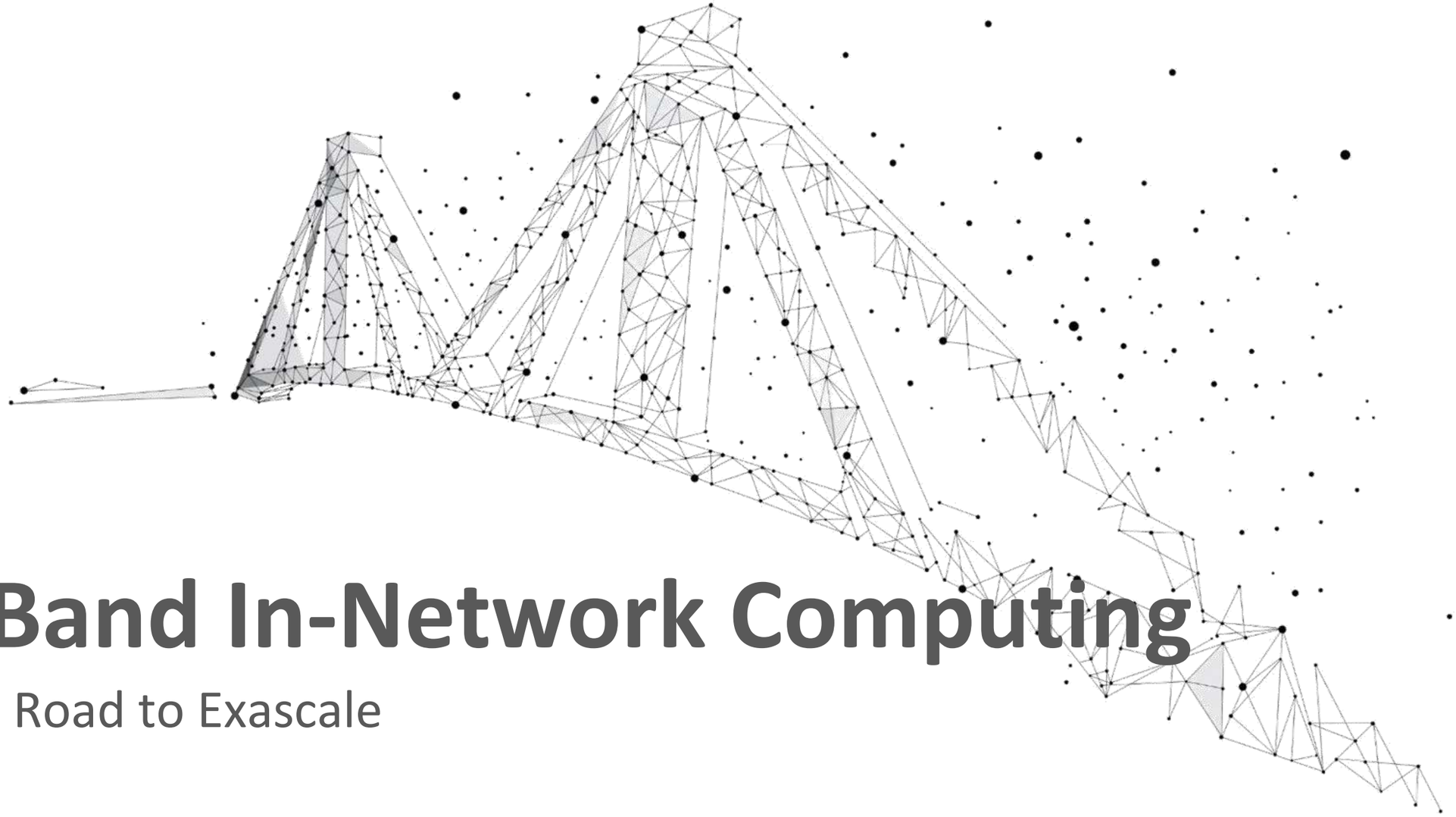


InfiniBand In-Network Computing

Paving the Road to Exascale

June 2019

Mellanox TECHNOLOGIES

SUPERCONNECTING the #1 Supercomputers

OAK RIDGE National Laboratory

1 TOP500 The List.

Lawrence Livermore National Laboratory

2 TOP500 The List.

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

3 TOP500 The List.

TACC TEXAS ADVANCED COMPUTING CENTER

5 TOP500 The List.

AIST ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

8 TOP500 The List.

Lawrence Livermore National Laboratory

10 TOP500 The List.

InfiniBand Accelerates 6 of Top 10 Supercomputers

Mellanox TECHNOLOGIES

SUPERCONNECTING the #1 Supercomputers

TACC
TEXAS ADVANCED COMPUTING CENTER

FRONTERA

5 TOP500 The List.

MISSISSIPPI STATE UNIVERSITY

NOAA

62 TOP500 The List.

CSC

FINNISH METEOROLOGICAL INSTITUTE

166 TOP500 The List.

筑波大学
University of Tsukuba

264 TOP500 The List.

UNIVERSITY OF MICHIGAN

World's First HDR InfiniBand Supercomputer

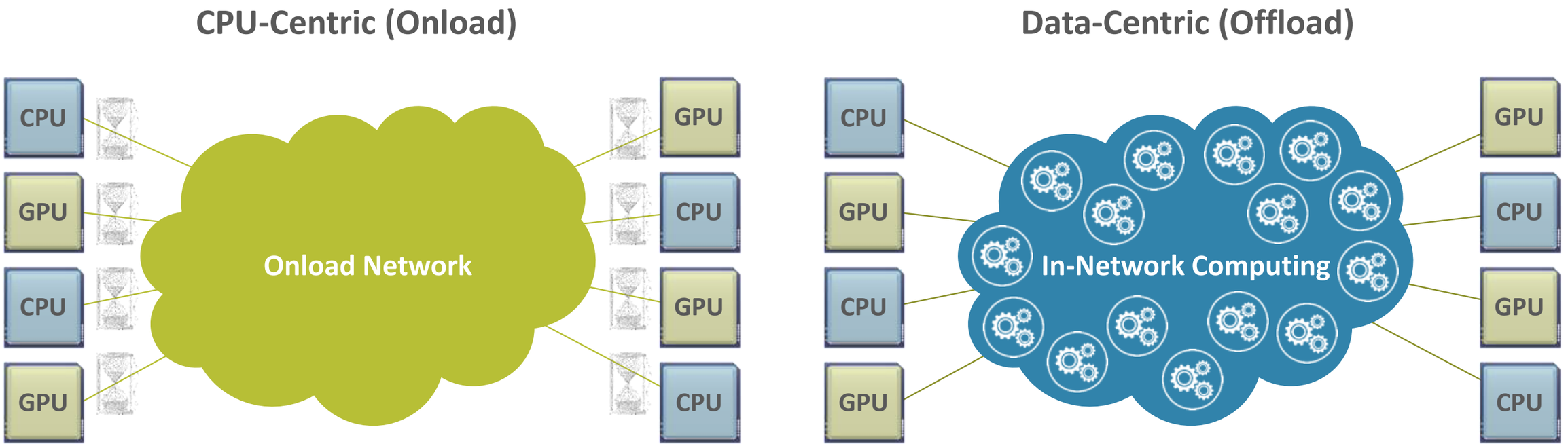
सी डैक
CDAC

India's National Supercomputing Program

HDR 200G InfiniBand Accelerated Supercomputers

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

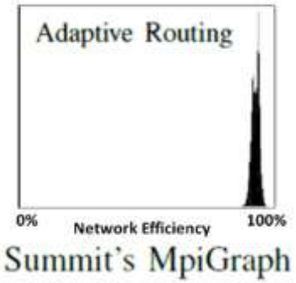
Analyze Data as it Moves!
Higher Performance and Scale

