MVAPICH/MVAPICH2 Update, Future Plans and Path Towards Exascale

MPI Panel at Open Fabrics Sonoma (March 2010)

by

Dhabaleswar K. (DK) Panda

The Ohio State University

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

http://www.cse.ohio-state.edu/~panda

MVAPICH/MVAPICH2 Software

- High Performance MPI Library for IB, 10GE/iWARP and RoCEE
- MVAPICH (MPI-1) and MVAPICH2 (MPI-2)
- First open-source version was demonstrated at SC '02
- Used by more than 1,075 organizations in 56 countries
  - Registered in a voluntary manner
- More than 38,000 downloads from OSU site directly
- Empowering many TOP500 clusters
  - 5th ranked 71,680-core cluster (Tianhe-1) in China
  - 9th ranked 62,976-core cluster (Ranger) at TACC
- Available with software stacks of many IB, 10GE/iWARP, RoCEE and server vendors including OFED
- http://mvapich.cse.ohio-state.edu

Latest Releases

- MVAPICH 1.2 (released 01/29/10)
  - Network Fault Resilience (NFR)
  - RoCEE (RDMAoE)
- MVAPICH2 1.4.1 (released 03/12/10)
  - Portable Hardware Locality (hwloc)
  - MPMD support
  - Multi-port support for iWARP
  - Scalability for large process counts with iWARP
  - Ring-based startup for RoCEE (RDMAoE)
- Both versions are available with OFED 1.5.1

MVAPICH/MVAPICH2 – Future Plans

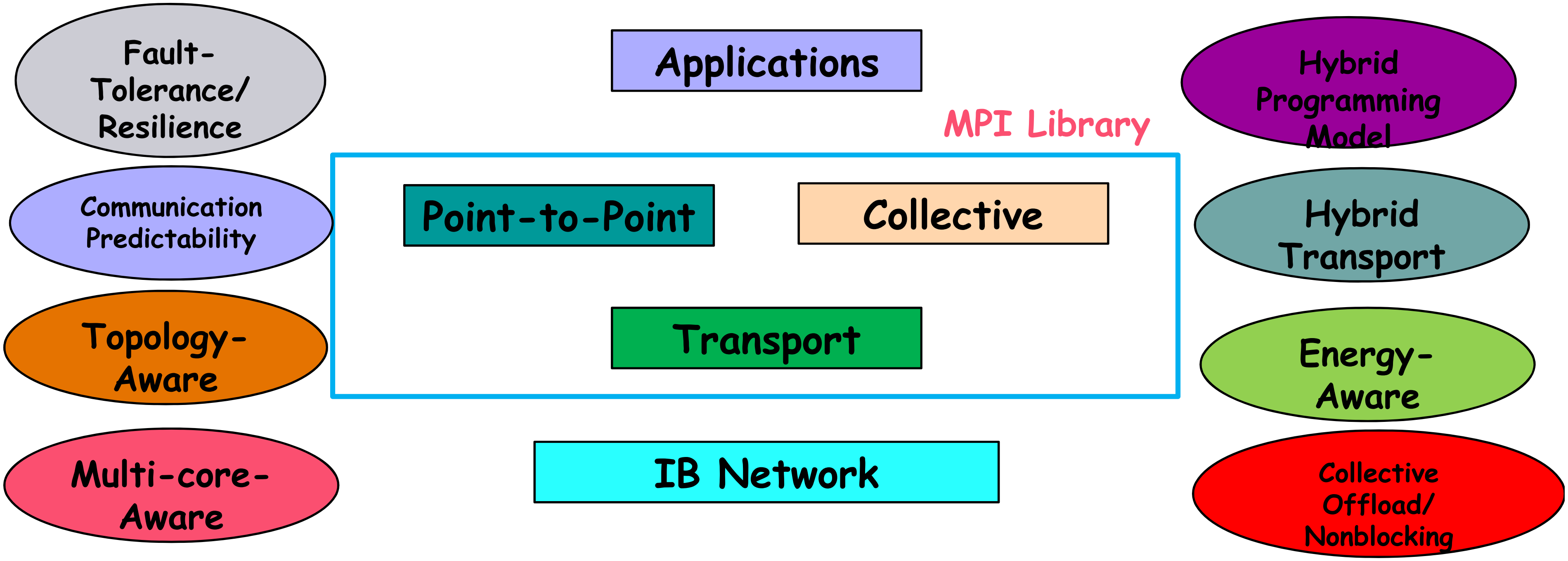
- More focus toward MVAPICH2
- Performance and Memory scalability toward 500K-1M cores
- Taking advantage of Collective Offload framework in ConnectX-2
  - Including non-blocking collectives
- Topology-aware Collectives
- Power-aware Collectives
- Flexible process binding for multi-rails
- Moving MVAPICH2 codebase to the new Nemesis-based design from Argonne (MPICH2 group)
  - Further performance enhancement and scalability for multi-core-based clusters
  - Supporting MPI 2.2 and upcoming 3.0 standard
- Checkpoint-Restart with incremental checkpointing
- QoS-aware I/O and checkpointing
- Job pause-migration-restart framework
- Support for PGAS Models and Languages (UPC, OpenShmem, X10, etc.)

Exascale Systems – How Large will they be?

