MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

The MVAPICH2 Project
Latest Status and Future Plans

Presentation at MPICH BoF (SC '17)

by

Hari Subramoni

The Ohio State University

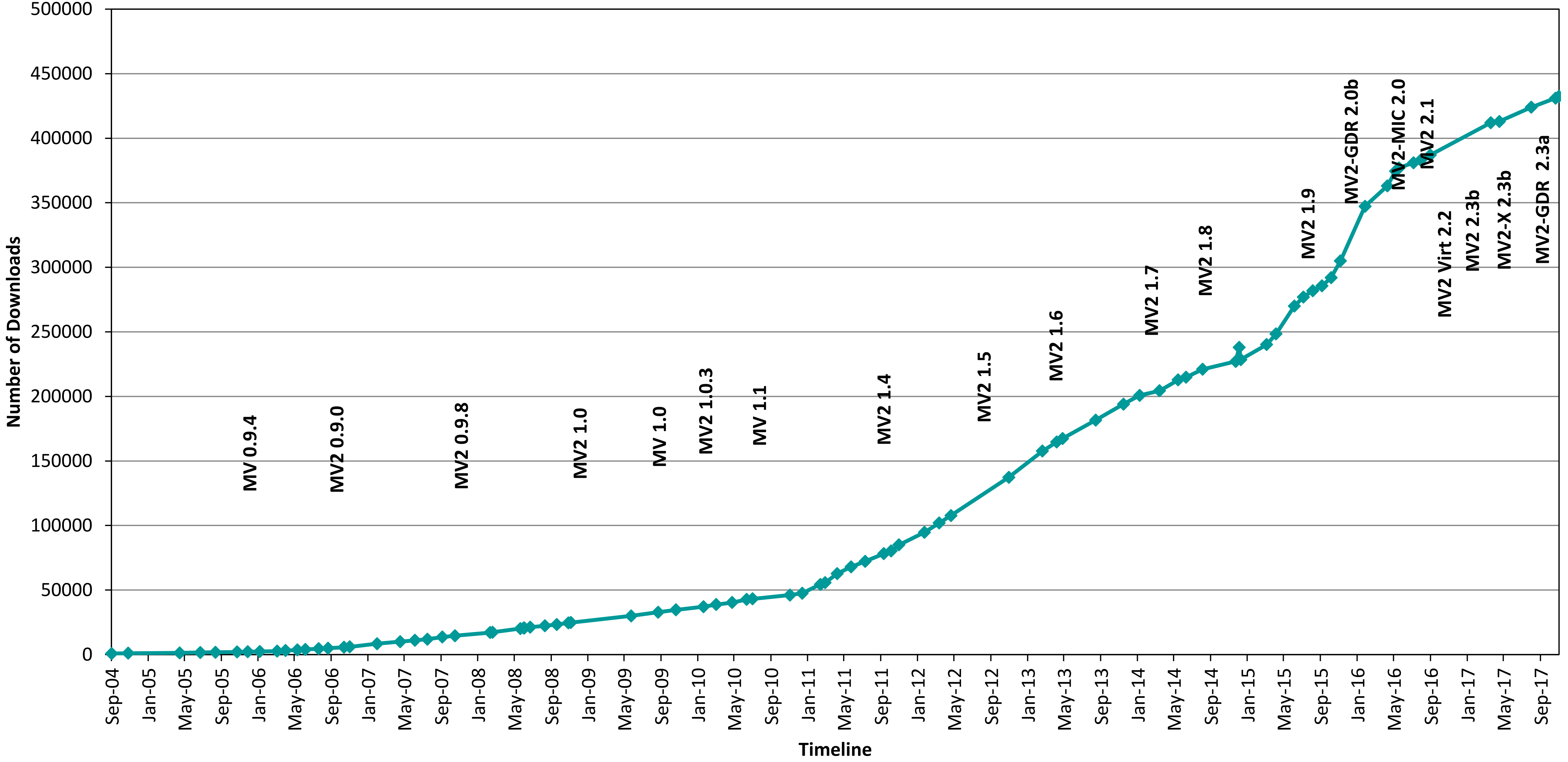
E-mail: subramon@cse.ohio-state.edu

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-3.0), 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 2,825 organizations in 85 countries**
  - **More than 433,000 (> 0.4 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (June '17 ranking)
    - **1st, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center in Wuxi, China**
    - 15th, 241,108-core (Pleiades) at NASA
    - 20th, 462,462-core (Stampede) at TACC
    - 44th, 74,520-core (Tsubame 2.5) at Tokyo Institute of Technology
  - Available with software stacks of many vendors and Linux Distros (RedHat and SuSE)
  - **http://mvapich.cse.ohio-state.edu**
- Empowering Top500 systems for over a decade
  - System-X from Virginia Tech (3rd in Nov 2003, 2,200 processors, 12.25 TFlops) ->
  - Sunway TaihuLight (1st in Jun'17, 10M cores, 100 PFlops)

MVAPICH2 Release Timeline and Downloads



MVAPICH2 Architecture

High Performance Parallel Programming Models

- Message Passing Interface (MPI)
- PGAS (UPC, OpenSHMEM, CAF, UPC++)
- Hybrid --- MPI + X (MPI + PGAS + OpenMP/Cilk)

High Performance and Scalable Communication Runtime

Diverse APIs and Mechanisms

- Point-to-point Primitives
- Collectives Algorithms
- Job Startup
- Energy-Awareness
- Remote Memory Access
- I/O and File Systems
- Fault Tolerance
- Virtualization
- Active Messages
- Introspection & Analysis

Support for Modern Networking Technology (InfiniBand, iWARP, RoCE, OmniPath)

Transport Protocols: RC, XRC, UD, DC

Modern Features: UMR, ODP*, SR-IOV, Multi Rail

Support for Modern Multi-/Many-core Architectures (Intel-Xeon, OpenPower, Xeon-Phi (MIC, KNL*), NVIDIA GPGPU)

