cisco Engage

Innovations in Data Center Compute

John Rice, CAI Solutions Engineer

Sam Aguirre, Solutions Engineer

October 15, 2025

Agenda

1 Compute Portfolio Overview
2 Introducing AI PODs
3 Intersight Demo
4 Q&A

Cisco Differentiation

The Security

Security-first architecture enables safe enterprise AI

The Network

High-performance integrated AI networking enables efficient model training and inferencing

The Assurance

Pre-validated AI infrastructure stack with flexible deployment options improves data scientists and developer productivity

Compute AI portfolio

Address AI workloads with visibility, consistency, and control

Validated solutions for AI with compute, network, storage, and software

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

RTX PRO SERVER
Supporting RTX PRO 6000 Blackwell Server Edition GPUs

Cisco UCS®
GPU-dense servers
PCIe and NVLink Servers

Cisco UCS blade (with GPU extensions) and rack servers

Enterprise AI edge

Dense compute for demanding AI

Full-stack AI with compute and networking

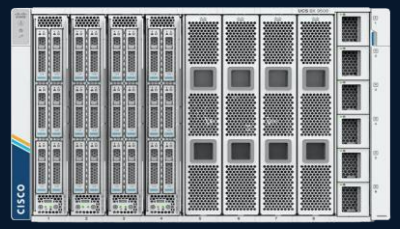
cisco

Cisco UCS Compute Portfolio

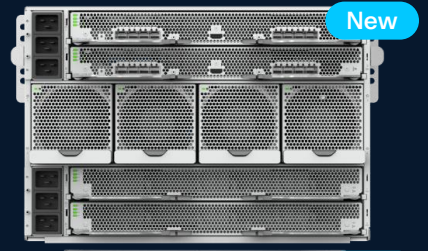
MAINSTREAM ENTERPRISE SERVERS

UCS X-Series X9508 Chassis

IFM Module



UCS X-Series Direct



UCS x580p M8

UCS X210c M7

UCS X210c M8

UCS X215c M8

UCS X410c M7

UCS C240 M8E3S
36 EDSFF E3.S1T

UCS C240 M8SX
28 HDD/SDD/NVMe

UCS C240 M8L
16 LFF + 4 SFF

UCS C240 M7SN
28 NVMe

UCS C240 M6S
14 SSD/HDD Media drive

UCS C240 M6N
14 NVMe Media Drive

UCS C220 M8E3S
16 EDSFF E3.S1T

UCS C220 M8S
10 HDD/SSD/NVMe

UCS C220 M7N
10 NVMe

UCS C245 M8SX
28 HDD/SDD

UCS C225 M8S
10 HDD/SSD

UCS C225 M8N
10 NVMe

AI SERVERS

UCS C885A M8
8RU Dense GPU Server

UCS C880A M8
10RU Dense GPU Server

UCS C845A M8
4RU MGX Server

cisco

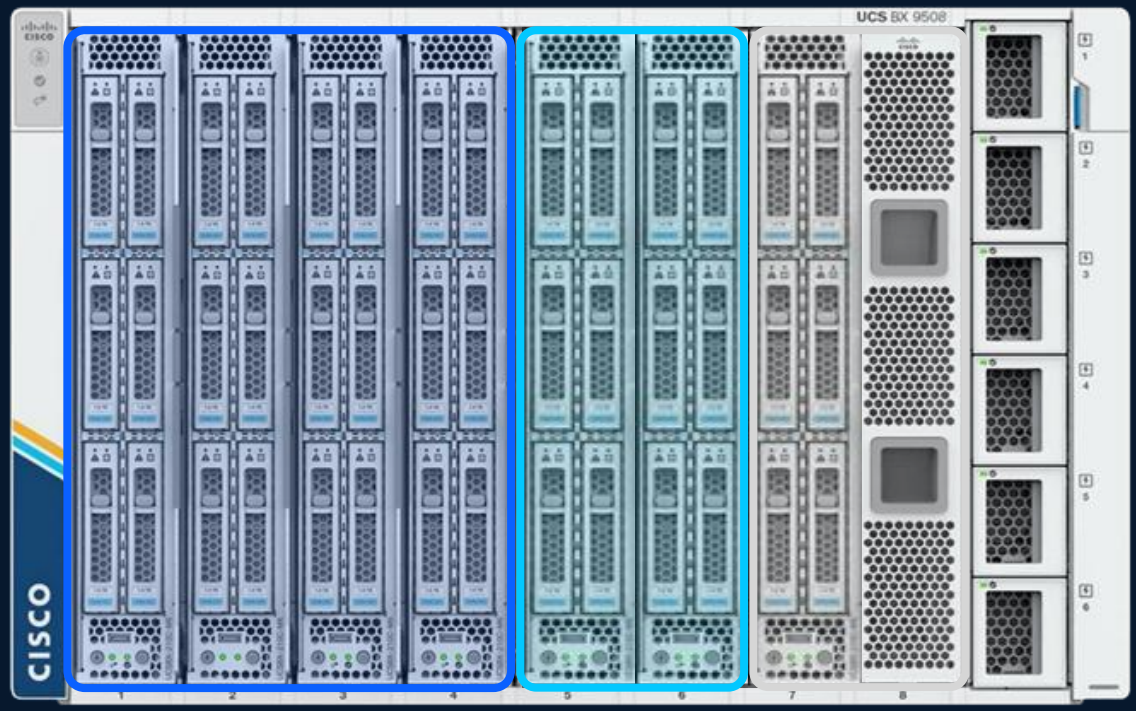
UCS® X-Series with X-Fabric

Consolidate rack workloads

- AI/ML
- Accelerated VDI
- Big Data, SDS, Containers

+

Traditional blade workloads



Up to 2,048 cores per chassis

24 GPUs per chassis

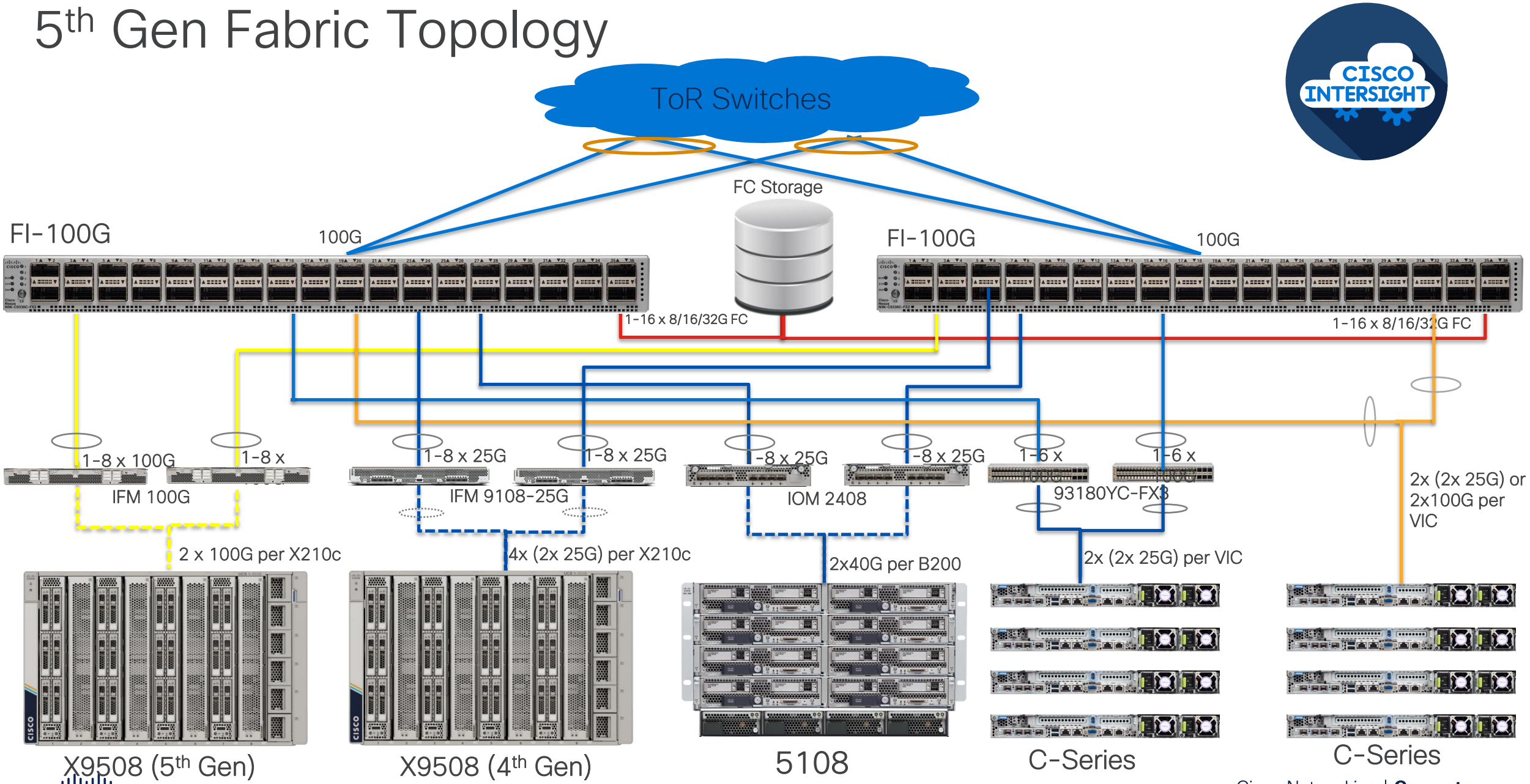
intel NVIDIA

200G bandwidth to compute node

736 TB of storage

cisco

5th Gen Fabric Topology



