Pushing the Limits of AI with In-Network Computing

APNET 2019

Gil Bloch

Mellanox Accelerates Leading HPC and AI Systems

World’s Top 3 Supercomputers

OAK RIDGE National Laboratory

U.S. DEPARTMENT OF ENERGY

1

Summit CORAL System
World’s Fastest HPC / AI System
9.2K InfiniBand Nodes

TOP 500 The List.

Lawrence Livermore National Laboratory

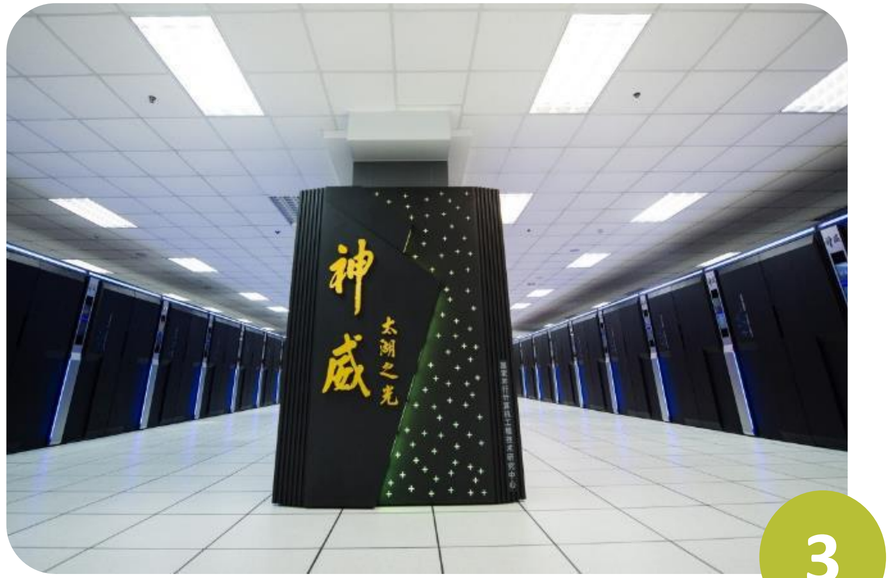
U.S. DEPARTMENT OF ENERGY

2

Sierra CORAL System
#2 USA Supercomputer
8.6K InfiniBand Nodes

TOP 500 The List.

国家超级计算无锡中心
National Supercomputing Center in Wuxi



3

Wuxi Supercomputing Center
Fastest Supercomputer in China
41K InfiniBand Nodes

TOP 500 The List.

Data is Growing Faster Than Ever

Autonomous vehicle generates 4000GByte per day

CAMERA

~20-40MB Per/sec

SONAR

~10-100KB Per/Sec

GPS

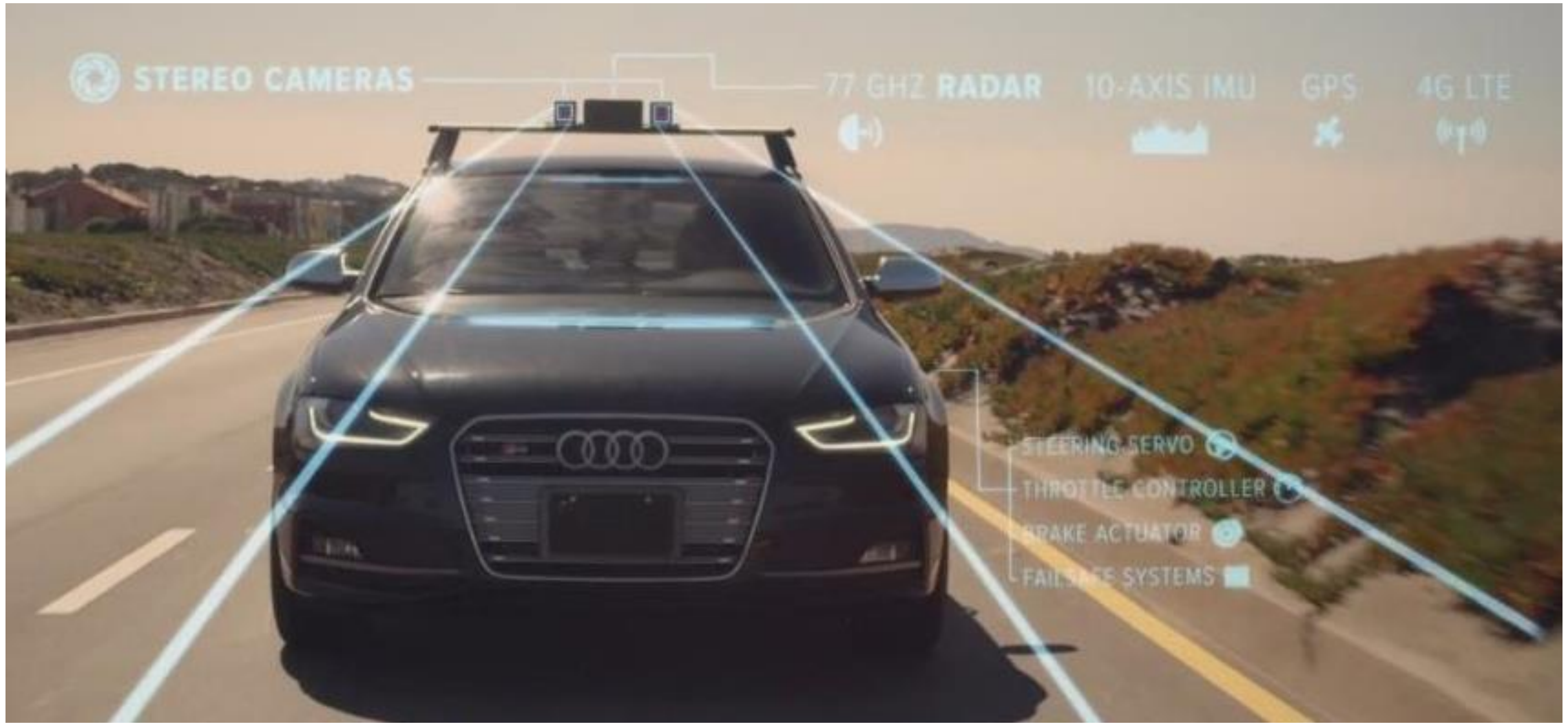
~50KB Per/Sec

RADAR

~10-100KB Per/Sec

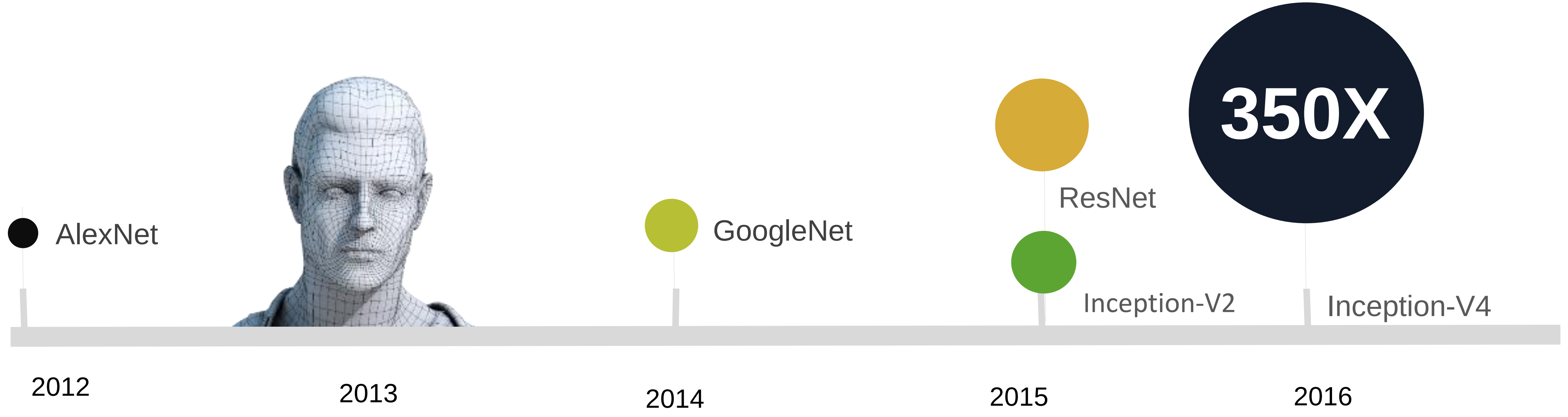
Light Detection & Ranging

~10-70MB Per/Sec



- Data will grow by a factor of 10 over the next decade to 163 Zeta Bytes in 2025 (source: IDC)
- Faster Data processing requires faster Interconnect speeds

