CISCO Connect

# Cisco Connect Compute

Joacim Pettersson
Solution Engineer CAI

Dec 2, 2025

# Computing for the next decade

Technology partnerships

Cisco CX services

Converged infrastructures

Validated designs

Cisco Plus

TCO and sizing tools

Cisco Intersight

Virtual infrastructures

Hybrid cloud

Enterprise applications

AI/ML

Containers

SAP HANA

Desktop virtualization

Data protection

**Cisco Solutions portfolio**

# UCS – Software Define Compute

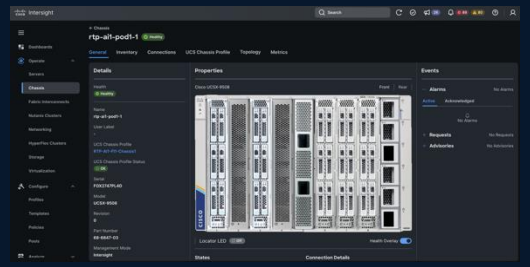
Model-Driven Framework to Abstract Resources

**Intersight**

Creates Object Model

Defines Model and Platform

Fabric Interconnects



## 1 Subject Matter Experts Define Policies

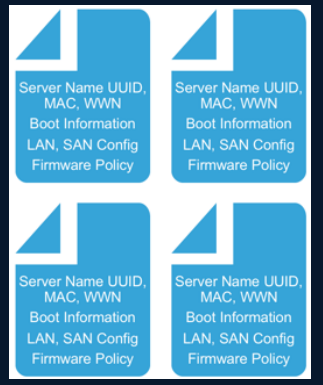
Server SME | Network SME | Storage SME

- Server Policy
- Storage Policy
- Network Policy
- Virtualization Policy
- Application Profiles

## 2 Policies Used to Create Service Profile Templates

Server Name UUID, MAC, WWN
Boot Information
LAN, SAN Config
Firmware Policy

## 3 Service Profile Templates Create Service Profiles




## 4 System Configures Hardware Elements Automatically and Eliminates Configuration Drift



cisco

# Cisco Intersight

Intuitive experience

Enhanced support

Proactive guidance

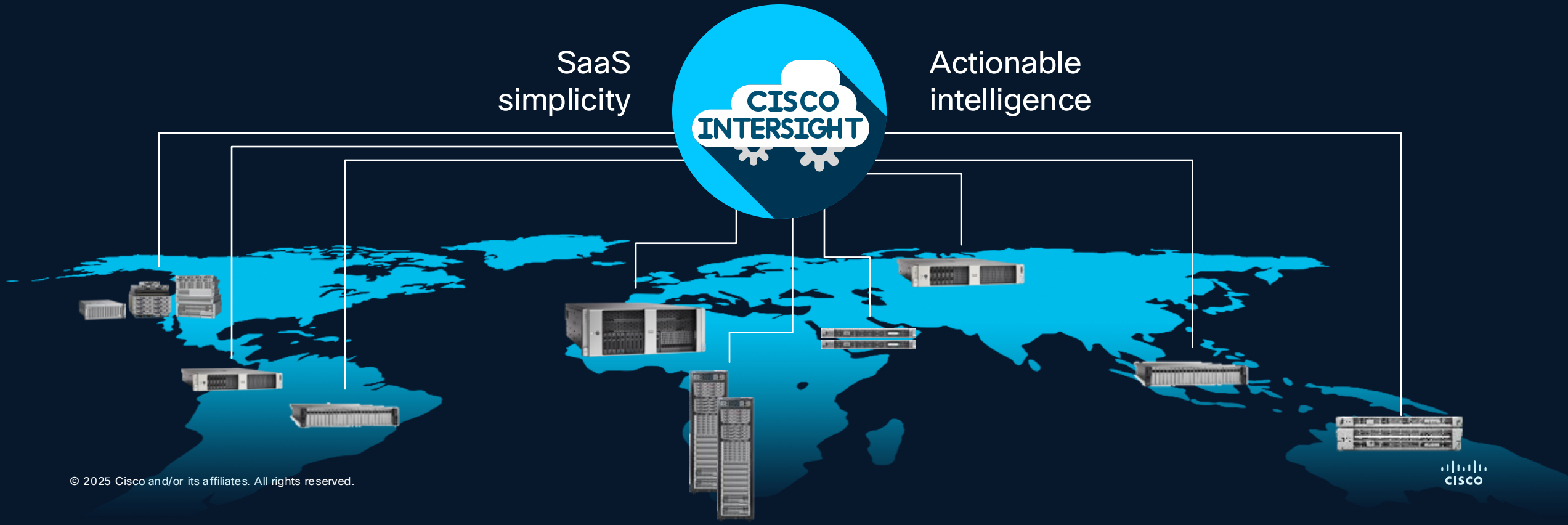
Secure and extensible

SaaS or connected appliance

SaaS simplicity

CISCO INTERSIGHT

Actionable intelligence





cisco

# UCS – Unifies Compute System: Fabric Centric Design

Simplicity, Resiliency and TCO Benefits

## High Performance

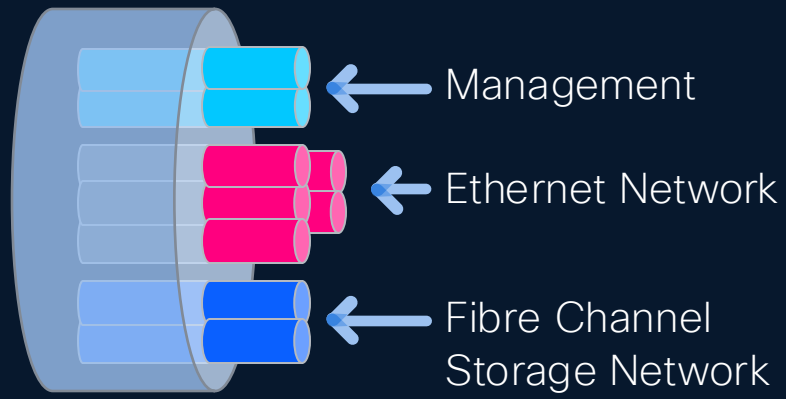
10/25/40/100 Gbps Ethernet
8/16/32/64G Fibre Channel
400/1600 Gbps per chassis

