

Why Cisco for AI/ML Infrastructure Networking

cisco Live !

Paresh Gupta
Principal Technical Marketing Engineer
Cisco

Cisco Webex App

Questions?

Use Cisco Webex App to chat with the speaker after the session

How

- 1 Find this session in the Cisco Live Mobile App
- 2 Click “Join the Discussion”
- 3 Install the Webex App or go directly to the Webex space
- 4 Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until 14 November 2025.



<https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-1226>

Agenda

- 01 AI cluster networks
- 02 Challenges and solutions
- 03 RoCEv2, PFC, ECN
- 04 GPU collective communication (NCCL, etc.)
- 05 Load-balancing Schemes
- 06 Cisco + NVIDIA Spectrum-X integration
- 07 Ultra Ethernet

This is a 1000 series basic session explaining networks of an AI cluster.
This session does not cover using AI, large-language models (LLM), or ChatGPT.

AI Clusters

Hyperscalers

Neoclouds

Sovereign Clouds

Enterprises

10s or 100s of thousands of GPUs

Scale

Hundreds or thousands of GPUs

Training, fine-tuning, and inference

Use case

Fine-tuning and inference

Scale-up, scale-out, and scale-across GPU clusters

Locality

Scale-up and scale-out GPU clusters

Significant research and development resources

R & D

Minimal research and development resources

Custom-built software

Software

Vendor-developed and supported software

Only a few in the world

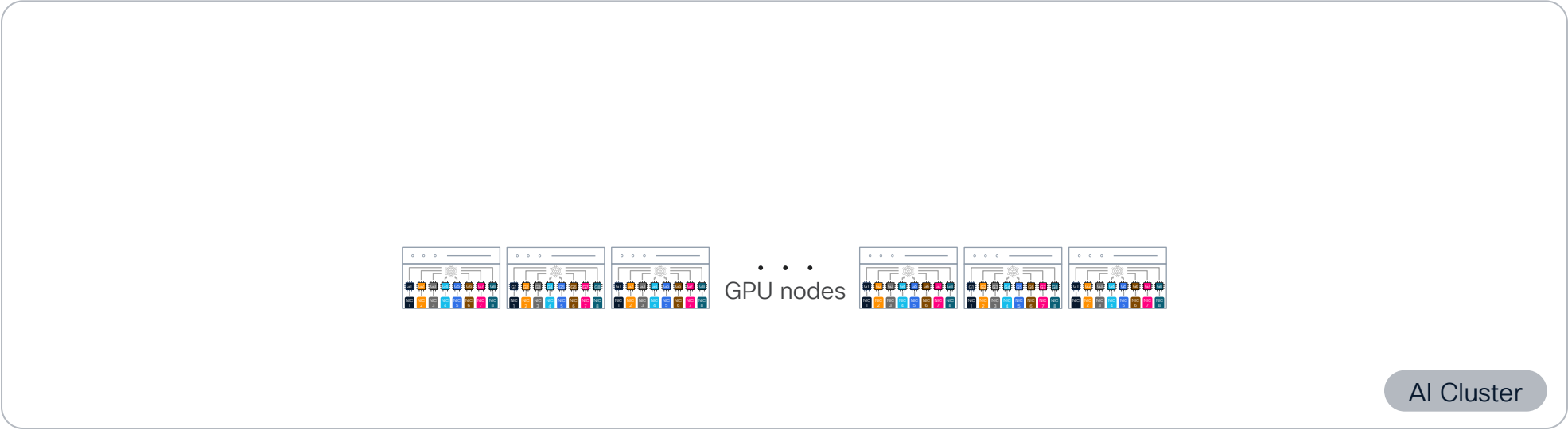
Opportunity

Thousands in the world

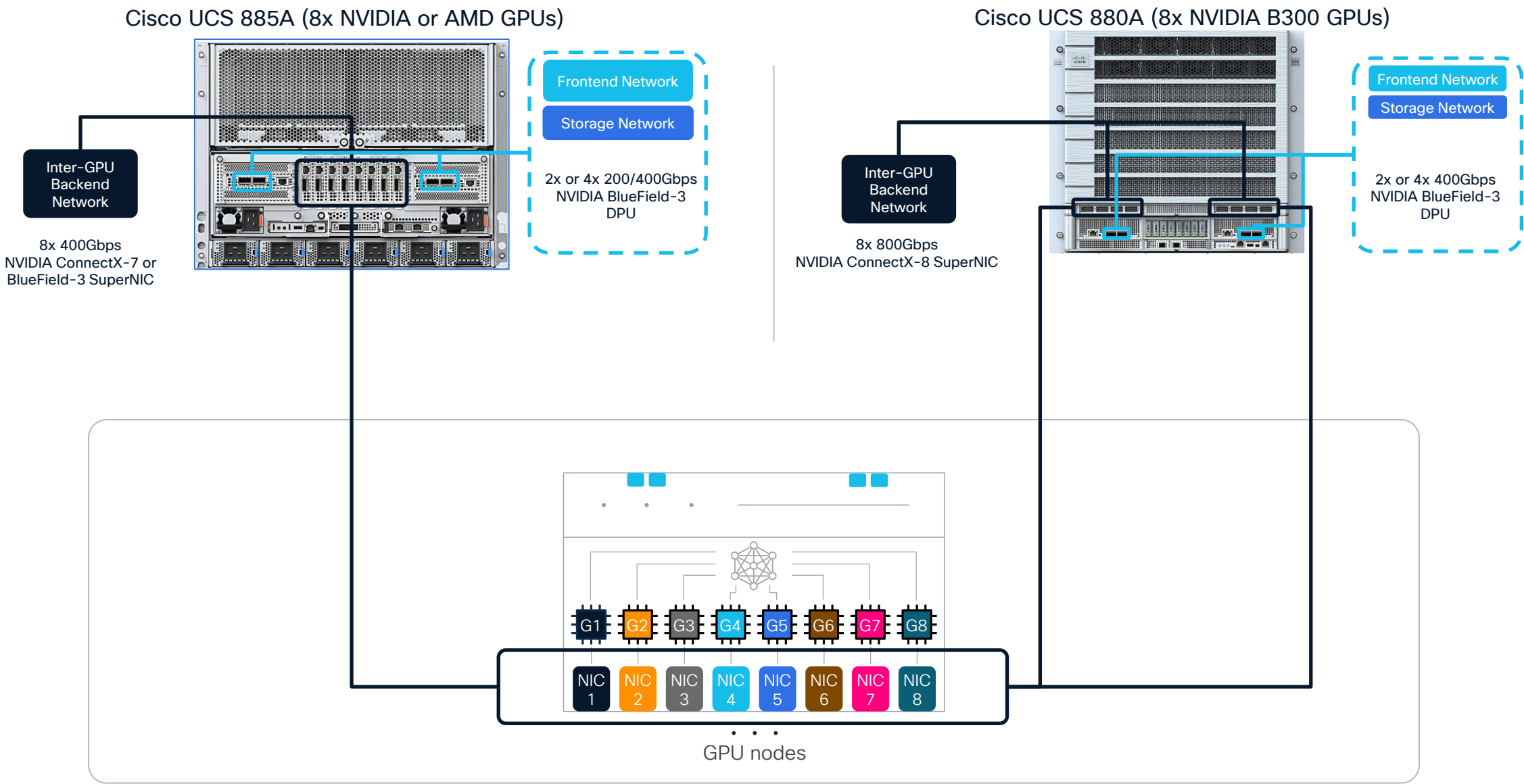
Environment are different,
Challenges are different, and therefore
Solutions are different



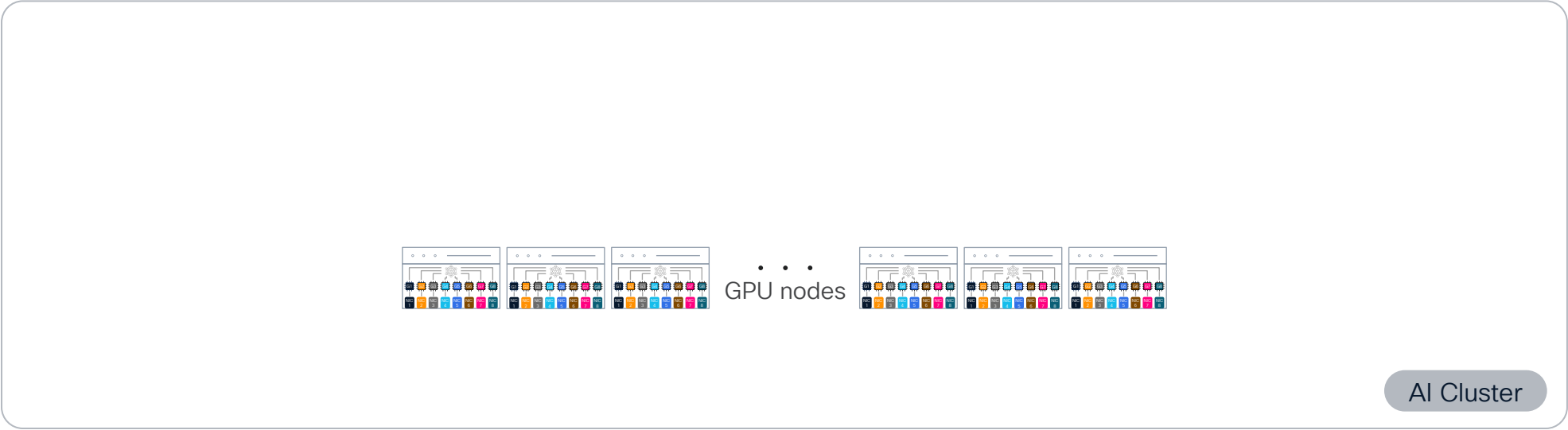
AI Cluster



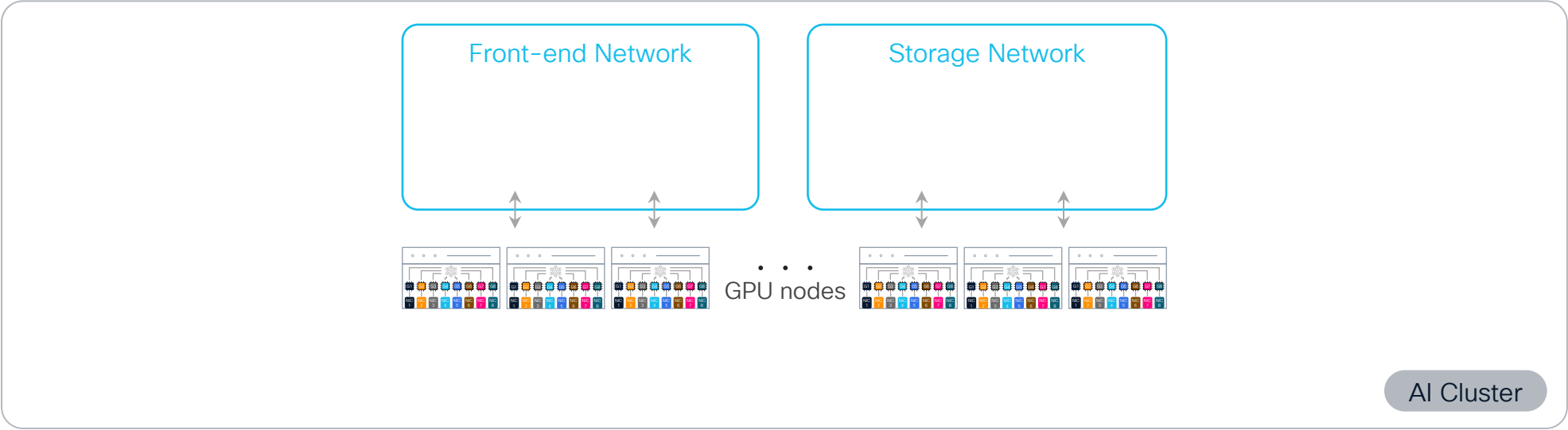
GPU Node Examples – Cisco UCS 885A and UCS 880A



