


IIT Kharagpur | IIT Madras | IIT Goa | IIT Palakkad

Design Principles for Building High Performance Clusters

Networking Fundamentals

Ashrut Ambastha
NVIDIA

National Supercomputing Mission

Centre for Development of Advanced Computing

Networks

Interlinking of Multiple Compute Elements

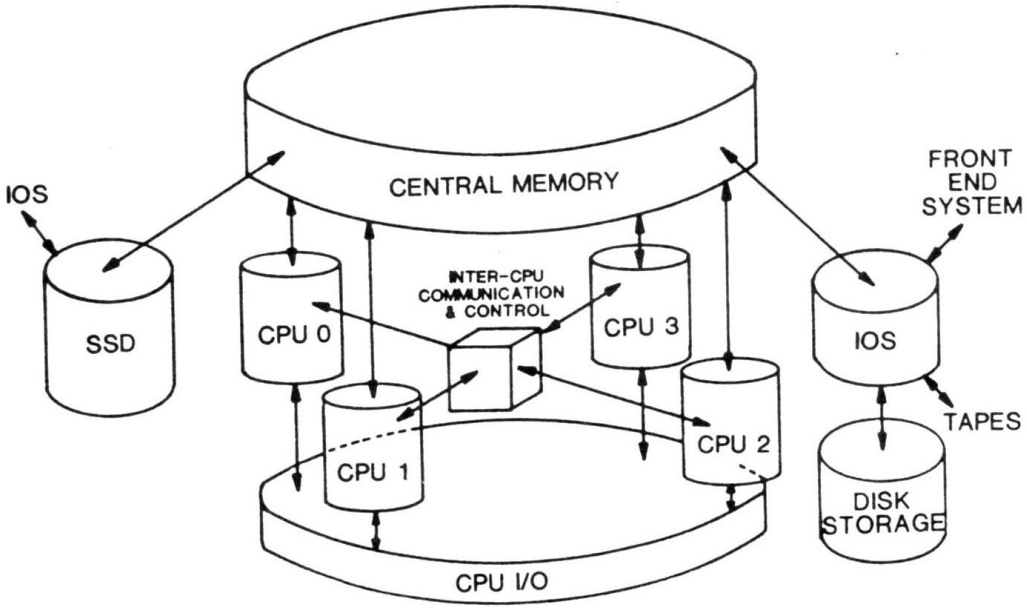
SMP?? Clusters??

Cache Coherency?? Message Passing??

Bandwidth?? Latency??

Traditional SMP Machines

*CRAY X-MP: https://commons.wikimedia.org/wiki/File:BSC-CRAY-X-MP-EA-A.JPG



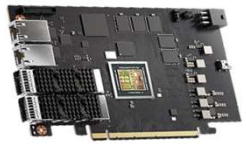
Modern Supercomputing Clusters

Network Components for High-Performance Clusters

Compute
Servers
GPUs, CPUs

Network
Switches

Storage
Subsystem
HDD/SSD/Flash



Network Switches, Network Adapters, Cables & Transceivers

HPC Interconnect History and Development

First Teraflop Supercomputer
Sandia ASCI Red
Intel

First Petaflop Supercomputer
LANL Roadrunner
IBM / Mellanox InfiniBand

Quadrics
QsNet
gnodal
QsNet with Gateway to Ethernet
Myricom
Myrinet
PathScale
InfiniPath (InfiniBand)
QLOGIC
TrueScale (InfiniBand)
intel
OmniPath
CRAY
Crossbar
Seastar
Gemini
Aries
Slingshot
Mellanox
InfiniBandSDR
DDR
QDR
FDR
EDR
HDR
NDR
XDR

1995 2000 2005 2010 2015 2020 2025

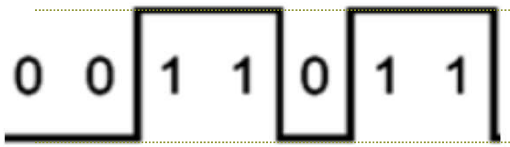
Fundamentals of the Physical Layer

Transmitting Bits and Bytes

NRZ
Non-Return to Zero

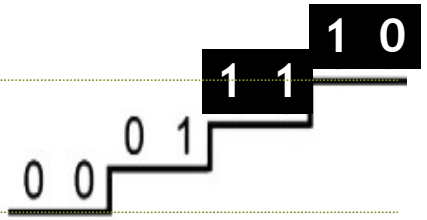
PAM4
Pulse Amplitude Modulation 4-levels

1G, 10G, 25G NRZ
(1 bit/clock)

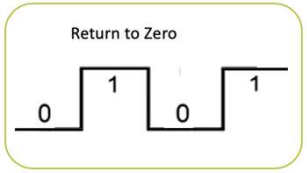


NRZ allows for multiple 1's in a row without having to return to zero

50G-PAM4 Signaling
(2 bits/clock)



Return-to-Zero
After each data pulse



NRZ used in:
1G 1x1G
10G 1x10G
40G 4x10G

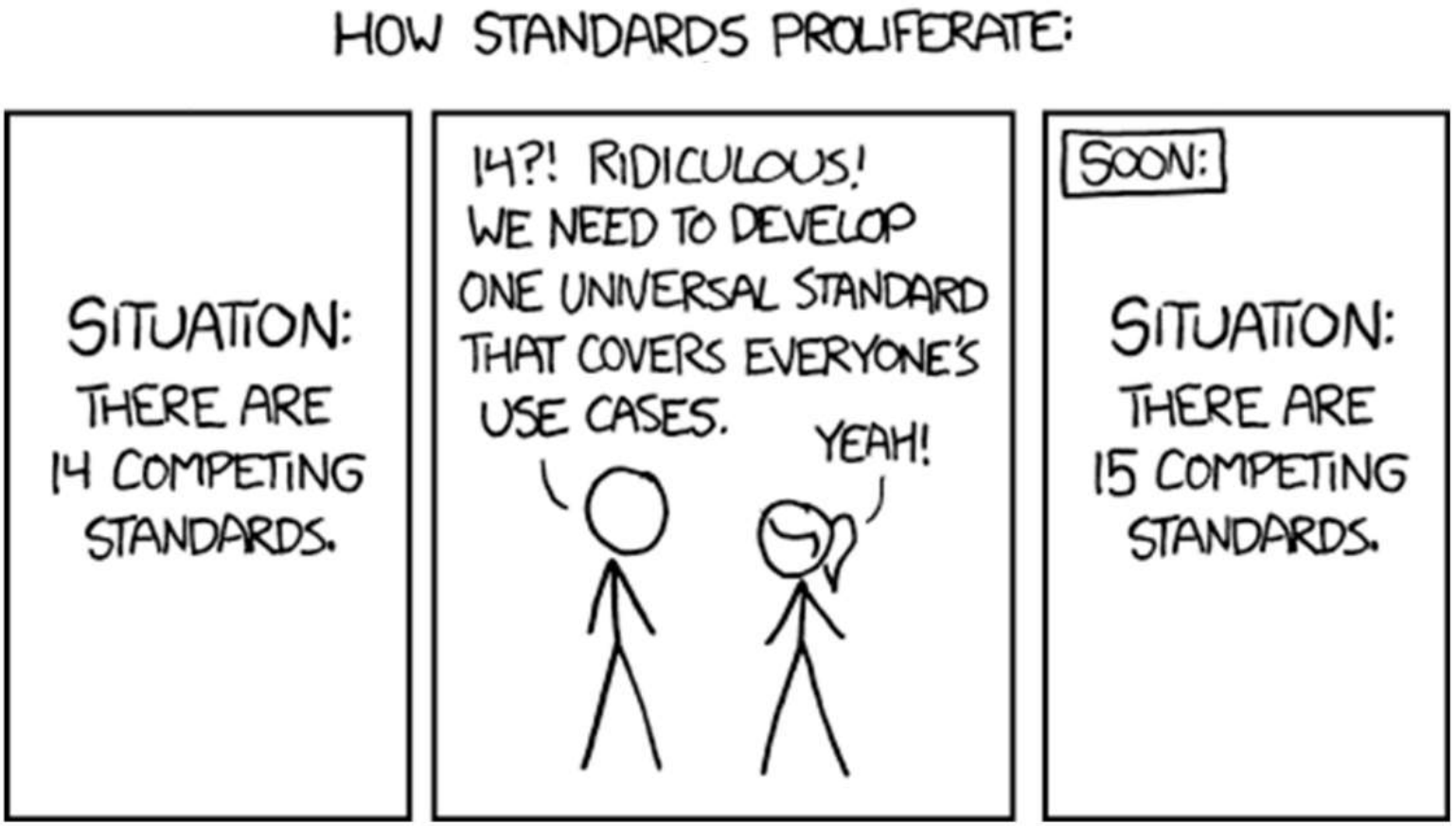
25G 1x25G
100G 4x25G

+ Doubles the data rate
+ Keeps same 25GHz clock
+ Enables lower cost material system

PAM4 used in:
200G 4x50G-PAM4
400G 8x50G-PAM4

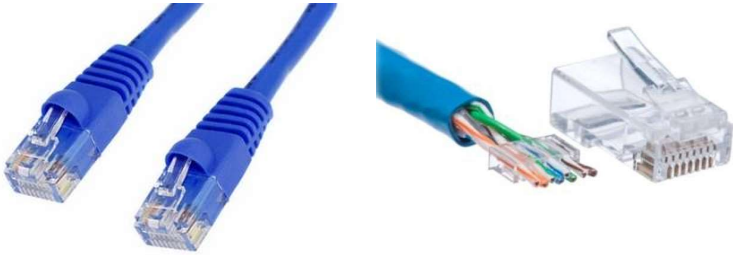
400G 4x100G-PAM4
800G 8x100G-PAM4

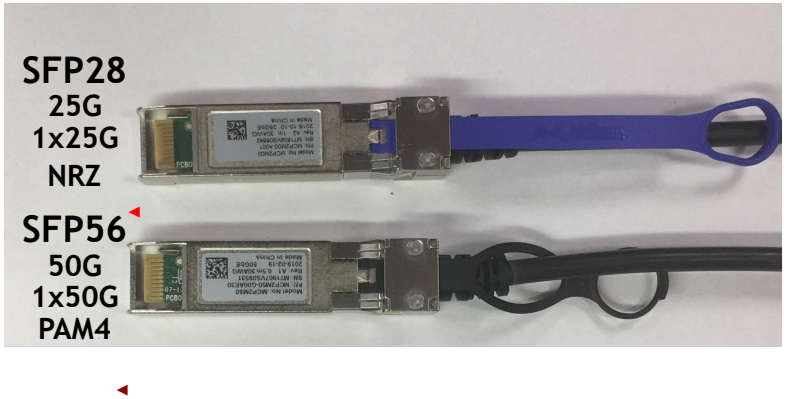
What We Don’t Want to Achieve



Connector Evolution and Need

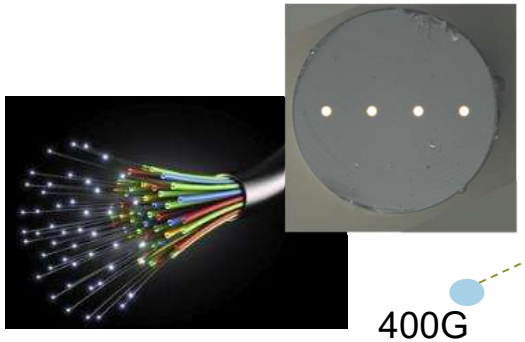
More space for optics & electronics; More heat dissipation

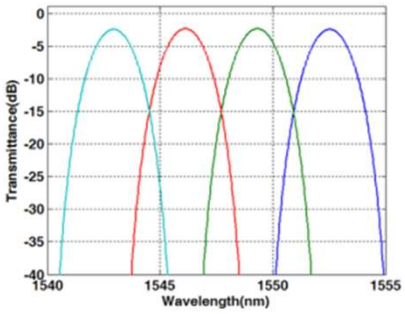




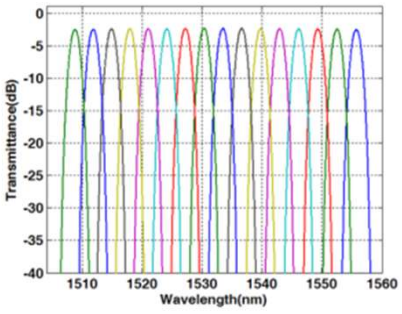
Interconnect Speeds and Modulation

Speeds and Feeds





4x25Gbps, channel spacing: 400GHz



16x25Gbps, channel spacing: 400GHz

Line Rate (Gb/s)

100

2x

"50"

