OPENFABRICS ALLIANCE

15th ANNUAL WORKSHOP 2019

In Network Computing

Tomislav (Tommy) Janjusic

March 19, 2019

Mellanox TECHNOLOGIES

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

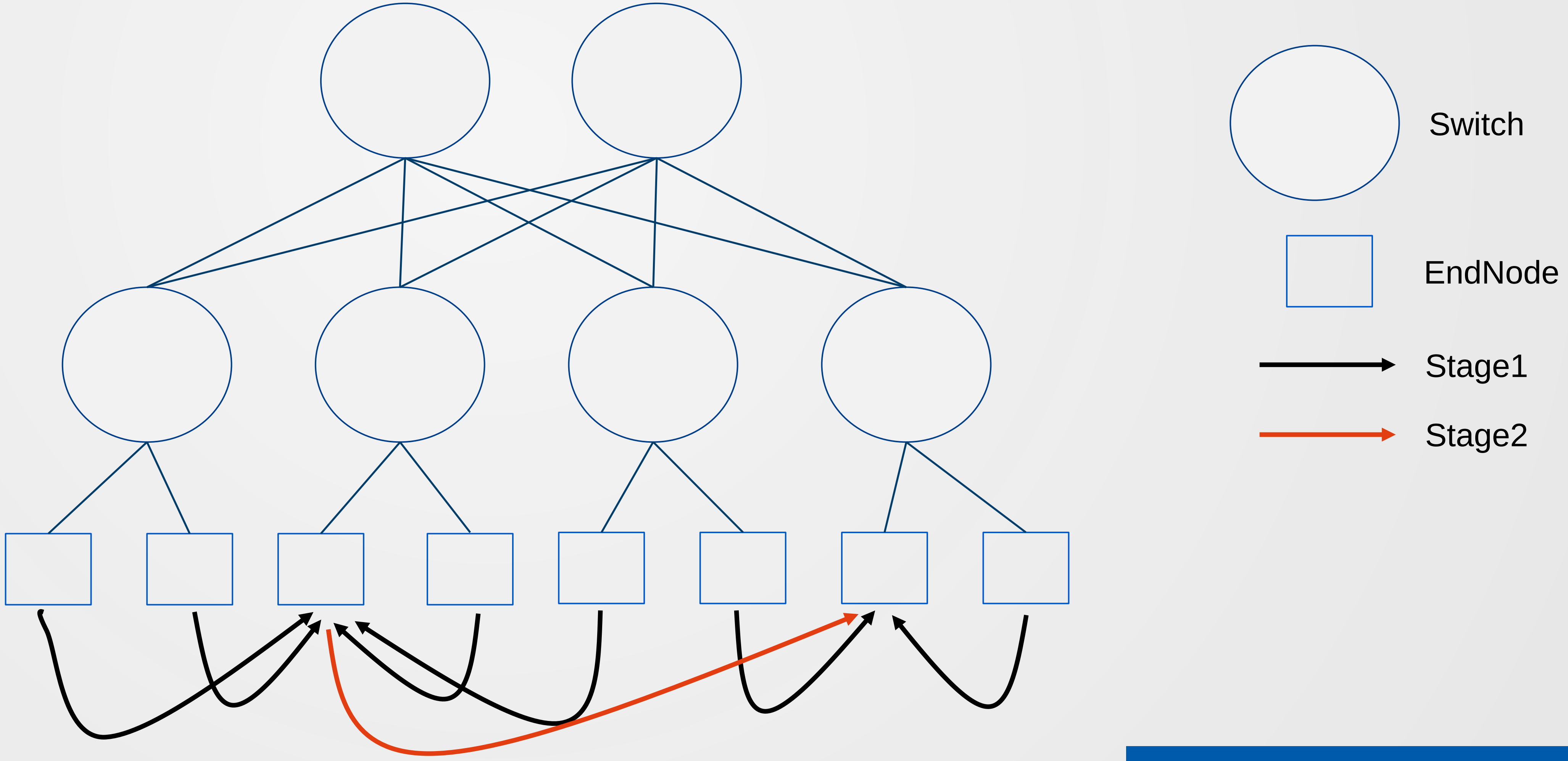
- Accelerating distributed machine learning

Aggregation Operations

- Allreduce – vector operation, reduce results and distribute to participating processing elements within the group (MPI Ranks)
- Reduce – similar to Allreduce, but result is sent to only one processing element (root).
- Gather / Allgather – vector concatenation operation

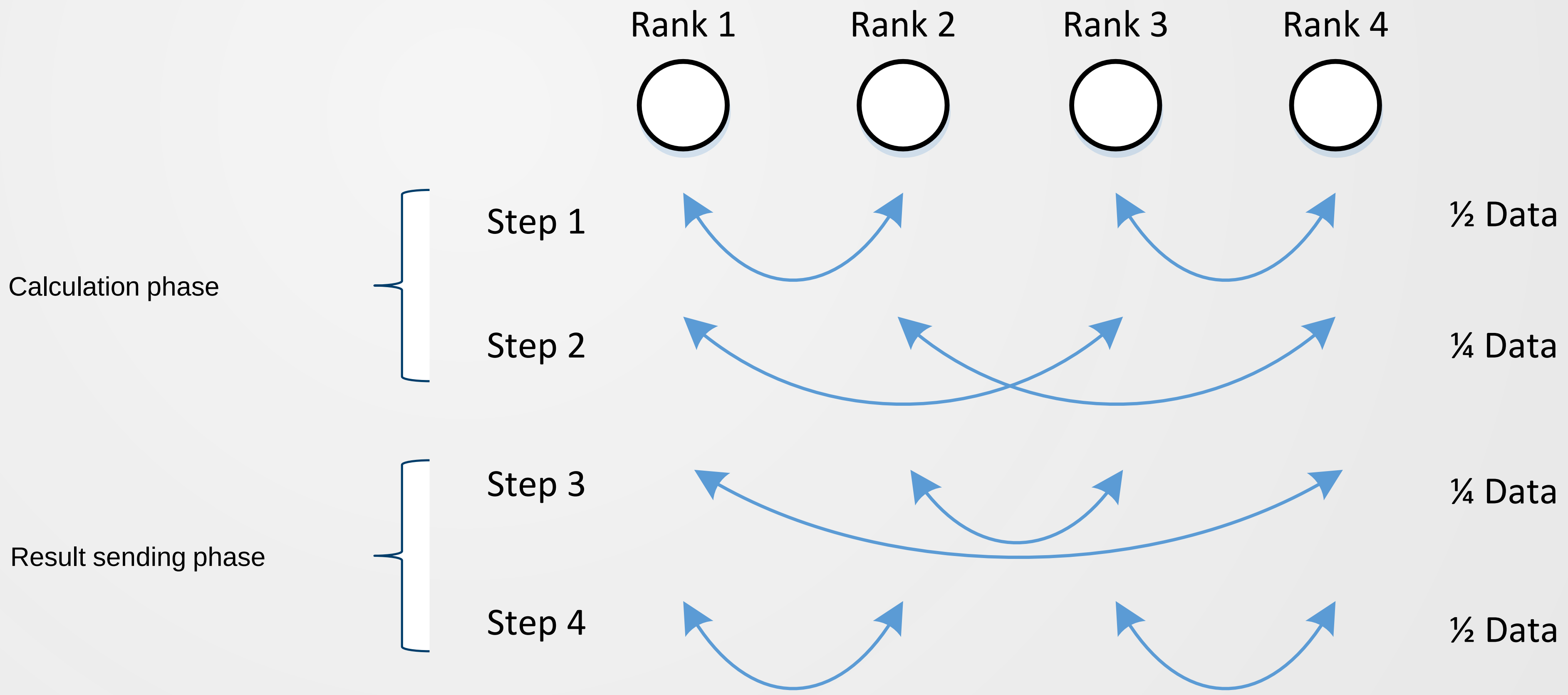
Collective (Example) – Trees

- 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



Collective (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**



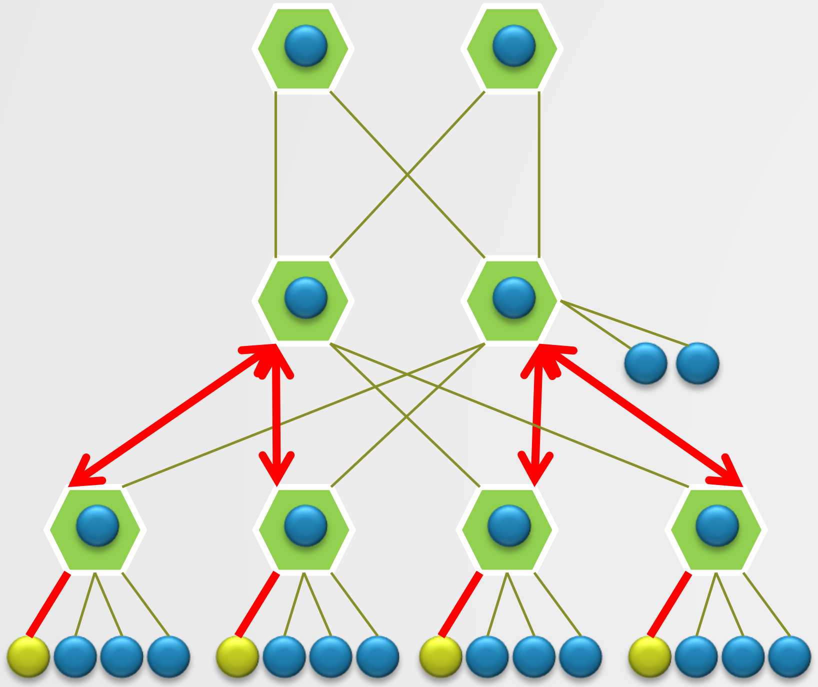
Which Offload Should We Suggest?