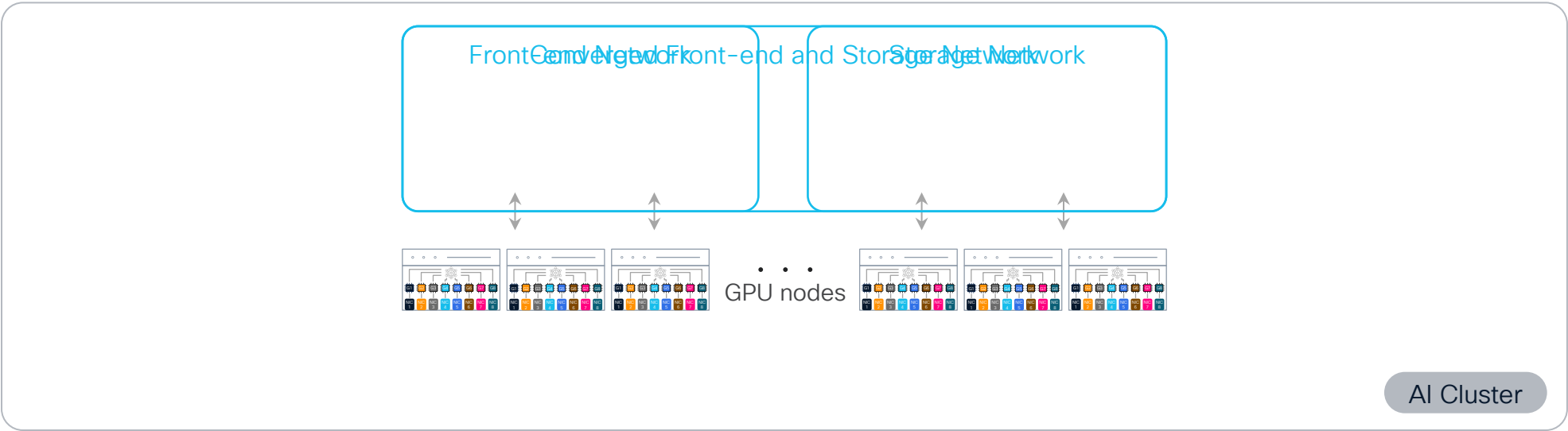
AI Cluster



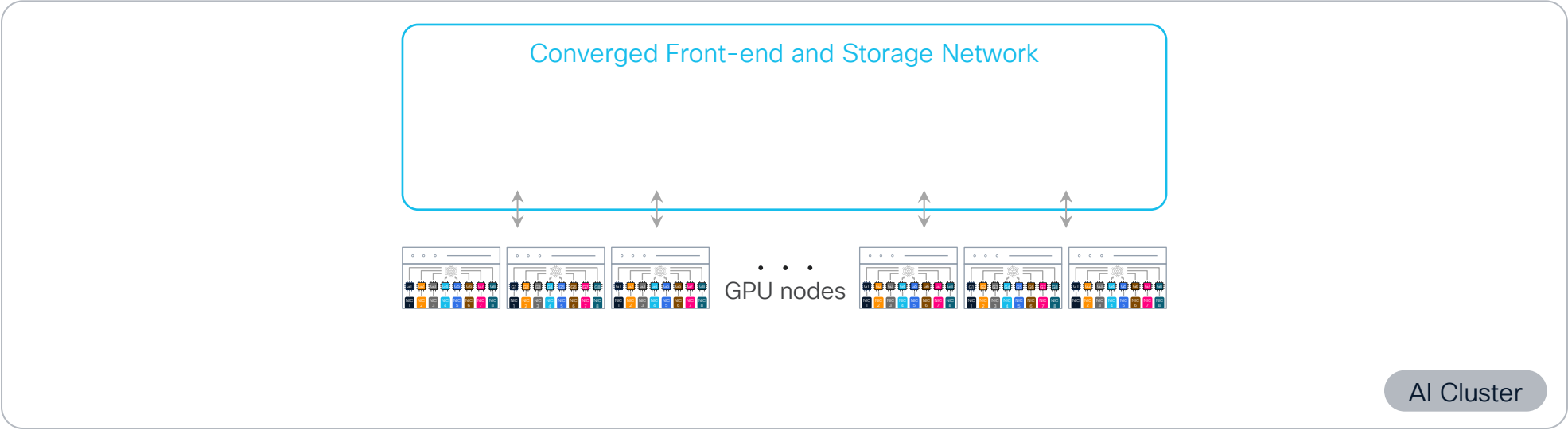
AI Cluster



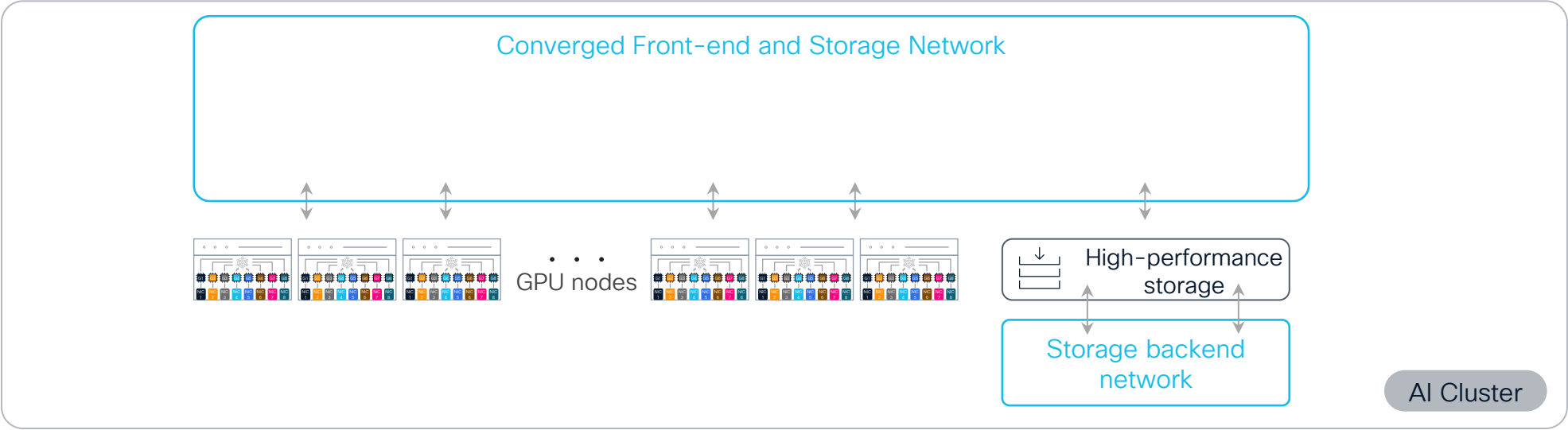
AI Cluster



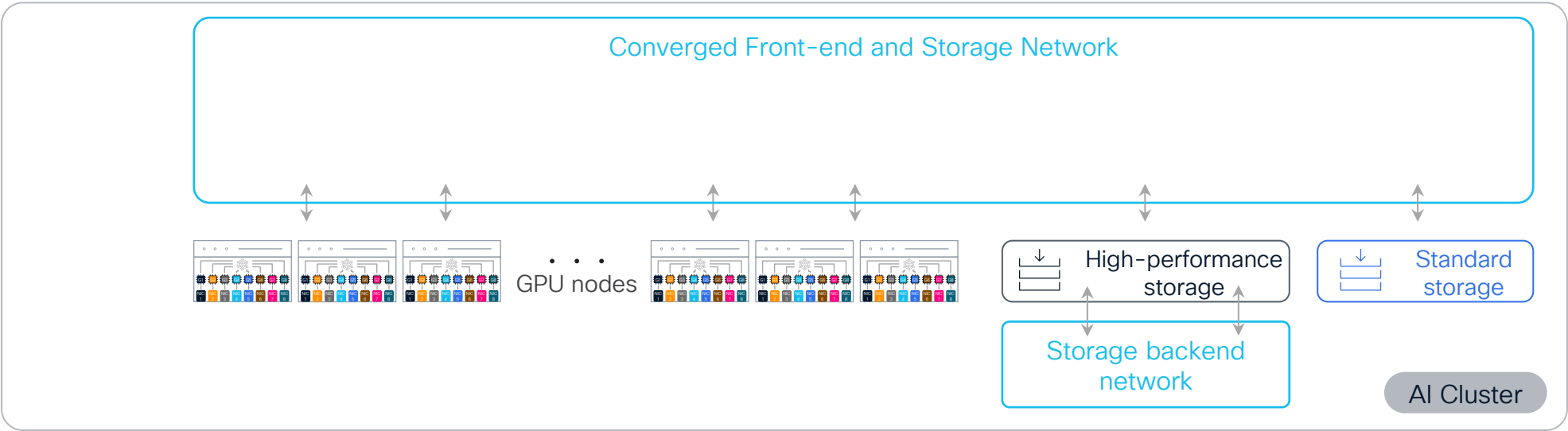
AI Cluster



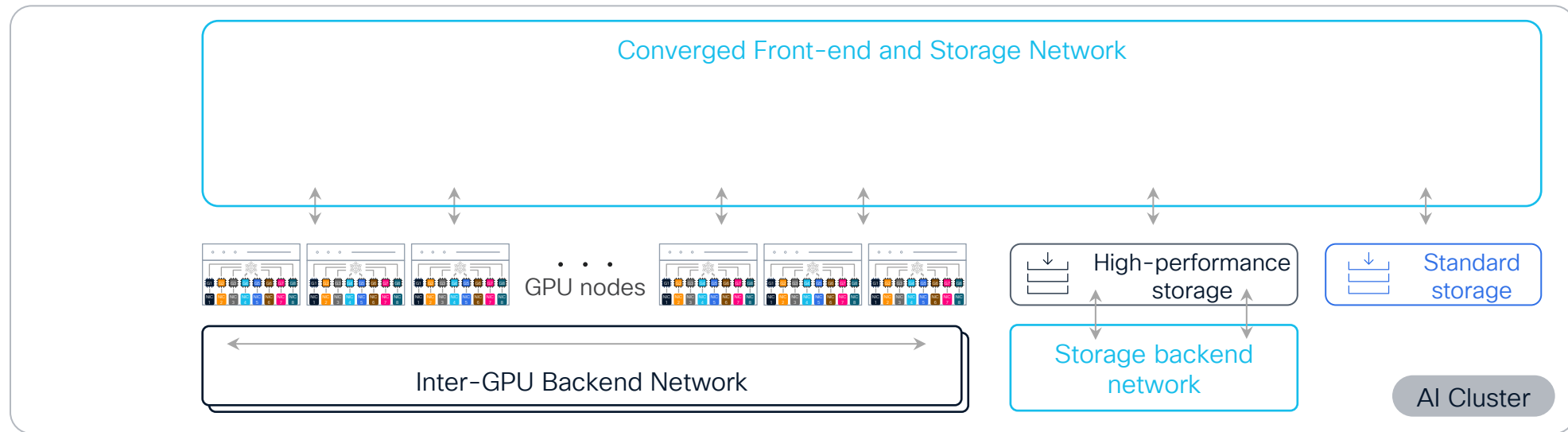
AI Cluster



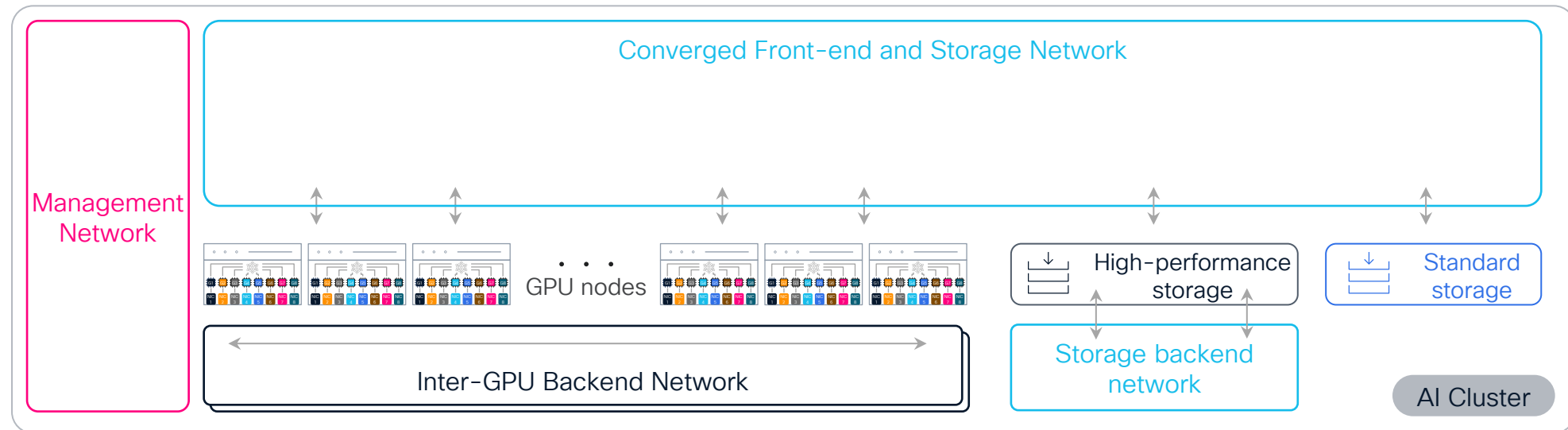
AI Cluster



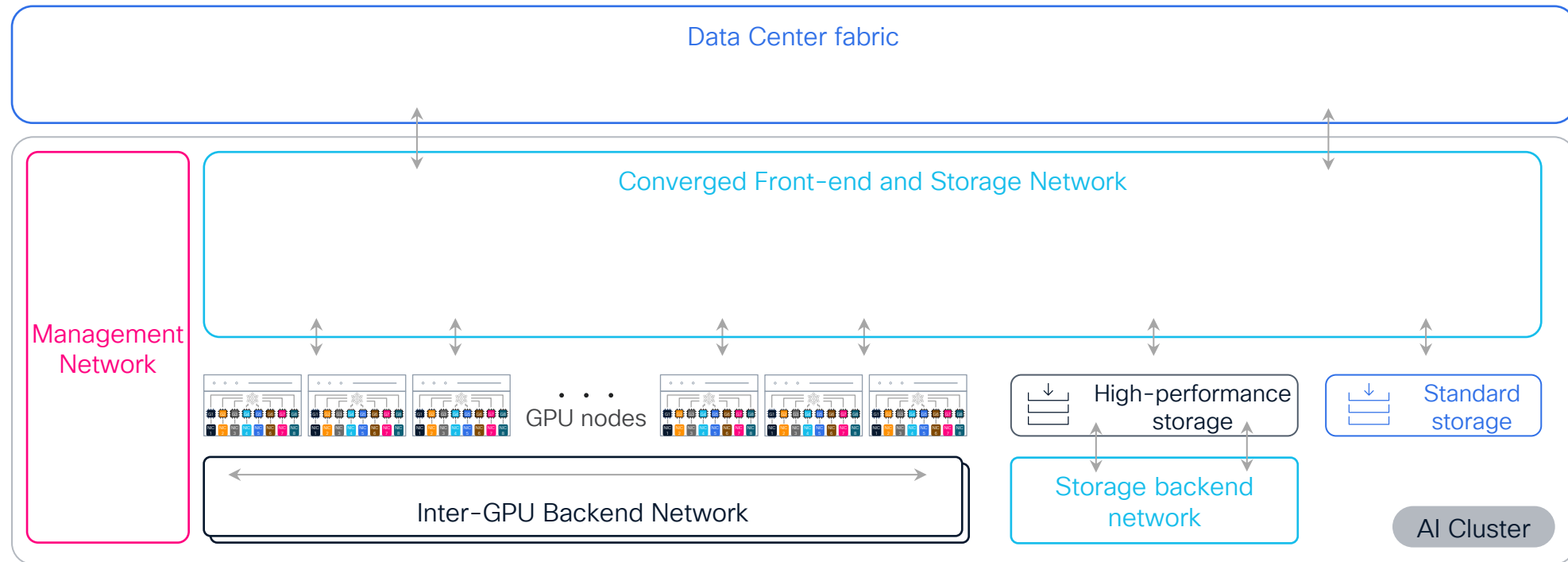
Multiple Networks of an AI Cluster



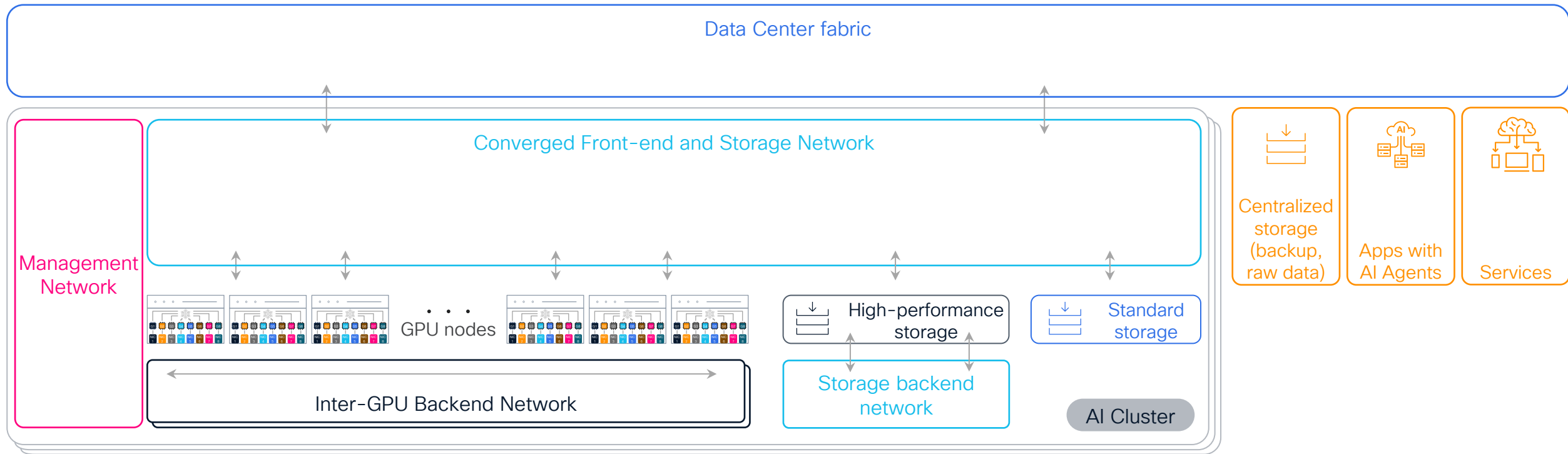
Multiple Networks of an AI Cluster



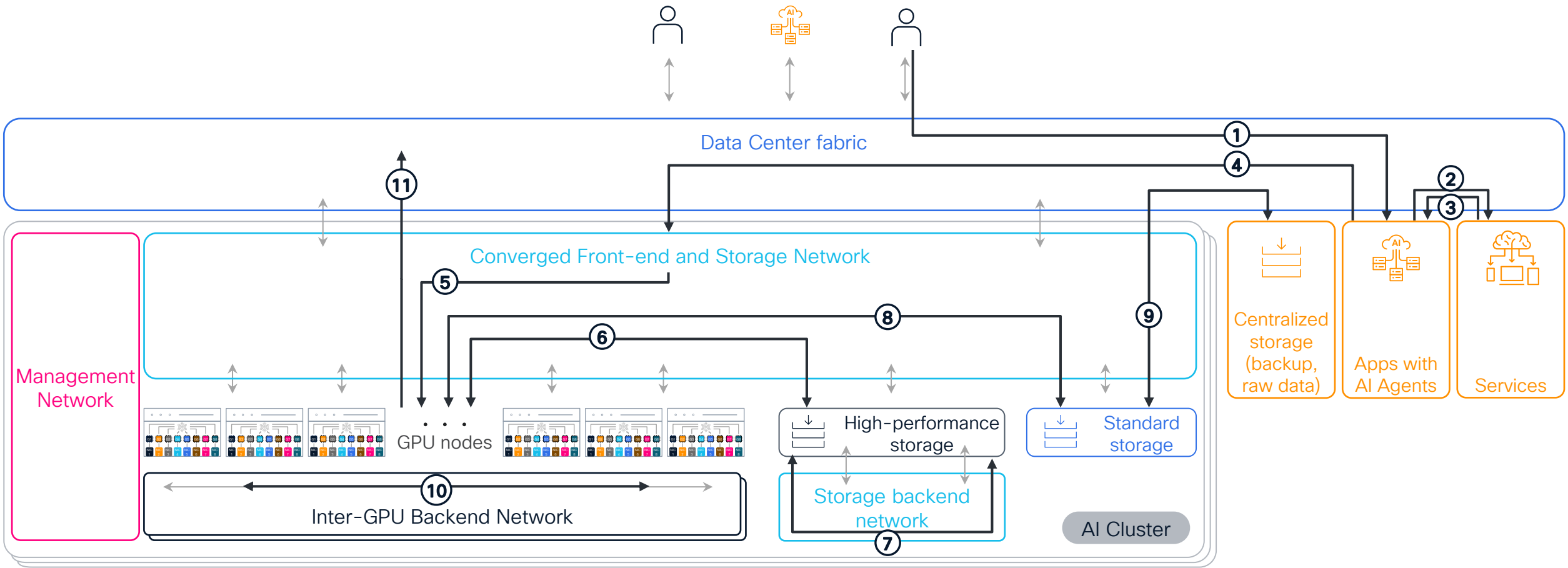
Multiple Networks of an AI Cluster



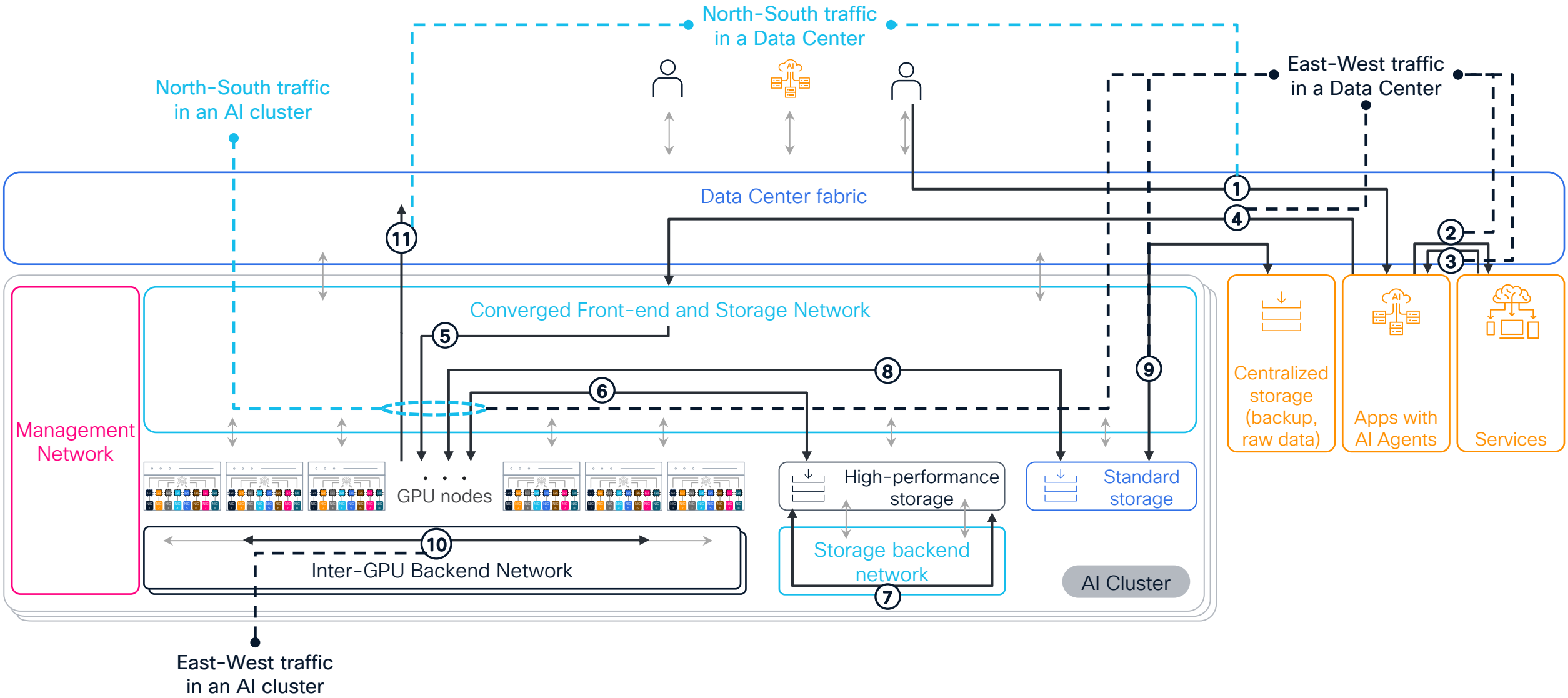
Multiple Networks of an AI Cluster



Traffic Flow in an AI Cluster



Traffic Flow in an AI Cluster



Data Center East-West traffic may be North-South traffic for AI Cluster

AI Networking Challenges

Reference Architectures and Validated Designs

Operational Simplicity

Observability

Converged Frontend
and Storage Network

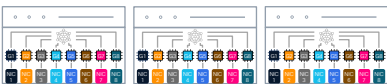
- Packet loss
- Network delay
- Network congestion
- Link failure

Storage traffic incast and
co-existence of many traffic types

Data Center fabric

Converged Front-end and Storage Network

Management
Network



GPU nodes



Inter-GPU Backend Network



High-performance
storage



Standard
storage

Storage backend
network

AI Cluster



Centralized
storage
(backup,
raw data)



Apps with
AI Agents



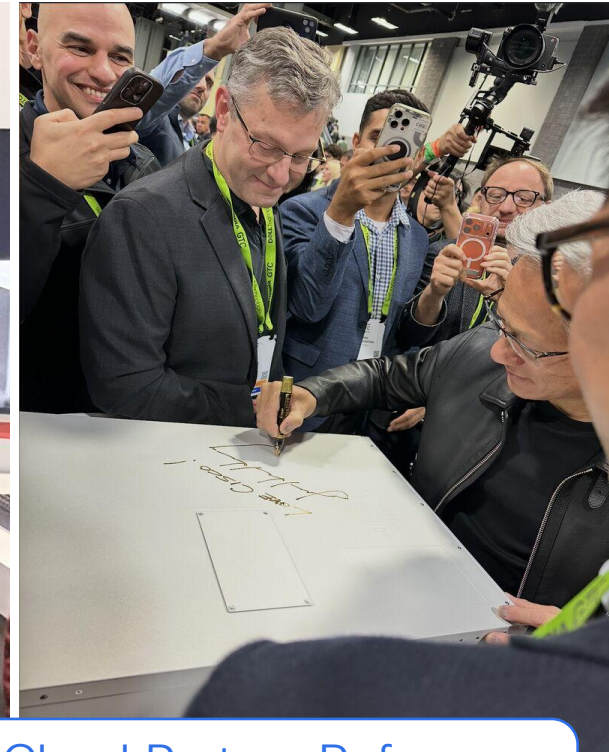
Services

Backend
Network

- Packet loss
- Network delay
- Network congestion
- Link failure

Large traffic burst from all
GPU NICs simultaneously

Cisco AI Reference Architectures with NVIDIA



Cisco Enterprise Reference Architecture (ERA)

Up to 1024 GPUs in a cluster

With Cisco Silicon One and Cloud Scale switches and NVIDIA Spectrum-X Ethernet technology

Available since March 2025 (NVIDIA GTC)

Cisco Cloud Reference Architecture (CRA)

1K+ GPUs in a cluster

With Cisco Silicon One switches and NVIDIA Spectrum-X Ethernet technology

Available since Oct 2025 (NVIDIA GTC DC)

NVIDIA Cloud Partner Reference Architecture (NCP RA)

1K+ GPUs in a cluster

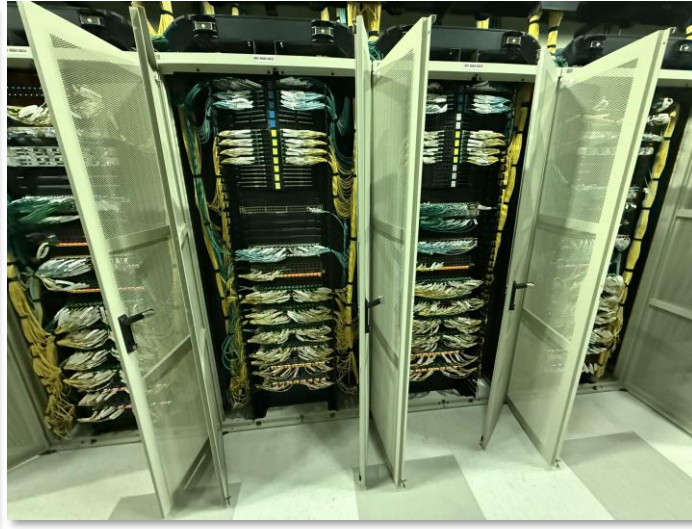
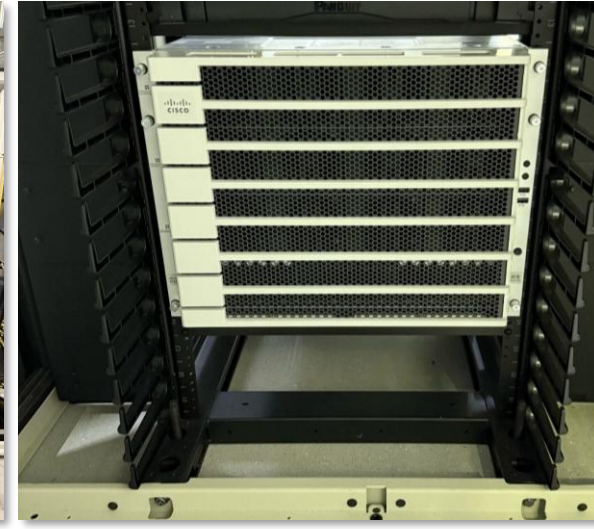
With Cisco N9100 switches (powered by NVIDIA Ethernet silicon) and NVIDIA Spectrum-X Ethernet technology

Available since Oct 2025 (NVIDIA GTC DC)

Operational simplicity of Cisco NX-OS, Nexus Dashboard, SONiC, and Nexus Hyperfabric AI

Hyperscalers, Neoclouds, Sovereign clouds, and Enterprises

Cisco AI Reference Architectures and Validations



Compute Accelerators

- Cisco Nexus 9364E-SG2 (Silicon One G200) switches validated with AMD MI300X and Pensando Pollara NICs
- Cisco Nexus 9364E-SG2 (Silicon One G200) switches validated with Intel Gaudi 3
- Cisco Nexus 9364D-GX2A (Cloud Scale) switches validated with Intel Gaudi 2

Storage

- VAST
- WEKA
- DDN
- Pure Storage
- NetApp
- Hitachi Vantara

Benchmarking

- IBPerf: Incast congestive tests, Bandwidth and Latency Tests
- PARAM/Chakra
- NCCL: Collectives (AllReduce, AlltoAll etc.), Split-mask for multiple workloads
- MLCommons Training: BERT, DLRM, GPT3, Llama2 70B, Llama3.1 405B, Stable Diffusion
- DeepSeek
- MLCommons Inference
- Customer models PoC hosting

Operational simplicity of Cisco NX-OS, Nexus Dashboard, SONiC, and Nexus Hyperfabric AI

Any GPU, Any Storage, Any NIC – Consistent Cisco networking experience

AI Cluster – Operational Models from Cisco

Nexus Dashboard

On-premises managed network



Compute

**Cisco UCS
NVIDIA HGX, MGX or
RTX PRO 6000 Servers**



Networking

**Cisco Nexus 9300
w/ Silicon One
w/ Cloud Scale
w/ NVIDIA Spectrum-X**



Storage

**NetApp | Pure Storage |
VAST Data | Hitachi
Vantara**

Training, fine-tuning, inference

Full-featured, massively deployed

Nexus Hyperfabric AI

Cisco-managed network cloud controller



Compute

**Cisco UCS
NVIDIA HGX
Future: RTX PRO Server**



Networking

**Cisco Nexus Hyperfabric
w/ Silicon One
w/ NVIDIA Spectrum-X**



Storage

VAST Data

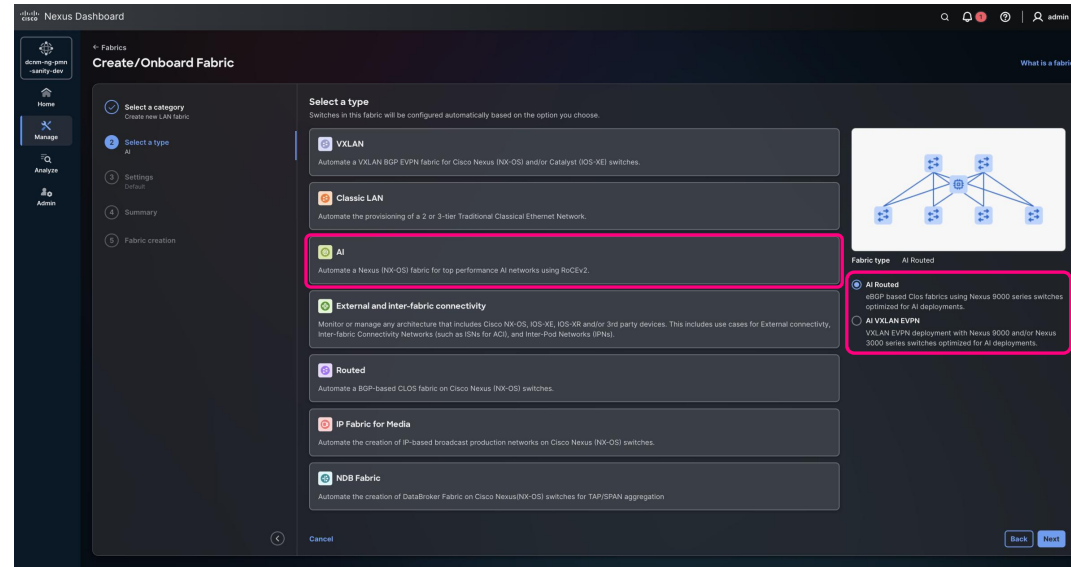
Training, fine-tuning, inference

Full stack: network, servers, GPU, storage

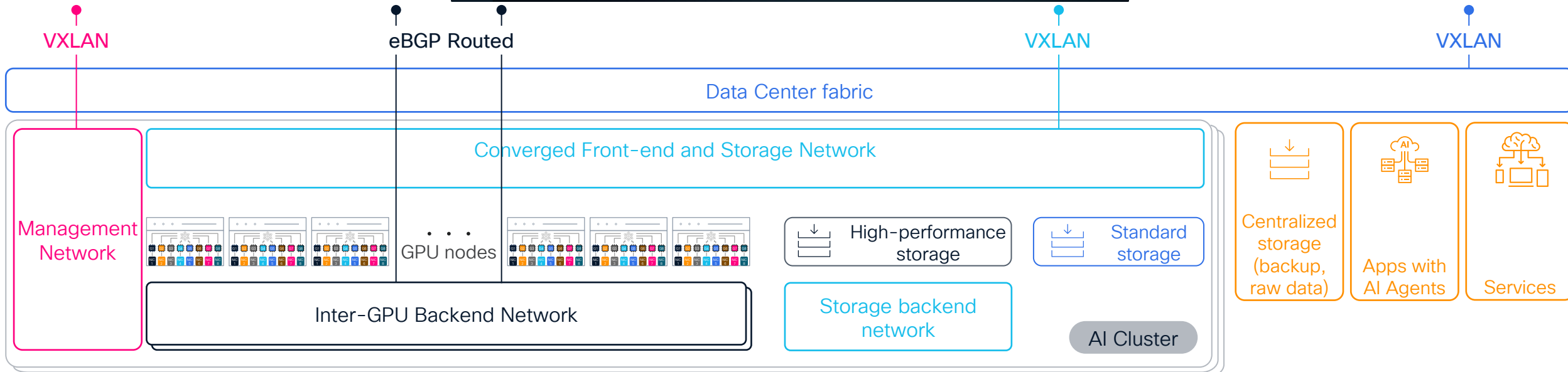
Nexus Dashboard – Common Platform for All Network Types

Visibility into AI job, GPUs, and NICs – correlated with network paths and anomalies

Common provisioning platform
Full-featured automation for all network types



Consistent APIs
Reuse existing automation frameworks and expertise



Nexus Dashboard – Built-in AI Fabric Types

 Nexus Dashboard

dcnm-ng-pmn
-sanity-dev

Home

Manage

Analyze

Admin

← Fabrics

Create/Onboard Fabric

What is a fabric?

✓ Select a category
Create new LAN fabric

2 Select a type
AI

3 Settings
Default

4 Summary

5 Fabric creation

Select a type

Switches in this fabric will be configured automatically based on the option you choose.

 **VXLAN**
Automate a VXLAN BGP EVPN fabric for Cisco Nexus (NX-OS) and/or Catalyst (IOS-XE) switches.

 **Classic LAN**
Automate the provisioning of a 2 or 3-tier Traditional Classical Ethernet Network.

 **AI**
Automate a Nexus (NX-OS) fabric for top performance AI networks using RoCEv2.

 **External and inter-fabric connectivity**
Monitor or manage any architecture that includes Cisco NX-OS, IOS-XE, IOS-XR and/or 3rd party devices. This includes use cases for External connectivity, Inter-fabric Connectivity Networks (such as ISNs for ACI), and Inter-Pod Networks (IPNs).

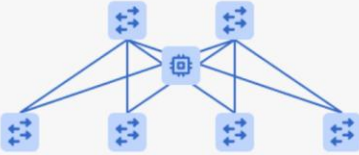
 **Routed**
Automate a BGP-based CLOS fabric on Cisco Nexus (NX-OS) switches.

 **IP Fabric for Media**
Automate the creation of IP-based broadcast production networks on Cisco Nexus (NX-OS) switches.

 **NDB Fabric**
Automate the creation of DataBroker Fabric on Cisco Nexus(NX-OS) switches for TAP/SPAN aggregation

Cancel

BackNext



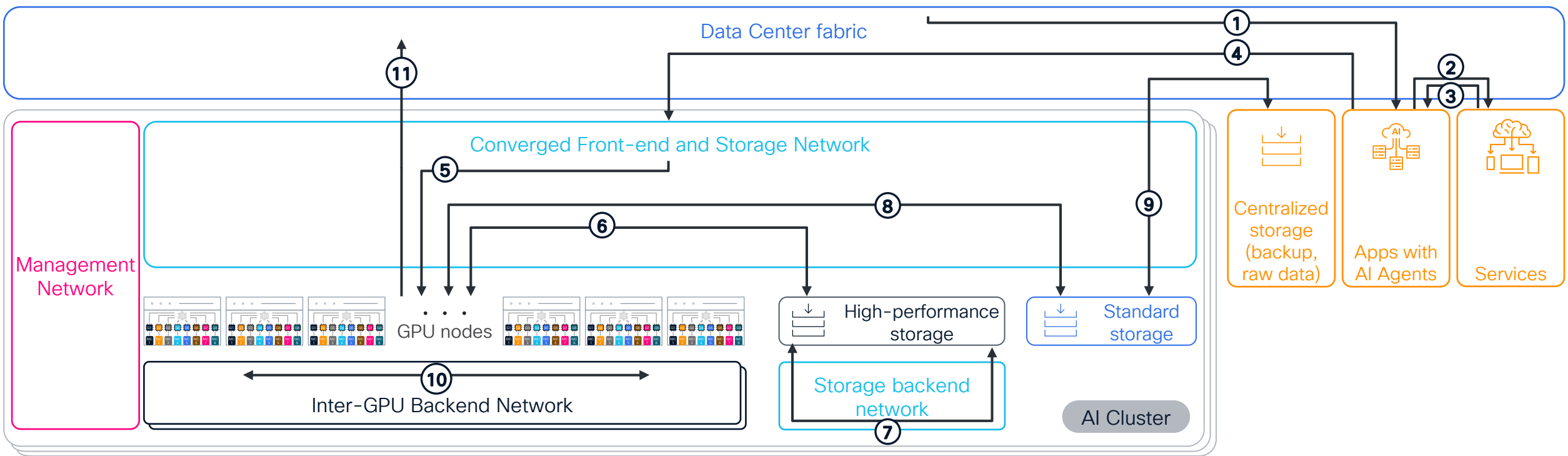
Fabric type

AI Routed

☒ **AI Routed**
eBGP based Clos fabrics using Nexus 9000 series switches optimized for AI deployments.

☐ **AI VXLAN EVPN**
VXLAN EVPN deployment with Nexus 9000 and/or Nexus 3000 series switches optimized for AI deployments.