Neural Networks Complexity Growth



AlexNet

GoogleNet

ResNet

Inception-V2

350X

Inception-V4

2012

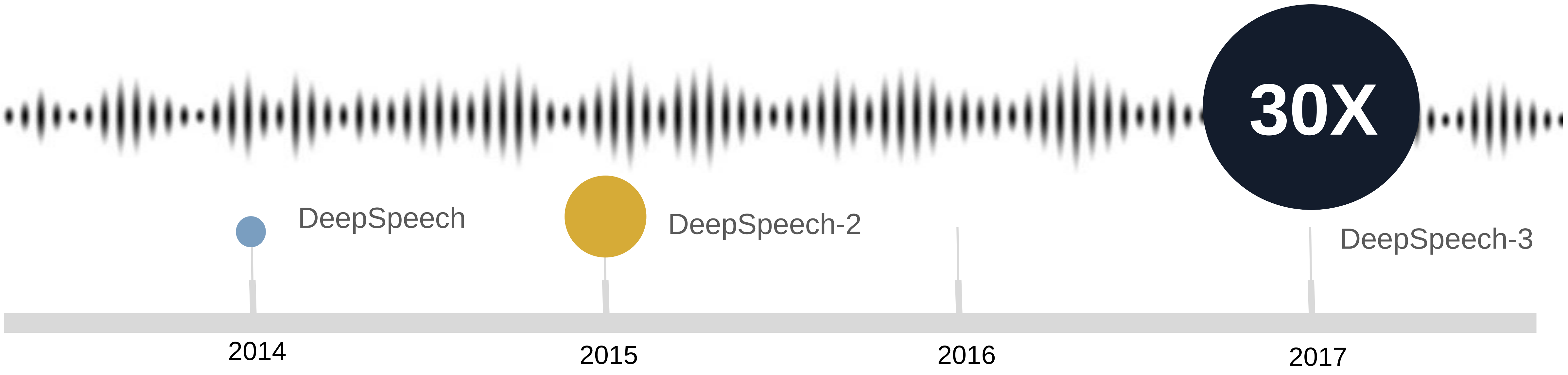
2013

2014

2015

2016

Image Recognition



30X

DeepSpeech

DeepSpeech-2

DeepSpeech-3

2014

2015

2016

2017

Speech Recognition

Complexity = GOPS X Bandwidth

Enabling World-Leading Artificial Intelligence Solutions

Mellanox Unleashes the Power of Artificial Intelligence

More Data

+

Better Models

+

Faster Interconnect

GPUs

CPUs

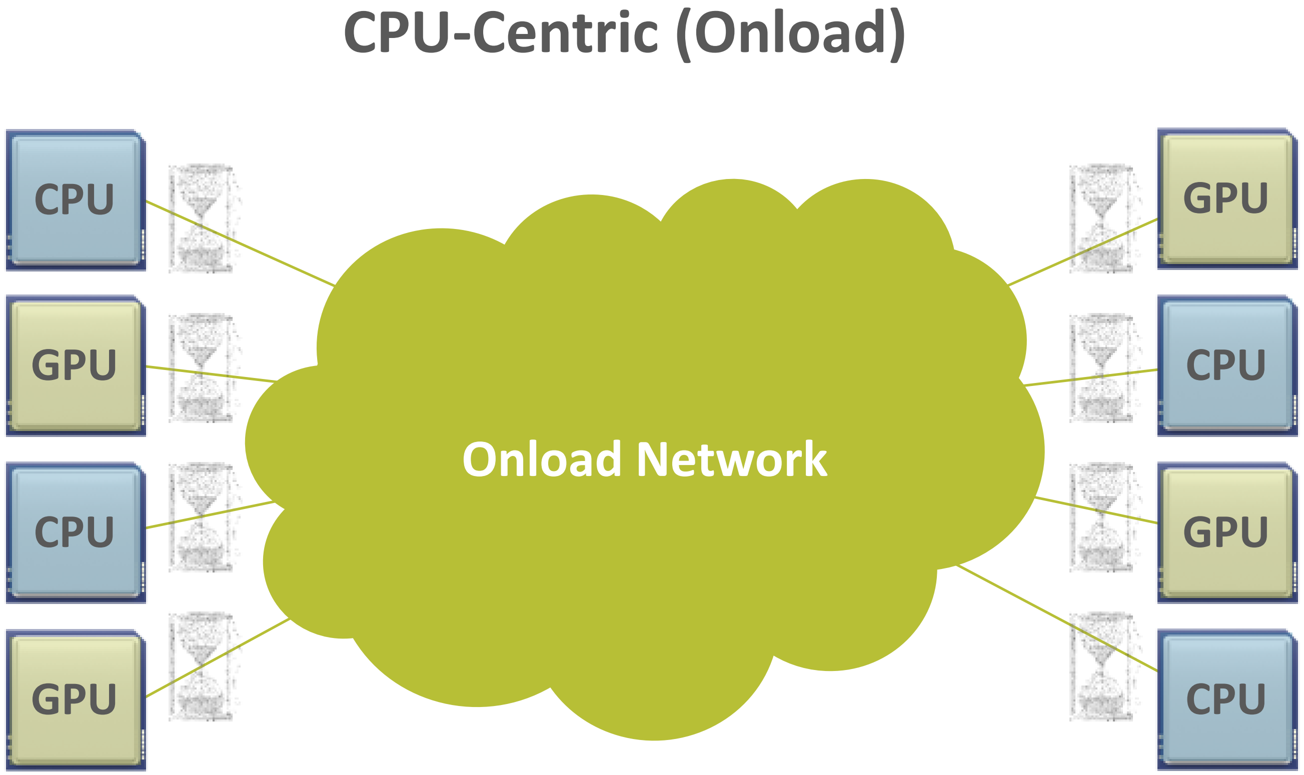
ASIC

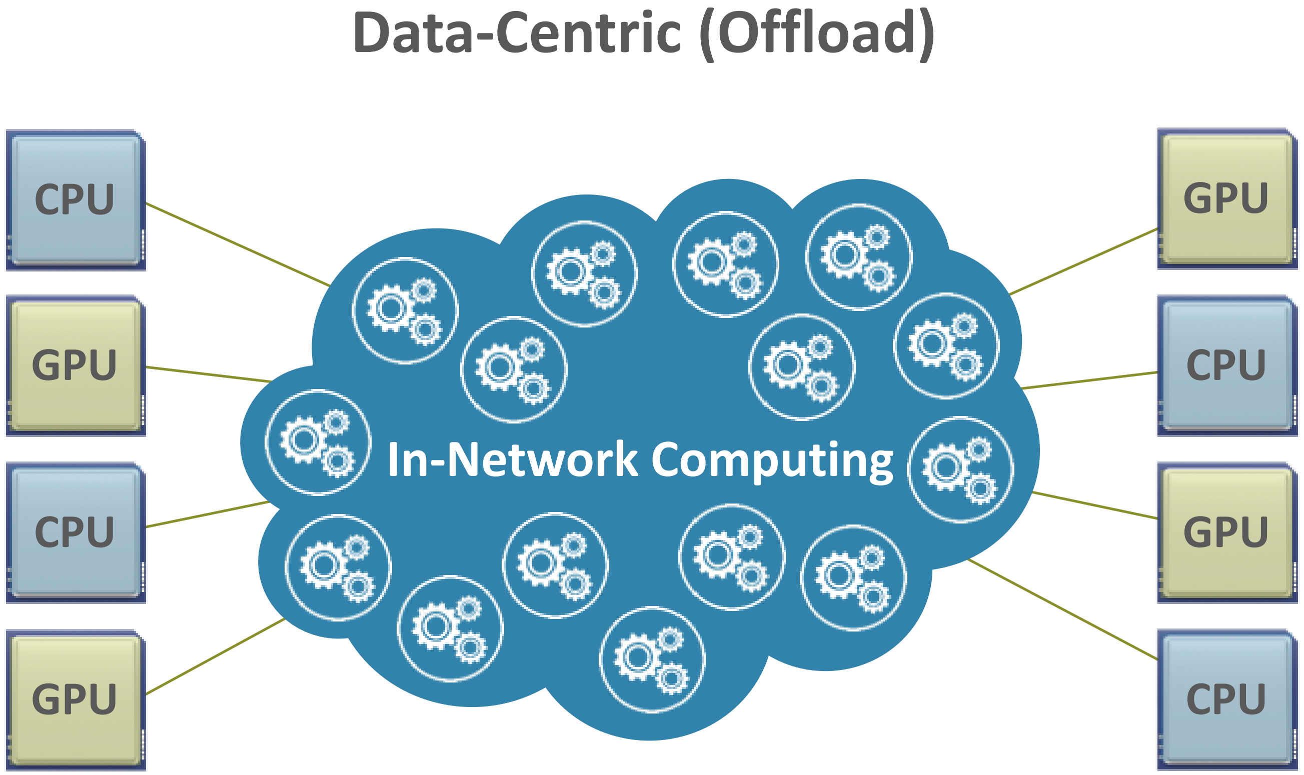
FPGAs

Storage

The Need for Intelligent and Faster Interconnect

Faster Data Speeds and In-Network Computing
Enable Higher Performance and Scale





Must Wait for the Data
Creates Performance Bottlenecks

→

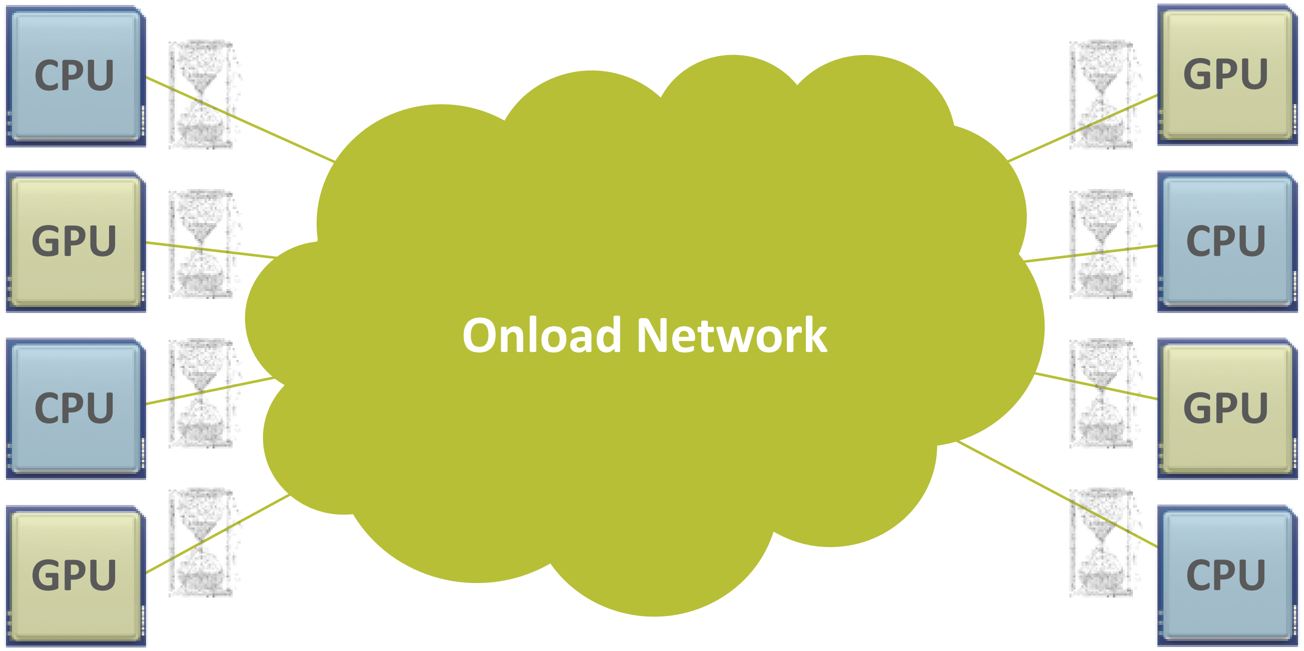
Analyze Data as it Moves!
Higher Performance and Scale

An Application Example – Pizza Processing

CPU 1 – Pizza Generation

CPU 2 – Pizza Consumption

CPU-Centric (Onload)



