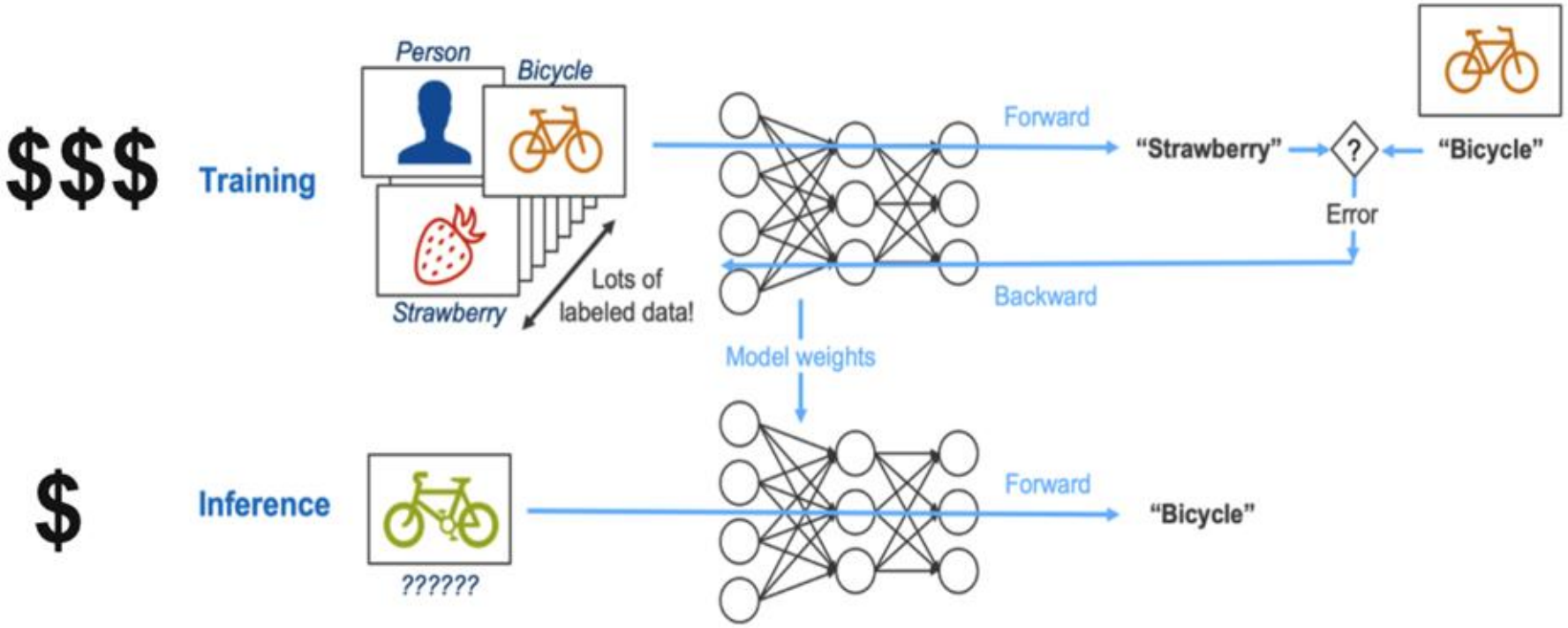
Building AI Back-End Fabric

Paul Ho

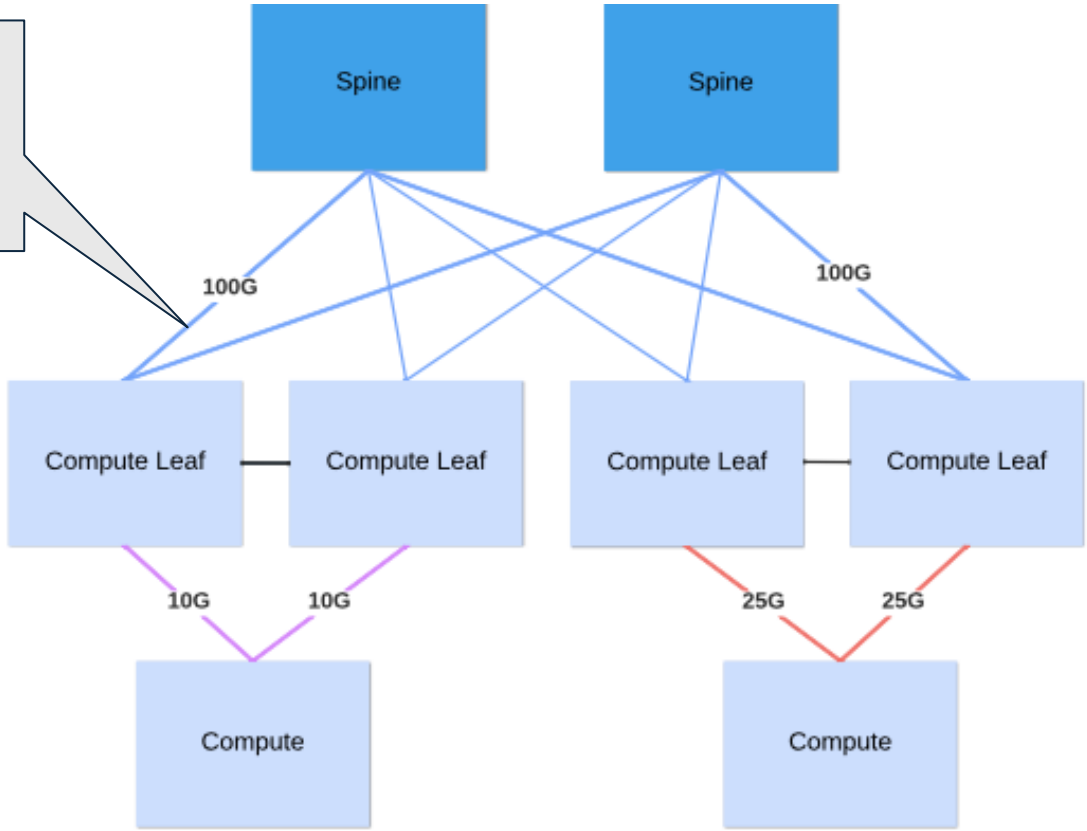
Tech Manager, System Engineering

Types of AI Networks



Standard Data Center Networks today

The average network focused on supporting CPUs and likely 100GbE Leaf/Spine



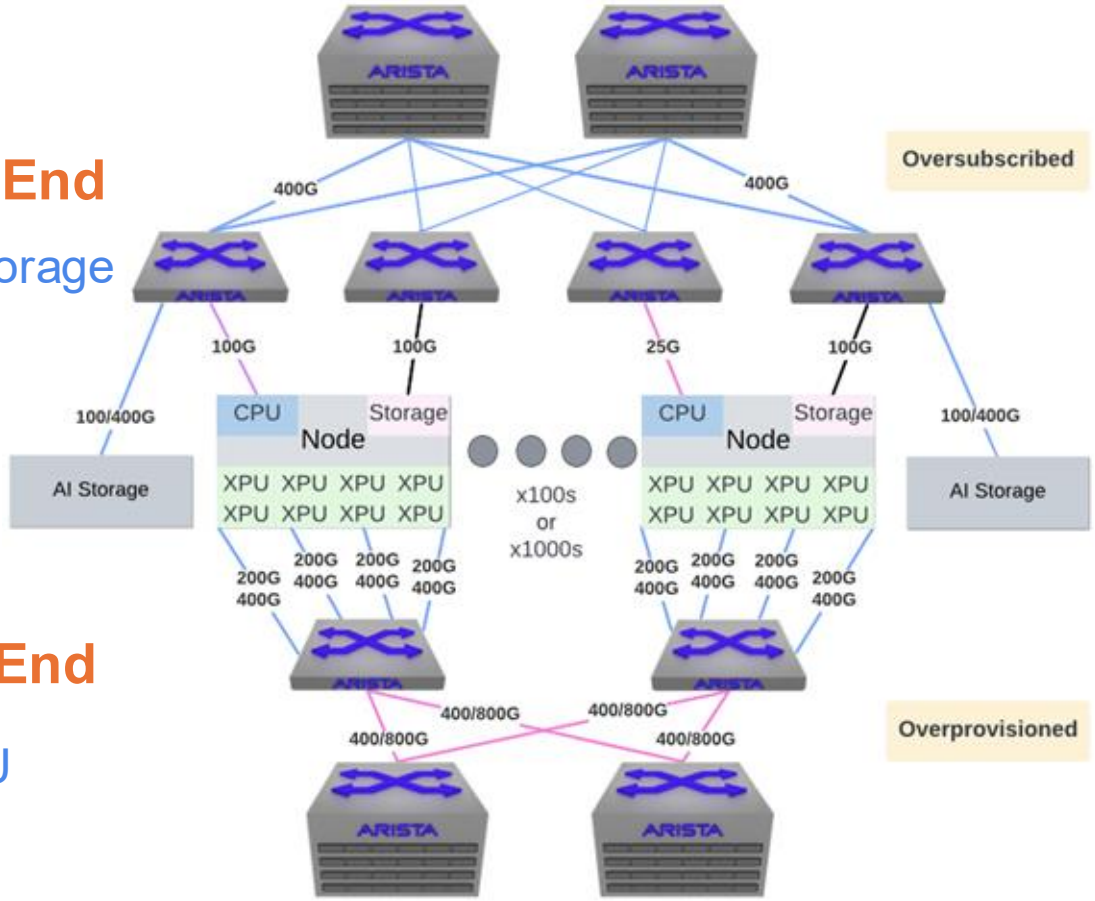
AI Center - What do these clusters look like?

