

Why are AI networks so special?

Because there are many!

Lukas Krattiger, Cisco Fellow
@CCIE21921

CISCO Live !

Cisco Webex App

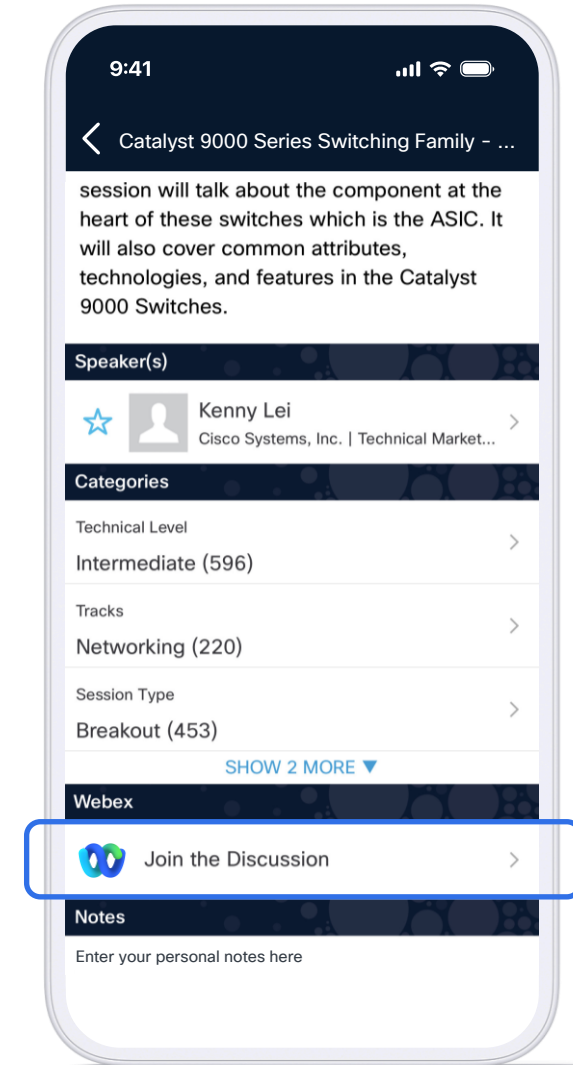
Questions?

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

How

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

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



<https://cislive.ciscoevents.com/cislivebot/#BRKDCN-2539>

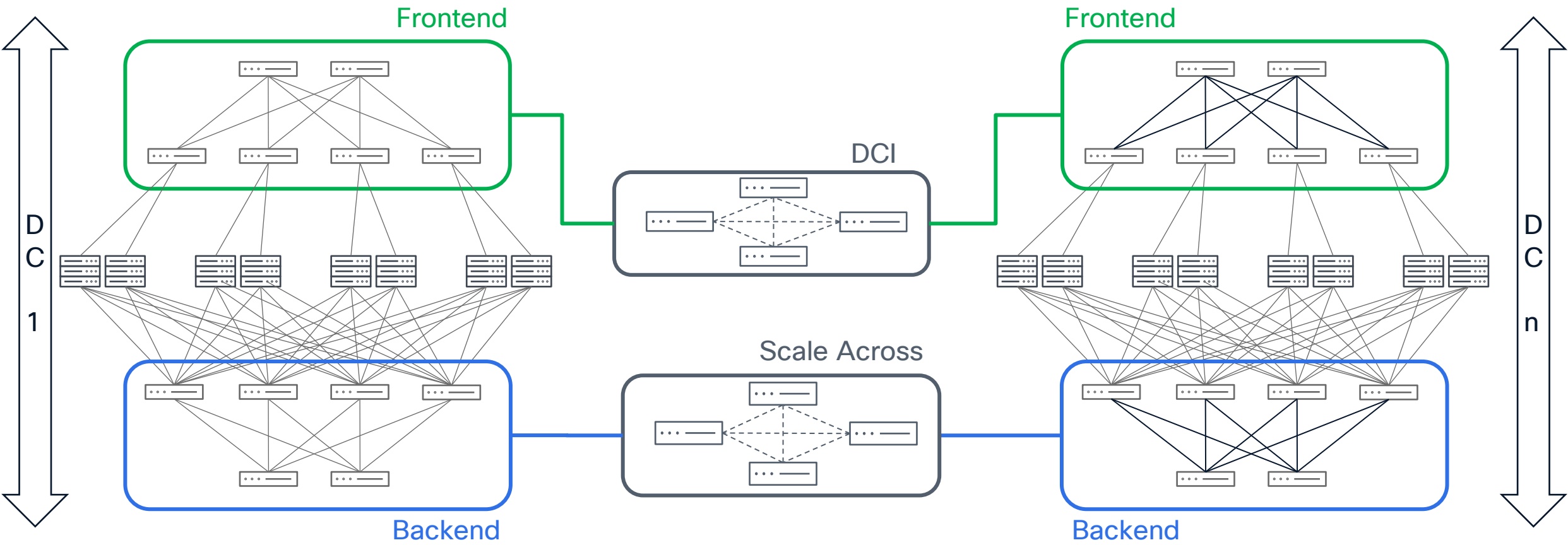
Introduction

- A Snapshot of the AI Network Reference Architecture – how it all comes together
- Server Connectivity Simplified – key requirements at a glance
- The Four Networks – understanding the pillars of AI infrastructure
- Attributes, Requirements & Real-World Challenges – what makes or breaks performance
- Practical Examples in Action – lessons from real implementations
- A Deeper Look into Nexus Dashboard

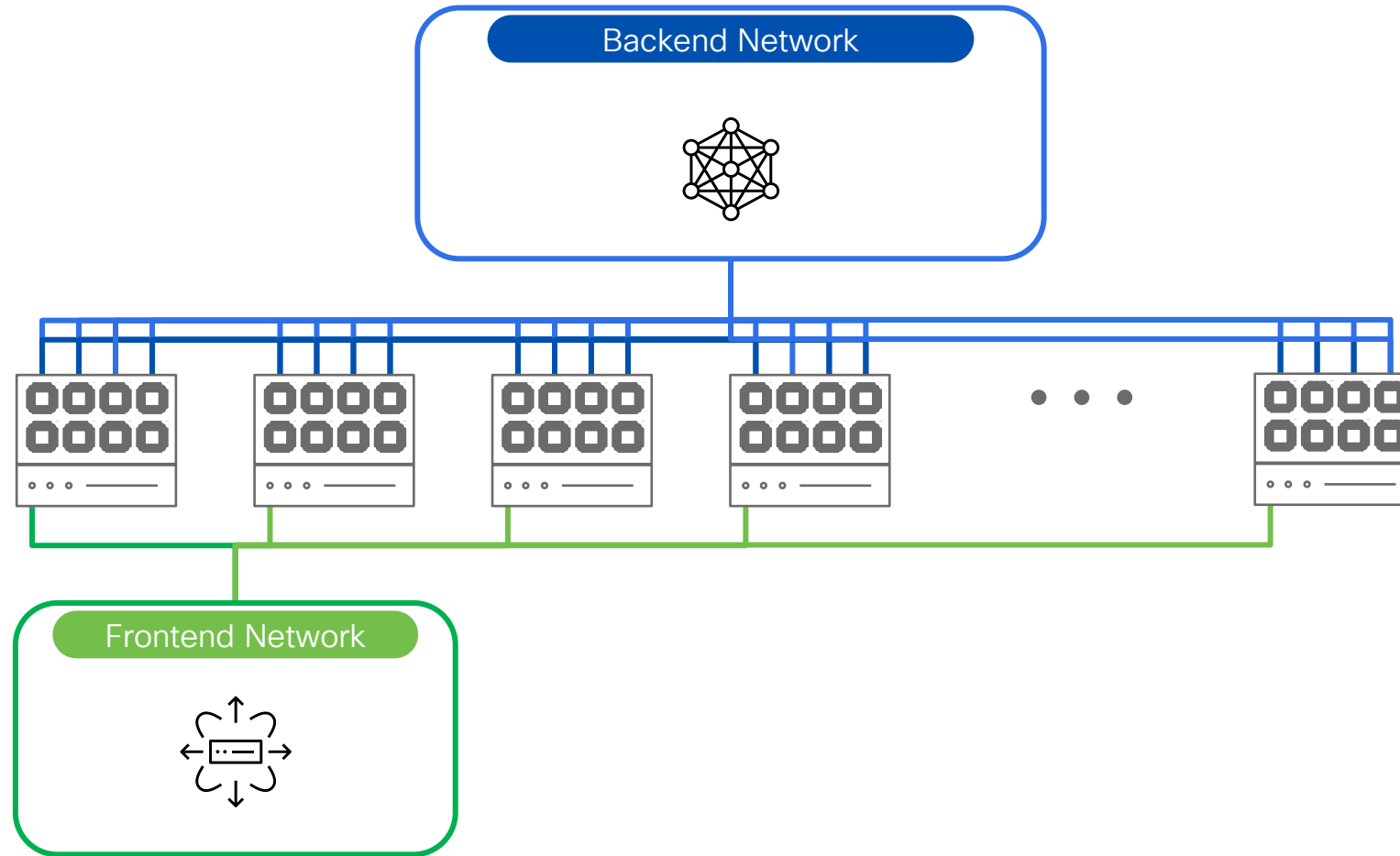
- 01 Overview
- 02 Frontend Network
- 03 Backend Network
- 04 Storage Network
- 05 Management Network
- 06 Operations
- 07 Conclusion

Overview

AI Network Reference Architecture



Multiple Networks for AI/ML Infrastructure

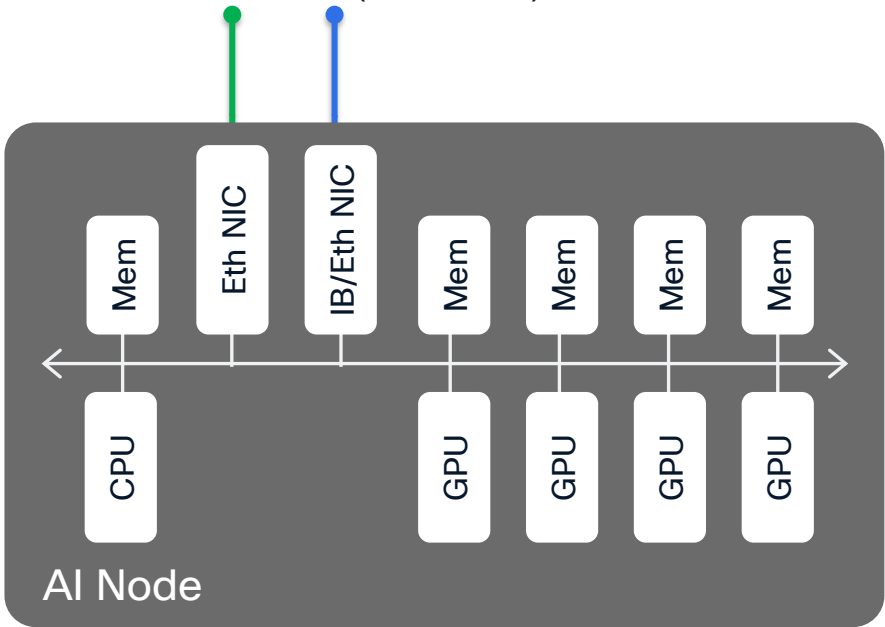


Simplified Server Connectivity



Frontend Network (Ethernet)
(2x 400GbE)

Backend Network (InfiniBand / Ethernet)
(8x 800GbE)



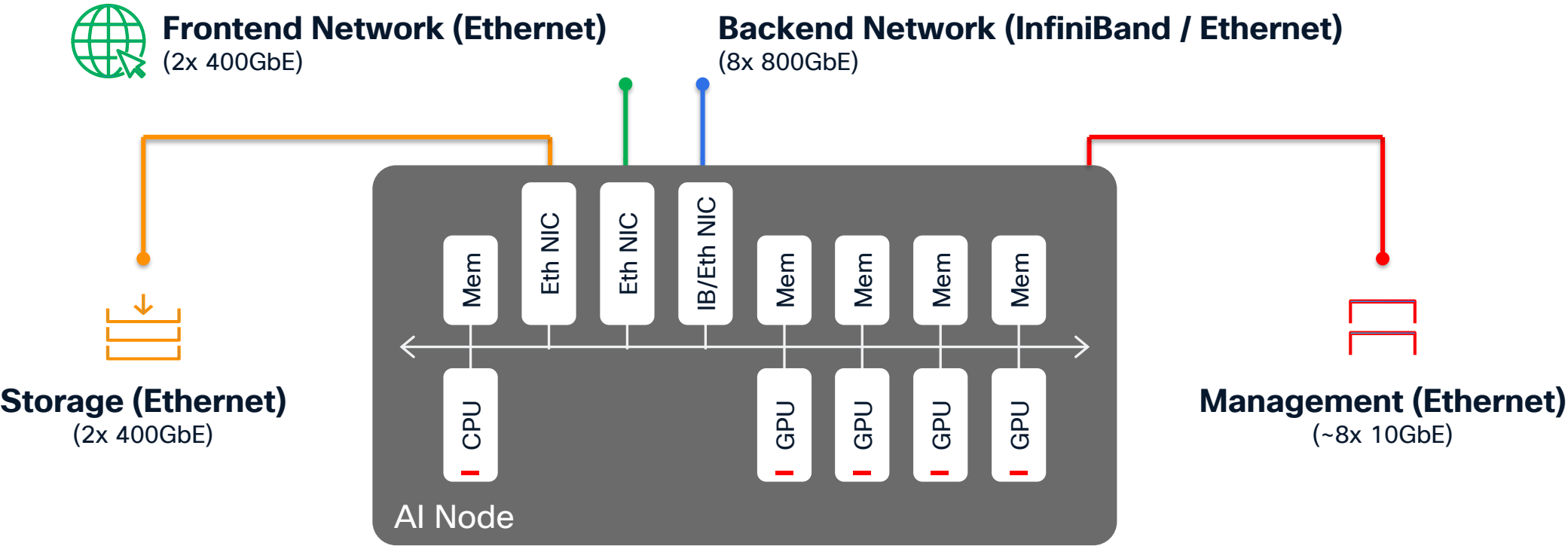
Frontend Network

User Access Connectivity to AI Nodes
Sometimes leveraged for Storage Connectivity