Transport Mechanisms: Shared Memory, CMA, IVSHMEM

Modern Features: MCDRAM*, NVLink*, CAPI*

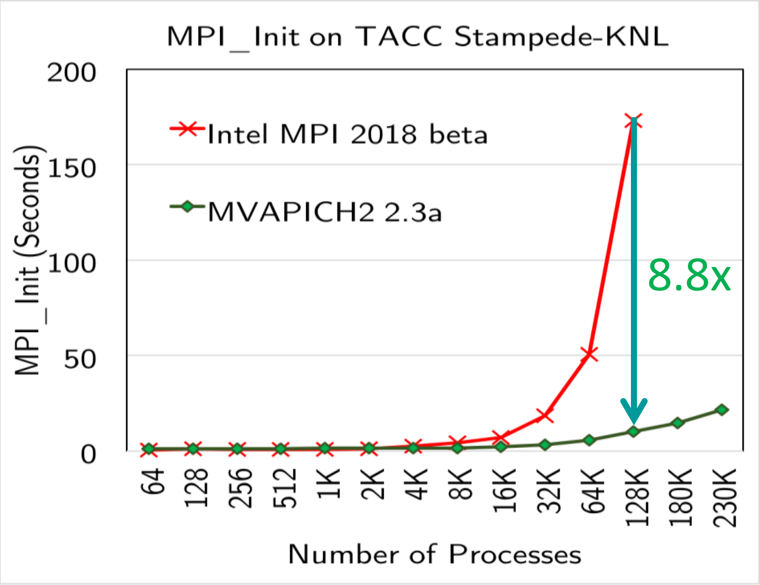
* Upcoming

MVAPICH2 Software Family

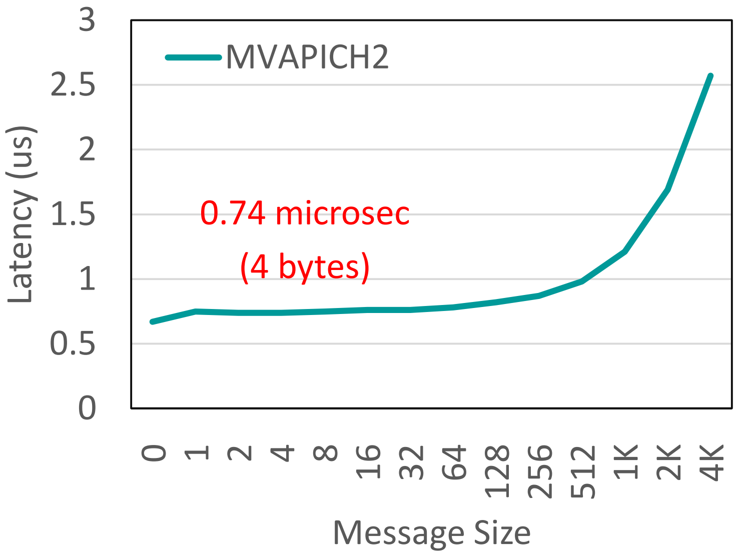
High-Performance Parallel Programming Libraries	
MVAPICH2	Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE
MVAPICH2-X	Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime
MVAPICH2-GDR	Optimized MPI for clusters 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

