CISCO Connect

Cisco Data Center Networking: Architecting the AI-Ready Infrastructure of Tomorrow

Erik Sklba- Account Executive, Architecture

March 24, 2026

Erik Sklba

- **AI Account Executive (~2 Years)**
- **Cisco Commercial East coverage**

- **C&AI Specialist @ Cisco (~8 Years)**
- **Financial Services, Healthcare, Manufacturing, Retail, Utility, & Software Verticals**

- **Systems Engineer @ Cisco (~8 Years)**

- ersklba@cisco.com
- 860.329.6603
- Linkedin.com/in/eriksklba

Agenda

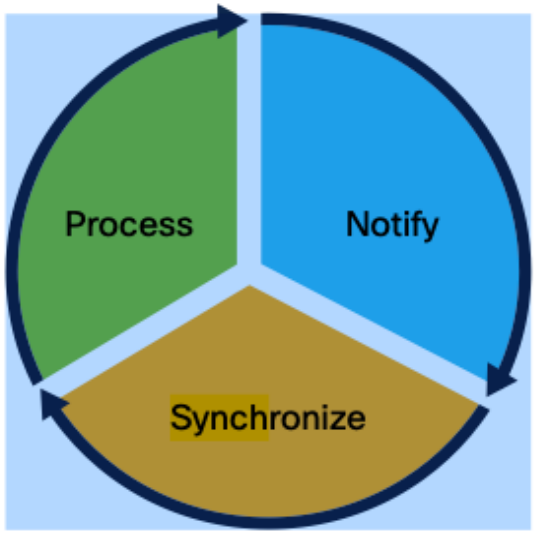
1. The “Why”: GPU Iteration and Data Transfer
   - GPU Iterations
   - NIC Transfer Pipeline
2. Topologies
   - GPU Backend
   - Network Requirements for AI/ML
   - Rail Optimized Design
3. Tuning Techniques
   - ECN & PFC
   - ECMP/DLB/DRE
4. Cisco Portfolio
   - Switches & Optics
   - Hyperfabric
   - Nexus Dashboard

The “Why”: GPU Iteration and Data Transfer

AI/ML Workload Characteristics

Execute instructions on GPU

High bandwidth compute can saturate network links



Send results of computation

Several methods, we'll focus just on one
All-to-All Collective (Everyone sends to everyone)

Wait for everyone to complete

Creates synchronization between GPUs
Computation stalls waiting for the slowest Path
Job Completion Time (JCT) is based on the worst-case tail latency

How a GPU Processes an Iterative Function

f(x) = sin(x) · e^(-x/10) — applied repeatedly until convergence

1 LOAD
Read from global memory

2 COMPUTE
All cores execute in parallel

3 WRITE BACK
Store results to memory

GPU DIE — 32 CUDA CORES

C0 0.48	C1 0.84	C2 0.96	C3 0.91	C4 0.76	C5 0.56	C6 0.33	C7 0.12	Loading
C8 sin·e⁻ˣ	C9 sin·e⁻ˣ	C10 sin·e⁻ˣ	C11 sin·e⁻ˣ	C12 sin·e⁻ˣ	C13 sin·e⁻ˣ	C14 sin·e⁻ˣ	C15 sin·e⁻ˣ	Computing
C16 0.42	C17 0.67	C18 0.73	C19 0.71	C20 0.63	C21 0.50	C22 0.31	C23 0.12	Complete
C24 —	C25 —	C26 —	C27 —	C28 —	C29 —	C30 —	C31 —	Idle

GLOBAL MEMORY BUS

ITERATION LOOP

Initialize x[0..N]
Assign starting values

Dispatch to cores
Each core gets x[i]

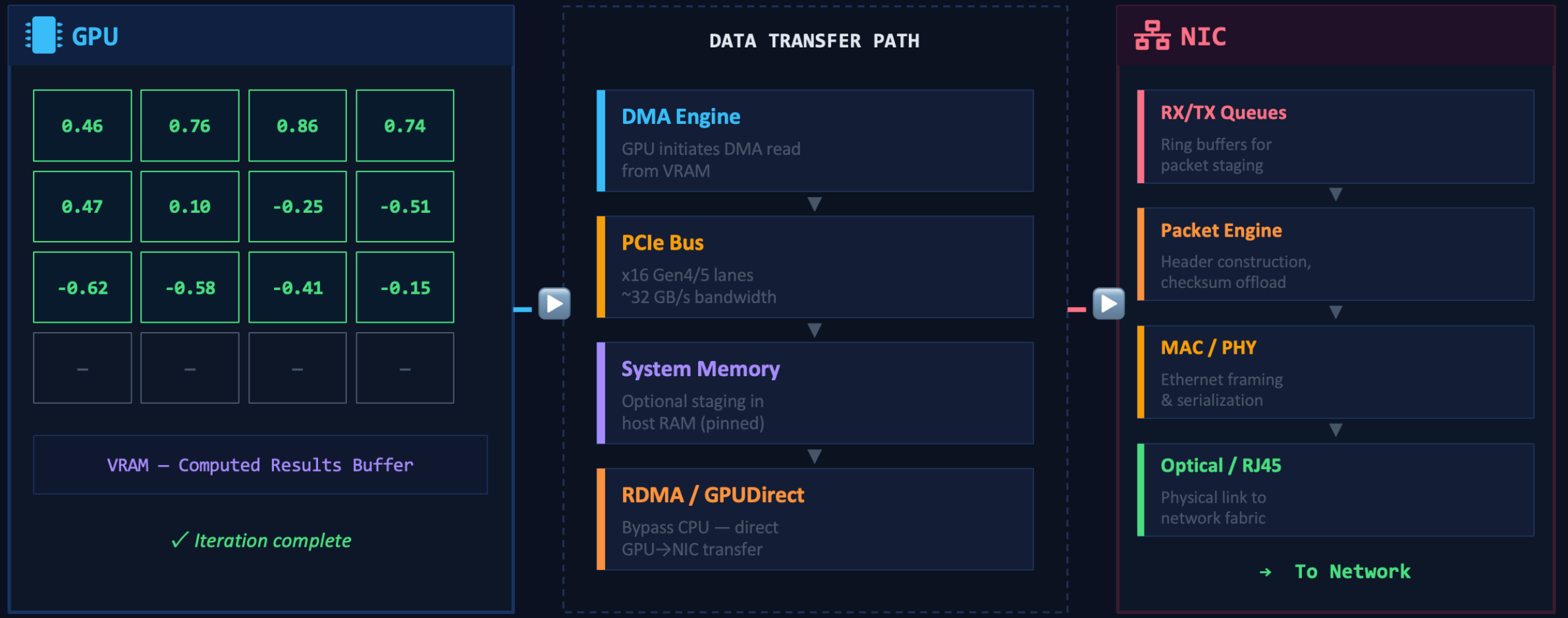
Apply f(x) in parallel
All cores compute at once

Sync & write back
x[i] ← f(x[i])

Repeat until converged

GPU → NIC: Data Transfer Pipeline

After iterative computation completes, results transfer from GPU memory through the PCIe bus to the NIC for network transmission



GPUDirect RDMA allows the NIC to read directly from GPU VRAM over PCIe, bypassing the CPU and system memory entirely — critical for low-latency AI inference serving.

