Accelerate Enterprise AI: Deploy, Scale, and Innovate with Cisco AI PODs

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. We empower organizations to deploy AI infrastructure quickly, scale as needed, and adapt to evolving AI use cases—without the usual complexity.

Overview

Enterprises today face a complex landscape when adopting AI. Traditional infrastructure is often difficult to scale, time-consuming to deploy, and rigid in adapting to evolving requirements. Many organizations are hindered by lengthy integration cycles and high operational overhead, forcing IT and AI teams to stitch together disparate compute, networking, and storage technologies. This results in operational silos, wasted resources, and slow time-to-value, often preventing AI initiatives from moving beyond pilot phases into full production.

Cisco AI PODs address these critical challenges by offering a new model for enterprise AI infrastructure. As pre-validated, full-stack building blocks, AI PODs are purpose-built for simplicity, flexibility, and modular deployment. They integrate Cisco® compute, high-performance networking, and leading GPU/AI platforms into a unified, configurable system. This modular design enables organizations to start small and expand incrementally, while standardized management and automation tools streamline operations from day one, ensuring faster time to value and lower operational risk.

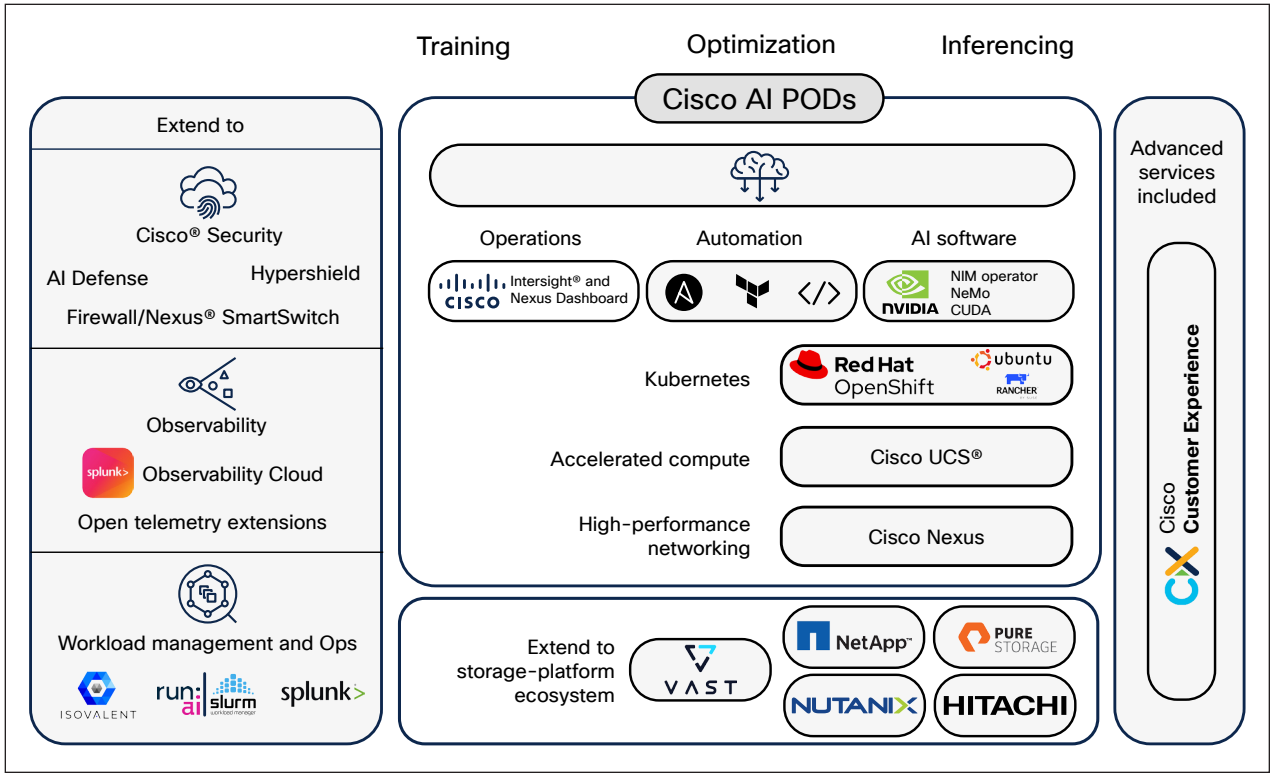
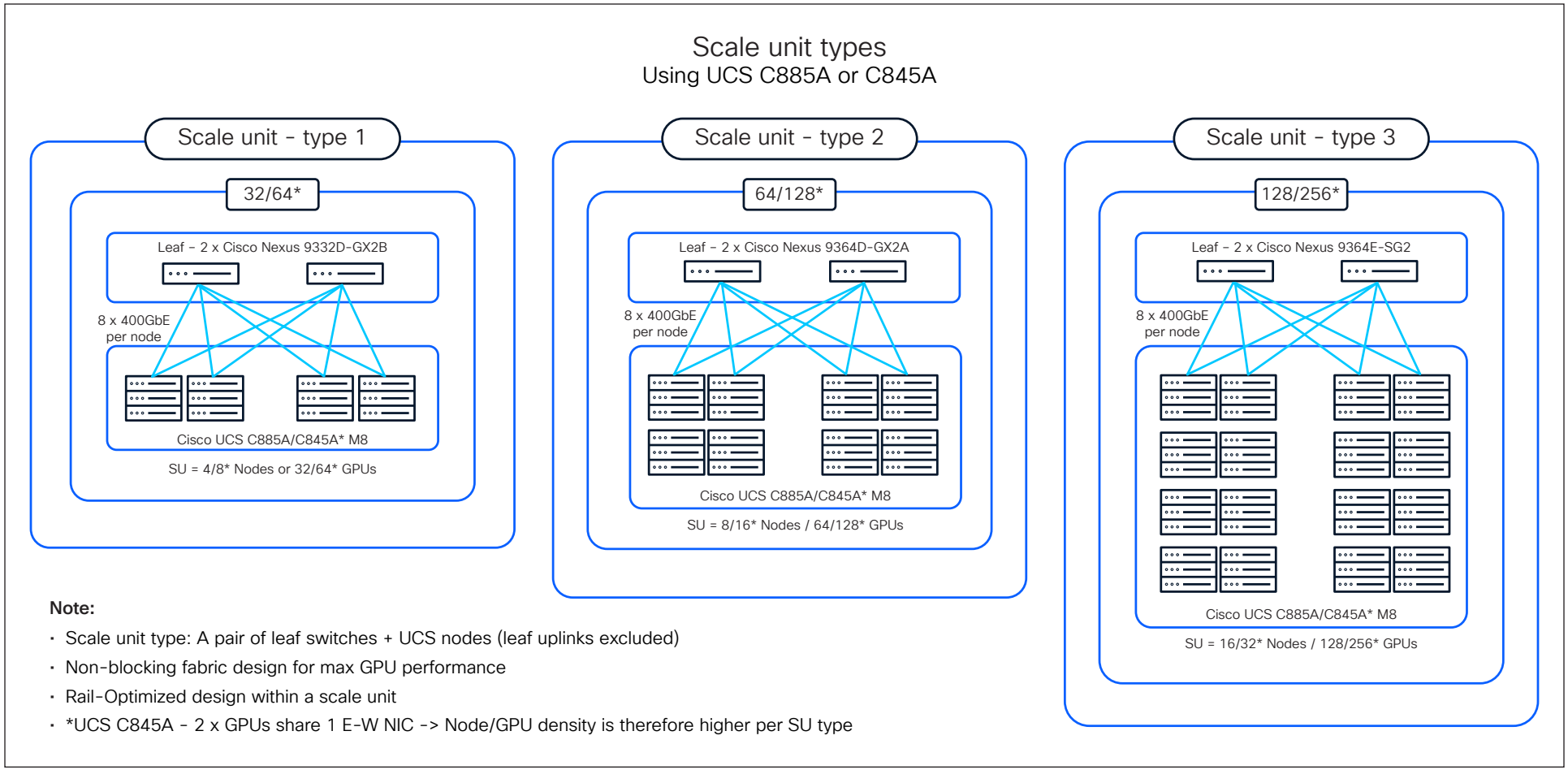


Figure 1. Cisco AI PODs solution overview

"Deploy AI infrastructure in days, not months, with Cisco AI PODs' pre-validated, full-stack solutions."