Highest Performance and Scalability for Exascale Platforms

OAK RIDGE National Laboratory

SHARP SHIELD SELF-HEALING

96%
Network Utilization



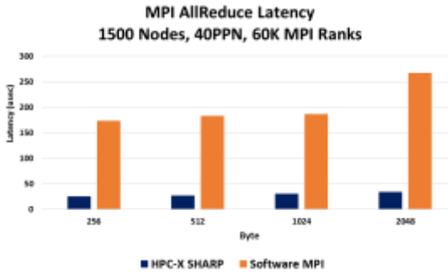
Summit's MpiGraph

UNIVERSITY OF TORONTO

SciNet

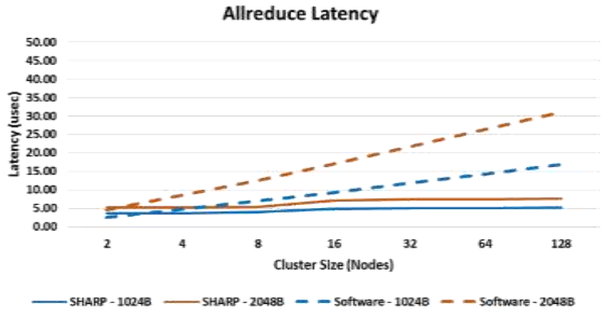
SHARP

7X
Higher Performance

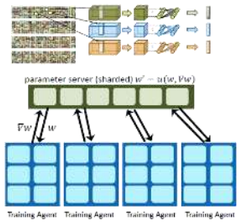


SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

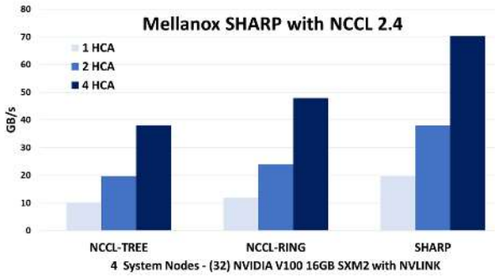
Flat Latency



Deep Learning

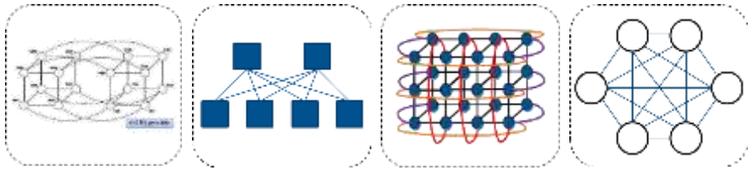


2X
Higher Performance



SHIELD SELF-HEALING

5000X
Higher Resiliency



XDR 1000G

NDR 400G

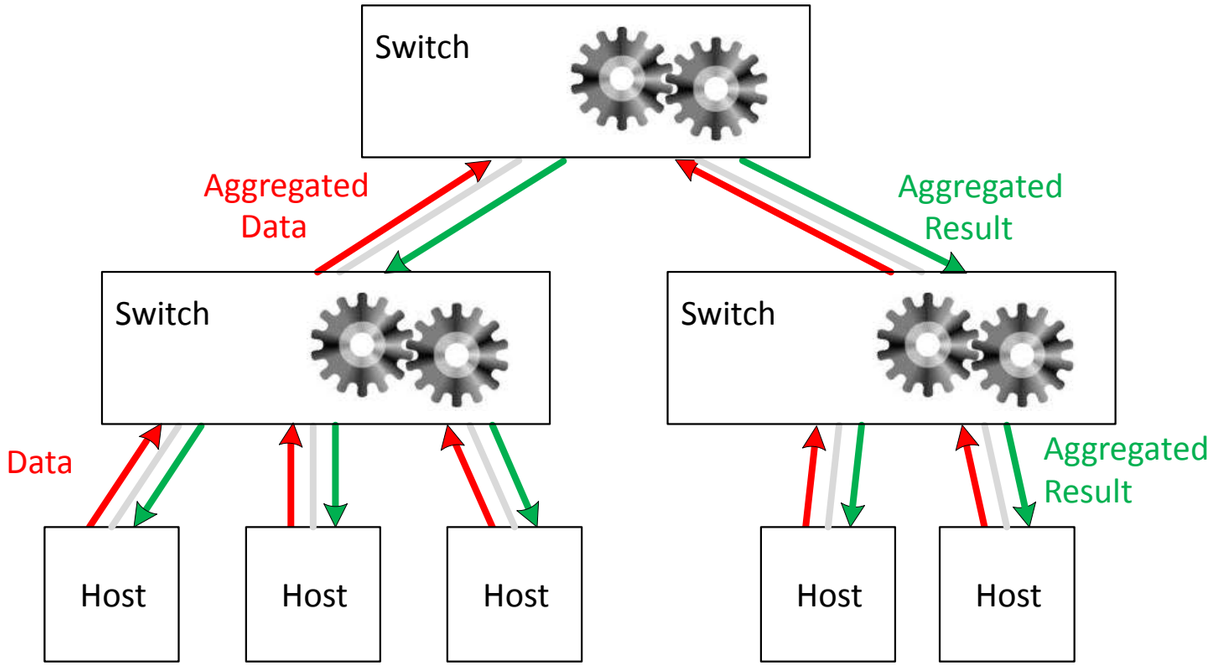
HDR 200G

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

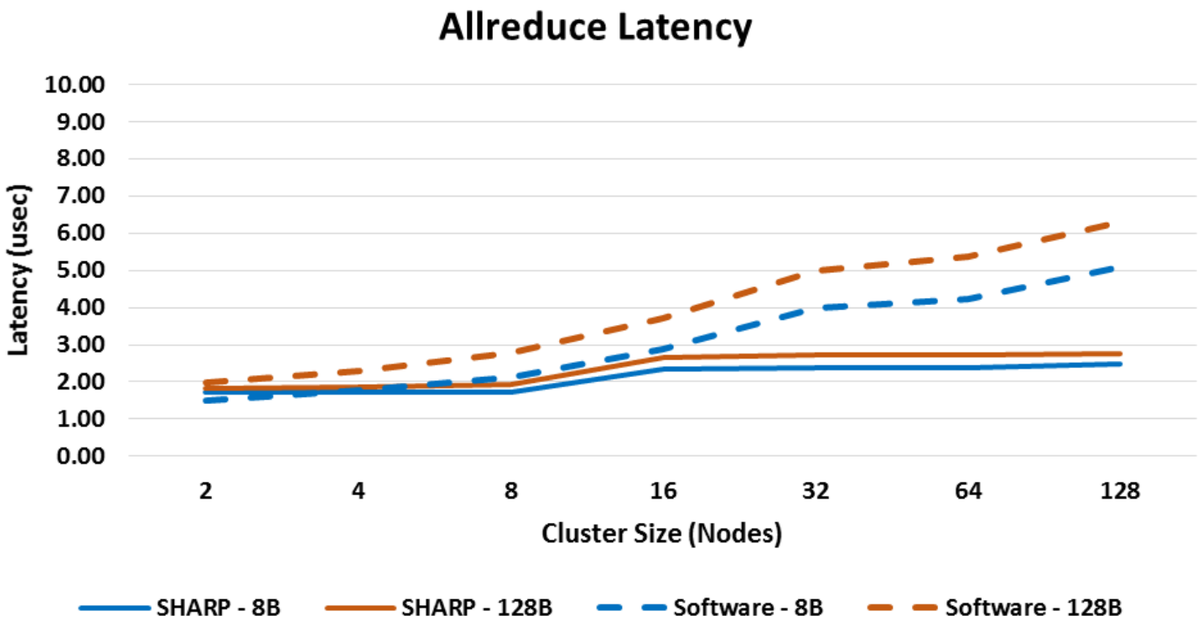
Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

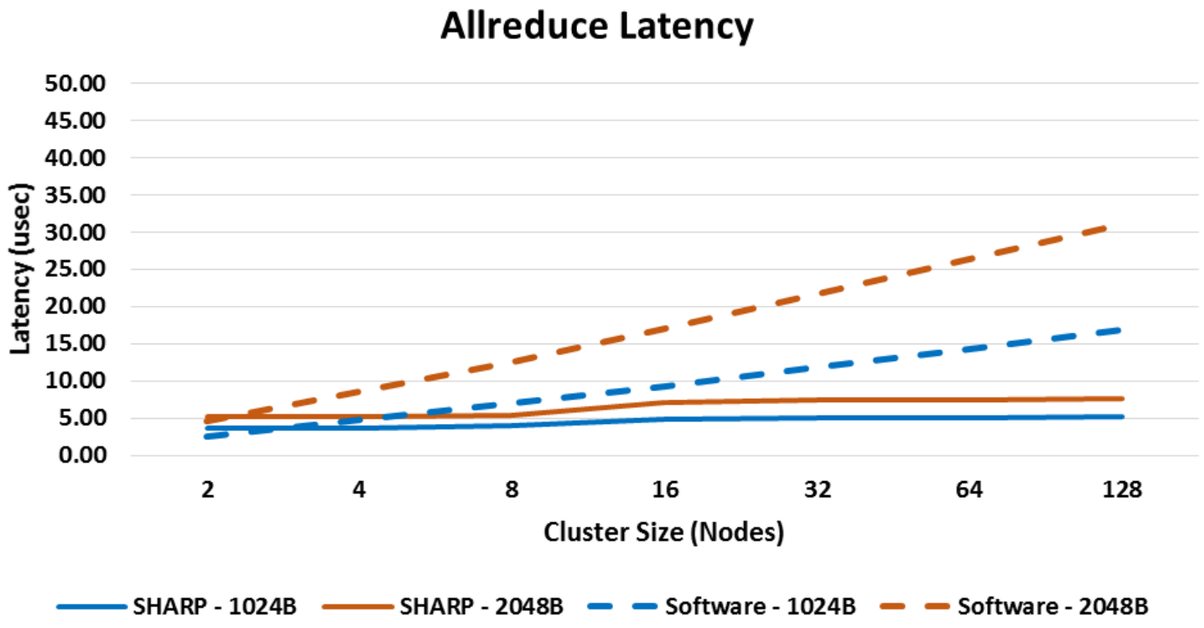
- Reliable Scalable General Purpose Primitive
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations
- Applicable to Multiple Use-cases
  - HPC Applications using MPI / SHMEM
  - Distributed Machine Learning applications
- Scalable High Performance Collective Offload
  - Barrier, Reduce, All-Reduce, Broadcast and more
  - Sum, Min, Max, Min-loc, max-loc, OR, XOR, AND
  - Integer and Floating-Point, 16/32/64 bits

SHARP
Scalable Hierarchical Aggregation and Reduction Protocol



SHARP AllReduce Performance Advantages (128 Nodes)



