



# MVAPICH2 Project: Latest Status and Future Plans

Presentation at MPICH2 BOF  
(Nov. '08)

by

Dhabaleswar K. (DK) Panda  
Department of Computer Science and Engg.  
The Ohio State University

E-mail: [panda@cse.ohio-state.edu](mailto:panda@cse.ohio-state.edu)  
<http://www.cse.ohio-state.edu/~panda>



# Overview

- Working closely with ANL on MVAPICH and MVAPICH2 projects for the last eight years
  - MVAPICH is a derivative of MPICH with MPI-1 semantics
  - MVAPICH2 is a derivative of MPICH2 with MPI-2 semantics
- Focus is on high performance implementation on emerging interconnects
  - InfiniBand
  - iWARP/10GigE
- Also supports a uDAPL device to work with any other interconnect supporting uDAPL interface
  - uDAPL over InfiniBand (Open Fabrics-Gen2 and Solaris)

## MVAPICH/MVAPICH2 Open-Source Software Distribution

- Both versions are available from OSU directly
  - <http://mvapich.cse.ohio-state.edu>
  - Source-tree available from public SVN
  - also available as tarballs
  - public mailing list (mvapich-discuss) with archives
- Also available with Open Fabrics Enterprise Distribution (OFED)
  - <http://www.openfabrics.org>
  - public mailing lists (openfabrics-general and openfabrics-ewg) with archives
- Also available from server vendors, interconnect vendors and Linux distributors

## MVAPICH/MVAPICH2 Open-Source Software Distribution (Cont'd)

- Latest releases
  - MVAPICH2 1.2
  - MVAPICH 1.1
- Included in the latest OFED 1.4
  - RC5 is out, [final release will be done soon](#)
- Used by more than 800 organizations in 42 countries
- More than 25,000 downloads from OSU site directly
- Empowering many TOP500 clusters
  - 6<sup>th</sup> ranked 62,976-core cluster (Ranger) at TACC



## MVAPICH2 Features



- Latest version is MVAPICH2 1.2
  - Based on MPICH2 1.0.7
- Unified design over Open Fabrics stack to support
  - InfiniBand
  - 10GigE/iWARP
- Scalable and robust daemon-less job startup with the new mpirun\_rsh framework
- High performance and scalable implementations
  - RDMA write
  - RDMA Read
- Multi-threading Support
- Optimized support for
  - two-sided operations
  - one-sided operations (Put, Get and Accumulate)
    - supports both active and passive synchronization



## MVAPICH2 Features (Cont'd)



- Integrated multi-rail support
  - Multiple adapters, queue pairs and multiple ports/adapters
- Efficient shared-memory point-to-point communication
  - For emerging multi-core architecture
- Optimized collectives
  - including optimizations for multi-core platforms
- Two different communication progress
  - Polling and Blocking
- On-demand connection management for large clusters
- Multiple solutions for Fault-Tolerance
  - Network-level FT support with Automatic Path Migration (APM)
  - Process-level FT with systems-level checkpoint-restart with shared-memory and shared-memory collectives
    - Uses BLCR
    - Two modes: automatic and user initiated
- Hot-spot avoidance mechanism for alleviating network congestion in large clusters
- Totalview Debugger support

