MVAPICH/MVAPICH2 Update

Presentation at Open Fabrics Sonoma Conference
(April '08)

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/MVAPICH2 Project
- Features of MVAPICH 1.0 and MVAPICH2 1.0
- Sample Performance Numbers
  - Point-to-point (Mellanox and Qlogic)
  - Scalable Startup
  - Multi-core-aware Optimized Collectives
  - UD-based Design
  - Lustre ADIO Support
- Upcoming Features and Issues
  - XRC support
  - Hybrid UD-RC Design
  - Asynchronous Progress
  - One-sided with Passive Synchronization
  - Checkpoint-Restart with Shared Memory
- Requirements from OpenFabrics
- Conclusions

Overview of MVAPICH/MVAPICH2 Project

- High Performance MPI Library for InfiniBand and 10GigE/iWARP Clusters
  - MVAPICH (MPI-1) and MVAPICH (MPI-2)
  - Used by more than 660 organizations in 42 countries (registered with OSU)
  - More than 18,000 downloads from OSU web site
  - Empowering many TOP500 clusters (Nov '07 listing)
    - 14,336-core cluster at State of New Mexico (3rd rank)
    - 5,848-core cluster at TACC (22nd rank)
    - 9,216-core cluster at LLNL (29th rank)
  - Running on a large number of production clusters including NCSA and TACC Ranger
  - Available with software stacks of many InfiniBand, iWARP and server vendors including Open Fabrics Enterprise Distribution (OFED)
  - http://mvapich.cse.ohio-state.edu/

New Features of MVAPICH 1.0

- Enhanced mpirun_rsh for scalable launching
  - Provides a two-level approach (nodes and cores within a node)
- Asynchronous Progress
  - Provides better overlap between computation and communication
- Flexible message coalescing
  - enable/disable coalescing
  - Allows varying degrees of coalescing
- UD-based support
  - Best performance and scalability with constant memory footprint for communication contexts
- Support for Automatic Path Migration (APM)
- Multi-core optimizations for Collectives
- Support for ConnectX
- Support for Qlogic/PSM
- Lustre ADIO support (contributed by ORNL)

New Features of MVAPICH2 1.0

- Message coalescing support
- Hot-spot avoidance mechanism for alleviating network congestion in large clusters
- Application-initiated systems-level checkpoint
  - in addition to the automatic systems-level checkpoint from 0.9.8
- Automatic Path Migration (APM) support
- RDMA Read
- Blocking
- Multi-rail support for iWARP
- RDMA CM-based connection management (Gen2-IB and Gen2-iWARP)
- On-demand connection management for uDAPL (including Solaris)

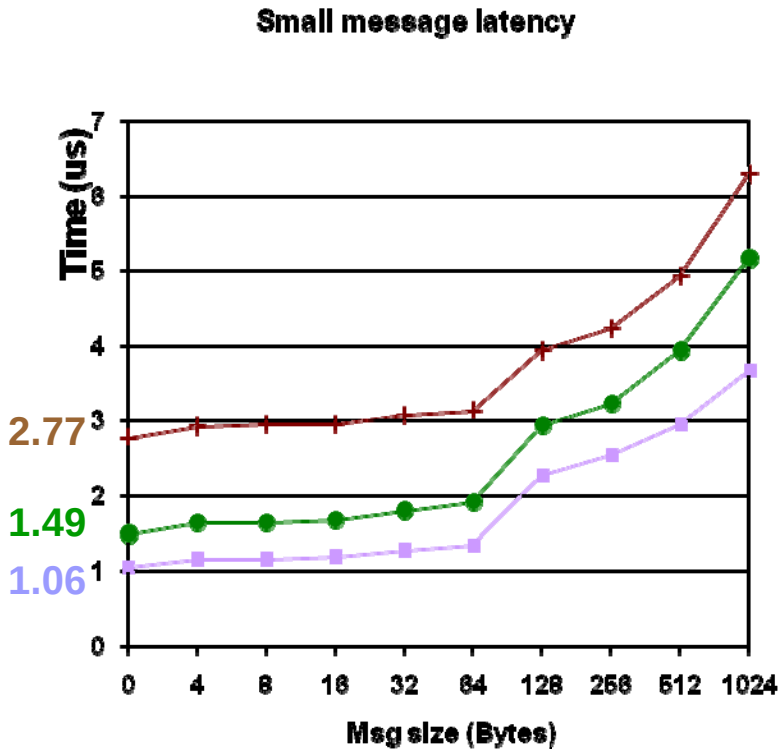
Support for Multiple Interfaces/Adapters

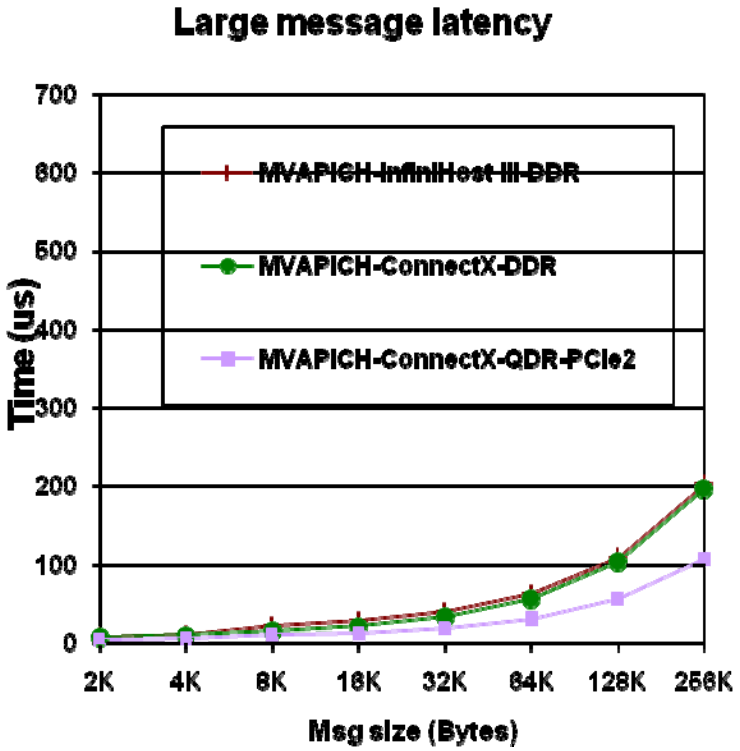
- OpenFabrics/Gen2-IB
  - All IB adapters supporting Gen2
- Qlogic/PSM
- uDAPL
  - Linux-IB
  - Solaris-IB
  - Other adapters such as Neteffect 10GigE
- OpenFabrics/Gen2-iWARP
  - Chelsio
- VAPI
  - All IB adapters supporting VAPI
- TCP/IP
  - Any adapter supporting TCP/IP interface
- Shared Memory Channel (MVAPICH), for running applications in a node with multi-core processors

Presentation Overview

- Overview of MVAPICH/MVAPICH2 Project
- Features of MVAPICH 1.0 and MVAPICH2 1.0
- Sample Performance Numbers
  - Point-to-point (Mellanox and Qlogic)
  - Scalable Startup
  - Multi-core-aware Optimized Collectives
  - UD-based Design
  - Lustre ADIO Support
- Upcoming Features and Issues
  - XRC support
  - Hybrid UD-RC Design
  - Asynchronous Progress
  - One-sided with Passive Synchronization
  - Checkpoint-Restart with Shared Memory
- Requirements from OpenFabrics
- Conclusions