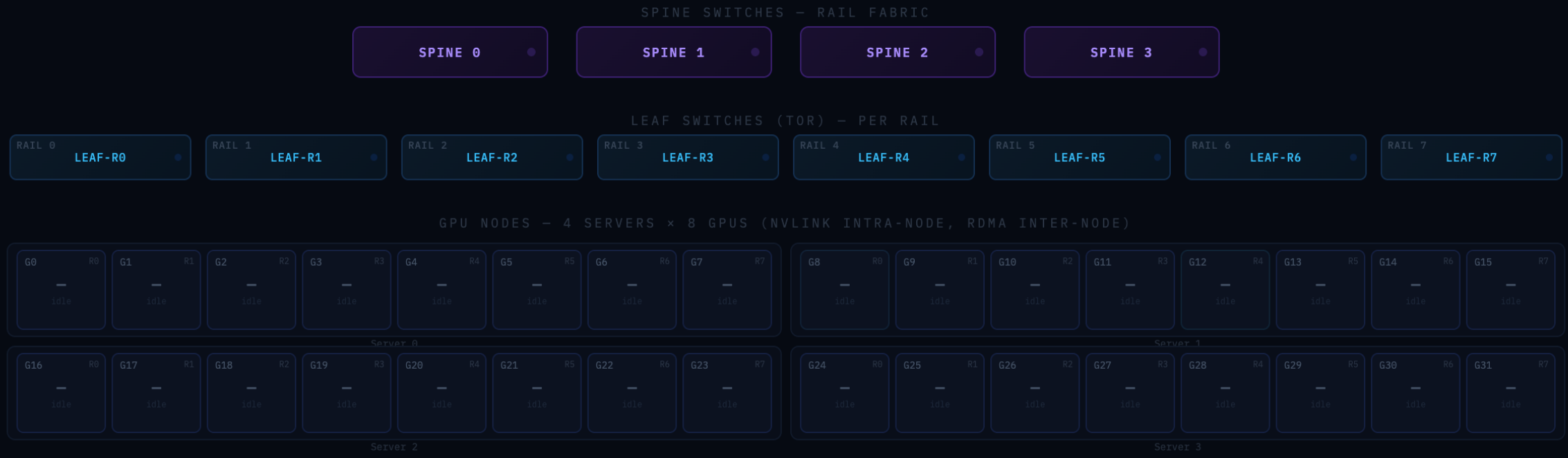
32-GPU Rail-Optimized Network Topology

4 Servers × 8 GPUs — Fat-tree with rail-optimized leaf/spine switching — AllReduce over RDMA

▶ Run AllReduce | ↺ Reset | Speed: Slow | Normal | Fast

● Compute ● Send ● Receive ● AllReduce

Idle	0 / 4	0 GB/s	32
Phase	Ring Step	Aggregate BW	Active GPUs



Rail-Optimized Topology

Each GPU NIC port (rail) maps to a dedicated leaf/spine path. GPU0 on every server connects to Leaf0→Spine0, GPU1→Leaf1→Spine1, etc. This ensures zero congestion between rails during AllReduce.

Ring AllReduce

Gradients are reduced in a ring pattern: each GPU sends a shard to the next GPU on the same rail across servers, and receives from the previous. After N-1 steps, all GPUs hold the fully reduced result.

GPUDirect RDMA

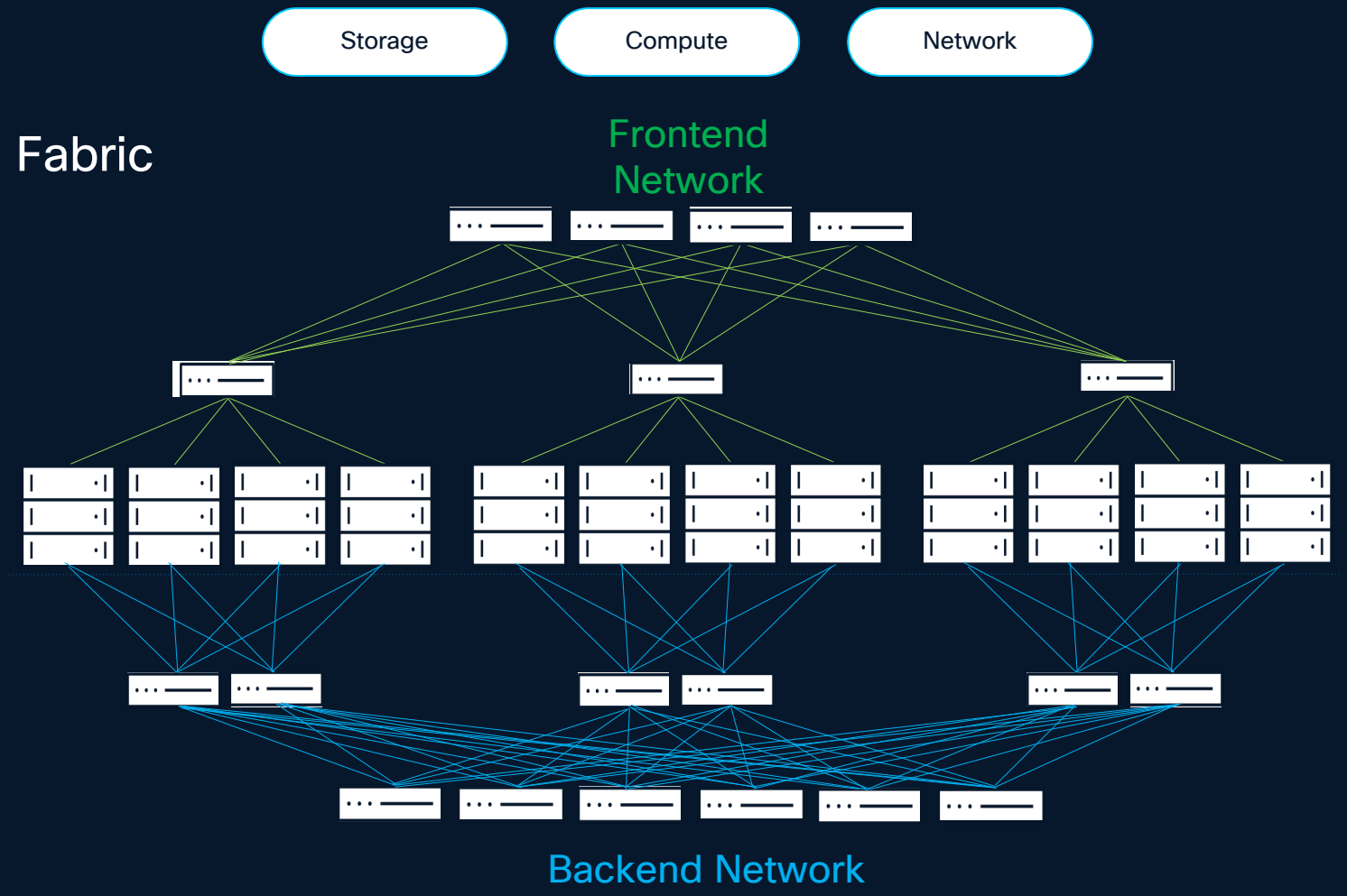
Data moves GPU→NIC→Switch→NIC→GPU without touching the CPU or system memory. Each rail operates at 200-400 Gb/s with InfiniBand or RoCEv2, giving ~50 GB/s per GPU bidirectional.

