Cisco N9000 Enterprise RA

Full Stack AI Infrastructure compliant to NVIDIA GB300 NVL72 Enterprise Reference Architecture

Contents

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 GB300 NVL72 with up to 1152 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.

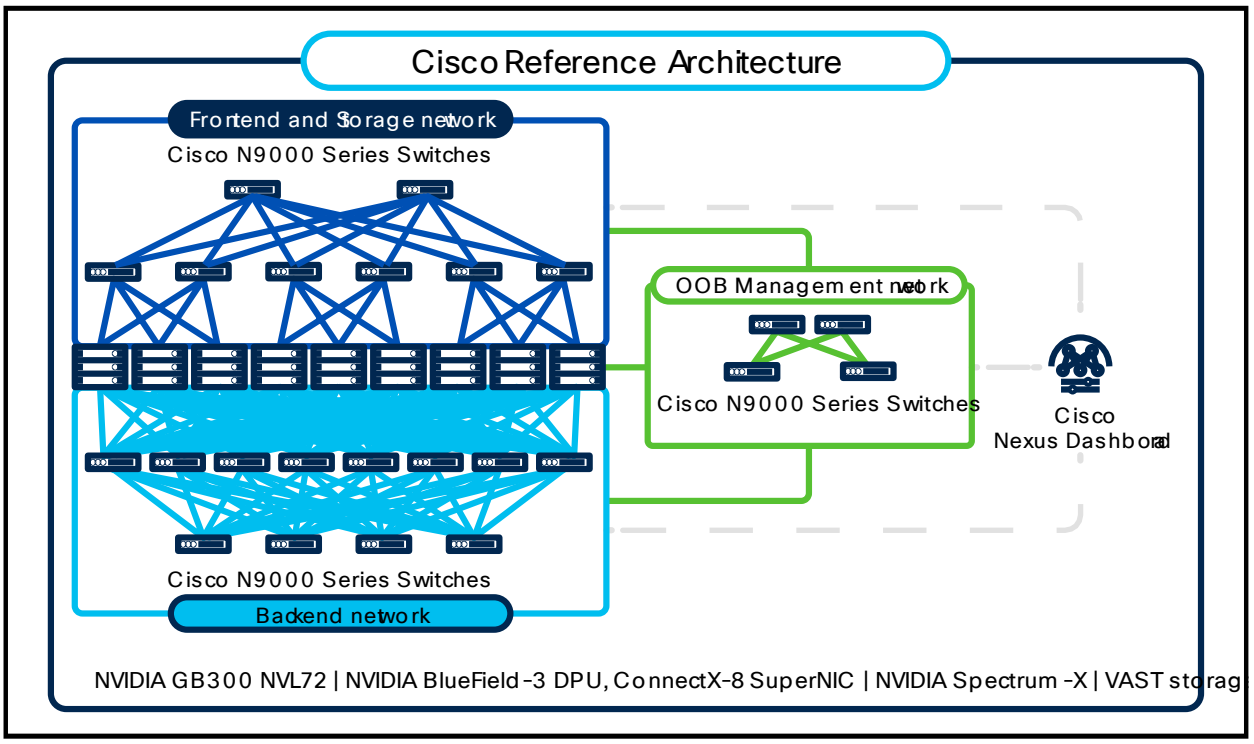


Figure 1.
Cisco N9000 Series Switches for networking the AI clusters, managed by Nexus Dashboard Platform

Hardware

GB300 NVL72 Rack

This RA uses NVIDIA-Certified™ GB300 NVL72 racks where each rack consists of 18 compute trays and 9 NVL5 switch trays. Each compute tray is in 2-4-5-800 (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).

Within each compute tray, there are 4x NVIDIA B300 GPUs paired with 2x Grace CPUs – one Grace CPU and two B300 GPUs together form a GB300 SuperChip. Each NVL5 switch tray consists of 2 NVSwitch ASICs per tray connecting to every GPU within the rack using NVL5 links. The 72 GPUs via the 18 NVSwitch ASICs together form a single coherent memory domain using the scale up network within the rack. Across the racks, GPU connectivity from every compute tray to other servers is via the use of 4x integrated NVIDIA ConnectX®-8 SuperNICs for East-West scale out traffic and via 1x NVIDIA BlueField®-3 B3240 DPU for North-South traffic. Supermicro SRS-GB300-NVL72 is an NVIDIA-Certified™ GB300 NVL72 rack supported in this RA.

Figure 2.
Supermicro SRS-GB300-NVL72 Rack

Cisco N9364E-SG2-O

The Cisco N9364E-SG2-O is a Cisco 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.

