Accelerating HPC and AI Applications with Data Processing Units (DPUs)

Donglai Dai

d.dai@x-scalesolutions.com

X-ScaleSolutions

http://x-scalesolutions.com

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) HPC and Cloud resources?

Presentation Outline

- **Overview of X-ScaleSolutions**
- Overview of the MVAPICH Project
- Offloading Strategies and Benefits:
  - Non-blocking Collectives (communication)
    - Ialltoall and P3DFFT
    - Ibcast and HPL
    - Ialltoallv and Xcompact3D
  - 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 and I/O)
- Conclusions

Overview of X-ScaleSolutions

- Started in 2018
- Bring innovative and efficient end-to-end solutions, services, support, and training to our customers
- Commercial support and training for the state-of-the-art communication libraries
  - High-Performance and Scalable MVAPICH2 Library and its families (MVAPICH2-X, MVAPICH2-GDR, MVAPICH2-Azure, MVAPICH2-AWS, MVAPICH, MVAPICH-Plus, and OSU INAM)
  - High-Performance Deep Learning/Machine Learning Libraries (MPI4DL and MPI4cuML)
  - High-Performance Big Data Libraries (RDMA-Hadoop, RDMA-Spark, RDMA-HBase, and RDMA-Memcached, MPI4Spark and MPI4Dask)

Commercial Support Features and Benefits

- Benefits:
  - Help and guidance with installation of the library
  - Platform-specific optimizations and tuning
  - Timely support for operational issues encountered with the library
  - Flexible Service Level Agreements
  - 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
- Support being provided to National Laboratories and International HPC centers
- Flexibility in providing such support
  - Directly to end organizations
  - Through third-party integrators

Value-Added Products

- Design and develop innovative and value-added products
- Winner of multiple U.S. DOE SBIR grants
- Market these products for HPC and AI applications with commercial support
- A Silver ISV member of the OpenPOWER Consortium

Overview of Products

- X-ScaleHPC: High-Performance Optimized Solution for HPC applications
- X-ScaleAI: High-Performance Solution with Deep Introspection for AI applications
- MVAPICH2-DPU: High-Performance MVAPICH2 for Accelerating Applications with NVIDIA's DPU technology
- X-ScaleHPL-DPU: Accelerating HPL with DPU Offload
- X-ScaleAI-DPU: Accelerating DL Training with DPU Offload

Presentation Outline

- Overview of X-ScaleSolutions
- **Overview of the MVAPICH Project**
- Offloading Strategies and Benefits:
  - Non-blocking Collectives (communication)
    - Ialltoall and P3DFFT
    - Ibcast and HPL
    - Ialltoallv and Xcompact3D
  - 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 and I/O)
- Conclusions

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, 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-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
  - MVAPICH2-GDR with support for NVIDIA (since 2014) and AMD (since 2020) GPUs
- Tools:
  - OSU MPI Micro-Benchmarks (OMB), since 2003
  - OSU InfiniBand Network Analysis and Monitoring (INAM), since 2015
- New Series
  - **MVAPICH 3.x and MVAPICH-Plus 3.x (since 2022)**

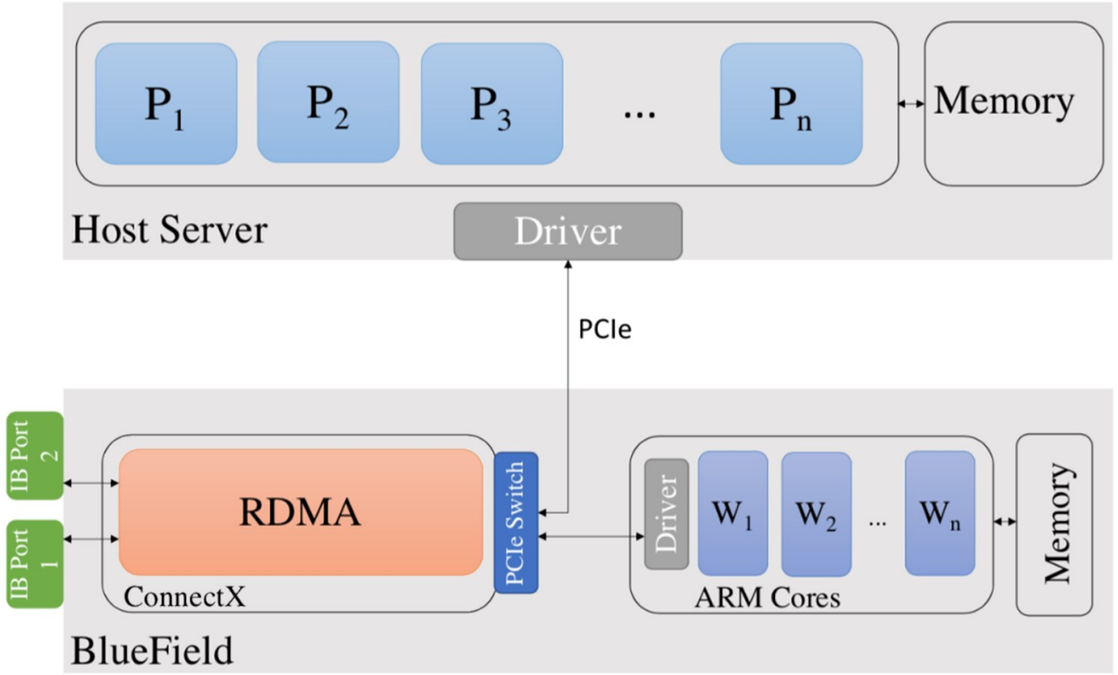
- Used by more than 3,325 organizations in 90 countries
- More than 1.73 Million downloads from the OSU site directly
- Empowering many TOP500 clusters (Jun '23 ranking)
  - 7th , 10,649,600-core (Sunway TaihuLight) at NSC, Wuxi, China
  - 21st , 448, 448 cores (Frontera) at TACC
  - 36th, 288,288 cores (Lassen) at LLNL
  - 49th, 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 21st ranked TACC Frontera system
- Empowering Top500 systems for more than 18 years

Presentation Outline

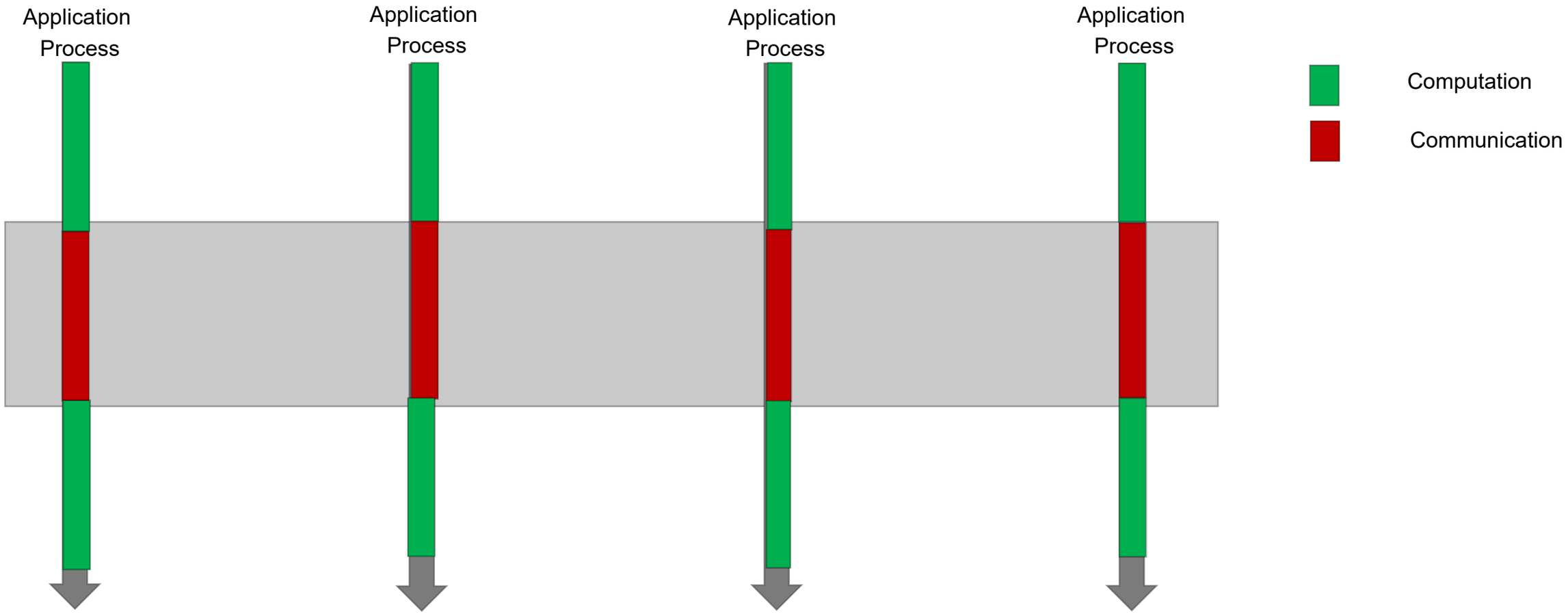
- Overview of X-ScaleSolutions
- Overview of the MVAPICH Project
- **Offloading Strategies and Benefits:**
  - **Non-blocking Collectives (communication)**
    - **Ialltoall and P3DFFT**
    - **Ibcast and HPL**
    - **Ialltoallv and Xcompact3D**
  - Non-blocking Point-to-point (communication)
    - Applications using 3D Stencils
  - Non-blocking Point-to-point and collective (communication and computation)
    - PETSc
  - Offloading DL training (computation and I/O)
- 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
- 32 GB of memory for the ARM cores

