NetApp Verified Architecture

FlexPod Enterprise Reference Architecture with Cisco AI POD

Reference Design for FlexPod

NVA-1186-DESIGN-0426

By NetApp Inc.

In partnership with

Abstract

Cisco AI POD with FlexPod aligns with the NVIDIA Enterprise Reference Architecture, delivering validated, performance-tested infrastructure designs for building scalable and reliable AI factories. By conforming to NVIDIA Enterprise RA design principles for GPU-accelerated compute, high-performance networking, and efficient data pipelines, the solution with FlexPod helps ensure predictable performance and consistent behavior across AI training and inference workloads. This alignment with NVIDIA and Cisco reference architectures enables customers and partners to deploy AI infrastructure with confidence accelerating innovation while reducing deployment complexity, operational risk, and time to value.

TABLE OF CONTENTS

Executive summary

This architecture brings together NVIDIA's software and GPUs, accelerated Cisco computing stack with Cisco's high-performance data center networking (400G/800G), unified operations, and zero-trust security, along with NetApp enterprise storage and data management capabilities, to make AI clusters faster to deploy, easier to operate, and more cost-effective to scale across data centers and distributed edge environments. Day-0 to Day-2 operations are streamlined through integrated management and observability, enabling consistent lifecycle management, visibility, governance, and data protection as AI environments grow.

Cisco AI POD with FlexPod represents an integrated, three-way offering from Cisco, NVIDIA, and NetApp combining Cisco AI POD infrastructure, NVIDIA accelerated computing and AI software, and the proven FlexPod converged infrastructure foundation with data services through NetApp AIDE (AI Data Engine) for the AI data pipeline. This modular design is intended to be replicated as a repeatable scale unit for training, fine-tuning, RAG, and inference workloads, enabling enterprises to deploy AI infrastructure with predictable performance, built-in resiliency, and simplified lifecycle management.

Scope

This blueprint with FlexPod is aligned to the NVIDIA Enterprise Reference Architecture and provides a set of standardized options: validated configurations that use Cisco UCS C885A M8, C845A M8, and C240 M8 rack servers with Cisco Nexus 9000 Series switches. The architecture specifies the design principles, logical building blocks, and deployment considerations required to deliver consistent performance, resiliency, and scalable growth for enterprise AI workloads.

Solution overview

This Enterprise RA provides a guidance for the following enterprise-grade, best-in-class components:

- Cisco UCS C240 M8 Rack servers with NVIDIA RTX 6000 PRO GPUs plus 1 * ConnectX®-7 NIC with 2 * 200GbE ports for N/S traffic
- Cisco UCS C845A M8 Rack servers with NVIDIA H200 GPUs using 4 * PCIe x16 HHHL NVIDIA BlueField-3 B3140H for E/W and 1 PCIe x16 FHHL NVIDIA BlueField-3 B3220 for N/S traffic, aligned with 2-8-5-200 architecture
- Cisco UCS C885A M8 Rack Servers with high-density NVIDIA HGX H200 systems (using 8 * PCIe x16 HHHL NVIDIA BlueField-3 B3140H for E/W and 1 PCIe x16 FHHL NVIDIA BlueField-3 B3220 for N/S traffic) aligned with the 2-8-9-400 pattern shown in the Cisco N9000 NVIDIA Enterprise reference architecture

The BlueField-3 B3220 which is used for the N/S traffic connecting to the NetApp systems should be configured in NIC mode to meet the connectivity requirement.

The servers could be managed through Cisco Intersight or Cisco IMC during the setup and configuration.

- Networking: Cisco Nexus 9000 Series 400/800GbE switches provide a dual-fabric design for FlexPod AI and Cisco AI POD deployments supporting non-blocking, high-performance east/west GPU traffic and north/south storage and client access. Centralized operations with Cisco Nexus Dashboard enable policy-based automation.
- Best practices-based deployment templates and streamlined Day 0–Day 2 lifecycle management and telemetry to get the insights.
- Storage: NetApp AFX delivers high-throughput, low-latency data layer required for AI datasets, feature stores, model checkpoints, and artifacts. Built-in enterprise data services and ransomware

protection help enable governance and cyber resiliency, while integration with AIDE accelerates the AI data pipeline and simplifies data mobility across on-premises and hybrid cloud environments.

Solution components

The following patterns are recommended for AI and high-performance computing (HPC) workloads in a FlexPod solution.

Cisco UCS C240 M8 Rack Server

Use cases : Inference

Cisco UCS C240 M8 is a 2RU, dual-socket rack server built for performance and expandability. It supports Intel ® Xeon ® 6 processors, large memory and storage configurations, PCIe GPUs for AI Acceleration, and virtualization.

Cisco UCS C845A M8 Rack Server

Use cases : Modular (starting with 4 * GPUs) Inference, Training, Fine Tuning, HPC

Cisco UCS C845A [AMD CPU] M8 is a flexible AI-optimized rack server that supports 2 to 8 NVIDIA PCIe GPUs.

Cisco UCS C885A M8 Rack Server

Use cases : Pre-configured (with 8 GPUs per node) for Large-scale model training, Fine tuning, Inference & HPC

Cisco UCS C885A M8 [AMD CPU] is an 8RU rack server with 8 NVIDIA SXM GPUs that provides the GPU performance needed for AI model training and fine-tuning.