Front End

CPU/Storage

Back End

XPU



Front End
Oversubscribed
(Typical DC BW)

2 or 4 to Spines



48 to Compute

Back End
Overprovisioning
(1:1 example)

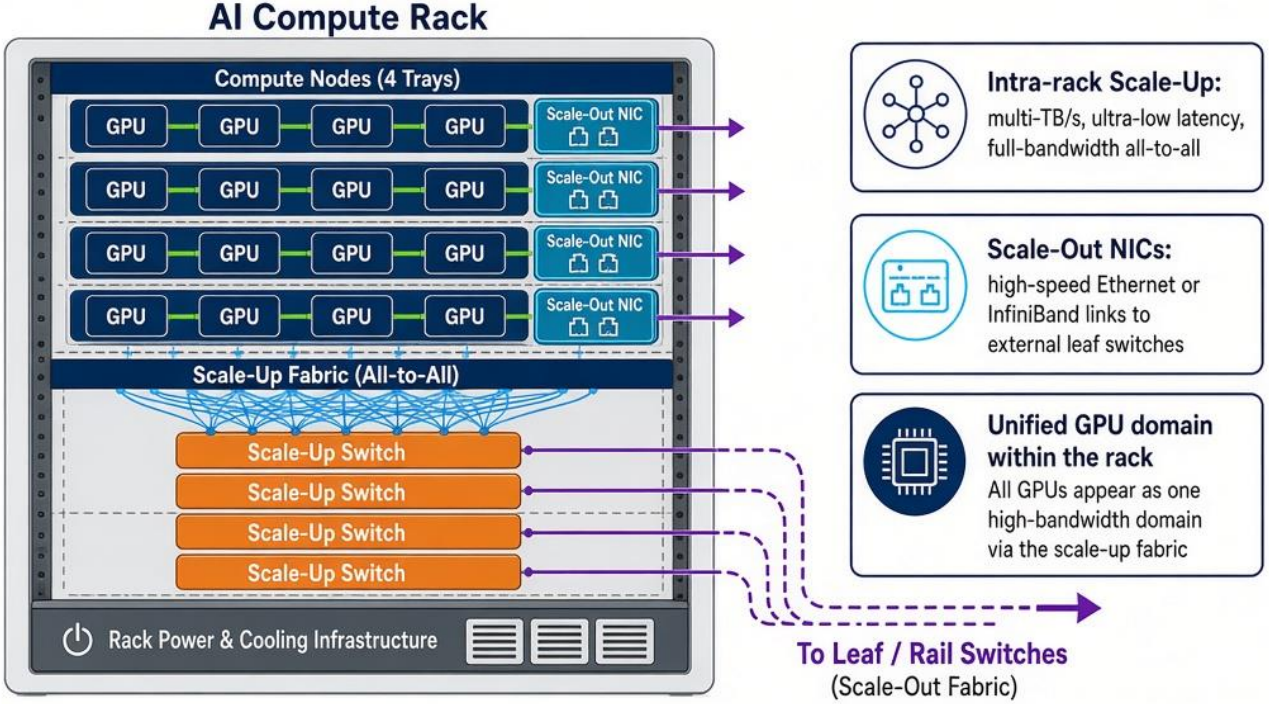
32 to Spines



32 to XPUs

64x 400GbE

AI Network Architecture Within a Single Rack



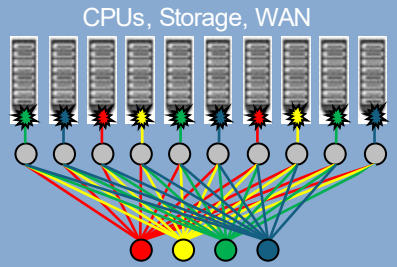
LEGEND Compute trays / GPUs · Scale-Up switches · Scale-Up interconnect (ultra-high bandwidth) · Intra-node links (within tray) · Scale-Out links leaving the rack (to leaf / rail switches)

Within the rack the scale-up fabric creates a single high-bandwidth domain; scale-out NICs connect the rack to the larger cluster fabric.

The 4 AI Fabrics for Training & Inference

Front End: Scale Out

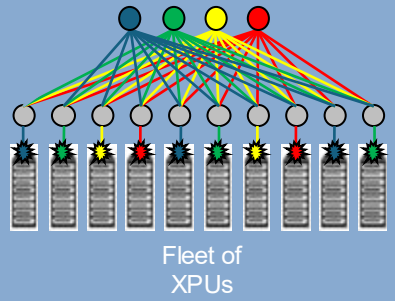
Connecting CPUs and Storage to the Data Center and Users



Thousands of XPU hosts, high speed storage and support applications

Back End: Scale Out

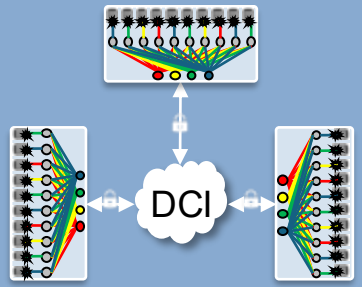
Back-End Connecting Racks of XPUs Together to Form a Cluster



Inter-rack scale out to ~100k XPUs

Back End: Scale Across

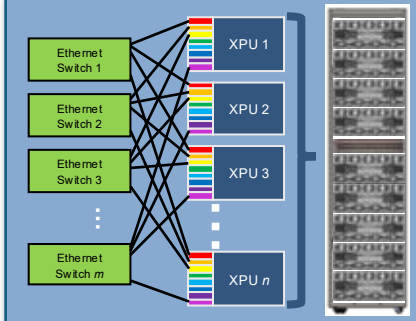
Connecting Distributed Clusters to Increase Scale



