Interconnect Your Future

Interconnect Topology Considerations Applied To Differing Applications And Clusters

December 2018

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.

Mellanox InfiniBand and Ethernet Accelerate World-Leading Supercomputers on the Nov'18 TOP500 List

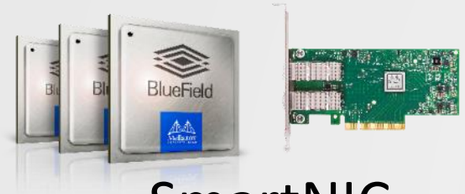
Mellanox connects 53% of overall TOP500 platforms or 265 systems (InfiniBand and Ethernet), Demonstrating 38% Growth in 12 months (Nov'17-Nov'18)

- InfiniBand accelerates the fastest HPC and AI supercomputer in the world – Oak Ridge National Laboratory 'Summit' system
- InfiniBand accelerates the top 3 supercomputers in the world - #1 (USA), #2 (USA), #3 (China)
- InfiniBand connects 135 supercomputers, or nearly 55% of overall HPC systems on the TOP500 list
- InfiniBand is the most used high-speed interconnect for the TOP500 systems
- Mellanox connects 130 Ethernet systems (25 Gigabit and faster), or 51% of total Ethernet systems
- The TOP500 list has evolved to include both HPC and cloud / hyperscale (non-HPC) platforms
- Nearly half of the platforms on the TOP500 list can be categorized as non-HPC application platforms (mostly Ethernet-based)

InfiniBand is the Interconnect of Choice for HPC and AI Infrastructures

Mellanox Ethernet is the Interconnect of Choice for Cloud and Hyperscale Platforms

HPC and AI Needs the Most Intelligent Interconnect



SmartNIC

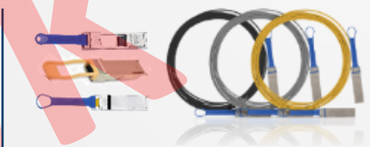
System on a Chip

Innova ConnectX

Adapters

Spectrum Quantum

Switches



LinkX

Cables & Transceivers

Higher Data Speeds

Faster Data Processing

Better Data Security

HDR 200G InfiniBand Accelerates Next Generation HPC/AI Systems

Highest Performance HDR 200G InfiniBand

ConnectX-6 Adapters

200Gb/s, 0.6us Latency
215 Million Messages per Second
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)

