MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiBD
High-Performance Big Data

HiDL
High-Performance Deep Learning

Accelerating HPC, Big Data and Deep Learning on OpenPOWER Platforms

Talk at OpenPOWER Academic Discussion Group Workshop 2019

by

Dhabaleswar K. (DK) Panda

The Ohio State University

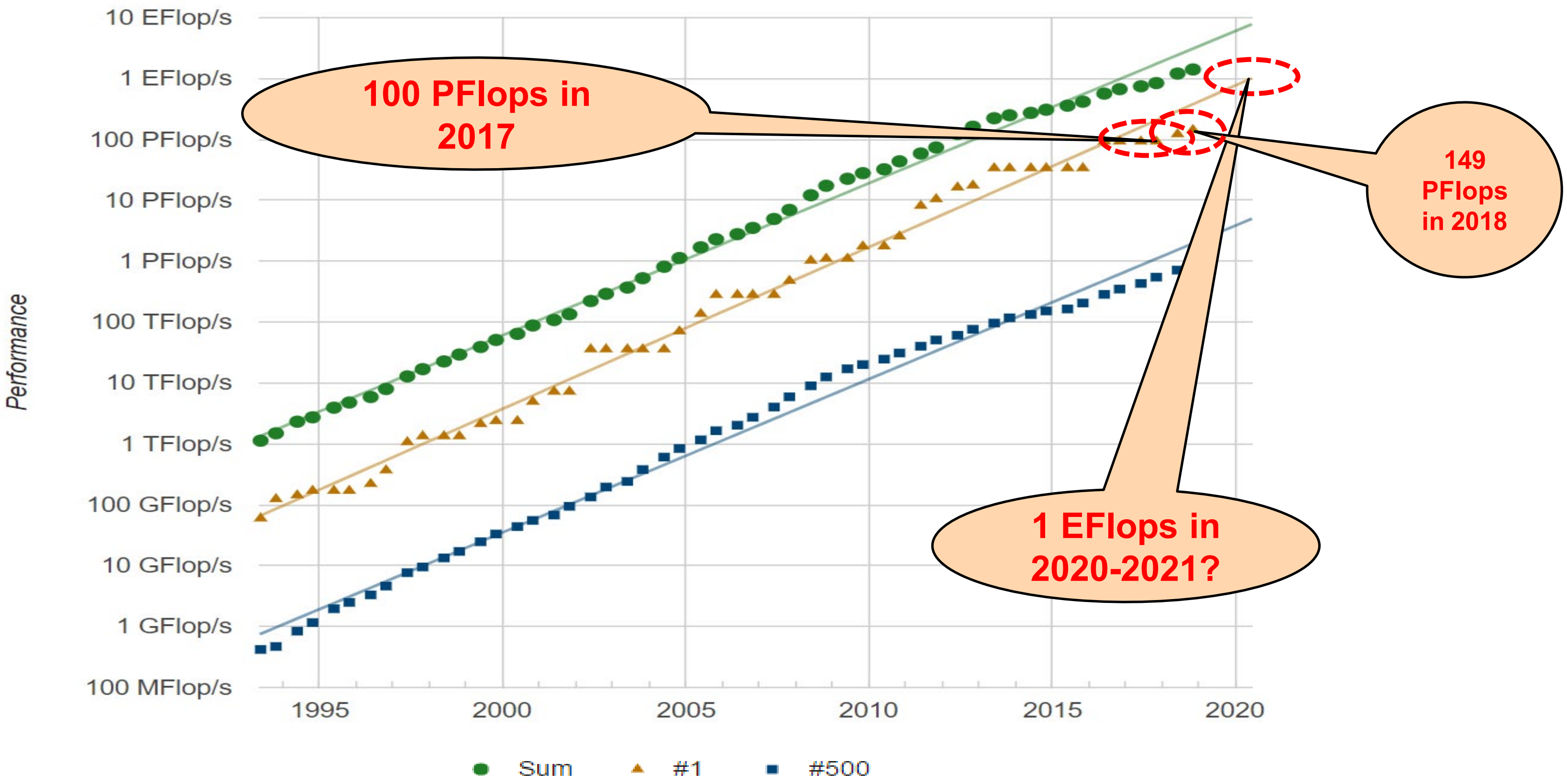
E-mail: panda@cse.ohio-state.edu

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

Follow us on

https://twitter.com/mvapich

High-End Computing (HEC): PetaFlop to ExaFlop

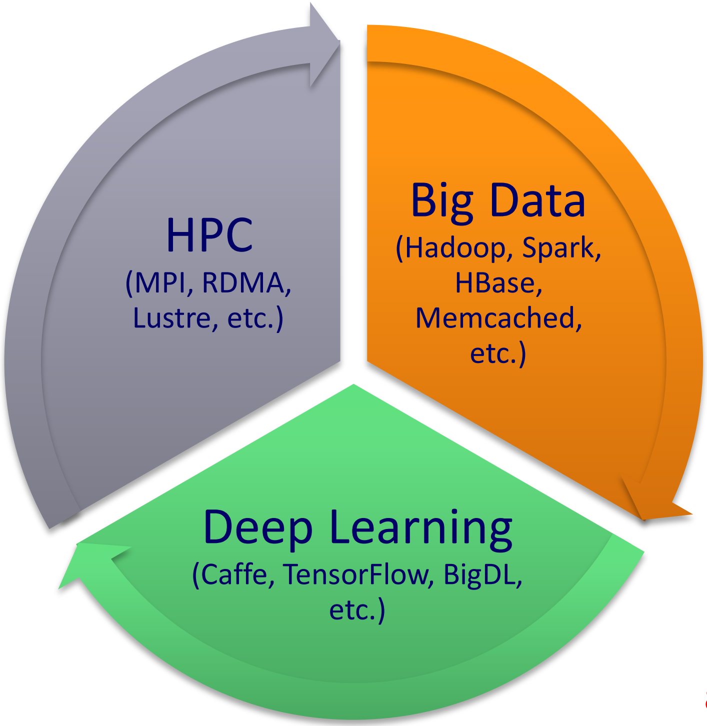


Expected to have an ExaFlop system in 2020-2021!

Presentation Overview

- Challenges in Designing Convergent HPC, Big Data and Deep Learning Architectures
- MVAPICH Project - MPI and PGAS (MVAPICH) Library with CUDA-Awareness
- HiDL Project – High-Performance Deep Learning
- HiBD Project – High-Performance Big Data Analytics Library
- Commercial Support from X-ScaleSolutions
- Conclusions and Q&A

Increasing Usage of HPC, Big Data and Deep Learning



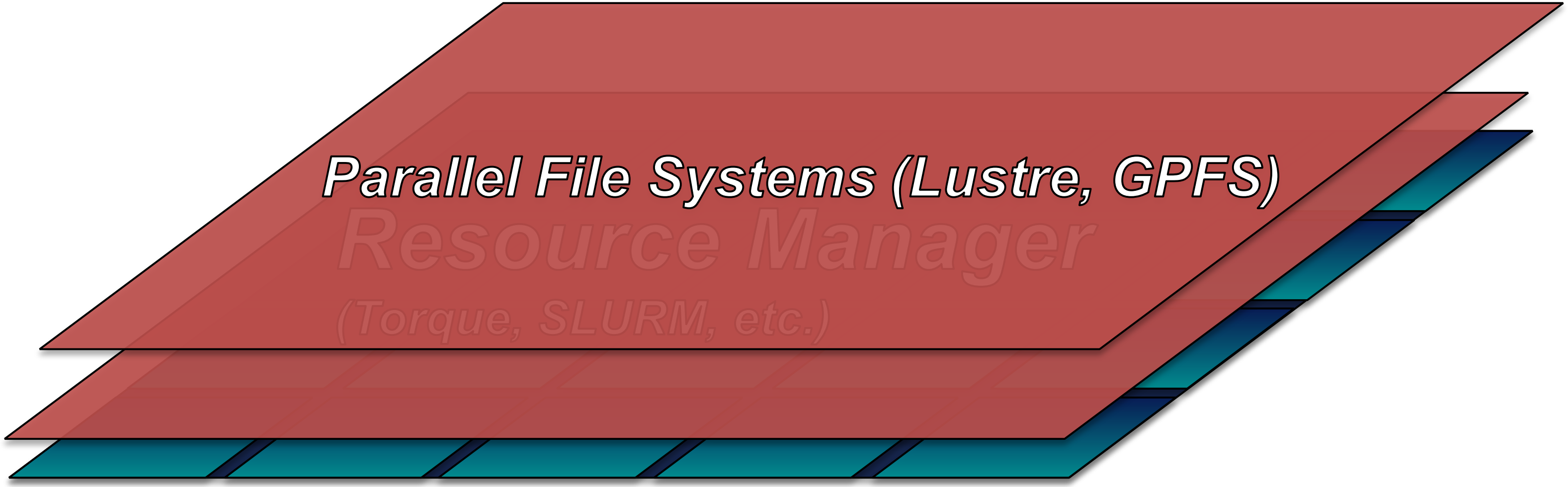
Convergence of HPC, Big Data, and Deep Learning!

Increasing Need to Run these applications on the Cloud!!

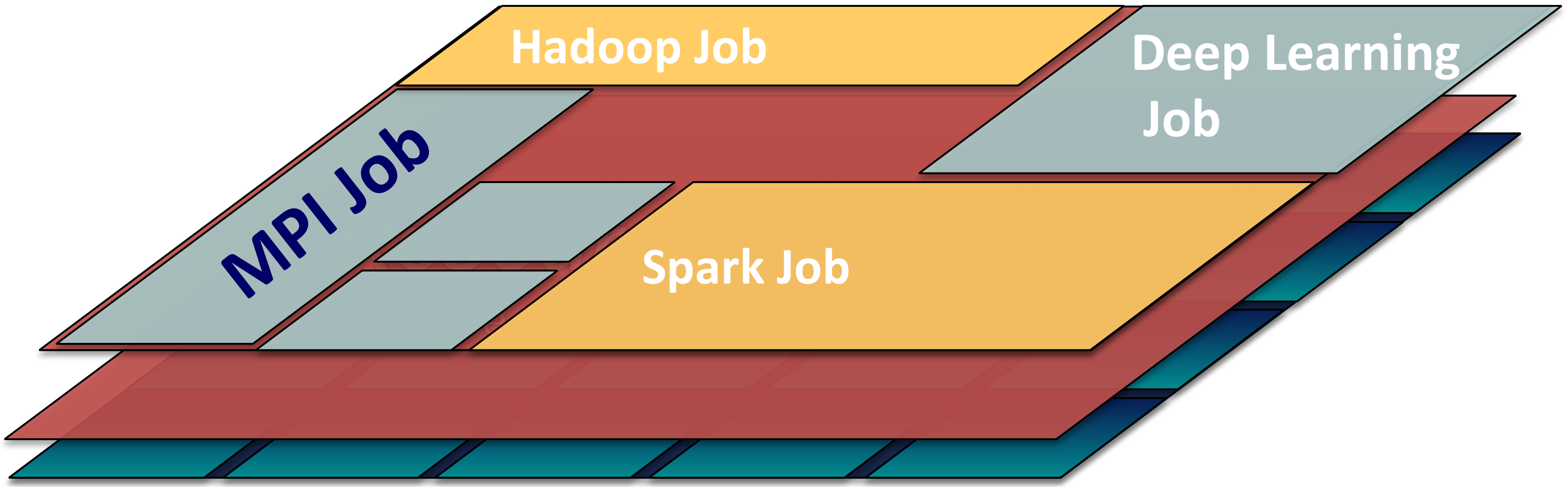
Can We Run HPC, Big Data and Deep Learning Jobs on Existing HPC Infrastructure?

Can We Run HPC, Big Data and Deep Learning Jobs on Existing HPC Infrastructure?

Can We Run HPC, Big Data and Deep Learning Jobs on Existing HPC Infrastructure?



Can We Run HPC, Big Data and Deep Learning Jobs on Existing HPC Infrastructure?



Presentation Overview

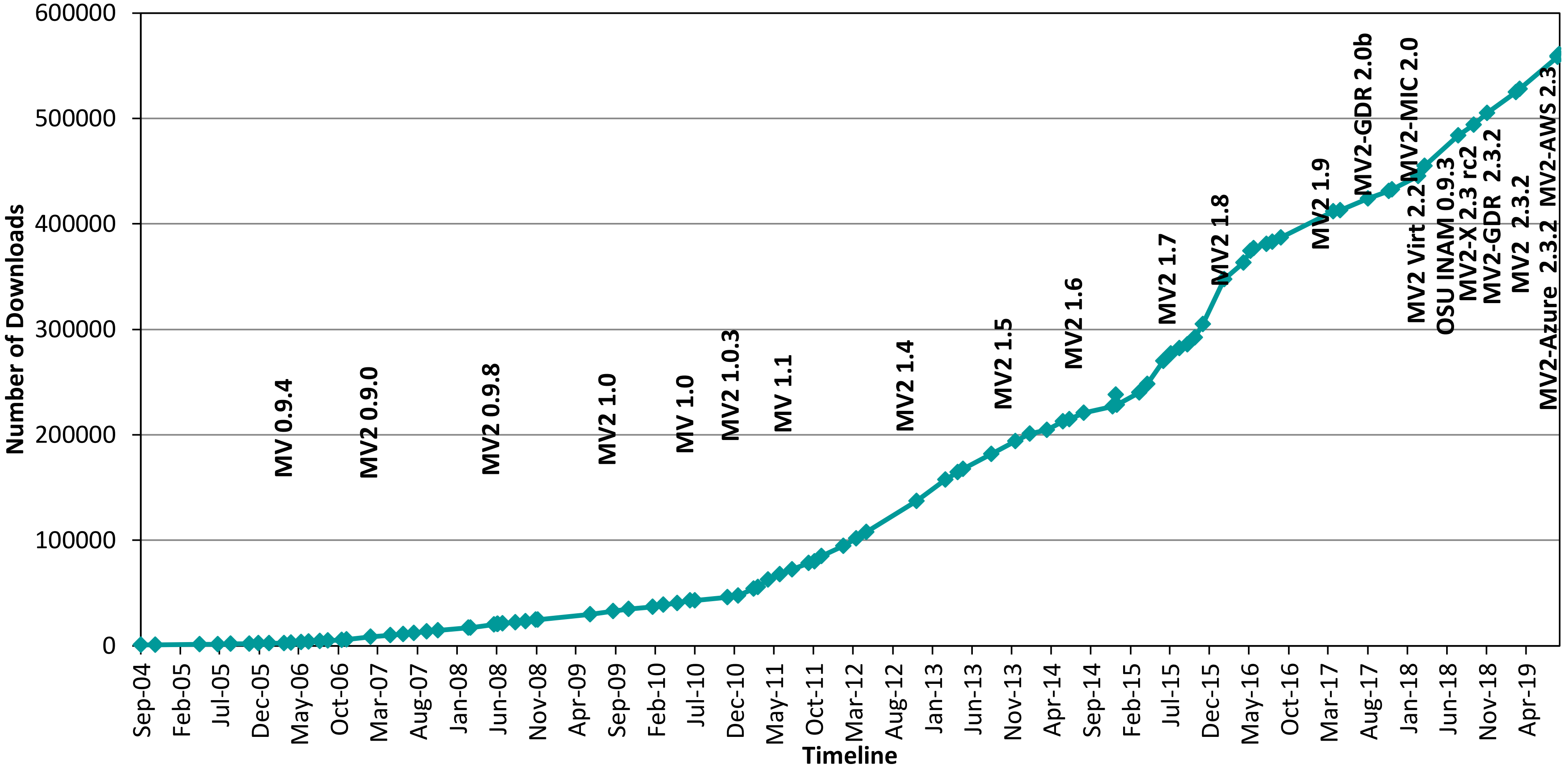
- Challenges in Designing Convergent HPC, Big Data and Deep Learning Architectures
- **MVAPICH Project – MPI and PGAS (MVAPICH) Library with CUDA-Awareness**
- HiDL Project – High-Performance Deep Learning
- HiBD Project – High-Performance Big Data Analytics Library
- Commercial Support from X-ScaleSolutions
- Conclusions and Q&A

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 (SC '02)
  - 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 3,050 organizations in 89 countries**
  - **More than 615,000 (> 0.6 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (June '19 ranking)
    - 3^{rd}, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center in Wuxi, China
    - 5^{th}, 448, 448 cores (Frontera) at TACC
    - 8^{th}, 391,680 cores (ABCI) in Japan
    - 15^{th}, 570,020 cores (Neurion) in South Korea and many others
  - Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, and OpenHPC)
  - **http://mvapich.cse.ohio-state.edu**
- Empowering Top500 systems for over a decade

Partner in the TACC Frontera System

MVAPICH2 Release Timeline and Downloads



Architecture of MVAPICH2 Software Family (MPI, PGAS and DL)

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, Omni-Path, Elastic Fabric Adapter)

Transport Protocols: RC, SRD, UD, DC

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

Support for Modern Multi-/Many-core Architectures (Intel-Xeon, OpenPOWER, Xeon-Phi, ARM, NVIDIA GPGPU)

Transport Mechanisms: Shared Memory, CMA, IVSHMEM, XPMEM

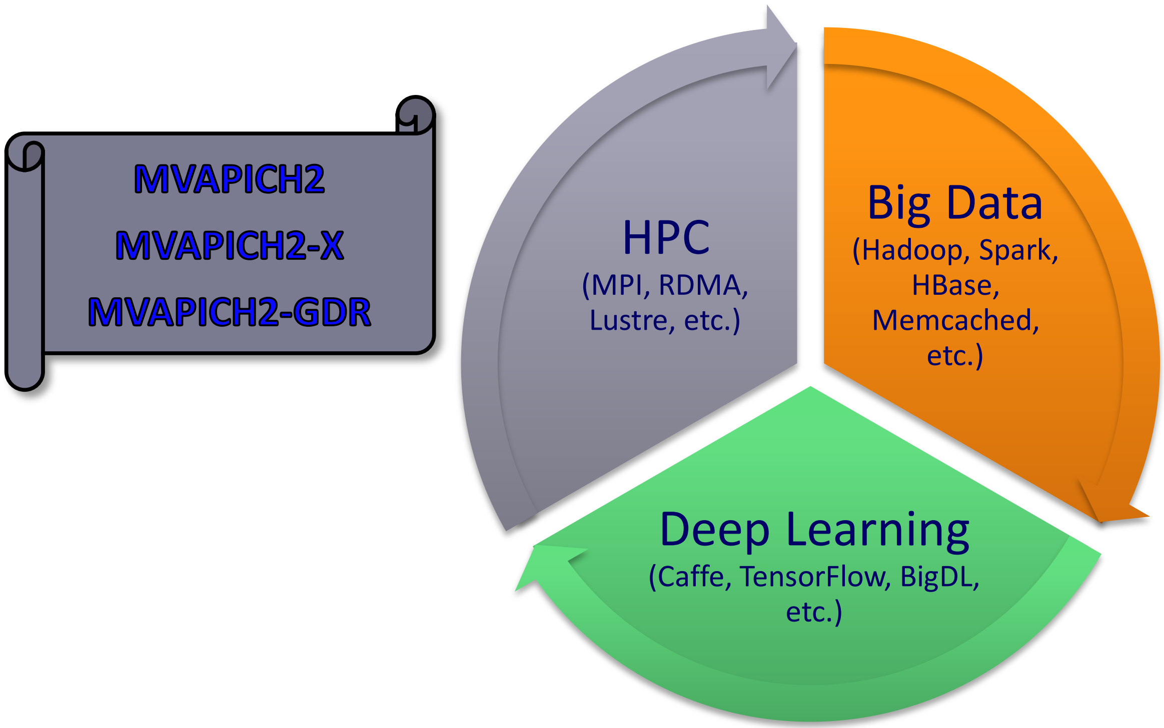
Modern Features: Optane*, NVLink, CAPI*

* Upcoming

MVAPICH2 Software Family

Requirements	Library
MPI with IB, iWARP, Omni-Path, and RoCE	MVAPICH2
Advanced MPI Features/Support, OSU INAM, PGAS and MPI+PGAS with IB, Omni-Path, and RoCE	MVAPICH2-X
MPI with IB, RoCE & GPU and Support for Deep Learning	MVAPICH2-GDR
HPC Cloud with MPI & IB	MVAPICH2-Virt
Energy-aware MPI with IB, iWARP and RoCE	MVAPICH2-EA
MPI Energy Monitoring Tool	OEMT
InfiniBand Network Analysis and Monitoring	OSU INAM
Microbenchmarks for Measuring MPI and PGAS Performance	OMB