- Will have ~100 millions of cores
- Take an example of population (from Wikipedia and other Google Searches)
  - Sonoma City ~ 9.8K
  - Sonoma Valley ~ 40K
  - Napa City ~ 74K
  - Napa County ~ 124K
  - Sonoma County ~ 467K
  - San Francisco City ~ 808K
  - Los Angeles City ~ 4M
  - California ~ 37M
  - Texas ~25M
  - New York ~20M
  - Florida ~19M
- Where does the state of InfiniBand cluster lie now?
  - ~100K cores
- Long way to go ....

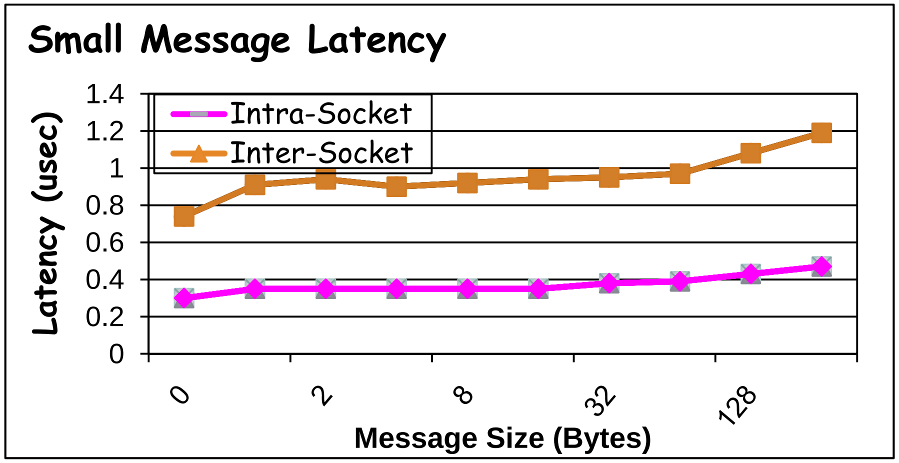
Challenges in Designing MPI at Exascale

- Example from aerospace industry
  - Designing a single-engine Cessna vs. Space Shuttle
  - Basic principles may remain the same
  - Performance, Scaling and Fault-tolerance aspects need to be taken into account in designing each and every component for a Space Shuttle design

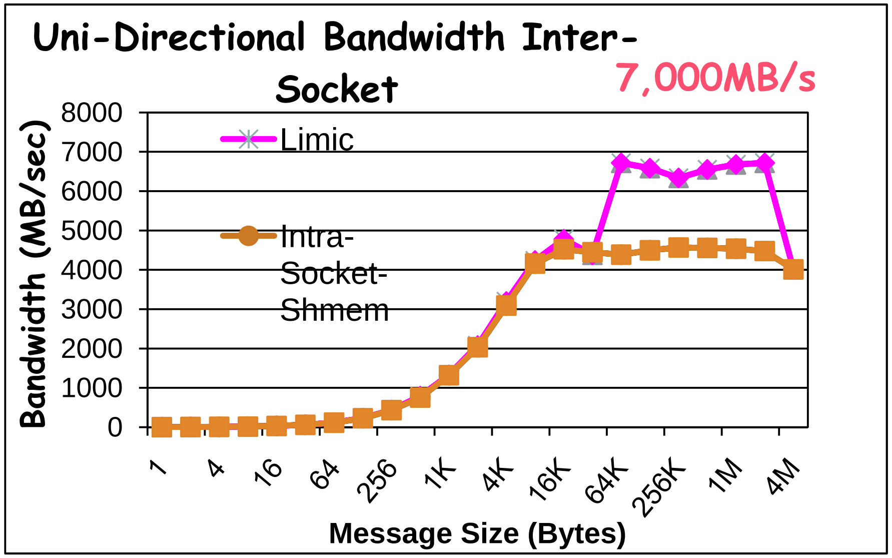


MVAPICH2 Two-Sided Intra-Node Performance (Intel Nehalem)

