MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiBD
High-Performance
Big Data

HiDL
High-Performance
Deep Learning

S61956- Accelerating HPC and AI Applications with Offloading to NVIDIA BlueField DPUs: Strategies and Benefits

NVIDIA GTC (March '24)

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

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 (GPUs from NVIDIA, AMD, and Intel)
- 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 HPC and AI systems while taking advantage of heterogeneous (CPU + GPU + DPU/IPU (xPU)) HPC and Cloud resources?

Presentation Outline

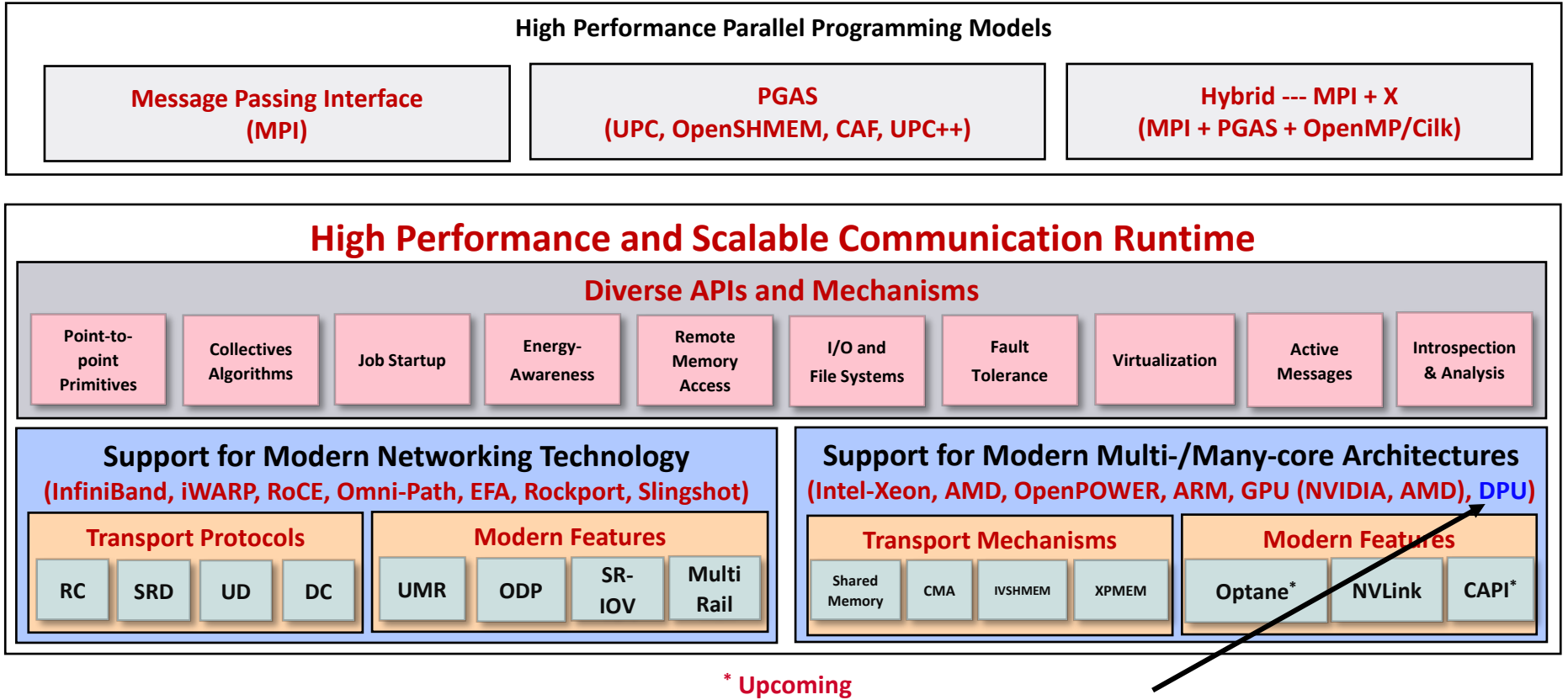
- **Overview of the MVAPICH Project**
- Offloading Strategies and Benefits:
  - Non-blocking Collectives (communication)
    - Ialltoall and P3DFFT
    - Ibcast and HPL
  - Non-blocking Point-to-point (communication)
    - Applications using 3D Stencils
  - Non-blocking Point-to-point and collective (communication and computation)
    - PETSc
  - Offloading DL training (computation)
- Conclusions

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 3,375 organizations in 91 countries**
- **More than 1.76 Million downloads from the OSU site directly**
- Empowering many TOP500 clusters (Nov '23 ranking)
  - 11th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 29th , 448, 448 cores (Frontera) at TACC
  - 46th, 288,288 cores (Lassen) at LLNL
  - 61st, 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 29th ranked TACC Frontera system
- **Empowering Top500 systems for more than 18 years**

