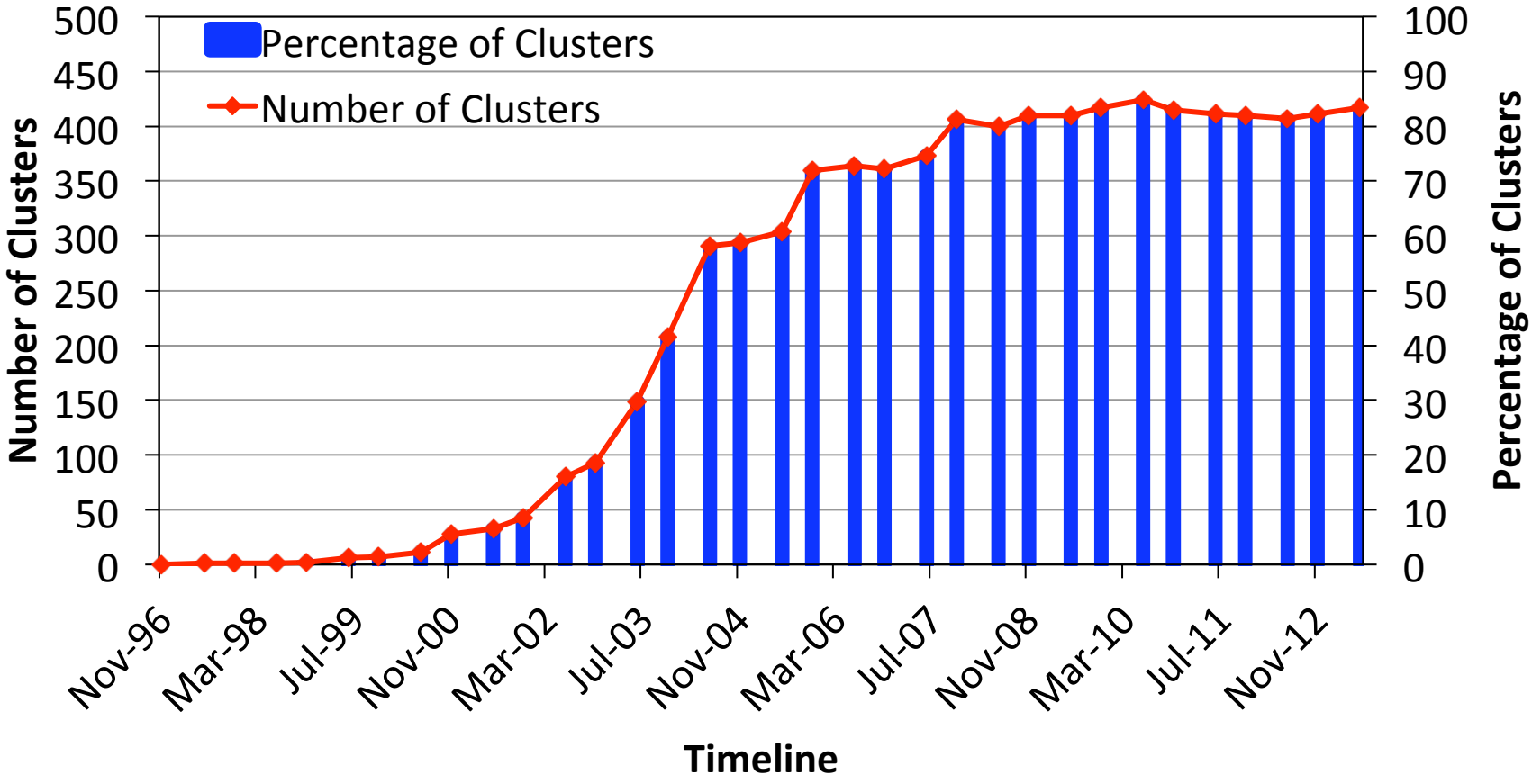
Overview of MVAPICH2 and MVAPICH2-X: Latest Status and Future Roadmap

MVAPICH2 User Group (MUG) Meeting

by

Dhabaleswar K. (DK) Panda
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

Trends for Commodity Computing Clusters in the Top 500 List (http://www.top500.org)



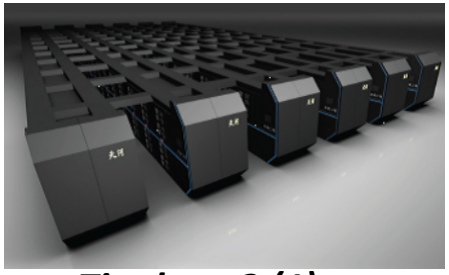
Drivers of Modern HPC Cluster Architectures

Multi-core Processors

High Performance Interconnects - InfiniBand
<1usec latency, >100Gbps Bandwidth

Accelerators / Coprocessors
high compute density, high performance/watt
>1 TFlop DP on a chip

- Multi-core processors are ubiquitous
- InfiniBand very popular in HPC clusters
- Accelerators/Coprocessors becoming common in high-end systems
- Pushing the envelope for Exascale computing



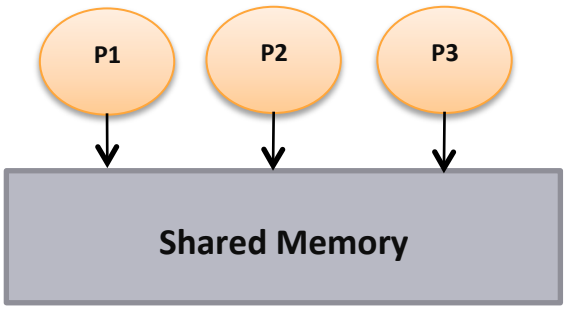
Tianhe – 2 (1)

Titan (2)

Stampede (6)

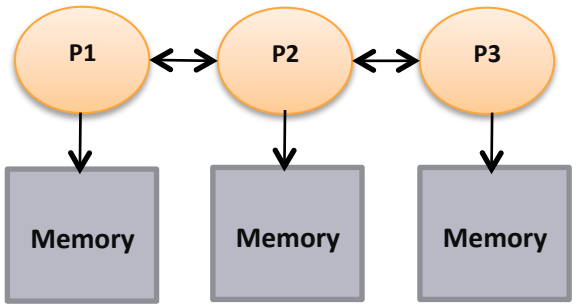
Tianhe – 1A (10)

Parallel Programming Models Overview



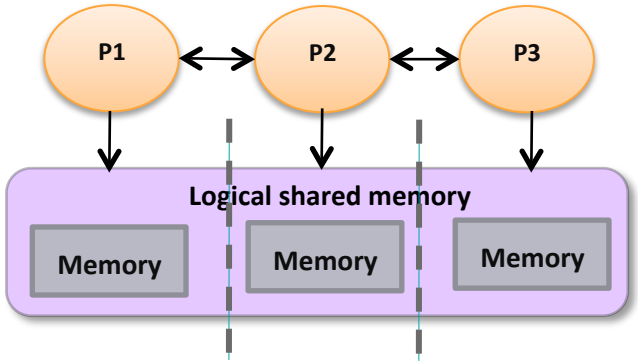
Shared Memory Model

SHMEM, DSM



Distributed Memory Model

MPI (Message Passing Interface)