Network: Cisco Nexus switches

In a Cisco AI POD with FlexPod, the network fabric is the critical foundation that interconnects GPU-dense Cisco UCS servers and delivers the high bandwidth, predictable low latency, and lossless transport required for distributed training and fine-tuning. The design commonly uses a dual-fabric approach: a backend (east–west) fabric optimized for GPU-to-GPU traffic and a frontend (north–south) fabric for connectivity to storage, management, shared services, and users. Cisco Nexus switches can be deployed as leaf and spine roles to provide the port density, switching capacity, and AI-optimized capabilities needed for these workloads, including lossless Ethernet and RDMA over Converged Ethernet (RoCE). Both fabrics can be deployed as repeatable, scalable building blocks and centrally automated, monitored, and managed with Cisco Nexus Dashboard.

FlexPod® AI supports both Cisco cloud-scale switching platforms and Cisco Silicon One based data center switches across the AI fabric. While this architecture supports either option end-End-to-end, the validation focuses on Silicon One based switches for the backend AI fabric and cloud-scale switches for the frontend and external connectivity, reflecting common enterprise deployment patterns.

Cisco Nexus 9364E-SG2 switch

The Cisco Nexus 9364E-SG2 is a high-density, 2RU 800GbE switch built for next-generation data centers. Powered by Cisco Silicon One® technology, it enables high-performance, power-efficient connectivity to support modern cloud architectures and the throughput demands of AI workloads. In a Cisco AI POD with FlexPod, the 9364E-SG2 can be deployed as a high-density leaf or a compact spine across both the backend and frontend fabrics, providing the port density and switching capacity required for scalable AI deployments.

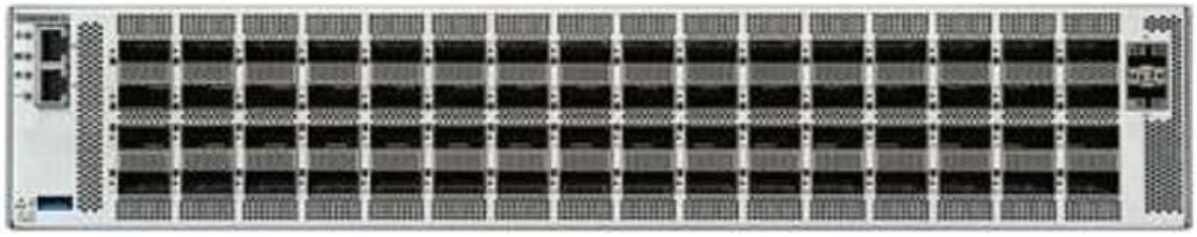

Figure 1) Cisco Nexus 9364E-SG2

The Nexus 9364E-SG2 offers 64 × 800GbE ports and is available with either QSFP-DD or OSFP interfaces. The ports provide flexible speed and density options, with the ability to operate at 400GbE, 200GbE, or 100GbE as required. With up to 51.2 Tbps of traffic forwarding capacity or packet forwarding, the switch is designed to meet high aggregate bandwidth demands. A 256 MB on-die buffer helps absorb traffic microbursts and reduce packet loss—an important requirement for AI/ML data flows.

For AI/ML fabrics such as those used in a Cisco AI POD with FlexPod the Nexus 9364E-SG2 delivers low latency, advanced congestion management, and rich telemetry. It supports dynamic load balancing (DLB), priority flow control (PFC), and explicit congestion notification (ECN), enabling lossless transport for RDMA over Converged Ethernet (RoCE). When integrated with Cisco Nexus Dashboard, it also provides centralized security, automation, visibility, analytics, and assurance to help operate high-performance AI/ML networks at scale.

Cisco Nexus 9364D-GX2A switch

The Cisco Nexus 9364D-GX2A (N9K-C9364D-GX2A) is a high-density 400GbE fixed-configuration data center switch in the Nexus 9300-GX2 family built for large-scale leaf/spine fabrics that need high throughput and predictable low latency. It provides 64×400GbE QSFP-DD ports plus 2×1/10GbE SFP+ ports, delivers up to 51.2 Tbps of switching capacity (up to 8.35 billion pps), and supports flexible speed options and common breakout modes (including 4×10G, 4×25G, 4×50G, 4×100G, and 2×200G per 400G port). The first 16 ports support wire-rate MACsec, and the switch can run in either NX-OS or ACI mode, making it a strong fit as a compact 2RU high-radix spine or high-density leaf for modern cloud, AI/HPC, east-west traffic, and high-bandwidth storage networks scaling to 100/200/400GbE.

.

