

Cisco N9000 Enterprise RA

Full Stack AI Infrastructure compliant with NVIDIA
HGX B300 Enterprise Reference Architecture

Contents

Introduction	3
Hardware	4
HGX B300 Rack Server	4
Cisco N9364E-SG2-O	4
Cisco N9164E-NS4-O	5
Cisco Nexus 9332D-GX2B	5
Cisco Nexus 93108TC-FX3	5
Cisco UCS C225 M8 Rack Server	5
Cisco Optics and cables	6
Networking topologies.....	6
Overview	6
Compute (Node East-West) Network	6
Converged (Node North-South) Network	7
Out-of-Band (OOB) Management Network	8
Topology with 16 compute nodes (128 GPUs)	8
Topology with 32 compute nodes (256 GPUs)	9
Topology with 64 compute nodes (512 GPUs)	10
Topology with 128 compute nodes (1024 GPUs)	11
Cluster BOM	12
High-Performance Storage	14
Software	15
Network controller	15
Compute controller	15
Storage controller	15
NVIDIA AI Enterprise	16
NVIDIA Spectrum-X	16
Security.....	17
Observability	17
Testing and certification.....	17
Summary.....	17
Appendix A – Compute server specifications	18
Appendix B – Control-node server specifications.....	18

Introduction

Cisco N9000® Enterprise Reference Architecture (RA) is a blueprint of an on-premises AI cluster that is managed by the on-premises Cisco [Nexus Dashboard](#) platform. It empowers and simplifies your AI initiatives and accelerates AI deployments with a comprehensive, integrated, on-premises managed solution. This reference architecture (RA) adheres to the NVIDIA Enterprise Reference Architecture (Enterprise RA) for NVIDIA HGX™ B300 with up to 1024 GPU scale.

Cisco [N9000](#) Series Switches provide high-speed, deterministic, low-latency, and power-efficient connectivity for AI and high-performance computing (HPC) workloads. With the availability of multiple form-factors, optics, and rich software features of the Cisco NX-OS operating system, N9000 switches provide a consistent experience for frontend, storage, backend, and out-of-band (OOB) management networks (see Figure 1).

Cisco Nexus Dashboard is the operations and automation platform for managing the N9000 switch-based fabrics. It complements the data plane features of the N9000 switches by simplifying their configuration using built-in templates. It detects network health issues, such as congestion, bit errors, and traffic bursts, in real time and automatically flags them as anomalies. These issues can be resolved faster using integrations with commonly used tools, such as ServiceNow and Ansible, allowing the networks of an AI cluster to be aligned with the existing workflows of an organization.

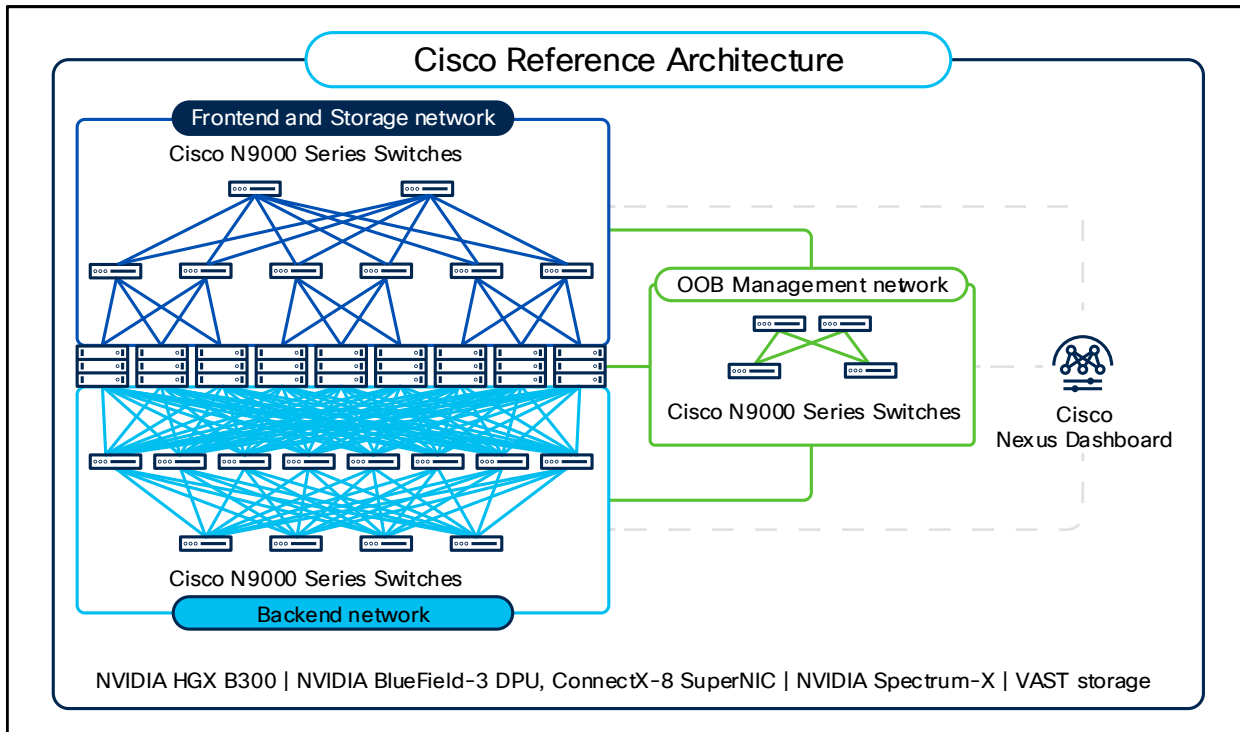


Figure 1. Cisco N9000 Series Switches for networking the AI clusters, managed by Nexus Dashboard Platform

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 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 N9364E-SG2-O

The Cisco N9364E-SG2-O is a Silicon One® Ethernet switch ASIC based 2RU switch supporting 64 800G OSFP modules allowing 64 800GE or 128 400GE ports. This switch will be used in both leaf and spine role.



