AI PODs for Inferencing

Contents

Cisco AI-ready Infrastructure Stacks offers tailored pods for performance and scalability, from cost-effective development and low-intensity tasks to advanced AI and enterprise applications, ensuring optimal productivity and growth.

In short, Cisco provides Scalable, Efficient AI Pods for Every Need.

Product overview

Cisco's AI-ready Infrastructure Stacks provide a comprehensive suite of solutions tailored for various AI and ML use cases. These validated designs simplify and automate AI infrastructure, enabling organizations to deploy AI models efficiently across different environments—from edge inferencing to large-scale data centers. Cisco's infrastructure stacks cater to a range of requirements:

- MLOps with Red Hat OpenShift AI to streamline machine learning operations and model deployment.
- Gen AI Modeling utilizing Intel Gaudi 2 processors and Nexus 9000, optimized for high-performance AI tasks.
- Data Modeling platforms powered by Cisco's robust networking and compute capabilities.
- Digital Twin frameworks, supported by NVIDIA AI Enterprise and integrated data systems.

Each solution is built on Cisco's validated designs, ensuring seamless integration and reliable performance for AI applications. Organizations can choose from converged infrastructure like FlashStack and FlexPod or extend their deployments with hyperconverged solutions like Nutanix.

Customers across different segments, from small businesses to large enterprises, will find immense value in Cisco AI Infrastructure. Small businesses can utilize cost-effective pods for development and basic workloads, while larger enterprises can harness advanced pods for intensive AI and data processing tasks. The primary business benefits include improved productivity, reduced costs, and enhanced scalability, while the primary technical value lies in the pods' ability to deliver high performance and reliability, ensuring seamless integration and operation within existing network infrastructures.

Features and benefits

Table 1. Features and benefits

Feature	Benefit
Scalability on Demand	Dynamically scale resources to meet evolving AI/ML workloads with configurable compute, memory, and accelerators tailored to deployment sizes.
Future Ready	Modular, future-proof system supporting next-generation processors, GPUs, and storage solutions, ensuring long-term investment protection.
Customization	Tailor configurations to specific AI models and workloads for optimal performance and efficiency across edge, medium, large, and scale-out deployments.
Comprehensive Platform	Integrated with Cisco Intersight and Nexus Dashboard, the AI PODs simplify infrastructure management with centralized provisioning, real-time visibility, and automated scaling.
Simplified Management	Streamline management with Cisco Intersight, offering centralized provisioning, real-time visibility, and automation for Cisco UCS servers and GPUs, reducing operational complexity.

Feature	Benefit
Flexible Configurations	Choose from multiple AI POD sizes, from small edge inferencing setups to large, multi-model deployments, to meet specific workload requirements.

Prominent feature

Transforming AI Architecture Design with Cisco's Full Stack Solution

This solution combines the best of hardware and software technologies to create a robust, scalable, and efficient AI-ready infrastructure tailored to diverse needs.

1. **NVIDIA AI Enterprise and NVIDIA Infrastructure Management Suite (NIMS):** These components provide a powerful foundation for managing AI workloads. They offer optimized performance, streamlined operations, and advanced management capabilities, enabling users to focus on developing and deploying AI models without worrying about underlying infrastructure complexities.
2. **OpenShift by Red Hat:** This Kubernetes-based container platform simplifies the orchestration and deployment of containerized AI applications. It provides flexibility and scalability, allowing users to easily manage and scale their AI projects as they grow.
3. **Cisco Unified Computing System (UCS):** Cisco UCS serves as the hardware backbone, delivering high performance and reliability with its multiple server configurations. This ensures that the compute power required for demanding AI workloads is readily available and scalable.
4. **Converged Infrastructure Solutions:**
    - FlashStack (Cisco UCS servers with Pure Storage) and FlexPod (Cisco UCS servers with NetApp storage) both provide high-performance storage and compute capabilities essential for handling large-scale AI and machine learning tasks. These solutions offer integrated, pre-validated configurations that simplify deployment and management, enhancing operational efficiency and reducing time to market.

