AI Infrastructure with Cisco Nexus 9000 Switches

Cisco Enterprise Reference Architecture

Contents

Featuring Cisco UCS® C885A M8 Rack Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X

Introduction

Cisco® Enterprise Reference Architecture (ERA) is based on Cisco Nexus® 9000 Series Switches for networking AI clusters managed by the on-premises Cisco Nexus Dashboard platform. It adheres to the NVIDIA Enterprise Reference Architecture for NVIDIA HGX H200 with NVIDIA Spectrum-X networking.

Cisco Nexus 9000 Series Switches, powered by Cisco Silicon One® and Cisco Cloud Scale architectures, provide high-speed, deterministic, low-latency, and power-efficient connectivity for AI and high-performance computing (HPC) workloads. With the availability of multiple form-factors, optics, and rich software features of the Cisco NX-OS operating system, Nexus 9000 switches provide a consistent experience for frontend, storage, backend, and out-of-band (OOB) management networks (see Figure 1).

Cisco Nexus Dashboard is the operations and automation platform for managing the Nexus 9000 switch-based fabrics. It complements the data plane features of the Nexus 9000 switches by simplifying their configuration using built-in templates. It detects network health issues, such as congestion, bit errors, and traffic bursts, in real time and automatically flags them as anomalies. These issues can be resolved faster using integrations with commonly used tools, such as ServiceNow and Ansible, allowing the networks of an AI cluster to be aligned with the existing workflows of an organization.

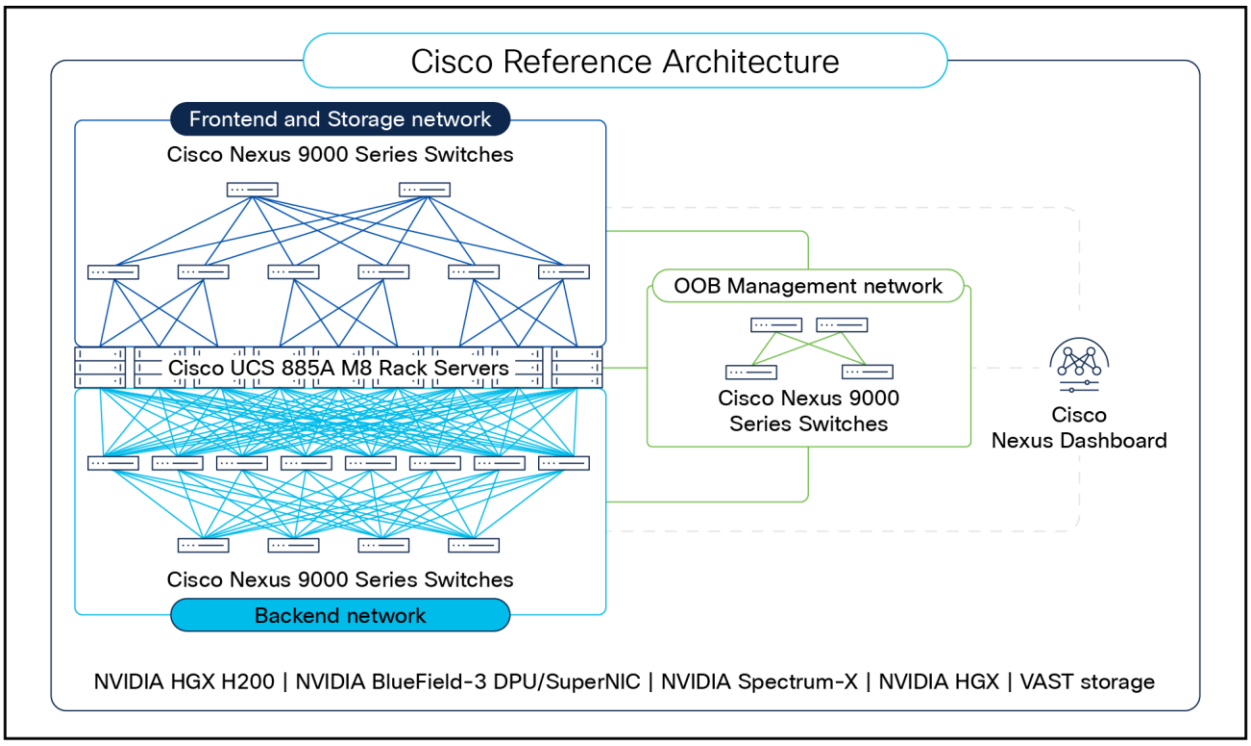


Figure 1.
Cisco Nexus 9000 Series Switches for networking the AI clusters, managed by Nexus Dashboard Platform

Hardware

Cisco UCS C885A M8 Rack Server

The Cisco UCS C885A M8 Rack Server is an 8RU, dense GPU server (node) that delivers massive, scalable performance for AI workloads such as Large Language Model (LLM) training, fine-tuning, inferencing, and Retrieval Augmented Generation (RAG). It is based on the NVIDIA HGX reference architecture in 2-8-10-400 (C-G-N-B) configuration where the C-G-N-B naming convention is defined as:

- C: number of CPUs in the node
- G: number of GPUs in the node
- N: number of network adapters (NICs), categorized into:
  - North-south: communication between nodes and external systems through the frontend network.
  - East-west: communication within the cluster through the backend network.
- B: average network bandwidth per GPU in gigabits per second (Gb/s)

The 8x NVIDIA H200 SXM GPUs within the server are interconnected using high-speed NVLink interconnects. GPU connectivity to other physical servers is performed using 8x NVIDIA BlueField-3 B3140H SuperNICs for East-West (E-W) traffic. For North-South (N-S) traffic, 1x NVIDIA BlueField-3 B3220 DPU NIC (in 2x200G mode) is available. Each server contains 2x AMD EPYC CPUs, up to 3 TB of DDR DRAM, 30 TB of NVMe local storage, and hot swappable fan trays and power supplies. Detailed specifications of the server are captured in Appendix A.

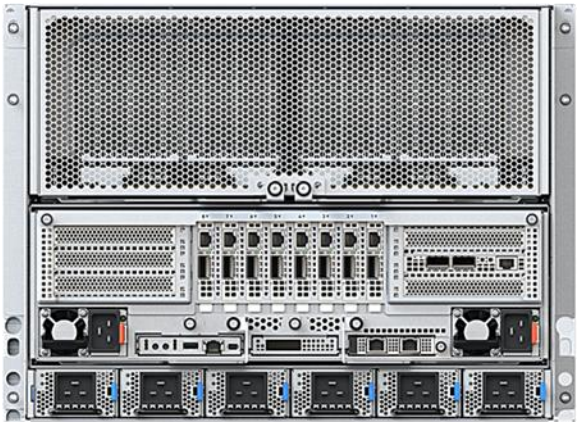


Figure 2.
Cisco UCS C885A M8 Rack Server with NVIDIA HGX

This ERA applies to any other server based on NVIDIA HGX in the 2-8-10-400 (C-G-N-B) configuration, although Cisco UCS C885A M8 Rack Server has been used for illustration.

Cisco Nexus 93108TC-FX3 switch

The Cisco Nexus 93108TC-FX3 switch (see Figure 3) provides 48 100-Mbps or 1/10-Gbps 10GBASE-T ports and six 1/10/25/40/100-Gbps QSFP28 ports in 1RU form-factor. This switch can be used in a management network.

Figure 3.
Cisco Nexus 93108TC-FX3 switch

Cisco Nexus 9332D-GX2B switch

The Cisco Nexus 9332D-GX2B switch (see Figure 4) provides 32 400G QSFP-DD ports with 10/25/50/100/200-Gbps breakout support in 1RU form-factor. This switch can be used in a leaf or spine role.

Figure 4.
Cisco Nexus 9332D-GX2B switch

Cisco Nexus 9364E-SG2 switch

The Cisco Nexus 9364E-SG2 switch (see Figure 5) provides 64 800G ports or 128 400G ports in 2RU form factor, available in both QSFP-DD and OSFP port types. This switch can be used in a leaf or spine role.

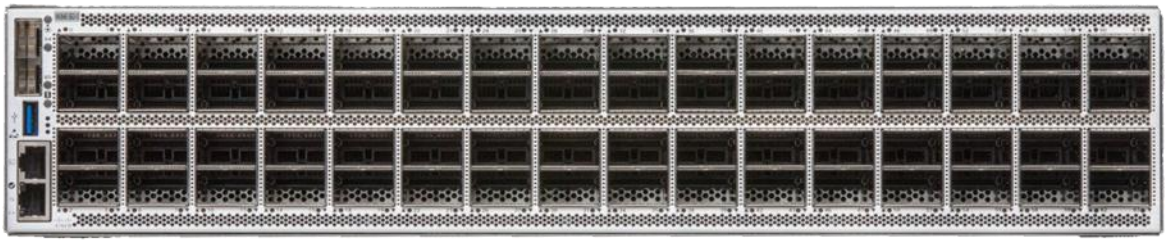


Figure 5.
Cisco Nexus 9364E-SG2 switch

Cisco UCS C225 M8 Rack Server

The Cisco C225 M8 Rack Server is a 1RU general-purpose server that can be used in many roles, such as application server, support server, control nodes for Kubernetes (K8s) and Slurm, or even for a storage platform. For example, these servers are also used to run the VAST storage solution as described in the "Storage architecture" section.

Figure 6.
Cisco UCS C225 M8 Rack Server

Networking topologies

Cisco Nexus 9000 Series Switches can be used for various types of data-center fabric designs. However, only the designs aligning closely to the NVIDIA ERA are considered part of this reference architecture.

Like the NVIDIA ERA, this Cisco ERA allows scale up to 128 NVIDIA eight-GPU HGX nodes for a total of 1024 H200 GPUs using a modular design of four-node Scalable Unit (SU), scalable up to 32 SUs to accommodate 128 nodes. Each SU is a discrete computation entity tied to the port availability size of the network devices. SUs can be replicated to adjust the scale of the environment with more ease.

