MVAPICH/MVAPICH2 Update

Presentation at Open Fabrics Sonoma Conference
(March '09)
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 1.1 and MVAPICH2 1.2
- Sample Performance Numbers
  - Point-to-point (Mellanox, Qlogic & Chelsio)
  - Scalable Startup
  - Hybrid UD-RC/UD-XRC Design
- Upcoming MVAPICH 1.2 and MVAPICH2 1.4 Features and Issues
  - Network Reliability
  - Dynamic Process Management
  - Kernel-based Single copy Intra-node Support
  - MVAPICH2-PSM Support
- Future Plans
- OpenFabrics Requirements
- Conclusions

Overview of MVAPICH/MVAPICH2 Project

- High Performance MPI Library for InfiniBand and 10GigE/iWARP Clusters
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2)
  - Available since 2002
  - Used by more than 870 organizations in 45 countries (registered with OSU)
  - More than 27,000 downloads from OSU web site
  - Empowering many TOP500 clusters in production environment (Nov '08 listing)
    - 62,976-core cluster (Ranger) at TACC (6th rank)
    - 18,176-core cluster (Chinook) at PNNL (20th rank)
    - Many others
  - 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 1.1

- Released on 11/14/08
- Part of OFED 1.4
- OpenFabrics-Gen2
  - eXtended Reliable Connection (XRC) Support
  - Lock-free Asynchronous Progress with RDMA Read for better overlap between computation and communication
  - Efficient intra-node shared memory support for diskless clusters
  - Optimized support for collectives including k-nomial-based broadcast and shared-memory-based algorithms
- OpenFabrics-Gen2-Hybrid
  - Newly introduced interface in 1.1
  - Replaces UD interface in 1.0
  - Targeted for emerging multi-thousand-core clusters to achieve the best performance with minimal memory footprint
  - Adaptive selection during run-time (based on application and systems characteristics) to switch between
    - RC and UD (or between XRC and UD) transports
  - Multiple buffer organization with XRC support

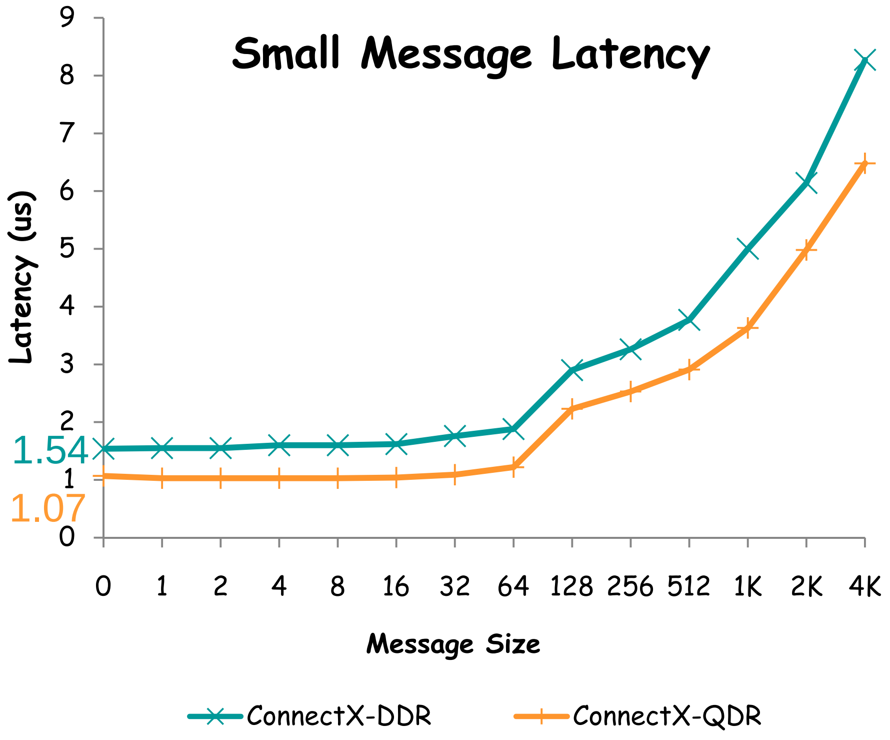
New Features of MVAPICH2 1.2

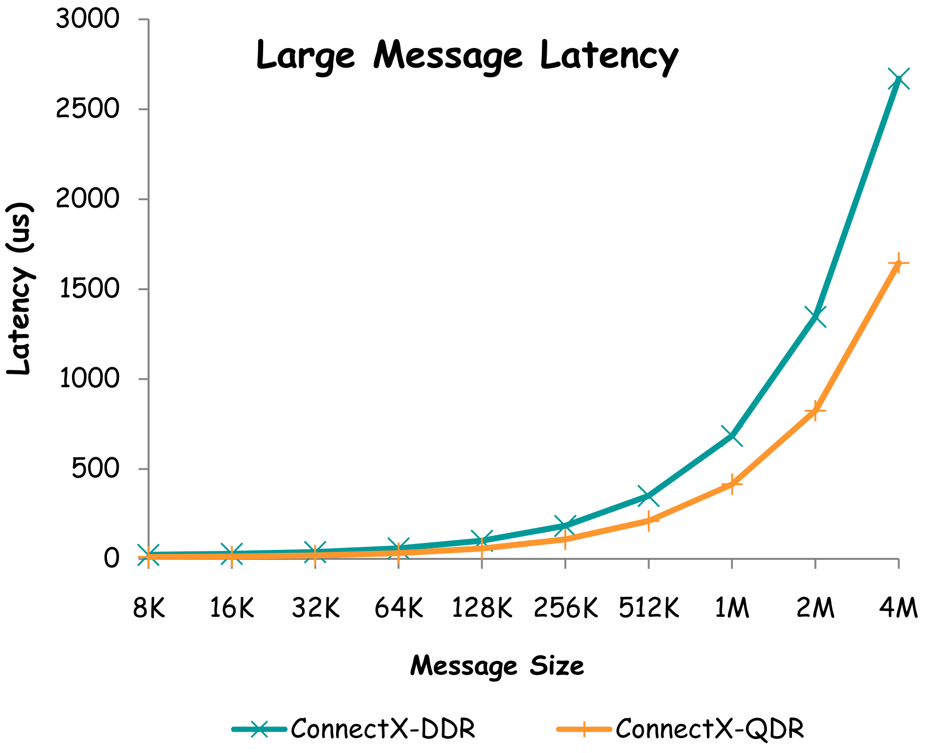
- Released on 11/11/08
- Part of OFED 1.4
- OpenFabrics-Gen2
  - Scalable startup with mpirun_rsh (no need for MPD)
  - Checkpoint-restart support with intra-node shared memory
  - Enhanced Processor Affinity using PLPA
  - Efficient intra-node shared memory support for diskless clusters
  - Scalable direct one-sided support
  - Shared memory-based MPI-Bcast and optimized collectives (including MPI-Alltoall)
  - Full autoconf-based configuration

