Accelerating HPC and DL Applications Using DPUs and Efficient Checkpointing

Donglai Dai

Aug 25, 2021

X-ScaleSolutions

http://x-scalesolutions.com

Outline

- **Overview of X-ScaleSolutions**
- X-ScaleHPC and X-ScaleAI packages
- MVAPICH2-DPU: High-Performance MVAPICH2 for Accelerating Applications with NVIDIA's DPU technology
- SCR-Exa: Efficient and Scalable Checkpointing for HPC and DL Applications

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
- **X-ScaleHPC and X-ScaleAI packages**
- MVAPICH2-DPU: High-Performance MVAPICH2 for Accelerating Applications with NVIDIA's DPU technology
- SCR-Exa: Efficient and Scalable Checkpointing for HPC and DL Applications

X-ScaleHPC Solution

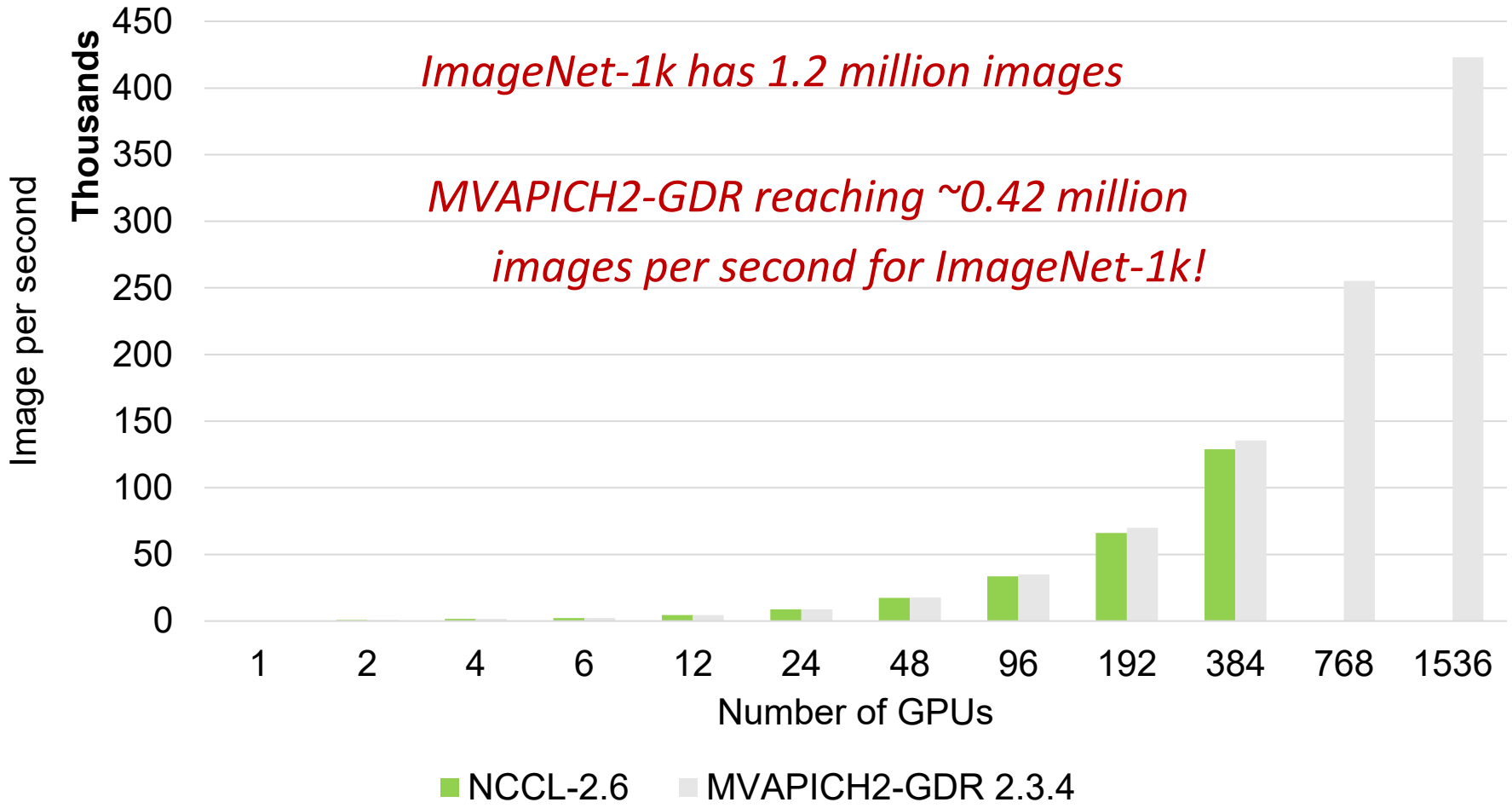
- Scalable solutions of communication middleware based on OSU MVAPICH2 libraries
- “out-of-the-box” fine-tuned and optimal performance on various HPC systems including CPUs and GPUs
- Professional technical support and customer services
- Stable and growing list of commercial customers

Features of X-ScaleAI Solution

- 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)
  - Provides cross-stack performance analysis in a visual manner and helps users to optimize their DL applications for higher performance and scalability
- Targeted for both CPU-based and GPU-based DL training
- Out-of-the-box optimal performance
  - Tuned for various CPU- and GPU-based HPC systems
- One-click deployment and execution
  - Do not need to struggle for many hours
- Support for OpenPOWER and x86 platforms
- Support for InfiniBand, RoCE and NVLink Interconnects

X-ScaleAI : Distributed TensorFlow on Summit (1,536 GPUs)

- ResNet-50 Training using TensorFlow benchmark on SUMMIT -- 1536 Volta GPUs!
- 1,281,167 (1.2 mil.) images
- Time/epoch = 3 seconds
- Total Time (90 epochs) = 3 x 90 = 270 seconds = 4.5 minutes!



*We observed issues for NCCL2 beyond 384 GPUs

Platform: The Summit Supercomputer (#2 on Top500.org) – 6 NVIDIA Volta GPUs per node connected with NVLink, CUDA 10.1

