CISCO Engage | Tech Day

Simplifying Compute Infrastructure for the AI Ready Data Center

Terence Gibson
C&AI Solutions Engineer – US Commercial

John Rice
C&AI Solutions Engineer – US Public Sector

Cisco Intersight®

Accelerate provisioning and deployment at scale

CISCO INTERSIGHT®

Strengthen compliance

Increase uptime

Bolster security

Improve productivity

Accelerate operations

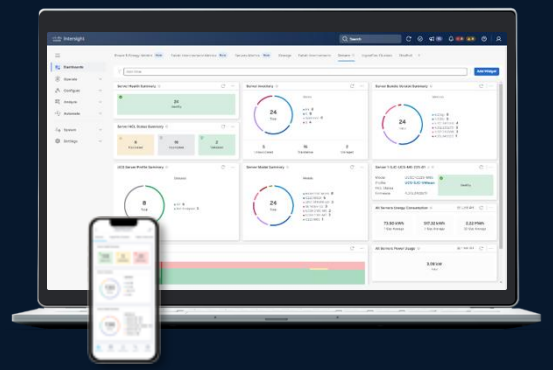
Control energy use

Control

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

See

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



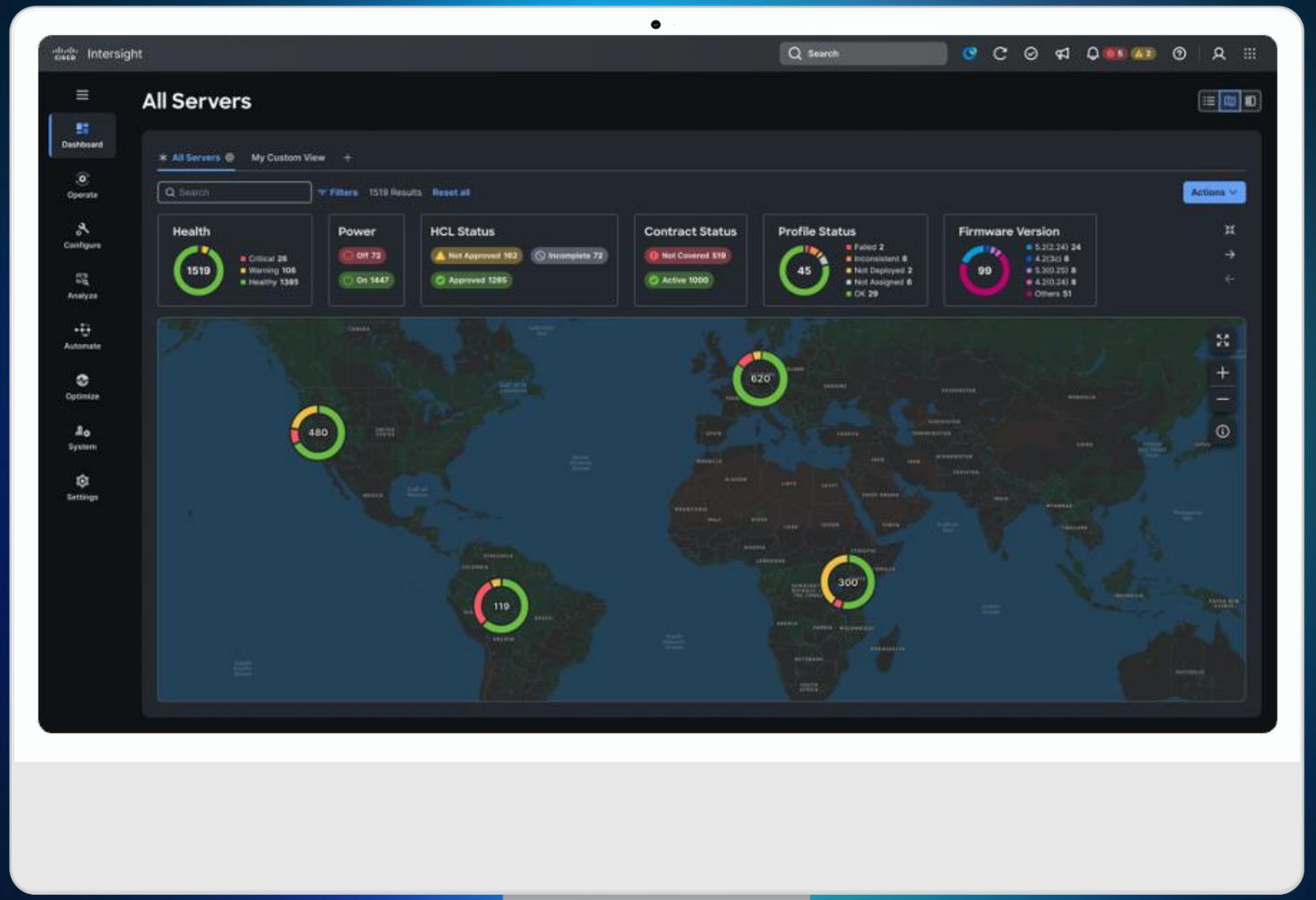
Automate

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

cisco

Intersight fleet management

- Simplified onboarding & Zero-touch provisioning
- Automated lifecycle management
- Global visualization