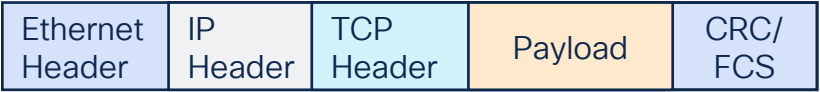
Traffic in Front-end, Storage, and Backend Networks



Transport Protocols in an AI Cluster

TCP

Transmission Control Protocol



Mostly used in front-end networks

Built-in congestion control
(mostly standardized –
based on network notifications (ECN) and RTT on end-devices)

Built-in lost-packet retransmission
(mostly runs on lossy network)

RoCEv2

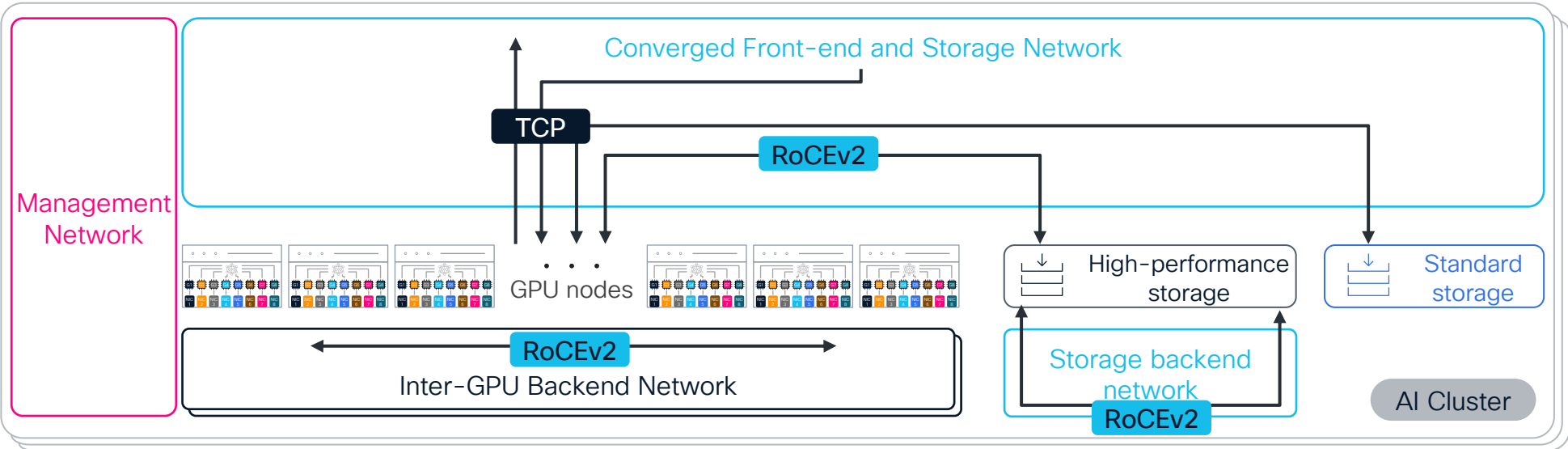
Remote Direct Memory Access (via UDP/IP) over Converged Ethernet



Mostly used in Storage and backend networks

RoCEv2 congestion management
(concept standardized, implementation dependent –
based on network notifications (ECN) and RTT on end-devices)

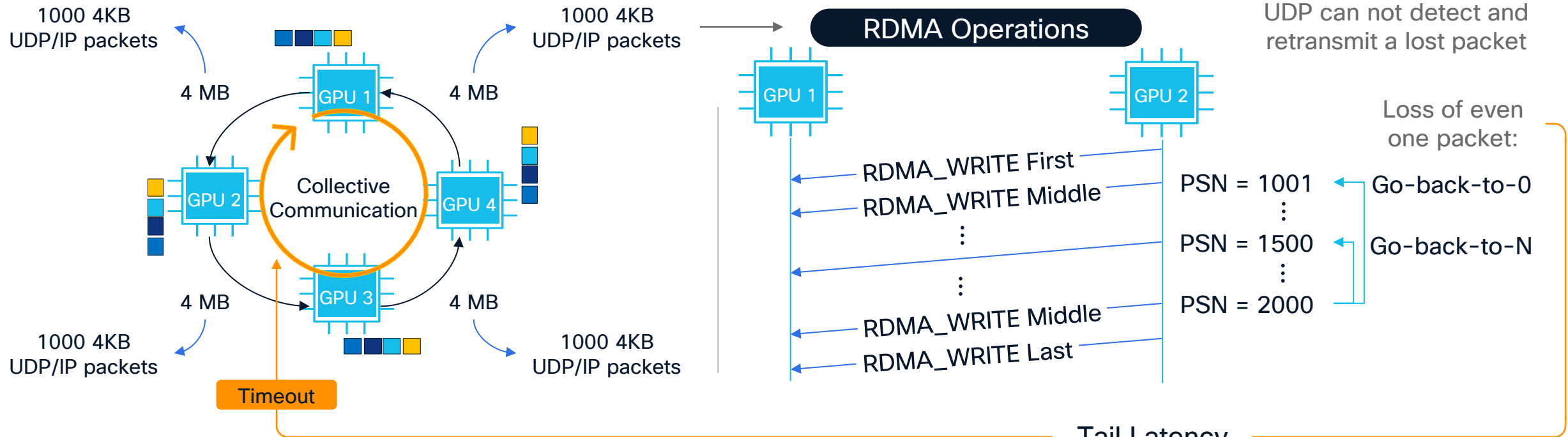
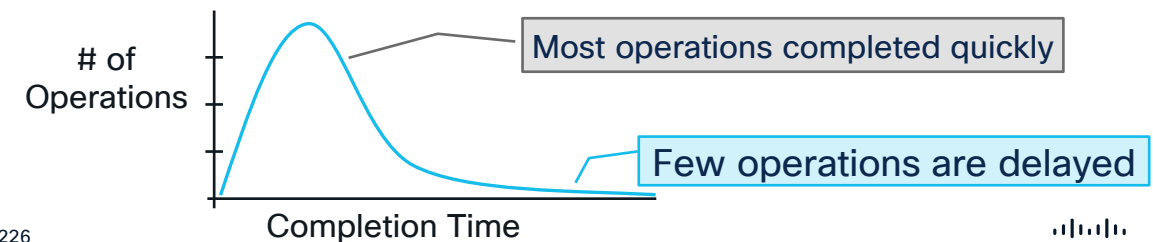
Packet loss leads to significant performance degradation
(mostly runs on lossless network)



Packet loss leads to significant performance degradation Why?
(mostly runs on lossless network)

caused by

Bit errors or Network congestion



RoCEv2

Packet loss leads to significant performance degradation
(mostly runs on lossless network)

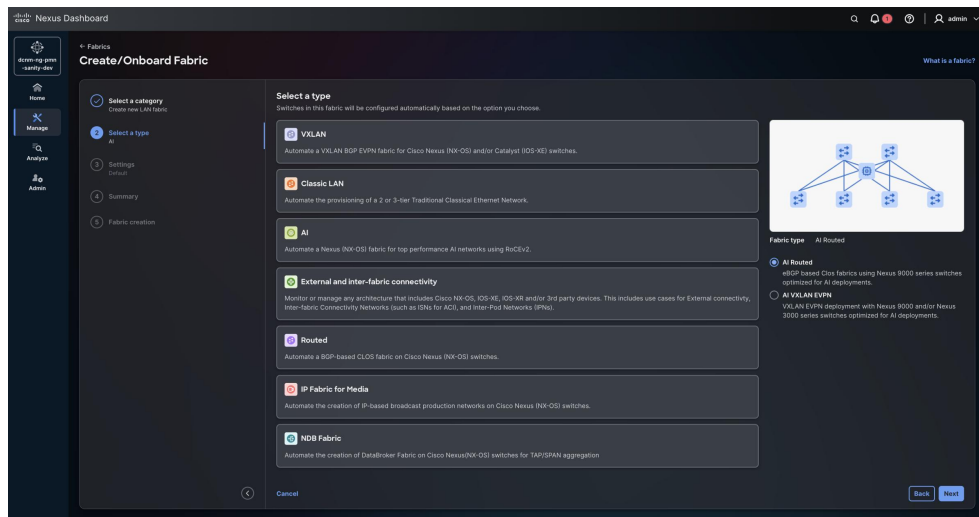
caused by

Bit errors or Network congestion

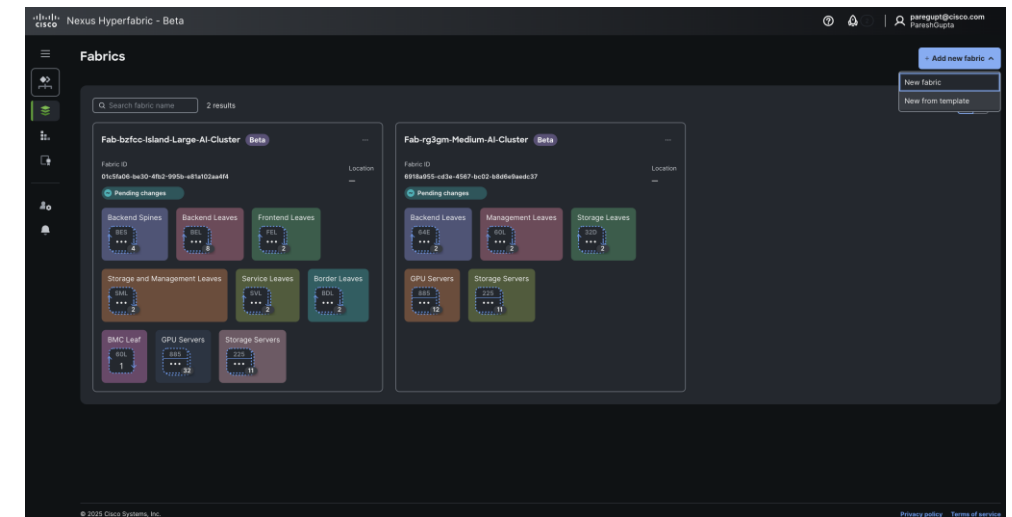
Priority-based Flow Control (PFC)

Fabric-wide enablement on all switches with fine-tuned ECN thresholds using

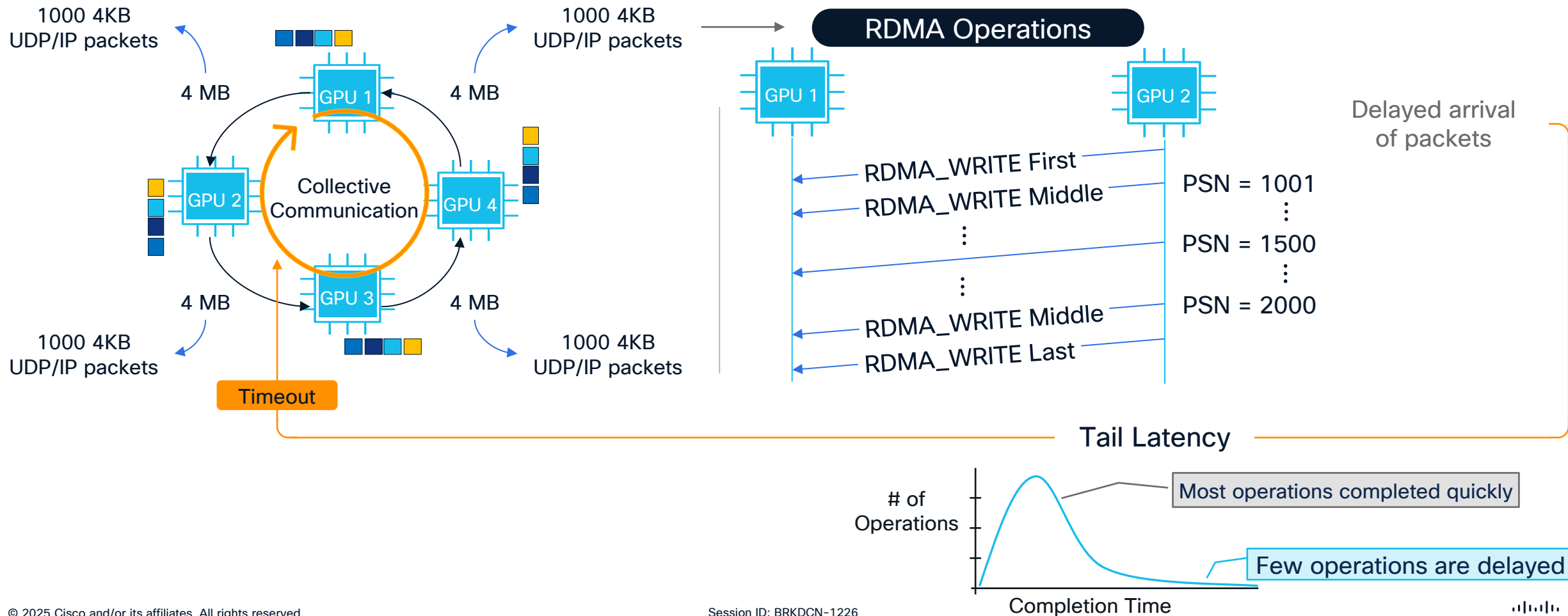
Nexus Dashboard
On-premises managed network



Nexus Hyperfabric AI
Cisco-managed network cloud controller



Network Delay leads to significant performance degradation Why?
(mostly runs on lossless network)



RoCEv2

Network Delay leads to significant performance degradation
(mostly runs on lossless network)

caused by

Forwarding delay

Propagation delay

Serialization delay

Queuing delay

addressed by

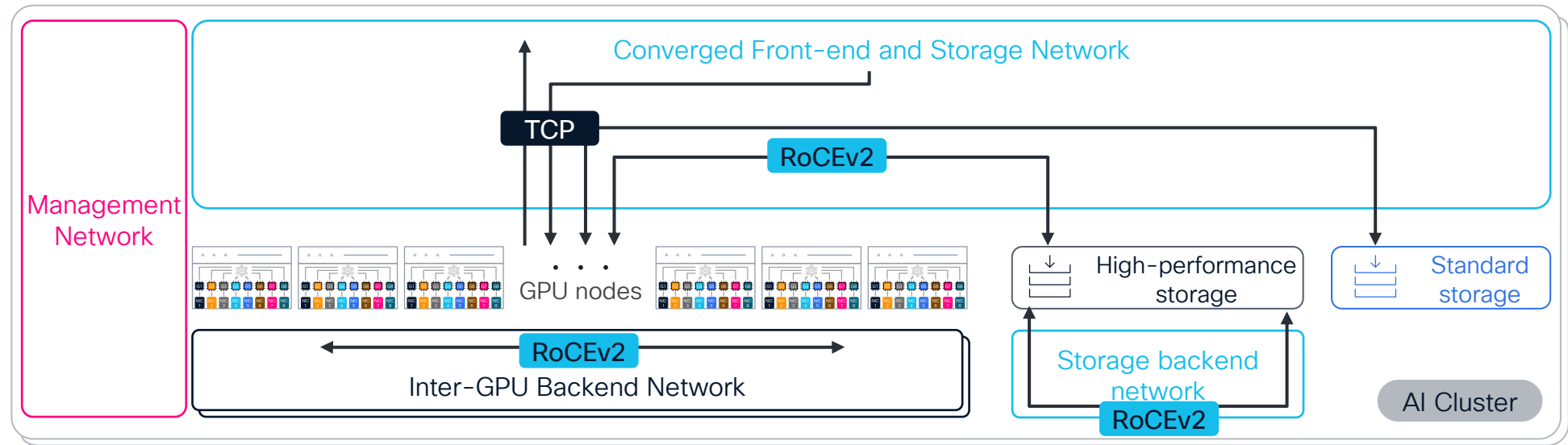
Low-latency switches

Minimum required and consistent cable lengths

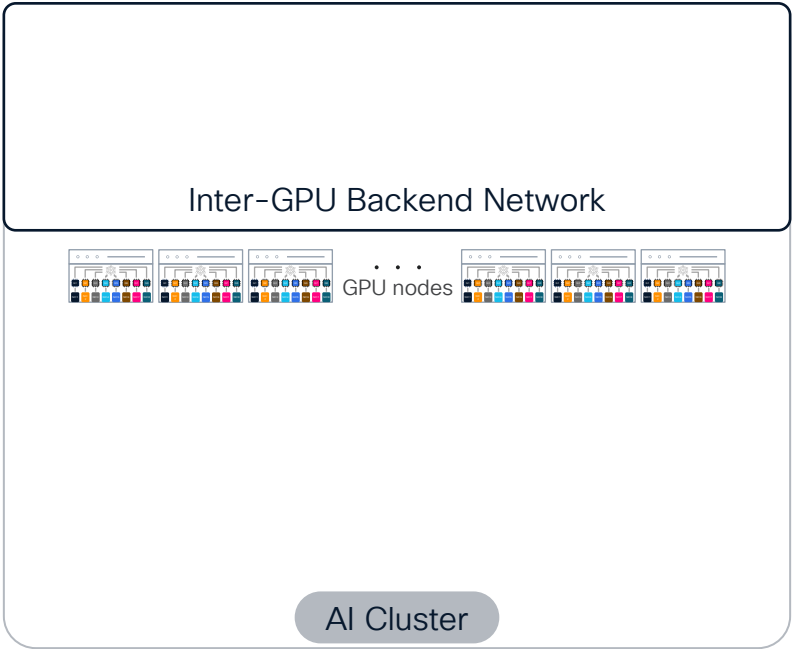
Faster speeds (400 Gbps / 800 Gbps)

Network congestion prevention

Efficient job placement and network design
(rail-optimized, non-blocking, faster-consistent speeds)



Non-Blocking and Rail-Optimized Backend Network



Non-Blocking and Rail-Optimized Backend Network

Non-Blocking

32 ports connected to spine switches



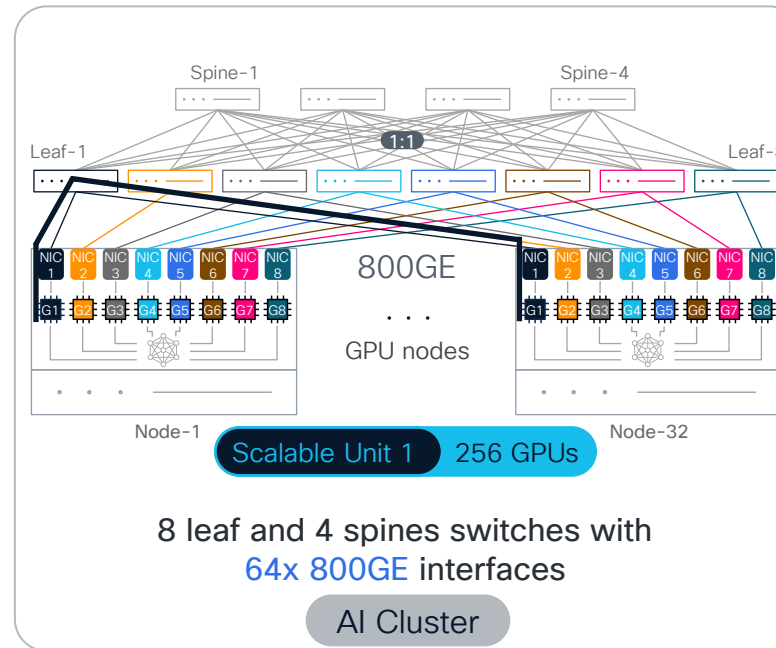
32 ports connected to GPU nodes

Provides enough network capacity for all GPUs to send and receive traffic at full capacity simultaneously

Rail-Optimized

Port 1 on all nodes connects to Leaf-1, Port 2 on all nodes connects to Leaf-2, and so on.

Allows single-hop forwarding among GPUs of a scalable unit and reduces or eliminates traffic to spines



Cisco Nexus 9364E-SG2

64x 800G or 128x 400G or 256x 200G or 512x 100G

Cisco Silicon One



Cisco N9164E-NS4

64x 800G or 128x 400G or 256x 200G

NVIDIA Spectrum-X Ethernet Silicon

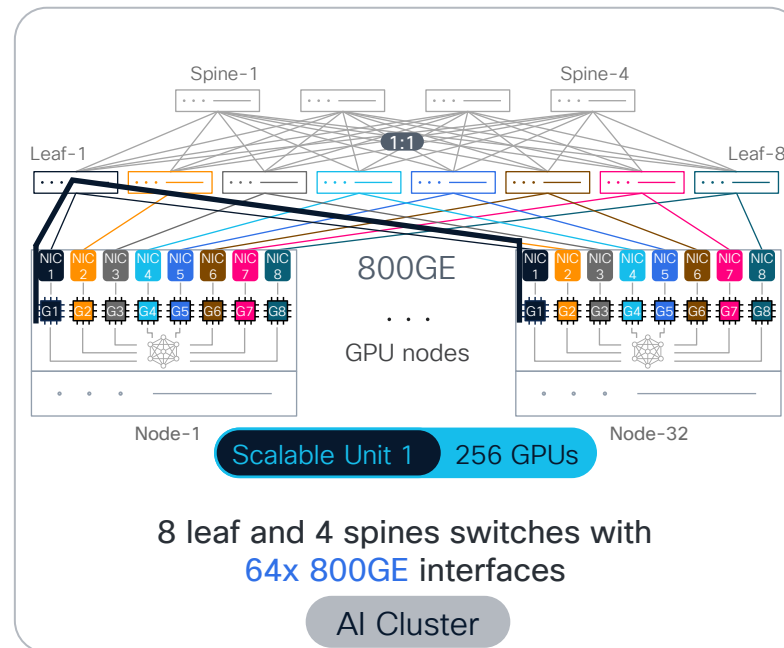
Rail-Optimized

Why?

Port 1 on all nodes connects to Leaf-1,
Port 2 on all nodes connects to Leaf-2,
and so on.

Allows single-hop forwarding among GPUs of
a scalable unit and reduces or eliminates
traffic to spines

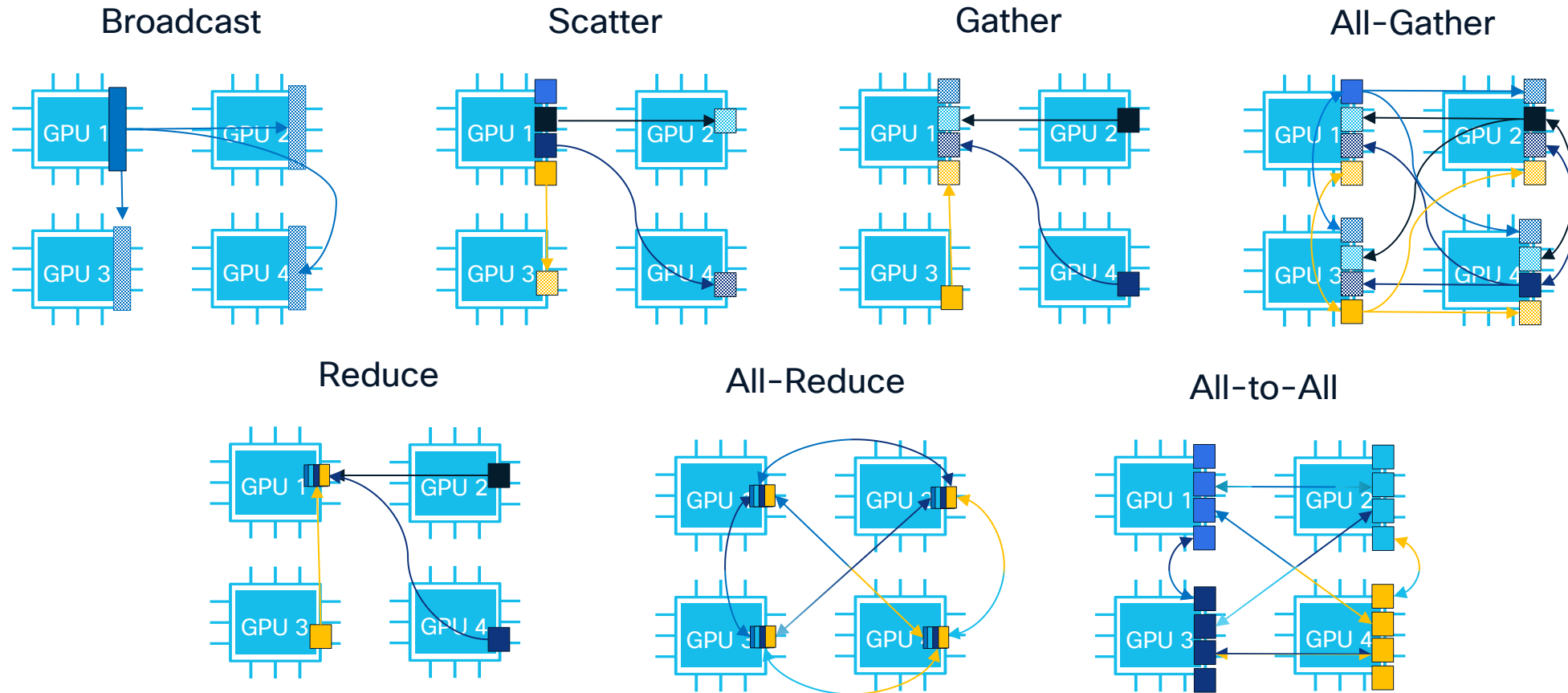
How?



Collective Communication

Parallel
distributed
processing

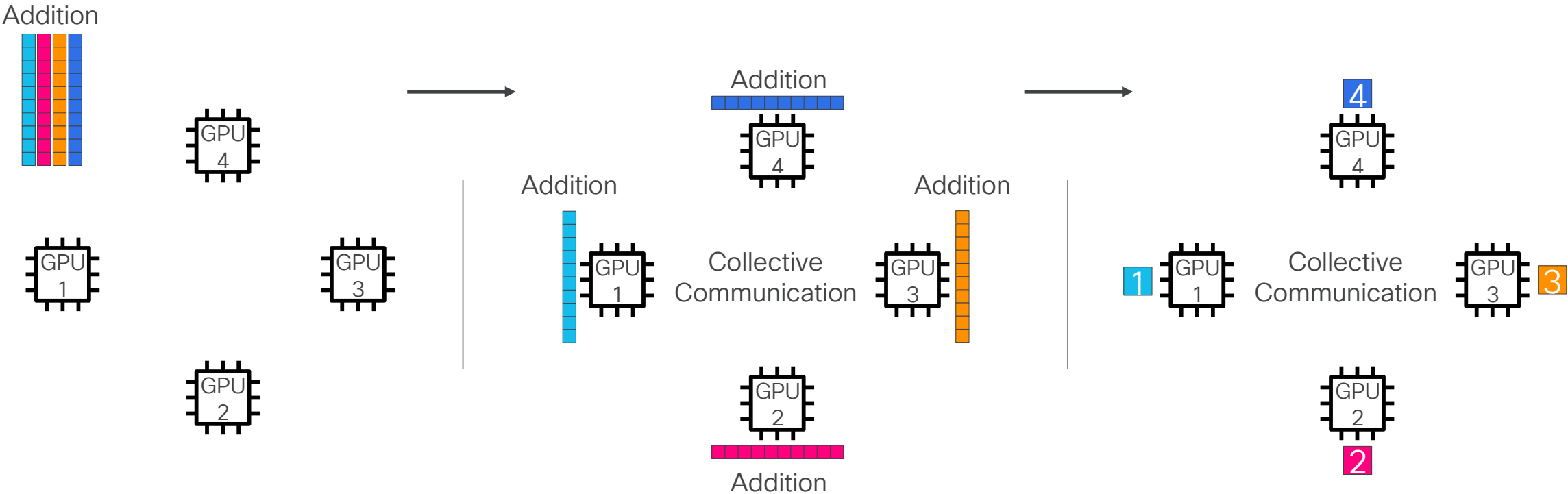
- NVIDIA Collective Communication Library (NCCL): <https://github.com/NVIDIA/nccl>
- AMD ROCm Communication Collectives Library: <https://github.com/ROCm/rccl>
- Intel Habana Collective Communication Library (HCCL): https://github.com/HabanaAI/hccl_demo



Collective Communication

Many types – AllReduce, ReduceScatter, AllGather, AlltoAll, etc.

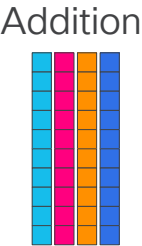
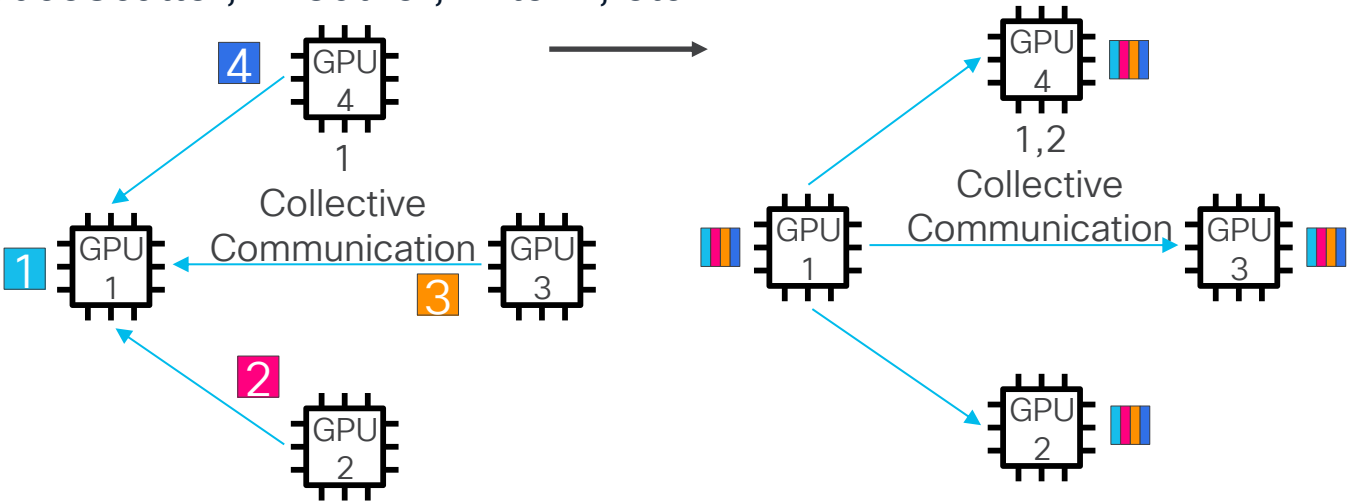
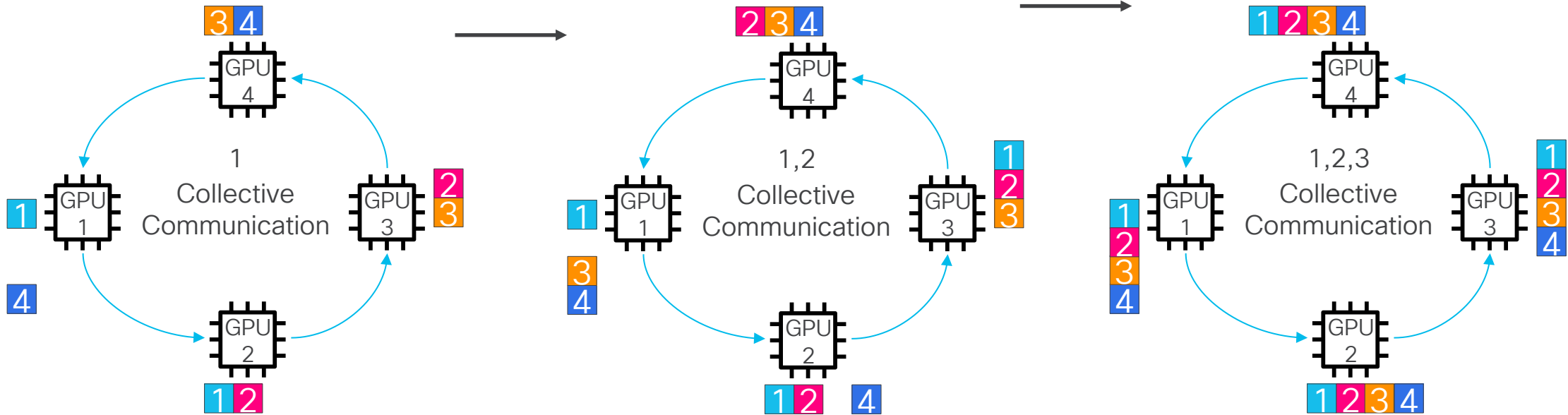
Parallel
distributed
processing



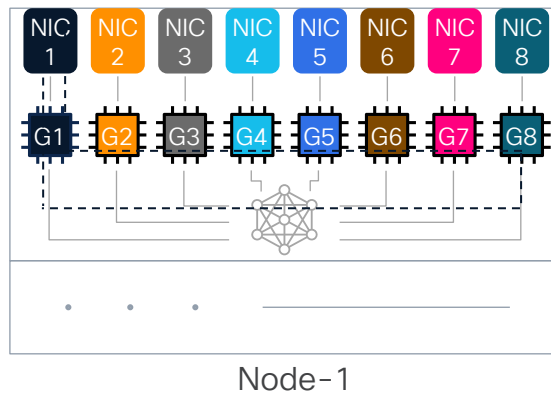
Collective Communication

Many types – AllReduce, ReduceScatter, AllGather, AlltoAll, etc.