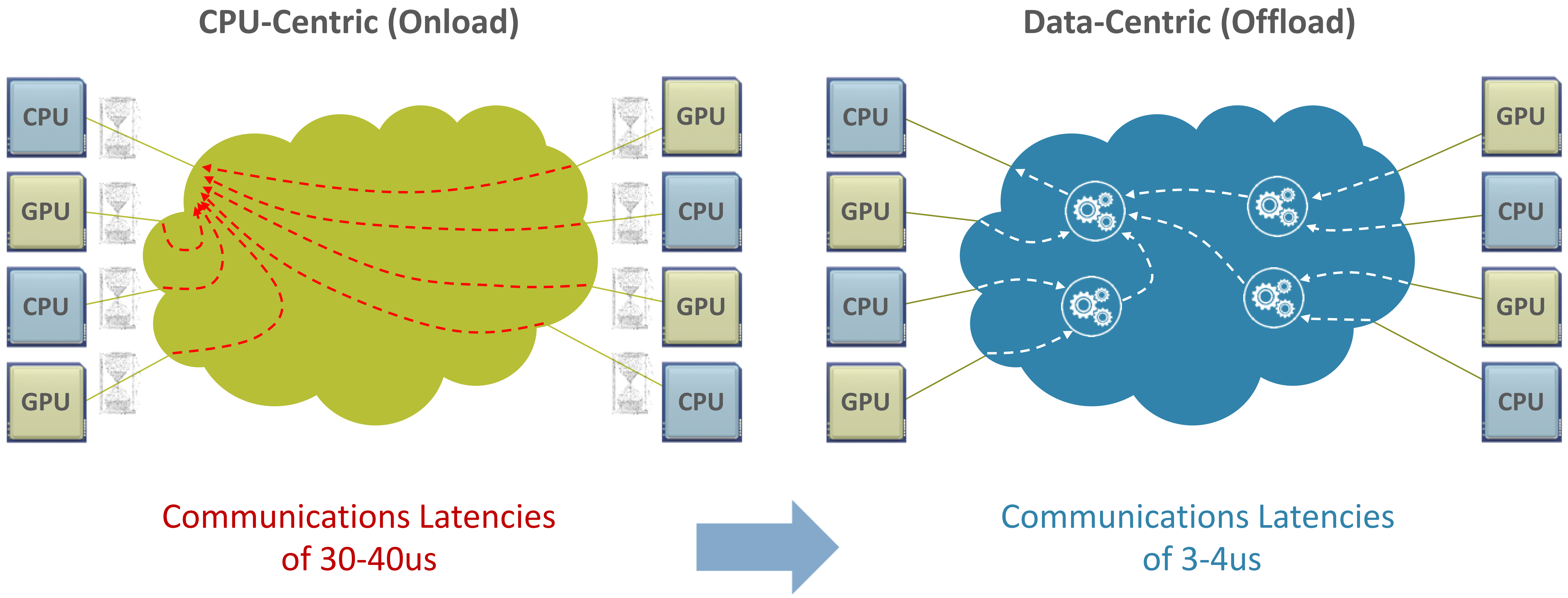
Must Wait for the Data
Creates Performance Bottlenecks

- Order Pizza
  - Call (or use Pizza application)
- CPU 1 – prepare Pizza
  - Tomato sauce, Cheese, Peperoni…
- CPU 1 – Put in the oven
  - And now we wait…
- CPU 1 – Pack and send
- Network (Pizza Delivery)

What if…

Data Centric Architecture to Overcome Latency Bottlenecks

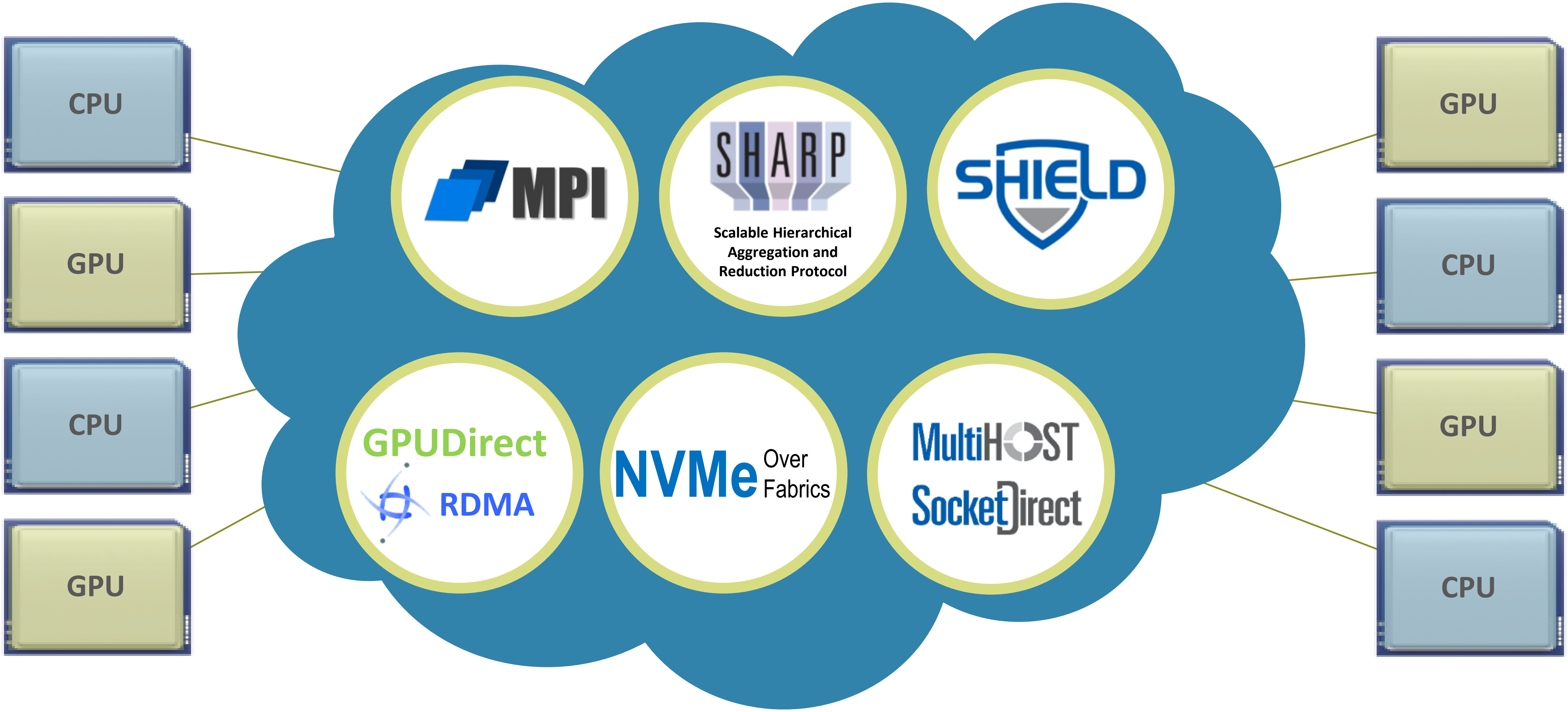
Intelligent Interconnect Paves the Road to Exascale Performance



Communications Latencies of 30-40us

Communications Latencies of 3-4us

In-Network Computing to Enable Data-Centric Data Centers



Accelerating All Levels of HPC/AI Frameworks

Application Framework

- Data Analysis
- Configurable Logic

Communication Framework

- SHARP – Data Aggregation
- MPI Tag Matching
- MPI Rendezvous
- SNAP - Software Defined Virtual Devices

MPI HPC-X™

UCF Unified Communication Framework

SHARP Scalable Hierarchical Aggregation and Reduction Protocol

Network Framework

- Network Transport Offload
- RDMA and GPU-Direct
- SHIELD (Self-Healing Network)
- Adaptive Routing and Congestion Control

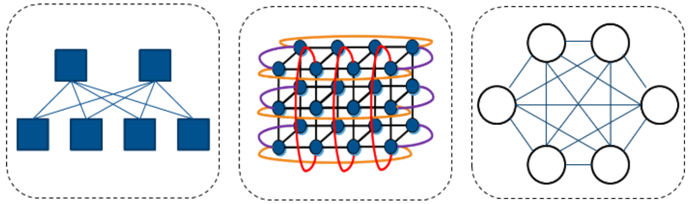
GPUDirect

RDMA

SHIELD

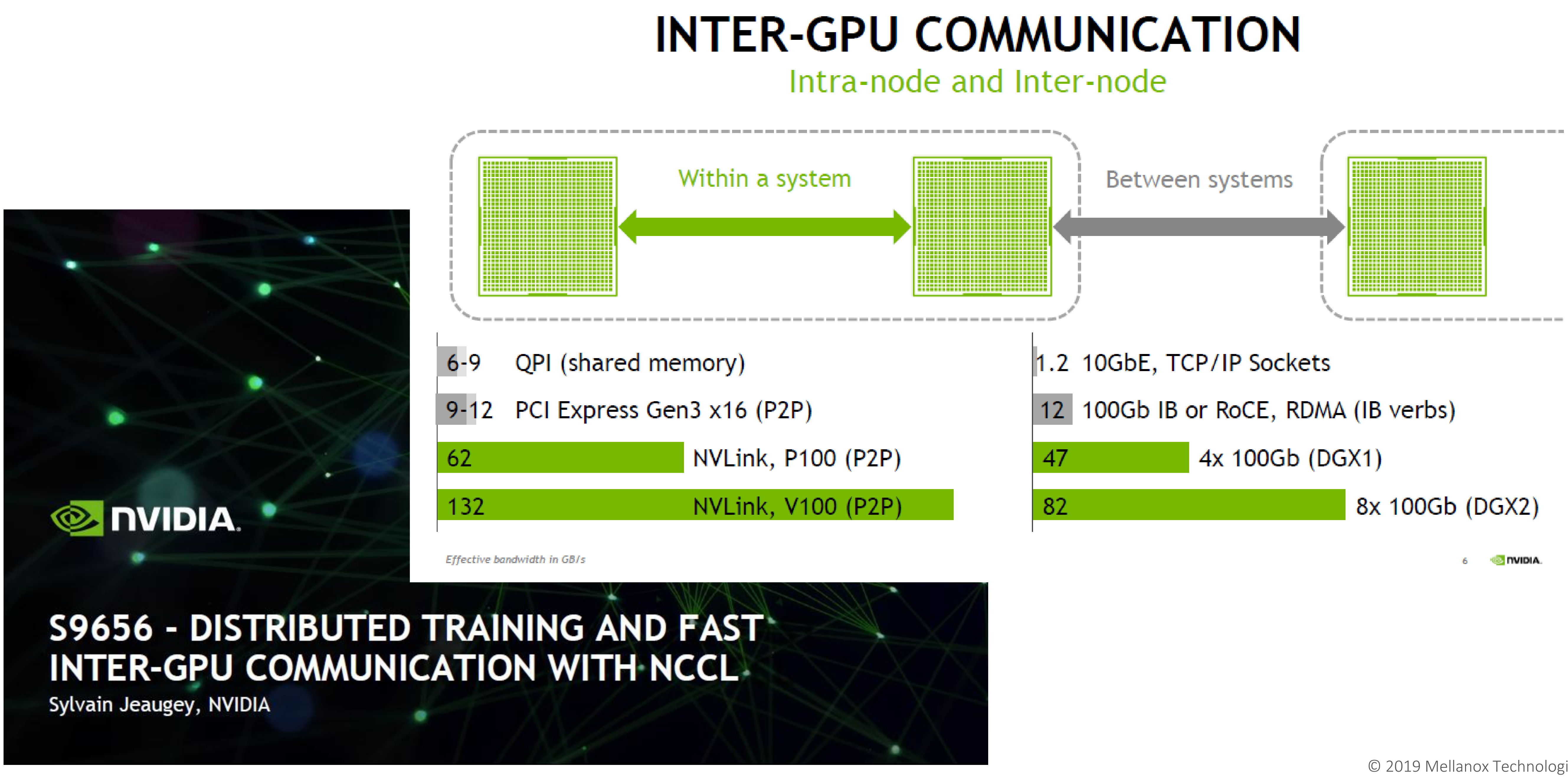
Connectivity Framework

- Multi-Host
- Enhanced Topologies
- Dragonfly+



The Need for Speed

Matching Inter and Intra Node Bandwidth



INTER-GPU COMMUNICATION

Intra-node and Inter-node

6-9	QPI (shared memory)
9-12	PCI Express Gen3 x16 (P2P)
62	NVLink, P100 (P2P)
132	NVLink, V100 (P2P)