Backend Network

GPU-to-GPU Connectivity between AI Nodes
Utilizes InfiniBand or high-performance Ethernet

Simplified Server Connectivity



Frontend Network

User Access Connectivity to AI Nodes
Sometimes leveraged for Storage Connectivity

Backend Network

GPU-to-GPU Connectivity between AI Nodes
Utilizes InfiniBand or high-performance Ethernet

Storage Network

Data Ingestion, Checkpointing, KV Storage

Management Network

AI Node Management, Image, Firmware, Upgrades

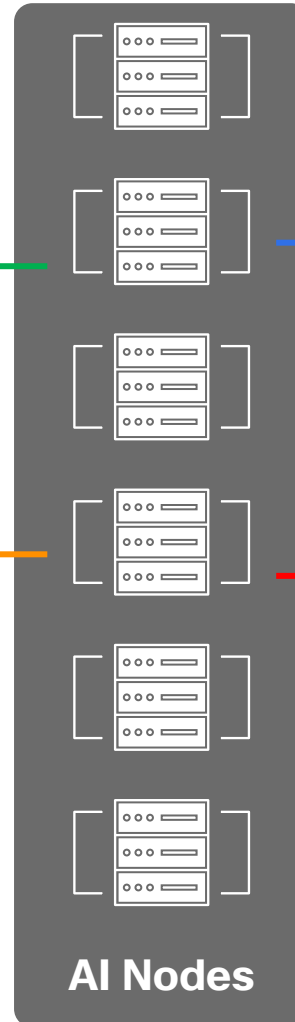
Connectivity Requirements - Detail

Frontend Network

Extended Requirements to the Network Paradigm
(Mobility, Multi-Homing, Bridging and Routing/Layer-2/3)
High Bandwidth, Segmentation and Security

Storage Network

Intermediate Network Paradigm
(Requirements driven by Storage Solution)
Focus on Lossless, Latency, Highest Throughput, and
Advanced Traffic Management



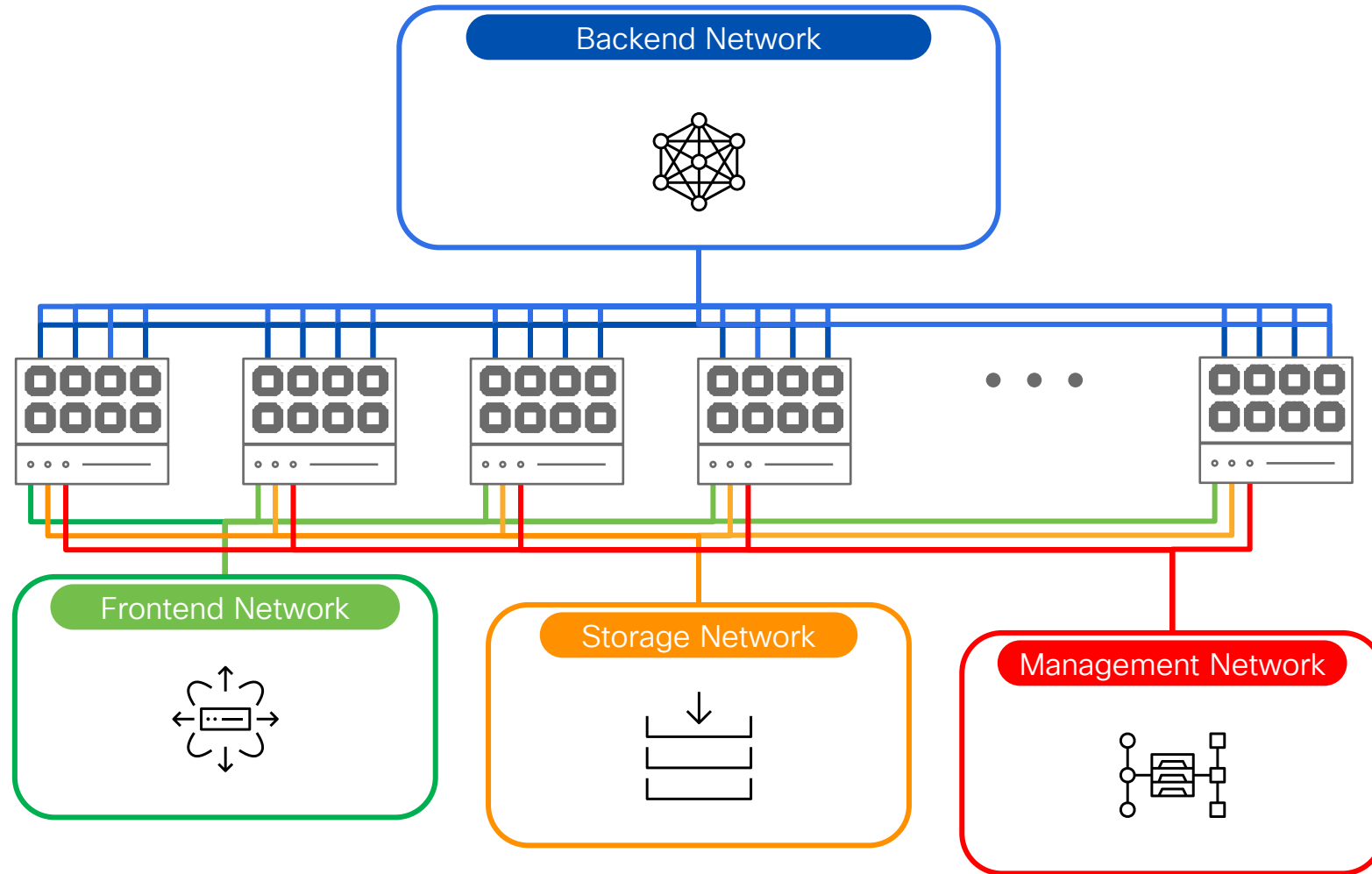
Backend Network

Simple Network Paradigm
(Minimal Requirements on Network Services)
Focus on Lossless, Latency, Highest Throughput, and
Advanced Traffic Management

Management Network

Extended Security Requirements to the Network Paradigm
(Efficient Isolation of Embedded Management Processors)
Highest Segmentation and Security, Intermediate Bandwidth

Multiple Networks for AI/ML Infrastructure



Frontend Network

“The AI Access and Application Plane”

The AI Access and Application Plane

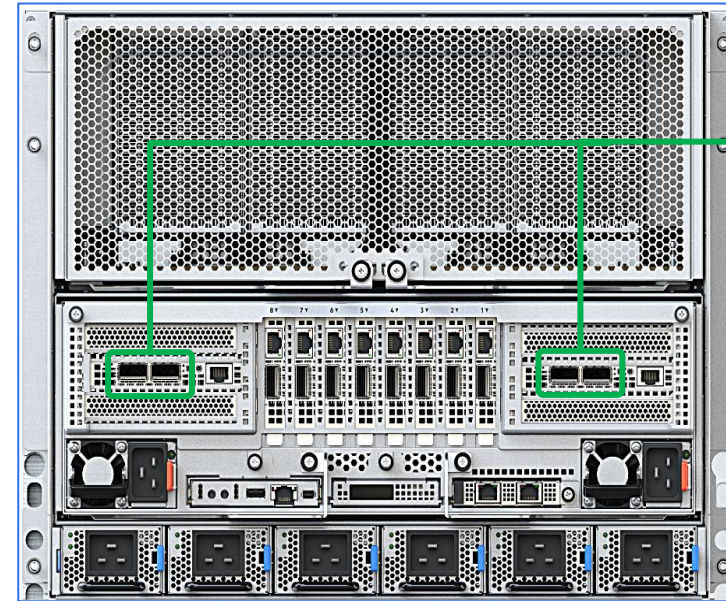
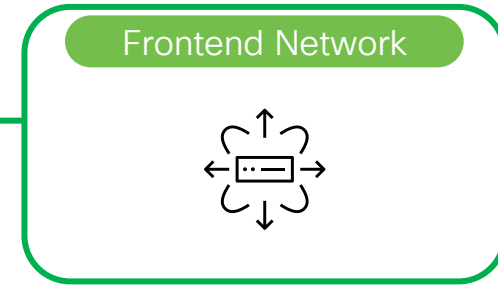
Frontend Network

User & Application Access: Connects users, APIs, and workloads to the AI compute fabric

Control & Coordination Layer: Manages data exchange between applications, GPUs, and storage

Security & Policy Enforcement: Provides segmentation, visibility, and protection for AI services

Operational Consistency: Ensures reliable, high-speed connectivity with unified management



Requirements

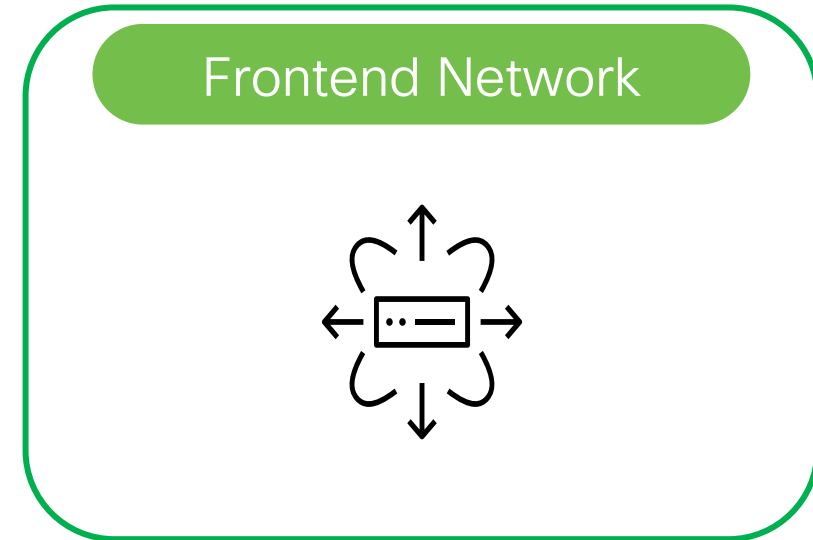
Frontend Network

High-Bandwidth Fabric: Upgraded to 25/100GbE connectivity to handle AI-scale traffic loads

Low-Latency & Resilient Design: Ensures fast convergence and redundancy for uninterrupted access

Cloud-Native Readiness: Supports Kubernetes, microservices, and dynamic workload mobility

Integrated Security & Segmentation: enforcing consistent policy across applications and clusters



Challenges

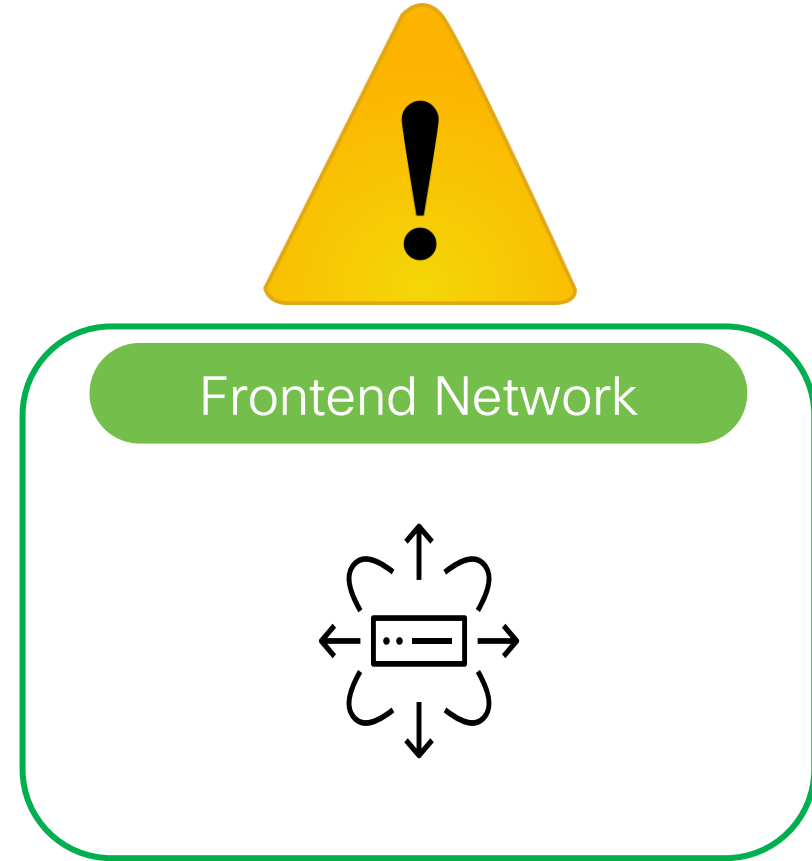
Frontend Network

Bandwidth Scaling: Upgrading from 1/10GbE to 25/100GbE to meet AI data and user traffic demands

Dynamic Workloads: Managing frequent changes accommodating the server environments

Consistent Security: Simplify the consistent enforcing of macro- or micro-segmentation and access policies across diverse applications and clusters

Visibility & Troubleshooting: Correlating app, service, and network performance across hybrid environments