As noted below, minor modifications are made to accommodate specific aspects of the Cisco design:

- Cisco UCS C885A M8 Rack Servers x86 management ports uses a speed of 10G instead of 1G.
- The VAST storage solution requires a minimum of 8x400G links to the storage network, although this requirement may change if a solution from a different NVIDIA-certified storage partner is used.
- The BMC ports of NVIDIA BlueField-3 SuperNICs are not connected, and they will be managed from an x86 host. However, the BMC ports of NVIDIA BlueField-3 DPUs are connected.

Topology for Cisco UCS C885A M8 Rack Servers with NVIDIA HGX

Figure 7 shows the cluster topology for up to 12 Cisco UCS C885A M8 Rack Servers with NVIDIA HGX, grouped in three SUs. At this small scale, a pair of Cisco Nexus 9364E-SG2 switches provide enough port density to converge frontend, backend, storage, and management network. This convergence increases affordability and simplicity while maintaining logical separation using VLANs.

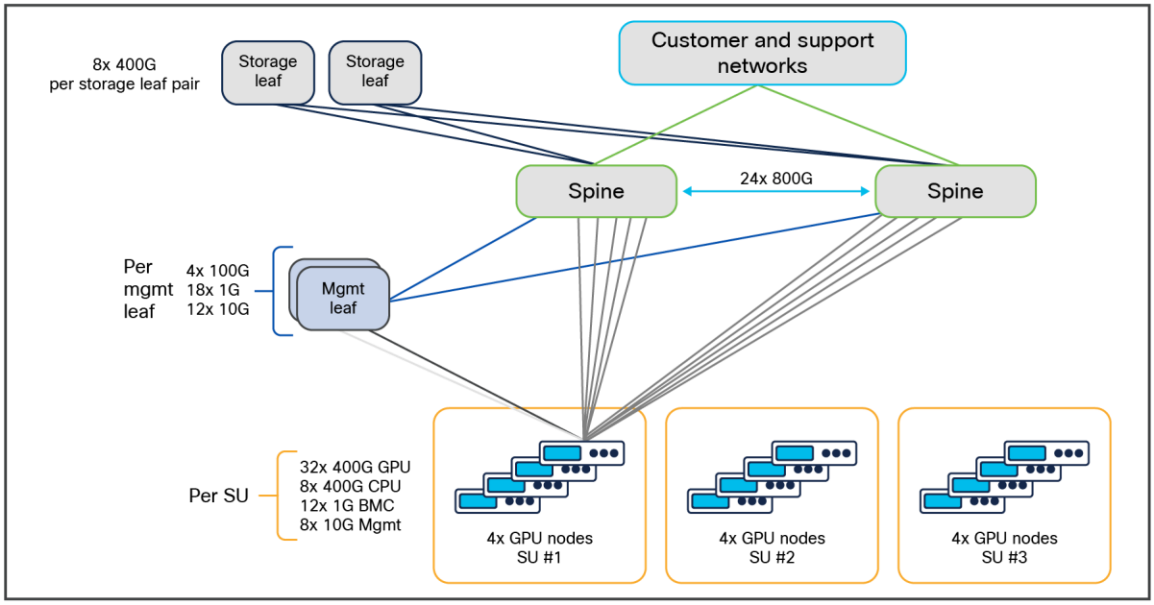


Figure 7.
Enterprise RA for 12 Cisco UCS C885A M8 Rack Servers with NVIDIA HGX (96 GPUs)

Table 1 shows the Bill of Materials (BOM) for a 12-node Cisco UCS C885A M8 Rack Server cluster with NVIDIA HGX.

Table 1. BOM for a 12-node Cisco UCS C885A Rack Server cluster with NVIDIA HGX (96 GPUs)

PID	Description	Quantity
UCSC-885A-M8-H12	Cisco UCS C885A M8 server with NVIDIA HGX	12
N9364E-SG2-O	Cisco Nexus switch, 64x800Gbps OSFP	2
N9K-93108TC-FX3	Cisco Nexus switch, 48 1/10G BASE-T 6 QSFP28	2
N9K-C9332D-GX2B	Cisco Nexus switch, 32x400Gbps QSFP-DD	2
OSFP-800G-DR8	OSFP, 800GBASE-DR8, SMF dual MPO-12 APC, 500m	114
QDD-400G-DR4-S	400G QSFP-DD transceiver, 400GBASE-DR4, MPO-12, 500m parallel	10
QSFP-400G-DR4	400G QSFP112 transceiver, 400GBASE-DR4, MPO-12, 500m parallel	118
QSFP-100G-DR-S	100GBASE DR QSFP transceiver, 500m over SMF	8
CB-M12-4LC-SMF	Cable, MPO12-4X duplex LC, breakout cable, SMF, various lengths	2
CB-M12-M12-SMF	MPO-12 cables	204
CAT6A	Copper cable for 10G	24
CAT5E	Copper cable for 1G	36

Nexus 9364E-SG2-O switches use OSFP-800G-DR8 twin-port transceivers (see Figure 8) with 2x400G MPO-12 connectors for 800G to 400G connections.

Figure 8.
Cisco OSFP-800G-DR8 transceiver module

Each connection independently supports 400G without the need for breakout cables (see Figure 9).