Mellanox Quantum Switch

40 HDR (200Gb/s) Ports
80 HDR100 (100Gb/s) Ports
16Tb/s Throughput, 15.6 Billion msg/sec

LinkX Interconnect

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

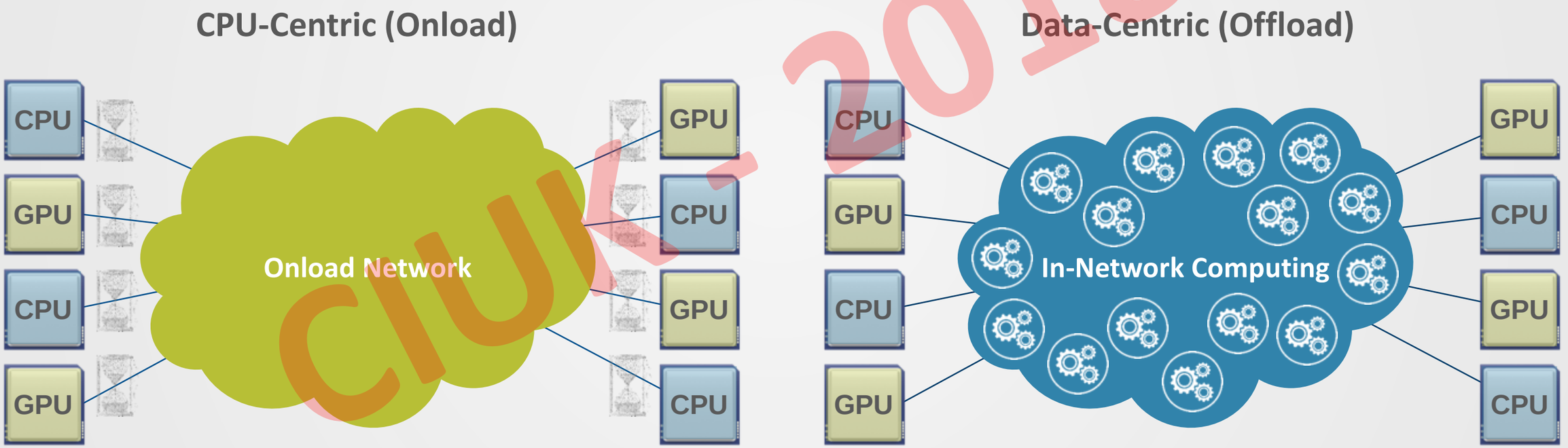
Lawrence Livermore National Laboratory

TACC
TEXAS ADVANCED COMPUTING CENTER

MICHIGAN

The Need for Intelligent and Faster Interconnect

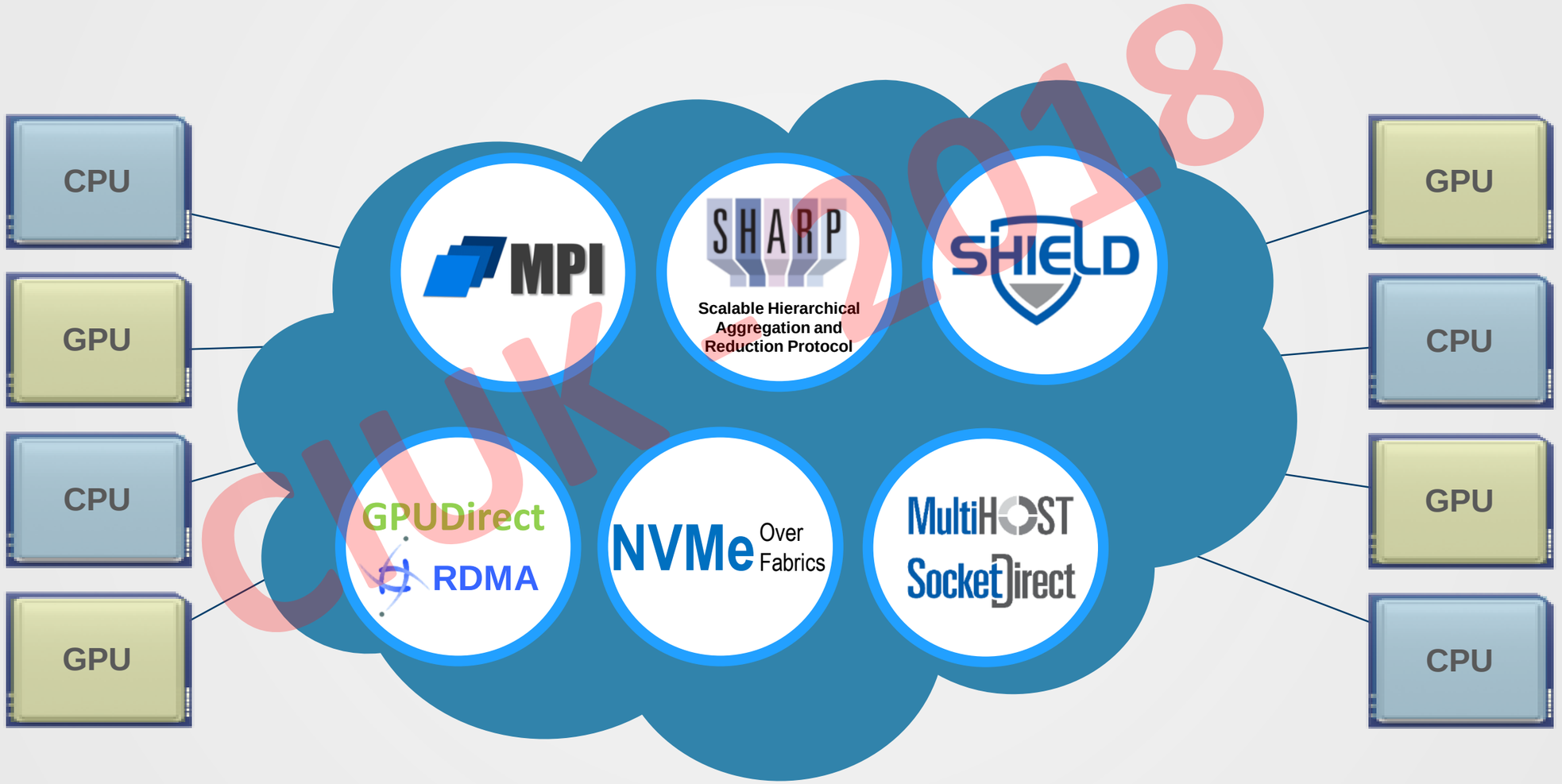
Faster Data Speeds and In-Network Computing
Enable Higher Performance and Scale