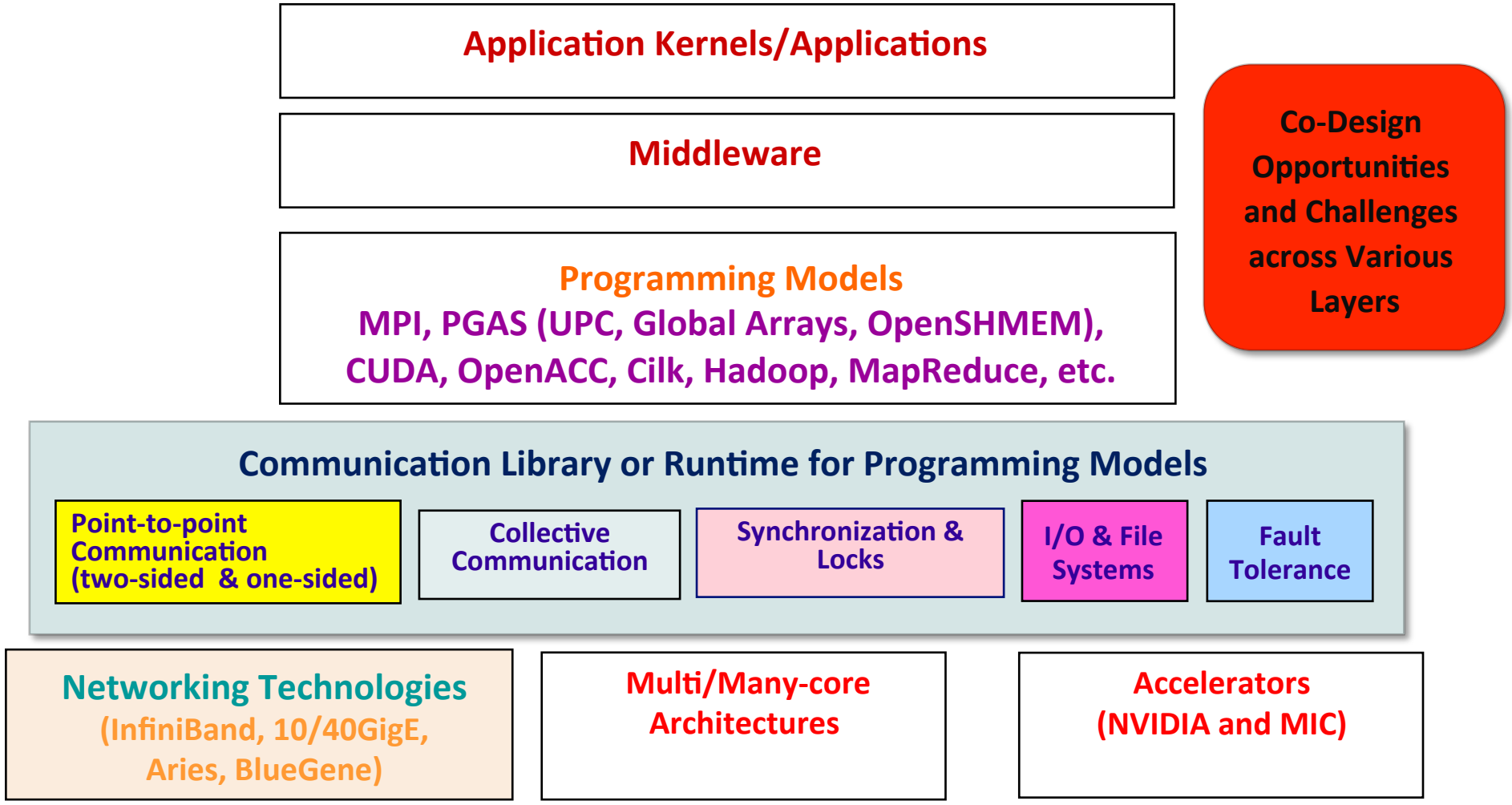


Partitioned Global Address Space (PGAS)

Global Arrays, UPC, Chapel, X10, CAF, …

- Programming models provide abstract machine models
- Models can be mapped on different types of systems
  - e.g. Distributed Shared Memory (DSM), MPI within a node, etc.
- PGAS models and Hybrid MPI+PGAS models are gradually receiving importance

