Cisco Nexus Hyperfabric Cloud RA with HGX B300

Contents

Featuring Networking Reference Architecture of NVIDIA HGX™ B300 with Cisco Nexus Hyperfabric

Introduction

The Cisco NCP compliant Cloud Reference Architecture (RA) is designed to be deployed with a high GPU scale 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 or 128K GPUs. The key technologies used in this RA include:

- NVIDIA HGX™ B300 paired with NVIDIA ConnectX®-8 SuperNICs, Bluefield®-3 DPUs.
- Cisco Nexus Hyperfabric™ Switches combined with Cisco compute, networking, and storage controllers.
- Cisco Optics and cables.
- Cisco provisioning, observability and security frameworks.

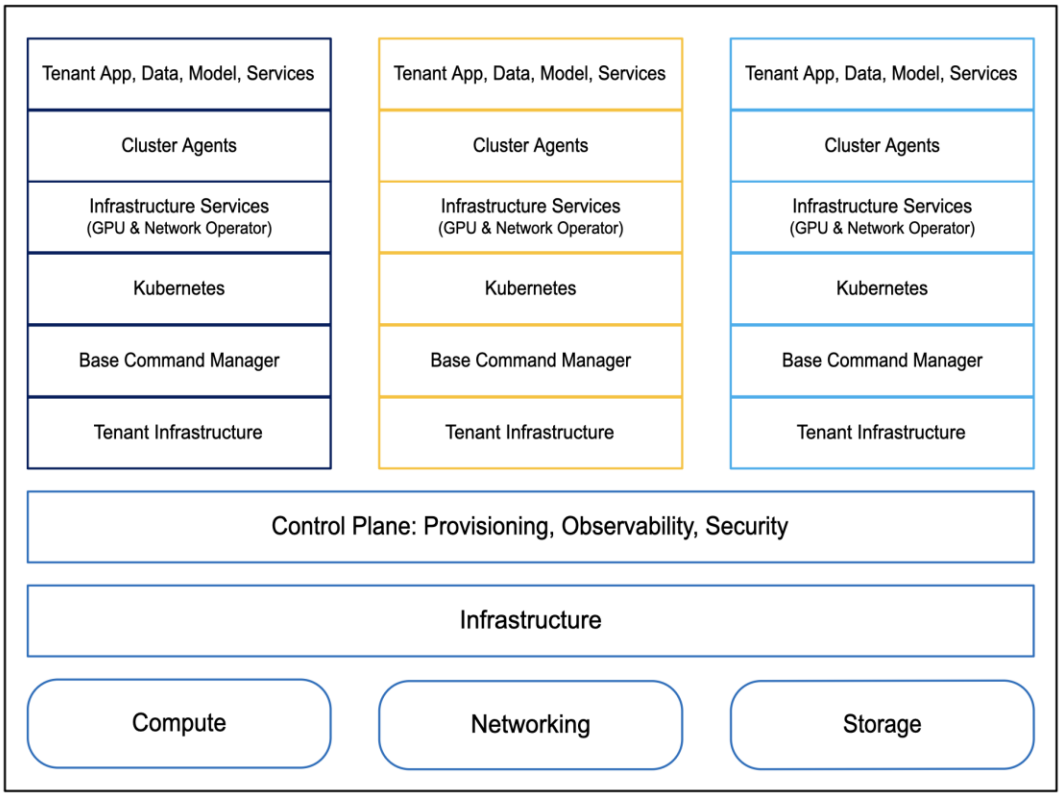

Figure 1.
Logical View of Cisco Cloud Reference Architecture

Support for the overall solution is provided by Cisco in collaboration with its partners. All customer cases are front end by Cisco and after triage, where applicable, routed to appropriate partner support on the backend driving case resolution.

Hardware

HGX B300 Rack Server

This RA uses NVIDIA-Certified® HGX™ B300 rack servers in 2-8-9-800 or 2-8-10-800 (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 B300 SXM GPUs within the server are interconnected using high speed NVLink interconnects. GPU connectivity to other physical servers is via the use of 8x integrated NVIDIA ConnectX®-8 SuperNICs for East-West traffic and via 1x or 2x NVIDIA BlueField®-3 B3240 DPU NICs (in 1x400G mode) for North-South traffic. Cisco UCS C880A M8, Supermicro NB3RT are NVIDIA-Certified® HGX™ B300 rack servers supported in this RA. Their required specification is captured in Appendix A.

Figure 2.
Cisco UCS C880A M8, Supermicro NB3RT Rack Servers with NVIDIA HGX™ B300

Cisco HF6100-64ED

The Cisco HF6100-64ED is a Cisco Silicon One™ Ethernet Switch ASIC-based 2RU switch supporting 64 800G OSFP modules allowing 64 800GE or 128 400GE ports. It will be used in both leaf and spine roles.

Figure 3.
Cisco HF6100-64ED switch

Cisco N9164E-NS4-O

The Cisco N9164E-NS4-O is a NVIDIA Spectrum™-4 Ethernet switch ASIC-based 2RU switch supporting 64 800G OSFP modules allowing 64 800GE or 128 400GE ports. This switch can be used in both leaf and spine role in East-West compute network, as an alternative to HF6100-64ED, for NCP compliance.

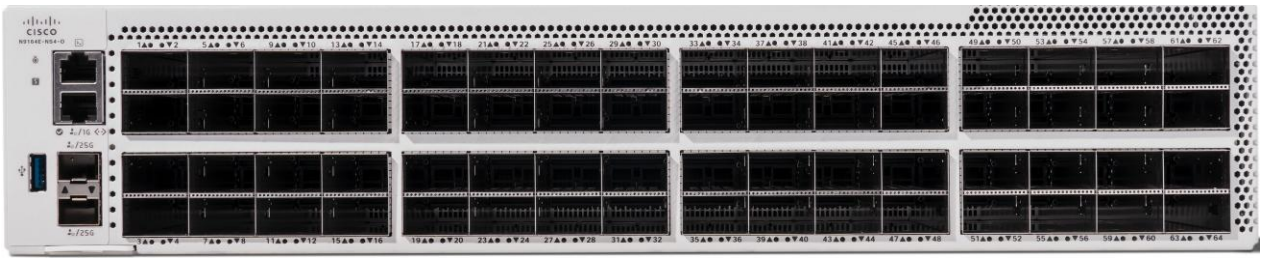

Figure 4.
Cisco N9164E-NS4-O switch

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 will be used in storage leaf and high-speed management leaf roles.

Figure 5.
Cisco HF6100-32D switch

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 400G QSFPDD with breakout support. This switch will be used in a low-speed management leaf role.

Figure 6.
Cisco HF6100-60L4D 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. Within this RA, these servers are also used to run the VAST Storage solution as described in the "High-Performance Storage" section, below.