file:///Users/ersklba/Downloads/gpu-farm-rail-network.html

Let’s Build a Topology!

What Does Infrastructure for AI/ML Look Like?

- Front End & Back End
- High Throughput Lossless Fabric
- Rail-Optimized



Why GPU Backend Networks Matter

THE CHALLENGE

- Modern AI workloads require massive parallel processing
- Single GPU memory insufficient for large models
- Data must move between GPUs at extreme speeds
- Training/inference bottlenecked by communication latency

THE SOLUTION

- High-bandwidth interconnects enable GPU-to-GPU communication
- Backend networks optimized for low latency and high throughput
- Scale from single-node to multi-node clusters
- Different technologies for different scale requirements

cisco

AI/ML: Key Networking Requirements

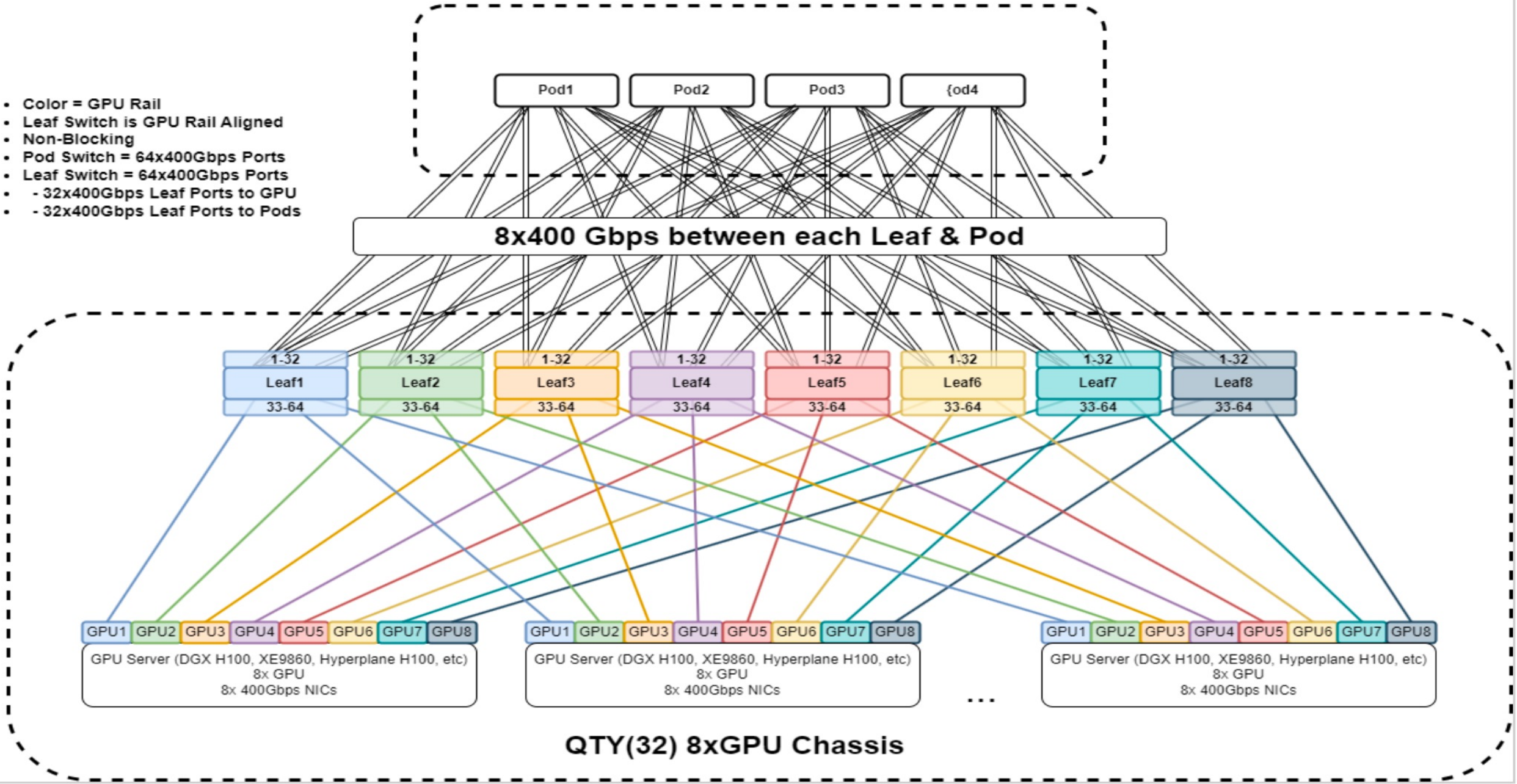
	Requirement	Solution
Latency	Low latency and jitter	• High speed ASICs • RDMA (Remote Direct Memory Access) NICs
Network Losses	Needs to be congestion less and lossless	• PFC (Priority Based Flow Control)/ECN • Scheduled fabric
Bandwidth/Scale	Network must support high bandwidth and scale	• ASIC's optimized for AI • High bandwidth NICs/optics
Load Balancing	Need efficient load balancing despite low entropy	• PBR rules/IBGP Pinning • ECMP with User Defined Fields • Dynamic Load Balancing
Visibility	Operator needs visibility to app perf, network utilization and adverse events	• Need well integrated application and network monitoring tools; telemetry

