MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiBD
High-Performance
Big Data

HiDL
High-Performance
Deep Learning

High-Performance and Scalable Middleware for HPC, AI and Data Science on Heterogenous Systems

Talk at the 2023 HPC-AI Advisory Council Japan Conference

by

Follow us on

https://twitter.com/mvapich

Dhabaleswar K. (DK) Panda

The Ohio State University

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

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

Drivers of Modern HPC Cluster Architectures

Multi-/Many-core Processors

High Performance Interconnects – InfiniBand (DPU), Slingshot <1usec latency, 200-400Gbps Bandwidth>

Accelerators high compute density, high performance/watt >9.7 TFlop DP on a chip

SSD, NVMe-SSD, NVRAM

- Multi-core/many-core technologies
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand, RoCE, Slingshot)
- Solid State Drives (SSDs), Non-Volatile Random-Access Memory (NVRAM), NVMe-SSD
- Accelerators (NVIDIA GPGPUs)
- Available on HPC Clouds, e.g., Amazon EC2, NSF Chameleon, Microsoft Azure, etc.

Frontier

Fugaku

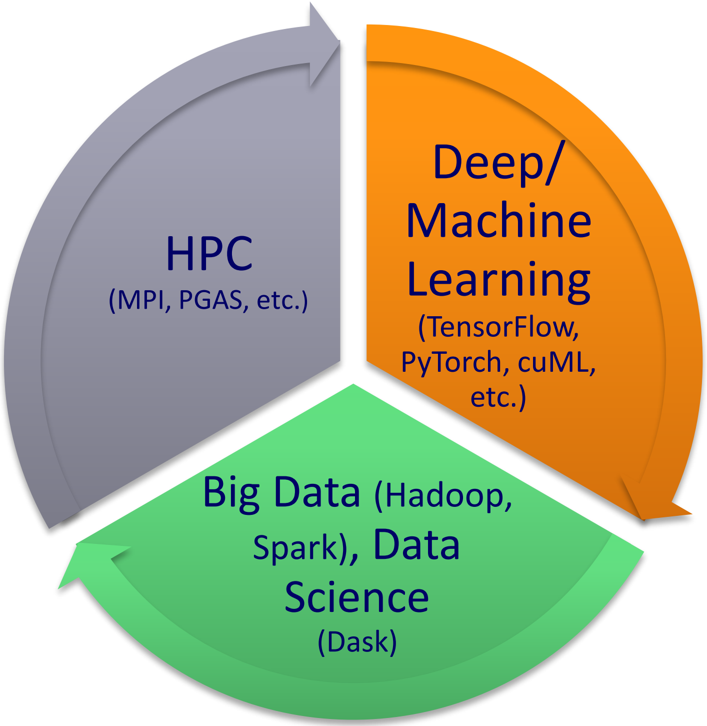
Summit

Lumi

Broad Challenge:

How to design high-performance and scalable middleware for HEC systems while taking advantage of heterogeneous (CPU + GPU + DPU/IPU (xPU)) HPC and Cloud resources?

Increasing Usage of HPC, AI, Big Data and Data Science



HPC
(MPI, PGAS, etc.)

Deep/ Machine Learning
(TensorFlow, PyTorch, cuML, etc.)

Big Data (Hadoop, Spark), Data Science
(Dask)

Convergence of HPC, Deep/Machine Learning, and Data Science!

Increasing Need to Run these applications on the Cloud!!

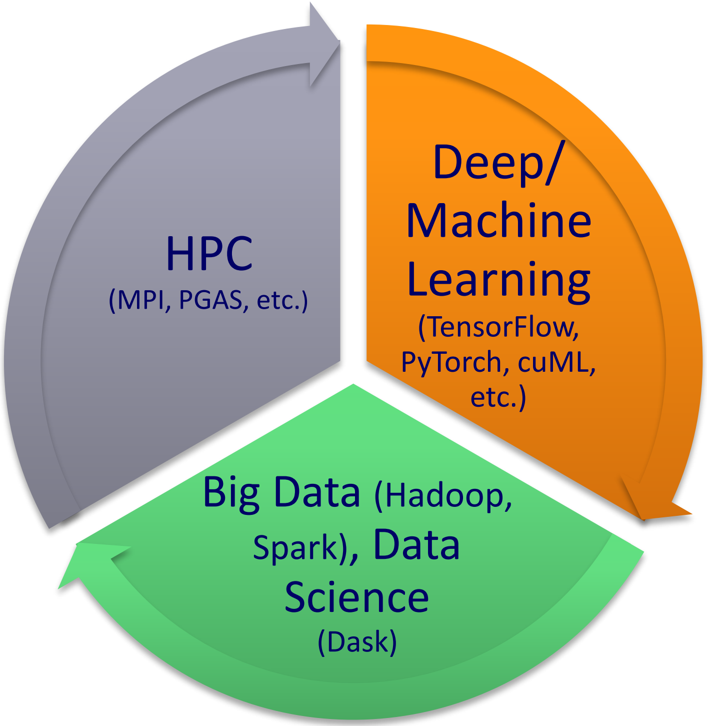
Can MPI-Driven Middleware be designed and used for all three domains?

Presentation Overview

- MVAPICH MPI Library Project
  - High-Performance Support for various CPU, GPU, DPU, and Networking Technologies
- HiDL Project
  - High-Performance Deep Learning and Machine Learning
  - Accelerating Deep Learning with DPU
- HiBD Project
  - Accelerating Spark with MPI
  - Accelerating Data Science Applications with Dask
- Commercial Support and Value-Added Products
- Conclusions

Converged Middleware for HPC, AI, Big Data and Data Science

MVAPICH2 &
MVAPICH2-DPU
Libraries



HPC
(MPI, PGAS, etc.)

Deep/
Machine
Learning
(TensorFlow,
PyTorch, cuML,
etc.)

Big Data (Hadoop,
Spark), Data
Science
(Dask)

Overview of the MVAPICH Project

- High Performance open-source MPI Library
- Support for multiple interconnects
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), AWS EFA, OPX, Broadcom RoCE, Intel Ethernet, Rockport Networks, Slingshot 10/11
- Support for multiple platforms
  - x86, OpenPOWER, ARM, Xeon-Phi, GPGPUs (NVIDIA and AMD)
- Started in 2001, first open-source version demonstrated at SC ‘02
- Supports the latest MPI-3.1 standard
- http://mvapich.cse.ohio-state.edu
- Additional optimized versions for different systems/environments:
  - MVAPICH2-X (Advanced MPI + PGAS), since 2011
  - MVAPICH2-GDR with support for NVIDIA (since 2014) and AMD (since 2020) GPUs
  - MVAPICH2-MIC with support for Intel Xeon-Phi, since 2014
  - MVAPICH2-Virt with virtualization support, since 2015
  - MVAPICH2-EA with support for Energy-Awareness, since 2015
  - MVAPICH2-Azure for Azure HPC IB instances, since 2019
  - MVAPICH2-X-AWS for AWS HPC+EFA instances, since 2019
- Tools:
  - OSU MPI Micro-Benchmarks (OMB), since 2003
  - OSU InfiniBand Network Analysis and Monitoring (INAM), since 2015

- **Used by more than 3,300 organizations in 90 countries**
- **More than 1.66 Million downloads from the OSU site directly**
- Empowering many TOP500 clusters (Nov ‘22 ranking)
  - 7th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 19th, 448, 448 cores (Frontera) at TACC
  - 34th, 288,288 cores (Lassen) at LLNL
  - 46th, 570,020 cores (Nurion) in South Korea and many others
- Available with software stacks of many vendors and Linux Distros (RedHat, SuSE, OpenHPC, and Spack)
- Partner in the 19th ranked TACC Frontera system
- **Empowering Top500 systems for more than 17 years**

MVAPICH-3.0a

- Released on 08/15/2022
- Based on MPICH 3.4.3
- Added support for the ch4:ucx and ch4:ofi devices
- Support for MVAPICH2 enhanced collectives over OFI and UCX
- Added support for the Cray Slingshot 11 interconnect over OFI
  - Supports Cray Slingshot 11 network adapters
- Added support for the Cornelis OPX library over OFI
  - Supports Intel Omni-Path adapters
- Added support for the Intel PSM3 library over OFI
  - Supports Intel Columbiaville network adapters
- Added support for IB verbs over UCX
  - Supports IB and RoCE network adapters

MVAPICH-Plus 3.0a

- Released on 11/11/2022
- Based on MVAPICH 3.0
- Advanced MPI with unified MVAPICH2-GDR and MVAPICH2-X features
- Support for both OFI and UCX
- Support for NVIDIA and AMD GPUs
- Support for all HPC interconnects
  - InfiniBand, Omni-Path, ROCE, Slingshot
- Optimized designs for HPC, DL, ML, Big Data and Data Science applications

Unify MVAPICH2-GDR and MVAPICH2-X

Exploit the benefits of in-network computing capabilities in UCX and OFI transparently

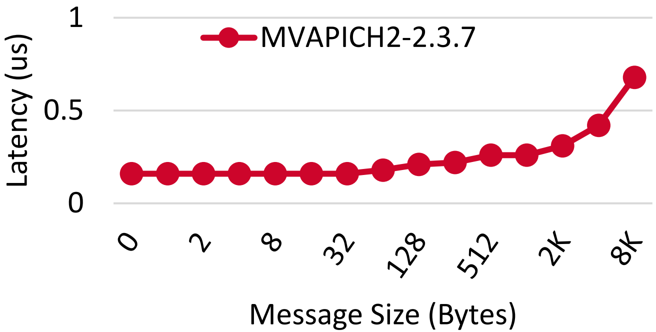
Presentation Overview

- MVAPICH MPI Library Project
  - RDMA-Enabled Designs
  - CUDA-Awareness Designs for GPUs
  - On-the-fly Compression for GPUs
  - Accelerating applications with DPU
- HiDL Project
  - High-Performance Deep Learning and Machine Learning
  - Accelerating Deep Learning with DPU
- HiBD Project
  - Accelerating Spark with MPI
  - Accelerating Data Science Applications with Dask
- Commercial Support and Value-Added Products
- Conclusions

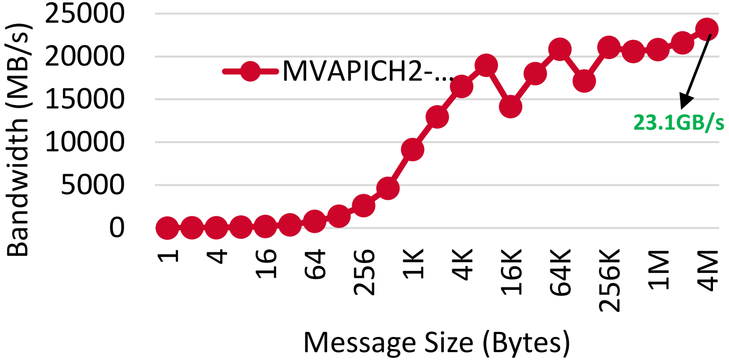
AMD Milan + HDR 200

Intra-Node CPU Point-to-Point

Latency



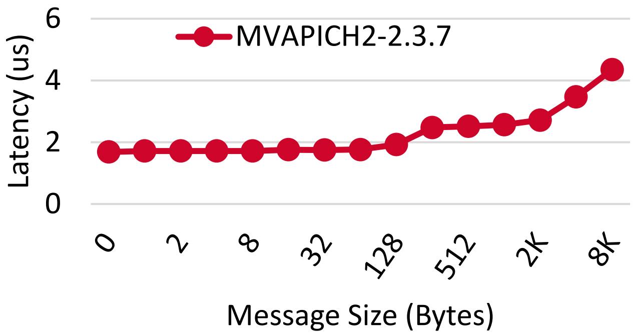
Bandwidth



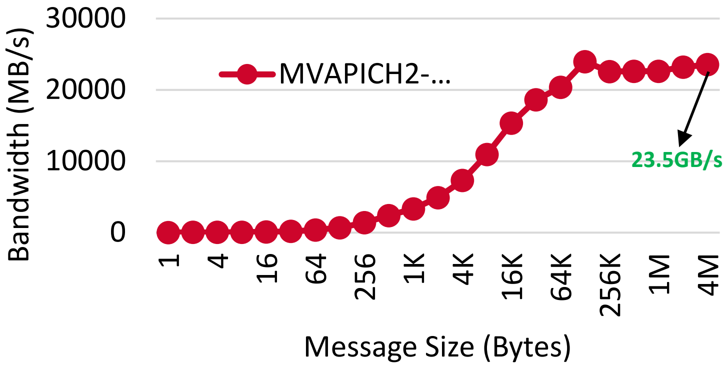
23.1GB/s

Inter-Node CPU Point-to-Point

Latency



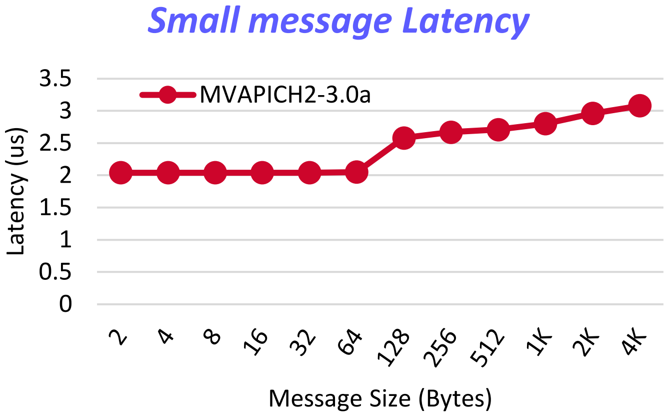
Bandwidth

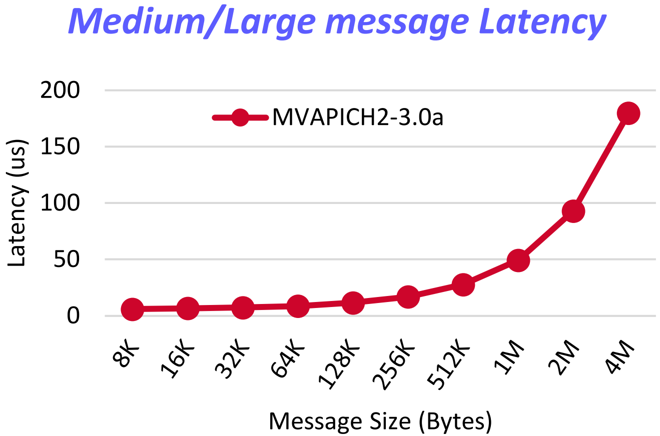


23.5GB/s

AMD EPYC 7V73X 64-Core Processor, Mellanox ConnectX-6 HDR HCA

MPI Level Latency on Slingshot 11





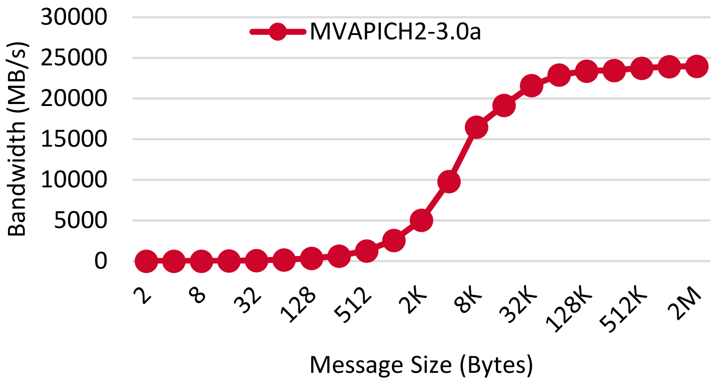
- **2us** inter-node point-to-point latency for small messages

