

Cisco 8000 Cloud Reference Architecture

Featuring Networking Reference Architecture of Cisco UCS C885A M8 Rack Servers with NVIDIA HGX™ H200 and NVIDIA Spectrum™-X



Contents

Introduction	3
Hardware	4
Cisco Optics and Cables	6
Networking Topologies	6
Cluster BOM.....	18
Multitenancy.....	19
Edge and Border connectivity	20
Storage architecture.....	20
Software	21
Security.....	22
Observability	22
Testing and certification	22
Summary.....	23
Appendix A – Compute server specifications.....	23
Appendix B – Management node server specifications.....	24

Introduction

The Cisco Cloud Reference Architecture (CRA) is designed to be deployed with a high GPU scale, ranging from 1K to 32K GPUs, at large Cloud Service Providers (CSPs) and high-performance Super Computing centers (SCCs) in order to solve the most computationally intensive problems without affecting ease of provisioning and operations. The overall design supports multitenancy in order to maximize the use of deployed hardware and, if required, can be scaled to 64K GPUs. Enterprises looking to deploy AI clusters with GPU scale less than 1K should refer to Cisco Enterprise Reference Architecture (ERA) at this [link](#).

The key technologies used in this CRA include:

- Cisco UCS® C885A Rack Servers with NVIDIA HGX™ H200 and Spectrum™-X Ethernet.
- Cisco® Silicon One® NPU-based 8000 Series SONiC Switches.
- Cisco Optics and cables.
- Cisco provisioning, observability and security frameworks.

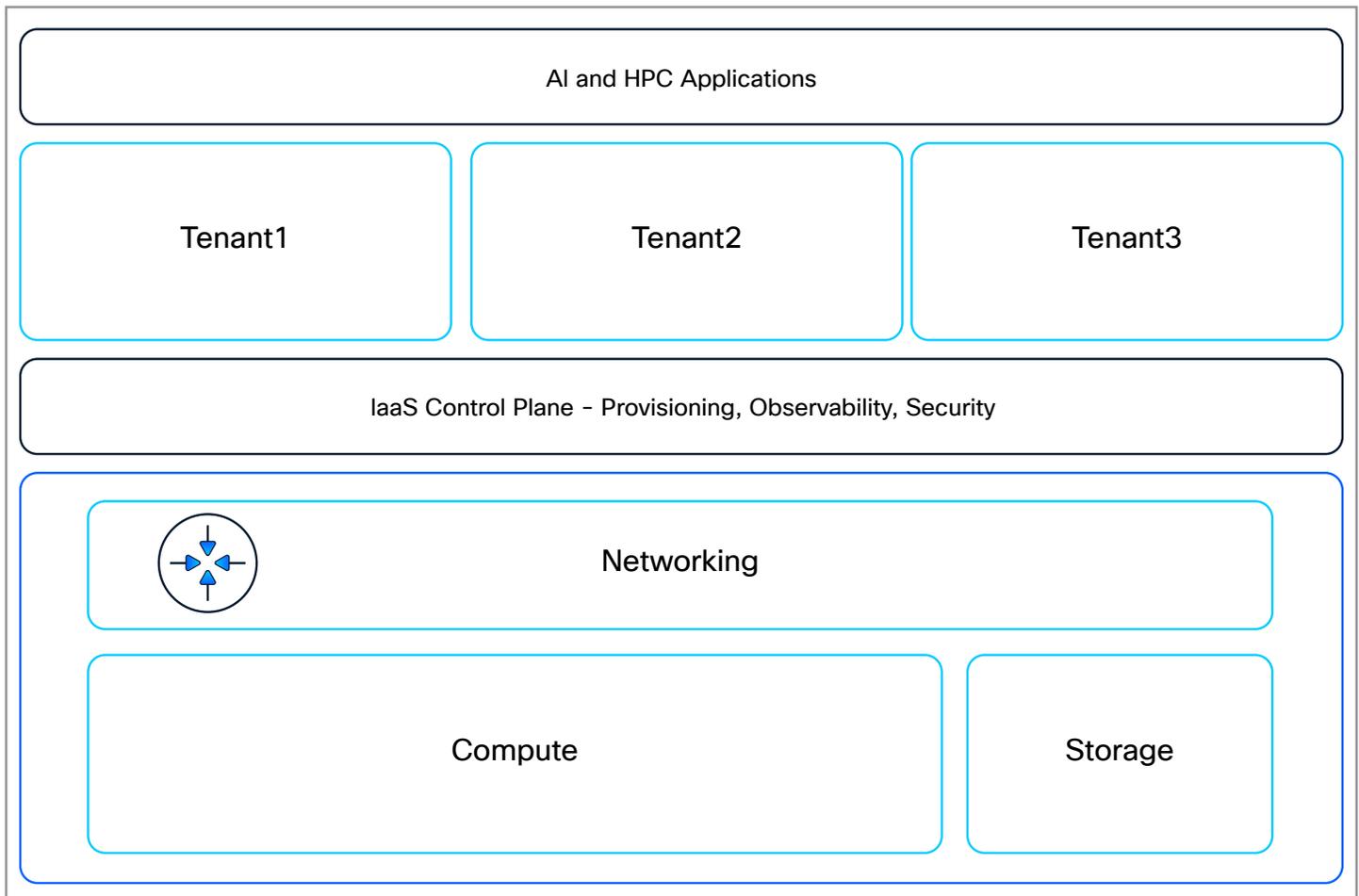


Figure 1. Logical View of Cisco Cloud Reference Architecture

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 HPC and AI workloads such as Large Language Model (LLM) training, fine-tuning, large model inferencing, and Retrieval Augmented Generation (RAG). 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 via the use of 8x NVIDIA BlueField®-3 B3140H SuperNICs for East-West traffic and via 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.

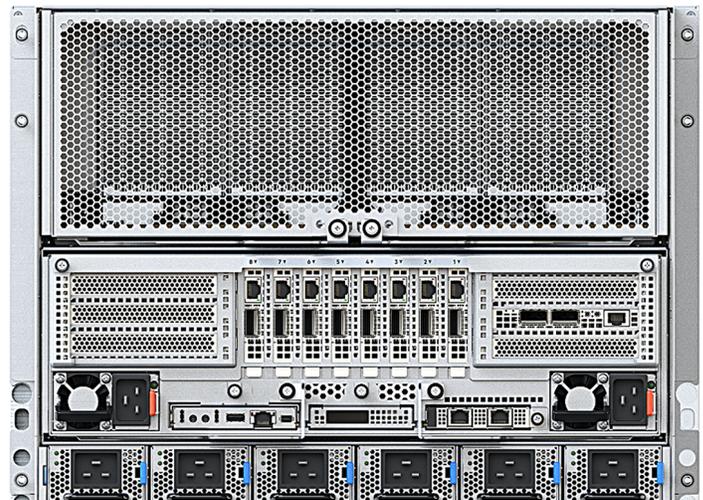


Figure 2. Cisco UCS C885A M8 Rack Server with NVIDIA HGX™

Cisco 8122-64EHF-O

The Cisco 8122-64EHF-O 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 400GE ports. It will be used in both leaf and spine roles.

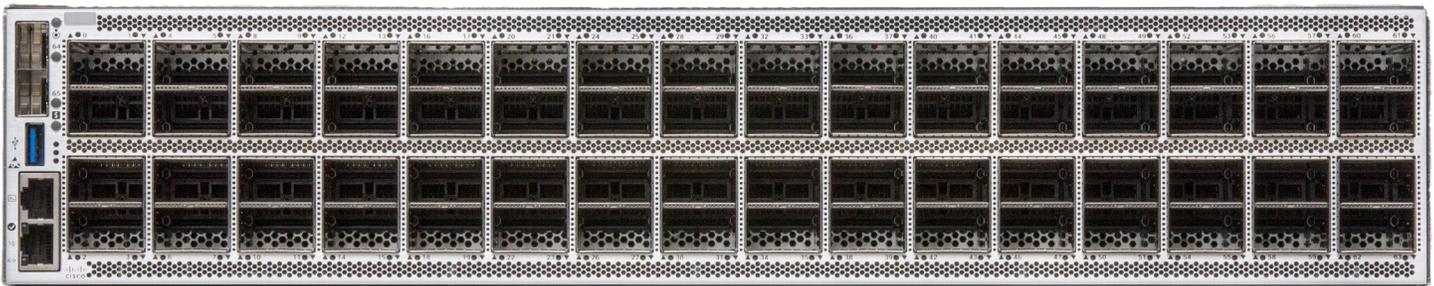


Figure 3. Cisco 8122-64EHF-O switch

Cisco 8101-32FH-O

The Cisco 8101-32FH-O is a 1RU Silicon One NPU-based high-density 400G port-capable switch supporting 32 ports of QSFPDD with breakout support. This switch will be used in storage leaf and high-speed management leaf roles.

