White paper

Enterprise Reference Architecture for Object Storage with VAST Data CNodes and DBoxes

Table of Contents

VAST Data collaboration with NVIDIA 3

NVIDIA-Certified Storage Overview 3

Executive Overview: 3

About the VAST AI Operating System 4

NVIDIA Enterprise Reference Architecture 5

Key building blocks of ERA 7

2-4-3-200: Foundation level 12

2-8-5-200: Foundation level 13

Summary 15

VAST Data collaboration with NVIDIA

The VAST Data partnership with NVIDIA establishes a powerful foundation for building a high-performance, AI-optimized infrastructure. A key aspect of this collaboration is the NVIDIA Enterprise Reference Architecture (ERA), the next generation of universal data center design — which leverages NVIDIA GPUs, NVIDIA Spectrum-X Ethernet platform, and seamlessly integrates with the VAST AI Operating System.

By combining the VAST AI Operating System with NVIDIA ERA™, enterprises gain a scalable, high-performance data storage solution purpose-built to accelerate Analytics, Classical AI, Generative AI, and small-scale High-Performance Computing (HPC) workloads. For Generative AI initiatives, this architecture is optimized to support Retrieval-Augmented Generation (RAG), model fine-tuning, and inference workloads, with a particular focus on single-tenant, Ethernet-based environments.

Organizations can further extend their AI solutions using frameworks such as the NVIDIA AI Enterprise software platform, along with other leading machine learning (ML) and deep learning (DL) stacks.

For AI and solution architects seeking deeper technical insights, the provided guide offers a comprehensive walkthrough of deploying and managing ERA systems — covering architecture components, networking configurations, and deployment best practices. With this proven foundation, enterprises can start with as few as four nodes and scale seamlessly as their AI demands evolve, ensuring their infrastructure remains future-ready and aligned with the latest AI advancements.

Learn more about the **VAST and NVIDIA collaboration**

NVIDIA-Certified Storage Overview

Executive Overview:

NVIDIA Enterprise Reference Architectures provide guidance for building high-performance, scalable data center infrastructure. As enterprises transform from traditional data centers into AI Factories, they are revolutionizing the way data is processed and analyzed — integrating advanced computing and networking technologies to meet the substantial computational demands of AI applications.

To support this transition, NVIDIA introduced the Enterprise Reference Architectures, offering clear, consolidated design recommendations for partners and enterprise customers building AI Factories based on NVIDIA-Certified Systems. Informed by years of expertise in designing and operating large-scale computing environments, these architectures streamline deployment, offering flexible, cost-effective configurations that minimize guesswork and reduce risk.

NVIDIA ERAs include prescriptive design guidance for building accelerated computing clusters, utilizing NVIDIA-Certified servers with optimized CPU-to-GPU-to-NIC ratios. This balanced design approach helps eliminate bottlenecks and ensures consistent, high-performance outcomes across AI workloads.

The VAST AI Operating System delivers a comprehensive solution for the enterprise market, addressing the rapidly growing multi-protocol storage needs of Accelerated Analytics, Classical AI, and Generative AI workloads.

VAST Data CNodes and DBoxes have achieved NVIDIA-Certified Storage under the Foundation level, which validates VAST Data object storage against AI workloads scaling to environments of up to 128 GPUs. . NVIDIA-Certified Storage Foundation level reference architecture #1139257 signifies the VAST Data AI Operating System meets stringent performance requirements with an architecture that delivers speed, scale, and operational efficiency. It is designed for ease of use, allowing any IT team to deploy and support large-scale AI initiatives with ERA designs seamlessly.

Business Benefits:

- Accelerate Time to Token: By leveraging NVIDIA's structured approach and recommendeddesigns, enterprises can deploy AI solutions faster, reducing time-to-token and accelerating time-to-business-value.
- Resource and Cost Efficiency: Optimized configurations ensure efficient use of resources,minimizing waste and reducing overall infrastructure and operational costs.
- Risk Mitigation: Deploying proven, tested design patterns increases customer confidence, mitigates deployment risks, and ensures consistent, predictable outcomes

IT Benefits:

- Performance: High-performance computing capabilities meet the demanding requirements ofAI workloads, ensuring consistent and optimal performance.
- Scale and Manageability: Flexible scaling options and streamlined configurations allow enterprises to grow their AI infrastructure efficiently and manage it with ease.
- Reduced Complexity and TCO: Simplified deployment processes and efficient architectural designs significantly reduce operational complexity and total cost of ownership (TCO).
- Supportability: Standardized design patterns enable consistent deployments across environments, reducing support overhead and accelerating issue resolution.

