Cisco AI POD Infrastructure for the NVIDIA 2-8-5-200 Enterprise Reference Architecture

Using Cisco UCS C845A M8 Rack Servers with NVIDIA GPUs and Cisco Nexus 9000 Series Switches.

v1.0, September 2026

Contents

Introduction

Enterprise AI is moving rapidly from experimentation into production. The workloads driving that transition are predominantly inference workloads (agentic, multimodal, generative, retrieval-augmented generation, etc.), served continuously to internal users, external customers, and downstream applications.

Enterprise inferencing requirements vary by workload, but several trends are changing requirements across every layer of the infrastructure. Deployments are seeing longer inputs and outputs per request, which increases KV-cache footprint and GPU memory pressure. Models that exceed single-GPU capacity require multi-GPU and multi-node serving patterns and high-bandwidth GPU-to-GPU communication on the backend fabric. Agentic pipelines add concurrent hosting of multiple models and sustained retrieval, tool, and inter-service traffic on the frontend fabric. As inference scales exponentially, enterprises may need to tier model weights, KV cache, and retrieval data beyond GPU memory across host DRAM, local NVMe, and shared external storage.

These requirements inform the compute, dual-fabric networking, and storage choices in this reference architecture.

Cisco and NVIDIA collaborate across a portfolio of AI infrastructure designs. Cisco AI PODs are the modular infrastructure building blocks for the Cisco Secure AI Factory with NVIDIA. Each is a pre-engineered, AI-ready, full-stack solution that combines Cisco UCS® compute, Cisco Nexus® networking, partner storage, and a curated software stack for enterprise AI inference, training, and fine-tuning. Cisco AI PODs are delivered as Cisco Validated Designs (CVD) or NVIDIA-endorsed Cisco Enterprise Reference Architectures (Enterprise RAs).

This document describes the Cisco Enterprise RA: a Cisco AI POD optimized for enterprise inference (the primary use case) and small-model training and fine-tuning. It delivers the accelerated compute, backend (east/west) and frontend (north/south) networking, and the management and orchestration required to support these workloads. This design is optimal for agentic AI deployments using small to medium-sized models, from reasoning models that perform multi-step inference per query to broader agentic applications. It aligns with the NVIDIA Enterprise Reference Architecture 2-8-5-200 node pattern: 2 CPUs, 8 GPUs, 5 NICs (1 north/south + 4 east/west), and 200 Gbps of east-west fabric bandwidth per GPU.

At its core is the Cisco UCS C845A M8 Rack Server, a 4RU, PCIe-based, NVIDIA-Certified server built on NVIDIA's MGX architecture, running NVIDIA RTX PRO™ 6000 Blackwell Server Edition or NVIDIA H200 NVL GPUs. These servers are interconnected using Cisco Nexus 9000 Series Switches in the backend and frontend fabrics, with Cisco Intersight® managing the compute and Cisco Nexus Dashboard the two fabrics.

NVIDIA endorsement

The design supports the following Cisco UCS server and GPU configurations at design points from 4 to 32 nodes:

The Cisco AI POD Infrastructure for Enterprises design in this document aligns with NVIDIA Enterprise Reference Architecture and is endorsed by NVIDIA for the infrastructure configuration based on NVIDIA's 2-8-5-200 Enterprise RA. The design supports the following UCS servers and GPU configurations at scale points from 4 to 32 nodes:

- Cisco UCS C845A M8 Rack Server with NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs
- Cisco UCS C845A M8 Rack Server with NVIDIA H200 NVL GPUs

Audience

This document is intended for IT architects, data-center engineers, AI/ML infrastructure specialists, and anyone responsible for designing and deploying high-performance enterprise infrastructure for AI workloads.

Scope

The purpose of this reference architecture is to provide a prescriptive Cisco AI POD design based on NVIDIA's PCIe-Optimized 2-8-5-200 Enterprise RA using Cisco UCS C845A M8 Rack Servers with NVIDIA RTX PRO 6000 Blackwell Server Edition or NVIDIA H200 NVL GPUs and Cisco Nexus 9000 Series Switches. The solution is designed to deliver predictable performance, reliability, and scalability for enterprise AI inferencing workloads. While the NVIDIA Enterprise Reference Architecture for 2-8-5-200 includes a range of modular designs, this document details one specific architecture endorsed by NVIDIA.

Cisco supports additional NVIDIA Enterprise RA-endorsed designs for different workload profiles, including an HGX-based configurations for large training and inferencing workloads. See complete list at: https://docs.nvidia.com/enterprise-reference-architectures/index.html#.

Solution overview

The Cisco AI POD design in this document is an NVIDIA-endorsed Cisco Enterprise RA, primarily for enterprise inferencing. The major elements of this solution are:

- **Compute:** Cisco UCS C845A M8 Rack Servers managed using Cisco Intersight. Each node is populated with up to a maximum of:
  - 8x NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs (or 8x NVIDIA H200 NVL GPUs)
  - 4x BlueField-3 B3140H SuperNICs (one per pair of GPUs, east/west)
  - 1x BlueField-3 B3220 DPU (data processing unit; one per chassis, north/south)
- **Network fabrics:** two non-blocking network fabrics deployed and managed using Cisco Nexus Dashboard. Both fabrics are BGP EVPN VXLAN fabrics, built on Cisco Nexus 9000 Series Switches in a spine-and-leaf topology:
  - 400GbE backend for GPU-to-GPU traffic, delivering 200 Gbps per GPU
  - 200GbE frontend for management, orchestration, storage, and customer/inference traffic
- **Storage:** connected to the frontend fabric and sized to meet customer requirements. External storage options include NVIDIA-certified solutions from NetApp, Everpure, and VAST Data.

The solution can scale from 4 nodes / 32 GPUs up to a maximum of 32 nodes / 256 GPUs. The topology, the operational model, and the management and orchestration stack stay the same at every scale point.

Target workloads and use cases

The NVIDIA RTX PRO 6000 Blackwell Server Edition GPU is purpose-built for enterprise AI inference. The Blackwell architecture's fifth-generation NVIDIA Tensor Cores natively support FP4 for high-throughput, low-precision inference, along with FP8 / FP16 / BF16 for higher-precision workloads. Each GPU carries 96 GB of GDDR7 memory (roughly double that of the PCIe-based L40S) and PCIe Gen5 connectivity, in an air-cooled, rack-density-friendly footprint. The NVIDIA H200 NVL GPU is an endorsed alternative within the same chassis for workloads that benefit from larger per-GPU memory (141 GB HBM3e), the higher memory bandwidth of HBM3e, and direct GPU-to-GPU NVLink connectivity through optional NVL2 / NVL4 bridges. This solution targets the following use cases:

- **Production enterprise inference:** Large Language Models (LLMs), multimodal models, vision models, speech models, and Small Language Models (SLMs) served behind an enterprise application or agent
- **Retrieval-Augmented Generation (RAG) and agentic AI** pipelines, including embedding generation, vector search, and re-ranking, along with the served model
- **Small to medium-sized model fine-tuning** using parameter-efficient methods (LoRA / QLoRA) and full fine-tuning
- **Multitenant AI platforms** that use GPU partitioning to run multiple isolated workloads per GPU, improving overall cluster utilization and consolidation
- **Hybrid inference and small-scale training** where the same infrastructure serves both workload types without standing up two separate clusters

Solution components

The key hardware and software components in this Cisco AI POD Enterprise RA are described below.

Cisco UCS compute

The Cisco UCS C845A M8 is a 4RU rack server designed for enterprise AI inference and small-scale training and fine-tuning. It supports up to 8 GPUs on a PCIe Gen5 fabric, with the GPU memory, mixed-precision performance, and bandwidth required for production AI workloads.