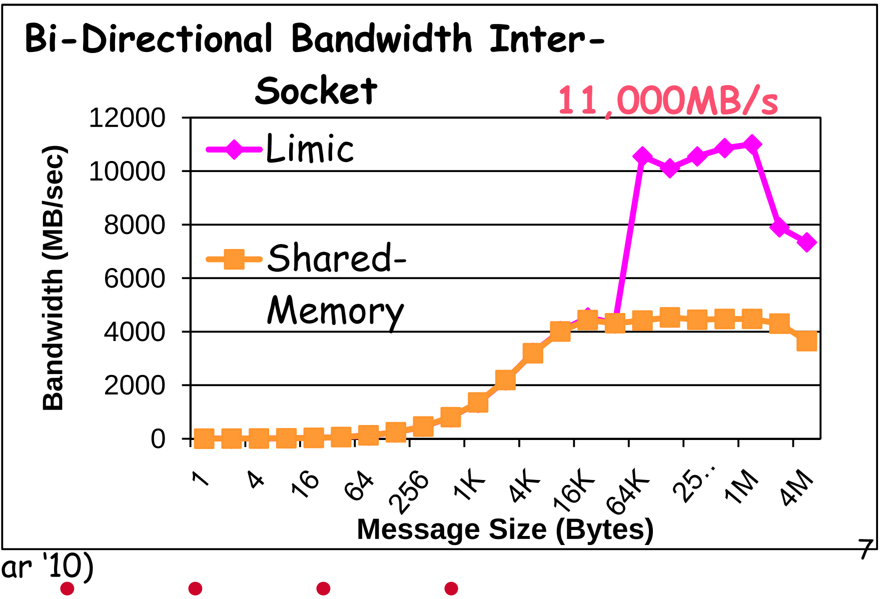
350 nsec


Small Message Latency

Available since MVAPICH2 1.4

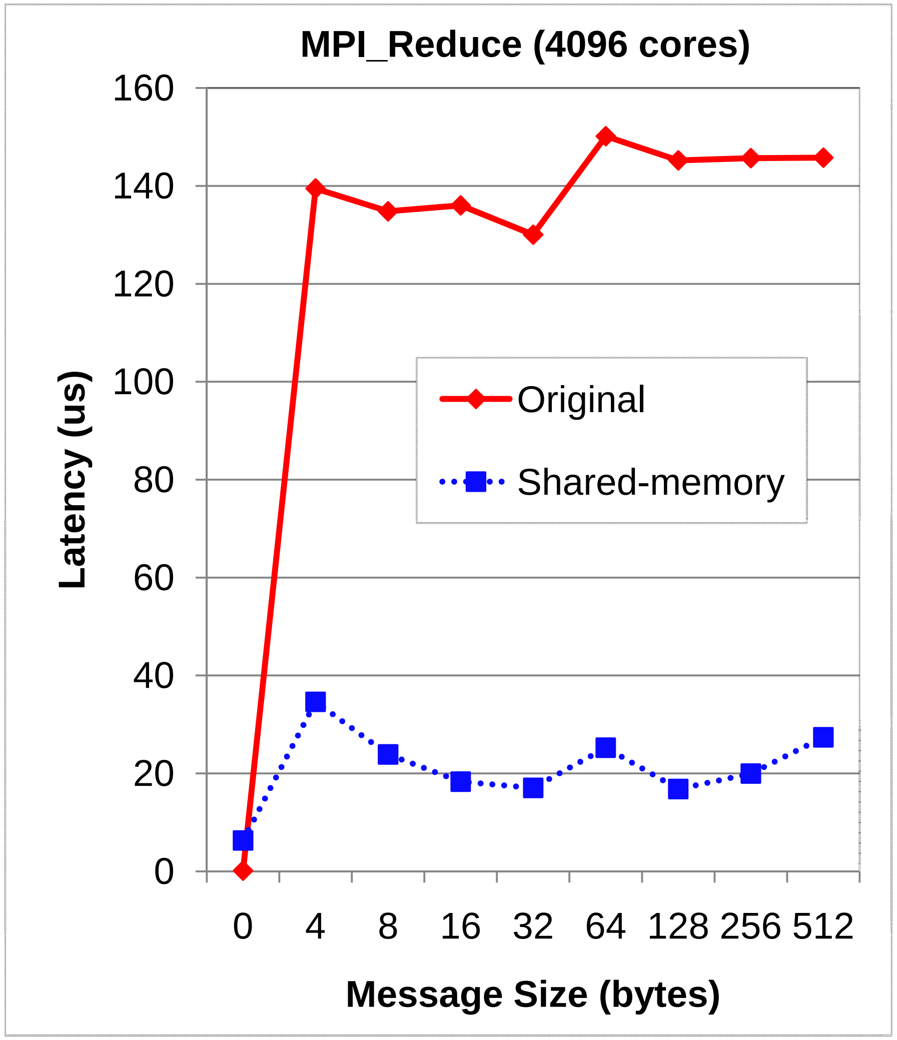

Uni-Directional Bandwidth Inter-Socket