- Lets aggregate the data while it is going through the network…
  - It will reduce the amount of data running through the network
  - It will reduce the latency because data will go through a shorter path
  - The operation will be fully offloaded

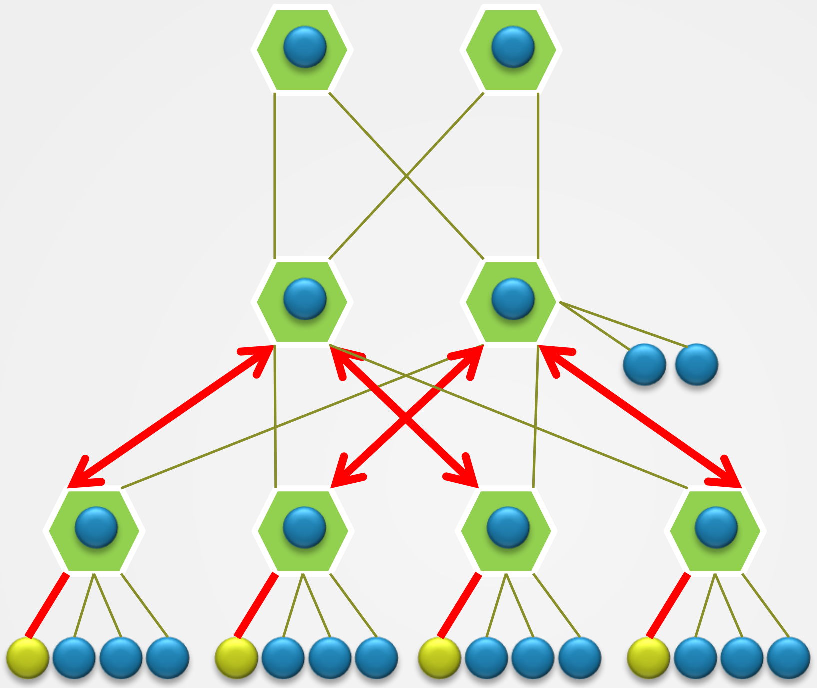
SHARP

Scalable Hierarchical Aggregation and Reduction Protocol

HCOLL: SHARP vs No-SHARP

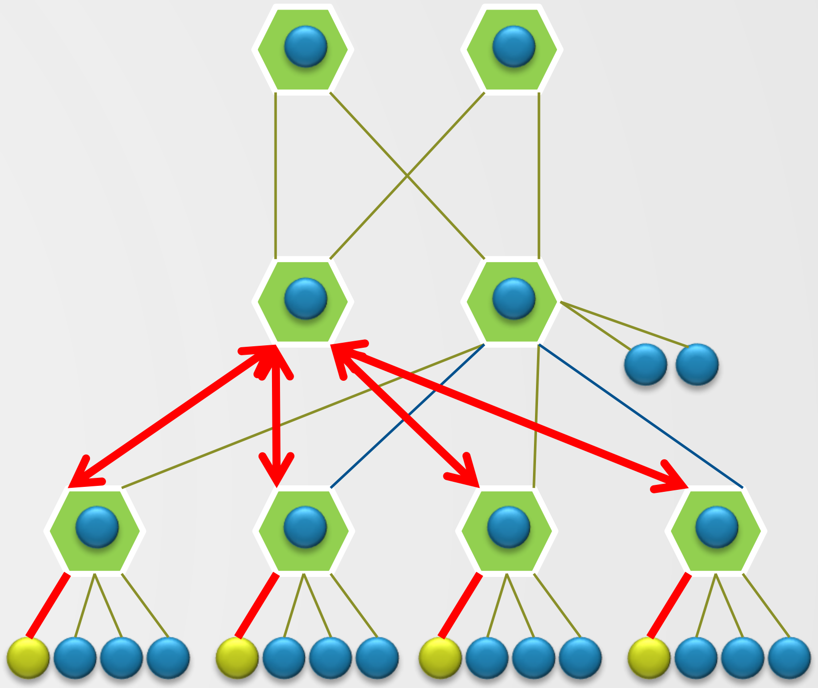


Step 1 `



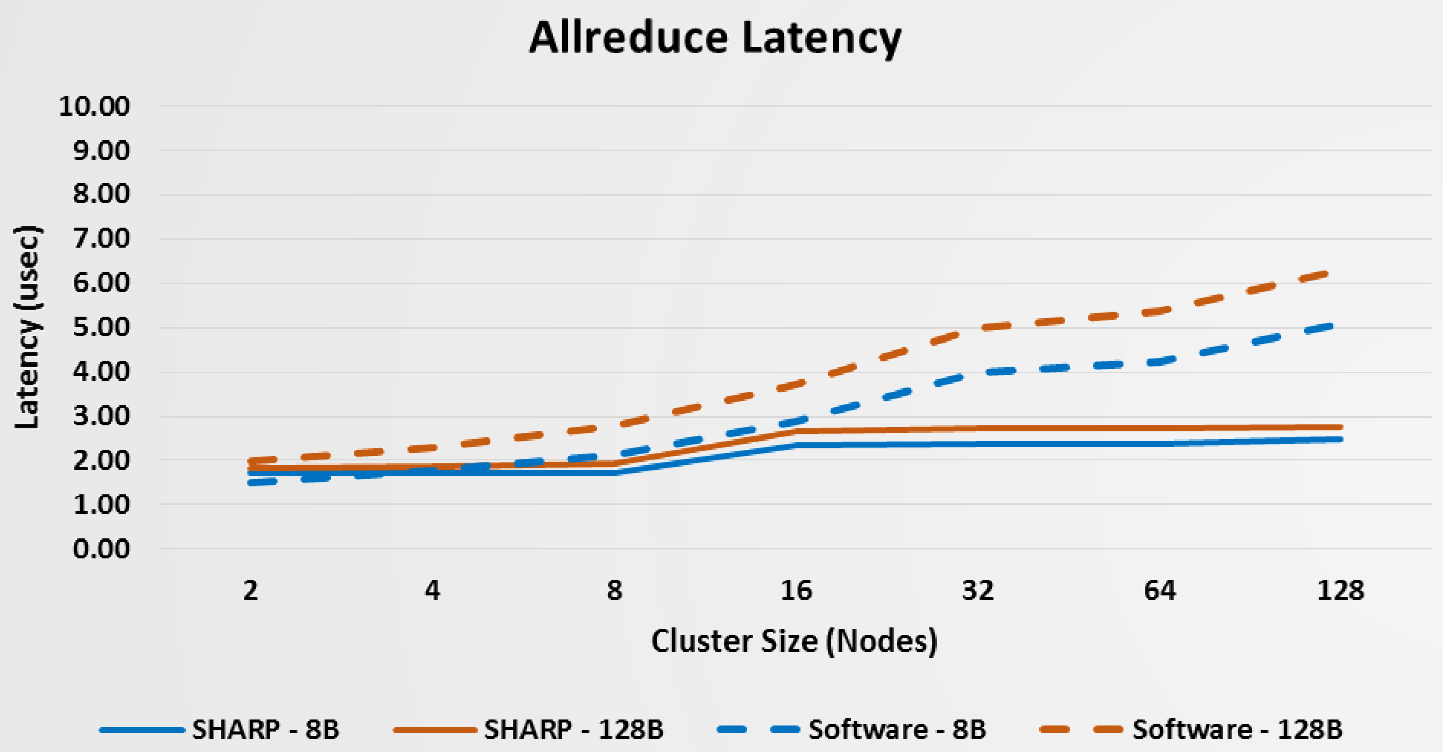
Step 2 `

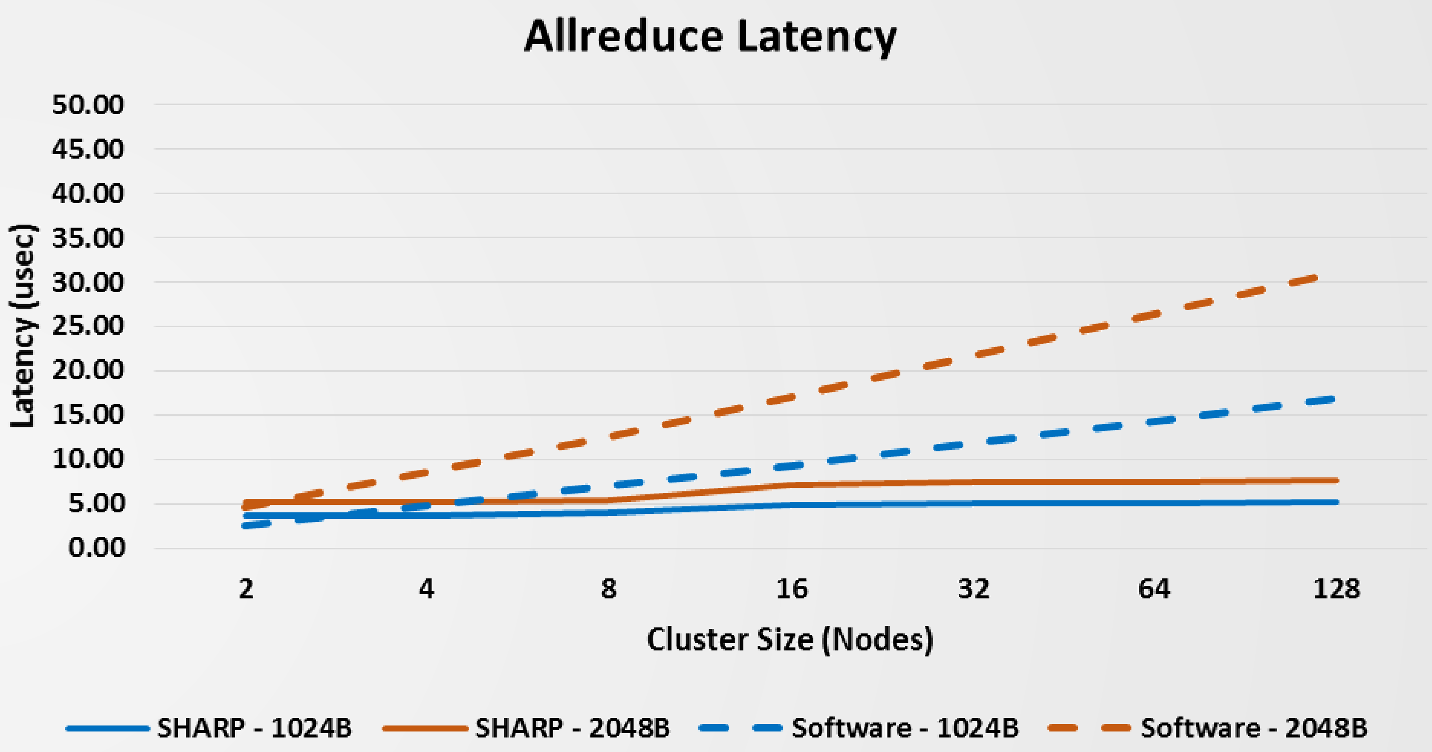
Recursive Doubling



SHARP

SHARP AllReduce Performance Advantages (128 Nodes)