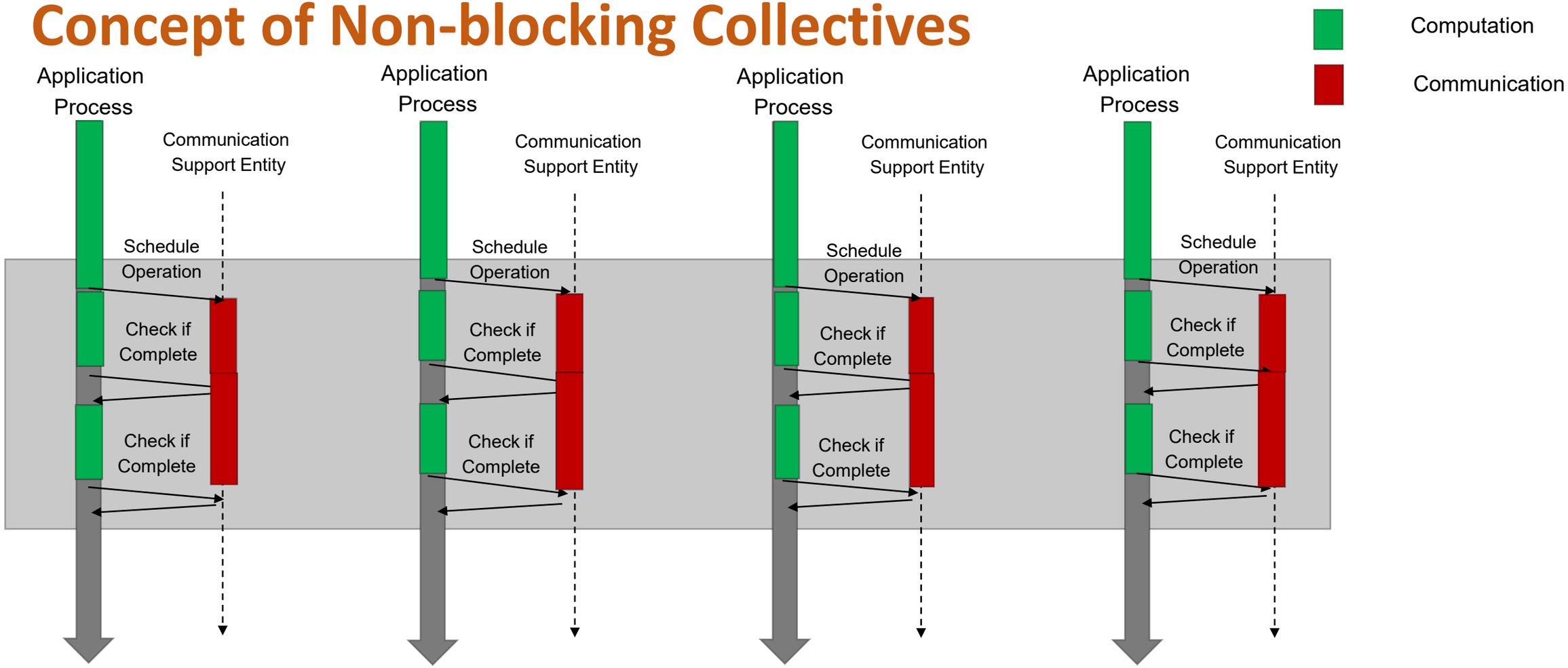


Problems with Blocking Collective Operations



- 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*

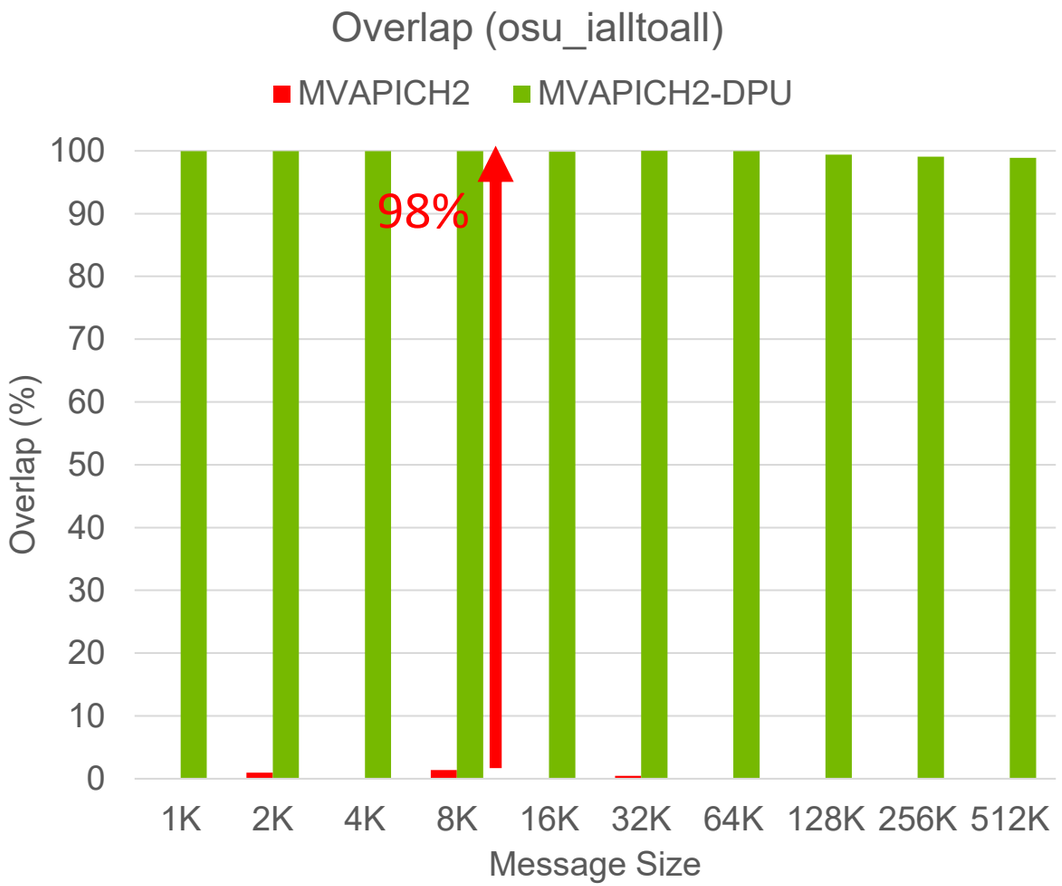
MVAPICH2-DPU Library 2023.10 Release

X-ScaleSolutions

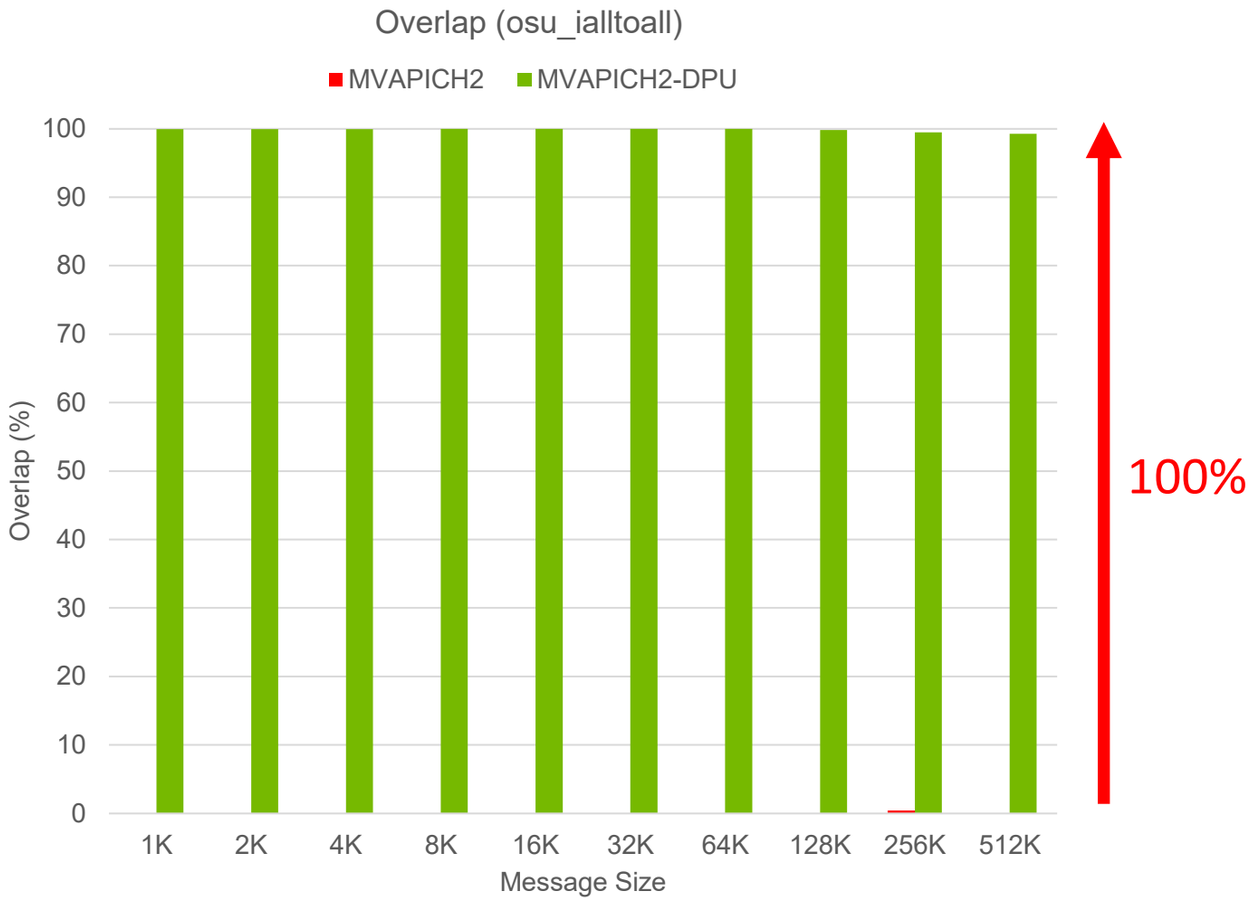
- Supports all features available with the MVAPICH2 2.3.7 release (http://mvapich.cse.ohio-state.edu)
- Novel framework to offload non-blocking collectives to DPU
- Offloads non-blocking Alltoall (MPI_Ialltoall) to DPU
- Offloads non-blocking Broadcast (MPI_Ibcast) to DPU
- Offloads non-blocking Alltoallv (MPI_Ialltoallv) to DPU
- Offloads non-blocking Point-to-Point (MPI_Isend, MPI_Irecv) 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)