Figure 3.
Cisco N9364E-SG2-O 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 Cisco Nexus 9364E-SG2-O switch.

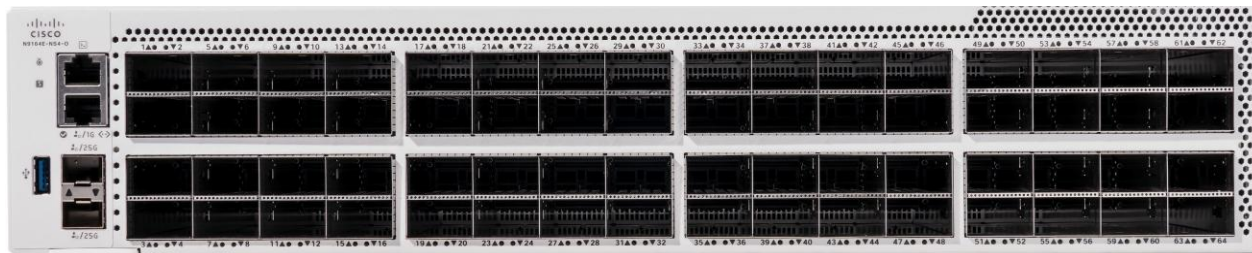


Figure 4.
Cisco Nexus N9164E-NS4-O switch

Cisco Nexus 9332D-GX2B

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



Figure 5.
Cisco Nexus 9332D-GX2B switch

Cisco Nexus 93108TC-FX3

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



Figure 6.
Cisco Nexus 93108TC-FX3 switch

Cisco UCS C225 M8 Rack Server

The Cisco C225 M8 Rack Server is a 1RU general-purpose server that can be used in many roles, such as application server, support server, control nodes for Kubernetes (K8s) and Slurm. In this RA, these servers are also used to run the VAST storage solution as described in the “High-Performance Storage” section.



Figure 7.
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 in various devices

Device	Optics and cable
B3220, B3240	QSFP-400G-DR4 with SMF MPO-12 cable
B3220L	QSFP-200G-SR4 with MMF MPO-12 cable
ConnectX-8	OSFP-800G-DR8 with dual CB-M12-M12-SMF cable
N9364E-SG2-O N9164E-NS4-O	OSFP-800G-DR8 with dual SMF MPO-12 cable
N9K-C9332D-GX2B	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
N9K-93108TC-FX3	QSFP-100G-DR-S with CB-LC-LC-SMF, CB-M12-4LC-SMF cable CAT5E cable CAT6A cable

Networking topologies

Overview

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

- Compute (Node East-West) Network
- Converged (Node North-South) Network
- Out-of-Band (OOB) Management Network

Compute (Node East-West) Network

This network is meant for communication between the GPUs via ConnectX-8 SuperNICs configured in 2 x 400G mode. Both single and dual plane fabrics are supported though dual plane fabric is recommended to avoid a single point of failure. In case of single plane, only 1 400G port per ConnectX-8 SuperNICs is used. However, in case of dual planes, 1 400G port per ConnectX-8 SuperNICs is connected to each of the two planes. Table 2, 3 shows the quantity of different units required for building a single and dual plane compute network respectively considering different cluster sizes.

Table 2. Single plane east-west compute fabric table – switch, transceivers, and cable counts

Compute counts		Switch counts		Transceiver counts			Cable counts	
Nodes	GPUs	Leaf	Spine	Node to leaf		Switch to switch (800G)	Node to leaf	Switch to switch
				Node (800G)	Leaf (800G)			
16	128	2	N/A	128	64	64	128	64
32	256	4	2	256	128	256	256	256
64	512	8	4	512	256	512	512	512
128	1024	16	8	1024	512	1024	1024	1024

Table 3. Dual plane east-west compute fabric table – switch, transceivers, and cable counts

Compute counts		Switch counts		Transceiver counts			Cable counts	
Nodes	GPUs	Leaf	Spine	Node to leaf		Switch to switch (800G)	Node to leaf	Switch to switch
				Node (800G)	Leaf (800G)			
16	128	4	N/A	128	64	128	256	128
32	256	8	4	256	128	512	512	512
64	512	16	8	512	256	1024	1024	1024
128	1024	32	16	1024	512	2048	2048	2048

Converged (Node North-South) Network

The converged north-south network is used for communication with compute, storage, in-band management, support, and end-customer connections. It connects compute nodes using 2 400G ports to two separate switches providing redundancy and high storage throughput. Enough storage facing ports are pre-allocated ensuring 12.5 gbps of network bandwidth per GPU. A total of 16 200G ports (4 800G ports) are also reserved for 8 support servers to allow running Slurm and Kubernetes control nodes, NVIDIA Base Command Manager head nodes, and additional control monitoring applications.