Convergent Software Stacks for HPC, Big Data and Deep Learning



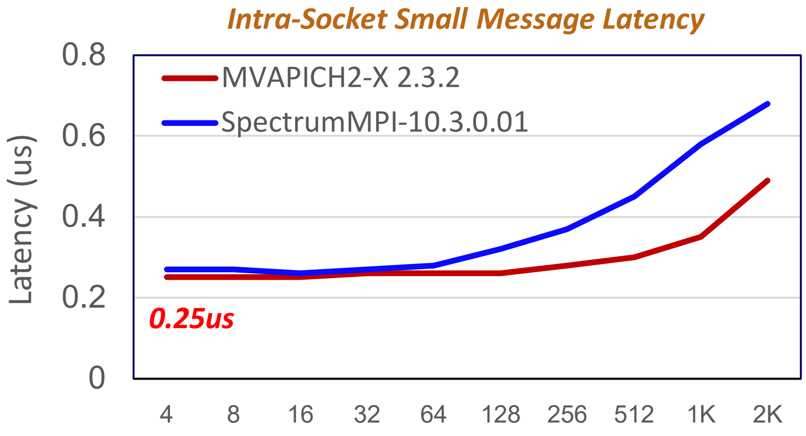
OpenPOWER Platform Support in MVAPICH2 Libraries

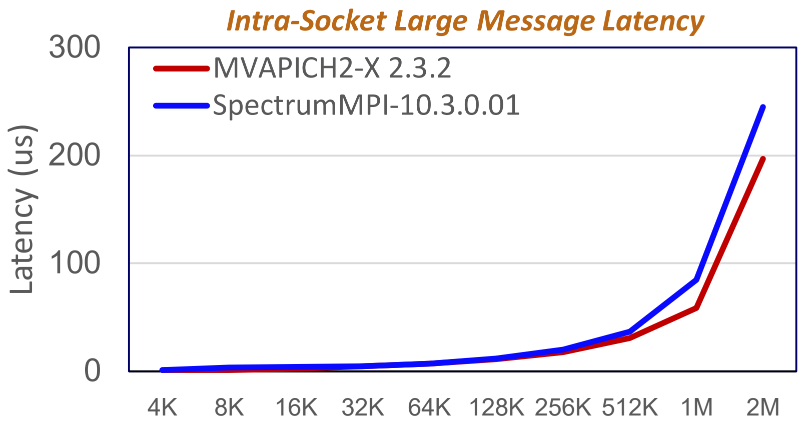
- MVAPICH2
  - Basic MPI support
    - since MVAPICH2 2.2rc1 (March 2016)
- MVAPICH2-X
  - PGAS (OpenSHMEM and UPC) and Hybrid MPI+PGAS support
    - since MVAPICH2-X 2.2b (March 2016)
  - Advanced Collective Support with CMA
    - Since MVAPICH2-X 2.3b (Oct 2017)
- MVAPICH2-GDR
  - NVIDIA GPGPU support with NVLink (CORAL systems like Summit and Sierra)
    - Since MVAPICH2-GDR 2.3a (Nov 2017)

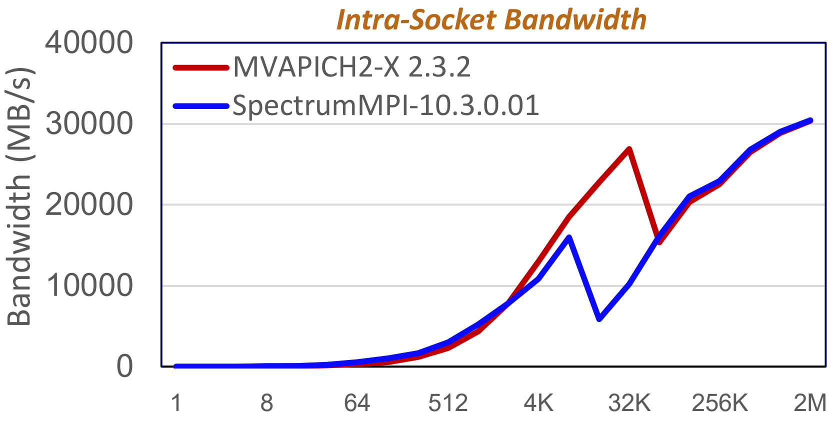
MPI, PGAS and Deep Learning Support for OpenPOWER

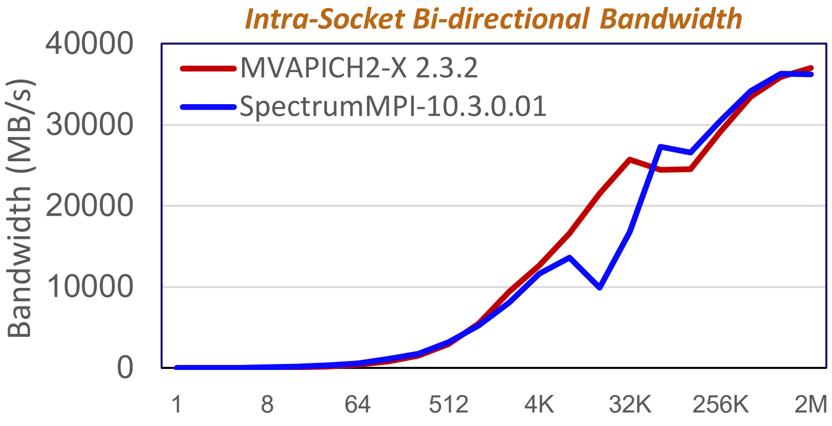
- **Message Passing Interface (MPI) Support**
  - **Point-to-point Inter-node and Intra-node**
  - **XPMEM-based collectives**
- Exploiting Accelerators (NVIDIA GPGPUs)
  - CUDA-aware MPI
  - Point-to-point
  - Applications
  - Integrated Support with TAU

Intra-node Point-to-Point Performance on OpenPOWER



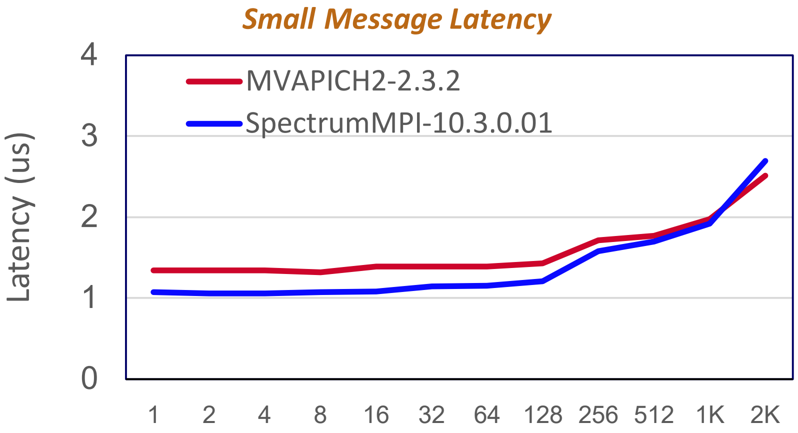


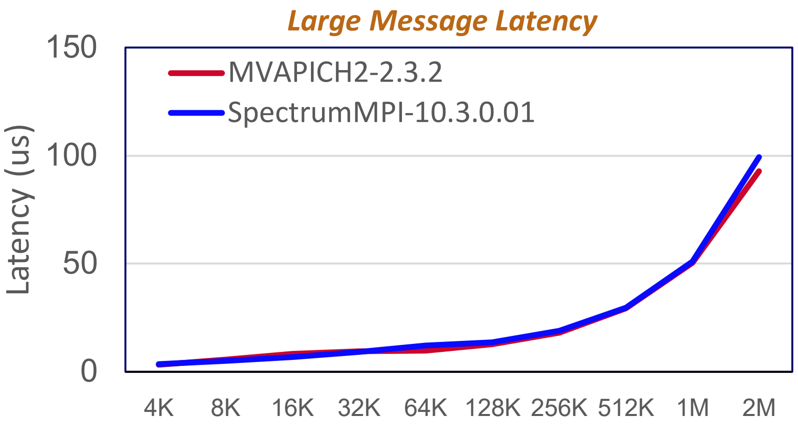


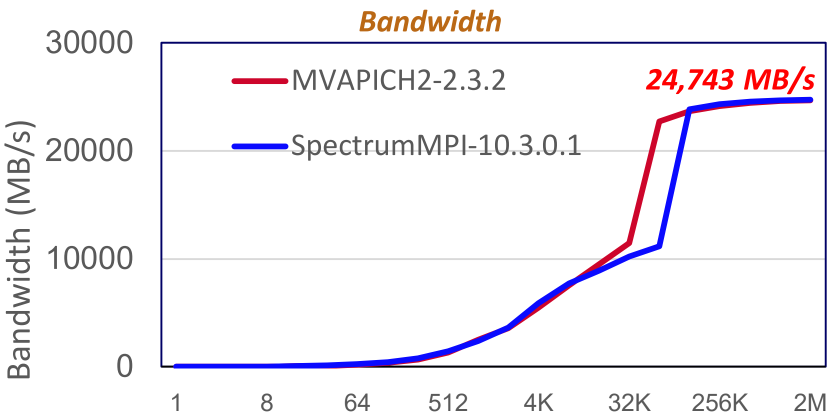


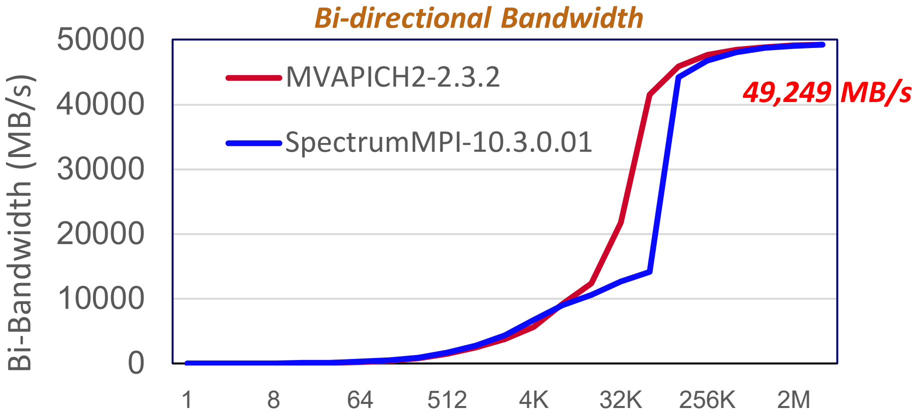
Platform: Two nodes of OpenPOWER (Power9-ppc64le) CPU using Mellanox EDR (MT4121) HCA

Inter-node Point-to-Point Performance on OpenPower



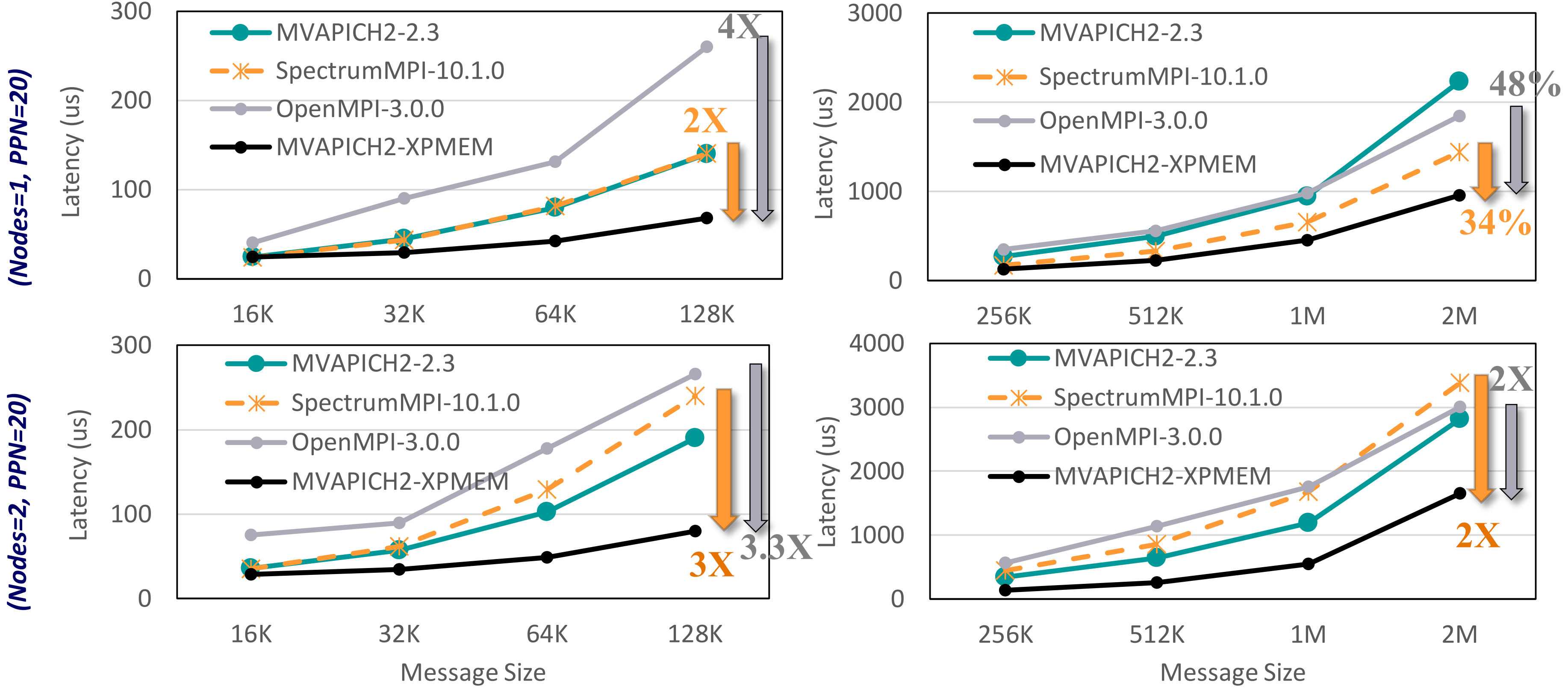






Platform: Two nodes of OpenPOWER (POWER9-ppc64le) CPU using Mellanox EDR (MT4121) HCA

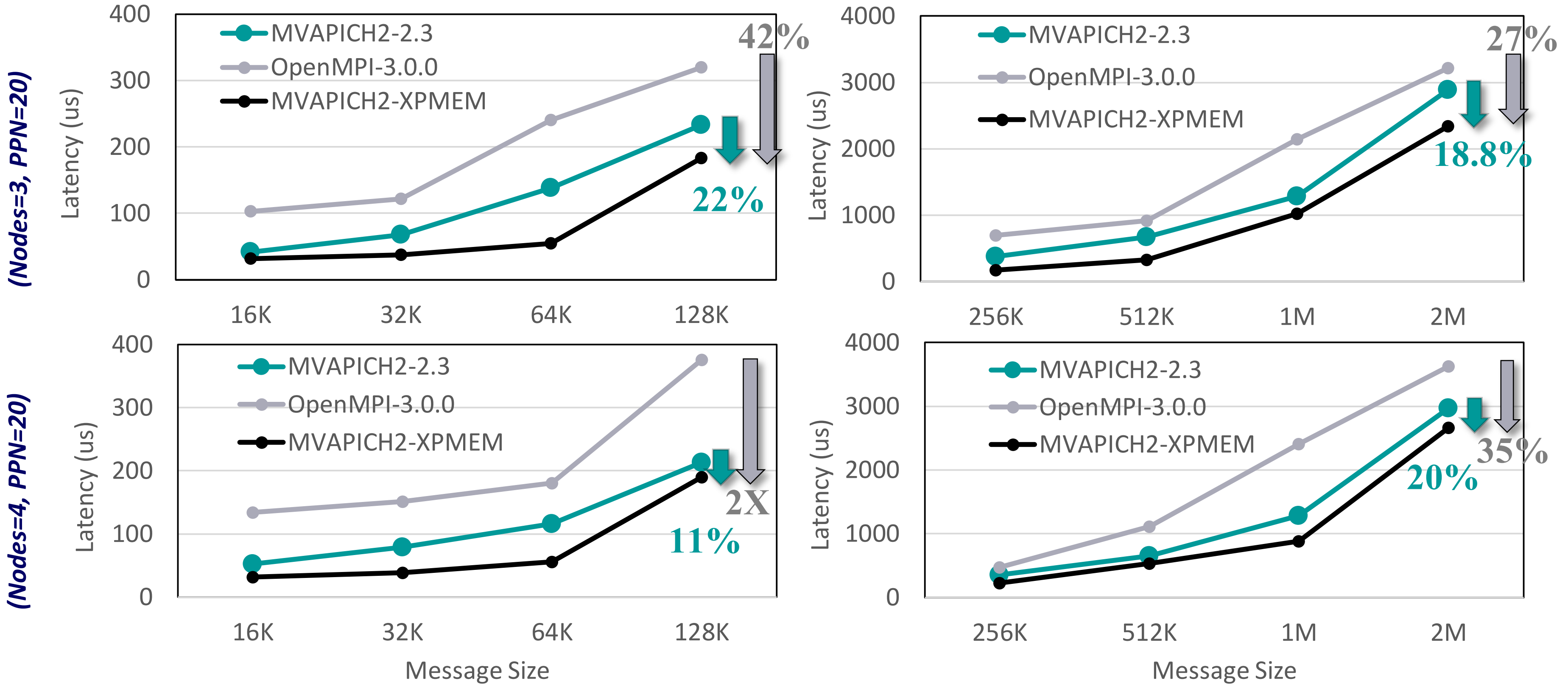
Optimized MVAPICH2 All-Reduce with XPMEM



- **Optimized MPI All-Reduce Design in MVAPICH2**
  - *Up to 2X* performance improvement over Spectrum MPI and 4X over OpenMPI for intra-node

Optimized Runtime Parameters: MV2_CPU_BINDING_POLICY=hybrid MV2_HYBRID_BINDING_POLICY=bunch

