Interconnect Your Future

Paving the Road to Exascale

November 2018

Highest-Performance 200Gb/s Interconnect Solutions

Adapters — ConnectX-6

200Gb/s Adapter, 0.6us latency
215 million messages per second
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)

Switch — Quantum (The smartest switch, became smarter)

40 HDR (200Gb/s) InfiniBand Ports
80 HDR100 InfiniBand Ports
Throughput of 16Tb/s, <90ns Latency

Switch — Spectrum-2

16 400GbE, 32 200GbE, 128 25/50GbE Ports
(10 / 25 / 40 / 50 / 100 / 200 GbE)
Throughput of 6.4Tb/s

SoC — BlueField

System on Chip and SmartNIC
Programmable adapter
Smart Offloads

Interconnect — LinkX

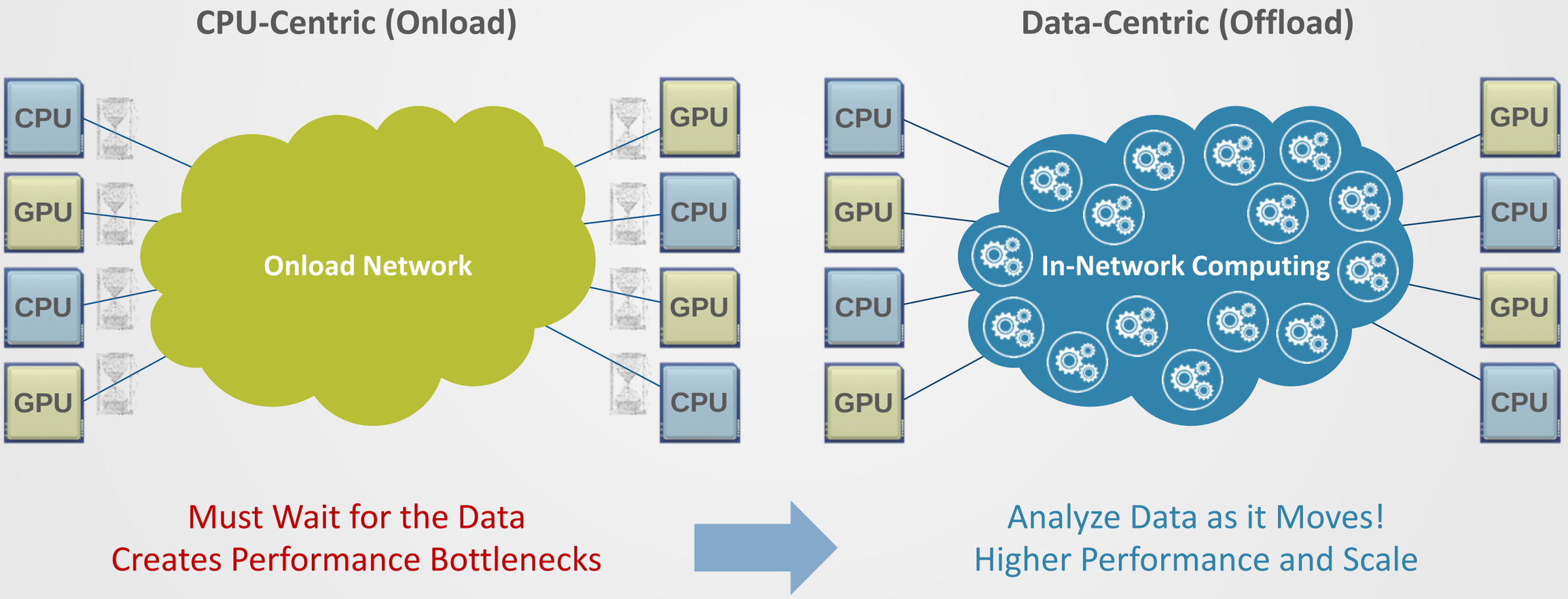
Transceivers
Active Optical and Copper Cables
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)

Software — HPC-X

MPI, SHMEM/PGAS, UPC
For Commercial and Open Source Applications
Leverages Hardware Accelerations