Cisco Networking | **Compute**

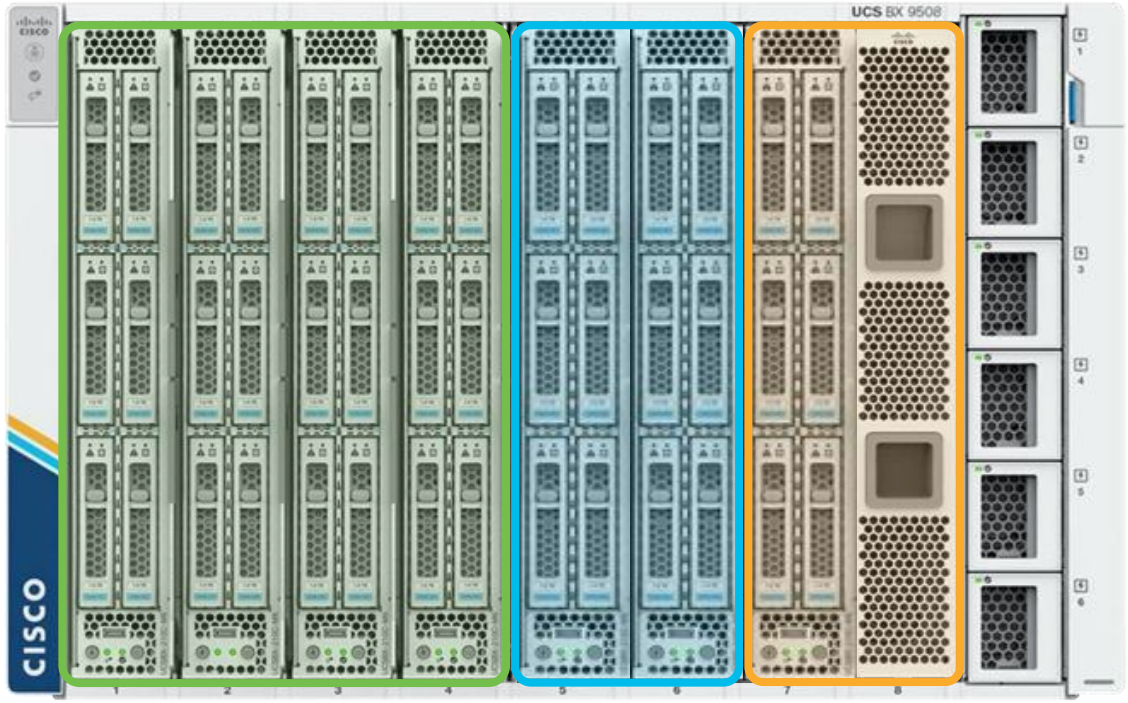
UCS® X-Series with X-Fabric

Consolidate rack workloads

- AI/ML
- Accelerated VDI
- Big Data, SDS, Containers

+

Traditional blade workloads



Up to 2,048 cores per chassis

24 GPUs per chassis

intel NVIDIA

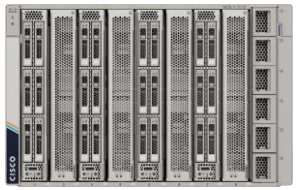
200G bandwidth to compute node

736 TB of storage

Cisco GPU-accelerated platforms offering

Plans are Subject to change

X-Series



Up to 24x HHHL GPUs or
8x FHFL GPUs per X9508 chassis

X210c M6/M7 2S Blades
2x NVIDIA T4 (MEZZ)

X210c M7 2S Blade (Q2'24)
Intel Flex140 (MEZZ)

X210c M7 2S Blade (Q3'24)
NVIDIA L4 (MEZZ)

X215c M8 2S Blade (Q3'24)
NVIDIA L4 (MEZZ)

X440p + X210c M6/M7
4x NVIDIA T4 (M6 Only)
2x NVIDIA A16
2x NVIDIA A40
2 x NVIDIA A100-80

X440p + M7 (X210c & X410c)
2x NVIDIA H100-80
2x NVIDIA L40
4x NVIDIA L4
2x NVIDIA L40S

X440p + M7 (X210c & X410c)
4x Intel Flex140
2x Intel Flex170

Plan (Q3'24)

X440p + X210c M7
2x NVIDIA H100-NVL

X440p + X215c M8 AMD
2x NVIDIA H100-NVL
2x AMD MI210
4x NVIDIA L4
2x NVIDIA L40S
2x NVIDIA L40
2x NVIDIA A16

C-Series Rack Servers

C240 M6 INTEL
C245 M6 AMD

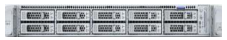
5x NVIDIA A10
3x NVIDIA A16
3x NVIDIA A30
3x NVIDIA A40
3x NVIDIA A100-80

8x NVIDIA L4
(C240 M6 only)

C240 M7 INTEL

3x NVIDIA A16
3x NVIDIA A30
3x NVIDIA A40
3x NVIDIA A100-80
2x NVIDIA H100-80
3x NVIDIA L40
8x NVIDIA L4
2x NVIDIA L40S
5x Intel Flex140
3x Intel Flex170

C220 M6 INTEL

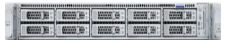


3x NVIDIA T4
3x NVIDIA L4

C225 M6 AMD

3x NVIDIA T4

C220 M7 INTEL



3x NVIDIA L4
3x Intel FLex140

C245 M8 AMD

Plan (2H'24)

NVIDIA H100-80
NVIDIA L40S
NVIDIA L40
NVIDIA L4
NVIDIA H100-NVL
NVIDIA A16
AMD MI210

C225 M8 AMD

Plan (2H'24)

3x NVIDIA L4

Please Refer to the Server Specifications and HCL for detailed configuration support:
C-Series: https://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-c-series-rack-servers/series.html#~tab-documents
X-Series: https://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-x-series-modular-system/series.html#~tab-documents
UCS HCL: https://ucshcltool.cloudapps.cisco.com/public/

