Cisco AI PODs: Pre-validated, Flexible and Modular Infrastructure for Cisco Secure AI Factory

Contents

Value statement

Cisco AI PODs deliver modular, secure, and pre-validated AI infrastructure, accelerating the full AI lifecycle with unmatched flexibility, simplicity, and rapid deployment.

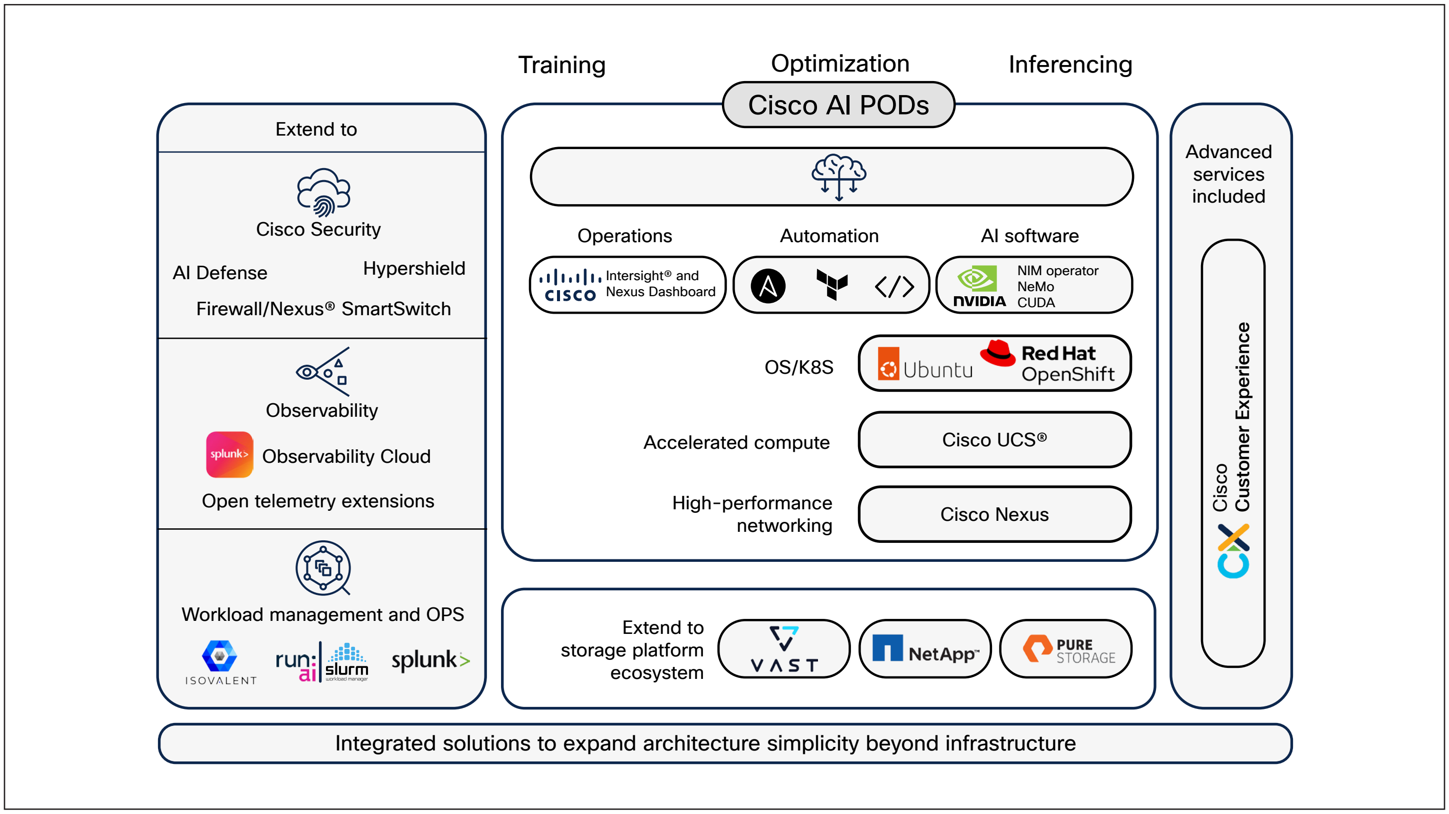
Training
Optimization
Inferencing
Cisco AI PODs
Extend to
Cisco Security
AI Defense
Hypershield
Firewall/Nexus® SmartSwitch
Observability
Observability Cloud
Open telemetry extensions
Workload management and OPS
ISOVALENT
run:ai
slurm
splunk>
Operations
Automation
AI software
Intersight® and Nexus Dashboard
NIM operator
NeMo
CUDA
NVIDIA
OS/K8S
Ubuntu
Red Hat OpenShift
Accelerated compute
Cisco UCS®
High-performance networking
Cisco Nexus
Extend to storage platform ecosystem
VAST
NetApp
PURE STORAGE
Advanced services included
Cisco Customer Experience
Integrated solutions to expand architecture simplicity beyond infrastructure

Figure 1. Cisco AI PODs pre-validated hardware and software stack

Product overview

Cisco AI PODs are the foundation of the Cisco Secure AI Factory, providing a full-stack, modular infrastructure platform for enterprise AI. Developed in partnership with NVIDIA and leading storage vendors, AI PODs combine Cisco UCS compute, Cisco Nexus networking, advanced GPUs, and robust software to support every stage of the AI lifecycle: from training and fine-tuning to high-throughput inferencing.