MVAPICH Latency (One-way): IBA (Mellanox)

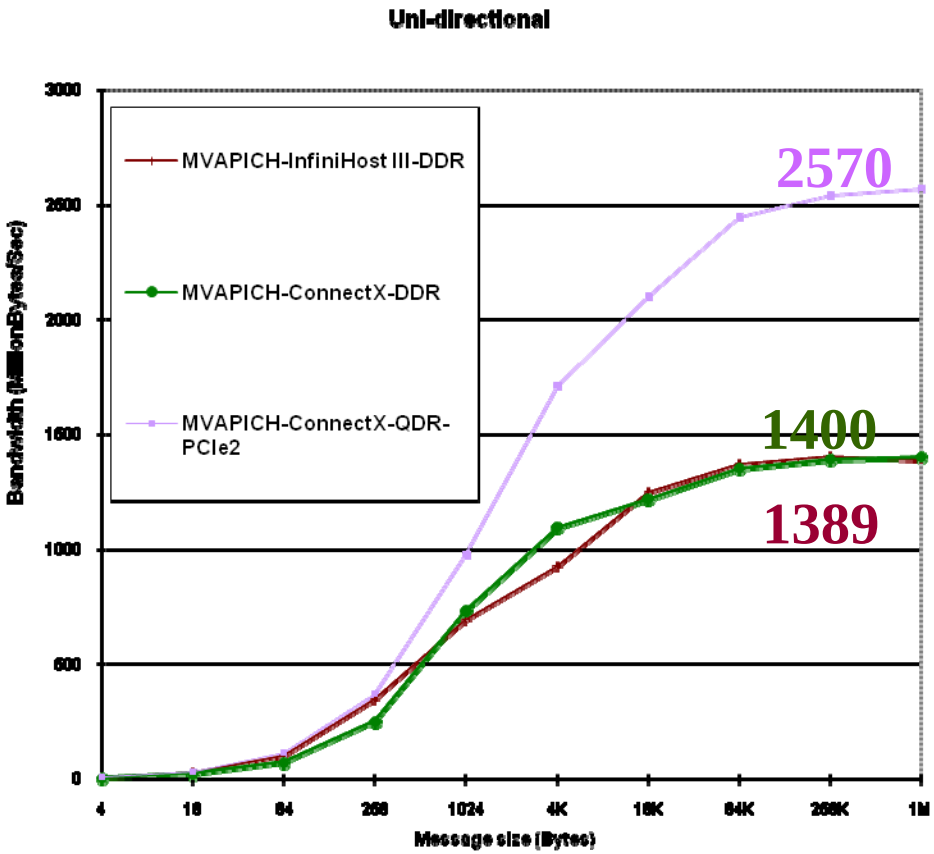




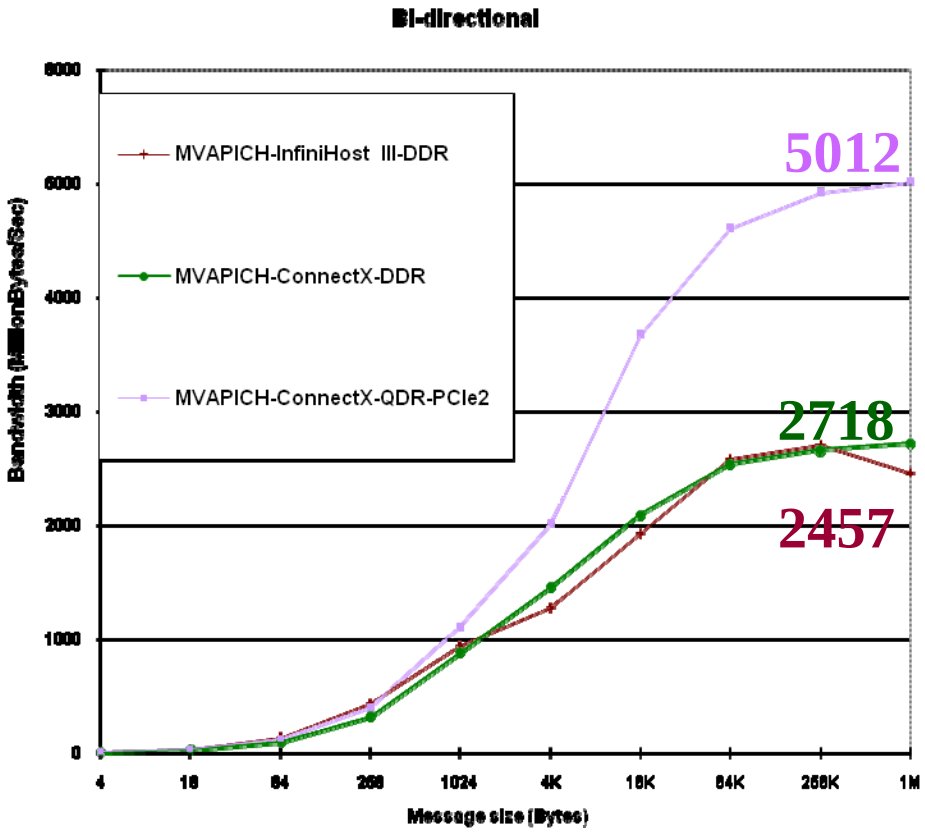
InfiniHost III and ConnectX-DDR: 2.33 GHz Quad-core (Clovertown) Intel with IB switch

ConnectX-QDR-PCIe2: 2.83 GHz Quad-core (Harpertown) Intel with back-to-back

MVAPICH Bandwidth: IBA (Mellanox)

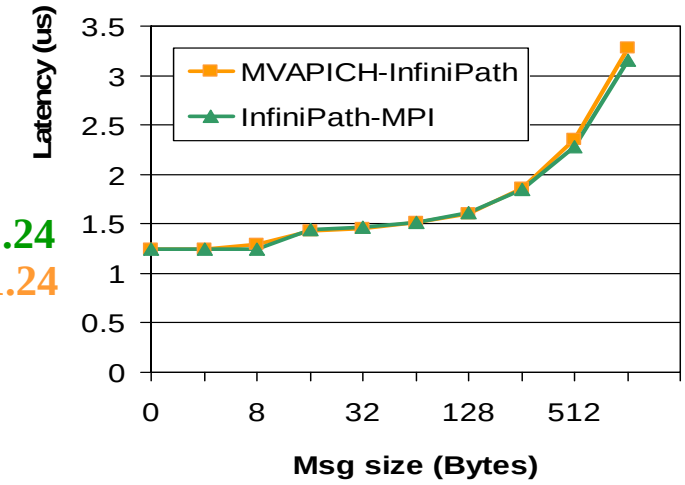


InfiniHost III and ConnectX-DDR: 2.33 GHz Quad-core (Clovertown) Intel with IB switch