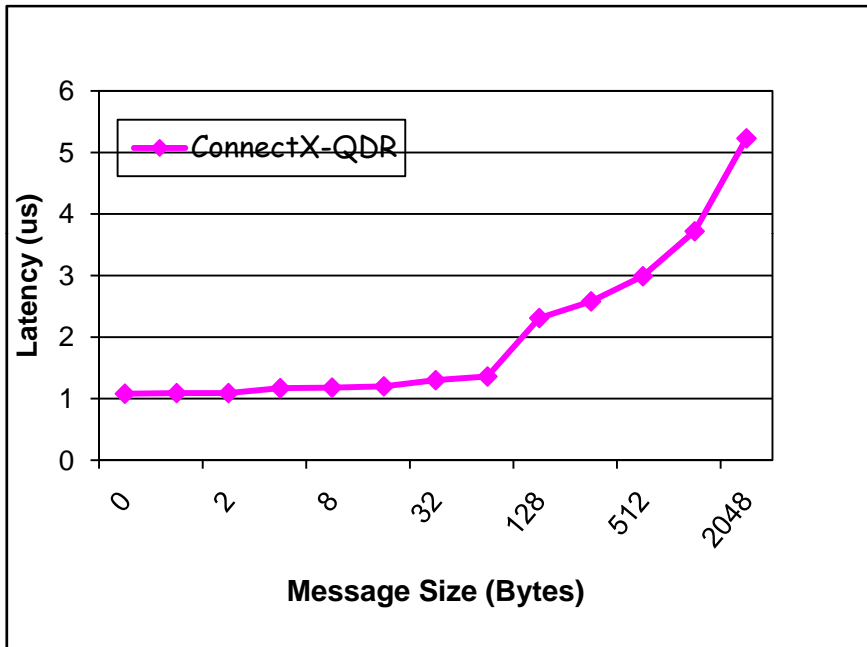
•  
•

## Support for Multiple Interfaces/Adapters

- OpenFabrics/Gen2-IB
  - All IB adapters (SDR, DDR and QDR) supporting Gen2
  - ConnectX
- uDAPL
  - Linux-IB
  - Solaris-IB
- OpenFabrics/Gen2-iWARP
  - Chelsio 10GigE
- Support for Qlogic/PSM
  - Will be available soon

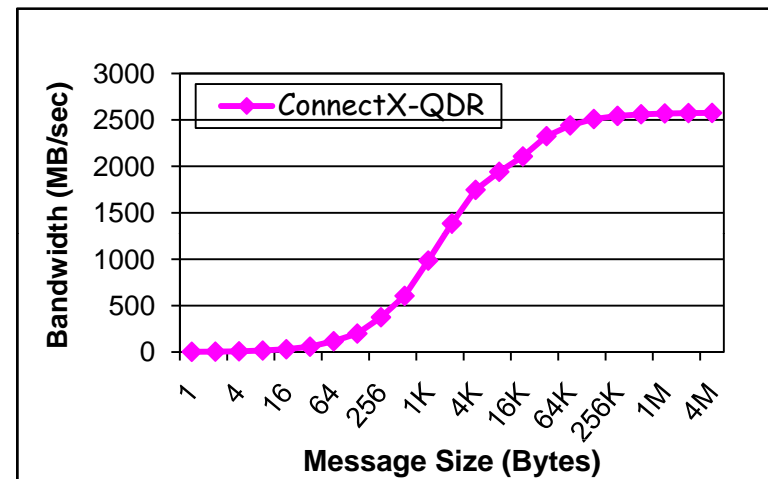
# MVAPICH2-1.2 Performance with MPI-Level Two-Sided Communication - IB Mellanox QDR

1.08

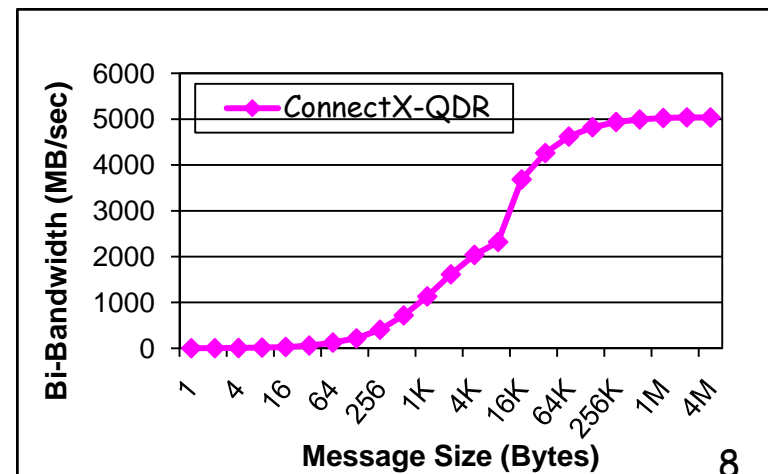


- Single port results only (EM64T, PCIE-Gen2)