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

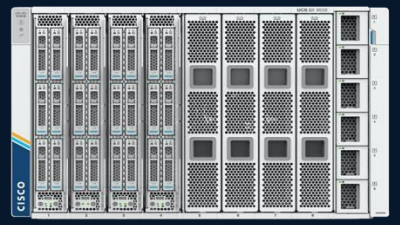
cisco

Cisco UCS Compute Portfolio

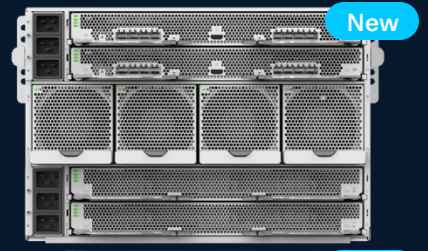
MAINSTREAM ENTERPRISE SERVERS

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

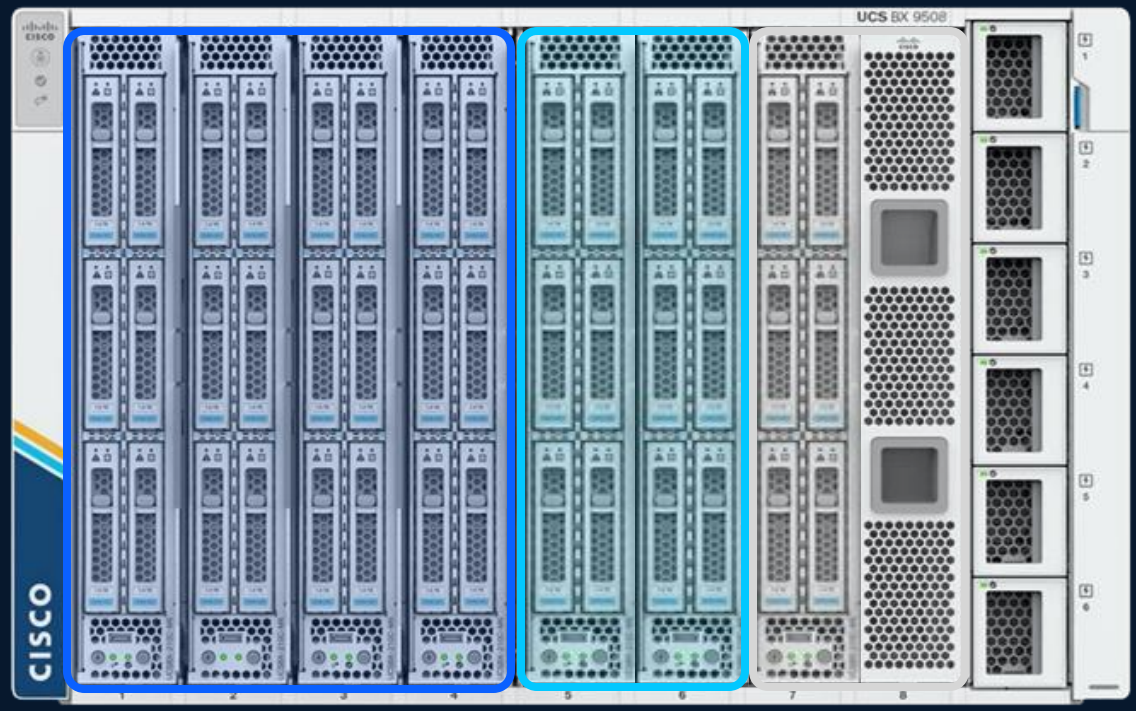
UCS® X-Series with X-Fabric

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

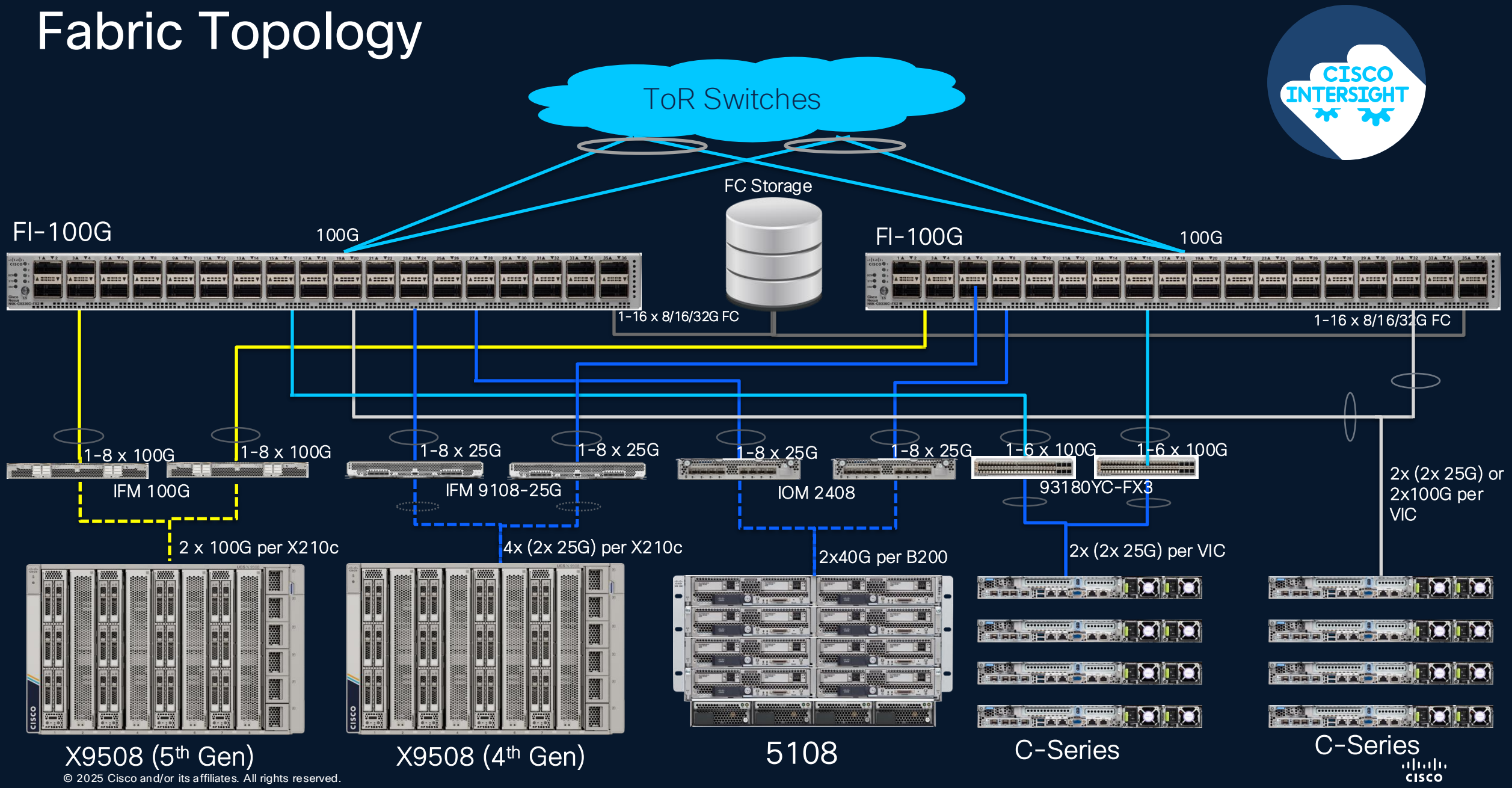
intel NVIDIA

200G
bandwidth to compute node

736 TB
of storage

cisco

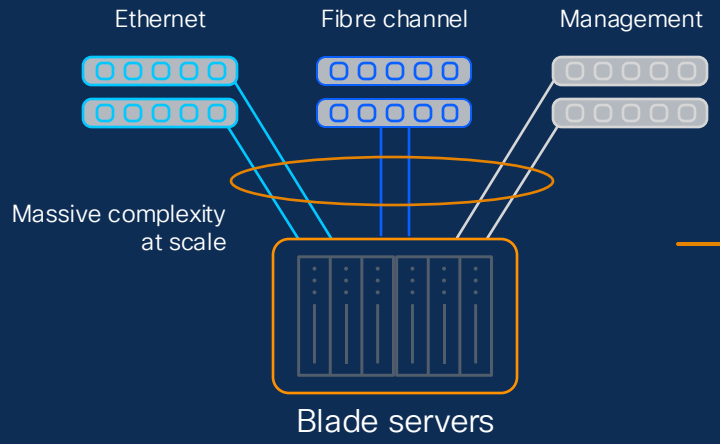
Fabric Topology



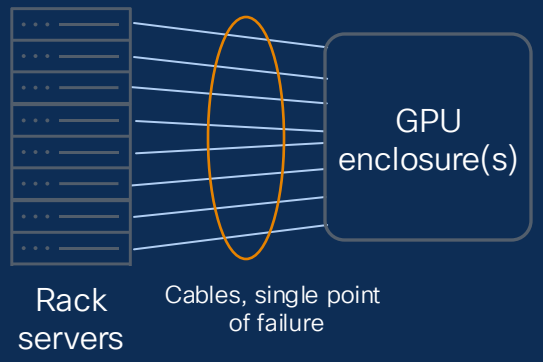
Industry-leading simplicity

Conventional approaches

1 | Silos of multiple Ethernet and SAN fabrics and adapters



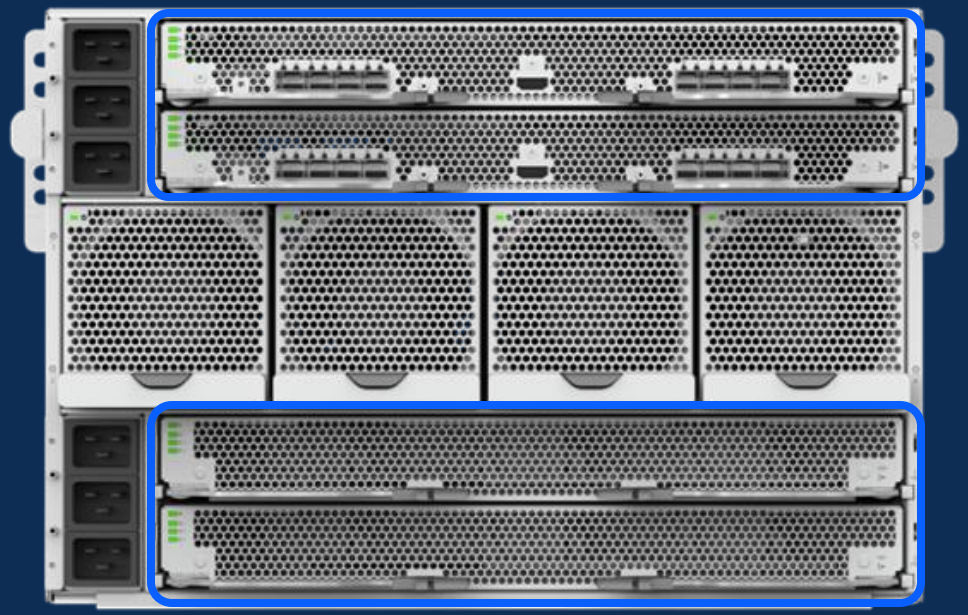
2 | Complex PCIe connectivity to external accelerators