Must Wait for the Data
Creates Performance Bottlenecks

Analyze Data as it Moves!
Higher Performance and Scale

In-Network Computing to Enable Data-Centric Data Centers



In-Network Computing Delivers Highest Performance

SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

In-Network Computing

10X Performance Acceleration
Critical for HPC and Machine Learning Applications

MPI In-Network Computing

35X Performance Acceleration
Delivers Highest Application Performance

SHIELD Self-Healing Technology

5000X Faster Network Recovery
Unbreakable Data Centers

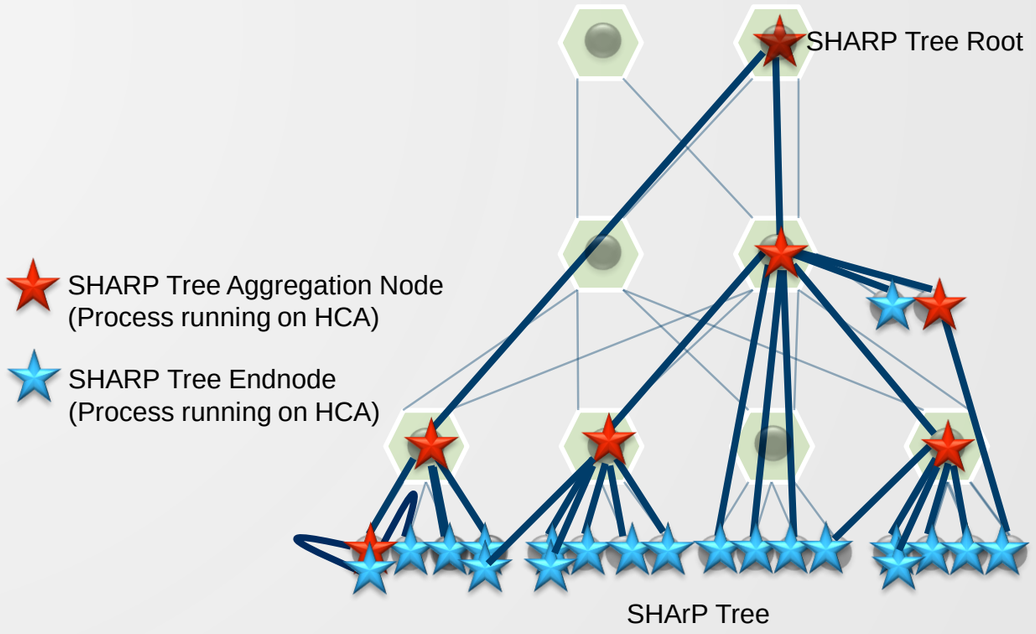
GPUDirect™ RDMA
GPU Acceleration Technology

10X Performance Acceleration
Critical for HPC and Machine Learning Applications

Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

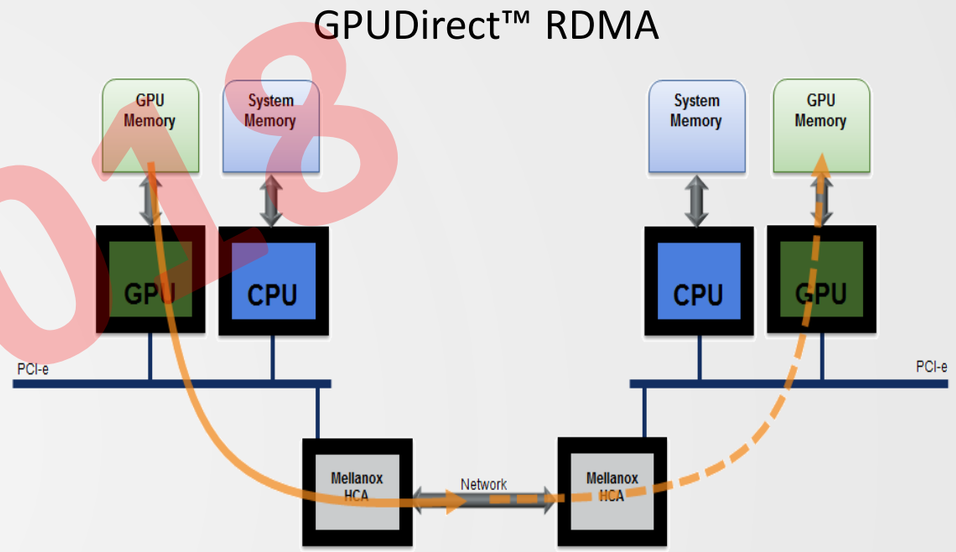
- Reliable Scalable General Purpose Primitive
  - In-network Tree based aggregation mechanism
  - Large number of groups
  - Multiple simultaneous outstanding operations
- Applicable to Multiple Use-cases
  - HPC Applications using MPI / SHMEM
  - Distributed Machine Learning applications
- Scalable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

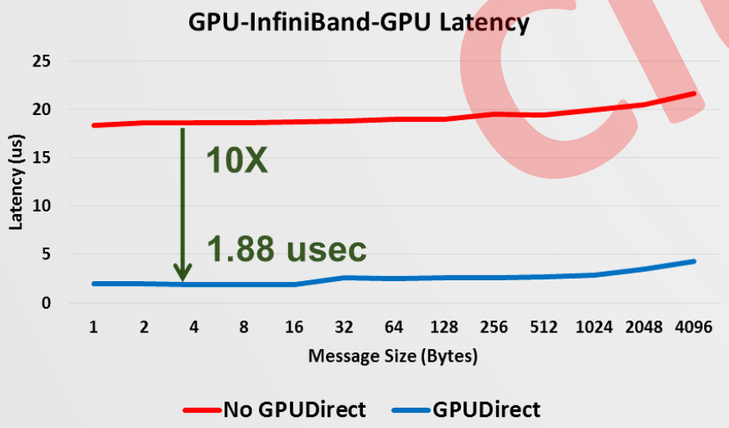
SHARP
Scalable Hierarchical Aggregation and Reduction Protocol

