0 (2.8 microsec)

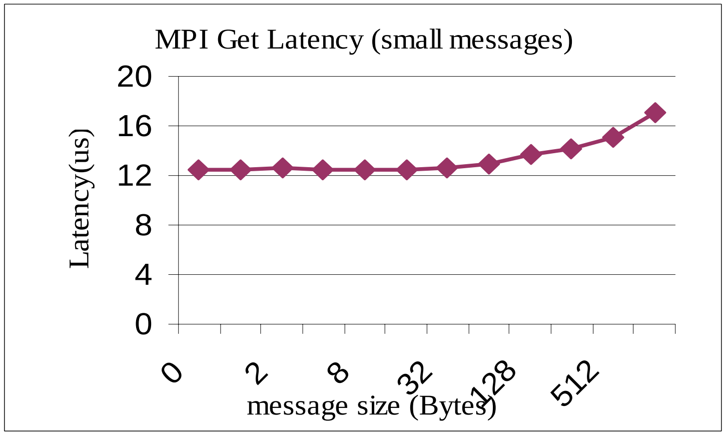
08/21/05

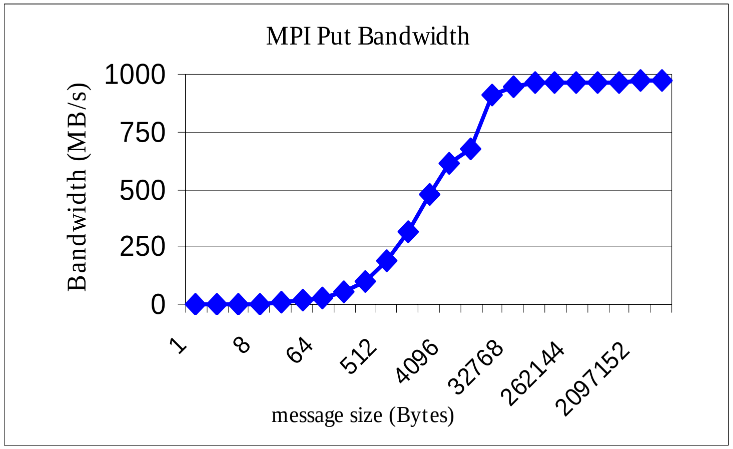
MVAPICH2-Gen2 with InfiniBand 4X SDR: MPI One Sided Performance

5.84



12.52





968

- Single port results only

Conclusions

- We analyze the complexity and performance trade-offs of designing MPI-2 over InfiniBand based on MPICH2's layered stack:
  - RDMA is the simplest interface to port
  - CH3 adds complexity to implement the progress engine, but increases throughput by 28%
  - ADI-3 is the most complex layer to implement, but more optimizations benefits latency and one sided communication

Current Work

- Coming up with a full fledged MPI2 design over InfiniBand based on ADI-3 layer
  - Supports multiple methods (shared memory, multi-rail, …)
  - Optimizes collectives operations
  - More optimized one sided communication
- MVAPICH2 0.7.0 with these features will be released soon

Acknowledgements

Our research is supported by the following organizations

- Current Funding support by

Office of Science U.S. DEPARTMENT OF ENERGY

NSF

intel.

Mellanox Technologies

Sun microsystems

- Current Equipment donations by

Mellanox Technologies

intel.

AMD

Sun microsystems

AMMASSO

Microway Technology you can count on

PathScale

SilverStorm

Web Pointers

Network Based Computing Laboratory

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