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

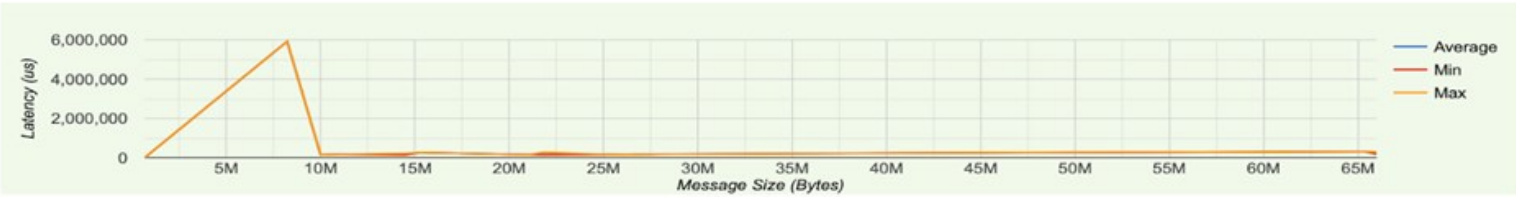
TOTAL CALLS
331

TOTAL TIME (US)
64,843,618

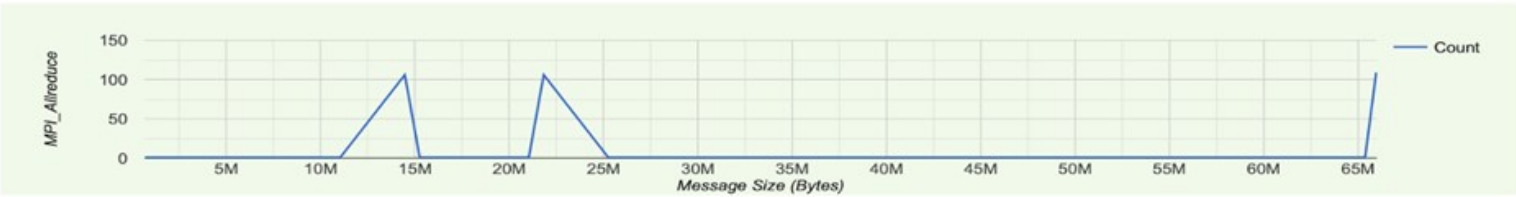
USAGE TAG
Parameter and Gradients

MPI OPERATION
MPI_Allreduce

Latency (us) by Message Size



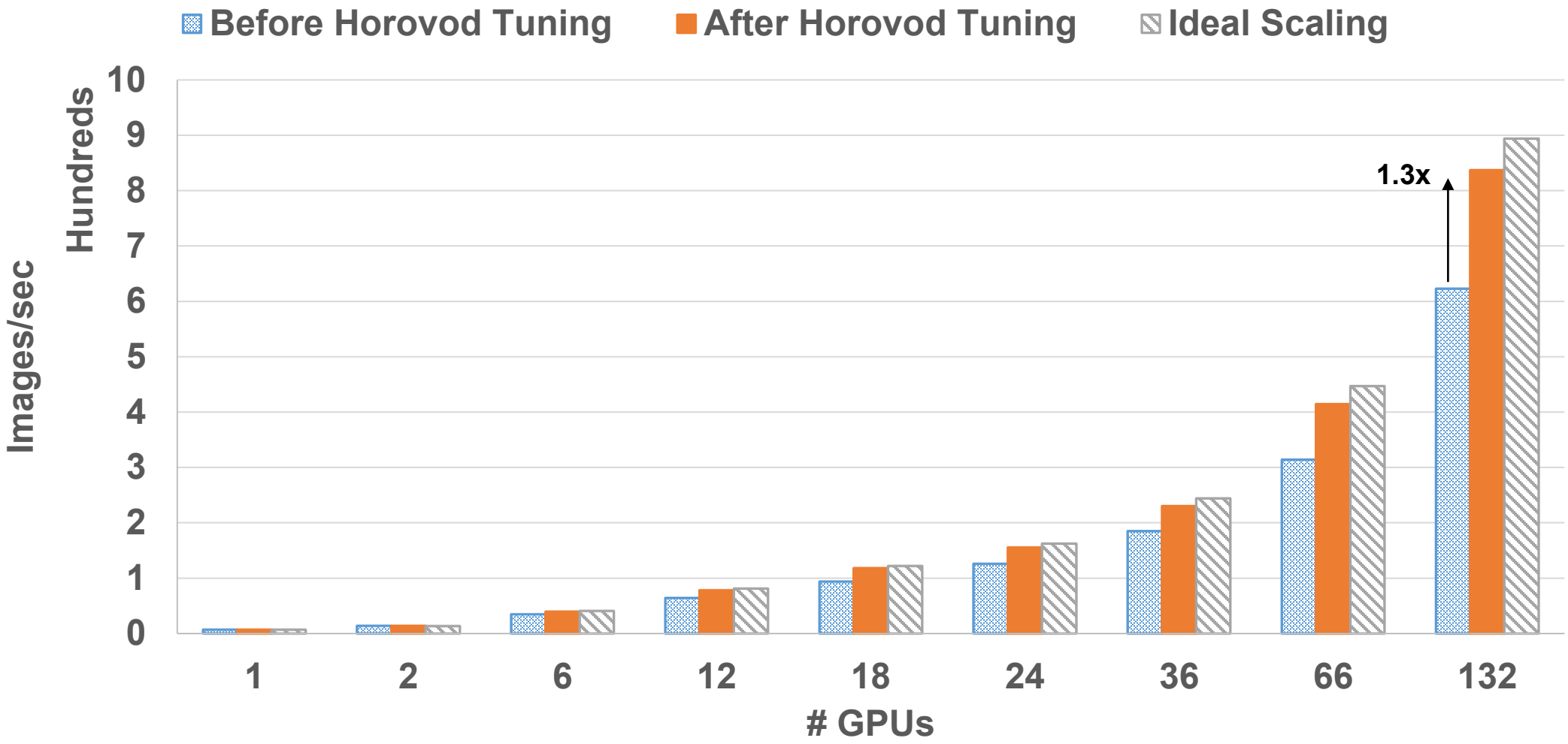
Count by Message Size



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!!

Outline

- Overview of X-ScaleSolutions
- X-ScaleHPC and X-ScaleAI packages
- **MVAPICH2-DPU: High-Performance MVAPICH2 for Accelerating Applications with NVIDIA's DPU technology**
- SCR-Exa: Efficient and Scalable Checkpointing for HPC and DL Applications

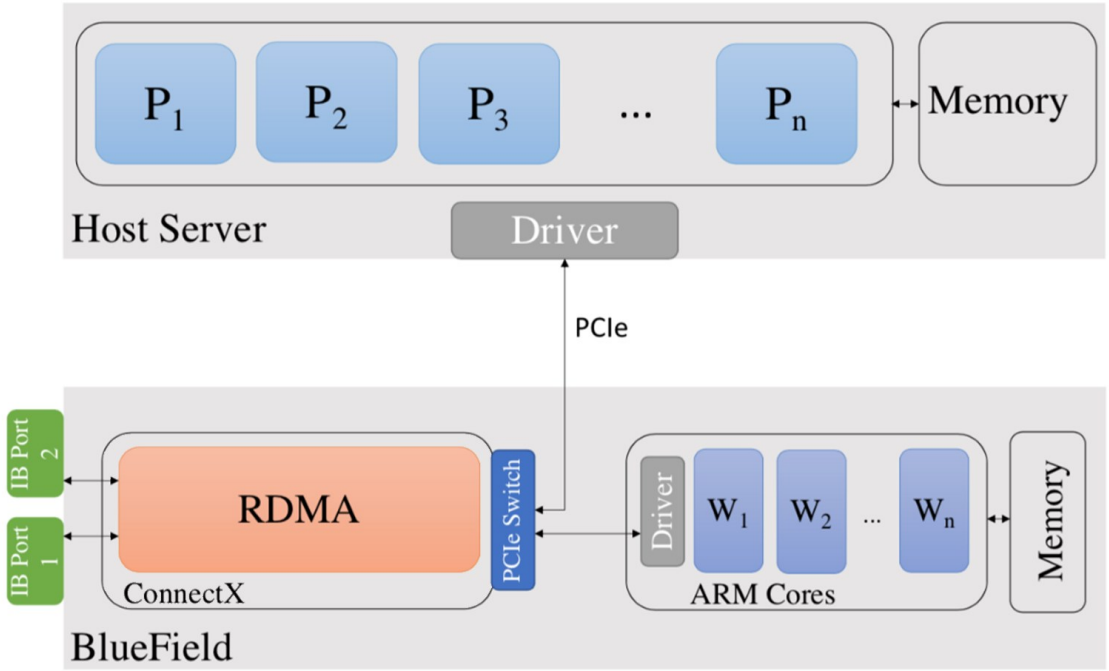
Requirements for Next-Generation MPI Libraries

- Message Passing Interface (MPI) libraries are used for HPC and AI applications
- Requirements for a high-performance and scalable MPI library:
  - Low latency communication
  - High bandwidth communication
  - Minimum contention for host CPU resources to progress non-blocking collectives
  - High overlap of computation with communication
- CPU based non-blocking communication progress can lead to sub-par performance as the main application has less CPU resources for useful application-level computation

Network offload mechanisms are gaining attraction as they have the potential to completely offload the communication of MPI primitives into the network

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



How to Re-design an MPI library to take advantage of DPUs and accelerate scientific applications?

MVAPICH2-DPU Library 2021.08 Release

- Based on MVAPICH2 2.3.6
- Released on 08/22/2021
- Supports all features available with the MVAPICH2 2.3.6 release (http://mvapich.cse.ohio-state.edu)
- Novel frameworks to offload non-blocking collectives to DPU
    - Alltoall (MPI_Ialltoall)
    - Allgather (MPI_Iallgather)
    - Broadcast (MPI_Ibcast)

MVAPICH2-DPU Library 2021.08 Release (Cont'd)

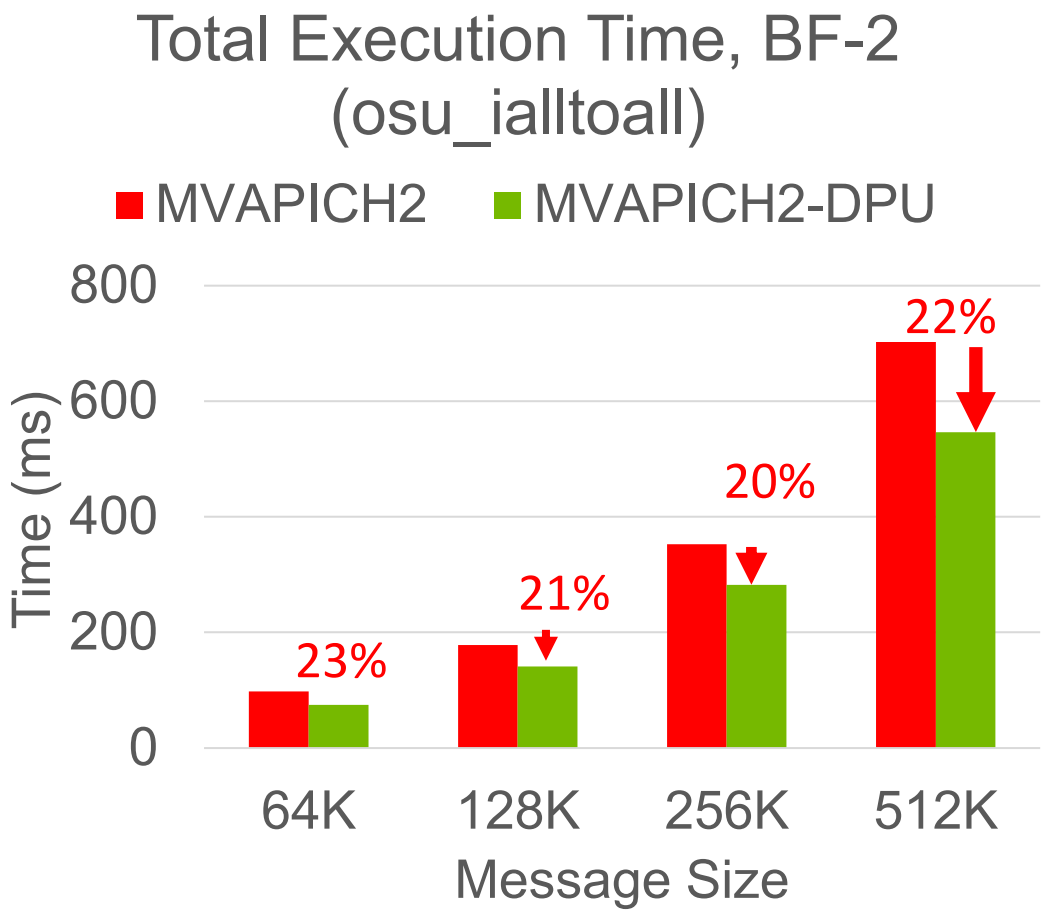
- Significantly increases (up to 100%) overlap of computation with any mix of MPI_Ialltoall, MPI_Iallgather, or MPI_Ibcast non-blocking collectives
- Accelerates scientific applications using any mix of MPI_Ialltoall , MPI_Iallgather, or MPI_Ibcast non-blocking collectives

