Tutorial and Demo

Overview of X-ScaleAI and X-ScaleSecureMPI Products for AI and HPC

Donglai Dai, Mohsen Gavahi, and Kyle Schaefer

X-ScaleSolutions

http://x-scalesolutions.com

Outline

- Overview of X-ScaleSolutions
- X-ScaleAI: High-Performance Solution for Various AI Applications/Models

  Pre-train, Finetune, and Inference
- X-ScaleSecureMPI: Secured Communication for HPC Applications with

  minimal overhead

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
  - Platform-specific optimizations and tuning
  - Application-specific optimizations and tuning
  - Obtaining guidelines on best practices
  - Timely support for installation and operational issues encountered with the library
  - Flexible Service Level Agreements
  - Web portal interface to submit issues and tracking their progress
  - Information on major releases and periodic information on major fixes and updates
  - Help with upgrading to the latest release
- Winner of multiple U.S. DOE SBIR grants
- Market these products for HPC and AI applications with commercial support

Overview of X-ScaleSolutions (cont'd)

- Currently, we offer five major products with commercial support:
  - MVAPICH2-DPU (https://x-scalesolutions.com/mvapich2-dpu/)
  - X-ScaleHPC (https://x-scalesolutions.com/x-scalehpc/)
  - X-ScalePETSc (https://x-scalesolutions.com/x-scalepetsc/)
  - X-ScaleAI (https://x-scalesolutions.com/x-scaleai/)
  - X-ScaleSecureMPI (https://x-scalesolutions.com/x-scalesecurempi/)
- More information about the specific features and capabilities of these products are available on the websites provided above
- Today's tutorial will focus on the last two products: X-ScaleAI and X-ScaleSecureMPI

X-ScaleSolutions will give a presentation at 2 pm ET on Wednesday Aug 21 that will go into more performance results of our products. Please come and join us.

Outline

- Overview of X-ScaleSolutions
- X-ScaleAI: High-Performance Solution for Various AI Applications/Models

  Pre-train, Finetune, and Inference
- X-ScaleSecureMPI: Secured Communication for HPC Applications with

  minimal overhead

X-ScaleAI: Features and Capabilities

Goal

- High-performance solution for AI problems on modern HPC and Cloud platforms
  - 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 will be available (as a service) for new users and organizations
- Pre-training, Fine Tuning, and Inference for multiple domains
- 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

- 30+ years of expertise in HPC technologies to enable scale-up and scale-out for AI applications
- 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

Ease-of-use in Configuration to Harness Performance

config_file.json (Example #1)

```
{“model”: “resnet50”,
“optimizer”: “SGD”,
“batch_size”: 64,
“exp_name”: “benchmark”,
“num_epochs”: 90,
“learning_rate”: 1e-3,
“data_loader” : {
"readers.file": { "random_shuffle":
true, "pad_last_batch": true,
"name": ”benchmark-read"},
"decoders.image_random_crop": {
"device": "mixed", "output_type":
"RGB”},
"resize": { "device": "gpu",
"resize_x": 224, "resize_y": 224}
}
}
```

config_file.json (Example #2)

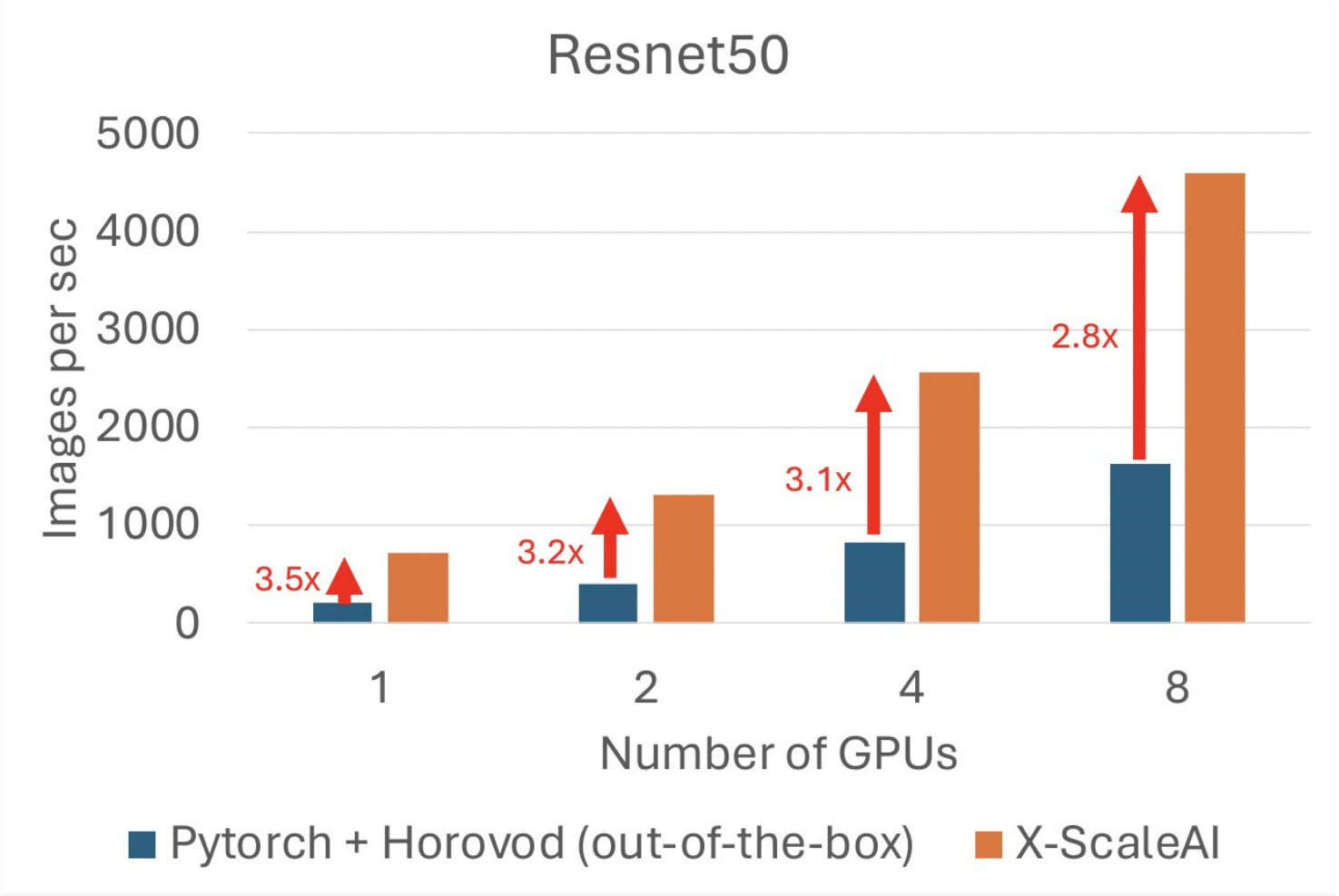
```
{“model”: “cheXpert”,
“optimizer”: AdamW,
“batch_size”: 2,
“exp_name”: “healthcare-ai”,
“num_epochs”: 40,
“learning_rate”: 2e-3,
“data_loader” : {
"readers.file": { "random_shuffle":
true, "pad_last_batch": true,
"name": ”healthcare-ai-read"},
"decoders.image_random_crop": {
"device": "mixed", "output_type":
"RGB”, "device_memory_padding":
211025920, "host_memory_padding" :
140544512 },
"resize": { "device": "gpu",
"resize_x": 224, "resize_y": 224}
}
}
```

