Transforming Bottlenecks into Breakthroughs: Scaling Secure AI Networks for AI Factories

Will Eatherton
SVP, Cisco Data Center, Internet & Cloud Infrastructure Engineering

Andrew Leece
Co-founder, Sharon AI

March 18, 2026

AI continues to drive unprecedented demands

Accelerating consumption[1]

$5.2T

AI data center investment by 2030

3.5x

increase in energy demand by 2030*

Massive traffic growth[2]

40%

CAGR in front-end network market through 2029

10x

growth in AI back-end network traffic every 2 yrs

Faster speeds[2]

800G

majority of front-end ports by 2029

1.6T

majority of back-end ports by 2027

1. McKinsey Quarterly
2. Dell'Oro

The Cisco & NVIDIA Partnership

Key innovations delivered in 2025

FEB

Expanded AI Infrastructure Partnership

Cisco UCS C845A M8 AI Rack Server

MAR

Cisco Secure AI Factory with NVIDIA

NVIDIA Enterprise Reference Architecture (ERA)

SEP

Cisco Silicon with NVIDIA Spectrum-X™

Cisco Silicon One (& Cloud Scale) becomes exclusive partner silicon for NVIDIA Spectrum-X networking

OCT

Cisco Nexus switch with Spectrum-X Ethernet silicon

NVIDIA Enterprise Reference Architecture (ERA)

NOV

Nexus Hyperfabric AI Clusters GA

Specialized networking infrastructure for generative AI deployments and high-performance computing

Multi-layer performance benchmarks

RDMA KPIs

out_of_sequence

packet_seq_error

roce_adp_retrans

np_ecn_marked_roce_packets

np_cnp_sent

rp_cnp_handled

Network KPIs

Congestion management

Packet Drop Ratio (Drops/Rx)

ECN Ratio (ECNs/Rx),
PFC Ratio (PFCs/Rx)