Interconnect : Cray HPE Slingshot 11
Library : MVAPICH2 3.0a
CPU : AMD EPYC 7763 (milan) Processor

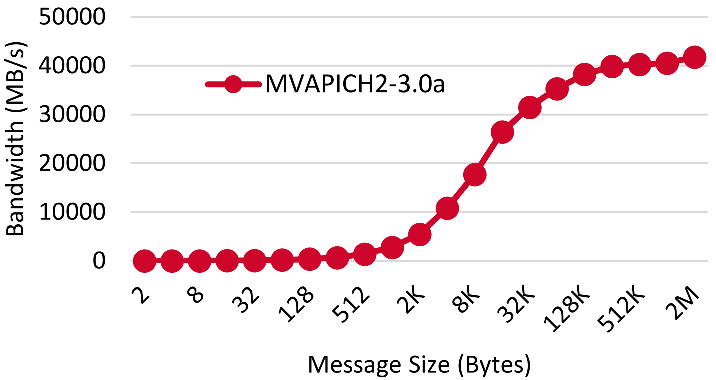
Available with MVAPICH2 3.0a Release

MPI Level Bandwidth on Slingshot 11

Uni-directional Bandwidth



Bi-Directional Bandwidth

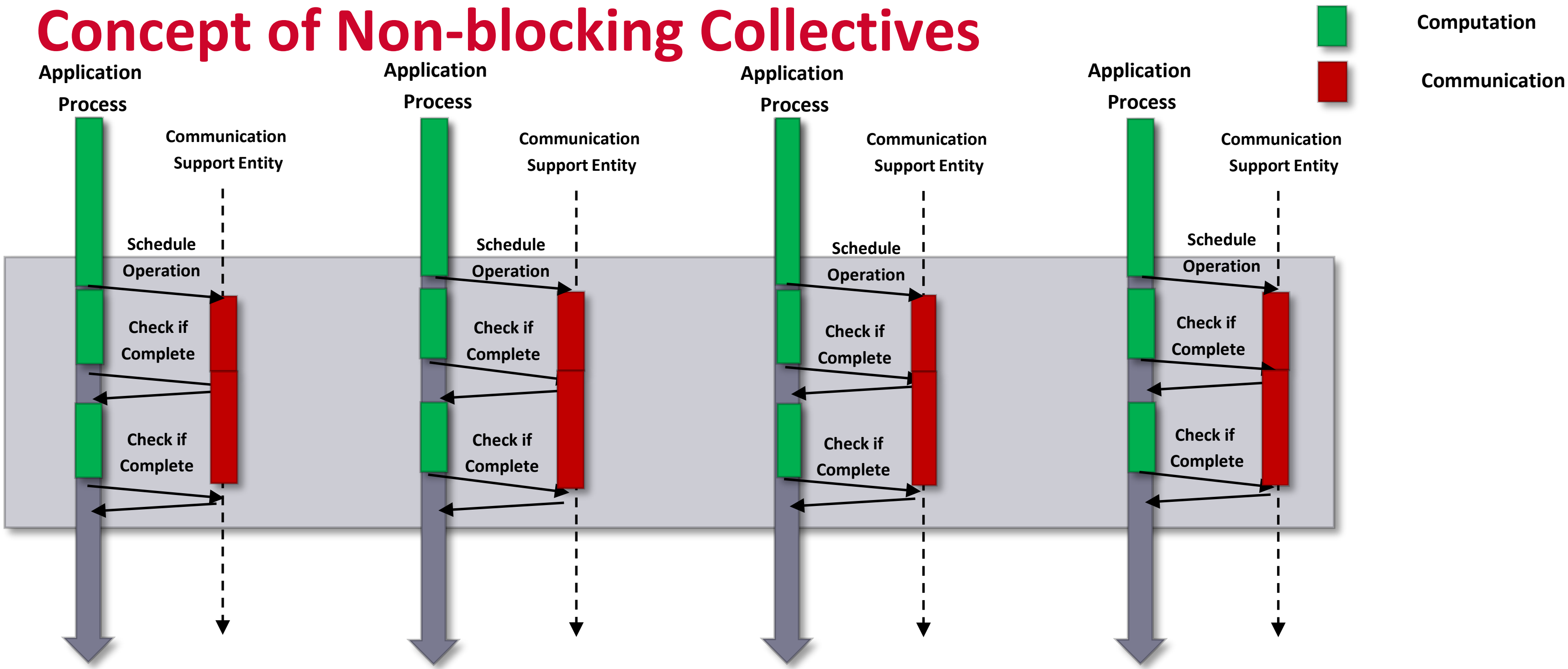


- **23,985 MB/s** uni-directional peak bandwidth
- **42,034 MB/s** bi-directional peak bandwidth

Available with MVAPICH2 3.0a Release

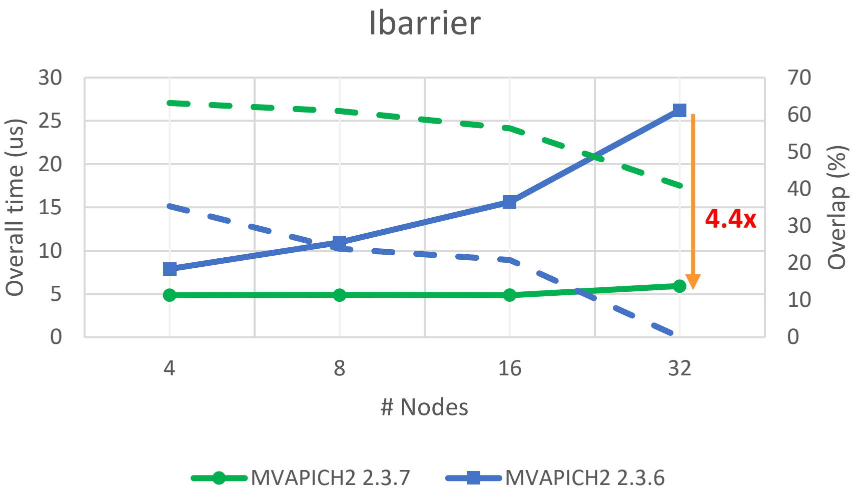
Interconnect : Cray HPE Slingshot 11 (200 Gbps)
Library : MVAPICH2 3.0a
CPU : AMD EPYC 7763 (milan) Processor

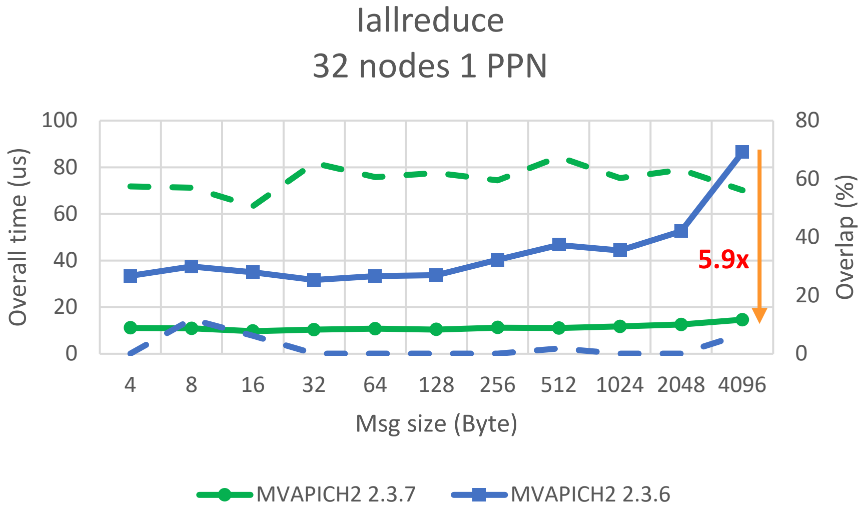
Concept of Non-blocking Collectives



- Application processes schedule collective operation
- Check periodically if operation is complete
- **Overlap of computation and communication => Better Performance**
- *Catch: Who will progress communication*

Non-blocking Collectives Support with In-Network Computing





- With SHARP:
  - Flat scaling in terms of overall time
  - High overlap between computation and communication

Platform: Dual-socket Intel(R) Xeon(R) Platinum 8280 CPU @ 2.70GHz nodes equipped with Mellanox InfiniBand, HDR-100 Interconnect

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

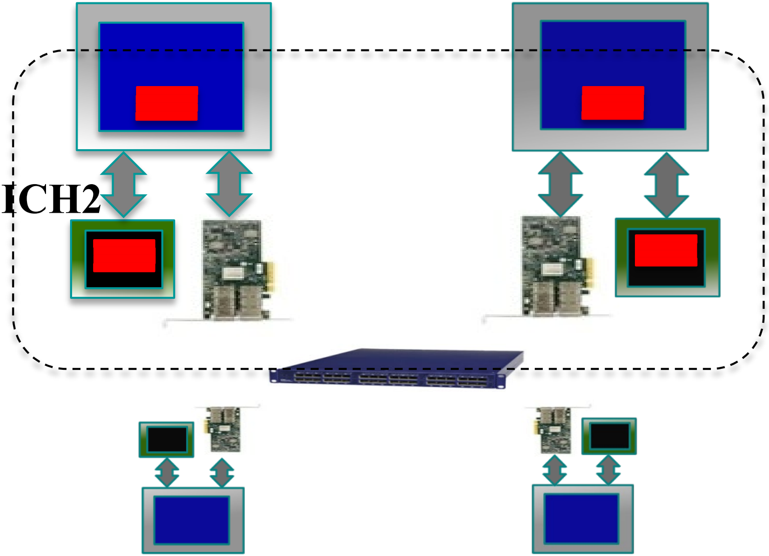
At Sender:

```
MPI_Send(s_devbuf, size, …);
```

At Receiver:

```
MPI_Recv(r_devbuf, size, …);
```

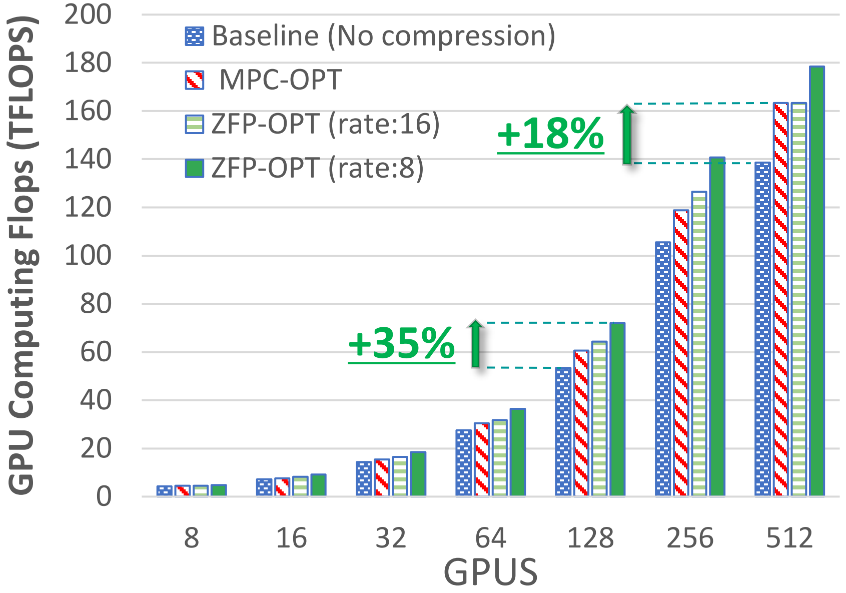
High Performance and High Productivity

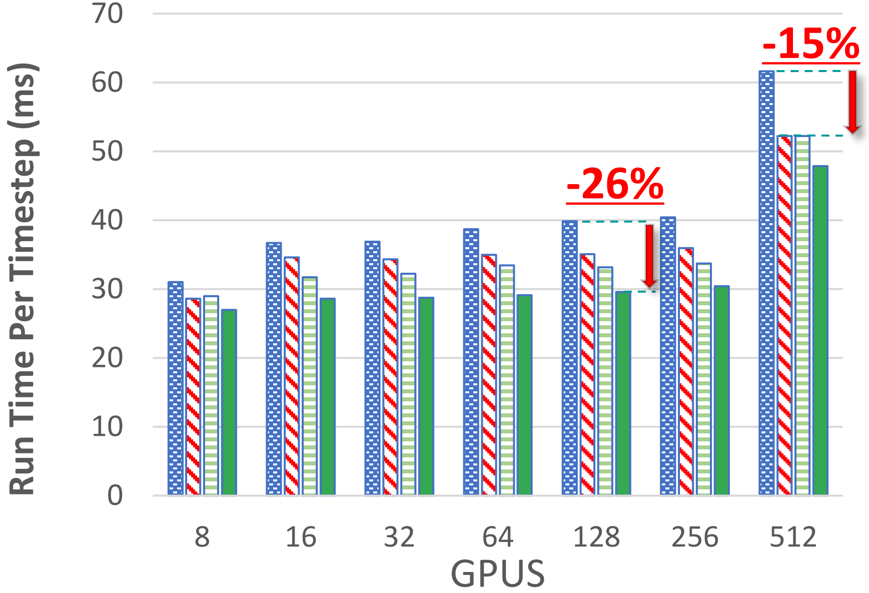


inside MVAPICH2

“On-the-fly” Compression Support in MVAPICH2-GDR

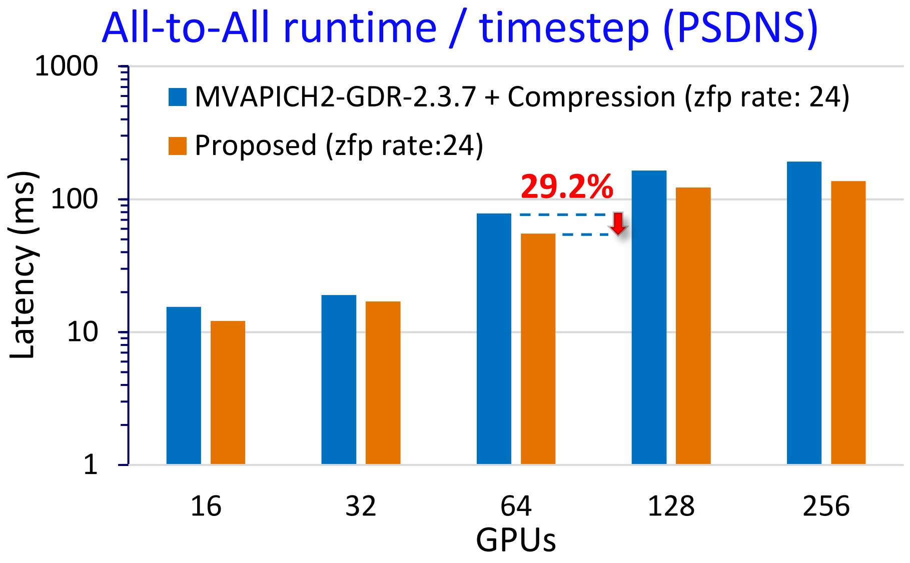
- Weak-Scaling of HPC application **AWP-ODC** on Lassen cluster (V100 nodes)
- MPC-OPT achieves up to **+18%** GPU computing flops, **-15%** runtime per timestep
- ZFP-OPT achieves up to **+35%** GPU computing flops, **-26%** runtime per timestep

