Accelerating HPC and AI Applications using Novel Products from X-ScaleSolutions

Kyle Schaefer, k.schaefer@x-scalesolutions.com

http://x-scalesolutions.com

X-ScaleSolutions

Overview of X-ScaleSolutions

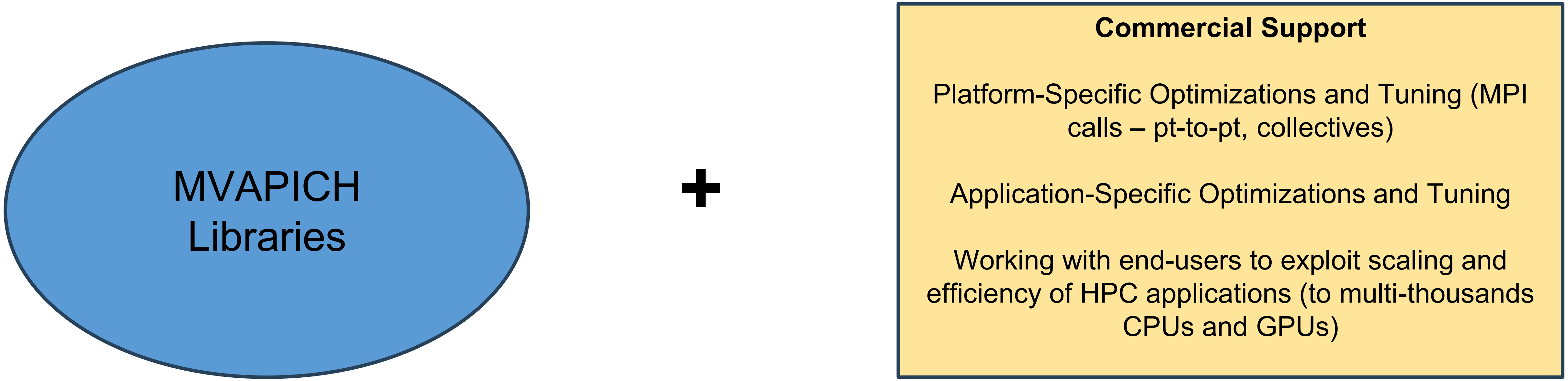
- Based on HPC and AI expertise for 30+ years
- Bring innovative and efficient end-to-end solutions, services, support, and training to HPC, AI, and Big Data customers
- Business Model:
  - Commercial Support (Optimization, tuning, and training) for the state-of-the-art communication libraries, designed and developed from the Ohio State University (OSU)
    - High-Performance and Scalable MVAPICH Library and its families
    - High-Performance Deep Learning/Machine Learning Libraries
    - High-Performance Big Data Libraries
  - Value-Added and New Products from X-ScaleSolutions
    - Licensing, commercial support and training

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-ScalePETSC: Accelerating PETSC Library (a common library for many scientific workloads) on clusters with CPUs and GPUs
- X-ScaleSecured-MPI: High-Performance MPI library with built-in security
- X-Scale Monitor: HPC/AI hardware monitoring

X-ScaleHPC: Features

- Application-aware and communication-focused optimization for common HPC applications



MVAPICH Libraries

+

Commercial Support

Platform-Specific Optimizations and Tuning (MPI calls – pt-to-pt, collectives)

Application-Specific Optimizations and Tuning

Working with end-users to exploit scaling and efficiency of HPC applications (to multi-thousands CPUs and GPUs)

X-ScaleHPC: Value Propositions

- User-level software
  - Can be installed by any user
  - Can be installed by a system administration and make it available as a module
- Vendor (CPU/GPU/Interconnect) Neutral Stack
- Performance portability across different platforms
- Continuous and sustained performance gain from next-generation hardware
- End benefits:
  - Reducing time-to-solution
  - Higher throughput on a given platform from multiple simulations
  - Running multiple simulations using less hardware
  - Reduction in TCO
  - Reduction in power usage (with reduced execution time) and carbon footprint

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-ScalePETSC: Accelerating PETSC Library (a common library for many scientific workloads) on clusters with CPUs and GPUs
- X-ScaleSecured-MPI: High-Performance MPI library with built-in security
- X-Scale Monitor: HPC/AI hardware monitoring

X-ScaleAI: Features and Capabilities

Goal

- High-performance solution for AI problems on modern HPC and Cloud platforms (supports MPI-driven approach)
  - Pre-Training, Inference, Fine-Tuning

Major Features

- End-to-end optimized software stack via container deployment
  - Bakes in all scaling and systems optimizations developed under the HiDL project
  - AWS cloud (AMI), apptainer for on-premise systems
- Supports models defined in PyTorch or HuggingFace
  - Large Language Models (LLMs)
    - E.g. Llama-3, OLMo, Pythia and BERT
  - Vision Models
    - E.g. ResNet, U-Net, ViT, Stable Diffusion
- Scalable model checkpoint and restart support for long-running training and fine-tuning applications
- “Out of the box” optimal performance for on-premise and cloud-based systems containing:
  - CPUs (x86, ARM)
  - GPUs (NVIDIA, AMD, and Intel)
  - Interconnects (EFA, InfiniBand, Ethernet, RoCE, Slingshot, and Omni-Path)

X-ScaleAI: Features and Capabilities (Cont’d)

- X-ScaleAI is a product *and* service
  - Supports public/private models with private/public data
  - Contains sample recipes to get started
  - On-boarding scheme is available (as a service) for new users and organizations

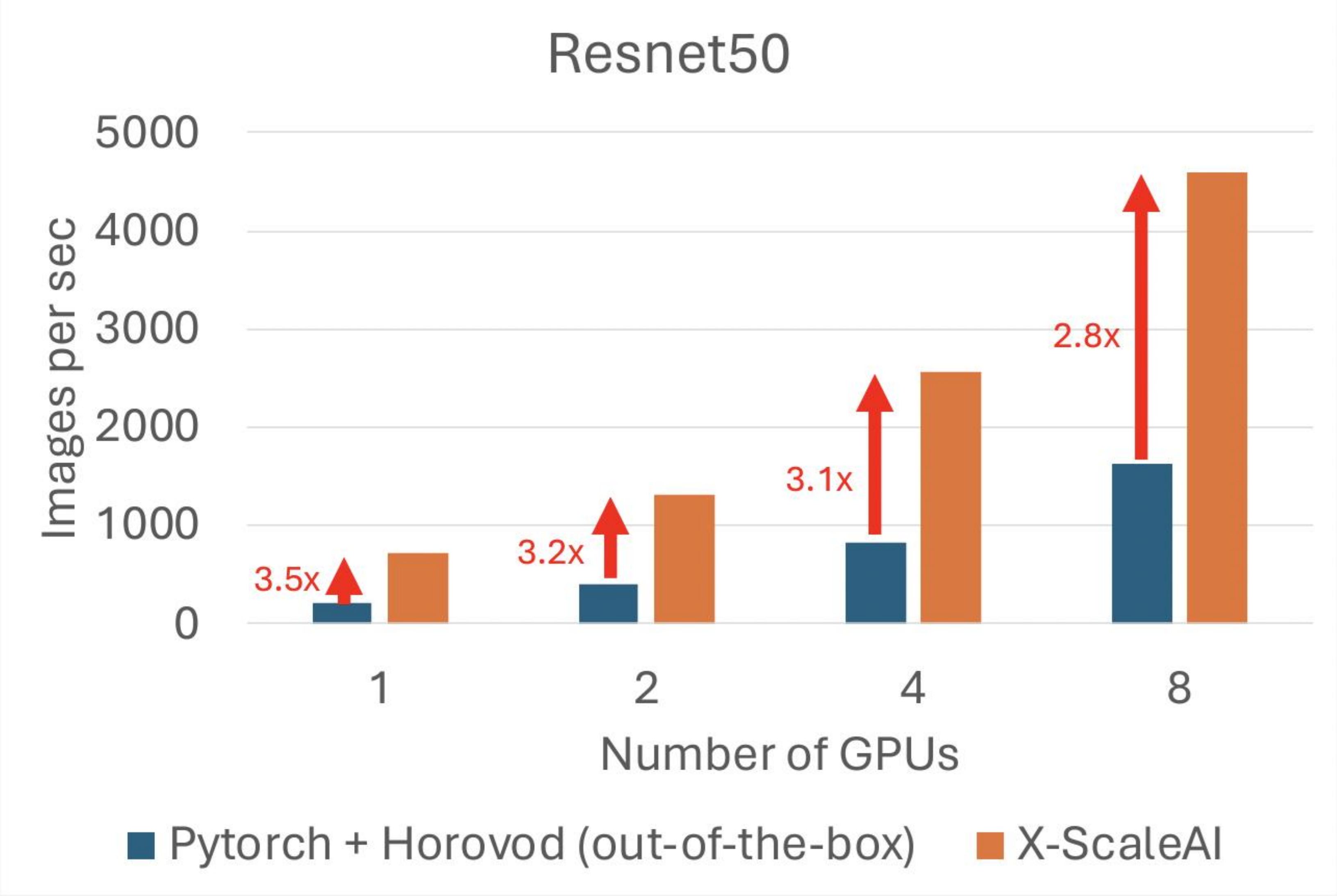
- Baked-in product support and team expertise across a wide range of use cases
  - Various language modeling tasks, Healthcare imaging, etc

X-ScaleAI: Value Propositions

- Reduction in Distributed Training, Fine-Tuning, and Inference time on a given hardware platform
  - CPU, GPU, and Interconnect
- Vendor (CPU/GPU/Interconnect) neutral stack
- Performance portability across different platforms
- Continuous and sustained performance gain from next-generation hardware
- End benefits:
  - Reducing time-to-solution
  - Higher throughput on a given platform from multiple AI applications
  - Reduction in power usage (with reduced training/inference time) and carbon footprint
  - Reduction in resource capacity to get similar or better performance
  - Helps with aiming for lower capacity of resources (CPUs and GPUs) for future deployments
  - Reduction in TCO and helps with investment multiplier