Buffer Utilization

Congestion avoidance

Jain's Fairness Index

Application KPIs

P01 BW

P99 Latency

JCT

Bus Bandwidth

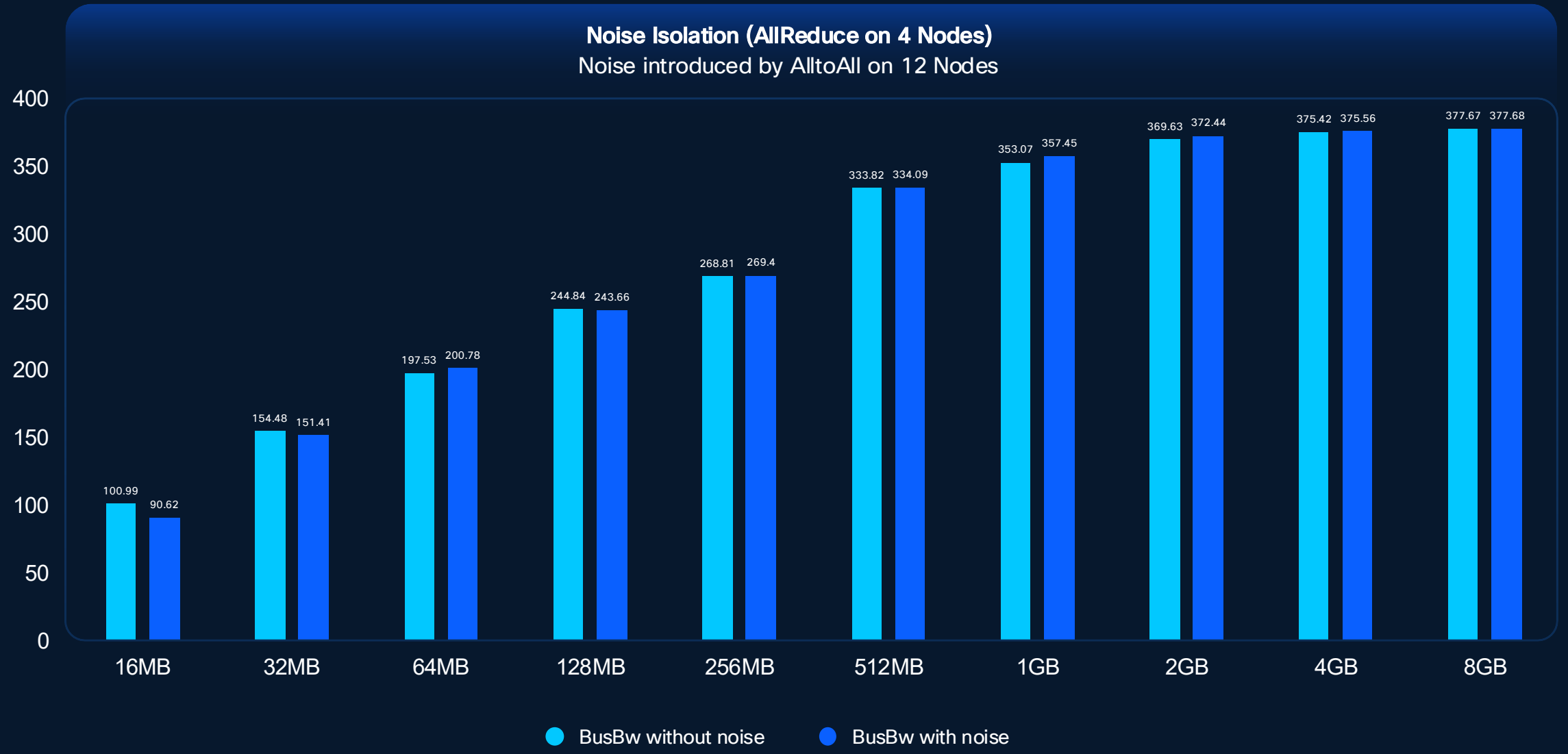
Tokens/second

Avg. Bus BW

RDMA bisection results

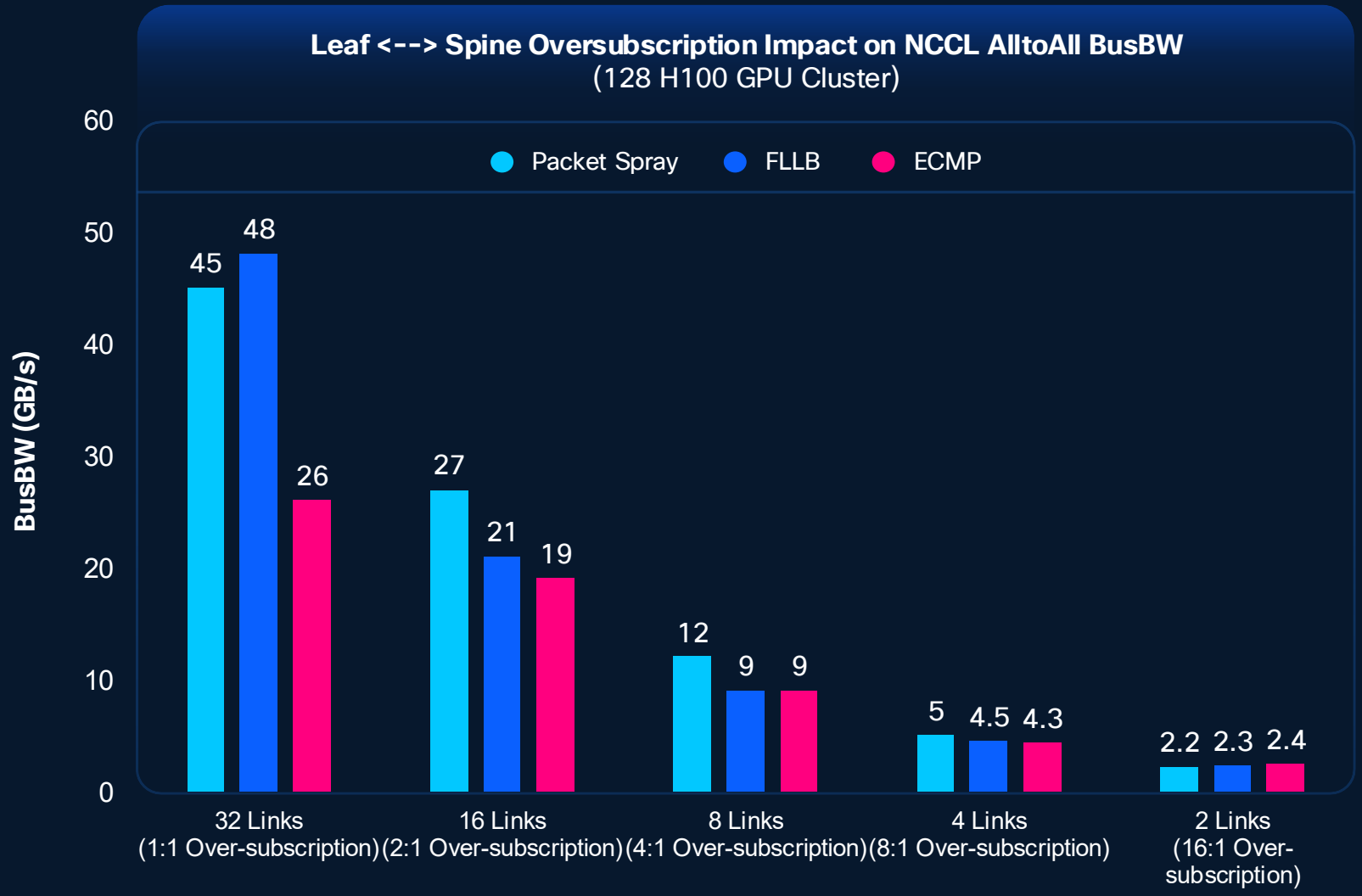
Queue Pairs	MTU	P01 Bandwidth (Gbps)	P01 Bandwidth (%)	P99 Bandwidth (Gbps)	P99 Bandwidth (%)	Tail Latency (µs)	Packet Drop	PFC	ECN	Jains Fairness Index
1	2048	388.4	97.1	389.3	97.3	~19	0	0	0	1
1	4096	388.4	97.1	389.3	97.3	~19	0	0	0	1
32	2048	389.1	97.2	389.5	97.3	~46	0	0	0	1
32	4096	389.1	97.2	389.5	97.3	~46	0	0	0	1