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

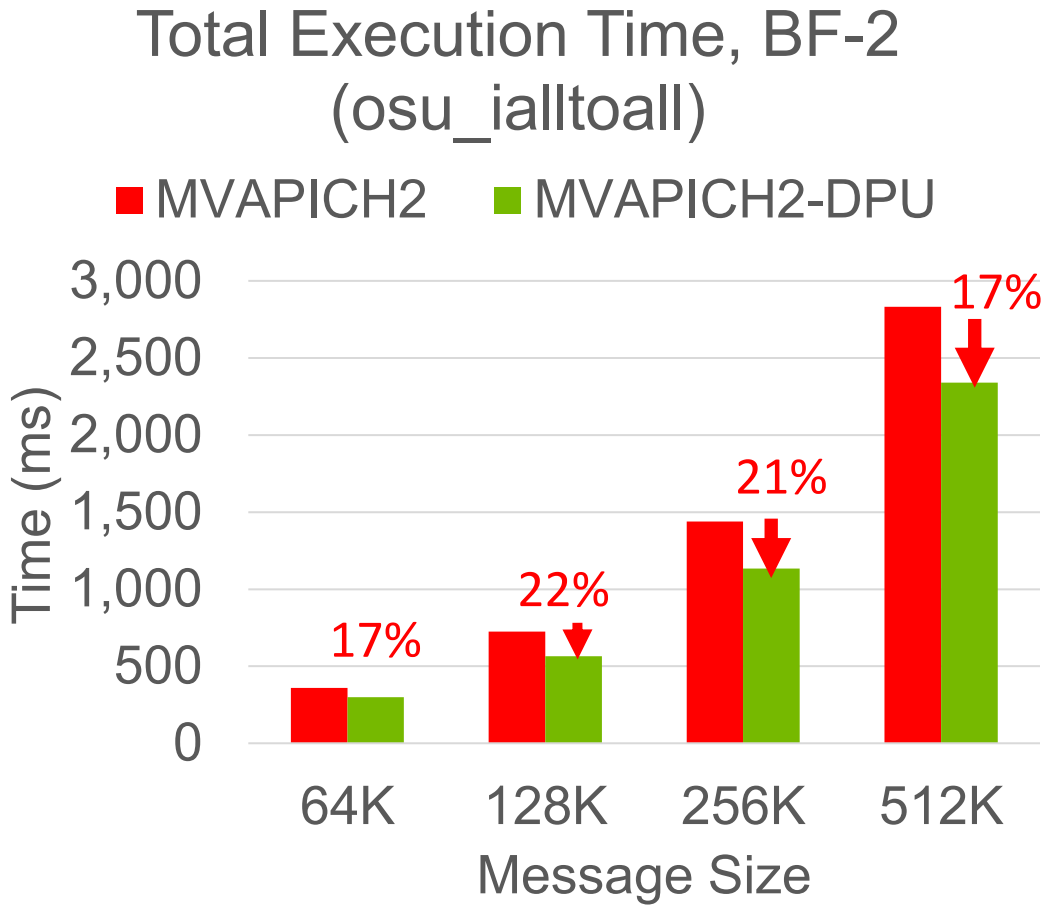
Evaluation Setup

- Being run on the HPC-AI Advisory Council cluster
  - 32 Xeon nodes connected with 32 DPUs over 200Gbps InfiniBand
  - 1,024 CPU cores (Xeons) and 256 ARM cores (DPUs)
- Configuration
  - Server HW:
    - CPU: Dual Socket Intel® Xeon® 16-core CPUs E5-2697A V4 @ 2.60 GHz
    - Adapter: Nvidia BlueField-2 DPU, 8 ARM cores 2.75 GHz, 16GB DDR4
  - Software/Firmware:
    - OS version: CentOS 8.3
    - Driver version: 5.2-1
    - Firmware version : 24.30.1004
  - MPI:
    - MVAPICH2-DPU 2021.08
  - OSU Micro-Benchmarks (OMB) 5.7.1

Total Execution Time with osu_Ialltoall (32 nodes)



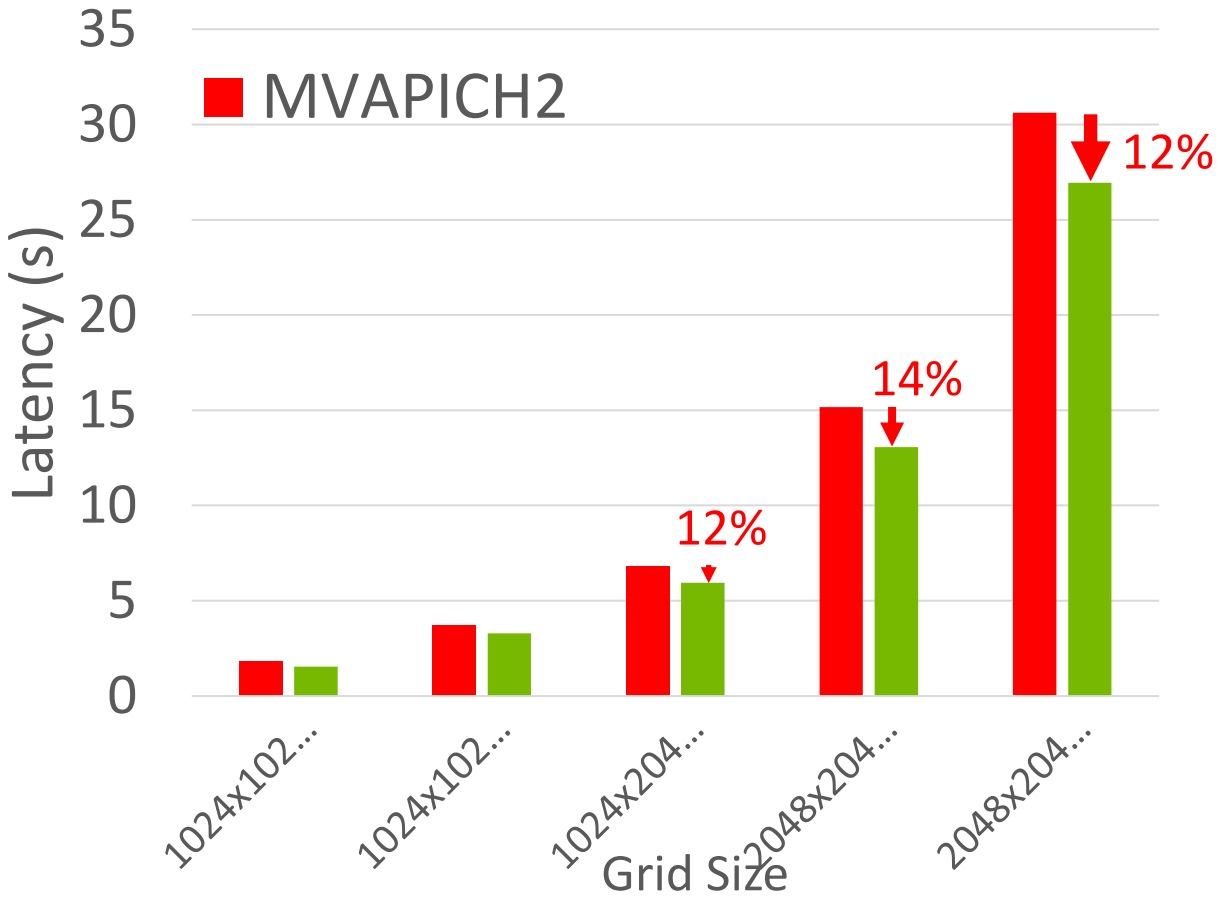
32 Nodes, 16 PPN



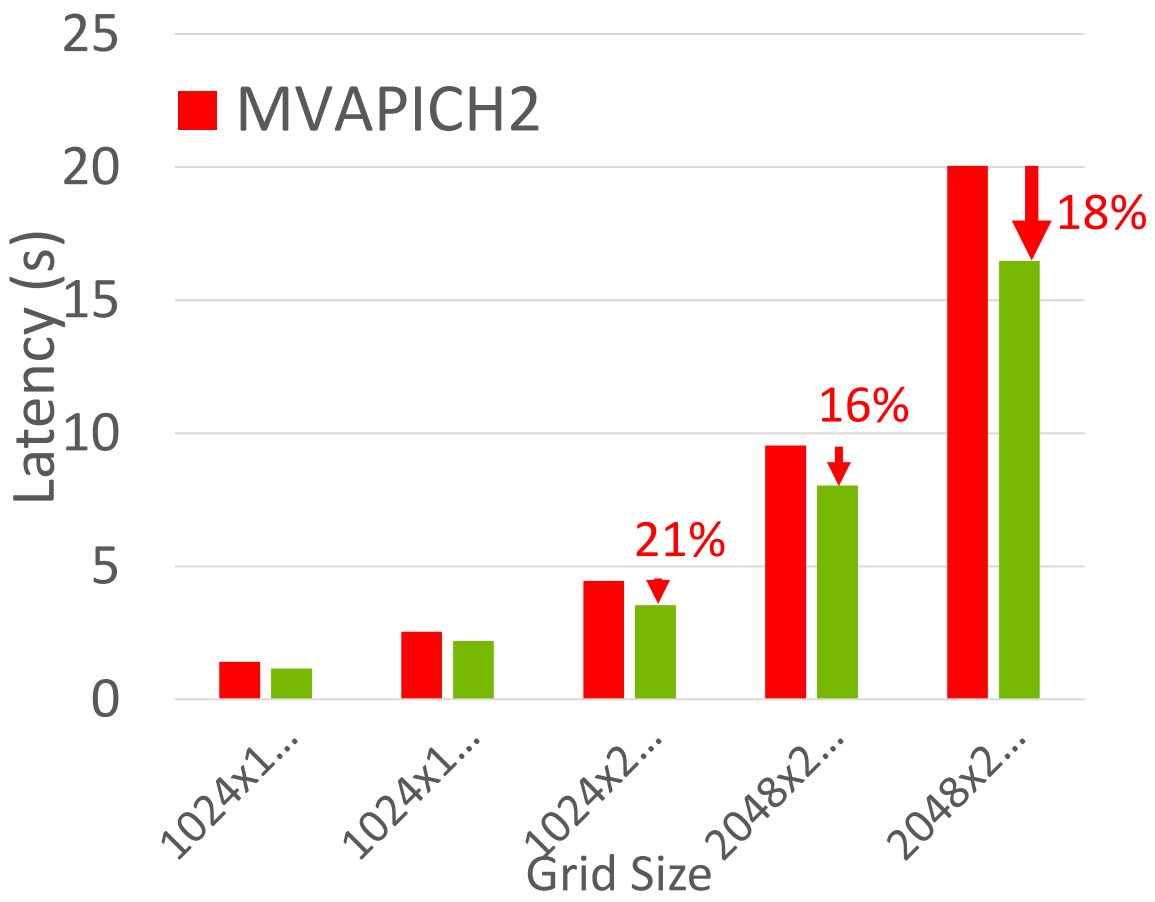
32 Nodes, 32 PPN

Benefits in Total execution time (Compute + Communication)