Supporting Programming Models for Multi-Petaflop and Exaflop Systems: Challenges



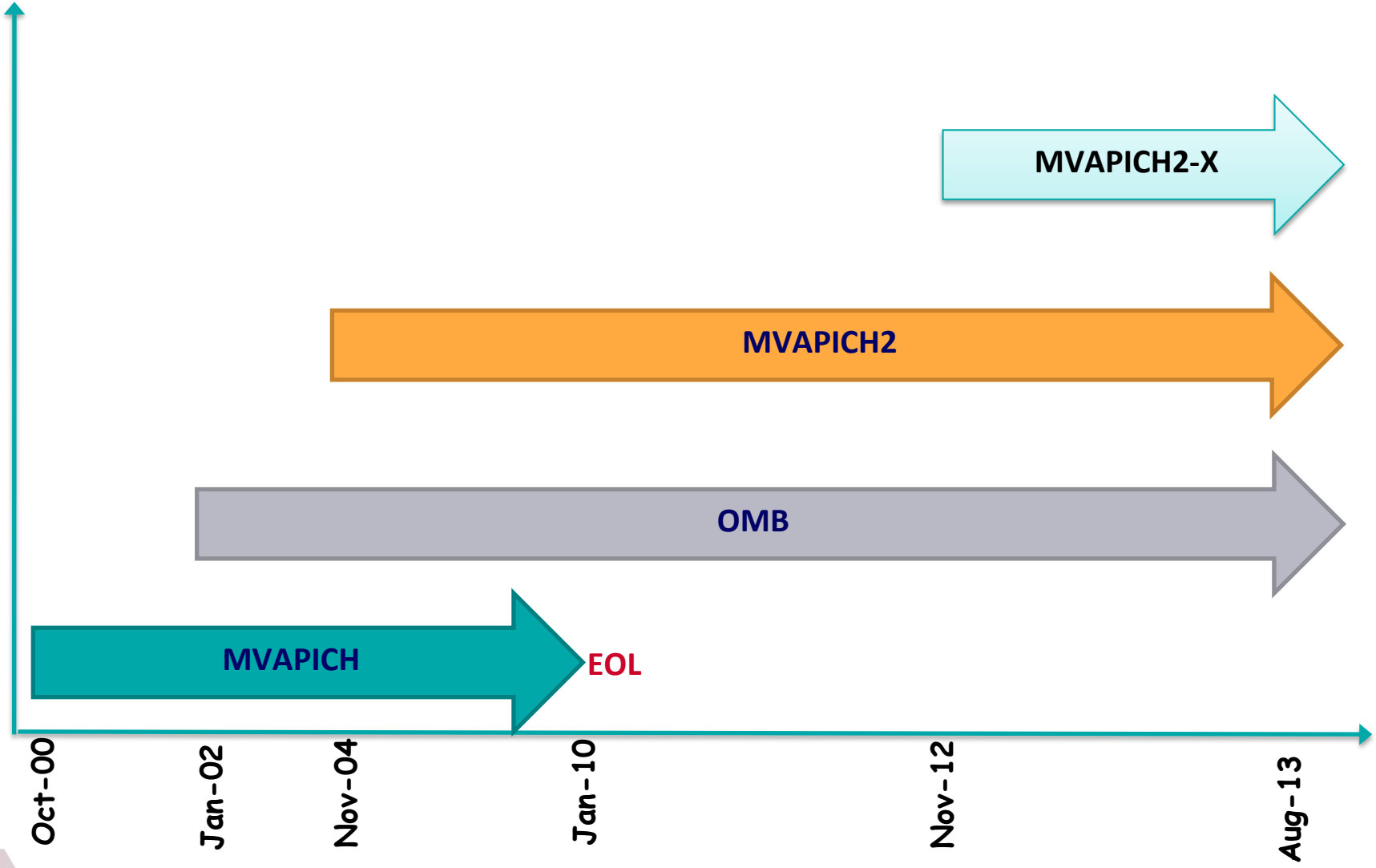
Designing (MPI+X) at Exascale

- Scalability for million to billion processors
  - Support for highly-efficient inter-node and intra-node communication (both two-sided and one-sided)
  - Extremely minimum memory footprint
- Hybrid programming (MPI + OpenMP, MPI + UPC, MPI + OpenSHMEM, …)
- Balancing intra-node and inter-node communication for next generation multi-core (128-1024 cores/node)
  - Multiple end-points per node
- Support for efficient multi-threading
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
  - Power-aware
- Support for MPI-3 RMA Model
- Support for GPGPUs and Accelerators
- Fault-tolerance/resiliency
- QoS support for communication and I/O

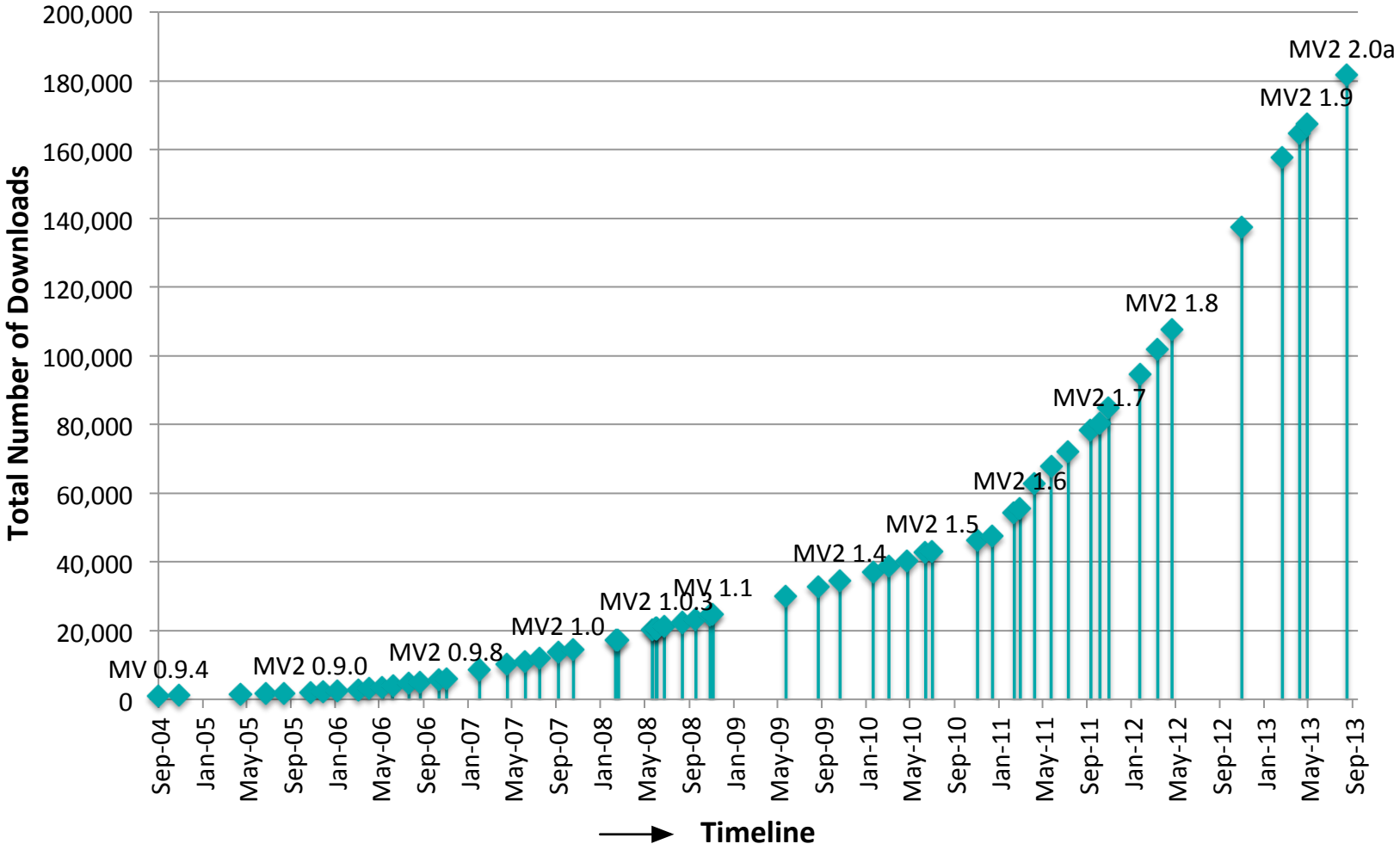
MVAPICH2/MVAPICH2-X Software

- http://mvapich.cse.ohio-state.edu
- High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP, and RDMA over Converged Enhanced Ethernet (RoCE)
  - MVAPICH (MPI-1) ,MVAPICH2 (MPI-2.2 and MPI-3.0), Available since 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2012
  - Used by more than 2,077 organizations (HPC Centers, Industry and Universities) in 70 countries

