Cisco EBox for NCP

High Performance Storage Reference Architecture for NCP with GB200, GB300 NVL72 Systems

Contents

Cisco EBox 3
Hardware 3
Software 4
Networking 4
Multi-Tenancy 5
Storage Sizing 6
Deployment Topology with 1,152 GPUs 6
Deployment Topology with 16,128 GPUs 7
Deployment Topology with 41,472 GPUs 8
Solution Performance Validation 8
Summary 8

The NVIDIA® Cloud Reference Design (NCP RD) for AI training and inference, featuring NVIDIA GB200 and GB300 NVL72, is the next generation of data center architecture for artificial intelligence (AI) for cloud service providers (CSPs). 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 high GPU scale AI factories that start with 1152 GPUs and scale to 41472 GPUs in NVL72 systems.

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.
- Zero-trust principles with per-tenant encryption, RBAC/ABAC, QoS controls, and audit logging. Tenant isolation with distinct directory hierarchies and integration with enterprise identity providers (LDAP/AD).

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

Item	Quantity	Description
QLC SSD	8	Minimum 15.36 TB QLC NVMe U.2 SSD
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.

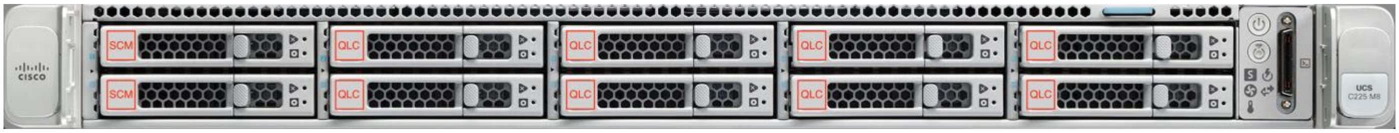

Figure 1: Front view of the server showing location of SCM and QLC drives

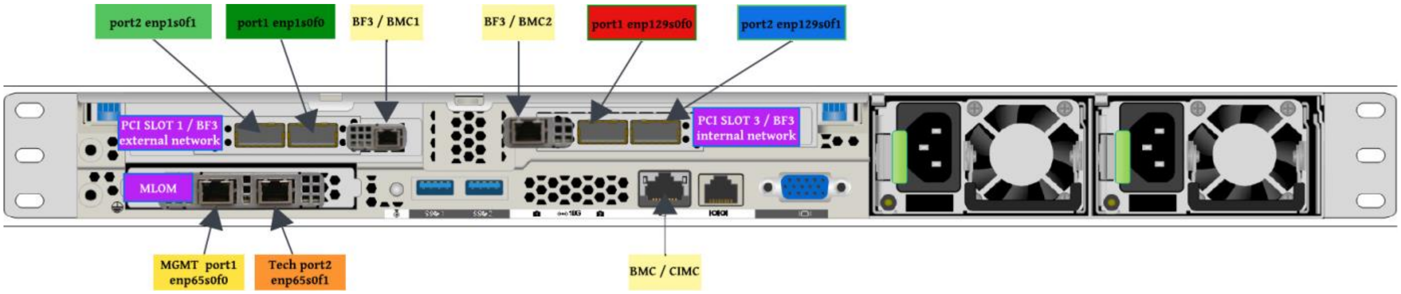

Figure 2: Rear view of the server showing location of different NIC ports

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 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.