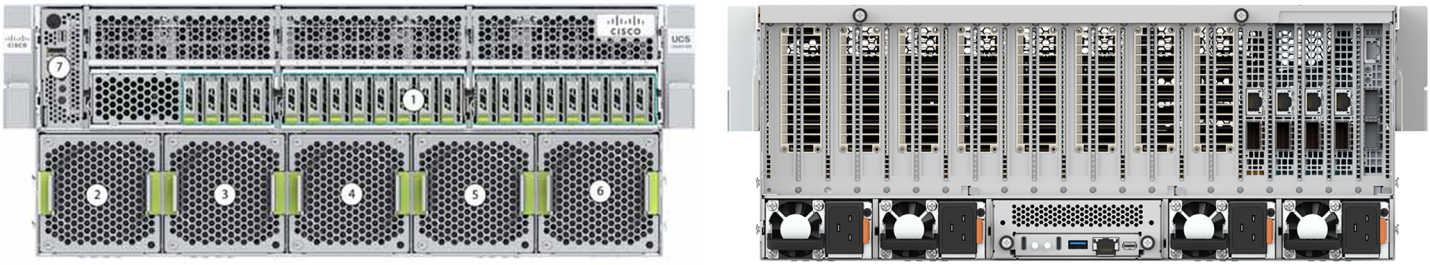

Figure 1. Cisco UCS C845A M8 Rack Server (front and rear views)

Each Cisco UCS C845A M8 Rack Server in this reference architecture is configured with:

- **CPUs:** 2x AMD EPYC 9555 (3.2 GHz, 64-core) processors
- **Memory:** 32x 64 GB Cisco DDR5-6400 RDIMMs (2 TB total), populating all 32 DIMM slots (2 DIMMs per channel). With 5th Gen AMD EPYC processors and a fully populated 2-DPC configuration, the memory operates at up to 4400 MT/s. (1-DPC configurations operate at up to 5200 MT/s; see the “Cisco UCS C845A M8 AI Memory Guide” in the References section for the full DIMM / speed matrix.)
- **GPUs:** 8x NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs (96 GB GDDR7 each, dual-slot full-height, full-length [FHFL], 600 W TDP). The Cisco UCS C845A M8 server also supports 8x NVIDIA H200 NVL GPUs (141 GB HBM3e each, dual-slot FHFL, 600 W TDP), with optional NVIDIA NVL2 / NVL4 bridges for multi-GPU NVLink pairing. See UCS C845A M8 data sheet in References section for a full list of currently supported GPUs.
- **Backend (east/west) NICs:** 4x NVIDIA BlueField-3 B3140H SuperNICs (1x 400GbE QSFP112 each, PCIe Gen5), one backend SuperNIC per pair of GPUs
- **Frontend (north/south) NIC:** 1x NVIDIA BlueField-3 B3220 DPU (2x 200GbE QSFP112, PCIe Gen5)
- **Local storage:** E1.S 15 mm NVMe PCIe Gen4 SSDs, distributed across the four E1.S cages and balanced across CPU sockets for NUMA affinity, with drives placed under the same PCIe root ports as the GPUs and NICs they serve. The NVIDIA Enterprise Reference Architecture recommends a minimum of 1 TB NVMe per CPU socket for inference and 2 TB NVMe per CPU socket for training and fine-tuning workloads. The recommended starting configuration for this design is 8x 3.8 TB E1.S NVMe SSDs per server (4 per CPU socket); see the BOM for the platform maximum and drive capacity options. Boot is through 2x 960 GB M.2 SATA SSDs configured as RAID-1.
- **Out-of-Band (OOB) management:** 1x Intel® X710-T2L 2x 10GbE RJ45 OCP 3.0 NIC for x86 host management
- **Power:** 4x 3.2 kW AC Titanium PSUs in N+1 redundancy (3 active + 1 redundant), supporting AC 220 V input

See the References section for up-to-date specifications for the product. The Cisco UCS C845A M8 server is managed through Cisco Intersight.

Cisco Nexus fabrics

This design uses two fabrics: a backend (east/west) for GPU-to-GPU traffic and a frontend (north/south) for management, orchestration, storage, and application/inference traffic. Both fabrics are built on Cisco Nexus 9000 Series Switches and managed through Cisco Nexus Dashboard, which provides best-practice templates for deploying and operating the fabric. The Cisco Nexus 9000 Series Switches deliver the AI fabric features that enterprise AI workloads require: RoCEv2, load-balancing mechanisms, Priority Flow Control (PFC), Explicit Congestion Notification (ECN), and telemetry.

Cisco Nexus 9364E-SG2 switch

The Cisco Nexus 9364E-SG2 is a 2RU 800GbE switch built on the Cisco® Silicon One® ASIC. It provides 64x 800GbE ports (QSFP-DD or OSFP) with support for 400 / 200 / 100GbE speeds, 51.2 Tbps of switching capacity, and a 256 MB on-die buffer for microburst absorption.

Figure 2. Cisco Nexus 9364E-SG2 switch

In this design, the Cisco Nexus 9364E-SG2 switch serves as both spine and leaf in the backend fabric. Each Cisco UCS C845A M8 Rack Server attaches at 400GbE per BlueField-3 SuperNIC, or 200 Gbps per GPU; the 800GbE switch ports are broken out to 400GbE toward the servers, leaving headroom and breakout flexibility as GPU and NIC link speeds increase.

Cisco Nexus 9364D-GX2A switch

The Cisco Nexus 9364D-GX2A is a 2RU fixed switch with 64x 400GbE QSFP-DD ports (with support for 200/100/50/25/10GbE speeds), 51.2 Tbps of switching capacity, 8.35 Bpps of forwarding, and a 120 MB shared buffer.

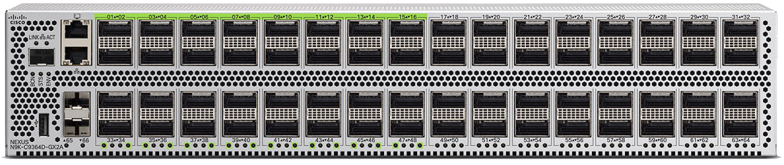

Figure 3. Cisco Nexus 9364D-GX2A switch

In this design, the Cisco Nexus 9364D-GX2A switch serves as both spine and leaf in the frontend fabric, carrying management, orchestration, storage, and customer/inference traffic. Each Cisco UCS C845A M8 Rack Server attaches at 2×200GbE per BlueField-3 DPU; the 400GbE switch ports are broken out to 200GbE toward the servers. The frontend in this solution uses Cisco Cloud Scale (the 9364D-GX2A); customers who prefer a single Cisco Silicon One fabric across both networks can use Silicon One switches on the frontend as well.

Unified Operations and Infrastructure Management

This solution uses Cisco Cloud Control (CCC), Cisco Intersight, and Cisco Nexus Dashboard for unified operations and infrastructure management. Cisco Nexus 93108TC-FX3 switches provide dedicated OOB management connectivity in the solution.

Cisco Cloud Control

Cisco Cloud Control provides unified, AI-enabled operations for the solution, integrating with both Cisco Intersight and Cisco Nexus Dashboard to provide a single sign-on experience, consolidated inventory and topology, and operational context across compute, network fabric, security, and observability. Cisco Intersight and Cisco Nexus Dashboard remain the domain managers for their respective domains, while Cisco Cloud Control enable a common over-arching operational experience across domains. Cisco Intersight manages Cisco UCS infrastructure, while Cisco Nexus Dashboard manages the frontend and backend network fabrics.

Cisco Intersight

Cisco Intersight is a cloud-based software-as-a-service (SaaS) platform that provides centralized lifecycle management for the Cisco UCS platforms in this solution, including Cisco UCS C845A M8 GPU compute nodes and the Cisco UCS X-Series management cluster. Its capabilities include inventory, hardware-health monitoring, policy-based configuration, firmware lifecycle management, and operational visibility. Available capabilities vary by platform.