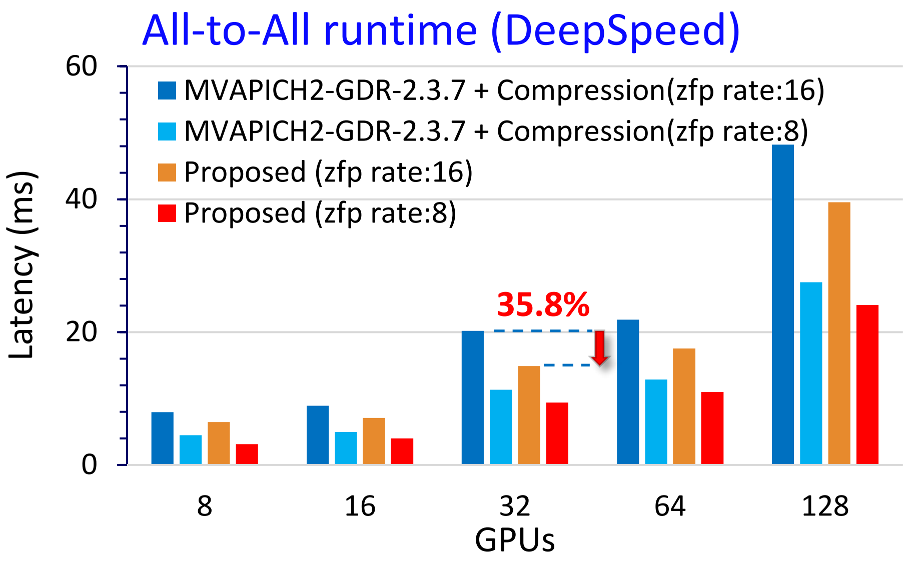




Q. Zhou, C. Chu, N. Senthil Kumar, P. Kousha, M. Ghazimirsaeed, H. Subramoni, and D.K. Panda, Designing High-Performance MPI Libraries with On-the-fly Compression for Modern GPU Clusters, 35th IEEE International Parallel & Distributed Processing Symposium (IPDPS), May 2021. [Best Paper Finalist]

Performance of All-to-All with Online Compression





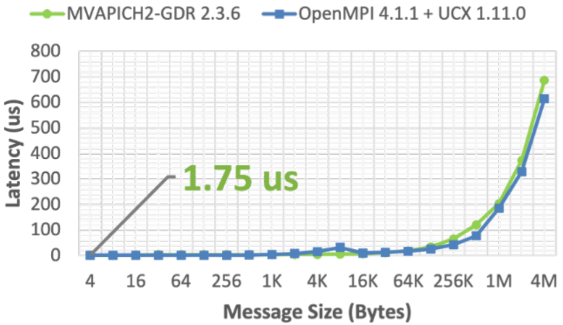
- Improvement compared to MVAPICH2-GDR-2.3.7 with Point-to-Point compression
  - 3D-FFT: Reduce All-to-All runtime by up to 29.2% with ZFP(rate: 24) on 64 GPUs
  - DeepSpeed benchmark: Reduce All-to-All runtime by up to 35.8% with ZFP(rate: 16) on 32 GPUs

Q. Zhou, P. Kousha, Q. Anthony, K. Khorassani, A. Shafi, H. Subramoni, and D.K. Panda, "Accelerating MPI All-to-All Communication with Online Compression on Modern GPU Clusters", ISC '22.

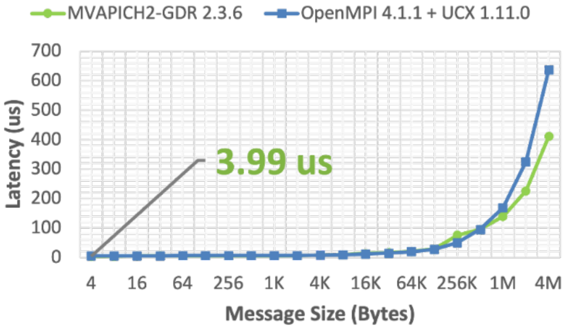
Available in MVAPICH2-GDR 2.3.7

ROCm-aware MVAPICH2-GDR - Support for AMD GPUs

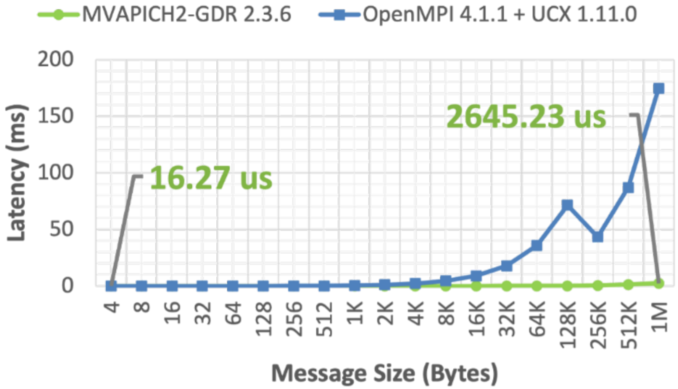
Intra-Node Point-to-Point Latency



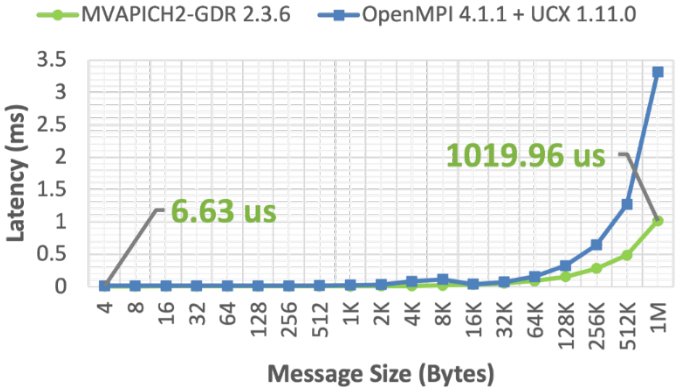
Inter-Node Point-to-Point Latency



Allreduce – 64 GPUs (8 nodes, 8 GPUs Per Node)



Bcast – 64 GPUs (8 nodes, 8 GPUs Per Node)



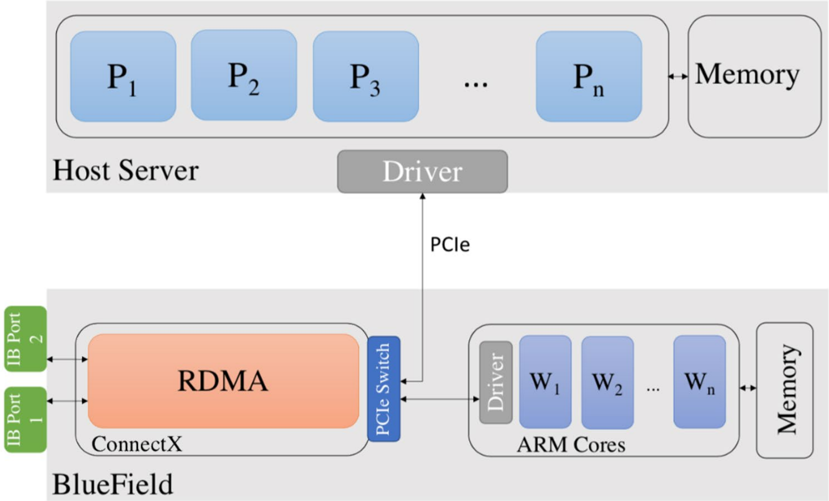
Corona Cluster @ LLNL - ROCm-4.3.0 (mi50 AMD GPUs)

Available with MVAPICH2-GDR 2.3.5+ & OMB v5.7+

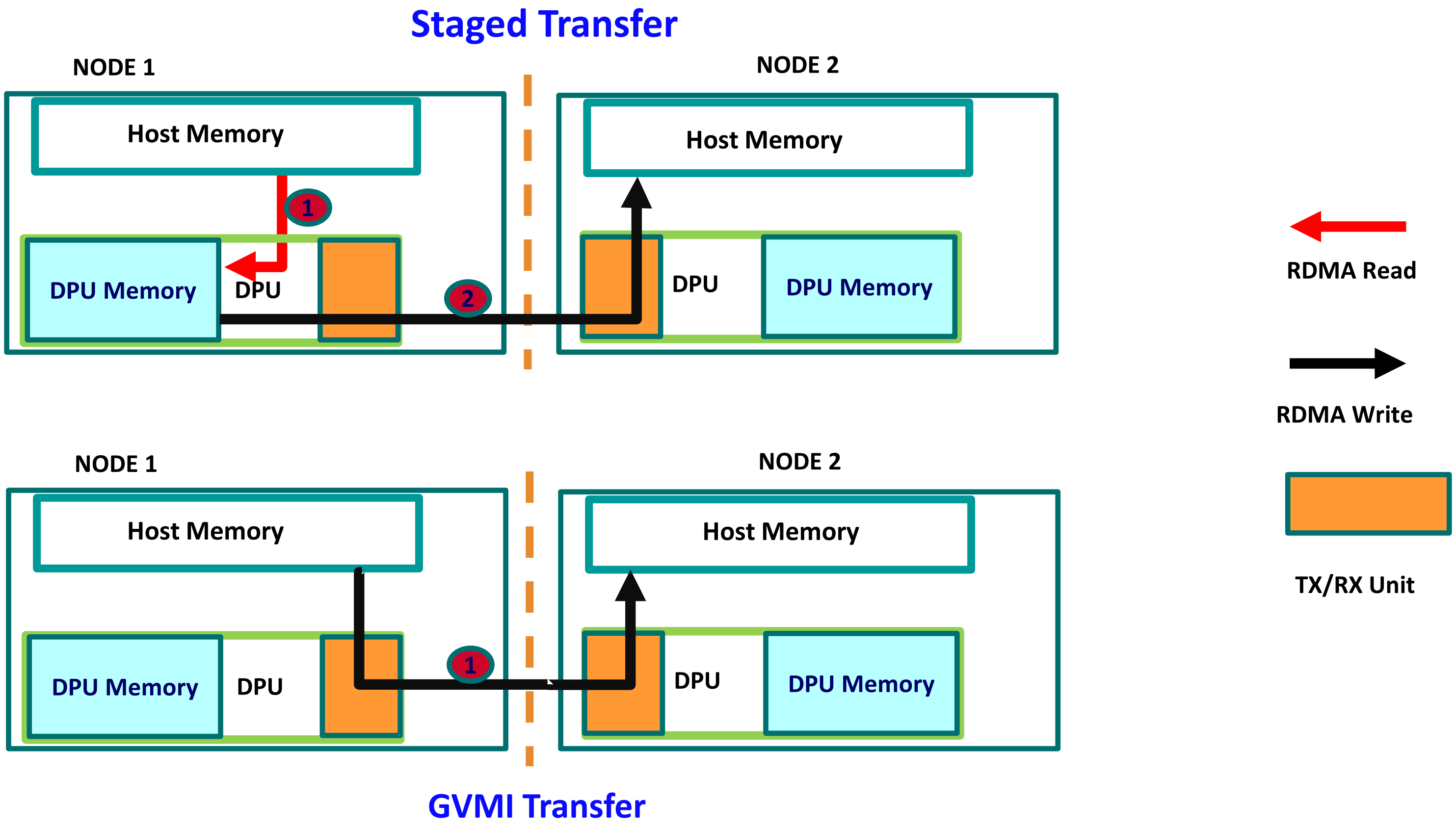
K. Khorassani, J. Hashmi, C. Chu, C. Chen, H. Subramoni, D. Panda Designing a ROCm-aware MPI Library for AMD GPUs: Early Experiences - ISC HIGH PERFORMANCE 2021, Jun 2021.

Accelerating Applications with BlueField-3 DPU

- InfiniBand network adapter with up to 400Gbps speed
- System-on-chip containing 16 64-bit ARMv8.2 A78 cores with 2.75 GHz each
- 16 GB of memory for the ARM cores



Staging vs. GVMI



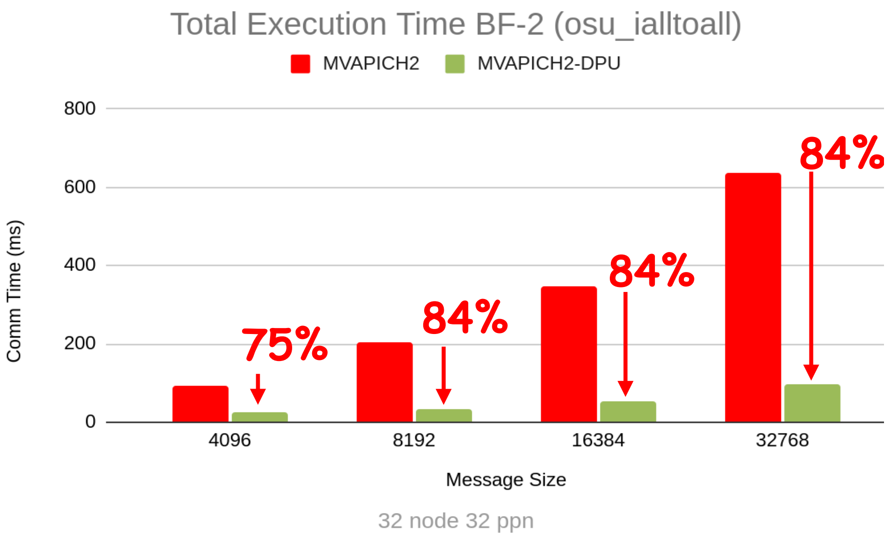
MVAPICH2-DPU 2023.04 Library Release

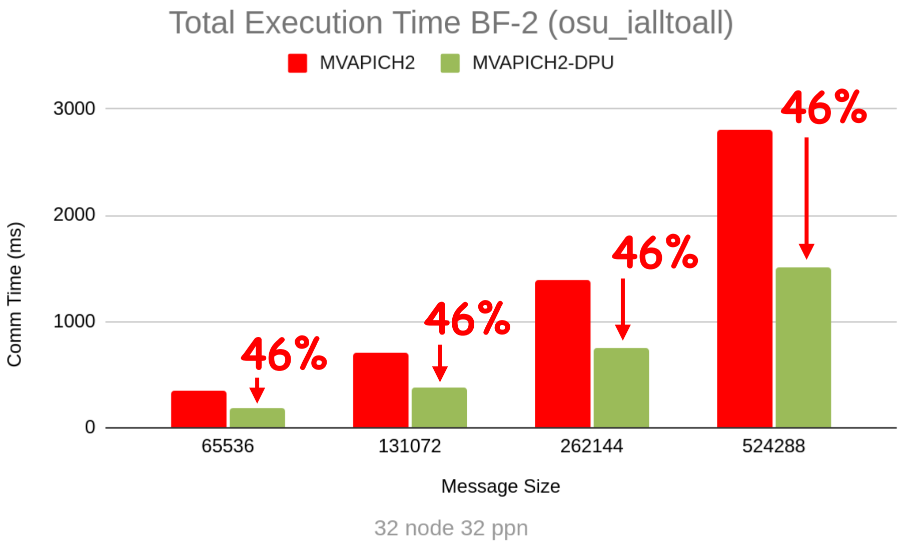
X-ScaleSolutions

- Released on 04/02/23
- Supports all features available with the MVAPICH2 release (http://mvapich.cse.ohio-state.edu)
- Novel GVMI-based framework to offload non-blocking collectives to DPU
- Offloads non-blocking Alltoall (MPI_Ialltoall) to DPU

Available from X-ScaleSolutions, please send a note to contactus@x-scalesolutions.com to get a trial license.

