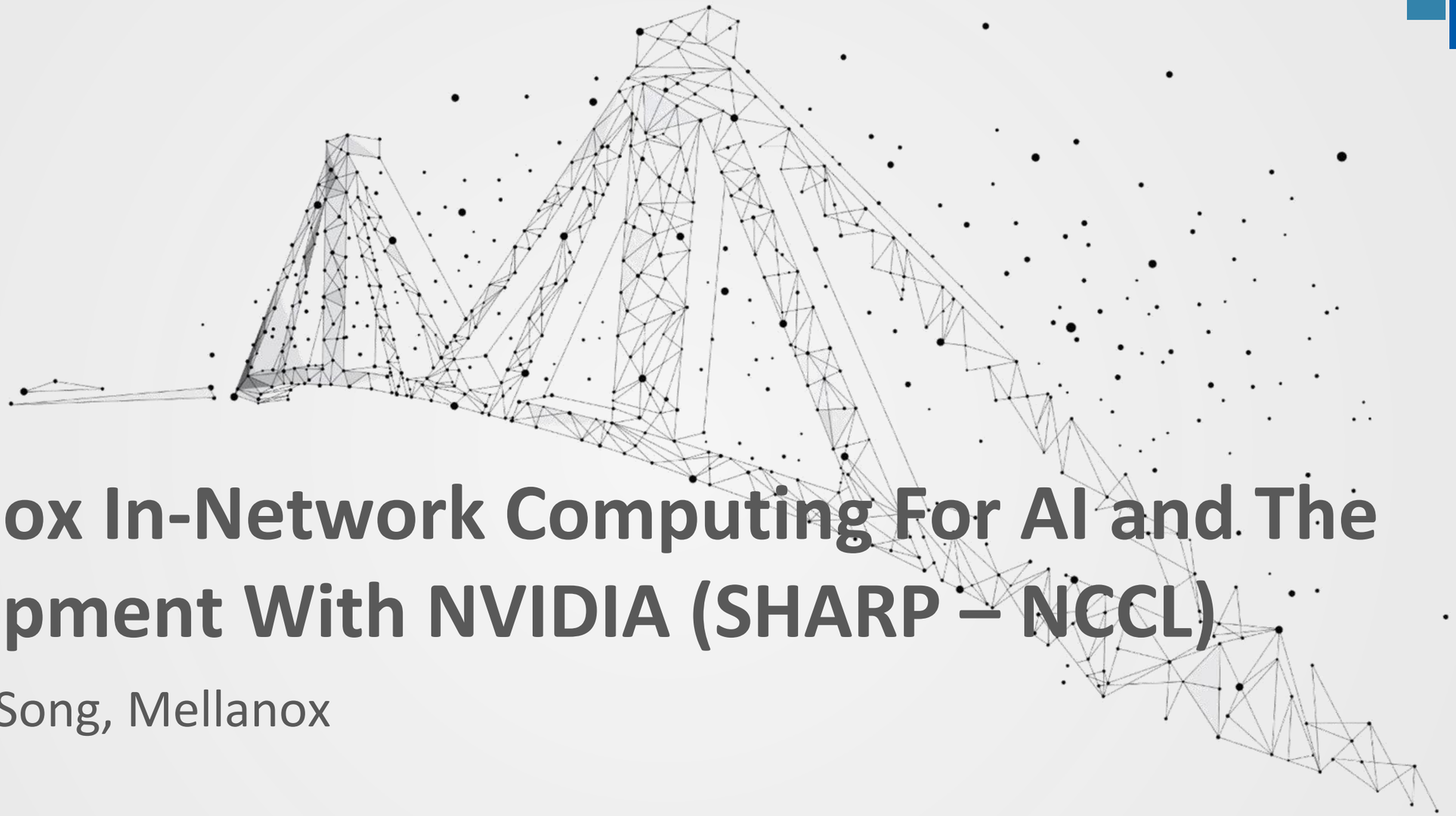


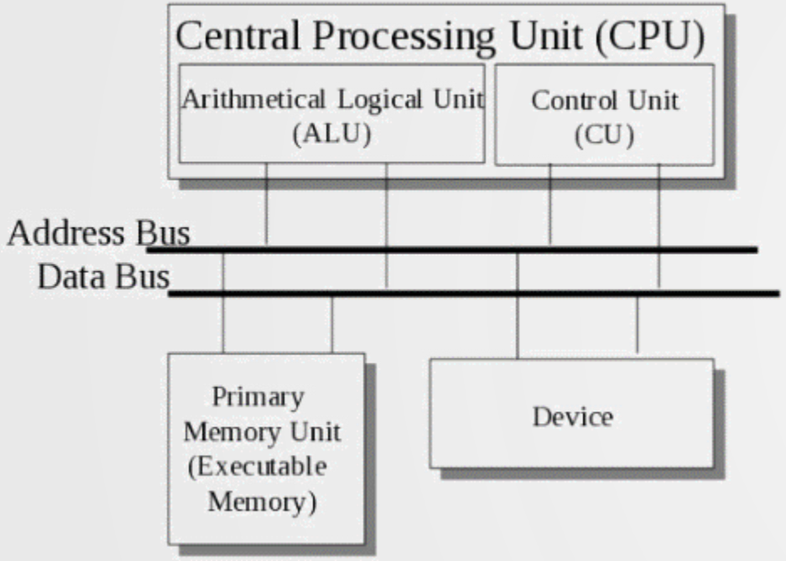
Mellanox In-Network Computing For AI and The Development With NVIDIA (SHARP – NCCL)

Qingchun Song, Mellanox

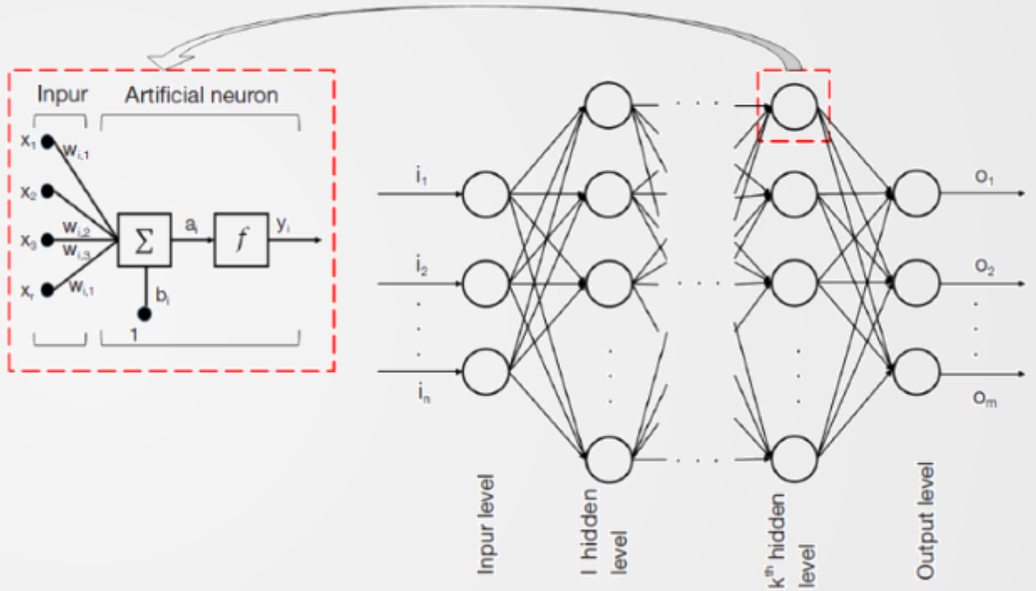
July 2, 2019

Data Processing Revolution – Data Centric

Compute-Centric



Data-Centric

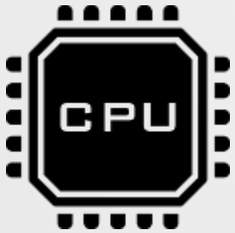


Von Neumann Machine

DataFlow Machine

CPU-Centric HPC/AI Center

Everything



CPU

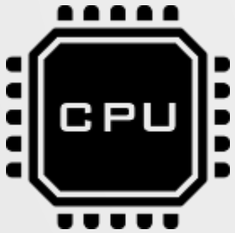
Network

Storage

Data-Centric HPC/AI Center

Workload

CPU Functions



In-CPU Computing

Workload

Communication Framework (MPI)