Operational

Frontend Network

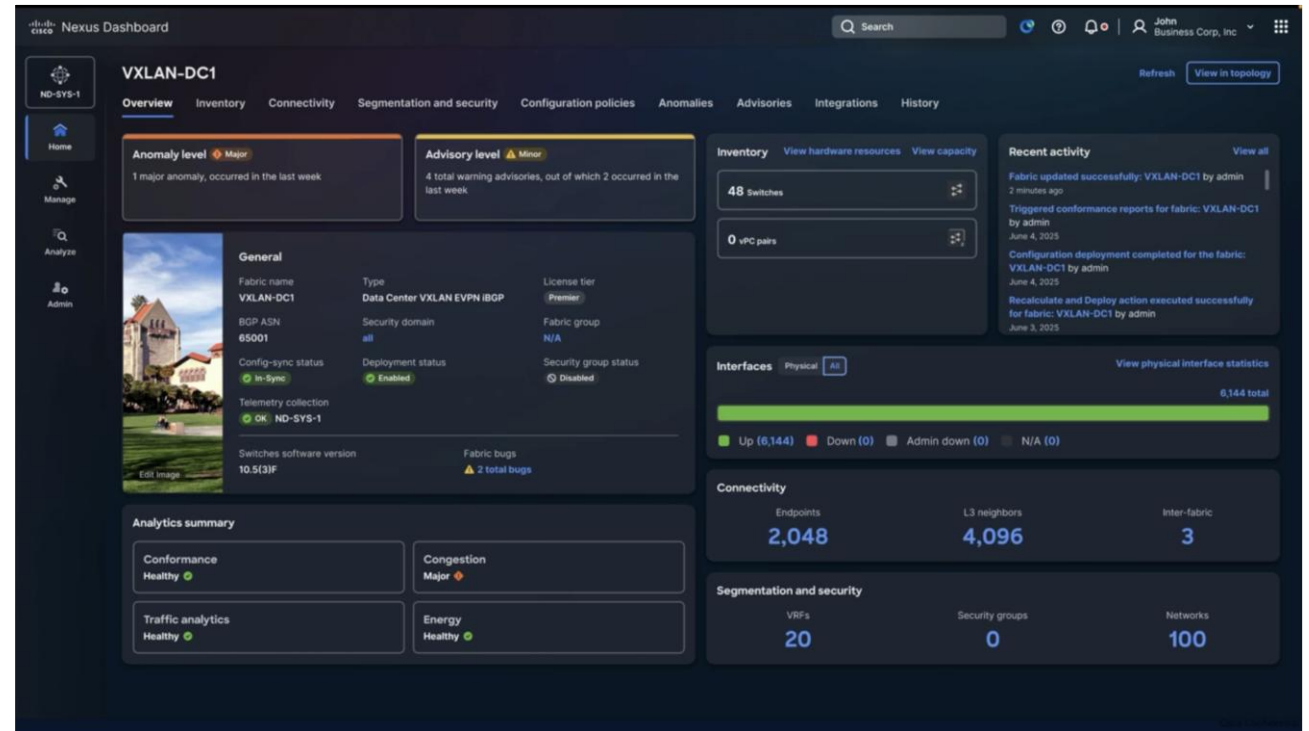
Scalable Fabric Management: Automates deployment and lifecycle of 25/100GbE networks

Unified Policy Model: Ensures consistent segmentation and security across all clusters

Full-Stack Visibility: Correlates app, service, and network telemetry for faster troubleshooting

Integrated Observability: Combines NetOps with tools from Isovalent for Kubernetes-level visibility

Cross-Domain Automation: Delivers consistent operations and compliance across on-prem and cloud sites



Backend Network

“The AI Compute Fabric”

The AI Compute Fabric

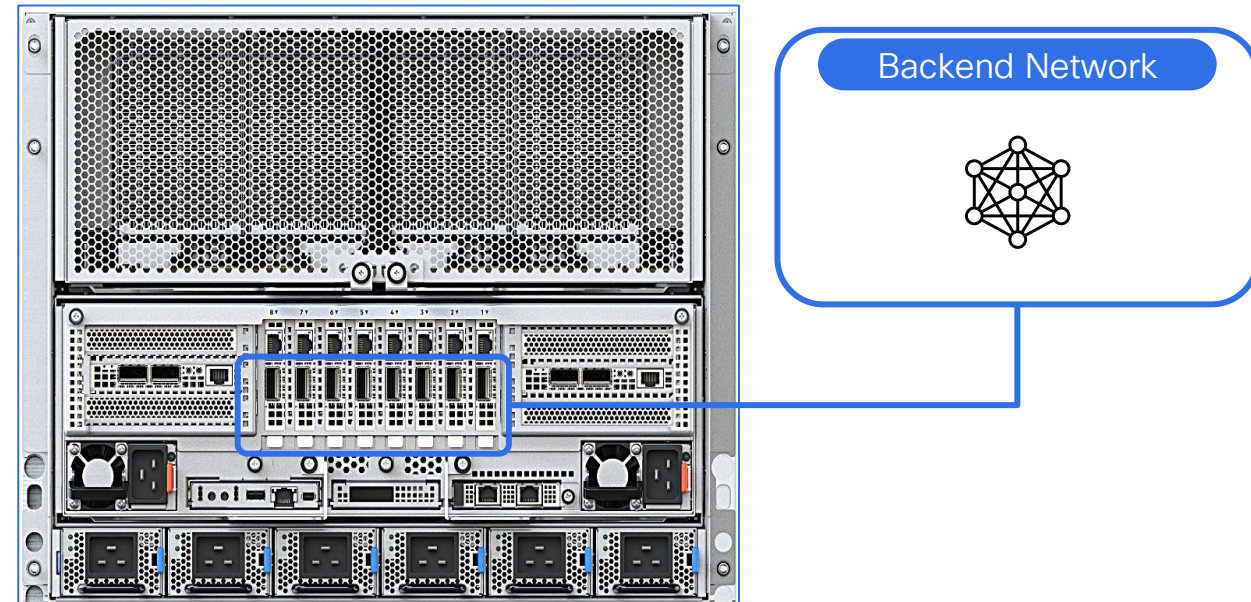
Backend Network

High-Speed Data Exchange: Enables massive, low-latency communication between GPUs during training and inference

Scalable Fabric Backbone: Provides the foundation for cluster growth – scaling up and out without performance loss

Performance Optimization: Ensures lossless, congestion-free traffic flow for deterministic AI workloads

Telemetry & Assurance: Delivers real-time visibility and closed-loop optimization

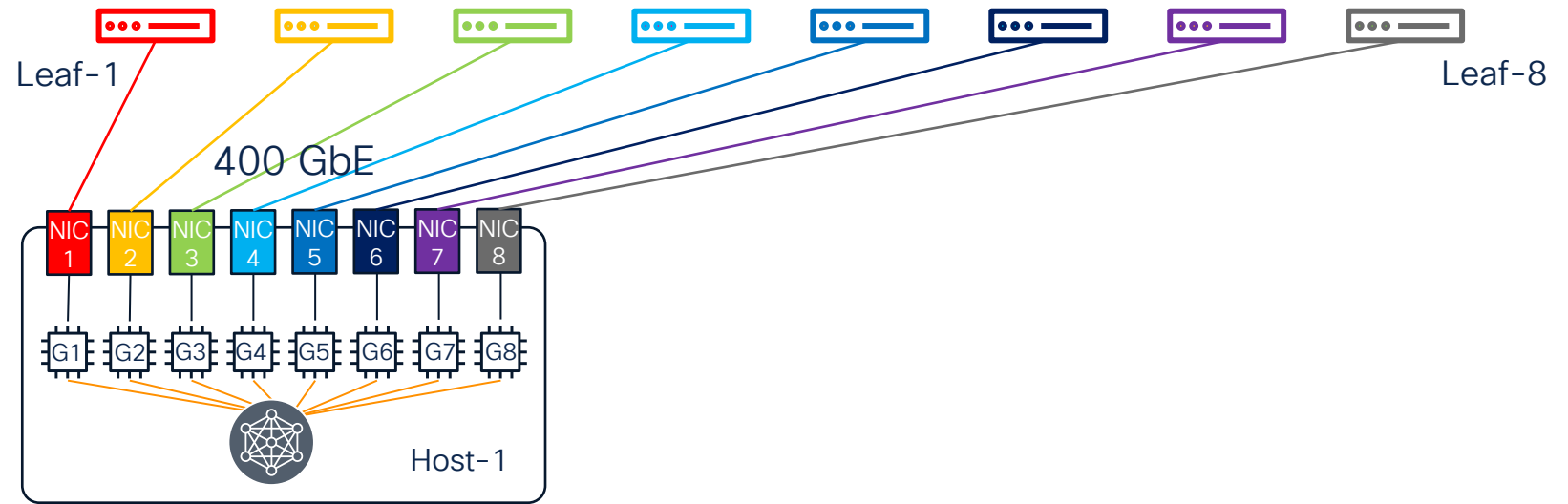


Non-Blocking Rail Only

Backend Network

Non-Blocking

32 ports connected
to GPU hosts



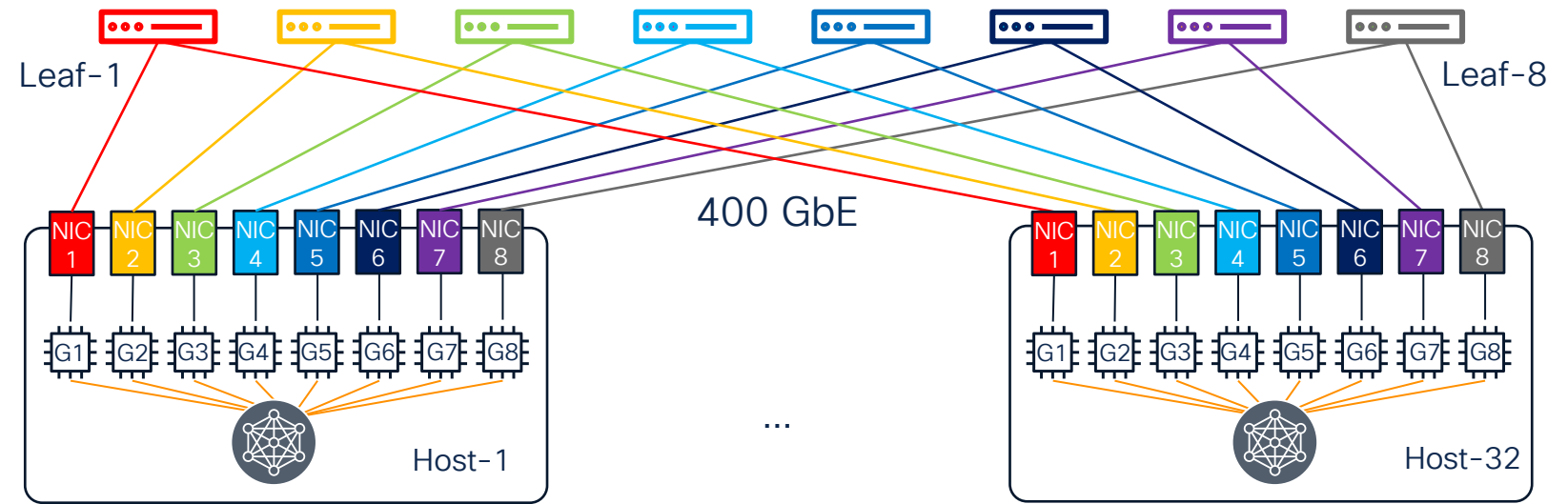
Port 1 on all hosts connects to Leaf-1
Port 2 on all hosts connects to Leaf-2, and so on

