From Datacenter to AI Center

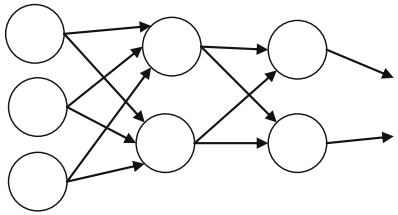
Building the networks that build AI

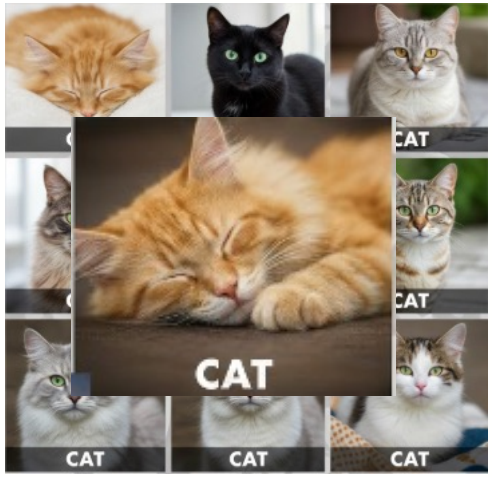
Tyler Conrad
Tech Lead, Systems Engineering

ARISTA

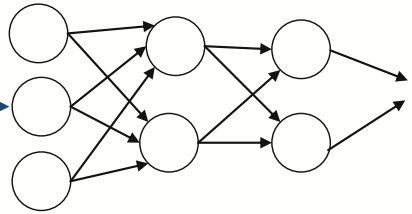
Types of AI Networks

Training
$$$





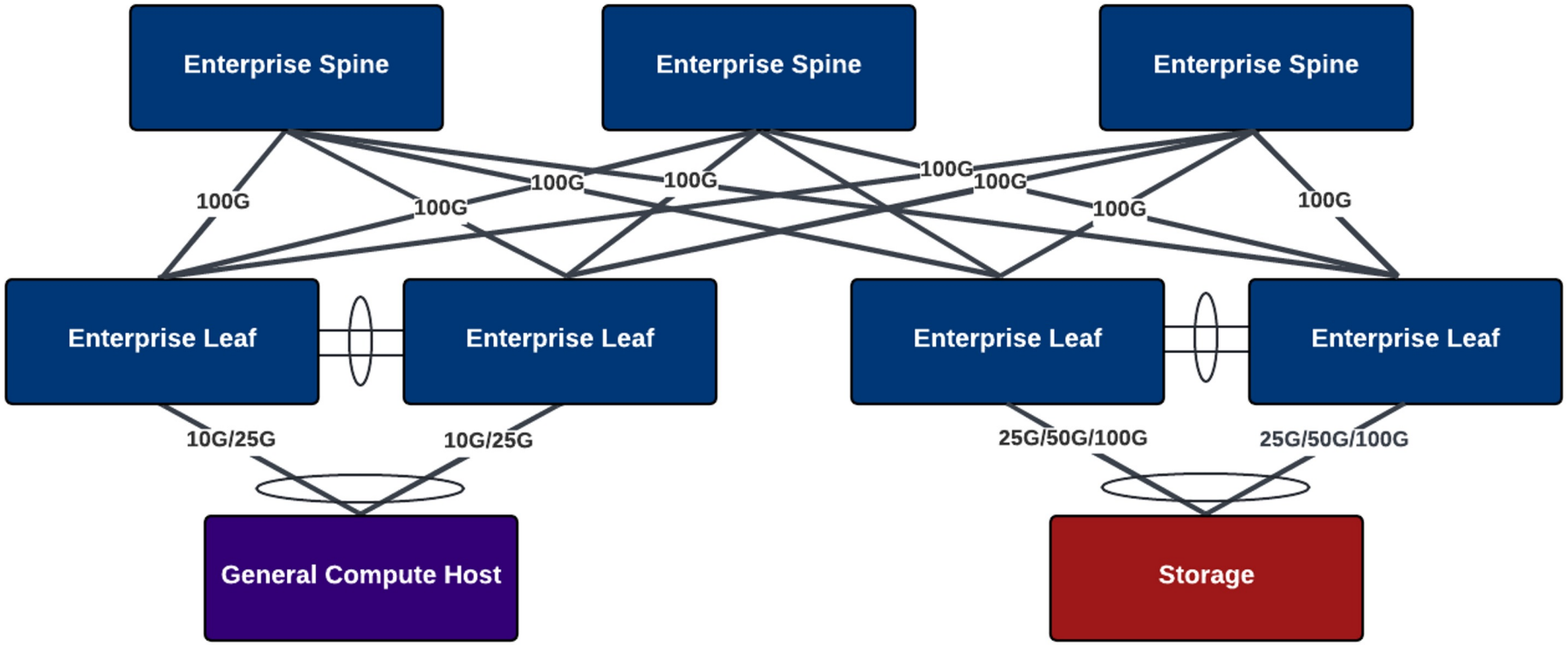
Inference
$



"Cat"

NANOG™

Traditional Networking

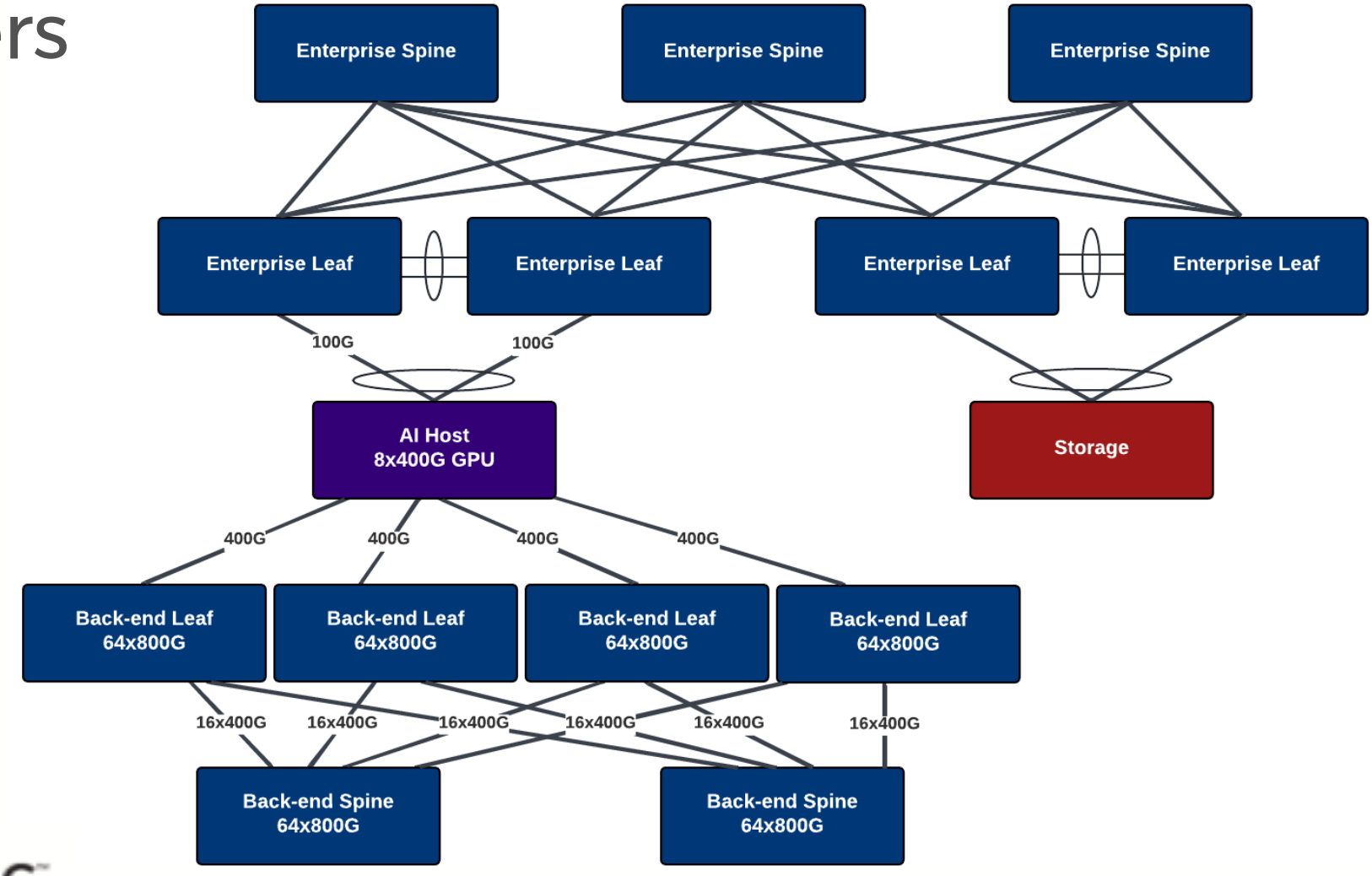


NANOG™

AI Centers

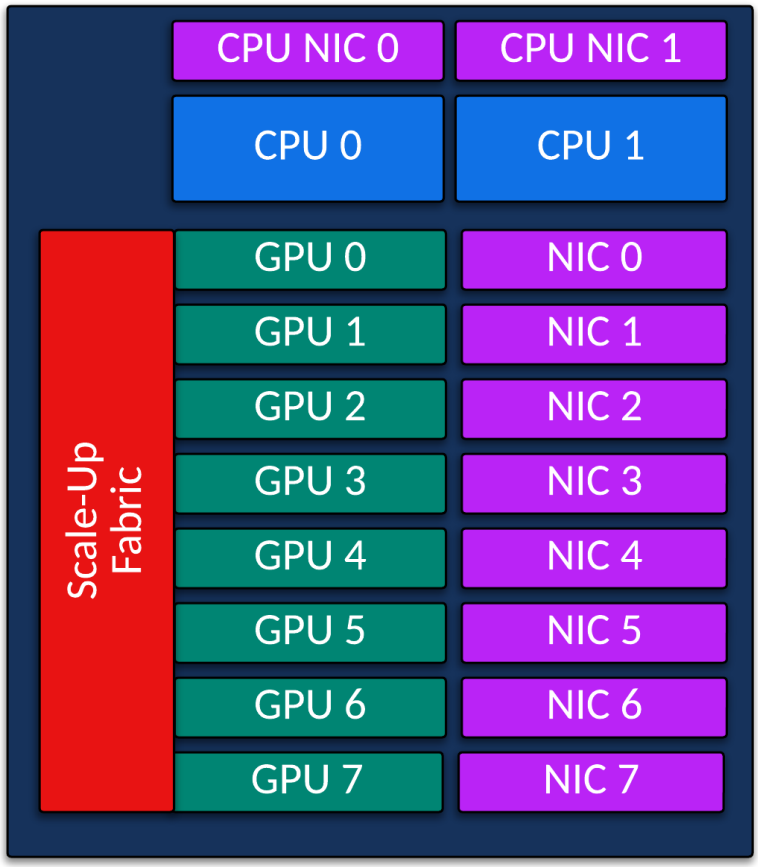
Front End
CPU

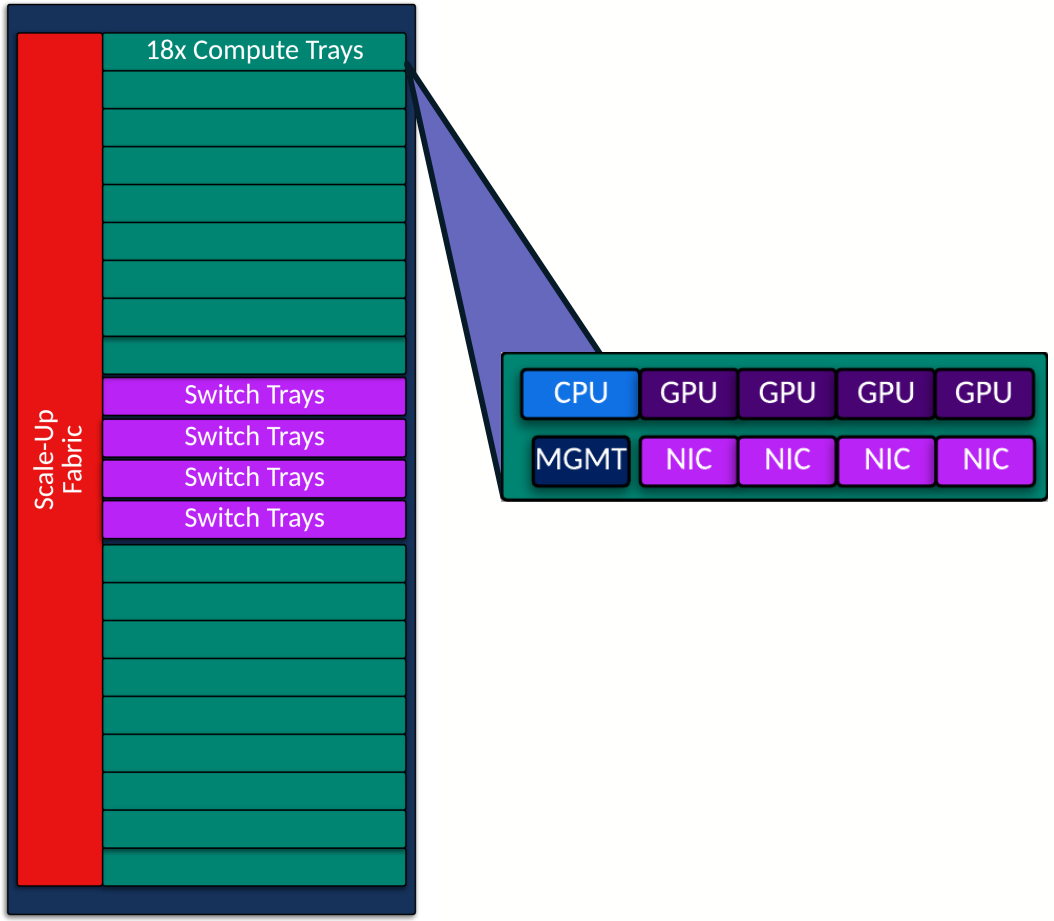
Back End
GPU



NANOG

What do these servers look like?