"Scale seamlessly from 32 to 128+ GPUs, ensuring predictable performance for any AI workload."



Note:

- Scale unit type: A pair of leaf switches + UCS nodes (leaf uplinks excluded)
- Non-blocking fabric design for max GPU performance
- Rail-Optimized design within a scale unit
- *UCS C845A - 2 x GPUs share 1 E-W NIC -> Node/GPU density is therefore higher per SU type

Figure 2. Cisco AI PODs modular scale units

Benefits

Cisco AI PODs provide a robust foundation for AI innovation, delivering tangible benefits:

- Accelerated deployment: Deploy AI-ready infrastructure in days, not months, with pre-integrated, full-stack solutions. Our Cisco Validated Designs (CVDs) and unified management through Cisco Intersight® and Nexus Dashboard drastically reduce setup time by up to 50 percent, minimizing deployment risk and the need for deep technical expertise.
- Superior performance and scalability: Drive AI workload performance with cutting-edge GPUs (from NVIDIA and AMD) and high-bandwidth, lossless networking. The modular scale-unit design allows seamless growth from 32 to 128+ GPUs per cluster, ensuring predictable performance and effortless expansion for diverse AI applications.
- Enterprise-grade security: Uniquely embed enterprise-grade security at every layer. With integrated Cisco AI Defense, Cisco Hypershield, and Isovalent® Enterprise Platform, organizations can proactively defend against AI-specific threats, ensure regulatory compliance, and maintain data sovereignty across the full stack.
- Operational simplicity: Simplify day-to-day management with unified automation and observability. Cisco Intersight and Nexus Dashboard provide centralized control, automated provisioning, and streamlined operations, empowering IT and AI teams to focus on delivering outcomes rather than managing infrastructure.
- Flexible and adaptable: Support any AI workload—from training and fine tuning to high-throughput inferencing—on a single, adaptable platform. Flexible deployment options support on-premises, hybrid, or edge environments, and broad ecosystem compatibility allows integration with preferred storage and software partners.

Trends and challenges

The AI imperative: growth and complexity

The landscape of enterprise technology is rapidly transforming, driven by the explosive growth of artificial intelligence. AI is no longer a futuristic concept but a strategic imperative, with organizations across every industry – from healthcare and finance to manufacturing and retail – seeking to leverage its power for innovation, efficiency, and competitive advantage. The demand for advanced AI capabilities, particularly in such areas as generative AI, Large Language Models (LLMs), and real-time analytics, is skyrocketing. This shift necessitates a robust, scalable, and secure infrastructure that can support the entire AI lifecycle, from data preparation and training to fine tuning and inferencing.

However, this rapid adoption presents significant challenges. Traditional IT infrastructures are often ill-equipped to handle the unique demands of AI workloads, which require immense computational power, high-bandwidth, and ultra-low-latency networking. Enterprises frequently struggle with:

- Infrastructure complexity: Designing, integrating, and deploying AI infrastructure from disparate components is notoriously complex, leading to lengthy project timelines and significant IT resource strain.
- Scaling and performance bottlenecks: Ensuring predictable performance and seamless scalability for diverse AI workloads, especially as they move from pilot to production, is a major hurdle. Inadequate networking can create bottlenecks, hindering GPU-to-GPU communication critical for training.
- Security and compliance: AI models and their underlying infrastructure introduce new attack surfaces and compliance considerations, demanding a security-first approach from the ground up.
- Operational overhead: Managing these complex environments manually results in high operational costs and diverts valuable IT talent from strategic initiatives.