2x

25G

100G Parallel

100G WDM

400G WDM

400G

4

80x

80

Wavelength Channel

NRZ

4-8x

SMF/MMF

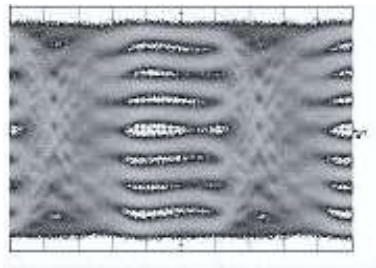
4x

PAM16

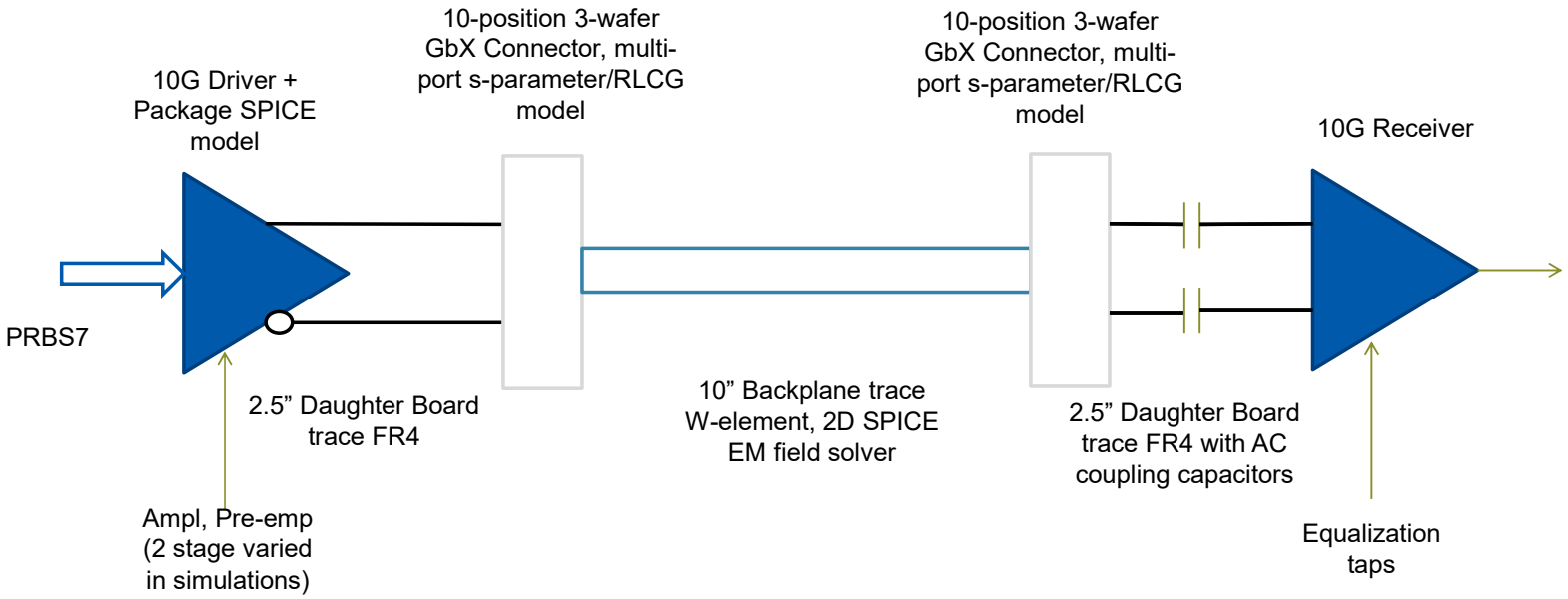
32

Physical Channel

Complex Modulations



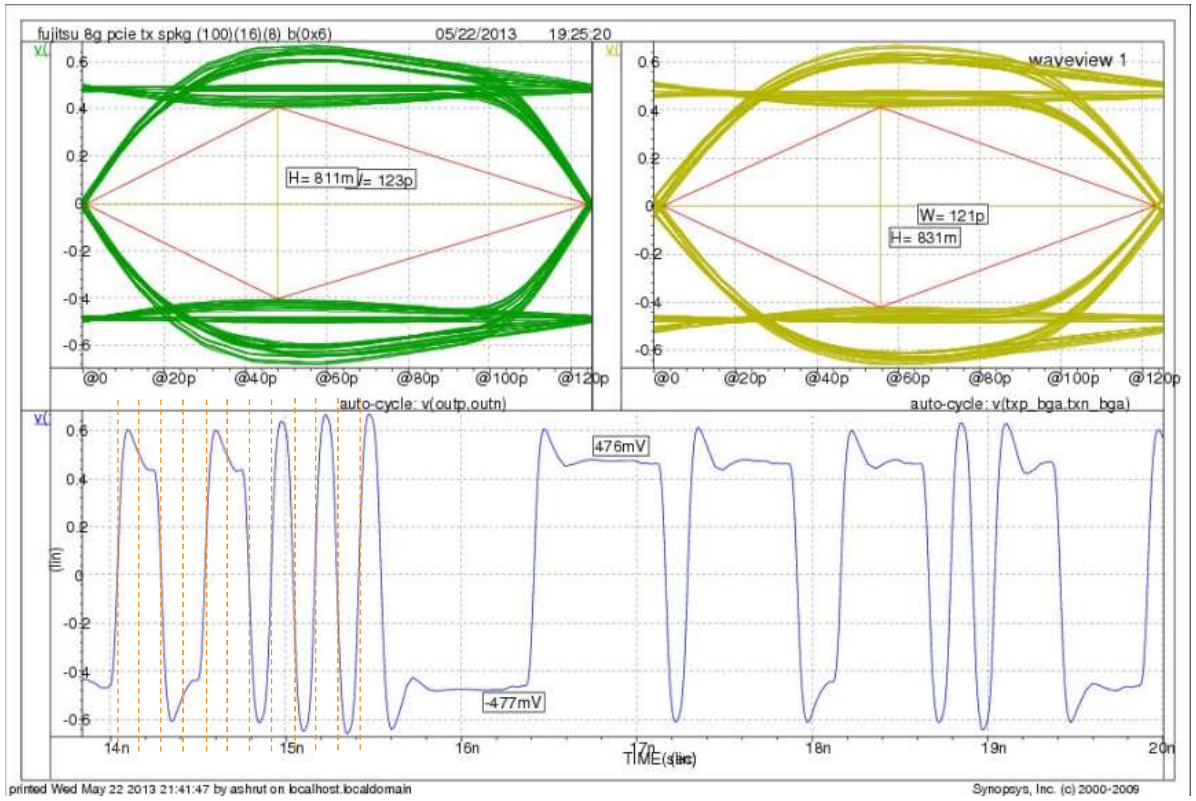
An Example High-Speed Channel



The Eye Diagram

- For high-speed signaling, the most fundamental unit of measure or visual-aid is the eye-diagram
- Nothing but a representation of 0's and 1's in a "folded" time space.
- X-axis represents 1-bit period
- Y-axis represents waveform amplitude
- But this does not look very much like an eye….

PCIe Gen3 eye-diagram at o/p buffer (before and after package model)

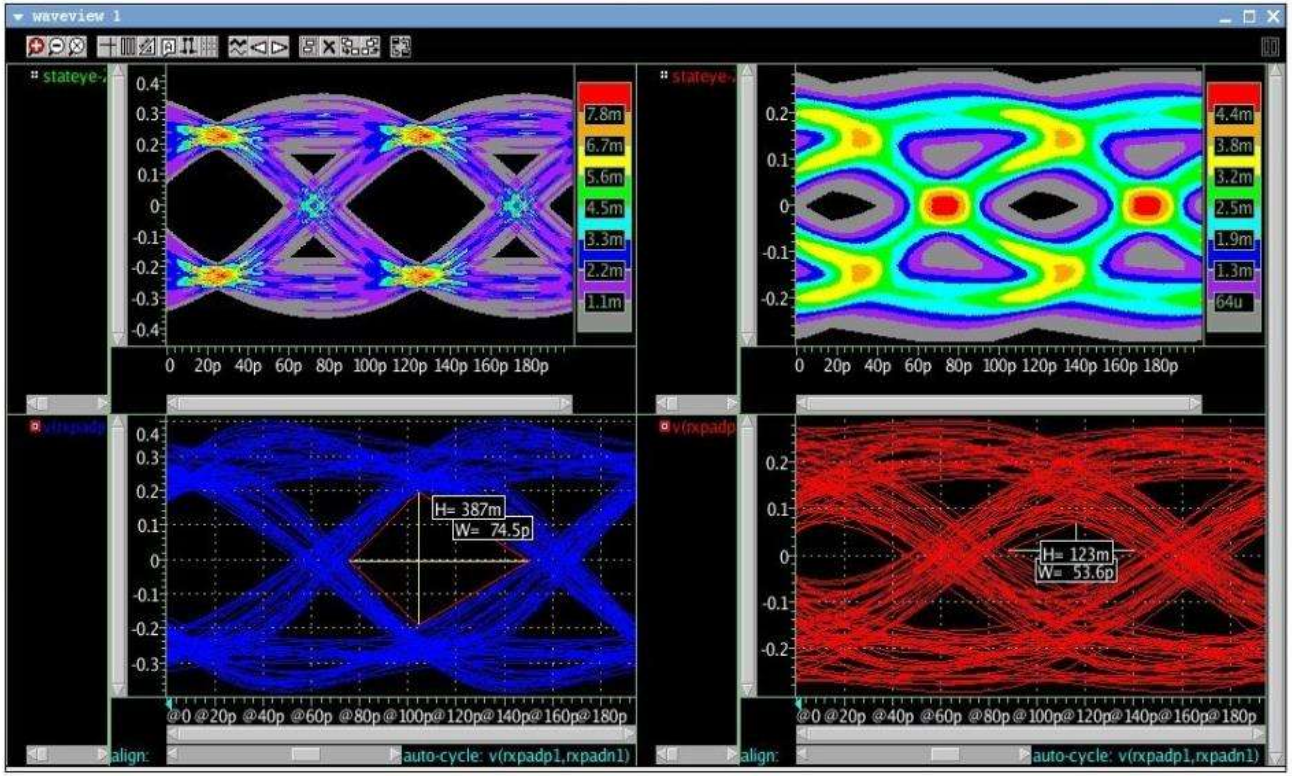


The Eye Diagram cont..

- Maybe now it resembles a bit more..
- Characteristics like amplitude, T_r/T_f, zero-crossing etc. of the transmitted bit pattern starts to change as the signal traverses the channel
- What receiver sees is a very degraded bit-pattern. Eye seems to be “closing”
- What causes this eye to close?

10G eye diagram (before and after “channel”)

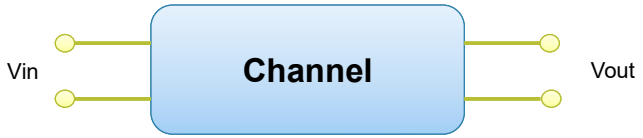
What was transmitted **What the Receiver sees**



Insertion Loss or Sdd21

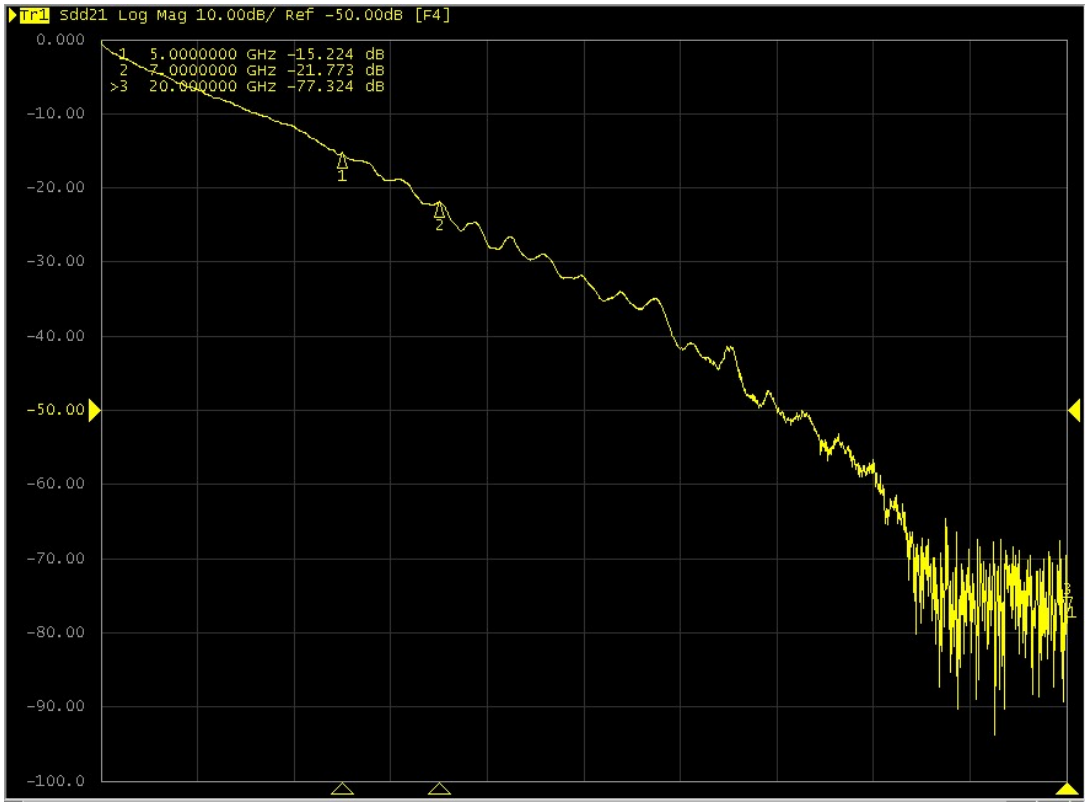
Without going into the incident/reflected power and scattering matrix of 2-port networks.

Insertion loss can be simply defined as $20*\log_{10}$ of “forward voltage attenuation”



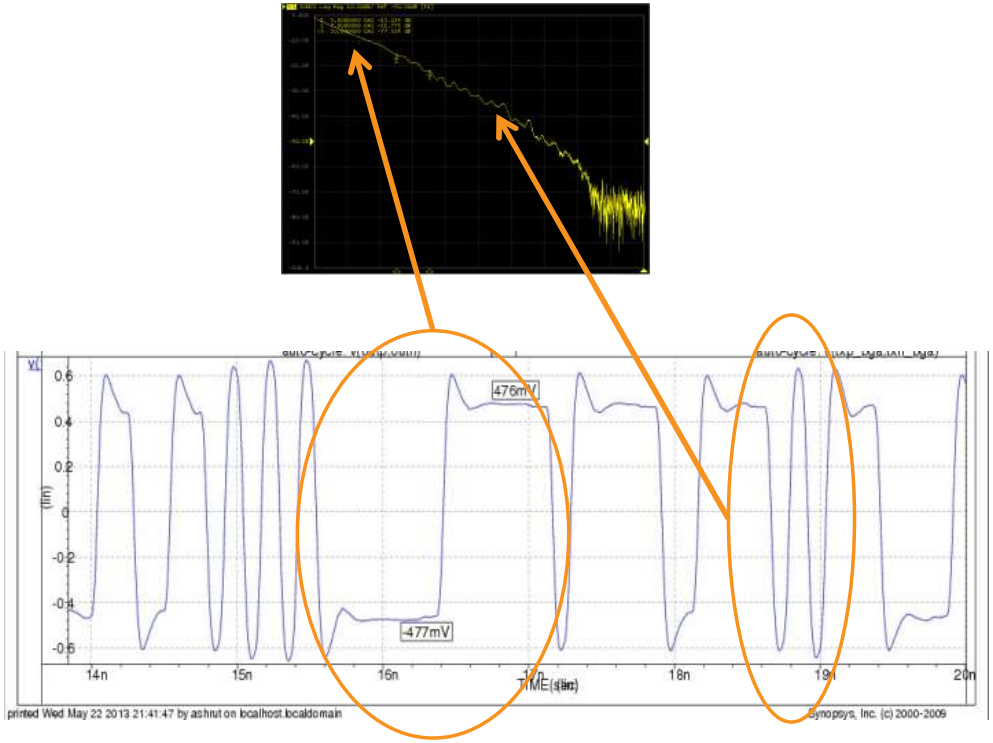
Freq sweep from DC-3rd harmonic

$$IL_{(dB)} = 20*\log(Vout/Vin)$$



Insertion Loss or Sdd21 cont..

- IL starts increasing as frequency increases



Will face lesser attenuation

Will be attenuated more

- IL is simply related to copper trace length and channel dielectric material properties
- Longer the channel (trace) length higher the loss

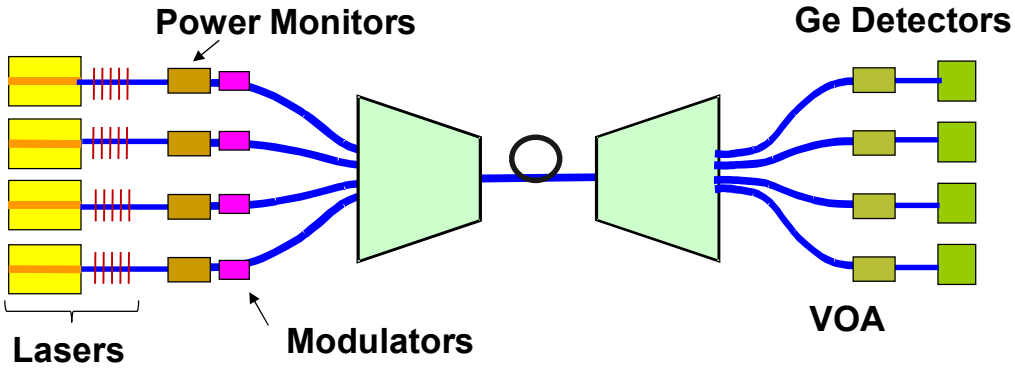
Cross-section view of a Transmission line

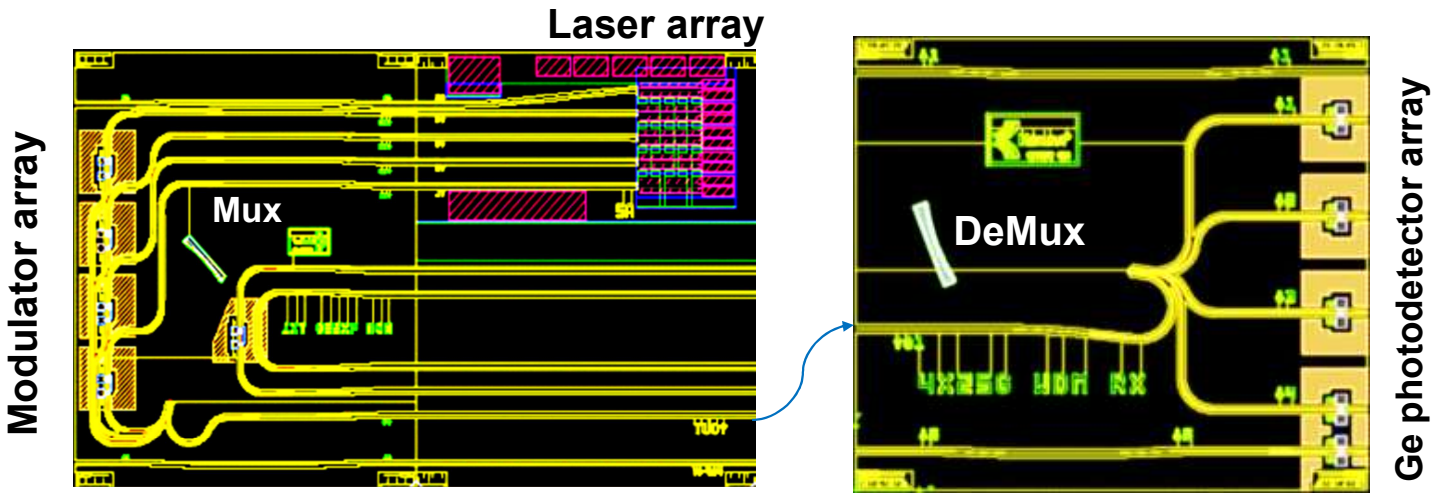
- Dielectric's ability to polarize/de-polarize causes freq dependence of Insertion Loss

Enter Optics.. VCSEL or Photonics

Electrical-Optical-Electrical Conversion

- Lasers
- Modulators
- Detectors





Network Layer and Routing

HDR InfiniBand Switch

QM8700, 1U Series

40 QSFP ports (PAM4, 50Gb/s per lane)

- 40 HDR (200Gb/s)
- 80 HDR100 (100Gb/s)

130nsec latency

390M messages per second (64Byte)