Sync the
states

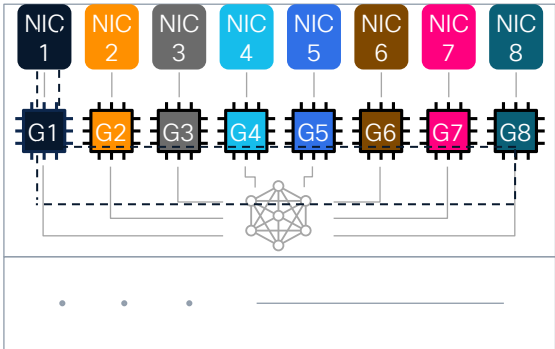


Collective Communication between GPU nodes

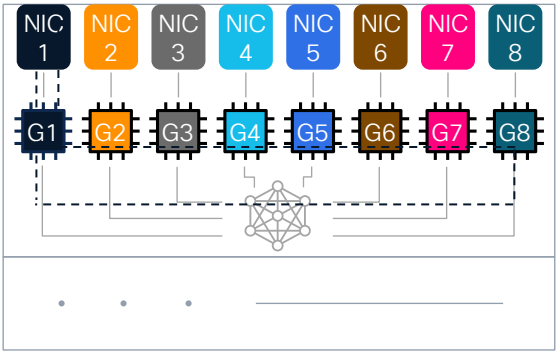


Only a conceptual representation for easier understanding

Collective Communication between GPU nodes



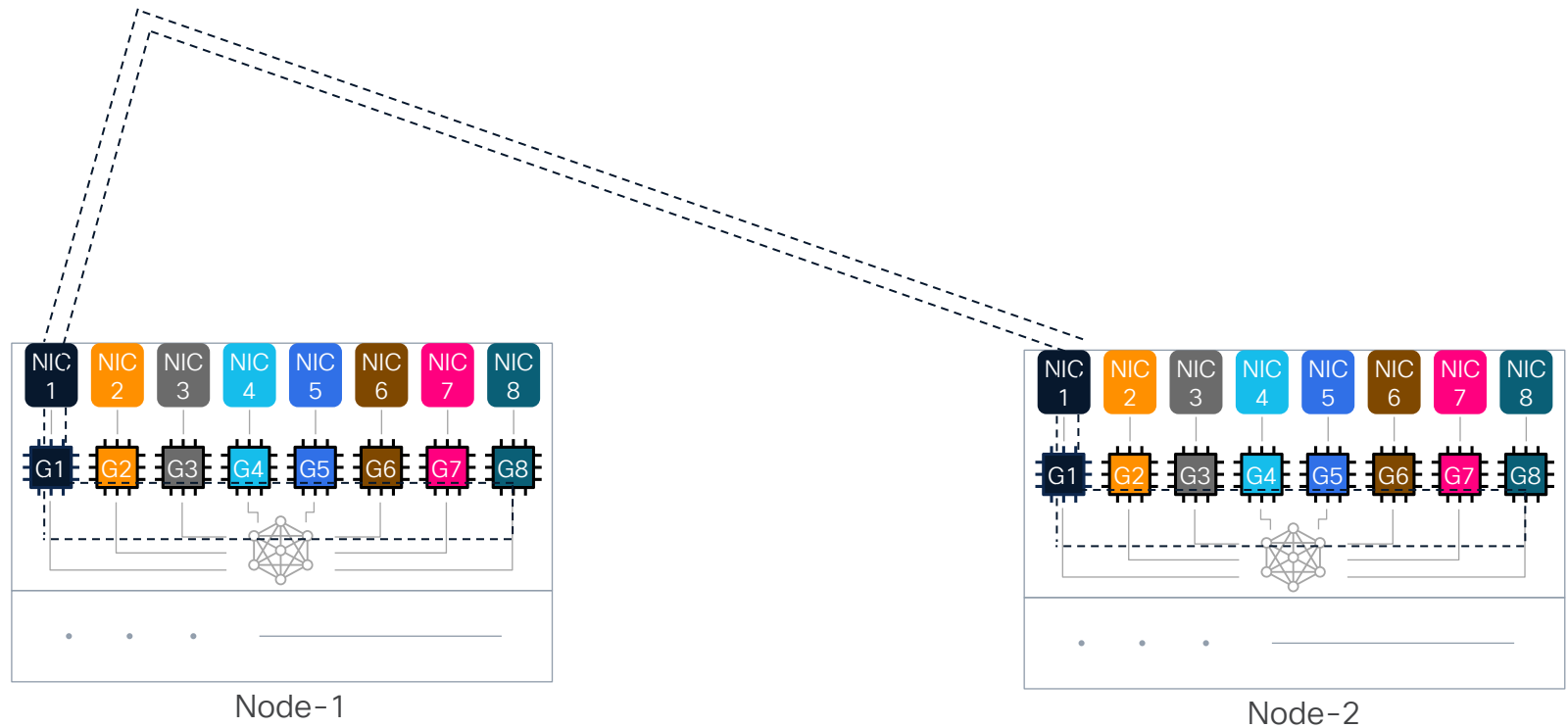
Node-1



Node-2

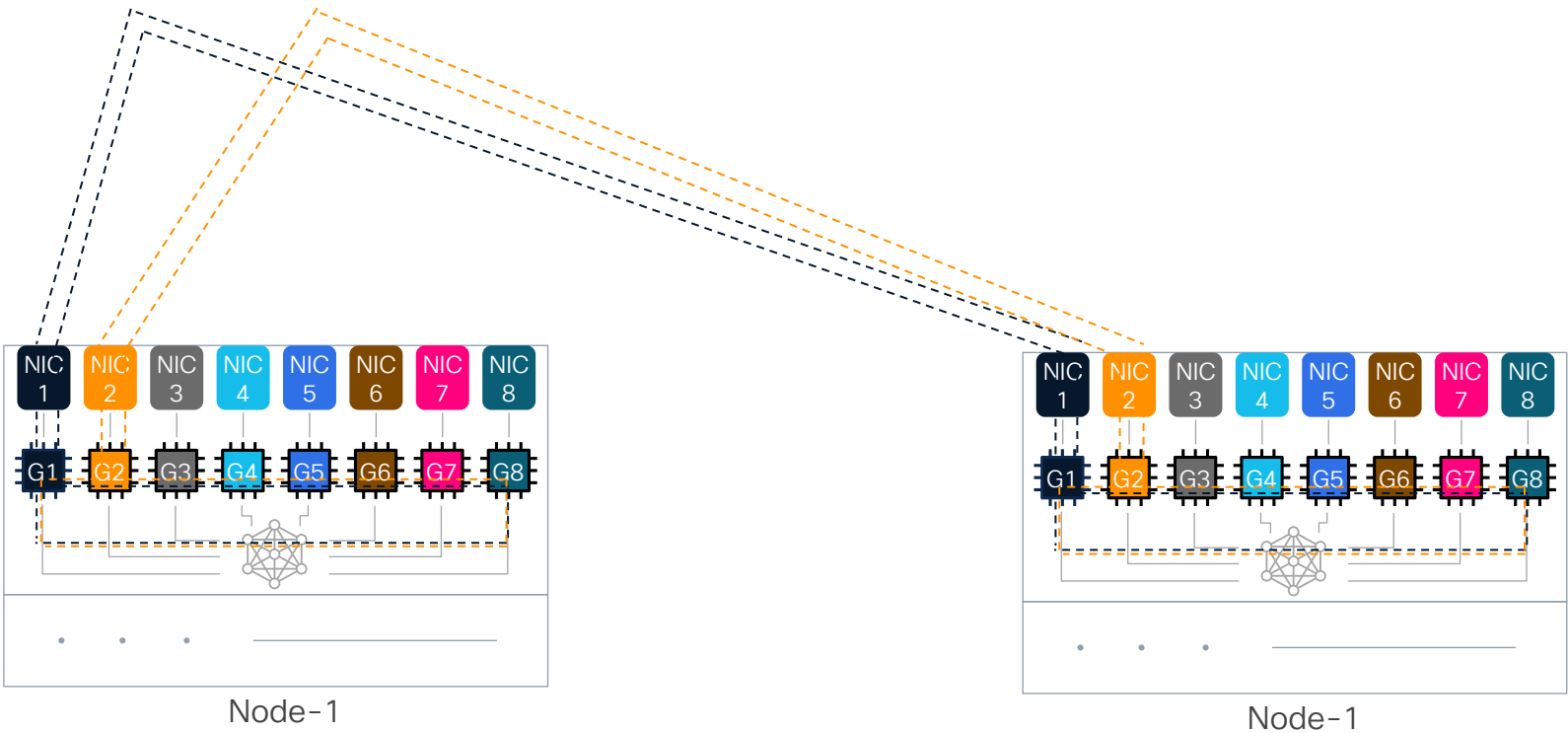
Only a conceptual representation for easier understanding

Collective Communication between GPU nodes



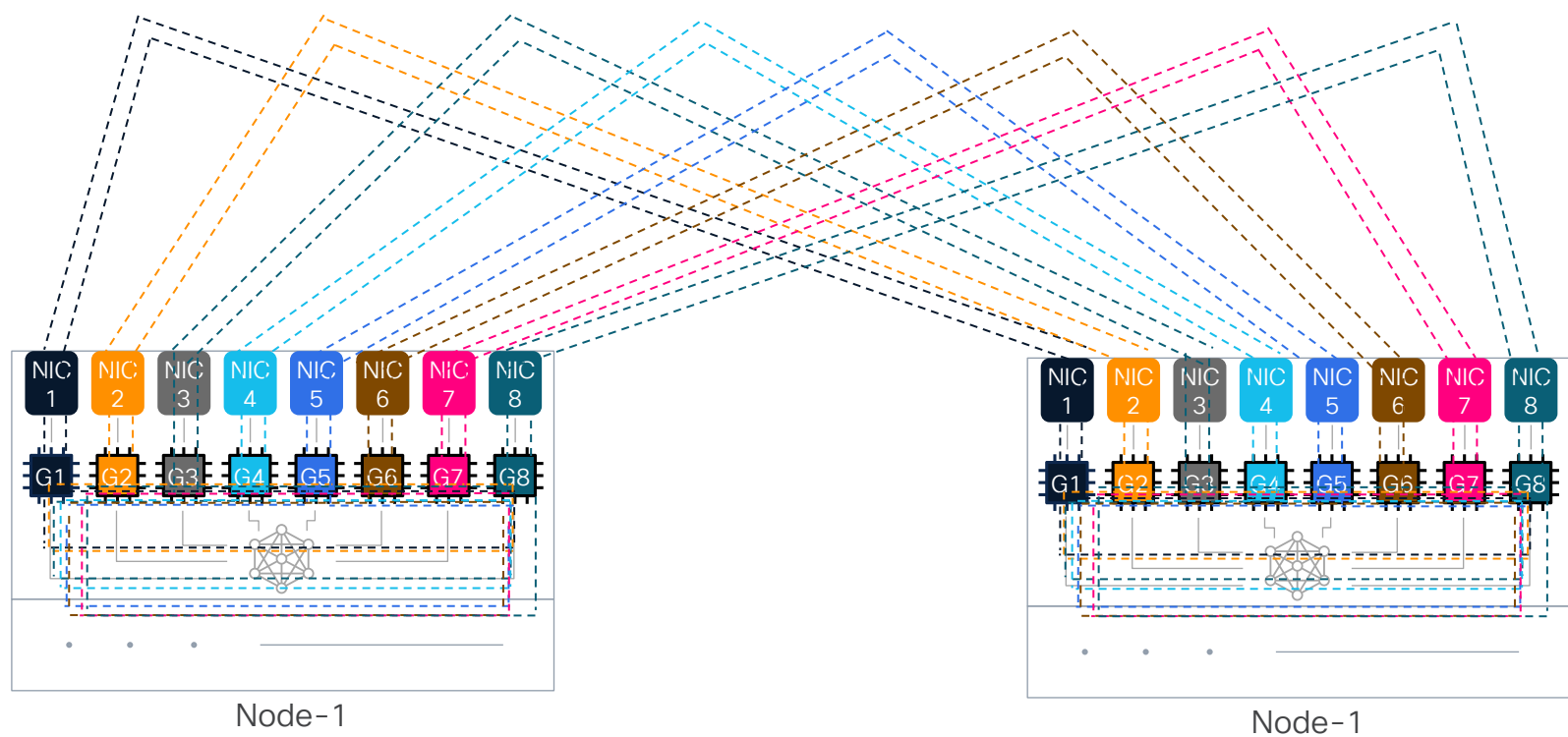
Only a conceptual representation for easier understanding

Collective Communication between GPU nodes



Only a conceptual representation for easier understanding

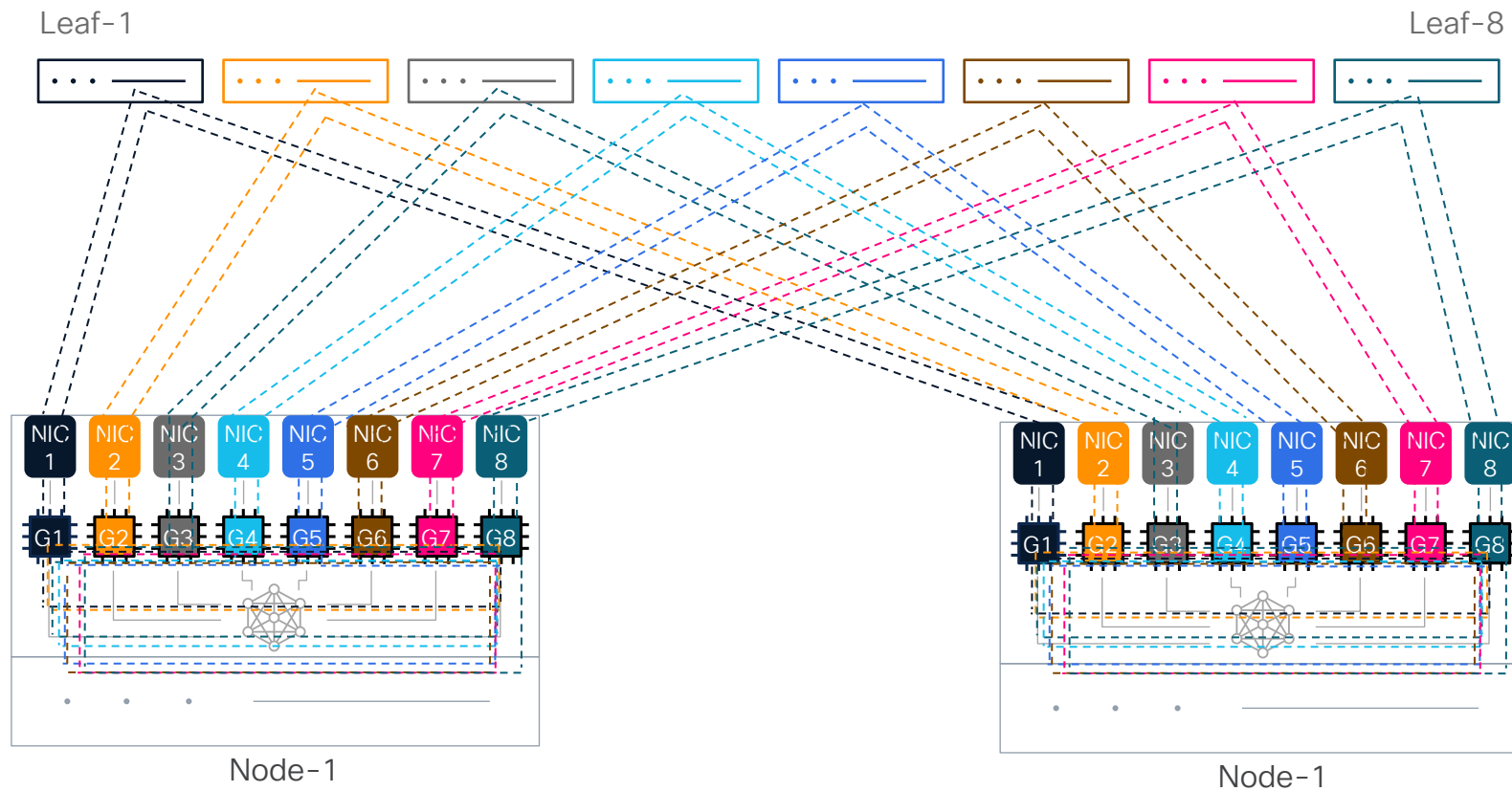
Collective Communication between GPU nodes



Only a conceptual representation for easier understanding

Rail-Optimized

Port 1 on all nodes connects to Leaf-1, Port 2 on all nodes connects to Leaf-2, and so on.



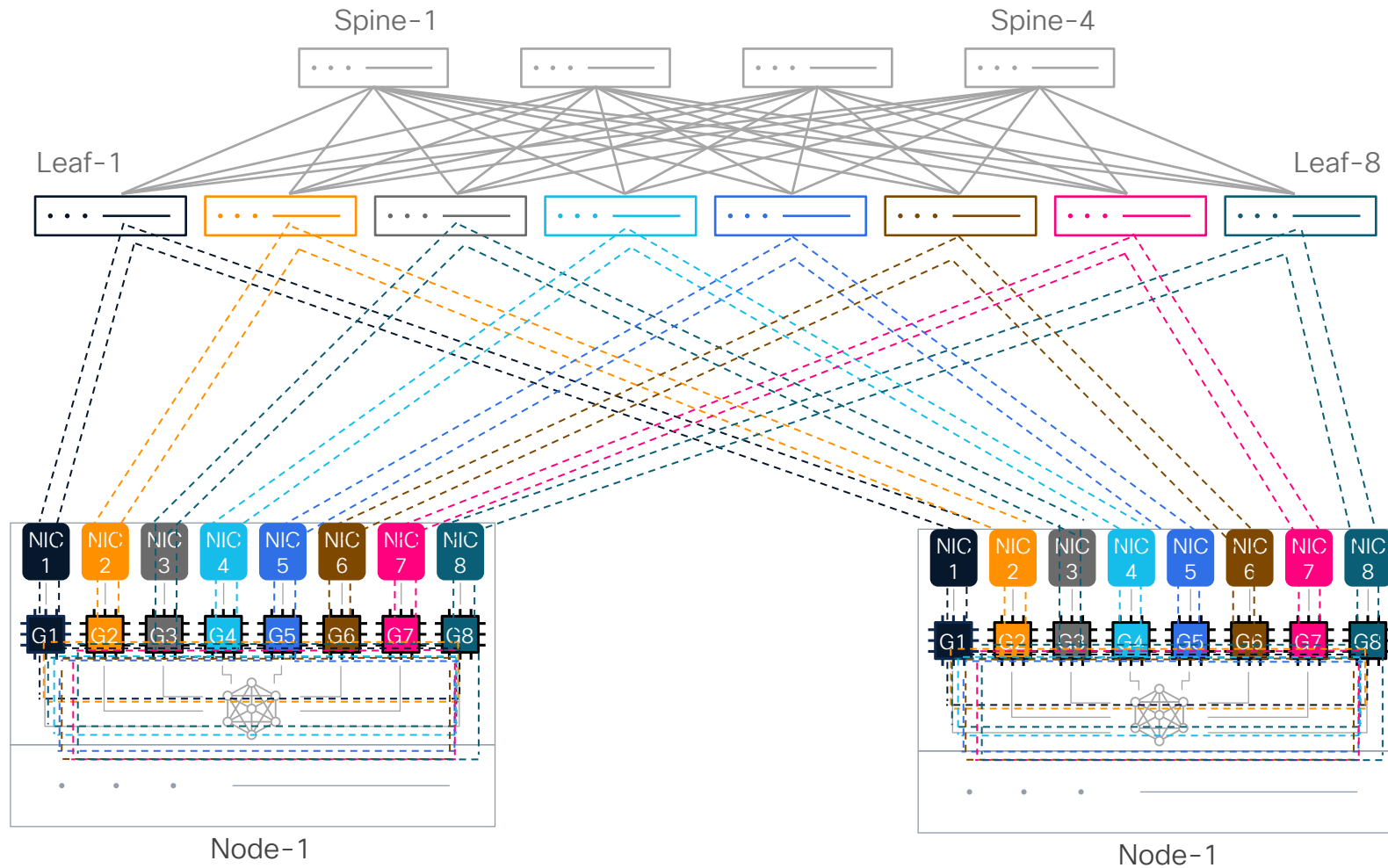
Port 1 on all nodes connects to Leaf-1, Port 2 on all nodes connects to Leaf-2, and so on.

Only a conceptual representation for easier understanding

Rail-Optimized

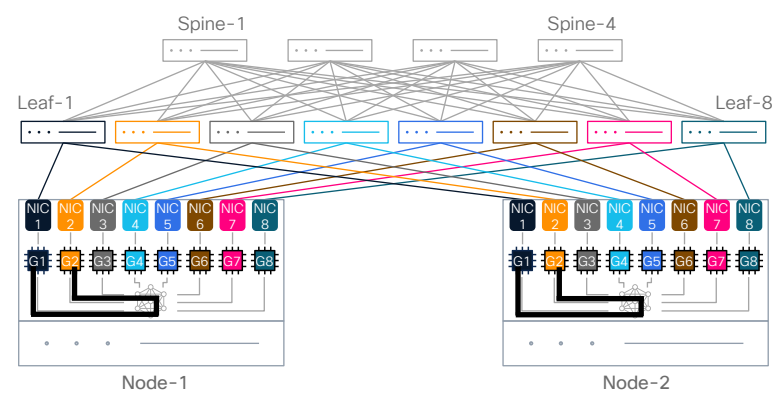
Port 1 on all nodes connects to Leaf-1, Port 2 on all nodes connects to Leaf-2, and so on.

Allows single-hop forwarding among GPUs of a scalable unit and reduces or eliminates traffic to spines



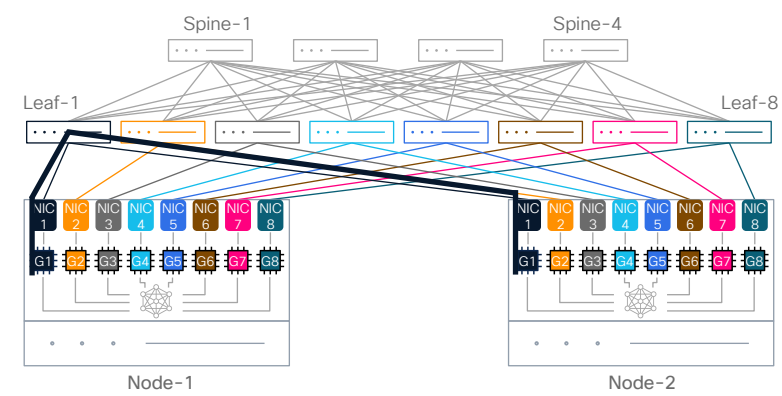
Only a conceptual representation for easier understanding

Rail-Optimized Design and GPU Communication



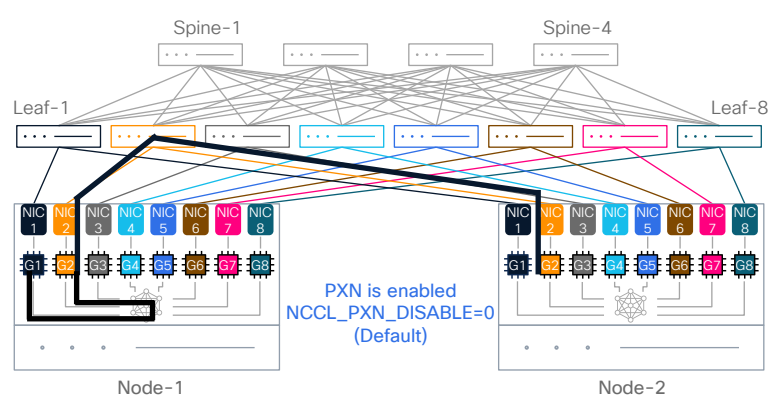
GPUs within a node

communicate via NVSwitch (1-Node-1 → NVSwitch → 2-Node-1)



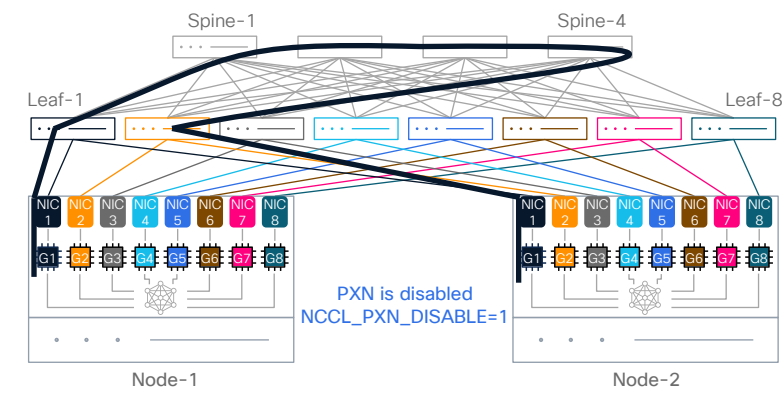
GPUs with the same IDs in different nodes

communicate via leaf switch (1-Node-1 → Leaf-1 → 1-Node-2).
Traffic does not go to spine



GPUs with different IDs in different nodes

first communicate via NVSwitch to the same GPU ID as the second node and
then via leaf switch to the second node
(1-Node-1 → NVSwitch → 2-Node-1 → Leaf-2 → 2-Node-2).
Traffic does not go to spine.



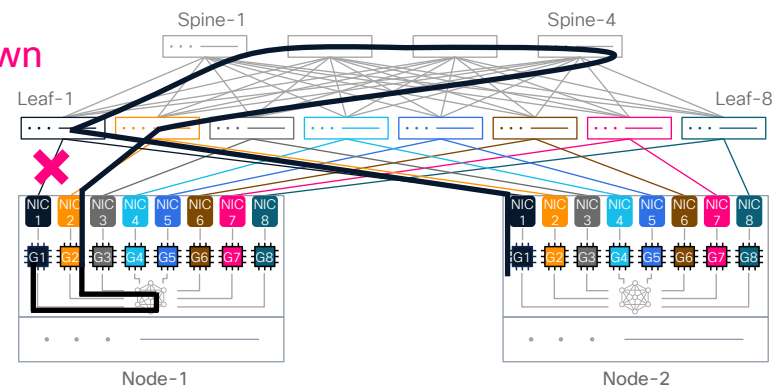
GPUs with different IDs in different nodes

communicate via leaf-spine-leaf switches
(1-Node-1 → Leaf-1 → Spines → Leaf-2 → 2-Node-2)

* Not all collectives work this way. Behavior changes with different NCCL versions. Some collectives may not send traffic to spines even with NCCL_PXN_DISABLE=1

Traffic Goes to Spines when...

