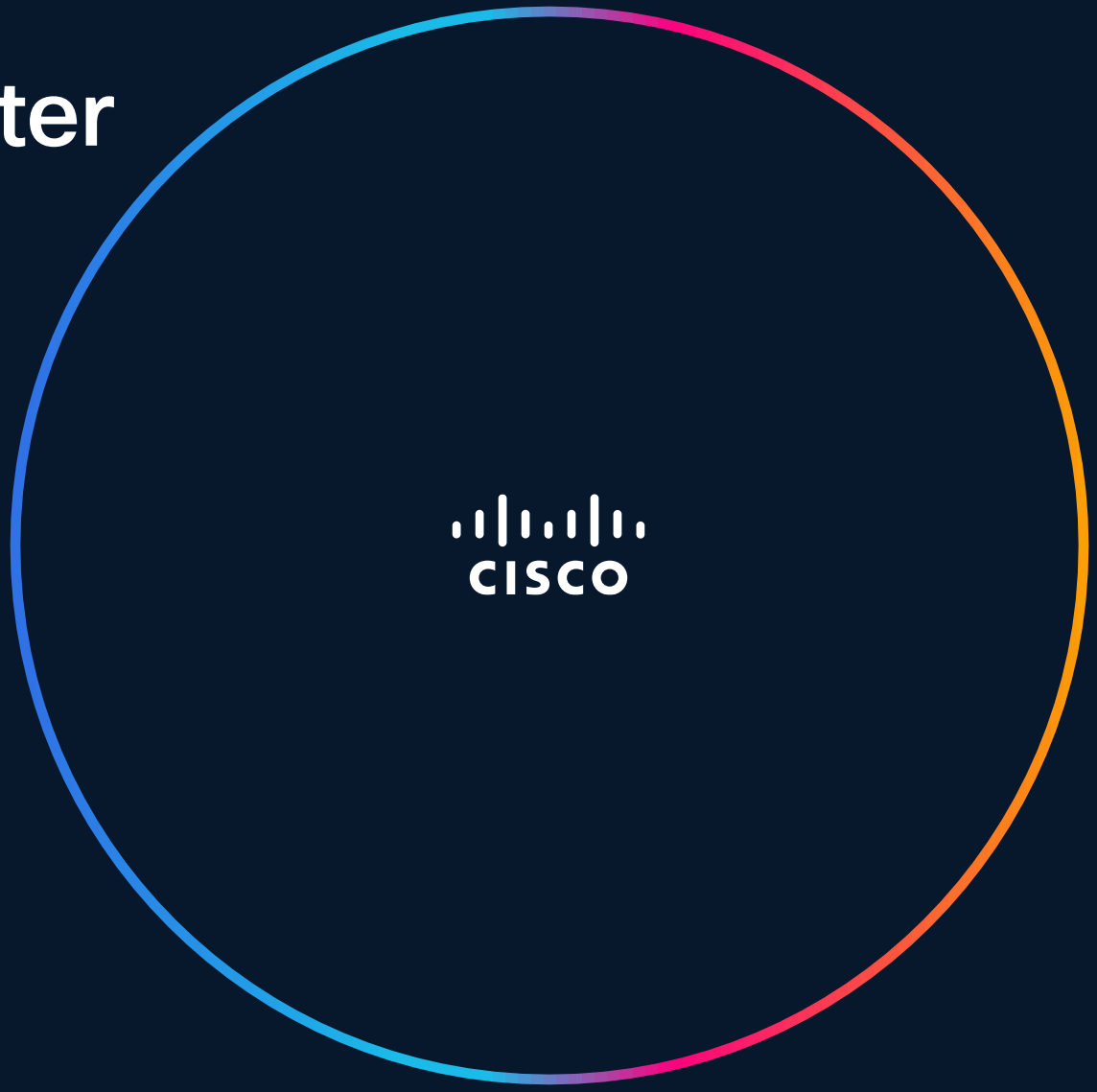
What Goes Into an AI Cluster

Discover how Cisco accelerates AI adoption with secure, scalable, and high-performance AI-ready datacenter infrastructure – Integrating Cisco's advanced technologies and NIVIDIA's innovations to simplify deployment and support of diverse AI use cases.



Nate Reid
AI Solutions Engineer

October 2025

Nate Reid

https://www.linkedin.com/in/natereid/

- Cisco Solutions Engineer with 25+ years of industry experience in tech.
- Blend of experience; including virtual infrastructure, cybersecurity, DevOps, MLOps, container orchestration, and machine learning/AI.
- Prior to Cisco, worked for an AI spinout from ETH Zurich. Focused on computer vision and LLM robustness.
- Detroit based, working with our Central region customers and partners.

© 2025 Cisco and/or its affiliates. All rights reserved.

Any information provided in this document regarding future functionalities is for informational purposes only and is subject to change including ceasing any further development of such functionality. Many of these future functionalities remain in varying stages of development and will be offered on a when-and-if available basis, and Cisco makes no commitment as to the final delivery of any of such future functionalities. Cisco will have no liability for Cisco's failure to deliver any or all future functionalities and any such failure would not in any way imply the right to return any previously purchased Cisco products.

© 2025 Cisco and/or its affiliates. All rights reserved.

Agenda

01 AI Infrastructure Considerations
02 Cisco AI PODs
03 AI Workload Orchestration
04 Operations and Automation
05 MLOps
06 Cisco AI PODs Extensibility

But first… a quick level-set on some terminology

- Foundational Model
- Pre-Training
- Fine-tuning
- Inferencing
- Token
- Context
- Parameters
- TTFT
- Transformer
- Attention Head
- Embedding
- RAG (Retrieval-Augmented Generation)
- MCP (Model Context Protocol)
- MLOps (Machine Learning Operations)
- Guardrails
- AI Factory
- …

© 2025 Cisco and/or its affiliates. All rights reserved.

But first... a QUICK level-set on some terminology

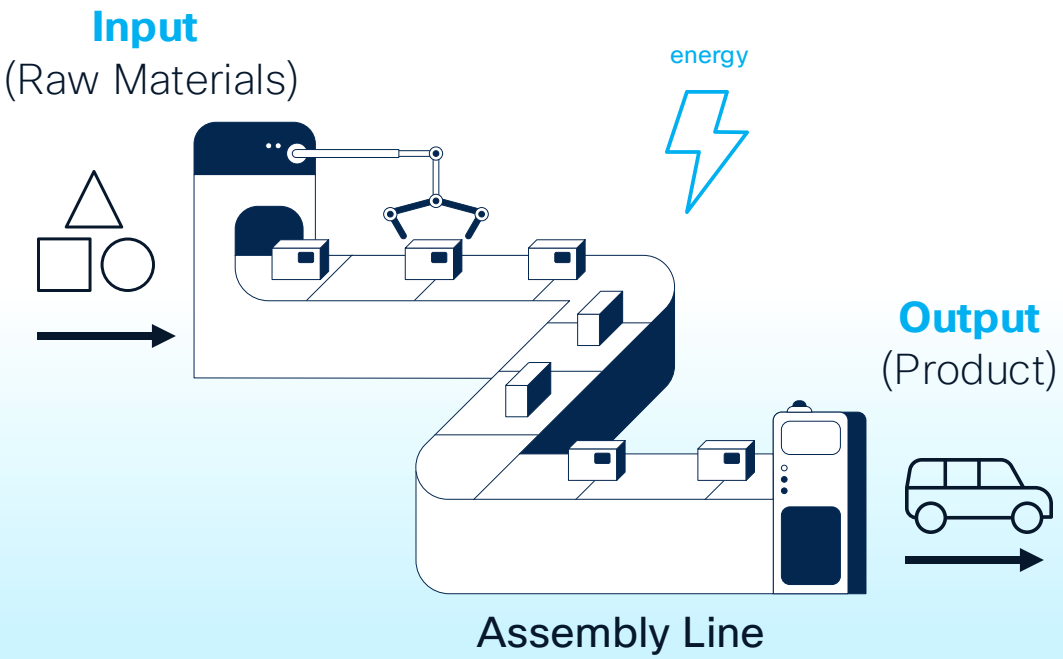
- Foundational Model
- Pre-Training
- Fine-tuning
- Inferencing
- Token
- Guardrails
- AI Factory

© 2025 Cisco and/or its affiliates. All rights reserved.

The Factory

Repeatable, scalable business capability

- Mass Production Efficiency
- Quality Control
- Supply Chain Integration

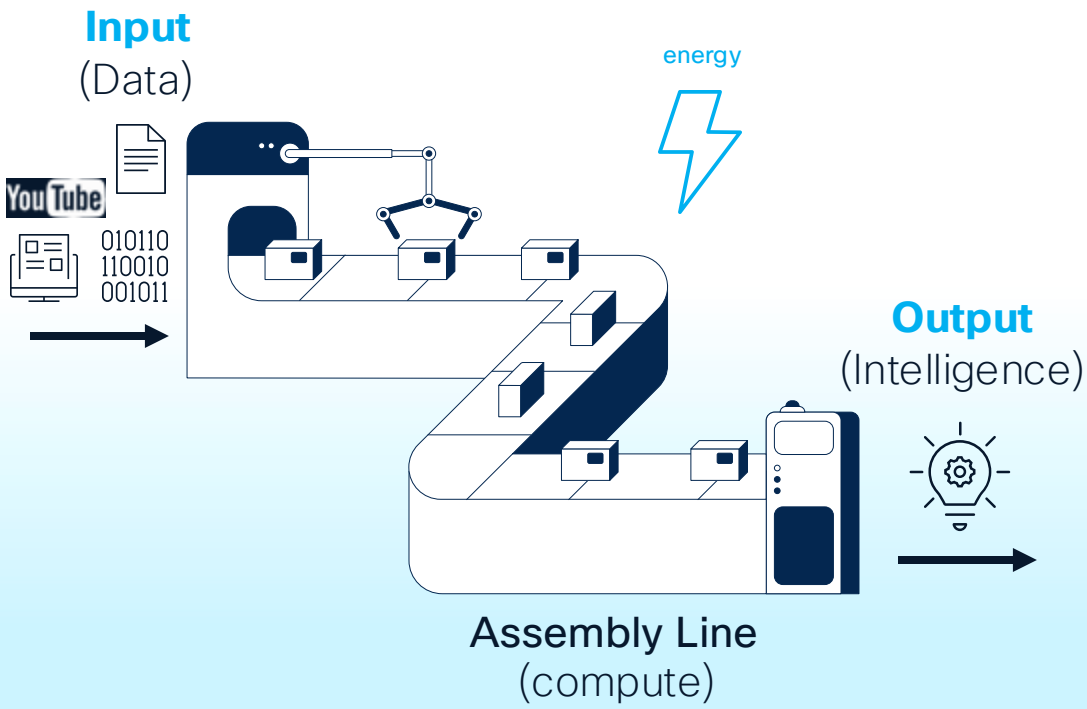


© 2025 Cisco and/or its affiliates. All rights reserved.

