Accelerating HPC and DL Applications Using MVAPICH2-DPU Library and X-ScaleAI Package

Donglai Dai

Aug 24, 2022

X-ScaleSolutions

http://x-scalesolutions.com

Outline

- **Overview of X-ScaleSolutions**
- MVAPICH2-DPU: High-Performance MPI for Accelerating Applications with NVIDIA's DPU technology
- X-ScaleAI package: High-Performance Toolkit for Accelerating DL Applications
- X-ScaleAI-DPU package: High-Performance Toolkit for Accelerating DL Applications with intelligent DPU offloading
- Conclusion

X-ScaleSolutions

- 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, and OSU INAM)
  - High-Performance Big Data Libraries (RDMA-Hadoop, RDMA-Spark, RDMA-HBase, and RDMA-Memcached)
- Provide commercial support of these Libraries to US federal national labs and international supercomputer centers

X-ScaleSolutions (Cont'd)

- Winner of multiple U.S. DOE SBIR grants to design and develop innovative and value added products
- A Silver ISV member of the OpenPOWER Consortium
- More details on all products in http://x-scalesolutions.com
  - contactus@x-scalesolutions.com

Outline

- Overview of X-ScaleSolutions
- **MVAPICH2-DPU: High-Performance MPI for Accelerating Applications with NVIDIA's DPU technology**
- X-ScaleAI package: High-Performance Toolkit for Accelerating DL Applications
- X-ScaleAI-DPU package: High-Performance Toolkit for Accelerating DL Applications with intelligent DPU offloading
- Conclusion

MVAPICH2-DPU Library

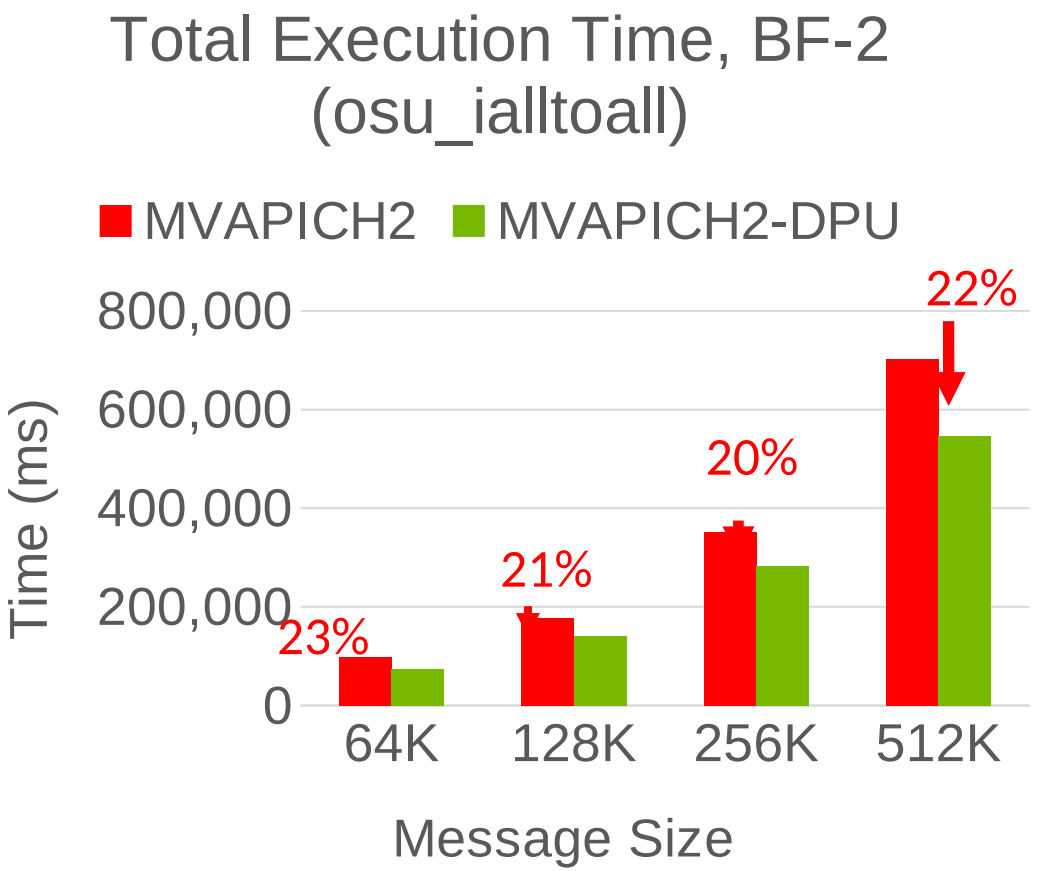
- Based on MVAPICH2 2.3.6
- Current version v2022.02; Next version will be released soon
- Supports all features available with the MVAPICH2 2.3.6 release (http://mvapich.cse.ohio-state.edu)
- Enables offloading non-blocking collectives to DPU
    - Alltoall (MPI_Ialltoall)
    - Allgather (MPI_Iallgather)
    - Broadcast (MPI_Ibcast)

MVAPICH2-DPU Library (Cont'd)

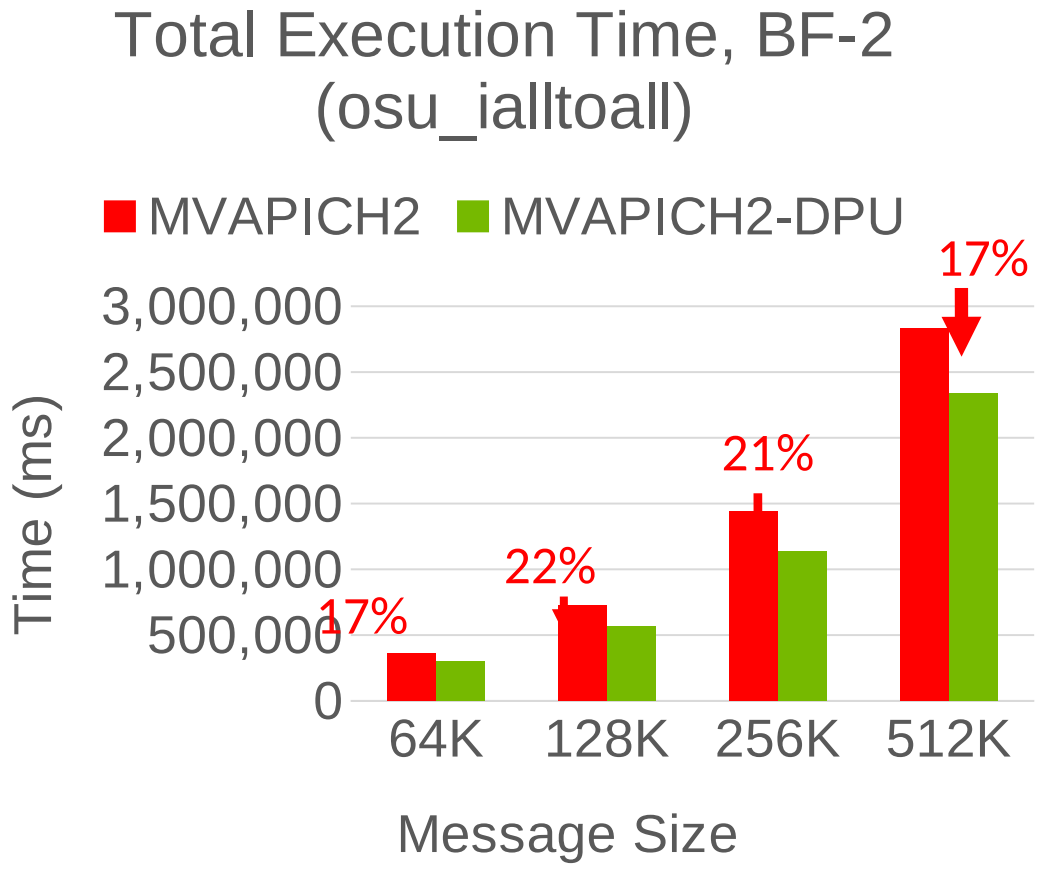
- Significantly increases (up to 100%) overlap of computation with any mix of MPI_Ialltoall, MPI_Iallgather, or MPI_Ibcast non-blocking collectives
- Accelerates scientific applications using any mix of MPI_Ialltoall , MPI_Iallgather, or MPI_Ibcast non-blocking collectives

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)



