

cisco Live !

Inside Cisco's Secure AI Factory: Practical Design Patterns and Insights from Real Deployments

Mehran Chowdhury, CCIE DC #52331
Sr. Customer Delivery Architect, Cisco

BRKXAR-2041

Cisco Webex App

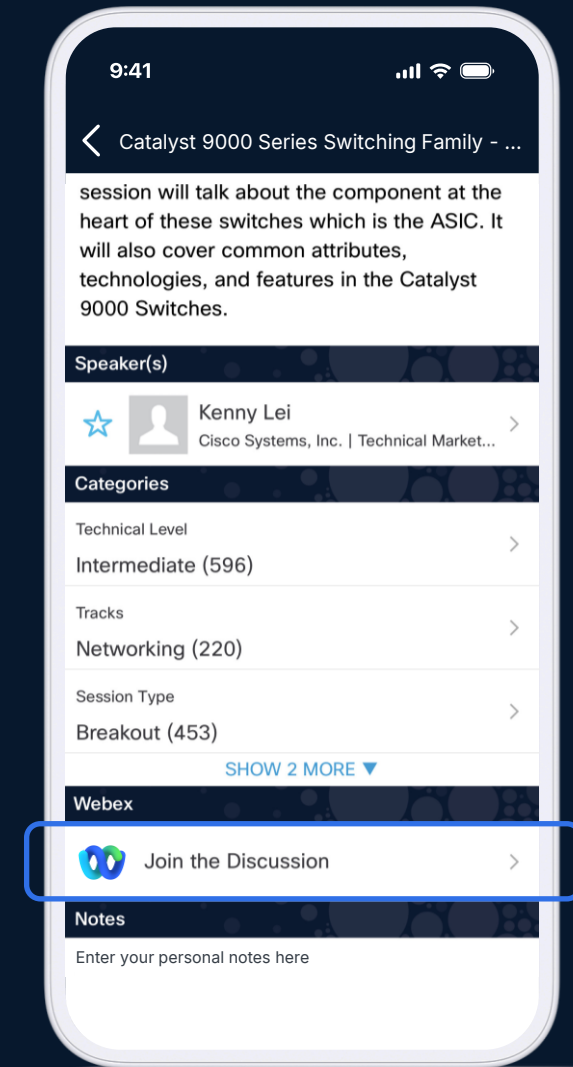
Questions?

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 June 13, 2026.

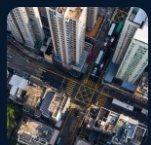


<https://ciscolive.ciscoevents.com/clamer26/BRKXAR-2041>

Agenda

- 01 AI Infrastructure Design Shift
- 02 Cisco Secure AI Factory Overview
- 03 Network Design and Considerations
- 04 Compute and Storage Design and Considerations
- 05 AI Defense
- 06 Observability & Day 2 Operations
- 07 Wrap-up and Q&A

Artificial Intelligence Outcomes Span Every Industry



Government

- Deliver Enhanced Citizen Services
- Data-Driven Policy Decisions and Creation
- Modernization & Streamline Operations
- Optimizing Infrastructure Management
- AI-Powered Traffic Design and Public Safety



Manufacturing

- Intelligent Quality Control
- Proactive Machine Maintenance
- Digital Twin Creation
- Supply-Chain Optimization and Tracking
- Optimizing Production Processes



Finance

- Predictive Trading Algorithms
- Fraud Detection and Prevention
- Personalized Financial Advice
- Investment Portfolio Optimization
- Virtual assistants, and seamless transaction experiences



Healthcare

- Medical Imaging Analysis
- Enhance Diagnosis and Treatment
- Patient management through predictive analytics
- Improved access to Healthcare with remote monitoring tools.
- Drug Research and Development



Retail

- Consumer Behavior Analytics
- Enhanced Customer Experiences
- Personalized Product Recommendations
- Dynamic Virtual Shopping Experience
- Demand Analysis and Prediction

Build the Model | Training

Optimize the Model | Fine-tuning & RAG

Use the Model | Inferencing

All organizations are defining their AI Services Stack

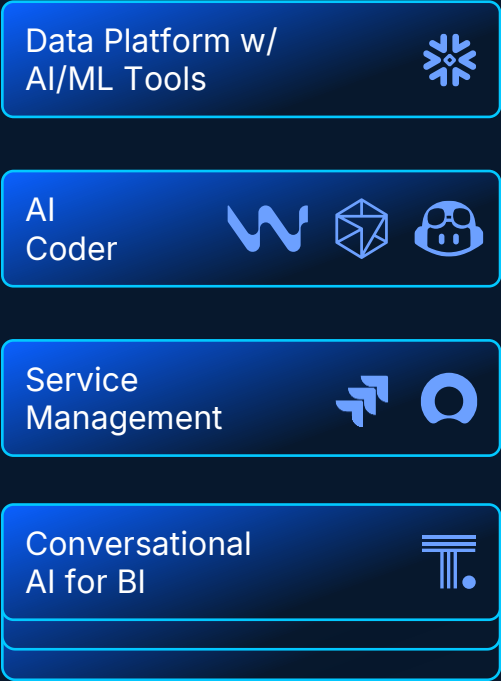
Agent Ecosystem



SaaS based AI Services



AI Platform Services





AI Infrastructure Design Shift

Shifting Traffic Patterns

Traditional Data Center



- Best-effort Ethernet; lossy, TCP-recoverable
- North-south dominant; modest, bursty east-west
- Oversubscribed fabric (with 2:1, 3:1 and 4:1 oversubscription at the leaf)
- Millisecond latency targets; tail latency tolerated
- Smaller, independent flows with variable demand and diverse traffic.
- Traffic patterns stay predictable and moderate

AI Data Center



- Lossless RoCEv2 with PFC and ECN (DCQCN) tuning
- East-west dominant; synchronized collectives generate massive, line-rate GPU-to-GPU and node-to-node traffic
- Non-blocking fabric, no oversubscription (strict 1:1 end-to-end)
- Microsecond latency targets; tail latency stalls collectives and idles GPUs
- Larger, tightly coupled flows synchronized across GPUs, where one slow path can impact the entire job
- Traffic patterns are bursty, synchronized, and bandwidth-intensive



Compute and Storage

Traditional Data Center



- x86 CPU-centric compute optimized for mixed enterprise workloads
- Workload Type: Mixed transactional, batch, database, mainframe and service-oriented applications
- SAN and NAS architectures using mixed SSD and HDD tiers, typically optimized for IOPS, latency, and data services.
- Heavy virtualization - VMs slice resources for multi-tenancy and consolidation
- 10/25/40/100G NIC connectivity sized for mixed application and storage traffic.

AI Data Center



- Accelerated dense compute with CPU, GPU, and DPU resources optimized for AI workloads
- Workload Type: Massively parallel, tightly coupled AI workloads using distributed GPU compute, collective communication, and synchronization
- All-flash/NVMe storage platforms optimized for parallel access, high throughput, metadata scale, and checkpoint performance.
- Bare metal Linux + Kubernetes/OpenShift
- Multi-NIC 100/200/400/800G connectivity optimized for GPU fabric, storage, and high-throughput AI data movement



Power and Cooling

Traditional Data Center



- Standard Density: 5–20 kW per rack, fitting within standard grid capacity.
- Air cooling is often sufficient for standard x86 compute environments
- Steady State: Power draw is relatively predictable based on uncorrelated user traffic cycles.
- Power, cabling, airflow, and cooling are often handled through standard capacity planning.
- ToR-based access is common for standard server racks

AI Data Center



- Extreme Density: Racks demand 40–150+ kW today, with industry projections of MW-class racks within 5–7 years
- High-density GPU racks may require rear-door heat exchangers, direct liquid cooling, or hybrid cooling designs
- High Volatility: Synchronous training causes massive load swings (10% to 100% in milliseconds)
- MoR/EoR switching helps preserve GPU rack power and cooling capacity while improving cabling, airflow, and serviceability.



Observability and Day-2 Ops

Traditional Data Center



- Monitoring : servers, network, storage, and applications in separate operational domains
- Primary Metrics : CPU, memory, disk, interface utilization, latency, errors, application health
- Operational Focus : Keep applications, VMs, services, and infrastructure available
- Impact : Issues often affect an application, service, VM, or infrastructure tier
- Troubleshooting : Siloed troubleshooting by server, network, storage, or application teams

AI Data Center



- Monitoring : Requires correlated visibility across GPU, network, storage, Kubernetes/OpenShift, scheduler, and application layers
- Primary Metrics : CPU utilization, GPU utilization, GPU memory, job status, congestion, latency, PFC/ECN behavior, storage throughput, and pod/node health
- Operational Focus : Keep GPUs utilized, jobs progressing, and AI pipelines performing end-to-end
- Impact : One slow GPU, NIC, network path, storage path, or job dependency can impact the full distributed AI workload
- Troubleshooting : Trace AI performance issues across the application, scheduler, Kubernetes/OpenShift, GPU, storage, and network layers to quickly identify bottlenecks and accelerate troubleshooting.

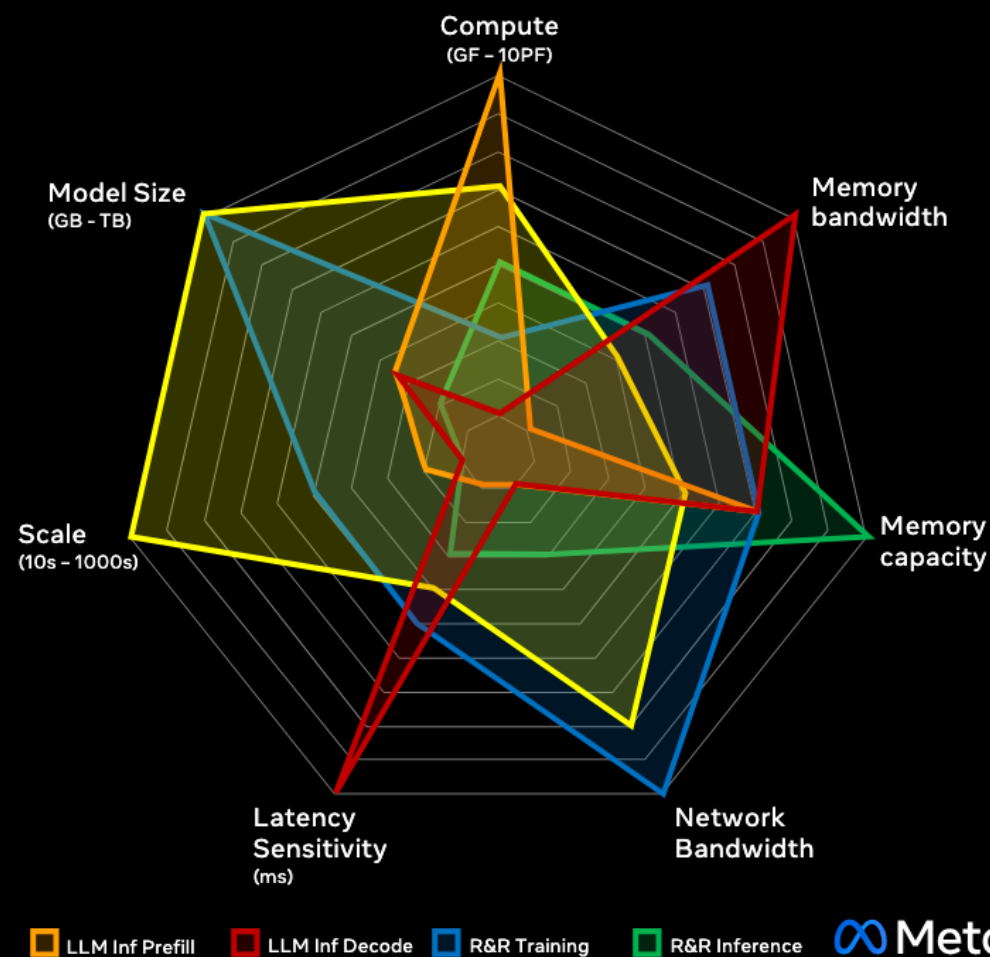


**Why is the network so
important?**

Diverse AI Requirements

Diversity of AI system requirements

- Difficult to serve all classes of models with a single system design point
- New models & parallelism techniques put unexpected pressures on AI systems
- The next frontier of innovation is in software/hardware co-design

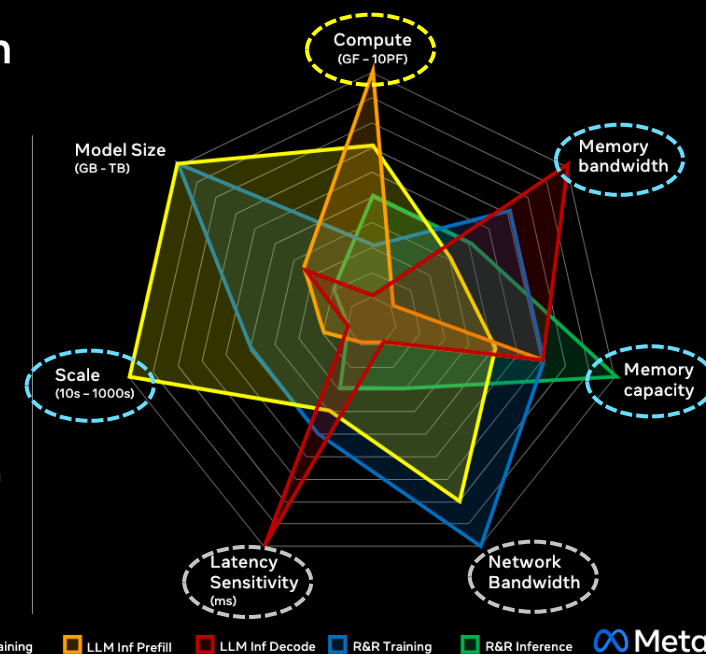


Source: [Meta AI Hardware Summit](#)

Training vs Inference

Diversity of AI system requirements

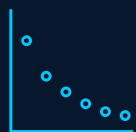
- Difficult to serve all classes of models with a single system design point
- New models & parallelism techniques put unexpected pressures on AI systems
- The next frontier of innovation is in software/hardware co-design



Meta AI Hardware Summit

	Distributed Training	Production Inference
Node to Node Bandwidth	High	Low
Key Metric	Training time of a model or Job Completion Time (JCT)	Network Latency, Time to First Token (TTFT)
Operational Mode	Model training is offline	Usually online, requires real time response
Infrastructure requirement	Large network with many GPU/CPU hosts	Smaller network with mid size of CPU/GPU hosts

AI/ML Demands More From the Network



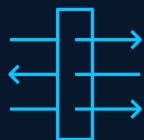
Low latency



**High, lossless
throughput**



High bandwidth



**Congestion
management**

*What it
means?*



1:1 over-subscription

RoCEv2

ECN

PFC

DCQCN (PFC+ECN)

WRED

**Dynamic Load balancing
(DLB)**

...

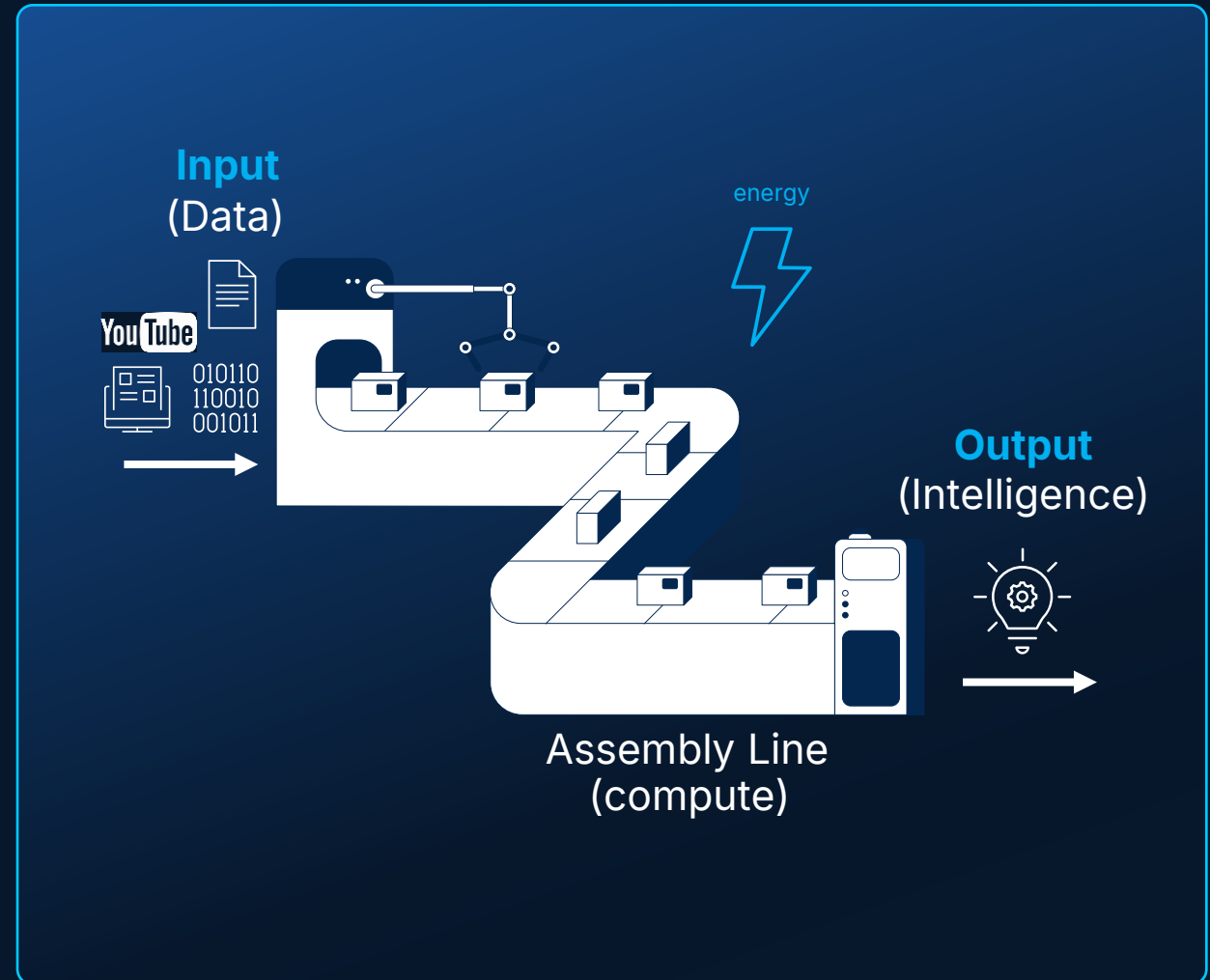


Cisco Secure AI Factory Overview

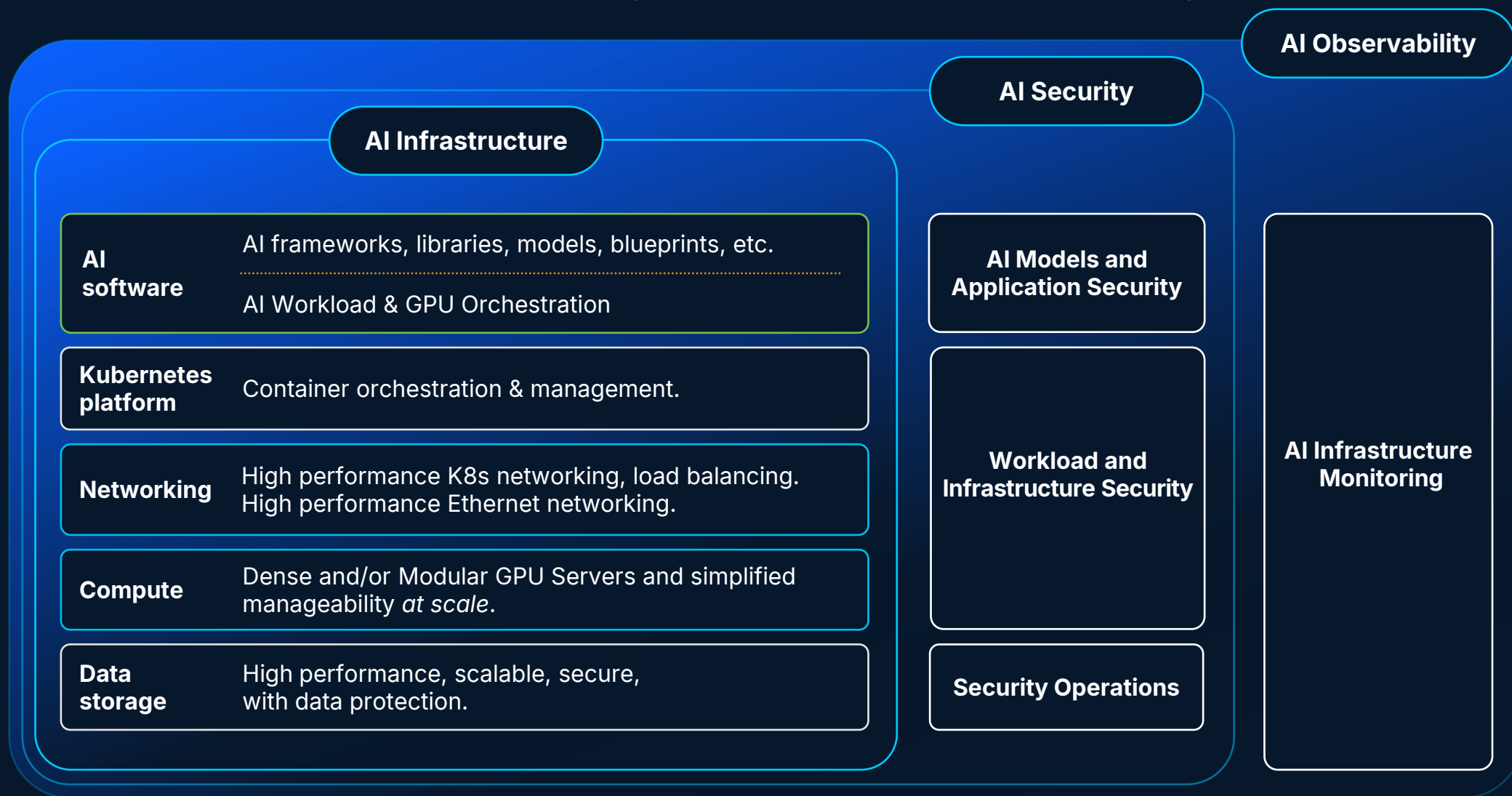
What is an AI Factory?

The processing plant for tokens

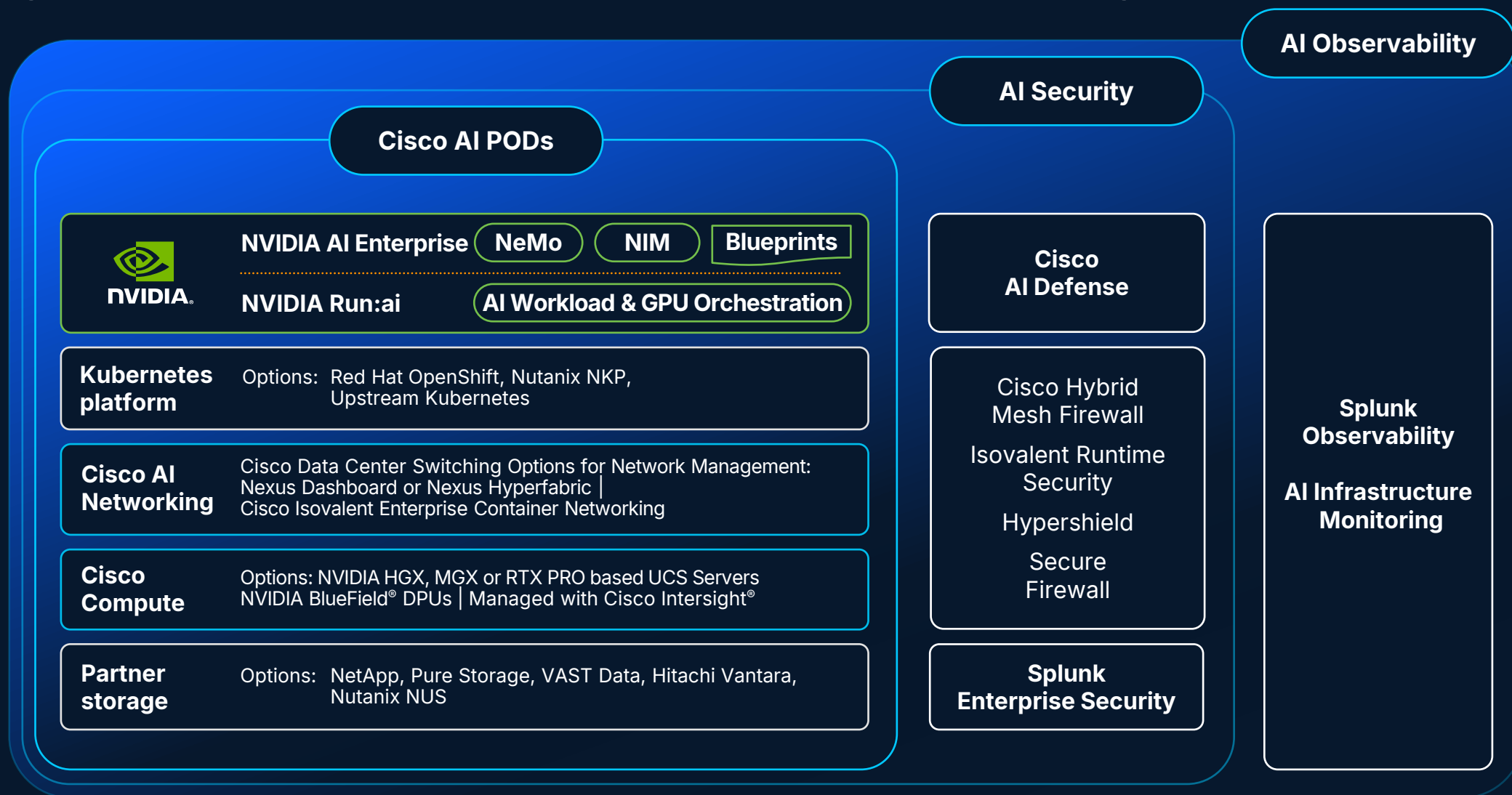
Organizations everywhere are thinking about how to **generate** tokens as quickly, safely, securely, and cost effectively as possible.



Cisco Secure AI Factory with NVIDIA - Key capabilities



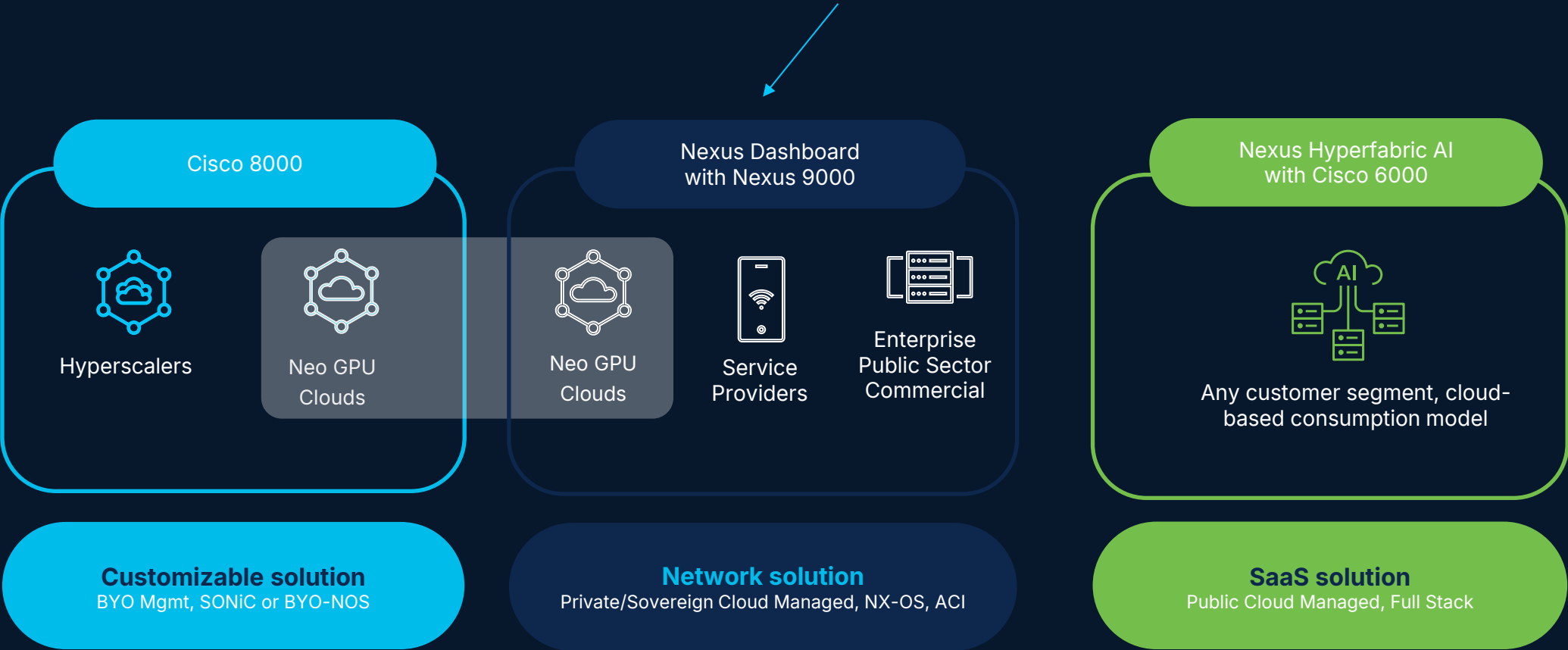
Key products in Cisco Secure AI Factory with NVIDIA





Network Design and Considerations

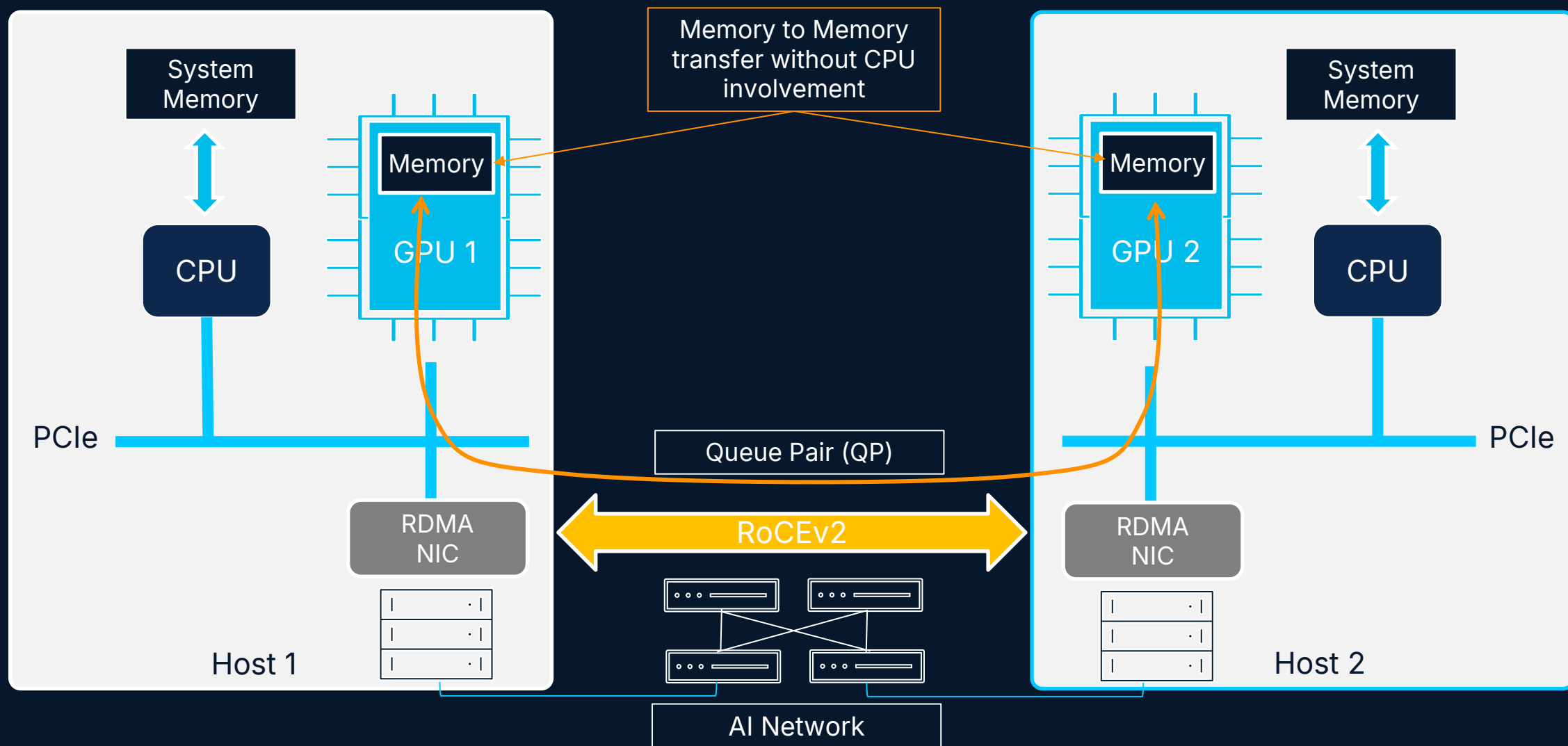
Cisco Networking options for AI/ML Infrastructure



What is Remote Direct Memory Access (RDMA)?

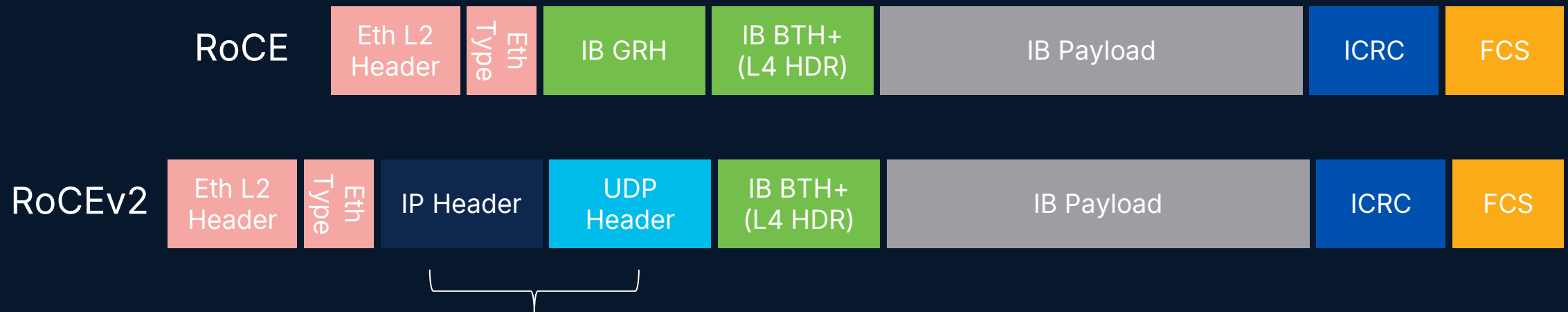
- Allows GPU direct memory to memory access thru RDMA NICs using RoCE stack (RoCE & RoCEv2)
- Requires lossless, non-blocking network with PFC and ECN
- High throughput low latency data transfers without CPU cycles
- Supports
 - **Zero Copy:** Apps transfer data between memory on two hosts bypassing CPU, Kernel network stack (TCP/IP) and System Cache
 - **Kernel Bypass:** Data transfer directly from user space without the need to perform context switches over the kernel

RDMA over Converged Ethernet Version 2 (RoCEv2)

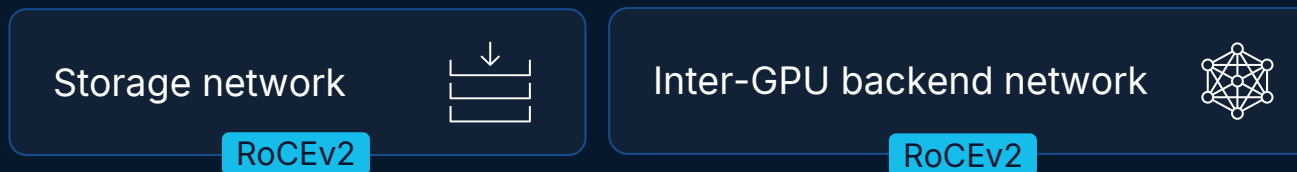


RoCEv2 Basics

- Extension of RoCE protocol that involves a simple modification of the RoCE packet format.
- Carry IP header and UDP header that serves as a stateless encapsulation layer for RDMA transport over IP, on UDP port 4791.
- Key benefits: IP routing, QoS and lossless networking based on IP DSCP where L2 header does not exist or not carried; congestion notification to end devices using IP ECN, and ability to load-share based on UDP header.
- "RoCEv2 header" – No such thing. This is InfiniBand payload in an Ethernet Frame with an IP/UDP Header.