Total Execution Time with osu_Ialltoall (32 nodes), GVMI, BF-2



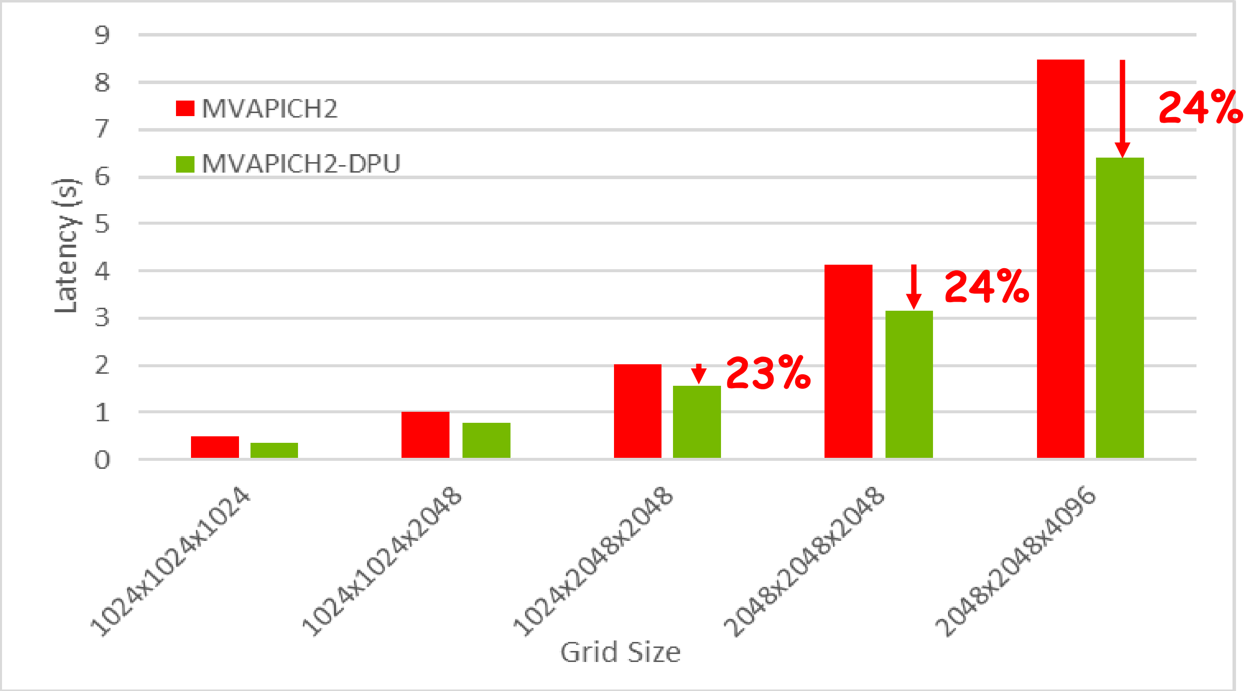


Benefits in Total execution time (Compute + Communication)

P3DFFT Application Execution Time (32 nodes), GVMI, BF-2

32 nodes with 32 ppn
(1,024 processes)

32x32 process grid



Benefits in application-level execution time

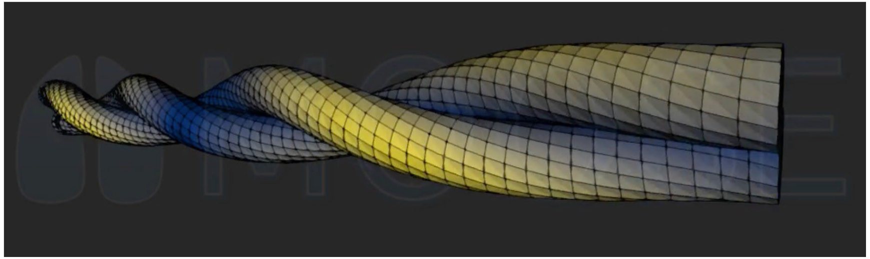
MVAPICH Drives Nuclear Energy Research at Idaho National Lab (INL)

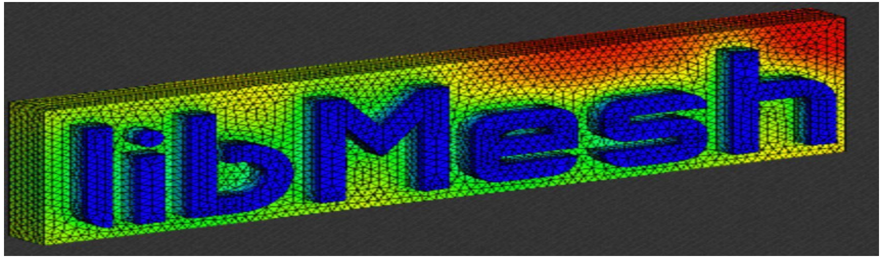
MOOSE

Multiphysics Object-Oriented Simulation Environment

PETSc

An open-source, parallel finite element framework



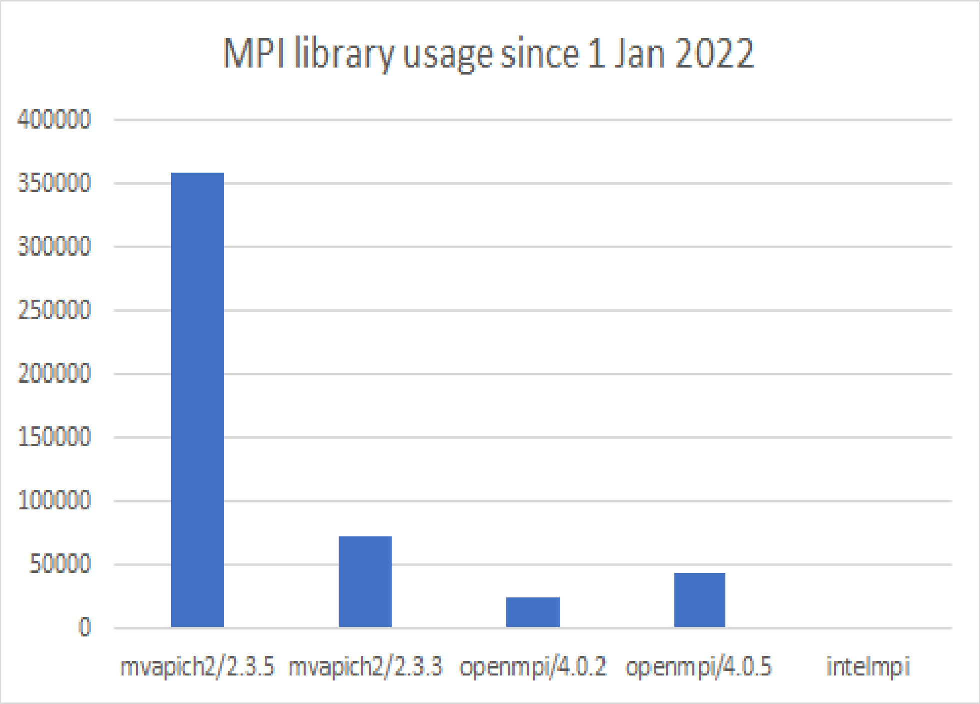


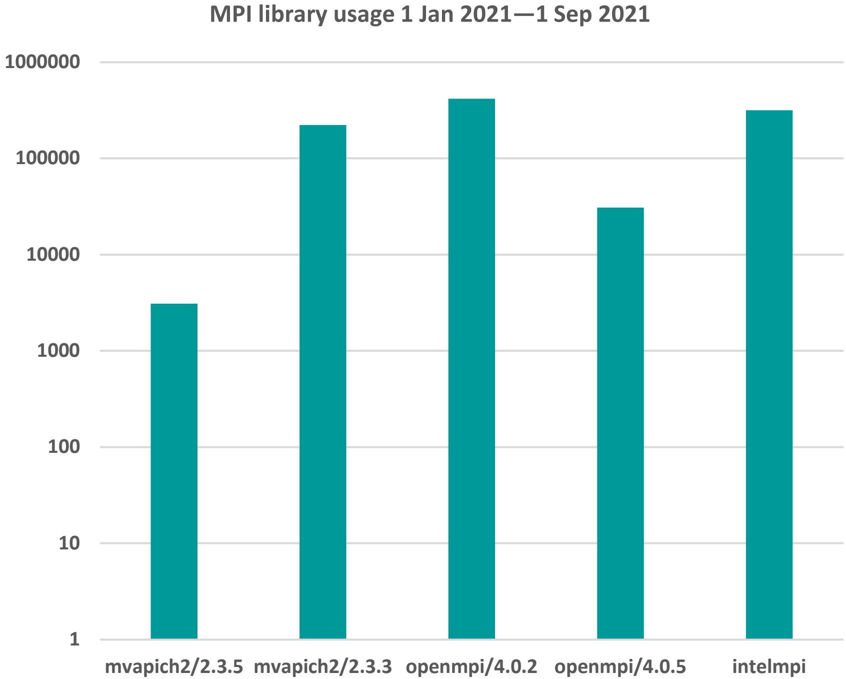
The MOOSE Multiphysics Computational Framework for Nuclear Power Applications: A Special Issue of Nuclear Technology (https://www.tandfonline.com/doi/full/10.1080/00295450.2021.1915487)

MVAPICH Integration for PBS Pro, HPC Team, Idaho National Laboratory

(http://mug.mvapich.cse.ohio-state.edu/static/media/mug/presentations/21/inl.pdf)

Rapid adoption of MVAPICH2 on INL HPC systems



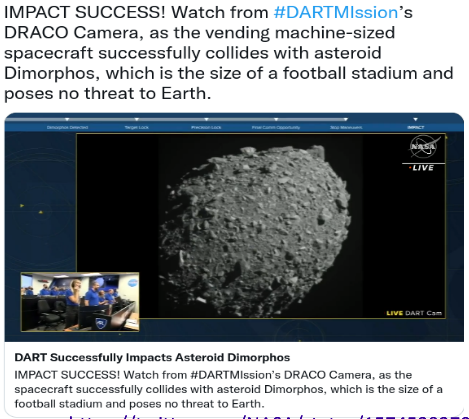


M. Anderson, Aggressive Asynchronous Communication in the MOOSE framework using MVAPICH2, 10th Annual MVAPICH User Group Conference (MUG), Aug 2022

MVAPICH2 enabling life-changing NASA's DART mission

- Near-Earth asteroids (NEAs) have caused recent and ancient global catastrophes
  - LLNL scientists research ways to prevent NEAs using methods known as asteroid deflection
  - Joint NASA-LLNL research modelled various asteroid deflection methods (NASA's DART mission)
- MVAPICH2 lived at the core of the (NASA-DART mission) and enabled scalability
  - Underneath large-scale hydrodynamical and gravitational simulations required to compute the impact such as Spheral models

NASA
@NASA



IMPACT SUCCESS! Watch from #DARTMIssion's DRACO Camera, as the vending machine-sized spacecraft successfully collides with asteroid Dimorphos, which is the size of a football stadium and poses no threat to Earth.

DART Successfully Impacts Asteroid Dimorphos

IMPACT SUCCESS! Watch from #DARTMIssion's DRACO Camera, as the spacecraft successfully collides with asteroid Dimorphos, which is the size of a football stadium and poses no threat to Earth.

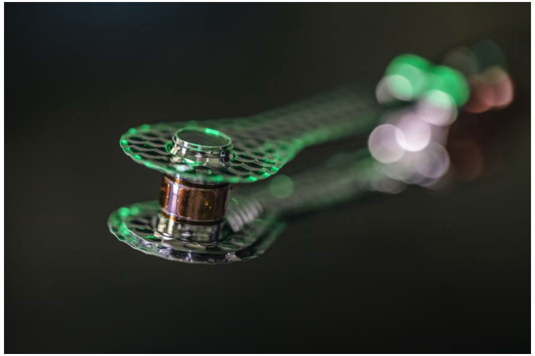
- https://twitter.com/NASA/status/1574539270987173903?s=20&t=u_4wIV9Cui2xyn9QLj286Q
- https://www.cbsnews.com/sanfrancisco/news/i-just-could-not-believe-it-livermore-team-celebrates-nasas-historic-strike-on-distant-asteroid/
- http://mug.mvapich.cse.ohio-state.edu/static/media/mug/presentations/18/moody-mug-18.pdf

MVAPICH enabling Nuclear Fusion Research

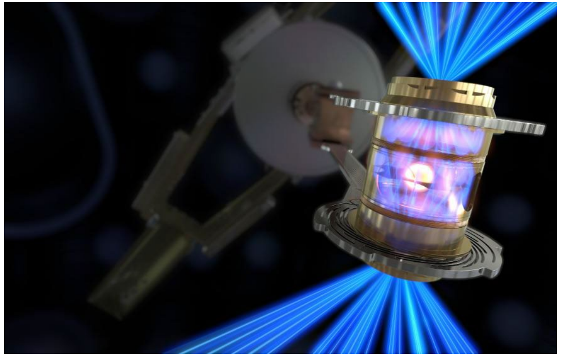
- LLNL’s National Ignition Facility (NIF) conducted the first controlled fusion experiment in history![1]
- MVAPICH, being the default MPI library on the LLNL systems, has been enabling the thousands of simulation jobs that have led to this amazing achievement!
- [1] https://www.llnl.gov/news/national-ignition-facility-achieves-fusion-ignition

IGNITION
U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration
Lawrence Livermore National Laboratory

The target chamber of LLNL’s National Ignition Facility, where 192 laser beams delivered more than 2 million joules of ultraviolet energy to a tiny fuel pellet to create fusion ignition on Dec. 5, 2022.



The hohlraum that houses the type of cryogenic target used to achieve ignition on Dec. 5, 2022, at LLNL’s National Ignition Facility.