Cisco solution

UCS® X-Series

Cisco Intersight®



Unified fabric

X-Fabric + PCIe nodes

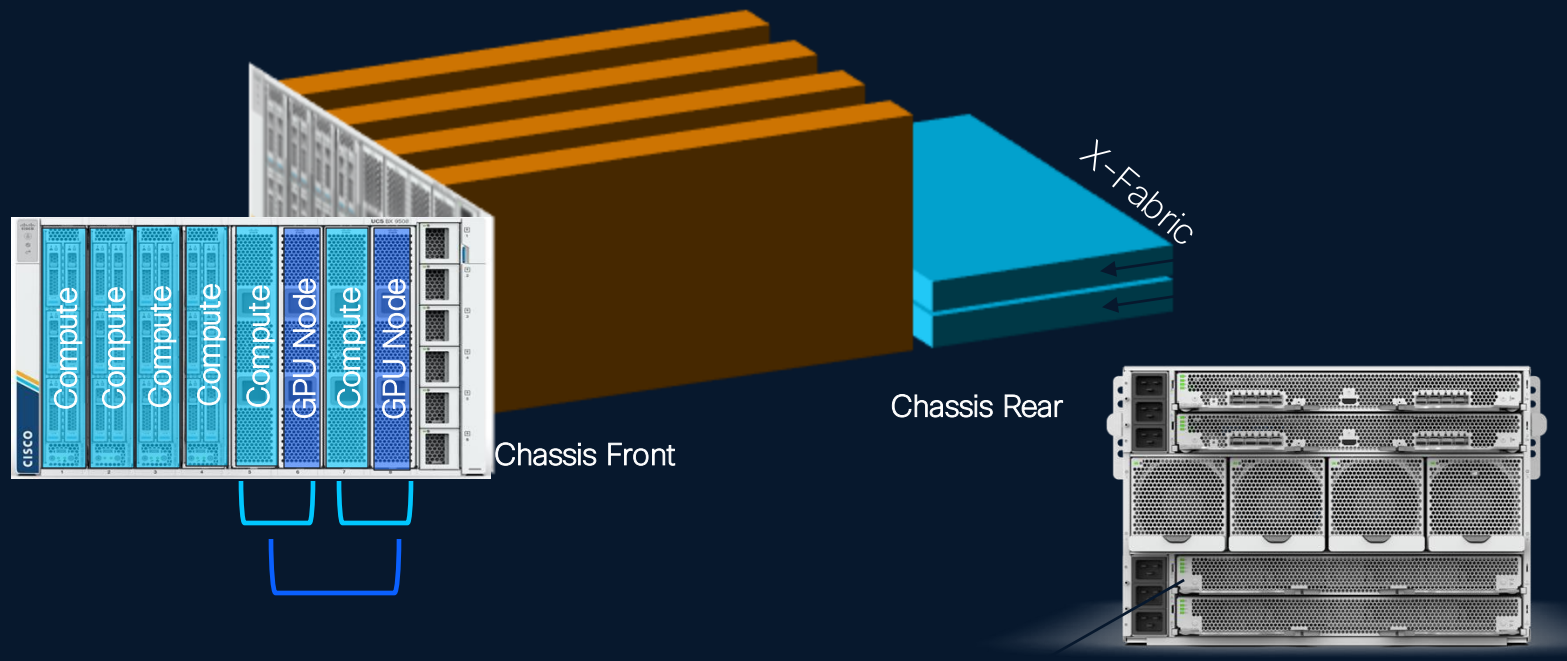
UCS X-Fabric Technology

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

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

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

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

Cisco AI PODs

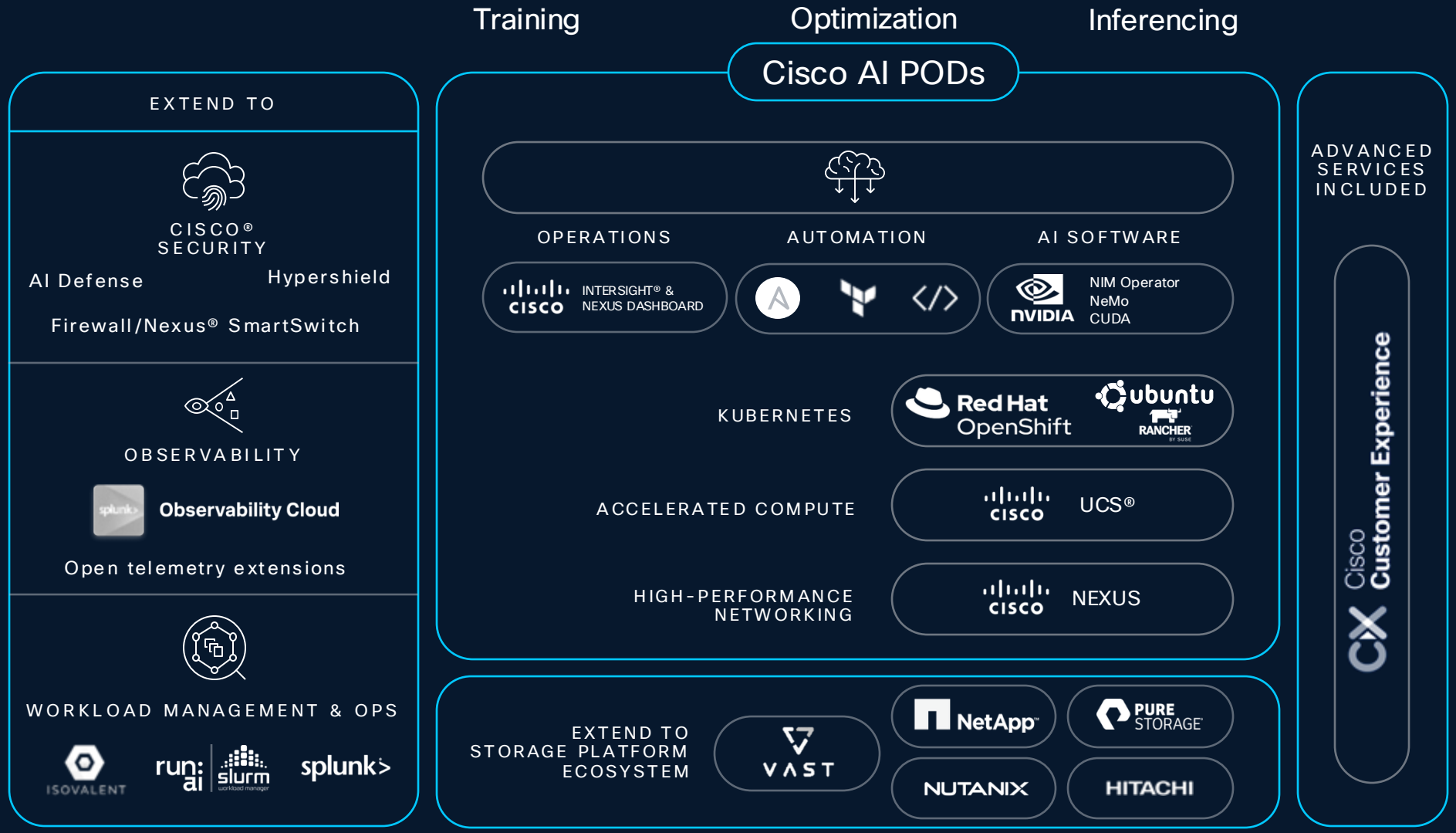
A scalable architecture, built to support any AI workload simply & efficiently

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



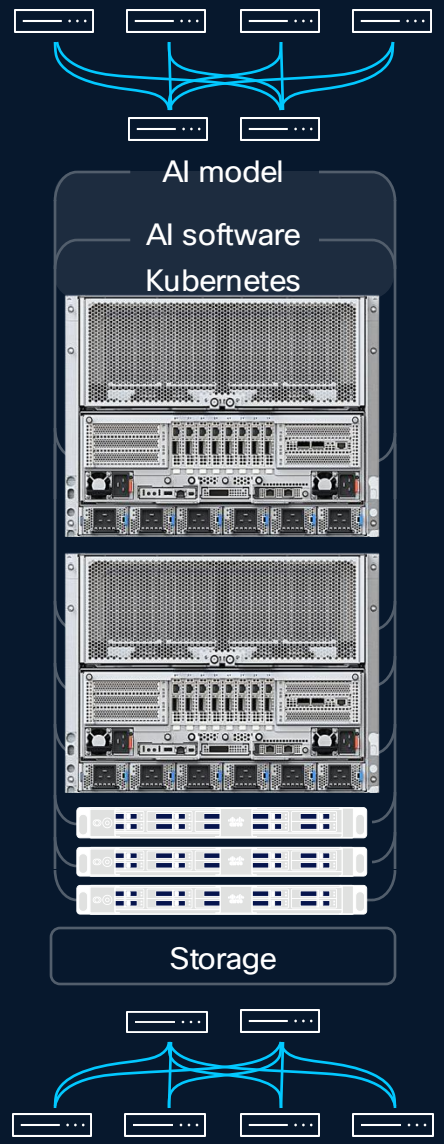
Cisco AI PODs

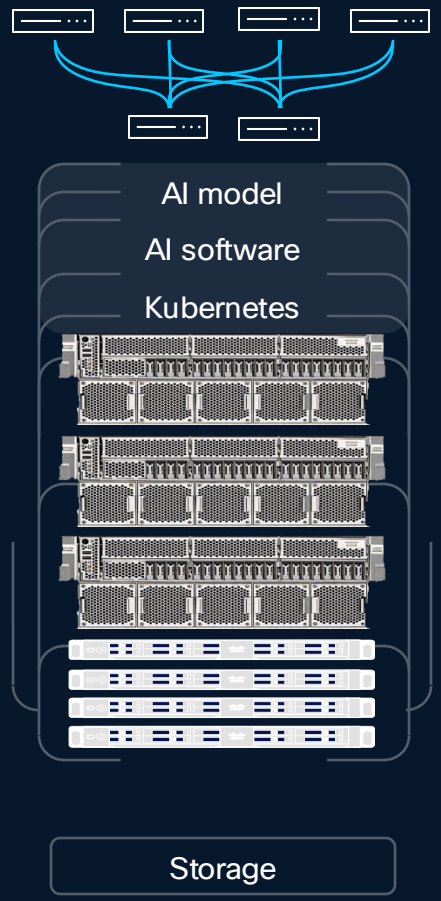
Expanded design portfolio

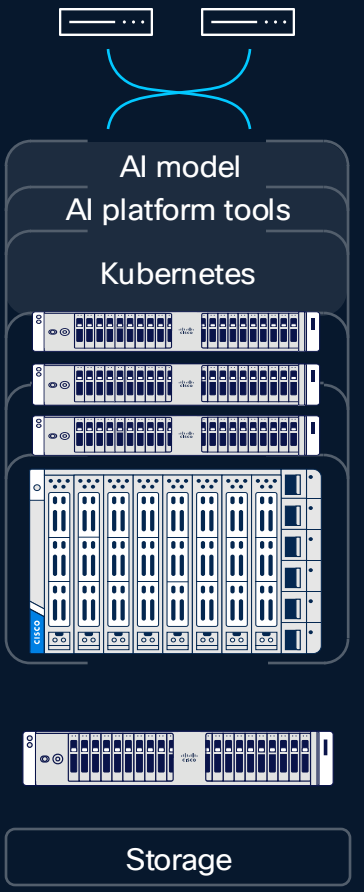
- Full AI lifecycle use-cases support
- Based on Cisco Validated Designs
- Pre-packaged AI-stack hardware + software platforms & automation
- Latest NVIDIA and AMD GPU Compute
- Adding VAST Data Storage
- Nexus 9000 Switches

Cisco Validated Architecture

Training, optimization and inferencing







Cisco AI PODs deployment options

Nexus Dashboard
On-premises managed network

- **Compute**: Cisco UCS — NVIDIA HGX, MGX or RTX PRO 6000 Servers
- **Networking**: Cisco Nexus 9300 — w/ Silicon One, w/ Cloud Scale, w/ NVIDIA Spectrum-X
- **Storage**: NetApp | Pure Storage | VAST Data | Hitachi Vantara

