

Secure AI Factory Compute

Cisco UCS GPU-Optimized Servers and AI PODs
for Scalable, Secure AI Infrastructure

Justin Blinn
Solutions Engineer

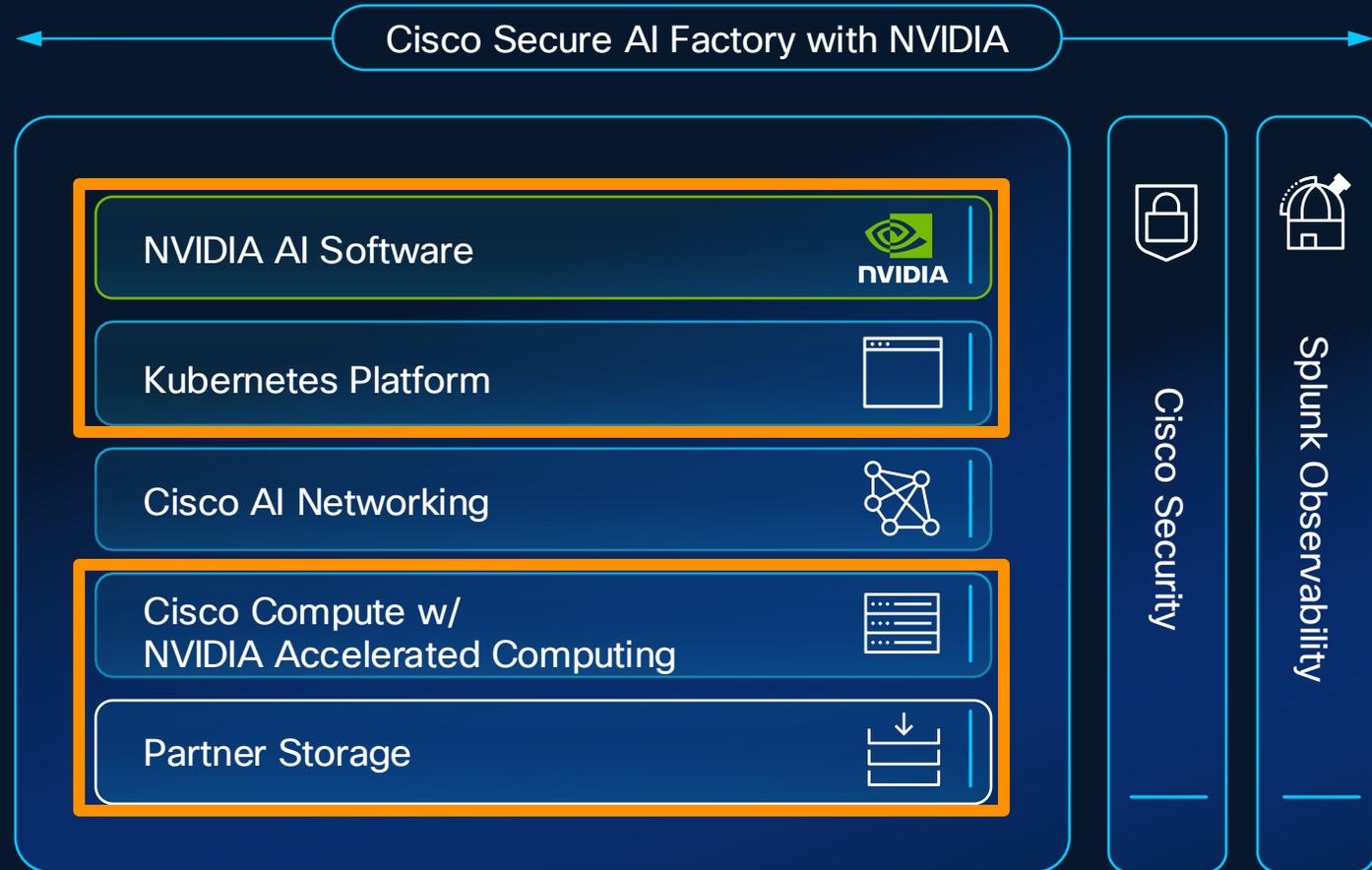
Andrew MacDonell
Solutions Engineer



Cisco Secure AI Factory with NVIDIA

What is it?

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

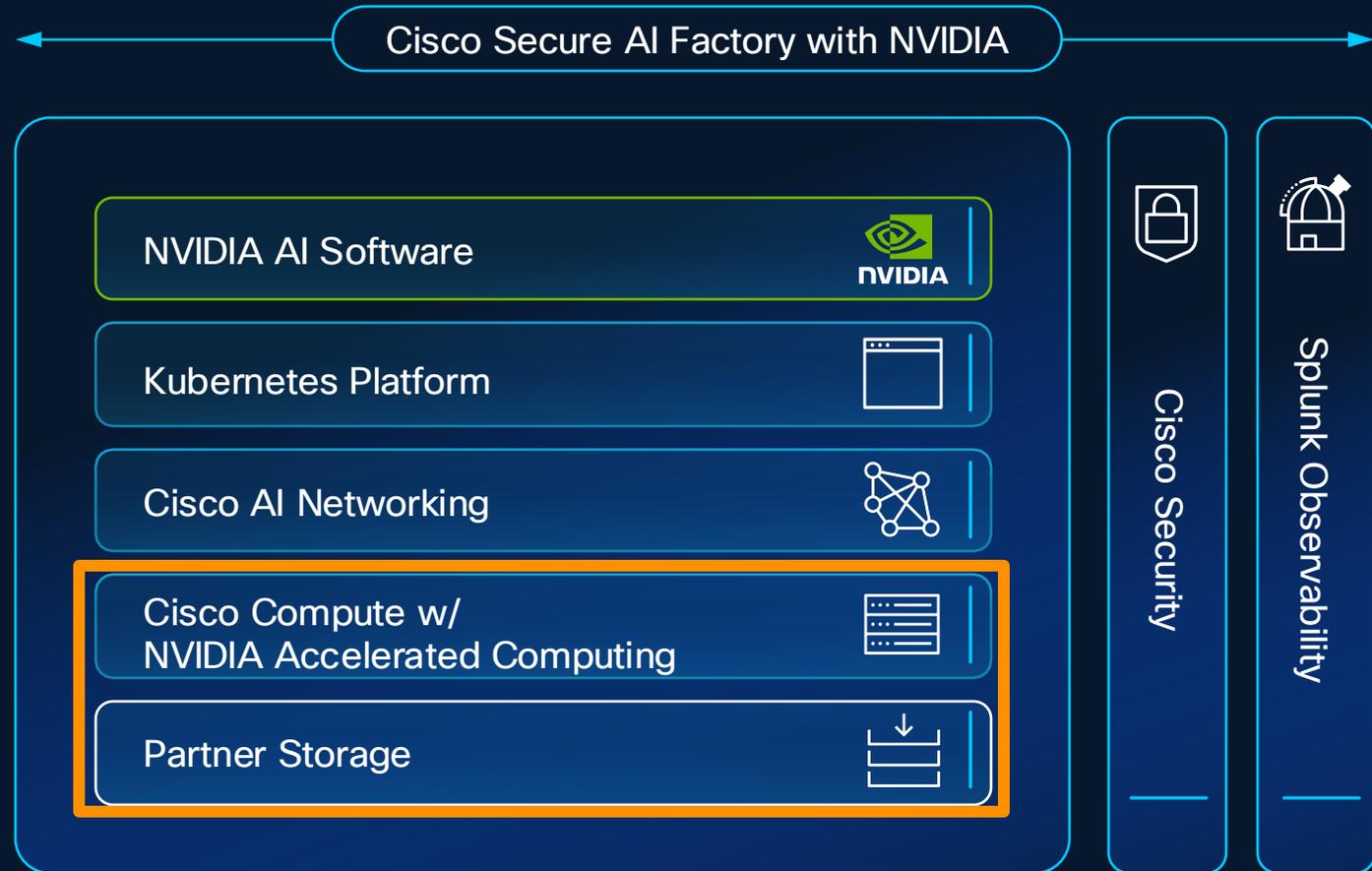


Compute and Storage

Cisco Secure AI Factory with NVIDIA

What is it?