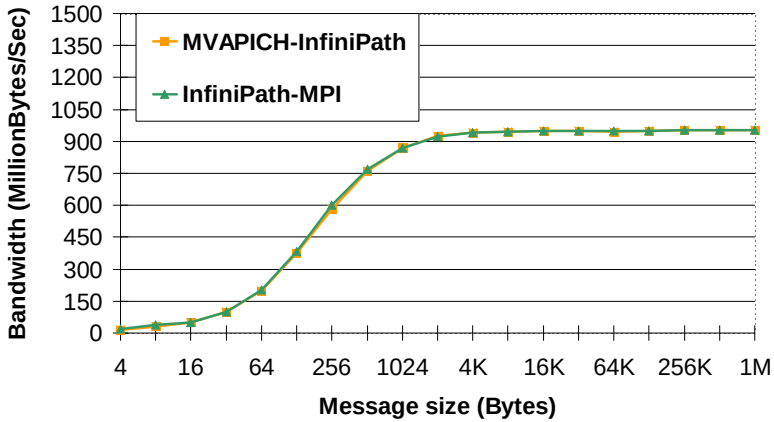
ConnectX-QDR-PCIe2: 2.83 GHz Quad-core (Harpertown) Intel with back-to-back

MVAPICH-PSM Performance: AMD Opteron (QLogic-SDR)