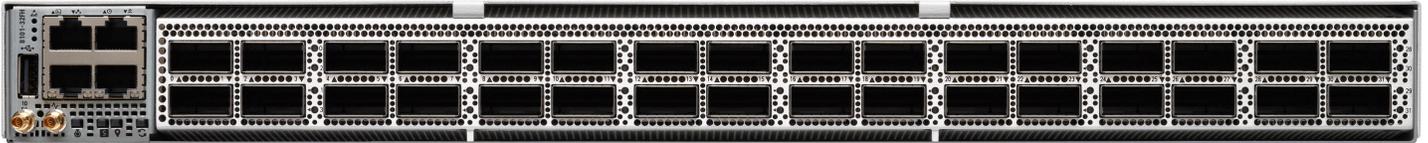


Figure 4. Cisco 8101-32FH-O switch

Cisco 8101-60Z4FH-O

The Cisco 8101-60Z4FH-O 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 400G QSFPDD with breakout support. This switch will be used in a low-speed management leaf role.



Figure 5. Cisco 8101-60Z4FH-O 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, management and support nodes, control nodes for Kubernetes (K8s) and Slurm, etc. Within this CRA, 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 as shown in Table 1 are being used on different devices in the solution.

Table 1. Supported List of Cisco Optics and Cables on different devices

Device	Optics and Cables
B3140H, B3240	QSFP-400G-DR4 with CB-M12-M12-SMF cable
B3220, B3220L	QSFP-200G-SR4 with CB-M12-M12-MMF cable
8122-64EHF-O	OSFP-800G-DR8 with dual CB-M12-M12-SMF cable
8101-32FH-O	QDD-400G-DR4 with CB-M12-M12-SMF cable QDD-400G-SR8-S with CB-M16-M12-MMF cable QDD-2Q200-CU3M passive copper cable QSFP-200G-SR4 with CB-M12-M12-MMF cable
8101-60Z4FH-O	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

Overview

Overall, the networking topology is split into two separate fabrics:

- East-West Compute Network
- Converged North-South Storage and Management Network

There are three additional networks for switch management, NVIDIA Bluefield[®]-3 management, and PDU connectivity, as described in the following sections.

East-West compute network

The compute network is meant for collective communications between the GPUs while solving a scientific problem or executing AI training. Customers looking to scale up to a maximum of 8K GPUs, can deploy the compute network in a two-tier topology with the use of 128 8122-64EHF-O leaf switches and 64 8122-64EHF-O spine switches, as shown in the following section. Customers interested to incrementally scale up to 32K GPUs should deploy the compute network with a three-tier topology from the beginning.

Two-tier East-West compute network

The two-tier topology can be incrementally deployed in units of a Scalable Unit (SU) where each SU consists of 32 NVIDIA HGX™ H200 systems for a total of 256 GPUs. As shown in Figure 7, the leaf switches are grouped into four Rail groups 1+5, 2+6, 3+7, and 4+8 deploying an SU in a rail-optimized manner. The number of leaf switches in each group can be incrementally increased upto a maximum of 32 leaf switches per group. For example, when SU31 is added, leaf switches L31, L63, L95, and L127 are added in Rail group 1+5, 2+6, 3+7, and 4+8, respectively. All GPUs within an SU, are one

hop away from each other. However, GPUs across SUs will have to communicate through the spine switches. This approach of grouping rails and rail-optimization per SU makes it efficient in allocating resources in a multi-tenant environment. It avoids one tenant’s excessive resource usage from degrading another’s performance and ensures that all tenants benefit from shared, yet isolated, infrastructure for large-scale AI training workloads.

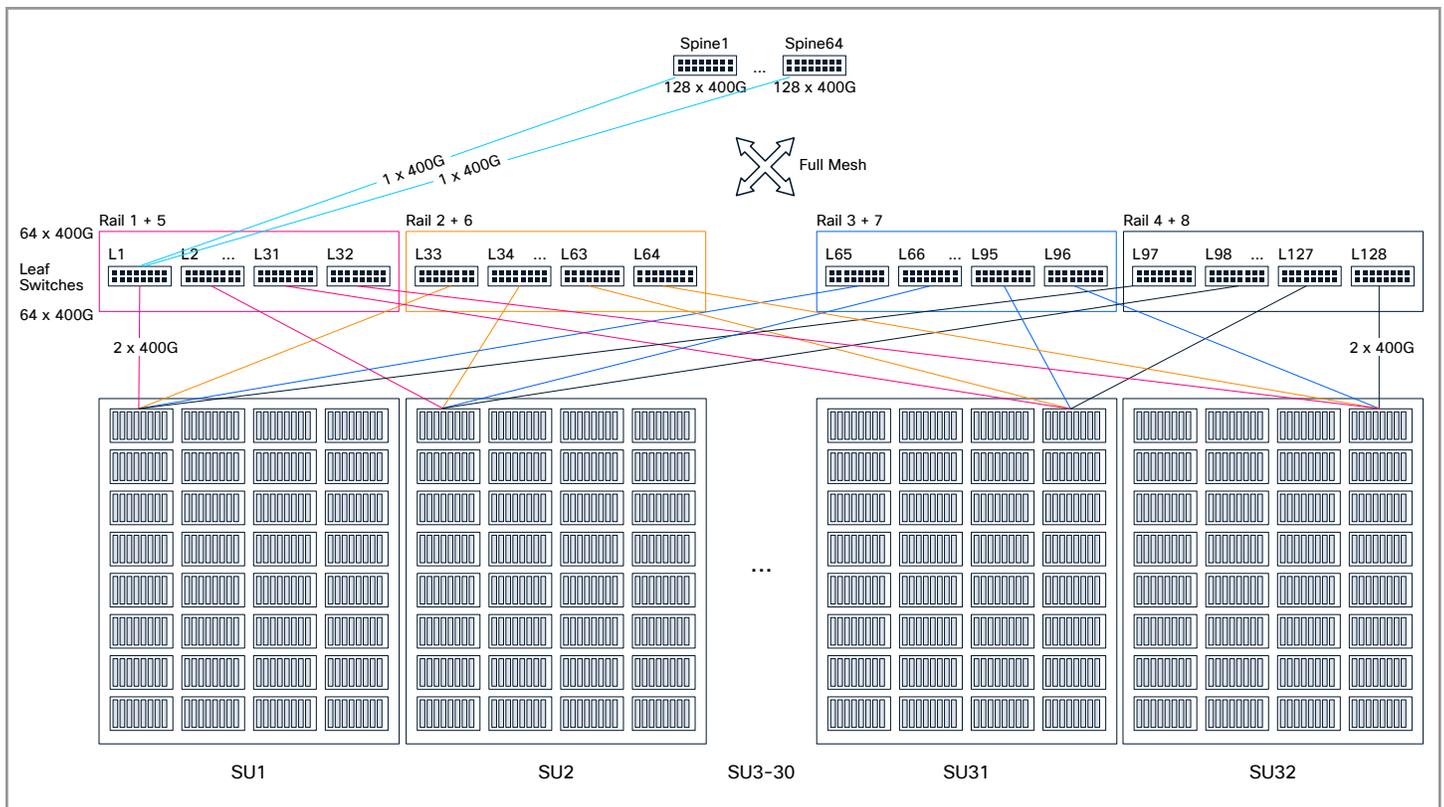


Figure 7. East-West two-tier Compute Network for 1024 HGX™ H200 nodes (8K GPUs)

Three-tier East-West compute network

As shown in Figures 8 and 9, a three-tier compute network is built in a modular way consisting of an SU-group of four Scalable Units (SUs), where each SU consists of 32 NVIDIA HGX™ H200 systems, for a total of 256 GPUs in an SU and 1024 GPUs in an SU-group. This allows incrementally deploying in units of SU-group of 1K GPUs. A three-tier leaf, spine, and super-spine topology is used so that incremental deployment doesn't require major re-cabling, and the whole fabric can scale upto 32 SU-groups, with a total of 32K GPUs. The super-spine layer consists of

four groups with switch count in each group ranging from 2 to 64. This network is built exclusively with 8122-64EHF-O switches optionally supporting NVIDIA Spectrum™-X Ethernet. Within an SU-group, GPUs in an SU are one hop away but GPUs across SUs will have to communicate through the spine switches. GPUs across SU-groups, will communicate through the super-spine switches. This architecture can also be scaled further to 64K or 128K GPUs with the use of 8-SU (2K GPUs) or 16-SU (4K GPUs) as building block within an SU-group respectively.

