Deploying High Performance and Low-Latency AI networks using Cisco Nexus 9000

Groq: An AI Success Story

Cameron Fredinands
Director of Network and Data Center Engineering, Groq

Swetha Velamuri
Leader, Product Management
Data Center Networking, Cisco

CISCO Live!

CSSDCN-2005

Cisco Webex App

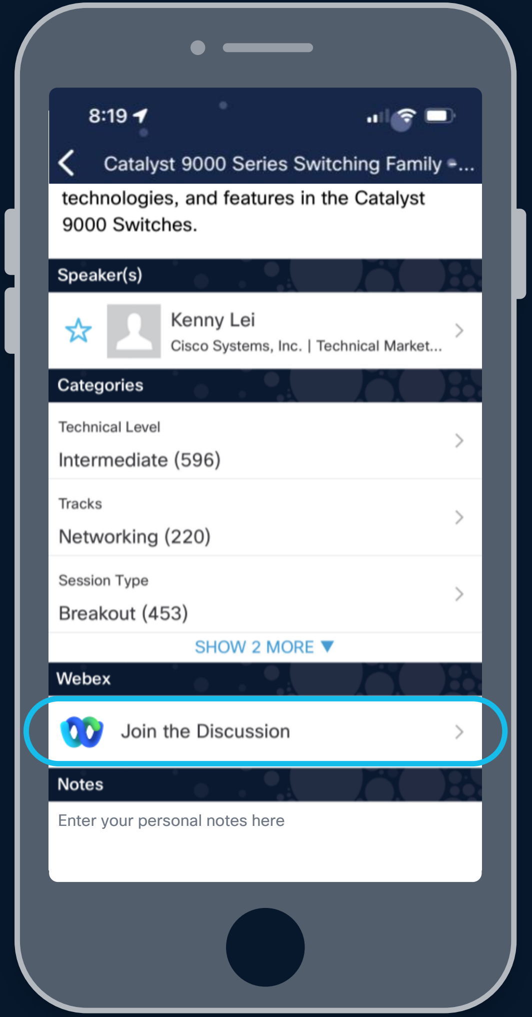
Questions?

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, 2025.



8:19

Catalyst 9000 Series Switching Family ...

technologies, and features in the Catalyst 9000 Switches.

Speaker(s)

Kenny Lei
Cisco Systems, Inc. | Technical Market...

Categories

Technical Level
Intermediate (596)

Tracks
Networking (220)

Session Type
Breakout (453)

SHOW 2 MORE ▼

Webex

Join the Discussion

Notes

Enter your personal notes here

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**

Optimized AI/ML Fabrics with Nexus

(Silicon, Systems, Software, Operations)

400/800G Ethernet Transition (25.6T & 51.2T switches)

High-bandwidth fabrics with reduced footprint and energy savings

RDMA over Ethernet (RoCEv2)
Non-Blocking Lossless network (PFC + ECN)

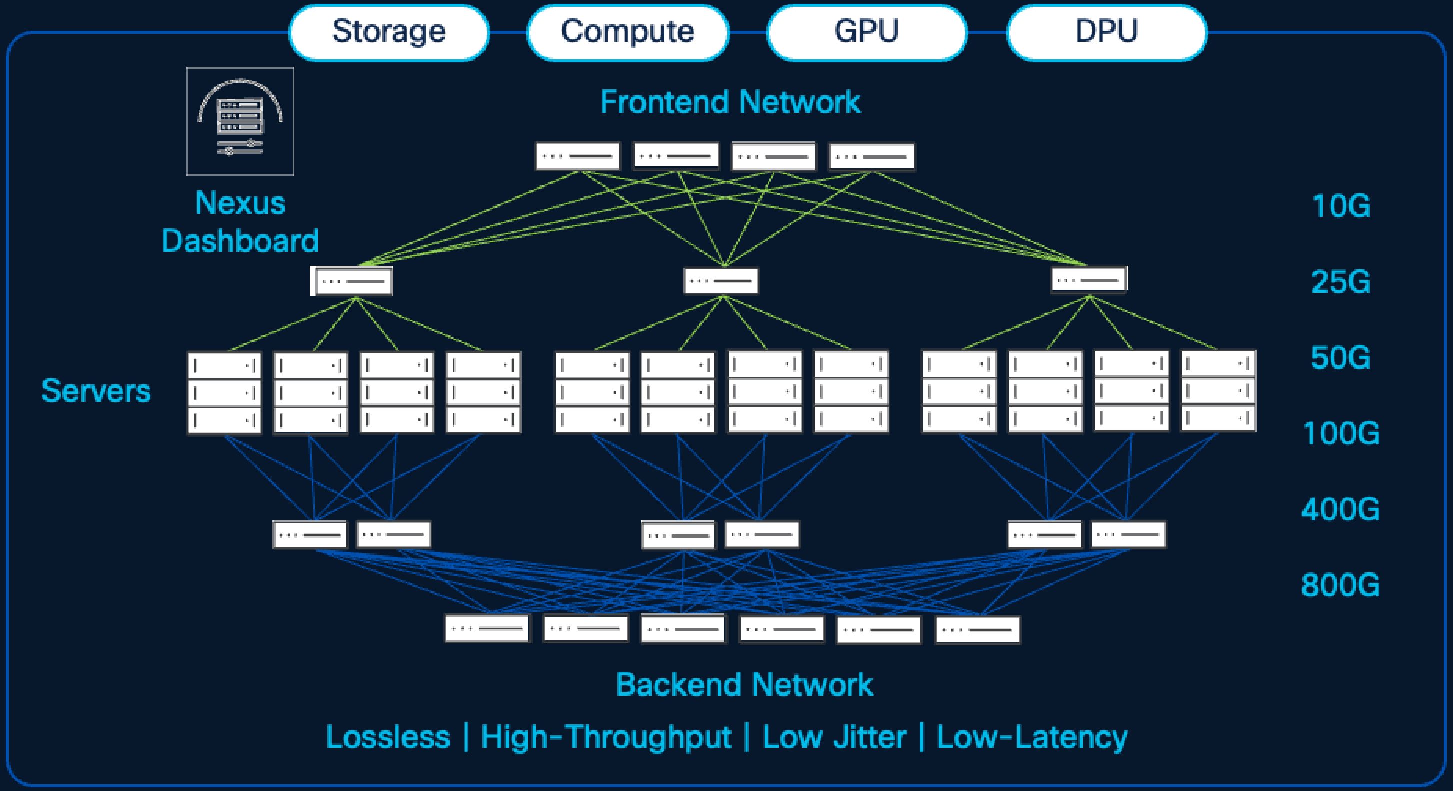
Advanced load balancing with congestion and fault aware traffic management powered by

