Cisco AI PODs

Benefits

- Accelerate AI adoption and deployment by leveraging pre-validated, full-stack infrastructure and unified management, drastically reducing setup time.
- Drive superior AI workload performance with cutting-edge GPUs and high-bandwidth networking, achieving faster training and seamless scalability for diverse AI applications.
- Ensure enterprise-grade security and compliance with robust data protection and sovereignty while also achieving sustainability goals through energy-efficient designs.
- Empower AI developers to build and deploy models faster with seamless integration of leading software stacks and robust APIs.

Overview

Companies around the world, in every industry, are keen to leverage AI to transform their business, improve customer satisfaction, and gain a competitive advantage. Deploying enterprise AI applications, especially generative AI, is a complex process that requires careful planning, evaluation of models and infrastructure, and execution. Enterprises often face challenges when scaling AI infrastructure, ensuring data privacy and security, and bridging the skills gap. Cisco AI PODs simplify and accelerate full AI lifecycle deployment with pre-validated, high-performance infrastructure, empowering developers and IT Ops to innovate sustainably in the Cisco Secure AI Factory with Nvidia. Cisco can help you to right-size your investment in AI-related infrastructure while balancing current business and IT needs, with a view to scalability in the future.

What is the full AI lifecycle?

The full AI lifecycle encompasses several critical stages:

- **On-premises training:** This involves developing and training AI models, such as Large Language Models (LLMs) or deep neural networks, using local infrastructure to process large datasets and optimize model parameters. It is compute-intensive and requires high-performance hardware and robust data management.
- **Model optimization (fine-tuning, RAG):** This stage refines pre-trained AI models to improve accuracy or adapt them for specific tasks (fine-tuning) and enhances LLMs with real-time data retrieval (Retrieval-Augmented Generation, RAG). Fine-tuning adjusts model weights, and RAG integrates external knowledge bases for context-aware responses.
- **Model inferencing:** Inferencing involves deploying trained AI models to generate predictions or decisions, either at scale in data centers or in real time at distributed-edge locations. Large-scale inferencing handles high-volume workloads, and edge inferencing prioritizes low latency for time-sensitive applications.