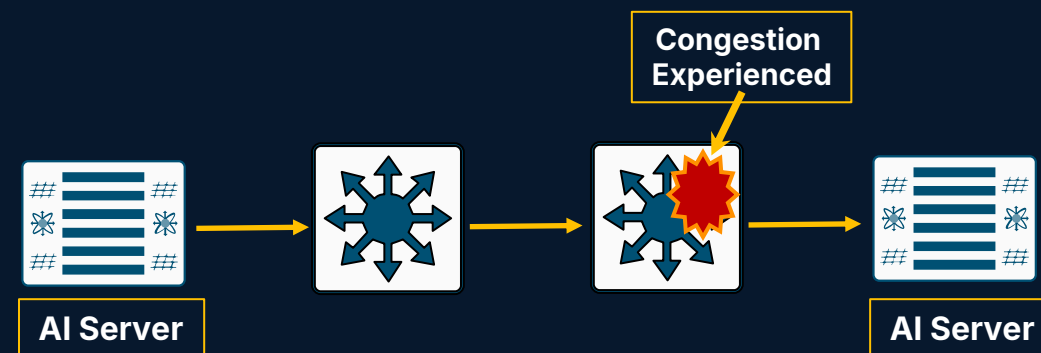
RoCEv2 – Key points



- RoCEv2 uses UDP/IP transport – All common IP/Ethernet networking knowledge applies
- Routing protocol (OSPF, IS-IS, eBGP) eBGP is recommended due to features (e.g. ECMP/WCMP)
- RoCEv2 requires lossless networks
 - Priority-based Flow Control (PFC) converts best-effort (lossy) Ethernet to lossless Ethernet
- Cisco NX-OS uses Modular QoS CLI (MQC) for PFC configuration (lossless Ethernet)
- RoCEv2 configuration for storage networks and backend networks is similar
 - Backend (storage and gpu) networks are dedicated, whereas front-end storage networks may be converged with other frontend networks
- Use Nexus Dashboard for fabric configuration – Strongly Recommended, but not mandatory

Explicit Congestion Notification (ECN)

- IP Explicit Congestion Notification (ECN) is used for congestion notification.
- ECN enables end-to-end congestion notification between two endpoints on IP network
- ECN uses 2 LSB of Type of Service (ToS) field in IP header

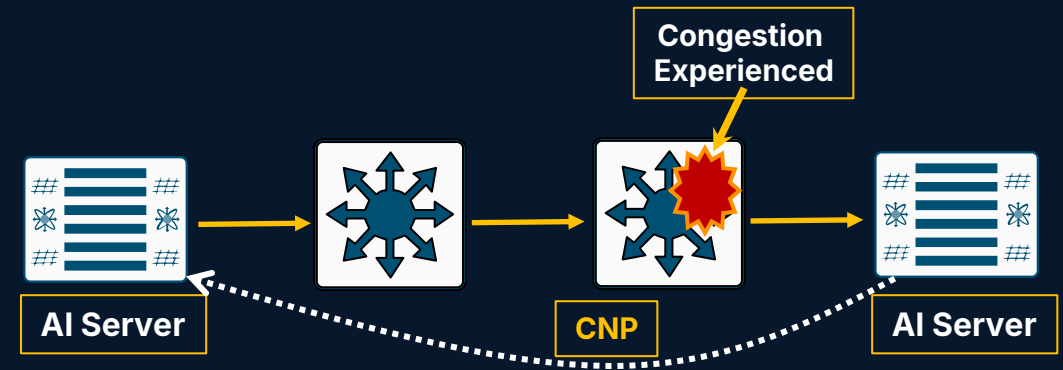


ECT	CE	ECN Behavior
0	0	Non ECN Capable
0	1	ECN Capable Transport
1	0	ECN Capable Transport
1	1	Congestion Encountered

ECT – ECN Capable Transport
CE – Congestion Experienced

Explicit Congestion Notification (ECN)

- IP Explicit Congestion Notification (ECN) is used for congestion notification.
- ECN enables end-to-end congestion notification between two endpoints on IP network
- ECN uses 2 LSB of Type of Service (ToS) field in IP header
- In case of congestion, ECN gets transmitting device to reduce transmission rate using Congestion Notification Packet (CNP) without pausing traffic.

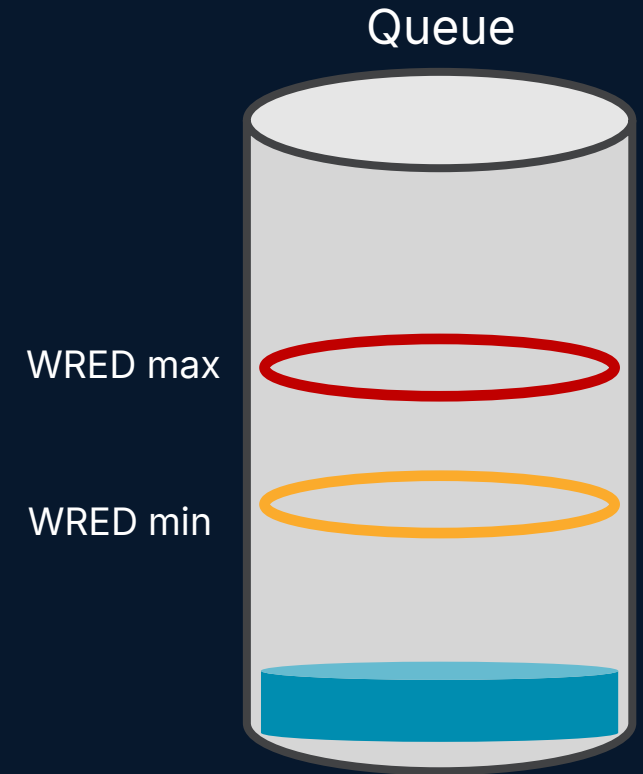


ECT	CE	ECN Behavior
0	0	Non ECN Capable
0	1	ECN Capable Transport
1	0	ECN Capable Transport
1	1	Congestion Encountered

ECT – ECN Capable Transport
CE – Congestion Experienced

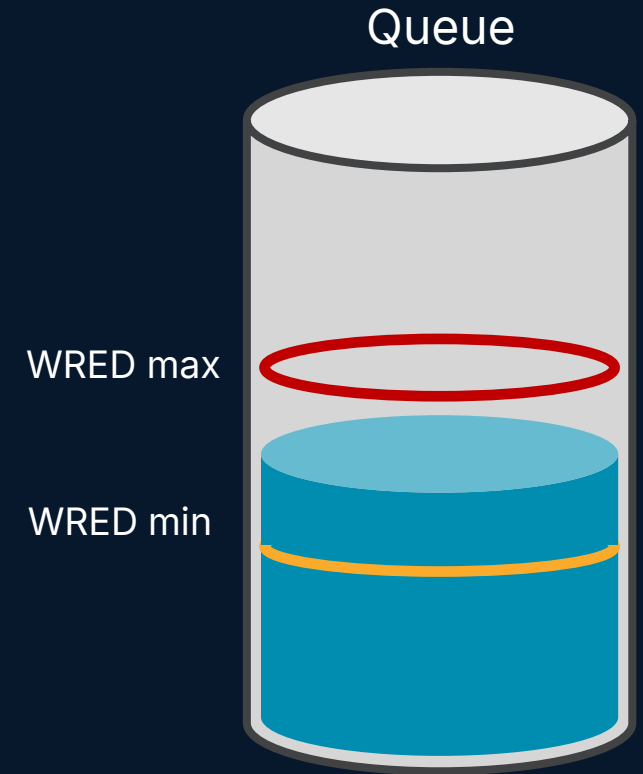
How does WRED ECN work?

- WRED (Weighted Random Early Detection) is used to signal the severity of congestion
- ECN is not marked when buffer usage is below WRED minimum threshold



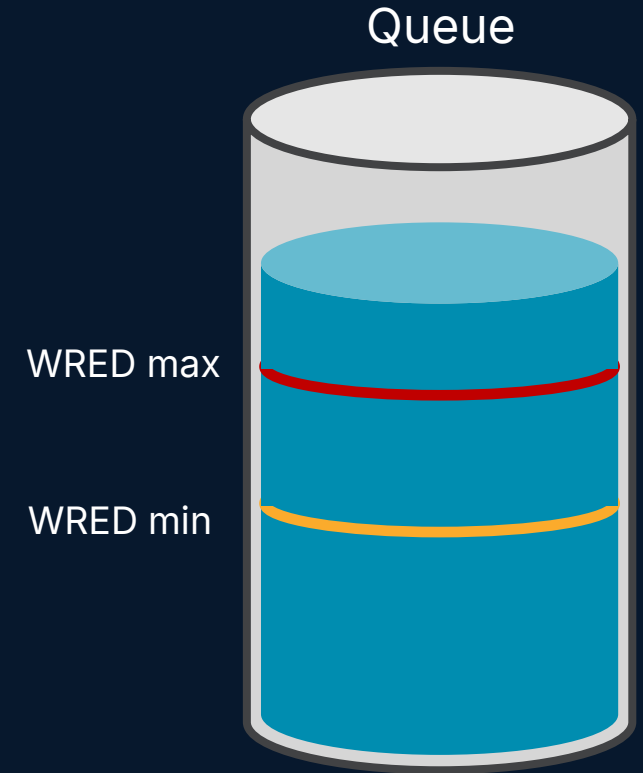
How does WRED ECN work?

- WRED (Weighted Random Early Detection) is used to signal the severity of congestion
- ECN is not marked when buffer usage is below WRED minimum threshold
- When buffer usage is minimal threshold, Congestion Encountered will be marked on **N** number of randomly selected packets (probability parameter)



How does WRED ECN work?

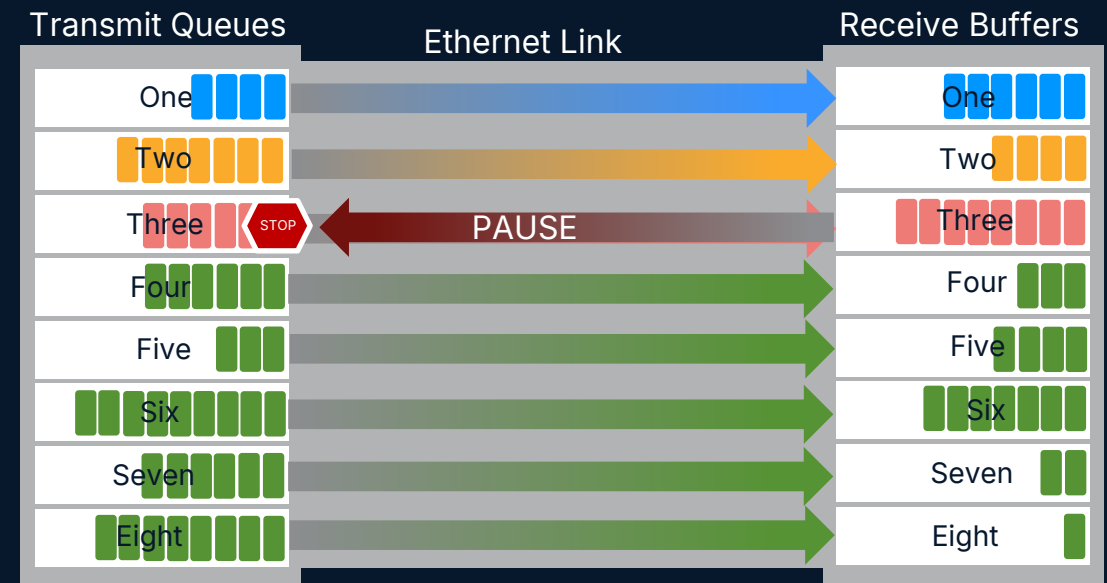
- WRED (Weighted Random Early Detection) is used to signal the severity of congestion
- ECN is not marked when buffer usage is below WRED minimum threshold
- When buffer usage is minimal threshold, Congestion Encountered will be marked on **N** number of randomly selected packets (probability parameter)
- After buffer usage crosses MAX threshold, **every** ECN capable packet will be marked with **Congestion Encountered**



Priority Flow Control (PFC)

Flow Control Mechanism – 802.1Qbb

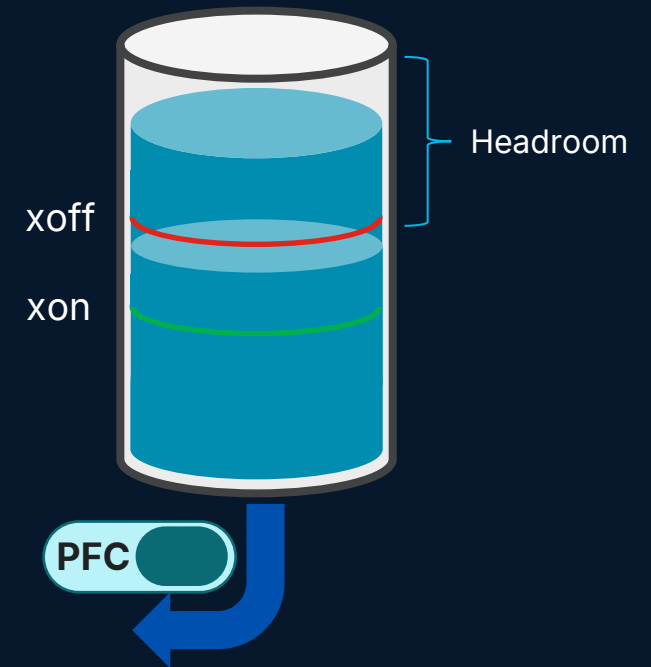
- Enables “lossless Ethernet” using PAUSE based Class of Service defined in 802.1p
- Enables Flow Control on Per-Priority basis
- Also called Per-Priority-Pause
- Have lossless and lossy priorities at the same time on the same wire
- When link gets congested, CoS assigned to “no-drop” class will be PAUSED
- Traffic assigned to other CoS will continue to transmit and rely on ULP for retransmission
- Different CoS of traffic operate independent of each other on the same link



Priority Flow Control (PFC) – When are they sent?

Flow Control Mechanism – 802.1Qbb

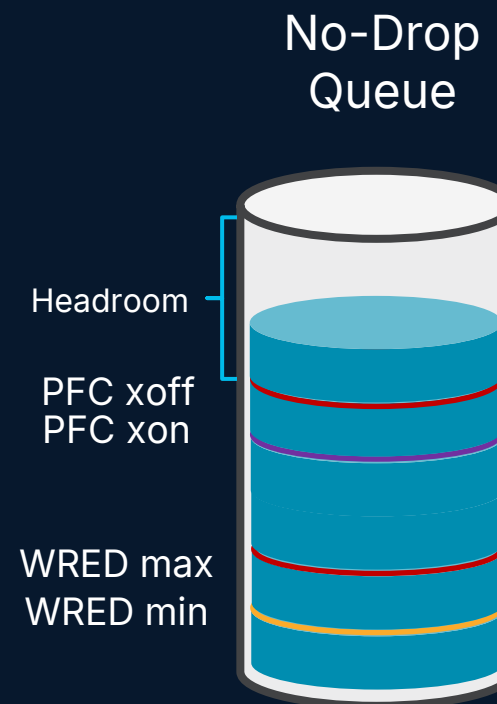
- PFC sets thresholds in no-drop queue *xon* and *xoff*
- Headroom is present to accommodate "in flight" packets
- Under congestion, traffic is buffered in non-drop queue
- PFC frames are sent toward sender after queue utilization exceeds *xoff* threshold
- While draining the queue, and when utilization drops below *xon* threshold, system will stop sending PFC frames



RoCEv2: PFC and ECN working to together

How does it work?

- WRED threshold are set low in no-drop queue
 - Signal early for congestion, give enough time for end points to react
- PFC threshold are set higher than ECN
 - In case oversubscription buffers can be filled quickly without giving time to ECN to react
 - PFC will react and mitigate congestion
 - PFC+ECN is Data Center Quantized Congestion Notification (DCQCN)

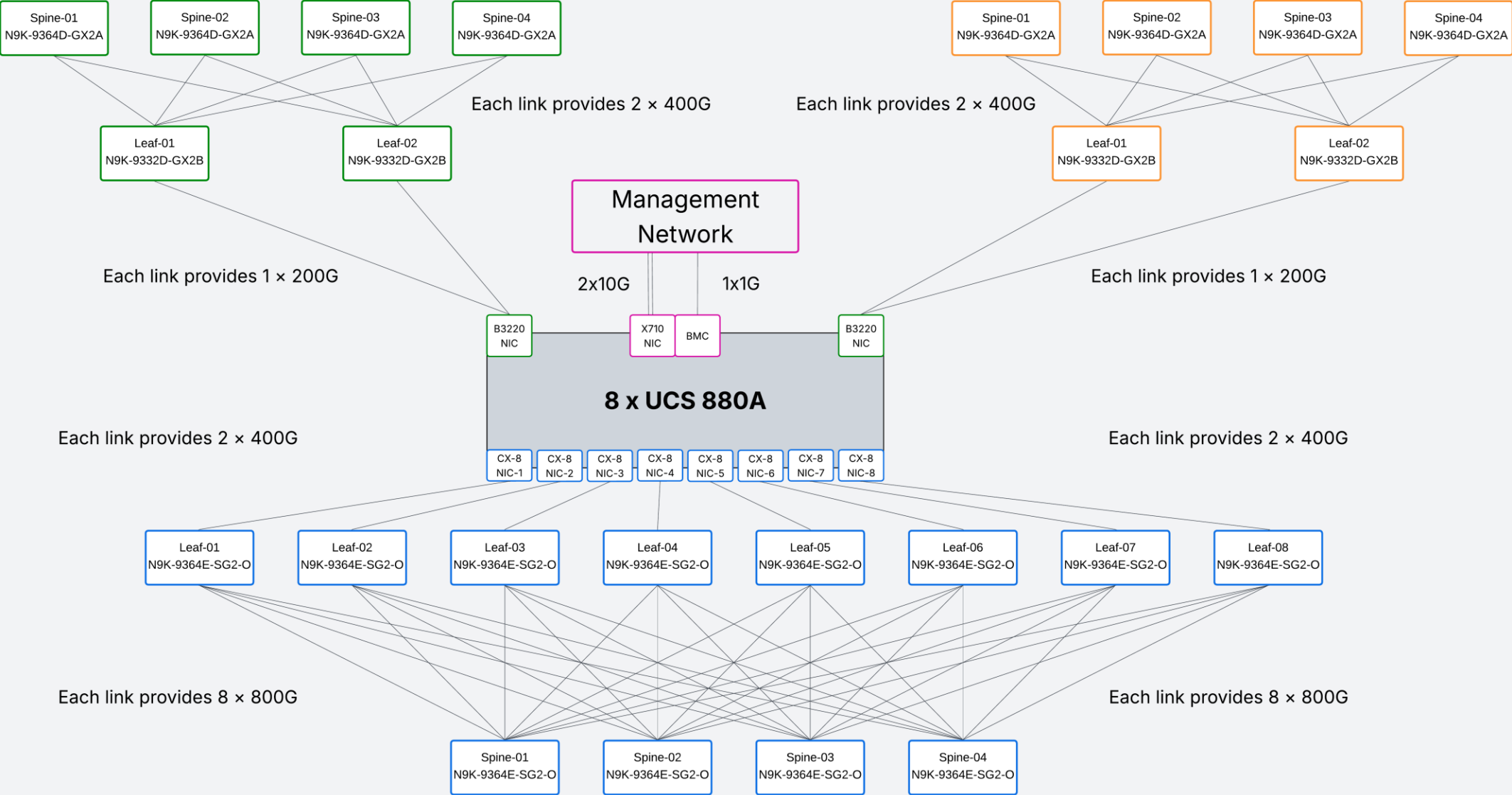


Reference Architecture

Based on a customer engagement!

Frontend Fabric

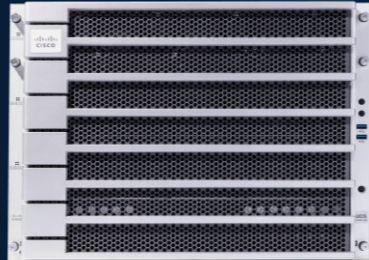
Frontend Storage Fabric



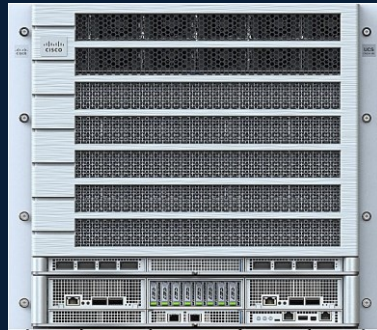
Backend GPU Fabric

Cisco AI/ML – Dense GPU Compute + Networking

AI Servers



UCS C885A M8
8RU Dense GPU Server



UCS C880A M8
10RU Dense GPU Server



UCS C845A M8
4RU MGX Server



AI Networks



Cisco N9364E-SG2
64 x 800G 2RU switch
Cisco Silicon



Cisco Nexus 9364D-GX2A
64 x 400G 2RU switch
Cisco Silicon



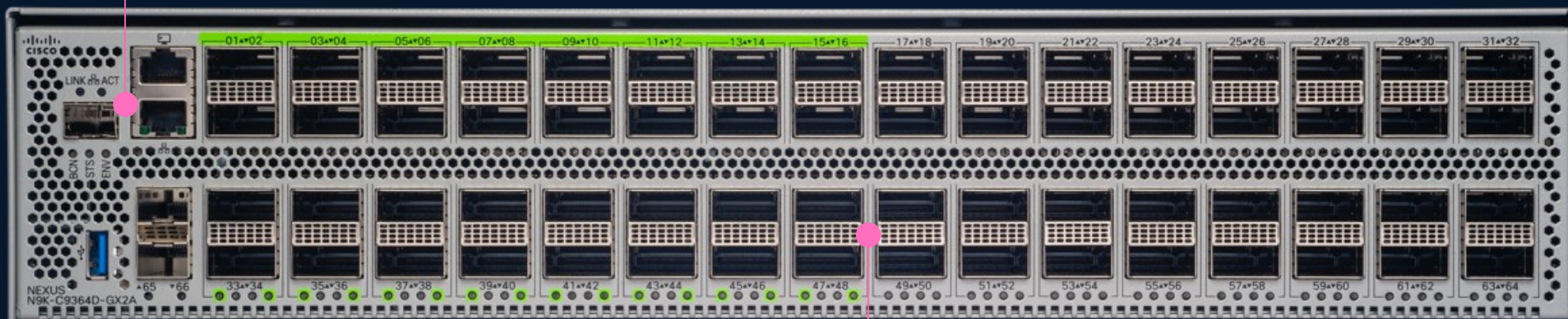
Cisco Nexus 9332D-GX2B
32 x 400G 1RU switch
Cisco Silicon



Nexus Dashboard

Nexus 9364D-GX2A

Console
Management
USB



2 RU

64p QSFP-DD 400G
2p 1/10G

- 64p QSFP-DD 400G + 2p 1/10G
- ACI and NXOS
- Support for 16 Port MACsec & CloudSec
- Enhanced Telemetry – FT, FTE, SSX, INT-XD
- Port-side Intake AC only

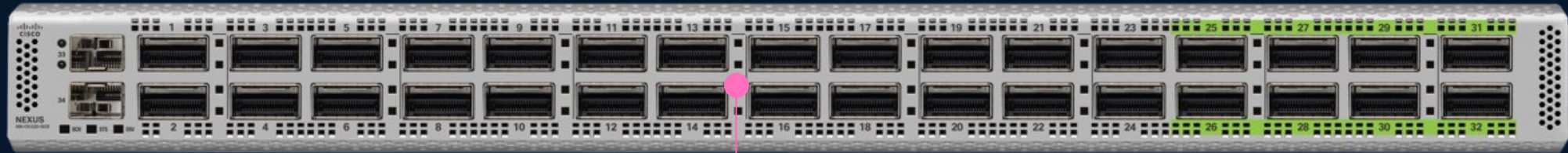
- Redundant PSU and fan support
- Breakout Options:
 - Up to 128x 200GE
 - Up to 256 Ports of 10G, 25G, 50G NRZ / 50G PAM4, 100G

Nexus 9364D-GX2A – Port to Slice Mappings

Front Panel Port	ASIC Slice Number	MACsec Ports
Ethernet 1/1-8	Slice 5	N/A
Ethernet 1/9-16	Slice 1	N/A
Ethernet 1/17-24	Slice 0	N/A
Ethernet 1/25-32	Slice 4	N/A
Ethernet 1/33-40	Slice 7	N/A
Ethernet 1/41-48	Slice 3	N/A
Ethernet 1/49-56	Slice 6	Capable
Ethernet 1/57-64	Slice 2	Capable

Slice Pair 

Nexus 9332D-GX2B



32p QSFP-DD 400G
2p 1/10G

1 RU

- 32p QSFP-DD 400G + 2p 1/10G
- ACI and NXOS spine
- Support for 8p MACsec & CloudSec
- Enhanced Telemetry – FT, FTE, SSX, INT-XD
- Port-side Intake

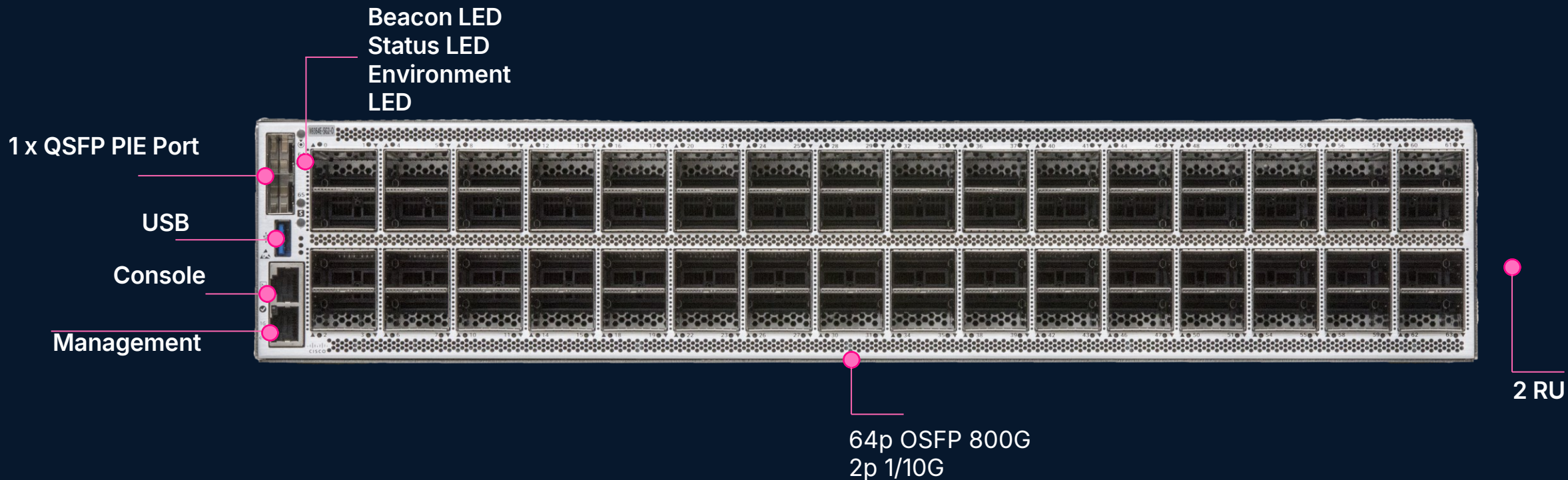
- Redundant PSU and fan support
- Cost optimized design
- Breakout Options
 - Up to 64 Ports of 200G
 - Up to 128 Ports of 10G, 25G, 50G NRZ / 50G PAM4, 100G

Nexus 9332-GX2B – Port to Slice Mappings

Front Panel Port	ASIC Slice Number	MACsec Ports
Ethernet 1/1-8	Slice 3	N/A
Ethernet 1/9-16	Slice 2	N/A
Ethernet 1/17-24	Slice 0	N/A
Ethernet 1/25-32	Slice 1	Capable

Slice Pair 

Nexus 9364E-SG2-O

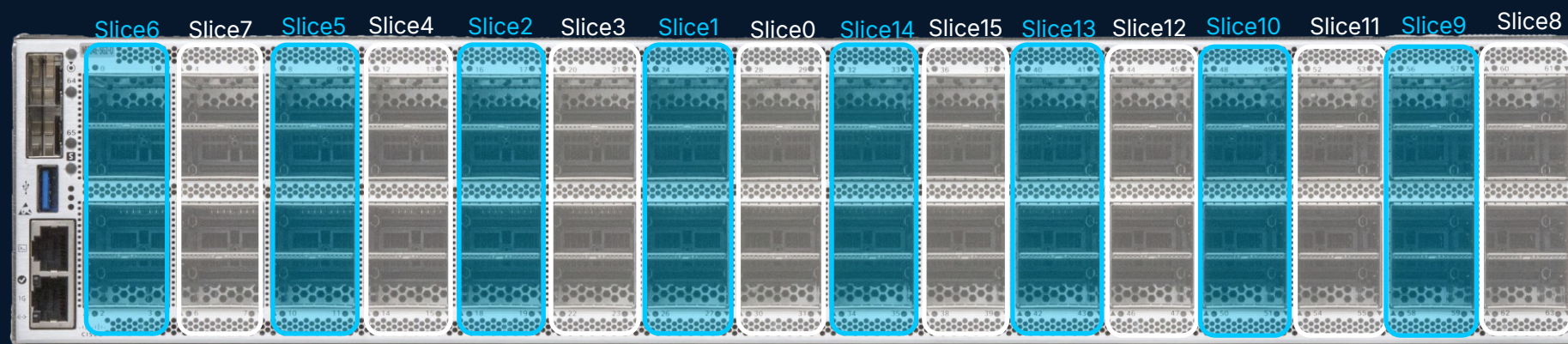


- 64x OSFP 800G ports
- All ports connected to ASIC (G200)
- Single Silicon One G200 ASIC operating @1.25 Ghz
- Line rate at packet sizes greater than 275 bytes
- Cooling/Power delivery for 17W OSFP Modules
- PTP 1588 Class C support
- Supports 2x400, and 8x100 breakout

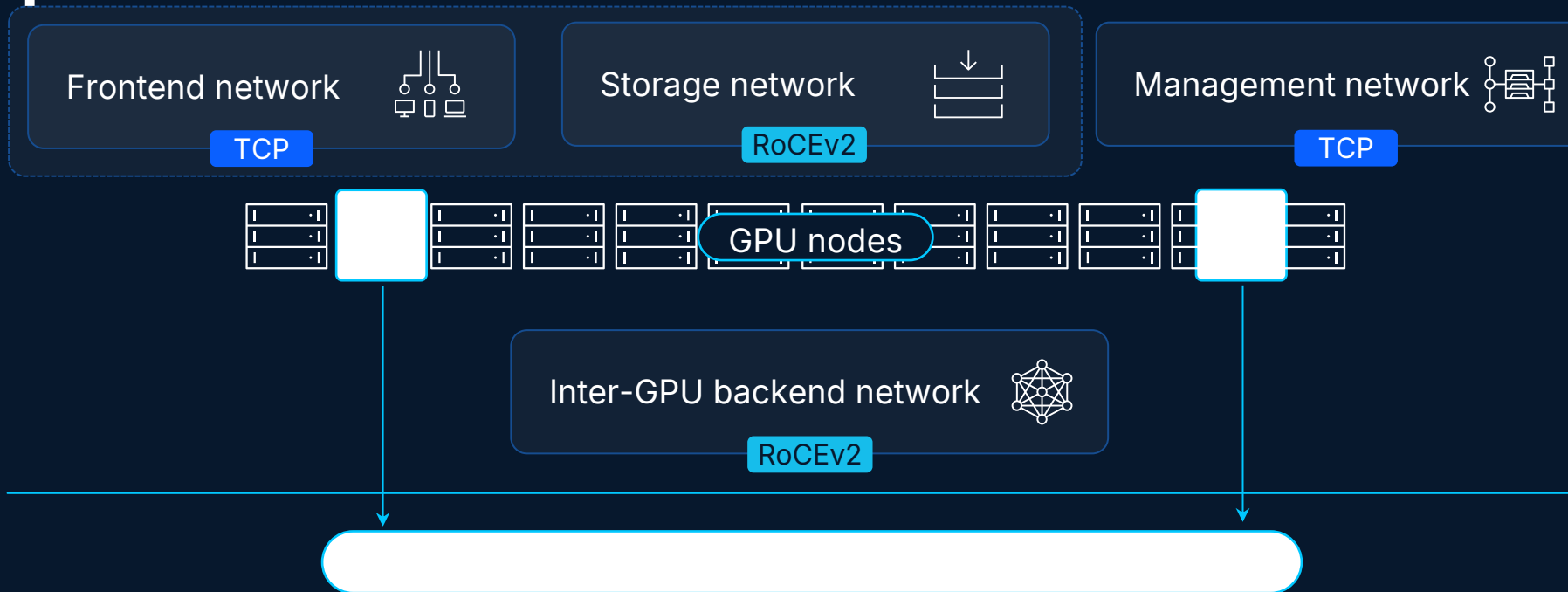
Nexus 9364E-SG2-O – Port to Slice Mappings

Front Panel Port	G200 Slice/IFG
1-4	6
5-8	7
9-12	5
13-16	4
17-20	2
21-24	3
25-28	1
29-32	0

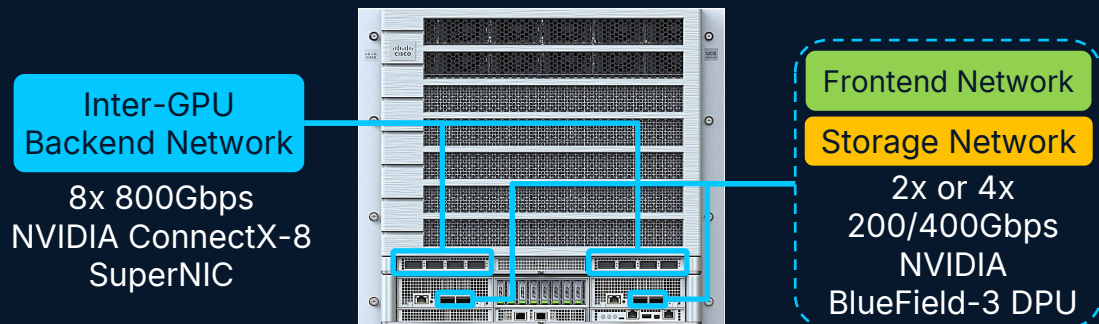
Front Panel Port	G200 Slice/IFG
33-36	14
37-40	15
41-44	13
45-48	12
49-52	10
53-56	11
57-60	9
61-64	8



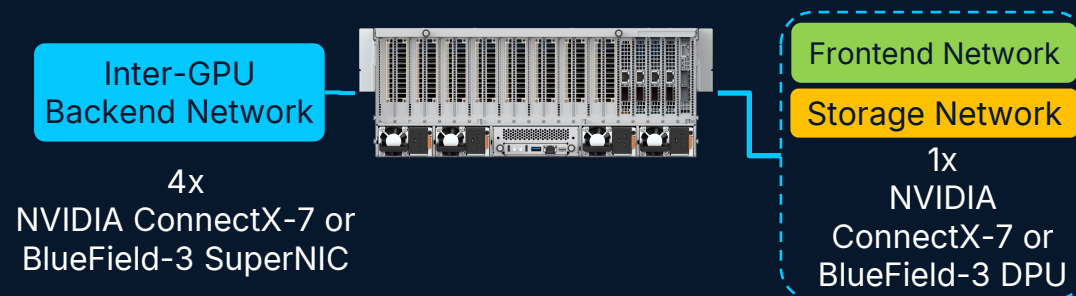
Multiple Networks for AI/ML Infrastructure



Cisco UCS 880A (8x NVIDIA B300 GPUs)

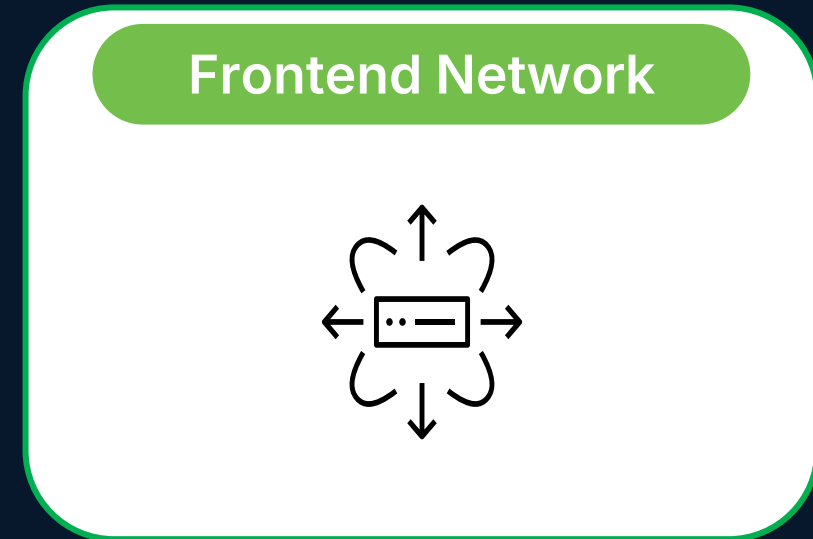


Cisco UCS 845A (8x NVIDIA H200 GPUs)



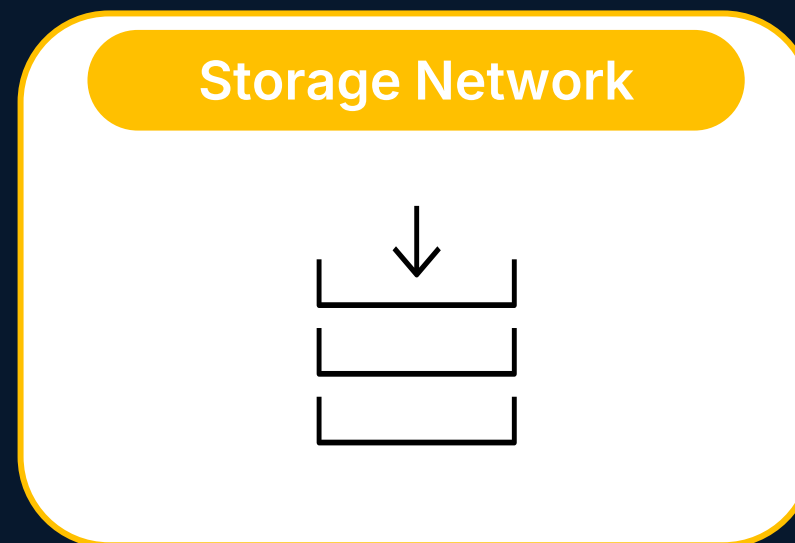
Frontend Network

- **North-South Connectivity:** the "front door" to the AI Factory; links the GPU cluster to the enterprise data center, campus, WAN, and internet
- **Data Ingestion:** moves training and fine-tuning datasets into the cluster from enterprise data lakes, repositories, and external sources
- **Inference & Application Access:** carries client and API request/response traffic to and from deployed models
- **Management & Orchestration:** control-plane traffic for Kubernetes/Scheduling, node provisioning, job scheduling, and CI/CD pipelines
- **Security boundary:** north-south segmentation and policy enforcement at the edge of the AI Factory
- Build or repurpose existing infrastructure for AI workloads.
Ex. Existing ACI or VXLAN Fabric



Storage Network

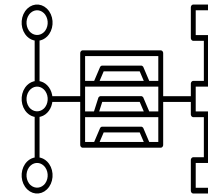
- **Dedicated storage fabric:** a standalone, non-blocking fabric dedicated entirely to GPU-to-storage data movement, providing low latency, high throughput access to training and inference data sets
- **GPUDirect Storage (GDS):** Direct DMA transfer of data from storage to GPU memory, bypassing CPU memory copies and reducing latency to keep GPUs continuously fed during training and inference.
- **RoCEv2 transport:** RDMA over Converged Ethernet for low-latency, high-throughput, CPU-bypass access to the scale-out / parallel storage
- **Sustained GPU Utilization:** keeps GPUs fed at line rate, eliminating data-starvation that stalls training throughput and job completion times
- **Scalable Performance:** scales storage bandwidth and capacity independently as datasets and model sizes grow, with no fabric re-architecture
- **Certified storage partners:** Hitachi Vantara, NetApp, Nutanix NUS, Pure Storage, and VAST Data



Management Network

- **Infrastructure Control Plane:** out-of-band (OOB) connectivity to every manageable element including switches, servers, GPUs, CIMC, BMC etc
- **Operational Reliability:** consistent, always-available reachability for monitoring, firmware/OS updates, automation etc
- **Segmentation & Security:** Segmentation & Security: Keeps management traffic isolated from workload networks, controls who can access the management plane, and limits lateral movement if a system is compromised.
- **Observability Backhaul:** exports metrics, logs, and traces (DCGM, OTel, Nexus Dashboard etc) out to monitoring platforms

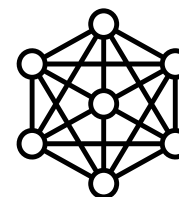
Management Network



Inter-GPU Backend Network

- **Dedicated GPU backend fabric:** Provides the high-speed backend network for GPU-to-GPU communication across servers
- **Non-blocking:** 1:1, zero-oversubscription leaf-to-spine, so every GPU can transmit at full line rate simultaneously during bursts that saturate the fabric
- **RDMA over Converged Ethernet Version 2 (RoCEv2):** GPUDirect RDMA moves data directly between GPU memory across nodes, bypassing the CPU and running over a Routed (L3) or VXLAN EVPN spine-leaf fabric
- **High-bandwidth low latency switching:** Uses 400G/800G-capable Cisco Nexus switching to keep GPU NICs fed at line rate and support scale-out AI/ML fabric requirements
- **Lossless Ethernet transport:** Uses Priority Flow Control (PFC) to provide pause-based no-drop behavior for RDMA traffic, with ECN/WRED (DCQCN) managing congestion before buffers are exhausted during bursty AI events
- **Scalable Clos topology:** Supports modular, pay-as-you-grow cluster expansion by adding leaf and spine capacity without re-architecting the backend fabric. Can be configured as eBGP Routed or VXLAN EVPN.