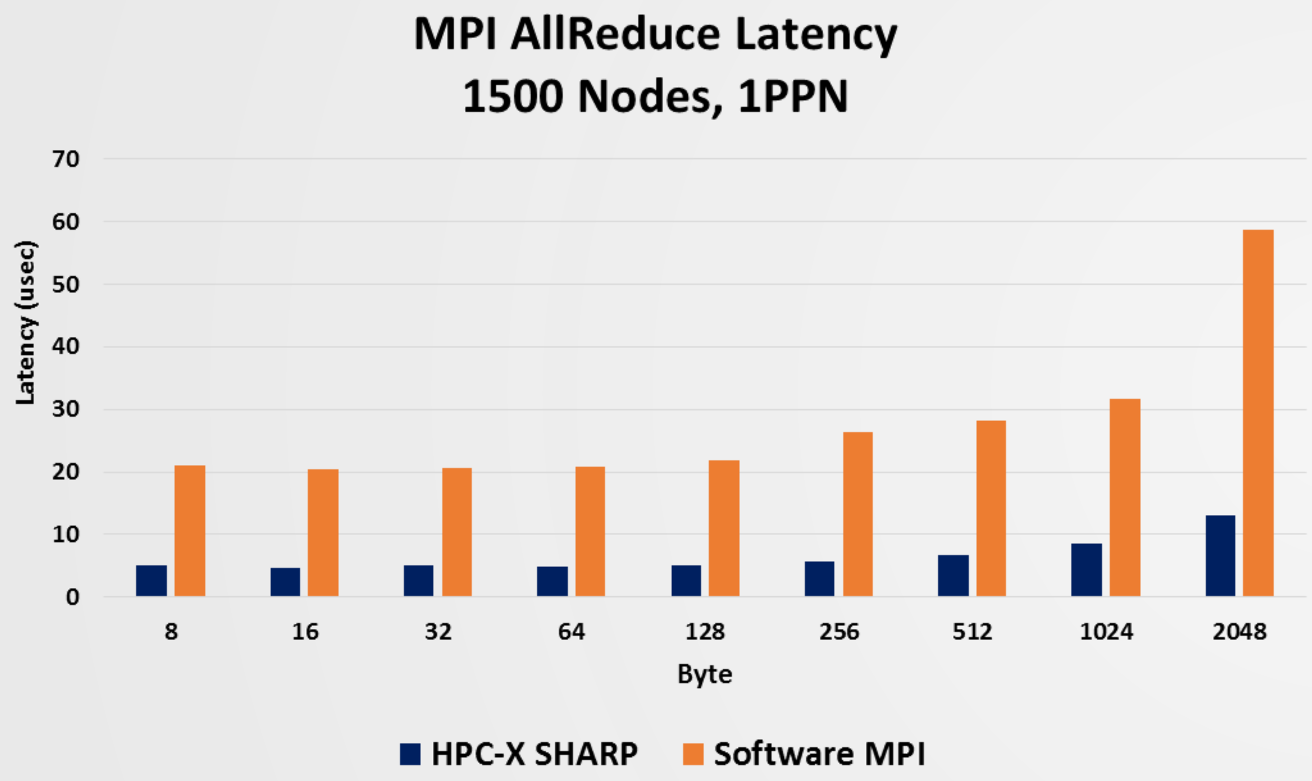
SHARP

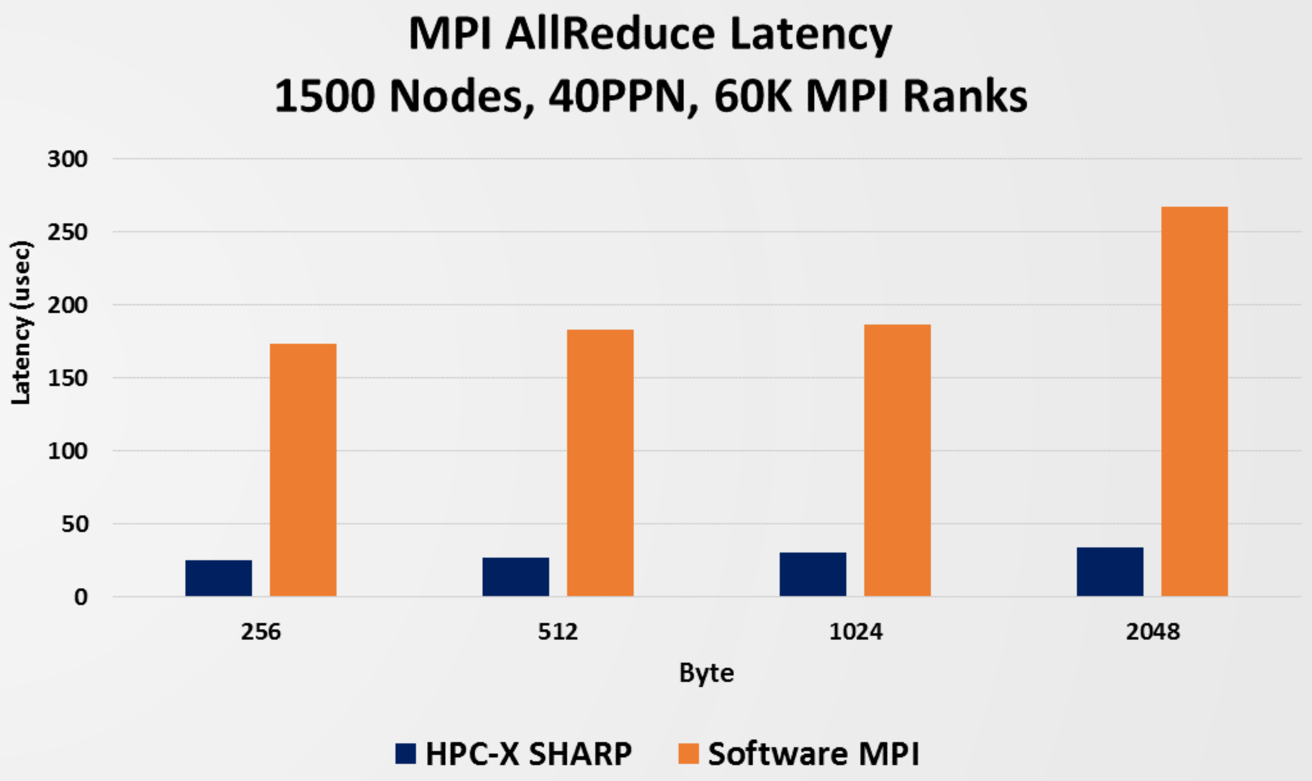
Scalable Hierarchical Aggregation and Reduction Protocol

SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

SHARP AllReduce Performance Advantages

1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology





Scalable Hierarchical Aggregation and Reduction Protocol

SHARP Enables Highest Performance

Scalable Hierarchical Aggregation Protocol

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

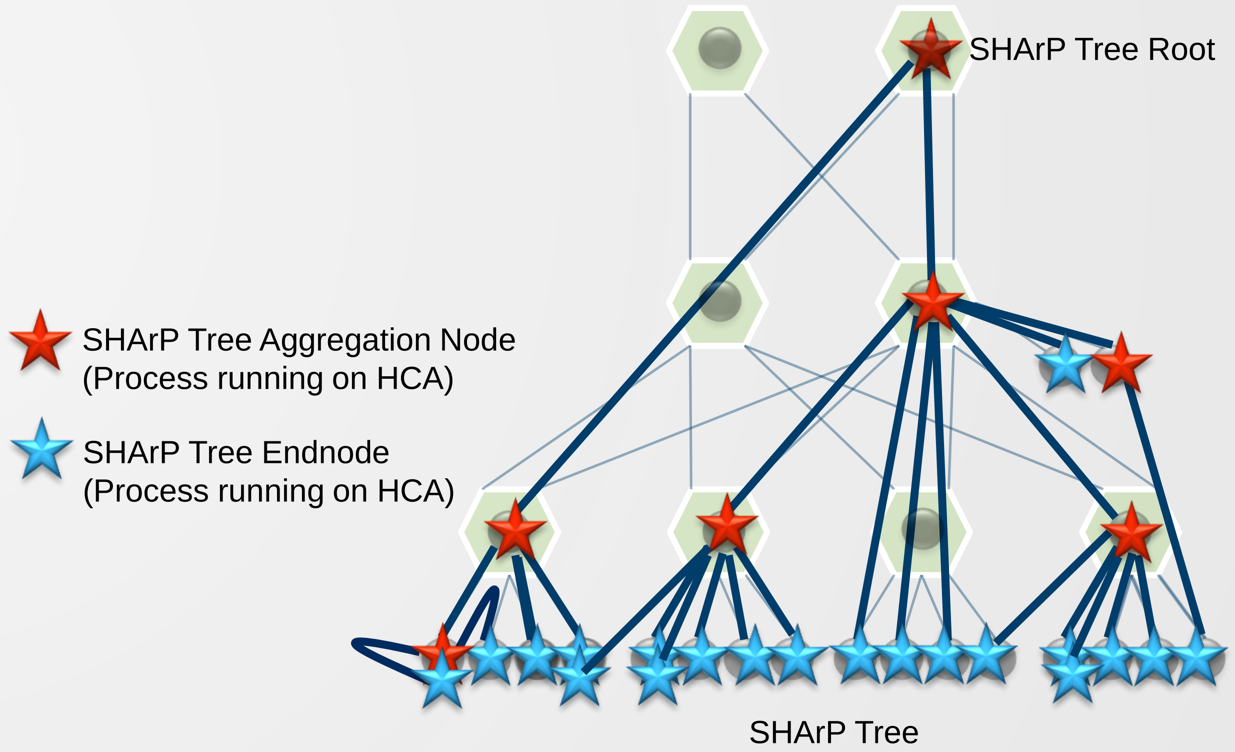
- In-network Tree based aggregation mechanism
- Large number of groups
- Multiple simultaneous outstanding operations
- Streaming aggregation