AI Infrastructure Requirements by Workload Type

Workload type	Small* model inferencing / RAG	Large* model inferencing / RAG	Fine Tuning	Small* model training	Large* model training
AI Accelerator (GPU)	Maybe 0-2 GPU per server	Yes 2-4 GPU per server	Yes 4-8 GPU per server	Yes 2-4 GPU per server	Yes 8 GPU per server
Cluster size	0 – 100's Scales horizontally with request rate	0 – 100's Scales horizontally with request rate	20-80 GPUs ~10's of servers	20-80 GPUs ~10's of servers	>1000 GPUs 100s of servers
Intra-host GPU interconnect (e.g. NVLink)	No	Maybe**	Maybe	Yes	Yes
Inter-host GPU interconnect ("Backend" network)	No	Probably Not**	Maybe 200-400G RoCE	Yes 200-400G RoCE	Yes 200-400G RoCE/Infiniband
"Frontend" network	"Any"	100-400G	100-400G	100-400G	100-400G

RoCE Rail-Optimized Network Topology
4 Pod Switches & 8x Bundles to Each Leaf

- Color = GPU Rail
- Leaf Switch is GPU Rail Aligned
- Non-Blocking
- Pod Switch = 64x400Gbps Ports
- Leaf Switch = 64x400Gbps Ports
- - 32x400Gbps Leaf Ports to GPU
- - 32x400Gbps Leaf Ports to Pods



Tuning Techniques

ECN & PFC

Normal = RoCE traffic at line rate, low buffer

- No marking or pausing

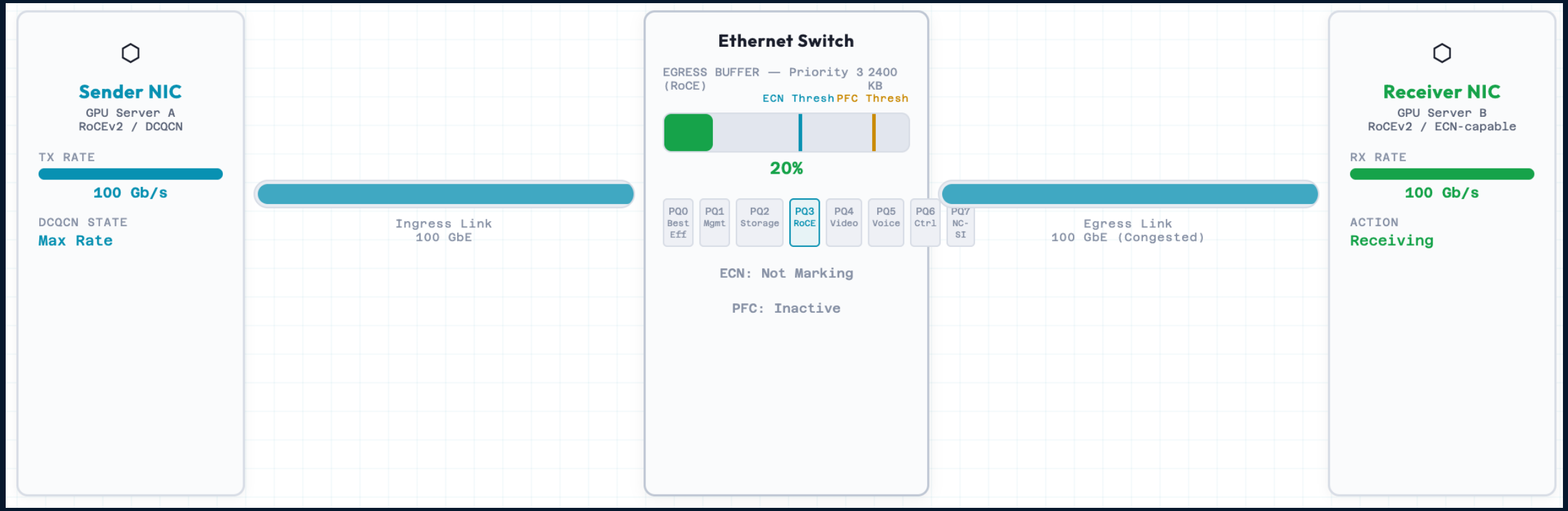
ECN Congestion – buffer fills past threshold

- Switch marks packets, NIC detects
- DCQCN algorithm cuts Tx rate (buffer drains, rate recovery)

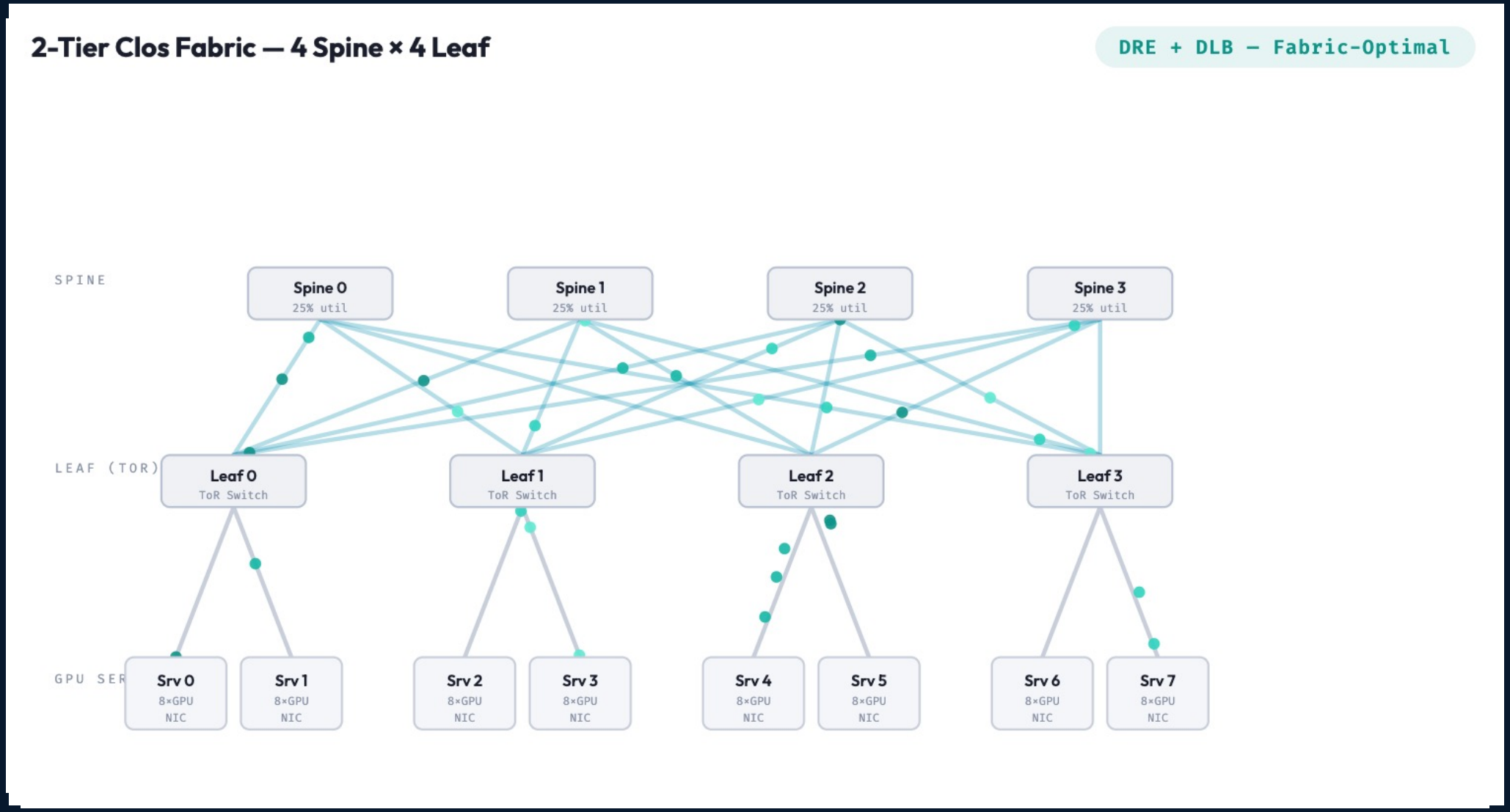
PFC Pause – a microburst fills the buffer before ECN/DCQCN responds

