MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

The MVAPICH2 Project
Latest Status and Future Plans

Presentation at MPICH BoF (SC '18)

by

Hari Subramoni

The Ohio State University

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

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

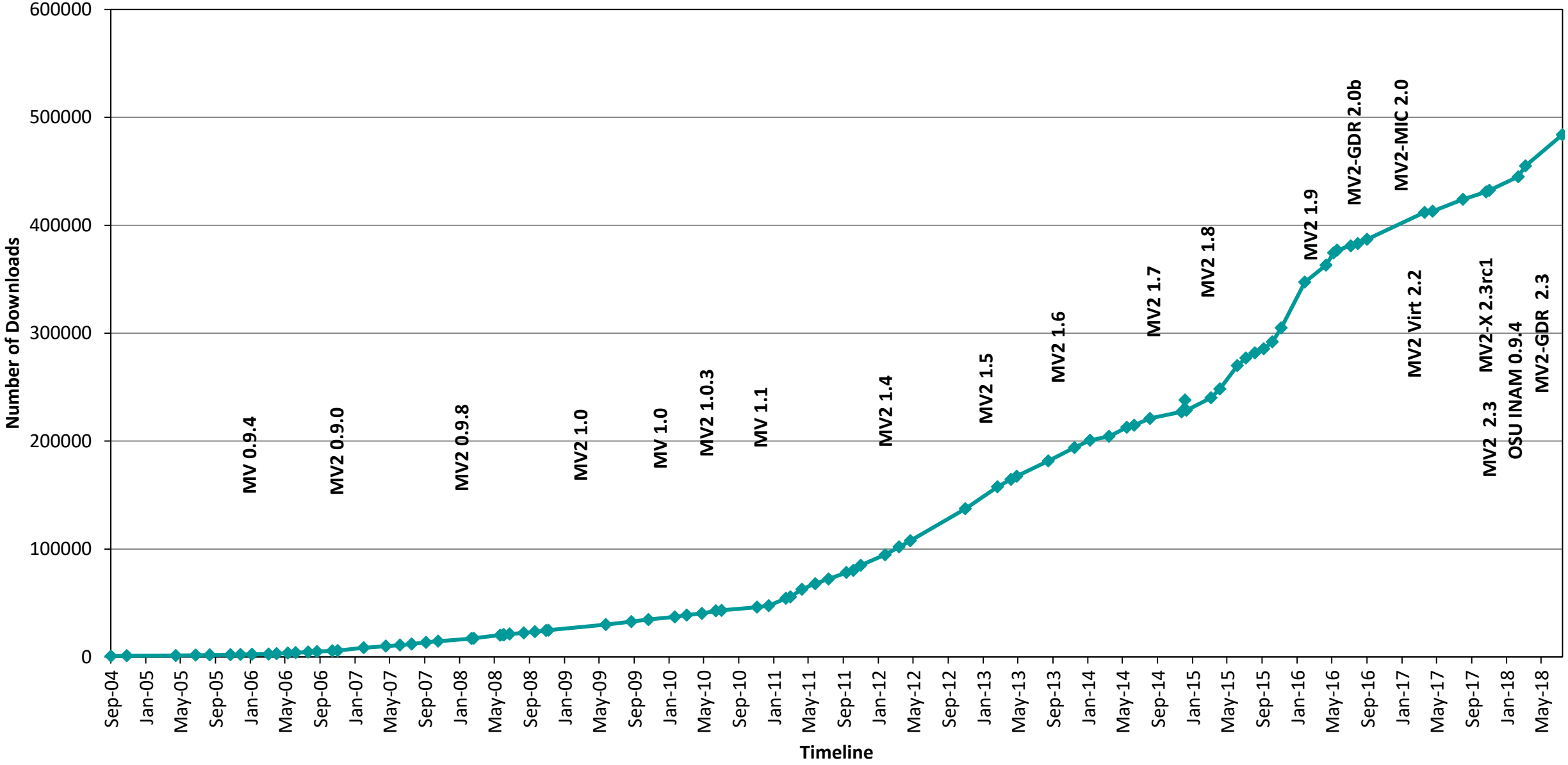
Overview of the MVAPICH2 Project

- High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.1), Started in 2001, First version available in 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2011
  - Support for GPGPUs (MVAPICH2-GDR) and MIC (MVAPICH2-MIC), Available since 2014
  - Support for Virtualization (MVAPICH2-Virt), Available since 2015