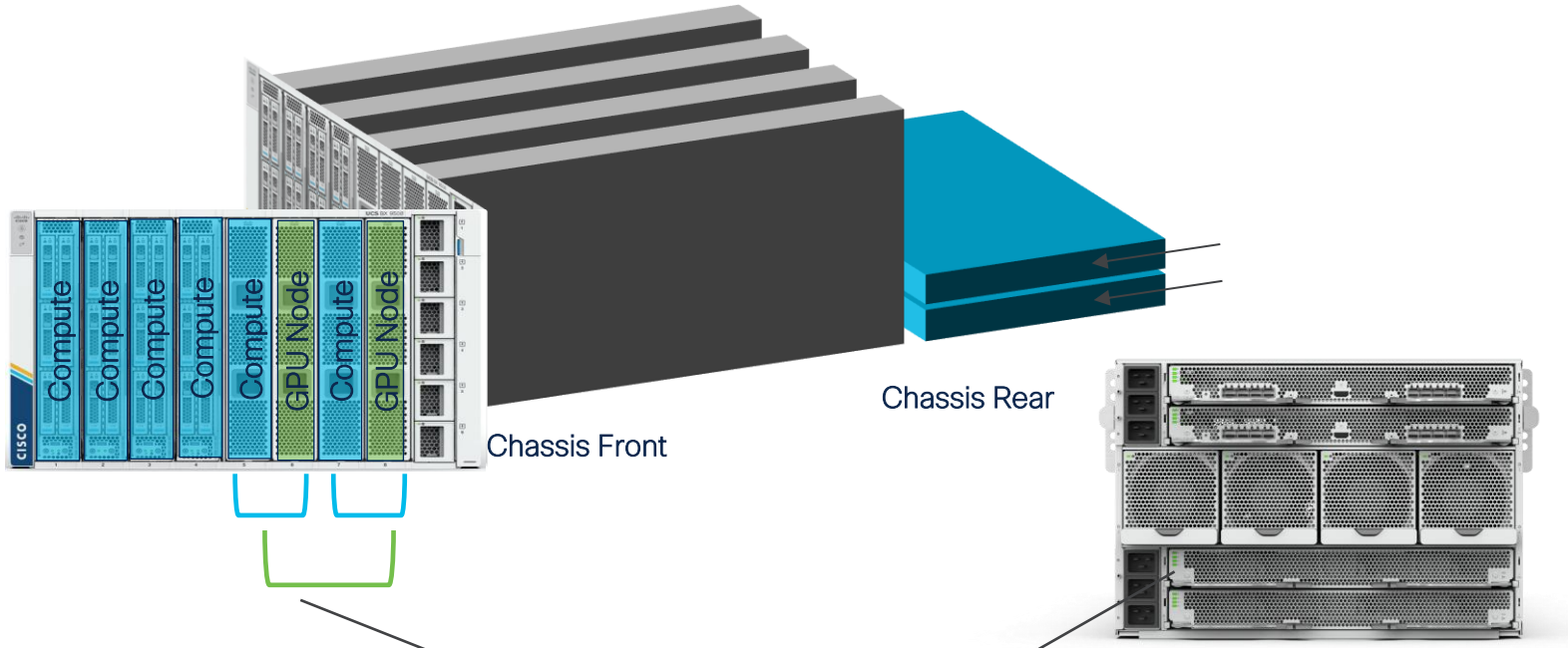
UCS X-Fabric Technology

Open, modular design enables compute and accelerator node connectivity

Open standards: PCIe 4/5/6, CXL*

No midplane nor cables = easy upgrades

Expandability to address new use cases in future (memory & storage nodes)



UCS X-Fabric Technology

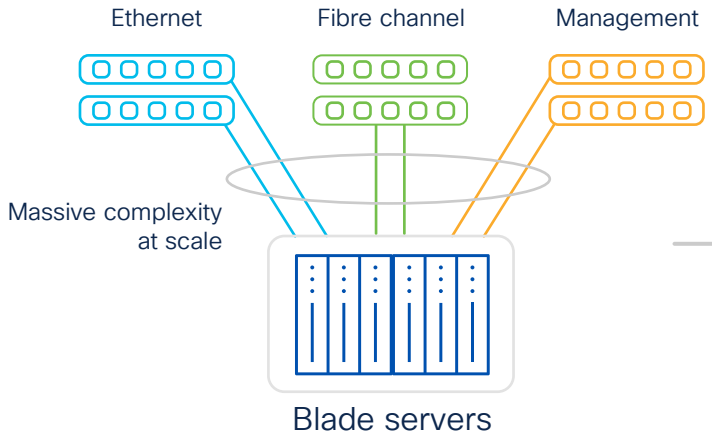
- Internal Fabric interconnects nodes
- Industry standard PCIe, CXL Traffic
- Upgrade to future generations

*CXL fabrics are dependent on future processors

Industry-leading simplicity

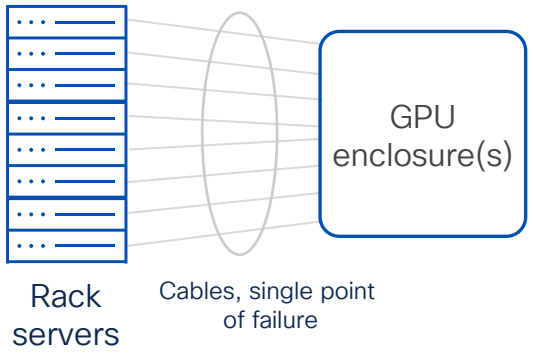
Conventional approaches

1 | Silos of multiple Ethernet and SAN fabrics and adapters



Unified fabric

2 | Complex PCIe connectivity to external accelerators

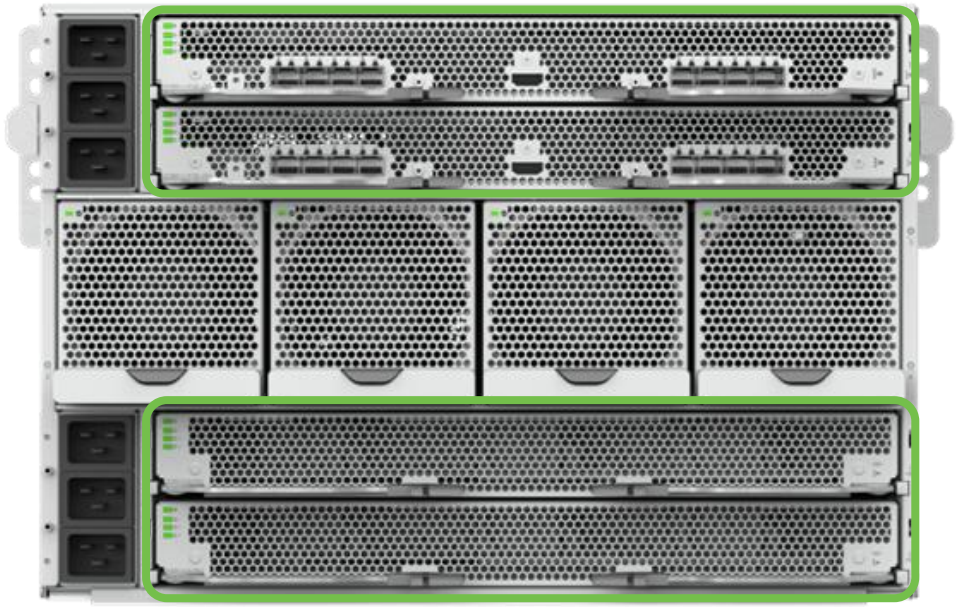


X-Fabric + PCIe nodes

Cisco solution

UCS® X-Series

Cisco Intersight®



Open Chassis Eliminates IO Midplane

Cisco UCS X9508 Modular System Chassis

Substantially higher performance per RU with flexibility for a long-term technology roadmap

