VAST Data Storage Reference Architecture

July 2025

Table of Contents

VAST Data Partnership 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 VAST Data's alignment with the NVIDIA Enterprise Reference Architectures (Enterprise RAs), which provide guidance for deploying data center infrastructure with NVIDIA accelerated computing and NVIDIA Spectrum-X™ Ethernet networking.

By integrating the VAST Data Platform with NVIDIA Enterprise RAs, enterprises gain a scalable, high-performance data platform specifically designed to enhance Analytics, Perception AI, Generative AI, Agentic AI, and small-scale High-Performance Computing (HPC) workloads. This architecture is optimized for Generative AI initiatives, supporting Retrieval-Augmented Generation (RAG), model fine-tuning, and inference workloads, with a focus on single-tenant, Ethernet-based environments.

Organizations can further extend their AI solutions using frameworks such as the NVIDIA AI Enterprise software platform and 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 enterprise AI Factories , 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 NVIDIA and VAST partnership.

NVIDIA-Certified Storage Overview

Executive Overview

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

The VAST Data Platform is an NVIDIA-Certified™ Storage system and delivers a comprehensive solution for the enterprise market, addressing the rapidly growing data storage needs of Accelerated Analytics, Classical AI, and Generative AI workloads. NVIDIA Enterprise RAs' architecture offers deployment guidance, cluster characterization, automated provisioning through BCM, and sizing guidelines tailored for common enterprise AI implementations.

This VAST Data Storage Reference Architecture aligns with the guidance in NVIDIA Enterprise RAs and aims to showcase the modular storage building blocks that support various compute clusters, emphasizing how storage infrastructure can scale seamlessly alongside expanding computational needs. Reference configurations include PCIe-optimized and NVIDIA HGX™ platforms that adhere to prescriptive design patterns, ensuring optimal performance when deployed in cluster environments, leveraging the NVIDIA Spectrum-X™ Ethernet platform.

Enterprise RAs are built on scalable units (SUs), consisting of four compute nodes. Every SU operates as a discrete computational entity aligned with the port availability of the network fabric. Enterprises can replicate SUs to scale their deployments efficiently and cost-effectively. These design principles provide a versatile, scalable foundation for enterprise AI environments, focusing on single-tenant, Ethernet-based architectures.

Business Benefits:

- **Accelerate Time to Token:** By leveraging an NVIDIA structured approach and recommended designs, 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 (HPC) capabilities meet the demanding requirements of AI workloads, ensuring consistent and optimal performance.
- **Scale and Manageability:** Flexible scaling options and streamlined configurations allow enterprises to grow their AI infrastructure efficiently and easily manage it.
- **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 Data Platform

The VAST Data Platform underwent a series of NVIDIA storage certification tests to validate its performance and ensure optimal configuration. It exceeded stringent performance requirements with an architecture that delivers speed, scale, and operational efficiency. Designed for ease of use, it enables any IT team to seamlessly deploy and support large-scale AI initiatives with NVIDIA-Certified Storage.

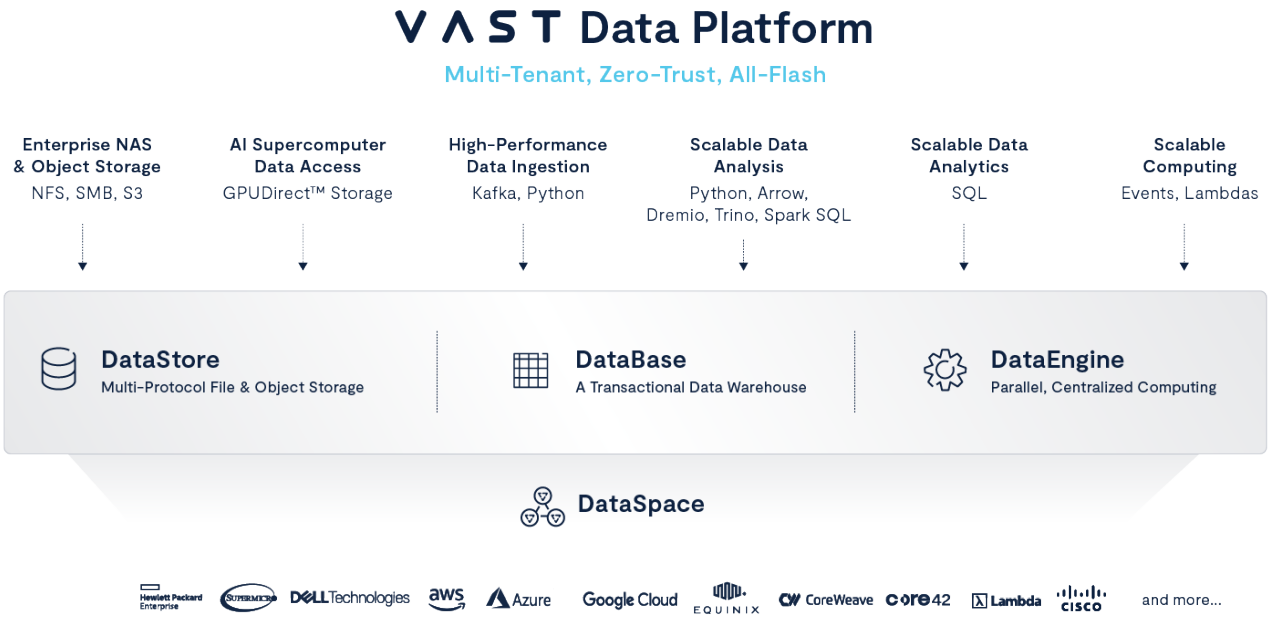


Figure 1. VAST's DASE architecture disaggregates CPUs and connects them to a globally accessible pool of SCM and ultra-dense NVMe flash SSDs.