Because this solution runs Red Hat OpenShift on bare-metal Cisco UCS infrastructure, two integrations between Cisco and Red Hat are particularly relevant:

- **OpenShift Assisted Installer integration:** In the OpenShift installation workflow, the Assisted Installer from Red Hat Hybrid Cloud Console cross-launches to Cisco Intersight. The administrator selects the Intersight-managed UCS nodes from a list of available servers. Intersight then mounts and boots the Red Hat discovery ISO on the selected servers. This removes the need for manual virtual-media mounting, a PXE boot network, or a separate provisioning host. The administrator then returns to the Assisted Installer to complete the OpenShift deployment. The workflow is used both to stand up a new cluster and to add bare-metal nodes to an existing one.
- **Cisco Intersight OpenShift Operator:** a Red Hat–certified operator deployed on the OpenShift cluster integrates Intersight to provide UCS server inventory, hardware health status, and lifecycle context from within OpenShift, giving administrators visibility into the physical compute layer from within the OpenShift console.

Cisco Nexus Dashboard

Cisco Nexus Dashboard provides a centralized automation and operations platform for the frontend and backend Nexus fabrics in the design. A single Nexus Dashboard cluster provisions and deploys both fabrics using Cisco best-practice templates and provides full fabric lifecycle management. Capabilities relevant to this design include:

- **AI best-practice fabric templates:** Built-in AI fabric templates apply Cisco's recommended settings for supported AI/ML fabric types, including lossless RoCE transport (PFC and ECN), load-balancing mechanisms, and QoS. Settings are applied fabric-wide, ensuring a consistent deployment based on a single template.
- **Multi-fabric management and federation:** A single Nexus Dashboard cluster can manage multiple fabrics. Multiple Nexus Dashboard clusters can also be interconnected to provide a single-pane-of-glass view as the deployment scales.

Cisco Nexus 93108TC-FX3 switch

The 1RU Cisco Nexus 93108TC-FX3 switch provides 48x 100M / 1 / 10GBASE-T ports and 6x 40 / 100GbE QSFP28 uplinks. The switches provide dedicated OOB management connectivity for Cisco UCS systems, Nexus switches, and Nexus Dashboard appliances in the solution. The 1GbE OOB management port on the NVIDIA BlueField-3 B3220 DPU in each Cisco UCS GPU compute node also connects to these switches for OOB management, providing access to the BlueField management plane for firmware and recovery operations. This connection is required for DPU-mode operation. OOB connections for the four BlueField-3 B3140H SuperNICs in each node can also be added for direct adapter management but are not required.

Figure 4. Cisco Nexus 93108TC-FX3 switch

Solution architecture

This section describes the solution architecture and design of its subsystems. The high-level solution topology for a 32-node / 256-GPU cluster is shown in Figure 5.

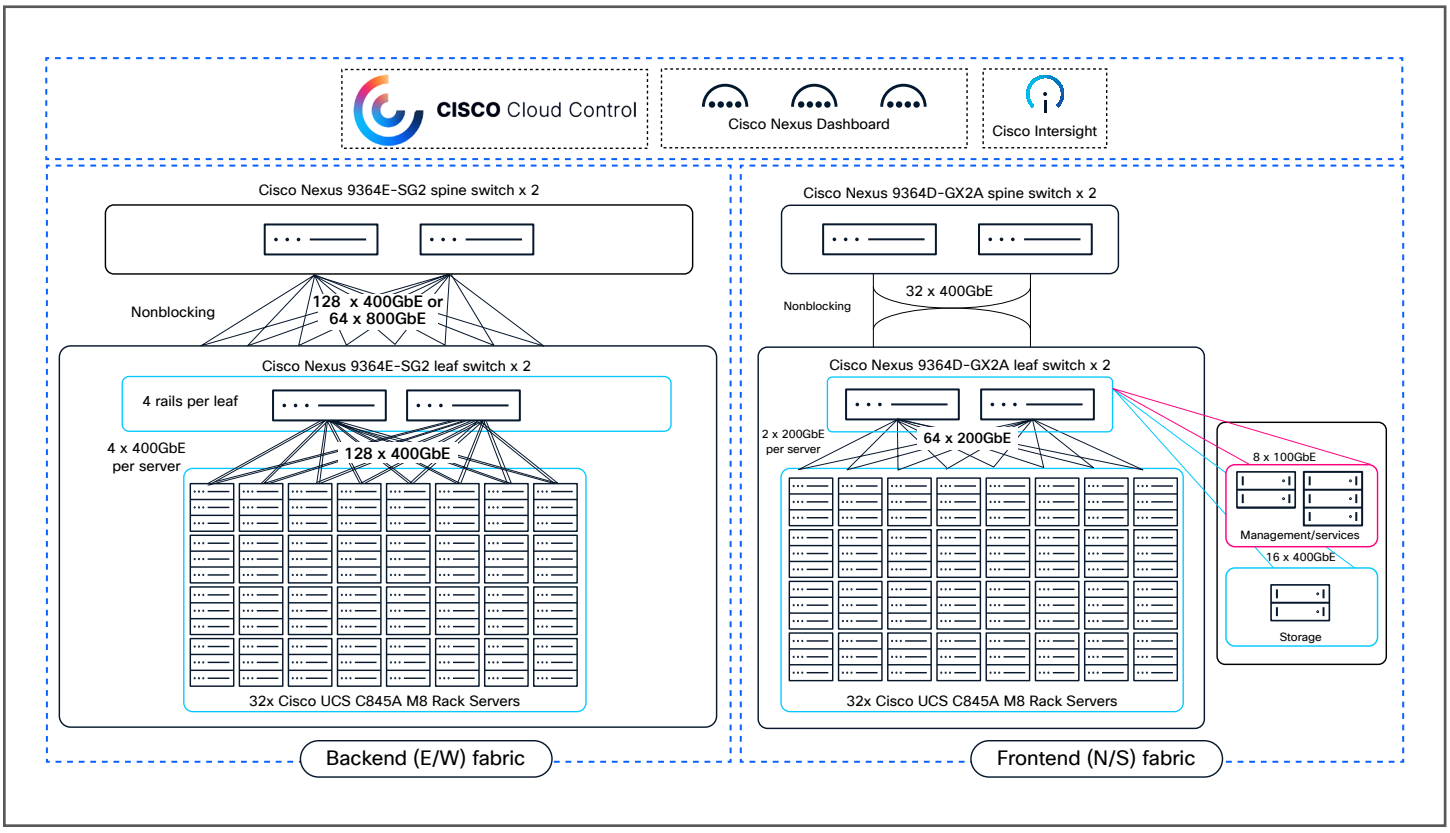

Figure 5. Cisco AI POD with Cisco UCS C845A M8 Rack Server solution topology

Key elements:

- **Backend (E/W) fabric:** Cisco Nexus 9364E-SG2 spine pair, with a single Cisco Nexus 9364E-SG2 leaf pair serving up to 32 GPU nodes (256 GPUs / 8 scalable units). Each Cisco UCS C845A M8 Rack Server attaches with 4x 400GbE (one BlueField-3 B3140H SuperNIC per GPU pair).
- **Frontend (N/S) fabric:** Cisco Nexus 9364D-GX2A spine pair with a single Cisco Nexus 9364D-GX2A leaf pair. Each Cisco UCS C845A M8 Rack Server attaches with 2x 200GbE (one BlueField-3 B3220 DPU). The Cisco UCS X-Series Direct management cluster attaches with 8x 100GbE.
- **Management:** A 3-node Cisco Nexus Dashboard cluster manages both fabrics, and Cisco Intersight manages the UCS compute nodes. For out-of-band management, Cisco Nexus 93108TC-FX3 1RU switches aggregate the management interfaces of every UCS, Nexus, and Nexus Dashboard appliance, along with the NVIDIA BlueField-3 B3220 DPU on each node; the 4x BlueField-3 B3140H SuperNICs per node can also be optionally connected to this OOB network.
- **Orchestration:** A Red Hat OpenShift (Kubernetes) control plane, hosted on a Cisco UCS X-Series Direct management cluster, schedules and manages containerized AI inference and training workloads across the GPU nodes.

The same physical topology, the same fabric templates, and the same management plane are used at every scale point, from a single Scalable Unit (SU) consisting of 4 nodes up through 8 SUs with 32 nodes and a 256-GPU maximum.