- Switch fires a PFC pause & sender hard stops
- Buffer drains, recovery but with Head of Line blocking

ECN & PFC in action



ECMP vs DLB vs DRE



Cisco’s Portfolio

Cisco’s Approach to AI/ML Networks

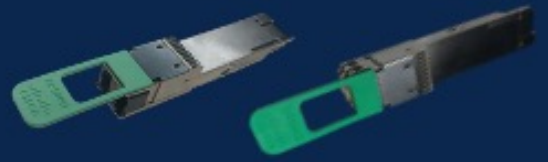
Proven full network stack

- A common architecture and consistent features across a variety of platforms
- Fabric automation and visibility made simple
- Widely deployed OS with rich feature set
- Extensive interop testing with major NIC vendors

Building blocks

Cisco® 8000 Series
Cisco Nexus® 9300 Series
800G, 400G or 100G Switches

800G, 400G, or 100G Optics



Infinitely customizable

- Open by Design
  Contributions across the open source community
- Broad API support
  BYO or Open Source tooling for management and visibility
- Co-developed and trusted by Hyperscalers
- Built for ultra-high scale

Cisco Silicon G200

Uniquely Efficient & Optimized for AI/ML

One architecture
A simpler and easier network to maintain

High performance
2x higher performance than G100

Sustainability via technology
2x more power efficient than G100

Ultra-low latency
2x Lower Latency than G100

Optimal network design
512-wide radix enables
flatter, more efficient networks

Fully shared packet buffer
Optimal fairness, burst performance,
Job Completion Time (JCT)

51.2 Tbps
5nm Technology

Advanced 112 Gbps SerDes
Cisco® designed next-generation ADC SerDes
Support for Optics, 4-meter DAC, LDO, and CPO

Advanced load balancing
Non-correlated Weighted Equal Cost Multi-Path (WECMP) avoids hash polarization Congestion-aware stateful load balancing Congestion-aware packet spraying

Link failure avoidance
HW-based traffic link failure redistribution optimizes real-world large-scale deployments

Programmable processor
Deterministic ultra-low latency processor
with run to completion for ultimate flexibility

435B+

Lookups per second
Enables advanced features like SRv6 uSID

Deep visibility and analytics
In-band telemetry including emerging protocols
Hardware analyzers enable post event debuggability

cisco

Cisco & NVIDIA Network Partnership

Cisco Reference Architectures, based on NVIDIA Enterprise & Cloud Partner design principles.

*ERA: Cisco Enterprise Reference Architecture

*CRA: Cisco Cloud Reference Architecture

*NCP: NVIDIA Cloud Partner Reference Architecture Compliant,

Cisco N9300 & HF6100 Series Switches

NVIDIA BlueField

Powered by Cisco Silicon with NVIDIA Spectrum X integration

NVIDIA BlueField

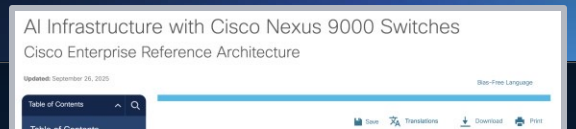
Cisco N9100 Series Switches

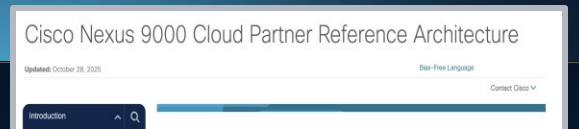
NVIDIA BlueField

Powered by NVIDIA Spectrum-X Ethernet Switch Silicon

NVIDIA BlueField

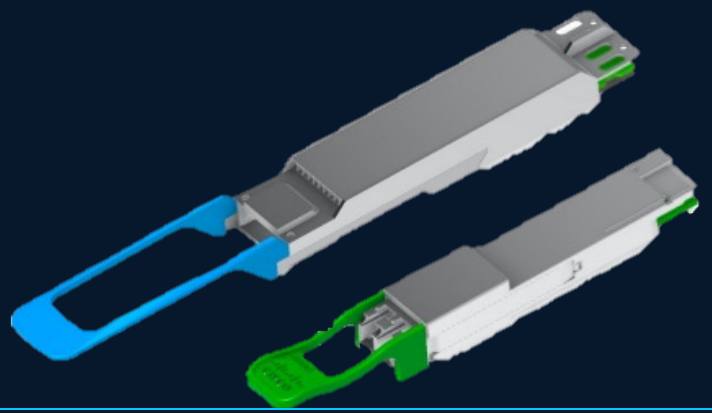
The only 3rd party switch validated as part of the reference architecture





800G Optics: OSFP vs. QSFP-DD

- Quad Small Form-Factor Pluggable Double-Density (QSFP-DD) and Octal Small Form Factor Pluggable (OSFP) – 2 transceiver form factors, same functionality
- Both have same optical, electrical, and management interfaces
- QSFP-DD and OSFP are optically interoperable
  - Same optical cables and connectors
  - Dissipate same power
- Physically different enough – not port compatible
- QSFP-DD ports are physically backward compatible with all QSFP modules
- OSFP comes in 3 physical forms (not compatible w/ each other)
  - Integrated Heat Sink (for switch and routers) 400G and 800G
  - Riding Heat Sink (for NICs and GPUs) 400G and 800G
  - OSFP-XD for 1.6T and 3.2T



GA in February 2026

Nexus Dashboard

AI Job Observability

End-to-end visibility

Monitor AI workloads across the entire stack. Track network, NICs, GPUs, and distributed compute nodes.

