2022 OFA Virtual Workshop

Accelerating MPI and Deep Learning Applications with the DPU Technology

Nick Sarkauskas, Arpan Jain, Nawras Alnaasan, Tu Tran, Bharat Ramesh, Aamir Shafi, Hari Subramoni, and Dhabaleswar K. (DK) Panda

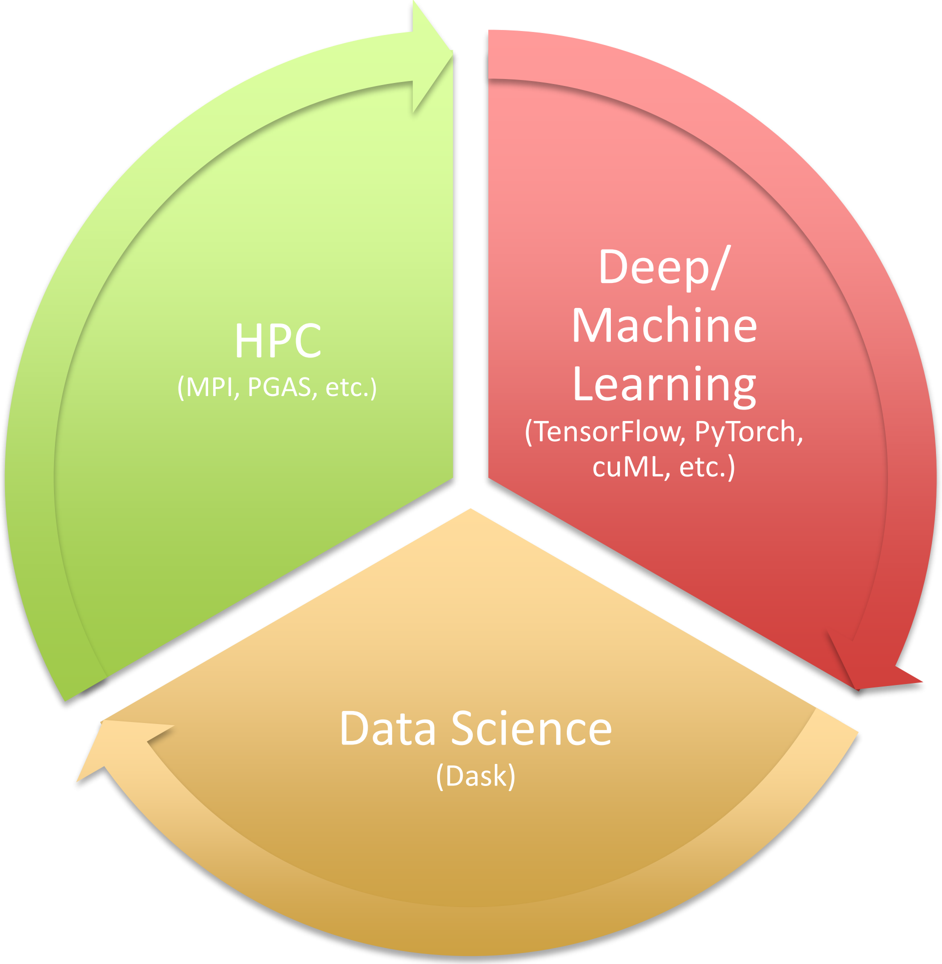
The Ohio State University

panda@cse.ohio-state.edu

OUTLINE

- **Introduction and Motivation**
- Overview of the MVAPICH2 Project
- Framework for Offloading Non-Blocking Collectives (NBC)
- Performance Benefits of Offloading NBC
- Offloading Deep Learning (DL) Applications
- Conclusion

INCREASING USAGE OF HPC, DEEP/MACHINE LEARNING, AND DATA SCIENCE



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

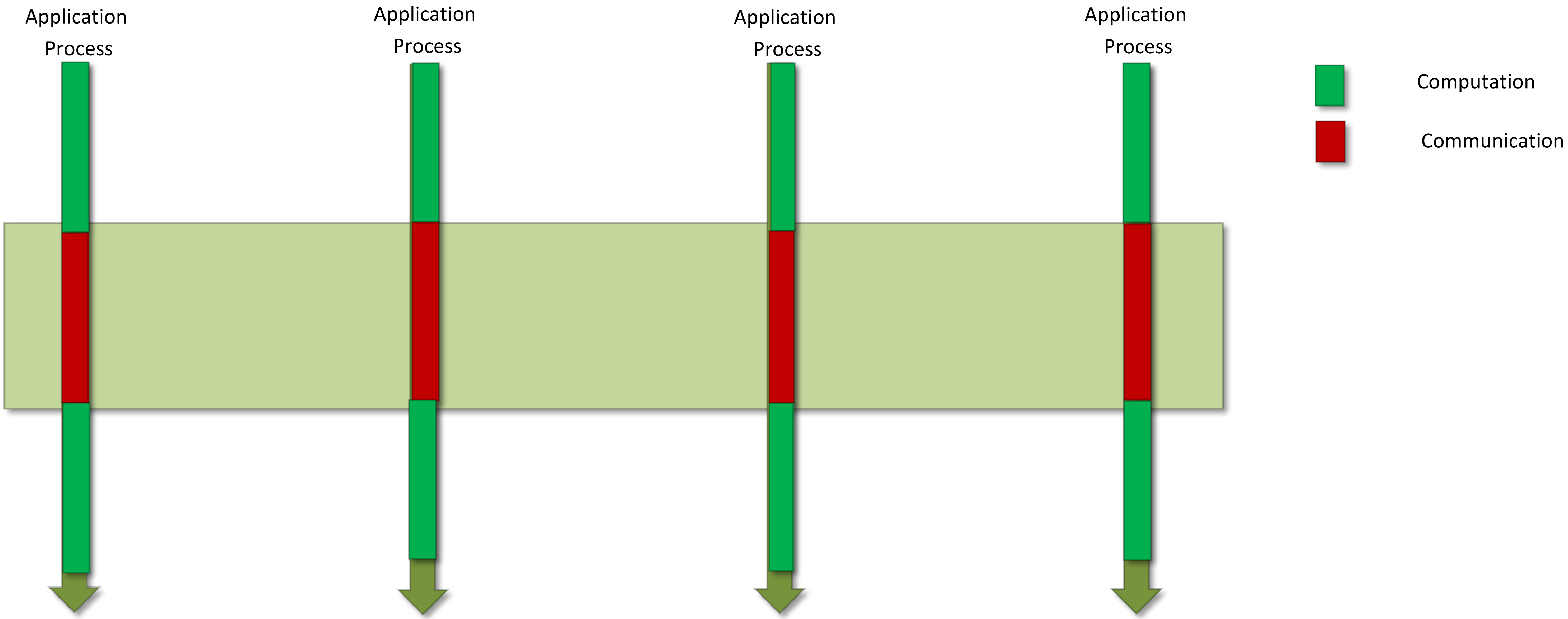
Increasing Need to Run these applications on the Cloud!!

MPI-Driven Middleware is a major component in this ecosystem!!

DESIGNING (MPI+X) FOR EXASCALE

- Scalability for million to billion processors
  - Support for highly-efficient inter-node and intra-node communication (both two-sided and one-sided)
- Scalable Collective communication
  - Offloaded
  - Non-blocking
  - Topology-aware
- Balancing intra-node and inter-node communication for next generation multi-/many-core (128-1024 cores/node)
  - Multiple end-points per node
- Support for efficient multi-threading
- Integrated Support for GPGPUs and Accelerators
- Fault-tolerance/resiliency
- QoS support for communication and I/O
- Support for Hybrid MPI+PGAS programming
  - MPI + OpenMP, MPI + UPC, MPI + OpenSHMEM, CAF, MPI + UPC++…
- Virtualization
- Energy-Awareness