Dual-fabric design

Enterprise AI clusters running a mix of inference, fine-tuning, and small-model training have two distinct traffic classes and therefore use two distinct network fabrics:

- **Backend (E/W) fabric:** An isolated fabric dedicated to low-latency, lossless inter-node GPU-to-GPU communication.
- **Frontend (N/S) fabric:** Carries management, control-plane, storage access, and customer/inference traffic to and from applications, users, and agents. The frontend fabric can be a dedicated fabric or an existing enterprise data-center network that meets the bandwidth and performance targets of inference workloads.

A backend fabric is not required for every inference deployment. Model size, the selected GPU, and performance requirements determine whether model serving can remain within one node. A backend fabric is required when model execution is across nodes. Single-node model-serving instances, including multiple independent replicas, do not use the backend fabric for model execution. Multi-node model serving in which model parallelism spans nodes and disaggregated serving in which prefill and decode worker pools run on different nodes require the backend fabric for inter-node model-serving traffic. This design includes the backend fabric to support those patterns.

Backend (E/W) network fabric

The backend fabric carries inter-node GPU traffic from distributed inference, including collective operations for tensor parallelism, activation transfers for pipeline parallelism, and KV cache transfers for disaggregated serving, as well as traffic from multi-node training and fine-tuning.

This backend fabric is a two-tier, non-blocking spine-and-leaf (Clos) fabric using Cisco Nexus 9364E-SG2 switches and managed by Cisco Nexus Dashboard. The fabric can be scaled by adding leaf pairs to the existing spine tier and adding spine capacity as needed. It uses an MP-BGP EVPN control plane and a VXLAN data plane:

- **Control plane:** MP-BGP advertises Layer 2 (MAC) and Layer 3 (IP) reachability across the fabric.
- **Data plane:** VXLAN encapsulation in IP/UDP carries the overlay networks.

This architecture provides native multitenancy, supports Layer 2 and Layer 3 overlays, and enables per-tenant segmentation. RoCEv2 is enabled fabric-wide to support GPUDirect RDMA between GPU nodes, with load balancing, Priority Flow Control (PFC) for lossless transport, and Explicit Congestion Notification (ECN) for congestion management across the backend fabric.

Frontend (N/S) network fabric

The frontend (north/south) fabric carries the cluster management and control plane traffic, GPU-server-to-storage traffic, and inference traffic between applications, users, agents, and the served models. Because the backend fabric is isolated, the frontend fabric is also the entry and exit point for all traffic into the GPU cluster where the models run.

The frontend fabric is also a two-tier, non-blocking spine-and-leaf (Clos) fabric using Cisco Nexus 9364D-GX2A or Cisco Nexus 9364E-SG2 switches and managed by the same Cisco Nexus Dashboard as the backend. Alternatively, an existing enterprise data center network that meets the bandwidth and performance targets of this design can serve as the frontend fabric.

The frontend fabric meets the following bandwidth targets outlined in the NVIDIA Enterprise RA:

- At least 12.5 Gbps per GPU for storage traffic
- At least 25 Gbps per GPU for user and inference traffic

Because the frontend fabric is shared for a mix of traffic types, QoS should prioritize latency-sensitive inference and storage traffic over best-effort traffic.

Node connectivity

This section describes how each Cisco UCS C845A M8 node connects internally across its GPUs, and externally to the backend and frontend fabrics.

Intra-node connectivity

Within a single Cisco UCS C845A M8 Rack Server, the 8 NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs are connected over a PCIe Gen5 fabric, with GPUs internally paired behind PCIe switches and each pair sharing an NVIDIA BlueField-3 B3140H SuperNIC for backend connectivity. When the server is configured with NVIDIA H200 NVL GPUs, the same PCIe topology applies, with optional NVIDIA NVL2 or NVL4 bridges providing direct GPU-to-GPU NVLink connectivity within the server for workloads that benefit from higher inter-GPU bandwidth than what PCIe provides. The intra-node PCIe, GPU, and NIC topology is shown in Figure 6. See the References section for a more detailed topology of the PCIe connectivity within the Cisco UCS C845A M8 Rack Server.

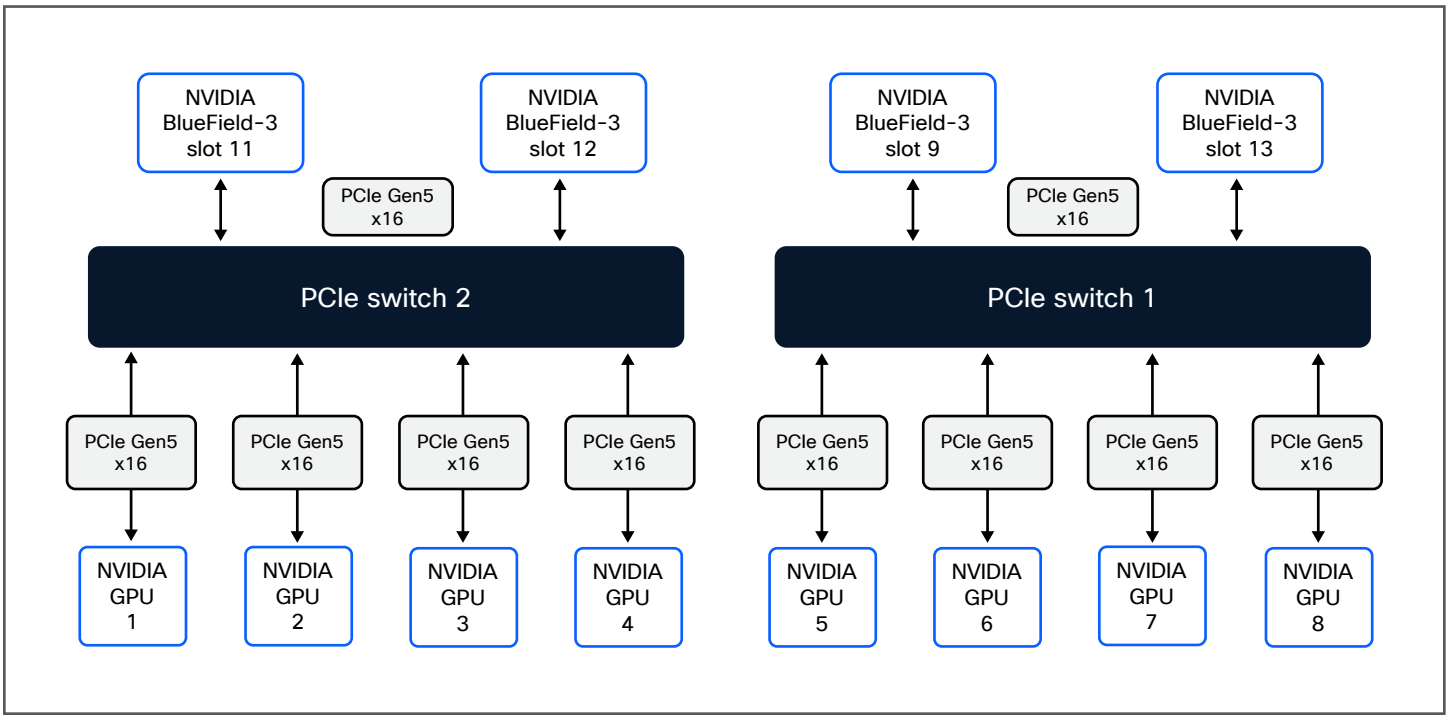

Figure 6. Cisco UCS C845A M8 Rack Server intra-node topology

Connectivity to the Backend Fabric

The Cisco UCS C845A M8 Rack Servers connect to the backend (east/west) fabric for GPU-to-GPU communication using a rail-optimized topology consistent with the NVIDIA's 2-8-5-200 Enterprise Reference Architecture. Each server's four NVIDIA BlueField-3 B3140H SuperNICs, one per GPU pair, are distributed across the leaf switches. Each GPU rank across nodes forms a rail, and multiple rails connected to the same leaf switch form a rail group. With the single backend leaf pair in this design, two SuperNICs from each node connect to each leaf switch, so each leaf switch carries a rail group comprising four rails. This mapping allows communication between the same GPU ranks across nodes to remain within one leaf switch, minimizing spine traversal.