Figure 2. Highly available NVIDIA BlueField® DPU-integrated NVMe enclosure.

VAST challenges the long-held assumption that NFS performance is inadequate for AI and HPC workloads. The VAST DASE architecture comprises two building blocks that scale across a standard NVMe fabric. First, the system's state and storage capacity are built from resilient, high-density NVMe-of storage enclosures. Second, the system's logic is implemented by stateless containers, each capable of connecting to and managing all the media in the enclosures. By disaggregating compute from storage, it is possible to distribute I/O across the system, achieving levels of parallelism for massive performance gains.

Other benefits of the VAST Data Platform 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 similarity 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.
- VAST Catalog, a built-in metadata index, enables customers to find and manage data via SQL queries.
- 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.

NVIDIA Enterprise Reference Architecture

- NVIDIA Enterprise RAs provide expert guidance for building high-performance, scalable data center infrastructure. As enterprises transform traditional data centers into AI Factories, they are revolutionizing how data is processed and analyzed, integrating advanced computing and networking technologies to meet the substantial computational demands of modern AI applications.
- NVIDIA introduced the Enterprise RAs to support this transition, 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 Enterprise RAs include prescriptive design guidance for building accelerated computing clusters, utilizing NVIDIA-Certified servers with optimized CPU-to-GPU-to-NIC-to-Bandwidth ratios. This balanced design approach helps eliminate bottlenecks and ensures consistent, high-performance outcomes across AI workloads.

Key Building Blocks

Accelerated Computing Clusters: NVIDIA provides validated design patterns for both GPU scale-out and scale-up cluster configurations with NVIDIA-Certified Systems. 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 NVIDIA HGH System). These configurations also offer scalable design guidance for expanding clusters by adding more servers, enhancing overall capacity and performance. This approach is well-suited for workloads that demand high performance from each server within the cluster.

- **HGX Reference Configuration:** This configuration is specifically designed on the **NVIDIA SXM baseboard architecture,** offering pre-validated designs for 4-GPU and 8-GPU systems, including the Hopper and Blackwell platform families. These systems are built to scale vertically up to the limits of NVLink interconnect capability. For instance, the NVIDIA GB200 NVL72 solution can scale up to 72 GPUs operating as a single, unified computational entity, delivering exceptional performance for the most demanding AI and HPC applications.

Both PCIe-optimized and SXM-based systems are designed to flexibly expand GPU and networking resources per node as workloads evolve. NVIDIA Enterprise RAs incorporate 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 integrate 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 NVIDIA 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 provides detailed guidance on network topology design, tailored for various 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, which are 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 compute nodes . 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. NVIDIA recommends deploying BlueField Data Processing Units (DPUs) to handle North-South traffic efficiently and securely. BlueField DPUs offload networking tasks from CPUs, enabling secure and efficient processing of incoming and outgoing network requests without introducing bottlenecks.

- **Switching:** NVIDIA Enterprise RAs designate **Ethernet-based switching** as the preferred networking option for enterprise AI and data workloads. Each RA provides validated Ethernet configuration recommendations to ensure consistent, high-performance switching that meets the rigorous demands of modern AI data centers.

VAST Data Storage

VAST Data storage is designed to integrate seamlessly with NVIDIA GPU compute configurations, ensuring it can meet the high-throughput demands of North-South networking while efficiently supplying compute nodes with data. This integration is critical for organizations aiming to build robust, high-performance systems that maximize the potential of NVIDIA accelerated computing with optimal storage performance.

The VAST Data Platform is fully compatible with NVIDIA GPUs and networking technologies, minimizing integration risks and deployment complexity. Its proven compatibility ensures a faster time-to-value for enterprise deployments and simplifies scaling as storage and compute needs grow.

VAST integrates seamlessly with NVIDIA accelerated computing clusters through standard Ethernet connectivity. To further boost performance, it supports RDMA over Converged Ethernet (RoCE) communication between these systems and VAST storage, enabling low-latency, high-bandwidth data transfers that are critical for AI, HPC, and data-intensive workloads.

The connection process is streamlined: Compute Nodes of the VAST Data Platform (CNodes) are linked to a pair of NVIDIA Spectrum SN5600 Ethernet switches using a standardized installation methodology. This proven approach has been successfully implemented in many deployments worldwide, ensuring operational reliability and performance consistency.

Software

NVIDIA Enterprise Reference Architectures offer a bare-metal foundation optimized for performance, reliability, and scalability across a wide range of AI workloads. Designed for flexibility, these Enterprise RAs integrate seamlessly with industry-standard orchestration tools such as Kubernetes and Slurm, enabling efficient cluster management and workload scheduling.

Beyond infrastructure, Enterprise RAs also establish the foundation for enterprise-grade AI software deployments. Systems built with Enterprise RAs can support comprehensive software stacks, including NVIDIA AI Enterprise, which includes NVIDIA NIM™ microservices to accelerate AI development and deployment.