NCCL multiple jobs (noise isolation) results



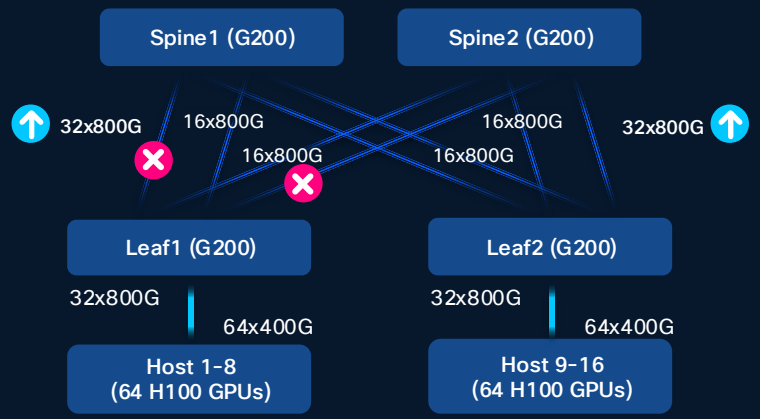
Resiliency link failure

Impact of oversubscription



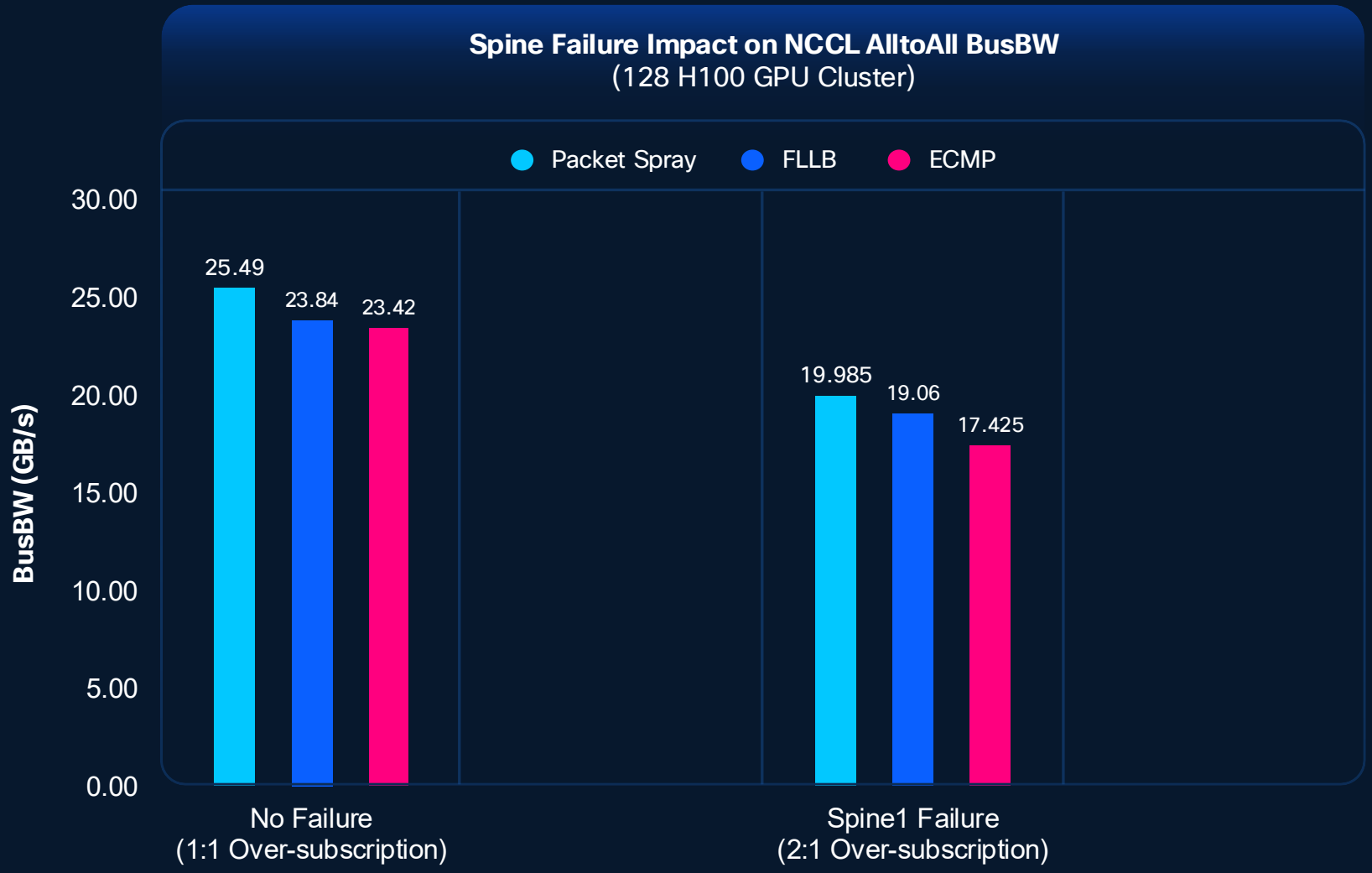
Leaf ←→ Spine

32 Links	Each Leaf has 16 links to each Spine
16 Links	Each Leaf has 8 links to each Spine
8 Links	Each Leaf has 4 links to each Spine
4 Links	Each Leaf has 2 links to each Spine
2 Links	Each Leaf has 1 links to each Spine