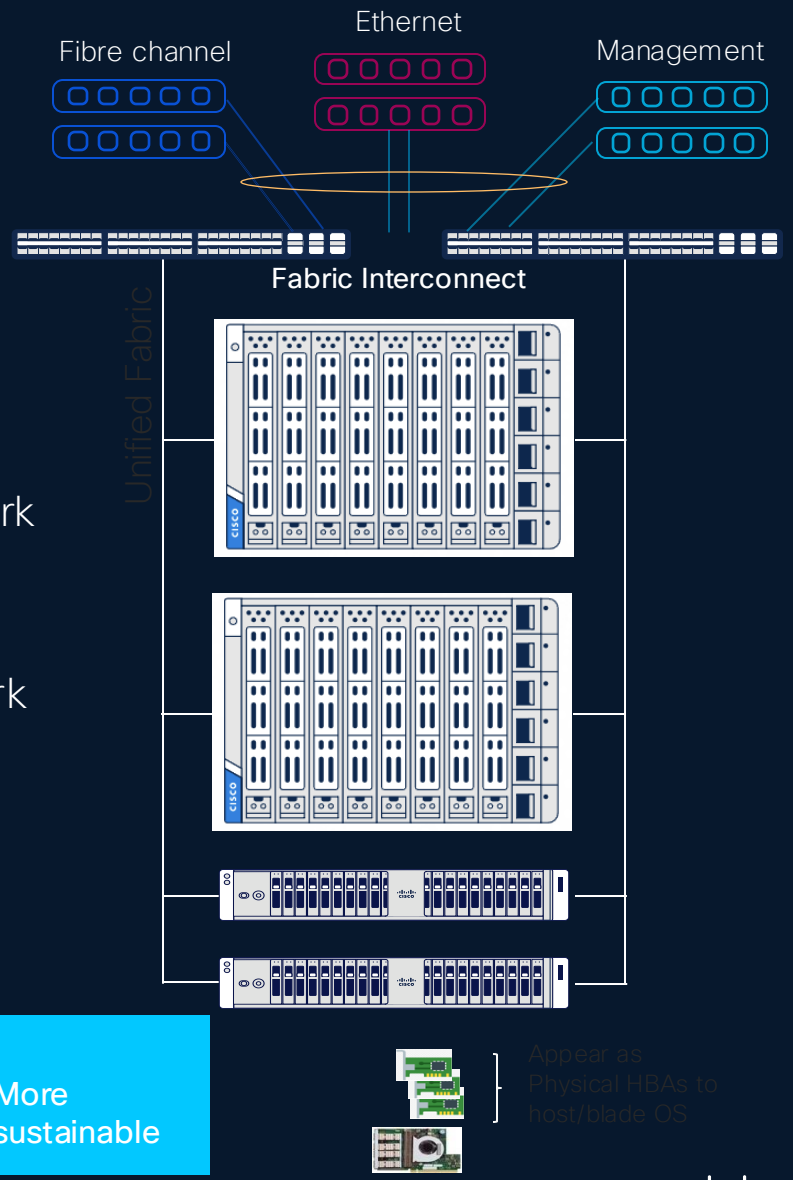
## Unified Fabric

Single Cable for Traffic
Multi-protocol IP/FC SAN
Dynamic I/O Virtualization

## Simplified Infra Management

Add cables for Bandwidth vs. Connectivity
Stateless infra Mgmt @ scale
Plug-n-play fabric




UP TO **66%** Less Cables Adapters Switches

UP TO **50%** OpEx, CapEx Savings

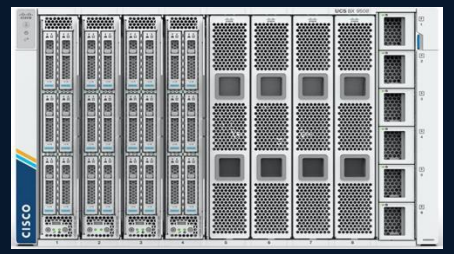
UP TO **30%** More sustainable

# Cisco UCS Compute Portfolio

## MAINSTREAM ENTERPRISE SERVERS

UCS X-Series
X9508 Chassis

IFM Module



UCS X-Series Direct

New

UCS X210c M7

UCS X210c M8

New

UCS X410c M7

UCS B200 M6

UCS X215c M8

New

UCS C240 M8E3S
36 EDSFF E3.S1T

New

UCS C240 M8SX
28 HDD/SDD/NVMe

New

UCS C240 M8L
16 LFF + 4 SFF

New



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

New

UCS C220 M8S
10 HDD/SSD/NVMe

New

UCS C220 M7N
10 NVMe

UCS C245 M8SX
28 HDD/SDD

New

UCS C225 M8S
10 HDD/SSD

New

UCS C225 M8N
10 NVMe

New

## AI SERVERS

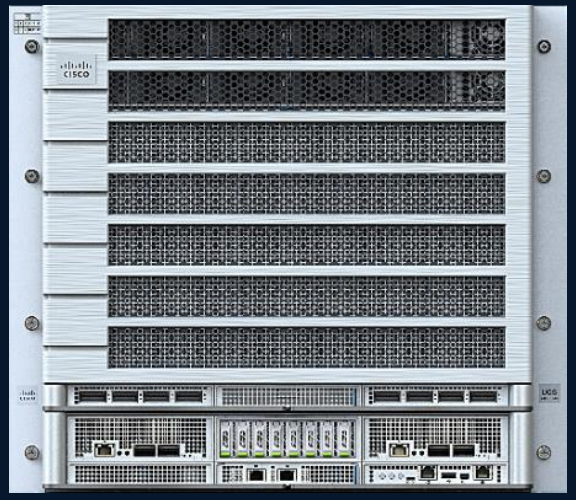
UCS C885A M8
8RU Dense GPU
Server

New

UCS C845A M8
4RU MGX Server

New

# UCS C880A Dense GPU Server Specifications



## Product Specifications

| | |
|---|---|
| Form Factor | • HGX 10RU 19” Rack Server |
| Compute + Memory | • 2x 6th Gen Intel Xeon CPUs (Select SKUs for AI and HPC workloads)<br>• Up to 32x DDR5 RDIMMS |
| Storage | • 2x M.2 SATA Boot Drives with HW RAID Controller (Boot)<br>• Up to 8x PCIe Gen5 x4 E1.S NVMe SSDs (Data) |
| GPUs | • 8x NVIDIA HGX B300 NVL8 air-cooled GPUs |
| Network Cards | • E-W: Integrated ConnectX-8<br>• N-S: 4x PCIe Gen5 x16 FHHL slots, 1x OCP TSFF Gen5 x8 |
| Cooling | • 20 Hot swappable FANs |
| Physical I/O | • 1 USB 3 type A, 1 mDP, 1 ID Button, 1 System Power Button, 1 Reset Button, 1 USB type C (for debugging), 1 RJ45 (OOB mgmt.), 1 RJ45 (LOM port) |
| Power Supply | • 12x 50V 3.2kW (N+N redundancy) |