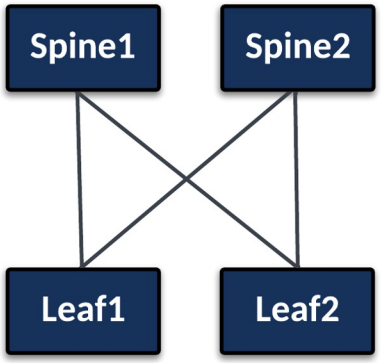


NANOG™

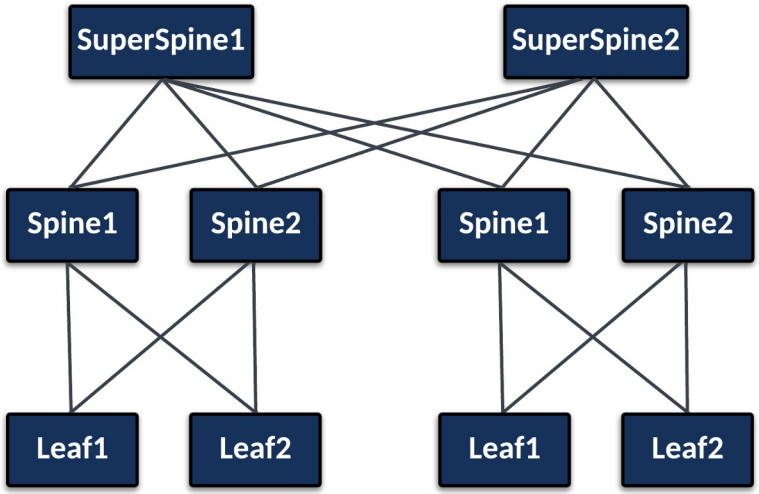
Topologies

Single-Box vs 2-Tier vs 3-Tier

<1K GPU



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



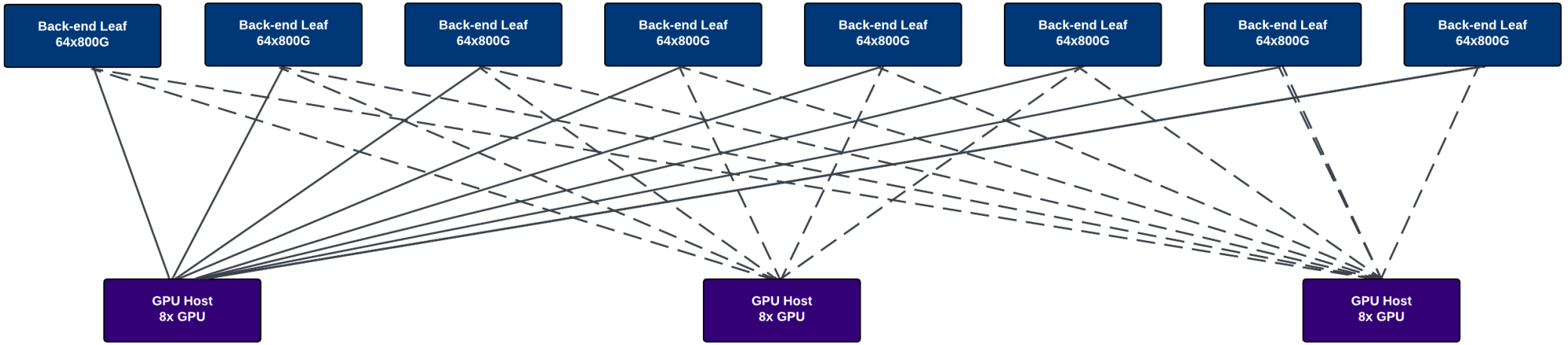
<524K Fixed
>100M Modular

NANOG™

Normalized for a 400G GPU

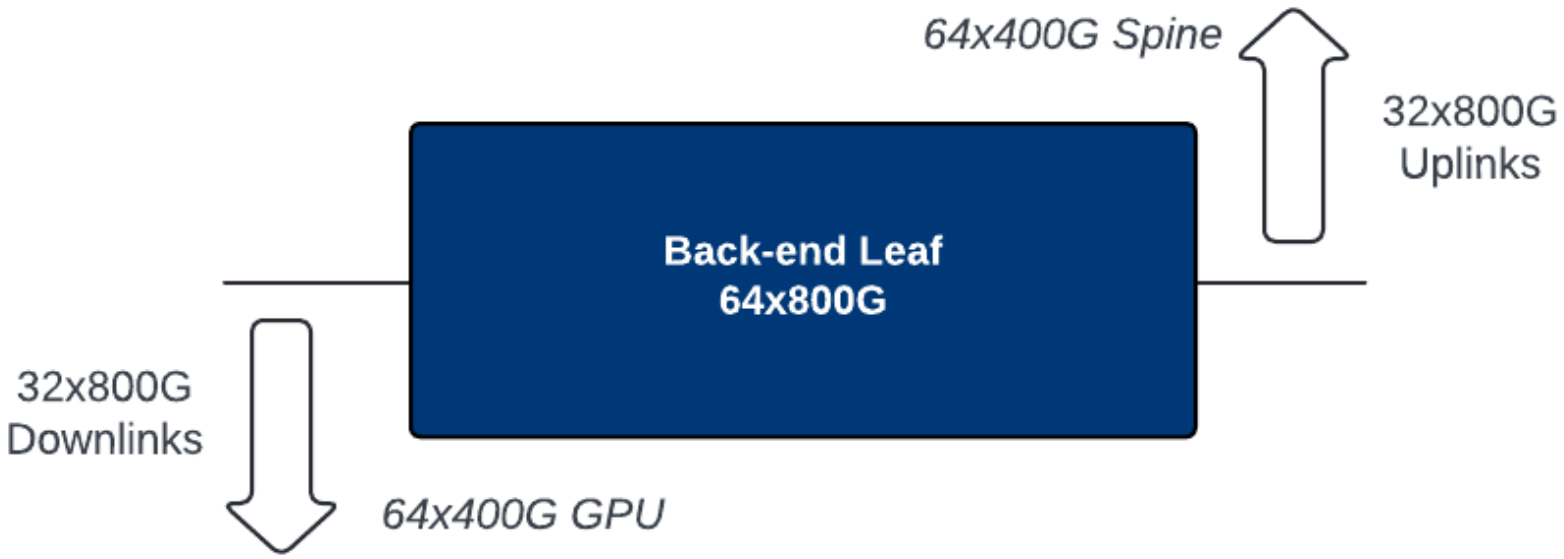
Rail Isolated Example

No spines in scope at low scale*



Each GPU in the chassis connects to one leaf.

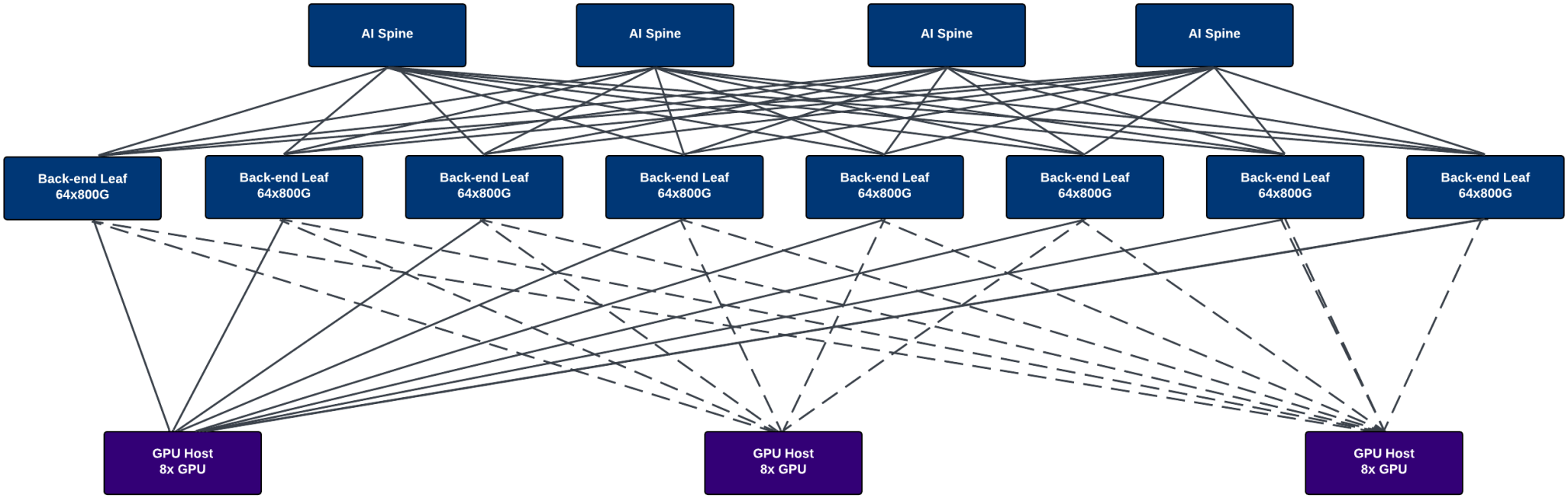
Subscription Ratios



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

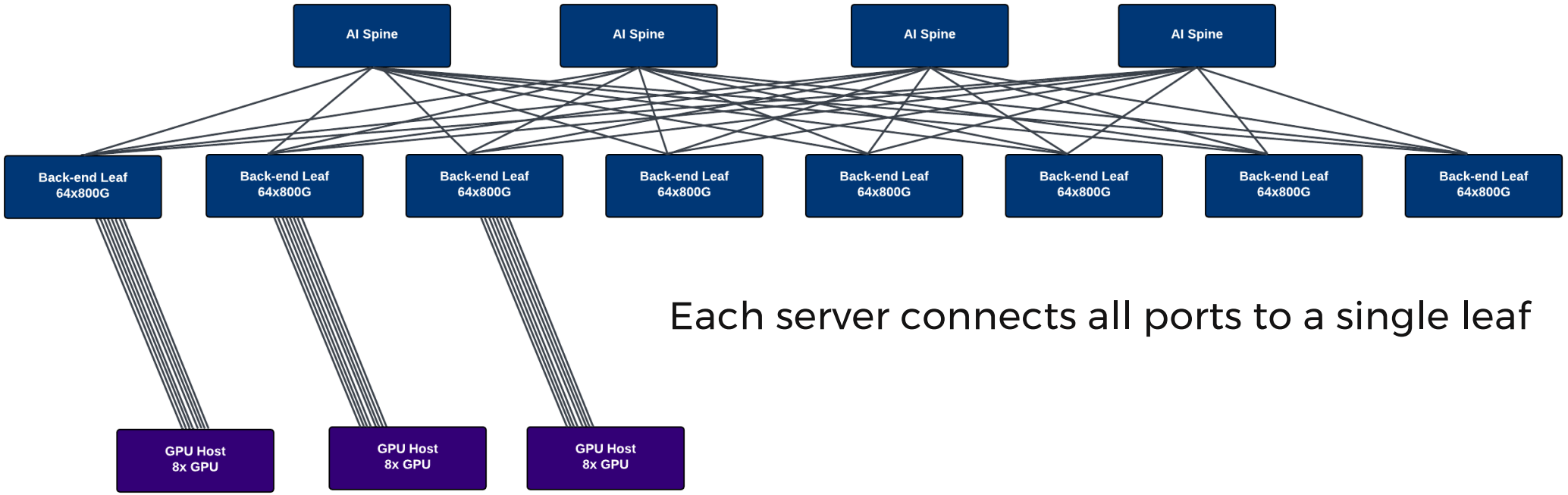
Rail Optimized Example

Spines available for scalability / future requirements.