Prevents technology lock-in

Reduced airflow restriction

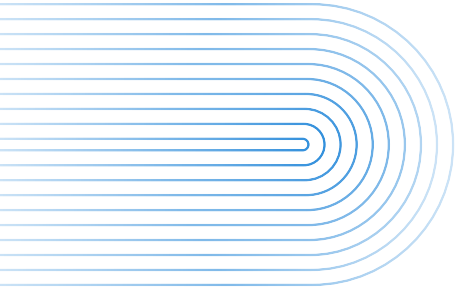
Maximize power efficiency

Configure as needed for workload

Energy Reduction Through Modernization

By replacing previous generation servers with UCS X-Series, a typical Cisco customer can expect:

70% reduction in total footprint



49% reduction in total power consumption

ner

More modernization benefits for customers

90% ↓

reduction in hardware operating costs

72% ↓

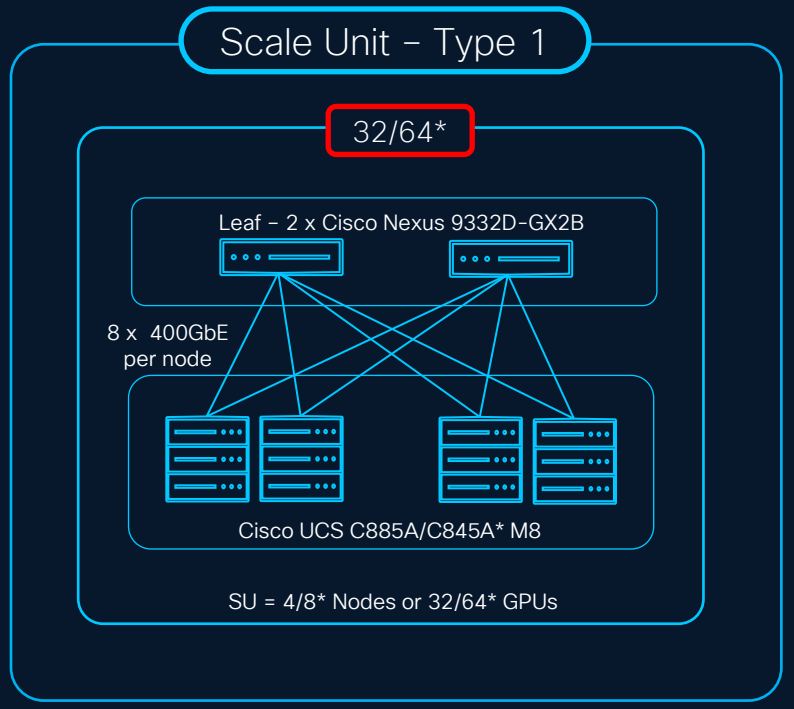
reduction in hardware maintenance costs

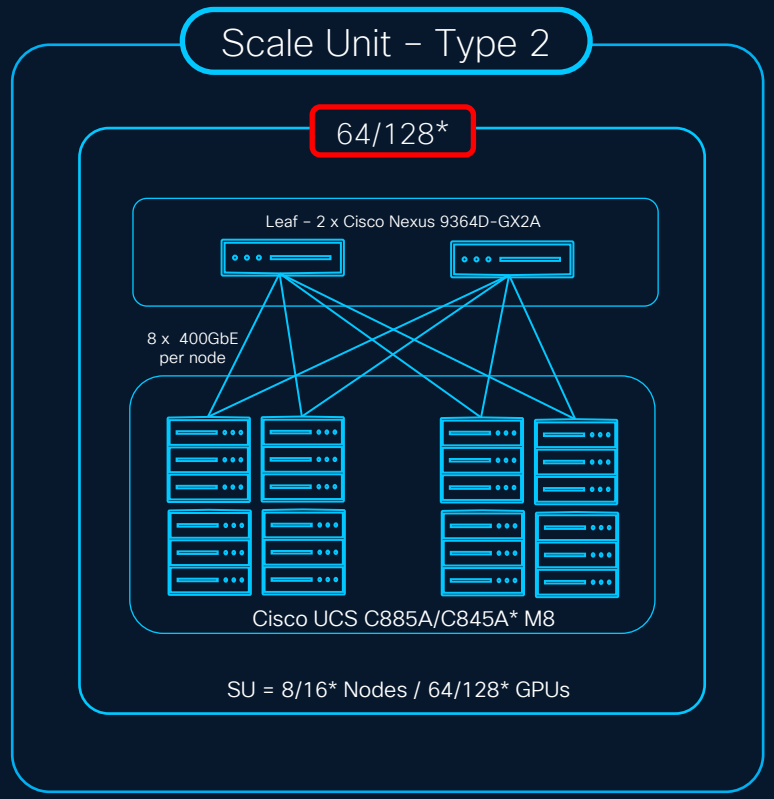
75% ↓

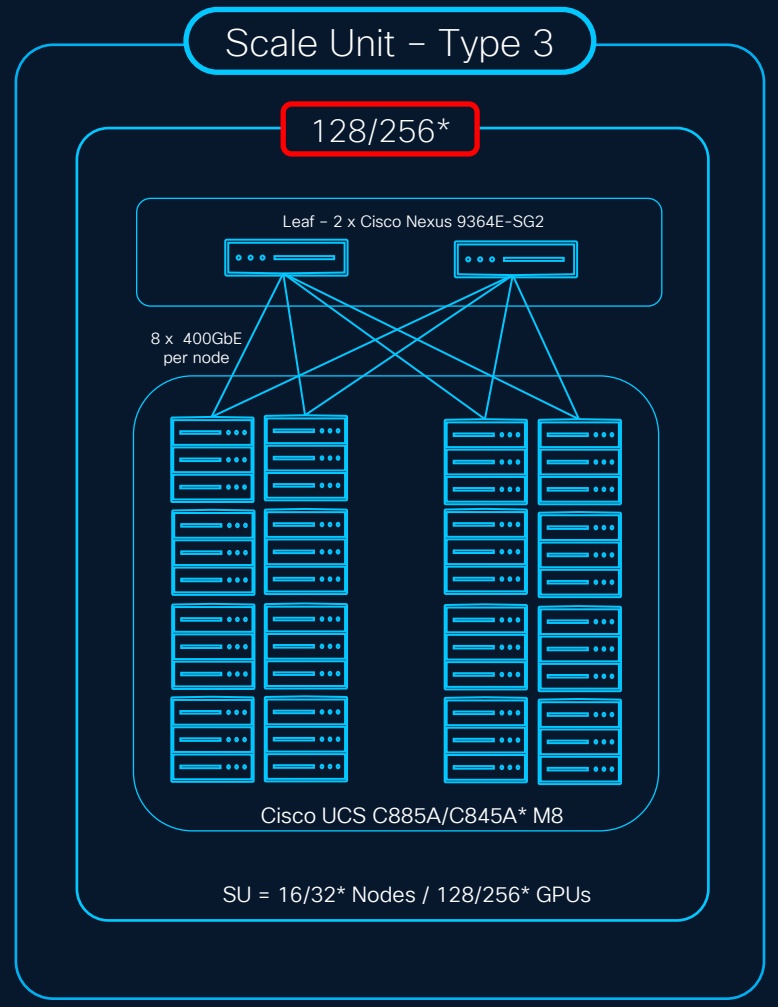
reduction in recurring software support costs

Scale Unit Types

Using UCS C885A or C845A







Note:

- Scale Unit Type: A pair of leaf switches + UCS nodes (leaf uplinks excluded)
- Non-blocking fabric design for max GPU performance
- Rail-Optimized Design within a scale unit
- *UCS C845A - 2 x GPUs share 1 E-W NIC → Node/GPU density is therefore higher per SU Type