By leveraging Cisco's full stack solution, customers benefit from:

- **Easy to Buy:**
    - Cisco AI PODs are available as pre-configured, validated bundles that are ready to order, ensuring that customers can quickly select the right setup for their specific AI needs. This eliminates the complexity of choosing individual components and reduces decision-making time.
    - By offering orderable bundles this quarter, Cisco enables faster time-to-value for customers, allowing them to accelerate their AI initiatives without unnecessary delays.
- **Easy to Deploy and Integrate:**
    - Cisco AI PODs are designed for plug-and-play deployment, ensuring that organizations can rapidly integrate them into existing data center or cloud environments with minimal effort. This seamless deployment capability is supported by Cisco's Intersight and Nexus Dashboard, providing centralized management, real-time visibility, and automation to reduce operational complexity.
    - The pre-configured bundles are pre-tested and validated, ensuring that all components work harmoniously together right out of the box. This reduces the risk of deployment issues and speeds up time-to-production for AI workloads.

- Whether deployed in a converged infrastructure setup like FlashStack and FlexPod or in a hyperconverged environment like Nutanix, Cisco AI PODs offer flexibility and scalability. This ensures easy integration into a wide range of existing infrastructure setups, providing a future-ready solution that can grow with organizational needs.

- **Easy to Manage:**
  - The pre-integrated and pre-tested nature of the solution reduces the complexity of designing AI infrastructure. It provides a clear, replicable formula for optimal performance, eliminating guesswork.
  - The solution supports a range of deployment sizes from small edge inferencing setups to large, high-performance environments. This scalability ensures that the infrastructure can grow with the organization's needs.
  - Integrated management tools from NVIDIA and Red Hat simplify the day-to-day operations of AI infrastructure, making it easier for IT teams to support AI projects.

For individuals and organizations at the ideation stage, Cisco's full stack solution provides a clear, structured path to developing a robust AI architecture. It combines the latest in AI management, container orchestration, and high-performance hardware, making it an ideal choice for anyone looking to build and scale their AI capabilities.

Scalability on Demand

Requirements change often, and you need a system that doesn't lock you into one set of resources when you find that you need another. Cisco AI-ready Infrastructure Stacks are by nature growth ready and seamlessly scales with your Generative AI inferencing needs because of the modular nature of the Cisco UCS X-Series. It's as simple as adding or removing servers, adjusting memory capacities, and configuring resources in an automated manner as your models evolve and workloads grow using Cisco Intersight®. Additionally, you even have the flexibility to vary the CPU to GPU ratio or choose between Intel Xeon Scalable or AMD EPYC processors within a single chassis, depending on your specific use case.

- **Dynamic Resource Allocation:** Easily add or reduce compute and storage resources.
- **Flexible CPU/GPU Ratios:** Customize the balance of compute power.
- **Automated Scaling:** Adjust resources automatically as workloads grow.

Future Ready

The Cisco AI-ready Infrastructure Stacks are built to be future ready, evolving with new technologies and adapting to emerging business objectives. The UCS X-Series modular system is designed to support future generations of processors, storage, nonvolatile memory, accelerators, and interconnects. It can be selectively upgraded piece by piece without the need to purchase, configure, maintain, power, and cool discrete management modules and servers. Cloud-based management is kept up to date automatically with a constant stream of new capabilities delivered by the Intersight software-as-a-service model. The resultant extended lifecycles for hardware will lower ongoing costs over time.

- **Modular Design:** Easily upgrade components without major overhauls.
- **Longevity:** Support for upcoming hardware generations.
- **Investment Protection:** Safeguard your technology investments with forward-compatible infrastructure.

Customization

Cisco's customizable configurations allow you to host models of your choice, edit configurations, and connect to diverse data sources and services. This flexibility ensures optimal performance and control over your AI workloads, tailored to your specific needs and deployment sizes.

- **Model Flexibility:** Host any AI model to meet your unique requirements.
- **Configurable Setups:** Modify hardware setups as needed.
- **Comprehensive Control:** Manage and control your deployment environment fully.

Comprehensive Platform

Operationalizing AI is a daunting task for any Enterprise. One of the biggest reasons AI/ML efforts fail to move from proof-of-concept to production is due to the complexity associated with streamlining and managing data and machine learning that deliver production-ready models. Instead of ad-hoc ML efforts that add technical debt with each AI/ML effort, it is important to adopt processes, tools, and best-practices that can continually deliver and maintain models with speed and accuracy. Cisco AI-ready Infrastructure Stacks provide a purpose-built, full stack solution that accelerates AI/ML efforts by reducing complexity. The design uses the Cisco UCS X-Series modular platform with the latest Cisco UCS M7 servers, Cisco UCS X440p PCIe nodes with NVIDIA GPUs, all centrally managed from the cloud using Cisco Intersight. Building on this accelerated and high-performance infrastructure, Red Hat OpenShift AI enhances the solution by integrating essential tools and technologies to accelerate and operationalize AI consistently and efficiently. NVIDIA AI Enterprise software further complements the solution, offering key capabilities such as virtual GPUs, a GPU operator, and a comprehensive library of tools and frameworks optimized for AI. These features simplify the adoption and scaling of AI solutions, making it easier for enterprises to operationalize AI technologies.