Presentation Overview

- Overview of MVAPICH/MVAPICH2 Project
- Features of MVAPICH 1.1 and MVAPICH2 1.2
- Sample Performance Numbers
  - Point-to-point (Mellanox, Qlogic & Chelsio)
  - Scalable Startup
  - Hybrid UD-RC/UD-XRC Design
- Upcoming MVAPICH 1.2 and MVAPICH2 1.4 Features and Issues
  - Network Reliability
  - Dynamic Process Management
  - Kernel-based Single copy Intra-node Support
  - MVAPICH2-PSM Support
- Future Plans
- OpenFabrics Requirements
- Conclusions

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

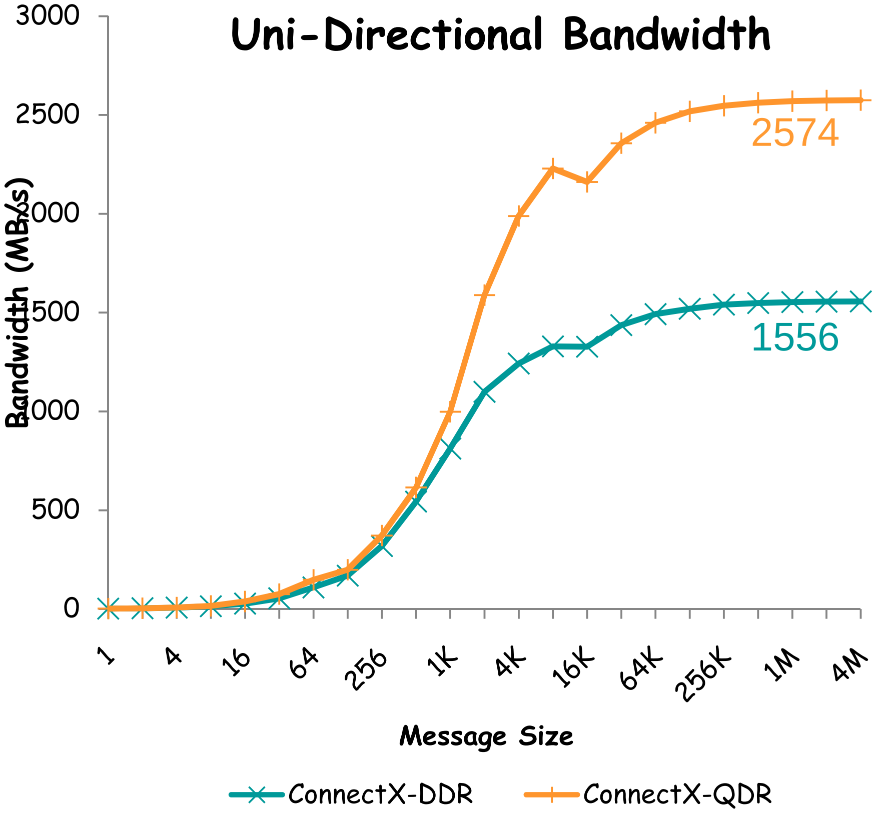


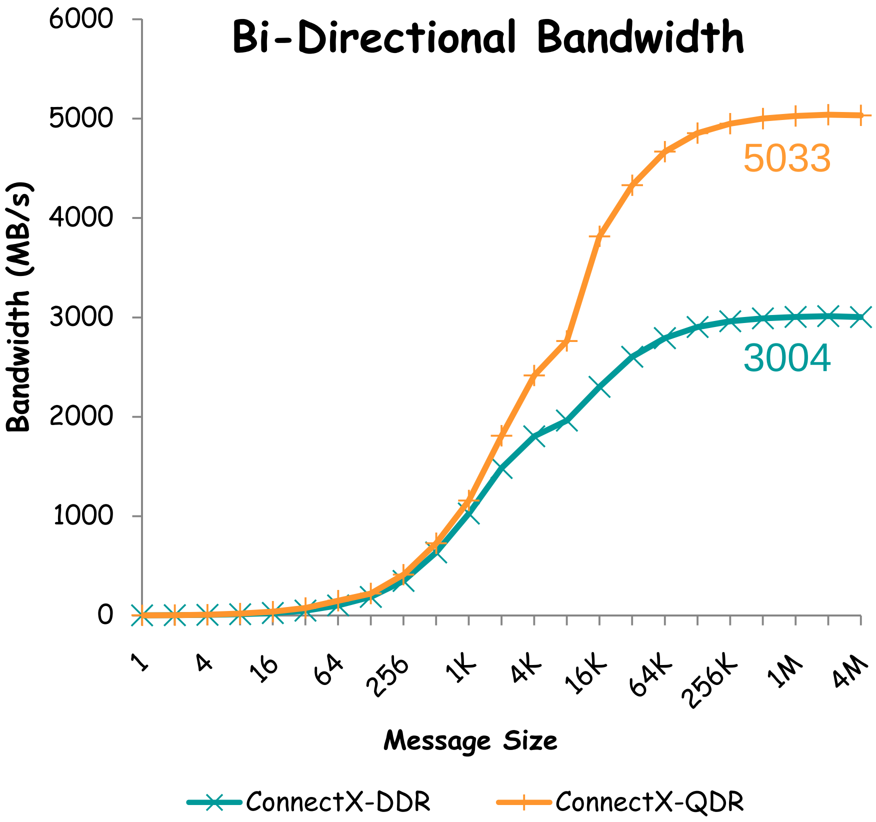


ConnectX-DDR: 2.33 GHz Quad-core (Clovertown) Intel with IB switch

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

MVAPICH Bandwidth:
IBA (Mellanox)