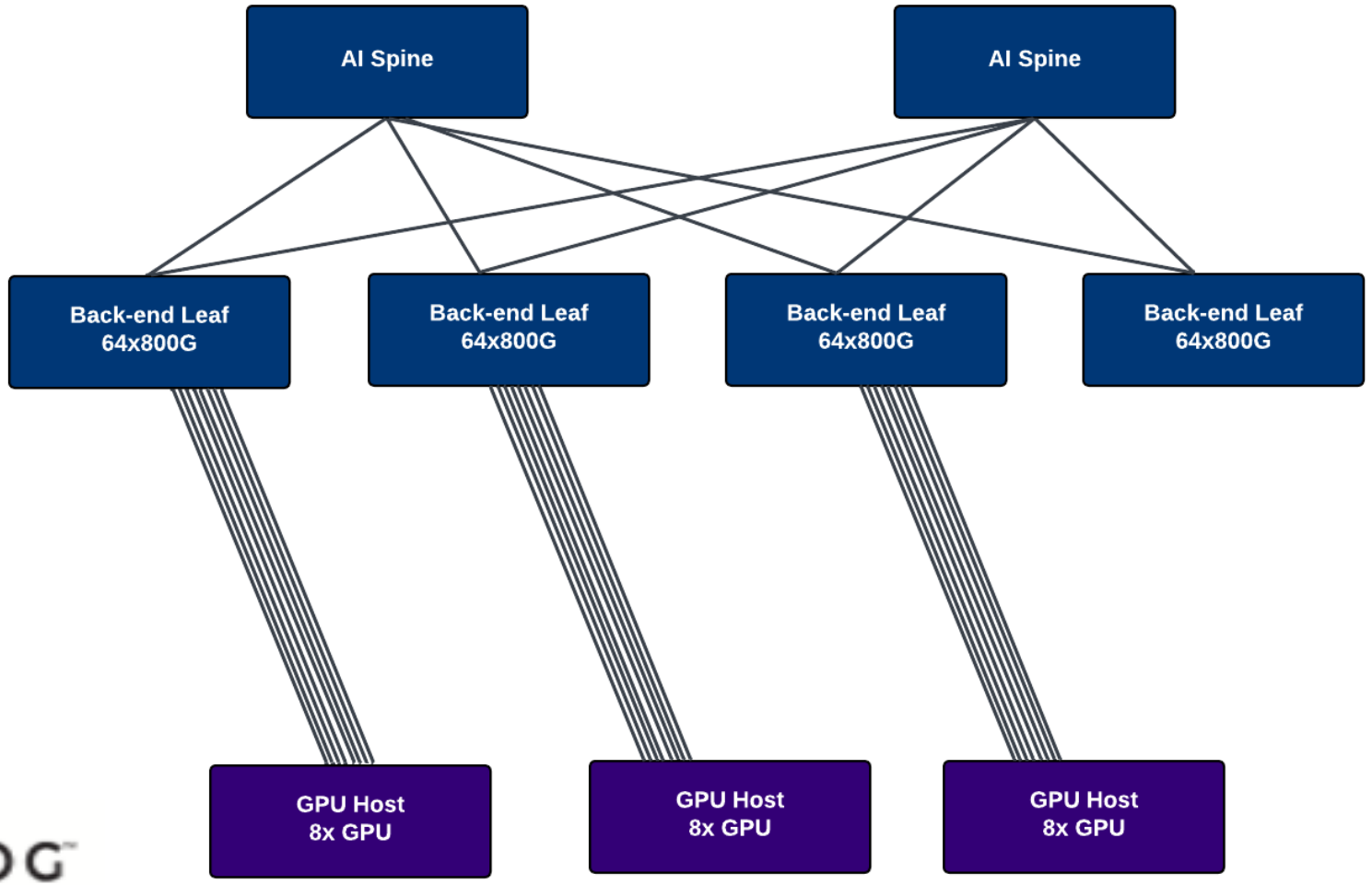
Each GPU in the chassis connects to one leaf.

Non-Rail / “Fat Tree” Example



Each server connects all ports to a single leaf

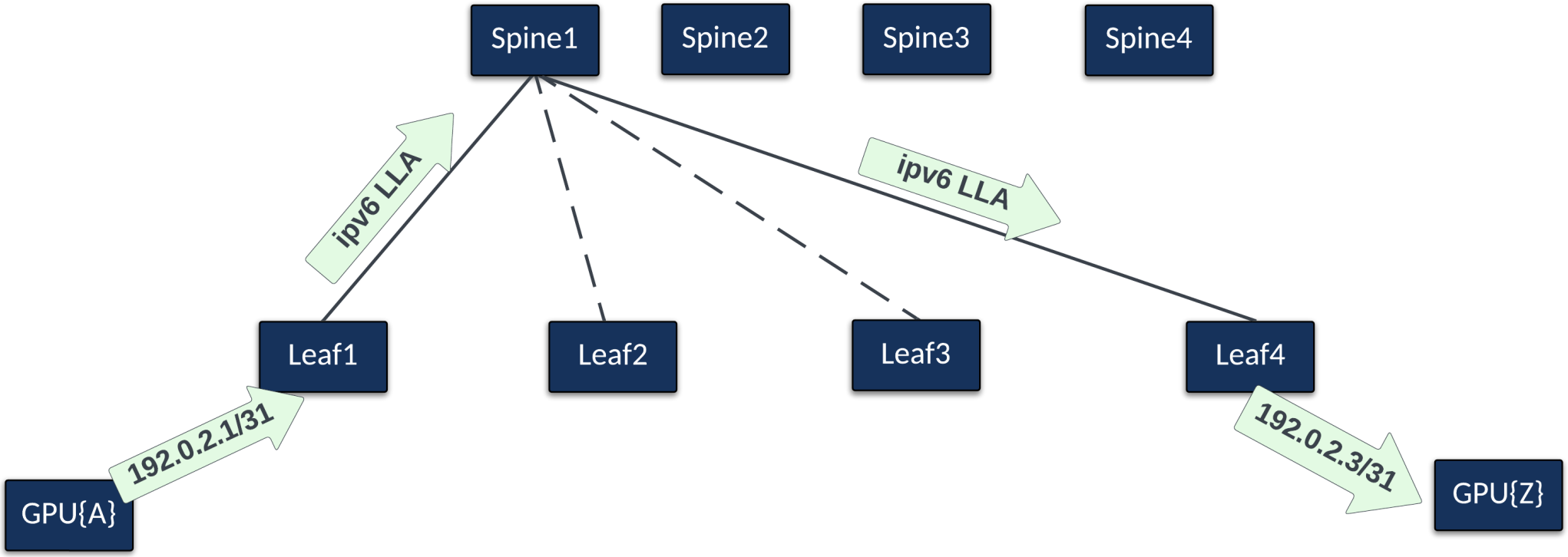
Scaling Topologies Down



NANOG

Routing in AI Fabrics

RFC-5549/8950 IPv4 over IPv6 Next-hops

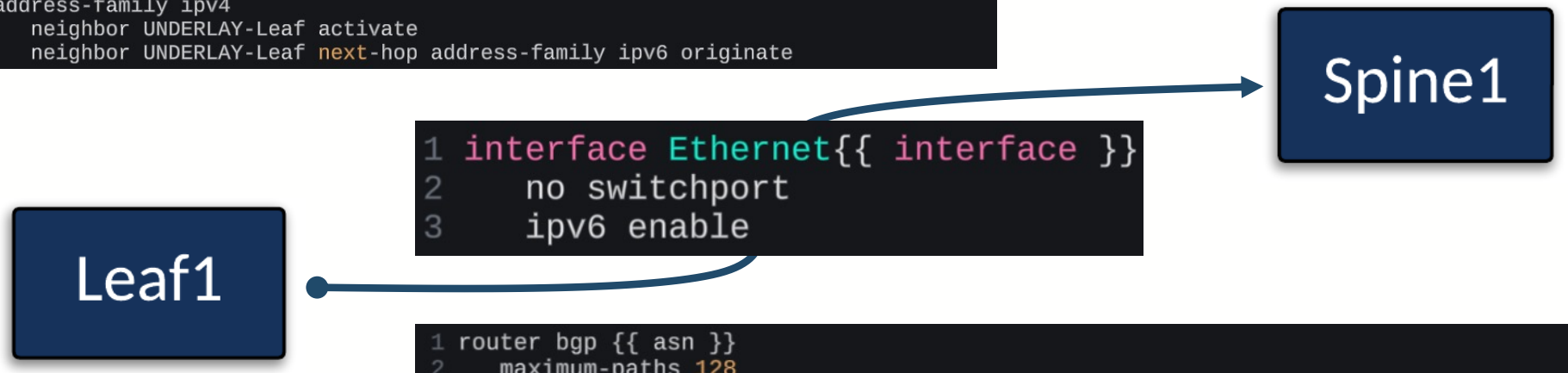


NANOG™

Routing in AI Fabrics

RFC-5549 Example

```
router bgp {{ asn }}
   bgp cluster-id {{ Spine_cluster_ID }}
   maximum-paths 128
   neighbor UNDERLAY-Leaf peer group
   neighbor UNDERLAY-Leaf next-hop-self
   neighbor UNDERLAY-Leaf route-reflector-client
   neighbor UNDERLAY-Leaf send-community
   neighbor interface {{ interfaces }} peer-group UNDERLAY-Leaf remote-as {{ asn }}
   !
   address-family ipv4
      neighbor UNDERLAY-Leaf activate
      neighbor UNDERLAY-Leaf next-hop address-family ipv6 originate
```

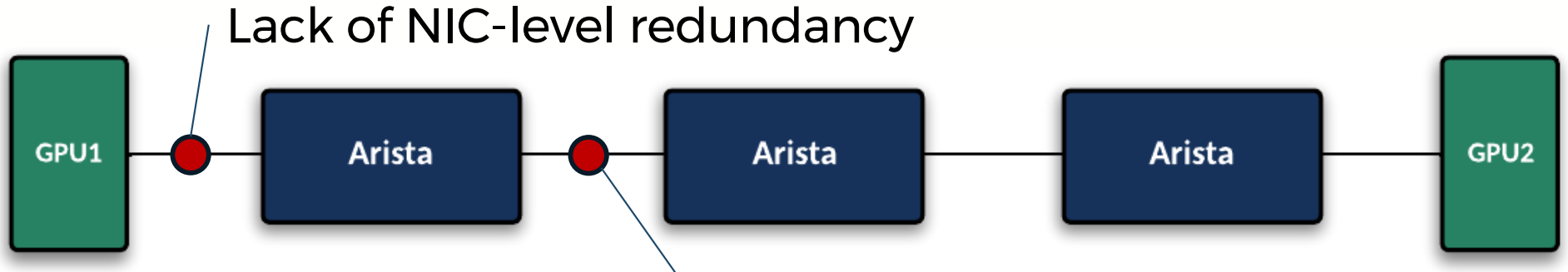



```
router bgp {{ asn }}
   maximum-paths 128
   neighbor UNDERLAY-Spine peer group
   neighbor UNDERLAY-Spine next-hop-self
   neighbor UNDERLAY-Spine send-community
   neighbor interface {{ interfaces }} peer-group UNDERLAY-Spine remote-as {{ asn }}
   !
   address-family ipv4
      neighbor UNDERLAY-Spine activate
      neighbor UNDERLAY-Spine next-hop address-family ipv6 originate
```