- **End-to-End Solution:** From infrastructure to AI frameworks.
- **Generative AI Support:** Optimized for AI model deployment at scale.
- **Enterprise-Grade:** Robust enough to meet large-scale enterprise demands.

Simplified Management

Optimize your infrastructure management with Cisco Intersight. Our platform offers centralized provisioning, real-time visibility, and automation, simplifying the management of UCS servers and GPUs. This streamlines operations and reduces manual intervention, enhancing efficiency and reducing complexity.

- **Centralized Management:** Manage all resources from a single interface.
- **Real-Time Visibility:** Monitor performance and status in real-time.
- **Automation:** Reduce manual intervention with automated management processes.

Product specifications

The Cisco AI Data Center and Edge Inference Pod is expertly engineered for edge inferencing applications, facilitating computation directly near the user at the network's edge, close to the data source. This strategic design minimizes latency and maximizes efficiency by processing data locally, rather than relying on a centralized cloud or data center. Supporting advanced models like Llama 2-7B, GPT-2B, and other Small Language Models (SLMs), the Data Center and Edge Inference Pod is highly versatile and capable. The inclusion of the integrated X-Series Direct fabric interconnect within the chassis eliminates the need for additional hardware, thereby reducing complexity and streamlining operations.

Table 2. Specifications for Cisco AI Data Center and Edge Inference Pod

Item	Specification
Compute Node	1 Cisco UCS X210c M7 compute node
Processors	Dual Intel 5th Generation 6548Y+ processors for each compute node
Memory	8x 64 GB DDR5 at 4800 MT/s for a total of 512 GB of memory per compute node, or 512GB in total
Internal Storage	2x 1.6 TB Non-volatile Memory Express (NVMe) 2.5-inch drives per compute node
mLOM	1x Cisco UCS VIC 15230 2x100G mLOM
PCIe Node	1x Cisco UCS X440p PCIe Node per compute node
GPU	1x Nvidia L40s GPU (dual slot) per compute node
Management	• Cisco Intersight software (SaaS, virtual appliance, and private virtual appliance) • Cisco UCS Manager (UCSM) 4.3(2) or later

The Cisco AI RAG Augmented Inference Pod is purpose-built for demanding AI workloads, supporting larger models such as Llama 2-13B and OPT 13B. Equipped with enhanced GPU and node capacity, it efficiently manages complex tasks, delivering increased computational power and efficiency. This advanced pod can accommodate optimizations like Retrieval-Augmented Generation (RAG), which leverages knowledge sources to provide contextual relevance during query service, significantly reducing hallucinations and improving response accuracy. This makes the Cisco RAG Augmented Inference Pod an ideal choice for advanced applications requiring robust AI capabilities and exceptional performance.

Table 3. Specifications for Cisco AI RAG Augmented Inference Pod

Item	Specification
Compute Node	2x Cisco UCS X210c M7 compute node
Processors	Dual Intel 5th Generation 6548Y+ processors for each compute node, or 4 CPUs in total
Memory	16x 64 GB DDR5 (4 DIMMS per CPU) at 4800 MT/s for a total of 512 GB of memory per compute node, or 1TB in total
Internal Storage	2x 1.6 TB Non-volatile Memory Express (NVMe) 2.5-inch drives per compute node
mLOM	1x Cisco UCS VIC 15230 2x100G mLOM per compute node
PCIe Node	1x Cisco UCS X440p PCIe Node per compute node
GPU	2x Nvidia L40S GPU (dual slot) per compute node, or 4 GPUs in total
Management	• Cisco Intersight software (SaaS, virtual appliance, and private virtual appliance) • Cisco UCS Manager (UCSM) 4.3(2) or later

The Cisco AI Scale Up Inference Pod for High Performance is meticulously optimized to support large-scale models like CodeLlama 34B and Falcon 40B. With its substantial GPU and node capacity, this pod is engineered to deliver exceptional performance for the most complex AI tasks. It significantly enhances Retrieval-Augmented Generation (RAG) capabilities by supporting larger vector databases and improving vector search accuracy through advanced algorithms.

