CISCO Engage | Tech Day NYC

AI-Ready Data Center:

Compute Infrastructure for the AI Era

Eric Rose

Solution Engineer – Compute Architect
GES East – Cloud & AI Infrastructure
erirose@cisco.com

March 4, 2026

Our customers tell us

Standing up enterprise-grade AI infrastructure is hard. Organizations face complexity, fragmentation, and security concerns.

Too many tools, not enough integration

Uncertain ROI to build AI-ready environments

Security risks across the AI lifecycle

The AI-ready data center

Powers all workloads

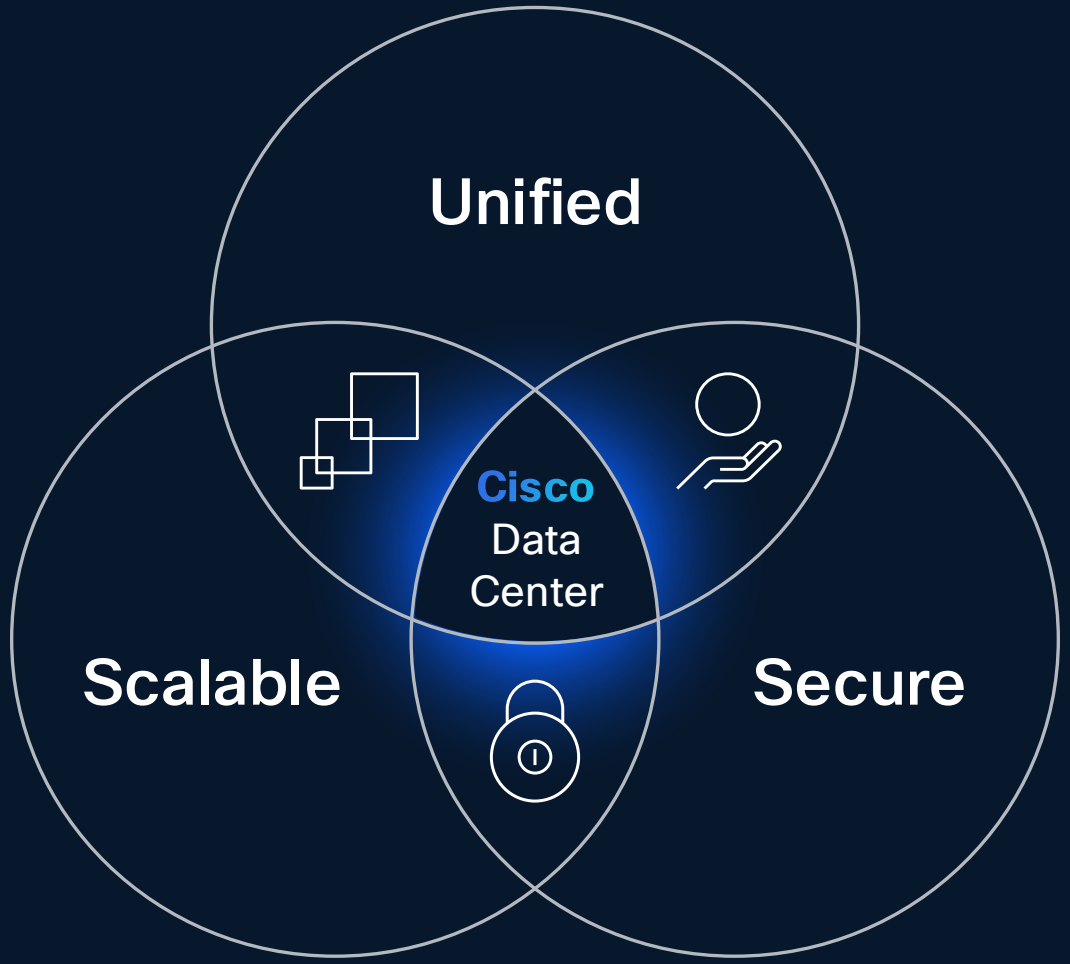
Scales for growth

Unifies management

Secures the entire stack

Keeps resilient

Whether you're modernizing traditional workloads, or scaling for AI, Cisco's approach is...



Cisco Networking

Cisco Unified Compute

Cisco Networking

Unified Architecture

The foundation for enterprise workloads

Cisco Unified Compute

Cisco Networking

Cisco Unified Edge

Cisco Silicon One

Cisco Optics

Compute for
every workload

UCS – AI use case focused servers

CISCO INTERSIGHT®

Validated solutions for AI

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Dense GPU

Modular (w/GPU Expansion) and Rack

Unified Edge

Demanding AI

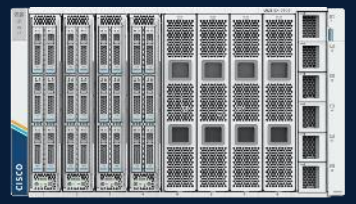
Mainstream and Edge AI

cisco

Infrastructure

Cisco UCS Compute Portfolio

Blade



UCS X-Series
X9508 Chassis

IFM Module

UCS X-Series Direct

UCS X210c M7

UCS X210c M8

UCS X410c M7

UCS X215c M8

UCS X580p
PCIe Gen5 node
PCIe Gen5 switch module

Rack

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 C220 M8E3S
16 EDSFF E3.S1T

UCS C220 M8S
10 HDD/SSD/NVMe

UCS C220 M7N
10 NVMe

UCS C245 M8SX
28 HDD/SDD

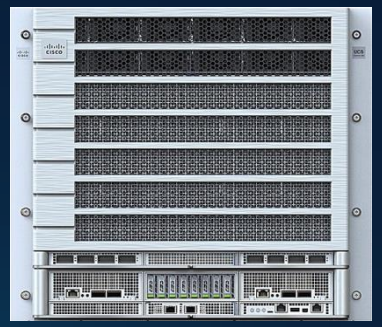
UCS C225 M8S
10 HDD/SSD

UCS C225 M8N
10 NVMe

AI (Dense GPU) Servers

UCS C885A M8
8RU Dense GPU Server

UCS C845A M8
4RU MGX Server



UCS C880A M8
10RU Dense GPU Server

Unified Edge

Unified Edge
UCS XE9305 Chassis
UCS XE130c M8
Compute Nodes

X-Series Portfolio

COMPUTE

X210c Compute Node

- 2-Socket, single slot servers
- **Three Generations: M6, M7 and M8**
- Intel 3rd Gen (Ice Lake)
  4th Gen (Sapphire Rapids)
  5th Gen (Emerald Rapids) and
  6th gen (Granite Rapids) Xeon CPUs

X410c Compute Node

- 4-Socket, dual slot servers
- Intel 4th Gen Xeon CPU
- Intel 6th Gen Xeon CPU
- Up to 64 DDR5 DIMMs

X215c Compute Node

- 2-Socket, single slot servers
- AMD 4th gen EPYC CPU (Genoa)
- AMD 5th gen EPYC CPU (Turin)

FABRIC

4th, 5th and 6th Gen FI

- 25/100G ports
- Unified ports: 32G FC (6536). 64G FC (66xx)
- Supports VIC 1400, 14000 and 15000 series

UCS X-Series Direct

- Scale at the edge
  with X-series advantage
  for 1-16 servers

25/100G IFM

- 8 x 25/100G connectivity

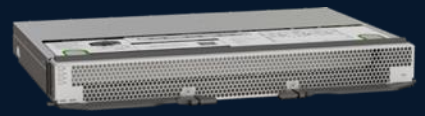
4th and 5th Gen VIC

- 25/100G connectivity
  for both blades and racks

X-FABRIC AND PCIE NODE

X-Fabric

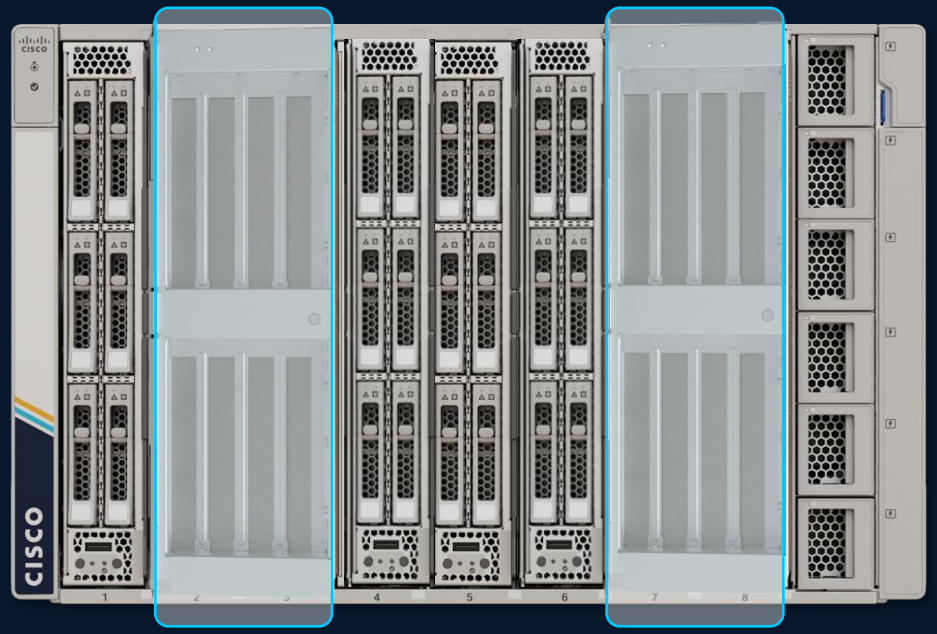
- Based on native PCIe Gen 4 or PCIe Gen 5
- Provides GPU acceleration to enterprise application
- No backplane or cables = Easy upgrades