Accelerating HPC applications

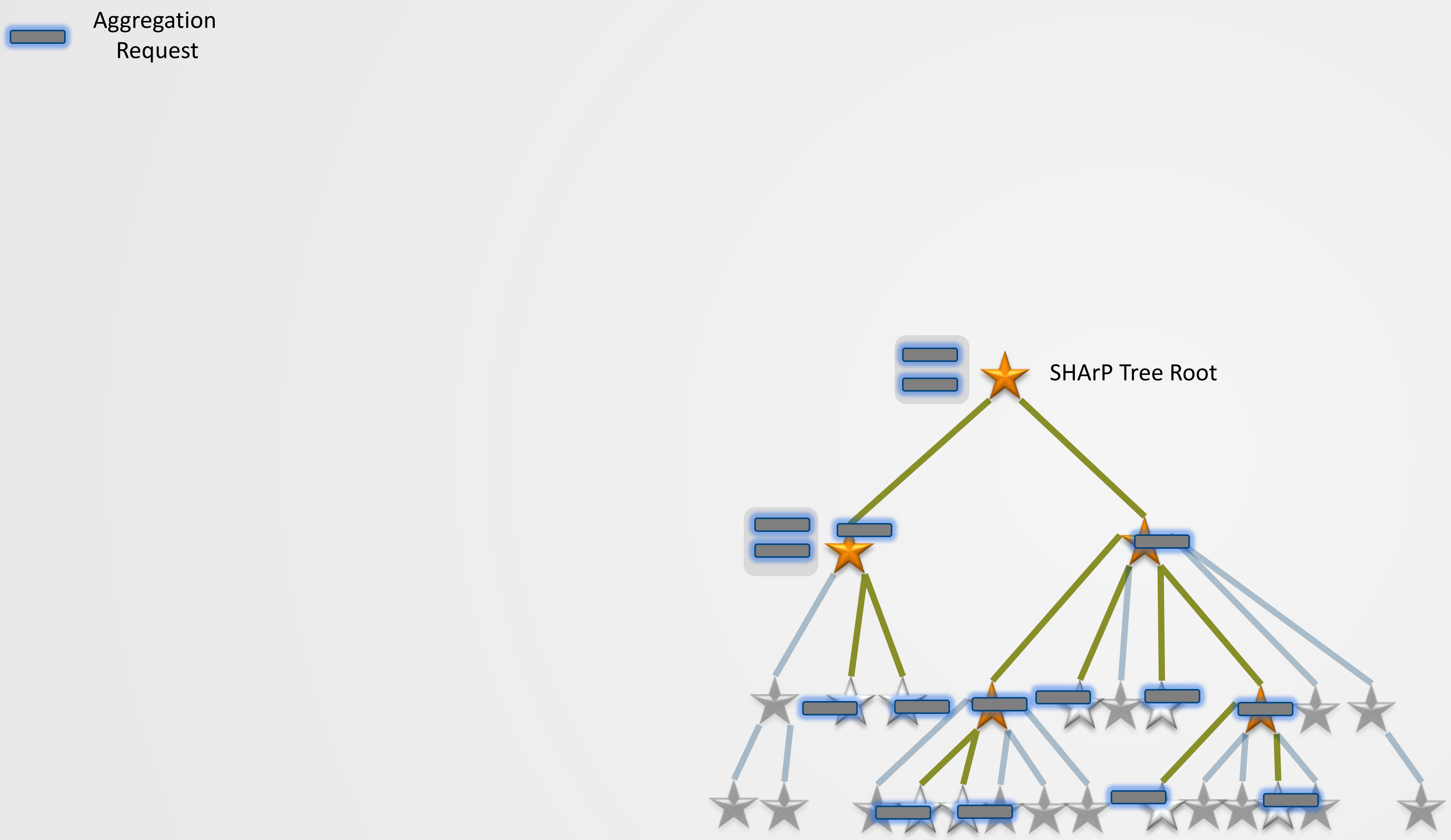
- Scalable High Performance Collective Offload
  - Barrier, Reduce, All-Reduce, Broadcast
  - Sum, Min, Max, Min-loc, Max-loc, OR, XOR, AND
  - Integer and Floating-Point, 16 / 32 / 64 bit
  - Up to 1KB payload size (in Quantum)
- Significantly reduce MPI collective runtime
- Increase CPU availability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

Accelerating Machine Learning applications

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

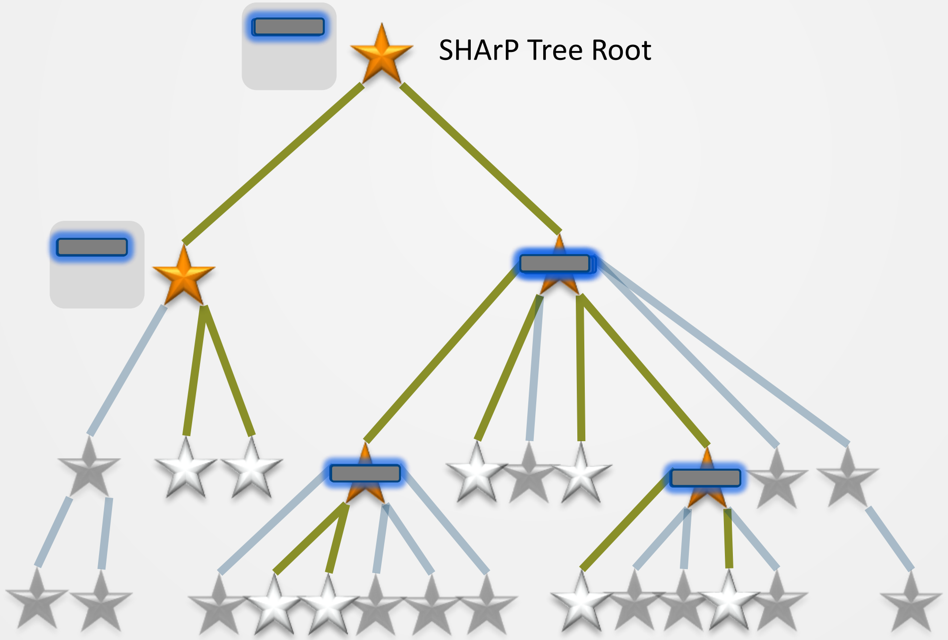


SHARP Principles of Operation - Request



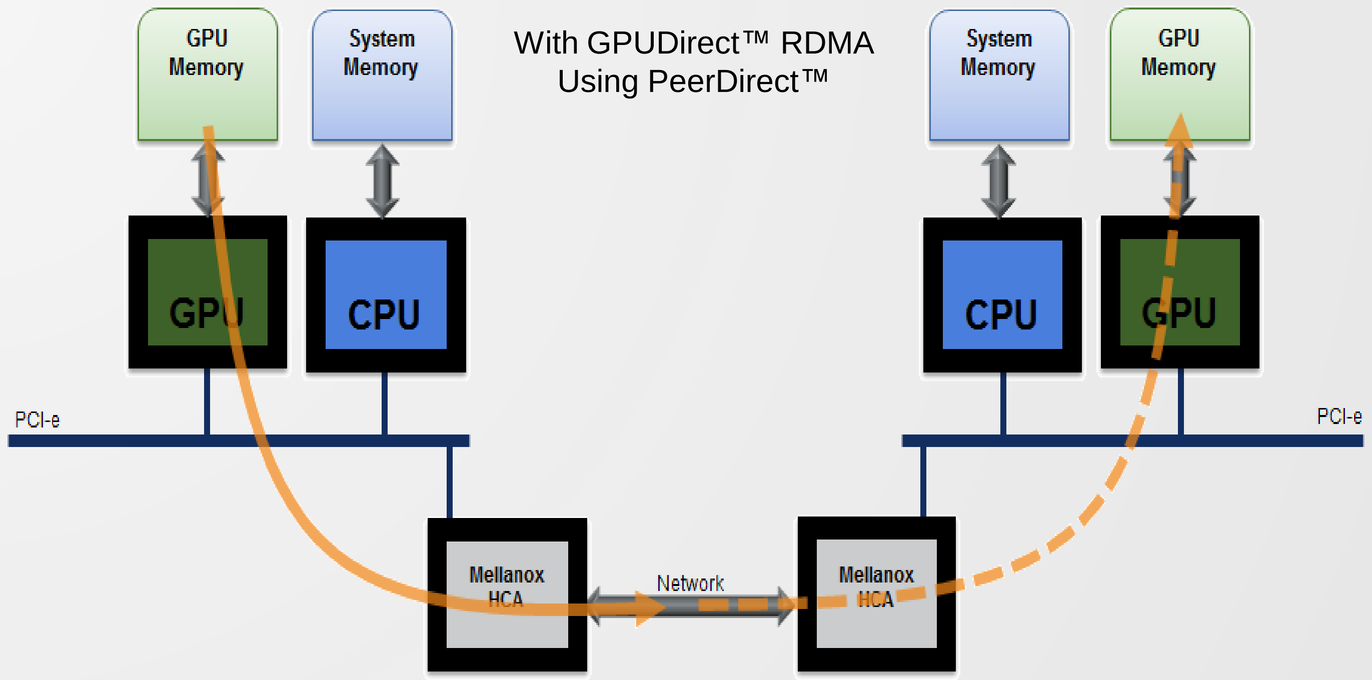
SHARP Principles of Operation – Response

