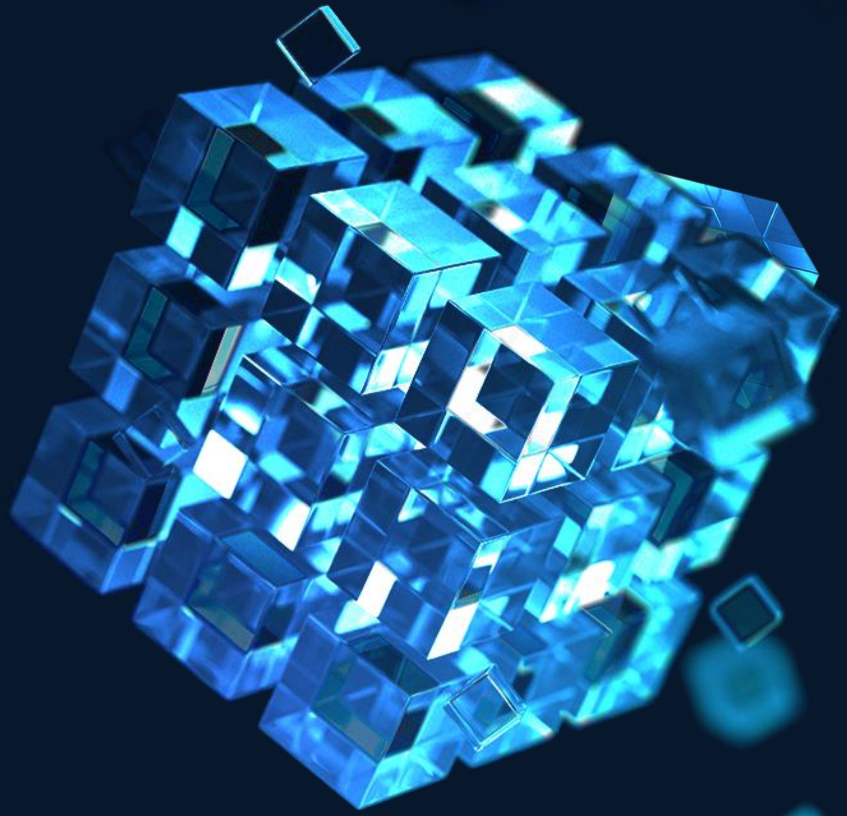
From Complexity to Confidence: Core-to-Edge Secure AI Factories with Consistent Operations

Kevin Wollenweber
SVP/GM, Data Center and Internet Infrastructure

Where does complexity live?

Hardware is only part of the story



Validation cycles

GPU cost

Incident debugging

Engineering hours

Compliance overlays

Hardware spend

Energy inefficiencies

Tuning time

Cisco is on a journey to making AI simpler

A full stack approach changes your starting point

Enables to start from a use case vantage point

Pre-validated architectures

Operationalize from day one

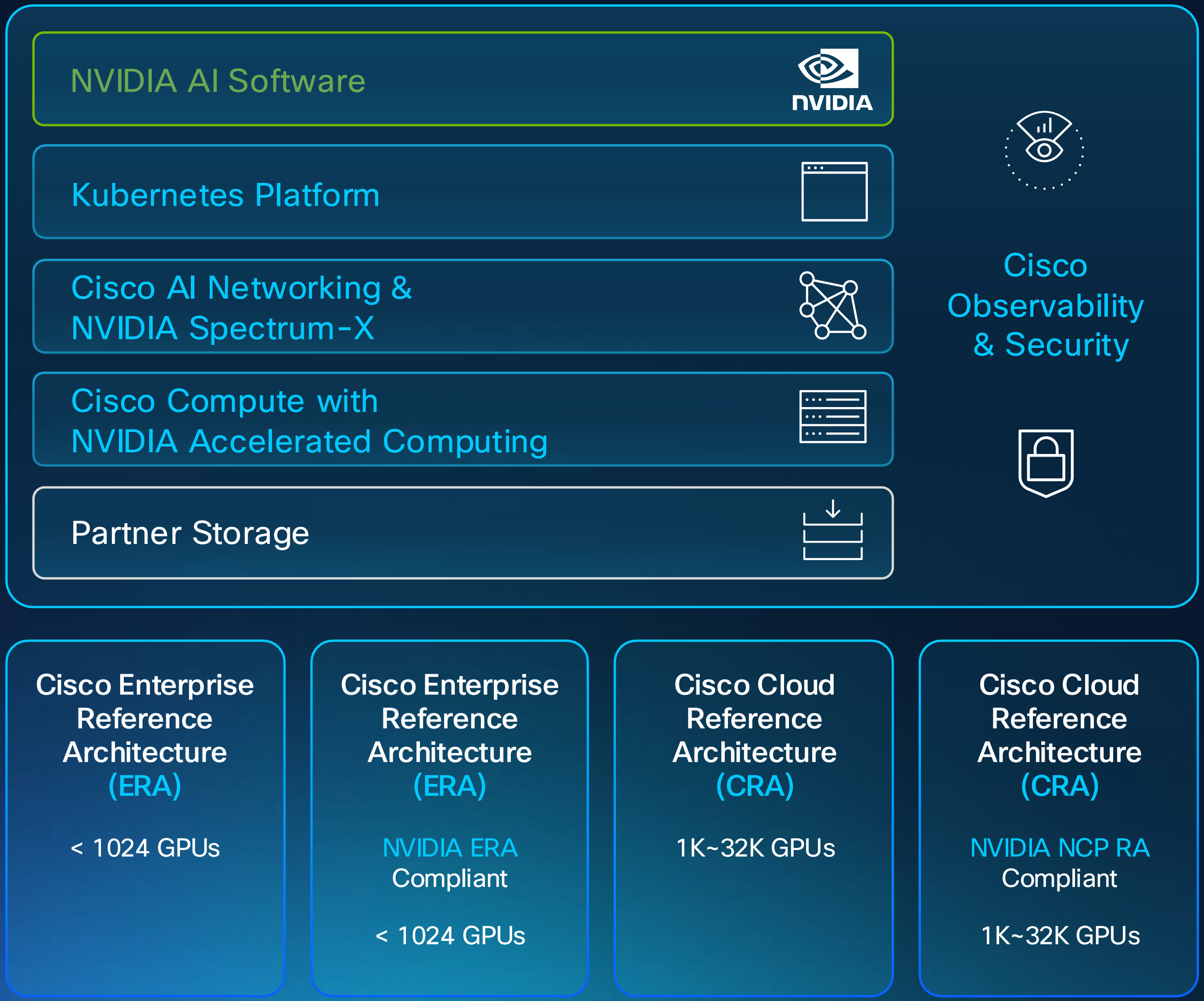
Reduced integration cycles

Accountability

CISCO | NVIDIA

Cisco Secure AI Factory with NVIDIA

Secure. Scalable. Simple.