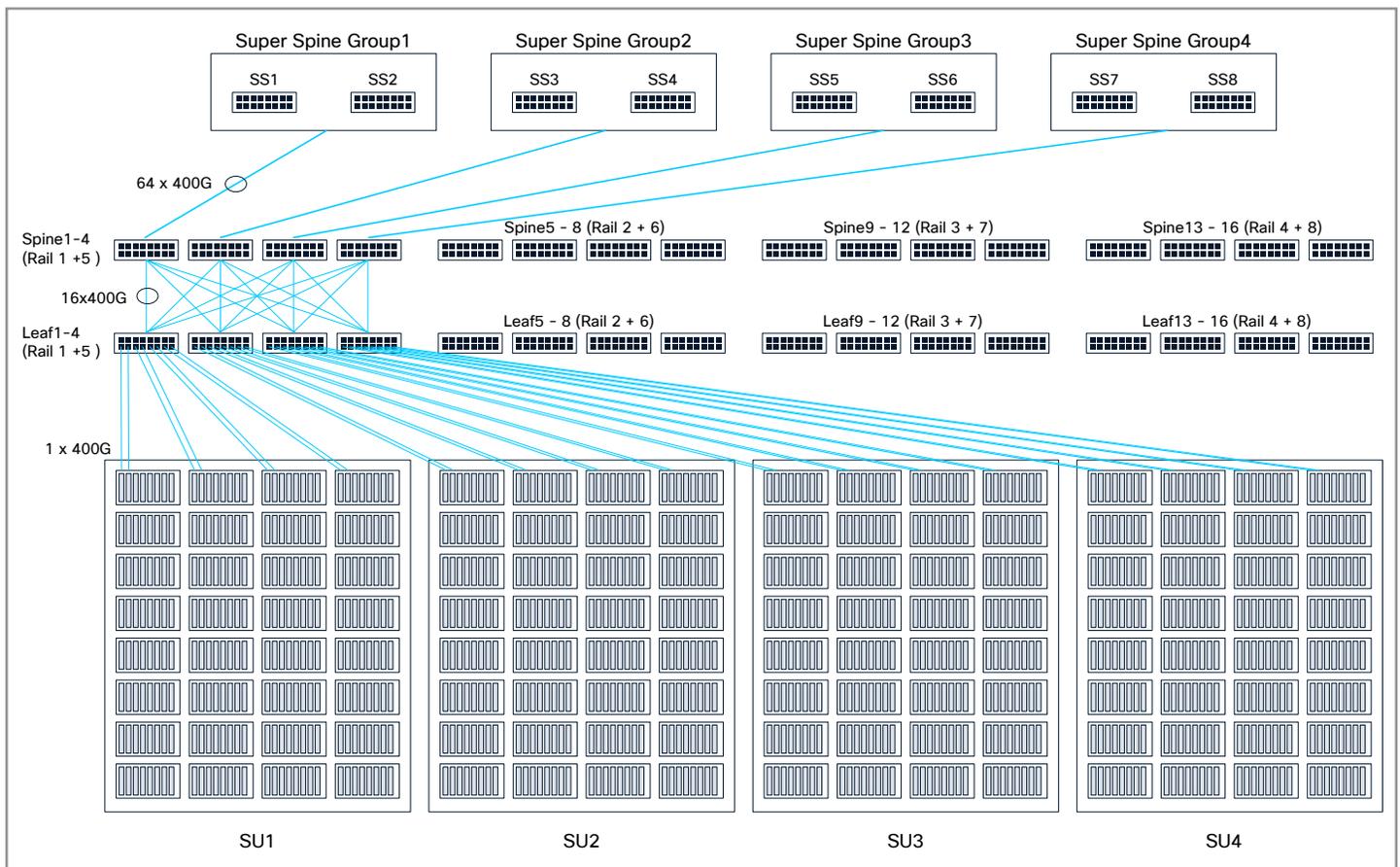


Figure 8. East-West three-tier Compute Network for 128 HGX™ H200 nodes (1K GPUs)

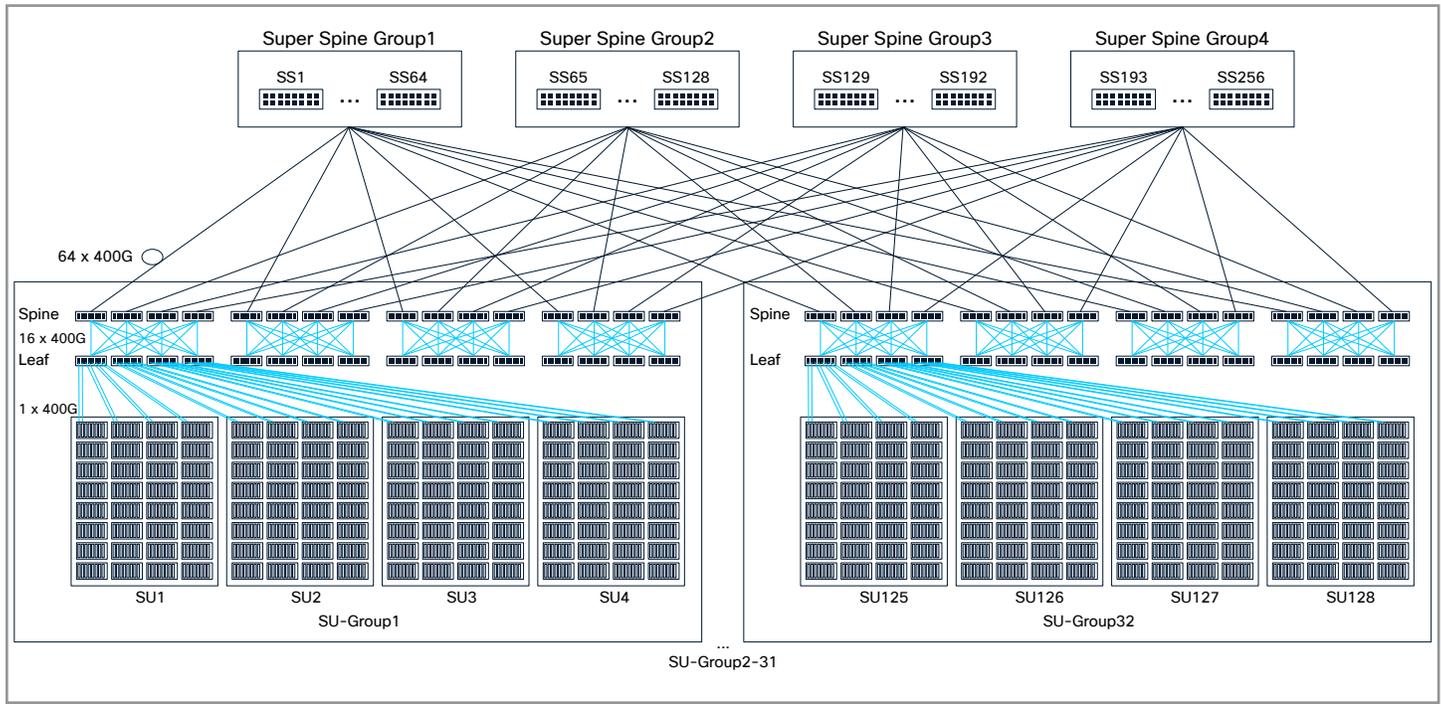


Figure 9. East-West three-tier Compute Network for 4096 HGX™ H200 nodes (32K GPUs)

The number of switches, transceivers, and cables required to build the compute network of different scale ranging from 1K to 32K GPUs is captured in Table 2.