The AI Factory

The AI Factory intuition is to unify hardware + software into a single system. Outcome: Resource sizing is not standalone – compute, network, and storage must be co-designed to avoid bottlenecks.

- Mass model/token production
- Model performance control
- Data and workflow integration



© 2025 Cisco and/or its affiliates. All rights reserved.

Agenda

01 AI Infrastructure Considerations

02 Cisco AI PODs

03 AI Workload Orchestration

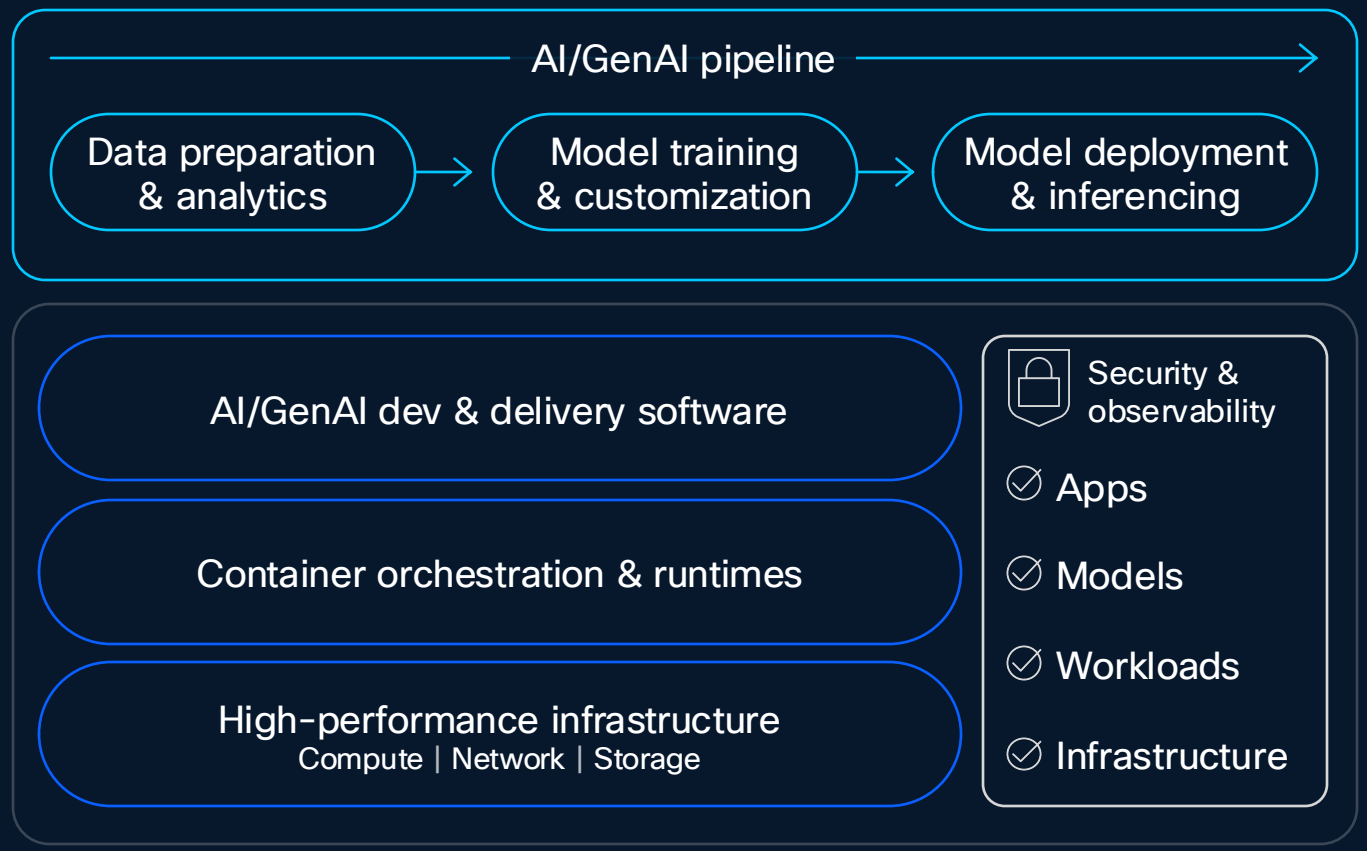
04 Operations and Automation

05 MLOps

06 Cisco AI PODs Extensibility

Enterprises need capable infrastructure to operationalize AI

AI applications are different, and they are driving demand for new architectures, security mechanisms, and lifecycle management



Faster time to business value

Mitigate risks

Simplify deployment

Reliability, availability, flexibility

© 2025 Cisco and/or its affiliates. All rights reserved.

Enterprise AI Infrastructure Requirements

Customizing foundational models

Training LLMs from scratch is cost-prohibitive for the average enterprise

Support multiple, smaller workloads

Enterprises can have many use cases spread across different LOBs, each using an LLM (worst case)

Integrate into existing data centers with ease

AI-enabled enterprise applications often need data, applications and other resources in existing data centers

© 2025 Cisco and/or its affiliates. All rights reserved.

Enterprise AI Infrastructure Requirements

Operational ease and consistency

Existing DC operational model to manage AI backend/frontend fabrics; simplify with fabric + server templates or IaC

Consistency and simplicity at scale

Building block approach using a spine-leaf architecture to scale-out with consistency and predictability

Fit-to-purpose AI infrastructure

Performant infrastructure without compromising on choice; established technologies

© 2025 Cisco and/or its affiliates. All rights reserved.

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 challenges

Model training and inferencing can generate a lot of traffic, slowing networks. Multi-node training demands significantly more from the network

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

Cisco AI-Ready Data Center

Transform data centers to power AI workloads anywhere

Solving key AI Challenges...

- Complexity of AI infrastructure deployments
- Security vulnerabilities
- Network performance bottlenecks

...by leveraging critical components:

- Purpose-built integrated AI infrastructure
- Full stack AI systems
- E2E AI infrastructure operations
- AI focused and equipped security

Digital Resilience: security, reliability, and performance

© 2025 Cisco and/or its affiliates. All rights reserved.

Agenda

01 AI Infrastructure Considerations

02 Cisco AI PODs

03 AI Workload Orchestration

04 Operations and Automation

05 MLOps

06 Cisco AI PODs Extensibility

Cisco AI PODs

Because no two organizations' AI requirements are identical, Cisco AI PODs deliver flexible, pre-validated infrastructure to meet diverse AI needs, from large-scale training to edge inferencing, with standardized, secure, and scalable solutions.

Large-scale training clusters

Extensive model optimization

Inferencing at scale and edge

Where **Cisco AI PODs** shine:

Large deployments with dedicated GPU backend networks

Multi-workload architectures and mixed use-cases to run AI models alongside enterprise apps

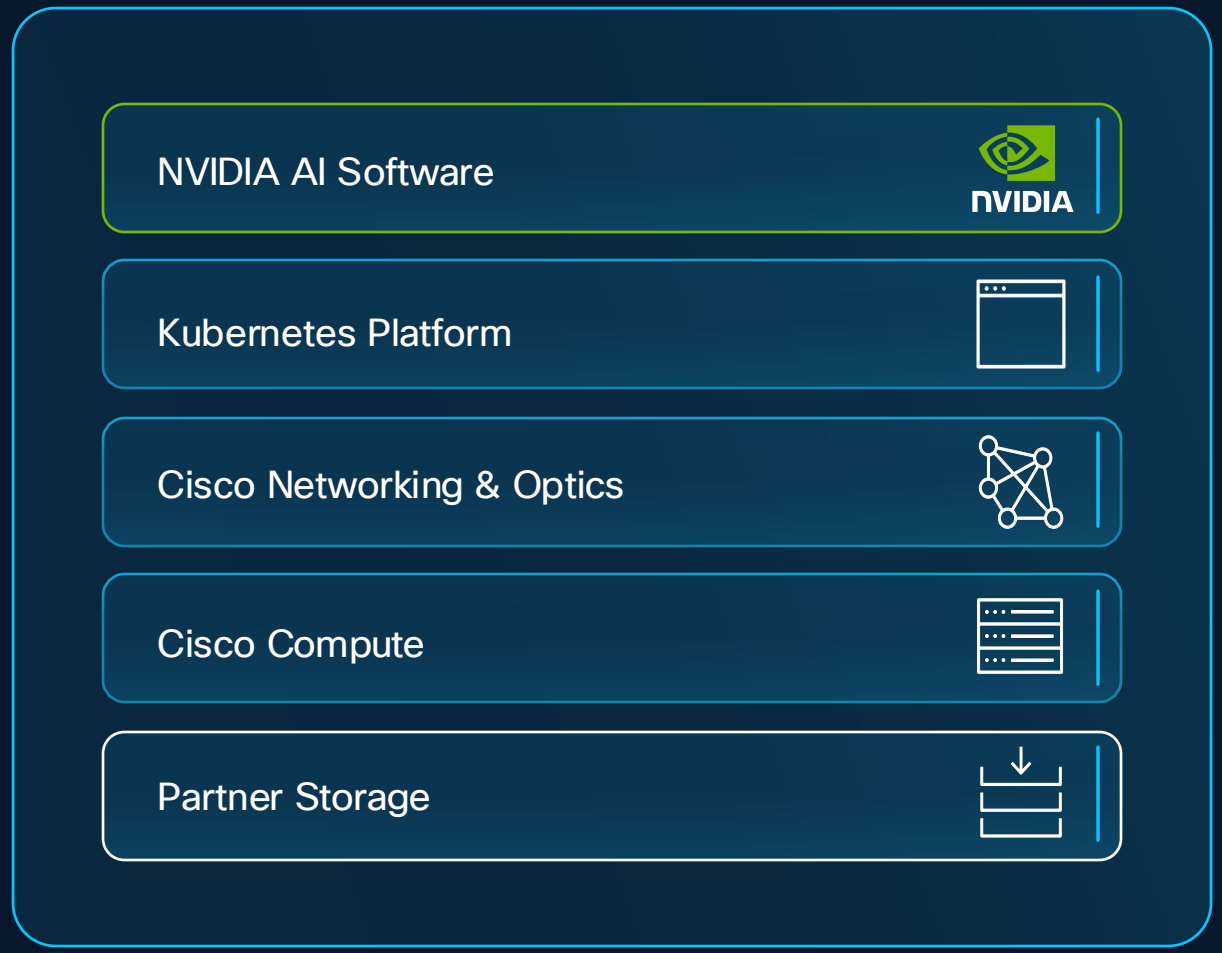
Inferencing for large AI models and high user concurrency at scale