Table 4. Specifications for Cisco AI Scale Up Inference Pod for High Performance

Item	Specification
Compute Node	2x Cisco UCS X210c M7 compute node
Processors	Dual Intel 5th Generation 6548Y+ processors for each compute node, or 4 CPUs in total
Memory	16x 64 GB DDR5 (4 DIMMS per CPU) at 4800 MT/s for a total of 512 GB of memory per compute node, or 1TB in total
Internal Storage	2x 1.6 TB Non-volatile Memory Express (NVMe) 2.5-inch drives per compute node
mLOM	1x Cisco UCS VIC 15230 2x100G mLOM per compute node
PCIe Node	1x Cisco UCS X440p PCIe Node per compute node
GPU	2x Nvidia H100 GPU (dual slot) per compute node, or 4 GPUs in total
Management	• Cisco Intersight software (SaaS, virtual appliance, and private virtual appliance) • Cisco UCS Manager (UCSM) 4.3(2) or later

The Cisco AI Scale Out Inference Pod for Large Deployment is designed to offer unparalleled flexibility, allowing multiple models to run concurrently within a single chassis. It is ideal for organizations that require robust lifecycle management or high availability of models. This pod enhances accuracy by allowing the optimal model to be selected for each specific task and effortlessly scales out to support multiple users. This makes the pod an exceptional choice for organizations seeking scalable, efficient, and reliable AI infrastructure to handle diverse and demanding AI workloads.

Table 5. Specifications for Cisco AI Scale Out Inference Pod for Large Deployment

Item	Specification
Compute Node	4x Cisco UCS X210c M7 compute node
Processors	Dual Intel 5th Generation 6548Y+ processors for each compute node, or 8 CPUs in total
Memory	64x 64 GB DDR5 (8 DIMMS per CPU) at 4800 MT/s for a total of 1 TB of memory per compute node, or 4TB in total
Internal Storage	2x 1.6 TB Non-volatile Memory Express (NVMe) 2.5-inch drives per compute node
mLOM	1x Cisco UCS VIC 15230 2x100G mLOM per compute node
PCIe Node	1x Cisco UCS X440p PCIe Node per compute node
GPU	2x Nvidia L40S GPU (dual slot) per compute node, or 8 GPUs in total
Management	• Cisco Intersight software (SaaS, virtual appliance, and private virtual appliance) • Cisco UCS Manager (UCSM) 4.3(2) or later

System requirements

Table 6. System requirements

Feature	Description
Disk Space	Minimum 1TB SSD per compute node for AI model storage and operations
Hardware	• 1x X210C Compute Node, 1x X440p PCIe, 1x Nvidia L40s for Data Center and Edge Inference Pod • 2x X210C Compute Nodes, 2x X440p PCIe, 4x Nvidia L40s for RAG Augmented Inference Pod • 2x X210C Compute Nodes, 2x X440p PCIe, 4x Nvidia H100 for Scale Up Inference Pod • 4x X210C Compute Nodes, 4x X440p PCIe, 8x Nvidia L40s for Scale Out Inference Pods
Memory	Minimum 128GB RAM per compute node, expandable based on workload requirements
Software	Cisco Intersight for centralized management
Network	High-speed network connectivity (minimum 10GbE) for data transfer between nodes and storage
Power and Cooling	Adequate power supply and cooling system to support high-performance compute nodes and GPUs

Feature	Description
Operating System	Compatible with Red Hat OpenShift AI for MLOps and AI/ML workloads
GPU Support	Nvidia GPU drivers and software (NVIDIA AI Enterprise) for optimal performance and management
Security	Secure boot, encryption, and access controls for data protection
Integration	Seamless integration with existing data center infrastructure and cloud services
Chassis	Cisco UCS X9508 Server Chassis
Fabric Interconnect	Cisco UCS 6454, 64108, and 6536 fabric interconnects
Cisco Intersight	Intersight Managed Mode (minimum Essentials license per server)

Ordering information

Table 7. Ordering Guide for Cisco AI Data Center and Edge Inference Pod

Item	Part #
Compute Node	UCSX-210C-M7
Chassis	UCSX-9508-ND-U
Processors	UCSX-CPU-I6548Y+
Memory	UCSX-MRX64G2RE1
Virtual Interface Card	UCSX-MLV5D200GV2D
mLOM	UCSX-S9108-100G
PCIe Node	UCSX-RIS-A-440P-D
GPU	UCSX-GPU-L40S
Fabric Interconnect	
XFM 2.0	UCSX-F-9416
M.2 Drive	UCSX-NVME4-1600-D