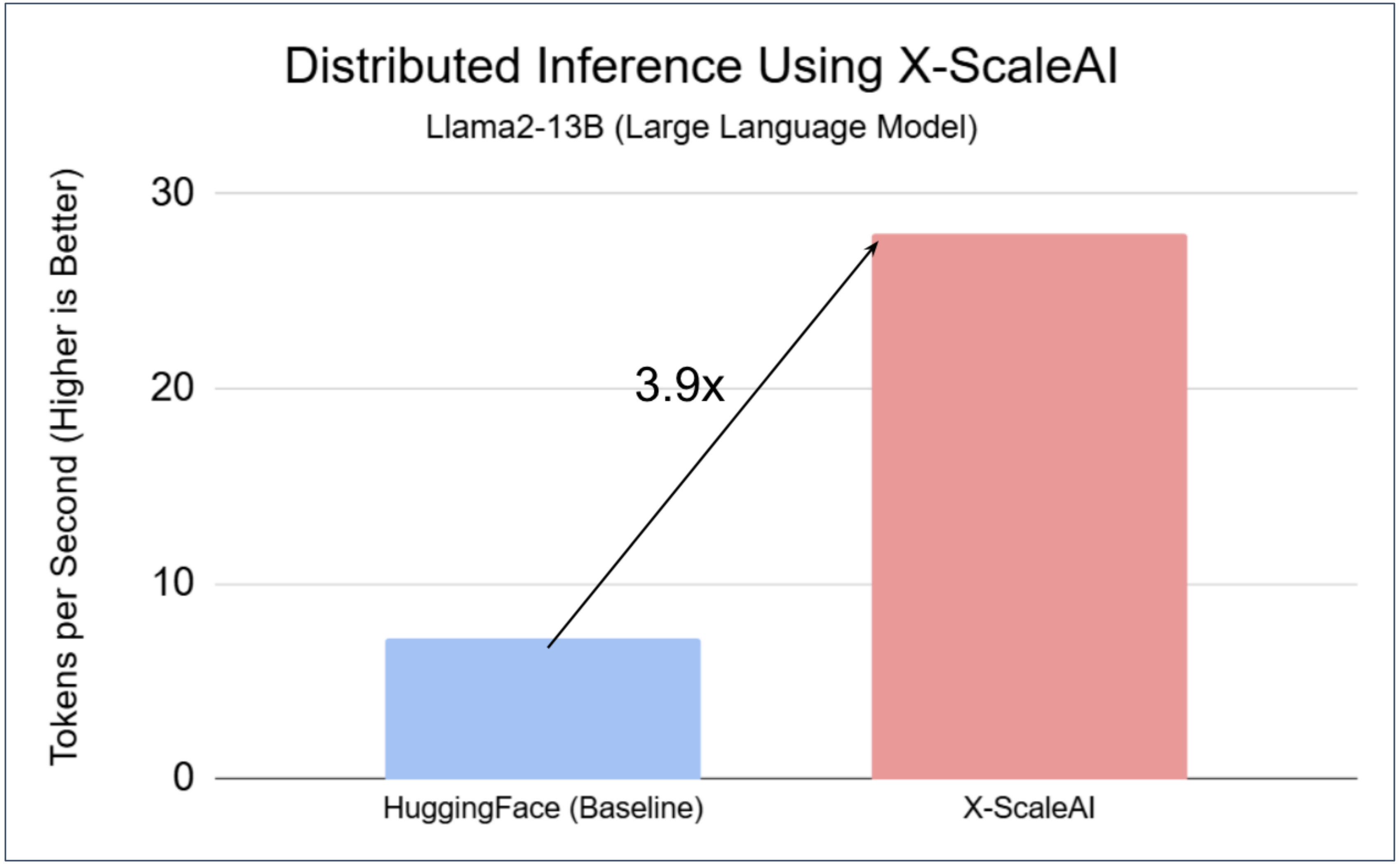
X-ScaleAI: Distributed PyTorch on Sample System #1

- Image classification
  - ResNet50
- On-premise:
  - Frontera (TACC)
- GPU:
  - NVIDIA Quadro RTX 5000
- Interconnect:
  - HDR100 - 100 Gb/s InfiniBand



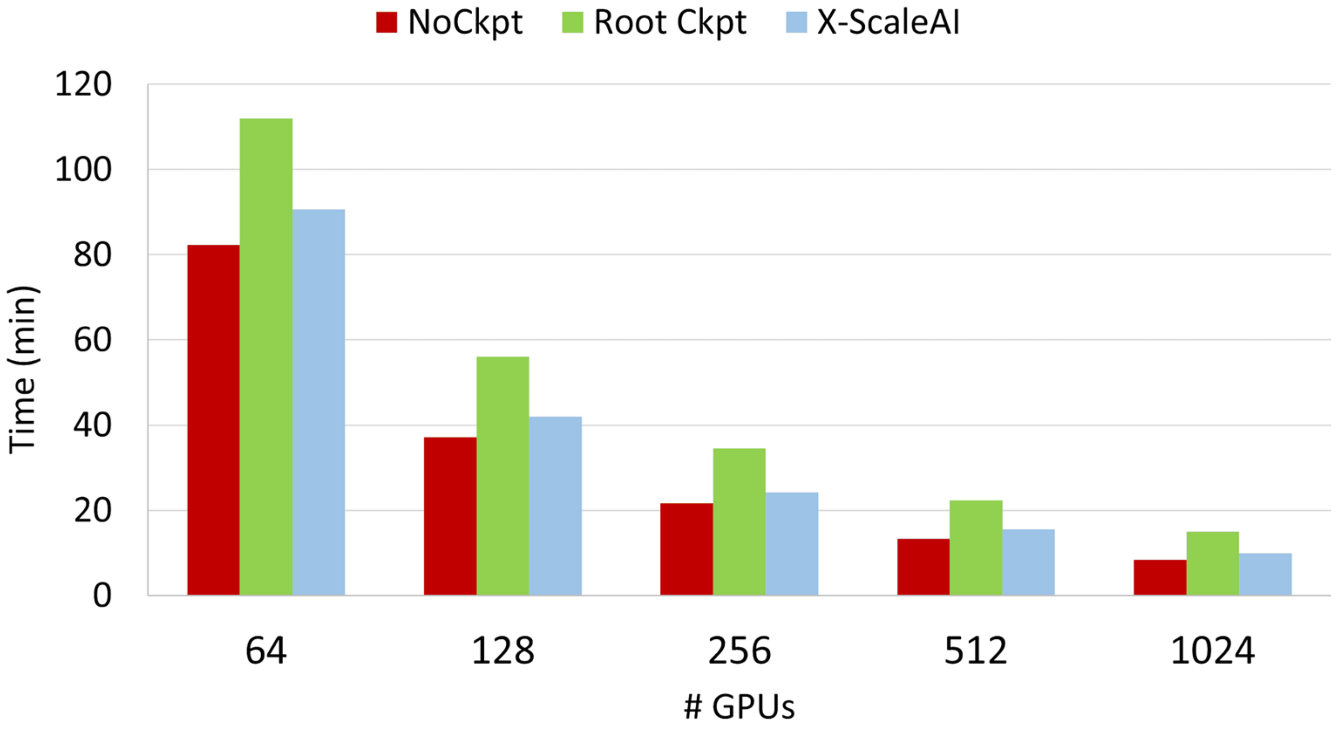
X-ScaleAI: LLM Inference on AWS

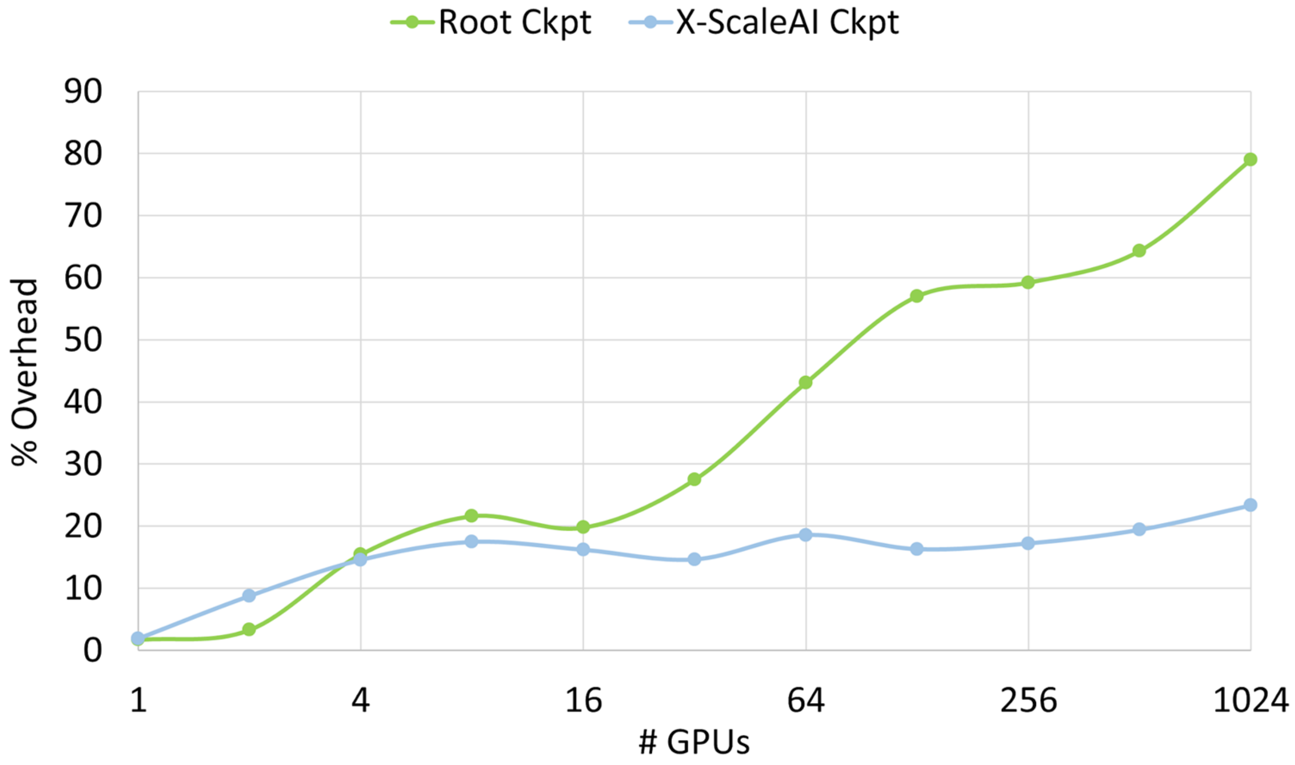
- Text Generation
  - Llama2-13B
- Cloud:
  - AWS
  - g4dn.12xlarge
- GPU:
  - 4x(NVIDIA T4)
  - 16GB VRAM/GPU
- Interconnect:
  - PCIe