MVAPICH Team Projects



MVAPICH/MVAPICH2 Release Timeline and Downloads



- Download counts from MVAPICH2 website

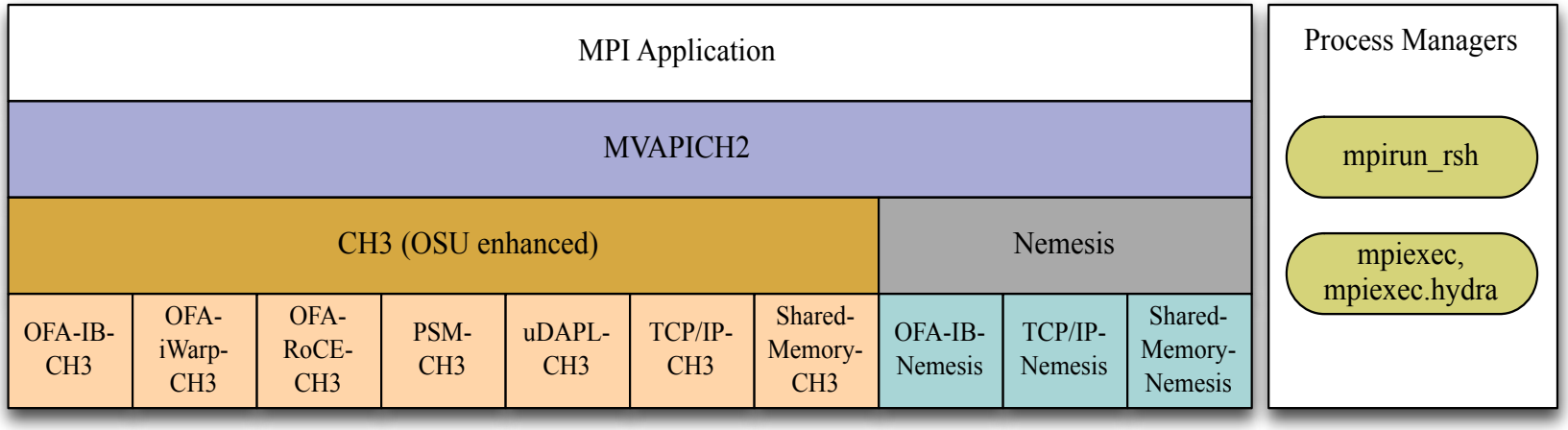
MVAPICH2/MVAPICH2-X Software (Cont'd)

- Available with software stacks of many IB, HSE, and server vendors including Linux Distros (RedHat and SuSE)
- Empowering many TOP500 clusters
  - 6th ranked 462,462-core cluster (Stampede) at TACC
  - 19th ranked 125,980-core cluster (Pleiades) at NASA
  - 21st ranked 73,278-core cluster (Tsubame 2.0) at Tokyo Institute of Technology and many others
- Partner in the U.S. NSF-TACC Stampede System

Strong Procedure for Design, Development and Release

- Research is done for exploring new designs
- Designs are first presented to conference/journal publications
- Best performing designs are incorporated into the codebase
- Rigorous Q&A procedure before making a release
  - Exhaustive unit testing
  - Various test procedures on diverse range of platforms and interconnects
  - Performance tuning
  - Applications-based evaluation
  - Evaluation on large-scale systems
- Even alpha and beta versions go through the above testing

MVAPICH2 Architecture (Latest Release 2.0)



All Different PCI interfaces

Major Computing Platforms: IA-32, EM64T, Nehalem, Westmere, Sandybridge, Opteron, Magny, ..