P3DFFT Application Execution Time (32 nodes)



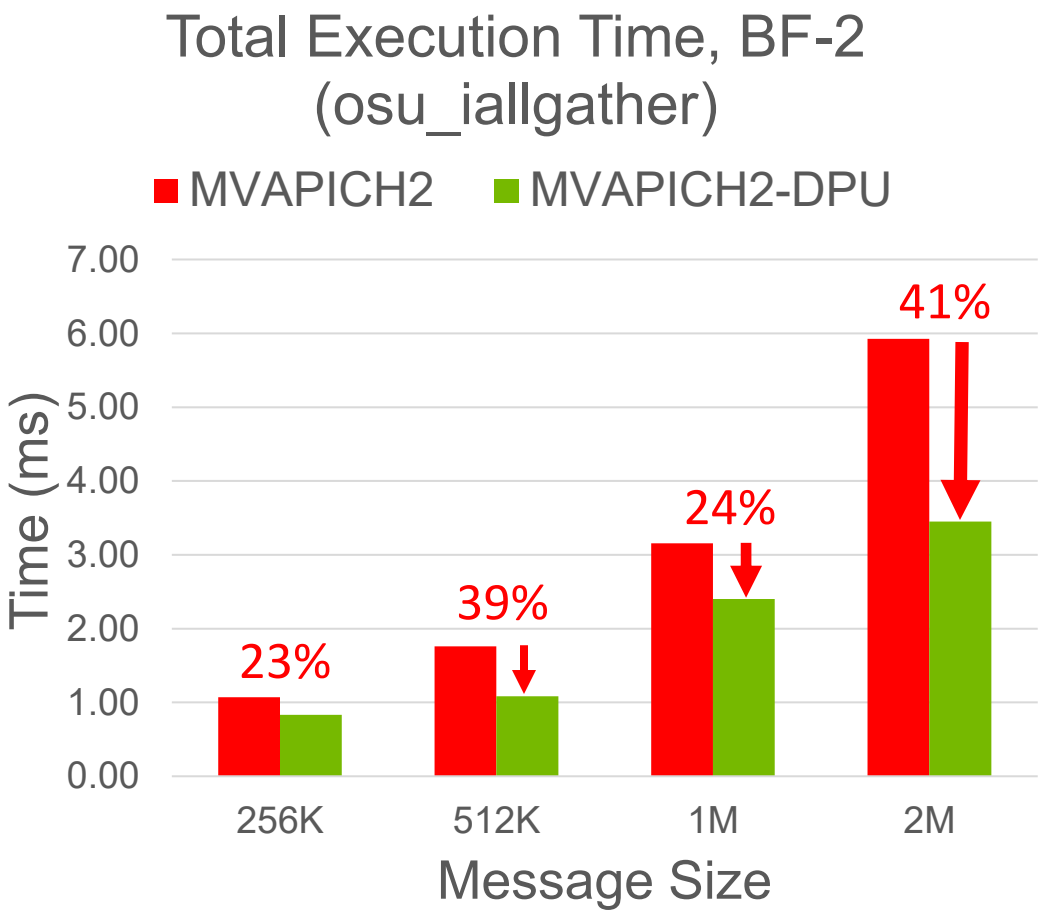
32 Nodes, 16 PPN



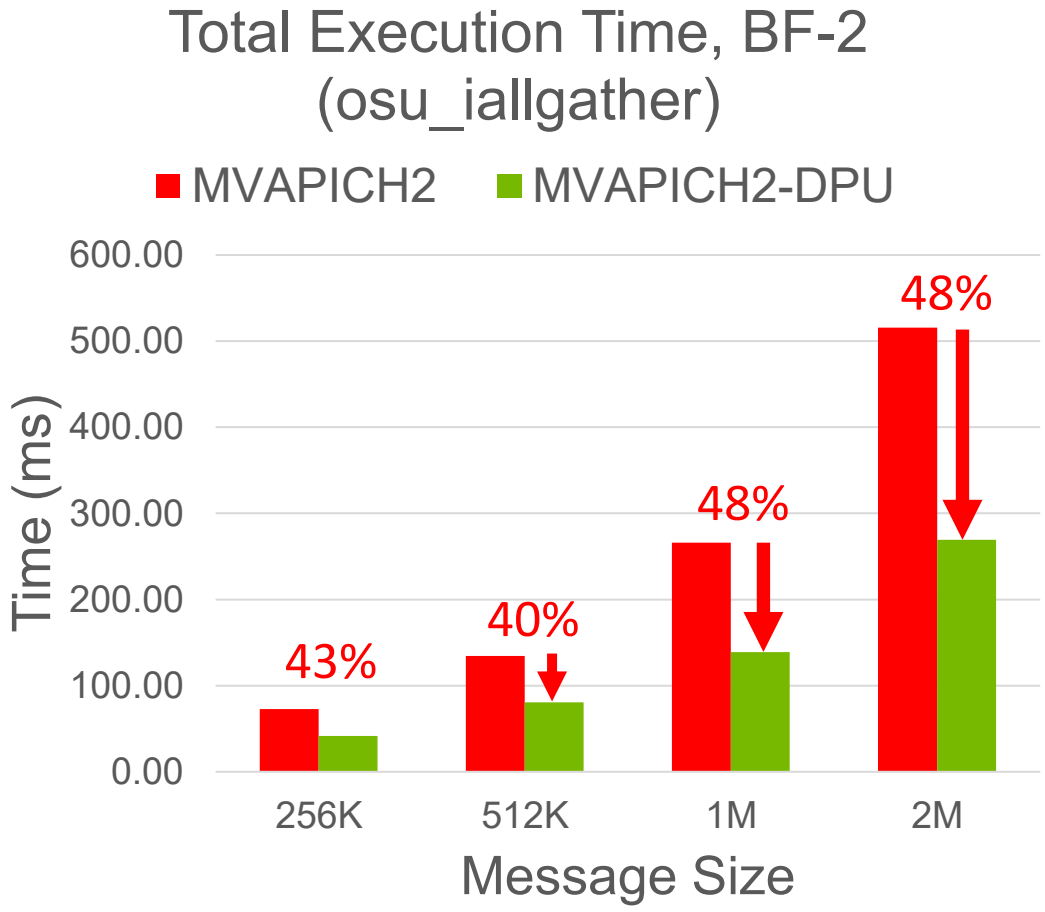
32 Nodes, 32 PPN

Benefits in application-level execution time

Total Execution Time with osu_Iallgather (16 nodes)