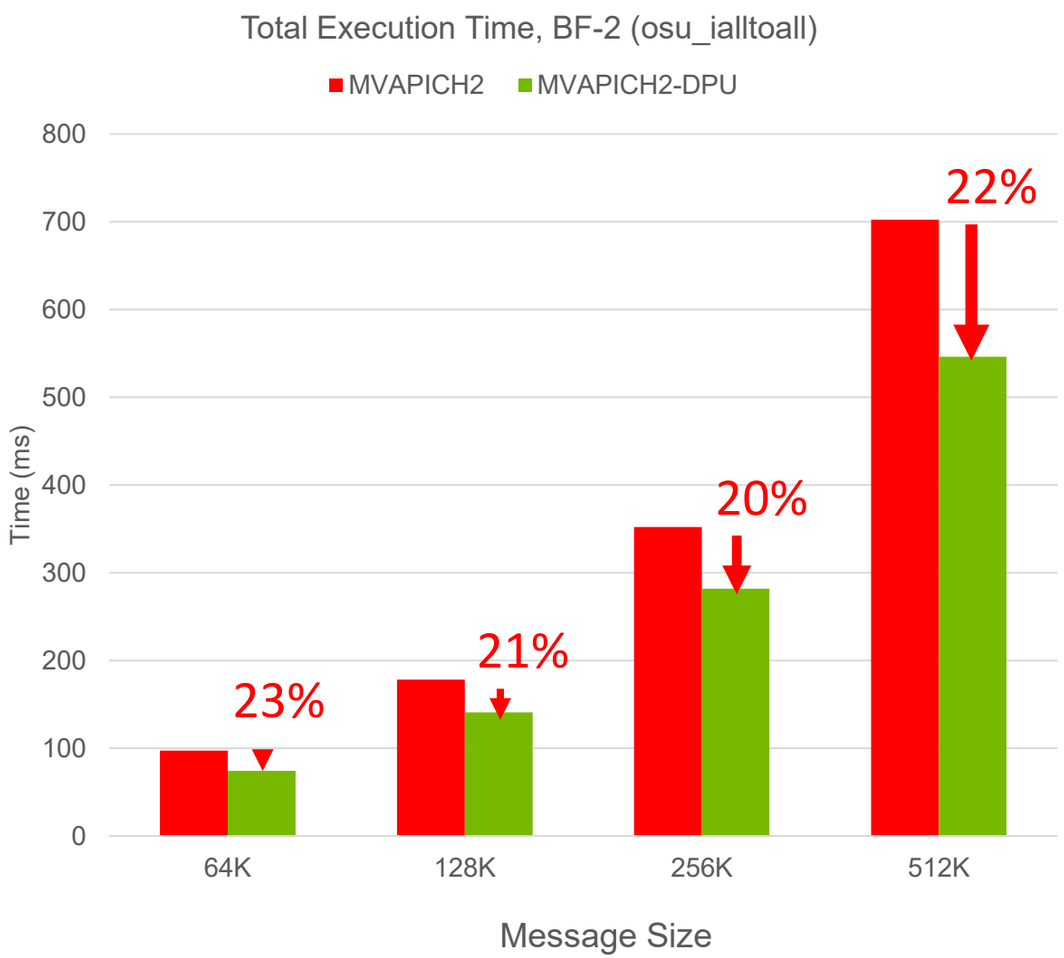
32 Nodes, 16 PPN



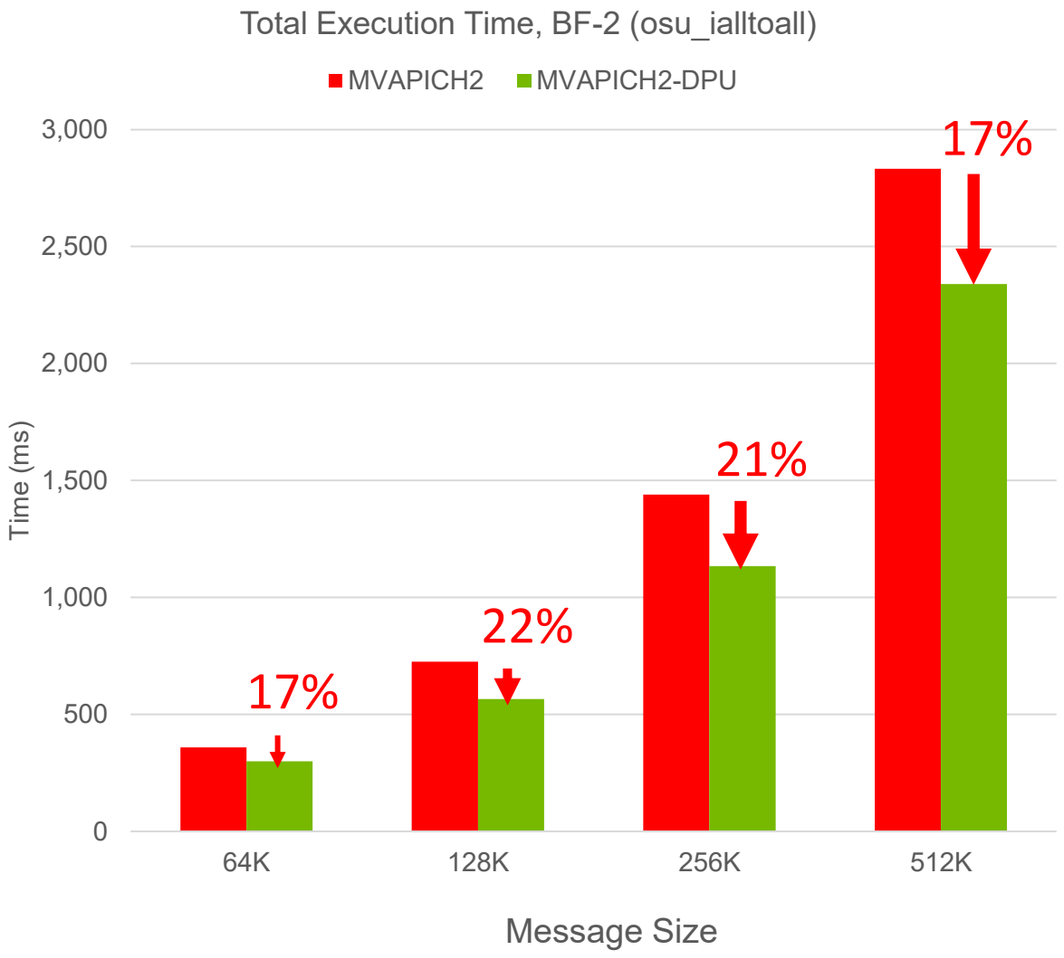
32 Nodes, 32 PPN

Delivers Peak Overlap

Total Execution Time with osu_lalltoall (BF-2, 32 Nodes)