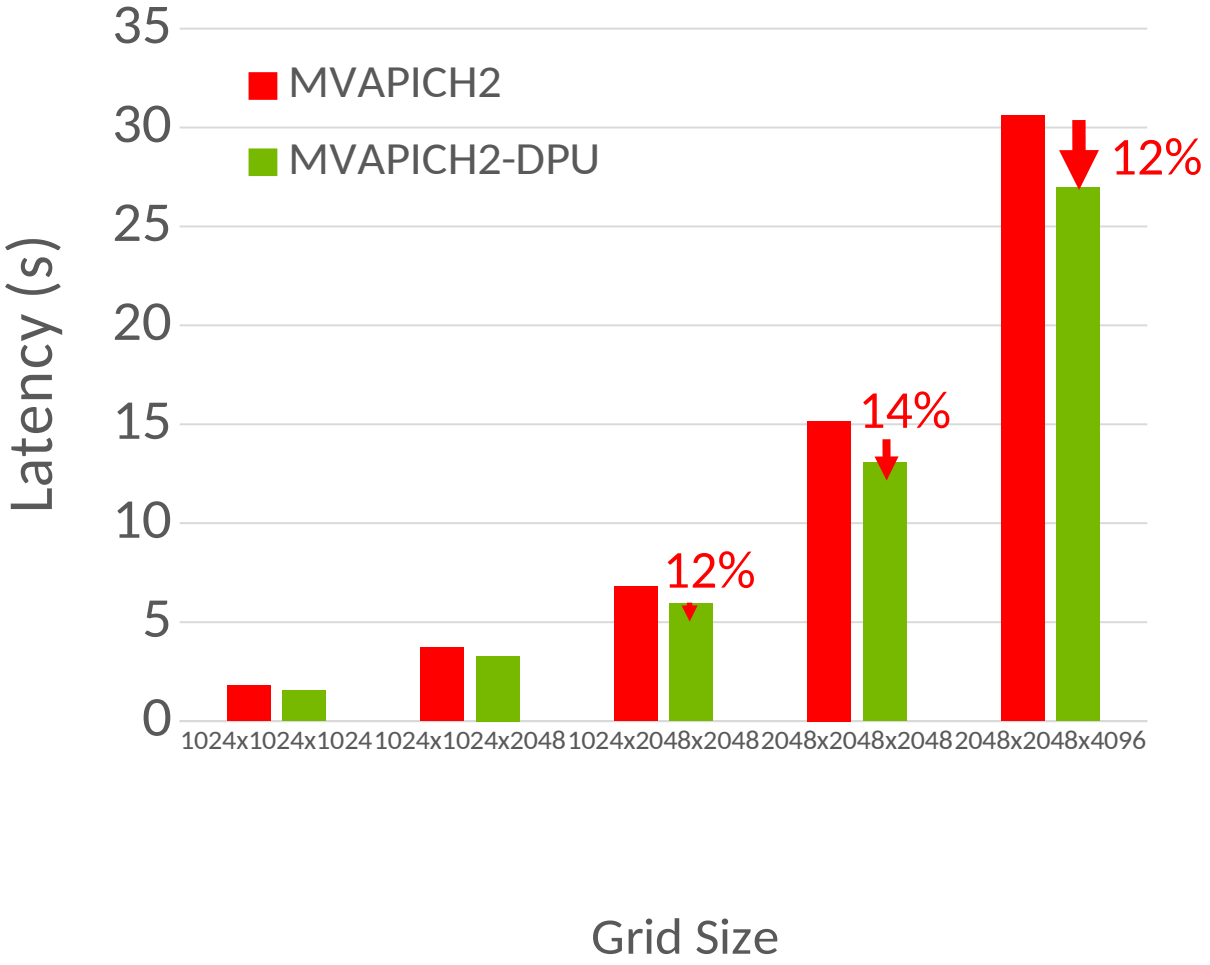
32 Nodes, 16 PPN



32 Nodes, 32 PPN

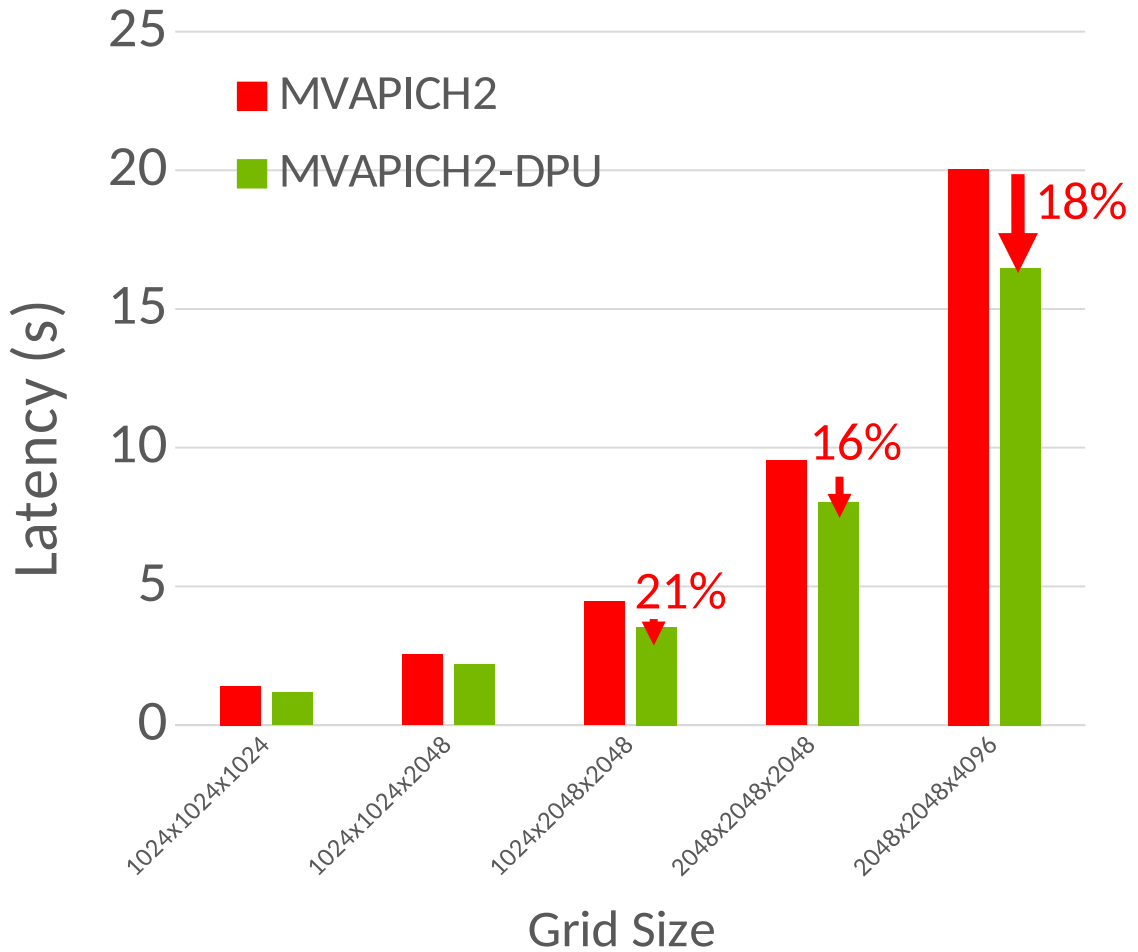
Benefits in Total execution time (Compute + Communication)

P3DFFT Application Execution Time (32 nodes)