A single pair of Cisco Nexus 9364E-SG2 leaf switches supports up to 32 Cisco UCS C845A M8 Rack Servers or 256 GPUs (8 scalable units) in a rail-optimized topology, when configured with sufficient leaf-to-spine bandwidth to preserve NIC-to-NIC non-blocking connectivity across the fabric. See the BOM notes for guidance on scaling beyond 32 nodes. Figure 7 shows the rail-optimized backend topology used in this design.

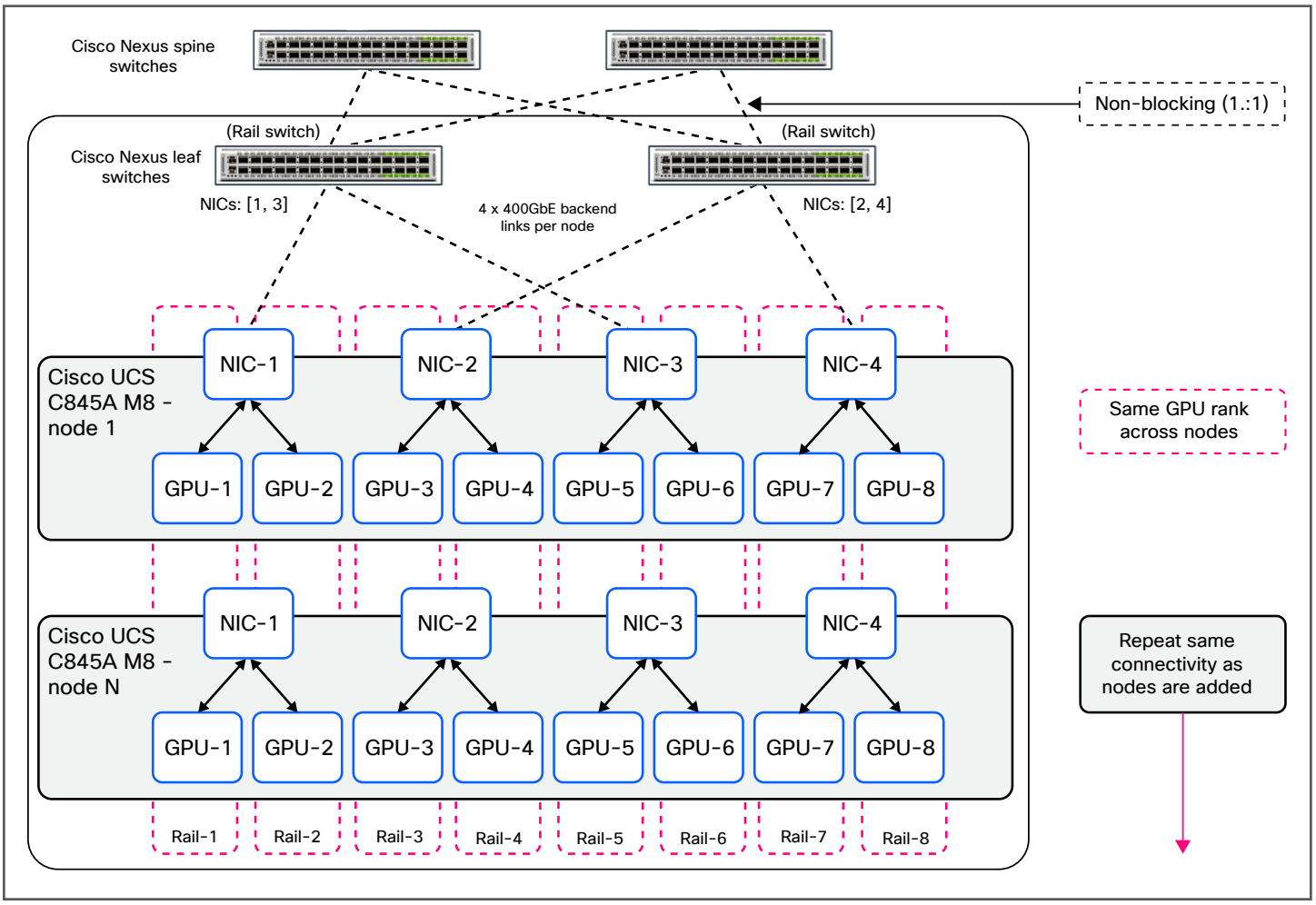

Figure 7. Rail-optimized backend topology

Connectivity to the frontend fabric

Each Cisco UCS C845A M8 Rack Server attaches to the frontend fabric through a single NVIDIA BlueField-3 B3220 DPU. The 2x 200GbE ports on the NIC are configured as an active/active port-channel toward the frontend leaf pair, providing up to 400 Gbps of frontend bandwidth per server. For an eight-GPU node, this provides 50Gbps of aggregate bandwidth per GPU, sufficient to meet the NVIDIA Enterprise RA targets of 25Gbps per GPU for customer and inference traffic and 12.5Gbps per GPU for storage traffic. Frontend traffic classes are separated into VLANs and carried across the bonded interfaces as tagged VLANs.

Orchestration and management

The Kubernetes orchestration and control-plane functions in this solution run on a separate management and services cluster. In this reference architecture, three Cisco UCS X-Series compute nodes installed in a Cisco UCS X9508 Chassis provide a highly available OpenShift control plane. The chassis is equipped with two Cisco UCS Fabric Interconnect 9108 100G modules to form a Cisco UCS X-Series Direct system. Together, the modules provide up to 16 × 100GbE uplinks, eight per Fabric Interconnect, to the frontend fabric. The Cisco UCS X-Series Direct is managed through Cisco Intersight and can be expanded by adding a second X9508 chassis with up to eight additional nodes and up to four Cisco UCS C-Series rack servers, for a total of up to 20 compute nodes per domain.

Additionally, a Nexus Dashboard cluster is deployed on three separate physical appliances, and a standalone management node provides OpenShift CLI access to the cluster as well as administrative access to devices through the OOB management network. OOB management connectivity is provided by Cisco Nexus 93108TC-FX3 switches in this design.

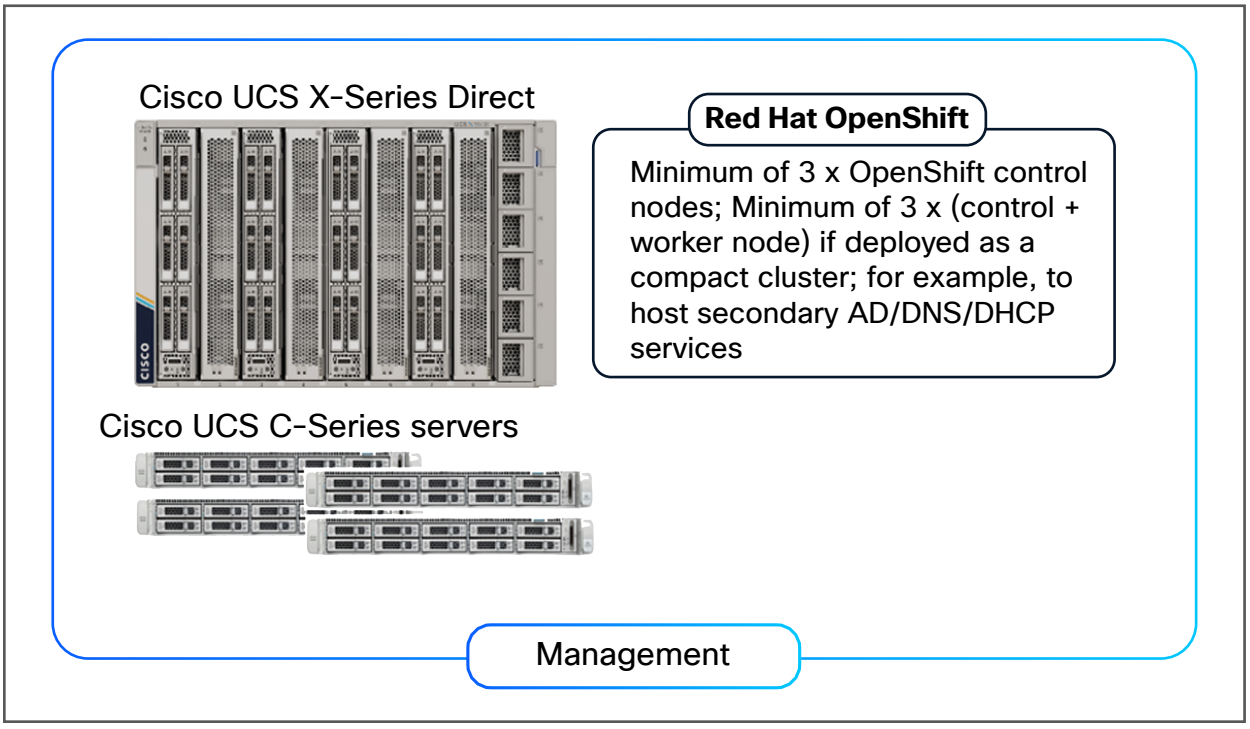

Figure 8. Cisco management cluster

Software stack

This section lists the software and firmware versions recommended for this reference design. For current supported versions and component combinations, use the compatibility resources listed in the Interoperability section.