Scalable Checkpoint-Restart for DL Applications

- We take the end-to-end training time of 100 epochs of EDSR training with X-ScaleAI
  - Competing frameworks save the checkpoint to the PFS on the root rank (Root Checkpoint)
  - X-ScaleAI has every rank save checkpoints to the local NVMe, and overlaps PFS writes with training
  - Greatly reduces checkpointing overhead at scale, and improves fault-tolerance





AWS Marketplace Availability

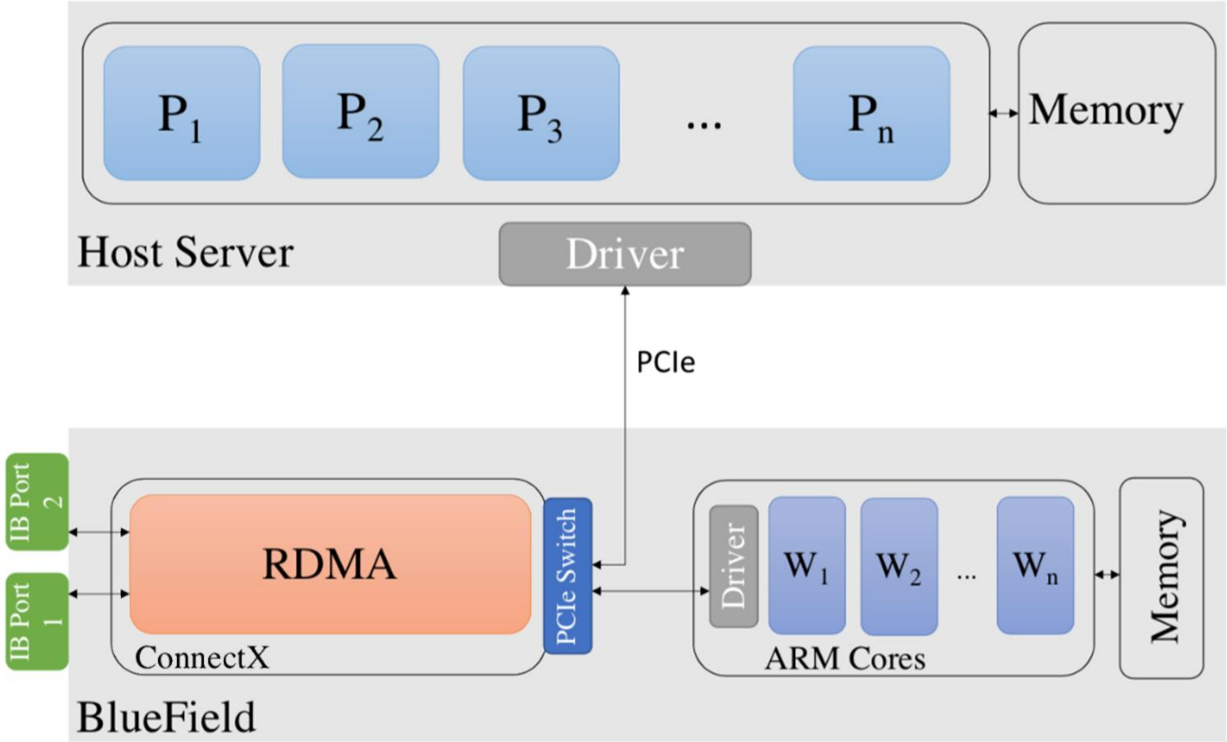
- X-Scale-AI product is available through AWS Marketplace
  - https://aws.amazon.com/marketplace/seller-profile?id=seller-2xz74f2owixfm
- Subscribe X-Scale-AI with a pay as you go or with a trial version

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-ScalePETSC: Accelerating PETSC Library (a common library for many scientific workloads) on clusters with CPUs and GPUs
- X-ScaleSecured-MPI: High-Performance MPI library with built-in security
- X-Scale Monitor: HPC/AI hardware monitoring