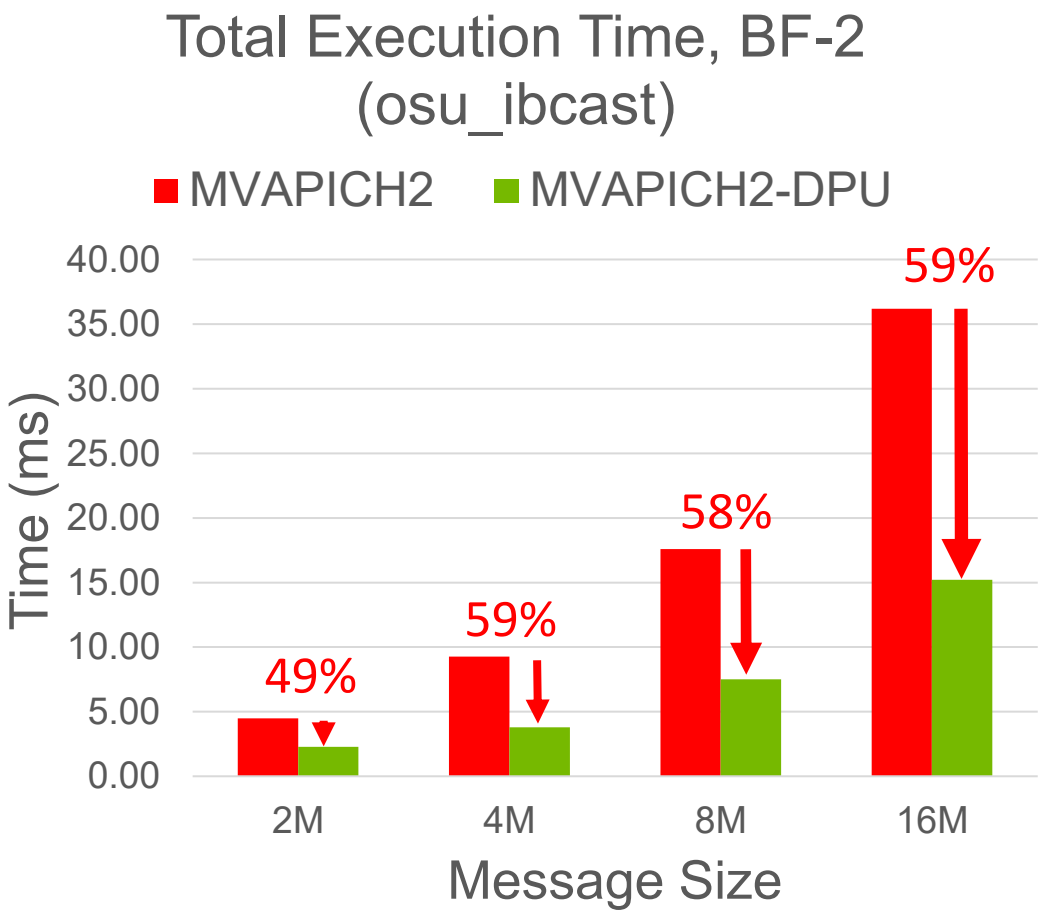
16 Nodes, 1 PPN



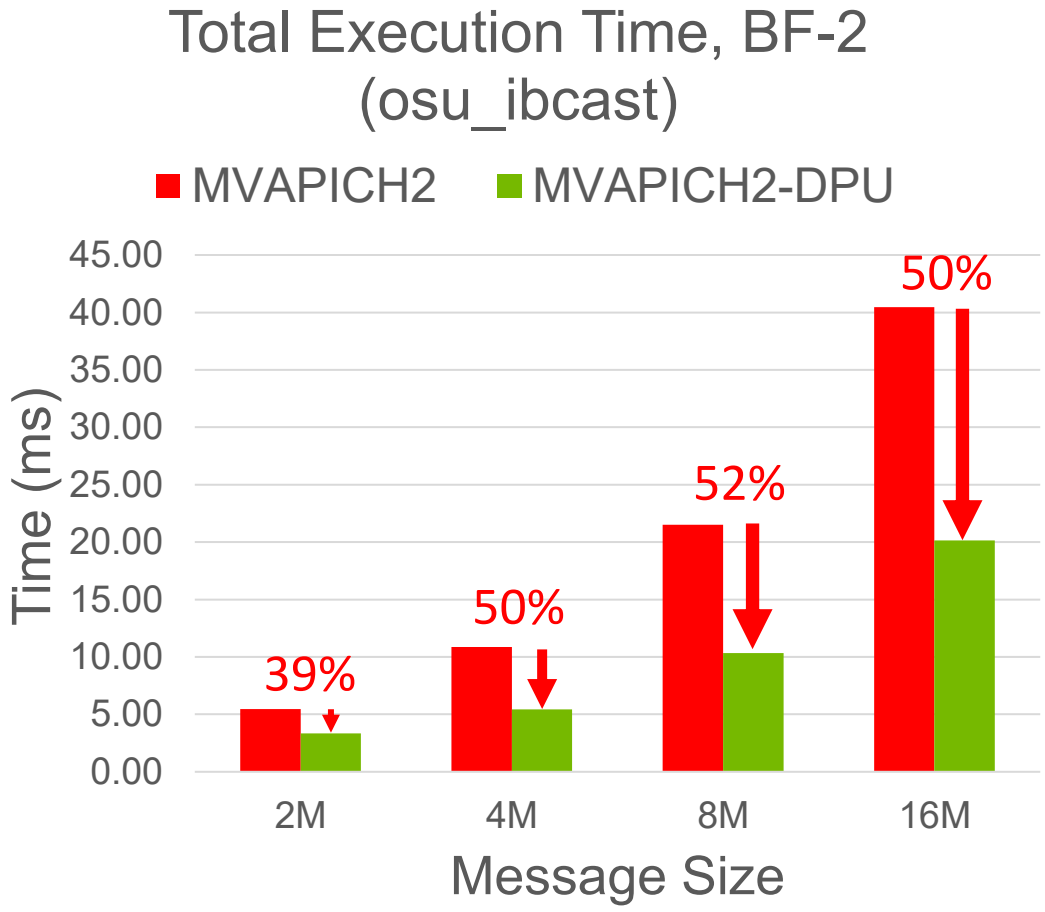
16 Nodes, 16 PPN

Benefits in 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)

Support to Accelerate DL Training in Next Release

- Support for distributed CPU-based DL training using NVIDIA Bluefiled-2 DPUs
- Intelligent designs to offload different phases of DL training
- Up to 15% performance improvement in DL training time compared to without DPU offloading
- Support for PyTorch/Torchvision and user defined DNN models and datasets

The design is based on a recent research paper “Accelerating CPU-based Distributed DNN Training on Modern HPC Clusters using BlueField-2 DPUs” by A. Jain, N. Alnaasan, A. Shafi, H. Subramoni, D. Panda, 28th IEEE Hot Interconnects, Aug 2021

Outline

- Overview of X-ScaleSolutions
- X-ScaleHPC and X-ScaleAI packages
- MVAPICH2-DPU: High-Performance MVAPICH2 for Accelerating Applications with NVIDIA's DPU technology
- SCR-Exa: Efficient and Scalable Checkpointing for HPC and DL Applications

Checkpointing for DL Frameworks & Applications

- Periodically saving snapshots of a DL model training a important for tolerating system failures
  - DL model training often requires long time to complete (sometimes, weeks or months)
  - Distributed DL model training at scale is more susceptible to system failures
- Single-machine DL training may simply load/store the DL model every *N* epochs
- For distributed training, the following (naïve) scheme is recommended:

```
0. for n in num_epochs:
1.     if rank == 0 and n % checkpoint_freq == 0:
2.         save_DNN()
3.     MPI_Barrier()
4.     …
5.     if rank == 0 and interruption:
6.         load_DNN()
7.     MPI_Bcast(DNN_params)
```

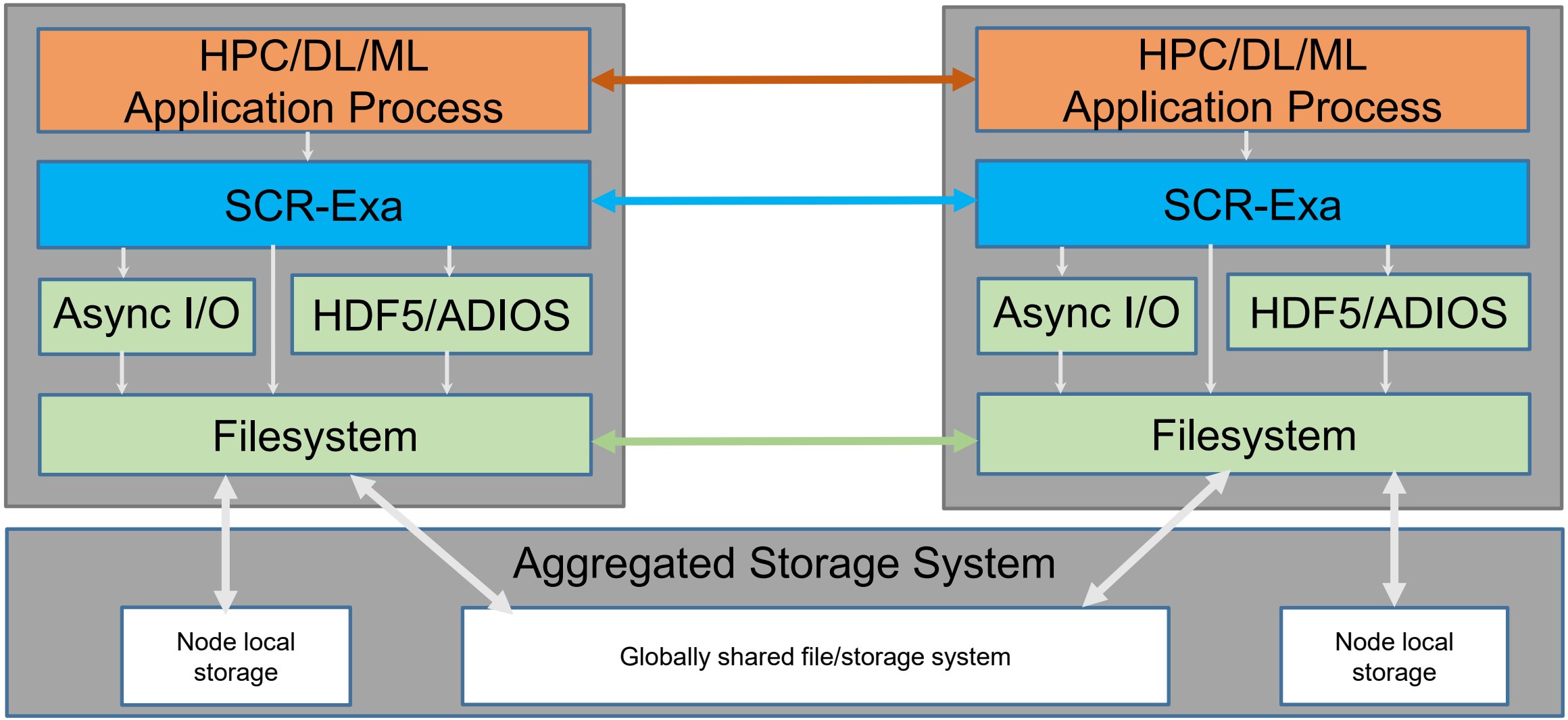
- This scheme requires all other ranks be blocked while rank 0 writes checkpoint info to the PFS

Challenge: Can we be more efficient in checkpointing for DL Training at scale?