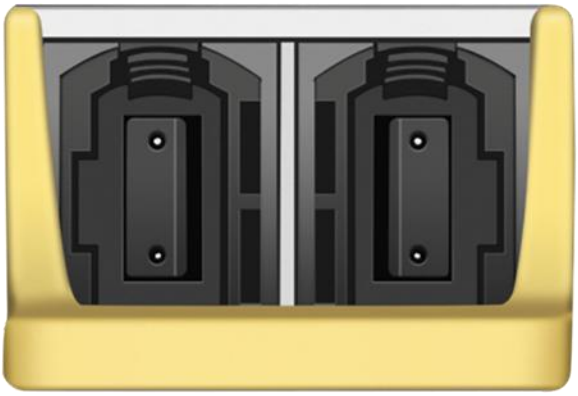


Figure 9.
Cisco OSFP-800G-DR8 plughole view

Figure 10 shows the cluster topology for up to 16 Cisco UCS C885A M8 Rack Servers with NVIDIA HGX. The E-W network is rail-aligned with Rails 1 to 4 falling on the left E-W spine and Rails 5 to 8 on the right E-W spine.

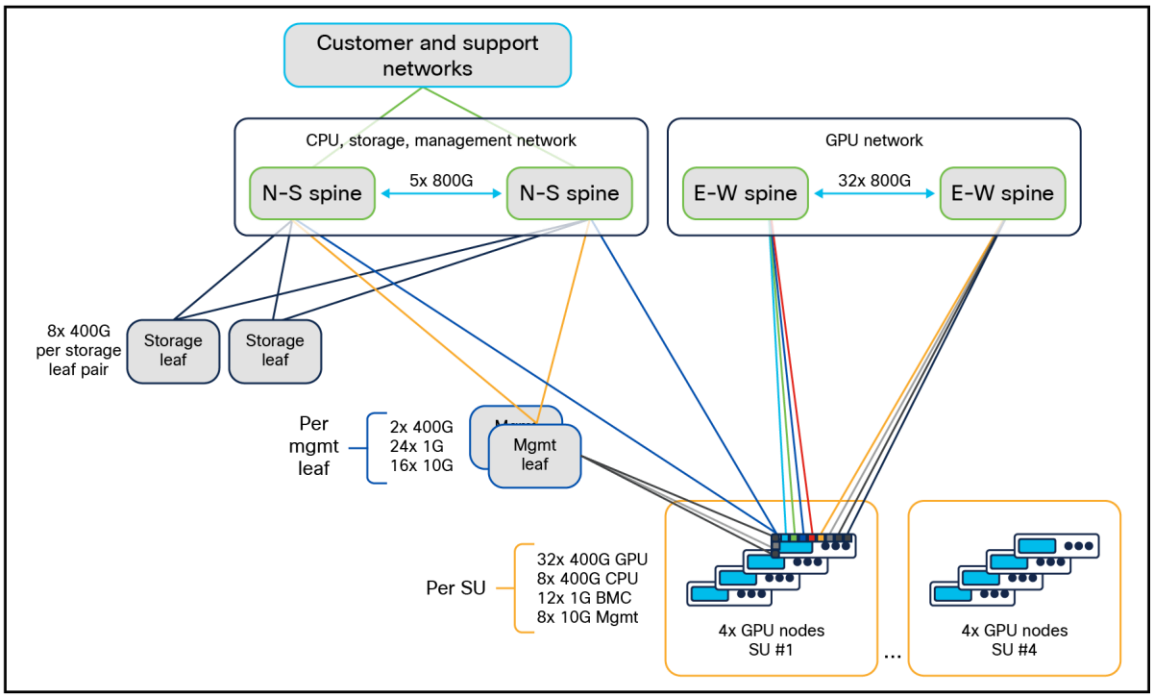


Figure 10.
Enterprise RA for 16 Cisco UCS C885A M8 Rack Servers with NVIDIA HGX (128 GPUs)

Table 2 shows the BOM for a 16-node Cisco UCS C885A M8 Rack Server cluster with NVIDIA HGX.

Table 2. BOM for a 16-node Cisco UCS C885A M8 Rack Server cluster with NVIDIA HGX (128 GPUs)

PID	Description	Quantity
UCSC-885A-M8-HC1	Cisco UCS C885A M8 Rack Server with NVIDIA HGX	16
N9364E-SG2-O	Cisco Nexus switch, 64x800Gbps OSFP	4
N9K-93108TC-FX3	Cisco Nexus switch, 48 1/10G BASE-T 6 QSFP28	2
N9K-C9332D-GX2B	Cisco Nexus switch, 32x400Gbps QSFP-DD	2
OSFP-800G-DR8	OSFP, 800GBASE-DR8, SMF dual MPO-12 APC, 500m	144
QDD-400G-DR4	400G QSFP-DD transceiver, 400GBASE-DR4, MPO-12, 500m parallel	12
QSFP-400G-DR4	400G QSFP112 transceiver, 400GBASE-DR4, MPO-12, 500m parallel	158
QSFP-100G-DR-S	100GBASE DR QSFP transceiver, 500m over SMF	8
CB-M12-4LC-SMF	Cable, MPO12-4X duplex LC, breakout cable, SMF, various lengths	2
CB-M12-M12-SMF	MPO-12 cables	198
CAT6A	Copper cable for 10G	32
CAT5E	Copper cable for 1G	48

For a cluster size greater than 16, the east-west compute network will expand into a spine-leaf fabric. For the largest cluster sizes, the north-south network will also be spine-leaf, as shown in Figure 11 of the 128 Cisco UCS C885A M8 Rack Server cluster. The E-W network is rail-aligned with each rail 1 to 8 falling on each E-W leaf 1 to 8.