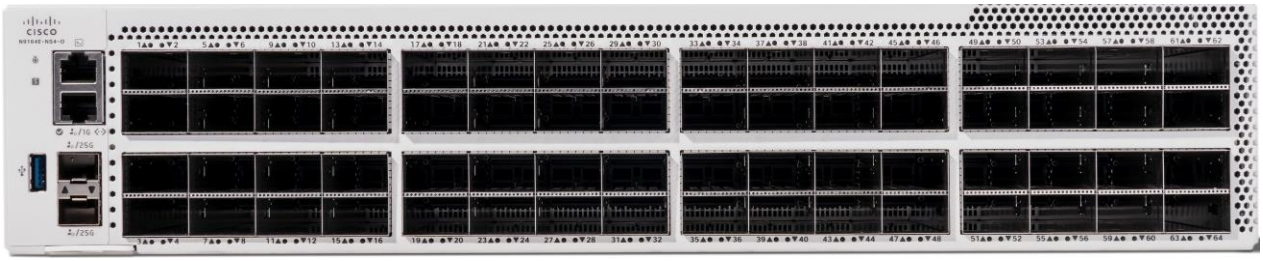


Figure 4.
Cisco Nexus N9164E-NS4-O switch

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	OSFPR-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 SuperNIC is used. However, in case of dual planes, 1 400G port per ConnectX-8 SuperNIC is connected to each of the two planes. Each GB300 NVL72 rack has 18 compute trays and each compute tray has 4 GPUs paired with 4 ConnectX-8 SuperNICs on a one-to-one basis representing 4 Rails 1, 2, 3, 4. This requires connecting 72 400G ports to plane 1 and 72 400G ports to plane 2 per rack split over 4 rail groups. Table 2, 3 shows the quantity of different units required for building a single and dual plane compute network with compute rack count ranging from 2 to 16.

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

Compute counts			Switch counts		Transceiver counts			Cable counts	
Racks	Nodes	GPUs	Leaf	Spine	Node to leaf		Switch to switch (800G)	Node to leaf	Switch to switch
					Node (800G)	Leaf (800G)			
2	**36**	144	4	2	144	72	144	144	144
4	**72**	288	8	3	288	144	288	288	288
8	**144**	576	16	6	576	288	576	576	576
16	**288**	1152	32	9	1152	576	1152	1152	1152

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

Compute counts			Switch counts		Transceiver counts			Cable counts	
Racks	Nodes	GPUs	Leaf	Spine	Node to leaf		Switch to switch (800G)	Node to leaf	Switch to switch
					Node (800G)	Leaf (800G)			
2	**36**	144	8	4	144	144	288	288	288
4	**72**	288	16	6	288	288	576	576	576
8	**144**	576	32	12	576	576	1152	1152	1152
16	**288**	1152	64	18	1152	1152	2304	2304	2304

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 each compute rack using 36 400G ports, over 18 compute trays, to two separate switches providing redundancy and high storage throughput. Each rack has 2 top-of-rack (ToR) switches each with 2 100G uplinks for redundancy and connect to the converged network. Enough storage facing ports are pre-allocated ensuring 12.5 gbps of network bandwidth per GPU. A total of 48 200G ports (12 800G ports) are also reserved for 12 support servers to allow running Slurm and Kubernetes control nodes, NVIDIA Base Command Manager™ head nodes, NVIDIA Mission Control™, NVIDIA NVLink Management Software Manager, 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
Racks	Nodes	GPUs	Leaf	Spine	Mgmt leaf	Storage leaf	Node to compute leaf		ISL ports	Rack ToR to spine		Mgmt leaf to spine		Storage leaf to spine		Spine to customer and support		
							Node (400G)	Leaf (800G)	800G	Node (100G)	Spine (800G)	Leaf (100G)	Spine (800G)	Leaf (400G)	Spine (800G)	Customer (800G)	Support (800G)	SMF MPO-12
2	**36**	144	2	N/A	1	2	72	36	32	8	2	2	2	8	4	6	12	148
4	**72**	288	4	2	1	2	144	72	144	16	2	2	2	12	6	10	12	344
8	**144**	576	8	4	1	2	288	144	288	32	4	4	4	20	10	18	12	656
16	**288**	1152	8	4	1	4	576	288	576	64	8	4	4	64	32	36	12	1312

Out-of-Band (OOB) Management Network

The OOB Management network is primarily used for node management connecting to the ToR switches inside the compute racks allowing access to the Base Management Controller (BMC) and host management ports of the different entities inside the compute rack and storage servers. The 100G uplinks of ToR switches inside the compute rack and 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 2 compute racks (144 GPUs)

Figure 8 shows the overall cluster topology interconnecting 2 compute racks with a total of 36 compute trays and 144 GPUs. Each plane of the compute network uses two spine and four leaf N9364E-SG2-O or N9164E-NS4-O switches. The four-leaf switches represent Rail 1, 2, 3, 4. The 4 800G NVIDIA ConnectX-8 SuperNICs on each compute tray are each plugged with OSFPR-800G-DR8 modules allowing 2 400G ports per NIC where the two ports connect to two different planes. With 2 compute racks consisting of 36 compute trays, there are a total of 144 800G or 288 400G ports, connecting 144 400G ports to compute

network plane1 and 144 400G ports to compute network plane2. Within a plane, 36 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 tray of every rack connect to two different switches for redundancy. With 2 compute racks consisting of 36 compute trays, there are a total of 72 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.