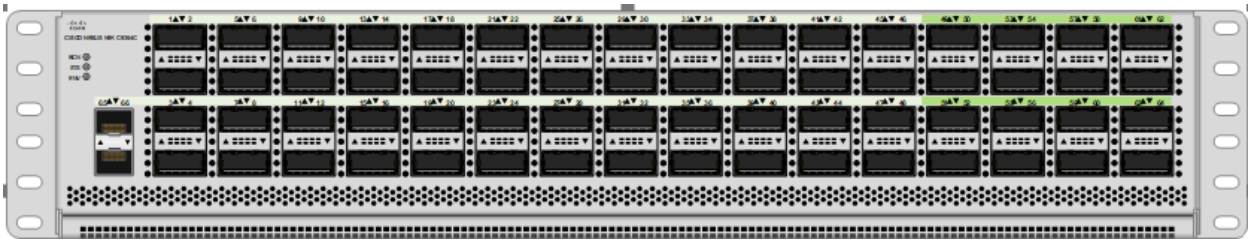

Figure 2) Nexus 9364D-GX2A

Cisco Nexus 9332D-GX2B switch

The Cisco Nexus 9332D-GX2B is a 1RU, fixed-configuration switch designed for high-density 400GbE deployments. It is typically deployed as a leaf switch in the frontend fabric, connecting multiple GPU servers to the enterprise network.

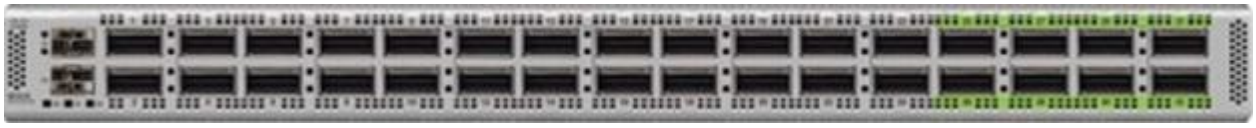

Figure 3) Cisco Nexus 9332D-GX2B

The Nexus 9332D-GX2B provides 32 x400GbE QSFP-DD ports. Each 400GbE port can be configured for multiple speeds, including 100GbE, 50GbE, 25GbE, and 10GbE, to meet a range of connectivity requirements. The switch delivers 25.6 Tbps of forwarding throughput and 8.5 billion packets per second, enabling efficient packet processing for latency-sensitive AI traffic. A 60 MB shared buffer helps absorb traffic bursts and reduce packet loss which is critical for maintaining lossless communication in AI/ML environments.

Cisco Intersight

Cisco Intersight® provides end-to-end infrastructure lifecycle management and is delivered as a cloud-Cisco Intersight® is a cloud-native IT operations platform that provides end-to-end infrastructure lifecycle management. Available as SaaS, a connected virtual appliance, and a private (air-gapped) virtual appliance, it provides IT teams with a unified, real-time view and centralized management of Cisco UCS across data centers, colocations, and edge environments through a single dashboard. Using policy-based automation, assisted operations capabilities, and an API-first architecture, Intersight streamlines deployment, configuration, and ongoing maintenance, while supporting consistent compliance and a strong security posture across the entire server estate. In this reference architecture Cisco Intersight provides unified, SaaS-based management for all Cisco UCS infrastructure.

Cisco Nexus Dashboard

Cisco Nexus Dashboard (ND) is the unified management and operations platform for deploying and operating the network fabrics in the AI POD architecture. It provides validated, best-practice blueprints for implementing both the high-speed backend (E/W) and frontend (N/S) fabrics commonly used in AI/ML environments. Through a software defined approach, ND enables centralized, fabric-level deployments rather than manual, switch-by-switch configuration, thereby promoting configuration consistency and reducing the likelihood of errors. This “one fabric” operating model simplifies administration and offers a single API endpoint for automation. ND also supports day-to-day operations through integrated monitoring, AI-driven traffic management, customizable dashboards, and multi-fabric lifecycle management. Its scale-out architecture and API-based automation capabilities provide a simplified, scalable, and intelligent platform for enterprise data center fabric operations.

NetApp AFX 1K

The NetApp AFX 1K has achieved NVIDIA-Certified Storage at the Enterprise Level. NetApp storage reference architecture supports high performance, scalability, and flexibility for AI workloads by optimizing data pipelines from ingestion to model training, fine tuning, and inference using NetApp data fabric. NetApp ONTAP and AI solutions manage large datasets across training, fine tuning, and inference. NetApp provides a unified, scalable platform to accelerate AI deployment with pre-configured infrastructure, ensuring seamless integration with NVIDIA Enterprise Reference Architectures. Disaggregated AFX architectures separate compute, storage, and networking so each layer can scale independently and be optimized for specific workloads. NetApp ONTAP provides the consistent storage software layer for this model delivering unified file and object access, proven data protection, and hybrid-cloud mobility so AFX teams can modernize infrastructure without compromising operational simplicity or resilience.

ONTAP with disaggregated AFX

Decouple performance and capacity: Scale storage resources independently from AFX compute, aligning cost to demand while maintaining predictable latency

Standardize data services: Use one storage operating environment across tiers to simplify provisioning, protection, and compliance

Enable workload placement: Present the right protocol to each application while keeping data management consistent

Support hybrid strategies: Extend ONTAP data services to cloud targets for backup, disaster recovery, or burst workloads

AI/ML and analytics pipelines: High throughput file access with scalable capacity for training data and feature stores

NVIDIA Enterprise Reference Architecture

Cisco UCS C240 M8 configuration - AI Acceleration

(Figure 4) UCS C240 M8 with 2 GPUs (NVIDIA Blackwell RTX™ PRO 6000 Server Edition, NVIDIA H200 NVL) is the smallest configuration which comes with PCIe optimized NVIDIA-Certified compute node with 1 * ConnectX®-7 NIC with 2 * 200GbE ports, and two CPUs.