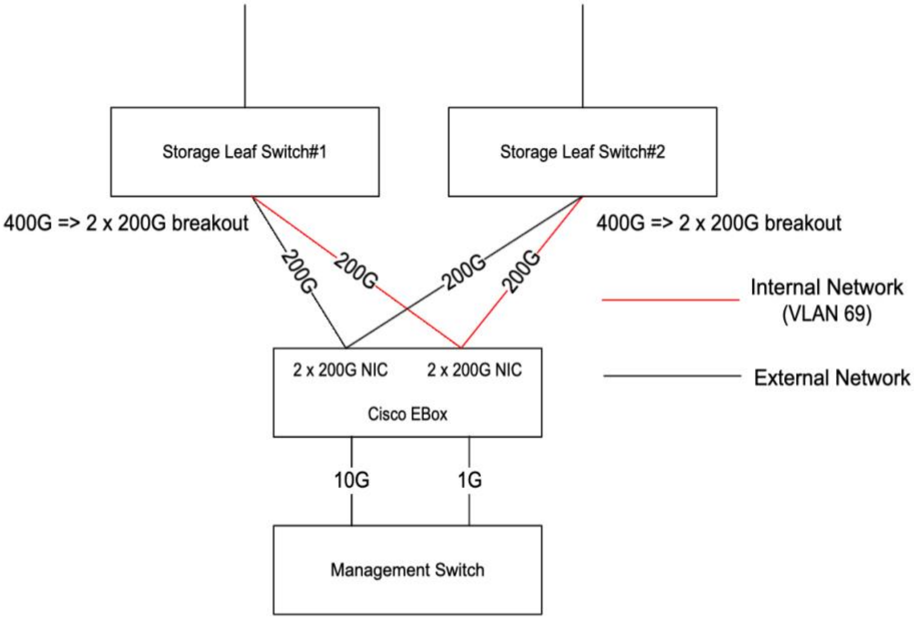

Figure 3: Network connectivity requirement for Cisco EBox

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 4x200G QSFP112 breakout option.

Multi-Tenancy

Ability to onboard multiple tenants on a shared storage cluster is supported where tenant isolation can be achieved via following two different approaches. Additionally, QoS controls on both throughput and IOPs are available to prevent impact of one tenants traffic on another.

- **Shared VIP pool with Tenant-IP filtering**: In this approach, a single shared VIP pool (with centralized DNS) per protocol serves all tenants but their access is allowed only from a pre-assigned non-overlapping source IP range list. This approach requires minimal configuration and is recommended when tenant scale is high.
- **Dedicated VIP pools**: In this approach, every tenant is assigned its own dedicated VIP pool(s) (with per-tenant DNS entries) which can also be isolated into a separate VLAN. This approach is recommended when hard network isolation is desired.

Storage Sizing

The storage performance target for training or inference can vary depending on the type of model and dataset. The guidelines in Table 2 provide standard throughput for the various GPU system sizes and HPS sizing. The final HPS requirements for throughput and capacity will be specified for each NCP opportunity.

Table 2: Guidance for standard HPS aggregate storage performance

Description	Number of GPUs						
	1152	2304	4608	8,064	16,128	29,952	41,472
Read Throughput (GB/s)	180	360	720	1,260	2,520	4,680	6,480
Write Throughput (GB/s)	90	180	360	630	1,260	2,340	3,240
Storage Configuration							
Number of appliances	12	24	48	83	165	306	424
Number of namespaces	1	1	1	1	1	1	1
Number of 400G storage ports	24	48	96	166	330	612	848
Number of in-band management connections	12	24	48	83	165	306	424
Number of out-of-band management connections	12	24	48	83	165	306	424
Number of rack units	12	24	48	83	165	306	424
Max Power (KW) – under extreme load and temperature	12	24	48	83	165	306	424
Max Cooling (KBTU/hr) - under extreme load and temperature	36	72	144	249	495	918	1272

Deployment Topology with 1,152 GPUs

Under standard storage throughput guidance, in order to support workloads in a cluster of 1,152 GPUs, 12 EBox servers are required. Within the converged network, as shown in Figure 4, they are connected to two storage leaf switches. Each storage leaf switch will require 12 400G ports (each in 2 x 200G breakout mode) as downlinks connecting to the storage servers.