Table 2. Three-Tier East-West Compute Network Switch, Transceivers, and Cable counts

Compute counts			Switch counts			Transceiver counts			Cable counts	
Nodes	GPUs	SUs	Leaf	Spine	SuperSpine	Node to Leaf		Switch to Switch (800G)	Node to Leaf	Switch to Switch
						Node (400G)	Leaf (800G)			
128	1024	4	16	16	8	1024	512	2048	1024	2048
256	2048	8	32	32	16	2048	1024	4096	2048	4096
512	4096	16	64	64	32	4096	2048	8192	4096	8192
1024	8192	32	128	128	64	8192	4096	16384	8192	16384
2048	16384	64	256	256	128	16384	8192	32768	16384	32768
4096	32768	128	512	512	256	32768	16384	65536	32768	65536

Converged North-South storage and management network

The converged North-South network is separate from the East-West Compute Network and serves the following key functions:

- Provides access to high performance storage from compute nodes.
- Provides host management related access to compute nodes from Management nodes.
- Interconnects with border leaf exit switches to forward traffic in and out of cluster.
- Allows interconnecting to additional customer infrastructure such as data lakes, and other nodes for support, monitoring, log collection, etc., that a cloud provider wishes to add.

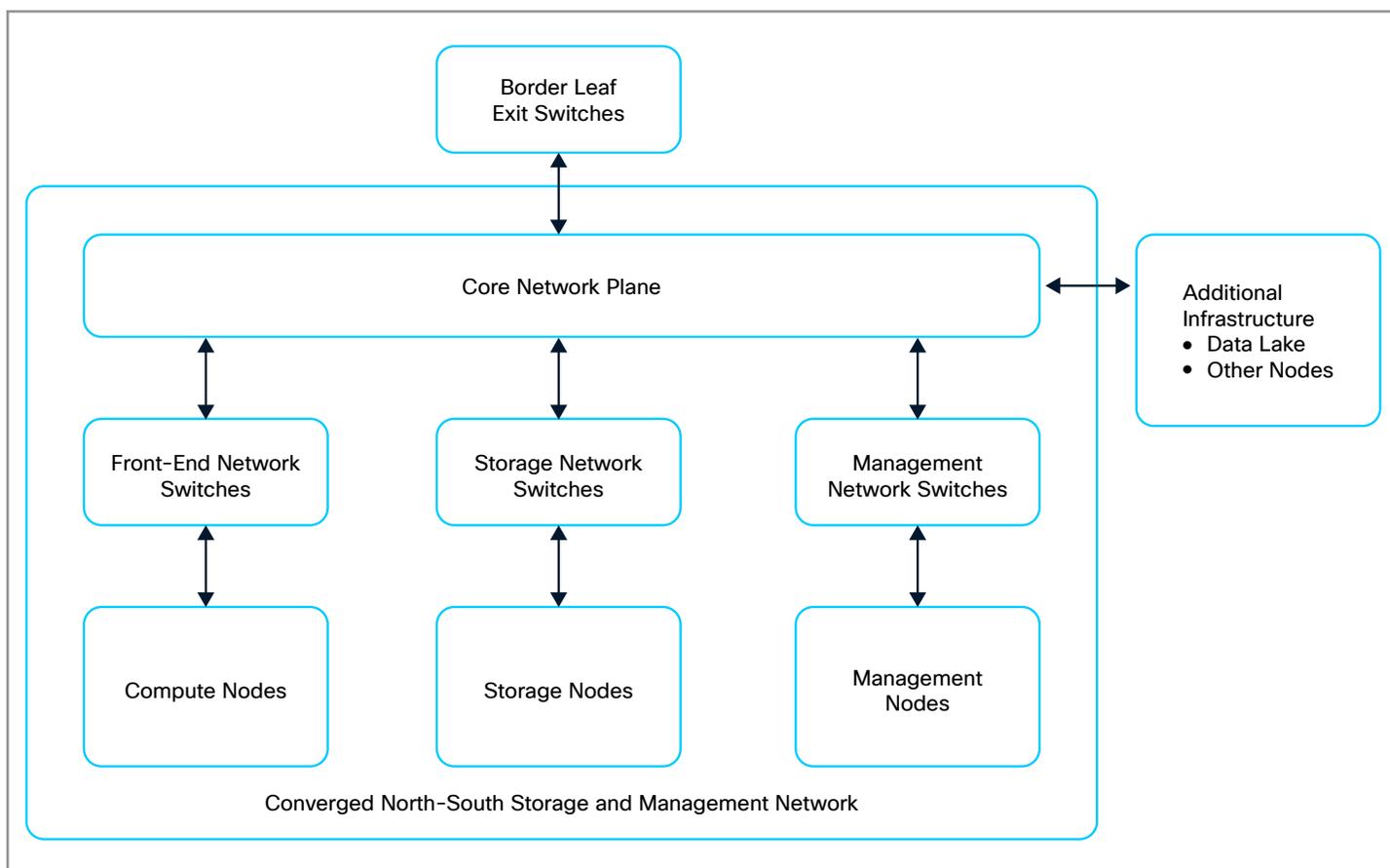


Figure 10. Logical view of Converged North-South Storage and Management Network

The physical design of this network follows different approaches based on the scale of the compute nodes involved.

Between 1K and 4K GPUs

As shown in Figure 11, at a scale of up to 4K GPUs, the converged network can be built using a two-stage leaf-spine topology, where the eight 8122-64EHF-O spine switches interconnect directly to the compute-node front-end leaf switches, low-speed out-of-band (OOB) management leaf switches, storage-node leaf switches, and in-band high-speed management node leaf switches. From a connectivity stand-point, the key design points are the following:

- The compute nodes are grouped into two SUs to create a half SU-group consisting of 64 nodes. Each half SU-group uses two 8122-64EHF-O front-end leaf switches. A total of eight half SU-groups (four SU-groups) are required to deploy 4K GPUs. Each front-end leaf switch 8122-64EHF-O uses 64 400GE downlinks to the compute nodes and 32 400GE uplinks (8 400GE to each spine switch) to provide greater than 12.5 Gbps target storage bandwidth per H200 GPU.
 - The host two 10G and BMC 1G ports of 512 compute nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access to these low-speed ports.
- On the storage side, networking will be pre-provisioned to deliver at least the amount of bandwidth requirement as mentioned above. Five redundant leaf-switch pairs, each with 64 200GE downlink ports per leaf pair, will be used for a total of 10 8101-32FH-O storage leaf switches and 320 200GE downlink ports. For uplinks, per storage leaf, 16 400GE ports will be used (2 400GE links per spine) for a total of 160 400GE ports. The number of storage nodes to connect to the storage leaf switches will vary based on cloud partners throughput and capacity needs on top of the minimum requirements.
 - The host 10G and BMC 1G ports of up to 80 storage nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access to these low-speed ports.
- The management node network consists of two parts:
 - A group of 24 management nodes (as per the configuration in Appendix B) are connected to a redundant pair of high-speed 8101-32FH-O management switches via 200GE links. These high-speed management switches connect via 1 400GE uplink to each of the 8 spines. The management nodes are used for equipment provisioning, workload orchestration, and monitoring.
 - The RA will allow connecting additional nodes to the cluster (via additional leaf switches) for any other operational needs of the customer.
 - The host 10G and BMC 1G ports of the management nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access to these low-speed ports.