Backend Network



Rail-Optimized Design

1:1 oversubscription

Inter-GPU backend network



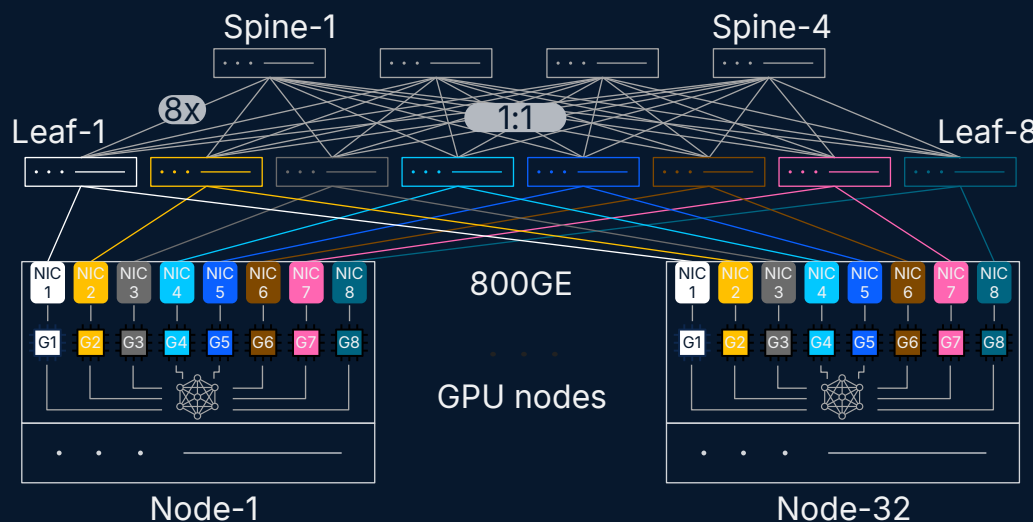
RoCEv2

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

Scalable Unit 1 256 GPUs

8 leaf and 4 spines switches with 64 x 800GE interfaces

Rail-Optimized Design

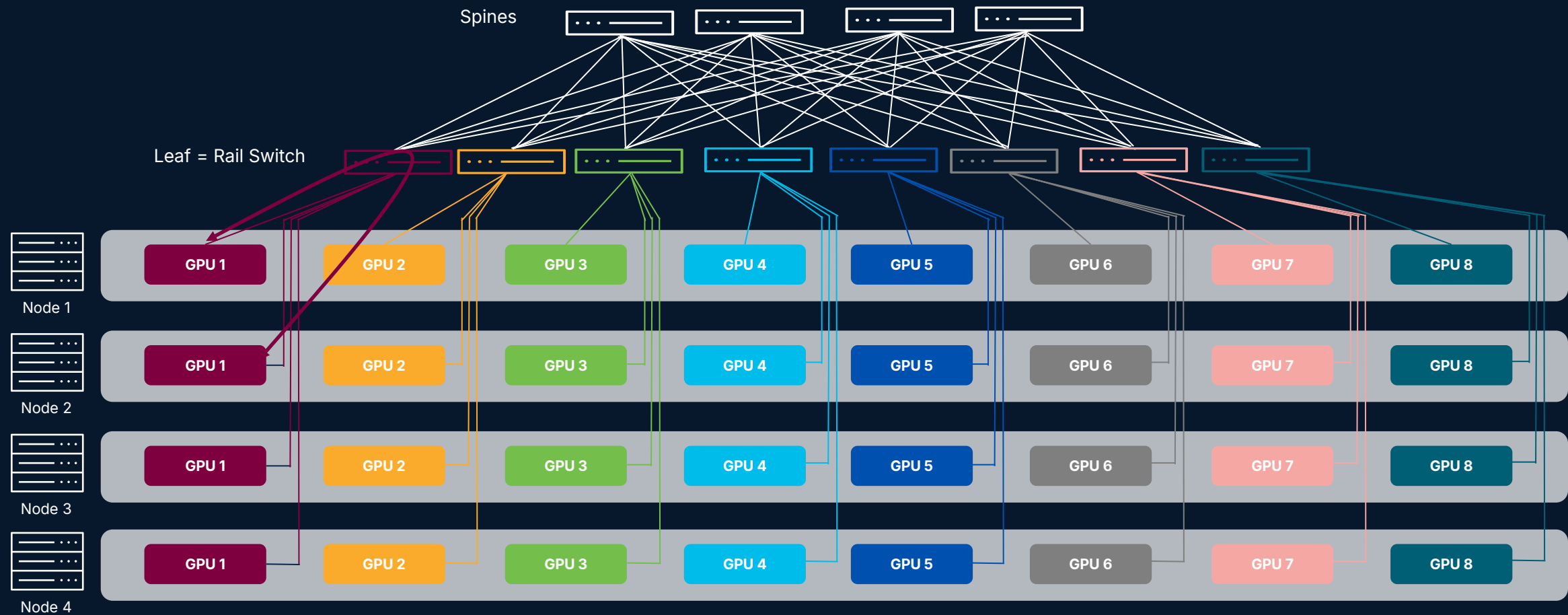
8 GPU:NIC pair mapped per Rail/Switch

Non-blocking / 1:1 Oversubscription

Up to 64/128-port 400G or 800G options for Leaf/Spine

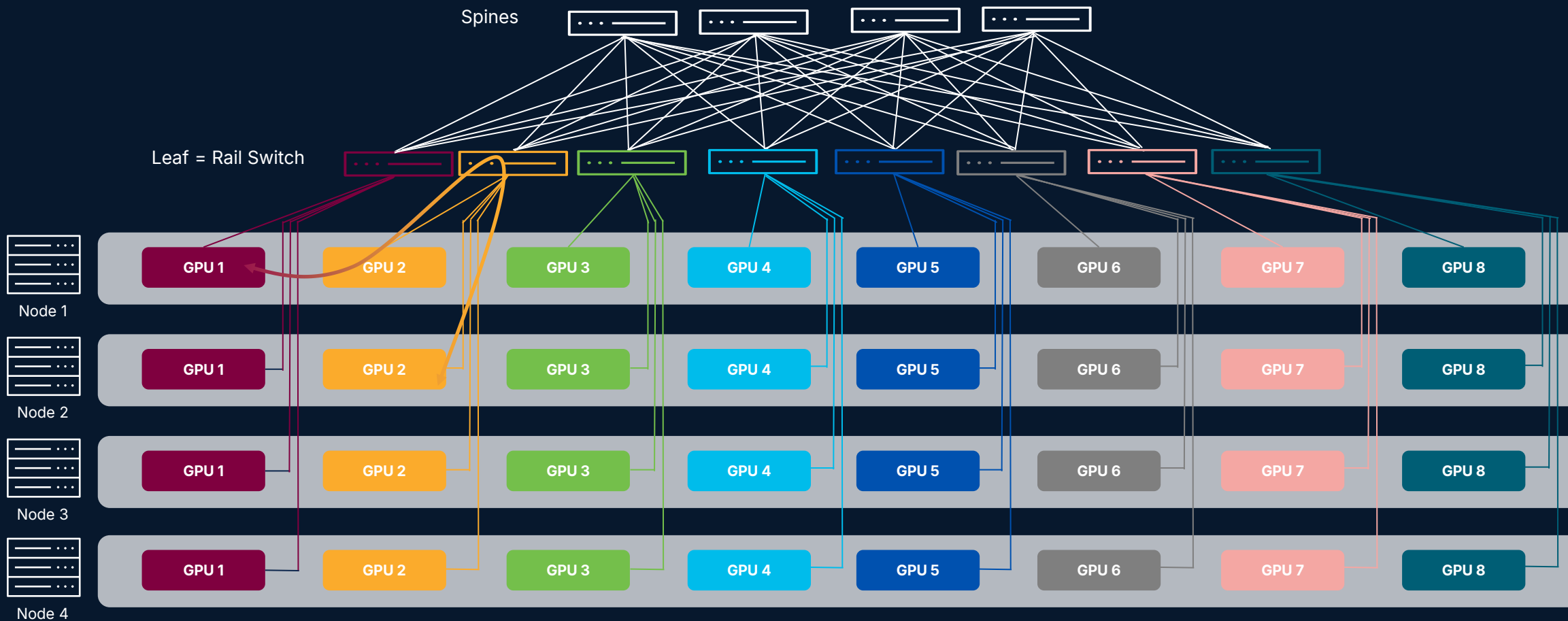
Backend Network

Rail-Optimized example with 4 nodes, 8-way GPUs



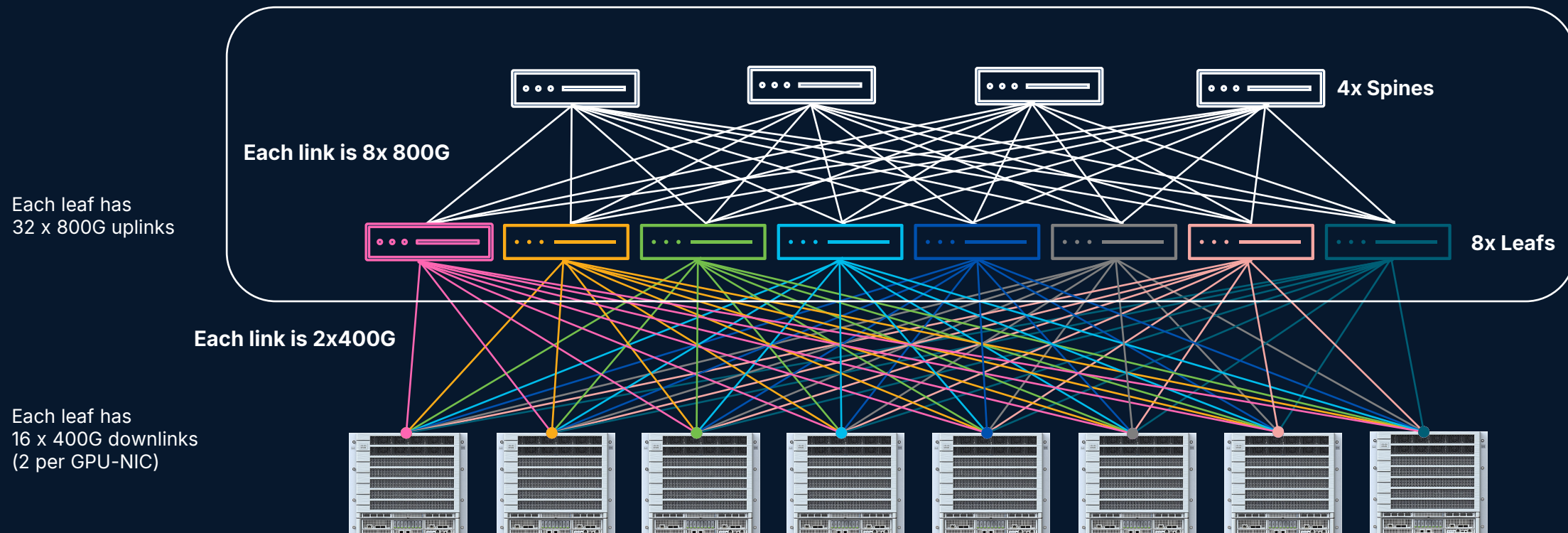
Backend Network

Rail-Optimized example with 4 nodes, 8-way GPUs



NVIDIA PXN (PCI x NVLink) allows connectivity between GPUs within the node to first move data on a GPU on the same rail as the destination, then send it to the destination without crossing rails.

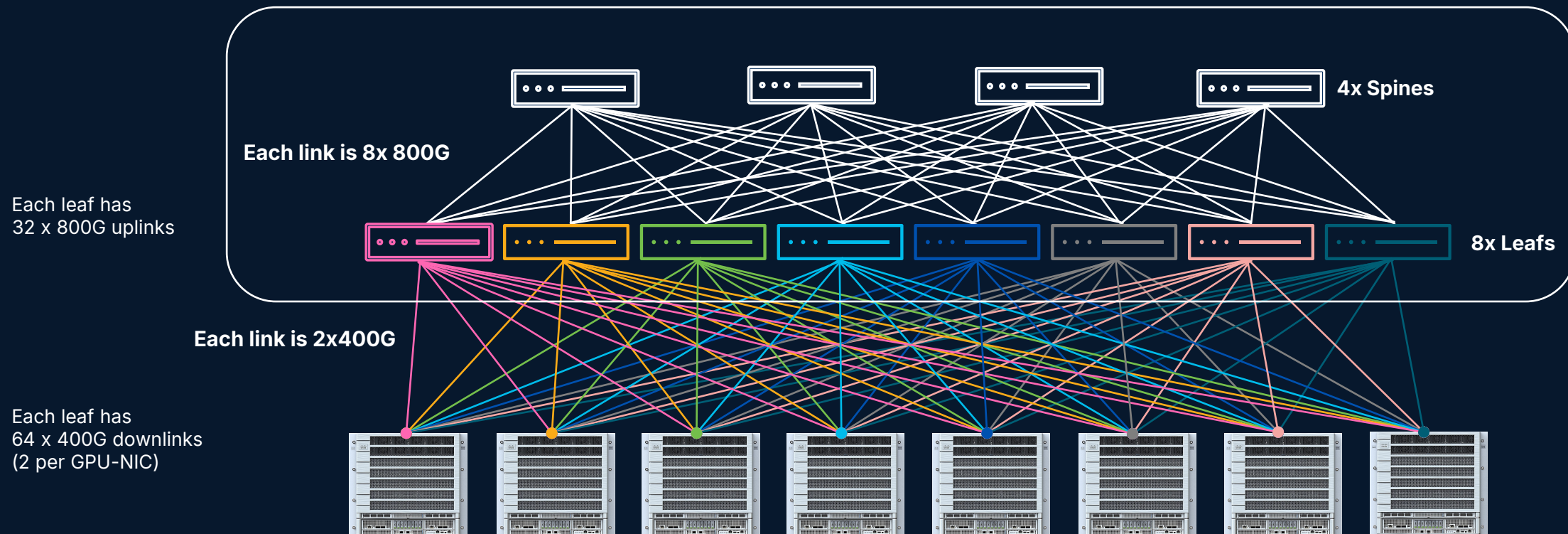
GPU Backend Network – Rail Optimized – 64 GPUs



- All inter-switch links use 800 Gbps transceivers (OSFP-800G-DR8=)
- 16x 400Gbps links per Leaf towards GPU Compute (8x RAILS) (OSFP-800G-DR8=)
- 32x 800 Gbps links per Leaf towards Spine Layer
- Dual MPO-12 SMF cables
- Non-blocking / 1:1 Oversubscription

8 x UCS 880A Nodes, 64 GPUs

GPU Backend Network – Rail Optimized – 256 GPUs



- All inter-switch links use 800 Gbps transceivers (OSFP-800G-DR8=)
- 16x 400Gbps links per Leaf towards GPU Compute (8x RAILS) (OSFP-800G-DR8=)
- 32x 800 Gbps links per Leaf towards Spine Layer
- Dual MPO-12 SMF cables
- Non-blocking / 1:1 Oversubscription

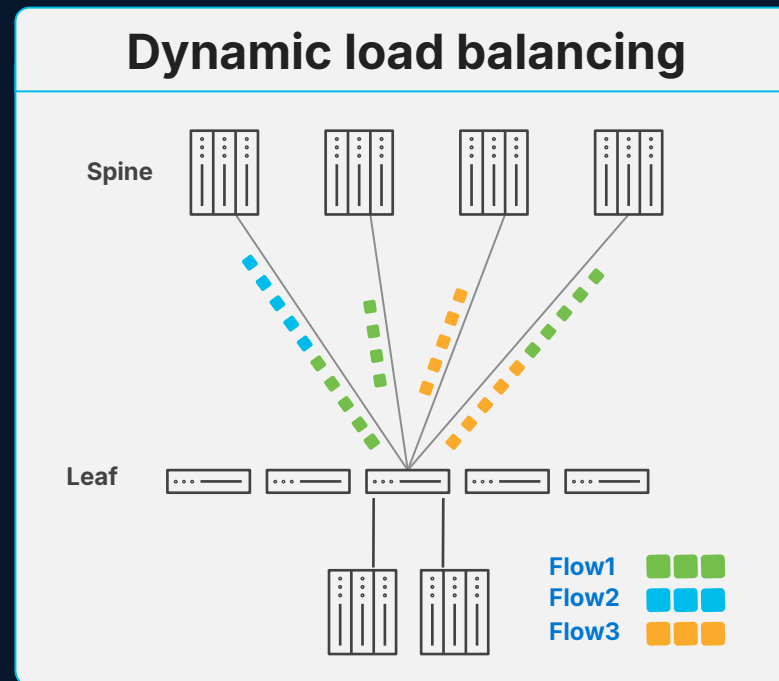
32 x UCS 880A Nodes, 256 GPUs

AI Backend Network – Key Requirements

- RDMA over Converged Ethernet (RoCEv2)
- Lossless network (PFC + ECN) = DCQCN
- Non-blocking (1:1 oversubscribed / non-oversubscribed)
- High Bandwidth / throughput
- Low Latency

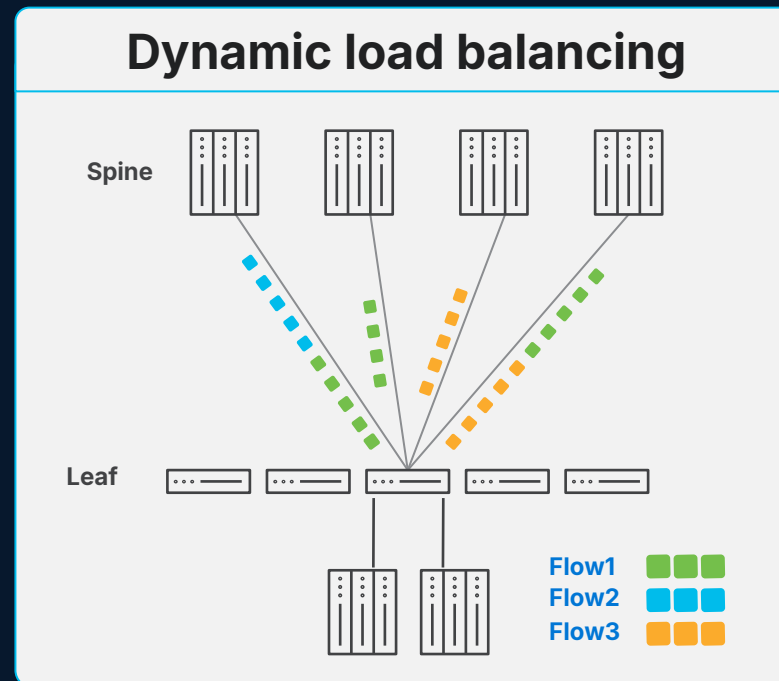
Dynamic Load Balancing (DLB)

- DLB distributes traffic across ECMP links in a fabric
- DLB operated at the IP layer (Layer 3)
- Implemented in hardware
- Enabled in NX-OS from release 10.5(1)F
- DLB monitors the traffic on ECMP links and selects the link with least utilization to distribute traffic
- Allows better traffic distribution, and prevents polarization of traffic over few links
- Layer 3 DLB is supported on RoCEv2 with spine-leaf architecture for AI/ML training networks



Dynamic Load Balancing (DLB) Modes

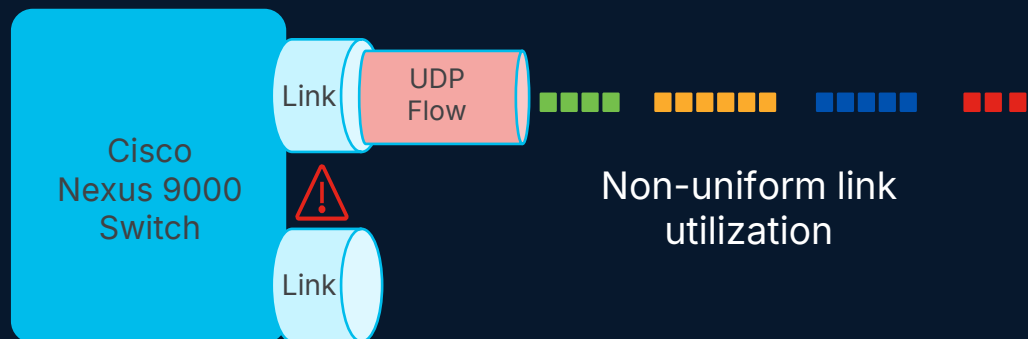
- DLB supports three modes of traffic load balancing
 - **Flowlet Load Balancing (FLB):** Flowlets are a group of packets in a flow identified by their 5-tuple with large enough gaps that allow them to be split and routed independently without causing reordering using DRE metrics
 - **Per-Packet Load balancing (PLB):** Load balancing decision is at a per-packet level instead of the flowlet level
 - **Static Pinning:** Allows for pairs of ingress and egress ports to be pinned together, user controls manually how the load is balanced over equal cost paths
- Advantages of DLB
 - Faster convergence and redundancy in case of link or node failures
 - Maximize the utilization of available network paths
 - Minimize congestion by evenly spreading traffic across all paths
 - Increase overall network performance without needing additional or specialized infrastructure



Improving Load Balancing in Inter-GPU Networks

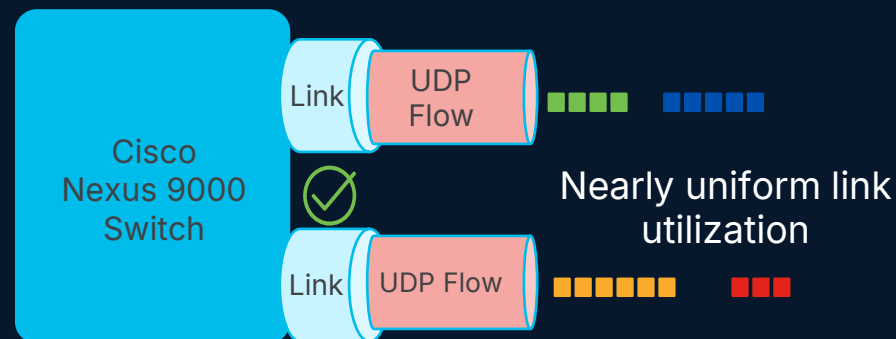
Dynamic Load Balancing (DLB)

Default ECMP load-sharing based
on Layer 4 UDP flow



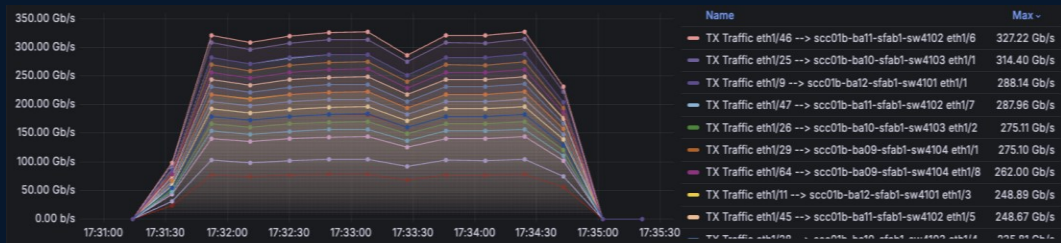
Non-uniform link
utilization

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

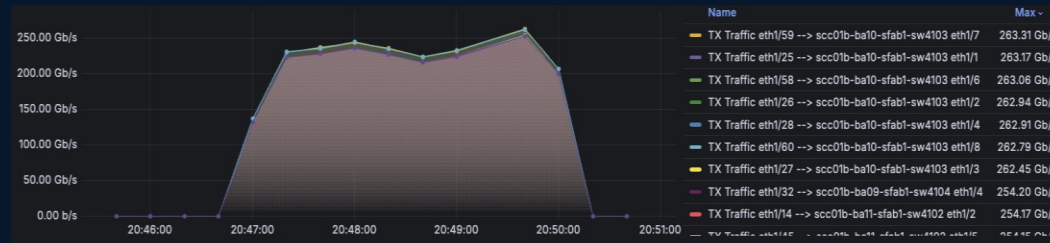


Nearly uniform link
utilization

Utilization of all equal-cost paths on a leaf switch



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

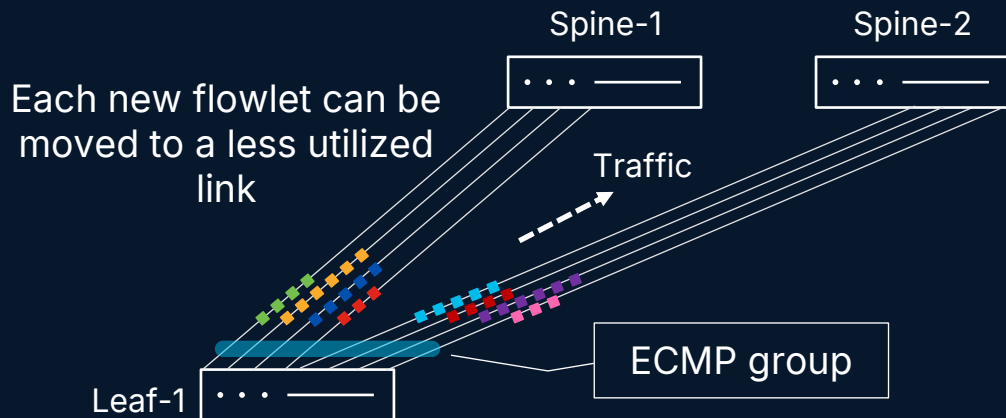


AllReduce 256 GPUs
(Max: 263.31 Gbps, Min: 253.66 Gbps)

Nearly uniform

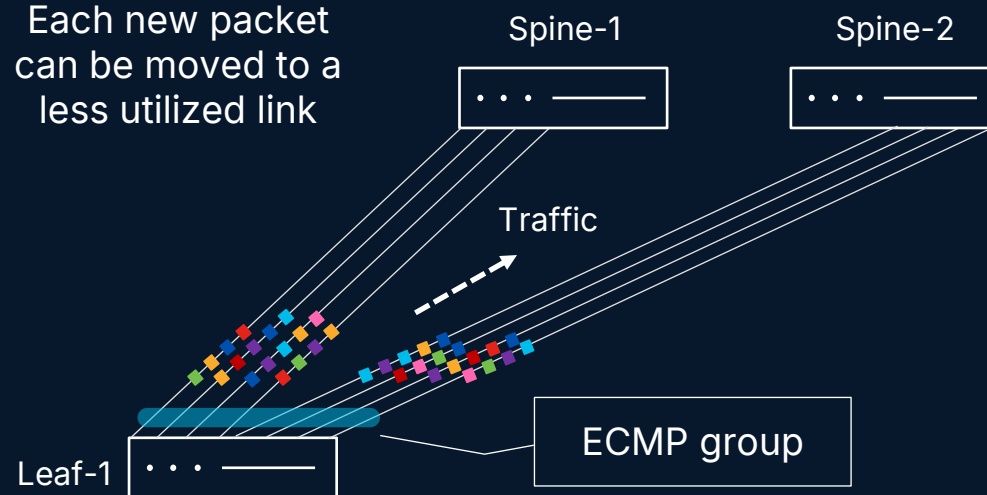
DLB Modes

Flowlet Load Balancing (FLB)



Flowlet – packet burst from a flow separated by large enough gap.
Flow is identified as per ip load-sharing command (ECMP or UDF)

Per-packet Load Balancing (PLB)



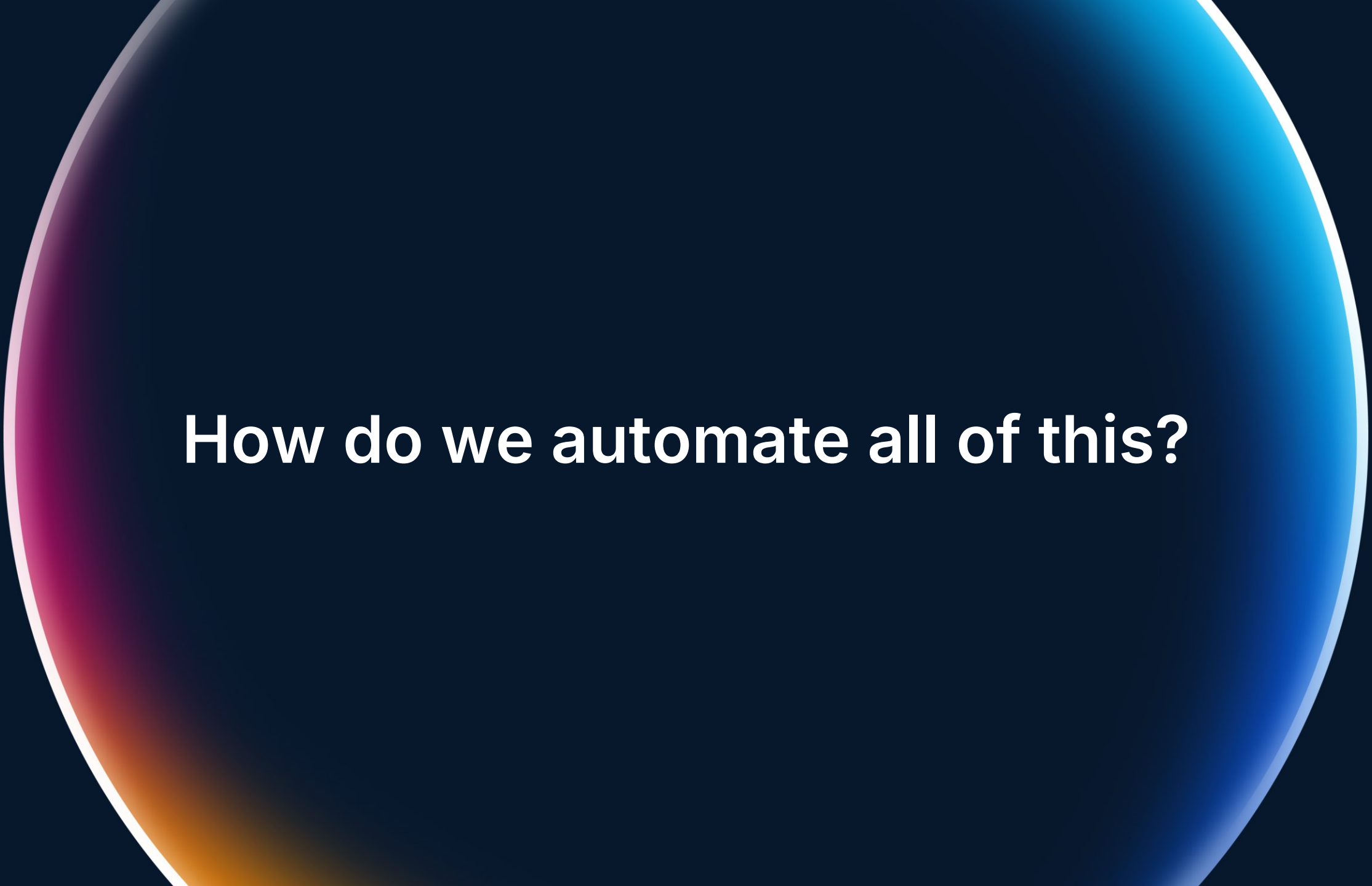
GPU Backend Networks – Key Points

Inter-GPU backend network



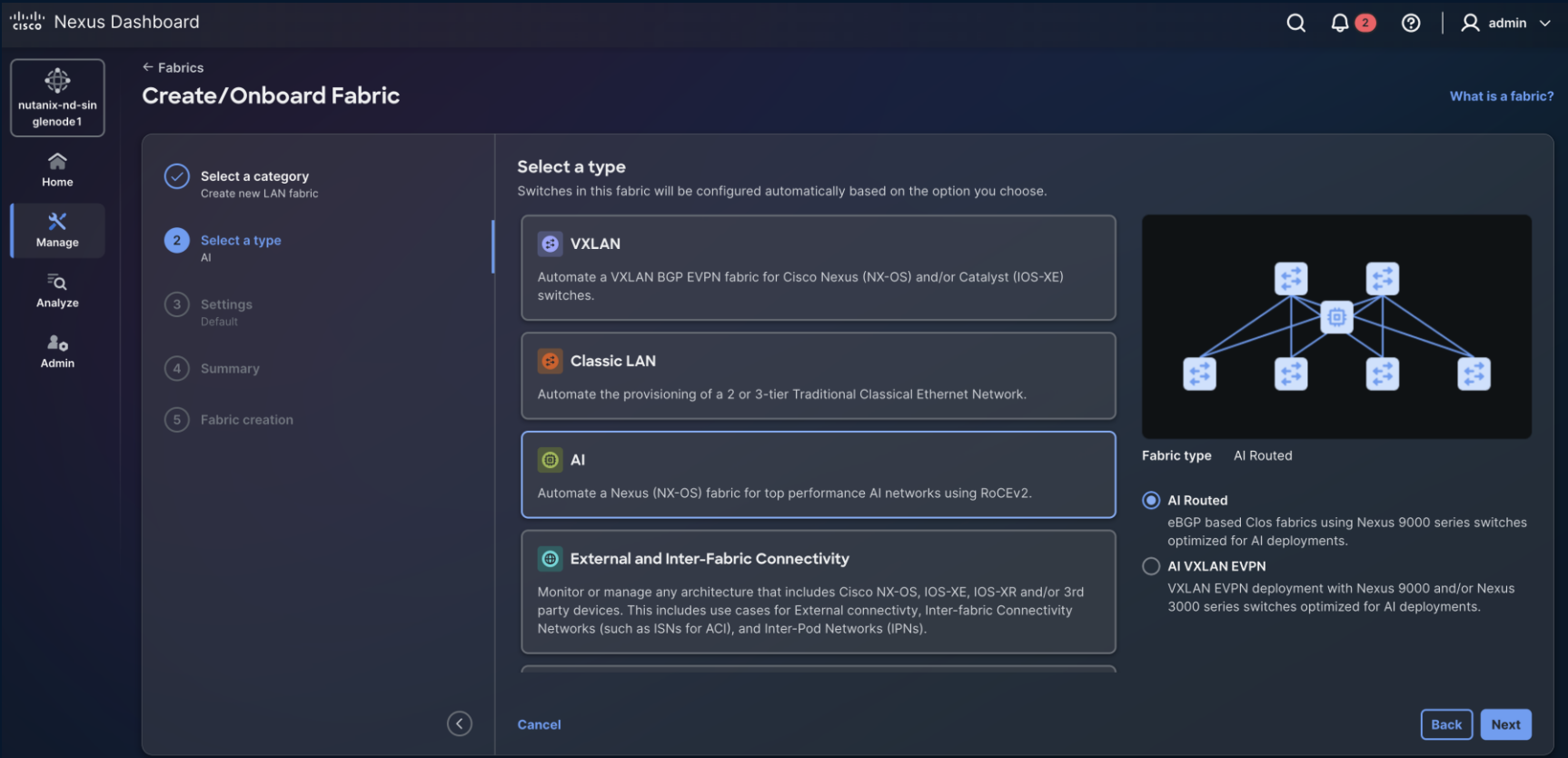
RoCEv2

1. **AI Routed:** Configure routing in the fabric: eBGP is preferred 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. (Optional) Enable WCMP, if using eBGP
2. **AI VXLAN EVPN:** Configure if you need flexible multi-tenancy and segmentation essential for GPU-as-a-Service (GPUaaS) models where multiple clients or projects share the same physical infrastructure.
3. **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
4. **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)
5. **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
6. **Dynamic Load Balancing:** Enable DLB Flowlet or per packet mode for use with NVIDIA Spectrum-X Adaptive Routing.
7. **Observability:** Network, GPU nodes, and storage nodes



How do we automate all of this?