NVIDIA AI Software

Kubernetes Platform

Cisco AI Networking & NVIDIA Spectrum-X

Cisco Compute with NVIDIA Accelerated Computing

Partner Storage

Cisco Observability & Security

Cisco Enterprise Reference Architecture (ERA)

< 1024 GPUs

Cisco Enterprise Reference Architecture (ERA)

NVIDIA ERA Compliant

< 1024 GPUs

Cisco Cloud Reference Architecture (CRA)

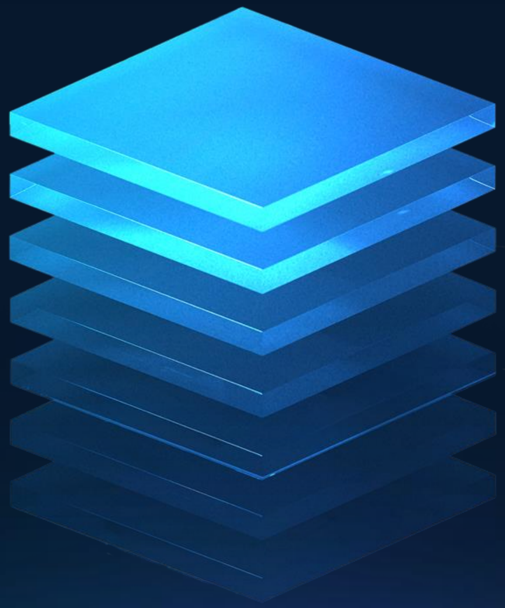
1K~32K GPUs

Cisco Cloud Reference Architecture (CRA)

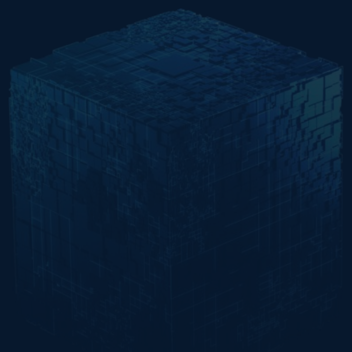
NVIDIA NCP RA Compliant

1K~32K GPUs

Scalable



Secure

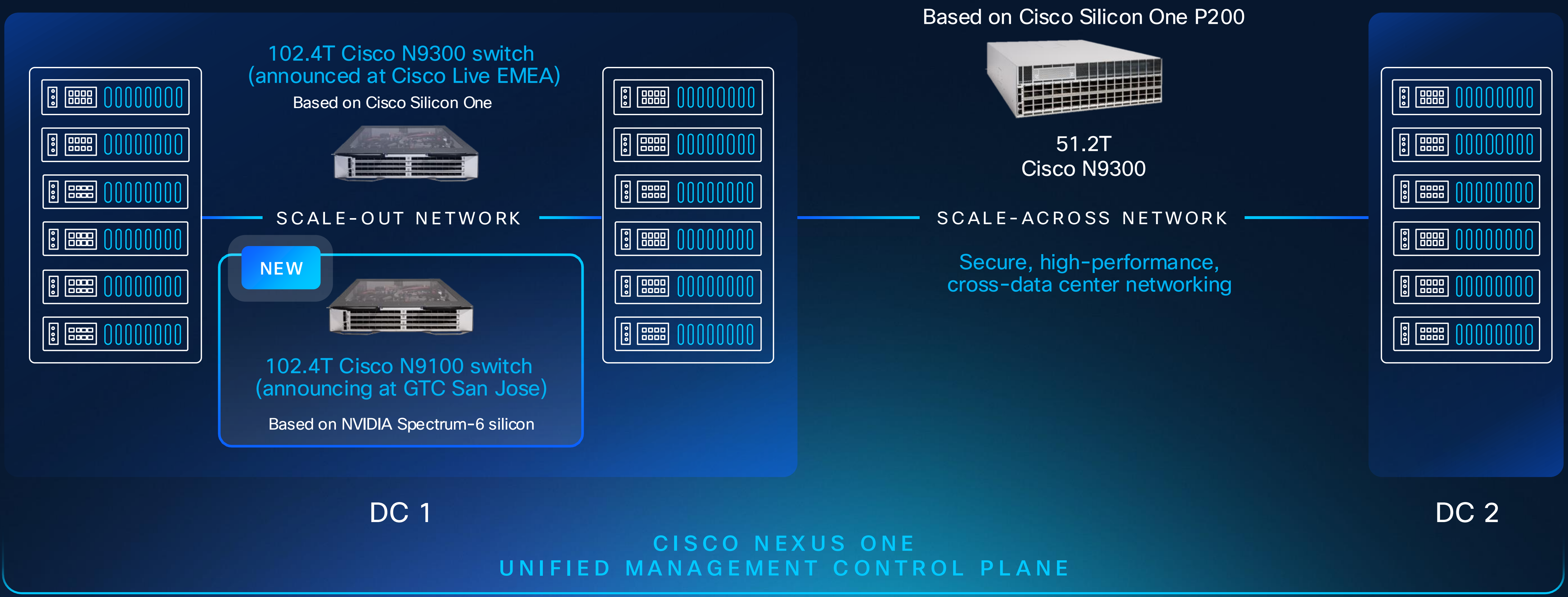


Simple

Sovereign & Neocloud AI at Giga-Scale

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

Choice of NCP and Cisco CRA Compliant architectures



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

Based on Cisco Silicon One

SCALE-OUT NETWORK

NEW

102.4T Cisco N9100 switch
(announcing at GTC San Jose)