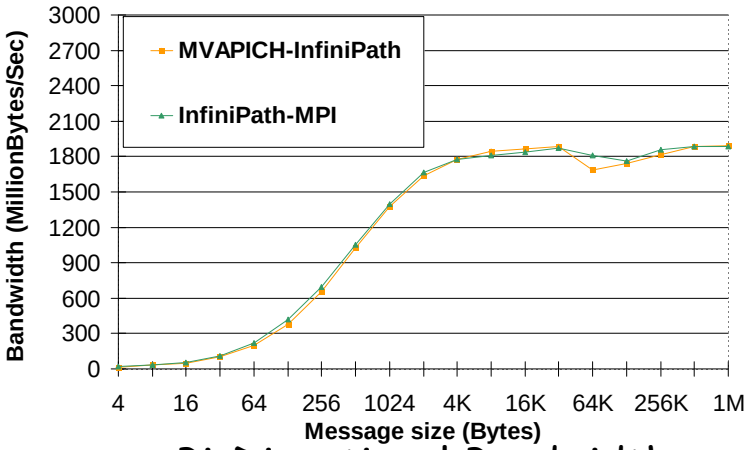
1.24
1.24

Small Message Latency



953
952

Uni-Directional Bandwidth

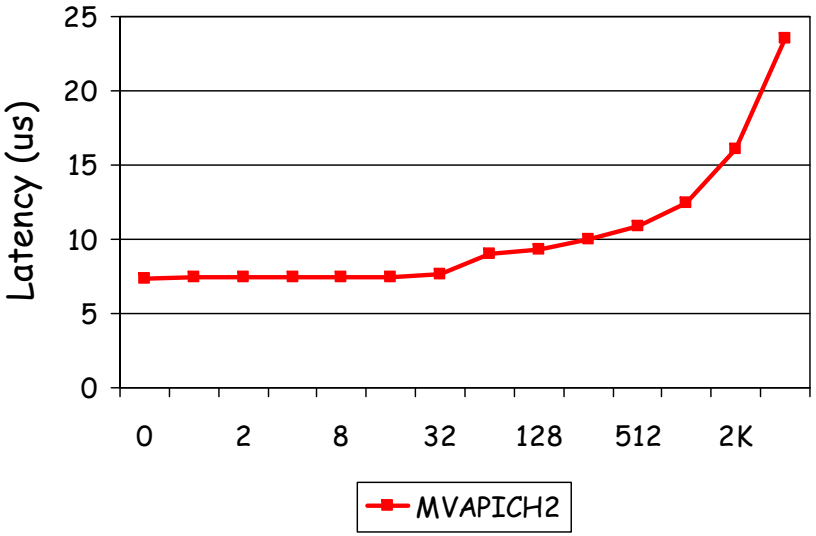


1888
1888

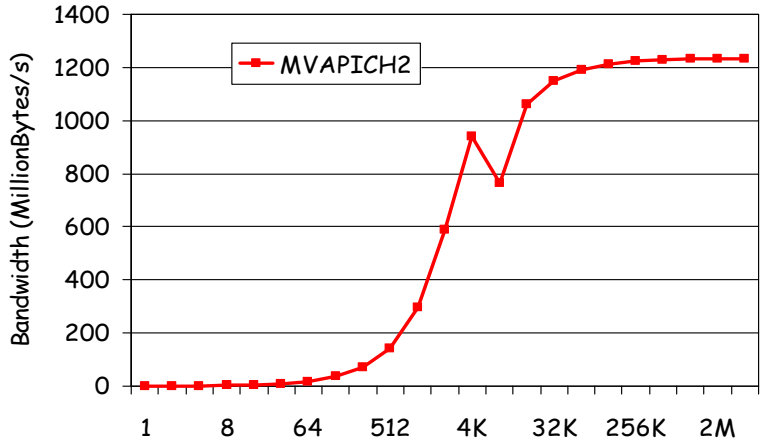
Bi-Directional Bandwidth

MPI-level Performance: iWARP with Chelsio

2.0 GHz Quad-core Intel
with 10GigE (Fulcrum) switch
NIC Firmware 5.0
OFED 1.3

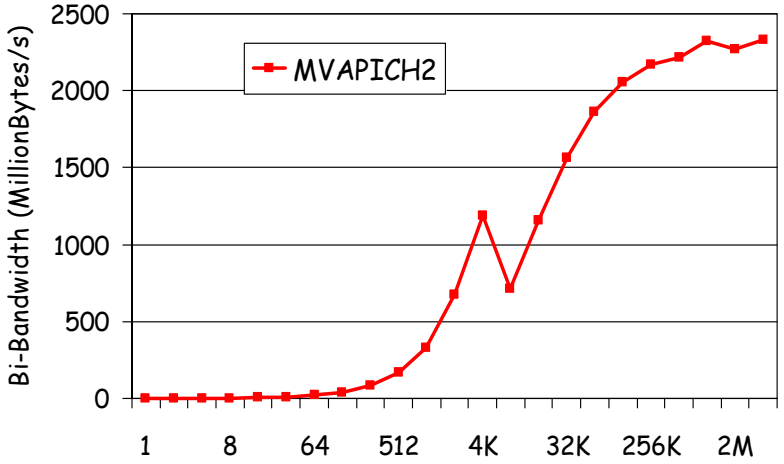


MVAPICH2 gives a latency of about 7.31us



1233

Uni-Directional Bandwidth

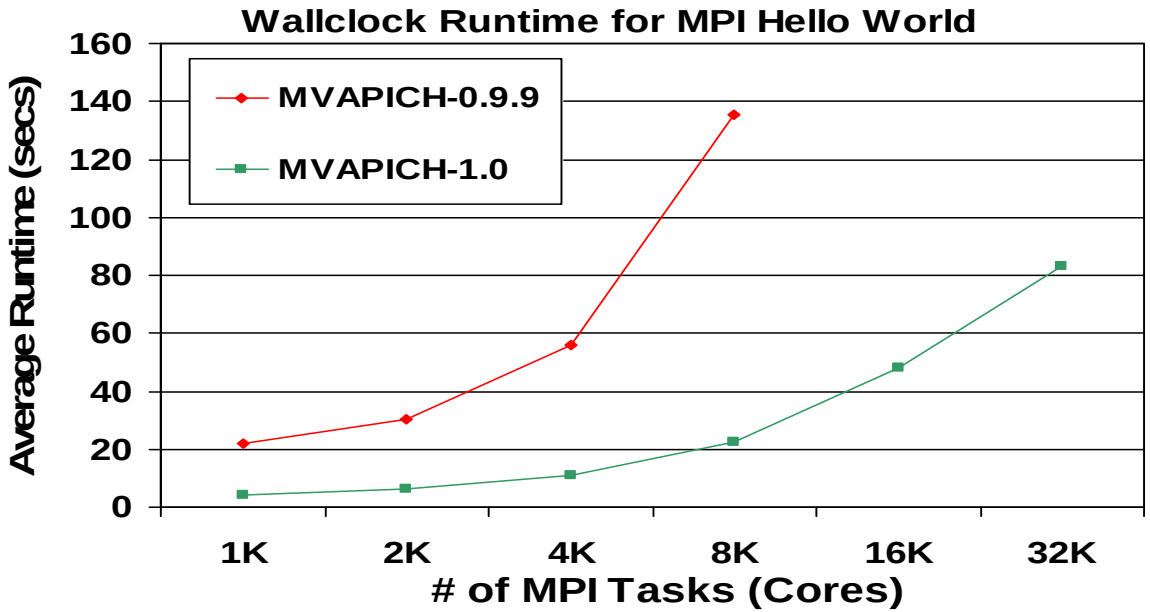


2328

Bi-Directional Bandwidth

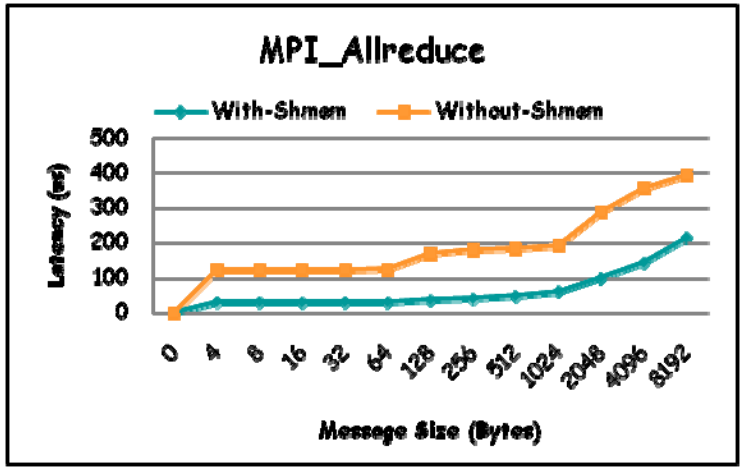
Scalable Startup

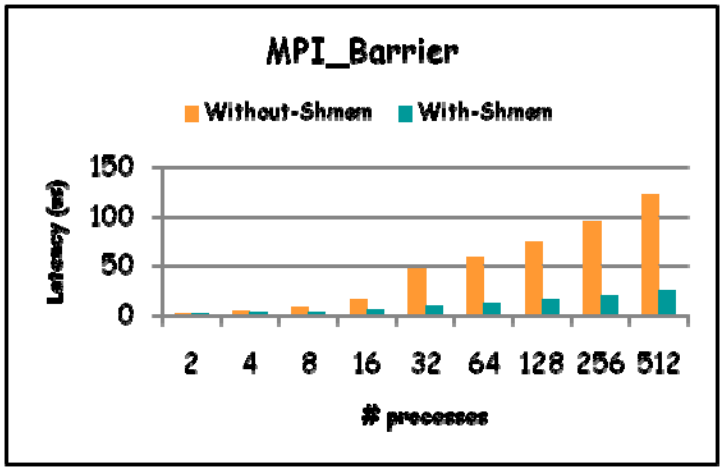
- An enhanced mpirun_rsh framework in MVAPICH 1.0 to significantly cut down job start-up on large clusters
- Will be enhanced further for MVAPICH 1.1
- Will also be available with MVAPICH2 1.1

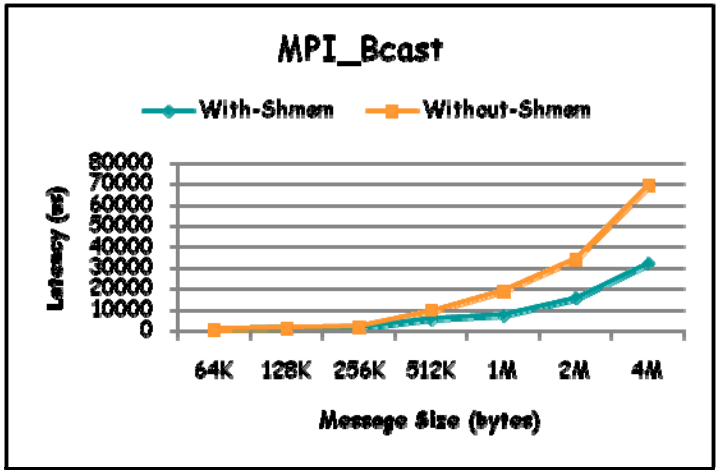


Courtesy TACC

Shared Memory Collectives in MVAPICH (512 cores)