Intelligent Packet Flow

AI fabric templates, AI analytics, telemetry, congestion scores, AI job monitoring, GPU, NIC visibility

Validated designs for networks and ecosystem partners (ERA)

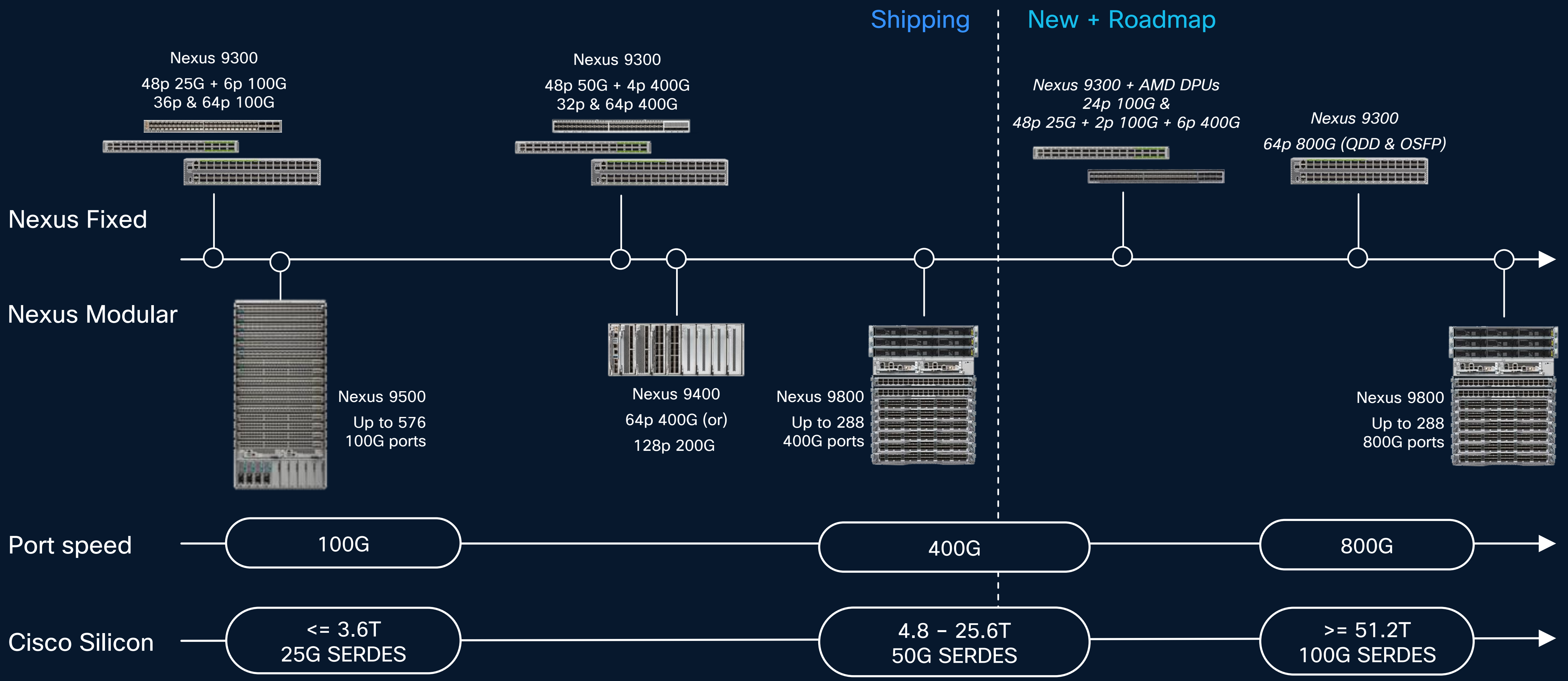
AI/ML Blueprint



Data center

Colocation

Cisco Nexus 9000 Series Switches



Shipping

New + Roadmap

Nexus Fixed

Nexus 9300
48p 25G + 6p 100G
36p & 64p 100G

Nexus 9300
48p 50G + 4p 400G
32p & 64p 400G

Nexus 9300 + AMD DPUs
24p 100G &
48p 25G + 2p 100G + 6p 400G

Nexus 9300
64p 800G (QDD & OSFP)