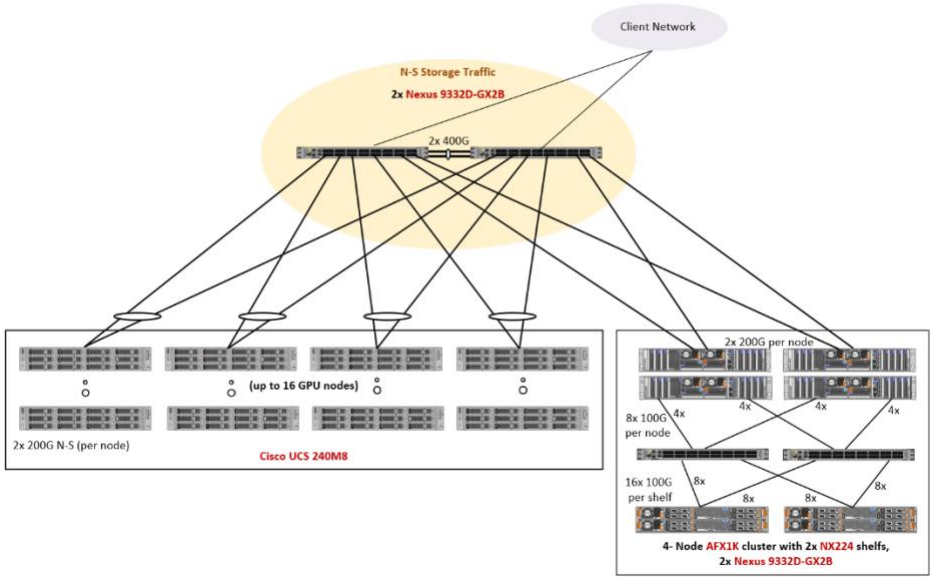

Figure 4) Cisco UCS C240 M8 configuration

This can scale from 1 SU up to 4 SUs in a cluster. Table 1 shows storage scaling requirements for up to four scale units and 32 GPUs. Each scalable unit provides the following connectivity components:

- For the N/S fabric: On 4 servers (SU), each equipped with ConnectX-7 NIC (2x 200GbE ports per server), providing a total of 8x 200Gb/s connections and an aggregate bandwidth of 1.6Tb/s.
- For the storage side: Each NetApp AFX 1K node will be equipped with 2x X50131B dual port adapters with only one port being used per adapter. Eight storage ports are used for this configuration to connect to the frontend switches, each providing 200Gb/s bandwidth. Only the uplinks from two controllers are shown in the diagram to minimize visual complexity.
- One pair of Cisco Nexus 9332D-GX2B 400GbE switches are used for N/S frontend (F/E) connectivity.
- The filesystem mounts should be using NFSoRDMA and pNFS should be configured on the NetApp storage which provides consistent latency as the solution is scaled.

- The front-end switches (N/S) are configured as vPC pairs. 2x 400GbE links are used for vPC peer-link. The C240 M8 servers connect to the front-end switches over vPC. The storage controllers connect to the front-end switches over non-vPC links.

Table 1) ERA for MGX with Cisco UCS C240 M8

Nodes C240 M8	GPU	Storage Nodes - AFX 1K	Server Ports N/S 200GbE	Storage Ports 200GbE	Total switch ports/ QDD-400G-SR8-S Transceivers	MPO16 Cables	QSFP-200G-SR4-S Transceivers	Cables [MPO16 - 2 * MPO12]	9332-GX2B (N/S)
4	8	4	8	8	12	2	16	8	2
8	16	4	16	8	16	2	24	12	2
12	24	4	24	8	20	2	32	16	2
16	32	4	32	8	24	2	40	40	2