AI PODs address the key challenges facing enterprises: scaling AI workloads, ensuring data security and compliance, and simplifying infrastructure management. The architecture features high-bandwidth, lossless networking, best-in-class GPU servers, and pre-validated designs to reduce setup time by up to 50 percent and accelerate time to value. Integrated with Cisco Intersight and Nexus Dashboard, AI PODs deliver unified management, automation, and operational visibility.

Flexible deployment options support on-premises, hybrid, or cloud integration, empowering organizations to innovate with AI confidently, securely, and at scale. Whether optimizing large language models, deploying real-time analytics at the edge, or supporting multitenant GPU cloud environments, Cisco AI PODs enable IT and AI teams to move from pilot to production with ease.

Features and benefits

Table 1. AI Pod expansion landscape

Feature	Benefit
Simple and flexible architecture	Enable design of entire AI clusters in a few clicks
Pre-validated, full-stack solutions	Reduces deployment time by up to 50%
Modular, scalable density	Grows seamlessly from 32 to 128+ GPUs per cluster
High-bandwidth, lossless networking	Delivers sub-millisecond latency for demanding AI tasks
Unified management automation	Simplifies operations with Cisco Intersight and Nexus Dashboard
Flexible deployment (on-premises/hybrid)	Supports a wide range of enterprise infrastructure needs
Integrated advanced storage options	Ensures high throughput for AI data pipelines

Prominent feature

Security-first architecture for enterprise AI

Cisco AI PODs uniquely embed enterprise-grade security at every layer. With integrated Cisco AI Defense, Hypershield, and Isovalent Enterprise Platform, organizations can proactively defend against AI-specific threats (for example, prompt injection, adversarial attacks, and unauthorized access) and ensure regulatory compliance and data sovereignty. Security is validated across the full stack, including networking, compute, and software, supporting even the most sensitive workloads.

High-performance, scalable infrastructure

AI PODs leverage platforms such as Cisco UCS C845A and C885A M8 servers with NVIDIA H100/H200 and NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, coupled with Nexus 9000 Series Switches supporting up to 800G bandwidth. This combination delivers lossless, low-latency networking and sustained throughput for AI model training, fine-tuning, and inferencing. The modular scale-unit design allows seamless growth from 32 to 128+ GPUs, making your investments future-ready as AI needs expand.

Deployment simplicity and flexibility

Cisco AI PODs feature pre-validated designs (Cisco Validated Designs [CVDs] and NVIDIA Enterprise Reference Architectures [ERAs]), plug-and-play deployment, and unified management through Cisco Intersight. This reduces setup complexity and accelerates time to value—enabling IT and AI teams to deploy, scale, and manage AI infrastructure with confidence, whether on premises, at the edge, or in hybrid environments.

Platform support

Table 2. Platform support for Cisco AI PODs

Product family	Platforms supported	IOS images (feature sets) supported
Cisco UCS C-Series	C845A M8, C885A M8, C240 M8, C245 M8	NVIDIA AI Enterprise, Red Hat OpenShift, Ubuntu, Rancher
Cisco UCS X-Series	X210c, X215c	Red Hat OpenShift, NVIDIA AI Ent.

Note: The platforms supported will be regularly updated.

Licensing

Cisco AI PODs offer flexible licensing for both hardware and software components.

- **NVIDIA AI Enterprise:** Per-GPU subscription (1-, 3-, 5-year terms)
- **Red Hat OpenShift:** Subscription per node/core (1-, 3-, 5-year)
- **Cisco Intersight:** Essentials/Advantage, per node
- **Optional:** Ubuntu or Rancher OS storage licenses as needed

Table 3. AI PODs licensing summary

Component	License model	Term options	Notes
NVIDIA AI Enterprise	Per GPU	1, 3, 5 years	Education discount available
Red Hat OpenShift	Per node/core	1, 3, 5 years	Optional
Cisco Intersight	Per compute node	1, 3, 5 years	Essentials required, Advantage optional
Storage (for example, VAST)	Per storage node	Varies	See partner data sheets

Note: The licensing summary highlighted above only concerns those software components that are available on Cisco Commerce. Clients can utilize other software licenses by directly purchasing from external vendors.