NANOG™

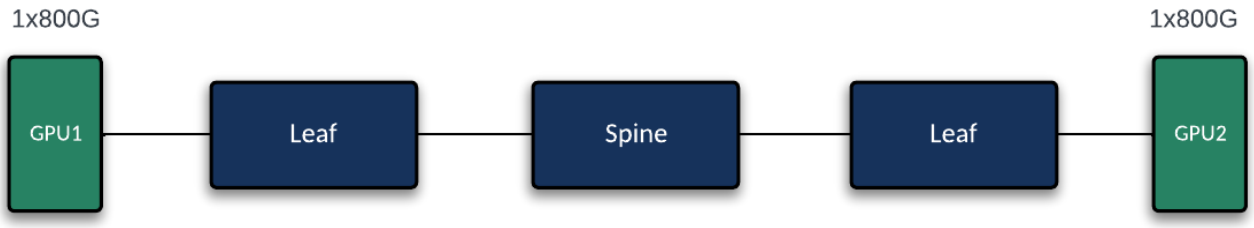
Introduction to Multi-Plane

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



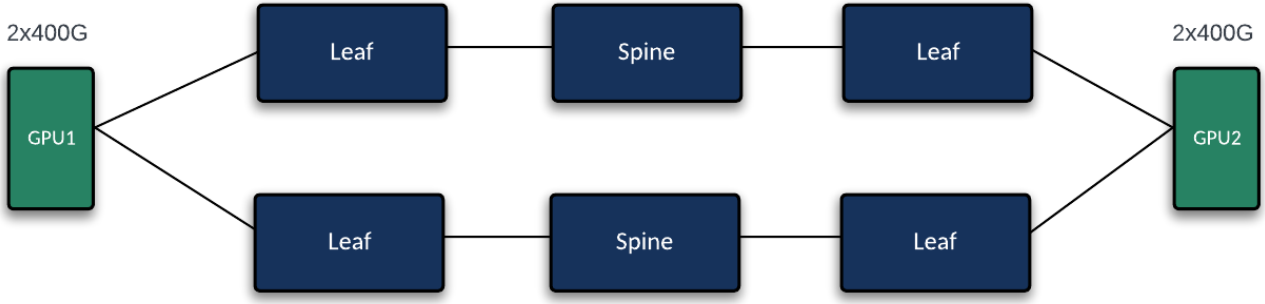
Higher speeds = lower radix.
64x800G Switch w/400G Uplinks in 2-tier = 8192 GPU
64x800G Switch w/800G Uplinks in 2-tier = 4096 GPU

Make like a banana and…



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

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

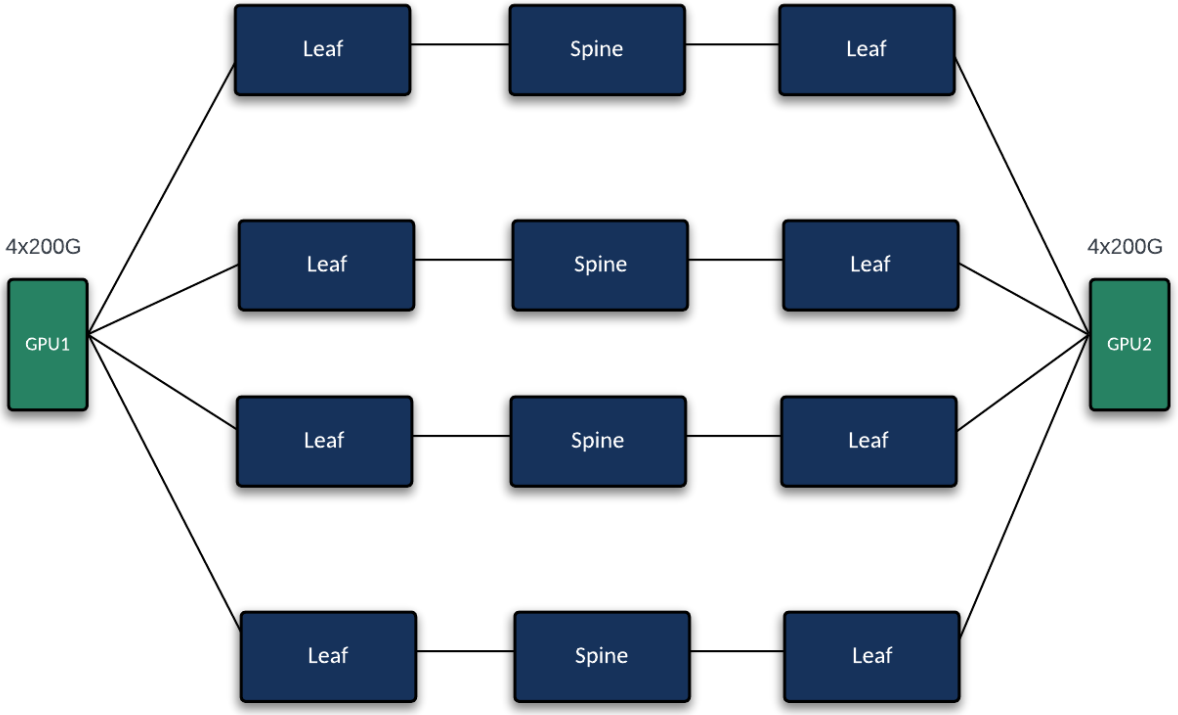


128x400G
•64 Spine per plane
•128 Leaf per plane (64 downlinks)

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

NANOG™

Taking it a step further...



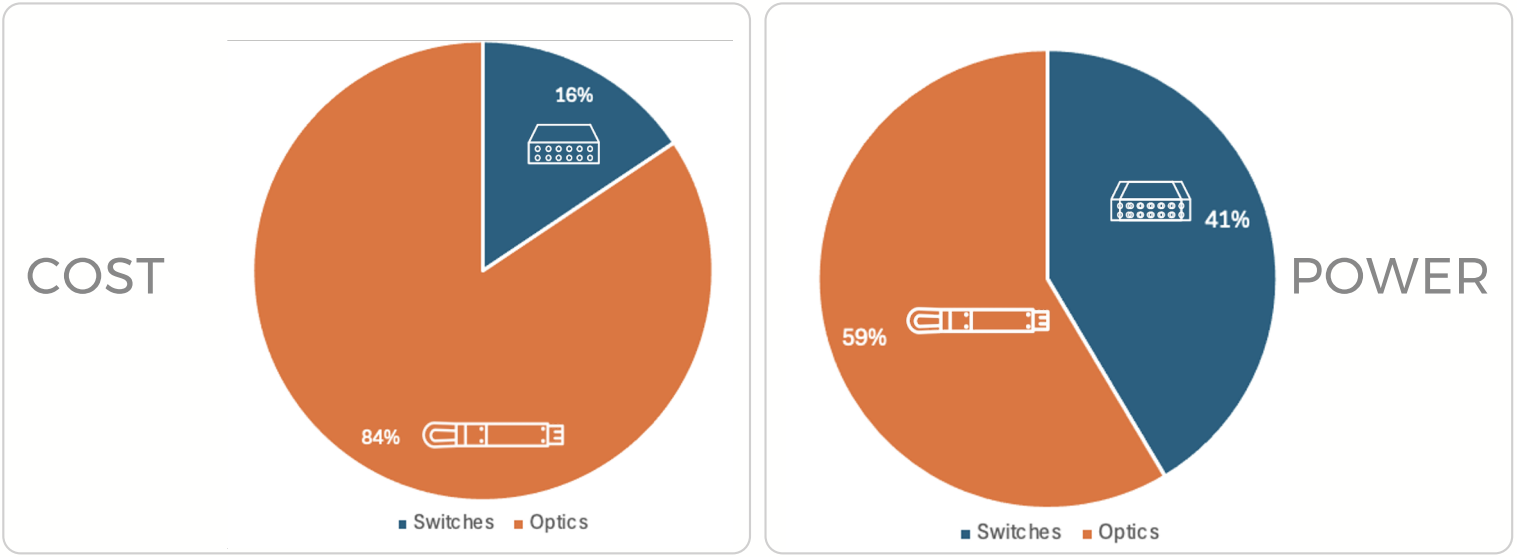
256x200G
•128 Spine per plane
•256 Leaf per plane (128 downlinks)

4x200G Max Scale (64x800G 2-tier) = 32768 GPU

Optics

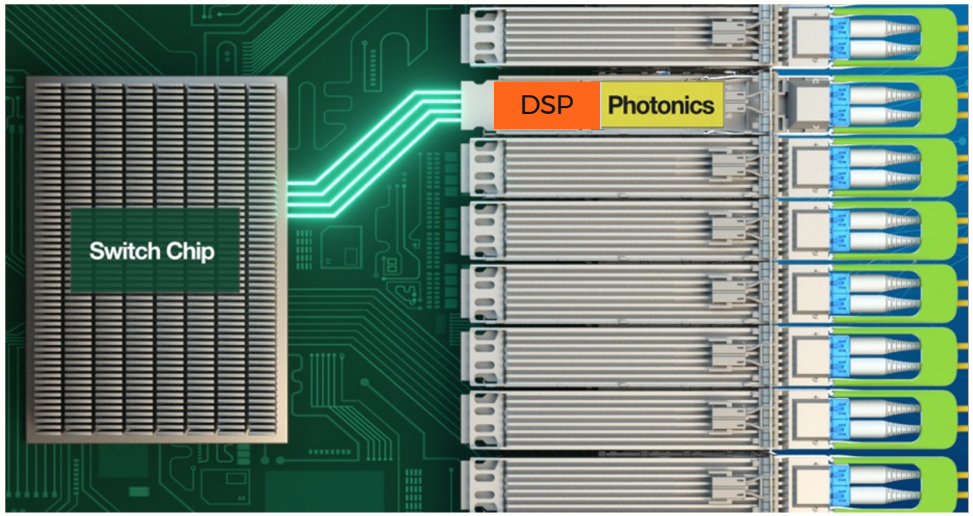
Optimizing optics

- Ex: 8k XPU ports / 2-tier AI-center infrastructure
  - 64x 64-port 800G AI Spine
  - 128x 64-port 800G AI Leaf
  - 16384x 800G optics
    - 8192x 800G optics (Fabric ports)
    - 8192x 800G optics (Host-facing ports)

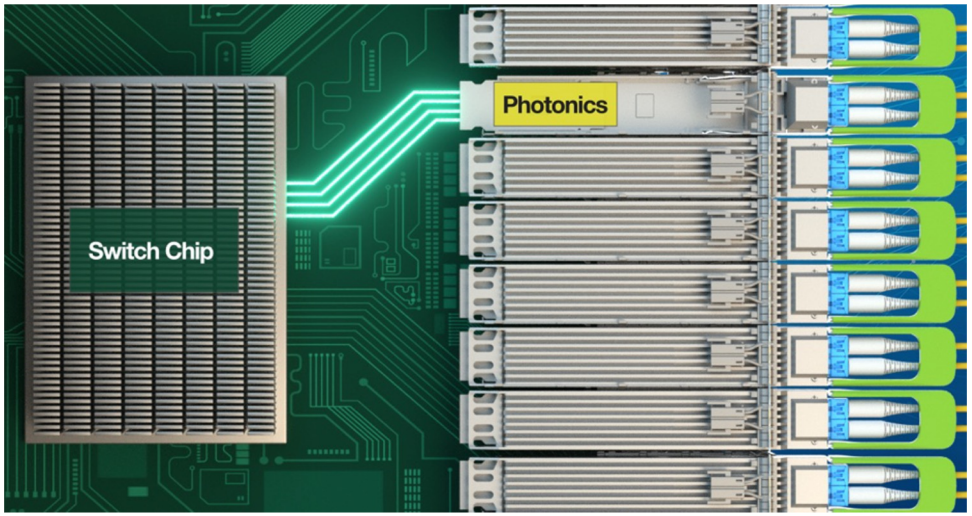


Linear-drive Pluggable Optics (LPO)