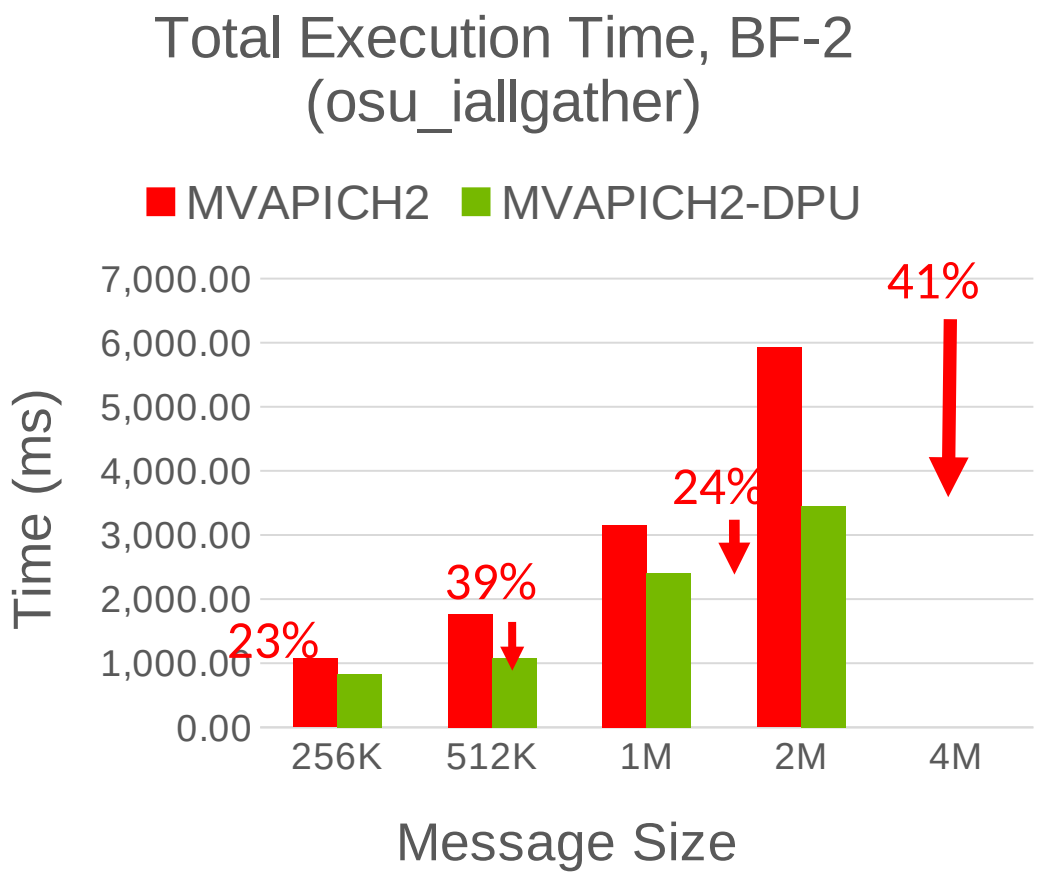
32 Nodes, 16 PPN

Benefits in application-level execution time

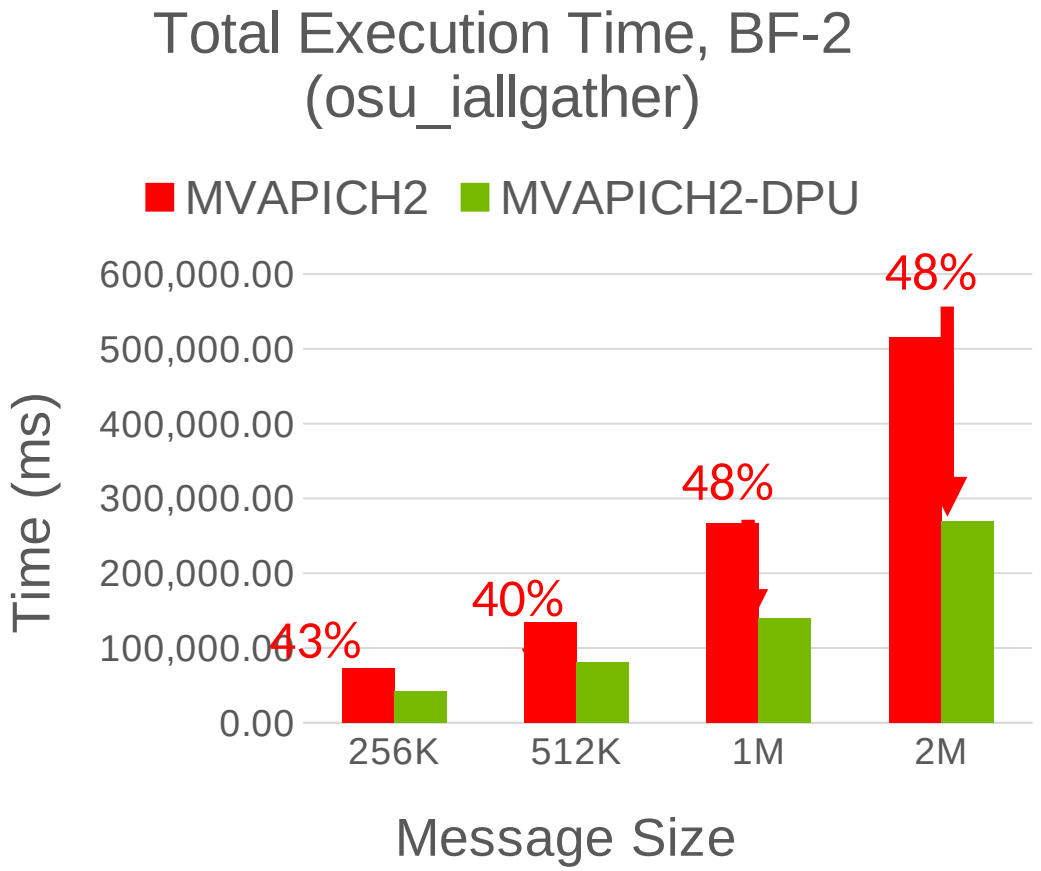


32 Nodes, 32 PPN

Total Execution Time with osu_Iallgather (16 nodes)