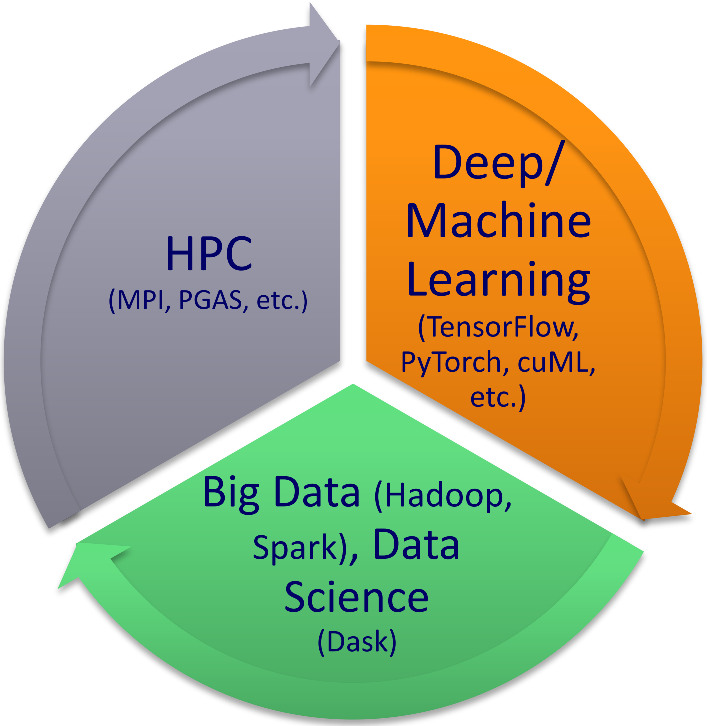
To create fusion ignition, the National Ignition Facility’s laser energy is converted into X-rays inside the hohlraum, which then compress a fuel capsule until it implodes, creating a high temperature, high pressure plasma.

Presentation Overview

- MVAPICH MPI Library Project
  - High-Performance Support for various CPU, GPU, DPU, and Networking Technologies
- **HiDL Project**
  - **High-Performance Deep Learning and Machine Learning**
  - **Accelerating Deep Learning with DPU**
- HiBD Project
  - Accelerating Spark with MPI
  - Accelerating Data Science Applications with Dask
- Commercial Support and Value-Added Products
- Conclusions

Converged Middleware for HPC, AI, Big Data and Data Science

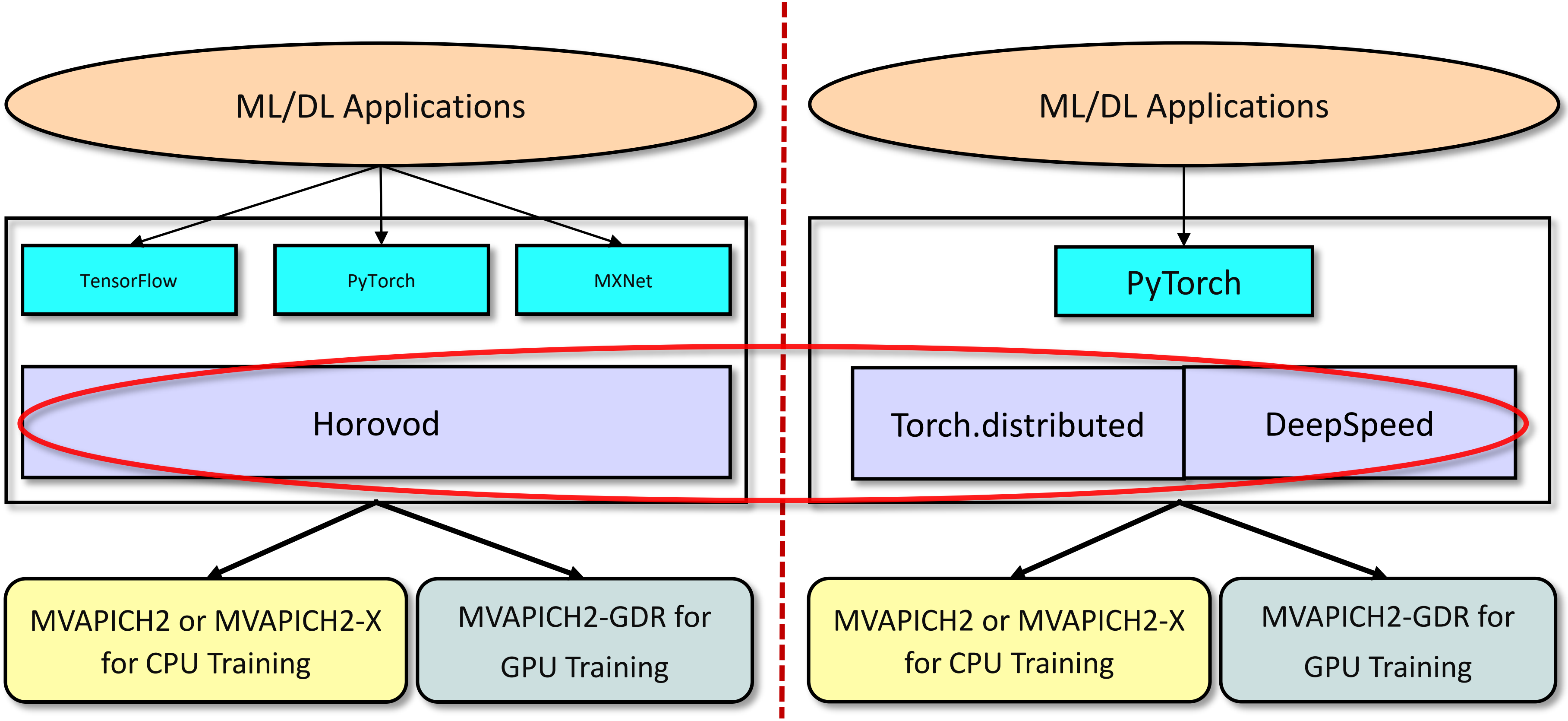
MVAPICH2 & MVAPICH2-DPU Libraries


HPC
(MPI, PGAS, etc.)
Deep/ Machine Learning
(TensorFlow, PyTorch, cuML, etc.)
Big Data (Hadoop, Spark), Data Science
(Dask)

High-Performance DL/ML Libraries with MVAPICH

X-Scale-AI-DPU Library With MVAPICH

MVAPICH2 (MPI)-driven Infrastructure for ML/DL Training


ML/DL Applications
TensorFlow
PyTorch
MXNet
Horovod
MVAPICH2 or MVAPICH2-X for CPU Training
MVAPICH2-GDR for GPU Training
ML/DL Applications
PyTorch
Torch.distributed
DeepSpeed
MVAPICH2 or MVAPICH2-X for CPU Training
MVAPICH2-GDR for GPU Training

More details available from: http://hidl.cse.ohio-state.edu

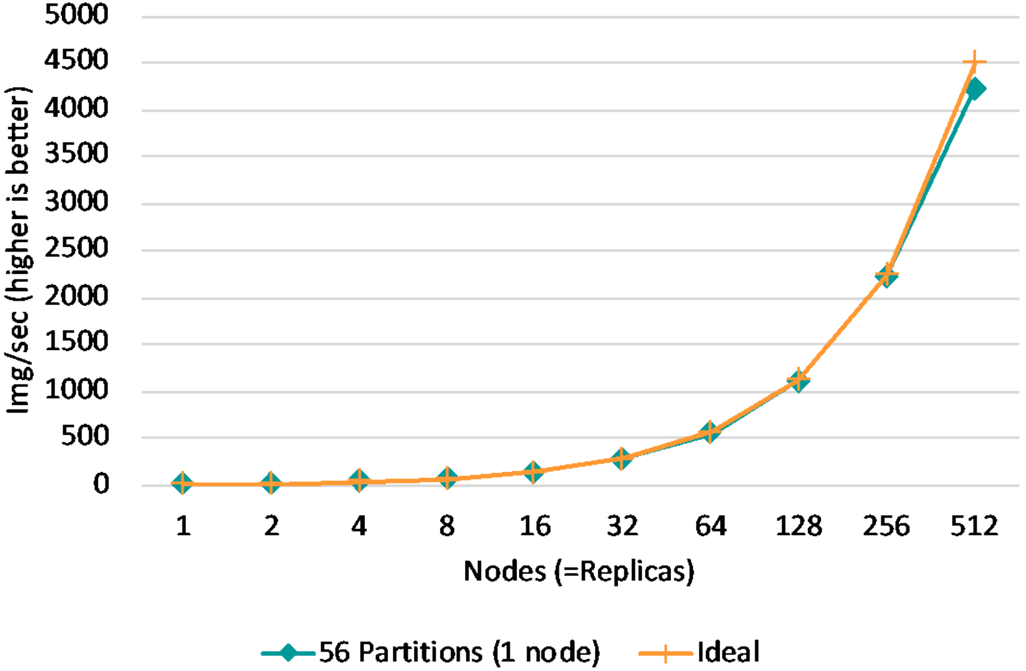
Sample Designs and Solutions

- Exploiting Hybrid (Data and Model) Parallelism for out-of-core training
- Accelerating Deep Learning with DPU
- Accelerating DL-Checkpointing with DPU

HyPar-Flow with Out-of-Core Training at Scale (512 nodes on TACC Frontera)

- ResNet-1001 with variable batch size
- Approach:
  - 48 model-partitions for 56 cores
  - 512 model-replicas for 512 nodes
  - Total cores: 56 x 512 = 28,672
- Speedup
  - **253X** on 256 nodes
  - **481X** on 512 nodes
- Scaling Efficiency
  - **98%** up to 256 nodes
  - **93.9%** for 512 nodes

481x speedup on 512 Intel Xeon Cascade Lake nodes (TACC Frontera)

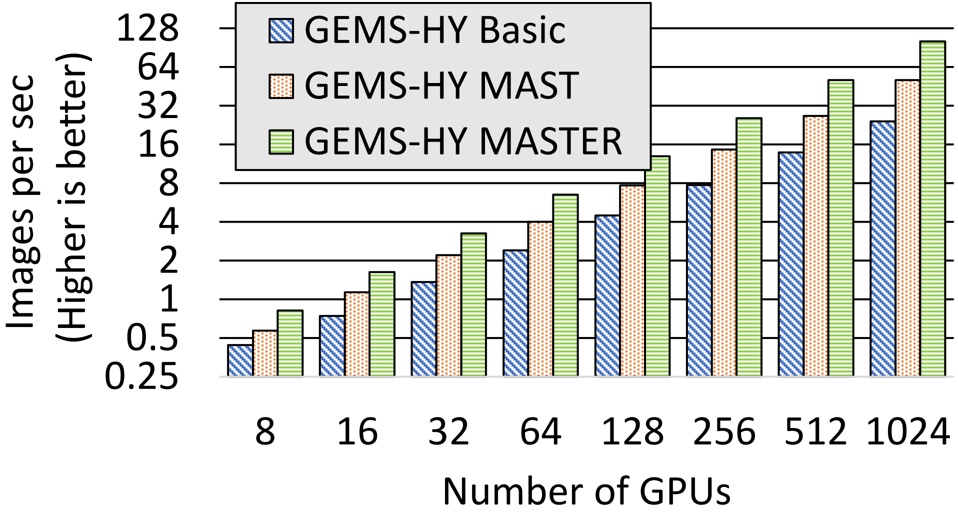


A. A. Awan, A. Jain, Q. Anthony, H. Subramoni, and DK Panda, "HyPar-Flow: Exploiting MPI and Keras for Hybrid Parallel Training of TensorFlow models", ISC '20, https://arxiv.org/pdf/1911.05146.pdf

Model Parallelism (GEMS) at Scale (1,024 V100 GPUs on LLNL Lassen)

- Two Approaches:
  - Memory Aware Synchronized Training (MAST)
  - Memory Aware Synchronized Training with Enhanced Replications (MASTER)
- Setup
  - ResNet-1k on 512 X 512 images
  - 128 Replications on 1024 GPUs
- Scaling Efficiency
  - **97.32%** on 1024 nodes

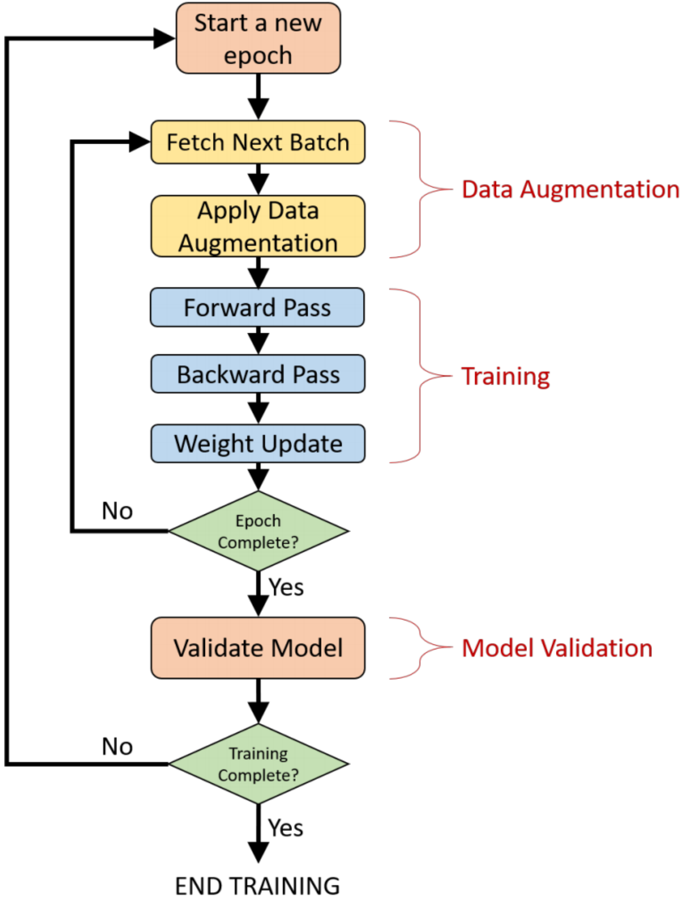
97.32% scaling efficiency on 1024 V100 GPUs (LLNL Lassen)



A. Jain, A. Awan, A. Aljuhani, J. Hashmi, Q. Anthony, H. Subramoni, D. Panda, R. Machiraju, A. Parwani, "GEMS: GPU Enabled Memory Aware Model Parallelism System for Distributed DNN", SC '20

Exploiting DPUs for Deep Neural Network Training

- There are several phases in Deep Neural Network Training
  - Fetching Training Data
  - Data Augmentation
  - Forward Pass
  - Backward Pass
  - Weight Update
  - Model Validation
- Different phases can be offloaded to DPUs to accelerate the training.