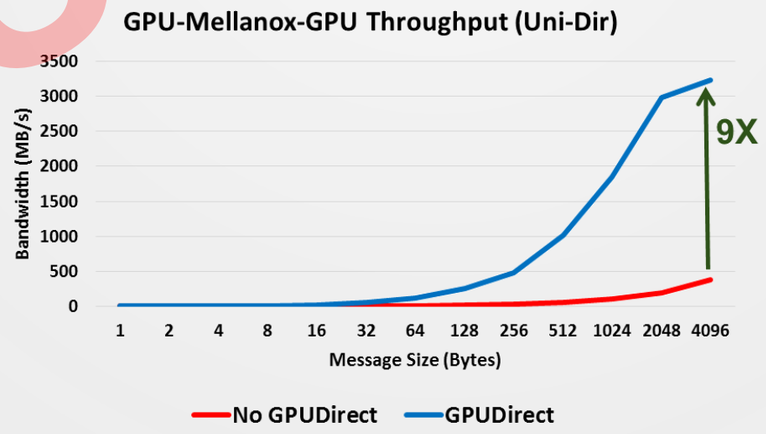


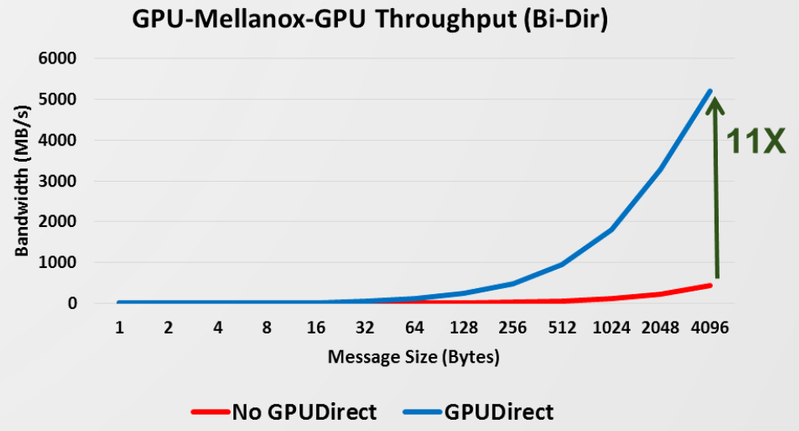
10X Higher Performance with GPUDirect™ RDMA