Architecture of MVAPICH2 Software Family for HPC, DL/ML, and Data Science

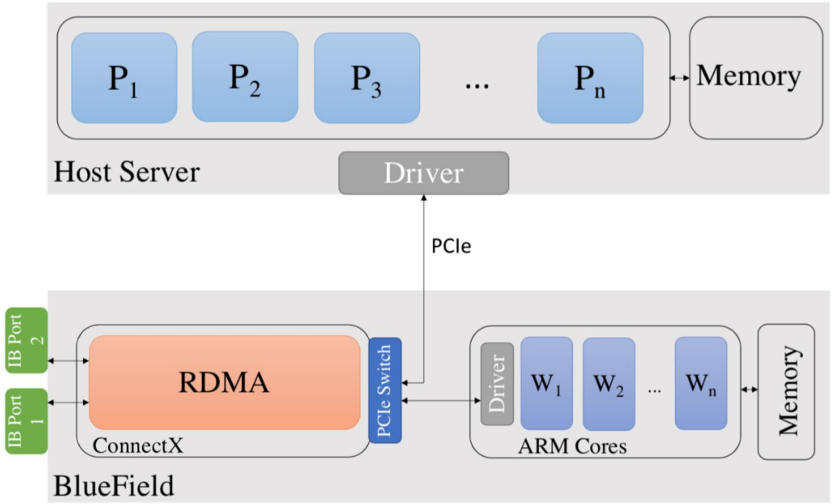


Presentation Outline

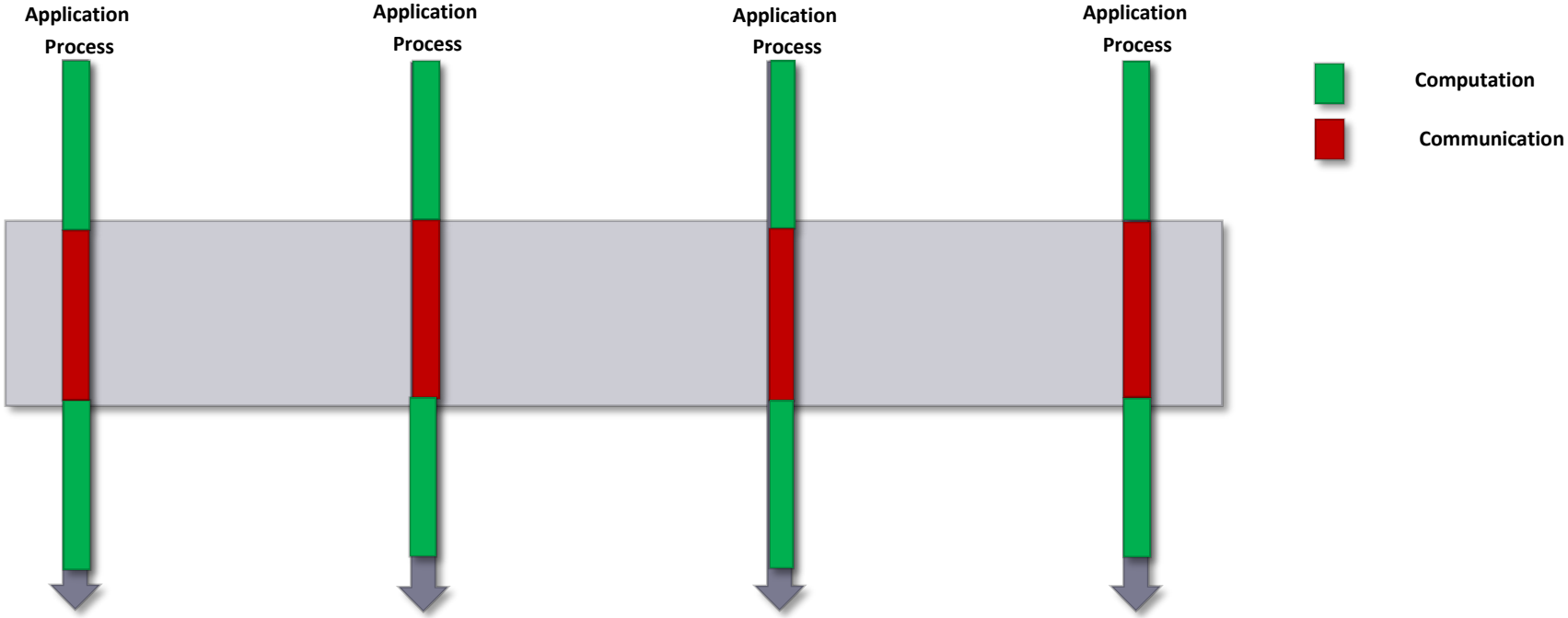
- Overview of the MVAPICH Project
- **Offloading Strategies and Benefits:**
  - **Non-blocking Collectives (communication)**
    - **Ialltoall and P3DFFT**
    - **Ibcast and HPL**
  - Non-blocking Point-to-point (communication)
    - Applications using 3D Stencils
  - Non-blocking Point-to-point and collective (communication and computation)
    - PETSc
  - Offloading DL training (computation)
- Conclusions

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

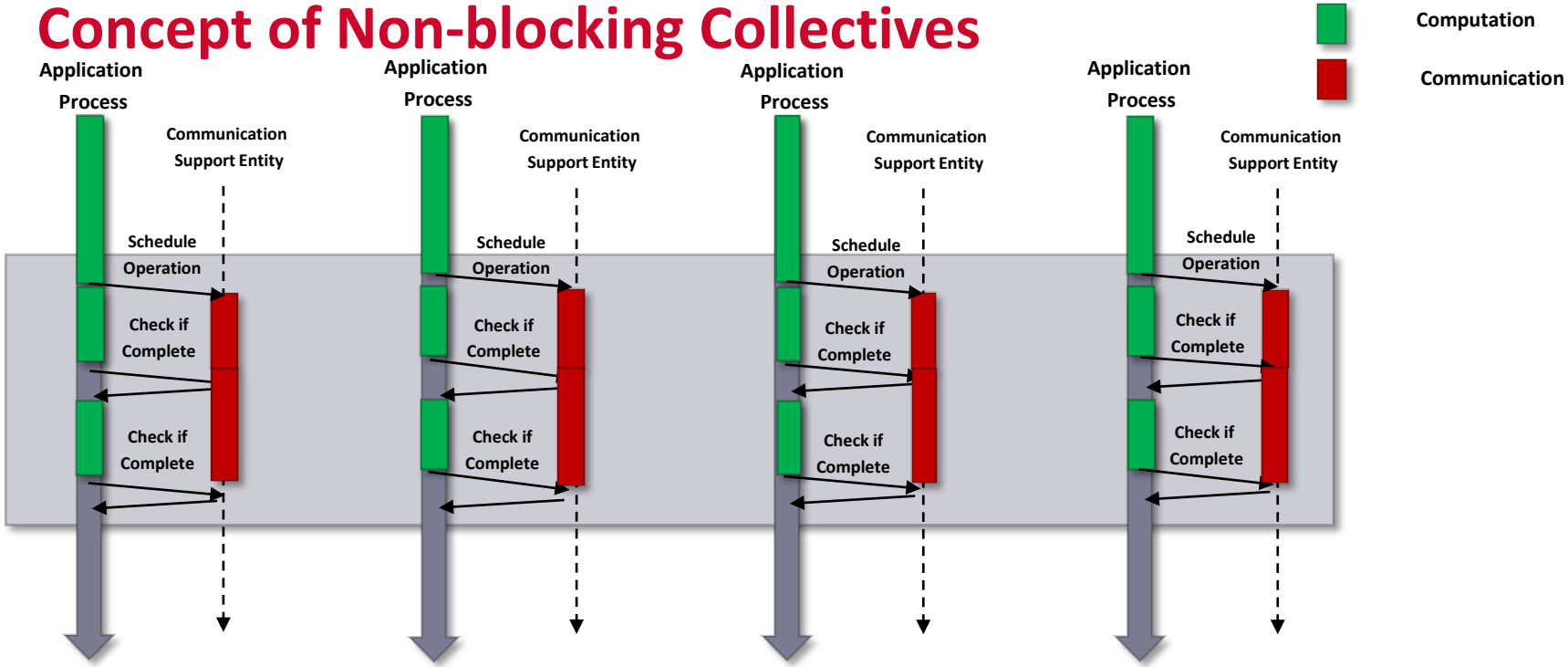


Problems with Blocking Collective Operations



- Communication time cannot be used for compute
  - No overlap of computation and communication
  - Inefficient