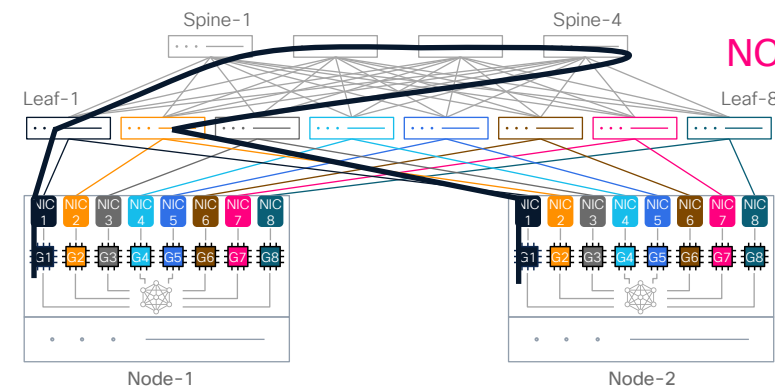
NIC link is down



GPUs with the same IDs in different nodes

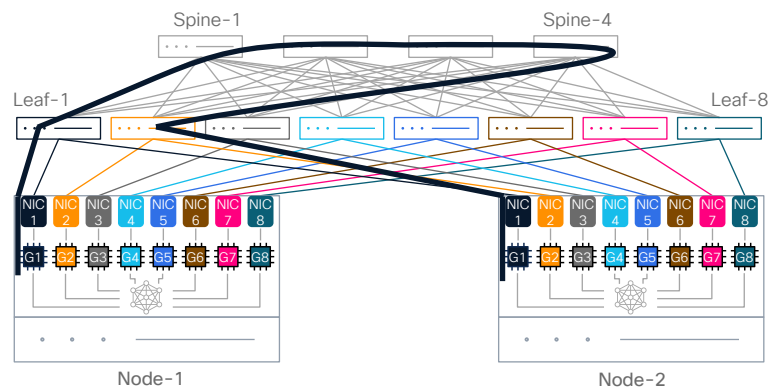
first communicate via NVSwitch to other GPUs within the node with failed link. The traffic then goes to network and crosses the rail boundary via spine switches to reach the second node (1-Node-1 → NVSwitch → 2-Node-1 → Leaf-2 → Spines → Leaf-1 → 1-Node-2).

NCCL is not used

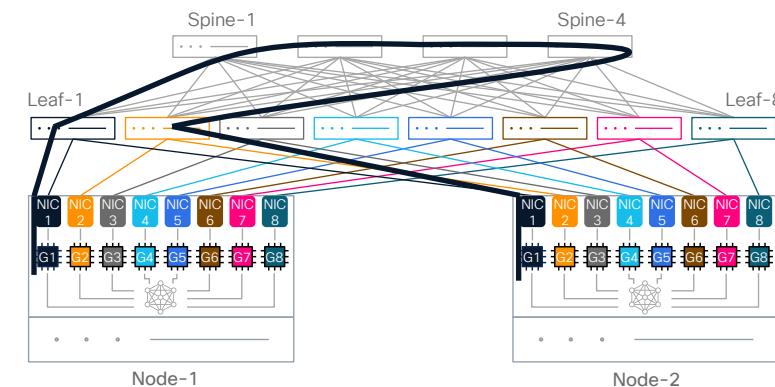


GPUs with different IDs in different nodes

communicate via leaf-spine-leaf switches (1-Node-1 → Leaf-1 → Spines → Leaf-2 → 1-Node-2).



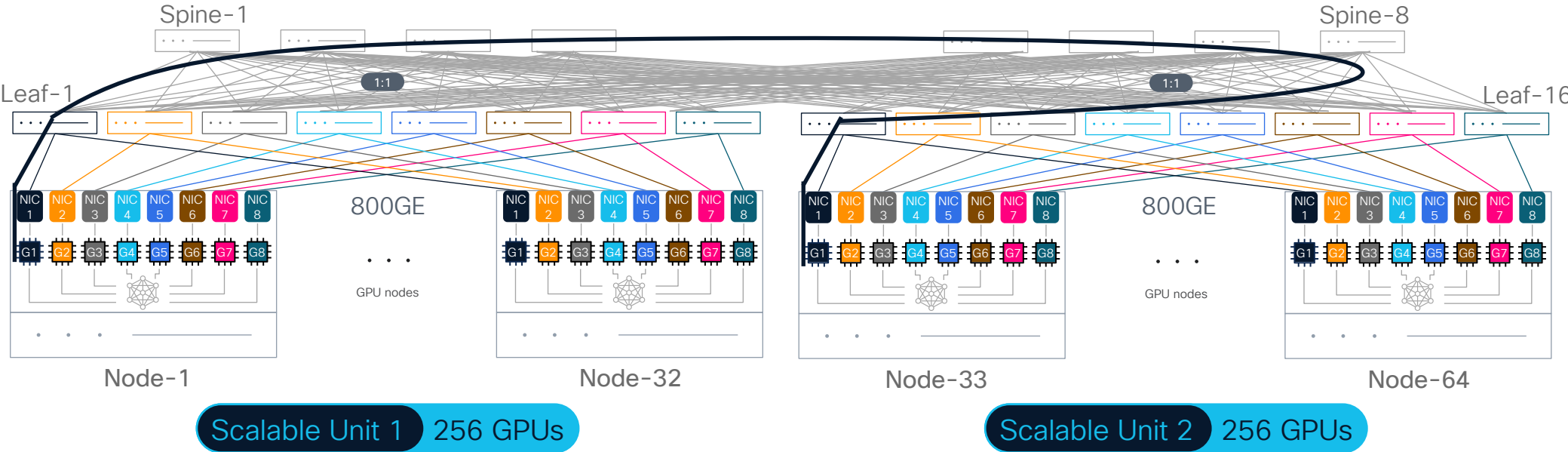
Affordable GPU nodes without NVSwitches



Older NCCL version is not optimized for PXN

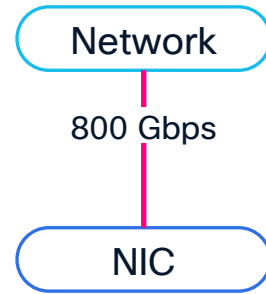
Traffic Goes to Spines when...

A job is large enough to run on multiple scalable units (256+ GPUs in this example)

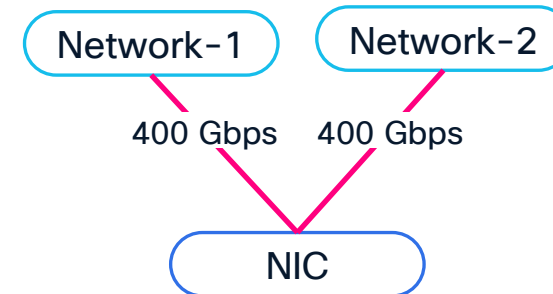
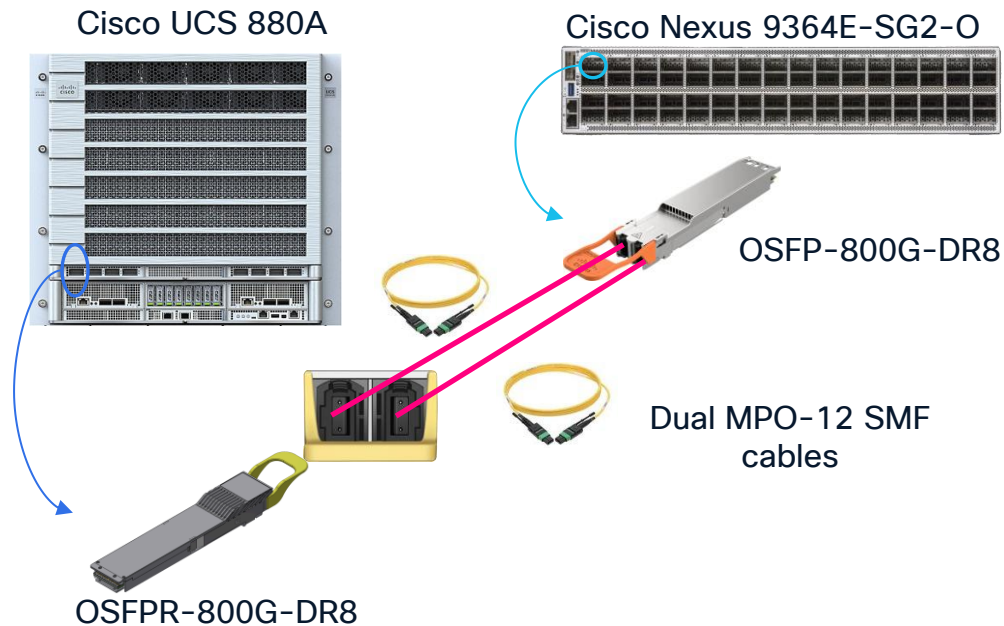
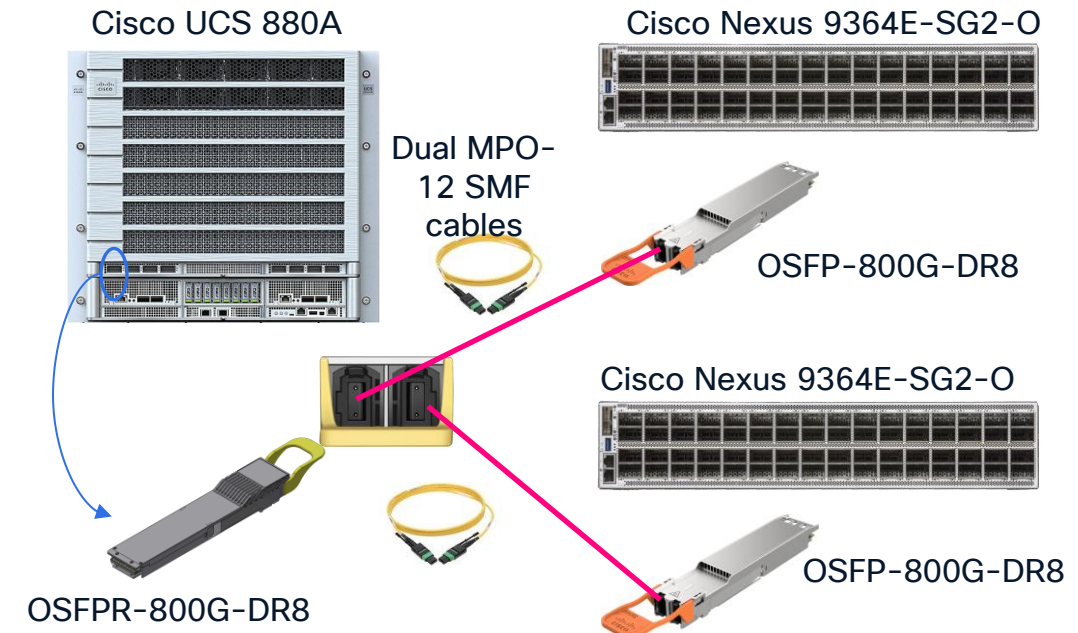


Evolution of Inter-GPU Backend Network Designs

Single-Plane Design

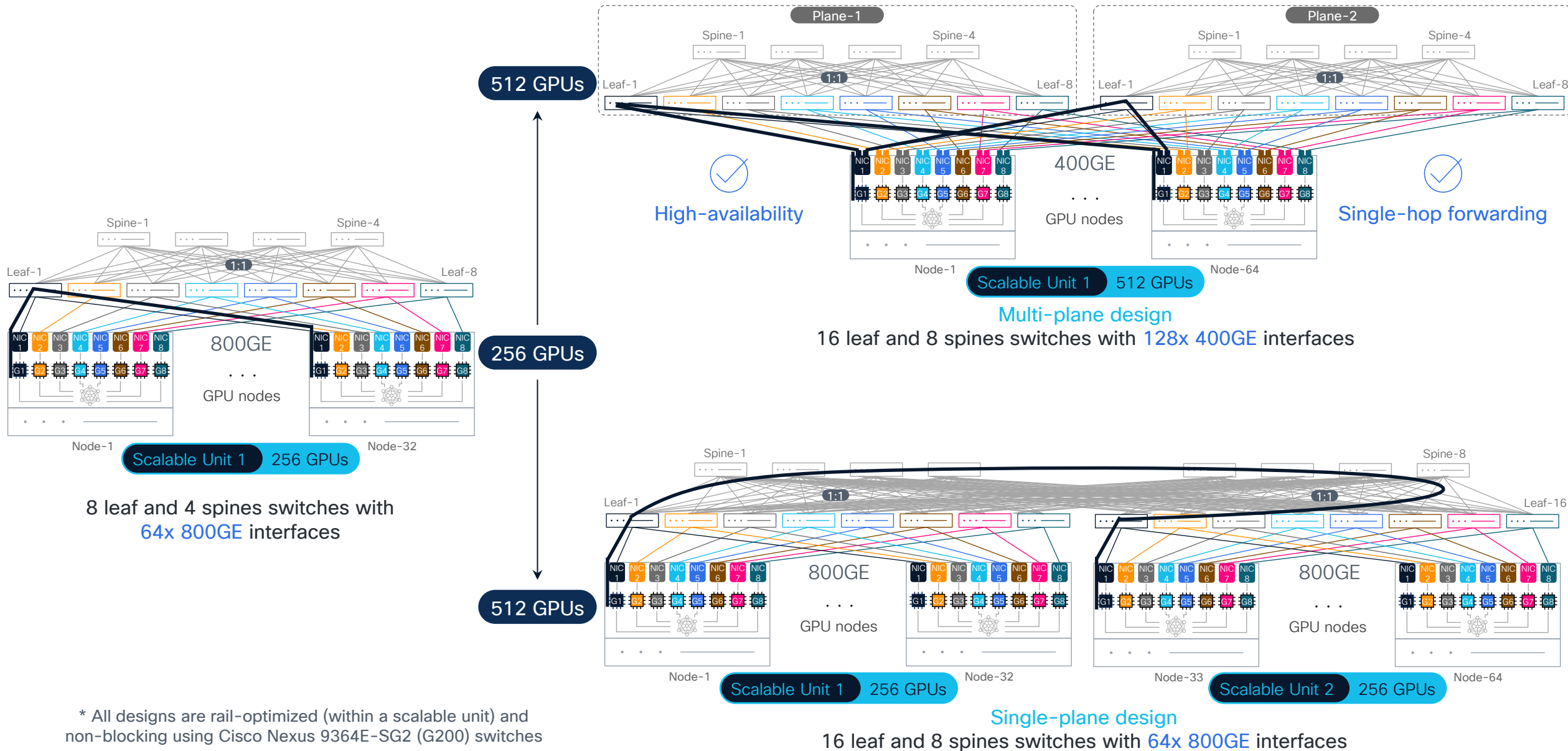


Dual-Plane Design



Many other options available on <https://tmgmatrix.cisco.com>

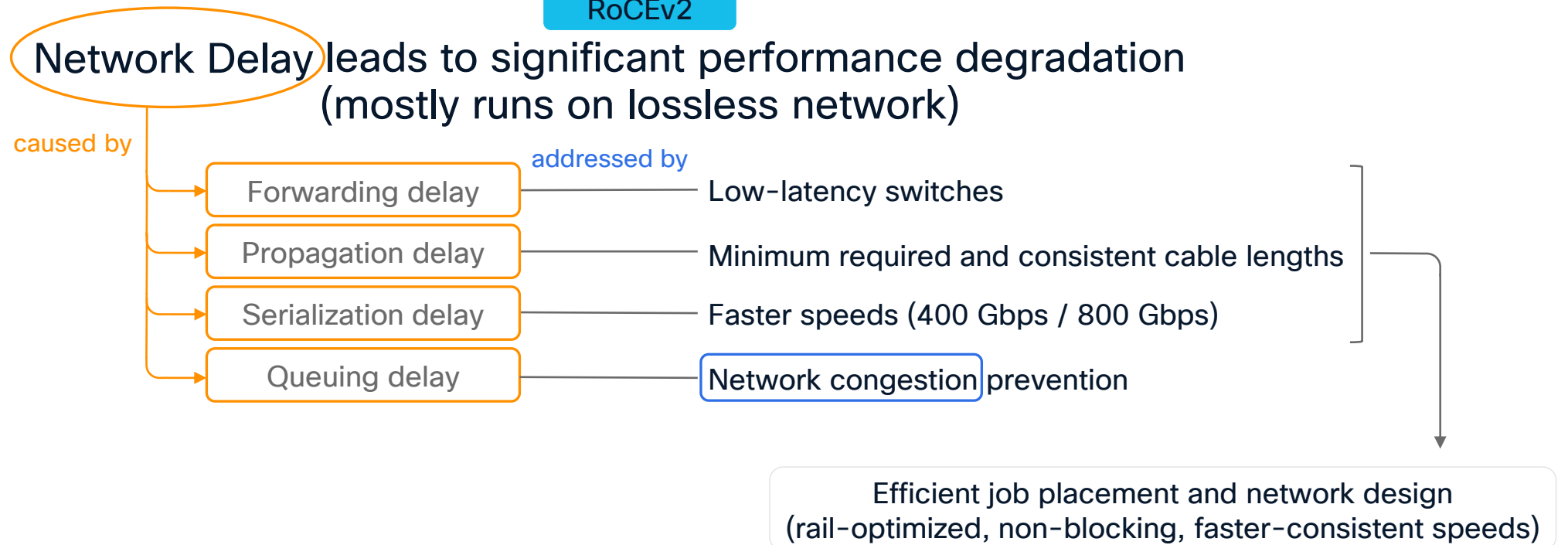
Evolution of Inter-GPU Backend Network Designs



RoCEv2



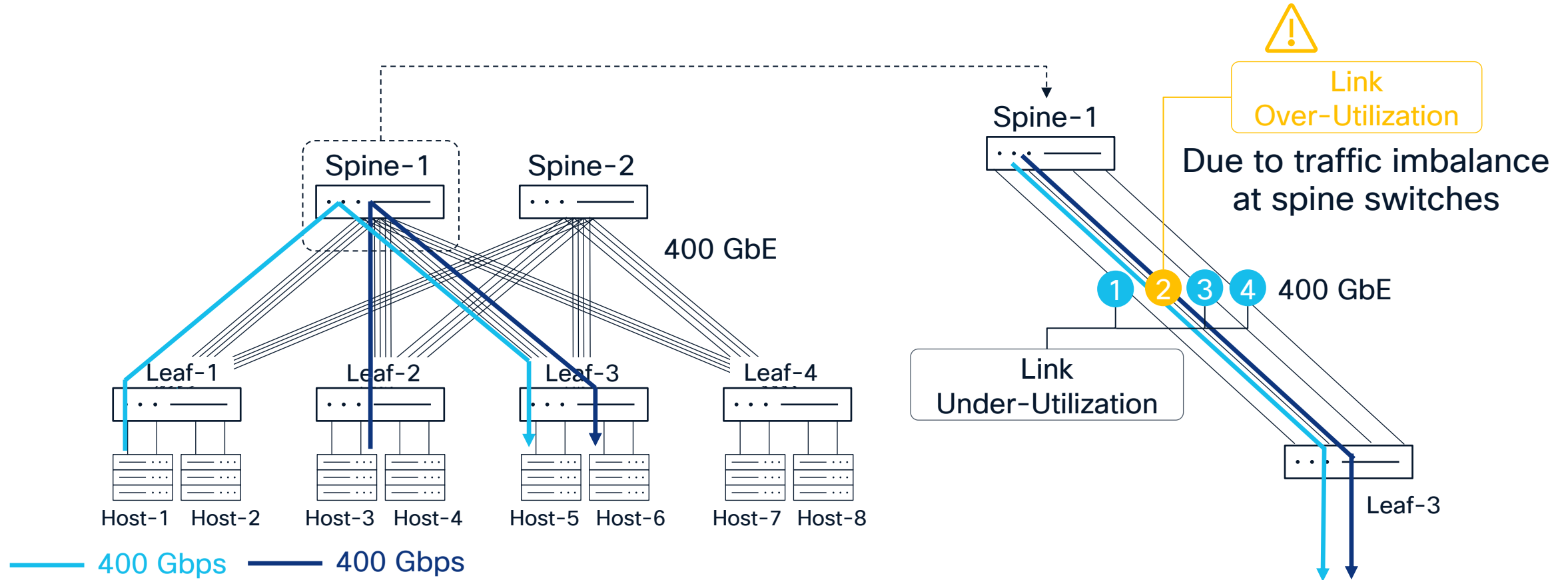
RoCEv2



Network congestion in AI Clusters

Network congestion in AI Clusters

Caused due to Traffic Imbalance



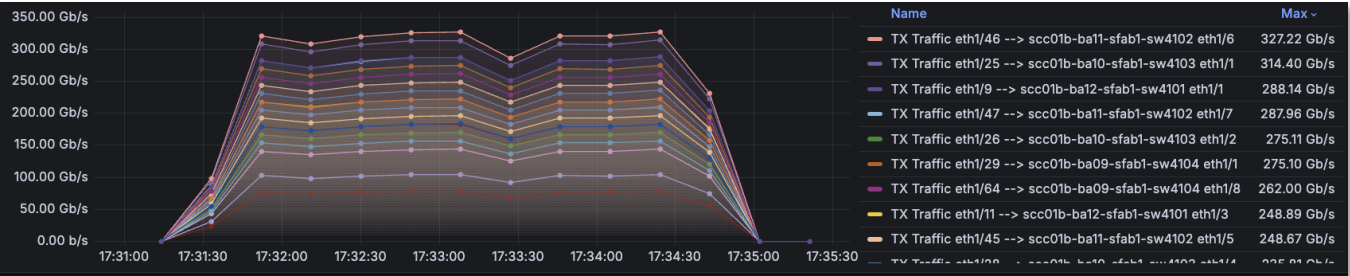
Default ECMP Load-Sharing based on 5-tuples

Source IP, Destination IP, UDP, Source Port, Destination Port

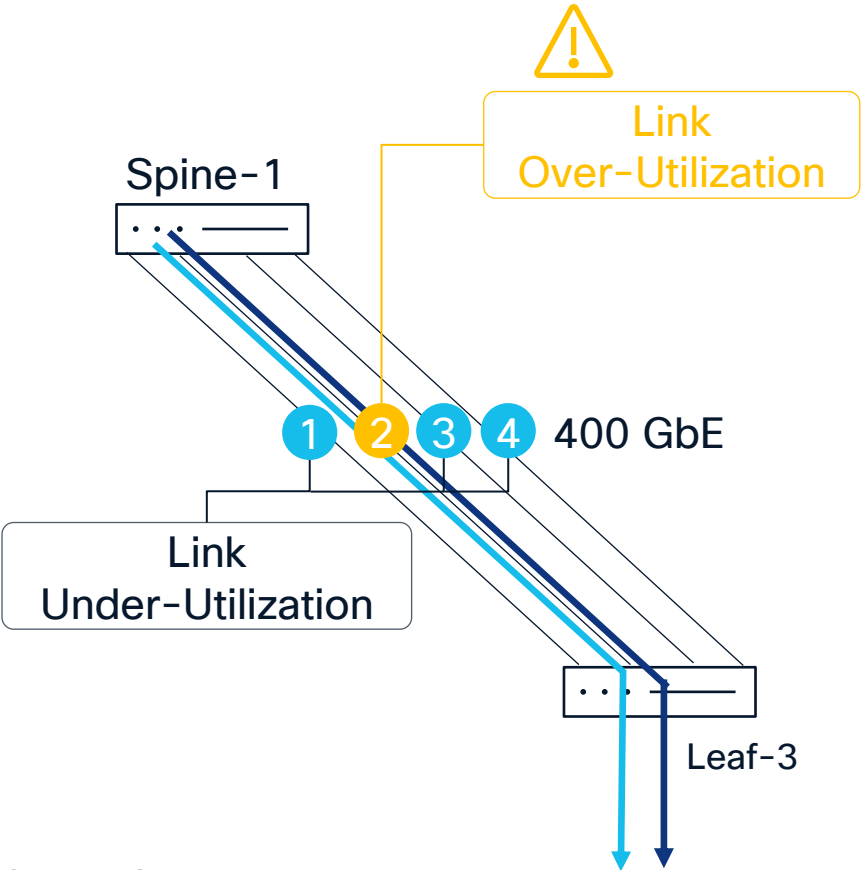
Congestion due to Over-Utilization

Flow Collision or Traffic Polarization

Default ECMP load-balancing is unaware of real-time link utilization



Behavior	Typical Data Center Network	Inter-GPU network
# of Layer-4 flows	Many	Fewer
Per-flow throughput	Many mouse flow and a few elephant flows	Mostly elephant flows
Link utilization	Not fully utilized	GPU send and receive traffic at the full capacity of their NIC ports



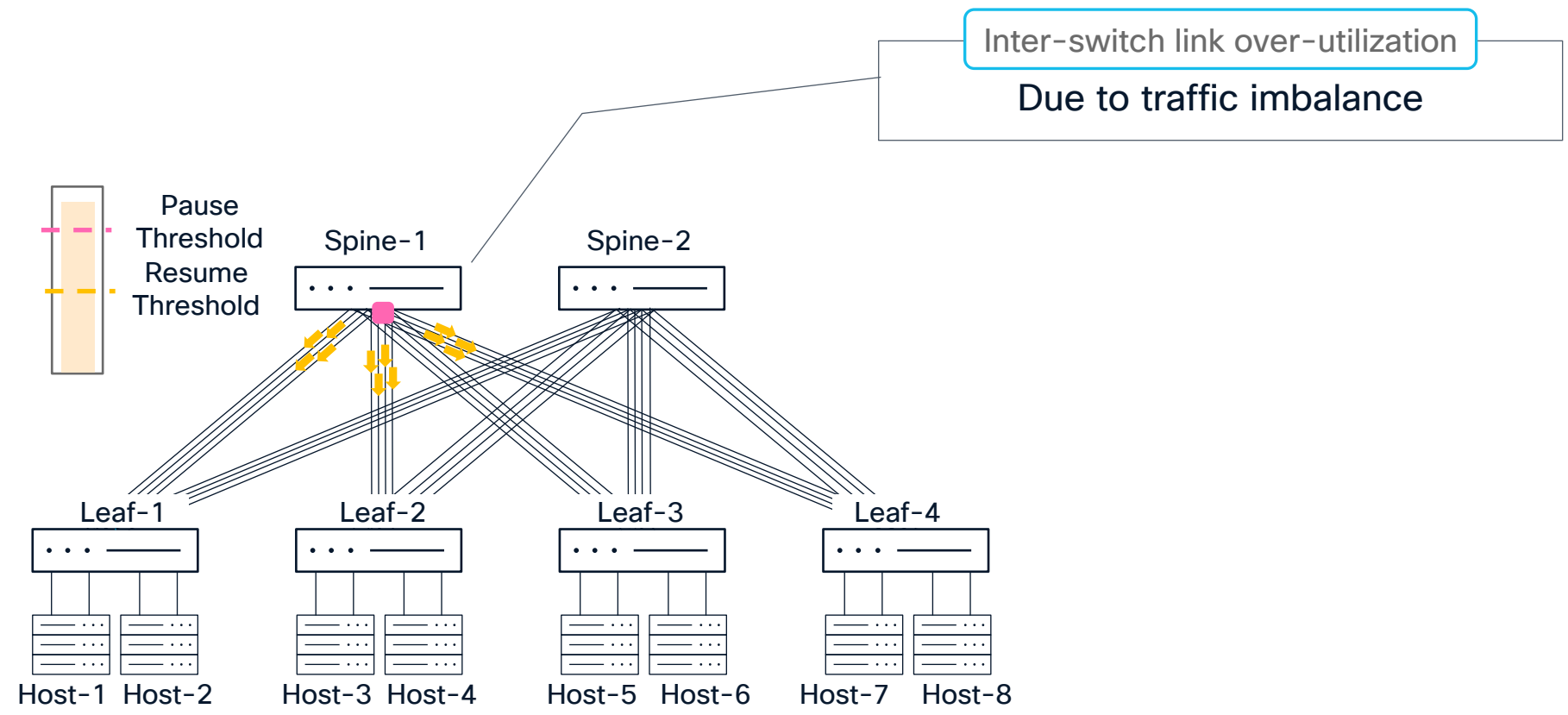
Low entropy in inter-GPU backend network:

- Layer-4 protocol (UDP), and destination port (4791) remain the same for RoCEv2
- # of IP addresses are limited, For example, a 256-GPU cluster uses only 256 IPs as source and destination
- Only UDP source port changes

Congestion Spreading in Lossless Networks

Inter-switch link Over-utilization

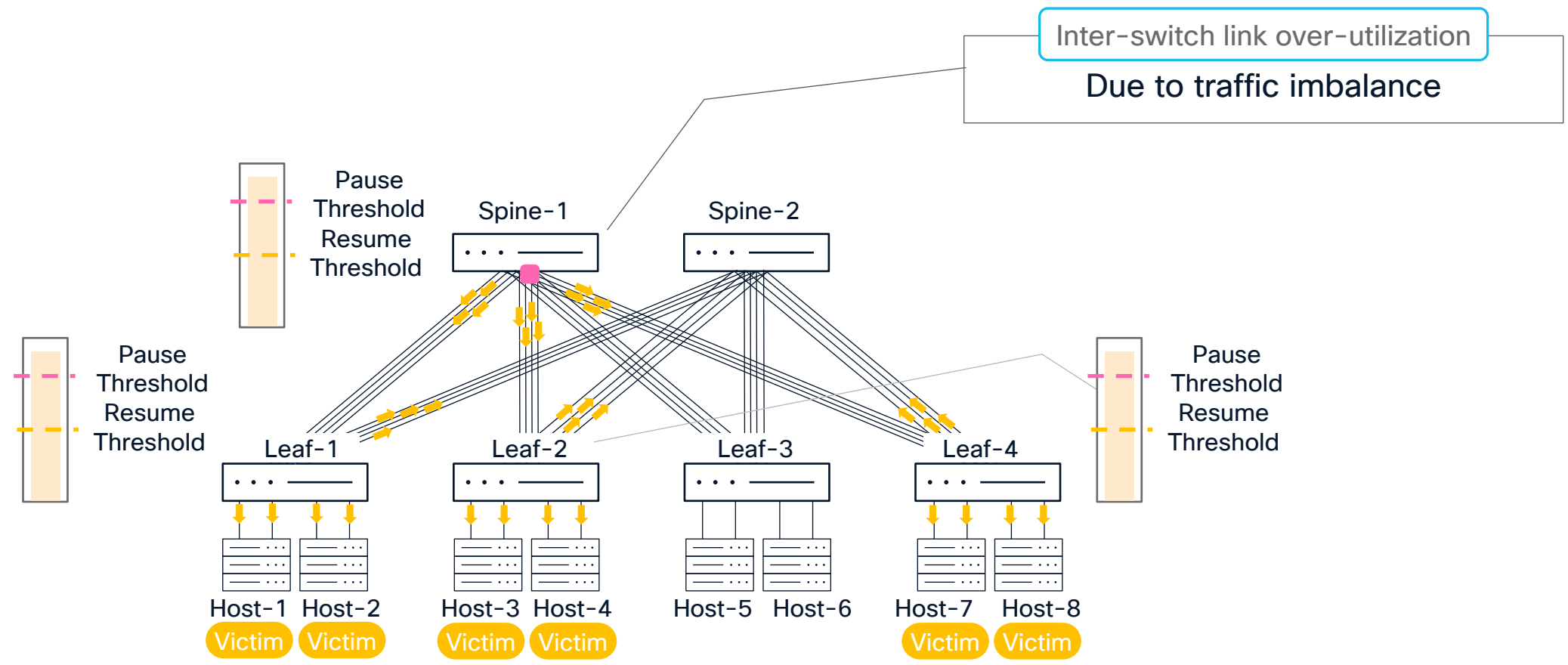
Pause frame 



Congestion Spreading in Lossless Networks

Inter-switch link Over-utilization

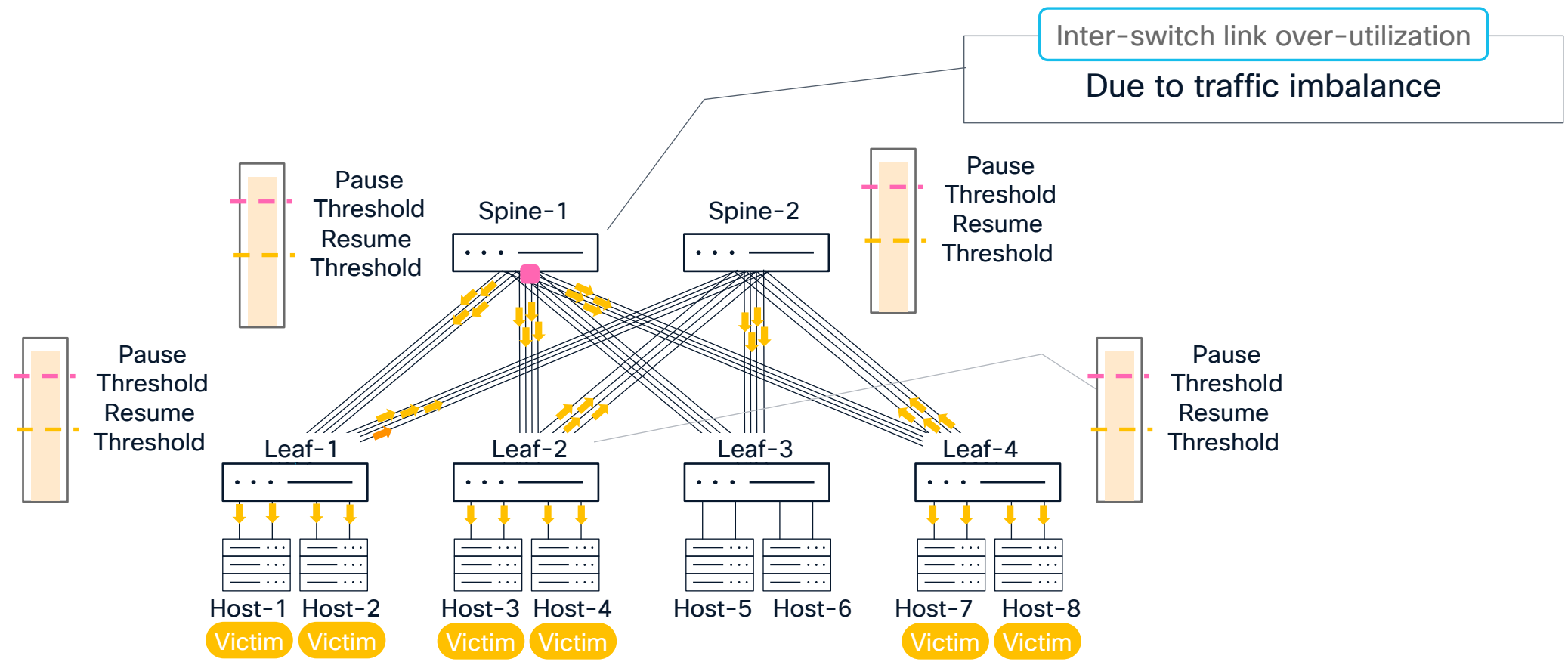
Pause frame 



Congestion Spreading in Lossless Networks

Inter-switch link Over-utilization

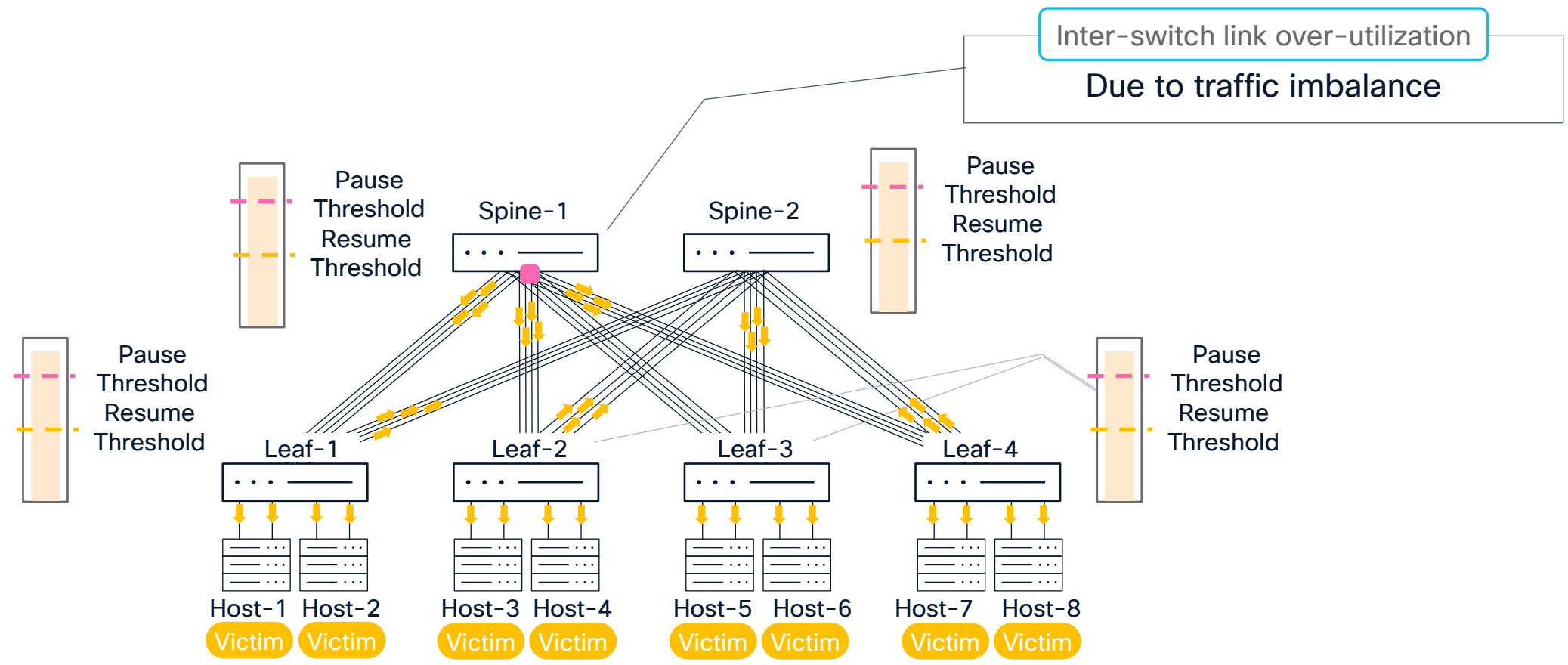
Pause frame 



Congestion Spreading in Lossless Networks

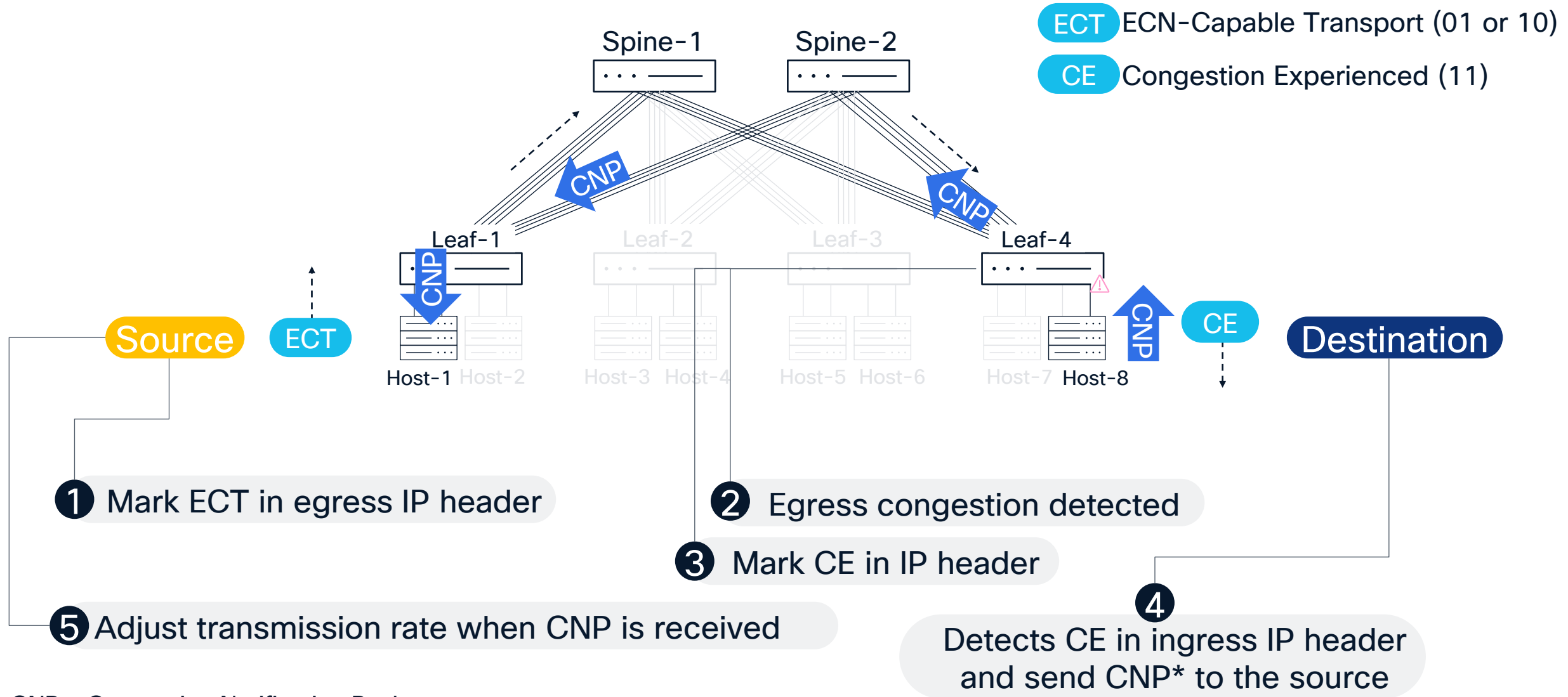
Inter-switch link Over-utilization

Pause frame 



RoCEv2 Congestion Management

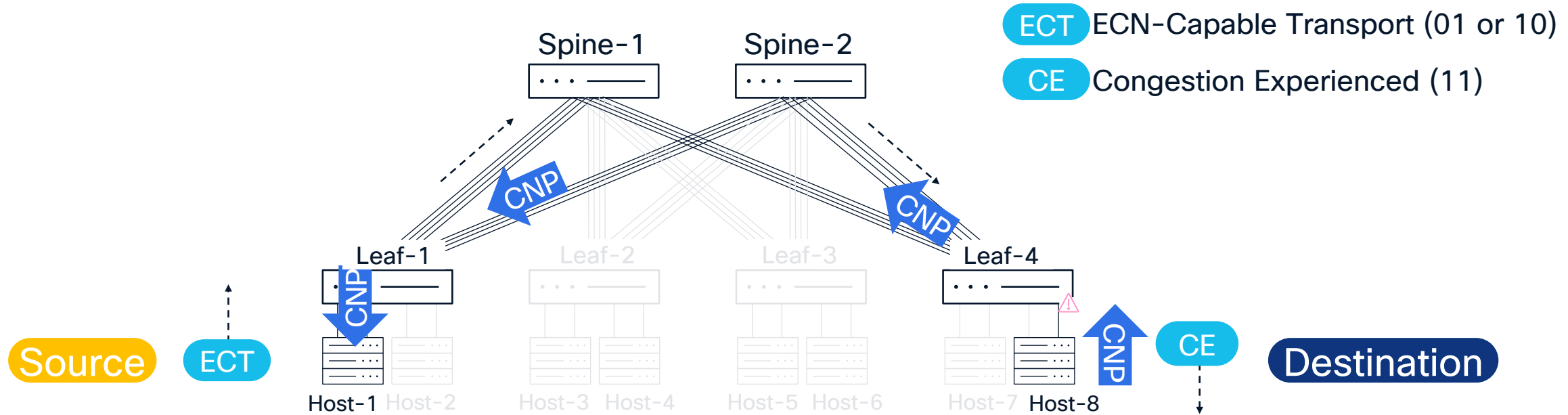
Data Center Quantized Congestion Notification (DCQCN)



* CNP - Congestion Notification Packet

RoCEv2 Congestion Management

DCQCN



ECN/CNP

Flow-level scope, **but its action is delayed**

PFC

Quick action and avoids packet loss, **but it has class-level scope**, and it leads to congestion spreading

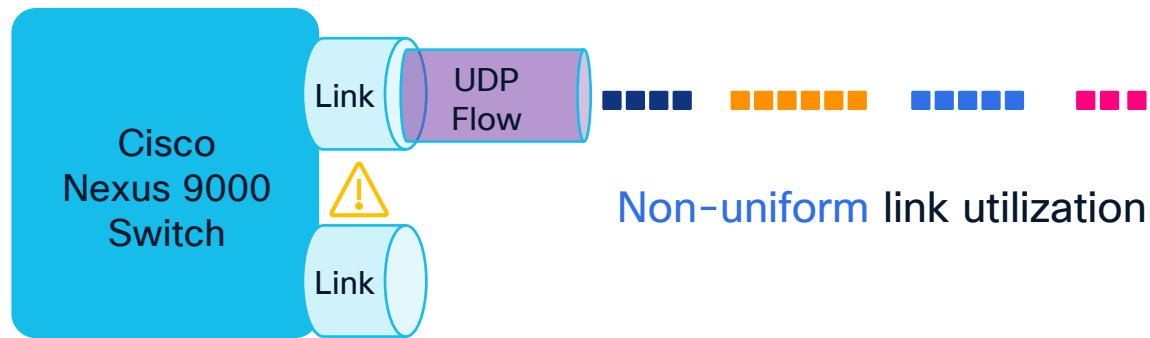
Together

Flow level scope, quick action, avoid packet loss.
Limits the spread and duration of congestion spreading.

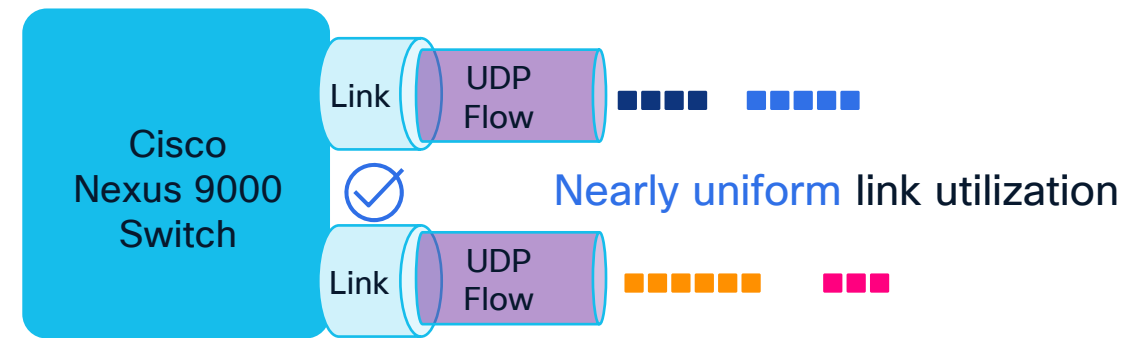
Improving Load Balancing in AI Networks

Flowlet Dynamic Load Balancing (DLB)

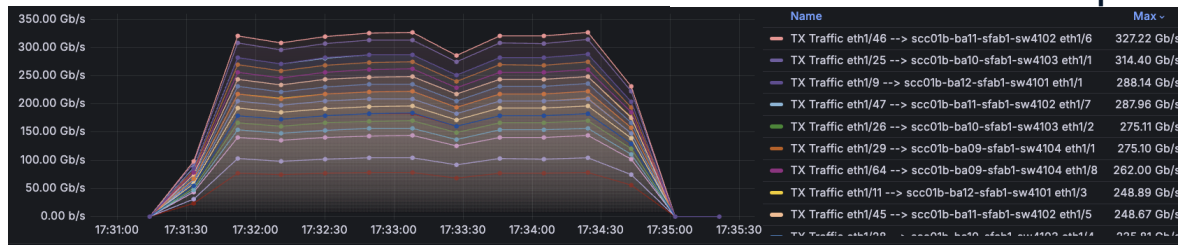
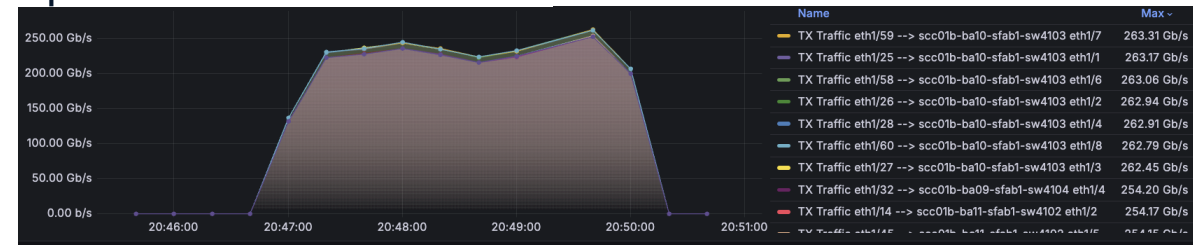
Default ECMP load-sharing based on Layer 4 UDP flow



DLB based on real-time usage and congestion states of links and flowlets identified by inter-packet gap



Utilization of all equal-cost paths on a leaf switch



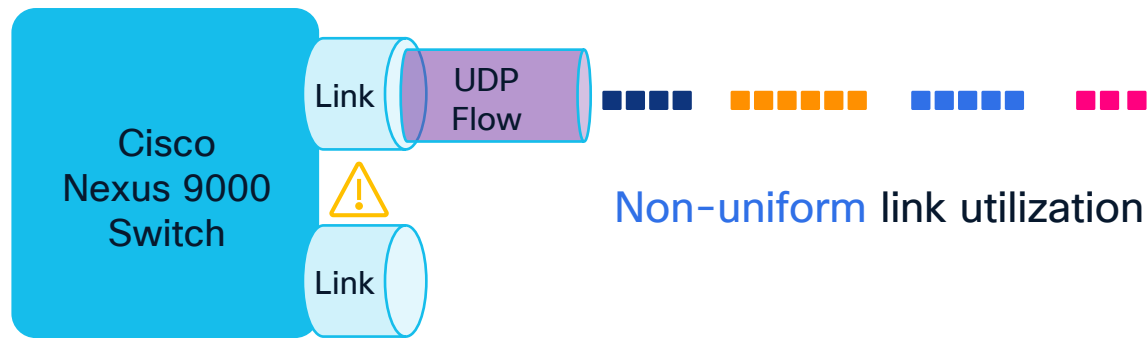
AIReduce 256 GPUs
(Max: 327.22 Gbps, Min: 76.68 Gbps)

AIReduce 256 GPUs
(Max: 263.31 Gbps, Min: 253.66 Gbps)
Nearly uniform

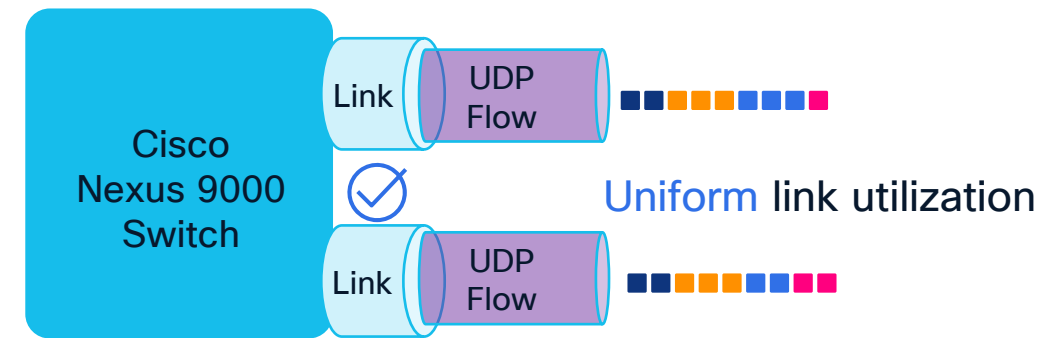
Improving Load Balancing in AI Networks

Cisco Per-Packet Dynamic Load Balancing (Packet Spray) + NVIDIA Spectrum-X Adaptive Routing

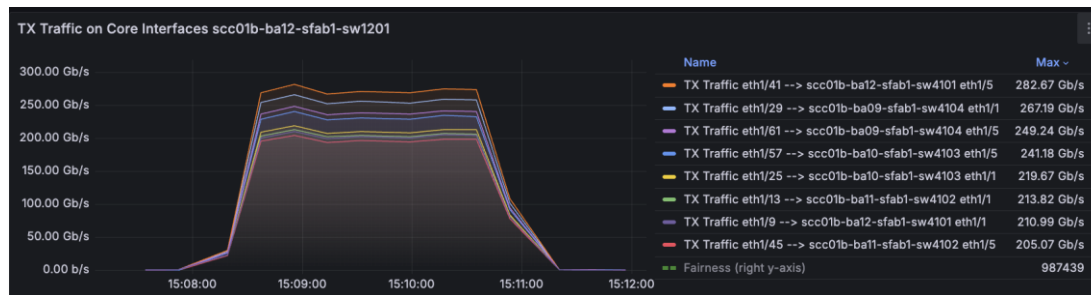
Default ECMP load-sharing based on Layer 4 UDP flow



Per-packet DLB based on real-time usage and congestion states of links

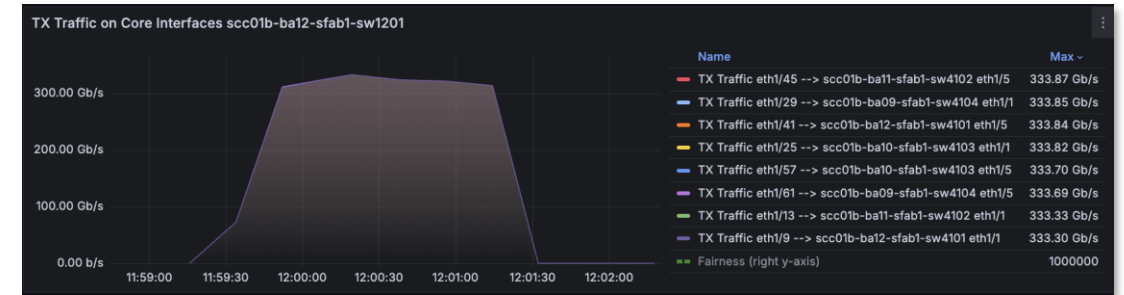


Utilization of all equal-cost paths on a leaf switch



AlltoAll 64 GPUs

(Max: 282 Gbps, Min: 205 Gbps)



AlltoAll 64 GPUs

(333 Gbps)

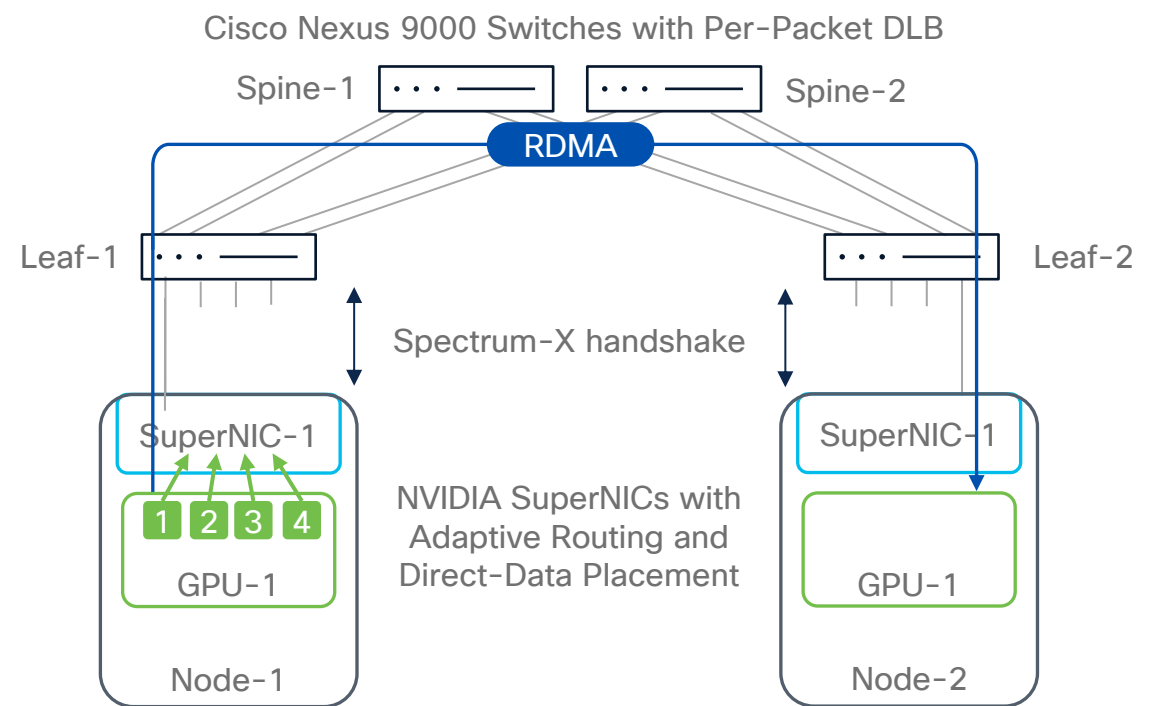
Uniform

Improving Load Balancing in AI Networks

Cisco Per-Packet Dynamic Load Balancing (DLB) + NVIDIA Spectrum-X Adaptive Routing