Non-Blocking Rail Only

Backend Network

Non-Blocking

32 ports connected
to GPU hosts



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

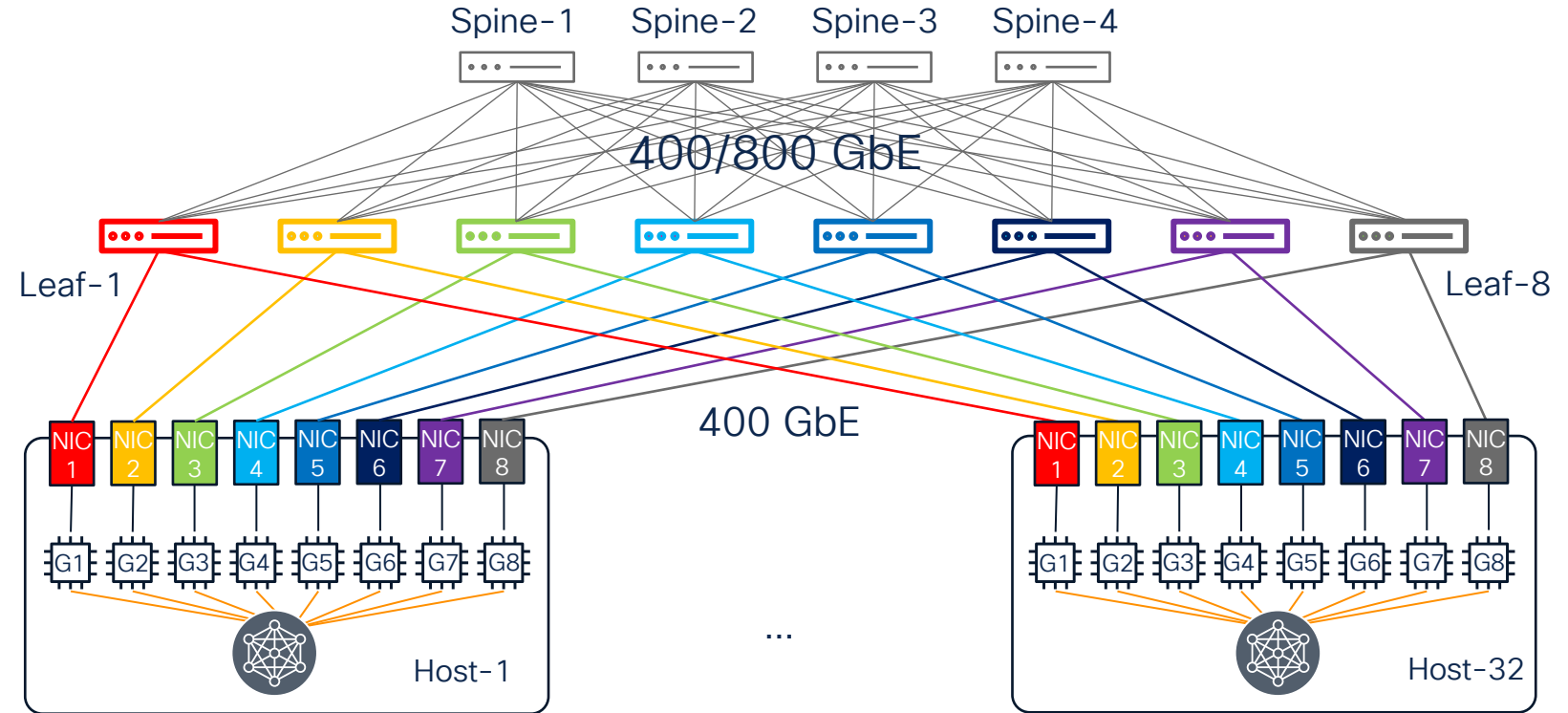
Non-Blocking Rail Optimized

Backend Network

Non-Blocking

32 ports connected
to spine switches

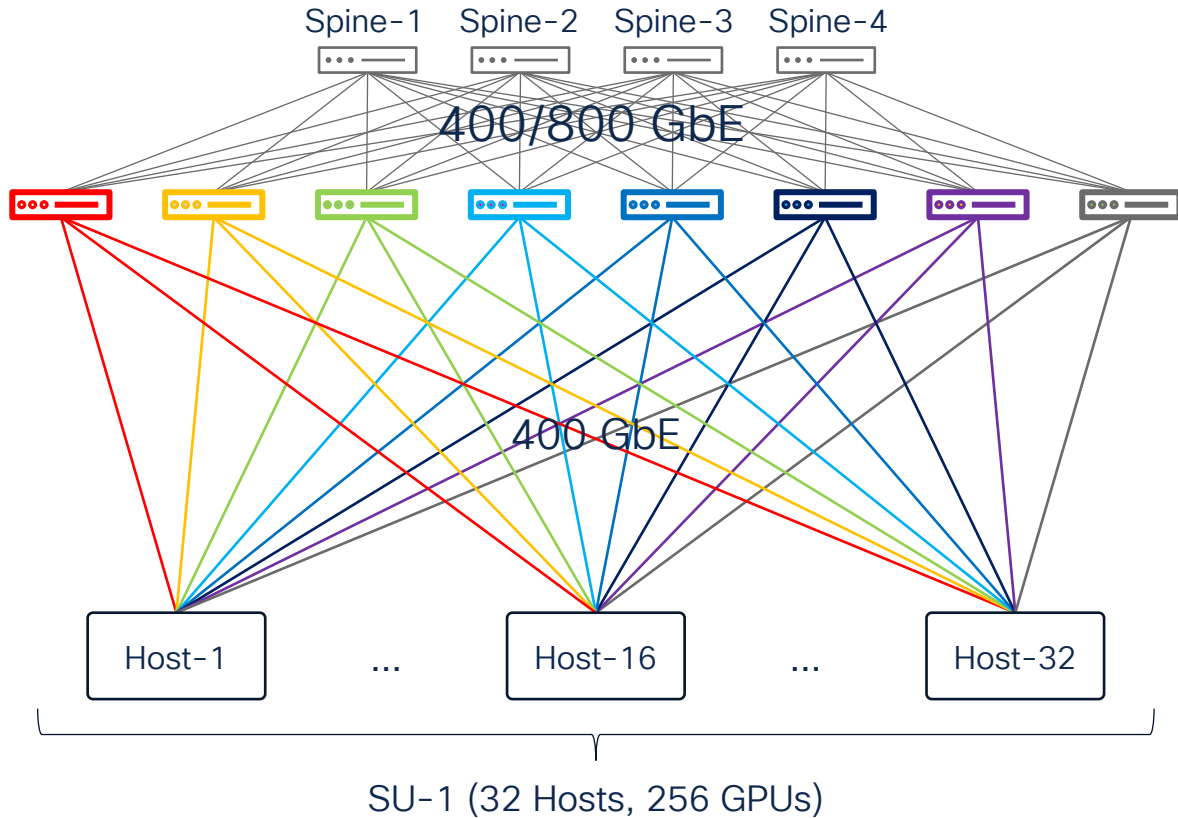
32 ports connected
to GPU hosts



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

Non-Blocking Rail Optimized

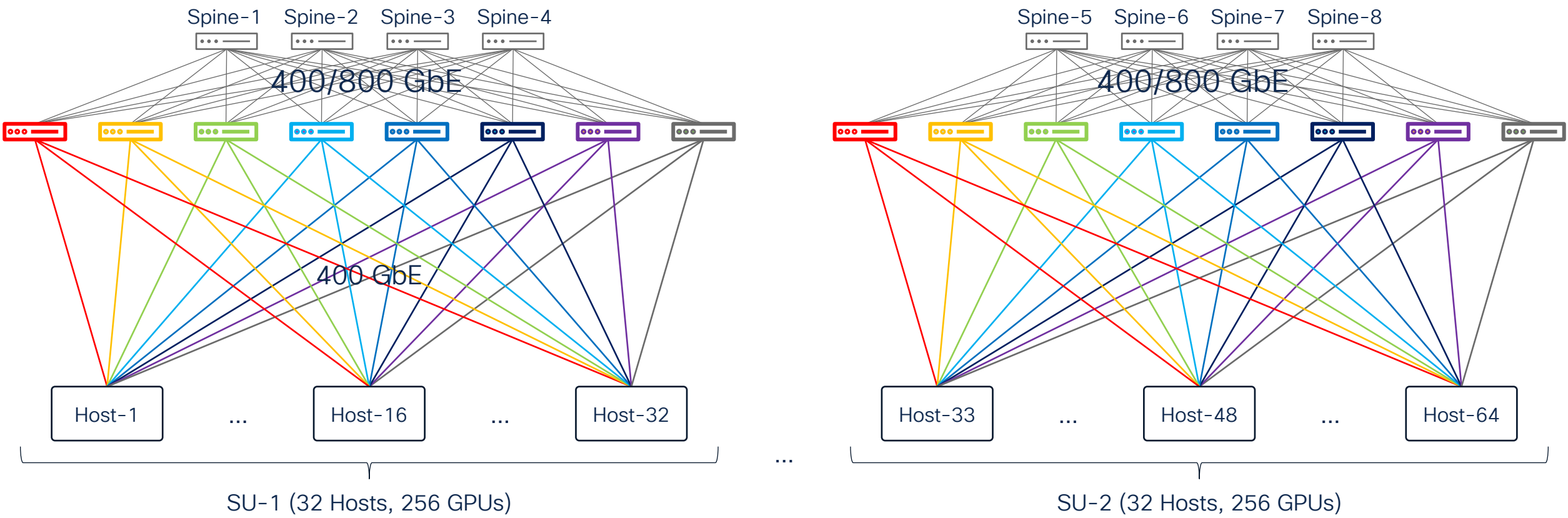
Backend Network



Add more Scale Unites (SU - each with 256 GPUs) and cables between leaf and spine switches for larger SuperPOD (up to 2048 GPUs)

Non-Blocking Rail Optimized

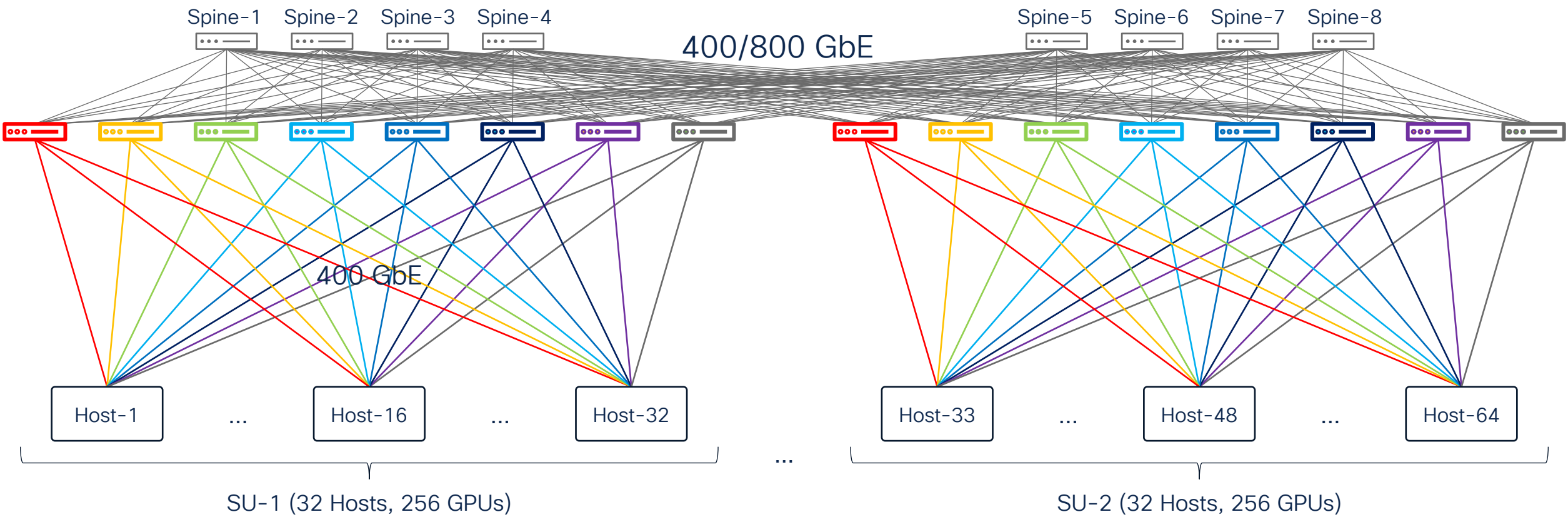
Backend Network



Add more Scale Unites (SU - each with 256 GPUs) and cables between leaf and spine switches for larger SuperPOD (up to 2048 GPUs)

Non-Blocking Rail Optimized

Backend Network



Add more Scale Unites (SU - each with 256 GPUs) and cables between leaf and spine switches for larger SuperPOD (up to 2048 GPUs)

Design Characteristics

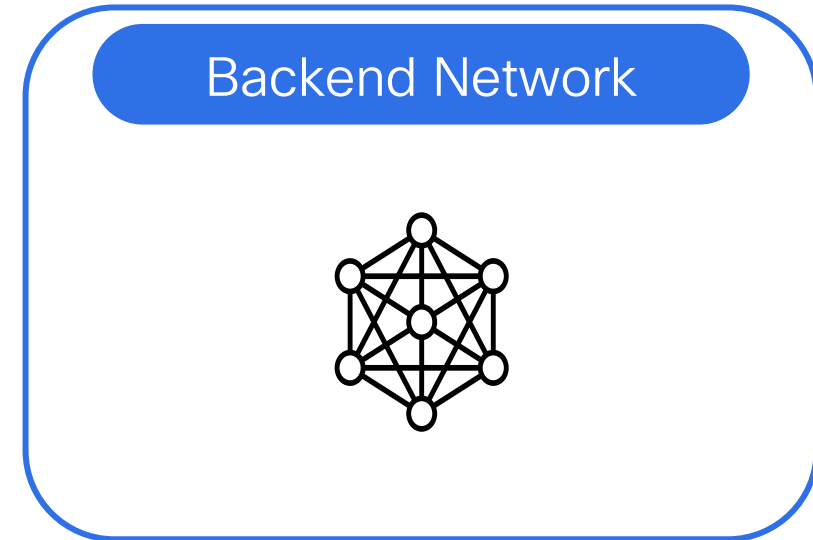
Backend Network

Ultra-Low Latency: Optimized for fast, lossless GPU-to-GPU communication using RoCEv2 on top of a routed or VXLAN EVPN fabric

Non-Blocking Fabric: Built with zero or minimal oversubscription to sustain peak throughput

Scalable Topology: Employs Clos designs to support large-scale “pay-as-you-grow” AI cluster

Deterministic Performance: Utilizes advanced congestion management and QoS tuning for predictable results



Challenges

Backend Network

Ultra-Low Latency Requirements:

Maintaining lossless, high-speed communication for GPU-to-GPU traffic

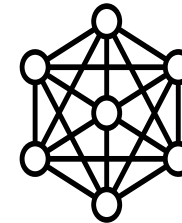
Congestion Control: Managing bursty, high-throughput flows without impacting training performance

Fabric Scalability: Expanding non-blocking architectures as GPU clusters grow in size and density

Operational Complexity: Tuning RoCE, QoS, and buffer settings consistently across thousands of links



Backend Network



Operational

Backend Network

Fabric Automation: Simplifies provisioning and lifecycle of large-scale GPU fabrics

Performance Insights: Monitors latency, loss, and congestion with real-time telemetry

Closed-Loop Optimization: Automates detection and remediation of hotspots

Consistent Policy Enforcement: Standardizes QoS and RoCE configurations across nodes