SCR-Exa: Efficient Checkpointing for HPC and DL Applications

- Based on open-source Scalable Checkpoint Restart (SCR) library
- Developed in collaboration with Lawrence Livermore National Lab (LLNL)
- Significantly increase portability and flexibility
  - Add support for diverse job launchers, resource managers, storage devices, etc.
  - Customized and optimized for a range of systems with different underlying protocols
- Enable fast and efficient restart and resume
  - Add support for launching applications with spare nodes
  - Automatically reconfigure to restart or resume using spare nodes after a failure (if possible)
- Significantly improve maintainability and extensibility
  - Add new python binding and python implementation of CLIs and APIs
- Expand support for DL/ML frameworks and applications
  - PyTorch, horovod, ResNet, EDSR, etc.
- Part of code enhancements are contributed back to the SCR open-source

Overview of SW Stack for Enabling SCR-Exa library

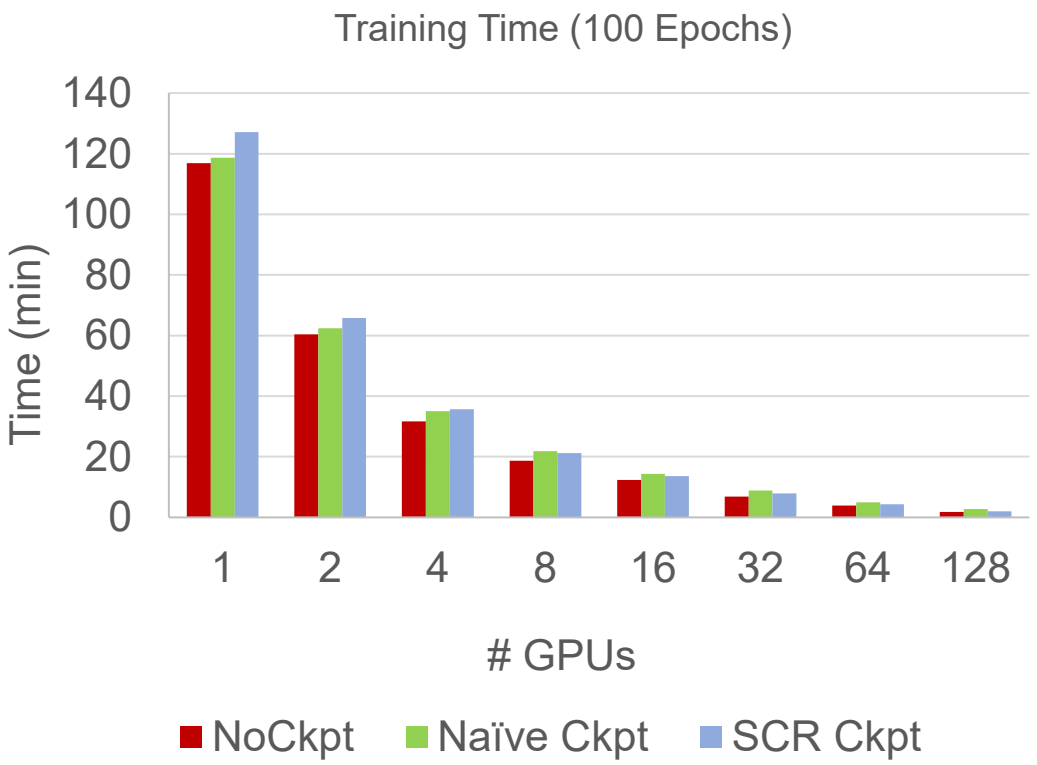


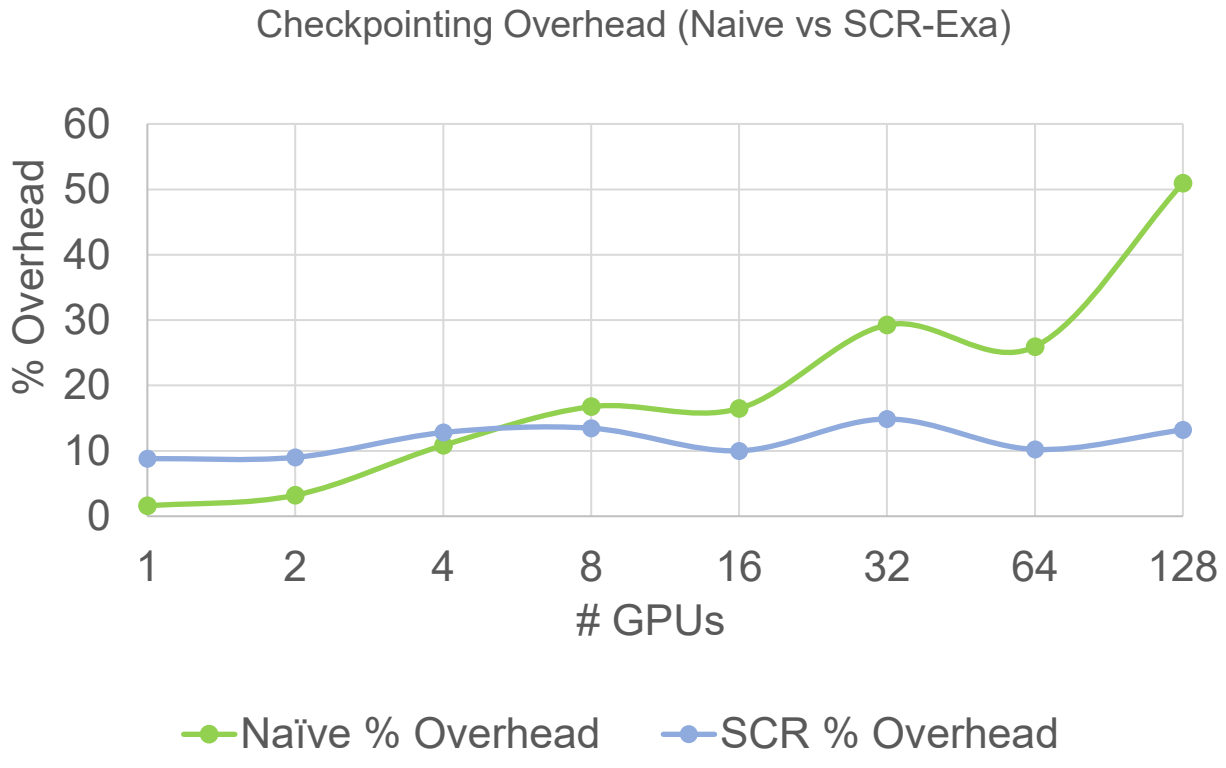
Enabling SCR-Exa for DL Frameworks and Applications

- Instrumented version of DL platforms to use SCR_Exa library:
  - PyTorch Distributed Data Parallel Model (DDP)
  - PyTorch Over Horovod
- DL Applications:
  - Residual Neural Network (ResNet)
  - Enhanced Deep Residual Networks for Single Image Super-Resolution (EDSR)

Use of SCR-Exa for DL Applications (ResNet-50)

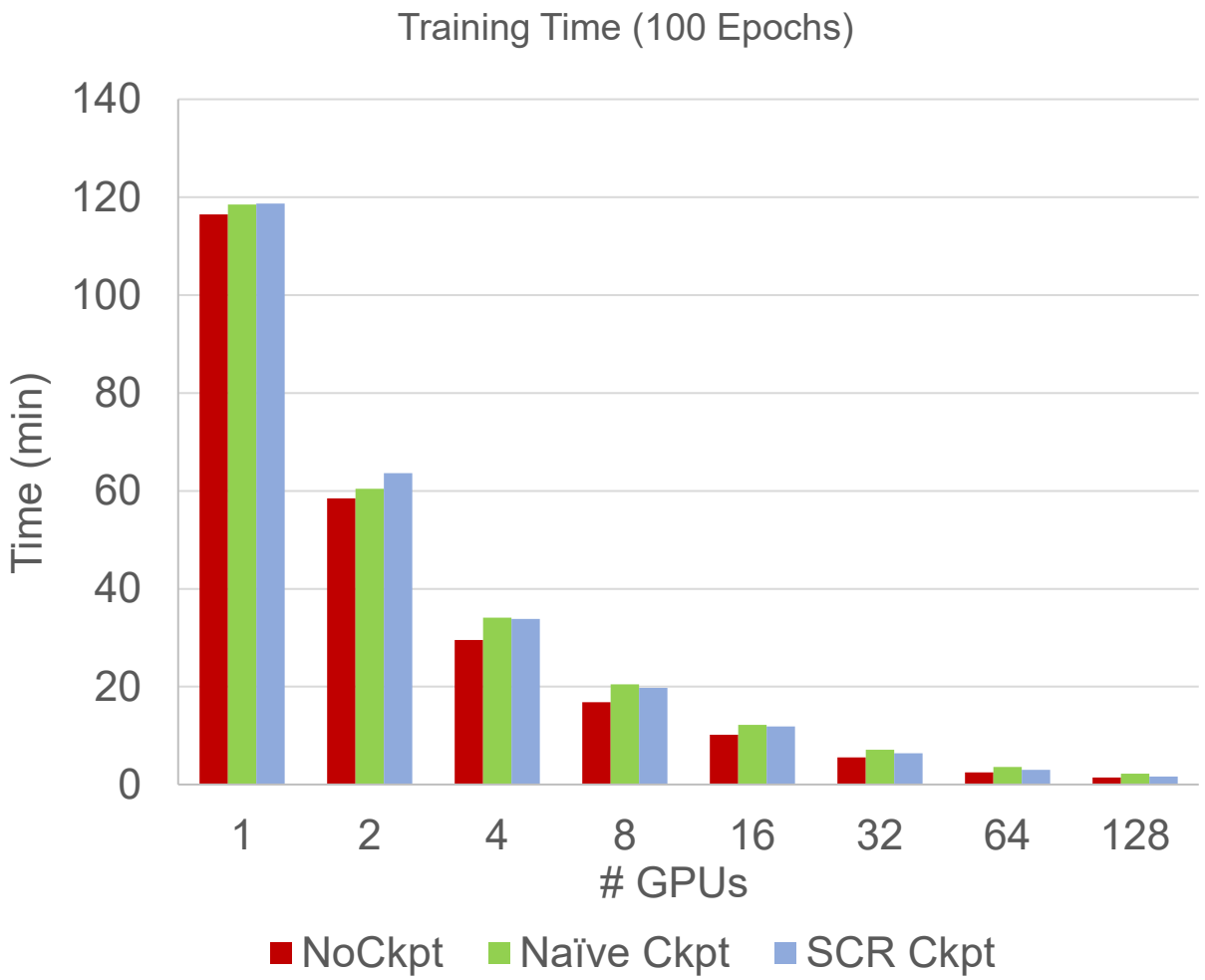
- Use PyTorch DDP platform
- Checkpoints saved every epoch in both naïve and SCR-Exa approaches
- SCR-Exa saves checkpoints to the local node, and only write to the PFS every 10 epochs.
- SCR-Exa is very efficient (about 10%-15% overhead) and scales very well (OpenPOWER9, V100 GPU)