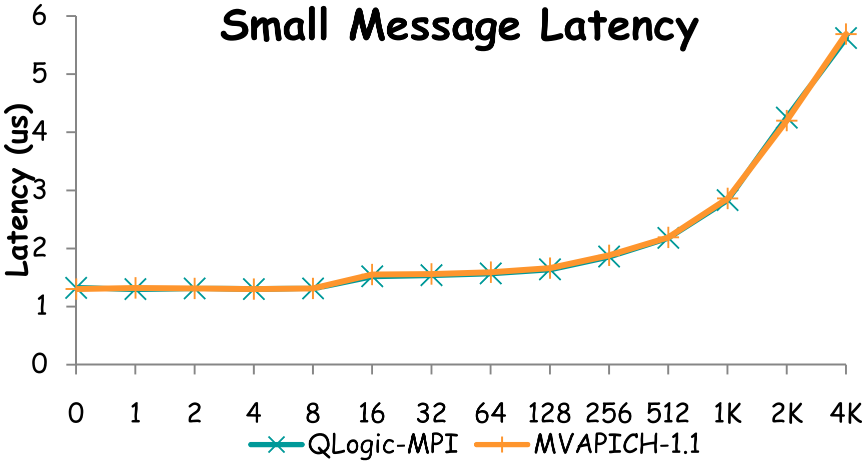


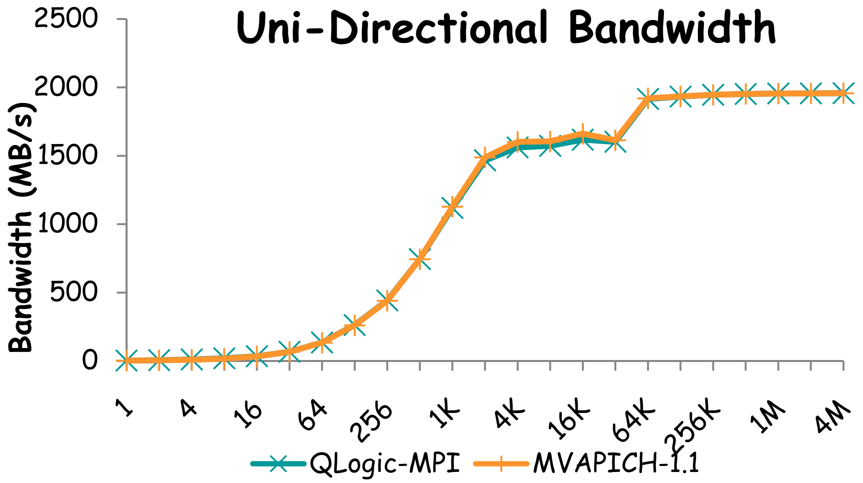


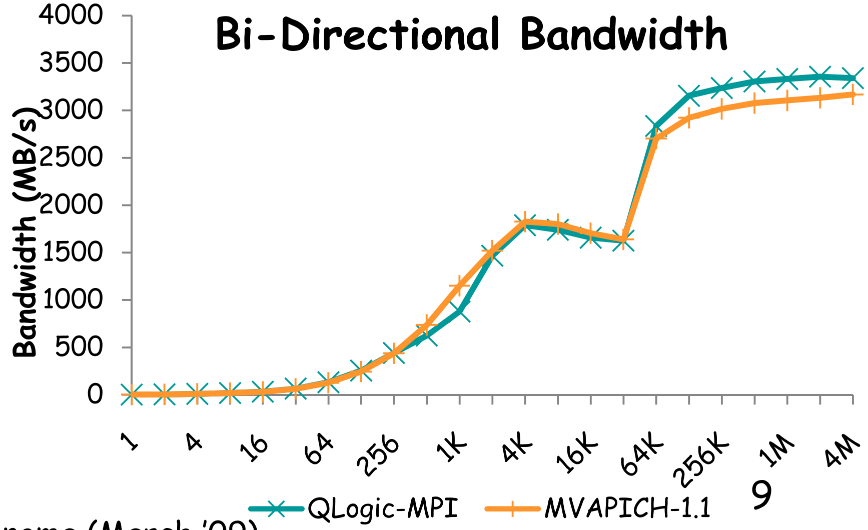
ConnectX-DDR: 2.33 GHz Quad-core (Clovertown) Intel with IB switch

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

MVAPICH-PSM Performance: Two-sided (QLogic-DDR)