1

Cisco Nexus 9000 Switches benefit from the NVIDIA Spectrum-X technology after a handshake with the NVIDIA SuperNICs

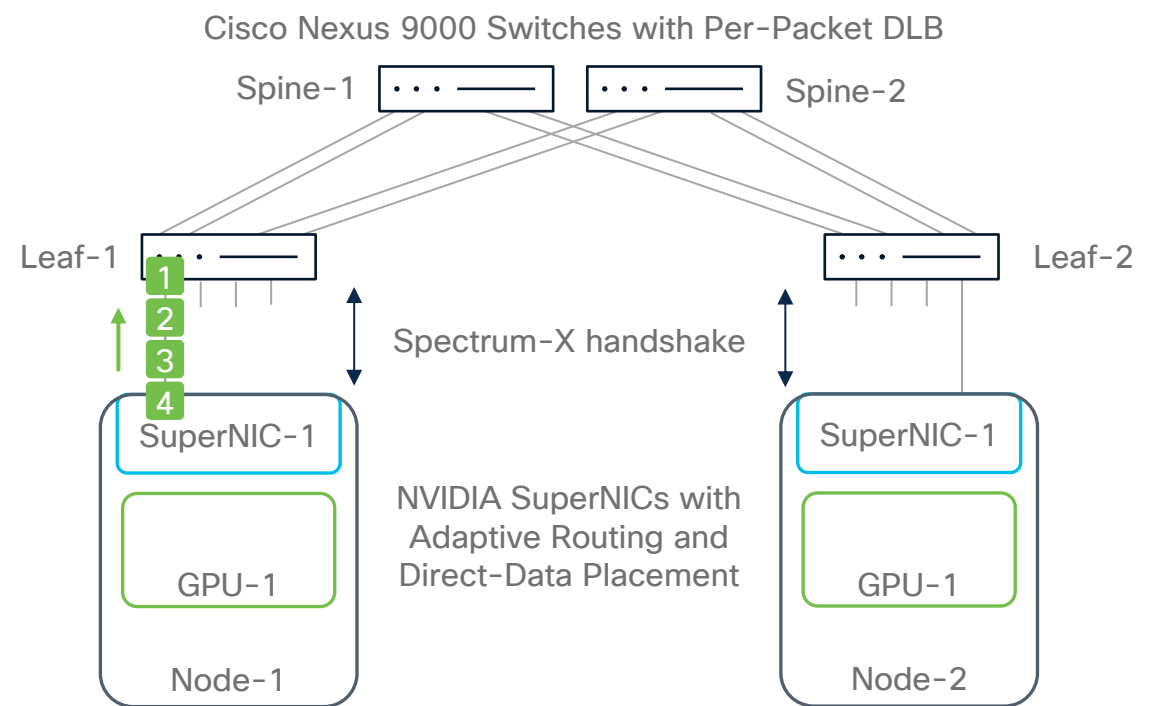


Improving Load Balancing in AI Networks

Cisco Per-Packet Dynamic Load Balancing (DLB) + NVIDIA Spectrum-X Adaptive Routing

2

Source SuperNIC on Node-1 sends data to destination SuperNIC on Node-2 in multiple packets

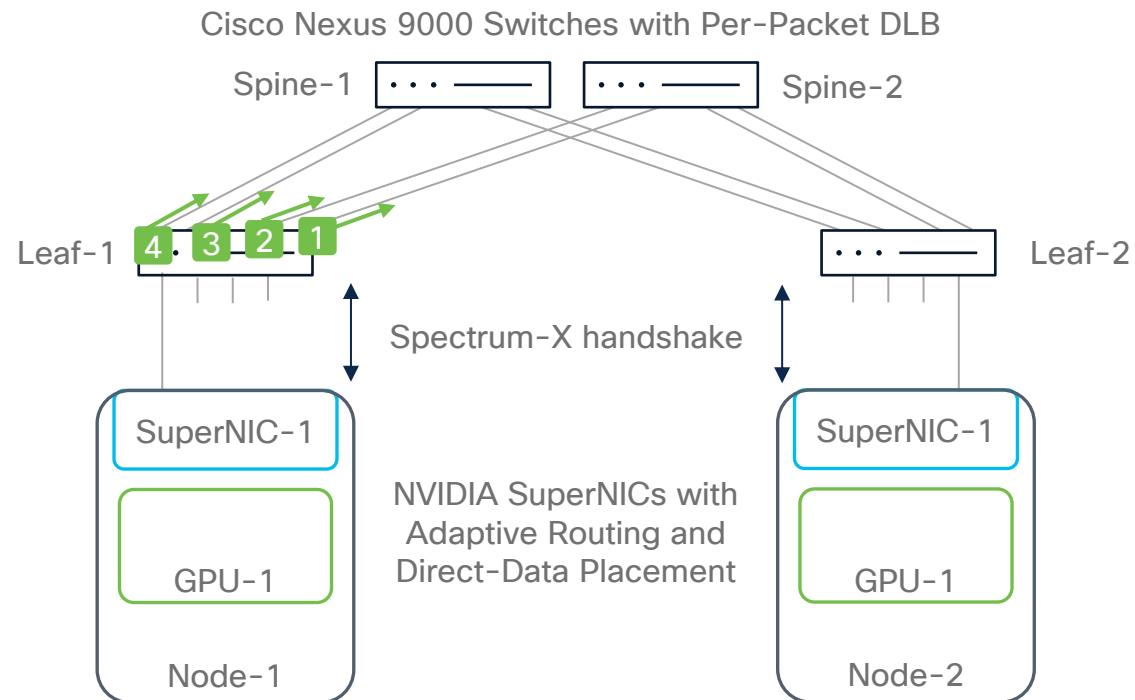


Improving Load Balancing in AI Networks

Cisco Per-Packet Dynamic Load Balancing (DLB) + NVIDIA Spectrum-X Adaptive Routing

3

Leaf-1 sprays packets (per-packet DLB) based on real-time utilization and congestion states of the links, thereby avoiding congestion due to over-utilization on leaf-spine links

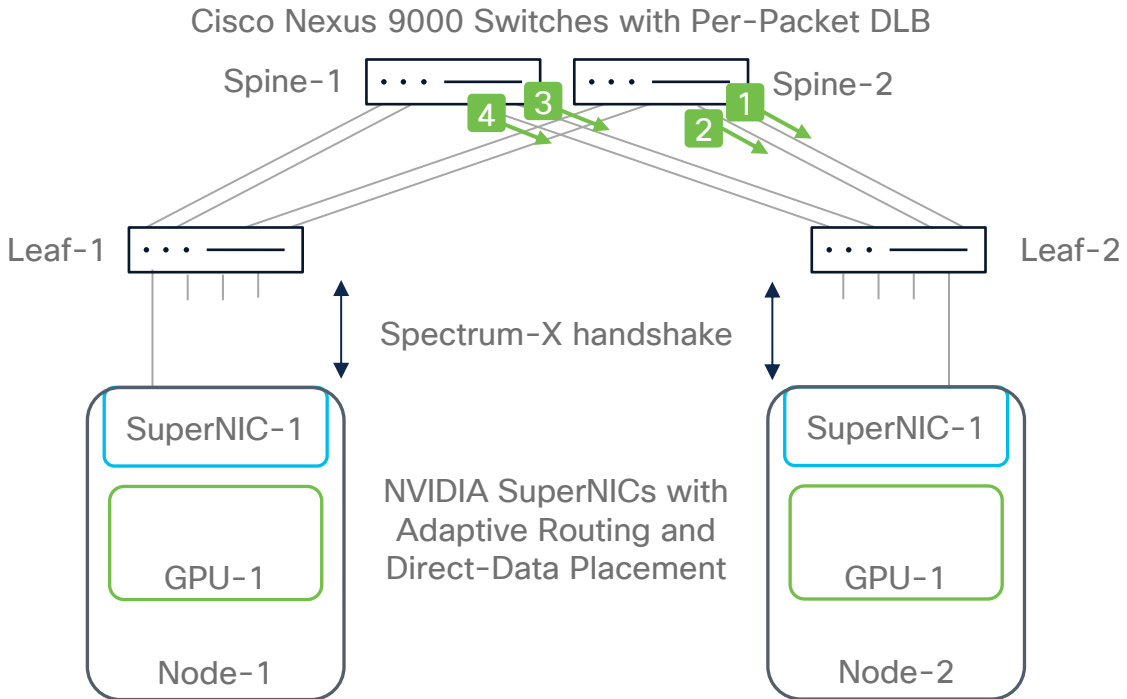


Improving Load Balancing in AI Networks

Cisco Per-Packet Dynamic Load Balancing (DLB) + NVIDIA Spectrum-X Adaptive Routing

4

Spine switches also spray packets based on real-time utilization and congestion states of the links, thereby avoiding congestion due to over-utilization on leaf-spine links

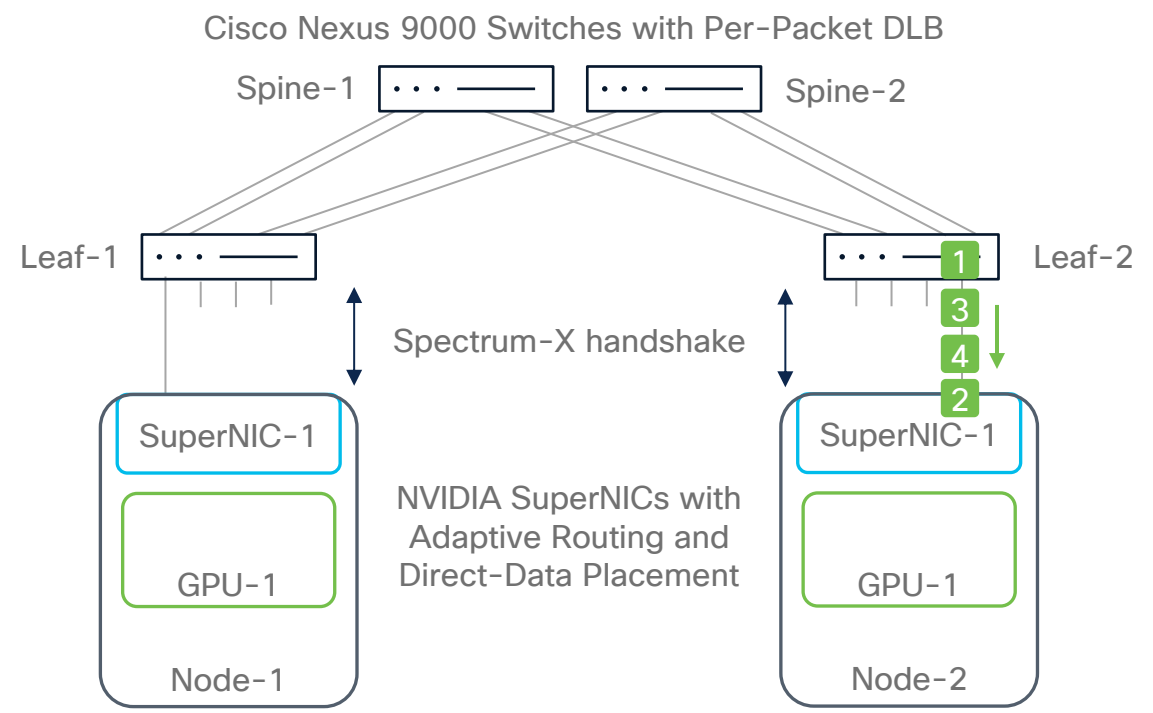


Improving Load Balancing in AI Networks

Cisco Per-Packet Dynamic Load Balancing (DLB) + NVIDIA Spectrum-X Adaptive Routing

5

Packets may arrive out-of-order on SuperNIC-1 on Node-2 due to packet spray by the switches

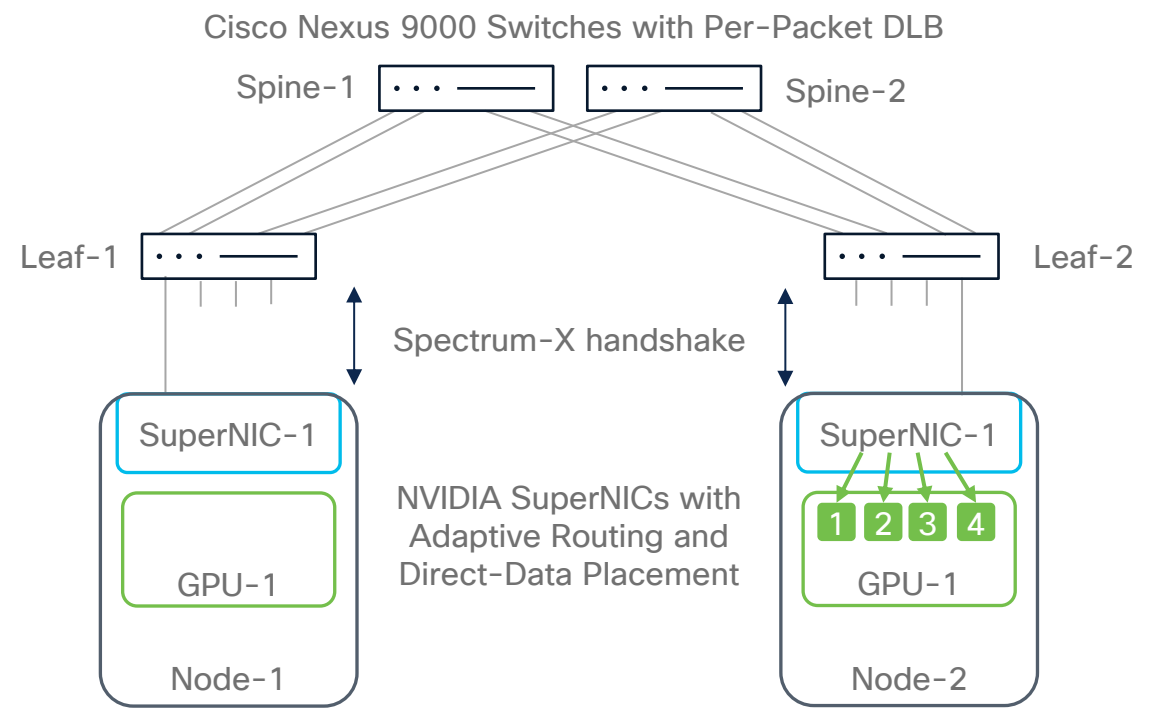


Improving Load Balancing in AI Networks

Cisco Per-Packet Dynamic Load Balancing (DLB) + NVIDIA Spectrum-X Adaptive Routing

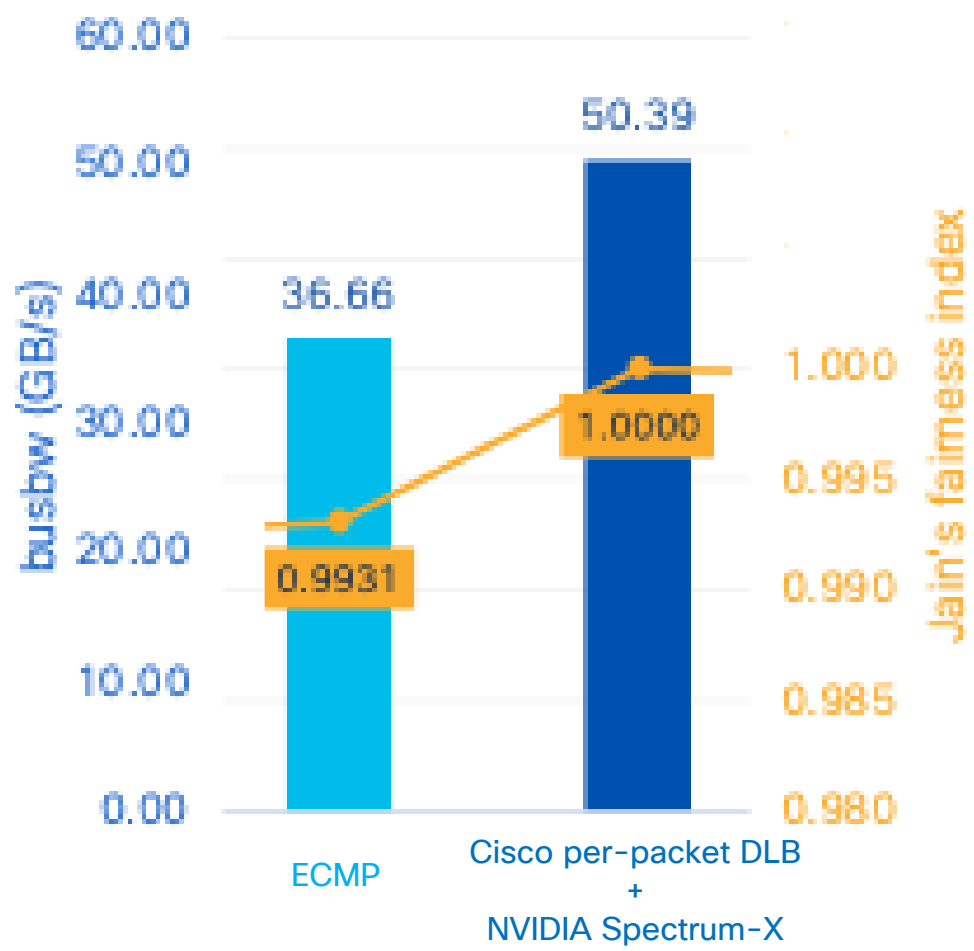
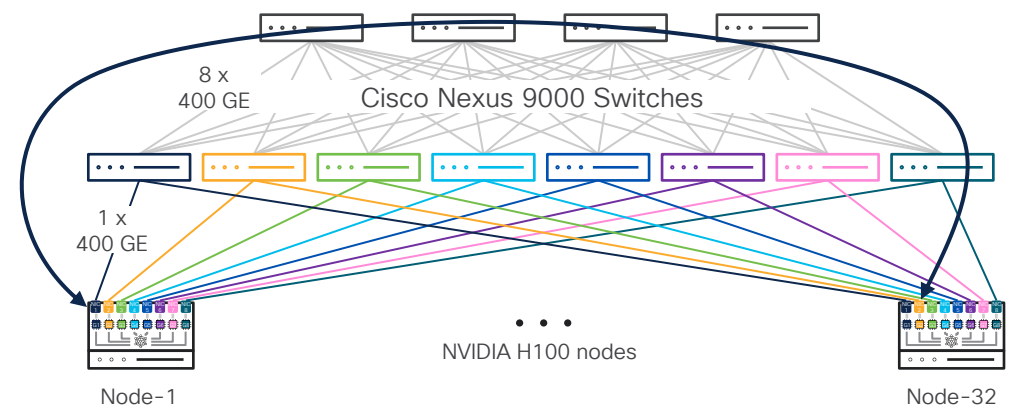
6

Finally, the receiving SuperNIC places packet payload in the correct order in the GPU memory using the NVIDIA Direct Data Placement (DDP) technology



Cisco Per-Packet DLB + NVIDIA Spectrum-X

Performance gain



AlltoAll collective among 64 NVIDIA H100 GPUs connected with Cisco Nexus 9000 switches

37% performance gain

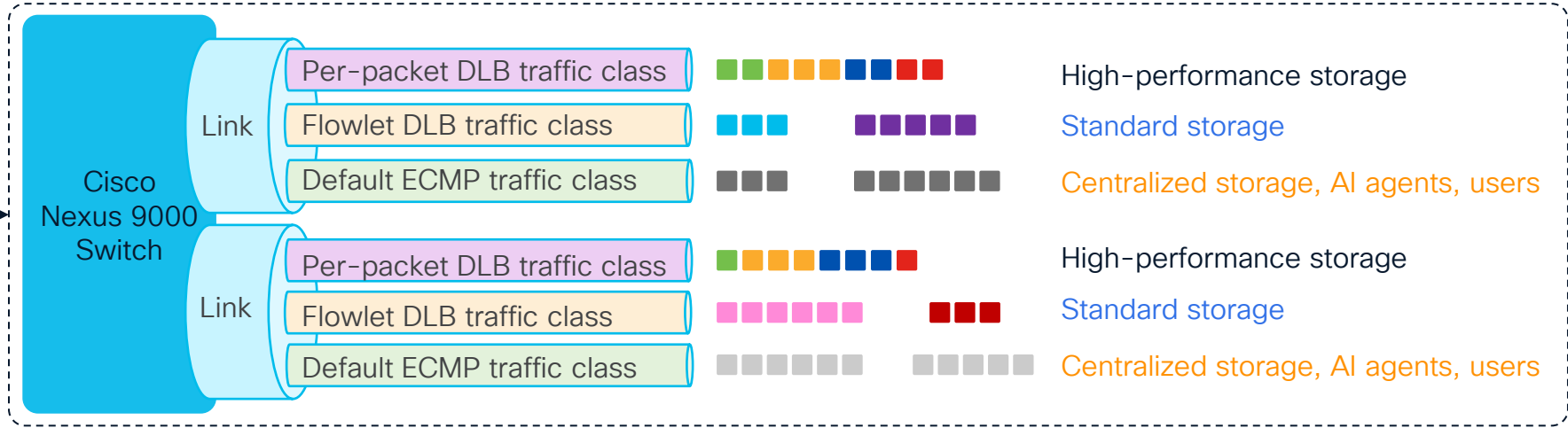
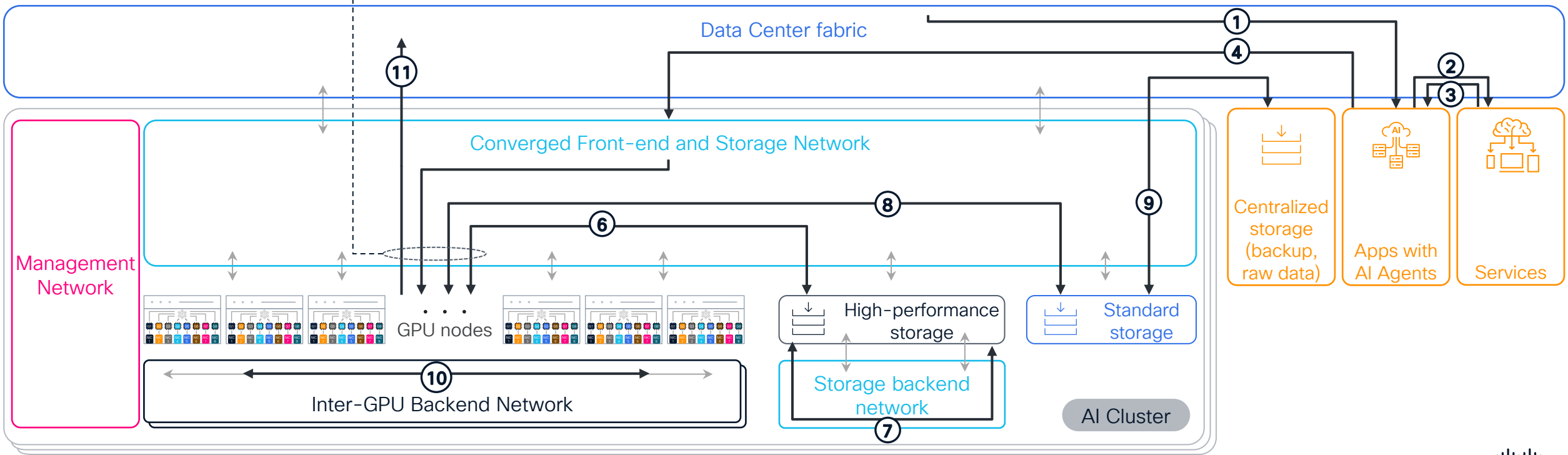
With Cisco per-packet DLB + NVIDIA Spectrum-X Adaptive Routing

Demo: <https://www.youtube.com/watch?v=FtKZOCzgP4>

Policy-Driven DLB on Cisco Nexus 9000 (Silicon One G200)

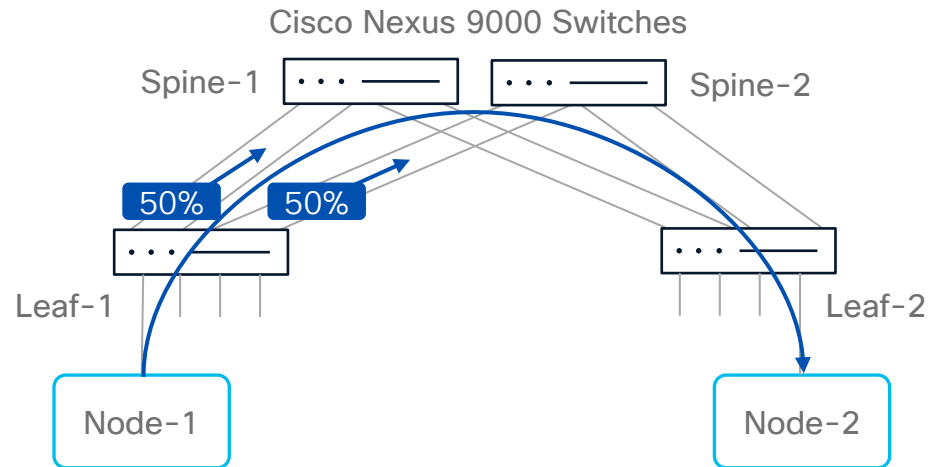
Proactive congestion prevention in backend and storage networks by simultaneous use of different load-balancing schemes, such as flowlet, per-packet, or ECMP, for different traffic types

Result
Increased network utilization and GPU utilization

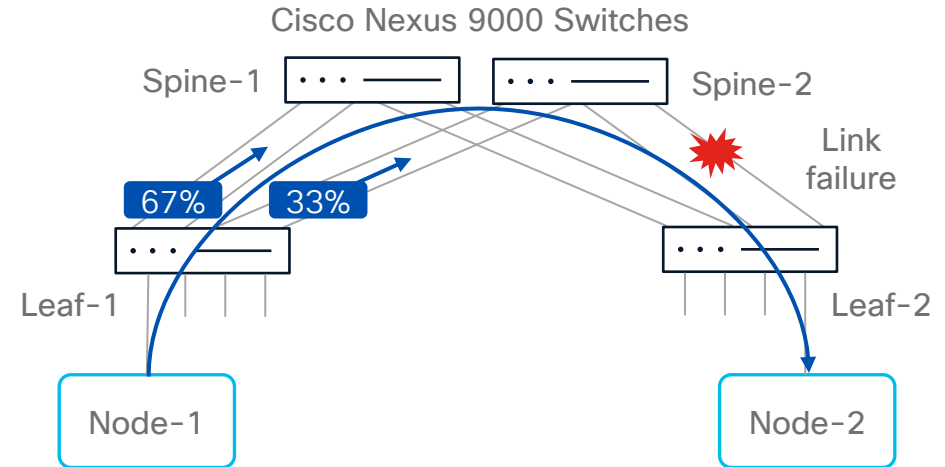


DLB + WCMP on Cisco Nexus 9000 (G200)

Uniform link utilization because of DLB on Leaf-1, when both spines have equal capacity to send traffic to Leaf-2