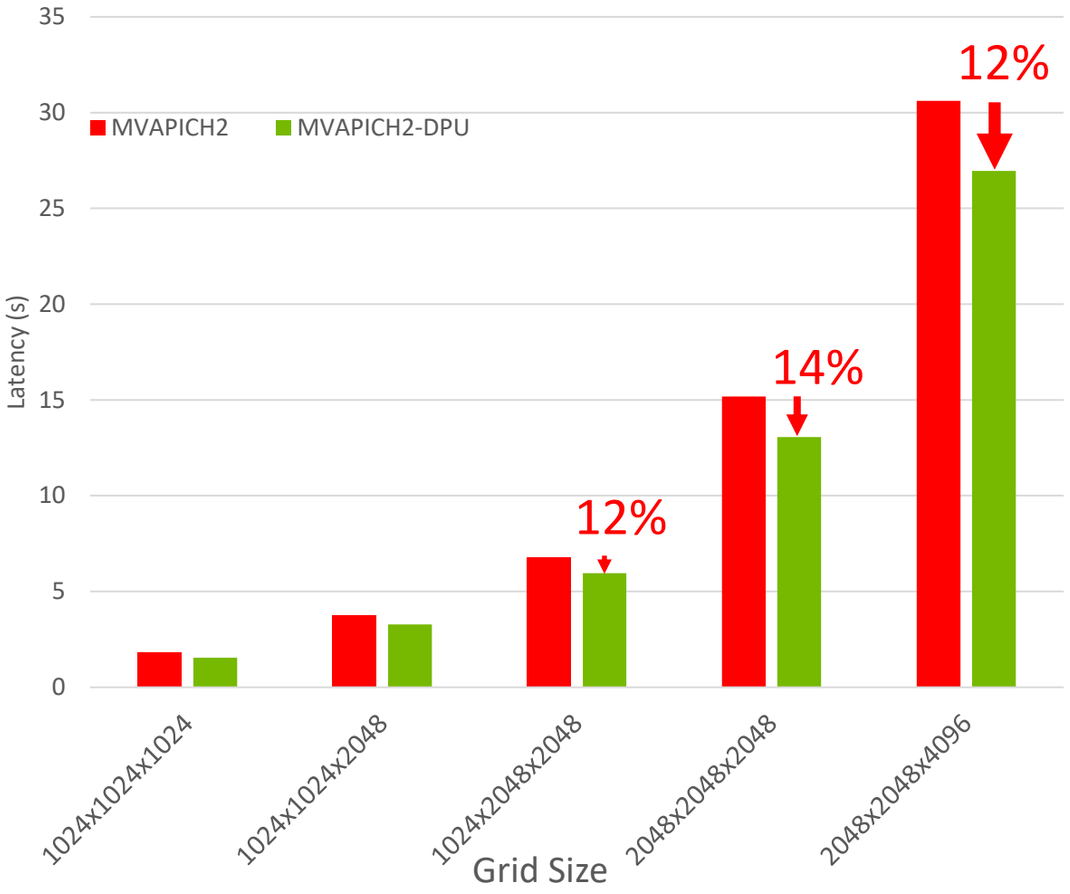
32 Nodes, 16 PPN



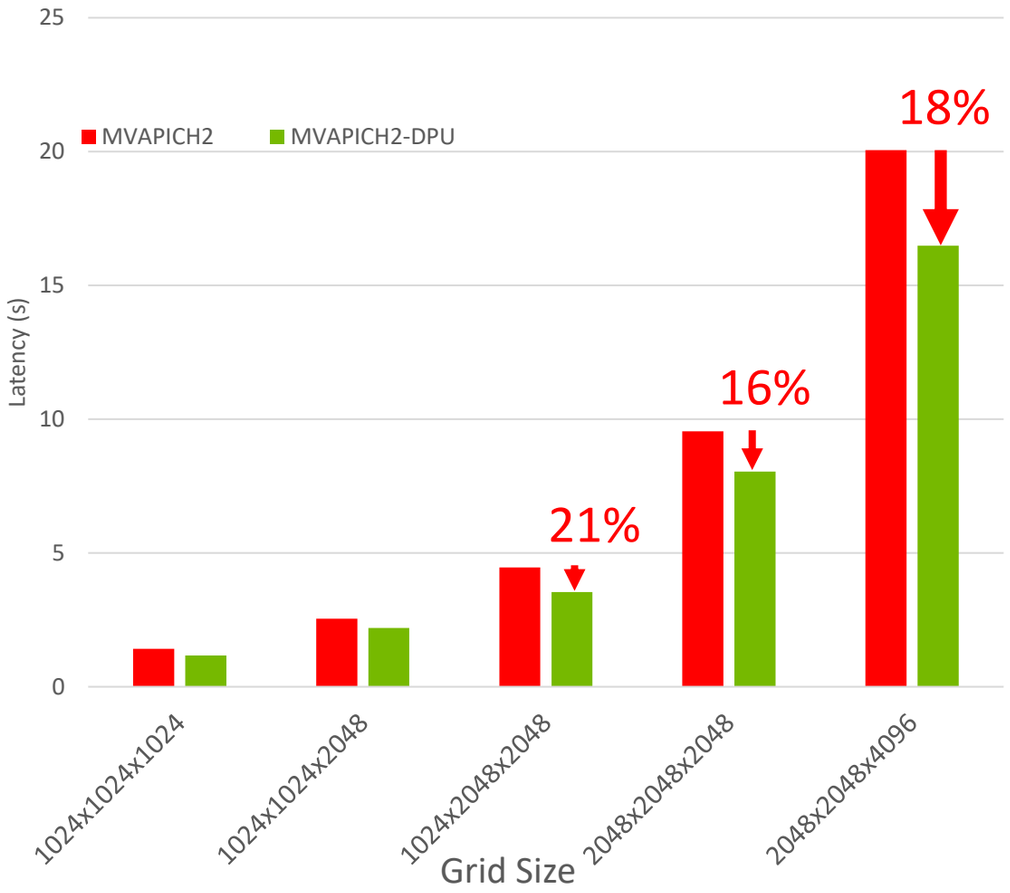
32 Nodes, 32 PPN

Benefits in Total execution time (Compute + Communication)

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



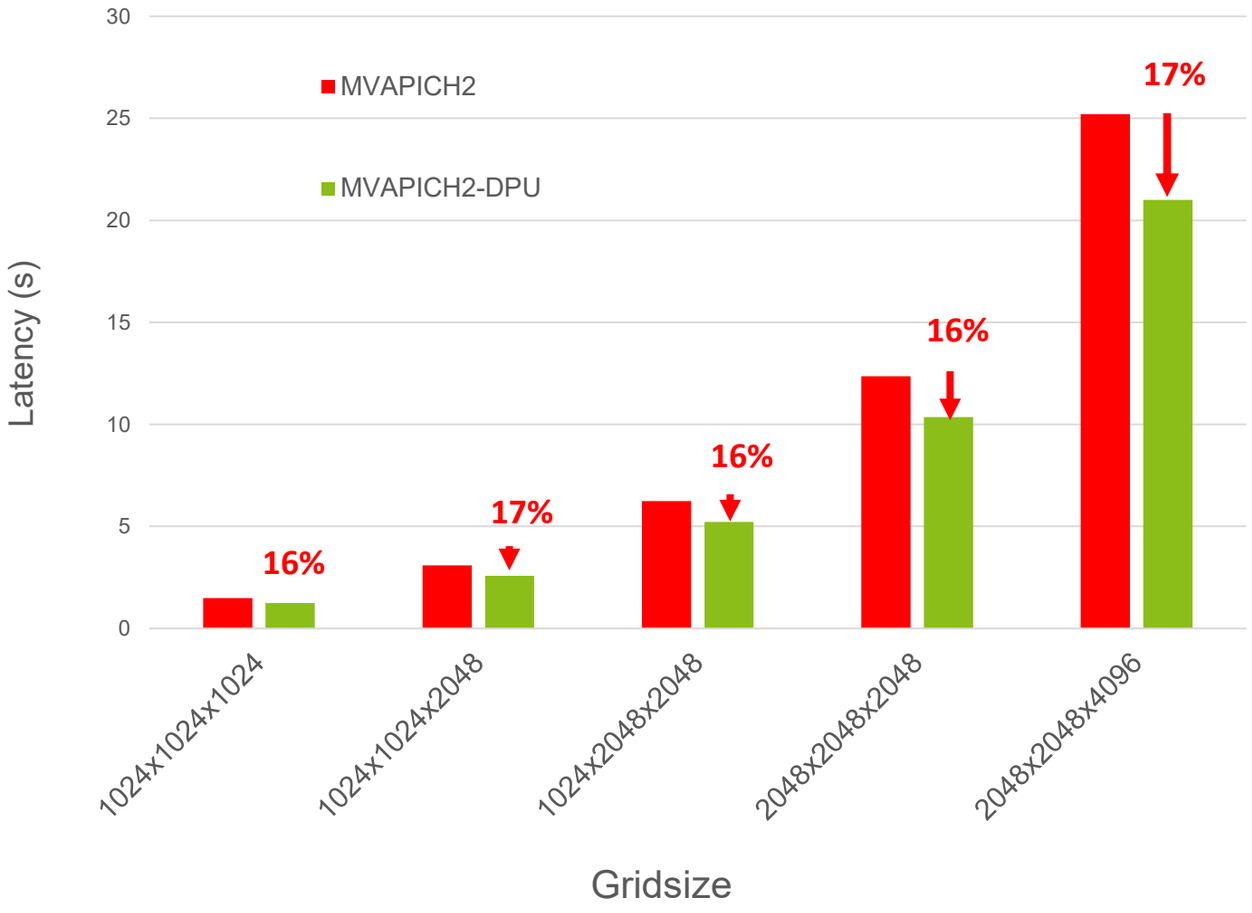
32 Nodes, 16 PPN



32 Nodes, 32 PPN

Benefits in application-level execution time

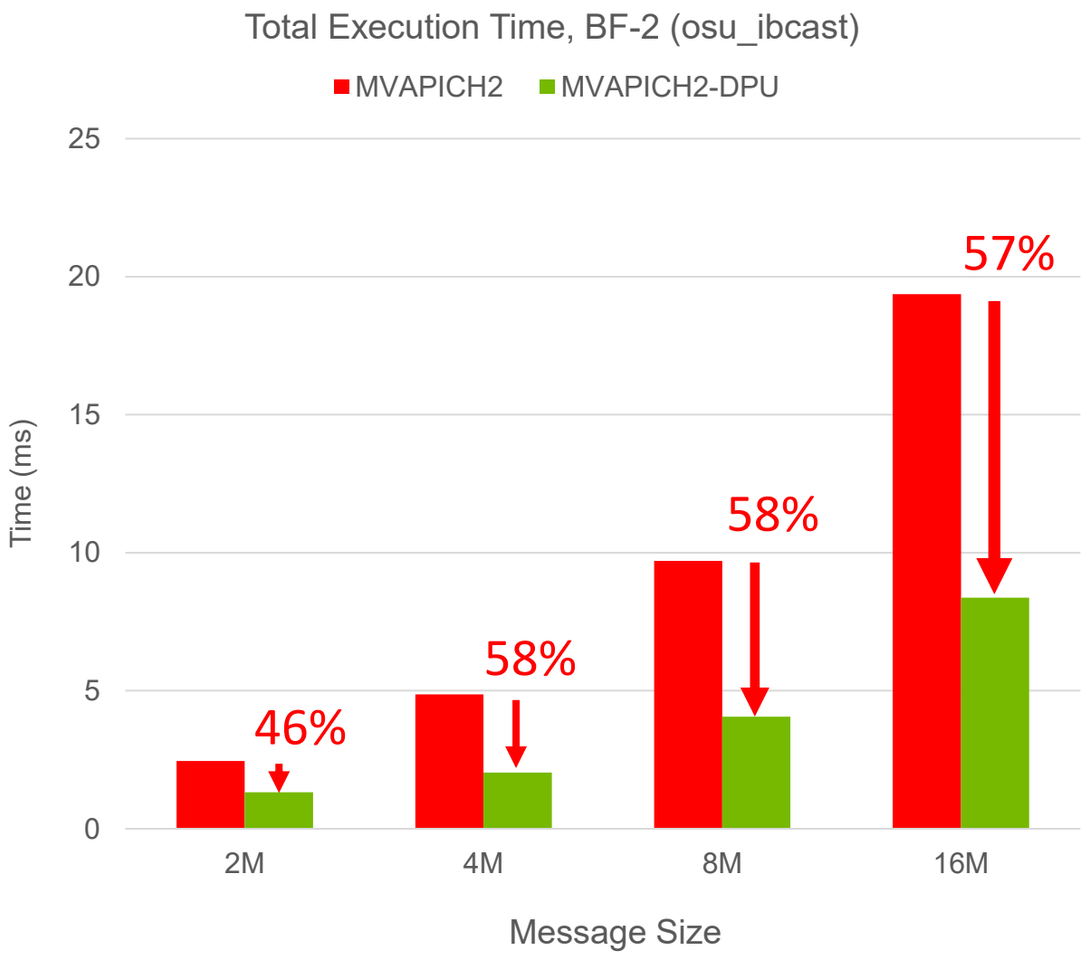
P3DFFT Application Execution Time (BF-3, 16 Nodes)