The 2 compute rack's ToR switches use a total of 8 100G uplinks connected to the converged network. Each Cisco EBox also requires 1 10G host management port and 1 1G server BMC port connected to a N93108TC-FX3 OOB management leaf switch.

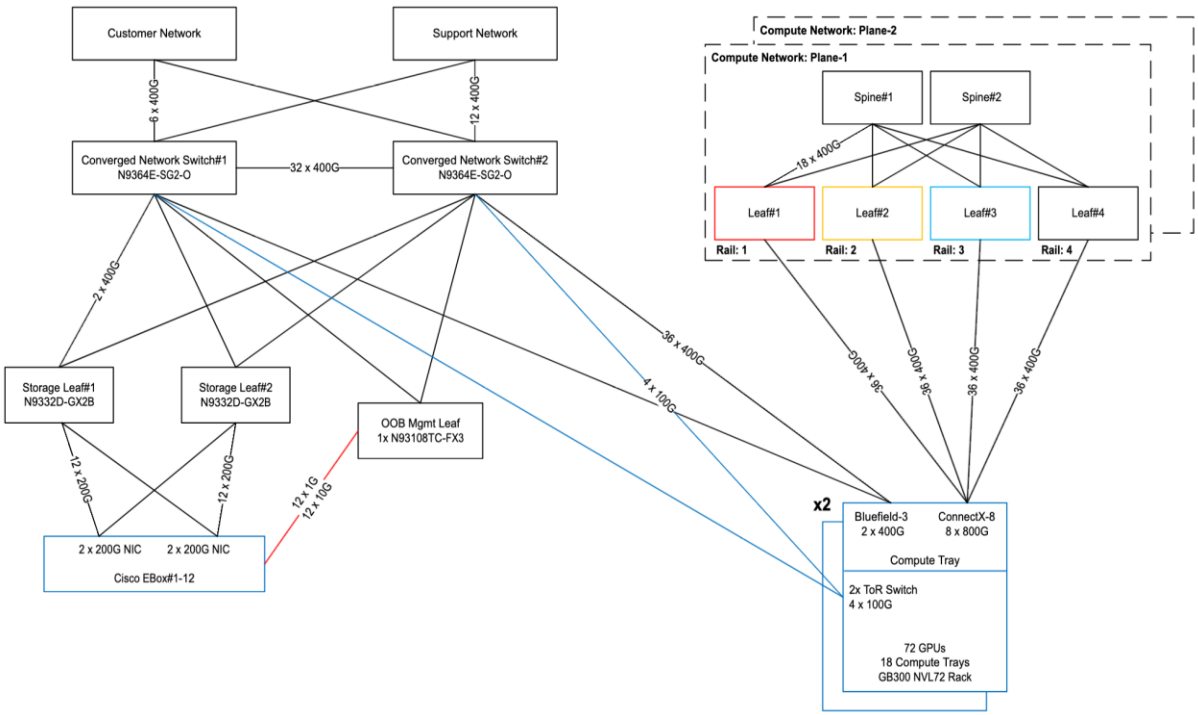


Figure 8.
Topology with 2 compute racks (144 GPUs)

Topology with 4 compute racks (288 GPUs)

Figure 9 shows the overall cluster topology interconnecting 4 compute racks with a total of 72 compute trays and 288 GPUs. Each plane of the compute network uses four spine and eight leaf N9364E-SG2-O or N9164E-NS4-O switches. The eight-leaf switches are divided into Rail 1, 2, 3, 4 with two leaf switches per rail-group. The 4 800G NVIDIA ConnectX-8 SuperNICs on each compute tray are each plugged with OSFPR-800G-DR8 modules allowing 2 400G ports per NIC where the two ports connect to two different planes. With 4 compute racks consisting of 72 compute trays, there are a total of 288 800G or 576 400G ports, connecting 288 400G ports to compute network plane1 and 288 400G ports to compute network plane2. Within a plane, 36 400G ports connect to every leaf switch.

The converged North-South network uses two spine and four leaf N9364E-SG2-O switches. The 2 400G NVIDIA Bluefield®-3 DPU ports per compute tray of every rack connect to two different switches for redundancy. With 4 compute racks consisting of 72 compute trays, there are a total of 144 400G DPU ports evenly split between four leaf switches.

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 12 400G uplinks evenly split between the two leaf switches.

The 4 compute rack's ToR switches use a total of 16 100G uplinks connected to the converged network. Each Cisco EBox also requires 1 10G host management port and 1 1G server BMC port connected to a N93108TC-FX3 OOB management leaf switch.

