Cisco EBox Foundation RA

High Performance Storage Foundation Reference Architecture

© 2026 Cisco and/or its affiliates. All rights reserved.

Contents

Cisco EBox 3
Hardware 3
Software 4
Networking 4
Deployment with 256 GPUs 5
Solution Performance Validation 6
Summary 6

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco, in collaboration with NVIDIA, has developed full-stack Cisco Secure AI Factory with NVIDIA solutions designed to accelerate Enterprise AI adoption, covering data ingestion, training, inference, retrieval-augmented generation (RAG). These solutions are based on NVIDIA-Certified® Cisco Enterprise Reference Architecture with the use of Nexus Hyperfabric and Nexus 9000 products. Beside compute and networking, storage is another critical component of the AI cluster that helps drive the overall workload performance. This document covers the high-performance storage (HPS) reference design for NVIDIA-Certified® Cisco EBox in AI clusters scaling to 256 GPUs with the use of compute systems as per design pattern 2-8-5-200 and 2-4-3-200.

Cisco EBox

The Cisco EBox storage solution has been developed in joint partnership between Cisco and VAST Data. It uses Cisco UCS C225-M8N 1RU rack servers running VAST Data AI OS. The key differentiating features of the overall storage solution are:

- Enhanced platform security with hardware anchored root-of-trust and tamper proof signed boot firmware.
- Single, global namespace that supports file, object, and tabular access while minimizing data copies and sprawl.
- Strong multi-protocol interoperability (NFS v3/v4, SMB v2/v3, S3) with high-performance RDMA and GPUDirect Storage support.
- Online, non-disruptive operations (expansion and upgrades) and enterprise-grade data protection (replication and large snapshot counts).
- Reliability, Availability, Serviceability (RAS) support with an ability to handle failure of multiple hardware components and their replacement in the field without cluster downtime.
- Wide erasure-code stripes across multiple storage enclosures and similarity-based data reduction to lower storage overhead and TCO.
- Integrated catalog, analytics, and observability to tie storage behavior to AI workload performance.
- Support for Quotas, ACLs, data encryption, key management, ransomware-proof indestructible snapshots.

Hardware

The required hardware configuration of the Cisco UCS C225-M8N server is shown in Table 1. It is pre-assembled and pre-configured at manufacturing when the solution is purchased to reduce the deployment time in the field.

Table 1: Hardware configuration of Cisco UCS C225-M8N EBox server

Item	Quantity	Description
CPU	1	AMD 9454P 2.75GHz 290W 48 cores
DRAM	384 GB	DDR5 (12 x 32 GB)
M.2 SSD	2	1 TB M.2 SATA SSD with HW RAID
SCM SSD	2	Micron XTR 960 GB NVMe U.2 SSD
QLC SSD	8	Minimum 15.36 TB QLC NVMe U.2 SSD

© 2026 Cisco and/or its affiliates. All rights reserved.

Item	Quantity	Description
Data Path NIC	2	NVIDIA Bluefield®-3 SuperNIC B3220L (2x200G)
x86 Mgmt NIC	1	2x10G RJ45 OCP 3.0 NIC
CIMC Mgmt NIC	1	1G RJ45
PSU	2	1200W Titanium Power Supply in (1 + 1) redundancy mode

For higher capacity deployments, one can also use 30.72/61.44/122.88 TB QLC SSDs paired with 1.92/3.84 TB SCM SSDs. NVIDIA Connectx®-7 (2x200G) as an alternate data path NIC is also supported.

Figure 1: Front view of the server showing location of SCM and QLC drives

Figure 2: Rear view of the server showing location of different NIC ports

Software

The server runs VAST Data AI OS supporting release-5.3.3 or later. Client nodes are required to install NFS driver for NFSoRDMA support. Within a vanilla Kubernetes or OpenShift environment, a CSI driver is also required. The overall solution is managed using a browser accessible UI that allows the operator to configure as well as monitor the overall storage cluster in realtime.

Networking

As shown in the server rear view in Figure. 2, each storage server has an internal NIC (closest to power supply) and external NIC (away from power supply). Ports on internal NIC are used for communication between storage servers. Ports on external NIC are used for communication to clients nodes. The x86 10G Mgmt port and 1G BMC/CIMC port should be connected to low speed switches.

As shown in Figure. 3, the two ports of the internal NIC should be connected to two different leaf switches for redundancy and should be configured to be in VLAN 69 on switch side. Similarly, the two ports of the external NIC should also be connected to two different redundant leaf switches. All 4 ports should be in default native VLAN to support IPv6 broadcast based neighbor discovery. All switches must also enable QoS configuration for RoCEv2 to support loss-less Ethernet. RDMA

© 2026 Cisco and/or its affiliates. All rights reserved.

traffic is set to use DSCP 26 and CNPs use DSCP 48. PFC must be enabled on queue3 and CNPs must be set to use queue6. Global link level flow control must not be enabled.

IP addresses from a Virtual IP Pool (VIP) is assigned to the external NIC ports of the storage server and used by client nodes for NFS mounting, S3 bucket access, SMB file share access. A good rule of thumb is to budget 4 to 8 virtual IPs per storage server with a maximum range of 2K for the whole storage cluster.

Figure 3: Network connectivity requirement for Cisco EBox

When deploying with NVIDIA SN5610, the switch side optics should be MMA4Z00-NS and server side optics should be MMA1Z00-NS400 both interconnected by MFP7E20-N0xx fiber. Using MCP7Y40-N003 copper cable is an alternate OSFP 800 to 4 x 200G QSFP112 breakout option.

Deployment with 256 GPUs

Under standard throughput guidance, an 11-EBox storage cluster (minimum offered) will meet the throughput demand of 256 GPUs. Table 3 shows the offered storage capacity when using drives of different sizes.

Table 2: Storage capacity with different drive types

GPUs	Nodes	EBox Count	Capacity (TB) (15.36 TB drives)	Capacity (TB) (30.72 TB drives)	Capacity (TB) (61.44 TB drives)
256	32	11	982	1965	3439

Figure 4 shows the network topology of 11 Cisco EBox storage servers connected to a pair of leaf switches each with 22 x 200G downlinks.

© 2026 Cisco and/or its affiliates. All rights reserved.

Figure 4: Cisco EBox storage cluster network topology supporting 256 GPUs

Solution Performance Validation

The overall storage solution has been rigorously verified in AI clusters using both training and inference workloads. A number of MLCommons training and inference benchmark results have been formally submitted. Synthetic test tools have also been utilized to stress the storage cluster to its maximum scale measuring consistent storage throughput, IOPs, and metadata performance. Additionally, the Cisco UCS C225-M8N EBox high-performance storage solution has achieved NVIDIA-Certified® Storage validation at the NCP level in both bonded and multirail (unbonded) configuration on client nodes with multiple North-South DPU ports.

Summary

The Cisco EBox is built on the foundations of a highly secure hardware platform that provides scalable storage performance for AI clusters of all sizes with significant economies of scale and maximum hardware utility to lower the TCO.

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

© 2026 Cisco and/or its affiliates. All rights reserved.