1.2	10GbE, TCP/IP Sockets
12	100Gb IB or RoCE, RDMA (IB verbs)
47	4x 100Gb (DGX1)
82	8x 100Gb (DGX2)

Effective bandwidth in GB/s

S9656 - DISTRIBUTED TRAINING AND FAST INTER-GPU COMMUNICATION WITH NCCL

Sylvain Jeaugey, NVIDIA

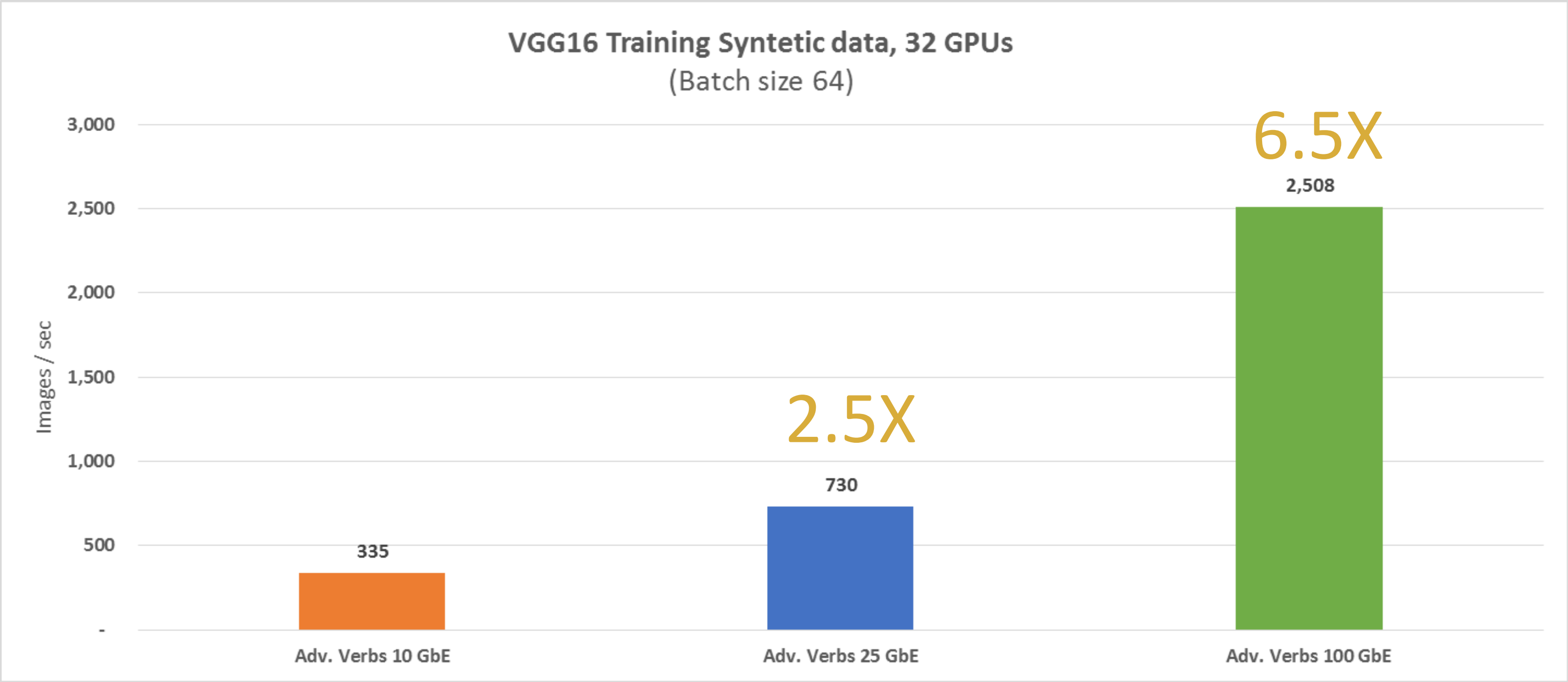
Mellanox Accelerates TensorFlow 1.5

TensorFlow

100G is a Must For Large Scale Models

6.5X

Faster Training with 100G

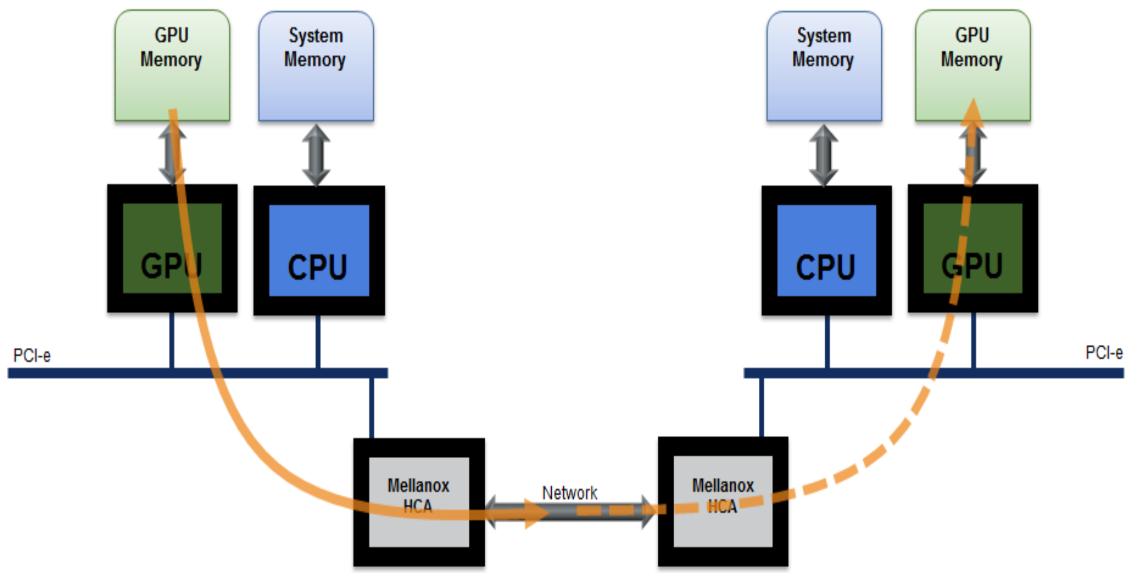


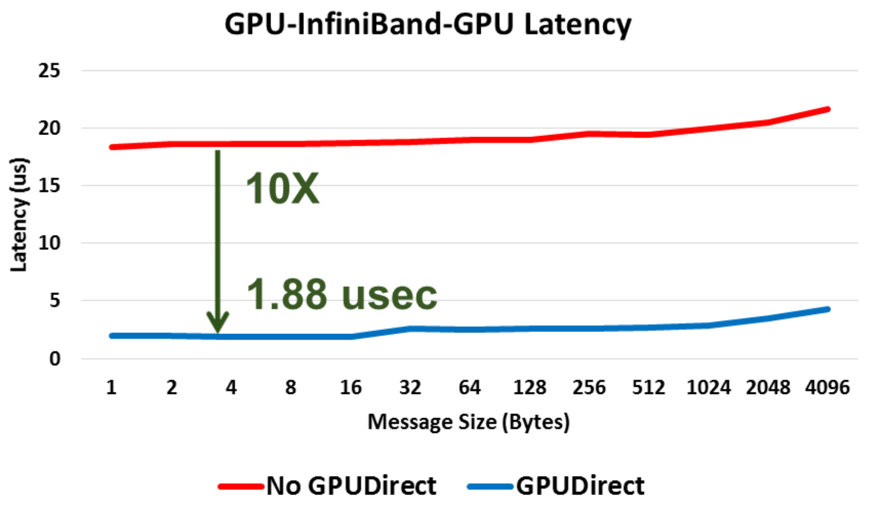
RDMA and GPUDirect

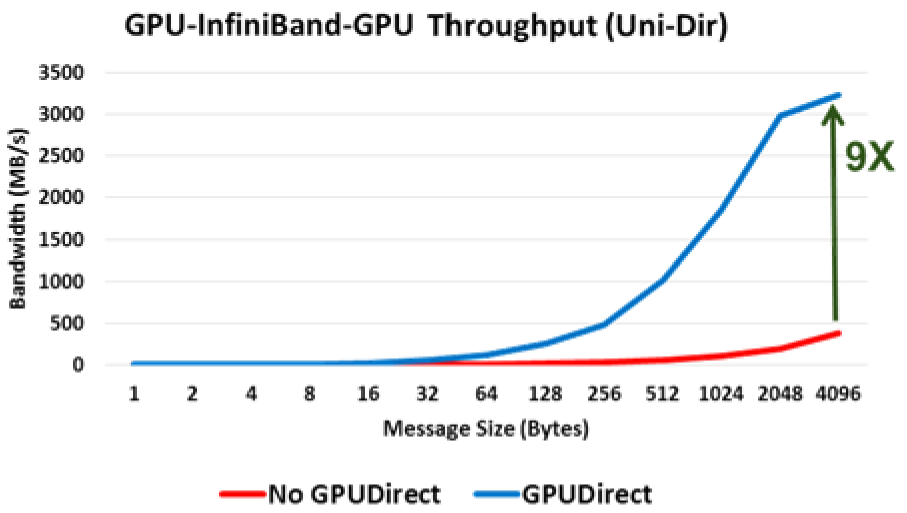
10X Higher Performance with GPUDirect™ RDMA