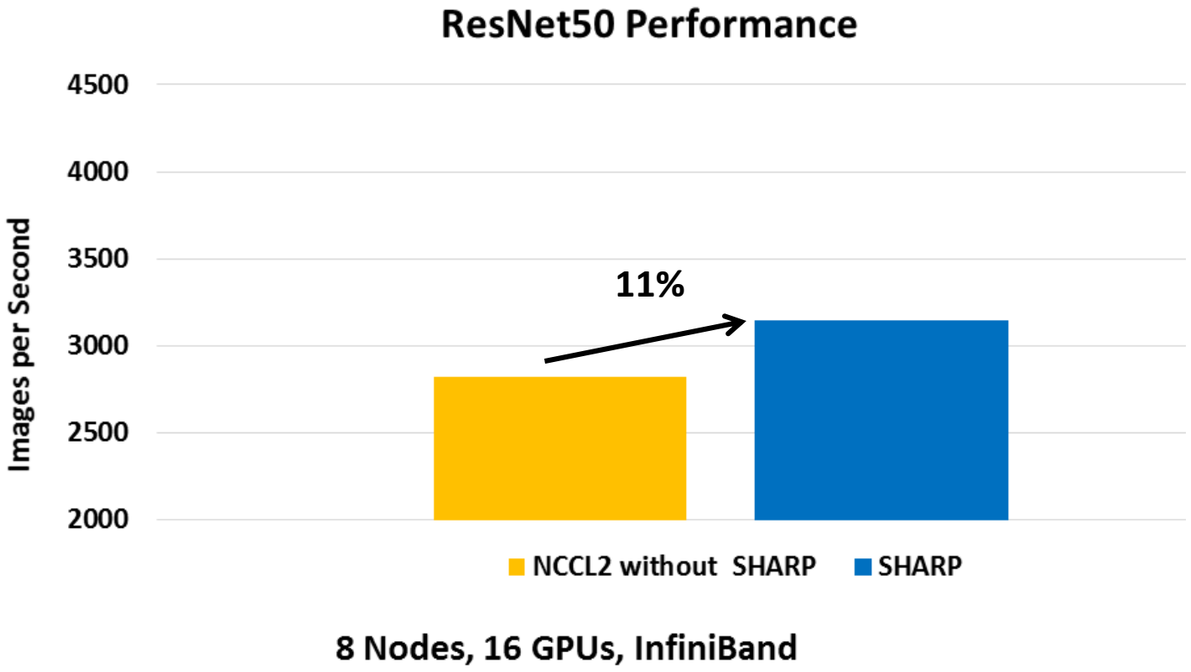


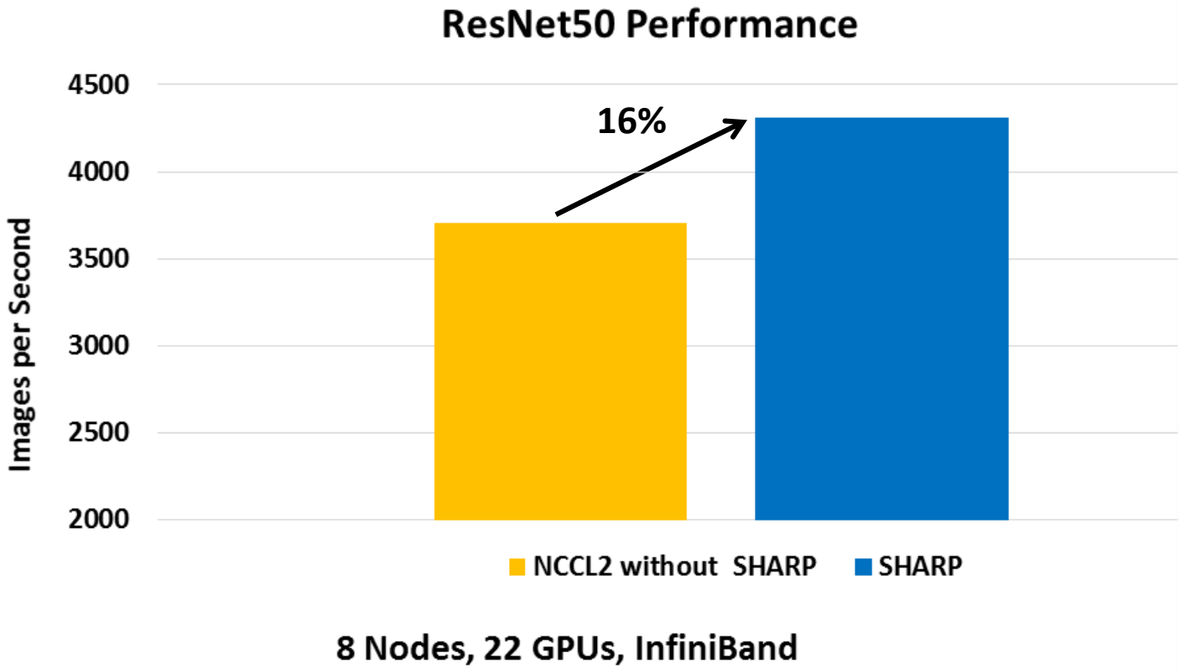
Scalable Hierarchical Aggregation and Reduction Protocol

SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

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)



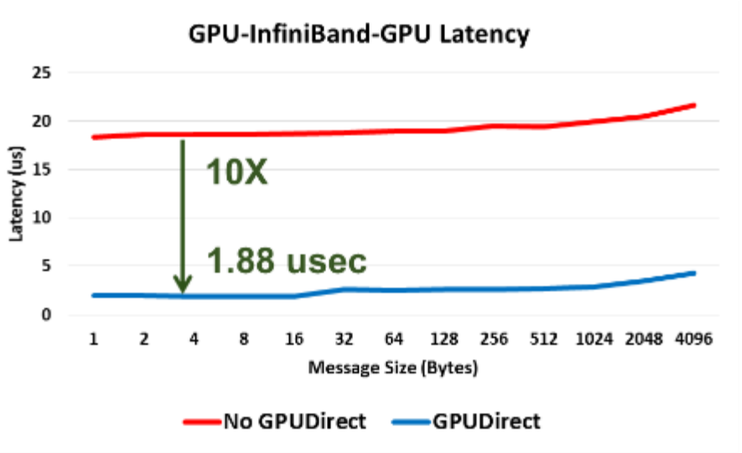


P100 NVIDIA GPUs, RH 7.5, Mellanox OFED 4.4, HPC-X v2.3, TensorFlow v1.11, Horovod 0.15.0