# UCS C845A M8 Front and Back Views

**CPU**

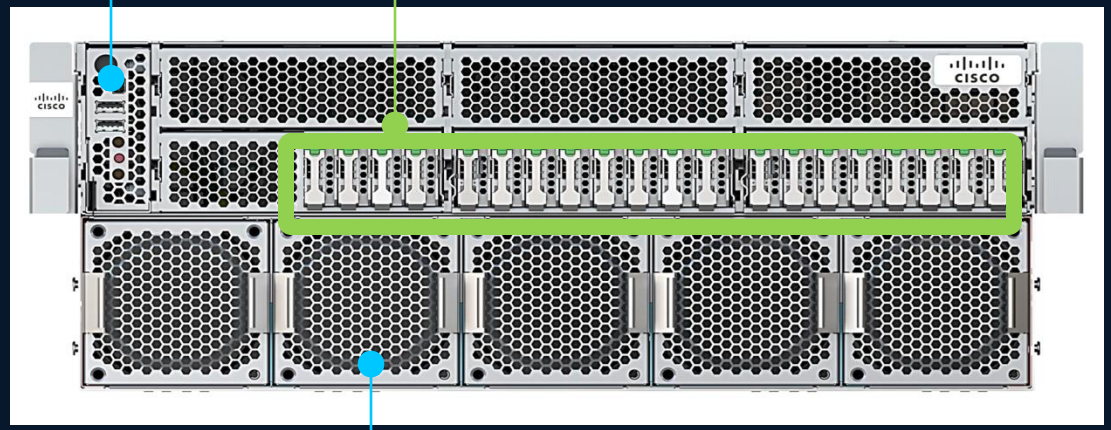
**20x**

Control Panel

Hot-Pluggable E1.S Drives Up to 152TB

**2x**

AMD 5th Gen CPUs
Up to 96 cores & up to 5GHz, 400W/CPU



**10x**

Hot-Pluggable Fans
Front to rear cooling

**Memory**

**Up to 32x**

64GB, 96GB or 128GB DDR5 RDIMMs
Up to 4TB

**GPU**

**2/4/6/8x**

Nvidia H100 NVL, H200 NVL, L40S, RTX Pro 6000 GPUs 350W/400W/600W GPUs

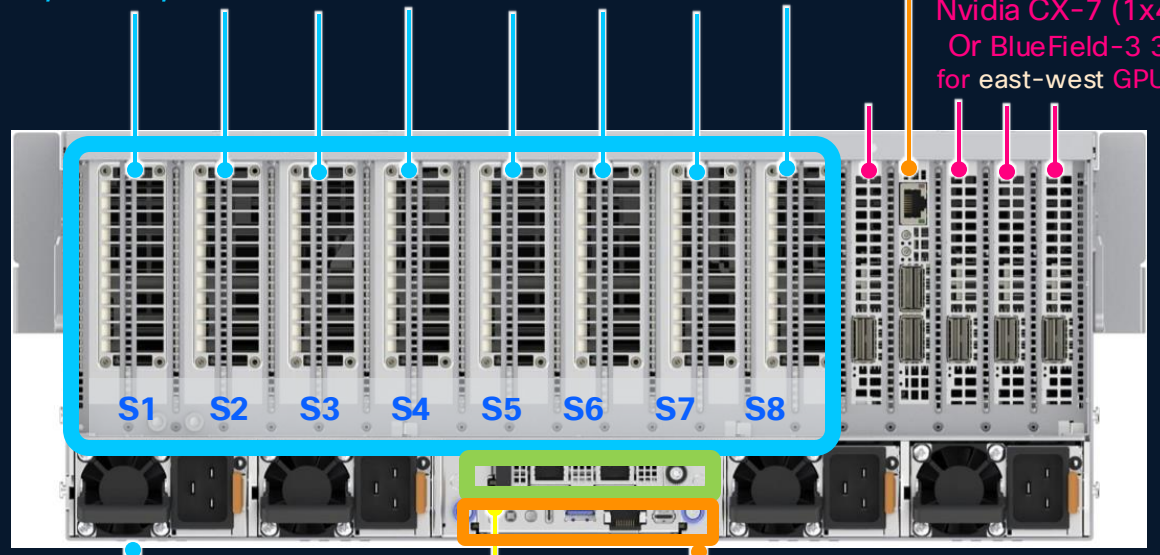
AMD MI210 GPUs 300W/GPU

**1x**

Nvidia CX-7 (2x200G) or BlueField-3 3220 (2x200G) for north-south front-end traffic

**4x**

Nvidia CX-7 (1x400G) Or BlueField-3 3140H for east-west GPU traffic




**4x**

3.2 kW Power Supplies N+1

**1x OCP 3.0**

Intel X710 for north-south host management traffic

**DC-SCM**

BMC for out-of-band server management

# UCS X-Series Direct: Update

FCS RELEASE (Q3CY24)

## Up to 8 compute nodes

- X210c M6, X210c M7, X215c M8
- X410c M7
- Support of X-Fabric + X440p
- 4th & 5th Gen VIC
- Full config support of X-series compute Node

Now Available

## Up to 16 X-Series compute nodes

- IFM-100G for 2nd Chassis
- Plus: Up to 4 rack server

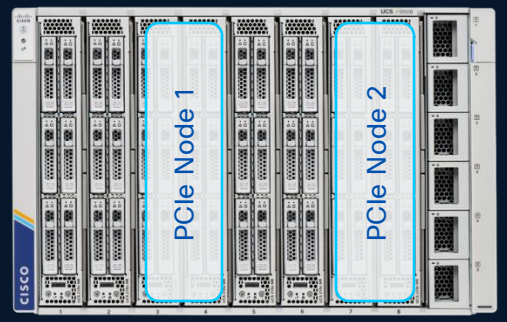


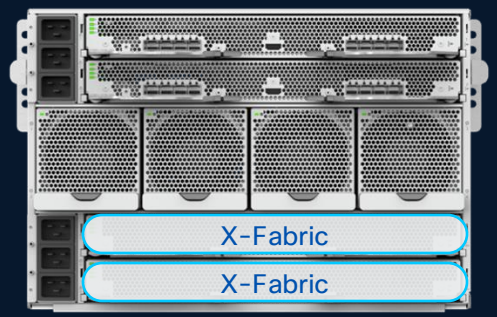
cisco

# 2nd Gen X580p PCIe Node and X9516 X-Fabric

## Cloud-Operated, Composable Infrastructure for AI and Traditional Workloads

- Solution for customer who needs higher GPU density
- Supports wide range of workloads
- Intersight managed solution
- Competitive differentiation with X-Fabric and X-Series

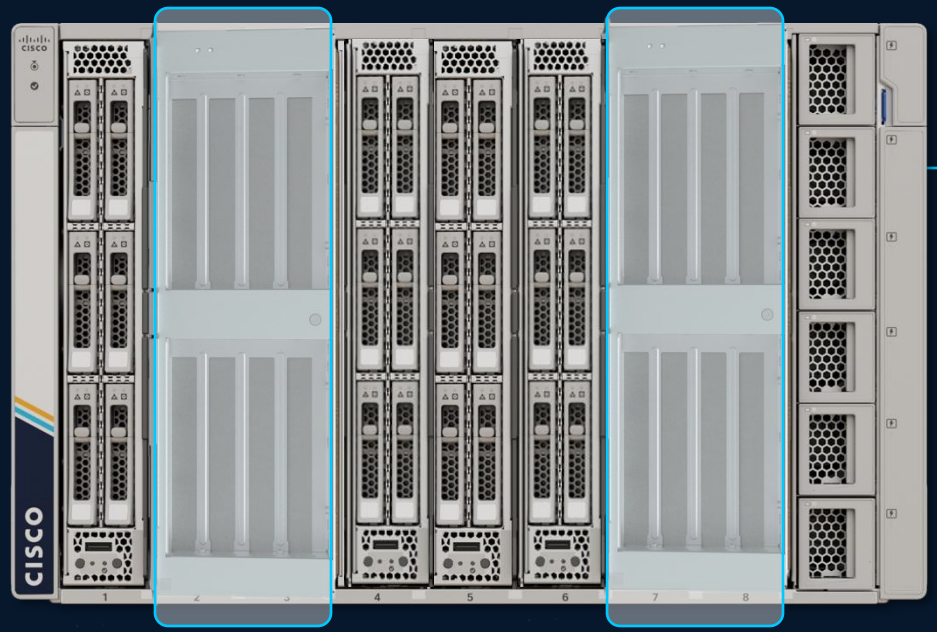



### UCS X-Fabric Technology with PCIe Node

- PCIe Switching with PCIe Gen 5 connectivity
- 4x FHFL or HHHL GPUs per PCIe node
- Intra-host GPU interconnect with NVLink
- Intersight policy-based Management
- Inter-host scaling with RDMA over AI Fabric