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco AI PODs

AI Practitioners / MLOps

IT Infrastructure & Operations



© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

Cisco AI PODs

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

Deploy AI with confidence

Cisco CVD, NVIDIA ERA

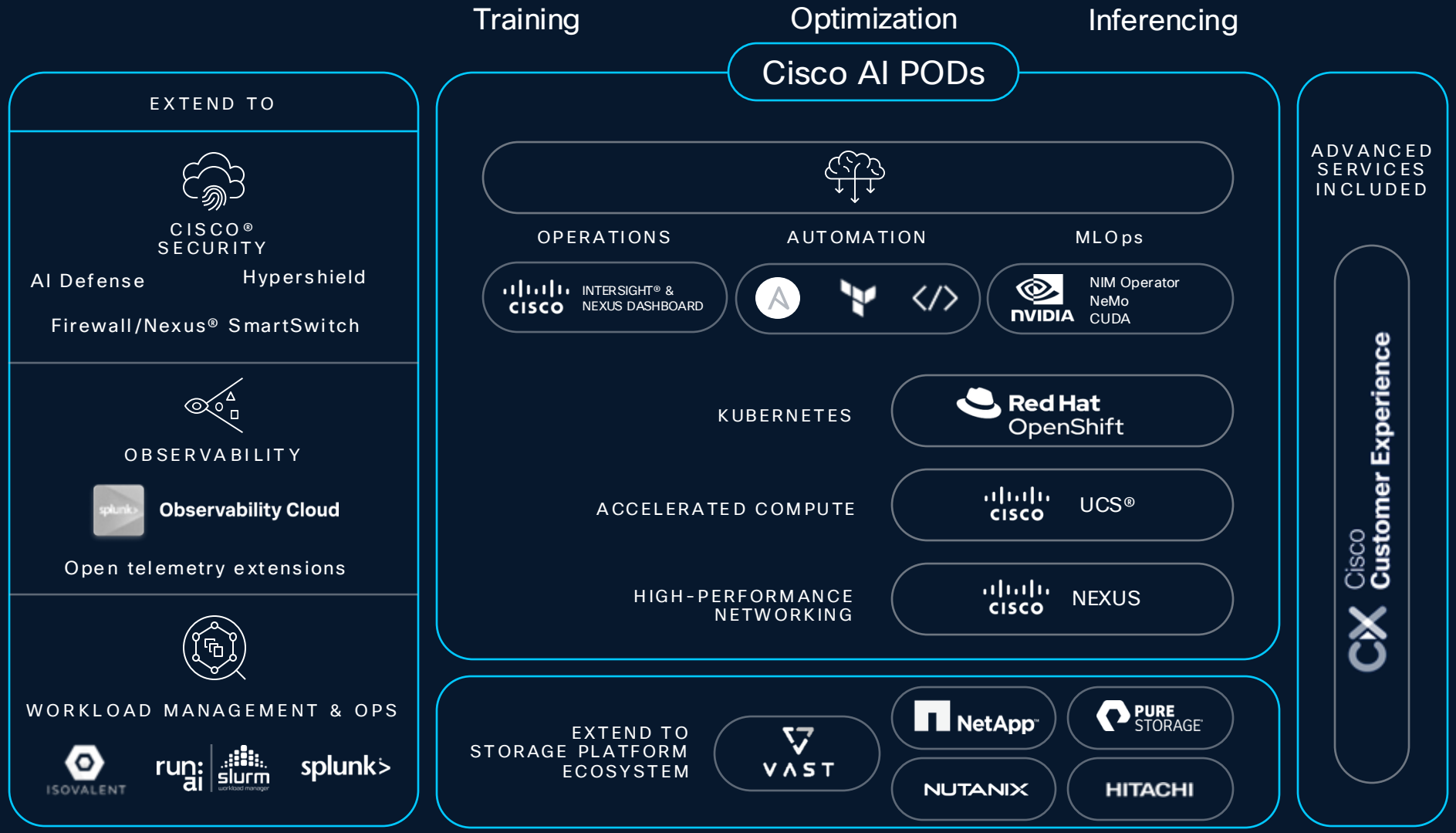
Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

Orderable, use case driven AI-Ready infrastructure stacks

Inferencing.
Optimization.
Training.

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



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco AI PODs

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

Deploy AI with confidence

Cisco CVD, NVIDIA ERA

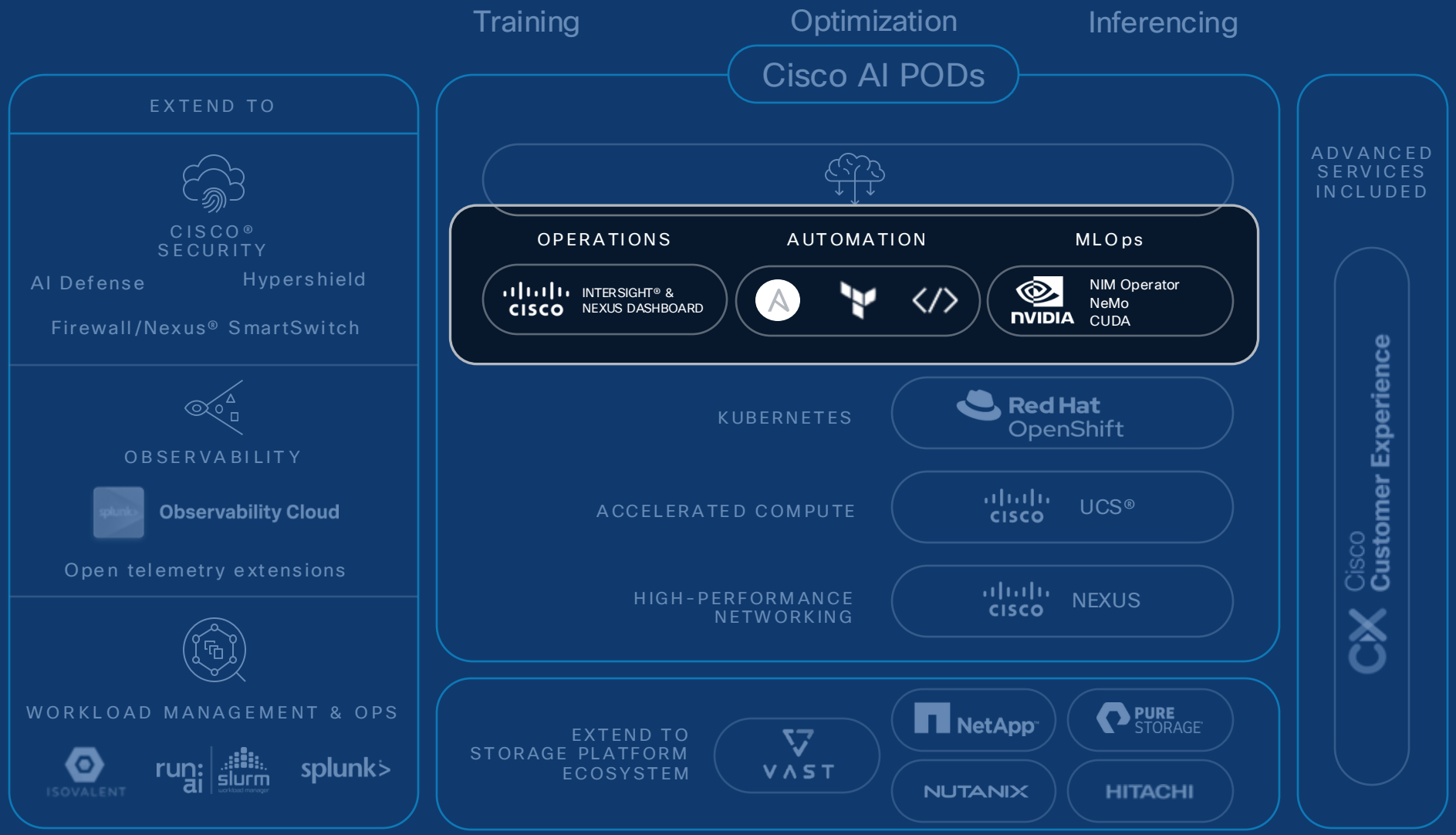
Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

Orderable, use case driven AI-Ready infrastructure stacks

Inferencing.
Optimization.
Training.