Figure 7.
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
B3220, B3240	QSFP-400G-DR4 with CB-M12-M12-SMF cable
B3220L	QSFP-200G-SR4 with CB-M12-M12-MMF cable
ConnectX-8	OSFPR-800G-DR8 with dual CB-M12-M12-SMF cable
HF6100-64ED **N9164E-NS4-O**	OSFP-800G-DR8 with dual CB-M12-M12-SMF cable
HF6100-32D	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
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

Overview

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

- East-West Compute Network.
- Converged North-South Storage and Management Network.
- Out-of-Band (OOB) Management Network.

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 256 N9164E-NS4-O leaf switches and 128 N9164E-NS4-O spine switches, as shown in the following section. Alternatively, HF6100-64ED switches can also be used without NCP compliance. Customers interested to incrementally scale beyond 8K GPUs should deploy the compute network with a three-tier topology from the beginning. Both the two-tier and three-tier use dual-plane topology where every E-W ConnectX-8 800G NIC (in 2 x 400G mode) is connected to two separate identical planes of switches for avoiding single point of failure. Traffic is optimally load-balanced over both the planes with global awareness instead of using traditional L2/L3

bonds like LAG that use static hash-based traffic distribution. The two globally aware load-balancing scheme that could be used are: (a) Software based Plane Load Balancing (SWPLB) managed by NCCL together with the use of Spectrum-X plugin; (b) Hardware based Plane Load Balancing (HWPLB) managed by ConnectX-8 firmware with Spectrum-X enabled.

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™ B300 systems for a total of 256 GPUs. As shown in Figure 8, within both identical plane 1 and 2, 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 up to a maximum of 32 leaf switches per group. For example, when SU32 is added, leaf switches L32 are added in every Rail group. 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. An alternate supported deployment model includes using 8 Rail groups 1 to 8 at the leaf layer where every rail on the server connects to a different leaf switch instead of connecting 2 rails per leaf switch in the case of 4 Rail groups.

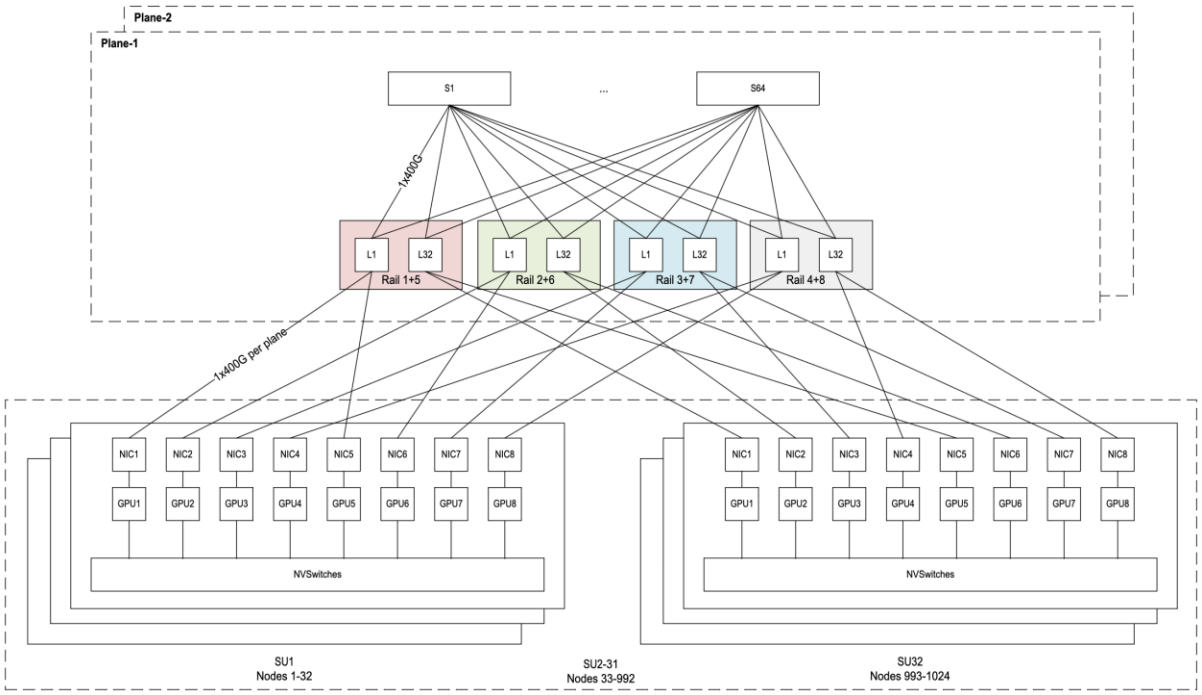

Figure 8.
East-West two-tier Compute Network for 1024 HGX™ B300 nodes (8K GPUs)

The number of switches, transceivers, and cables required to build the two-tier compute network of different scale ranging from 1K to 8K GPUs is captured in Table 2.

Table 2. Two-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 (800G)	Leaf (800G)			
128	1024	4	32	16	0	1024	1024	2048	2048	2048
256	2048	8	64	32	0	2048	2048	4096	4096	4096
512	4096	16	128	64	0	4096	4096	8192	8192	8192
1024	8192	32	256	128	0	8192	8192	16384	16384	16384

Three-tier East-West compute network

As shown in Figures 9 and 10, 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™ B300 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 up to 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 can be built with N9164E-NS4-O switches for NCP compliance or HF6100-64ED switches without NCP compliance. 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.