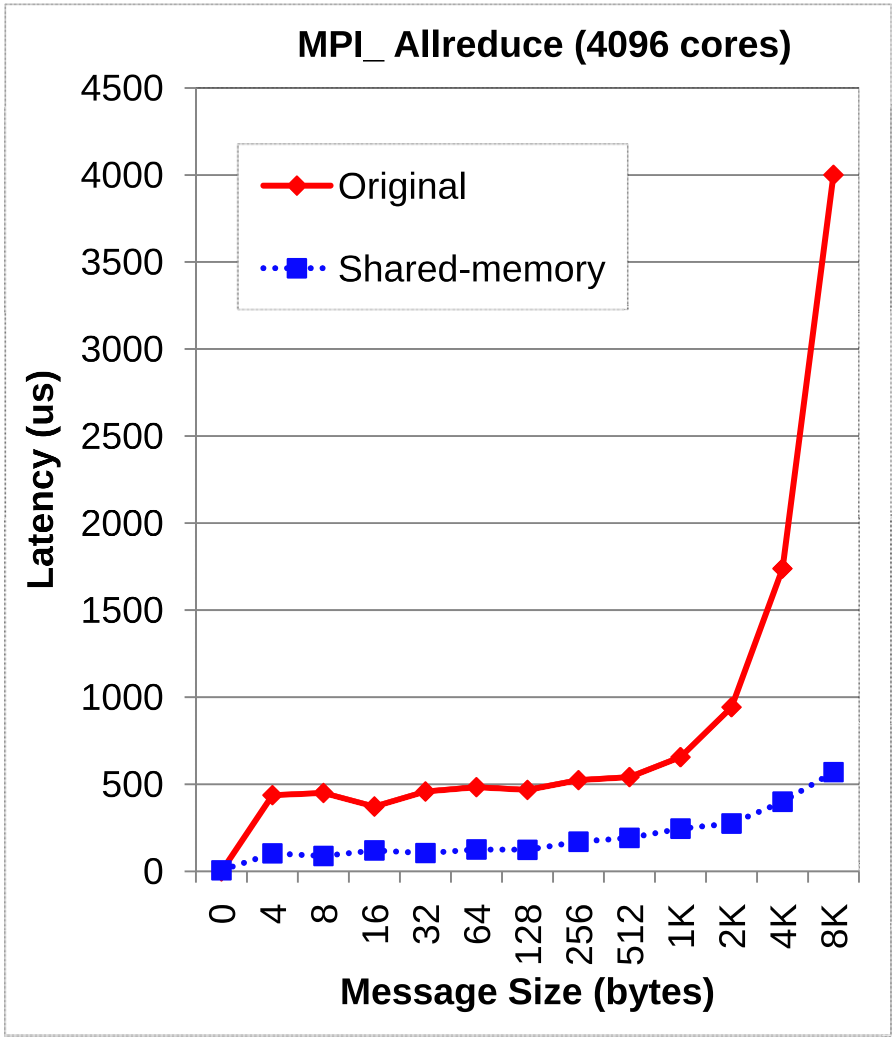
7,000MB/s


Bi-Directional Bandwidth Inter-Socket

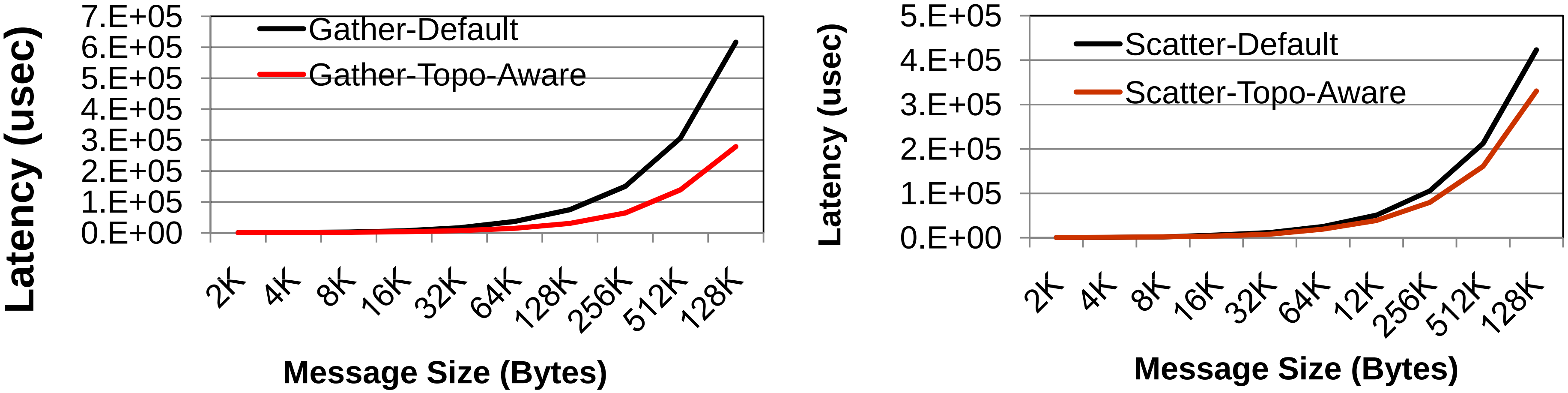
11,000MB/s

Multi-core-Aware Collectives (4K cores on TACC Ranger with MVAPICH2)