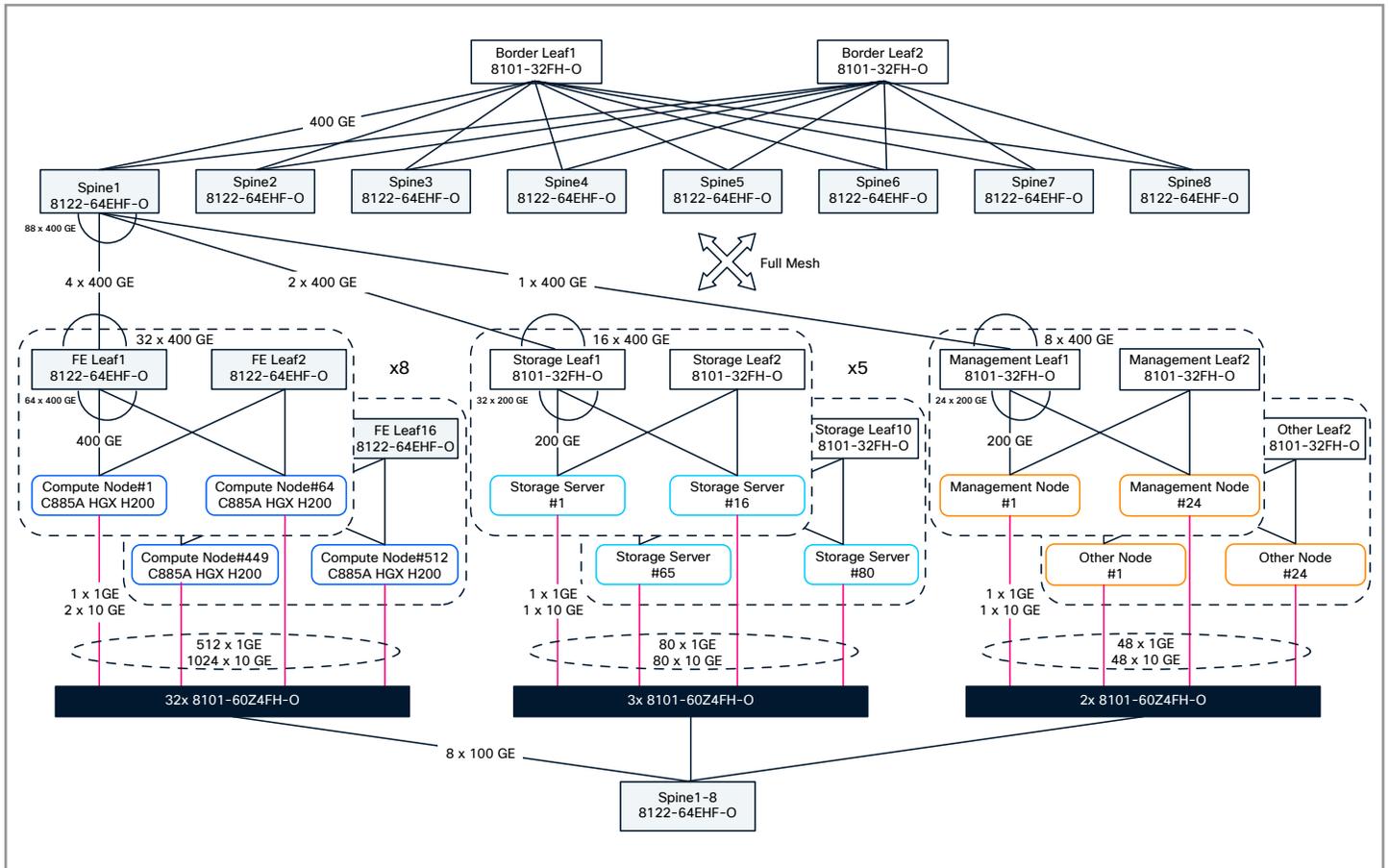


Figure 11. Converged North-South Storage and Management Network for 512 HGX™ H200 nodes (4K GPUs)

Between 4K and 8K GPUs

Beyond a scale of 4K switches, as shown in Figure 12, the converged network uses a three-tier Clos topology with leaf, spine, and the third stage constructed of a parallel plane with four core-group switches. The core-group switches connect to the compute node front-end network's spine switches, storage network spine switches, and management network spine switches.

- Each core-group will consist of a single 128 400GE port 8122-64EHF-O switch (for a total of four switches in four core-groups) to meet the scale of 8K GPUs.
- 128 compute nodes (1K GPUs) are grouped into an SU-group connected to four leaf switches. Two SU-groups (2K GPUs) connect to eight 8122-64EHF-O leaf switches, which in turn connect to four 8122-64EHF-O spine switches. Four such parallel planes each with two SU-groups, are required to deploy 8K GPUs. In this design, each leaf switch with 64 400GE downlink ports connects to each spine switch via 8 400GE ports, and will still allow 12.5 Gbps of target storage bandwidth to each GPU. Each of the four spine switches connect to their respective core-group switches via 16 400GE ports.
 - The host two 10G and BMC 1G ports of 1024 compute nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access to these low-speed ports.
- The storage network consists of 10 pairs of 8101-32FH-O leaf switches, each connected to four 8122-64EHF-O spine switches. Each redundant leaf-switch pair has 64 200GE downlink ports and 32 400GE uplink ports (8 to each spine switch) for a total of 20 leaf switches, 640 200GE downlink ports and 320 400GE uplink ports. Each spine switch connects to its corresponding core-group switch via 40 400GE ports. On an aggregate basis, this provides around 64 Tbps of storage bandwidth, or 8 Gbps to each GPU.
 - The host 10G and BMC 1G ports of up to 160 storage nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access to these low-speed ports.
- The management node network consists of two parts:
 - A group of 48 management nodes (as per configuration in Appendix B) are connected to two redundant pairs of high-speed 8101-32FH-O management switches via 200GE links. Each of these high-speed management switches connects via 6 400GE uplinks to 2 8122-64EHF-O spine switches. Each of the spine switches connects to two dedicated core-group switches via 12 400GE links (6 400GE links to each core switch). The management nodes are used for equipment provisioning, workload orchestration, and monitoring.
 - The RA will allow connecting additional nodes to the cluster (via additional leaf switches) for any other operational needs of the customer.
 - The host 10G and BMC 1G ports of management nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access to these low-speed ports.