Table 1. Hardware and software matrix

Component	Version	Notes
Backend fabric		
Cisco Nexus 9364E-SG2	10.6(3)	Spine and leaf switches
Frontend fabric		
Cisco Nexus 9364D-GX2A	10.6(3)	Spine and leaf switches
Out-of-band management		
Cisco Nexus 93108TC-FX3	10.6(3)	OOB management switch
Cisco UCS GPU compute		
Cisco UCS C845A M8 Rack Server	Latest firmware per UCS HCL	Managed by Cisco Intersight
NVIDIA RTX PRO 6000 Blackwell Server Edition	Same as above	Requires add-on NVAIE license – see SKU below.
NVIDIA H200 NVL	Same as above	5-year NVAIE subscription included – no add-on license required.
Management/services		
Cisco Nexus Dashboard	4.2.1	3-node physical cluster
Cisco Intersight	N/A	SaaS-based
Cisco UCS X9508 Chassis (UCSX-9508-D-U)	N/A	Server chassis for management
Cisco UCS X-Series Direct 100G (UCSX-S9108-100G)	N/A	8x100GbE uplinks, 2x per chassis for HA

Component	Version	Notes
Cisco UCS X-Series M8 compute nodes (UCSX-215C-M8)	Latest firmware per UCS HCL	Minimum 3 nodes as OpenShift control plane nodes
Cisco VIC 15230 mLOM (UCSX-MLV5D200GV2D)	Latest firmware per UCS HCL	2x 100GbE mLOM per compute node
Orchestration/software		
Red Hat AI Enterprise (RHAIE)	N/A	Umbrella licensing bundle for bare-metal OpenShift; per-GPU AI Accelerator entitlement
Red Hat OpenShift Container Platform (OCP)	4.21 or later	Verify support in UCS HCL
Red Hat OpenShift AI (RHOAI)	3.4 or later	MLOps platform
Cisco Intersight Operator	Latest available in OperatorHub	Lifecycle context for Intersight-managed Cisco UCS servers
NVIDIA AI Enterprise	8.1	Umbrella software and licensing suite for GPUs, AI software and tools
NVIDIA GPU Operator	26.3.1 or later	GPU Driver, GPUDirect RDMA
NVIDIA Network Operator	26.1.1 or later	DOCA-OFED Driver, RDMA/SR-IOV, GPUDirect RDMA

NVIDIA Spectrum-X: The NVIDIA Spectrum-X Networking technology significantly improves the performance and efficiency of Ethernet-based GPU and storage networks. Its benefits are available with Cisco Silicon One switches when connected to NVIDIA ConnectX-8 and BlueField-3 SuperNICs and Fine Grain Load Balancing (FGLB) license enabled.

Interoperability

The interoperability information for the components in this Reference Architecture is summarized in Table 2.

Table 2. Interoperability

Component	Interoperability matrix and other relevant links
Cisco UCS Hardware Compatibility List (HCL)	https://ucshcltool.cloudapps.cisco.com/public/
NVIDIA AI Enterprise licensing	https://docs.nvidia.com/ai-enterprise/planning-resource/licensing-guide/latest/licensing.html
NVIDIA-Certification	https://www.nvidia.com/en-us/data-center/products/certified-systems/
NVIDIA-Certified Systems and NVIDIA AI Enterprise compatibility	https://docs.nvidia.com/certification-programs/latest/nvidia-certified-systems.html
NVIDIA driver lifecycle, release, and CUDA support	https://docs.nvidia.com/datacenter/tesla/drivers/index.html#lifecycle
NVIDIA AI Enterprise Infrastructure Support Matrix	https://docs.nvidia.com/ai-enterprise/support-matrix/latest/index.html

Ordering

Bill of Materials (BOM)

The BOM in Table 3 describes a single Scalable Unit (SU; 4 nodes / 32 GPUs), which is the granular building block of this design. Larger deployments scale by adding additional scalable units onto the same backend leaf pair, up to the maximum supported cluster size of 32 nodes / 256 GPUs. See the BOM Notes section following Table 3 for additional context on scaling beyond the 32-node design point.

The transceivers and cables shown are representative; alternates are available from the Cisco Transceiver Module Group portfolio referenced below.

Table 3. Bill of materials (four Cisco UCS C845A M8 Rack Servers with 32 NVIDIA GPUs)

#	Type	PID	Description	Qty
Cisco UCS GPU server				
A1	Cisco UCS server bundle	UCS-MGPUM8-MLB	Cisco UCS C845A M8 AI Server Major Line Bundle (MLB)	1
A2	Cisco UCS AI server	CAI-845A-M8	Cisco UCS C845A M8 base server (no CPU, memory, drives); 4RU rack server	4
A3	CPU	CAI-CPU-A9555	AMD EPYC 9555 3.2 GHz 360 W 64-core / 256 MB cache	8
A4	Memory	CAI-MRX64G2RE5	64 GB DDR5-6400 RDIMM 2Rx4 (16 Gb); 2 TB per node; operates at 4400 MT/s in 2-DPC config	128
A5	Boot drive	CAI-M2-960G	960 GB M.2 SATA SSD	8
A6	Boot RAID	CAI-M2-HWRAID	Cisco boot-optimized M.2 RAID controller	4
A7	Power supply	CAI-845A-PSU	Cisco UCS C845A M8 3.2 kW AC Titanium PSU	16
A8	Power cable (chassis)	CAB-C19-CBN	Cabinet jumper power cord, 250 VAC 16 A, C20–C19	16

#	Type	PID	Description	Qty
A9	Power cable (GPU)	CAI-CBL-GPU-N	Cisco UCS C845A NVIDIA GPU power cable (Assumes 8 GPUs per node)	32
A10	Power cable (N-S NIC)	CAI-CBL-BF3-N-S	Cisco UCS C845A BlueField-3 N-S NIC power cable	4
A11	Heat sink	CAI-HS-C845A	Cisco UCS C845A heat sink	8
A12	Local storage	CAI-NVES3T8K1V	3.8 TB E1.S Gen4 NVMe SSD (KIOXIA XD7P). Recommended configuration: 8 per server (4 per CPU socket for NUMA affinity) = ~30 TB. Platform max: 20x E1.S NVMe SSDs per server (1.9 / 3.8 / 7.6 TB options).	32
A13	GPU bracket	CAI-BRK-GPU	Cisco UCS C845A GPU accessory bracket	32
A14	Rail kit	CAI-845A-RAIL	Cisco UCS C845A ball-bearing rail kit	4
A15	GPU - option 1	CAI-GPU-RTXP6000	NVIDIA RTX PRO 6000 Blackwell Server Edition; 600 W, 96 GB GDDR7, 2-Slot FHFL	32
A16	GPU - option 2	CAI-GPU-H200-NVL	NVIDIA H200 NVL; 600 W, 141 GB HBM3e, 2-Slot FHFL. Includes 5-year NVIDIA AI Enterprise software license.	32
A17	GPU - option 2 (NVL2 bridge)	CAI-NVL2-H200	NVIDIA NVL-2-way Bridge for H200 NVL GPU. Optional and varies based on the workload.	16
A18	GPU - option 2 (NVL4 bridge)	CAI-NVL4-H200	NVIDIA NVL-4-way Bridge for H200 NVL GPU. Optional and varies based on the workload.	8
A19	OOB NIC	CAI-O-ID10GC	Intel X710-T2L 2x 10GbE RJ45 OCP 3.0 NIC	4