Nexus Modular

Nexus 9500
Up to 576
100G ports

Nexus 9400
64p 400G (or)
128p 200G

Nexus 9800
Up to 288
400G ports

Nexus 9800
Up to 288
800G ports

Port speed

100G

400G

800G

Cisco Silicon

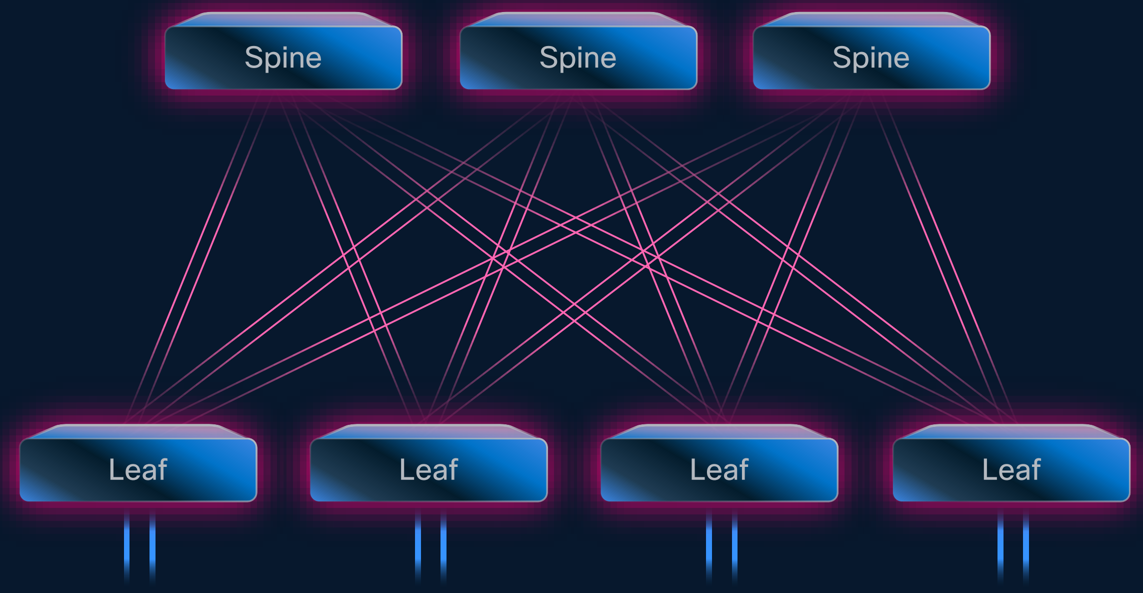
<= 3.6T
25G SERDES

4.8 – 25.6T
50G SERDES

>= 51.2T
100G SERDES

Unified Architecture

Powering AI Fabrics with Intelligent Packet Flow



Ultra Ethernet Consortium READY

Meet Intelligent Packet Flow

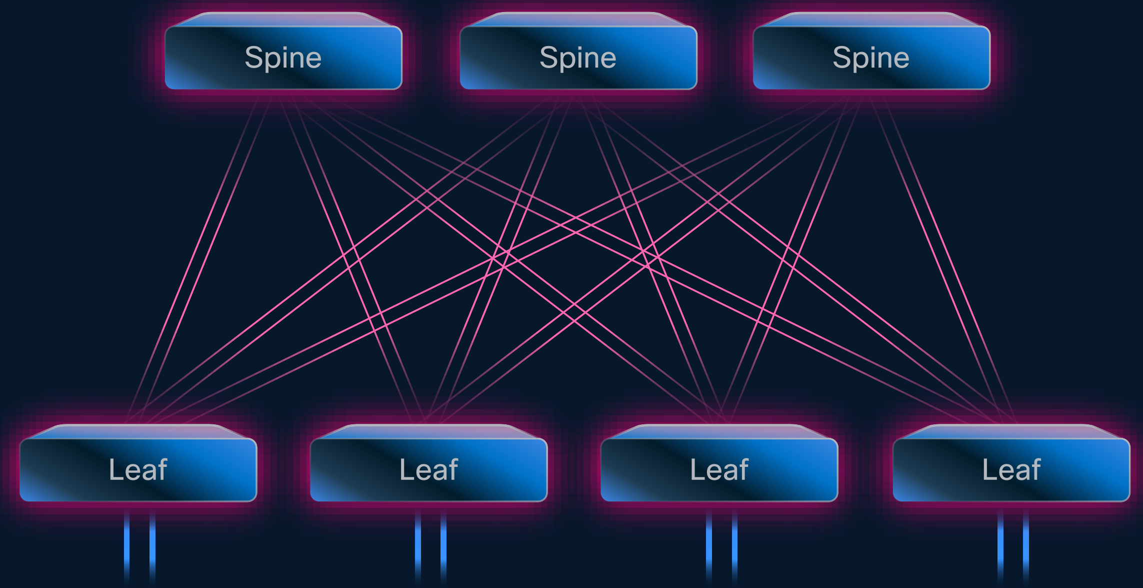
Advanced load balancing

Hardware accelerated telemetry

Fault aware recovery

Unified Architecture

Powering AI Fabrics with Intelligent Packet Flow



Ultra Ethernet Consortium READY

Advanced load balancing

FEATURES

Dynamic Load Balancing (DLB) with Load and Congestion Awareness

Per Packet and Selective Packet Spray (Ex: RDMA vs. non-RDMA)

Equal Cost Multi Path Load Balancing (ECMP) Static Pinning

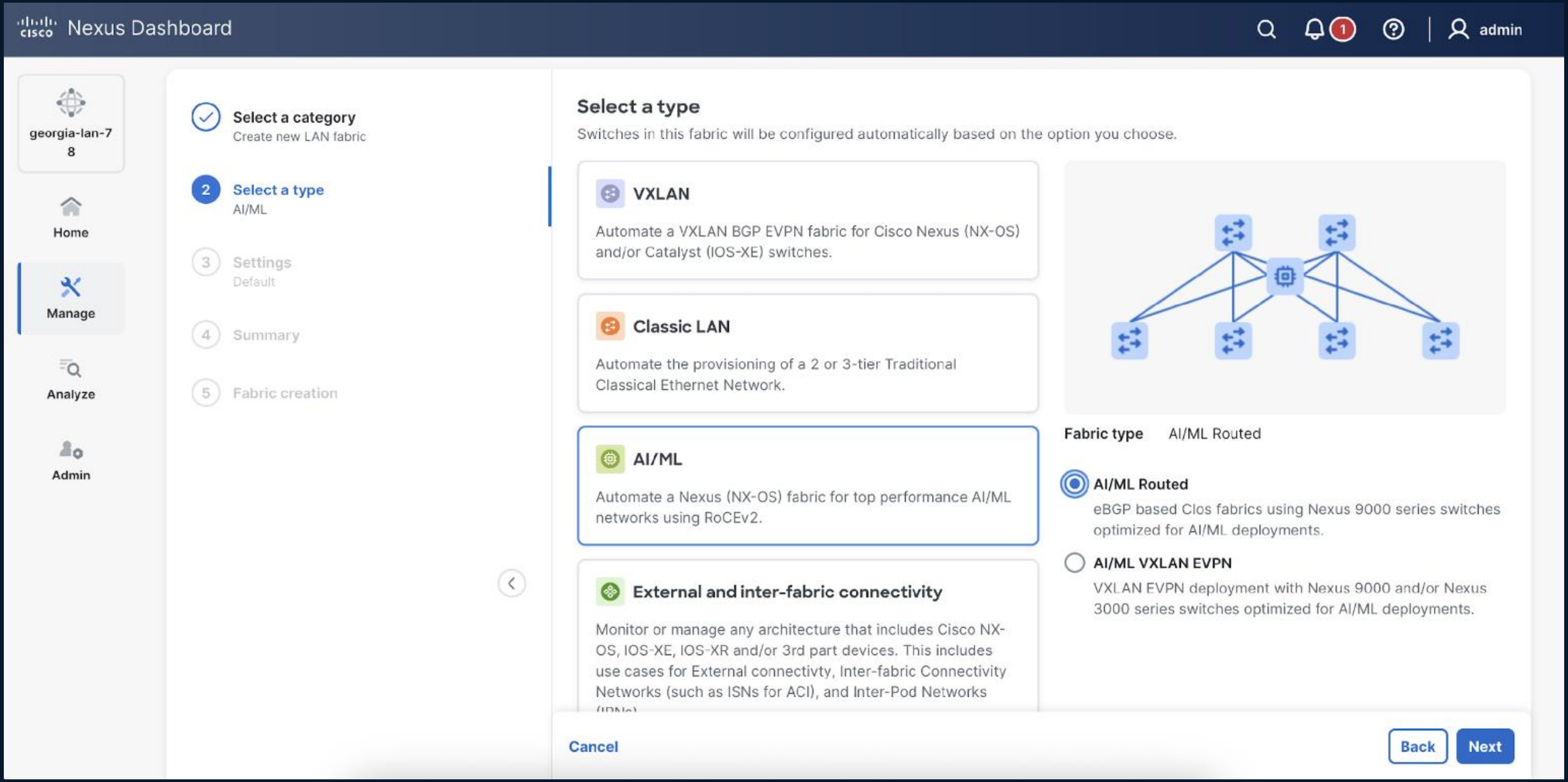
Weighted Cost Multi Path Load Balancing (WCMP) with Dynamic Load Balancing