About the VAST AI Operating System

The VAST AI Operating System, engineered with the industry's first Disaggregated, Shared-Everything (DASE) architecture, comprises three key infrastructure components: CNodes (compute) & DBoxes (capacity) communicating with each other over a low-latency, RDMA-enabled, NVMe fabric (NVMe-oF).

The NVMe fabric, in this reference design, is provided by NVIDIA Spectrum SN5600/SN5610 leaf switches. This architecture enables a groundbreaking approach to storage system design by enabling the ability to independently scale performance (via CNodes) and capacity (via DBox). This flexibility for service providers— especially those with growing or expanding infrastructure needs—alleviates common capacity forecasting challenges while simultaneously meeting the needs of evolving AI workloads.

VAST CNodes

x86-based compute nodes each running a collection of stateless containers that handle all aspects data access and management.



VAST DBoxes

High-density NVMe enclosures in a high-availability configuration to store all file and object storage data as well as metadata in a single namespace.



VAST CNodes

VAST CNodes constitute the core processing units of the VAST AI Operating System. CNodes are stateless, containerized servers that manage all critical functions, including protecting user data via erasure coding, performing background data reduction, processing complex database queries, and handling file and object storage requests.

CNodes handle every processing request without ever persisting data or metadata locally (in fact, there are no local disks except for boot drives). Instead, every CNode has access to every remote SSD in the cluster as if it were locally attached. The stateless nature of the CNode design is what enables truly linear performance scaling (no east-west communication in the data path), dramatically better reliability (failed CNodes do not trigger disk rebuilds), as well as simplified management (all maintenance operations—including upgrades—are done completely online).

Each CNode today leverages industry-standard x86 processors, equipped with PCIe Gen 5 interfaces to maximize data throughput and minimize latency. The design is optimized to support intense I/O operations, with each node contributing processing power to the aggregate throughput of an entire cluster.



Key specifications of each VAST CNode include:

CPU:
Single AMD Turin 9555P (64 Cores/128 Threads) CPU 3.2Ghz

I/O Connectivity:
Dual Port Connect-X7 for frontend (Client-Facing and Internal Facing)

Memory:
384GB (12 x 32GB) DDR5-6400

Form Factor:
1 Rack Units (RU)

Power Consumption:
0.55 kW Active

VAST DBoxes (High Availability NVMe Enclosures)

VAST DBoxes are high-density, 100% NVMe, highly-redundant systems that store all the persistent state (both data and metadata) of the cluster. VAST DBoxes house a combination of Storage Class Memory (SCM) and hyperscale NVMe flash SSDs and present each drive—via multiple NVIDIA Spectrum-X Ethernet network interfaces—onto an internal, ultra-low-latency NVMe fabric.

One specific DBox model, "CeresV2", leverages NVIDIA BlueField®-3 DPUs to enhance networking capabilities and offload essential data processing tasks, enhancing security and efficiency. These DPUs are instrumental in managing the network flow and securing the data paths, ensuring robust HPS operations.



Key specifications of each 1RU CeresV2 DBox include:

Servers:
Each DBox contains two trays

Processors:
Each tray supports one BlueField-3 DPU

I/O connectivity:
2 ports per BlueField-3 DPU per tray

Total Capacity:
Up to 2,700TB maximum raw capacity per DBox

Drives:
22 x E1.L (9.5 mm) Hyperscale QLC drives and 8 x SCM drives

Form Factor:
1 rack unit (RU)

Power Consumption:
0.75 kW Active

The Converged Network

In the Converged Data Network design, the Front-End (FE) and Back-End (BE) network traffic for the CNodes & DBoxes are hosted on the same Spectrum SN5600/SN5610 Leaf switches.