Unified Operations: Provides a single view of fabric health, capacity, and performance trends



Storage Network

“The Data Lifeline of AI”

The Data Lifeline of AI

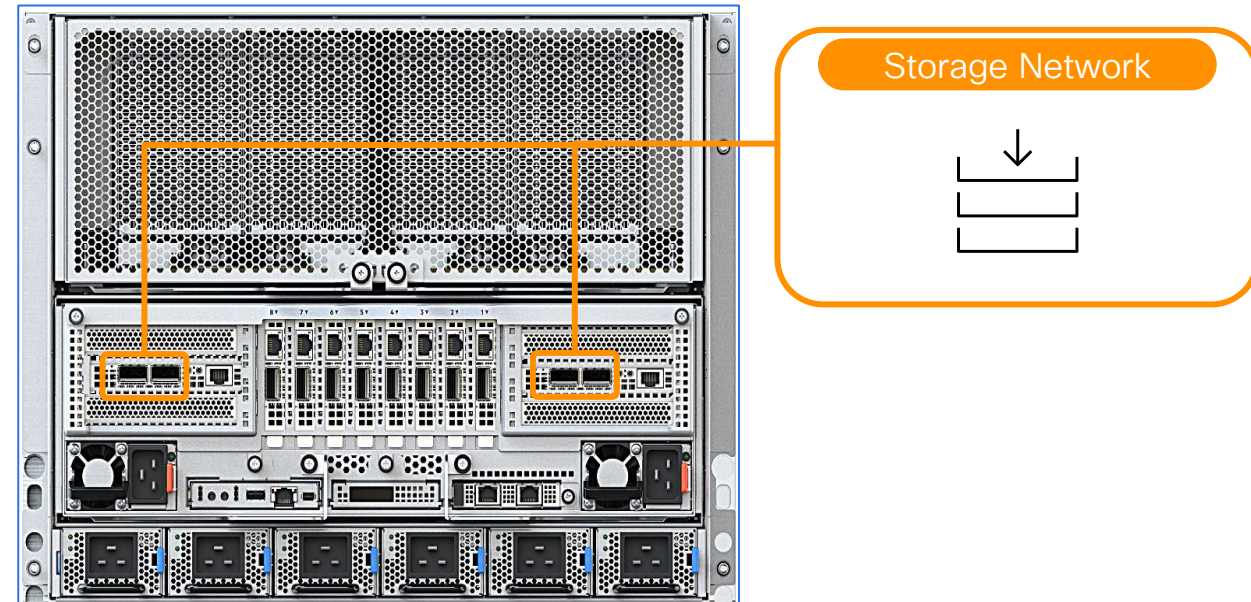
Storage Network

Data Pipeline Backbone: Delivers high-throughput, low-latency access to training and inference datasets

Sustained GPU Utilization: Ensures GPUs stay fully fed with data to maximize compute efficiency

Scalable Performance: Supports expanding storage tiers and capacity as AI workloads grow

End-to-End Visibility: Enables correlation of storage and network performance via tooling



Design Characteristics

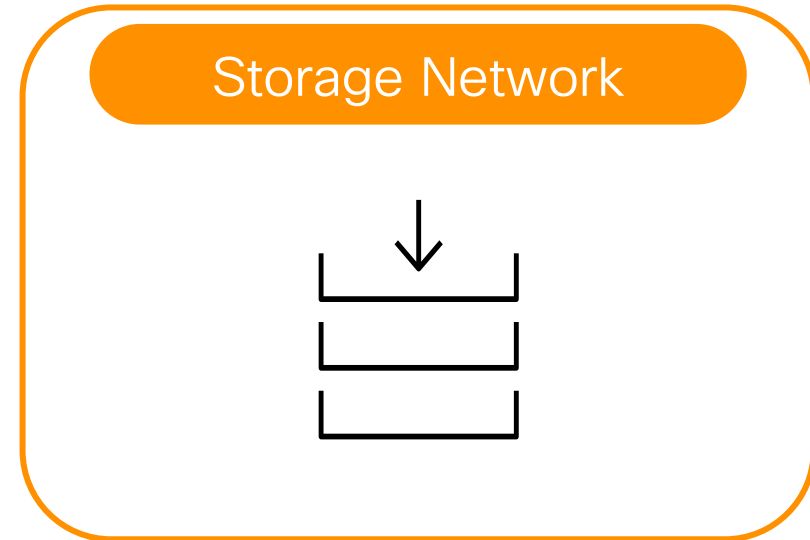
Storage Network

High Throughput & Low Latency: Designed to sustain continuous data flow for training and inference

Lossless Fabric: Optimized for NVMe-oF and RDMA protocols to ensure reliable, efficient data access

Scalable Architecture: Supports expansion as data volumes and AI workloads grow

Balanced Performance: Implements QoS and traffic management to prevent storage or I/O hotspots



Challenges

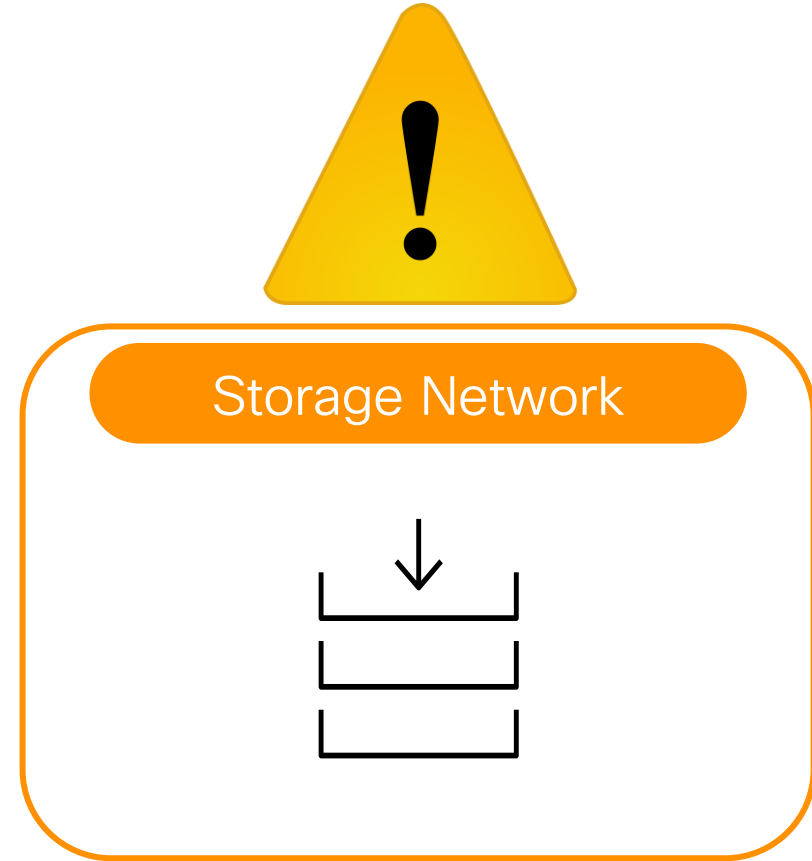
Storage Network

Performance Bottlenecks: Maintaining high throughput and low latency for data-intensive AI workloads

I/O Imbalance: Managing uneven traffic patterns that create hotspots across storage nodes

Scalability Limits: Expanding capacity and bandwidth seamlessly as datasets and clusters grow

Complex Tuning: Optimizing RDMA, QoS, and buffer parameters to ensure consistent performance



Operational

Storage Network

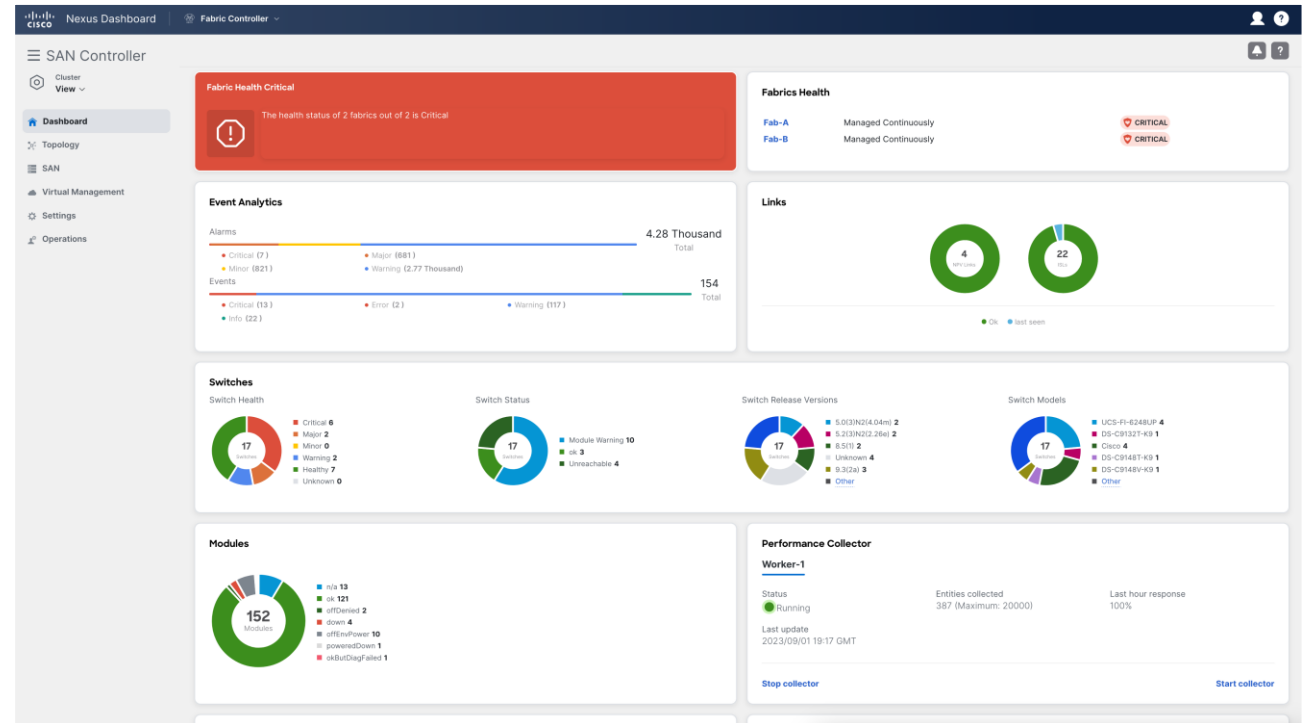
Performance Monitoring: Tracks throughput, latency, and packet loss across storage paths

Hotspot Detection: Identifies I/O bottlenecks and traffic imbalances in real time

Automated Optimization: Enforces consistent QoS, ECN, and buffer policies across fabrics

Scalable Orchestration: Simplifies deployment and expansion of storage networks as demand grows

Unified Visibility: Correlates storage, compute, and network telemetry for end-to-end assurance



Management Network

“The Hidden Backbone”

The Hidden Backbone

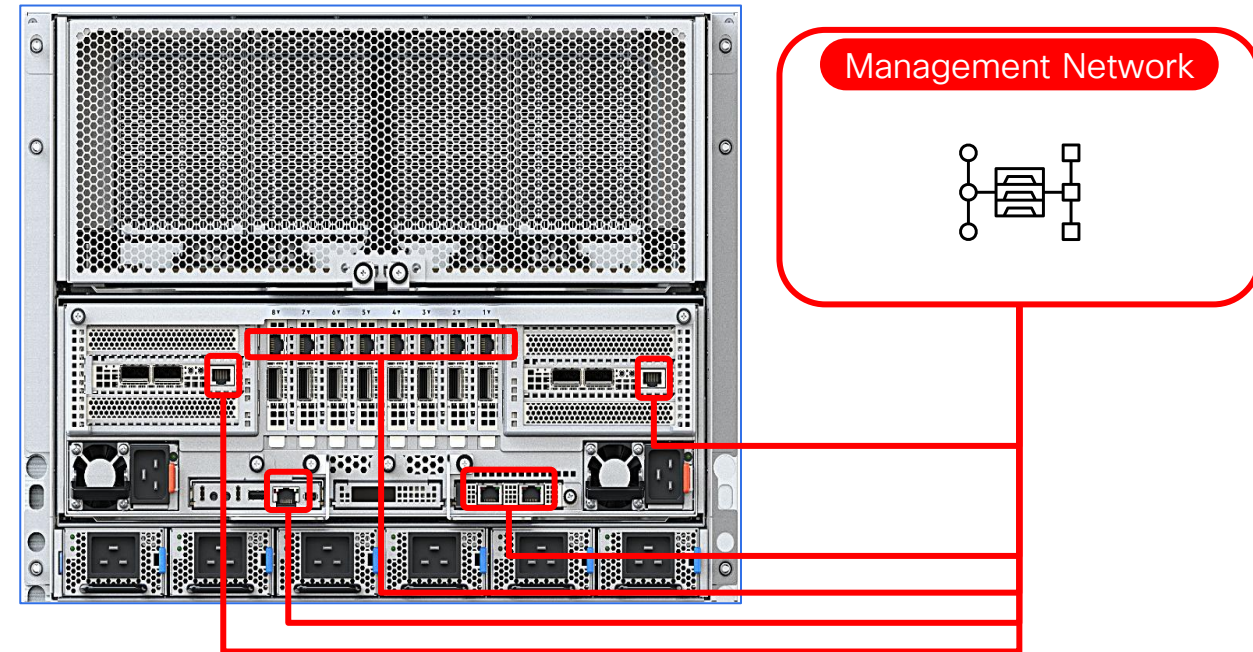
Management Network

Infrastructure Control Plane: Connects and manages servers, GPUs, DPUs, and baseboard controllers