Policy Based Flowlet Load Balancing (DSCP, ACL...)

Packet Trimming *

Nexus Dashboard – AI/ML fabric deployment

Simplifying Network Operations



✓ Select a category
Create new LAN fabric

2 Select a type
AI/ML

3 Settings
Default

4 Summary

5 Fabric creation

Select a type

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

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

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

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

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

Fabric type AI/ML Routed

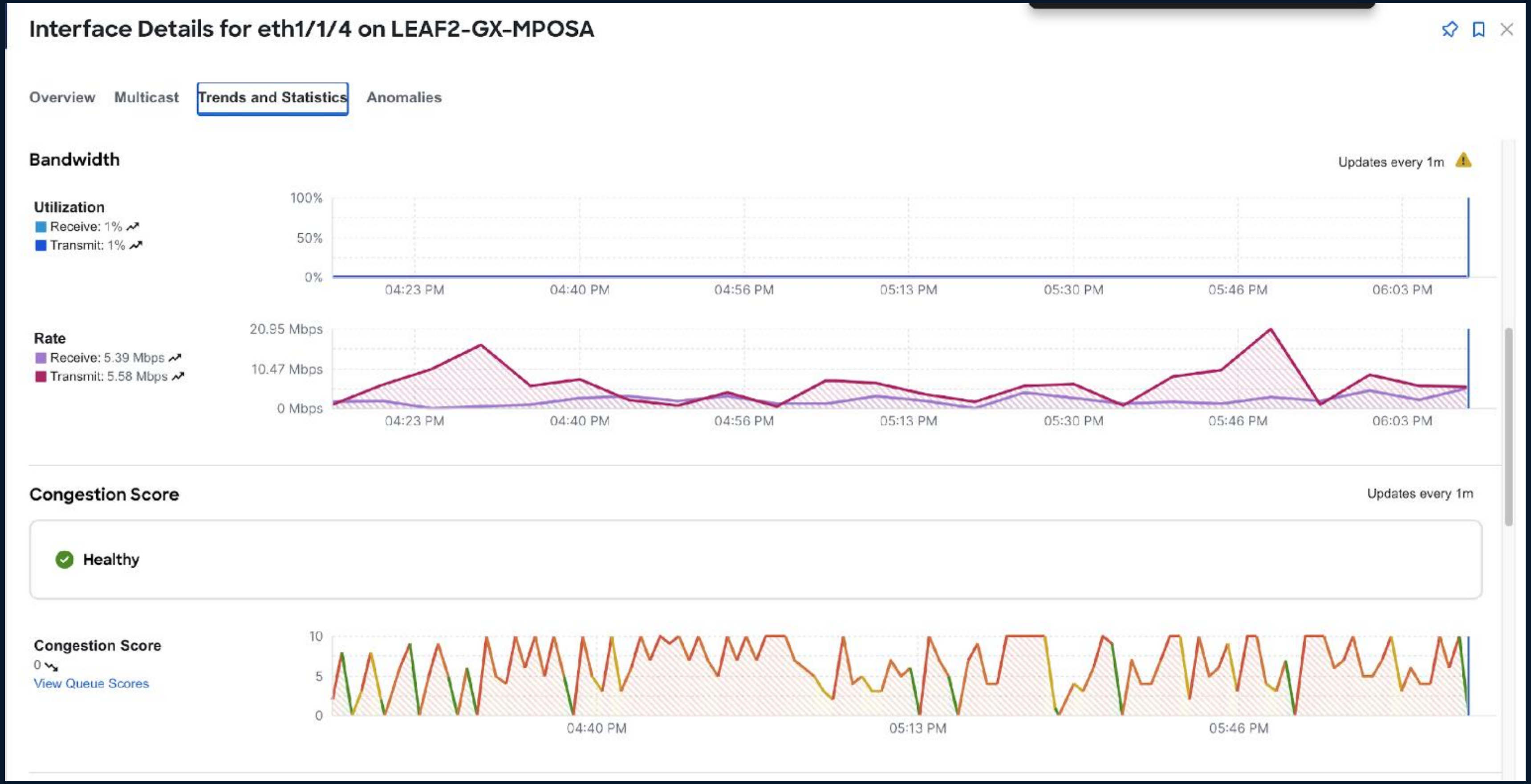
● **AI/ML Routed**
eBGP based Clos fabrics using Nexus 9000 series switches optimized for AI/ML deployments.