Inter-site scale across to >1M XPUs

Back End: Scale Up

Connecting XPUs within an Enclosure



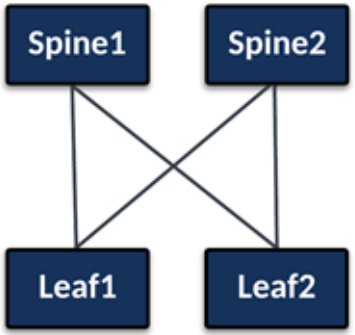
Intra-rack scale up for ~1k XPUs

Backend Network - Topologies

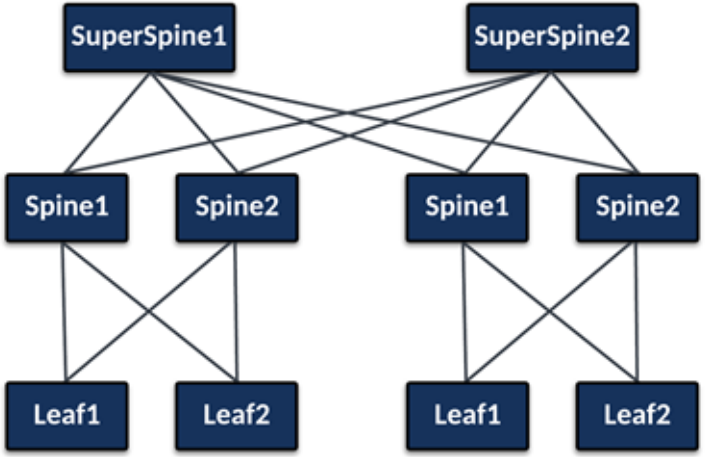
Single-Box vs 2-Tier vs 3-Tier

Modular Leaf

<1K GPU



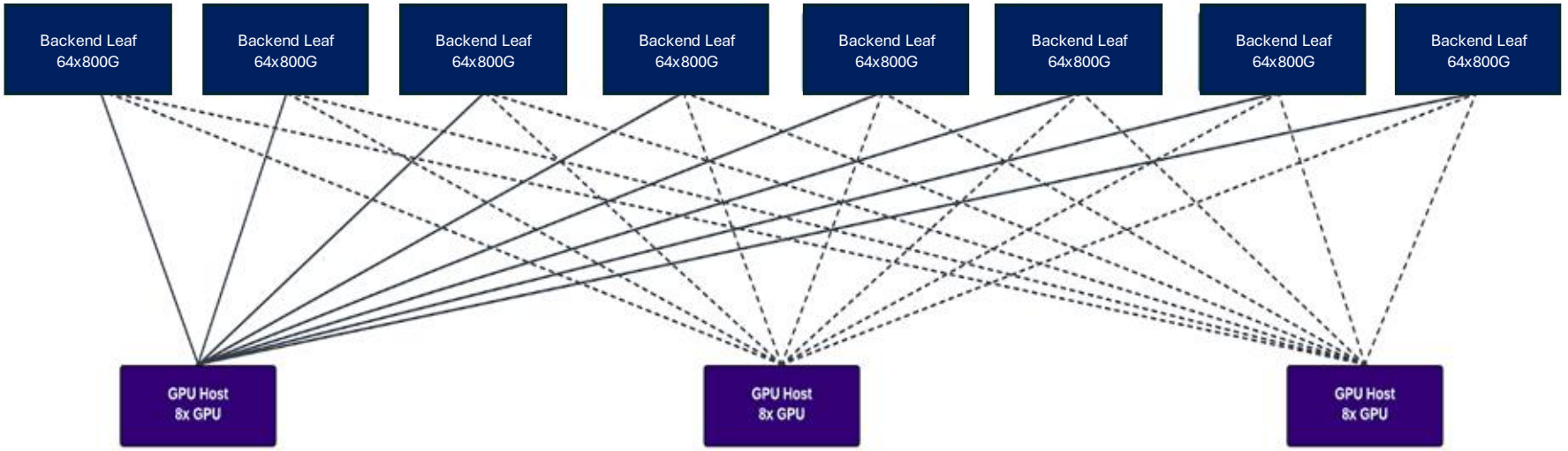
<8K All Fixed
<73K Hybrid Modular/Fixed
<663K All Modular



<524K All Fixed
191M Modular

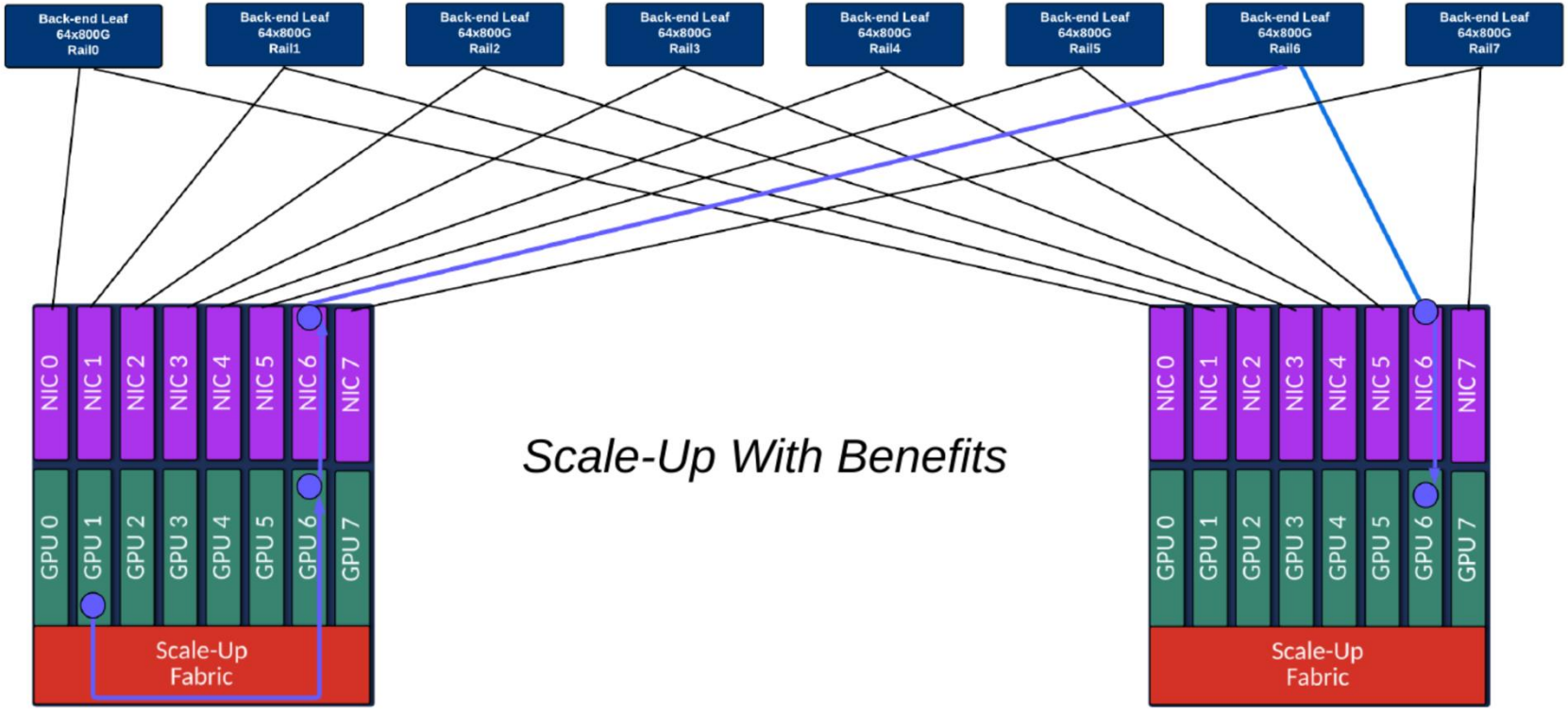
Rail Only Example

No Spines in Scope at lower scales*



Each GPU in the chassis connects to one leaf.

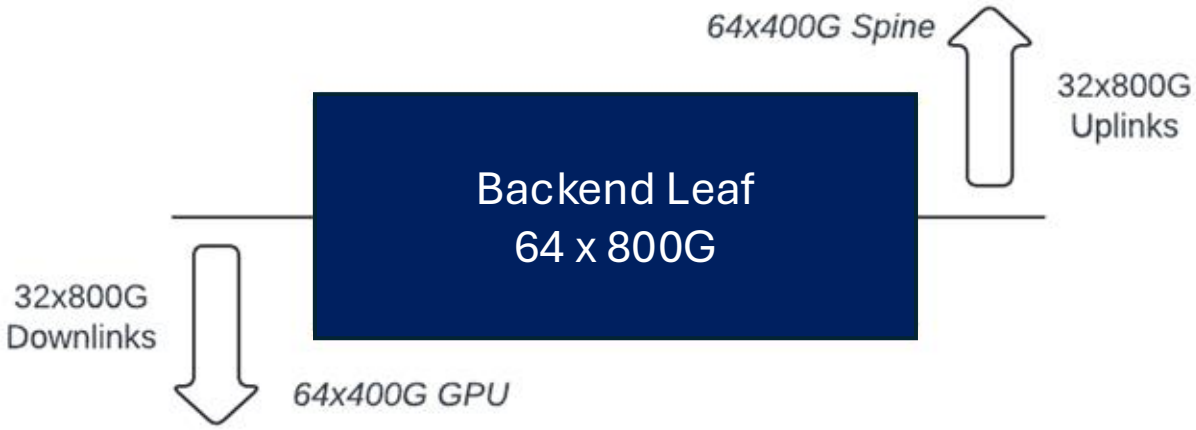
PXN - PCI-E Extended Networking



Use another GPU NIC to reach other Rails

PXN: GPU 1 hops through the scale-up fabric to a sibling NIC on the same host to reach a remote rail-6 leaf