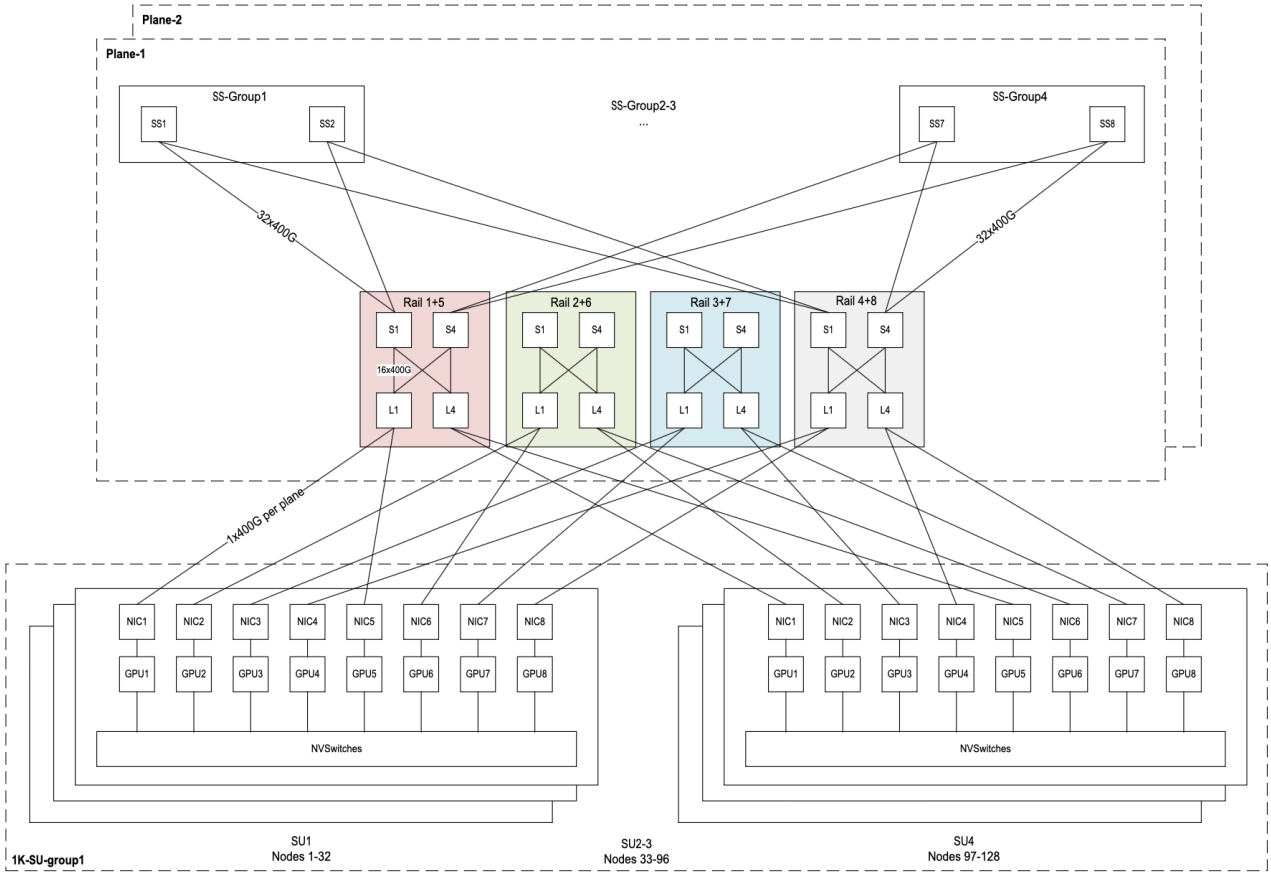

Figure 9.
East-West three-tier Compute Network for 128 HGX™ B300 nodes (1K GPUs)

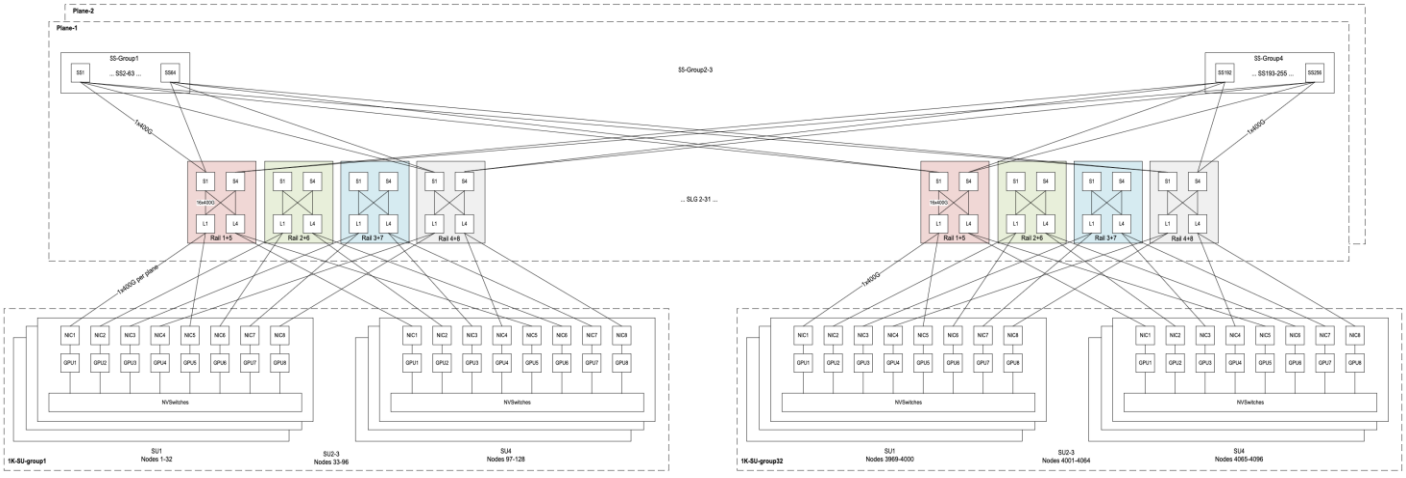

Figure 10.
East-West three-tier Compute Network for 4096 HGX™ B300 nodes (32K GPUs)

The number of switches, transceivers, and cables required to build the three-tier compute network of different scale ranging from 1K to 32K GPUs is captured in Table 3.

Table 3. 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 (800G)	Leaf (800G)			
128	1024	4	32	32	16	1024	1024	4096	2048	4096
256	2048	8	64	64	32	2048	2048	8192	4096	8192
512	4096	16	128	128	64	4096	4096	16384	8192	16384
1024	8192	32	256	256	128	8192	8192	32768	16384	32768
2048	16384	64	512	512	256	16384	16384	65536	32768	65536
4096	32768	128	1024	1024	512	32768	32768	131072	65536	131072
8192	65536	256	2048	2048	1024	65536	65536	262144	131072	262144

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.