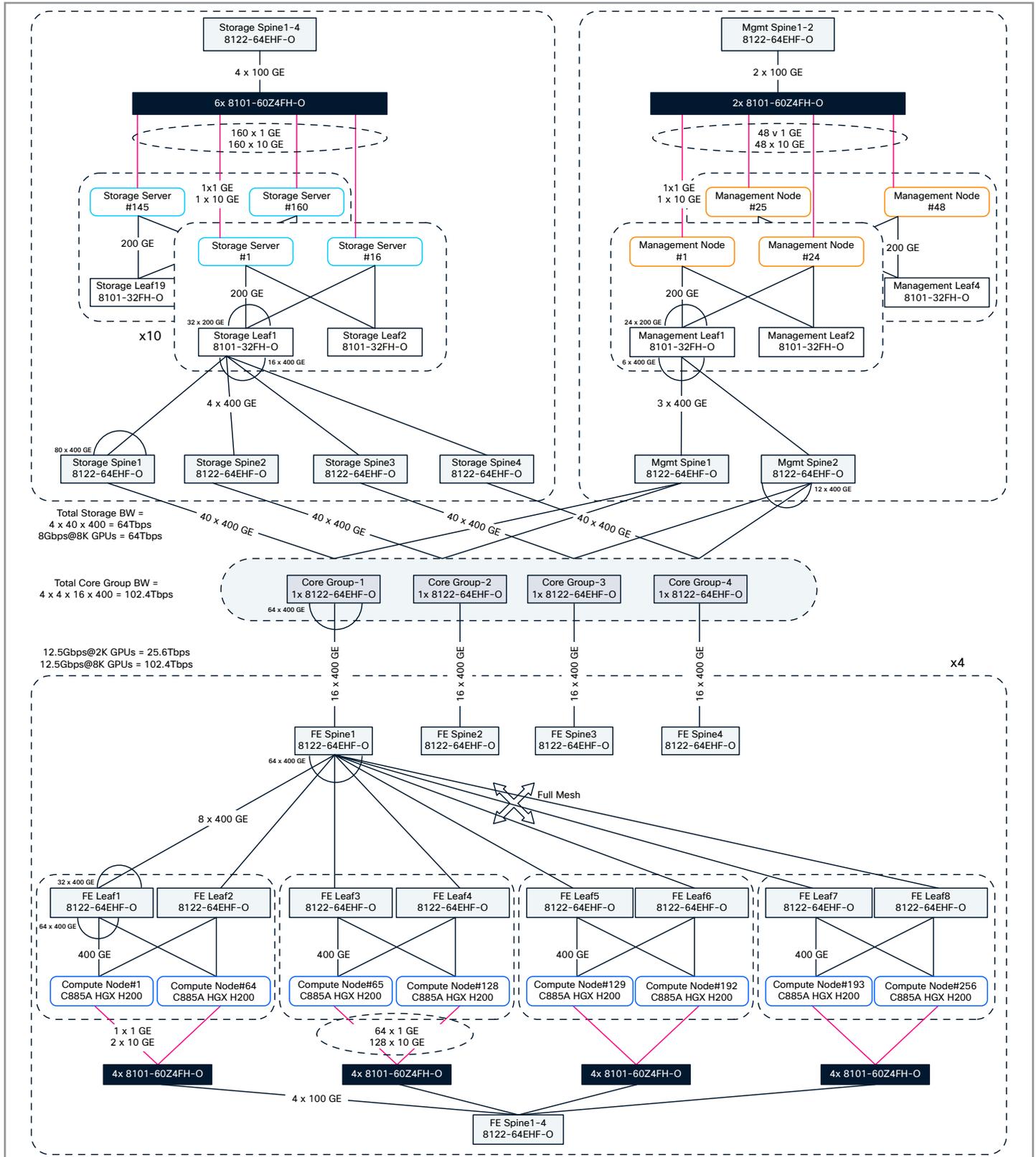


Figure 12. Converged North-South Storage and Management Network for 1024 HGX™ H200 nodes (8K GPUs)

Between 8K and 16K GPUs

As shown in Figure 13, to accommodate the scale increase to 16K GPUs:

- The number of switches within a core-group are doubled to two for a total of eight core switches.
- Eight parallel planes each with two SU-groups of compute nodes (2K GPUs) are used to deploy 16K GPUs. Spine switches in each of these parallel planes connect to their respective core-group switches via 16 400GE links (8 400GE per core switch).
 - The host two 10G and BMC 1G ports of 2048 compute nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access these to low-speed ports.
- The storage network consists of 20 pairs of 8101-32FH-O leaf switches, each connected to 8 8122-64EHF-O spine switches. Each redundant leaf-switch pair has 64 200GE downlink ports and 32 400GE uplink ports (4 to each spine switch) for a total of 40 leaf switches, 1280 200GE downlink ports and 640 400GE uplink ports. Each spine switch connects to its corresponding core-group switch via 40 400GE ports. On an aggregate basis, this provides around 128 Tbps of storage bandwidth, or 8 Gbps to each GPU.
 - The host 10G and BMC 1G ports of up to 320 storage nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access to these low-speed ports.
- The management node network consists of two parts:
 - A group of 96 management nodes (as per the configuration in Appendix B) are connected to four redundant pairs of high-speed 8101-32FH-O management switches via 200GE links. Each of these high-speed management switches connect via 3 400GE uplinks to 2 8122-64EHF-O spine switches. Each of the spine switches connect to two dedicated core-group switches via 24 400GE links (12 400GE links to each core switch). The management nodes are used for equipment provisioning, workload orchestration, and monitoring.
 - The RA will allow connecting additional nodes to the cluster (via additional leaf switches) for any other operational needs of the customer.
 - The host 10G and BMC 1G ports of management nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access to these low-speed ports.



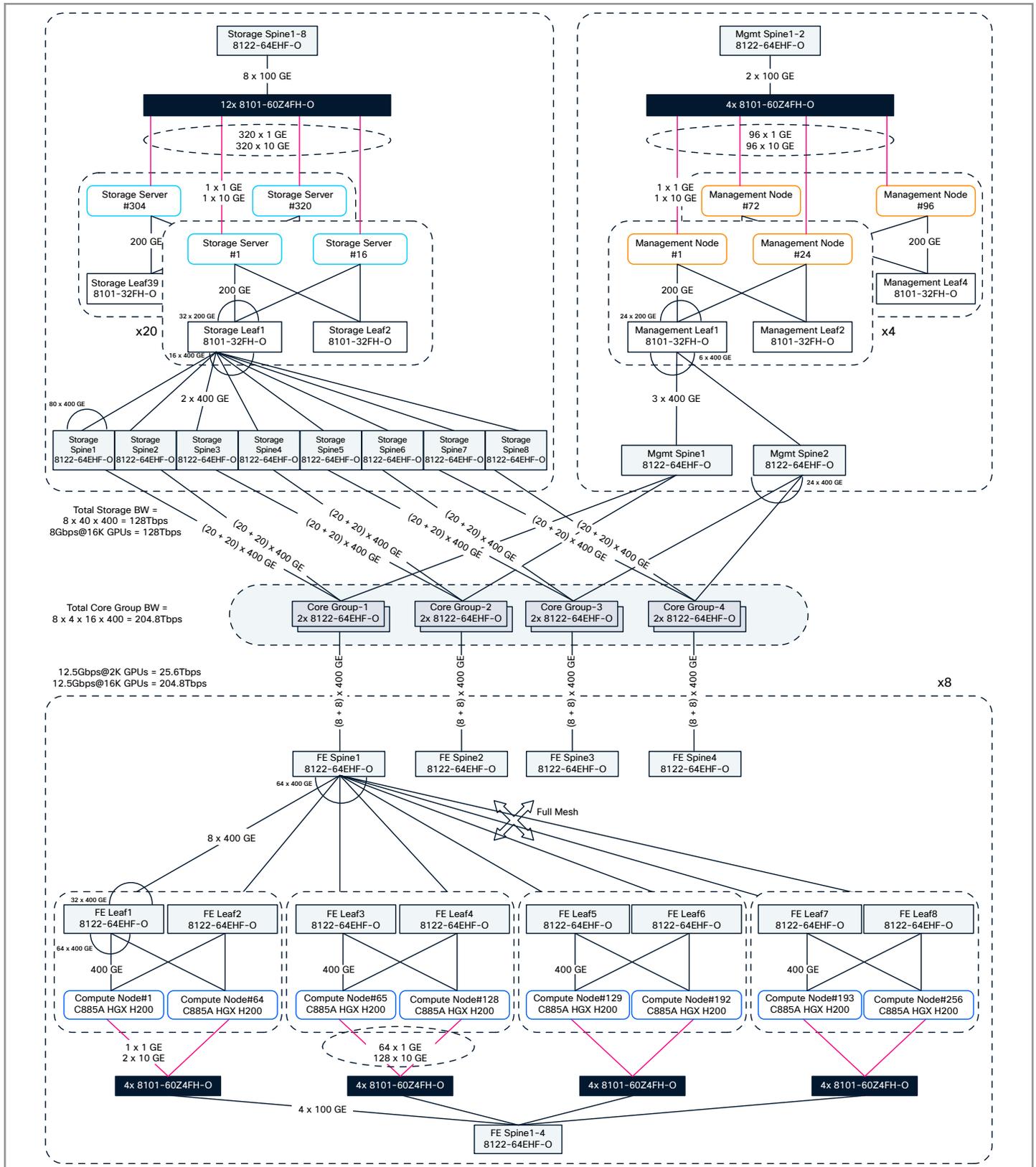


Figure 13. Converged GE North-South Storage and Management Network for 2048 HGX™ H200 nodes (16K GPUs)

Between 16K and 32K GPUs

To accommodate the scale increase to 32K GPUs:

- The number of switches within a core-group is doubled to 4 for a total of 16 core switches.
- 16 parallel planes each with two SU-groups of compute nodes (2K GPUs), are used to deploy 32K GPUs. Spine switches in each of these parallel planes connect to their respective core-group switches via 16 400GE links (4 400GE links per core switch).
 - The host two 10G and BMC 1G ports of 4096 compute nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access to these low-speed ports.
- The storage network consists of 40 pairs of 8101-32FH-O leaf switches, each connected to 16 8122-64EHF-O spine switches. Each redundant leaf-switch pair has 64 200GE downlink ports and 32 400GE uplink ports (2 to each spine switch) for a total of 80 leaf switches, 2560 200GE downlink ports, and 1280 400GE uplink ports. Each spine switch connects to its corresponding core-group switches via 40 400GE ports. On an aggregate basis, this provides around 256 Tbps of storage bandwidth, or 8 Gbps to each GPU.
 - The host 10G and BMC 1G ports of up to 640 storage nodes are connected to high-density low-speed port 8101-60Z4FH-O switches with a 100G uplink to each spine switch. Tenants are not given access to these low-speed ports.
- The management node network consists of two parts:
 - A group of 192 management nodes (as per the configuration in Appendix B) are connected to

8 pairs of high-speed 8101-32FH-O management switches via 200GE links. Each of these high-speed management switches connect via 6 400GE uplinks to 2 8122-64EHF-O spine switches. Each of the spine switches connect to two dedicated core-group switches via 48 400GE links (6 400GE links to each core switch). The management nodes are used for equipment provisioning, workload orchestration, and monitoring.