NVIDIA Omniverse enables advanced simulation, collaboration, and digital twin applications. By leveraging Enterprise RAs, organizations can deploy full-stack AI solutions—from infrastructure to application layer—ensuring fast, reliable, and scalable AI innovation.

By leveraging these Enterprise RAs, organizations can deploy full-stack AI solutions—from the infrastructure to the application layer—ensuring fast, reliable, and scalable AI innovation.

4 GPU System Configuration (2-4-3-200)

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 the L40S and H100 NVL)
- 3 high-performance network adapters/SuperNICs (including ConnectX®-7 and BlueField-3 SuperNICs)
- 200 GbE per GPU of E-W network bandwidth

This Enterprise Reference Architecture (RA) utilizes PCIe based systems, providing robust AI inference, training, and visualization capabilities to enhance the next generation of AI-enabled enterprise workloads in the data center.

The modular 2-4-3-200 design can be deployed using a four-node scalable unit (SU) and can scale up to 32 nodes in a cluster, totaling up to 128 GPUs. The flexible, rail-optimized end-of-row network architecture supports adaptable rack layouts and server densities, offering enterprises the ability to adjust configurations based on space, scaling needs, and workload demands.

2-4-3-200 System Configuration

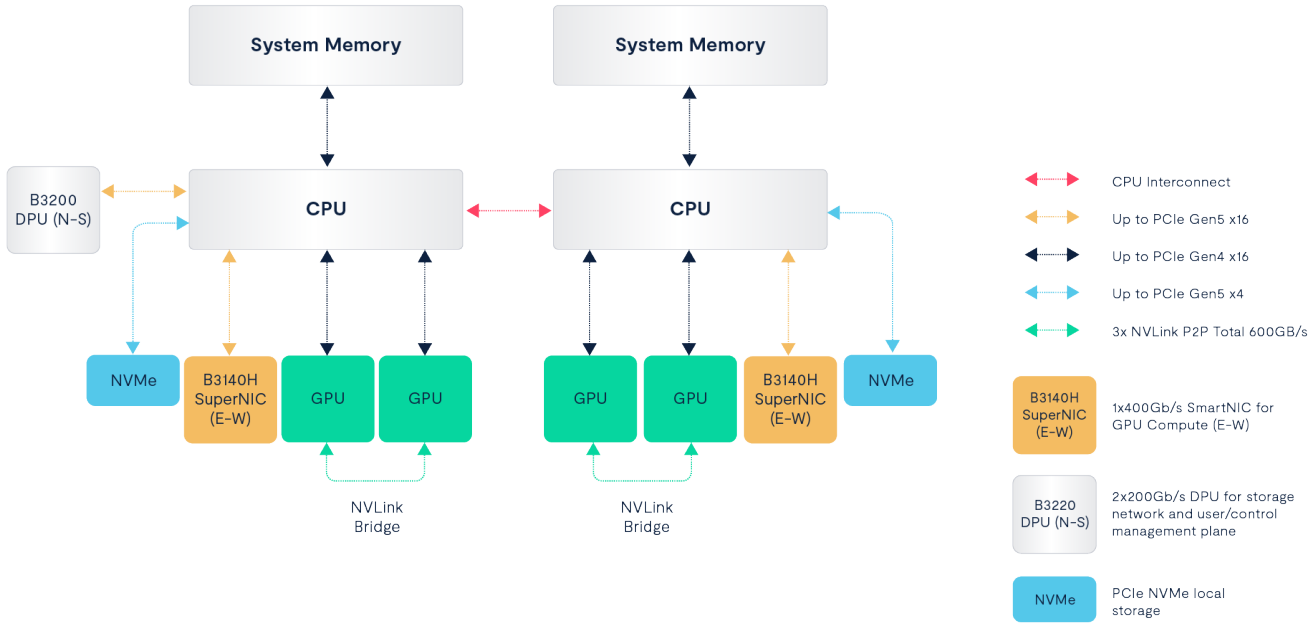


Figure 3 Example diagram of 4-Node 4-GPU 3-NIC SU

Compute Counts				VAST Data Storage			Converged Network Ports			
SUs	Nodes	GPUs	CPUs (Node N-S)	CNodes	DBox	800 Gbe Port Utilization	Storage	Mgmt Uplinks	Customer	Support
1	4	16	2	3	1	2	4	4	4	4
2	8	32	4	3	1	2	4	4	4	4
3	12	48	6	3	1	2	4	4	4	4
4	16	64	8	5	2	3	4	4	4	4
6	24	96	12	5	2	3	4	4	4	4
8	32	128	16	5	2	3	4	4	8	4

Table 1: Scalable Configuration of 243 Design from 1 SU to 8SU with VAST Data

The 4-Node 1 scalable unit (SU) provides a high-performance, modular building block for computing, storage, and networking. Each SU is designed with the connectivity building blocks mentioned below:

- For the Compute (E-W) fabric, four servers, each equipped with four B3140H SuperNICs, provide sixteen 400Gb/s connections and a total aggregate bandwidth of 6.4Tb/s. This setup enables ultra-high-bandwidth, low-latency communication between compute nodes to support AI model training and inference at scale.
- For the Converged (N-S) fabric: four servers, each with one 1x B3220 DPU providing eight 200Gb/s connections and a total aggregate bandwidth of 1.6 Tb/s. It supports external data movement, storage access, and secure north-south traffic processing.
- For the Out-of-Band (OOB) Management fabric, four servers, each with six 1Gb/s connections, provide 48 1Gb/s connections for management.
- Table 1 above provides a 243 Design with VAST Data Platform scaling from a single SU to eight SUs. As the scalable unit expands, VAST Data CNodes (Compute Nodes) and DBoxes (Data Nodes) can be seamlessly augmented to deliver additional performance scaling and meet increasing AI and enterprise workload demands.