Based on NVIDIA Spectrum-6 silicon

Based on Cisco Silicon One P200

51.2T
Cisco N9300

SCALE-ACROSS NETWORK

Secure, high-performance,
cross-data center networking

DC 1

DC 2

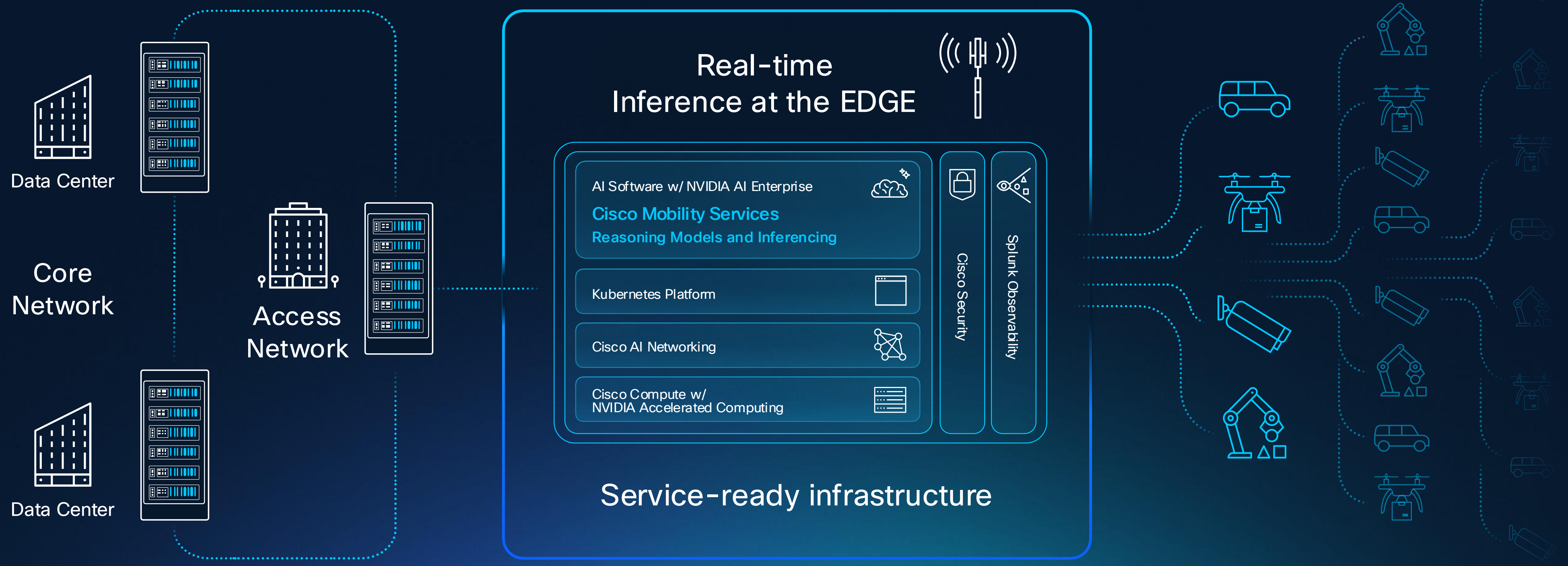
CISCO NEXUS ONE
UNIFIED MANAGEMENT CONTROL PLANE

Extending AI Factory to the Edge

AI Factories at the Telco Edge for distributed inferencing

Cisco Secure AI Grid powered by NVIDIA

AT&T



Data Center

Core Network

Data Center

Access Network

Real-time Inference at the EDGE

AI Software w/ NVIDIA AI Enterprise

Cisco Mobility Services

Reasoning Models and Inferencing

Kubernetes Platform

Cisco AI Networking

Cisco Compute w/ NVIDIA Accelerated Computing

Cisco Security

Splunk Observability

Service-ready infrastructure

End Point Swarm
5G/6G Devices

SIMPLE, SECURE INTELLIGENT CONNECTIVITY

Broad Spectrum of Cisco accelerated computing from Core to Edge

Cisco Intersight Unified Fleet Management

TRAINING

Dense GPU Server

C845A M8

GPUs supported:
NVIDIA RTX Pro 6000,
H100, H200,
AMD MI210, L40S

OPTIMIZATION & INFERENCING

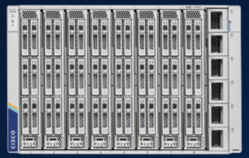
Rack Server

C245 M8 /
C240 M8

GPUs supported:
NVIDIA RTX Pro 6000,
H100, H200,
A16, L40S, L4

Modular Servers

X-Series



GPUs supported:
NVIDIA RTX Pro 6000,
H100, H200,
A16, L40S, L4

INFERENCING

Unified Edge

GPUs supported:
NVIDIA L40S

NEW

NVIDIA RTX Pro 4500 GPUs
Price, performance efficiency for inferencing and data analytics