Nexus Dashboard – AI Fabric Creation



Nexus Dashboard – AI Routed Fabric

nutanix-nd-sin

glenode1

Home

Manage

Analyze

Admin

← Fabrics

Create/Onboard Fabric

What is a fabric?

✓ Select a category

Create new LAN fabric

✓ Select a type

AI

3 Settings

Default

✓ Summary

5 Fabric creation

Settings

These are the recommended settings for configuring the parameters and capabilities of the new fabric.

Configuration mode ⓘ

☒ Default

☐ Advanced

Name *

Cisco_AI_Fabric

Location *

San Francisco, US

BGP ASN for spines *

65000

1-4294967295 | 1-65535[0-65535]

License tier for fabric ⓘ

☐ Essentials

☐ Advantage

☒ Premier

Enabled features

☒ Telemetry ⓘ

Cancel

Back

Next



Fabric type

AI Routed

© 2026 Cisco and/or its affiliates. All rights reserved.

BRKXAR-2041

 CISCO

Cisco
Confidential

Nexus Dashboard – AI VXLAN EVPN Fabric

nutanix-nd-sin

glenode1

Home

Manage

Analyze

Admin

← Fabrics

Create/Onboard Fabric

What is a fabric?

✓ Select a category
Create new LAN fabric

✓ Select a type
AI

3 Settings
Advanced

✓ AI Settings

5 Advanced settings

6 Summary

7 Fabric creation

Settings

These are the recommended settings for configuring the parameters and capabilities of the new fabric.

Configuration mode ⓘ
☐ Default ☒ Advanced

Name *

Location *

Overlay routing protocol
☐ iBGP ☒ eBGP

BGP ASN for spines *

1-4294967295 | 1-65535[0-65535]

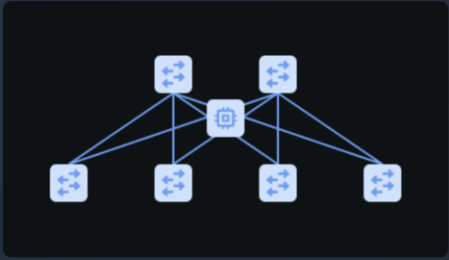
License tier for fabric ⓘ
☐ Essentials ☐ Advantage ☒ Premier

Enabled features
☒ Telemetry ⓘ

Cancel

Back

Next



Fabric type AI Data Center VXLAN EVPN - eBGP

© 2026 Cisco and/or its affiliates. All rights reserved.

BRKXAR-2041

 CISCO

Cisco
Confidential

Nexus Dashboard – AI QoS Policies

Nexus Dashboard

🔍

🔔 2

?

admin

nutanix-nd-sin
glenode1

Home

Manage

Analyze

Admin

← Fabrics

Create/Onboard Fabric

What is a fabric?

✓ Select a category
Create new LAN fabric

✓ Select a type
AI

✓ Settings
Advanced

4 AI Settings

5 Advanced settings

6 Summary

7 Fabric creation

AI Settings

These are the recommended settings for configuring the parameters and capabilities of the new fabric.

AI QoS & queuing policy

Policy selection mode ⓘ

800G

User-defined

800G ✓

400G

100G

25G

Flowlet aging timer ⓘ

500

Cancel

Back

Next

© 2026 Cisco and/or its affiliates. All rights reserved.

BRKXAR-2041

 CISCO

Cisco
Confidential

Nexus Dashboard – DLB Modes

Nexus Dashboard

Q

🔔 2

?

admin

nutanix-nd-sin

glenode 1

Home

Manage

Analyze

Admin

← Fabrics

Create/Onboard Fabric

What is a fabric?

✓ Select a category
Create new LAN fabric

✓ Select a type
AI

✓ Settings
Advanced

4 AI Settings

5 Advanced settings

6 Summary

7 Fabric creation

AI Settings

These are the recommended settings for configuring the parameters and capabilities of the new fabric.

AI QoS & queuing policy

Policy selection mode ⓘ
800G

Dynamic load balancing

Enable

Mode selection ⓘ
Flowlet
Flowlet ✓
Per packet
Policy-driven-flowlet
Policy-driven-per-packet
Policy-driven-mixed-mode

Back

Next

© 2026 Cisco and/or its affiliates. All rights reserved.

BRKXAR-2041

 CISCO

Cisco
Confidential

Nexus Dashboard – Advanced Settings

Nutanix-nd-sin

glenode1

Home

Manage

Analyze

Admin

← Fabrics

Create/Onboard Fabric

What is a fabric?

✓ Select a category

Create new LAN fabric

✓ Select a type

AI

✓ Settings

Advanced

✓ AI Settings

5 Advanced settings

6 Summary

7 Fabric creation

Advanced settings

The following optional settings will be deployed and/or used when deploying this fabric.

General Parameters

EVPN

vPC

Protocols

Authentication

Security

Advanced

Freeform

Manageability

Hypershield

Bootstrap

Configuration Backup

BGP ASN for super spines

1-4294967295 | 1-65535[.0-65535]

BGP ASN for leafs

1-4294967295 | 1-65535[.0-65535], applicable when 'BGP AS mode' is same-tier-as

BGP ASN for border and border gateway switches

1-4294967295 | 1-65535[.0-65535], applicable when 'BGP AS mode' is same-tier-as

BGP AS mode*

multiAS

Multi-as: unique ASN per leaf/border/border gateway (borders and border gateways are allowed to share ASN). Same-tier-as: leafs share one ASN, borders/border gateways share one ASN.

☒ ASN auto allocation

Automatically allocate and track BGP ASN for leafs, borders and border gateways in multi-as mode. It cannot be enabled when 'allow same ASN on leafs' is enabled

BGP ASN range

Cancel

Back

Next

© 2026 Cisco and/or its affiliates. All rights reserved.

BRKXAR-2041

Cisco

Confidential

Nexus Dashboard – Create Fabric

nutanix-nd-sin

glenode1

Home

Manage

Analyze

Admin

← Fabrics

Create/Onboard Fabric

[What is a fabric?](#)

✓ Select a category

Create new LAN fabric

✓ Select a type

AI

✓ Settings

Advanced

✓ AI Settings

✓ Advanced settings

6 Summary

7 Fabric creation

Summary

Review your selections below.

Category

Fabric categoryNew LAN fabric

Type

Fabric typeAI

Fabric subtypeAI Data Center VXLAN EVPN - eBGP

Settings

NameCisco_AI_Pod

LocationSan Francisco, US

License tier for fabricPremier

Security domainall

Overlay routing protocolebgp

BGP ASN for spines65000

Enabled featuresTelemetry

Cancel

Back

Submit

© 2026 Cisco and/or its affiliates. All rights reserved.

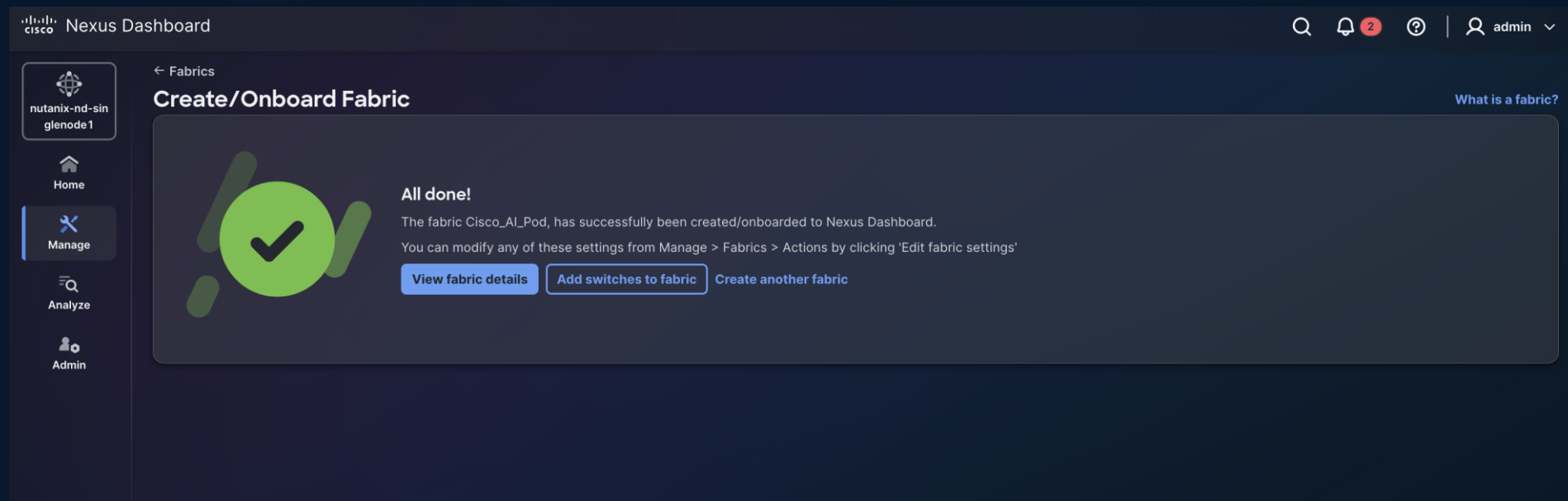
BRKXAR-2041

CISCO

Cisco

Confidential

Nexus Dashboard – Fabric Created!



Cisco NX-OS Configuration

QoS for RoCEv2

All switches

```
policy-map type network-qos qos_network
  class type network-qos c-8q-nq3
    mtu 9216
    pause pfc-cos 3
  class type network-qos c-8q-nq-default
    mtu 9216
class-map type qos match-any CNP
  match dscp 48
class-map type qos match-any ROCEv2
  match dscp 24
policy-map type qos QOS_CLASSIFICATION
  class ROCEv2
    set qos-group 3
  class CNP
    set qos-group 7
  class class-default
    set qos-group 0

! continued...
```

Enable PFC for COS 3

Match with the host config for prioritizing Congestion Notification Packet (CNP) for DCQCN congestion control

Match with the host config for enabling PFC for packets marked with DSCP 24, which maps to COS 3

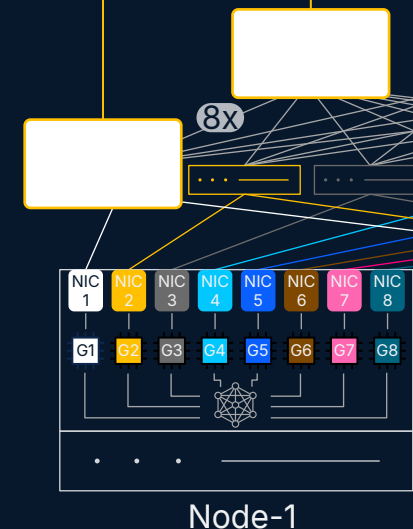
Assign RoCEv2/lossless traffic to qos-group 3

Assign CNP traffic to qos-group 7

Inter-GPU backend network



RoCEv2



Or

Use Nexus Dashboard
"AI Routed" fabric

Cisco NX-OS Configuration

QoS for RoCEv2

All switches

```
! ...continued...
policy-map type queuing QOS_EGRESS_PORT
  class type queuing c-out-8q-q6
    bandwidth remaining percent 0
  class type queuing c-out-8q-q5
    bandwidth remaining percent 0
  class type queuing c-out-8q-q4
    bandwidth remaining percent 0
  class type queuing c-out-8q-q3
    bandwidth remaining percent 99
    random-detect minimum-threshold 950
    kbytes maximum-threshold 3000 kbytes
    drop-probability 7 weight 0 ecn
  class type queuing c-out-8q-q2
    bandwidth remaining percent 0
  class type queuing c-out-8q-q1
    bandwidth remaining percent 0
  class type queuing c-out-8q-q-default
    bandwidth remaining percent 1
  class type queuing c-out-8q-q7
    priority level 1
! ... continued...
```

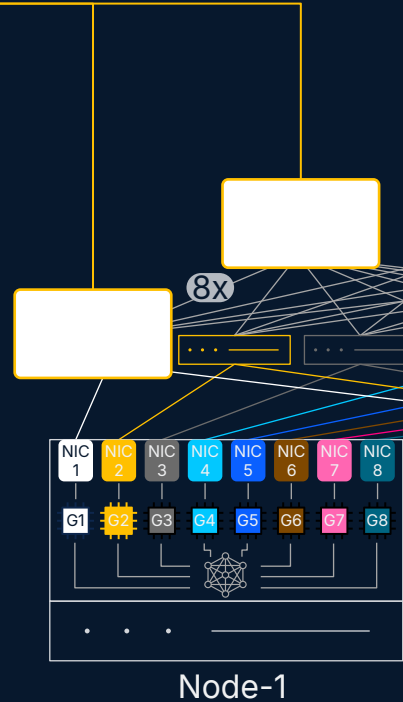
Enable ECN marking for RoCEv2/lossless traffic using WRED

Assign CNP to priority queue for faster congestion control action

Inter-GPU backend network



RoCEv2



Or

Use Nexus Dashboard
"AI Routed" fabric

Cisco NX-OS Configuration

QoS for RoCEv2

Inter-GPU backend network



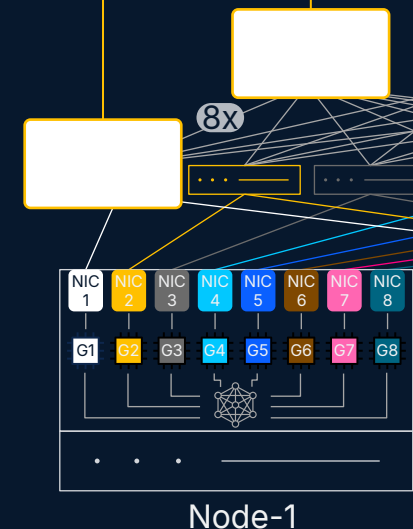
RoCEv2

All switches

```
! ...continued...
system qos
  service-policy type network-qos qos_network
  service-policy type queuing output QOS_EGRESS_PORT
!
interface Ethernet1/1-64
  service-policy type qos input QOS_CLASSIFICATION
  priority-flow-control mode on
!
```

Apply QoS policies on all interfaces

Enable PFC on all interfaces



Or

Use Nexus Dashboard
"AI Routed" fabric

IP DSCP, IP ToS, and Ethernet CoS/PCP Mapping

- Cisco switches use DSCP-to-CoS mapping for lossless behavior in L2-L3 boundaries (e.g., VXLAN)
- NVIDIA NICs use ToS values for configuring PFC for RoCEv2 traffic

DSCP Name	DSCP Binary	DSCP Decimal	ToS Decimal (DSCP binary << 2)	CoS/PCP Binary (DSCP binary >> 3)	CoS/PCP Decimal
CS0	000000	0	0	000	0
CS1 / AF11 / AF12 / AF13	001000 / 001010 / 001100 / 001110	8 / 10 / 12 / 14	32 / 40 / 48 / 56	001	1
CS2 / AF21 / AF22 / AF23	010000 / 010010 / 010100 / 010110	16 / 18 / 20 / 22	64 / 72 / 80 / 88	010	2
CS3 / AF31 / AF32 / AF33	011000 / 011010 / 011100 / 011110	24 / 26 / 28 / 30	96 / 104 / 112 / 120	011	3
CS4 / AF41 / AF42 / AF43	100000 / 100010 / 100100 / 100110	32 / 34 / 36 / 38	128 / 136 / 144 / 152	100	4
CS5 / / / EF	101000 / 101010 / 101100 / 101110	40 / 42 / 44 / 46	160 / 168 / 176 / 184	101	5
CS6	110000 / 110010 / 110100 / 110110	48 / 50 / 52 / 54	192 / 200 / 208 / 216	110	6
CS7	111000 / 111010 / 111100 / 111110	56 / 58 / 60 / 62	224 / 232 / 240 / 248	111	7

Considerations for Nexus 9300 GX2A/2B Switches

- Upgrade switches to NX-OS 10.5(2) or later
- Enable hardware qos pfc static on all switches and reload
 - Introduced in NX-OS 10.(5)2 on Nexus 9300 GX2A/2B switches
 - Dedicates buffers per port for efficient handling of large bursts on all ports simultaneously
 - Increases traffic throughput in no-drop/lossless traffic class

Cisco NX-OS

```
N9364D-GX2A(config)# hardware qos pfc static
Warning: Please save running config and reload system for configuration to take effect.
N9364D-GX2A(config)# end
N9364D-GX2A # copy running-config startup-config
[#####] 100%
Copy complete, now saving to disk (please wait)...
Copy complete.
N9364D-GX2A # reload
This command will reboot the system. (y/n)? [n] y
```

Considerations for Nexus 9364E-SG2 Switches

- Switch support requires NX-OS 10.5(3)F or later
- Spine to Leaf connections utilizing 9364E-SG2 switches can be native 800G
- Breakout configuration is required on server facing ports towards GPU compute CX-8 SuperNIC - 2 x 400G (OSFP single optic with 2 ports). A reload is required post breakout.

Cisco NX-OS

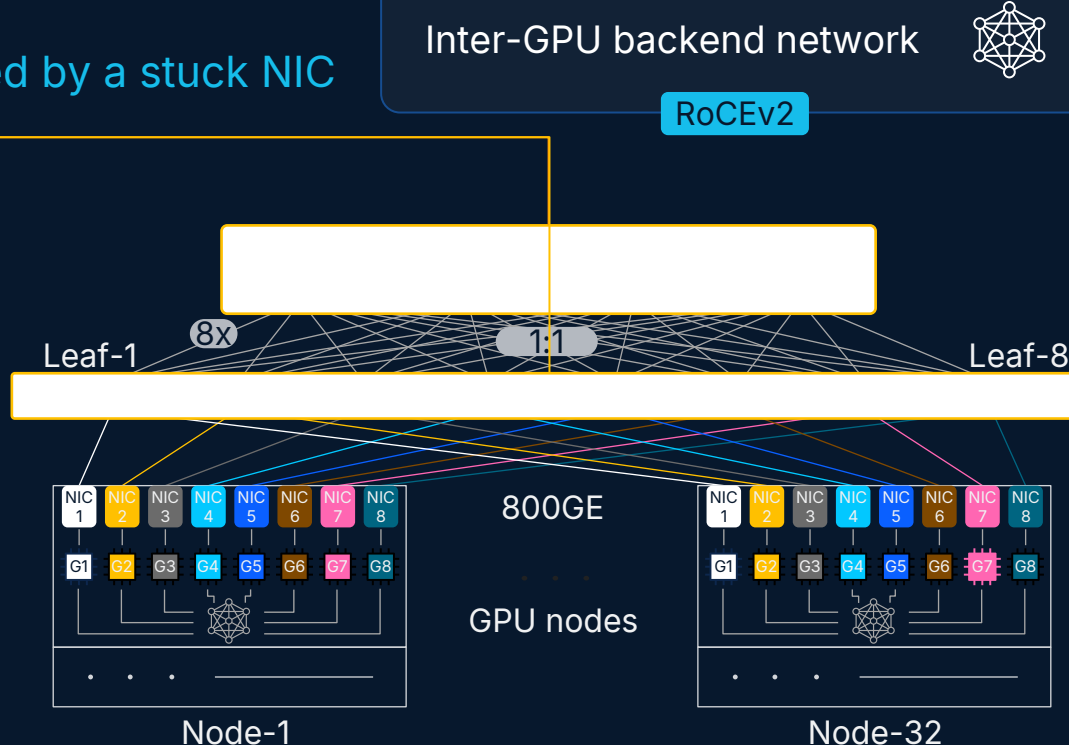
```
N9364E-SG2-O(config)#
interface breakout module 1 port 1-2,5-6,9-10,13-14 map 400g-2x
!
interface Ethernet1/1/1
  description GPUWORKER-NIC1
  switchport
  switchport access vlan 10
  priority-flow-control mode on
  priority-flow-control watch-dog-interval on
  spanning-tree port type edge
  spanning-tree bpduguard enable
  mtu 9216
  service-policy type qos input QOS_CLASSIFICATION
  no shutdown
!
interface Ethernet1/1/2
  description GPUWORKER2-NIC1
  switchport
  switchport access vlan 10
  priority-flow-control mode on
  priority-flow-control watch-dog-interval on
  spanning-tree port type edge
  spanning-tree bpduguard enable
  mtu 9216
  service-policy type qos input QOS_CLASSIFICATION
  no shutdown
```

NX-OS Configuration - PFC

Enable PFC Watchdog – Avoids congestion spreading caused by a stuck NIC

All switches

```
!
priority-flow-control watch-dog-interval on
!
interface Ethernet1/1 - 64
    priority-flow-control watch-dog-interval on
!
interface Ethernet1/1/1
    description gpuwkr001-NIC1
    switchport access vlan 100
    priority-flow-control mode on
    priority-flow-control watch-dog-interval on
    spanning-tree port type edge
    spanning-tree bpduguard enable
    mtu 9216
    service-policy type qos input QOS_CLASSIFICATION
    no shutdown
```



Use Nexus Dashboard
"AI Routed" fabric
or
"AI VXLAN Fabric"

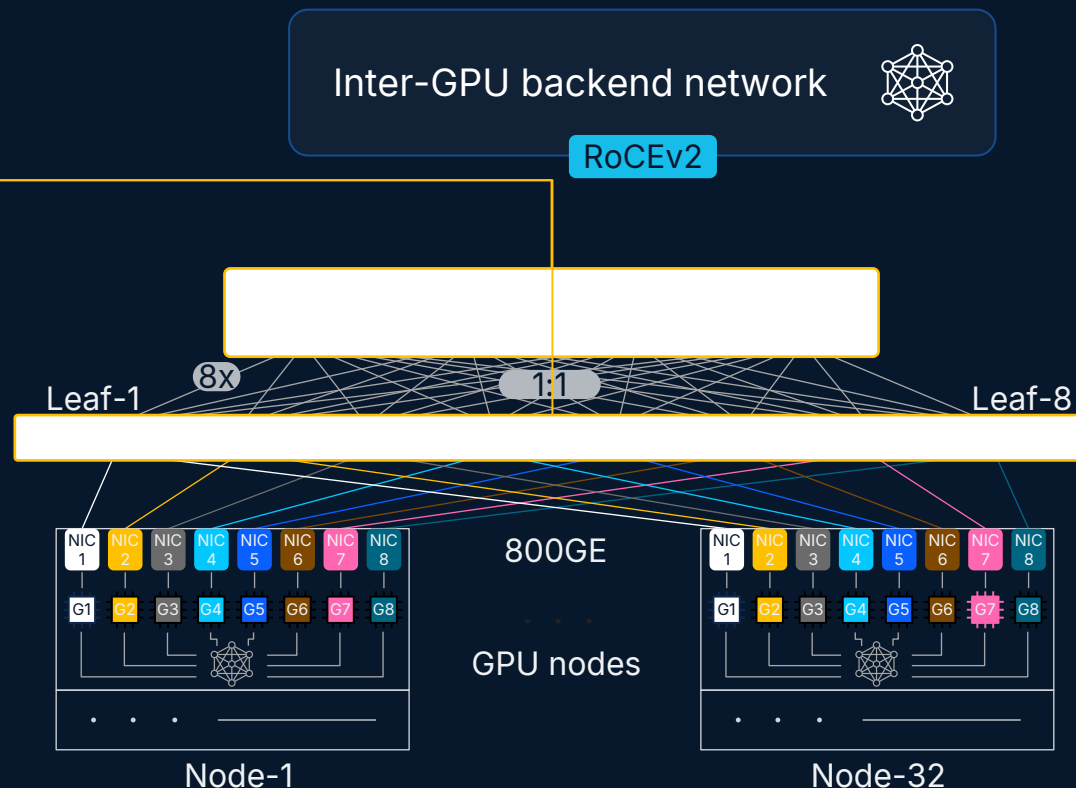
NX-OS Configuration - DLB

Dynamic Load Balancing (DLB)

All Switches

```
Spine#  
hardware profile dib  
flowlet-aging 500  
dlb-interface Ethernet1/1-64  
  
Leaf#  
hardware profile dlb  
flowlet-aging 500  
dlb-interface Ethernet1/3-4, Ethernet1/7-8, Ethernet1/11-12,  
Ethernet1/15-16, Ethernet1/19-20, Ethernet1/23-24, Ethernet1/27-28,  
Ethernet1/31-32, Ethernet1/35-36, Ethernet1/39-40, Ethernet1/43-44,  
Ethernet1/47-48, Ethernet1/51-52, Ethernet1/55-56, Ethernet1/59-60,  
Ethernet1/63-64
```

- Apply to all Spine Interfaces and all Spine facing interfaces on Leafs
- Requires a free-form template for Silicon One based switches in ND 4.1 and earlier
- Included in ND 4.2 and above in the AI setting tab under Fabric Settings
- If using OSPF underlay with more than 8 uplinks from Leaf to Spine, increase OSPF maximum paths from default value of 8
- For flowlet mode, adjust flowlet aging time if traffic is not evenly distributed between spine and leaf links





Compute and Storage Design and Considerations

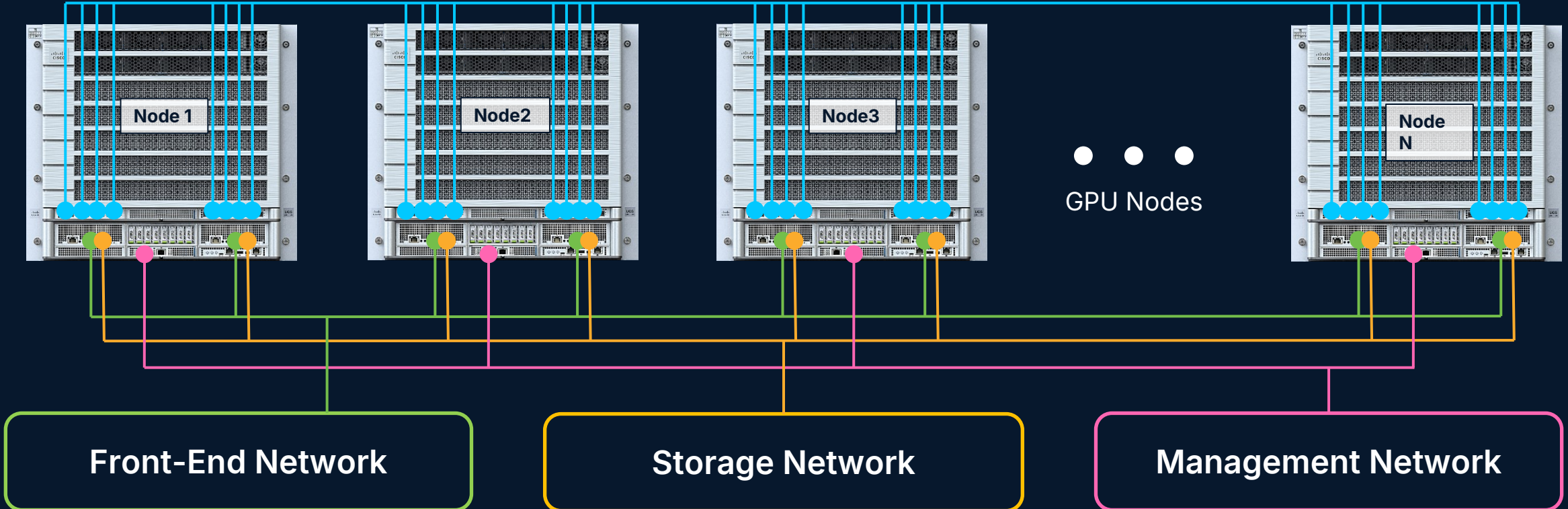
UCS C880A M8 Network Definitions

- **Inter-GPU backend network:** An Inter-GPU backend network connects the dedicated GPU ports for running distributed training. This network is also known as the back-end network, compute fabric, or scale-out network.
- **Frontend network:** A frontend network connects the GPU nodes to the data center network for inferencing, logging, managing in-band devices, and so on.
- **Storage network:** A storage network connects the GPU nodes to the shared storage devices providing parallel file system access to all the nodes for loading (reading) the data sets for training, and checkpointing (writing) the model parameters as they are learned. Some users may share the front-end network to connect storage devices, eliminating a dedicated storage network.
- **Management network:** A management network provides out-of-band connectivity to the devices of the AI/ML infrastructure, such as GPU nodes, network switches, and storage devices.