#	Type	PID	Description	Qty
A20	E-W NIC (SuperNIC)	CAI-P-N3140H	NVIDIA OEM BlueField-3 B3140H SuperNIC, 1x 400GbE QSFP112, PCIe Gen5 (E-W slots 9, 11, 12, 13)	16
A21	N-S NIC	CAI-P-N3220	NVIDIA BlueField-3 B3220 DPU, 2x 200GbE	4
A22	E-W transceivers	QSFP-400G-DR4	400G QSFP112, 400GBASE-DR4, SMF MPO-12 APC, 500 m	16
A23	N-S transceivers	QSFP-200G-SR4-S	QSFP56, 200GbE, MMF, MPO-12 UPC, 100 m OM4	8
Compute/GPU licenses				
A24	Cisco Intersight licenses	DC-MGT-SAAS	Cisco Intersight Advantage SaaS – per Cisco UCS C845A M8 GPU server	4
A25	License bundles	RHNV-AI-FAC-SW-LIC	Red Hat AI Factory with NVIDIA Software. (It combines Red Hat AI Enterprise and NVIDIA AI Enterprise into a single offering with simplified node-based pricing on NVIDIA hardware.)	-
A25.1	Red Hat AI	RH-RHAIE-P3S	Red Hat AI Enterprise, Premium 3-yr SnS[1] (1 physical or virtual node)	4
A25.2	NVIDIA AI Enterprise	NV-NVAIE-8GPUC-3Y	NVAIE subscription per node (8 GPU) business critical support; 3-year	4
Backend (E/W) fabric switches				
B1	Spine-leaf switches	N9K-C9364E-SG2-Q	Cisco Nexus 9364E-SG2, 64x 800G QSFP-DD switch (2x spine + 2x leaf)	4
B2	Transceivers (spine→leaf)	QDD-8X100G-FR	Dual-port QSFP-DD module, SMF, dual MPO-12 (APC), 2 km, parallel (2-port module; half the port count)	8

#	Type	PID	Description	Qty
B3	Transceivers (leaf→spine)	QDD-8X100G-FR	Dual-port QSFP-DD module, SMF, dual MPO-12 (APC), 2 km, parallel (2-port module; half the port count)	8
B4	Cables (leaf↔spine)	CB-M12-M12-SMF5M=	MPO-12 to MPO-12 SMF cables	16
B5	Transceivers (leaf→server)	QDD-8X100G-FR	Dual-port QSFP-DD module, SMF, dual MPO-12 (APC), 2 km, parallel (2-port module; half the port count)	8
B6	Cables (leaf↔server)	CB-M12-M12-SMF5M=	MPO-12 to MPO-12 SMF cables	16
B7	Licenses	C1A1TN9300XF3-3Y	Cisco Data Center Networking (DCN) Advantage subscription for 800GbE Fixed Platforms (N9300 XF3); 3-year	4
B8	Licenses	SVS-L2N9KA-XF3-3Y	Mandatory Cisco Support Signature (SVS) license for base DCN Advantage license above; 3-year	4
B9	Licenses	C1N9K-FGLB-XF3-3Y	Adaptive routing (fine-grain load balancing) subscription for Cisco Nexus 9000 switches; requires Cisco DCN Advantage base subscription and NVIDIA NICs; 3-year	4
Frontend (N/S) fabric switches				
C1	Spine-leaf switches	N9K-C9364D-GX2A	Cisco Nexus 9300 Series, 64p 400G switch	4
C2	Transceivers (spine→leaf)	QDD-400G-FR4-S=	400G QSFP-DD, 400G-FR4, Duplex LC, 2 km SMF	4
C3	Transceivers (leaf→spine)	QDD-400G-FR4-S=	400G QSFP-DD, 400G-FR4, Duplex LC, 2 km SMF	4

#	Type	PID	Description	Qty
C4	Cables (spine↔leaf)	CB-LC-LC-SMF5M=	LC SMF (UPC) cables	4
C5	Transceivers (leaf→server)	QDD-400G-SR8-S=	400G QSFP-DD, MMF, MPO-16 APC, 100 m OM4; breaks out to 2× 200G-SR4 toward the node	4
C6	Cables (leaf↔server)	M16-2xM12-OM4-7M	MPO-16 APC to 2× MPO-12 UPC OM4 MMF breakout trunk cable, 7 m	4
C7	Transceivers (leaf↔mgmt)	QDD-4X100G-LR-S=	QSFP-DD, 4× 100GBASE-LR1, MPO-12 APC, 10 km parallel SMF (4×100G breakout, 1 module per leaf)	2
C8	Cables (leaf↔mgmt)	CB-M12-4LC-SMF5M=	MPO-12 to 4× LC duplex SMF breakout cable, 5 m (paired with the C7 4×100G breakout module)	2
C9	Licenses	C1A1TN9300XF2-3Y	Cisco Data Center Networking (DCN) Advantage subscription for Cisco Nexus 9364C and 9300-GX fixed platforms (N9300 XF2); 3-year	4
C10	Licenses	SVS-L2N9KA-XF2-3Y	Mandatory Cisco Support Signature (SVS) license for base DCN Advantage license above; 3-year	4
Orchestration and management				
D1	Fabric management	ND-CLUSTER-L4	Cisco Nexus Dashboard 3-node cluster	1
D2	Management	UCSX-M8-MLB	Cisco UCS X9508 blade server chassis with 3x control nodes for HA	1
D3	Chassis	UCSX-9508-D-U	Cisco UCS X9508 blade server Chassis	1

#	Type	PID	Description	Qty
D4	Server	UCSX-215C-M8	Cisco UCS X-Series compute node (no CPU/memory)[2]	3
D5	CPU	UCSX-CPU-A9375F	AMD EPYC 9375F 3.8 GHz 320 W 32-core / 256 MB DDR5-6000 MT/s	6
D6	Memory	UCSX-MRX32G1RE5	32 GB DDR5-6400 RDIMM 1Rx4 (16 Gb) – 12 DIMMs/blade (6/ socket balanced); 384 GB/blade	36
D7	Boot drive	UCSX-NVM2-960GB	960 GB M.2 boot NVMe	6
D8	NIC	UCSX-MLV5D200GV2D	Cisco VIC 15230 2x 100G mLOM X-Series w/ secure boot	3
D9	Fabric interconnect	UCSX-S9108-100G	Cisco UCS X-Series Direct fabric interconnect (FI)	2
D10	Transceivers (X-Direct uplink)	QSFP-100G-FR-S	Cisco 100GbE transceiver, LC, Duplex, SMF, 2 km	8
D11	Licenses	DC-MGT-SAAS	Cisco Intersight Advantage SaaS license for K8s control nodes	3
D12	Licenses	RH-OCP-B-P3S=	Red Hat OpenShift Container Platform (bare metal), Premium 3-year SnS	0[3]
D13	OOB management switch	N9K-C93108TC-FX3	Cisco Nexus 9300, 48p 100M / 1 / 10GBASE-T + 6p 40 / 100G QSFP28 (HA pair)	2
D14	Cables (OOB 10G)	CAT6A	Copper cable for 10G (2× OCP NIC per server)	8
D15	Cables (OOB 1G)	CAT5E	Copper cable for 1G	22[6]
D16	Management node	UCSC-C220-M8S	Cisco UCS C220 M8 1U rack server – OpenShift Management Node	1

#	Type	PID	Description	Qty
D17	Cisco Cloud Control	N/A	No separate CCC license is required; Included with Cisco Intersight subscriptions	A/R
Racks				
F1	Rack	Cisco R42612	Cisco R-Series 42U rack for UCS and Nexus (1× GPU rack + 1× mgmt/network rack at the 4-node design point)	2
F2	PDU (3-phase)	N/A	Customer-selected 3-phase PDU (2 per rack – A and B feeds)	4
Support/services				
G1	DC support services	CON-CXP-DCC-SAS	Solution-attached services for DC – cloud and compute	A/R[5]
G2	Solution+ services	MINT-COMPUTE	DC compute mentored installation – MINT	A/R[5]

[1] SnS = Software and support

[2] Intel-based Cisco UCS X210c M8 Compute Node is also supported.