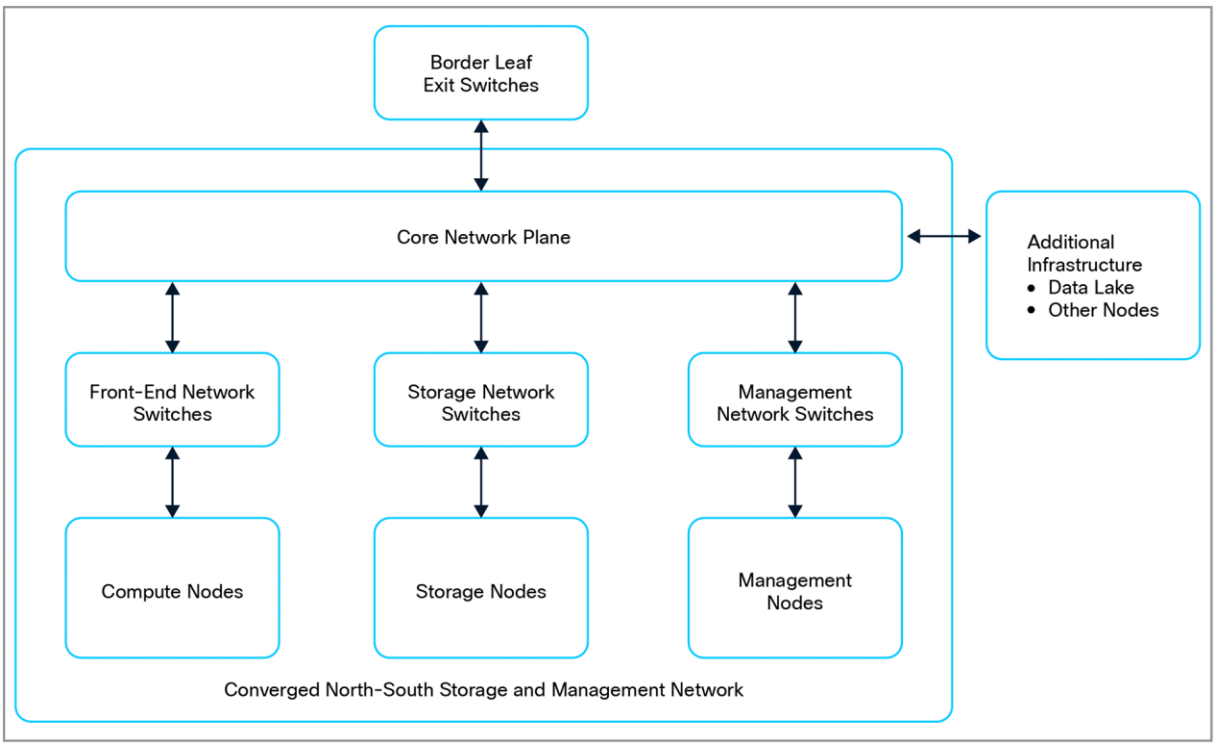

Figure 11.
Logical view of Converged North-South Storage and Management Network

Each compute node has 2 400G DPU ports that need to be connected to a redundant pair of HF6100-64ED leaf switches. With 64 compute nodes, there are a total of 128 400G ports evenly split across two different leaf switches. These 64 compute nodes connected to a pair of leaf switches form a Lego building where each leaf switch has 64 400G downlinks and 32 uplinks – the 64 400G uplinks between two leaf switches provide a network bandwidth of 50gbps per GPU within the 64 compute nodes consisting of a total of 512 GPUs. Each compute node has a 1G port to allow access to BMC and 2 10G ports for host management.

Each storage server has 2 dual 200GE port NICs for a total of 4 200GE ports. The 2 200GE port per NIC needs to be connected to a redundant pair of HF6100-32D storage leaf switches. A group of 16 storage servers connected to two storage leaf switches together form a Lego building block where each leaf switch has 32 200GE downlinks and 16 400GE uplinks. Enough network bandwidth to storage is provisioned ensuring 12.5gbps per GPU. Each storage server also requires connecting a 1GE port to server BMC and a 10GE port for server host management.

Each management node has a dual 200GE port NIC connecting to a redundant pair of HF6100-32D management leaf switches. About five management nodes are provisioned for every 32 compute nodes of 256 GPUs. Additional eight management nodes are allocated for overall cluster management. The number of management nodes can be adjusted as per tenant and cloud provider requirements. The reference design uses a maximum of 48 management nodes per management leaf pair as a Lego block with additional blocks added with increase in the number of compute nodes beyond 256. Each management node also requires connecting a 1GE port to server BMC and a 10GE port for server host management.

The physical design of the converged network uses a two-tier topology till 1024 compute node (8192 GPUs) and a three-tier topology beyond that scale.

Two-tier Topology

The two-tier design uses a two-stage leaf-spine Clos topology with the spine layer shared between compute, storage, and management layers and also interconnects to the border leaf switches. The entire topology is built using the Lego block concept and sizing rules described previously where the number of storage and management nodes are adjusted as per the number of compute nodes (GPUs) deployed in the cluster.

As shown in Figure 12, a 512-compute node (4096 GPUs) cluster requires 8 HF6100-64ED spine switches, 16 HF6100-64ED front-end (FE) leaf switches. Each FE leaf uses 64 400GE downlink ports and 32 400GE uplink ports, with a total of 512 400GE across all FE leaf switches providing 50gbps of network bandwidth per GPU. This is paired with eight pairs of storage leaf switches each with 16 400GE uplinks, for a total of 256 400GE uplinks over 16 storage leaf switches, providing 12.5gbps per GPU of network bandwidth to storage with additional 50% bandwidth reserved for inter storage node communication. A total of 96 management nodes with two pairs of management leaf switches each with 8 400GE uplinks, one to each spine, are provisioned. The 1GE BMC and 10GE host management ports from compute, storage, and management nodes are connected to the OOB management network.

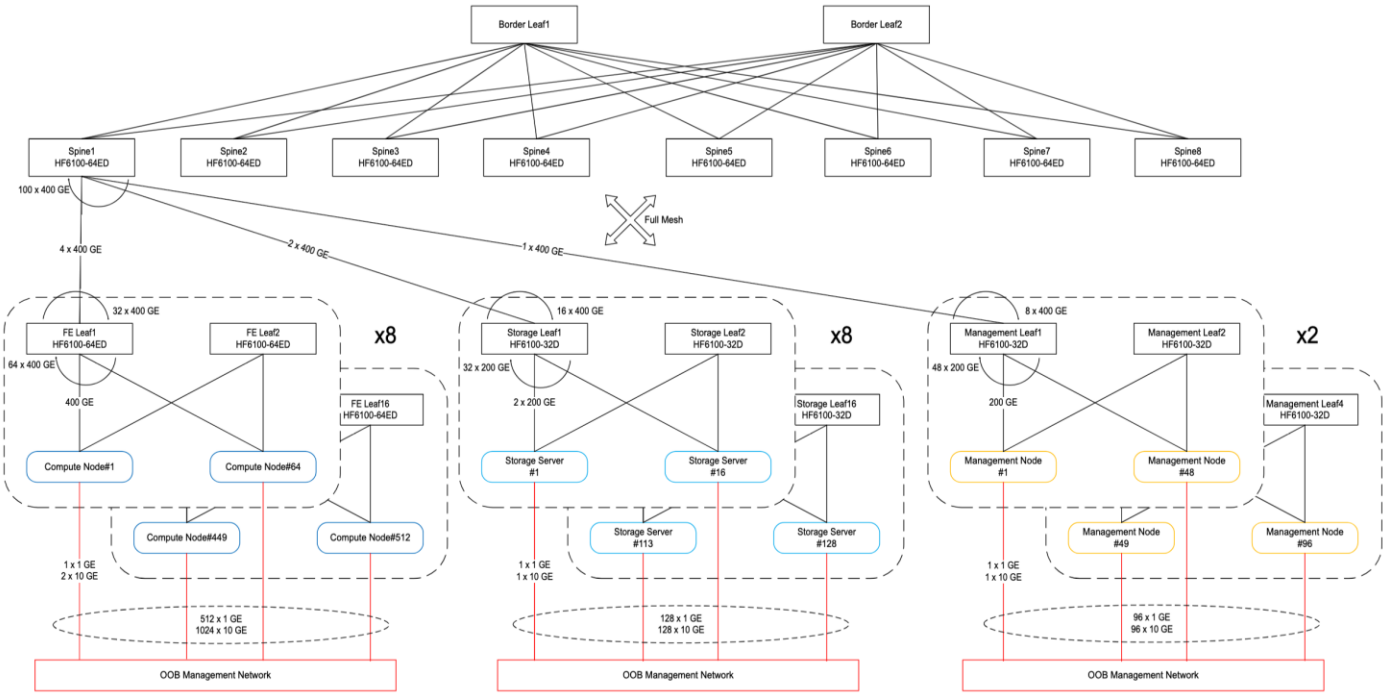

Figure 12.
Converged North-South Storage and Management Network for 512 HGX™ B300 nodes (4K GPUs)