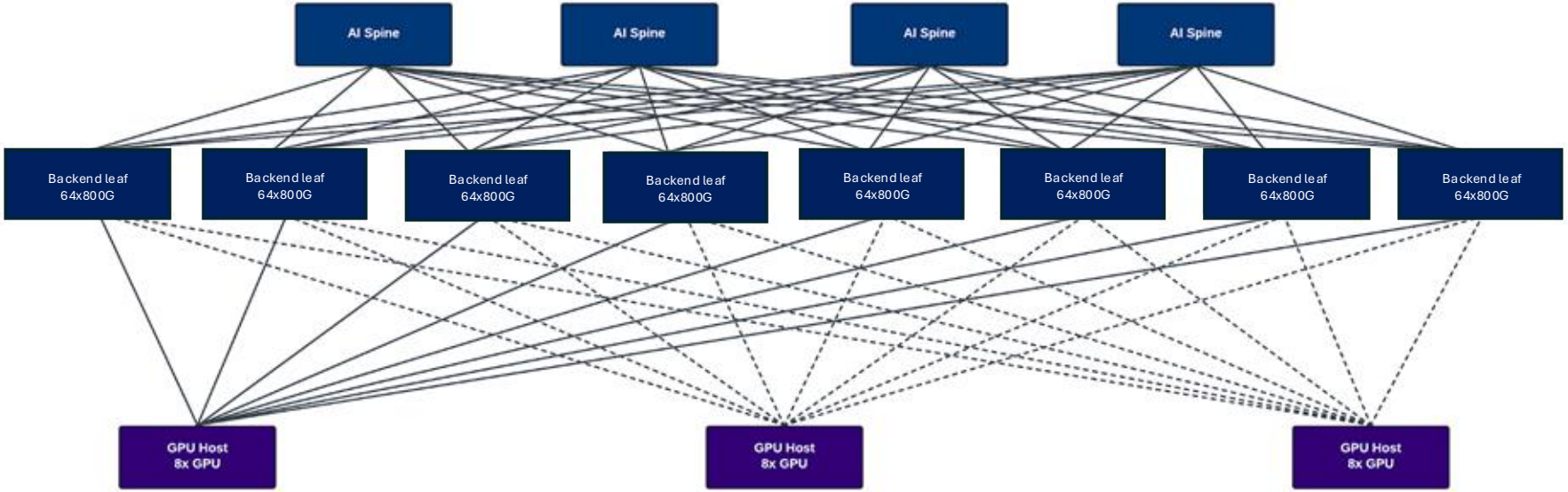
Subscription Ratios



Target is 1:1 or better for uplink to downlink ratios.

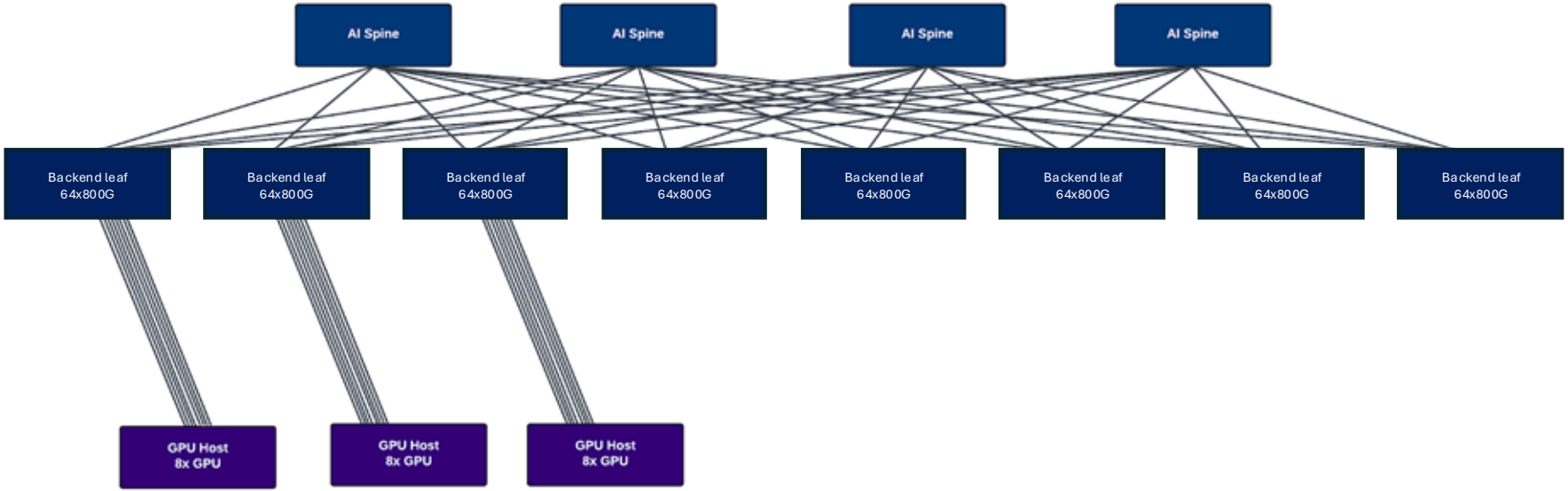
Rail Optimized Example

Spines available for scalability / future requirements.



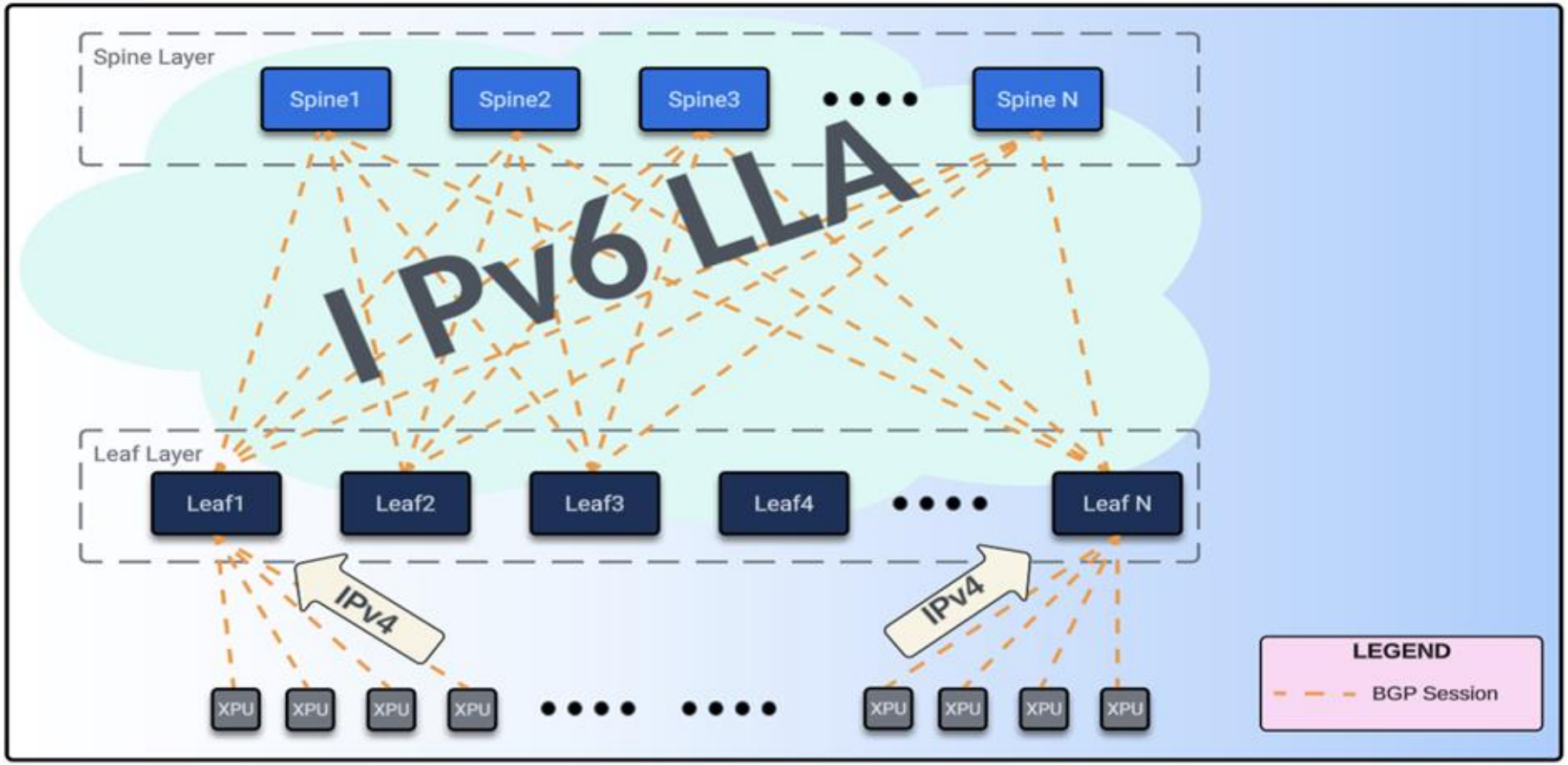
Each GPU in the chassis connects to one leaf.