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





HITACHI

NUTANIX



Qumulo

IMPORTANT

SAIF is intentionally storage-agnostic but vendor-validated

Customers choose the platform that best fits their workload and regulatory environment

This partner flexibility is a major competitive advantage compared to monolithic cloud stacks

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

UCS – AI Use Case Focused Servers



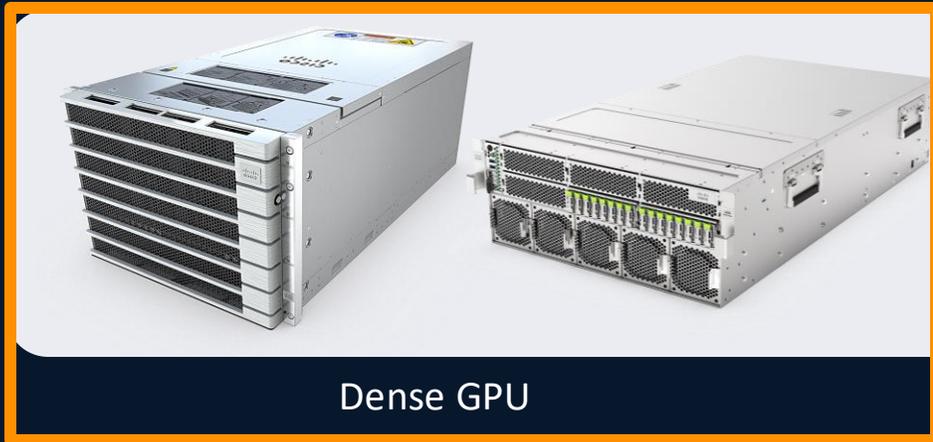
CISCO INTERSIGHT®

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

High-Density Blackwell GPU Server

Built for LLM training, deep learning, fine-tuning, and HPC



UCS Accelerated | UCS C880A M8

NEW

AVAILABLE NOW



2 CPUs

Intel Xeon 6th Gen Scalable Processor

NVIDIA HGX with 8 GPUs

NVIDIA B300 with NVL8 Air Cooled

Network

(8) NVIDIA ConnectX-8 GPU Board
Integrated (E-W)

(2) NVIDIA BF3 B3220, NVIDIA BF3240,
NVIDIA ConnectX-7 (N-S)

Power

(12) 50V 3200W (N+N redundancy)

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

BlueField-3 or ConnectX-7

Flexible, Modular AI Servers

“Start small and scale up” with AI

UCS accelerated | Cisco UCS C845A



NVIDIA MGX™ reference design

With NVIDIA RTX 6000,
H100, H200, L40S GPUs

BlueField-3 or ConnectX-7

High performance in a compact form factor

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

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

UCSC (Rack) – C240/245



**NVIDIA PCIe
reference design**

With NVIDIA H100, H200,
L40S GPUs

*RTX 6000 (throttled)



**NVIDIA PCIe
reference design**

With NVIDIA H100, L40S
GPUs

*RTX 6000 (throttled)

Available Now

UCSX - High-Density GPU Nodes

AI Enable Your Blade Infrastructure

UCS accelerated | Cisco UCX X580p & UCS 9516 X-Fabric



NVIDIA HGX™ reference design

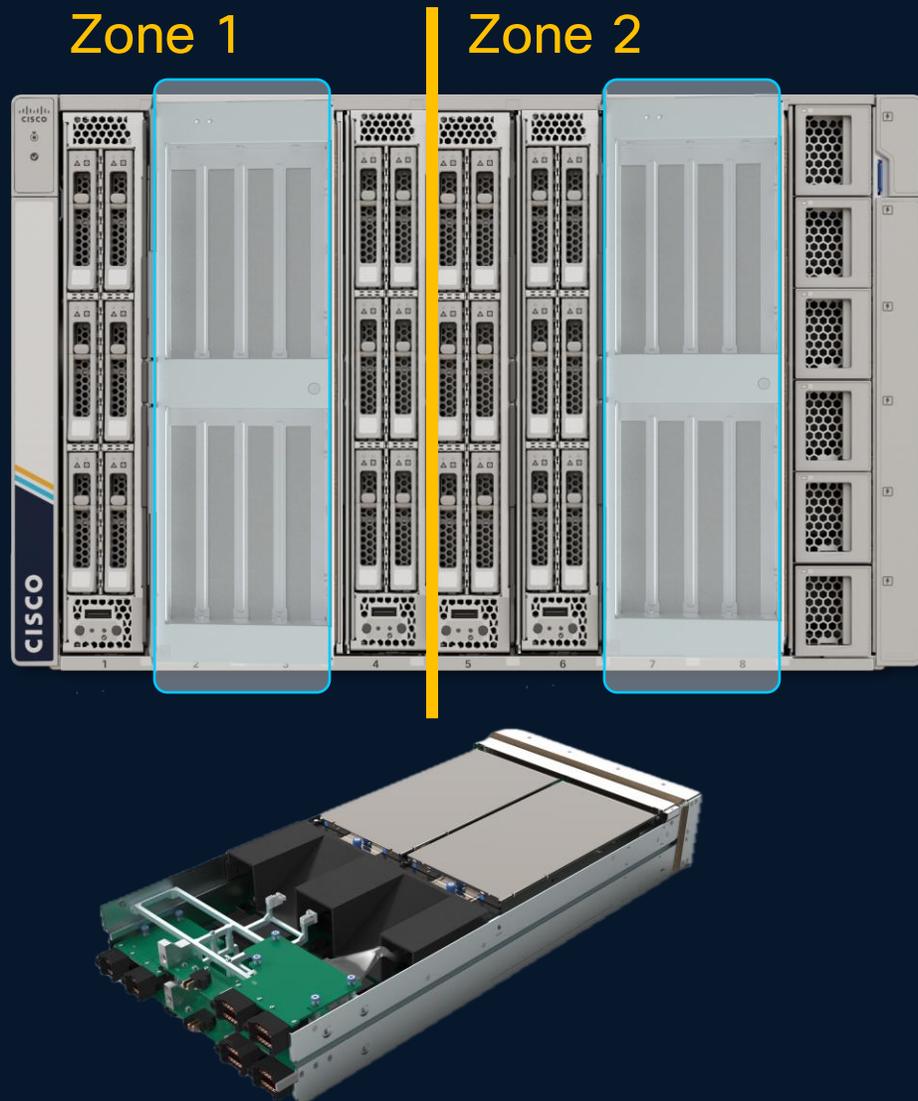
Double wide PCIe node for 4x
FHFL GPU
and PCIe G5 GPU support

NVLink bridge support

ConnectX-7 1x 400GB or 2x200GB

Full 600W FHFL GPU Powering
RTX 6000, H200, L40S

UCSX - GPU Sharing Within Chassis



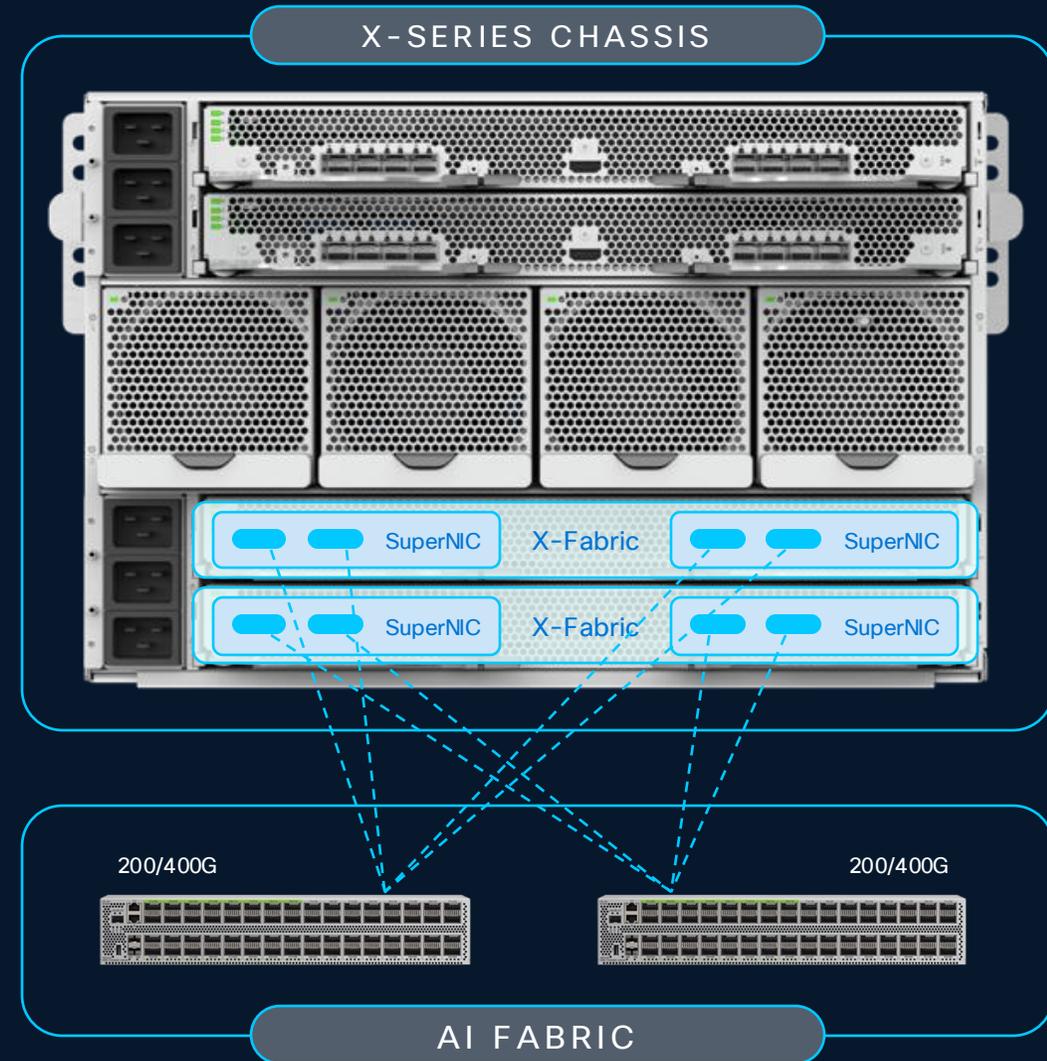
- Double wide PCIe node for 4x FHFL GPU and PCIe G5 GPU support
 - Nvidia H200, RTX PRO 6000 & L40S
- Support multiple vendors: Nvidia, AMD*/Intel*
- NVLink bridge support
- Support up to 600W FHFL GPU
- Managed PCIe node with BMC support
- Policy based GPU management
- Ability to share GPUs across two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