Table 4 shows the quantity of different units required for building a converged network considering different cluster sizes.

Table 4. Converged north-south fabric table – switch, transceivers, and 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		SMF MPO-12	CAT6A + CAT5E
						Node (400G)	Leaf (800G)	800G	Node	Leaf	Leaf (100G)	Spine (800G)	Leaf (400G)	Spine (800G)	Customer (800G)	Support (800G)		
16	128	2	N/A	3	2	32	16	16	N/A	N/A	8	2	8	4	4	4	72	80
32	256	2	N/A	4	2	64	32	32	N/A	N/A	16	4	8	4	8	4	128	160
64	512	4	2	8	2	128	64	128	N/A	N/A	32	4	16	8	16	4	312	320
128	1024	8	4	14	4	256	128	256	N/A	N/A	56	8	64	32	32	4	648	640

Out-of-Band (OOB) Management Network

The OOB Management network is primarily used for node management connecting to the 1G ports of Base Management Controller (BMC) of the servers and NVIDIA Bluefield[®]-3 DPUs, and 10G host ports of the servers. The 100G uplinks of OOB management leaf switches are connected to converged north-south spine switches.

Additionally, the 1G management ports of the networking switches need to be connected separately to allow for their configuration and monitoring.

Topology with 16 compute nodes (128 GPUs)

Figure 8 shows the overall cluster topology interconnecting 16 compute nodes with a total of 128 GPUs. Each plane of the compute network uses a pair of N9364E-SG2-O or N9164E-NS4-O switches with odd rail groups on switch#1 and even rail groups on switch#2. The 8 800G NVIDIA ConnectX-8 NICs on each server are each plugged with OSFPR-800G-DR8 modules allowing 2 400G ports per NIC where the two ports connect to two different planes. With 16 compute nodes, there are a total of 128 800G or 256 400G ports, connecting 128 400G ports to compute network plane1 and 128 400G ports to compute network plane2. Within a plane, 64 400G ports connect to switch#1 and 64 400G ports to switch#2.

The converged north-south network uses a pair of N9364E-SG2-O switches. The 2 400G NVIDIA Bluefield[®]-3 DPU ports per compute server connect to two different switches for redundancy. With 16 compute nodes, there are a total of 32 400G DPU ports evenly split between switch#1 and switch#2.

A total of 12 Cisco EBox nodes are used as high-performance storage connected to 2 N9332D-GX2B storage leaf switches for a total of 48 200G downlinks and 8 400G uplinks evenly split between the two leaf switches.

Every compute node uses 2 10G host management ports, 1 1G server BMC port, 1 or 2 1G DPU BMC ports depending upon if the server has 1 or 2 DPUs. Similarly, each Cisco EBox also requires 1 10G host management port and 1 1G server BMC port. All these ports from 16 compute nodes and 12 storage nodes are connected to 3 N93108TC-FX3 OOB management leaf switches.