DPU Offloading Strategy

- Offloads data augmentation and model validation to DPUs.
- Creates three types of processes
  - Training processes (on CPU)
  - Data Augmentation processes (On DPU)
  - Testing processes (On DPU)

A. Jain, N. Alnaasan, A. Shafi, H. Subramoni, D. K. Panda, “Accelerating CPU-based Distributed DNN Training on Modern HPC Clusters using BlueField-2 DPUs”, Hot Interconnect ‘21

X-ScaleAI-DPU Package

X-ScaleSolutions

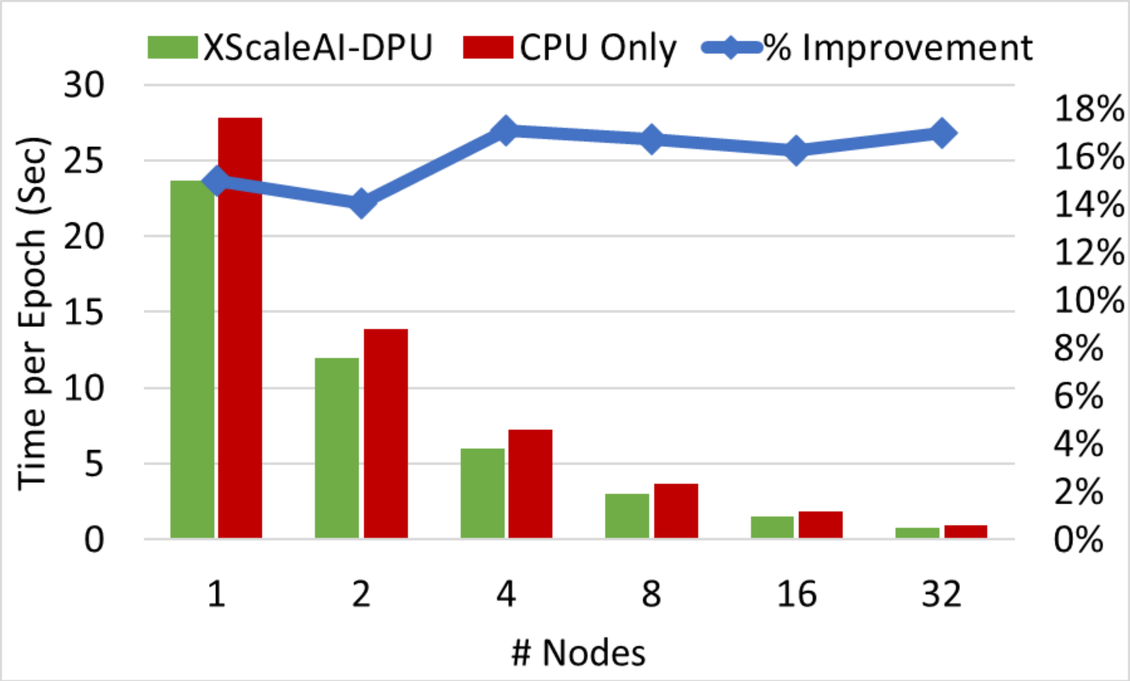
- Accelerating CPU-based DNN training with DPU support
- Based on MVAPICH2 2.3.7 with Horovod 0.25.0
- Supports all features available with the MVAPICH2 2.3.7 release (http://mvapich.cse.ohio-state.edu)
- Supports PyTorch framework for Deep Learning

Available from X-ScaleSolutions, please send a note to contactus@x-scalesolutions.com to get a trial license.

Training of ResNet-20v1 model on the CIFAR10 dataset

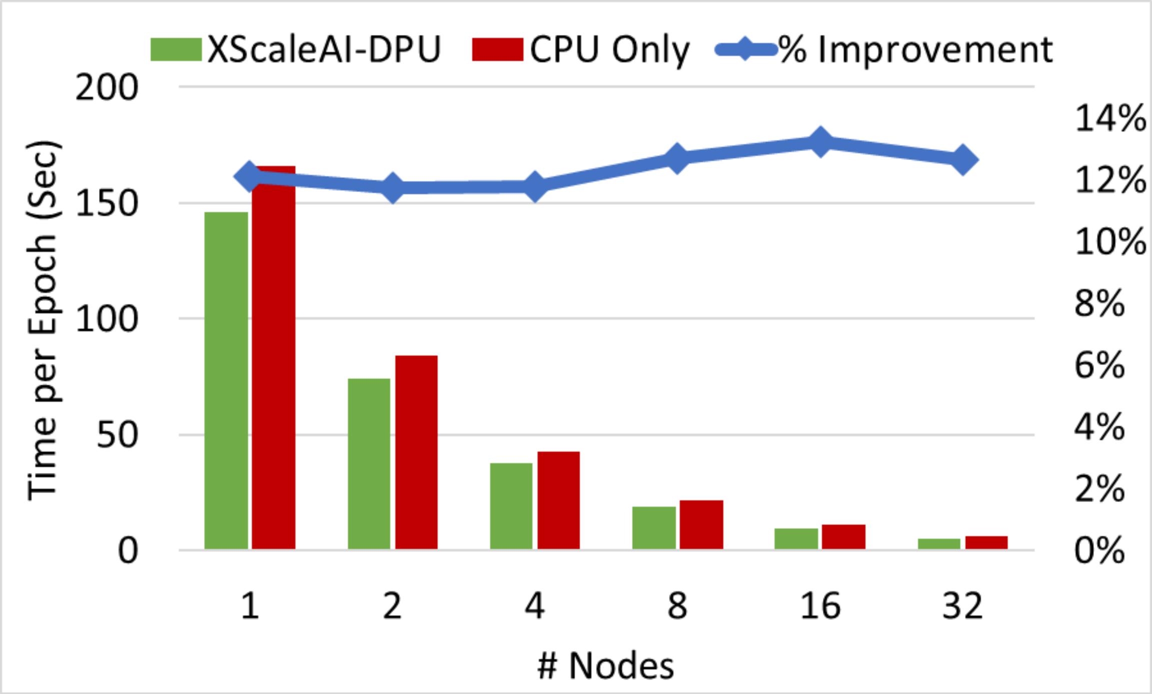
System Configuration

- Two Intel(R) Xeon(R) 16-core CPUs (32 total) E5-2697A V4 @ 2.60 GHz
- NVIDIA BlueField-2 SoC, HDR100 100Gb/s InfiniBand/VPI adapters
- Memory: 256GB DDR4 2400MHz RDIMMs per node
- 1TB 7.2K RPM SSD 2.5" hard drive per node
- NVIDIA ConnectX-6 HDR/HDR100 200/100Gb/s InfiniBand/VPI adapters with Socket Direct



Performance improvement using X-ScaleAI-DPU over CPU-only training on the ResNet-20v1 model on the CIFAR10 dataset

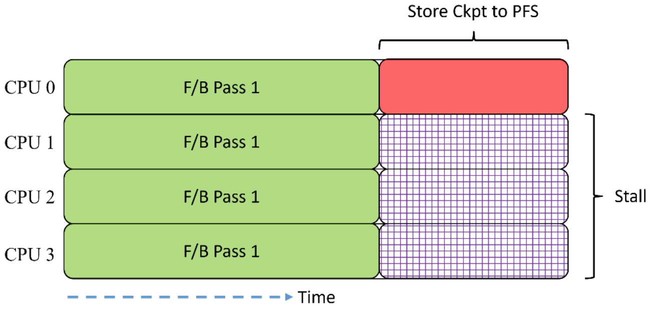
Training of the ShuffleNet model on the TinyImageNet Dataset



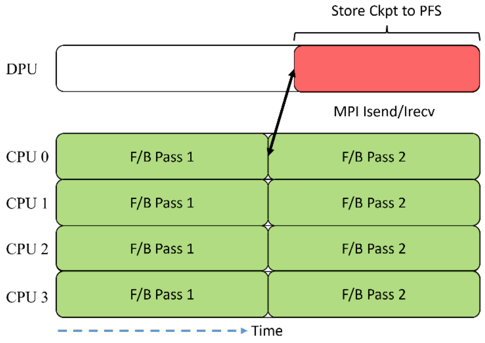
Performance improvement using X-ScaleAI-DPU over CPU-only training on the ShuffleNet model on the TinyImageNet dataset

X-ScaleAI DPU Checkpointing

- All DNN training runs must save snapshots of in-progress snapshots of the model parameters called a *checkpoint*
  - Unstable HPC/Cloud clusters require frequent checkpointing
  - The checkpoint cost scales with the number of model parameters
- Typically, the root rank saves the checkpoint while all other ranks stall (Called a *root ckpt*)
- By offloading checkpointing to the DPU, we can overlap checkpoint I/O with compute



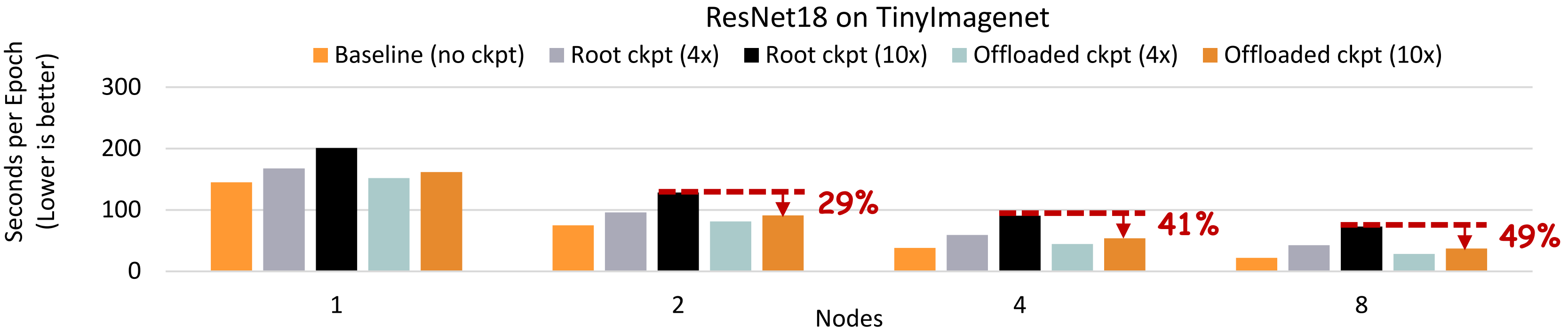
Default Root Checkpointing



Checkpointing with DPU offload

X-ScaleAI DPU Checkpointing

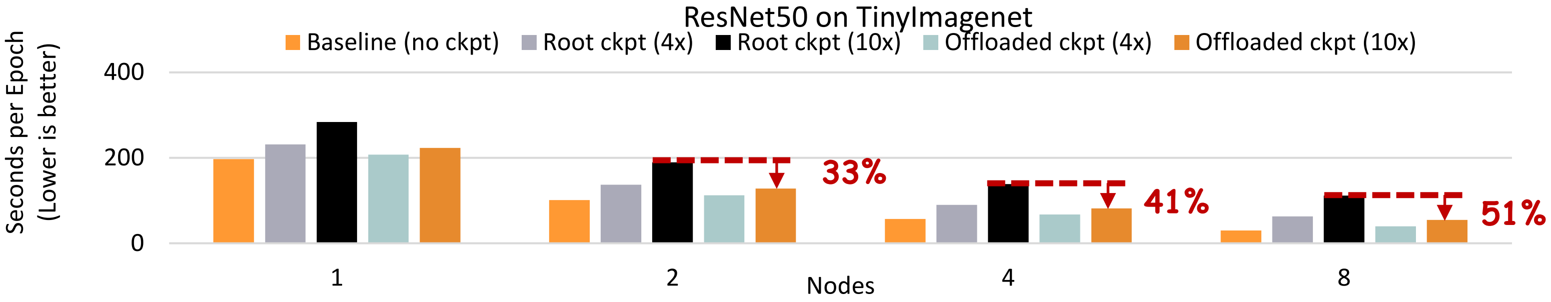
- We measure the time per epoch with ResNet18 on the TinyImageNet dataset
  - *Root* and *Offloaded* checkpoints refer to the default and DPU checkpointing schemes, respectively
  - TinyImageNet is a subset of ImageNet containing 100k images downsized to 64x64 pixels
  - We use *4x* and *10x* to refer to checkpointing 4 and 10 times within an epoch, respectively
  - We expect users to checkpoint less frequently (4x) on stable HPC/Cloud systems, and more frequently (10x) on unstable HPC/Cloud systems
- DPU-Offloaded checkpointing outperforms root checkpointing at all node scales
- Up to **49%** reduction in epoch time for ResNet18 on an unstable system where frequent checkpointing is required

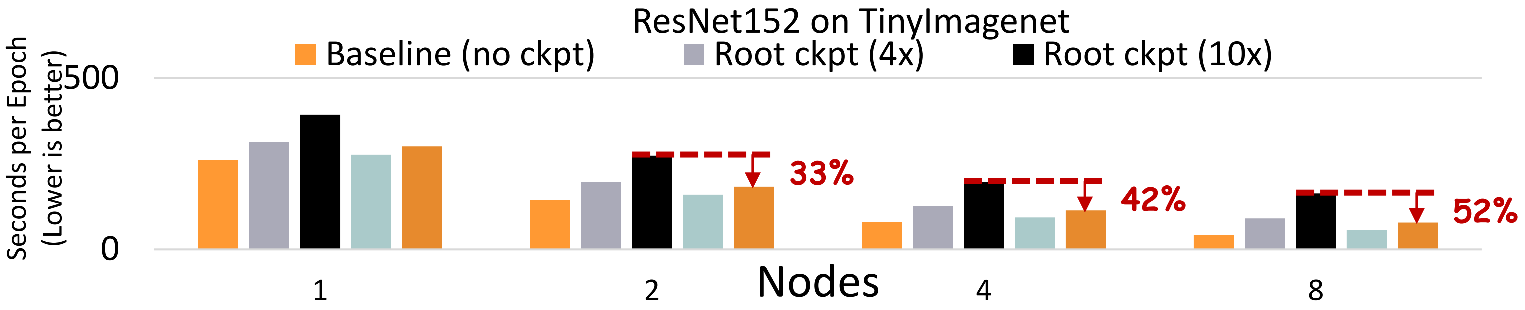


X-ScaleAI DPU Checkpointing

Will be available in upcoming X-ScaleAI-DPU Release

- Similarly for larger CNN models like ResNet50 and ResNet152, DPU-Offloaded checkpointing outperforms root checkpointing at all node scales
- Up to **51% and 52%** reduction in epoch time for ResNet50 and ResNet152, respectively on an unstable system where frequent checkpointing is required