Product specifications

Table 4. Cisco AI PODs specifications (The example given is an UCS-AIPOD-POD2scale unit .)

Specification	Value/details
GPU density	32, 64, 128+ (expandable scale units)
CPU options	Dual Intel® Xeon® Scalable, AMD EPYC
Memory	Up to 4 TB DDR5 per node
Local storage	NVMe SSD, up to 30.7 TB per node
Networking	400G/800G Nexus 9000 series, RoCEv2, lossless, sub-ms latency
Power supply	80+ platinum, redundancy supported
Management	Cisco Intersight, Nexus Dashboard
Supported OS	RHEL, Ubuntu, Rancher, Red Hat OpenShift
GPU options	NVIDIA H100, H200, RTX Pro 6000, L40S, A100, MI300X
Storage options	VAST Data, NetApp AFF, Pure Storage FlashArray, Nutanix
Security	Cisco Hypershield, AI Defense, Isovalent, SSO, AAA

System requirements

Table 5. Cisco AI PODs system requirements

Feature	Benefit
Disk space	Varies by configuration–minimum 2 TB per node for system ops
Hardware	Cisco UCS C845A/C885A/X-Series/C-Series, Nexus 9000 switches, VAST/NetApp/Pure storage
Memory	512 GB–4 TB per node recommended for training workloads
Software	NVIDIA AI Enterprise, Red Hat OpenShift, Cisco Intersight, supported Linux distro

Ordering information

To order Cisco AI PODs, visit the Cisco Ordering Home Page or contact your Cisco account representative. For configuration and quoting, use the unified AI POD MLB (major line bundle) SKUs.

Table 6. Cisco AI PODs ordering information (sample)

Part #	Product description
UCS-AIPOD-POD1	AI POD for Inferencing/Edge (pre-trained models)
UCS-AIPOD-POD2	AI POD for Training/Fine-Tuning (customizable, scalable)

Warranty information

Cisco AI POD hardware products carry a standard 90-day limited hardware warranty; extended support and service contracts are available.

Product sustainability

Information about Cisco's Environmental, Social and Governance (ESG) initiatives and performance is provided in Cisco's CSR and sustainability reporting.

Cisco AI PODs are engineered for sustainability, supporting customer ESG goals with energy efficiency, recyclable packaging, and materials transparency.

Table 7. Cisco environmental sustainability information

Sustainability Topic		Reference
General	Information on product-material-content laws and regulations	Materials
	Information on electronic waste laws and regulations, including our products, batteries, and packaging	WEEE Compliance
	Information on product takeback and reuse program	Cisco Takeback and Reuse Program
	Sustainability inquiries	Contact: csr_inquiries@cisco.com
Power	Energy consumption	Idle/typical/max values in product specs
	Hardware energy-saving features	Power-efficient PSUs, thermal management
Material	Product packaging weight and materials	Contact: environment@cisco.com Packaging Recyclable materials, reduced foam, high pallet density
	Contact for sustainability inquiries	csr_inquiries@cisco.com

Cisco and partner services

Services from Cisco and our certified partners accelerate deployment and maximize the value of Cisco AI PODs. Cisco CX Success Tracks provide full-stack support, solution adoption services, and ongoing optimization. Mint partners offer pre-sales assessments, proof-of-concept support, and tailored implementation for advanced AI workloads. Learn more at Cisco Services.

Cisco Capital

Flexible payment solutions to help you achieve your objectives

Cisco Capital makes it easier to get the right technology to achieve your objectives, enable business transformation and help you stay competitive. We can help you reduce the total cost of ownership, conserve capital, and accelerate growth. In more than 100 countries, our flexible payment solutions can help you acquire hardware, software, services and complementary third-party equipment in easy, predictable payments. Learn more.

For more information

Accelerate your AI transformation today

Ready to scale AI securely and efficiently? Discover how Cisco AI PODs can help you operationalize AI across your enterprise—from pilot to production.

Request a demo, explore detailed specifications, or contact your Cisco representative at: https://www.cisco.com/site/us/en/solutions/artificial-intelligence/infrastructure/ai-pods.html.

Document history

Table 8. Document history

New or Revised Topic	Described In	Date
Added support for NVIDIA RTX PRO 6000 Blackwell Server Edition GPU	Product specifications	July 2025
Expanded licensing options for OpenShift	Licensing	July 2025
Initial release of AI PODs data sheet	All sections	July 2025

Legal notice

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. A listing of Cisco's trademarks can be found at www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company.

Specifications and product availability subject to change without notice. Cisco reserves the right to make changes to this document and the products described herein at any time, without notice. The information in this document is provided "as is" without warranty of any kind, either expressed or implied.