Network Functions

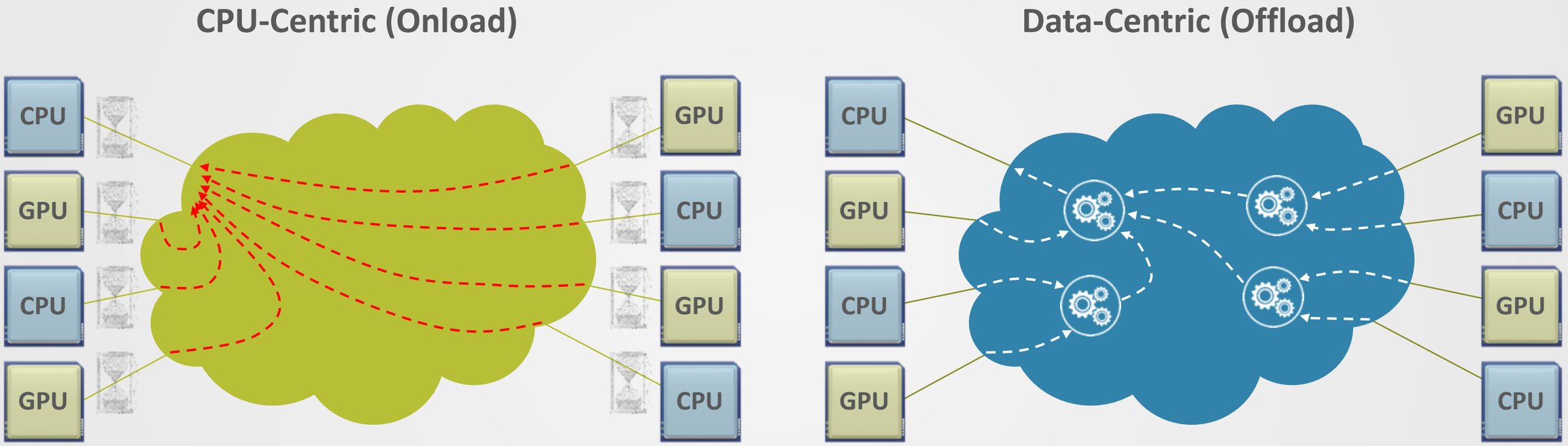
In-Network Computing

Workload

Storage Functions

In-Storage Computing

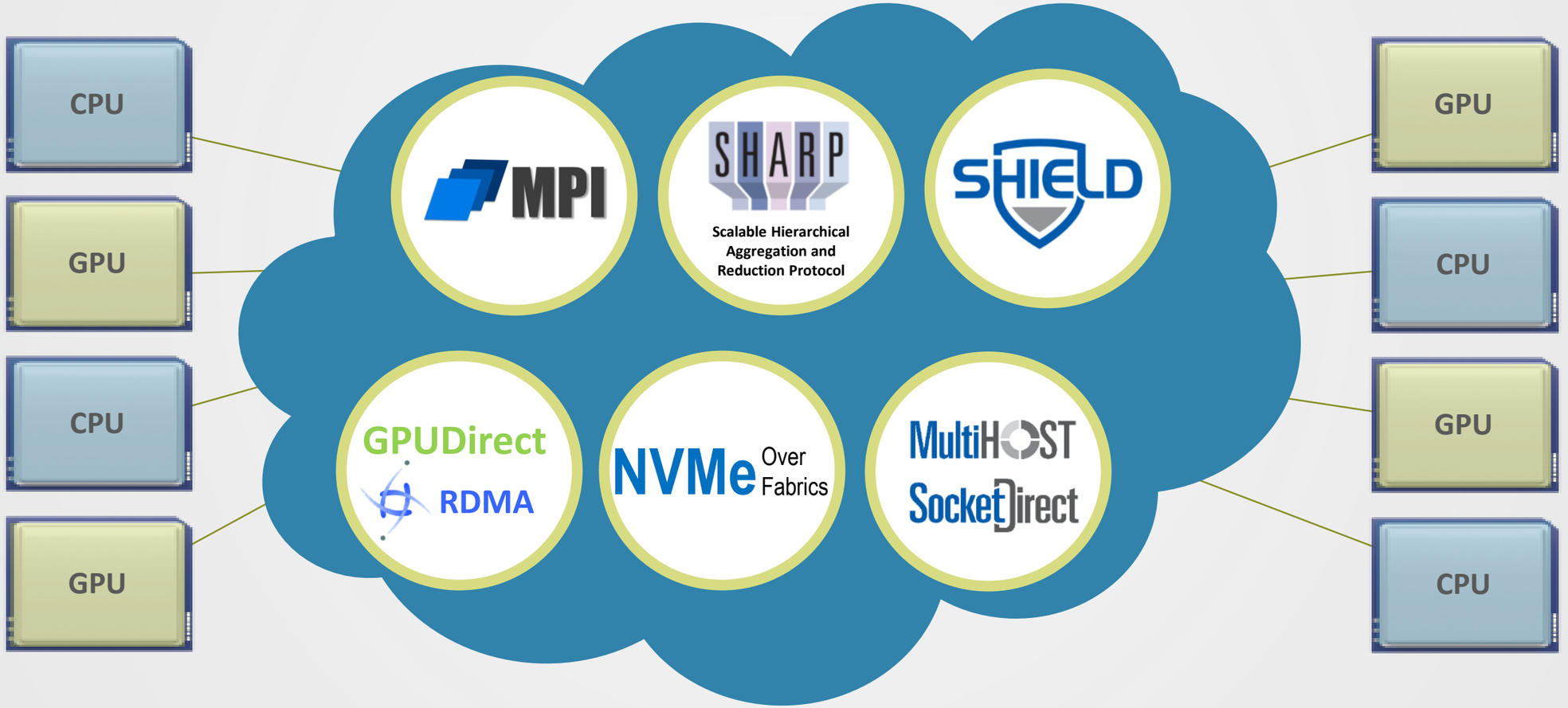
In-Network Computing Architecture



Communications Latencies of 30-40us

Communications Latencies of 3-4us

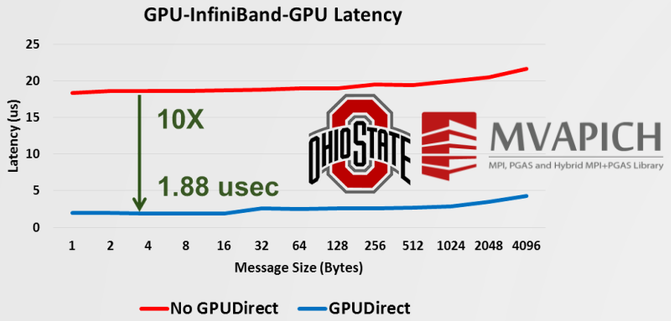
In-Network Computing to Enable Data-Centric Data Centers

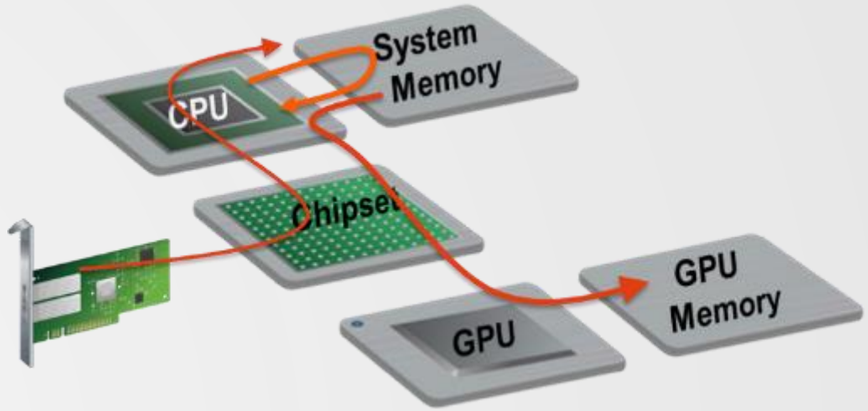


In-Network Computing Connects the World's Fastest HPC and AI Supercomputer

- Summit CORAL System, World's Fastest HPC / AI System
- Nvidia V100 GPU + InfiniBand HCA + In-Network Computing Fabric