Table 8. Ordering Guide for Cisco AI RAG Augmented Inference Pod

Item	Part #
Compute Node	UCSX-210C-M7
Chassis	UCSX-9508-ND-U
Processors	UCSX-CPU-I6548Y+
Memory	UCSX-MRX64G2RE1
Virtual Interface Card	UCSX-MLV5D200GV2D
mLOM	UCSX-I9108-100GND
PCIe Node	UCSX-RIS-A-440P-D
GPU	UCSX-GPU-L40S
Fabric Interconnect	UCSX-FI-6536-ND-U
XFM 2.0	UCSX-F-9416
M.2 Drive	UCSX-NVME4-1600-D

Table 9. Ordering Guide for Cisco AI Scale Up Inference Pod for High Performance

Item	Part #
Compute Node	UCSX-210C-M7
Chassis	UCSX-9508-ND-U
Processors	UCSX-CPU-I6548Y+
Memory	UCSX-MRX64G2RE1
Virtual Interface Card	UCSX-MLV5D200GV2D
mLOM	UCSX-I9108-100GND
PCIe Node	UCSX-RIS-A-440P-D
GPU	UCSX-GPU-H100-80
Fabric Interconnect	UCSX-FI-6536-ND-U
XFM 2.0	UCSX-F-9416
M.2 Drive	UCSX-NVME4-1600-D

Table 10. Ordering Guide for Cisco AI Scale Out Inference Pods for Large Deployment

Item	Part #
Compute Node	UCSX-210C-M7
Chassis	UCSX-9508-ND-U
Processors	UCSX-CPU-I6548Y+
Memory	UCSX-MRX64G2RE1
Virtual Interface Card	UCSX-MLV5D200GV2D
mLOM	UCSX-I9108-100GND
PCIe Node	UCSX-RIS-A-440P-D
GPU	UCSX-GPU-L40S
Fabric Interconnect	UCSX-FI-6536-ND-U
XFM 2.0	UCSX-F-9416
M.2 Drive	UCSX-NVME4-1600-D

Warranty information

The Cisco UCS X210c Compute Node has a three-year Next-Business-Day (NBD) hardware warranty and a 90-day software warranty.

Augmenting the Cisco Unified Computing System™ (Cisco UCS) warranty, Cisco Smart Net Total Care® and Cisco Solution Support services are part of Cisco's technical services portfolio. Cisco Smart Net Total Care combines Cisco's industry-leading and award-winning foundational technical services with an extra level of actionable business intelligence that is delivered to you through the smart capabilities in the Cisco Smart Net Total Care portal. For more information, please refer to https://www.cisco.com/c/en/us/support/services/smart-net-total-care/index.html

Cisco Solution Support includes both Cisco® product support and solution-level support, resolving complex issues in multivendor environments on average 43 percent more quickly than with product support alone. Cisco Solution Support is a critical element in data center administration, helping rapidly resolve issues encountered while maintaining performance, reliability, and return on investment.

This service centralizes support across your multivendor Cisco environment for both our products and solution partner products that you have deployed in your ecosystem. Whether there is an issue with a Cisco product or with a solution partner product, just call us. Our experts are the primary point of contact and own the case from first call to resolution. For more information, please refer to https://www.cisco.com/c/en/us/services/technical/solution-support.html.

Product sustainability

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

Table 11. 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
Material	Product packaging weight and materials	Contact: environment@cisco.com

Cisco and partner services

Cisco and partner services for AI-ready Infrastructure

Cisco and our certified partners provide comprehensive services to support your AI Infrastructure, ensuring seamless integration and optimal performance. Our planning and design services align technology with your business goals, increasing deployment accuracy and efficiency. Technical services enhance operational efficiency, reduce costs, and mitigate risks. Optimization services continuously improve performance and help your team succeed with new technologies. For more details, please visit our Cisco AI Services page.

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

Unlock the Future of AI with Cisco AI Pods

Discover how Cisco AI-ready Infrastructure can revolutionize your business. Request a demo, sign up for a free assessment, or contact our virtual sales rep to explore our tailored solutions. For more insights, visit our Cisco AI Solutions Overview page and access white papers, case studies, and more. Ensure your business stays ahead with cutting-edge technology designed to optimize productivity and growth.

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA C78-4662163-00 09/24