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



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco AI PODs

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

Deploy AI with confidence

Cisco CVD, NVIDIA ERA

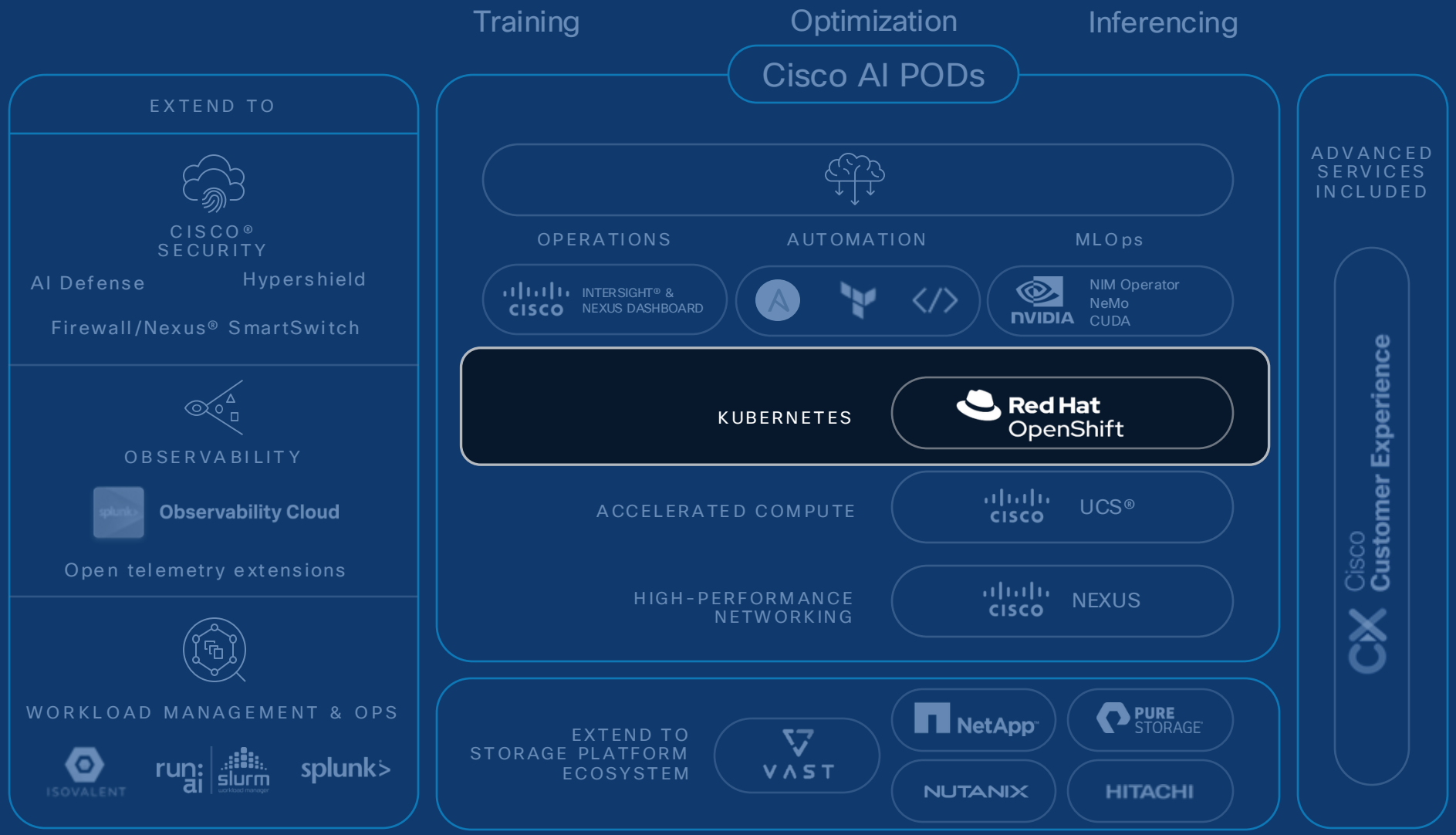
Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

Orderable, use case driven AI-Ready infrastructure stacks

Inferencing.
Optimization.
Training.

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



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco AI PODs

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

Deploy AI with confidence

Cisco CVD, NVIDIA ERA

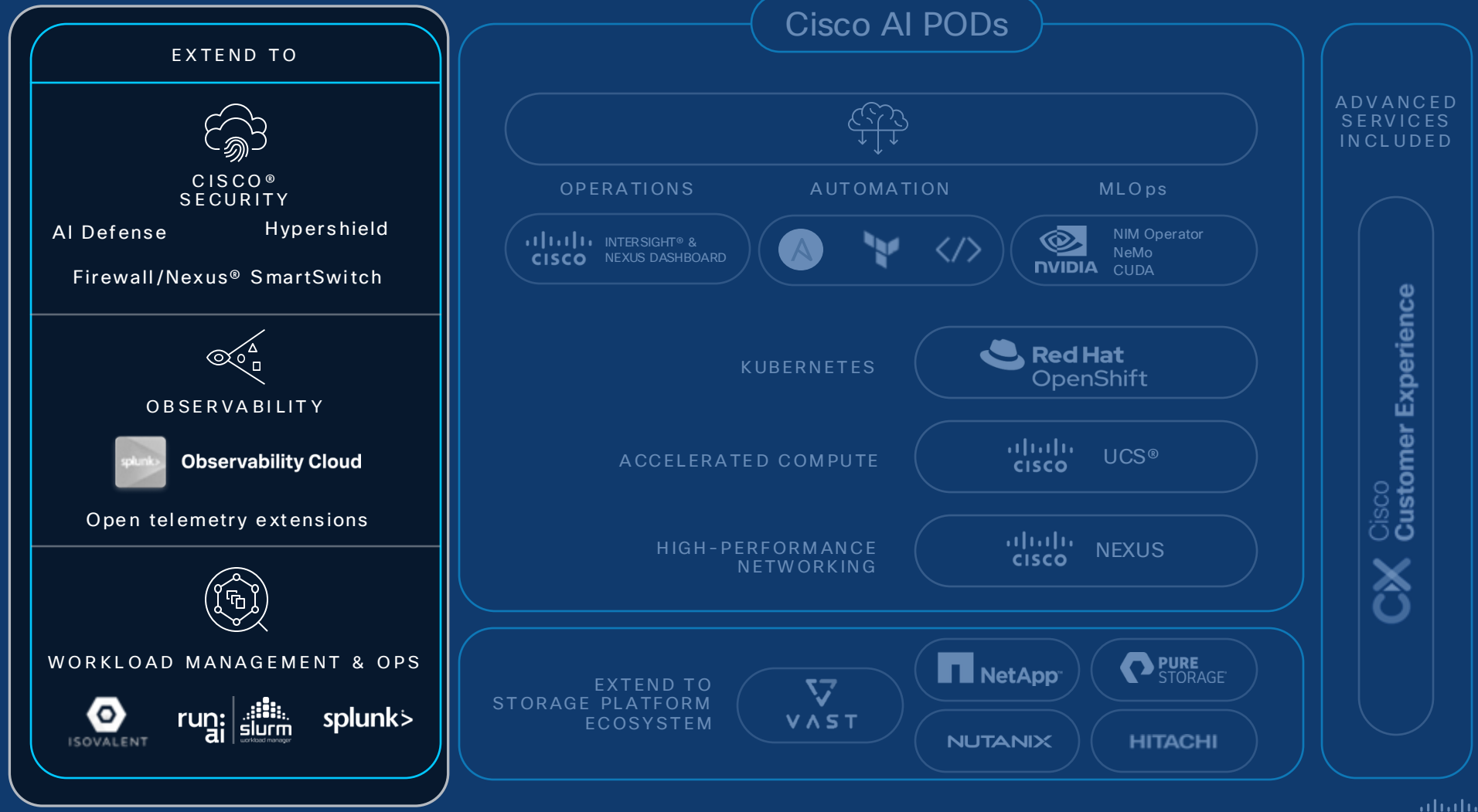
Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

Orderable, use case driven AI-Ready infrastructure stacks

Inferencing.
Optimization.
Training.

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



© 2025 Cisco and/or its affiliates. All rights reserved.