16Tb/s aggregated bandwidth

SHARPv2 low latency data reduction and streaming aggregation

22” depth

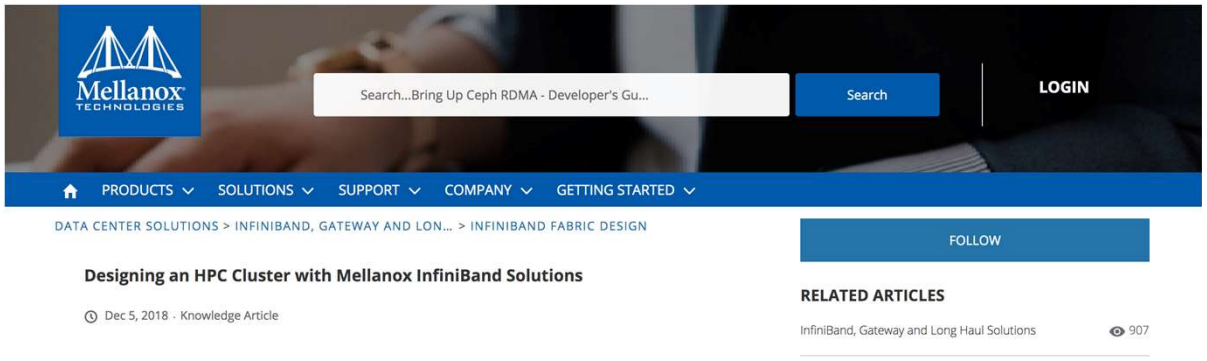
6 fans (5+1), hot swappable

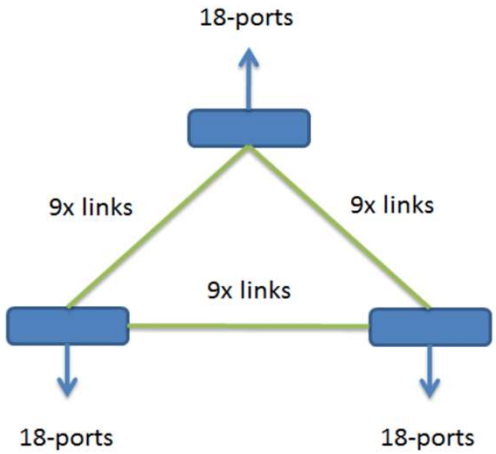
2 power supplies (1+1), hot swappable

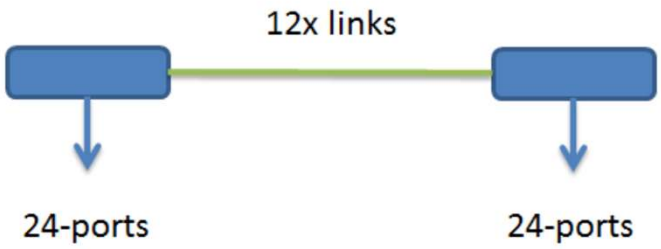
Mellanox Quantum

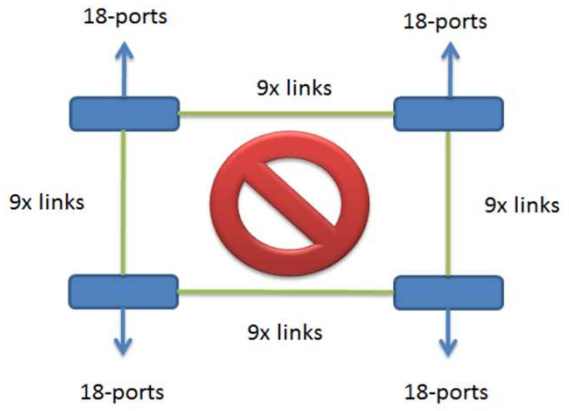
https://www.nvidia.com/en-in/networking/infiniband/qm8700/

Designing Small HPC/AI Clusters





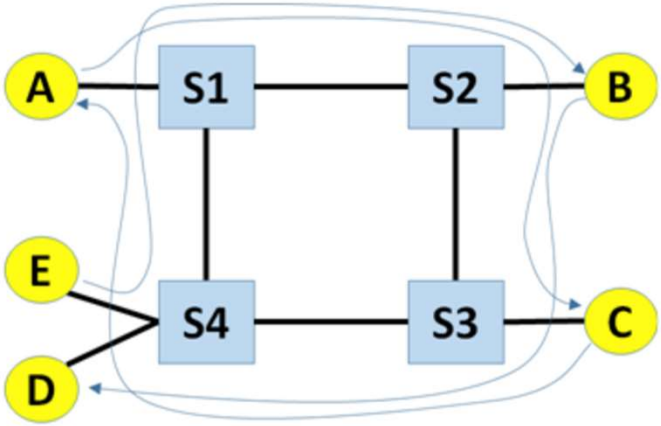




Credit Loops

Like many other networks, **InfiniBand dislikes loops**. Specifically, it dislikes **logical loops** where link back pressure can create a deadlock situation. These are called **credit loops**. Although the HOQ timers periodically clear such a deadlock, performance suffers. A credit loop only represents a potential deadlock, which depends on the traffic at each link in the loop. However, at InfiniBand speeds such a deadlock can occur very quickly given the right traffic pattern.

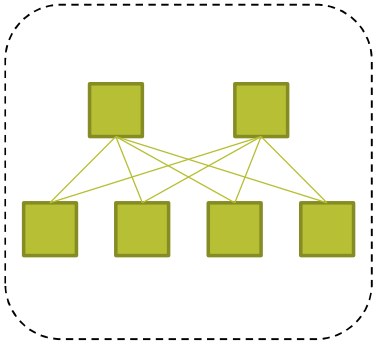
The following diagram illustrates a very simple credit loop.



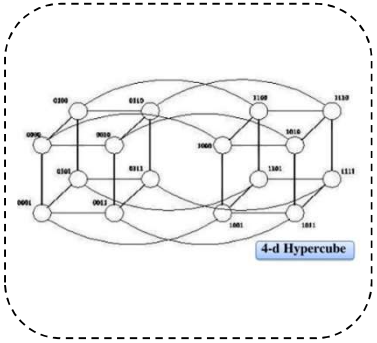
In this example:

- Four Switch ASICs are connected in a ring topology.
- A host adapter (HCA) connects a server (Node) to each Switch, with an additional HCA on Switch 4.

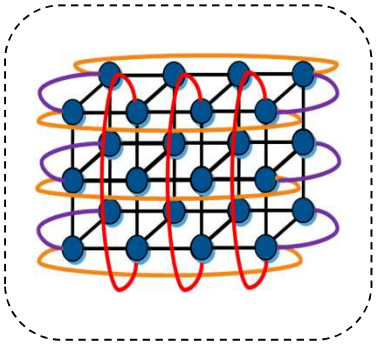
Network Topologies



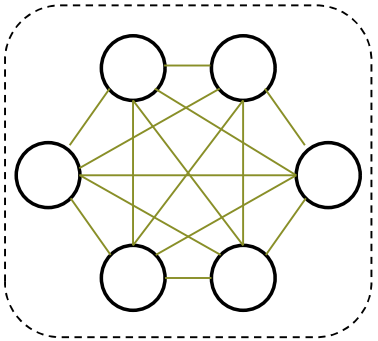
Fat Tree



Hypercube

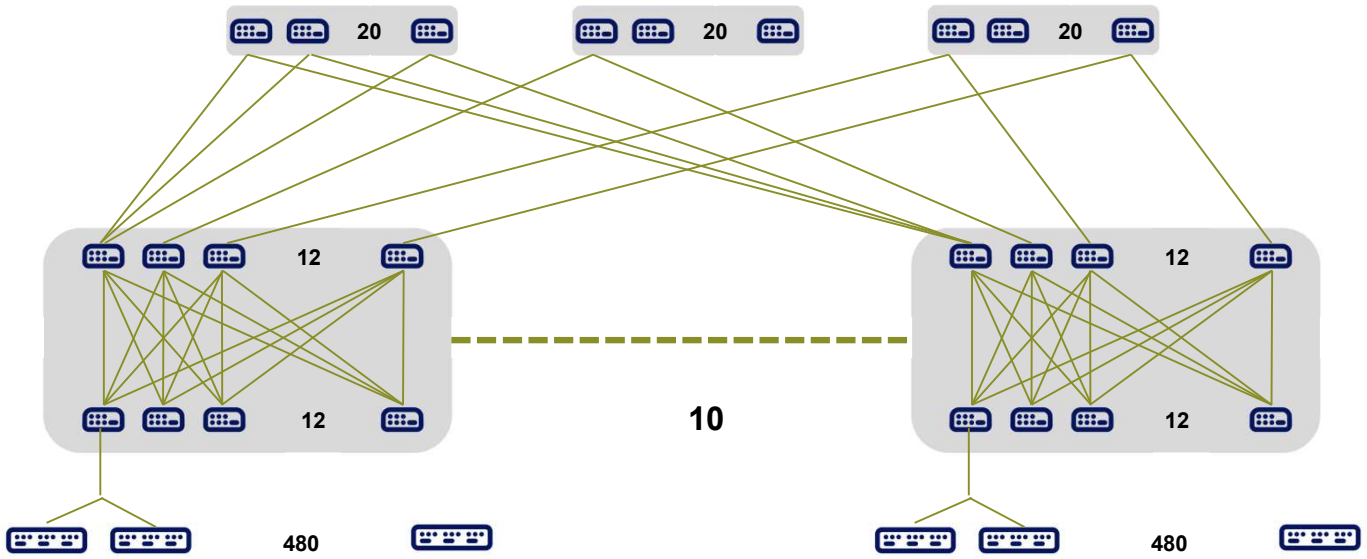


Torus



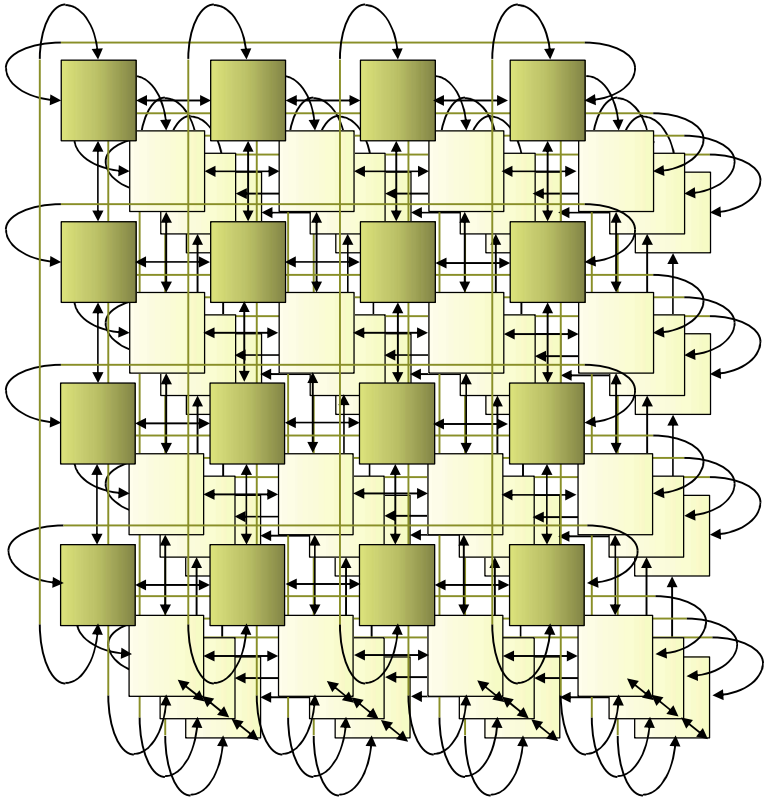
Dragonfly

3-level Clos with 1U HDR switches

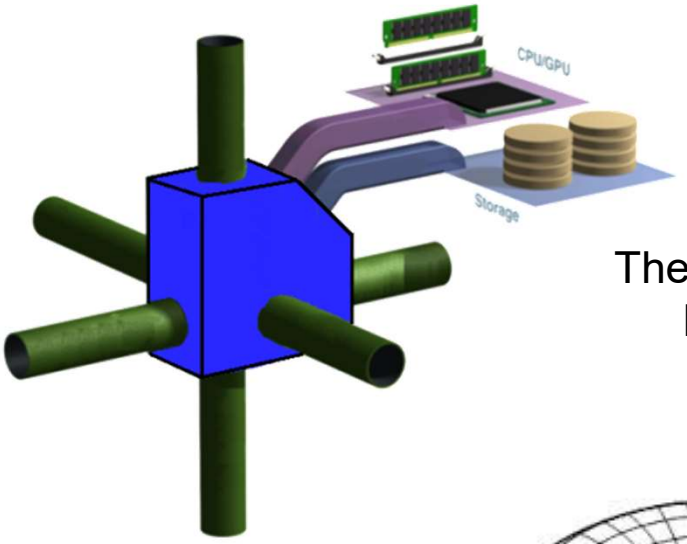


- **300 ASICs**
- 40 x 12 = 480 node sub-cluster (S1-S10)
- 10 Sub-clusters = 4800 port system
- Expansion requires re-design
- 5-hop network across any node-pair

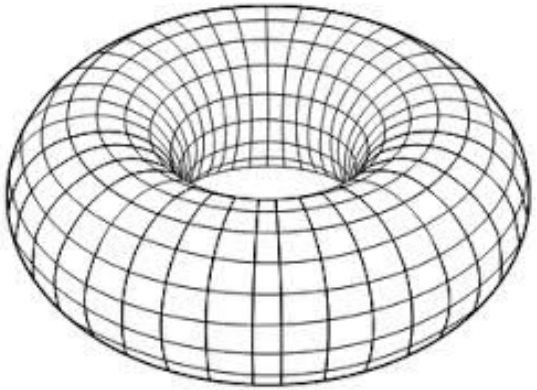
The 3-D Torus



A 4x4x4 Torus

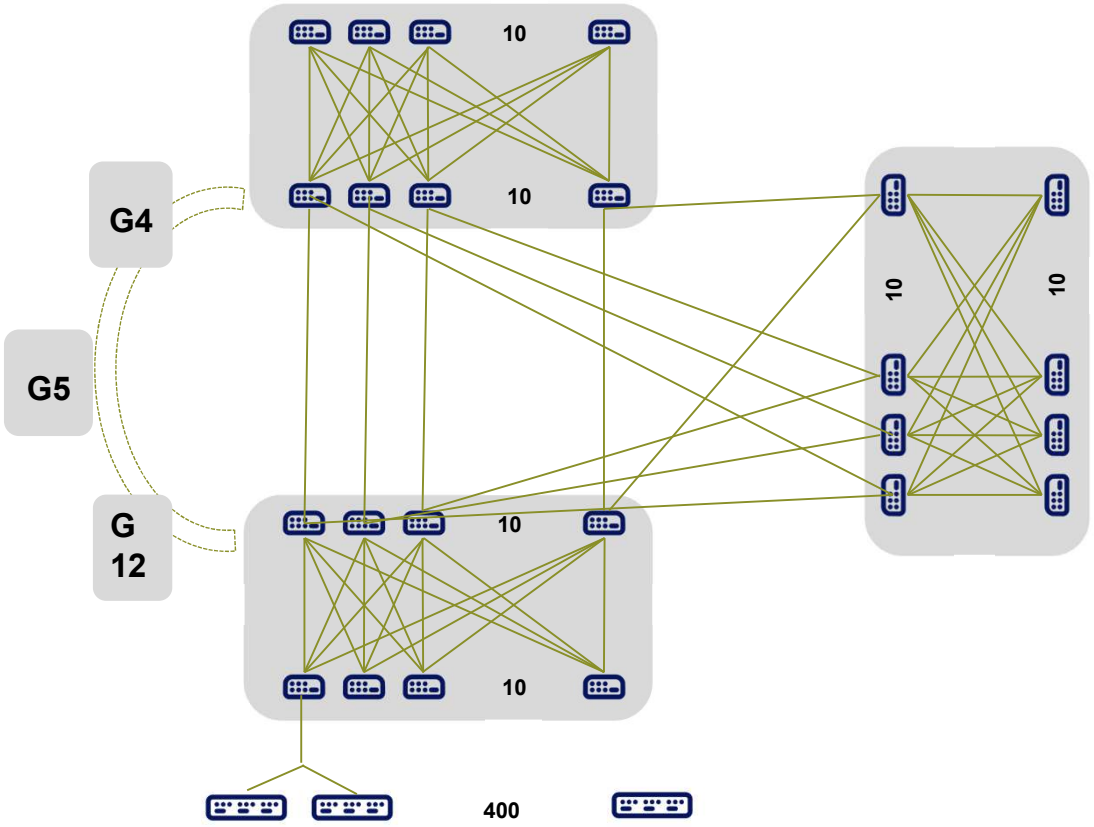


The Building Block



Natural “folding” to create a Torus

Dragonfy+ {240 ASICs}



- All-connect “Mesh” of “Groups”
- Each Group is full bi-partite
  - 20x 40p 200G switch per group
  - 200x 200G intra-group links per switch
  - 200x 200G inter-group links per switch
- 400 servers per group
- System can accommodate up-to 20 groups {8000 servers}
- Min-hop routing will have 4-hops across any node pair