GPUDirect RDMA Technology and Advantages

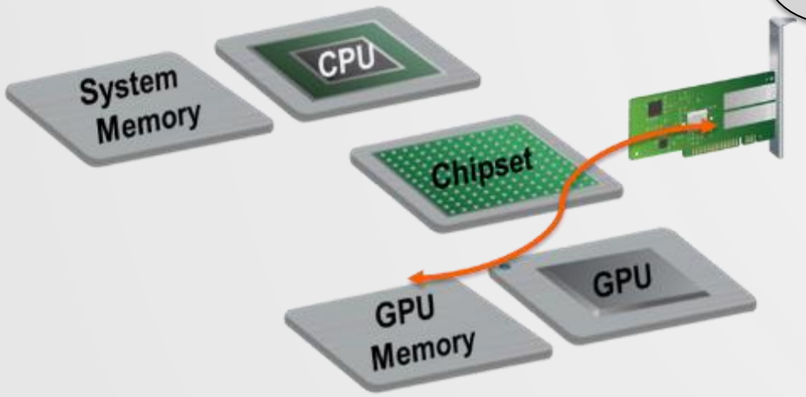


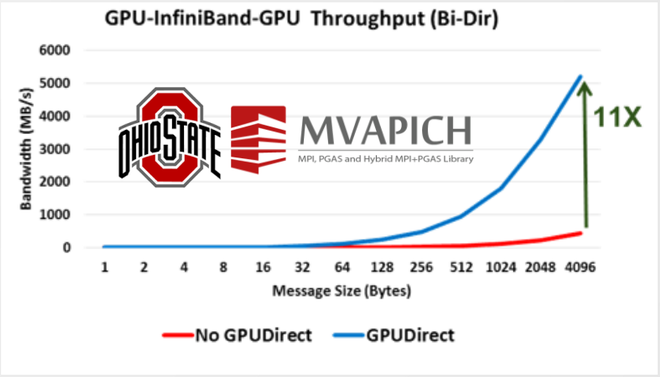


RDMA Enabled Network

Without GPU Direct
- Same Data Copied 3 Times

With GPUDirect



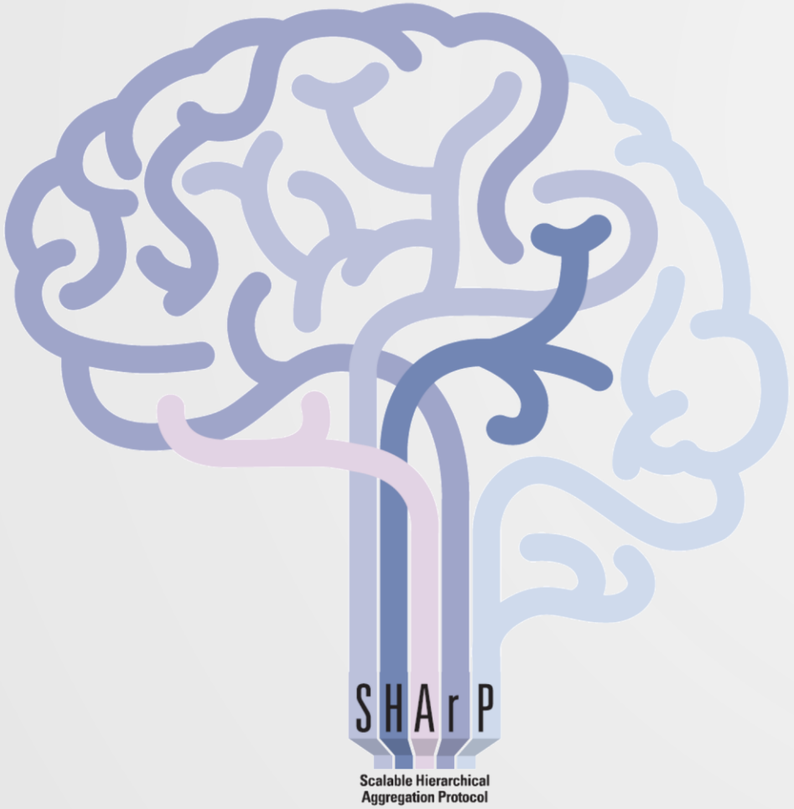


Scalable Hierarchical Aggregation And Reduction Protocol (SHARP)

Accelerating HPC and AI Applications

Accelerating HPC Applications

- Significantly reduce MPI collective runtime
- Increase CPU availability and efficiency
- Enable communication and computation overlap

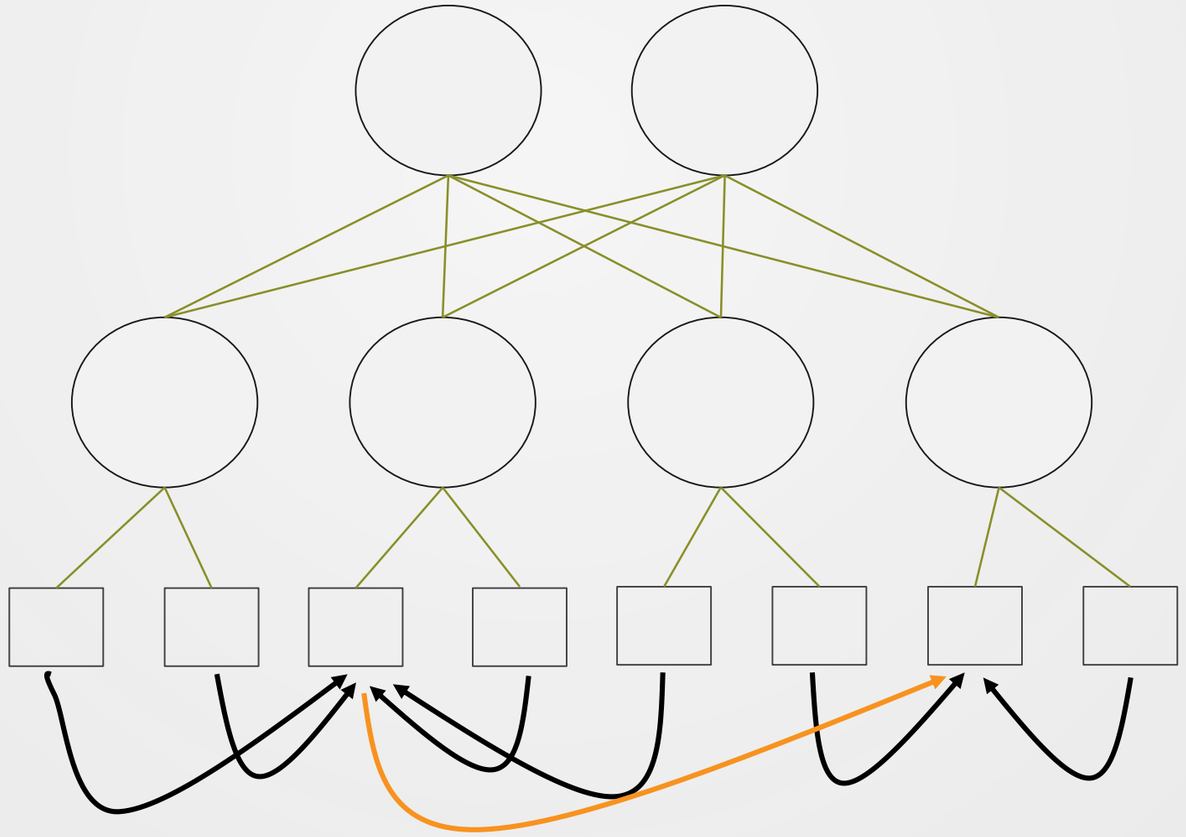


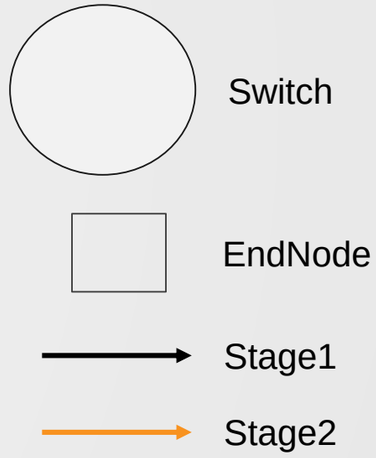
Enabling Artificial Intelligence Solutions to Perform Critical and Timely Decision Making

- Accelerating distributed machine learning

AllReduce Example – Trees

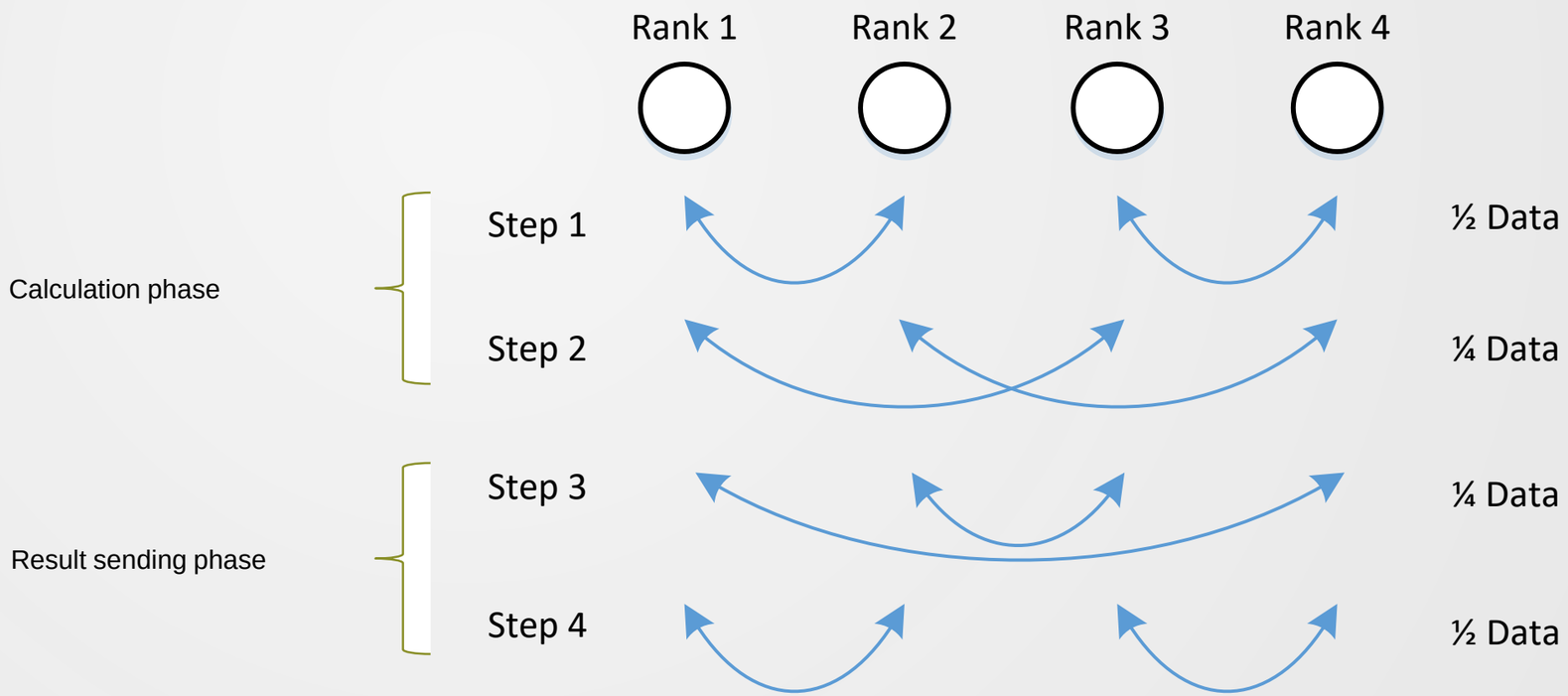
- Many2One and One2Many traffic patterns – possible network congestion
- Probably not a good solution for large data
- Large scale requires higher tree / larger radix
- Result distribution – over the tree / MC





