Data sheet
Cisco public

NVIDIA Certified Cisco Nexus Hyperfabric AI Enterprise Reference Architecture

Contents

Cisco Nexus Hyperfabric AI Enterprise Reference Architechure certified by NVIDIA, featuring Cisco® cloud-managed AI/ML networking of Cisco UCS® C885A M8 Rack Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X

Introduction

Cisco Nexus® Hyperfabric AI is an on-premises AI cluster that is managed by a cloud-hosted controller. It empowers and simplifies your AI initiatives and accelerates AI deployments with a comprehensive, integrated, cloud-managed solution. Cisco Nexus Hyperfabric AI Reference Architecture is based on Cisco Silicon One® switches and adheres to the NVIDIA Enterprise Reference Architecture (Enterprise RA) for NVIDIA HGX H200 and Spectrum-X.

Figure 1 shows the key components of the solution. The key hardware components used in the cluster are described in the next section.

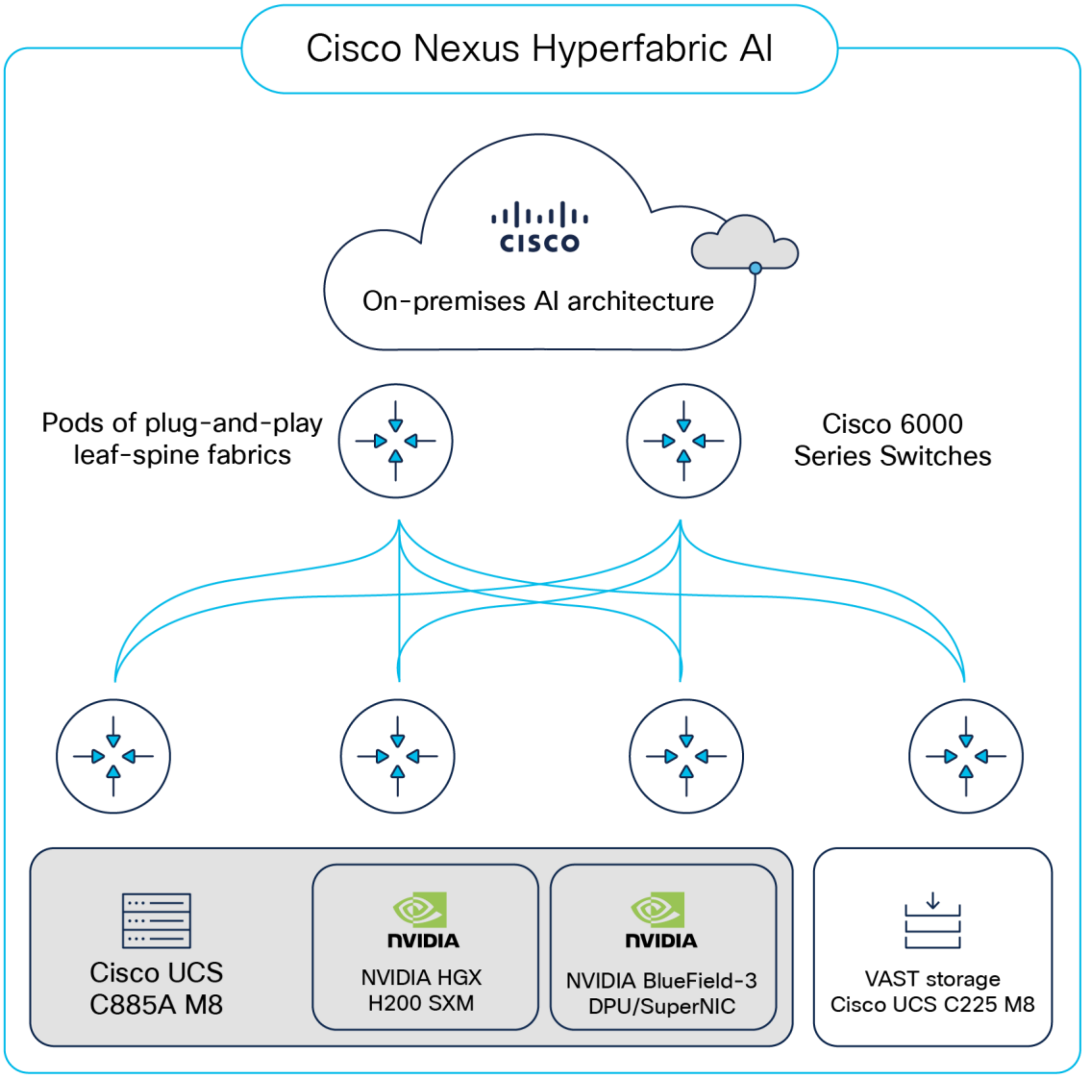

Cisco Nexus Hyperfabric AI
cisco
On-premises AI architecture
Pods of plug-and-play leaf-spine fabrics
Cisco 6000 Series Switches
Cisco UCS C885A M8
NVIDIA
NVIDIA HGX H200 SXM
NVIDIA
NVIDIA BlueField-3 DPU/SuperNIC
VAST storage Cisco UCS C225 M8

Figure 1.
Key components of Cisco Nexus Hyperfabric AI

Hardware

Cisco UCS C885A M8 Rack Server

The Cisco UCS C885A M8 Rack Server is an 8RU, dense GPU server that delivers massive, scalable performance for AI workloads such as large language model (LLM) training, fine-tuning, large model 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 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