  - Support for Energy-Awareness (MVAPICH2-EA), Available since 2015
  - Support for InfiniBand Network Analysis and Monitoring (OSU INAM) since 2015
  - **Used by more than 2,950 organizations in 86 countries**
  - **More than 505,000 (> 0.5 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (Nov '18 ranking)
    - 3rd ranked 10,649,640-core cluster (Sunway TaihuLight) at NSC, Wuxi, China
    - 14th, 556,104 cores (Oakforest-PACS) in Japan
    - 17th, 367,024 cores (Stampede2) at TACC
    - 27th, 241,108-core (Pleiades) at NASA and many others
  - Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, and OpenHPC)
  - **http://mvapich.cse.ohio-state.edu**

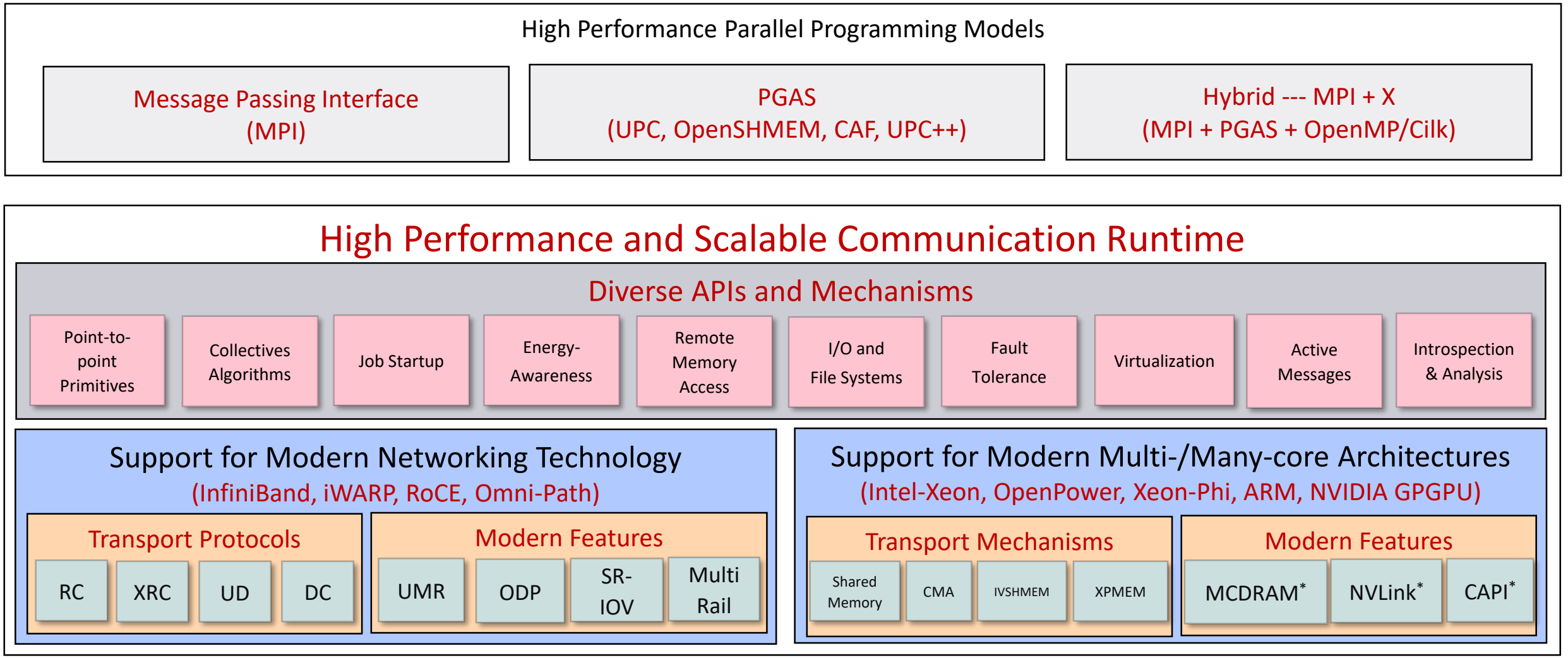
Partner in the upcoming TACC Frontera System