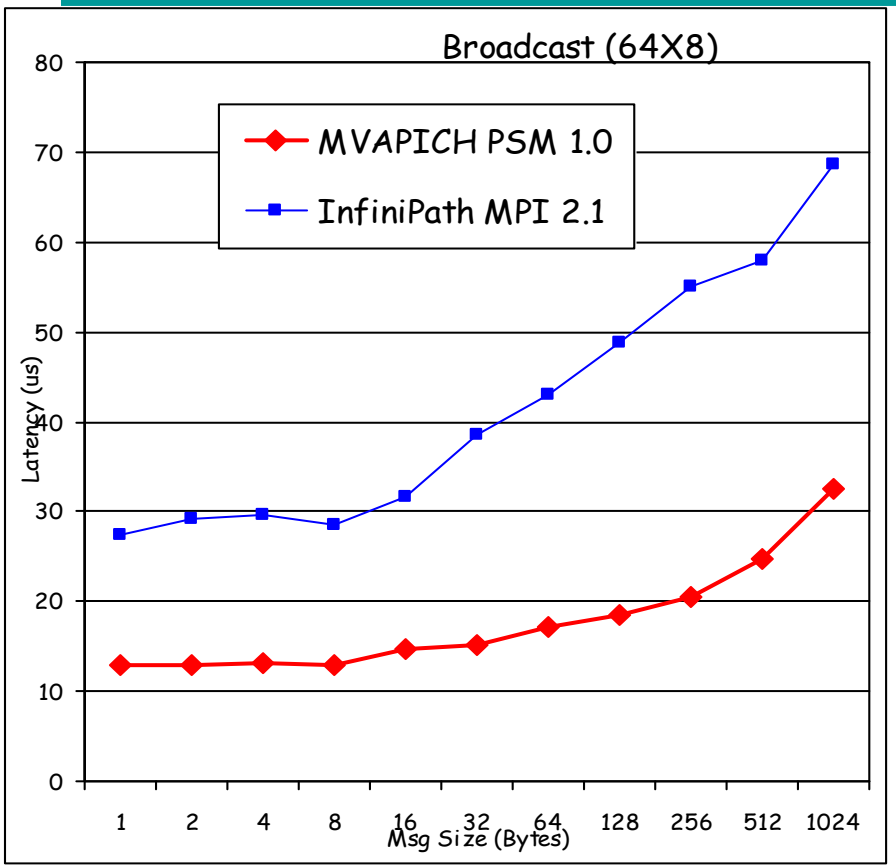


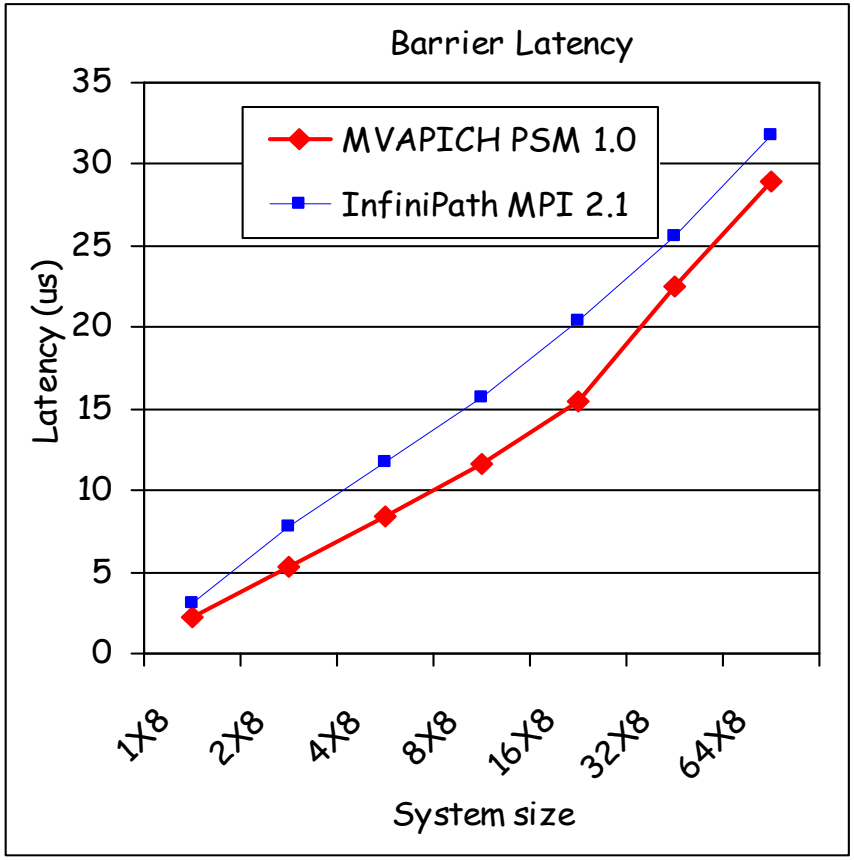
- Shared Memory hierarchical scheme improves latency of MPI_Allreduce by four times and MPI_Barrier by six times respectively

- Shared Memory scheme for MPI_Bcast improves latency by a factor of two

- Tuned for the message sizes and # cores

MVAPICH-PSM Collective Performance (512-cores)

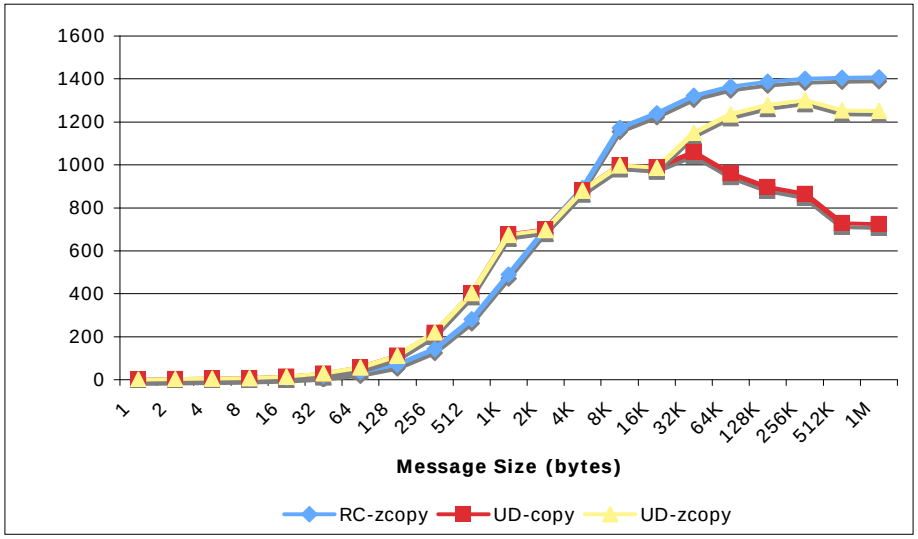




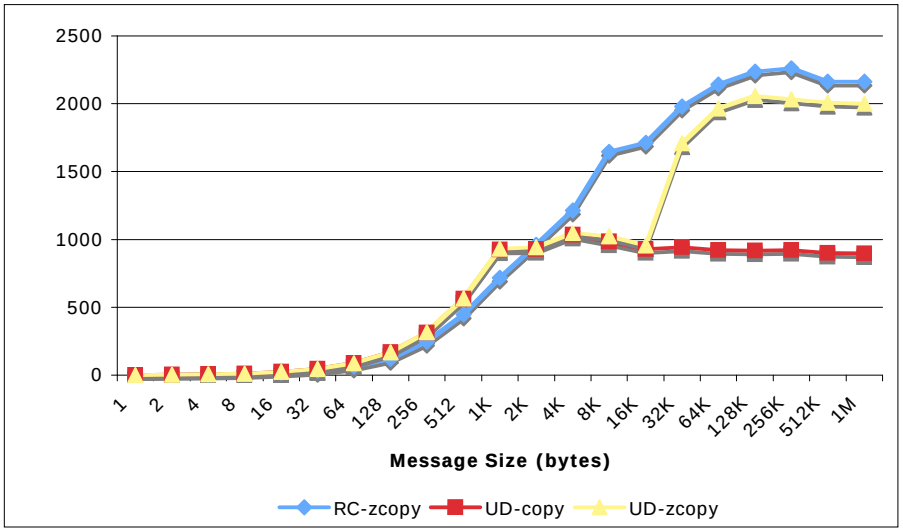
- 64 Intel Quad-core systems with dual sockets; PCIe InfiniPath Adapters
- Significant performance improvement for MPI_Bcast and MPI_Barrier

Zero-Copy over Unreliable Datagram (UD)