“Fat Tree” Example



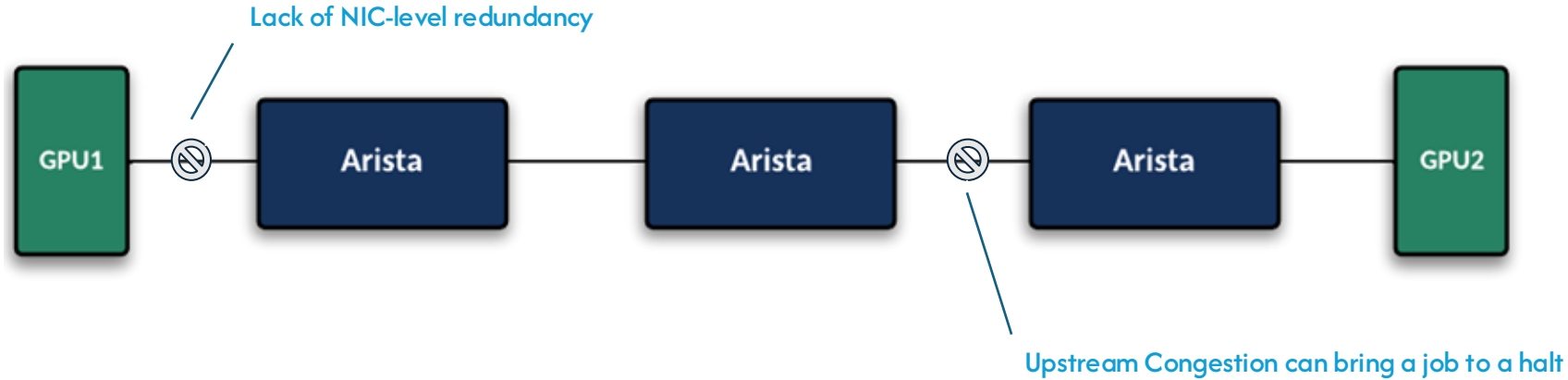
Each server connects all ports to a single leaf

Routing in AI Fabric – RFC-5549/8950 IPv4 over IPv6 NH



Intro to Multi-Plane Designs

2-Tier vs 3-Tier:
- Lower Latency
- Less Power
- Fewer(ish) Optics/Switches



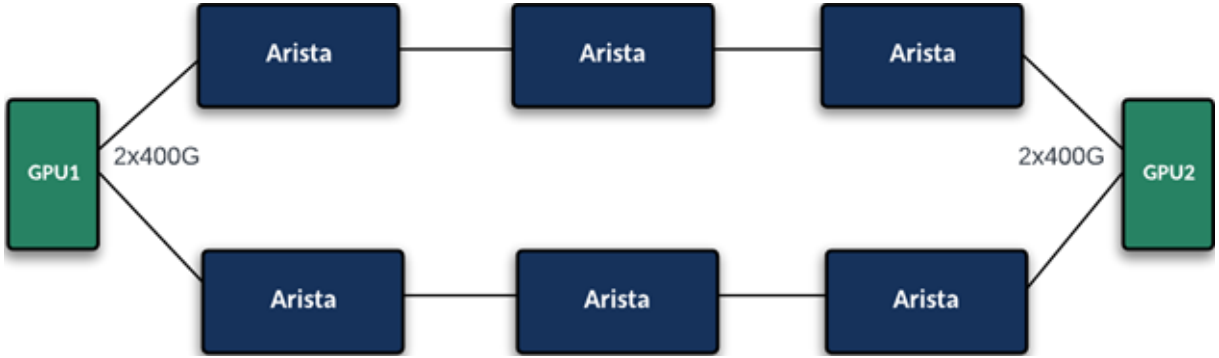
Higher speeds = low radix.
64x800G Switch w/400G Uplink in 2-tier = 4096 GPU
64x800G Switch w/800G Uplink in 2-tier = 2048 GPU

Make like a banana and…



64x800G
- 32 Spine per plane
- 64 Leaf per plane (32 downlinks)

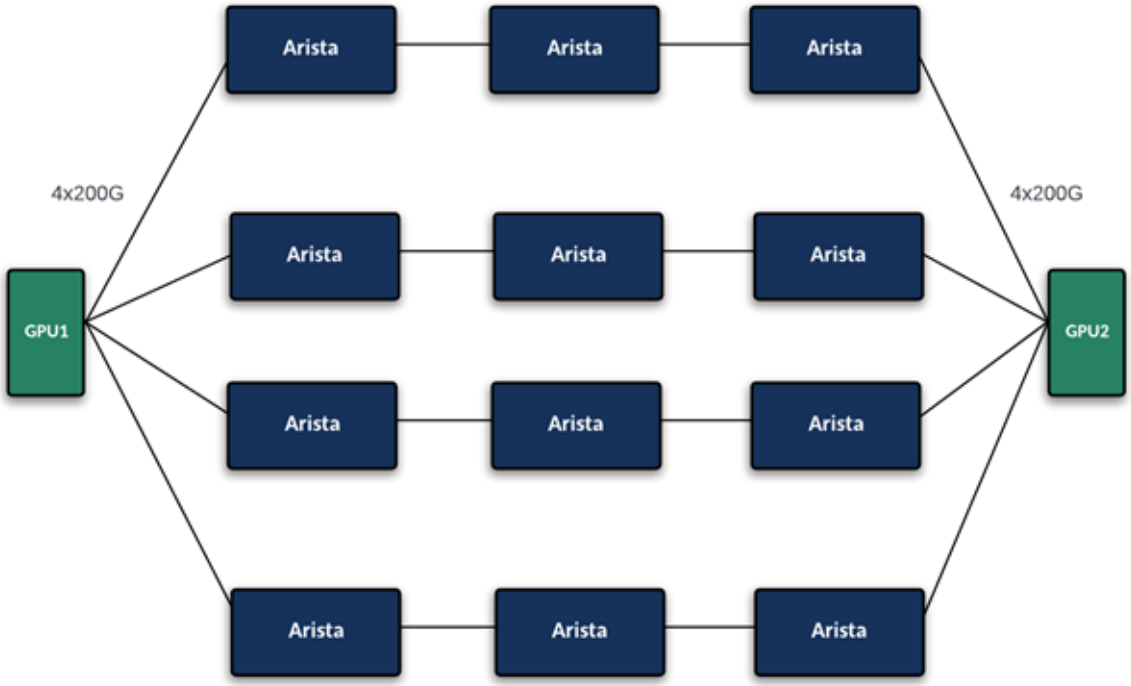
1x800G Max Scale (64x800G 2-tier) = 2048 GPU



128x400G
- 64 Spine per plane
- 128 Leaf per plane (64 downlinks)

2x400G Max Scale (64x800G 2-tier) = 8192 GPU

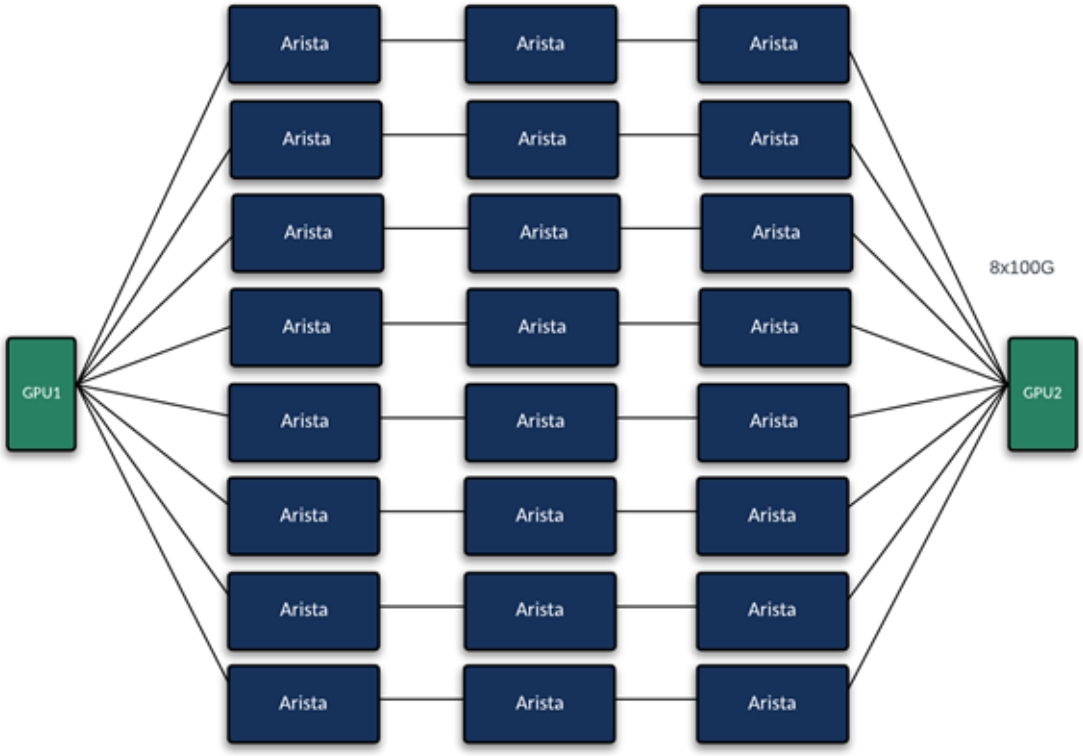
Taking it a step further…



256x200G
- 128 Spine per plane
- 256 Leaf per plane (128 downlinks)