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



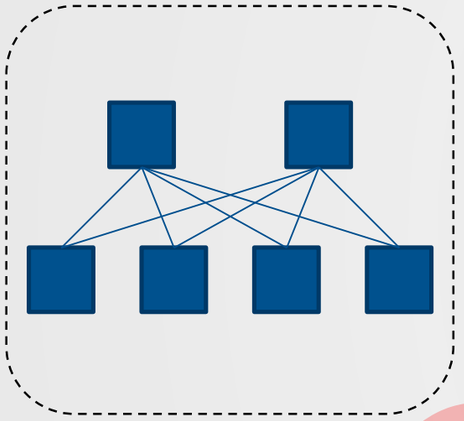




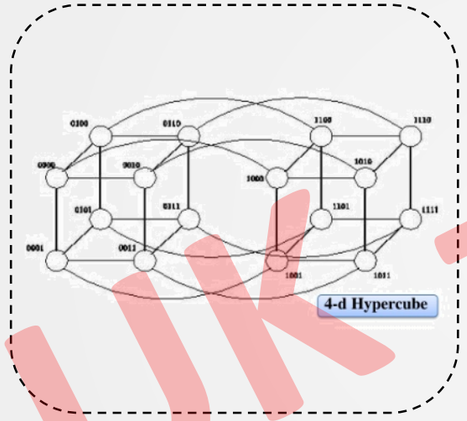


Network Topologies

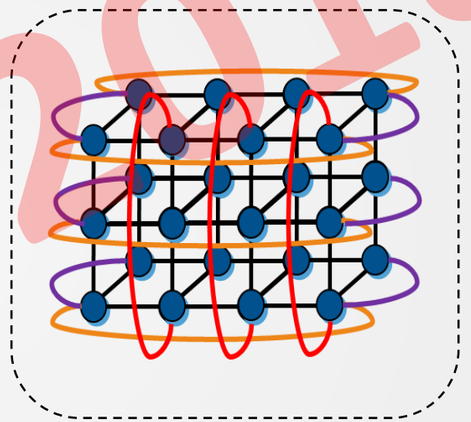
Supporting Variety of Topologies



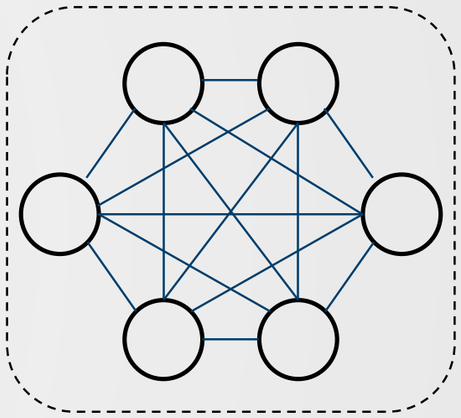