  - East-west: communication within the cluster
- B: average network bandwidth per GPU in Gigabits per second (GbE)

The 8x NVIDIA H200 SXM GPUs within the server are interconnected using high-speed NVLink interconnects. GPU connectivity to other physical servers is through the use of 8x NVIDIA BlueField-3 B3140H SuperNICs for east-west traffic and through 2x NVIDIA BlueField-3 B3240 DPU NICs (in 1x400G mode) for north/south traffic. For compute, 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.

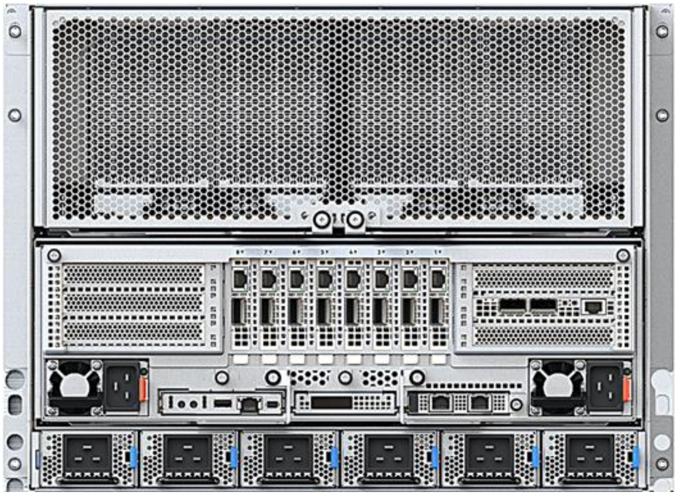


Figure 2.
Cisco UCS C885A M8 Rack Server with NVIDIA HGX

Cisco HF6100-60L4D

The Cisco HF6100-60L4D is a 1RU Silicon One NPU-based high-density switch supporting 60 SFP56 ports capable of 1/10/25/50GE speeds, plus 4 ports of 400 QSFPDD with breakout support. This switch is used in many different roles, such as management network, connectivity to applications and support servers, etc.

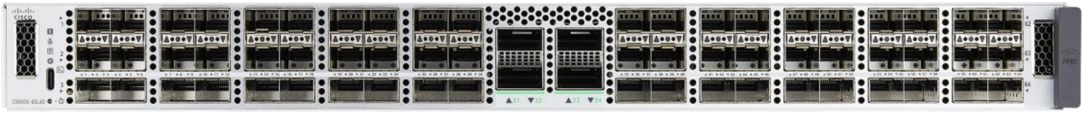


Figure 3.
Cisco HF6100-60L4D switch

Cisco HF6100-32D

The Cisco HF6100-32D is a 1RU Silicon One NPU-based high-density 400G-port-capable switch supporting 32 ports of QSFPDD with breakout support. This switch can be used in the leaf or spine role depending on the requirements of the cluster.

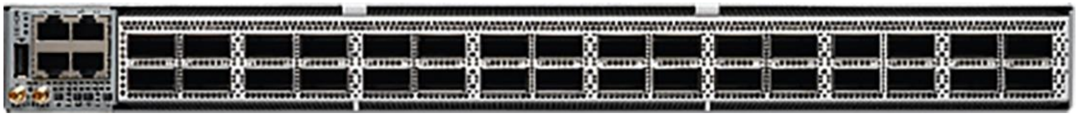


Figure 4.
Cisco HF6100-32D switch

Cisco HF6100-64E

The Cisco HF6100-64E is a 2RU Silicon One NPU-based high-density 800G-port-capable switch supporting 64 ports of OSFP 800G and allowing 64 800GE or 128 400G GE ports. It can be used in both leaf and spine roles.