- Empowering Top500 systems for over a decade

MVAPICH2 Release Timeline and Downloads



Architecture of MVAPICH2 Software Family



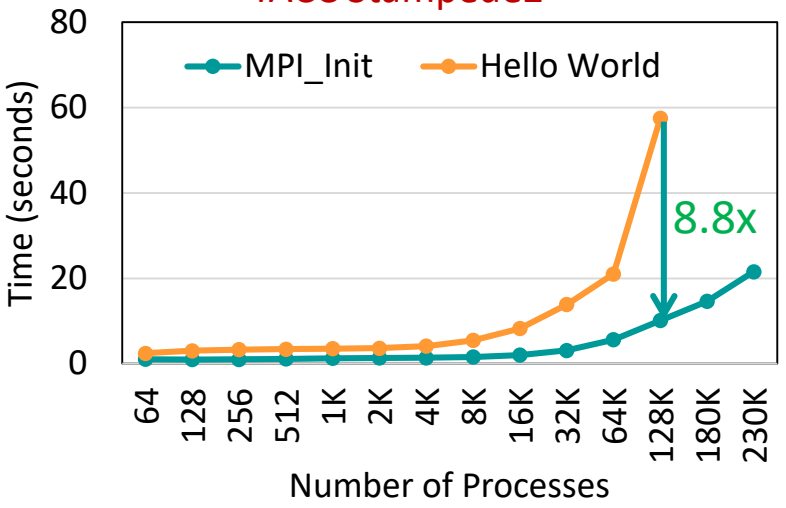
* Upcoming

MVAPICH2 Software Family

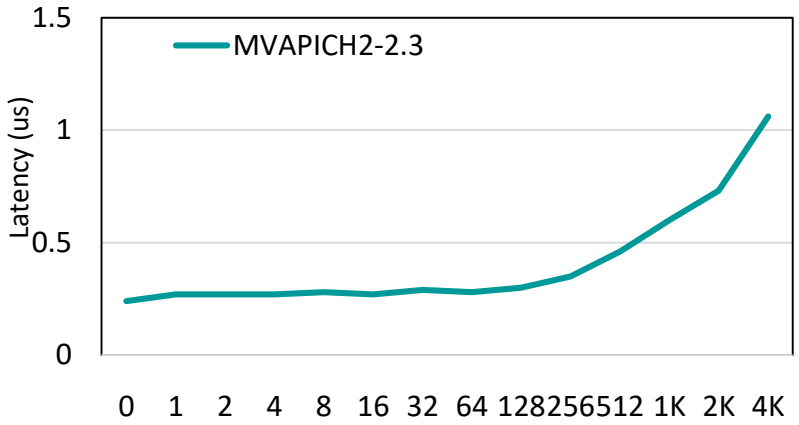
High-Performance Parallel Programming Libraries	
MVAPICH2	Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE
MVAPICH2-X	Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime
MVAPICH2-GDR	Optimized MPI for clusters with NVIDIA GPUs
MVAPICH2-Virt	High-performance and scalable MPI for hypervisor and container based HPC cloud
MVAPICH2-EA	Energy aware and High-performance MPI
MVAPICH2-MIC	Optimized MPI for clusters with Intel KNC
Microbenchmarks	
OMB	Microbenchmarks suite to evaluate MPI and PGAS (OpenSHMEM, UPC, and UPC++) libraries for CPUs and GPUs
Tools	
OSU INAM	Network monitoring, profiling, and analysis for clusters with MPI and scheduler integration
OEMT	Utility to measure the energy consumption of MPI applications

MVAPICH2 – Basic MPI

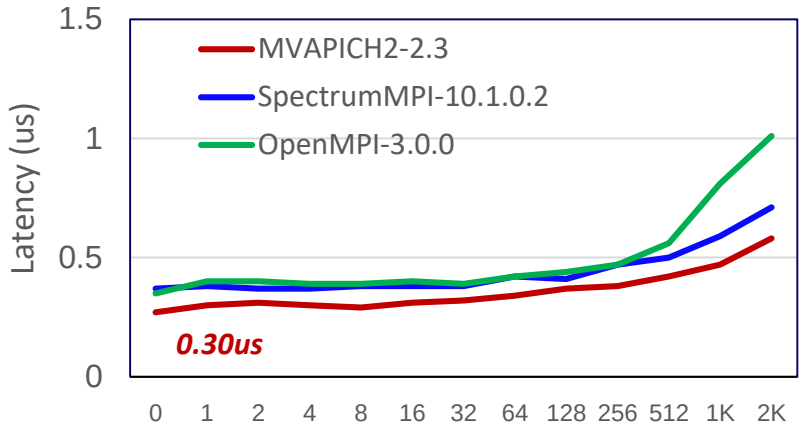
Fast Startup on Emerging Many-Cores
TACC Stampede2