As shown in Figure 13, a 1024-compute node (8192 GPUs) cluster requires 16 HF6100-64ED spine switches, 32 front-end (FE) leaf switches. Each FE leaf uses 64 400GE downlink ports and 32 400GE uplink ports, with a total of 1024 400GE across all FE leaf switches providing 50gbps of network bandwidth per GPU. This is paired with 16 pairs of storage leaf switches each with 16 400GE uplinks, for a total of 512 400GE uplinks over 32 storage leaf switches, providing 12.5gbps per GPU of network bandwidth to storage with additional 50% bandwidth reserved for inter storage node communication. A total of 192 management nodes with four pairs of management leaf switches each with 16 200GE uplinks, one to each spine, are provisioned. The 1GE BMC and 10GE host management ports from compute, storage, and management nodes are connected to the OOB management network.

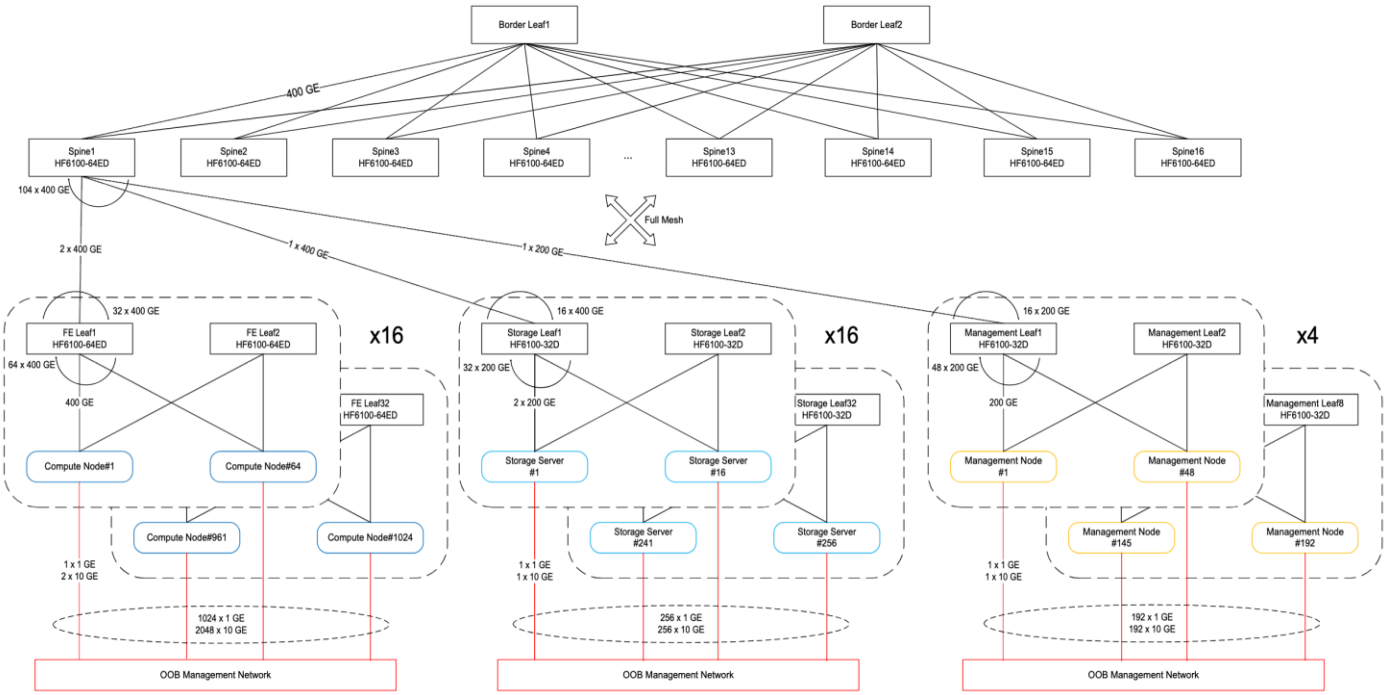

Figure 13.
Converged North-South Storage and Management Network for 1024 HGX™ B300 nodes (8K GPUs)

Three-tier Topology

Beyond 1024 compute nodes (8192 GPUs), a three-stage Clos topology is used with the third stage represented by four core groups of switches also known as core fabric. The compute, storage and management groups each have their own dedicated spine layer connecting to the core fabric. The HF6100-64ED switches are used in all spine layers and core fabric.

For the compute layer, a larger Lego super building block is used with 4 smaller compute Lego blocks described before with each of them interconnected by 4 FE spine switches – together this bigger Lego block consists of 256 compute nodes (2048 GPUs), 4 pairs of FE leaf switches (total of 8), and 4 FE spine switches. Each of the 4 FE spine switches have 64 400GE downlinks and their uplinks connect to their respective core group via 16 400 GE ports for a total of 64 400GE uplinks across all 4 FE spine switches providing 12.5gbps of network bandwidth per GPU.

Similar to compute layer, the storage layer uses the same Lego block concept described before with storage leaf switches in each block connecting to all four-storage spine group of switches. Each storage spine connects to its respective core group via 64 400GE uplinks. The management layer uses a similar approach with uplinks of management spine switches connecting to a pair of core groups.

As shown in Figure 14, a 2048-compute node (16384 GPUs) cluster requires 4 switches per core group and 2 switches per storage spine group. The aggregate storage spine to core group uses 512 400GE ports providing network bandwidth of 12.5gbps per GPU. Overall, there are 64 FE leaf, 32 FE spine, 8 storage spine, 32 storage leaf, 2 management spine, 16 management leaf, and 16 core group switches. The 1GE BMC and 10GE host management ports from compute, storage, and management nodes are connected to the OOB management network.