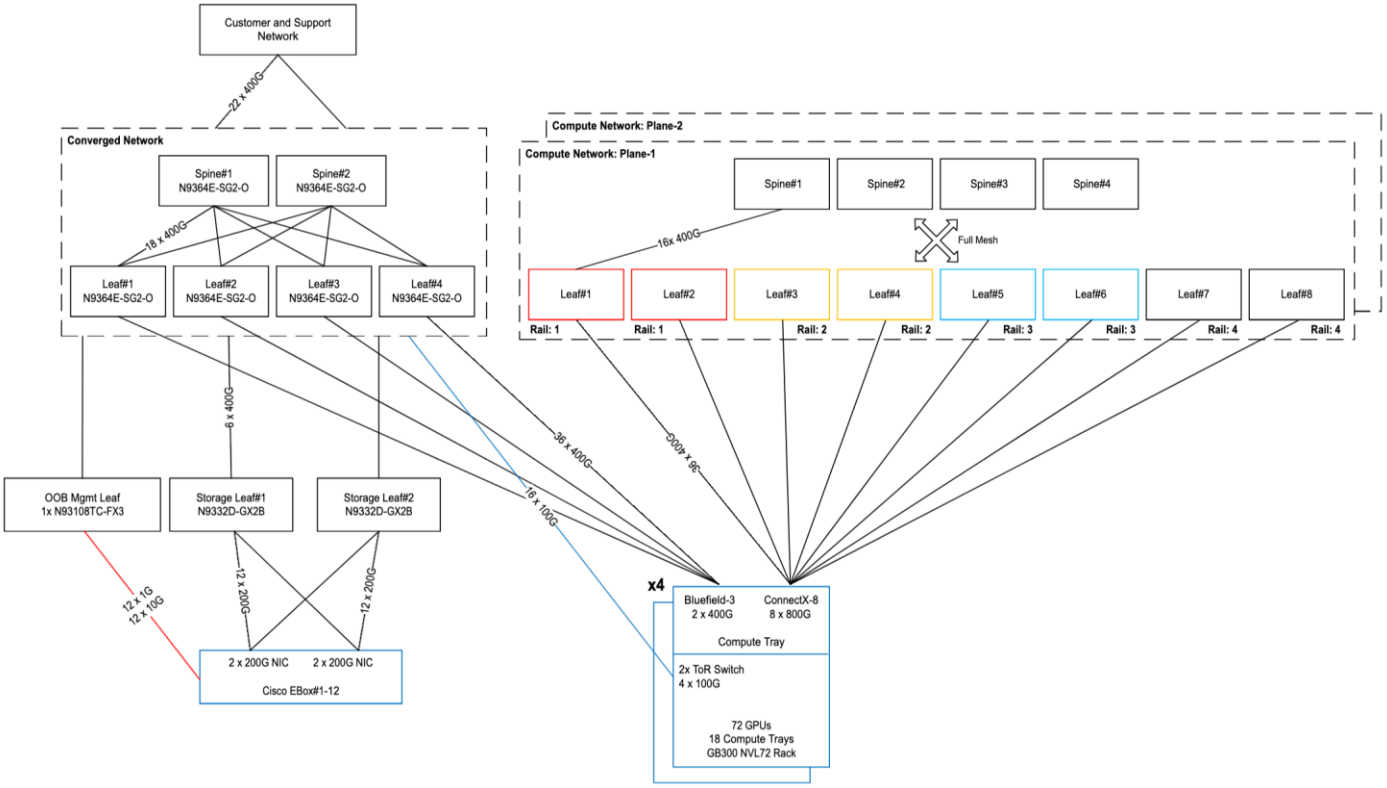


Figure 9.
Topology with 4 compute racks (288 GPUs)

Topology with 8 compute racks (576 GPUs)

Figure 10 shows the overall cluster topology interconnecting 8 compute racks with a total of 144 compute trays and 576 GPUs. Each plane of the compute network uses six spine and sixteen leaf N9364E-SG2-O or N9164E-NS4-O switches. The sixteen-leaf switches are divided into Rail 1, 2, 3, 4 with four leaf switches per rail-group. The 4 800G NVIDIA ConnectX-8 SuperNICs on each compute tray are each plugged with OSFPR-800G-DR8 modules allowing 2 400G ports per NIC where the two ports connect to two different planes. With 8 compute racks consisting of 144 compute trays, there are a total of 576 800G or 1152 400G ports, connecting 576 400G ports to compute network plane1 and 576 400G ports to compute network plane2. Within a plane, 36 400G ports connect to every leaf switch.

The converged North-South network uses four spine and eight leaf N9364E-SG2-O switches. The 2 400G NVIDIA Bluefield®-3 DPU ports per compute tray of every rack connect to two different switches for

redundancy. With 8 compute racks consisting of 144 compute trays, there are a total of 288 400G DPU ports evenly split between eight leaf switches.