The Need for Intelligent and Faster Interconnect

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

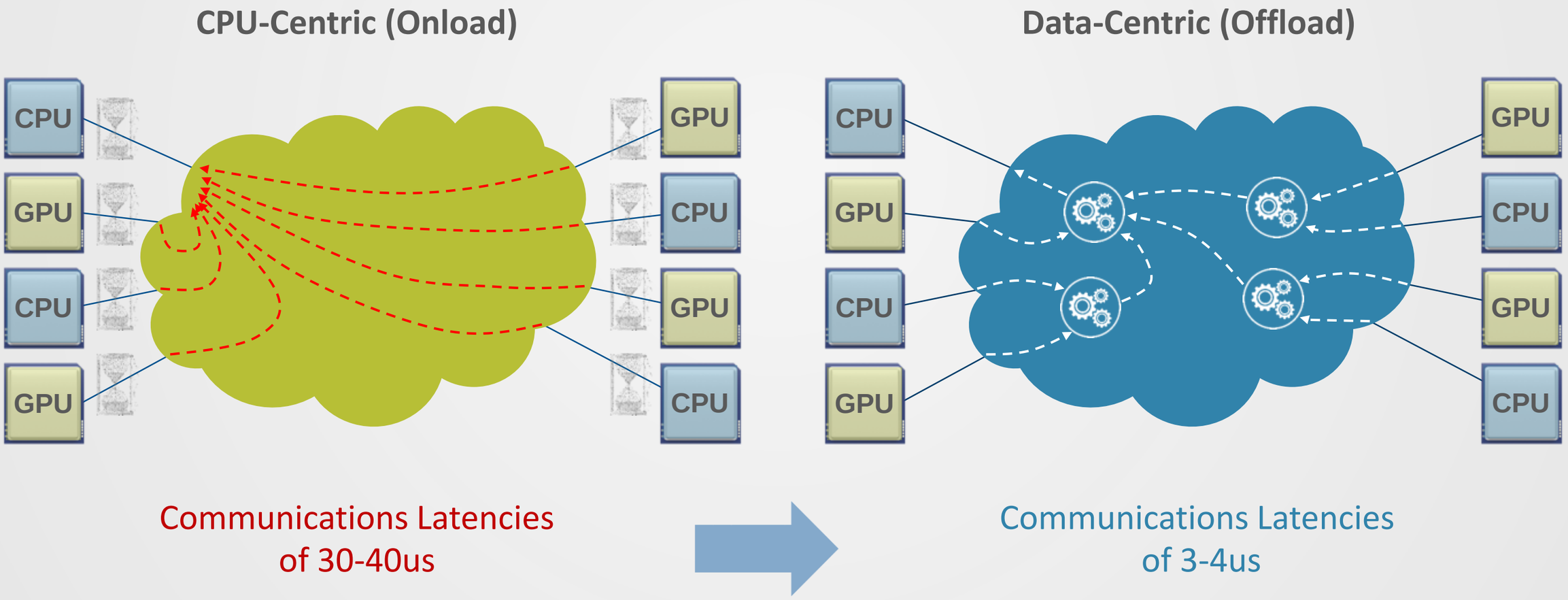


Must Wait for the Data
Creates Performance Bottlenecks

Analyze Data as it Moves!
Higher Performance and Scale

Data Centric Architecture to Overcome Latency Bottlenecks

Intelligent Interconnect Paves the Road to Exascale Performance

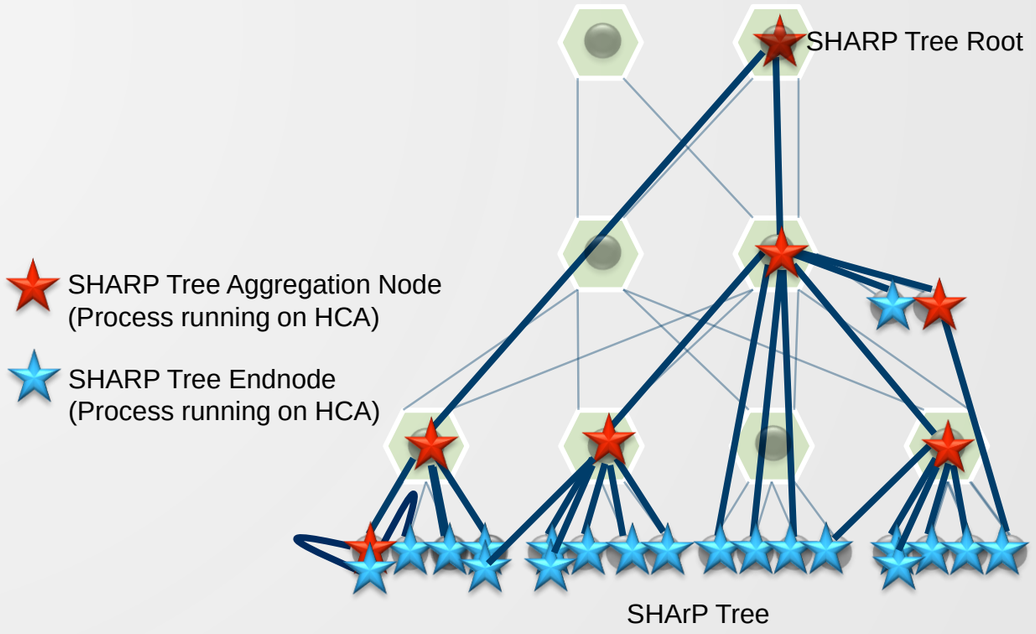


Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

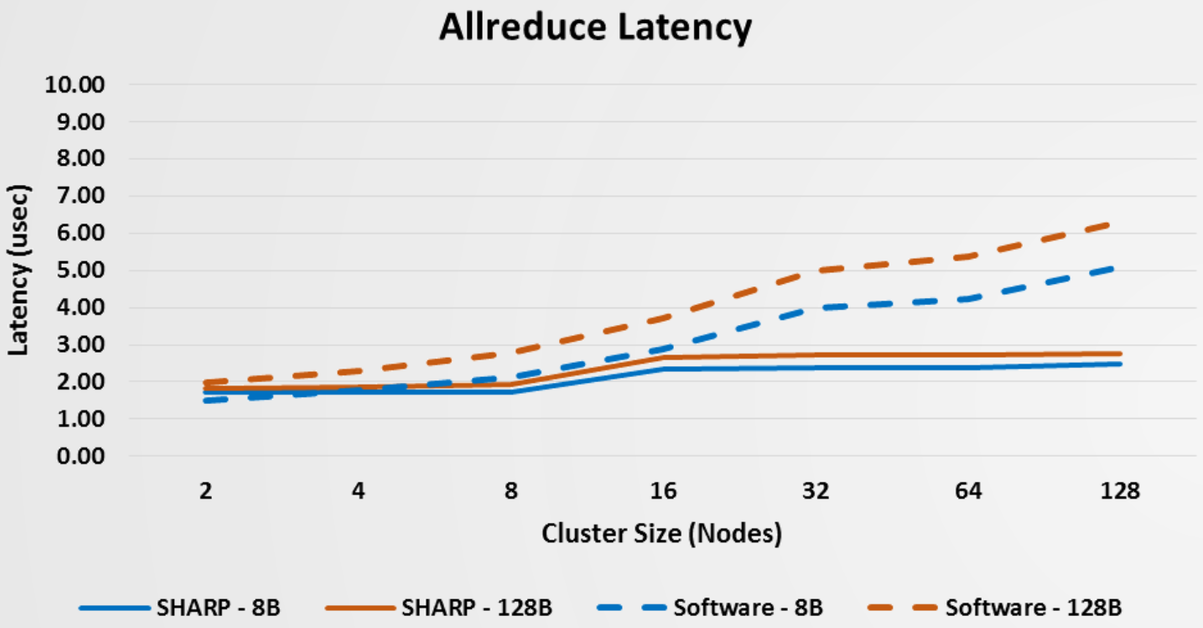
- Reliable Scalable General Purpose Primitive
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations
- Applicable to Multiple Use-cases
  - HPC Applications using MPI / SHMEM
  - Distributed Machine Learning applications
- Scalable 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

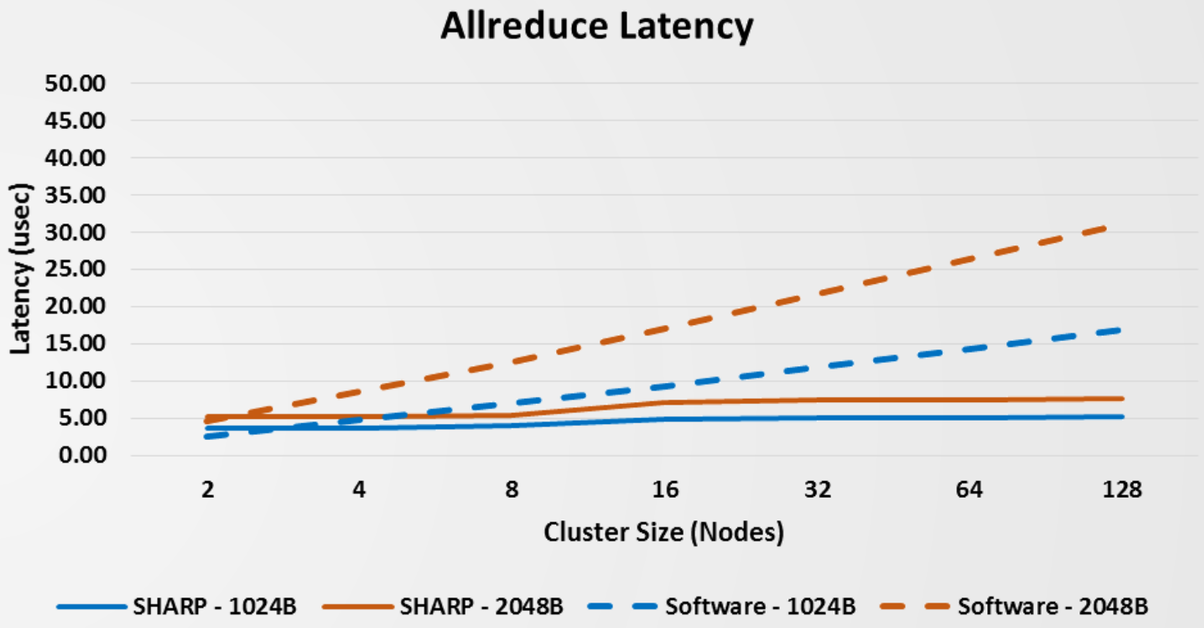
SHARP

Scalable Hierarchical Aggregation and Reduction Protocol



SHARP AllReduce Performance Advantages (128 Nodes)