These challenges create a costly gap between AI ambition and business results, limiting innovation and competitive advantage. Organizations need a simplified, pre-validated, and highly performant solution that can accelerate their AI journey without compromising security or operational efficiency.

How it works

Cisco AI PODs are purpose-built, full-stack solutions designed to simplify and accelerate enterprise AI deployments. They combine Cisco's leadership in compute and networking with advanced GPU technologies and integrated software, delivered as pre-validated building blocks.

Core components and architecture

At the heart of Cisco AI PODs are best-in-class hardware components:

- **Cisco UCS compute:** Leveraging high-performance Cisco UCS servers, including the UCS C845A M8 and C240 and C245 M8 (Cisco RTX PRO Servers) and C885A M8 (HGX and OAM servers), alongside and UCS X-Series (X210c and X215c compute nodes) platforms. These servers are optimized for AI workloads, supporting a wide range of CPU options (Intel® Xeon® Scalable and AMD EPYC processors) and up to 4 TB DDR5 memory per node.
- **Advanced GPUs:** AI PODs support the latest NVIDIA GPUs (H100, H200, RTX PRO 6000 Blackwell Server Edition, L40S, L4, and A16) and AMD GPUs (MI300X, MI350X, and MI210), ensuring that customers can power the most demanding AI model training, fine-tuning, and inferencing tasks.

- **Cisco Nexus networking:** High-bandwidth, lossless networking is crucial for AI. AI PODs utilize Cisco Nexus 9000 Series Switches, providing up to 800G bandwidth, RoCEv2 support, and sub-millisecond latency. A leaf-and-spine architecture ensures non-blocking, high-throughput communication, essential for GPU-to-GPU synchronization in training workloads.

Modular scale units for predictable growth
Cisco AI PODs are built on a modular "Scale Unit" (SU) concept, enabling predictable and seamless scalability:

- **SU1:** consists of 4 nodes and supports 32 GPUs
- **SU2:** consists of 8 nodes and supports 64 GPUs
- **SU3:** consists of 16 nodes and supports 128 GPUs

These units form the foundational building blocks, allowing organizations to start small and incrementally expand their AI infrastructure by adding more SUs. While different SU types can coexist within a POD, each individual scale unit is designed for a consistent hardware profile to ensure validated performance.

Full-stack software integration Beyond hardware, AI PODs include a comprehensive software stack for streamlined operations:

- **NVIDIA AI Enterprise (NVAIE):** provides a production-ready software suite for building, deploying, and managing AI workloads
- **Red Hat OpenShift:** a Kubernetes-based container platform that simplifies the orchestration and deployment of containerized AI applications, offering flexibility and scalability
- **Cisco Intersight:** a cloud-based management platform for lifecycle management, providing unified API automation points and server profile templates for cloud-scale operations
- **Cisco Nexus Dashboard:** centralizes configuration, monitoring, analytics, and management for the network fabric, optimizing network performance and streamlining operations using AI/ML fabric templates
- **Automation tools:** integration with Red Hat, Ansible, and Terraform for Infrastructure-as-Code (IaC) automation, enabling automated provisioning and lifecycle management

Security-first architecture: Cisco AI PODs embed enterprise-grade security at every layer. They can be extended with solutions such as Cisco AI Defense, Cisco Hypershield, and Isovalent Enterprise Platform to proactively defend against AI-specific threats (for example, prompt injection and adversarial attacks) and ensure regulatory compliance and data sovereignty.

Flexible Deployment and Storage Options
AI PODs support flexible deployment models, including on-premises, hybrid cloud integration, and edge deployments. They integrate with a broad ecosystem of leading storage vendors, including VAST Data, NetApp AFF, Pure Storage FlashArray, and Nutanix, ensuring high throughput for demanding AI data pipelines.

Cisco Interconnects: Tested and Trusted for AI PODs

Cisco interconnects power AI PODs with the latest high-performance specifications, combining ultra-low latency, high bandwidth, and advanced fabric technologies. Each deployment undergoes rigorous end-to-end validation, ensuring predictable performance, seamless scalability, and enterprise-grade reliability—so organizations can build AI infrastructure with confidence.