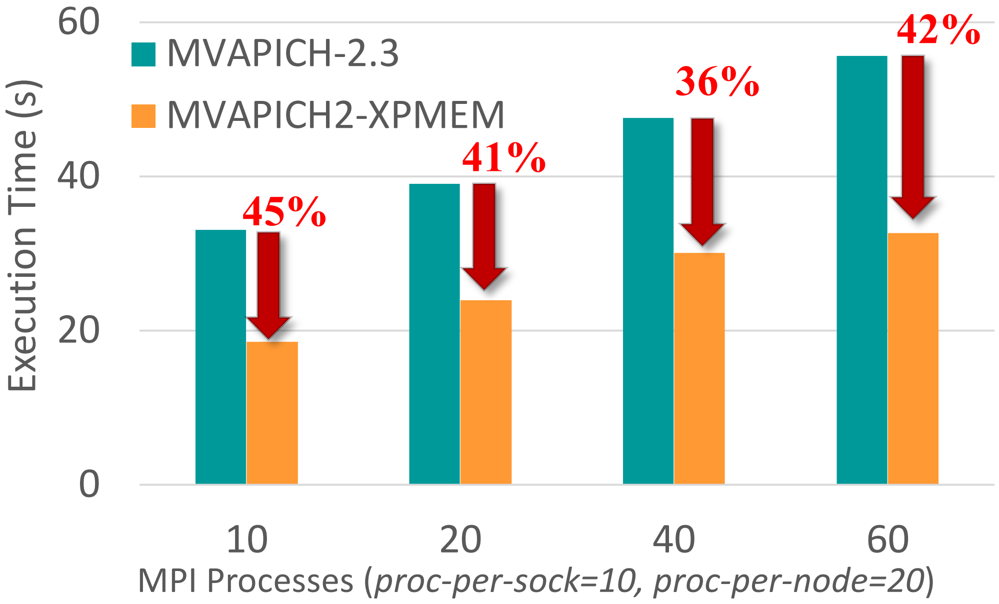
Optimized MVAPICH2 All-Reduce with XPMEM



- **Optimized MPI All-Reduce Design in MVAPICH2**
  - *Up to 2X* performance improvement over OpenMPI for inter-node. (Spectrum MPI didn't run for >2 processes)

Optimized Runtime Parameters: MV2_CPU_BINDING_POLICY=hybrid MV2_HYBRID_BINDING_POLICY=bunch

MiniAMR Performance using Optimized XPMEM-based Collectives



- **MiniAMR application execution time comparing MVAPICH2-2.3rc1 and optimized All-Reduce design**
  - ***Up to 45% improvement over MVAPICH2-2.3rc1 in mesh-refinement time of MiniAMR application for weak-scaling workload on up to four POWER8 nodes.***

Optimized Runtime Parameters: MV2_CPU_BINDING_POLICY=hybrid MV2_HYBRID_BINDING_POLICY=scatter

MPI, PGAS and Deep Learning Support for OpenPOWER

- Message Passing Interface (MPI) Support
  - Point-to-point Inter-node and Intra-node
  - XPMEM-based collectives
- **Exploiting Accelerators (NVIDIA GPGPUs)**
  - **CUDA-aware MPI**
  - **Point-to-point**
  - **Applications**
  - **Integrated Support with TAU**

GPU-Aware (CUDA-Aware) MPI Library: MVAPICH2-GPU

- Standard MPI interfaces used for unified data movement
- Takes advantage of Unified Virtual Addressing (>= CUDA 4.0)
- Overlaps data movement from GPU with RDMA transfers

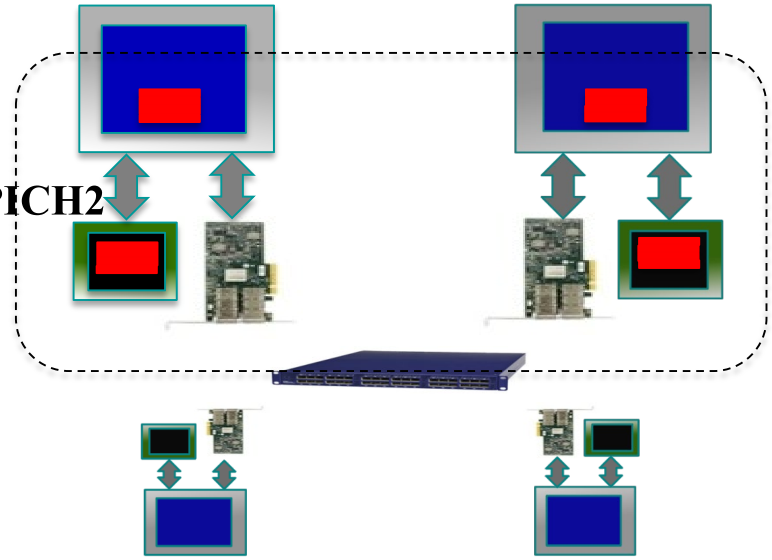
At Sender:

```
MPI_Send(s_devbuf, size, …);
```

At Receiver:

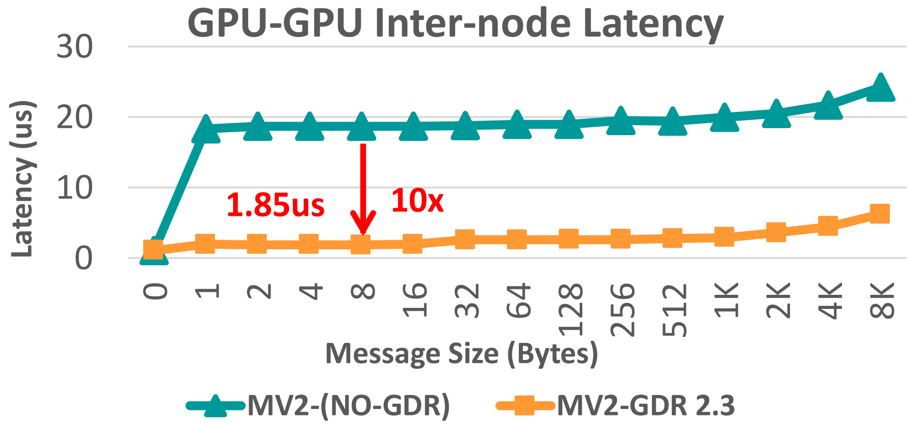
```
MPI_Recv(r_devbuf, size, …);
```

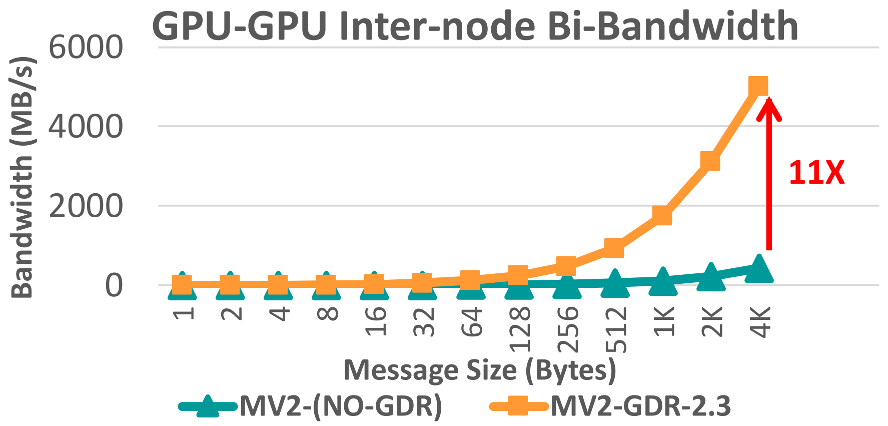
inside MVAPICH2

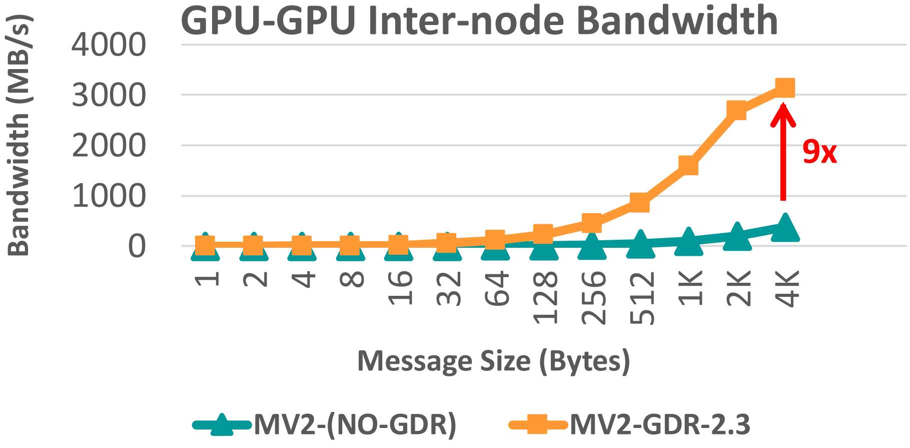


High Performance and High Productivity

Optimized MVAPICH2-GDR Design (D-D) Performance

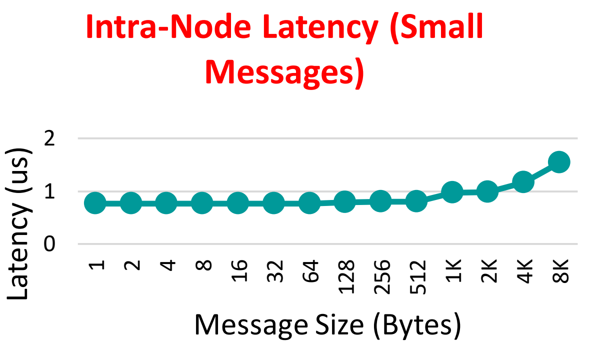


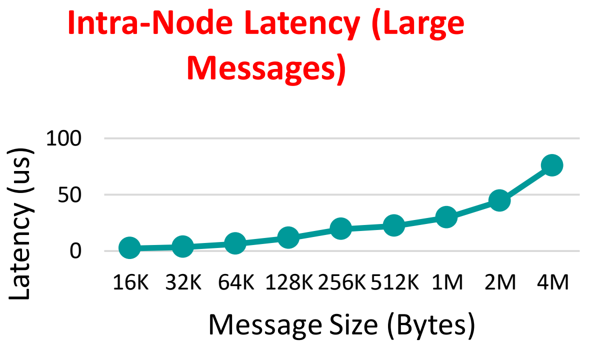


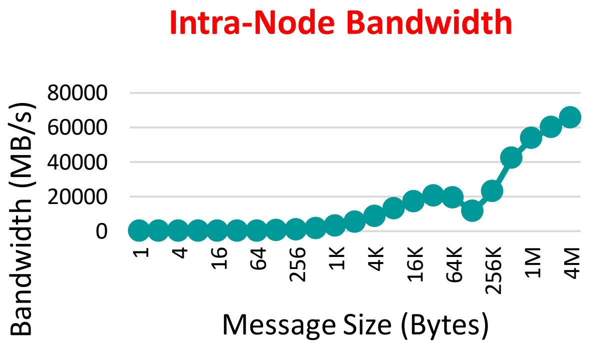


MVAPICH2-GDR-2.3
Intel Haswell (E5-2687W @ 3.10 GHz) node - 20 cores
NVIDIA Volta V100 GPU
Mellanox Connect-X4 EDR HCA
CUDA 9.0
Mellanox OFED 4.0 with GPU-Direct-RDMA

D-to-D Performance on OpenPOWER w/ GDRCopy (NVLink2 + Volta)

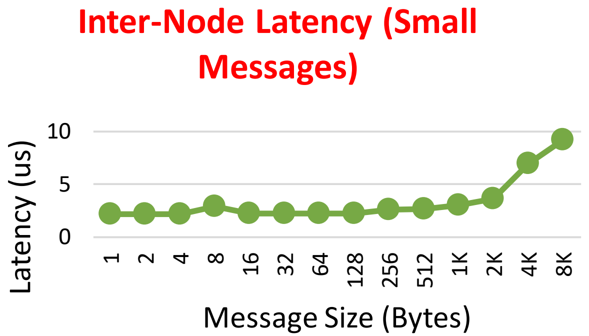


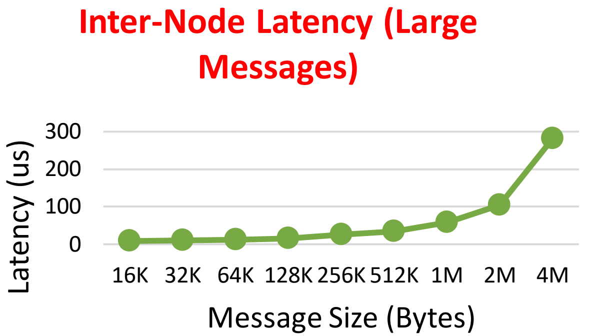


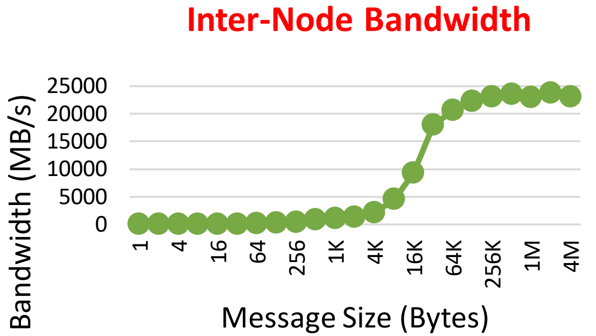


*Intra-node Latency: **0.76 us** (with GDRCopy)*

*Intra-node Bandwidth: **65.48 GB/sec** for 4MB (via NVLINK2)*







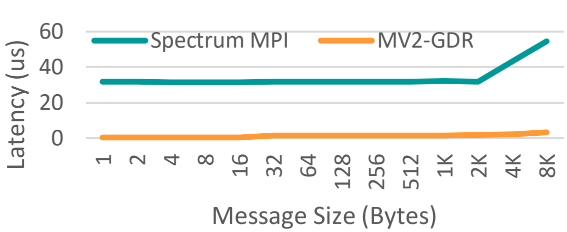
*Inter-node Latency: **2.18 us** (with GDRCopy 2.0)*

*Inter-node Bandwidth: **23 GB/sec** for 4MB (via 2 Port EDR)*

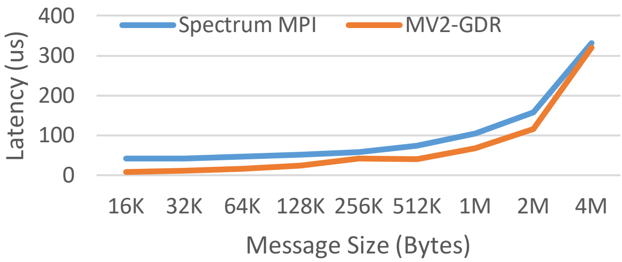
Platform: OpenPOWER (POWER9-ppc64le) nodes equipped with a dual-socket CPU, 4 Volta V100 GPUs, and 2port EDR InfiniBand Interconnect

D-to-H & H-to-D Performance on OpenPOWER w/ GDRCopy (NVLink2 + Volta)

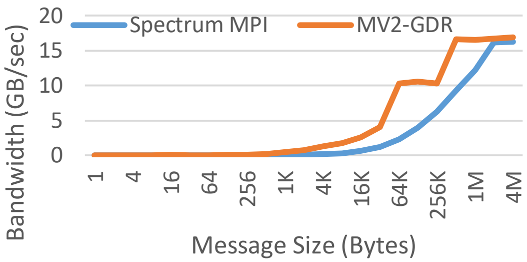
D-H INTRA-NODE LATENCY (SMALL)



D-H INTRA-NODE LATENCY (LARGE)



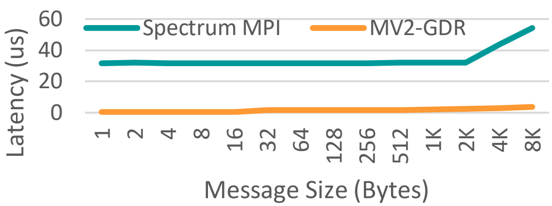
D-H INTRA-NODE BW



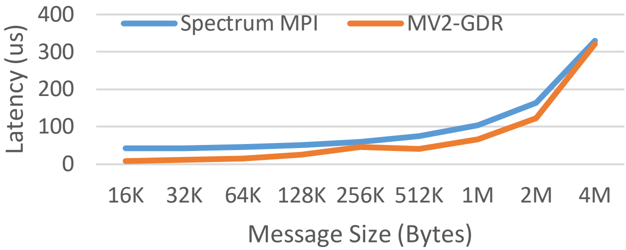
Intra-node D-H Latency: 0.49 us (with GDRCopy)

Intra-node D-H Bandwidth: 16.70 GB/sec for 2MB (via NVLINK2)

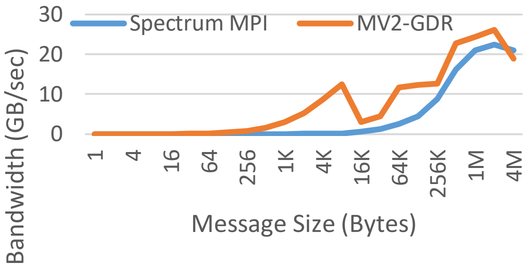
H-D INTRA-NODE LATENCY (SMALL)



H-D INTRA-NODE LATENCY (LARGE)



H-D INTRA-NODE BW

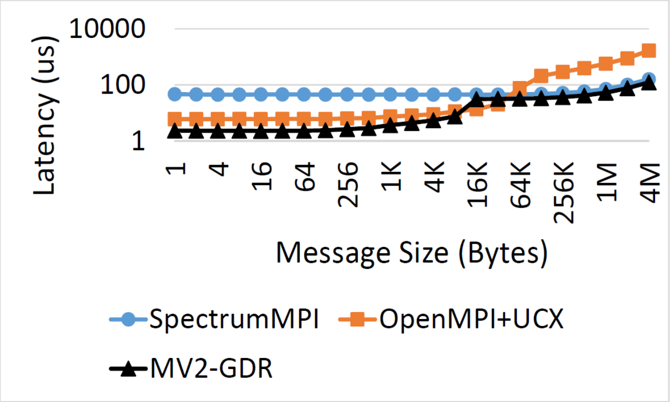


Intra-node H-D Latency: 0.49 us (with GDRCopy 2.0)

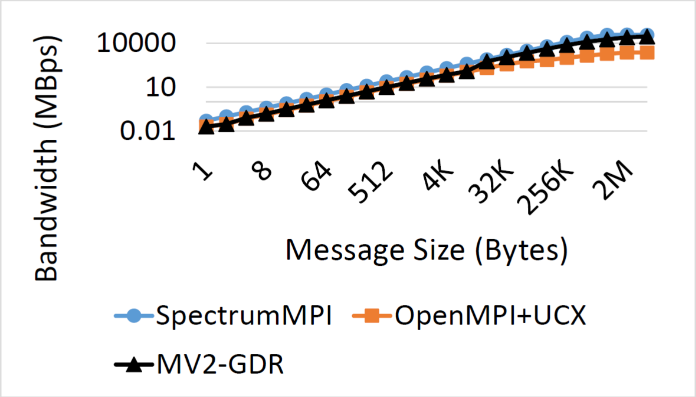
Intra-node H-D Bandwidth: 26.09 GB/sec for 2MB (via NVLINK2)

Platform: OpenPOWER (POWER9-ppc64le) nodes equipped with a dual-socket CPU, 4 Volta V100 GPUs, and 2port EDR InfiniBand Interconnect