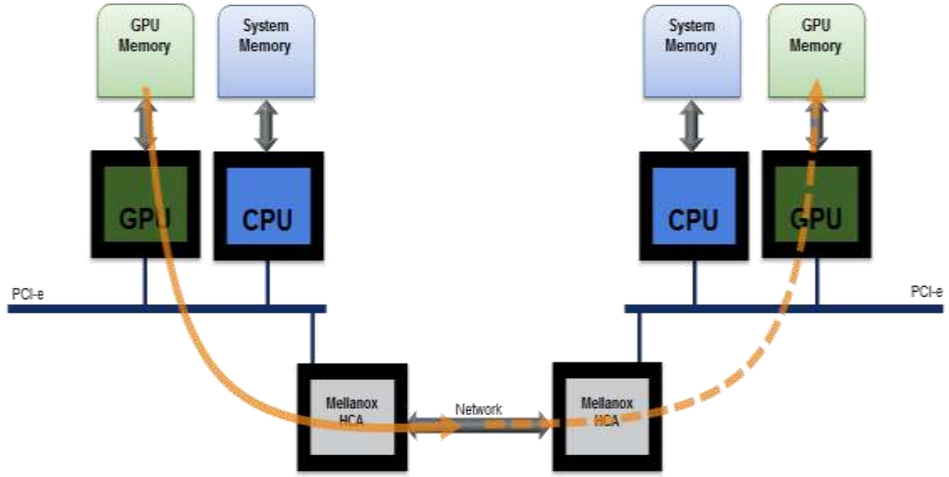
GPUDirect

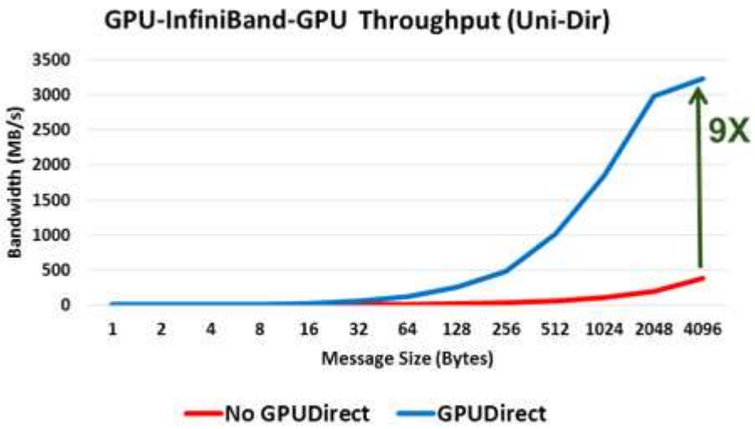
10X Higher Performance with GPUDirect™ RDMA

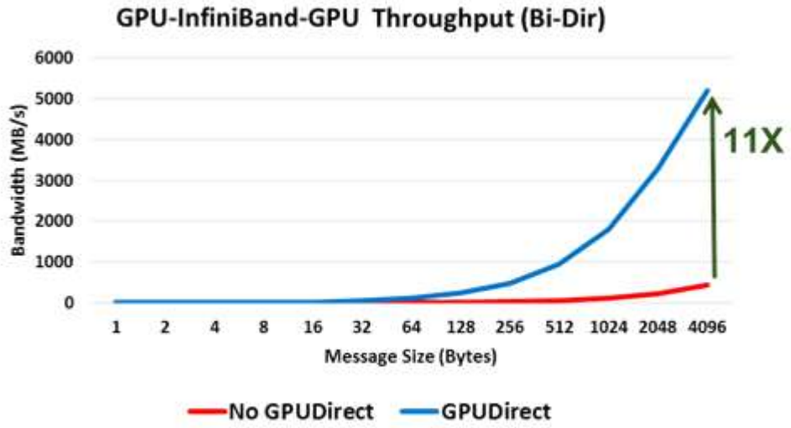
- Accelerates HPC and Deep Learning performance
- Lowest communication latency for GPUs



GPUDirect™ RDMA







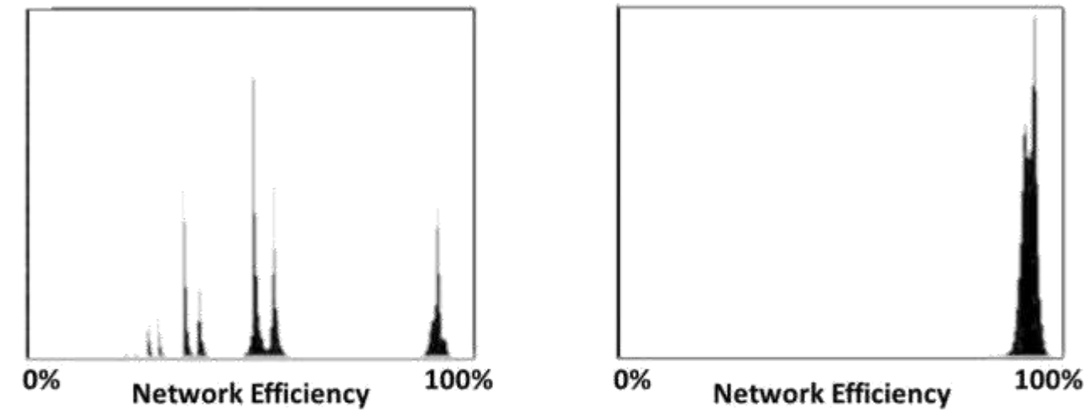
Courtesy of Dhabaleswar K. (DK) Panda
Ohio State University

Adaptive Routing

InfiniBand Proven Adaptive Routing Performance

- Oak Ridge National Laboratory – Coral Summit supercomputer
- Bisection bandwidth benchmark, based on mpiGraph
  - Explores the bandwidth between possible MPI process pairs
- AR results demonstrate an average performance of 96% of the maximum bandwidth measured

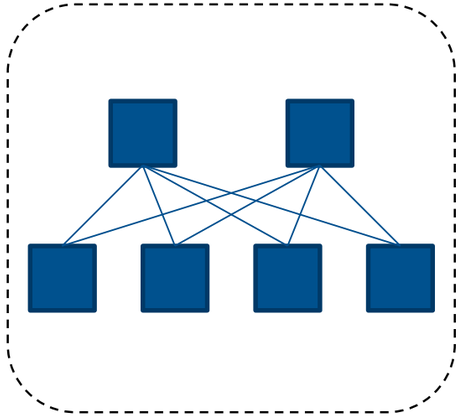
mpiGraph explores the bandwidth between possible MPI process pairs. In the histograms, the single cluster with AR indicates that all pairs achieve nearly maximum bandwidth while single-path static routing has nine clusters as congestion limits bandwidth, negatively impacting overall application performance.