UCS C880A M8 Networking

Inter-GPU Backend Network

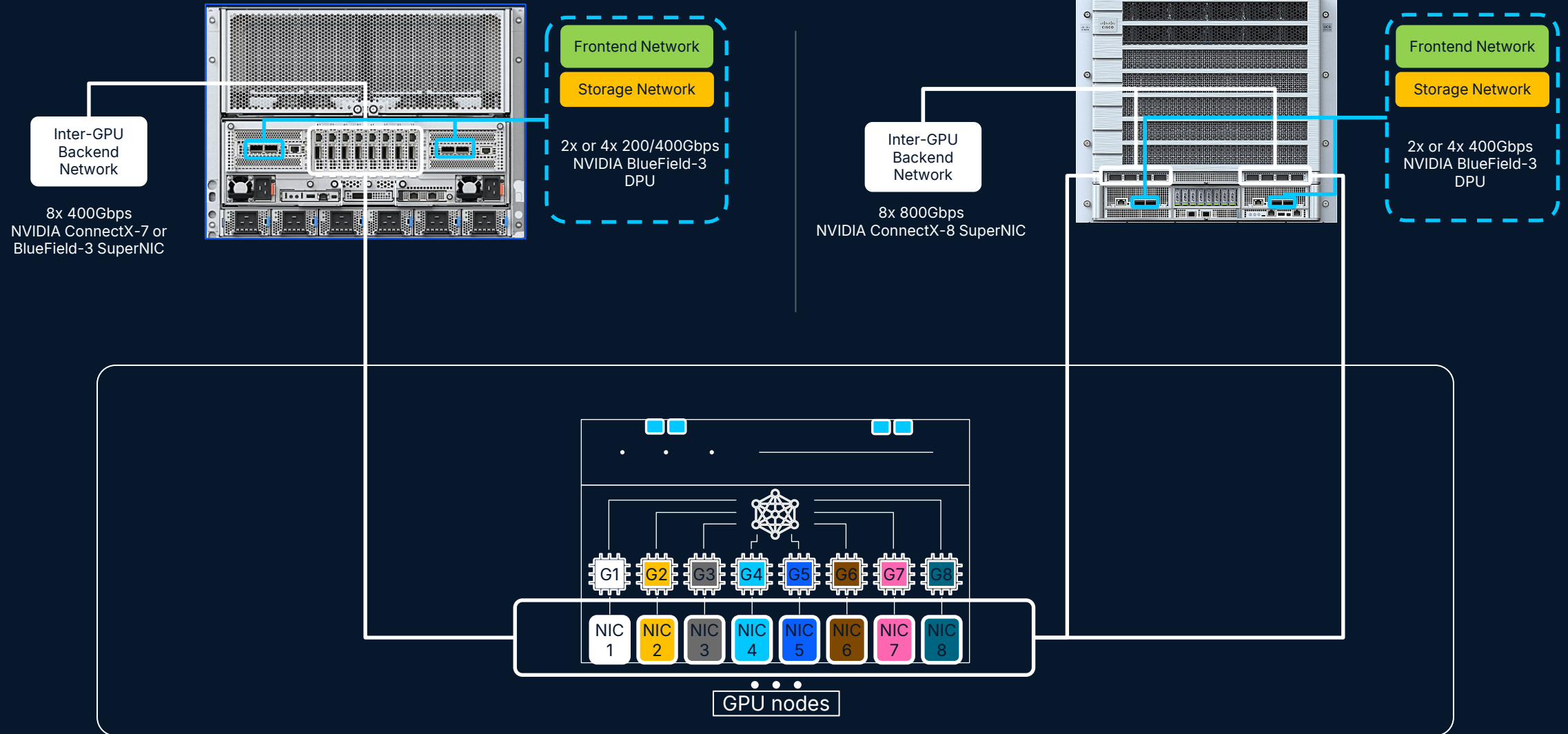
← GPUs sync their distributed training states via inter-GPU backend network →



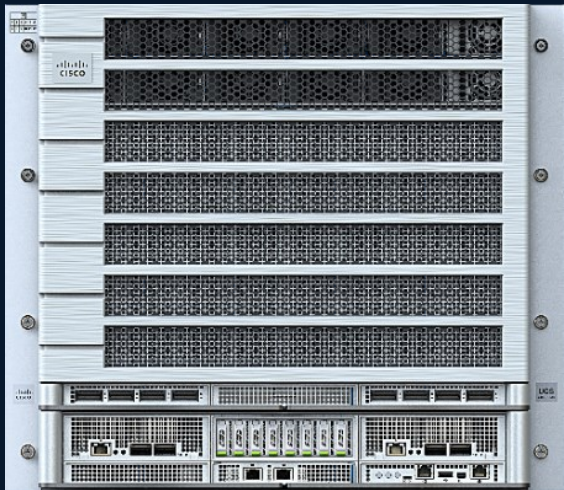
GPU Servers – Cisco UCS 885A and UCS 880A

Cisco UCS 885A (8x NVIDIA or AMD GPUs)

Cisco UCS 880A (8x NVIDIA B300 GPUs)



UCS C880A Dense GPU Server Specifications



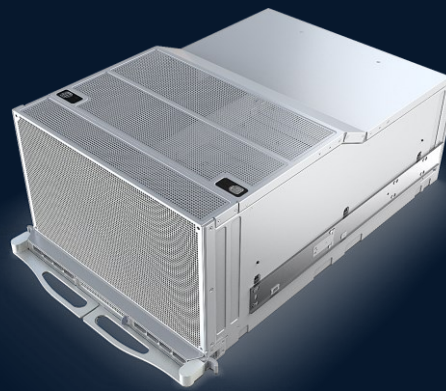
Product Specifications	
Form Factor	• HGX 10RU 19" Rack Server
Compute + Memory	• 2x 6th Gen Intel Xeon CPUs (Select SKUs for AI and HPC workloads) • Up to 32x DDR5 RDIMMS
Storage	• 2x M.2 SATA Boot Drives with HW RAID Controller (Boot) • Up to 8x PCIe Gen5 x4 E1.S NVMe SSDs (Data)
GPUs	• 8x NVIDIA HGX B300 NVL8 air-cooled GPUs
Network Cards	• E-W: Integrated ConnectX-8 • N-S: 4x PCIe Gen5 x16 FHHL slots, 1x OCP TSFF Gen5 x8
Cooling	• 20 Hot swappable FANs
Physical I/O	• 1 USB 3 type A, 1 mDP, 1 ID Button, 1 System Power Button, 1 Reset Button, 1 USB type C (for debugging), 1 RJ45 (OOB mgmt.), 1 RJ45 (LOM port)
Power Supply	• 12x 50V 3.2kW (N+N redundancy)

UCS C880A M8 Modular Sled Design

Main Chassis



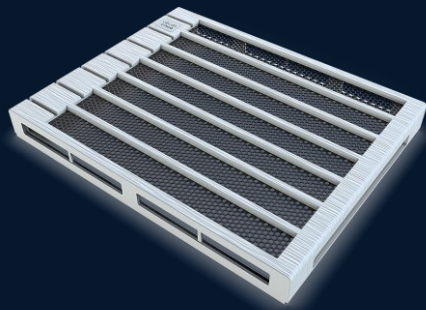
GPU Sled with CX8



CPU Sled



Front Bezel



GPU Block



FAN Module

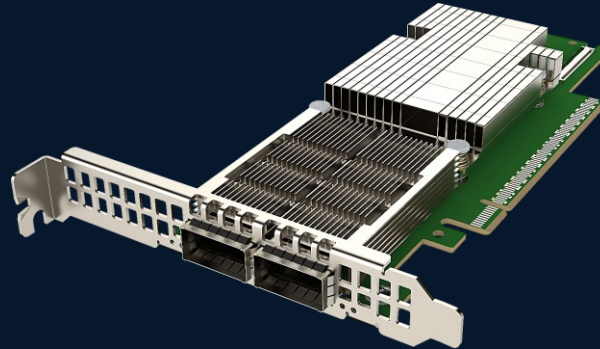


UCS C880A M8 Modular Sled Design

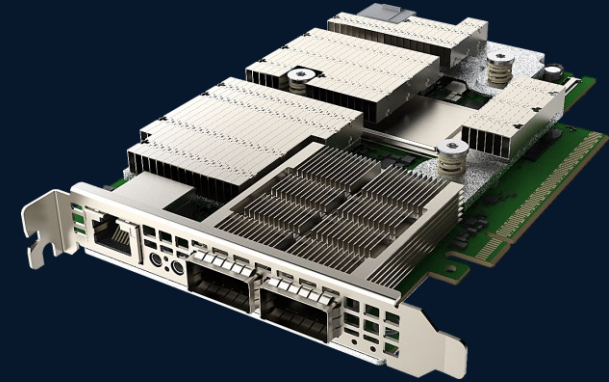
PSU



CX7 2x200G



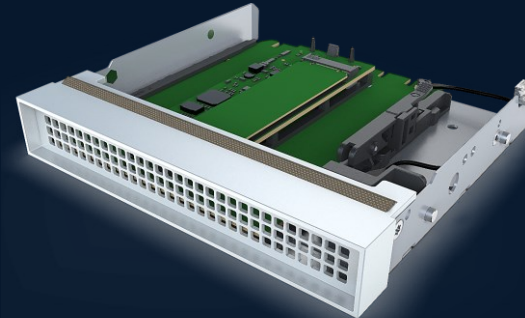
B3220/B3240



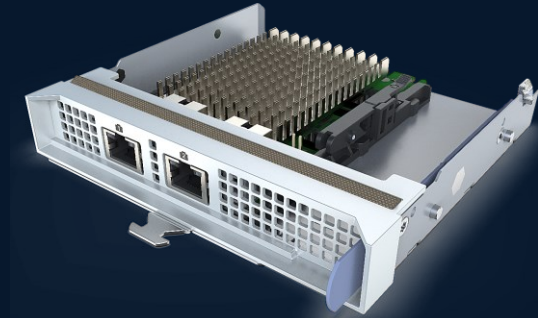
E1.S Drive



M.2 Module



OCP Module



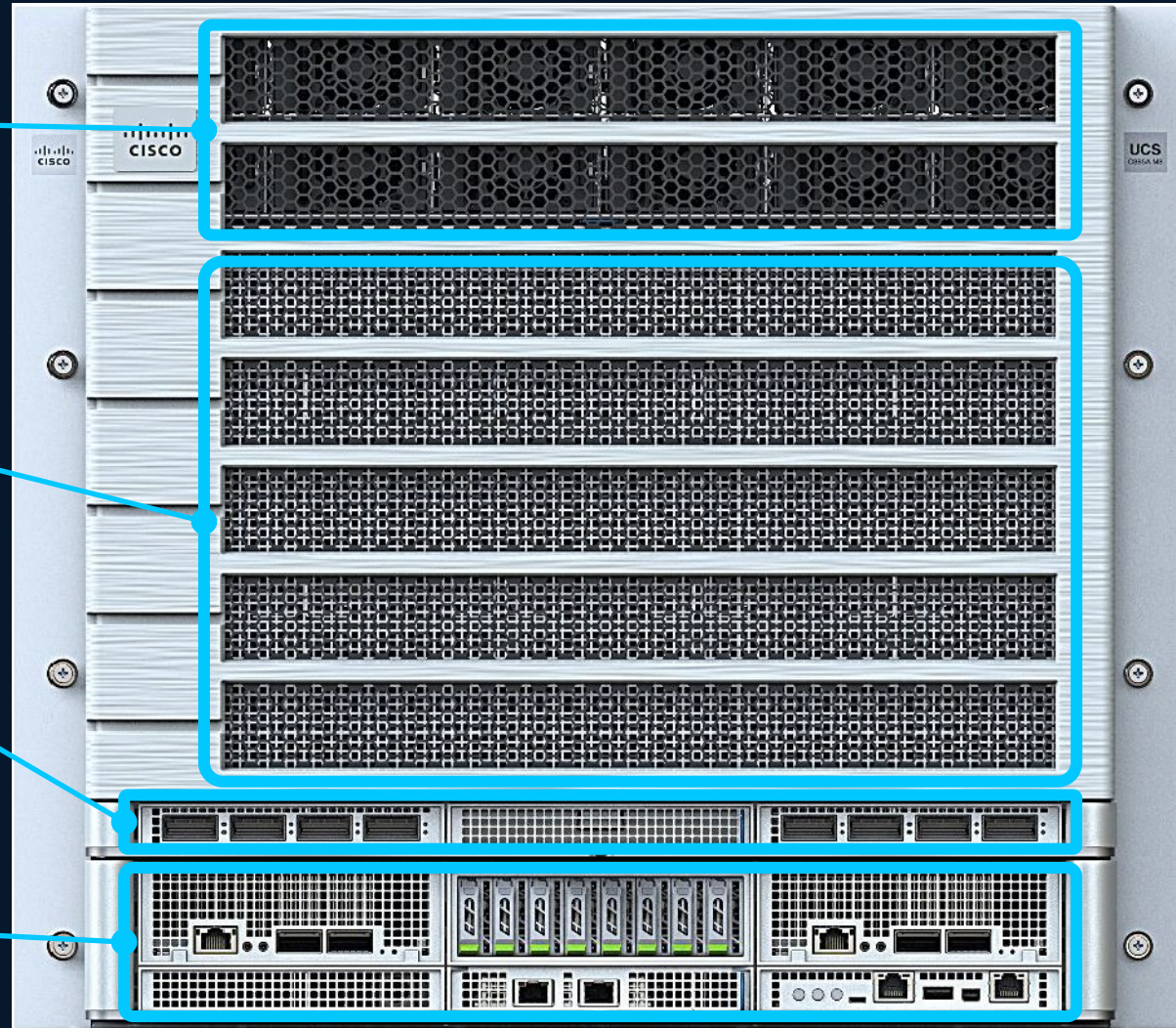
UCS C880A Front View

PSU
Exhaust

GPU
Exhaust

CX8 East /
West

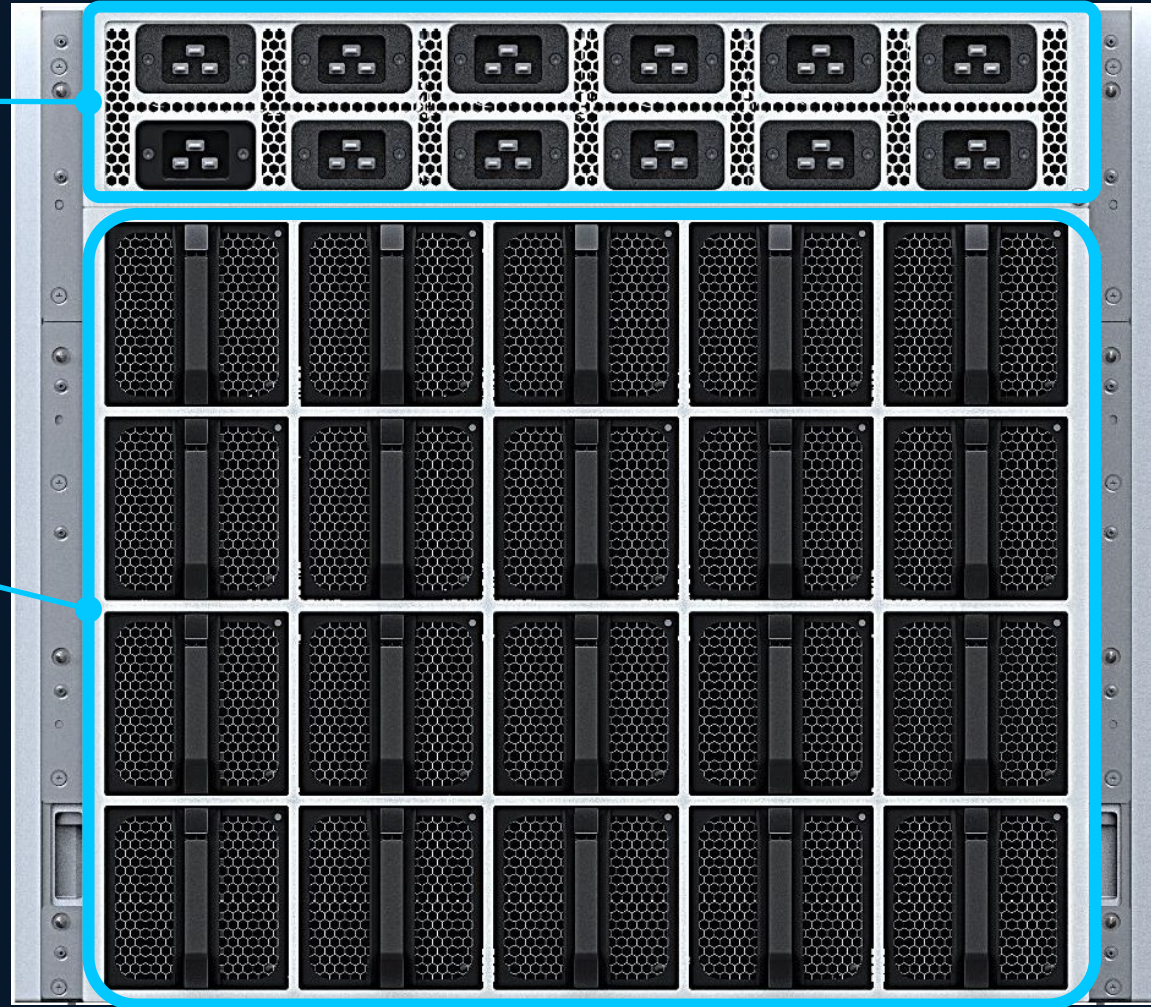
CPU
Sled



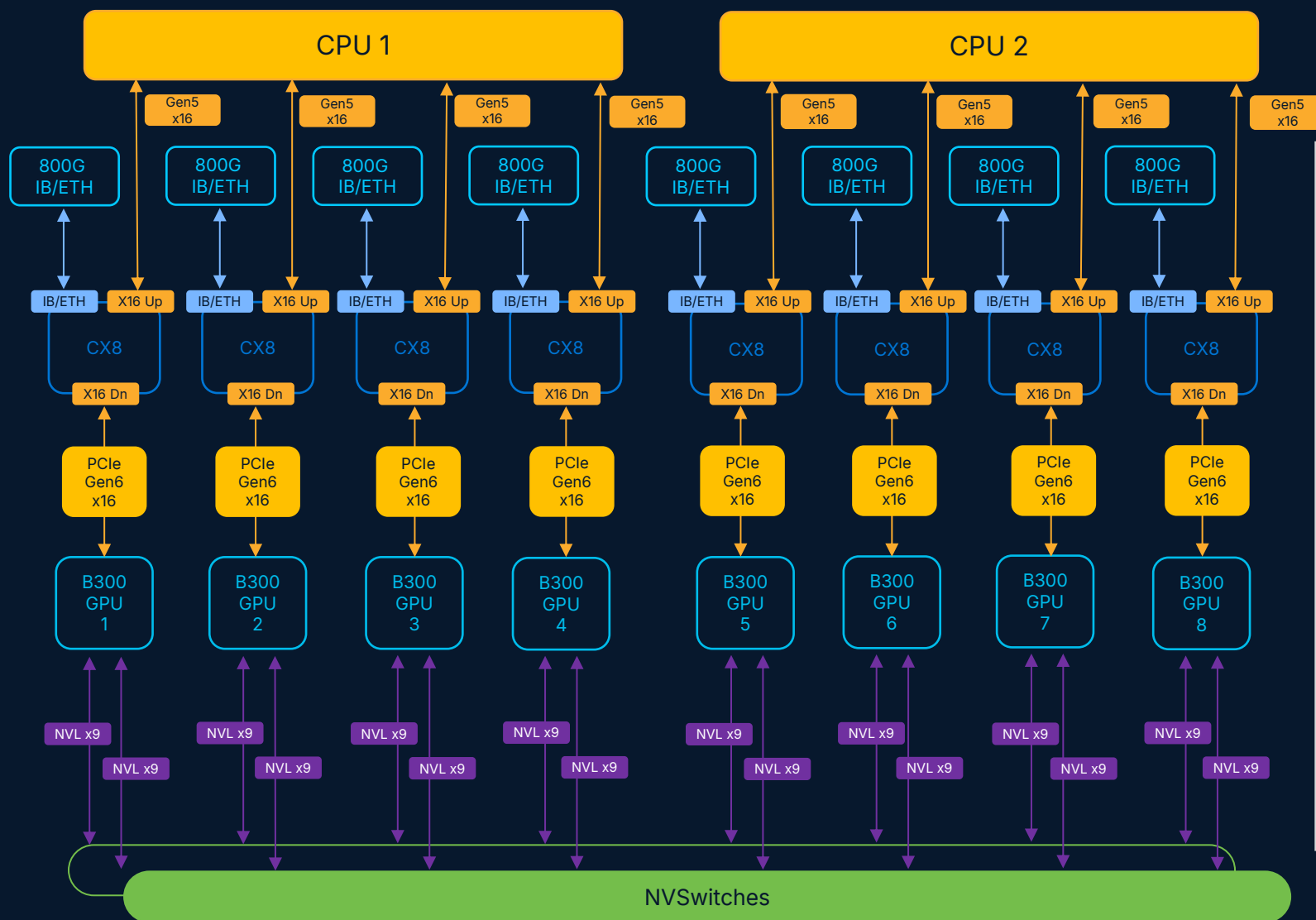
UCS C880A Rear View

PSUs

FANs

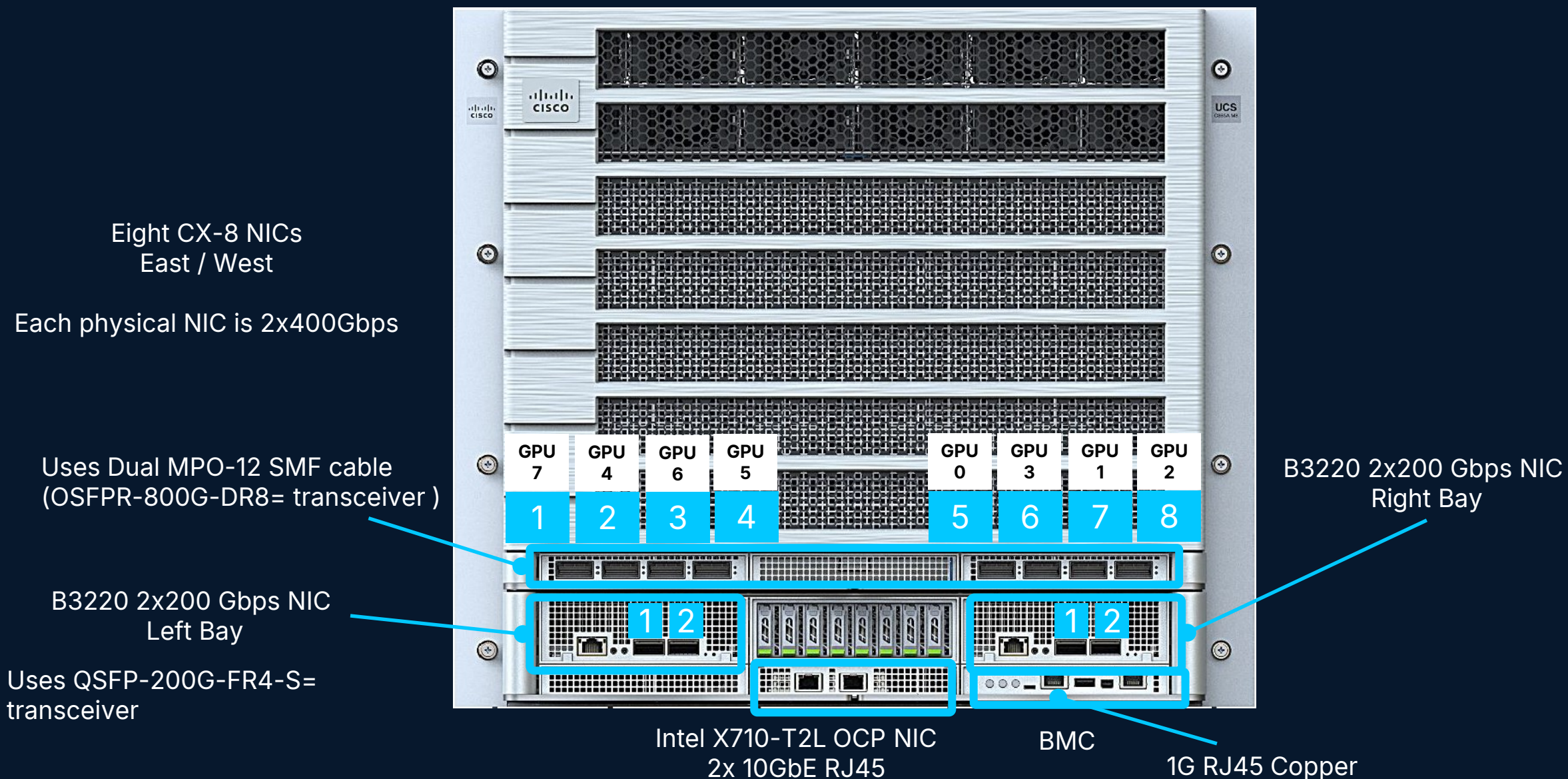


UCS C880A – GPU / E-W Connectivity

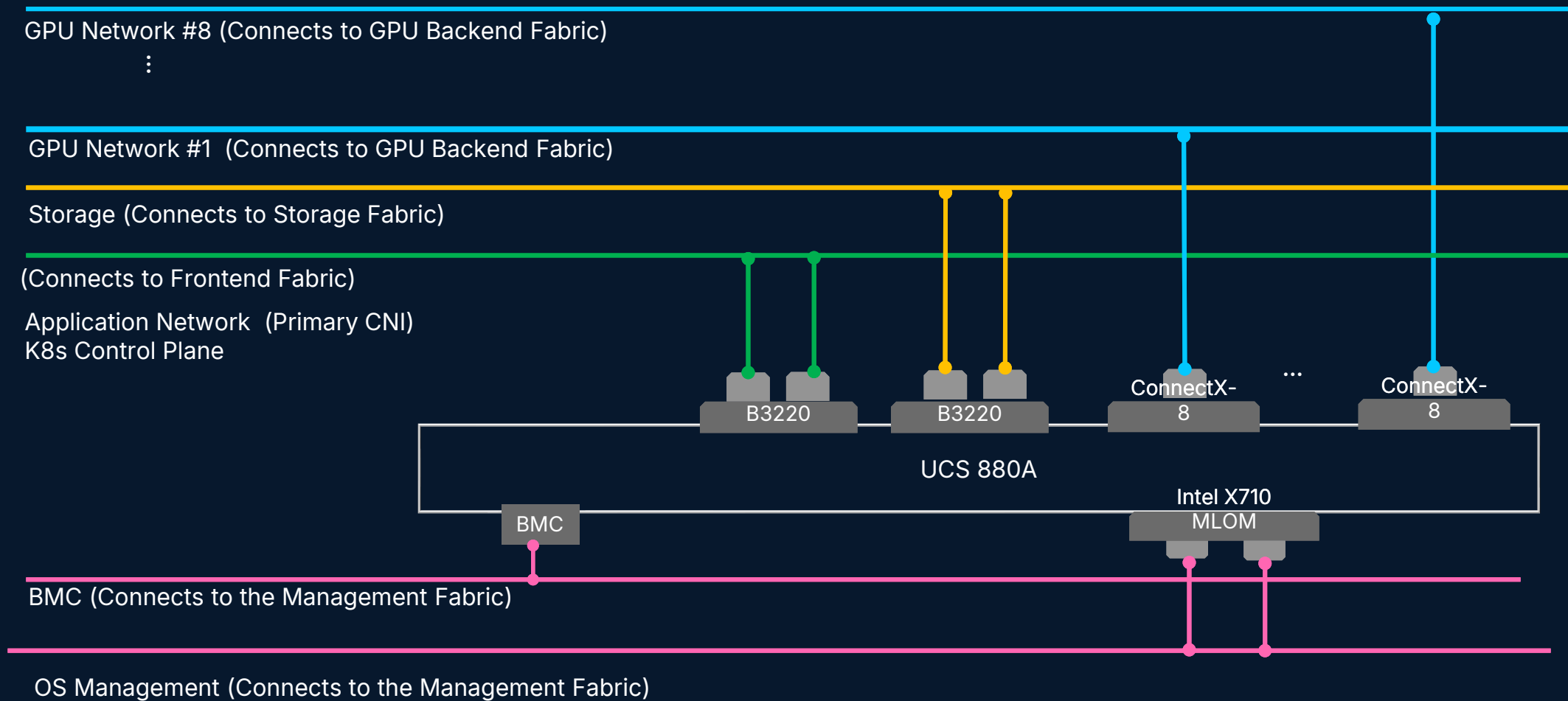


- 8x Nvidia B300 SXM GPUs
- Each B300 GPU has 18 physical NVLink connections to NVLink fabric (100 GB/s each)
- 2 x fully non-blocking NVLink switches connecting all 8 GPUs provide bidirectional bandwidth of 1800GB/s between any pair of Blackwell GPUs in the same node
- Each B300 GPU also has a dedicated CX8 NIC directly connected via Gen6 x16 PCI for GPU-to-GPU connectivity across nodes

UCS C880A M8 – Connectivity



UCS C880A NIC Mapping



UCS 880A GPU Worker Deployment Workflow

Hardware / BCM Baseline

- 1. Bring up BMC access for all UCS nodes**
Establish BMC/CIMC access for the UCS C2XX control plane/infra nodes and UCS 880A GPU worker nodes.
- 2. Validate UCS 880A hardware inventory through BMC**
Confirm CPU, memory, disks, GPUs, NICs/SuperNICs, power supplies, fans, and PCIe device visibility.
- 3. Baseline firmware, BIOS, NICs, GPUs, and boot settings**
Apply the required firmware/BIOS baseline, validate NIC/GPU firmware, confirm UEFI boot settings, and set the correct boot order.
- 4. Bring up the initial cluster using NVIDIA BCM for baseline testing**
Use NVIDIA Base Command Manager to perform initial provisioning, firmware validation, GPU visibility checks, and baseline performance testing.
- 5. Run initial GPU, NIC, and fabric validation**
Validate GPUs, high-speed NICs, RDMA capability, RoCEv2 fabric behavior, and initial NCCL performance while the cluster is under BCM control.

OpenShift Foundation

- 6. Wipe the UCS 880A worker nodes and prepare them for OpenShift**
Remove the BCM-managed host image/configuration from the GPU worker nodes before rebuilding them with RHCOS/OpenShift.
- 7. Install OpenShift Control Plane Node**
Deploy the OpenShift control plane nodes and establish the Kubernetes API, etcd, scheduler, controller manager, authentication, and cluster lifecycle services.
- 8. Validate OpenShift control plane health**
Confirm the OpenShift API, etcd quorum, DNS, ingress, monitoring, authentication, and cluster operators are healthy.
- 9. Install and join OpenShift infra nodes on UCS C2XX servers**
Join the infra nodes and place platform services such as ingress, registry, monitoring, logging, and management services as required.

GPU Worker Onboarding

- 10. Install and join UCS 880A GPU worker nodes**
Use BMC virtual media discovery ISO to install RHCOS on the UCS 880A nodes and join them to the OpenShift cluster as workers.
- 11. Approve CSRs and confirm GPU workers are Ready**
Approve the required certificate signing requests and verify that all UCS 880A worker nodes reach Ready state.
- 12. Blacklist the IRDMA drivers**
Blacklist the unwanted irdma drivers if required by the validated design to avoid driver conflicts with the NVIDIA/Mellanox RDMA stack.
- 13. Move UCS 880A nodes into a dedicated GPU worker role**
Apply labels and, if needed, taints to classify the UCS 880A nodes as dedicated GPU workers.

Discovery / Node Networking

14. Install the Node Feature Discovery Operator
Deploy NFD so OpenShift can discover and label hardware capabilities such as GPUs, PCIe devices, CPU features, and NIC capabilities.

15. Install the NMState Operator
Use NMState to declaratively configure node-level networking, including bonds, VLANs, MTU, static IPs, routes, and secondary network interfaces.

16. Install the NVIDIA GPU Operator
Deploy the NVIDIA driver, container toolkit, Kubernetes device plugin, GPU Feature Discovery, and DCGM exporter.

17. Install the NVIDIA Network Operator
Deploy and configure the NVIDIA networking/RDMA stack required for RoCEv2, GPUDirect RDMA, ConnectX/SuperNIC support, and high-performance AI networking.

AI Platform

18. Apply the NIC configuration template
Apply the required NIC configuration template so QoS, RDMA, RoCEv2, PFC/ECN-related behavior, and interface-level settings are correctly applied to the NICs.

19. Validate GPUs, NICs, RDMA, RoCEv2, and NCCL
Confirm GPU visibility, NIC visibility, RDMA devices, MTU, RoCEv2 behavior, QoS consistency, and NCCL performance across the UCS 880A GPU workers.

20. Configure NVIDIA AI Enterprise, NGC, and licensing
Configure NVAIE entitlement, NGC access, image pull secrets, licensing, and supported NVIDIA AI Enterprise container images.

21. Configure storage CSI integration
Install and validate the appropriate CSI driver and storage classes for VAST

22. Install Run:ai
Install Run:ai after the GPU and network stack has been validated. Configure departments, projects, quotas, node pools, queues, and GPU scheduling policies.

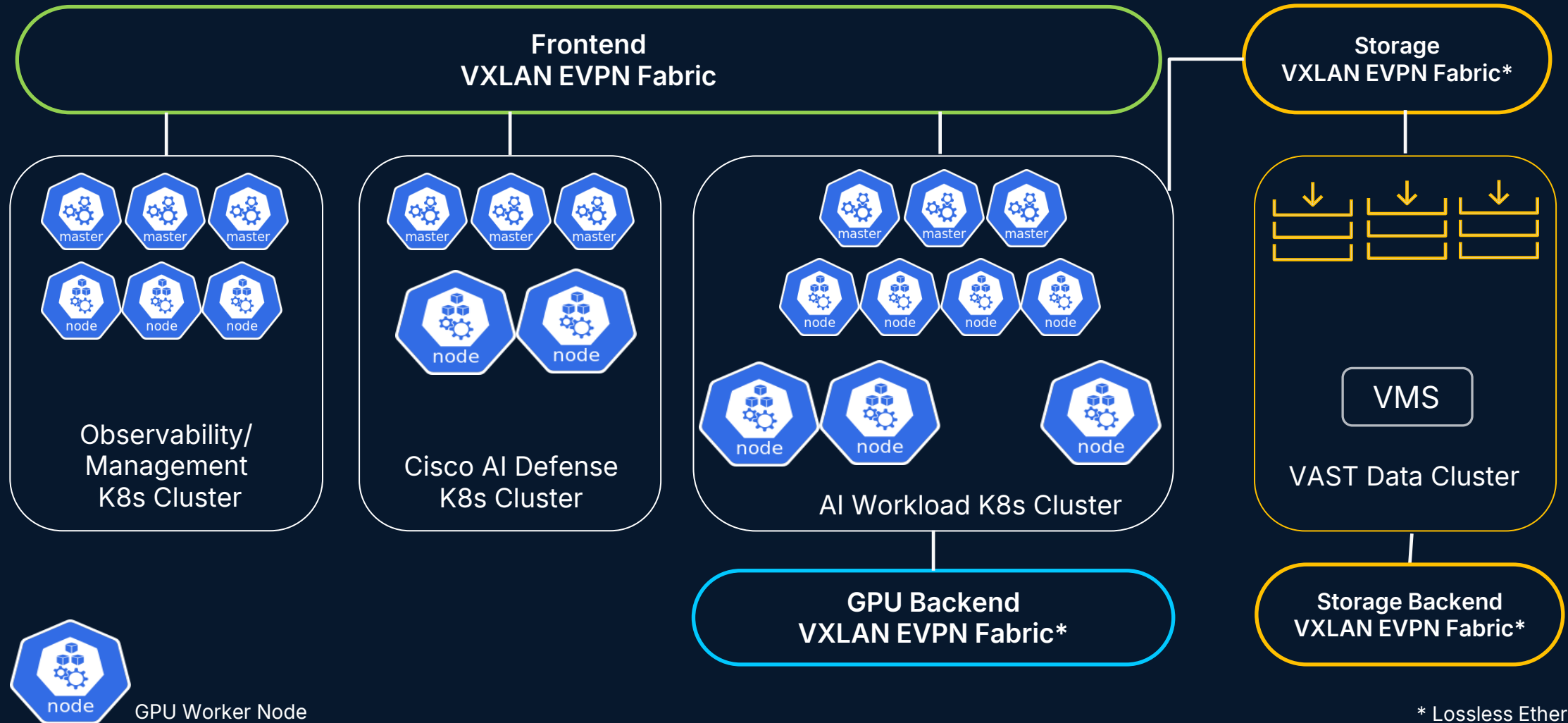
Storage / Ops / Validation

23. 24. Configure observability
Integrate Cisco Intersight, Nexus Dashboard, DCGM, OpenShift monitoring, storage telemetry, and Run:ai observability.

24. 25. Apply security controls
Configure RBAC, image policy, network policy, secrets management, audit logging, vulnerability scanning, and any required AI security controls.

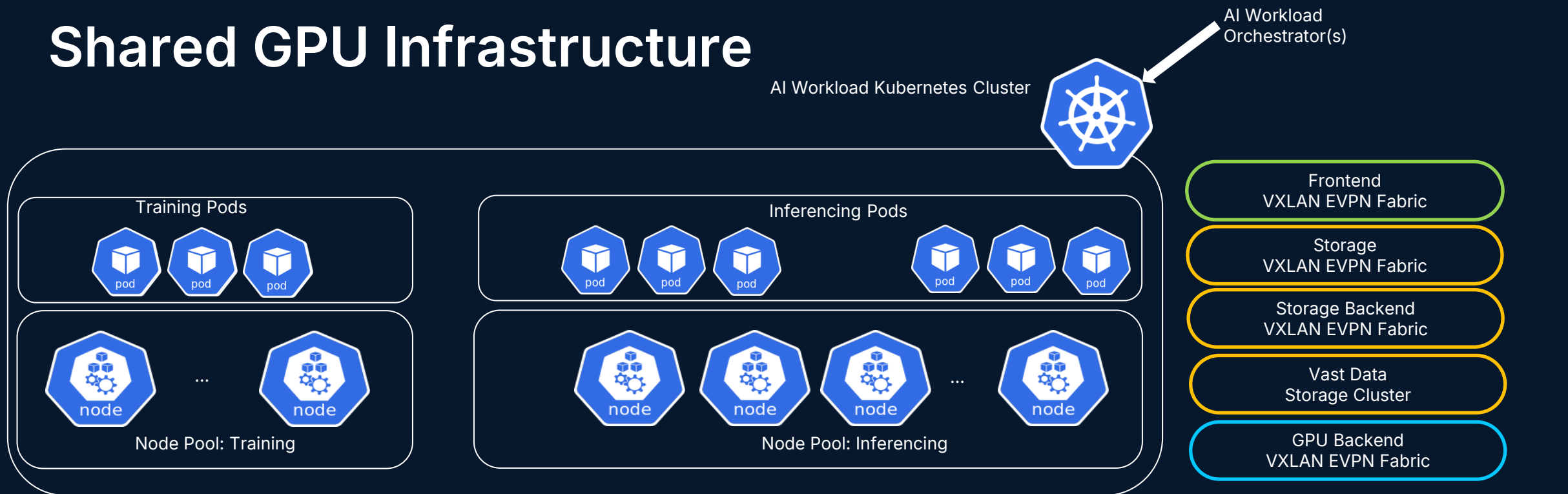
25. 26. Run end-to-end AI workload validation
Validate the complete stack using GPU workloads, multi-node NCCL tests, storage performance tests, Run:ai job scheduling, observability checks, and failure/recovery scenarios.

Shared GPU Infrastructure



- * Lossless Ethernet Fabric

Shared GPU Infrastructure



Training Jobs

Multi-GPU (Single Node or Multi Node)

Backend Networking required

Storage access via Worker Node (NFSoRDMA)

Inferencing Jobs

- Fractional or Full GPUs or MIG (Single Node or Multi Node)
- Backend Networking not required
- Distributed Inferencing Jobs
- Multi-GPU (Single Node or Multi Node)
- Backend Networking required
- Storage access via Worker Node (NFSoRDMA)

Segmentation Controls

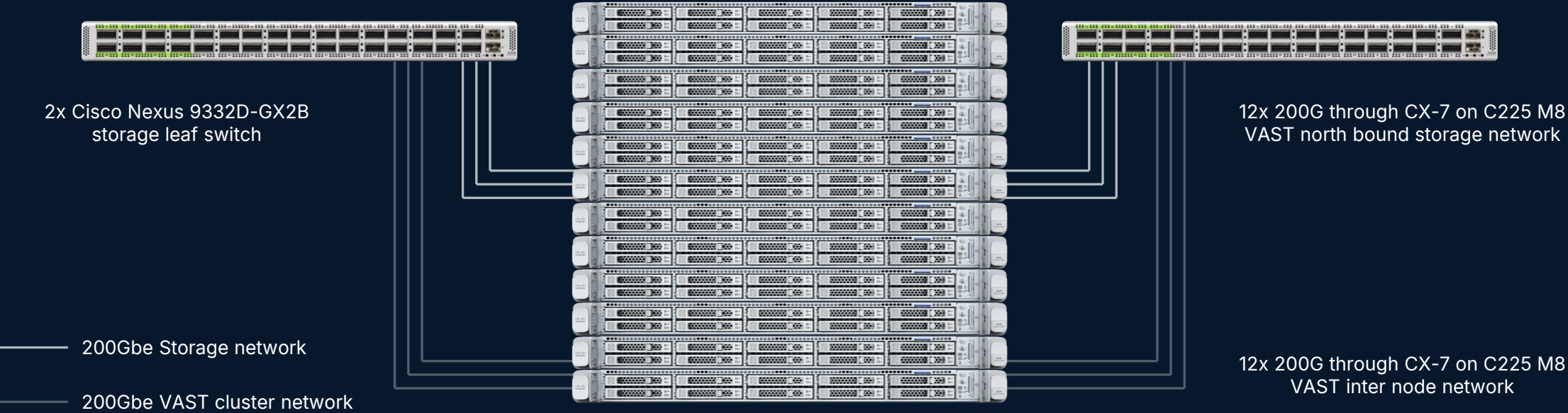
- VXLAN Overlay Networks
- VXLAN GPO SGACL
- Kubernetes Network Policy
- Kubernetes Multi Network Policy
- VAST Client Access Control

Storage Network Function

- Connects GPU nodes to shared storage platforms for high-throughput dataset and model access.
- Supports dataset reads during training, including samples, metadata, and preprocessing outputs.
- Supports checkpoint writes for model weights, optimizer state, intermediate checkpoints, and final artifacts.
- Provides storage access for inference and RAG workloads, including models, embeddings, indexes, and source documents.
- May include multiple tiers, with Tier 1 optimized for distributed training/checkpointing and Tier 2 used for general file storage and less latency-sensitive workloads.
- Storage access may use TCP-based protocols such as NFS over TCP, or RDMA/RoCE depending on the storage platform.
- Lossy versus lossless fabric requirements depend on the storage protocol and vendor design.
- If RDMA/RoCE is used for storage, the same core RoCE concepts apply: PFC, ECN, QoS, MTU consistency, congestion management, and telemetry.