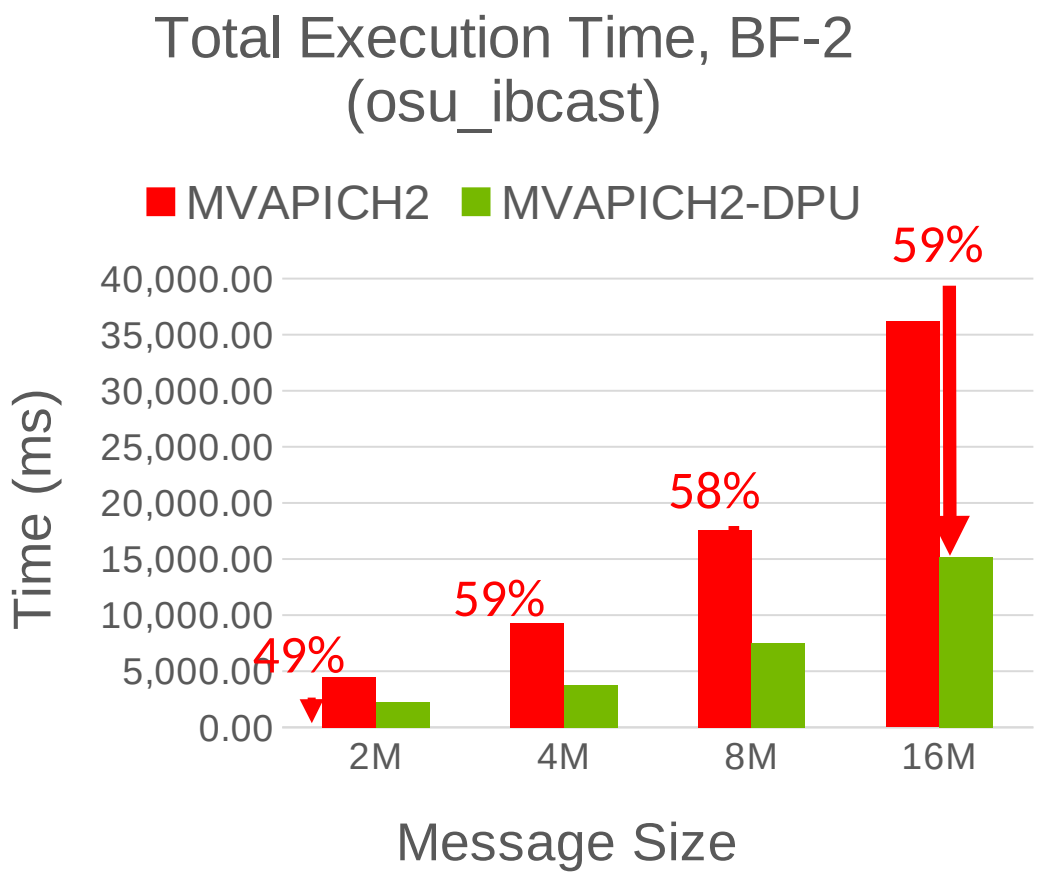
16 Nodes, 1 PPN



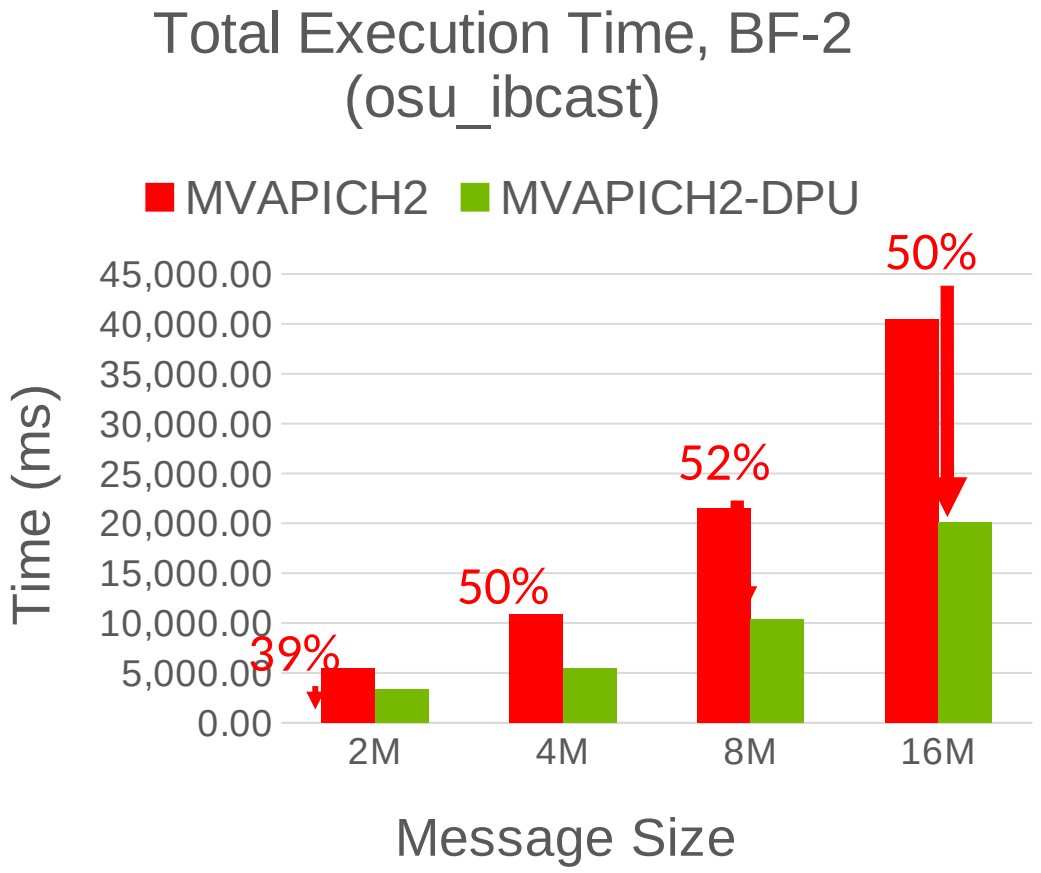
16 Nodes, 16 PPN

Benefits in Total execution time (Compute + Communication)

Total Execution Time with osu_Ibcast (16 nodes)



16 Nodes, 16 PPN



16 Nodes, 32 PPN

Benefits in Total execution time (Compute + Communication)

Outline

- Overview of X-ScaleSolutions
- MVAPICH2-DPU: High-Performance MPI for Accelerating Applications with NVIDIA's DPU technology
- **X-ScaleAI package: High-Performance Toolkit for Accelerating DL Applications**
- X-ScaleAI-DPU package: High-Performance Toolkit for Accelerating DL Applications with intelligent DPU offloading
- Conclusion

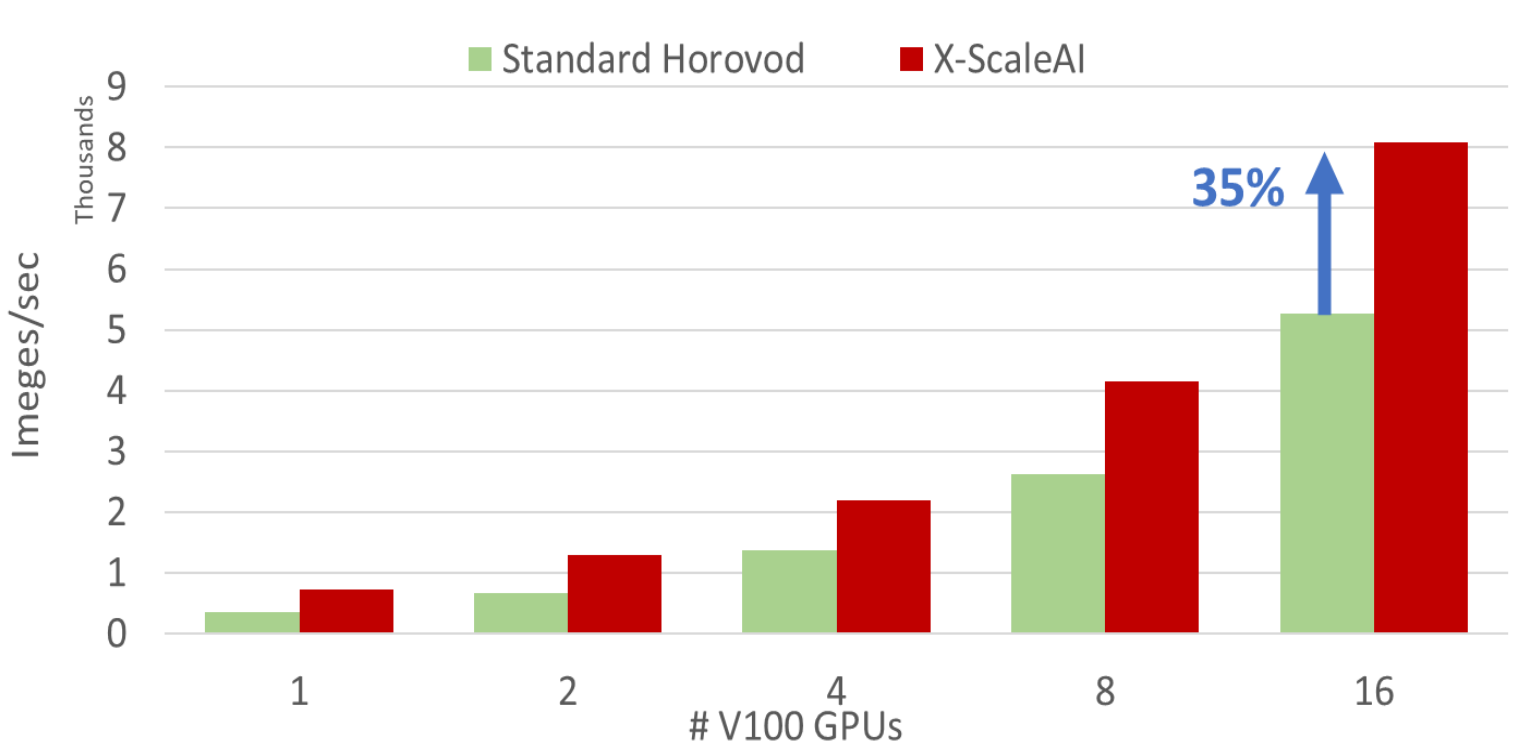
Features of X-ScaleAI Package

- Built on top of MVAPICH2 libraries
- Integrated packaging to support popular DL frameworks
  - TensorFlow, PyTorch, MXNet, etc
- Integrated profiling and introspection support for DL applications across the stacks (DeepIntrospect)
  - Helps users to optimize their DL applications for higher performance and scalability
- Integrated efficient checkpointing restart support
- Targeted for both CPU-based and GPU-based DL training
- One-click deployment and Out-of-the-box optimal performance
- Support for OpenPOWER and x86 platforms
- Support for InfiniBand, RoCE and NVLink Interconnects

X-ScaleAI : Distributed PyTorch on Sample System Configuration #1

System Configuration #1:

- Eight NVIDIA Tesla V100-32GB SXM2 GPUs (per node)
- Two Intel Xeon Gold 6248 "Cascade Lake" CPUs:
  - 20 cores, 2.50–3.90GHz, 27.5MB LLC, 6 memory channels
- 512GB of RAM: DDR4-2933
- 7.68TB NVMe SSD
- Two Mellanox ConnectX-6 InfiniBand HDR 200Gb/s Adapter