Infiniband Packet Format

- Packets are routable end-to-end fabric unit of transfer
  - Link management packets: train and maintain link operation
  - Data packets
    - Send
    - Read
    - Write
    - Acks

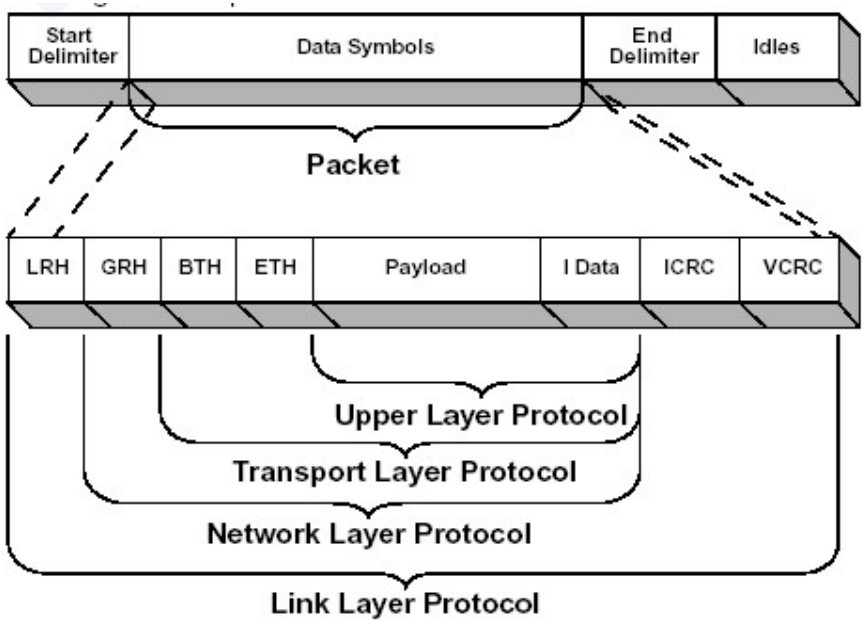
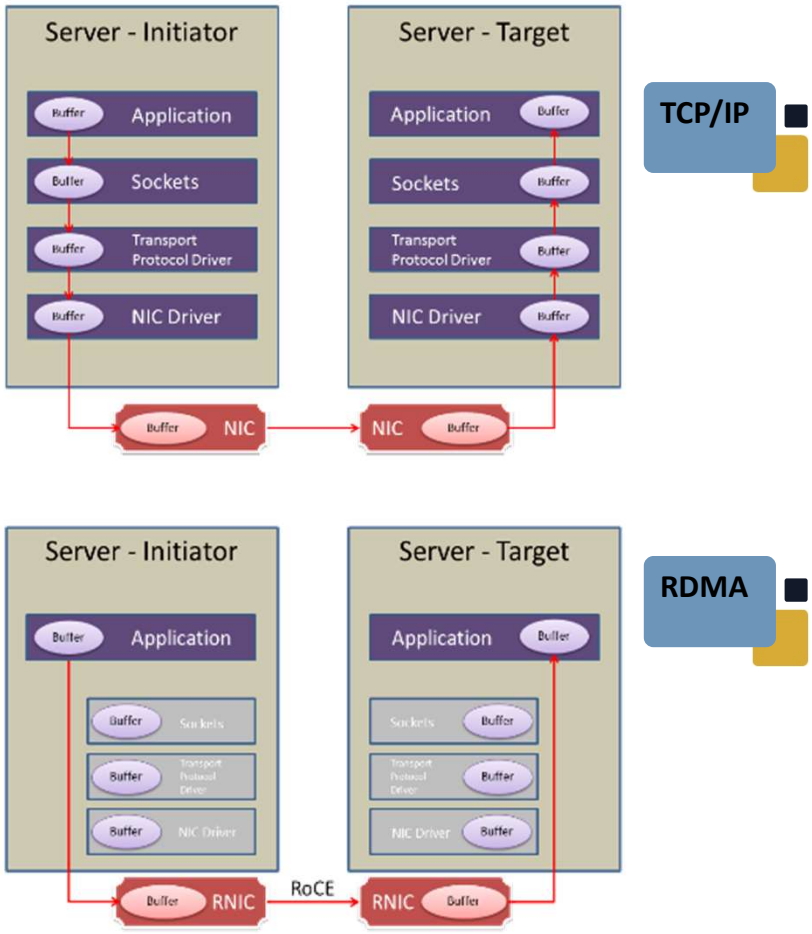


Figure 27 IBA Data Packet Format

Transport Layer and RDMA

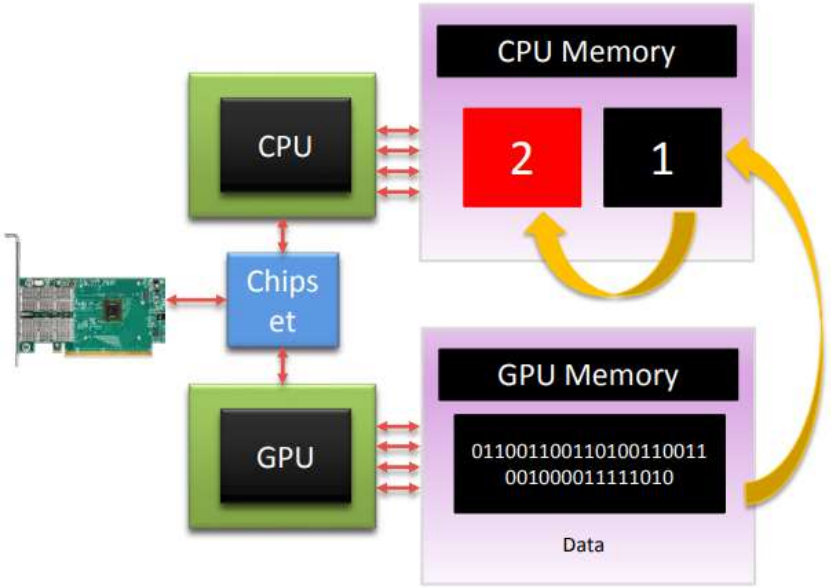
RDMA in SuperComputing Clusters

- Remote Direct Memory Access (RDMA)
- Advance transport protocol (same layer as TCP and UDP)
- Main features
  - Remote memory read/write semantics in addition to send/receive
  - Kernel bypass / direct user space access
  - Full hardware offload for network stack
  - Secure, channel based IO
- Application Advantage
  - Lowest latency
  - Highest bandwidth
  - Lowest CPU consumption
- Verbs: RDMA SW Interface (Equivalent to Sockets)



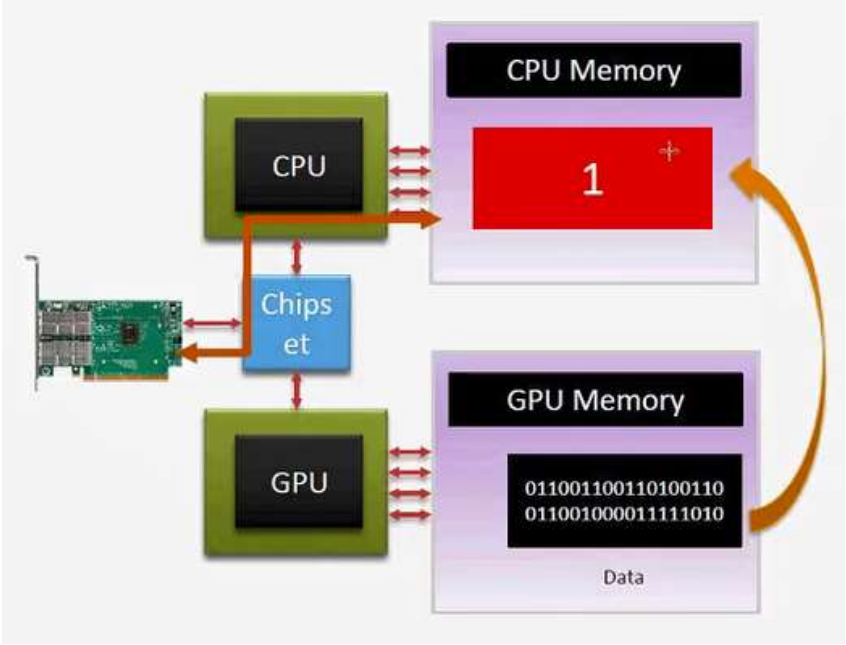
GPUDirect RDMA

- **Prior to GPUDirect**
  - GPU use driver-allocated pinned memory buffer
  - RDMA use pinned buffers for zero-copy kernel-bypass communication
- Impossible for RDMA drivers to pin memory allocated by GPU
- Two copies
  - GPU copies data from GPU internal memory to GPU driver system pinned memory (1)
  - User space needs to copy data between the GPU driver system pinned memory (1) and RDMA system pinned memory (2)
  - RDMA device sends data to network



GPUDirect RDMA Evolution (cont.)

- **GPUDirect / GPUDirect P2P** (Peer-to-Peer)
  - GPU and RDMA devices share the same pinned memory buffer
- One copy
  - GPU copies data from GPU internal memory to system pinned memory (1)
  - RDMA device sends data to network

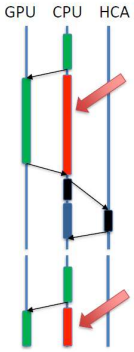


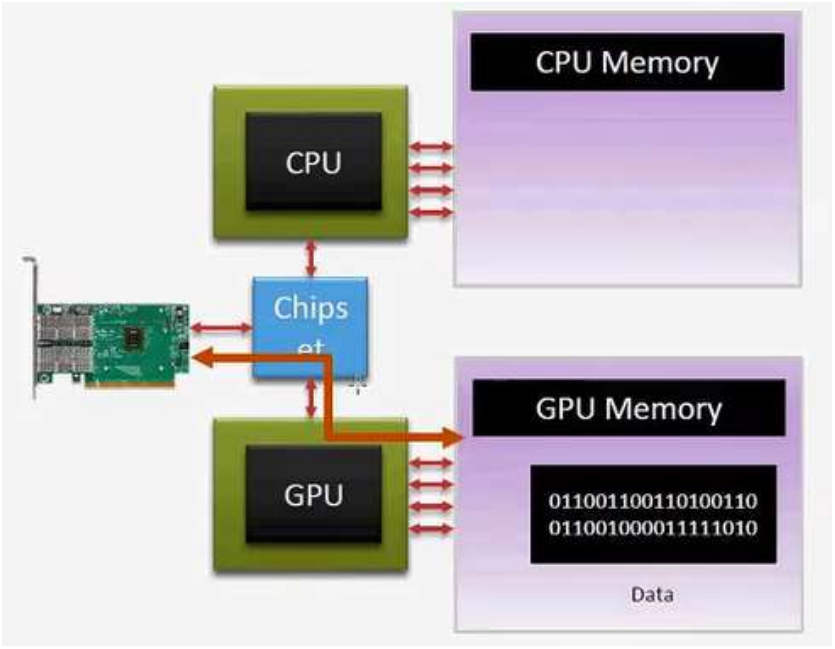
GPUDirect RDMA Evolution (cont.)

- **GPUDirect RDMA**
  - GPU memory is exposed to RDMA NIC
  - Direct data path from GPU to network - Data path is zero copy
  - **CPU** is involved in the control path - WQE preparation, ring doorbell, handles completions for incoming packets to GPU
- Zero copy
  - RDMA device sends data to network from GPU memory
  - RDMA device receive data from network to GPU memory
- The CPU still synchronizes between GPU tasks and data transfers

```
while(fin) {
    gpu_kernel <<<... , stream>>>(buf);
    cudaStreamSynchronize(stream);
    ibv_post_send(buf);
    ibv_poll_cq(cqe);
}
```

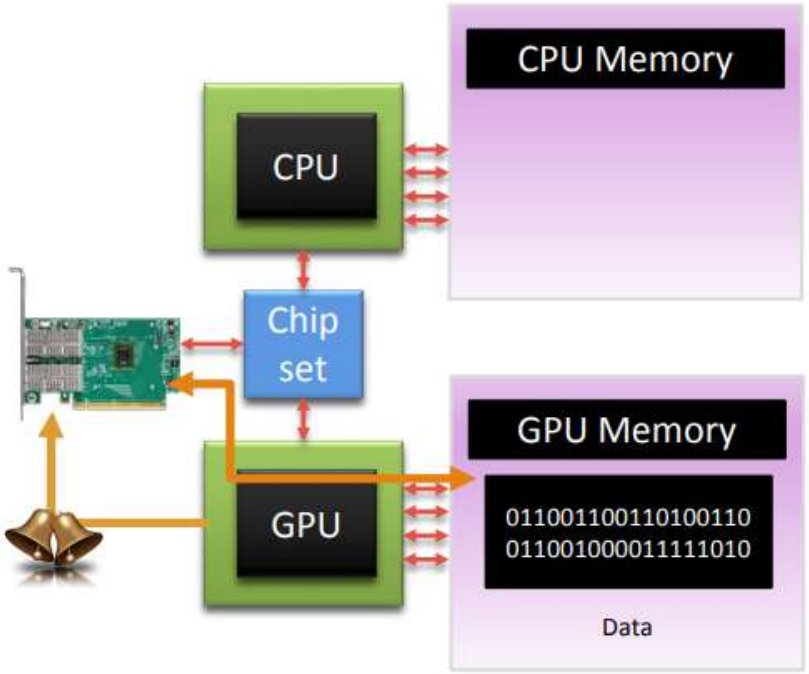
100% CPU Utilization



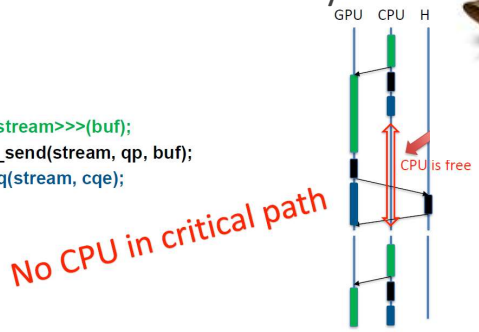


GPUDirect RDMA Evolution (cont.)