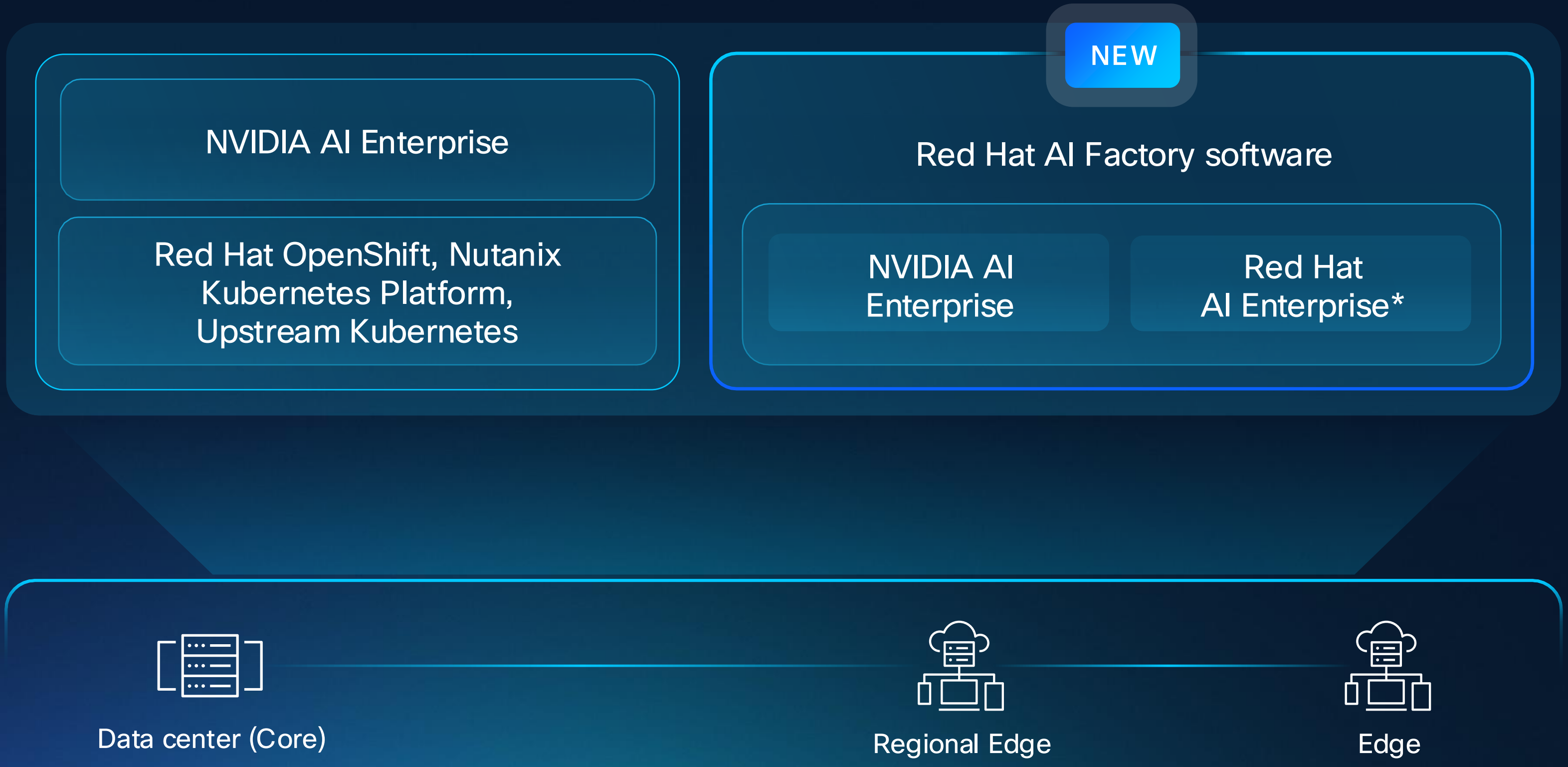
Data center (Core)

Regional Edge

Edge

Extending Choice of AI Software

Broader choice of tooling for AI practitioners to speed up development and delivery of AI applications



NVIDIA AI Enterprise

Red Hat OpenShift, Nutanix Kubernetes Platform, Upstream Kubernetes

NEW

Red Hat AI Factory software

NVIDIA AI Enterprise

Red Hat AI Enterprise*

* Includes Red Hat OpenShift Container Platform

Secure

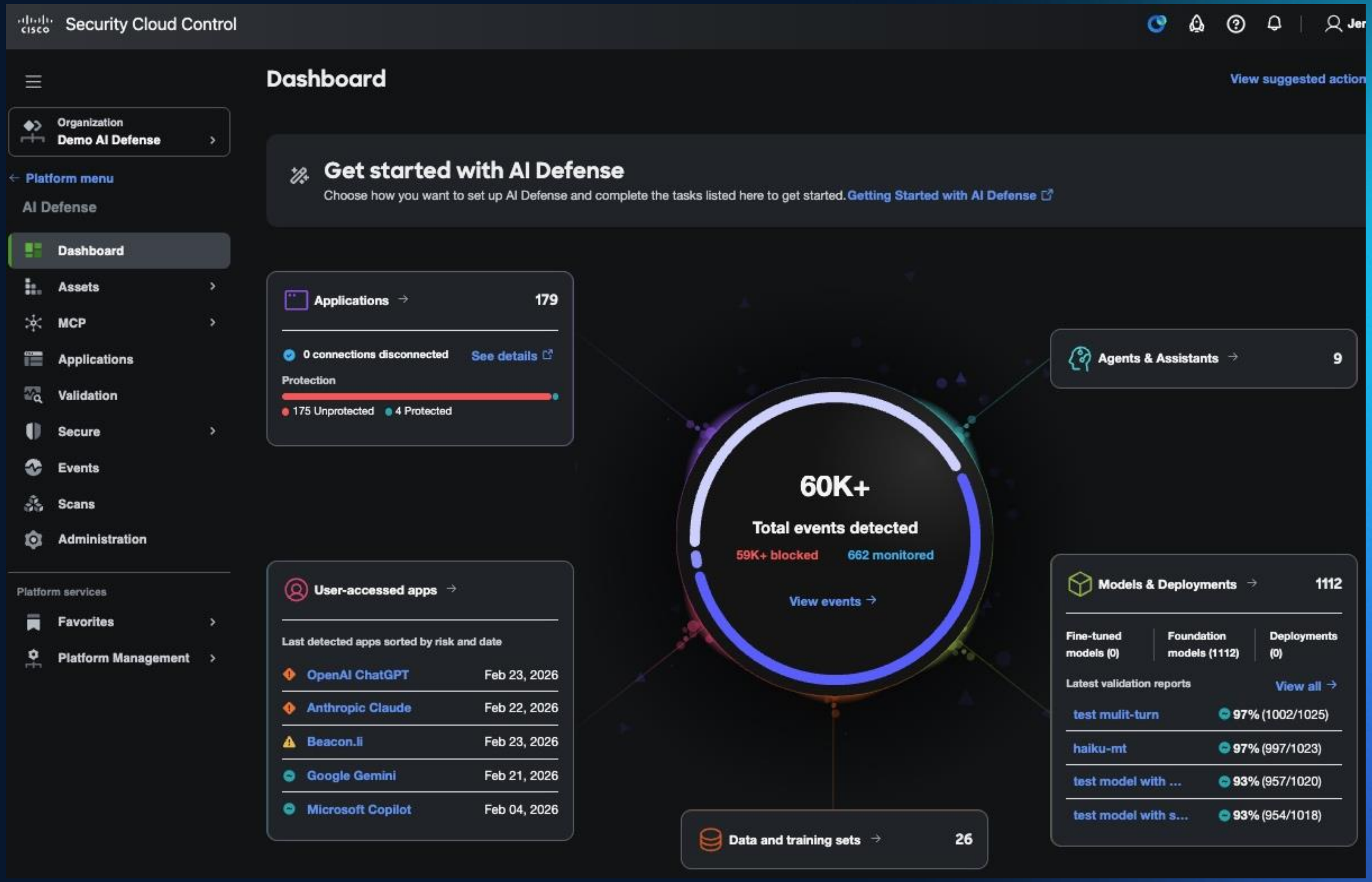
Scalable

Simple

Safeguard AI models, agents, and applications from safety and security risks with Cisco AI Defense