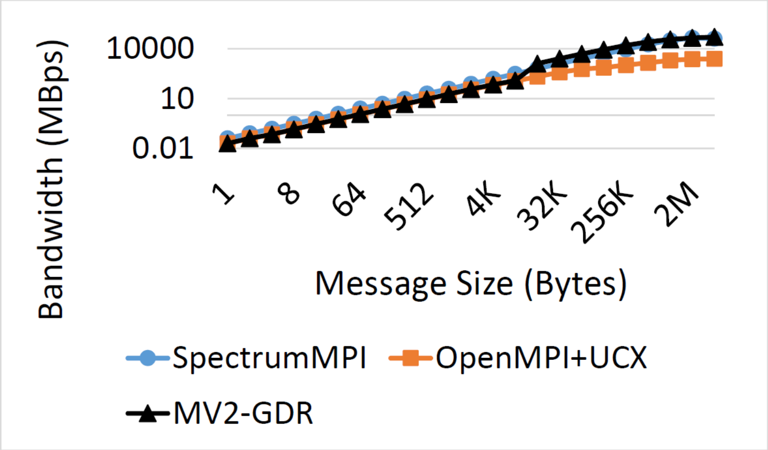
Managed Memory Performance (OpenPOWER Intra-node)



Latency MD MD

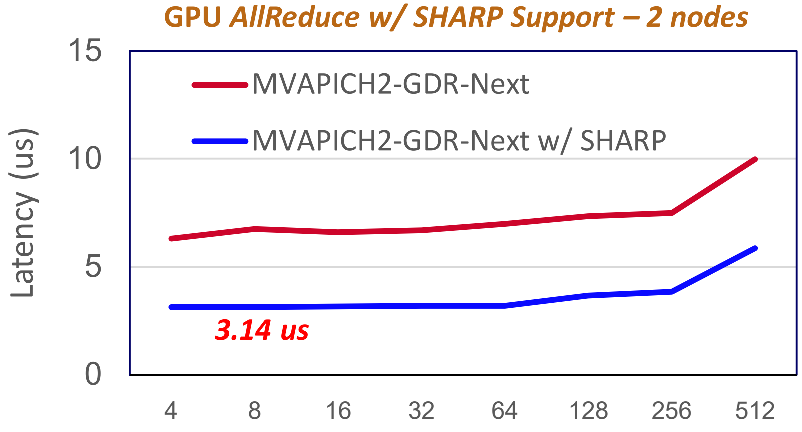


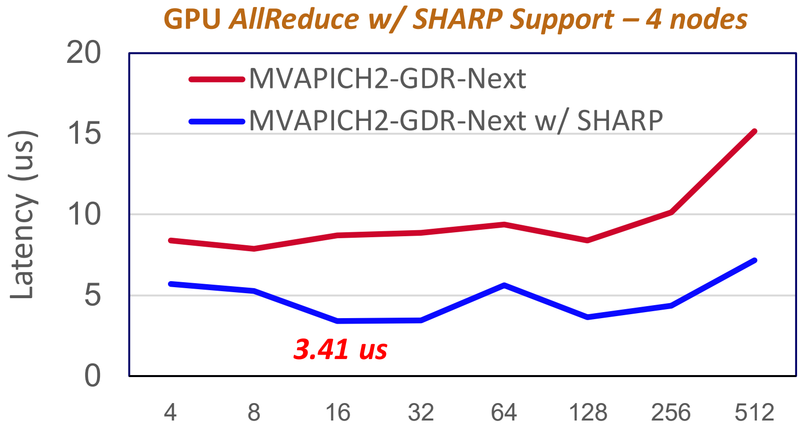
Bandwidth MD MD

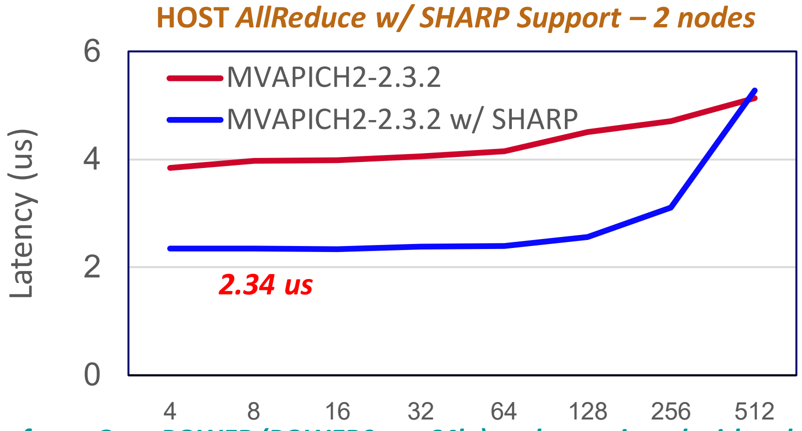


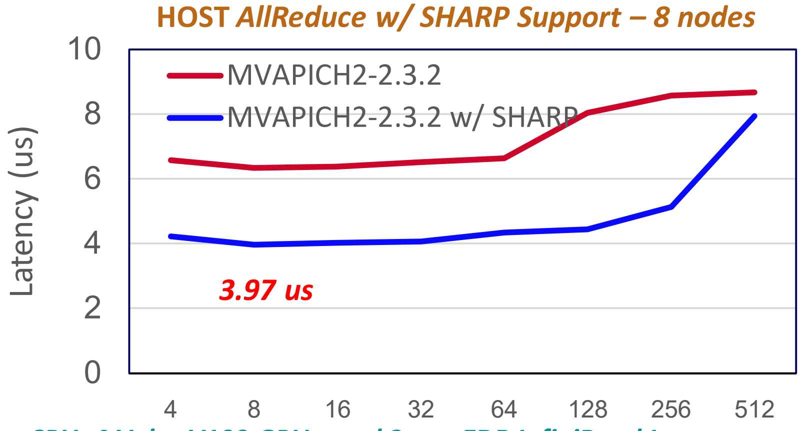
Bi-Bandwidth MD MD

MVAPICH2 with SHARP Support (Preliminary Results)



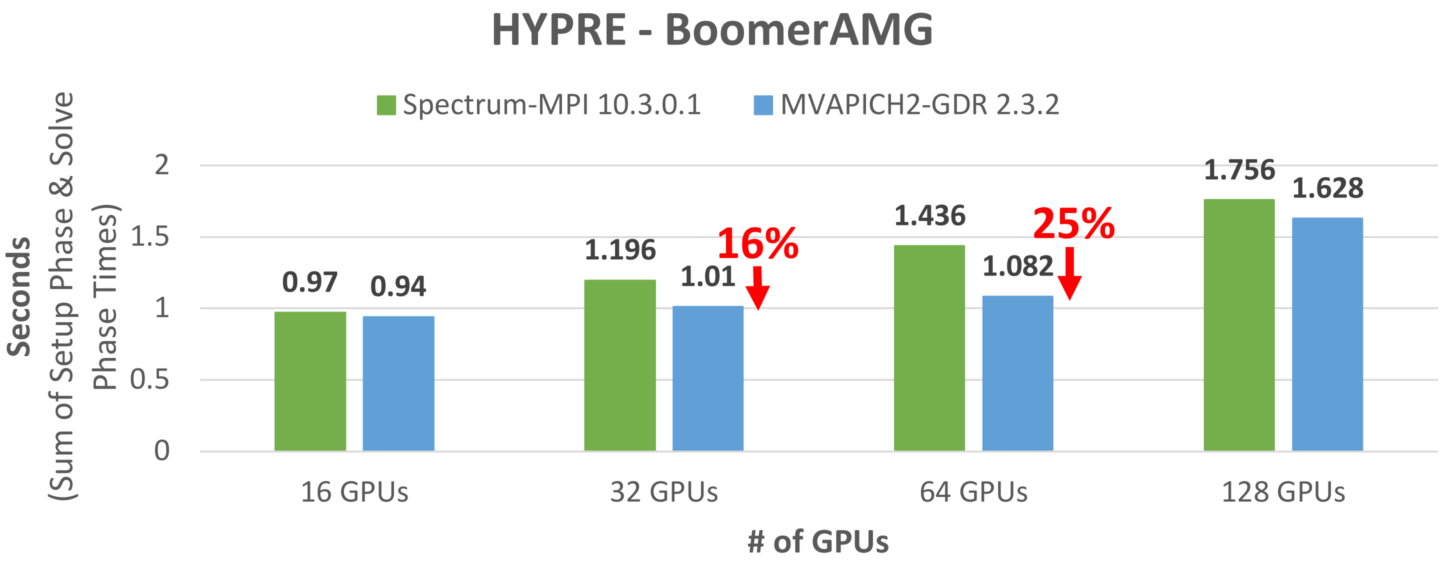






Platform: OpenPOWER (POWER9-ppc64le) nodes equipped with a dual-socket CPU, 4 Volta V100 GPUs, and 2port EDR InfiniBand Interconnect

Application: HYPRE - BoomerAMG



RUN MVAPICH2-GDR 2.3.2:

export MV2_USE_CUDA=1 MV2_USE_GDRCOPY=0 MV2_USE_RDMA_CM=0

export MV2_USE_GPUDIRECT_LOOPBACK=0 MV2_HYBRID_BINDING_POLICY=spread MV2_IBA_HCA=mlx5_0:mlx5_3

OMP_NUM_THREADS=20 lrun -n 128 -N 32 mpibind ./ij -P 8 4 4 -n 50 50 50 -pmis -Pmx 8 -keepT 1 -rlx 18

RUN Spectrum-MPI 10.3.0.1:

OMP_NUM_THREADS=20 lrun -n 128 -N 32 --smpiargs "-gpu --disable_gdr" mpibind ./ij -P 8 4 4 -n 50 50 50 -pmis -Pmx 8 -keepT 1 -rlx 18

Application: COMB

Run Scripts pushed to COMB Github repo: https://github.com/LLNL/Comb/pull/2

16 GPUs on POWER9 system (test Comm mpi Mesh cuda Device Buffers mpi_type)									
	pre-comm	post-recv	post-send	wait-recv	wait-send	post-comm	start-up	test-comm	bench-comm
Spectrum MPI 10.3	0.0001	0.0000	1.6021	1.7204	0.0112	0.0001	0.0004	7.7383	**83.6229**
MVAPICH2-GDR 2.3.2	0.0001	0.0000	0.0862	0.0871	0.0018	0.0001	0.0009	0.3558	**4.4396**
MVAPICH2-GDR 2.3.3 (Upcoming)	0.0001	0.0000	0.0030	0.0032	0.0001	0.0001	0.0009	0.0133	**0.1602**

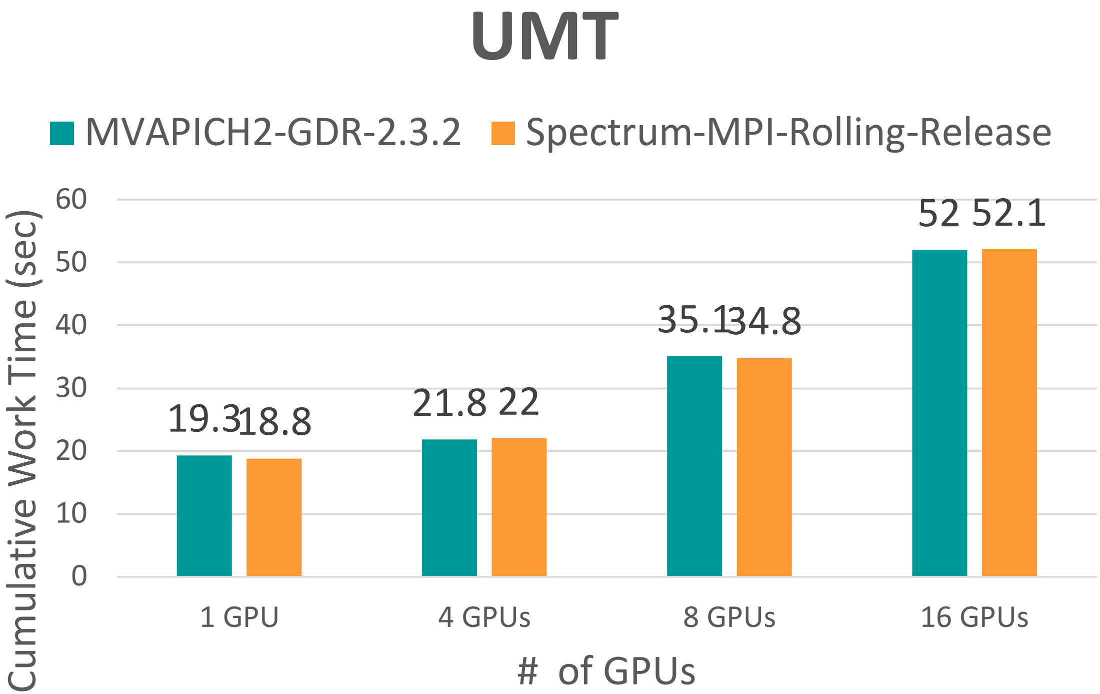
18x

27x

- **Improvements due to enhanced support for GPU-kernel based packing/unpacking routines**

Application: UMT - GPU

- Use MV2-GDR pgi/18.7 w/ jsrun rpm
- Use TAU for profiling application



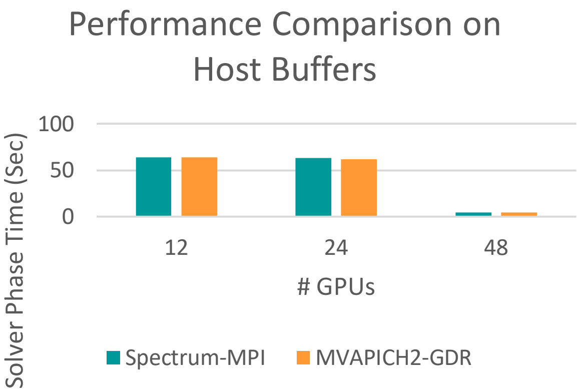
PREPARE MV2-GDR

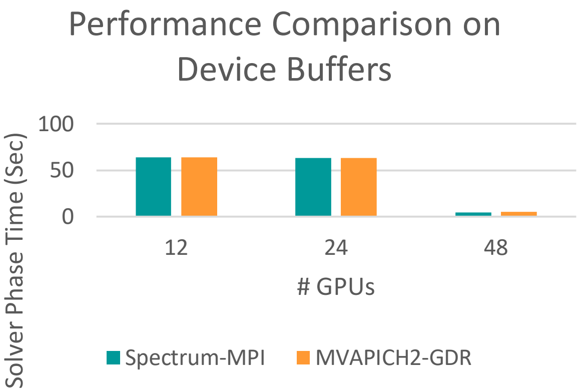
```
export MV2_USE_GDRCOPY=0
export MV2_USE_CUDA=1
export MV2_SUPPORT_TENSOR_FLOW=1
export MV2_ENABLE_AFFINITY=0
export OMPI_LD_PRELOAD_PREPEND=$HOME/software/mvapich232-jsrun-pgi/install/lib/libmpi.so
```

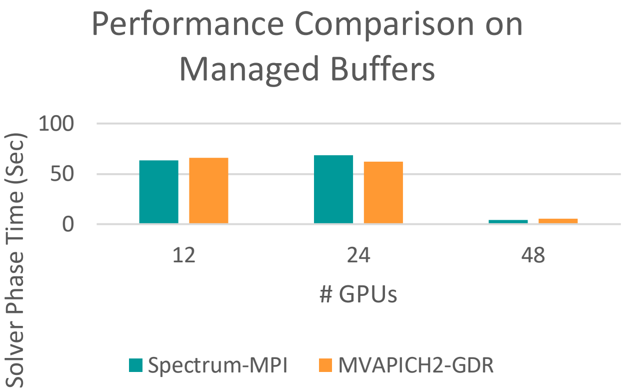
RUN Spectrum-MPI/MV2-GDR

```
jsrun -r 4 -p 16 mpibind tau_exec -ebs ./SuOlsonTest ../sierra-runs/2x2x4_20.cmg 16 2 16 8 4
```

Application: SW4







NPROCS=<#gpus per node>*<#allocated nodes>
Input_file= hayward-att-h200-ref.in

RUN MV2-GDR

$MPI_HOME/bin/mpirun_rsh -export-all -np $NPROCS --hostfile <hostfile> $mv2-gdr-flags LD_PRELOAD=$MPI_HOME/lib/libmpi.so ./sw4 $Input_file

RUN Spectrum-MPI

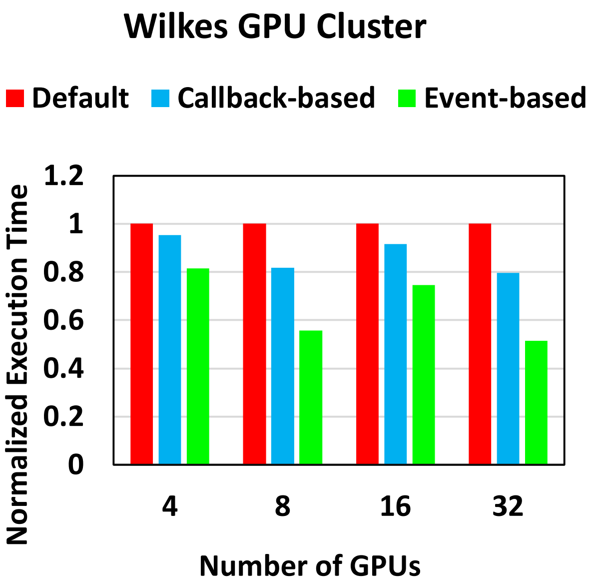
GPU_SUPPORT='-M "-gpu"' (for device & managed buffers)

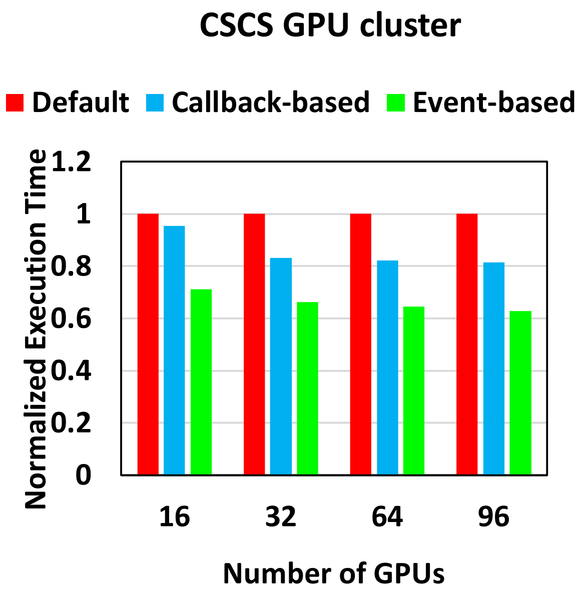
jsrun -n $NPROCS -g1 -c7 -a1 $GPU_SUPPORT ./sw4 $Input_file

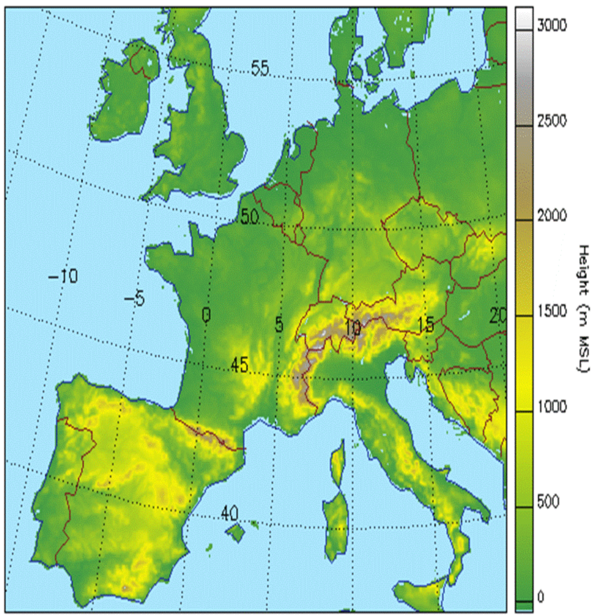
MV2-GDR-flags

MV2_USE_CUDA=1
MV2_USE_GPUDIRECT_RDMA=1
MV2_USE_GPUDIRECT_GDRCOPY=0
MV2_USE_RDMA_CM=0
MV2_DEBUG_SHOW_BACKTRACE=1
MV2_SHOW_CPU_BINDING=1
MV2_SHOW_ENV_INFO=2
MV2_USE_GPUDIRECT_LOOPBACK=0
MV2_CPU_BINDING_POLICY=HYBRID
MV2_HYBRID_BINDING_POLICY=SPREAD
MV2_IBA_HCA=mlx5_0:mlx5_3