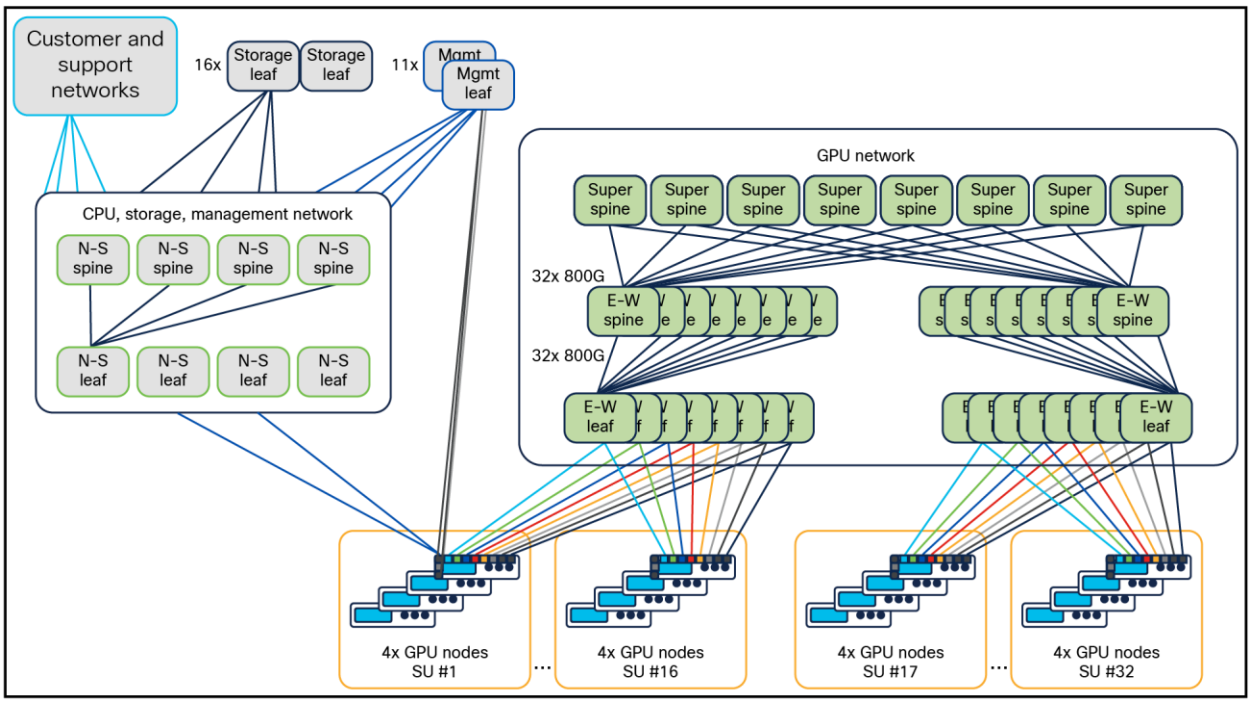


Figure 11.
Enterprise RA for 128 Cisco UCS C885A M8 Rack Servers with NVIDIA HGX (1024 GPUs)

Cluster fabric sizing tables

Sizing with Cisco UCS C885A M8 Rack Server

Tables 3 and 4 show the quantity of different units required for different cluster sizes using Cisco UCS C885A M8 Rack Server with NVIDIA HGX, eight E-W B3140H NVIDIA BlueField-3 SuperNICs, and two N-S B3240 NVIDIA BlueField-3 DPU NICs.

Table 3. East-west compute fabric table for Cisco UCS C885A M8 Rack Server with NVIDIA HGX –switch, transceivers, and cable counts

Compute counts		Switch counts			Transceiver counts			Cable counts	
Nodes	GPUs	Leaf	Spine	SuperSpine	Node to leaf		Switch to switch (800G)	Node to leaf	Switch to switch
					Node (400G)	Leaf (800G)			
12	96	2	N/A	N/A	96	48	48	96	48
16	128	2	N/A	N/A	128	64	64	128	64

Compute counts		Switch counts			Transceiver counts			Cable counts	
32	256	4	2	N/A	256	128	256	256	256
64	512	8	8	N/A	512	256	1024	512	1024
128	1024	16	16	8	1024	512	2048	1024	2048

Table 4. North-south fabric table for Cisco UCS C885A M8 Rack Server with NVIDIA HGX - switch, transceivers, and cable counts