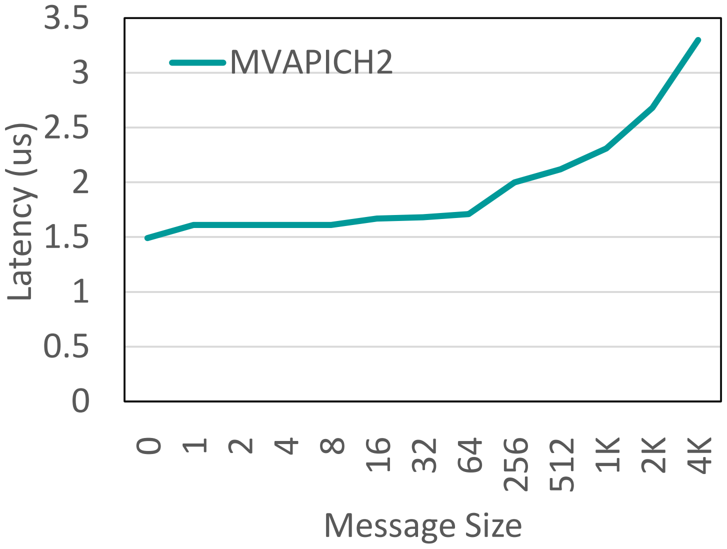
Fast Startup on Emerging Many-Cores



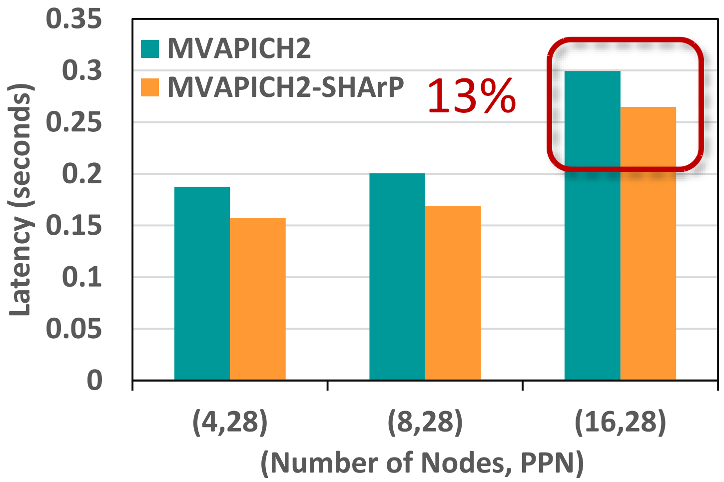
Enhanced Intra-node Performance for ARM



Enhanced Inter-node Performance for OpenPOWER



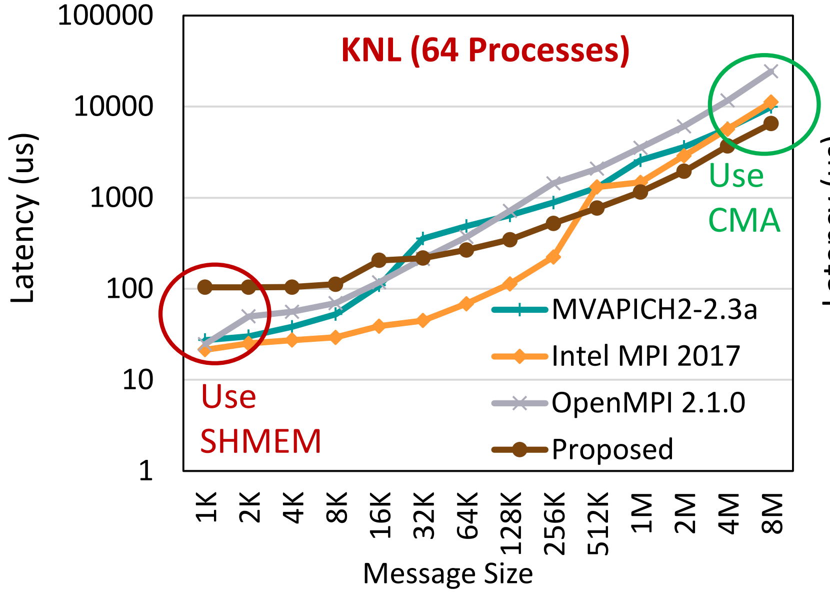
Advanced Allreduce with SHARP

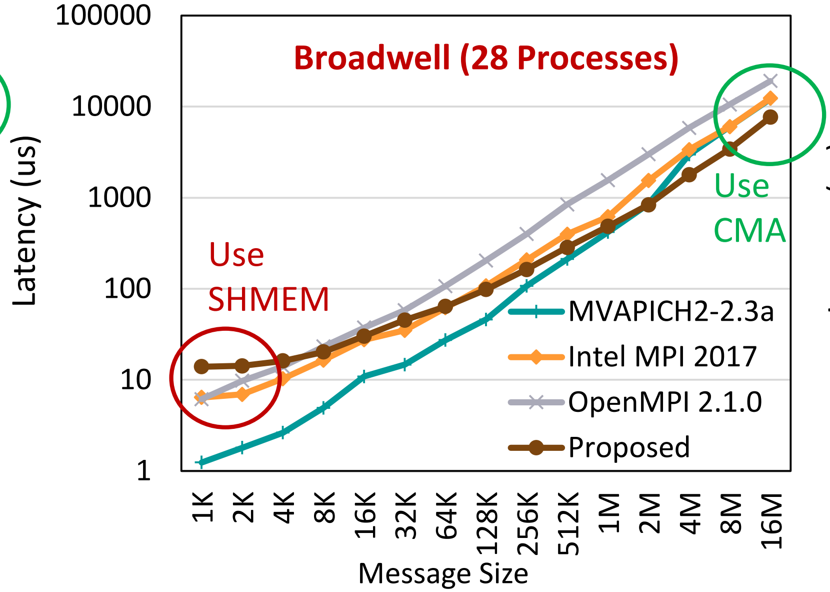


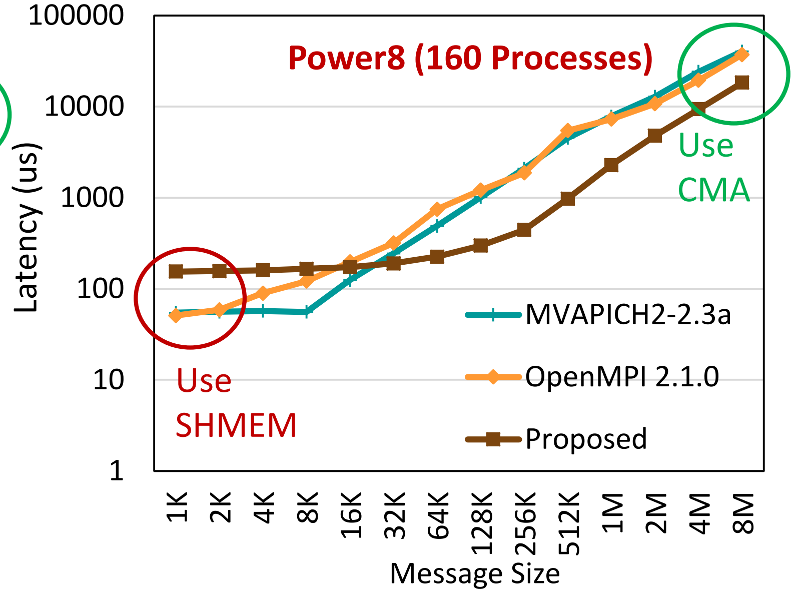
- Major Features and Enhancements in MVAPICH2 2.3b released on 08/10/2017
  - **Enhanced performance for point-to-point and RMA operations**
  - **Enhanced process to core mapping for many-cores**