- The RA will allow connecting additional nodes to the cluster (via additional leaf switches) for any other operational needs of the customer.
- An additional part of the management node network is to connect the host 10G and BMC 1G ports of the management nodes via high-density, low-speed port switches with a 100G uplink to each spine switch.

Switch management network

The switch management network is used to connect the 1G management ports of the 8000 series switches to high 1G port density Cisco® Catalyst® C9350-48T switches. The uplinks of many C9350-48T switches are aggregated to redundant modular switches. This network is not exposed to the tenants. However, it is made accessible to controllers for provisioning, observability, and overall cluster management.

NVIDIA Bluefield®-3 Management network

The NVIDIA Bluefield®-3 management network is used to connect the 1G BMC port of NVIDIA Bluefield®-3 DPUs to high 1G port density Cisco® Catalyst® C9350-48T switches. This network is also not exposed to the tenants and is mainly intended for out-of-band management of NVIDIA Bluefield®-3 DPUs.

Miscellaneous service network

The miscellaneous service network is used to connect PDUs across all racks to high 1G port density Cisco® Catalyst® C9350-48T switches. This network is also not exposed to the tenants and is mainly intended for overall power management including such activities

as AC power cycle or shutdown of equipment, power measurement etc. Additional equipment that can be connected to this network include terminal servers for console access to switches and other equipment.

Cluster BOM

Table 4 shows the bill-of-materials (BOM) for building cluster of different sizes, from 4K to 32K GPUs.

Table 4. Minimum BOM for cluster with 4K to 32K GPUs

PID	Description	4K GPUs	8K GPUs	16K GPUs	32K GPUs
UCSC-885A-M8-HC1	Cisco UCS C885A M8 Rack Server with NVIDIA HGX™	512	1024	2048	4096
8122-64EHF-O	Cisco 8000 series switch, 64x800Gbps OSFP	184	378	754	1506
8101-32FH-O	Cisco 8000 series switch, 32x400Gbps QSFP-DD	12	26	50	98
8101-60Z4FH-O	Cisco 8000 series switch 60x50G SFP28 4x400G QSFP-DD	36	72	144	288
OSFP-800G-DR8	800G OSFP transceiver, 800GBASE-DR8, SMF dual MPO-12 APC, 500m	11500	23128	46256	92512
QDD-400G-DR4	400G QSFP-DD transceiver, 400GBASE-DR4, MPO-12, 500m parallel	324	640	1264	2528
QSFP-400G-DR4	400G QSFP112 transceiver, 400GBASE-DR4, MPO-12, 500m parallel	5120	10240	20480	40960
QDD-400G-SR8-S	400G QSFP-DD transceiver, 400GBASE-SR8, MPO-16 APC, 100m	184	368	736	1472
QSFP-200G-SR4-S	200G QSFP transceiver, 200GBASE-SR4, MPO-12, 100m	368	736	1472	2944
SFP-1G-T-X	1G SFP	616	1232	2464	4928
SFP-10G-T-X	10G SFP	1128	2256	4512	9024
CB-M12-M12-SMF	MPO-12 cables	14164	28696	57392	114784

PID	Description	4K GPUs	8K GPUs	16K GPUs	32K GPUs
CB-M16-M12-MMF	MPO-16 to dual MPO-12 breakout cables	184	368	736	1472
CAT6A	Copper cable for 10G	1128	2256	4512	9024
CAT5E	Copper cable for 1G	616	1232	2464	4928
UCSC-C225-M8N (storage server)	Cisco UCS C225-M8 1RU Rack Server	80	160	320	640
UCSC-C225-M8N UCSC-C245-M8SX (management node)	Cisco UCS C225-M8 1RU Rack Server Cisco UCS C245-M8 2RU Rack Server	24	48	96	192

Multitenancy

The entire networking fabric is configured using VXLAN data-plane and BGP EVPN control-plane enabling native support for multitenancy. All resources assigned to the tenants such as bare metal or virtualized compute nodes, management nodes, and access to storage, are completely isolated. Multitenancy is supported throughout the fabric via L2 or L3 segmentation. Every leaf switch host facing port can be assigned to the appropriate VLAN, VNI and VRF to isolate tenant traffic.

There are two VRFs whose access is limited to the CSP and direct access to them is not allowed to the tenants:

1. Out-of-Band (OOB) Management Network: vrfOOBMgmt VRF.
2. Storage Internal Network: vrfStorageInternal VRF.

Out-of-Band (OOB) Management network

The BMC 1G and Host 10G ports of compute, storage, and management nodes are part of the OOB management network. The switch ports connecting to these 1G and 10G ports are put into a separate logical network (untagged VLANs) and are part of vrfOOBMgmt VRF. Tenants are not given access to this network for security reasons.

Storage Internal network

As described in the “Storage Architecture” section, every storage node has 2 DPUs where NIC-0 is used for internal storage server to storage server communication and NIC-1 is used for external communication to clients such as compute nodes, management nodes etc. The ports on NIC-0 of all storage servers are part of a separate vrfStorageInternal VRF whose access is limited only to the CSP and not allowed to the tenants.

Every tenant is allocated at least two VRFs:

1. Compute Network: vrf<Tenant>Backend VRF.
2. Converged Storage and Management Network: vrf<Tenant>Frontend VRF.

Compute network

The routes in the backend East-West network of compute nodes assigned to a tenant are isolated into vrf<Tenant>Backend VRF.

Converged storage and management network

For every tenant, a separate tenant account is created in the high-performance storage, thereby allowing further provisioning of storage resources assigned to the tenant. A separate VLAN is also assigned to isolate tenant's storage access from the compute and management nodes assigned to the tenant. This VLAN's VxLAN VNI is part of vrf<Tenant>Frontend VRF.

Workload orchestration

Each tenant is assigned a group of management nodes that can be used for:

- Provisioning the compute nodes either via Cisco Intersight or NVIDIA Base Command Manager (BCM) or additional provisioning tools/frameworks.
- Setup Slurm and/or Kubernetes control nodes for orchestrating jobs on worker compute nodes.
- Additional infrastructure for observability, monitoring, and logs collections.

These management nodes are accessible to the tenant via the vrf<Tenant>Frontend VRF.

Edge and Border connectivity

The converged fabric connects to two or more border leaf switches with a redundant number of links to allow forwarding data into and out of the cluster for both the CSP as well as the tenants. The border leaf switches perform L3 routing while all VxLAN encapsulation and decapsulation are done inside the converged network fabric. The number of border leaf switches, the number of links between them and the converged network fabric, the networking feature sets enabled on them would vary as per customer use case and are beyond the scope of this RA.

Storage architecture

Cisco has partnered with [VAST Data](#) to onboard their AI OS on Cisco UCS C225-M8N Rack Servers in EBox architecture: together, they provide the storage subsystem for this RA. This product is called Cisco EBox and it is NVIDIA-Certified high-performance storage for both NCP and Cisco Cloud Reference Architecture based large GPU scale Secure AI Factory. 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 a single namespace. This allows building clusters of different sizes with varying number of storage servers. Additional features include native support for multitenancy, multiprotocol (NFS, S3, SMB), data reduction, data protection, cluster high availability, serviceability of failed hardware components etc.

Figure 14 shows the overall network connectivity of storage servers. For data path, each server uses two NVIDIA BlueField®-3 B3220L 2x200G DPUs – NIC0 is used for internal network within the servers allowing any server to access storage drives from any other server, NIC1 is used for external network supporting client traffic such as NFS, S3, SMB. The 1G BMC and 10G x86 management ports are connected to a management leaf switch.