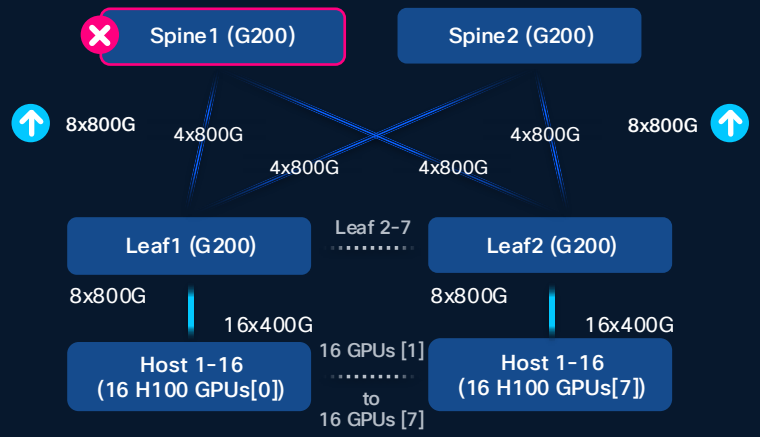
Resiliency spine failure impact

Spine Failure impact on 8G bytes message size



Leaf ↔ Spine

8 Links	Each Leaf has 4 links to Spine1 & Spine2
4 Links	Each Leaf has links to Spine2 Only



PXN was disabled on the compute nodes to force rail traffic over the Spine and simulate a fabric Link failure; otherwise, with default NCCL settings, traffic would remain within the rails.

AI innovations continue unabated

1

Hardware & Optics

100T

CPO/LPO

Liquid cooling

2

Architecture

Scale up

Scale out

Scale across

3

Software Services

Routing protocols

Network OS

Services

Hardware & Optics

100 Tbps is the new normal

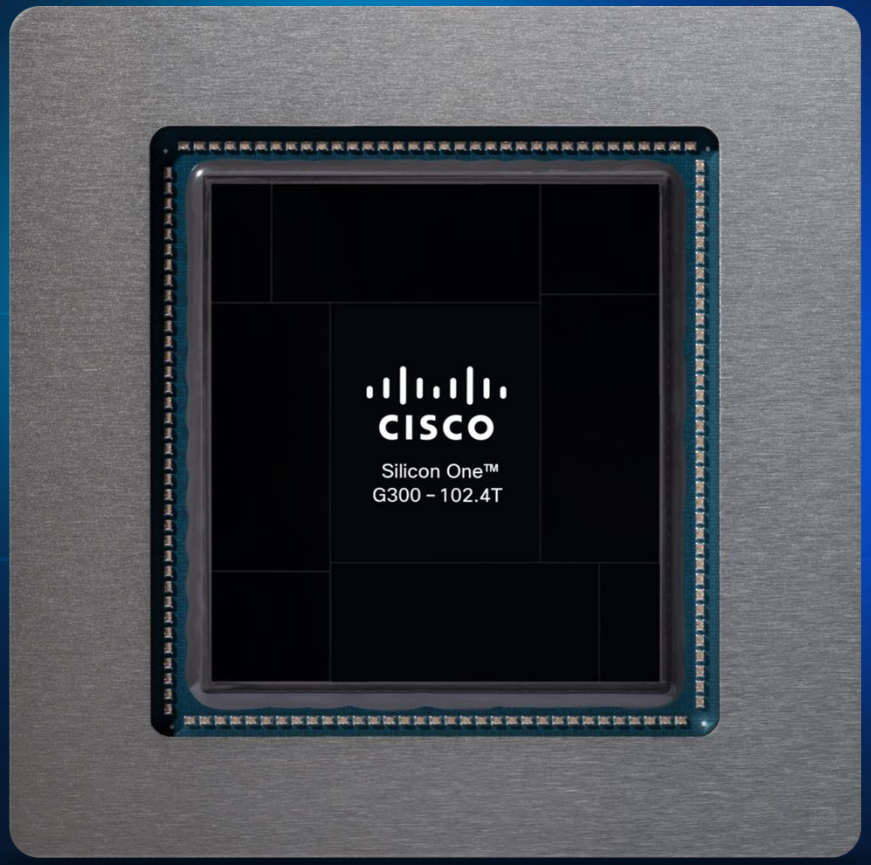
Introducing

Cisco Silicon One G300

102.4 Tbps
AI Network Switch

1.6T Ethernet
Port Switch

256MB
Fully-Shared Packet Memory



200 Gbps
Integrated Serdes

512 ports
High Radix Scaling

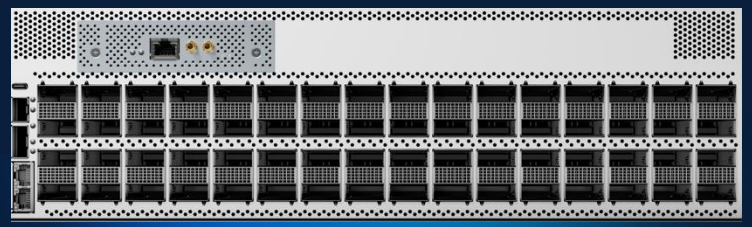
Optimizing TCO
for Scale AI Deployments

Cisco Advanced 102.4T Systems Portfolio

Seamless deployment flexibility for optimized AI performance

H2'26

Cisco N9364F-SG3 Cisco 8133
Air Cooled



NX-OS

Features

3RU, 64p 1.6T OSFP

EIA 19"