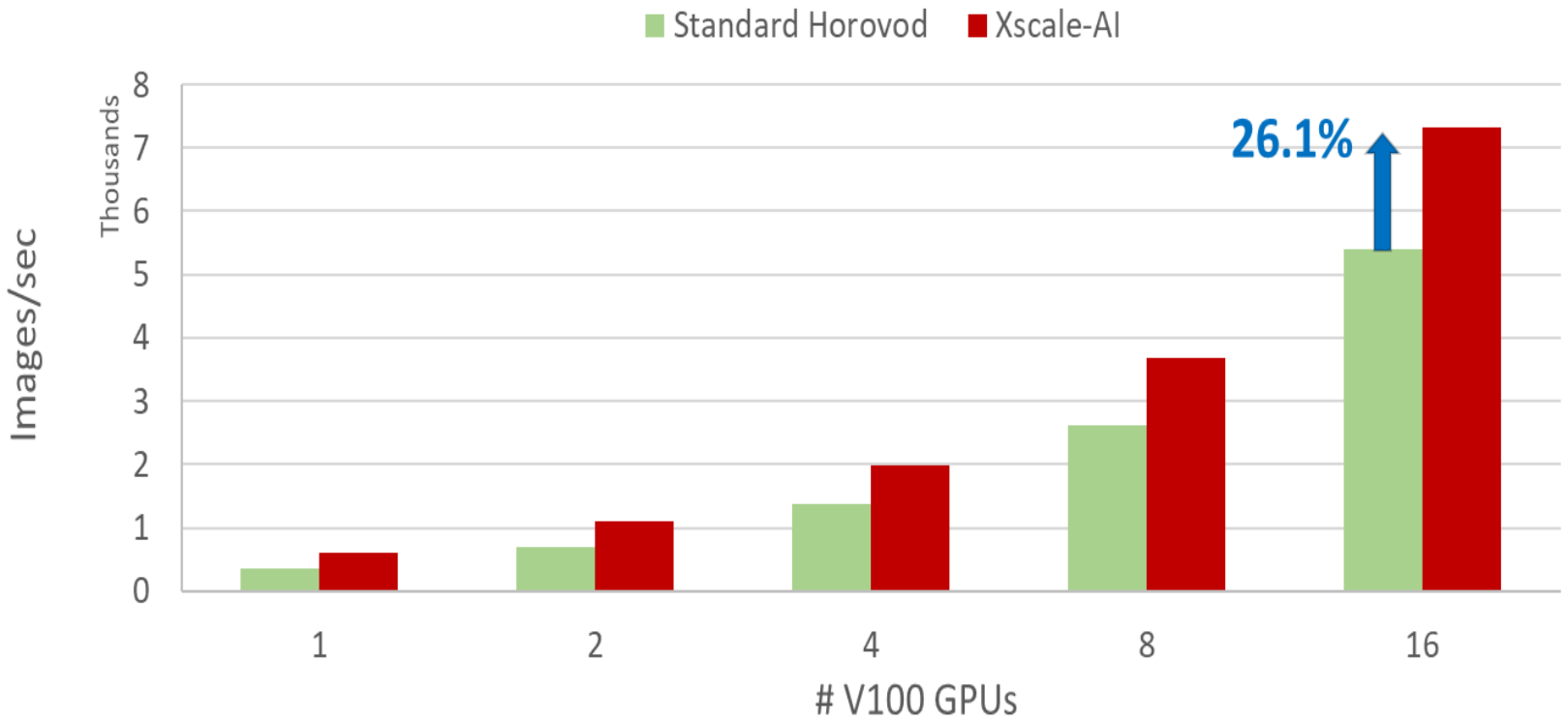


Comparison of the performance of training ResNet-50 model on the Imagenet dataset using up to 16 GPUs (2 nodes, 8 GPUs per node) on system configuration #1.

X-ScaleAI : Distributed PyTorch on Sample System Configuration #2

System Configuration #2:

- Four NVIDIA Tesla V100-SXM2 GPUs (per node), connected with NVLink
- Two Intel Xeon Gold 6248 “Cascade Lake” CPUs:
  - 20 cores, 2.50GHz
- 384GB of RAM: DDR4
- 1.6TB Samsung PM1745b NVMe PCIe SSD
- Two Mellanox ConnectX-6 InfiniBand HDR 200Gb/s Adapter



Comparison of the performance of training ResNet-50 model on the Imagenet dataset using up to 16 GPUs (4 nodes, 4 GPUs per node) on system configuration #2

X-ScaleAI DI GUI Profiler View (Expended)

Deep Introspect Profiler

X-ScaleSolutions

DEEP INTROSPECT (DI) DASHBOARD:

NUMBER OF PROCESSES (NP): 1024

PROCESSES PER NODE (PPN): 4

PROMPT: xscale-ai-run -np 1024 --hostfile ./hfile ./xscale-ai/install/miniconda/bin/python ./xscale-ai/install/benchmarks/horovod_benchmarks/pytorch/pytorch_synthetic_benchmark.py --batch-size=64