Change config_file.json and simply run in the jobscript:

```
xscale-ai-run image_xscaleai.py --config config_file.json
```

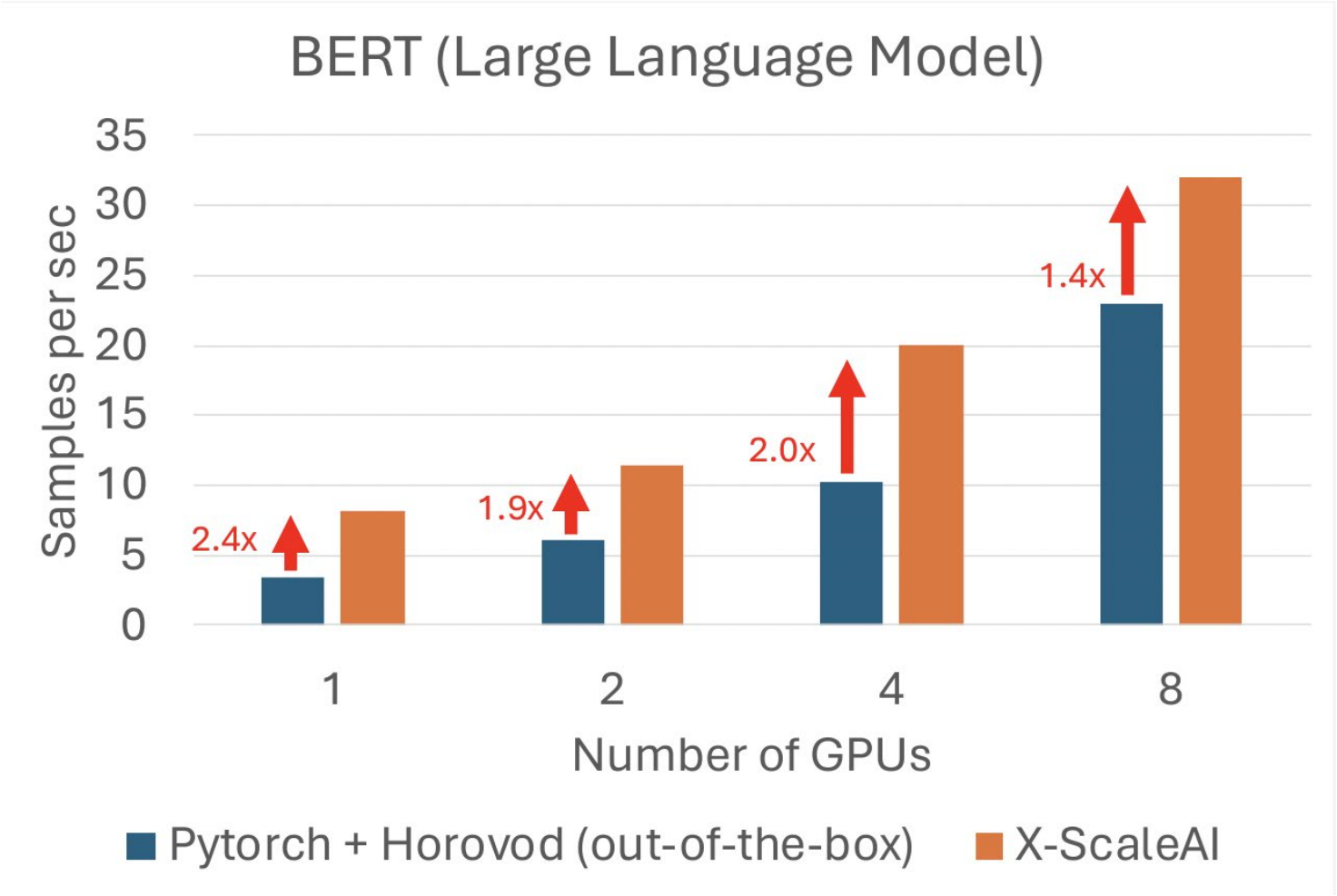
X-ScaleAI: Distributed PyTorch on Sample System #1

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



X-ScaleAI: Distributed PyTorch on Sample System #2

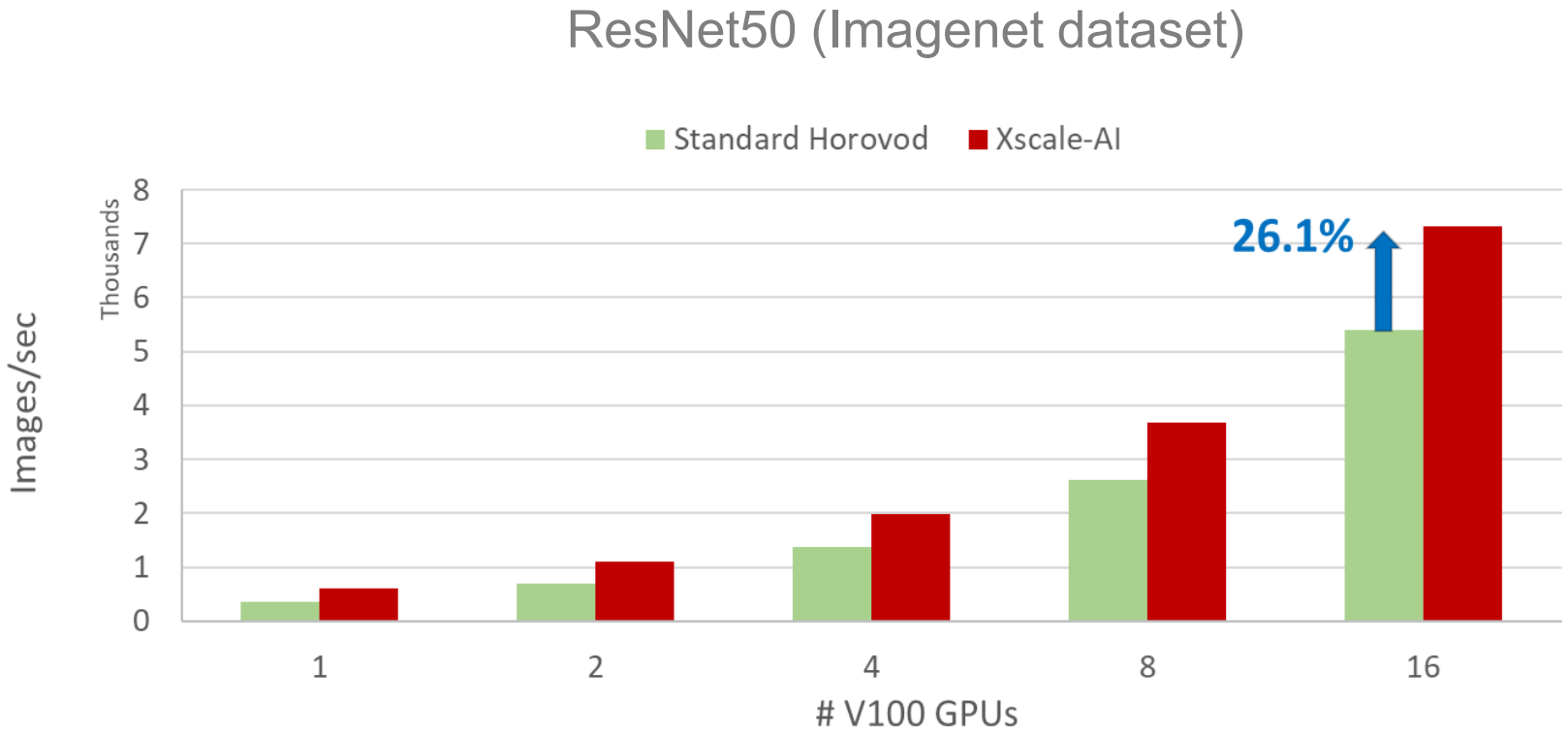
- Sentence classification
  - BERT
- Cloud:
  - 2 AWS g4dn.16xlarge instances
- GPU:
  - 4 NVIDIA T4's per instance
- Interconnect:
  - Elastic Fabric Adapters (EFA) – 50 Gb/s