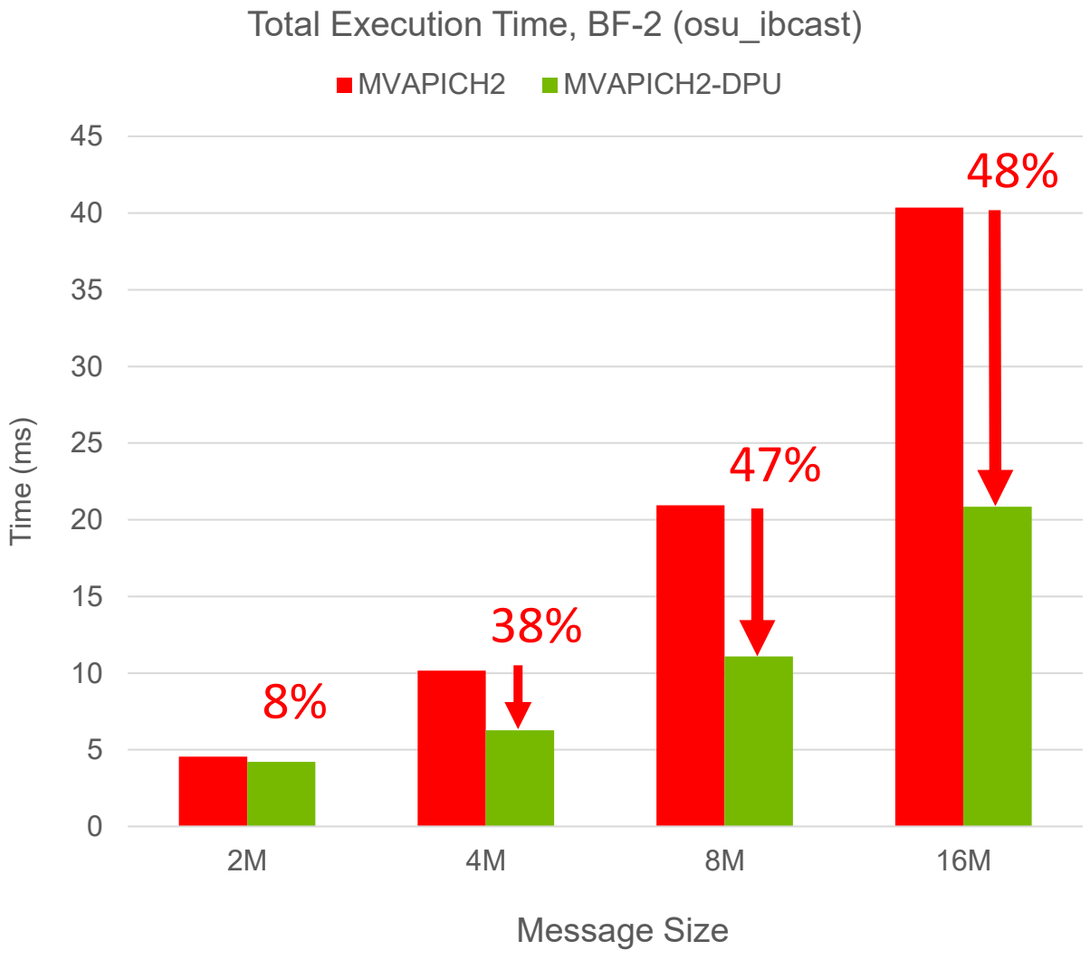
16 Nodes, 32 PPN

Benefits in application-level execution time

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



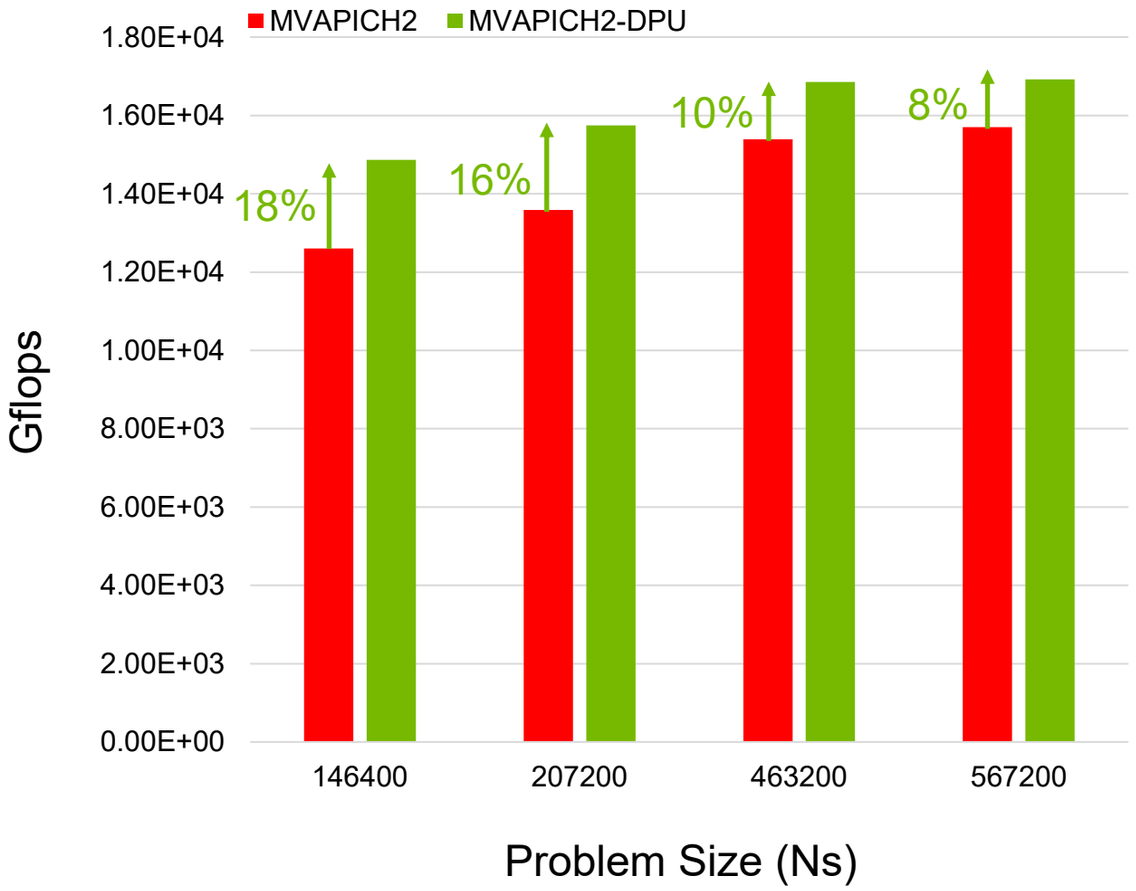
32 Nodes, 1 PPN



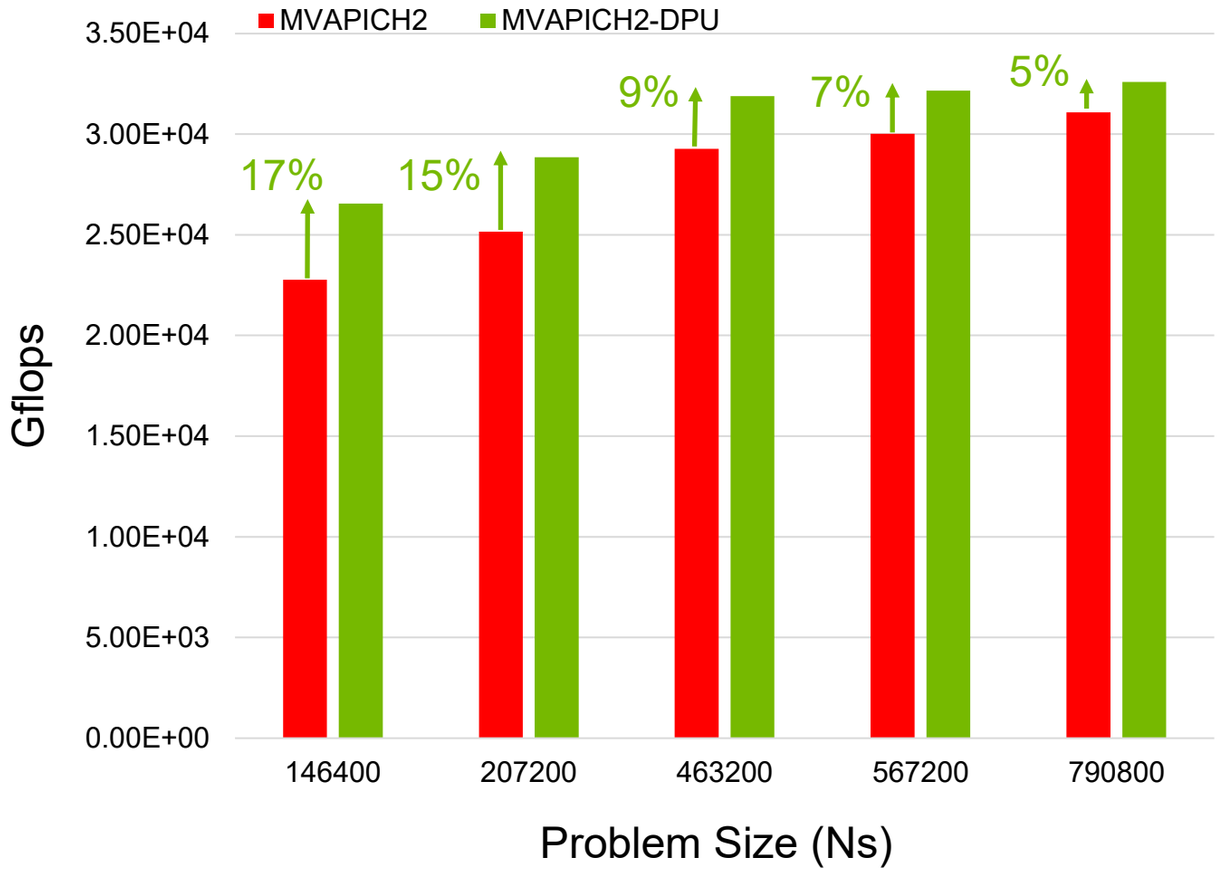
32 Nodes, 16 PPN

Benefits in Total execution time (Compute + Communication)

Accelerating HPL with MVAPICH2-DPU and X-ScaleHPL-DPU (BF-2)