Aggregation Response



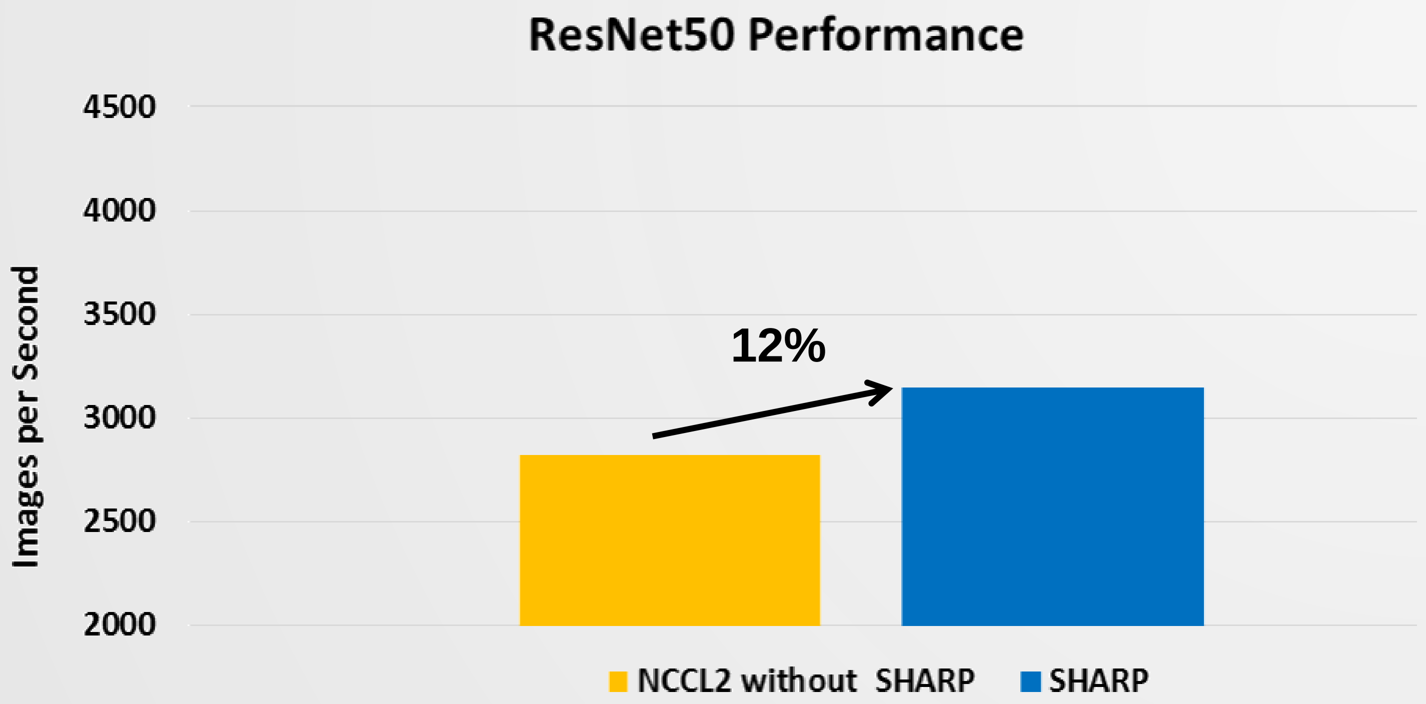
GPU Direct™ RDMA

- Network adapter can directly read data from GPU device memory
- Avoids copies through the host
- Eliminates CPU bandwidth and latency bottlenecks
- Uses remote direct memory access (RDMA) transfers between GPUs
- Resulting in significantly improved MPISendRecv efficiency between GPUs in remote nodes
- Fastest possible communication between GPU and other PCI-E devices
- Allows for better asynchronous communication

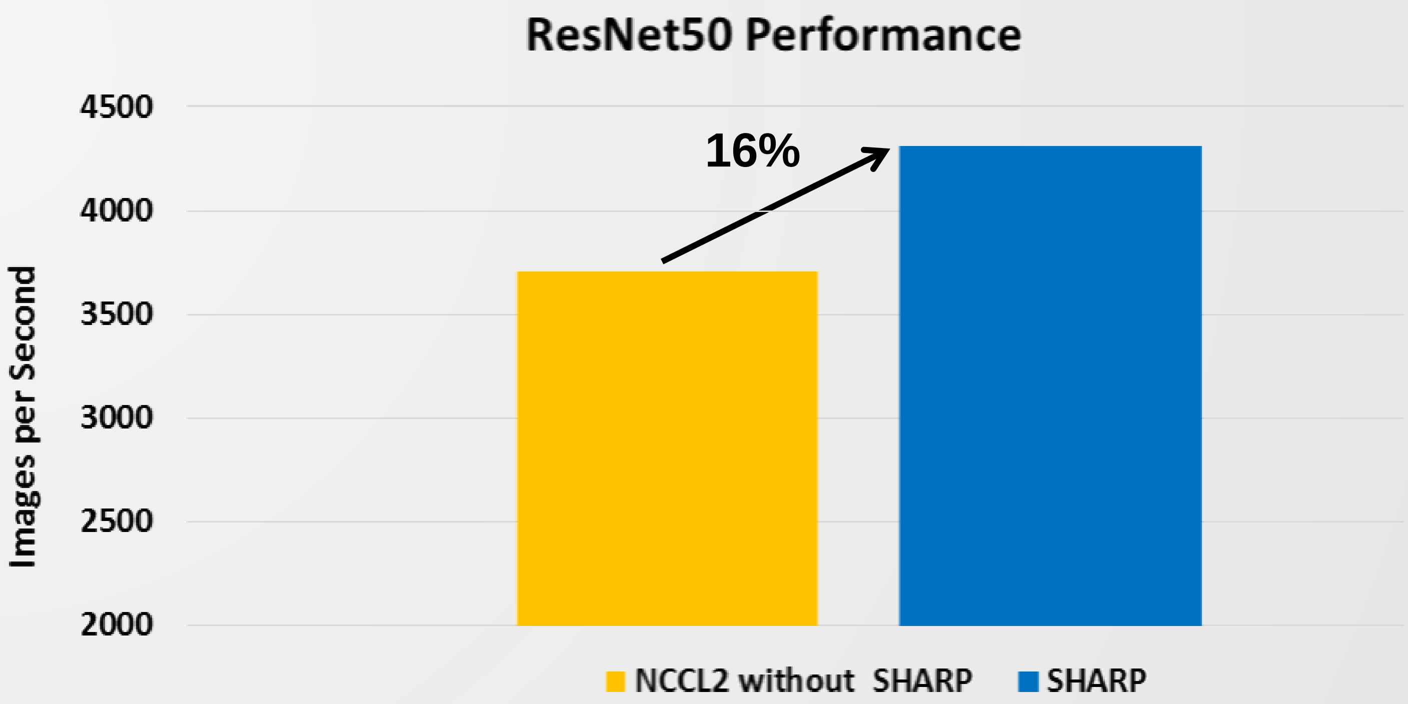


GPUDirect & SHARP Performance Advantage for AI

- TensorFlow Horovod running ResNet50 benchmark
- E5-2650V4, 12 cores @ 2.2GHz, 30M L2 cache, 9.6GT QPI, 256GB RAM: 16 x 16 GB DDR4
- P100 NVIDIA 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



8 Nodes, 16 GPUs, InfiniBand



8 Nodes, 22 GPUs, InfiniBand

SHARP SW Overview

Mellanox HPC-X™ Scalable HPC Software Toolkit

- Complete MPI, PGAS OpenSHMEM and UPC package
- Maximize application performance
- For commercial and open source applications
- Best out of the box experience

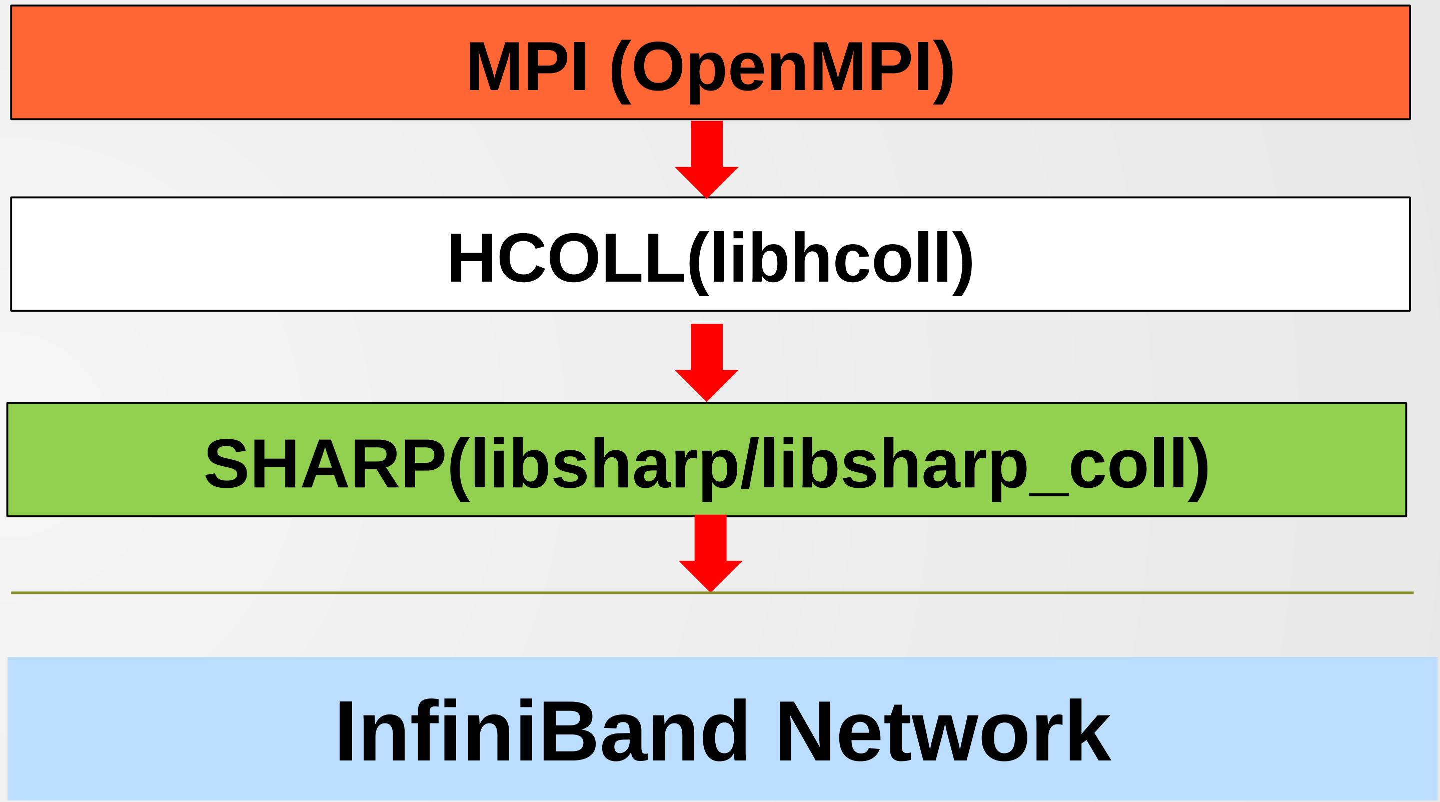
Mellanox HPC-X™ Scalable HPC Software Toolkit