  - **Improved support for emerging many-core architectures (ARM, OpenPOWER, KNL)**
  - **Improve launch time for large-scale jobs with mpirun_rsh**
  - **Add support for non-blocking Allreduce using Mellanox SHARP**
  - **Enhanced collective tuning for various Knight's Landing and Intel Omni-Path based systems**
    - **Bebop@ANL, Bridges@PSC, and Stampede2@TACC systems**
  - **Enhance support for MPI_T PVARs and CVARs**

MVAPICH2-X – Advanced MPI + PGAS + Tools

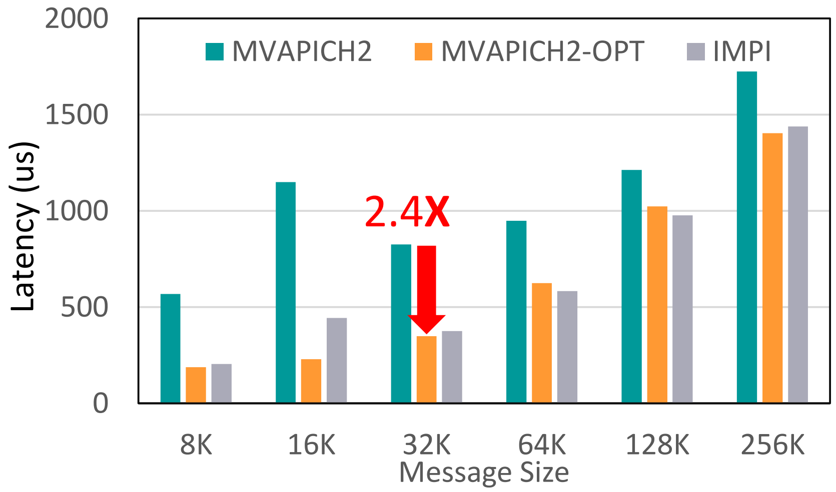
Enhanced MPI_Bcast for Emerging Many-Core Platforms with Optimized CMA-based Design