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

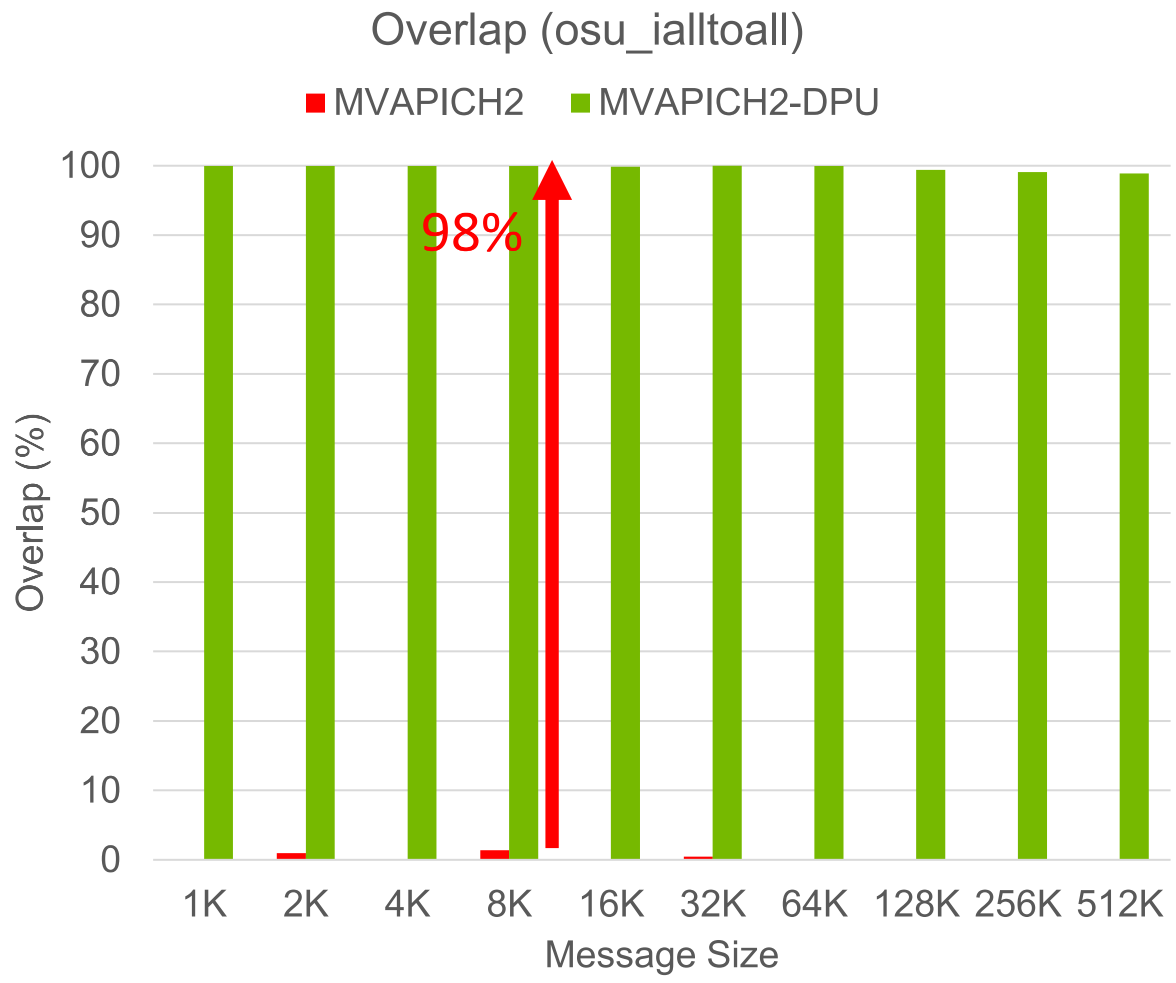


MVAPICH2-DPU Library Release

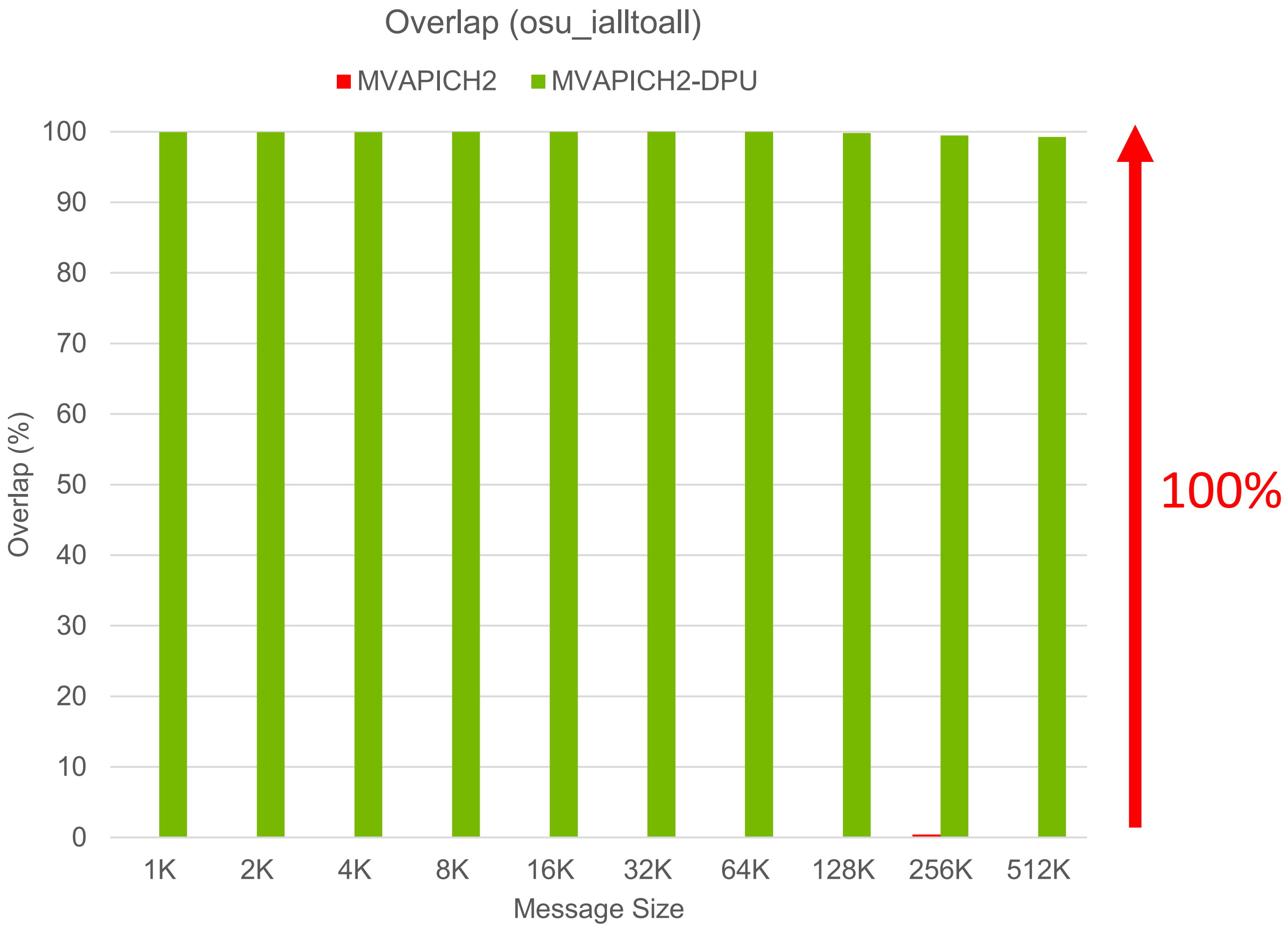
X-ScaleSolutions

- 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/v (MPI_Ialltoall/v) to DPU
- Offloads non/blocking point-to-point to the DPU
- Offloads non-blocking Broadcast (MPI_Ibcast) to DPU