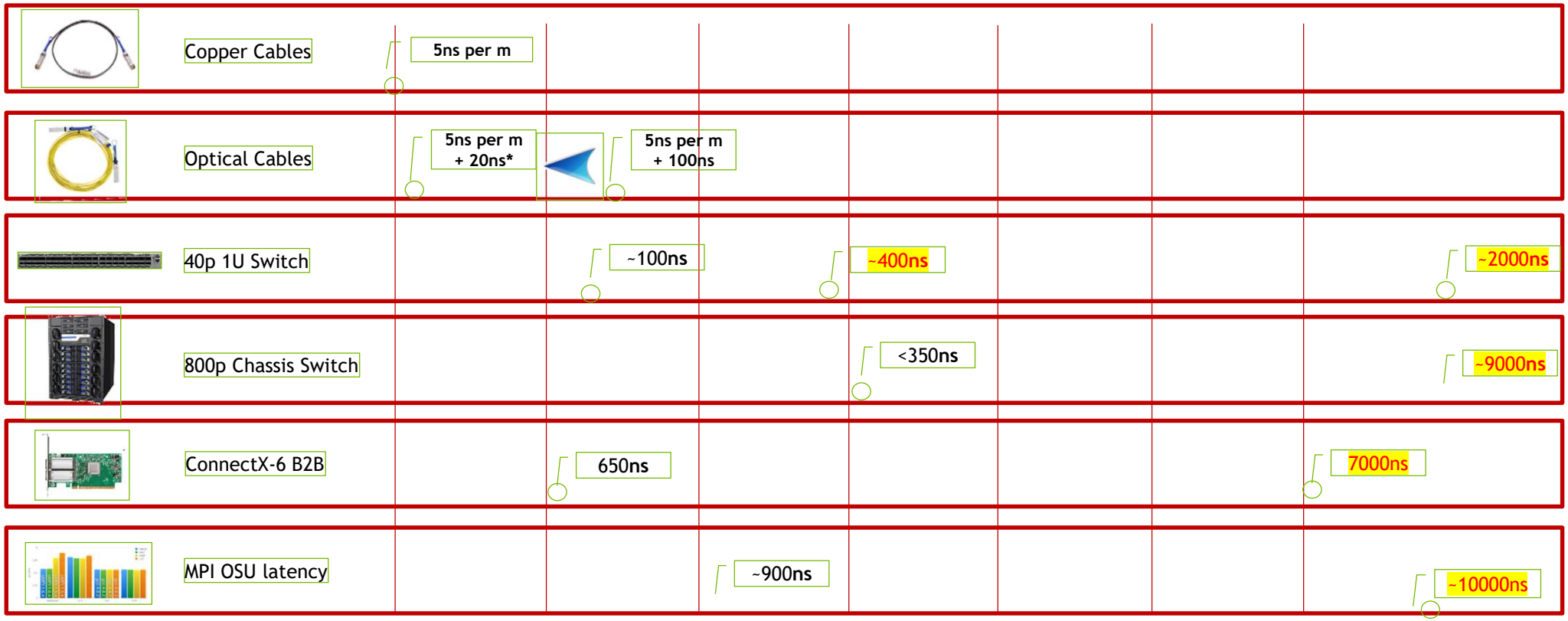
- **GPUDirect RDMA Async**
  - GPU memory is exposed to RDMA NIC
  - Direct data path from GPU to network - Data path is zero copy
  - **CPU** is involved in WQE preparation and release completed WQEs
  - **GPU** is involved in Ring Doorbell, Handles completions for incoming packets to GPU
- Zero copy
  - RDMA device sends data to network from GPU memory
  - RDMA device receive data from network to GPU memory
- Reduce CPU utilization




```
while(fin) {
    gpu_kernel <<<... , stream>>>(buf);
    gds_stream_queue_send(stream, qp, buf);
    gds_stream_wait_cq(stream, cqe);
}
```



Latency Contribution of Network Components



Need to Accelerate All Levels of HPC / AI Frameworks

Application

- Data Analysis
- Real Time
- Deep Learning

Communication

- 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

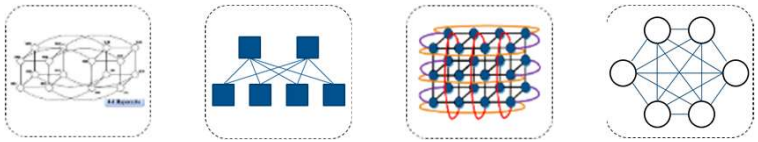
Network

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

RDMA GPUDirect · SHIELD

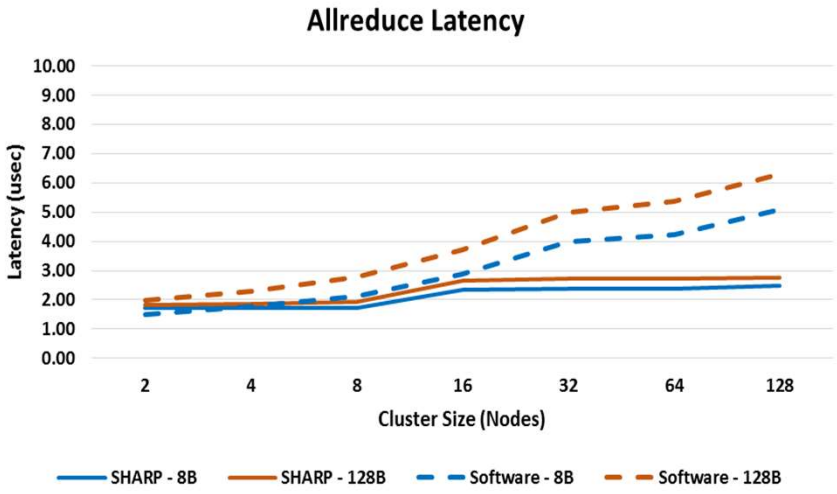
Connectivity

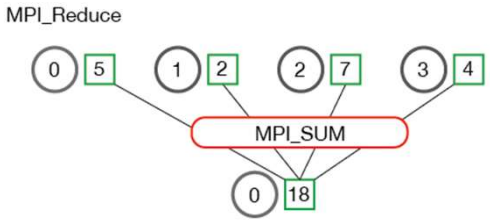
- Multi-Host Technology
- Socket-Direct Technology
- Enhanced Topologies

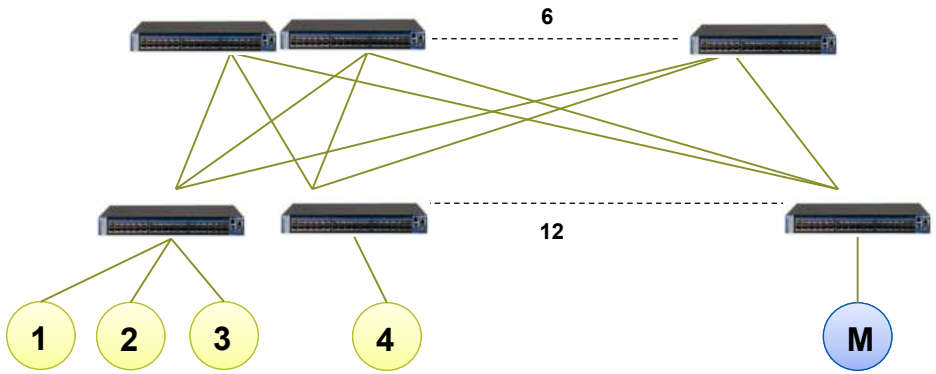


Advance Technologies and “Intelligent” Networks

Collectives like MPI all_reduce



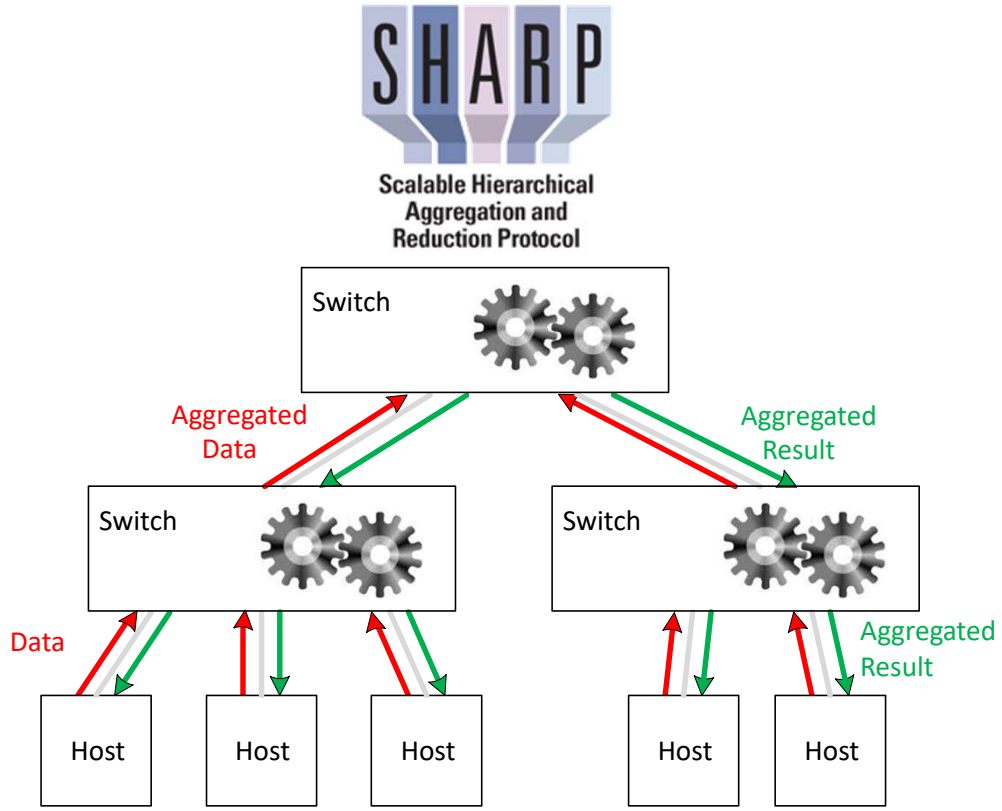




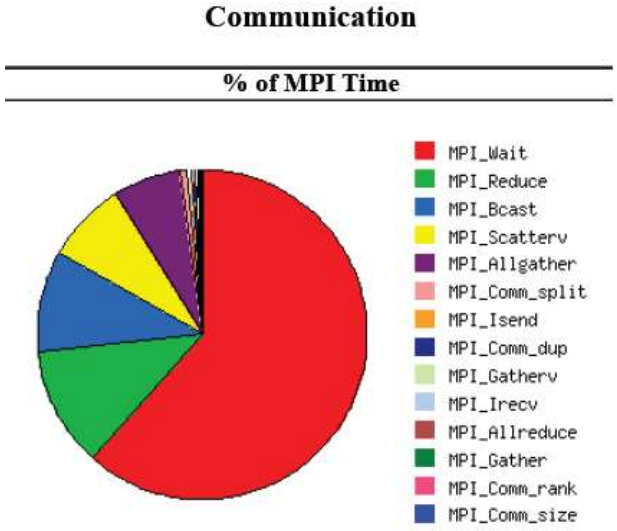
SHARP enables 75% Reduction in Latency
Providing Scalable Flat Latency

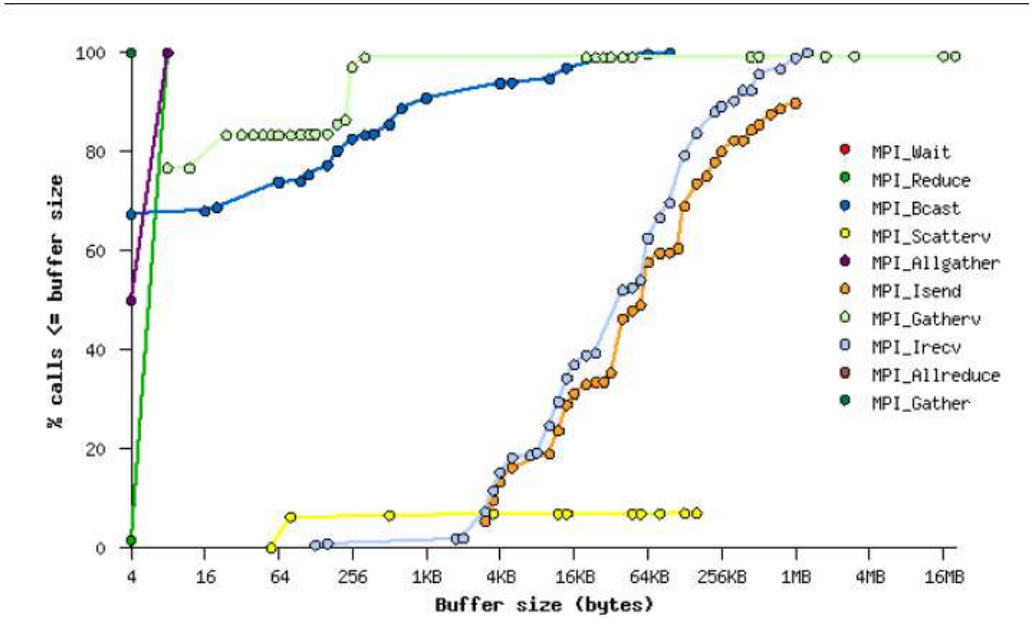
Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)

- 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



MPI profile of an application (eg. WRF)

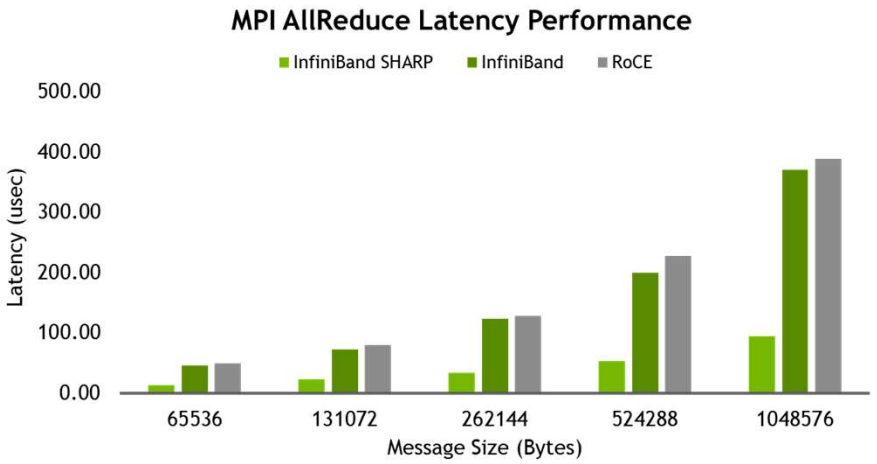


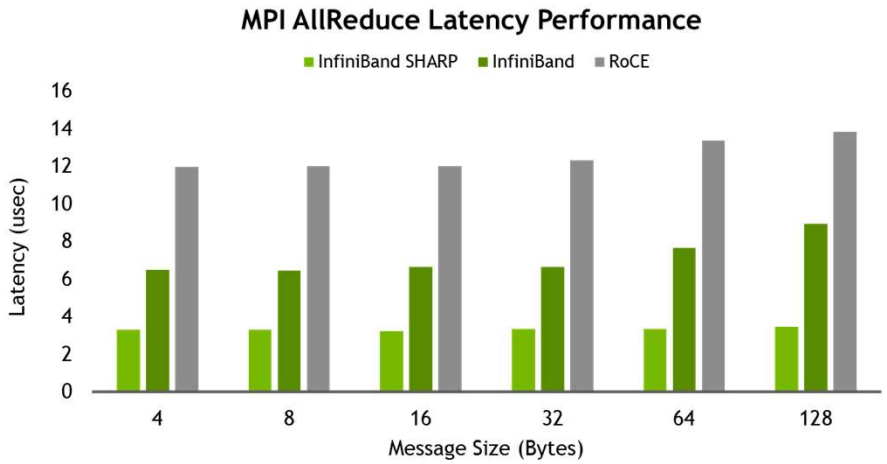


- Take MPI_Reduce for example
- In this application for eg; 4-8 byte MPI_Reduce has a significant component
- SHARP can reduce this latency from ~100us to <10us for a 512 node run

InfiniBand SHARP Performance Advantage

4X Higher Performance

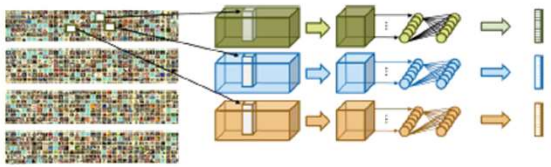


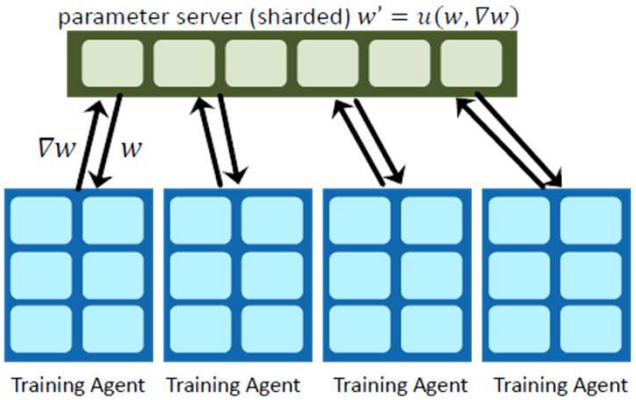


SHARP Accelerates AI Performance

The CPU in a parameter server becomes the bottleneck

SHARP Scalable Hierarchical Aggregation and Reduction Protocol



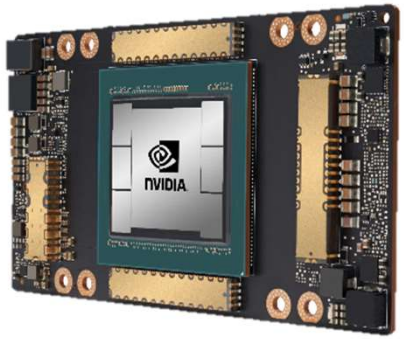


Performs the Gradient Averaging
Replaces all physical parameter servers
Accelerate AI Performance