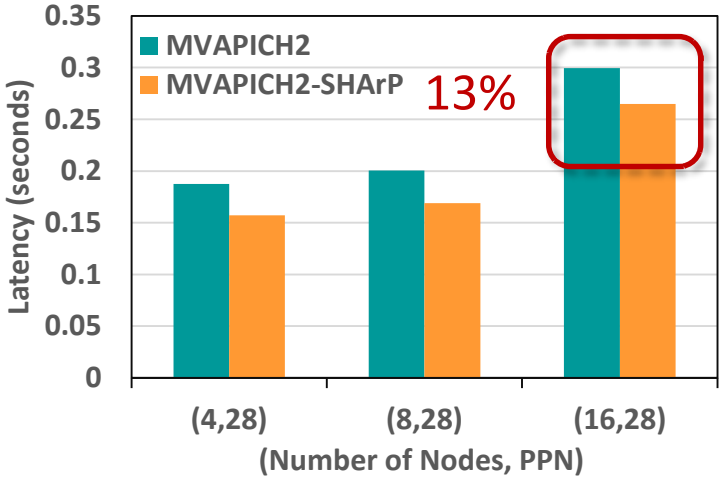
Enhanced Intra-node
Performance for ARM



Enhanced Intra-node
Performance for OpenPOWER



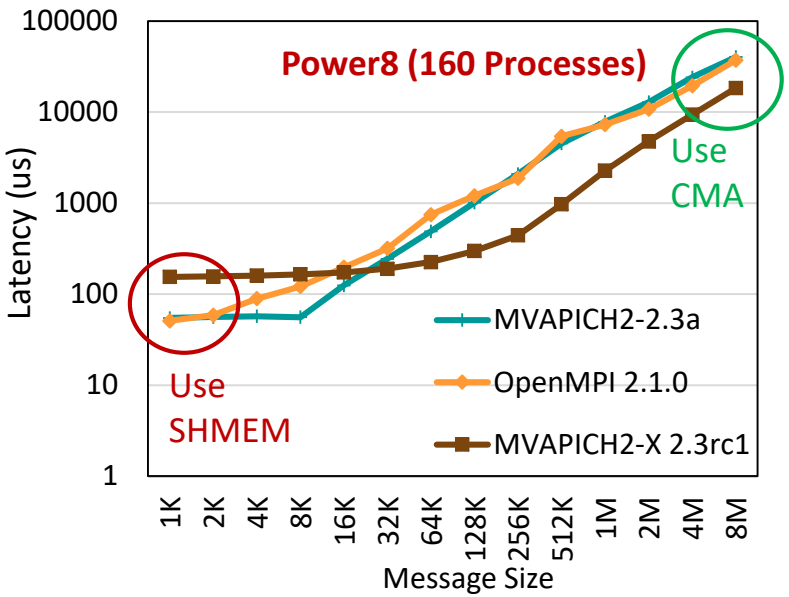
Advanced Allreduce with SHARP



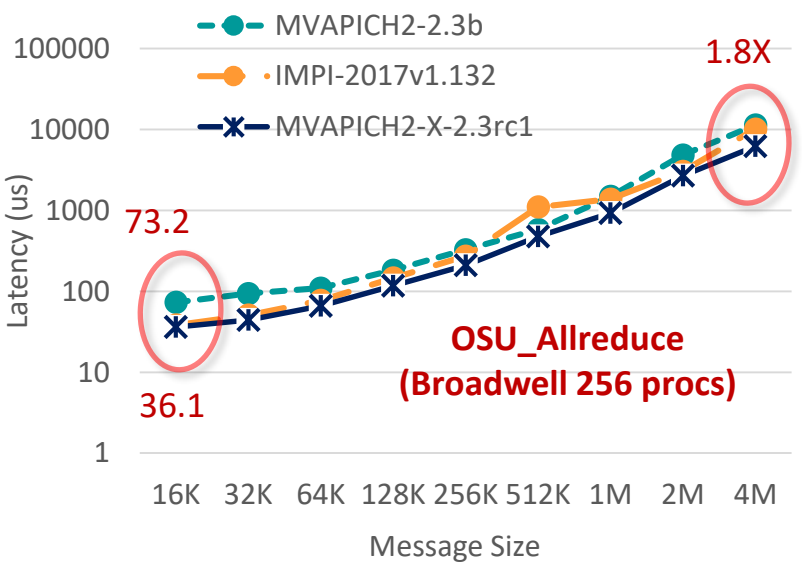
- Major Features and Enhancements in MVAPICH2 2.3 released on 07/23/2018
  - **Improved job startup time for OFA-IB-CH3, PSM-CH3, and PSM2-CH3**
  - **Enhanced performance of point-to-point operations for CH3-Gen2 (InfiniBand), CH3-PSM, and CH3-PSM2 (Omni-Path) channels**
  - **Enhanced performance for Allreduce, Reduce_scatter_block, Allgather, Allgatherv, Iallreduce**
  - **Enhanced process mapping strategies and automatic architecture/network detection**
  - **Enhanced support for MPI_T PVARs, CVARs and debugging abilities**