AllReduce (Example) - Recursive Doubling

- The data is recursively divided, processed by CPUs and distributed
- The rank's CPUs are occupied performing the reduce algorithm
- The data is sent at least 2x times, consumes at least twice the BW



Scalable Hierarchical Aggregation Protocol

Reliable Scalable General Purpose Primitive, Applicable to Multiple Use-cases

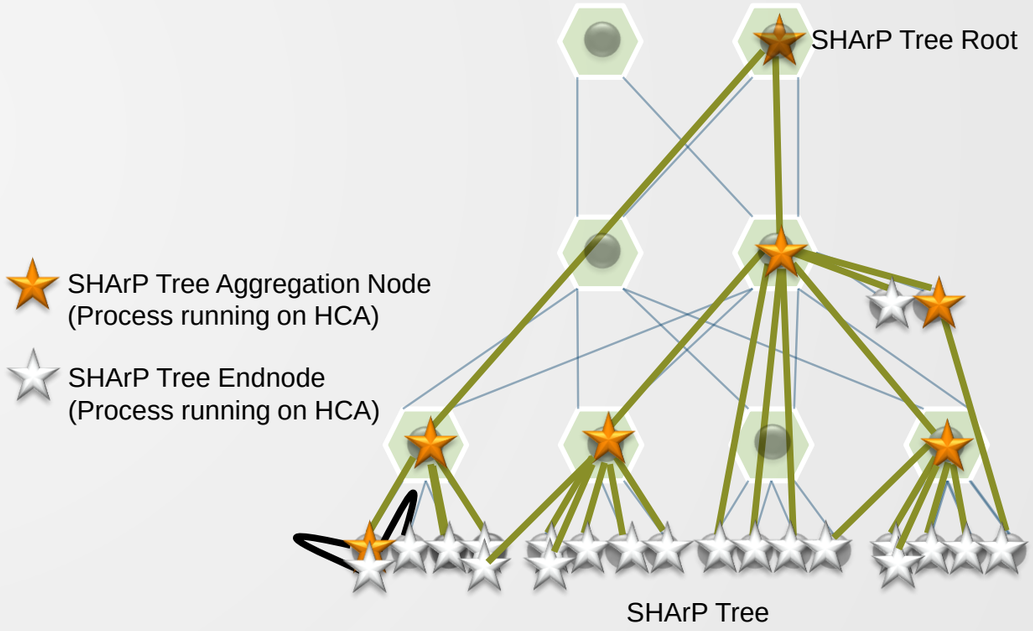
- In-network Tree based aggregation mechanism
- Large number of groups
- Multiple simultaneous outstanding operations
- Streaming aggregation

Accelerating HPC applications

- Scalable High Performance Collective Offload
  - Barrier, Reduce, All-Reduce, Broadcast
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
  - Integer and Floating-Point, 16 / 32 / 64 bit
  - Up to 1KB payload size (in Quantum)
- Significantly reduce MPI collective runtime
- Increase CPU availability and efficiency
- Enable communication and computation overlap

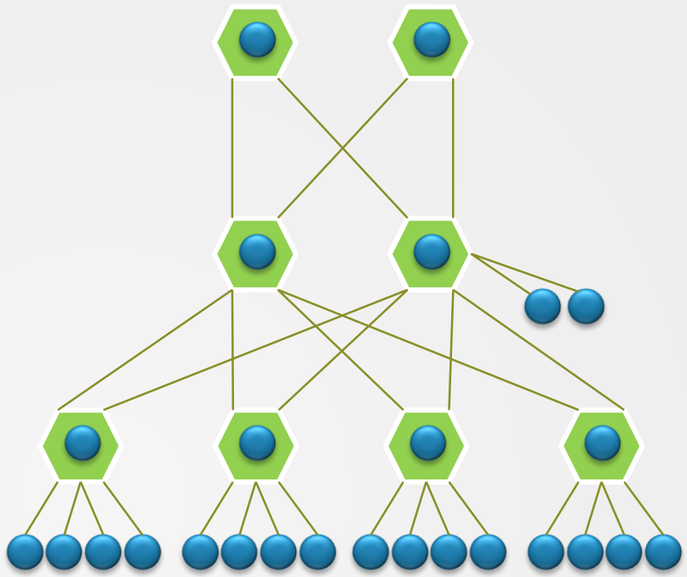
Accelerating Machine Learning applications

- Proven the *many-to-one* Traffic Pattern
- CUDA , GPUDirect RDMA

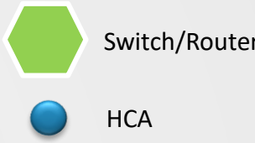


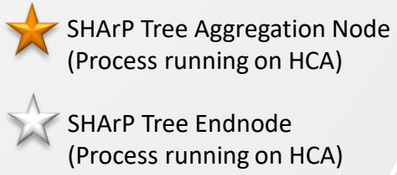
Scalable Hierarchical Aggregation Protocol

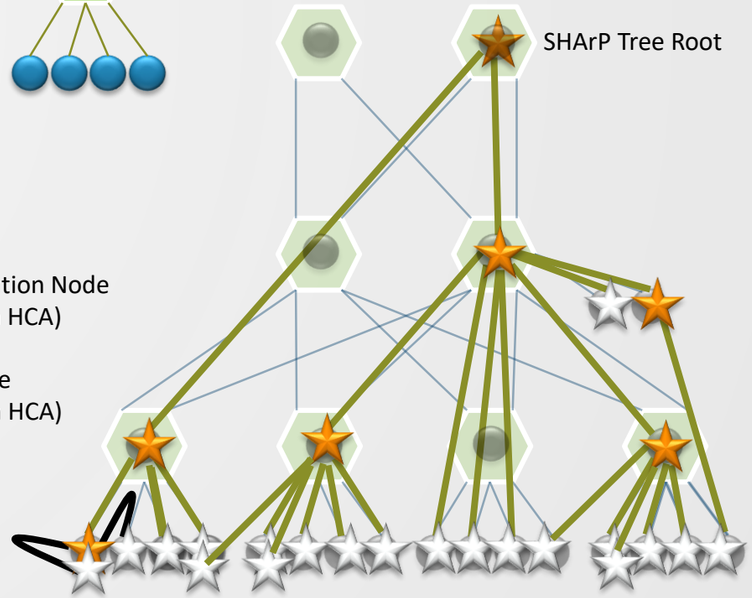
- SHARP Tree is a Logical Construct
  - Nodes in the SHArP Tree are IB Endnodes
  - Logical tree defined on top of the physical underlying fabric
  - SHArP Tree Links are implemented on top of the IB transport (Reliable Connection)
  - Expected to follow the physical topology for performance but not required
- SHARP Operations are Executed by a SHARP Tree
  - Multiple SHArP Trees are Supported
  - Each SHArP Tree can handle Multiple Outstanding SHArP Operations
  - Within a SHArP Tree, each Operation is Uniquely Identified by a SHArP-Tuple
    - GroupID
    - SequenceNumber