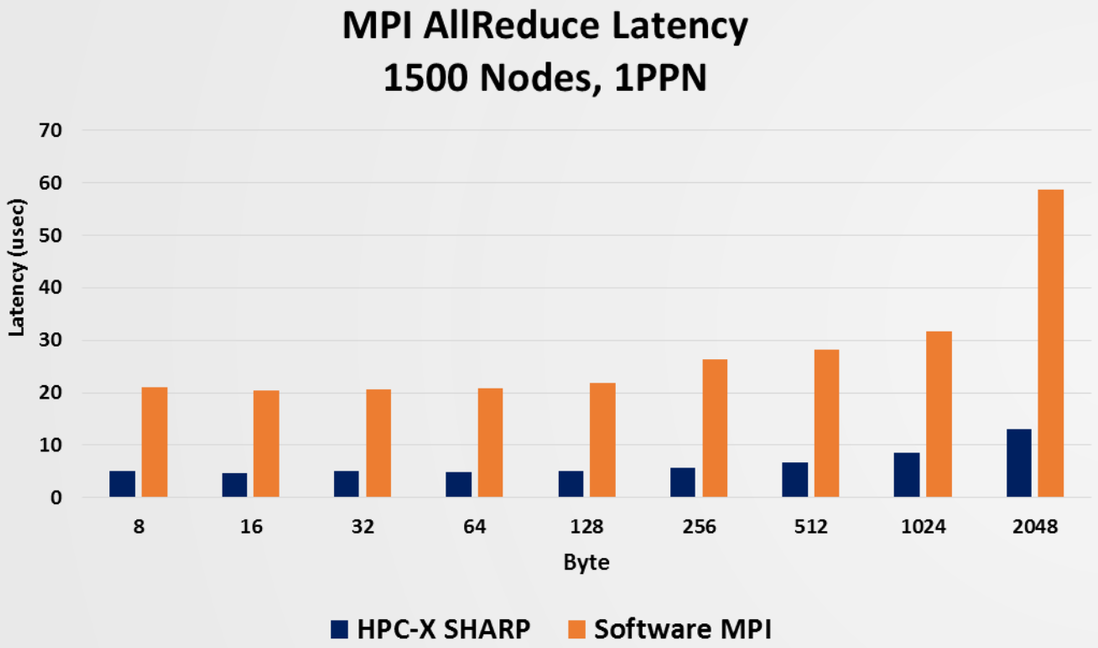


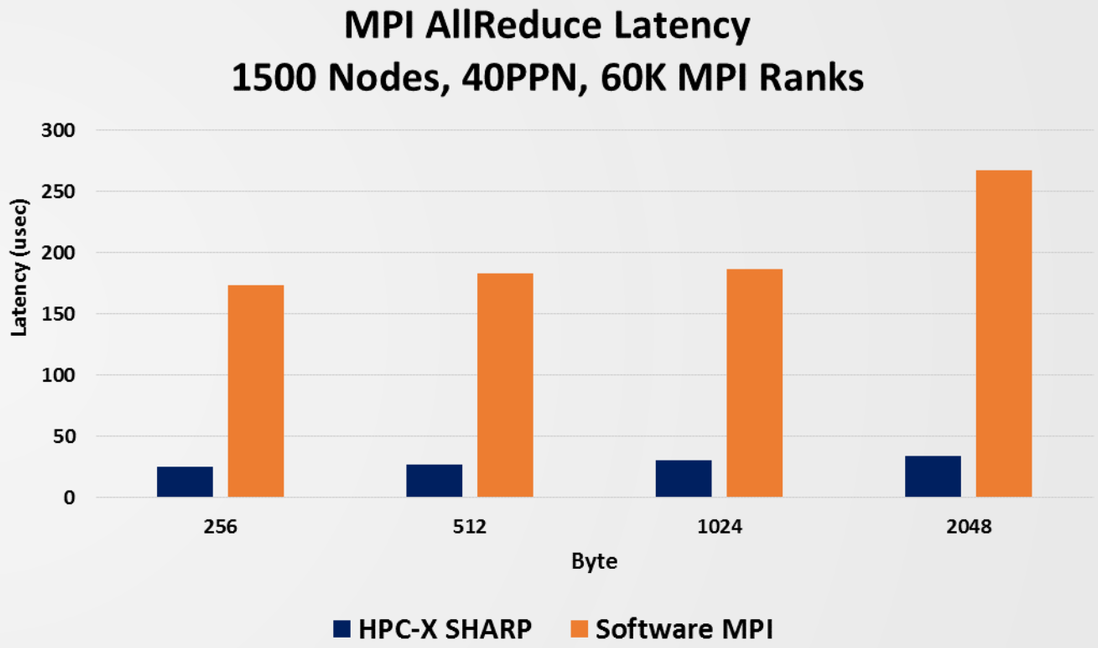


SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

SHARP AllReduce Performance Advantages

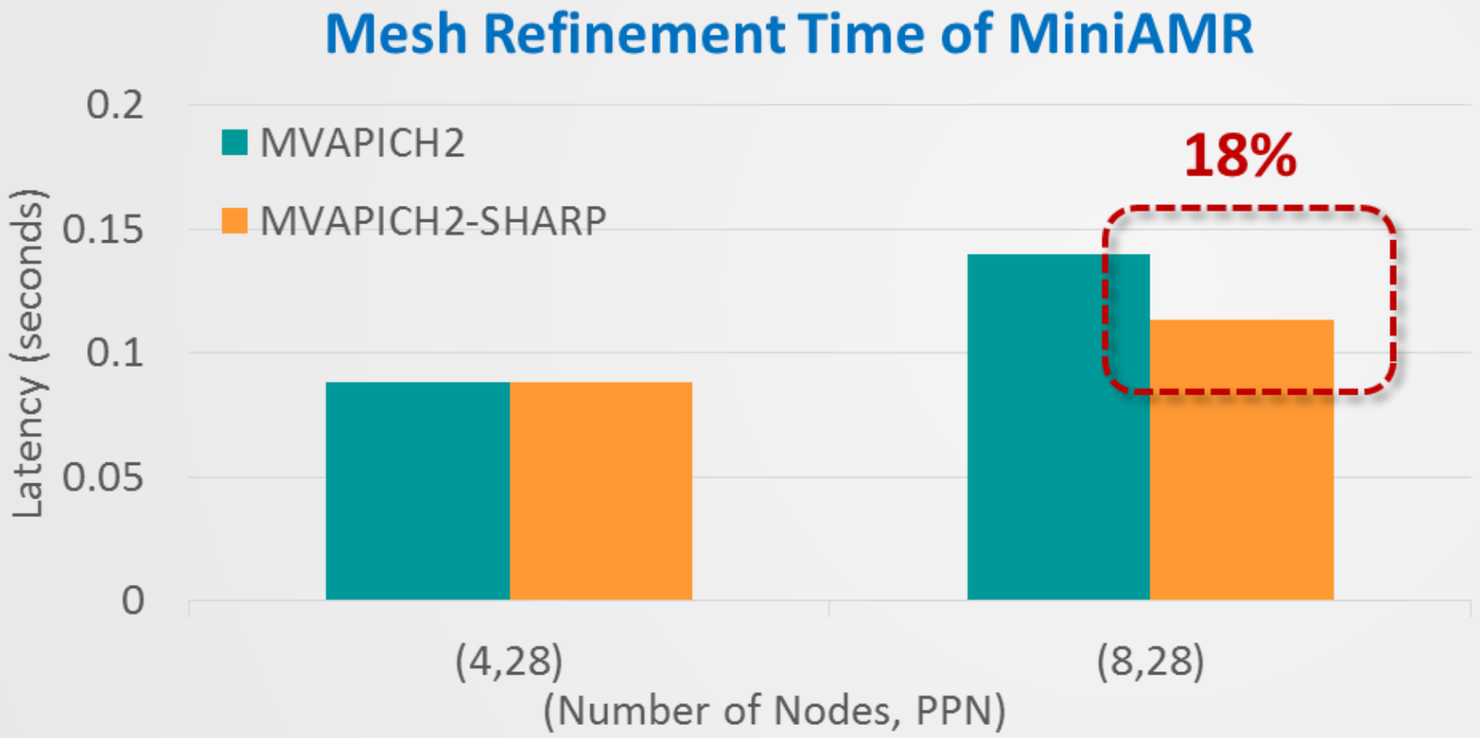
1500 Nodes, 60K MPI Ranks, Dragonfly+ Topology