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

Nexus Dashboard – AI Analytics

Simplifying Network Operations

- AI Network Visibility
- UX/UI Dashboard
- Visibility – Lossless Ethernet
- Monitoring (ECN,PFC)
- Congestion Score
- Application to Network Performance Correlation
- Telemetry and NetOps



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

With the granular visibility provided by Cisco Nexus Dashboard Insights the network administrator can observe drops

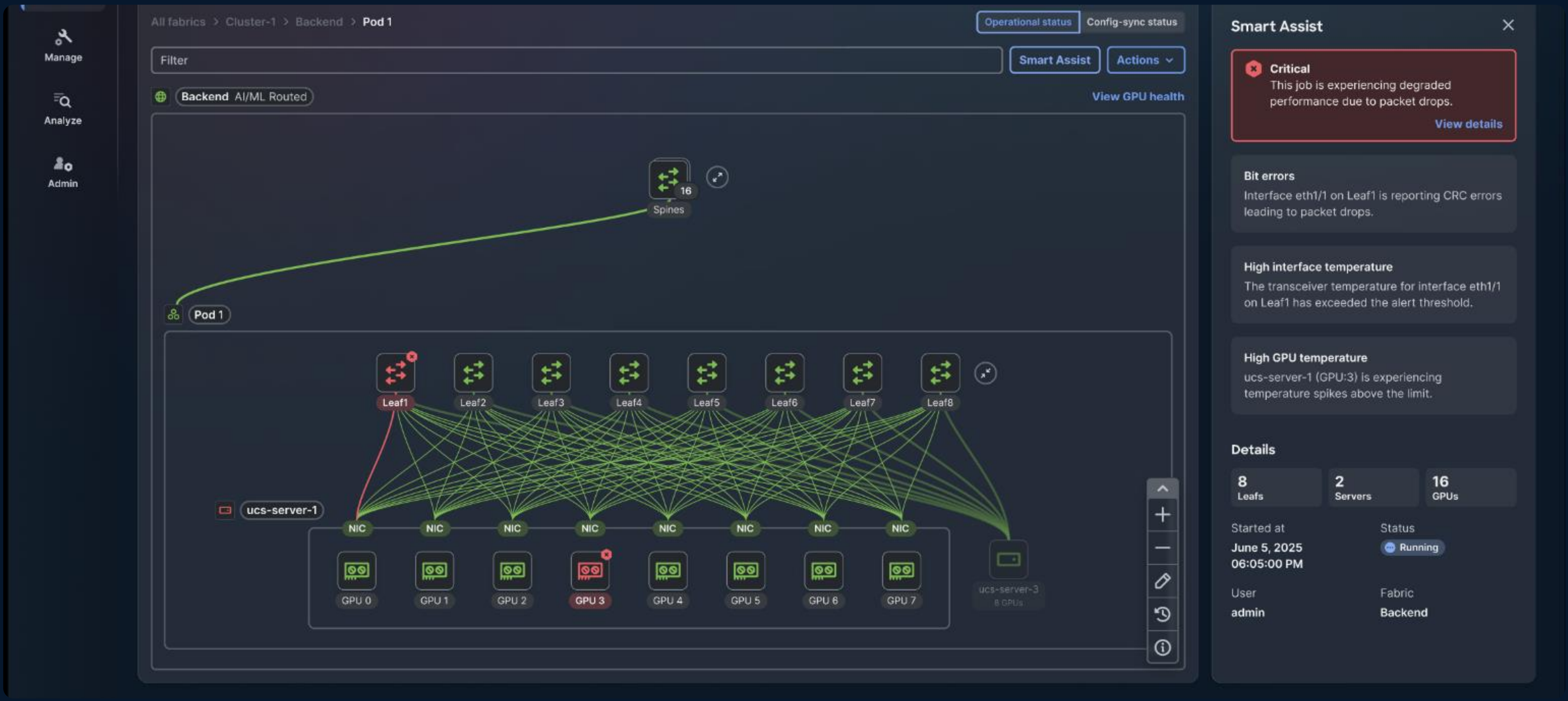
Tune thresholds until congestion hot spots clear and packet drops stop in normal traffic conditions

This is the first and most important step to ensure that the AI/ML network will cope with regular traffic congestion occurrences effectively

Nexus Dashboard – AI Analytics

Visibility into AI Jobs, GPU, NICS

*** Coming Soon*



All fabrics > Cluster-1 > Backend > Pod 1

Operational status | Config-sync status

Filter

Smart Assist | Actions

Backend AI/ML Routed

View GPU health

Smart Assist

Critical
This job is experiencing degraded performance due to packet drops.
View details

Bit errors
Interface eth1/1 on Leaf1 is reporting CRC errors leading to packet drops.

High interface temperature
The transceiver temperature for interface eth1/1 on Leaf1 has exceeded the alert threshold.

High GPU temperature
ucs-server-1 (GPU:3) is experiencing temperature spikes above the limit.