Uni-Directional Bandwidth



Bi-Directional Bandwidth

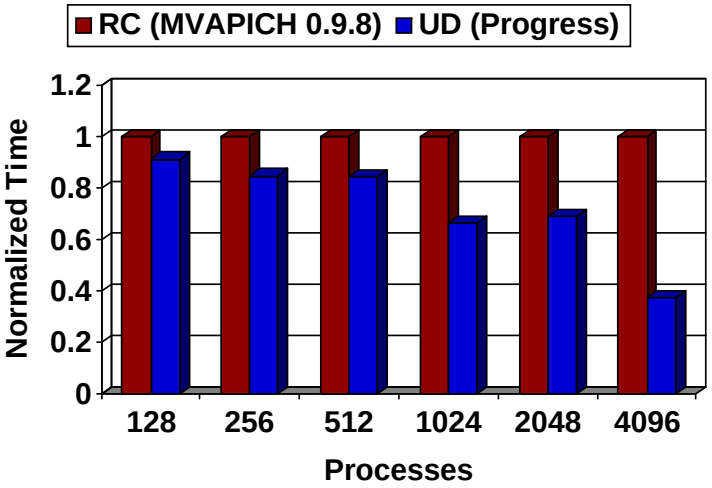


- **Using a novel technique, zero-copy transfers can be made over UD.**
- **Performance very close to that of RC**
- **Supported in MVAPICH 1.0**

M. Koop, S. Sur and D. K. Panda, Zero-Copy Protocol for MPI using InfiniBand Unreliable Datagram, Cluster 2007

SMG2000

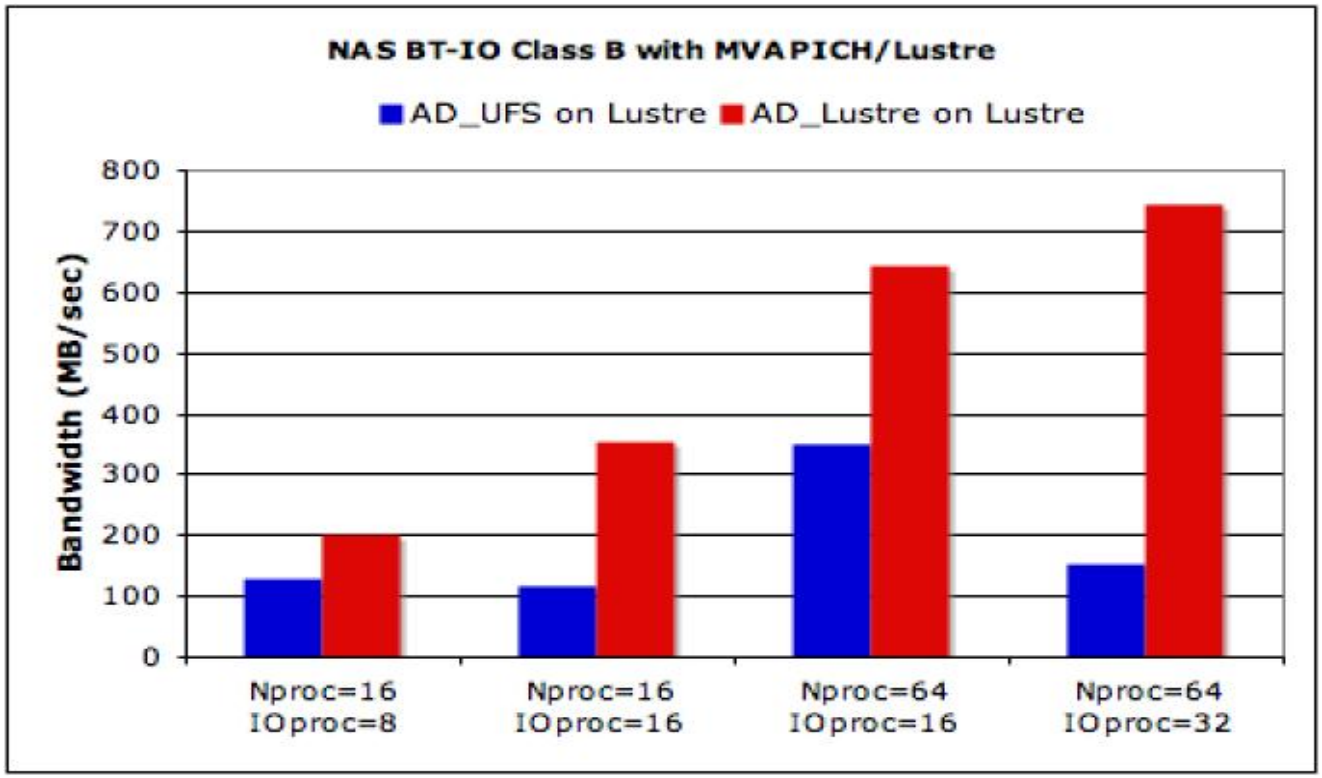
	RC (MVAPICH 0.9.8)				UD Design			
	Conn.	Buffers	Struct.	Total	Conn	Buffers	Struct	Total
512	22.9	65.0	0.3	88.2	0	37.0	0.2	37.2
1024	29.5	65.0	0.6	95.1	0	37.0	0.4	37.4
2048	42.4	65.0	1.2	107.4	0	37.0	0.9	37.9
4096	66.7	65.0	2.4	**134.1**	0	37.0	1.7	**38.7**



- Performance is enhanced considerably with UD
- Large number of communicating peers per process (992 at maximum)
  - UD reduces HCA cache thrashing
  - Very communication intensive
- 27 packet drops at 4K processes with 1.4 billion MPI messages
- Large difference in memory consumption, even only 1/4 of connections made

Lustre ADIO Support

- Contributed by Weikuan Yu (Future Technologies Group at ORNL)

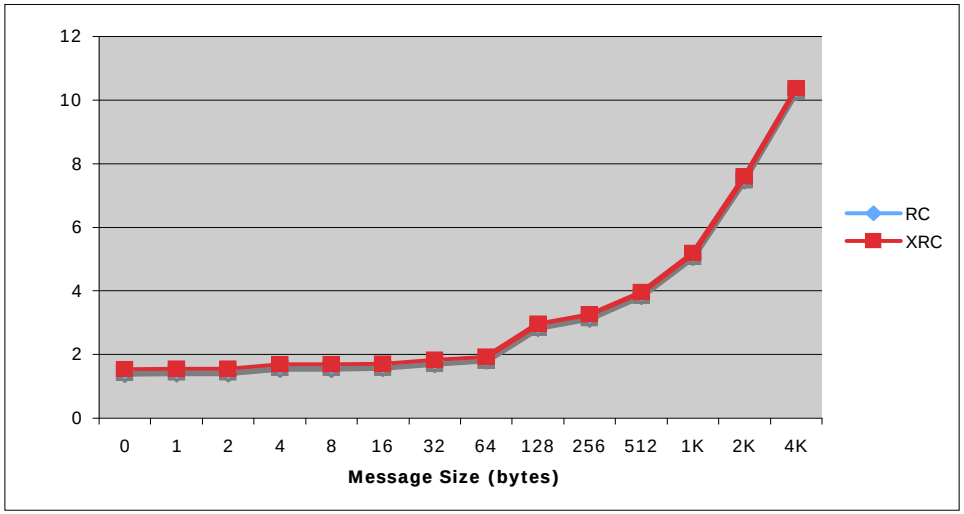


Presentation Overview

- Overview of MVAPICH/MVAPICH2 Project
- Features of MVAPICH 1.0 and MVAPICH2 1.0
- Sample Performance Numbers
  - Point-to-point (Mellanox and Qlogic)
  - Scalable Startup
  - Multi-core-aware Optimized Collectives
  - UD-based Design
  - Lustre ADIO Support
- Upcoming Features and Issues
  - XRC support
  - Hybrid UD-RC Design
  - Asynchronous Progress
  - One-sided with Passive Synchronization
  - Checkpoint-Restart with Shared Memory
- Requirements from OpenFabrics
- Conclusions