Cisco UCS C845 M8 configuration – 2-8-5-200

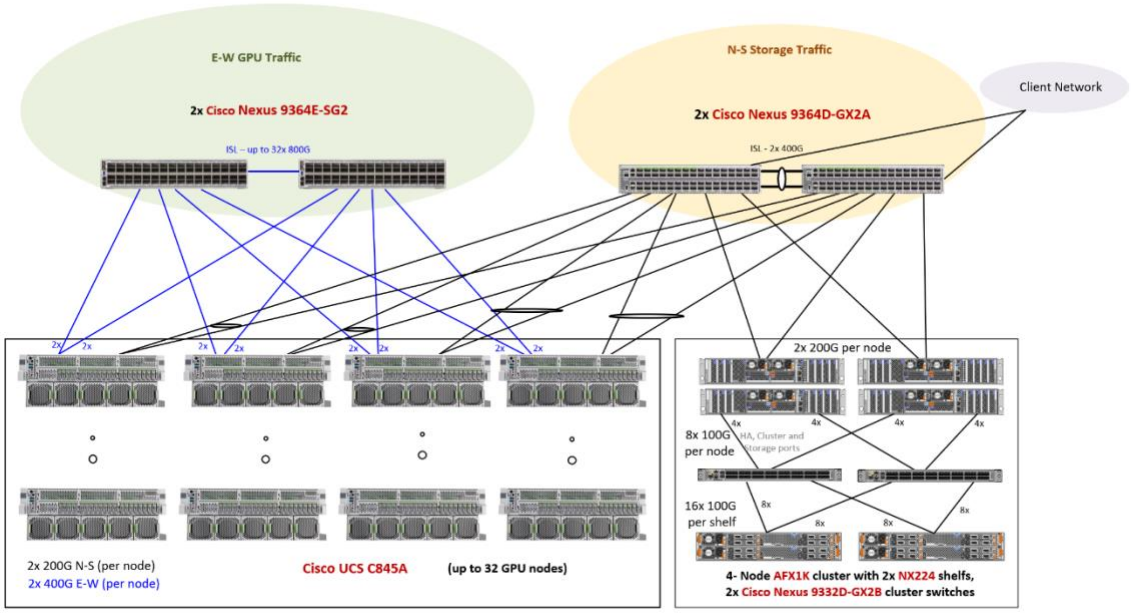

Figure 5) Cisco UCS C845A M8 in MGX configuration

A PCIe optimized NVIDIA-Certified compute node supporting eight GPUs (NVIDIA RTX 6000 PRO, H200 NVL), five network adapters (4 * PCIe x16 HHHL NVIDIA BlueField-3 B3140H for E/W and 1 PCIe x16 FHHL NVIDIA BlueField-3 B3220 N/S), and two CPUs. This can scale from 1 SU up to 8 SUs in a cluster as shown in Figure 5.

Table 2 shows storage scaling requirements for up to eight scale units and 256 GPUs. Each scalable unit provides the following connectivity components:

- For the E/W fabric: On 4 servers (SU), each equipped with 4x B3140H BlueField-3 SuperNICs (1x 400GbE port), providing a total of 16x 400Gb/s connections and an aggregate bandwidth of 6.4Tb/s.
- For the Converged N/S fabric for storage: On 4 servers, each equipped with 1 * PCIe x16 FHHL NVIDIA BlueField-3 B3220 for N/S, providing a total of 8x 200Gb/s connections and an aggregate

bandwidth of 1.6Tb/s. Eight storage ports are used for this configuration to connect to the frontend switches, each providing 200Gb/s bandwidth. Only the uplinks from two controllers are shown in the diagram to minimize visual complexity.

- Cisco Nexus 9364E-SG2-O is used as spine and leaf for backend network. The front-end fabric uses Cisco Nexus 9364D-GX2A switches for spine and leaf.
- For the storage side: Each NetApp AFX 1K node will be equipped with 2 * X50131B dual port adapter
- Eight storage ports are used for 2-8-5-200 configuration, each port providing 200Gb/s bandwidth.
- The GPU nodes and storage controllers connect to the front-end and backend switches over non-vPC links.
- The filesystem mounts should be using NFSoRDMA and pNFS should be configured on the NetApp storage which provides consistent latency as solution is scaled
- Table 3 shows the details of N/S topology which covers the compute to network to storage connections
- The bandwidth between the spine and leaf must equal the bandwidth coming into the leaf from the servers to have a fully non-blocking back-end network

Table 2) ERA for MGX with Cisco UCS C845A M8 in 2-8-5-200 listing E/W Topology

Nodes Cisco C845 M8	GPUS	QDD-400G-DR4-S on the GPU node (400GbE)	Leaf-Leaf Total ports (800GbE)	Total switch ports/OSFP-800G-DR8	CB-M12-M12-SMF MPO-12 Cable	E/W Networking 9364E-SG2	
						Leaf	Spine
4	32	16	8	16	32	2	N/A
8	64	32	16	32	64	2	N/A
12	96	48	24	48	96	2	N/A
16	128	64	32	64	128	2	N/A
20	160	80	40	80	160	2	N/A
24	192	96	48	96	192	2	N/A
28	224	112	56	112	224	2	N/A
32	256	128	64	128	258	2	N/A

Table 3) ERA for MGX with UCS C845A M8 in 2-8-5-200 listing N/S Topology

GPU Nodes	GPU	Storage Nodes - AFX 1K	Server Ports N/S (200GbE)	Storage Ports (200GbE)	Total switch ports N/S (400GbE)	QSFP-200G-SR4-S	Cables [MPO16 - 2 * MPO12]	Transceivers QDD-400G-SR8-S	Leaf to leaf Cables MPO16	N/S Networking 9364D-GX2A	
										Leaf	Spine
4	32	4	8	8	12	16	8	12	2	2	N/A
8	64	4	16	8	16	24	12	16	2	2	N/A
12	96	4	24	8	20	32	16	20	2	2	N/A
16	128	4	32	8	24	40	20	24	2	2	N/A
20	160	4	40	8	28	48	24	28	2	2	N/A
24	192	4	48	8	32	56	28	32	2	2	N/A
28	224	4	56	8	36	64	32	36	2	2	N/A
32	256	4	64	8	40	72	36	40	2	2	N/A