Details

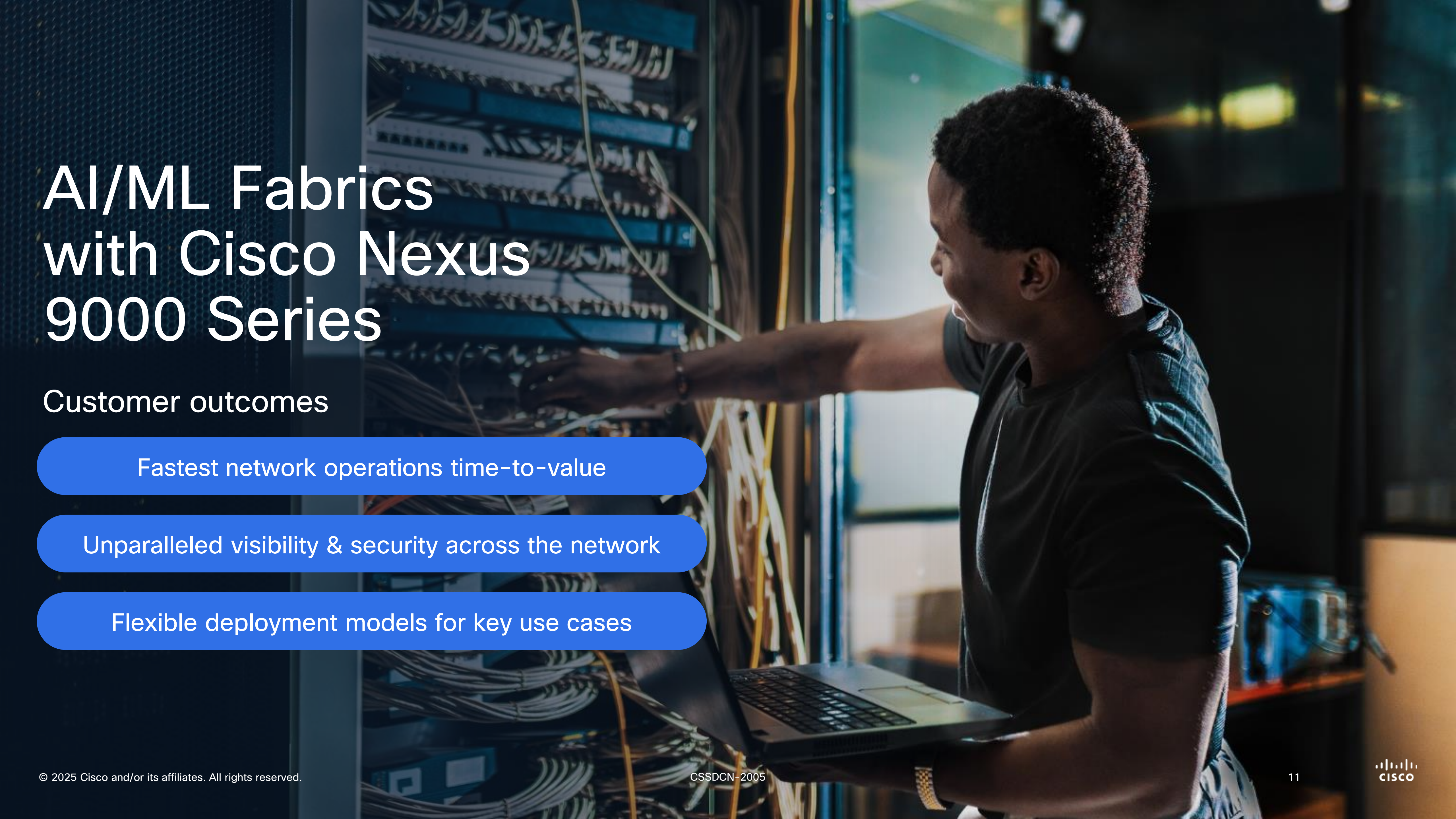
8 Leafs	2 Servers	16 GPUs

Started at
June 5, 2025
06:05:00 PM

Status
Running

User
admin

Fabric
Backend



AI/ML Fabrics with Cisco Nexus 9000 Series

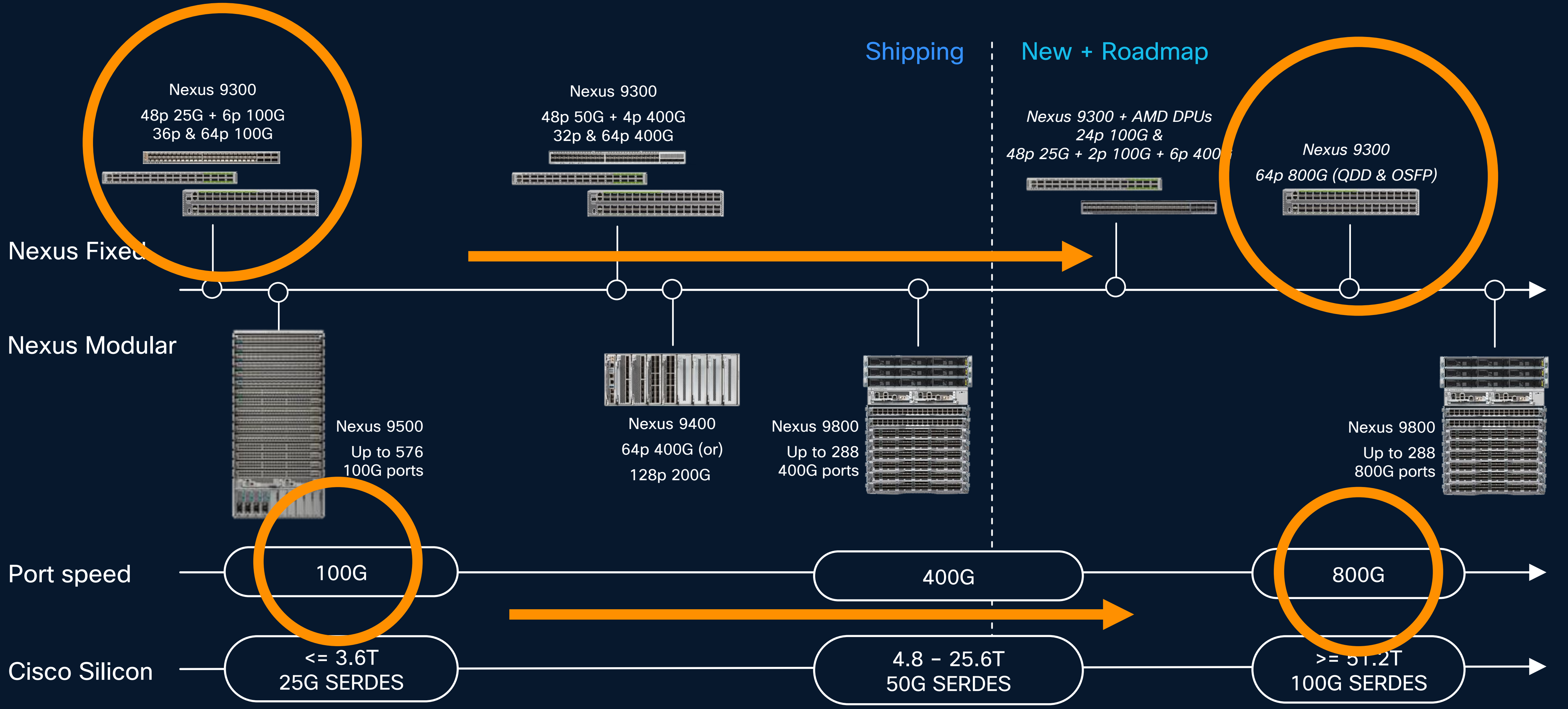
Customer outcomes

- Fastest network operations time-to-value
- Unparalleled visibility & security across the network
- Flexible deployment models for key use cases

In an era where AI workloads and real-time data processing define the competitive edge, high-performance data center switching is no longer a luxury—it's a necessity. Powered by Cisco Nexus 9000 series switching, we are able to move massive volumes of data with unparalleled speed and low latency that is foundational to unlocking the full potential of Groq's innovative fast inferencing solutions, ensuring organizations stay ahead in a data-driven world.