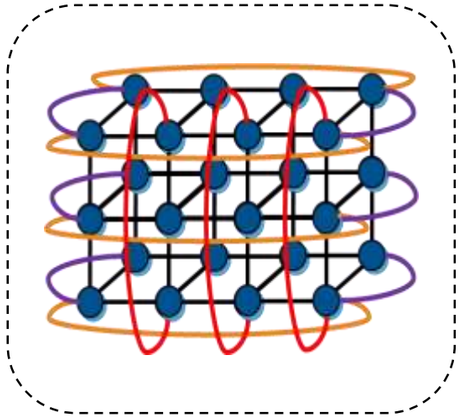
"The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems", Sudharshan S. Vazhkudai, Arthur S. Bland, Al Geist, Christopher J. Zimmer, Scott Atchley, Sarp Oral, Don E. Maxwell, Veronica G. Vergara Larrea, Wayne Joubert, Matthew A. Ezell, Dustin Leverman, James H. Rogers, Drew Schmidt, Mallikarjun Shankar, Feiyi Wang, Junqi Yin (Oak Ridge National Laboratory) and Bronis R. de Supinski, Adam Bertsch, Robin Goldstone, Chris Chambreau, Ben Casses, Elsa Gonsiorowski, Ian Karlin, Matthew L. Leininger, Adam Moody, Martin Ohmacht, Ramesh Pankajakshan, Fernando Pizzano, Py Watson, Lance D. Weems (Lawrence Livermore National Laboratory) and James Sexton, Jim Kahle, David Appelhans, Robert Blackmore, George Chochia, Gene Davison, Tom Gooding, Leopold Grinberg, Bill Hanson, Bill Hartner, Chris Marroquin, Bryan Rosenburg, Bob Walkup (IBM)