X-ScaleAI: Distributed PyTorch on Sample System #3

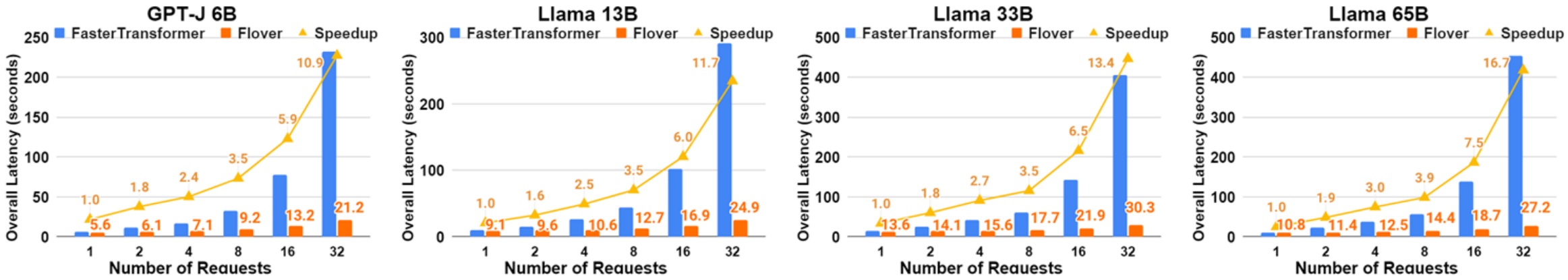
System Configuration #3:

- GPUs:
  - Four V100s per node with NVLink
- CPUs:
  - 20 cores, 2.50GHz
  - Two Intel Xeon Gold 6248 “Cascade Lake”
- 384GB of RAM: DDR4
- 1.6TB NVMe PCIe SSD
- Interconnect:
  - Two Mellanox ConnectX-6 InfiniBand HDR 200Gb/s Adapters



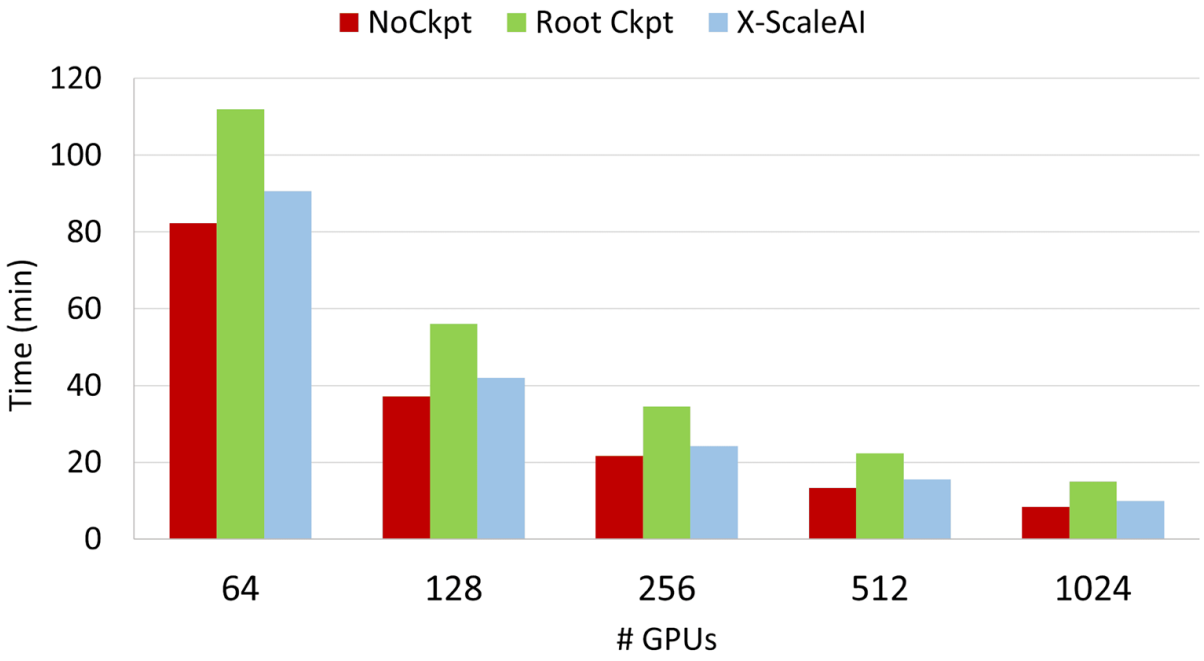
Parallel Inference on LLMs with Temporal Fusion

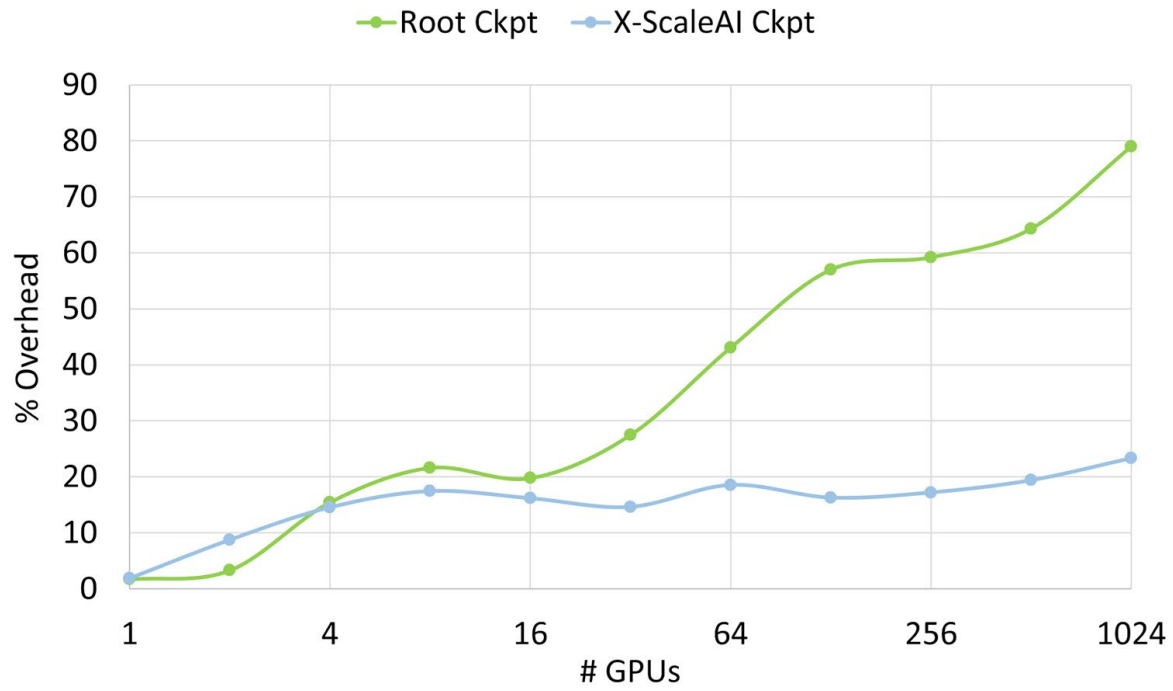
- Developed based on Flover, an inference framework when processing multiple requests in parallel.
- For all models, the average inference latency for a single request is >> default time interval
  - For GPT-J 6B and Llama 13B, we run on 1 GPU without tensor parallelism
  - For Llama 33B, we run on 2 GPUs with tensor parallelism of size 2
  - For Llama 65B, we use 4 GPUs to perform degree-4 tensor parallelism
- **Compared to FasterTransformer, our method achieves up to 16.7x speedup in latency to finish all requests**
- **We are integrating these optimizations among others into the X-ScaleAI package for efficient inference**