Application-Level Evaluation (Cosmo) and Weather Forecasting in Switzerland







- 2X improvement on 32 GPUs nodes
- 30% improvement on 96 GPU nodes (8 GPUs/node)

Cosmo model: http://www2.cosmo-model.org/content/tasks/operational/meteoSwiss/

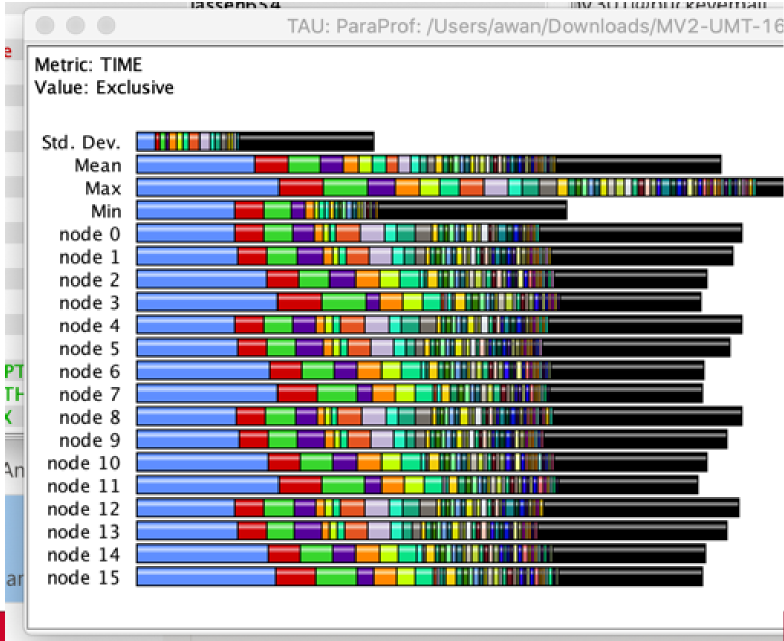
On-going collaboration with CSCS and MeteoSwiss (Switzerland) in co-designing MV2-GDR and Cosmo Application

C. Chu, K. Hamidouche, A. Venkatesh, D. Banerjee , H. Subramoni, and D. K. Panda, Exploiting Maximal Overlap for Non-Contiguous Data Movement Processing on Modern GPU-enabled Systems, IPDPS'16

TAU Profile with MVAPICH2-GDR

TAU: ParaProf: Statistics for: node 0 - /Users/awan/Downloads/MV2-UMT-16-LATEST.ppk

Name △	Exclusive TIME	Inclusive TIME	Calls	Child Calls
.TAU application	38.424	64.567	1	84,672
[CONTEXT] .TAU application	0	38.15	1,232	0
MPI_Allreduce()	6.951	6.951	437	0
MPI_Allreduce() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	6.951	6.951	437	0
MPI_Barrier()	9.424	9.424	241	0
MPI_Barrier() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	9.424	9.424	241	0
MPI_Bcast()	0.001	0.001	24	0
MPI_Bcast() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	0.001	0.001	24	0
MPI_Comm_rank()	0.082	0.082	79,297	0
MPI_Comm_rank() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	0.082	0.082	79,297	0
MPI_Comm_size()	0	0	4	0
MPI_Finalize()	0.392	0.392	1	0
MPI_Gather()	0	0	1	0
MPI_Gather() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	0	0	1	0
MPI_Gatherv()	0	0	2	0
MPI_Gatherv() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	0	0	2	0
MPI_Init()	0.674	0.674	1	0
MPI_Irecv()	0.001	0.001	84	0
MPI_Irecv() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	0.001	0.001	84	0
MPI_Isend()	0.004	0.004	84	0
MPI_Isend() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	0.004	0.004	84	0
MPI_Recv_init()	0.001	0.001	168	0
MPI_Recv_init() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	0.001	0.001	168	0
MPI_Reduce()	0	0	1	0
MPI_Reduce() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	0	0	1	0
MPI_Request_free()	0.001	0.001	336	0
MPI_Send_init()	0	0	168	0
MPI_Send_init() [<comm> = <ranks: 0, 1, 2, 3, 4, 5, 6, 7 ...> <addr=0x44000000>]	0	0	168	0
MPI_Start()	0.148	0.148	1,848	0
MPI_Wait()	8.458	8.458	1,932	0
MPI_Waitall()	0.003	0.003	42	0
cudaFreeHost	0.314	0.314	2	0
cudaGetDevice	0	0	1	0

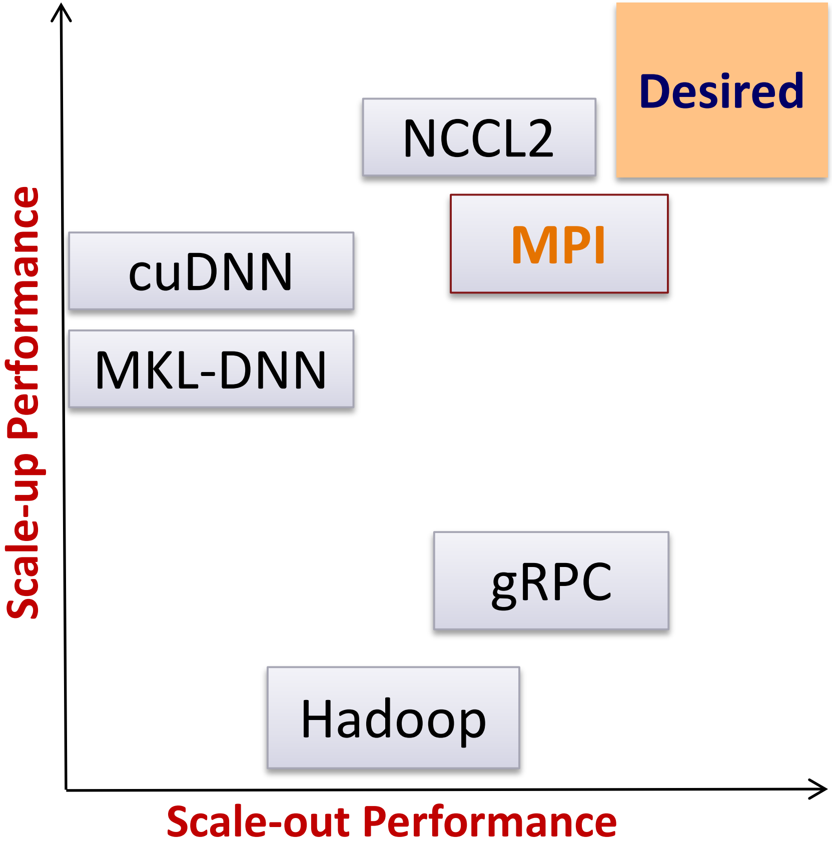


Presentation Overview

- Challenges in Designing Convergent HPC, Big Data and Deep Learning Architectures
- MVAPICH Project - MPI and PGAS (MVAPICH) Library with CUDA-Awareness
- **HiDL Project – High-Performance Deep Learning**
- HiBD Project – High-Performance Big Data Analytics Library
- Commercial Support from X-ScaleSolutions
- Conclusions and Q&A

Deep Learning: New Challenges for MPI Runtimes

- Deep Learning frameworks are a different game altogether
  - Unusually large message sizes (order of megabytes)
  - Most communication based on GPU buffers
- Existing State-of-the-art
  - cuDNN, cuBLAS, NCCL --> **scale-up** performance
  - NCCL2, CUDA-Aware MPI --> **scale-out** performance
    - For small and medium message sizes only!
- Proposed: Can we **co-design** the MPI runtime (**MVAPICH2-GDR**) and the DL framework (**Caffe**) to achieve both?
  - Efficient **Overlap** of Computation and Communication
  - Efficient **Large-Message** Communication (Reductions)
  - What **application co-designs** are needed to exploit **communication-runtime co-designs**?



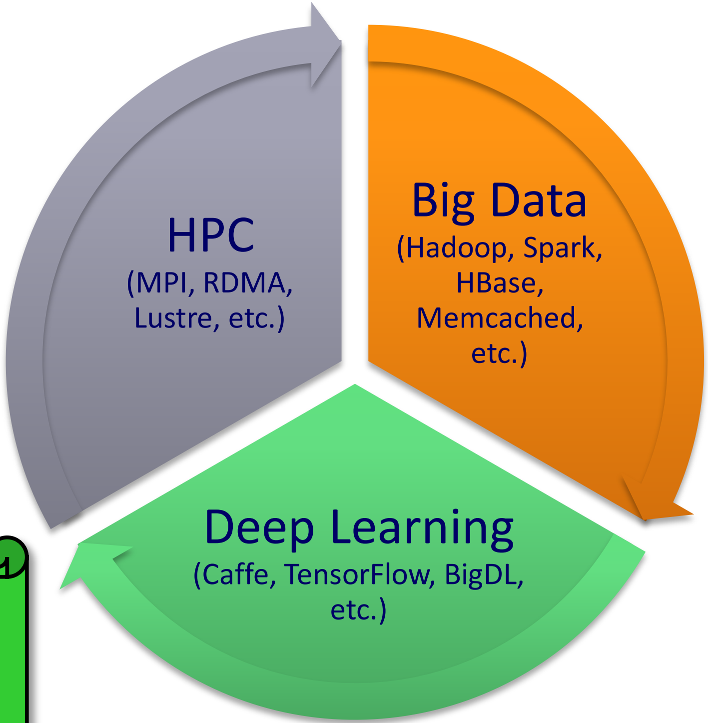
A. A. Awan, K. Hamidouche, J. M. Hashmi, and D. K. Panda, S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters. In *Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPoPP '17)

Convergent Software Stacks for HPC, Big Data and Deep Learning

MVAPICH2

MVAPICH2-X

MVAPICH2-GDR



MVAPICH2-X

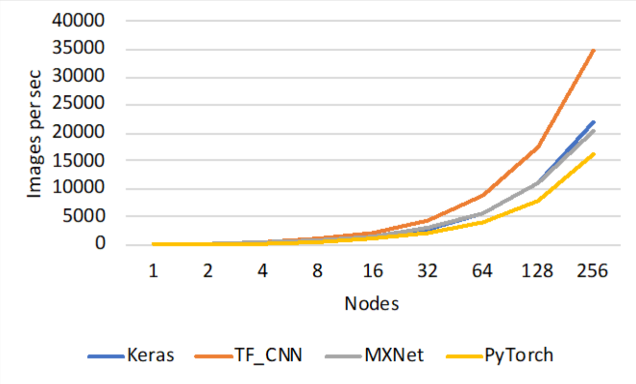
MVAPICH2-GDR

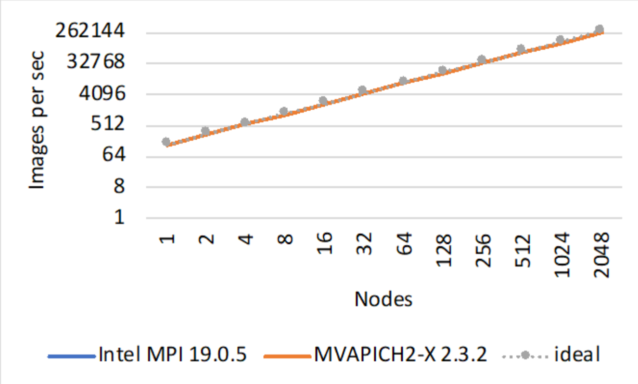
High-Performance Deep Learning

- **CPU-based Deep Learning**
  - **Using MVAPICH2-X**
- GPU-based Deep Learning
  - Using MVAPICH2-GDR

ResNet-50 using various DL benchmarks on Frontera

- Observed 260K images per sec for ResNet-50 on 2,048 Nodes
- Scaled MVAPICH2-X on 2,048 nodes on Frontera for Distributed Training using TensorFlow
- ResNet-50 can be trained in 7 minutes on 2048 nodes (114,688 cores)





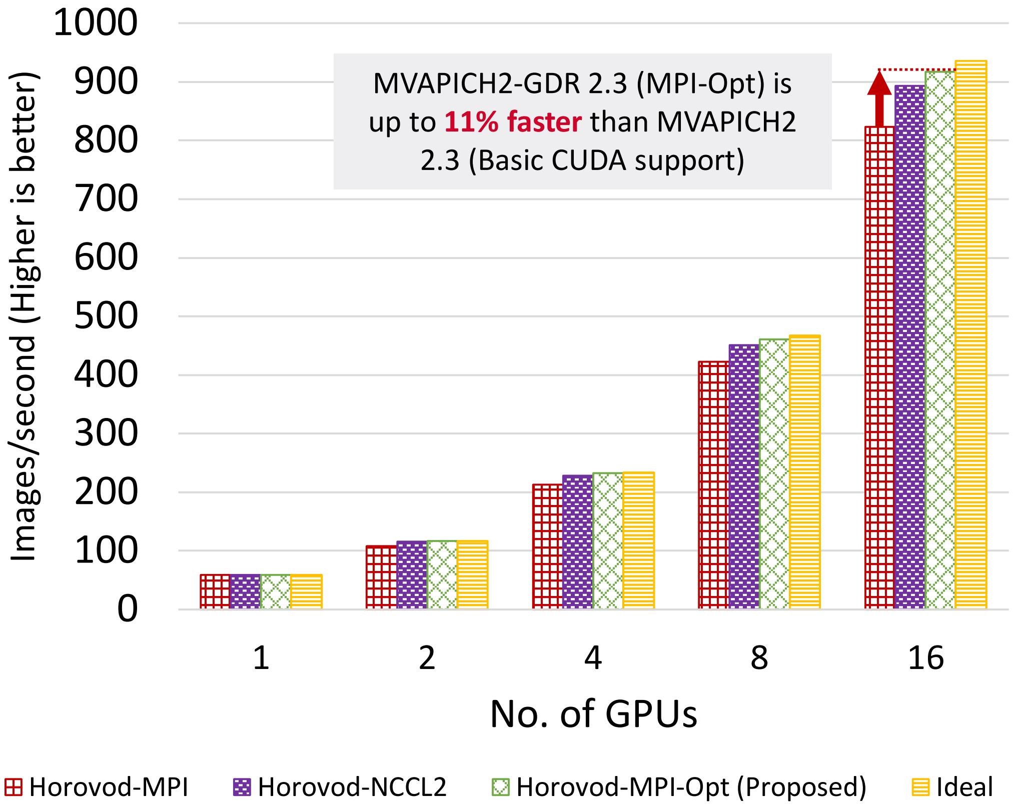
*Jain et al., "Scaling TensorFlow, PyTorch, and MXNet using MVAPICH2 for High-Performance Deep Learning on Frontera", DLS '19 (in conjunction with SC '19).

High-Performance Deep Learning

- CPU-based Deep Learning
  - Using MVAPICH2-X
- **GPU-based Deep Learning**
  - **Using MVAPICH2-GDR**

Exploiting CUDA-Aware MPI for TensorFlow (Horovod)

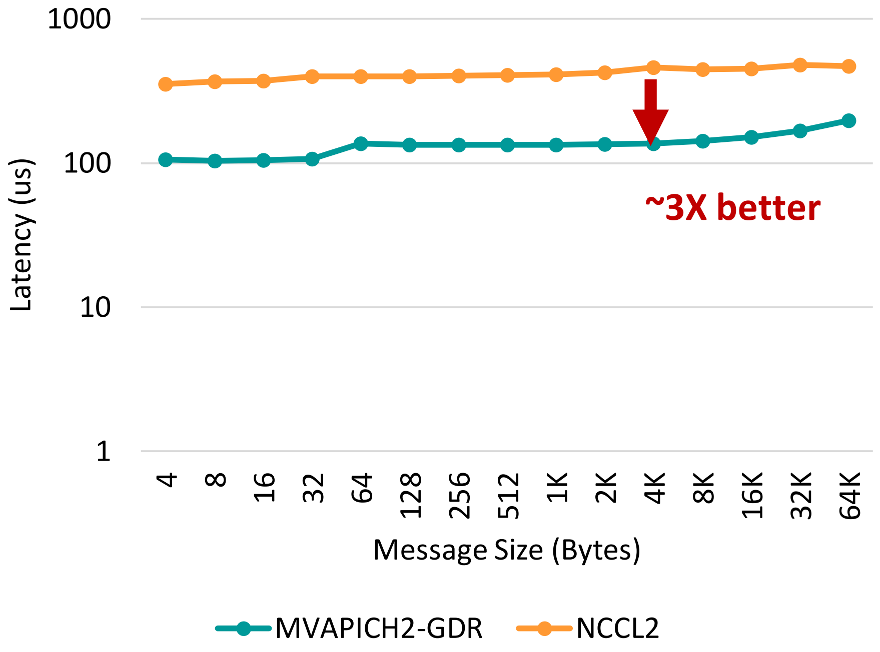
- MVAPICH2-GDR offers excellent performance via advanced designs for MPI_Allreduce.
- Up to **11% better** performance on the RI2 cluster (16 GPUs)
- Near-ideal – **98% scaling efficiency**

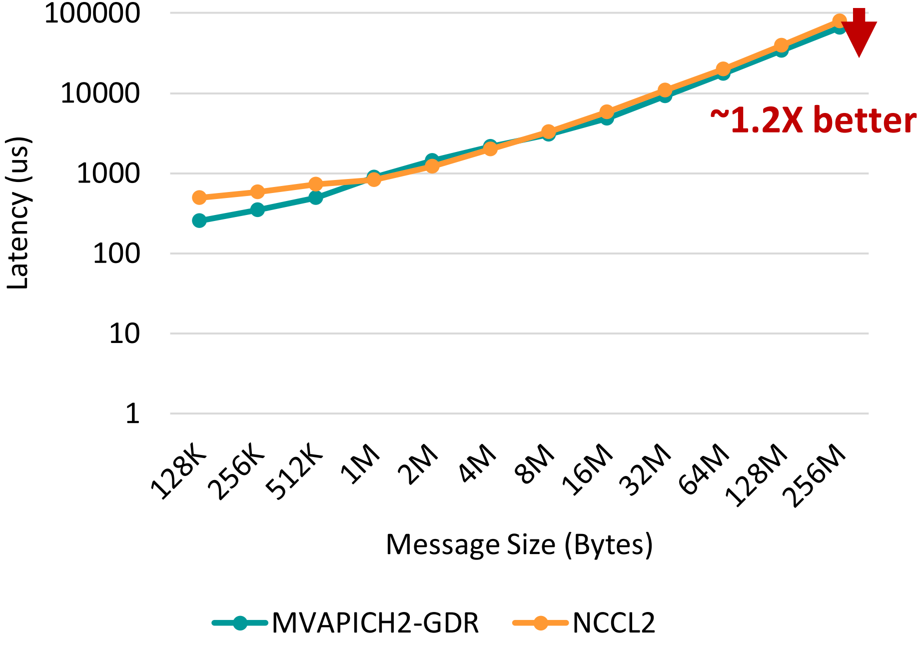


A. A. Awan et al., "Scalable Distributed DNN Training using TensorFlow and CUDA-Aware MPI: Characterization, Designs, and Performance Evaluation", CCGrid '19, https://arxiv.org/abs/1810.11112