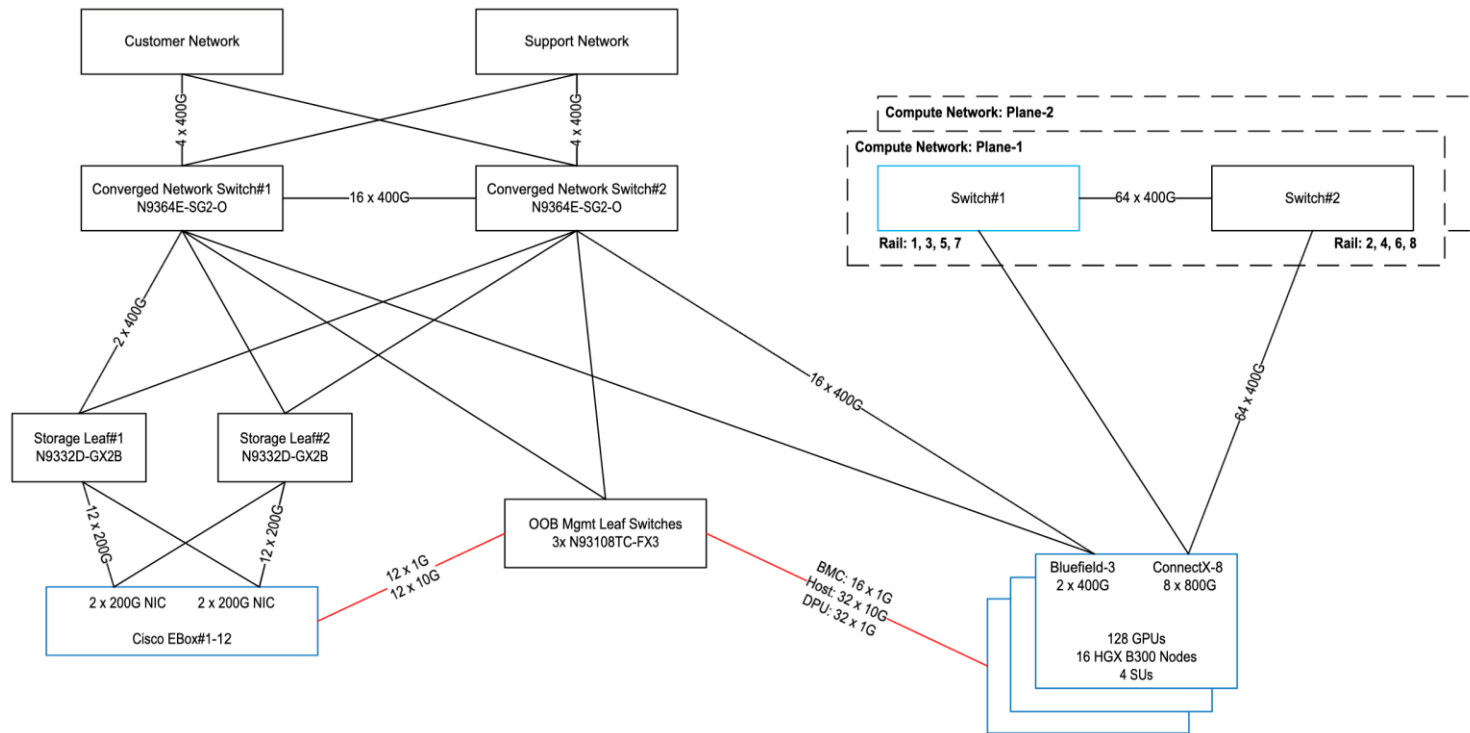


Figure 8.
Topology with 16 compute nodes (128 GPUs)

Topology with 32 compute nodes (256 GPUs)

Figure 9 shows the overall cluster topology interconnecting 32 compute nodes with a total of 256 GPUs. Each plane of the compute network uses 4 leaf switches and 2 spine switches. Both N9364E-SG2-O or N9164E-NS4-O switches can be used. The 4 leaf switches respectively map to rail group 1+5, 2+6, 3+7, 4+8. The 8 800G NVIDIA ConnectX-8 NICs on each server are each plugged with OSFPR-800G-DR8 modules allowing 2 400G ports per NIC where the two ports connect to two different planes. With 32 compute nodes, there are a total of 256 800G or 512 400G ports, connecting 256 400G ports to compute network plane1 and 256 400G ports to compute network plane2. Within a plane, 64 400G ports connect to every leaf switch.

The converged north-south network uses a pair of N9364E-SG2-O switches. The 2 400G NVIDIA Bluefield®-3 DPU ports per compute server connect to two different switches for redundancy. With 32 compute nodes, there are a total of 64 400G DPU ports evenly split between switch#1 and switch#2.

A total of 12 Cisco EBox nodes are used as high-performance storage connected to 2 N9332D-GX2B storage leaf switches for a total of 48 200G downlinks and 8 400G uplinks evenly split between the two leaf switches.

Every compute node uses 2 10G host management ports, 1 1G server BMC port, 1 or 2 1G DPU BMC ports depending upon if the server has 1 or 2 DPUs. Similarly, each Cisco EBox also requires 1 10G host management port and 1 1G server BMC port. All these ports from 32 compute nodes and 12 storage nodes are connected to 4 N93108TC-FX3 OOB management leaf switches.

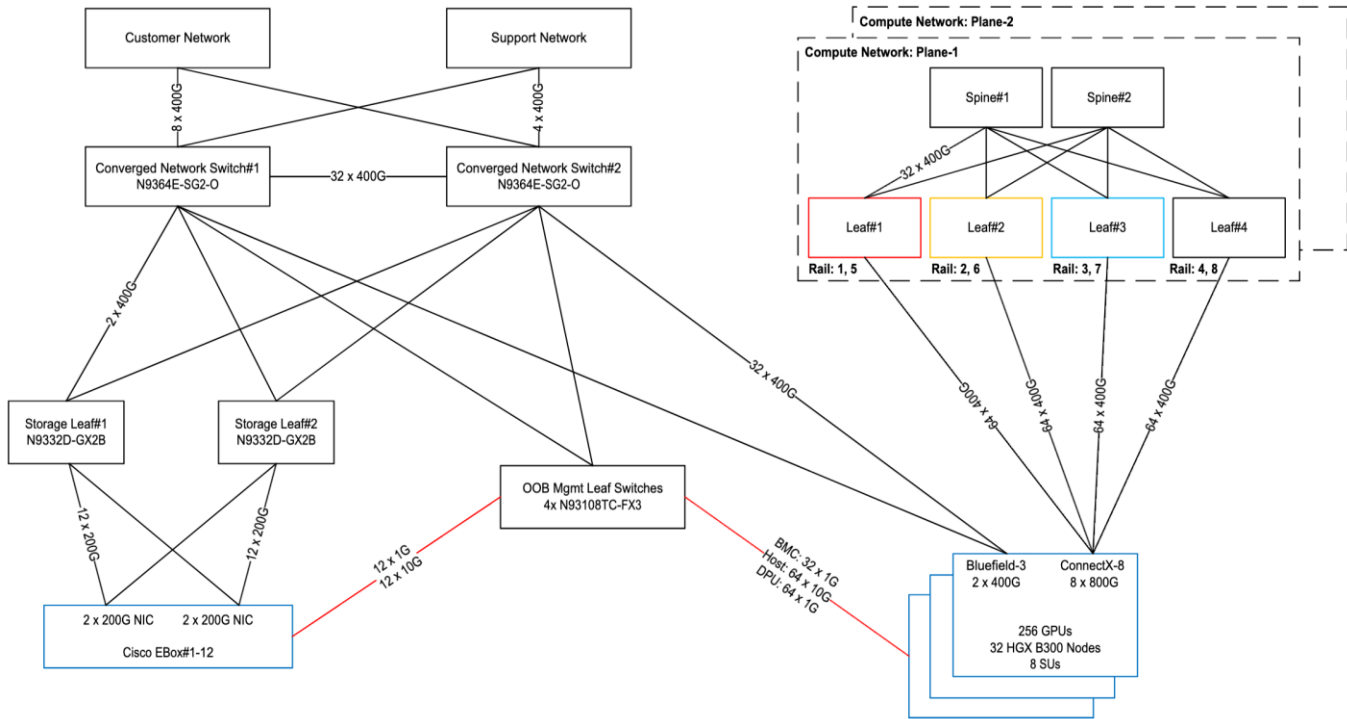


Figure 9.
Topology with 32 compute nodes (256 GPUs)