Scalable Checkpoint-Restart for DL Applications (ResNet50)

- 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





User Step #1: AWS Marketplace and EC2 Launch

1) Subscribe to the XScaleAI product through AWS Marketplace

- find XScaleAI by searching for the X-ScaleSolutions marketplace page
- subscribe XScaleAI with a pay as you go or with a trial version

2) Once subscribed, there are two ways to launch an instance of XScaleAI

First option is:

- Launch regular instances through the EC2 interface by naming the instance. Select the AWS Instance type with matching AMI. Currently any g4dn instance type is supported (more will be added soon)
- Select your SSH keys. Configure everything else as desired. Select the number of instances. Click Launch
- Once the instance is launched, wait for the initialization to complete. Then you can use the public up address listed and ssh into your machine.

```
ssh -i pkey ubuntu@<ip-address>
```

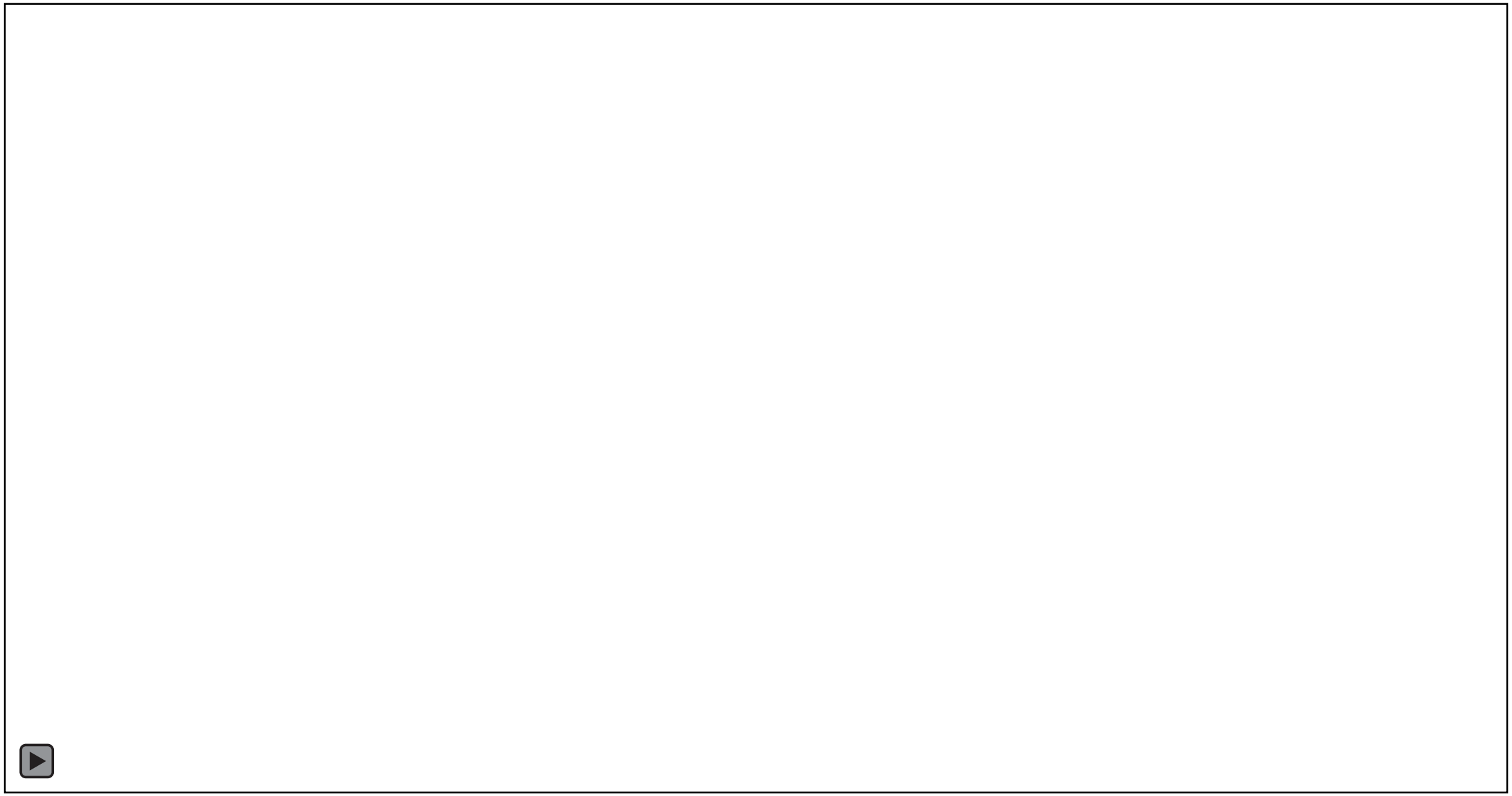
User Step #1: AWS Marketplace and EC2 Launch (cont.)

User Step #2: PCLUSTER Configuration and Creation

The second option to launch XScaleAI jobs is to use Pcluster (ie., parallel cluster)

- create a Pcluster configuration file
  - define the head node, cluster manager, compute nodes, etc.
- use Pcluster CLI to launch an XScaleAI run using the Pcluster configuration file. (This is very similar to running applications on HPC systems.)

User Step #2: PCLUSTER Configuration and Creation (cont.)



User Step #3: PCLUSTER slurm Allocation and Run

Connect to the head node and use slurm to manage your resources and jobs

- Allocate a compute node: "salloc -p gpu -N 1"
- Connect to the compute node: "ssh <ip-address>"
- Run application: "xscale-ai-run <application> <application parameters>"

  for example: "xscale-ai-run python example-vision.py --train-dir ./data/imagenette2-320/train --val-dir ./data/imagenette2-320/val --epochs 1 --batch-size 128"

Note: the "xscale-ai-run" command will automatically resolve things like your hostfile, number of processes, etc. By default it will use all of the resources configured in the job allocation.

User Step #3: PCLUSTER slurm Allocation and Run (cont.)