Cisco UCS C885 M8 configuration – 2-8-9-400

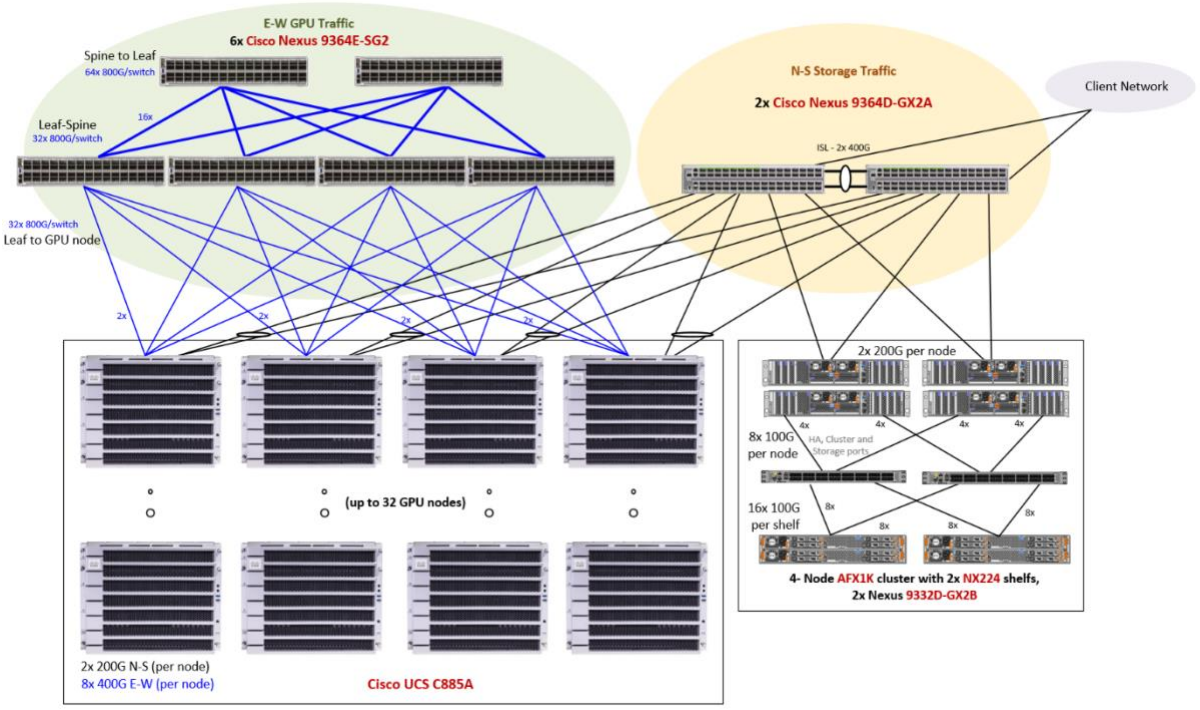

Figure 6) Cisco UCS C885A M8 HGX Configuration

2-8-9-400 is a SXM (NVLink-connected GPUs) supporting eight GPUs, nine network adapters (8 * PCIe x16 HHHL NVIDIA BlueField-3 B3140H for E/W and 1 PCIe x16 FHHL NVIDIA BlueField-3 B3220 N/S), and two CPUs. It scales from 1 SU to 8 SUs in a cluster. The topology of the 2-8-9-400 is shown in Figure 6.

Table 4 shows storage scaling for up to eight scale units and 256 GPUs. Each scalable unit provides the following connectivity building blocks:

- For the E/W fabric: On 4 servers (SU), each equipped with 8x B3140H BlueField-3 SuperNICs (1x 400GbE port), providing a total of 32x 400Gb/s connections and an aggregate bandwidth of 12.8Tb/s
- For the Converged N/S fabric for storage: On 4 servers, each equipped with 1 * PCIe x16 FHHL NVIDIA BlueField-3 B3220 for N/S, providing a total of 8x 200Gb/s connections and an aggregate bandwidth of 1.6Tb/s. Eight storage ports are used for this configuration to connect to the frontend switches, each providing 200Gb/s bandwidth. Only the uplinks from two controllers are shown in the diagram to minimize visual complexity
- For the storage side: Each NetApp AFX 1K node will be equipped with 2 * X50131B dual port adapter
- Cisco Nexus 9364E-SG2-O is used as spine and leaf for backend network. The front-end fabric uses Cisco Nexus 9364D-GX2A switches for spine and leaf.
- The GPU nodes and storage controllers connect to the front-end and backend switches over non-vPC links.
- The filesystem mounts should be using NFSoRDMA and pNFS should be configured on the NetApp storage which provides consistent latency as solution is scaled
- Table 5 shows the details of N/S topology which covers the compute to network to storage connections
- The bandwidth between the spine and leaf must equal the bandwidth coming into the leaf from the servers to have a fully non-blocking back-end network

Table 4) ERA for HGX with Cisco UCS C885A M8 in 2-8-9-400 listing E/W Topology

Nodes Cisco C885A M8	GPU	Server Ports E/W (400GbE)	Total switch to server ports E/W (800GbE)	Leaf-Leaf 800G ports	OSFP-800G-DR8 OSFP	CB-M12-M12-SMF MPO-12 Cable	QDD-400G-DR4-S on the GPU node	E/W Networking 9364E-SG2	
								Leaf	Spine
4	32	32	16	16	32	48	32	2	N/A
8	64	64	32	32	64	128	64	2	N/A
12	96	96	48	48	96	144	96	2	N/A
16	128	128	64	64	128	192	128	2	N/A
20	160	160	80	-	240	320	160	4	2
24	192	192	96	-	288	384	192	4	2
28	224	224	112	-	336	448	224	4	2
32	256	256	128	-	384	512	256	4	2