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



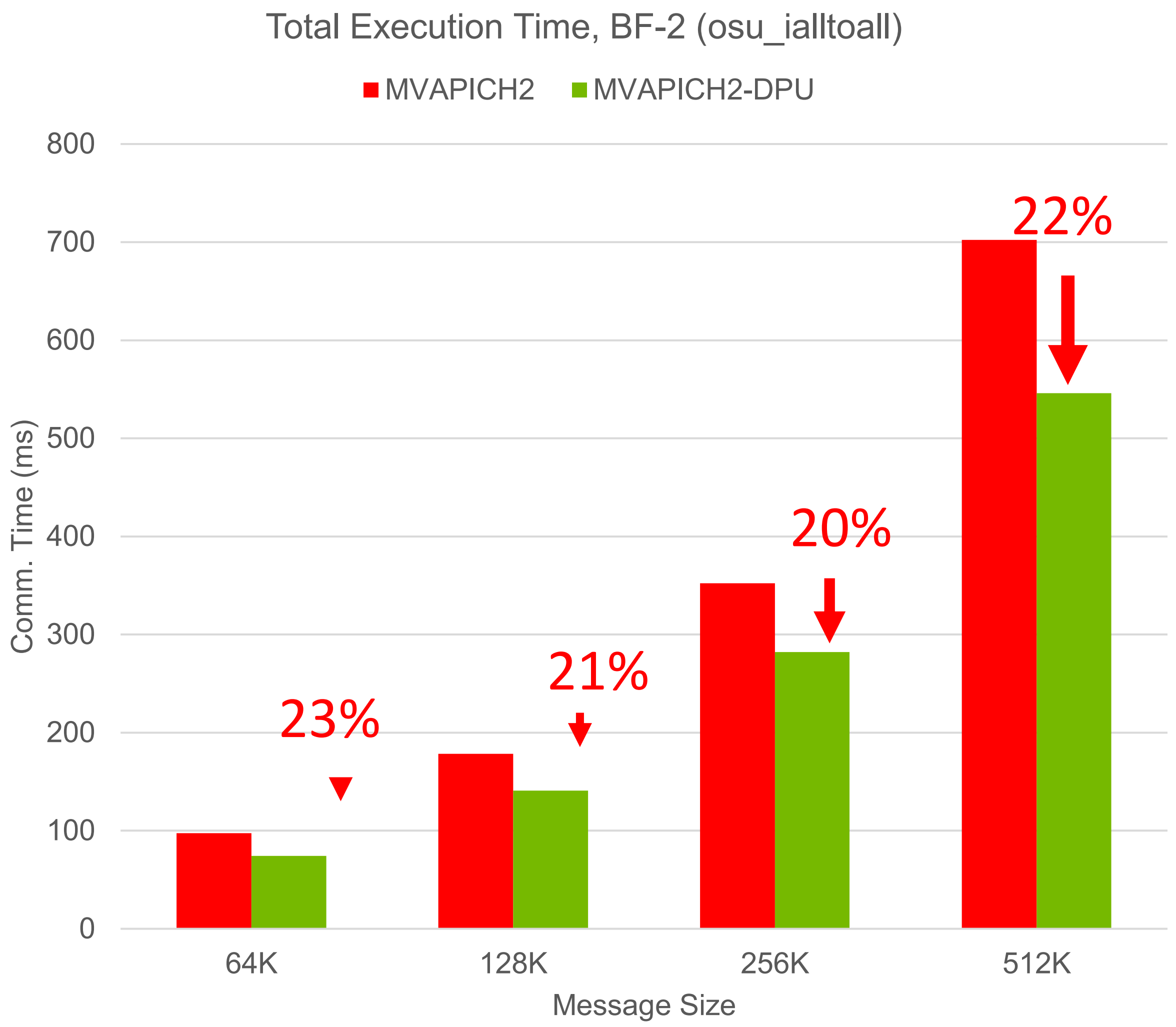
32 Nodes, 16 PPN



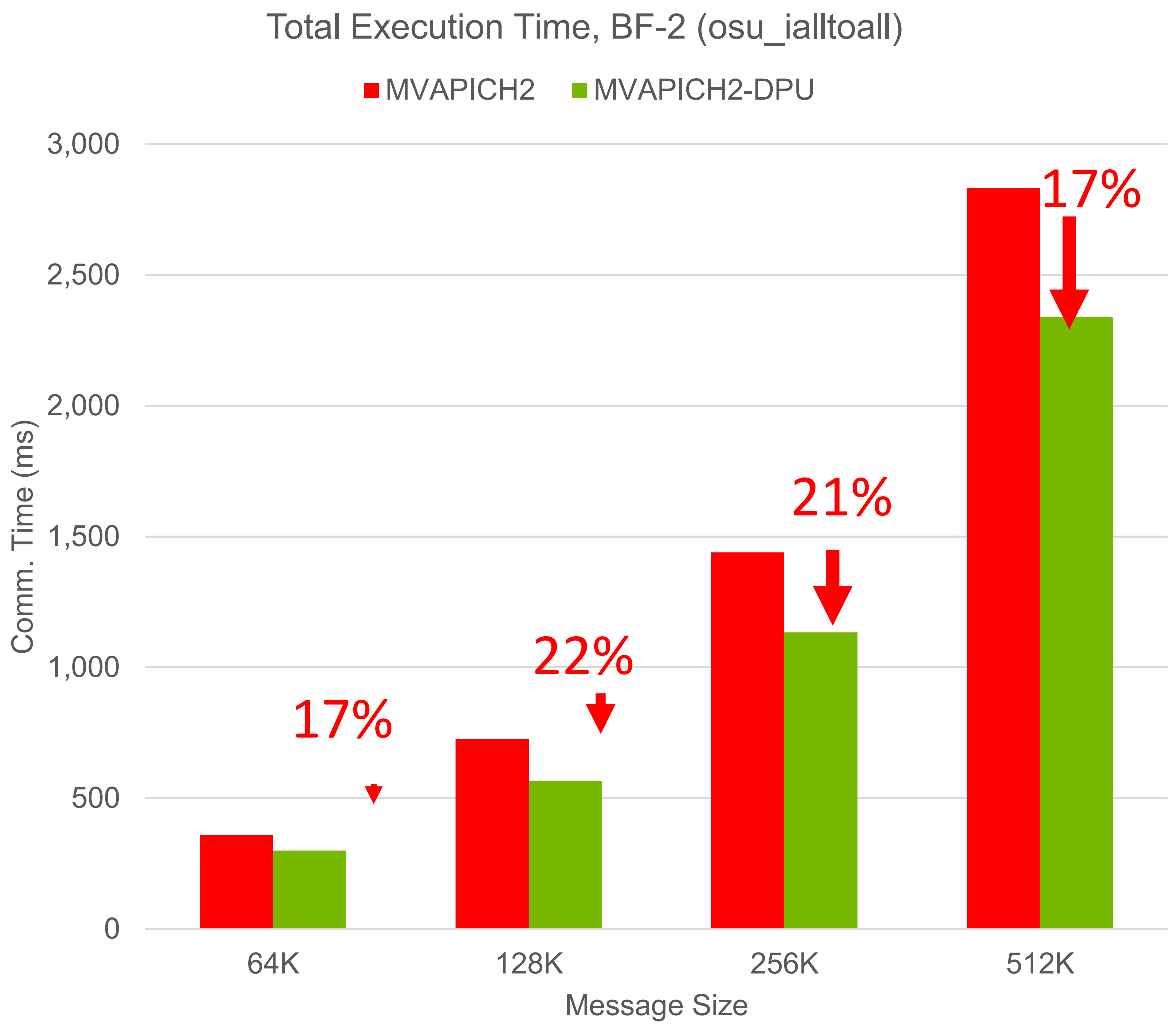
32 Nodes, 32 PPN

Delivers Peak Overlap

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



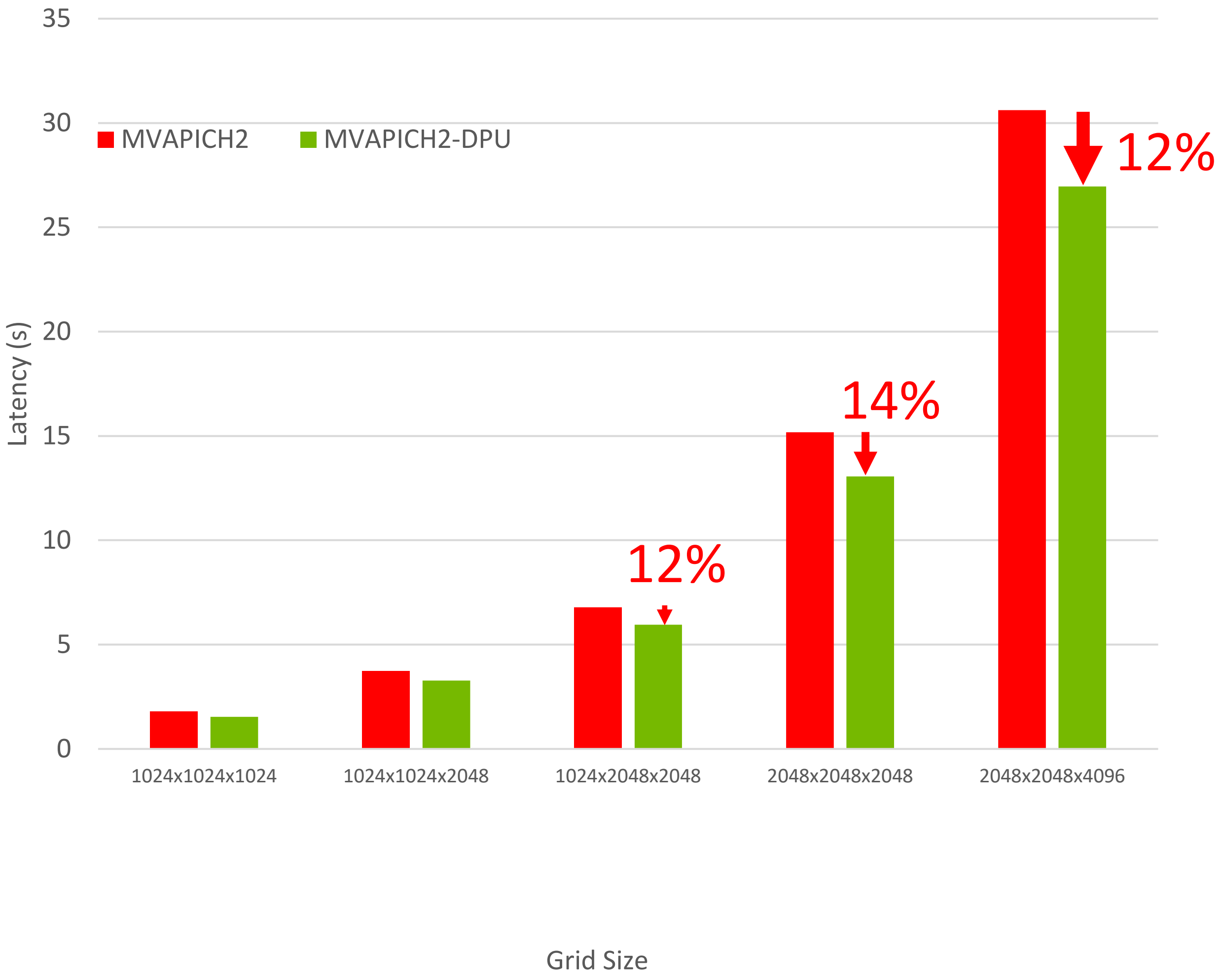
32 Nodes, 16 PPN



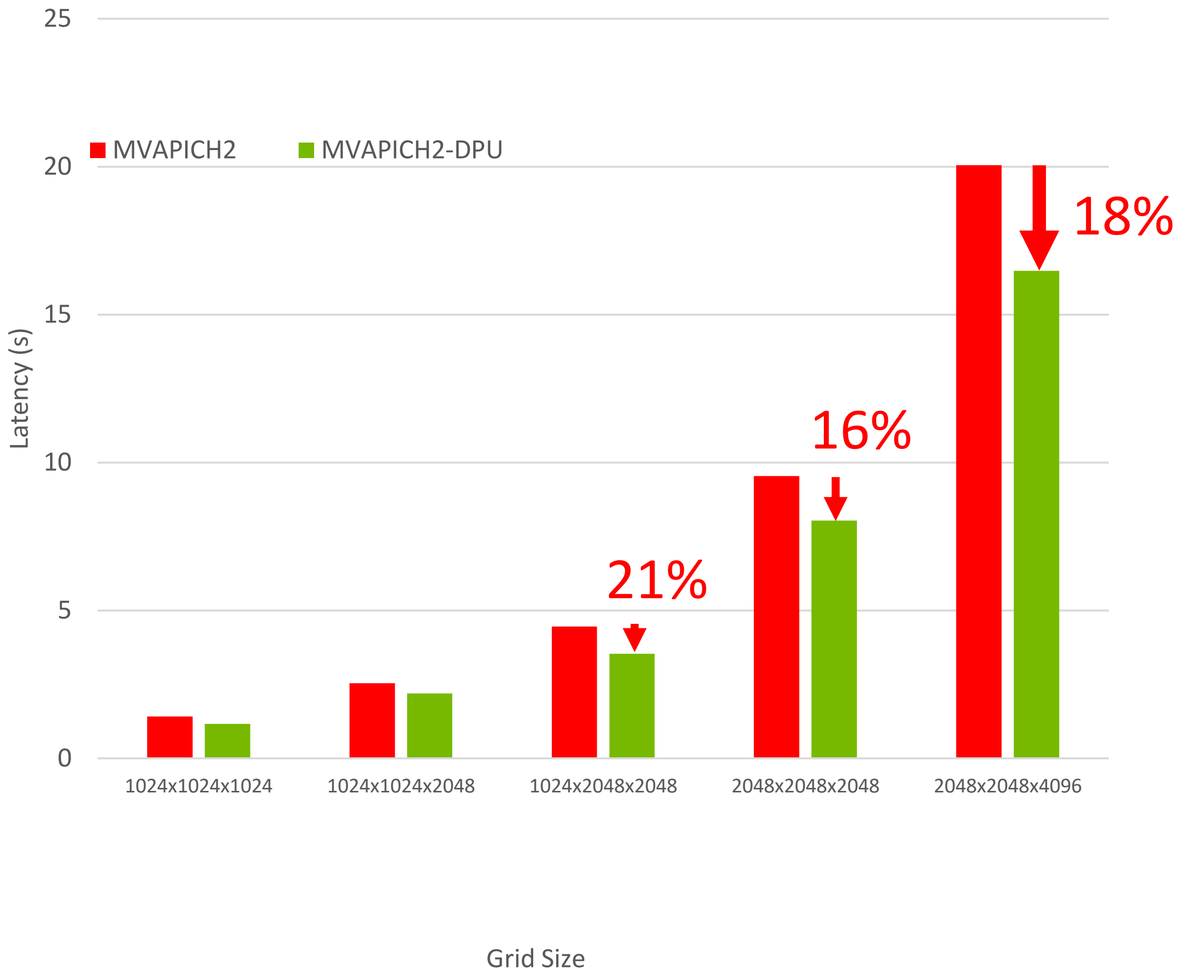
32 Nodes, 32 PPN

Benefits in Total execution time (Compute + Communication)