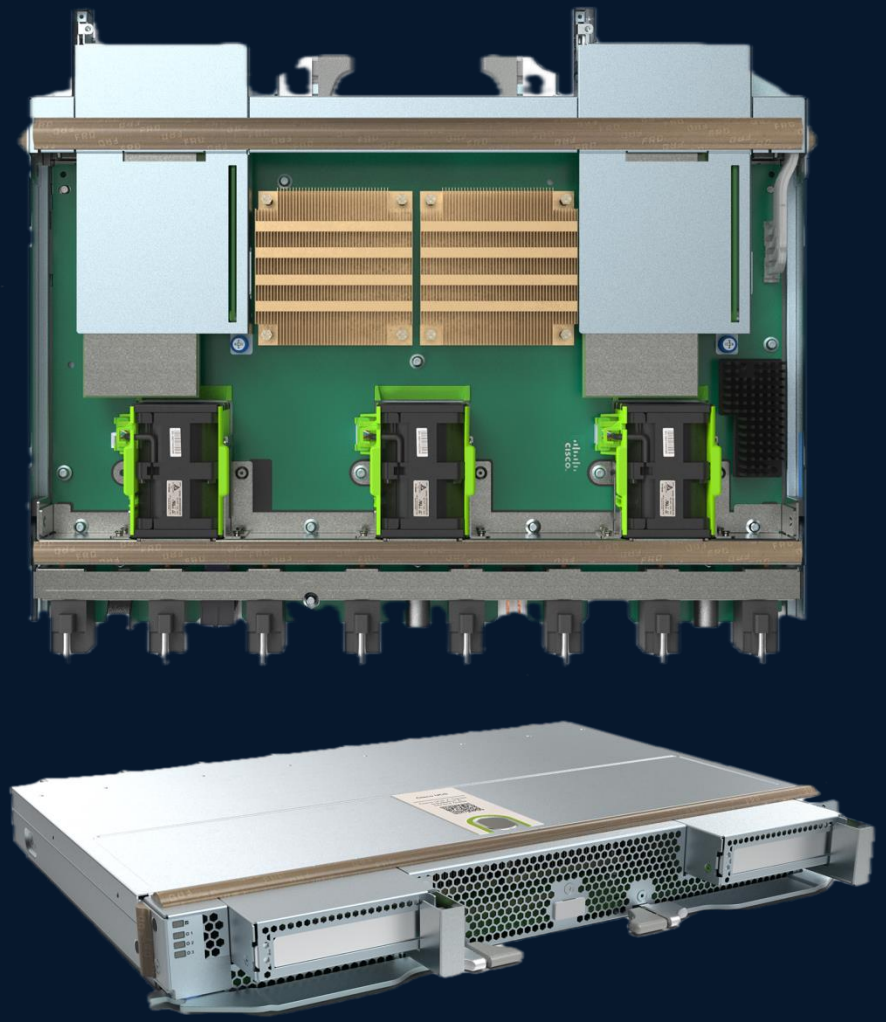
GPU Node and Front Mezz GPUs

- Nvidia and AMD GPU
- GPUs in various configurations

X580p PCIe Node



UCS 9516 X-Fabric

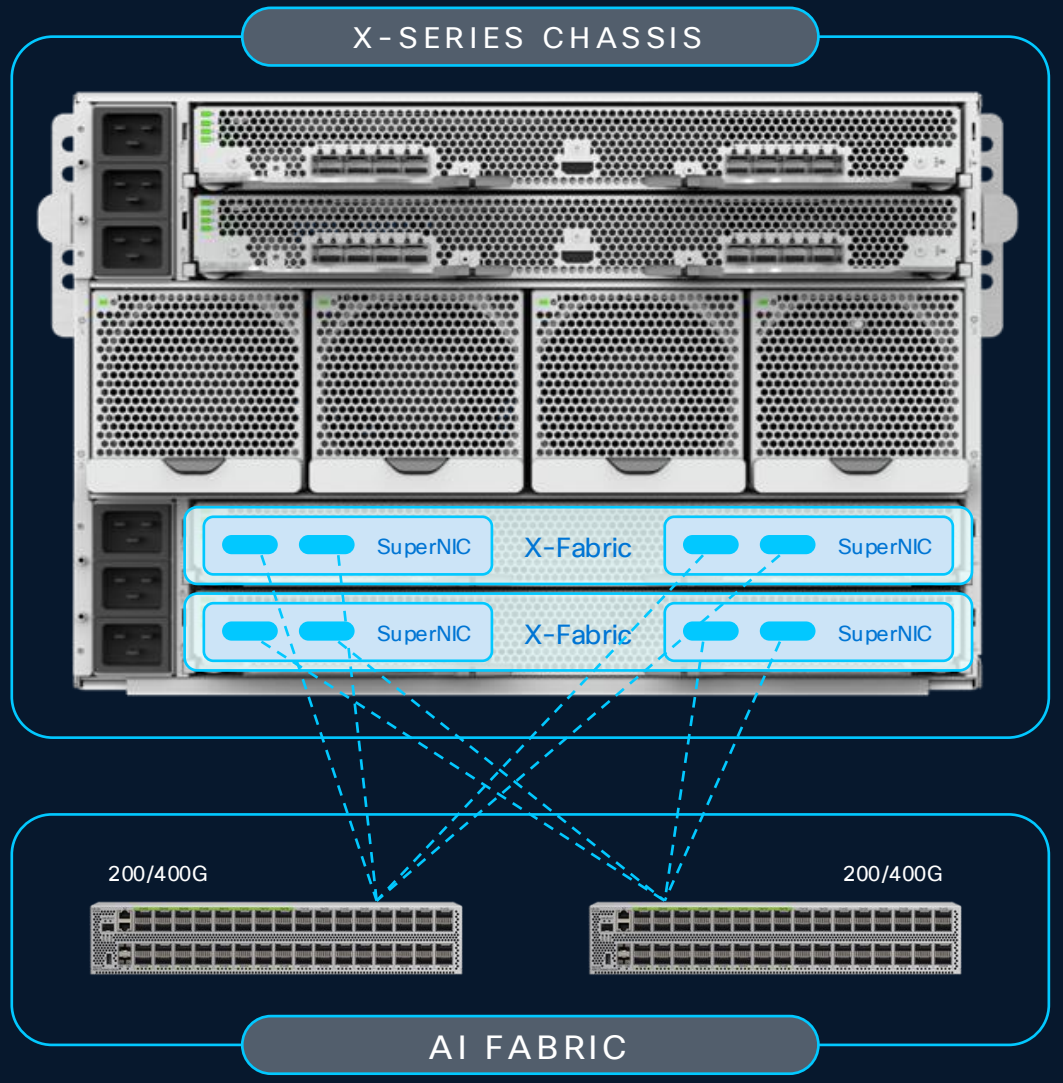


AI Cluster Expansion

GPU-to-GPU connectivity

with XFM external ports

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

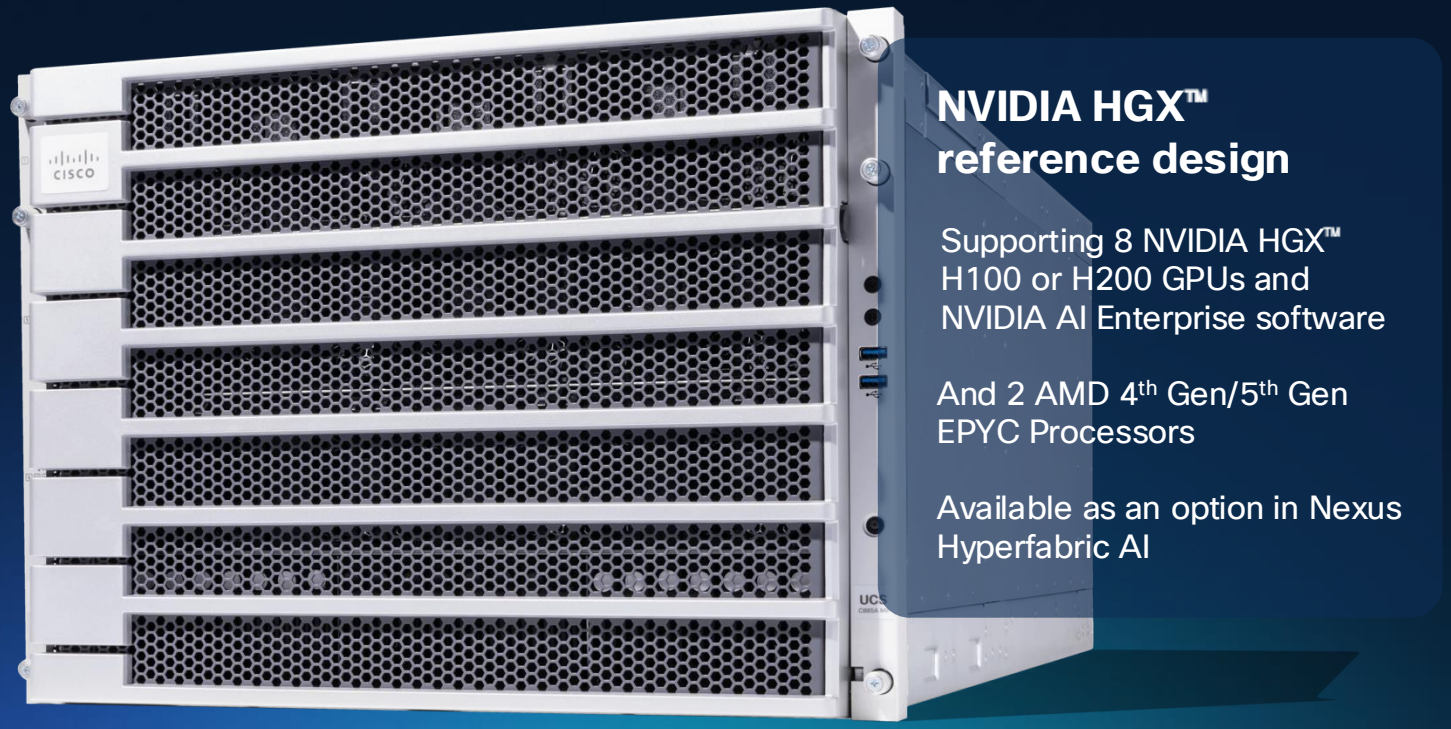


Shipping Now

High-density GPU servers

For data-intensive use cases like model training and deep learning

UCS accelerated | Cisco UCS C885A



NVIDIA HGX™ reference design

Supporting 8 NVIDIA HGX™ H100 or H200 GPUs and NVIDIA AI Enterprise software

And 2 AMD 4th Gen/5th Gen EPYC Processors

Available as an option in Nexus Hyperfabric AI

Shipping Now

Flexible, modular AI servers

"Start small and scale up" with AI

UCS accelerated | Cisco UCS C845A

NVIDIA MGX™ reference design

With NVIDIA H100, H200, L40S, AMD MI210 GPUs

Included as an option in Nexus Hyperfabric AI

High performance in a compact form factor

Enhanced power delivery, fewer PCBs, and better cable routing for optimal airflow and thermal management

with NVIDIA RTX PRO 6000 Blackwell GPUs

Reliability when it matters most

Cisco tops list for server reliability

According to a survey by **the Information Technology Intelligence Consulting Corp (ITIC)** reported by Tech Channel

TechChannel®

Server Reliability Survey: IBM, Lenovo, Cisco Top ITIC's List

Unified Operations

We’re making AI easier than ever to operate

Intersight – Transform IT operations

One platform:

Unified, intelligent management of all Cisco UCS® compute infrastructure.

All locations:

Across data centers, co-location facilities, and edge locations.

Simplifies operations:

Cloud-native platform consolidates tools, automates tasks, delivers proactive insights.

Business outcomes:

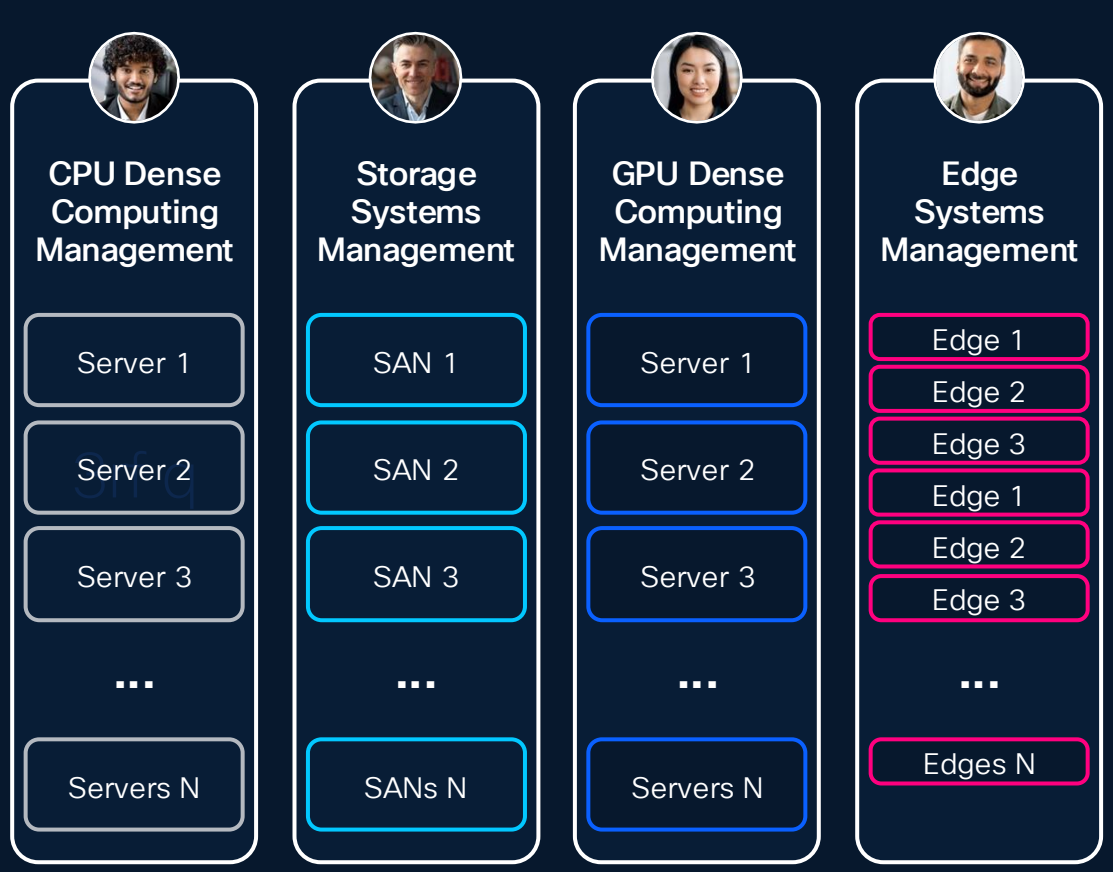
Accelerate traditional deployments and new AI initiatives, strengthen security, free up IT resources.