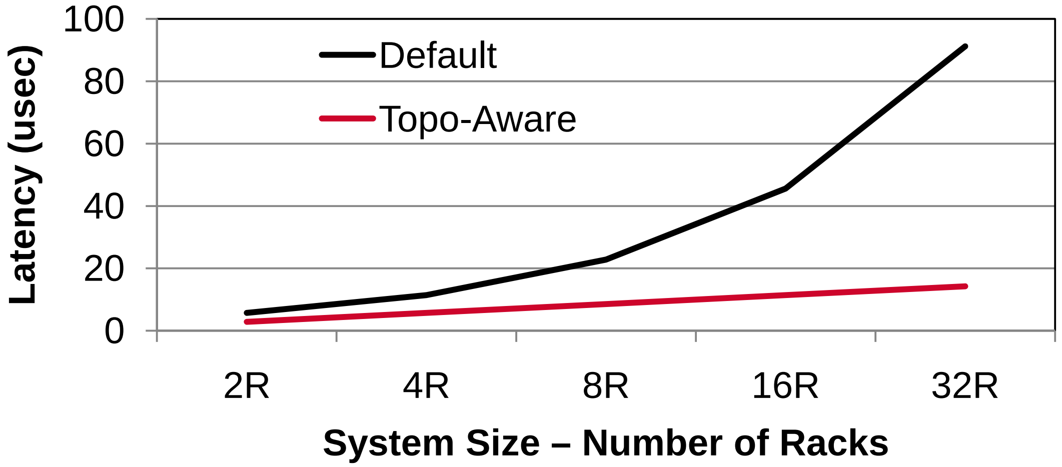




Topology-Aware Collectives



Default (Binomial) Vs Topology-Aware Algorithms with 296 Processes

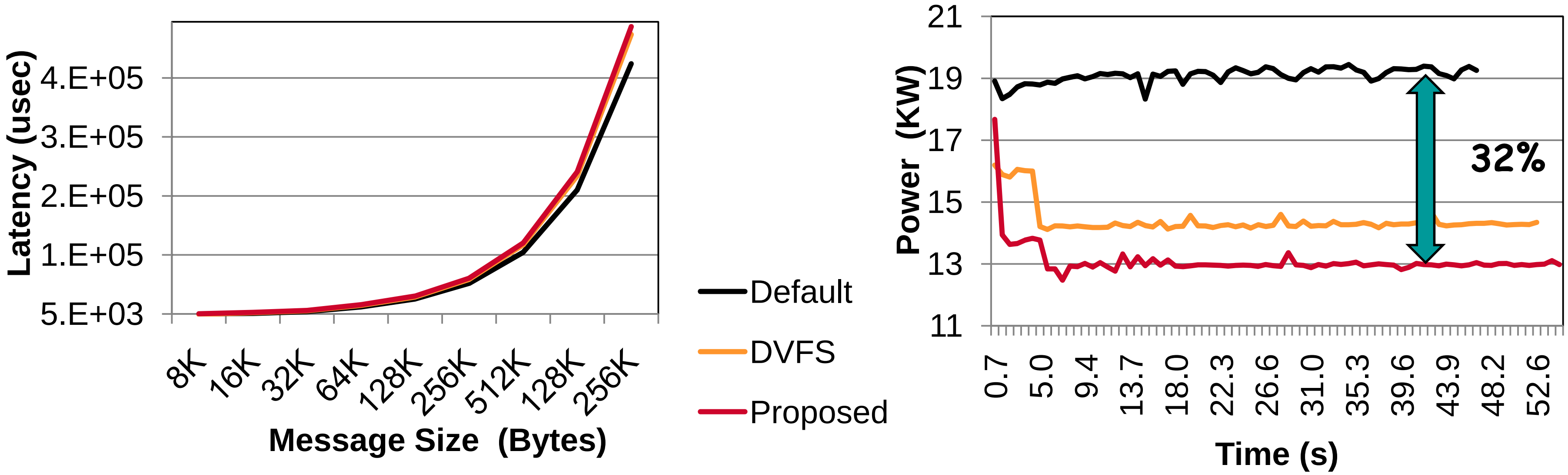


Estimated Latency
Of Default and Topology
Aware Algorithms
for small messages
And Varying System
Sizes

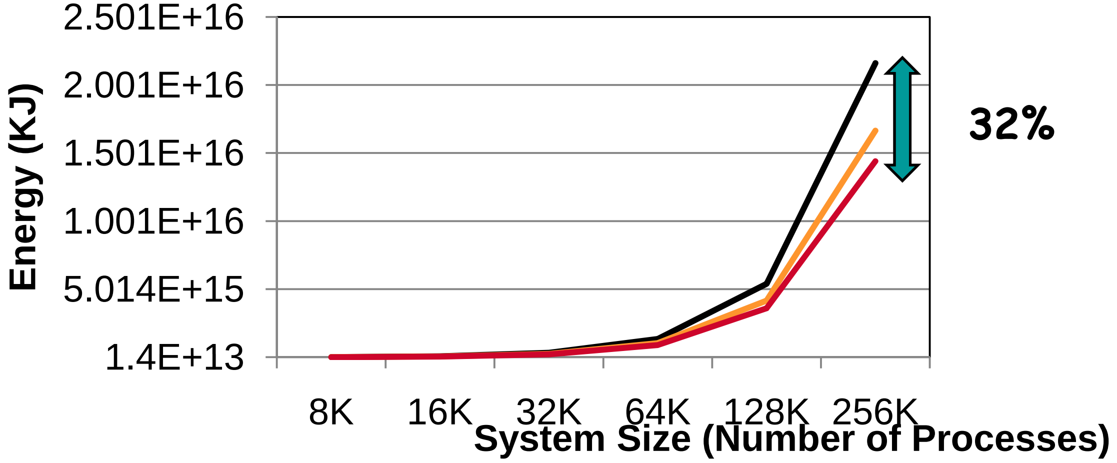
Joint NSF Project between OSU, TACC and SDSC

K. Kandalla, H. Subramoni, A. Vishnu and D. K. Panda, "*Designing Topology-Aware Collective Communication Algorithms for Large Scale Infiniband Clusters: Case Studies with Scatter and Gather*," CAC '10

Power-Aware Collectives



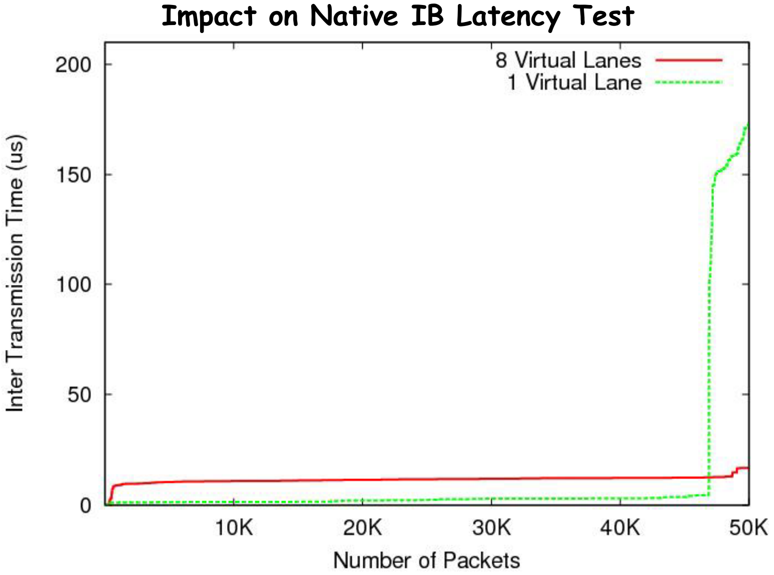
Performance and Power Comparison : MPI_Alltoall with 64 processes on 8 nodes