Topology with 64 compute nodes (512 GPUs)

Figure 10 shows the overall cluster topology interconnecting 64 compute nodes with a total of 512 GPUs. Each plane of the compute network uses 8 leaf and 4 spine switches. Both N9364E-SG2-O or N9164E-NS4-O switches can be used. The 4 pairs of leaf switches respectively map to rail group 1+5, 2+6, 3+7, 4+8. The 8 800G NVIDIA ConnectX-8 NICs on each server are each plugged with OSFPR-800G-DR8 modules allowing 2 400G ports per NIC where the two ports connect to two different planes. With 64 compute nodes, there are a total of 512 800G or 1024 400G ports, connecting 512 400G ports to compute network plane1 and 512 400G ports to compute network plane2. Within a plane, 64 400G ports connect to every leaf switch.

The converged north-south network uses 4 leaf and 2 spine N9364E-SG2-O switches. The 2 400G NVIDIA Bluefield®-3 DPU ports per compute server connect to two different switches for redundancy. With 64 compute nodes, there are a total of 128 400G DPU ports evenly split between 4 leaf switches (32 x 400G per leaf switch).

A total of 12 Cisco EBox nodes are used as high-performance storage connected to 2 N9332D-GX2B storage leaf switches for a total of 48 200G downlinks and 16 400G uplinks evenly split between the two leaf switches. If required, the number of Cisco EBox nodes can be expanded to 24 by using 2 x 200G breakout on storage leaf switch downlink ports.

Every compute node uses 2 10G host management ports, 1 1G server BMC port, 1 or 2 1G DPU BMC ports depending upon if the server has 1 or 2 DPUs. Similarly, each Cisco EBox also requires 1 10G host management port and 1 1G server BMC port. All these ports from 64 compute nodes and 12 storage nodes are connected to 8 N93108TC-FX3 OOB management leaf switches.

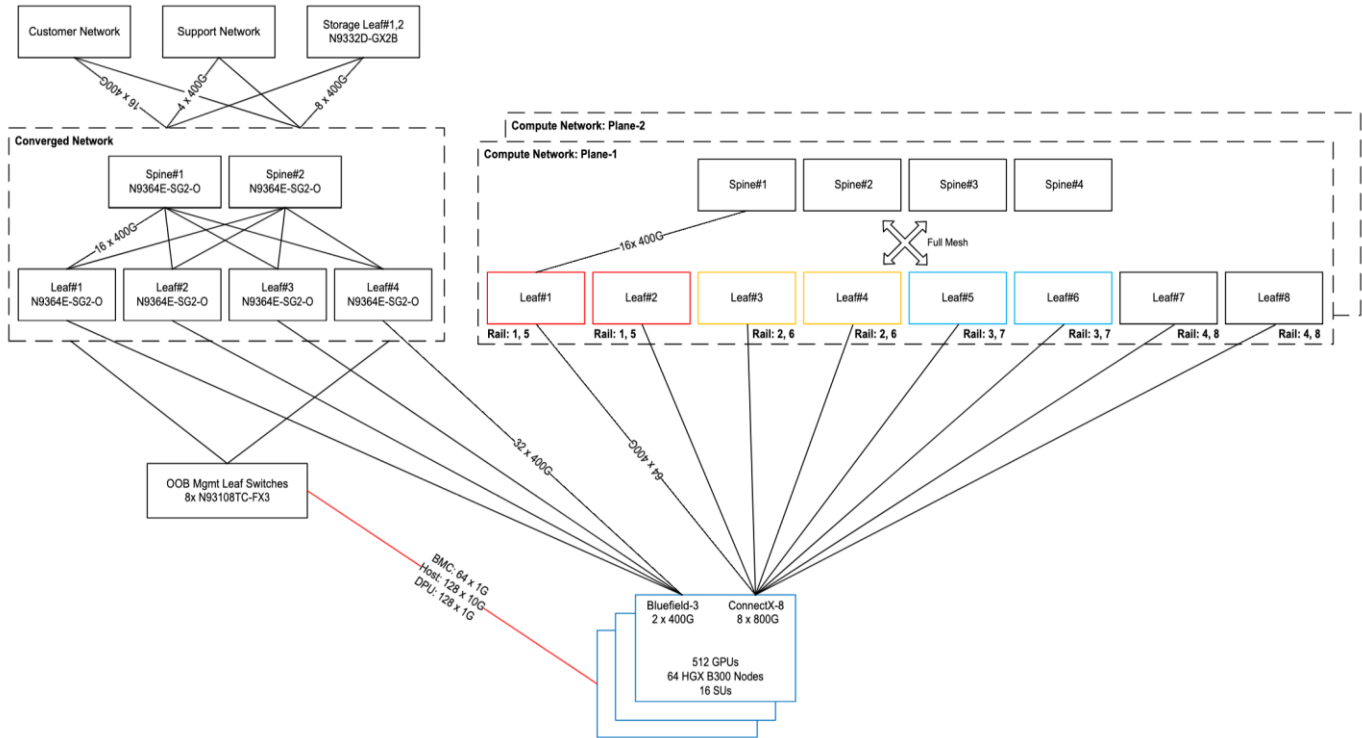


Figure 10.
Topology with 64 compute nodes (512 GPUs)