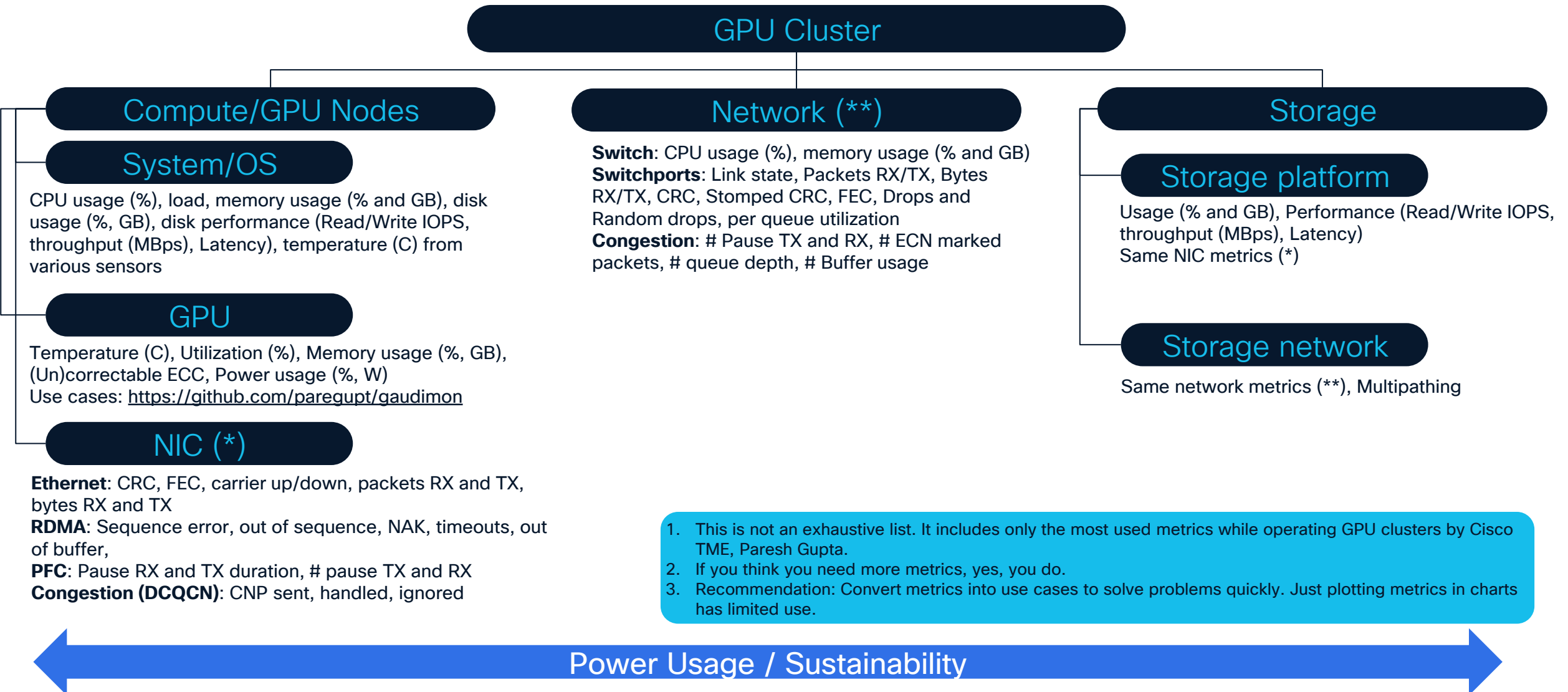
DLB with WCMP allows Leaf-1 to send traffic to each spine proportional to that spine's capacity to send traffic to the Leaf-2



Configuring Inter-GPU Backend Networks

1. **Routing**: Configure routing in the fabric: Prefer eBGP because of features, but other routing protocols can also be used
 - a. Configure host-connected switchports as routed (no switchport) in /30 or /31
 - b. Redistribute these subnets for all-to-all reachability among the GPU NICs
 - c. Enable WCMP, if using eBGP
2. **MTU**: Increased MTU on host interfaces (NICs) to 4KB. Match the same on switches (+ header overhead, typically 4200 bytes), or just increase to maximum of 9216 bytes
3. **PFC**: Enable lossless networking for RoCEv2 traffic by enabling Priority-based Flow Control (PFC)
 - a. Enable PFC watchdog to isolate a stuck NIC (source of congestion)
4. **Congestion management**: If using DCQCN
 - a. Enable WRED to mark ECN bits for RoCEv2 traffic
 - b. Prioritize sending of Congestion Notification Packet (CNP) via priority queue on switches
5. **DLB**: Enable DLB, if needed.
6. **Observability**: Network, GPU nodes, and storage nodes

GPU Metrics

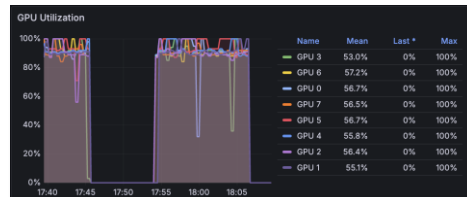
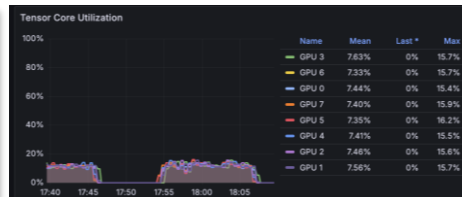


1. This is not an exhaustive list. It includes only the most used metrics while operating GPU clusters by Cisco TME, Paresh Gupta.
2. If you think you need more metrics, yes, you do.
3. Recommendation: Convert metrics into use cases to solve problems quickly. Just plotting metrics in charts has limited use.

GPU Cluster Observability – Metrics

NVIDIA DCGM for Metrics Export

GPU Temperature from two nodes in the same cluster



```
# nvidia-smi
Mon Feb  3 18:02:01 2025
```

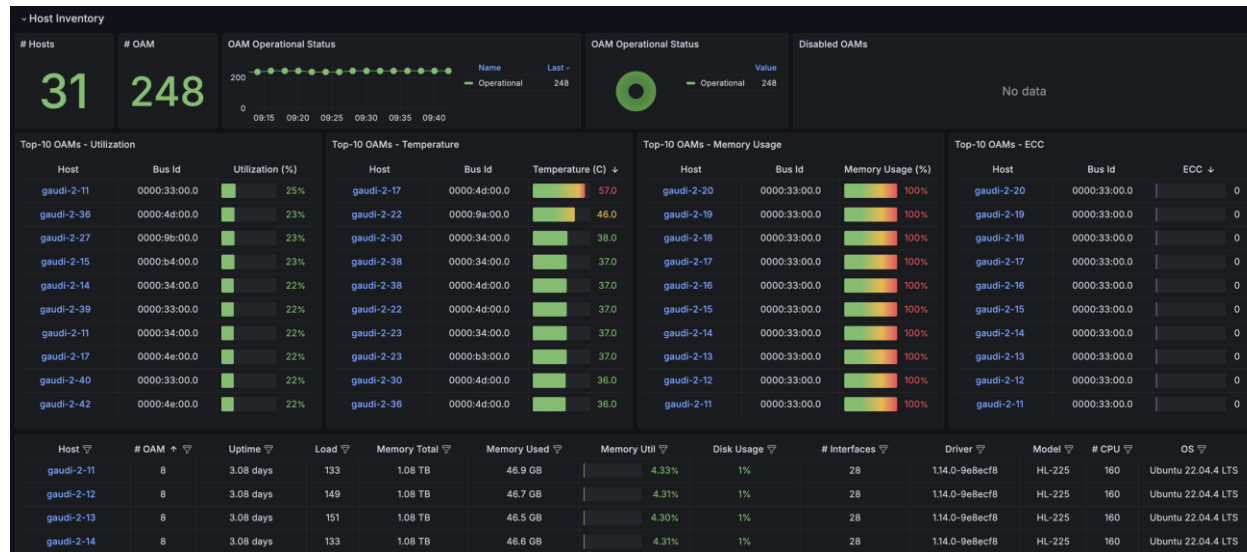
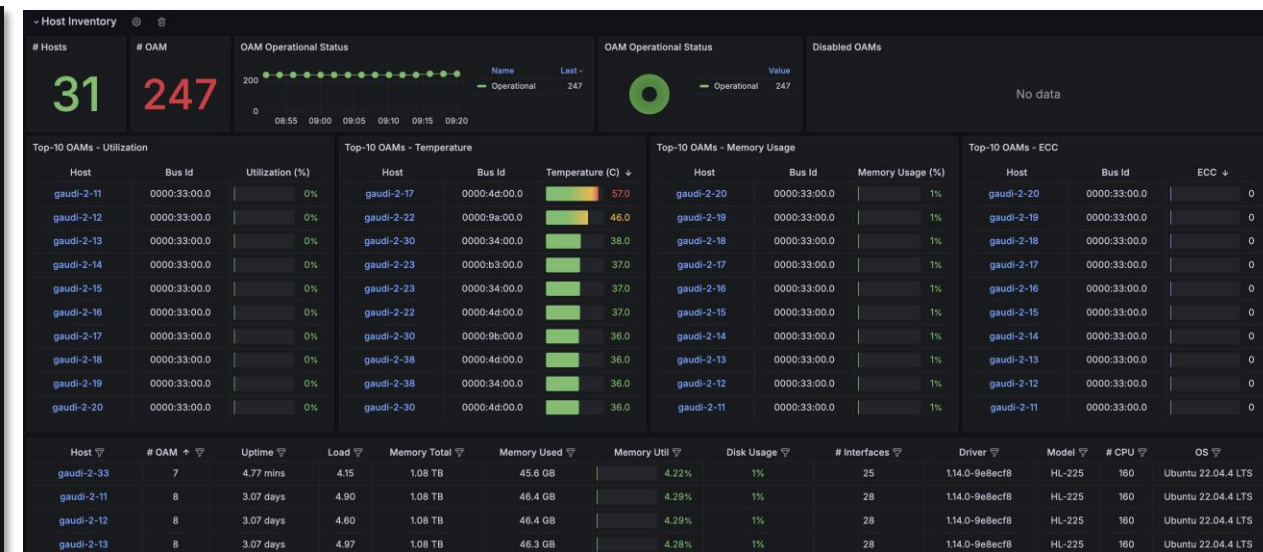
NVIDIA-SMI 535.129.03		Driver Version: 535.129.03		CUDA Version: 12.2	
GPU	Name	Persistence-M	Bus-Id	Disp.A	Volatile Uncorr. ECC
Fan	Temp	Pwr:Usage/Cap	Memory-Usage	GPU-Util	Compute M. MIG M.
Perf					
0	NVIDIA H100 80GB HBM3	On	00000000:1B:00:0	Off	0
N/A	44C P0	307W / 700W	73557MiB / 81559MiB	92%	Default Disabled
1	NVIDIA H100 80GB HBM3	On	00000000:43:00:0	Off	0
N/A	45C P0	297W / 700W	71467MiB / 81559MiB	90%	Default Disabled
2	NVIDIA H100 80GB HBM3	On	00000000:52:00:0	Off	0
N/A	55C P0	318W / 700W	71467MiB / 81559MiB	91%	Default Disabled
3	NVIDIA H100 80GB HBM3	On	00000000:61:00:0	Off	0
N/A	54C P0	317W / 700W	71467MiB / 81559MiB	91%	Default Disabled
4	NVIDIA H100 80GB HBM3	On	00000000:9D:00:0	Off	0
N/A	45C P0	302W / 700W	71467MiB / 81559MiB	91%	Default Disabled
5	NVIDIA H100 80GB HBM3	On	00000000:C3:00:0	Off	0
N/A	44C P0	304W / 700W	71467MiB / 81559MiB	92%	Default Disabled
6	NVIDIA H100 80GB HBM3	On	00000000:D1:00:0	Off	0
N/A	55C P0	309W / 700W	71467MiB / 81559MiB	92%	Default Disabled
7	NVIDIA H100 80GB HBM3	On	00000000:DF:00:0	Off	0
N/A	57C P0	320W / 700W	70987MiB / 81559MiB	92%	Default Disabled

Processes:						
GPU	GI ID	CI ID	PID	Type	Process name	GPU Memory Usage
0	N/A	N/A	1918069	C	/opt/conda/bin/python3.11	73504MiB
1	N/A	N/A	1918444	C	/opt/conda/bin/python3.11	71414MiB
2	N/A	N/A	1918445	C	/opt/conda/bin/python3.11	71414MiB
3	N/A	N/A	1918446	C	/opt/conda/bin/python3.11	71414MiB
4	N/A	N/A	1918447	C	/opt/conda/bin/python3.11	71414MiB
5	N/A	N/A	1918448	C	/opt/conda/bin/python3.11	71414MiB
6	N/A	N/A	1918449	C	/opt/conda/bin/python3.11	71414MiB
7	N/A	N/A	1918450	C	/opt/conda/bin/python3.11	70934MiB

GPU Cluster Observability – Missing GPU

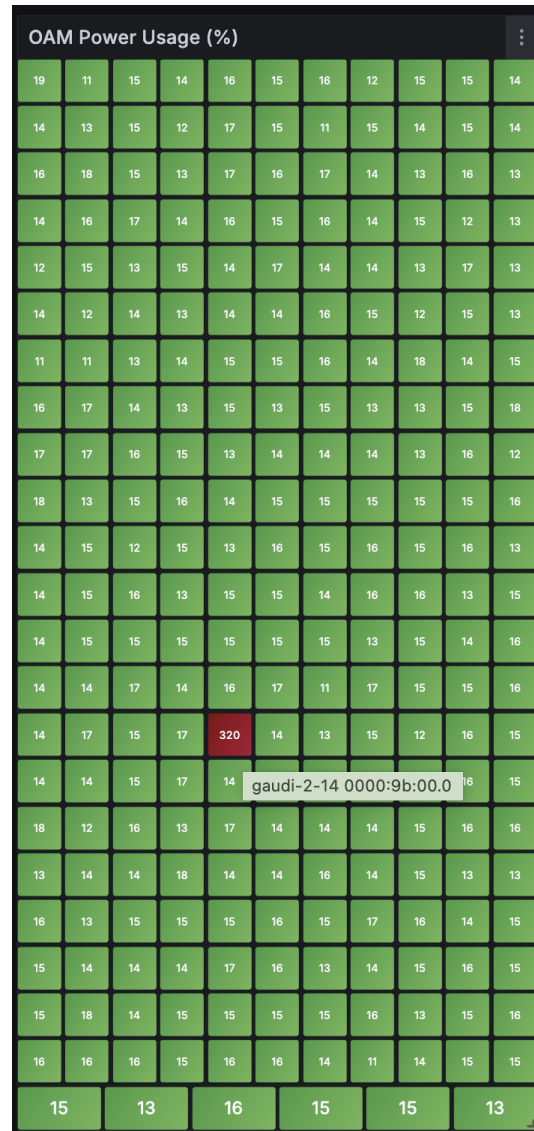
<https://github.com/paregupt/gaudimon>

Spot the difference to find the problem



GPU Monitoring Use Cases – Faulty GPU

<https://github.com/paregupt/gaudimon>



RoCEv2 Network Monitoring

Nexus Traffic Monitoring (NTM)

Purposefully built for Inter-GPU
RoCEv2 Lossless Ethernet networks

Uses NX-API and gNMI for metric
collection from NX-OS Switches

Open source. No fee, no support

Use Cases:

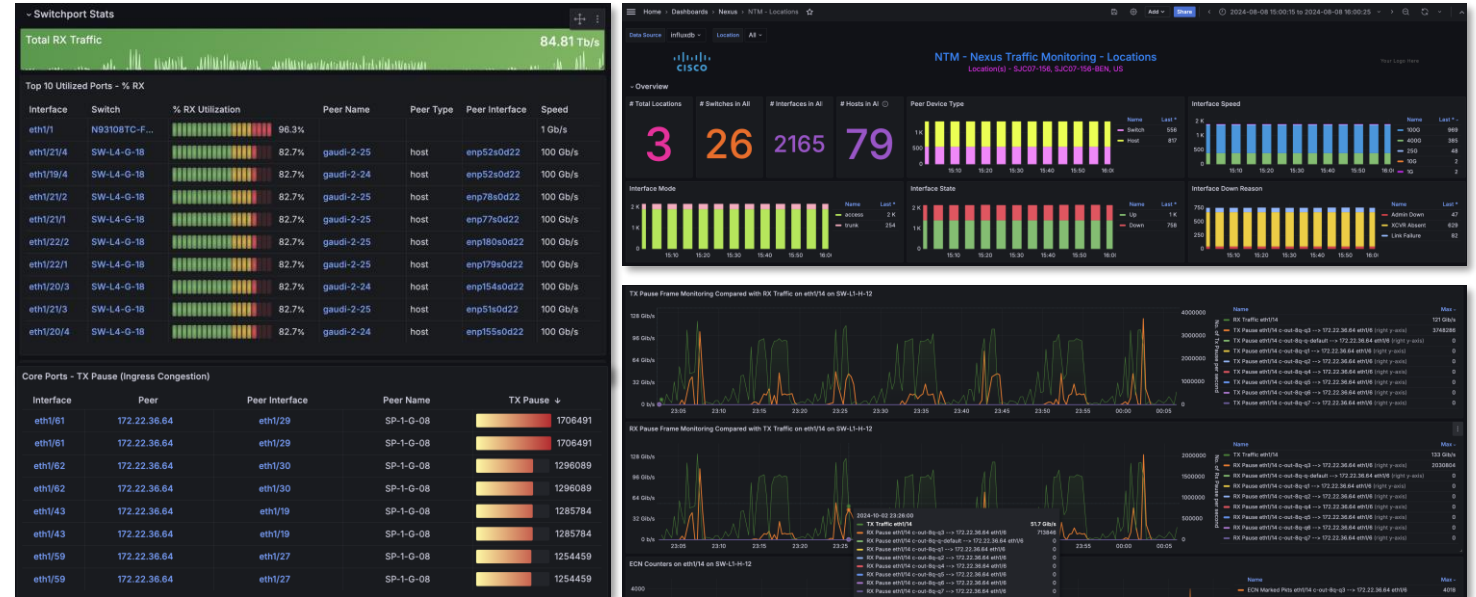
PFC Pause Frames

ECN

Correlation with GPU NICs

Cable connectivity

Inventory

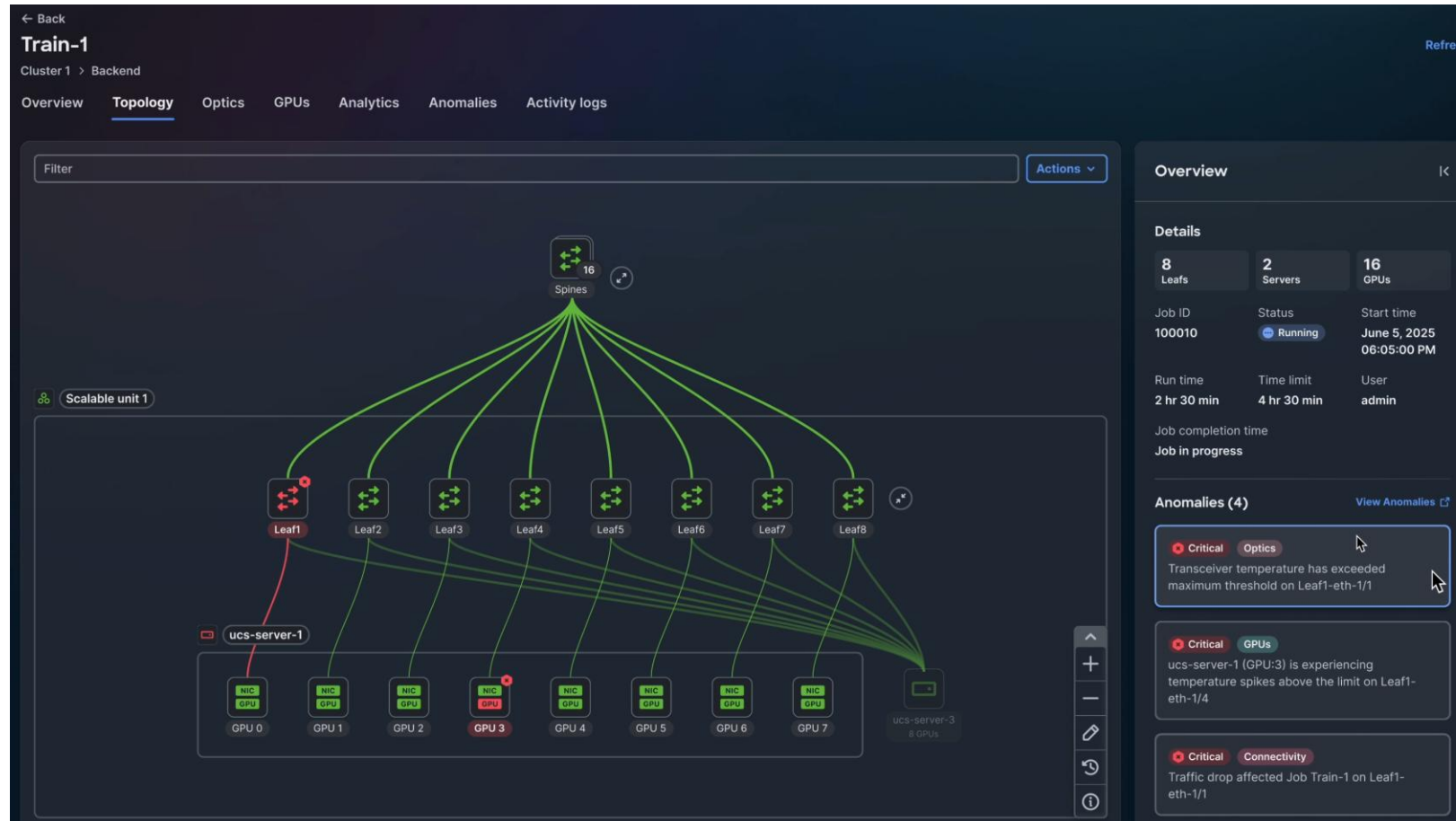


https://github.com/paregupt/nexus_traffic_monitor

Nexus Dashboard – Visibility into GPU Cluster and Jobs

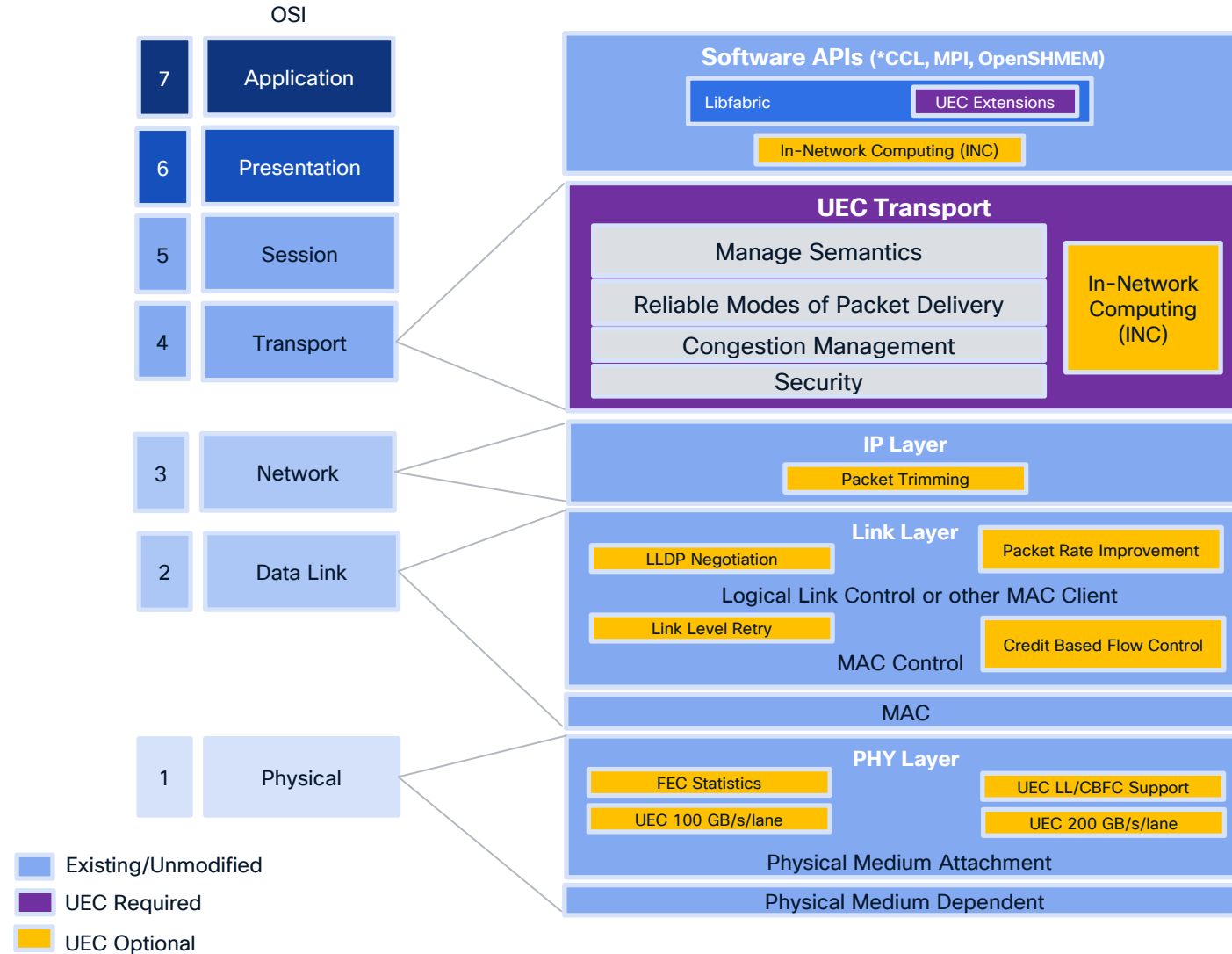
Coming soon – SLURM integration

Correlation of job health with GPU metrics, NIC metrics, and network anomalies



Ultra Ethernet Consortium (UEC) and Cisco

Focused on Inter-GPU backend networks



- UEC Mission – Deliver an Ethernet based open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI & HPC at scale
- Cisco is a founding and steering member of UEC
- UE from a switch perspective has no additional Layer-1 and Layer-2 requirements
- Cisco Silicon complies with UE fabric baseline requirements (PFC, ECN, Multiple traffic classes, etc.) for AI Ethernet network
- Cisco Silicon One is future-proof because of its programmable architecture



1

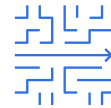


Extend NVIDIA's Spectrum-X architecture to include Cisco Silicon and Cisco Optics

Cisco Enterprise Reference Architecture (ERA) based on NVIDIA ERA, available since March 2025

Cisco Cloud Reference Architecture (CRA) based on NVIDIA Cloud Partner (NCP) reference architecture available since Oct 2025

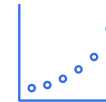
2



Build Cisco Nexus Switches powered by Cisco Software and NVIDIA Spectrum

Orderable since Nov 2025
Included with NVIDIA NCP RA

3



Deliver the highest performance Ethernet solutions available for AI workloads with tighter integration

Adaptive Routing, part of NVIDIA Spectrum-X, is available with NX-OS 10.6(1) onwards

Cisco AI Networking

1



Systems

Cisco Silicon, NVIDIA Spectrum Silicon
+
Cisco Switches (Nexus or 8000)
+
Cisco Transceivers
+
NX-OS or SONiC

2



Operating model

On-premises
Nexus Dashboard

or

Cloud-hosted
Nexus Hyperfabric

3



Reference Architecture

AI cluster validations and
benchmarking with leading GPU, NIC,
and storage vendors



Hyperscalers, Neoclouds, Sovereign clouds, and Enterprises

Any GPU, Any Storage, Any NIC – Consistent Cisco Networking Experience

Complete your session evaluations



Complete a minimum of 4 session surveys and the Overall Event Survey to claim a Cisco Live T-Shirt.



Earn up to 800 points by completing all surveys and climb the Cisco Live Challenge leaderboard.



Level up and earn exclusive prizes!



Complete your surveys in the Cisco Live Events app.

Continue your education



Visit the Cisco Stand for related demos



Book your one-on-one Meet the Expert meeting



Attend the interactive education with Capture the Flag, and Walk-in Labs



Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Thank you

CISCO Live !