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*

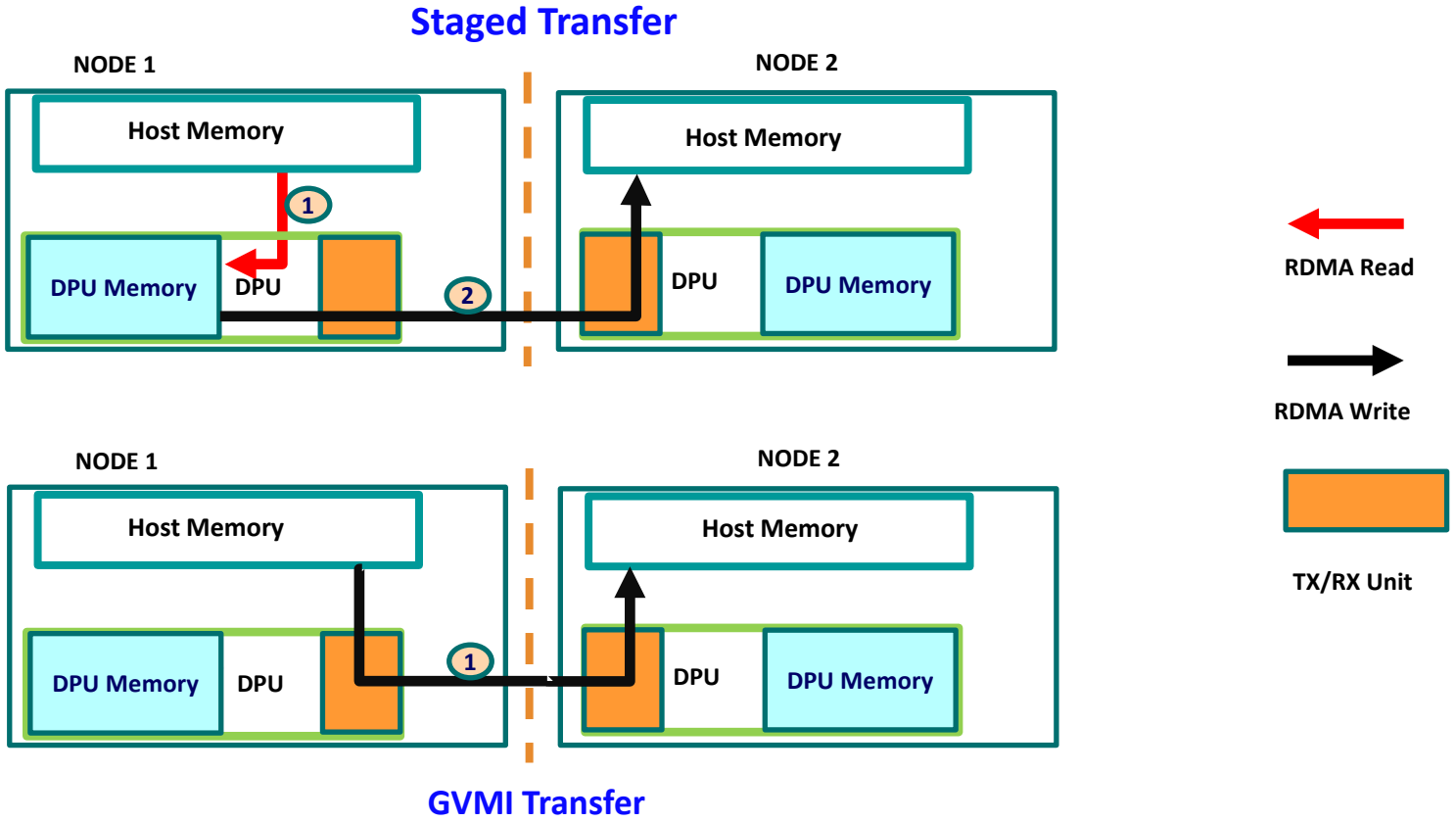
Major Opportunity and Benefits

- Overlap of Computation with Communication
- Reducing the overall application execution time

Four Major Challenges

- Host-DPU-Network communication mechanisms
- Non-blocking Collective Algorithm offload
- Load balancing across ARM cores to take care of the offloading tasks
- Re-designing applications using non-blocking collectives

Staging vs. GVMI (Guest Virtual Machine ID)