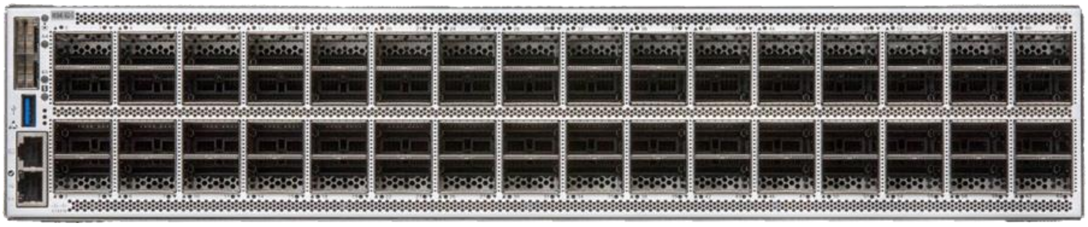


Figure 5.
Cisco HF6100-64E switch

Cisco UCS C225 M8 Rack Server

The Cisco UCS 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, etc. In Cisco Nexus Hyperfabric AI, these servers are also used to run the VAST storage solution as described in the “Storage Architecture” section, below.

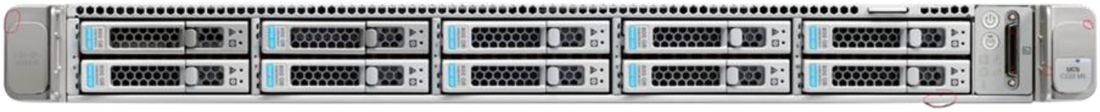


Figure 6.
Cisco UCS C225 M8 Rack Server

Cisco Optics and cables

The following Cisco optics and cables shown in Table 1 are being used on the listed devices.

Table 1. Supported list of optics and cables on different devices

Device	Optics and cables
B3140H, B3240	QSFP-400G-DR4 with CB-M12-M12-SMF cable
B3220	QSFP-200G-SR4 with CB-M12-M12-MMF cable
HF6100-64E	OSFP-800G-DR8 with dual CB-M12-M12-SMF cable
HF6100-32D	QDD-400G-DR4 with CB-M12-M12-SMF cable QSFP-200G-SR4 with CB-M12-M12-MMF cable
HF6100-60L4D	QDD-400G-DR4 with CB-M12-M12-SMF cable SFP-1G-T-X for 1G with CAT5E cable SFP-10G-T-X for 10G with CAT6A cable

Networking topologies

Hyperfabric is a flexible, multipurpose fabric design tool that can be used to specify any type of fabric design. To facilitate AI cluster deployments, Hyperfabric for AI will come with preconfigured templates for different "T-shirt" sizes of AI clusters. Customers will have the choice of using a template "as-is" (with no choices or customizations outside of the length of fiber and associated optics needed to connect devices) or designing their own custom network. Only the templated designs will be considered part of the reference architecture aligning closely to NVIDIA Enterprise Reference Architecture (ERA). As noted below, minor modifications will be made to accommodate specific aspects of the Cisco® design:

- Cisco UCS C885A M8 Rack Servers with NVIDIA HGX contain 2x400G frontend ports instead of NVIDIA's 2x200G ports. Also, the x86 management ports will use a speed of 10G instead of 1G.
- The VAST storage solution requires a minimum of 8x400G to the storage network.
- The BMC ports of NVIDIA BlueField-3 SuperNICs will not be connected, and they will be managed from x86 host. However, the BMC ports of NVIDIA BlueField-3 DPUs will be connected.

Topology for Cisco UCS C885A M8 Rack Server with NVIDIA HGX

The diagram in Figure 7 shows the cluster topology for up to 12 Cisco UCS C885A M8 Rack Servers with NVIDIA HGX.