From reactive to proactive with Cisco Intersight®

Centralized infrastructure management

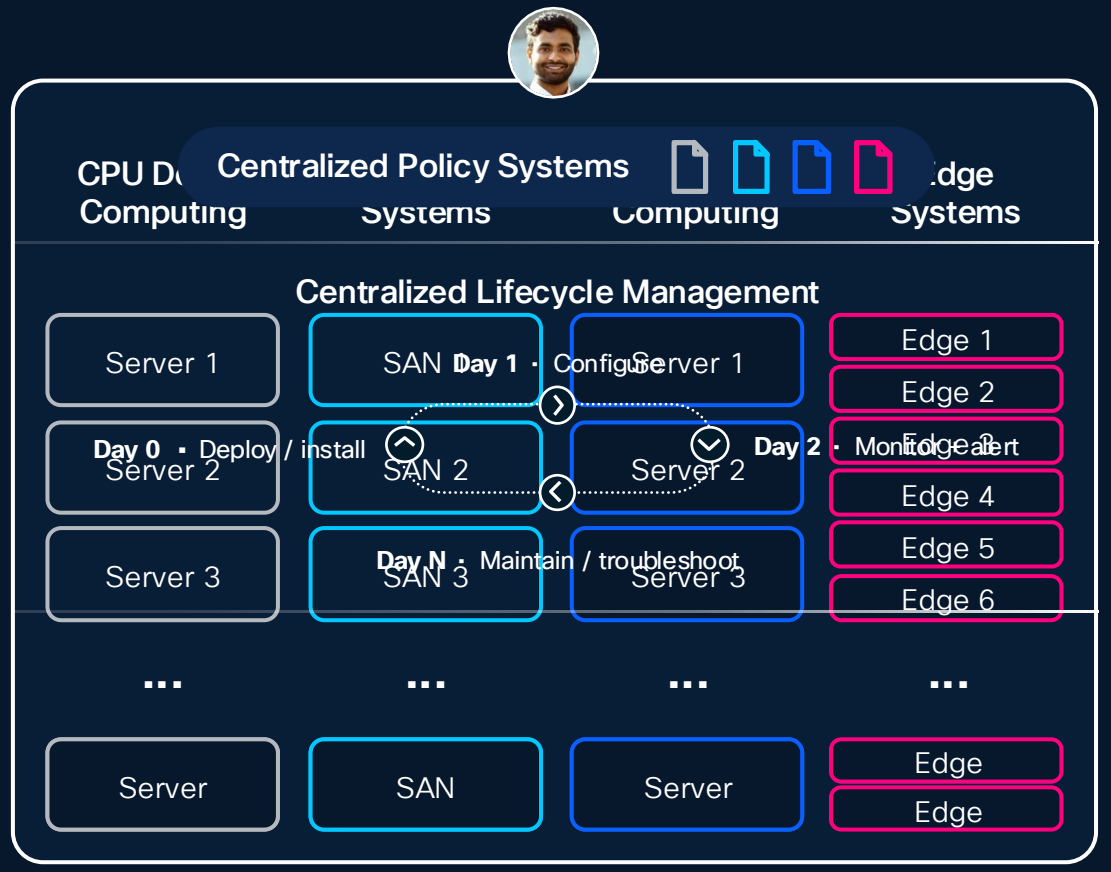
Conventional approach

Fragmented and Siloed Systems Management



Cisco Computing System

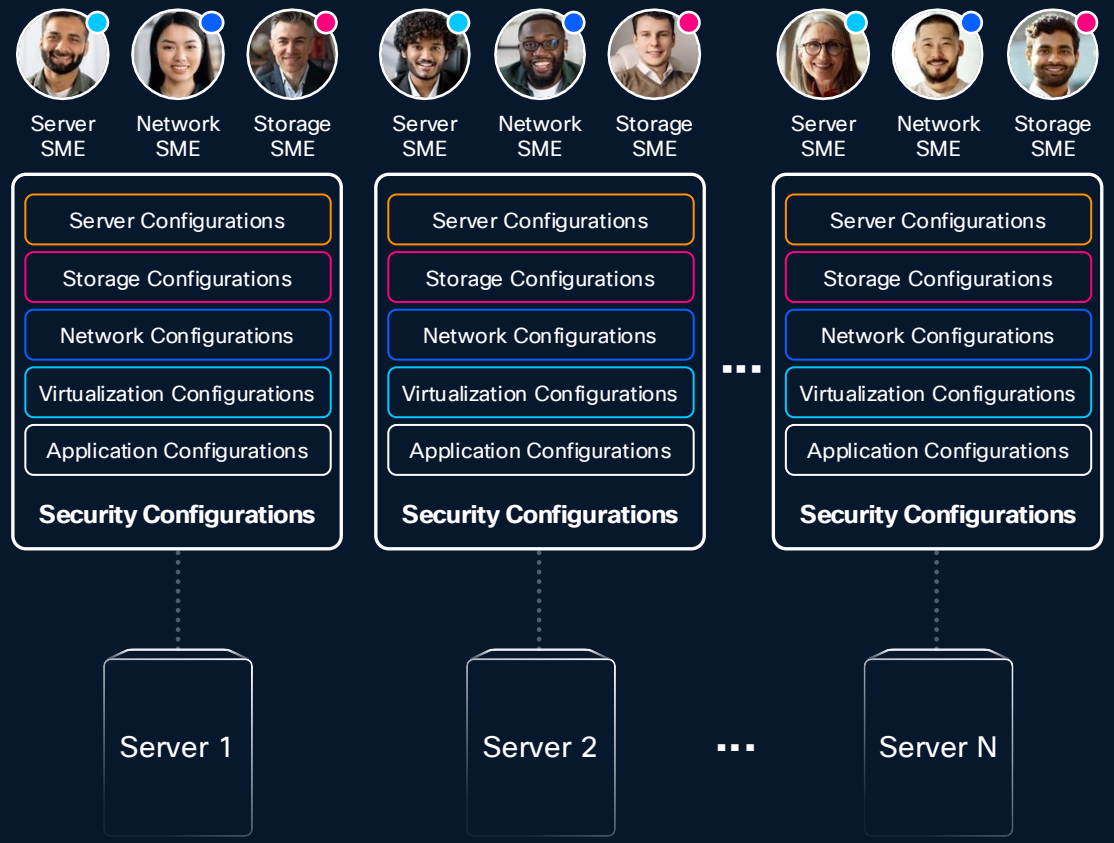
Centralized visibility, policy, automation across entire footprint



Cisco software-defined computing

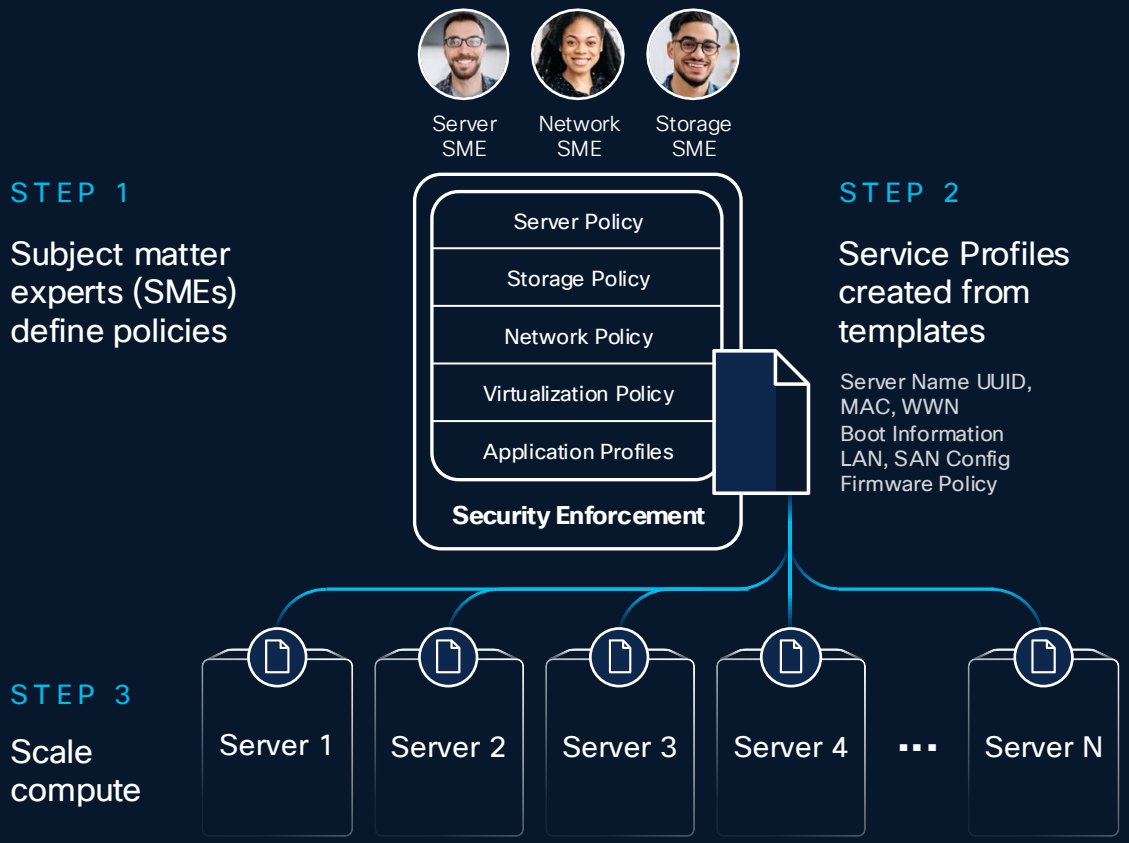
Conventional Approach

Massive complexity and time to scale compute



Cisco Computing System

State-less to State-full compute in minutes



Management at scale

Unified control, global reach

Comprehensive visibility and control

Gain real-time insights into and control of your distributed infrastructure

Support for Cisco® Compute solutions

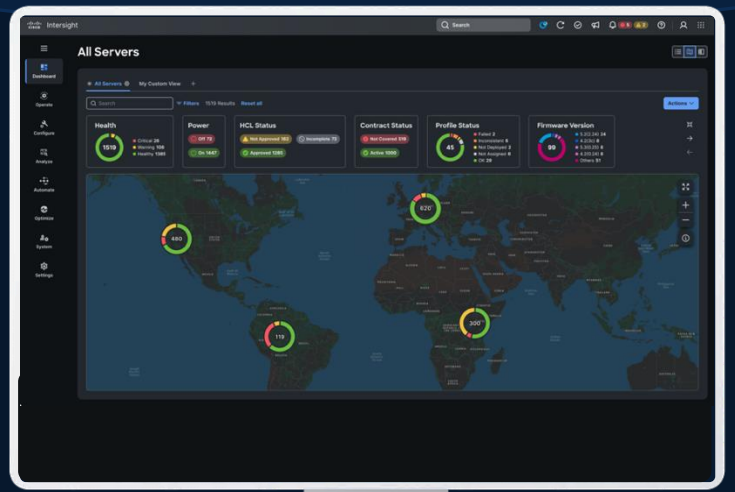
Manage any Cisco UCS® modular or rack server, including AI-accelerated GPUs, and all Cisco Compute solutions

Automated lifecycle management

Streamline day-0, day-1 operations and ongoing maintenance

Policy-driven automation

Define "golden configurations" once and apply across thousands of servers, ensuring consistency and compliance everywhere



Comprehensive infrastructure support

1

Cisco UCS®

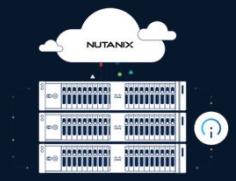
Configure, deploy, operate, and maintain Cisco UCS C-Series and X-Series rack and blade servers, anytime and anywhere

FlexPod FlashStack HITACHI

2

Converged infrastructure with Cisco UCS X-Series

See converged infrastructure inventory and incorporate into orchestrated workflows



3

Cisco Compute Hyperconverged with Nutanix

See and control your entire hyperconverged infrastructure fleet in one place—spanning clusters globally

4

AI-ready infrastructure

Deploy, monitor, and maintain AI training and inferencing solutions (Cisco Secure AI Factory, Cisco AI PODs, C885A, C845A)

5

Cisco Unified Edge

Deploy and operate consistent, repeatable AI-ready infrastructure across edge environments

Cisco Intersight® provides the most comprehensive set of infrastructure management capabilities for any Cisco UCS server form factor and most generations in one place.