Small Message Latency

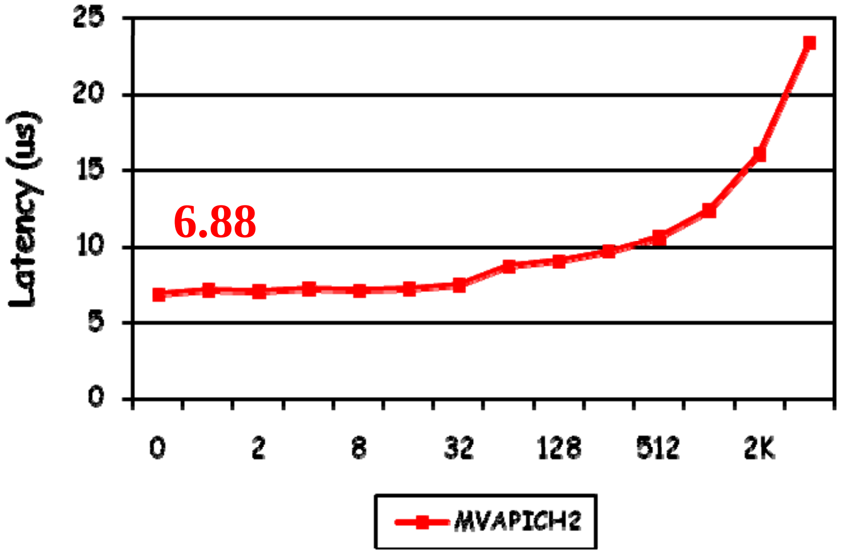

Uni-Directional Bandwidth


Bi-Directional Bandwidth

2.0 GHz Dual-core Opteron with PCIe and IB switch

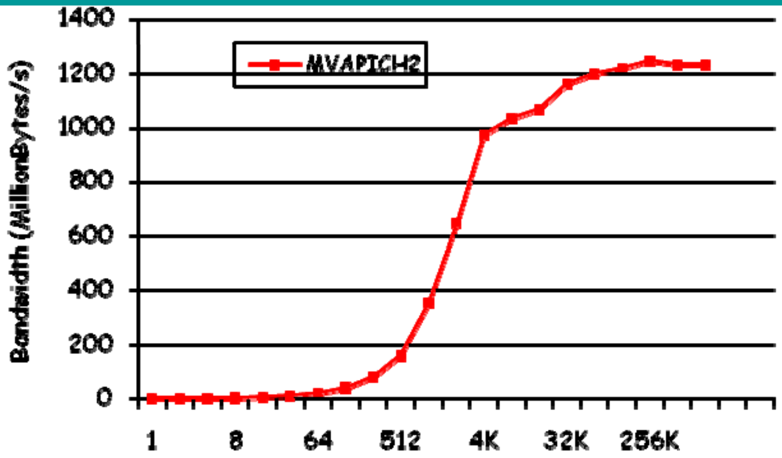
MPI-level Performance: iWARP with Chelsio