Physical Topology



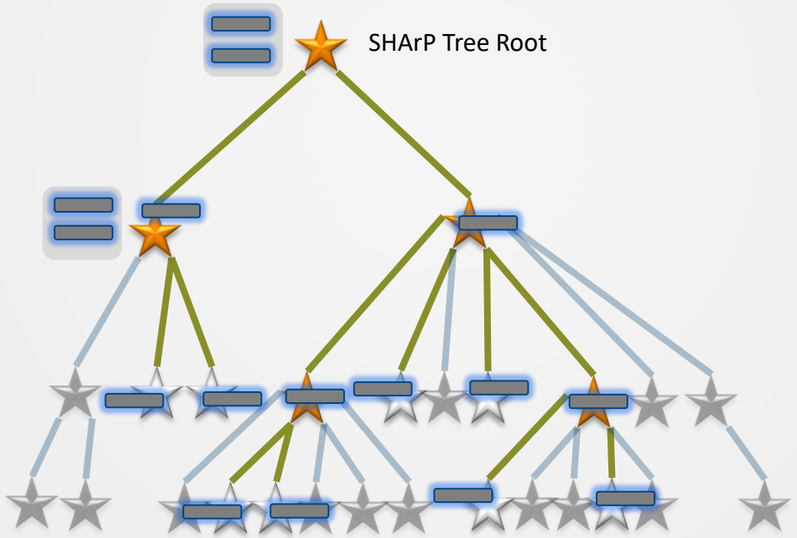




One SHArP Tree

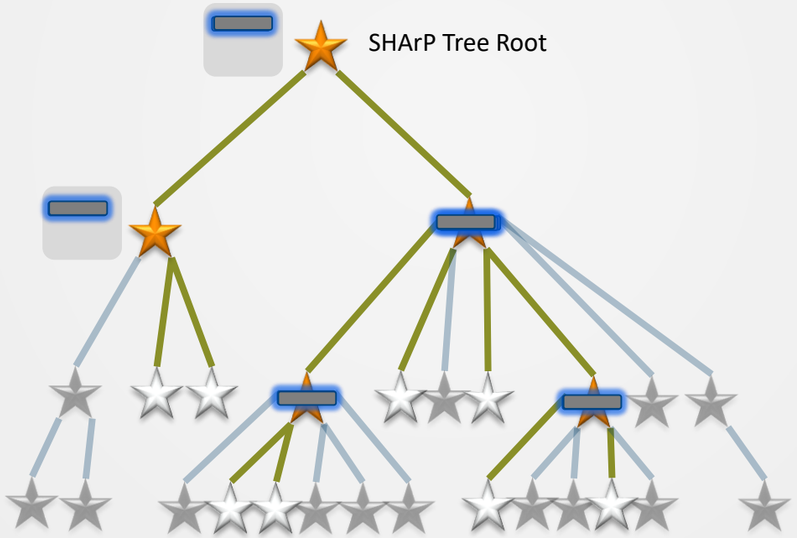
SHARP Principles of Operation - Request

Aggregation Request



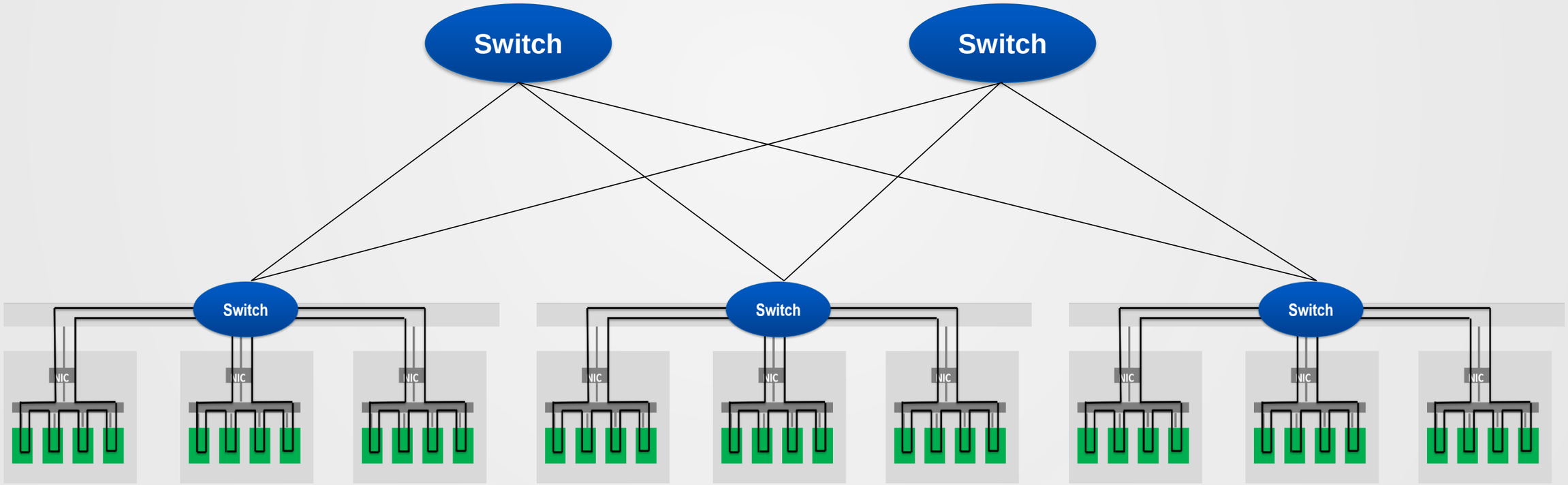
SHARP Principles of Operation – Response

Aggregation
Response



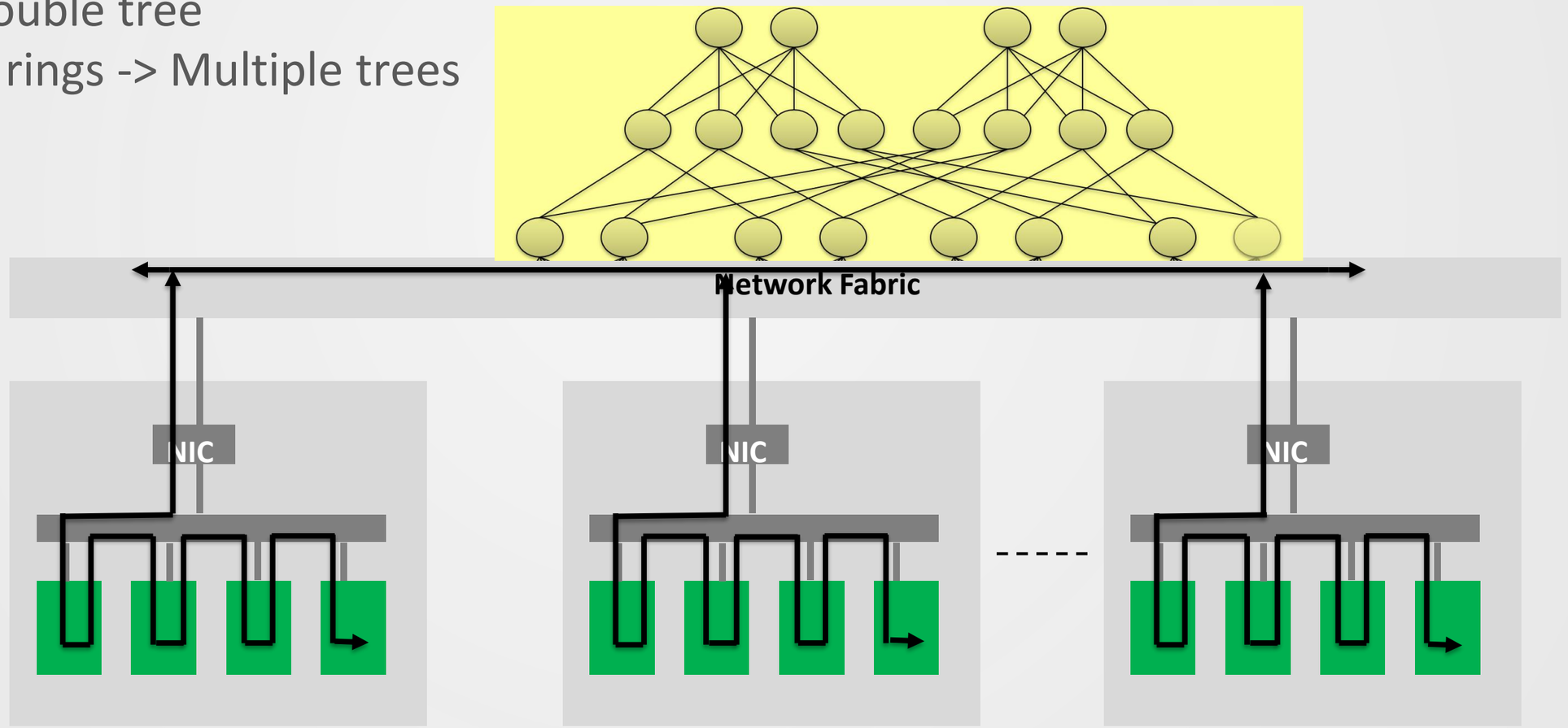
NCCL Ring

- Simple
- Linear Latency
- Support in NCCL-2.3 & previous version
- Multiple rings

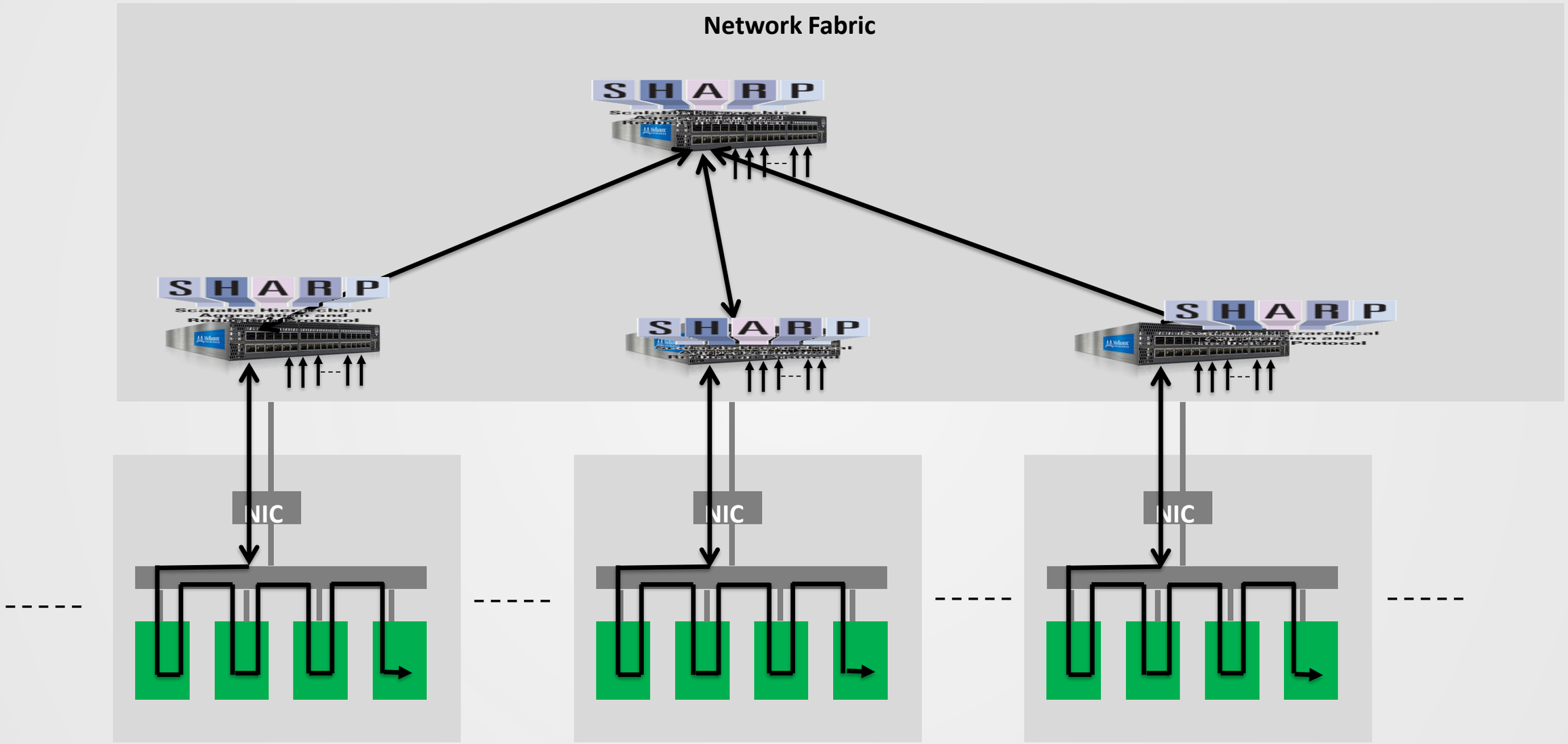


NCCL Tree

- Support added in NCCL-2.4
- Keep Intra-node chain
- Node leaders participate in tree
- Binary double tree
- Multiple rings -> Multiple trees