Other vendors either provide less functionality, cannot cover their full server suite, or require you to use multiple tools.

cisco

Assisted operations

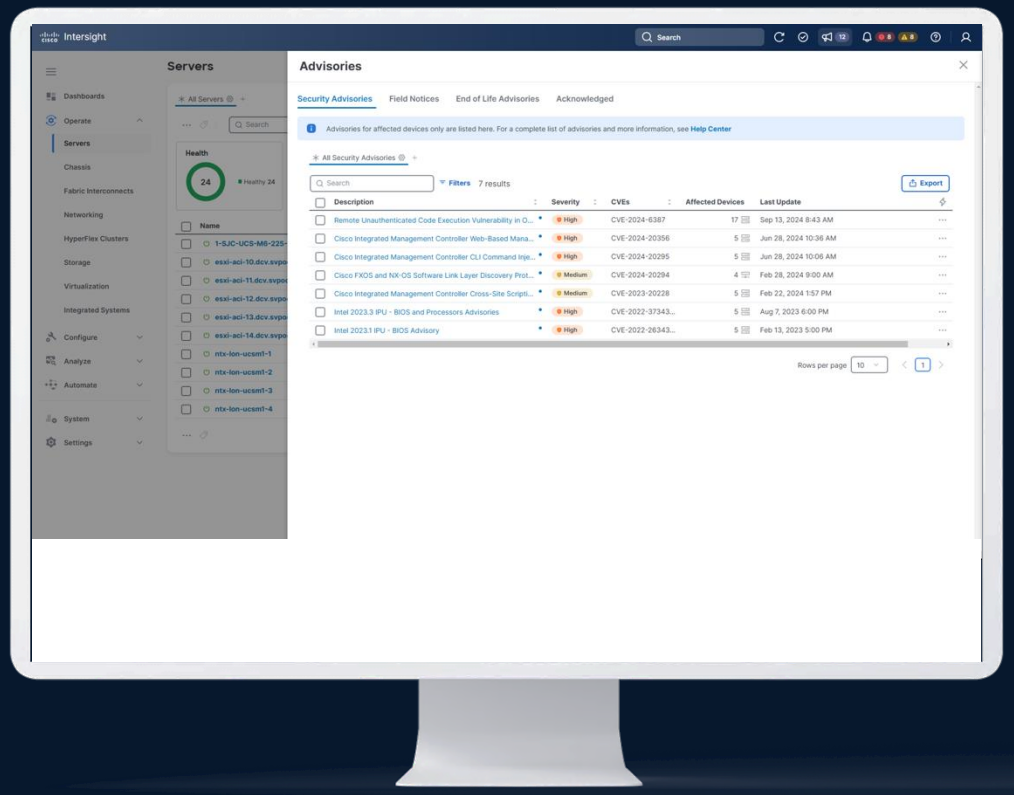
Proactive insights, automated resolution

Intelligent predictive analytics:

Anticipate issues before they impact operations

Proactive security and compliance:

Continuously strengthen your security posture



Automated issue resolution:

Accelerate troubleshooting and reduce manual effort

Workflow automation:

Automate complex tasks across your IT stack

Unified intelligence with Cisco Intersight

Next-gen metrics

Uncover hidden trends, optimize performance, and supercharge operational awareness

Enhanced topology view

Gain a holistic perspective of UCS domains for clarity, context, and control

Complete visibility and assurance

Gain deeper visibility into server health and performance, and ensure high-speed networking with proactive monitoring

Data center

On premises edge

AI increasing demand for local processing and networking

25%

2027

75%

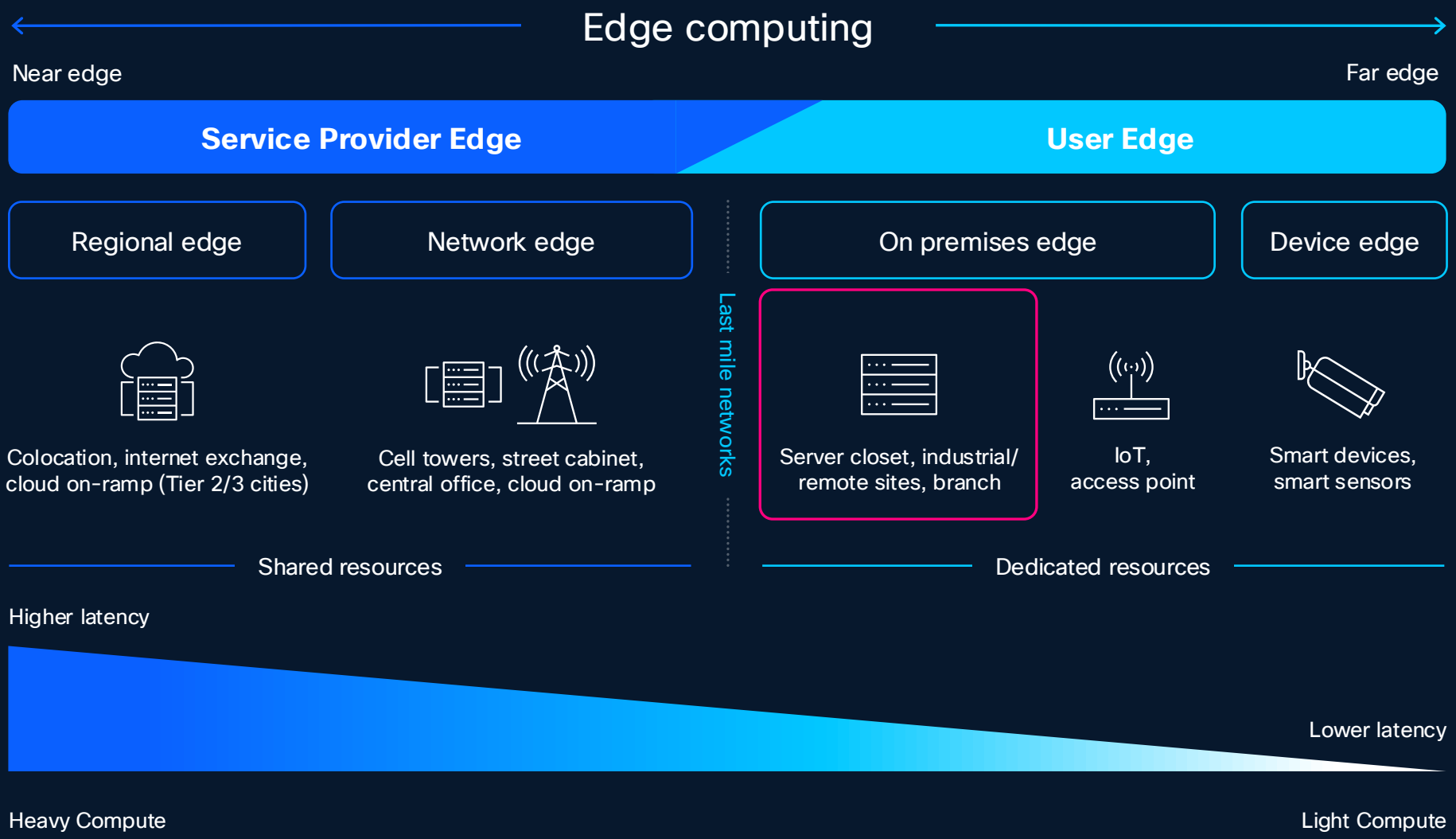
DATA PROCESSING

90%

2021

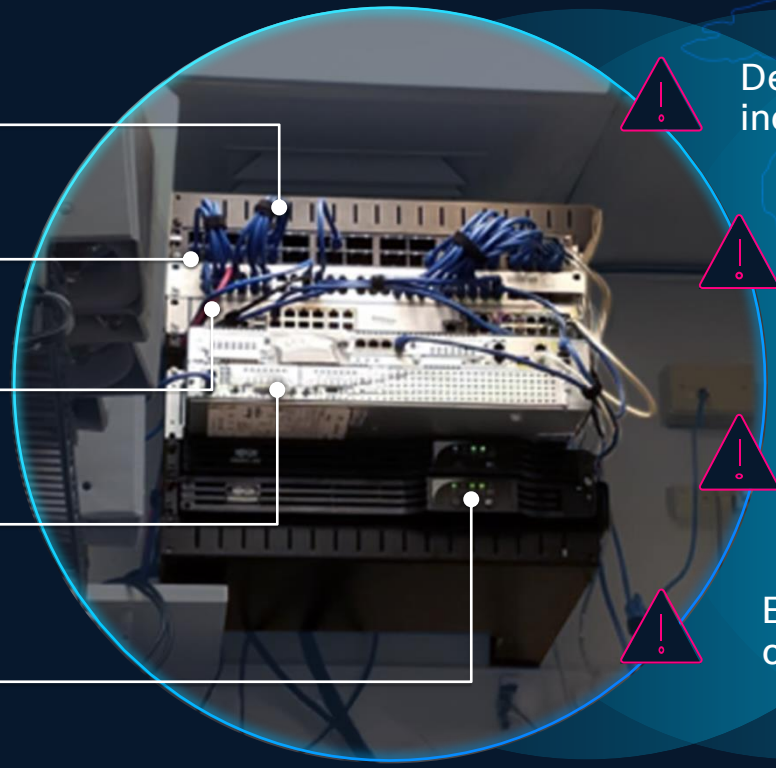
10%

The edge computing spectrum



However, the status quo is not ready for this future

Legacy edge infrastructure



Deployment inconsistency

Product incompatibility

Limited technical expertise

Environmental constraints

Deployed **everywhere**

Operational complexity

Security risks

Introducing Cisco Unified Edge

AI-ready edge

Compute | Storage | Networking | Software | SaaS Management | Analytics | Security

NUTANIX | Red Hat | vmware by Broadcom | intel | cisco | SUSE

cisco

Cisco Unified Edge: Future-Ready Performance

Integrates compute, networking, storage, and security

AI-ready edge

Compute node

- Compute
- Storage
- Software

GPU

Half-height/half-length GPU

NVIDIA L4 first

Additional GPUs on roadmap

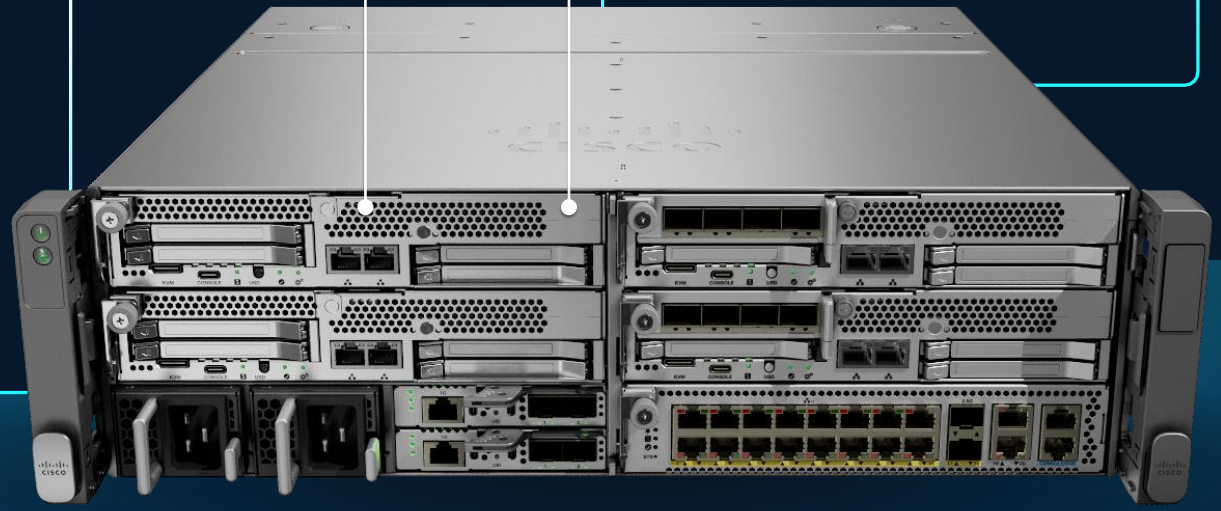
Intel Xeon 6 SoC

CPU native AI inferencing (AMX)

Confidential compute (TDX & SGX)

Integrated Ethernet

Scalable multithreaded cores (12, 20, 32)



NUTANIX · Red Hat · vmware by Broadcom · Canonical Ubuntu · Microsoft · intel · cisco · SUSE