MVAPICH2-GDR vs. NCCL2 – Allreduce Operation

- **Optimized designs in MVAPICH2-GDR 2.3 offer better/comparable performance for most cases**
- **MPI_Allreduce (MVAPICH2-GDR) vs. ncclAllreduce (NCCL2) on 16 GPUs**

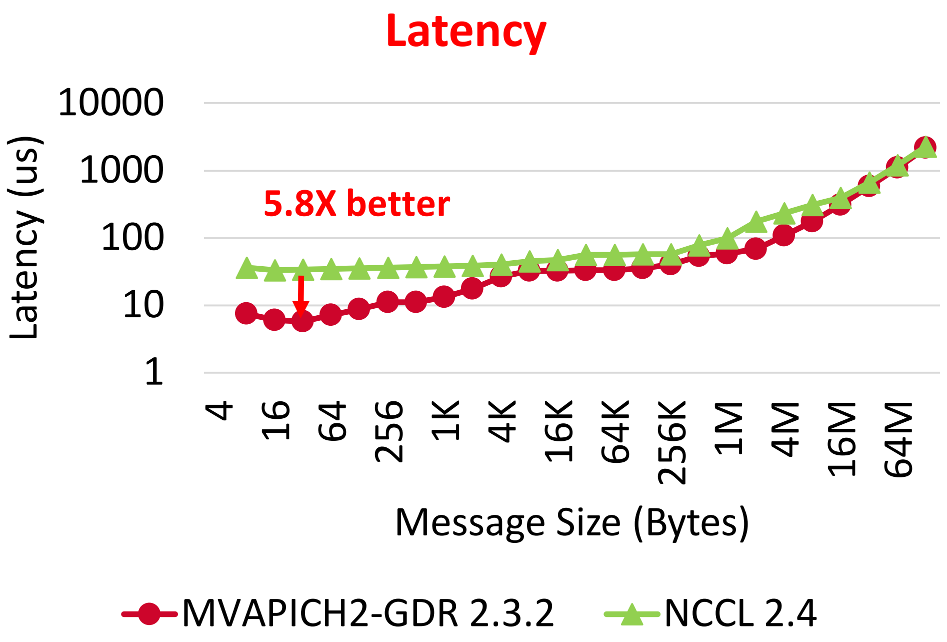


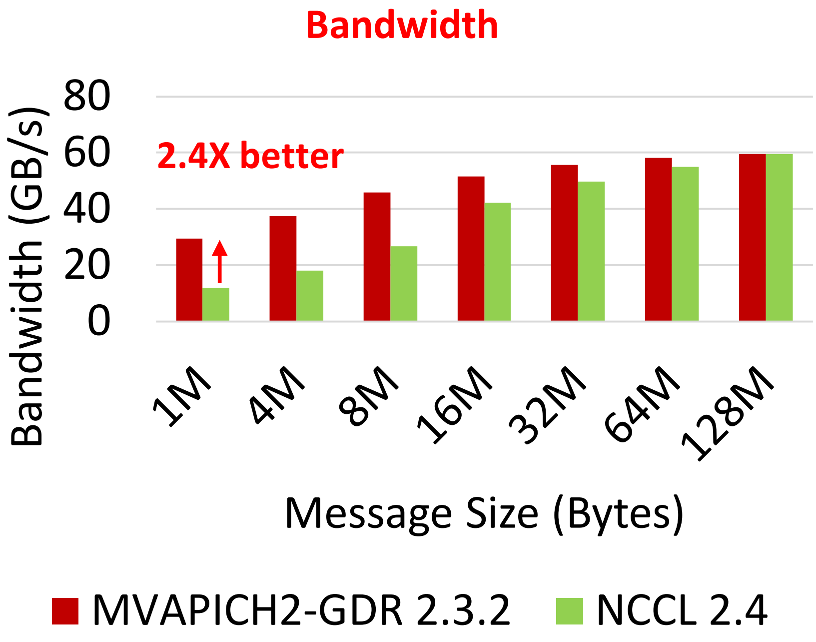


Platform: Intel Xeon (Broadwell) nodes equipped with a dual-socket CPU, 1 K-80 GPUs, and EDR InfiniBand Inter-connect

MVAPICH2-GDR vs. NCCL2: Allreduce Optimization (DGX-2)

- Optimized designs in upcoming MVAPICH2-GDR offer better performance for most cases
- MPI_Allreduce (MVAPICH2-GDR) vs. ncclAllreduce (NCCL2) on a DGX-2 machine

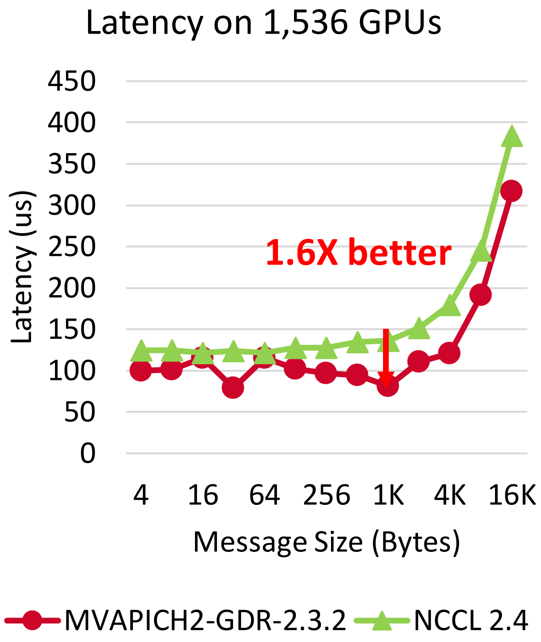


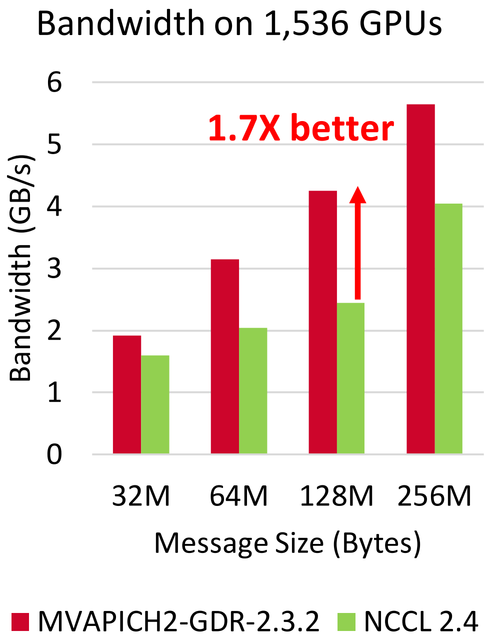


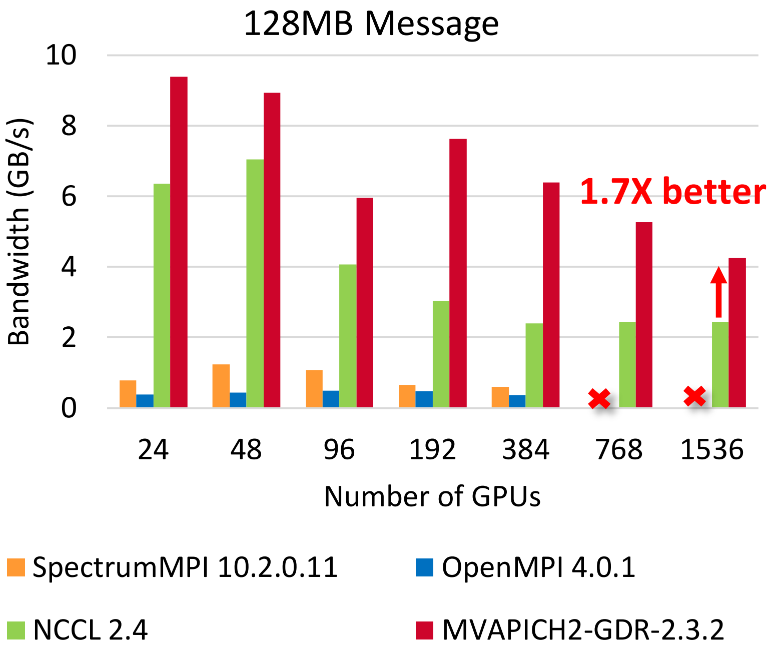
Platform: Nvidia DGX-2 system @ PSC (16 Nvidia Volta GPUs connected with NVSwitch), CUDA 9.2

MVAPICH2-GDR: MPI_Allreduce (Device Buffers) on Summit

- **Optimized designs in MVAPICH2-GDR offer better performance for most cases**
- **MPI_Allreduce (MVAPICH2-GDR) vs. ncclAllreduce (NCCL2) up to 1,536 GPUs**



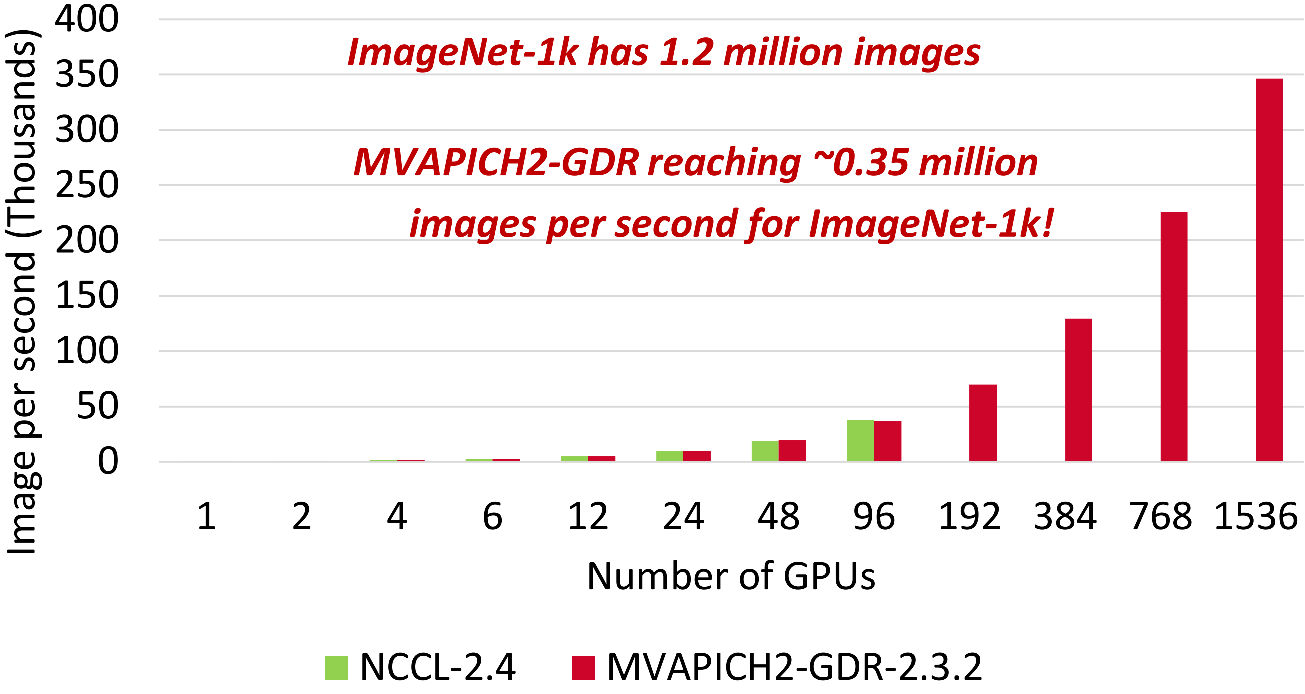




Platform: Dual-socket IBM POWER9 CPU, 6 NVIDIA Volta V100 GPUs, and 2-port InfiniBand EDR Interconnect

Distributed Training with TensorFlow and MVAPICH2-GDR on Summit

- ResNet-50 Training using TensorFlow benchmark on SUMMIT -- 1536 Volta GPUs!
- 1,281,167 (1.2 mil.) images
- Time/epoch = 3.6 seconds
- Total Time (90 epochs) = 3.6 x 90 = 332 seconds = **5.5 minutes!**

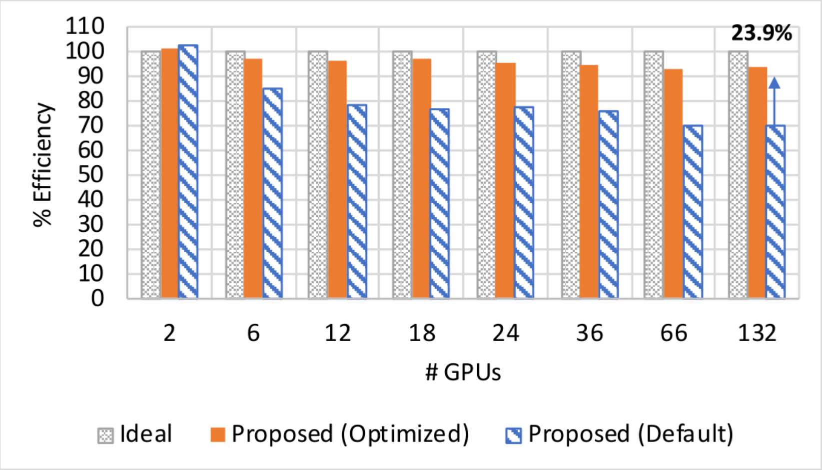


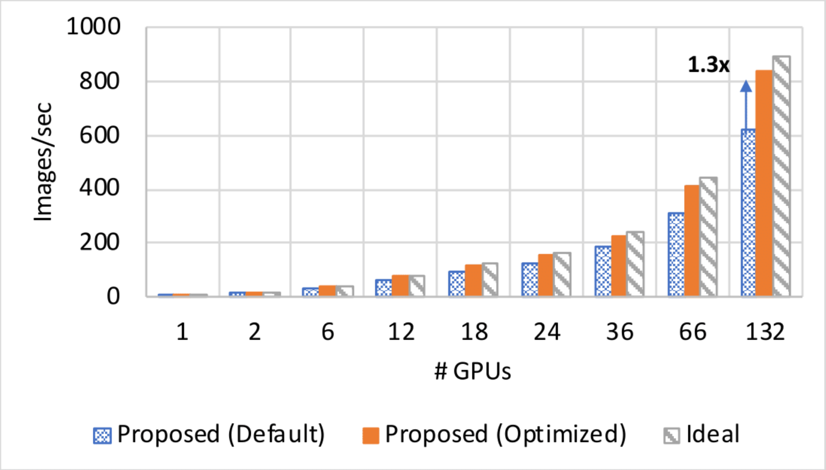
*We observed errors for NCCL2 beyond 96 GPUs

Platform: The Summit Supercomputer (#1 on Top500.org) – 6 NVIDIA Volta GPUs per node connected with NVLink, CUDA 9.2

New Benchmark for Image Segmentation on Summit

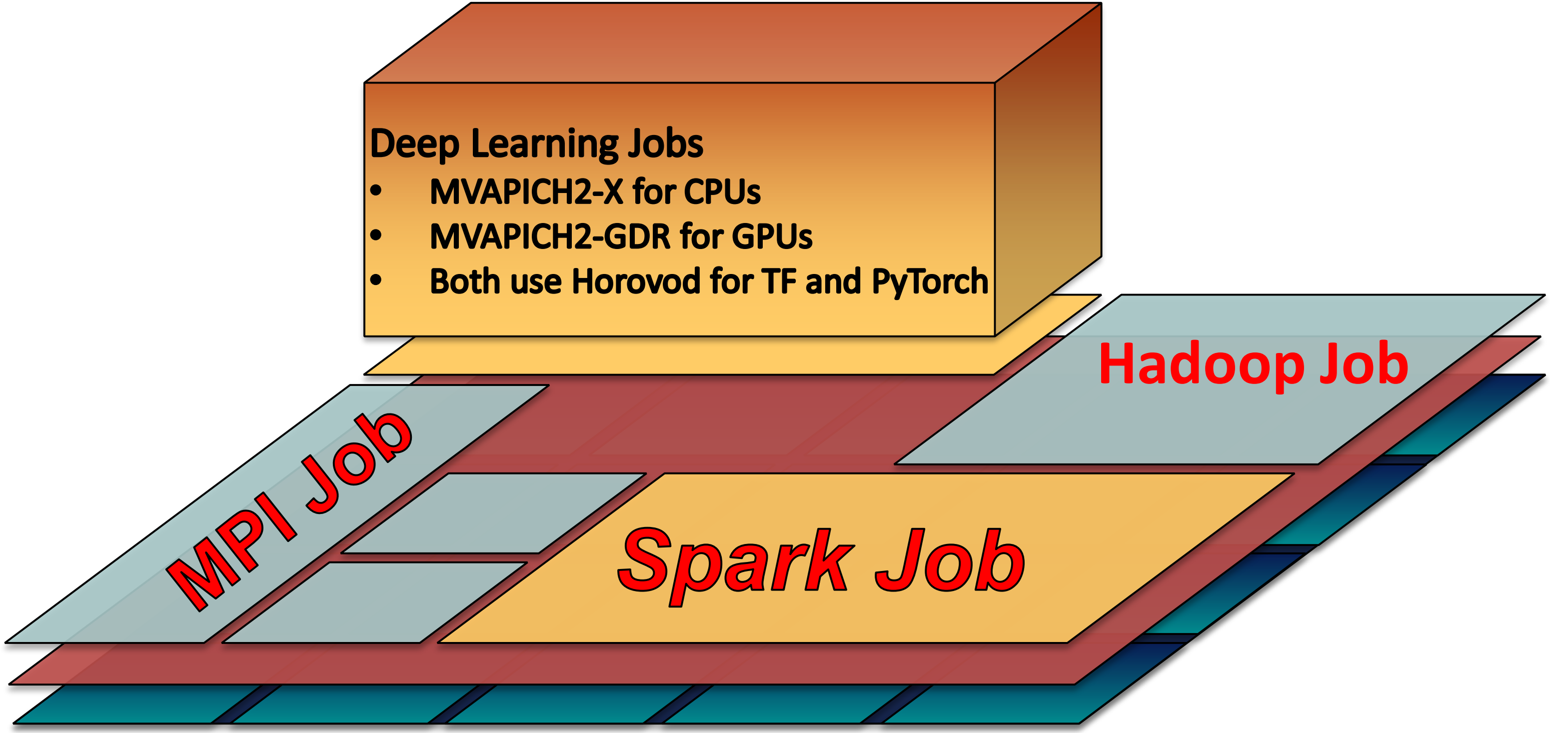
- Near-linear scaling may be achieved by **tuning Horovod/MPI**
  - Optimizing MPI/Horovod towards large message sizes for high-resolution images
- Develop a generic Image Segmentation benchmark
- Tuned DeepLabV3+ model using the benchmark and Horovod, up to **1.3X** better than default





*Anthony et al., "Scaling Semantic Image Segmentation using Tensorflow and MVAPICH2-GDR on HPC Systems" (Submission under review)