The Front-End traffic from the CNodes serves all protocol access (NFS/S3/SMB/Block/Kafka) to the GPU nodes and enables the VAST Management system (VMS) to be reached on the data network via a dedicated mgmt-data Virtual IP (VIP). This dedicated mgmt-data VIP facilitates API access for eg. Kubernetes CSI/COSI, Audit Catalog.

NVMe-oF is the main Back-End connectivity between CNodes and DBoxes and forms the backbone of the VAST DASE architecture. This technology allows VAST CNodes to access storage devices (within DBoxes) across a high-speed network, as if they were locally connected in the same physical server. By leveraging Remote Direct Memory Access (RDMA), NVMe-oF provides the essential high-speed data transfer capabilities required to meet the demands of data-intensive applications. NVMe-oF is integral to ensure that each CNode can efficiently interact with every SSD across all DBoxes in the deployment, regardless of physical location within the data center. This setup supports the disaggregated storage architecture of the VAST AI Operating System, where data processing and capacity resources are scaled and managed independently.

Software Architecture

The VAST AI Operating System (AI OS) is the first platform built from the ground up to manage the complete AI lifecycle. One unified foundation for storage, database, and compute runtime – eliminating the need to stitch together dozens of tools for training, inference, autonomous agents, and everything in between as shown in figure 1.

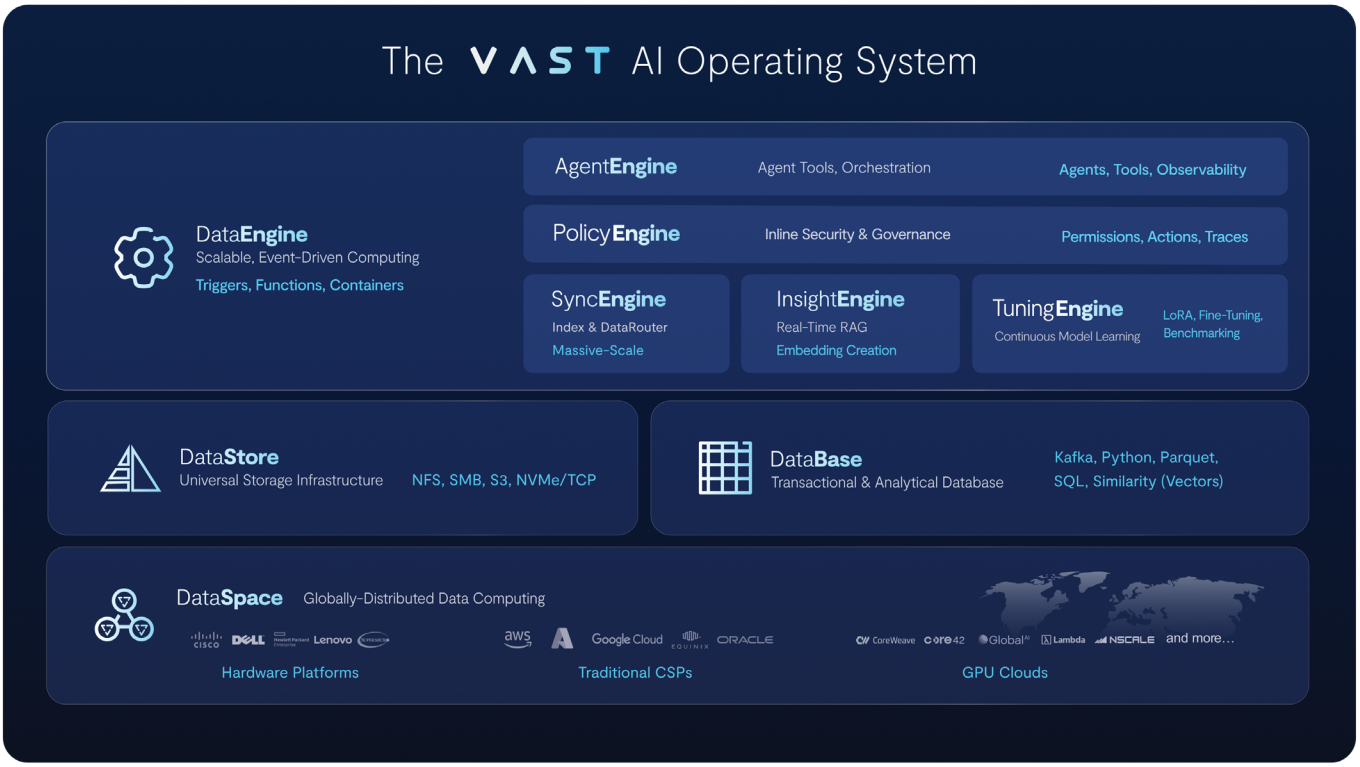


Figure 1. The VAST AI Operating System

Other benefits of the VAST AI Operating System for ERA designs include:

- Independent scaling of performance and capacity, with support for mixed generations of hardware in a single exabyte-scale namespace.
- Archive-tier economics enabled by next-generation global data reduction algorithms, support for hyperscale QLC flash with ten years of endurance, and ultra-efficient locally decodable erasure codes.
- Non-disruptive, online system expansions and software upgrades.
- Encryption, authentication, and external key management.
- Enterprise-grade data protection with support for n-1 and 1-n replication topologies and up to 1 million ransomware-proof snapshots.
- The VAST Data Management System (VMS) GUI provides thousands of metrics via an API-first architecture, unlocking real-time visibility into performance metrics.
- The VAST DataStore is a scalable HPS solution with support for multiple file (NFS, SMB) and object (S3) interfaces. This multi-protocol support and interchangeability enable simultaneous access to data for reading and writing from either NFS or S3 storage protocols.

- The VAST DataBase is an exabyte-scale data warehouse capable of delivering robust performance for both transactional and analytical workloads. It excels in environments with intensive data preprocessing needs by accelerating data access for frameworks such as Apache Spark, Trino, and NVIDIA Data Science (via RAPIDS-enabled applications).
- VAST DataSpace extends the platform's capabilities into a cohesive environment for global data management, offering a unified namespace across on-prem and multiple region data centers. It ensures consistent and reliable data access with policies that optimize data placement and synchronization while maintaining high performance.
- VAST DataEngine is a real-time, serverless compute and orchestration engine that transforms data into actionable intelligence. It enables continuous, event-driven processing across a globally distributed infrastructure.

Framework	Purpose	Key Benefit
VAST SyncEngine	Ingestion & Sync	Keeps datasets current by synchronizing external repositories automatically—eliminating the need for manual scripts or scheduled jobs.
VAST InsightEngine	Real-Time Enrichment	Automatically extracts structure and meaning from new data as it arrives, maintaining a live index for immediate AI application use.
VAST AgentEngine	Agent Operationalization	Provides a governed runtime for AI agents, allowing long-running tasks to resume reliably without restarting from scratch.

NVIDIA Enterprise Reference Architecture:

Key building blocks of ERA:

Accelerated Computing Clusters: NVIDIA provides validated design patterns for both GPU scale-out and scale-up cluster configurations. These patterns leverage NVIDIA® NVLink® technology to enable high-speed, low-latency, all-to-all GPU communication, critical for performance-intensive AI, Analytics and HPC workloads.

- PCIe-Optimized Reference Configurations Enterprise Reference Architectures (RAs) based on PCIe-Optimized Configurations integrate NVLink technology within individual servers (where supported, such as in the H100 NVL). These configurations also provide scalable design guidance for expanding clusters by adding more servers, enhancing both overall capacity and performance. This approach is ideal for workloads demanding high performance from each server within the cluster.

Building blocks are engineered to flexibly expand GPU and networking resources per node as workloads evolve. NVIDIA Enterprise RAs incorporates the best practices to maintain a balanced architecture across CPUs, GPUs, and Network Interface Cards (NICs), minimizing bottlenecks and ensuring consistent, optimal system performance. This balanced design enables organizations to scale computational resources seamlessly, supporting the growing demands of complex, distributed workloads while sustaining high throughput across the entire cluster.

Networking

NVIDIA Enterprise Reference Architectures (RAs) integrates best practices from AI cloud data center deployments to optimize network traffic flow, maximize AI performance, and maintain cloud-scale manageability and security. Each Enterprise RA includes network design recommendations based on the Spectrum-X Ethernet platform, combining Spectrum-4 Ethernet switches with NVIDIA BlueField-3 SuperNICs to deliver optimal throughput, low latency, and secure operations.

Each architecture also provides detailed guidance on network topology design, tailored for different scales and use cases, ensuring flexibility and future-proof scalability.