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 20 400G uplinks evenly split between the two leaf switches. The number of Cisco EBox nodes can be increased to 20 for higher storage throughput and/or capacity with the use of QDD-400G-SR8 optics on storage leaf downlinks in 2 x 200G breakout mode.

The 8 compute rack's ToR switches use a total of 32 100G uplinks connected to the converged network. Each Cisco EBox also requires 1 10G host management port and 1 1G server BMC port connected to a N93108TC-FX3 OOB management leaf switch.

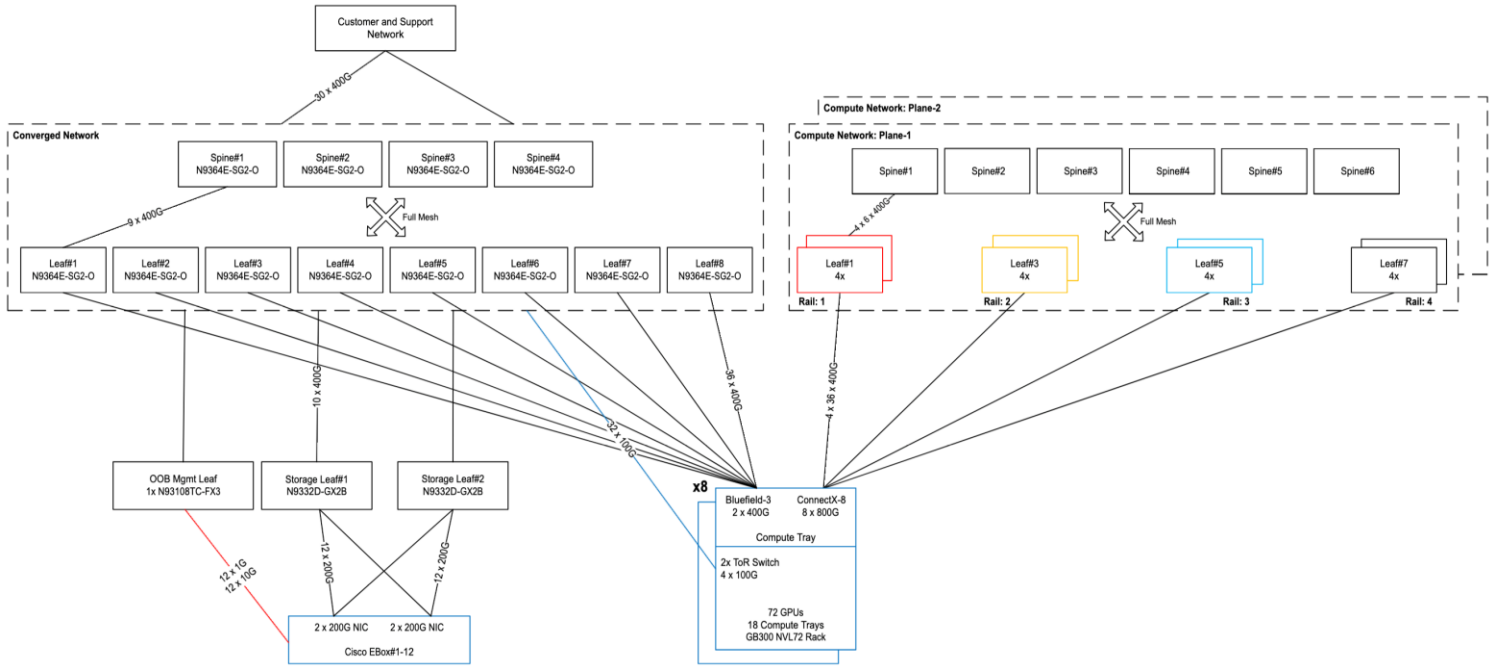


Figure 10.
Topology with 8 compute racks (576 GPUs)

Topology with 16 compute racks (1152 GPUs)

Figure 11 shows the overall cluster topology interconnecting 16 compute racks with a total of 288 compute trays and 1152 GPUs. Each plane of the compute network uses nine spine and thirty-two leaf N9364E-SG2-O or N9164E-NS4-O switches. The thirty-two leaf switches are divided into Rail 1, 2, 3, 4 with eight leaf switches per rail-group. The 4 800G NVIDIA ConnectX-8 SuperNICs on each compute tray 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 racks consisting of 288 compute trays, there are a total of 1152 800G or 2304 400G ports, connecting 1152 400G ports to compute network plane1 and 1152 400G ports to compute network plane2. Within a plane, 36 400G ports connect to every leaf switch.

The converged North-South network uses four spine and eight leaf N9364E-SG2-O switches. The 2 400G NVIDIA Bluefield®-3 DPU ports per compute tray of every rack connect to two different switches for redundancy. With 16 compute racks consisting of 288 compute trays, there are a total of 576 400G DPU ports evenly split between eight leaf switches. Unlike prior topologies, each leaf has a 2:1 oversubscription with 72 400G downlinks and 36 400G uplinks.