Outline

- Overview of X-ScaleSolutions
- X-ScaleAI: High-Performance Solution for Various AI Applications/Models

  Pre-train, Finetune, and Inference
- X-ScaleSecureMPI: Secured Communication for HPC Applications with minimal overhead

X-ScaleSecureMPI: Features and Capabilities

Goal

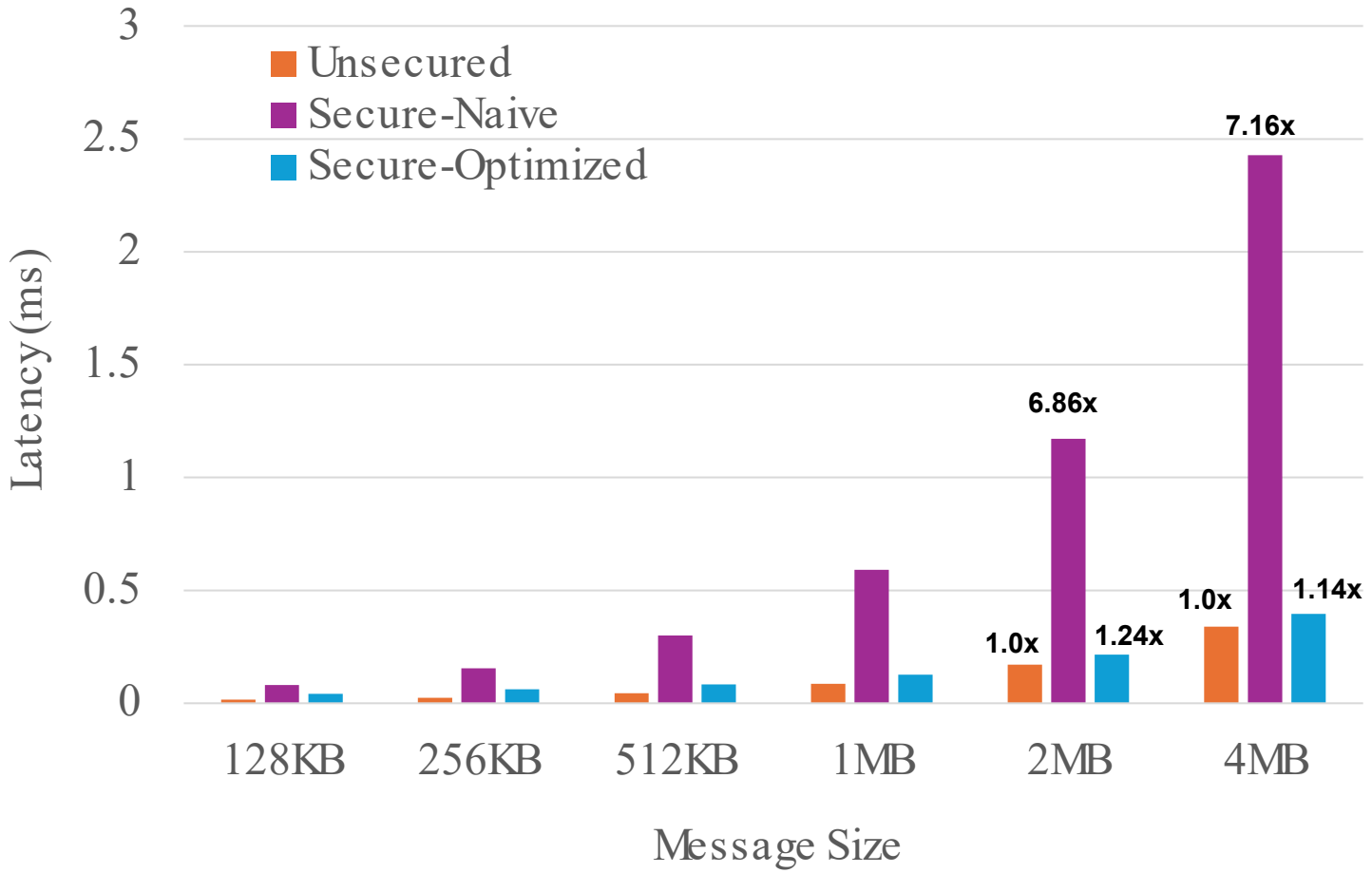
- Ensure secure communication in security-sensitive MPI applications running on modern HPC and cloud platforms with minimal performance overhead

Major Features

- Scalable solutions of secure communication middleware based on the OSU MVAPICH libraries
- 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

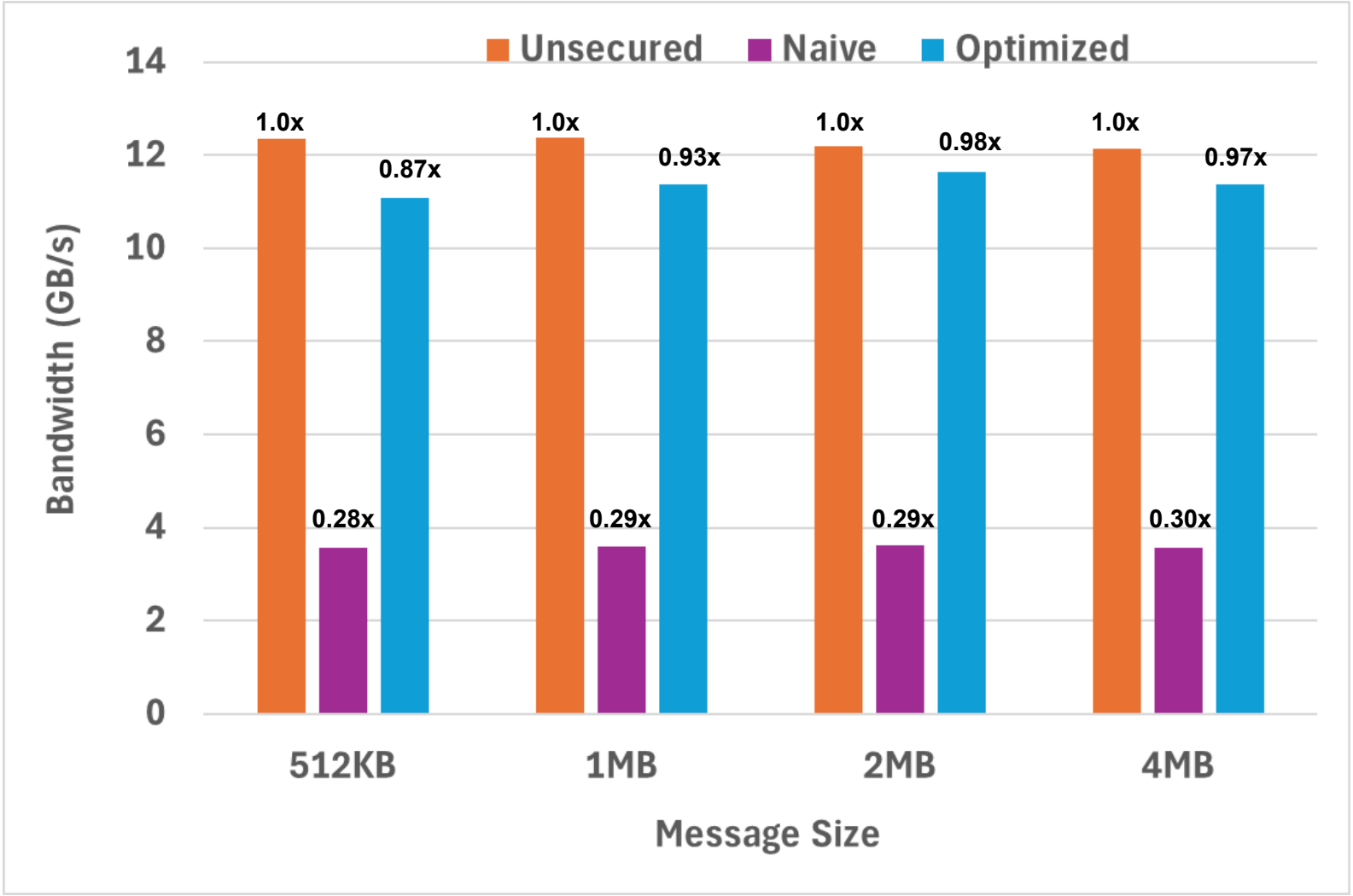
Performance: OSU_Latency Micro-benchmark