- East-West Networking: East-West traffic refers to internal data transfers within the data center—critical for AI model training and scaling. NVIDIA's East-West networking recommendations focus on delivering high bandwidth and ultra-low latency connections between GPUs, CPUs, and storage systems. These solutions ensure seamless communication within AI clusters, preventing internal bottlenecks that could slow training times and reduce the efficiency of AI pipelines. Properly optimized East-West traffic is essential for scaling AI workloads as data moves across multiple layers and computational units.
- North-South Networking: North-South traffic manages external communications—specifically for data ingestion, storage access, and result delivery. For handling North-South traffic efficiently and securely, NVIDIA recommends the deployment of BlueField Data Processing Units (DPUs). BlueField DPUs offload networking tasks from CPUs, enabling secure, efficient processing of incoming and outgoing network requests without introducing bottlenecks.
- Switching: NVIDIA Enterprise RAs specify Ethernet-based switching as the preferred networking choice for enterprise AI and data workloads. Each RA includes validated Ethernet configuration recommendations to ensure consistent, high-performance switching that meets the demanding requirements of modern AI data centers.

2-4-3-200: Foundation and Enterprise level

The 2-4-3-200 (CPU–GPU–Network Adapter–Network Bandwidth) configuration defines a standardized, NVIDIA-Certified scale-out compute node optimized for enterprise AI and visual computing workloads. Each node supports:

- 2 CPUs
- 4 PCIe GPUs (such as L40S, RTX PRO 6000 Server Edition, H100 NVL, or H200 NVL)
- 3 high-performance network adapters (including ConnectX-7 and BlueField-3 SuperNICs)
- 200 Gbps of total network bandwidth per node

Complete details of this high-performance, modular building block for compute, storage, and networking can be found in NVIDIA-Certified Storage Foundation level reference architecture design guide – #113925.

Compute Counts				VAST Config		Converged Network Ports	
SUs	Nodes	GPUs	CPUs (Node N-S)	CNodes	DBox	400 Gbe Port Utilization	Mgmt Uplinks
1	4	16	2	3	1	5	5
2	8	32	4	3	1	5	5
3	12	48	4	3	1	5	5
4	16	64	6	3	1	5	5

Table 1: Scalable Configuration of 243 Design from 1 SU to 4 SU with VAST Data.

Table 1 above provides 243 Design with VAST AI Operating System scaling based on a single SU to four SUs. As the scalable unit grows, VAST Data CNodes (Compute Nodes) and DBoxes (Data Nodes) can be seamlessly expanded to provide additional performance and capacity scaling to meet growing AI and enterprise workload demands.