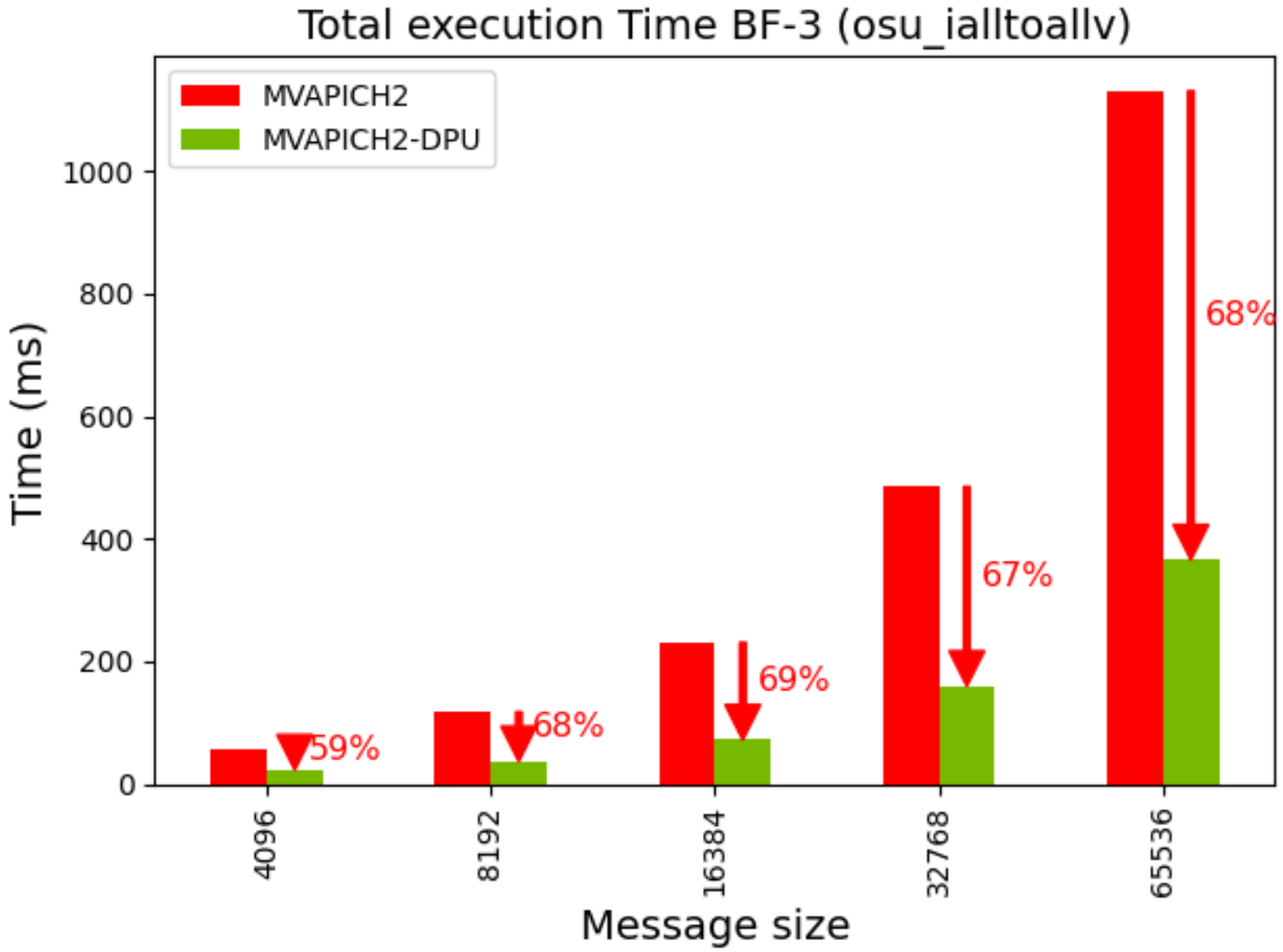
16x32 process grid



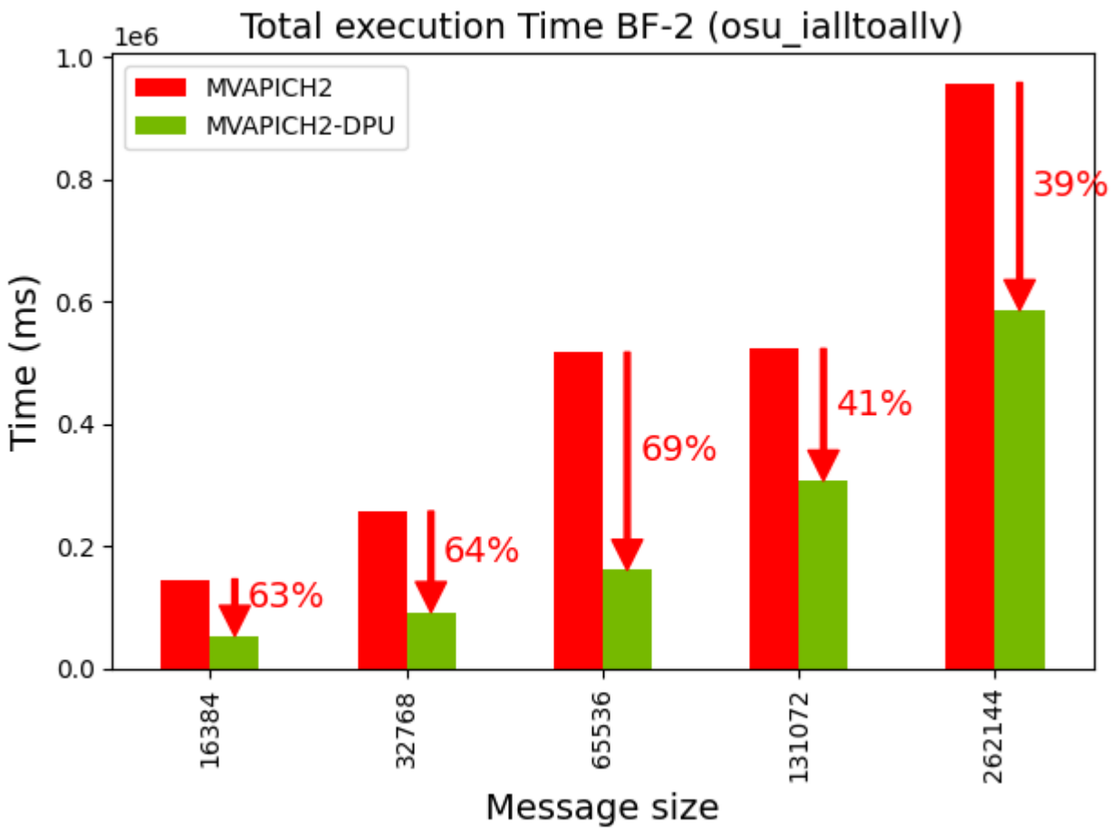
31x32 process grid

Benefits in application-level execution time

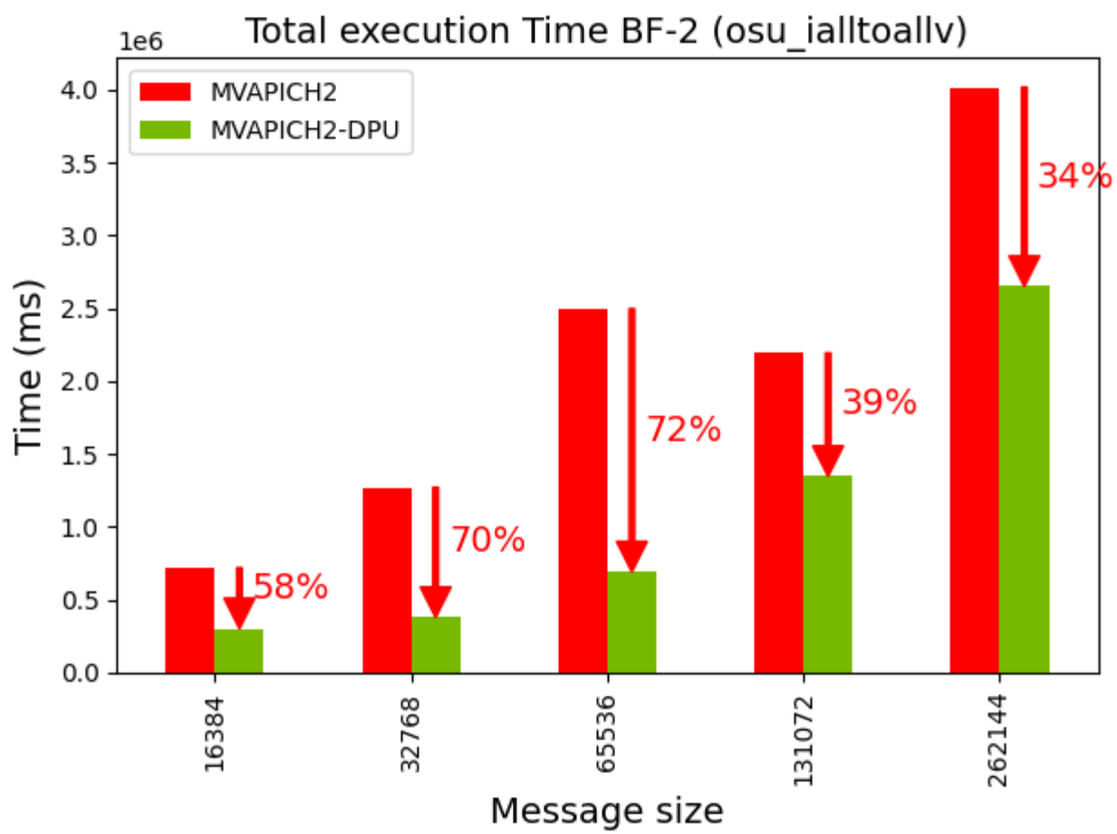
Total Execution Time with osu_Ialltoallv (BF-3, 32 Xeon Nodes, 1K Processes)



Total Execution Time with osu_Ialltoallv (BF-2, 8 AMD EPYC Nodes)