Nvidia A100 GPU

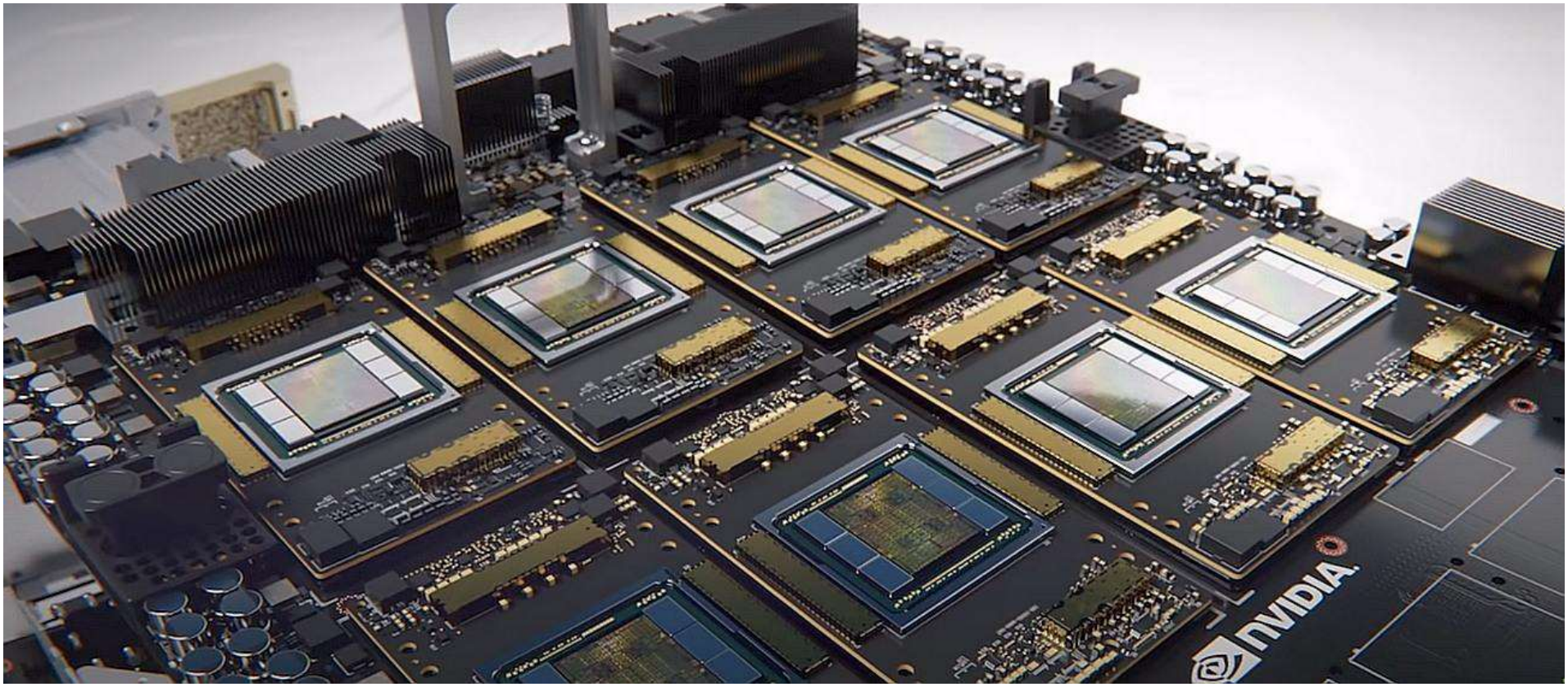
	A100 40GB PCIe	A100 80GB PCIe	A100 40GB SXM	A100 80GB SXM
FP64	9.7 TFLOPS			
FP64 Tensor Core	19.5 TFLOPS			
FP32	19.5 TFLOPS			
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*			
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*			
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*			
INT8 Tensor Core	624 TOPS | 1248 TOPS*			
GPU Memory	40GB HBM2	80GB HBM2e	40GB HBM2	80GB HBM2e
GPU Memory Bandwidth	1,555GB/s	1,935GB/s	1,555GB/s	2,039GB/s
Max Thermal Design Power (TDP)	250W	300W	400W	400W
Multi-Instance GPU	Up to 7 MIGs @ 5GB	Up to 7 MIGs @ 10GB	Up to 7 MIGs @ 5GB	Up to 7 MIGs @ 10GB
Form Factor	PCIe		SXM	
Interconnect	NVIDIA® NVLink® Bridge for 2 GPUs: 600GB/s ** PCIe Gen4: 64GB/s		NVLink: 600GB/s PCIe Gen4: 64GB/s	
Server Options	Partner and NVIDIA-Certified Systems™ with 1-8 GPUs		NVIDIA HGX™ A100-Partner and NVIDIA-Certified Systems with 4,8, or 16 GPUs NVIDIA DGX™ A100 with 8 GPUs	

A100 PCIe

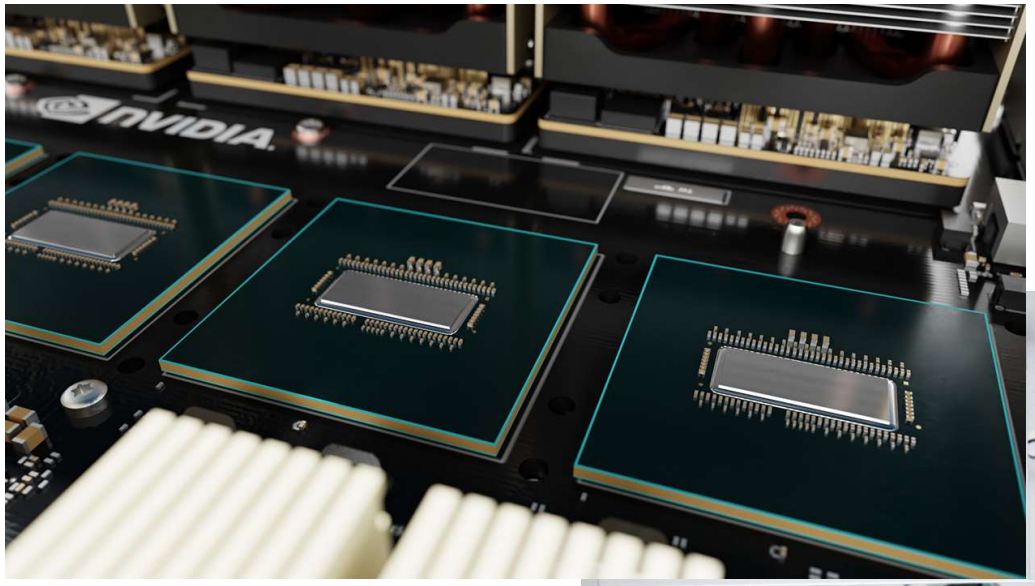


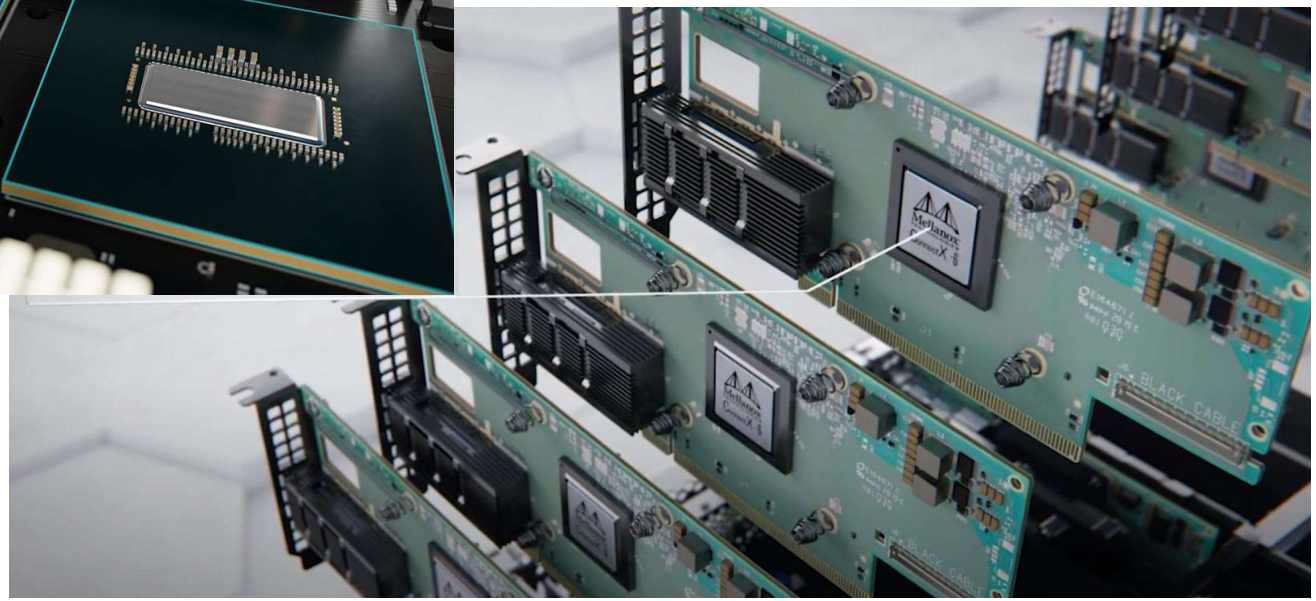
A100 SXM

8x A100 GPUs



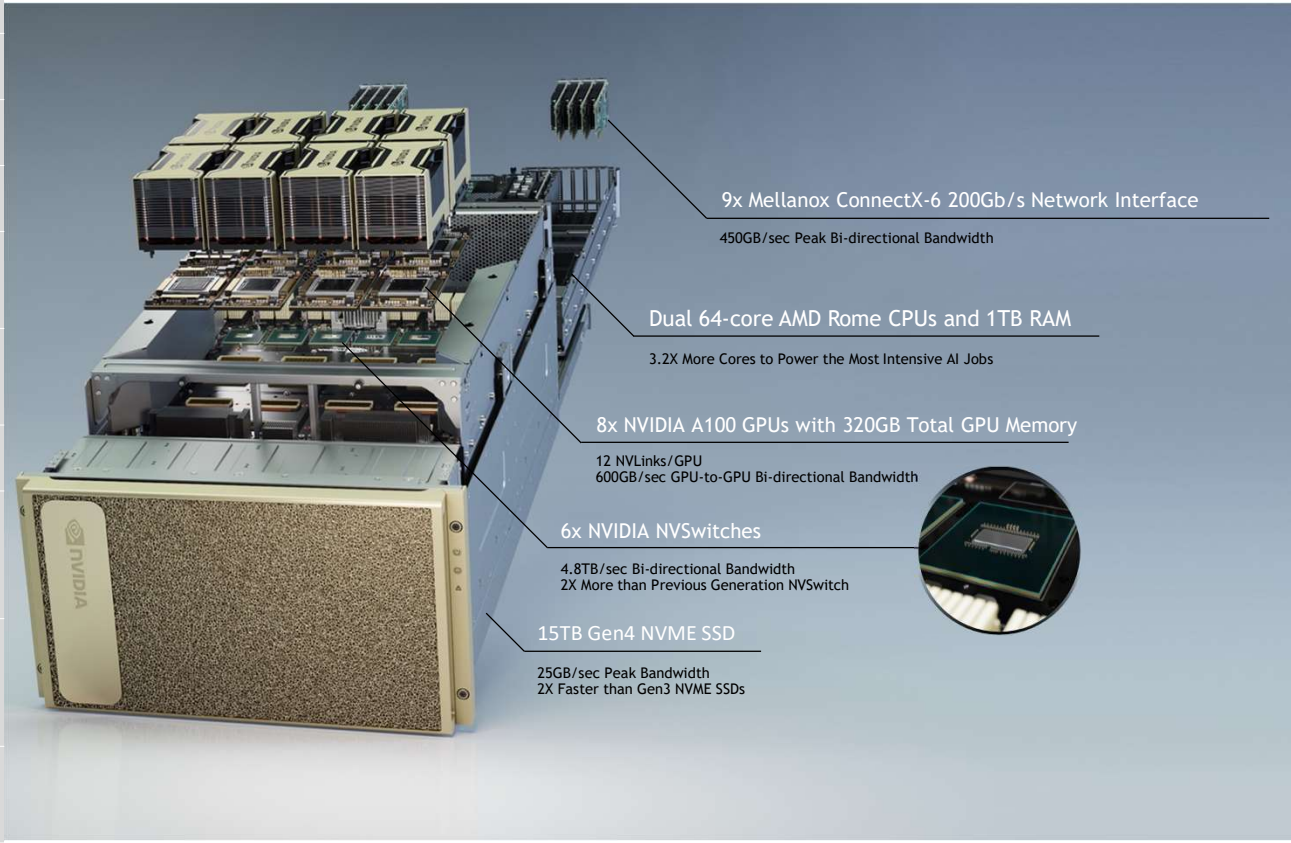
Internal and External Networking



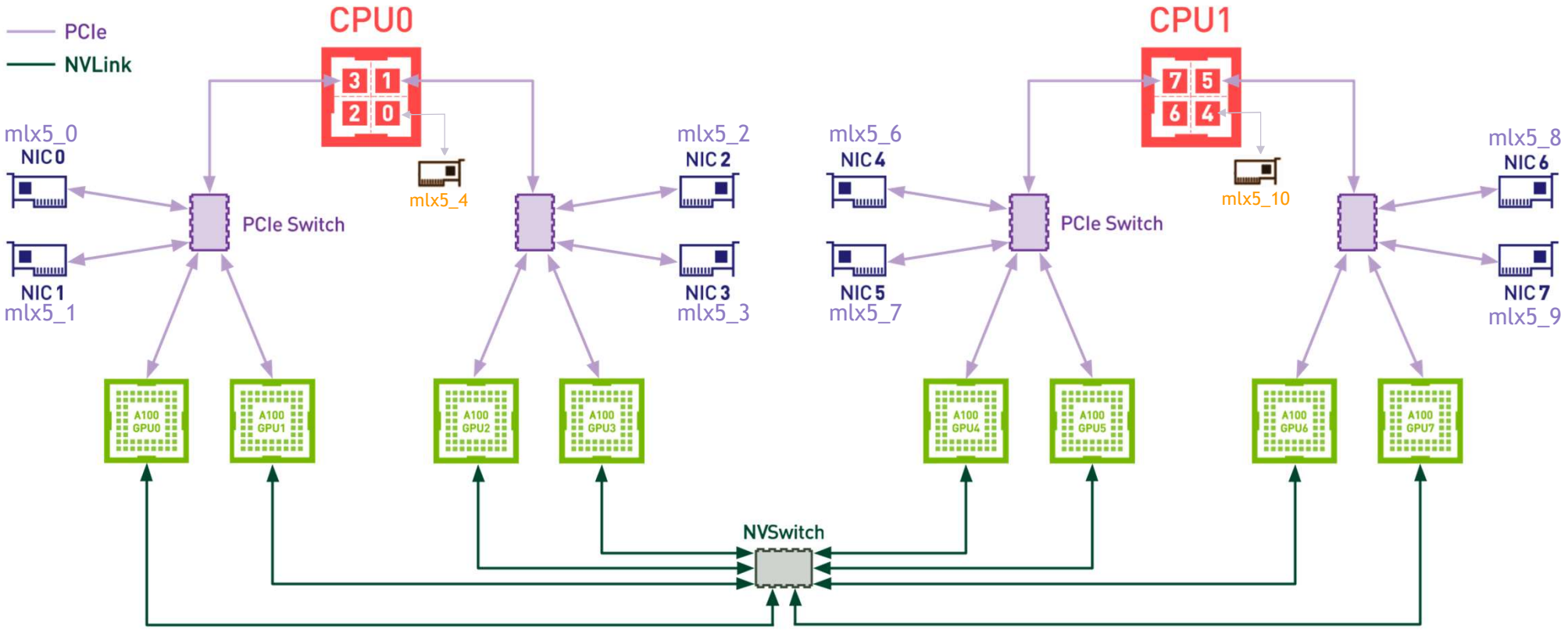


Game-changing performance for innovators

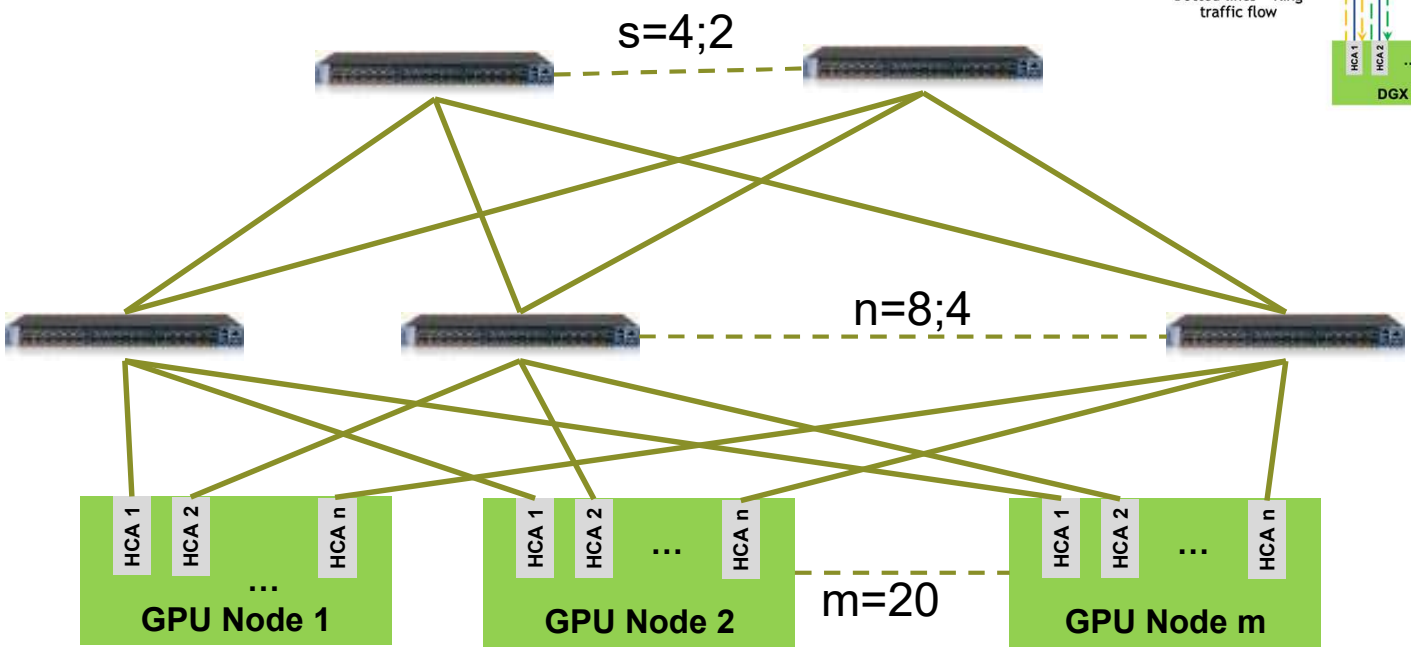
	App Focus Components
GPUs	8x NVIDIA A100 Tensor Core GPUs
GPU Memory	320GB Total
NVIDIA NVSwitch	6
Performance	5 petaFLOPS AI 10 petaOPS, INT8
CPU	Dual AMD Rome, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)
System Memory	1TB
Networking	9x Mellanox ConnectX-6 VPI HDR InfiniBand/200GigE 10th Dual-port ConnectX-6 optional
Storage	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives
System Power Usage	6.5 kW Max
System Weight	271 lbs (123 kgs)