Results for other platforms at <http://mvapich.cse.ohio-state.edu>



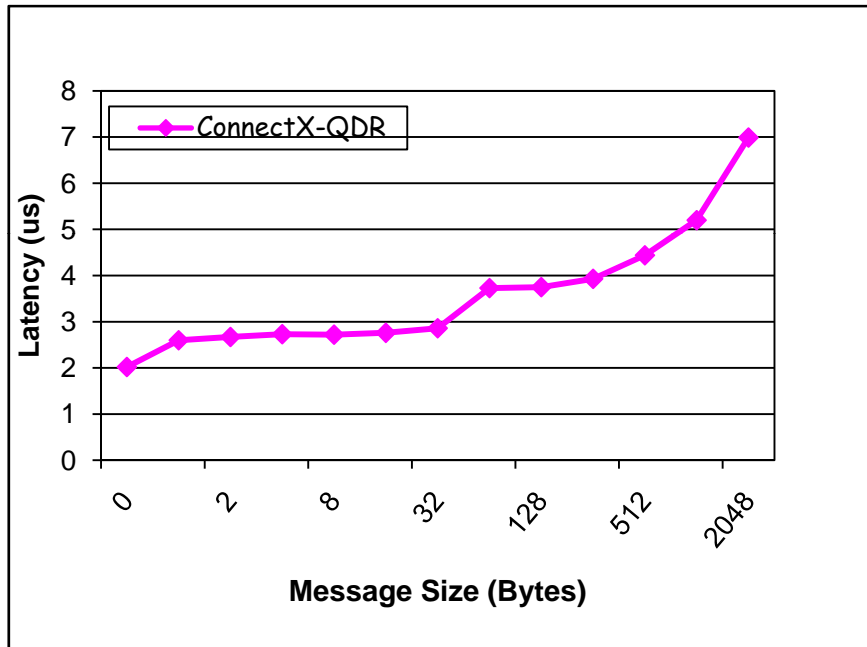
2576



5040

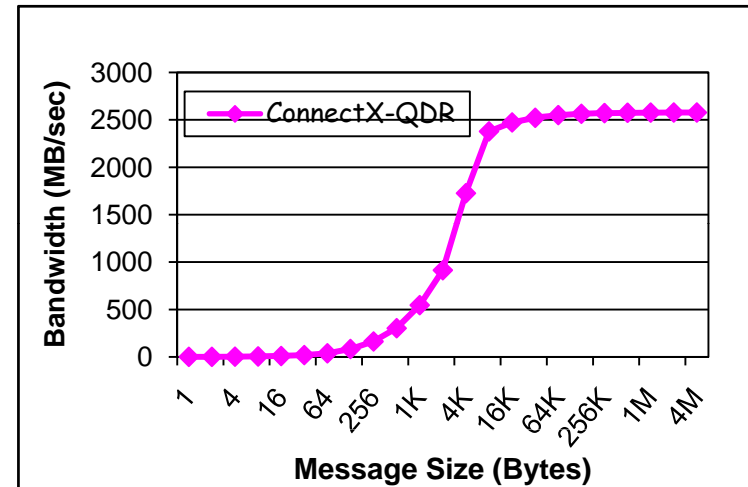
# MVAPICH2-1.2 Performance with MPI One Sided Put (Active Target) - IB Mellanox QDR