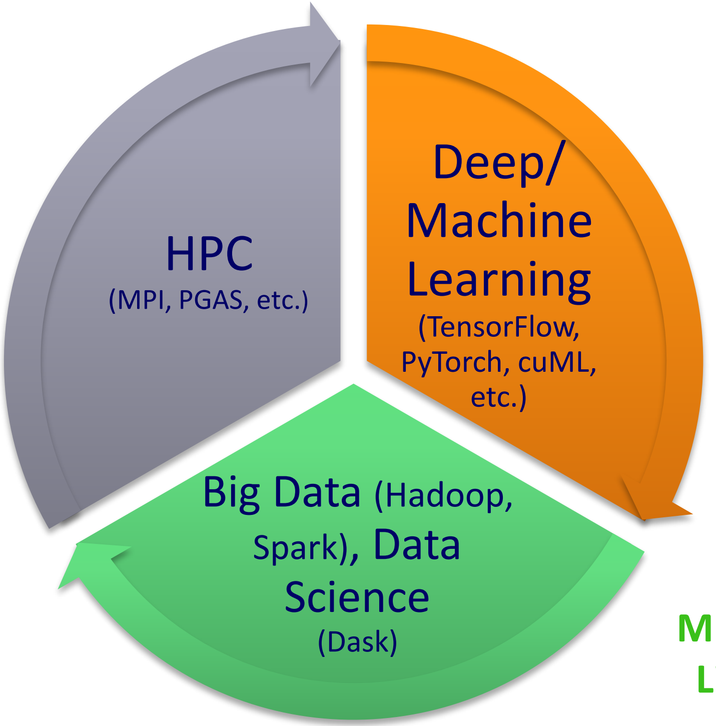


Presentation Overview

- MVAPICH MPI Library Project
  - High-Performance Support for various CPU, GPU, DPU, and Networking Technologies
- HiDL Project
  - High-Performance Deep Learning and Machine Learning
  - Accelerating Deep Learning with DPU
- **HiBD Project**
  - **Accelerating Spark with MPI**
  - **Accelerating Data Science Applications with Dask**
- Commercial Support and Value-Added Products
- Conclusions

Converged Middleware for HPC, AI, Big Data and Data Science

MVAPICH2 & MVAPICH2-DPU Libraries


HPC
(MPI, PGAS, etc.)
Deep/ Machine Learning
(TensorFlow, PyTorch, cuML, etc.)
Big Data (Hadoop, Spark), Data Science
(Dask)

High-Performance DL/ML Libraries with MVAPICH

X-Scale-AI-DPU

MPI4Spark and MPI4Dask Libraries with MVAPICH

The High-Performance Big Data (HiBD) Project

- Since 2013
- 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)
- **MPI4Dask**
- **MPI4Spark**
- OSU HiBD-Benchmarks (OHB)
  - HDFS, Memcached, HBase, and Spark Micro-benchmarks
- **http://hibd.cse.ohio-state.edu**
- **Users Base: 355 organizations from 39 countries**
- **More than 47,350 downloads from the project site**

Available for InfiniBand and RoCE

Also run on Ethernet

Available for x86 and OpenPOWER

Support for Singularity and Docker

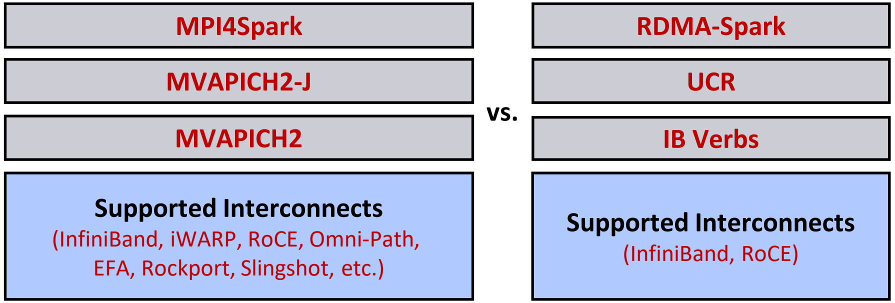
HiBD
High-Performance
Big Data

Network Based Computing
Laboratory

THE OHIO STATE
UNIVERSITY

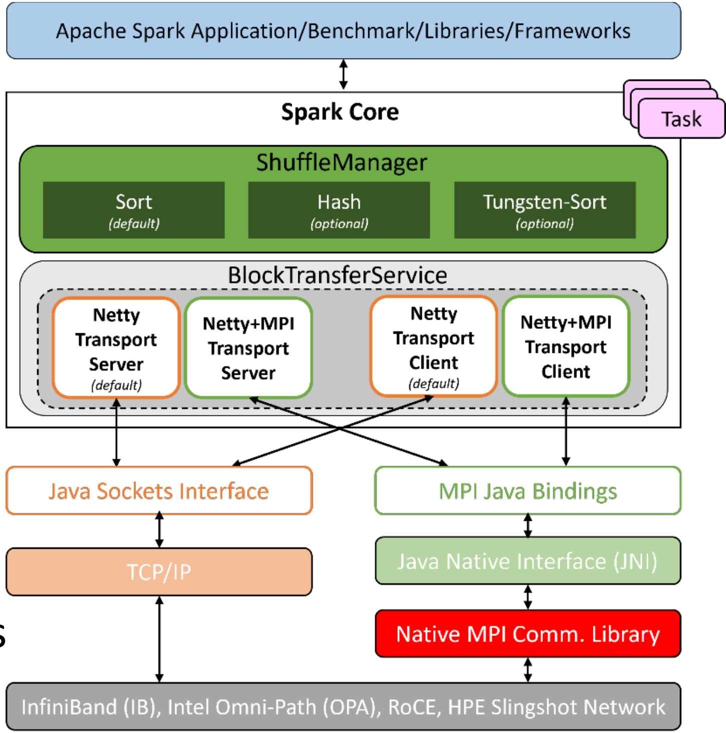
MPI4Spark Interconnect Support

- The current approach is different from its predecessor design, RDMA-Spark (http://hibd.cse.ohio-state.edu)
  - RDMA-Spark supports only InfiniBand and RoCE
  - Requires new designs for new interconnect
- MPI4Spark supports multiple interconnects/systems through a common MPI library
  - Such as InfiniBand (IB), Intel Omni-Path (OPA), HPE Slingshot, RoCE, and others
  - No need to re-design the stack for a new interconnect as long as the MPI library supports it



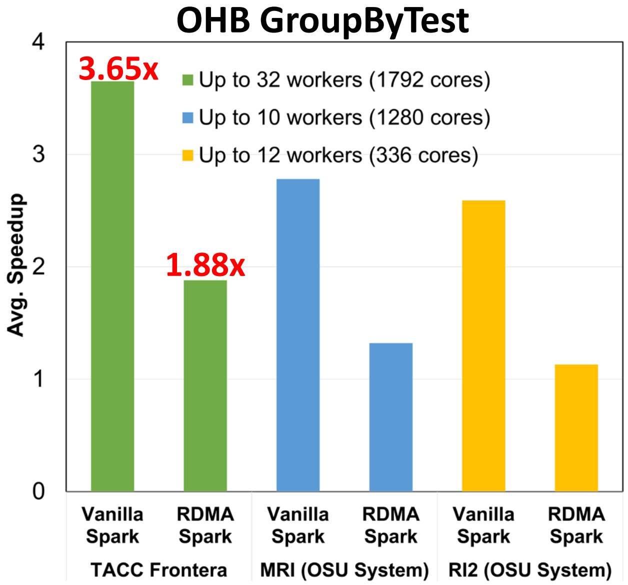
MPI4Spark: Using MVAPICH2 to Optimize Apache Spark

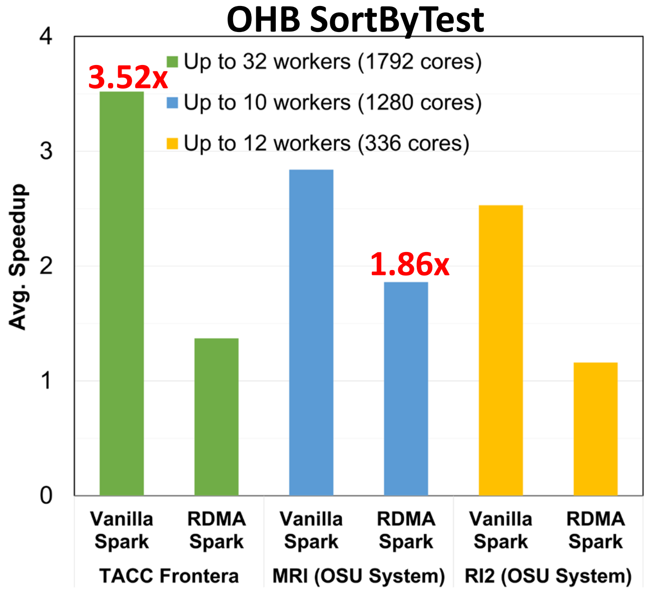
- The main motivation of this work is to utilize the communication functionality provided by MVAPICH2 in the Apache Spark framework
- MPI4Spark relies on Java bindings of the MVAPICH2 library
- Spark's default ShuffleManager relies on Netty for communication:
  - Netty is a Java New I/O (NIO) client/server framework for event-based networking applications
  - The key idea is to utilize MPI-based point-to-point communication inside Netty


Apache Spark Application/Benchmark/Libraries/Frameworks
Spark Core
Task
ShuffleManager
Sort (default)
Hash (optional)
Tungsten-Sort (optional)
BlockTransferService
Netty Transport Server (default)
Netty+MPI Transport Server
Netty Transport Client (default)
Netty+MPI Transport Client
Java Sockets Interface
MPI Java Bindings
TCP/IP
Java Native Interface (JNI)
Native MPI Comm. Library
InfiniBand (IB), Intel Omni-Path (OPA), RoCE, HPE Slingshot Network

MPI4Spark: Relative Speedups to Vanilla Spark and RDMA-Spark on Four HPC Systems

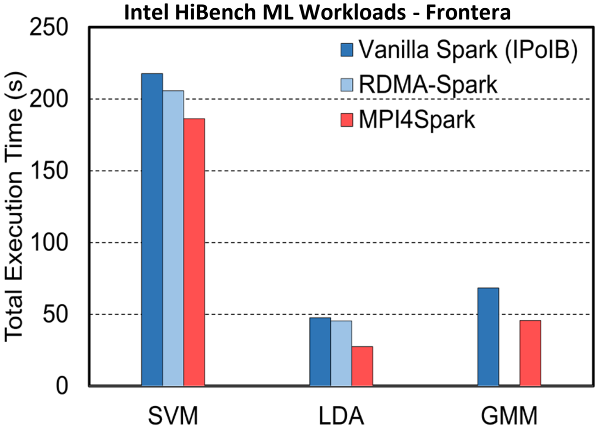
System Name	Nodes Used	Processor	Cores Used	Sockets	Cores/socket	RAM	Interconnect
TACC Frontera	34	Xeon Platinum	1792	2	28	192 GB	HDR (100G)
RI2 (OSU System)	14	Xeon Broadwell	336	2	14	128 GB	EDR (100G)
MRI (OSU System)	12	AMD EPYC 7713	1280	2	64	264 GB	200 Gb/sec (4X HDR)
TACC Stamepede2	10	Xeon Platinum	384	2	28	192 GB	Intel Omni-Path (100G)





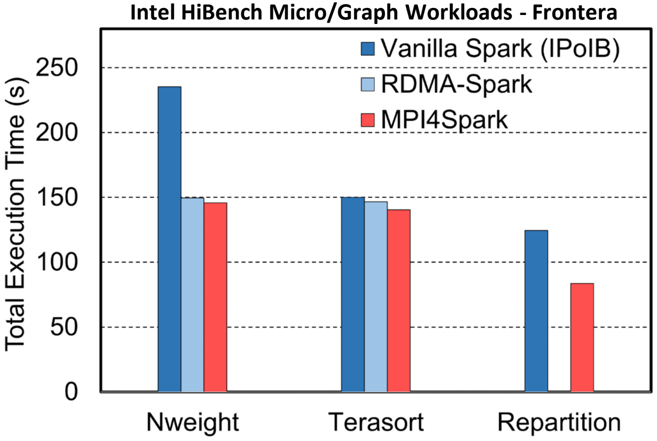
Performance Evaluation with Intel HiBench Workloads

1.4x on average than RDMA-Spark



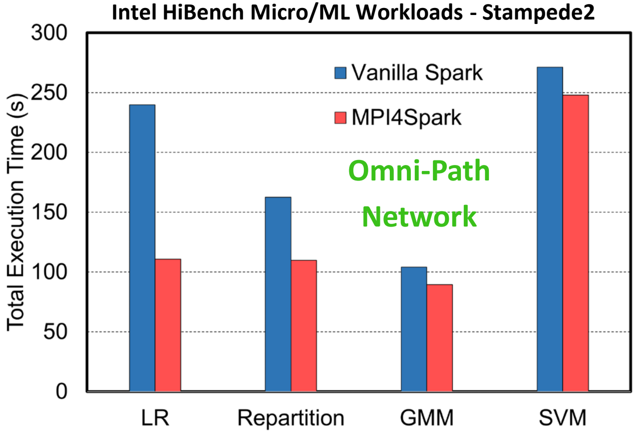
Intel HiBench ML Workloads - Frontera

1.4x on average than Vanilla Spark



Intel HiBench Micro/Graph Workloads - Frontera

1.5x on average than Vanilla Spark

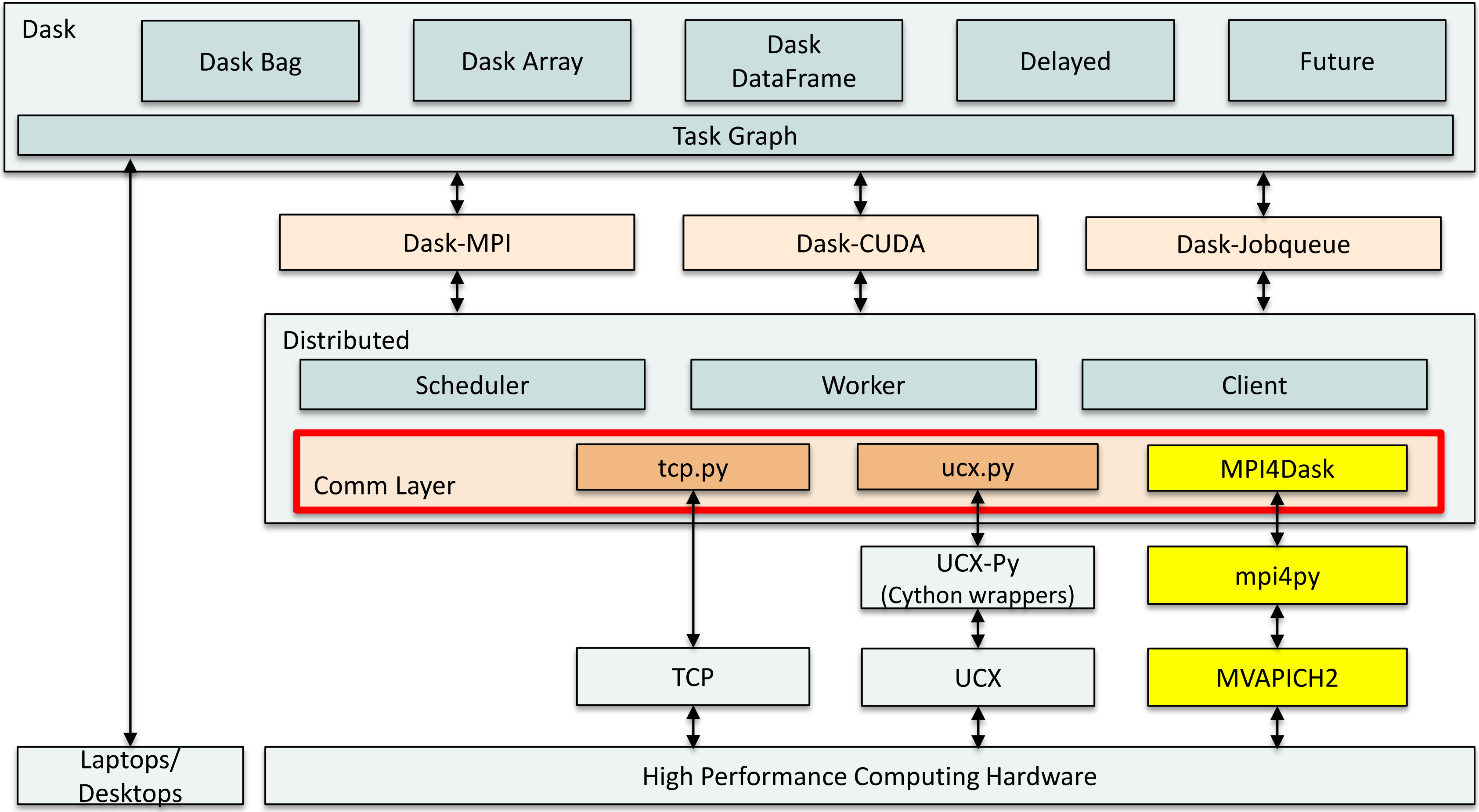


Intel HiBench Micro/ML Workloads - Stampede2

- This evaluation was done on the TACC Frontera (IB) and the TACC Stampede2 (OPA) Systems
- This illustrates the portability of MPI4Spark on different interconnects
- We see a speed-up for the LR machine learning workload on Stampede2 of about **2.2x**
- Speed-ups for the LDA machine learning workload on Frontera are **1.7x** for both IPoIB and RDMA

K. Al Attar, A. Shafi, M. Abduljabbar, H. Subramoni, D. Panda, Spark Meets MPI: Towards High-Performance Communication Framework for Spark using MPI, IEEE Cluster '22, Sep 2022.