Using HiDL Packages for Deep Learning on Existing HPC Infrastructure

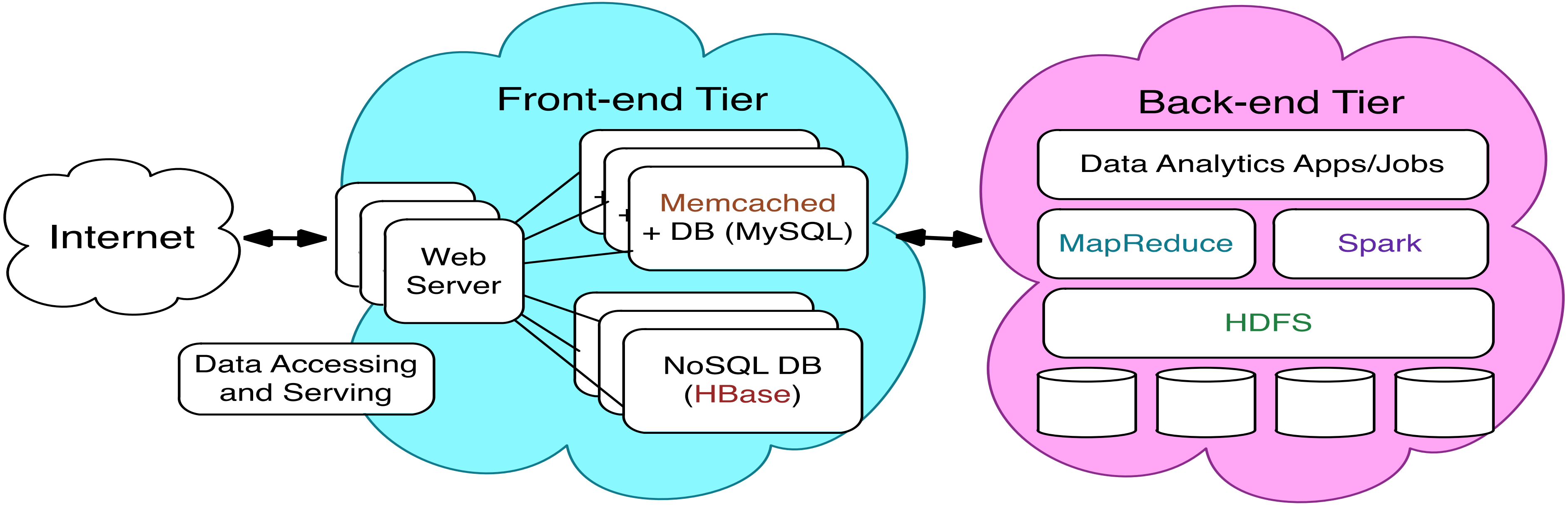


Presentation Overview

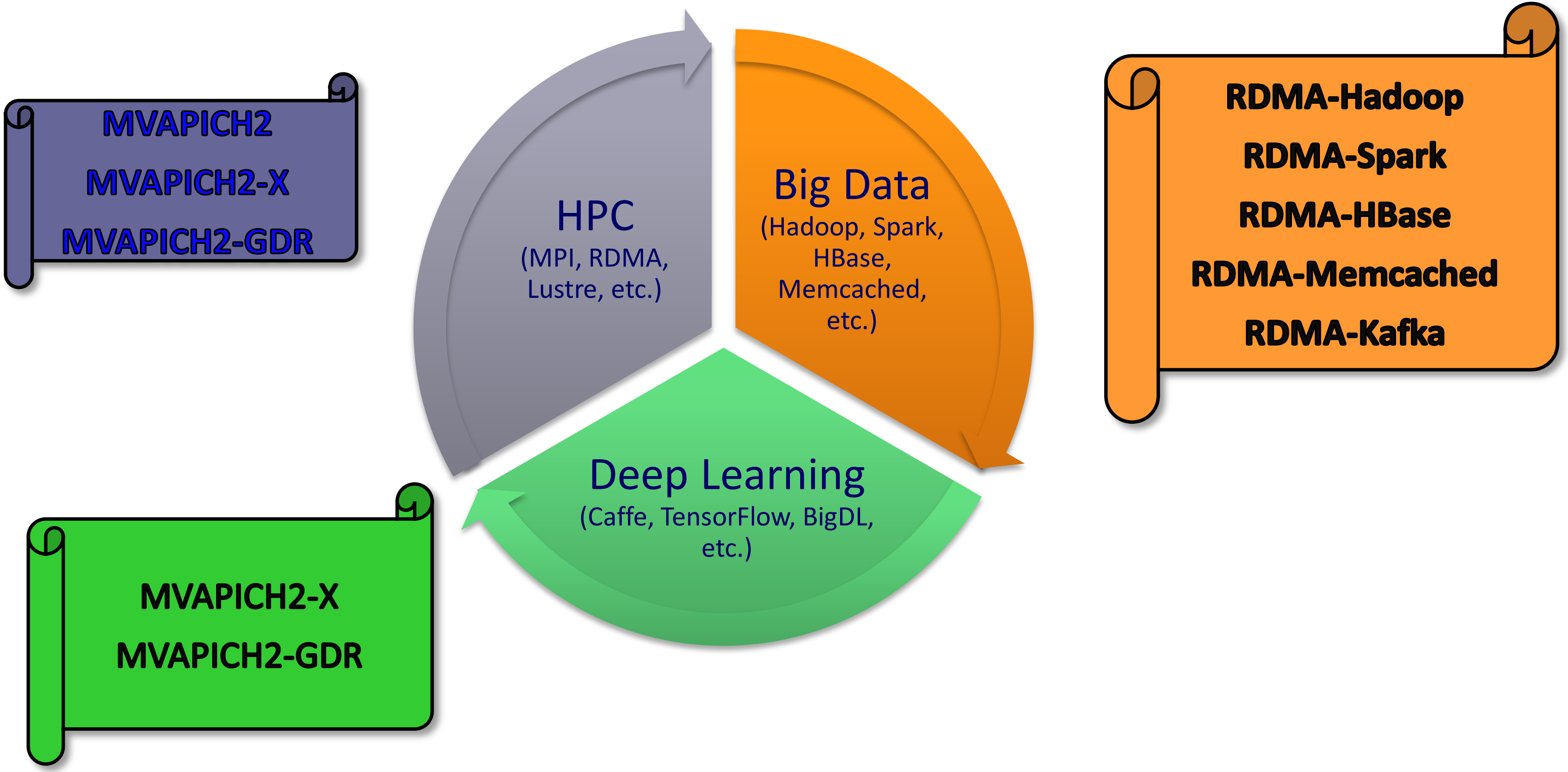
- Challenges in Designing Convergent HPC, Big Data and Deep Learning Architectures
- MVAPICH Project - MPI and PGAS (MVAPICH) Library with CUDA-Awareness
- HiDL Project – High-Performance Deep Learning
- **HiBD Project – High-Performance Big Data Analytics Library**
- Commercial Support from X-ScaleSolutions
- Conclusions and Q&A

Data Management and Processing on Modern Datacenters

- Substantial impact on designing and utilizing data management and processing systems in multiple tiers
  - Front-end data accessing and serving (Online)
    - Memcached + DB (e.g. MySQL), HBase
  - Back-end data analytics (Offline)
    - HDFS, MapReduce, Spark



Convergent Software Stacks for HPC, Big Data and Deep Learning



The High-Performance Big Data (HiBD) Project

- RDMA for Apache Spark
- RDMA for Apache Hadoop 3.x (RDMA-Hadoop-3.x)
- RDMA for Apache Hadoop 2.x (RDMA-Hadoop-2.x)
  - Plugins for Apache, Hortonworks (HDP) and Cloudera (CDH) Hadoop distributions
- RDMA for Apache Kafka
- RDMA for Apache HBase
- RDMA for Memcached (RDMA-Memcached)
- RDMA for Apache Hadoop 1.x (RDMA-Hadoop)
- **OSU HiBD-Benchmarks (OHB)**
  - **HDFS, Memcached, HBase, and Spark Micro-benchmarks**
- **http://hibd.cse.ohio-state.edu**
- Users Base: 315 organizations from 35 countries
- More than 31,600 downloads from the project site

Available for InfiniBand and RoCE

Also run on Ethernet

Available for x86 and OpenPOWER

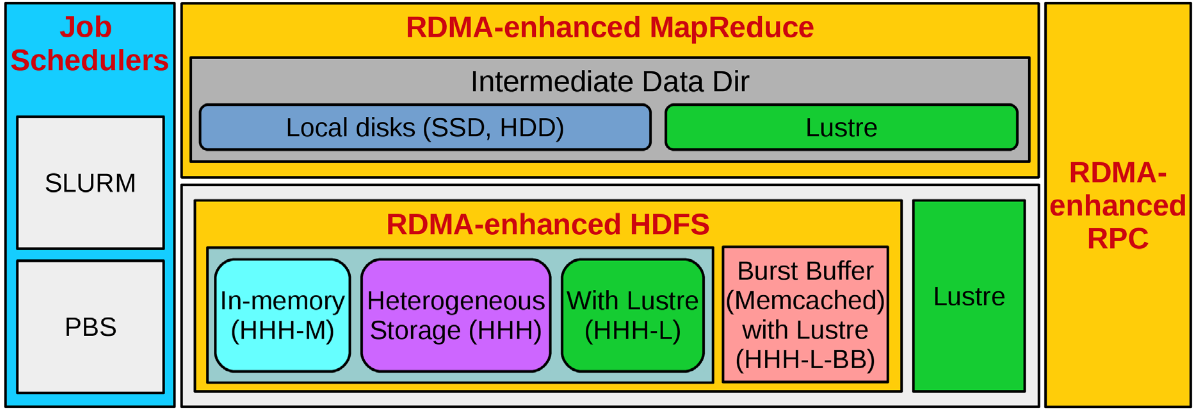
Support for Singularity and Docker

HiBD High-Performance Big Data

Network Based Computing Laboratory

THE OHIO STATE UNIVERSITY

Different Modes of RDMA for Apache Hadoop 2.x



- **HHH**: Heterogeneous storage devices with hybrid replication schemes are supported in this mode of operation to have better fault-tolerance as well as performance. This mode is enabled by **default** in the package.
- **HHH-M**: A high-performance in-memory based setup has been introduced in this package that can be utilized to perform all I/O operations in-memory and obtain as much performance benefit as possible.
- **HHH-L**: With parallel file systems integrated, HHH-L mode can take advantage of the Lustre available in the cluster.
- **HHH-L-BB**: This mode deploys a Memcached-based burst buffer system to reduce the bandwidth bottleneck of shared file system access. The burst buffer design is hosted by Memcached servers, each of which has a local SSD.
- **MapReduce over Lustre, with/without local disks**: Besides, HDFS based solutions, this package also provides support to run MapReduce jobs on top of Lustre alone. Here, two different modes are introduced: with local disks and without local disks.
- **Running with Slurm and PBS**: Supports deploying RDMA for Apache Hadoop 2.x with Slurm and PBS in different running modes (HHH, HHH-M, HHH-L, and MapReduce over Lustre).

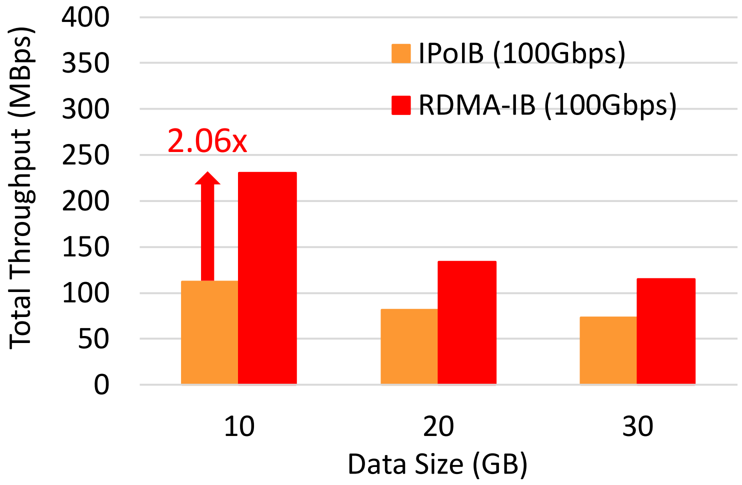
RDMA for Apache Hadoop 2.x Distribution

- High-Performance Design of Hadoop over RDMA-enabled Interconnects
  - High performance RDMA-enhanced design with native InfiniBand and RoCE support at the verbs-level for HDFS, MapReduce, and RPC components
  - Enhanced HDFS with in-memory and heterogeneous storage
  - High performance design of MapReduce over Lustre
  - Memcached-based burst buffer for MapReduce over Lustre-integrated HDFS (HHH-L-BB mode)
  - Plugin-based architecture supporting RDMA-based designs for Apache Hadoop, CDH and HDP
  - Support for OpenPOWER, Singularity, and Docker
- Current release: 1.3.5
  - Based on Apache Hadoop 2.8.0
  - Compliant with Apache Hadoop 2.8.0, HDP 2.5.0.3 and CDH 5.8.2 APIs and applications
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR, FDR, and EDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms (x86, POWER)
    - Different file systems with disks and SSDs and Lustre

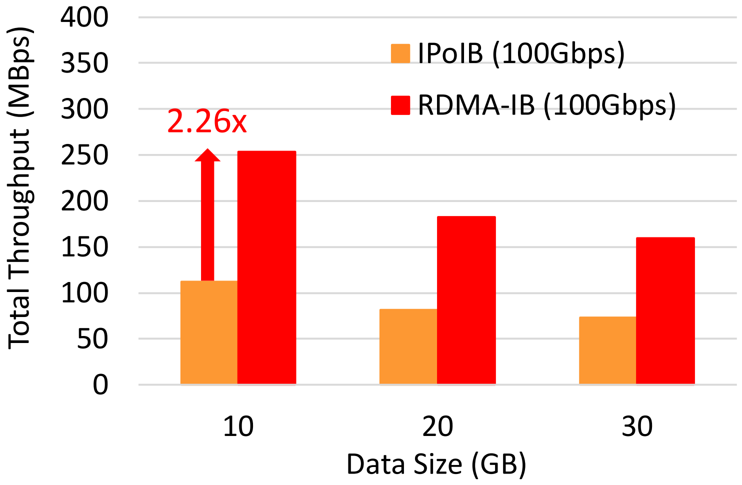
http://hibd.cse.ohio-state.edu

Performance of RDMA-Hadoop on OpenPOWER

TestDFSIO Throughput



HHH Mode

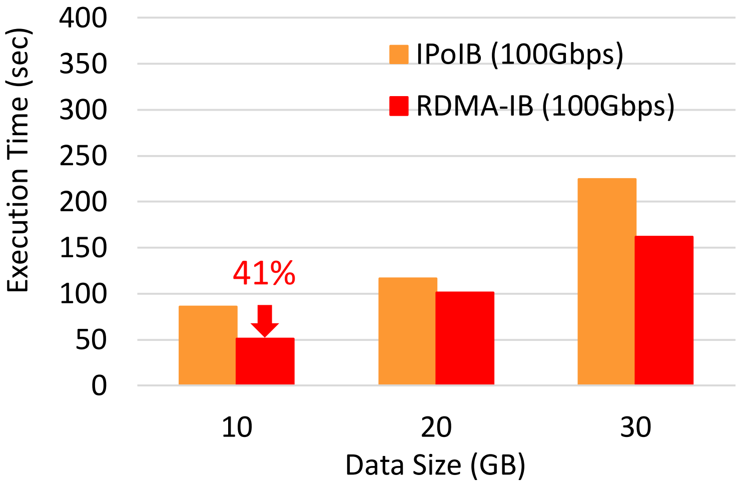


HHH-M Mode

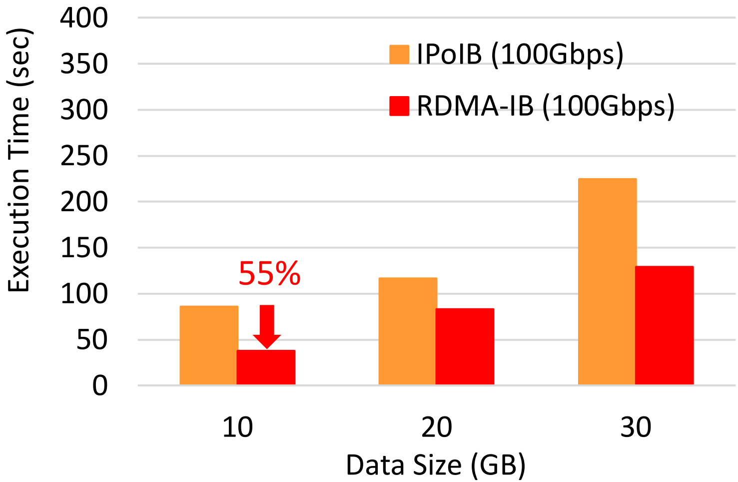
- For TestDFSIO throughput experiment, RDMA-IB design of HHH mode has an improvement of 1.57x - 2.06x compared to IPoIB (100Gbps).
- In HHH-M mode, the improvement goes up to 2.18x - 2.26x compared to IPoIB (100Gbps).

Performance of RDMA-Hadoop on OpenPOWER

Sort Execution Time



HHH Mode

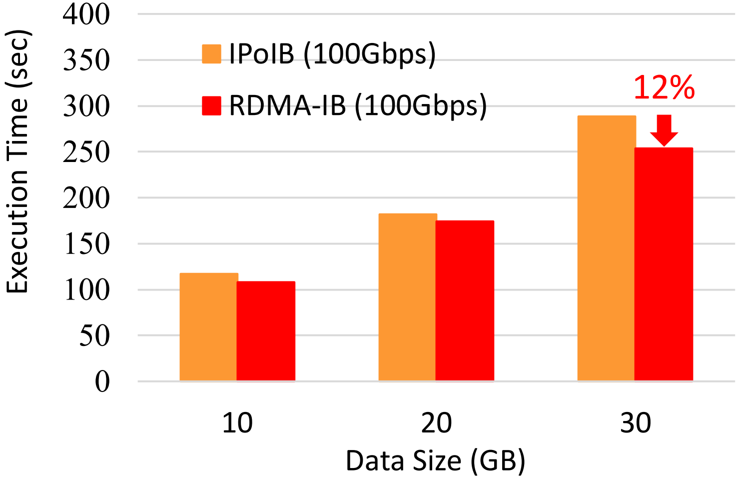


HHH-M Mode

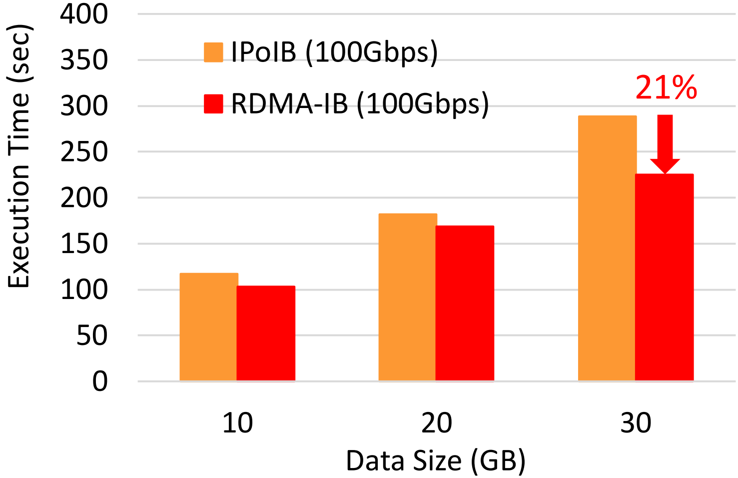
- The RDMA-IB design of HHH mode reduces the job execution time of Sort by a maximum of 41% compared to IPoIB (100Gbps).
- The HHH-M design reduces the execution time by a maximum of 55%.