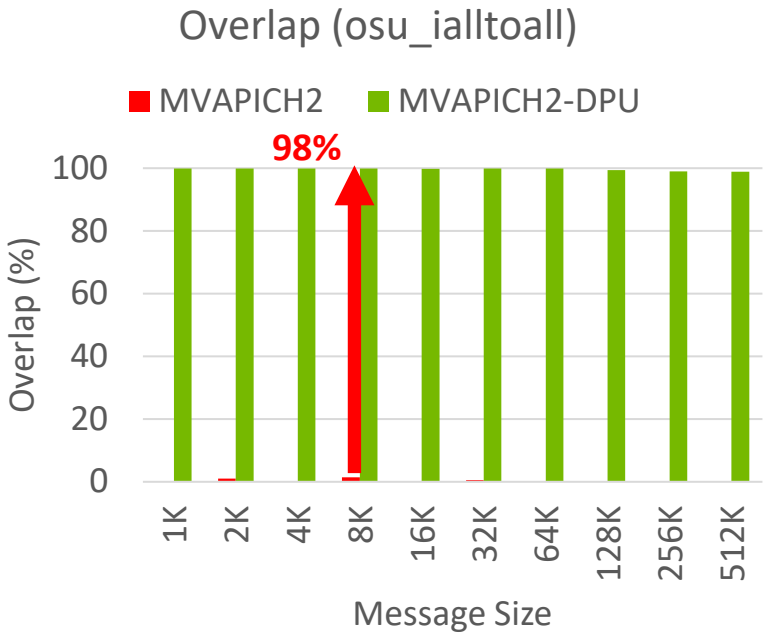
MVAPICH2-DPU Library Release

X-ScaleSolutions

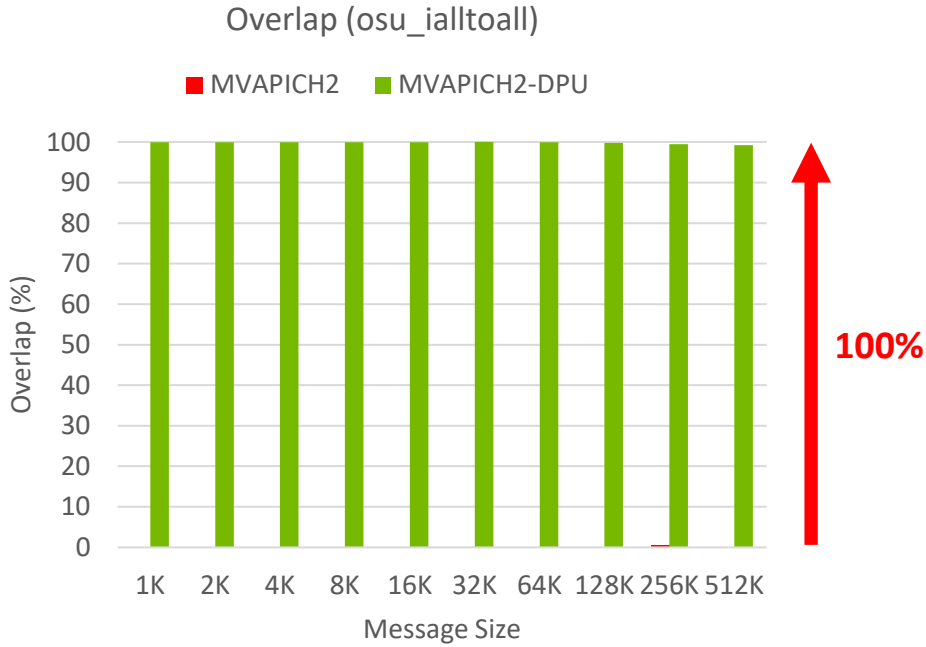
- Supports all features available with the MVAPICH2 release (http://mvapich.cse.ohio-state.edu)
- Novel framework to offload non-blocking collectives to DPU
- Offloads non-blocking Alltoall (MPI_Ialltoall) to DPU
- Offloads non-blocking Broadcast (MPI_Ibcast) to DPU

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

Overlap of Communication and Computation with osu_Ialltoall (BF-2, 32 nodes)