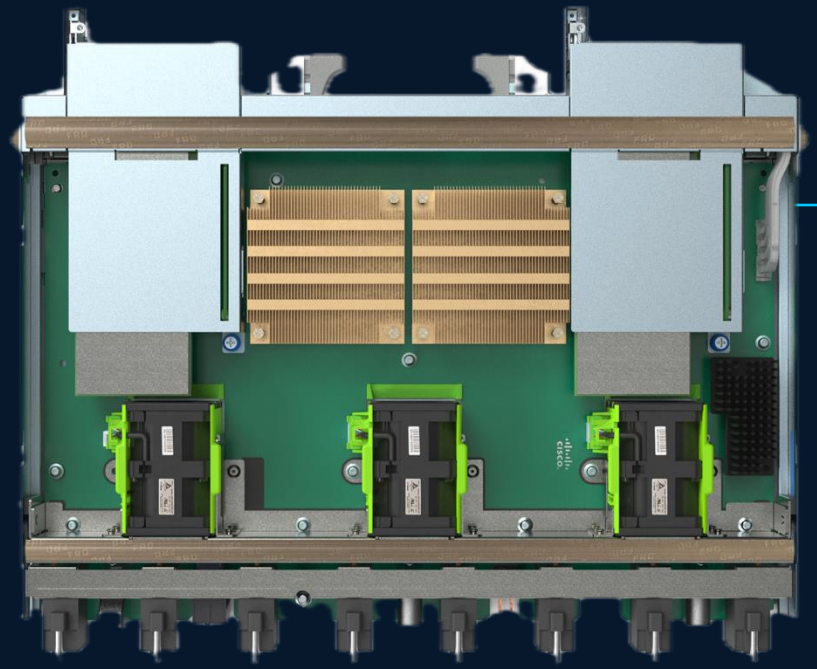
# UCS X580p PCIe Node




- Double wide PCIe node for 4x FHFL GPU and PCIe G5 GPU support
  - Nvidia H200-NVL, RTX PRO 6000 & L40S
- Support multiple vendors: Nvidia, AMD*/Intel*
- NVLink bridge support
- Support up to 600W FHFL GPU
- Managed PCIe node with BMC support
- Policy based GPU management
- Ability to share GPUs across two Compute nodes

* AMD & Intel GPUs support will be post FCS

# UCS X9516 X-Fabric



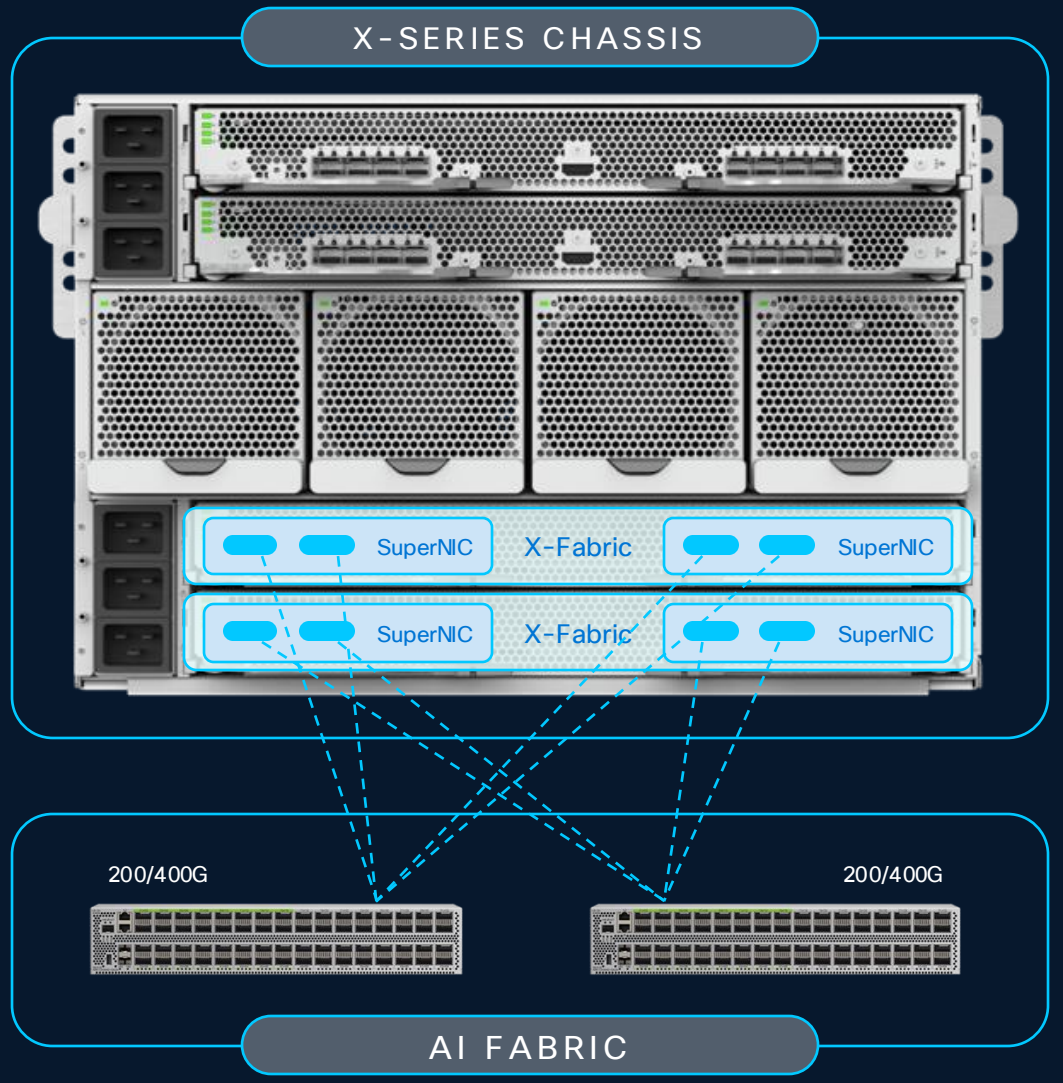
- PCIe Gen5 Switching
- 2x CEM Slots to support HHHL NIC cards
  - ConnectX 7 (2x 200GB & 1x 400G)
- Managed XFM Modules with BMC support
- GPU Direct Support over RDMA
- GPU Backend (East-West Traffic) network support