Compute counts		Switch counts				Transceiver counts											
Nodes	GPUs	Leaf	Spine	Mgmt leaf	Storage leaf	Node to compute leaf		ISL ports	Node to mgmt leaf (1/10G)		Mgmt leaf to spine		Storage leaf to spine		Spine to customer and support			
						Node (400G)	Leaf (800G)	800G	Node	Leaf	Leaf (100G)	Spine (800G)	Leaf (400G)	Spine (800G)	Customer (800G)	Support (800G)	SMF MPO-12	CAT6A + CAT5E
12	96	Converged in east-west		2	2	24	12	N/A	N/A	N/A	8	2	8	4	8	4	60	60
16	128	2	N/A	2	2	32	16	10	N/A	N/A	8	2	8	4	8	4	78	80
32	256	2	N/A	4	4	64	32	16	N/A	N/A	16	4	16	8	16	4	144	160
64	512	2	N/A	7	8	128	64	30	N/A	N/A	28	7	32	16	32	4	274	320
128	1024	4	4	14	16	256	128	256	N/A	N/A	56	14	64	32	64	4	756	640

Optics and cables

In the first phase, the optics and cables listed in Table 5 are used in various devices in the system.

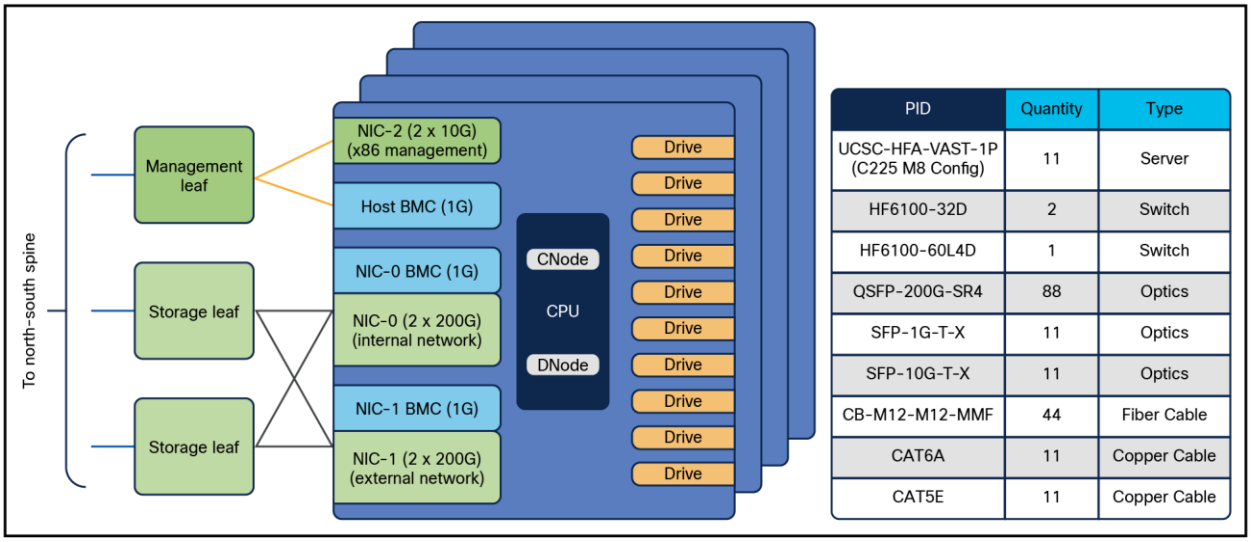
Table 5. Supported list of optics and cables in various devices

Device	Optics and cable
B3140H, B3240	QSFP-400G-DR4 with SMF MPO-12 cable
B3220	QSFP-200G-SR4 with MMF MPO-12 cable
N9364E-SG2-O	OSFP-800G-DR8 with dual SMF MPO-12 cable
N9K-C9332D-GX2B	QDD-400G-DR4 with SMF MPO-12 cable QSFP-200G-SR4 with MMF MPO-12 cable
N9K-93108TC-FX3	QSFP-100G-DR-S with SMF duplex LC cable CAT5E cable CAT6A cable

Storage architecture

Cisco has partnered with VAST Data to onboard their storage software on Cisco UCS C225 M8 Rack Servers in EBOX architecture. VAST Data supports a “Distributed And Shared Everything” (DASE) architecture that allows for horizontally scaling storage capacity and read/write performance by incrementally adding servers. To support all stages of an AI data pipeline, all protocol servers, such as NFS, S3, and SMB, are enabled.

Figure 12 shows the overall network connectivity of storage servers and the BOM for a single EBOX with two storage leafs. For Data Path, each server uses two NVIDIA BlueField-3 B3220L 2x200G NICs: NIC0 is used for internal network within the servers, allowing any server to access storage drives from any other server, and NIC1 is used for the external network, supporting client traffic such as NFS, S3, and SMB. Note: The internal network traffic is switched locally at the leaf (it never goes to the spine) because every server connects to every leaf. For client-facing external traffic, per EBOX, the minimum requirement over the spine is 11x 200G or 6x 400G. The 1G BMC and 10G x86 management ports are connected to a management leaf switch.