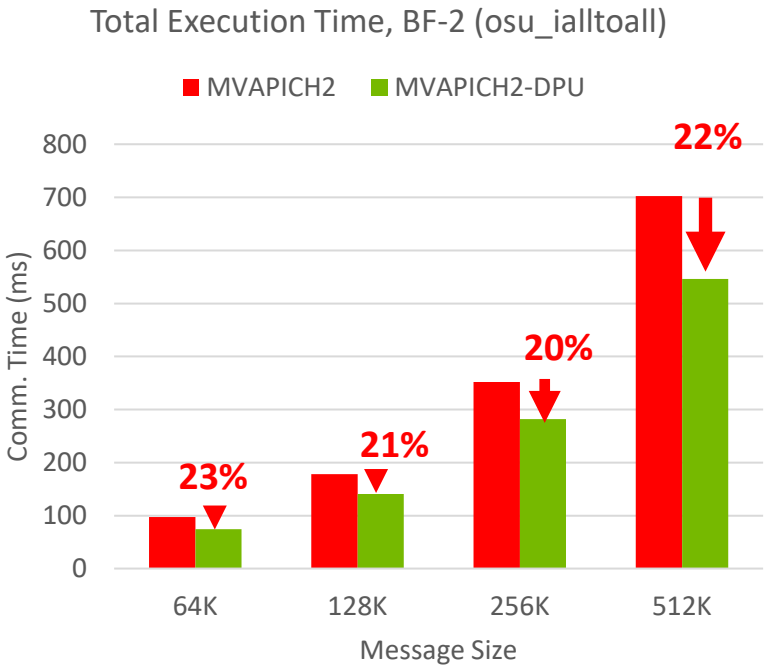
32 Nodes, 16 PPN



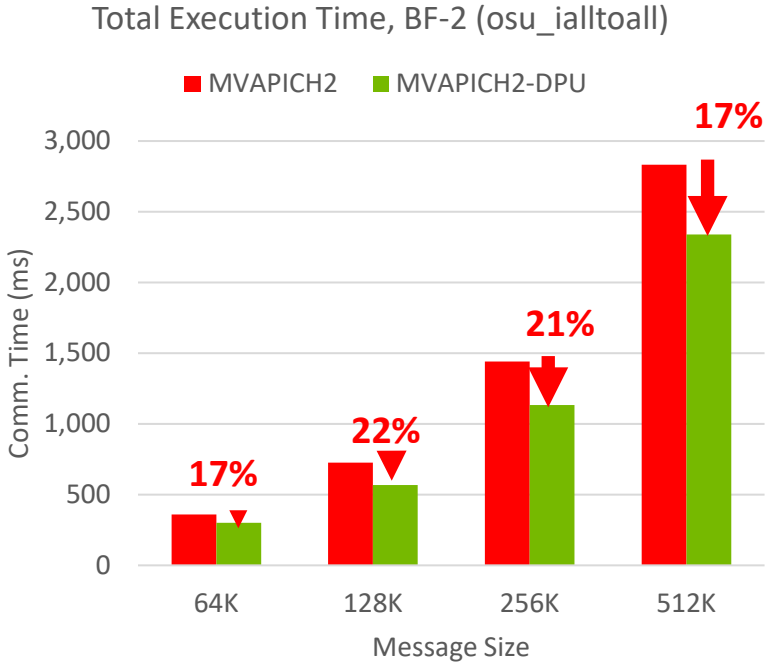
32 Nodes, 32 PPN

Delivers Peak Overlap

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



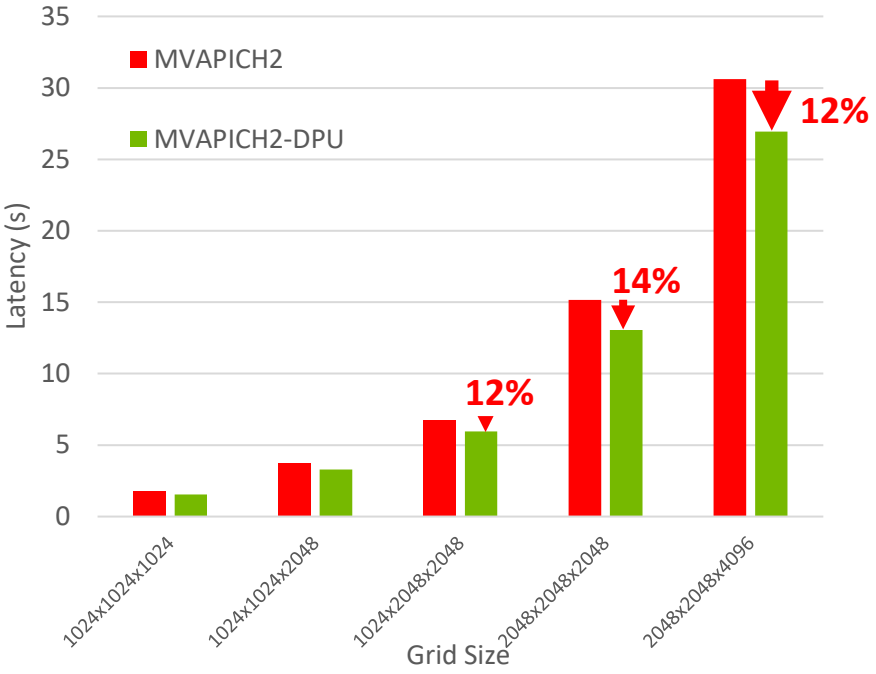
32 Nodes, 16 PPN



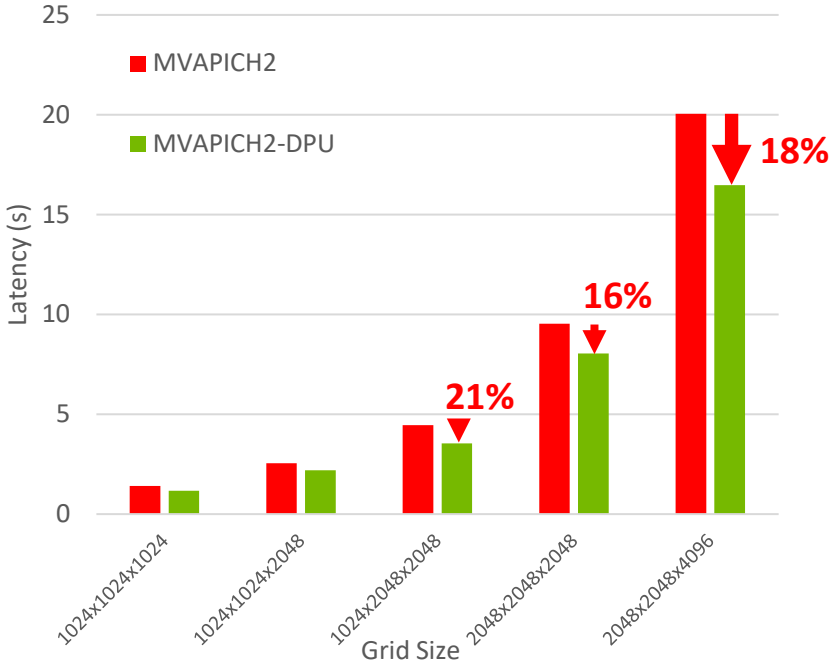
32 Nodes, 32 PPN

Benefits in Total execution time (Compute + Communication)

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



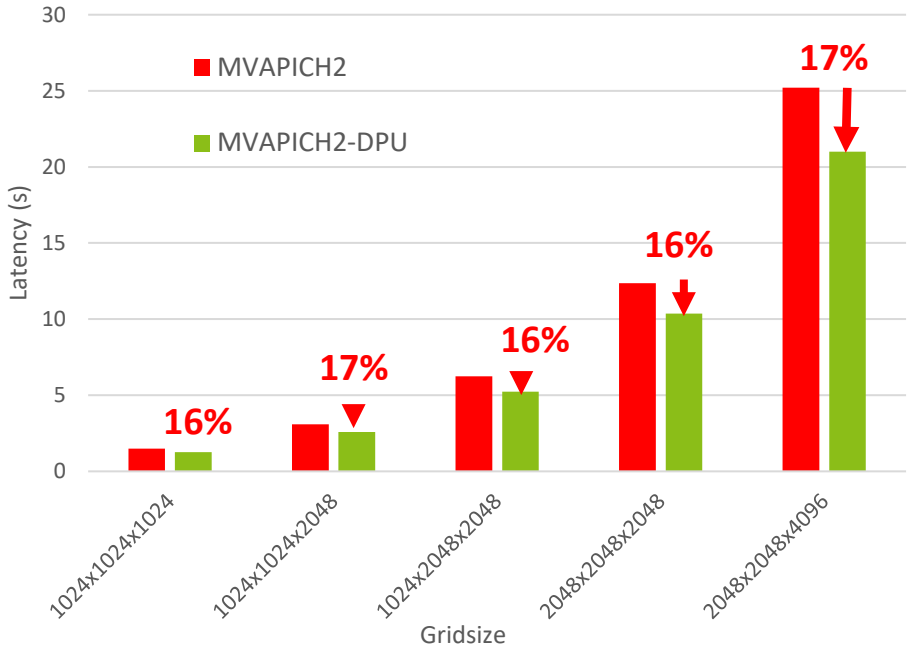
32 Nodes, 16 PPN



32 Nodes, 32 PPN

Benefits in application-level execution time

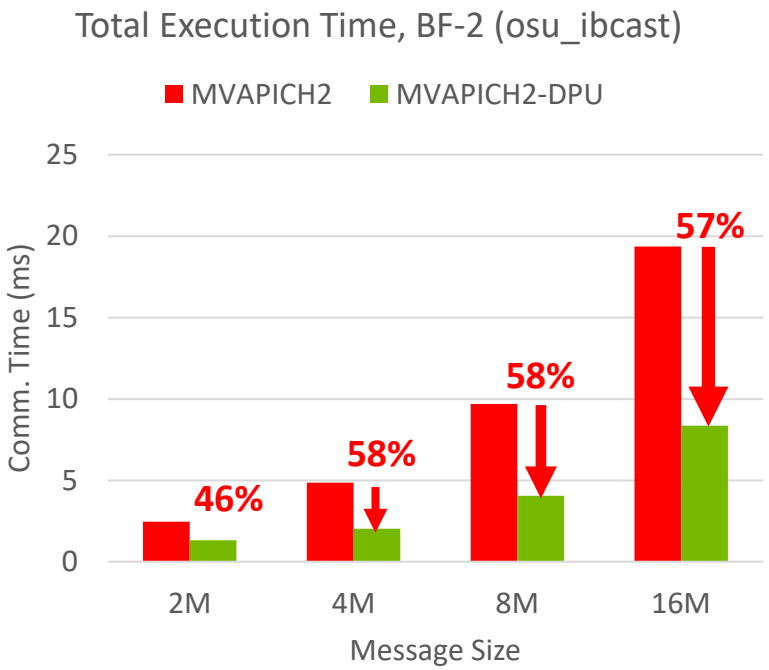
P3DFFT Application Execution Time (BF-3, 16 nodes)