Nexus Hyperfabric
Cisco-managed network cloud controller

- **Compute**: Cisco UCS — NVIDIA HGX, MGX or RTX PRO 6000 Servers
- **Networking**: Cisco Nexus Hyperfabric — w/ Silicon One
- **Storage**: NetApp | Pure Storage | VAST Data | Hitachi Vantara

Nexus Hyperfabric AI
Cisco-managed network cloud controller

- **Compute**: Cisco UCS — NVIDIA HGX (885a with 8x H200), Future: RTX PRO Server
- **Networking**: Cisco Nexus Hyperfabric — w/ Silicon One
- **Storage**: VAST Data

Customizable ⟷ *Prescriptive*

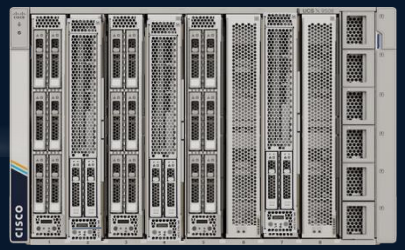
AI Platform Considerations: UCS GPU Options

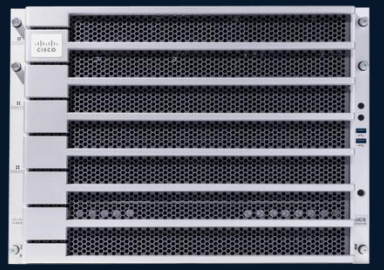
Potential Workload Type

- 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*		NVIDIA H200 NVL*	NVIDIA H100 / H200 SXM
	AMD MI210*			AMD MI300X OAM
	NVIDIA H100 NVL*			
	NVIDIA RTX PRO 6000 *			
	NVIDIA L40S*			
Max GPUs	3	2-8*	2-8*	2-8	8



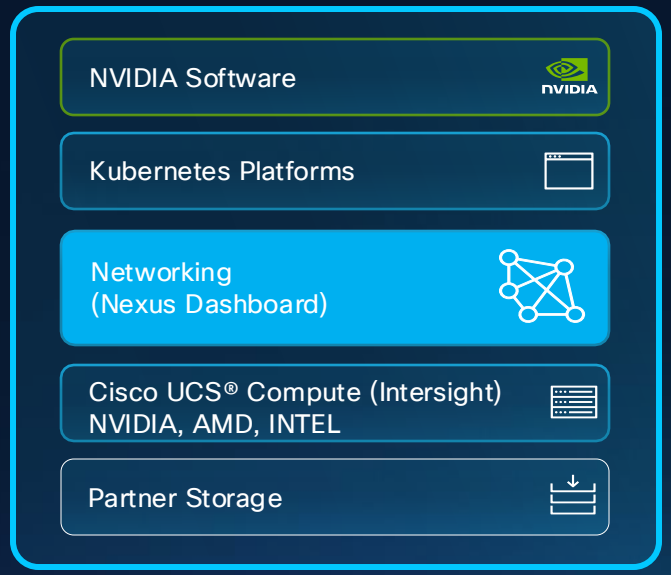


** NOTE: GPU Form Factor and GPU model support may vary between AMD and Intel Platforms (i.e. c220/c225, c240/c245, and x210c/x215c). Check the spec sheet for each platform to determine maximum GPU support based on GPU selection*

cisco

Cisco AI PODs: Flexible Operating Models

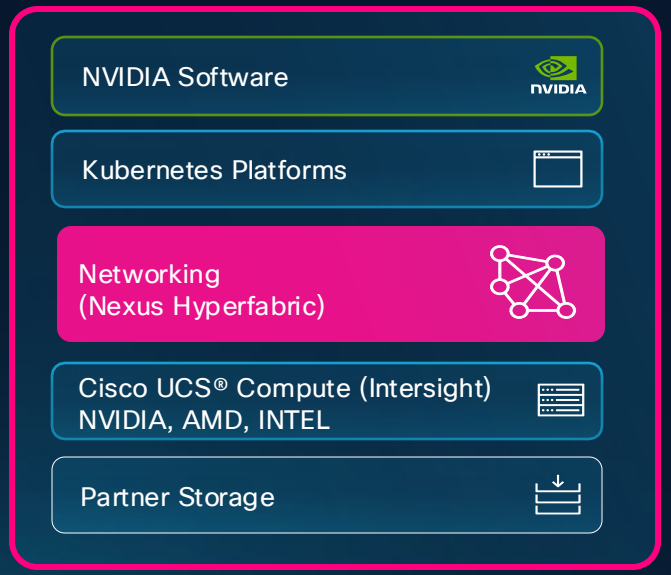
AI POD w/ On-prem management



Modular, pre-validated infrastructure:

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

AI POD w/ Cloud management

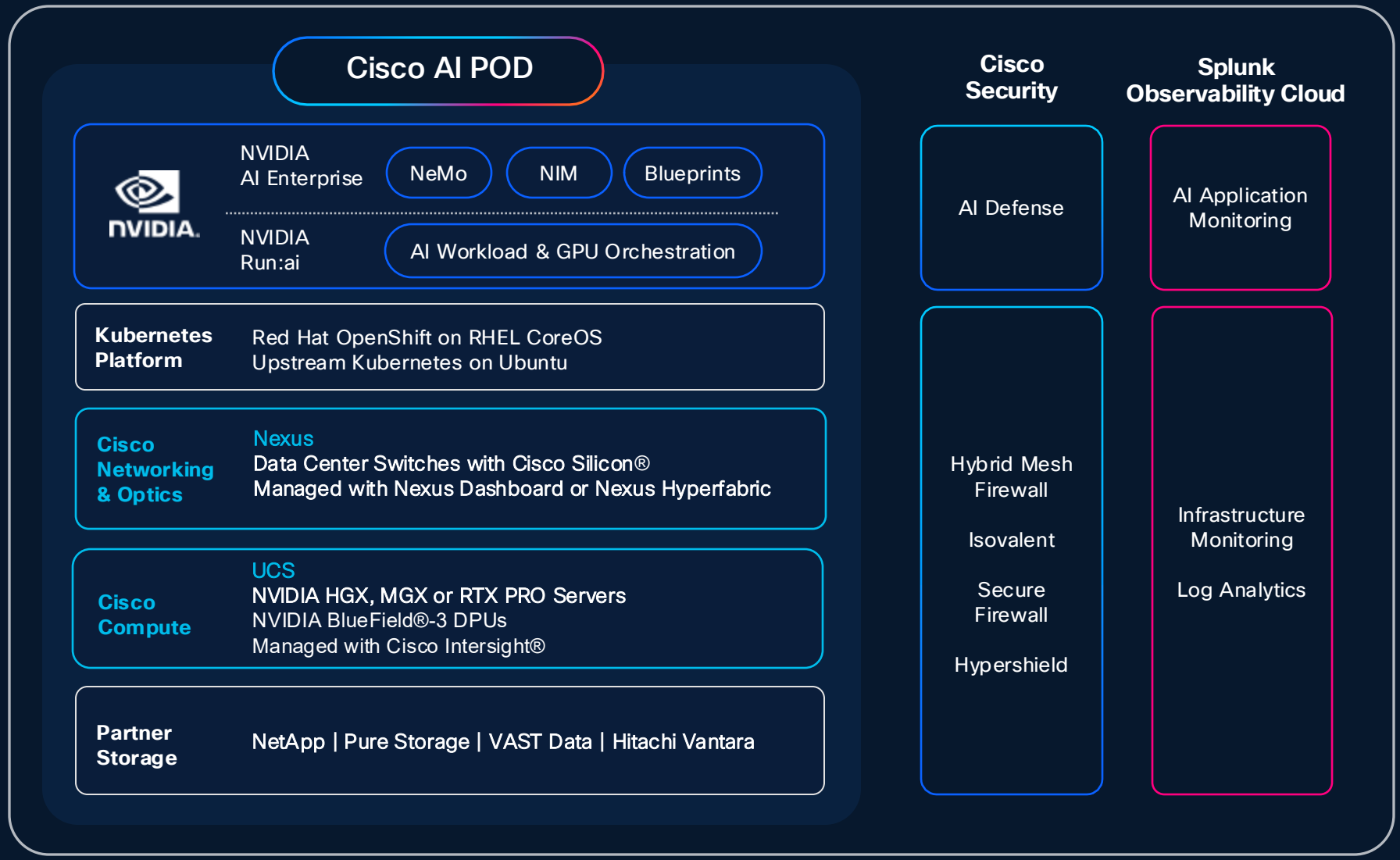


Turnkey infrastructure:

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

Cisco Secure AI Factory with NVIDIA

Delivering trusted AI outcomes



Bringing observability to the Factory

Cisco AI Defense on AI POD – Summary

End-to-end security

Continuous visibility and governance

Sovereign ready

DATA CONTROL

SECURITY

USE CASES

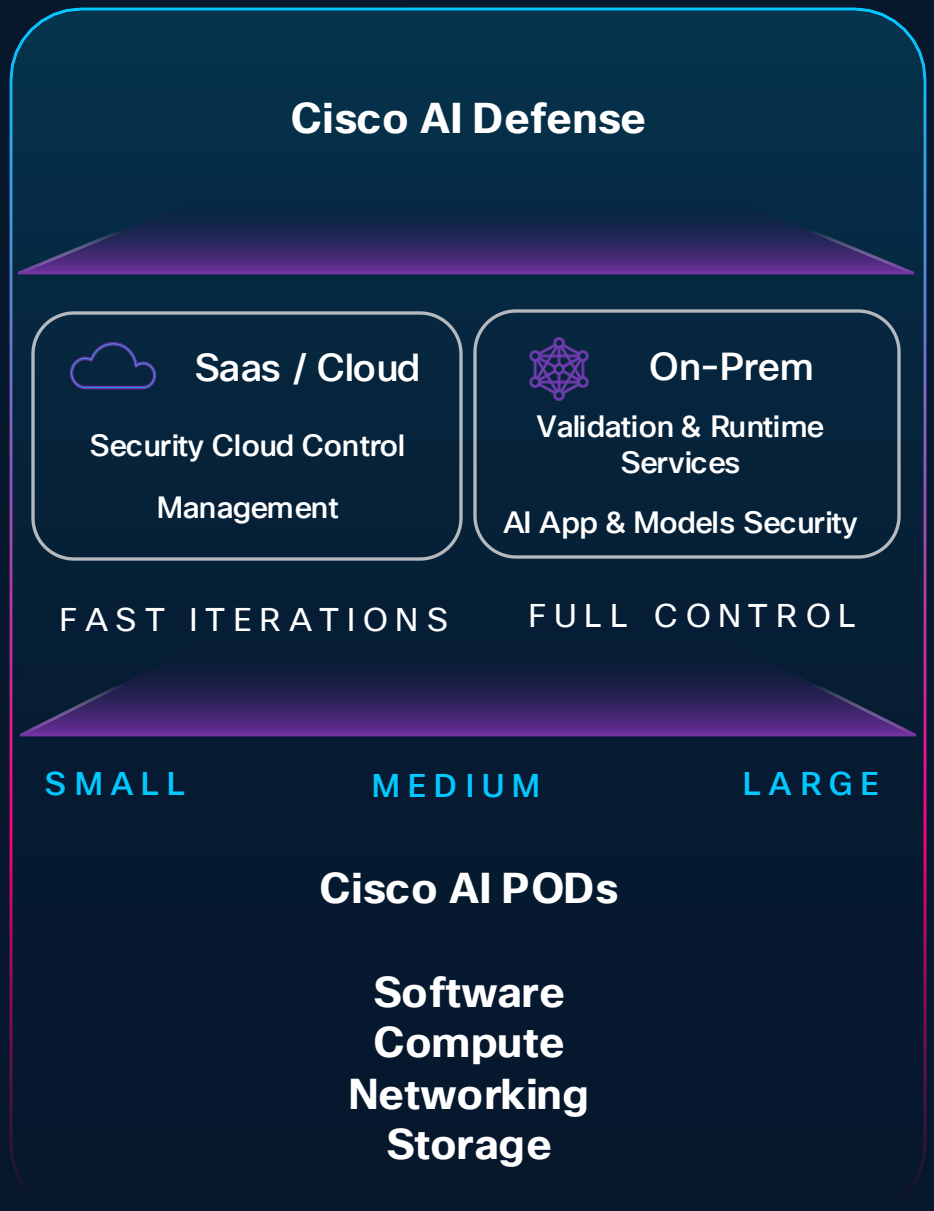
MODEL AND CUSTOMIZATION

Beta

Available

Controlled Availability

Expected in 90 days



Full Application Security

AI Application and Model Validation

AI Runtime Application Protection

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

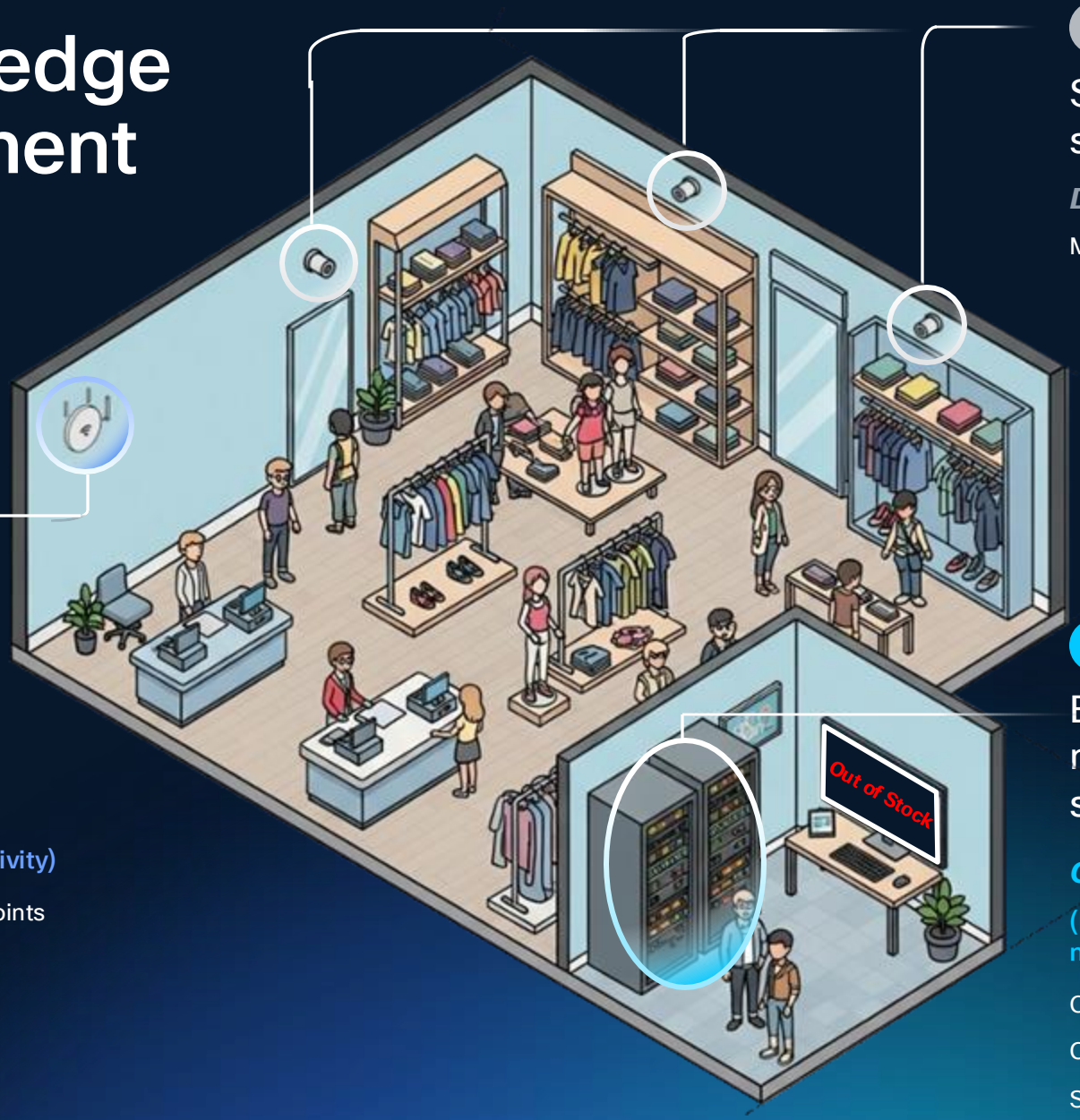
Pre-validated AI infrastructure stack for simplified deployment drastically reduces set-up time

ILLUSTRATIVE ONLY

Typical edge deployment

Retail

User Edge



A

Sensors and smart cameras

Device Edge

Meraki Smart Cameras

B

Gateways / access points

On Premises Edge

(Light edge and connectivity)

Meraki/Catalyst Access Points

Far Edge

C

Edge server, networking and security

On Premises Edge

(Local heavy compute, networking)

Cisco Unified Edge

Catalyst

Security, Observability

Inventory Management

A Track items on shelves

Send data to gateways

B Aggregate raw data

Do local pre-processing

Perform light computation*

Send data to edge servers