Identify vulnerabilities

Mitigates threats in real time



Dashboard

Get started with AI Defense

Choose how you want to set up AI Defense and complete the tasks listed here to get started. Getting Started with AI Defense

Applications → 179

0 connections disconnected See details

Protection

● 175 Unprotected ● 4 Protected

60K+

Total events detected

59K+ blocked 662 monitored

View events →

Agents & Assistants → 9

User-accessed apps →

Last detected apps sorted by risk and date

OpenAI ChatGPT	Feb 23, 2026
Anthropic Claude	Feb 22, 2026
Beacon.li	Feb 23, 2026
Google Gemini	Feb 21, 2026
Microsoft Copilot	Feb 04, 2026

Data and training sets → 26

Models & Deployments → 1112

Fine-tuned models (0)	Foundation models (1112)	Deployments (0)

Latest validation reports View all →

test mulit-turn	97% (1002/1025)
haiku-mt	97% (997/1023)
test model with ...	93% (957/1020)
test model with s...	93% (954/1018)

NEW

Security close to every workload

Hybrid Mesh Firewall

Cisco Hybrid Mesh Firewall extends security policy enforcement on NVIDIA BlueField DPUs on AI servers to ensure protection without negatively impacting CPU & GPU performance.

Hybrid Mesh Firewall

Cisco Firewalls

Physical | Virtual | Cloud | FWaaS

3rd Party Firewalls

NVIDIA Bluefield DPU Firewall on AI Servers

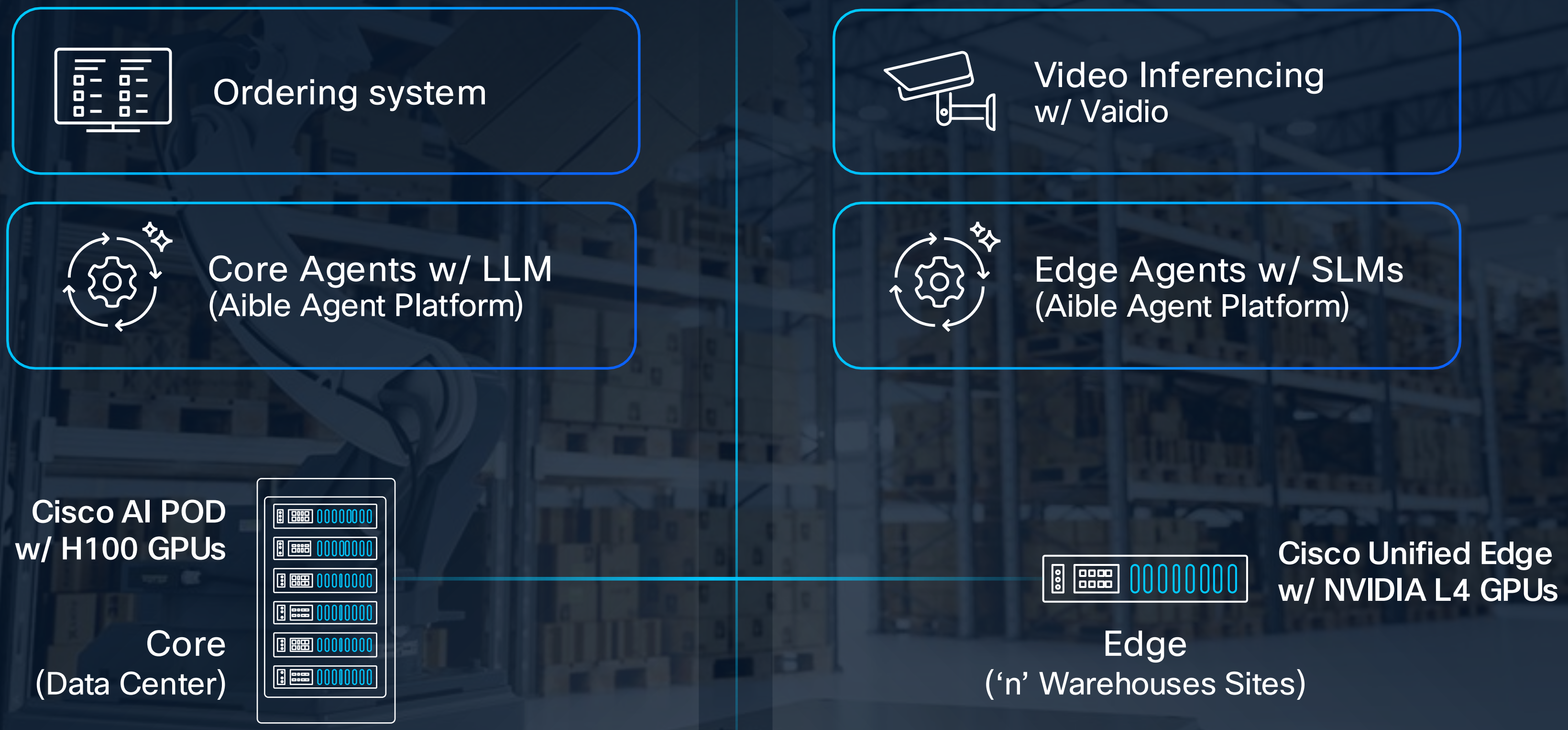
Smart Switches

Workload Agents

Define policy once, enforce everywhere

New solution to showcase secure multi-agent system use case

Use Case: Intelligent Warehouse with secure multi-agent operations



Ordering system

Core Agents w/ LLM
(Aible Agent Platform)

Video Inferencing
w/ Vaidio

Edge Agents w/ SLMs
(Aible Agent Platform)