4 x 200G Max Scale (64x800 2-tier) = 32768 GPU

Max Radix 8x100G



512x100G

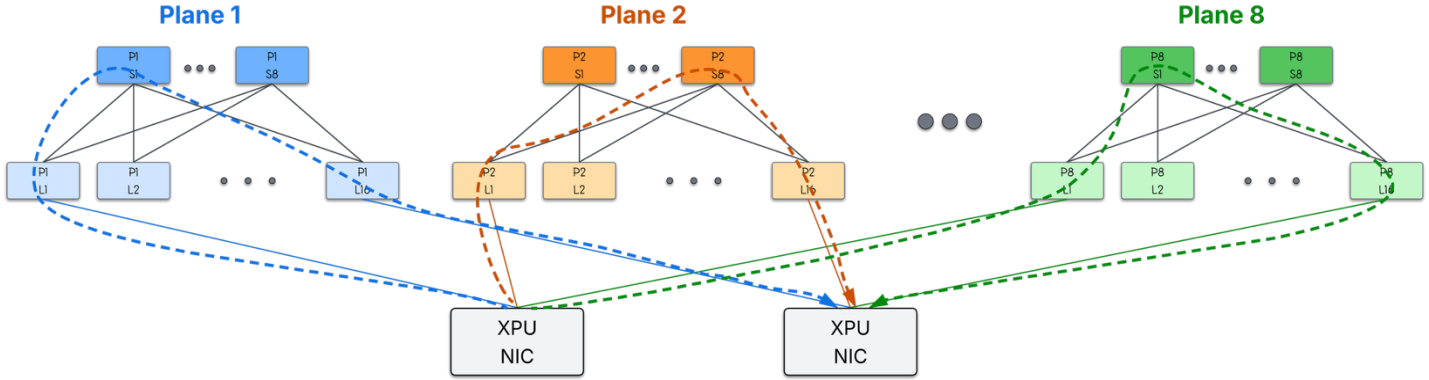
- 256 Spine per plane
- 512 Leaf per plane (256 downlinks)

Note that the number of devices doesn't grow by nature of adding more planes. It increases due to the lowering of the link speed between gpu and leaf or leaf and spine; the planes just add that bandwidth back to the port, making it so it's not a net loss by dropping speed.

MultiPath Reliable Connection- MRC

Multipath Reliable Connection - (MRC)

- Open OCP specification - new RDMA-based transport protocol, MRC, sprays across many paths and actively load-balances between them, eliminating the issue of flow collisions
- the use of multi-plane Clos topologies to get the benefits of high switch radix and redundancy, allowing training clusters well over 100K GPUs to be built as two-tier topologies while increasing physical redundancy, and
- the use of static source-routing using SRv6 to allow MRC the freedom to bypass failures by itself.
- End station NICs stripe their traffic across multiple links and paths to the receiver, with out of order packets automatically handled.



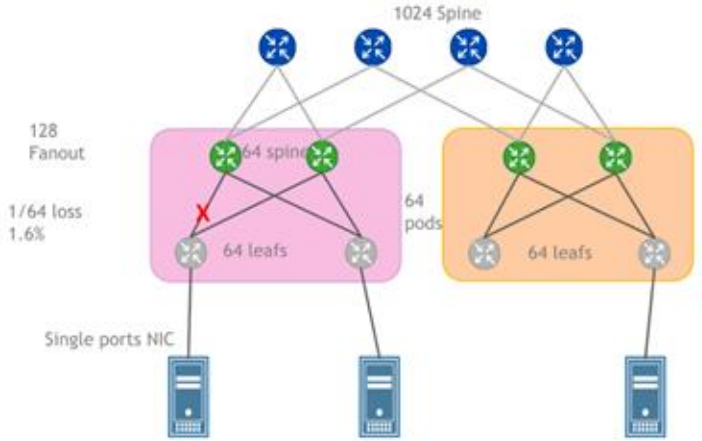
MRC Operation

- Restricted RDMA operations, targeting AI collective specifically
  - Leverage and extends Verbs API
  - At transport Level, only RDMA write and write-with-immediate operations
- Packet spraying for optimized load-balancing
- Out-of-order packet delivery
- Adaptive load-balancing/re-balancing based on ECN, packet trimming and probes
- Selective retransmission
- Typical deployment is an IPv6 underlay with SRV6

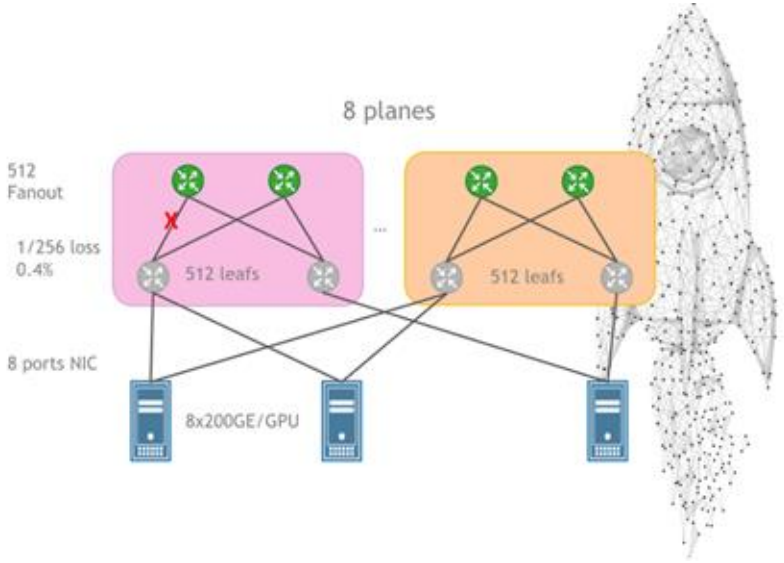
Innovation from OpenAI

Multipath Reliable Connection MRC May 2026

Legacy 3 Tier single plane



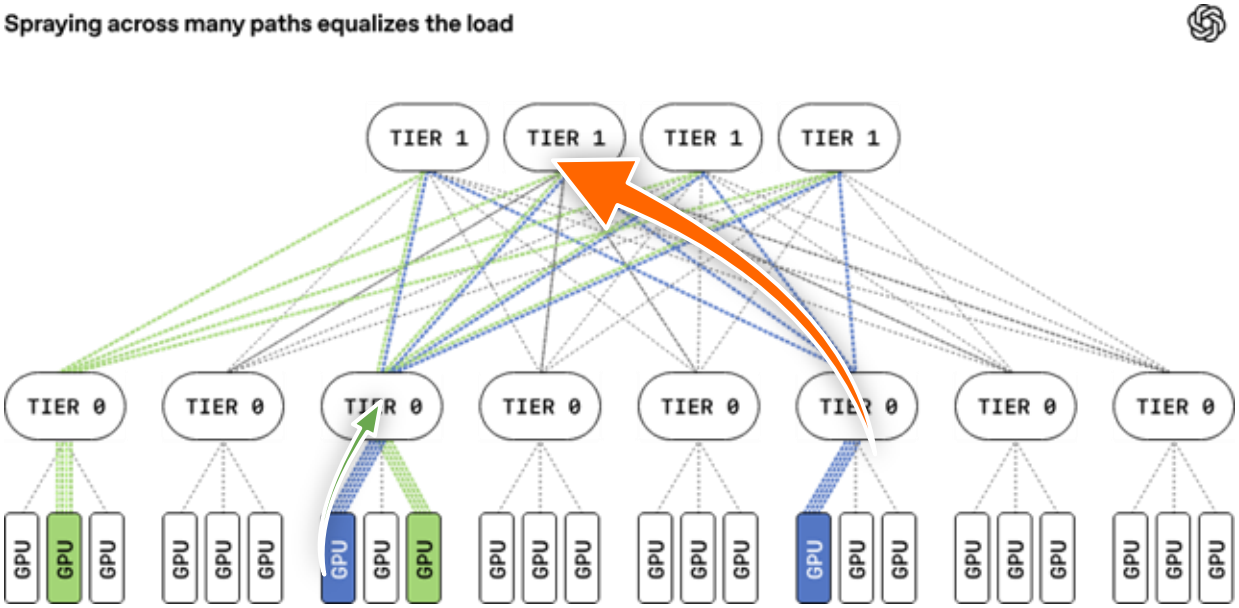
New Gen 2 Tier multi-plane



Cost and Power:

- 1/3 Optics and 2/5 switches saved
- Over 40% TCO

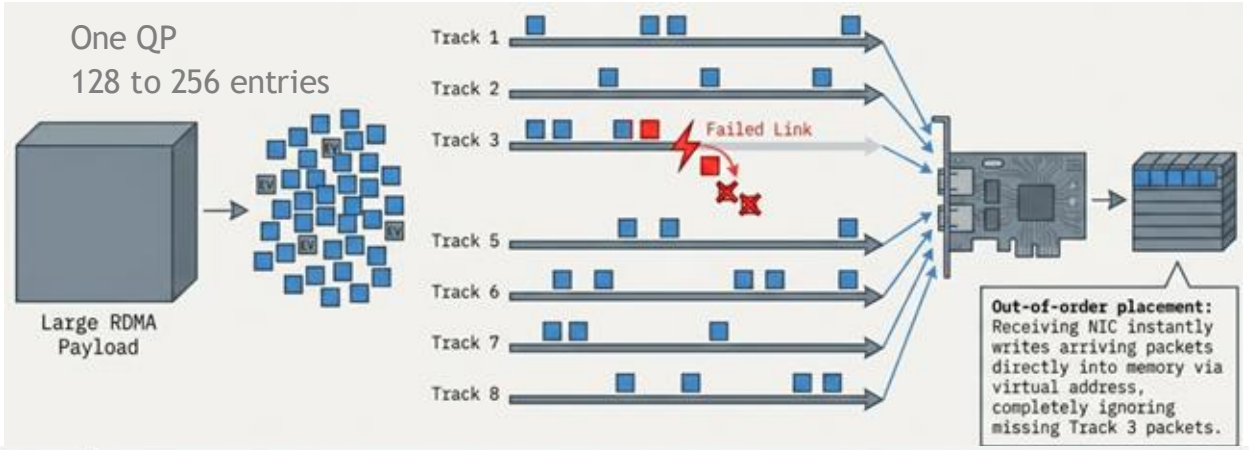
OpenAI Multipath Reliable Connection MRC

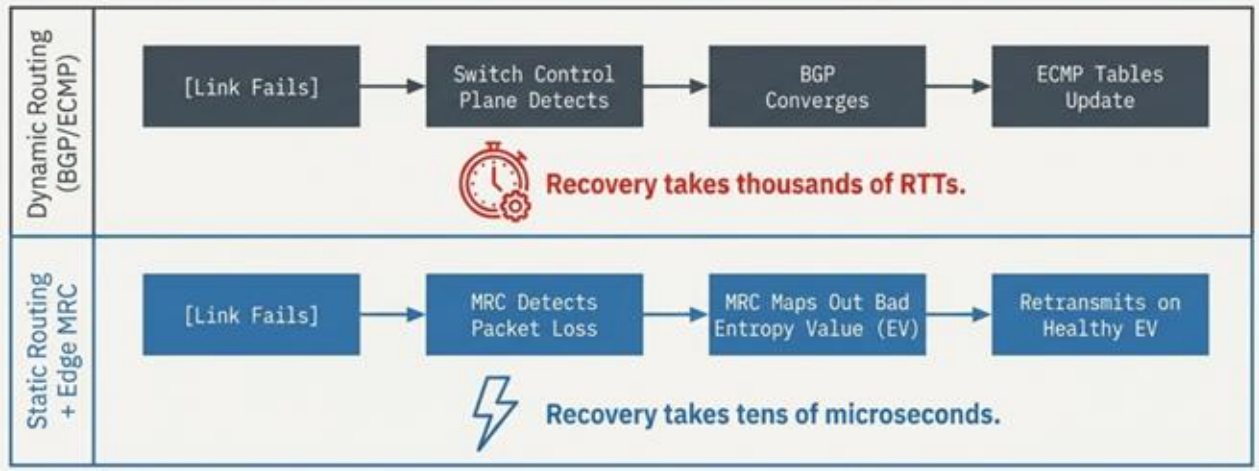


- Very Fast Convergence
- Very good performance
- No BFD, detection SRv6
- No Adptive Routing
- No ECMP explosion
- No BGP
- Link/Switch detection, no Flow detection