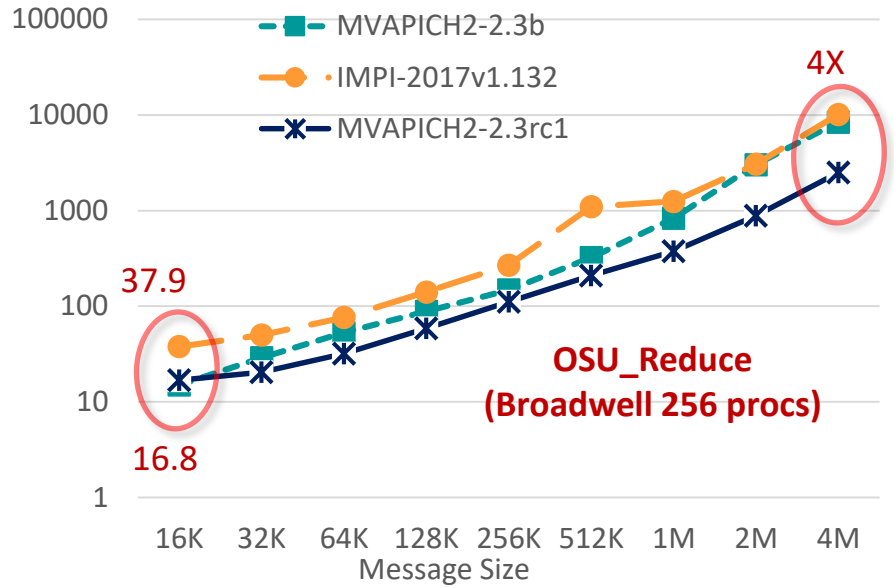
MVAPICH2-X – Advanced MPI + PGAS + Tools