Topology with 128 compute nodes (1024 GPUs)

Cluster topology design interconnecting 128 compute nodes with a total of 1024 GPUs is very similar to the topology used for 64 compute nodes with both compute and converged network using leaf-spine architecture. Each plane of the compute network uses 16 leaf and 8 spine switches. Both N9364E-SG2-O or N9164E-NS4-O switches can be used. The 4 quad groups of leaf switches respectively map to rail group 1+5, 2+6, 3+7, 4+8. The 8 800G NVIDIA ConnectX-8 NICs on each server are each plugged with OSFPR-800G-DR8 modules allowing 2 400G ports per NIC where the two ports connect to two different planes. With 128 compute nodes, there are a total of 1024 800G or 2048 400G ports, connecting 1024 400G ports to compute network plane1 and 1024 400G ports to compute network plane2. Within a plane, 64 400G ports connect to every leaf switch.

The converged north-south network uses 8 leaf and 4 spine N9364E-SG2-O switches. The 2 400G NVIDIA Bluefield®-3 DPU ports per compute server connect to two different switches for redundancy. With 128 compute nodes, there are a total of 256 400G DPU ports evenly split between 8 leaf switches (32 x 400G per leaf switch).

A minimum of 12 Cisco EBox nodes are used as high-performance storage connected to 4 N9332D-GX2B storage leaf switches for a total of 48 200G downlinks and 64 400G uplinks evenly split between the two pairs of storage leaf switches. If required, the number of Cisco EBox nodes can be expanded to 32 by using 2 x 200G breakout on storage leaf switch downlink ports instead of native 200G ports.

Every compute node uses 2 10G host management ports, 1 1G server BMC port, 1 or 2 1G DPU BMC ports depending upon if the server has 1 or 2 DPUs. Similarly, each Cisco EBox also requires 1 10G host management port and 1 1G server BMC port. All these ports from 128 compute nodes and 12 storage nodes are connected to 14 N93108TC-FX3 OOB management leaf switches.

Cluster BOM

The BOM for clusters of different sizes using single compute plane is shown in Table 5.

Table 5. BOM of clusters with GPU scale 128 to 1024 using single compute plane

PID	Description	128 GPUs	256 GPUs	512 GPUs	1024 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	16	32	64	128
N9364E-SG2-O <small>(Converged Network)</small>	Cisco N9000 switch, 64x800Gbps OSFP	2	2	6	12
N9364E-SG2-O N9164E-NS4-O <small>(Compute Network)</small>	Cisco N9000 switch, 64x800Gbps OSFP	2	6	12	24
N9364E-SG2-O <small>(Both Converged and Compute Network)</small>	Cisco N9000 switch, 64x800Gbps OSFP	4	8	18	36
N9K-93108TC-FX3	Cisco Nexus switch, 48 1/10G BASE-T 6 QSFP28	2	4	8	14
N9K-C9332D-GX2B	Cisco Nexus switch, 32x400Gbps QSFP-DD	2	2	2	4
OSFP-800G-DR8	OSFP, 800GBASE-DR8, SMF dual MPO-12 APC, 500m (integrated heat sink)	174	466	992	1996
OSFPR-800G-DR8	OSFP, 800GBASE-DR8, SMF dual MPO-12 APC, 500m (riding heat sink)	128	256	512	1024
QDD-400G-DR4	400G QSFP-DD transceiver, 400GBASE-DR4, MPO-12, 500m parallel	8	8	16	64
QSFP-400G-DR4	400G QSFP112 transceiver, 400GBASE-DR4, MPO-12, 500m parallel	48	80	144	272
QSFP-200G-SR4-S	200G QSFP transceiver, 200GBASE-SR4, MPO-12, 100m	96	96	96	96
QSFP-100G-DR-S	100GBASE DR QSFP transceiver, 500m over SMF	8	16	32	56
CB-M12-4LC-SMF	Cable, MPO12-4X duplex LC, breakout cable, SMF, various lengths	2	4	8	14
CB-M12-M12-SMF	MPO-12 single mode cables	264	640	1336	2696
CB-M12-M12-MMF	MPO-12 multi-mode cables	48	48	48	48
CAT5E	Copper cable for 1G	60	108	204	396

PID	Description	128 GPUs	256 GPUs	512 GPUs	1024 GPUs
CAT6A	Copper cable for 10G	44	76	140	268
UCSC-C225-M8N (storage server)	Cisco UCS C225-M8 1RU Rack Server	12	12	Min: 12 Max: 24	Min: 12 Max: 32
UCSC-C240-M8	Cisco UCS C240-M8 2RU Rack Server	8	8	8	8
UCSC-C245-M8SX (support server)	Cisco UCS C245-M8 2RU Rack Server				

The BOM for clusters of different sizes using dual compute plane is shown in Table 6.