Traditional pluggable optical modules



LPO modules

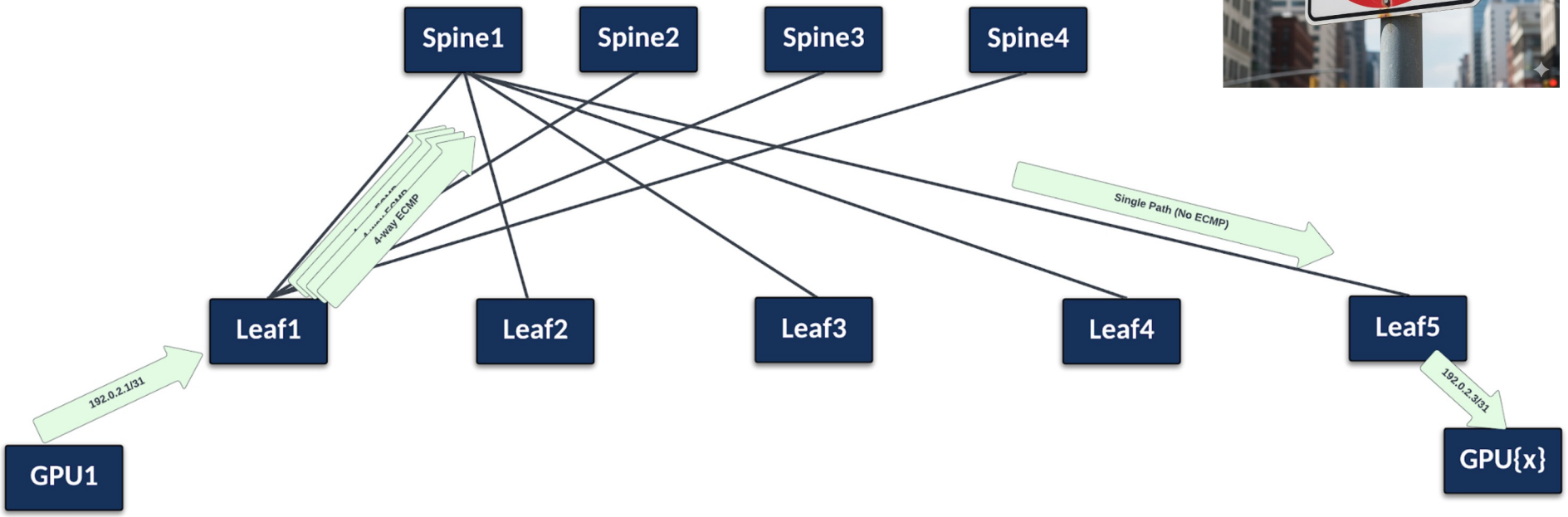


- LPO means no DSP in the optic module
- How is this possible ?
  - Certain switch silicon has advanced DSP technology on-chip
  - Requires careful system design and SerDes tuning
- Lower power (~0.5x), cost (~0.7x) and latency (~0.01x) with higher reliability

NANOG™

Load Balancing / Congestion Control

Equal Cost MultiPath (ECMP) / Fan-Out



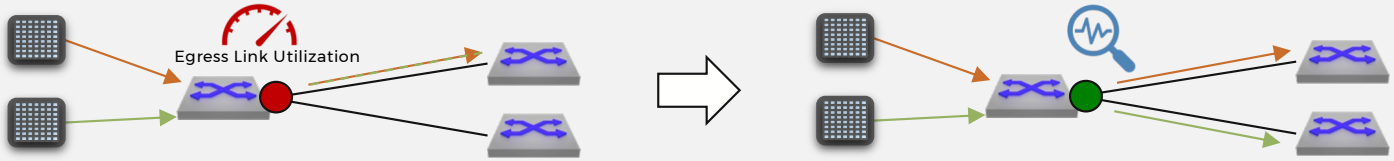
NANOG™

Load-Balancing Options

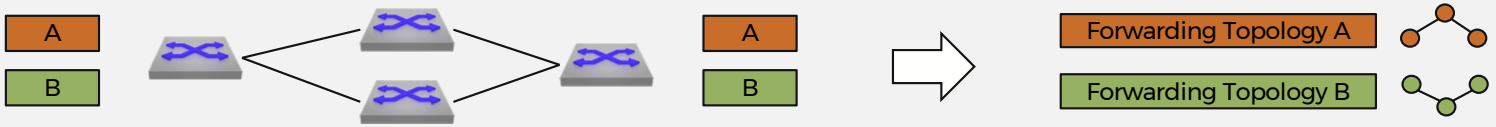
- Classic ECMP hashing:

- RDMA aware static ECMP hashing:

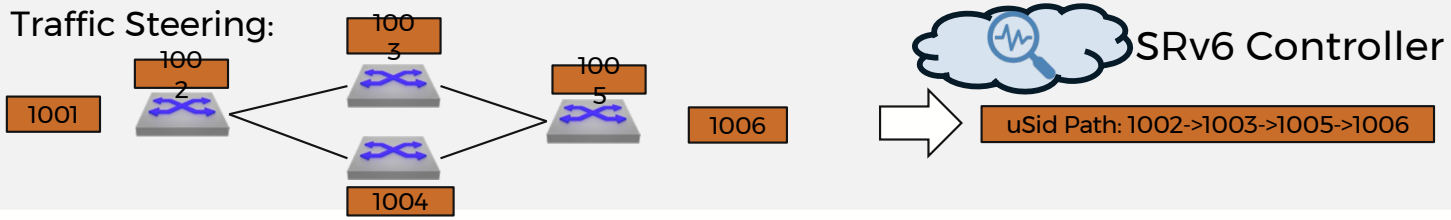
- Dynamic load balancing:



- Cluster Load Balancing:

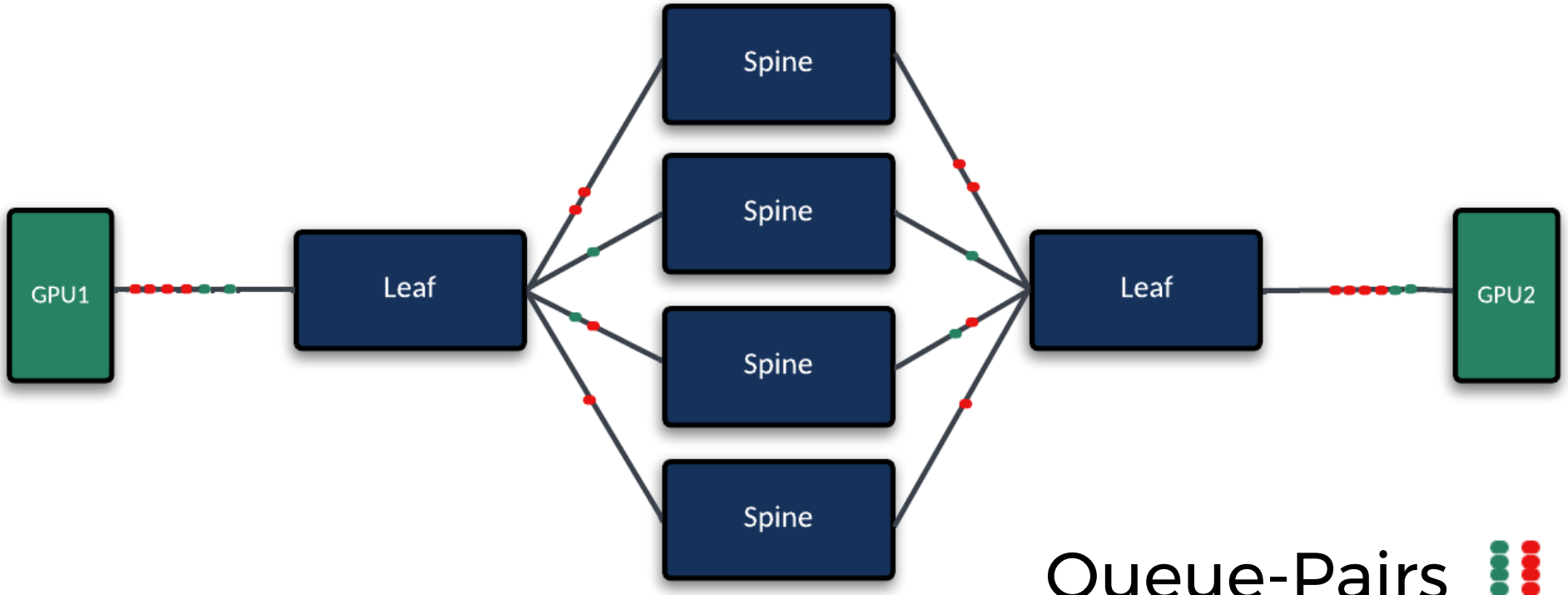


- Traffic Steering:



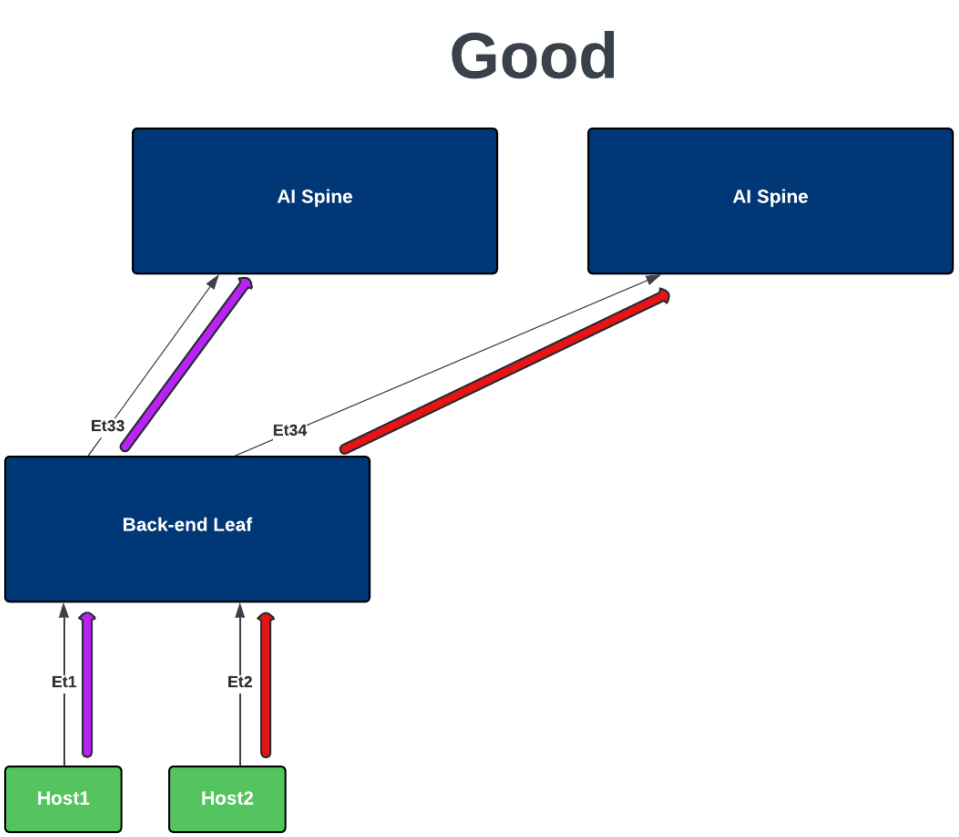
NANOG™

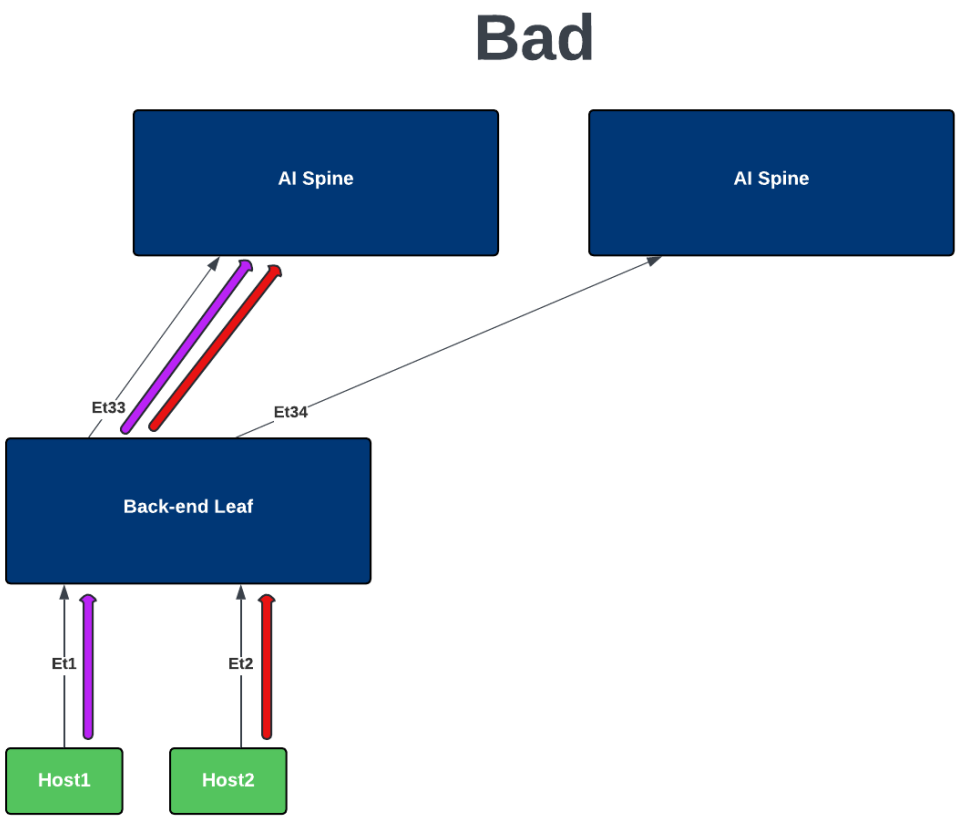
Packet Spraying



Queue-Pairs

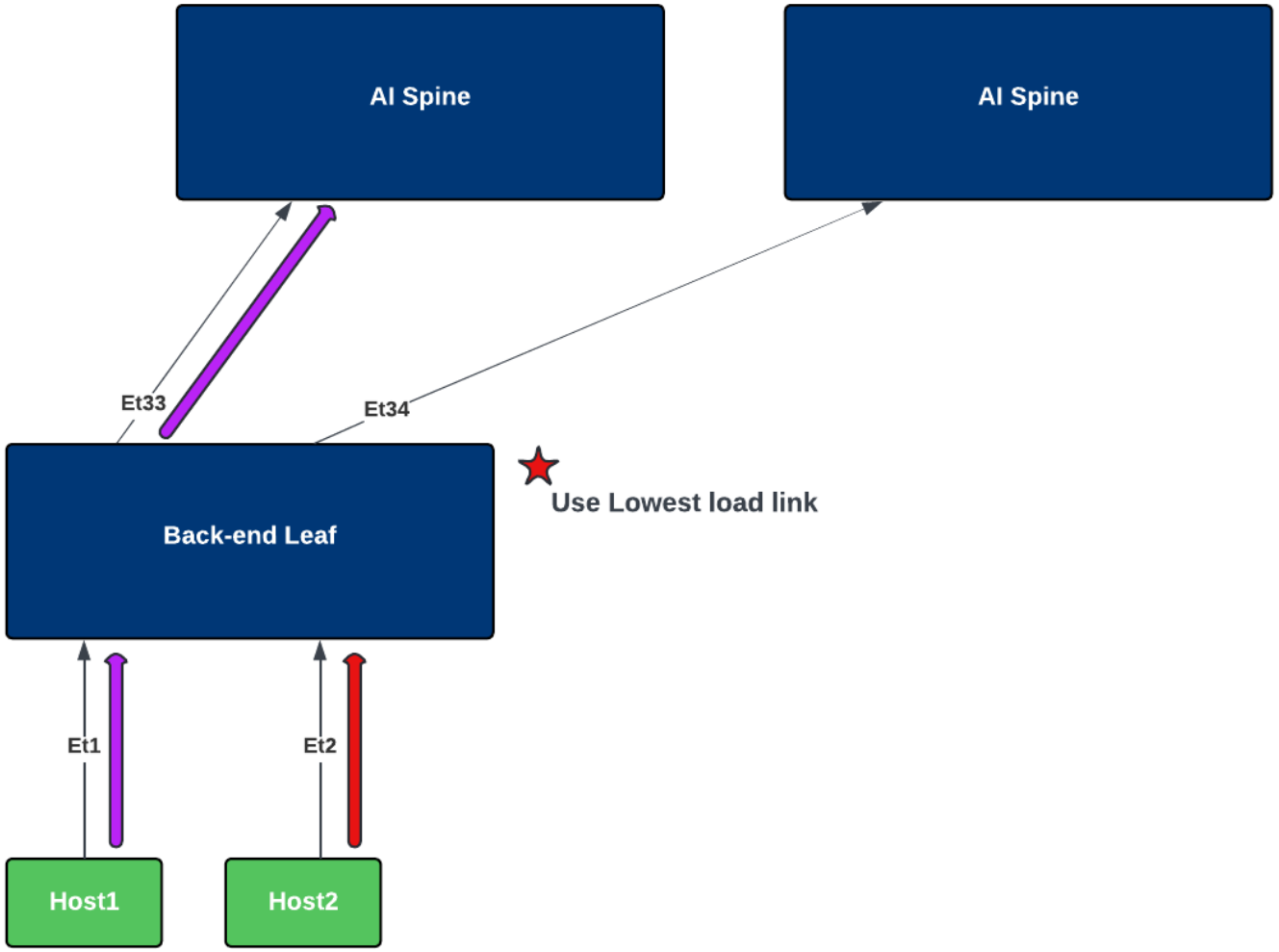
ECMP Hashing





NANOG™

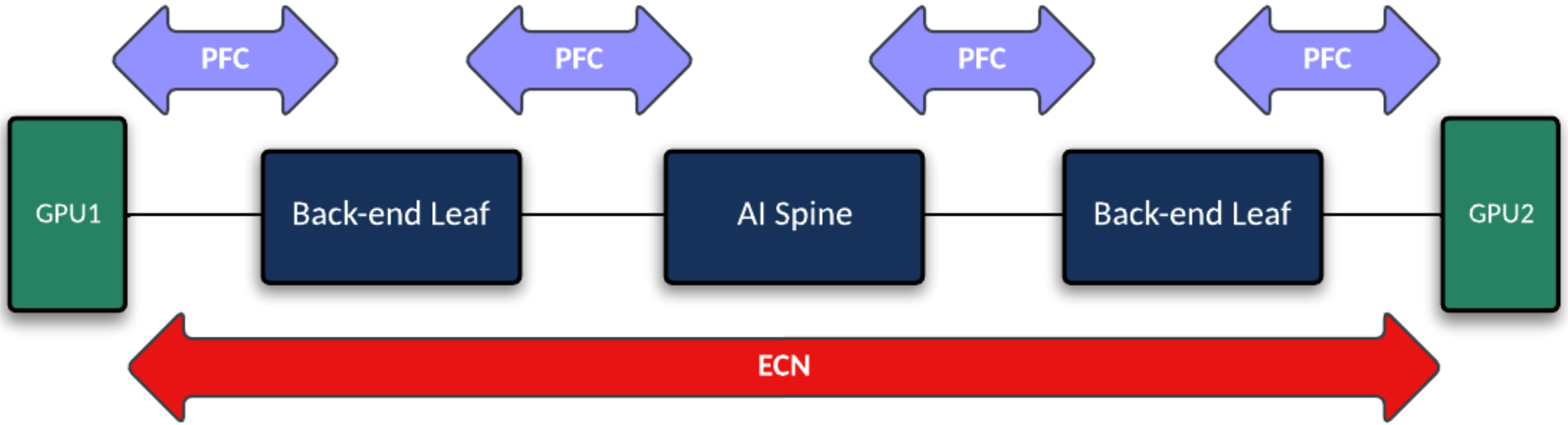
Dynamic Load-Balancing (DLB)



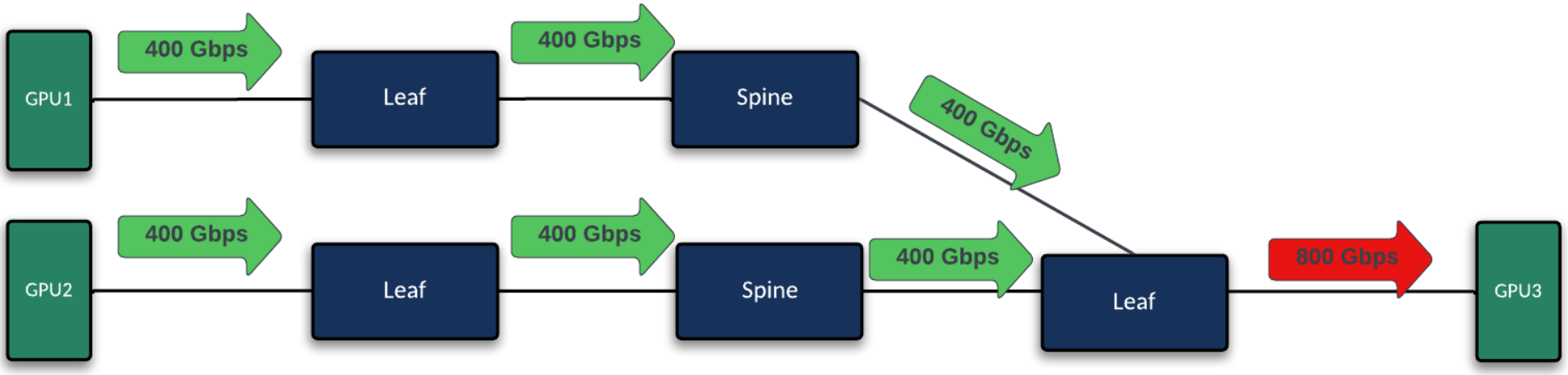
RDMA / RoCEv2

RDMA - Remote Direct Memory Access
DCQCN - Datacenter Quantized Congestion Notification
RoCEv2 – RDMA over Converged Ethernet (version 2)
PFC – Priority Flow Control
ECN – Explicit Congestion Notification