Future-ready performance

Personalized

Serviceable

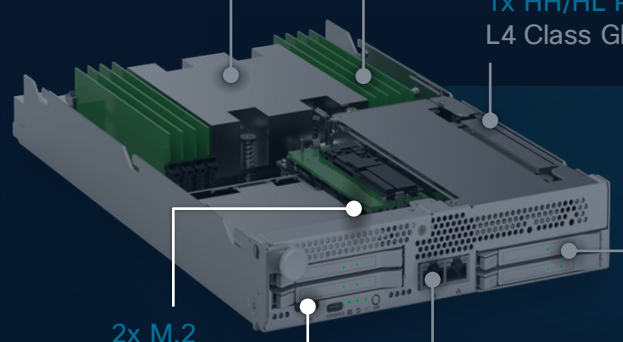
Modular

Storage-optimized configuration

Xeon-6 SoC
P-core processor
12/20/32 Core options

8x Memory DIMMs
Up to 768GB

1x HH/HL PCIe G5
L4 Class GPU



4x NVMe
E3.S drives
Up to 120TB

2x M.2
SATA drives
HW RAID

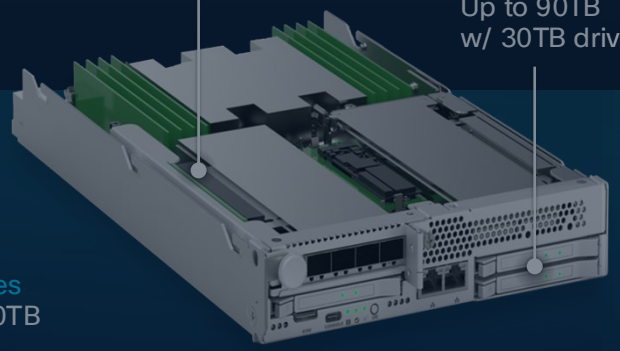
2x 1/10
G NIC ports

Console / KVM ports
For local access

IO-optimized configuration

1x HH/HL PCIe G5
NIC / DPU

3x NVMe E3.S drives
Up to 90TB
w/ 30TB drives



High-performance networking

25G internal network
Node to node
connectivity

Up to 5 modules
Mix and match
nodes

Node level personalization

Chassis level personalization

cisco

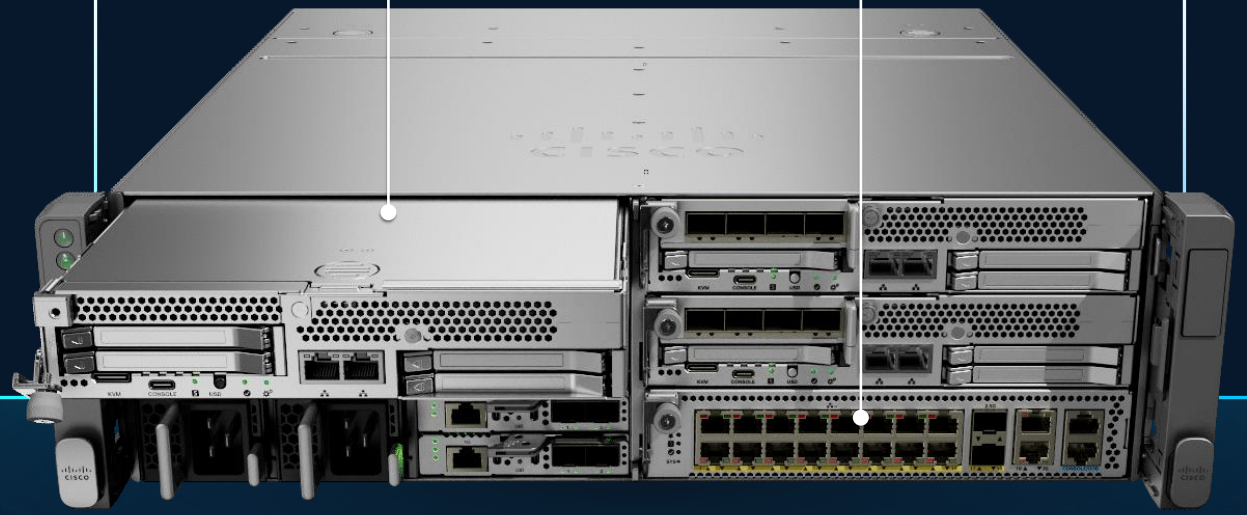
Cisco Unified Edge

AI-ready edge

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

Compute node

- Compute
- Storage
- Software



Networking node

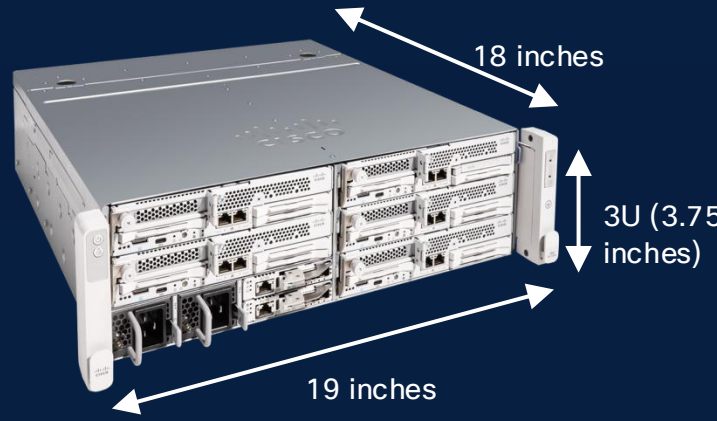
- Network
- Security

NUTANIX · Red Hat · vmware by Broadcom · Canonical Ubuntu · Microsoft · intel · cisco · SUSE

cisco

XE9305 chassis mounting options

4-post rack with sliding rails



Mount brackets for horizontal positioning

2-post rack with center mount brackets

Wall mount bracket

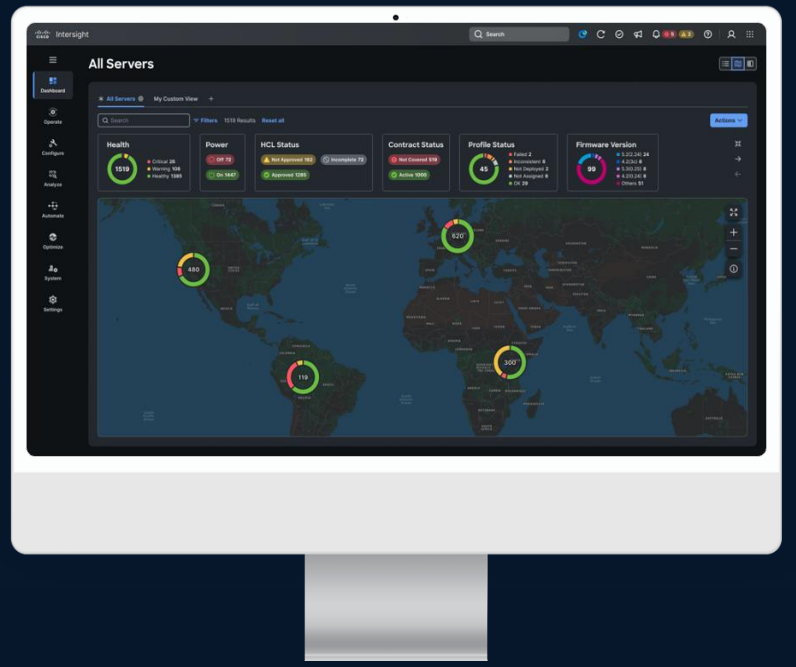
Mount brackets for vertical positioning

*Planned for post GA

Intersight fleet management

Consistent, repeatable AI-ready infrastructure deployments across multiple sites

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



Cisco Edge blueprints
(Retail, manufacturing, healthcare)

- Application
- Cluster
- OS
- Network config
- 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

14:43

84

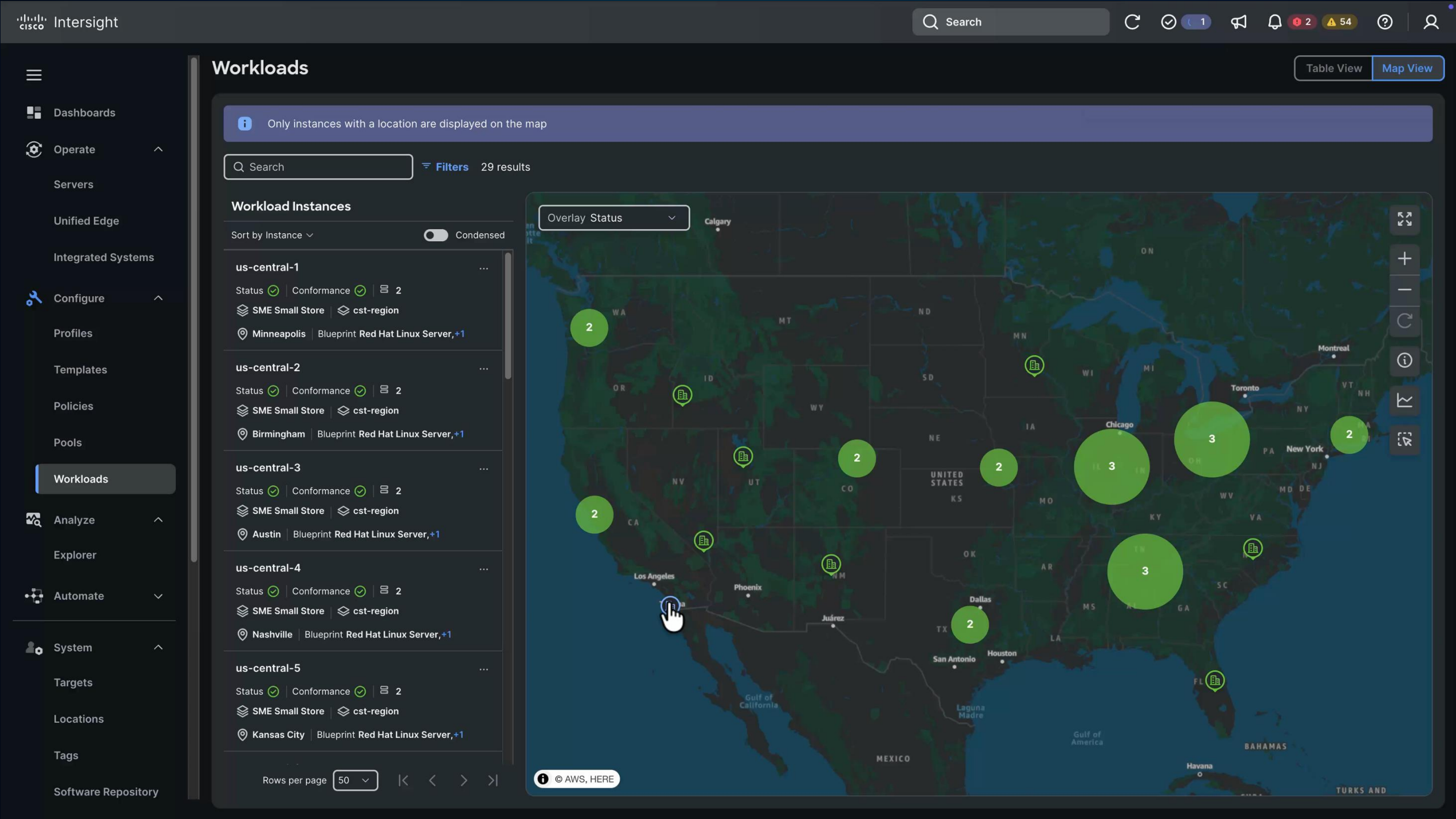
Claim with NFC

Optionally set Location and Resource Group. Tap Claim, then hold your mobile phone against the NFC marker on the Unified Edge device until the claim completes.

OPTIONAL

Location	San Diego
Resource Groups	Select

Claim


Intersight
Search
1
2
54
Workloads
Table View
Map View
Dashboards
Operate
Servers
Unified Edge
Integrated Systems
Configure
Profiles
Templates
Policies
Pools
Workloads
Analyze
Explorer
Automate
System
Targets
Locations
Tags
Software Repository
Only instances with a location are displayed on the map
Search
Filters
29 results
Workload Instances
Sort by Instance
Condensed
us-central-1
Status
Conformance
2
SME Small Store
cst-region
Minneapolis
Blueprint Red Hat Linux Server,+1
us-central-2
Status
Conformance
2
SME Small Store
cst-region
Birmingham
Blueprint Red Hat Linux Server,+1
us-central-3
Status
Conformance
2
SME Small Store
cst-region
Austin
Blueprint Red Hat Linux Server,+1
us-central-4
Status
Conformance
2
SME Small Store
cst-region
Nashville
Blueprint Red Hat Linux Server,+1
us-central-5
Status
Conformance
2
SME Small Store
cst-region
Kansas City
Blueprint Red Hat Linux Server,+1
Rows per page
50
Overlay Status
Calgary
Montreal
Toronto
Chicago
New York
Los Angeles
Phoenix
Juárez
Dallas
San Antonio
Houston
Gulf of California
Laguna Madre
Gulf of America
MEXICO
BAHAMAS
Havana
© AWS, HERE

Dashboards
Operate
Servers
Unified Edge
Integrated Systems
Configure
Profiles
Templates
Policies
Pools
Workloads
Analyze
Explorer
Automate
System
Targets
Locations
Tags
Software Repository

← Workloads

Instances: us-west-6

General | Configuration | Inventory

Details

Name
us-west-6

Health
Healthy

Conformance
Not OK

Status
In Progress

Workload
SME Small Store

Workload Deployment
pst-region

Running Workload Version
V1

Desired Workload Version
V1

Organization
default

Tags Set

workload/S... Global/Amer...