Table 6. BOM of clusters with GPU scale 128 to 1024 using dual compute plane

PID	Description	128 GPUs	256 GPUs	512 GPUs	1024 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	16	32	64	128
N9364E-SG2-O (Converged Network)	Cisco N9000 switch, 64x800Gbps OSFP	2	2	6	12
N9164E-NS4-O (Compute Network)	Cisco N9000 switch, 64x800Gbps OSFP	4	12	24	48
N9364E-SG2-O (Both Converged and Compute Network)	Cisco N9000 switch, 64x800Gbps OSFP	6	14	30	60
N9K-93108TC-FX3	Cisco Nexus switch, 48 1/10G BASE-T 6 QSFP28	2	4	8	14
N9K-C9332D-GX2B	Cisco Nexus switch, 32x400Gbps QSFP-DD	2	2	2	4
OSFP-800G-DR8	OSFP, 800GBASE-DR8, SMF dual MPO-12 APC, 500m (integrated heat sink)	238	722	1504	3020
OSFPR-800G-DR8	OSFP, 800GBASE-DR8, SMF dual MPO-12 APC, 500m (riding heat sink)	128	256	512	1024
QDD-400G-DR4	400G QSFP-DD transceiver, 400GBASE-DR4, MPO-12, 500m parallel	8	8	16	64
QSFP-400G-DR4	400G QSFP112 transceiver, 400GBASE-DR4, MPO-12, 500m parallel	48	80	144	272
QSFP-200G-SR4-S	200G QSFP transceiver, 200GBASE-SR4,	96	96	96	96

PID	Description	128 GPUs	256 GPUs	512 GPUs	1024 GPUs
	MPO-12, 100m				
QSFP-100G-DR-S	100GBASE DR QSFP transceiver, 500m over SMF	8	16	32	56
CB-M12-4LC-SMF	Cable, MPO12-4X duplex LC, breakout cable, SMF, various lengths	2	4	8	14
CB-M12-M12-SMF	MPO-12 single mode cables	456	1152	2360	4744
CB-M12-M12-MMF	MPO-12 multi-mode cables	48	48	48	48
CAT5E	Copper cable for 1G	60	108	204	396
CAT6A	Copper cable for 10G	44	76	140	268
UCSC-C225-M8N (storage server)	Cisco UCS C225-M8 1RU Rack Server	12	12	Min: 12 Max: 24	Min: 12 Max: 32
UCSC-C240-M8	Cisco UCS C240-M8 2RU Rack Server	8	8	8	8
UCSC-C245-M8SX (support server)	Cisco UCS C245-M8 2RU Rack Server				

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 NVIDIA and Cisco Enterprise 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. 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 13 shows the overall network connectivity and BOM with 12 storage servers. For data path, each server uses two NVIDIA BlueField-3 B3220L 2x200G NICs: NIC0 is used for internal network within the servers, allowing any server to access storage drives from any other server, and NIC1 is used for the external network, supporting client traffic such as NFS, S3, and SMB. The 1G BMC and 10G x86 management ports are connected to a management leaf switch.

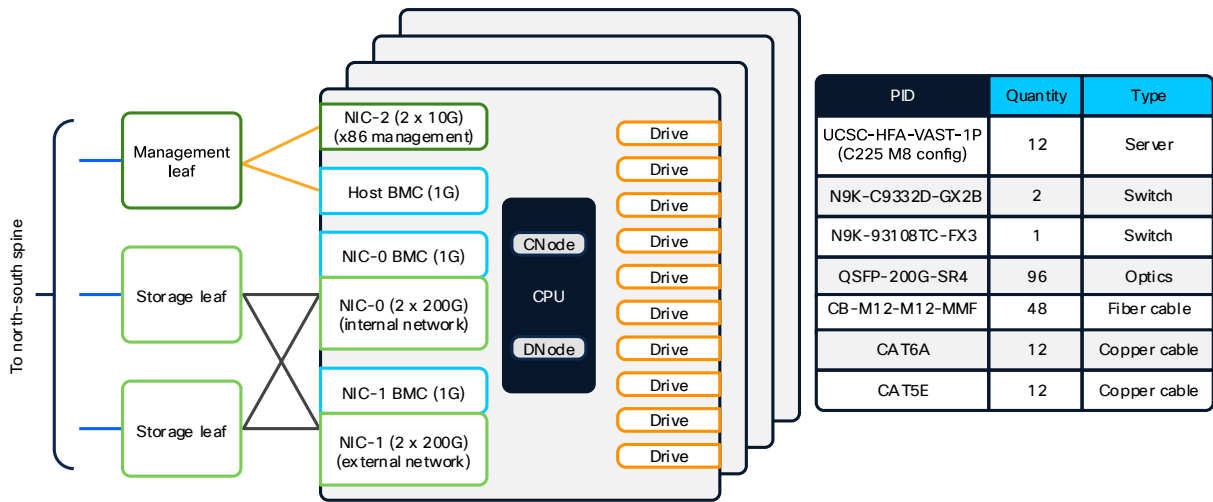


Figure 11.
Block diagram and BOM of storage subsystem

Beside Cisco EBox, other NVIDIA-Certified storage partners can also be used in this reference architecture.