2.0 GHz Quad-core Intel
with 10GigE (Fulcrum) switch
NIC Firmware 6.1
OFED 1.3.1



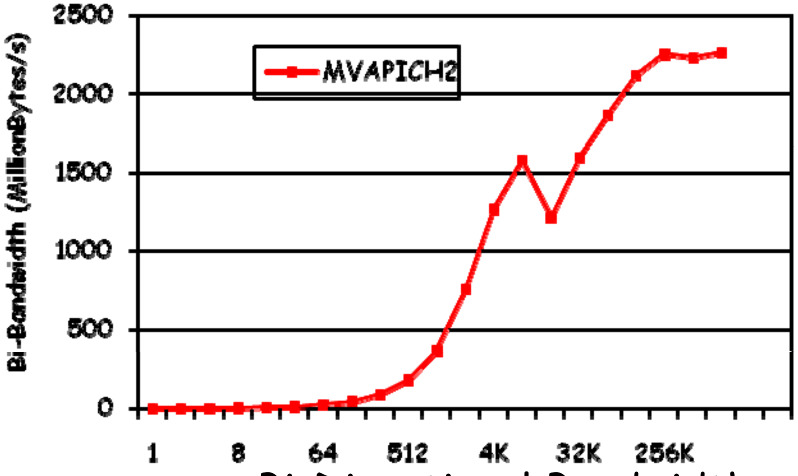
6.88

MVAPICH2 gives a latency of about 6.88us



1232

Uni-Directional Bandwidth

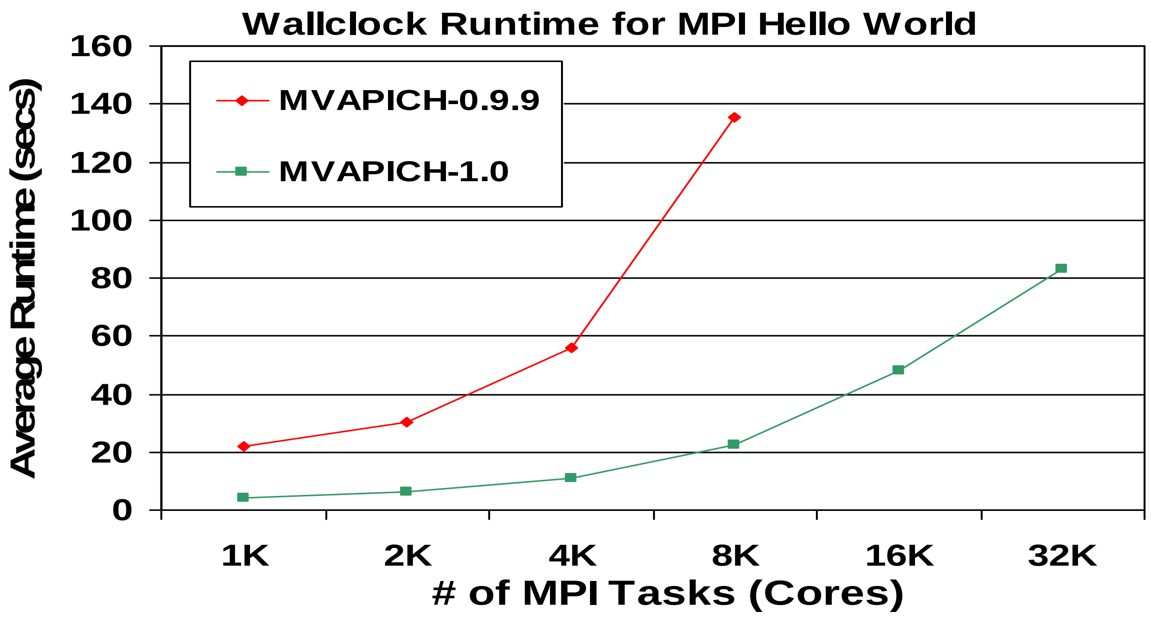


2261

Bi-Directional Bandwidth

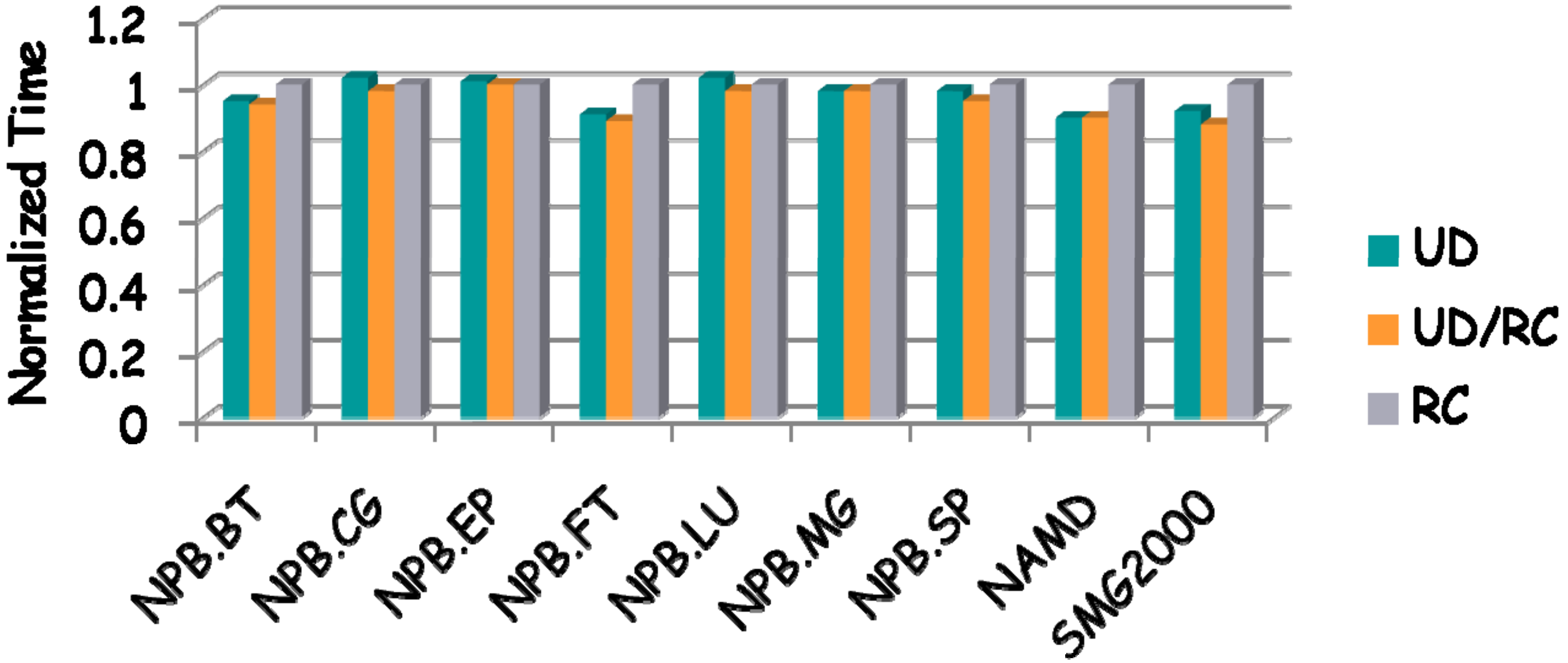
Scalable Startup

- An enhanced mpirun_rsh framework
- Available since MVAPICH 1.0 and MVAPICH2 1.2
- Enhanced further in MVAPICH2 1.4



Courtesy TACC

Impact of Hybrid RC/UD 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," IPDPS '08

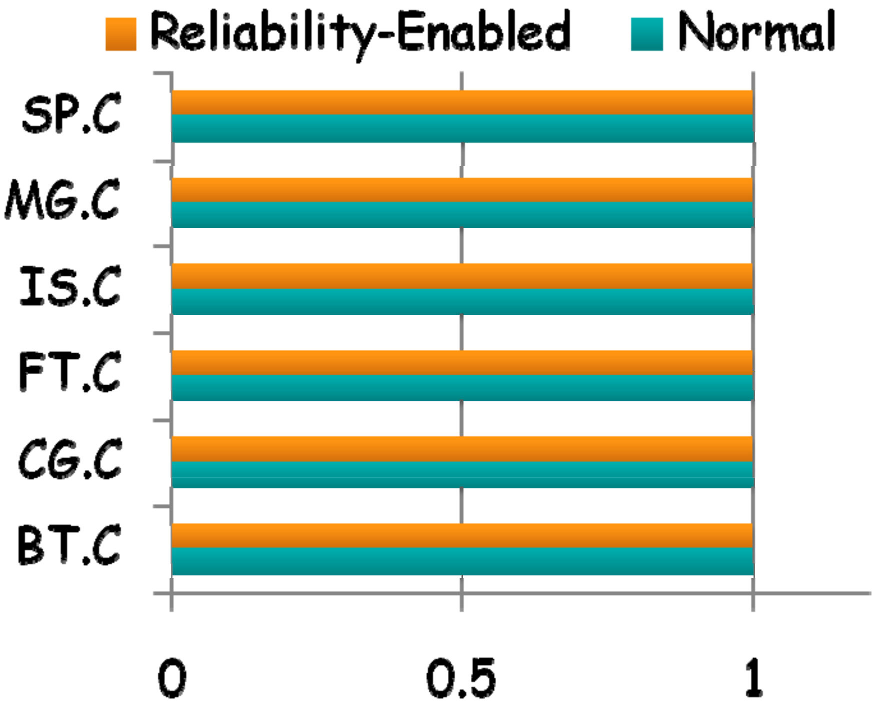
Available in the latest MVAPICH 1.1 Release (Gen2-Hybrid)

Presentation Overview

- Overview of MVAPICH/MVAPICH2 Project
- Features of MVAPICH 1.1 and MVAPICH2 1.2
- Sample Performance Numbers
  - Point-to-point (Mellanox, Qlogic & Chelsio)
  - Scalable Startup
  - Hybrid UD-RC/UD-XRC Design
- Upcoming MVAPICH 1.2 and MVAPICH2 1.4 Features and Issues
  - Network Reliability
  - Dynamic Process Management
  - Kernel-based Single copy Intra-node Support
  - MVAPICH2-PSM Support
- Future Plans
- OpenFabrics Requirements
- Conclusions

Network Reliability

- Protection against various network failures
  - Switch reboot/failure
  - HCA failure
- Option to stall instead of abort job while component fixed
- No significant performance change
- Designed and developed with Mellanox
- Will be available in MVAPICH 1.2
- Will be part of OFED 1.5