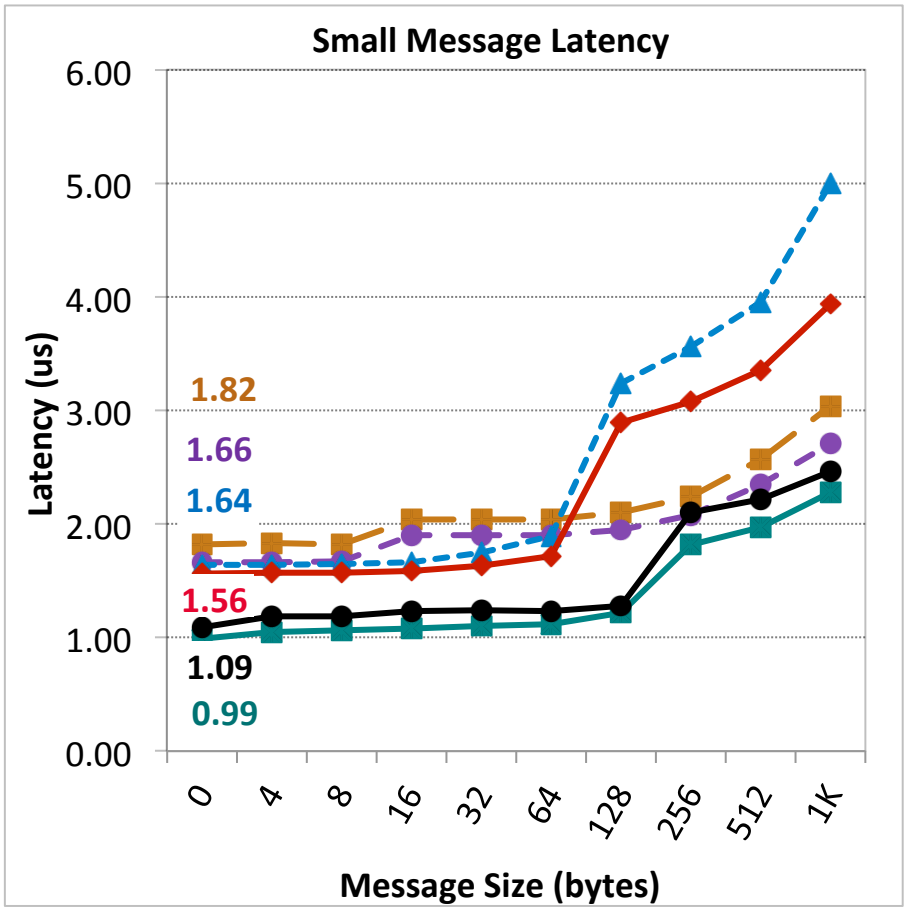
MVAPICH2 1.9 and MVAPICH2-X 1.9

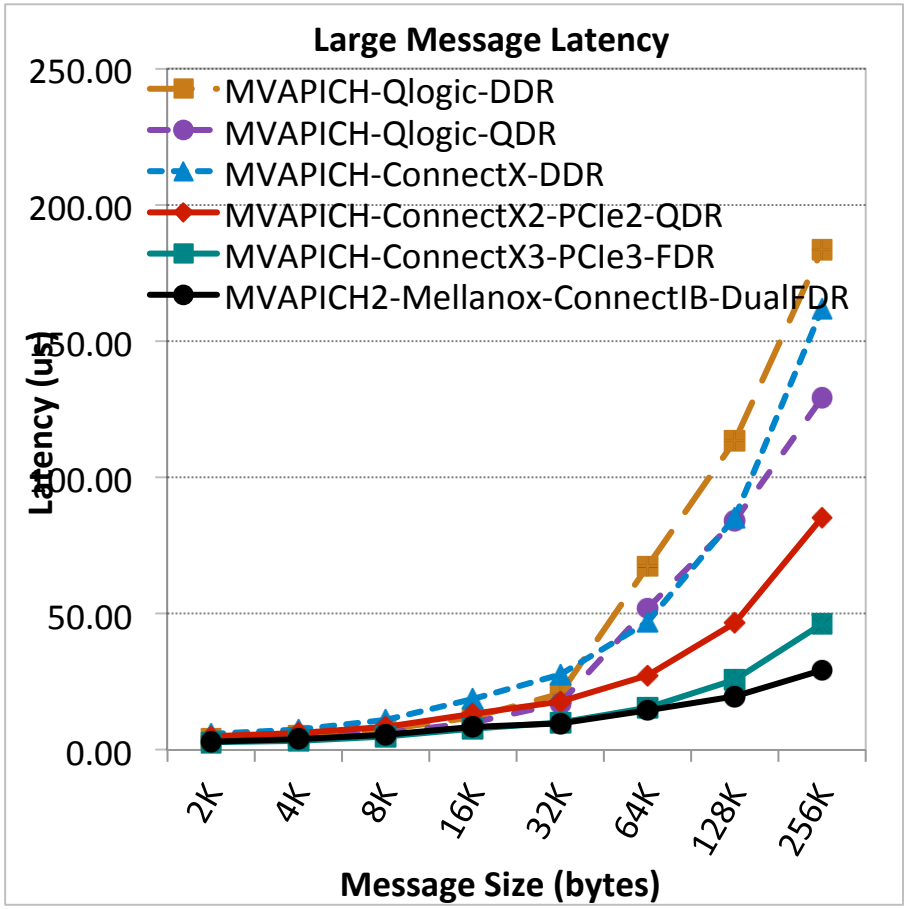
- Released on 05/06/13
- Major Features and Enhancements
  - Based on MPICH-3.0.3
    - Support for MPI-3 features
  - Support for single copy intra-node communication using Linux supported CMA (Cross Memory Attach)
    - Provides flexibility for intra-node communication: shared memory, LiMIC2, and CMA
  - Checkpoint/Restart using LLNL's Scalable Checkpoint/Restart Library (SCR)
    - Support for application-level checkpointing
    - Support for hierarchical system-level checkpointing
  - Scalable UD-multicast-based designs and tuned algorithm selection for collectives
  - Improved and tuned MPI communication from GPU device memory
  - Improved job startup time
    - Provided a new runtime variable MV2_HOMOGENEOUS_CLUSTER for optimized startup on homogeneous clusters
  - Revamped Build system with support for parallel builds
- MVAPICH2-X 1.9 supports hybrid MPI + PGAS (UPC and OpenSHMEM) programming models.
  - Based on MVAPICH2 1.9 including MPI-3 features; Compliant with UPC 2.16.2 and OpenSHMEM v1.0d

MVAPICH2 2.0a and MVAPICH2-X 2.0a

- Released on 08/24/13
- Major Features and Enhancements
  - Based on MPICH-3.0.4
  - Dynamic CUDA initialization. Support GPU device selection after MPI_Init
  - Support for running on heterogeneous clusters with GPU and non-GPU nodes
  - Supporting MPI-3 RMA atomic operations and flush operations with CH3-Gen2 interface
  - Exposing internal performance variables to MPI-3 Tools information interface (MPIT)
  - Enhanced MPI_Bcast performance
  - Enhanced performance for large message MPI_Scatter and MPI_Gather
  - Enhanced intra-node SMP performance
  - Reduced memory footprint
  - Improved job-startup performance
- MVAPICH2-X 2.0a supports hybrid MPI + PGAS (UPC and OpenSHMEM) programming models.
  - Based on MVAPICH2 2.0a including MPI-3 features; Compliant with UPC 2.16.2 and OpenSHMEM v1.0d
  - Improved OpenSHMEM collectives