PROBLEMS WITH BLOCKING COLLECTIVE OPERATIONS



Application Process

Application Process

Application Process

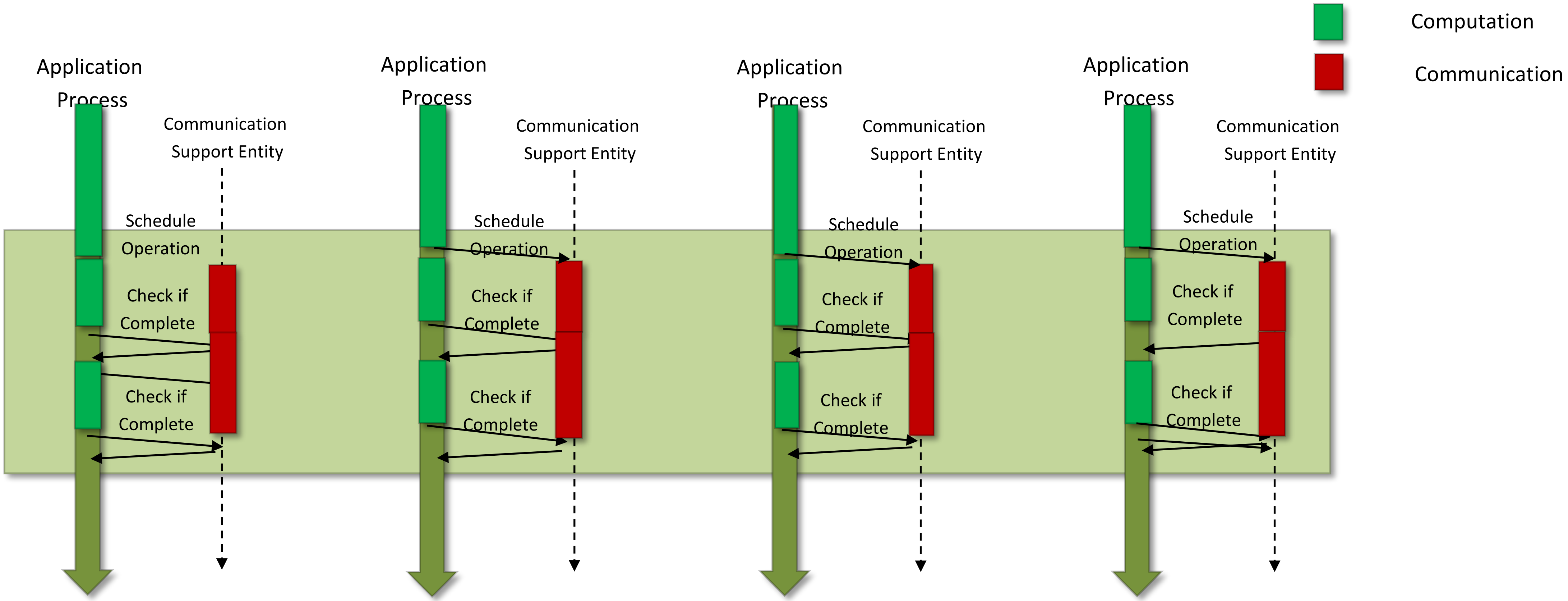
Application Process

🟩 Computation

🟥 Communication

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

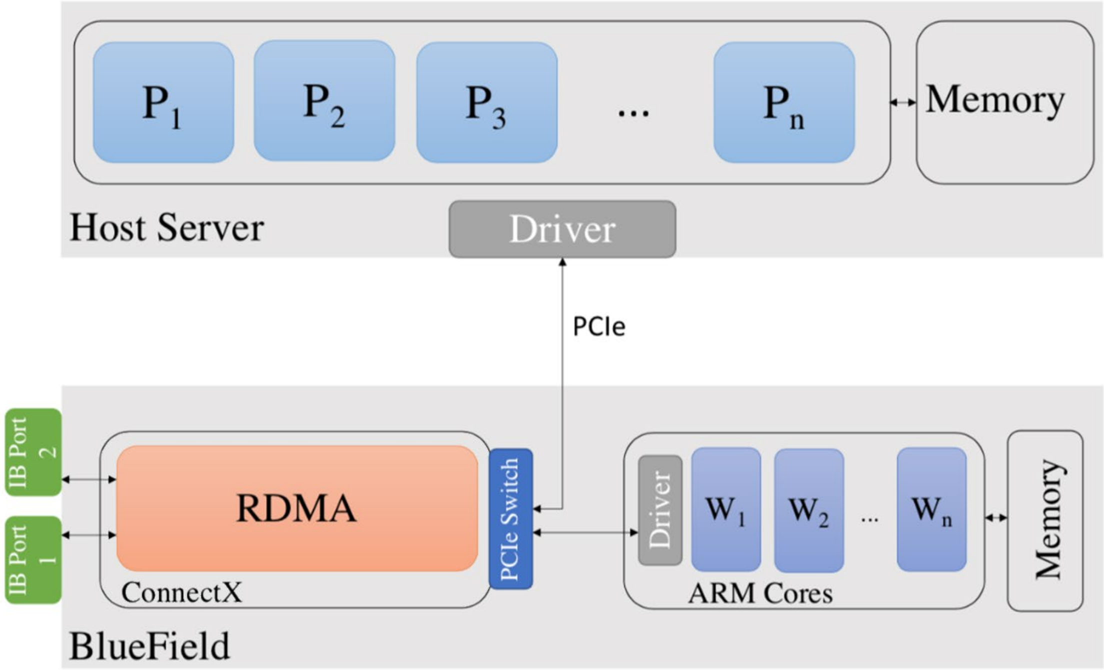
CONCEPT OF NON-BLOCKING COLLECTIVES



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

OVERVIEW OF BLUEFIELD-2 DPU

- ConnectX-6 network adapter with 200Gbps InfiniBand
- System-on-chip containing eight 64-bit ARMv8 A72 cores with 2.75 GHz each
- 16 GB of memory for the ARM cores



CAN MPI FUNCTIONS BE OFFLOADED TO BLUEFIELD-DPU?

- **Can we exploit additional compute capabilities of modern BlueField DPUs into existing MPI middleware to extract**
  - Peak pure communication performance
  - Overlap of communication and computation

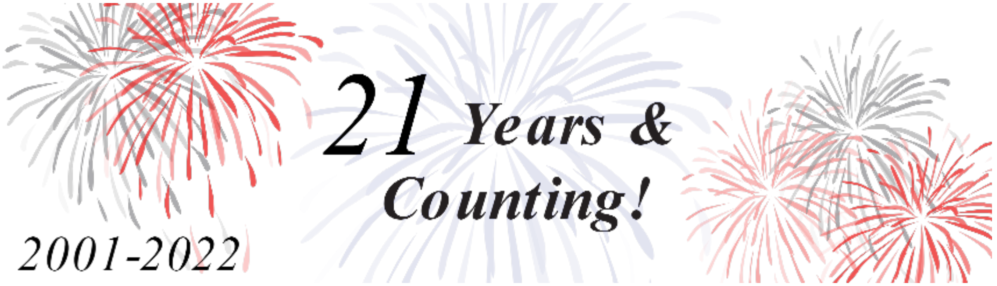
  For non-blocking collective communications?
- **What will be the benefits at the applications level?**

OUTLINE

- Introduction and Motivation
- **Overview of the MVAPICH2 Project**
- Framework for Offloading Non-Blocking Collectives (NBC)
- Performance Benefits of Offloading NBC
- Offloading Deep Learning (DL) Applications
- Conclusion

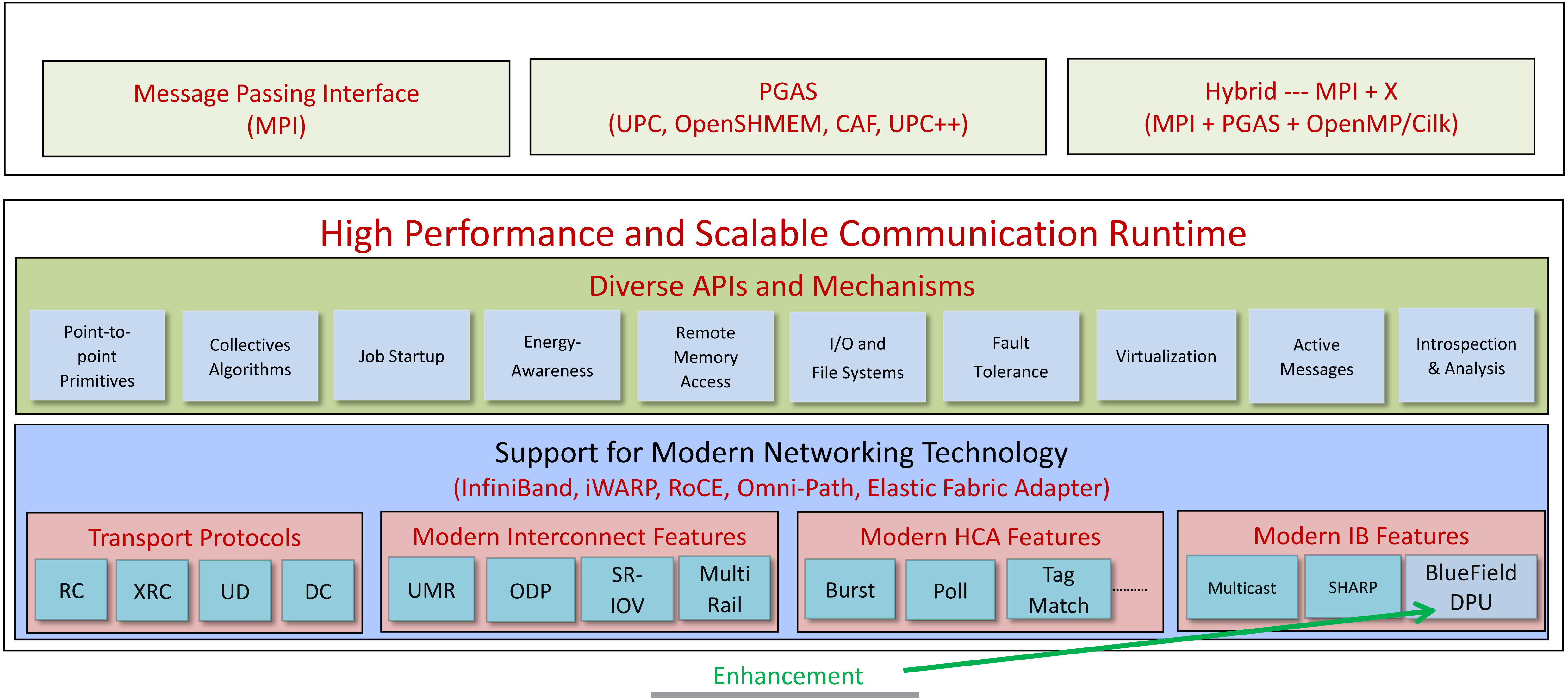
OVERVIEW OF THE MVAPICH2 PROJECT

- **High Performance open-source MPI Library**
- **Support for multiple interconnects**
  - InfiniBand, Omni-Path, Ethernet/iWARP, RDMA over Converged Ethernet (RoCE), AWS EFA, and Rockport Networks
- **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 3,200 organizations in 89 countries
- More than 1.57 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov ‘21 ranking)
  - 4th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 13th, 448, 448 cores (Frontera) at TACC
  - 26th, 288,288 cores (Lassen) at LLNL
  - 38th, 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 13th ranked TACC Frontera system
- Empowering Top500 systems for more than 16 years

ENHANCING MVAPICH2 SOFTWARE ARCHITECTURE WITH DPU



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
- XRC
- UD
- DC

Modern Interconnect Features

- UMR
- ODP
- SR-IOV
- Multi Rail

Modern HCA Features

- Burst
- Poll
- Tag Match
- ....

Modern IB Features

- Multicast
- SHARP
- BlueField DPU

Enhancement

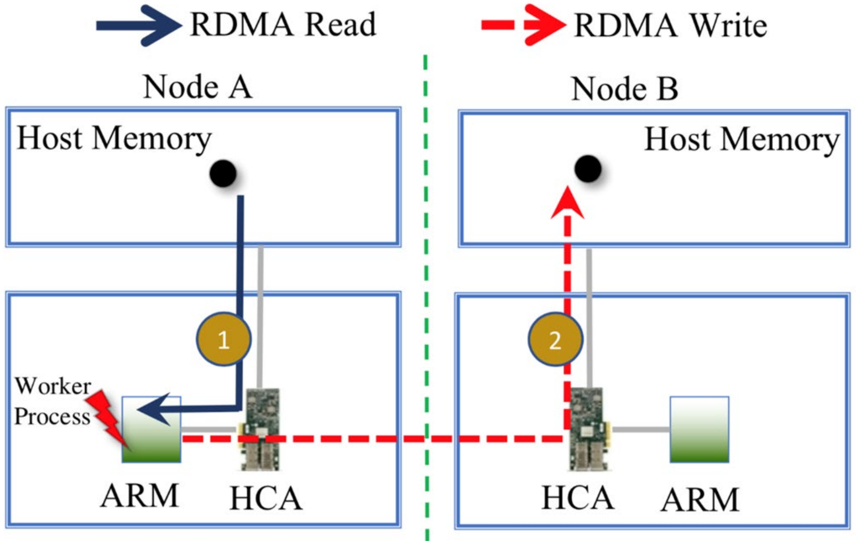
OUTLINE

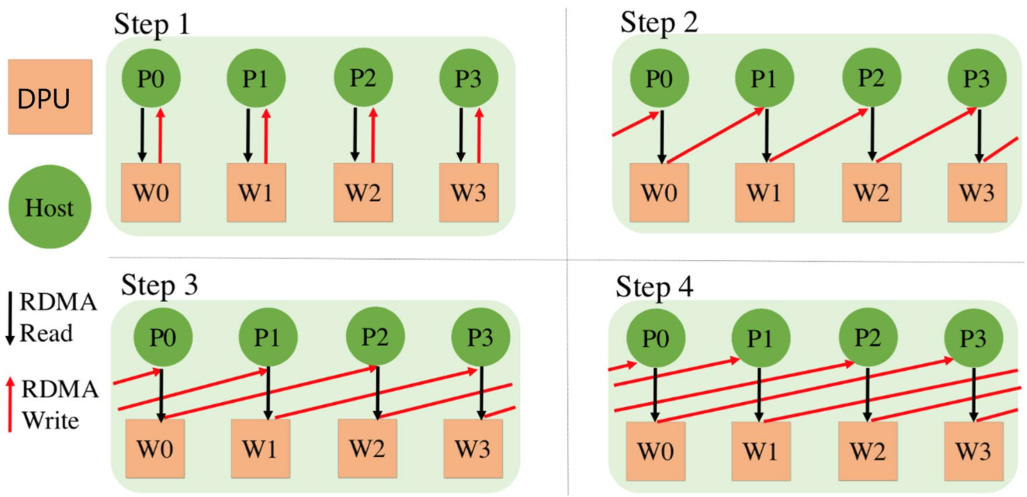
- Introduction and Motivation
- Overview of MVAPICH2 Project
- **Framework for Offloading Non-Blocking Collectives (NBC)**
- Performance Benefits of Offloading NBC
- Offloading Deep Learning (DL) Applications
- Conclusion

PROPOSED OFFLOAD FRAMEWORK

- Non-blocking collective operations are offloaded to a set of "worker processes"
- BlueField is set to separated host mode
  - Worker processes are spawned to the ARM cores of BlueField
- Worker processes progress the collective on behalf of the host processes
- Once message exchanges are completed, worker processes notify the host processes about the completion of the non-blocking operation

PROPOSED NONBLOCKING ALLTOALL DESIGN





- Non-blocking collective operations are offloaded to a set of Worker processes running on the ARM cores of BF-2 (BlueField-2)

- Alltoall is realized by an efficient proposed scatter destination algorithm

MVAPICH2-DPU LIBRARY 2022.02 RELEASE

X-ScaleSolutions

- Based on MVAPICH2 2.3.6
- Released on 02/15/22
- Supports all features available with the MVAPICH2 2.3.6 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 Allgather (MPI_Iallgather) 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.