NVIDIA H100 NVL Enterprise Reference Architecture Optimized 2-4-3 Design

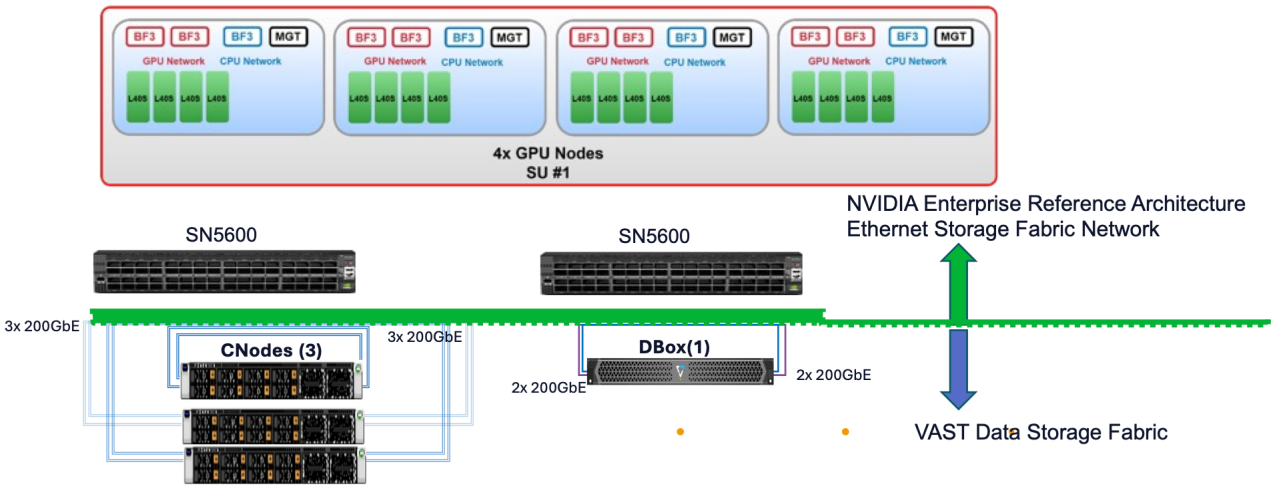


Figure 2 VAST Data Integration with 243 Design

The key important Use Cases for Enterprises with 243 Design are

Visual Computing: It empowers enterprises to process, generate, and interact with rich visual content — including 3D graphics, image rendering, simulations, and augmented/virtual reality (AR/VR) experiences. These capabilities drive innovation across a wide range of industries:

- Design and Engineering: Utilized in Compute Aided Design (CAD) for industries such as automotive, aerospace, and architecture, enabling detailed modeling, prototyping, and simulation workflows. CAD applications for automotive, aerospace, and architecture industries.
- Retail and E-commerce: Supports virtual product modeling and augmented reality (AR) shopping experiences, enhancing customer engagement and accelerating purchasing decisions.
- Healthcare: Powers 3D medical imaging, diagnostic modeling, and surgical simulation platforms, improving diagnostic accuracy and supporting advanced medical training.
- Media and Entertainment: Used by movie studios, animation houses, and game developers for 3D animation, visual effects (VFX), and immersive gaming experiences.

AI Inference: Enables enterprises to deploy medium-sized pre-trained models for tasks such as predictions, classifications, and content generation. Ideal for real-time and batch inference use cases where medium-scale model sizes (typically hundreds of millions to a few billion parameters) deliver high accuracy with manageable compute and latency requirements.

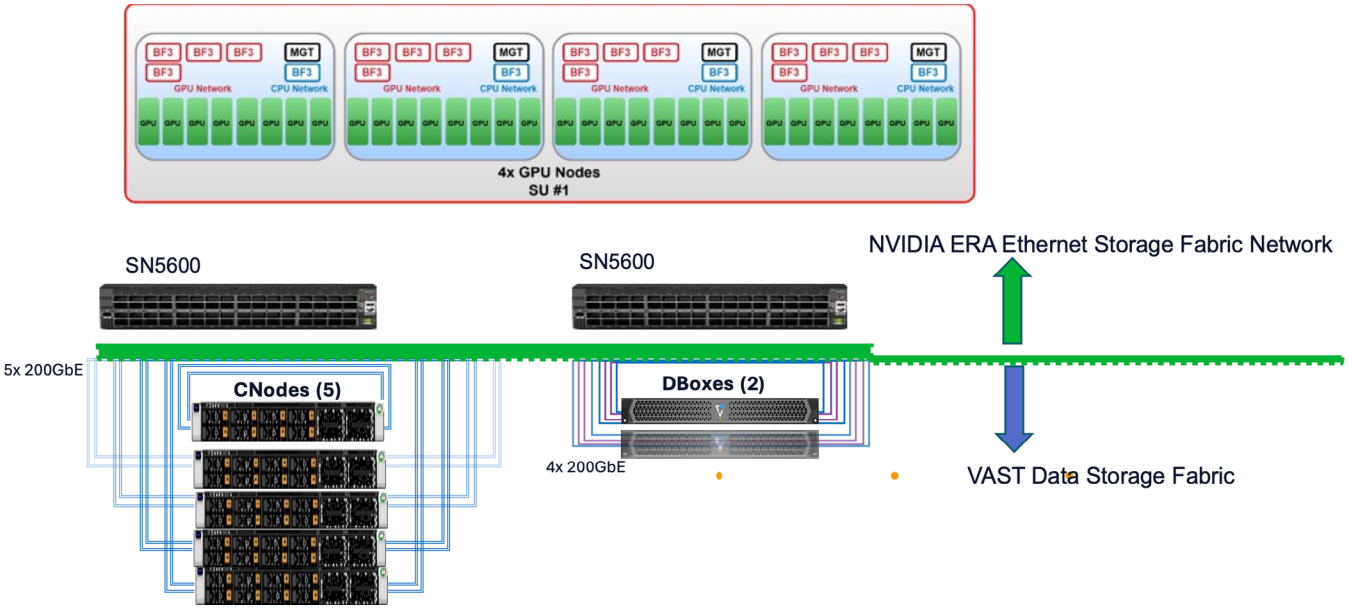
AI Training: Supports training new small-scale models or fine-tuning pre-trained models (typically models with a few million parameters) on specialized tasks or custom datasets. This capability is crucial for organizations developing domain-specific models, optimizing existing models for new data patterns, or enabling rapid iteration in R&D environments.

