AI PODs for Inferencing

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 a generative AI application is a complex process that requires careful planning, evaluation of models and infrastructure, and execution. It is as much an opportunity to succeed as to fail.

Many organizations are struggling to define a winning strategy for their investment in AI projects, while addressing the risk of costly and complex point AI solutions. The infrastructure needs can vary significantly based on the type and size of the AI model. Cisco® can help you to right-size your investment in AI-related infrastructure while balancing current business and IT needs, with a view for scalability in the future.

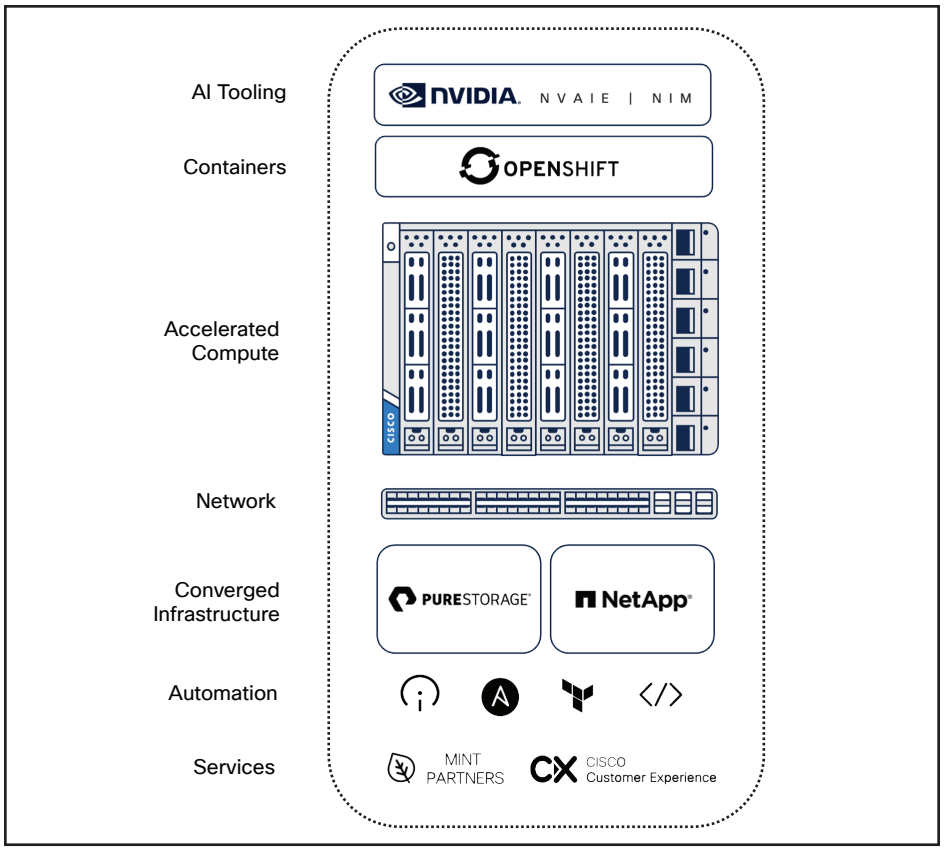

Figure 1. AI Infrastructure POD for Inferencing solution

What is AI inferencing?

AI inferencing involves taking a pre-trained model (e.g., GPT-4, Claude 3, Llama 3) and using it to analyze new data, generating inferences or the most probable outcomes based on that data. This process is widely used in applications like chatbots, coding assistance, and image recognition. While effective for general knowledge questions, traditional AI models can struggle with queries requiring specific data that wasn't part of their training, such as proprietary company data.

This is where Retrieval-Augmented Generation (RAG) comes in. RAG enhances the accuracy and relevance of AI inferencing by incorporating external data sources that the original model wasn't trained on. It connects the model to domain-specific data, enabling it to generate more precise and relevant outputs. For example, consider an insurance model trained on a country's population data—by adding your customer-specific data, the model can provide more accurate and business-relevant insights.

Benefits

- Confidently deploy AI-ready infrastructure with performance assurance and seamless scalability, ensuring your systems are prepared for advanced AI workloads.
- Accelerate AI model deployment and shorten time to production-ready inferencing by leveraging full-stack validation of infrastructure, software, and AI toolsets.
- Operate with best-in-class single-support for your AI deployment architecture, streamlining operations and enhancing reliability across your AI systems.

Learn more

- For more in-depth information on the Cisco AI Infrastructure PODs for Inferencing, refer to the data sheet.
- For Information on all Cisco AI native infrastructure for Data Center, visit **Cisco.com**.
- For more information on the Cisco UCS X-Series Modular System, visit **https://www.cisco.com/go/ucsx**.

What it does

Cisco has been developing and providing Validated Designs for over 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 (TAC). This support and integration provide customers with a reliable and efficient path to achieving their business objectives.

Cisco AI Infrastructure PODs for Inferencing are CVD-based solutions for Edge Inference, RAG, and Large-Scale Inferencing. It provides 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 Infrastructure PODs for Inferencing have independent scalability at each layer of infrastructure and are perfect for DC or Edge AI deployments. There are four configurations that vary the amount of CPU and GPUs in the POD.

Regardless of the configuration, they all contain:

- Cisco UCS X-Series Modular System
  - Cisco UCS X9508 Chassis
  - Cisco UCS-X-Series M7 Compute Nodes
  - Cisco UCS X440p PCIe Node with Nvidia GPUs
  - Cisco UCS 9108 Intelligent Fabric Modules
  - Cisco UCS 6536 Fabric Interconnect or Cisco UCS Fabric Interconnect 9108 100G
  - Cisco UCS X9416 X-Fabric Modules
- Cisco Intersight®
- Cisco Services
- Nvidia NVAIE Subscription
- NVIDIA HPC-X Software Toolkit
- RedHat OpenShift licensing

Optional storage is also available from NetApp (FlexPod) and Pure Storage (FlashStack). Both provide DataOps toolkits to help developers and data scientists perform numerous data management tasks.

C45-4661663-00 09/24