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 32 400G uplinks evenly split between the two leaf switches. An additional pair of storage leaf switches with 32 400G uplinks are pre-provisioned to allow future expansion of high-performance storage servers. The 64 400G uplinks allow 12.5 gbps network bandwidth per GPU with the remaining network bandwidth meant for inter-storage server communication across the two pairs of storage leaf switches. The number of Cisco EBox nodes can be increased to 32 for higher storage throughput and/or capacity with the use of QDD-400G-SR8 optics on storage leaf downlinks in 2 x 200G breakout mode.

The 16 compute rack's ToR switches use a total of 64 100G uplinks connected to the converged network. Each Cisco EBox also requires 1 10G host management port and 1 1G server BMC port connected to a N93108TC-FX3 OOB management leaf switch.

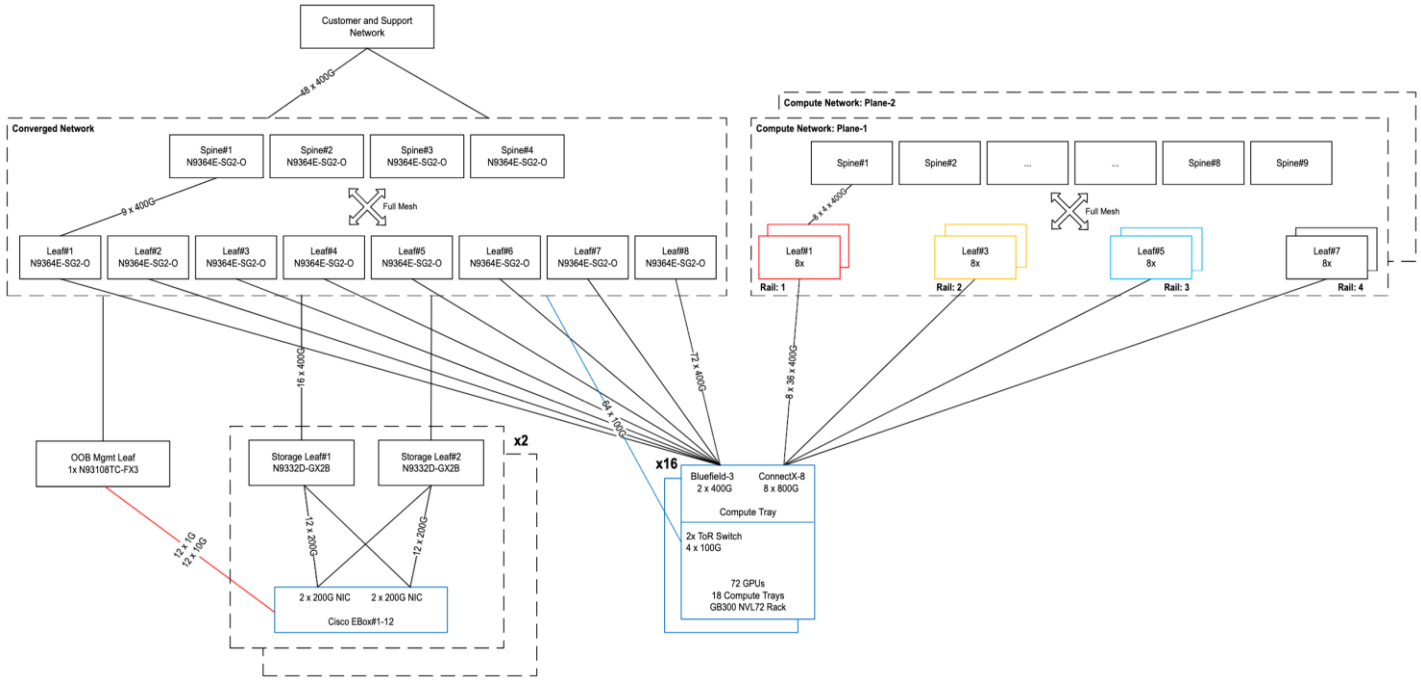


Figure 11.
Topology with 16 compute racks (1152 GPUs)

Cluster BOM

The BOM for clusters with varying compute racks (GPUs) using single compute plane is shown in Table 5.