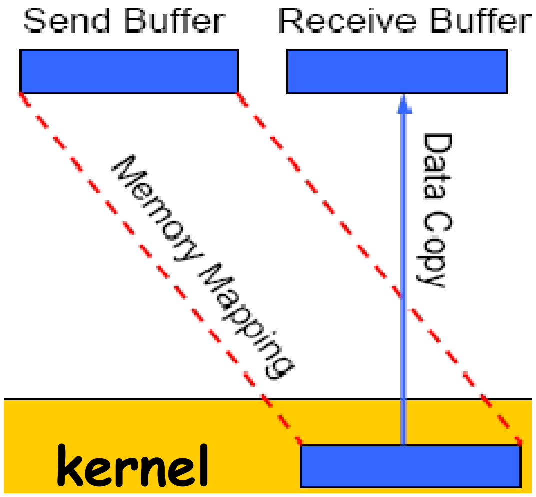


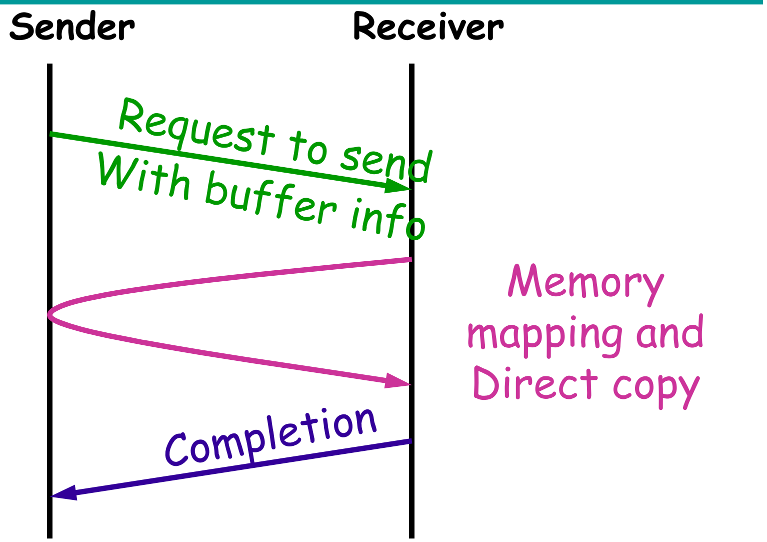
Normalized Time
(256 cores, DDR HCAs)

Features of Upcoming MVAPICH2 1.4

- Multiple New Features
  - MPI 2.1 compliant
  - Reducing job startup time further with mpirun_rsh
  - Checkpoint-restart support with mpirun_rsh (no need to use MPD)
  - Dynamic Process Management
  - Kernel-based Intra-node Shared Memory Communication
  - Support for Qlogic-PSM
  - Enhanced optimized collectives
- RC1 will be released in a few weeks
- Will be part of OFED 1.5

Kernel-Based Single Copy Intra-node Support in MVAPICH2: MVAPICH2-LiMIC2

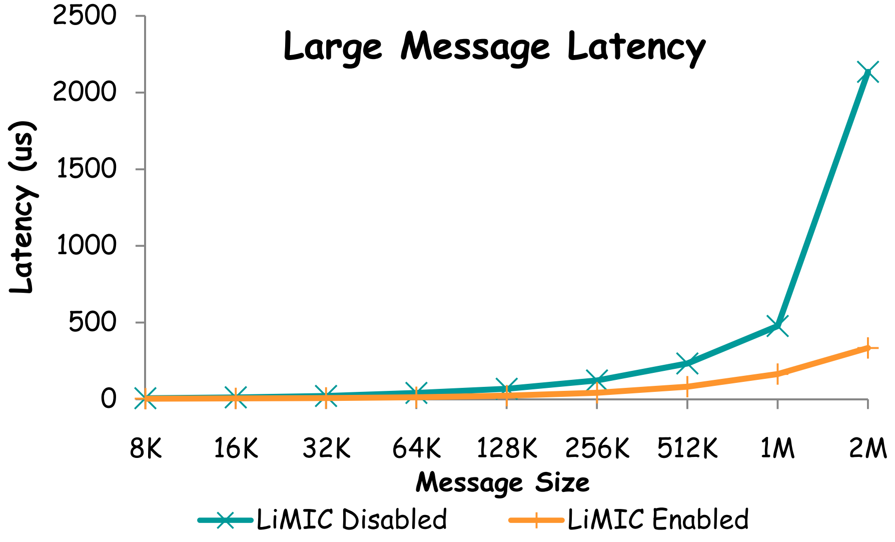


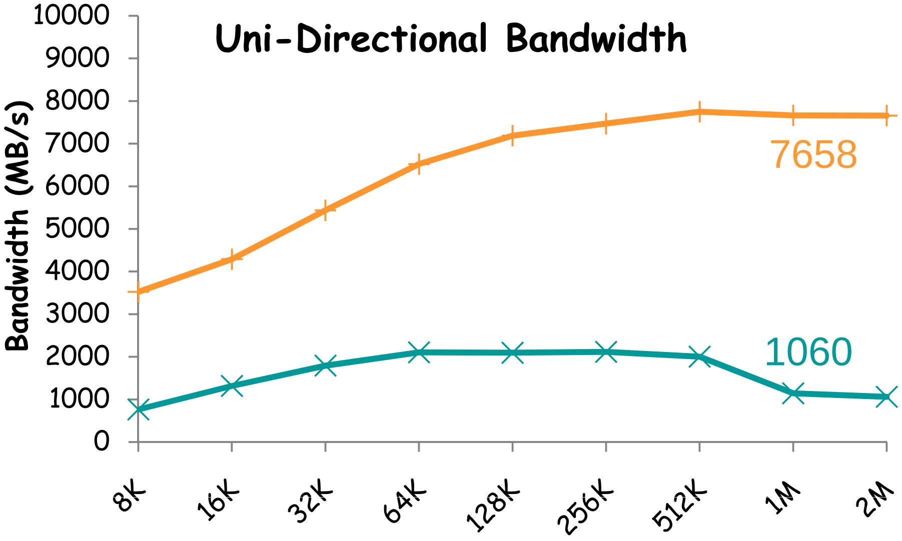


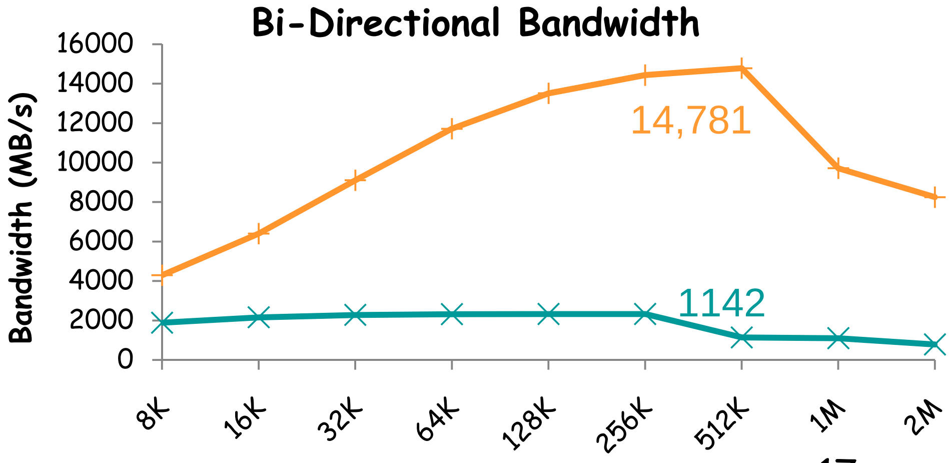
- LiMIC2 (Linux kernel module for MPI Intra-node Communication)
  - Light weight communication primitives
  - Implements memory mapping and data movement primitives
  - Depends on the MPI library for message matching etc
  - Designed for Linux kernel 2.6