C Run heavier AI analytics / models locally

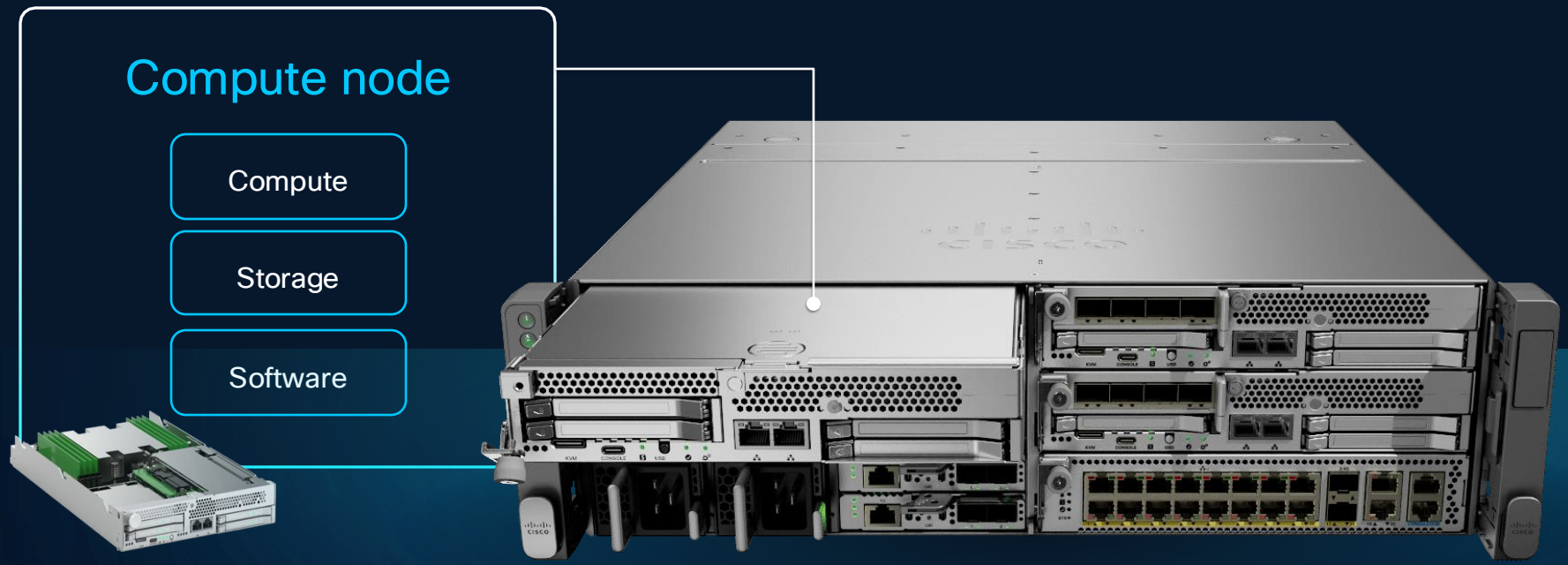
Send real-time alerts on low inventory, out-of-stock

* Gateways can also run lightweight AI models and send real-time alerts

Cisco Unified Edge: Future-Ready Performance

Future-ready performance

Integrates compute, networking, storage, and security



NUTANIX — Nutanix Stack for Edge

Red Hat — Red Hat Edge Stack

vmware by Broadcom — VMware Edge Compute Stack

intel — Intel Tiber

cisco — Virtual Networking

RANCHER BY SUSE — SUSE Edge

cisco

Cisco Unified Edge

Fully validated, full-stack environment that integrates advanced network, compute, storage and security

Future-ready performance

Compute node

- Compute
- Storage
- Software

Networking node

- Network
- Security

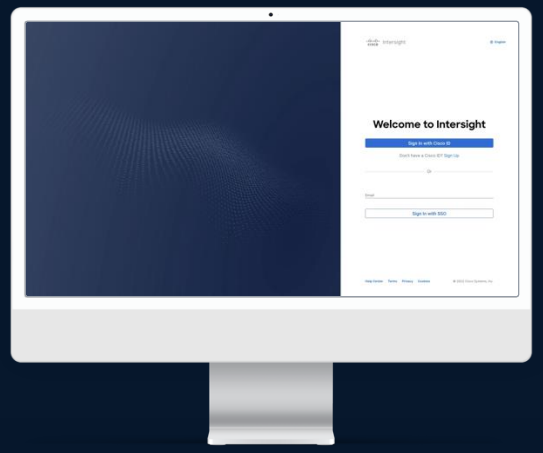
NUTANIX	Red Hat	vmware by Broadcom	intel.	cisco	RANCHER BY SUSE
Nutanix Stack for Edge	Red Hat Edge Stack	VMware Edge Compute Stack	Intel Tiber	Virtual Networking	SUSE Edge

cisco

Day 0: Zero touch provisioning

Unified operations

Consistent, repeatable AI-ready infrastructure deployments across multiple sites



Cisco Intersight

Cisco edge blueprints

- Application
- Cluster
- O S
- Network configuration
- Server configuration
- GPU configuration
- Storage configuration
- Security configuration