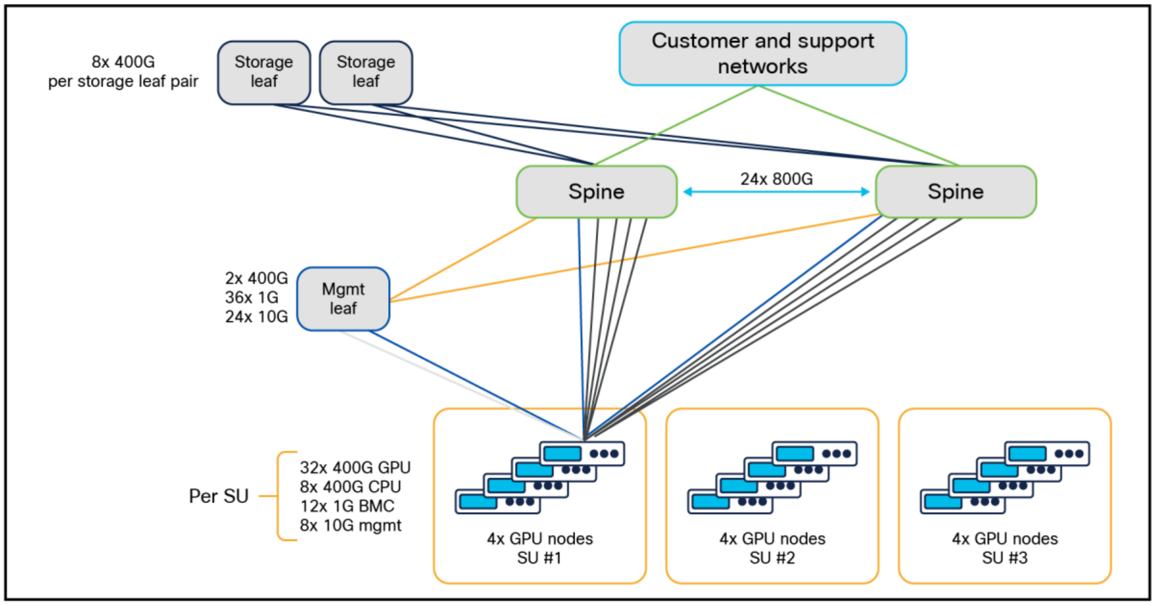


Figure 7.
Enterprise RA for 12 Cisco UCS C885A M8 Rack Servers with NVIDIA HGX (96 GPUs)

The BOM for a 12-node Cisco UCS C885A M8 Rack Server cluster with NVIDIA HGX server is shown in Table 2.

Table 2. BOM for a 12-node Cisco UCS C885A M8 Rack Server cluster with NVIDIA HGX (96 GPUs)

PID	Description	Quantity
UCSC-885A-M8-HC1	Cisco UCS C885A M8 Rack Server with NVIDIA HGX	12
HF6100-64ED	Cisco Hyperfabric switch, 64x800Gbps OSFP	2
HF6100-60L4D	Cisco Hyperfabric switch 60x50G SFP28 4x400G QSFP-DD	1
HF6100-32D	Cisco Hyperfabric 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	120
SFP-1G-T-X	1G SFP	36
SFP-10G-T-X	10G SFP	24
CB-M12-M12-SMF	MPO-12 cables	204
CAT6A	Copper cable for 10G	24
CAT5E	Copper cable for 1G	36

The 800G to 400G connections will use optics with dual 2x400G MPO-12 connectors on the switch side, as shown below.

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

Each connection independently supports 400G without the need for breakout cables.

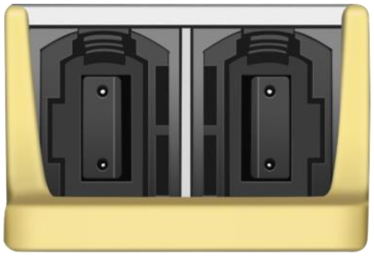


Figure 9.
Cisco OSFP-800G-DR8 plughole view

The diagram in Figure 10 shows the cluster topology for up to 16 Cisco UCS C885A M8 Rack Servers with NVIDIA HGX. The east-west (E-W) network is rail-aligned with 4 rails on the left east/west spine and 4 rails on the right east-west spine.