Topology-aware visualization

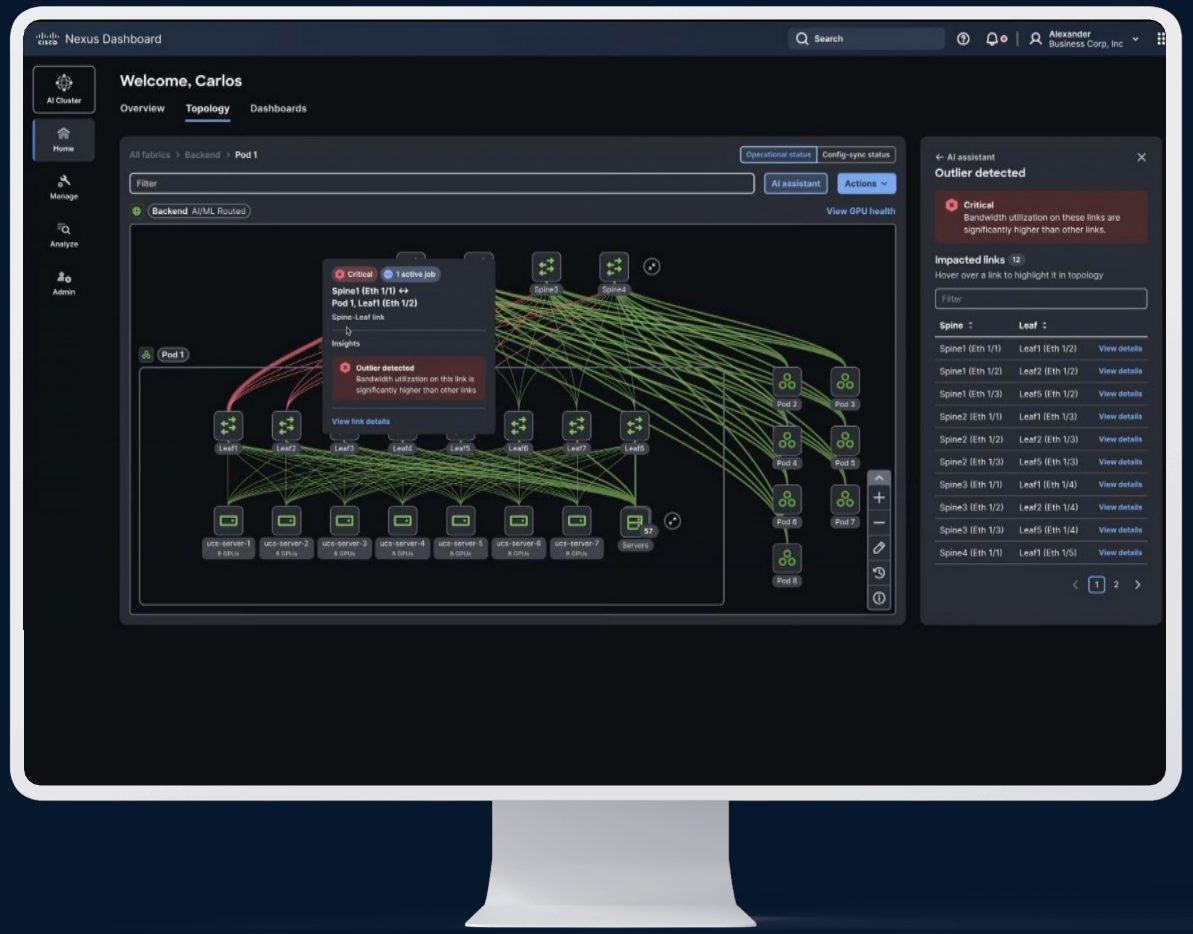
Correlate network elements with hardware components. See connections between infrastructure and AI job health.

Real-time metrics

Track throughput, latency, and GPU utilization. Get actionable insights into distributed AI workloads.