Fat Tree



Hypercube

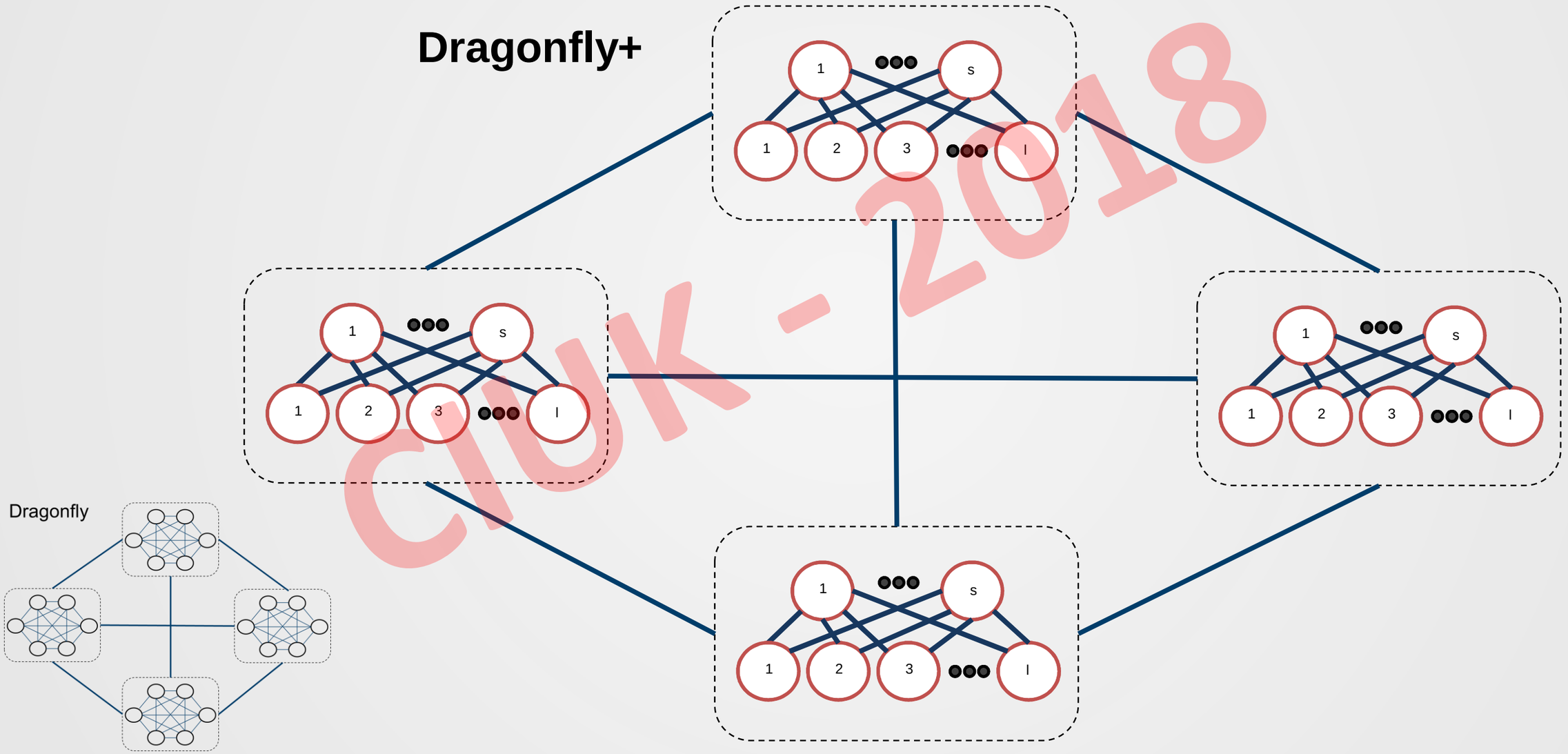


Torus



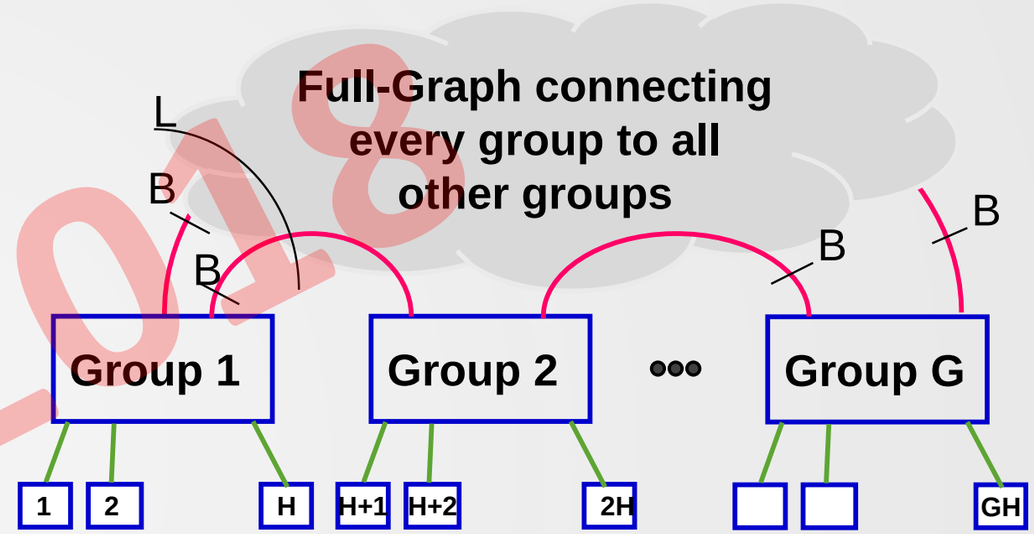
Dragonfly

Traditional Dragonfly vs Dragonfly+

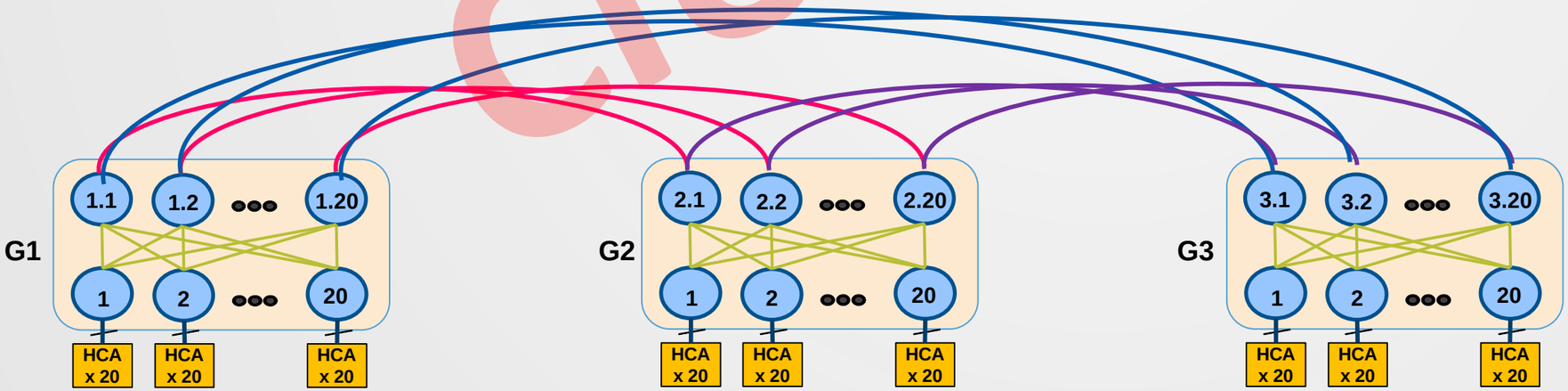


Dragonfly+ Topology

- 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