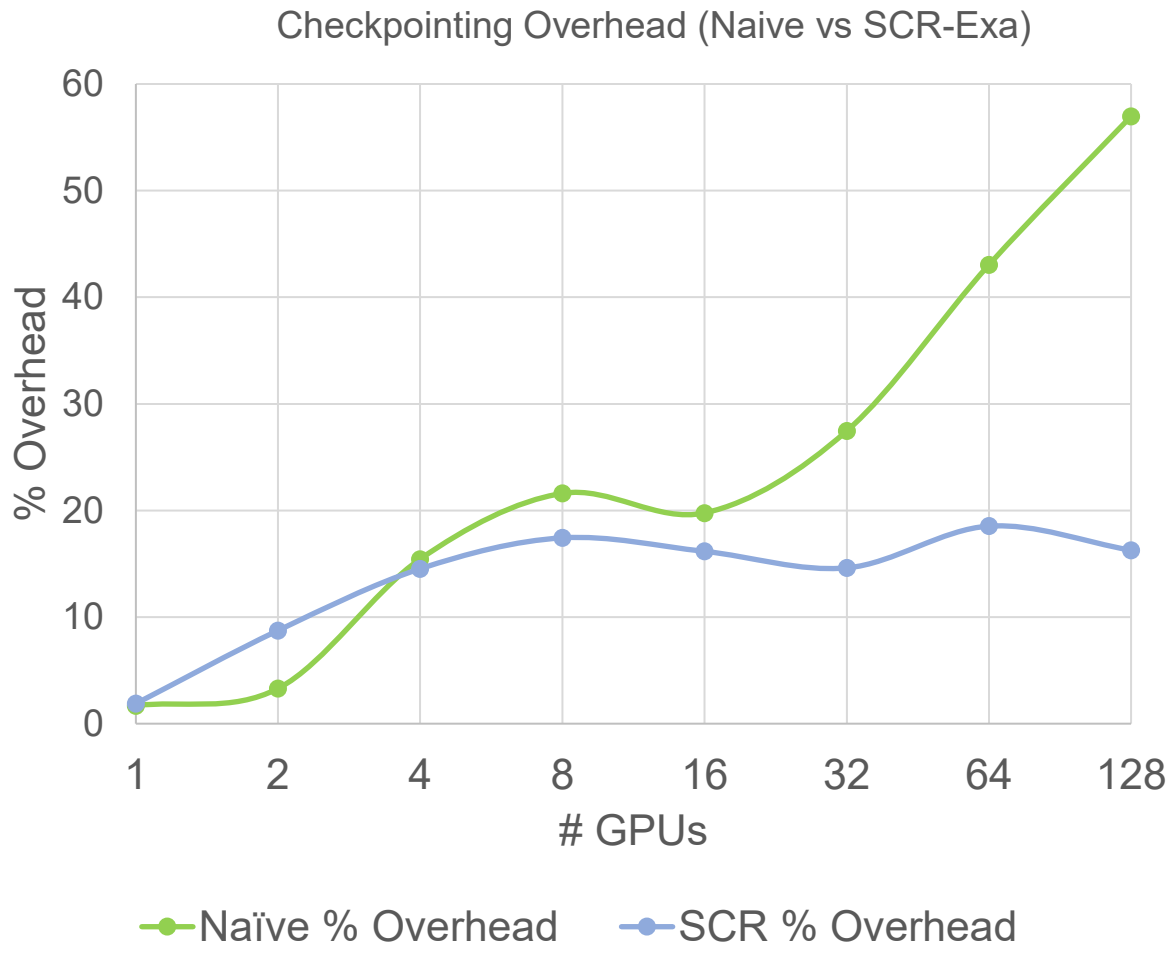




Enabling SCR-Exa for DL Applications (ResNet-50, Cont'd)

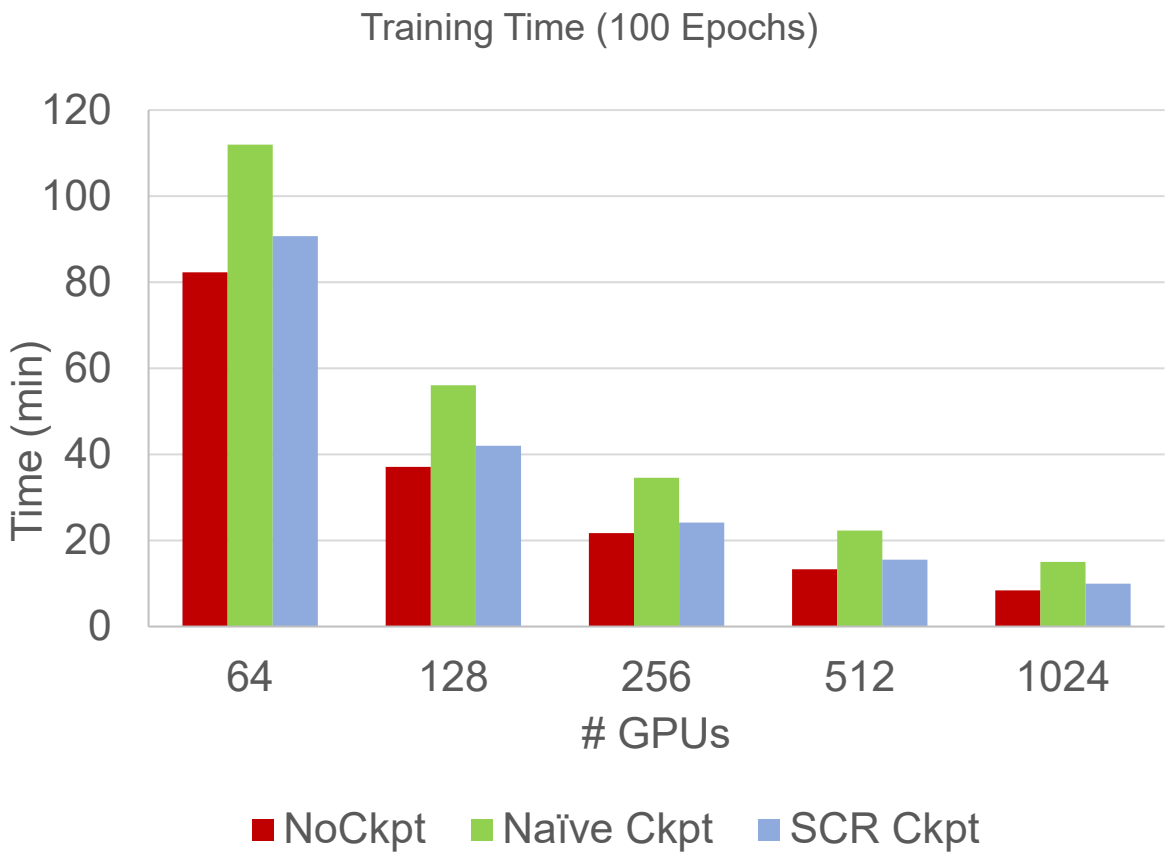
- Similar performance trends observed for the PyTorch over Horovod platform