One-way Latency: MPI over IB



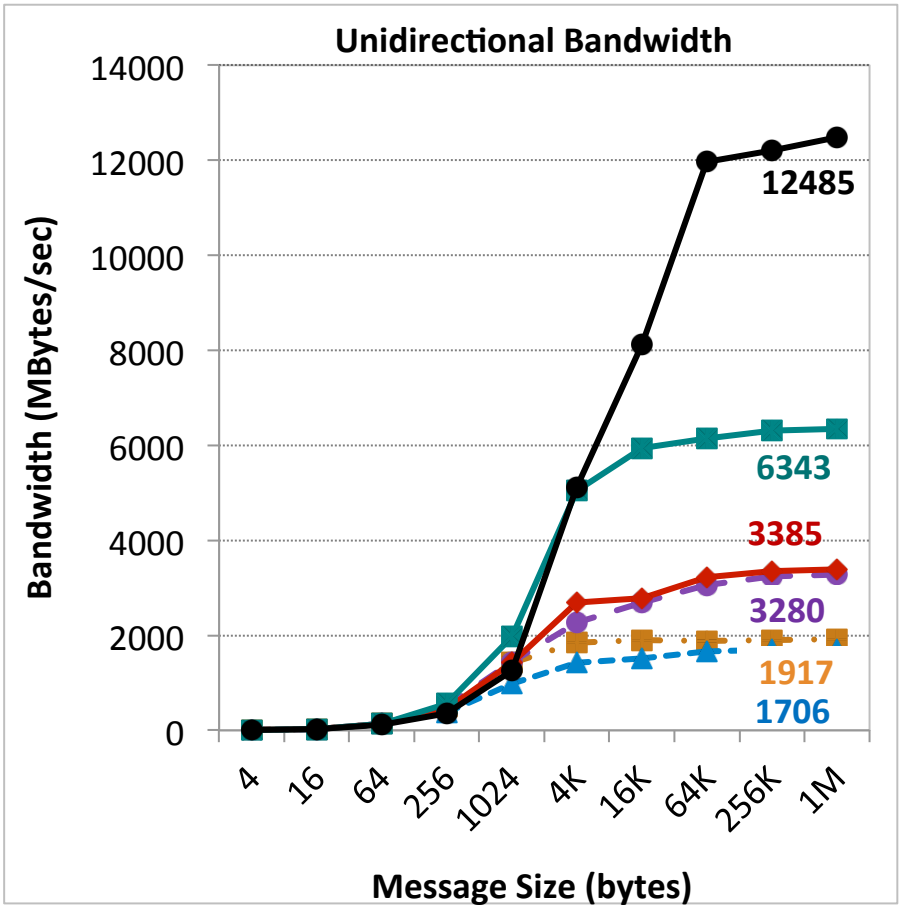


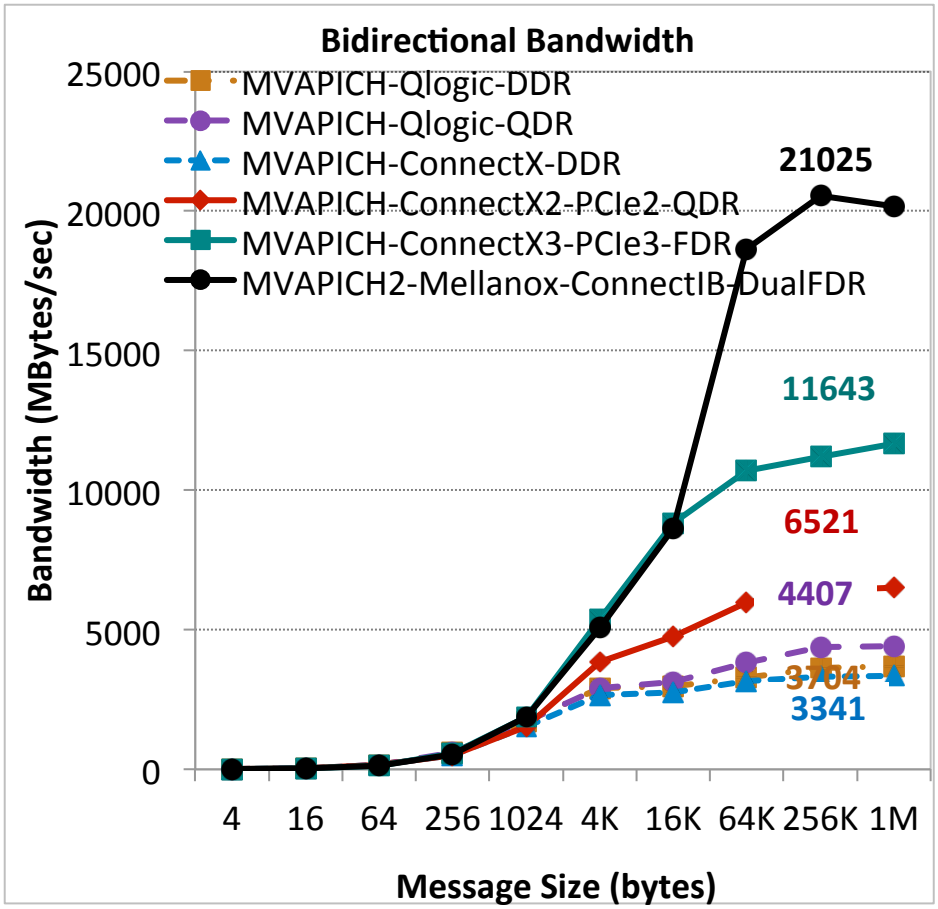
DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch

FDR - 2.6 GHz Octa-core (Sandybridge) Intel PCI Gen3 with IB switch

ConnectIB-Dual FDR - 2.6 GHz Octa-core (Sandybridge) Intel PCI Gen3 with IB switch

Bandwidth: MPI over IB



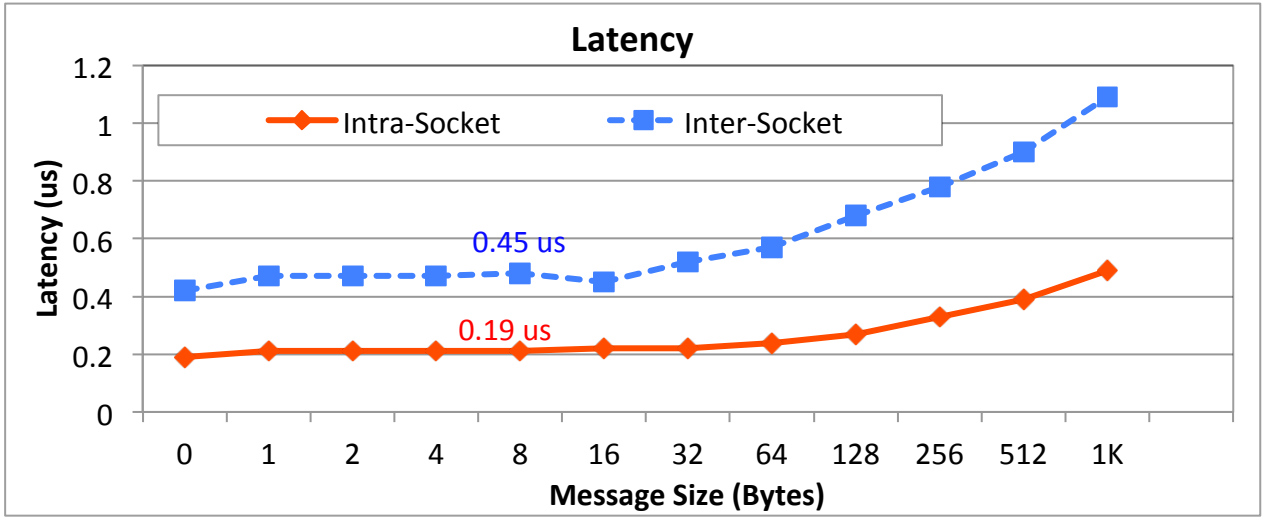


DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch

FDR - 2.6 GHz Octa-core (Sandybridge) Intel PCI Gen3 with IB switch

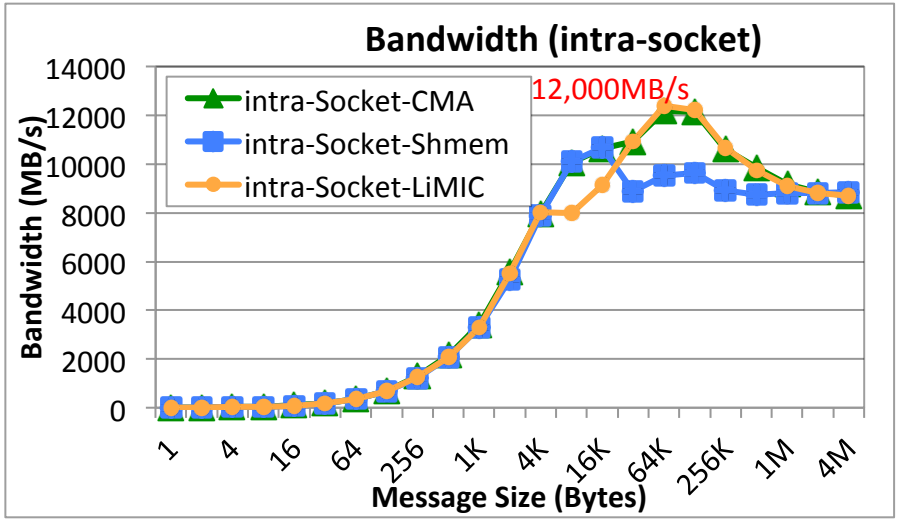
ConnectIB-Dual FDR - 2.6 GHz Octa-core (Sandybridge) Intel PCI Gen3 with IB switch