Cameron Fredinands,
Head of Network Operations, Groq

Groq’s Journey with Nexus 9000



Shipping

New + Roadmap

Nexus Fixed

Nexus 9300

48p 25G + 6p 100G
36p & 64p 100G

Nexus 9300

48p 50G + 4p 400G
32p & 64p 400G

Nexus 9300 + AMD DPUs
24p 100G &
48p 25G + 2p 100G + 6p 400G

Nexus 9300

64p 800G (QDD & OSFP)

Nexus Modular

Nexus 9500

Up to 576
100G ports

Nexus 9400

64p 400G (or)

128p 200G

Nexus 9800

Up to 288
400G ports

Nexus 9800

Up to 288
800G ports

Port speed

100G

400G

800G

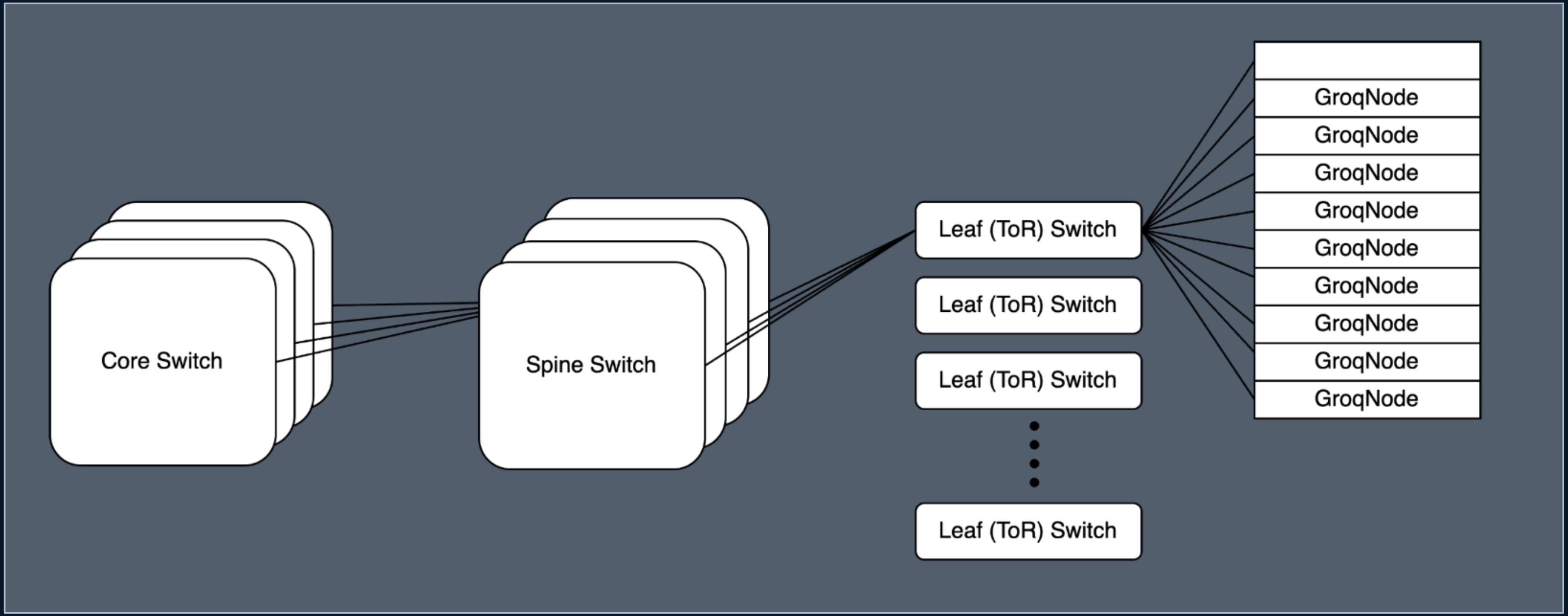
Cisco Silicon

<= 3.6T
25G SERDES

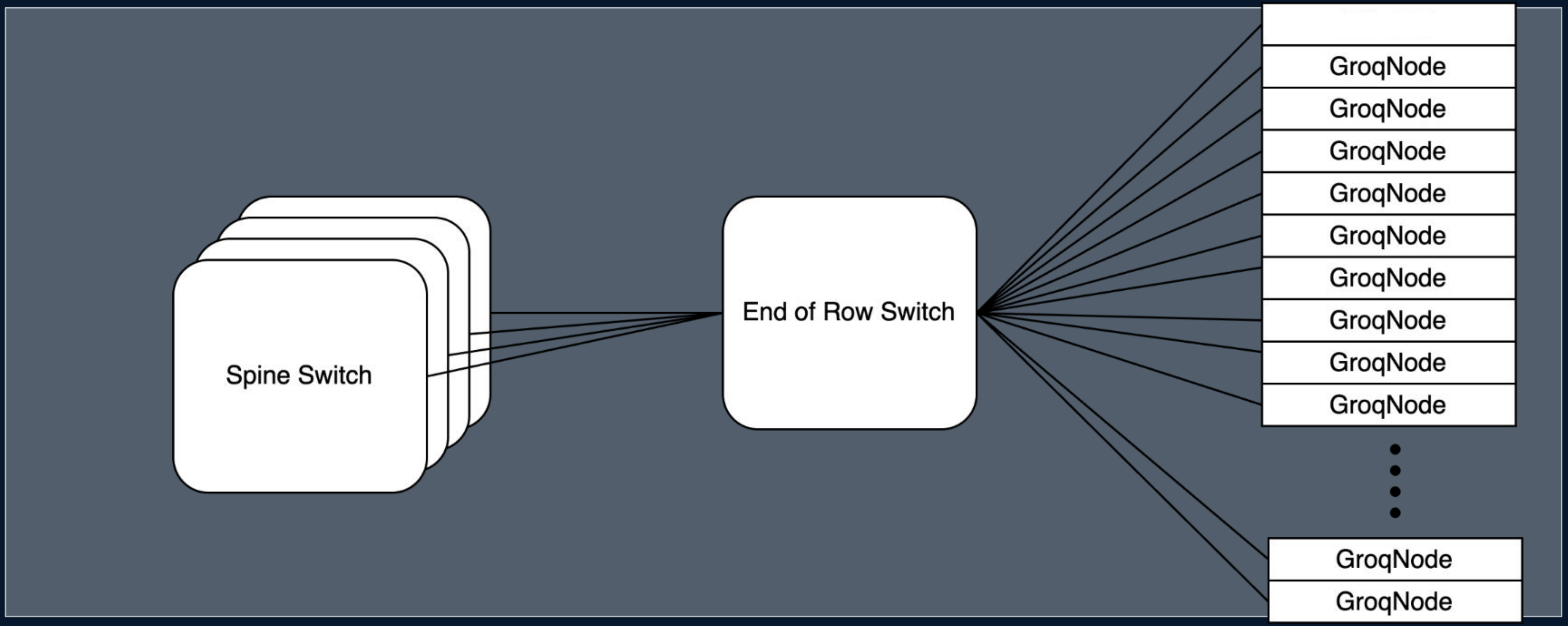
4.8 – 25.6T
50G SERDES

>= 51.2T
100G SERDES

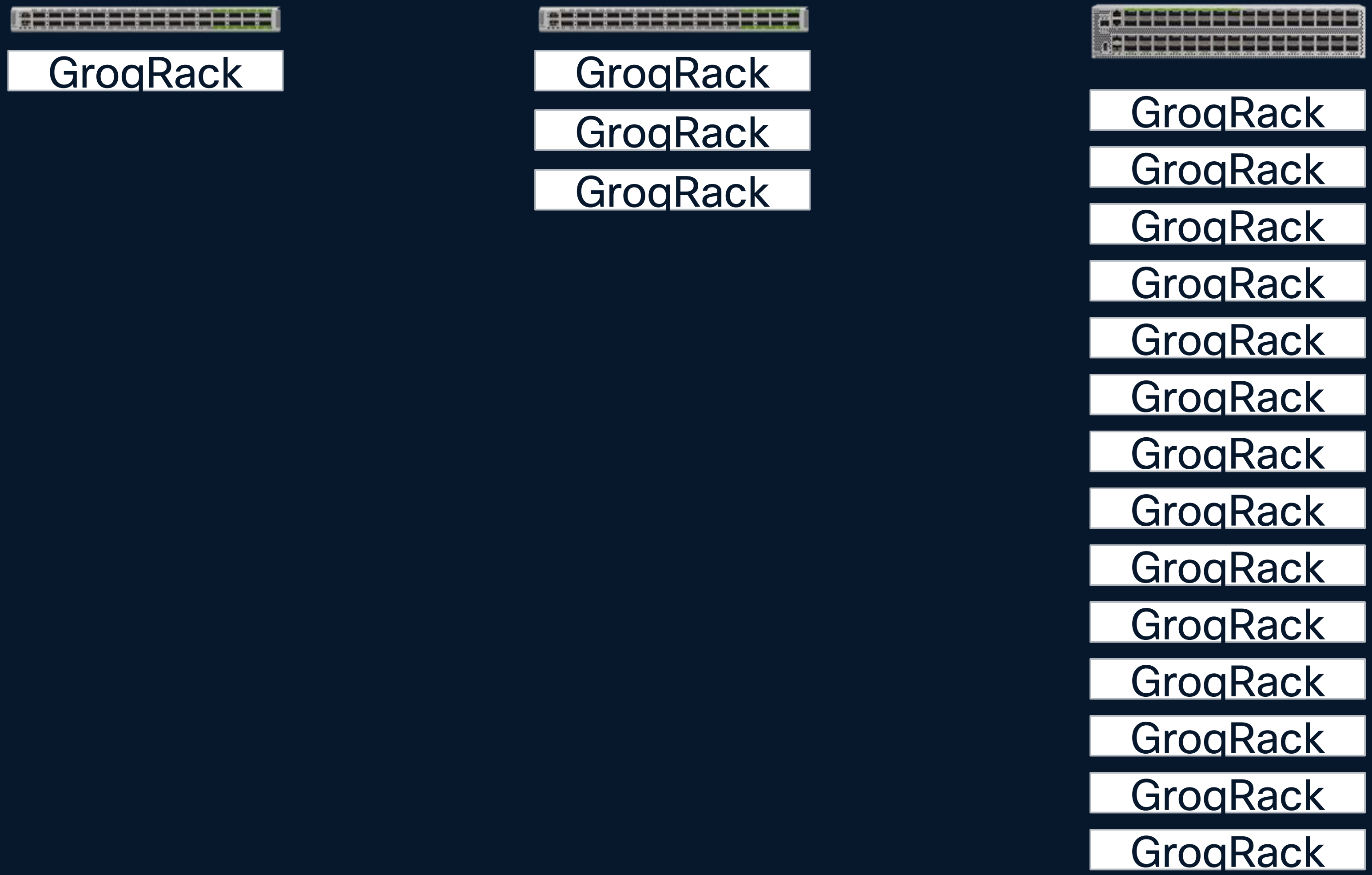
100G architecture