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



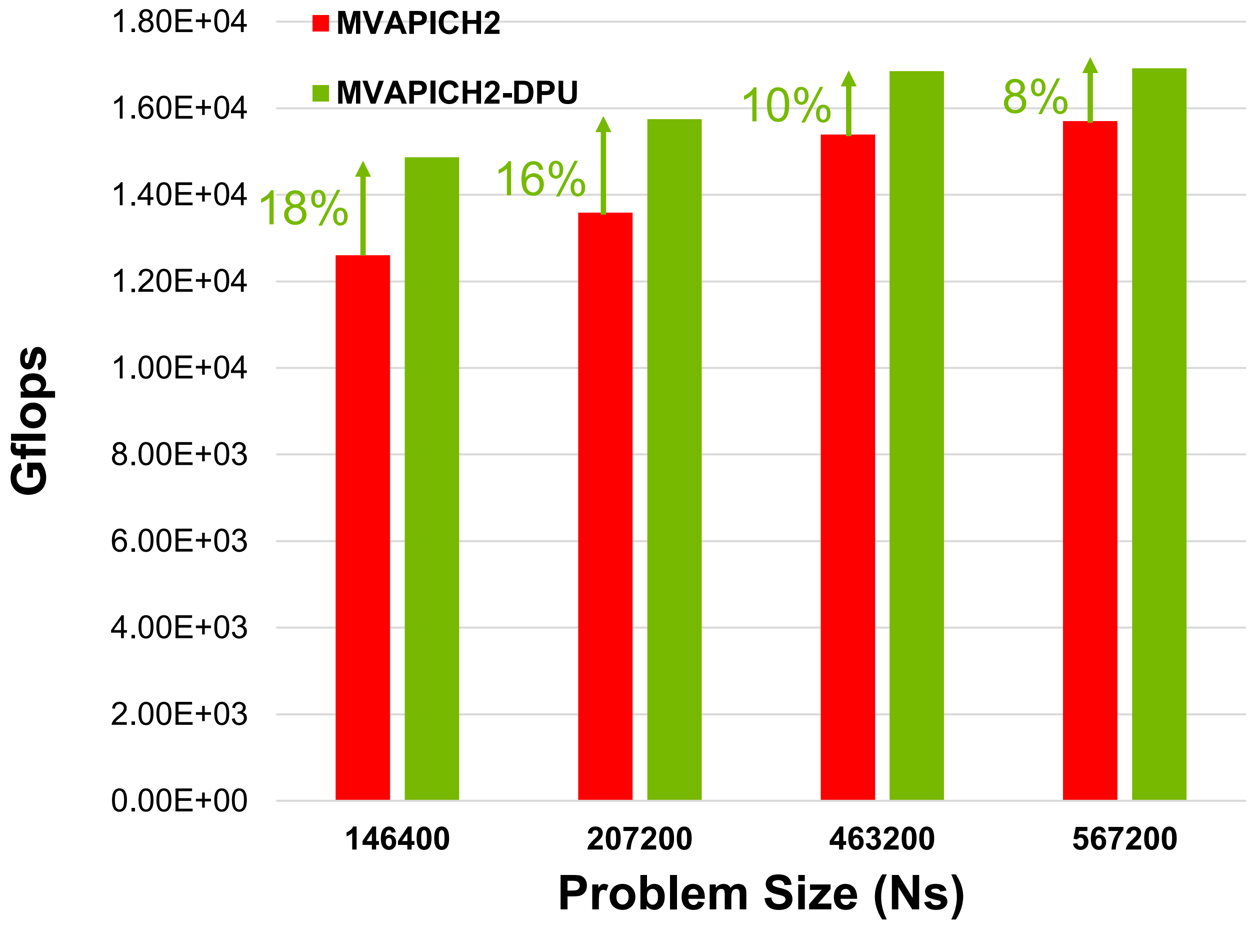
32 Nodes, 16 PPN



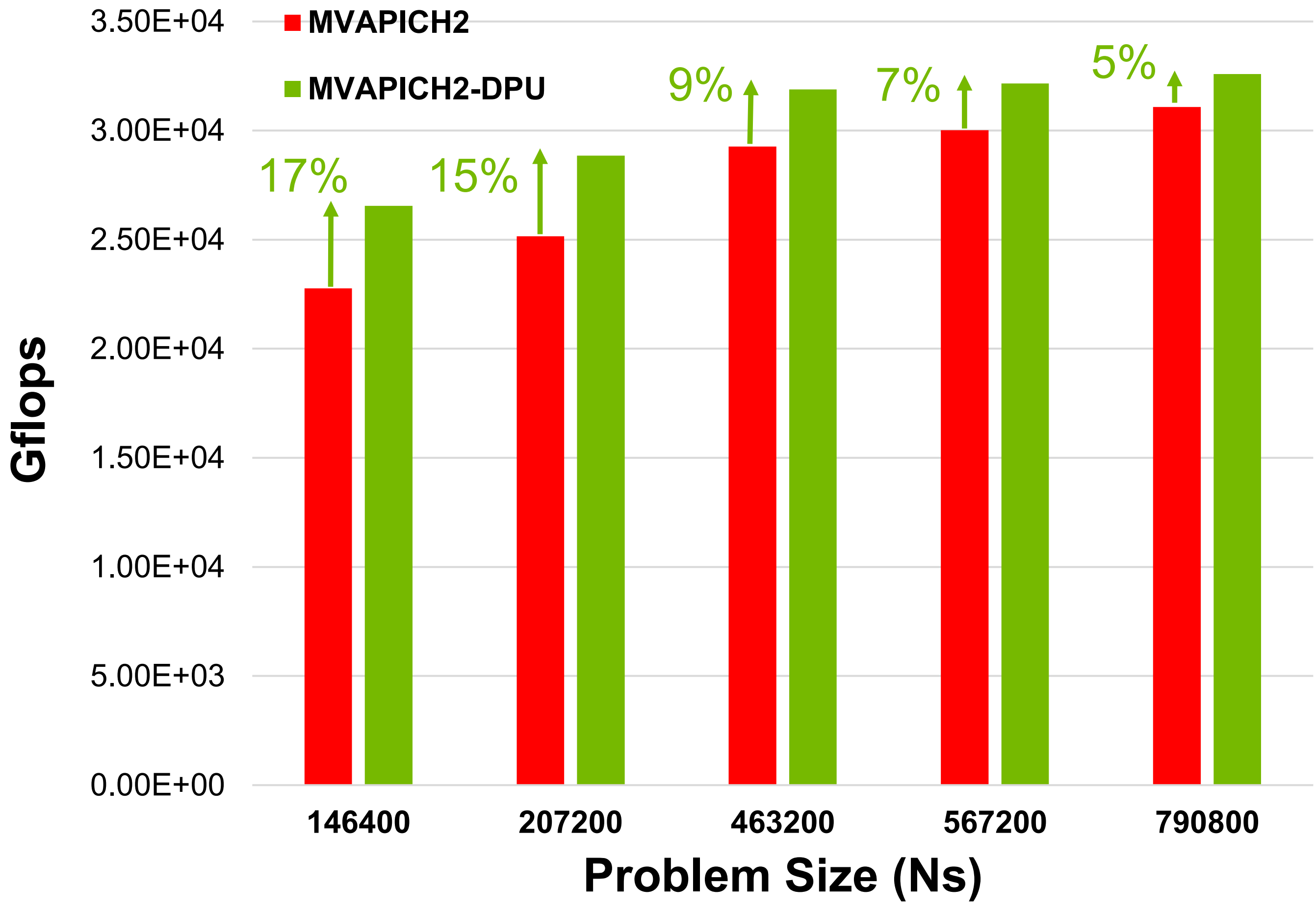
32 Nodes, 32 PPN

Benefits in application-level execution time

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



16x32 process grid



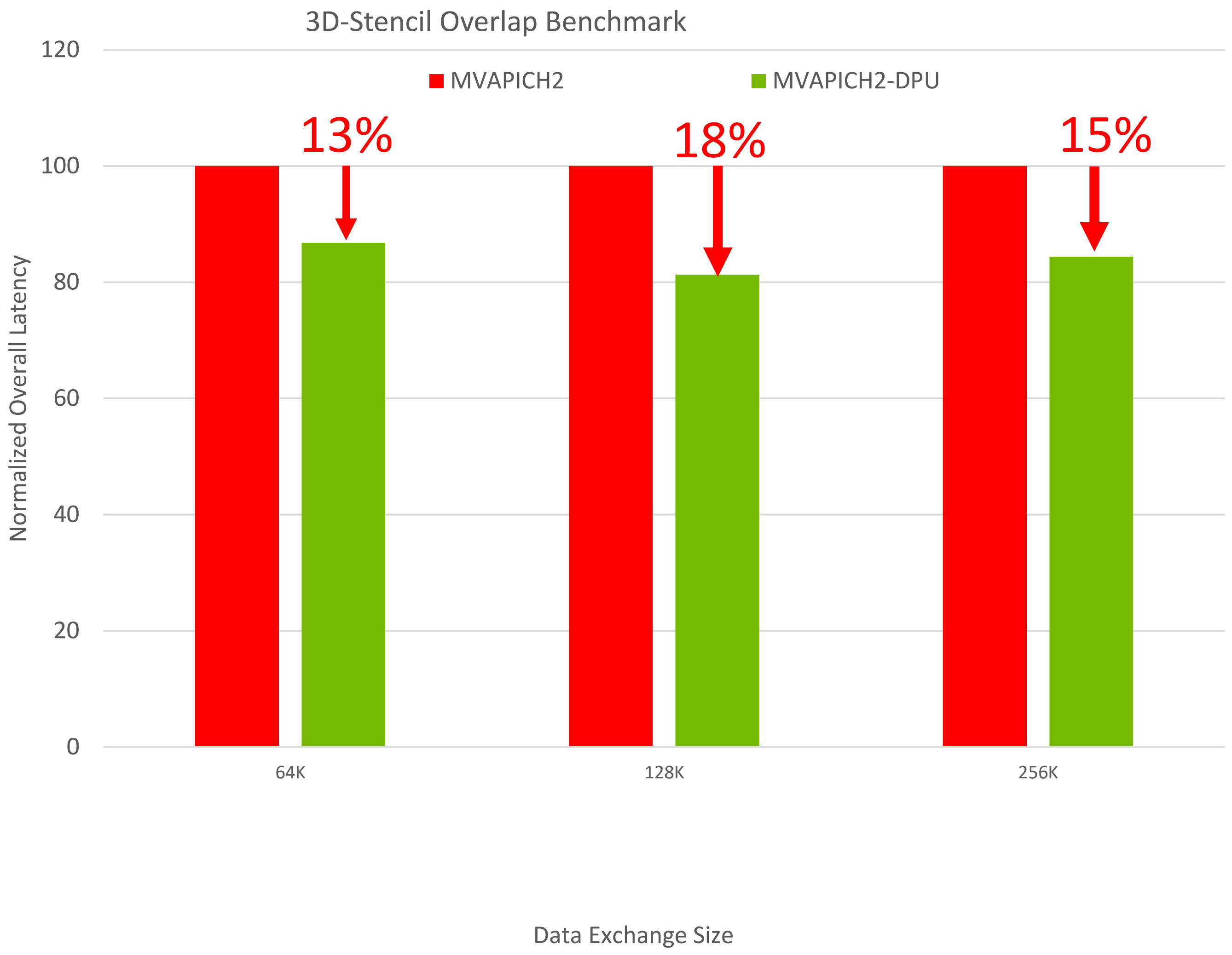
31x32 process grid

Benefits in application-level execution time

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



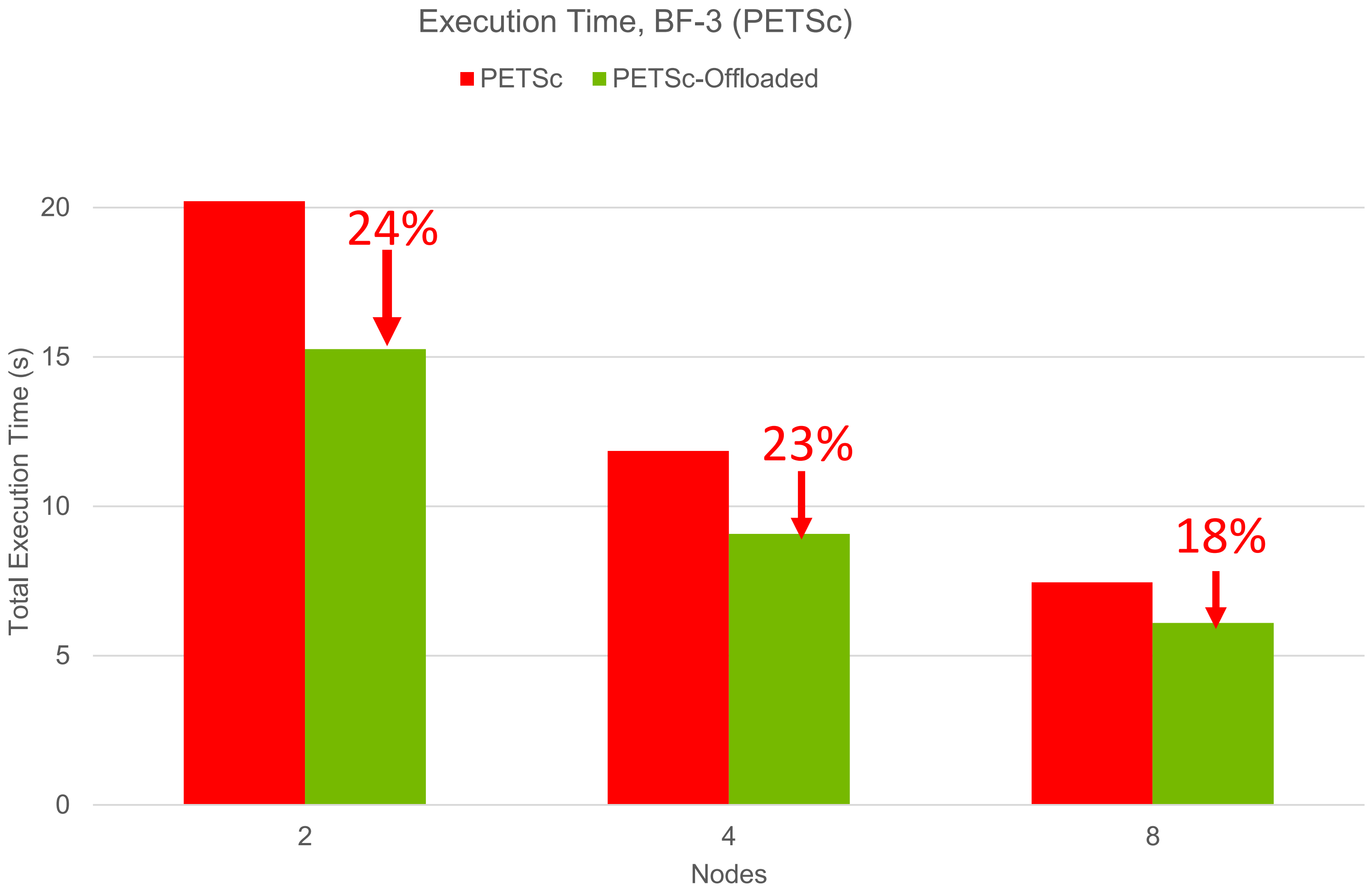
16 Nodes, 32 PPN

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)

Overview of Products