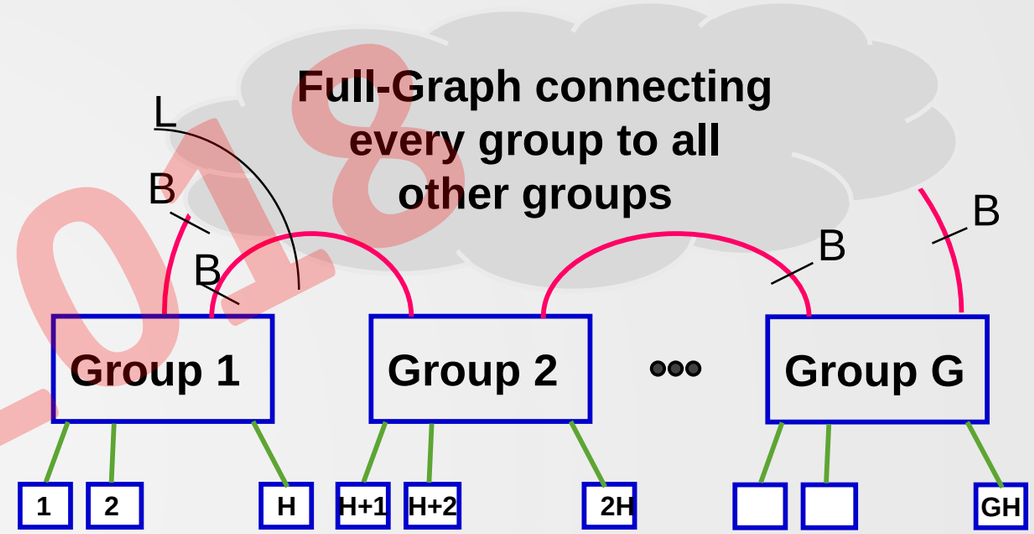


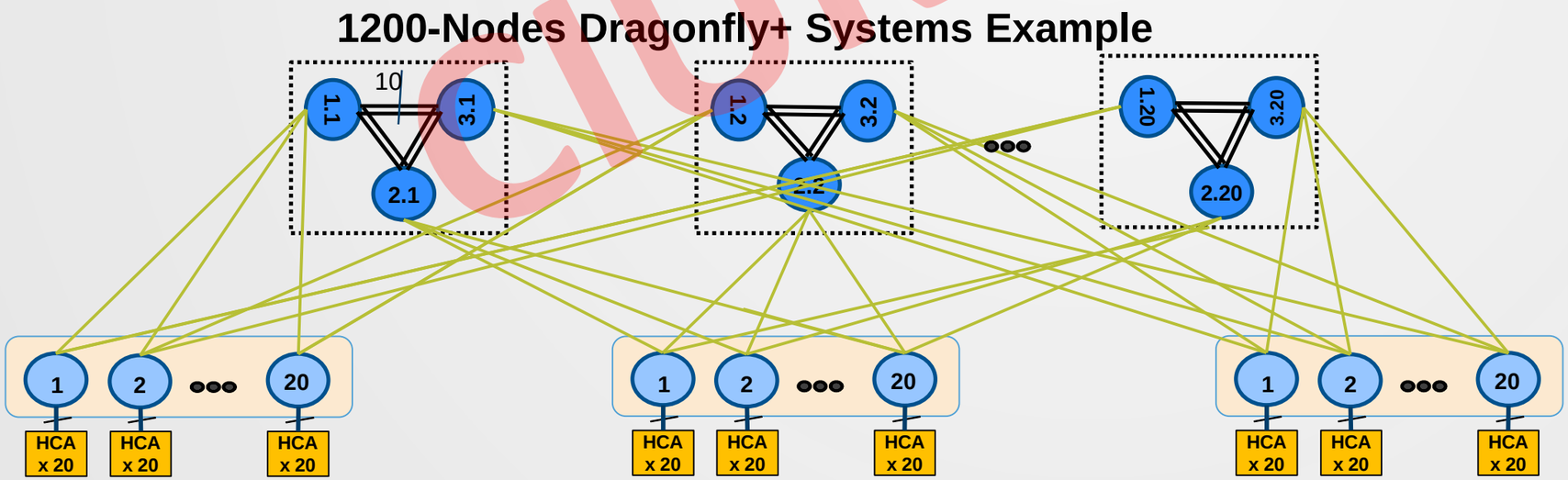
1200-Nodes Dragonfly+ Systems Example



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

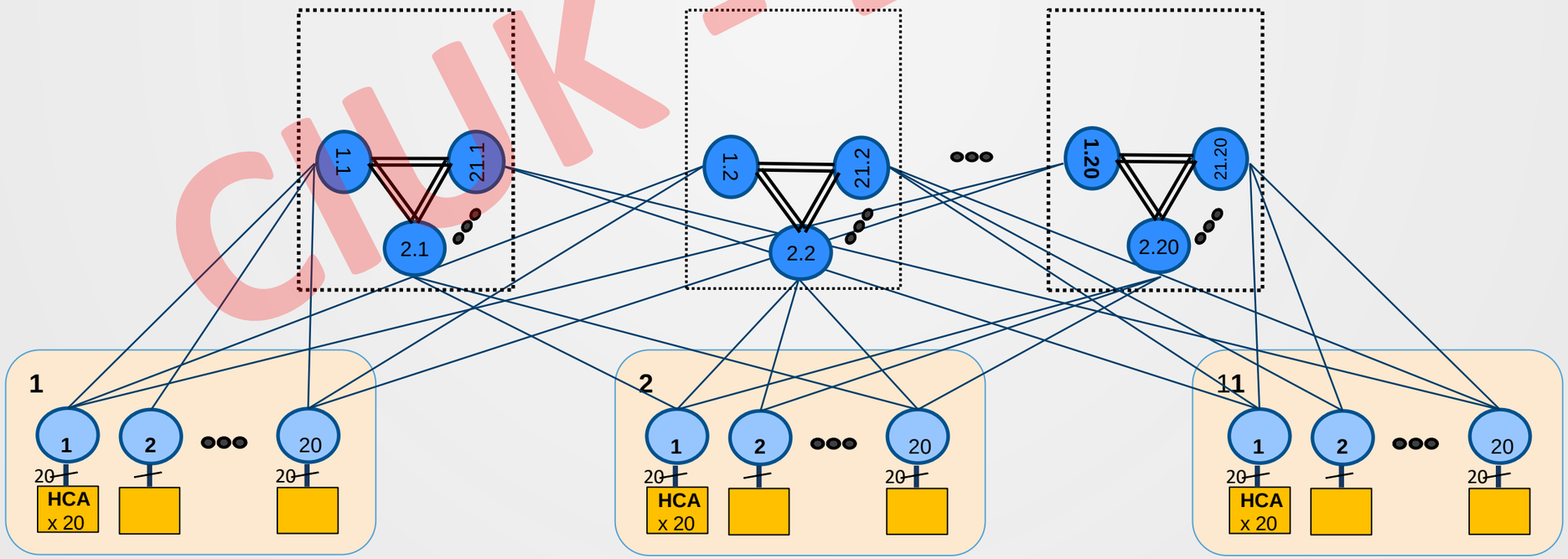




Future Expansion of Dragonfly+ Based System