CMA-Aware MPI_Bcast

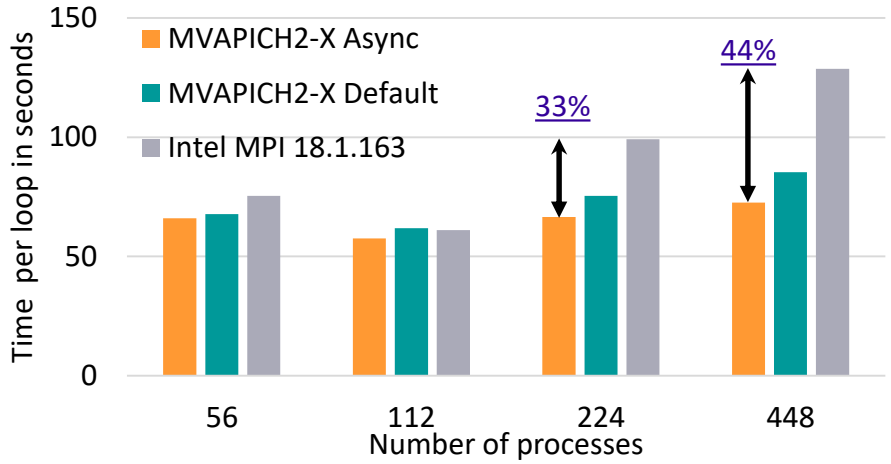


Shared Address Space (XPMEM)-based Collectives Design





Performance of P3DFFT Optimized Async Progress

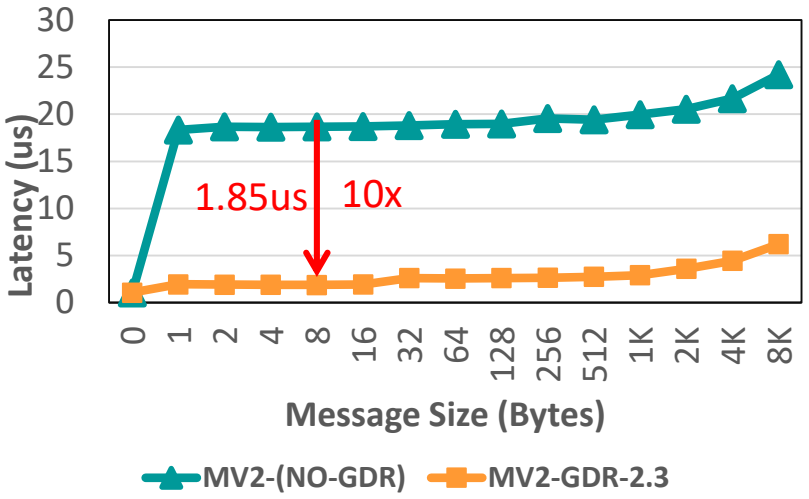


Major Features and Enhancements in MVAPICH2-X 2.3rc1 released on 09/21/2018

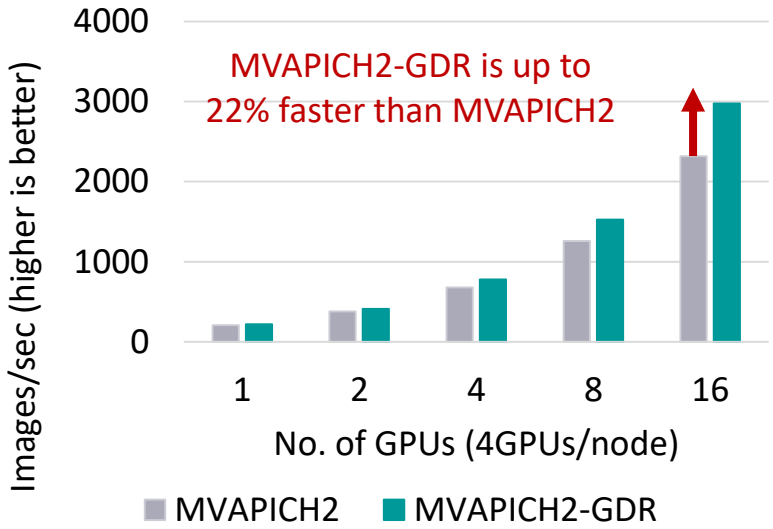
- **MPI Features**
  - Based on MVAPICH2 2.3GA
  - Optimized support for Skylake, ARM, and OpenPOWER architecture
- **MPI (Advanced) Features**
  - Support for XPMEM-based point-to-point operations and collective operations (Reduce and All-Reduce)
  - Enhanced asynchronous progress designs for progressing non-blocking point-to-point and collective operations
- **UPC Features**
  - Support Contention Aware Kernel-Assisted MPI collectives
- **OpenSHMEM Features**
  - Support Contention Aware Kernel-Assisted MPI collectives

MVAPICH2-GDR – Optimized MPI for clusters with NVIDIA GPUs

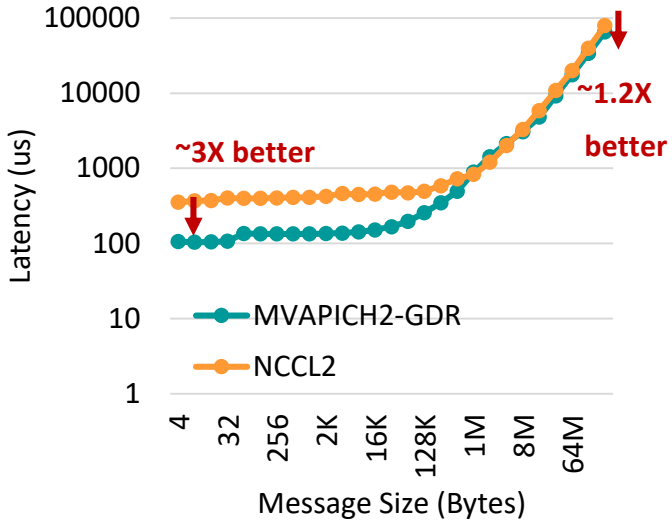
Best Performance for GPU-based Transfers