VAST Data with Cisco UCS Ebox

Solution overview



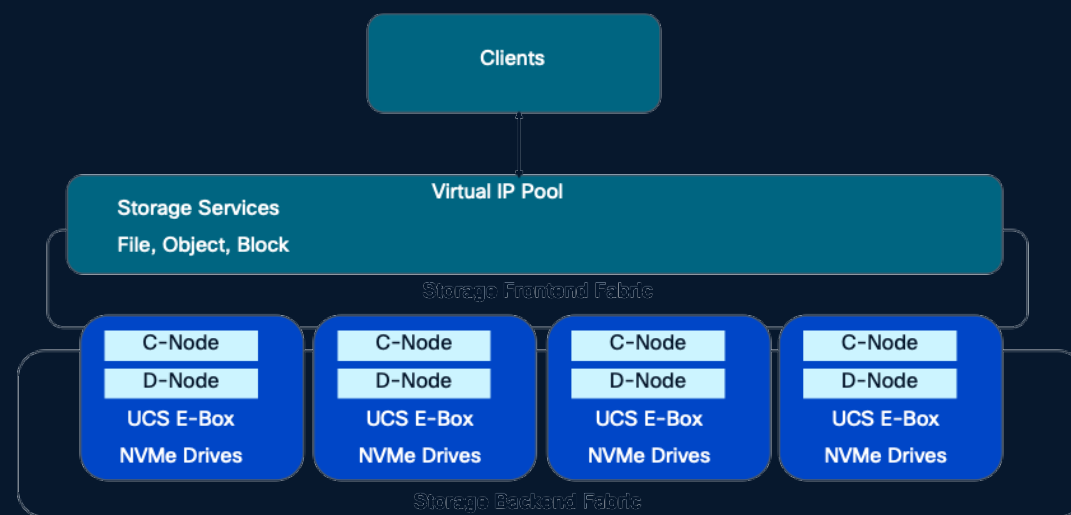
12x UCS C225 M8

Cluster Specification - 12 x VAST Ebox	
Useable Capacity	1.1 PB @ 1:1 Data Reduction
Effective Useable Capacity	2.19 PB @ 2:1 Data Reduction
Servers	12 Ebox
Rack Units	18U
Read Bandwidth	192 GB/S
Write Bandwidth	120 GB/S
READ IOPS	1,920,000
WRITE IOS	264,000

VAST Data

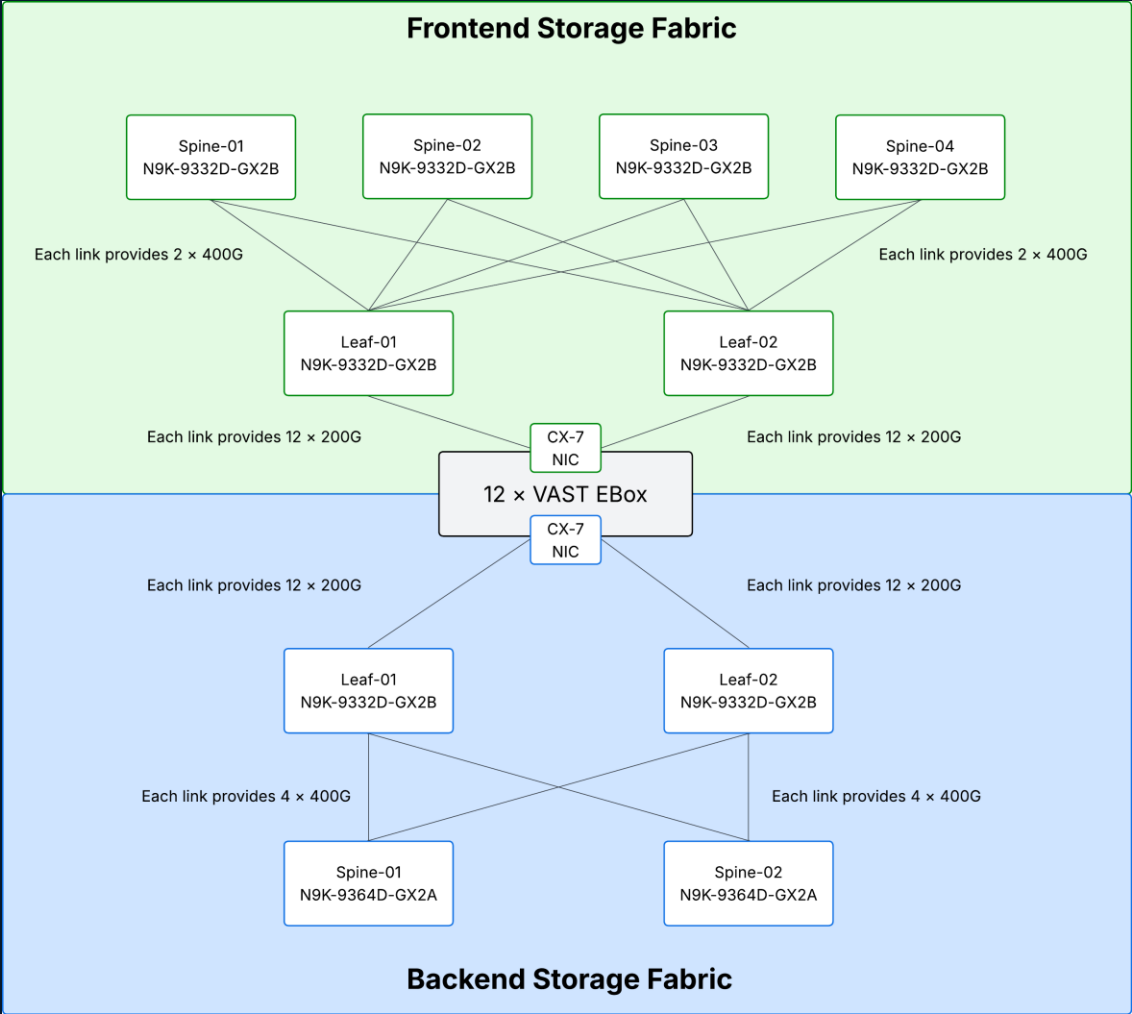
Overview

- Unified AI storage layer for training, fine-tuning, inference, RAG, and analytics.
- Built on VAST DASE architecture, separating storage services from persistent NVMe media.
- C-Nodes handle stateless compute, protocol services, metadata, and client I/O.
- D-Nodes provide persistent NVMe flash capacity for data and metadata.
- Cisco UCS EBox uses a combined C-Node/D-Node model in a single UCS-based appliance.
- Provides global namespace consistency and parallel access across GPU workers.
- Integrates with OpenShift through CSI drivers, dynamic provisioning, PVs, and PVCs.
- Supports NFS, S3-compatible object access, and NVMe/TCP block access.
- Supports GPUDirect Storage / RDMA-oriented data paths where validated, reducing CPU and host-memory involvement in storage-to-GPU data movement.
- RDMA-based access requires validated RoCEv2 design with QoS, PFC, ECN/CNP, MTU consistency, and congestion visibility.



VAST Data Storage Fabric Connectivity

Fabric Connectivity



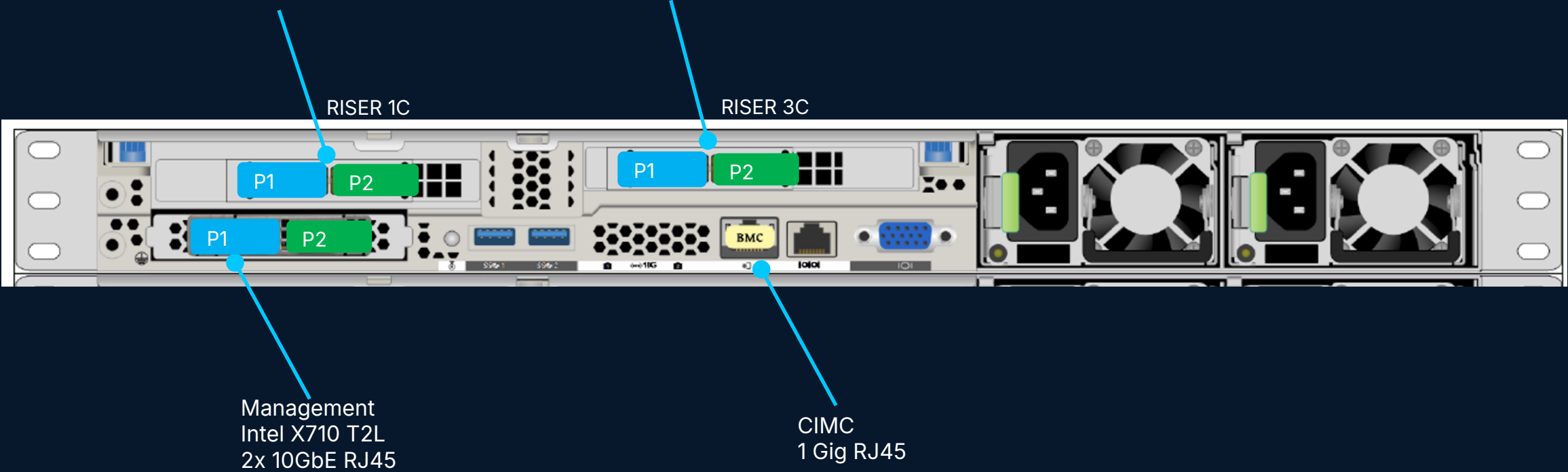
VAST Ebox – UCS C225 M8

Uses QSFP-200G-FR4-S= transceiver

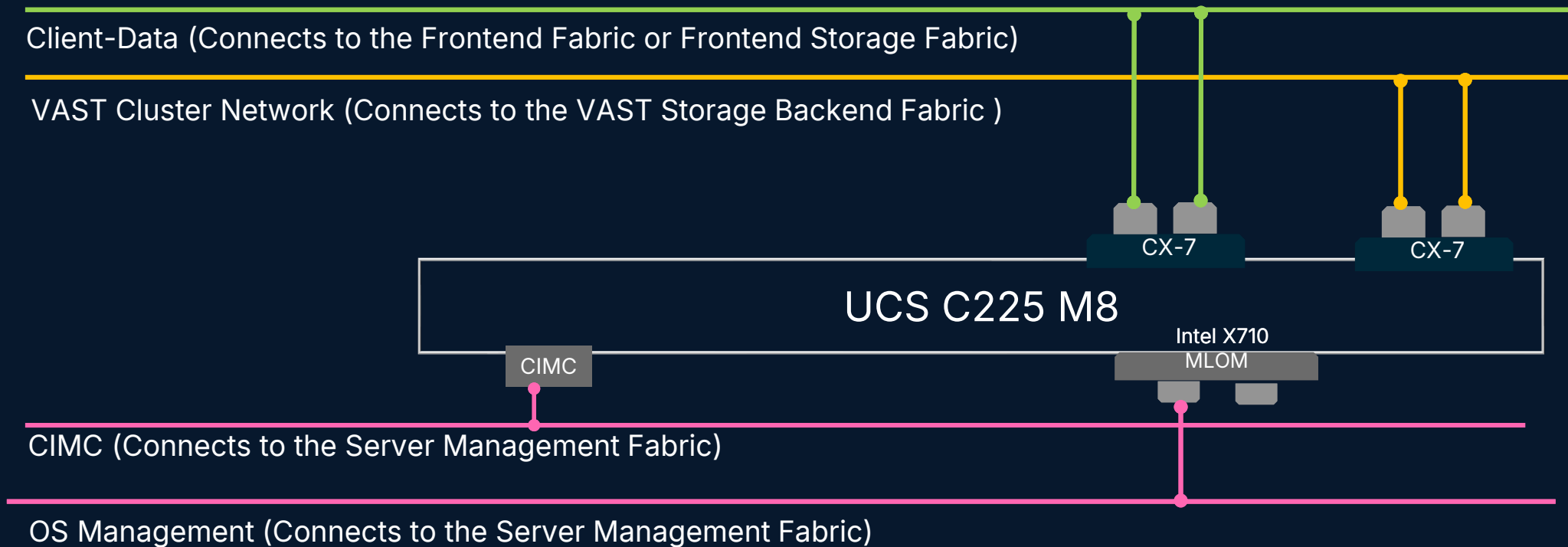
Storage (Client-Data) Traffic
ConnectX-7 2x200Gbps

Uses QSFP-200G-FR4-S= transceiver

Cluster Traffic
ConnectX-7 2x200 Gbps



VAST Ebox – UCS C225 M8 – NIC Mapping



Lessons Learned – Intel RDMA Drivers

- **Blacklist or Uninstall Unused Drivers:** Disable any unnecessary kernel modules that are not actively required on your nodes.
- **Resolve Multi-Adapter Conflicts:** Systems containing both Intel and NVIDIA RDMA-capable adapters will experience driver collisions if both stacks attempt to load simultaneously.
- **Prevent Discovery and Validation Failures:** Active Intel drivers (irdma) will interfere with NVIDIA MOFED device discovery, disrupting network interface behavior and complicating troubleshooting.
- The **NVIDIA Network Operator** automates Kubernetes management for ConnectX network cards. It installs **MOFED** (Mellanox Open Fabrics Enterprise Distribution), the driver stack required for high-performance networking and RDMA.
- **Mandatory Action for Containerized MOFED:** You must completely blacklist the Intel irdma driver on all worker nodes to allow the Network Operator to successfully load the containerized MOFED stack without kernel conflicts.

Lessons Learned - Maintain Backdoor Node Access

- Every AI node should have a reliable out-of-band or management-network access path that remains available even when the high-speed data, storage, or GPU fabric is down
- AI infrastructure troubleshooting often involves disruptive changes: NIC firmware upgrades, DPU mode changes, driver reloads, OS reboots, OpenShift node rebuilds, kernel updates, and storage/network reconfiguration.
- If the only usable path to the node depends on the same high-speed network being modified, recovery becomes much harder
- Access through CIMC, Redfish, KVM console, SSH over management, or a dedicated management VLAN should be validated before production cutover.

Lessons Learned – Bluefield NIC DPU Mode

Why you may disable DPU mode:

- The deployment only requires standard high-speed Ethernet NIC functionality
- The software stack, OS image, or OpenShift deployment model is not using DPU-hosted services
- DPU mode introduces additional dependencies around firmware, embedded OS, provisioning, security policy, and lifecycle management.

Why you may need DPU mode:

- You plan to offload infrastructure services from the host CPU.
- You want to run security, firewalling, telemetry, service chaining, or storage/network acceleration on the DPU
- You are using a design that depends on BlueField SNAP, DOCA-based services, GPUDirect-related acceleration patterns, or infrastructure offload

Lessons Learned - Storage Design

- Use Both Ports from the Same NIC for Storage Multipathing
- For the storage front-end network, storage multipathing should be designed using both ports from the **same physical NIC**, not one port from each separate NIC
- Each server has multiple high-speed NICs, but the storage path design needs to account for how the NICs, PCIe topology, NUMA locality, and operating system pathing behaved together
- Using ports from the same NIC provides a more predictable and supportable storage multipathing design
- In AI environments, storage path design directly impacts data loading, checkpointing, and GPU utilization, so NIC placement and pathing must be validated early

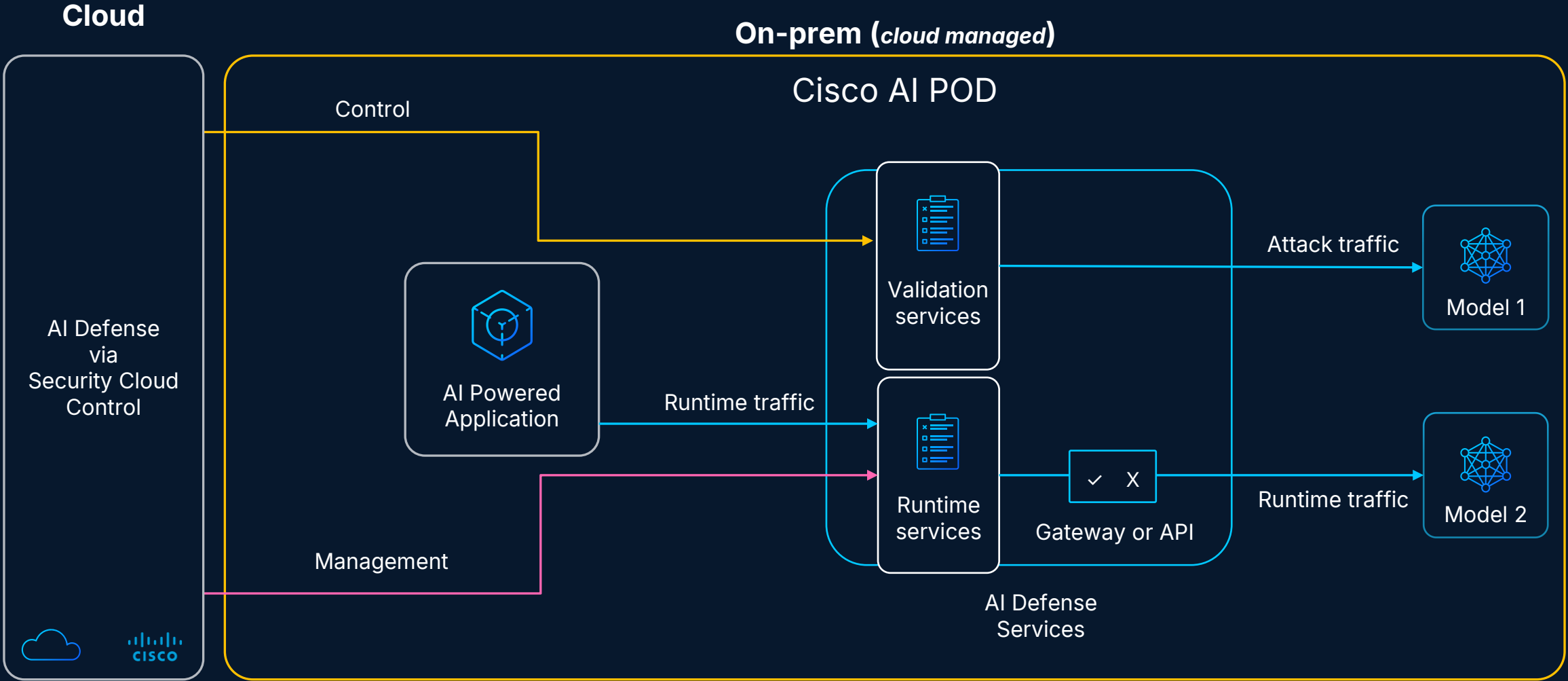
Lessons Learned – RoCE/RDMA Is Required for High-Performance AI Storage

- AI storage front-end networks require RoCE/RDMA because GPU workloads are highly sensitive to storage throughput, latency, and CPU overhead
- Poor storage performance can appear as a GPU utilization issue, even when the root cause is actually storage or network related
- NIC settings, driver versions, firmware versions, switch QoS policy, and storage platform guidance must be aligned
- Lossless configuration mistakes may not show up immediately; they often appear underload during storage benchmarking or AI job execution
- PFC and ECN counters should be monitored continuously.

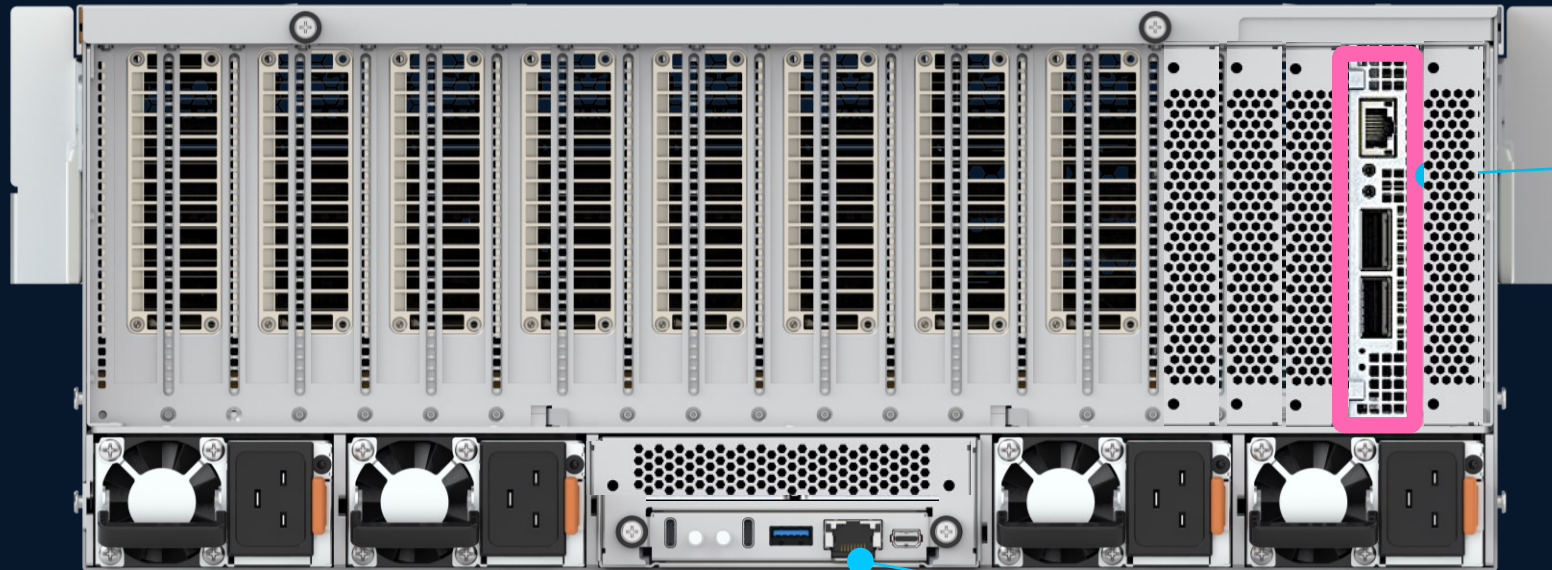


AI Security

AI Defense – AI PODs Architecture



UCS 845A AI Defense Worker Node



CX-7 2x200
for north-south front-end traffic
QSFP-200G-FR4-S=

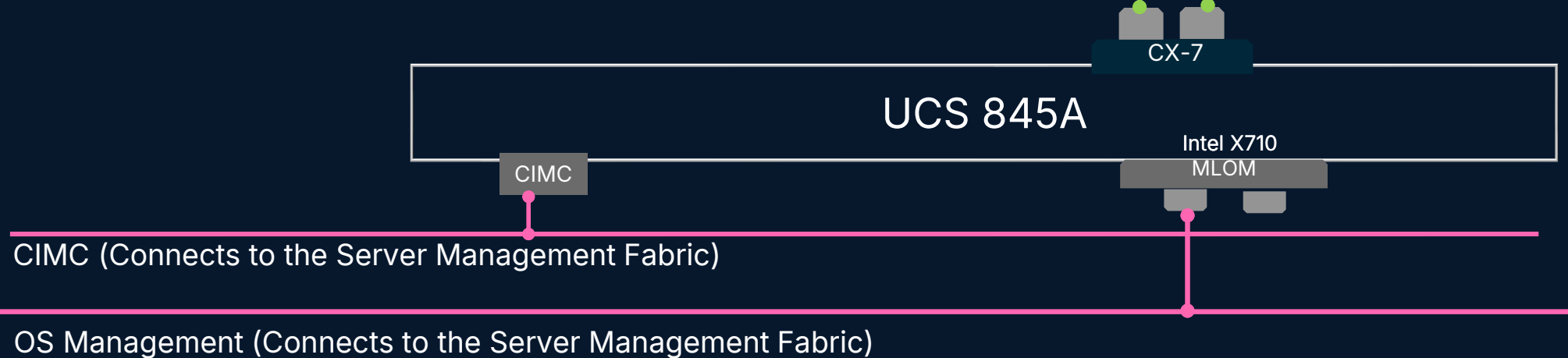
BMC RJ45 Copper

CX-7 ports are numbered Top to Down as seen in the rear view

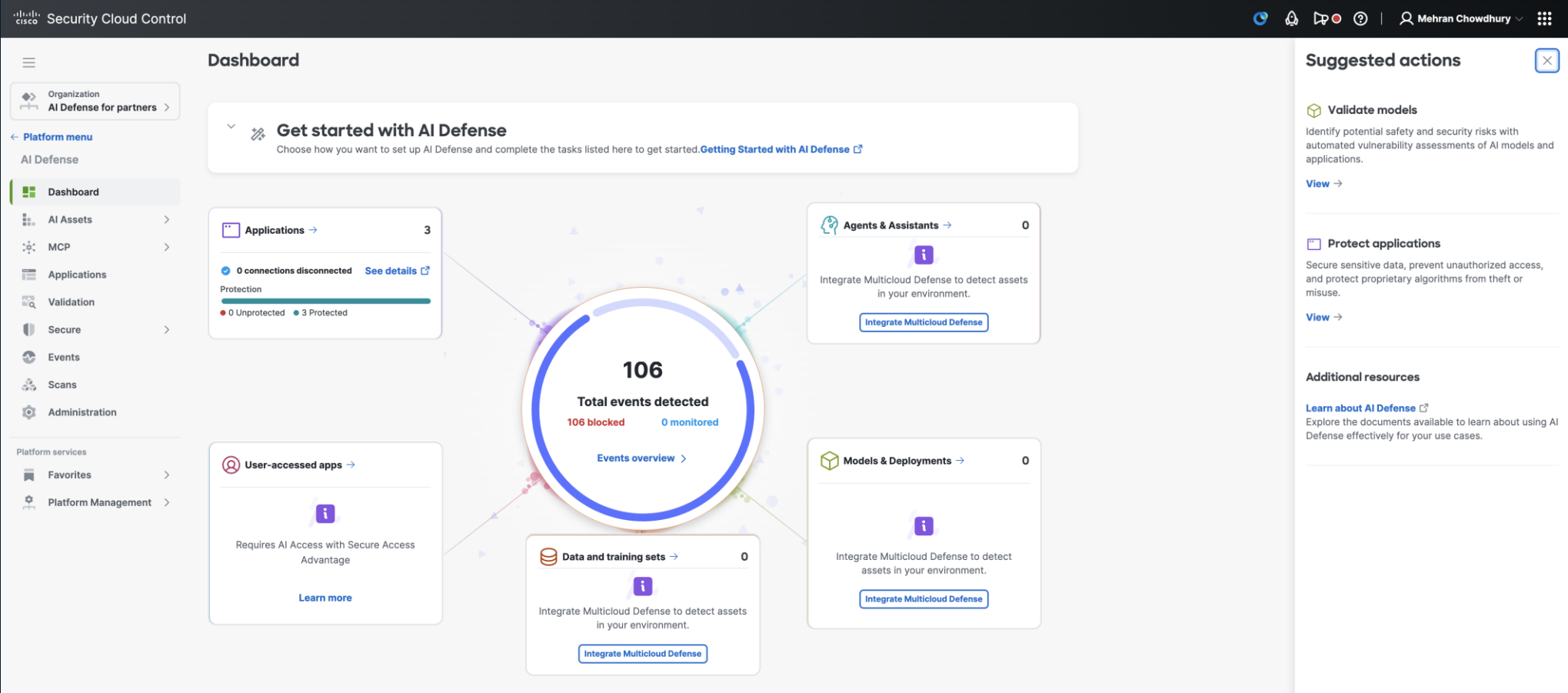
UCS C845A AI Defense Worker Node NIC Mapping

Openshift Control Plane and Application Traffic (Connects to Frontend Fabric)

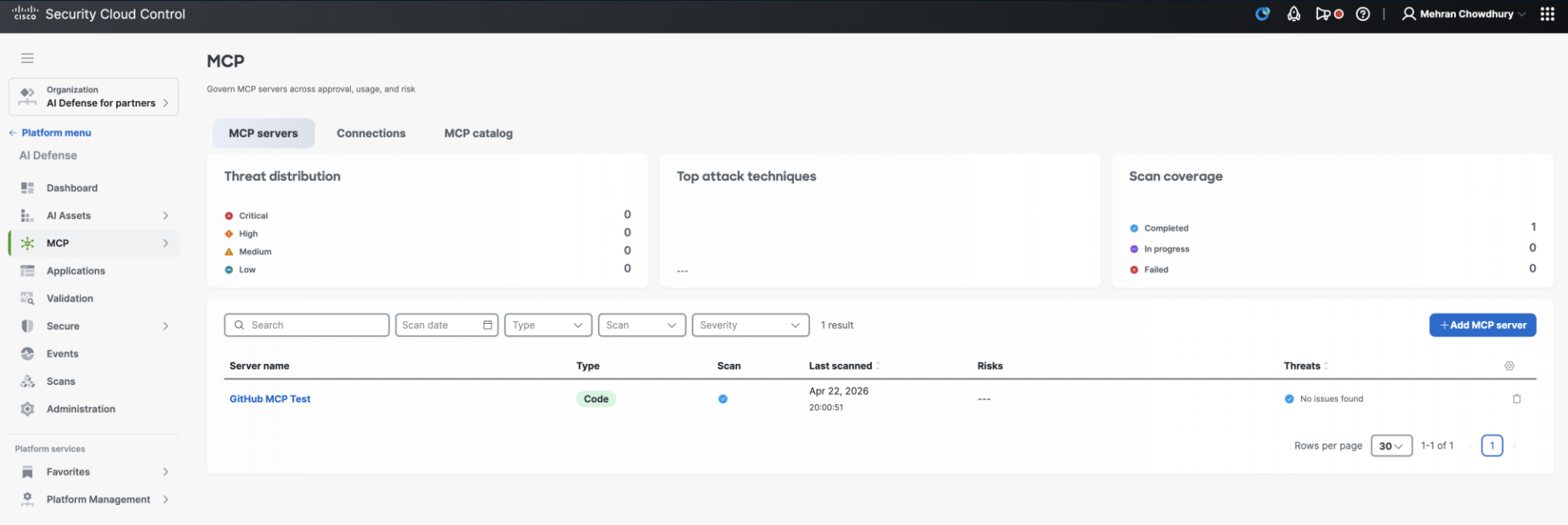
Application Network (Primary CNI)



AI Defense - Dashboard



AI Defense – MCP Servers



AI Defense - Applications

Security Cloud Control

Organization

AI Defense for partners

Platform menu

AI Defense

Dashboard

AI Assets

MCP

Applications

Validation

Secure

Events

Scans

Administration

Platform services

Favorites

Platform Management

Applications

The Applications dashboard lists all the generative AI applications and their associated connections protected by AI Defense.

Applications

Connections

Q Name

Q Description

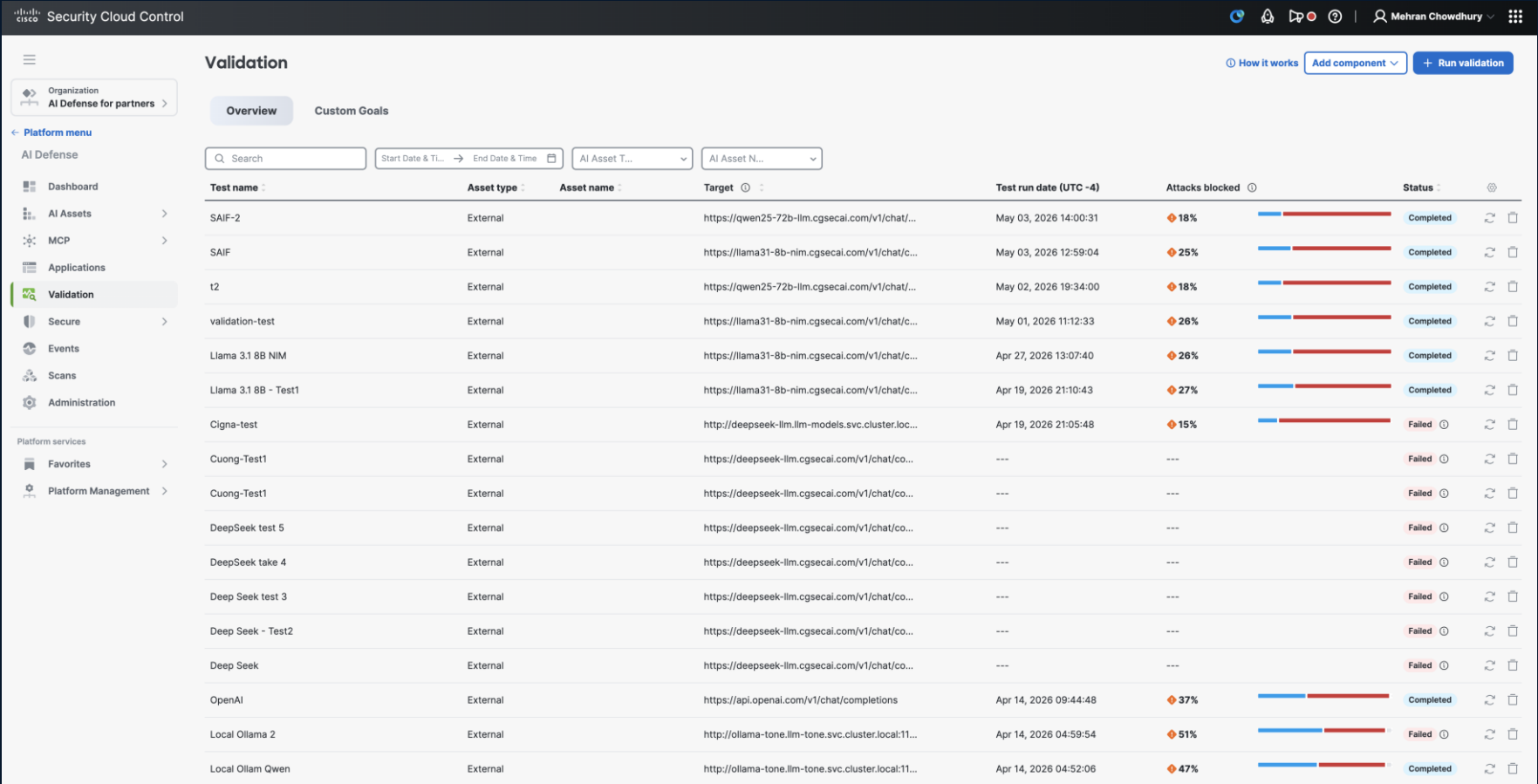
Connection type

Q Connection name

3 results

Name	Description	Connection type	Last modified (UTC -4)	
Cilium-chatbot		Gateway	Apr 20, 2026 11:53:24 Isaiah Laughlin	
MyChatApp-API		API	Apr 13, 2026 07:31:02 Tilen Miakar	
MyChatApp-Gateway		Gateway	Apr 13, 2026 07:31:50 Tilen Miakar	

AI Defense - Validation



AI Defense - Policies

Organization

AI Defense for partners

Platform menu

AI Defense

Dashboard

AI Assets

MCP

Applications

Validation

Secure

Policies

The Policies dashboard lists all the policies and the connections to which they are applied.[Learn more about policies.](#)

Search by policy name

Language config

Connection type

Status

7 results

Policy name	Language config	Connection type	Connections applied	Last modified (UTC -4)	Status	
greg	English	Gateway	---	May 01, 2026 13:56:55 Isaiah Laughlin	Draft	
Safety-Policy-API	English	API	1 connection	Apr 13, 2026 11:17:53 Tilen Mlakar	Enabled	
My-App-Policy	English	Unspecified	---	Apr 06, 2026 10:25:51 Cuong Tran	Draft	
MCP-Security-Policy	English	MCP API	---	Apr 15, 2026 08:49:59 Tilen Mlakar	Draft	
Implement-Safety-Policy	English	Gateway	1 connection	Apr 21, 2026 12:07:22 Tilen Mlakar	Enabled	

+ Create policy

AI Defense - Events

Organization
AI Defense for partners

Platform menu

AI Defense

- Dashboard
- AI Assets
- MCP
- Applications
- Validation
- Secure
- Events
- Scans
- Administration

Platform services

- Favorites
- Platform Management

Events

Event logs

Start date and ... → End date and ti... Rule action Message type Applications Model Policy Filters 106 results

Application	Rule action	Message type	Enforcement point	Policy applied	Match	Event timestamp (UTC -4)
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security Safety	Apr 22, 2026 02:58:33
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security Safety	Apr 22, 2026 02:58:30
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Safety	Apr 22, 2026 02:58:27
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security Safety	Apr 21, 2026 12:05:07
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security Safety	Apr 21, 2026 12:05:04
MyChatApp-Gateway gpt-4o-mini	Block	Response	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security	Apr 21, 2026 12:01:36
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security Safety	Apr 21, 2026 05:28:17
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security Safety	Apr 21, 2026 05:28:15
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security Safety	Apr 21, 2026 04:46:48
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Safety	Apr 21, 2026 04:46:45
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security Safety	Apr 21, 2026 04:43:16
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Safety	Apr 21, 2026 04:43:13
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security Safety	Apr 15, 2026 05:14:36
MyChatApp-Gateway gpt-4o-mini	Block	Prompt	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Safety	Apr 15, 2026 05:14:33
MyChatApp-Gateway gpt-4o-mini	Block	Response	AI-Defense - SaaS Gateway	Implement-Safety-Policy	Security	Apr 15, 2026 05:14:28
MyChatApp-API	Block	Response	AI-Defense - SaaS API	Safety-Policy-API	Security	Apr 13, 2026 11:24:40
MyChatApp-API	Block	Response	AI-Defense - SaaS API	Safety-Policy-API	Security	Apr 13, 2026 11:24:25

AI Defense – Model Scans

Security Cloud Control

Organization

AI Defense for partners

Platform menu

AI Defense
Dashboard
AI Assets
MCP
Applications
Validation
Secure
Events
Scans
Administration
Platform services
Favorites
Platform Management

Model scans Beta

Scan your individual model files or model repositories for known security vulnerabilities and to identify harmful files and scripts in model inputs.