Cisco Intersight®

Accelerate AI PODs provisioning and deployment at scale

CISCO INTERSIGHT®

Strengthen compliance

Increase uptime

Bolster security

Improve productivity

Accelerate operations

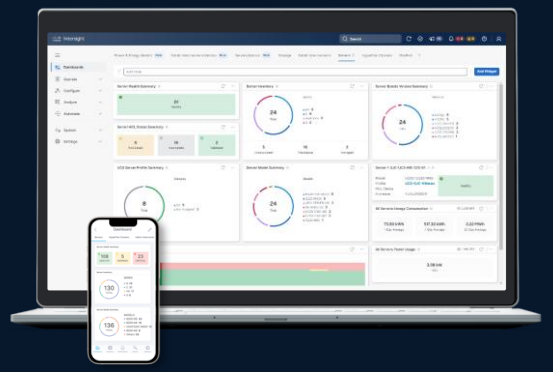
Control energy use

Control

your Cisco AI PODs and UCS® solutions from one place, 24x7x365

See

your Cisco AI PODs and other UCS® infrastructure in one dashboard



Automate

Compute deployments, configuration, workflows, and day-0 to day-N tasks

cisco

Common AI Challenges

Unclear business objectives & priorities

Unclear direction hinders cross team collaboration, creates confusion, and hampers acquisition of necessary skills

Complex AI infrastructure deployment

Lack of high-performance infrastructure with integrated compute, network, storage, and AI software can stall AI projects

Security vulnerabilities

AI models, frameworks, apps, and supporting infrastructure represent a new cyberattack surface

Network performance & Security challenges

Model training and inferencing generates a lot of traffic, slowing networks and also results in new attack surface

cisco

Introducing: Cisco AI PODs

A scalable architecture, built to support any AI workload simply & efficiently

Training ..... Optimization ..... Inference

AI PODs

BUILD TUNE DEPLOY

DATA PIPELINE

AI FACTORY

AI TOOLING

One solution for your AI infrastructure journey

CPU

GPU

ADVANCED SERVICES

ACCELERATED COMPUTE

KUBERNETES

ADVANCED NETWORKING

DATA STORAGE

OPERATIONS

Cisco AI PODs

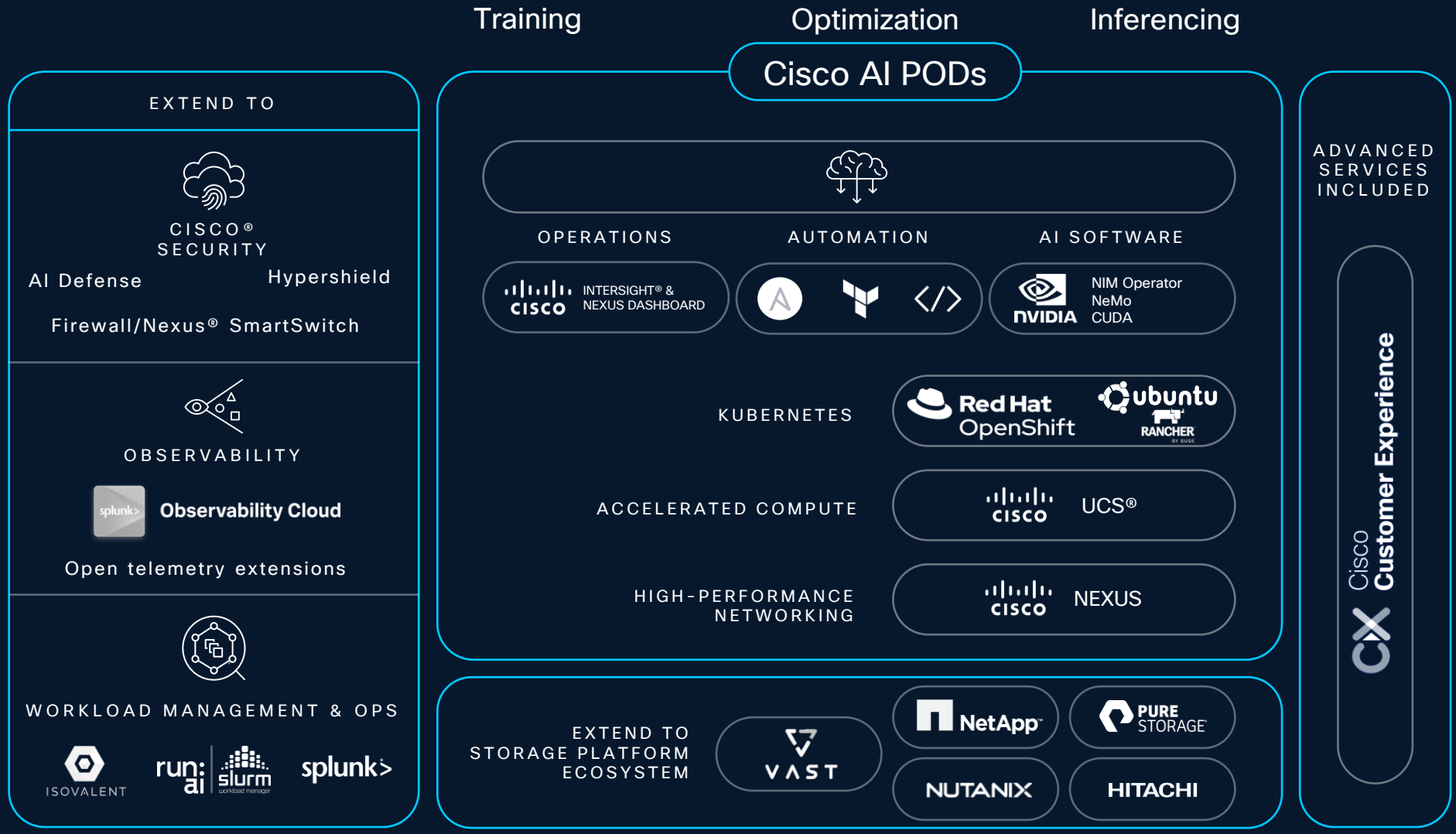
A scalable architecture, built to support any AI workload simply & efficiently

Deploy AI with confidence
Cisco CVD, NVIDIA ERA

Fully supported stack including Cisco and 3rd party components
Cisco CX Success Track

Orderable, use case driven AI-ready infrastructure stacks
Inferencing.
Optimization.
Training.