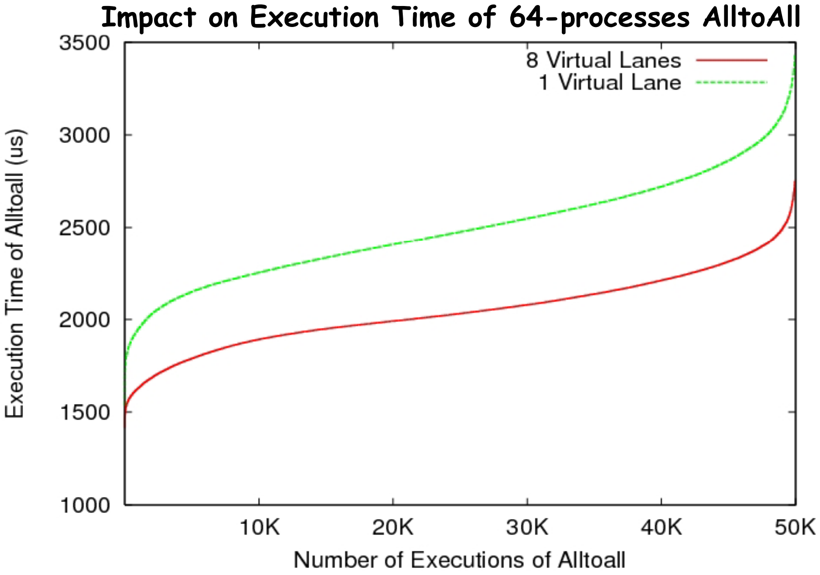


Estimated Energy Consumption during an MPI_Alltoall operation with 128K Message size and Varying System Size

Communication Predictability

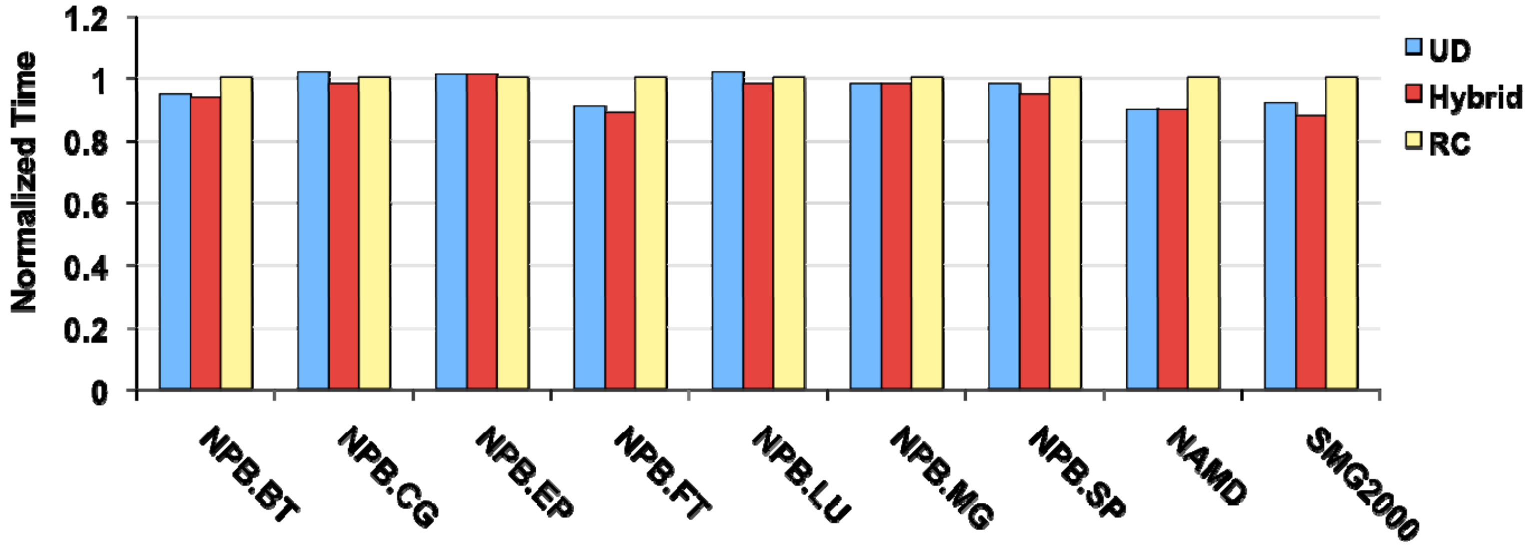
- IB QoS component: Virtual Lane (VL) VL0 … VL15
- Distribute traffic across all the VLs
- Reduce communication contention





- Average latency is decreased
- More stable minimum latency
- Performance predictability is improved

Handling Memory Scalability – Hybrid Transport Design (UD/RC/XRC)