7 Total scans

0 Completed scans in the last 7 days

1 Critical vulnerabilities

7 results

Name	Scan date (UTC)	Type	Files	Issues by severity	Status
sharktide/recyclebot0	Apr 14, 2026 08:46:45 Tilen Miakar	Repository	8	1 High	Completed
macedonizer/mk-roberta-base	Apr 14, 2026 08:45:55 Tilen Miakar	Repository	13	No issues found	Completed
mscgo77/pickle-picklescan-bypass-poc	Apr 14, 2026 08:36:31 Tilen Miakar	Repository	5	5 High 12 Low	Completed
optimus-fulcra/modelscan-pickle-bypass-poc	Apr 14, 2026 08:36:01 Tilen Miakar	Repository	7	1 Critical 3 High 5 Low 1 Medium	Completed
mscgo77/pickle-picklescan-bypass-poc	Apr 14, 2026 08:34:43 Tilen Miakar	Repository	5	5 High 12 Low	Completed
bombman/HunyuanOCR-GGUF	Apr 14, 2026 08:15:04 Tilen Miakar	Repository	6	No issues found	Completed
Qwen/Qwen3.5-9B	Apr 14, 2026 08:07:39 Tilen Miakar	Repository	16	No issues found	Completed



Observability and Day-2 Ops

The image features a dark blue background with a large, stylized circular graphic on the left side. This graphic has a gradient, transitioning from a deep purple at the top to a bright blue at the bottom, with a white outline. Centered on the right side of the image is the text "Nexus Dashboard" in a white, bold, sans-serif font.

Nexus Dashboard

Cisco Nexus Dashboard

Advanced Congestion detection

Nexus Dashboard provides critical insights into network health, detecting congestion even without Dynamic Load Balancing (DLB) enabled. This deep visibility ensures optimal performance for your AI workloads.



PFC Pattern Analysis

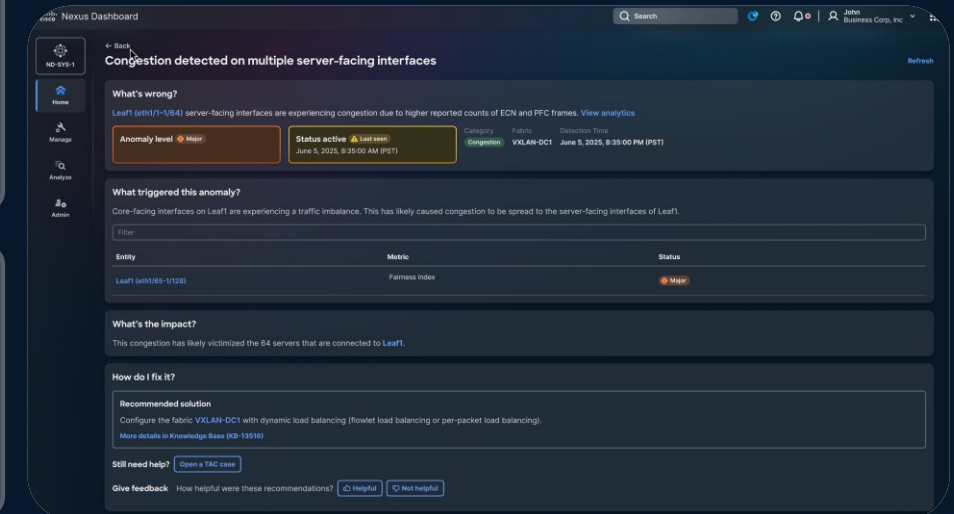
Detects congestion through detailed Priority Flow Control (PFC) patterns.

ECN Packet Monitoring

Analyzes Explicit Congestion Notification (ECN) packets for early warnings.

Hierarchical Scoring

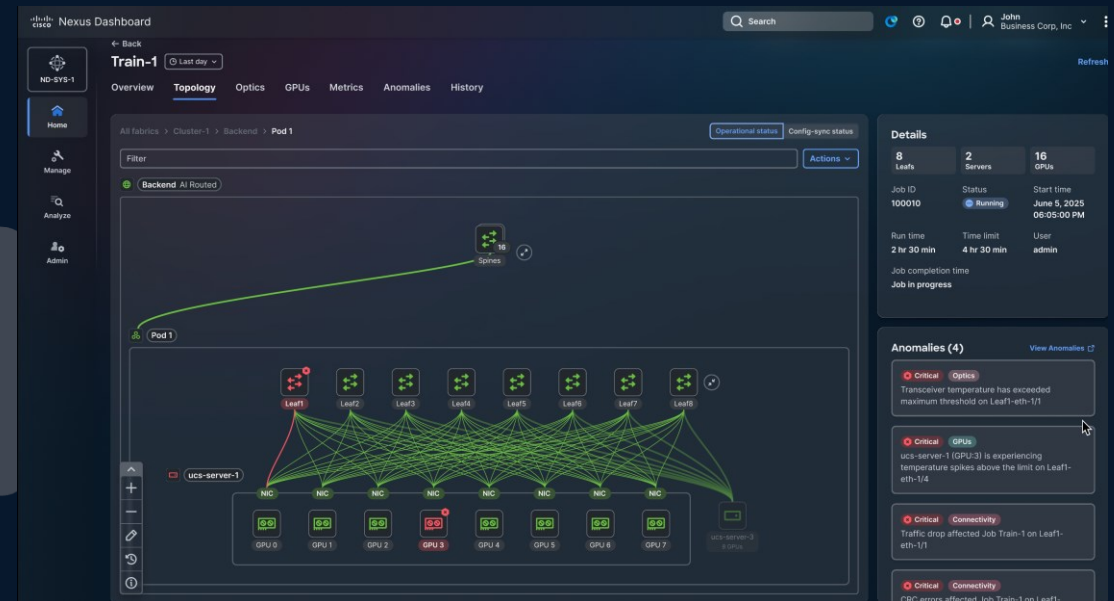
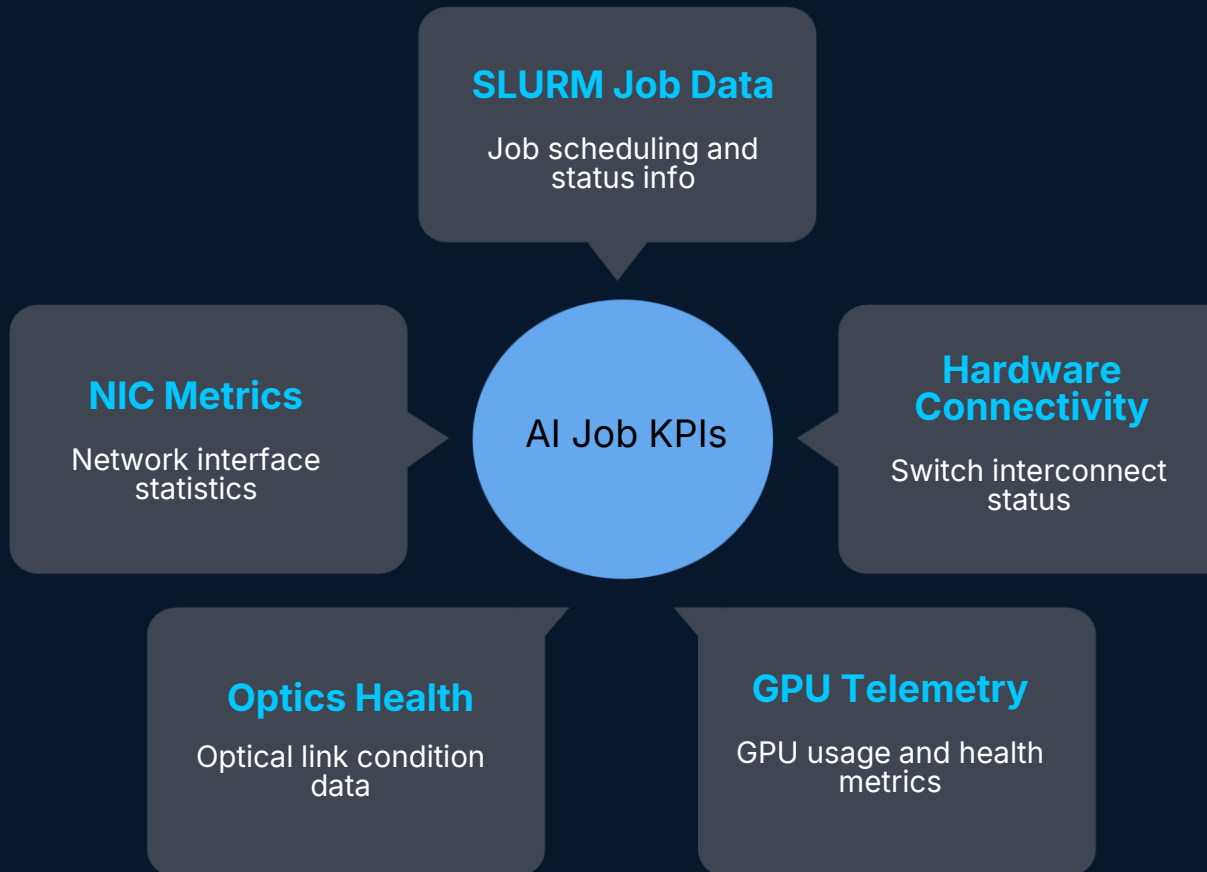
Applies a hierarchical congestion score from interface to switch to the entire fabric.



Cisco Nexus Dashboard

Integration with SLURM for AI Job monitoring, Job-Centric Correlation for AI Fabrics

Optimizing performance for complex AI workloads requires a comprehensive understanding of interactions across the entire fabric, achieved by correlating diverse data streams.



Nexus Dashboard 4.2 – GPU Inventory

ND-SYS-1

Home

Manage

Analyze

Admin

Nexus Dashboard

Search

John Business Corp, Inc

Inventory

Switches

AI resources

Clusters 1

Scalable units 4

Servers 256

GPUs 2,048

Jobs 12

12 Critical

4 Major

4 Minor

0 Warning

2028 Healthy

Filter - Note: click on the "3" to see GPU details :-)

List view

Health view

Cluster 1 > Backend 6 critical

Scalable unit 1

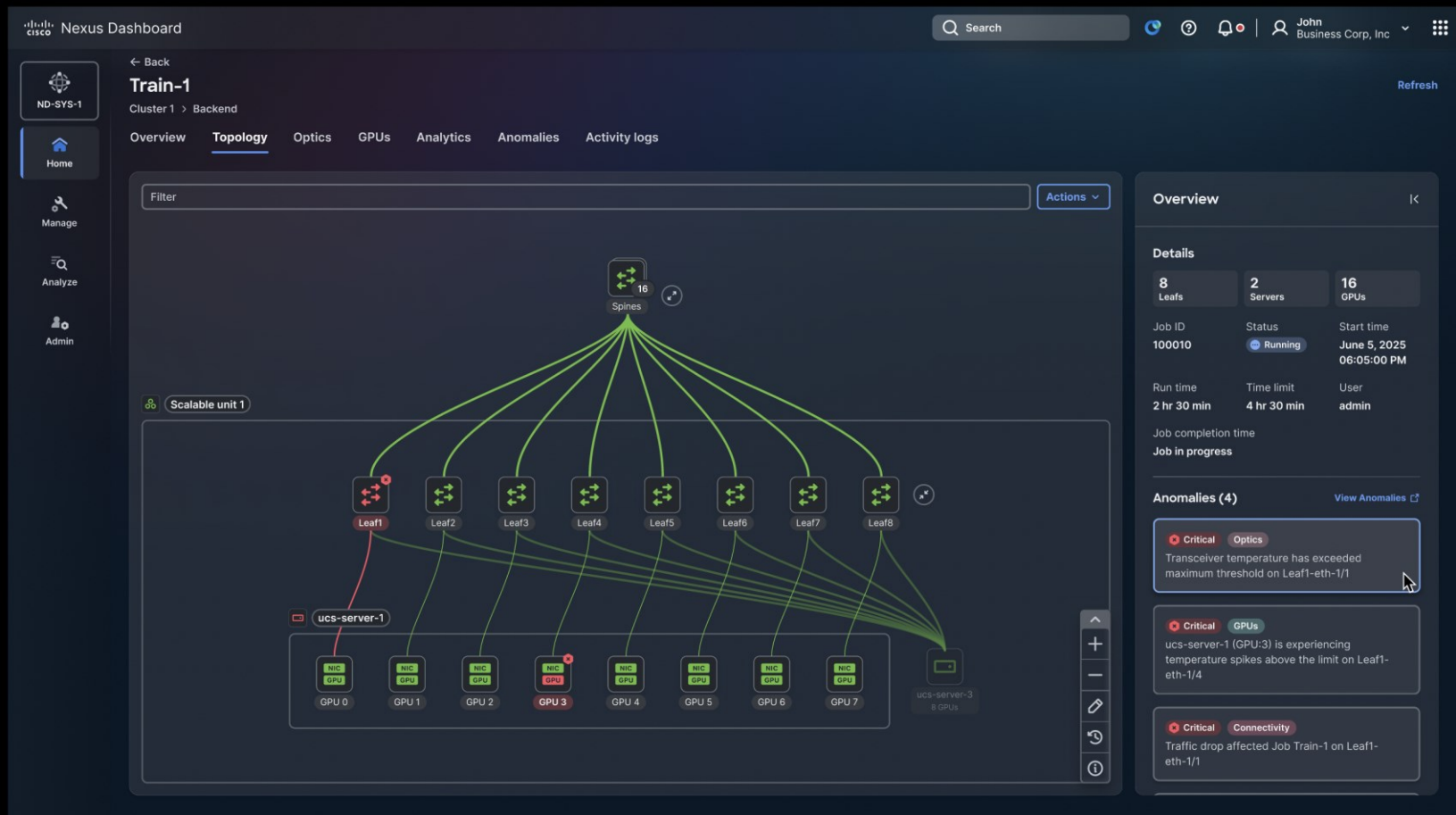
Anomaly level	GPU ID	Servers	Vendor	Model	Switch-interface
Critical	3	ucs-server-1	Nvidia	A100	Leaf1-eth1/4
Healthy	0	ucs-server-1	Nvidia	A100	Leaf1-eth1/1
Healthy	1	ucs-server-1	Nvidia	A100	Leaf1-eth1/2
Healthy	2	ucs-server-1	Nvidia	A100	Leaf1-eth1/3
Healthy	4	ucs-server-1	Nvidia	A100	Leaf1-eth1/5
Healthy	5	ucs-server-1	Nvidia	A100	Leaf1-eth1/6
Healthy	6	ucs-server-1	Nvidia	A100	Leaf1-eth1/7
Healthy	7	ucs-server-1	Nvidia	A100	Leaf1-eth1/8
Healthy	0	ucs-server-3	Nvidia	A100	Leaf1-eth1/9
Healthy	1	ucs-server-3	Nvidia	A100	Leaf1-eth1/10

Rows per page 10

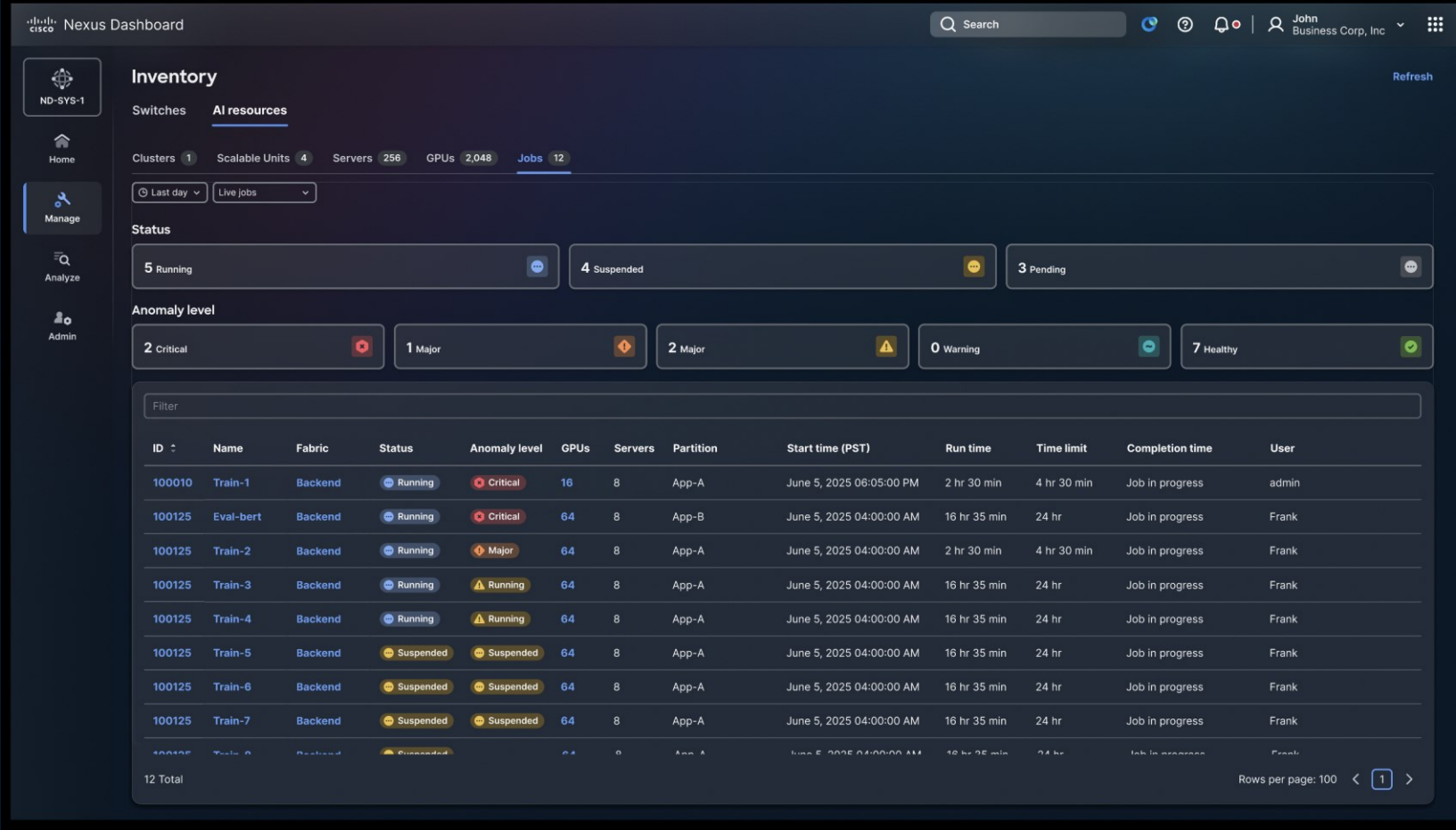
1-10 of 512

1 2 ... 52

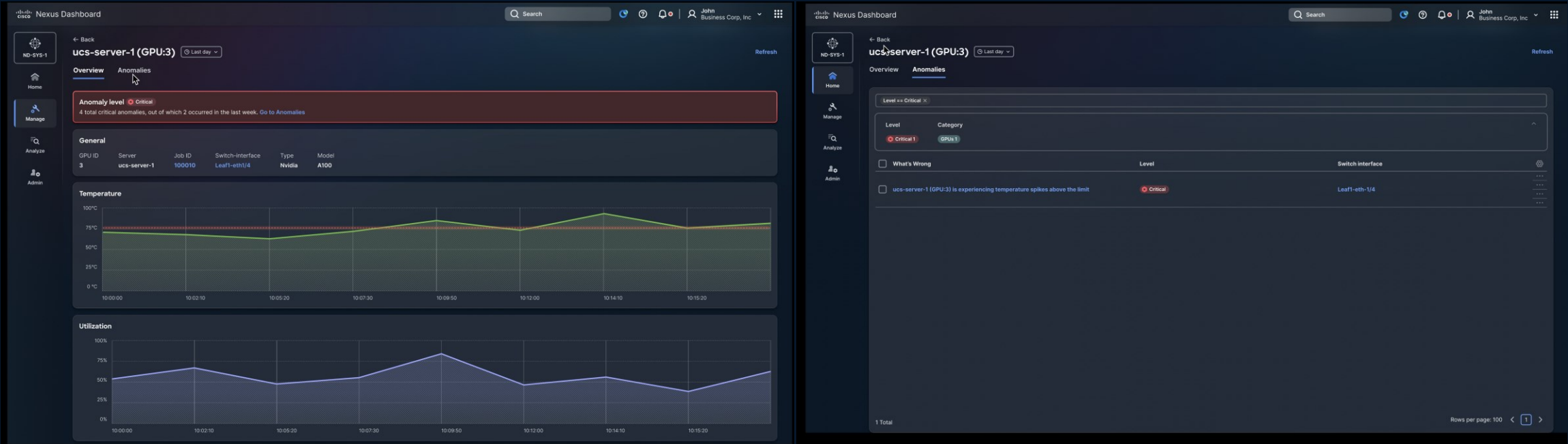
Nexus Dashboard 4.2– AI Fabric Topology



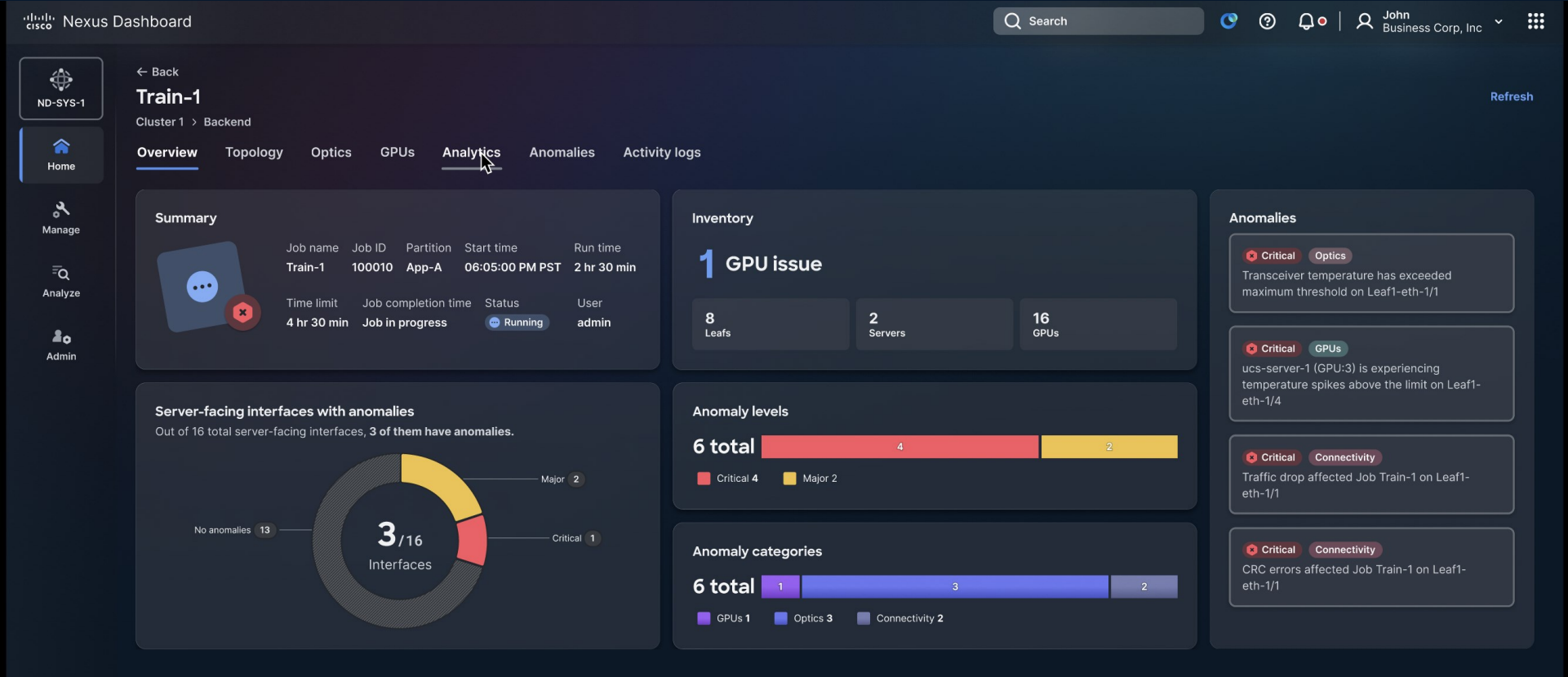
Nexus Dashboard 4.2– AI Job Monitoring



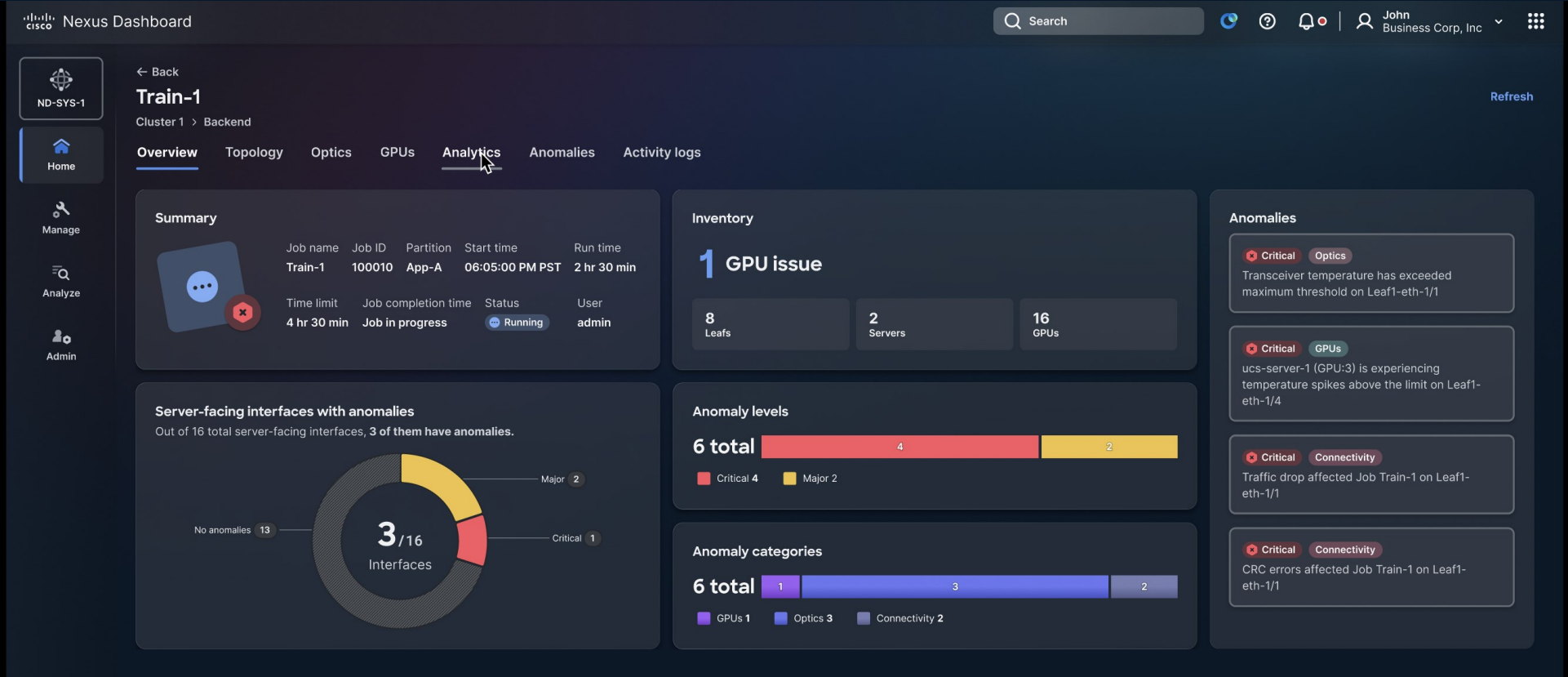
Nexus Dashboard 4.2 – GPU Health



Nexus Dashboard 4.2 – AI Fabric Analytics



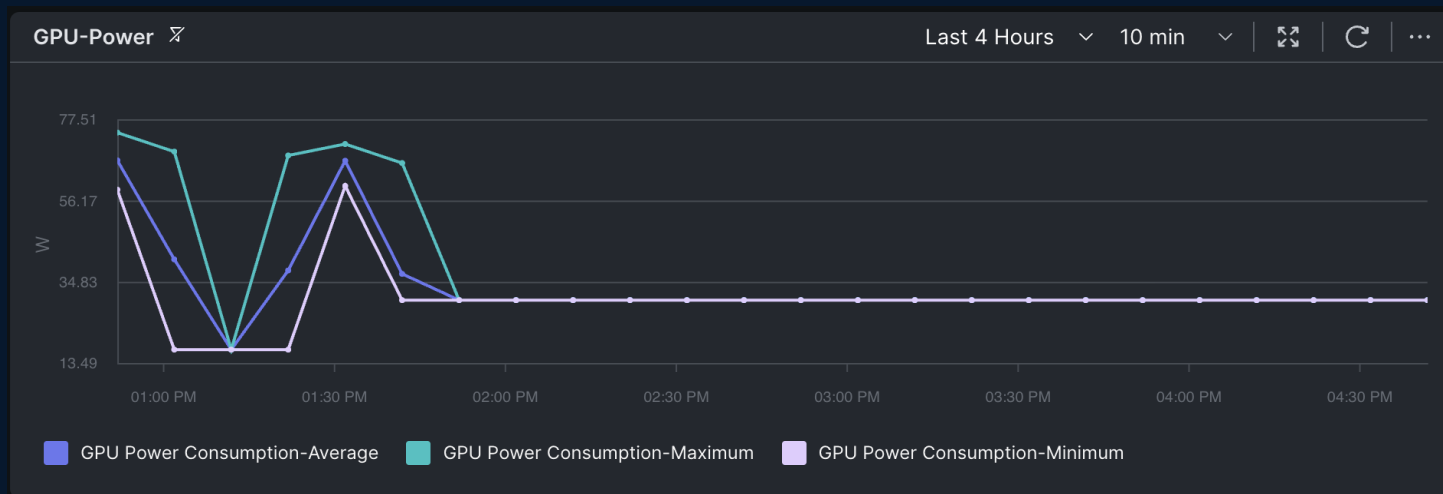
Nexus Dashboard 4.2 – AI Fabric Analytics



A large, stylized graphic of a sphere or circle with a gradient from dark blue to light blue, with a white outline, positioned behind the text.