Incremental, atomic-level -or- fabric-based cluster scale

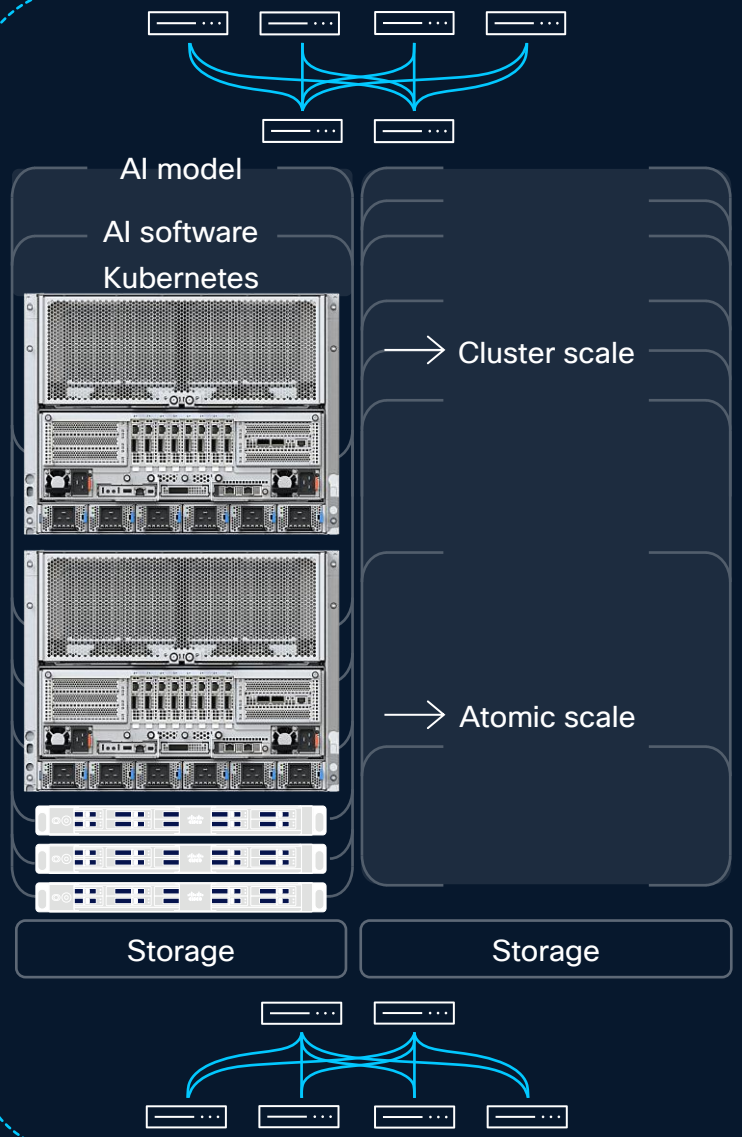


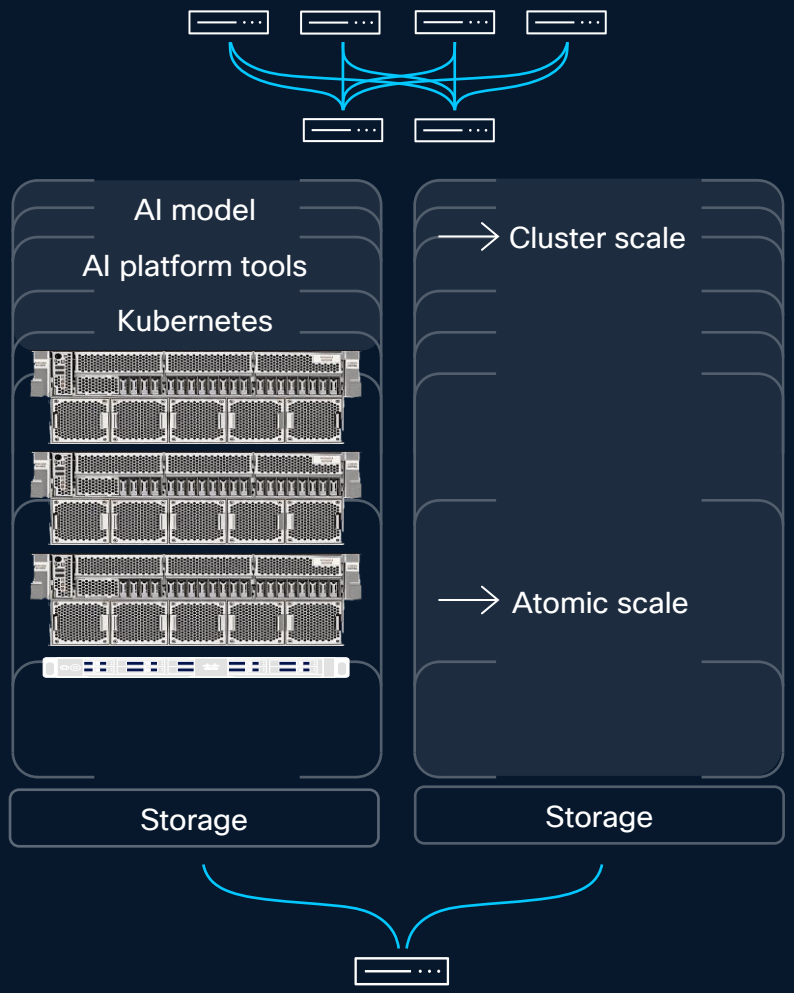
Cisco AI PODs

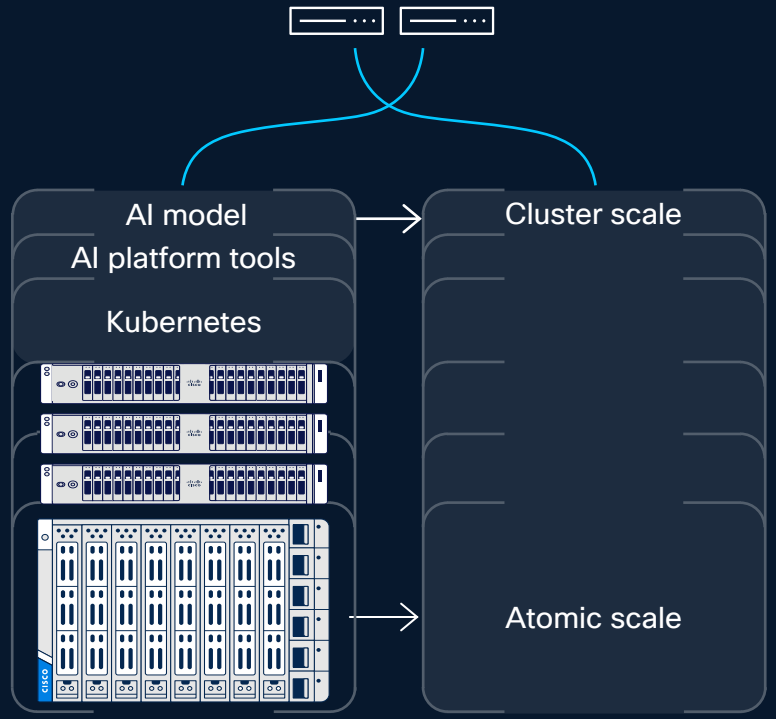
Cisco validated architectures

Scale Units *preview

Training, optimization & inferencing

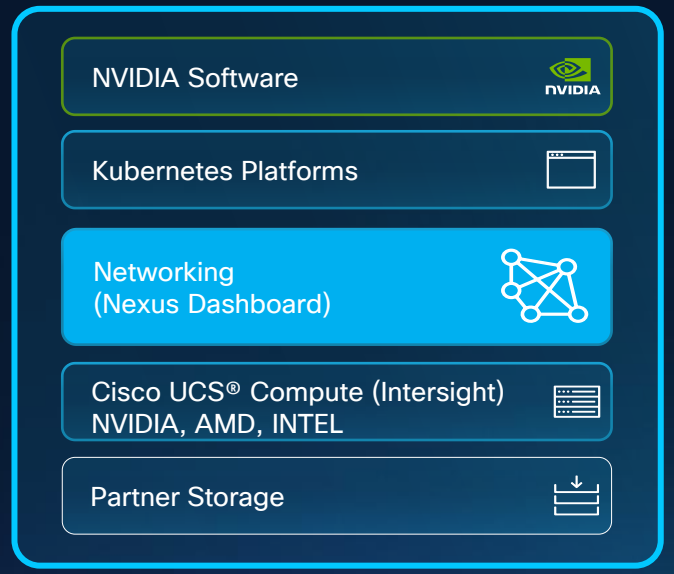






Cisco AI PODs: Flexible Operating Models

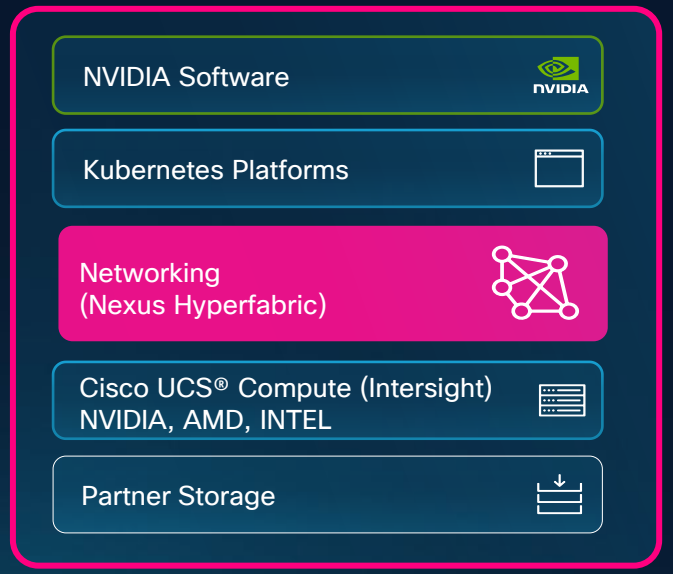
AI POD w/ On-prem management



Modular, pre-validated infrastructure:

- Full stack, buy & deploy
- Nexus Dashboard: On-prem networking management

AI POD w/ Cloud management



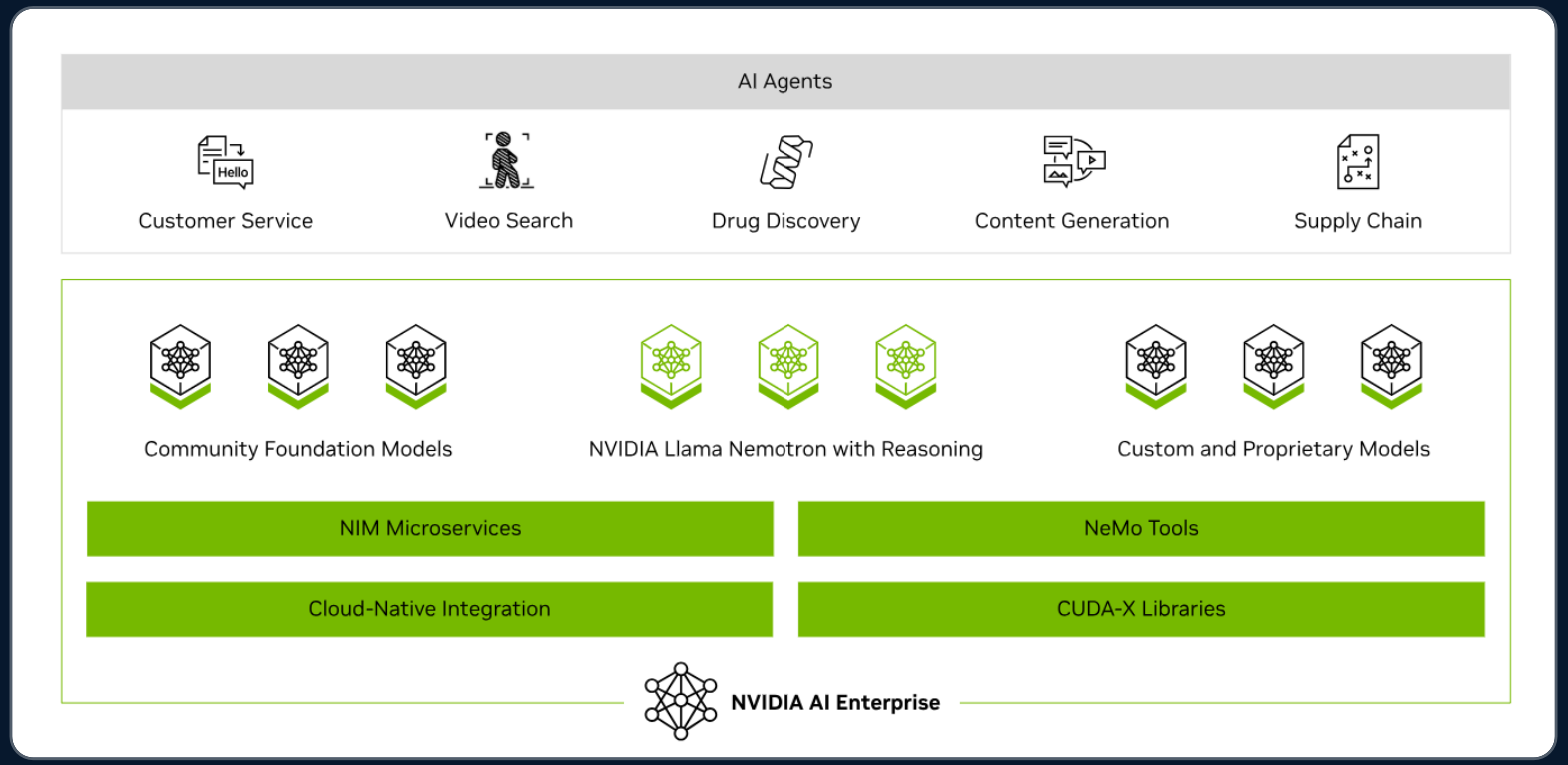
Turnkey infrastructure:

- Full stack, buy & deploy
- Nexus Hyperfabric: Cloud-managed Networking
- Nexus Hyperfabric AI: Cloud-managed physical infrastructure

NVIDIA AI Enterprise included on Cisco AI-PODs

Delivering building blocks for enterprise AI

Production-ready software for agentic AI



The NVIDIA AI Enterprise tools on Cisco AI-PODs provide support for each step in the training, optimization, and deployment of AI agents.

Deploy the latest state-of-the-art AI models

Explore the NVIDIA NIMs catalog of enterprise-ready, performance-optimized models for efficient inference and reasoning.

Build and manage data flywheels with NeMo

Discover powerful, ready-to-use model training, evaluation, and guard railing tools and RAG building blocks for optimizing agentic AI.

Customizable blueprints for your use case

Reference workflows for building fast, high-performance, and secure agentic systems using the latest machine learning best practices.

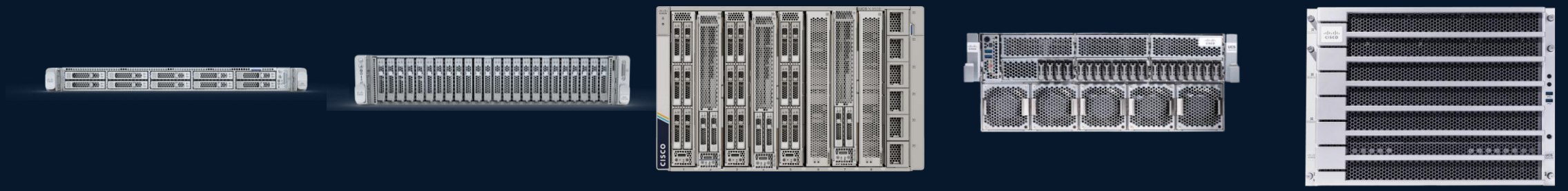
AI Platform Considerations: UCS GPU Options

Potential Workload Type

- Entry Tier & Edge
- Universal AI, Text to Image/Video
- AI Training / Inferencing (PCIe)
- AI Training, HPC, Data Analytics

GPUs are subject to change as new peripherals are added to the portfolio

UCS C220/C225	UCS C240/C245	UCS X-Series	UCS C845A	UCS C885A
NVIDIA L4*	AMD MI355*	AMD MI355*	NVIDIA H200 NVL*	NVIDIA H100 / H200 SXM
	AMD MI210*	AMD MI210*		AMD MI300X OAM
	NVIDIA H100 NVL*	NVIDIA H100 NVL*		
	NVIDIA RTX PRO 6000 *	NVIDIA RTX PRO 6000 *		
	NVIDIA L40S*	NVIDIA L40S*		



Max GPUs	3	2-8*	2-8*	2-8	8

** NOTE: GPU Form Factor and GPU model support may vary between AMD and Intel Platforms (i.e. c220/c225, c240/c245, and x210c/x215c). Check the spec sheet for each platform to determine maximum GPU support based on GPU selection*