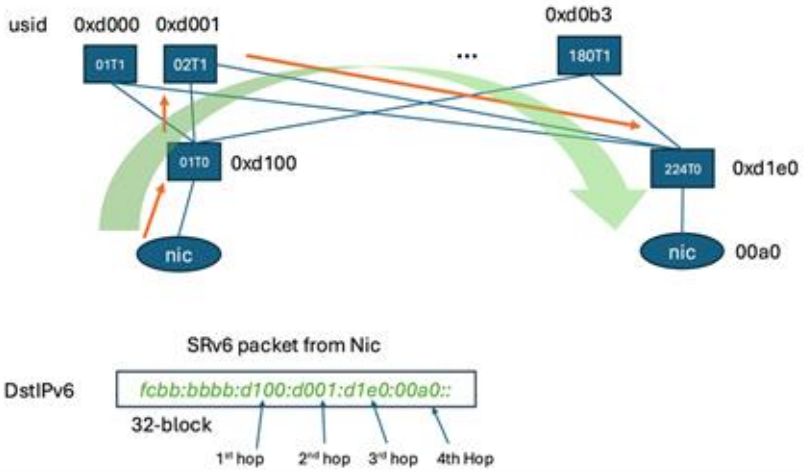
Multipath Reliable Connection MRC

BGP vs Static SRv6





Failure Protection



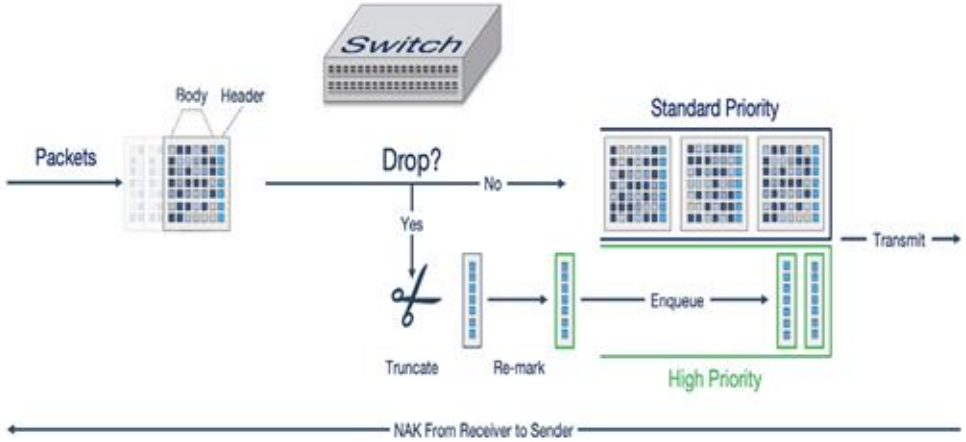
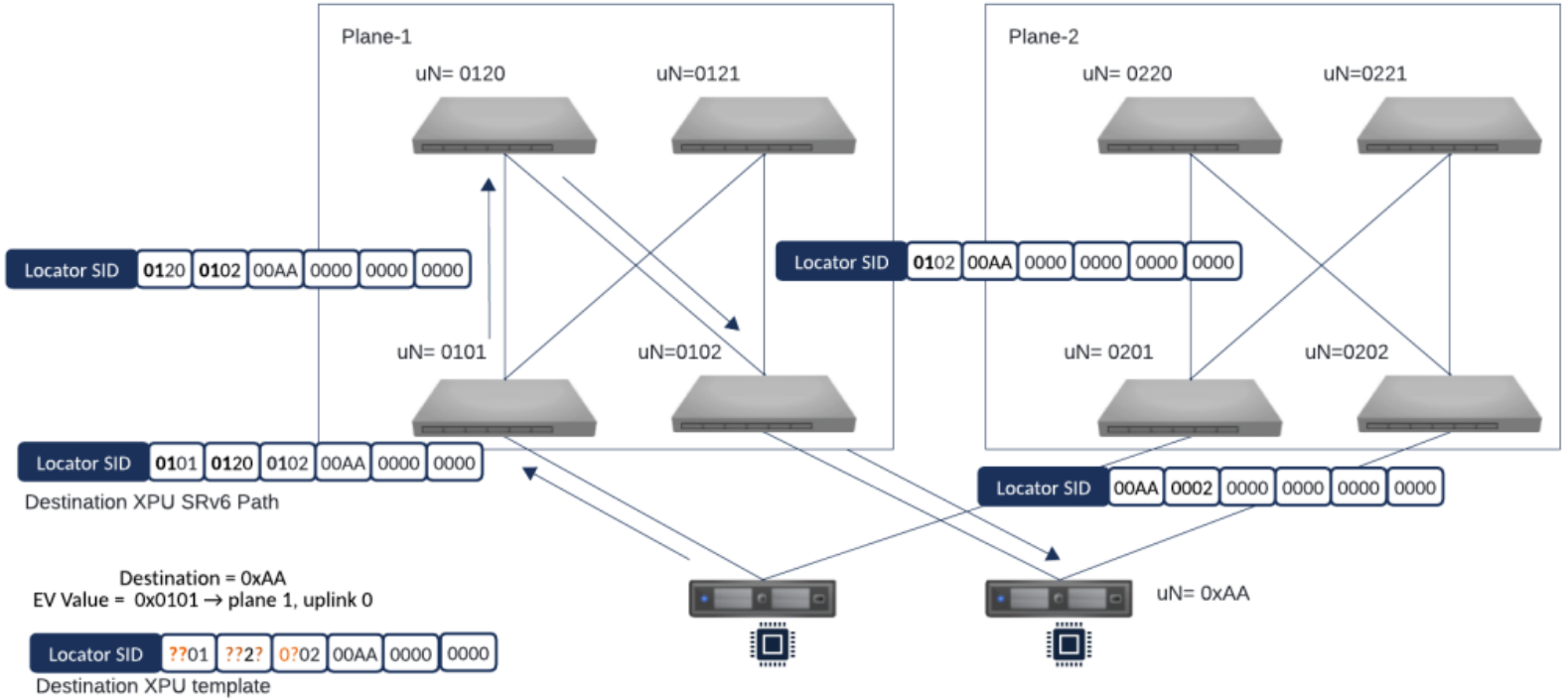
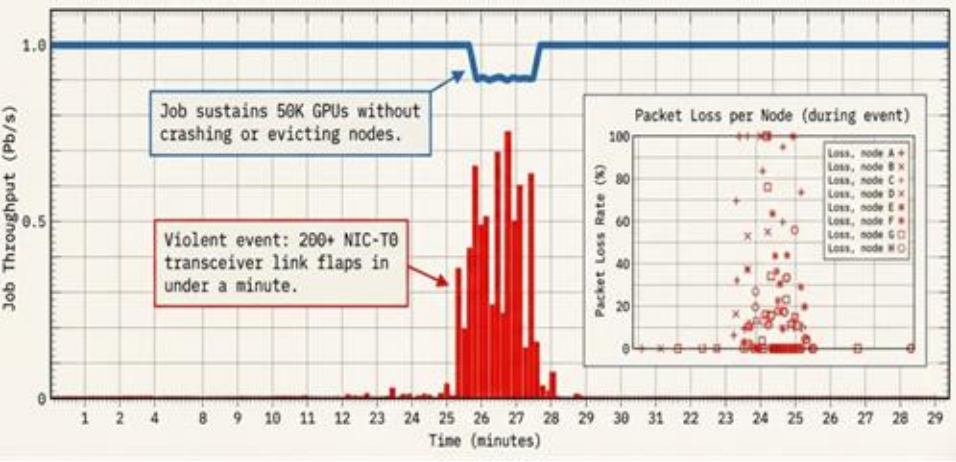


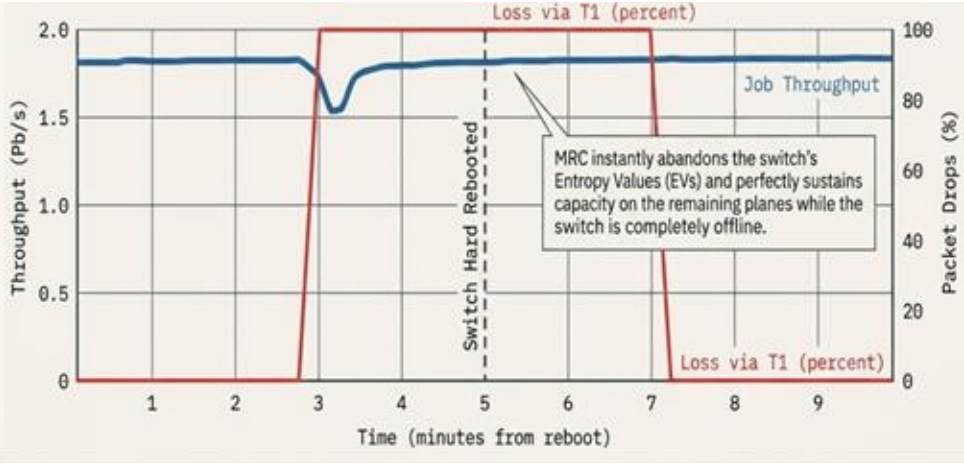
Figure 4: Packet Trimming Mechanism

MRC operation – SRV6



MRC Failure Protection





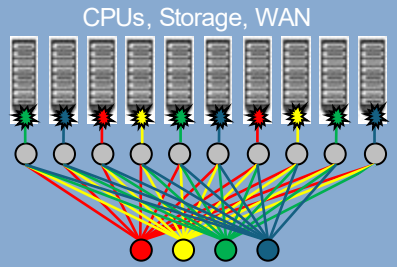
Link flaps between TO and T1 switches no longer crash training jobs. They are simply mapped out by MRC, turning urgent IT tickets into low-priority maintenance.

Switch software bugs cause gray failures-the control plane is up, but packets drop. Because MRC relies on static SRv6, operators can hard-reboot T1 switches without coordinating with active training jobs.

The 4 AI Fabrics for Training & Inference

Front End: Scale Out

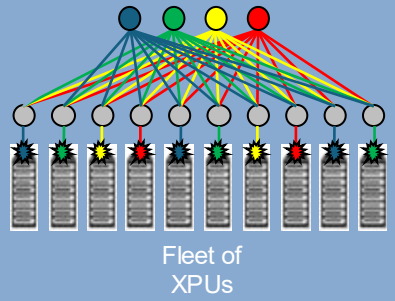
Connecting CPUs and Storage to the Data Center and Users



Thousands of XPU hosts, high speed storage and support applications

Back End: Scale Out

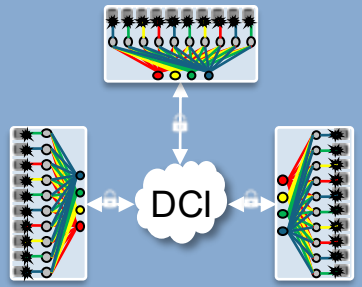
Back-End Connecting Racks of XPUs Together to Form a Cluster



Inter-rack scale out to ~100k XPUs

Back End: Scale Across

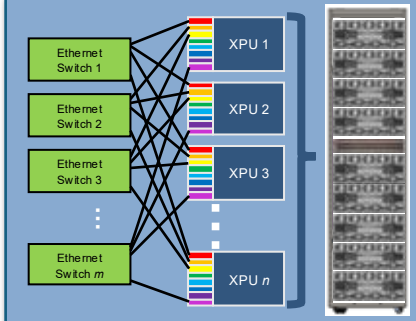
Connecting Distributed Clusters to Increase Scale



Inter-site scale across to >1M XPUs

Back End: Scale Up

Connecting XPUs within an Enclosure



Intra-rack scale up for ~1k XPUs

Thank You