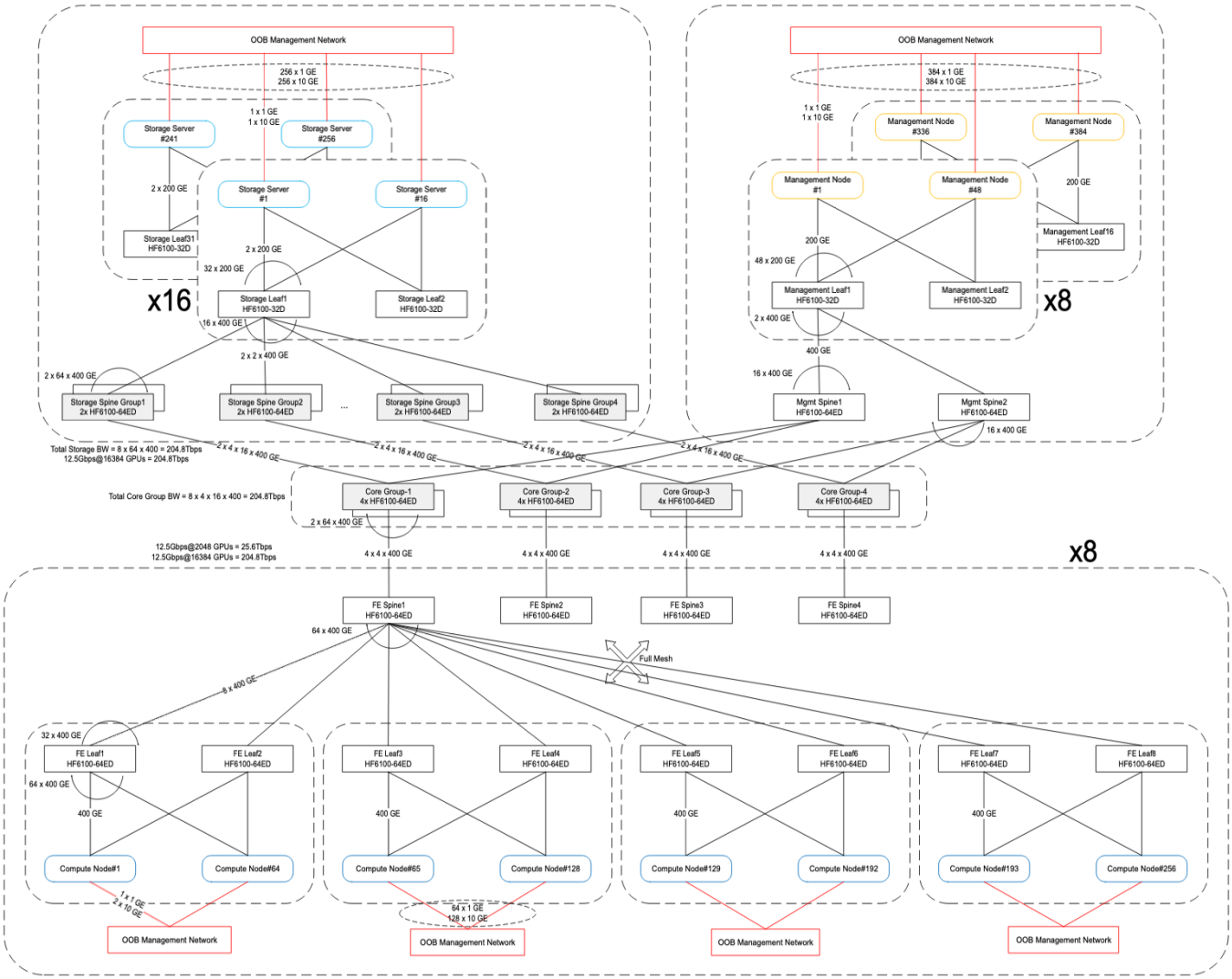

Figure 14.
Converged North-South Storage and Management Network for 2048 HGX™ B300 nodes (16K GPUs)

As shown in Figure 15, a 4096-compute node (32768 GPUs) cluster requires 8 switches per core group and 4 switches per storage spine group. The aggregate storage spine to core group uses 1024 400GE ports providing network bandwidth of 12.5gbps per GPU. Overall, there are 128 FE leaf, 64 FE spine, 16 storage spine, 64 storage leaf, 2 management spine, 32 management leaf, and 32 core group switches. The 1GE BMC and 10GE host management ports from compute, storage, and management nodes are connected to the OOB management network.

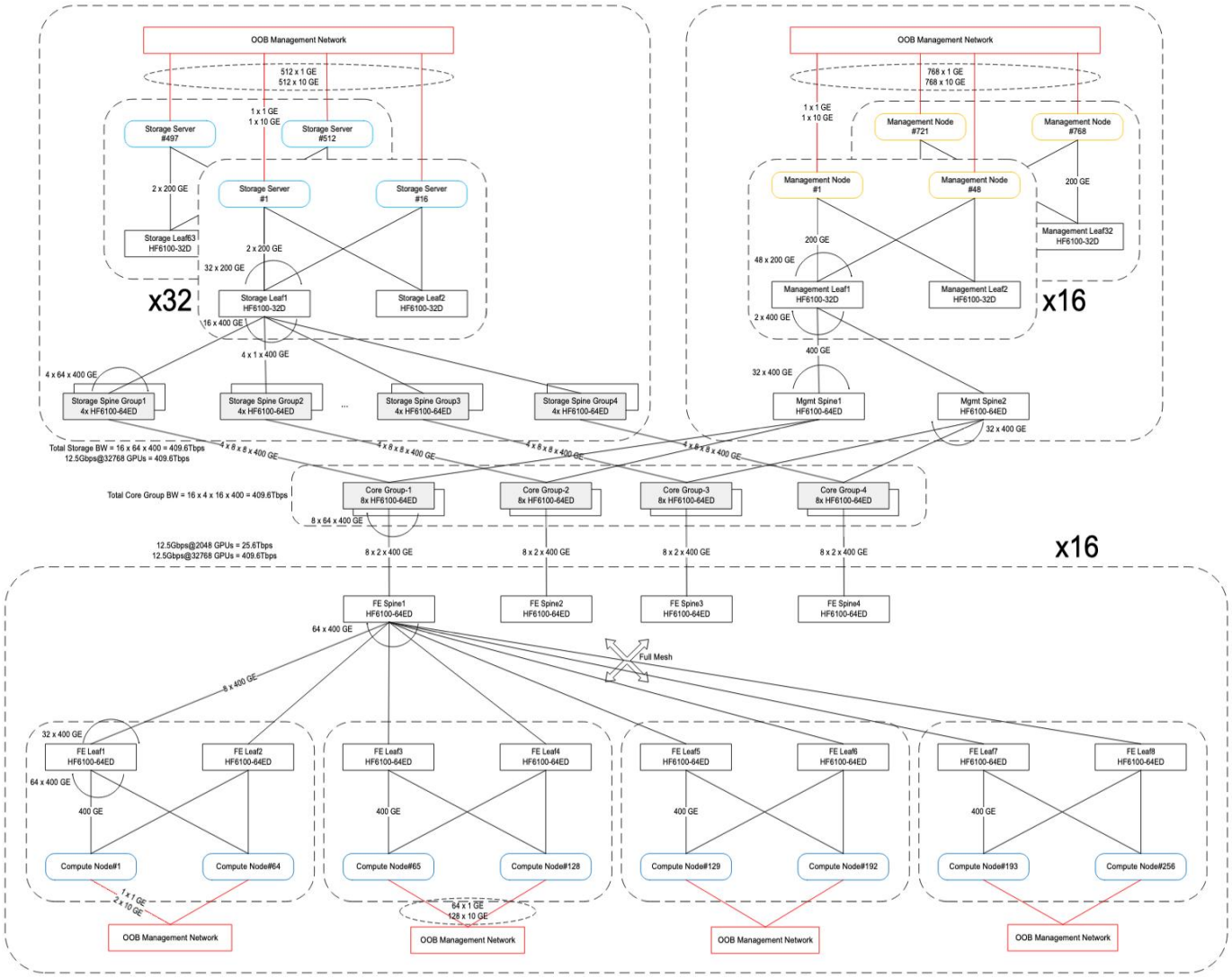

Figure 15.
Converged North-South Storage and Management Network for 4096 HGX™ B300 nodes (32K GPUs)