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

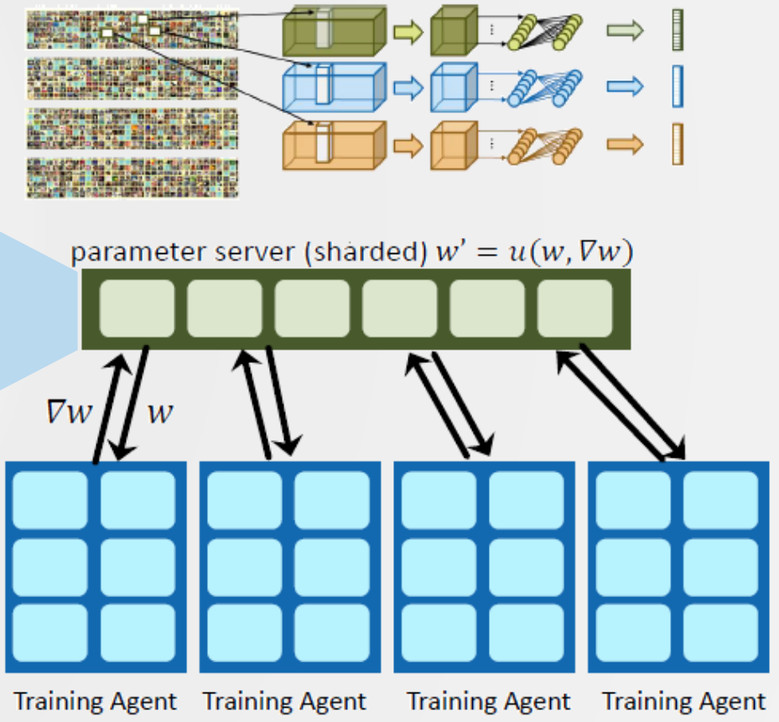
The MVAPICH2 Project
http://mvapich.cse.ohio-state.edu/

Source: Prof. DK Panda, Ohio State University

SHARP Accelerates AI Performance

The CPU in a parameter server becomes the bottleneck

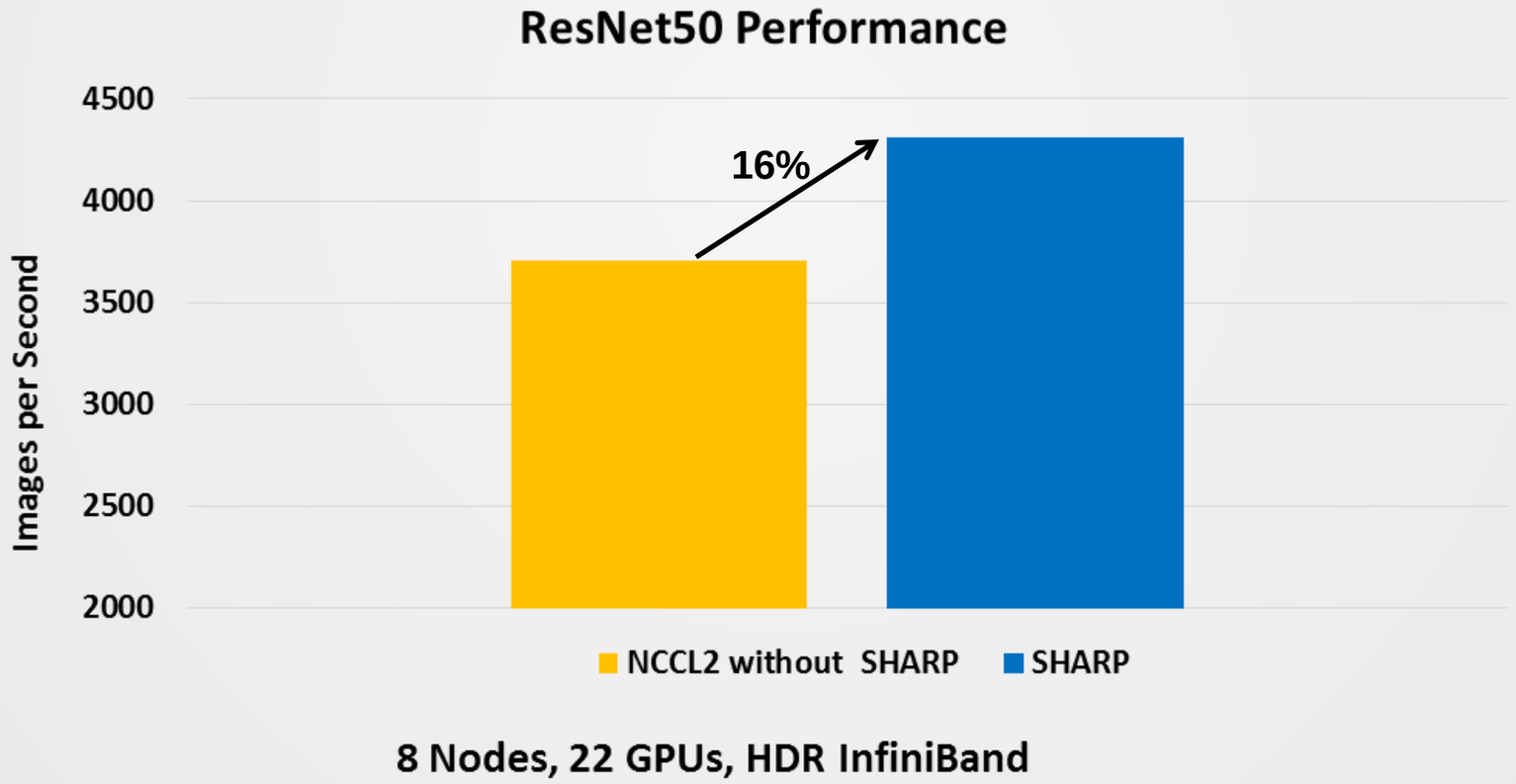
SHARP Scalable Hierarchical Aggregation and Reduction Protocol



Performs the Gradient Averaging
Replaces all physical parameter servers
Accelerate AI Performance

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)