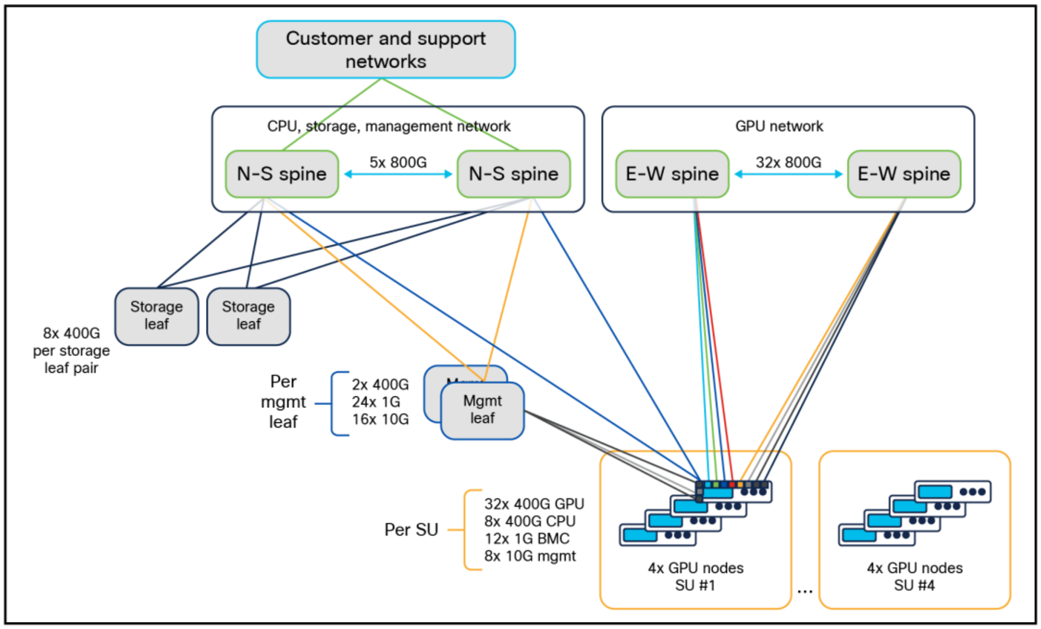


Figure 10.
Enterprise RA for 16 Cisco UCS C885A M8 Rack Servers with NVIDIA HGX (128 GPUs)

The BOM for a 16-node Cisco UCS C885A M8 Rack Server cluster with NVIDIA HGX servers is shown in Table 3.

Table 3. 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
HF6100-64ED	Cisco Hyperfabric switch, 64x800Gbps OSFP	4
HF6100-60L4D	Cisco Hyperfabric switch 60x50G SFP28 4x400G QSFP-DD	2
HF6100-32D	Cisco Hyperfabric 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	160
SFP-1G-T-X	1G SFP	48
SFP-10G-T-X	10G SFP	32
CB-M12-M12-SMF	MPO-12 cables	198
CAT6A	Copper cable for 10G	32
CAT5E	Copper cable for 1G	48

For cluster size greater than 16, the east-west (E-W) compute network will expand into a spine-leaf fabric. For the largest cluster sizes, the north/south (N-S) network will also be spine-leaf, as shown in Figure 11 showing 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.

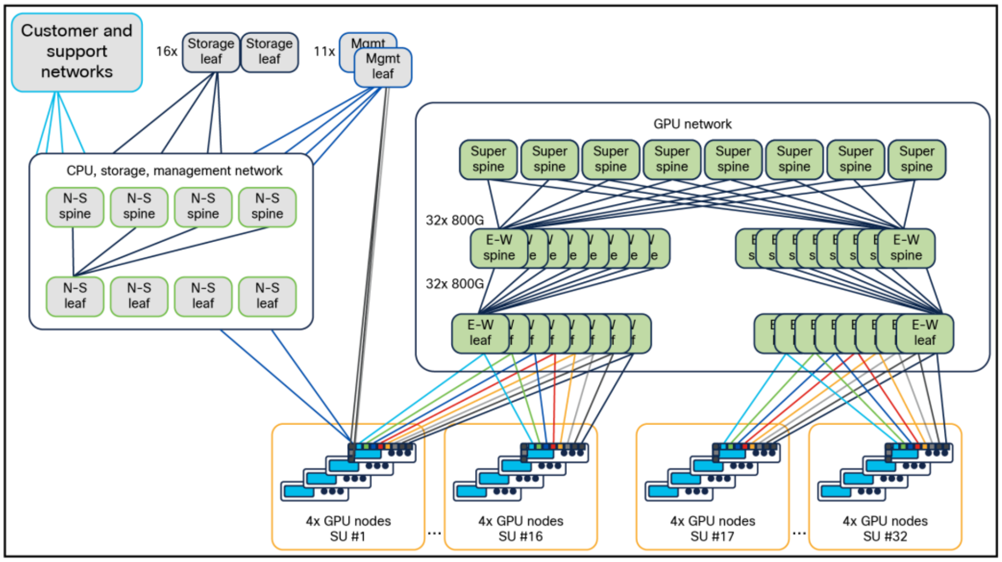


Figure 11.
Enterprise RA for 128 Cisco UCS C885A M8 Rack Servers with NVIDIA HGX (1024 GPUs)