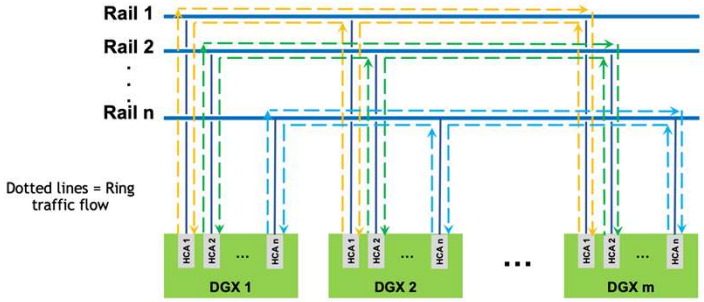


Typical GPU accelerated server - DGX A100

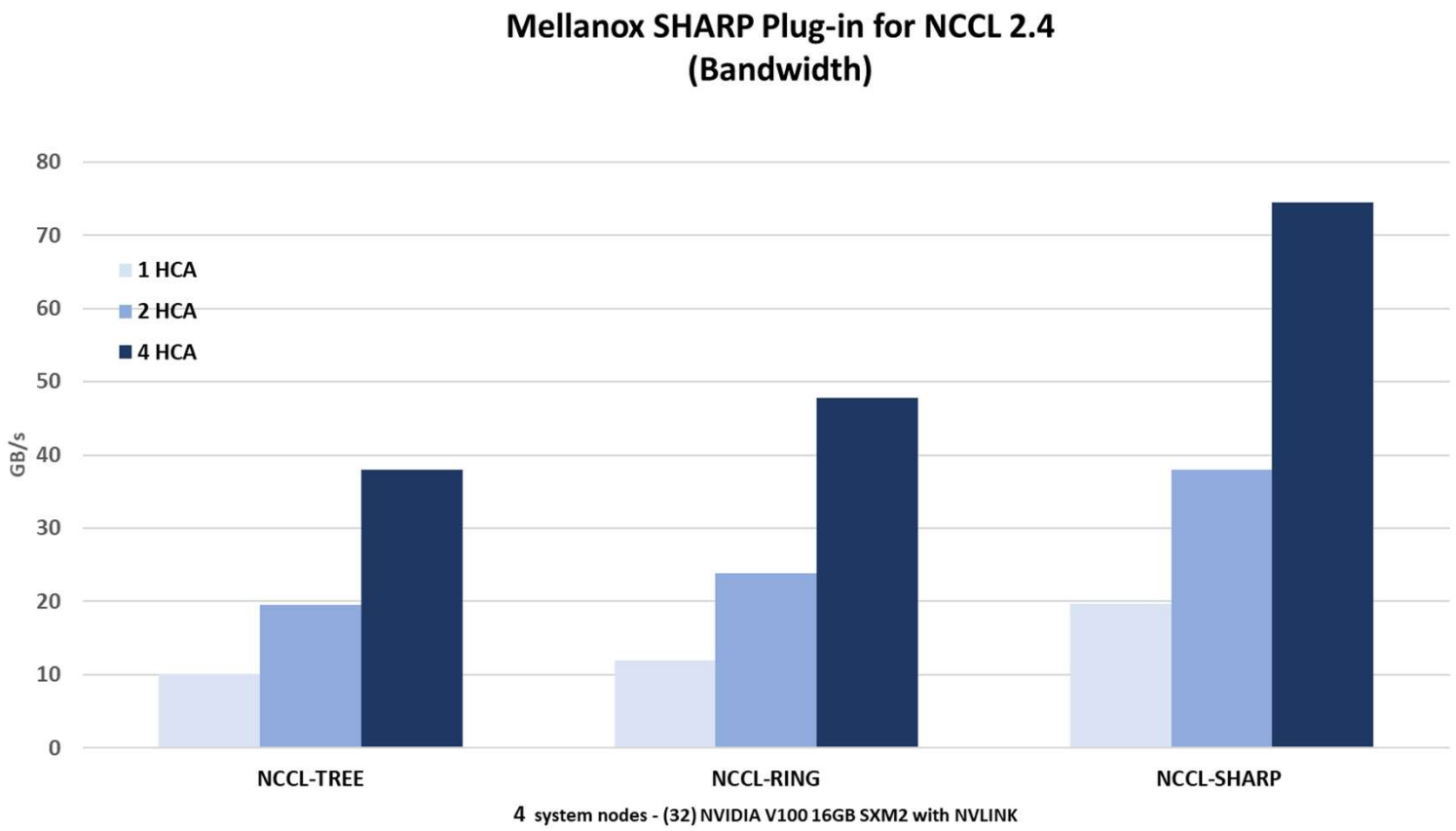


Mapping Rails To IB Trees

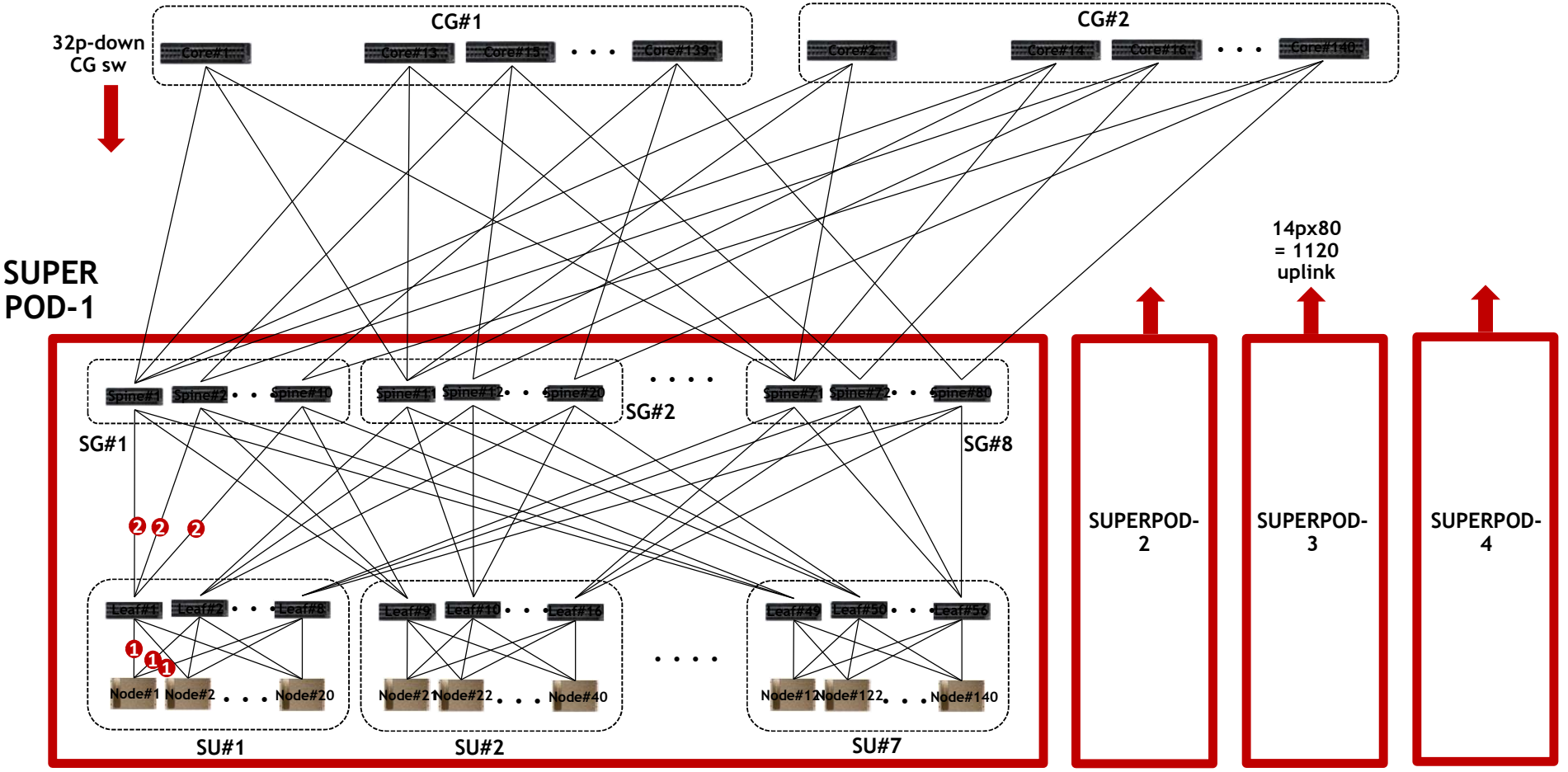




SHARP Delivers Highest Performance for AI

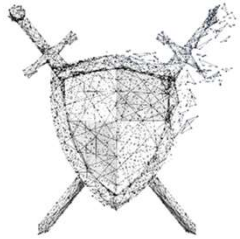


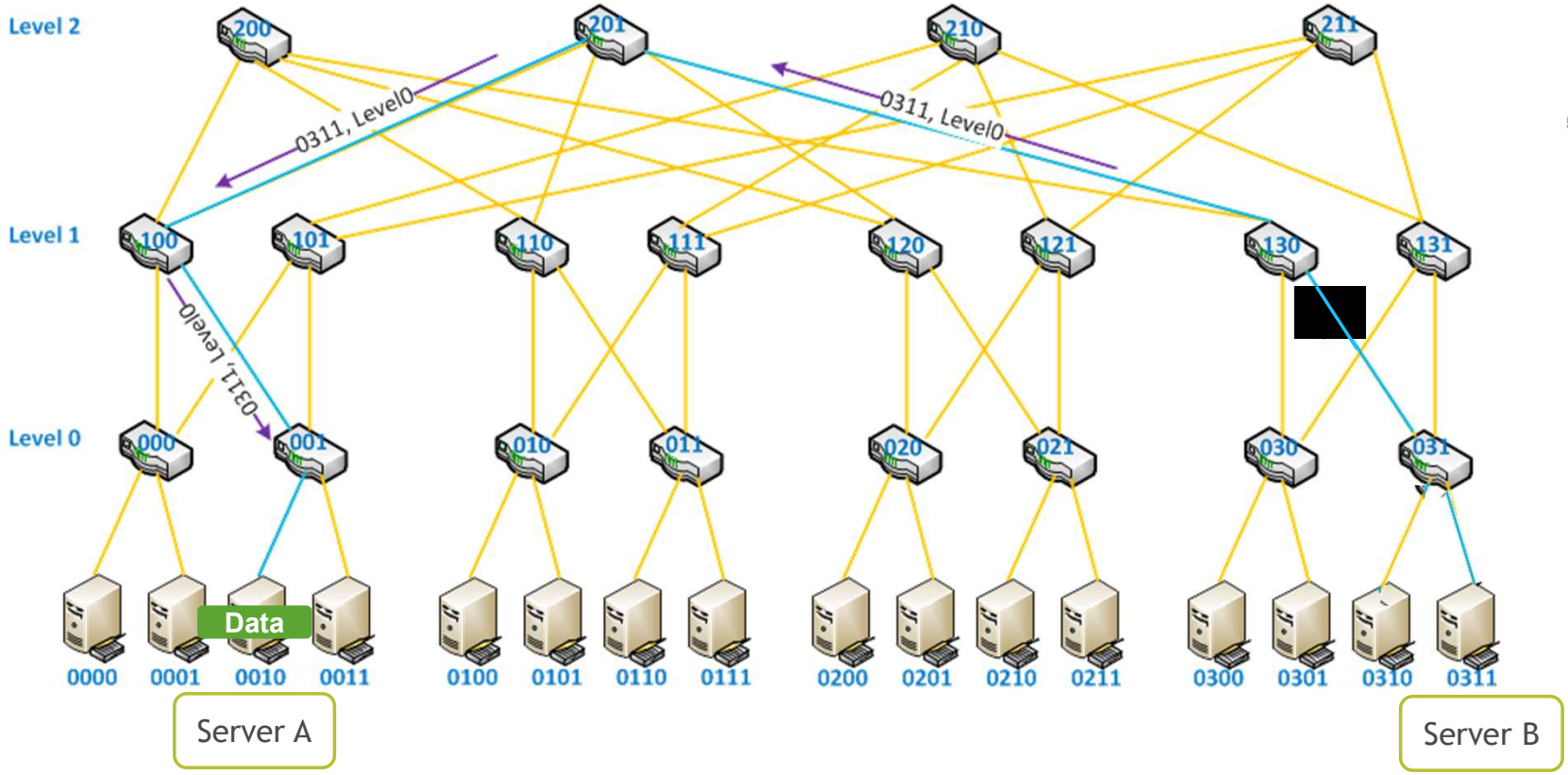
4x SuperPod network topology



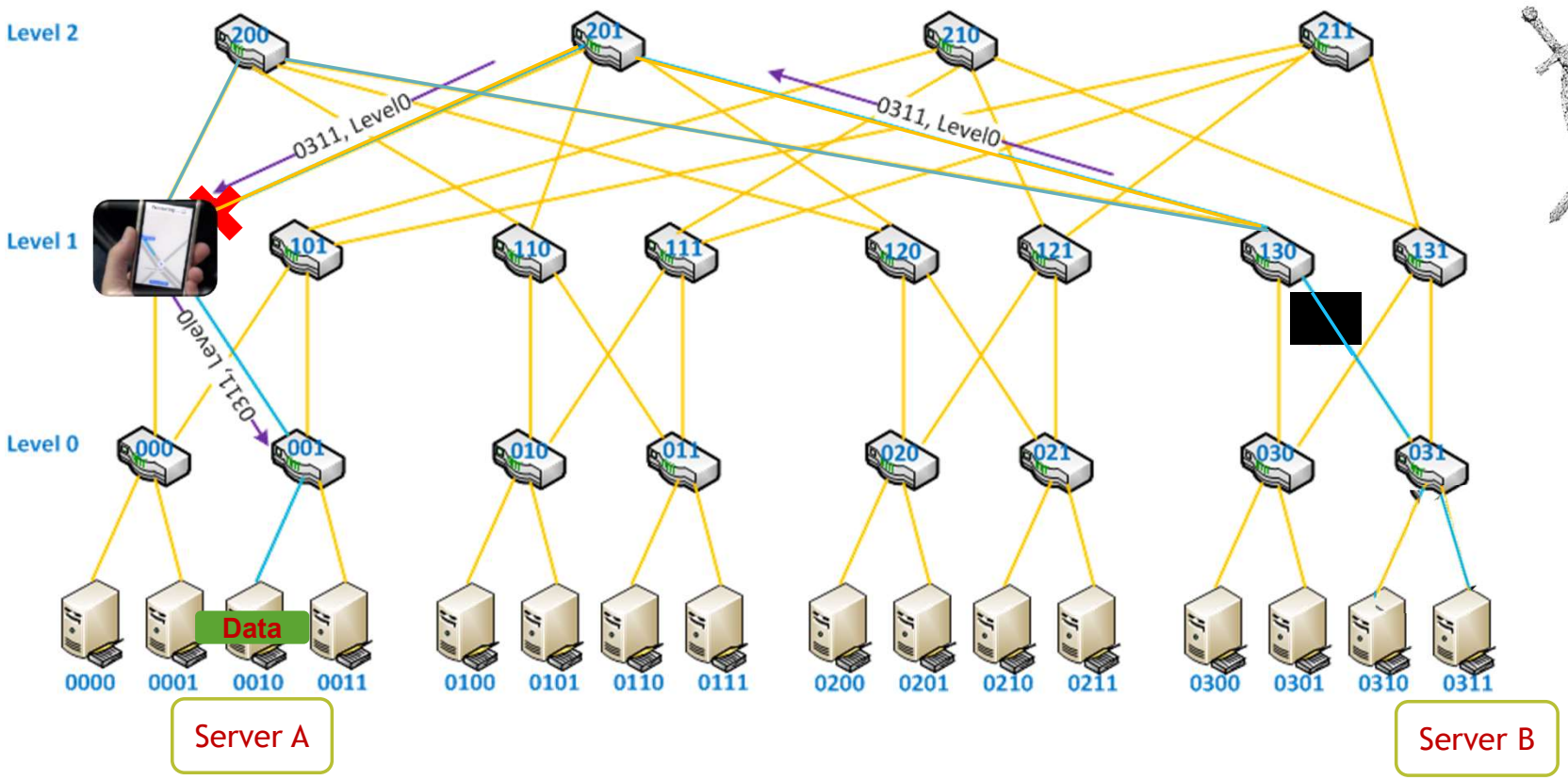
Resilient and “Self Healing” Networks

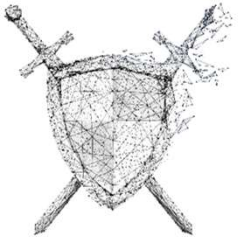
Example: Consider a Flow From A to B



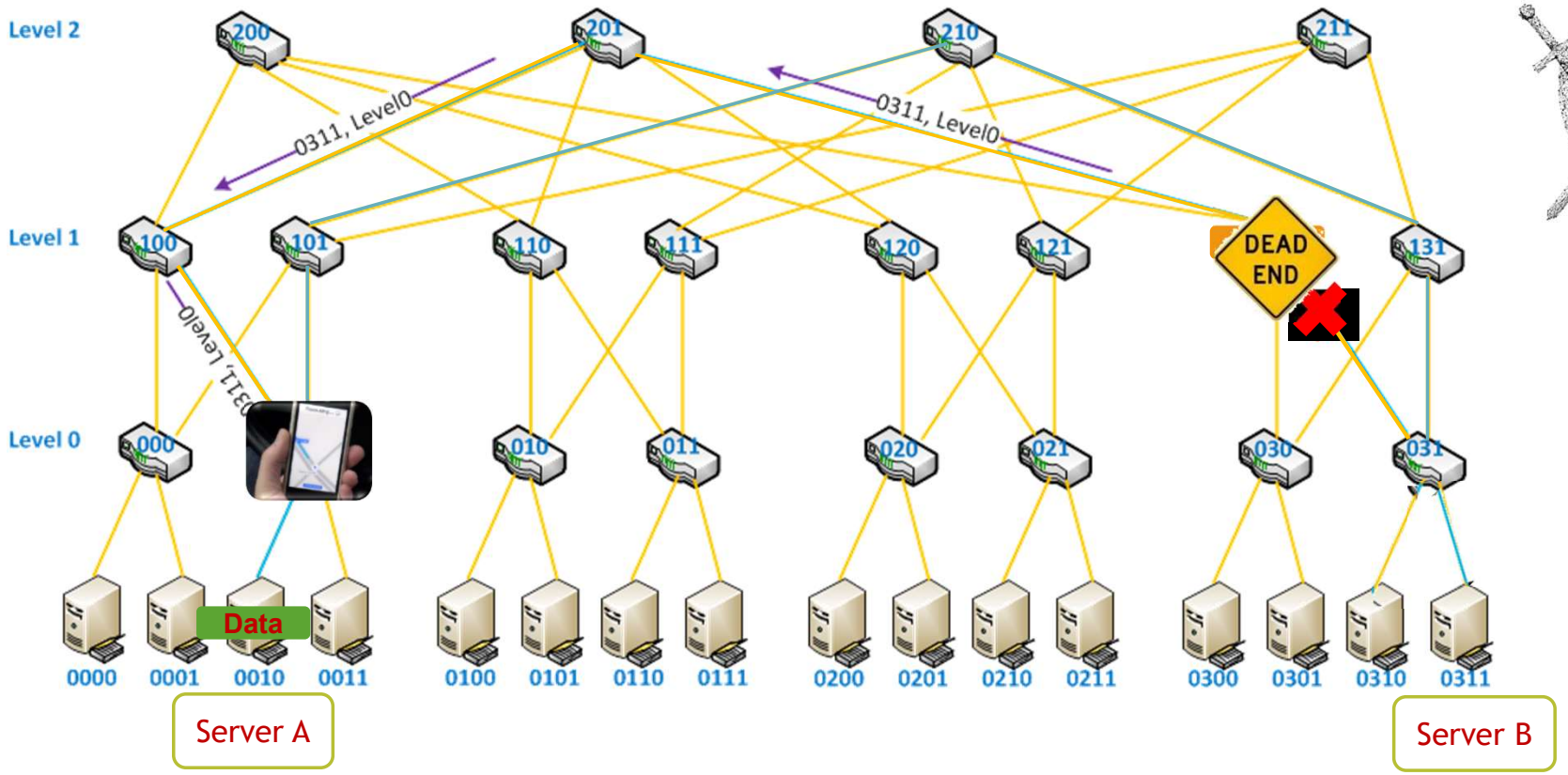


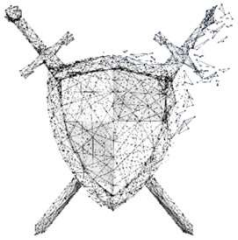
Example: The Simple Case: Local Fix



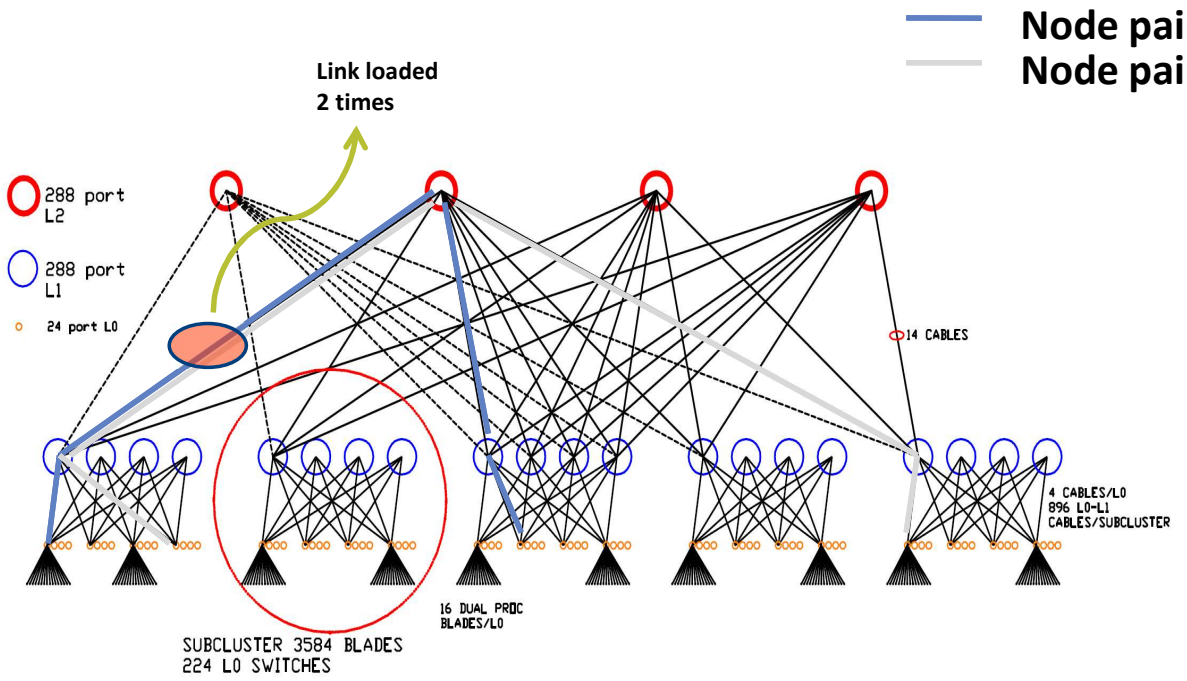


Example: The Remote Case - Using Fault Recovery Notifications



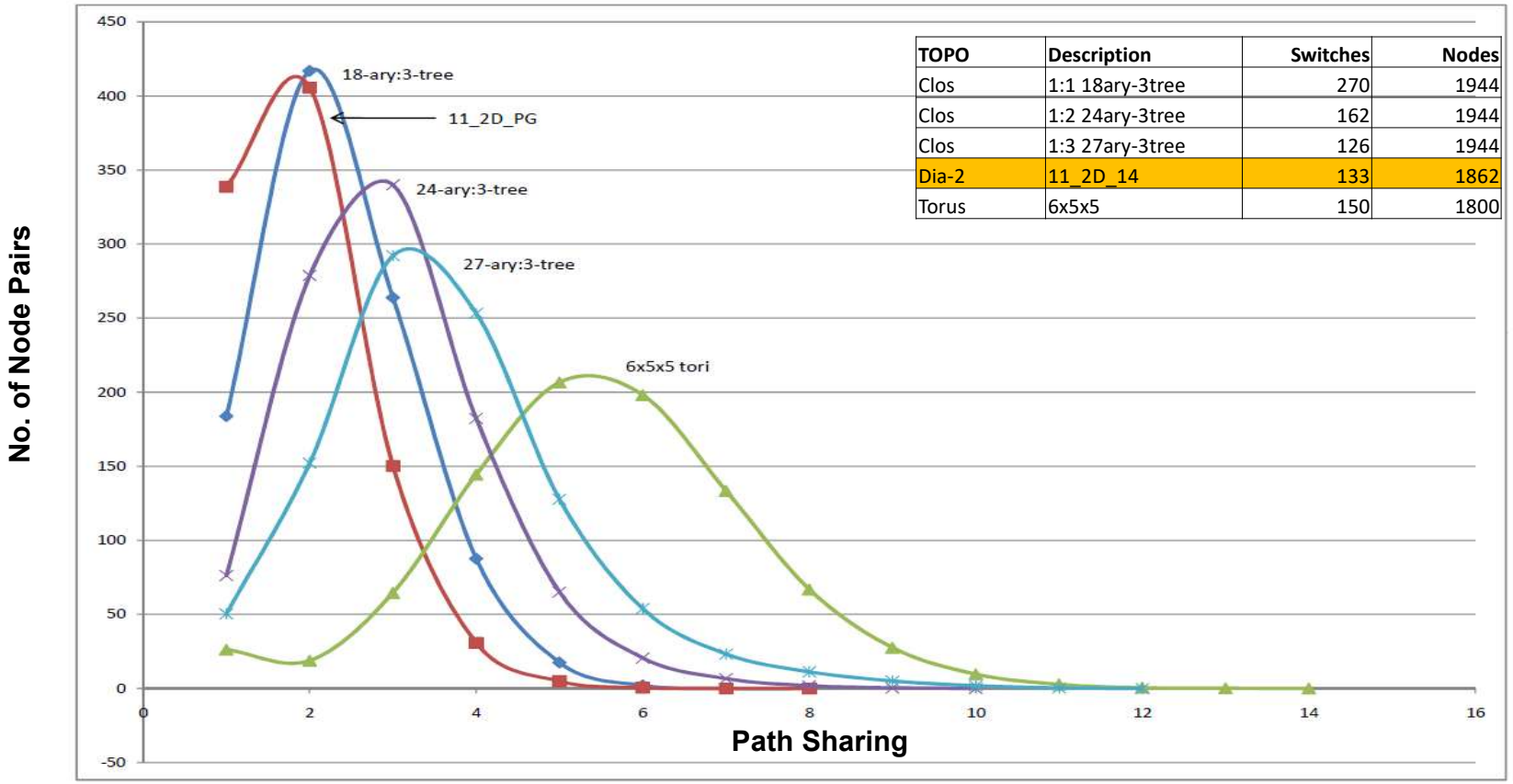


Over Subscription and Bisection in Static Routing



- Divide cluster into two subset of nodes (random)
- Subset-1 = Tx and subset-2 =Rx
- Trace data-flow path between each communicating pair
- Increment “subscription bucket” if any segment in flow path overlaps