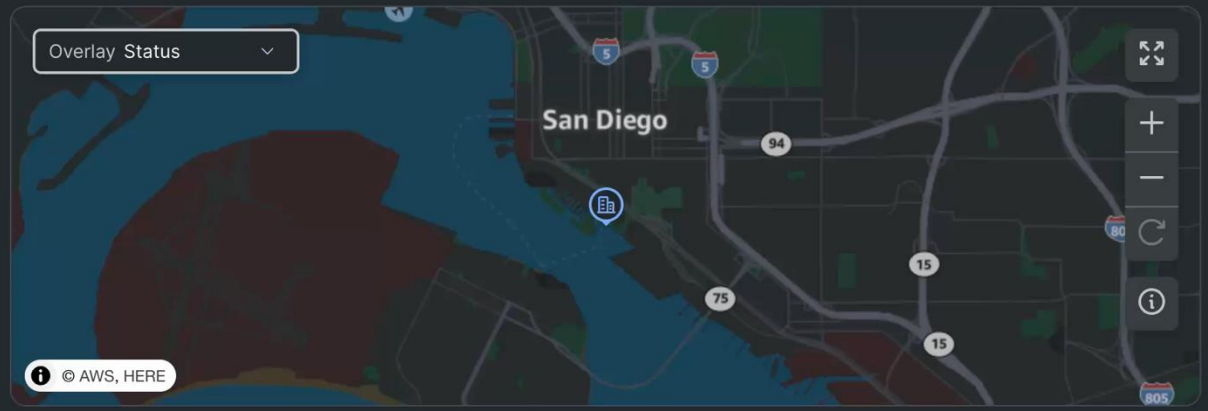
Properties

Cisco UCSXE-9305 Front | Rear

Locator LED Off

Health Overlay

Location



Events

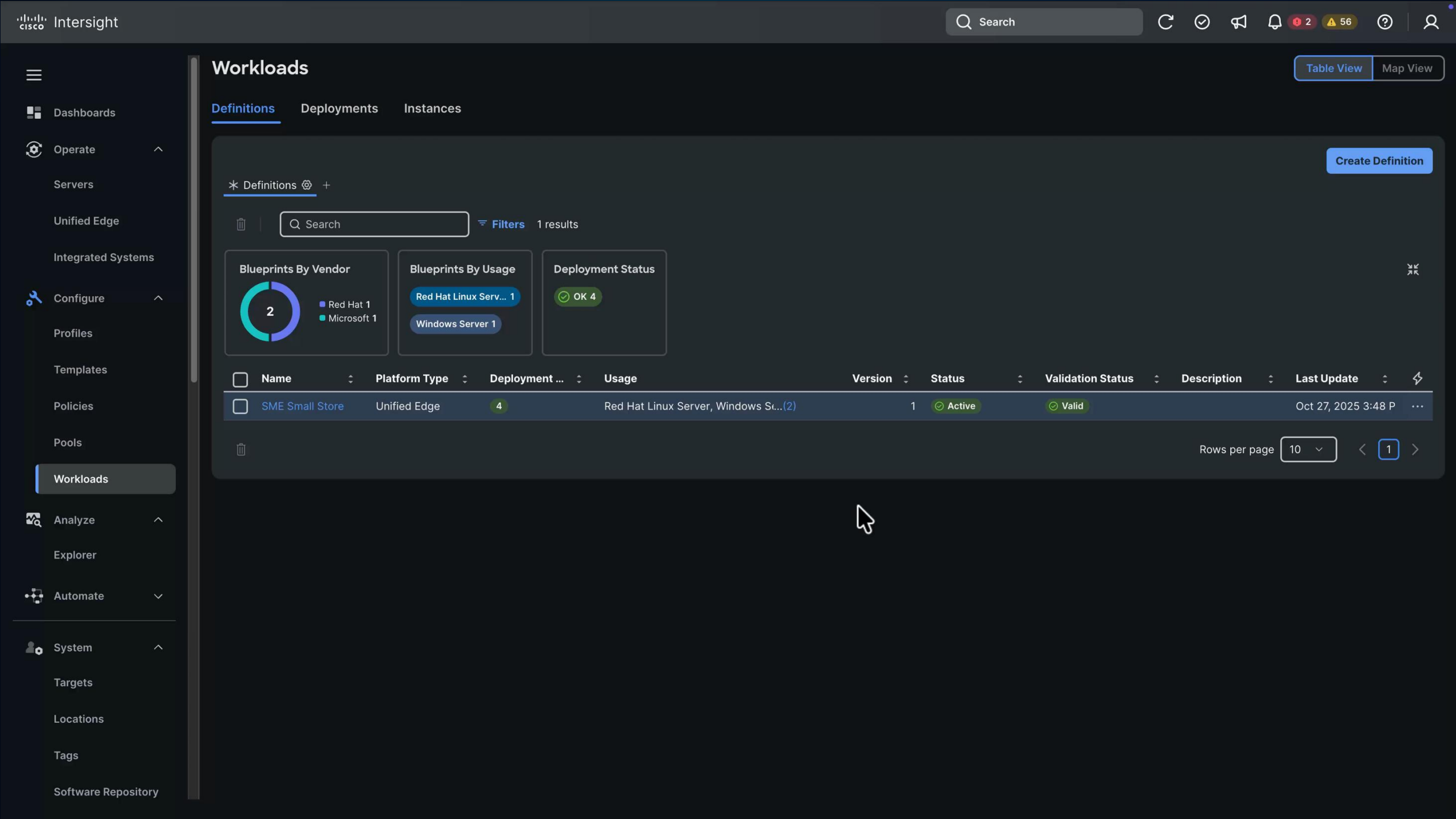
Alarms No Alarms

Requests 1

Requests for last 7 days

Deploy Workload Instance In Progress
us-west
a minute ago

Advisories No Advisories


Intersight
Search
2
56
Workloads
Table View
Map View
Definitions
Deployments
Instances
Create Definition
Definitions
Search
Filters
1 results
Blueprints By Vendor
2
Red Hat 1
Microsoft 1
Blueprints By Usage
Red Hat Linux Serv... 1
Windows Server 1
Deployment Status
OK 4
Name
Platform Type
Deployment ...
Usage
Version
Status
Validation Status
Description
Last Update
SME Small Store
Unified Edge
4
Red Hat Linux Server, Windows S...(2)
1
Active
Valid
Oct 27, 2025 3:48 P
Rows per page
10
1
Dashboards
Operate
Servers
Unified Edge
Integrated Systems
Configure
Profiles
Templates
Policies
Pools
Workloads
Analyze
Explorer
Automate
System
Targets
Locations
Tags
Software Repository

Dashboards
Operate
Servers
Unified Edge
Integrated Systems
Configure
Profiles
Templates
Policies
Pools
Workloads
Analyze
Explorer
Automate
System
Targets
Locations
Tags
Software Repository

← Definitions: Manage SME Small Store Versions

Create a New Version of Definition: SME Small Store

General
2 Blueprint Selection
Red Hat Server 1
Windows Server 1
System Configuration
Optional Settings
Summary
6 Validation Result

Validation Result

Check for validation errors and warnings.

Definition Validated

Next, **Create Deployments** to apply this definition to a fleet of systems. Deployments allow optional regional customizations, maintenance schedules, target resource selection, and specific rollout strategies for new versions.