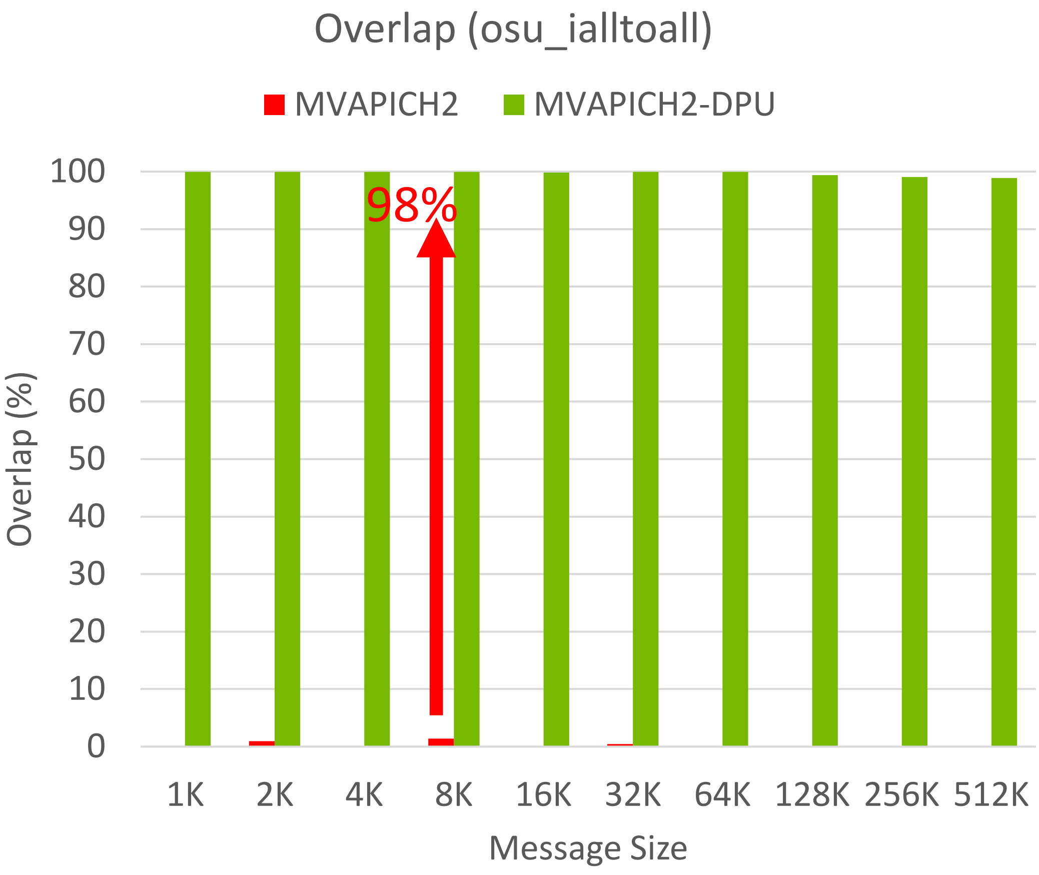
OUTLINE

- Introduction and Motivation
- Overview of MVAPICH2 Project
- Framework for Offloading Non-Blocking Collectives (NBC)
- **Performance Benefits of Offloading NBC**
- Offloading Deep Learning (DL) Applications
- Conclusion

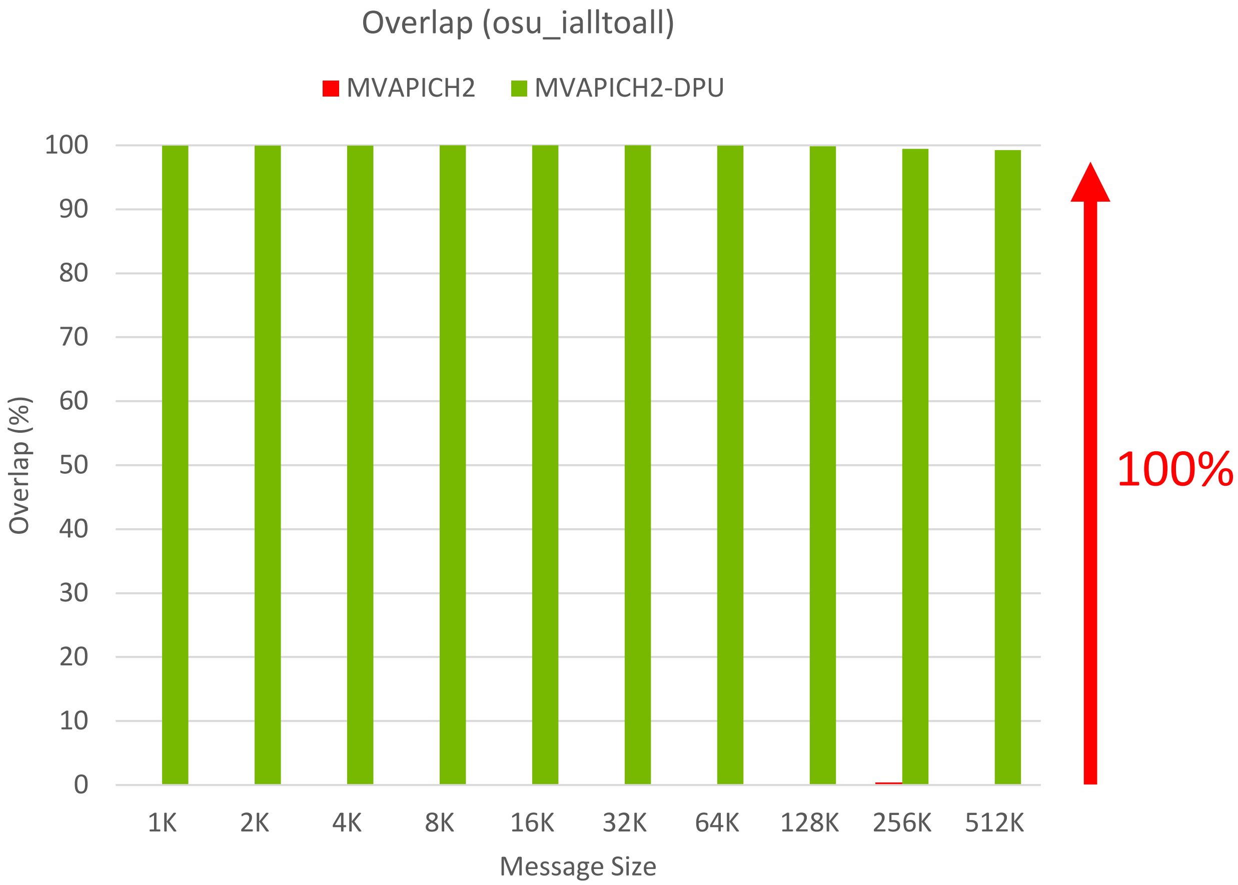
EXPERIMENTAL SETUP FOR PERFORMANCE EVALUATION

- **HPC Advisory Council High-Performance Computing Center**
  - Cluster has 32 compute-node with Broadwell series of Xeon dual-socket, 16-core processors operating at 2.60 GHz with 256 GB RAM
  - NVIDIA BlueField-2 adapters are equipped with 8 ARM cores operating at 2.0 GHz with 16 GB RAM
- **Based on the MVAPICH2-DPU MPI library**
- **OSU Micro Benchmark for nonblocking Alltoall, Allgather, Bcast, and P3DFFT Application**

OVERLAP OF COMMUNICATION AND COMPUTATION WITH OSU_IALLTOALL (32 NODES)