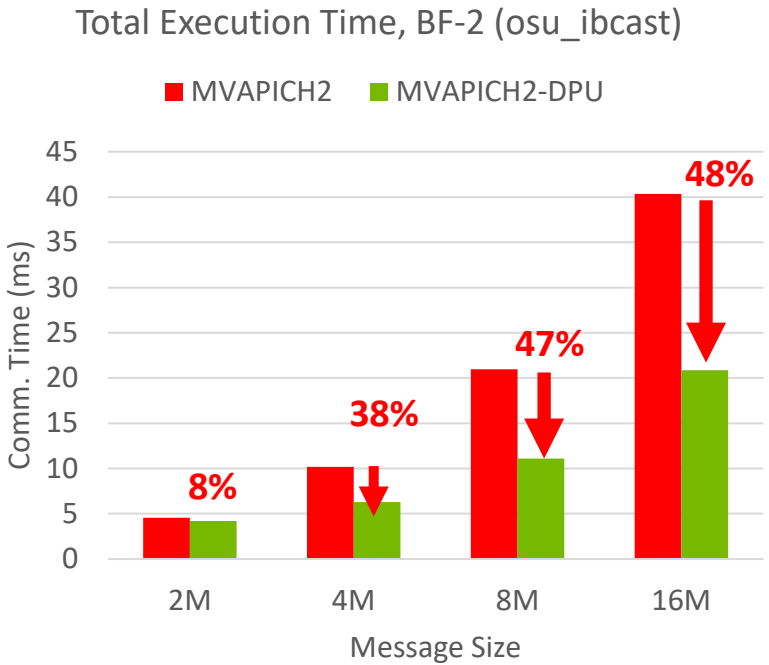
16 Nodes, 32 PPN

Benefits in application-level execution time

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



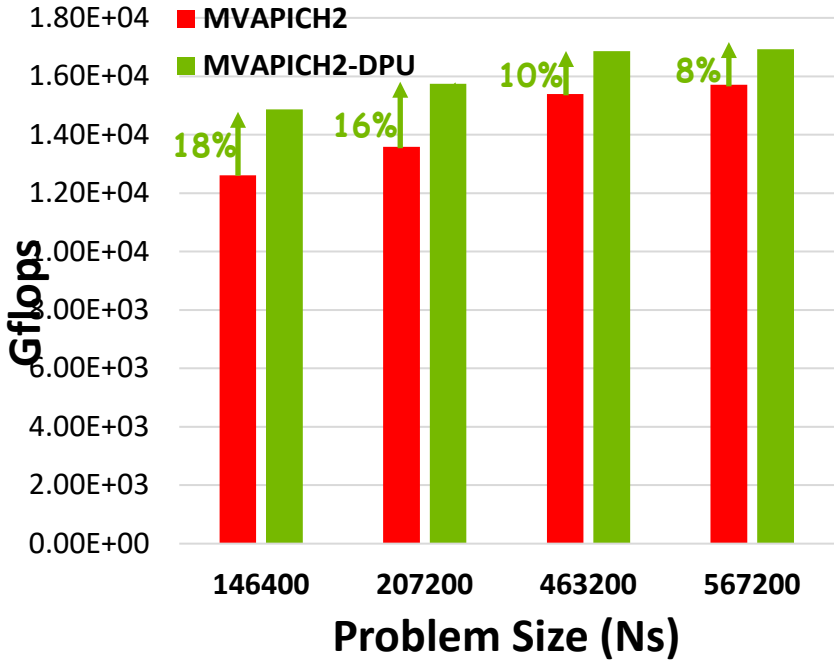
32 Nodes, 1 PPN



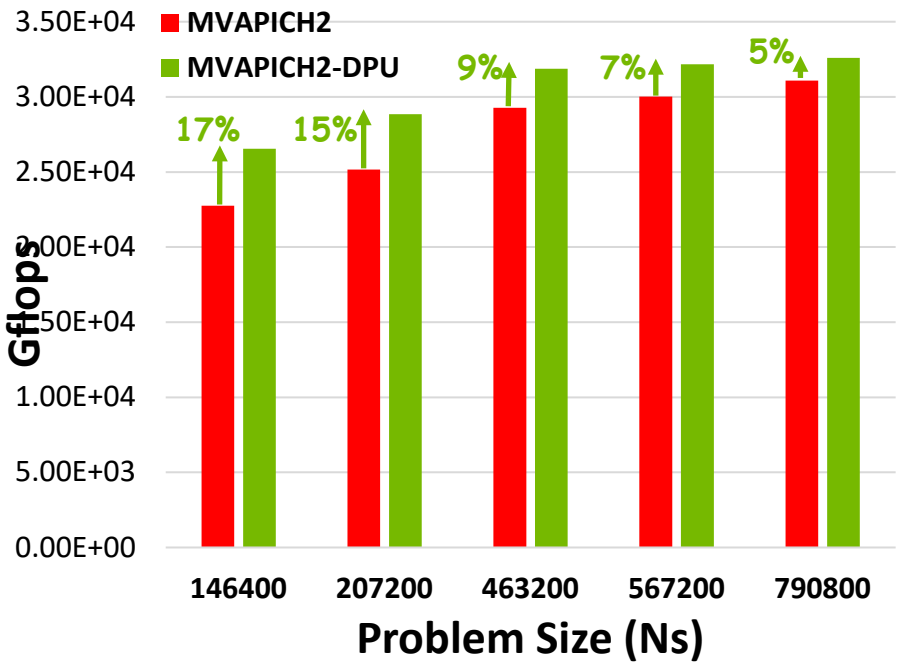
32 Nodes, 16 PPN

Benefits in Total execution time (Compute + Communication)

Accelerating HPL with MVAPICH2-DPU and XScaleHPL-DPU (BF-2)