Cisco AI POD
w/ H100 GPUs

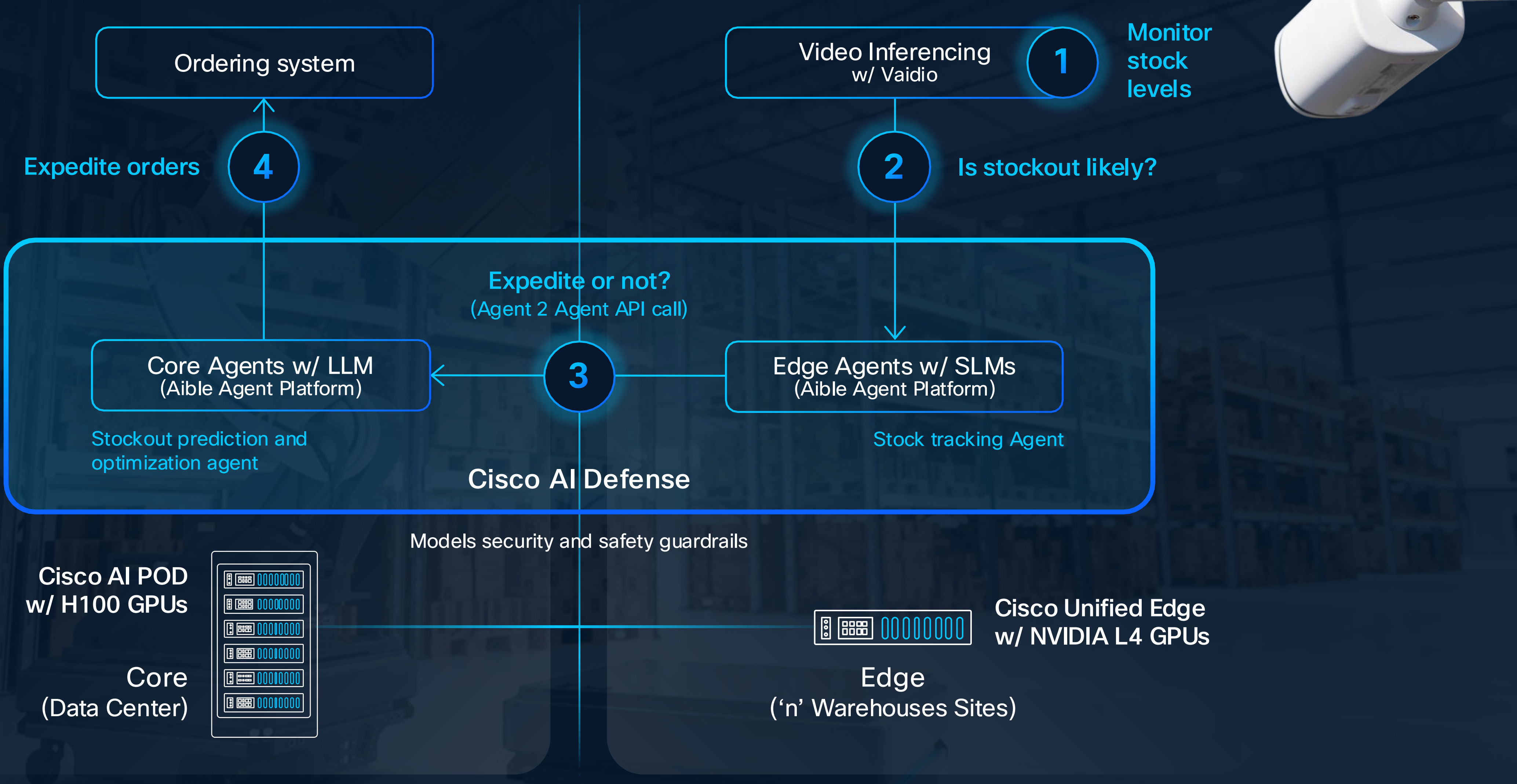
Core
(Data Center)

Cisco Unified Edge
w/ NVIDIA L4 GPUs

Edge
('n' Warehouses Sites)

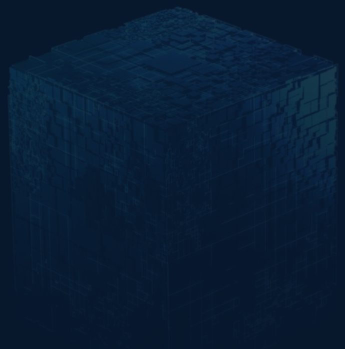
New solution to showcase secure multi-agent system use case