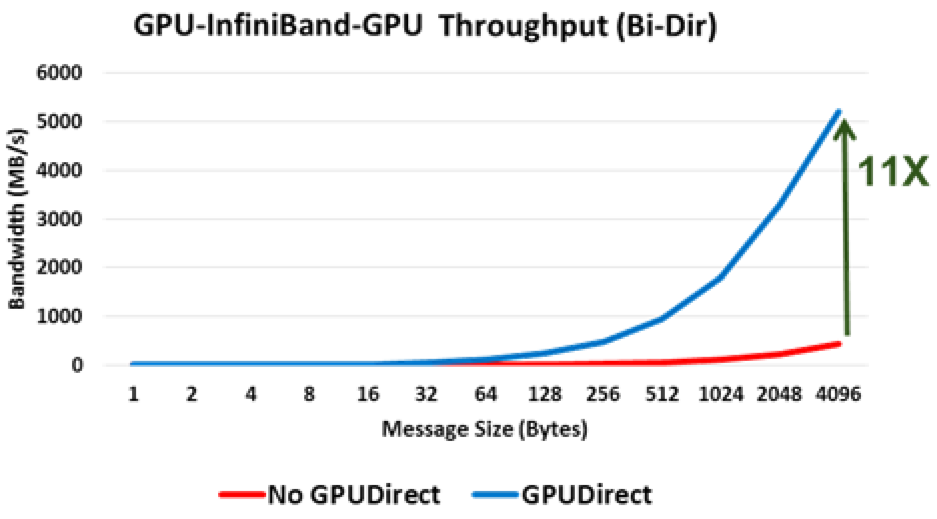
- Accelerates HPC and Deep Learning performance
- Lowest communication latency for GPUs

GPUDirect™ RDMA




GPU-InfiniBand-GPU Latency

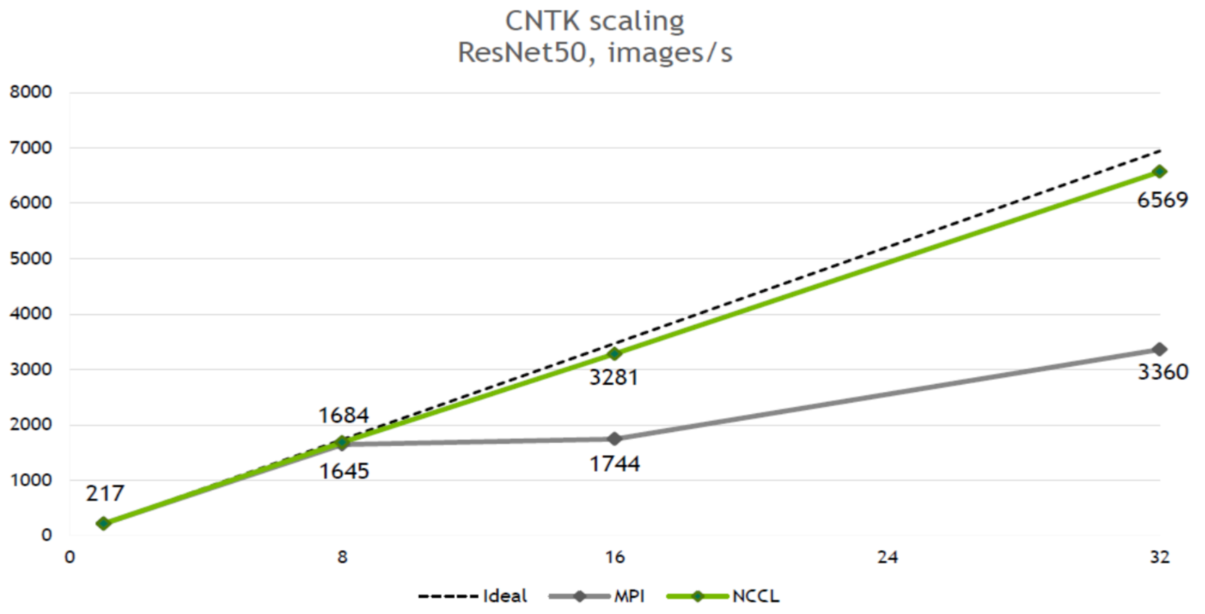

GPU-InfiniBand-GPU Throughput (Uni-Dir)

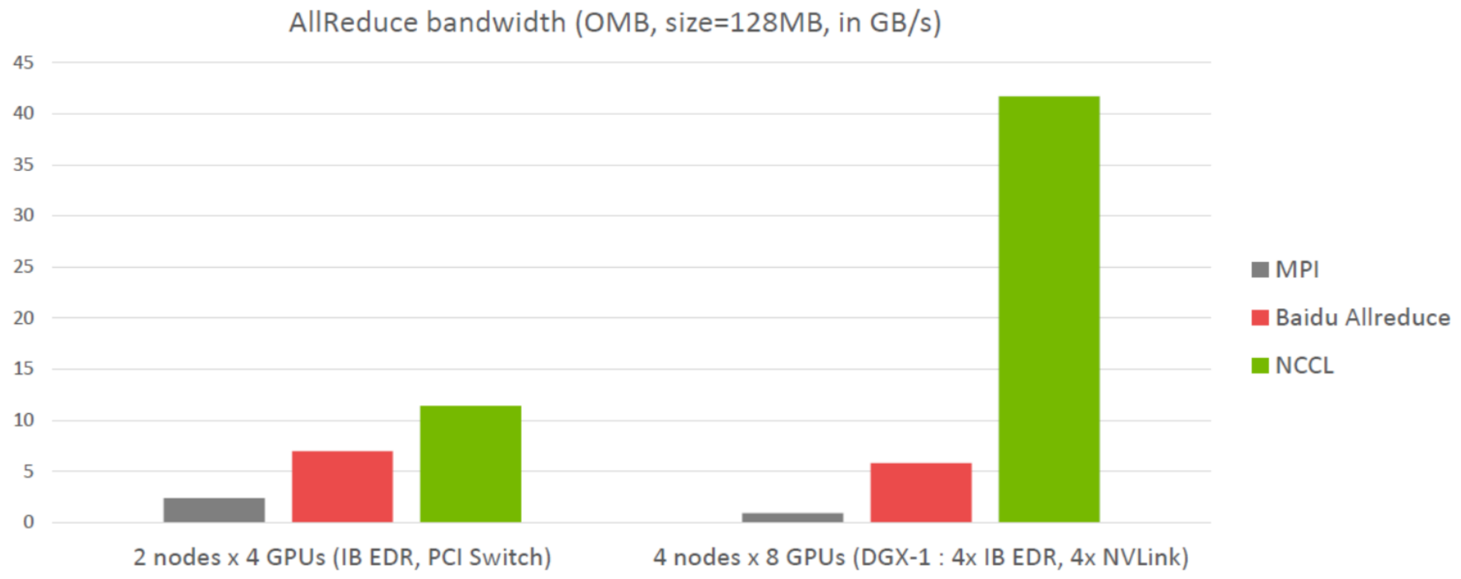

GPU-InfiniBand-GPU Throughput (Bi-Dir)

Mellanox Accelerates NVIDIA NCCL 2.0

50% Performance Improvement

with NVIDIA® DGX-1 across
32 NVIDIA Tesla V100 GPUs
Using InfiniBand RDMA
and GPUDirect™ RDMA





Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

SHARP

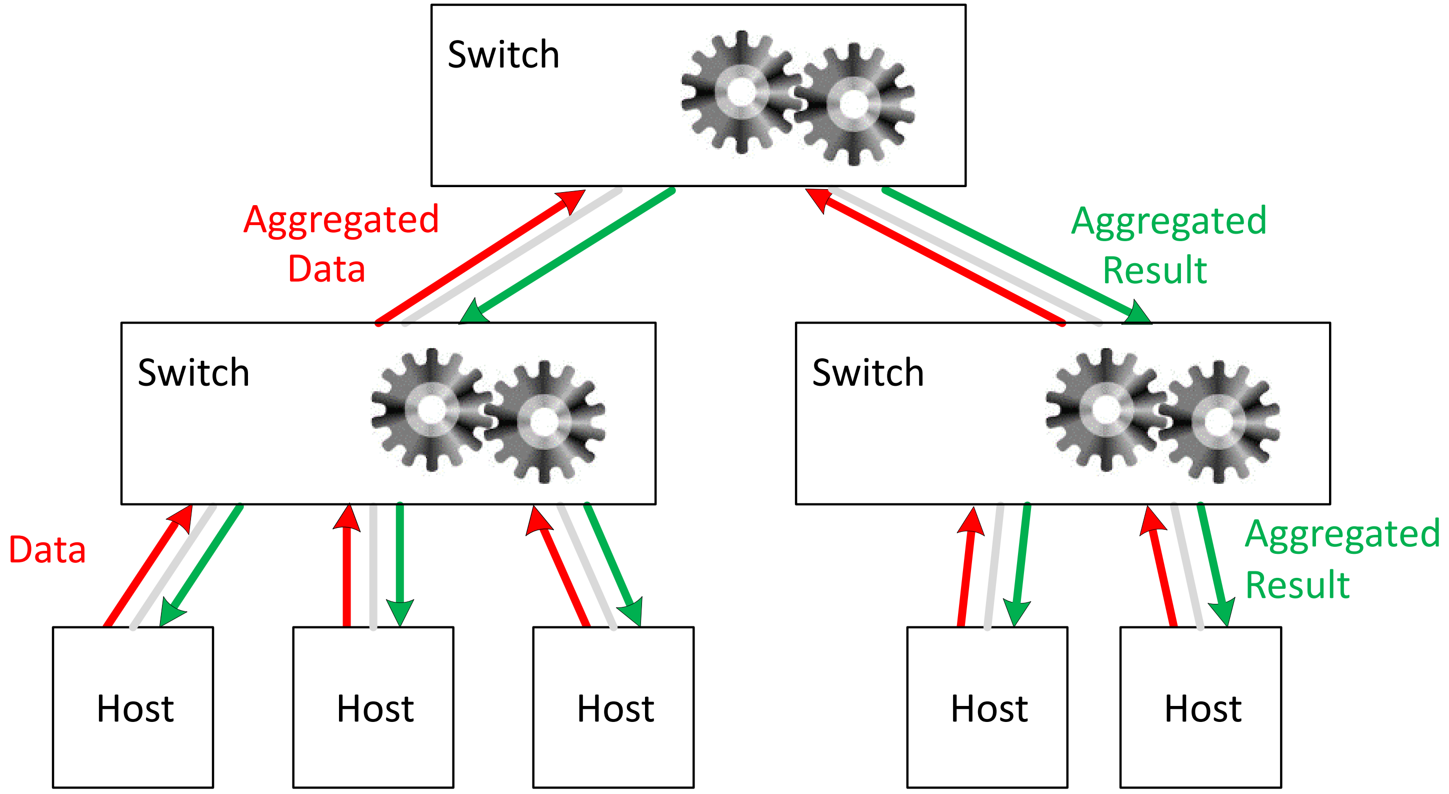
Scalable Hierarchical
Aggregation and
Reduction Protocol