Agenda

01 AI Infrastructure Considerations

02 Cisco AI PODs

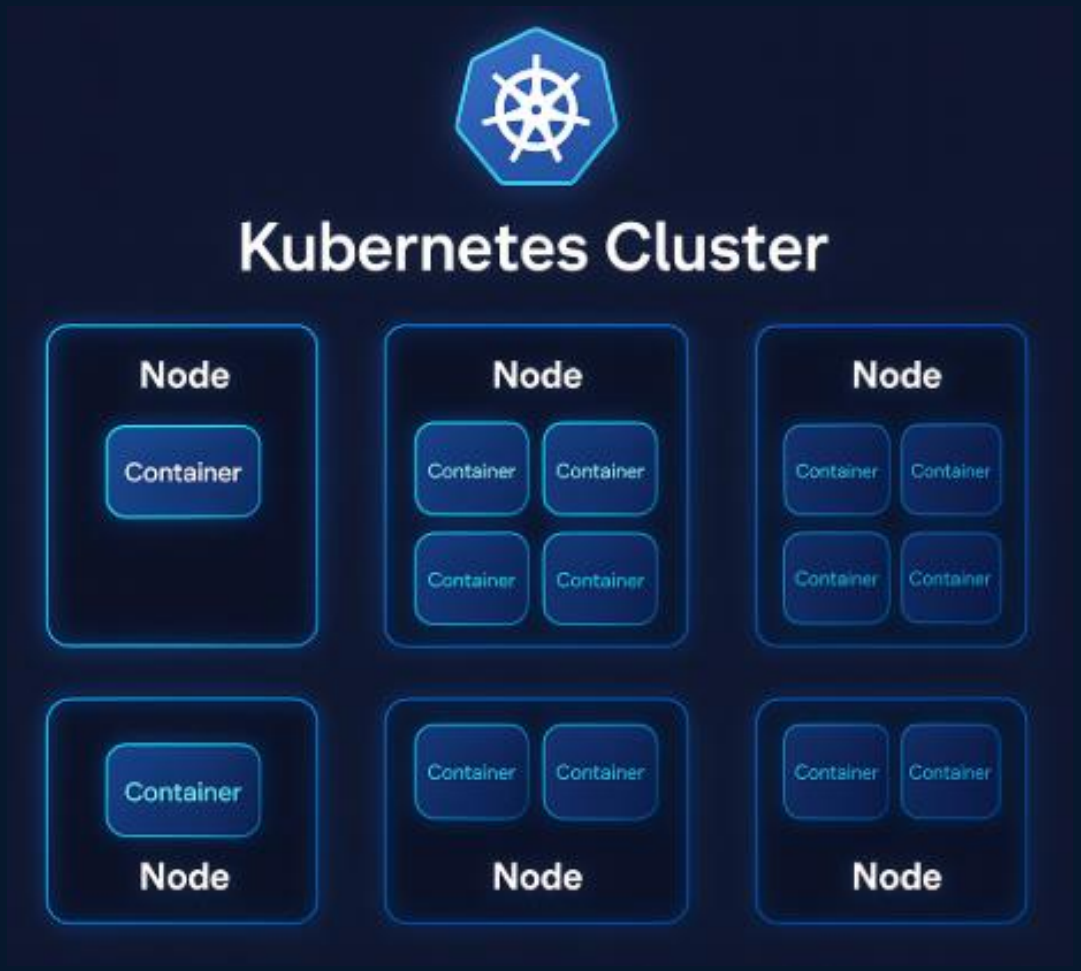
03 AI Workload Orchestration

04 Operations and Automation

05 MLOps

06 Cisco AI PODs Extensibility

Kubernetes



- Automation & orchestration – CI/CD
- Robust ecosystem of AI tooling
- GPU scheduling
- Portability
- Scalability
- Fault tolerance & reliability

© 2025 Cisco and/or its affiliates. All rights reserved.

Kubernetes with Red Hat OpenShift

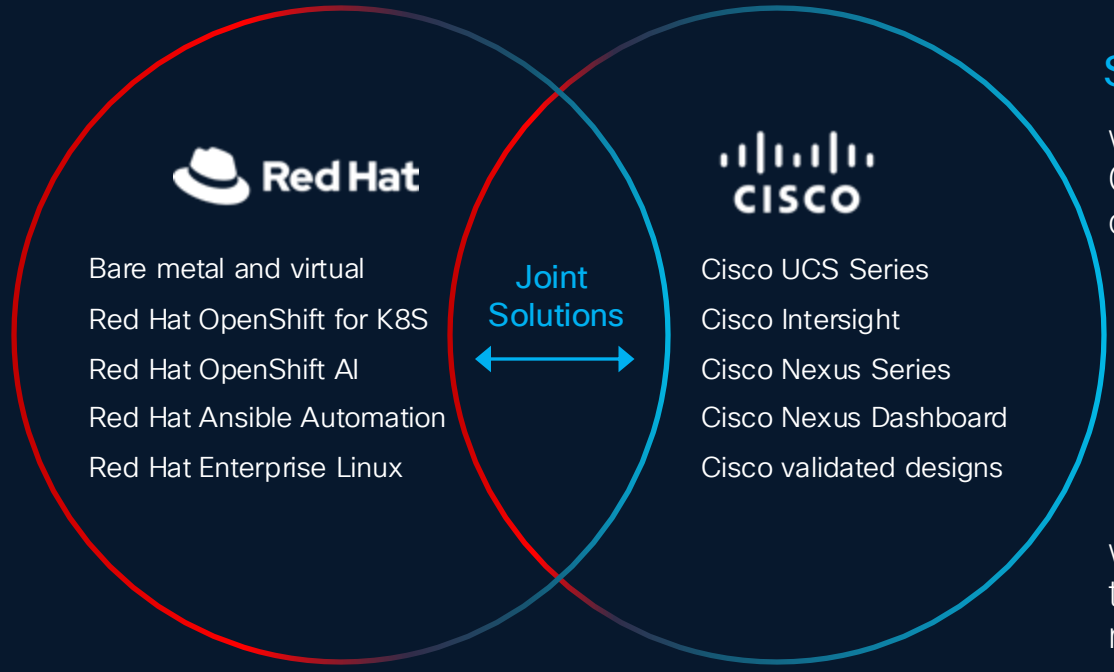
on-premises and cloud

Open Cloud Infrastructure

platform built on open-source innovation

Accelerated Time to value

with turnkey experience and integrated automation



Simplified Operations and Support

with Cloud managed infrastructure and Cisco Solution Support across Red Hat on converged infrastructure stacks

Reduced Risk

with Cisco Validated Designs, delivering tested architectures for standardized, repeatable deployments.

Operate across hybrid multicloud	More choice and flexibility	20+ Cisco Validated Designs	Consistent app dev experience	Increased sustainability

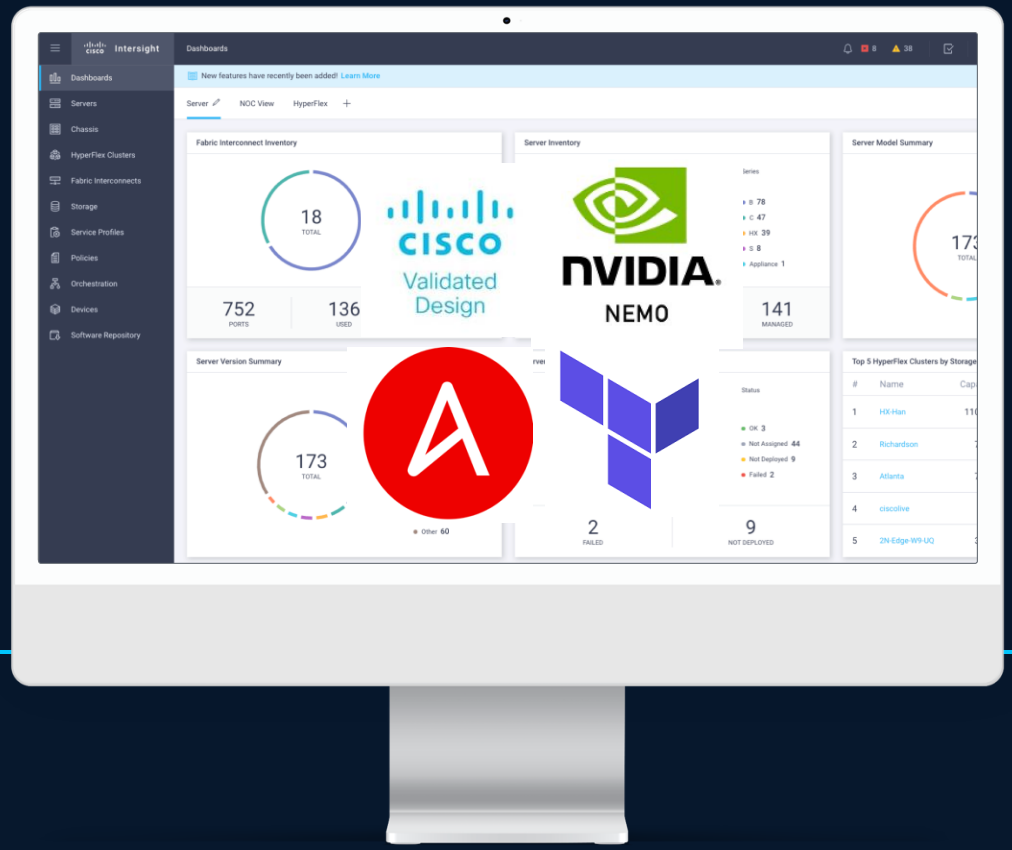
© 2025 Cisco and/or its affiliates. All rights reserved.

Agenda

01 AI Infrastructure Considerations
02 Cisco AI PODs
03 AI Workload Orchestration
04 Operations and Automation
05 MLOps
06 Cisco AI PODs Extensibility

Cisco Intersight

Unified Operating Model



Intersight Dashboard

Simplified operation with AI-driven capabilities including Connected TAC, and Predictive Insight

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

Consistent operational model globally, from DC to edge, at cloud scale

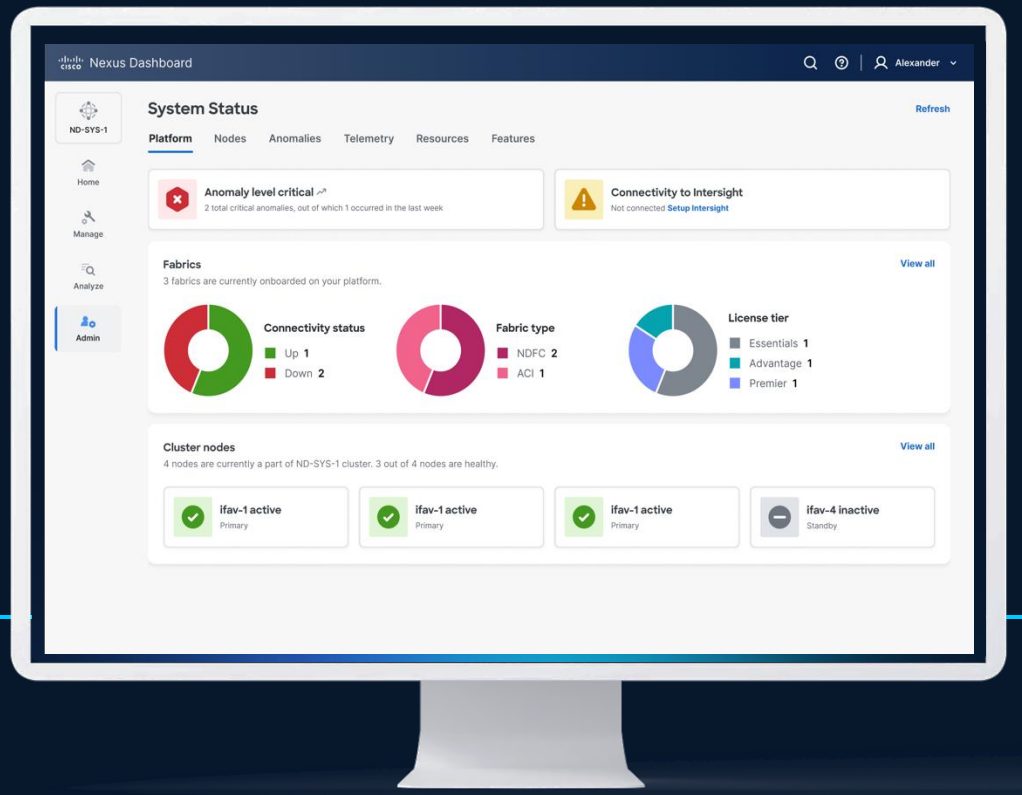
Secure operations with built-in advisories and continuous risk mitigation

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Nexus Dashboard

Simplify Data Center Network Operations

Common policy across NX-OS and ACI fabrics



Nexus Dashboard

Configure, operate and analyze your network from one place across data center networks

Minimize downtime through increased visibility and resolve problems with fix recommendations

Track **Power and Cooling** by surfacing your networks impact on KWh and CO2 emissions

Accelerate innovation with built in infrastructure-as-code and popular automation tooling integration

© 2025 Cisco and/or its affiliates. All rights reserved.