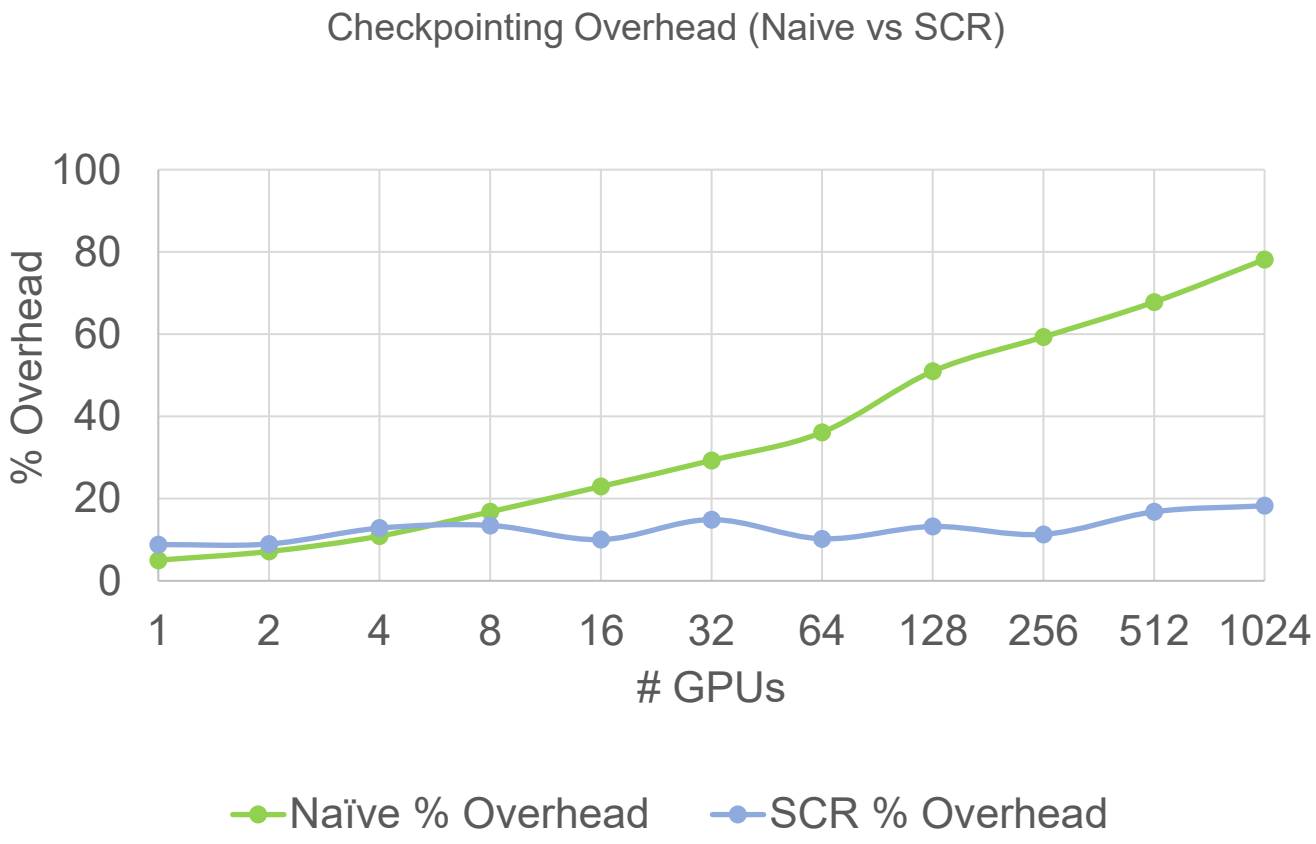




Enabling SCR-Exa for DL Applications (EDSR)

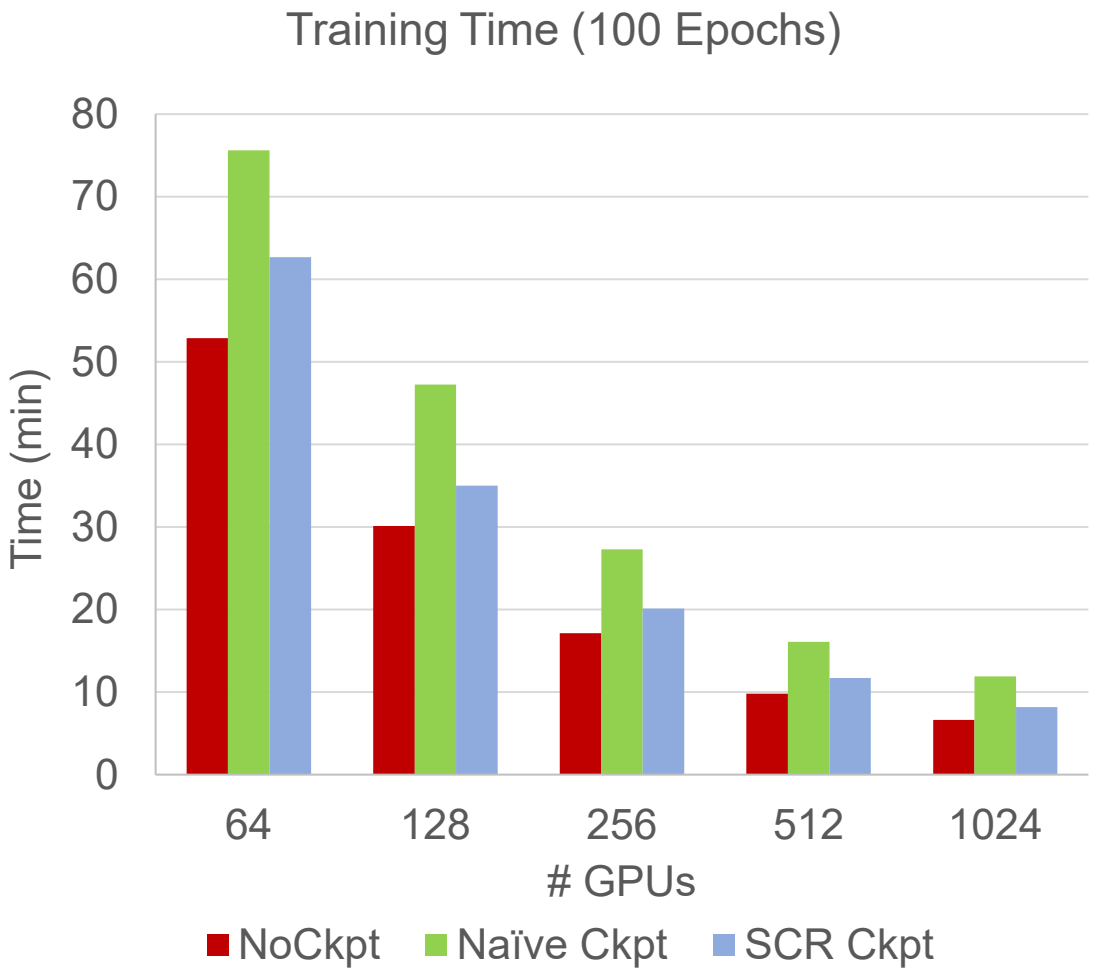
- Use TyTorch DDP platform
- Same training parameters as ResNet-50, but scaled to 1024 GPUs

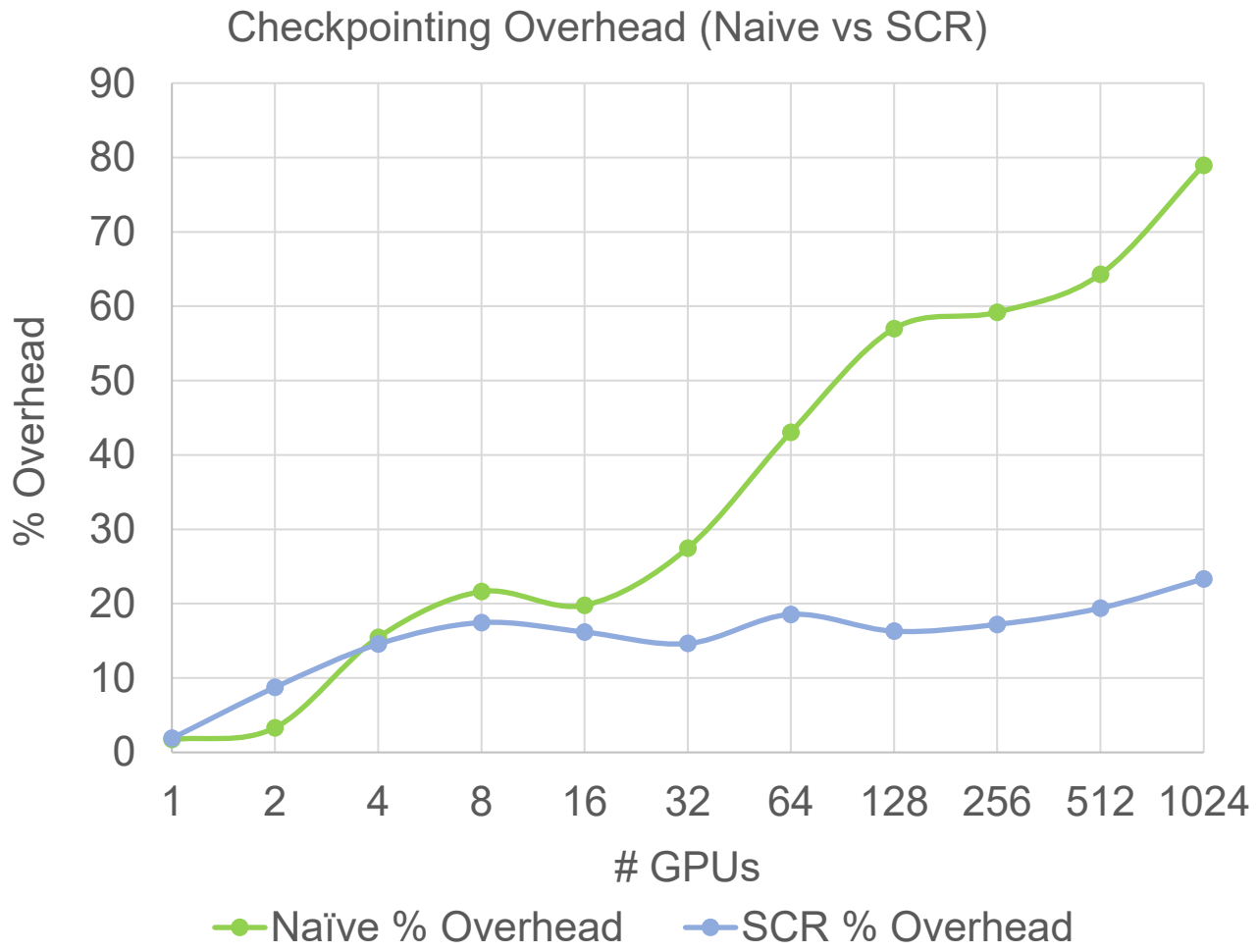




Enabling SCR-Exa for DL Applications (EDSR, Cont'd)

- Similar performance trends observed for the PyTorch over Horovod platform





Thank You!

d.dai@x-scalesolutions.com

http://x-scalesolutions.com/