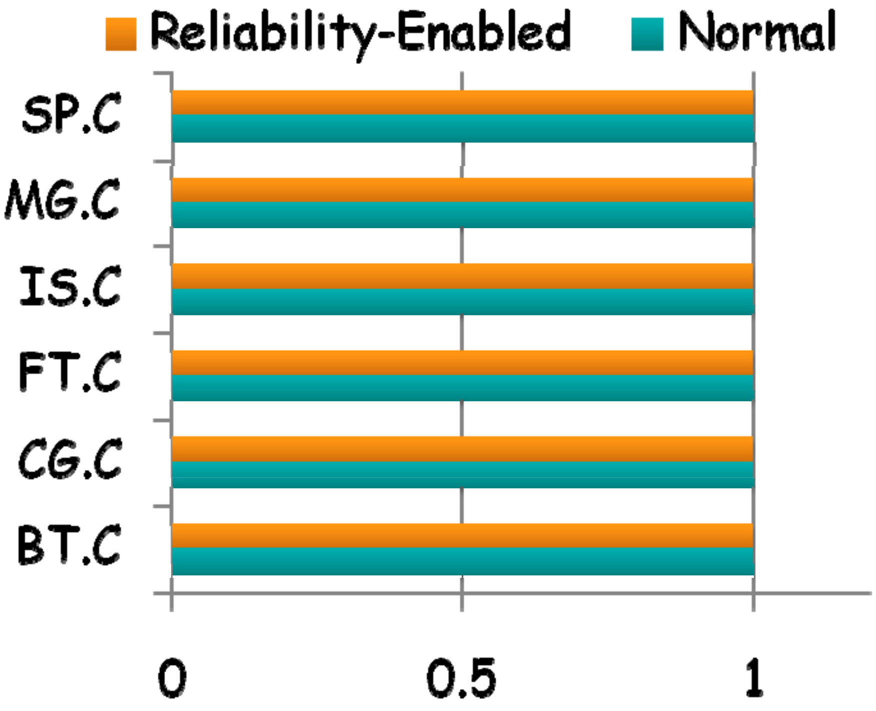


- Both UD and RC/XRC have benefits
- Automatic adaptation based on application characteristics
- Delivers best performance with reduced memory footprint
- Available in MVAPICH 1.1 and 1.2
- Will be available in MVAPICH2 soon

M. Koop, T. Jones and D. K. Panda, “MVAPICH-Aptus: Scalable High-Performance Multi-Transport MPI over InfiniBand,” IPDPS ‘08

Network Fault Resiliency

- MPI jobs typically abort if there is a failure in link, adapter or switch
- Can we stall a job instead of aborting it while the failed component is fixed?
- Protection against various network failures
  - Switch reboot/failure
  - HCA failure
  - Severe congestion
- Available in MVAPICH 1.2 and OFED 1.5.1

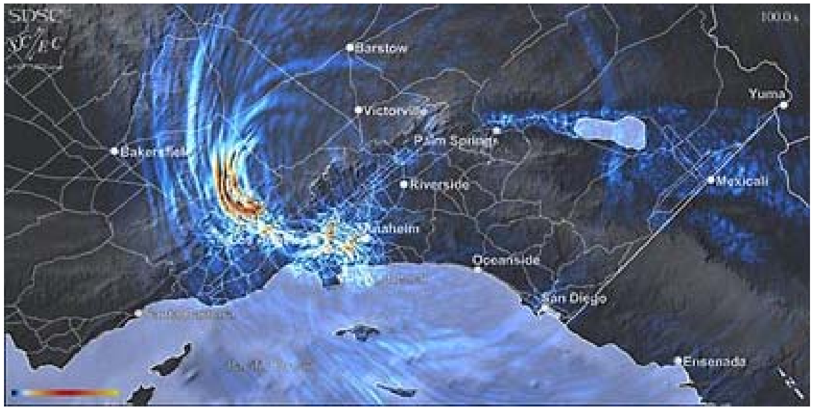


Normalized Time
(256 cores, DDR HCAs)

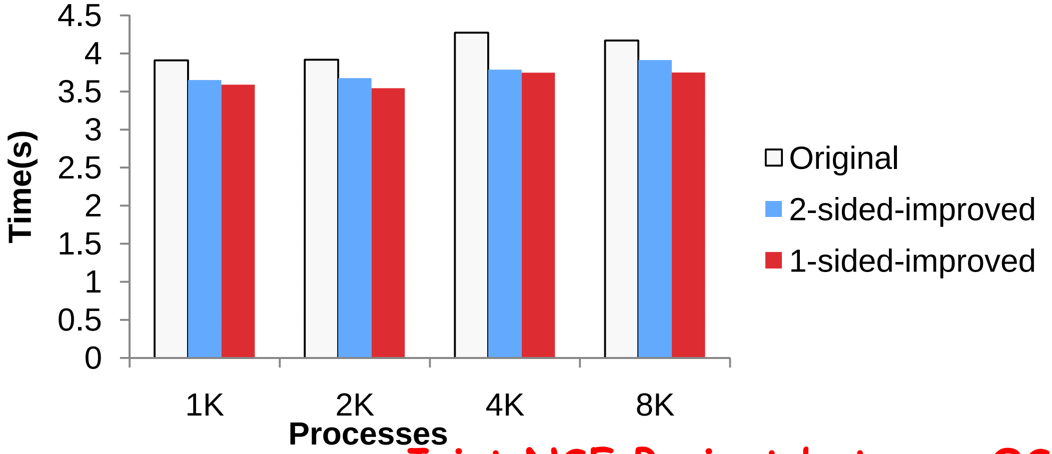
M. Koop, P. Shamis, I. Rabinovitz and D. K. Panda, Designing High-Performance and Resilient Message Passing on InfiniBand, The 10th Workshop on Communication Architecture for Clusters (CAC 10), May 2010.