- Blocking point-to-point send/recv
- 2 nodes,1 ppn on an Intel cluster (Inter-node communication)
- Intel x86 CPU
- InfiniBand network
- Two popular encryption libraries supported:
  - BoringSSL+OpenSSL
  - Intel IPP+OpenSSL



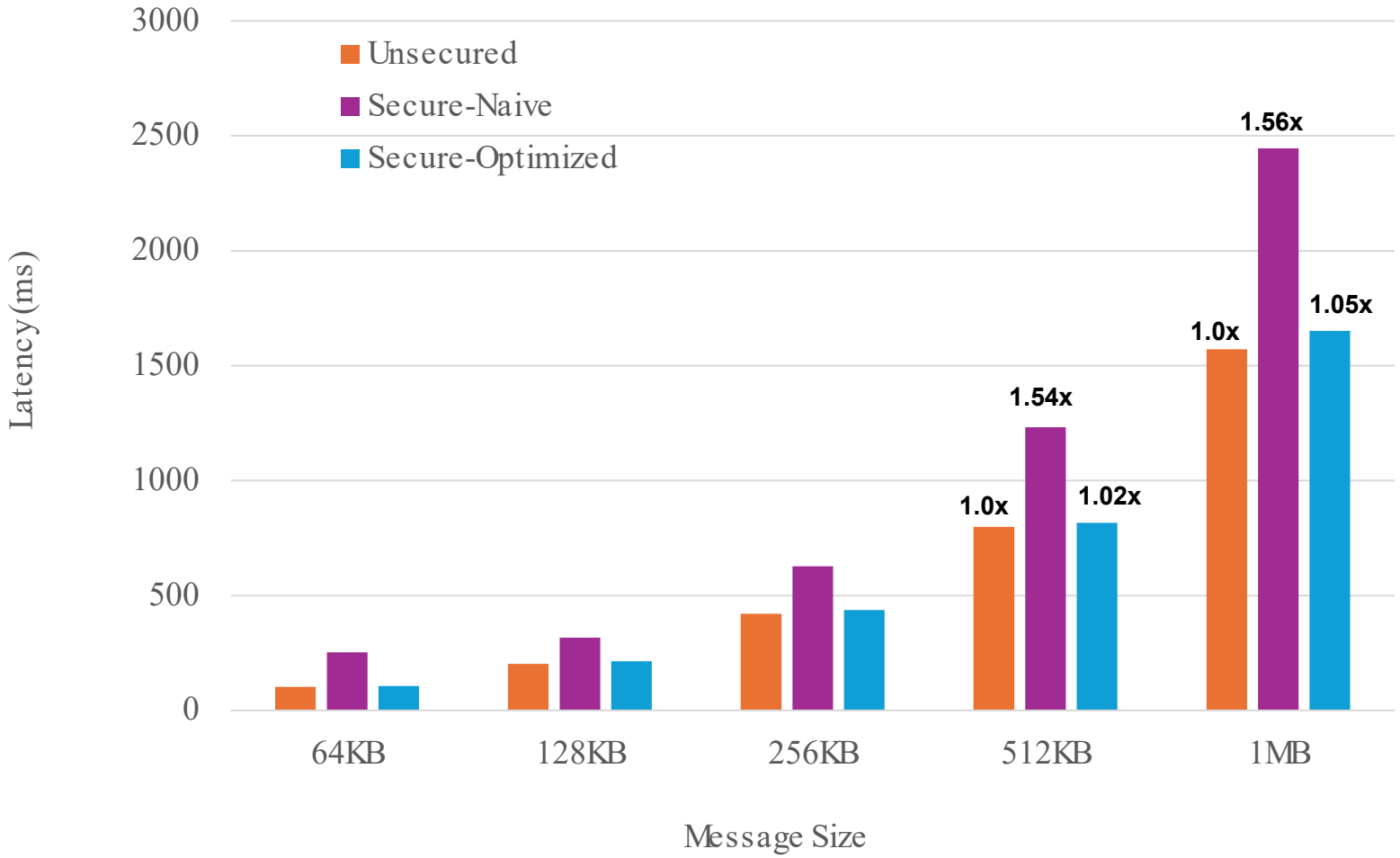
Performance: OSU_BW Micro-benchmark

- Non-Blocking point-to-point Isend/Irecv
- 2 nodes,1 ppn on an Intel cluster (Inter-node communication)
- Intel x86 CPU
- InfiniBand network
- Two popular encryption libraries supported:
  - BoringSSL+OpenSSL
  - Intel IPP+OpenSSL



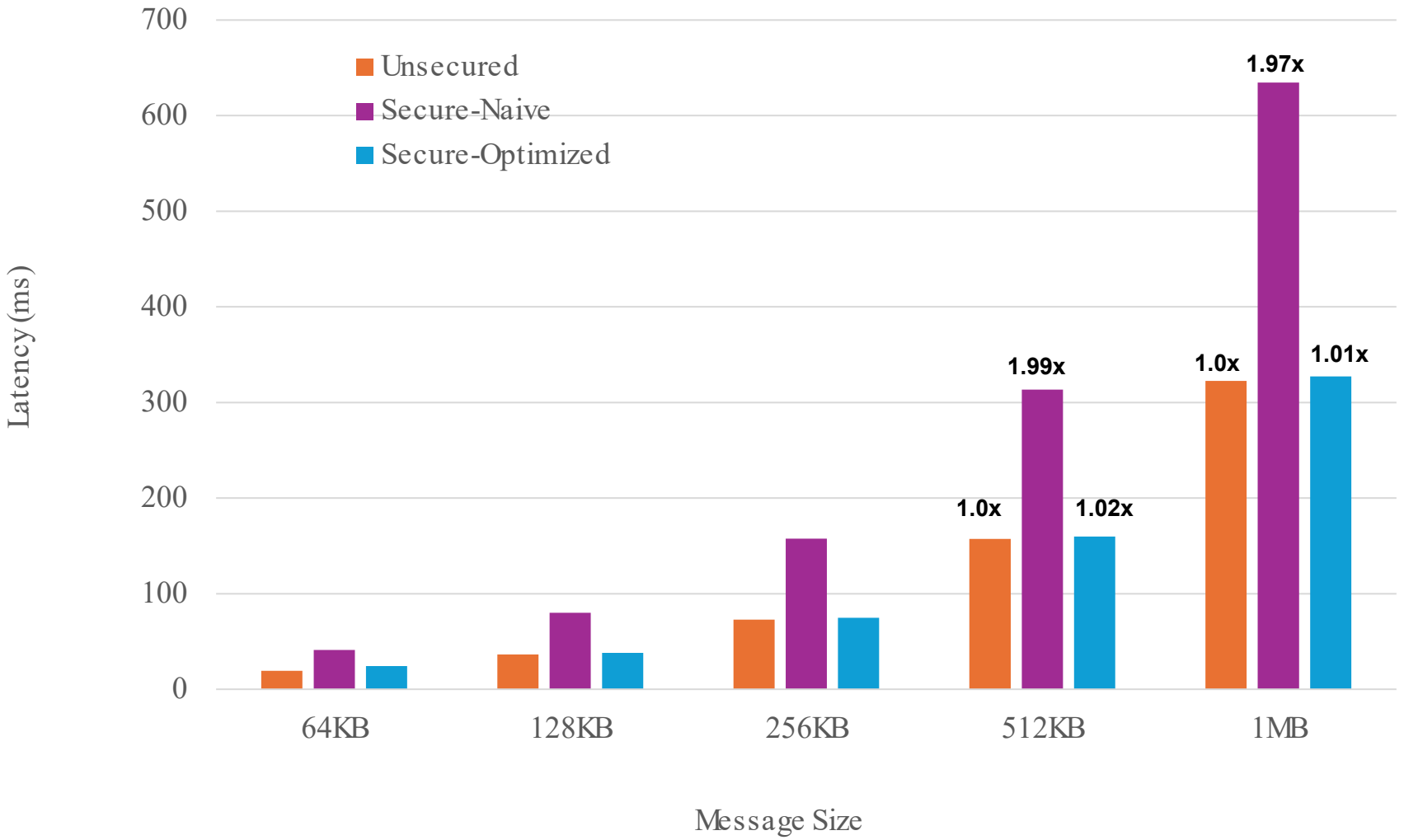
Performance: OSU_Alltoall Micro-benchmark