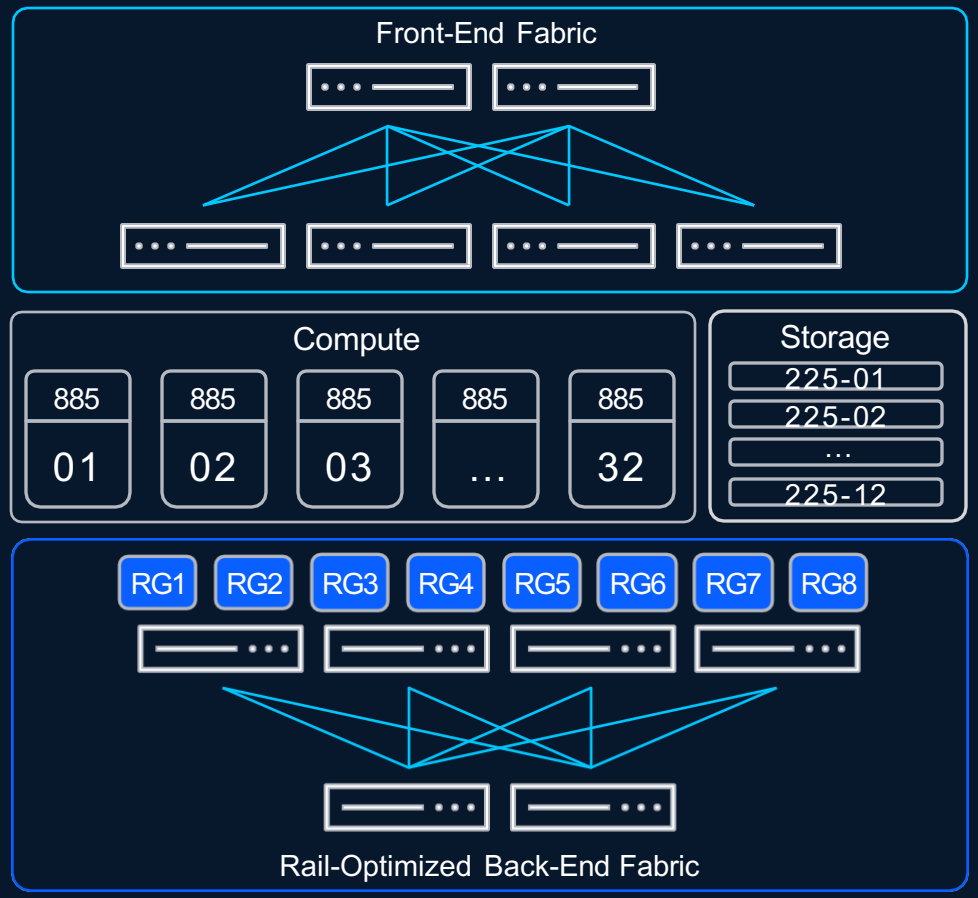
Proactive troubleshooting

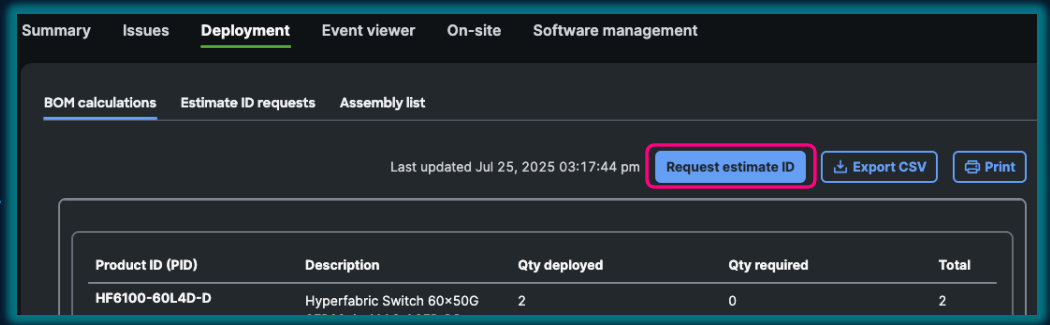
Detect anomalies against historical baselines. Accelerate root-cause analysis with automated recommendations.

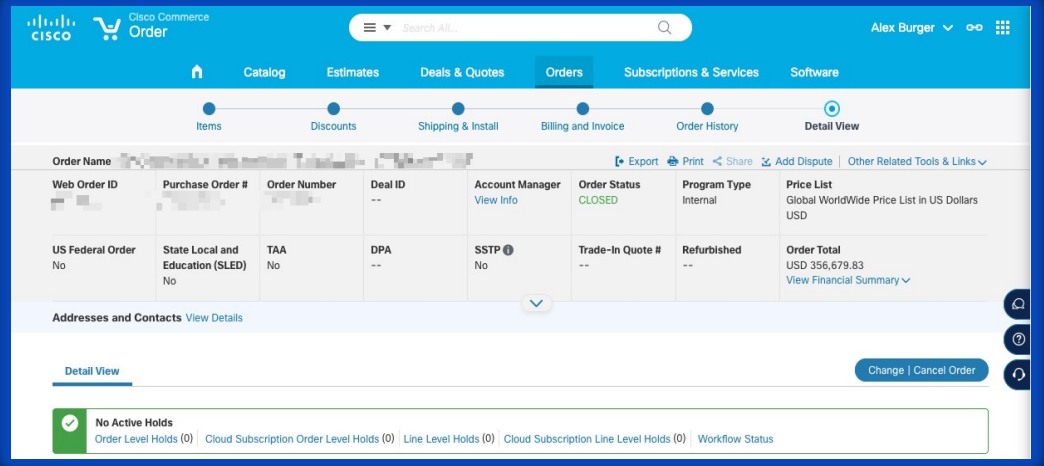


Hyperfabric AI Differentiation – Day 0

Directly export blueprint to Cisco Commerce Workspace, speeding up design-to-order process and reducing errors