Cluster fabric sizing tables

Sizing with Cisco UCS C885A M8 Rack Server

Tables 4 and 5 show the quantity of different units required for different cluster sizes using Cisco UCS C885A M8 Rack Server system with NVIDIA HGX, 8 E-W B3140H NVIDIA BlueField-3 SuperNICs and 2 N-S B3240 NVIDIA BlueField-3 DPU NICs.

Table 4. East-west compute fabric table for Cisco UCS C885A M8 Rack Server with NVIDIA HGX – compute, switch, transceivers, 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
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

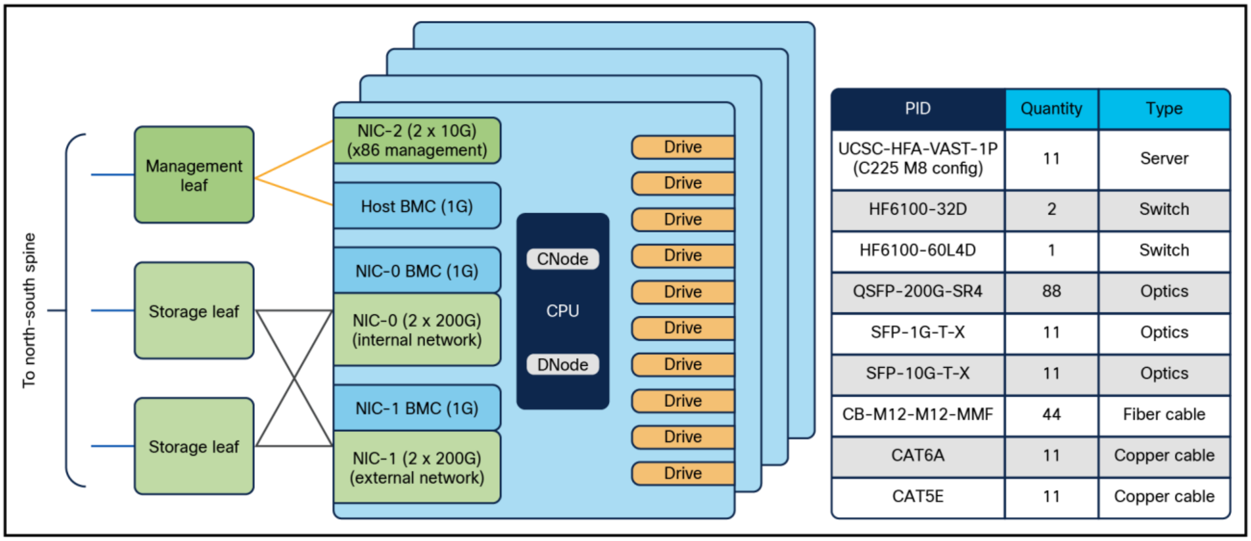
Table 5. North-south fabric table for Cisco UCS C885A M8 Rack Server with NVIDIA HGX – compute, switch, transceivers, cable counts

Compute counts		Switch counts				Transceiver counts											Cable 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 (400G)	Spine (800G)	Leaf (400G)	Spine (800G)	Customer (800G)	Support (800G)	SMF MPO-12	CAT6A + CAT5E
12	96	Converged in East-west		1	2	24	12	N/A	N/A	60	2	2	8	4	8	4	60	60
16	128	2	N/A	2	2	32	16	10	N/A	80	4	2	8	4	8	4	78	80
32	256	2	N/A	3	4	64	32	16	N/A	160	6	4	16	8	16	4	144	160
64	512	2	N/A	6	8	128	64	30	N/A	320	12	6	32	16	32	4	274	320
128	1024	4	4	11	16	256	128	256	N/A	640	44	22	64	32	64	4	756	640