XRC Support with ConnectX

- XRC (eXtended Reliable Connection) is being proposed for large-scale clusters
- We have designed and implemented an initial prototype of MVAPICH with XRC support

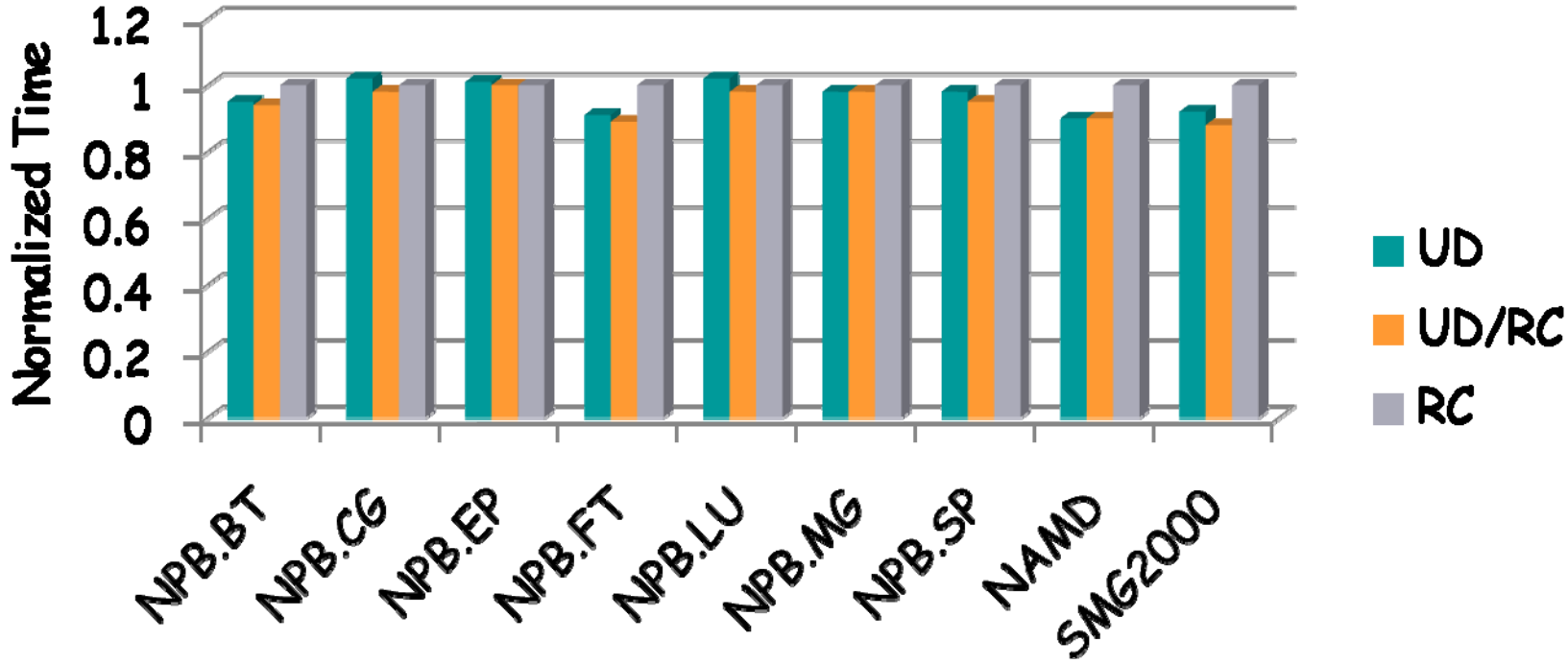


- Results for latency are nearly identical between the use of RC and XRC transports
- 1.49usec for RC, 1.54usec for XRC

Hybrid UD-RC Design

- Current UD-based design in MVAPICH 1.0 delivers good performance
  - Some overheads on large-scale clusters for some applications
- Working on a new hybrid UD-RC design
- Delivers equal to or better performance than RC or UD design
- Will be available in MVAPICH 1.1

Impact of RC/UD Hybrid Design



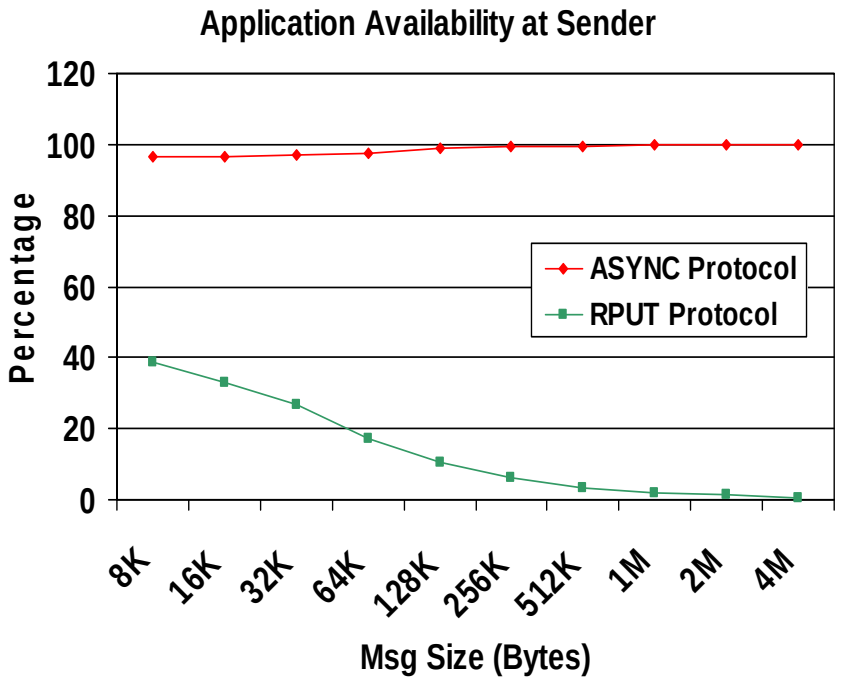
Application benchmark results on 512-core system