UCSX - GPU Sharing Across Fabric

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



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

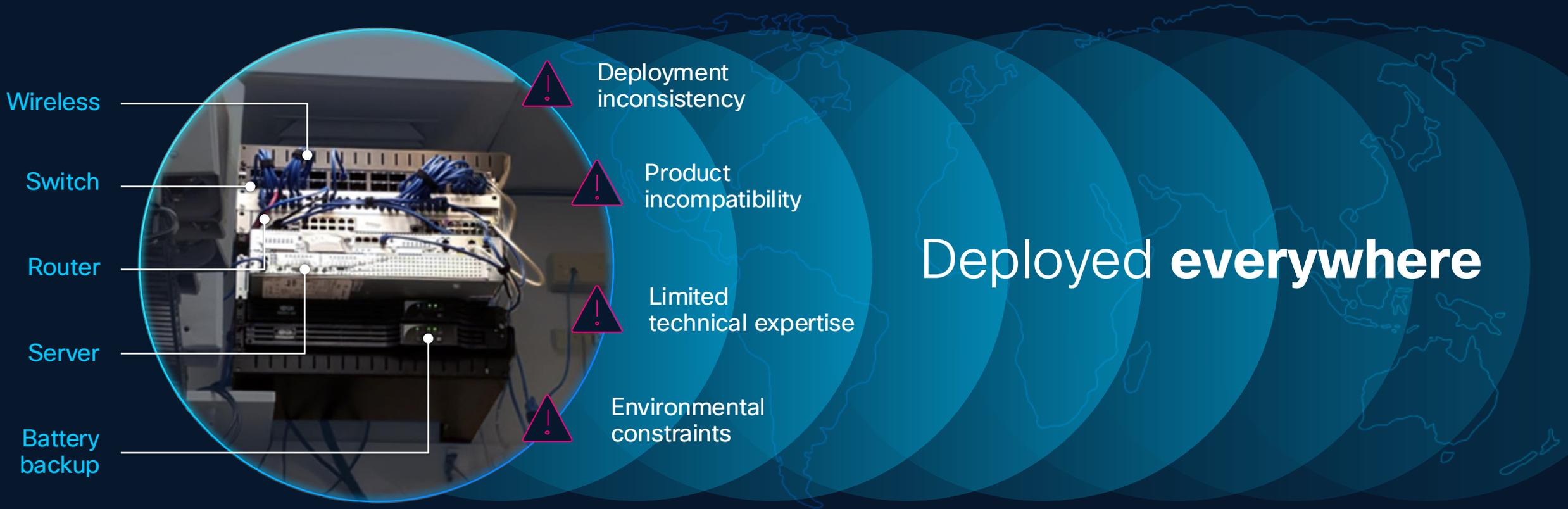
Modular (w/GPU Expansion) and Rack

Unified Edge

Demanding AI

Mainstream and Edge AI

Legacy Edge Infrastructure



Operational complexity



Security risks

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



vmware
by Broadcom



Microsoft

intel



SUSE



Intersight – Fleet Deploy, Operations and Support

Centralized Management
Global Policies

Intuitive Experience



Enhanced Support



Proactive Guidance



Secure and Extensible



SaaS Delivered



Comprehensive Automation
Single Pane of Glass



SaaS Consumption Model
Free customers from care and feeding of management tools and eliminate upgrade dependencies



Seamless Extensibility
Simplify management across technologies and geography

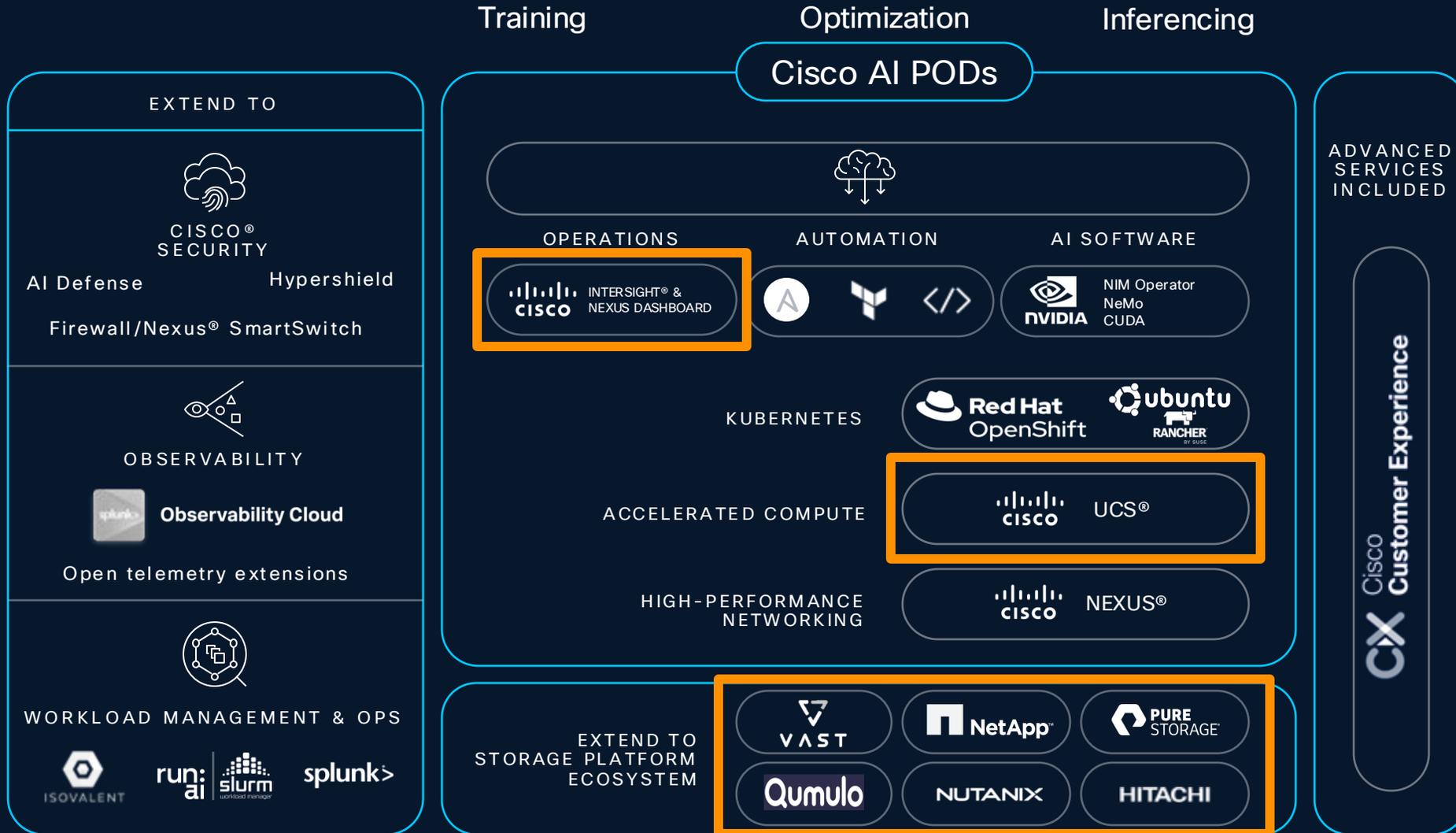


Continuous Feature Integration
Rapid development, delivery and customer feedback

Cisco AI PODs

Introducing AI POD “Integrated Offerings”

BYO AI tools:



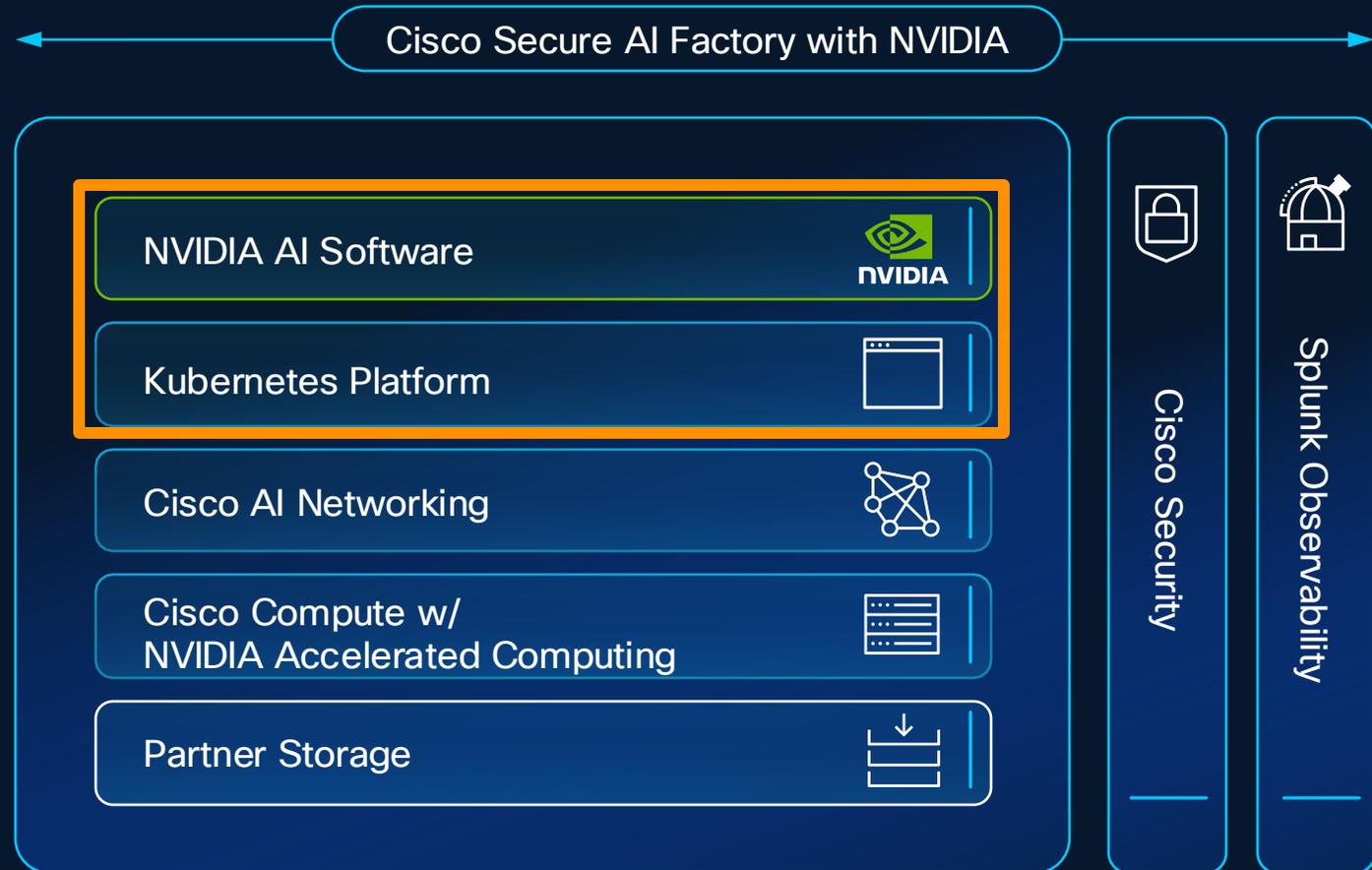
- RAFAY
- Kubeflow
- jupyter
- Apache Airflow
- Weights & Biases
- mlflow
- neptune.ai
- kedro
- comet
- ZenML
- CLEARML
- PREFECT
- Flyte
- mongoDB

Containers and AI Software

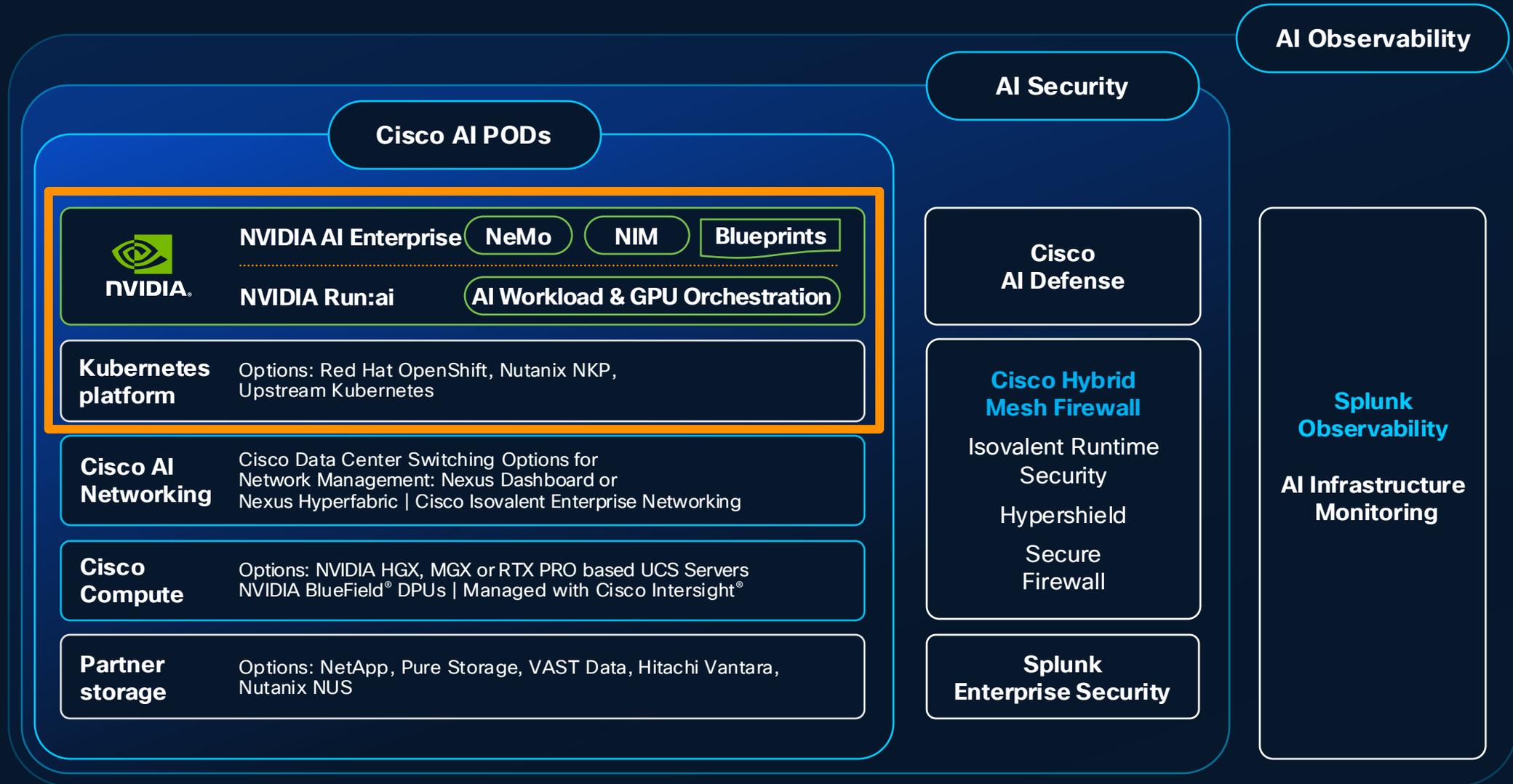
Cisco Secure AI Factory with NVIDIA

What is it?

- AI Software Stack
- OS & Kubernetes Platform



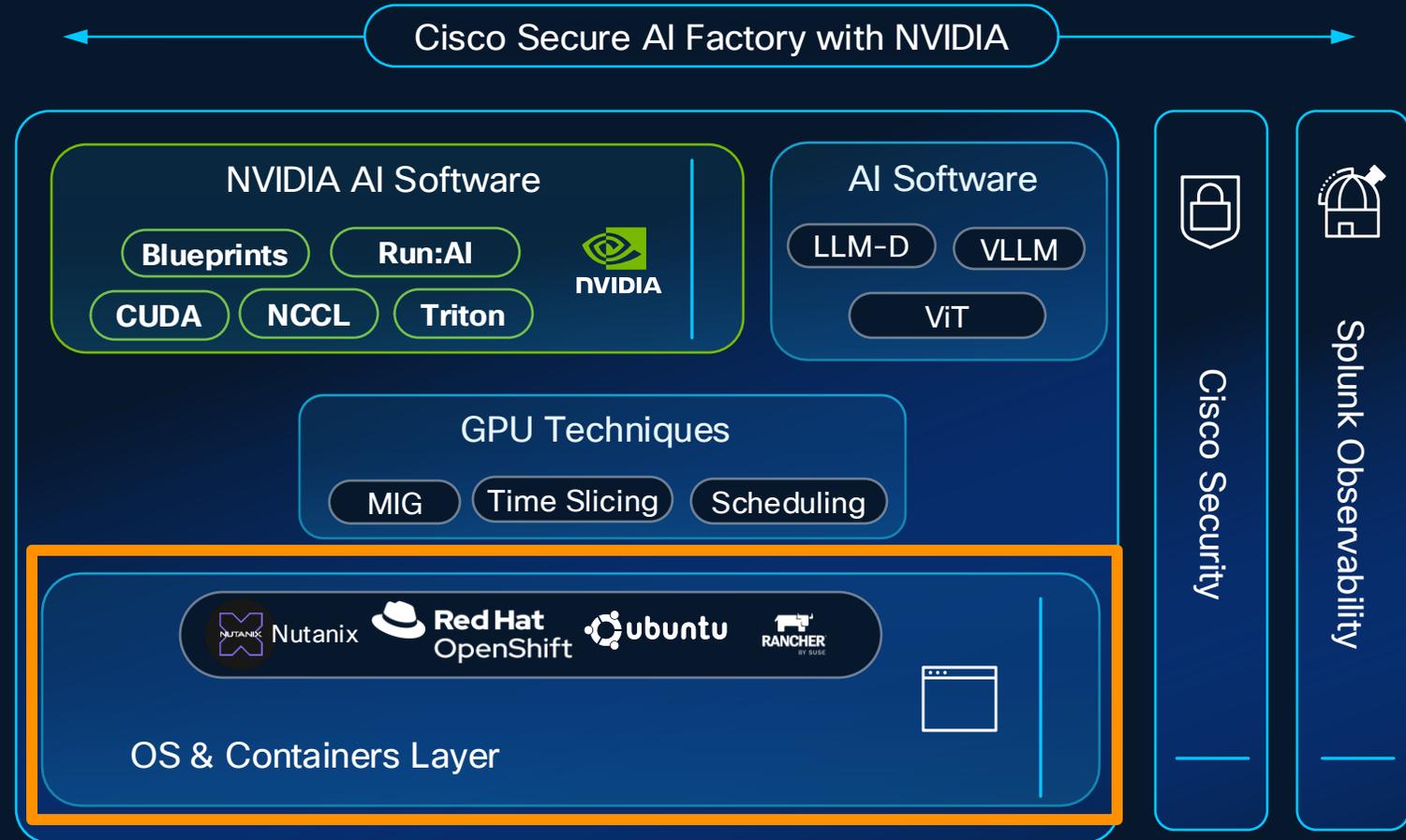
Key Elements in Cisco Secure AI Factory with NVIDIA



Cisco Secure AI Factory with NVIDIA

What is it?