Storage architecture

Cisco has partnered with VAST Data to onboard their storage software on Cisco UCS C225 M8 Rack Servers in EBOX architecture and together provide the storage subsystem for Cisco Nexus Hyperfabric AI cluster. 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 the 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 BOM for a single EBOX with two storage leafs. For Data Path, each server uses 2 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 external network supporting client traffic such as NFS, S3, and SMB. Note that the internal network traffic is switched locally at the leaf (it never goes to spine) because every server connects to every leaf. For client-facing external traffic, per EBOX, the minimum requirement over spine is 11 x 200G or 6 x 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 sub-system

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

Software

Hyperfabric controller

Hyperfabric is a cloud-hosted, multitenant controller whose primary function is to control the network fabric. It fully manages the configuration target state, software versions, and so on for the Hyperfabric switches.

Network controllers benefit from the additional visibility of the network behavior observed by devices connected to them. For the Cisco UCS-based compute and storage servers, Hyperfabric will provide IPM telemetry and minimally manage some port-level options, such as packet reordering, buffer adjustments, and IPG tuning. Otherwise, servers and NICs are visibility-only devices.

To that end, Hyperfabric 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 to 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. 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 accelerates models, improves TCO, and makes 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, GPUs, 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 with both Kubernetes and Slurm.

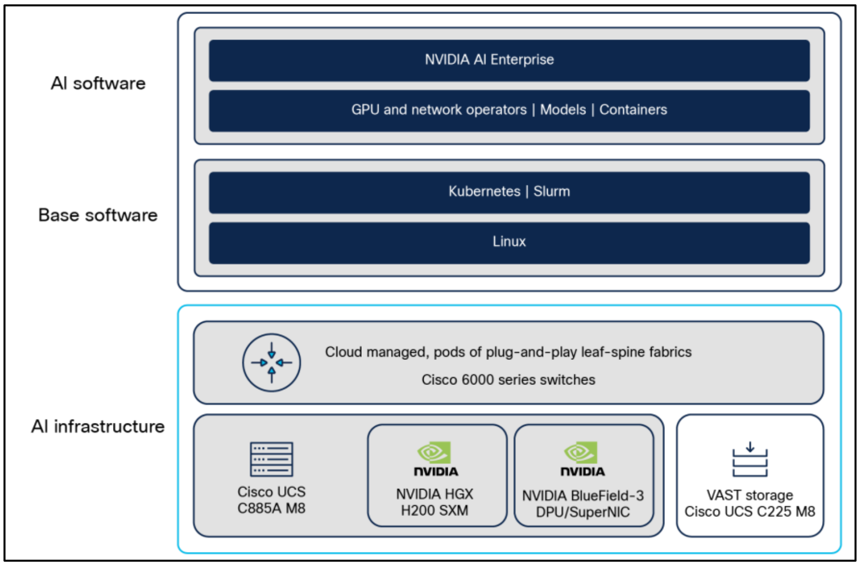

AI software
NVIDIA AI Enterprise
GPU and network operators | Models | Containers
Base software
Kubernetes | Slurm
Linux
AI infrastructure
Cloud managed, pods of plug-and-play leaf-spine fabrics
Cisco 6000 series switches
Cisco UCS C885A M8
NVIDIA
NVIDIA HGX H200 SXM
NVIDIA
NVIDIA BlueField-3 DPU/SuperNIC
VAST storage Cisco UCS C225 M8

Figure 13.
Compute-server software stack

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

Security

Beyond phase one of the solution, additional Cisco network security and observability services will be integrated in different hardware (switches, hosts, NICs) and software components of the cluster.

Testing and certification

The overall solution has been thoroughly tested on all aspects of management plane, control plane, and dataplane, combining 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 the Cisco UCS C885A M8 Rack Server.

Summary

In short, Cisco Nexus Hyperfabric AI is a fully integrated, end-to-end tested AI cluster solution, offering customers a one-stop shopping place for their AI infrastructure deployment needs.

Appendix A – Compute server specifications

Cisco UCS C885A M8 Rack Server

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 6,000 MT/S (Recommended Memory config) 24x 128GB DDR5 RDIMMs, up to 6,000 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 H200 GPUs (700W each)
Network cards	8 PCIe x16 HHHL NVIDIA BlueField-3 B3140H East-West NIC 2 PCIe x16 FHHL NVIDIA BlueField-3 B3240 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 will be used as a support server and 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.

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 C78-5049901-00 06/25