Cisco AI Defense on AI POD - Summary

End-to-end security

Continuous visibility and governance

Sovereign ready

DATA CONTROL

SECURITY

USE CASES

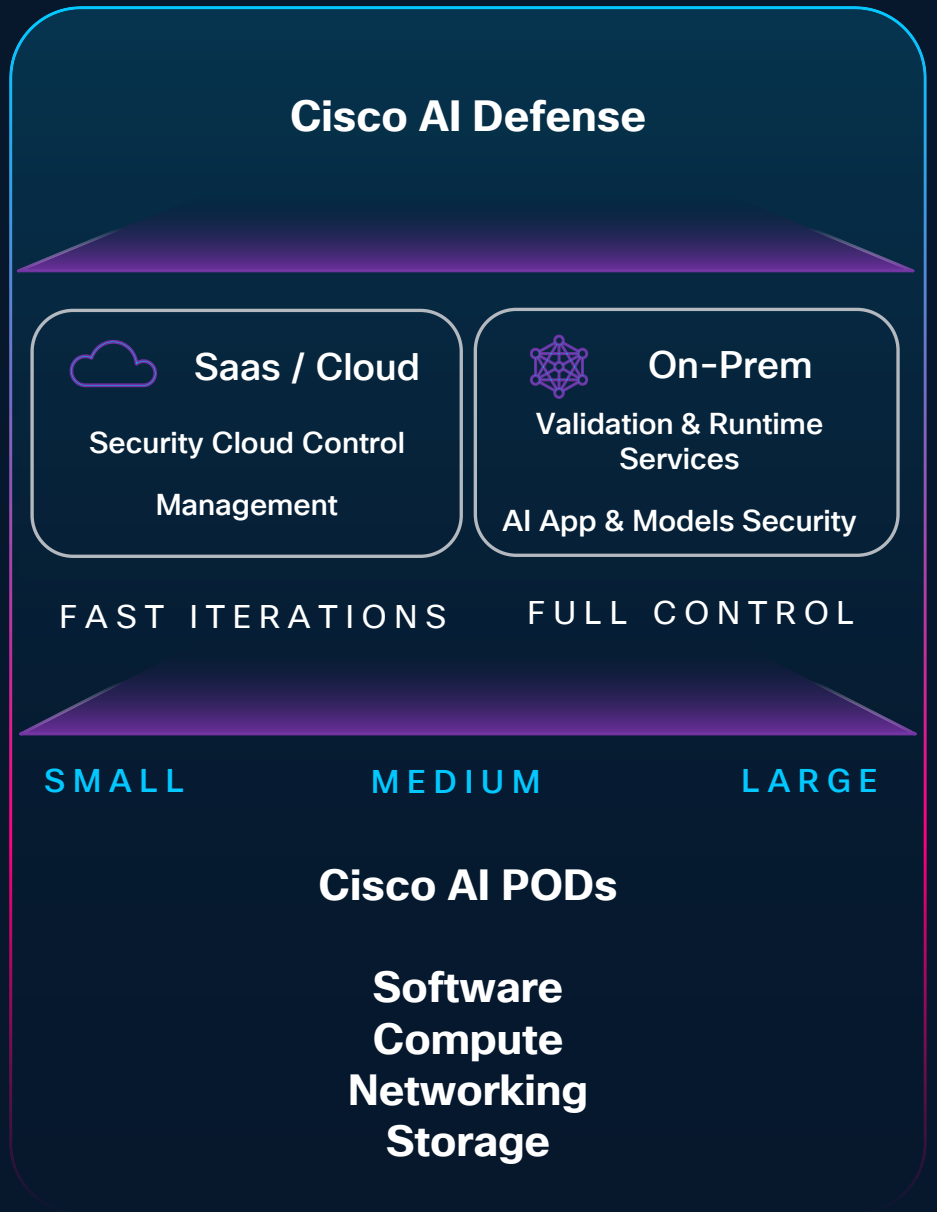
MODEL AND CUSTOMIZATION

Beta

Available after GSX

Controlled Availability

Expected in 90 days



Full Application Security

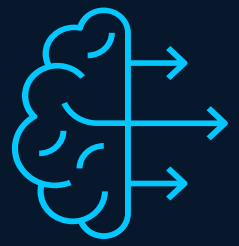
AI Application and Model Validation

AI Runtime Application Protection

Transforming AI Infrastructure with RAFAY GPUaaS

Delivering Sovereign & Enterprise GPU Clouds for Accelerated AI Innovation

Maximize GPU Utilization & ROI



Accelerate AI Innovation with Self-Service

Establish Secure, Governed Multi-Tenant AI Clouds

DEVELOPERS & DATA SCIENTISTS

RAFAY

AUTONOMY CONTROL EFFICIENCY

ARCHITECTURE & PLATFORM TEAMS

Cisco AI PODs

AIPOD-POD1

For Running AI Models, not Building Them

Required

AI POD 1 is like buying a car that's ready to drive. The model's already trained – you just need to use it. Whether you're classifying documents, detecting images, or answering questions with a chatbot, this POD gives you everything you need to run AI at the edge or in a small data center. It's simple to deploy, doesn't need a lot of space, and works well for companies who want fast, reliable AI results without heavy compute or complex wiring. Think of it as "plug-and-go" AI.

Compute

AMD intel

- C-885A
- C-845A
- C-240A
- C-245A
- X-Series (M8 and Beyond)

Front-end fabric

- Nexus 9364D-GX2A
- Nexus 9332D-GX2B
- Nexus 9324C-SE1U
- Nexus 9364D-GX2A (Spine)

Management software

- Cisco Intersight®
- Nexus Dashboard (Physical, virtual)

Management cluster

- X-Series
- Nexus 93108TC-FX3
- C-Series

OS/Kubernetes

Red Hat OpenShift

RANCHER BY SUSE

Storage

MLB Included:

VAST NUTANIX

Other Options:

NetApp HITACHI

Adoption services

- Cisco CX
- MINT Service

cisco

Cisco

AIPOD-POD2

For companies that want to customize or build AI

AI POD 2 is more like a garage full of high-end parts and tools—built for people who want to *train, fine-tune,* or *customize* their AI models. It supports large-scale operations where GPUs need to talk to each other at high speeds, like training a model on your proprietary data or refining a foundation model to your industry. This POD is ideal if you need serious computing power, are managing big datasets, and want full control over how your AI behaves. It's not just using AI—it's building the AI engine itself.

Required

SU1 **Node<=4**

SU2 **Node<=8**

SU3 **Node<=16**

Compute

AMD intel

- C-885A
- C-845A

Back-end fabric

- Nexus 9332D-GX2B
- Nexus 9364D-GX2A
- Nexus 9364E-SG2
- Nexus 9364D-GX2A (Spine)

Front-end fabric

- Nexus 9364D-GX2A
- Nexus 9332D-GX2B
- Nexus 9324C-SE1U
- Nexus 9364D-GX2A (Spine)

Management software

- Cisco Intersight
- Nexus Dashboard (Physical, virtual)

Management cluster

- X-Series
- Nexus 93108TC-FX3
- C-Series

OS/Kubernetes

Red Hat OpenShift

RANCHER BY SUSE

Storage

MLB Included:

VAST NUTANIX

Other Options:

NetApp HITACHI

Adoption services

- Cisco CX
- MINT Service

Cisco AI PODs

Included in Cisco Secure AI Factory with NVIDIA

Why Cisco AI PODs?

- Security-first architecture enables safe enterprise AI
- Unmatched performance AI infrastructure enables efficient model training, customization, and inferencing
- Pre-validated AI infrastructure stack for simplified deployment drastically reduces set-up time

Intersight Demo

Thank you to our sponsors!

Q&A


CISCO