- Blocking alltoall operation on 512 cores
- 16 nodes, 32 ppn on an Intel cluster (Inter-node & intra-node communication)



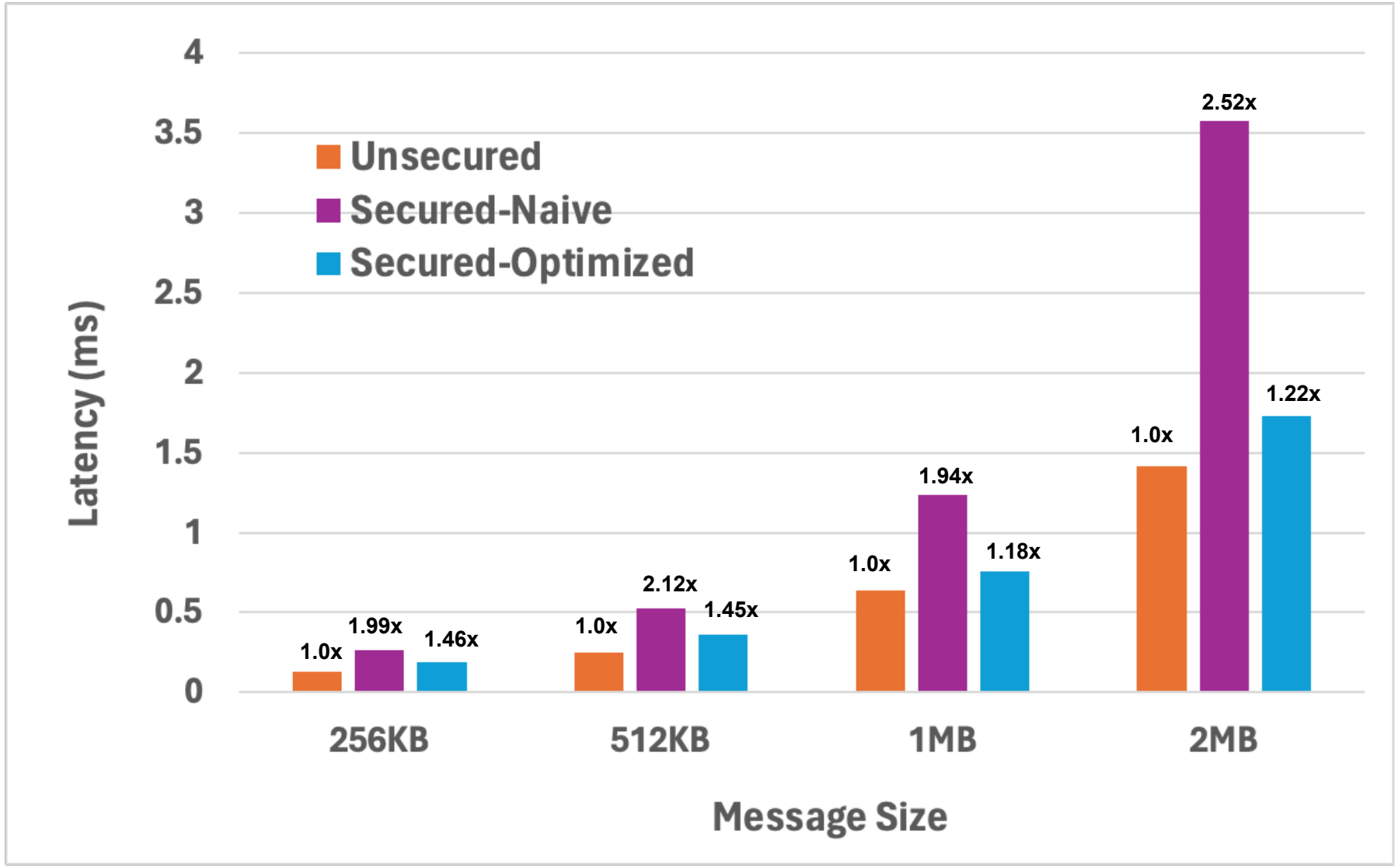
Performance: OSU_Allgather Micro-benchmark

- Blocking allgather operation on 512 cores
- 16 nodes,32 ppn on an Intel cluster (Inter-node & intra-node communication)