MPI_Allreduce

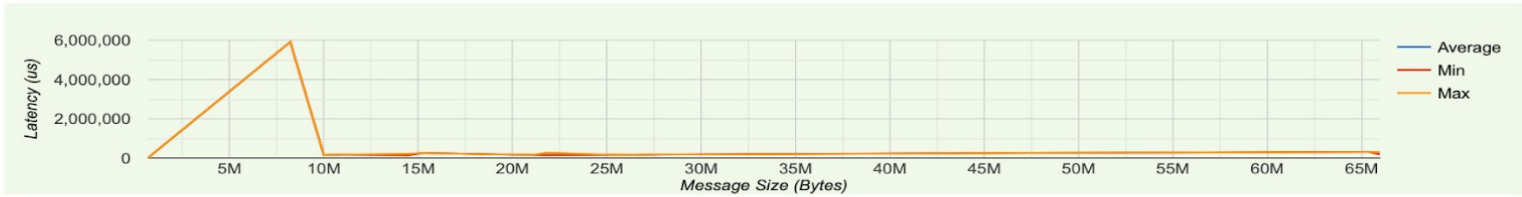
TOTAL CALLS
331

TOTAL TIME (US)
64,843,618

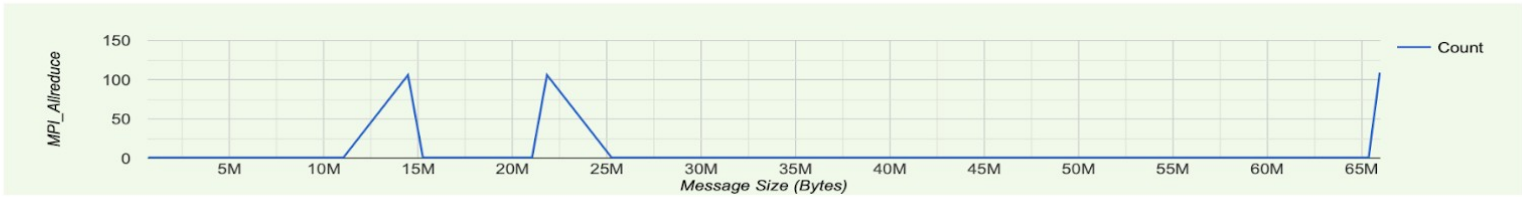
USAGE TAG
Parameter and Gradients

MPI OPERATION
MPI_Allreduce

Latency (us) by Message Size



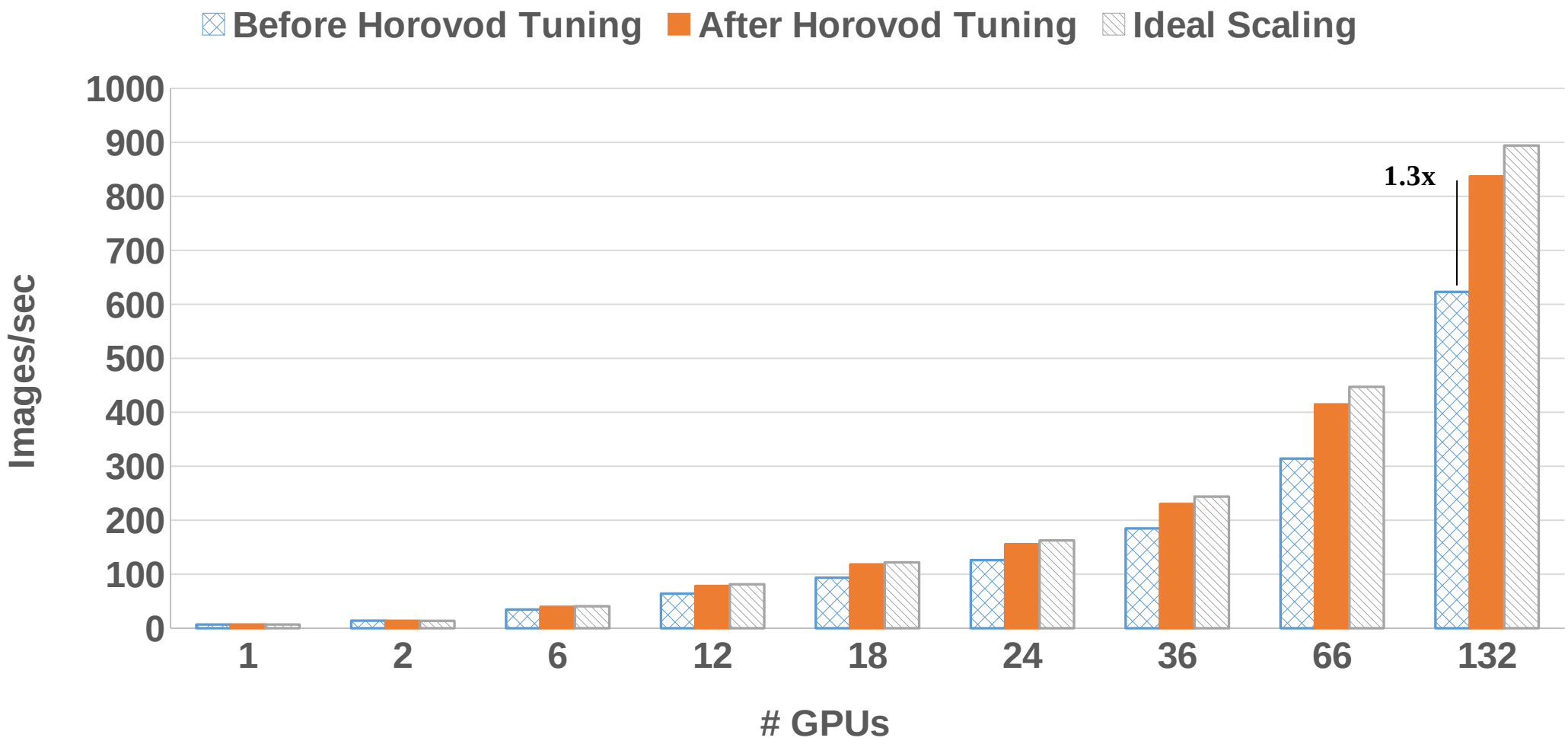
Count by Message Size



Latency (us) and Count

Message Size	Count	Latency (us)		
		Average	Min	Max
685,312	1	19,256	19,256	19,256
8,212,384	1	5,912,414	5,912,414	5,912,414
9,985,024	1	155,680	155,680	155,680
11,034,624	1	178,562	178,562	178,562
14,452,736	106	155,870	140,740	218,396
15,240,448	1	250,050	250,050	250,050
21,029,888	1	161,634	161,634	161,634
21,817,600	106	158,618	146,251	269,381
25,235,712	1	167,044	167,044	167,044

X-ScaleAI Use Case #1: Application Benefits (DeepLabv3+)



Harness 30% higher performance and better scaling on DeepLabv3+ (using TF) with the X-ScaleAI Tool

X-ScaleAI Use Case #2: Application Benefits (ResNet-50)

- As a result of tuning the MPI layer, the user can vastly improve application performance

# GPUs	Images/sec (Expected)*	Images/sec (Obtained Initially)	Images/sec (Obtained Finally)
1024	**~370,000**	**181,020**	**341,590**

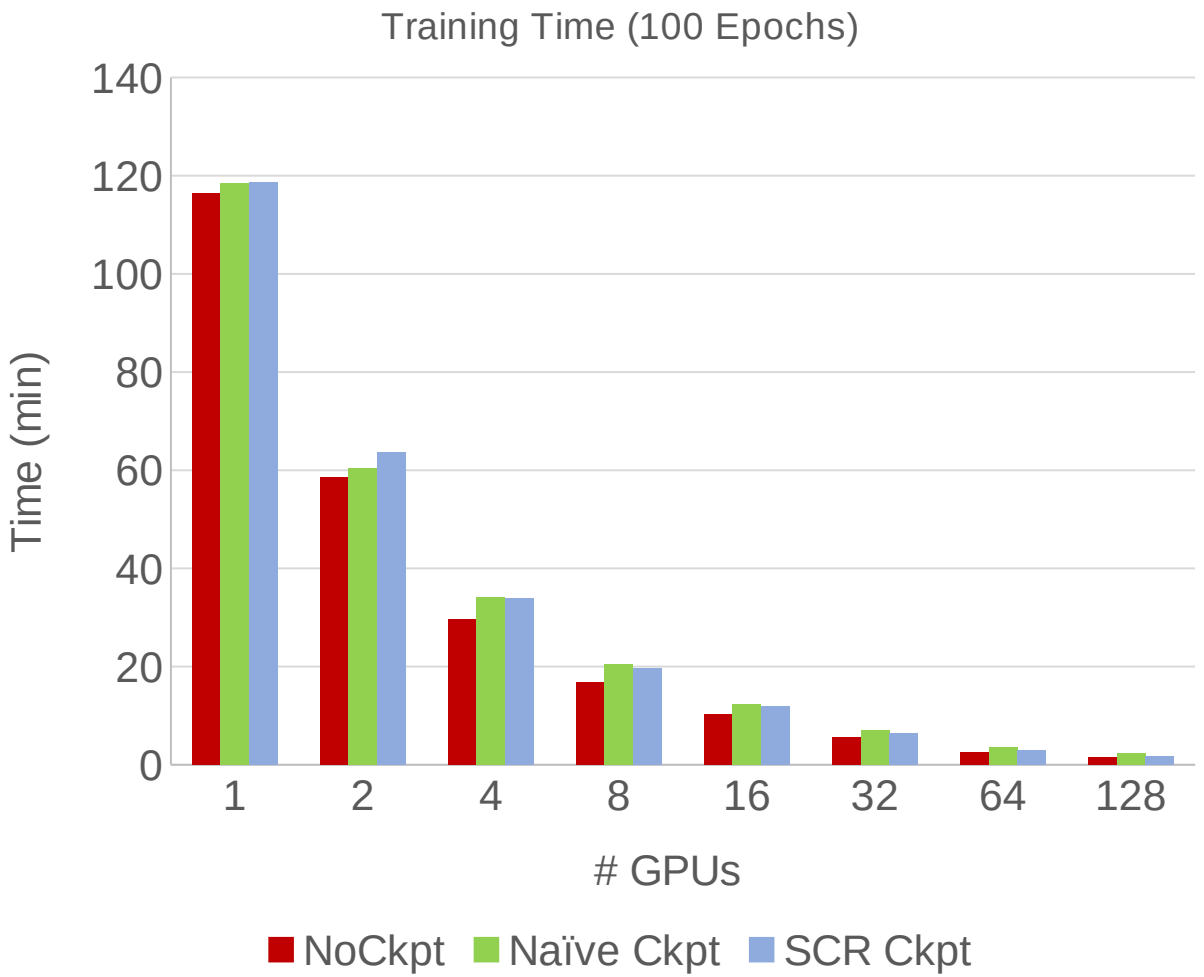
1.9x speedup in ResNet-50 (using PyTorch) throughput, while reducing debugging time for the DL scientist considerably!!