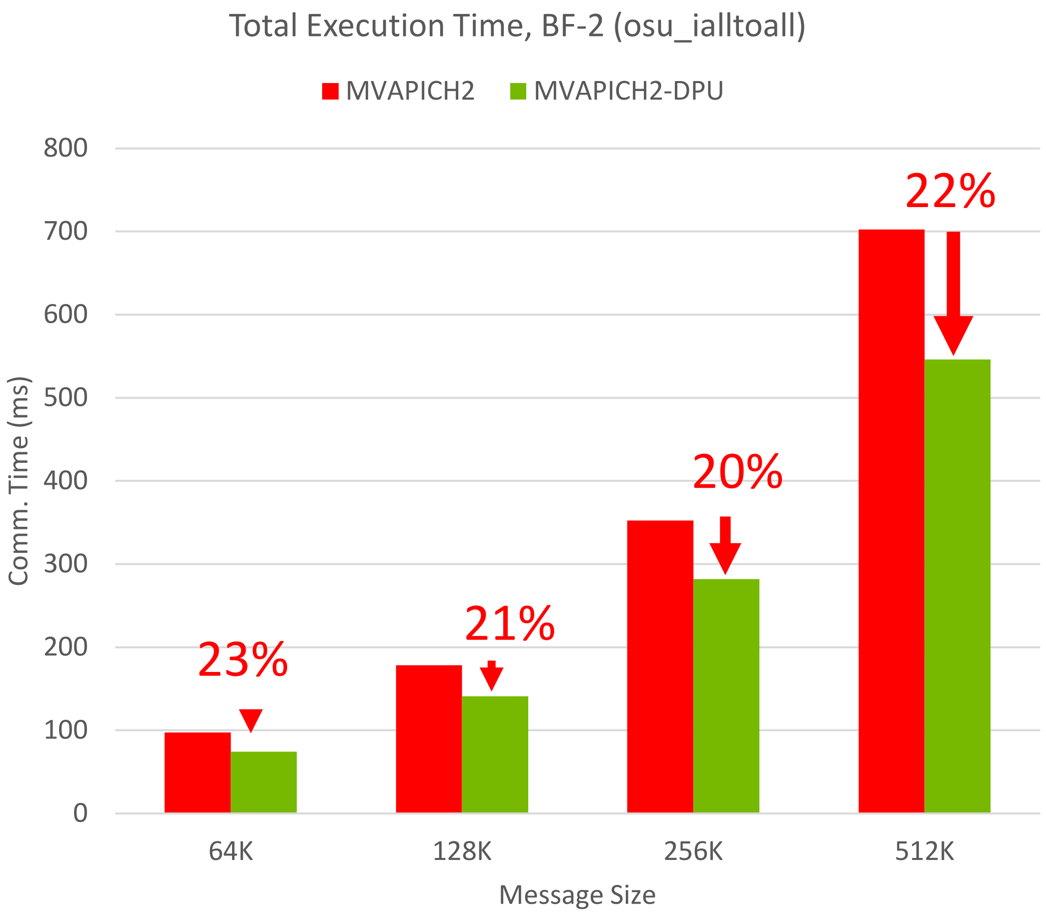
32 Nodes, 16 PPN



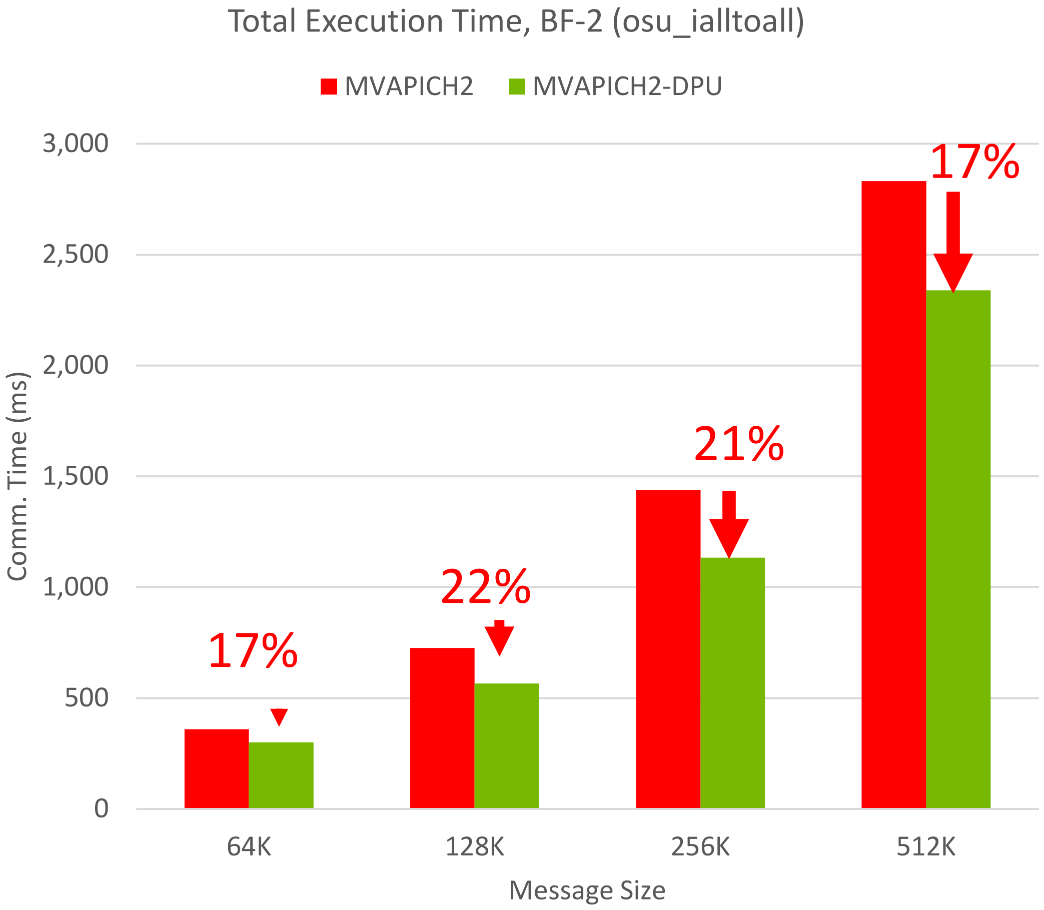
32 Nodes, 32 PPN

Delivers peak overlap

TOTAL EXECUTION TIME WITH OSU_IALLTOALL (32 NODES)



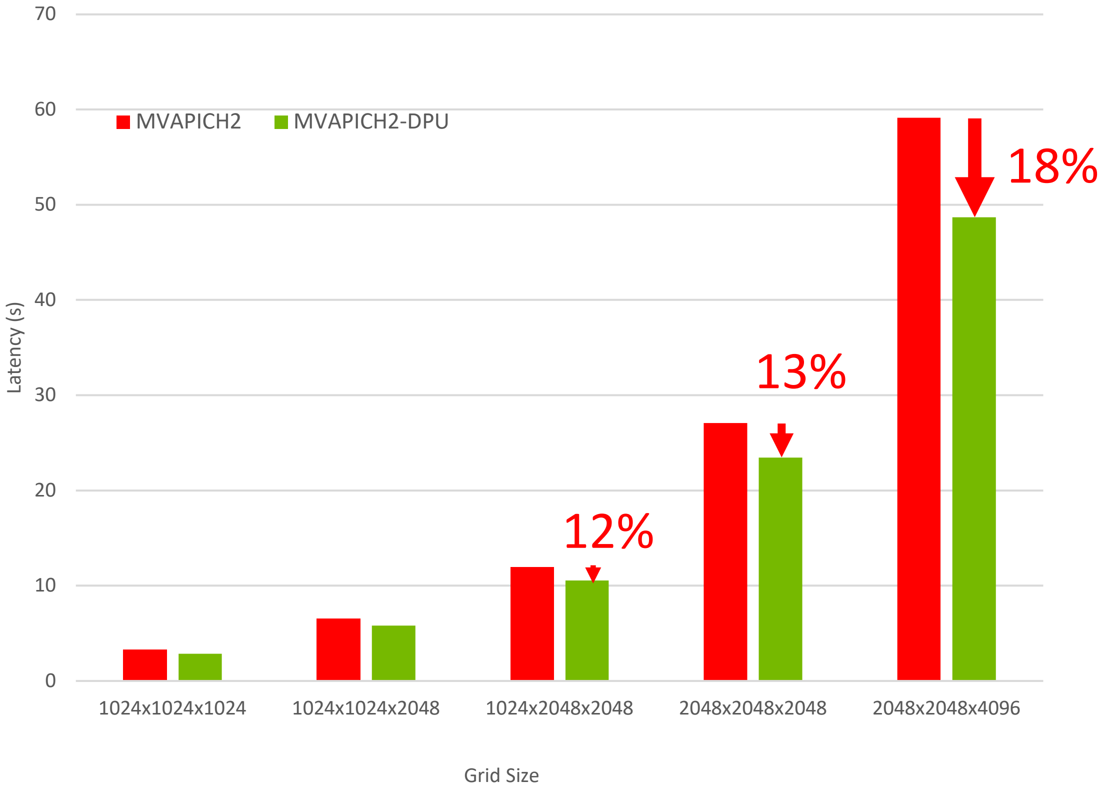
32 Nodes, 16 PPN



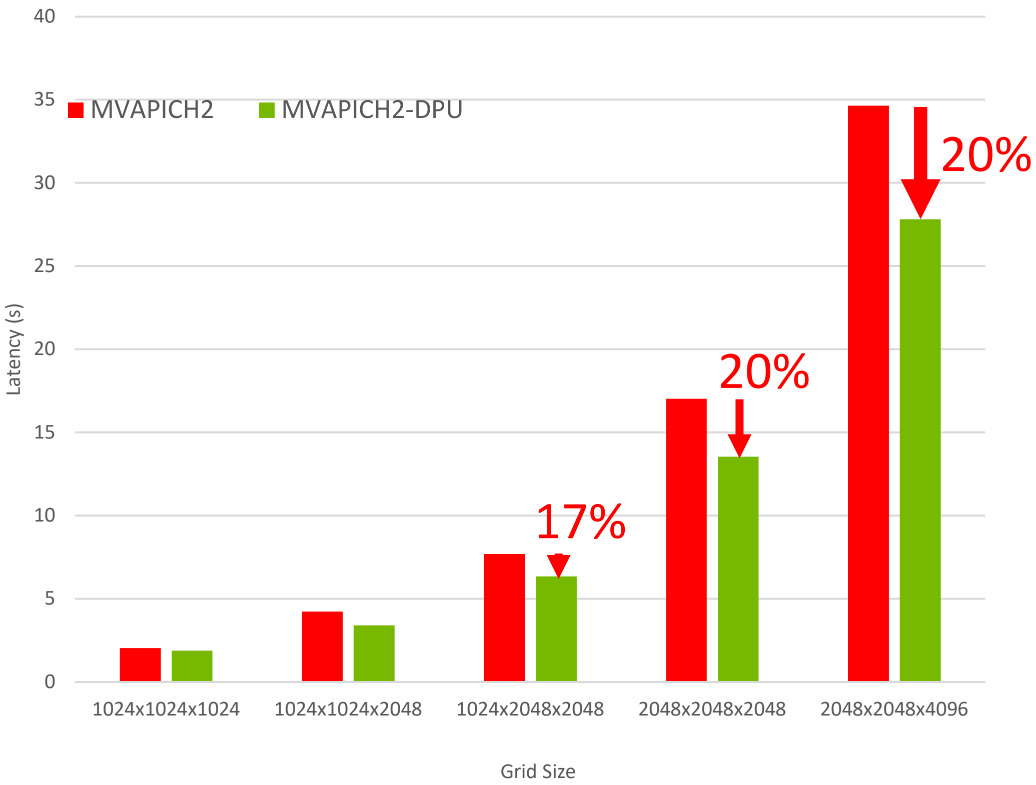
32 Nodes, 32 PPN

Benefits in Total execution time (Compute + Communication)

P3DFFT APPLICATION EXECUTION TIME (16 NODES)