Network Topologies

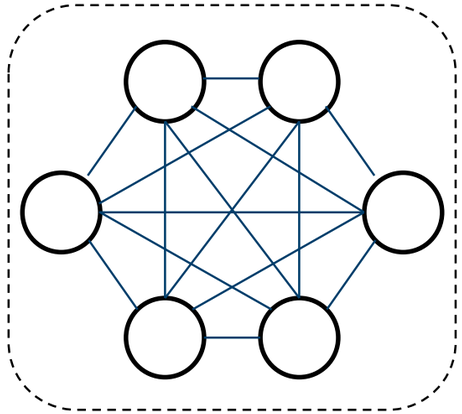
Supporting Variety of Topologies



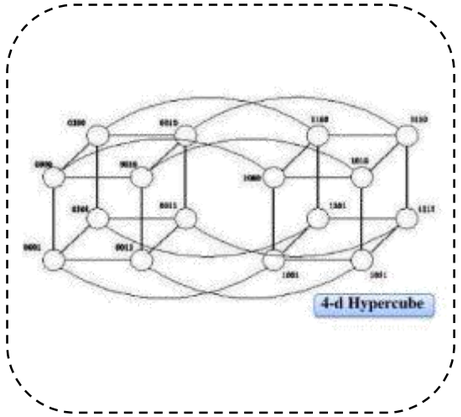
Fat Tree



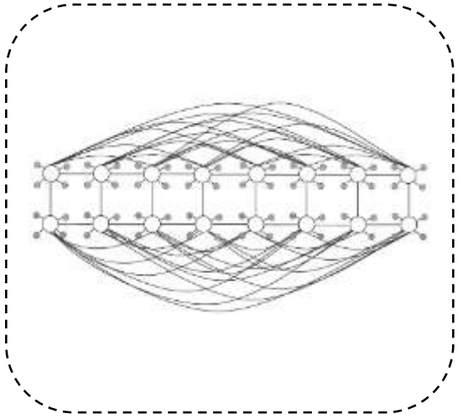
Torus



Dragonfly

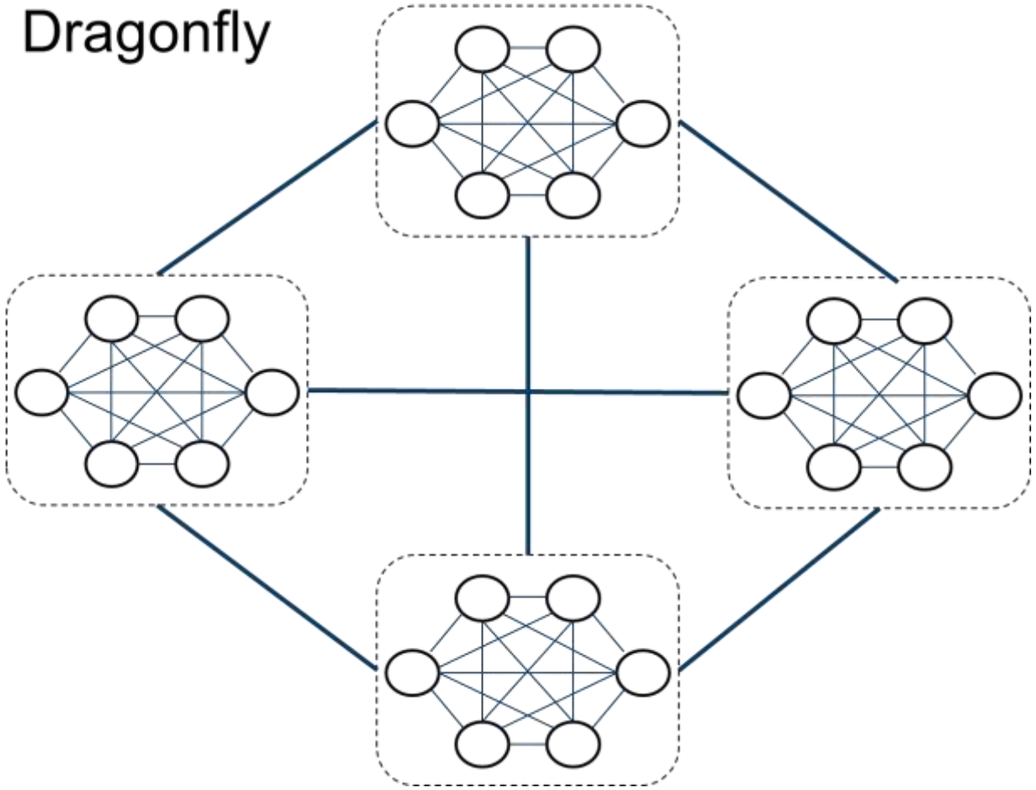


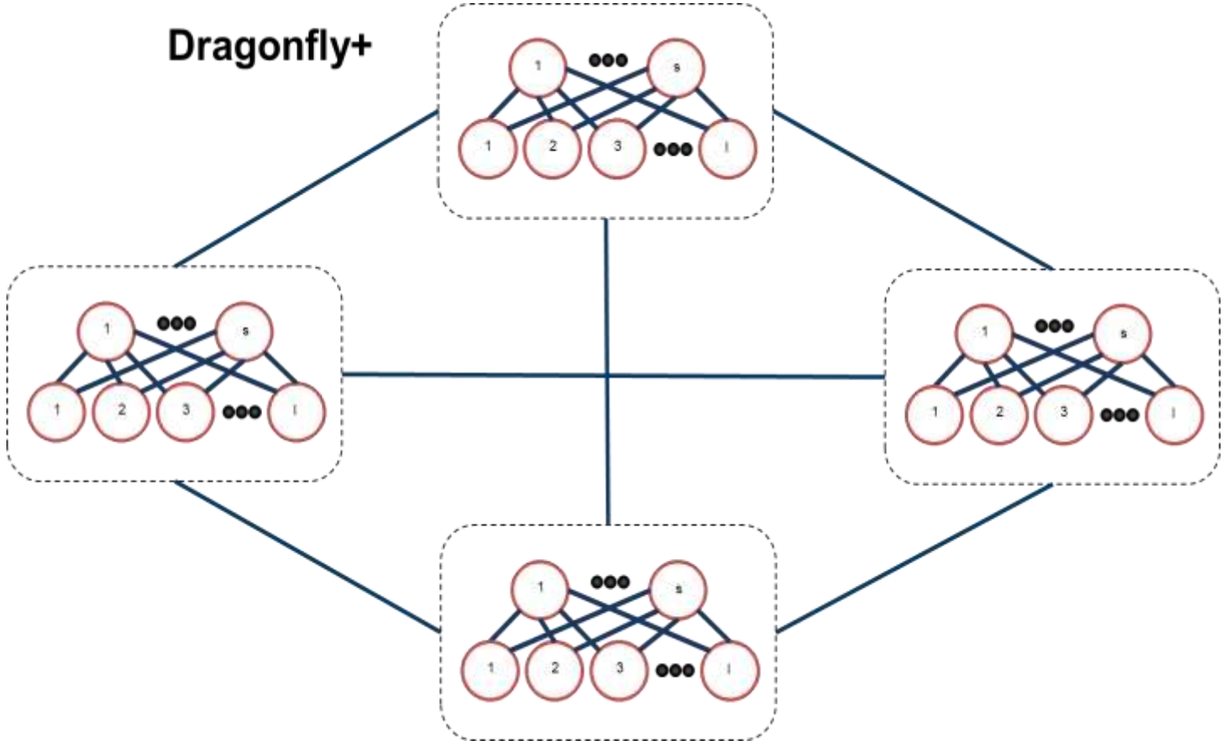
Hypercube



HyperX

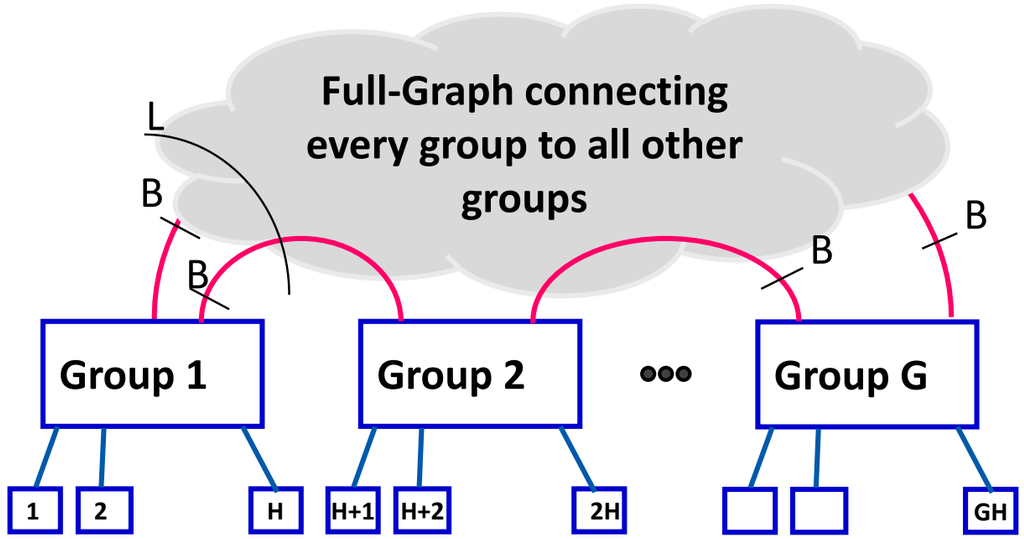
Dragonfly+ vs Traditional Dragonfly



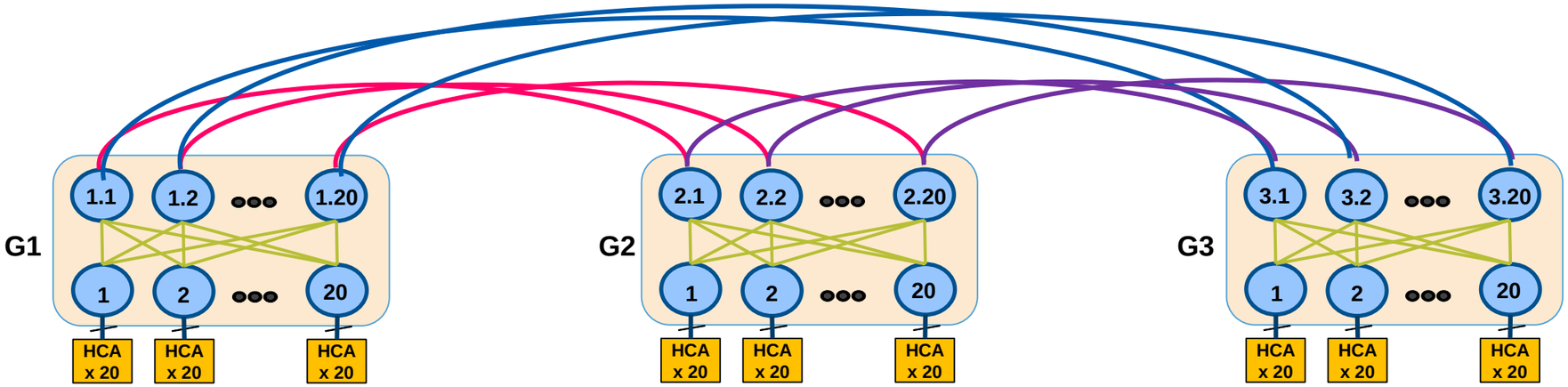


Dragonfly+ Topology