Redundant Pluggable Power

102.4Tbps

Cisco N9364F-SG3 Cisco 8132
Fully Liquid Cooled

NX-OS

Features

2OU, 64p 1.6T OSFP

ORv3N 21" blind mate

Redundant Integrated Power

Advanced Leak Detection

102.4Tbps

Cisco Silicon One Spectrum-X License

Hyperscale | Neocloud | Service Providers | Sovereign Cloud | Enterprise

NEW

Introducing

102.4T Cisco N9100

Scale Out Fabric, powered by NVIDIA Spectrum-6 Ethernet

The next major leap in AI data center capacity with NCP RA-compliant Cisco Secure AI Factory with NVIDIA

FUTURE

N9164F-NS6

64p 1.6T OSFP, 2OU

Cisco N9164F-NS6 OSFP

Ultimate performance

Extraordinary network scale for AI to meet the demands of next generation secure AI factory

100% liquid cooled system with direct to chip liquid cooling

NVIDIA technology and partnership

Addresses large scale AI infrastructures with NCP Reference Architecture (RA)-compliant network designs

NVIDIA Spectrum-X Ethernet technology enables scale-out AI fabric

Sovereign & Neocloud AI at Giga-Scale

Enabling customer need for scale-out and scale-across networking

Choice of NCP RA-compliant and Cisco CRA Compliant architectures

102.4T Cisco N9300 switch (announced at Cisco Live EMEA 2026)

Based on Cisco Silicon One

SCALE-OUT NETWORK

NEW

102.4T Cisco N9100 switch

Based on NVIDIA Spectrum-6 Ethernet silicon

Reliable connectivity for AI data centers

Cisco OSFP 1.6T Pluggable Optics

Q3’26

High-performance and reliability

Cisco Silicon Photonics technology

Advanced DSP-based diagnostics and monitoring

High bandwidth AI scale-out

1.6T switch-to-NIC connectivity and high-density fabrics for massive AI cluster builds (2x800G, 4x400G, and 8x200G links)

Diversified supply chain

Seamless AI cluster scaling with diversified supply chain and interoperability across Cisco and 3rd party DSPs

Efficiency and flexibility

Optics for power-efficiency and deployment flexibility in liquid cooled and air-cooled platforms