Table 5) ERA for HGX with Cisco UCS C885A M8 in 2-8-9-400 listing N/S Topology

GPU Nodes	GPU	Storage Nodes - AFX 1K	Server Ports N/S (200GbE)	Storage Ports (200GbE)	Total switch ports N/S (400GbE)	QSFP-200G-SR4-S Transceivers for GPU node	Cables [MPO16 - 2 * MPO12	QDD-400G-SR8-S Transceivers	Leaf to Leaf Cables MPO16	N/S Networking 9364D-GX2A	
										Leaf	Spine
4	32	4	8	8	12	16	8	12	2	2	N/A
8	64	4	16	8	16	24	12	16	2	2	N/A
12	96	4	24	8	20	32	16	20	2	2	N/A
16	128	4	32	8	24	40	20	24	2	2	N/A
20	160	4	40	8	28	48	24	28	2	2	N/A
24	192	4	48	8	32	56	28	32	2	2	N/A
28	224	4	56	8	36	64	32	36	2	2	N/A
32	256	4	64	8	40	72	36	40	2	2	N/A

Data Services – with NetApp AIDE

NetApp AIDE – is a storage integrated AI data service designed to make enterprise data "AI ready" for GenAI use cases like RAG, agentic AI, and AI factories by reducing data sprawl, keeping data fresh, and enforcing governance/security close to where the data live.

AIDE is supported on Cisco C240 M8 server (minimum 3 * C240 M8) and can be configured to any of the three Enterprise reference architecture mentioned above (Cisco UCS C240 M8 configuration - AI Acceleration, NVIDIA Enterprise reference architecture 2-8-5-200 or NVIDIA Enterprise reference architecture 2-8-9-400). AIDE is positioned as part of NetApp's AI portfolio alongside (Figure 7, sample configuration using 2-8-9-400) systems like AFX and is built on ONTAP to bring AI data services "to the data", rather than copying data into many separate tools and in practical terms:

- Builds a global, structured view of your data estate (especially unstructured file data) via a large-scale metadata catalog/engine, so teams can discover and understand what data exists
- Curates data into trusted collections so you can pick the "right" documents/objects for an AI use case rather than dumping everything into a pipeline
- Applies policy driven "guardrails" classification, monitoring, and controls to help protect sensitive data and support compliance
- Enables real-time, continuous vectorization and semantic retrieval so curated data can be published as retrieval endpoints for RAG systems, while staying current through automated change detection and sync