# AI Cluster Expansion

## GPU-to-GPU connectivity

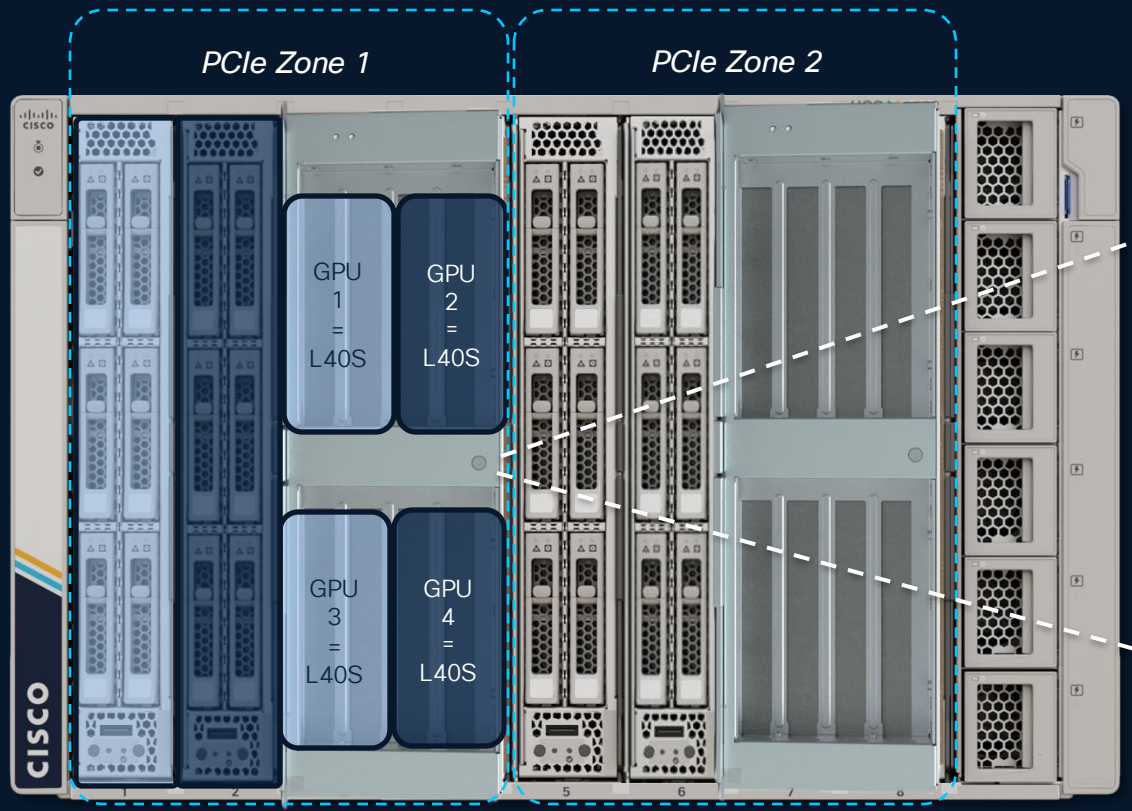
with XFM external ports

- X-Fabric Module with Gen5 PCIe switch
- SuperNIC Adapter for GPU East-to-West traffic
- 1 or 2 external ethernet ports based on adapter



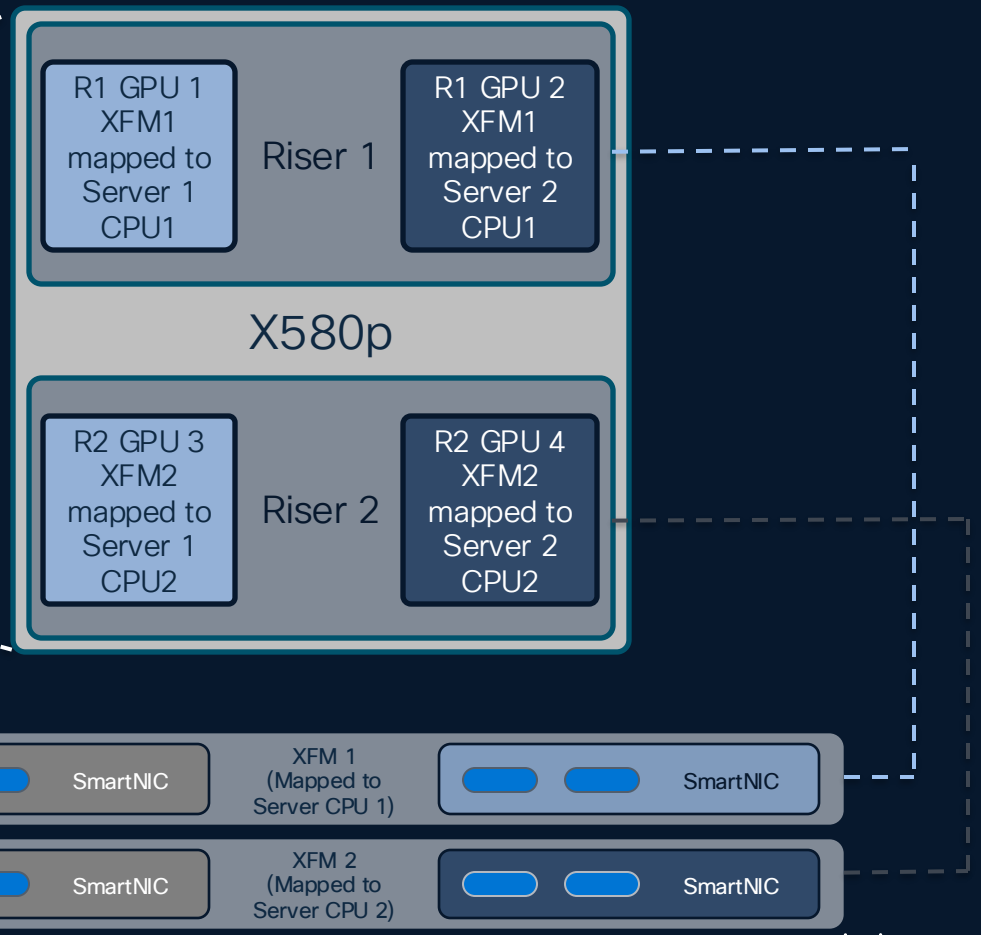
# X580p – GPUs 1/3 Mapped to Server 1 and GPUs 2/4 Mapped to Server 2