RDMA / RoCEv2


PFC
PFC
PFC
PFC
GPU1
Back-end Leaf
AI Spine
Back-end Leaf
GPU2
ECN

PFC

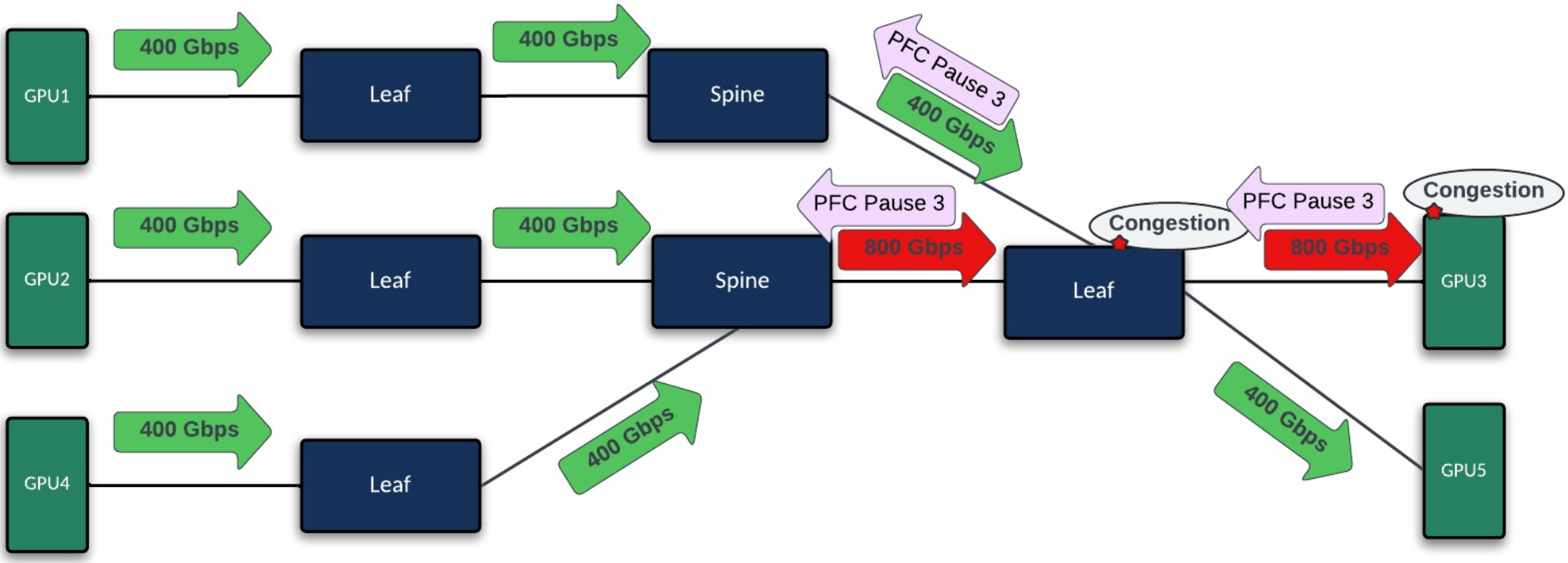


Ingress Queue

Egress Queue

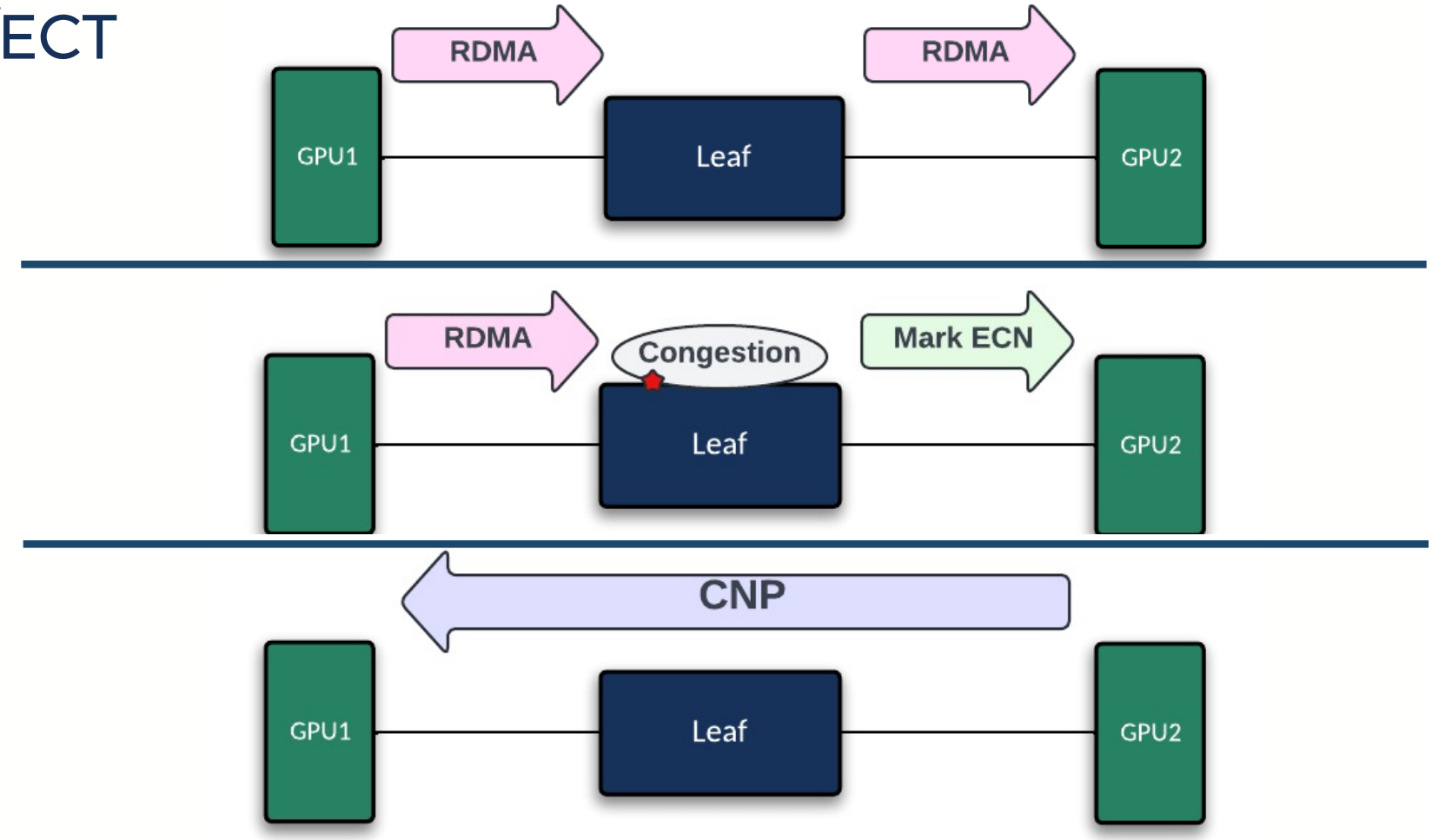
Queue Backpressure in Egress causes Ingress to start building.

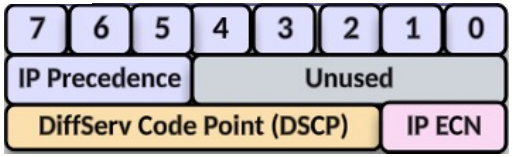
PFC - Victim Flows



NANOG™

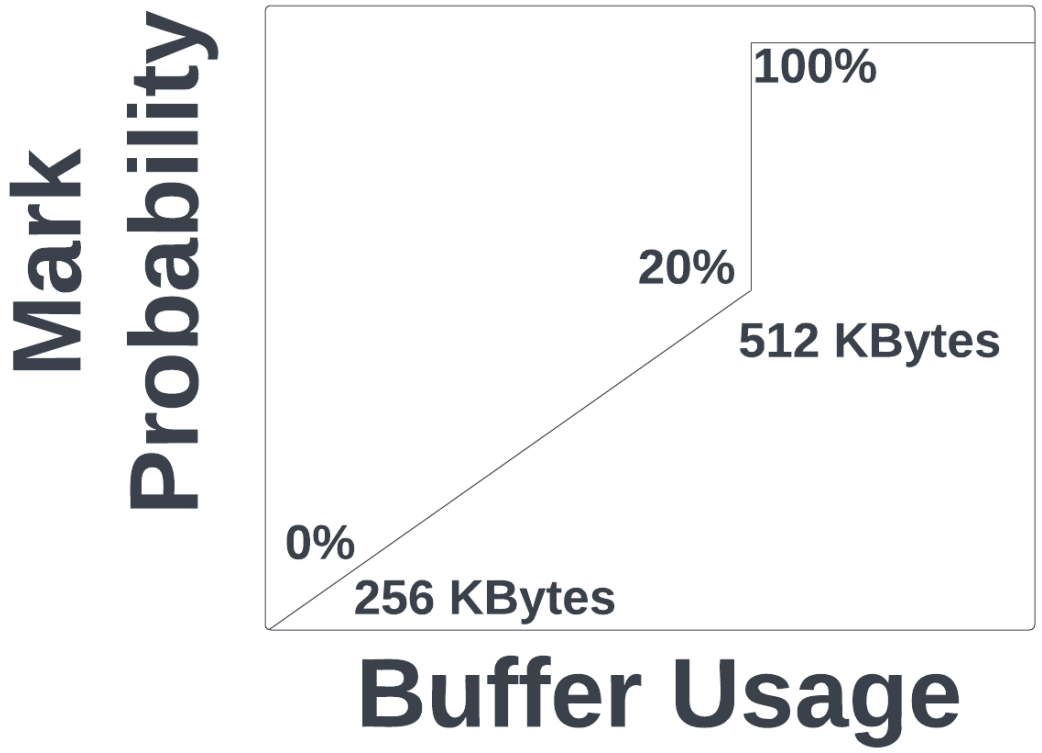
ECN/ECT



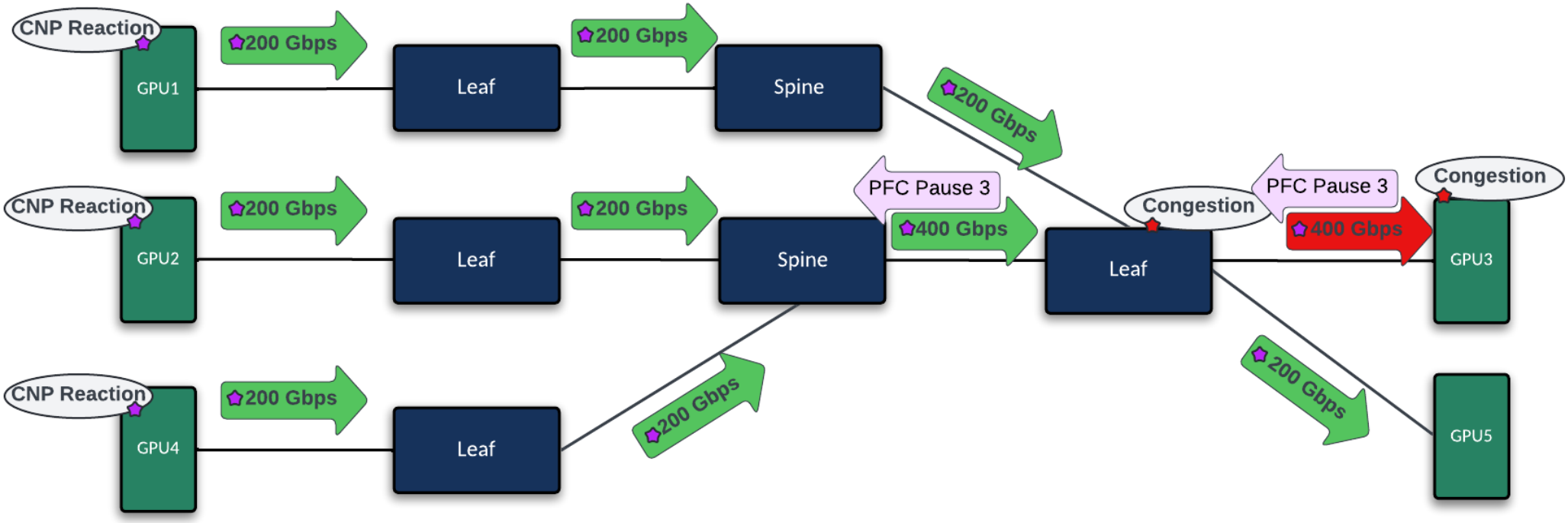


NANOG™

Probabilistic ECN



PFC w/ECN



NANOG™

Quality of Service

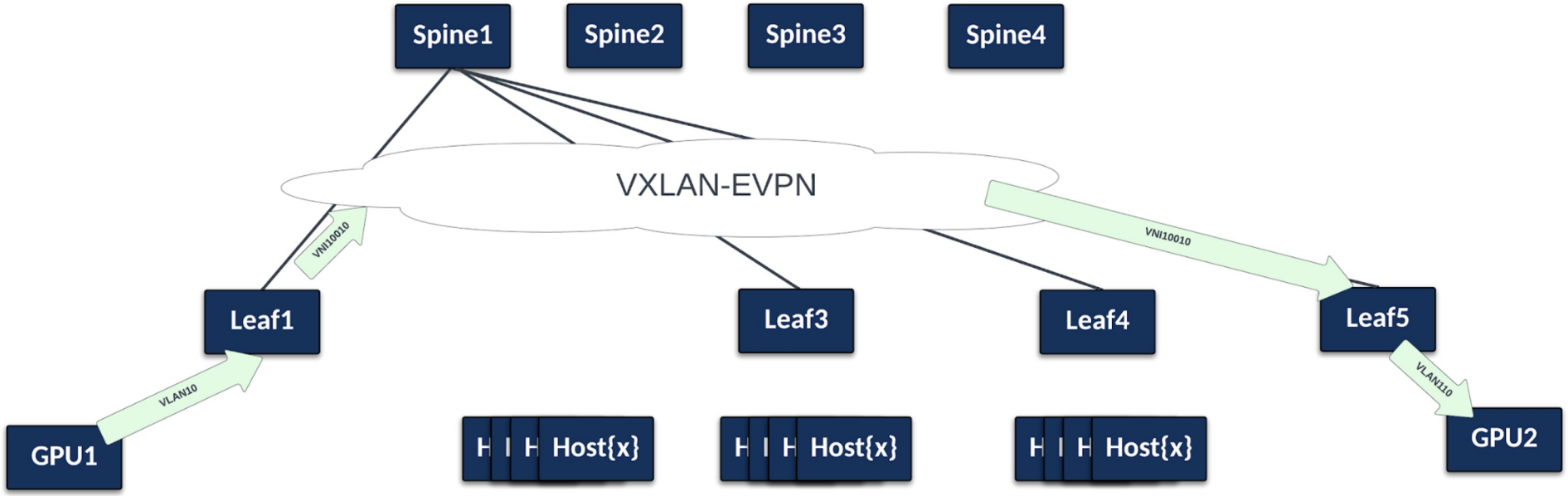
QoS Basics for AI

SP Queue 6/7 - CNP
SP Queue 4 - RDMA Control
WRR Lossless Queue 3 - RDMA Data
WRR Queue 1 - Everything Else

