INFERENCE PODS

This document compares different Cisco Inference PODs designed for edge inferencing and AI workloads, highlighting their key features and specifications.

DATA CENTER AND EDGE INFERENCE POD

EDGE INFERENCE FOCUS

Designed for edge applications, processing computations near the user.

SMALLER MODELS

Supports 7B - 13B parameter advanced models like Llama 2-7B and GPT-2B.

SPECIFICATIONS

- 1x UCS X210c M7 compute node
- 2 CPUs
- 512 GB memory
- 1x Nvidia L40s GPU

RAG AUGMENTED INFERENCE POD

DEMANDING AI WORKLOADS

This POD can accommodate **Retrieval-Augmented Generation (RAG)**, which leverages knowledge sources to provide contextual relevance during query service.

LARGER MODELS

Designed for 13B - 40B+ AI models like Llama 2-13B and OPT 13B.

SPECIFICATIONS

- 2x UCS X210c M7 compute nodes
- 4 CPUs
- 1 TB of Memory
- 4x NVIDIA L40s GPUs

ALL CONFIGURATIONS OF CISCO AI INFRASTRUCTURE PODS SHARE COMMON COMPONENTS, INCLUDING:

- Cisco UCS X-Series Modular System
- Cisco Intersight®
- Cisco Services
- Nvidia NVAIE Subscription
- NVIDIA HPC-X Software Toolkit
- RedHat OpenShift licensing

*Optional storage is available from NetApp (FlexPod) and Pure Storage (FlashStack)

SCALE-UP INFERENCE POD

HIGH PERFORMANCE

Optimized to support large-scale models (70B+ Parameters) like Llama 3.3 70B.

ENHANCED RAG

Supports larger vector databases and improves vector search accuracy.

SPECIFICATIONS

- 2x Cisco UCS X210c M7 compute nodes
- 4x CPUs
- 1 TB memory
- 4x Nvidia H100 GPUs

SCALE-OUT INFERENCE POD

FLEXIBILITY AND SCALABILITY

Designed for running multiple models concurrently within a single chassis.

ROBUST

Ideal for organizations that require robust lifecycle management or high availability of models.

SPECIFICATIONS

- 4x Cisco UCS X210c M7 compute nodes
- 8x CPUs
- 4 TB memory
- 8x Nvidia L40s GPUs

CISCO'S INFERENCE PODS SUMMARY

This comparison highlights the different capabilities and specifications of Cisco's Inference PODs for various AI workloads, enabling users to choose the optimal solution.