- AI Software Stack
- OS & Kubernetes Platform

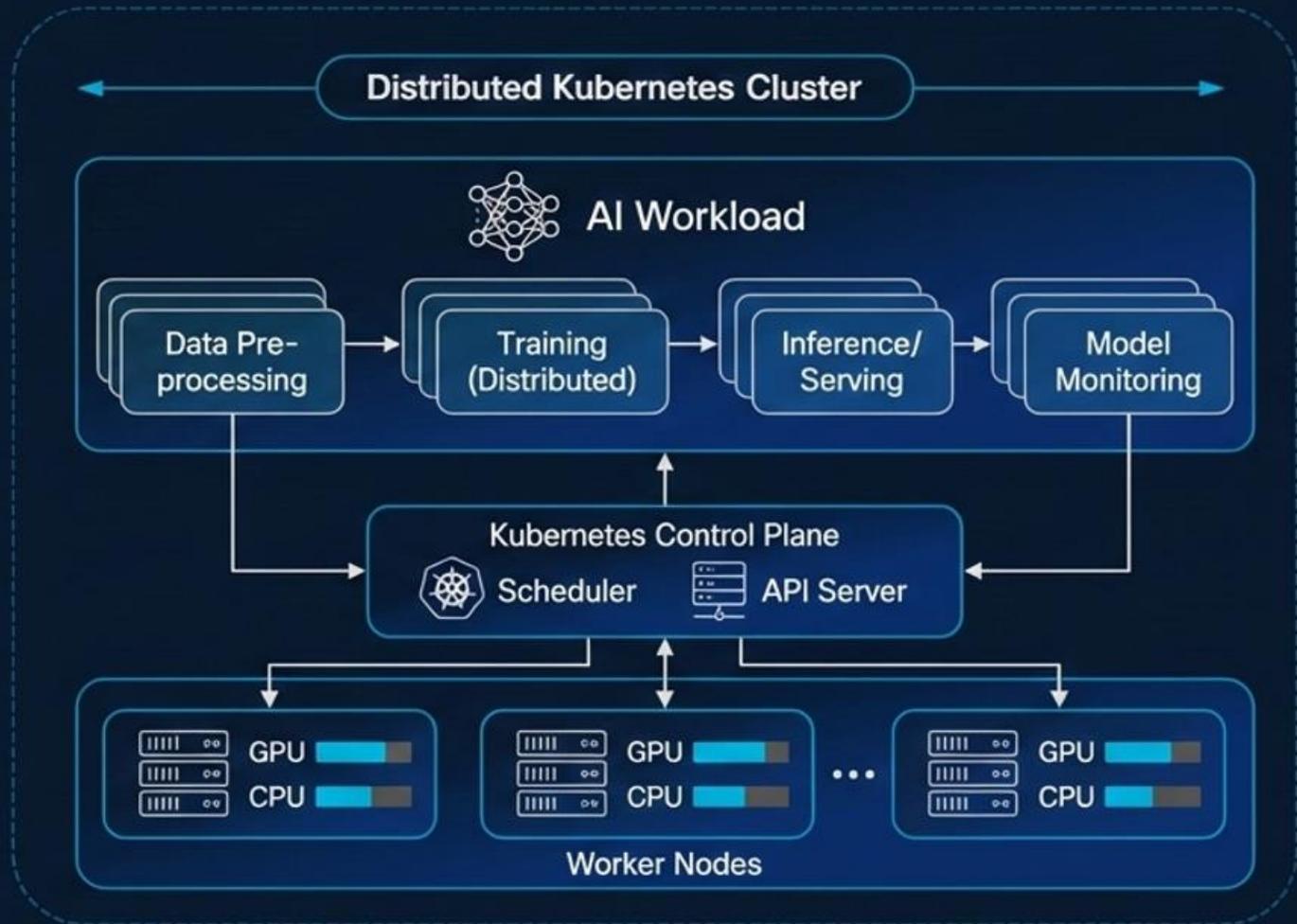


Distributed & Containerized Design for AI Workloads

Why it Matters for AI

Key Advantages

- Modularity & Portability
- Scalability (horizontal)
- Resource Utilization (GPUs/CPU)
- Fault Tolerance & Resilience
- Accelerated Experimentation



The Future of Cisco and Isovalent

Workload Portability

Modern Applications

Virtualization

AI

Cisco Security & Observability

Application

Virtualization

Modern Apps

AI/ML

Kubernetes

L7 Service Mesh



Virtual
Networking



Firewall /
Microsegmentation



Load
Balancing



Network
Visibility



VPN/
Encryption



Runtime
Security

Enterprise Platform

the creators of



Cisco Networking

Nexus

Nexus Dashboard

Cisco Compute

UCS

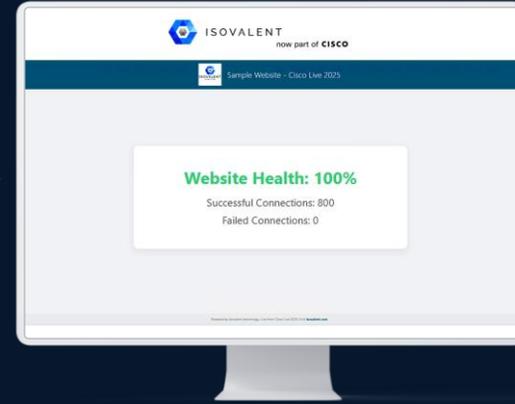
Intersight

Partner Storage

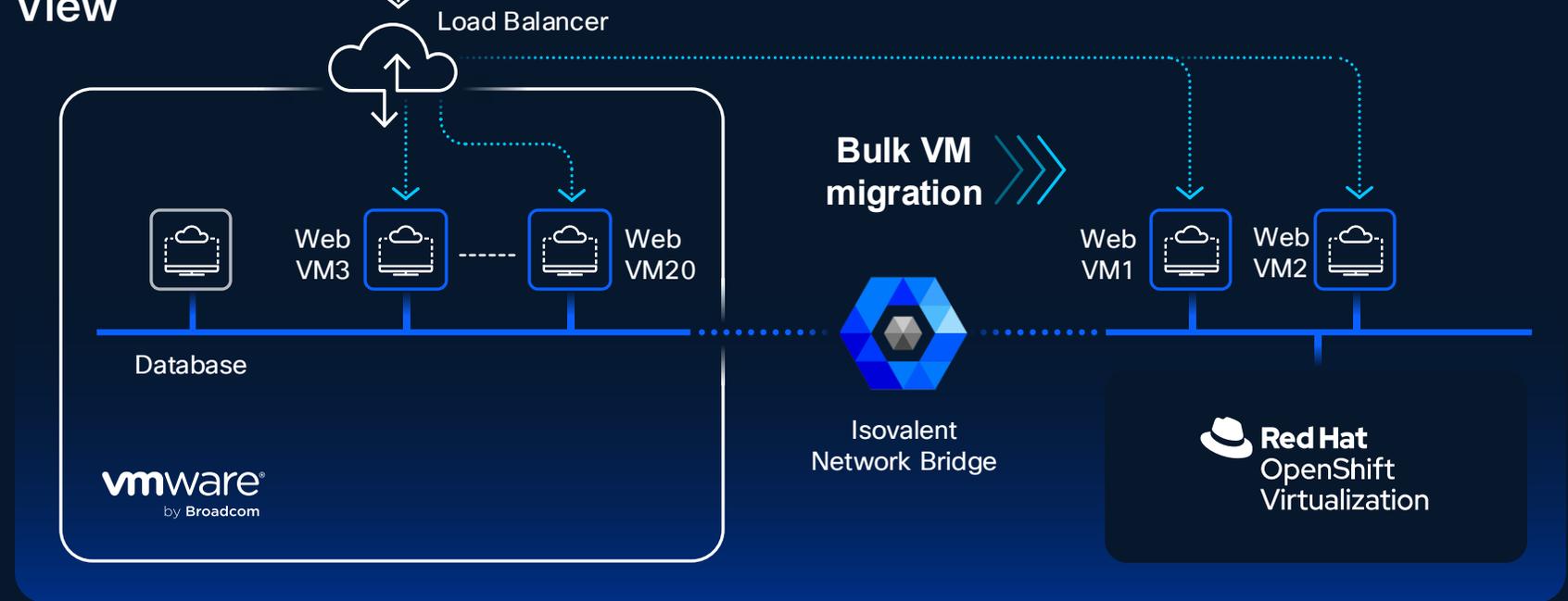
DEMO

VM Workload Portability Before Migration

Application User Experience



Infrastructure View





Navigation pane showing folder structure:

- vcasa-388579.af41878c.us-central1.gve.g...
- Datacenter
 - cisco-xldxx
 - demo-actors**
 - actors
 - db
 - Discovered virtual machine
 - HCX Management VMs
 - Lab infrastructure
 - Templates
 - TGW
 - vCLS
 - Workload VMs

demo-actors | ACTIONS

Summary Monitor Configure Permissions **VMs** Updates

Virtual Machines | VM Templates | vApps | VM Folders

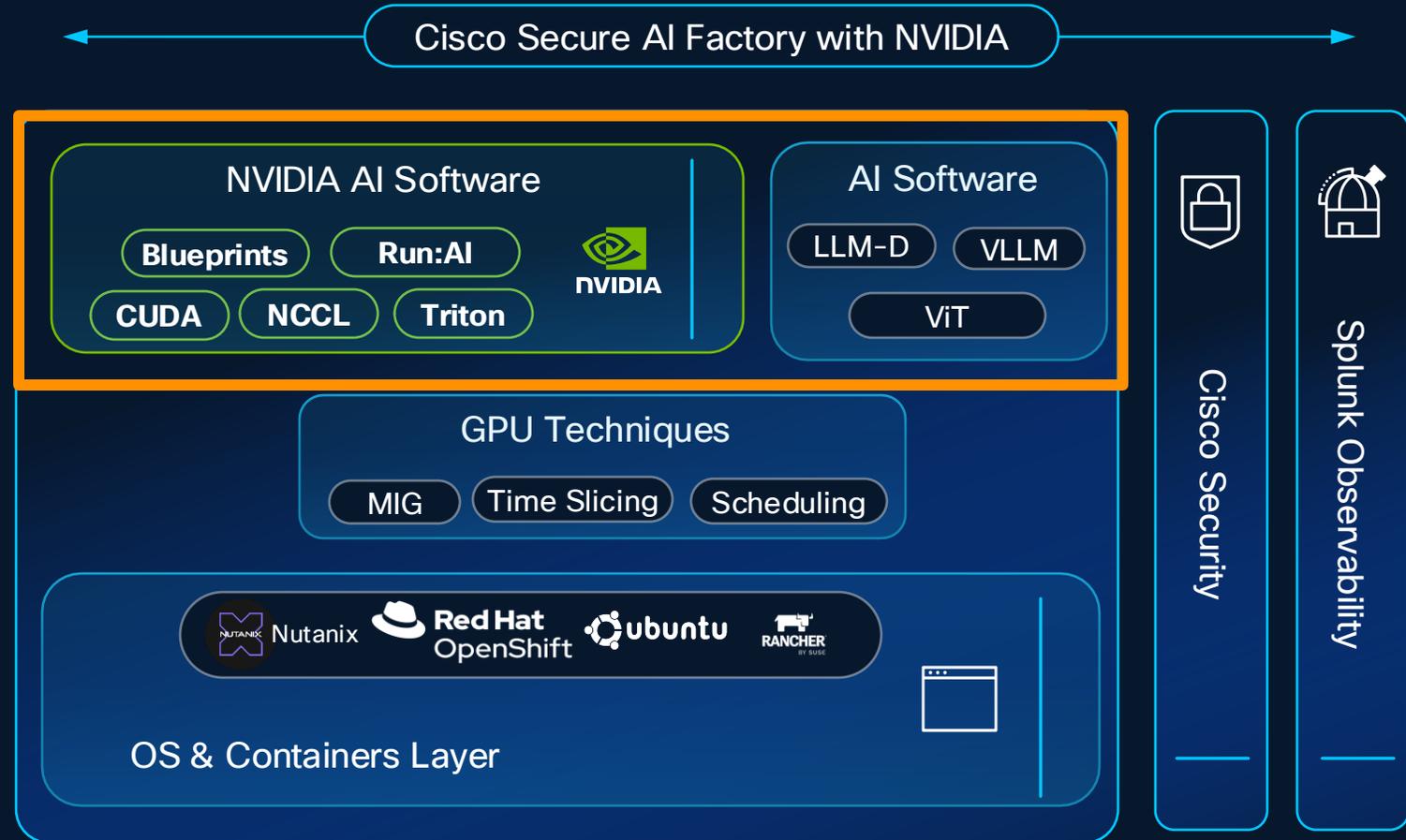
<input type="checkbox"/>	Name	↑	State	Status	Provisioned Space	Used Space	Host CPU	Host Mem
<input type="checkbox"/>	actor-001		Powered ...	✓ Normal	21.39 GB	7.77 GB	0 Hz	0 B
<input type="checkbox"/>	actor-002		Powered ...	✓ Normal	21.38 GB	9.5 GB	51 MHz	514 MB
<input type="checkbox"/>	actor-003		Powered ...	✓ Normal	21.4 GB	9.95 GB	0 Hz	0 B
<input type="checkbox"/>	actor-004		Powered ...	✓ Normal	21.38 GB	9.93 GB	0 Hz	0 B
<input type="checkbox"/>	actor-005		Powered ...	✓ Normal	21.43 GB	9.48 GB	51 MHz	502 MB
<input type="checkbox"/>	actor-006		Powered ...	✓ Normal	21.42 GB	10.41 GB	0 Hz	0 B
<input type="checkbox"/>	actor-007		Powered ...	✓ Normal	21.39 GB	8.98 GB	0 Hz	0 B
<input type="checkbox"/>	actor-008		Powered ...	✓ Normal	21.45 GB	9.52 GB	0 Hz	0 B
<input type="checkbox"/>	actor-009		Powered ...	✓ Normal	21.42 GB	10.42 GB	0 Hz	0 B
<input type="checkbox"/>	actor-010		Powered ...	✓ Normal	21.4 GB	9.27 GB	0 Hz	0 B
<input type="checkbox"/>	actor-011		Powered ...	✓ Normal	21.4 GB	9.66 GB	51 MHz	520 MB
<input type="checkbox"/>	actor-012		Powered ...	✓ Normal	21.39 GB	9.69 GB	0 Hz	0 B
<input type="checkbox"/>	actor-013		Powered ...	✓ Normal	21.42 GB	9.46 GB	0 Hz	0 B
<input type="checkbox"/>	actor-014		Powered ...	✓ Normal	21.39 GB	9.33 GB	0 Hz	0 B
<input type="checkbox"/>	actor-015		Powered ...	✓ Normal	21.39 GB	9.5 GB	0 Hz	0 B
<input type="checkbox"/>	actor-016		Powered ...	✓ Normal	21.38 GB	9.8 GB	0 Hz	0 B
<input type="checkbox"/>	actor-017		Powered ...	✓ Normal	21.43 GB	9.87 GB	0 Hz	0 B
<input type="checkbox"/>	actor-018		Powered ...	✓ Normal	21.39 GB	8.89 GB	0 Hz	0 B
<input type="checkbox"/>	actor-019		Powered ...	✓ Normal	21.43 GB	9.88 GB	0 Hz	0 B
<input type="checkbox"/>	actor-020		Powered ...	✓ Normal	21.39 GB	10.24 GB	0 Hz	0 B

20 items | EXPORT

Cisco Secure AI Factory with NVIDIA

What is it?

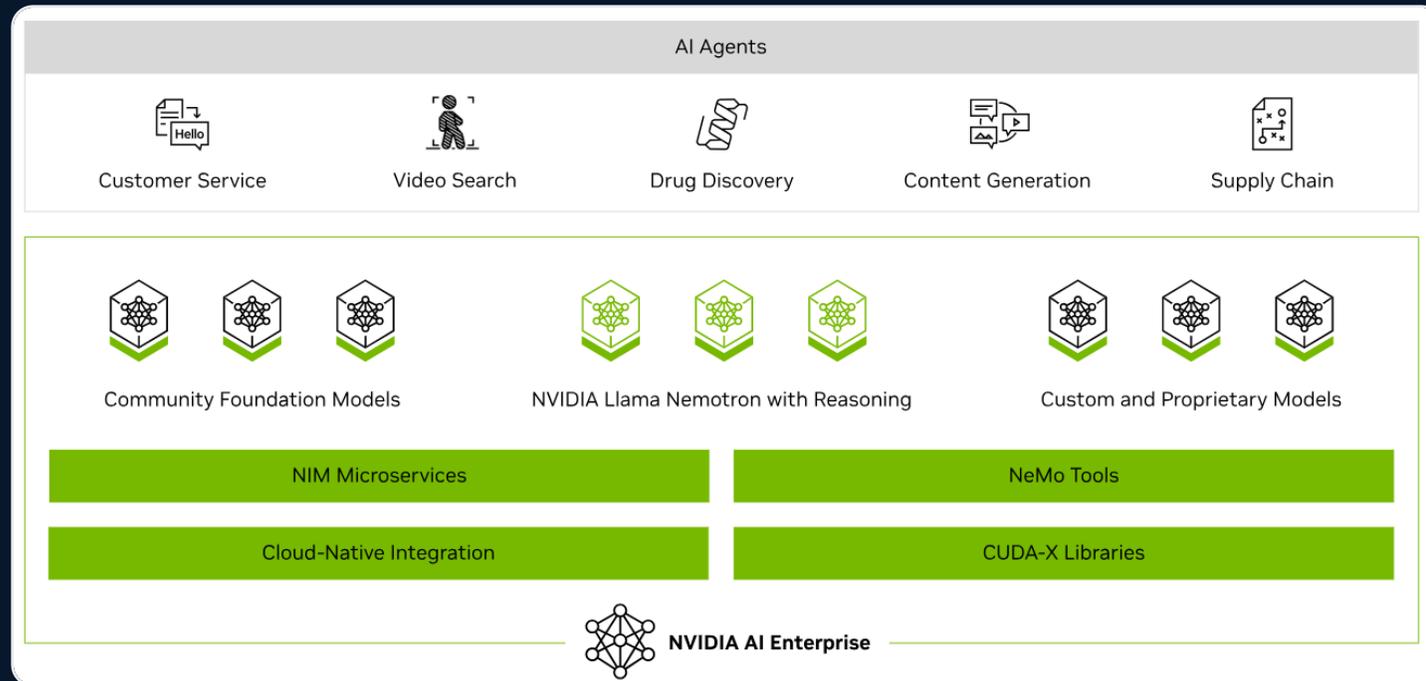
- AI Software Stack
- OS & Kubernetes Platform



NVIDIA Enterprise Software Included on Cisco AI-PODs

The NVIDIA Enterprise tools in the Cisco Secure AI Factory with NVIDIA provide support for each step in the training, optimization, and deployment of AI agents.