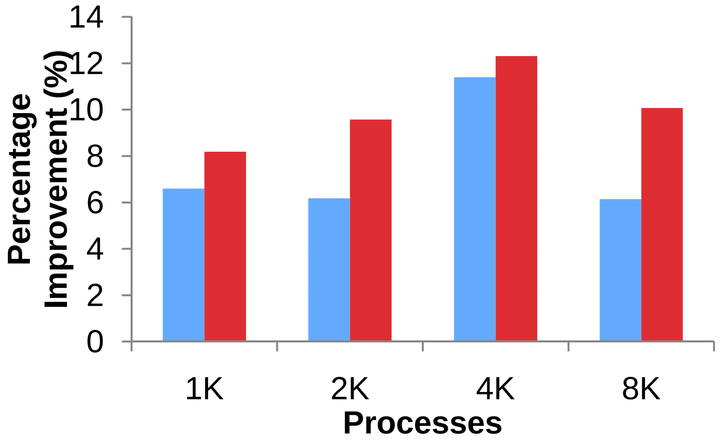
Benefits of MPI-2 One-Sided Communication

- AWM-Olsen, a fourth-order finite difference code for seismic simulation
- Consumes 10’s of millions of SU’s every year on the TeraGrid Network
- Spends 31% of time in MPI_Waitall() due to blocking communication design
- Improved using MPI-2 One-sided semantics for overlap – can save up to 65,500 core-hours in a single real-world run on 32K processes



Shakeout Earthquake Simulation
Visualization Credits: Amit Chourasia - Visualization Services, SDSC
Simulation Credits: Kim Olsen et. al., SCEC; Yifeng Cui et. al., SDSC



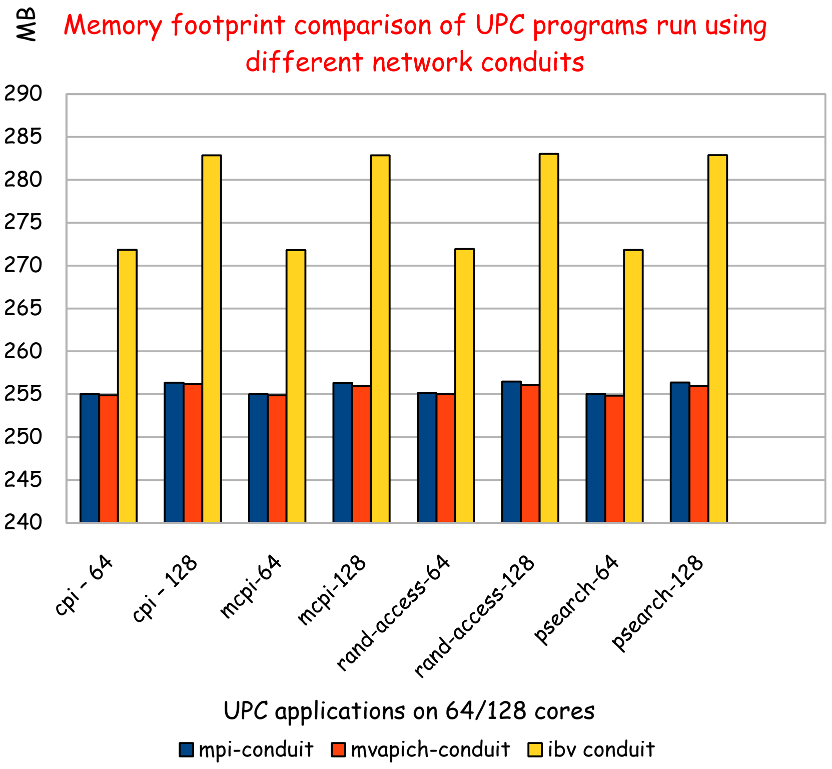


Joint NSF Project between OSU, TACC and SDSC

Supporting Hybrid Programming Models

- No single software stack with support for hybrid programming models (MPI+PGAS) yet
- Using MVAPICH/MVAPICH2 lower layers to support PGAS programming models (UPC, X10, OpenShmem, etc.)
- Can make use of the already available high performance features in MPI stack for PGAS models
- Multiple benefits
    - Better scalability and performance for PGAS
    - Will allow users to explore hybrid programming models

An Early Prototype in Supporting MPI + UPC


Memory footprint comparison of UPC programs run using different network conduits

Conclusions

- Designing MPI at Exascale brings multiple challenges
- MVAPICH/MVAPICH2 project is already addressing some of these challenges
- Plan to scale MVAPICH2 stack to 500K-1M cores during the next few years
- OpenIB 2003 (DOE Workshop on InfiniBand)
  - Can 100K-core IB clusters with MPI be operational by 2010?
- OpenFabrics 2010
  - Can 1-10-100M core IB clusters with MPI be operational by 2018?

Web Pointers

MVAPICH

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

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