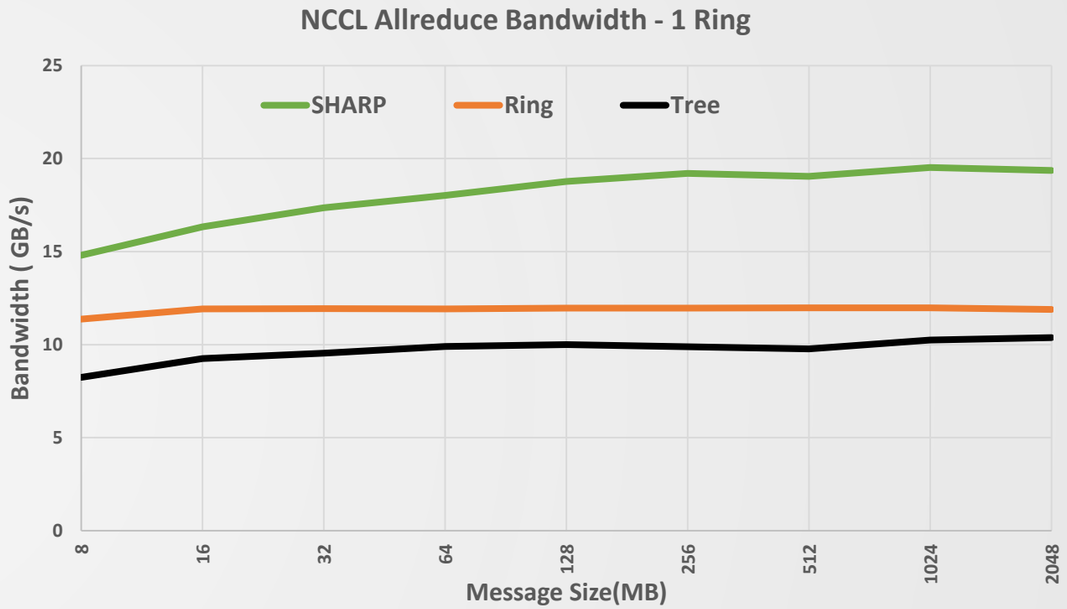


NCCL SHARP



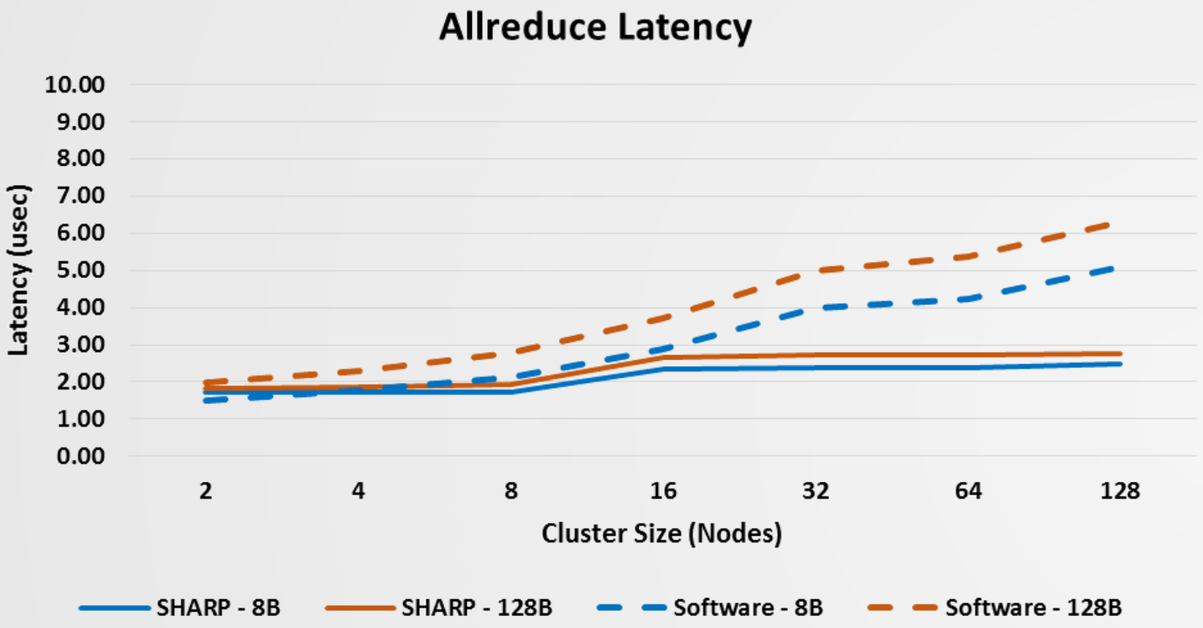
NCCL SHARP

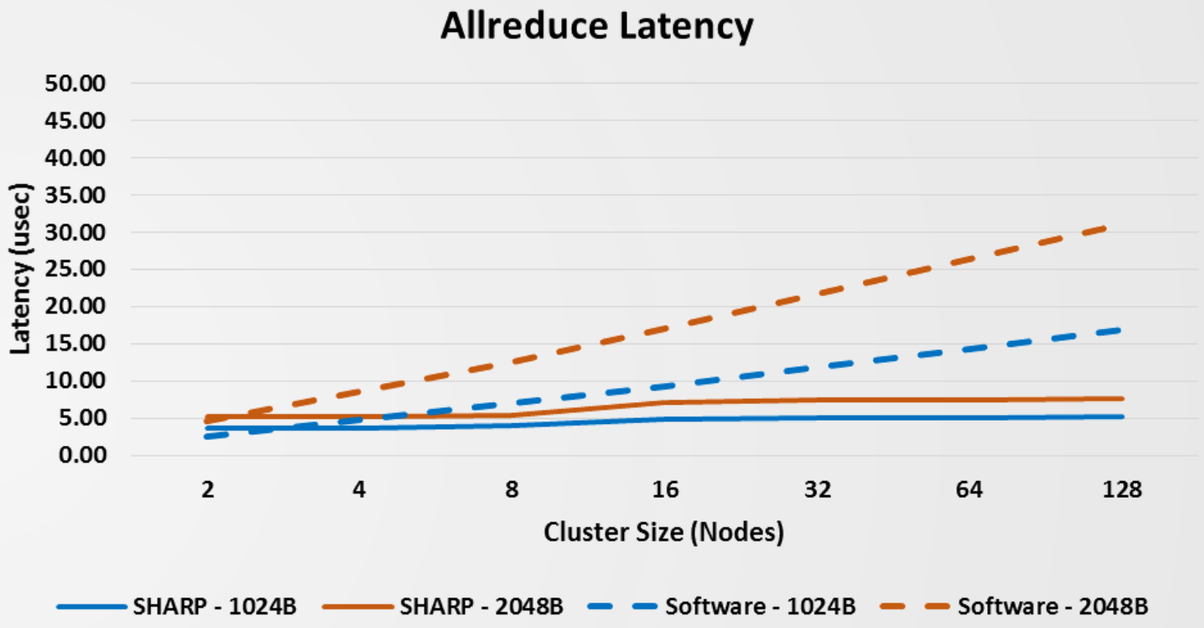
- Collective network Plugin
- Replace Inter-node tree with SHARP Tree
- Keeps Intra-node ring
- Aggregation in network switch
- Streaming from GPU memory with GPU Direct RDMA
- 2x BW compared to NCCL-TREE



SHARP Enables 2X Higher Data Throughput for NCCL

SHARP AllReduce Performance Advantages (128 Nodes)