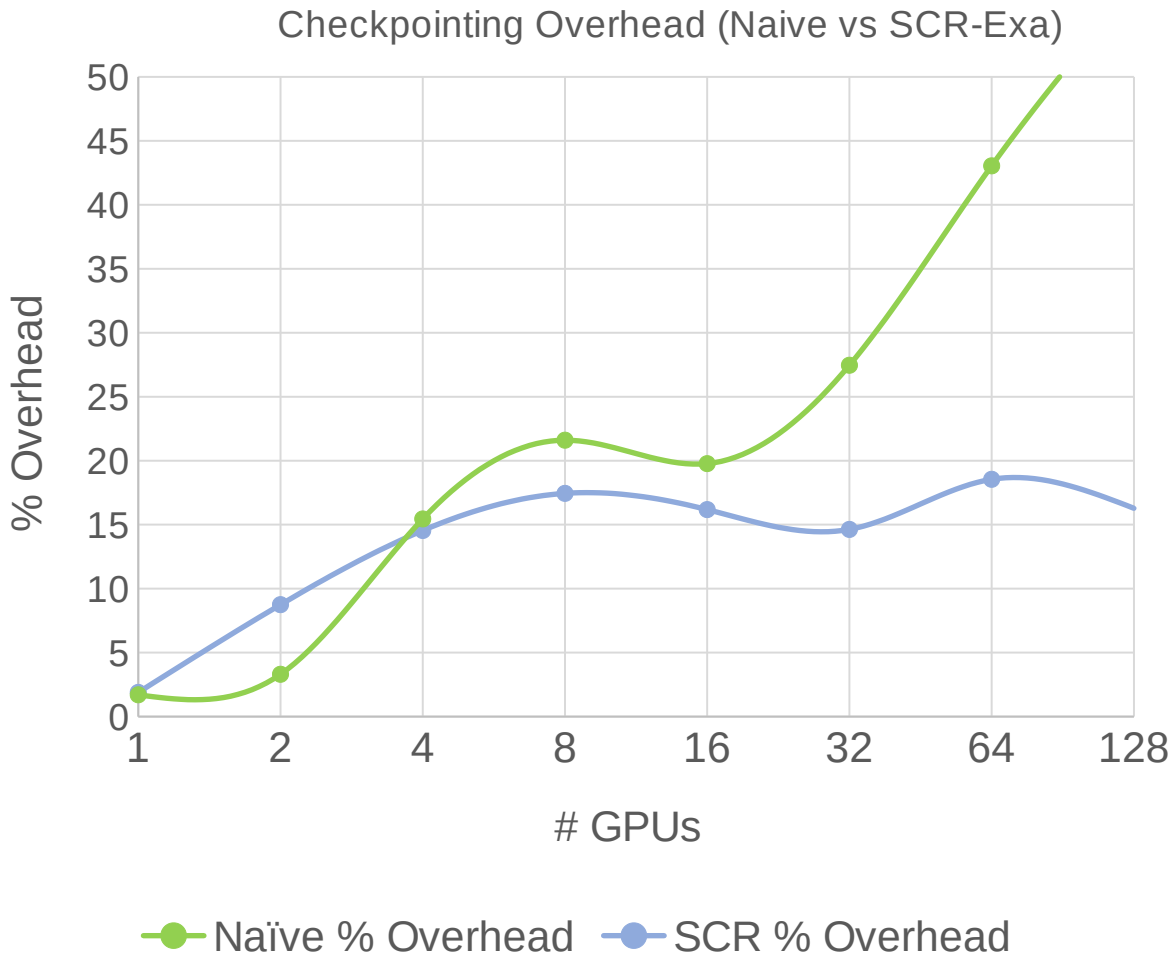
Efficient Checkpointing for DL Applications

- Based on open-source SCR library from Lawrence Livermore National Laboratories (LLNL)
- Support efficient checkpointing for any DL models and applications using popular DL frameworks
  - PyTorch Distributed Data Parallel Model (DDP)
  - PyTorch Over Horovod
- Example DL Applications:
  - Residual Neural Network (ResNet)
  - Enhanced Deep Residual Networks for Single Image Super-Resolution (EDSR)

Checkpointing performance for ResNet-50

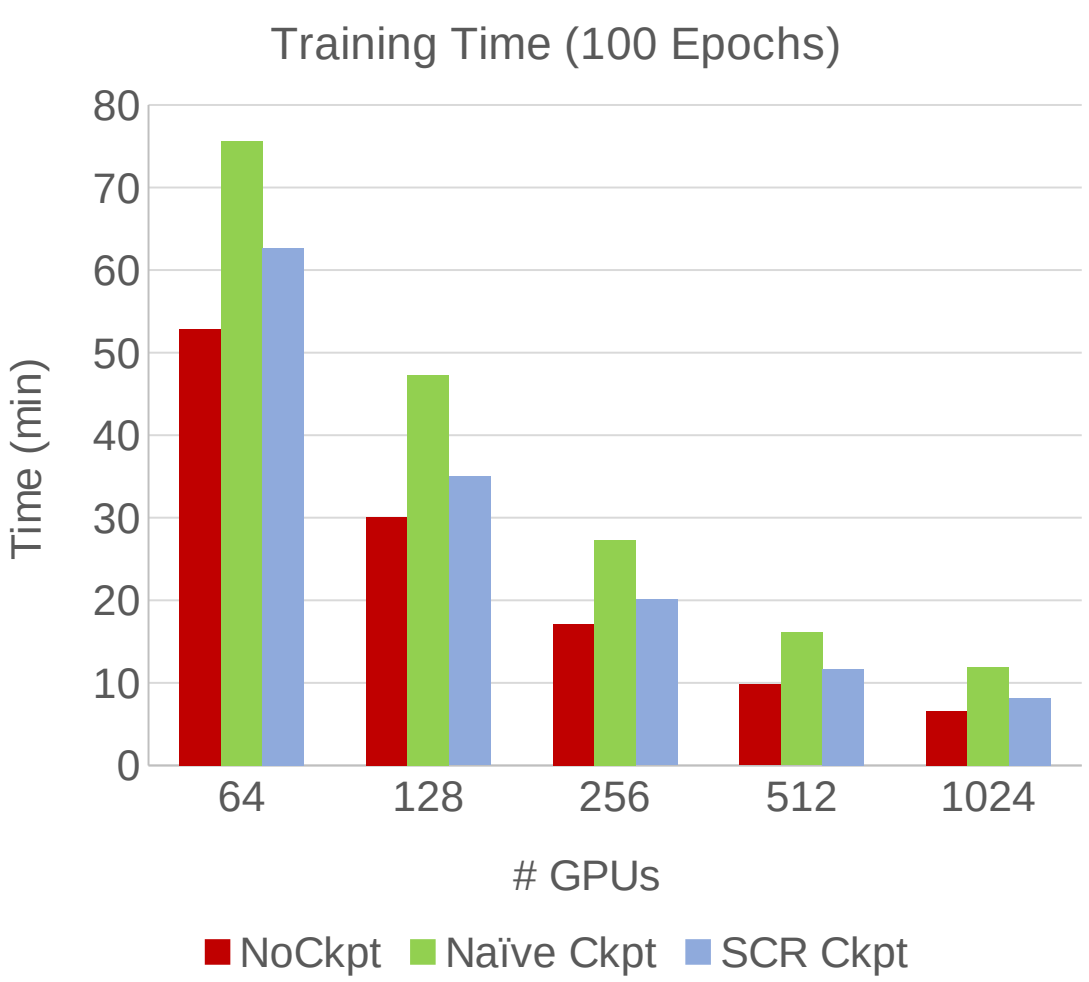
- Similar performance trends observed for the PyTorch over Horovod platform