16x32 process grid



31x32 process grid

Benefits in application-level execution time

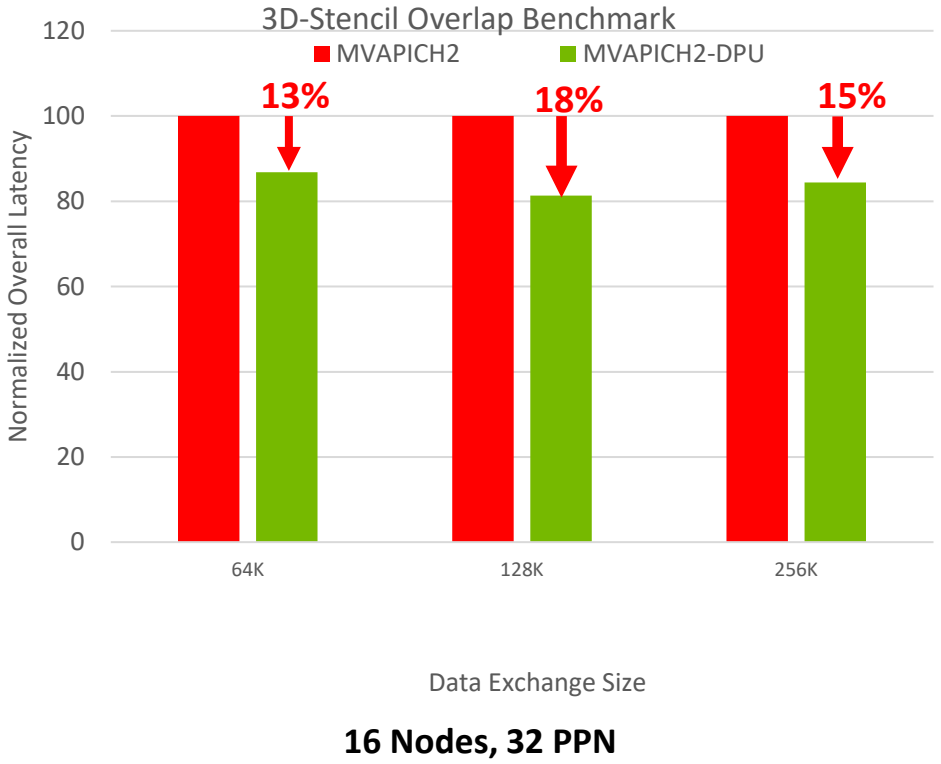
Presentation Outline

- Overview of the MVAPICH Project
- **Offloading Strategies and Benefits:**
  - Non-blocking Collectives (communication)
    - Ialltoall and P3DFFT
    - Ibcast and HPL
  - **Non-blocking Point-to-point (communication)**
    - **Applications using 3D Stencils**
  - **Non-blocking Point-to-point and collective (communication and computation)**
    - **PETSc**
  - Offloading DL training (computation)
- Conclusions

Offloading MPI Point-to-Point with 3D Stencil (BF-3)

- Use GVMI to Offload MPI_Isend/MPI_Irecv to the DPU

- 3D Stencil Overlap Benchmark :
  - Perform data exchange with 6 peers. (Similar to 7-point stencil)
  - Overlap computation with data-exchange
  - Up to 18% benefits



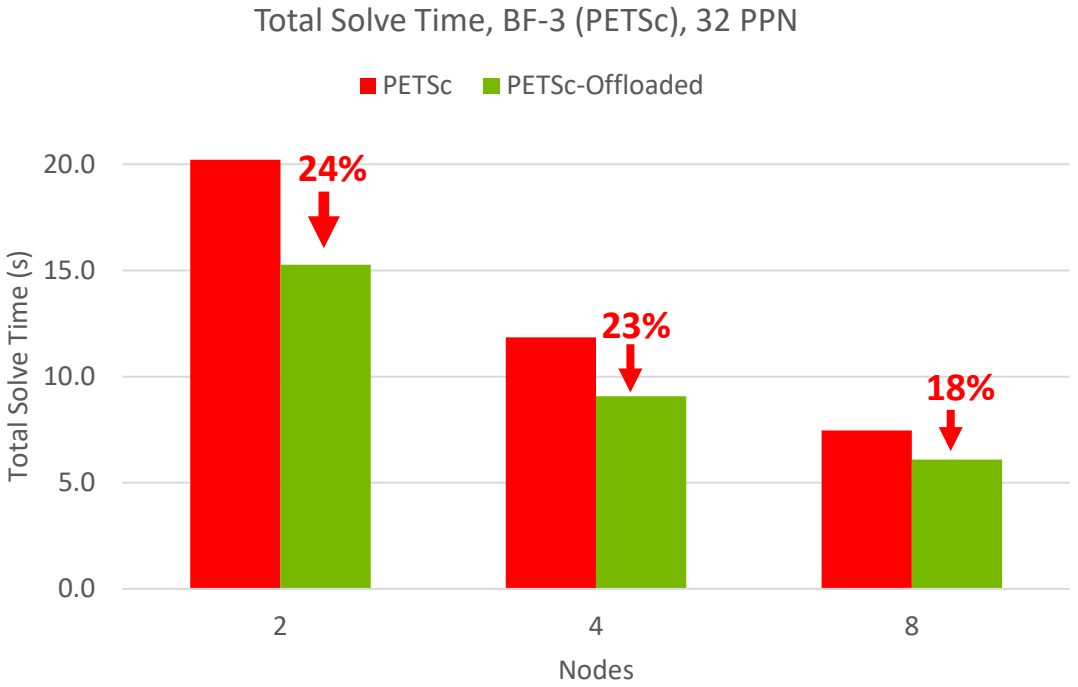
16 Nodes, 32 PPN

Offloading MPI Point-to-Point and Reduction with PETSc

- PETSc, the Portable, Extensible Toolkit for Scientific Computation
  - Includes a large suite of scalable parallel linear and nonlinear equation solvers, ODE integrators, and optimization algorithms for application codes written in C, C++, Fortran, and Python.
  - Includes support for managing parallel PDE discretization including parallel matrix and vector assembly routines
  - https://petsc.org/release/overview/
- Used in many different toolkits and libraries
  - Adflow, DAFoam, FreeFEM, MFEM, MOOSE, OpenFoam, etc.

Offloading MPI Point-to-Point and Reduction with PETSc (BF-3)

- PETSc:
  - Solves 3D Laplacian with 27-point finite difference stencil
- Modified Solver Algorithm to efficiently offload reduction (compute + communication) operations to the DPU
- Problem Size: 256X256X256
  - Strong Scaling Run
  - % Host-to-DPU data exchange cost increases for large node count
  - Up to 24% benefits