Use Case: Intelligent Warehouse with secure multi-agent operations



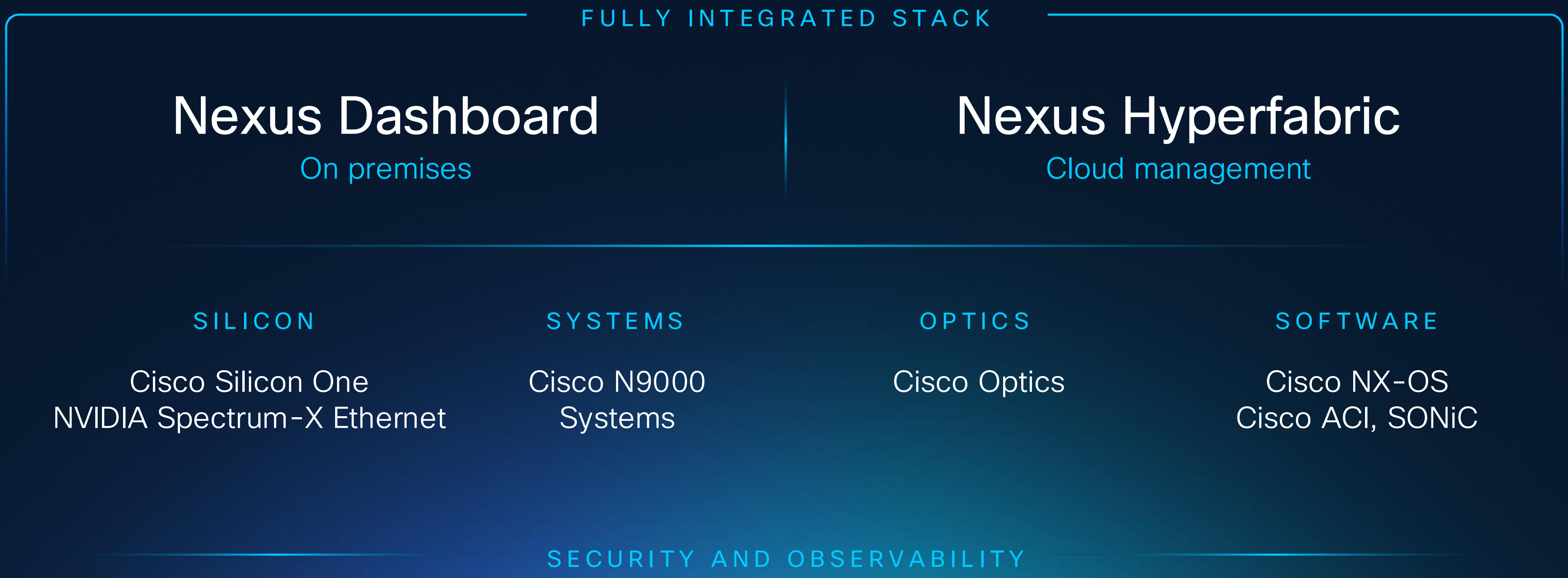
Simple

Secure



Scalable

Unified operations with Cisco Nexus One



FULLY INTEGRATED STACK

Nexus Dashboard

On premises

Nexus Hyperfabric

Cloud management

SILICON

Cisco Silicon One
NVIDIA Spectrum-X Ethernet

SYSTEMS

Cisco N9000
Systems

OPTICS

Cisco Optics

SOFTWARE

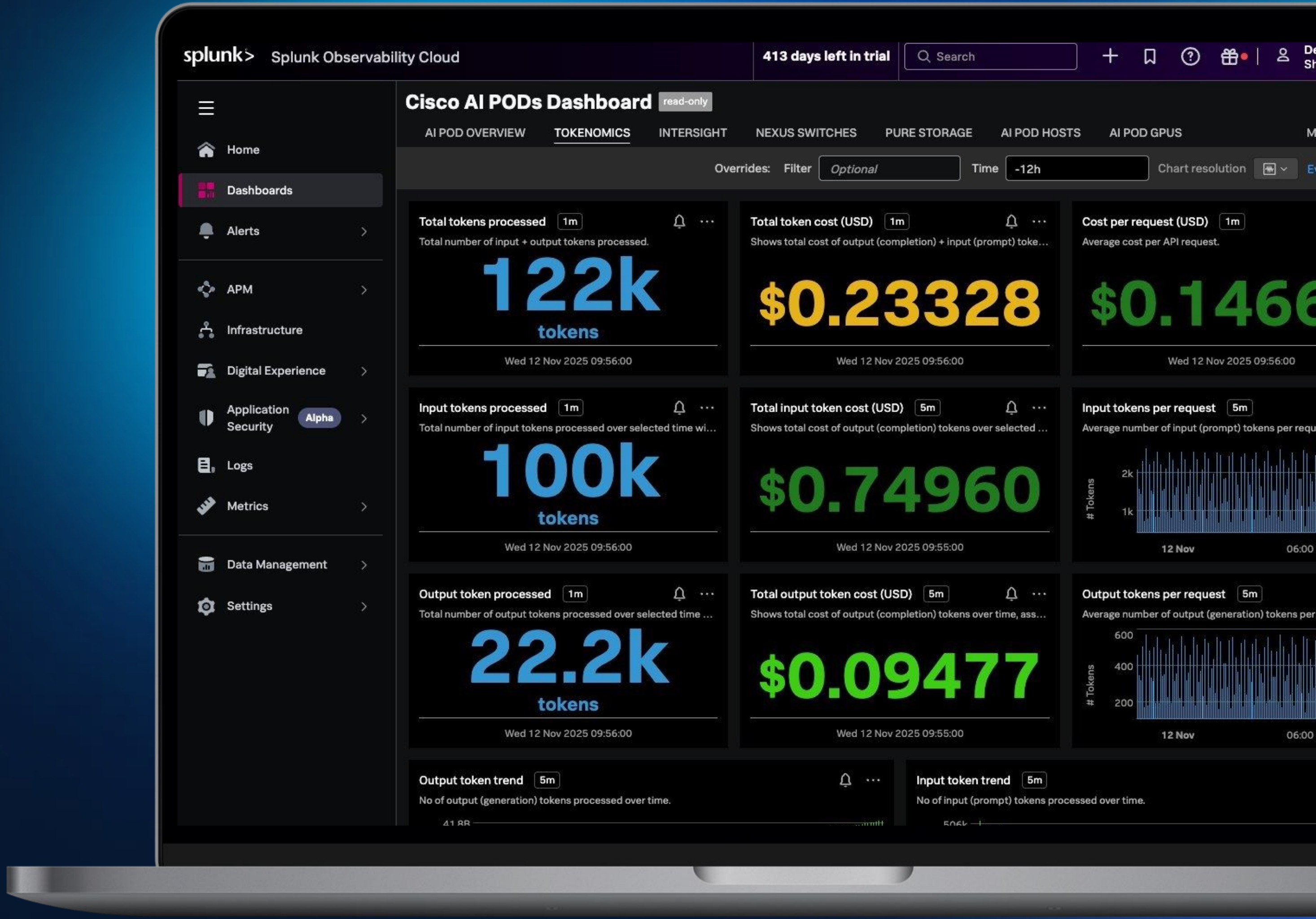
Cisco NX-OS
Cisco ACI, SONiC

SECURITY AND OBSERVABILITY

Splunk Observability for full stack monitoring and reporting

Tokenomics

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



Cisco AI PODs Dashboard read-only

Overrides: Filter Optional Time -12h Chart resolution

Total tokens processed 1m
Total number of input + output tokens processed.
122k tokens
Wed 12 Nov 2025 09:56:00

Total token cost (USD) 1m
Shows total cost of output (completion) + input (prompt) toke...
$0.23328
Wed 12 Nov 2025 09:56:00

Cost per request (USD) 1m
Average cost per API request.
$0.1466
Wed 12 Nov 2025 09:56:00

Input tokens processed 1m
Total number of input tokens processed over selected time wi...
100k tokens
Wed 12 Nov 2025 09:56:00

Total input token cost (USD) 5m
Shows total cost of output (completion) tokens over selected ...
$0.74960
Wed 12 Nov 2025 09:55:00

Input tokens per request 5m
Average number of input (prompt) tokens per requ

Output token processed 1m
Total number of output tokens processed over selected time ...
22.2k tokens
Wed 12 Nov 2025 09:56:00

Total output token cost (USD) 5m
Shows total cost of output (completion) tokens over time, ass...
$0.09477
Wed 12 Nov 2025 09:55:00

Output tokens per request 5m
Average number of output (generation) tokens per

Output token trend 5m
No of output (generation) tokens processed over time.