[3] Qty=0 for control-plane nodes; for compact (control + worker) nodes, increase quantity.

[4] CIMC = Cisco Integrated Management Controller

[5] A/R = As required

[6] Per host: 1× Server CIMC[4] + 1× B3220 N-S DPU OOB; plus 14 shared mgmt cables – 2× FI mgmt + 3× ND CIMCs + 1× Mgmt Node CIMC + 8× fabric switch mgmt; split across both OOB switches for HA

BOM Notes:

- **Scaling.** This design is endorsed by NVIDIA up to 32 nodes / 256 GPUs. Beyond that endorsed scale point in this document, the same architecture extends to 64 nodes / 512 GPUs by adding a second pair of leaf switches (the same model) to the existing spine pair in the backend fabric. On the frontend fabric, no additional leaf switches are necessary, because the existing leaf pair can cover this scale.
- **Storage.** External storage is not included in this BOM but will connect to the frontend (north/south) fabric.

Transceivers

The BOM lists one transceiver and cable set for each link in this design. The Cisco Transceiver Module Group (TMG) portfolio supports additional options at the same speeds and form factors; the resources below help identify alternates between Cisco Nexus switches, Cisco UCS adapters, and the NVIDIA BlueField-3 SuperNICs and DPUs.

- Cisco TMG Compatibility Matrix: https://tmgmatrix.cisco.com
- COPI — Cisco Optics Product Information: https://copi.cisco.com
- OPTSEL — Cisco Optics Selector: https://optsel.cisco.com
- Cisco QSFP-DD800 Transceiver Modules Data Sheet: https://www.cisco.com/c/en/us/products/collateral/interfaces-modules/transceiver-modules/qsfp-dd800-transceiver-modules-ds.html

Cabling options

The frontend leaf-to-server link (Cisco Nexus 9364D-GX2A leaf to the BlueField-3 B3220 DPU on each Cisco UCS C845A M8 Rack Server) can be implemented in three ways, listed below in order of increasing reach and cost. The BOM in this document uses the MMF optical breakout (default); the alternatives apply to deployments with different physical layouts or cost constraints.

Option	Leaf side	Node side	Cable	Notes
DAC breakout (≤3 m, passive copper)	N/A (integrated into cable)	N/A (integrated into cable)	QDD-2Q200-CU1M / CU2M / CU2.5M / CU3M	Within rack; lowest power and latency
MMF optical breakout (≤100 m, OM4)	QDD-400G-SR8-S (MPO-16 APC, OM4)	2x QSFP-200G-SR4-S (MPO-12 UPC, OM4)	MPO-16 APC to 2x MPO-12 UPC OM4 breakout trunk	Intra-row; field-replaceable optics, one 400G leaf port serves two 200G server ports
SMF point-to-point (≤2 km)	QSFP-200G-FR4-S (LC duplex SMF)	QSFP-200G-FR4-S (LC duplex SMF)	LC duplex SMF patch cord (PC/UPC)	Across rows / SMF cabling standard; shares the frontend spine-leaf optic; uses two 400G leaf ports per node (each in 200G mode)

Licenses

Additional information on licensing for the components in the Cisco AI POD stack is available at:

- Cisco Intersight licensing: https://intersight.com/help/saas/getting_started/licensing_requirements/lic_intro
- Cisco NX-OS licensing options guide: https://www.cisco.com/c/en/us/td/docs/switches/datacenter/licensing-options/cisco-nexus-licensing-options-guide.html
- NVIDIA licensing (Cisco ordering reference): https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/release/notes/nvidia_ordering-guide.html
- Red Hat AI Enterprise licensing: https://www.redhat.com/en/products/ai

Conclusion

This white paper documents a specific Cisco AI POD Enterprise Reference Architecture, which is NVIDIA endorsed for Infrastructure Configuration. The design is built on Cisco UCS C845A M8 Rack Servers with NVIDIA RTX PRO 6000 Blackwell Server Edition (or NVIDIA H200 NVL) GPUs, managed by Cisco Intersight and Cisco Nexus 9000 Series network fabrics managed using Cisco Nexus Dashboard. This design is endorsed for scale points of 4 to 32 UCS nodes and up to 256 GPUs.

The design is purpose-built for enterprise inference and small-model training and fine-tuning workloads. It is intended as a starting blueprint that partners and customers can deploy and grow predictably from a 4-node / 32-GPU to a 32-node / 256-GPU cluster, without changing the topology or the operational model as they scale.

References

The primary references for this Reference Architecture are listed below.

Cisco UCS and Intersight

- Cisco UCS C845A M8 Rack Server Data Sheet – https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c845a-m8-rack-server-ds.html
- Cisco UCS C845A M8 Rack Server Spec Sheet – https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c845a-m8-rack-server-spec-sheet.pdf
- Cisco UCS C845A M8 Getting Started Guide – https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/ucs-c845a-m8-rack-server-og.html
- Intersight Help: Managing UCS C845A M8 Server – https://intersight.com/help/saas/resources/managing_ucs_c845a_m8_server
- Cisco UCS X-Series Direct Data Sheet – https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/ucs-x-series-direct-ds.html
- Cisco Intersight Data Sheet – https://www.cisco.com/c/en/us/products/collateral/cloud-systems-management/intersight/intersight-ds.html

NVIDIA GPUs, SuperNICs, DPUs, and software

- NVIDIA RTX PRO 6000 Blackwell Server Edition – https://www.nvidia.com/en-us/data-center/rtx-pro-6000-blackwell-server-edition/
- NVIDIA H200 NVL – https://www.nvidia.com/en-us/data-center/h200/
- NVIDIA Ethernet SuperNICs (BlueField-3 SuperNIC product page) – https://www.nvidia.com/en-us/networking/products/ethernet/supernic/
- NVIDIA BlueField-3 Networking Platform Specifications (SuperNICs and DPUs) – https://docs.nvidia.com/networking/display/BF3DPU/Specifications
- NVIDIA AI Enterprise Support Matrix – https://docs.nvidia.com/ai-enterprise/latest/product-support-matrix/

NVIDIA Enterprise RA

- NVIDIA Enterprise Reference Architectures (NVIDIA documentation hub) – https://docs.nvidia.com/enterprise-reference-architectures/index.html. Canonical landing page for published Enterprise RAs and partner-endorsed designs based on Enterprise RA patterns
- **ERA-00004-001** – NVIDIA 2-8-5-200 GPU Node Configuration and NVIDIA Spectrum-X Platforms (NVIDIA partner portal). This is the canonical NVIDIA reference document for the GPU node pattern used in this design.
- **ERA-00008-001** – Network Deployment Guide for NVIDIA Spectrum-X Platforms (NVIDIA partner portal). This is the canonical NVIDIA reference for backend-fabric configuration (RoCEv2, DLB, PFC, ECN, and adaptive routing) supporting GPUDirect RDMA.

Cisco Nexus switches and Cisco Nexus Dashboard

- Cisco Nexus 9364E-SG2 Switch Data Sheet – https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/nexus-9364e-sg2-switch-ds.html
- Cisco Nexus 9300 GX2 Series Fixed Switches Data Sheet (includes 9364D-GX2A) – https://www.cisco.com/site/us/en/products/collateral/networking/switches/nexus-9000-series-switches/nexus-9300-gx2-series-fixed-switches-data-sheet.html
- Cisco Nexus Dashboard Data Sheet – https://www.cisco.com/c/en/us/products/collateral/data-center-analytics/nexus-dashboard/datasheet-c78-744371.html