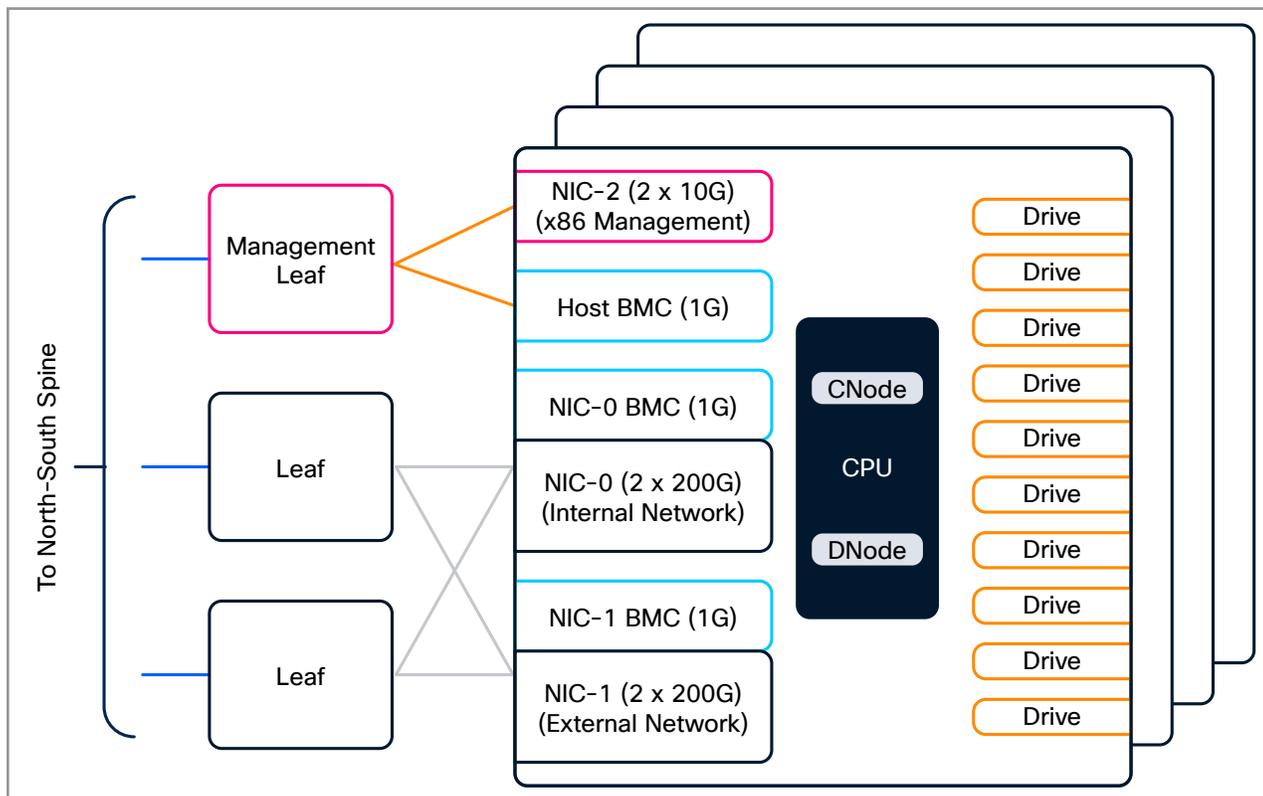


Figure 14. Cisco EBox Storage Logical Block Diagram

Software

To deploy and manage a high-scale AI cluster, a robust software stack is required with an automation-first approach. The use of controllers along with their programmability interfaces can tremendously simplify day-0 resource provisioning, day-1 configuration, and day-N operationalization. The following sub-sections cover the key software components involved in this reference architecture.

Network Management

SONiC supports a number of programmability interfaces such as CLIs, REST-APIs, gNMI with native YANG models that can be used by CSPs in their automation

frameworks to fully configure, monitor, and manage the switches at scale.

Compute controller

[Cisco Intersight](#) is used to do provisioning of the Cisco UCS C885A M8 Rack Servers as well as their end-to-end life cycle management. It also supports integration with other automation frameworks via [RESTful APIs](#). Cloud partners or tenants can also choose to use on-prem NVIDIA Base Command Manager or additional open-source or custom tools or frameworks via the management nodes for compute node provisioning.

Storage controller

The Cisco EBox storage controller (also known as VAST Management Service) will be used for provisioning and managing the attached high-performance storage. Besides this, cloud partner and every tenant is also allocated a storage management URL, a user login, and a dashboard for configuration, monitoring, and overall management. RESTful APIs are supported for integration with automation frameworks.

NVIDIA AI Enterprise and Spectrum-X

This reference architecture includes NVIDIA AI Enterprise, deployed and supported on NVIDIA-Certified Cisco UCS C885A M8 servers. NVIDIA AI Enterprise is a cloud-native suite of software tools, libraries, and frameworks designed to deliver optimized performance, robust security, and stability for production AI deployments. Easy-to-use microservices enhances model performance with enterprise-grade security, support, and stability, ensuring a smooth transition from prototype to production for enterprises that run their businesses on AI.

NVIDIA NIM™ is a set of easy-to-use microservices designed for secure, reliable deployment of high-performance AI model inferencing across clouds, data centers, and workstations. Supporting a wide range of AI models, including open-source community and NVIDIA AI foundation models, it ensures seamless, scalable AI inferencing on premises and in the cloud with industry-standard APIs.

NVIDIA® Spectrum™-X Ethernet Networking Platform, featuring Spectrum™-X Ethernet switches and Spectrum™-X Ethernet SuperNICs, is the world's first Ethernet fabric built for AI, accelerating generative AI network performance by 1.6x. It's benefits are available with Cisco SiliconOne based 8122-64EHF-O switches used in this RA when connected to NVIDIA BlueField®-3 SuperNICs.

Security

Security in a multitenant AI infrastructure is very crucial to ensure confidentiality, integrity, and high availability against adversarial attacks by implementing robust access controls and host and network isolation to prevent unauthorized access or manipulation. A number of Cisco security technologies, as enumerated below, are available that can be deployed by CSPs and tenants to configure, monitor, and enforce end-to-end security right from applications to overall infrastructure. The complete integration of these technologies into the end-to-end workflow is beyond the scope of this RA.

- [Cisco Secure Firewall](#)
- [Cisco Isovalent](#)
- [Cisco Hypershield](#)
- [Cisco AI Defense](#)

Observability

Observability is a key element of AI infrastructure to ensure continuous visibility and reliability and to provide high-performance by tuning as well as proper infrastructure scaling. It also facilitates debugging, aids security, and helps maintain trustworthy and effective AI systems. Cisco [Splunk](#)® is an industry-leading observability solution for cloud partners as well as for tenants to ingest significant amounts of telemetry and gain in-depth visibility. It's integration within the end-to-end workflow is beyond the scope of this RA.

Testing and certification

The overall solution has been thoroughly tested considering all aspects of management plane, control plane, and data plane combining compute, storage, and networking together. The compute nodes are NVIDIA-Certified Systems™. The Cisco EBox high-performance storage solution has achieved NVIDIA-Certified Storage validation at the NCP level. A number of benchmark test suites such as HPC Benchmark, single and multi-hop IB PerfTest, NCCL collective communications tests, and high-availability (across switch and link failure) tests, MLCommons Training and Inference benchmarks have

also been run to evaluate end-to-end performance and assist with tuning. Different elements and entities of the NVIDIA AI Enterprise ecosystem have been brought up with use cases around Model Training, Fine-tuning, Inferencing, and RAG.

Summary

In short, the Cisco Cloud Reference Architecture is a fully integrated, end-to-end tested, high GPU-scale multitenant AI cluster solution offering cloud partners a one-stop shop place for their AI infrastructure deployment needs.

Appendix A – Compute server specifications

Cisco UCS C885A M8 Rack Server

Area	Details
Form Factor	8RU Rack Server (Air Cooled)
Compute + Memory	2x 5 th 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

Area	Details
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 – Management node server specifications

The versatile Cisco UCS C225 M8 1RU Rack Server will be used as support server, control node server for Slurm and Kubernetes (K8s), etc. The following table shows the minimum specifications of the server.

Area	Details
Form Factor	1RU Rack Server (Air Cooled)
Compute + Memory	1x 4 th 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 C240 M8 Rack Server or Cisco UCS C245 M8 Rack Server 2RU variant along with B3220 DPU NICs.