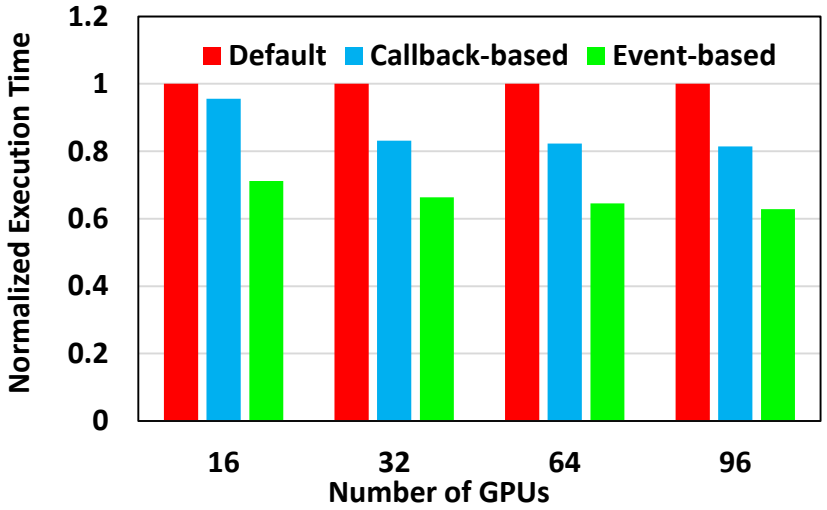
Exploiting CUDA-Aware MPI for TensorFlow (Horovod)