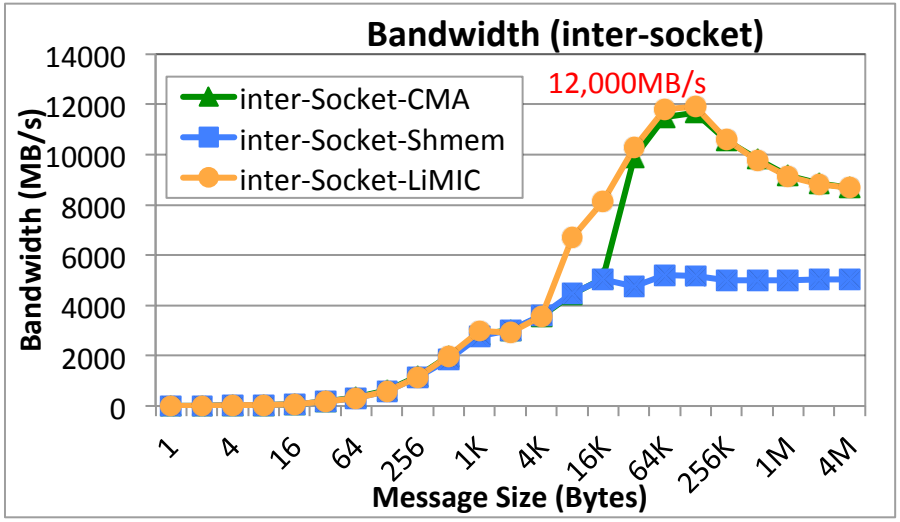
MVAPICH2 Two-Sided Intra-Node Performance
(Shared memory and Kernel-based Zero-copy Support (LiMIC and CMA))



Latest MVAPICH2 2.0a

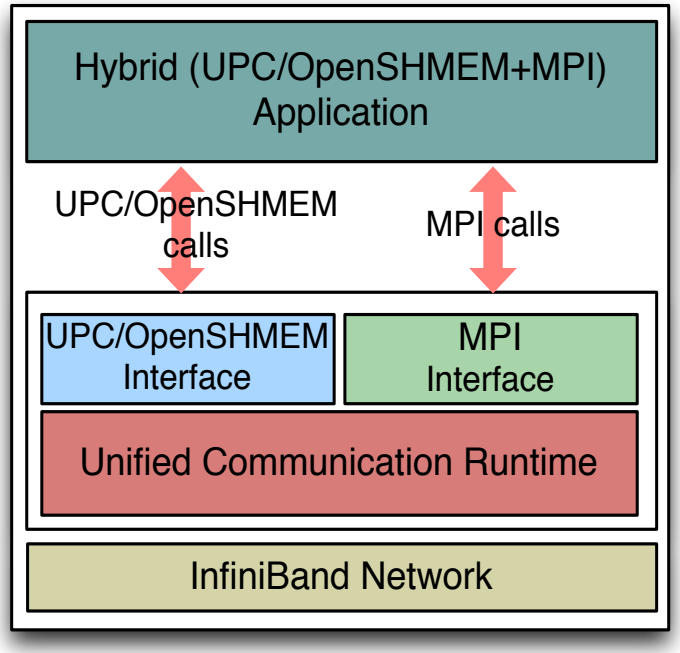
Intel Sandy-bridge





Scalable OpenSHMEM/UPC and Hybrid (MPI, UPC and OpenSHMEM) designs

- Based on OpenSHMEM Reference Implementation (http://openshmem.org/) & UPC version 2.14.2 (http://upc.lbl.gov/)
  - Provides a design over GASNet
  - **Does not take advantage of all OFED features**
- Design Scalable and High-Performance OpenSHMEM & UPC over OFED
- Designing a Hybrid MPI + OpenSHMEM/UPC Model
  - Current Model – Separate Runtimes for OpenSHMEM/UPC and MPI
    - Possible deadlock if both runtimes are not progressed
    - Consumes more network resource
  - Our Approach – Single Unified Runtime for MPI and OpenSHMEM/UPC



Hybrid MPI+OpenSHMEM/UPC

Available since MVAPICH2-X 1.9

OSU Micro-Benchmarks (OMB)

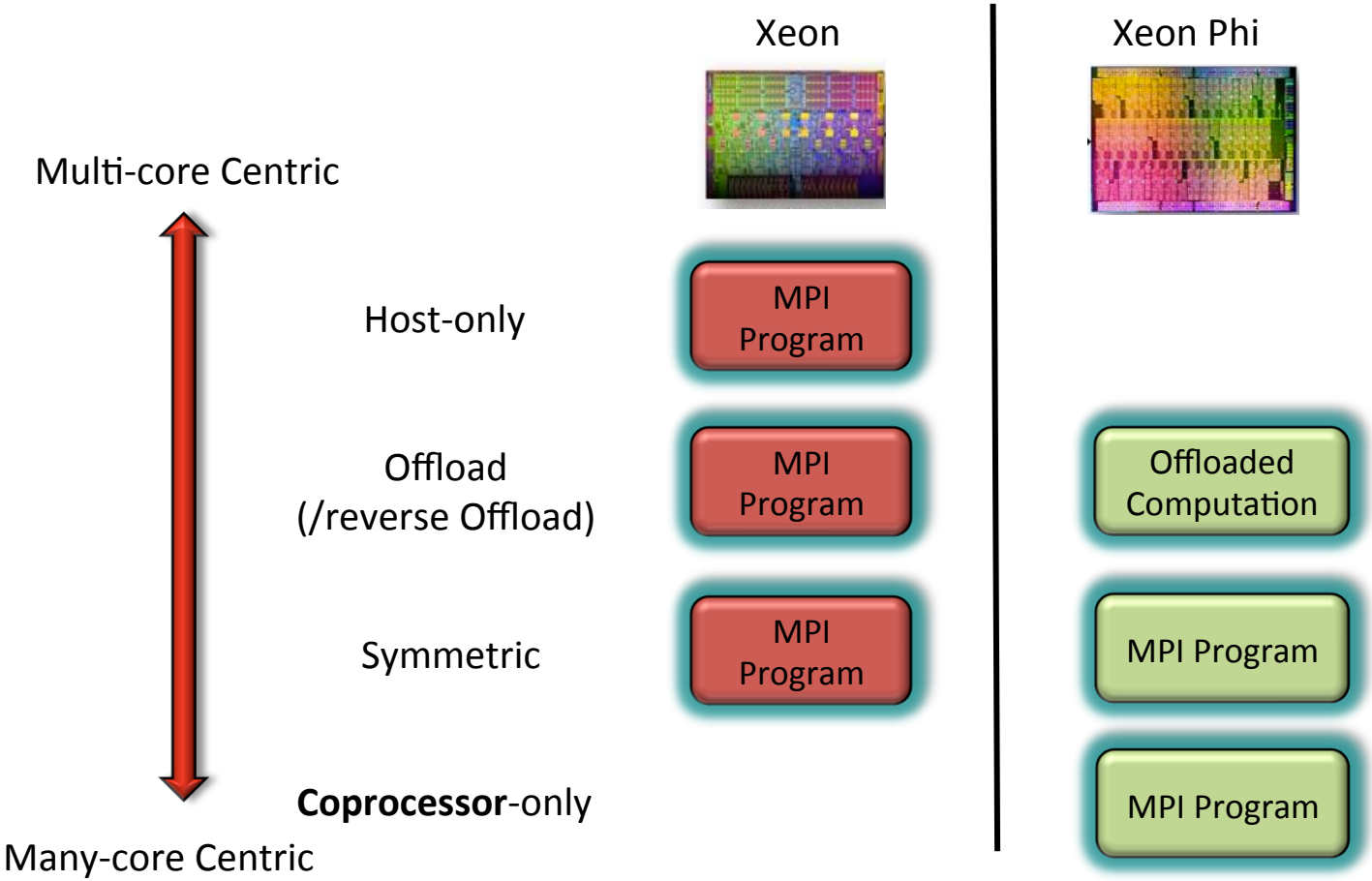
- Started in 2004 and continuing steadily
- Allows MPI developers and users to
  - Test and evaluate MPI libraries
- Has a wide-range of benchmarks
  - Two-sided (MPI-1, MPI-2 and MPI-3)
  - One-sided (MPI-2 and MPI-3)
  - RMA (MPI-3)
  - Collectives (MPI-1, MPI-2 and MPI-3)
  - Extensions for GPU-aware communication (CUDA and OpenACC)
  - UPC (Pt-to-Pt)
  - OpenSHMEM (Pt-to-Pt and Collectives)
- Widely-used in the MPI community