Golden configurations base on Cisco validated designs

Faster deployments with fewer errors

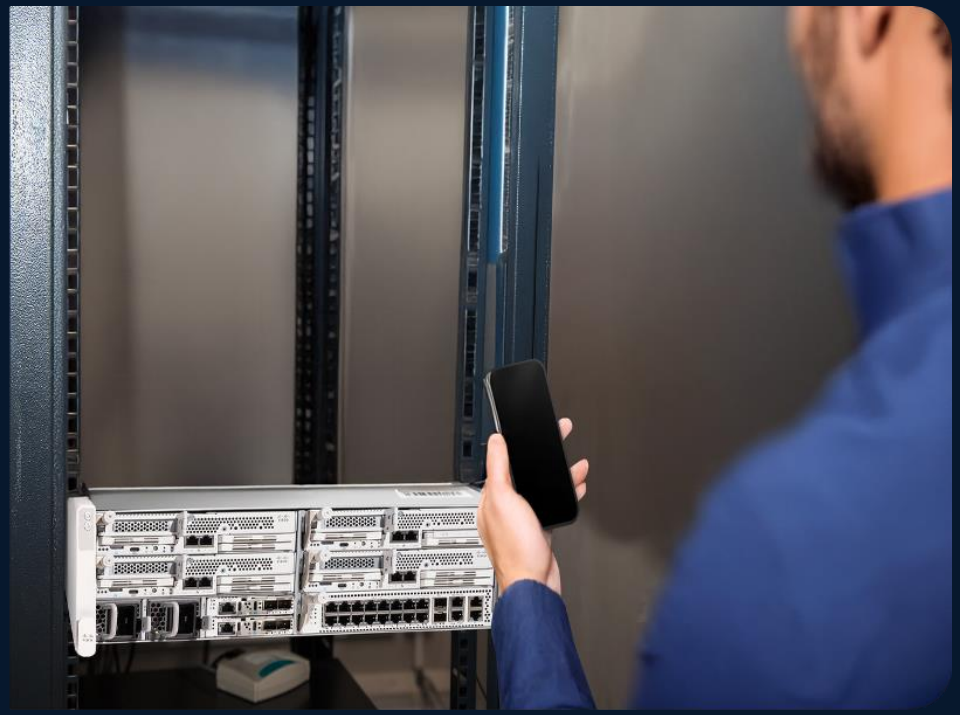
Supported by Cisco TAC

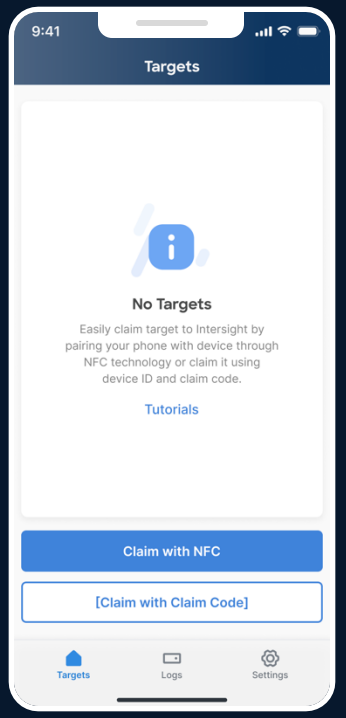
Available in a marketplace

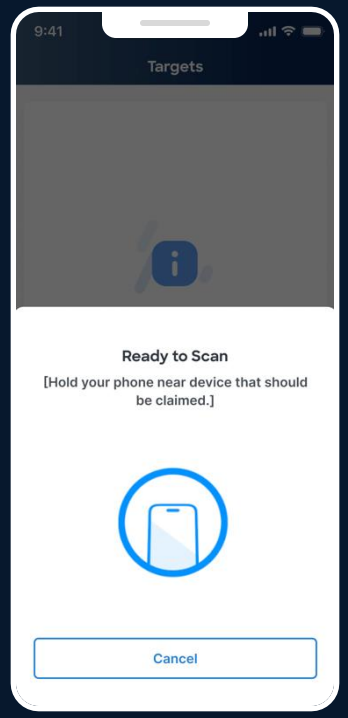
cisco

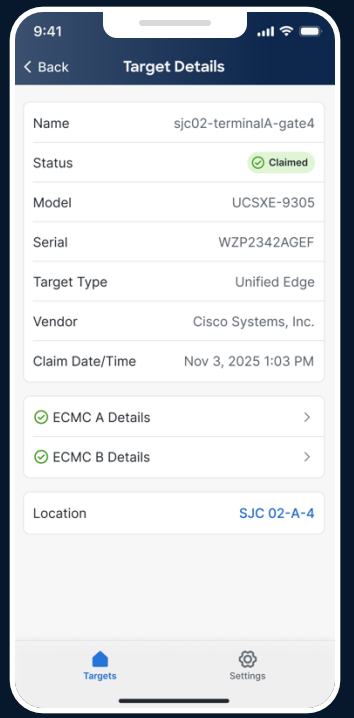
Day 1: NFC-based claim via mobile app

Unified operations









Fewer truck rolls – no need to onboard systems centrally

No technical expertise needed to claim device

Pre-claimed capabilities accessible to authorized on-site staff

Day 2: Deployment at scale

Consistent, repeatable AI-ready infrastructure deployments across multiple sites

Unified operations

Cisco Intersight

Cisco Edge blueprints
(Retail, manufacturing, healthcare)

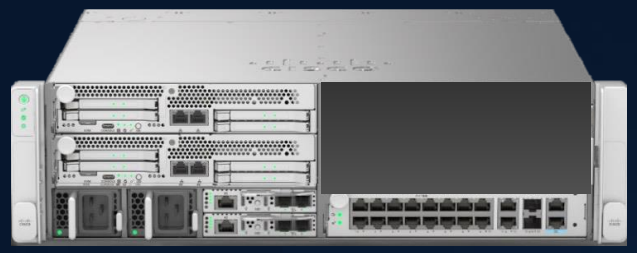
- Application
- Cluster
- O S
- Network configuration
- Server configuration
- GPU configuration
- Storage configuration
- Security configuration

Day N: Seamless, scalable operations

Fleet management at global scale

Unified operations

Deploy a new service



1. New AI application requires GPU
2. Install a new compute node with GPU
3. No forklift upgrade

Avoiding unscheduled down-time

1. Intersight identifies predictive maintenance item
2. Replacement node sent out automatically
3. Easy on-site install

End to end visibility & observability

Intersight + splunk> + ThousandEyes

- Fleet management
- Observability
- Security

cisco

Thank you


CISCO