2.02

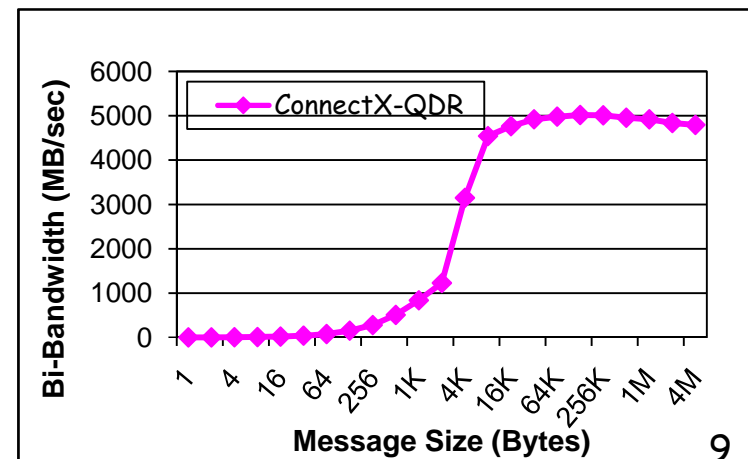


- Single port results only (EM64T, PCIE-Gen2)

Results for other platforms at <http://mvapich.cse.ohio-state.edu>

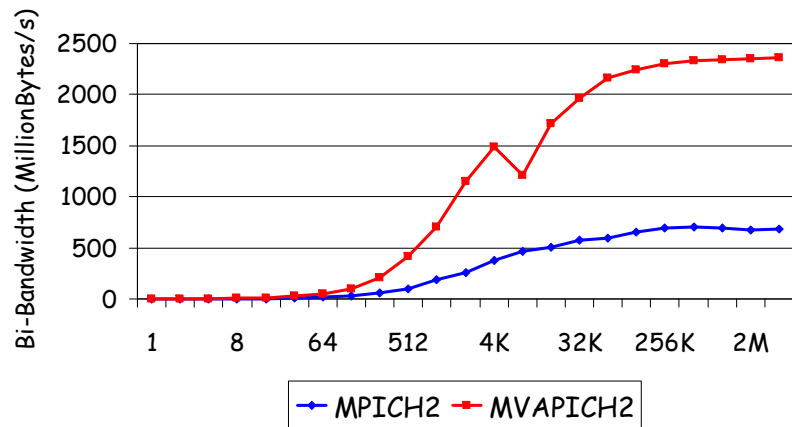
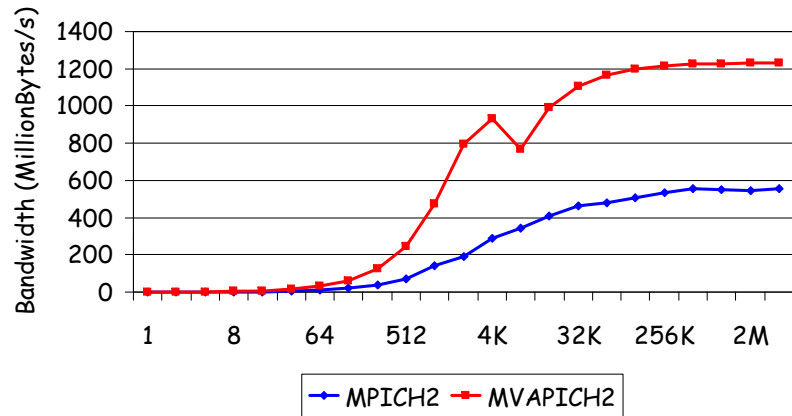
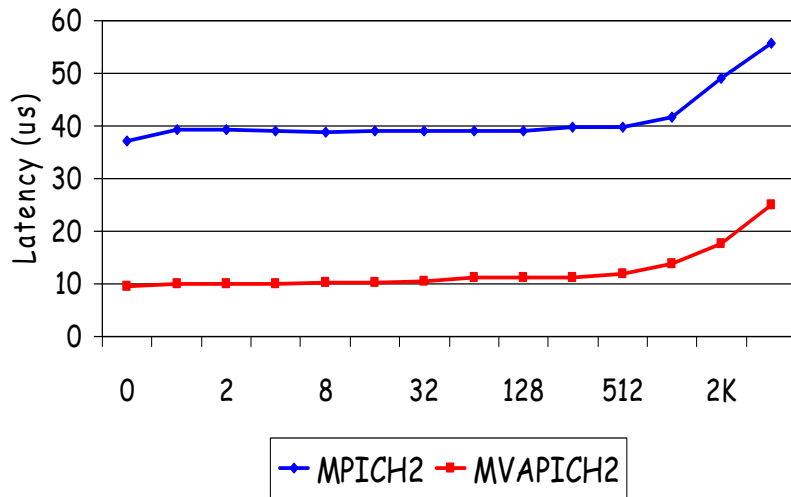