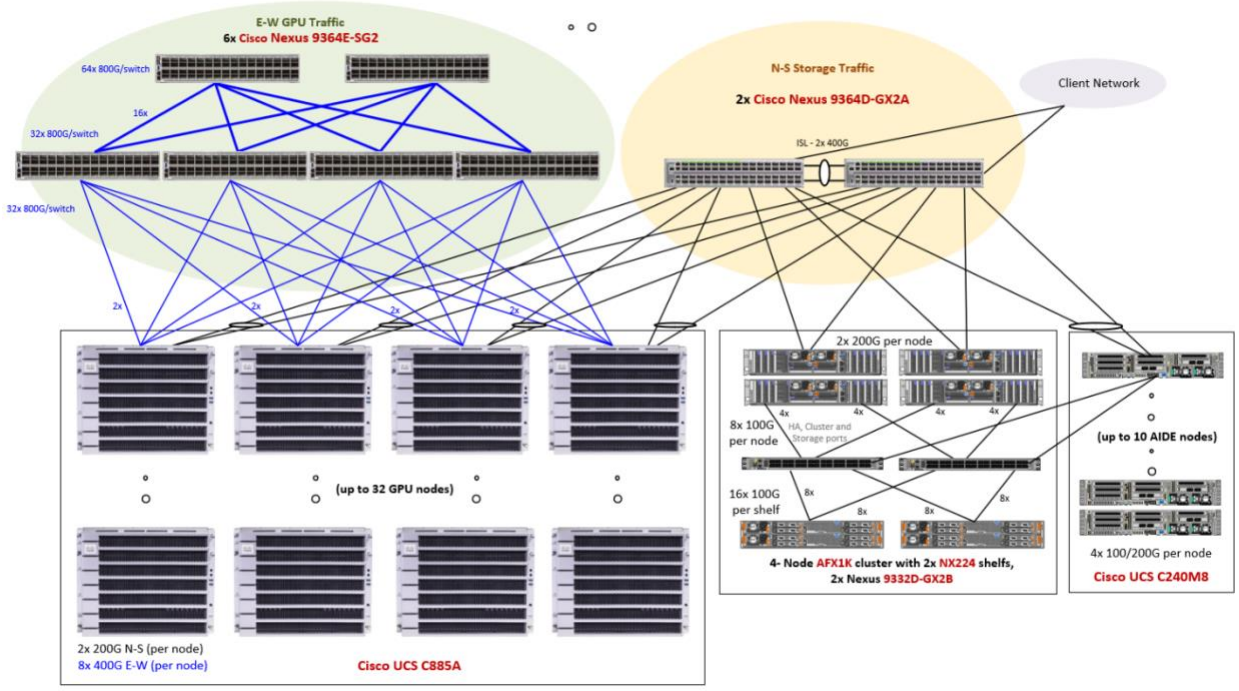

Figure 7) Data Services with AIDE with Cisco UCS C240 M8 in a HGX configuration

Table 6) AIDE Node Scaling

AIDE Nodes C240 M8	GPU RTX 6K	Backend Server Ports (200GbE)	Frontend Server Ports (200GbE)	Total switch ports (E/W) (400GbE)	Total switch ports (N/S) (400GbE)	QSFP-200G-SR4-S	Cables [MPO16 - 2 * MPO12]	QDD-400G-SR8-S
3	6	6	6	4	4	12	8	8
4	8	8	8	4	4	16	8	8
5	10	10	10	6	6	20	12	12
6	12	12	12	6	6	24	12	12
7	14	14	14	8	8	28	16	16
8	16	16	16	8	8	32	16	16
9	18	18	18	9	9	36	20	20
10	20	20	20	10	10	40	20	20

Table 6 shows scaling of AIDE nodes [right side of the topology] which can go up to ten nodes and any increments. The scaling of AIDE nodes depends on the data set size, number of volumes, and ONTAP cluster size. The current architecture supports up to 2 * NVIDIA RTX 6000 PRO GPUs per Cisco UCS C240 M8 server.

- 2 * ConnectX-7 NIC (2x 200GbE ports per server) per AIDE server is a requirement to connect to N/S switches and NetApp storage cluster switches
- For the AIDE frontend fabric: On each server equipped with 1 dual port ConnectX-7 NIC (2x 200GbE ports per server) provides 400Gb/s of aggregated bandwidth and is connected to the 400GbE N/S storage switch
- For the AIDE backend fabric: On each server equipped with 1 dual port ConnectX-7 NIC (2x 200GbE ports per server) provides 400Gb/s of aggregated bandwidth and is connected to the 400GbE AFX cluster switch
- Each AIDE node will have a 4 * port 10/25/50G adapter for any management, services etc

Licensing

Additional licensing information for the various components of the Cisco AI POD stack is available here:

- Cisco Intersight
- Cisco Nexus NX-OS licensing options
- NVIDIA licensing
- Red Hat OpenShift licensing
- ONTAP One Licensing
- AIDE Licensing
- NVIDIA Spectrum Licensing

Conclusion

The NVIDIA Enterprise reference architecture for FlexPod in Cisco AI Pod configuration provides a scalable, validated foundation for enterprise AI workloads, combining NVIDIA accelerated computing with Cisco and NetApp data management. With AIDE integrated into FlexPod end-users will have the capability of having a storage integrated data service designed to simplify, secure, and accelerate AI data pipelines, from ingestion to application serving. The end-to-end solution addresses security, data management, and resource utilization while delivering high availability and storage efficiency to reduce bottlenecks across the AI data pipeline enabling data scientists and ML engineers to focus on innovation and achieve faster time-to-value with the benefits:

- Validated, repeatable design that reduces deployment risk and accelerates time to production
- High throughput, low latency network fabric to move data efficiently between GPU compute, storage, and users
- Optimized GPU utilization through scalable compute building blocks and consistent configuration across nodes
- Storage performance and efficiency designed for the AI data pipeline (ingest, feature engineering, training, inference, and archival) with simplified data management
- High availability and resiliency across compute, network, and storage to minimize downtime and protect long- running training jobs
- Security controls aligned to enterprise requirements, including segmentation and policy driven access to data and infrastructure

- Operational consistency with unified monitoring and lifecycle management to simplify upgrades and scale out expansion

Where to find additional information

To learn more about the information that is described in this document, review the following documents and/or websites:

NetApp Documentation

- NetApp Documentation
- NetApp AI Solutions Documentation
- NetApp Install and Maintain AFX Systems
- NFS over RDMA
- What is pNFS

Cisco UCS M8 Rack Servers & Management Software

- Cisco UCS C885A M8 Rack Server Data Sheet
- Cisco UCS C845A M8 Rack Server Data Sheet
- Cisco UCS C240 M8 Rack Server Data Sheet
- Cisco Intersight Data Sheet
- Intersight At-a-Glance
- Cisco UCS C845A M8 Rack Server at a Glance
- Cisco UCS C885A M8 Rack Server at a Glance

Cisco Nexus Switches

- Cisco Nexus 9332D-GX2B and Nexus 9364D-GX2A Switch Data Sheet
- Cisco Nexus 9364E-SG2 Switch Data Sheet
- Cisco Nexus Dashboard Data Sheet
- Nexus Dashboard, Release 4.2.x User Content
- Cisco Data Center Networking (DCN) Licensing Ordering Guide

Transceivers

To identify the supported transceivers for any component in the Cisco AI POD stack, including interoperability across components, refer to Cisco's Transceiver Matrix Group site. The links below include the main site and the relevant subsites.

- Cisco Optics-to-Device Compatibility Matrix
- Cisco Optics Product Information
- Cisco Optics Selector
- Transceiver data sheets

Version history

As an option, use the NetApp Table style to create a Version History table. Do not add a table number or caption.

Version	Date	Document version history
Version 1.0	May 2026	New document

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright information

Trademark information