Table 5. BOM of clusters with GPU scale 144 to 1152 using single compute plane

PID	Description	144 GPUs	288 GPUs	576 GPUs	1152 GPUs
SRS-GB300-NVL72	Supermicro GB300 NVL72 rack	2	4	8	16
N9364E-SG2-O **(Converged Network)**	Cisco N9000 switch, 64x800Gbps OSFP	2	6	12	12
N9364E-SG2-O	Cisco N9000 switch, 64x800Gbps OSFP	6	11	22	41

PID	Description	144 GPUs	288 GPUs	576 GPUs	1152 GPUs
N9164E-NS4-O (Compute Network)					
N9364E-SG2-O (Both Converged and Compute Network)	Cisco N9000 switch, 64x800Gbps OSFP	8	17	34	53
N9K-93108TC-FX3	Cisco Nexus switch, 48 1/10G BASE-T 6 QSFP28	1	1	1	1
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)	310	680	1344	2684
OSFPR-800G-DR8	OSFP, 800GBASE-DR8, SMF dual MPO-12 APC, 500m (riding heat sink)	144	288	576	1152
QDD-400G-DR4	400G QSFP-DD transceiver, 400GBASE-DR4, MPO-12, 500m parallel	8	12	20	64
QSFP-400G-DR4	400G QSFP112 transceiver, 400GBASE-DR4, MPO-12, 500m parallel	120	192	336	624
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	10	18	36	68
CB-M12-4LC-SMF	Cable, MPO12-4X duplex LC, breakout cable, SMF, various lengths	4	6	10	18
CB-M12-M12-SMF	MPO-12 single mode cables	436	920	1808	3616
CB-M12-M12-MMF	MPO-12 multi-mode cables	48	48	48	48
CAT5E	Copper cable for 1G	12	12	12	12
CAT6A	Copper cable for 10G	12	12	12	12
UCSC-C225-M8N (storage server)	Cisco UCS C225-M8 1RU Rack Server	12	12	Min: 12 Max: 20	Min: 12 Max: 32
UCSC-C240-M8 **UCSC-C245-M8SX (support server)**	Cisco UCS C240-M8 2RU Rack Server Cisco UCS C245-M8 2RU Rack Server	12	12	12	12

The BOM for clusters with varying compute racks (GPUs) using dual compute plane is shown in Table 6.

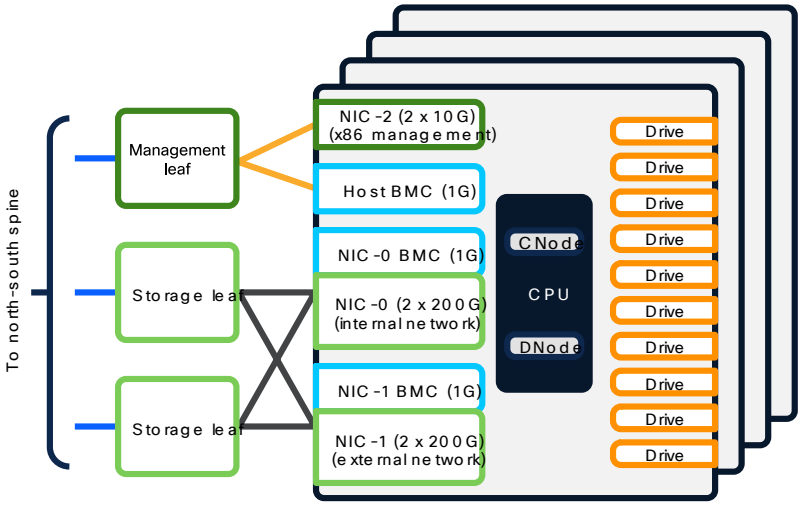
Table 6. BOM of clusters with GPU scale 144 to 1152 using dual compute plane

PID	Description	144 GPUs	288 GPUs	576 GPUs	1152 GPUs
SRS-GB300-NVL72	Supermicro GB300 NVL72 rack	2	4	8	16
N9364E-SG2-O (Converged Network)	Cisco N9000 switch, 64x800Gbps OSFP	2	6	12	12
N9164E-NS4-O (Compute Network)	Cisco N9000 switch, 64x800Gbps OSFP	12	22	44	82
N9364E-SG2-O (Both Converged and Compute Network)	Cisco N9000 switch, 64x800Gbps OSFP	14	28	56	94
N9K-93108TC-FX3	Cisco Nexus switch, 48 1/10G BASE-T 6 QSFP28	1	1	1	1
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)	526	1112	2208	4412
OSFPR-800G-DR8	OSFP, 800GBASE-DR8, SMF dual MPO-12 APC, 500m (riding heat sink)	144	288	576	1152
QDD-400G-DR4	400G QSFP-DD transceiver, 400GBASE-DR4, MPO-12, 500m parallel	8	12	20	64
QSFP-400G-DR4	400G QSFP112 transceiver, 400GBASE-DR4, MPO-12, 500m parallel	120	192	336	624
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	10	18	36	68
CB-M12-4LC-SMF	Cable, MPO12-4X duplex LC, breakout cable, SMF, various lengths	4	6	10	18
CB-M12-M12-SMF	MPO-12 single mode cables	724	1496	2960	5920
CB-M12-M12-MMF	MPO-12 multi-mode cables	48	48	48	48
CAT5E	Copper cable for 1G	12	12	12	12
CAT6A	Copper cable for 10G	12	12	12	12
UCSC-C225-M8N (storage server)	Cisco UCS C225-M8 1RU Rack Server	12	12	Min: 12 Max: 20	Min: 12 Max: 32
UCSC-C240-M8 **UCSC-C245-M8SX (support server)**	Cisco UCS C240-M8 2RU Rack Server Cisco UCS C245-M8 2RU Rack Server	12	12	12	12

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 12 shows the overall network connectivity and BOM with 12 storage servers. For data path, each server uses two NVIDIA BlueField-3 B3220L 2x200G SuperNICs: 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.