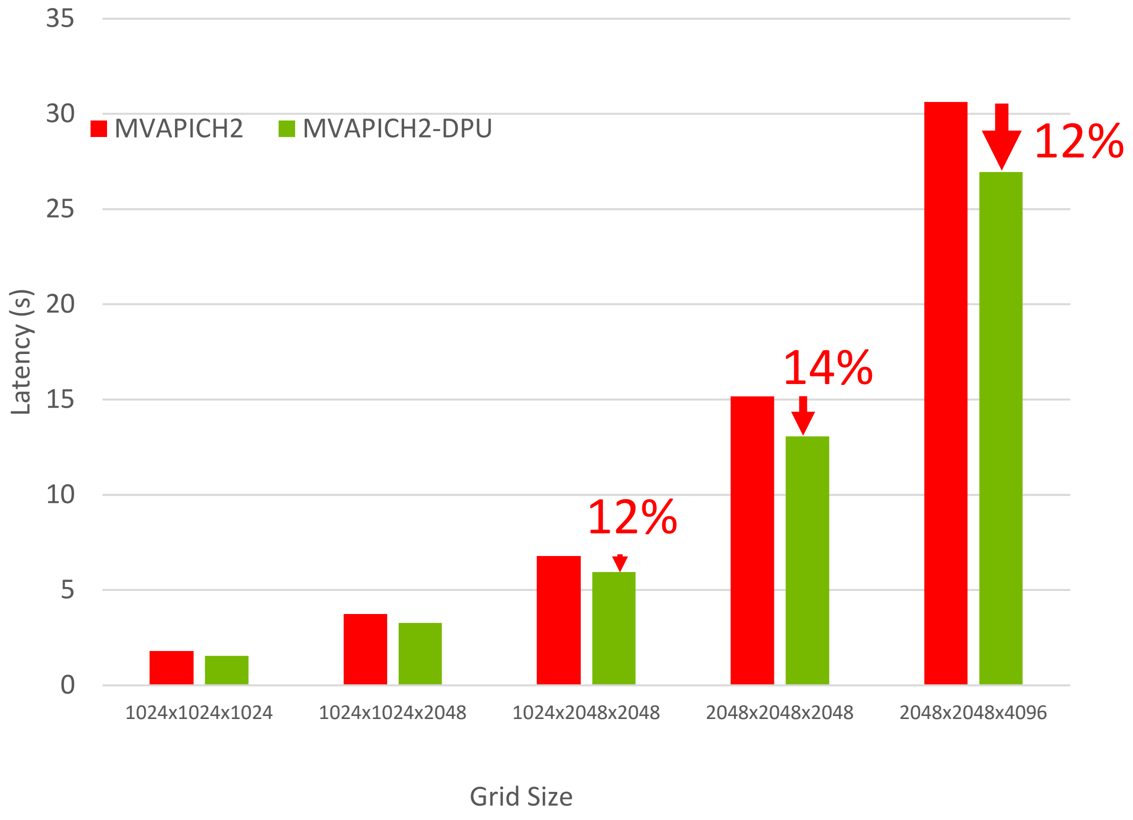
16 Nodes, 16 PPN



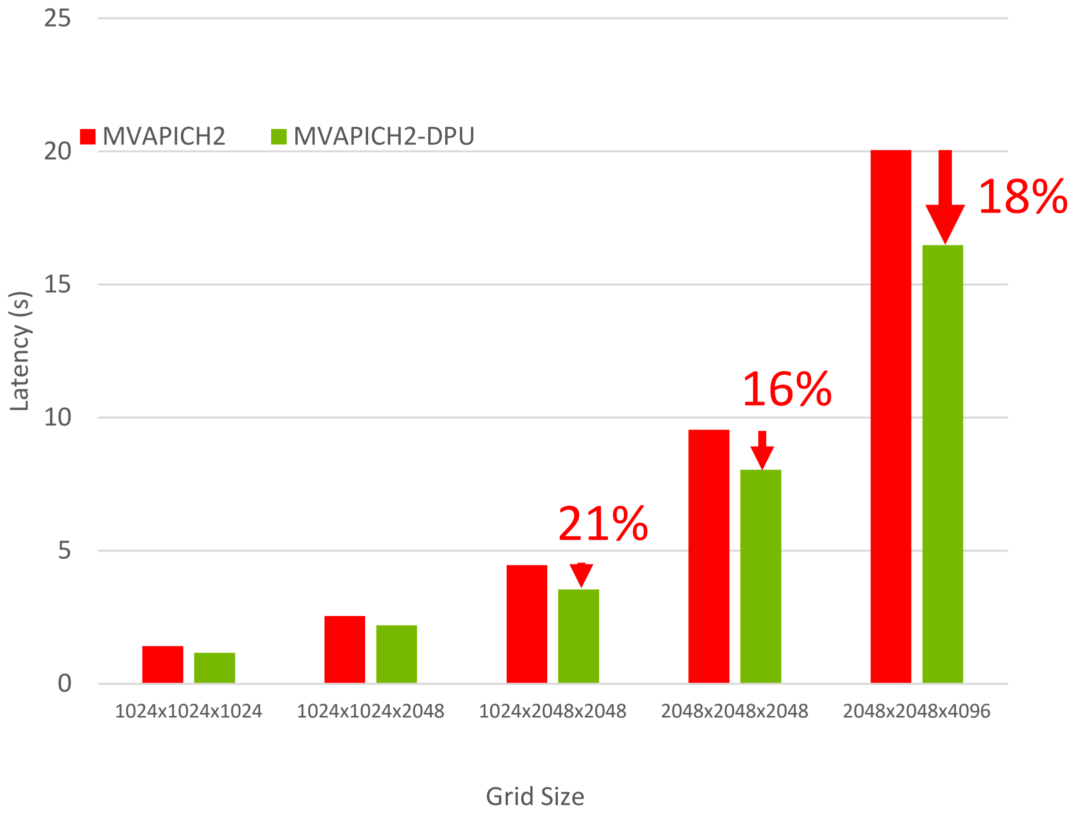
16 Nodes, 32 PPN

Benefits in application-level execution time

P3DFFT APPLICATION EXECUTION TIME (32 NODES)



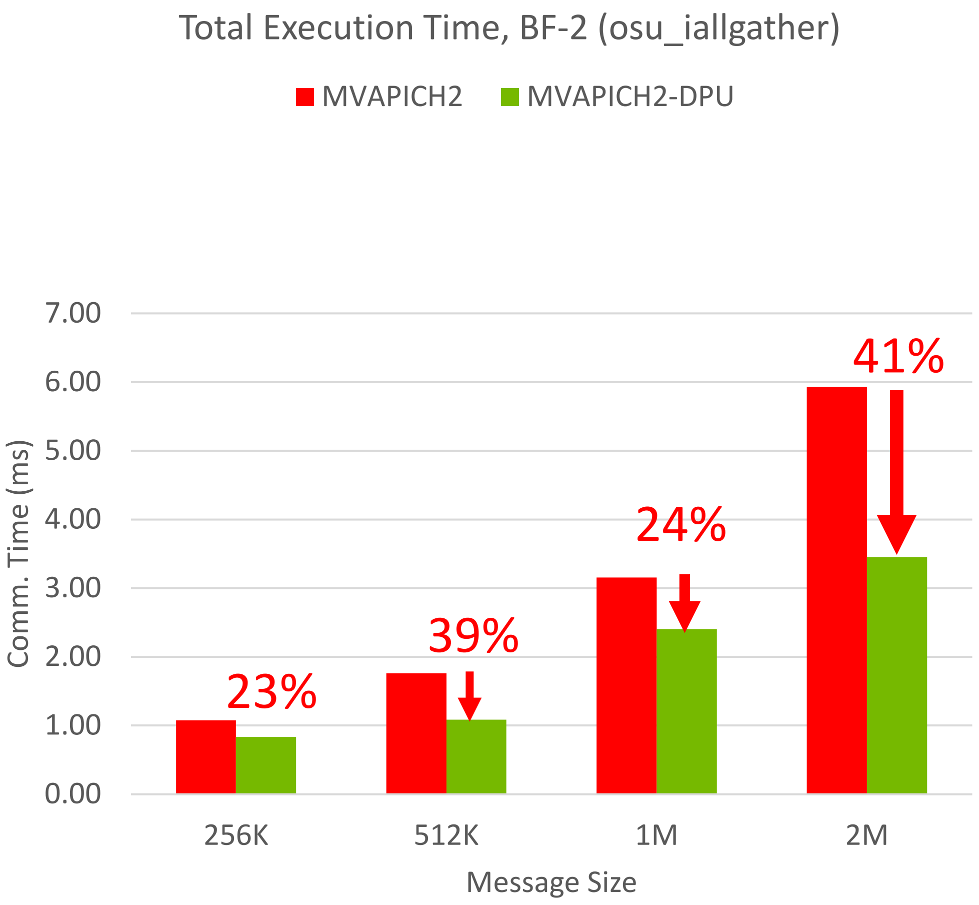
32 Nodes, 16 PPN



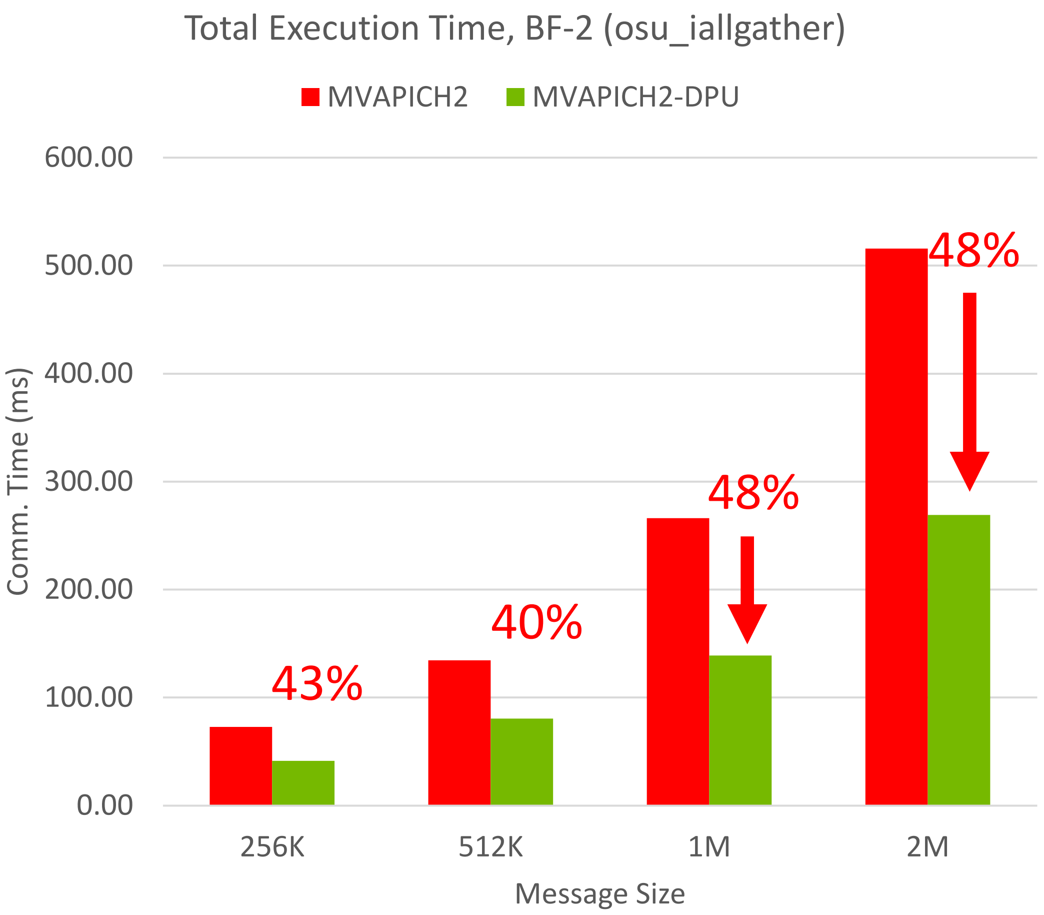
32 Nodes, 32 PPN

Benefits in application-level execution time

TOTAL EXECUTION TIME WITH OSU_IALLGATHER (16 NODES)