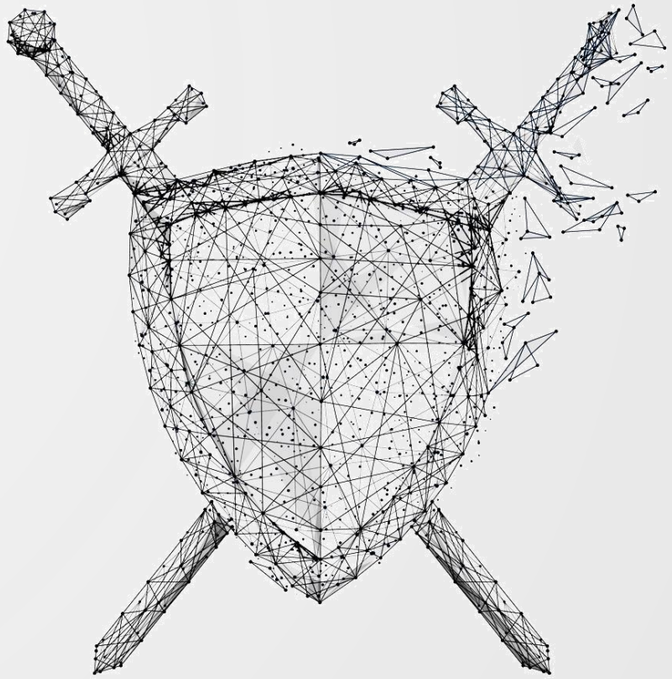


SHIELD - Self Healing Interconnect Technology

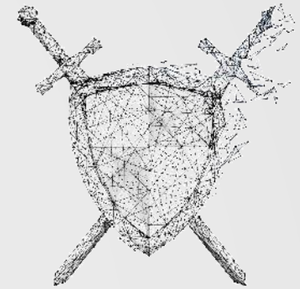
Enables Unbreakable Data Centers

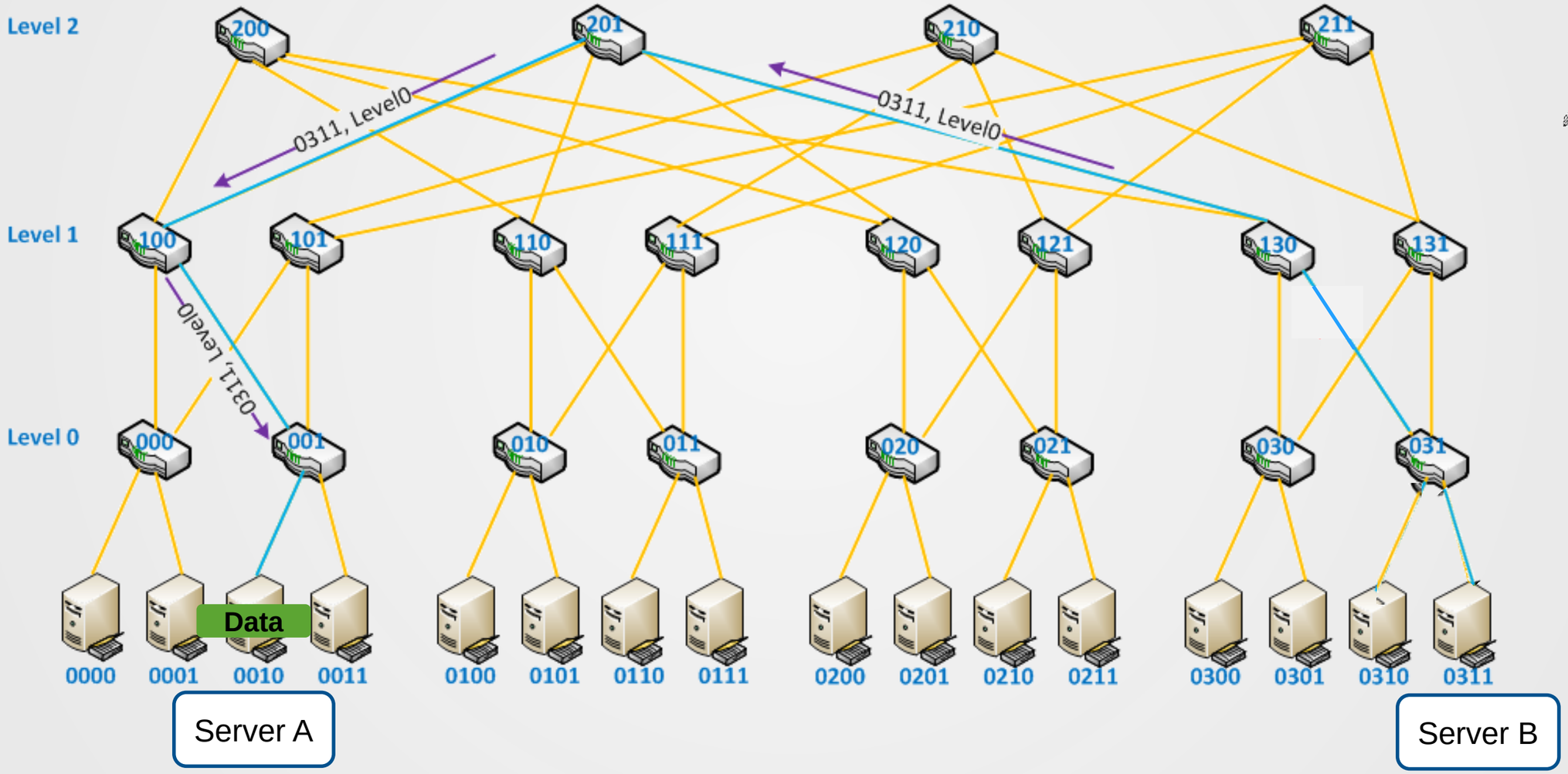
- The ability to overcome network failures, locally, by the switches
- Software-based solutions suffer from long delays detecting network failures
  - 5-30 seconds for 1K to 10K nodes clusters
  - Accelerates network recovery time by 5000X
  - The higher the speed or scale the greater the recovery value
- Available with EDR and HDR switches and beyond

SHIELD
SELF-HEALING INTERCONNECT

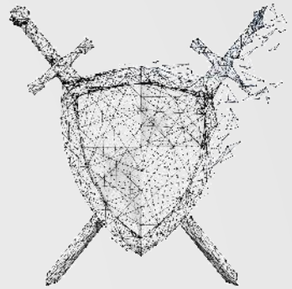


SHIELD: Consider a Flow From A to B





SHIELD: The Simple Case: Local Fix



Level 2

Level 1

200 201 210 211

100 101 110 111 120 121 130 131

001 010 011 020 021 030 031

0311, Level0

0311, Level0

0311, Level0

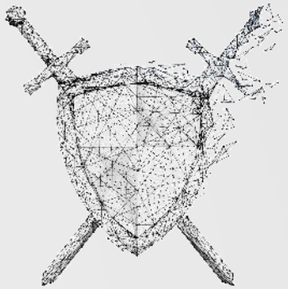
Data

0000 0001 0010 0011 0100 0101 0110 0111 0200 0201 0210 0211 0300 0301 0310 0311

Server A

Server B

SHIELD: The Remote Case - Using Fault Recovery Notifications



Level 2

Level 1

200 201 210 211

100 101 110 111 120 121 FRN 131

010 011 020 021 030 031

0311, Level0

0311, Level0

0311, Level0

Data

Data

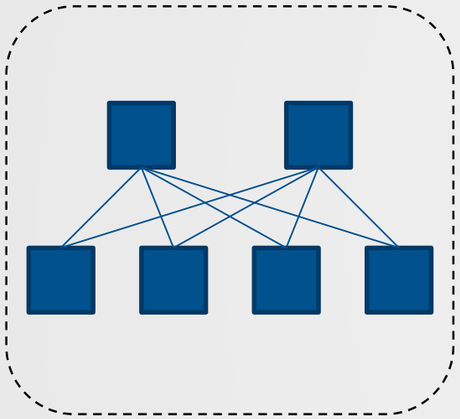
0000 0001 0010 0011 0100 0101 0110 0111 0200 0201 0210 0211 0300 0301 0310 0311