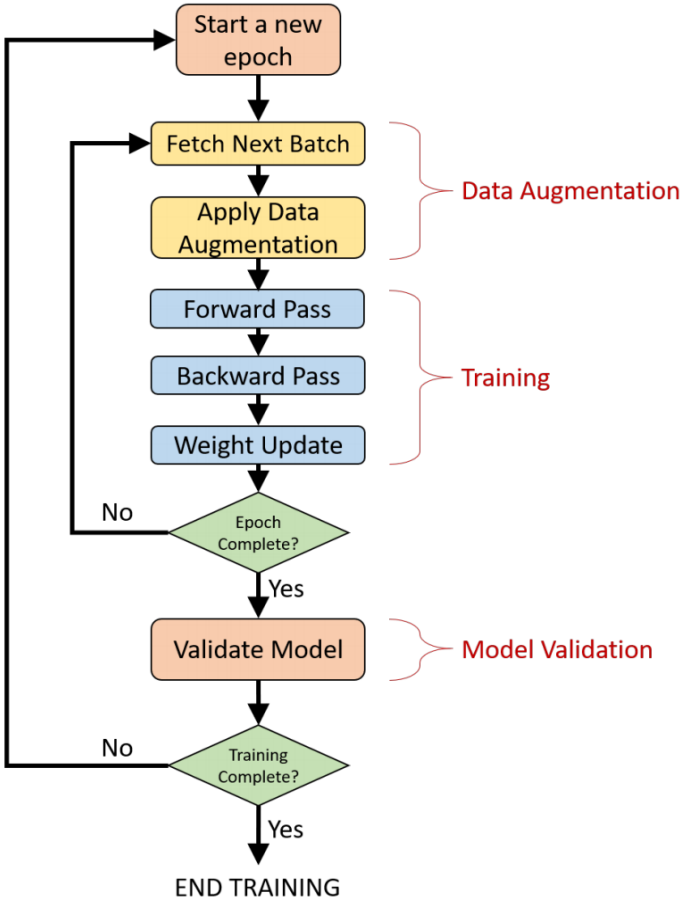
Benefits in Total execution time (Compute + Communication)

Presentation Outline

- Overview of the MVAPICH Project
- **Offloading Strategies and Benefits:**
  - Non-blocking Collectives (communication)
    - Ialltoall and P3DFFT
    - Ibcast and HPL
  - Non-blocking Point-to-point (communication)
    - Applications using 3D Stencils
  - Non-blocking Point-to-point and collective (communication and computation)
    - PETSc
  - **Offloading DL training (computation)**
- Conclusions

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

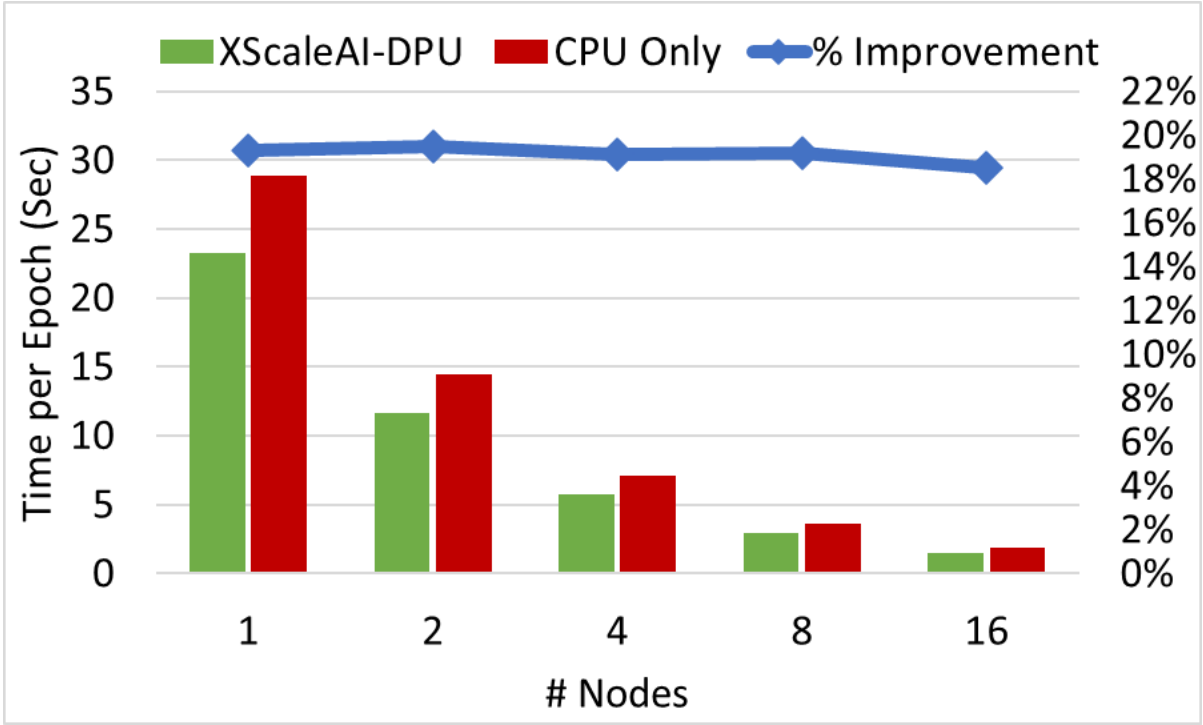
- Accelerating CPU-based DNN training with DPU support
- Based on MVAPICH2 2.3.7 with Horovod 0.25.0
- Supports all features available with the MVAPICH2 2.3.7 release (http://mvapich.cse.ohio-state.edu)
- Supports PyTorch framework for Deep Learning

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 (BF-3)

System Configuration

- Two Intel(R) Xeon(R) 16-core CPUs (32 total) E5-2697A V4 @ 2.60 GHz
- NVIDIA BlueField-3 SoC, HDR100 100Gb/s InfiniBand 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 adapters with Socket Direct



Up to 19% Performance improvement using X-ScaleAI-DPU over CPU-only training on the ResNet-20v1 model on the CIFAR10 dataset

Presentation Outline

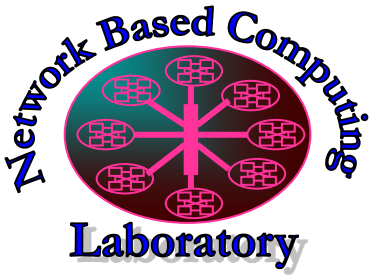
- Overview of the MVAPICH Project
- Offloading Strategies and Benefits:
  - Non-blocking Collectives (communication)
    - Ialltoall and P3DFFT
    - Ibcast and HPL
  - Non-blocking Point-to-point (communication)
    - Applications using 3D Stencils
  - Non-blocking Point-to-point and collective (communication and computation)
    - PETSc
  - Offloading DL training (computation)
- **Conclusions**

Conclusions

- DPU technology provides novel ways to offload computation and communication from host CPUs to DPU cores
- Demonstrated two ways to take advantage of the DPU technology to accelerate MPI and Deep Learning applications
- The proposed new GVMI interface allows to further optimize the overhead associated with offloading communication
- Promises potential for accelerating application performance further

Thank You!

panda@cse.ohio-state.edu



Follow us on

https://twitter.com/mvapich

Network-Based Computing Laboratory

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

The High-Performance MPI/PGAS Project

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

The High-Performance Big Data Project

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

The High-Performance Deep Learning Project

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