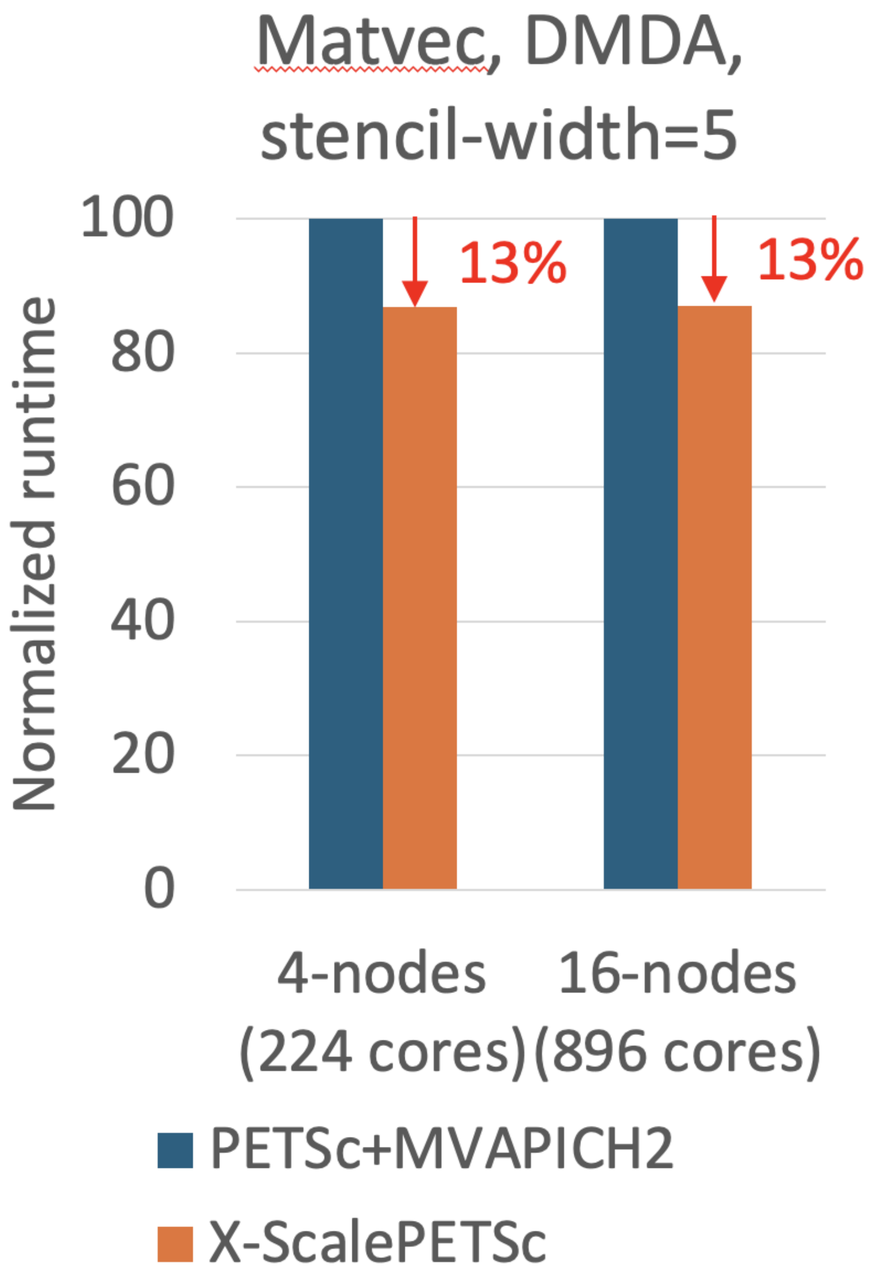
- X-ScaleHPC: High-Performance Optimized Solution for HPC applications
- X-ScaleAI: High-Performance Solution with Deep Introspection for AI applications
- MVAPICH2-DPU: High-Performance MVAPICH2 for Accelerating Applications with NVIDIA’s DPU technology
- X-ScalePETSC: Accelerating PETSC Library (a common library for many scientific workloads) on clusters with CPUs and GPUs
- X-ScaleSecured-MPI: High-Performance MPI library with built-in security
- X-Scale Monitor: HPC/AI hardware monitoring

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.