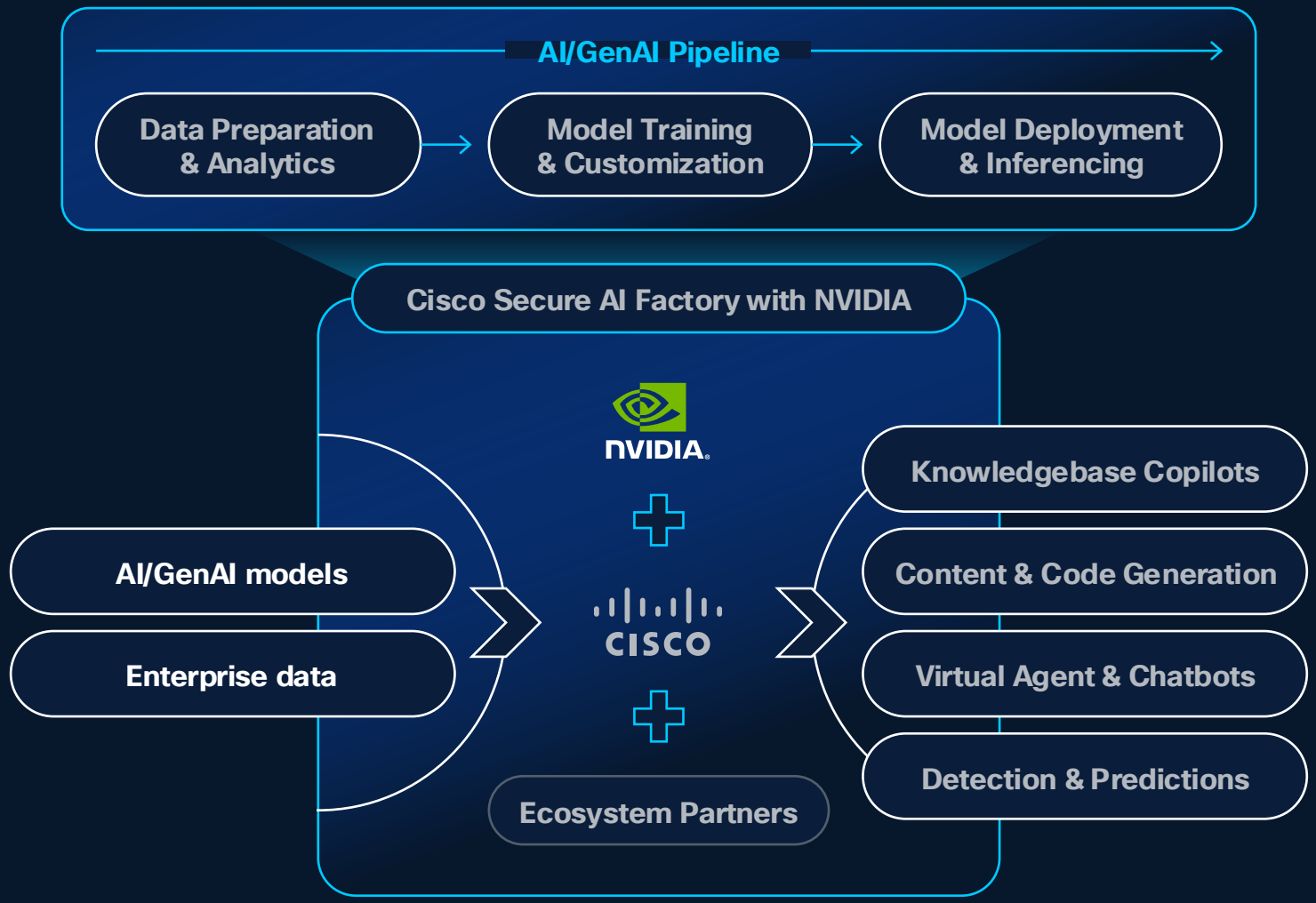
Unified solutions

We’re expanding our solutions to meet your needs

cisco

Cisco Secure AI Factory with NVIDIA

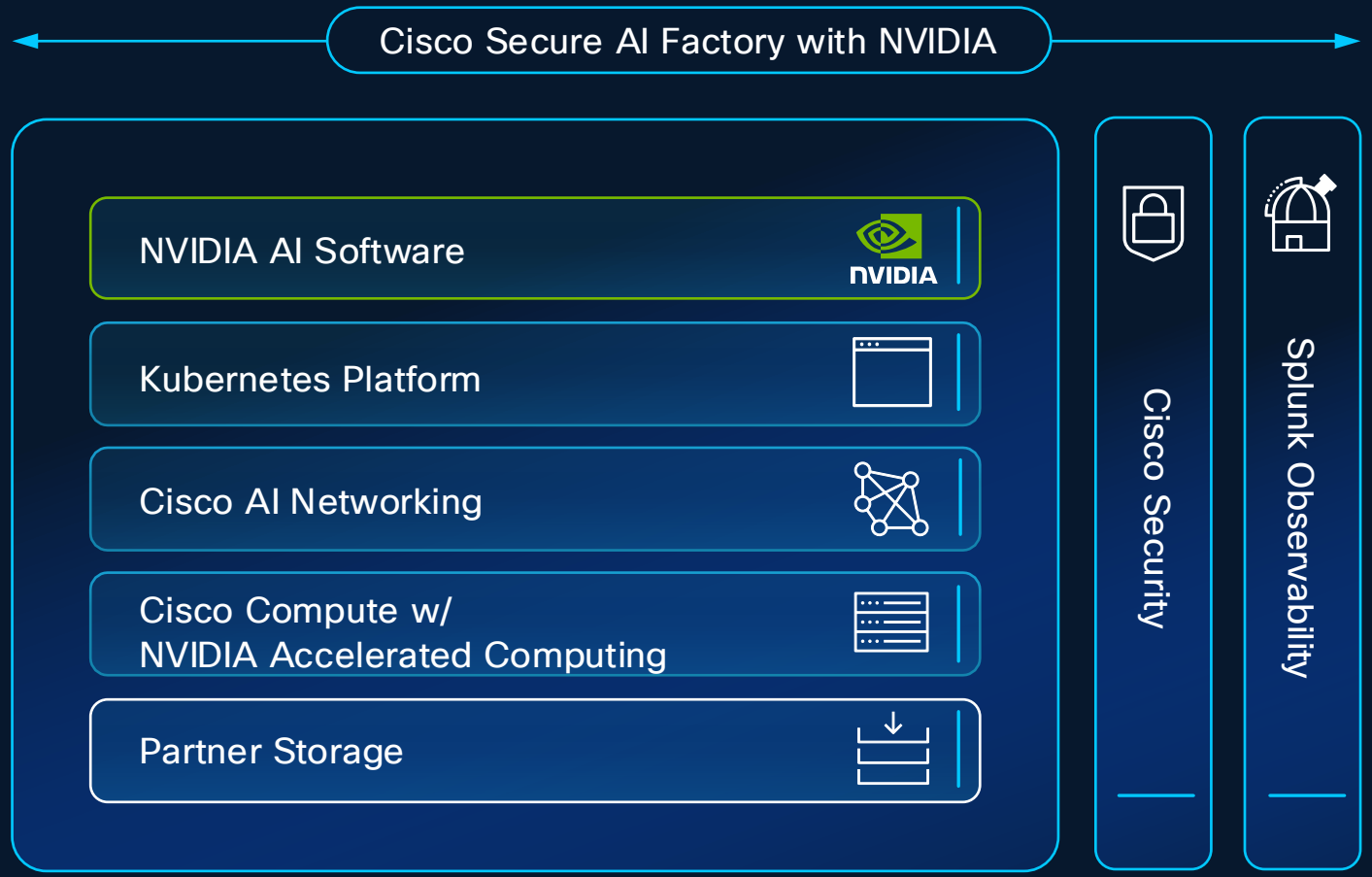
Accelerate delivery of trusted, transformative AI applications



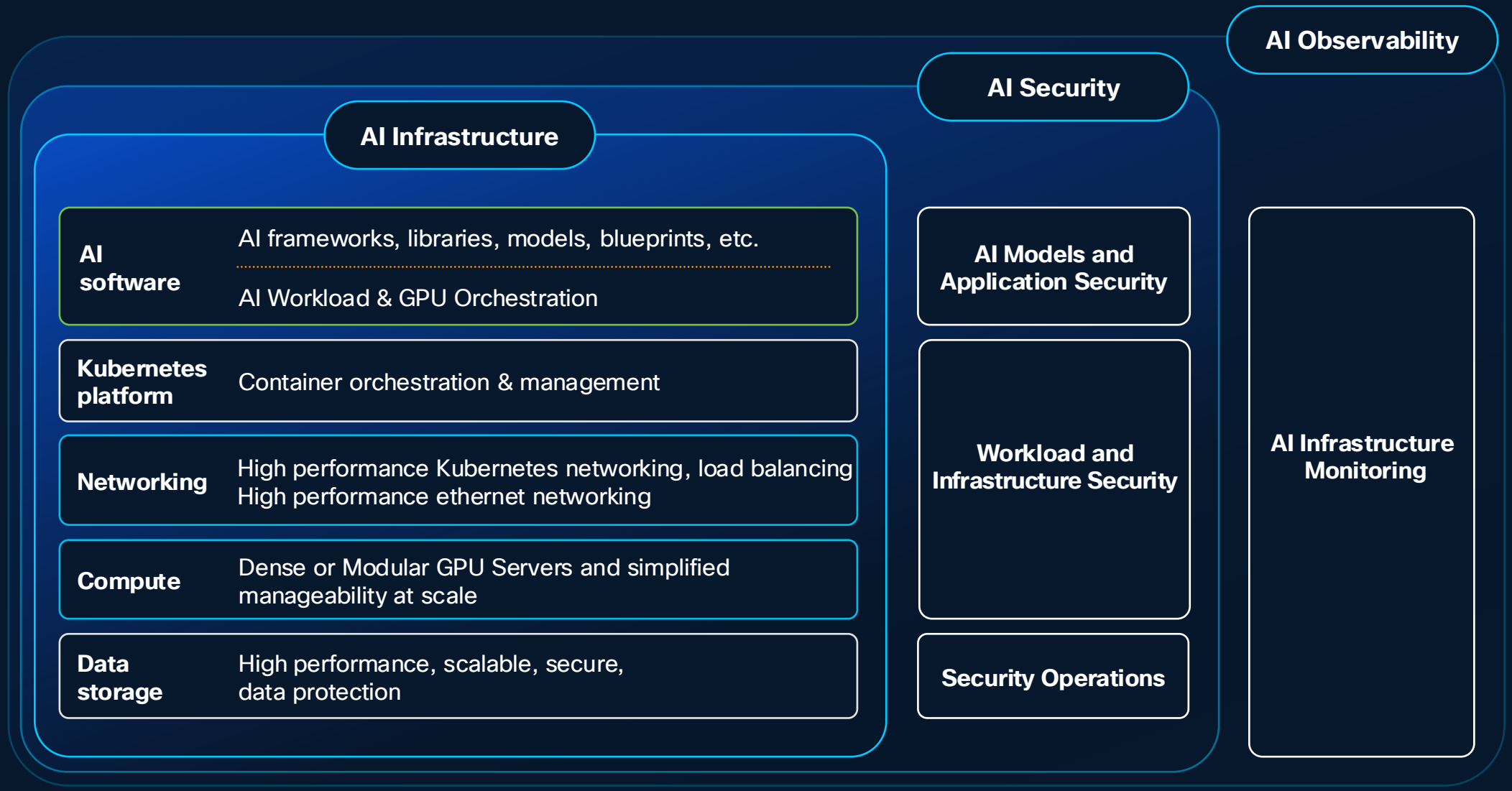
Cisco Secure AI Factory with NVIDIA

What is it?

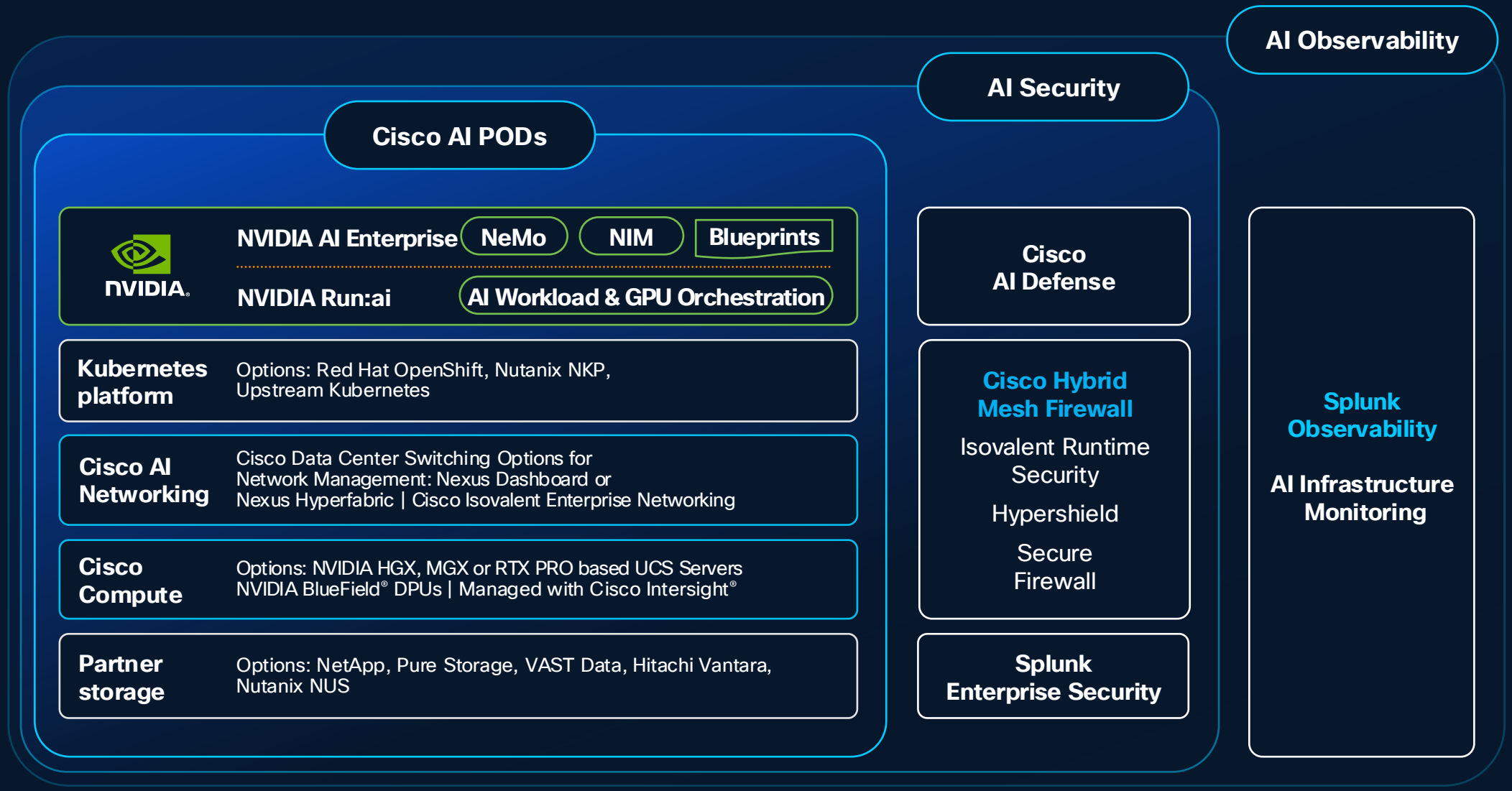
A modular reference design that combines high-performance infrastructure with full-stack security and observability



Key capabilities of Cisco Secure AI Factory with NVIDIA



Key products in Cisco Secure AI Factory with NVIDIA



Security-first architecture enables safe Enterprise AI

Security at all layers of the stack

Securing the Applications

Cisco AI Defense: Testing and runtime security of LLMs and GenAI applications, integrated with NVIDIA AI.

Securing the Workloads and Infrastructure

Cisco Hybrid Mesh Firewall: Unified management, consistent, pervasive policies.

Cisco Isovalent: Enhanced visibility into cloud native interactions, consistent policy definition and enforcement.

Cisco Hypershield: Protection against lateral movement, proactive vulnerability mitigation.