Cisco 800G Linear Pluggable Optics (LPO) Solutions

Powered by **Cisco Silicon Photonics**

AI Scale-out power efficiency and reliability for 800G Networks

Shift power consuming optics DSP functionality to Cisco Silicon One platform

High signal integrity driven by Cisco Silicon Photonics and Cisco Silicon One

High reliability with lower thermal stress

Increase port density without PDU upgrades

Q3’26

Cisco L800G OSFP LPO

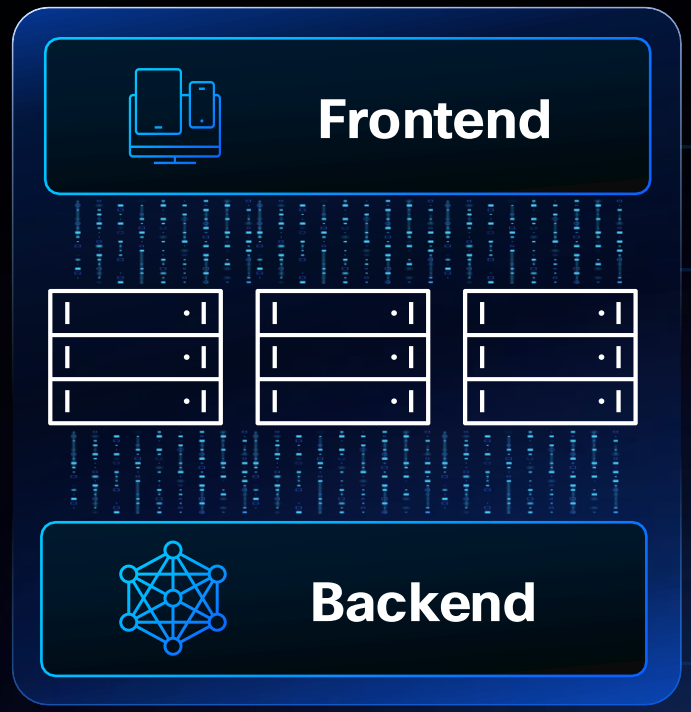
50% less power per module

30% less power for integrated systems

Architecture

SRv6-Powered Data Center Interconnect

DCI growth for inference and training is massive



Data Center 1

System design choices are changing over time

Modular system for 1.6T

Dis-aggregate with 51.2T deep-buffered switches

High capacity ZRP and ZR+ optics

SRv6 for AI DCI roles

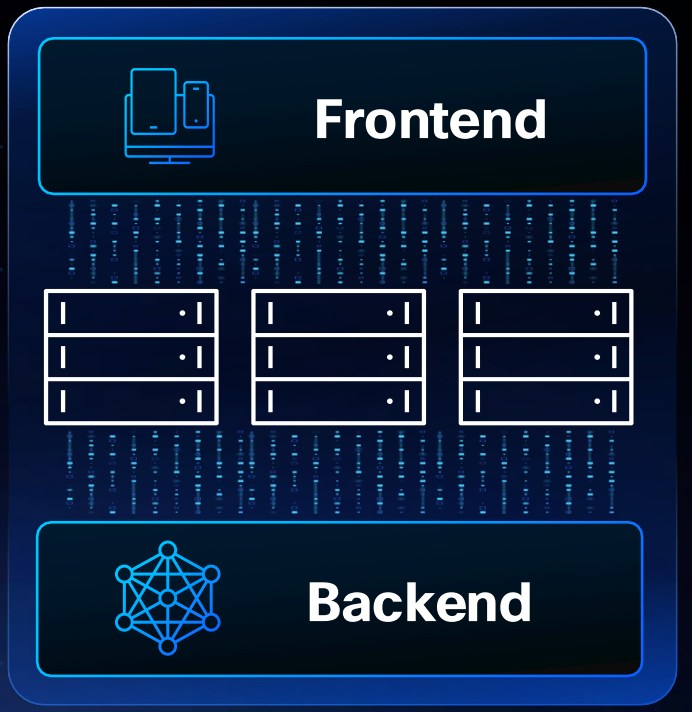
Utilizes Micro-SID (uSID) to reduce header overhead, improving performance for AI and high-throughput connectivity

Enhanced Traffic Engineering

Improved scalability

Reference Architecture

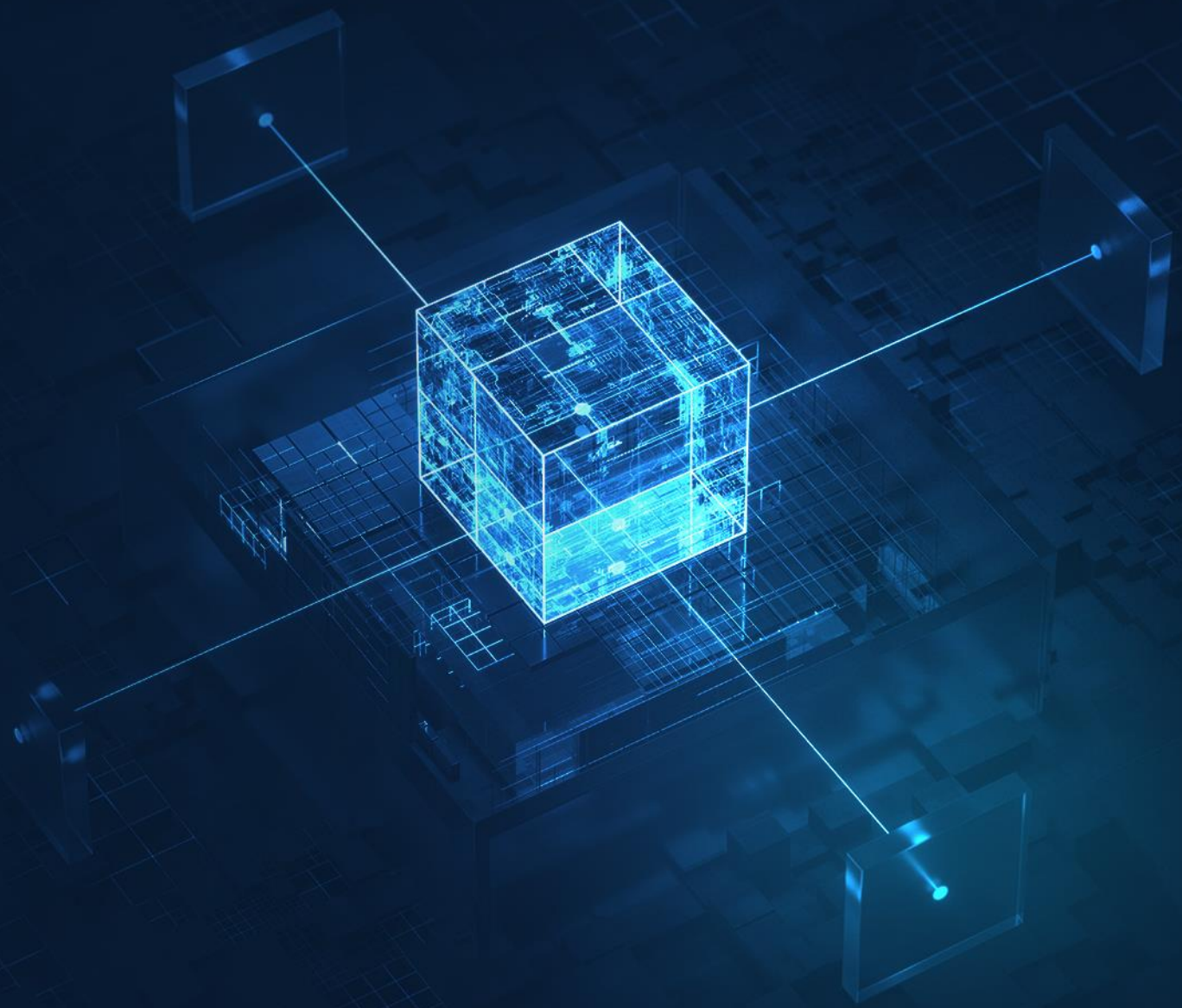
Expanding Cisco CRA to cover DC connectivity