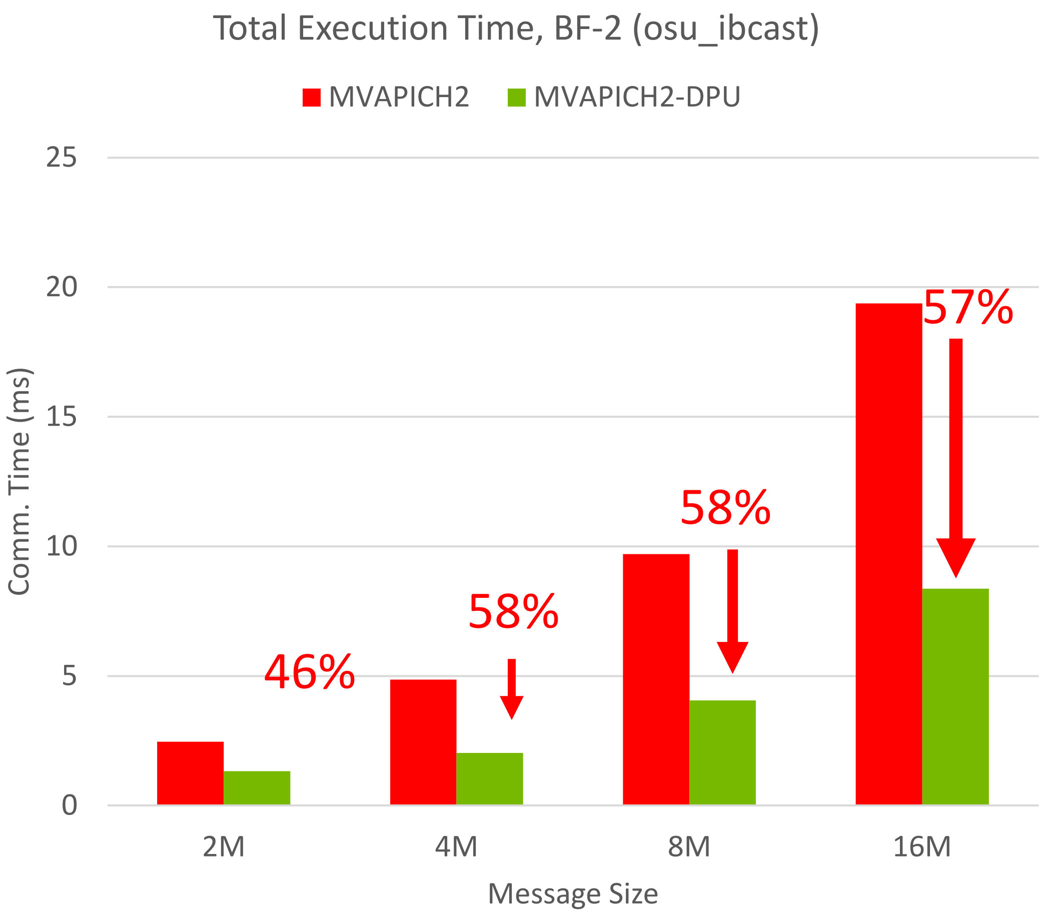
16 Nodes, 1 PPN



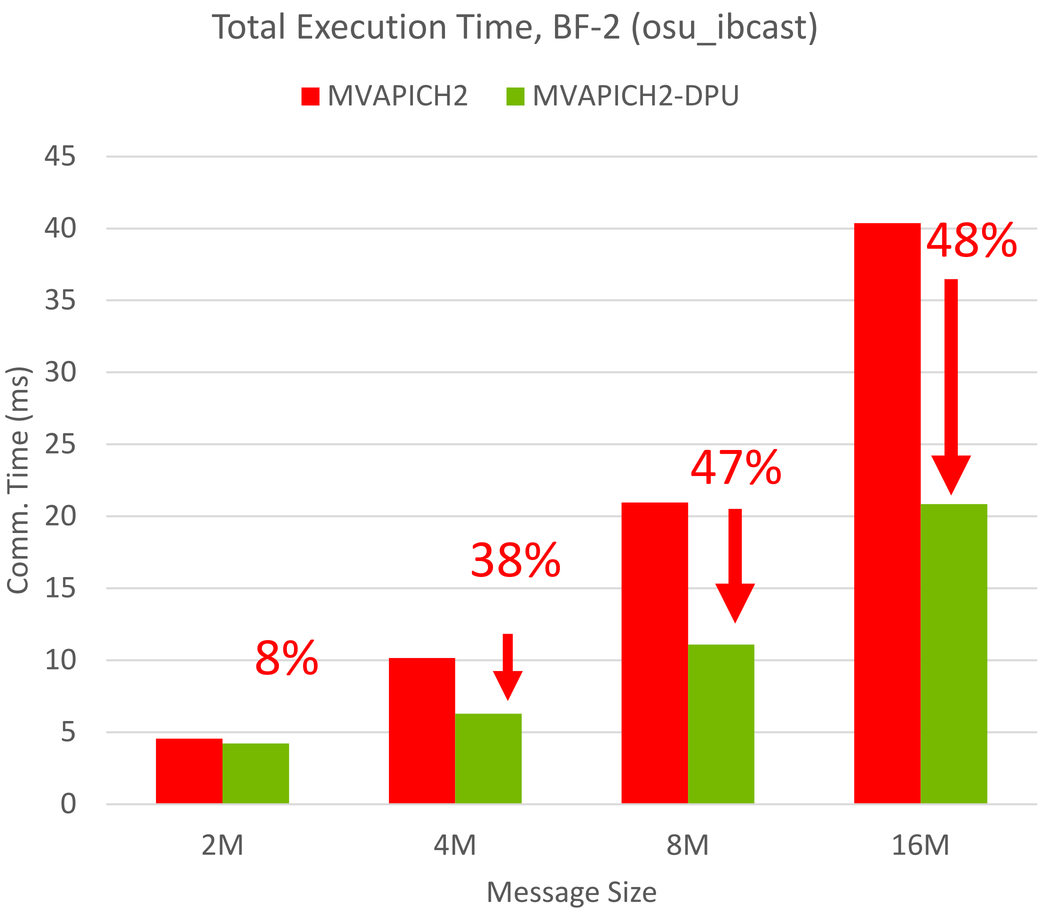
16 Nodes, 16 PPN

Benefits in Overall Iallgather (Computation and Communication)

TOTAL EXECUTION TIME WITH OSU_IBCAST (NODES)



32 Nodes, 1 PPN



32 Nodes, 16 PPN

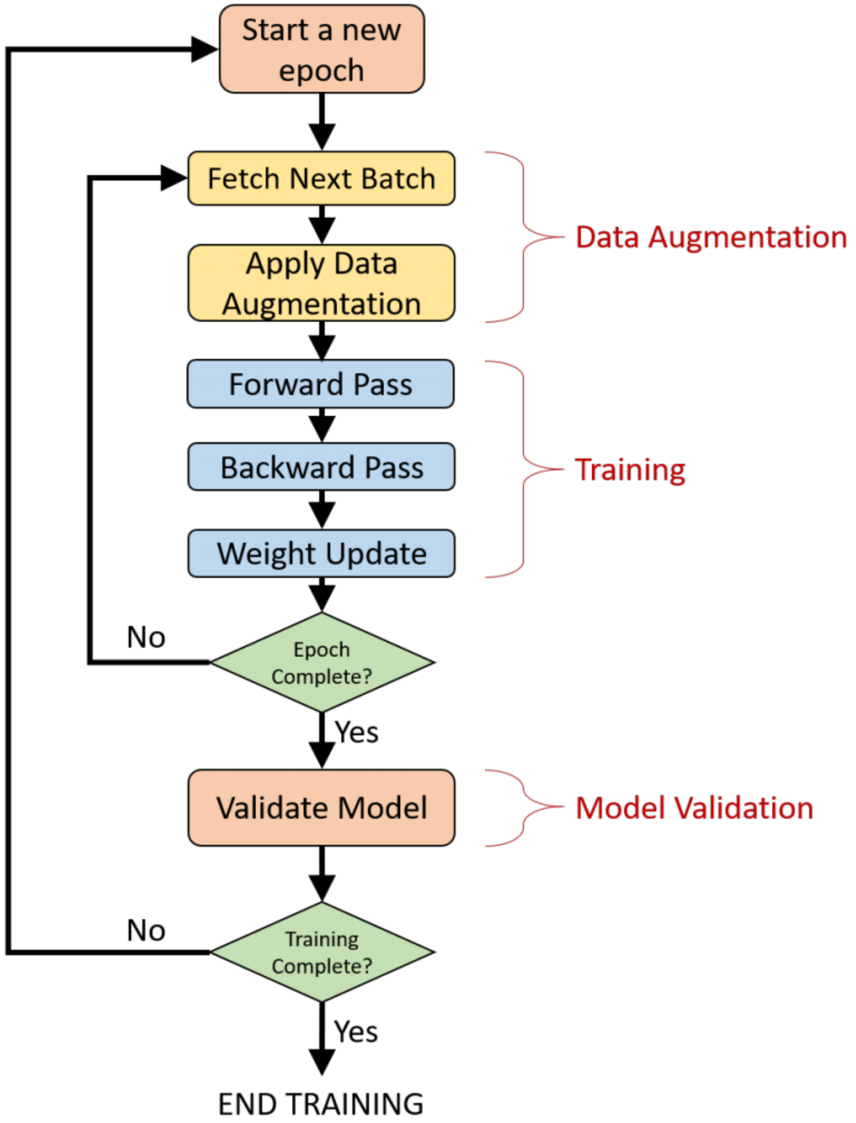
Benefits in Overall Ibcast (Computation and Communication)

OUTLINE

- Introduction and Motivation
- Overview of MVAPICH2 Project
- Framework for Offloading Non-Blocking Collectives (NBC)
- Performance Benefits of Offloading NBC
- **Offloading Deep Learning (DL) Applications**
- Conclusion

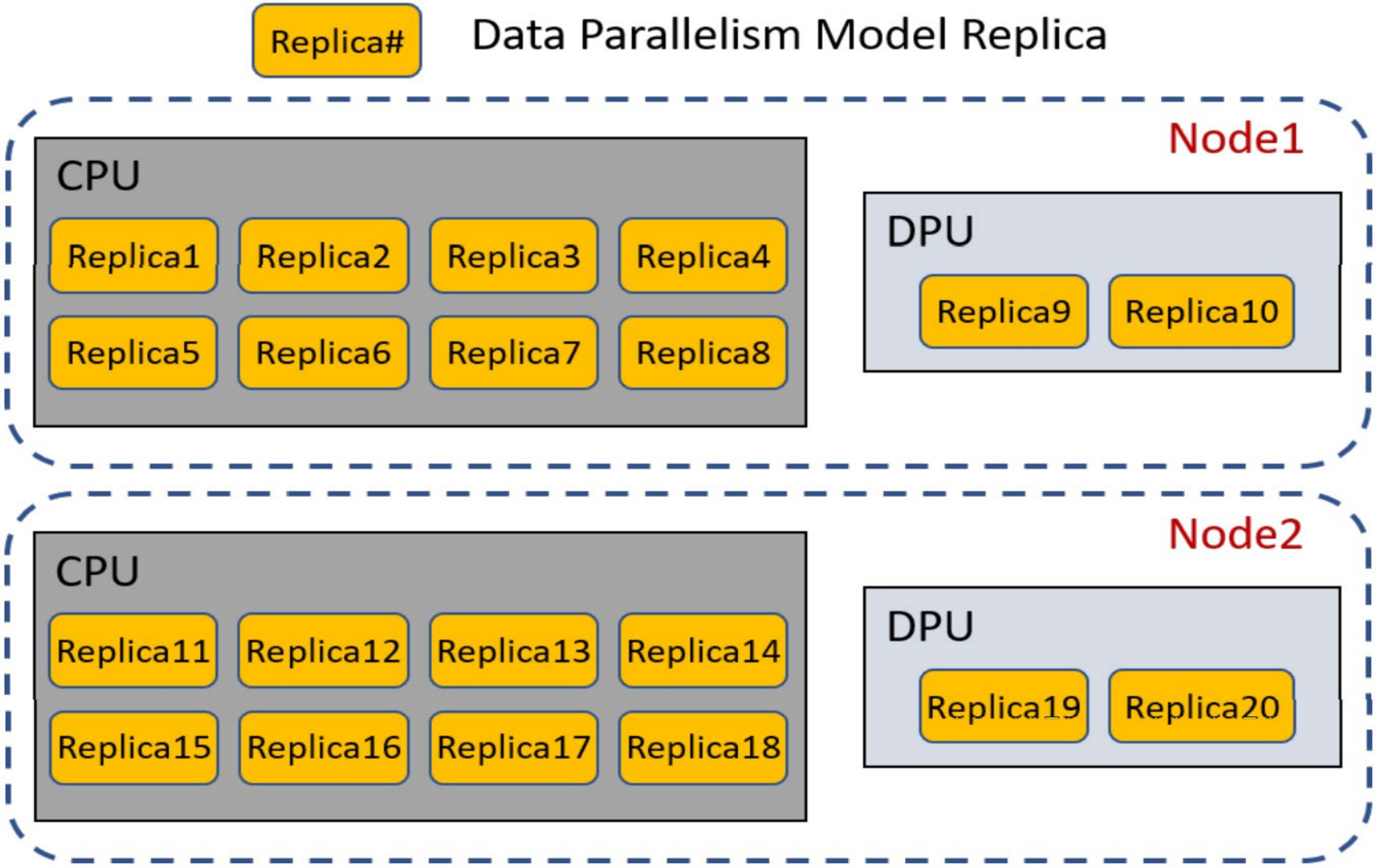
EXPLOITING DPUS FOR DEEP NEURAL NETWORK TRAINING