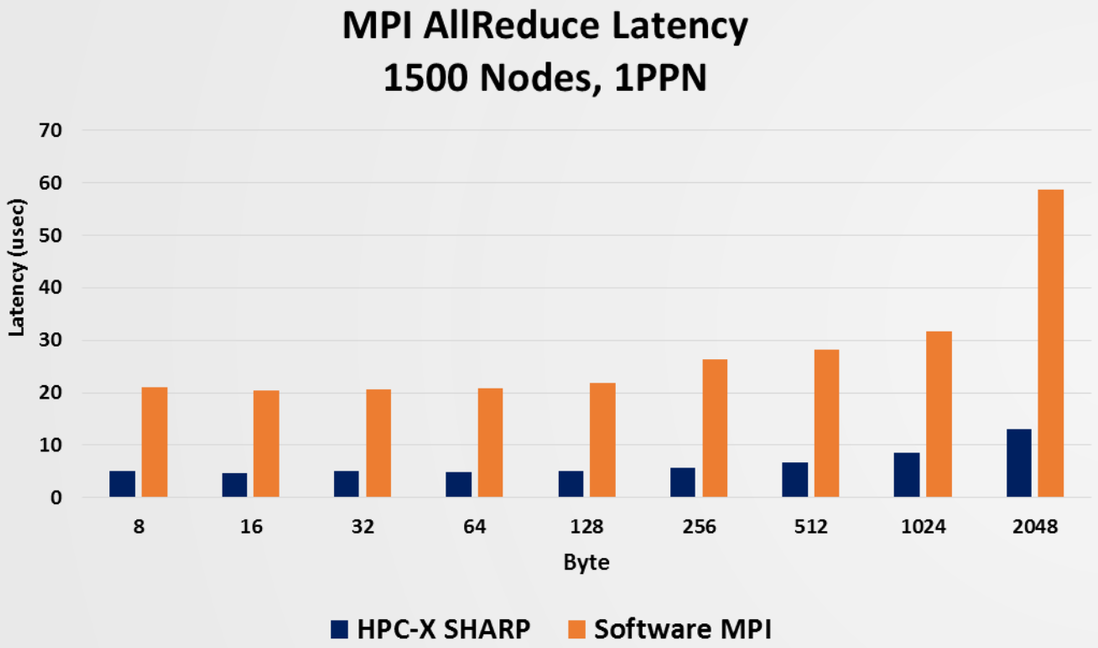
SHARP

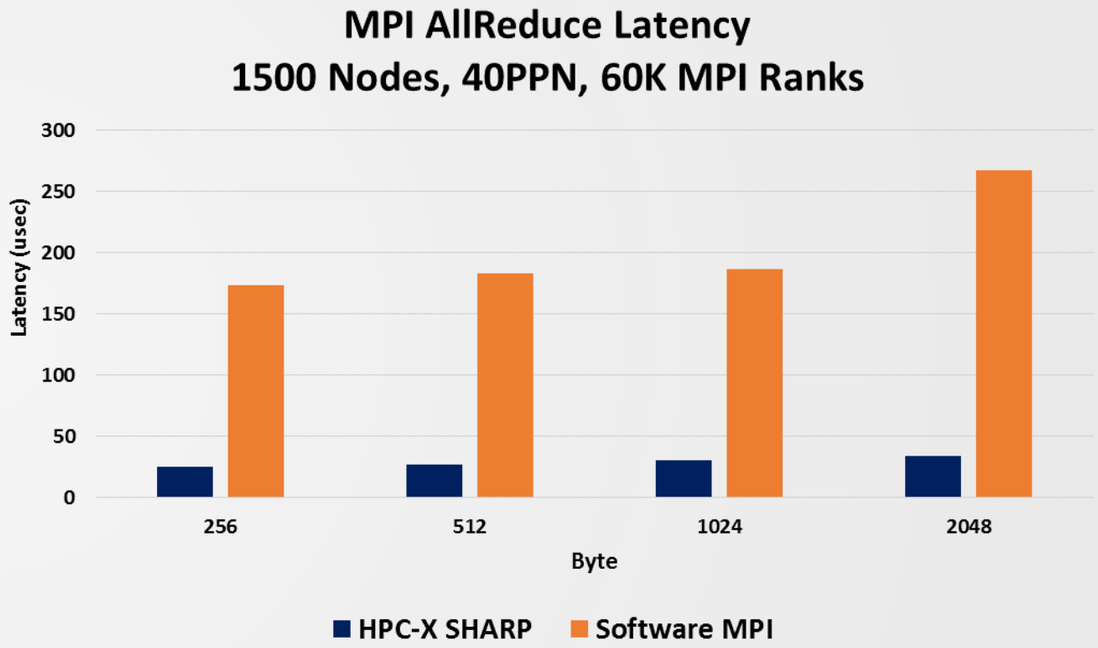
Scalable Hierarchical Aggregation and Reduction Protocol

SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

SHARP AllReduce Performance Advantages

1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology



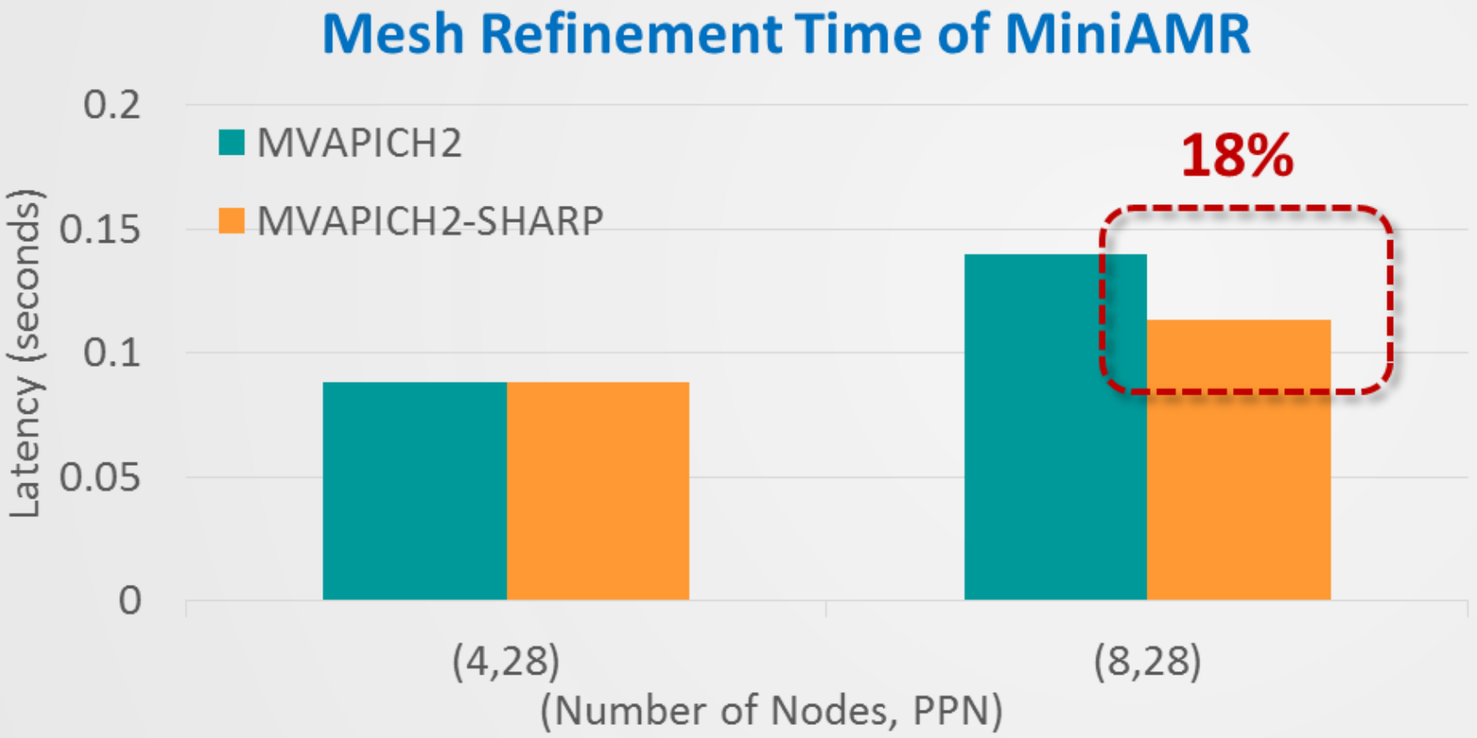


SHARP

Scalable Hierarchical Aggregation and Reduction Protocol

SHARP Enables Highest Performance

SHARP Performance – Application (OSU)



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

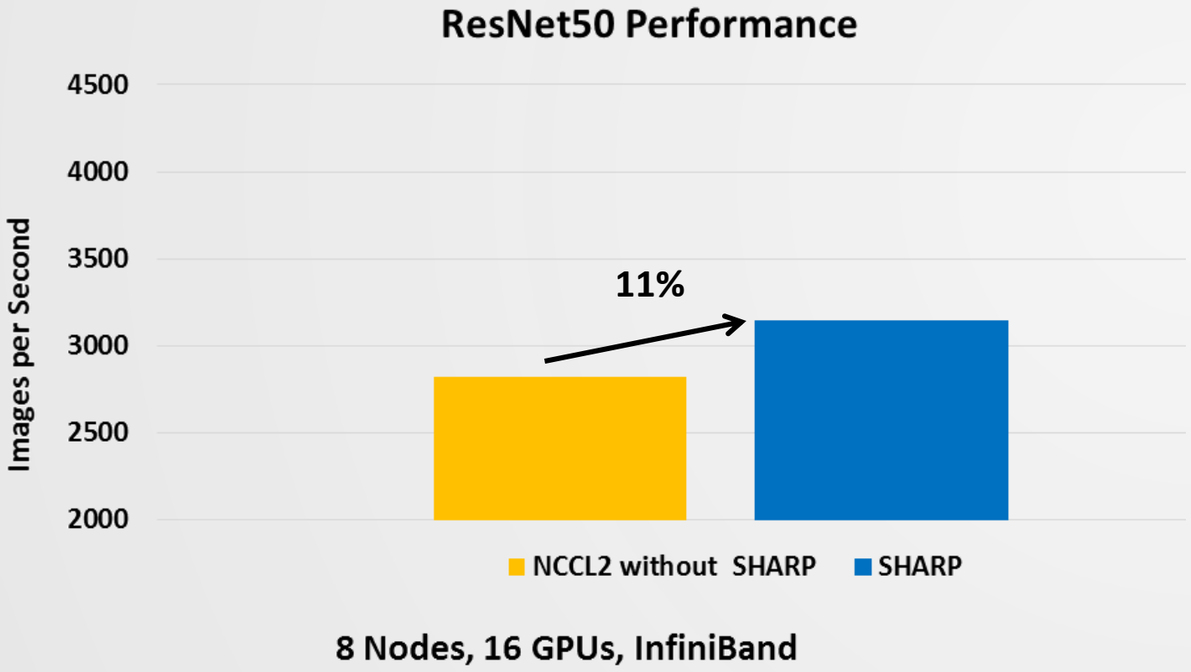
MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