Data Center 2


DC 1
DC 2
SCALE ACROSS
OPTION 1
Dynamic distance-aware NCCL routing
Topology aware networking
Increased buffering at server NICs to account for round-trip delay
Plus, the microburst within network
Intelligent collective networking
Local congestion Aware
Fabric congestion Aware
Cisco G200/G300-based systems for low-buffer scale-across buildout
OPTION 2
Use deep buffered devices to connect for traffic transitioning
No compromise solution for variable distance
Reliability and tolerant for transition and failures with additional buffering
Cisco P200-based 51.2T system

Software Services



Routing Protocol Evolution for AI Infrastructure

OPEN ROUTING ENGINE

FRRouting (FRR)

Linux-native BGP/IS-IS/OSPF suite powering hyperscaler data centers

FRR 10.3: SRv6 uSID Static — enables AI Backend & SDN deployments

FRR 10.5: multi-locator, F4816 format, explicit BGP SID allocation

Contributors: Cisco, NVIDIA, 6WIND

FRR 10.5
Latest Release

DATA PLANE PROTOCOL

SRv6 uSID

Cisco pioneering SR since 2012; SRv6 at IETF 2017

Deterministic GPU-to-GPU path control — outperforms VXLAN & MPLS

Deployed: Alibaba Cloud, Microsoft AI Backend

Eliminates overlay gateways — pure IPv6, stateless, end-to-end

60+
Global Operators

NETWORK OS

SONiC

Open disaggregated NOS adopted by hyperscalers at scale

SONiC 202505: 122 PRs merged — Cisco #1 contributor (37 PRs)

PhoenixWing Initiative (2024): Cisco & Alibaba drive SRv6 into SONiC

Ecosystem: Alibaba, Google, Intel, Microsoft, NVIDIA

350+ PRs
Cisco OSS Commits since 2021

SONiC on Cisco 8000

DC Fabric Architectures

High Performance switches & optics
(100G / 400G / 800G)

Optimized AI/ML network buildouts
Back-end / Front-end / Storage

Highly scalable and flexible fabric architectures

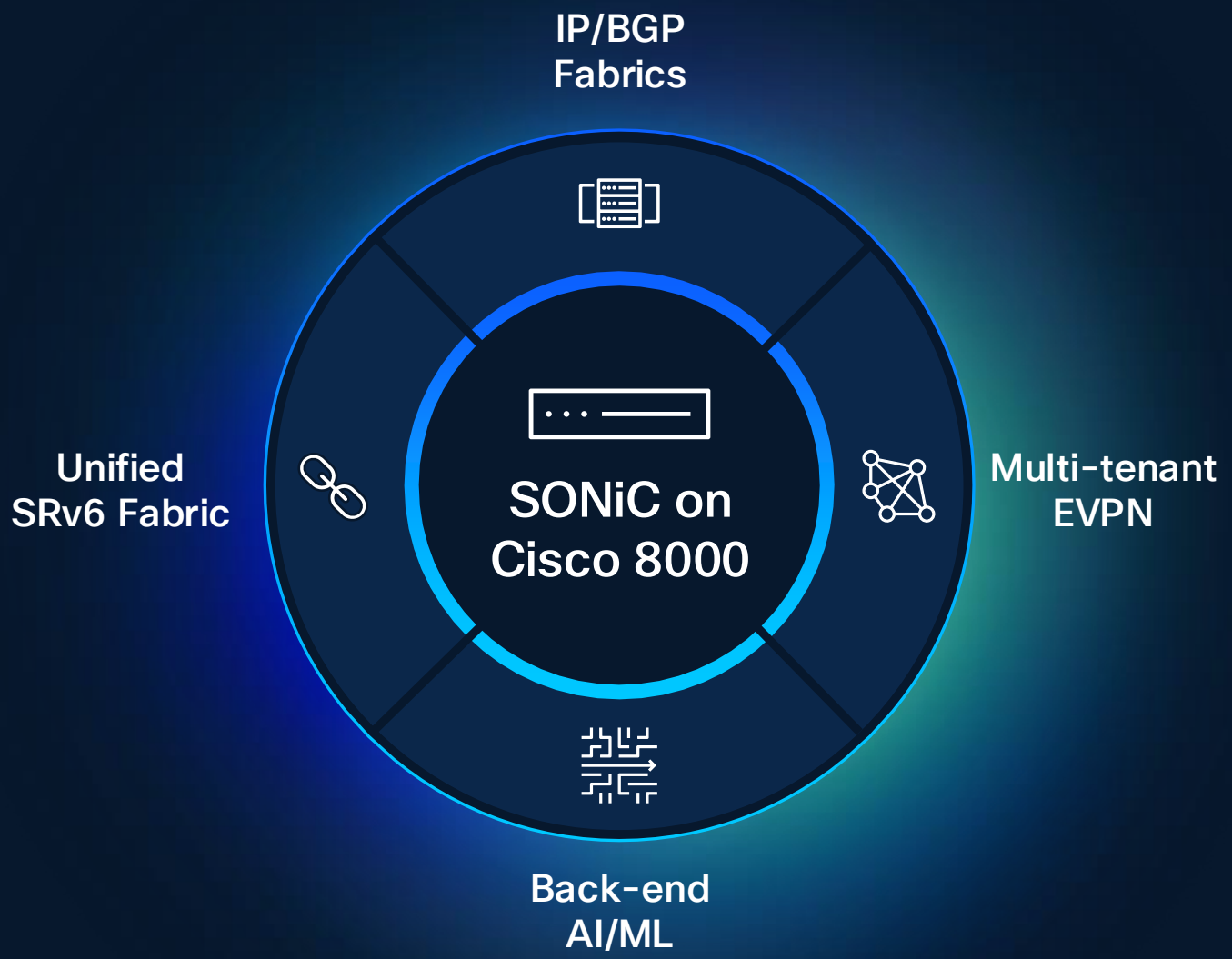
Enhanced programmability & open APIs for automation