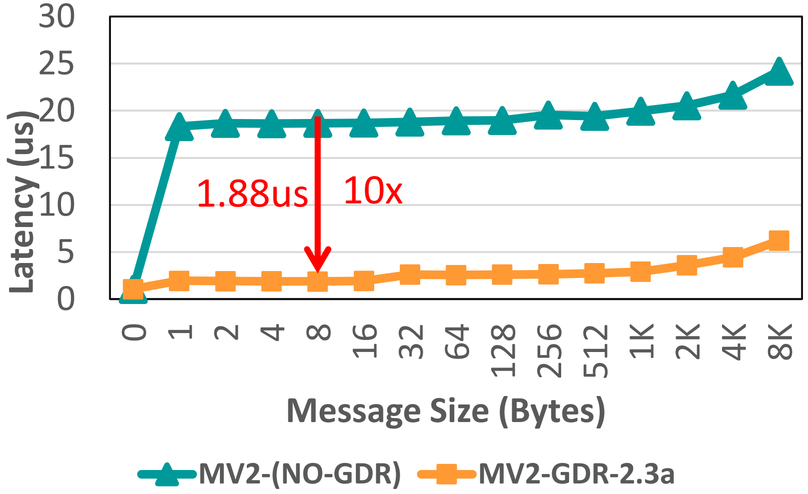
MPI_Allreduce On Stampede2 (10,240 Processes)



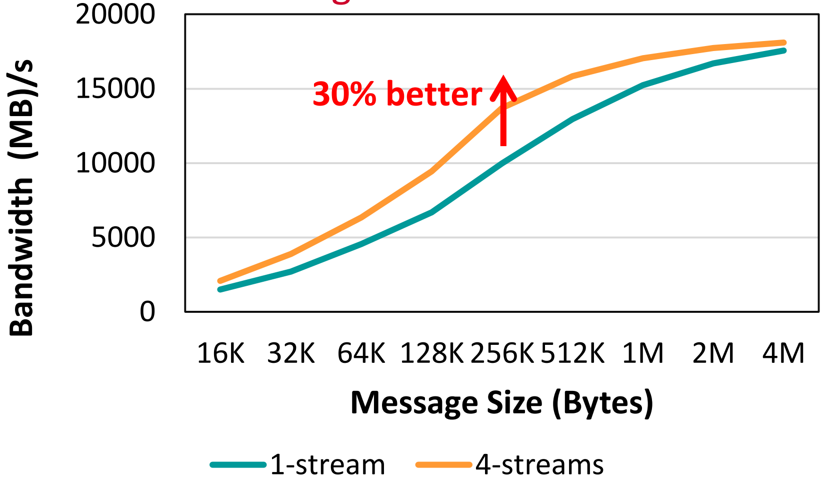
- **Major Features and Enhancements in MVAPICH2-X 2.3b released on 10/30/2017**
  - **MPI Features**
    - **Based on MVAPICH2 2.3b**
    - **Optimized support for Skylake, ARM, and OpenPOWER architecture**
  - **MPI (Advanced) Features**
    - **Support Data Partitioning-based Multi-Leader Design (DPML) for MPI collectives**
    - **Support Contention Aware Kernel-Assisted MPI collectives**
    - **Support for OSU InfiniBand Network Analysis and Management (OSU INAM) Tool v0.9.2**
  - **OpenSHMEM Features**
    - **Based on OpenSHMEM reference implementation 1.3**
    - **Support Non-Blocking remote memory access routines**

MVAPICH2-GDR – Optimized MPI for clusters with NVIDIA GPUs

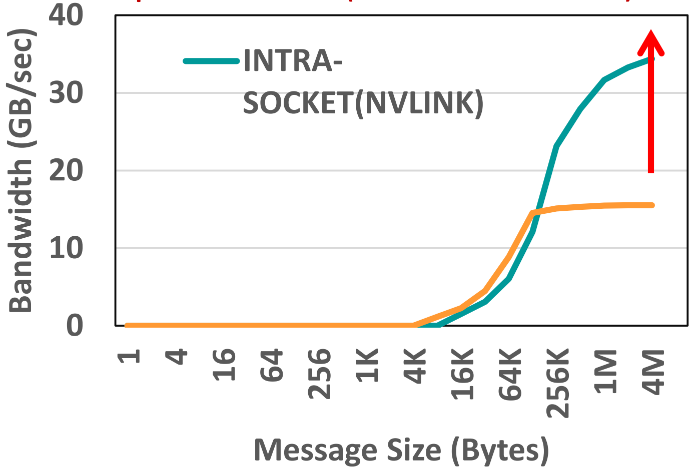
Best Performance for GPU-based Transfers