- MVAPICH2-LiMIC2
  - Uses LiMIC2 for intra-node communication for medium and large messages
  - Rendezvous protocol
- Benefits
  - One copy
  - Reducing cache pollution
  - Reducing memory usage

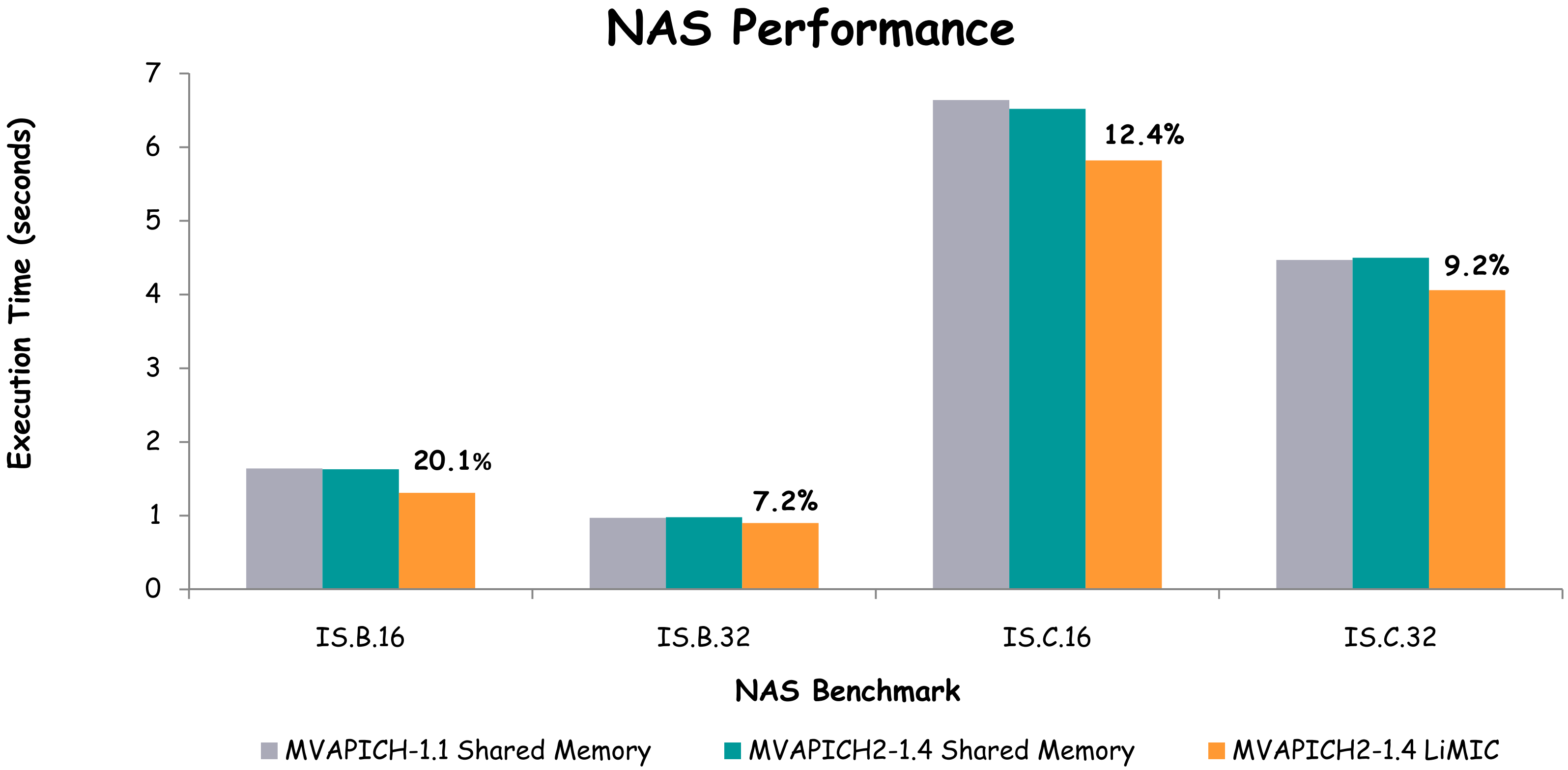
MVAPICH2-LiMIC2 Performance: Two Sided Communication