Security & Isolation: Segments management traffic to protect against lateral movement and attacks

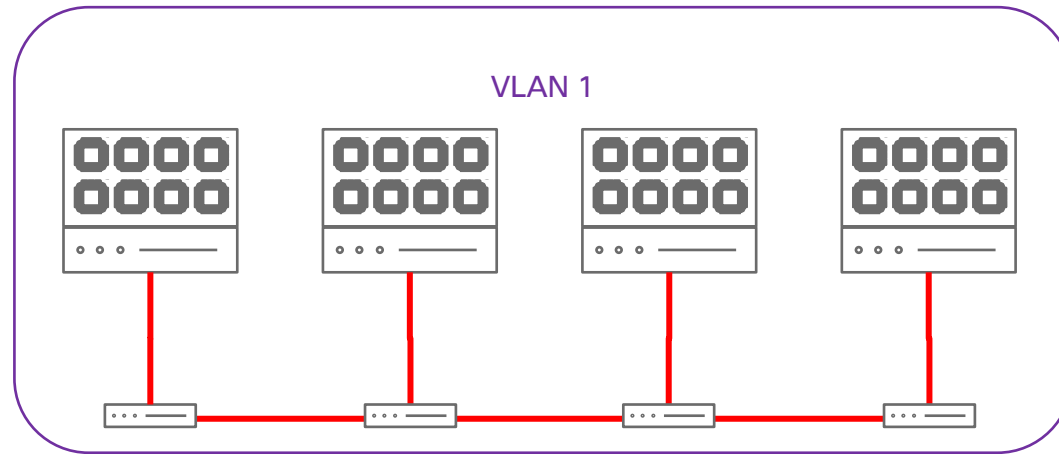
Operational Reliability: Ensures consistent access for monitoring, updates, and automation tools

Centralized Oversight: Provides unified visibility and compliance assurance through tooling

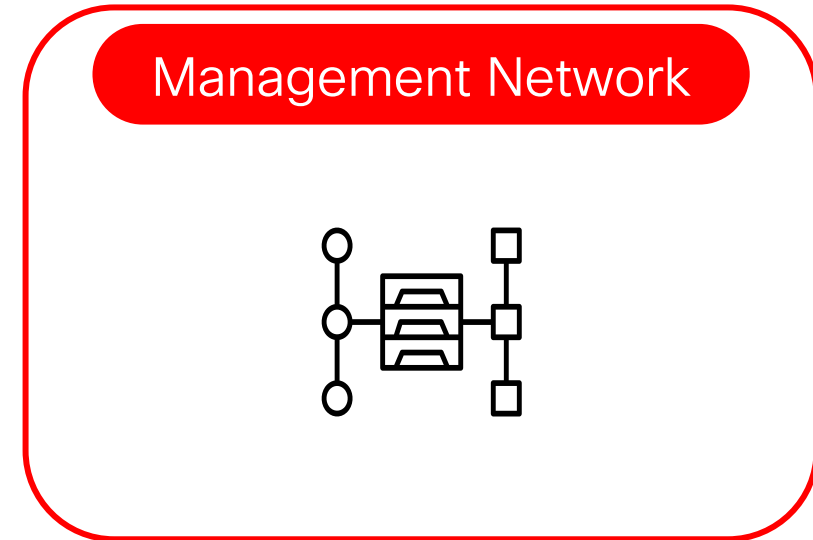


Design Characteristics

Management Network

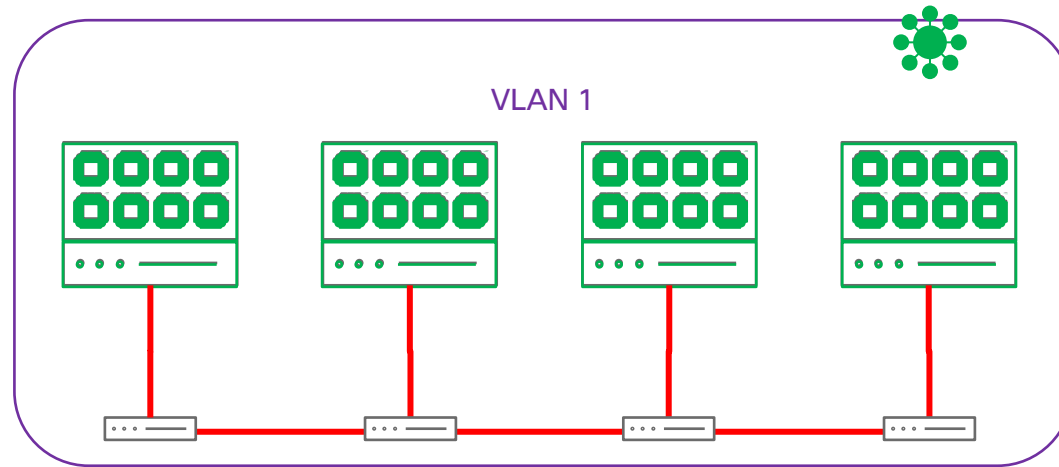


Traditional Layer-2 OOB Network

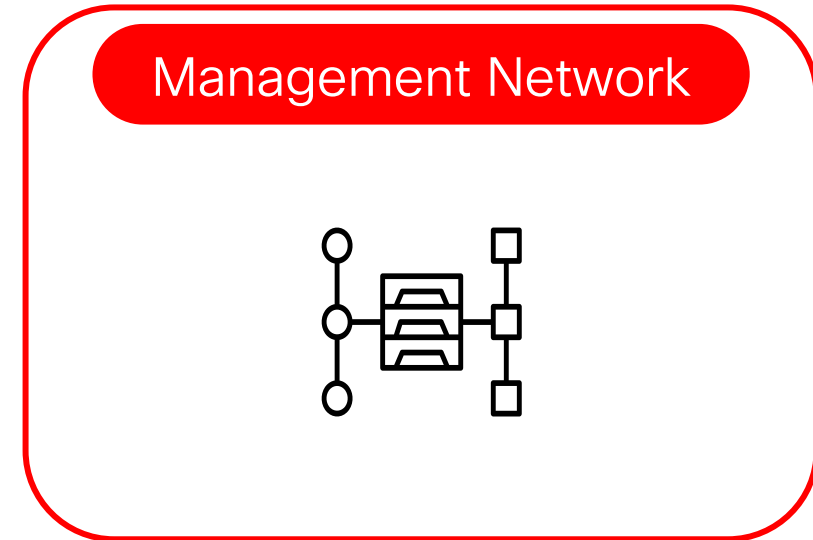


Design Characteristics

Management Network



Traditional Layer-2 OOB Network



Design Characteristics

Management Network

Design Evolution: Well structured L3-based network (no longer flat L2)

Segmentation & Security: Built with isolated security zones to separate management traffic and prevent lateral movement

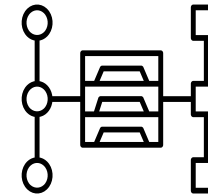
Management Processers can't be secured inside (SW exposure)

East-West isolation between GPUs' Management ports

North-South access control on the Network

High Port Density: Supports multiple interfaces per server (BMCs, GPUs, DPUs, baseboards) for full infrastructure control

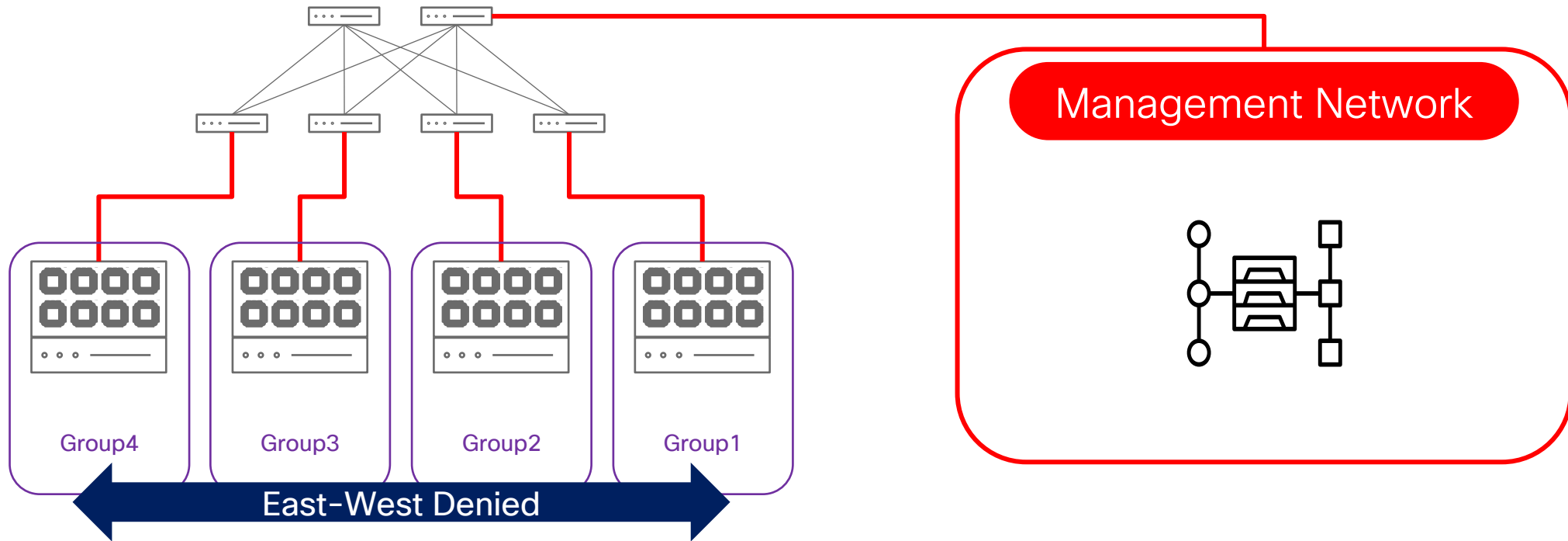
Management Network



Design Characteristics

Management Network

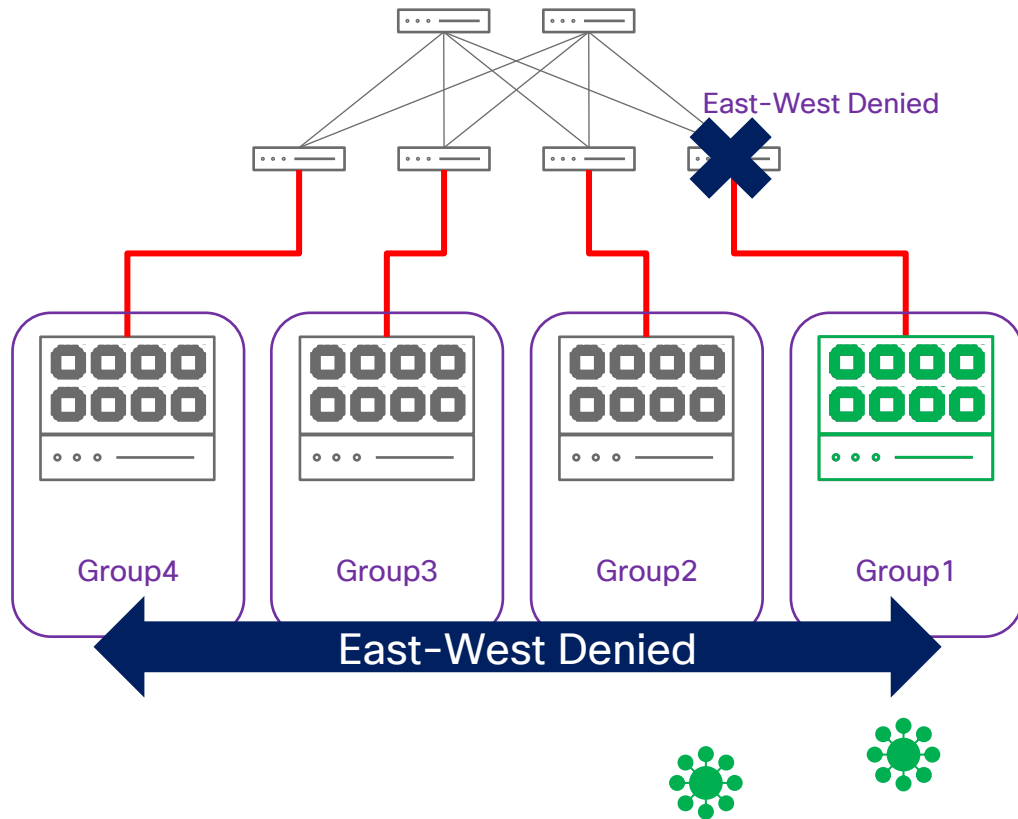
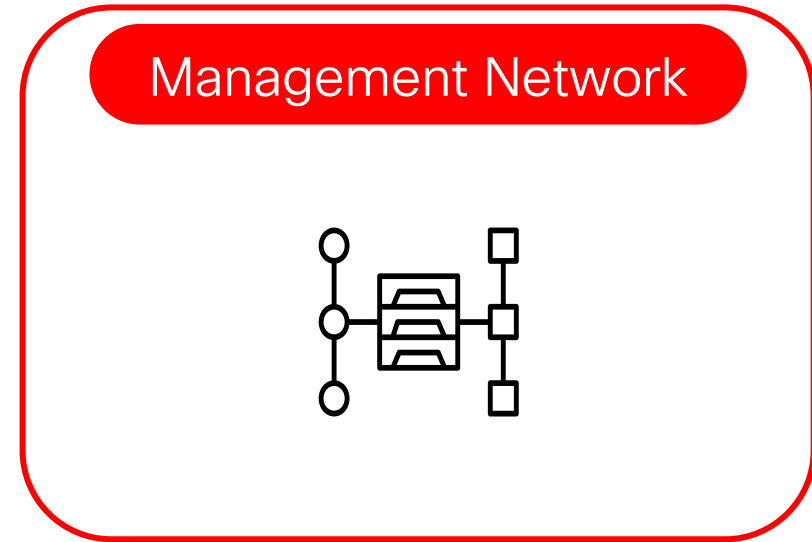
Structured DC Fabric (with Security)