PID	Quantity	Type
UCSC-HFA-VAST-1P (C225 M8 Config)	11	Server
HF6100-32D	2	Switch
HF6100-60L4D	1	Switch
QSFP-200G-SR4	88	Optics
SFP-1G-T-X	11	Optics
SFP-10G-T-X	11	Optics
CB-M12-M12-MMF	44	Fiber Cable
CAT6A	11	Copper Cable
CAT5E	11	Copper Cable

Figure 12.
Block diagram and BOM of storage subsystem

As the cluster size increases, the number of storage leaf switches and EBOX will linearly increase, as captured in the cluster-sizing tables.

Besides VAST Data, other NVIDIA-certified storage partners can also be used in this reference architecture.

Software

Cisco NX-OS and Nexus Dashboard

As mentioned earlier, Cisco NX-OS is the operating system running on Nexus 9000 Series Switches, whereas Nexus Dashboard is the operations and automation platform for managing the fabrics.

Cisco NX-OS and Nexus Dashboard will not:

- Configure compute or storage in any way
- Manage the BMC or host CPU software lifecycle of servers
- Manage the kernel and distribution on the NVIDIA BlueField-3 NICs.

Server configuration and management functions must be managed through some other means (Cisco Intersight® is an option). The customer will be solely responsible for deploying and using these tools. In addition to being the appropriate scope for a network controller, this separation of concerns aligns with the dominant operational paradigm that segments network operations from compute and storage.

NVIDIA AI Enterprise

This reference architecture includes NVIDIA AI Enterprise, deployed and supported on NVIDIA-certified Cisco UCS C885A M8 Rack Servers with NVIDIA HGX NVIDIA AI Enterprise is a cloud-native software platform that streamlines development and deployment of production-grade AI solutions, including AI agents, generative AI, computer vision, speech AI, and more. Enterprise-grade security, support, and API stability ensure a smooth transition from prototype to production.

NVIDIA NIM™ microservices provide a complete inference stack for production deployment of open-source community models, custom models, and NVIDIA AI Foundation models. Their scalable, optimized inference engine and ease of use accelerate models, improve TCO, and make production deployment faster.

Compute server stack

The entire cluster solution has been verified with compute nodes running Ubuntu Linux 22.04 LTS and NVIDIA Cloud Native Stack (CNS) version 12.3, which includes the compatible drivers, GPU and network operators within Kubernetes (K8s) environment. Slurm version 24.11.1 has been verified as a workload orchestration engine. Containers under NVIDIA NGC™ catalog can be launched both with Kubernetes and with Slurm.

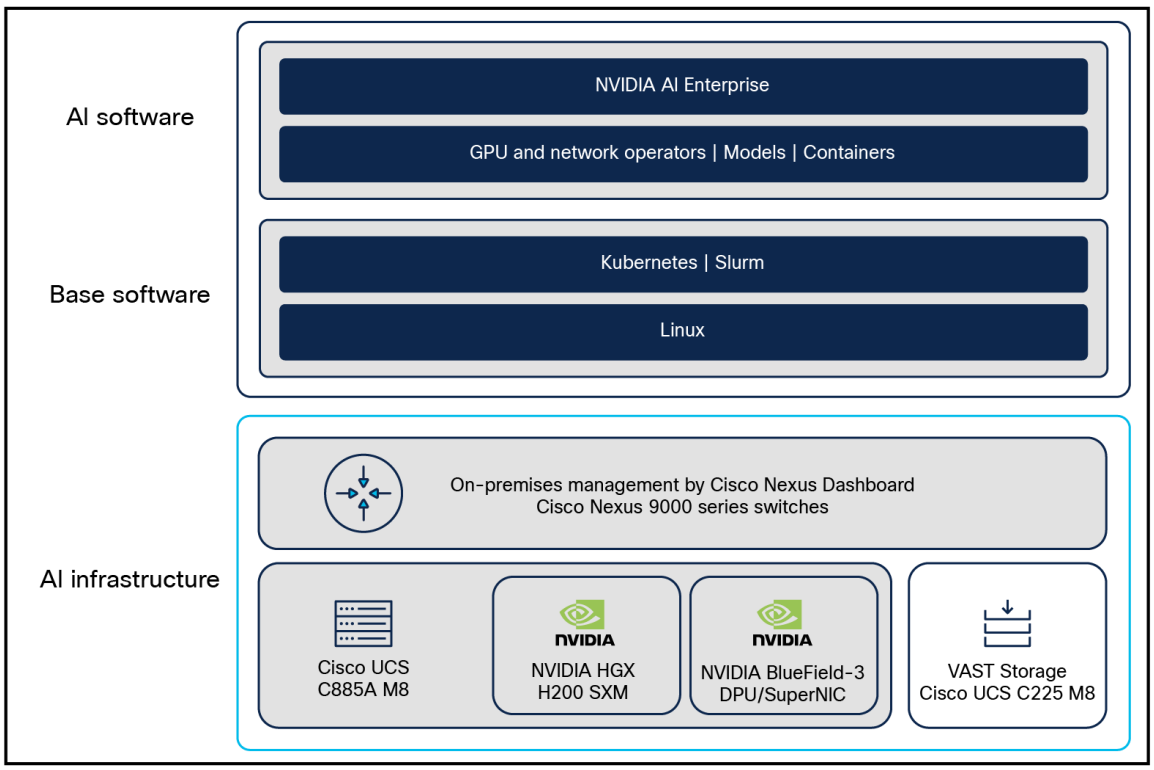


Figure 13.
Compute server software stack

Customers can run their choice of OS distribution and software versions as per NVIDIA AI Enterprise, drivers, and the CNS compatibility matrix published by NVIDIA.