- Allow fast and simple deployment of HPC libraries
  - Both Stable & Latest Beta are bundled
  - All libraries are pre-compiled
  - Includes scripts/module files to ease deployment

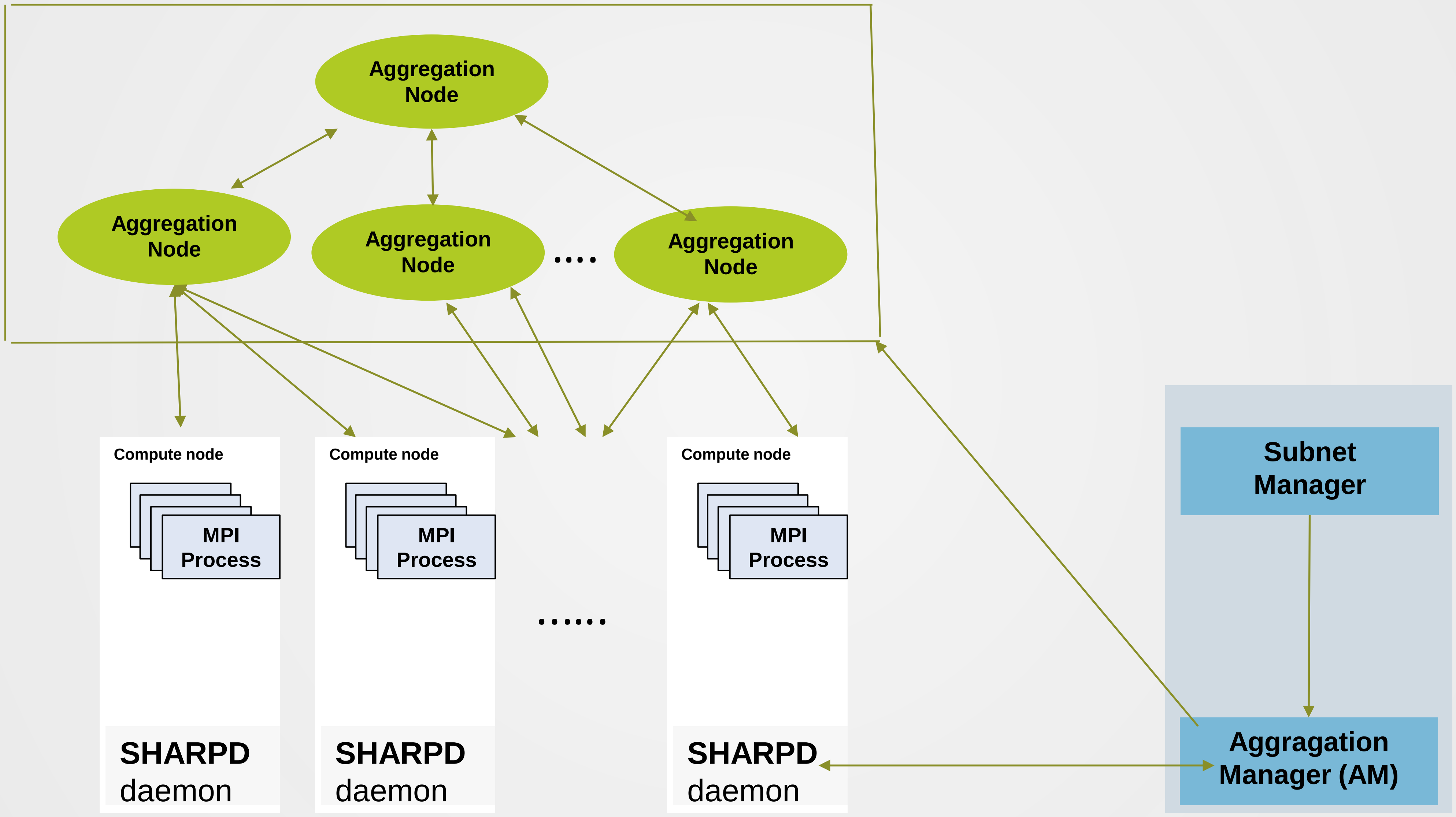
- Package Includes
  - OpenMPI / OpenSHMEM
  - BUPC (Berkeley UPC)
  - UCX
  - FCA/HCOLL
  - SHARP
  - KNEM
    - Allows fast intra-node MPI communication for large messages
  - Profiling Tools
    - Libibprof
    - IPM
  - Standard Benchmarks
    - OSU
    - IMB

HPCX/SHARP SW architecture

- HCOLL
  - optimized collective library
  - Easy to integrate with multiple MPIs(OpenMPI, MPICH, MVAPICH*)
- Libsharp.so
  - Implementation of low level sharp API
- Libsharp_coll.so
  - Implementation of high level sharp API for enabling sharp collectives for MPI
  - uses low level libsharp.so API
  - Easy to integrate with multiple MPIs(OpenMPI, MPICH, MVAPICH*)