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

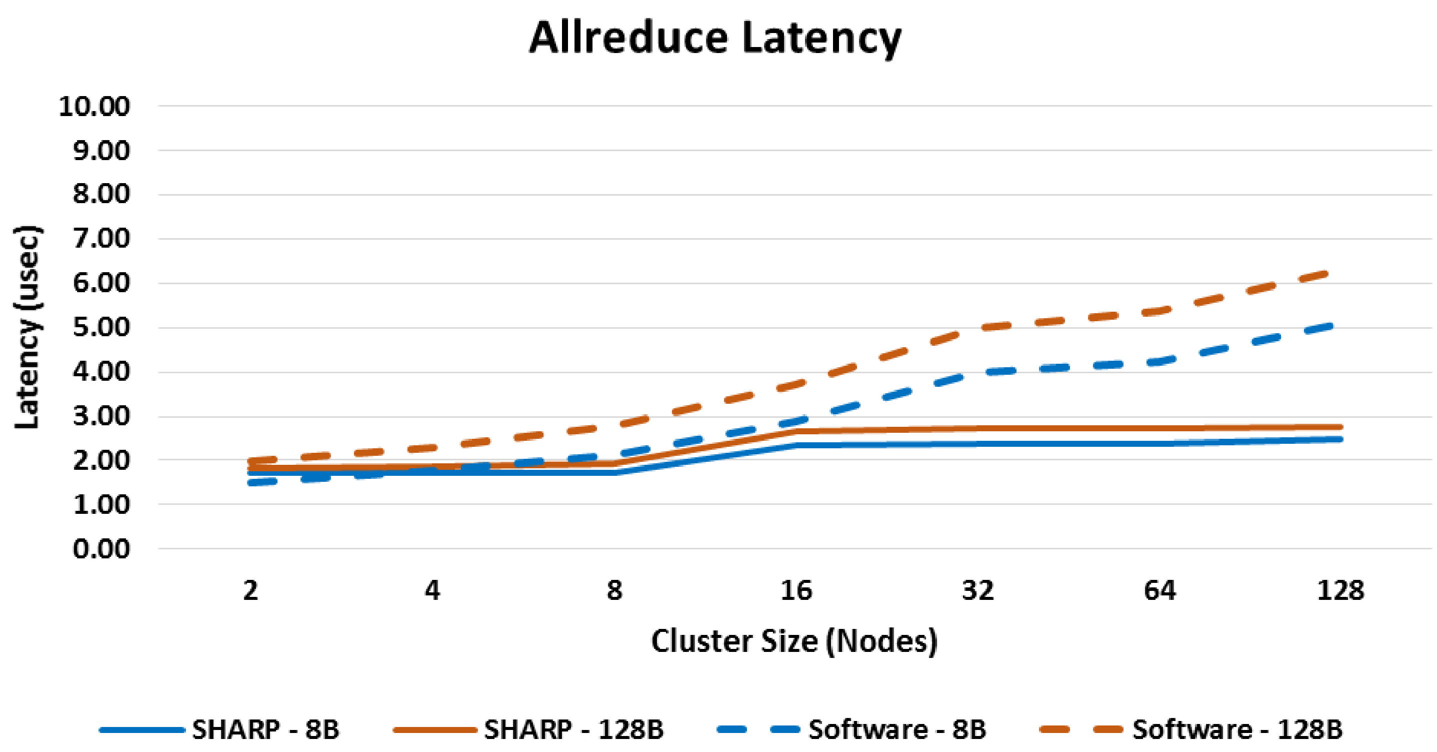
- Reliable Scalable General Purpose Primitive
- Applicable to Multiple Use-cases in ML/HPC
- Scalable High Performance Collective Offload

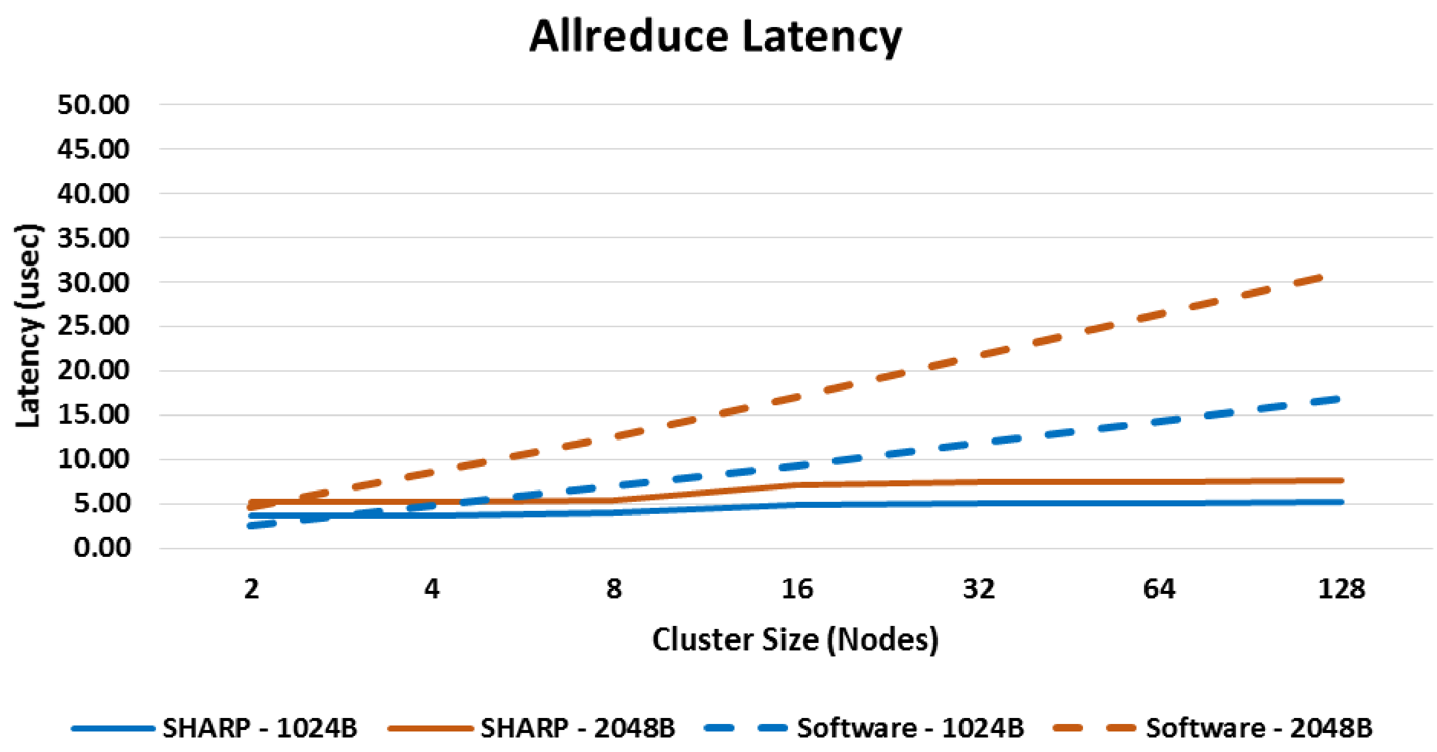
SHARP

Scalable Hierarchical Aggregation and Reduction Protocol



SHARP AllReduce Performance Advantages (128 Nodes)