2-8-5-200: Foundation level

This design incorporates a flexible, rail-optimized, end-of-row network architecture, accommodates modifications to rack layouts and server densities — providing enterprises with the flexibility to customize infrastructure according to their space and workload demands.

- PCIe Optimized 2-8-5-200 (CPU-GPU-Network Adaptor-Network Bandwidth) are NVIDIA-Certified scale-out compute nodes that support eight PCIe GPUs (such as L40S, RTX PRO 6000 Server Edition, H100 NVL, H200 NVL) with five network adaptors (such as ConnectX-7, BlueField-3 SuperNIC) and two CPUs. This configuration supports scaling from four up to 64 nodes within a cluster, enabling enterprises to flexibly expand compute capacity to meet evolving AI application demands.
- Complete details of this high-performance, modular building block for compute, storage, and networking can be found in NVIDIA-Certified Storage Foundation level reference architecture design guide - 113925

NVIDIA Enterprise Reference Architecture Optimized 2-8-5 Design



Compute Counts				VAST Config		Converged Network Ports	
SUs	Nodes	GPUs	CPUs (Node N-S)	CNodes	DBox	400 Gbe Port Utilization	Mgmt Uplinks
1	4	32	2	3	1	5	5
2	8	64	4	3	1	5	5
3	12	96	6	3	1	5	5
4	16	128	8	5	2	9	9

Table 2: Scalable Configuration of 285 Design from 1 SU to 4 SU with VAST Data.

- Table 2 above provides 285 Design with VAST AI Operating System scaling based on a single SU to 8 SUs. As the scalable unit grows, VAST Data CNodes and DBoxes can be seamlessly expanded to provide additional performance and capacity scaling to meet increasing data storage capacity requirements.

The 2-8-5 Design architecture is purpose-built to support critical enterprise AI workloads, providing the compute density, networking, and storage integration required for production-grade deployments:

- AI Inference: Medium Model Parameter Workloads Empowers enterprises to deploy medium-sized pre-trained models for real-time predictions, classifications, and content generation use cases. This enables high-throughput inference for applications such as intelligent virtual assistants, recommendation engines, image/video analysis, and more.
- AI Training: Small Model Training and Fine-Tuning Supports training new small-scale models or fine-tuning pre-trained models (typically a few million parameters) on specialized tasks or custom datasets. Ideal for enterprises building domain-specific AI models, adapting foundation models to proprietary datasets, or accelerating machine learning R&D workflows.

Summary

The Storage Reference Architecture for NVIDIA provides the framework to integrate Vast Data storage solutions into a scalable Enterprise Reference architecture. This helps to build the next generation of data center scale architecture to meet the demanding and growing needs of enterprise AI. The optimized combination of all these components keeps systems running reliably, with maximum performance, and enables users to push the bounds of state-of-the-art.

The VAST AI Operating System meets the ERA architecture system performance and functionality requirements. As an enterprise object storage platform, the VAST AI Operating System, when paired with ERA, enables customers to tackle demanding AI workloads without the complexity and specialized skills typically associated with HPC storage solutions.

As requirements grow, the VAST AI Operating System can be seamlessly scaled by adding compute and/or storage resources tailored to meet performance and capacity targets. It supports multiple generations of infrastructure in a cluster, allowing customers to mitigate supply chain issues and select from a range of VAST-certified hardware solutions.

NVIDIA ERA customers can be confident that the VAST AI Operating System will meet their most challenging AI workloads

©2026 VAST Data, Inc. All rights reserved. All trademarks belong to their respective owners.

06/26 I 00000