- 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

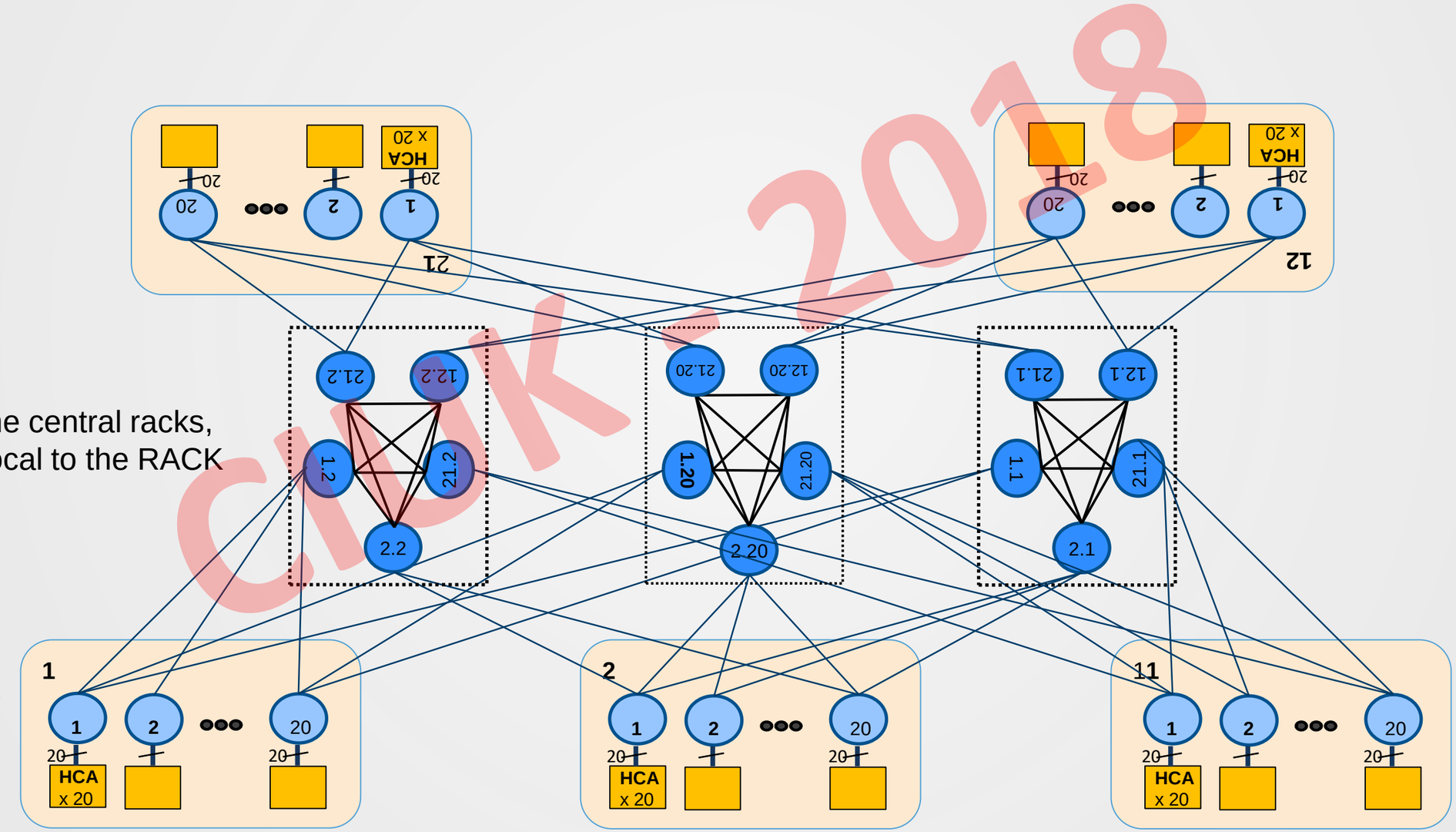


Future Expansion of Dragonfly+ Based System

Phase 2:
+10x400 =
8400 hosts

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

Phase 1:
11x400 =
4400 hosts



Questions?

Thank You