PID	Quantity	Type
UCSC-HFA-VAST-1P (C225 M8 config)	12	Server
N9K-C9332D-GX2B	2	Switch
N9K-93108TC-FX3	1	Switch
QSFP-200G-SR4	96	Optics
CB-M12-M12-MMF	48	Fiber cable
CAT6A	12	Copper cable
CAT5E	12	Copper cable

Figure 12.
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

NVIDIA Mission Control™ paired with NVIDIA Base Command Manager and NVLink Management software manager are essential for deploying and managing GB300 NVL72 systems, providing centralized provisioning, configuration, autonomous hardware recovery, and deep observability. They automate firmware updates, monitor liquid cooling, and manage the high-speed NVLink fabric, maximizing uptime for the overall compute infrastructure.

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 Supermicro GB300 NVL72 racks. NVIDIA AI Enterprise brings together optimized microservices, frameworks and libraries for AI development with advanced GPU orchestration and infrastructure management into a production-grade software suite. It enables the deployment of leading open-source tools and AI models, lowers infrastructure costs and reduces time to market while ensuring reliable, secure, and scalable operations.

NVIDIA NIM™ gives developers ready-to-run containers and standard APIs for serving AI models on NVIDIA GPUs.

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.

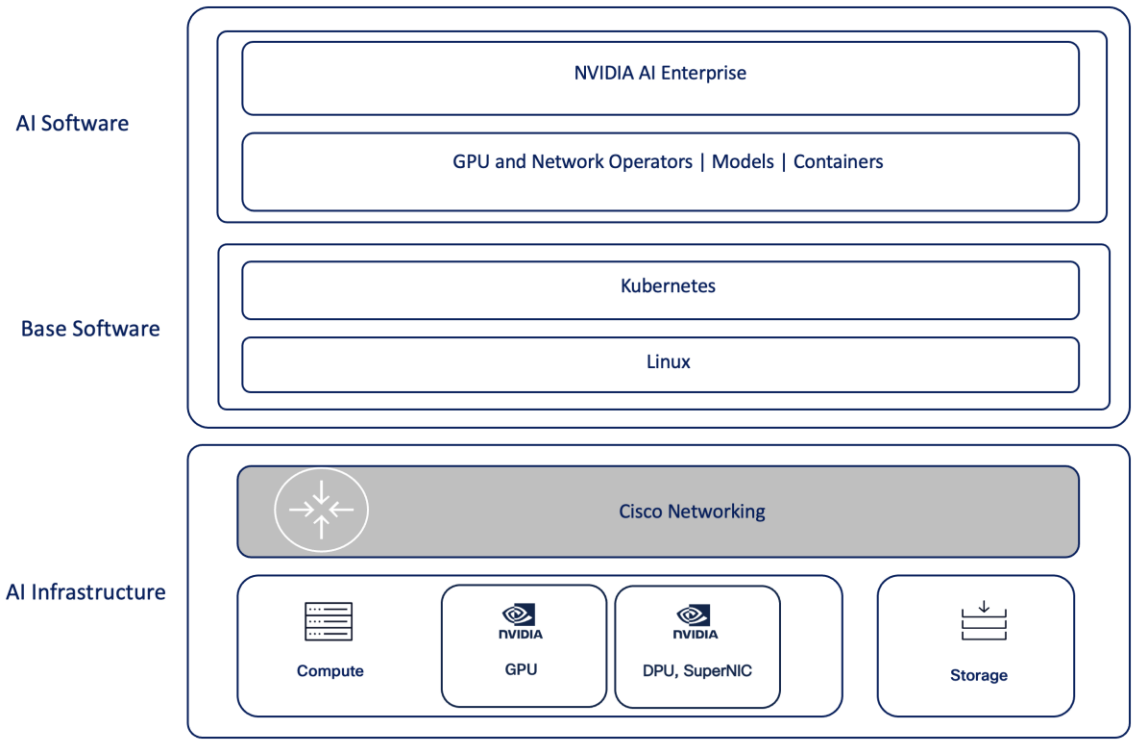


Figure 13.
Compute server software stack

Customers can run their choice of OS distribution and software versions as per the NVIDIA AI Enterprise 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 racks are NVIDIA-Certified Systems™. The Cisco EBox high-performance storage solution has achieved NVIDIA-Certified Storage validation at the NCP level. Several 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 several 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 – 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	2 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**.