Server A

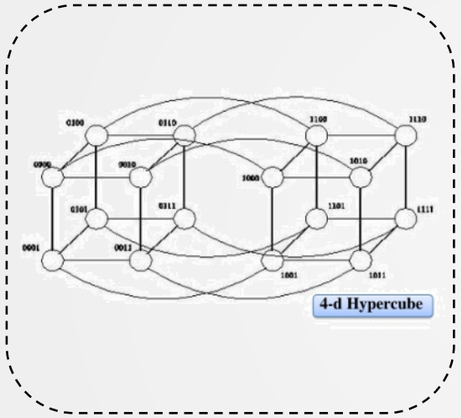
Server B

Network Topologies

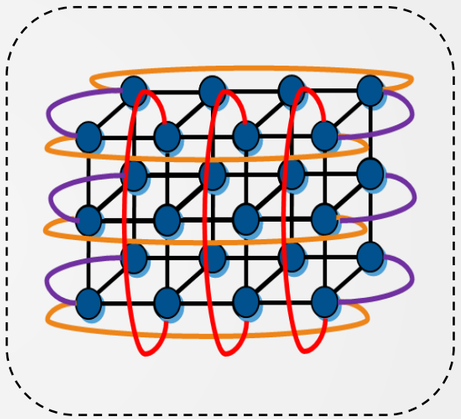
Supporting Variety of Topologies



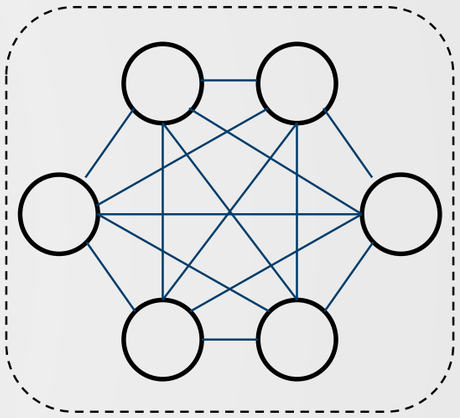
Fat Tree



Hypercube

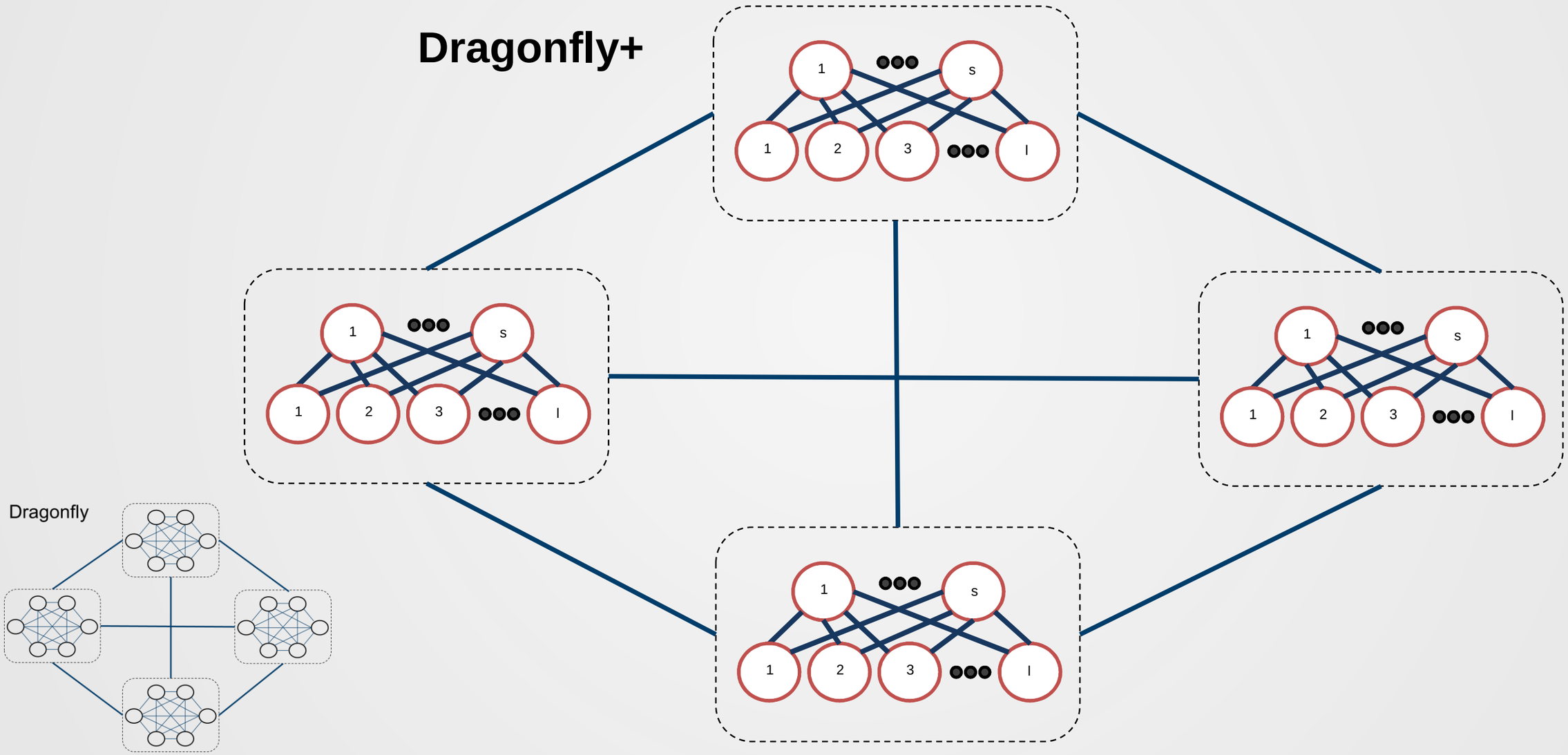


Torus



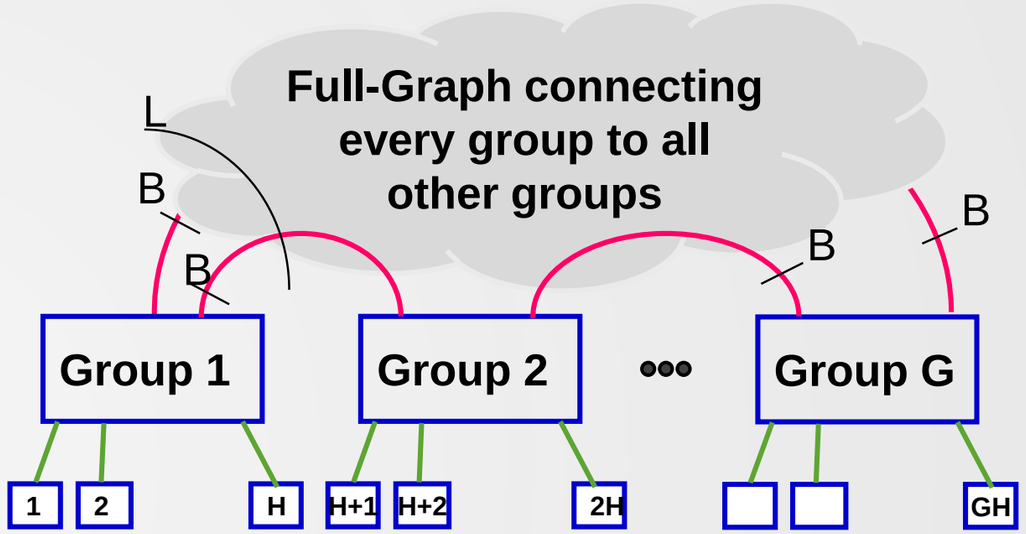
Dragonfly

Traditional Dragonfly vs Dragonfly+

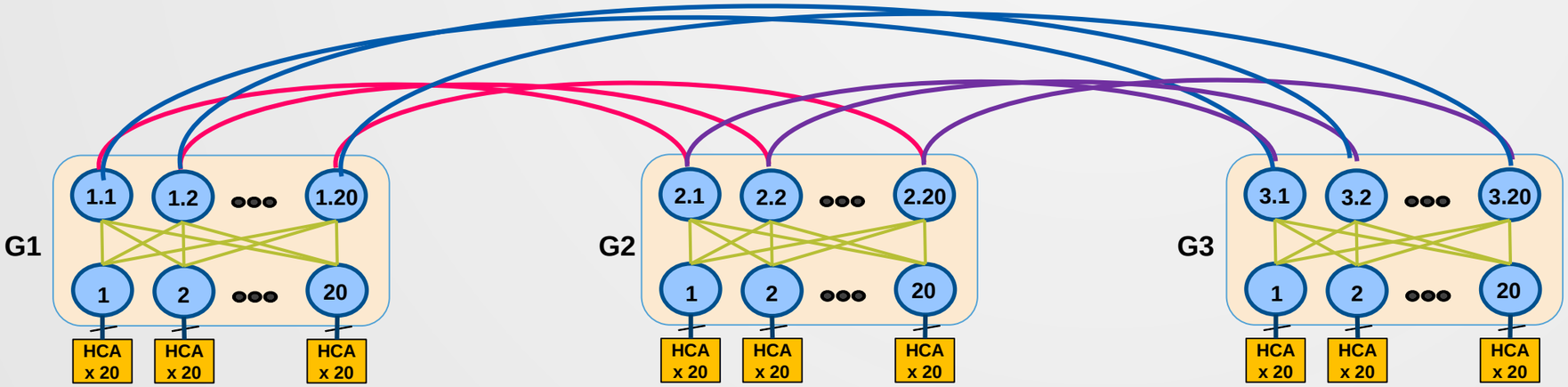


Dragonfly+ Topology

- 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