- Repeat n times for statistical accuracy and plot

And Some Simple Results

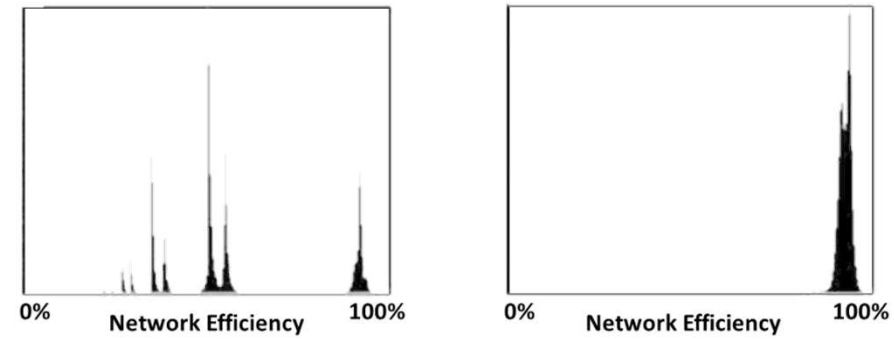


TOPO	Description	Switches	Nodes
Clos	1:1 18ary-3tree	270	1944
Clos	1:2 24ary-3tree	162	1944
Clos	1:3 27ary-3tree	126	1944
Dia-2	11_2D_14	133	1862
Torus	6x5x5	150	1800

Adaptive Routing (AR) 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.



Summit's MpiGraph Output

Accelerating All Levels of HPC / AI Frameworks

Application

- Data Analysis
- Real Time
- Deep Learning

Communication

- 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

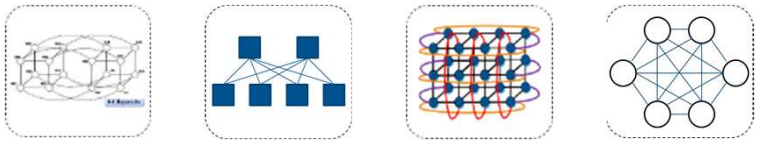
Network

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

RDMA GPUDirect · SHIELD

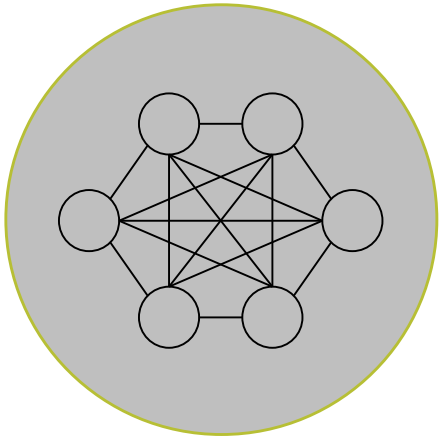
Connectivity

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

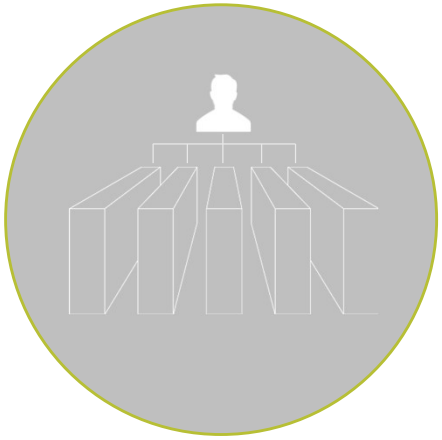


InfiniBand Technology Fundamentals

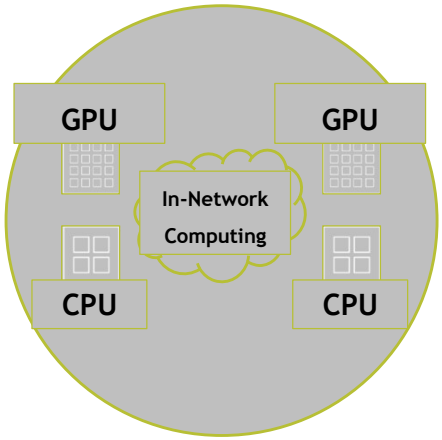
Standard



Architected to Scale



Software Defined Network
Centralized Management



In-Network Computing

Thank You