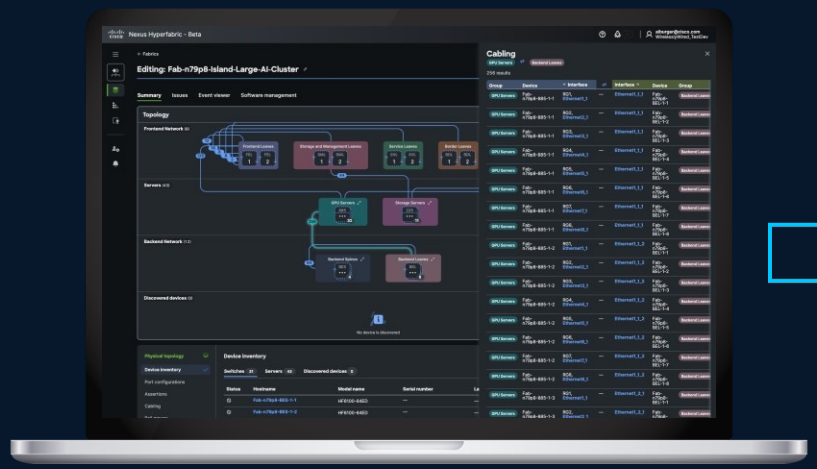


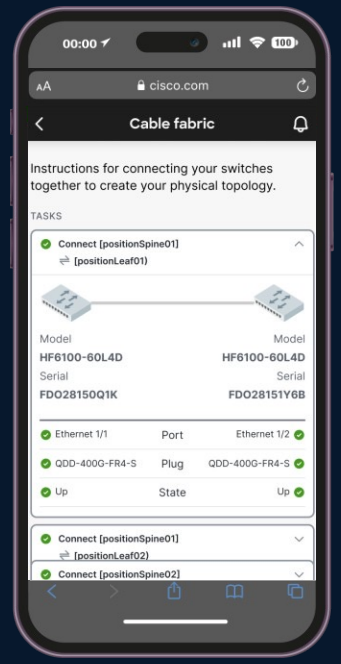


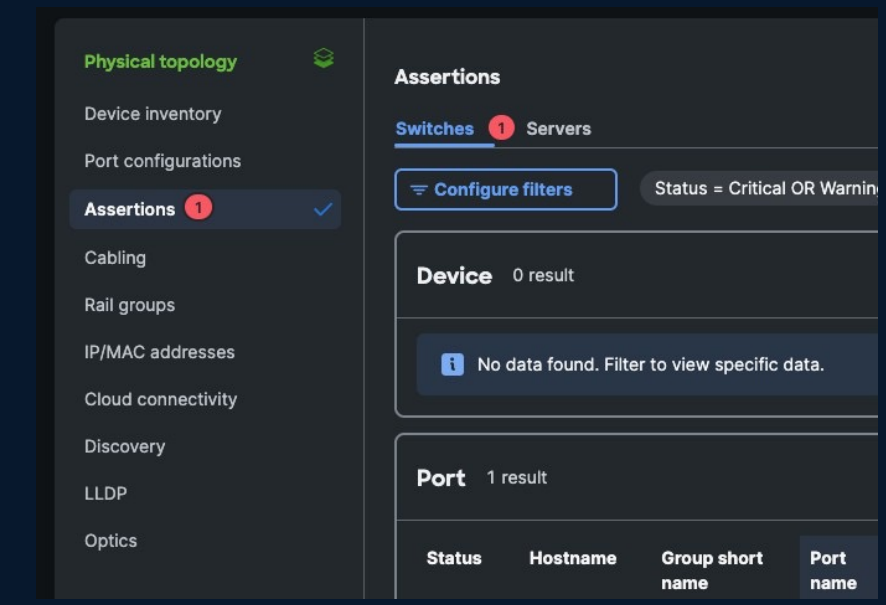


Hyperfabric AI Differentiation – Day 1

From design to deployment and validation







Blueprint

On-Site

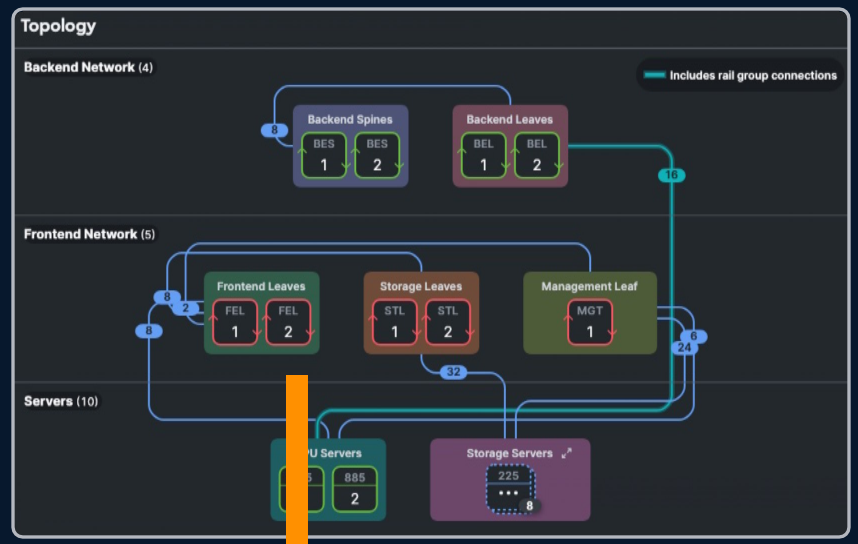
Validation

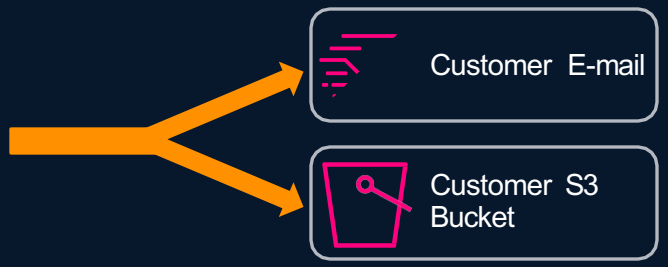
Hyperfabric AI Differentiation – Day N

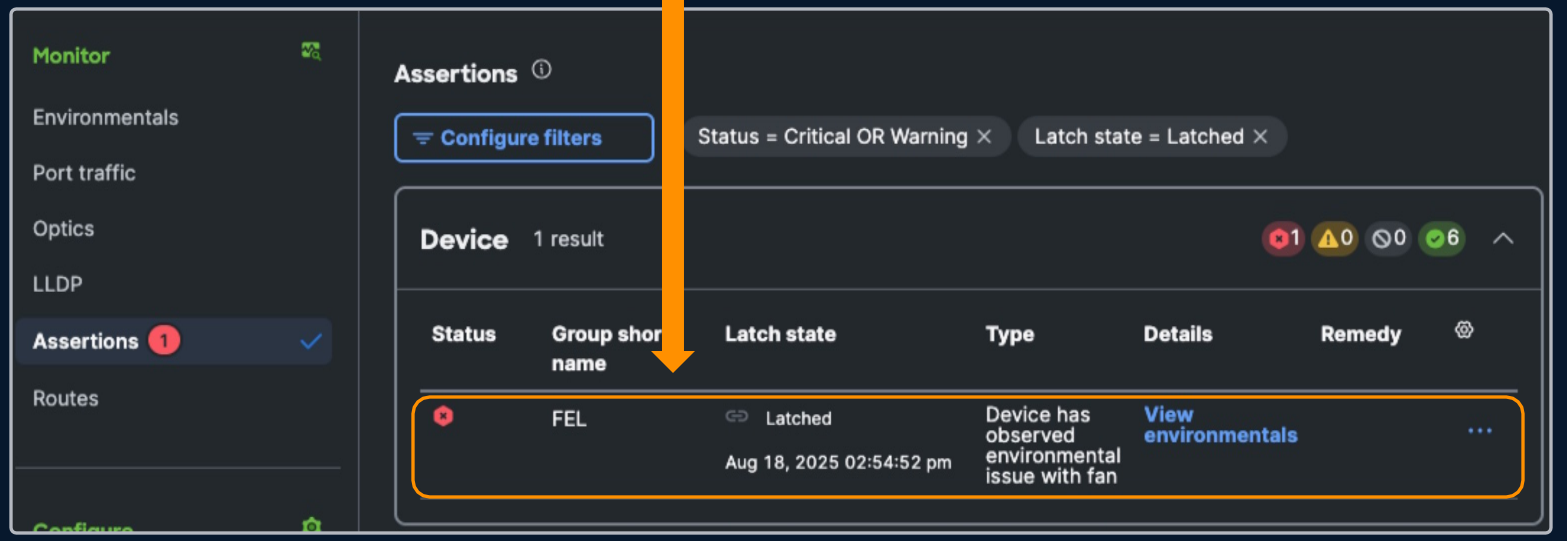
Vertical stack assurance and assertion-based event correlation and proactive alerting

Assurance metrics:

- Environmental
- Port details and statistics
- Digital Optical Monitoring
- L2 and L3 network statistics
- Interconnectivity







Proactive Alerting

- Hyperfabric Dashboard
- Customer E-Mail
- Customer S3 Bucket
- Webhooks (coming soon)

Competitive Reality in DCN

cisco

Yes, it’s a lot…Q/A?

https://www.cisco.com/c/en/us/td/docs/dcn/whitepapers/cisco-data-center-networking-blueprint-for-ai-ml-applications.html

https://www.cisco.com/c/en/us/td/docs/dcn/whitepapers/cisco-addressing-ai-ml-network-challenges.html

CISCO Connect

Thank you

Interested in Learning More?


CISCO