MPI4Dask in the Dask Architecture



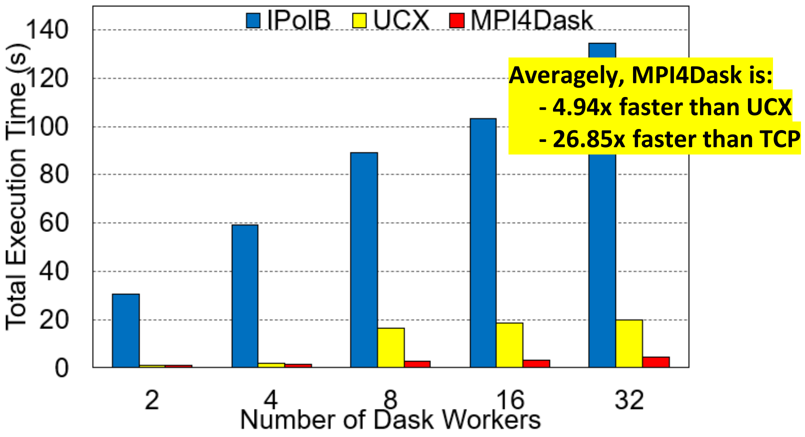
cuDF Merge Benchmark on the Cambridge Wilkes-3 System

- GPU-based Operation: $ddf1.merge(ddf2)$, using persist
  - Merge two GPU data frames, each with length of 32*1e8
  - Compute() will gather the data from all worker nodes to the client node, and make a copy on the host memory.
  - Persist() will leave the data on its current nodes without any gathering

Wilke3 GPU System:

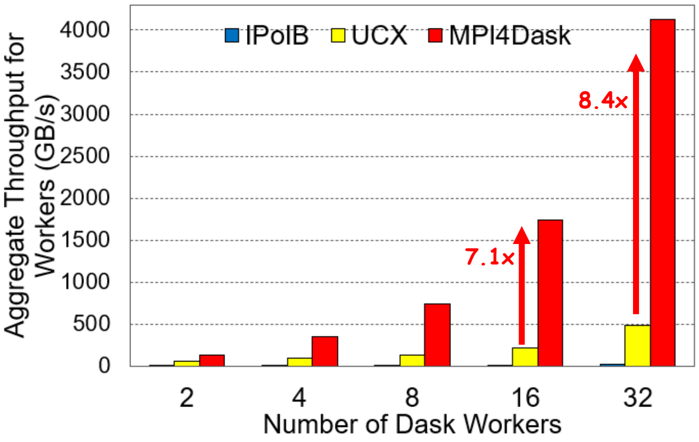
- **80 nodes**
- **2x AMD EPYC 7763 64-core Processors**
- **1000 GiB RAM**
- **Dual-rail Mellanox HDR200 IB**
- **4x NVIDIA A100 SXM4 80 GB**

Execution Time



Averagely, MPI4Dask is:
- 4.94x faster than UCX
- 26.85x faster than TCP

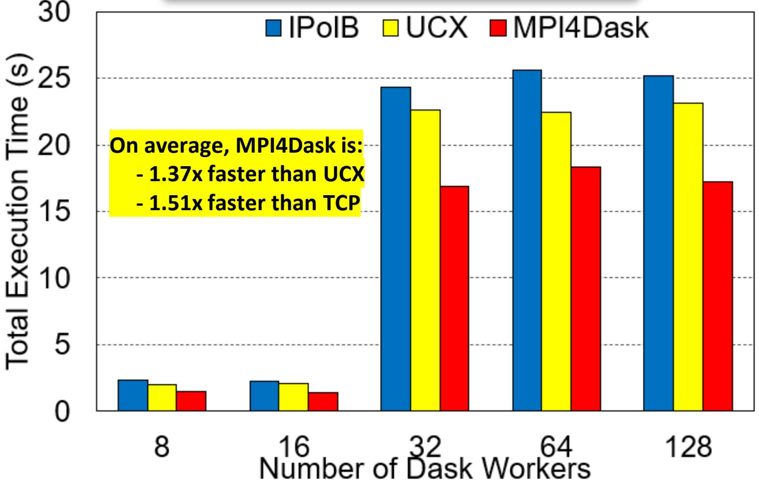
Aggregated Throughput



MPI4Dask 0.3, Dask 2022.8.1, Distributed, 2022.8.1, MVAPICH2-GDR 2.3.7, UCX v1.13.1, UCX-py 0.27.00

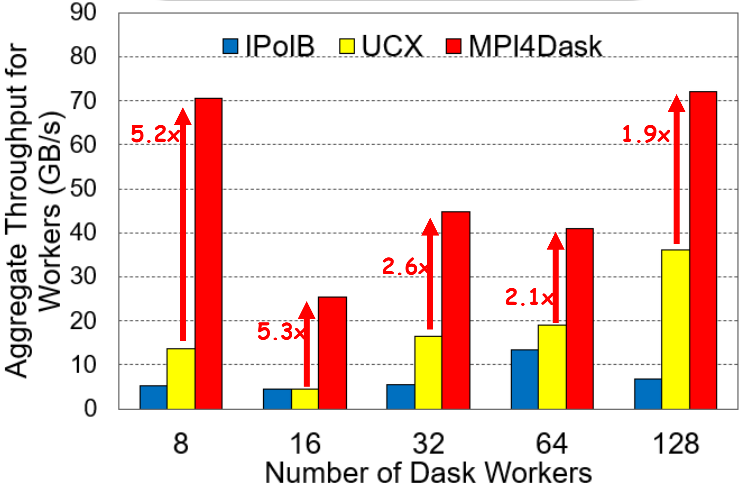
NumPy Array Slicing Benchmark on TACC Frontera CPU System

1.26x better on average



On average, MPI4Dask is:
- 1.37x faster than UCX
- 1.51x faster than TCP

3.17x better on average



From 32 workers, we increase array size by 16 times

A. Shafi , J. Hashmi , H. Subramoni , and D. K. Panda, Efficient MPI-based Communication for GPU-Accelerated Dask Applications, CCGrid '21
https://arxiv.org/abs/2101.08878

MPI4Dask 0.3 release

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

Presentation Overview

- MVAPICH MPI Library Project
  - High-Performance Support for various CPU, GPU, DPU, and Networking Technologies
- HiDL Project
  - High-Performance Deep Learning and Machine Learning
  - Accelerating Deep Learning with DPU
- HiBD Project
  - Accelerating Spark with MPI
  - Accelerating Data Science Applications with Dask
- **Commercial Support and Value-Added Products**
- Conclusions

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 being provided to National Laboratories and International HPC centers

X-ScaleSolutions

Value-Added Products with Support

- Multiple value-added products with commercial support
  - X-ScaleHPC
  - X-ScaleAI
  - MVAPICH2-DPU
  - X-ScaleAI-DPU
- These products have additional
  - Features
  - Performance Optimizations
  - Profiling and Introspection Support
- Send an e-mail to contactus@x-scalesolutions.com for free trial!!

X-ScaleSolutions

Presentation Overview

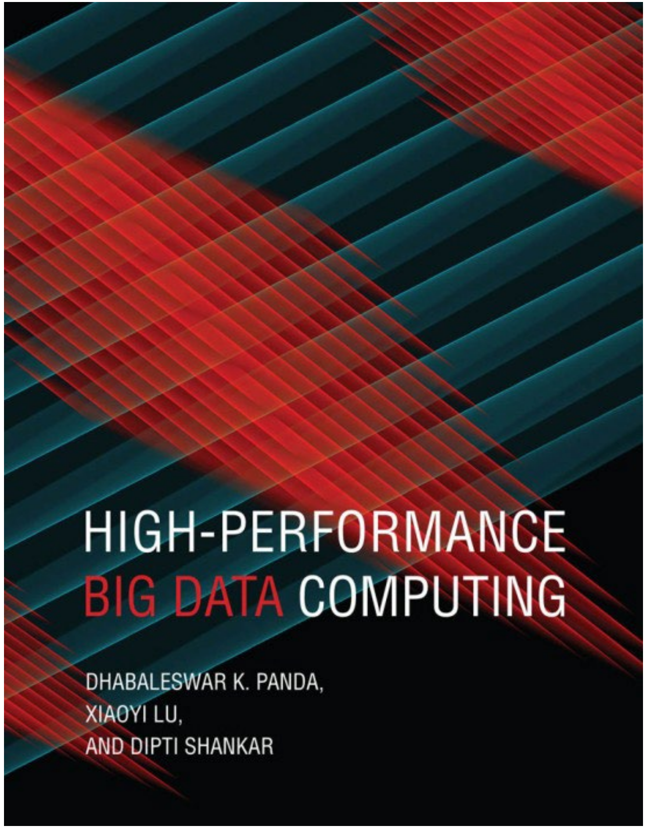
- MVAPICH MPI Library Project
  - High-Performance Support for various CPU, GPU, DPU, and Networking Technologies
- HiDL Project
  - High-Performance Deep Learning and Machine Learning
  - Accelerating Deep Learning with DPU
- HiBD Project
  - Accelerating Spark with MPI
  - Accelerating Data Science Applications with Dask
- Commercial Support and Value-Added Products
- **Conclusions**

Concluding Remarks

- Upcoming Exascale systems and Cloud need to be designed with a holistic view of HPC, Big Data, Deep/Machine Learning, Data Science, and computing continuum
- Presented an overview of opportunities and challenges in designing scalable and high-performance middleware for such systems
- Presented a set of solutions which enable these communities to take advantage of current and next-generation systems with latest technologies
- Next-generation Zetascale systems will need continuous innovations in designing converged software architectures …..

A New Book on High-Performance Big Data Computing

- By MIT Press
- Released in August 2022
- An in-depth overview of an emerging field that brings together high-performance computing, big data processing, and deep learning.
- https://mitpress.mit.edu/books/high-performance-big-data-computing



HIGH-PERFORMANCE
BIG DATA COMPUTING

DHABALESWAR K. PANDA,
XIAOYI LU,
AND DIPTI SHANKAR

Funding Acknowledgments

Funding Support by

Office of Science U.S. DEPARTMENT OF ENERGY
NSF
Mellanox TECHNOLOGIES
Microsoft
QLOGIC
arm
CRAY THE SUPERCOMPUTER COMPANY
BROADCOM
intel
LINUX NETWORX
NVIDIA
AMD
NetApp
Sun microsystems
CISCO
ORACLE
ROCKPORT
ETRI
CORNELIS NETWORKS

Equipment Support by

Mellanox TECHNOLOGIES
intel
arm
Sun microsystems
APPRO
NVIDIA
AMD
QLOGIC
advanced clustering technologies, inc.
Microway Technology you can count on
ROCKPORT
BROADCOM

Acknowledgments to all the Heroes (Past/Current Students and Staffs)

Current Students (Graduate)

- N. Alnaasan (Ph.D.)
- Q. Anthony (Ph.D.)
- C.-C. Chun (Ph.D.)
- N. Contini (Ph.D.)
- K. S. Khorassani (Ph.D.)
- P. Kousha (Ph.D.)
- B. Michalowicz (Ph.D.)
- B. Ramesh (Ph.D.)
- K. K. Suresh (Ph.D.)
- A. H. Tu (Ph.D.)
- S. Xu (Ph.D.)
- Q. Zhou (Ph.D.)
- K. Al Attar (M.S.)
- L. Xu (Ph.D.)
- G. Kuncham (Ph.D.)
- R. Vaidya (Ph.D.)
- J. Yao (Ph.D.)
- M. Han (M.S.)
- A. Guptha (M.S.)

Current Research Scientists

- M. Abduljabbar
- A. Shafi

Current Students (Undergrads)

- T. Chen

Current Faculty

- H. Subramoni

Current Software Engineers

- B. Seeds
- N. Pavuk
- N. Shineman
- M. Lieber

Current Research Specialist

- R. Motlagh

Past Students

- A. Awan (Ph.D.)
- A. Augustine (M.S.)
- P. Balaji (Ph.D.)
- M. Bayatpour (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.)
- C.-H. Chu (Ph.D.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- J. Hashmi (Ph.D.)
- W. Huang (Ph.D.)
- A. Jain (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (Ph.D.)
- M. Kedia (M.S.)
- 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.)
- K. Raj (M.S.)
- R. Rajachandrasekar (Ph.D.)
- D. Shankar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- N. Sarkauskas (B.S. and M.S)
- N. Senthil Kumar (M.S.)
- A. Singh (Ph.D.)
- J. Sridhar (M.S.)
- S. Srivastava (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 Post-Docs

- D. Banerjee
- X. Besseron
- M. S. Ghazimeersaeed
- H.-W. Jin
- J. Lin
- M. Luo
- E. Mancini
- K. Manian
- S. Marcarelli
- A. Ruhela
- J. Vienne
- H. Wang

Past Research Scientists

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

Past Senior Research Associate

- J. Hashmi

Past Programmers

- A. Reifsteck
- D. Bureddy
- J. Perkins

Past Research Specialist

- M. Arnold
- J. Smith

Thank You!

panda@cse.ohio-state.edu

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/