NVIDIA Enterprise Reference Architecture

Optimized 2-4-3 Design

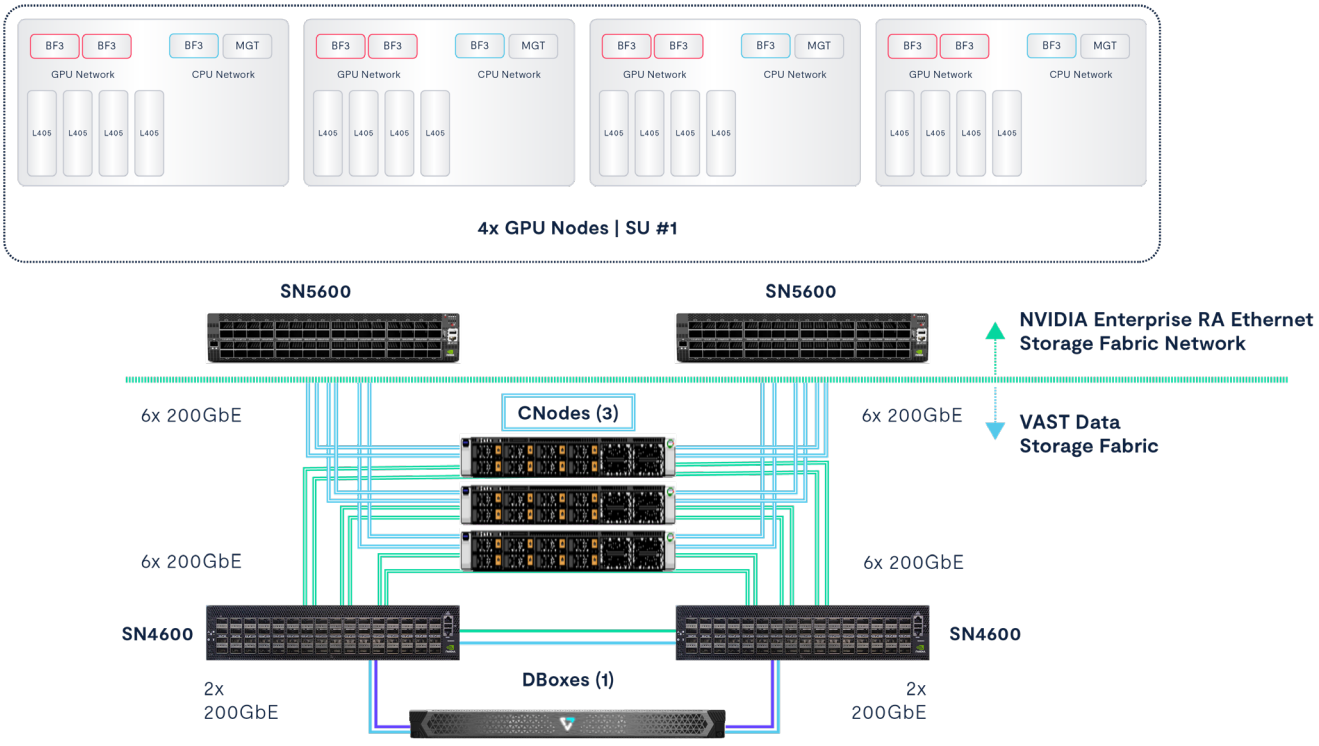


Figure 4 VAST Data Integration with 2-4-3-200 Design Pattern

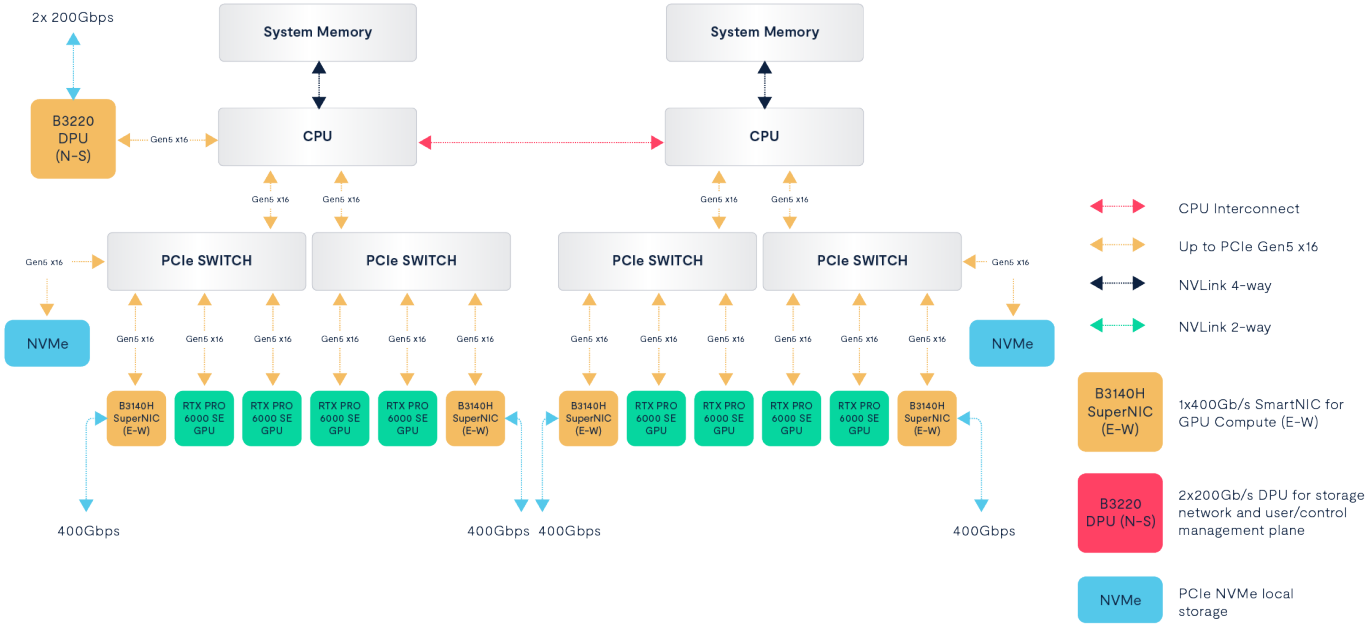
The key important Use Cases for Enterprises with 2-4-3-200 design patterns are:

- **Visual Computing** 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 Computer Aided Design (CAD) for industries such as automotive, aerospace, and architecture, it enables detailed modeling, prototyping, and simulation workflows. CAD applications are typically found in the automotive, aerospace, and architecture industries.
  - **Retail and E-commerce:** Facilitates virtual product modeling and augmented reality (AR) shopping experiences, improving customer engagement and speeding up purchasing decisions.
  - **Healthcare:** Powers 3D medical imaging, diagnostic modeling, and surgical simulation platforms, enhancing diagnostic accuracy and supporting advanced medical training.
  - **Media and Entertainment:** Utilized 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. It is ideal for real-time and batch inference use cases where medium-scale model sizes (typically hundreds of millions to a few billion parameters) provide high accuracy while maintaining manageable compute and latency requirements.
- **AI Training:** Supports training of new small-scale models or fine-tuning of 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.

8 GPU System Configuration (2-8-5-200)

The PCIe Optimized 2-8-5-200 (CPU-GPU-Network Adapte r-Network Bandwidth) are NVIDIA-certified compute nodes that support up to eight PCIe GPUs (such as the NVIDIA RTX PRO 6000 Blackwell Server Edition and H200 NVL), five network adapters/SuperNICs (such as ConnectX-7, BlueField-3 SuperNIC), and two CPUs. This configuration allows scaling from four to 32 nodes within a cluster, enabling enterprises to flexibly expand compute capacity to meet evolving AI application demands. The design features a flexible, rail-optimized, end-of-row network architecture that accommodates modifications to rack layouts and server densities, offering enterprises the ability to customize infrastructure according to their space and workload demands.