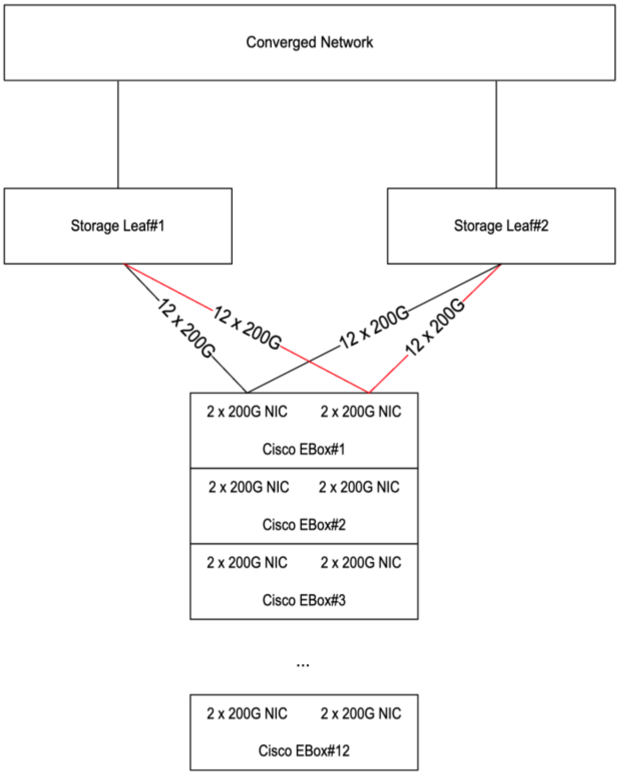

Figure 4: Cisco EBox storage cluster network topology supporting 1,152 GPUs

Deployment Topology with 16,128 GPUs

Under standard storage throughput guidance, in order to support workloads in a cluster of 16,128 GPUs, 165 EBox servers are required. Within the converged network, as shown in Figure 5, they are distributed over 3 pairs of 128 x 400G storage leaf switches - the first 2 pairs each connect to 64 storage servers on the downlink and the last pair connects to 37 storage servers. One can consider a pair of storage leaf switches together with 64 storage servers as a lego building block which is replicated enough number of times to match the desired appliance count as per Table 2 for a given GPU cluster size.