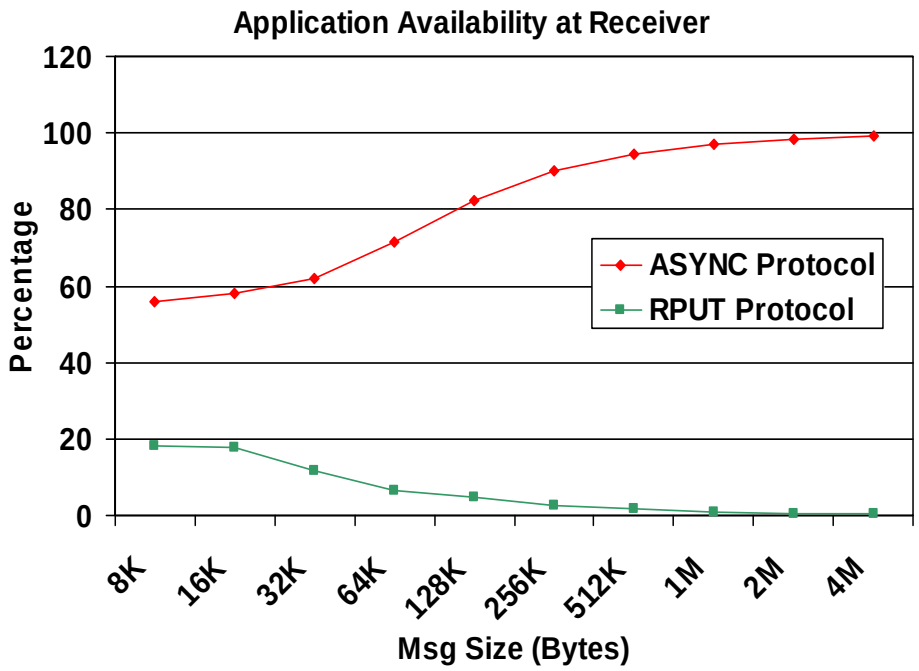
Combine the benefits of both RC and UD together

M. Koop, T. Jones and D. K. Panda, "MVAPICH-Aptus: Scalable High-Performance Multi-Transport MPI over InfiniBand," Int'l Parallel and Distributed Processing Symposium (IPDPS '08)

Asynchronous Progress (Mellanox DDR)

- A preliminary design for asynchronous progress (both at sender and receiver) is available in MVAPICH 1.0
- An enhanced design is being worked out for MVAPICH 1.1

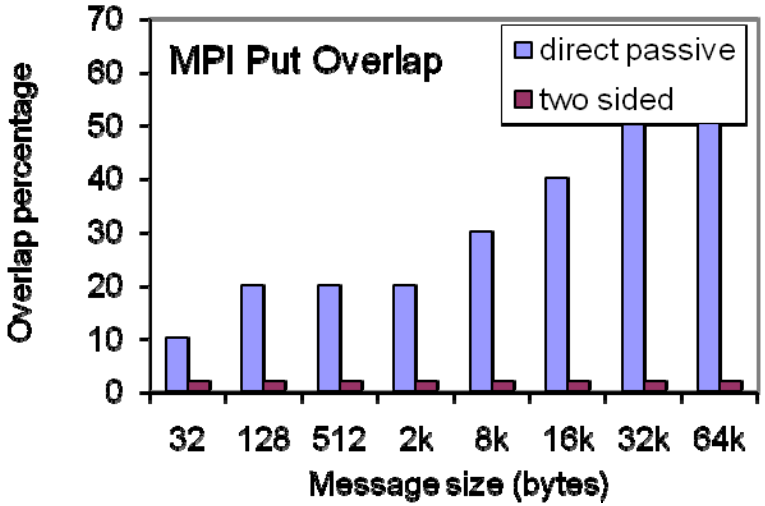


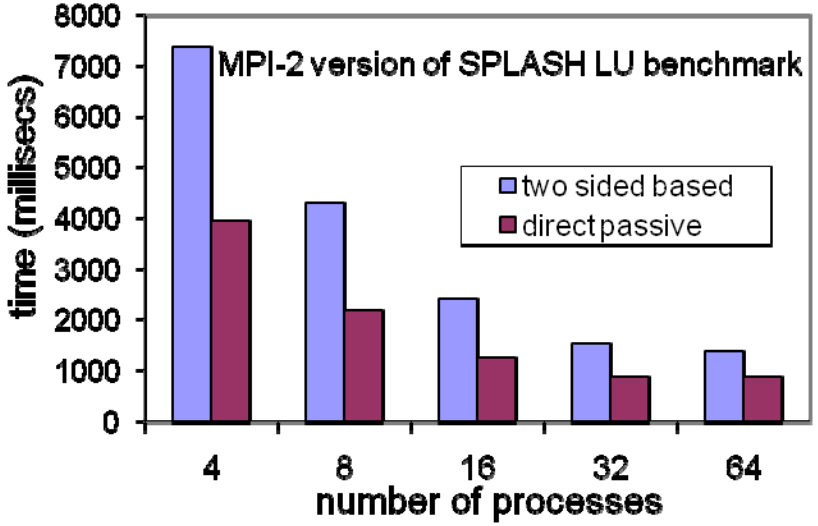


Results for SMB Benchmark

Passive Synchronization for One-Sided Operations with Atomic Operations

- One-sided operations in MPI-2 semantics have two synchronization schemes (Active and Passive)
- Have taken InfiniBand atomic operations into account to implement high performance and scalable passive synchronization
- Will be available in MVAPICH2 1.1





G. Santhanaraman, S. Narravula, and D. K. Panda, Designing Passive Synchronization for MPI-2 One-Sided Communication to Maximize Overlap , IEEE International Parallel and Distributed Processing Symposium (IPDPS '08), Miami, Florida, April, 2008.

Checkpoint-Restart with Shared Memory

- Checkpoint-Restart (CR) with BLCR is already available with MVAPICH2 1.0
- Since BLCR didn't support checkpointing shared memory, MVAPICH2 1.0 with CR support uses network loopback for intra-node communication
- A new solution has been worked out to provide CR support with the latest release of BLCR
  - Supports shared-memory collectives too
- Will be available with MVAPICH2 1.1

Requirements from OpenFabrics

- Fast Memory Registration
- Reliable Multicast

Conclusions

- MVAPICH and MVAPICH2 are being widely used in stable production IB clusters delivering best performance and scalability
- Also enabling clusters with iWARP support
- The user base stands at more than 660 organizations
- New features for scalability, high performance and fault tolerance support are aimed to deploy large-scale clusters (50-100K) nodes in the near future

Acknowledgements

Our research is supported by the following organizations

- Current Funding support by

- Current Equipment support by


Acknowledgements

- Current Students
  - L. Chai (Ph.D.)
  - W. Huang (Ph.D.)
  - M. Koop (Ph.D.)
  - R. Kumar (M.S.)
  - A. Mamidala (Ph.D.)
  - S. Narravula (Ph.D.)
  - R. Noronha (Ph.D.)
  - J. Sridhar (M. S.)
  - G. Santhanaraman (Ph.D.)
  - K. Vaidyanathan (Ph.D.)
- Current Programmers
  - B. Curtis
  - J. Perkins
- Past Students
  - P. Balaji (Ph.D.)
  - D. Buntinas (Ph.D.)
  - S. Bhagvat (M.S.)
  - B. Chandrasekharan (M.S.)
  - W. Jiang (M.S.)
  - S. Kini (M.S.)
  - S. Krishnamoorthy (M.S.)
  - J. Liu (Ph.D.)
  - S. Sur (Ph.D.)
  - A. Vishnu (Ph.D.)
  - J. Wu (Ph.D.)
  - W. Yu (Ph.D.)

Web Pointers

MVAPICH Web Page
http://mvapich.cse.ohio-state.edu/

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