NVIDIA Enterprise Reference Architecture

Optimized 2-8-5 Design

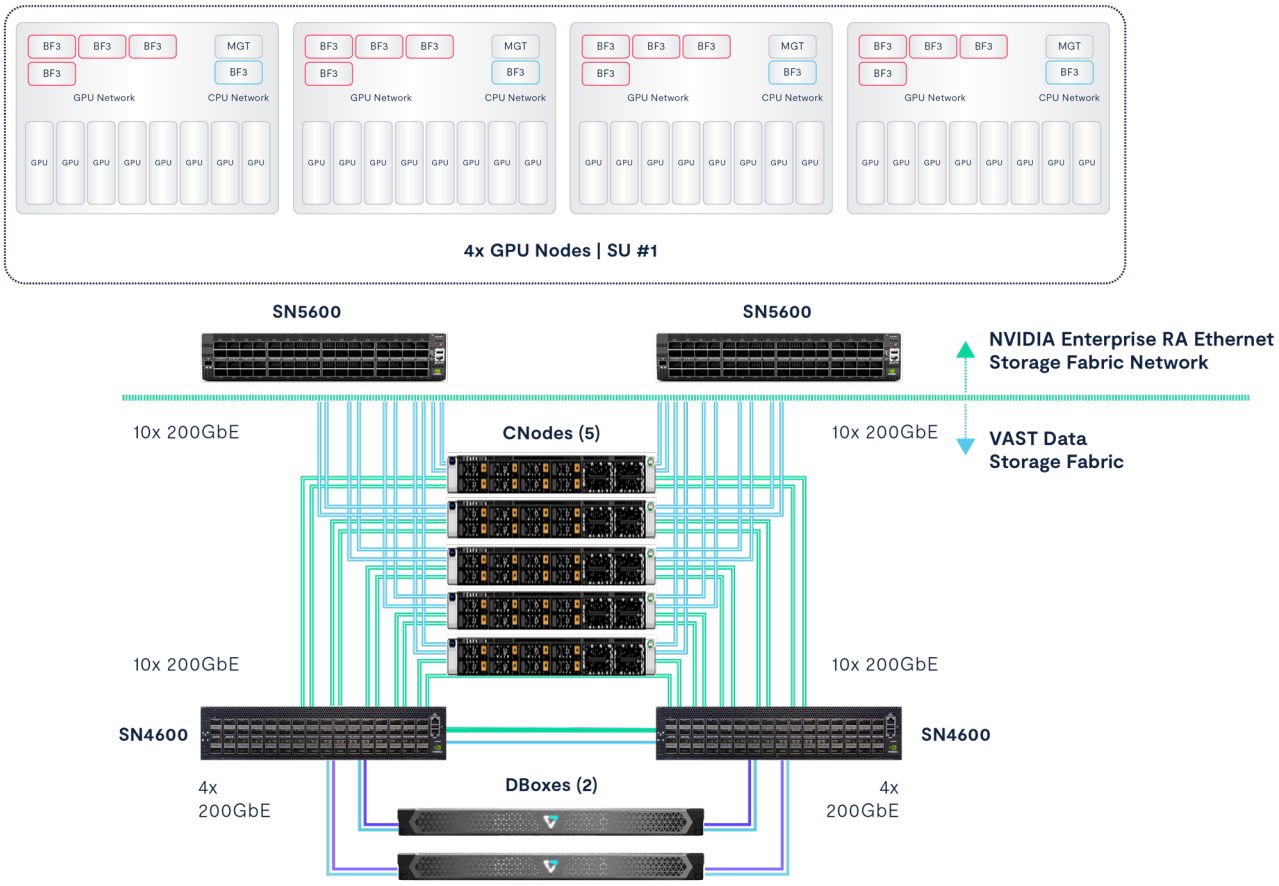


Figure 5 VAST Data Integration with 285 Design

Compute Counts				VAST Data Storage			Converged Network Ports			
SUs	Nodes	GPUs	CPUs (Node N-S)	CNodes	DBox	800 Gbe Port Utilization	Storage	Mgmt Uplinks	Customer	Support
1	4	32	2	3	1	2	4	4	4	4
2	8	64	4	3	1	2	4	4	4	4
3	12	96	6	3	1	2	4	4	4	4
4	16	128	8	5	2	3	4	4	8	4
6	24	192	12	5	2	3	4	4	8	4
8	32	256	16	5	2	3	4	4	8	4

Table 2: Scalable Configuration of 285 Design from 1 to 8 SU with VAST Data.

The 4-Node Scalable Unit (SU) for the 2-8-5 Design provides high-performance building blocks for compute, storage, and management networking. The connectivity components are organized as follows:

- For the Compute (E-W) fabric: 4 servers, each with 4x B3140H SuperNICs, providing 16x 400Gb/s connections and a total aggregate bandwidth of 6.4Tb/s
- For the Converged (N-S) fabric: 4 servers, each with 1x B3220 DPU providing 8x 200Gb/s connections and a total aggregate bandwidth of 1.6Tb/s
- For the OB Management fabric, 4 servers, each with 6x 1Gb/s connections providing 48x 1Gb/s for management
- Table 2 above presents the 285 Design with VAST Data Platform scaling from a single SU to 8 SUs. As the number of scalable units increases, VAST Data CNodes and DBoxes can be seamlessly expanded to deliver enhanced performance and accommodate growing data storage capacity requirements.

The 2-8-5 Design architecture is specifically designed to support critical enterprise AI workloads, offering the necessary compute density, networking, and storage integration for production-grade deployments:

- **AI Inference:** Medium Model Parameter Workloads empower 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 and video analysis, and more.
- **AI Training:** Small Model Training and Fine-Tuning facilitates training of new small-scale models or fine-tuning of pre-trained models (typically a few million parameters) on specialized tasks or custom datasets. It is ideal for enterprises developing domain-specific AI models, adapting foundational models to proprietary datasets, or accelerating machine learning R&D workflows.

8 GPU System Configuration (2-8-9-400): Enterprise Level Exclusively

The NVIDIA HGX ™ GPUs and NVIDIA Spectrum-X Networking Enterprise RA is optimized for multi-node AI or hybrid applications. This modular architecture is based on NVIDIA-Certified HGX systems s, each equipped with eight NVIDIA H100, H200, or B200 SXM GPUs. Using a four-node scalable unit (SU), this can scale up to 32 nodes , totaling 256 GPUs. Fully tested systems can scale to thirty-two SUs, with the potential for larger clusters based on customer requirements. The flexible, rail-optimized end-of-row network architecture accommodates modifications in rack layout and the number of servers per rack.

Design pattern 2-8-9-400 (CPU-GPU-Network Adapte r-Network Bandwidth) consists of NVIDIA-Certified NVIDIA HGX systems configured 2 CPUs, 8 GPUs, plus 9 NICs with East-West traffic of 400 GbE per GPU s