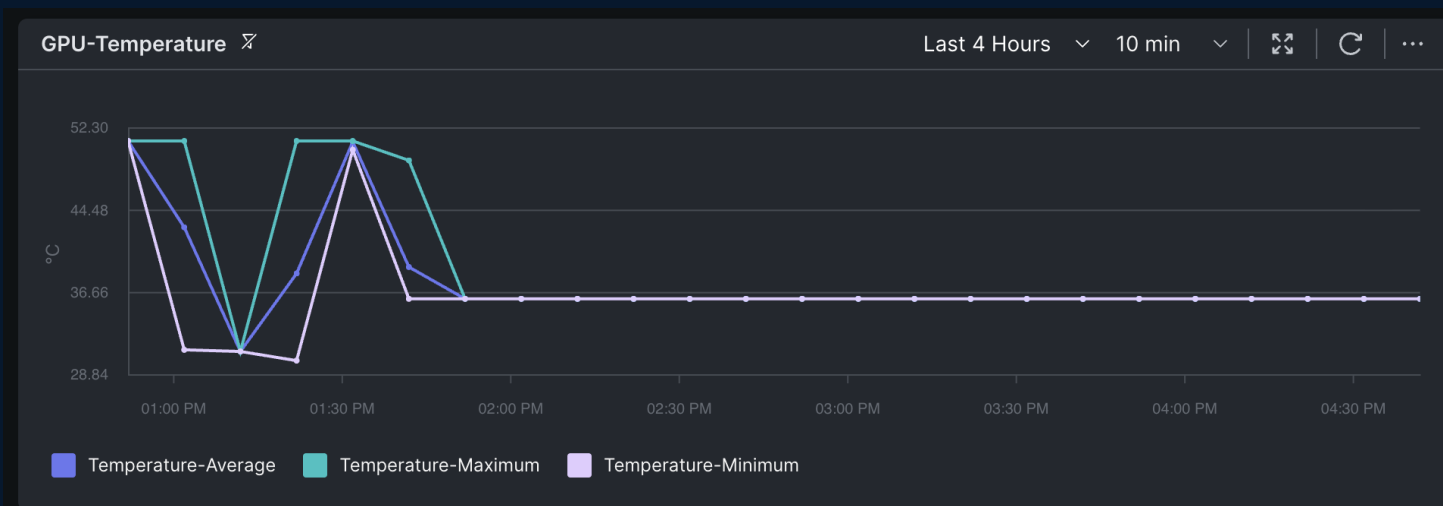
Cisco Intersight

Intersight - GPU Power and Temperature



GPU Power Metric

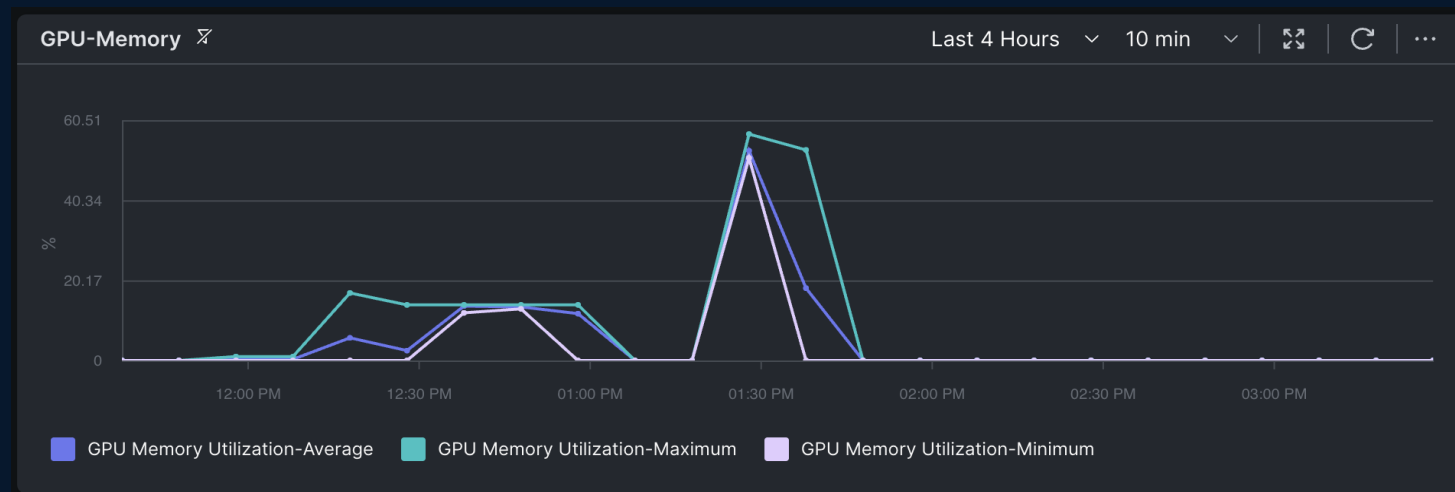
- Measures real time power draw to assess hardware health performance efficiency and energy usage
- Identifies power spikes thermal stress and workload imbalance
- Validates power delivery and cooling capacity for sustained GPU demand



GPU Temperature Metric

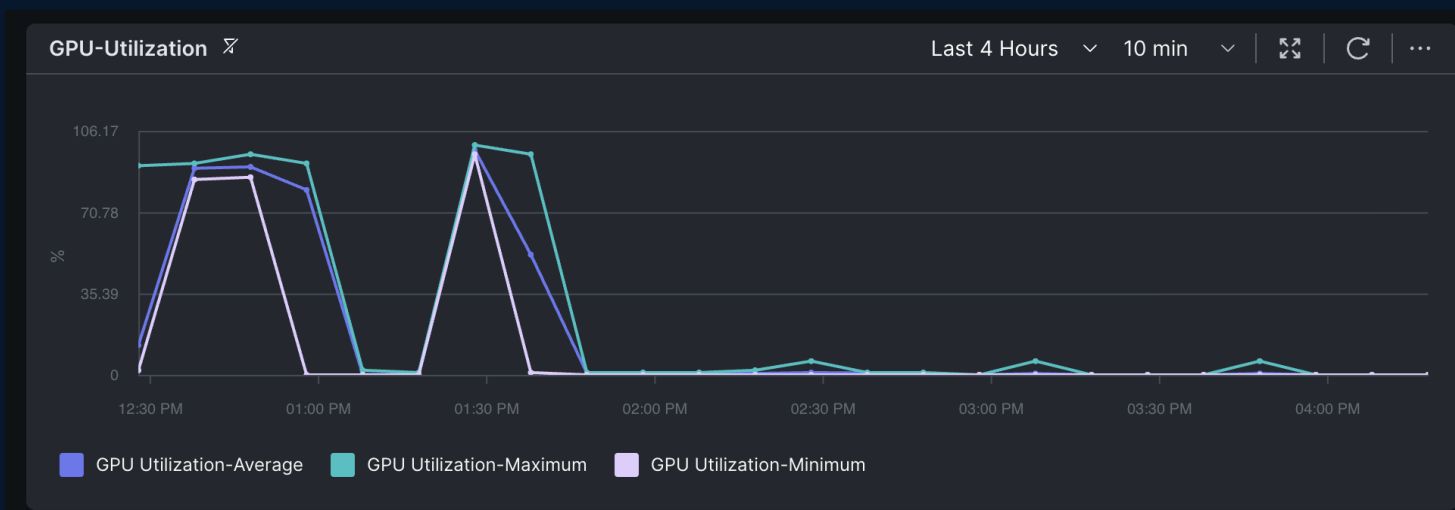
- Tracks GPU and memory temperatures to prevent thermal throttling and performance degradation
- Highlights when elevated temperatures reduce clock speeds and efficiency
- Triggers alerts in Cisco Intersight when thresholds are exceeded

Intersight - GPU Memory and GPU Utilization



GPU Memory Metric

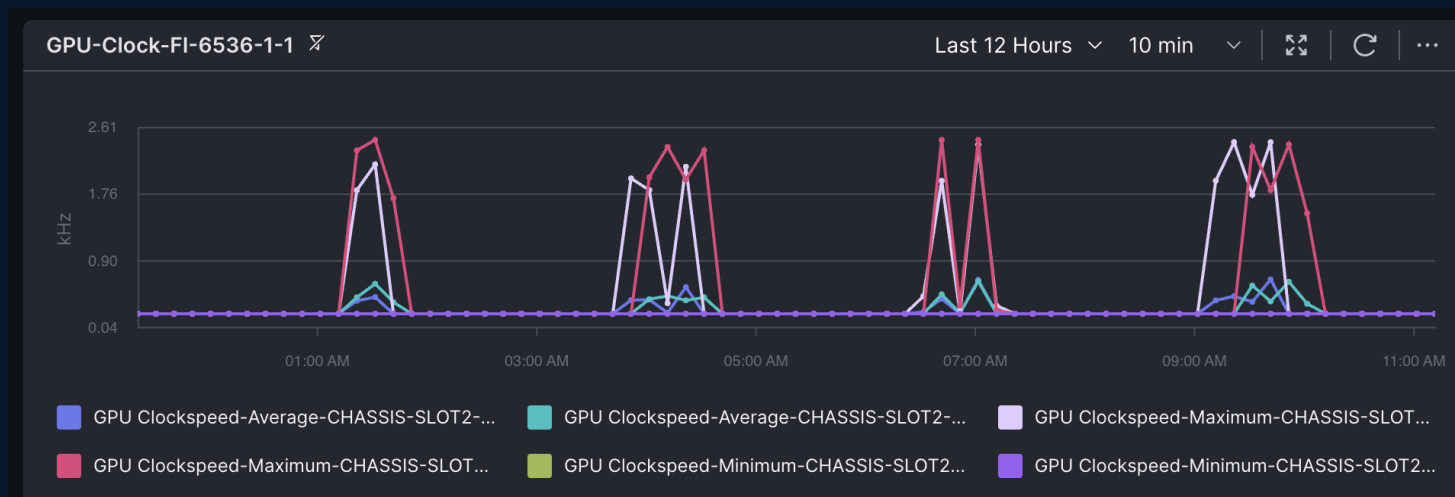
- Measures GPU memory usage in bytes or percentage
- Indicates memory pressure from model size activations and batch configuration
- High utilization can lead to out of memory errors performance degradation or memory bound workloads and helps determine if workloads are limited by memory capacity and should be correlated with bandwidth metrics



GPU Utilization Metric

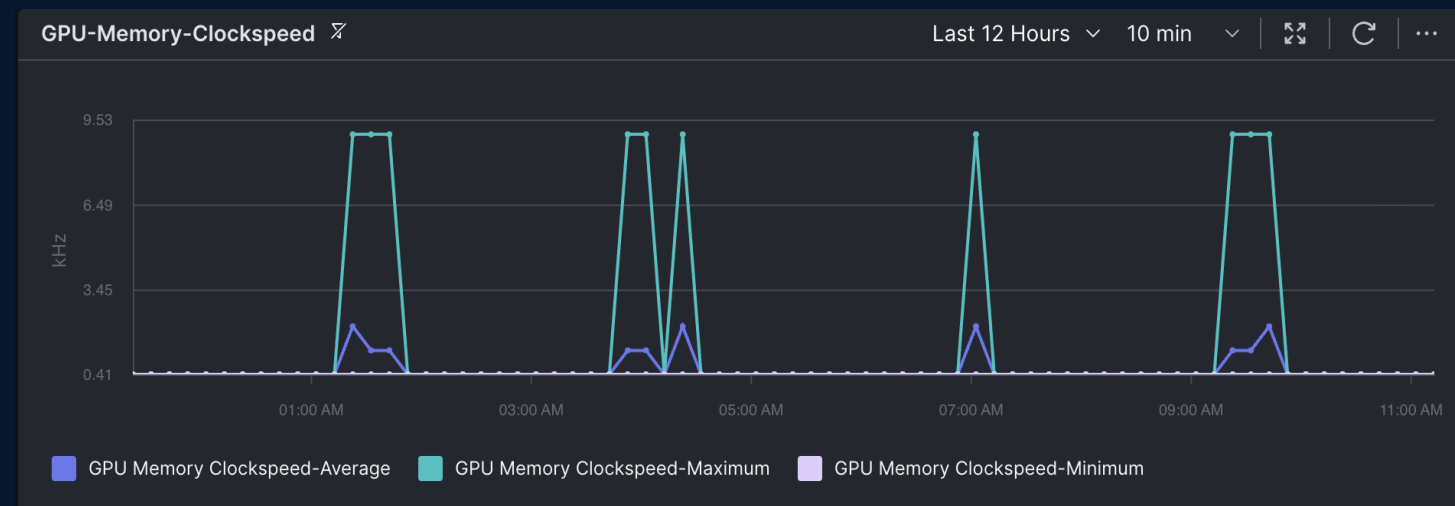
- Measures GPU compute activity based on Streaming Multiprocessor usage
- Indicates how busy the GPU is during workload execution
- Identifies underutilized GPUs or bottlenecks in CPU storage or data pipeline
- Represents activity not efficiency and should be correlated with memory and other metrics

Intersight - GPU and GPU Memory Clock Speed



GPU Clockspeed Metric (SM Clock)

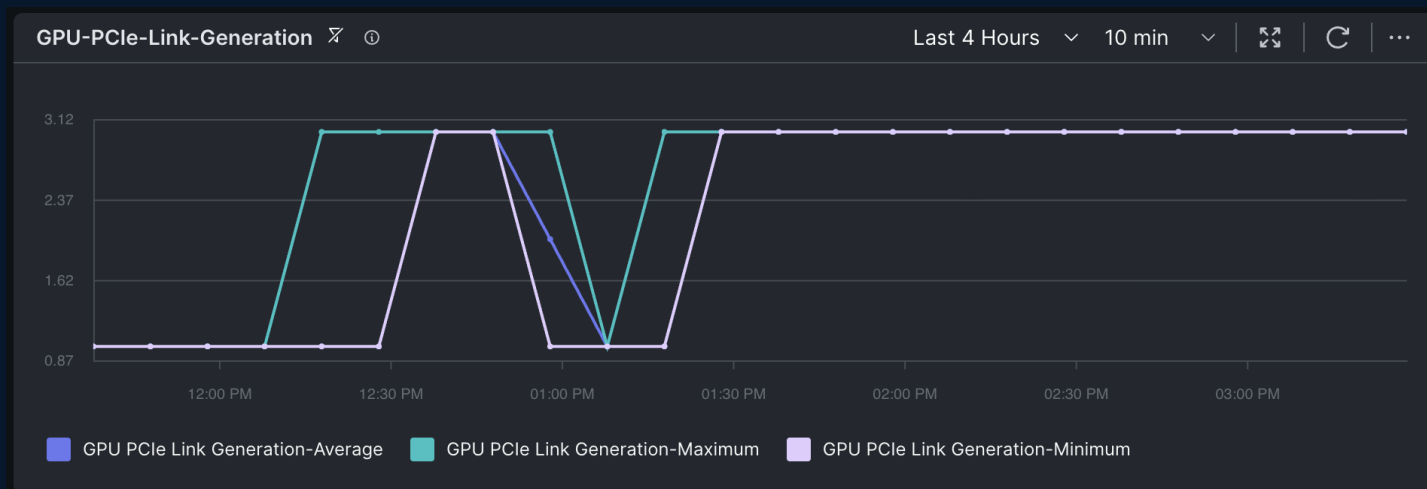
- Measures the operating frequency of GPU cores
- Indicates whether GPUs are reaching expected boost clocks under workload
- Drops in clockspeed can signal thermal throttling power limits or upstream bottlenecks
- Should be correlated with utilization and temperature to identify performance constraints



GPU Memory Clockspeed Metric

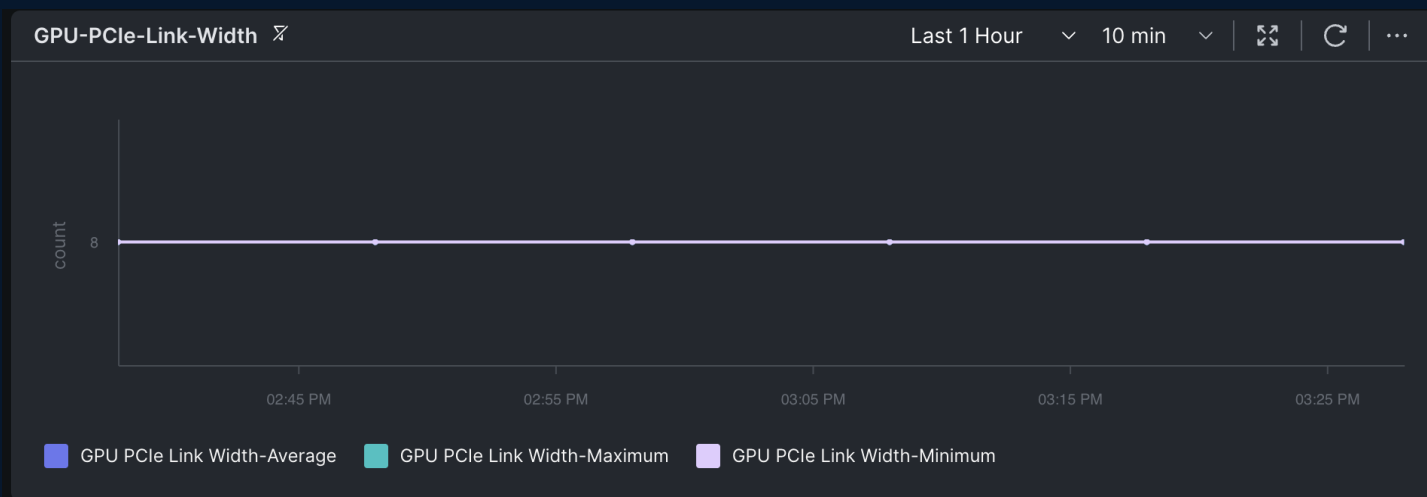
- Measures memory clock frequency
- Represents the rate at which data is accessed from GPU memory
- Reduced clockspeed can indicate memory bandwidth constraints or power and thermal limits
- Should be correlated with memory utilization and workload behavior to detect data movement bottlenecks

Intersight - GPU PCIe Link Metrics



GPU PCIe Link Generation Metric

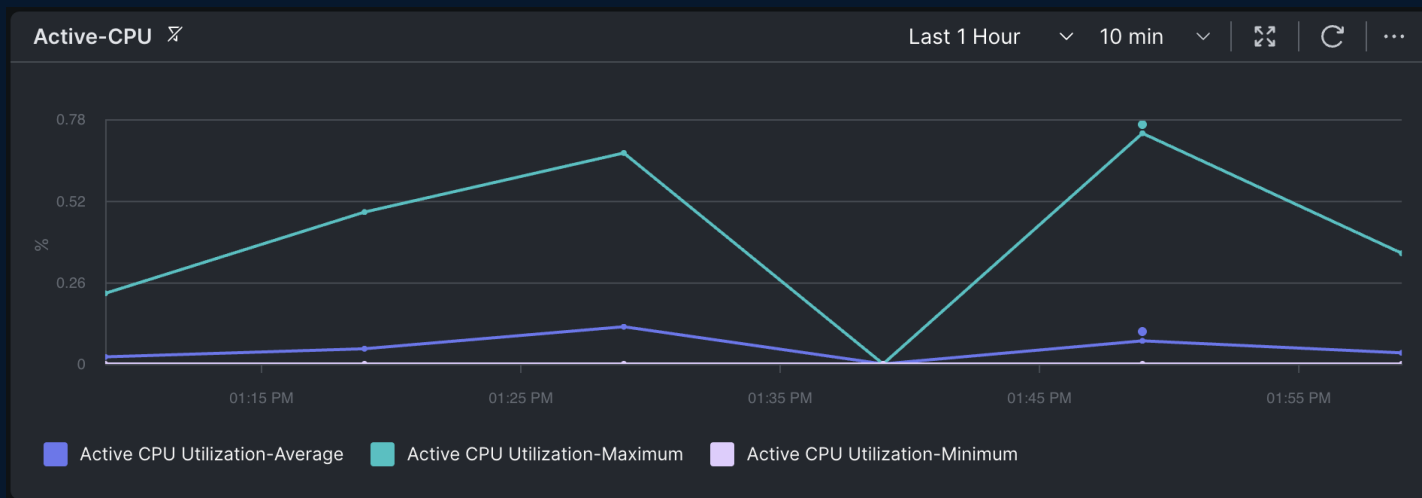
- Measures the negotiated PCIe link speed between GPU and host and indicates whether the system is operating at expected performance capability
- May reduce during idle due to power management but should scale under workload
- Sustained low link generation under load can indicate configuration issues or hardware limitations



GPU PCIe Link Width Metric

- Measures the negotiated number of PCIe lanes between GPU and host
- Indicates available bandwidth for data transfer between CPU and GPU
- Reduced link width can signal hardware faults misconfiguration or lane availability issues
- Should be monitored to ensure full bandwidth for GPU workloads

Intersight - Active CPU Utilization

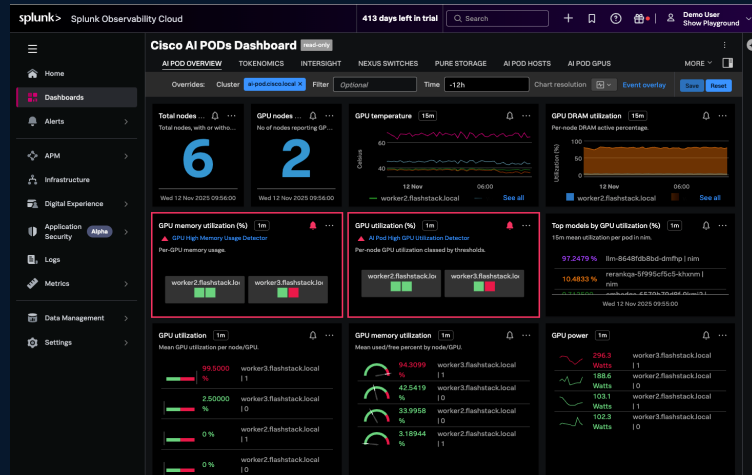


Active CPU Utilization Metric

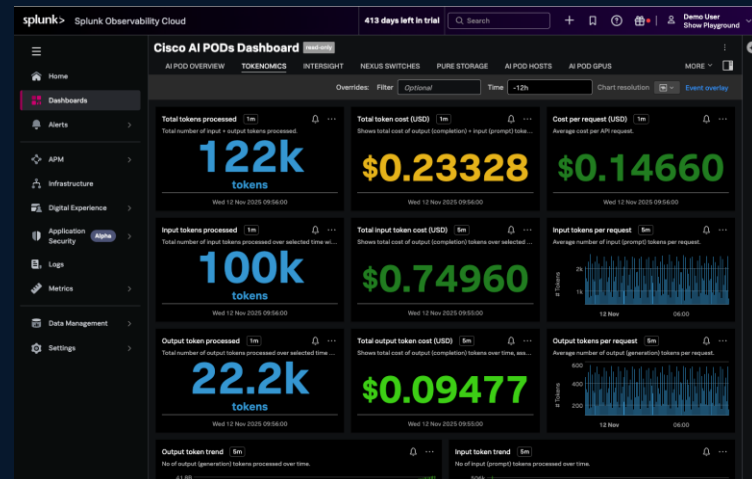
- Measures CPU activity based on time spent in active processing state (C0)
- Represents aggregated utilization across all CPU cores
- Indicates CPU demand for data preparation scheduling and system operations
- High CPU utilization can signal bottlenecks impacting GPU efficiency and data pipeline performance
- Should be correlated with GPU utilization to identify upstream bottlenecks

Splunk AI Observability

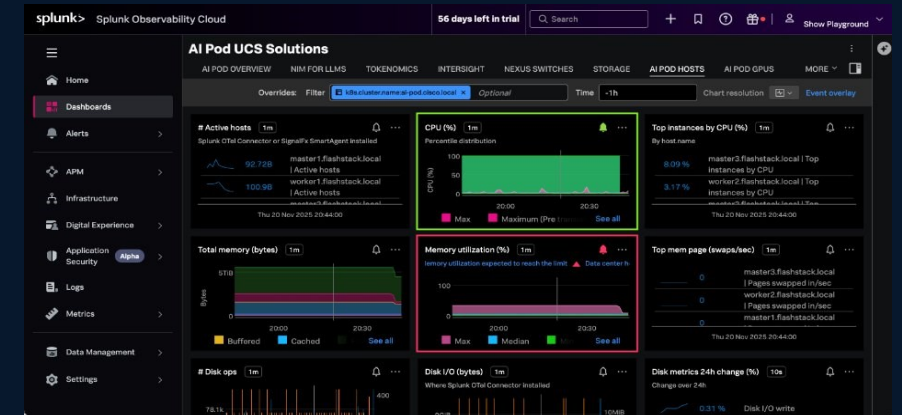
Splunk Observability Cloud – AI Dashboard Views



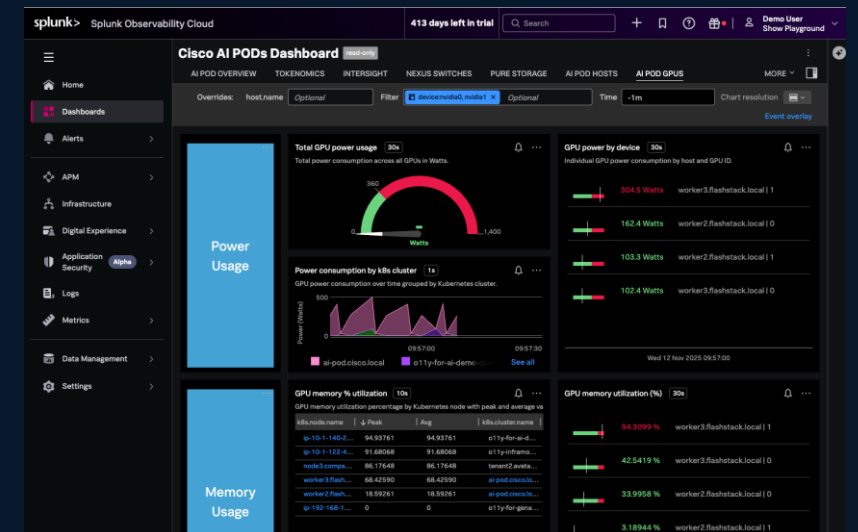
AI POD Overview - Total nodes, GPU power and utilization (%), etc...



Tokenomics - Total tokens processed, Total token cost, etc...

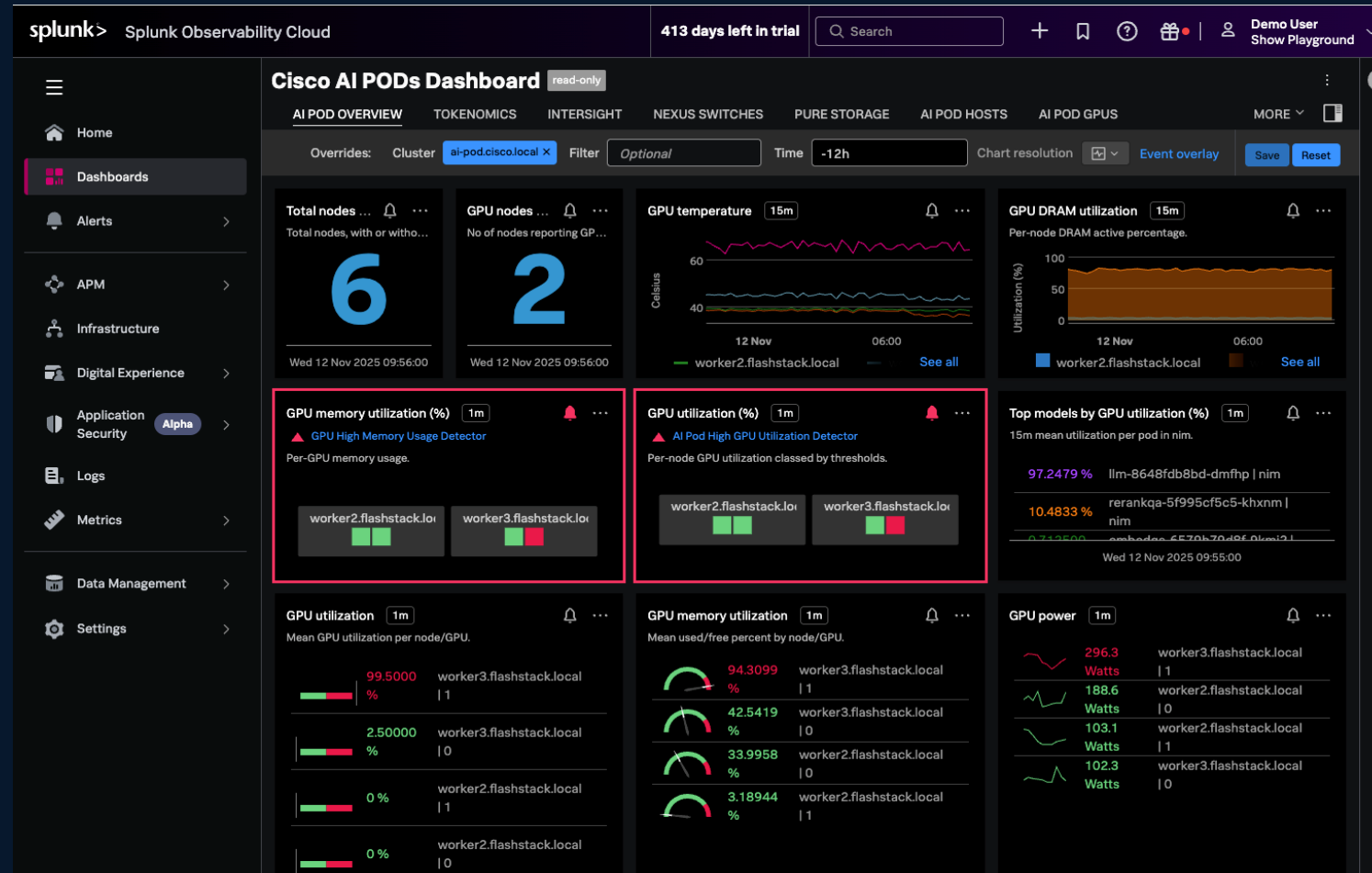


AI POD Hosts - # of active hosts, CPU (%), memory utilization (%), etc...



AI POD GPUs- GPU power usage, GPU memory utilization (%), etc...

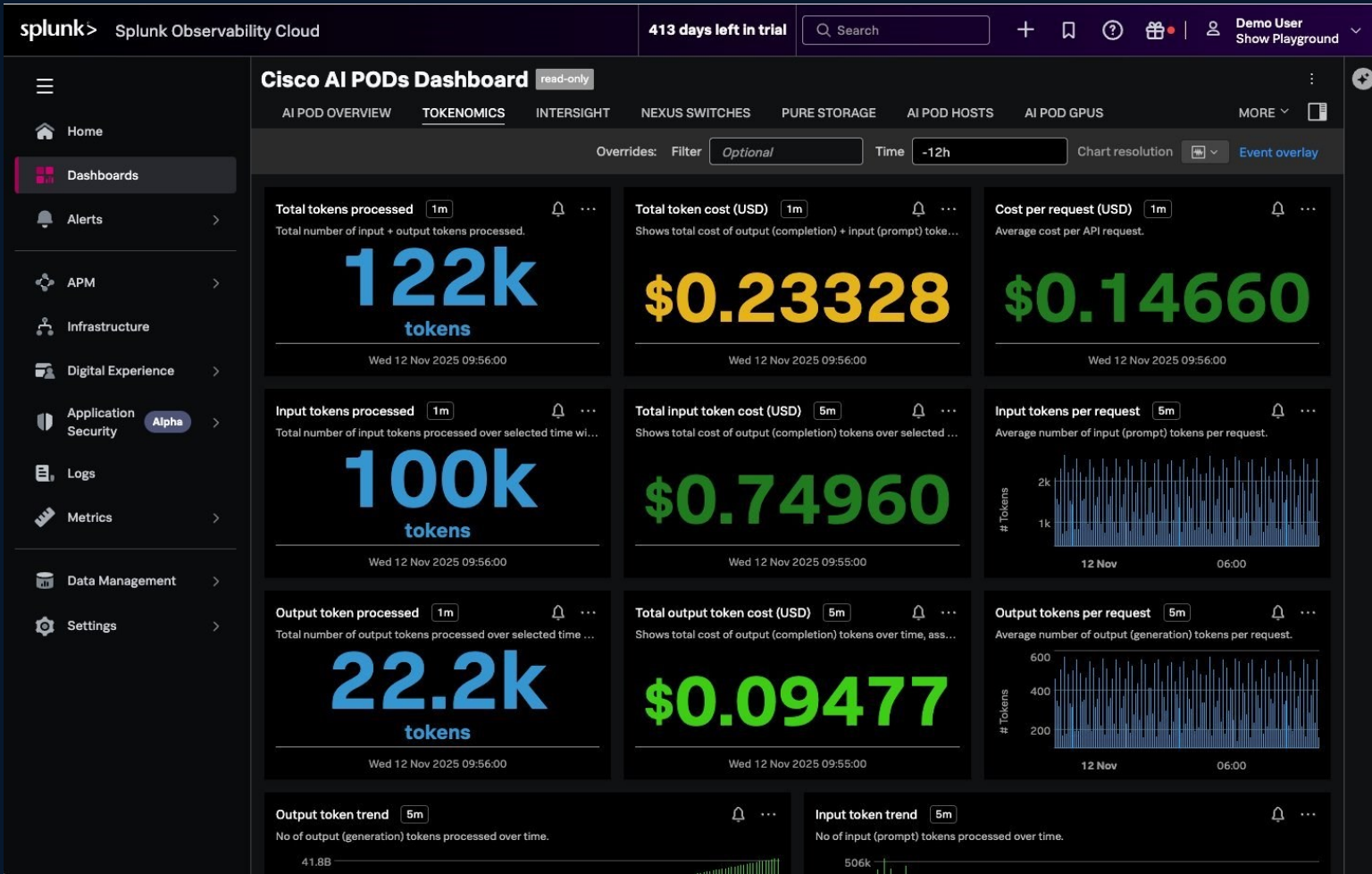
Splunk Observability Cloud – AI POD Overview



AI POD Overview

Total nodes, GPU power and utilization (%), etc...

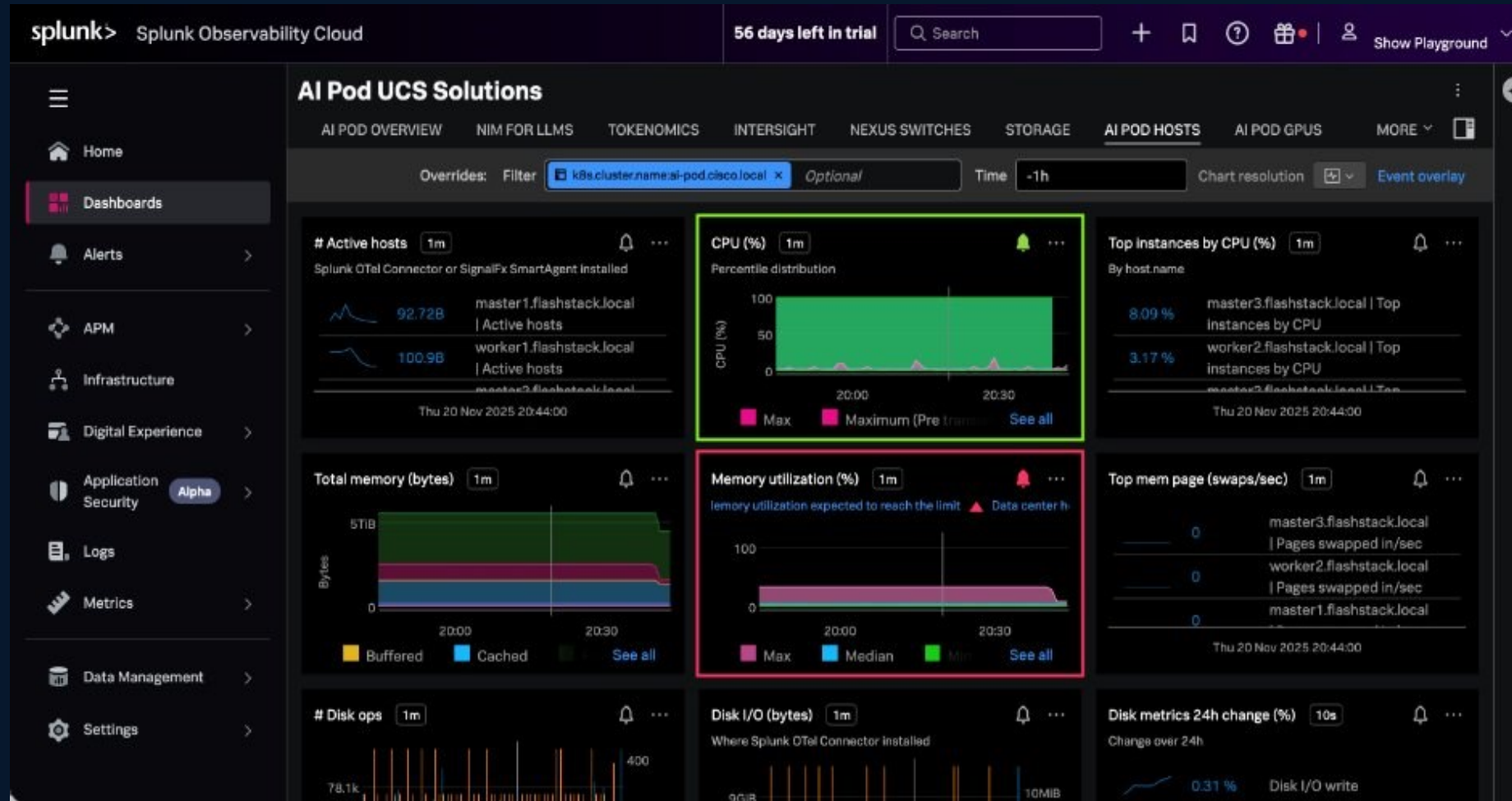
Splunk Observability Cloud – Tokenomics



Tokenomics

Total tokens processed, Estimated token cost, etc...

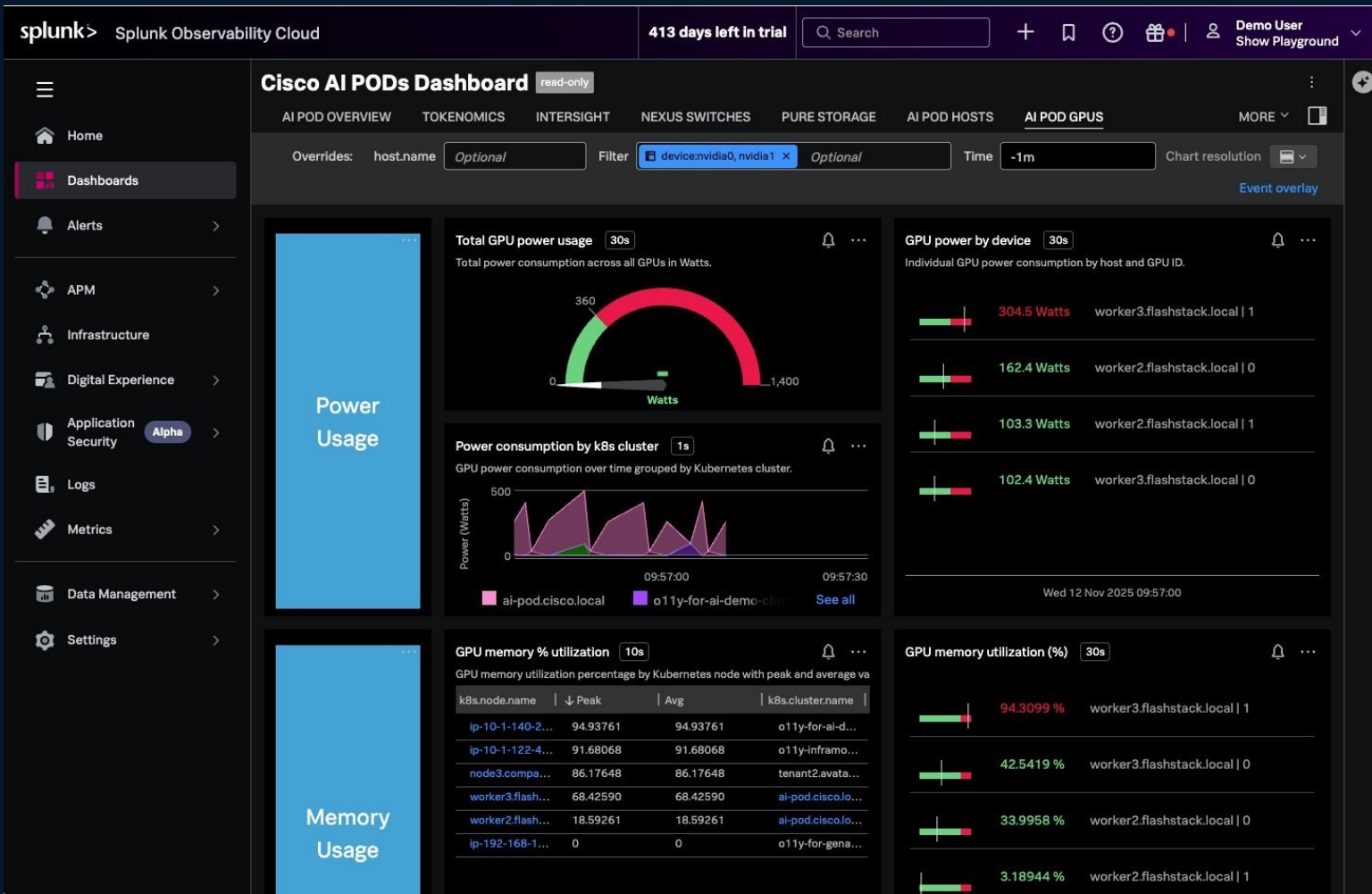
Splunk Observability Cloud – AI POD Hosts



AI POD Hosts

of active hosts, CPU (%), memory utilization (%), etc...

Splunk Observability Cloud – AI POD GPUs



AI POD GPUs

GPU power usage, GPU memory utilization (%), etc...

Complete your session evaluations



Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2027.



Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.



Level up and earn exclusive prizes!



Complete your surveys in the Cisco Live mobile app.

Continue your education



Visit the Cisco Showcase for related demos



Book your one-on-one Meet the Engineer meeting



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



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

CISCO Live !

Thank you