Production-ready software for agentic AI



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.

Software
for AI



NVIDIA
Enterprise

NVIDIA
Run:ai

NeMo - Triton

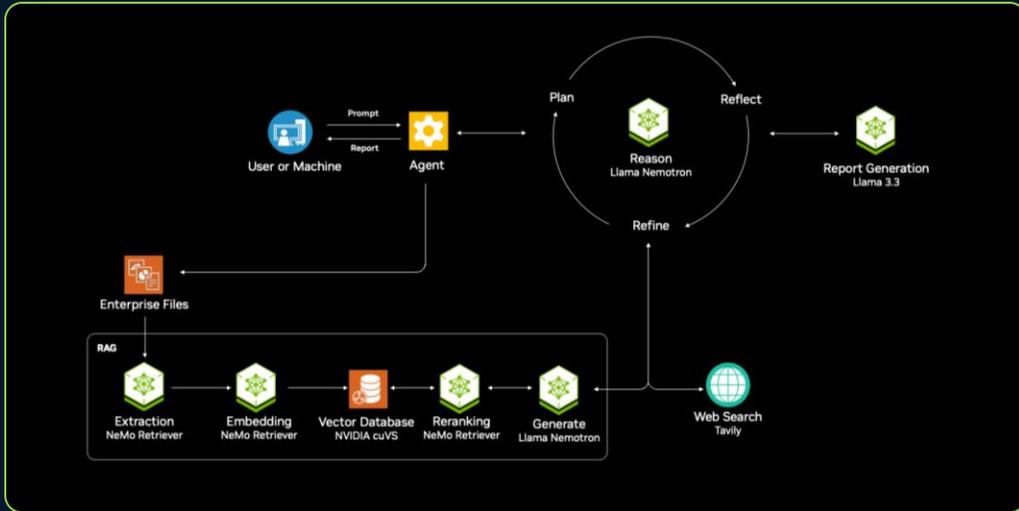
NIM

Blueprints

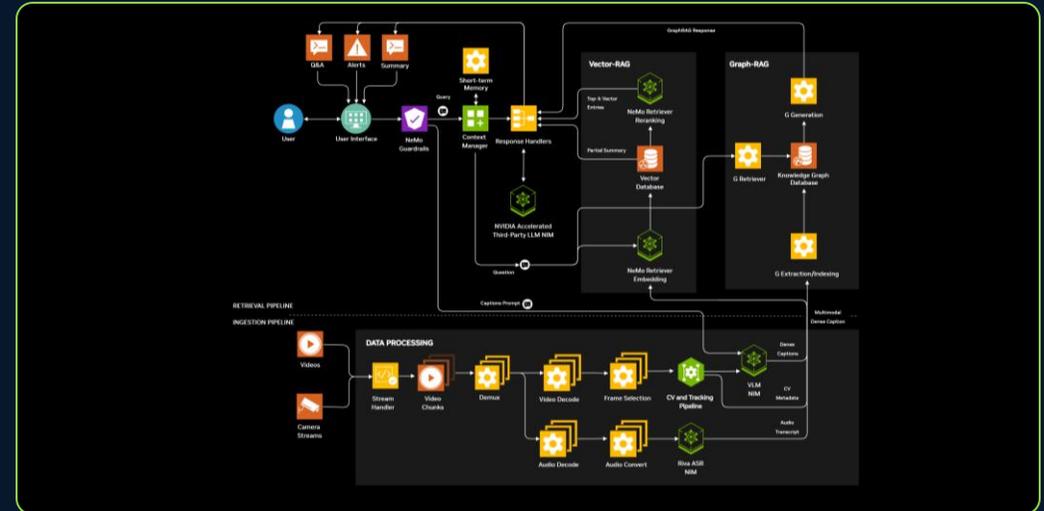
AI Workload & GPU Orchestration

NVIDIA Enterprise – Blueprints for Use Cases

Research Assistant



Video Search & Summarization



Blueprints offer sample workload designs for common AI use cases. These blueprints leverage technology available in the NVIDIA Enterprise software suite. These blueprints are but a few of infinite use cases that can be developed with AI software.

Software
for AI



NVIDIA
Enterprise

NVIDIA
Run:ai

NeMo

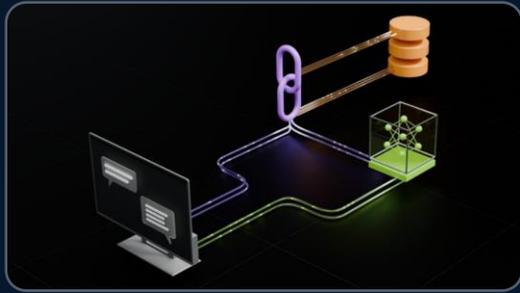
NIM

Blueprints

AI Workload & GPU Orchestration

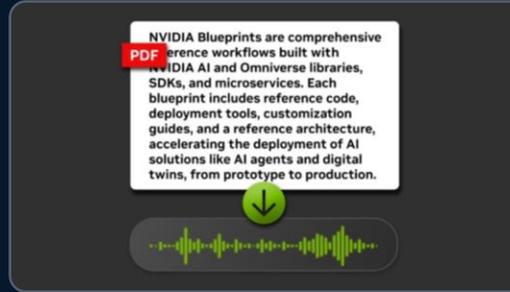
NVIDIA Blueprints—Cisco AI PODs Use Cases

Reference workflows for agentic AI, accelerated on Cisco AI-PODs



Enterprise RAG

Connect AI applications to multimodal enterprise data with a retrieval-augmented generation (RAG) pipeline built on NIM microservices for scalable data extraction and accurate information retrieval.



PDF to podcasts

Transform your PDF data into personalized audio content—including educational presentations, technical guides, and product documentation.



Video search and summarization

Build an agent that processes massive volumes of live or archived videos and extract insights for summarization and interactive Q&A.



Digital human

Create intelligent, interactive avatars for customer service across industries with NVIDIA AI Blueprint for digital humans to enhance customer service.



Security vulnerability analysis

Using NVIDIA NIM, NVIDIA NeMo Retriever, and NVIDIA Morpheus, this event-driven RAG application dramatically decreases CVE analysis and remediation time from days to seconds.