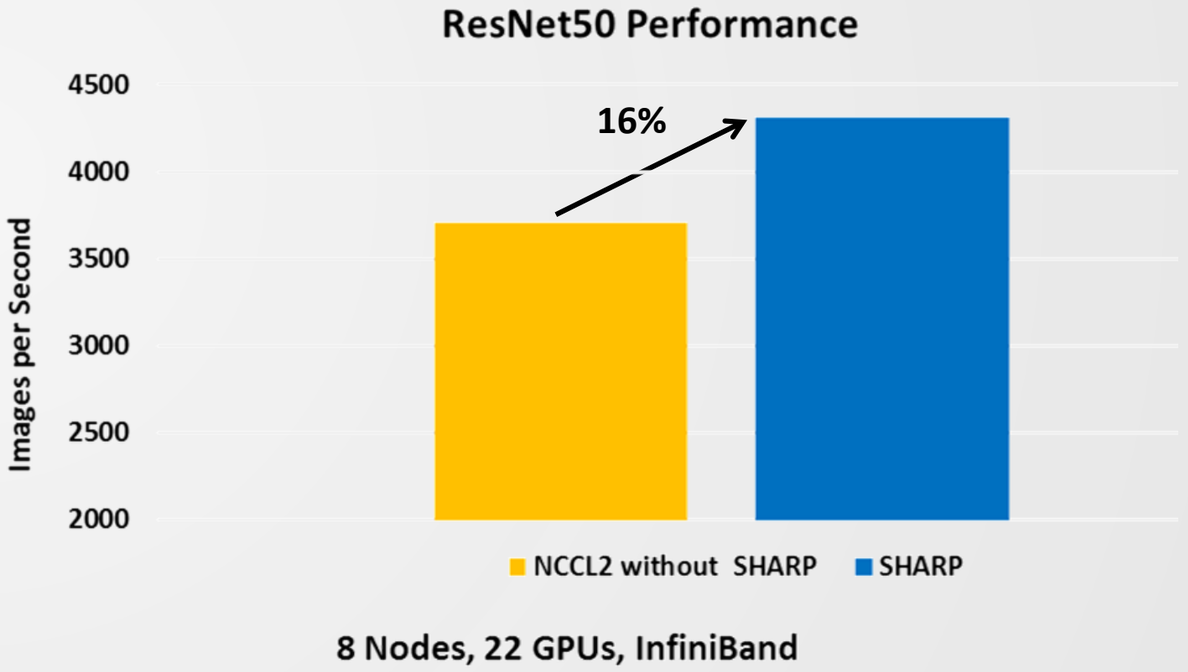
The MVAPICH2 Project
http://mvapich.cse.ohio-state.edu/

Source: Prof. DK Panda, Ohio State University

SHARP Performance Advantage for AI

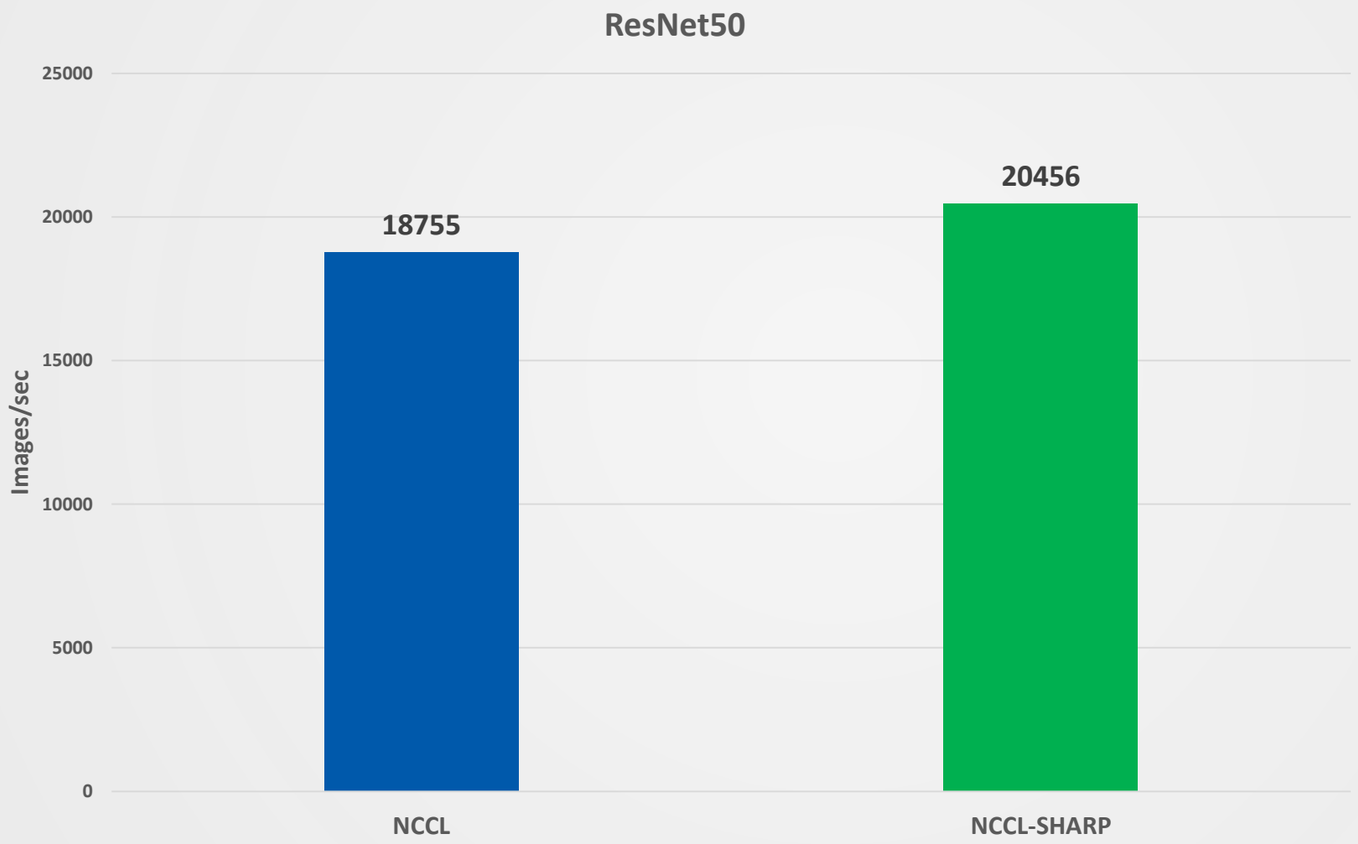
- SHARP provides 16% Performance Increase for deep learning, initial results
- TensorFlow with Horovod running ResNet50 benchmark, HDR InfiniBand (ConnectX-6, Quantum)





P100 NVIDIA GPUs, RH 7.5, Mellanox OFED 4.4, HPC-X v2.3, TensorFlow v1.11, Horovod 0.15.0

NCCL-SHARP Performance – DL Training



System Configuration: (4) HPE Apollo 6500 systems configured with (8) NVIDIA Tesla V100 SXM2 16GB, (2) HPE DL360 Gen10 Intel Xeon-Gold 6134 (3.2 GHz/8-core/130 W) CPUs, (24) DDR4-2666 CAS-19-19-19 Registered Memory Modules
HPE 1.6 TB NVMe SFF (2.5") SSD, ConnectX-6 HCA, IB Quantum Switch (EDR speed), Ubuntu 16.04

Accelerating All Levels of HPC / AI Frameworks

Application

- Data Analysis
- Real Time
- Deep Learning

Communication

- Mellanox SHARP In-Network Computing
- MPI Tag Matching
- MPI Rendezvous
- Software Defined Virtual Devices

MPI HPC-X™ UCF Unified Communication Framework SHARP Scalable Hierarchical Aggregation and Reduction Protocol

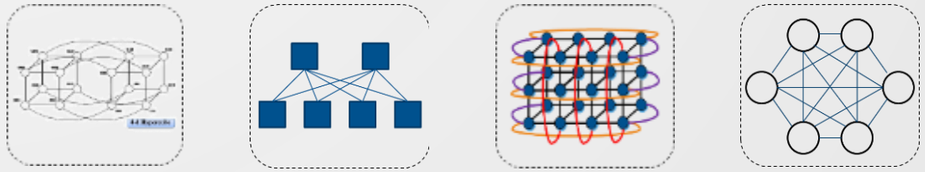
Network

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

GPUDirect RDMA SHIELD

Connectivity

- Multi-Host Technology
- Socket-Direct Technology
- Enhanced Topologies



Thank You