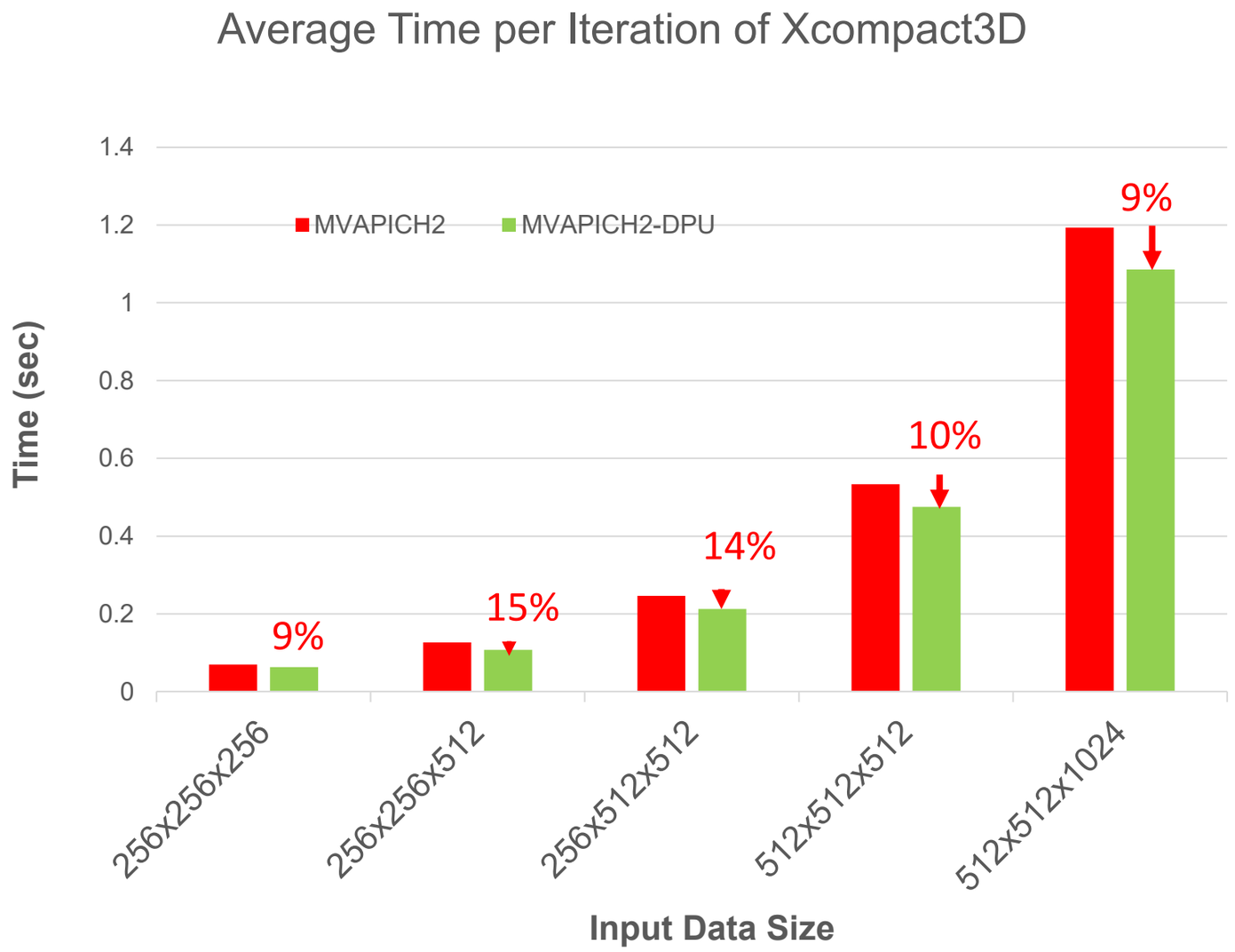


64 PPN



128 PPN

XCompact3D Application Execution Time (8 AMD EPYC Nodes)



AMD EPYC cluster
8 nodes 128 ppn

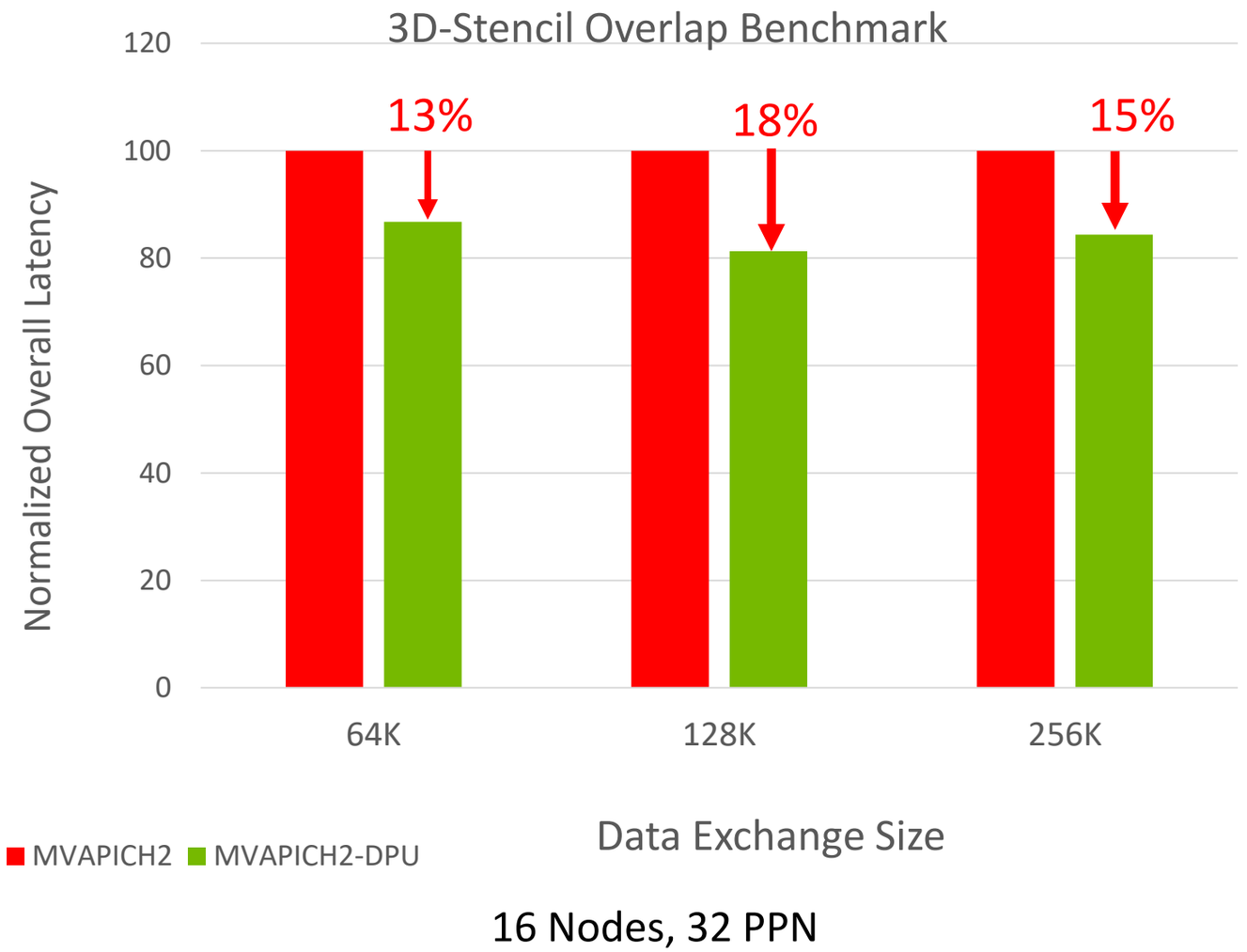
Presentation Outline

- Overview of the MVAPICH Project
- **Offloading Strategies and Benefits:**
  - Non-blocking Collectives (communication)
    - Ialltoall and P3DFFT
    - Ibcast and HPL
    - Ialltoallv and Xcompact3D
  - **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 and I/O)
- Conclusions

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

- Use GVMI to Offload MPI_Isend/MPI_Irecv to the DPU

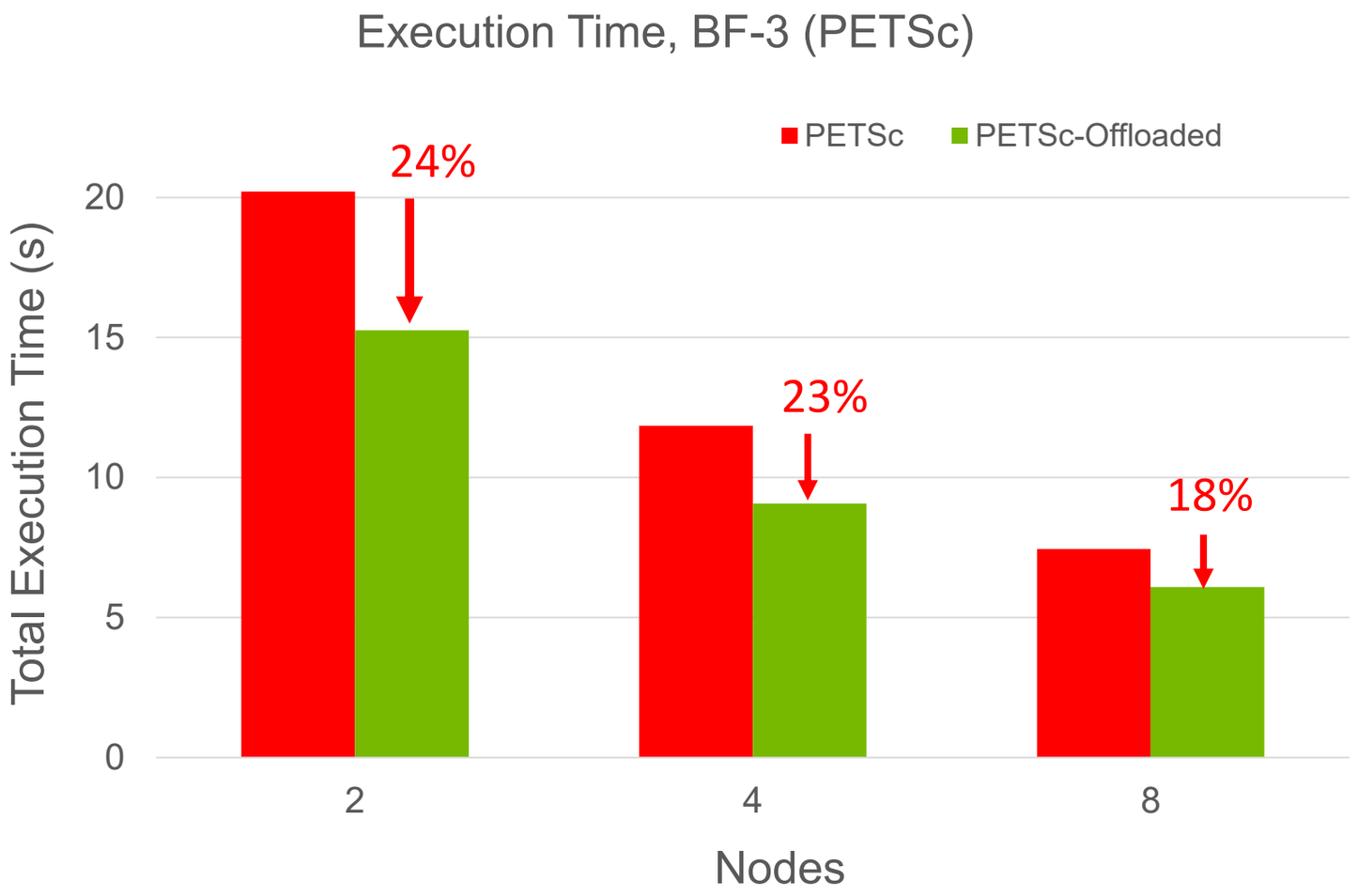
- 3D Stencil Overlap Benchmark :
  - Perform data exchange with 6 peers. (Similar to 7-point stencil)
  - Overlap computation with data-exchange
  - Up to 18% benefits