Designing GPU-Aware MPI Library

- OSU started this research and development direction in 2011
- Initial support was provided in MVAPICH2 1.8a (SC '11)
- Since then many enhancements and new designs related to GPU communication have been incorporated in 1.8, 1.9 and 2.0a series
- Have also extended OSU Micro-Benchmark Suite (OMB) to test and evaluate GPU-aware MPI communication
  - CUDA
  - OpenACC
- MVAPICH2 Design for GPUDirect RDMA (GDR)
  - Available based on 1.9

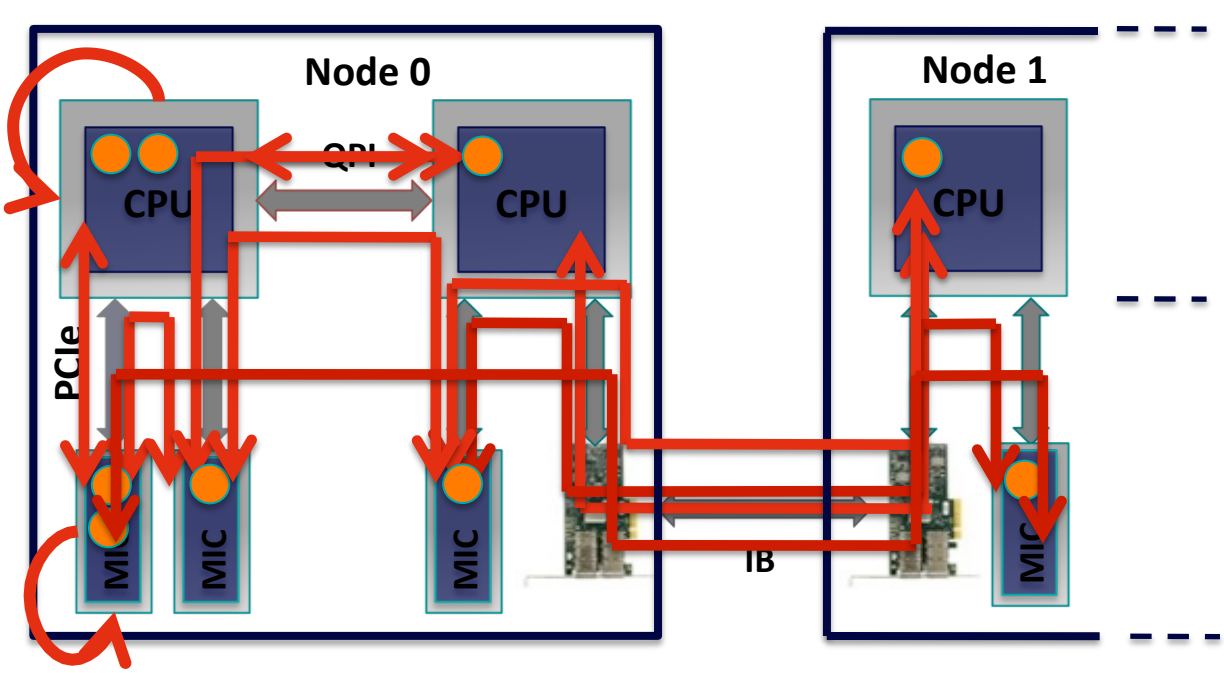
MPI Applications on MIC Clusters

- Flexibility in launching MPI jobs on clusters with Xeon Phi



Data Movement on Intel Xeon Phi Clusters

- Connected as PCIe devices – Flexibility but Complexity



MPI Process

1. Intra-Socket
2. Inter-Socket
3. Inter-Node
4. Intra-MIC
5. Intra-Socket MIC-MIC
6. Inter-Socket MIC-MIC
7. Inter-Node MIC-MIC
8. Intra-Socket MIC-Host
9. Inter-Socket MIC-Host
10. Inter-Node MIC-Host

11. Inter-Node MIC-MIC with IB adapter on remote socket and more . . .

- Critical for runtimes to optimize data movement, hiding the complexity

MVAPICH2-MIC Design for Clusters with IB and Xeon Phi

- Offload Mode
- Intranode Communication
  - Coprocessor-only Mode
  - Symmetric Mode
- Internode Communication
  - Coprocessors-only
  - Symmetric Mode
- Multi-MIC Node Configurations
- Based on MVAPICH2 1.9
- Being tested and deployed on TACC Stampede

MVAPICH2 – Plans for Exascale

- Performance and Memory scalability toward 500K-1M cores
  - Dynamically Connected Transport (DCT) service with Connect-IB
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF …)
- Enhanced Optimization for GPU Support and Accelerators
- Taking advantage of Collective Offload framework
  - Including support for non-blocking collectives (MPI 3.0)
  - Core-Direct support
- Extended RMA support (as in MPI 3.0)
- Extended topology-aware collectives
- Power-aware collectives
- Extended Support for MPI Tools Interface (as in MPI 3.0)
- Extended Checkpoint-Restart and migration support with SCR

Web Pointers

NOWLAB Web Page

http://nowlab.cse.ohio-state.edu

MVAPICH Web Page

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