Security

Cisco network security and observability services can be optionally integrated in different hardware (switches, hosts, NICs) and software components of the cluster.

Testing and certification

The overall solution has been thoroughly tested considering all aspects of management plane, control plane, and data plane combining together compute, storage, and networking. A number of benchmark test suites such as HPC Benchmark, IB PerfTest, NCCL Test, MLCommons Training, and Inference benchmarks have also been run to evaluate performance and assist with tuning. Different elements and entities of the NVIDIA AI Enterprise ecosystem have been brought up and tested to evaluate a number of enterprise-centric customer use cases around fine-tuning, inferencing, and RAG. Results from running NVIDIA-Certified Systems Test Suite version 3.5 for both single-node and multi-node with networking have passed for Cisco UCS C885A M8 Rack Servers.

Summary

Cisco Nexus 9000 Series Switches and the Nexus Dashboard platform provide scalable, easy-to-manage, and high-performance networking for AI Infrastructure powered by NVIDIA-accelerated computing .

Appendix A – Compute server specifications

Table 6 provides the specification of the Part# UCSC-885A-M8-H12 for Cisco UCS C885A M8 Rack Server used as GPU node in this ERA.

Table 6. Cisco UCS C885A M8 8RU rack server

Area	Details
Form factor	8RU rack server (air-cooled)
Compute + memory	2x 5th Gen AMD EPYC 9575F (400W, 64 core, up to 5GHz) 24x 96GB DDR5 RDIMMs, up to 6000 MT/S (Recommended Memory config) 24x 128GB DDR5 RDIMMs, up to 6000 MT/S (Max Supported Memory config)
Storage	Dual 1 TB M.2 NVMe with RAID support (boot device) Up to 16 PCIe5 x4 2.5” U.2 1.92 TB NVMe SSD (data cache)
GPUs	8x NVIDIA HGX™ H200 (700W each)
Network cards	8 PCIe x16 HHHL NVIDIA BlueField-3 B3140H east-west NIC 1 PCIe x16 FHHL NVIDIA BlueField-3 B3220 north-south NIC 1 OCP 3.0 X710-T2L for host management
Cooling	16 Hot swappable (N+1) fans for system cooling
Front IO	2 USB 2.0, 1 ID Button, 1 power button
Rear IO	1 USB 3.0 A, 1 USB 3.0 C, mDP, 1 ID button, 1 power button, 1 USB 2.0 C, 1 RJ45
Power supply	6x 54V 3kW MCRPS (4+2 redundancy) and 2x 12V 2.7kW CRPS (1+1 redundancy)

Appendix B – Control-node server specifications

The versatile Cisco UCS C225 M8 1RU rack server can be used as a support server (also called a control-node server) for Slurm and Kubernetes (K8s), etc. Table 7 shows the minimum specifications of the server.

Table 7. Cisco UCS C225 M8 1RU rack server

Area	Details
Form factor	1RU rack server (air-cooled)
Compute + memory	1x 4th Gen AMD EPYC 9454P (48-cores) 12x 32GB DDR5 RDIMMs 4800MT/s
Storage	Dual 1 TB M.2 SATA SSD with RAID (boot device) Up to 10x 2.5-inch PCIe Gen4 x4 NVMe PCIe SSDs (each with capacity 1.9 to 15.3 TB) – Optional
Network cards	1 PCIe x16 FHHL NVIDIA BlueField-3 B3220L configured in DPU mode Or 1 PCIe x16 FHHL NVIDIA BlueField®-3 B3140H configured in DPU mode 1 OCP 3.0 X710-T2L (2 x 10G RJ45) for x86 host management
Cooling	8 Hot swappable (N+1) fans for system cooling
Power supply	2x 1.2KW MCRPs PSU with N+1 redundancy
BMC	1G RJ45 for host management

Deployments looking for 2-socket CPUs can use the Cisco UCS C245 M8 2RU rack server variant along with B3220 DPU NICs.

References

- NVIDIA AI Enterprise Software Reference Architecture: https://docs.nvidia.com/ai-enterprise/reference-architecture/latest/index.html
- NVIDIA HGX: https://www.nvidia.com/en-us/data-center/hgx
- NVIDIA Spectrum-X networking: https://www.nvidia.com/en-us/networking/spectrumx/
- Cisco Nexus 9000 Series Switches: https://www.cisco.com/site/us/en/products/networking/cloud-networking-switches/nexus-9000-switches/index.html
- Cisco Nexus Dashboard: https://www.cisco.com/site/us/en/products/networking/cloud-networking/nexus-platform/index.html

Document history

New or revised topic	Described in	Date
Topology for Cisco UCS C885A M8 Rack Servers with NVIDIA HGX	Networking Topologies	June 2025

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 C78-5054501-01 07/25