also 100G architecture!



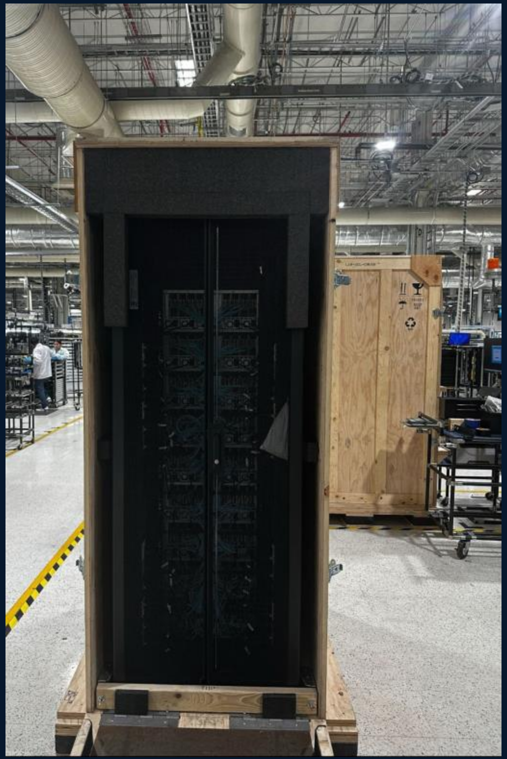
Our Journey from ToR - Tri-Rack - End of Row



GroqRack

GroqRack
GroqRack
GroqRack

GroqRack
GroqRack
GroqRack
GroqRack
GroqRack
GroqRack
GroqRack
GroqRack
GroqRack
GroqRack
GroqRack
GroqRack
GroqRack
GroqRack



> *A major differentiating factor in our DC switching choice was the extremely high radix inside Cisco Silicon One - supporting upto 512 x 100G interfaces on a single switch allowed us to make the 800G investment to support our 100G workloads today.*

Cameron Fredinands,
Director of Network and Datacenter Engineering, Groq

Thank you

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 2026.

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

Thank you