High bandwidth fabrics with reduced footprint

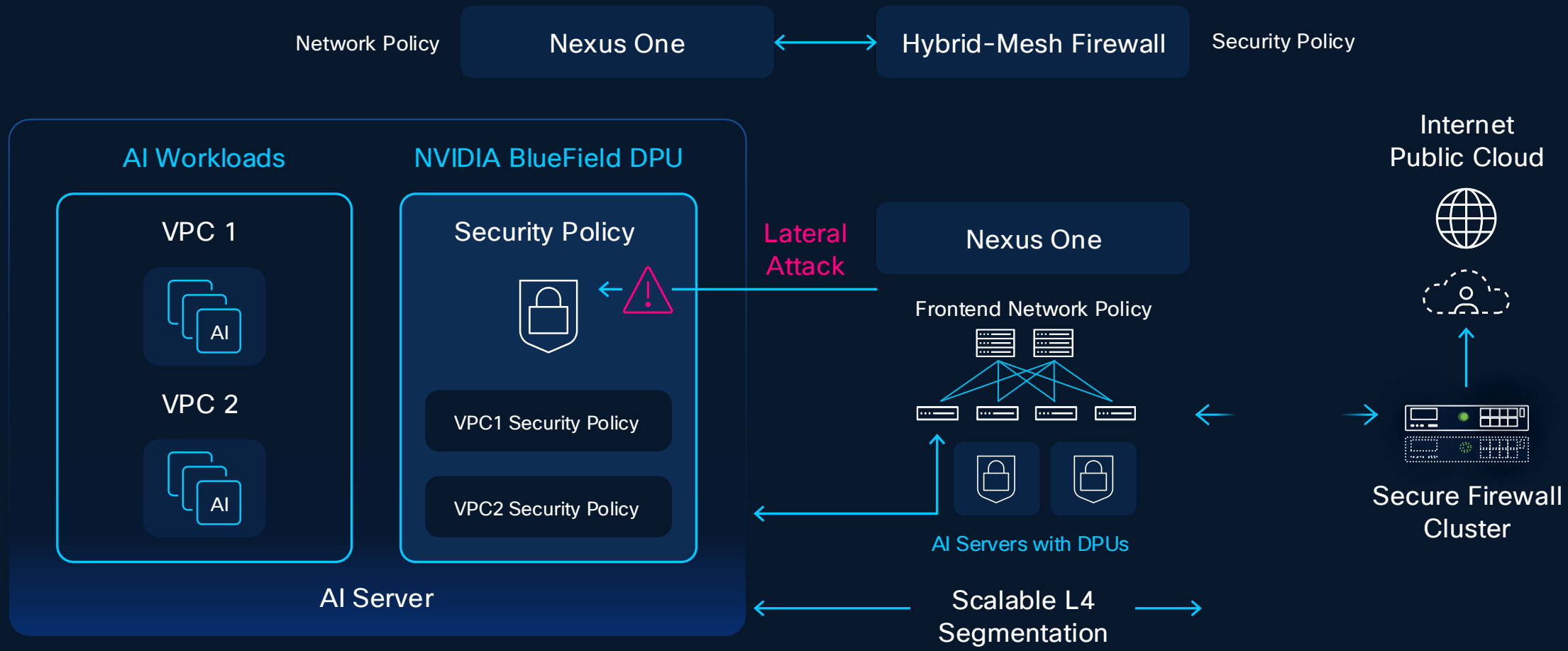
Observability, deep visibility with hardware and software telemetry

350,000+ Units

SONiC on Cisco 8000 deployment footprint



AI Workload Protection for Frontend Fabric Servers



Distribute policies close to workloads without requiring an additional firewall

Automate policy deployment based on VPC attachment

Enable scalable network segmentation

Offload network functions to preserve CPU resources for AI workloads

Fireside chat

Will Eatherton

SVP, Infrastructure Engineering

Andrew Leece

Co-founder

Cisco & NVIDIA Partnership

1. Multi-year partnership
2. Ethernet for AI
3. Cisco Nexus Hyperfabric is shipping
4. Security services fused into the AI infrastructure


CISCO