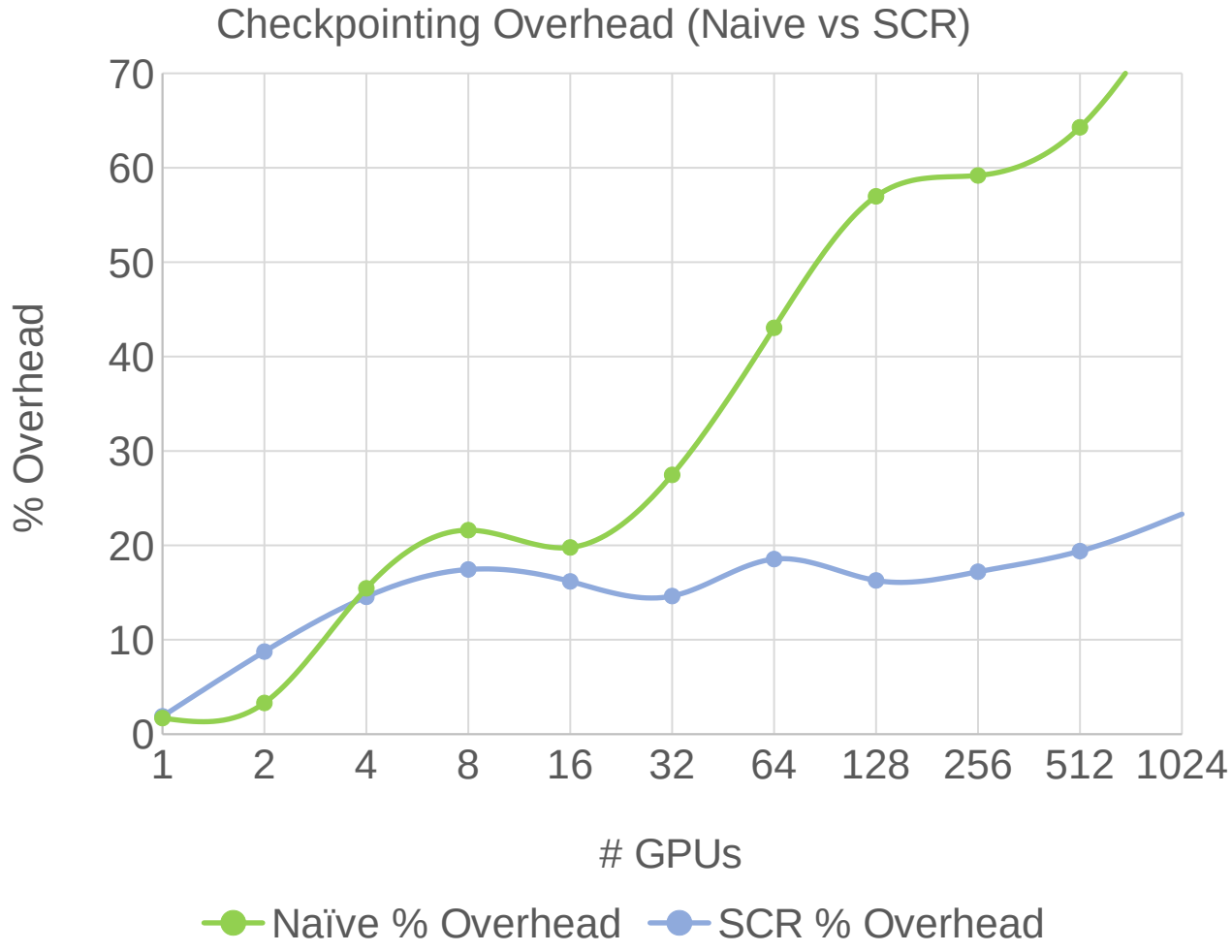




Checkpointing performance for EDSR

- Similar performance trends observed for the PyTorch over Horovod platform





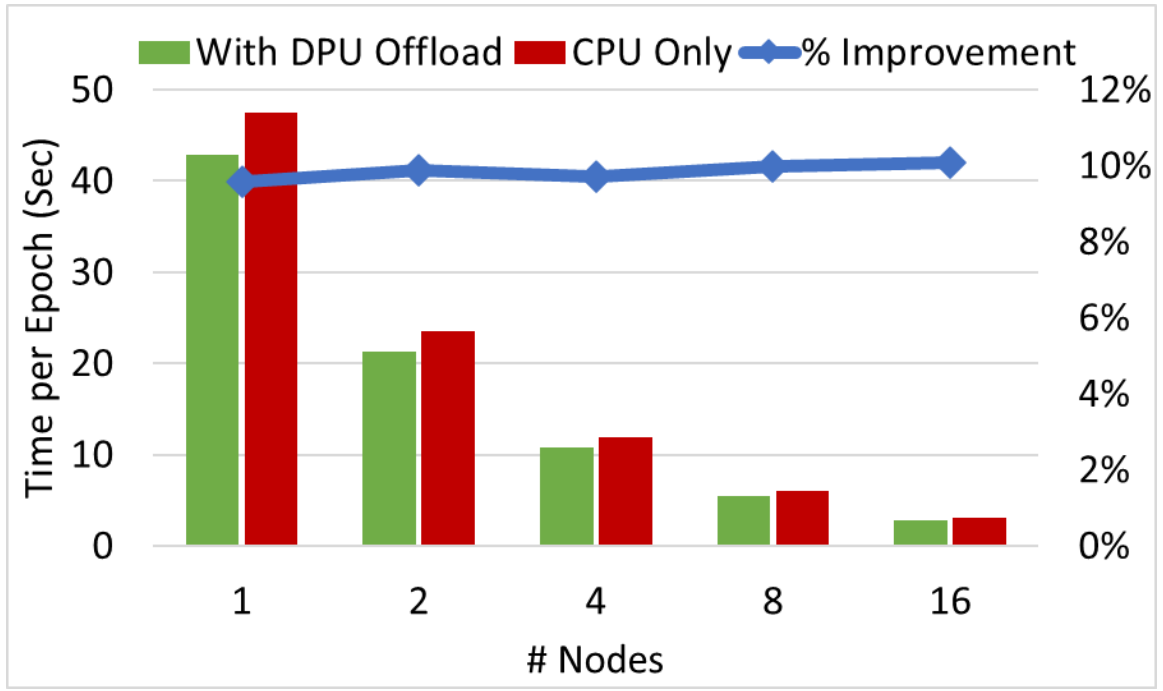
Outline

- Overview of X-ScaleSolutions
- MVAPICH2-DPU: High-Performance MPI for Accelerating Applications with NVIDIA's DPU technology
- X-ScaleAI package: High-Performance Toolkit for Accelerating DL Applications
- **X-ScaleAI-DPU package: High-Performance Toolkit for Accelerating DL Applications with intelligent DPU offloading**
- Conclusion

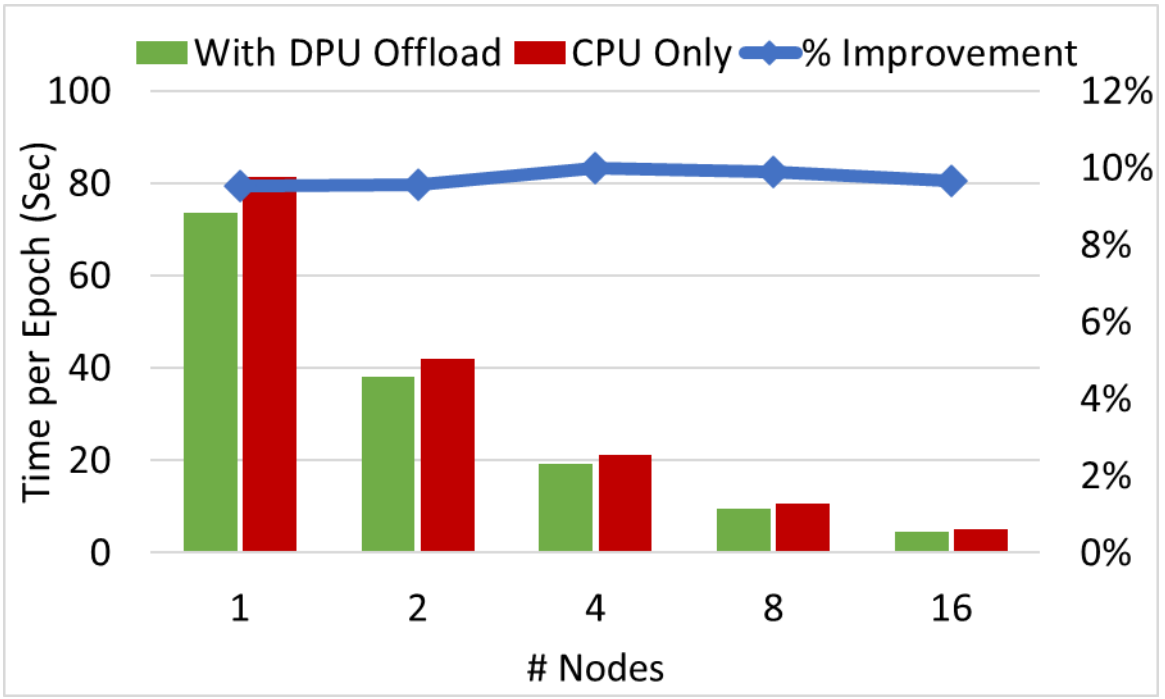
X-ScaleAI-DPU Package

- Support for distributed CPU-based DL training by offloading different DNN stages to NVIDIA DPUs.
- Intelligent designs to noticeably accelerate DNN training.
- Support for PyTorch/Torchvision and user defined DNN models and datasets.
- User friendly and simple Python API with automatic DNN training function.
- One-click deployment on both x86 and ARM architectures and Out-of-the-box optimal performance
  - Do not need to struggle for many hours

X-ScaleAI-DPU : Distributed DNN Training Examples



Comparison between DPU offload and CPU only training ResNet-18 model on the CIFAR10 dataset



Comparison between DPU offload and CPU only training ShuffleNet model on the SVHN dataset

Conclusion

- Exponential growth is projected in HPC and AI market in the decades ahead
- HPC software are critical for HPC and DL applications to take full advantages of advanced hardware technologies
  - x86_64/OpenPOWER CPU; NVIDIA/AMD/Intel GPU; NVIDIA DPU, Intel IPU; InfiniBand/High-Speed Ethernet/NVLink networks, etc.
- MVAPICH2-DPU, X-ScaleAI, and X-ScaleAI-DPU from X-ScaleSolutions provide tailored high-performance solutions for your complex HPC and AI applications on your target hardware systems.
- Contact us for a demo and free-trial! (contactus@x-scalesolutions.com)

Thank You!

d.dai@x-scalesolutions.com

http://x-scalesolutions.com/