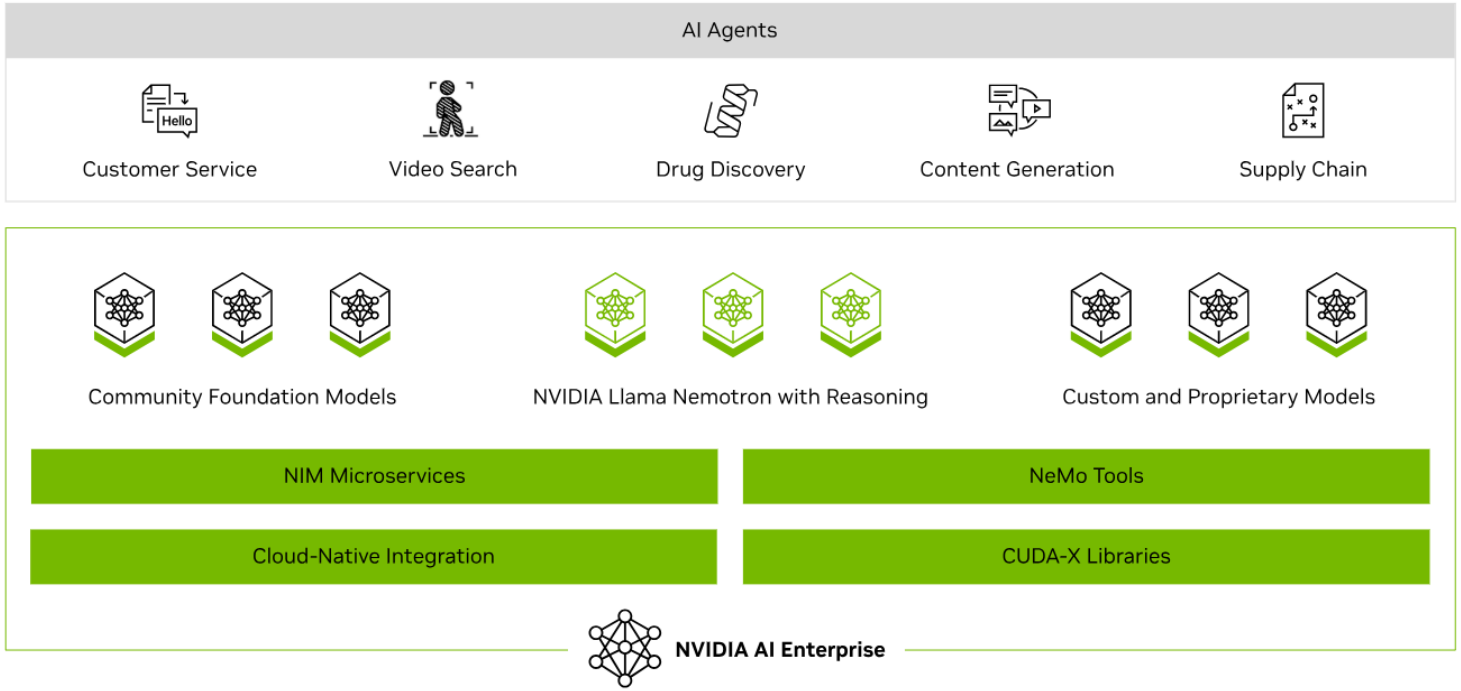
Agenda

01 AI Infrastructure Considerations
02 Cisco AI PODs
03 AI Workload Orchestration
04 Operations and Automation
05 MLOps
06 Cisco AI PODs Extensibility

NVIDIA AI Enterprise

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.

© 2025 Cisco and/or its affiliates. All rights reserved.

What will my support experience be?

Support in a multi-vendor solution

"If something breaks, which support team do I call?"

"We don't have the resources to manage multiple product support teams."

"Even minor maintenance changes can cause serious issues."

"Our support experience with some vendors has been inconsistent."

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Solution Support

One Service, Broad Coverage

Service Features	Cisco Solution Support
Global 24x7 product-level technical support	●
24-hour access to Cisco® online resources	●
Hardware replacement (2- and 4-hour, next business day)	●
Network management / operating system software updates and upgrades	●
Proactive diagnostics and immediate alerts on devices through Cisco Smart Call Home	●
Web-based user community for self-service support of smart capabilities	●
Cisco software application support	●
Primary point of contact with solution-level expertise	●
Accountability for issue resolution, no matter where it resides	●
Coordination between Cisco TAC and solution partner product support teams	●
Case management from first call to resolution	●

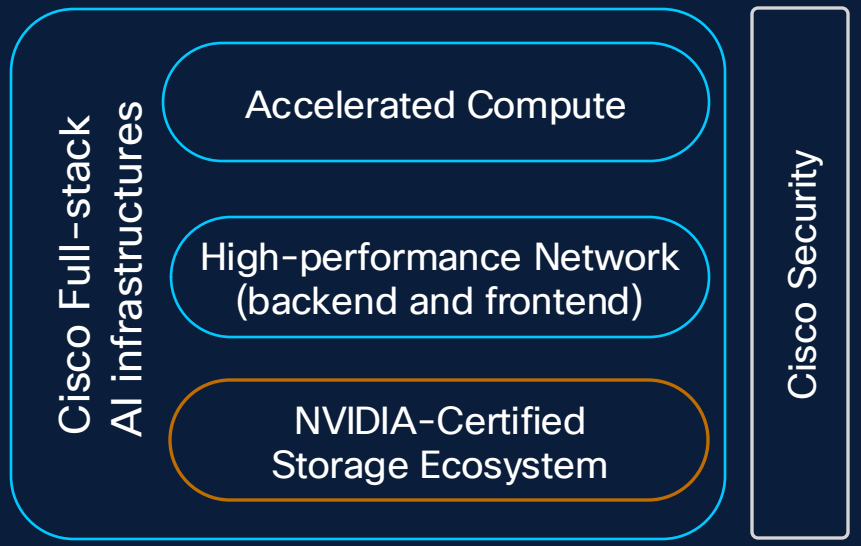
© 2025 Cisco and/or its affiliates. All rights reserved.

Agenda

01 AI Infrastructure Considerations
02 Cisco AI PODs
03 AI Workload Orchestration
04 Operations and Automation
05 MLOps
06 Cisco AI PODs Extensibility

Expanded partnership to accelerate AI adoption in the enterprise

Partnership focus on Enterprise data centers



Cisco is now included in NVIDIA Spectrum™-X Ecosystem, and a validated partner for NVIDIA Enterprise Reference Architectures

Cisco Secure AI Factory with NVIDIA (Secure AI Factory) enables enterprises to accelerate the adoption of AI use cases

Jointly deliver Secure AI Factory with NVIDIA Spectrum™-X Silicon integrated Cisco Silicon One® switches.

© 2025 Cisco and/or its affiliates. All rights reserved.