GPU-Based MPI_Allreduce

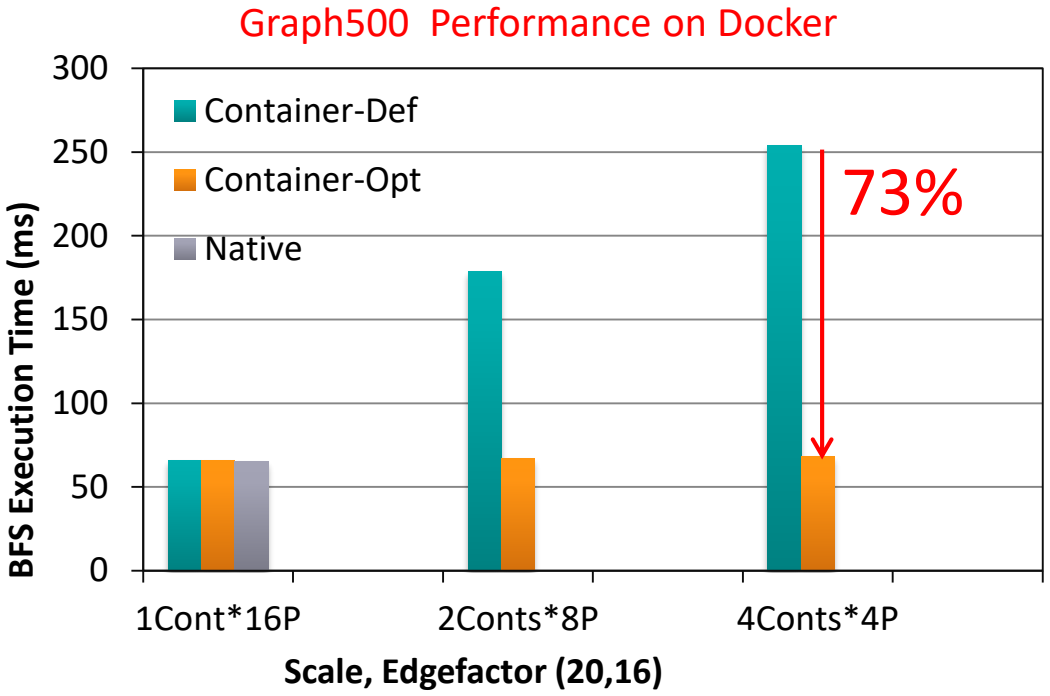


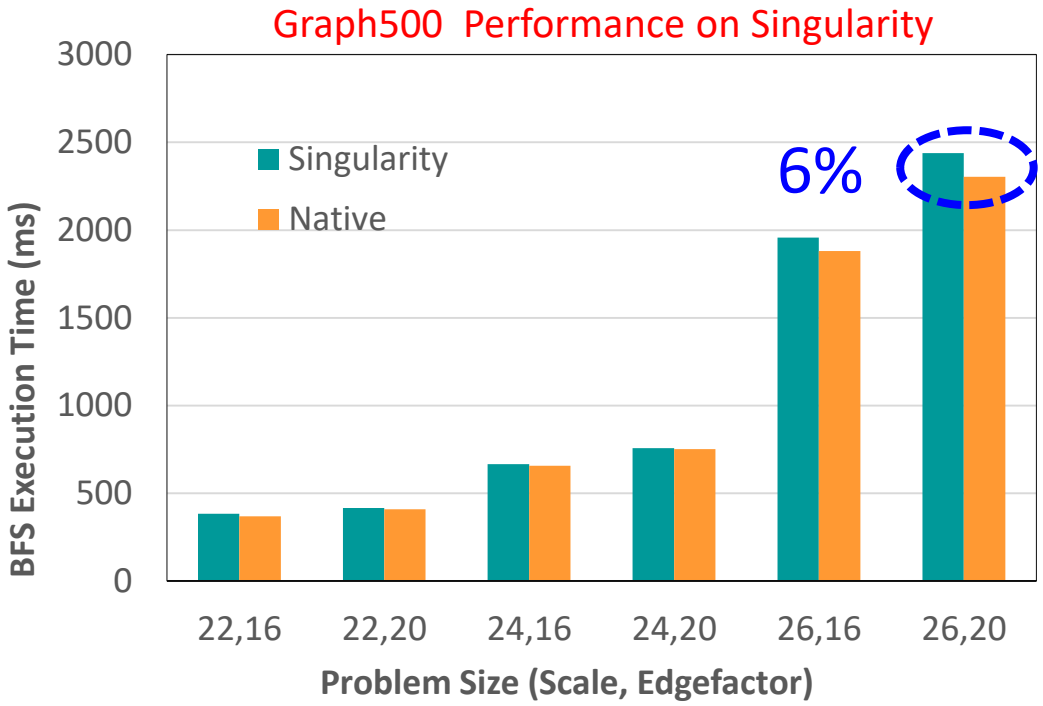
Enhanced Kernel-based Datatype Processing



- Major Features and Enhancements in MVAPICH2-GDR 2.3 released on 11/10/2018
  - **Support for CUDA 10, 9.2, 9.0, Volta (V100) GPU, and OpenPOWER with NVLink**
  - **Efficient Multiple CUDA stream-based IPC communication**
  - **Enhanced performance of GPU-based point-to-point communication**
  - **Leverage Linux CMA feature for enhanced host-based communication**
  - **InfiniBand Multicast based designs for GPU-based broadcast and streaming applications**
  - **Efficient reduce, allreduce, and broadcast designs for Deep Learning applications**
  - **Enhanced collective tuning on Xeon, OpenPOWER, and NVIDIA DGX-1 systems**

MVAPICH2-Virt – Advanced Support for HPC-Clouds





- Virtualization has many benefits
  - Fault-tolerance
  - Job migration
  - Compaction
- Have not been very popular in HPC due to overhead associated with Virtualization
- New SR-IOV (Single Root – IO Virtualization) support available with Mellanox InfiniBand adapters changes the field
- Enhanced MVAPICH2 support for SR-IOV
- MVAPICH2-Virt 2.2 supports:
  - OpenStack, Docker, and singularity

MVAPICH2 – Plans for Exascale

- Performance and Memory scalability toward 1M-10M cores
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF …)
  - MPI + Task*
- Enhanced Optimization for GPUs and FPGAs*
- Taking advantage of advanced features of Mellanox InfiniBand
  - Tag Matching*
  - Adapter Memory*
- Enhanced communication schemes for upcoming architectures
  - NVLINK*
  - CAPI*
- Extended topology-aware collectives
- Extended Energy-aware designs and Virtualization Support
- Extended Support for MPI Tools Interface (as in MPI 3.0)
- Extended FT support
- Support for * features will be available in future MVAPICH2 Releases

Thank You!

subramoni.1@osu.edu

http://web.cse.ohio-state.edu/~subramon

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

The High-Performance Deep Learning Project
http://hidl.cse.ohio-state.edu/