Cisco AI PODs solution

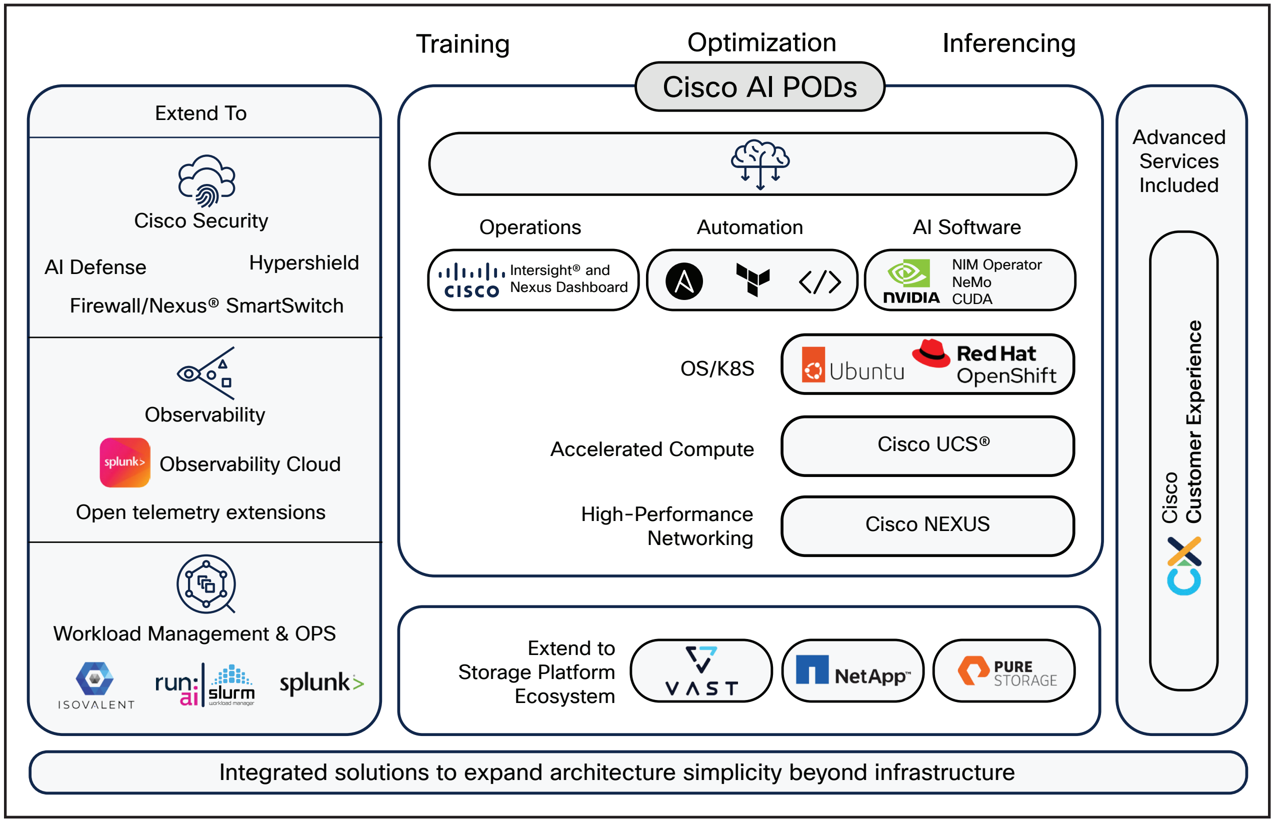

Figure 1. Cisco AI PODs Solution

Cisco AI PODs, initially launched for inferencing solutions, are expanding to support the full technical lifecycle of AI deployments, including training, fine-tuning, and inferencing. They are based on validated architectural stacks and fabric designs, providing a blueprint for deployment that is referenceable and essentially “off the shelf.” They include software tools and components, with extensions for observability (Splunk®) and security (AI Defense, Hypershield), making them orderable as part of one ecosystem with a consolidated support model.

Why Cisco® AI PODs?

- **Simplified deployment:** Pre-validated designs (Cisco Validated Designs, or CVDs) and Cisco Intersight reduce setup time by up to 50 percent, enabling AI developers to focus on innovation and ITOps to streamline operations.
- **High performance and scalability:** Powered by NVIDIA, AMD GPUs, and Cisco high-performance networking, AI PODs accelerate training by up to 30 percent and scale inference seamlessly, meeting diverse workload demands.
- **Enterprise-grade security:** Hypershield and Cisco AI Defense ensure compliance and data sovereignty, giving I Ops peace of mind and developers a trusted platform.
- **Developer empowerment:** Integration with RedHat Openshift, PyTorch, TensorFlow, and NVIDIA AI Enterprise empowers AI developers to build and deploy models faster, supported by robust APIs and tutorials.
- **Sustainability and cost-effectiveness:** Energy-efficient designs cut power costs, maximizing ROI for ITOps while aligning with environmental, social, and governance (ESG) goals.

What it does

Cisco has been developing and providing validated designs for more than 20 years. Cisco Validated Designs (CVDs) are comprehensive, rigorously tested guidelines that help customers deploy and manage IT infrastructure effectively. They include detailed implementation guides, best practices, and real-world use cases, often incorporating Cisco technology partner products. CVDs reduce deployment risk, optimize performance, and ensure scalability, all while being supported by the Cisco Technical Assistance Center (Cisco TAC). This support and integration provide customers with a reliable and efficient path to achieving their business objectives.

Cisco AI PODs for the full AI lifecycle are CVD-based solutions for inferencing, model optimization, and training, covering edge and core deployments. They provide accelerated deployment with centralized management and automation. The solution has been performance-tested and demonstrates linear scalability through benchmark tests on real-life model simulation, showcasing consistent performance even with varying dataset sizes. Cisco AI PODs have independent scalability at each layer of infrastructure and are perfect for data-center and edge AI deployments. They are designed to fit customer cost-models and are easy to manage, with use of Cisco smart switches for demarcation and segmentation in modern data centers.

Regardless of the configuration, they all contain:

- Cisco UCS C845A (PCIE GPU) and 885A (HGX&OAM) M8, and X-Series servers.
- Cisco Nexus 9000 Switches.
- Cisco Intersight (SaaS management).
- NVIDIA AI Enterprise.
- RedHat OpenShift.
- Ansible, Terraform Automation.
- Cisco CX Services.

Optional storage is also available from NetApp, Pure Storage, and VAST Data. These partners provide data-platform solutions to help developers and data scientists perform numerous data management tasks.

Learn more

- For more in-depth information on the Cisco AI PODs, refer to the data sheet.
- For information on all Cisco AI-native infrastructure for the data center, visit Cisco.com.
- For more information on the Cisco UCS X-Series Modular System, visit https://www.cisco.com/go/ucsx.
- For more information on the Cisco UCS C-Series Rack Servers for AI, visit https://www.cisco.com/site/us/en/products/computing/servers-unified-computing-systems/ucs-c-series-rack-servers/index.html.

For more information

Book an expert consultation to start your AI-ready infrastructure journey

Receive expert guidance on modernizing your network and compute infrastructure with AI-ready infrastructure—combining technologies, products, and Cisco Validated Designs to support and scale AI workloads, all while advancing sustainability initiatives.

Get Started

Our experts recommend BIOS Performance and workload: Tuning guide for Cisco UCS M8 Platforms White Paper.

Cisco UCS Servers with Intel® Xeon® 6 CPUs FAQ.