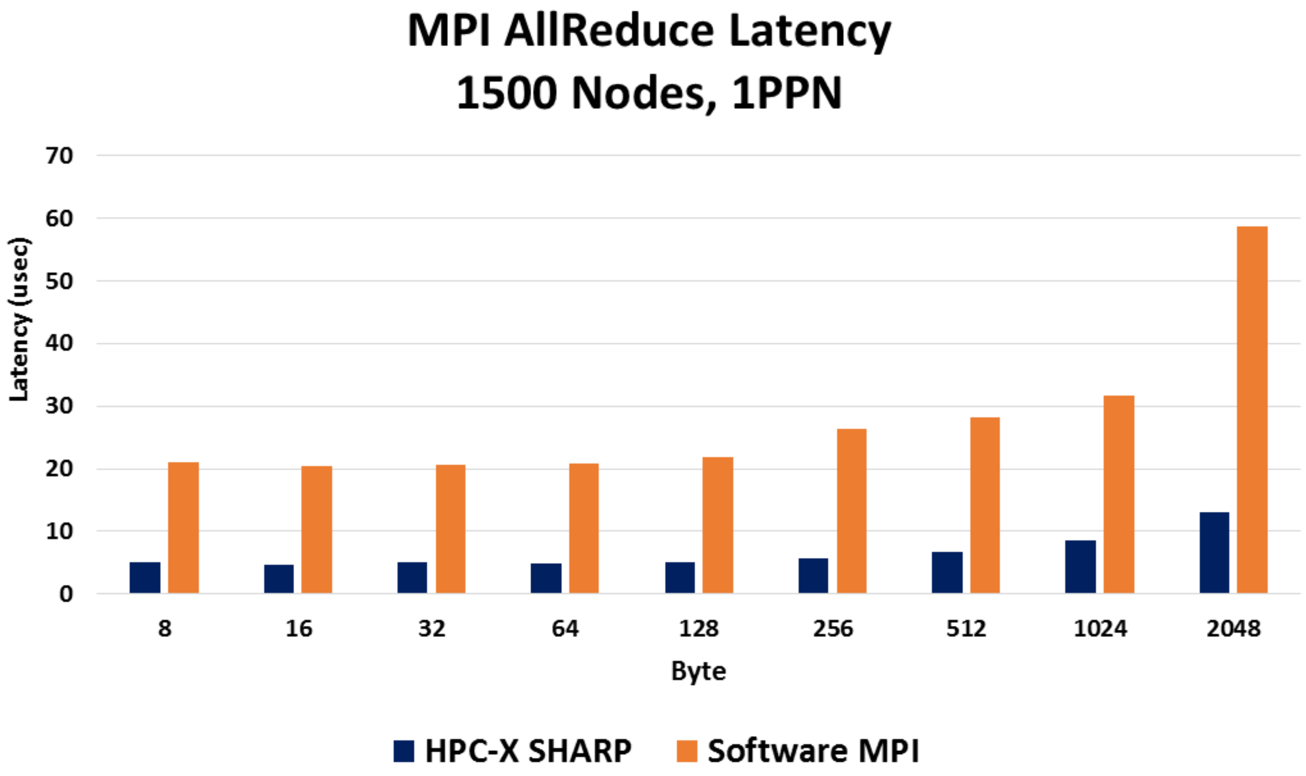
SHARP

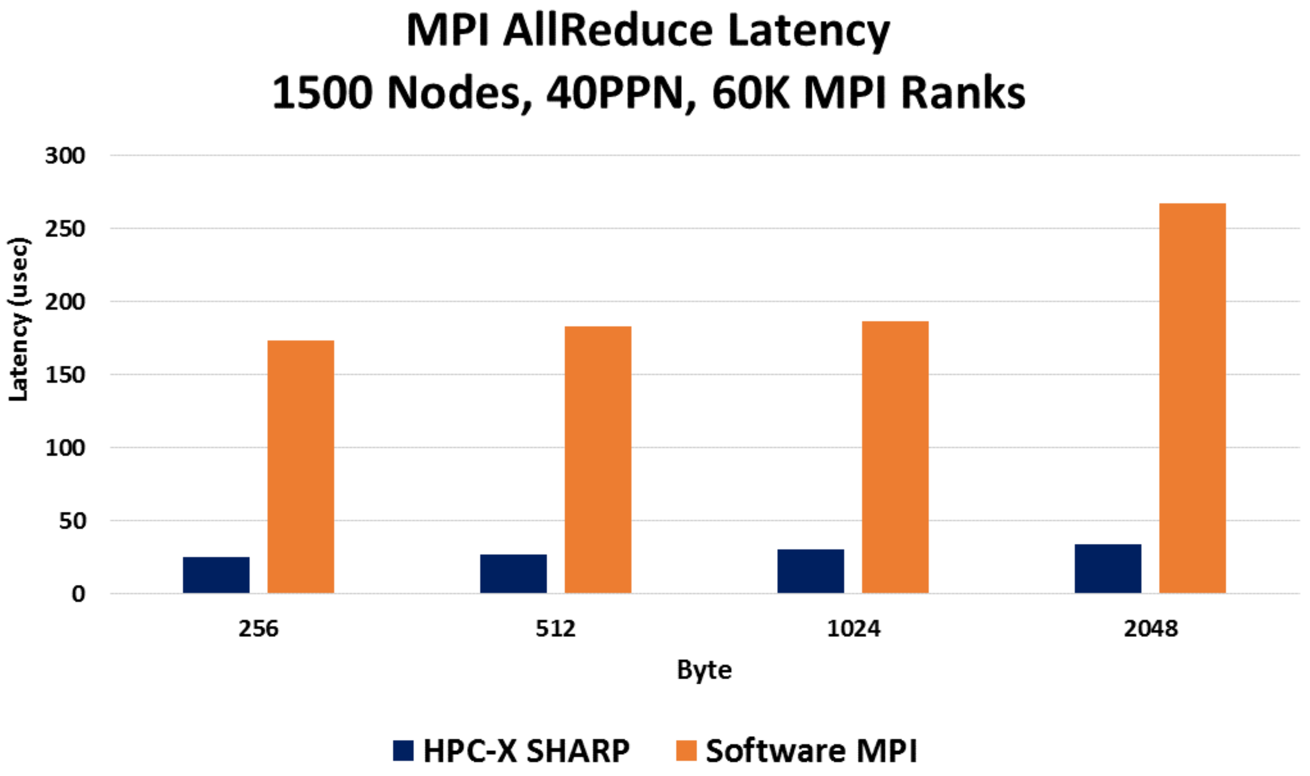
Scalable Hierarchical Aggregation and Reduction Protocol

SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

SHARP AllReduce Performance Advantages

1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology

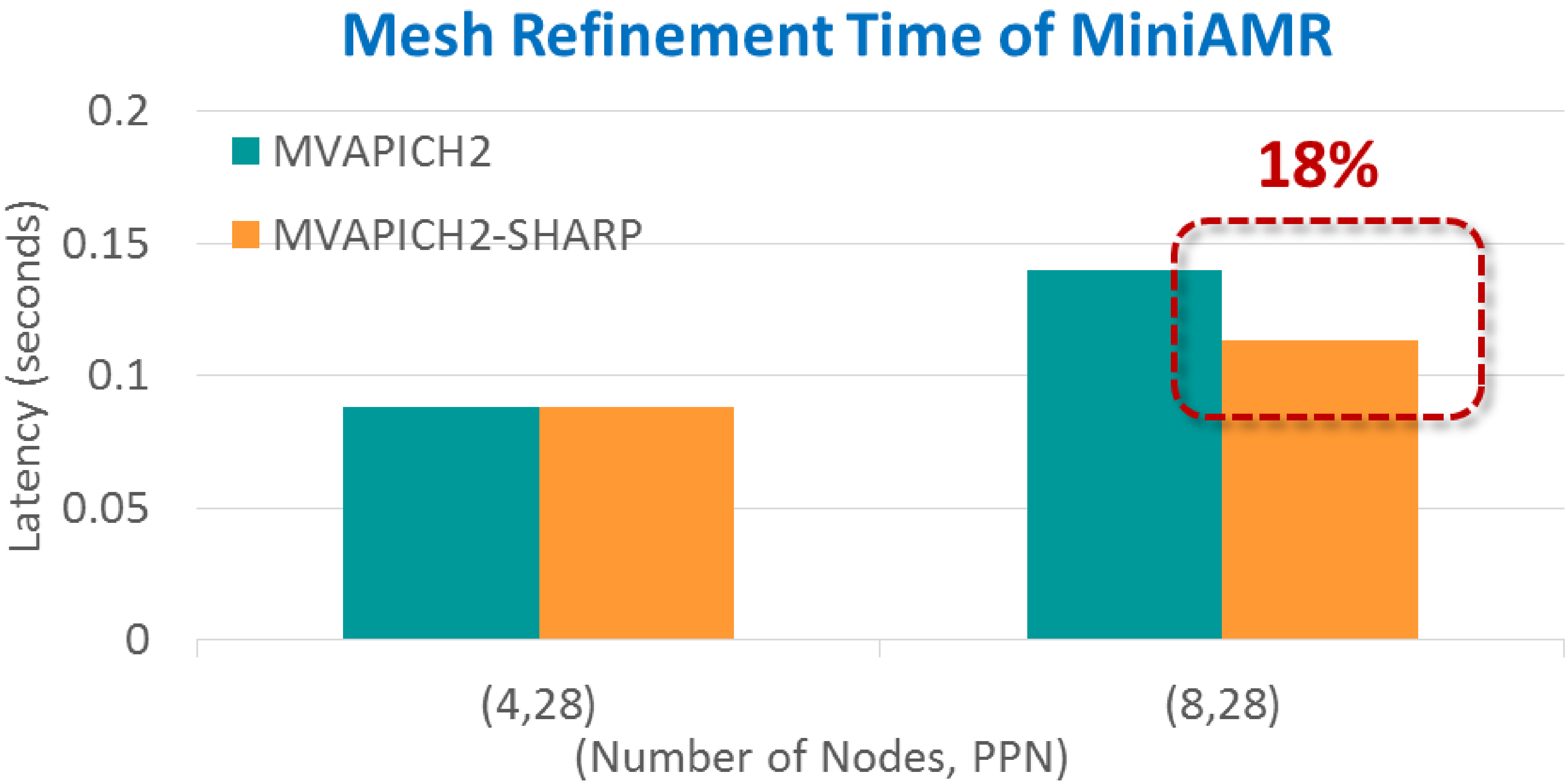

MPI AllReduce Latency
1500 Nodes, 1PPN


MPI AllReduce Latency
1500 Nodes, 40PPN, 60K MPI Ranks

Scalable Hierarchical
Aggregation and
Reduction Protocol

SHARP Enables Highest Performance

SHARP Performance – Application (OSU)



Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

The MVAPICH2 Project
http://mvapich.cse.ohio-state.edu/

Source: Prof. DK Panda, Ohio State University

SHARP Accelerates AI Performance

The CPU in a parameter server
becomes the bottleneck

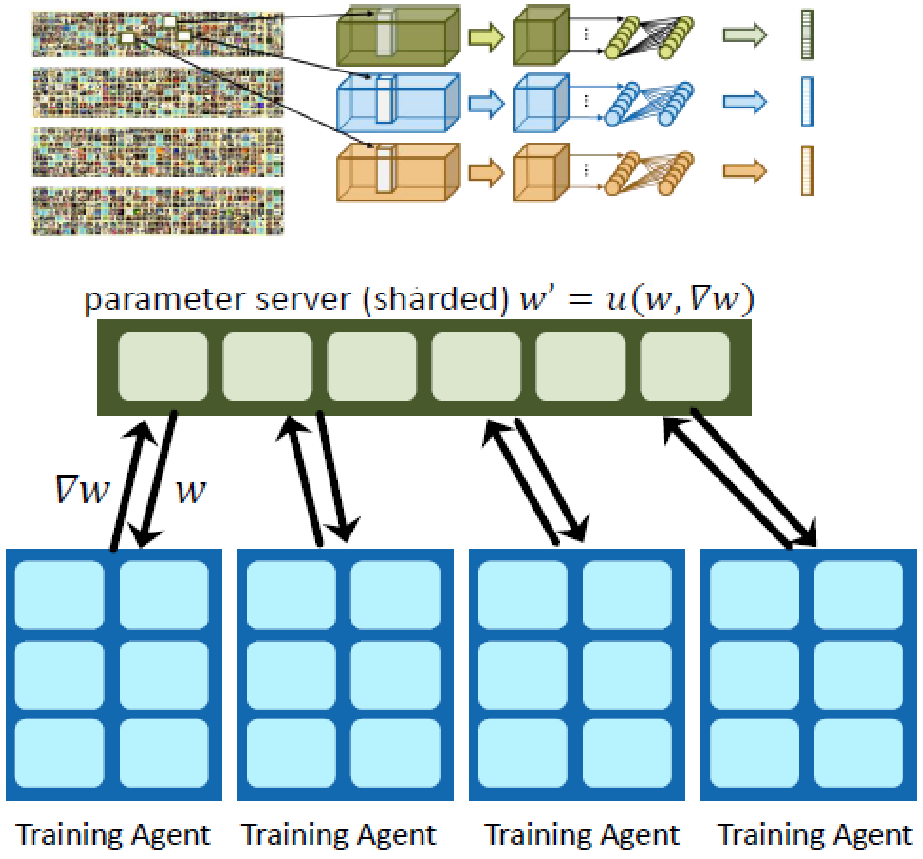
SHARP

Scalable Hierarchical
Aggregation and
Reduction Protocol

Performs the Gradient Averaging

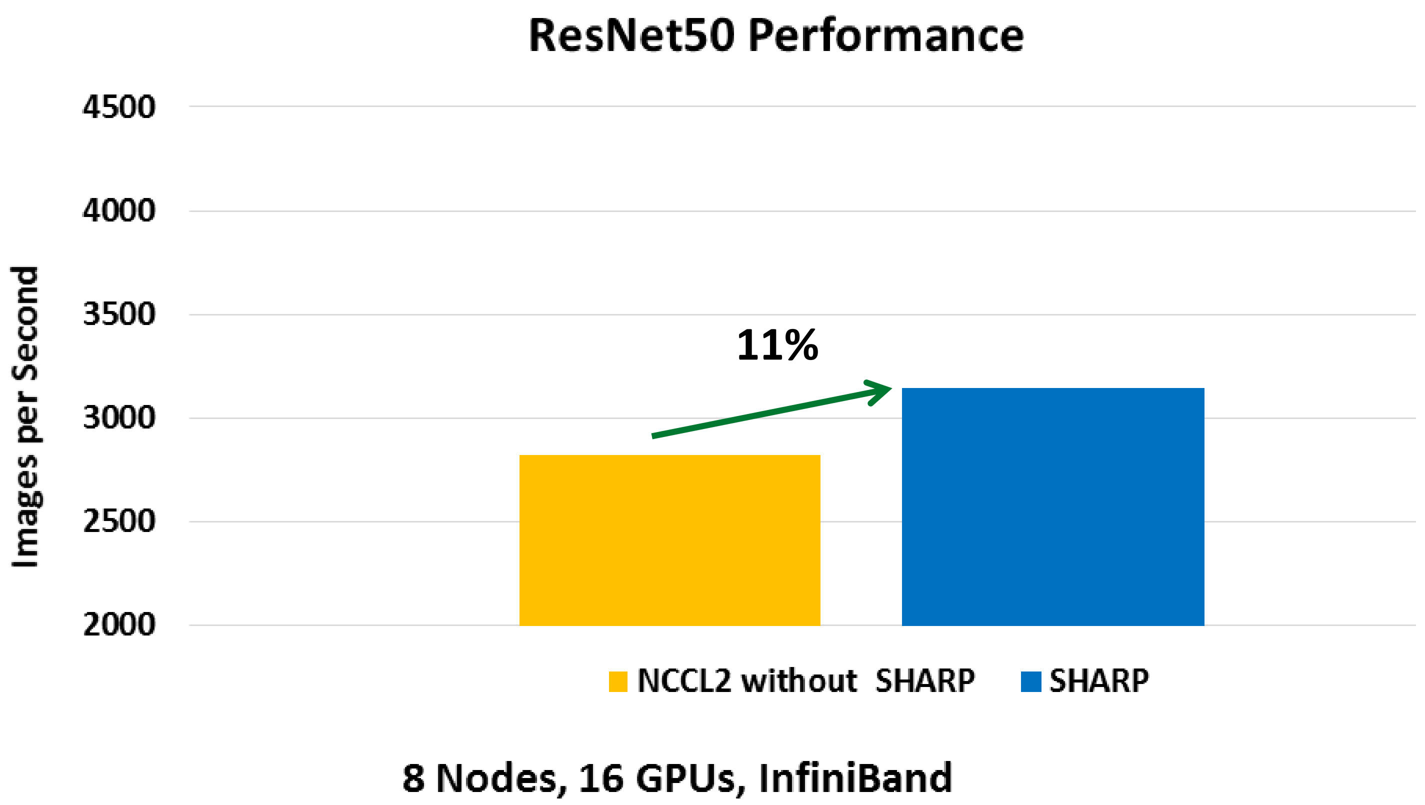
Replaces all physical parameter servers

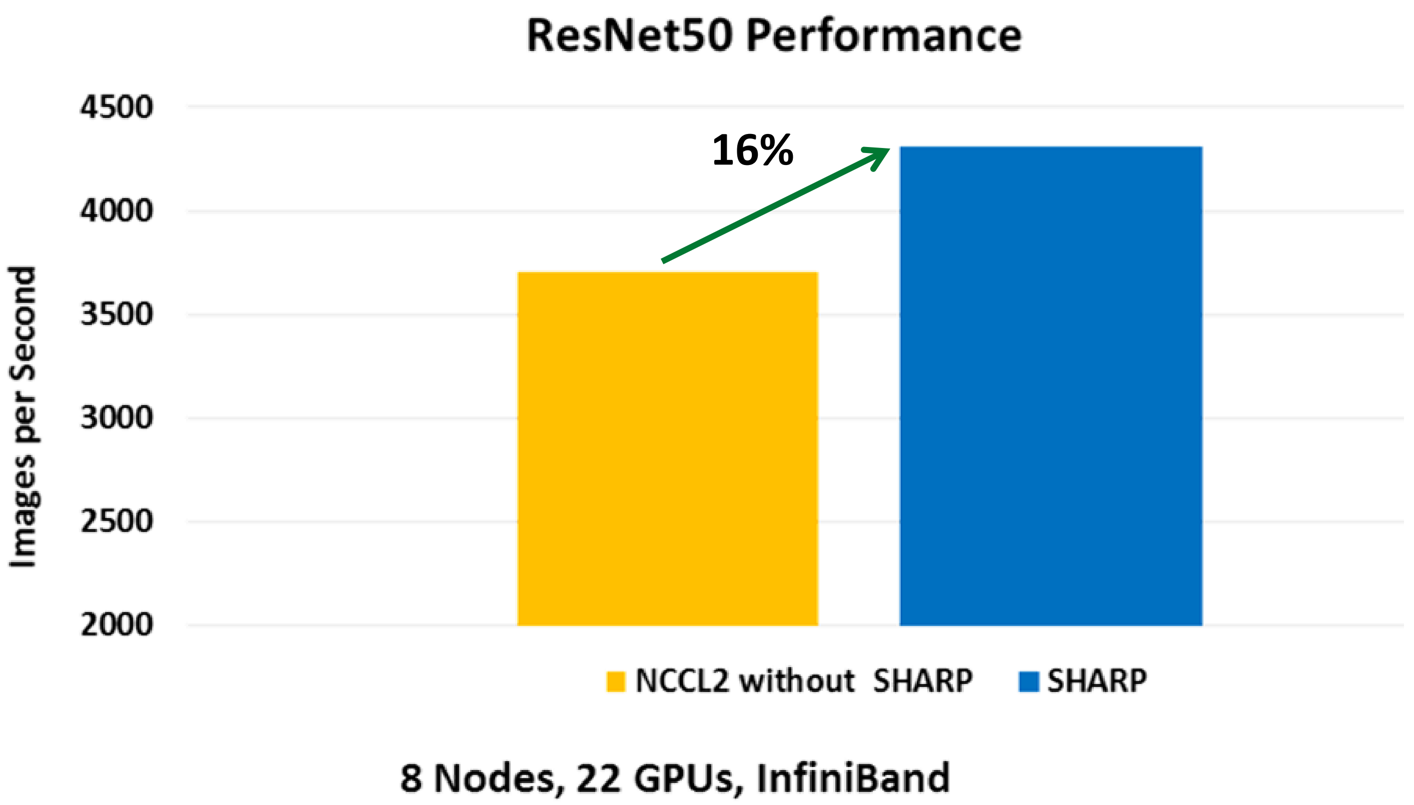
Accelerate AI Performance



InfiniBand SHARP Advantage for Deep Learning

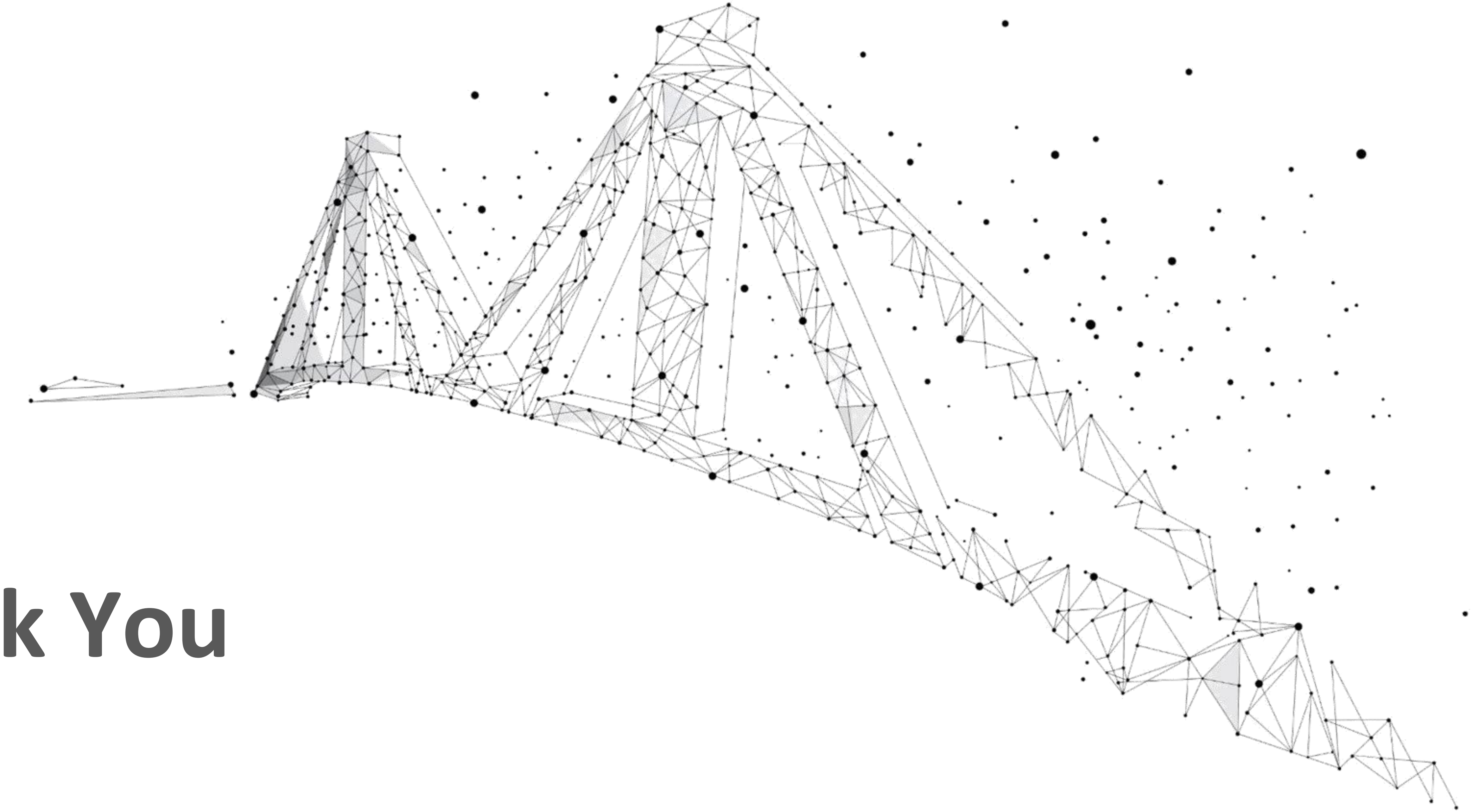
- Increase System Performance
- Better Scalability
- Reduces amount of data traversing the network





Scalable Performance for Distributed AI

System Configuration: Intel E5-2650V4, 12 cores @ 2.2GHz, 30M L2 cache, 9.6GT QPI, 256GB RAM: 16 x 16 GB DDR4, NVIDIA P100 GPUs, ConnectX-6 HCA, IB Quantum Switch (EDR speed), RH 7.5, Mellanox OFED 4.4, HPC-X v2.3, TensorFlow v1.11, Horovod 0.15.0



Thank You