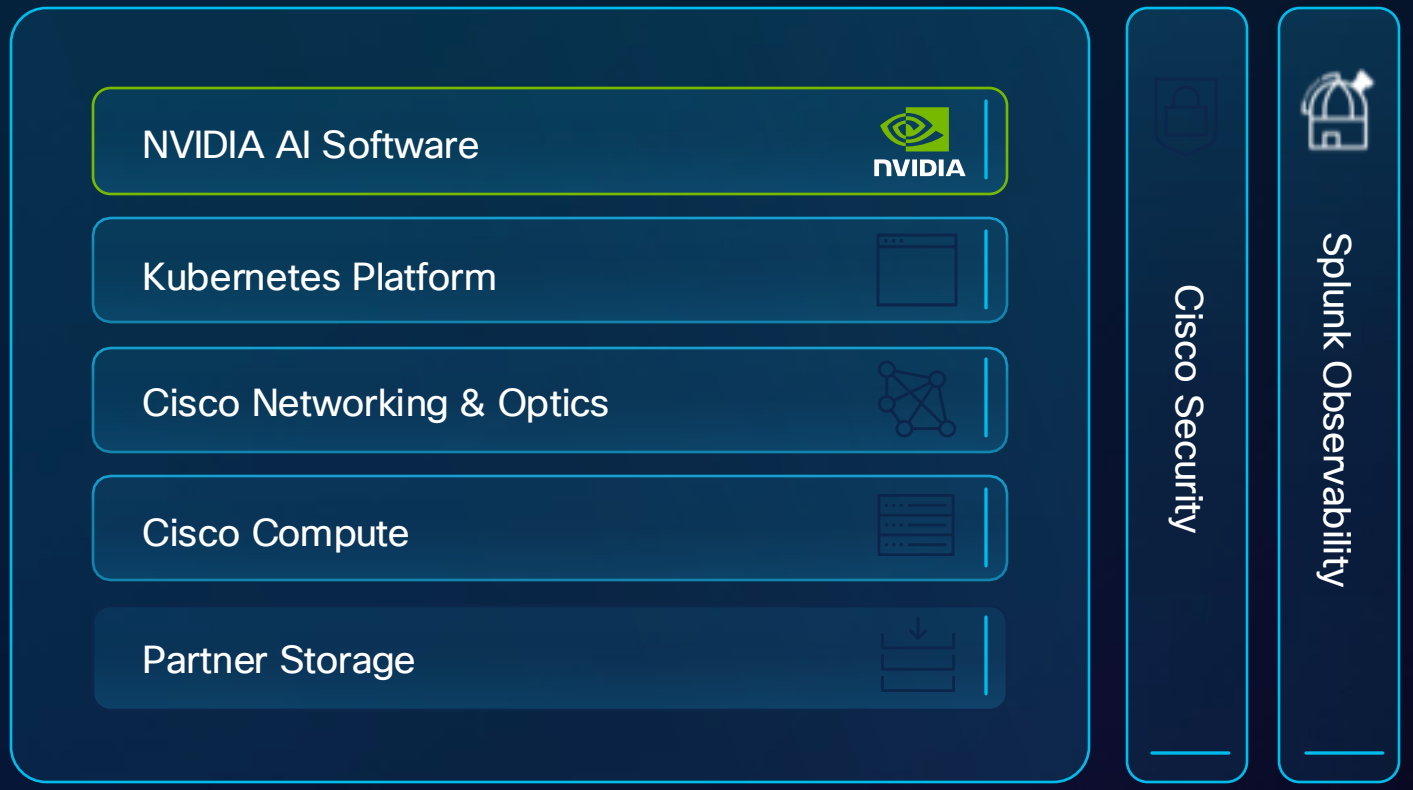
Cisco Secure AI Factory with NVIDIA

What is it?

Reference architecture

Validated solutions and turnkey offerings

Differentiated with Security and Observability



© 2025 Cisco and/or its affiliates. All rights reserved.

Security-first architecture enables safe Enterprise AI

Security at all layers of the stack

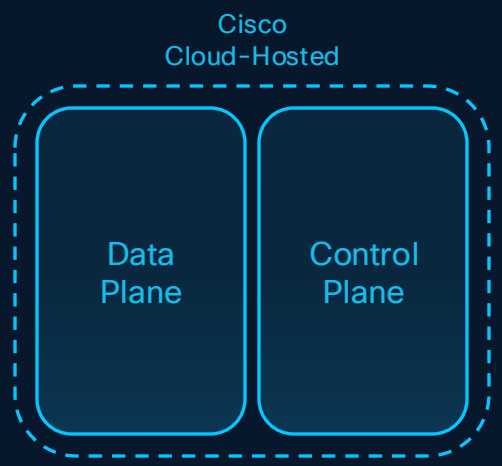
Securing the Applications	Cisco AI Defense—Robust testing and runtime security of LLMs and generative AI applications, integrated with NVIDIA AI.
Securing the Workloads	Cisco Hypershield—Protection against adversary lateral movement and proactive vulnerability mitigation without the need for patching, all from a single management interface, integrated with NVIDIA AI.
	Integration with NVIDIA Bluefield-3's DOCA AppShield for intrusion detection in AI-focused VMs and containers. **Future**
Securing the Infrastructure	Cisco Hybrid Mesh Firewall—Unified security management and consistent and pervasive policy across multiple enforcement points.
	Cisco Isovalent: Enhanced visibility into cloud native interactions, enabling smooth policy definition and enforcement across software defined networks.
	To include management of **NVIDIA BlueField®-3 DPUs** for enabling AI Cluster perimeter firewalls. **Future**

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Secure AI Factory – Layered Defense

© 2025 Cisco and/or its affiliates. All rights reserved.

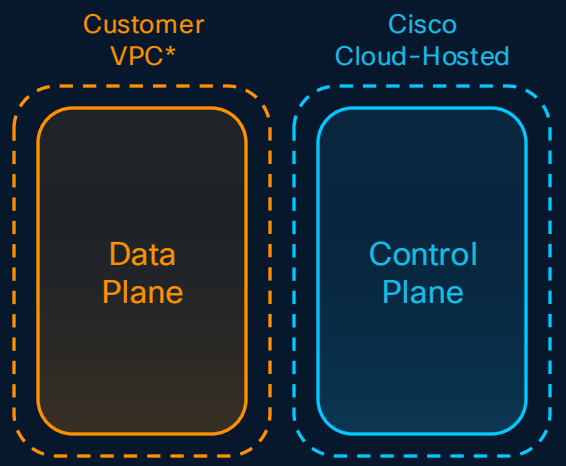
AI Defense Deployment options for every situation



SaaS

Fully hosted and managed in the cloud

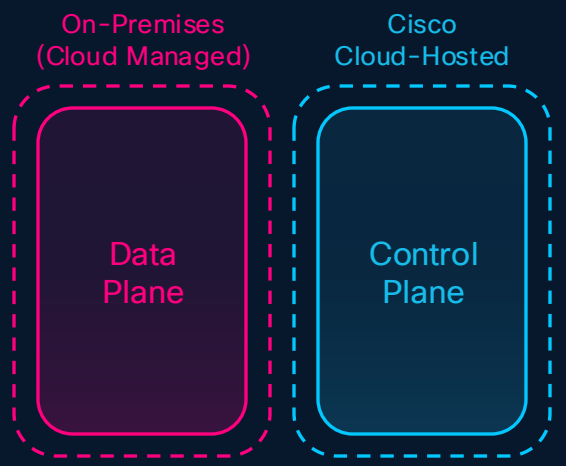
***Best for** customers looking for a simple, flexible deployment with zero infrastructure to manage*



VPC

Virtual private cloud environments with a cloud-hosted control plane

***Best for** customers looking to balance data control and compliance with cloud scalability*



On-Premises

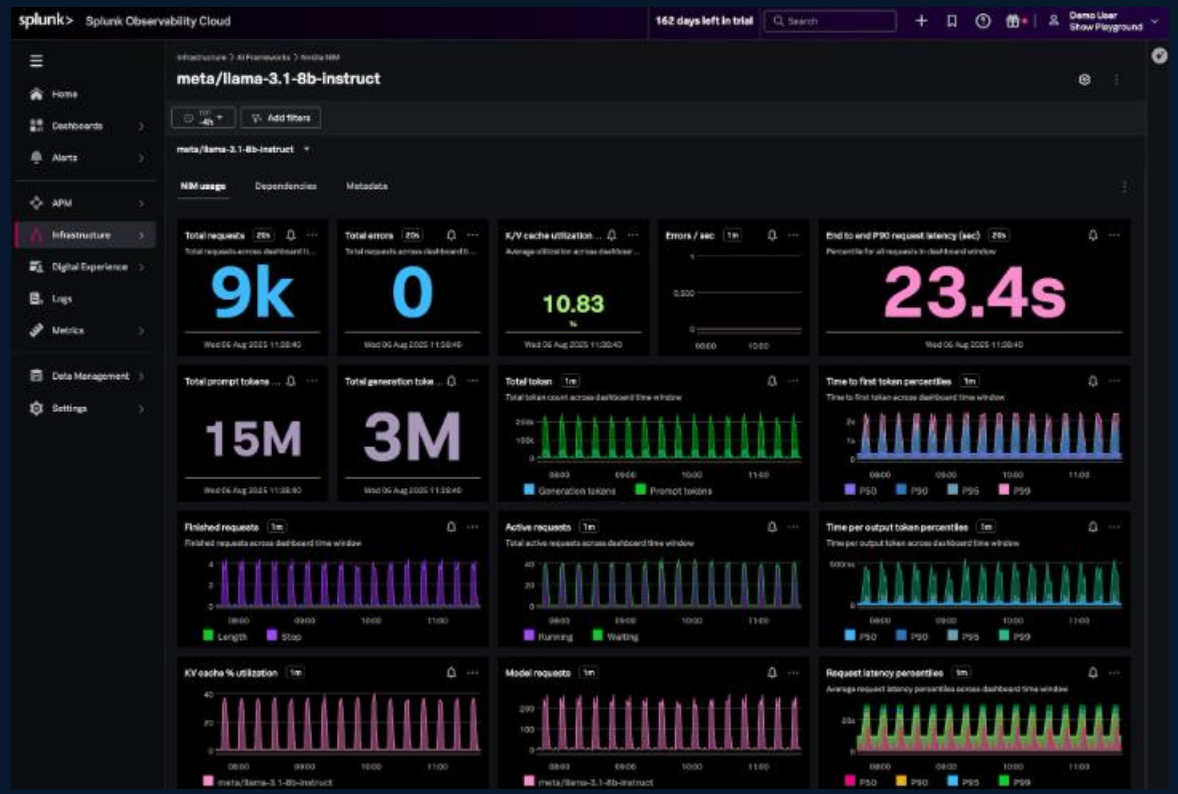
Combines Cisco UCS hardware with a cloud-hosted control plane

***Best for** customers that want to manage AI workloads themselves rather than relying on hyperscalers*

© 2025 Cisco and/or its affiliates. All rights reserved.

+ = Better Together

Splunk acts as the **observability + analytics layer** for Cisco AI PODs. While Intersight manages lifecycle, Splunk provides **cross-domain correlation, security, and operational intelligence**



- AI Infrastructure Observability
- AI/ML Cluster Monitoring
- Security & Compliance
- Capacity & Planning
- Networking & Storage Visibility

© 2025 Cisco and/or its affiliates. All rights reserved.