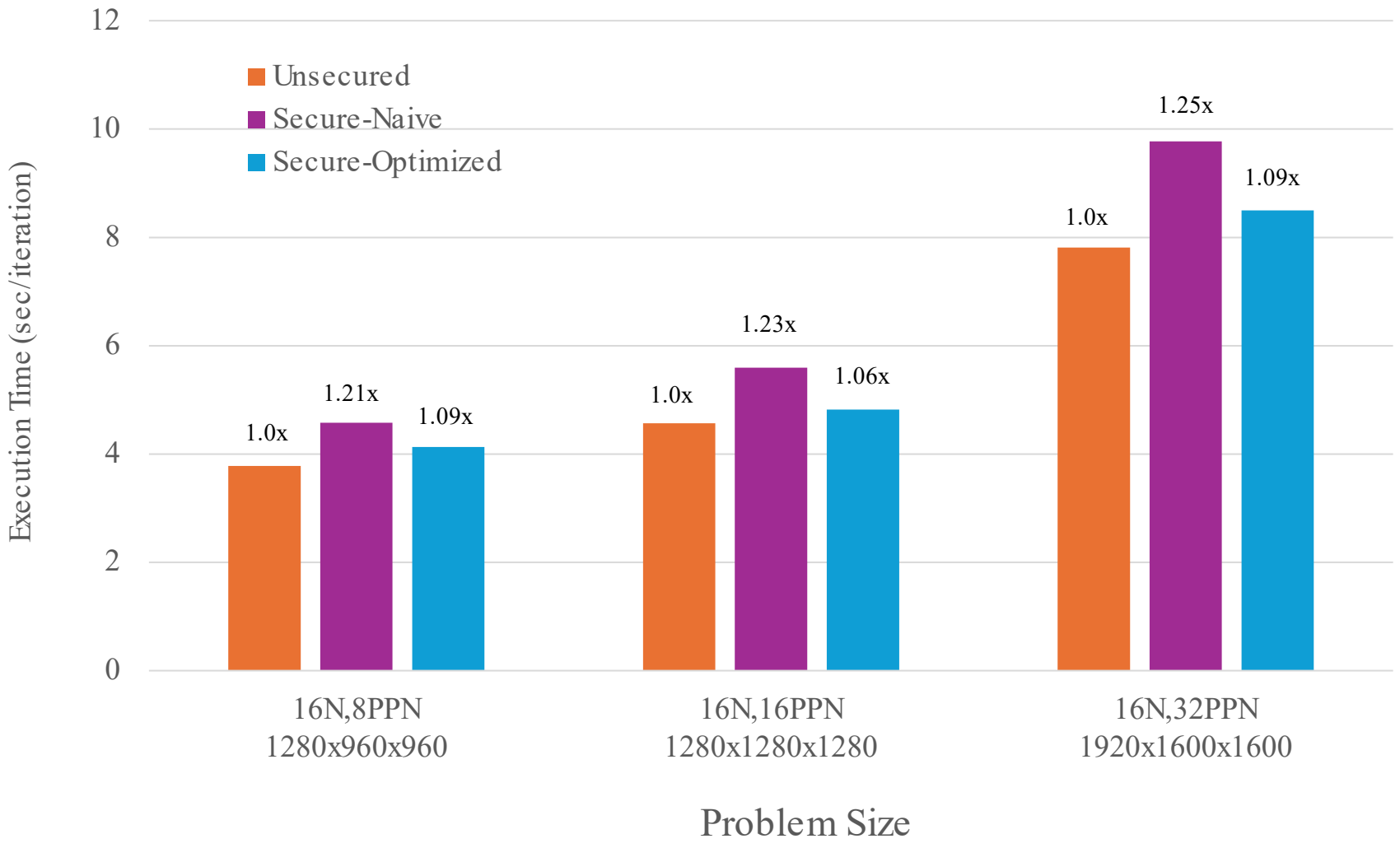
Performance: OSU_Bcast Micro-benchmark

- Blocking bcast operation on 512 cores
- 16 nodes,32 ppn on an Intel cluster (Inter-node & intra-node communication)



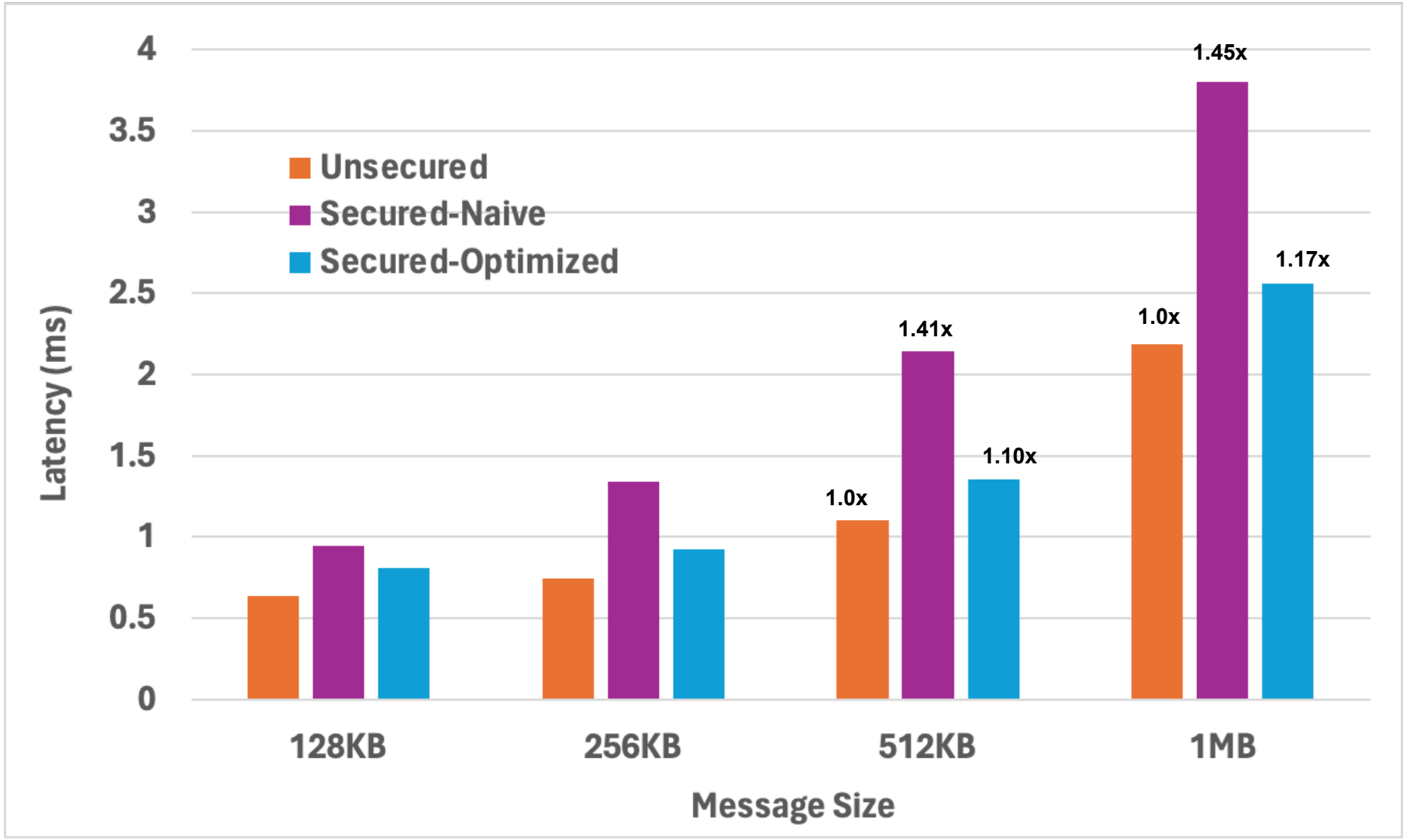
SecureMPI Performance: P3DFFT Application Kernel

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



SecureMPI Performance: 3DStencil Application Kernel

- Parallel 3DStencil application kernel with various problem sizes
- Up to 8 nodes,8 ppn on an Intel cluster (Inter-node & intra-node communication)



Future Releases and Engagement Plan

- Upcoming release for X-ScaleAI will support more AWS instance types, other cloud provides (Azure, Oracle, and Google), on-premise systems, and improved fine-tuning and inference capabilities
- Upcoming release for X-ScaleSecureMPI will support more MPI collectives

- X-ScaleSolutions will be happy to get engaged with end users and collaborators, please send a note to contactus@x-scalesolutions.com

Thank You!

contactus@x-scalesolutions.com

X-ScaleSolutions

http://x-scalesolutions.com/