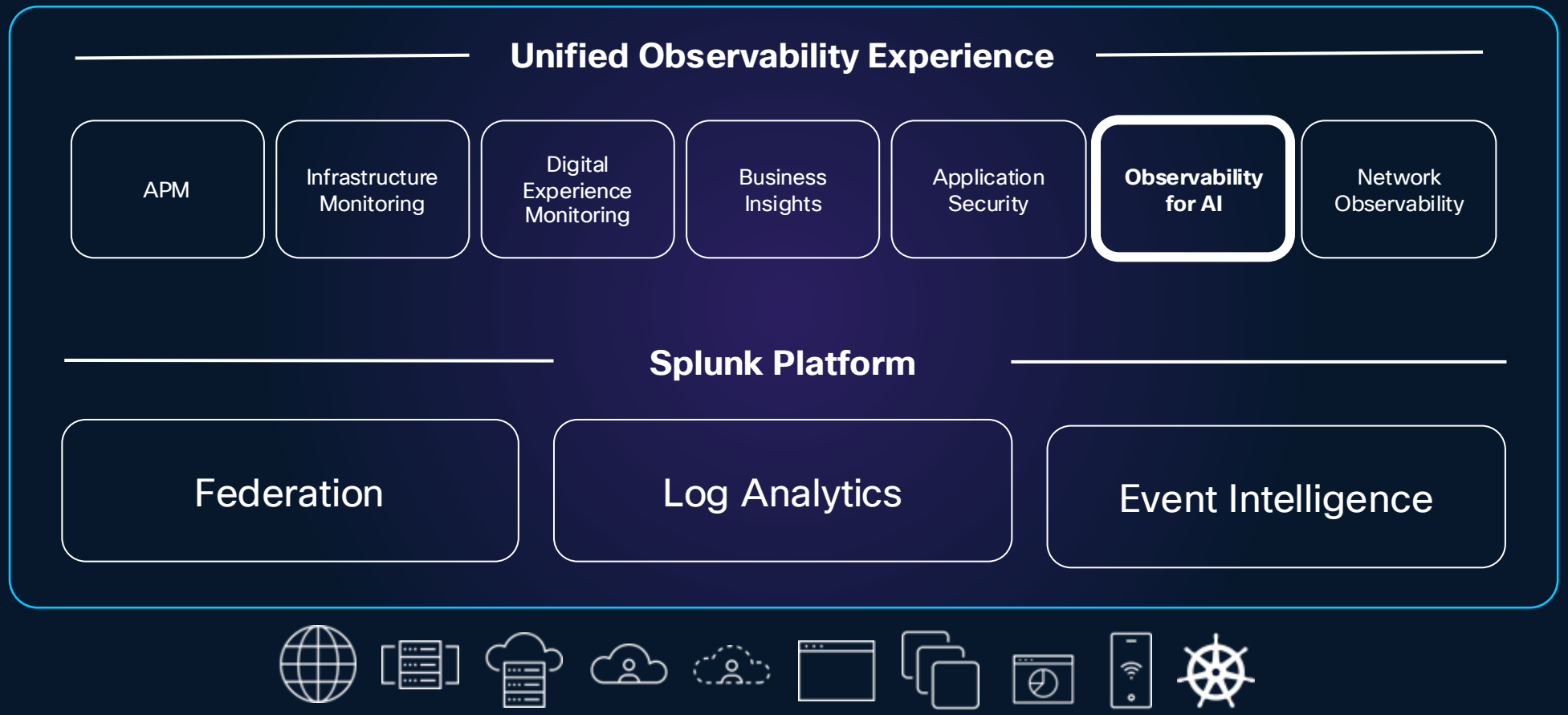
Cisco Secure Firewall: Threat protection at scale without compromising performance.

Security Operations

Splunk Enterprise Security: Real-time threat detection, investigation, and response through analytics, automation, insights.

Splunk Observability

Building a leading observability practice in the AI era



Cisco AI POD Dashboard Views Using AI Infrastructure Monitoring

Automatic attribution to instrumented services, customized data slicing, and platform-dependent attribution models

Additional Dashboard Views

Intersight

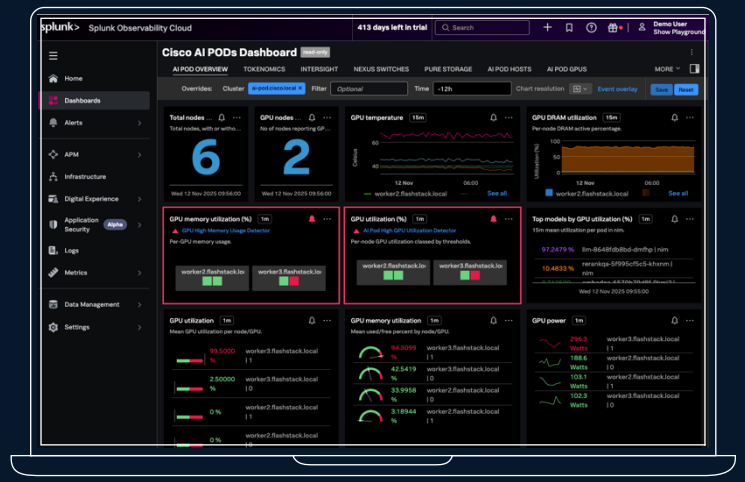
Nexus Switches

Storage

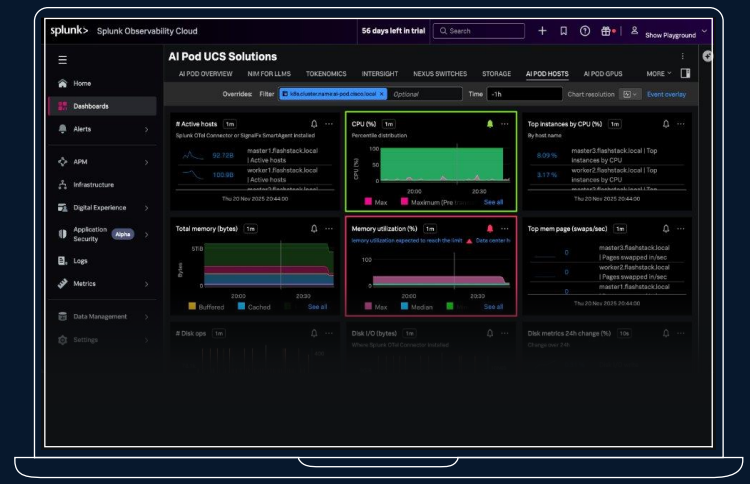
NIM for LLMs

Clusters

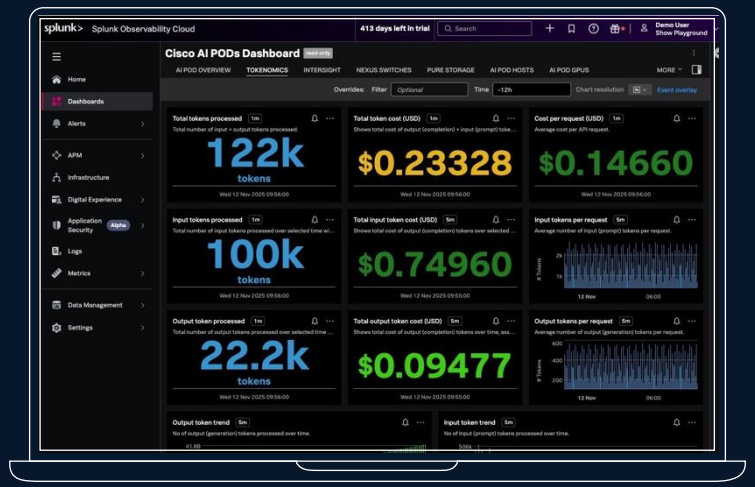
LLM Model Costs



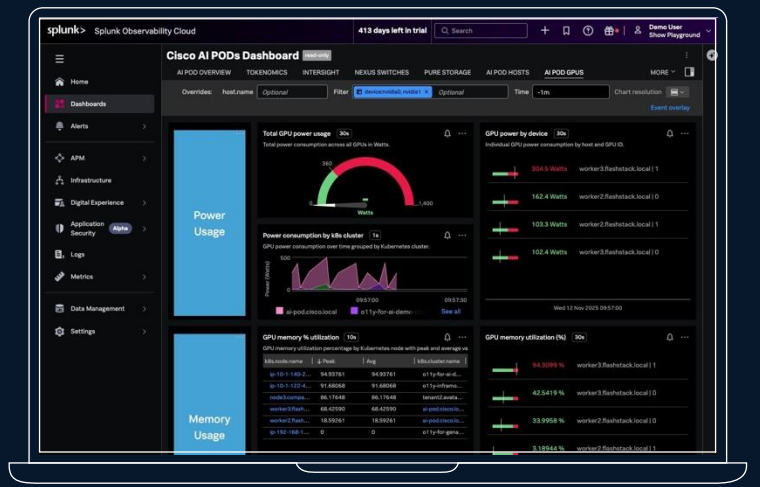
AI POD Overview - Total nodes, GPU power and utilization (%), etc...



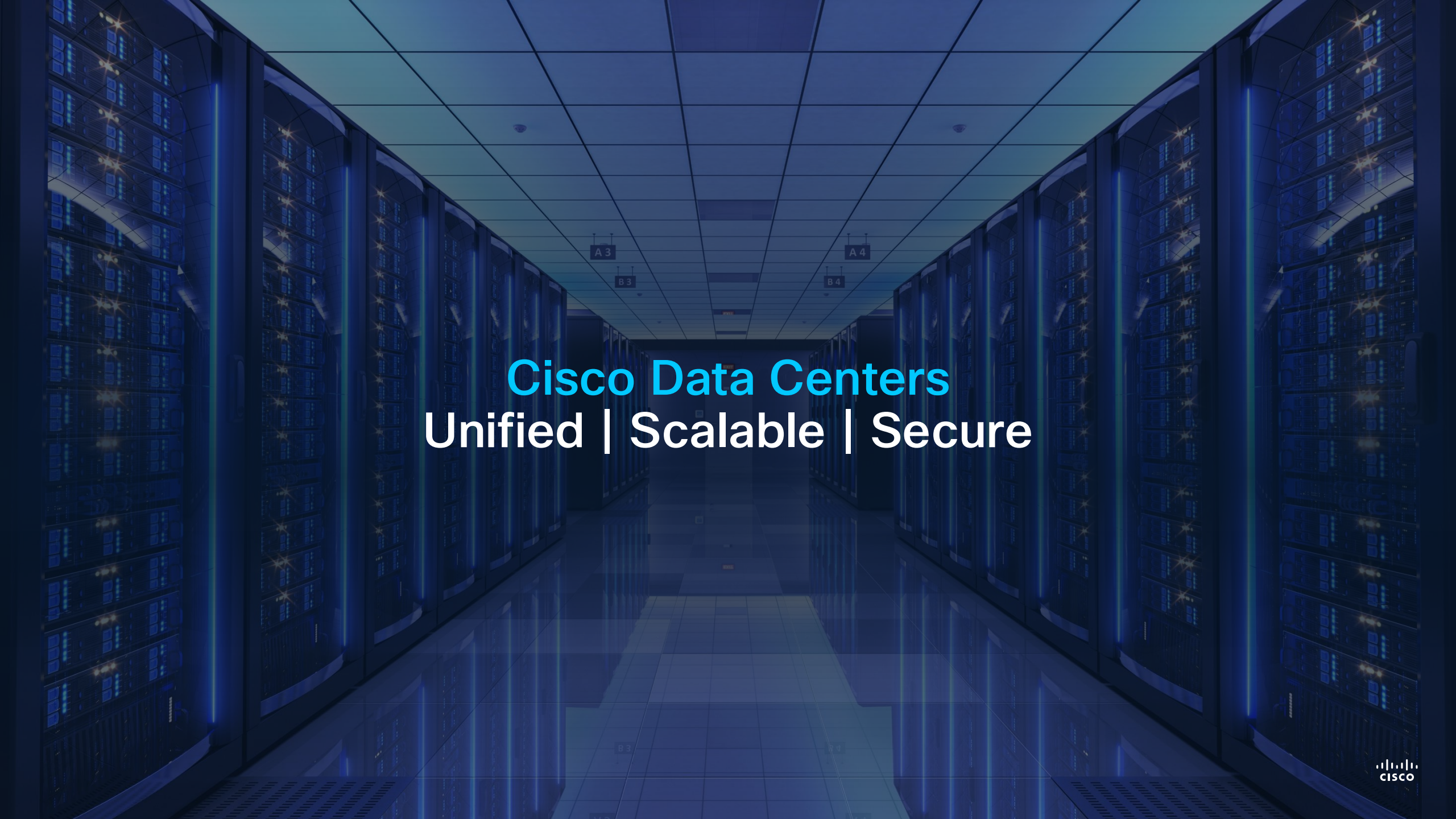
AI POD Hosts - # of active hosts, CPU (%), memory utilization (%), etc...



Tokenomics - Total tokens processed, Total token cost, etc...



AI POD GPUs- GPU power usage, GPU memory utilization (%), etc...


Cisco Data Centers
Unified | Scalable | Secure

Infrastructure to power AI

Services to accelerate the value of AI

Security for AI, AI for security

Data to drive insights and context

Software to unlock productivity

Cisco is bringing these together to make your enterprise AI journey easier

Resources to learn more

Cisco Compute

View on cisco.com/go/ucs

AI-Ready Infrastructure

View on cisco.com

Isovalent Enterprise Platform

View on Isovalent.com (now part of Cisco)

Cisco Compute YouTube channel

Visit youtube.com

Blogs

Visit blogs.cisco.com/datacenter

Online community

Visit Data Center and Cloud online community

Thank you