Out-of-Band management network

The OOB management network is used to connect the:

- 1G BMC port of NVIDIA Bluefield®-3 DPUs
- 1G BMC and 10G host ports of compute, storage, and management nodes - these can alternatively be part of converged network as per CSPs discretion
- 1G management port of Power Distribution Units (PDUs)
- 1G management port of Terminal servers used for equipment console connectivity

This network is not exposed to the tenants. However, it is made accessible to controllers for provisioning, observability, and overall cluster management. Additionally, the 1G management ports of switches need to be connected to a separate switch management network and made accessible to the cloud network controller to allow configuration and monitoring. Switch HF6100-60L4D is used as OOB management leaf with 100GE uplinks using QSFP-100G-DR-S optics and HF6100-64ED as OOB management spine switch.

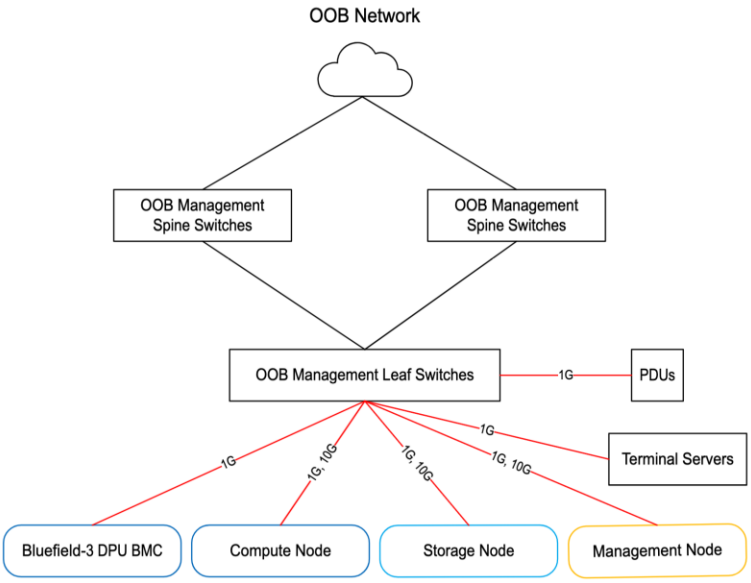

Figure 16.
OOB Management Network

Cluster BOM

Table 4 shows the Bill-of-Materials (BOM) for building clusters of 4K and 8K GPUs using two-tier compute network.

Table 4. Minimum BOM for clusters with 4K and 8K GPUs using two-tier compute network

PID	Description	4K GPUs	8K GPUs
UCSC-880A-M8-B306 **SYS-822GS-NB3RT** **AS-8126GS-NB3RT** **SYS-422GS-NB3RT-ALC**	Cisco, Supermicro HGX B300 air and liquid cooled rack server	512	1024
HF6100-64ED **(Converged network only use case)**	Cisco Hyperfabric switch, 64x800Gbps OSFP	24	48
N9164E-NS4-O **(East-West compute network only use case)**	Cisco N9000 switch, 64x800Gbps OSFP	192	384
HF6100-64ED **(Both East-West compute & Converged network use case)**	Cisco Hyperfabric switch, 64x800Gbps OSFP	216	432
HF6100-32D **(Storage and Management node leaf)**	Cisco Hyperfabric switch, 32x400Gbps QSFP-DD	20	40
HF6100-60L4D **(OOB Management leaf)**	Cisco Hyperfabric switch 60x50G SFP28 4x400G QSFP-DD	51	101

PID	Description	4K GPUs	8K GPUs
HF6100-64ED (OOB Management spine)	Cisco Hyperfabric switch, 64x800Gbps OSFP	2	2
OSFPR-800G-DR8	800G OSFP transceiver, 800GBASE-DR8, SMF dual MPO-12 APC, 500m (riding heat sink)	4096	8192
OSFP-800G-DR8	800G OSFP transceiver, 800GBASE-DR8, SMF dual MPO-12 APC, 500m (integrated heat sink)	13470	26970
QDD-400G-DR4	400G QSFP-DD transceiver, 400GBASE-DR4, MPO-12, 500m parallel	288	576
QSFP-400G-DR4	400G QSFP112 transceiver, 400GBASE-DR4, MPO-12, 500m parallel	1024	2048
QDD-400G-SR8-S	400G QSFP-DD transceiver, 400GBASE-SR8, MPO-16 APC, 100m	352	704
QSFP-200G-SR4-S	200G QSFP transceiver, 200GBASE-SR4, MPO-12, 100m	704	1408
QSFP-100G-DR-S	100G QSFP transceiver, 100GBASE-DR, LC, 500m	102	202
SFP-1G-T-X	1G SFP	1760	3520
SFP-10G-T-X	10G SFP	1248	2496
CB-M12-M12-SMF	MPO-12 cables	18208	36480
CB-M16-M12-MMF	MPO-16 to dual MPO-12 breakout cables	352	704
CB-M12-4LC-SMF	MPO-12 to 4x Duplex LC, SMF	26	52
CAT6A	Copper cable for 10G	1248	2496
CAT5E	Copper cable for 1G	1760	3520
UCSC-C225-M8N (storage server)	Cisco UCS C225-M8 1RU Rack Server	Min: 42 Max: 128	Min: 84 Max: 256
UCSC-C240-M8 UCSC-C245-M8SX (management node)	Cisco UCS C240-M8 2RU Rack Server Cisco UCS C245-M8 2RU Rack Server	96	192

Table 5 shows the Bill-of-Materials (BOM) for building clusters of different sizes, from 4K to 32K GPUs, with three-tier compute network.