16 Nodes, 32 PPN

Offloading MPI Point-to-Point and Reduction with PETSc (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
  - Up to 24% benefits



Benefits in Total execution time (Compute + Communication)

Presentation Outline

- Overview of X-ScaleSolutions
- Overview of the MVAPICH Project
- Offloading Strategies and Benefits:
  - Non-blocking Collectives (communication)
    - Ialltoall and P3DFFT
    - Ibcast and HPL
    - Ialltoallv and Xcompact3D
  - 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 and I/O)**
- Conclusions

X-ScaleAI-DPU Package

X-ScaleSolutions

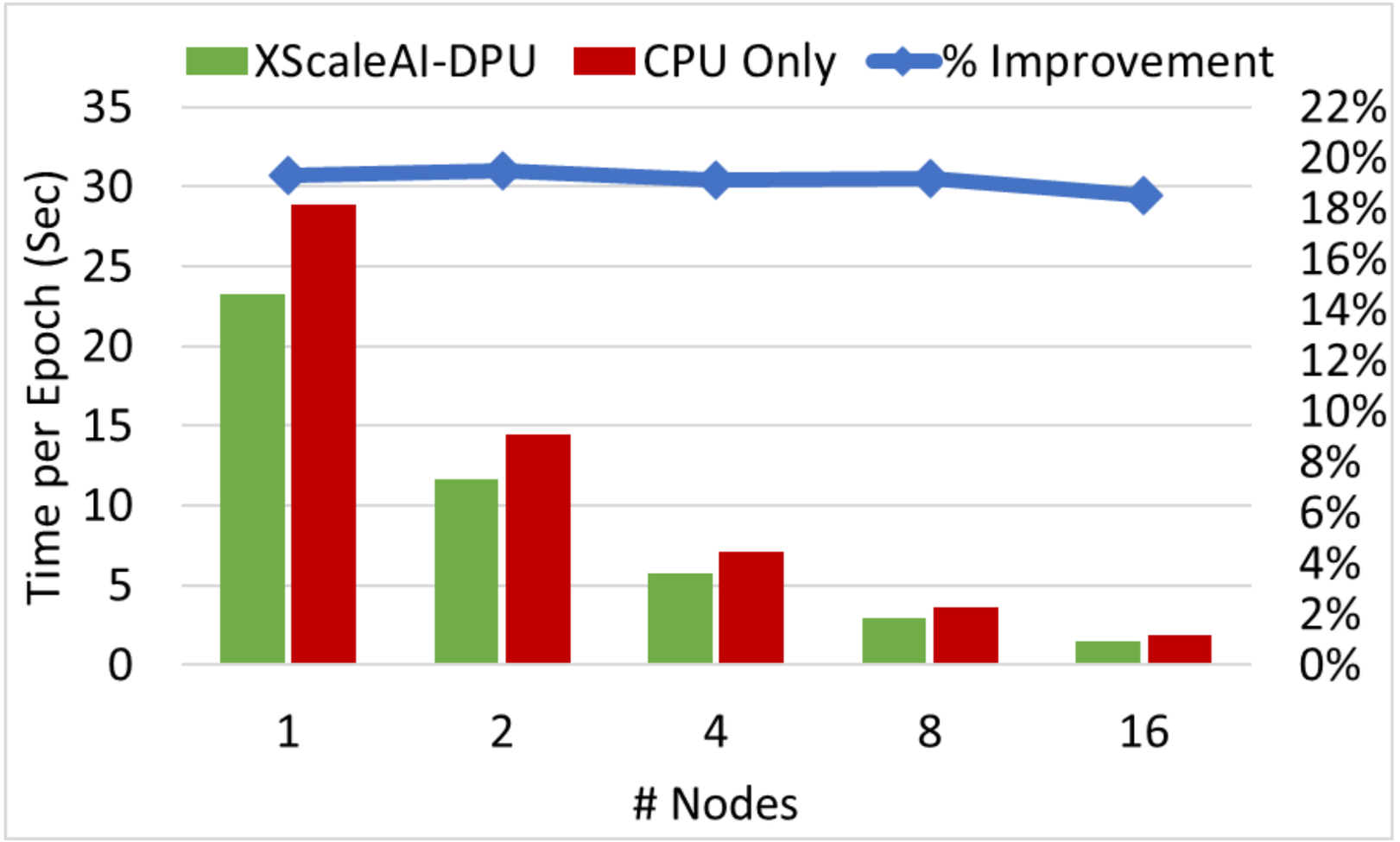
- Accelerating CPU-based DNN training with DPU support
- Based on MVAPICH2 2.3.7 with Horovod 0.25.0
- Supports all features available with the MVAPICH2 2.3.7 release (http://mvapich.cse.ohio-state.edu)
- Supports PyTorch framework for Deep Learning with offloaded checkpointing

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

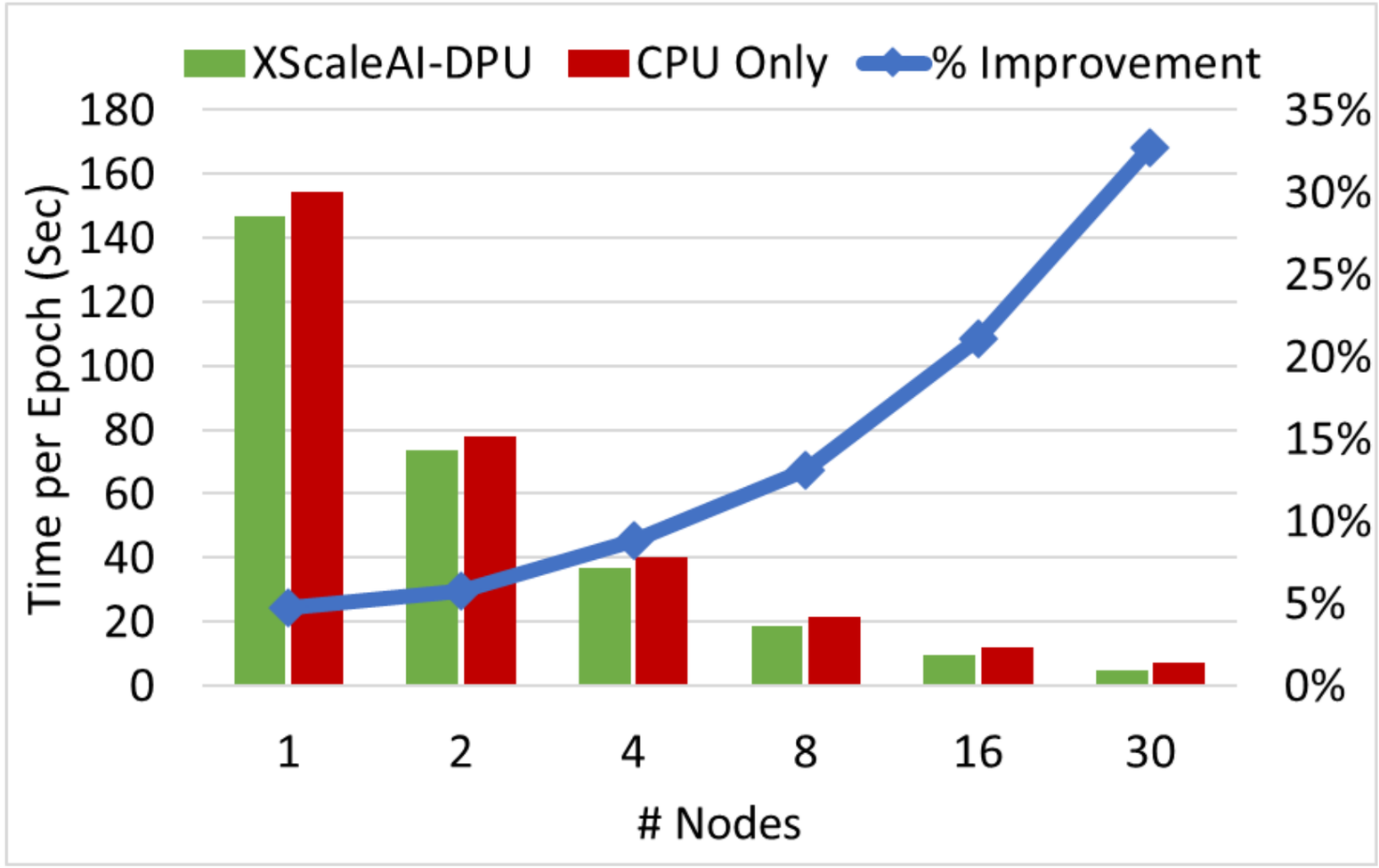
- Two Intel(R) Xeon(R) 16-core CPUs (32 total) E5-2697A V4 @ 2.60 GHz
- NVIDIA BlueField-3 SoC, HDR100 100Gb/s InfiniBand adapters
- Memory: 256GB DDR4 2400MHz RDIMMs per node
- 1TB 7.2K RPM SSD 2.5" hard drive per node
- NVIDIA ConnectX-6 HDR/HDR100 200/100Gb/s InfiniBand adapters with Socket Direct



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

X-ScaleAI-DPU: Checkpointing Offload for DNN Training

- New X-ScaleAI-DPU feature: offload DNN checkpointing during training to the DPU.
- Up to 33% improvement in epoch time on the ResNet-34 model using X-ScaleAI-DPU compared to CPU only.
- Improvement percentage using X-ScaleAI-DPU for checkpointing increases as number of nodes increases.
- Improvement observed across different DL models.



Performance improvement for checkpointing using X-ScaleAI-DPU over CPU-only training on the ResNet-34 model on the CIFAR10 dataset

Conclusions

- DPU technology provides novel ways to offload computation, communication, and I/O from host CPUs to DPU cores
- Demonstrated two ways to take advantage of the DPU technology to accelerate MPI and Deep Learning applications
- Promises potential for accelerating application performance further
- X-ScaleSolutions will be happy to get engaged with collaborators

Thank You!

Donglai Dai

contactus@x-scalesolutions.com

X-ScaleSolutions

http://x-scalesolutions.com/