2576

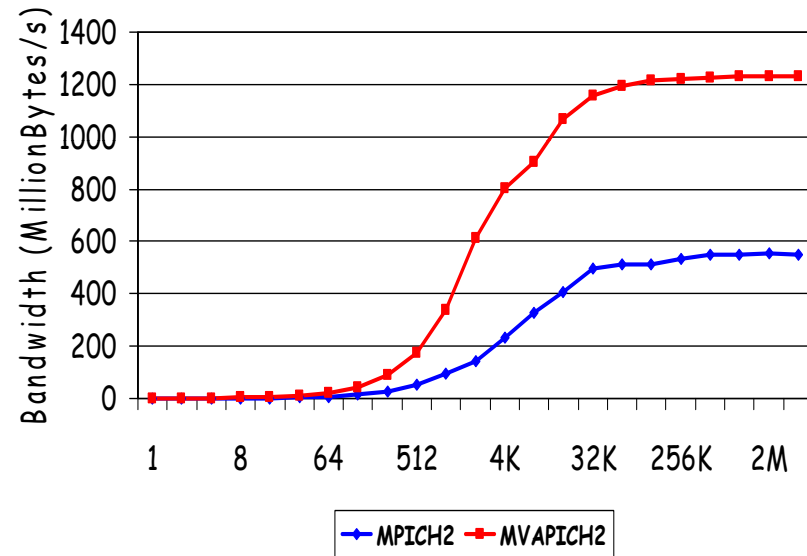
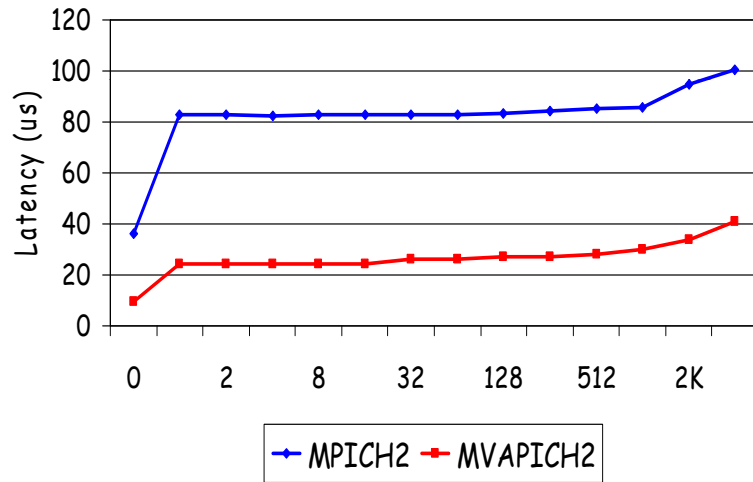


5018

# MVAPICH2 Performance with MPI-Level One-Sided Put Communication - 10GigE/iWARP Chelsio



# MVAPICH2 Performance with MPI-Level One-Sided Get Communication - 10GigE/iWARP Chelsio



# Continuing and Future Work

- Will continue to work closely with ANL to incorporate the latest features and updates from the MPICH2 stack
  - MPICH2 1.08 and the new 1.1 series
    - Nemesis support
    - Dynamically loadable network modules
  - Support for Qlogic/PSM
  - A bunch of additional features
    - IB-XRC, Dynamic Process Management, LiMIC2, etc.

•  
•  
•

# Web Pointers



**MVAPICH**

MVAPICH Web Page

<http://mvapich.cse.ohio-state.edu/>

E-mail: [panda@cse.ohio-state.edu](mailto:panda@cse.ohio-state.edu)