*(1x NIC mapped to each server)*




Server 1: 2x GPUs 1x NIC

Server 2: 2x GPUs 1x NIC



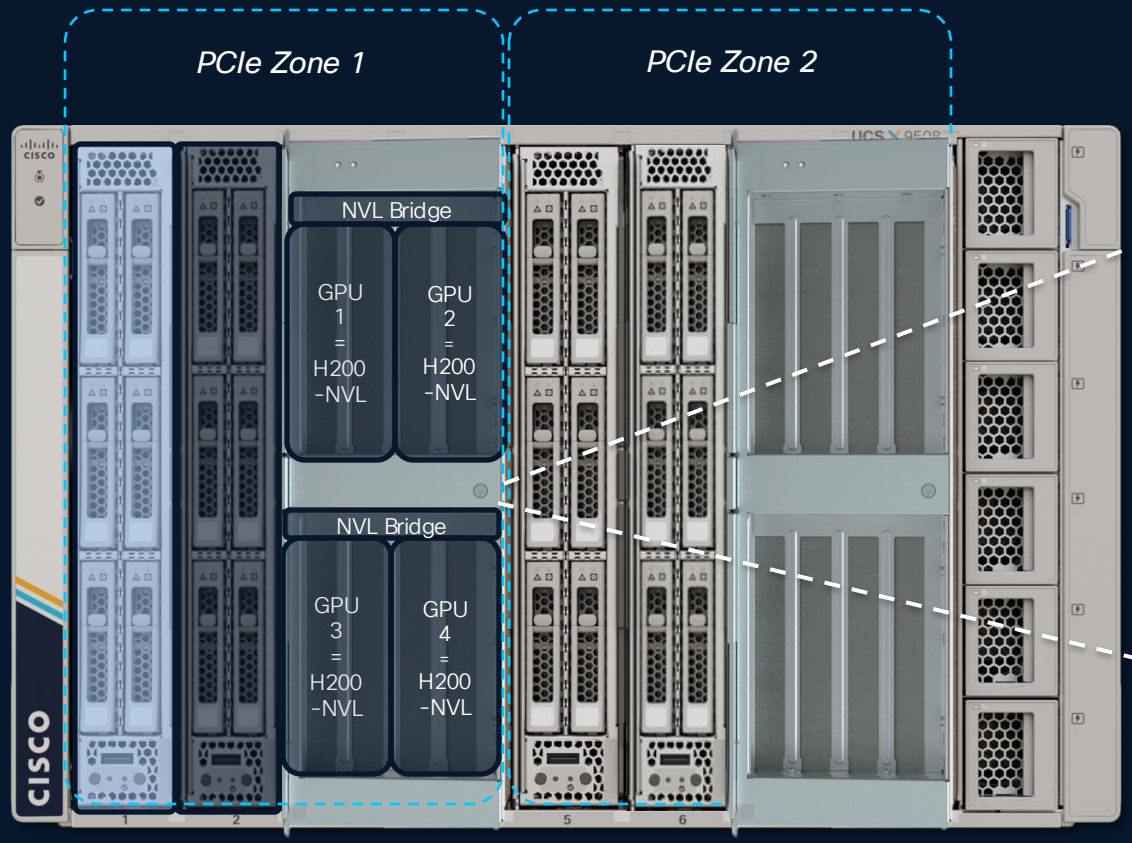

*Note: X580p default in 3/4 or 7/8 and is not supported in slots 2/3 and 6/7. It can be installed in slots 1/2 or 5/6 as well.*



cisco

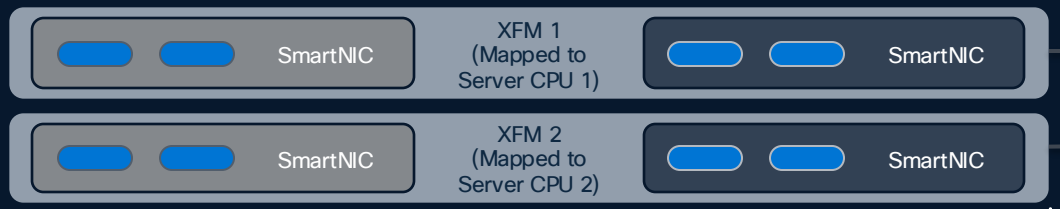
# X580p – All GPUs Allocated to Server 2 w/NVL Bridge

*(2x NICs mapped to one server)*




Server 1: No GPU

Server 2: 4x GPUs 2x NICs


*Note: X580p default in 3/4 or 7/8 and is not supported in slots 2/3 and 6/7. It can be installed in slots 1/2 or 5/6 as well.*



cisco

# 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

# Cisco AI PODs

## Introducing AI POD “Integrated Offerings”

Training | Optimization | Inferencing

### Cisco AI PODs

**EXTEND TO**

CISCO® SECURITY

AI Defense | Hypershield

Firewall/Nexus® SmartSwitch

OBSERVABILITY

Observability Cloud

Open telemetry extensions

WORKLOAD MANAGEMENT & OPS

ISOVALENT | run:ai | slurm workload manager | splunk>

OPERATIONS | AUTOMATION | AI SOFTWARE

INTERSIGHT® & NEXUS DASHBOARD

NIM Operator
NeMo
CUDA

KUBERNETES: Red Hat OpenShift | ubuntu RANCHER

ACCELERATED COMPUTE: UCS®

HIGH-PERFORMANCE NETWORKING: NEXUS®

EXTEND TO STORAGE PLATFORM ECOSYSTEM: VAST | NetApp | PURE STORAGE | NUTANIX | HITACHI

ADVANCED SERVICES INCLUDED

Cisco Customer Experience

**BYO AI tools:**

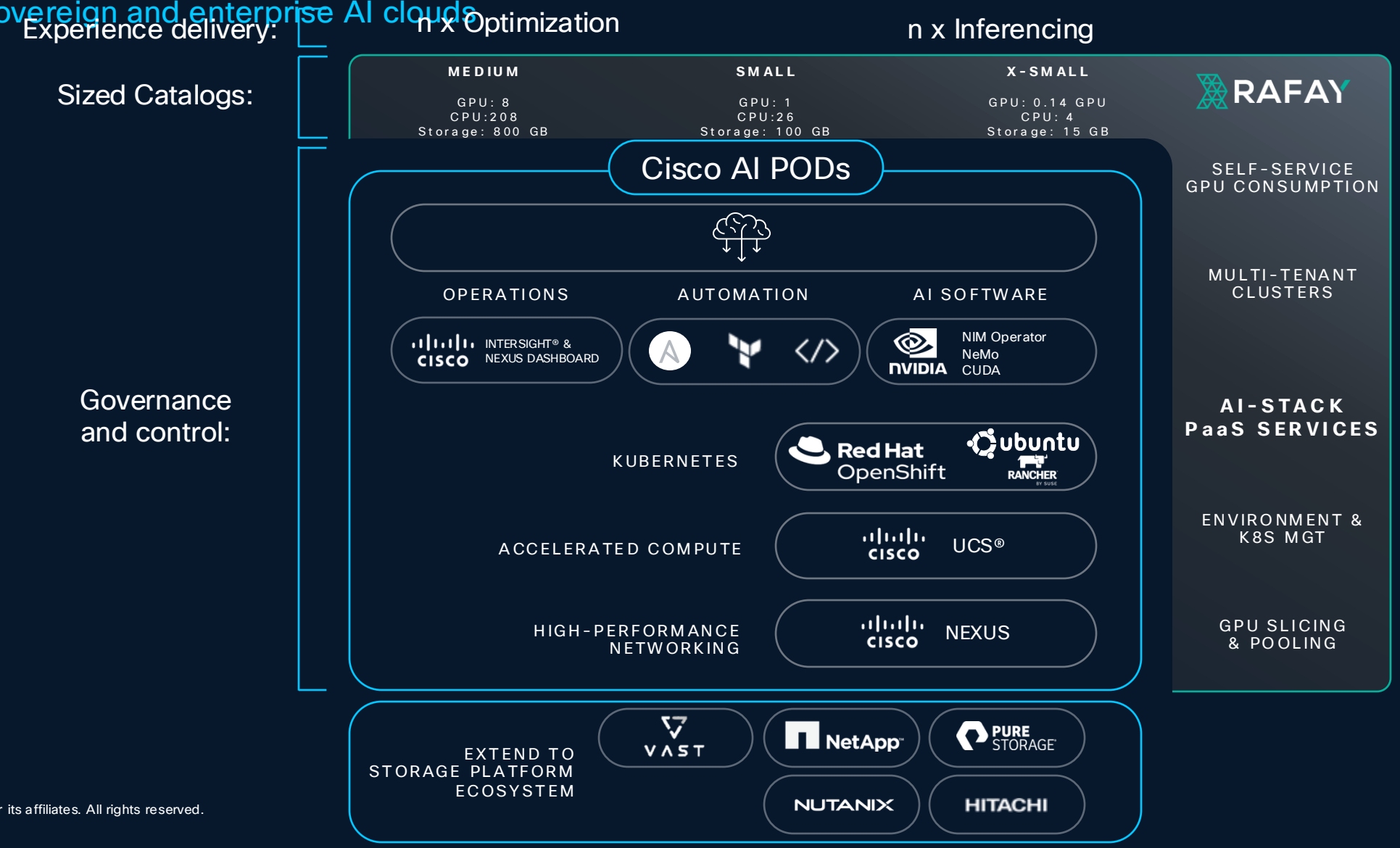
RAFAY | Kubeflow | jupyter | Apache Airflow | Weights & Biases | mlflow | neptune.ai | kedro | comet | ZenML | CLEARML | PREFECT | Flyte | mongoDB