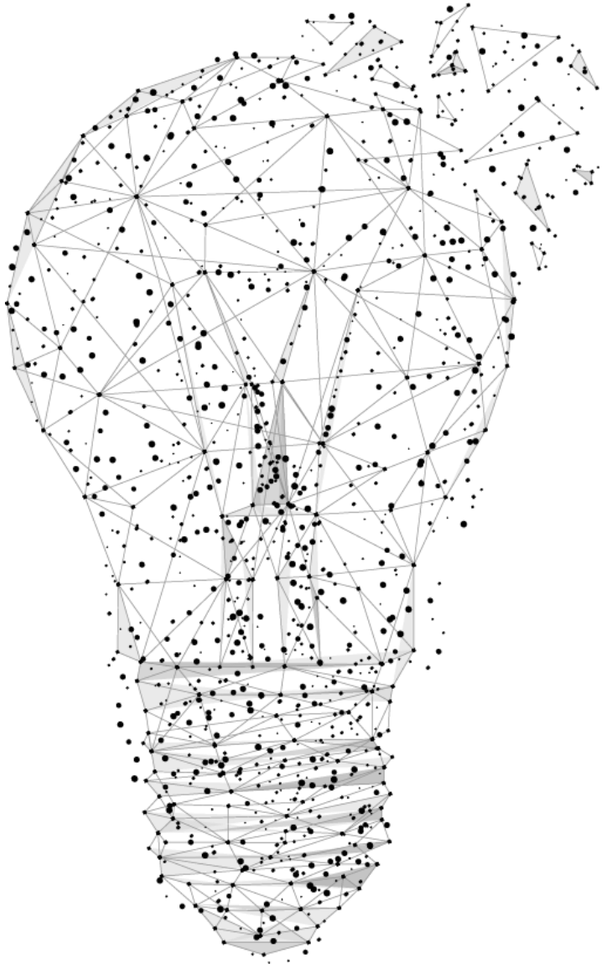
- Several “groups”, connected using all to all links
- The topology inside each group can be any topology
- Reduce total cost of network (fewer long cables)
- Utilizes Adaptive Routing to for efficient operations
- Simplifies future system expansion



1200-Nodes Dragonfly+ Systems Example



BlueField SoC Programmable Network



BlueField for Smart Solutions

BlueField SoC (System on Chip)

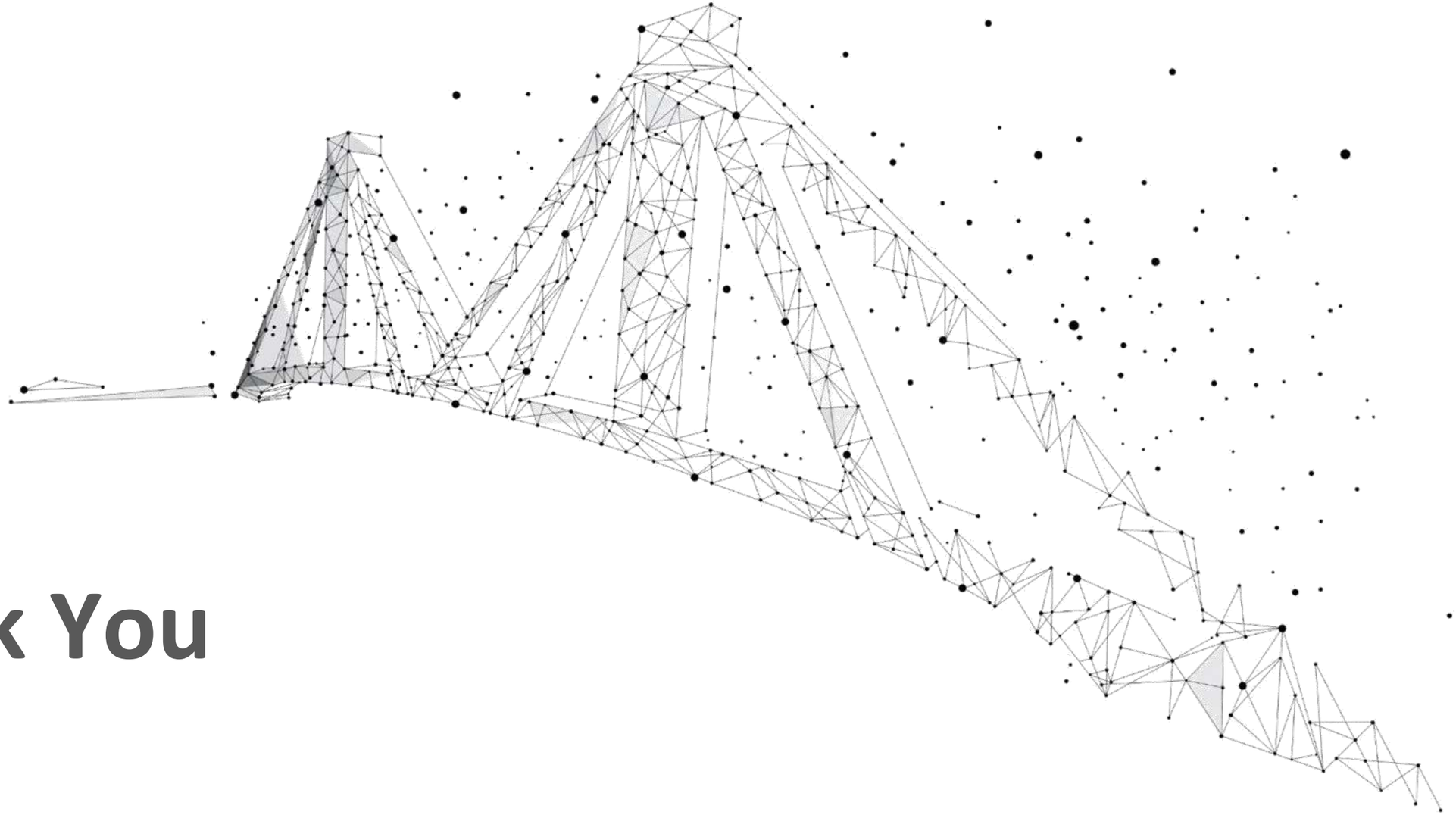
- SoC: Compute, networking and PCIe connectivity
  - Dual port VPI EDR/100GbE
  - 16 Arm cores
  - 32 lanes of PCIe switch gen3/4

Storage Solutions

- NVMe-based storage platforms
  - RDMA, NVMe over Fabrics, RAID, Signature offload
- Partner's solutions based on BlueField storage controller

Smart Adapters

- In-network computing and collective offloads
- Co-processor running proprietary smart algorithms
- Security and privacy algorithms



Thank You