Segmentation

Segmentation in AI Fabrics

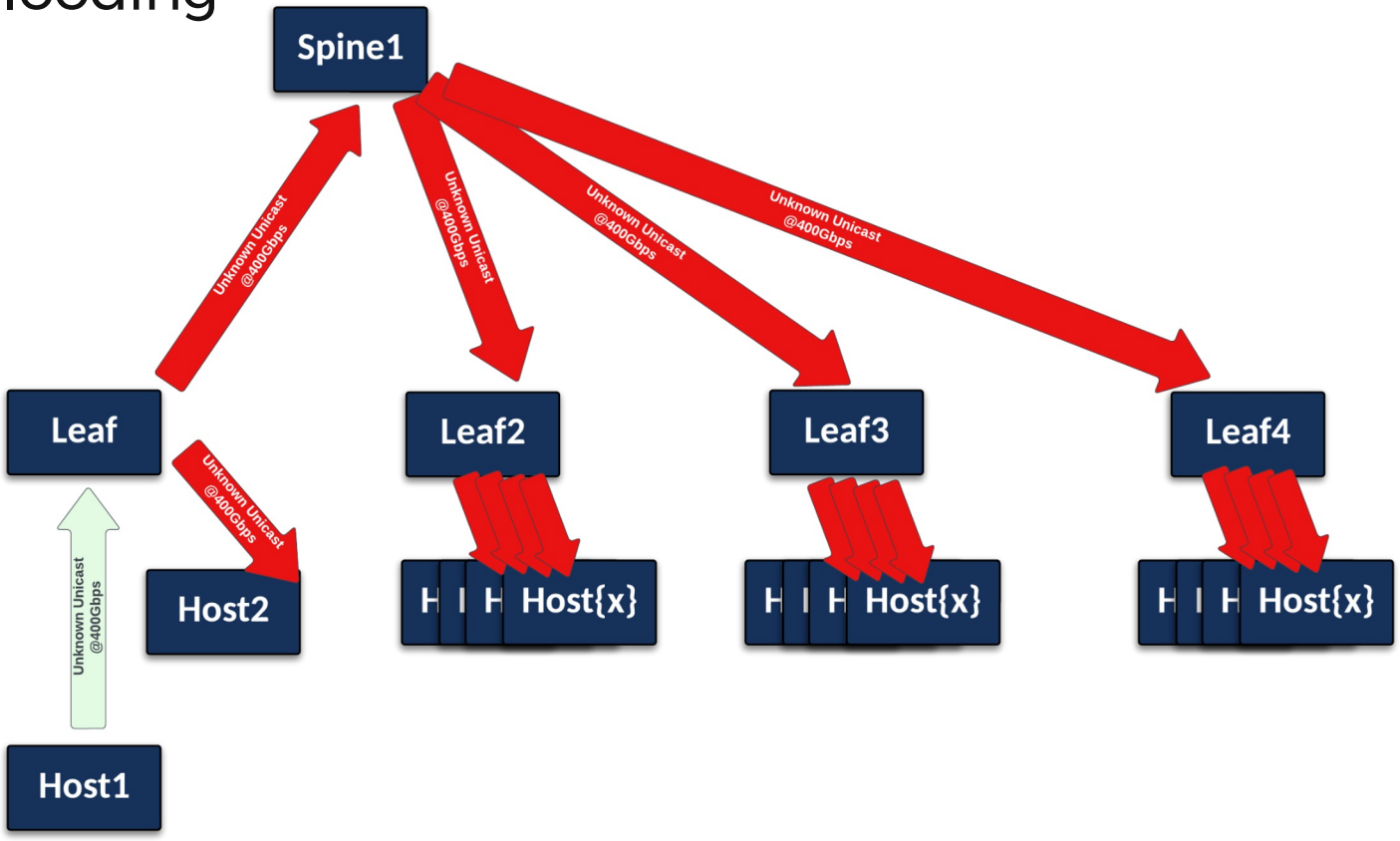
VLANs - Encapsulation



NANOG™

Segmentation in AI Fabrics

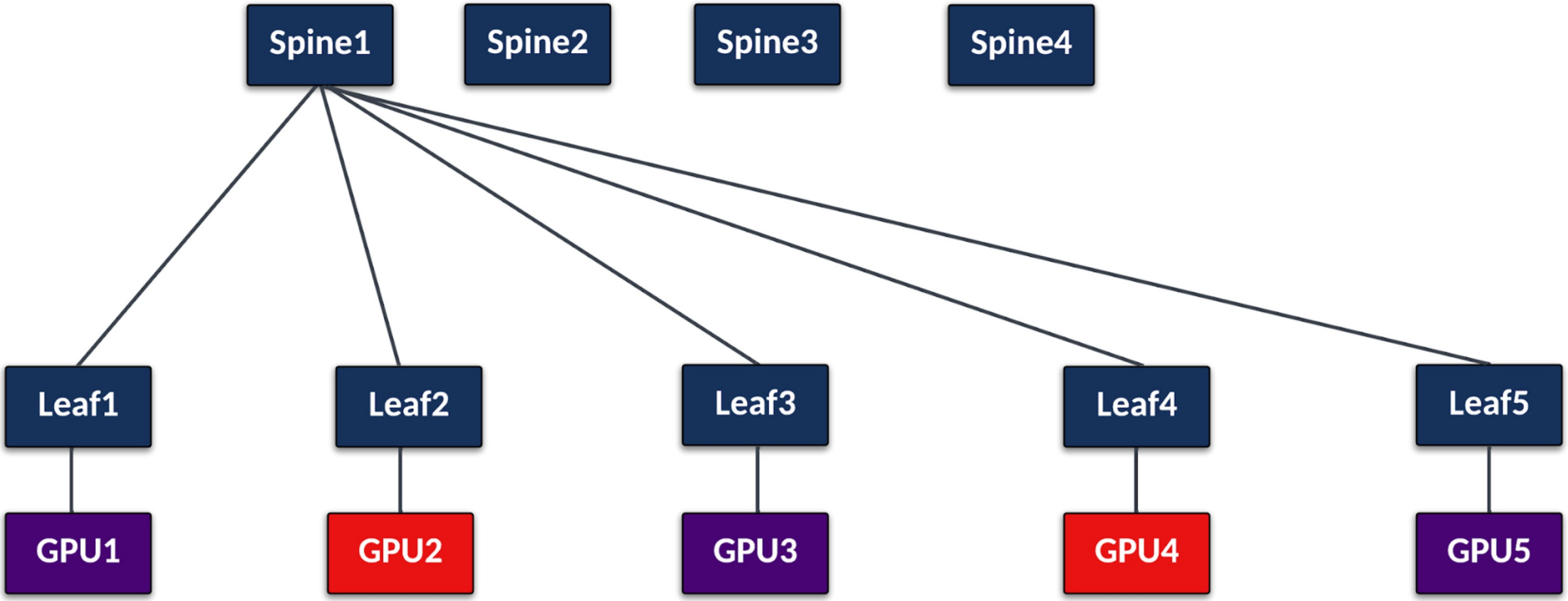
L2 Flooding



NANOG™

Segmentation in AI Fabrics

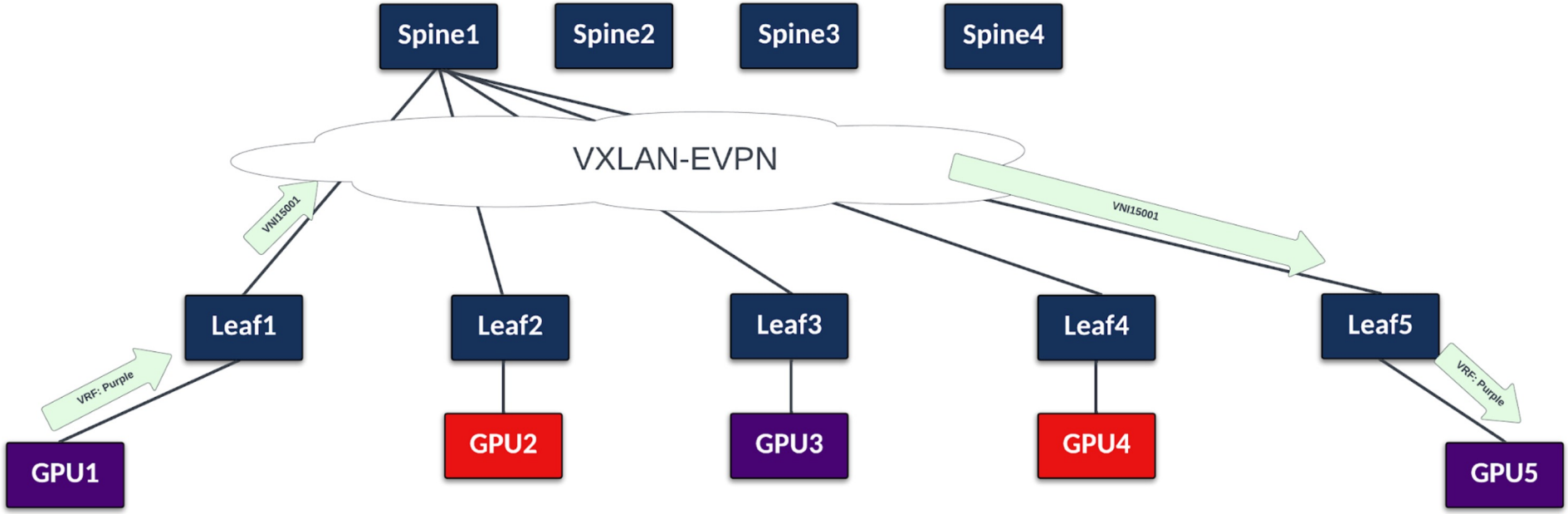
VRFs



NANOG™

Segmentation in AI Fabrics

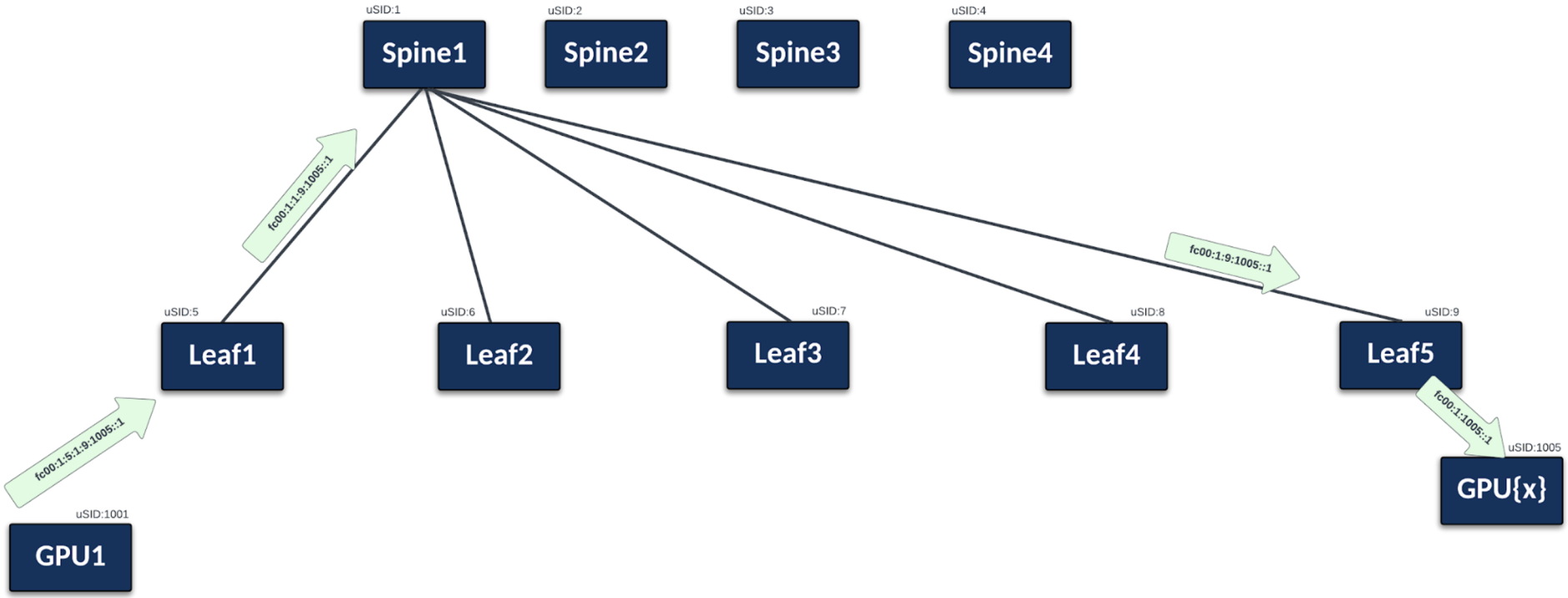
VRFs - Encapsulation



NANOG™

Segmentation in AI Fabrics

SRv6 uSID



Containerlab Topology:
https://github.com/brokenpackets/clab_Topos/tree/main/srv6_uSID

NANOG™

Common Problems

Common Problems - Networking

QoS Mismatches

Incorrect Cabling (8k cluster = ~16-20k optics)

Dirty Fiber/Optics

NIC Speed/FEC/Autoneg mismatch

MMU Tuning

Congestion Control Tuning

Soft/Hard Failures

Optics

For a hypothetical 100k GPU build:

- Total Links: 257,904
- Total Optics: 515,808

For JUST the Back-end.

Assuming a MTBF of 2.3M hours on optics...

2.3m hours / 515808 = 4.45 hours between failures. ~5 failures per day

Key Takeaways

Front-end is likely similar to how you're building Datacenter networks today, albeit with potentially higher speeds/subscription ratios.

Back-end (usually) requires a dedicated network and will require net-new equipment; re-use of existing gear is unlikely. Radix, speed, and subscription ratios are usually the largest drivers in back-end networking design.

Power/Heat are usually the primary constraint in building these networks, followed closely by capex dollars for all the gear required. If you can solve for power, others will solve the capex problem for you.

Careful monitoring of the network is a hard requirement. There are a large number of potential failure points, and a slow job completion time or job failure has a very high impact.

THANK YOU

ARISTA