Table 5. Minimum BOM for clusters with 4K to 32K GPUs using three-tier compute network

PID	Description	4K GPUs	8K GPUs	16K GPUs	32K GPUs
UCSC-880A-M8-B306 **SYS-822GS-NB3RT** **AS-8126GS-NB3RT** **SYS-422GS-NB3RT-ALC**	Cisco, Supermicro HGX B300 air and liquid cooled rack server	512	1024	2048	4096
HF6100-64ED (Converged network only use case)	Cisco Hyperfabric switch, 64x800Gbps OSFP	24	48	122	242
N9164E-NS4-O (East-West compute network only use case)	Cisco N9000 switch, 64x800Gbps OSFP	320	640	1280	2560
HF6100-64ED (Both East-West compute & Converged network use case)	Cisco Hyperfabric switch, 64x800Gbps OSFP	344	688	1402	2802
HF6100-32D (Storage and Management node leaf)	Cisco Hyperfabric switch, 32x400Gbps QSFP-DD	20	40	48	96
HF6100-60L4D (OOB Management leaf)	Cisco Hyperfabric switch 60x50G SFP28 4x400G QSFP-DD	51	101	192	384
HF6100-64ED (OOB Management spine)	Cisco Hyperfabric switch, 64x800Gbps OSFP	2	2	2	2
OSFPR-800G-DR8	800G OSFP transceiver, 800GBASE-DR8, SMF dual MPO-12 APC, 500m (riding heat sink)	4096	8192	16384	32768
OSFP-800G-DR8	800G OSFP transceiver, 800GBASE-DR8, SMF dual MPO-12 APC, 500m (integrated heat sink)	21648	43328	87344	174688
QDD-400G-DR4	400G QSFP-DD transceiver, 400GBASE-DR4, MPO-12, 500m parallel	288	576	544	1088
QSFP-400G-DR4	400G QSFP112 transceiver, 400GBASE-DR4, MPO-12, 500m parallel	1024	2048	4096	8192
QDD-400G-SR8-S	400G QSFP-DD transceiver, 400GBASE-SR8, MPO-16 APC, 100m	352	704	896	1792
QSFP-200G-SR4-S	200G QSFP transceiver, 200GBASE-SR4, MPO-12, 100m	704	1408	1792	3584
QSFP-100G-DR-S	100G QSFP transceiver, 100GBASE-DR, LC, 500m	102	202	384	768
SFP-1G-T-X	1G SFP	1760	3520	6784	13568
SFP-10G-T-X	10G SFP	1248	2496	4736	9472
CB-M12-M12-SMF	MPO-12 cables	26400	52864	106048	212096

PID	Description	4K GPUs	8K GPUs	16K GPUs	32K GPUs
CB-M16-M12-MMF	MPO-16 to dual MPO-12 breakout cables	352	704	896	1792
CB-M12-4LC-SMF	MPO-12 to 4x Duplex LC, SMF	26	51	96	192
CAT6A	Copper cable for 10G	1248	2496	4736	9472
CAT5E	Copper cable for 1G	1760	3520	6784	13568
UCSC-C225-M8N (storage server)	Cisco UCS C225-M8 1RU Rack Server	Min: 42 Max: 128	Min: 84 Max: 256	Min: 168 Max: 256	Min: 336 Max: 512
UCSC-C240-M8 UCSC-C245-M8SX (management node)	Cisco UCS C240-M8 2RU Rack Server Cisco UCS C245-M8 2RU Rack Server	96	192	384	768

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 computes 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 “High-Performance Storage” 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, and the networking feature sets enabled on them would vary as per customer use case and are beyond the scope of this RA.

High-Performance Storage

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 "Disaggregated 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 17 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.

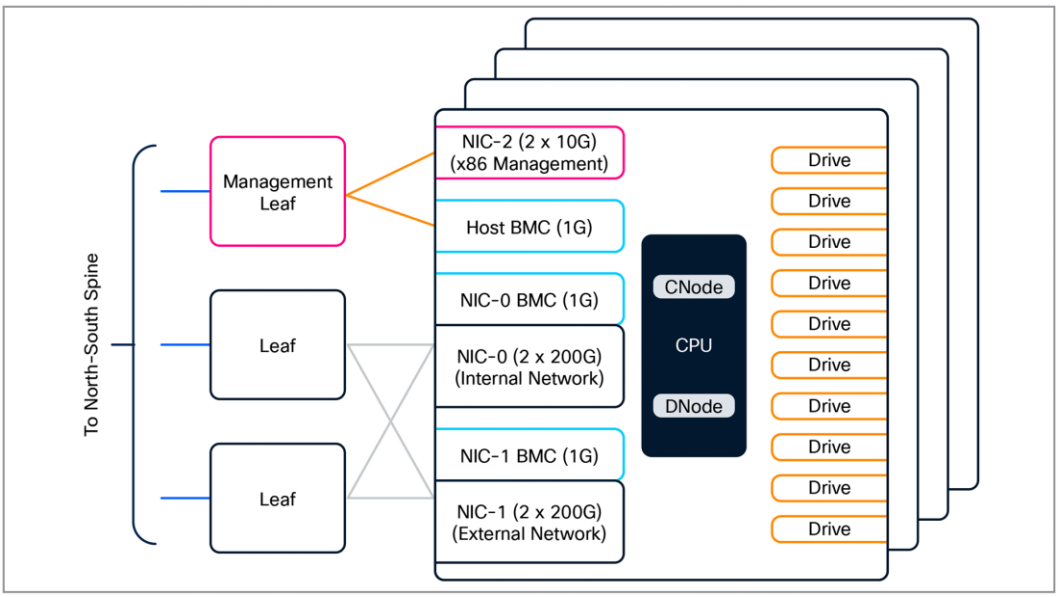

Figure 17.
Cisco EBox Storage Logical Block Diagram

Beside the Cisco EBox, this RA will support all NVIDIA-Certified high-performance storage solutions certified till the NCP level and supporting multi-tenancy.

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 Controller

Cisco Nexus® Hyperfabric controller is required to provision and manage the entire networking fabric including provisioning for tenants. It fully manages the configuration target state, switch software versions etc. It ingests telemetry from switches and NICs on compute and storage nodes for end-to-end network visibility and optimization of network performance. CSPs can also utilize the available programmability interfaces to fully integrate with the controller via their automation frameworks.

Compute Controller

Cisco Intersight is used to do provisioning of the Cisco UCS C880A 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-premises 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 C880A 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 HF6100-64ED switches used in this RA when connected to NVIDIA ConnectX®-8 and enabled with Fine Grain Load Balancing (FGLB) license. The Spectrum-X license is not required when deploying East-West compute network with the use of N9164E-NS4-O switch.

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

Area	Details
Compute + Memory	2x 6th Gen Intel Xeon or AMD Turin CPUs each with 64 cores 32x 128GB DDR5 RDIMMs, up to 6,000 MT/S (max supported memory config)
Storage	2x 960GB M.2 SATA or NVMe boot drives with HW RAID controller Up to 8 PCIe Gen 5 x4 E1.S NVMe SSDs
GPUs	8x NVIDIA B300 GPUs with 8x ConnectX-8 (OSFP based) integrated on the board
Network Cards	1x or 2x PCIe x16 FHHL NVIDIA BlueField®-3 B3240 crypto enabled North-South NIC 1 OCP 3.0 X710-T2L for host management