- **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.**



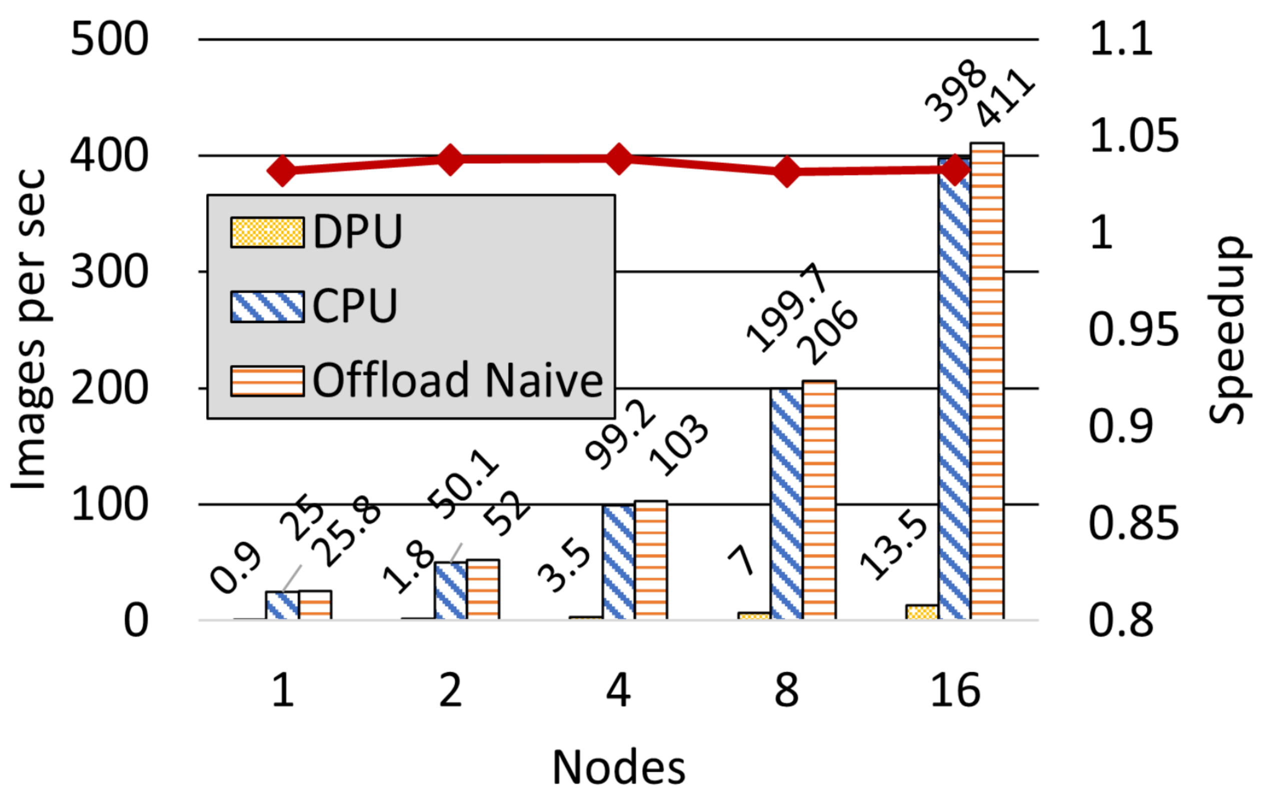
OFFLOAD NAIVE (O-N): OFFLOADING DL TRAINING USING DATA PARALLELISM

- **Data parallelism can used to train DNN on DPUs**



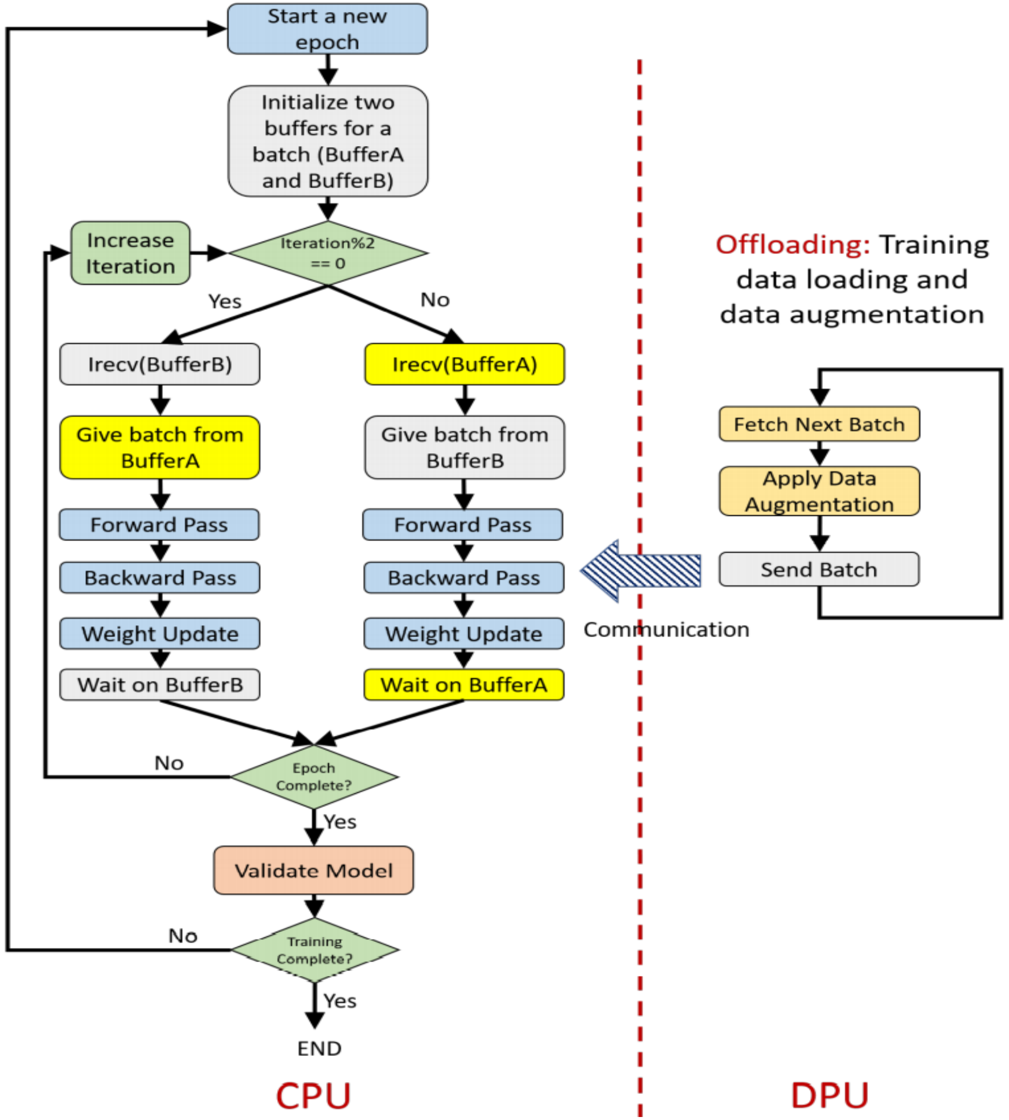
ACCELERATING DNN TRAINING USING OFFLOAD-NAIVE

- **Time per iteration can be used to distribute the work (batch size) between CPU and DPU**
- **Speedup:**
  - We report up to 1.03X speedup
  - Maximum speedup possible: 1.04X
- **Offload-Naive does not give significant speedup as forward and backward pass are compute-intensive tasks and DPUs are not as powerful as CPUs**



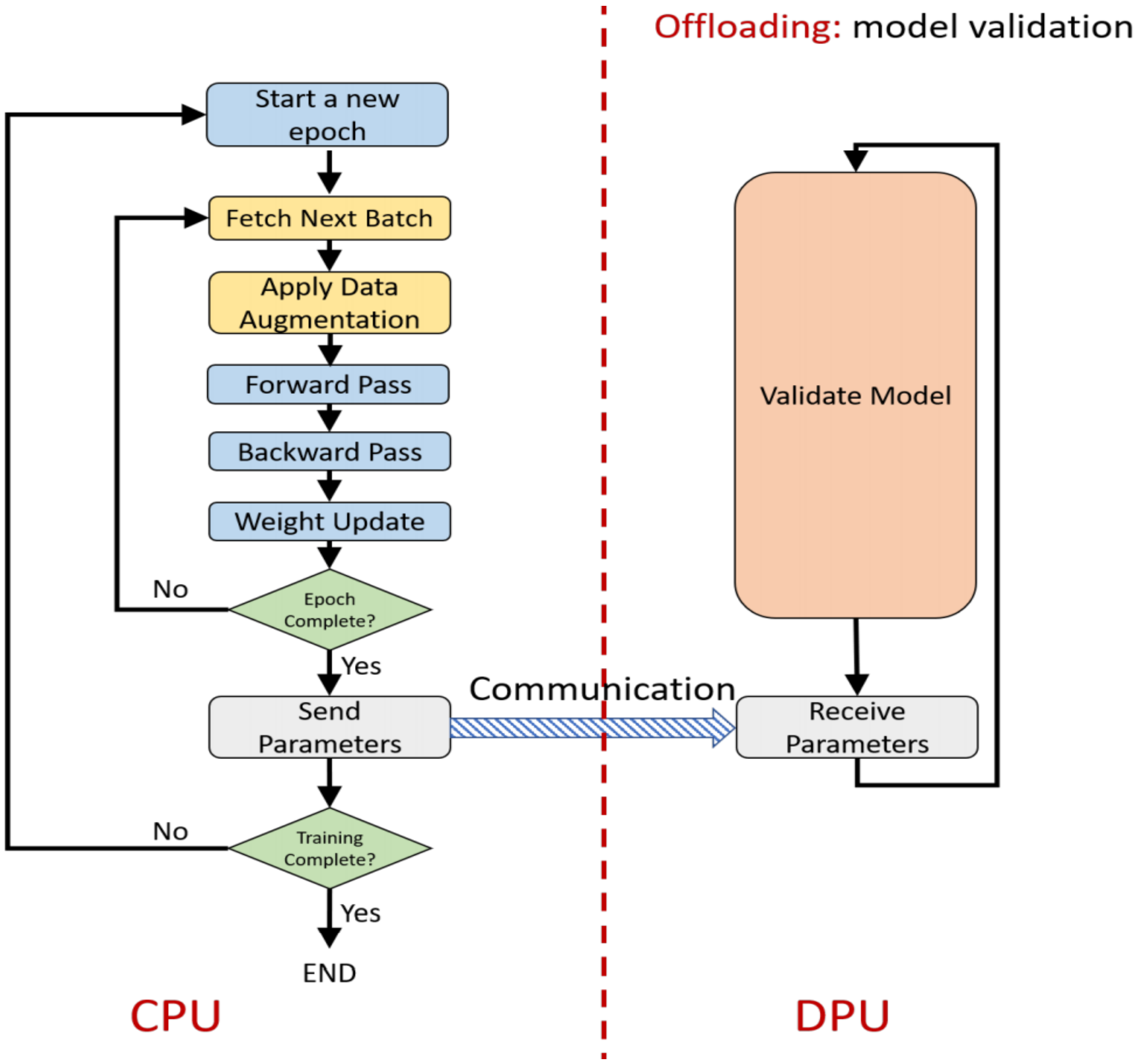
DESIGN 1: OFFLOAD DATA AUGMENTATION (O-DA)