X-ScalePETSc 2024.4: Matvec Kernel

- Matrix-vector multiplication kernel
- Cartesian structured mesh using DMDA
- Stencil width of 5
- Compare the matvec kernel time, using PETSc+MVAPICH2 vs. using X-ScalePETSc
- Up to 13% performance improvement gained by X-ScalePETSc for different number of nodes



Frontera (TACC)

Processors	Intel 8280 “Cascade Lake”
Cores/Node	56 (28 per socket)
Memory/Node	192GB DDR-4
Network	Mellanox InfiniBand, HDR-100

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-ScalePETSC: Accelerating PETSC Library (a common library for many scientific workloads) on clusters with CPUs and GPUs
- X-ScaleSecured-MPI: High-Performance MPI library with built-in security
- X-Scale Monitor: HPC/AI hardware monitoring

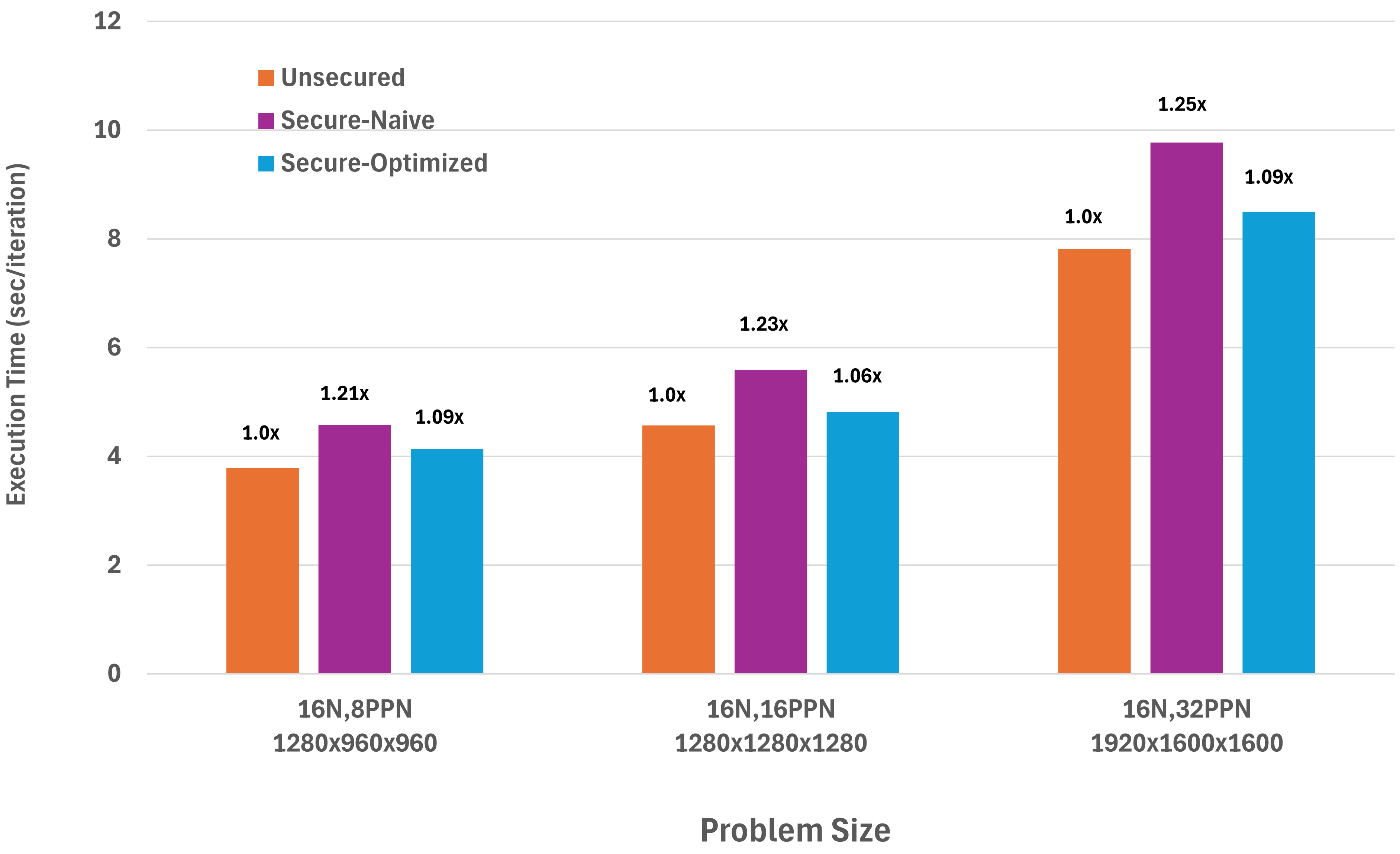
X-ScaleSecureMPI

Main Features

- Scalable solutions of secure communication middleware based on MVAPICH2
- Flexible support for multiple cryptographic libraries and encryption schemes, configurable per request
- Compliant to TLS/SSL security key management protocol
- Supports secured point-to-point communication operations, blocking and non-blocking
- Simple installation and execution in one command
- Supports widely used collective operations including broadcast, alltoall and allgather
- Tested with MPI micro-benchmarks and MPI applications up to 1,024 ranks

SecureMPI Performance: P3DFFT Application Kernel

- Parallel 3D FFT application kernel with various problem sizes
- Up to 16 nodes,32 ppn on an Intel cluster (Inter-node & intra-node communication)

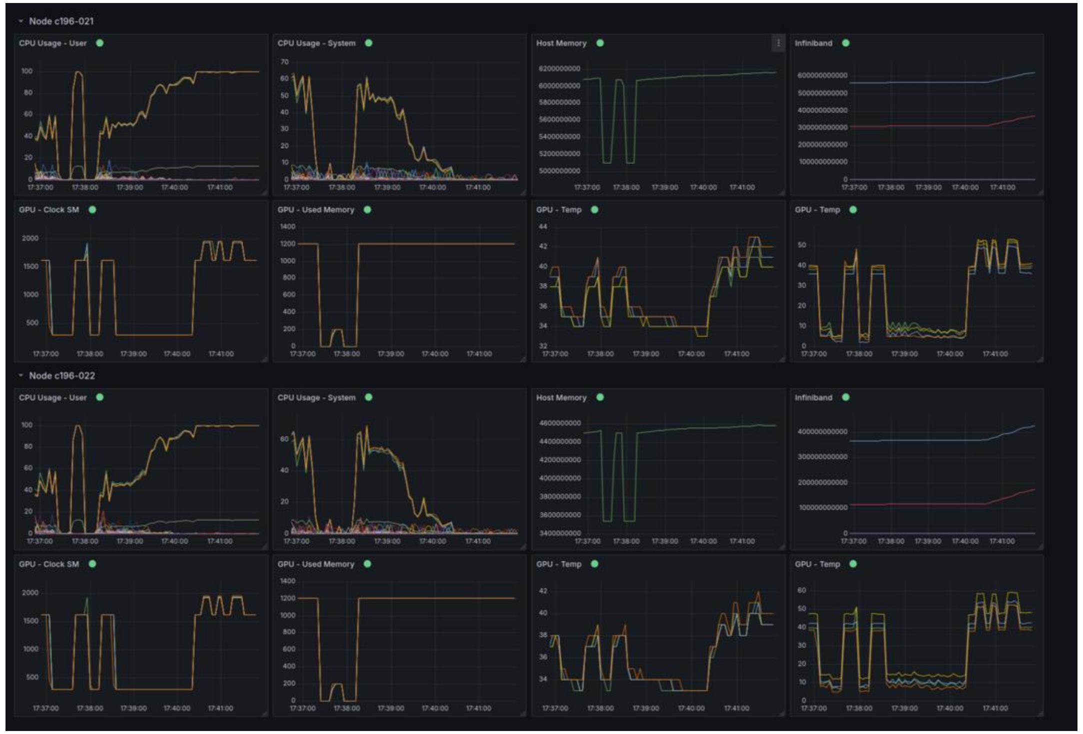


Overview of Products

- 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-ScalePETSC: Accelerating PETSC Library (a common library for many scientific workloads) on clusters with CPUs and GPUs
- X-ScaleSecured-MPI: High-Performance MPI library with built-in security
- X-Scale Monitor: HPC/AI hardware monitoring

X-Scale Monitor

- **X-Scale Monitor Features**
  - Rich monitoring on every node in the job
    - Metrics for the CPU, GPU, and network
  - InfluxDB database integration
    - All logs can be backed up to a database for convenient metric storage and offline processing
  - X-Scale monitoring is a standalone tool
    - Can be used for HPC jobs, AI training, AI inference, etc
  - Containerized deployment
    - We ship X-Scale Monitor as a container (Docker or Apptainer), which users or cluster managers can immediately deploy
  - Extremely low-overhead
    - We have novel polling designs to reduce the overhead of querying hardware state
    - This enables high polling frequencies (<1s) without affecting the application



Conclusions

- X-ScaleSolutions offer a suite of software for accelerating HPC and AI applications with a focus on efficient communication, overlapping between communication and computation, and communication security
- Innovative value-added products provide high-performance and scalable solutions for HPC and AI applications while exploiting modern CPU, GPU, and DPU technologies
- Promises potential for exploiting higher performance, significant scalability and reduced TCO for your HPC and AI applications
- X-ScaleSolutions will be happy to get engaged with end customers, collaborators, resellers, and integrators

Thank You!

contactus@x-scalesolutions.com

X-ScaleSolutions

http://x-scalesolutions.com/