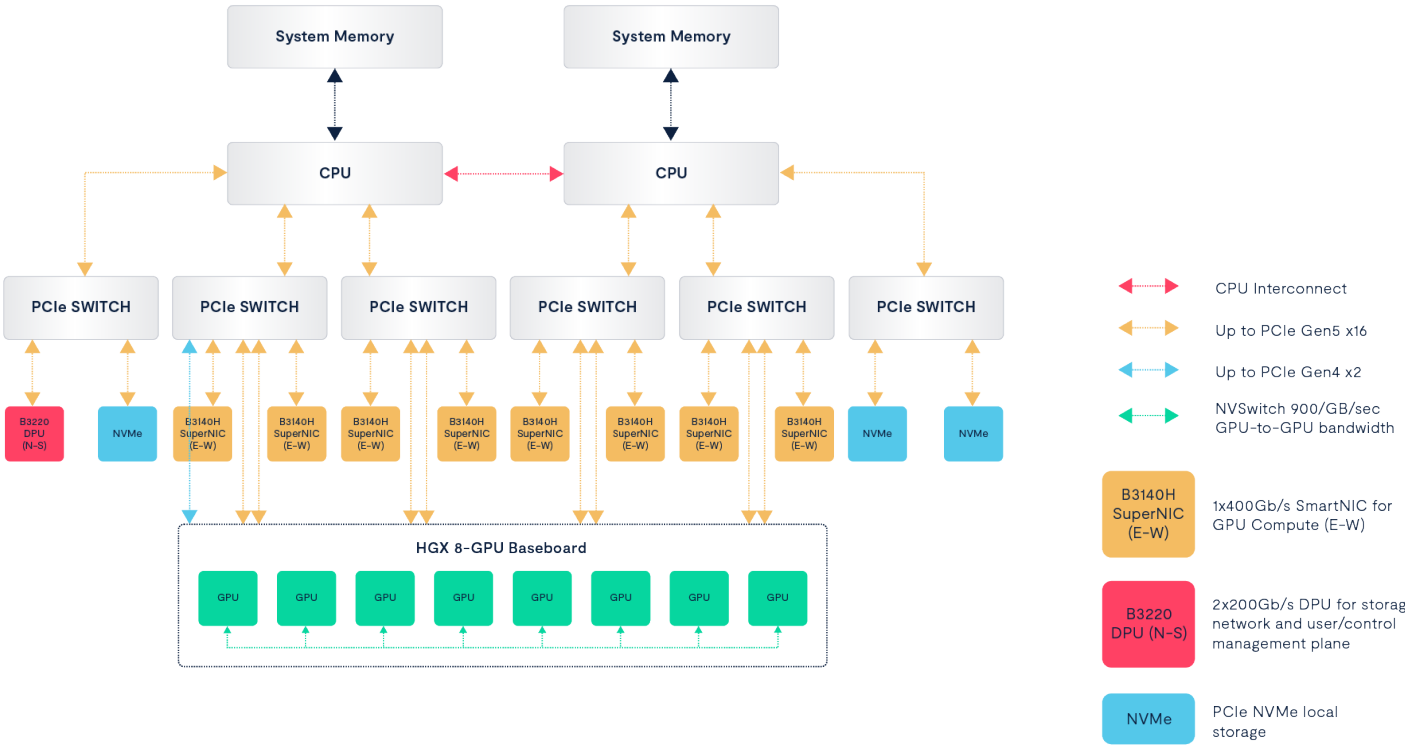


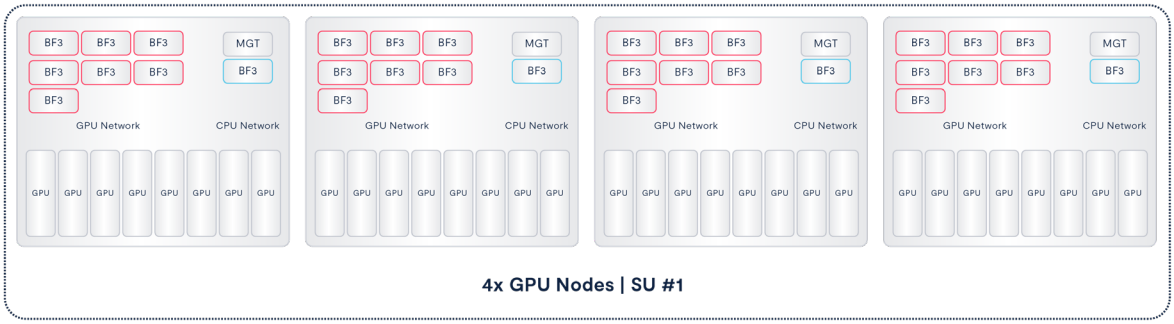
Figure 4 VAST Data Integration with 289 Architecture