Input token trend 5m
No of input (prompt) tokens processed over time.

NEW

AgenticOps for Data Center with AI Canvas

Cisco Deep Network Model

Domain-Specific LLM · 20% higher precision on networking tasks

Next frontier of operations

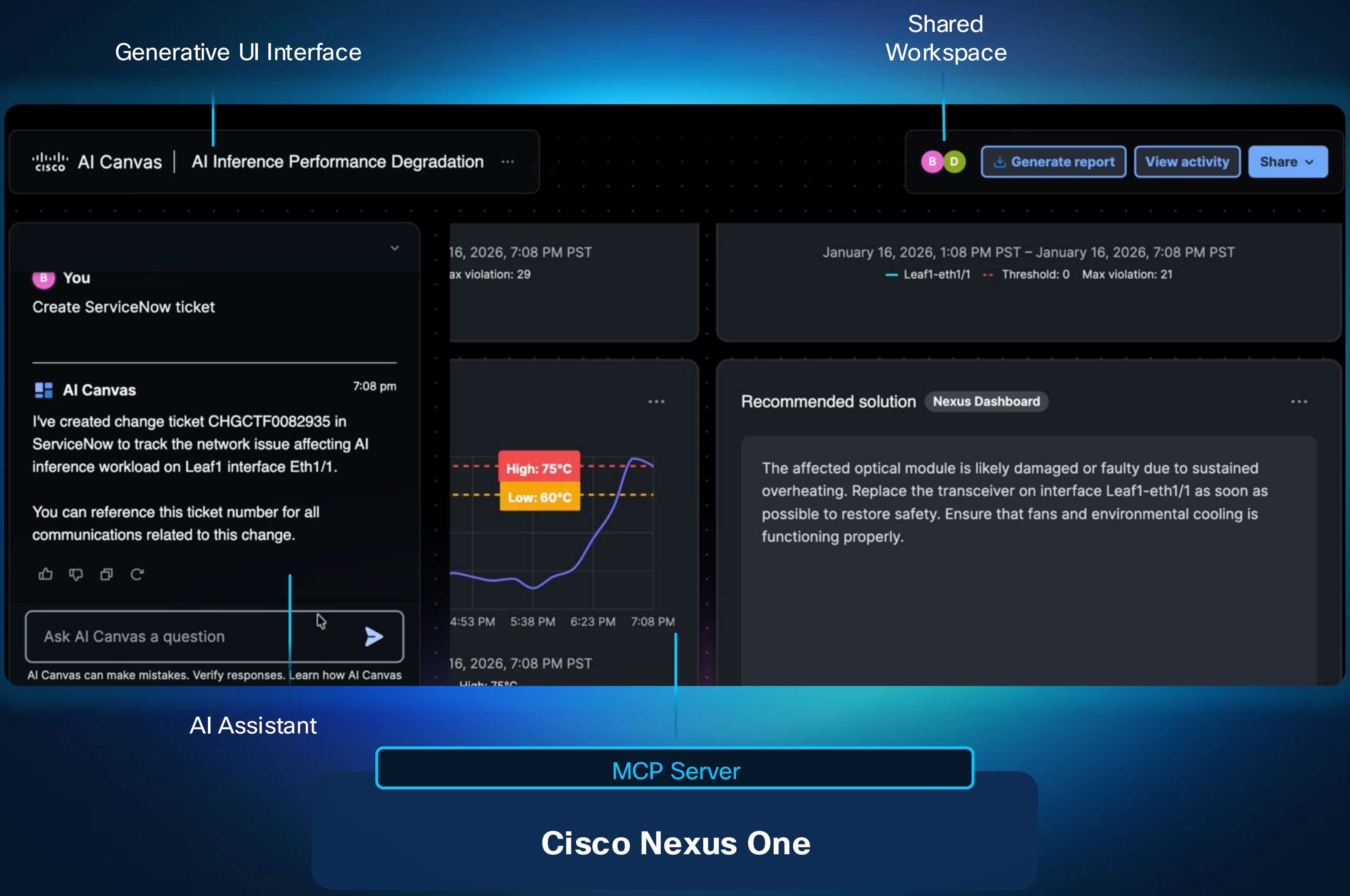
Contextual data center operations with AgenticOps capabilities for troubleshooting

Precision reasoning

Powered by purpose-built deep network models

Breaking silos

Collaborate and correlate across entire infrastructure using natural language



16, 2026, 7:08 PM PST
ax violation: 29

January 16, 2026, 1:08 PM PST – January 16, 2026, 7:08 PM PST
Leaf1-eth1/1 Threshold: 0 Max violation: 21

You
Create ServiceNow ticket

AI Canvas 7:08 pm
I've created change ticket CHGCTF0082935 in ServiceNow to track the network issue affecting AI inference workload on Leaf1 interface Eth1/1.

You can reference this ticket number for all communications related to this change.

Ask AI Canvas a question

AI Canvas can make mistakes. Verify responses. Learn how AI Canvas

16, 2026, 7:08 PM PST
High: 75°C

Recommended solution

The affected optical module is likely damaged or faulty due to sustained overheating. Replace the transceiver on interface Leaf1-eth1/1 as soon as possible to restore safety. Ensure that fans and environmental cooling is functioning properly.

Flexible deployment options

Build your own

Buy and deploy individual products, as needed

AI POD with on-prem network management

Buy & deploy full stack

Modular, pre-validated

Backed by CVDs

NVIDIA ERA compliant

AI POD with cloud based network management

Buy and deploy full stack

Nexus Hyperfabric for cloud-managed networking or turnkey physical infrastructure

NVIDIA ERA and NCP RA compliant

Faster time to first intelligent and ROI with professional services from Cisco and partners, globally

Fireside chat

Kevin Wollenweber

SVP/GM, Data Center and
Internet Infrastructure

Cisco

Saul Mankes

VP, Digital Platforms and
Infrastructure

CVS

cisco