Design Characteristics

Management Network

Structured DC Fabric (with Security)



Challenges

Management Network

Port Proliferation: Managing a rapid increase from a few to 10–12 management interfaces per server

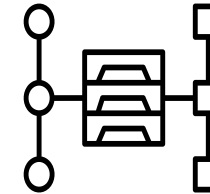
Security Exposure: Handling numerous endpoints, often running unmanaged or lightweight OSs

Segmentation Complexity: Moving from flat Layer-2 designs to secure, scalable segmented architectures

Operational Overhead: Maintaining consistency, reliability, and isolation across thousands of management links



Management Network



Operational

Management Network

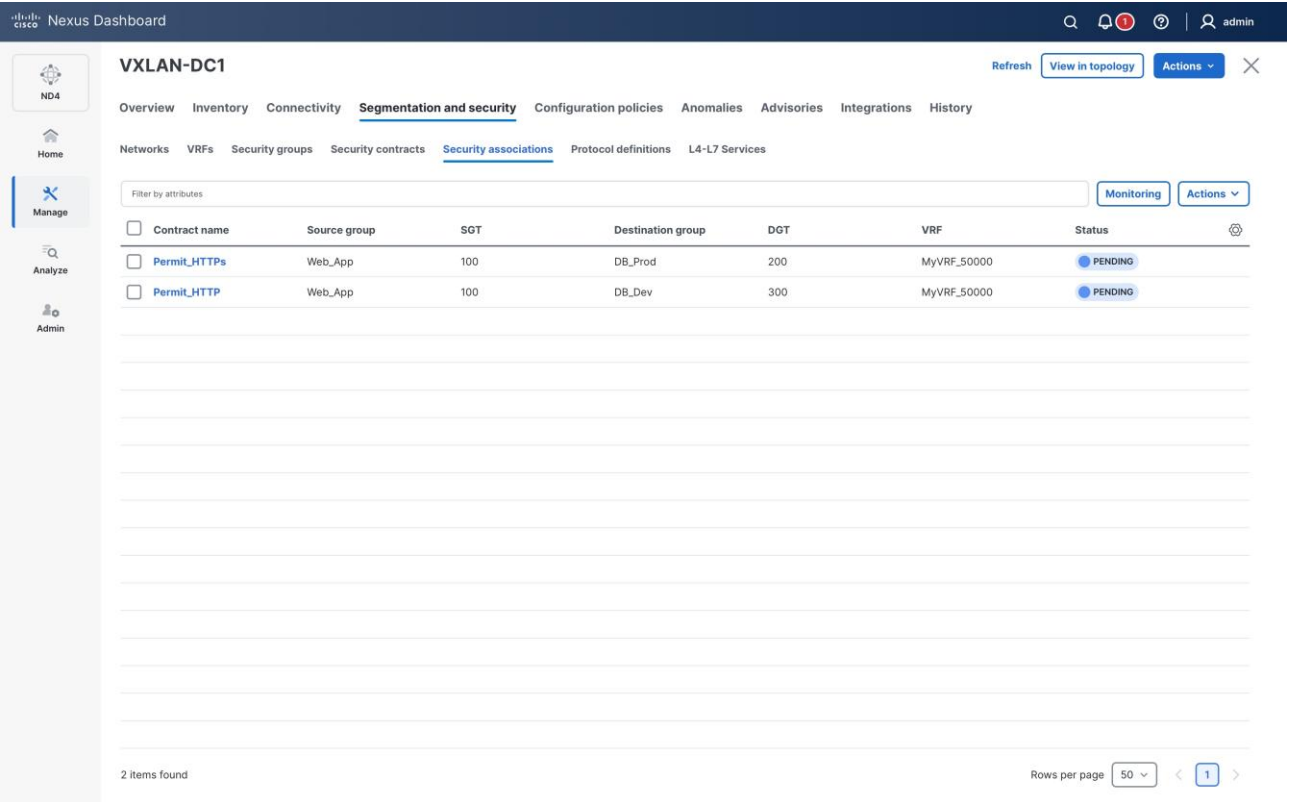
Segmentation at Scale: Automates VRF, VLAN, and ACL deployment for secure isolation

Anomaly Detection: Identifies unauthorized or unexpected management traffic

Compliance Validation: Continuously audits configurations against security baselines

Centralized Lifecycle Control: Simplifies firmware, config, and network updates across domains

Unified Visibility: Provides a single view of all management connections and device states



The screenshot shows the Cisco Nexus Dashboard interface for VXLAN-DC1. The 'Segmentation and security' tab is active, and the 'Security associations' sub-tab is selected. The table displays the following data:

Contract name	Source group	SGT	Destination group	DGT	VRF	Status
Permit_HTTPs	Web_App	100	DB_Prod	200	MyVRF_50000	PENDING
Permit_HTTP	Web_App	100	DB_Dev	300	MyVRF_50000	PENDING

Operations

Nexus Dashboard

Provision

Classic

2- or 3-tier NX-OS architectures

VXLAN

BGP EVPN fabrics for Nexus & Catalyst (IOS XE)

AI/ML

Fine-tuned network provisioning to transport AI/ML apps with the best performance

External and Inter-Fabric Networks

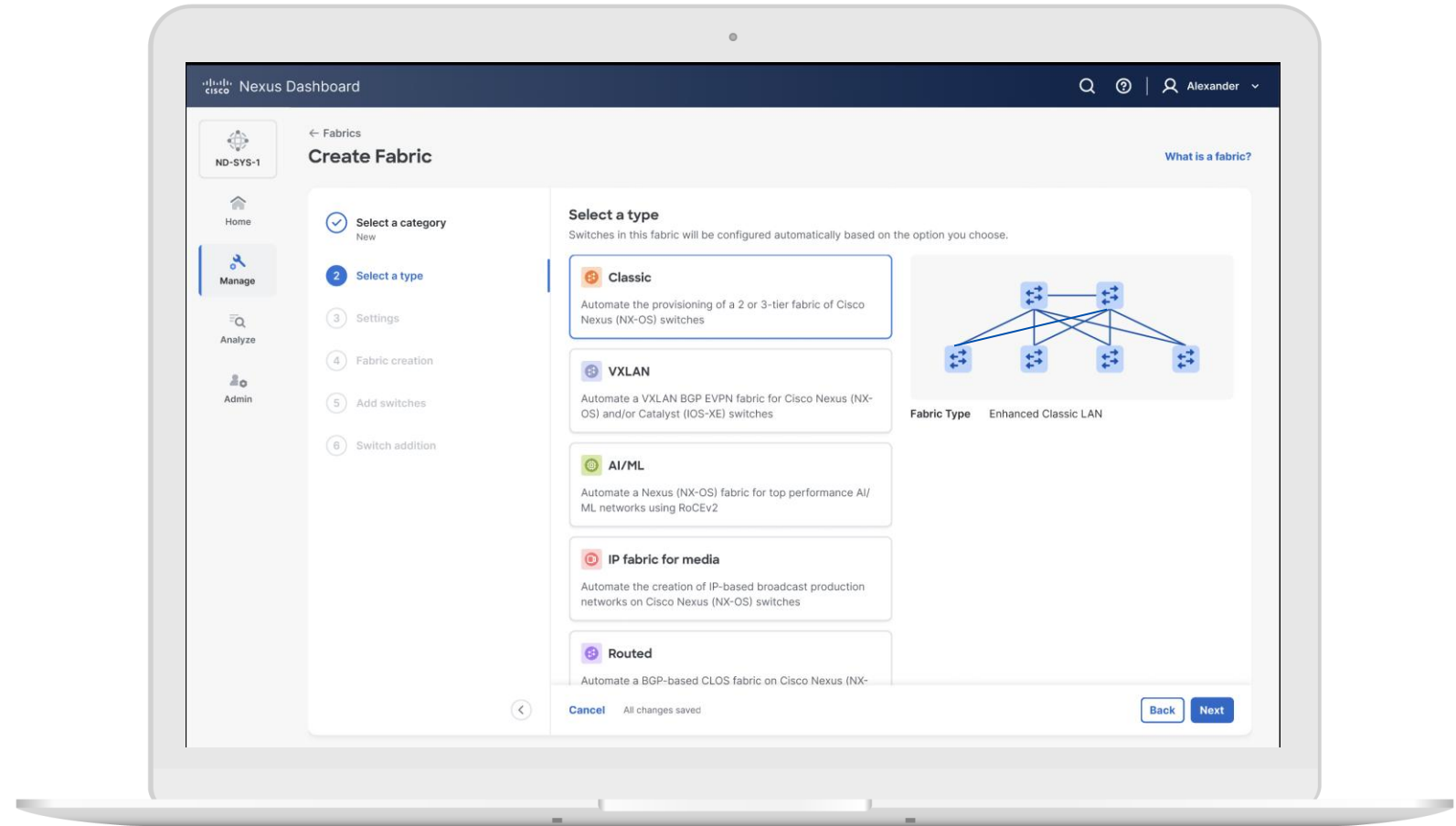
NX-OS, IOS XE, XR, third-party networks mainly used to interconnect ACI, NX-OS, and Campus

IP Fabric for Media

Fine-tuned network provisioning and monitoring for broadcasting and media

Routed

BGP-based CLOS NX-OS fabrics



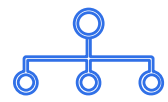
Provision all your fabrics from a single place

Nexus Dashboard



Visualize

Nexus Dashboard



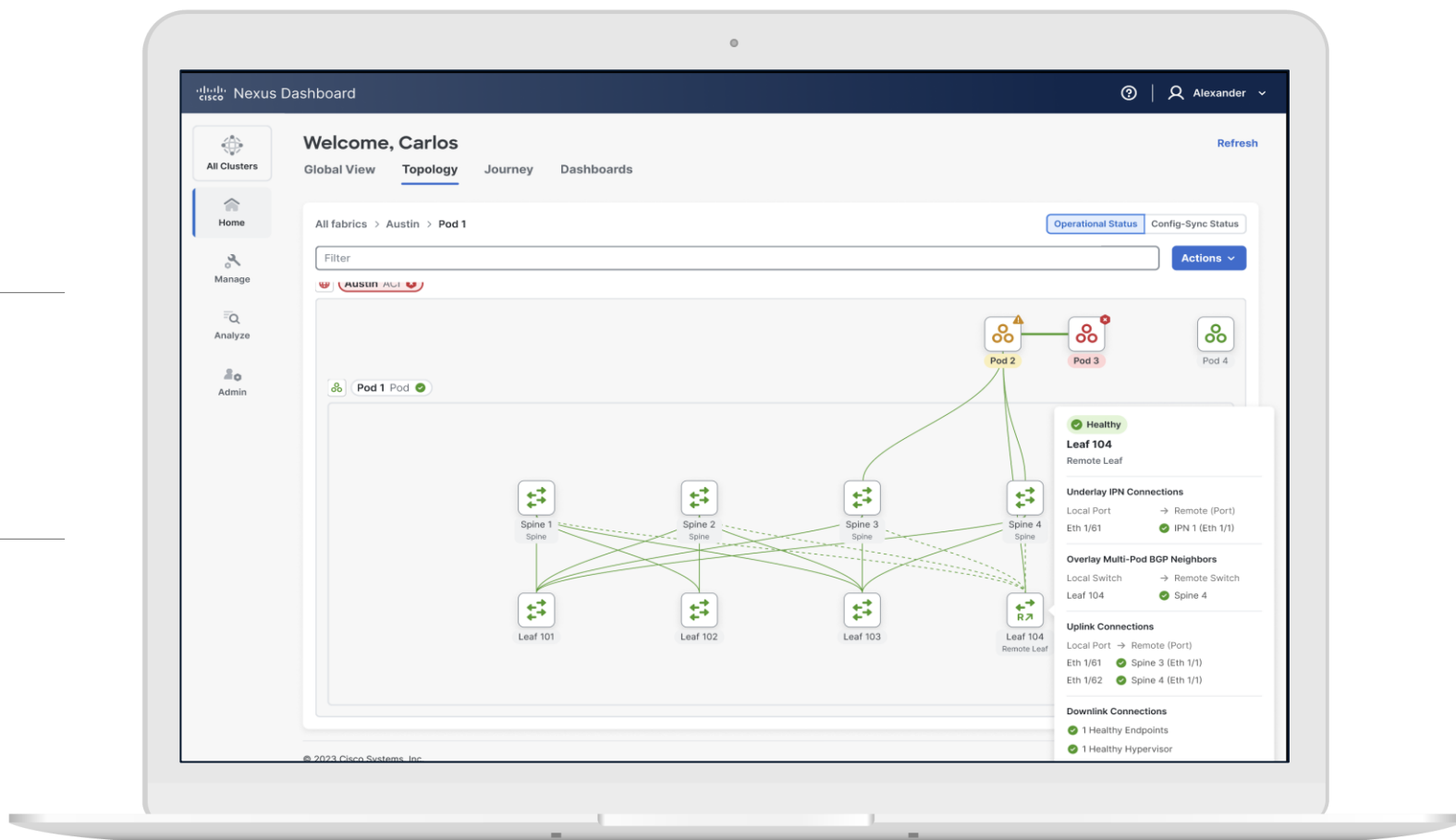
Updated topology
View fabrics, switches, interfaces, and endpoints with their corresponding anomaly scores