cisco

# Transform Cisco AI PODs into a GPU cloud

## Deliver sovereign and enterprise AI clouds



# 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

### Storage

MLB Included:

VAST NUTANIX

Other Options:

NetApp HITACHI

### Adoption services

- Cisco CX
- MINT Service



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

### Storage

MLB Included:

VAST NUTANIX

Other Options:

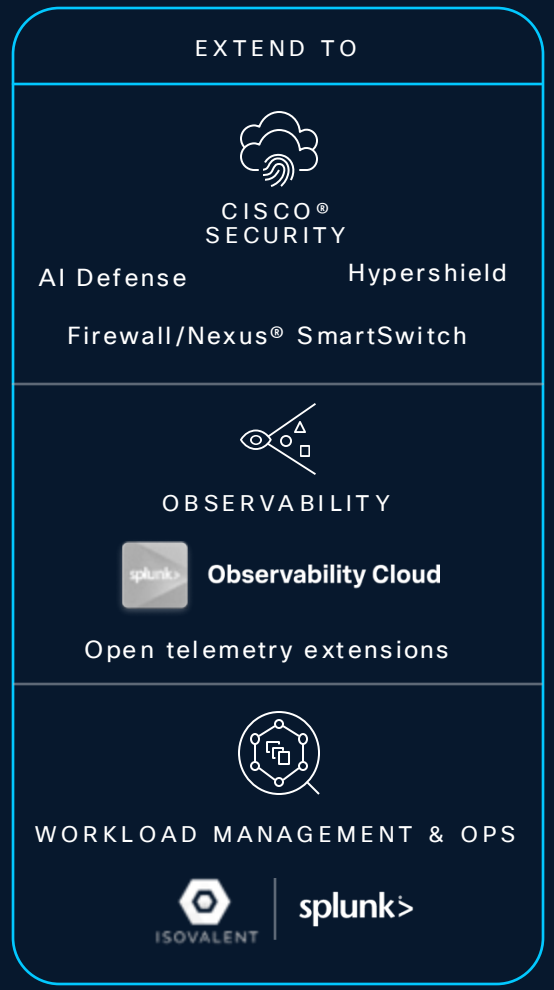
NetApp HITACHI

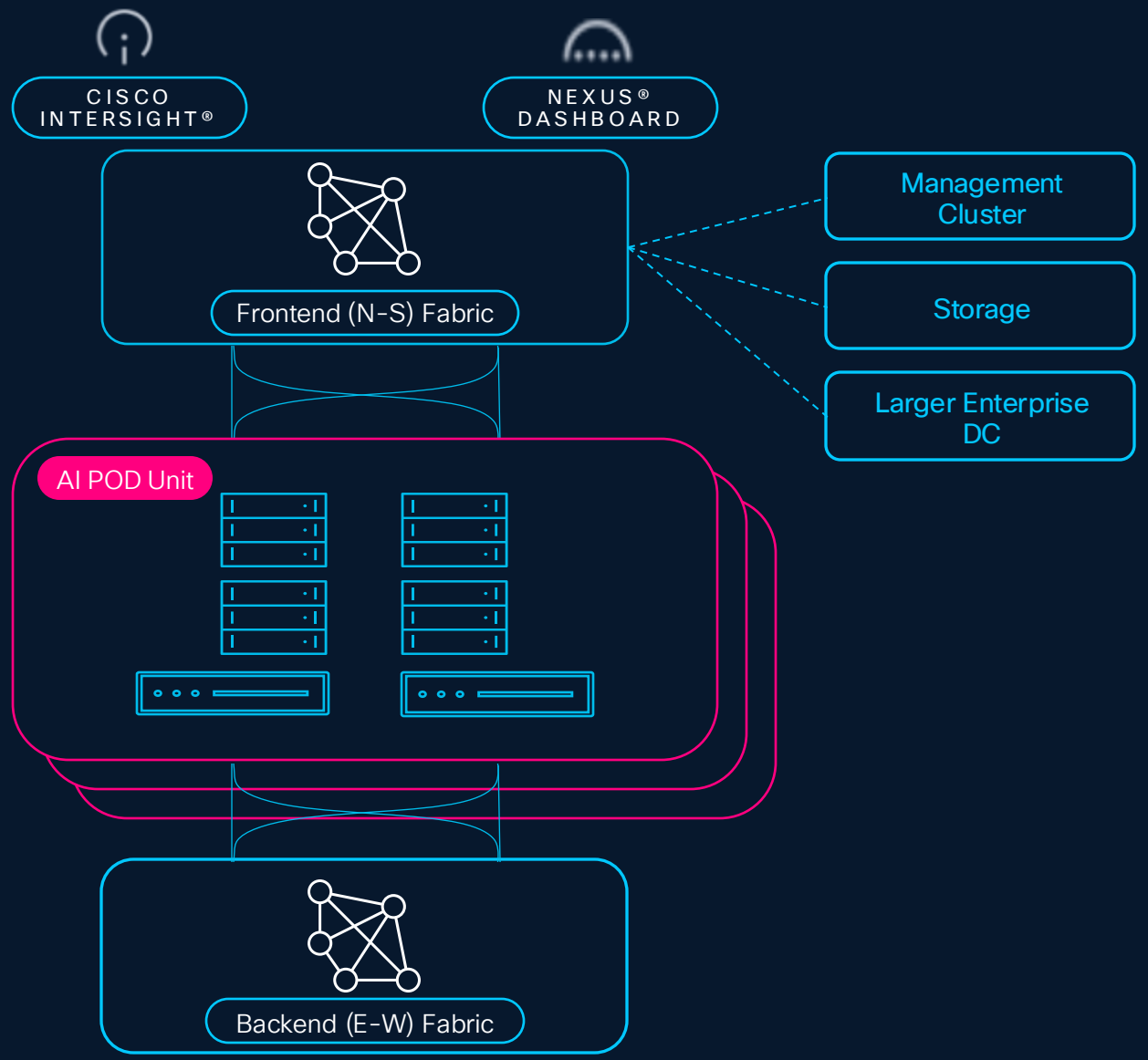
### Adoption services

- Cisco CX
- MINT Service

# Cisco AI POD

## An AI-ready Infrastructure



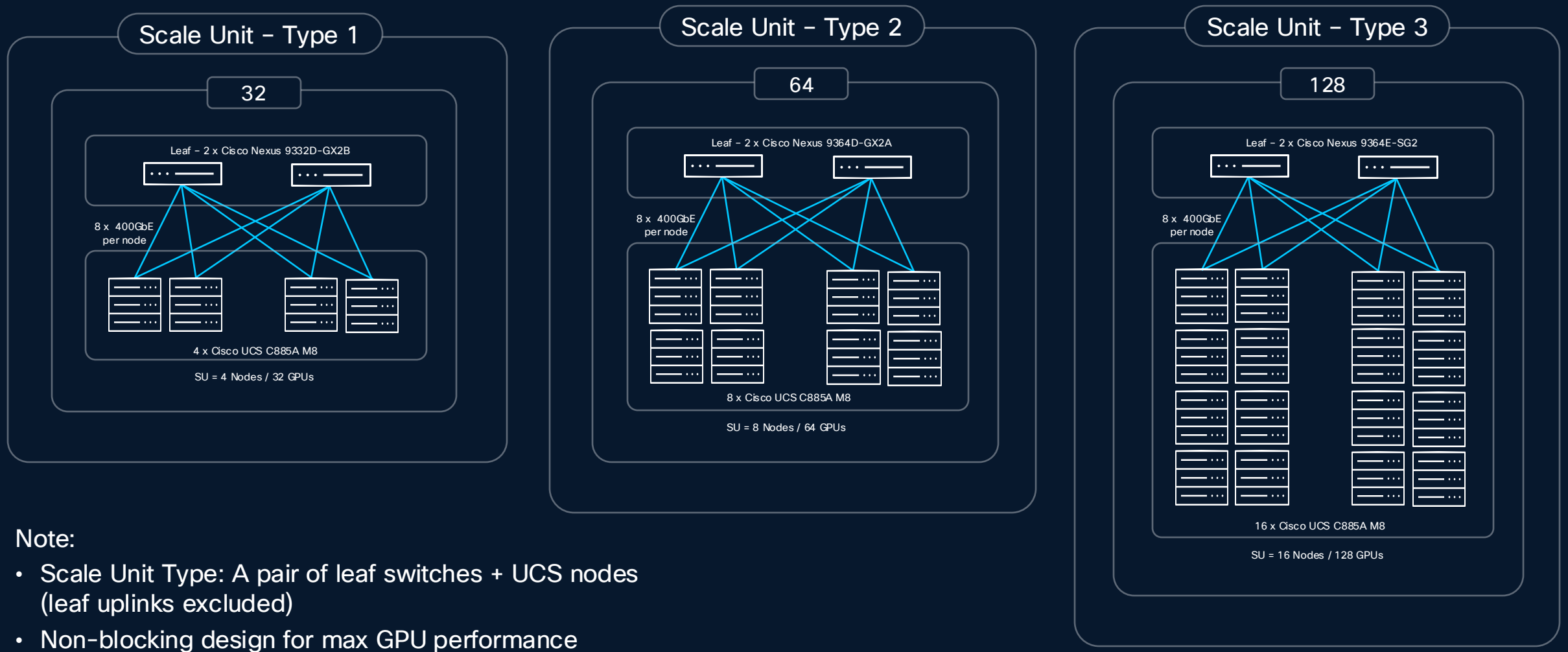






cisco

# Scale Unit definition

## Example UCS C885A–GPU Compute Fabric (E-W)

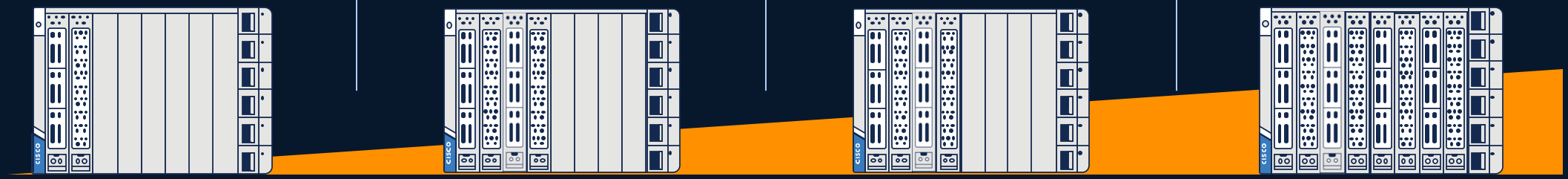



Note:

- Scale Unit Type: A pair of leaf switches + UCS nodes (leaf uplinks excluded)
- Non-blocking design for max GPU performance