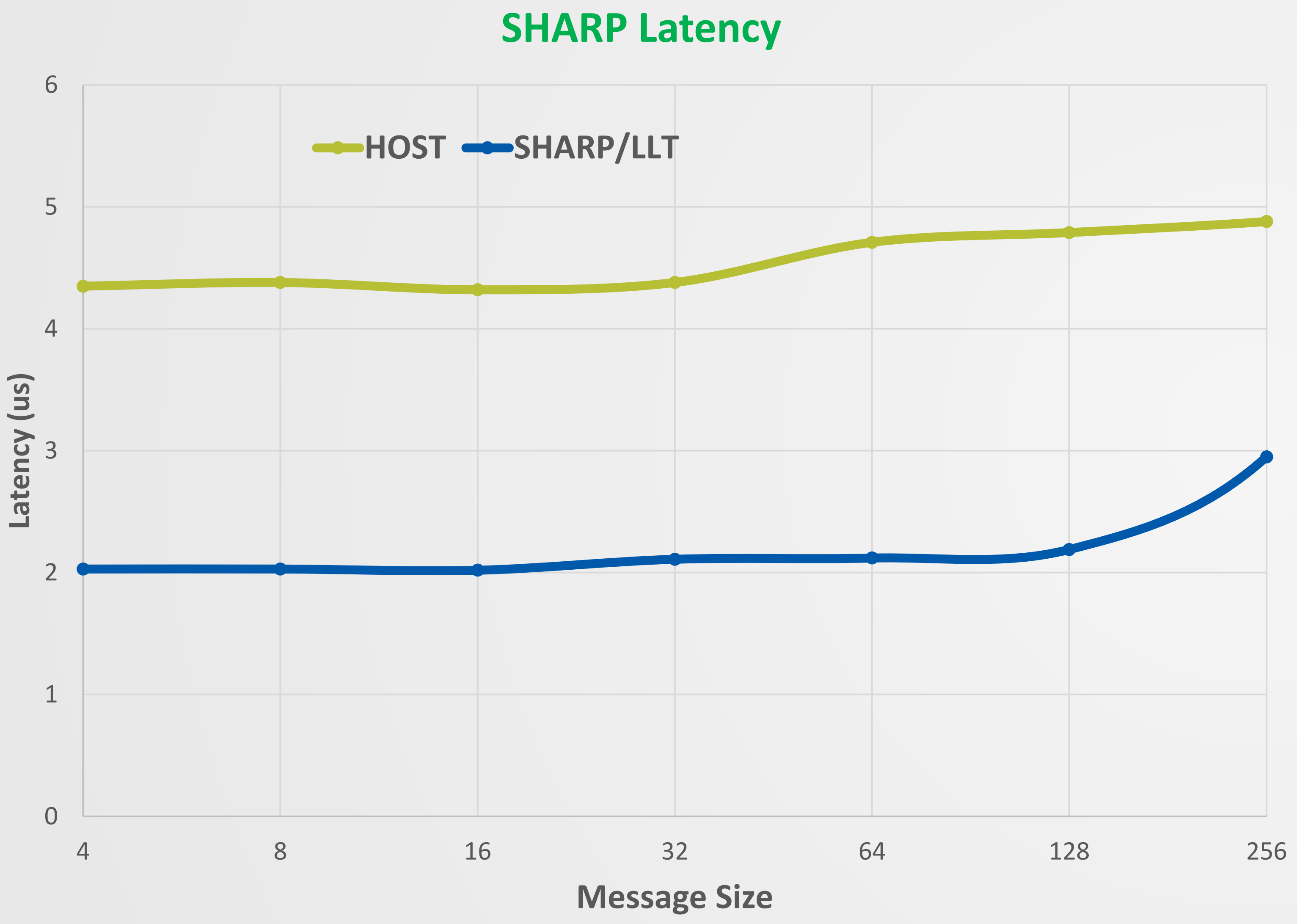
SHARP Software Architecture

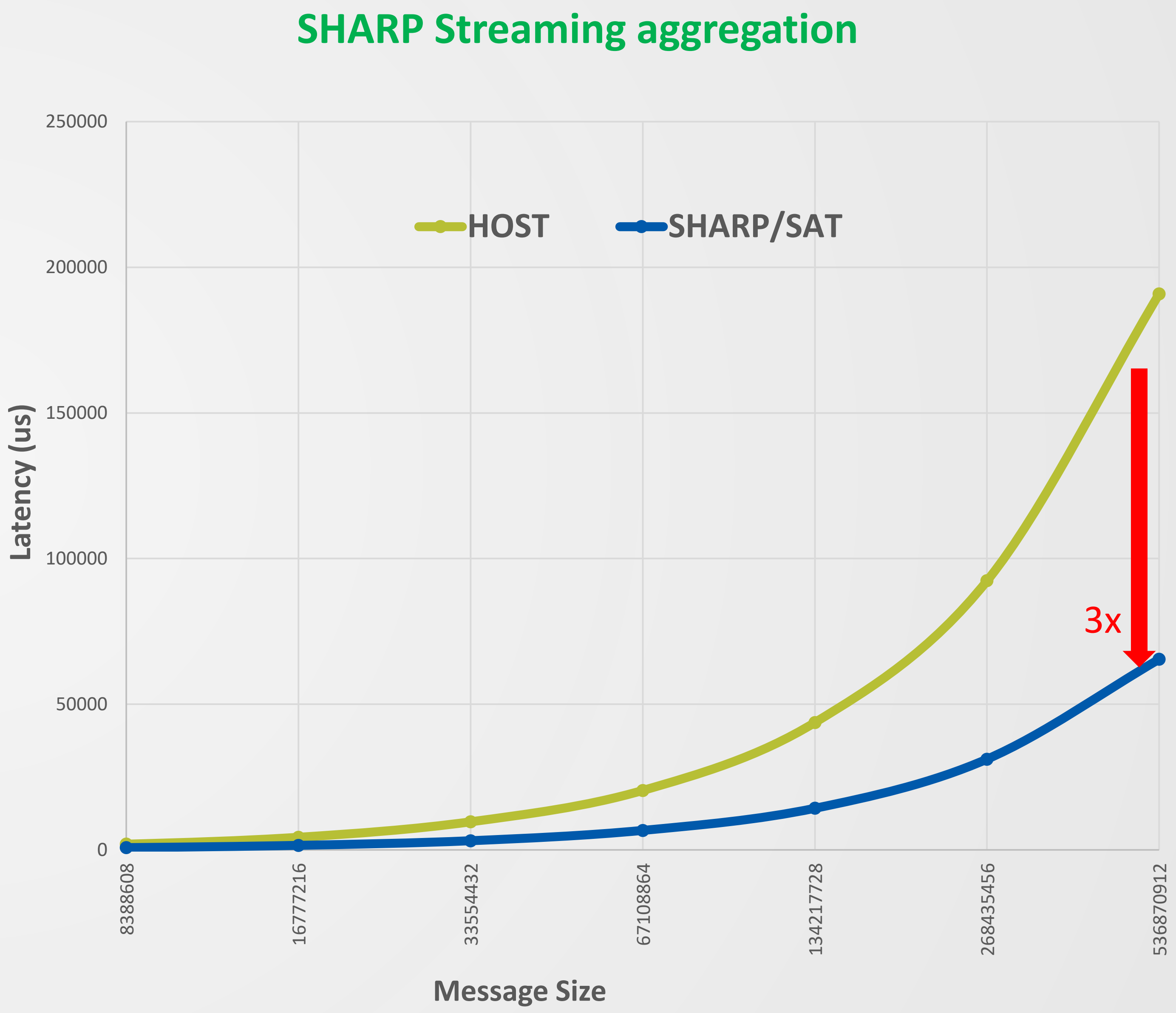


Setup

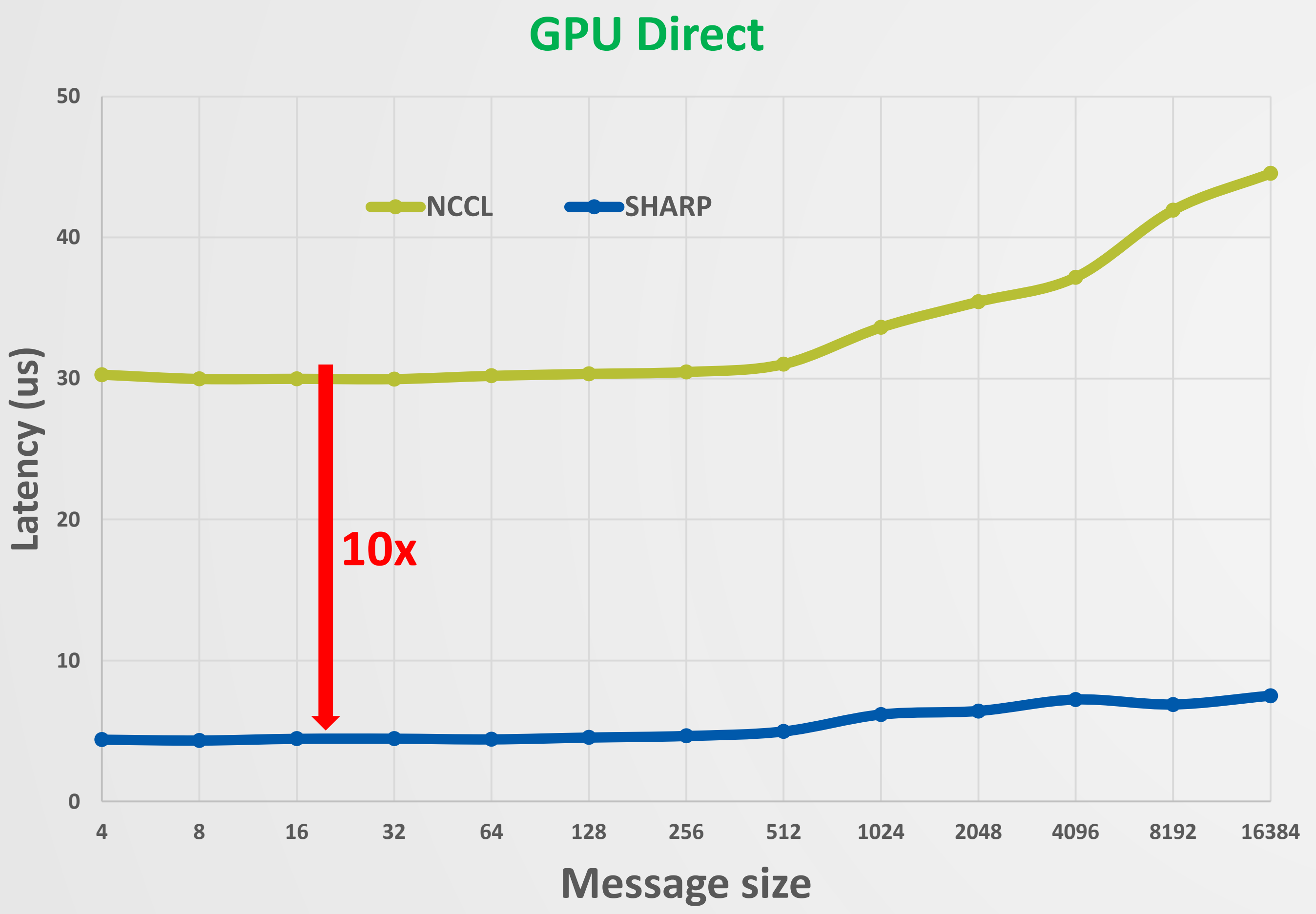
- 4 nodes, 16 GPUs
- TensorFlow/Horovod running ResNet50 benchmark
- Intel(R) Xeon(R) Gold 6150 CPU @ 2.70GHz
- Volta NVIDIA GPUs, ConnectX-6 HCA, IB Quantum Switch (EDR speed)
- Ubuntu-16.04, Mellanox OFED 4.5, HPC-X v2.3, TensorFlow v1.12, Horovod 0.15.2
- NCCL : 1 Ring, NVLink with in the Node.
- SHARP: Using 4 channels (4 ports) directly participating in SAT operation
- Topology:

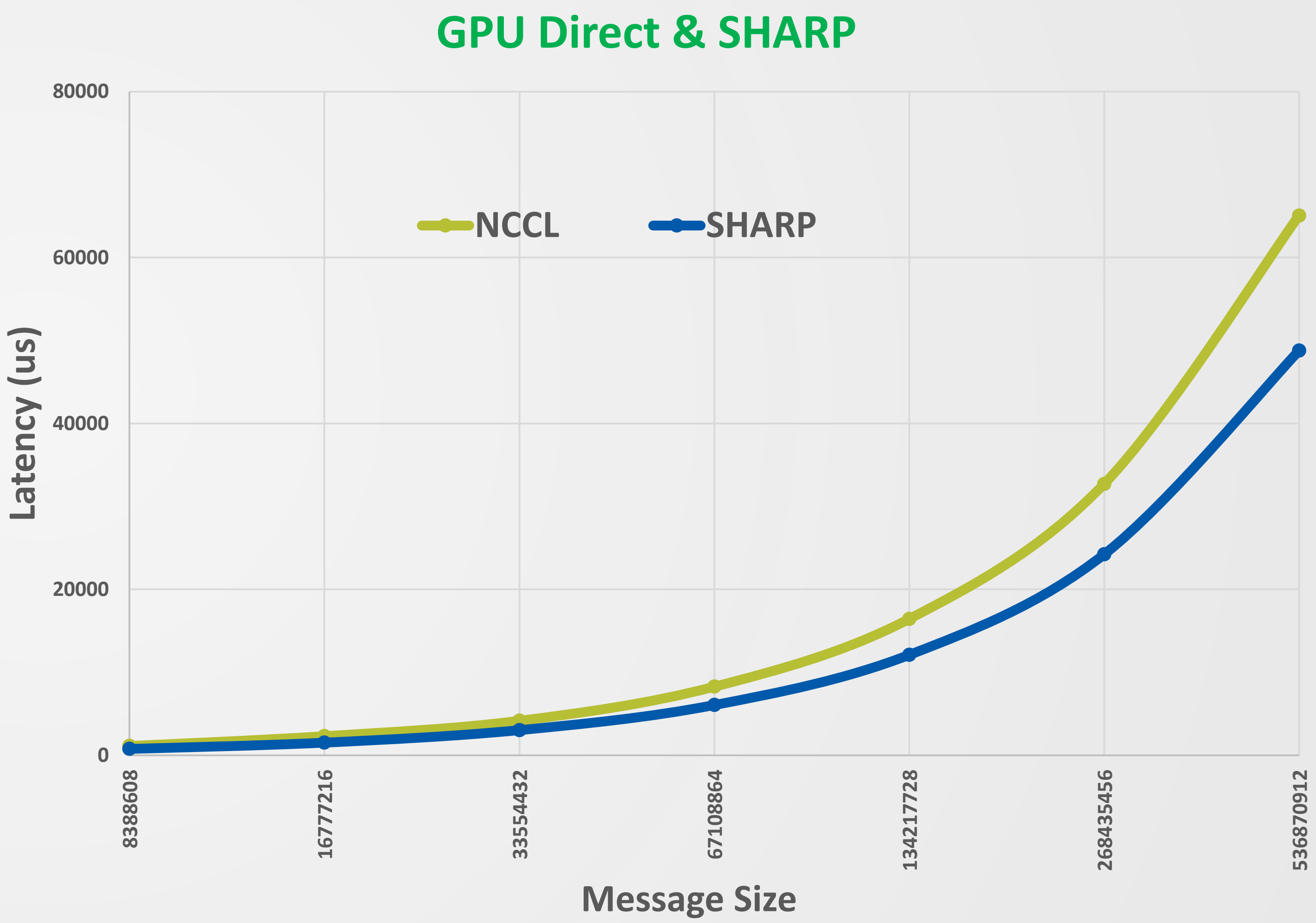
Allreduce - SHARP



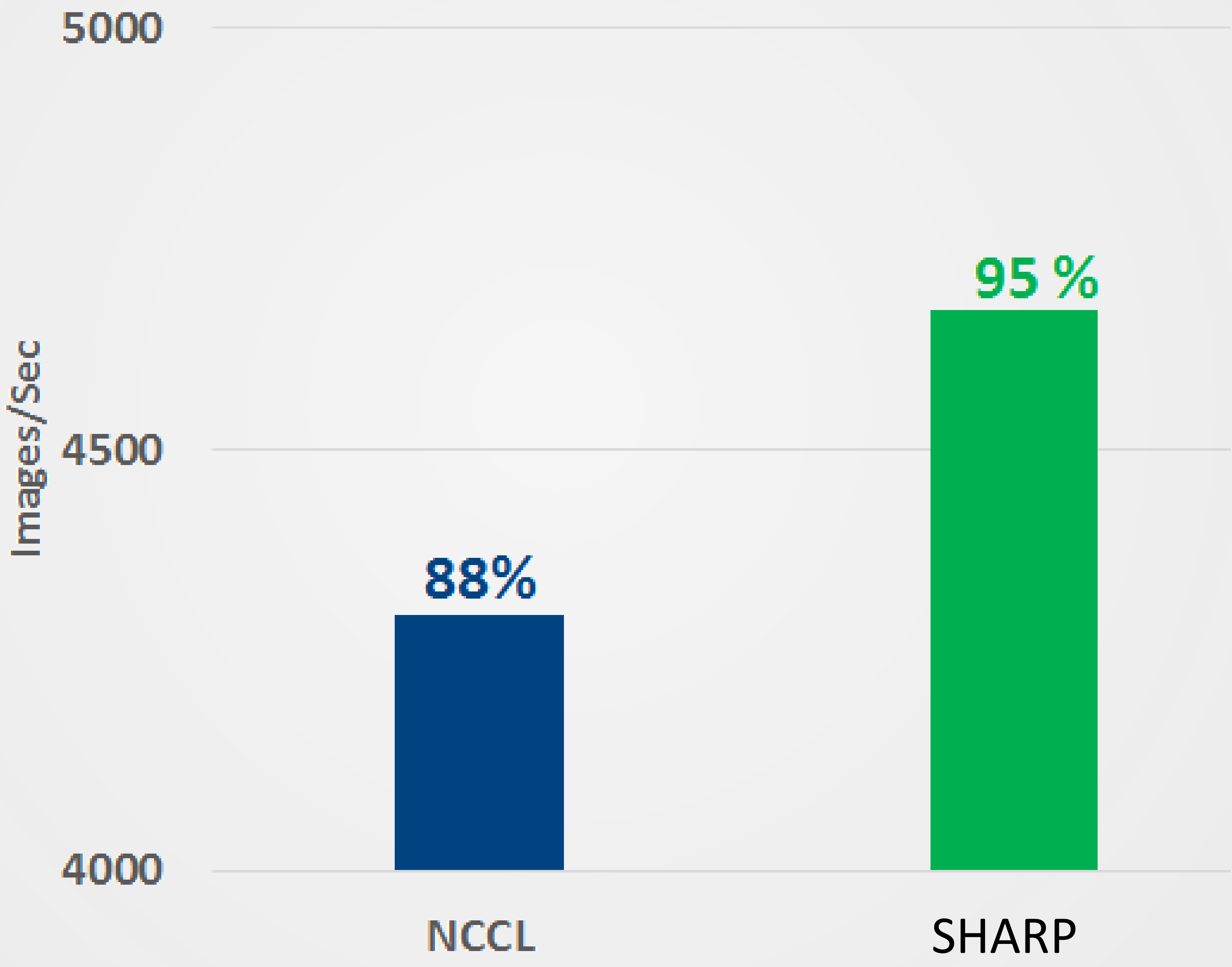


Allreduce - GPU Direct & SHARP

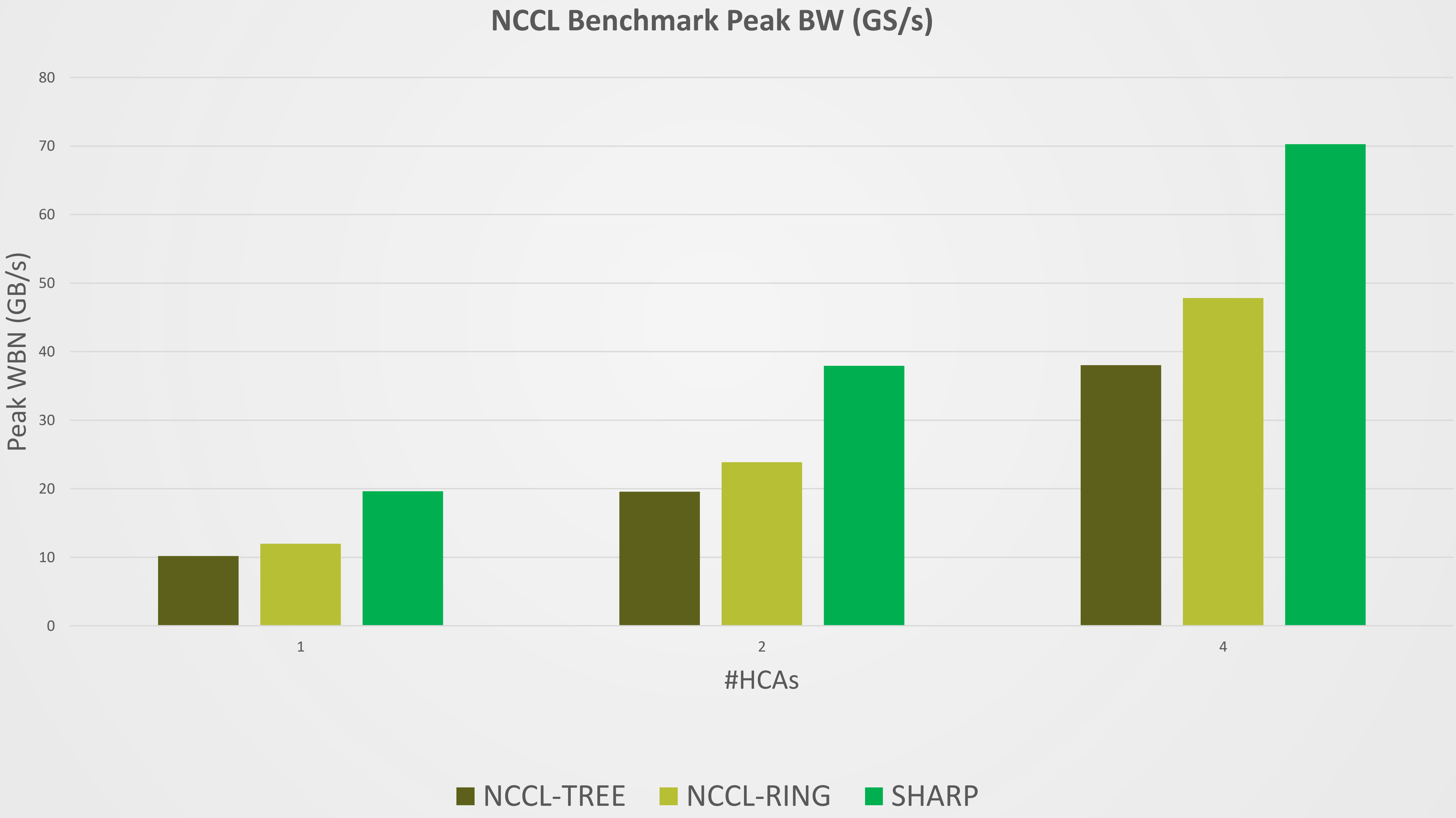




Horovod – Resnet50



NCCL benchmarks



Thank You