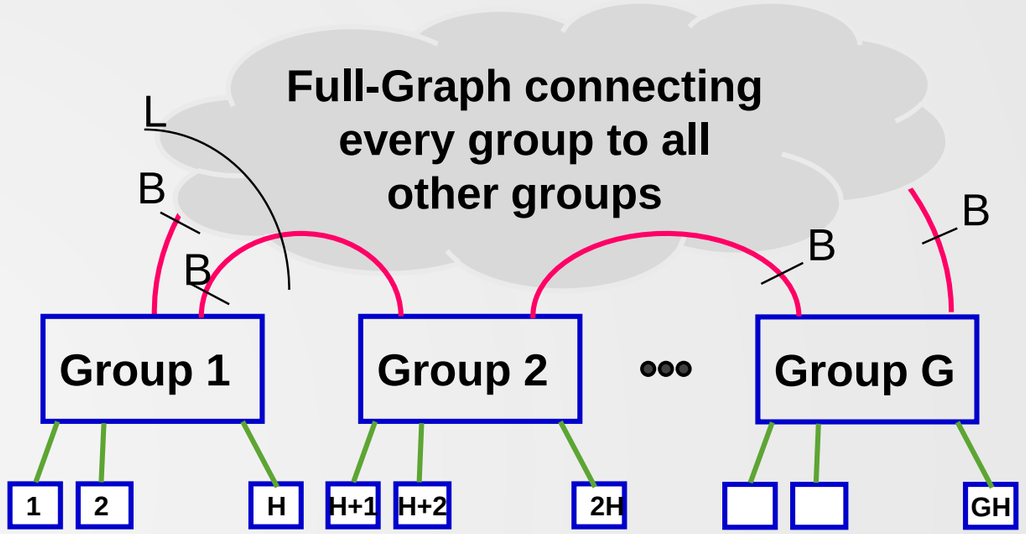


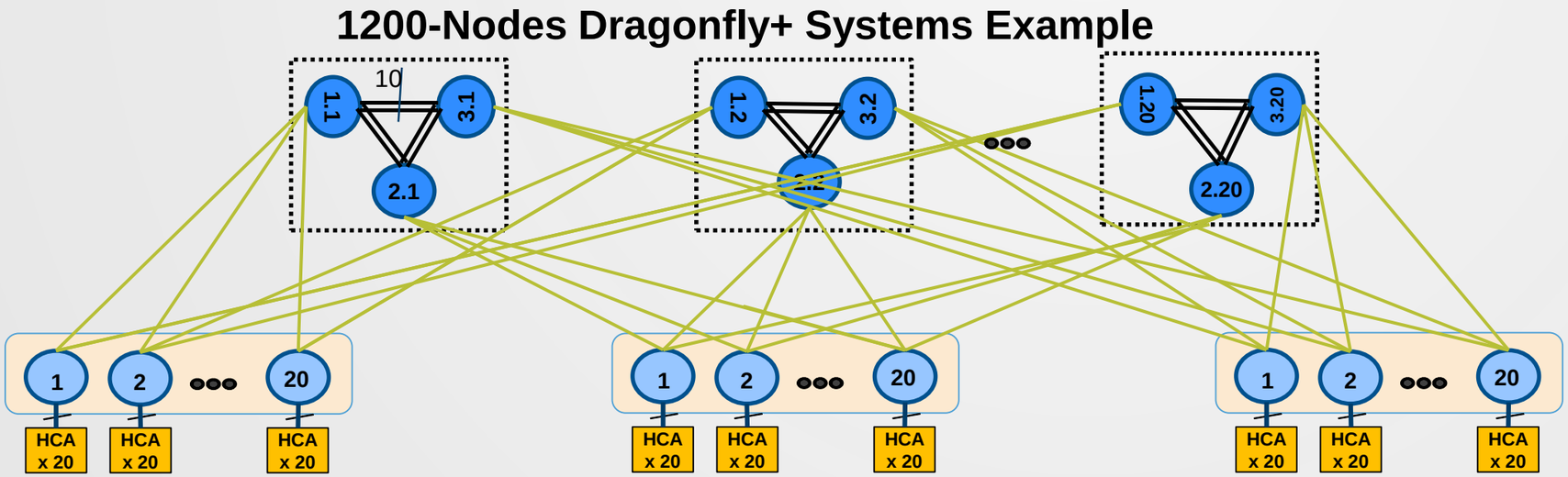
1200-Nodes Dragonfly+ Systems Example



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

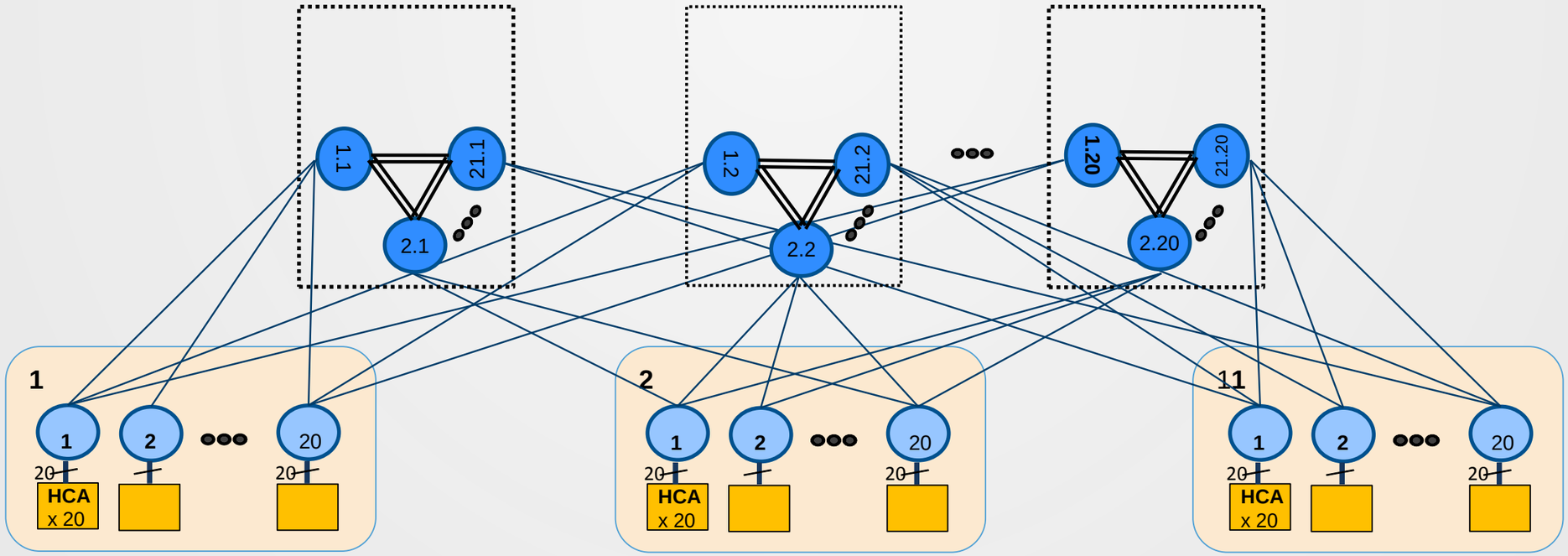




Future Expansion of Dragonfly+ Based System

- Topology expansion of a Fat Tree, or a regular/Aries like Dragonfly requires one of the following
  - Reduction of early phase bisection bandwidth due to reservation of ports on the network switches