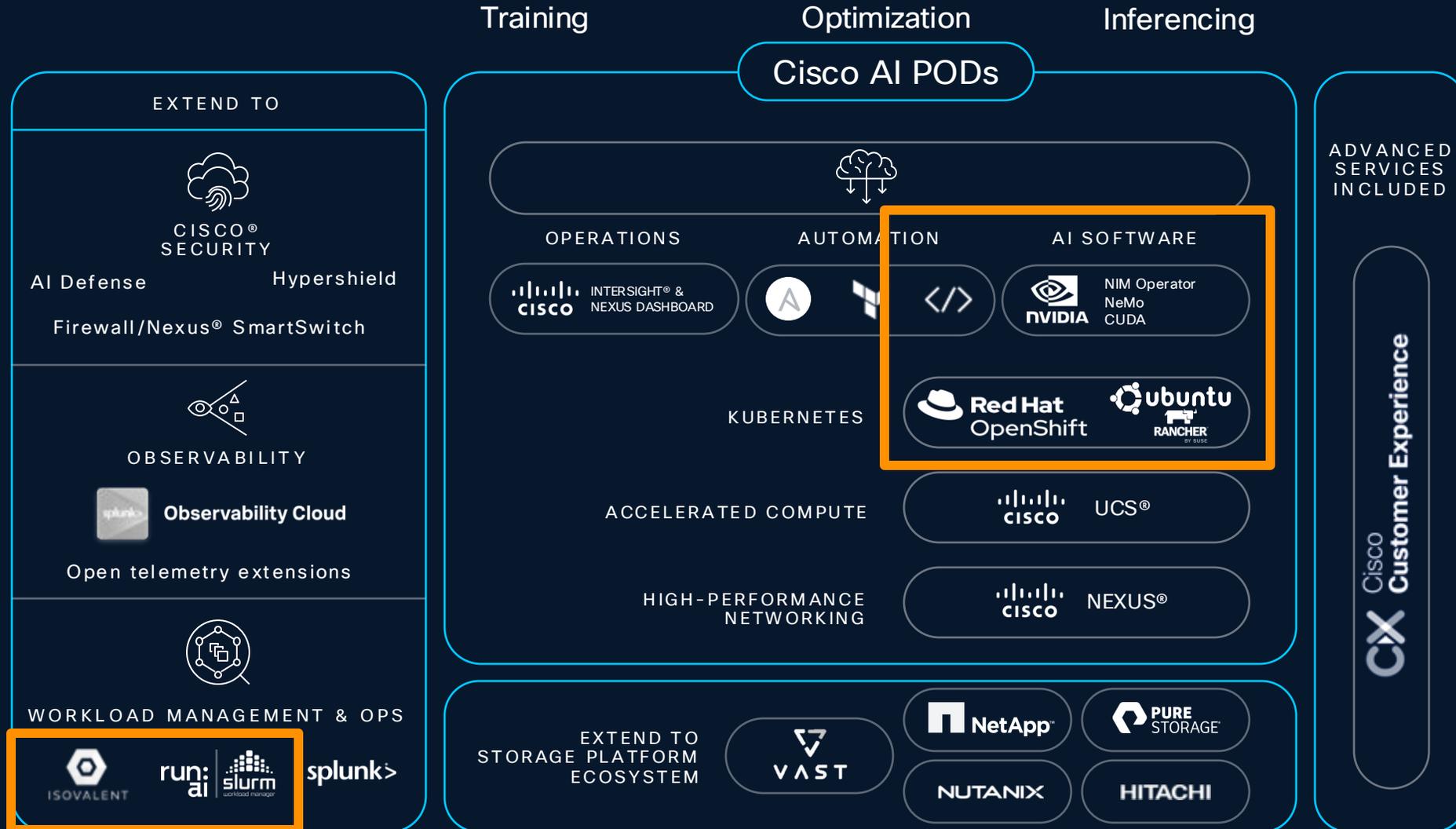
Virtual assistant

Whether for customer service, employee support, or business analytics, use this Blueprint to build an interactive AI agent to accelerate your growth.

Cisco AI PODs

Introducing AI POD “Integrated Offerings”

BYO AI tools:



- RAFAY
- Kubeflow
- jupyter
- Apache Airflow
- Weights & Biases
- mlflow
- neptune.ai
- kedro
- comet
- ZenML
- CLEAR ML
- PREFECT
- Flyte
- mongoDB

Cisco Reference Architectures

Aligned with NVIDIA Reference Architectures

Foundational Cisco Reference Architectures

- NVIDIA AI Software
- NVIDIA certified servers (HGX, MGX, RTX Pro)
- AI Optimized Backend GPU Networking with Nexus, BlueField, & Spectrum-X
- Standard Scale Unit deployments

Extended With

- Security & Observability solutions

AI Infrastructure with Cisco Nexus 9000 Switches
Cisco Enterprise Reference Architecture

Updated: October 26, 2025

Table of Contents

Introduction

Featuring Cisco UCS® C885A M8 Rack Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X

Introduction

Cisco® AI Infrastructure Reference Architecture (IRAA) is based on Cisco Nexus™ 9000 Series Switches for networking AI clusters managed by the on-premises Cisco Nexus Dashboard platform. It adheres to the NVIDIA Enterprise Reference Architecture for NVIDIA HGX™ H200 with NVIDIA Spectrum™-X networking.

Cisco Nexus 9000 Series Switches, powered by Cisco Silicon One™ and Cisco Cloud Scale architecture, provide high-speed, deterministic, low latency, and power-efficient connectivity for AI and high-performance computing (HPC) workloads. With the availability of multiple form factors, optics, and rich software features of the Cisco NX-OS operating system, Nexus 9000 switches provide a consistent experience for hardware, storage, backend, and on-premises (OPEX) management networks (see Figure 1).

Cisco Nexus Dashboard is the operations and automation platform for managing the Nexus 9000 switch-based fabric. It complements the data plane features of the Nexus 9000 switches by simplifying their configuration using built-in templates. It alerts on health issues, such as configuration, software, and health issues, in real time and automatically fixes them as needed. These issues can be resolved faster using integrations with commonly used tools, such as ServiceNow and Ansible, allowing the network of an AI cluster to be aligned with the existing workflows of an organization.

Cisco Reference Architecture

link

Cisco Nexus 9000 Cloud Partner Reference Architecture

Updated: October 26, 2025

Table of Contents

Introduction

Featuring Networking Reference Architecture of Cisco UCS C885A M8 Rack Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X

Introduction

The Cisco Cloud Partner Reference Architecture (CPRA) is designed to be deployed with a high GPU scale, ranging from 1K to 32K GPUs, at large Cloud Service Providers (CSPs) and high-performance Super Computing Centers (SCCs) in order to solve the most computationally intensive problems without affecting user of provisioning and operations. The overall design supports multi-tenancy in order to maximize the use of deployed hardware and, if required, can be scaled to 64K GPUs. Enterprises looking to deploy AI clusters with GPU scale less than 1K, should refer to Cisco Enterprise Reference Architecture (IRA) at this link.

The key technologies used in this CRA include:

- Cisco UCS® C885A Rack Servers with NVIDIA HGX™ H200 and Spectrum™-X E-Series
- Cisco® Silicon One™ and Cloud Scale NPV-based Nexus® 9000 Series Switches combined with Cisco compute, networking, and storage controllers
- Cisco Optics and cables
- Cisco provisioning, observability and security frameworks

AI and HPC Applications

Tenant1 Tenant2 Tenant3

Inf Control Plane - Provisioning, Observability, Security

link

Cisco Nexus Hyperfabric AI Enterprise Reference Architecture
Compliant with NVIDIA Enterprise Reference Architectures

Updated: October 23, 2025

Table of Contents

Introduction

Featuring Cisco® AI Infrastructure Reference Architecture (IRAA) compliant with NVIDIA Enterprise Reference Architectures, featuring Cisco® cloud-managed AI/ML networking of Cisco UCS® C885A M8 Rack Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X

Introduction

Cisco Nexus™ Hyperfabric AI is an on-premises AI cluster that is managed by a cloud-based controller. It empowers and simplifies your AI initiatives and accelerates AI deployments with a comprehensive, integrated, cloud-managed solution. Cisco Nexus Hyperfabric AI Reference Architecture is based on Cisco Silicon One™ switches and adheres to the NVIDIA Enterprise Reference Architecture (Enterprise RA) for NVIDIA HGX™ H200 and Spectrum™-X.

Figure 1 shows the key components of the solution. The key hardware components used in the cluster are described in the next section.

Cisco Nexus Hyperfabric AI

On-premises AI architecture

Pods of plug-and-play leaf-spine fabrics

Cisco 6000 Series Switches

link

Cisco Hyperfabric AI Cloud Partner Reference Architecture

Updated: October 26, 2025

Table of Contents

Introduction

Featuring Networking Reference Architecture of Cisco UCS C885A M8 Rack Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X

Introduction

The Cisco Cloud Partner Reference Architecture (CPRA) is designed to be deployed with a high GPU scale, ranging from 1K to 32K GPUs, at large Cloud Service Providers (CSPs) and high-performance Super Computing Centers (SCCs) in order to solve the most computationally intensive problems without affecting user of provisioning and operations. The overall design supports multi-tenancy in order to maximize the use of deployed hardware and, if required, can be scaled to 64K GPUs. Enterprises looking to deploy AI clusters with GPU scale less than 1K, should refer to Cisco Enterprise Reference Architecture (IRA) at this link. The key technologies used in this CRA include:

- Cisco UCS® C885A M8 Rack Servers with NVIDIA HGX™ H200 and Spectrum™-X E-Series
- Cisco® Silicon One™ NPV-based Cisco Nexus Hyperfabric™ Switches combined with Cisco compute, networking, and storage controllers
- Cisco Optics and cables
- Cisco provisioning, observability and security frameworks

AI and HPC Applications

Tenant1 Tenant2 Tenant3

Inf Control Plane - Provisioning, Observability, Security

link

AI dCloud Demo Collection

Cisco dCloud /

Artificial Intelligence (AI)

Elevate your sales strategy by unlocking the potential of artificial intelligence. Explore curated resources, including dCloud demos, that bring intelligent UX to customers.

AI ON Cisco

Cisco plays a pivotal role in providing fundamental infrastructure solutions that are revolutionary and highly adaptable. Scale your infrastructure to meet the increasing demands of AI workloads.

[Infrastructure Demos & Labs](#) [Use Cases](#)

Solution Demo

Accelerate AI Innovation with Cisco AI-Ready Data Center

Accelerate and simplify AI/ML deployments at scale.

[Explore](#)

Solution Demo

Simplify AI Deployments with Cisco AI-PODs using UCS X-Series

Accelerate customer initiatives from infrastructure to actionable insights.

[Schedule](#)

Scheduled Demo

Run Gen AI and LLMs on Cisco UCS X-Series

Meet today's business requirements and future AI demands.

[Schedule](#)

AI IN Cisco

...tics, assurance, consistency, and

On this page

- [AI ON Cisco](#)
- [AI IN Cisco](#)

Talking Points

- [Cisco AI/ML Whitepaper >](#)
- [Cisco DC Networking Blueprint for AI/ML Applications >](#)
- [CVD for Data Center Networking Blueprint for AI/ML >](#)
- [Cisco DNA Center AI-Enhanced RRM Deployment Guide >](#)
- [Explore AI-ready infrastructure >](#)
- [Unleashing Creativity with Generative AI At-a-Glance >](#)
- [Cisco Converged Infrastructure Design Navigator >](#)

[Give us feedback](#)



Infrastructure
to power AI



Security for AI,
AI for security



Services to accelerate
the value of AI



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