- **Offloads the reading of training data from memory and data augmentation on input data to DPUs.**
- **Creates two types of processes**
  - Training processes (on CPU)
  - Data Augmentation processes (On DPU)
- **Initializes two buffers to enable asynchronous communication**
- **Each training processes has one data augmentation processes on DPU.**



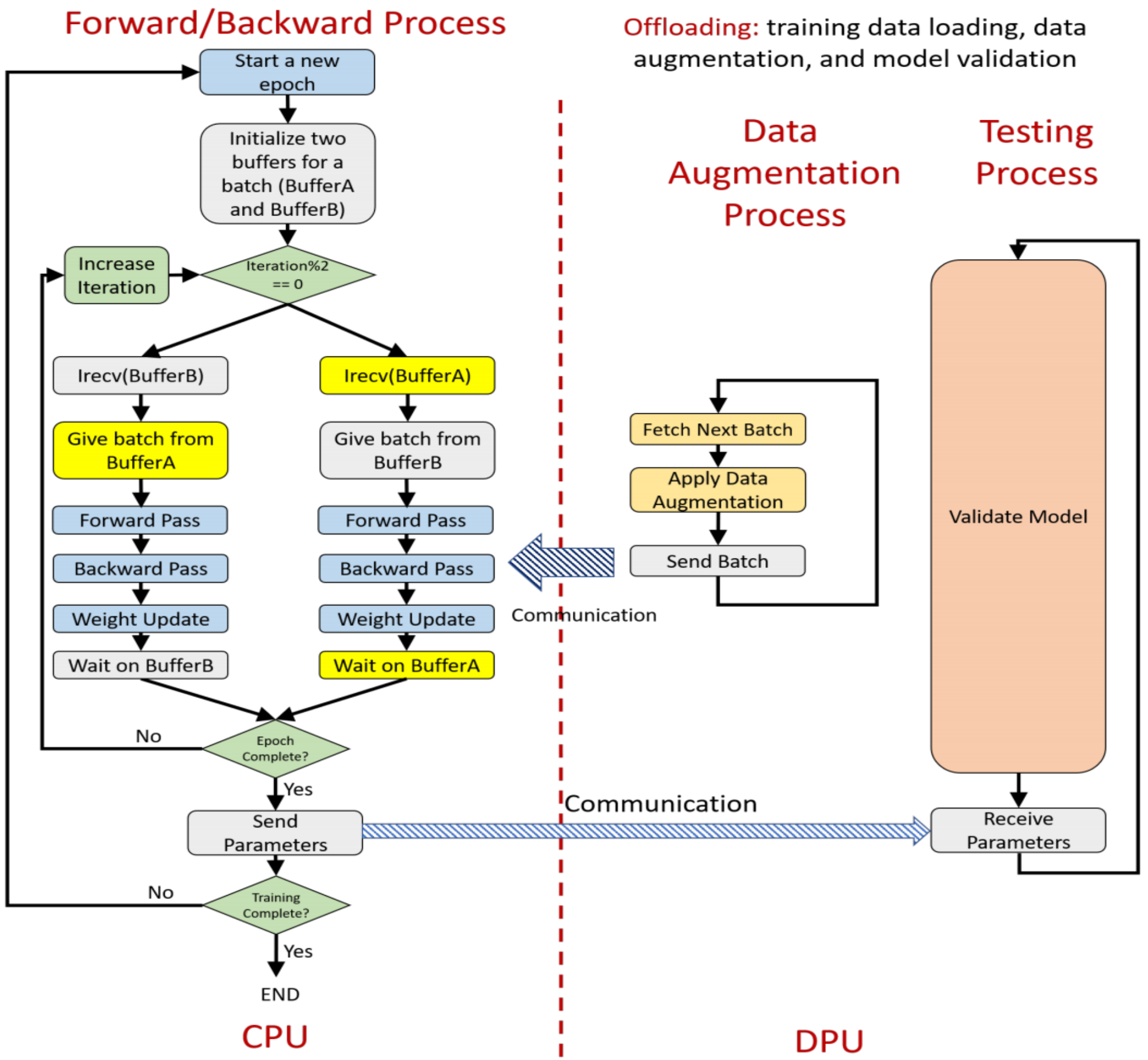
DESIGN 2: OFFLOAD MODEL VALIDATION (O-MV)

- Offloads validation of model after each epoch to DPUs.
- Model validation is a less compute-intensive task as it has only forward pass
- Creates two types of processes
  - Training processes (on CPU)
  - Testing processes (On DPU)
- One communication operation per epoch
- Validation data is equally divided among testing processes.



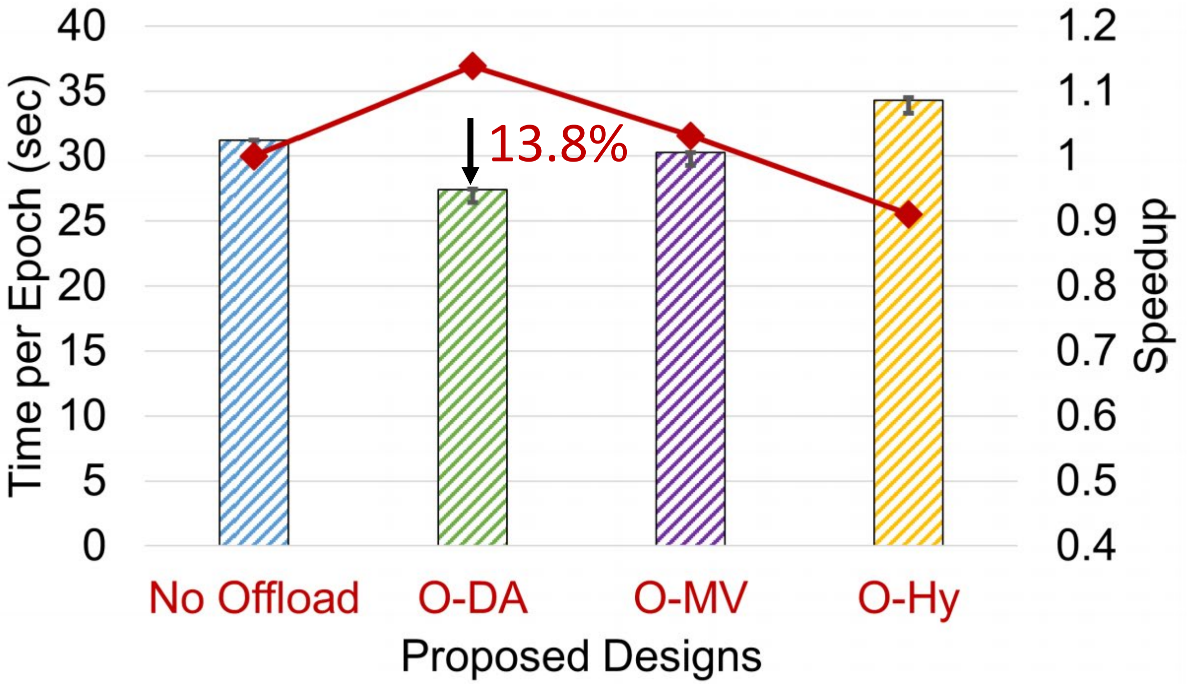
DESIGN 3: OFFLOAD HYBRID (O-HY)

- **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)
- **Each Data Augmentation process on DPU supports multiple training processes.**
- **Data Augmentation processes does asynchronous communication and Testing processes does synchronous communication**

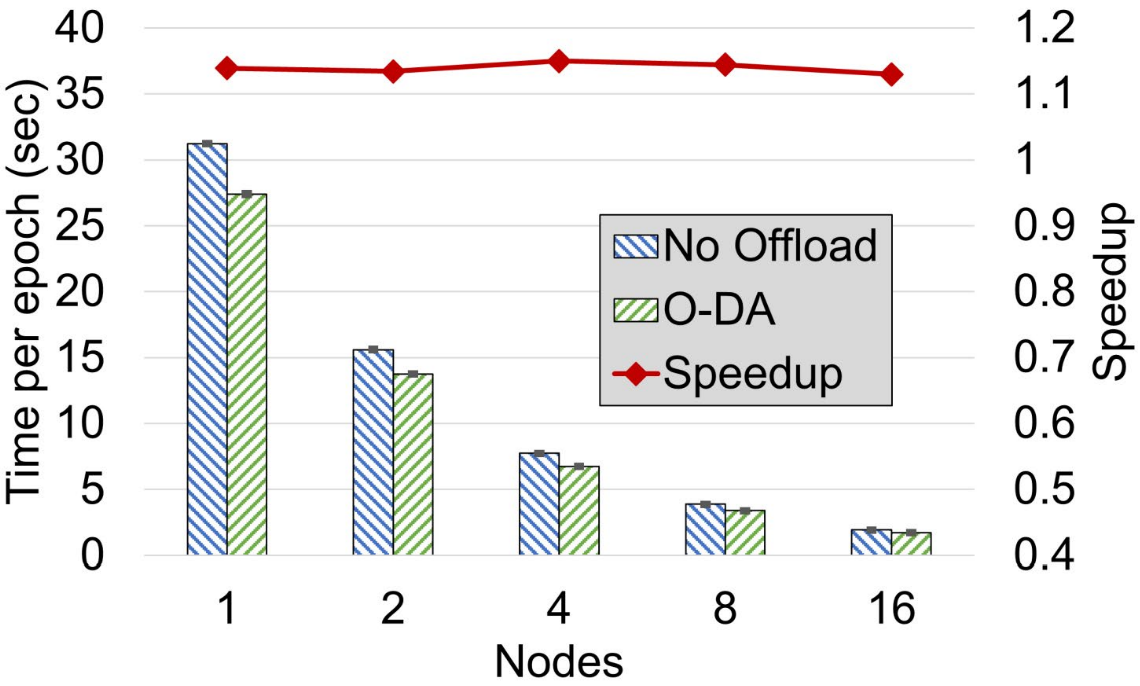


ACCELERATING DNN TRAINING RESNET-20 ON CIFAR-10 DATASET WITH BLUEFIELD-2 DPU

- **Speedup**
  - Single node: O-DA (13.8%) and O-MV (3.1%)
  - Multi-node: Achieves average 13.9% speedup on 1-16 nodes



Single Node Experiments



Multi-Node Experiments

PUBLICATIONS

- M. Bayatpour, N. Sarkauskas, H. Subramoni, J. M. Hashmi, and Dhabaleswar K. (DK) Panda, BluesMPI: Efficient MPI Non-blocking Alltoall offloading Designs on Modern BlueField Smart NICs, Int'l Supercomputing Conference (ISC '21)
- A. Jain, N. Alnassan, A. Shafi, H. Subramoni, and Dhabaleswar K. (DK) Panda, Accelerating CPU-based Distributed DNN Training on Modern HPC Clusters using BlueField-2 DPUs, Hot Interconnects (HotI '21)
- M. Bayatpour, N. Sarkauskas, M. Bayatpour, A. Tran, B. Ramesh, H. Subramoni, and Dhabaleswar K. (DK) Panda, Large-Message Nonblocking MPI_Iallgather and MPI_Ibcast Offload via BlueField-2 DPU, Int'l Conference on High-Performance Computing, Data Analytics, and Data Science (HiPC '21)

CONCLUSION

- Proposed efficient designs for the MVAPICH2 MPI library that utilize the BlueField DPU to progress MPI non-blocking collective operations
- Design provides close to 100% overlap of communication and computation for non-blocking Alltoall, Allgather and Bcast
- Reduces the total execution time of P3DFFT application up to 21% on 1,024 processes
- Working on offloading designs for other non-blocking collective and MPI operations
- Demonstration of how AI (DL/ML) workloads can take advantage of DPU technology

THANK YOU!

Network Based Computing Laboratory

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/