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 Dashboard](#) can be used to provision and manage the entire networking fabric. It offers a unified platform that integrates key services – Insights (visibility and telemetry), Orchestrator (orchestration), and Fabric Controller (automation) – to deliver comprehensive network visibility, automation, and operational simplicity. Enterprises can also manage the switches and overall networking via open-source tools such as Ansible, Chef, and Puppet, integrating them with available [programmability interfaces](#).

Compute controller

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

Storage controller

The Cisco EBox storage controller (also known as VAST Management Service) will be used for configuration and monitoring of the high-performance storage. RESTful APIs are supported for integration with automation frameworks.

NVIDIA AI Enterprise

This reference architecture includes NVIDIA AI Enterprise, deployed and supported on NVIDIA-certified Cisco UCS C880A M8 and Supermicro NB3RT 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 accelerate models, improve TCO, and make production deployment faster.

NVIDIA Spectrum-X

The NVIDIA Spectrum-X Networking technology significantly improves the performance and efficiency of Ethernet-based GPU and storage networks. Its benefits are available with Cisco N9000 switches running NX-OS 10.6(1)F onwards when connected to NVIDIA ConnectX-8 and BlueField-3 SuperNICs. Fine Grain Load Balancing (FGLB) license is needed in addition to DCN Advantage License to enable NVIDIA Spectrum-X on Cisco N9000 switches. The Spectrum-X license is not required when deploying East-West compute network with the use of N9164E-NS4-O switch.

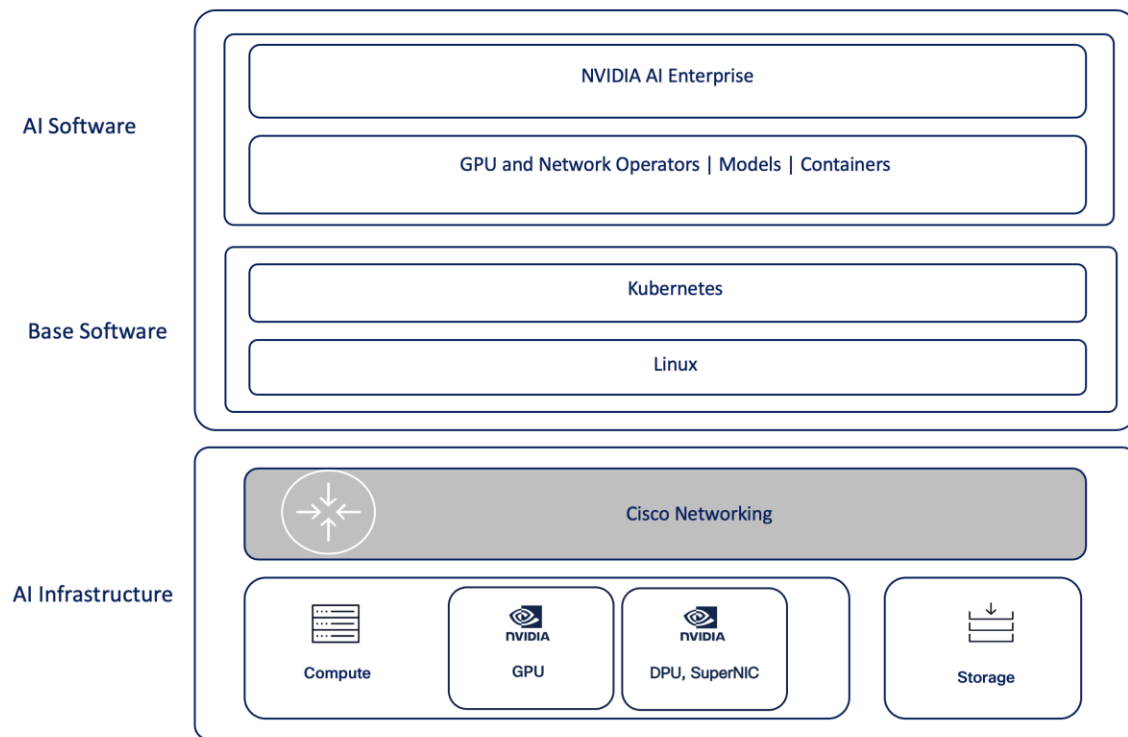


Figure 12.
Compute server software stack

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

Security

Security in 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 Enterprises 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 Enterprises 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 together compute, storage, and networking. 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, IB PerfTest, NCCL Test, 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 and tested to evaluate a number of enterprise-centric customer use cases around fine-tuning, inferencing, and RAG.

Summary

Cisco N9000 Series Switches and the Nexus Dashboard platform provide scalable, easy-to-manage, and high-performance networking for AI Infrastructure powered by NVIDIA-accelerated computing.

Appendix A – Compute server specifications

Area	Details
Compute + Memory	2x 6 th 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

Appendix B – Control-node server specifications

The following table shows the minimum specifications of control, management, and support server.

Table 7. Minimum specification of control, management, and support rack server

Area	Details
Compute + memory	1 or 2 Latest AMD or Intel x86 CPUs with minimum 64-cores total 512GB DDR5 RDIMMs
Storage	Dual 1 TB M.2 SATA or NVMe SSD with RAID (boot device)
Network cards	1 PCIe x16 FHHL NVIDIA BlueField-3 B3220 configured in DPU mode 1 OCP 3.0 X710-T2L (2 x 10G RJ45) for x86 host management
Power supply	2x PSU with N+1 redundancy
BMC	1G RJ45 for host management

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>.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: <https://www.cisco.com/go/trademarks>. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

Printed in USA