Visualize and configure
Verify health and configuration-sync status, configure vPC pairs, assign roles, and more



Single and multi-fabric
Drill down into a fabric or visualize connections across fabrics, including external and inter-fabric networks such as IPN and ISN



Visualize all your fabrics in a single place



Analyze

Nexus Dashboard



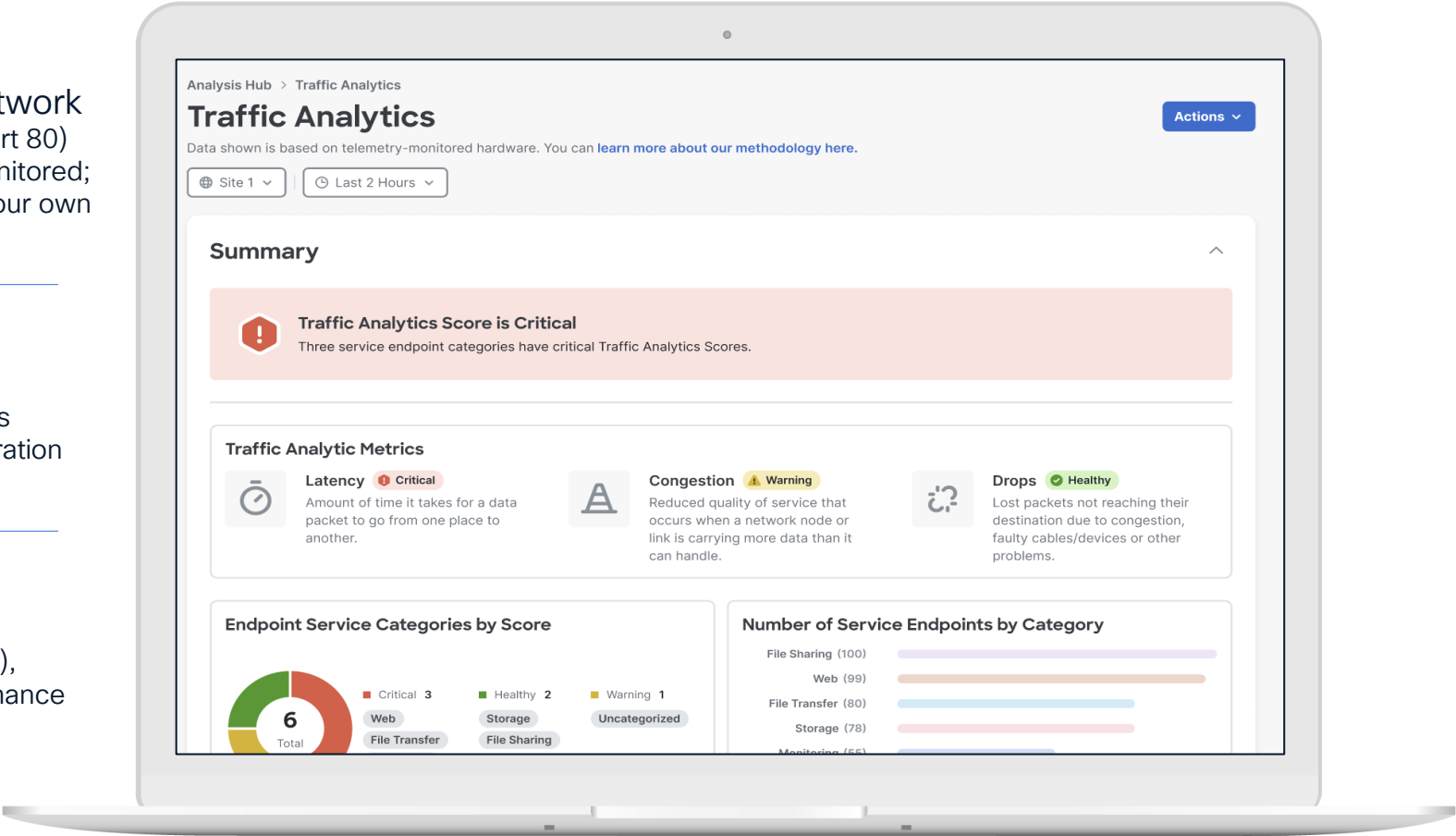
Automatically identify services in the network
Through well-known L4 ports (e.g., Web – TCP port 80) pre-loaded service categories are learned and monitored; category customization is also allowed based on your own preferences



Pervasive across switches and fabrics
Learn about clients and services connecting across multiple fabrics without any additional rule configuration



Analytics
Visibility (Lossless Ethernet), Monitoring (ECN,PFC), Congestion Score, Application to Network Performance Correlation, Telemetry and NetOps



Identify latency, congestion, and drops in your network

Analytics

Nexus Dashboard

AI Network Visibility

Visibility – Lossless Ethernet

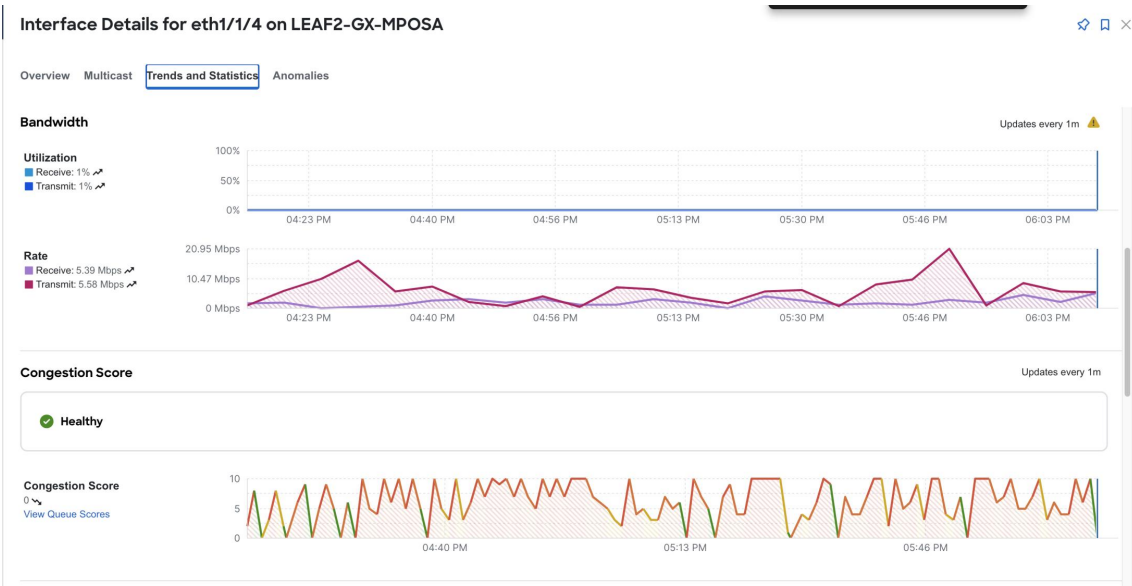
Monitoring (ECN,PFC)

Congestion Score

Application to Network Performance

Correlation

Telemetry and NetOps



Analytics

Nexus Dashboard

Interface Details for eth1/1/4 on LEAF2-GX-MPOSA



Overview Multicast **Trends and Statistics** Anomalies

Bandwidth

Updates every 1m

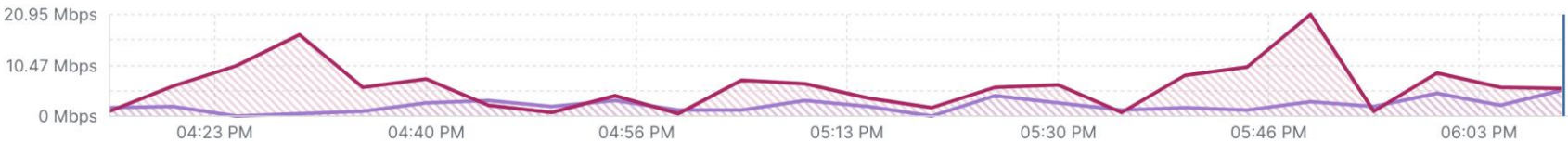
Utilization

- Receive: 1%
- Transmit: 1%



Rate

- Receive: 5.39 Mbps
- Transmit: 5.58 Mbps



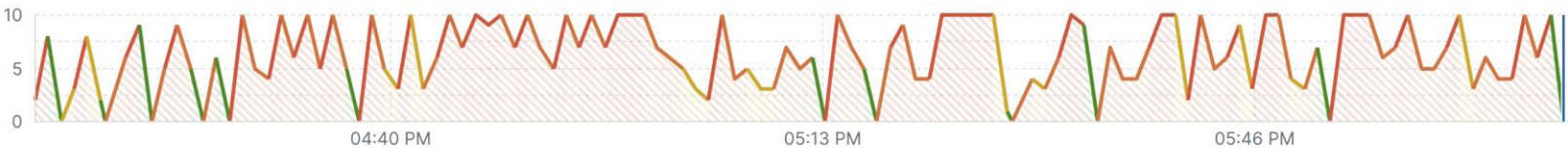
Congestion Score

Updates every 1m

Healthy

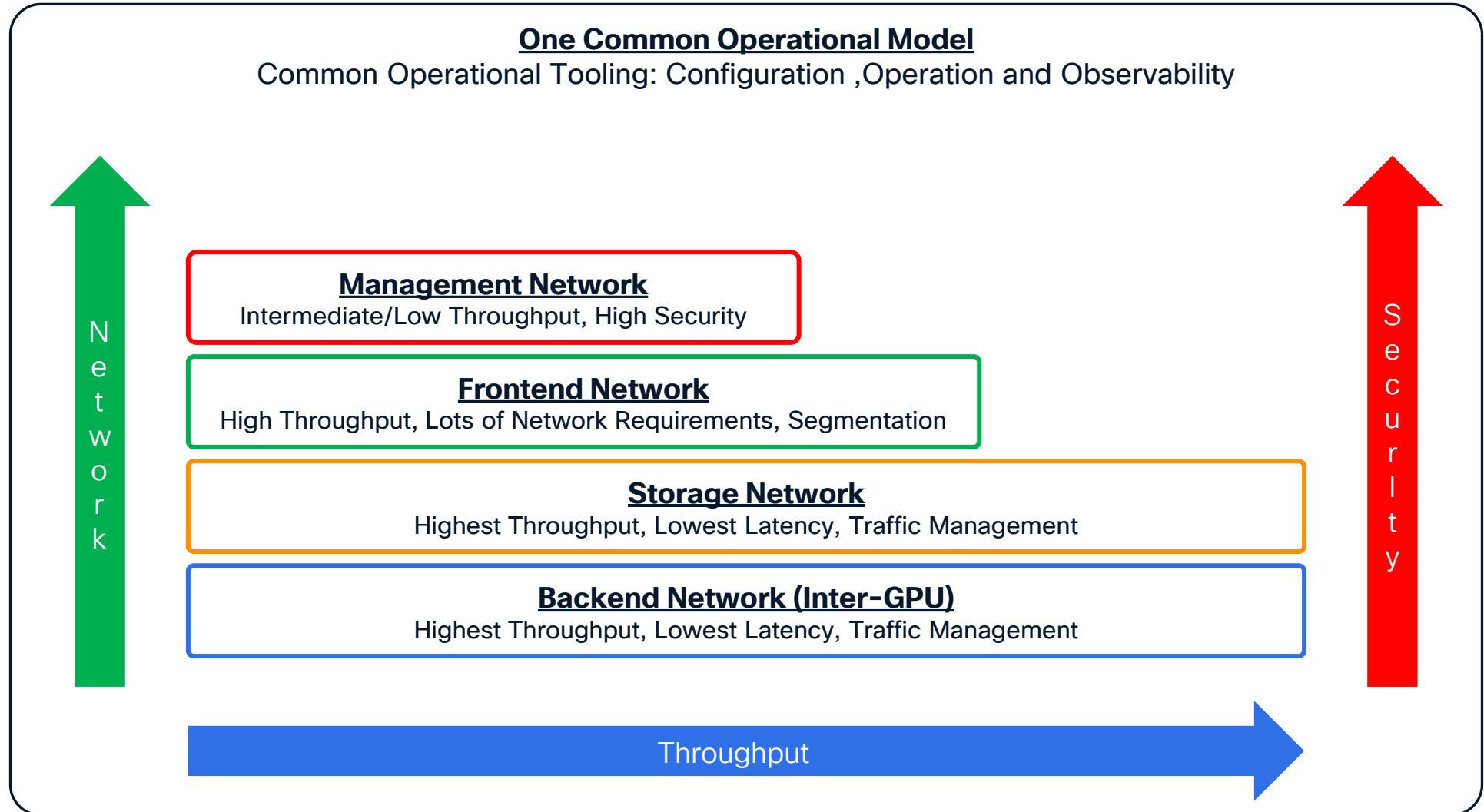
Congestion Score

- 0
- [View Queue Scores](#)



Conclusion

Different Requirements – One Operational Model



Conclusion

- **AI infrastructure** demands a **range of network types** to support different workloads and operational needs.
- While Backend networks tend to be more **straightforward**, Frontend networks require **broader capabilities and greater versatility**.
- Despite **diverse networking paradigms**, customers consistently seek unified operational tools for **streamlined management**.
- **Integrated operational tooling** for Frontend, Backend, Management, and Storage networks is increasingly seen as **essential**.

Complete your session evaluations



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



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



Level up and earn exclusive prizes!



Complete your surveys in the Cisco Live Events app.

Continue your education



Visit the Cisco Stand for related demos



Book your one-on-one Meet the Expert meeting



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



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

Contact me at: Insert preferred comms method

Thank you

CISCO Live !