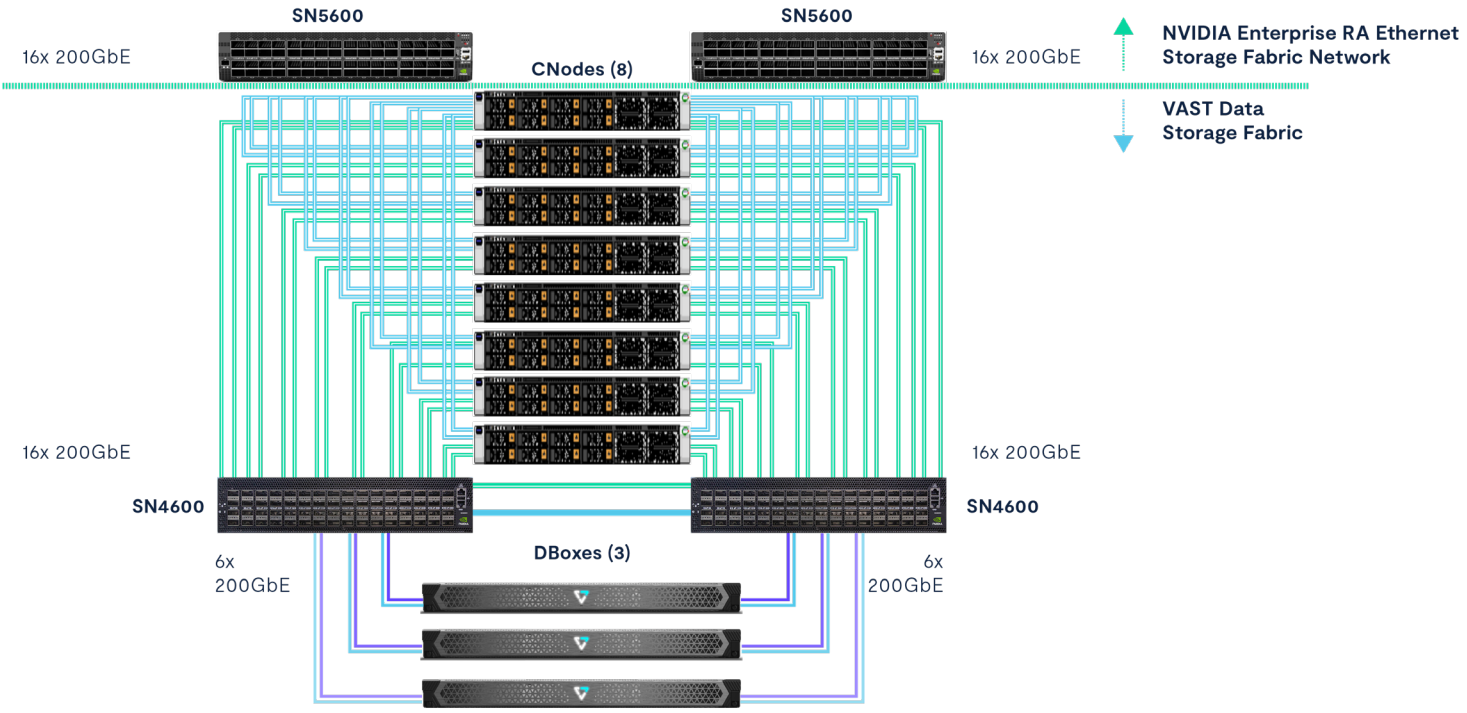
Compute Counts				VAST Data Storage			Converged Network Ports			
SUs	Nodes	GPUs	CPUs (Node N-S)	CNodes	DBox	800 Gbe Port Utilization	Storage	Mgmt Uplinks	Customer	Support
1	4	32	2	3	1	4	1	4	4	4
2	8	64	4	3	1	4	1	4	4	4
3	12	96	6	3	1	4	1	4	8	4
4	16	128	8	5	2	4	1	4	8	4
6	24	192	12	5	2	8	1	4	12	4
8	32	256	16	5	2	8	1	4	16	4
12	48	384	24	8	3	12	2	8	24	4
16	64	512	32	10	4	16	2	8	32	4
24	96	768	48	10	4	12	2	12	48	4
32	128	1024	64	13	5	16	3	16	64	4

Table 3: Scalable Configuration of 285 Design from 1 SU to 32 SU with VAST Data

NVIDIA Enterprise Reference Architecture

Optimized 2-8-9 Design





The 4-Node scalable unit provides the following connectivity building blocks:

- For the Compute (E-W) fabric: 4 servers, each with 8x B3140H SuperNICs, offering 32x 400Gb/s connections and a total aggregate bandwidth of 12.8Tb/s.
- For the Converged (N-S) fabric: four servers, each with one 1x B3220 DPU providing eight 200Gb/s connections and a total aggregate bandwidth of 1.6Tb/s.
- For the OOB Management fabric, four servers, each with twelve 1Gb/s connections, providing a total of forty-eight 1Gb/s for management.
- Table 3 above presents 289 Design with the VAST Data Platform scaling from a single SU to 8 SUs. As the scalable unit increases, VAST Data CNodes and DBoxes can be effortlessly expanded to enhance performance and accommodate the growing data storage capacity needs.

The 2-8-9-400 Design is engineered to support the most demanding enterprise AI workloads, providing exceptional compute density, high bandwidth, and flexible scalability:

- **AI Inference:** Large (per node) and medium (per GPU) model parameter inference workloads enable enterprises to deploy large-scale, multi-model architectures for multi-modal content generation use cases—such as image synthesis, video generation, text-to-image, and audio-visual models. Each node can host large, aggregated models across multiple GPUs while still supporting high-throughput inference at the individual GPU level.
- **AI Training:** Large to Small model training and fine-tuning on cluster sizing. It supports the training of both large-scale and small to medium-sized models, providing flexibility according to cluster size and enterprise needs. Enterprises can train completely new models or fine-tune pre-trained foundation models on custom datasets for specialized tasks across industries such as healthcare, finance, manufacturing, and media.

Summary

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

The NVIDIA-Certified VAST Data Platform meets the NVIDIA Enterprise Reference Architecture system performance and functionality requirements. As an enterprise NAS solution, the VAST Data Platform, when combined with Enterprise RAs, allows customers to address demanding AI workloads without the complexity and specialized skills typically associated with HPC storage solutions.

As requirements grow, the VAST Data Platform can be easily scaled by adding compute and/or storage resources designed to meet performance and capacity targets. It supports multiple generations of infrastructure in a cluster, enabling customers to alleviate supply chain issues and choose from a variety of VAST-certified hardware solutions.

NVIDIA Enterprise RA customers can trust that the VAST Data Platform will scale to handle their most demanding AI workloads.