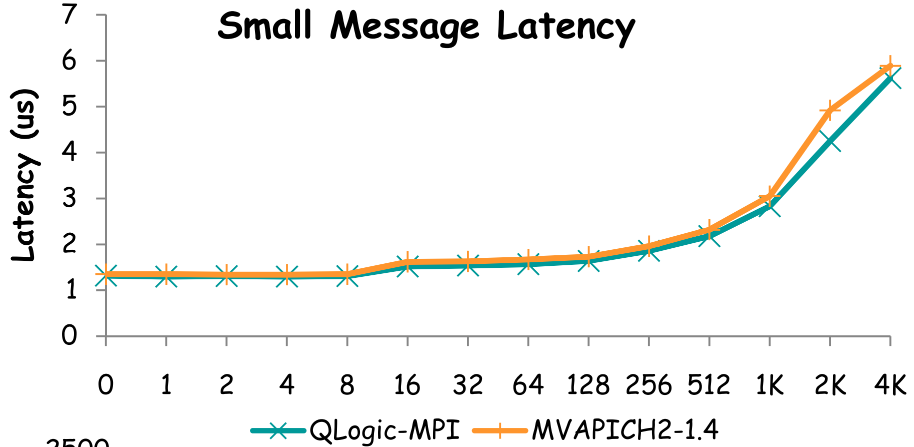


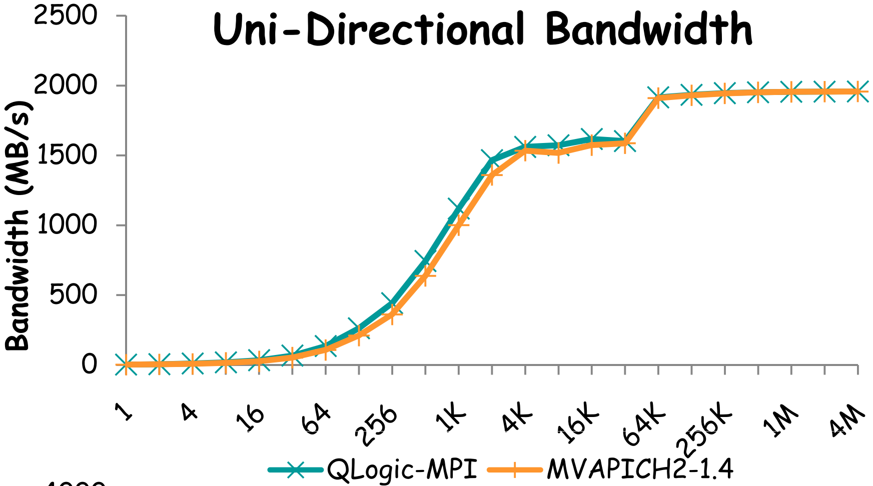
MVAPICH2-LiMIC2: NAS Performance

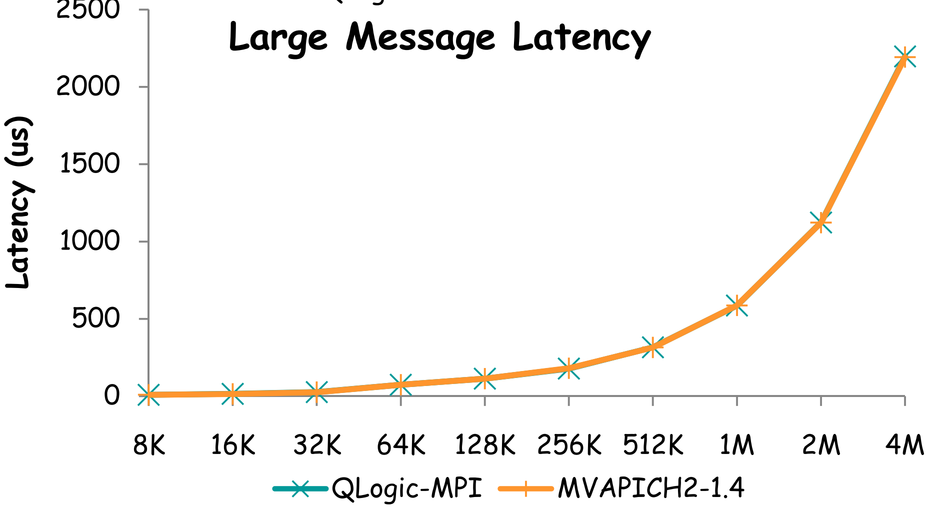


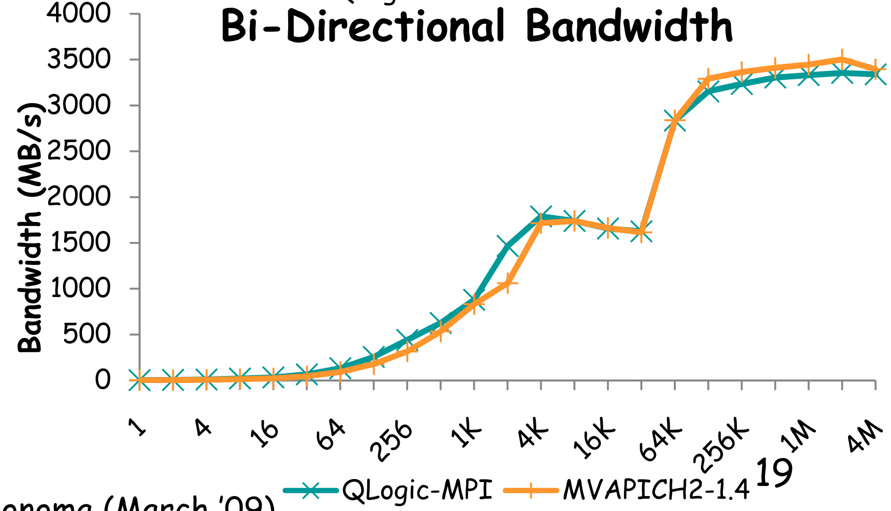
ConnectX-DDR: 2.33 GHz Quad-core (Clovertown) Intel with IB switch

MVAPICH2-PSM Performance: Two-Sided (QLogic-DDR)


Small Message Latency


Uni-Directional Bandwidth


Large Message Latency


Bi-Directional Bandwidth

Presentation Overview

- Overview of MVAPICH/MVAPICH2 Project
- Features of MVAPICH 1.1 and MVAPICH2 1.2
- Sample Performance Numbers
  - Point-to-point (Mellanox, Qlogic & Chelsio)
  - Scalable Startup
  - Hybrid UD-RC/UD-XRC Design
- Upcoming MVAPICH 1.2 and MVAPICH2 1.4 Features and Issues
  - Network Reliability
  - Dynamic Process Management
  - Kernel-based Single copy Intra-node Support
  - MVAPICH2-PSM Support
- Future Plans
- OpenFabrics Requirements
- Conclusions

Future Plans

- MPI-level QoS Support
  - Intra-MPI and Inter-MPI
- Incorporating MPICH2 Nemesis-based Design in MVAPICH2
  - Core-to-core MPI-level latency of 240nsec
- High Performance and Scalable Collectives based on new HCA features
  - Reliable Multicast
  - Offload
- Topology-aware Collectives
- Automatic Tuning of Pt-to-point and Collectives
- Job-Pause Resume Framework for Fault Tolerance

Requirements from OpenFabrics

- Fast Memory Registration
  - User Level
- Reliable Datagram
- Adding additional features to UD
  - RDMA with UD
  - Offloaded segmentation
- 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 870 organizations
- New features for scalability, high performance and fault tolerance support are aimed to deploy large-scale clusters (~100K) cores in the near future

Web Pointers

MVAPICH

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

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