# Cisco UCSX AI PODs

| Typical use case | Edge Inferencing (7B-13B Parameter) | RAG Augmented Inferencing (13B-40B+ Parameter) | Large-Scale RAG Augmented Inferencing | Scale-Out Inferencing Cluster (Inferencing Multiple Models) |
|---|---|---|---|---|
| Hardware specification | **Small**<br>1x X210C compute node<br>• 2x Intel 5th Gen 6548Y+<br>• 512 GB System Memory<br>• 5x 1.6 TB NVMe drives<br>• 1x X440p PCIe<br>• 1x NVIDIA L40S | **Medium**<br>2x X210C compute nodes<br>• 4x Intel 5th Gen 6548Y+<br>• 1 TB System Memory<br>• 10x 1.6 TB NVMe drives<br>• 2x X440p PCIe<br>• 4x NVIDIA L40S | **Large**<br>2x X210C compute nodes<br>• 4x Intel 5th Gen 6548Y+<br>• 1 TB System Memory<br>• 10x 1.6 TB NVMe drives<br>• 2x X440p PCIe<br>• 4x NVIDIA H100 NVL | **Scale-Out**<br>4x X210C compute nodes<br>• 8x Intel 5th Gen 6548Y+<br>• 1.5 TB System Memory<br>• 12x 1.9 TB NVMe drives<br>• 4x X440p PCIe<br>• 8x NVIDIA L40S |



Performance and Scale

Inferencing Suite


CISCO