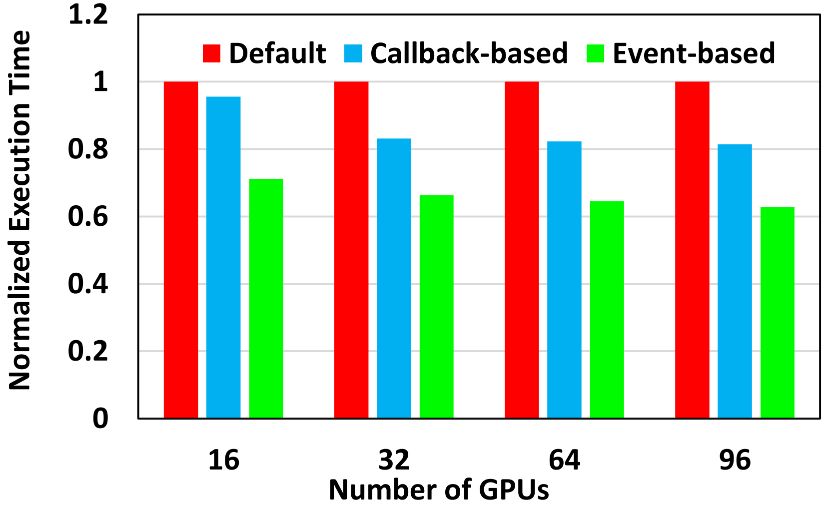
Multi-stream Communication using CUDA IPC on DGX-1



Intra-node Performance on OpenPOWER (NVLink + Pascal)