Performance of RDMA-Hadoop on OpenPOWER

TeraSort Execution Time



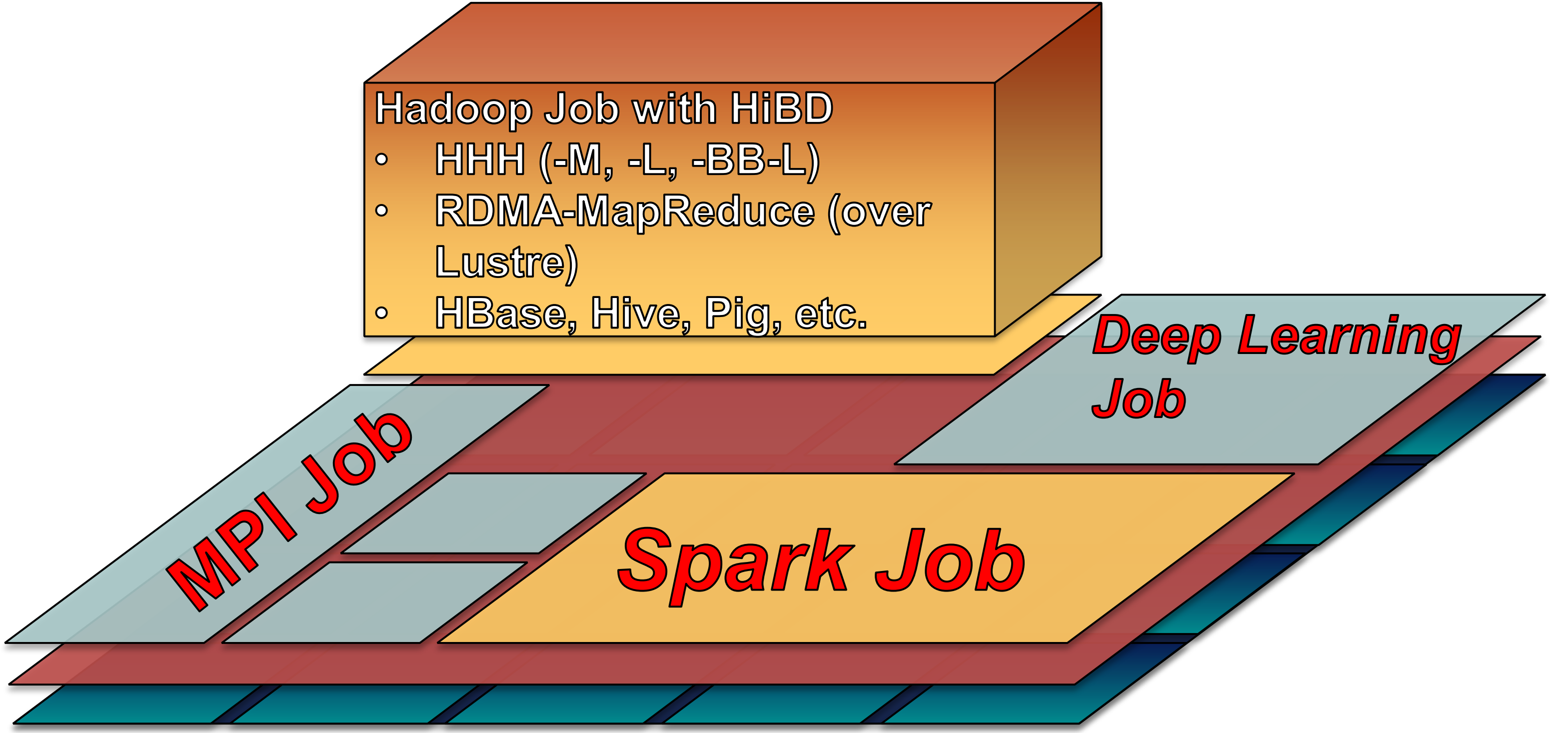
HHH Mode



HHH-M Mode

- The RDMA-IB design of HHH mode reduces the job execution time of TeraSort by a maximum of 12% compared to IPoIB (100Gbps).
- In HHH-M mode, the execution time of TeraSort is reduced by a maximum of 21% compared to IPoIB (100Gbps).

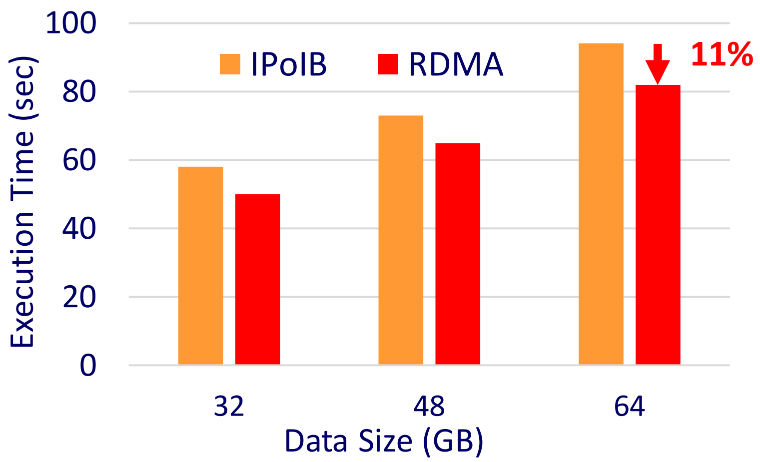
Using HiBD Packages for Big Data Processing on Existing HPC Infrastructure



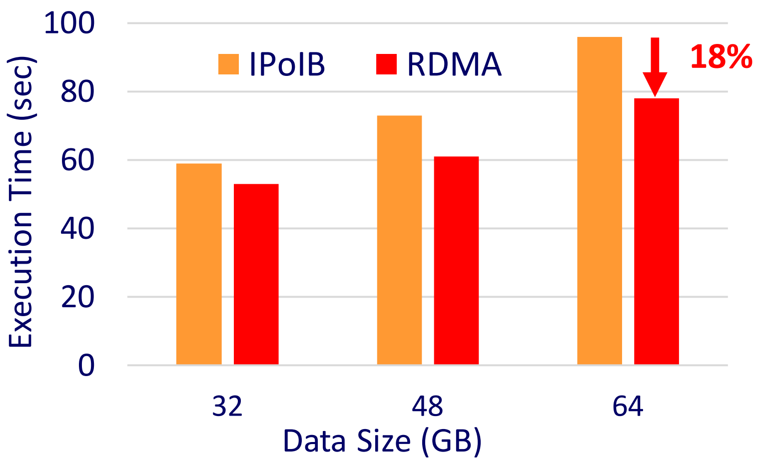
RDMA for Apache Spark Distribution

- High-Performance Design of Spark over RDMA-enabled Interconnects
  - High performance RDMA-enhanced design with native InfiniBand and RoCE support at the verbs-level for Spark
  - RDMA-based data shuffle and SEDA-based shuffle architecture
  - Non-blocking and chunk-based data transfer
  - Off-JVM-heap buffer management
  - Support for OpenPOWER
  - Easily configurable for different protocols (native InfiniBand, RoCE, and IPoIB)
- Current release: 0.9.5
  - Based on Apache Spark 2.1.0
  - Tested with
    - Mellanox InfiniBand adapters (DDR, QDR, FDR, and EDR)
    - RoCE support with Mellanox adapters
    - Various multi-core platforms (x86, POWER)
    - RAM disks, SSDs, and HDD
  - http://hibd.cse.ohio-state.edu

Performance of RDMA-Spark on OpenPOWER



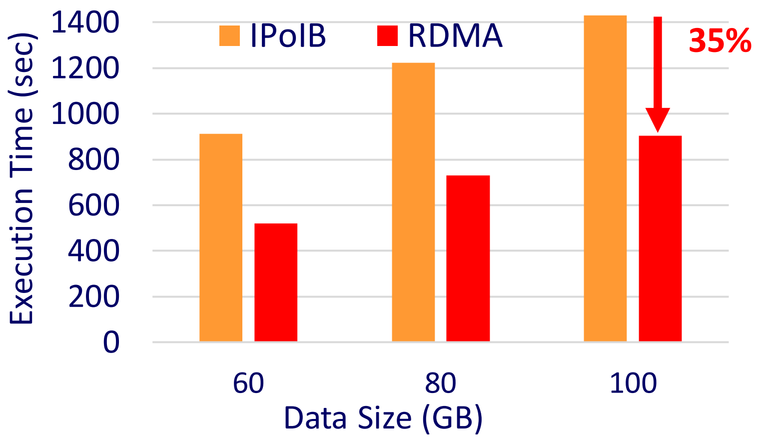
GroupBy



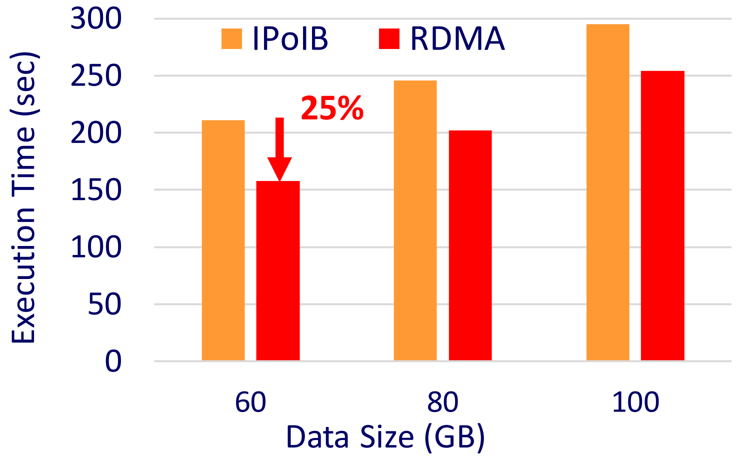
SortBy

- GroupBy: RDMA design outperforms IPoIB by a maximum of 11%
- SortBy: RDMA design outperforms IPoIB by a maximum of 18%

Performance of RDMA-Spark on OpenPOWER



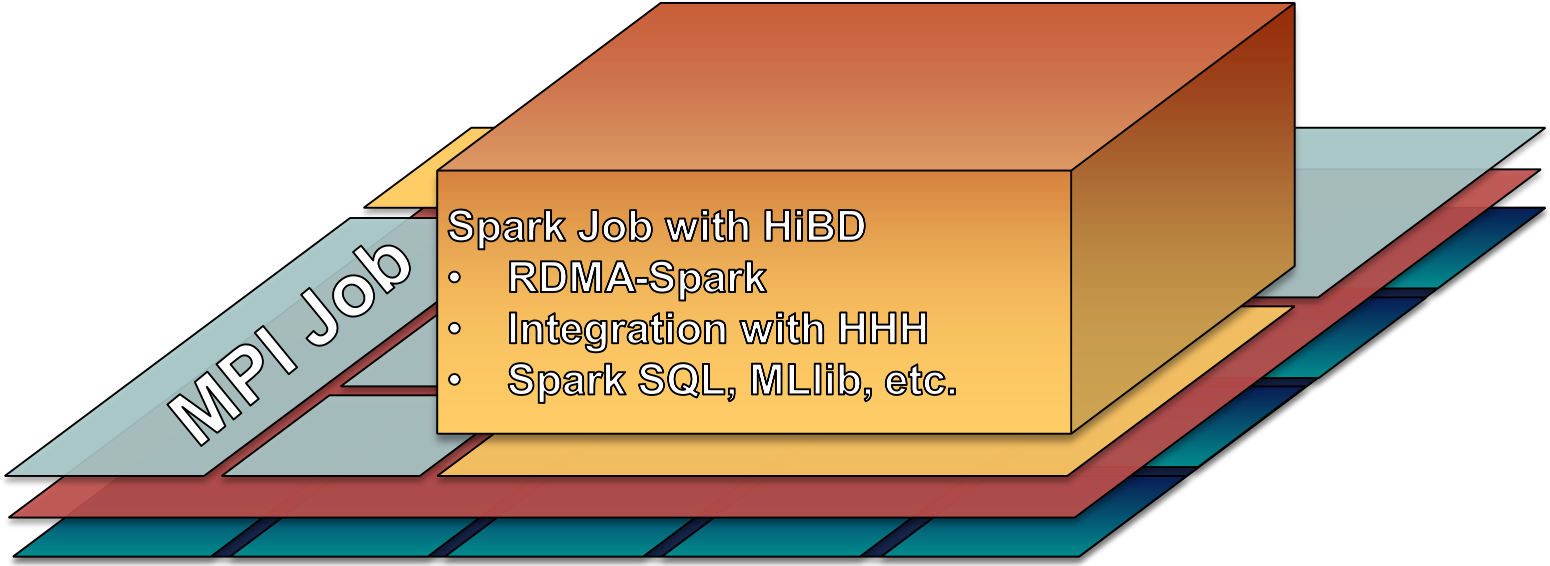
TeraSort



Sort

- TeraSort: RDMA design outperforms IPoIB by a maximum of 35%
- Sort: RDMA design outperforms IPoIB by a maximum of 25%

Using HiBD Packages for Big Data Processing on Existing HPC Infrastructure



Presentation Overview

- Challenges in Designing Convergent HPC, Big Data and Deep Learning Architectures
- MVAPICH Project - MPI and PGAS (MVAPICH) Library with CUDA-Awareness
- HiDL Project – High-Performance Deep Learning
- HiBD Project – High-Performance Big Data Analytics Library
- **Commercial Support from X-ScaleSolutions**
- Conclusions and Q&A

Commercial Support for MVAPICH2, HiBD, and HiDL Libraries

- Supported through X-ScaleSolutions (http://x-scalesolutions.com)
- Benefits:
  - Help and guidance with installation of the library
  - Platform-specific optimizations and tuning
  - Timely support for operational issues encountered with the library
  - Web portal interface to submit issues and tracking their progress
  - Advanced debugging techniques
  - Application-specific optimizations and tuning
  - Obtaining guidelines on best practices
  - Periodic information on major fixes and updates
  - Information on major releases
  - Help with upgrading to the latest release
  - Flexible Service Level Agreements
- **Support provided to Lawrence Livermore National Laboratory (LLNL) for the last two years**

X-ScaleSolutions

Silver ISV Member for the OpenPOWER Consortium + Products

- Has joined the OpenPOWER Consortium as a silver ISV member
- Provides flexibility:
  - To have MVAPICH2, HiDL and HiBD libraries getting integrated into the OpenPOWER software stack
  - A part of the OpenPOWER ecosystem
  - Can participate with different vendors for bidding, installation and deployment process
- Introduced two new integrated products with support for OpenPOWER systems (Presented at the OpenPOWER North America Summit)
  - X-ScaleHPC
  - X-ScaleAI
  - Send an e-mail to contactus@x-scalesolutions.com for free trial!!

X-ScaleSolutions

X-ScaleHPC Package

- Scalable solutions of communication middleware based on OSU MVAPICH2 libraries
- **“out-of-the-box”** fine-tuned and optimal performance on various HPC systems including OpenPOWER platforms
- Contact us for more details and a free trial!!
  - contactus@x-scalesolutions.com
- Stop by X-ScaleSolutions booth (#2094) for a Demo!!

X-ScaleAI Package

- High-Performance and scalable solutions for deep learning
  - Fully exploiting HPC resources using our X-ScaleHPC package
- "out-of-the-box" optimal performance on OpenPOWER (POWER9) + GPU platforms such as #1 Summit system
- What's in the X-ScaleAI package?
  - Fine-tuned CUDA-Aware MPI library
  - Google TensorFlow framework built with OpenPOWER system
  - Distributed training using Horovod on top of TensorFlow
  - Simple installation and execution in one command!
- Contact us for more details and a free trial!!
  - contactus@x-scalesolutions.com
- Stop by X-ScaleSolutions booth (#2094) for a Demo!!

Concluding Remarks

- Upcoming Exascale systems need to be designed with a holistic view of HPC, Big Data, Deep Learning, and Cloud
- OpenPOWER, InfiniBand, and NVIDIA GPGPUs are emerging technologies for such systems
- Presented a set of solutions from OSU to enable HPC, Big Data and Deep Learning through a convergent software architecture for OpenPOWER platforms
- X-ScaleSolutions is an ISV provider in the OpenPOWER consortium to provide commercial support, optimizations, tuning and training for the OSU solutions
- OpenPOWER users are encouraged to take advantage of these solutions to extract highest performance and scalability for their applications on OpenPOWER platforms

Multiple Events at SC '19

- Presentations at OSU and X-Scale Booth (#2094)
  - Members of the MVAPICH, HiBD and HiDL members
  - External speakers
- Presentations at SC main program (Tutorials, Workshops, BoFs, Posters, and Doctoral Showcase)
- Presentation at many other booths (Mellanox, Intel, Microsoft, and AWS) and satellite events
- Complete details available at

  http://mvapich.cse.ohio-state.edu/conference/752/talks/

Funding Acknowledgments

Funding Support by

Equipment Support by

Personnel Acknowledgments

Current Students (Graduate)

- A. Awan (Ph.D.)
- M. Bayatpour (Ph.D.)
- C.-H. Chu (Ph.D.)
- J. Hashmi (Ph.D.)
- A. Jain (Ph.D.)
- K. S. Kandadi (M.S.)
- Kamal Raj (M.S.)
- K. S. Khorassani (Ph.D.)
- P. Kousha (Ph.D.)
- A. Quentin (Ph.D.)
- B. Ramesh (M. S.)
- S. Xu (M.S.)
- Q. Zhou (Ph.D.)

Current Research Scientist

- H. Subramoni

Current Students (Undergraduate)

- V. Gangal (B.S.)
- N. Sarkauskas (B.S.)

Current Post-doc

- M. S. Ghazimeersaeed
- A. Ruhela
- K. Manian

Current Research Specialist

- J. Smith

Past Students

- A. Augustine (M.S.)
- P. Balaji (Ph.D.)
- R. Biswas (M.S.)
- S. Bhagvat (M.S.)
- A. Bhat (M.S.)
- D. Buntinas (Ph.D.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- S. Chakraborthy (Ph.D.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- W. Huang (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (Ph.D.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- K. Kulkarni (M.S.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (Ph.D.)
- M. Li (Ph.D.)
- P. Lai (M.S.)
- J. Liu (Ph.D.)
- M. Luo (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- V. Meshram (M.S.)
- A. Moody (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- X. Ouyang (Ph.D.)
- S. Pai (M.S.)
- S. Potluri (Ph.D.)
- R. Rajachandrasekar (Ph.D.)
- D. Shankar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- A. Singh (Ph.D.)
- J. Sridhar (M.S.)
- S. Sur (Ph.D.)
- H. Subramoni (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)
- J. Zhang (Ph.D.)

Past Research Scientist

- K. Hamidouche
- S. Sur
- X. Lu

Past Programmers

- D. Bureddy
- J. Perkins

Past Research Specialist

- M. Arnold

Past Post-Docs

- D. Banerjee
- X. Besseron
- H.-W. Jin
- J. Lin
- M. Luo
- E. Mancini
- S. Marcarelli
- J. Vienne
- H. Wang

Thank You!

panda@cse.ohio-state.edu

Follow us on

https://twitter.com/mvapich

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/

HiBD
High-Performance Big Data

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/

HiDL
High-Performance Deep Learning

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