Services

Maximize AI success with Cisco CX AI lifecycle services

Cisco Services accelerates your AI journey with tailored support across the AI lifecycle. Our experts assess your infrastructure, design and deploy Cisco AI PODs, and optimize performance with continuous monitoring and management. From planning to production, we ensure rapid deployment, seamless integration, and operational efficiency, thereby minimizing risks and maximizing ROI. Contact your Cisco Services representative to customize a service plan that aligns with your AI goals.

Use cases

Table 1. Use cases

AI lifecycle stage	Technical use case examples
Data preparation	Clean and preprocess large datasets for AI training using automated data pipelines.
Model training	Train large-scale generative AI models for Natural Language Processing (NLP) or vision with high-performance computing.
Model evaluation	Evaluate AI model performance using automated testing and validation on diverse datasets.
Model deployment	Deploy real-time AI inferencing for customer service chatbots in retail environments.
Model monitoring	Monitor AI model drift in fraud detection systems with real-time performance analytics.
Data annotation	Automate image labeling for medical diagnostics using AI-driven annotation tools.
Model fine-tuning	Fine-tune LLMs with domain-specific datasets and optimization.
Inference optimization	Optimize low-latency AI inferencing for autonomous vehicle decision-making systems.

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.

Industry name	Use case examples by vertical
Healthcare	Accelerate drug discovery with AI-driven molecular modeling and patient-data analysis.
Financial services	Enhance fraud detection using real-time AI inferencing on transaction data.
Manufacturing	Optimize supply chains with predictive analytics and real-time demand forecasting.
Retail	Personalize customer experiences through RAG-based conversational AI and recommendations.
Automotive	Train autonomous driving models with large-scale data processing and simulation.
Life sciences	Streamline genomic sequencing with high-performance AI training and analysis.
Government	Improve public safety with AI-powered video surveillance and threat detection.
Telecommunications	Enhance network optimization with AI-driven traffic analysis and predictive maintenance.

Learn more

Start your AI journey with confidence

Ready to simplify and accelerate your AI deployments? Discover how Cisco AI PODs can transform your infrastructure. Visit https://www.cisco.com/site/us/en/solutions/artificial-intelligence/infrastructure/ai-pods.html to explore solutions or contact your Cisco Services representative for a personalized assessment.

Healthcare provider accelerates AI-driven diagnostics

Challenge: A global healthcare provider needed to scale AI for medical imaging analysis but faced long deployment times and integration challenges.

Solution: The healthcare provider deployed Cisco AI PODs (SU2) with NVIDIA GPUs, Cisco Nexus 9000 switches, and Red Hat OpenShift for containerized AI workloads.

Benefits:

- Reduced deployment time by 60 percent, enabling rapid diagnostics
- Scaled from 64 to 128 GPUs seamlessly, supporting larger datasets
- Enhanced security with Cisco AI Defense, ensuring HIPAA compliance

Financial firm boosts fraud detection with AI PODs

Challenge: A financial institution struggled with real-time fraud detection due to legacy infrastructure bottlenecks.

Solution: The institution implemented Cisco AI PODs (SU1) with NVIDIA H100 GPUs, integrated with Cisco Intersight for automated management.

Benefits:

- Achieved sub-millisecond inferencing for real-time fraud detection
- Reduced operational overhead by 40 percent with unified automation
- Ensured compliance with Cisco Hypershield security integration

The Cisco Advantage

Cisco's leadership in networking, compute, and security, combined with our AI-optimized PODs, delivers unmatched simplicity and performance. Our pre-validated, modular designs and ecosystem partnerships (including NVIDIA, AMD, Red Hat, VAST, and others) ensure rapid deployment and scalability. With integrated security and automation, Cisco empowers you to innovate confidently, bridging the gap between AI ambition and business success.

C22-5394481-00 09/25