  - Re-cabling the long cables
- Dragonfly+ is the only topology that allows system expansion at zero cost
  - While maintaining bisection bandwidth
  - No port reservation
  - No re-cabling

Phase 1:
11x400 =
4400 hosts

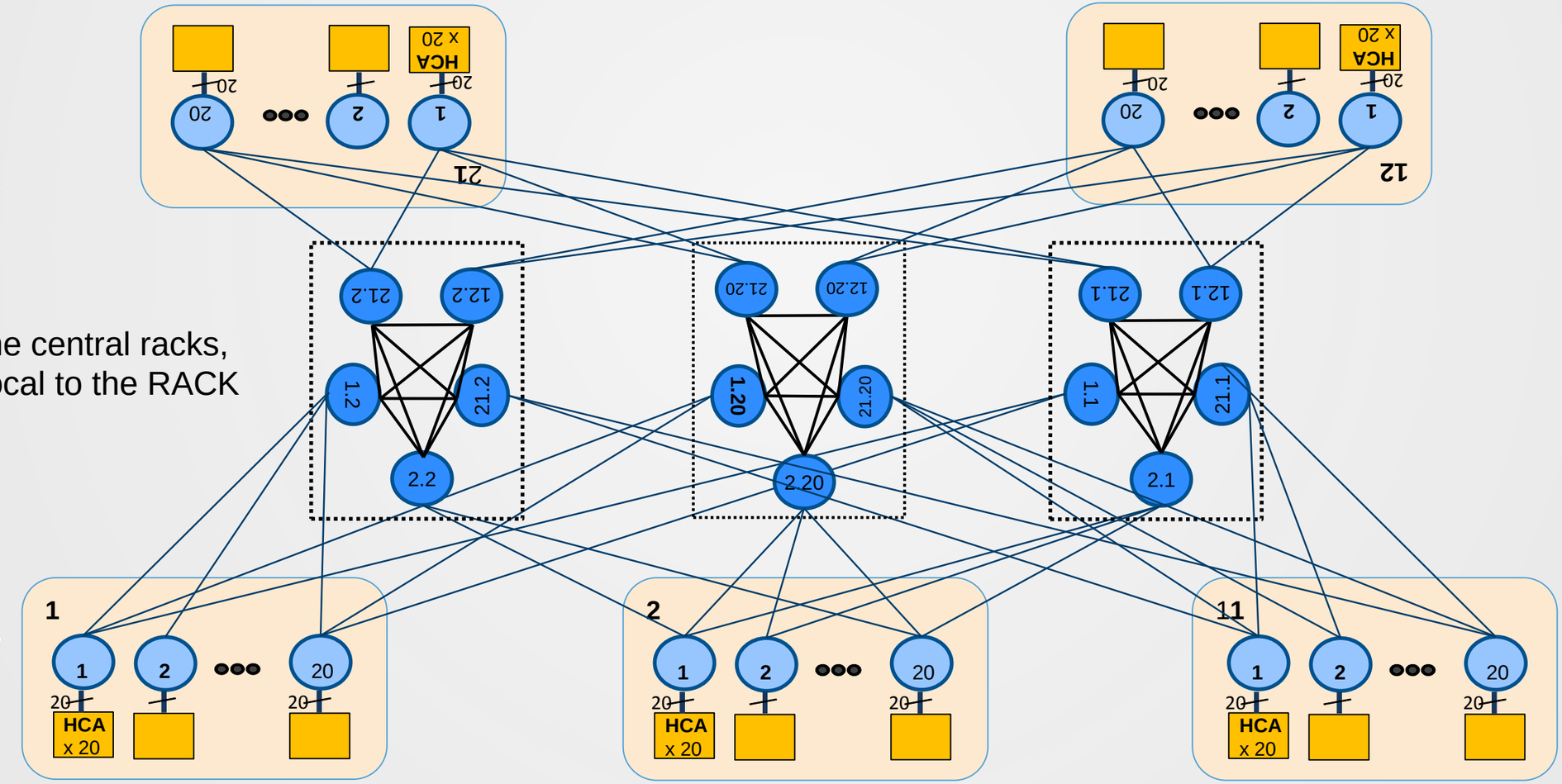


Future Expansion of Dragonfly+ Based System

Phase 2:
+10x400 =
8400 hosts

Re-cable the central racks,
a change local to the RACK

Phase 1:
11x400 =
4400 hosts



Thank You