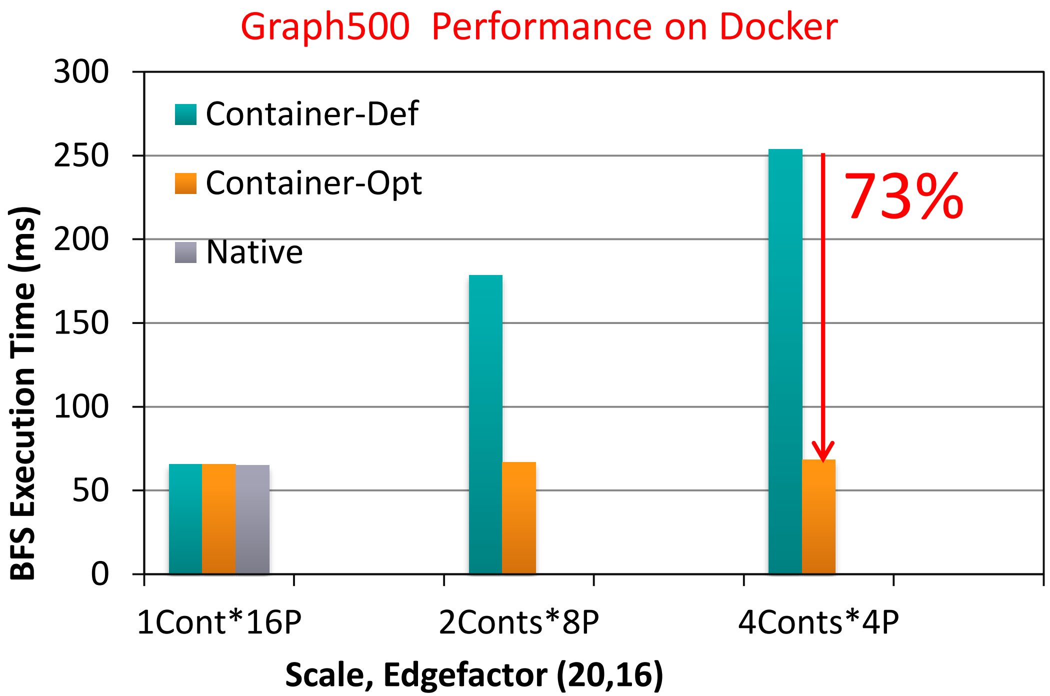


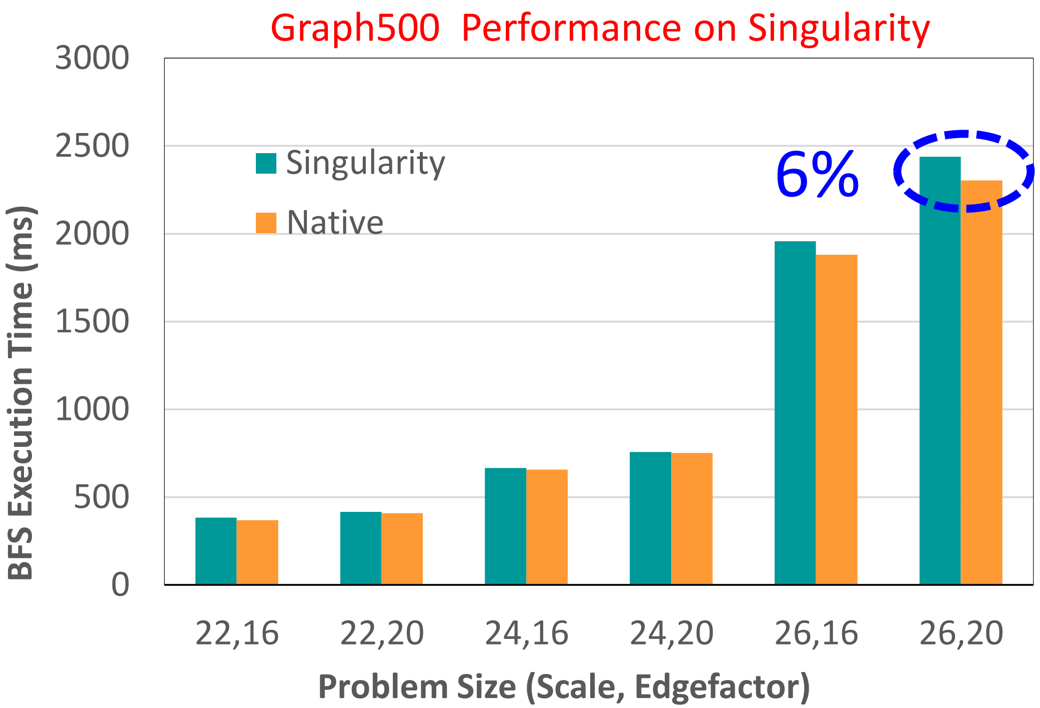
Enhanced Kernel-based Datatype Processing



- Major Features and Enhancements in MVAPICH2-GDR 2.3a released on 11/09/2017
  - **Support for CUDA 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**
  - **Enhanced performance of MPI_Allreduce for GPU-resident data**
  - **InfiniBand Multicast based designs for GPU-based broadcast and streaming applications**
  - **Efficient 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 cores
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF …)
  - MPI + Task*
- Enhanced Optimization for GPU Support and Accelerators
- Taking advantage of advanced features of Mellanox InfiniBand
  - Multi-host Adapters*
  - Hardware-based Tag Matching*
- Enhanced communication schemes for upcoming architectures
  - Knights Landing with MCDRAM*
  - CAPI*
- Extended topology-aware collectives
- Extended Energy-aware designs and Virtualization Support
- Extended Support for MPI Tools Interface (as in MPI 3.1)
- Extended Checkpoint-Restart and migration support with SCR
- 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/

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

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