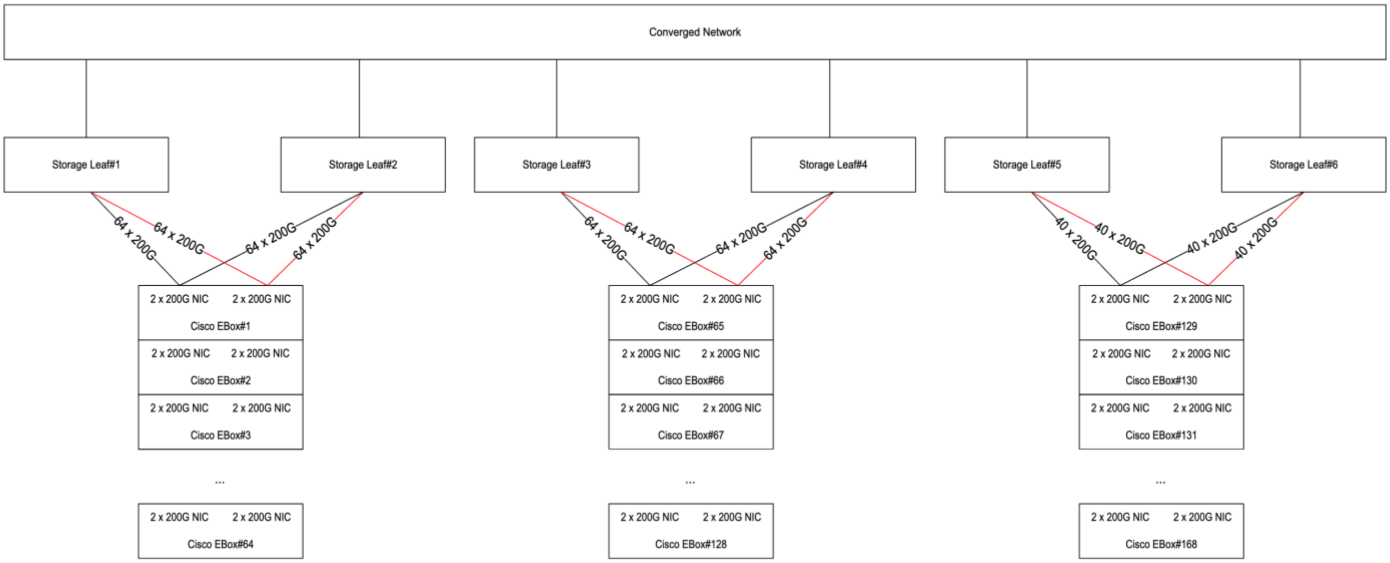

Figure 5: Cisco EBox storage cluster network topology with 128 x 400G switches supporting 16,128 GPUs

Deployment Topology with 41,472 GPUs

Under standard storage throughput guidance, in order to support workloads in a cluster of 41,472 GPUs, 424 EBox servers are required. Within the converged network, as shown in Figure 7, they are distributed over 7 pairs of 128 x 400G storage leaf switches - the first 6 pairs each connect to 64 storage servers on the downlink and the last pair connects to 40 storage servers. An alternate way to visualize, utilizing the lego block concept in previous section, is to use 6 such lego blocks (each with 64 storage servers) and the last block with 40 storage servers.

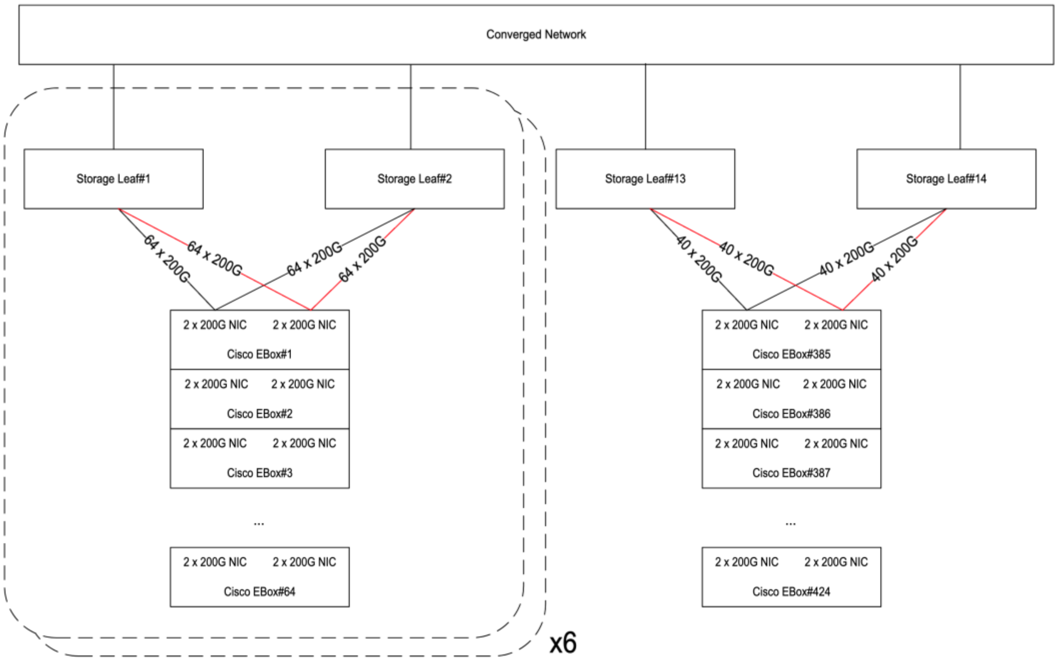

Figure 6: Cisco EBox storage cluster network topology with 128 x 400G switches supporting 41,472 GPUs

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**.

Printed in USA