VAST Data on Cisco AI PODs

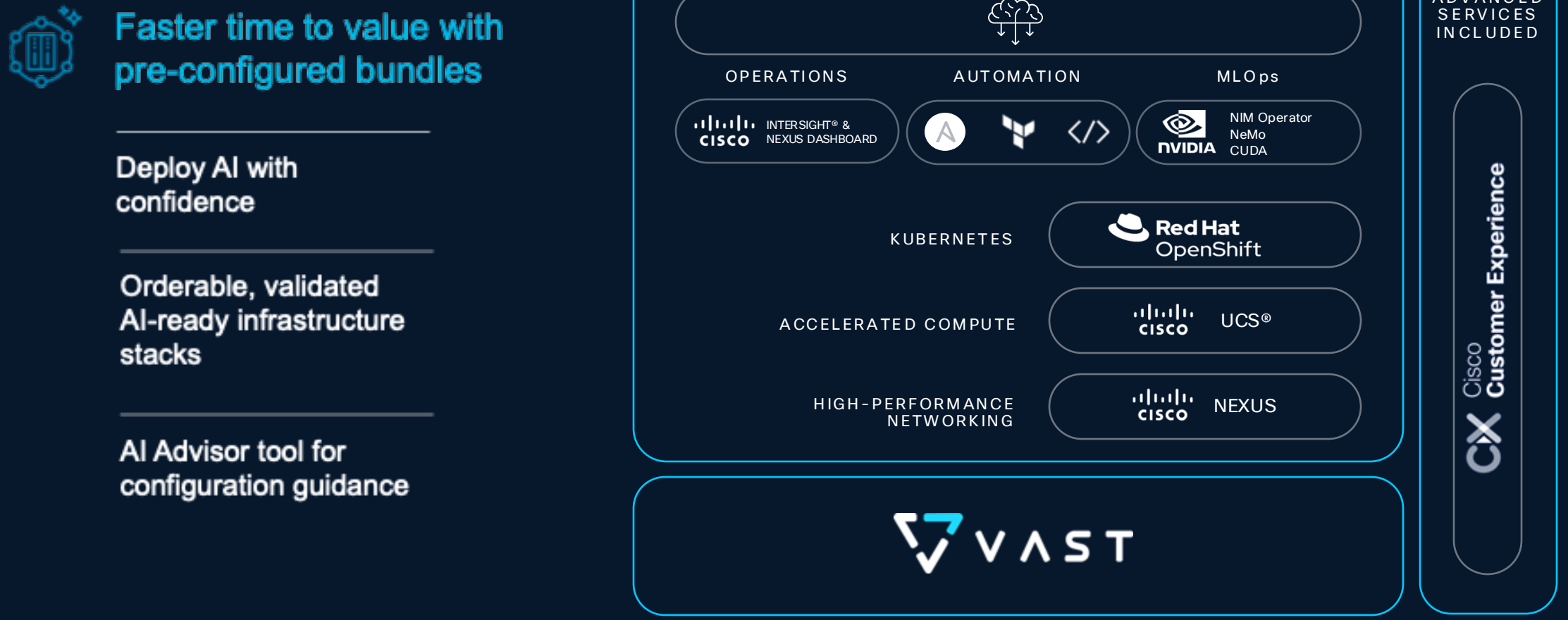
Simplified Full Stack Orderability

Faster time to value with pre-configured bundles

Deploy AI with confidence

Orderable, validated AI-ready infrastructure stacks

AI Advisor tool for configuration guidance



© 2025 Cisco and/or its affiliates. All rights reserved.

The VAST AI Operating System

DataEngine
Scalable, Event-Driven Computing
Triggers, Functions, Containers

SyncEngine
Data Router
Migrations

InsightEngine
Real-Time RAG
Embeddings

AgentEngine
Agent Tools & Orchestration
Agents, Tools, Observability

DataStore
Universal Storage & Data Protection Infrastructure
NFS, SMB, S3, NVMe/TCP, VAAI, CSI, vLLM

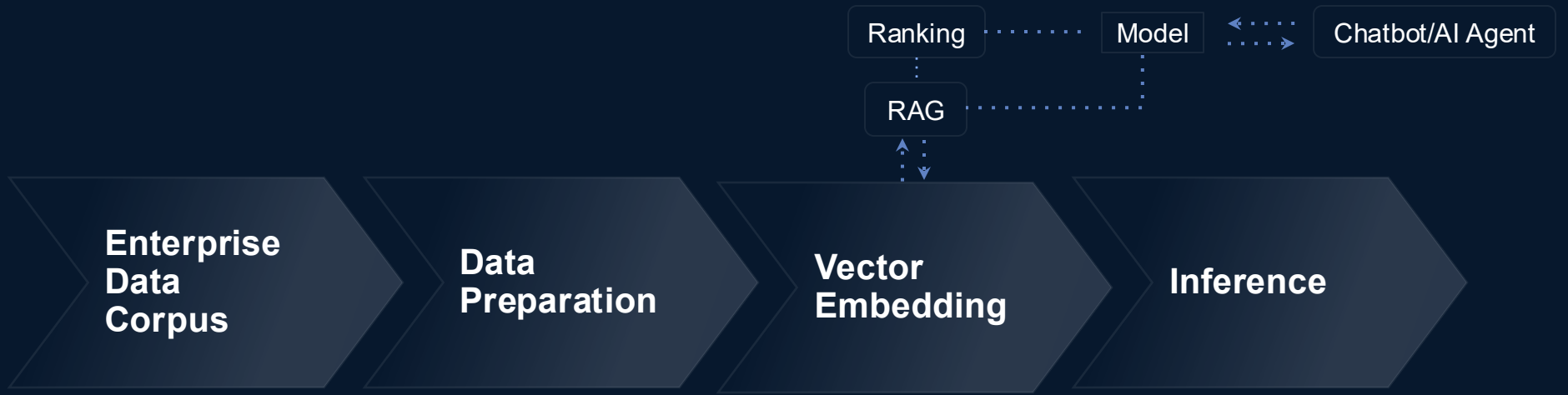
DataBase
Transactional, Analytical, & Vector Database
Kafka, Python, Parquet, SQL

DataSpace Globally-Distributed Data Computing

© 2025 Cisco and/or its affiliates. All rights reserved.

Understanding Enterprise AI Pipelines and RAG

Cisco, NVIDIA, and VAST working as a single solution



© 2025 Cisco and/or its affiliates. All rights reserved.

Summary – Cisco AI PODs

Deploy AI with confidence

Cisco CVD, NVIDIA ERA

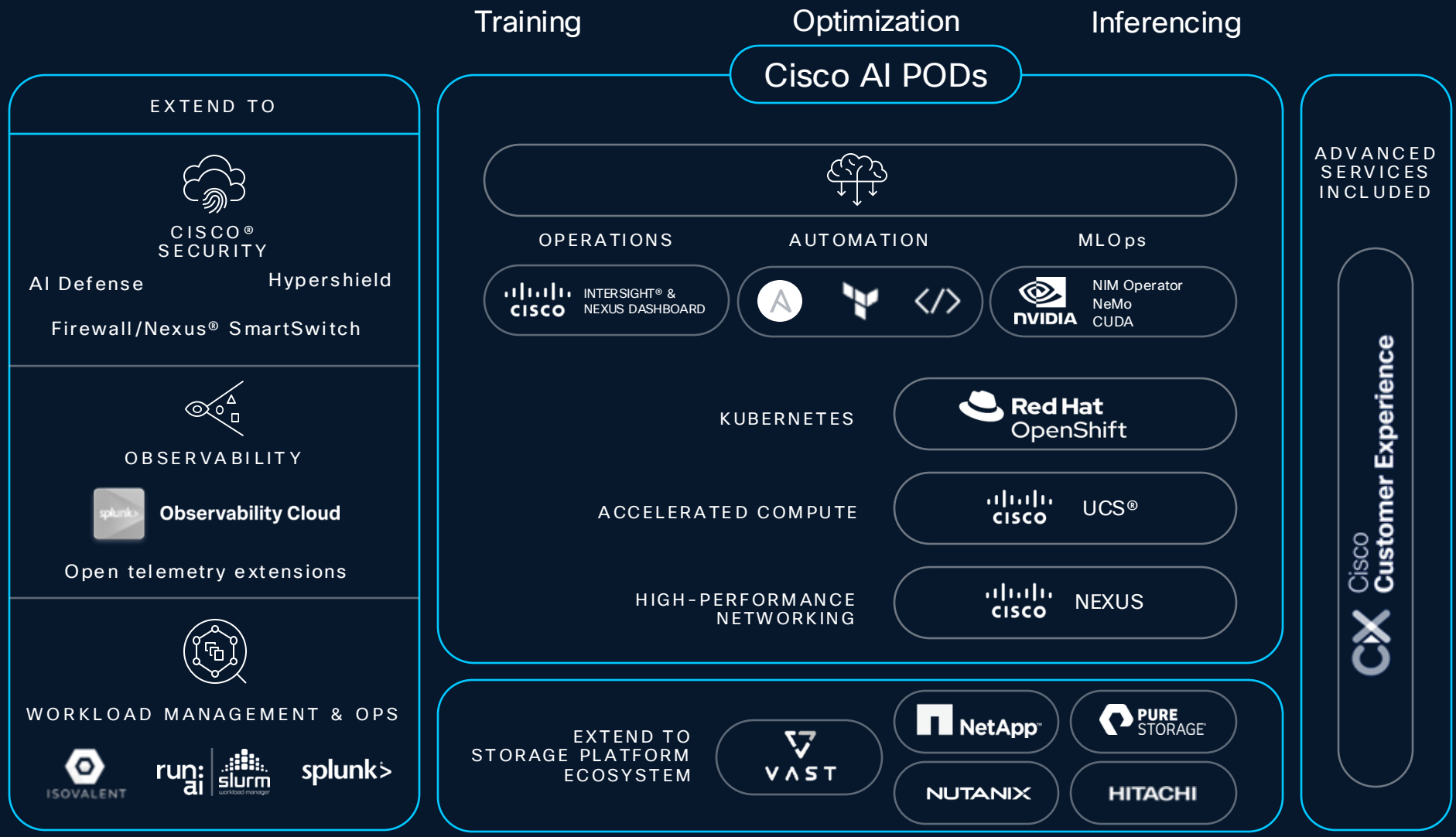
Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

Orderable, use case driven AI-Ready infrastructure stacks

Inferencing.
Optimization.
Training.

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



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Differentiation

The Security

Security-first architecture enables safe enterprise AI

The Infrastructure

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

The Assurance

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

© 2025 Cisco and/or its affiliates. All rights reserved.

Thank you

Making AI work for you.

© 2025 Cisco and/or its affiliates. All rights reserved.