



# NVIDIA Enterprise Reference Architecture Overview

Whitepaper

# Table of Contents

<b>Building AI Factories for the Enterprise</b> .....	<b>3</b>
<b>Introducing the NVIDIA Enterprise Reference Architectures</b> .....	<b>4</b>
Methodology for Bringing Reference Architectures to Market.....	4
NVIDIA Enterprise Reference Architecture.....	5
NVIDIA Cloud Partner (NCP) Reference Architecture.....	6
<b>Key building blocks of Enterprise Reference Architectures</b> .....	<b>7</b>
Accelerated Computing Clusters.....	7
Networking.....	11
Storage.....	12
Software.....	13
Why This Matters.....	13
Business Benefits.....	13
IT Benefits.....	14
<b>Appendix: A</b> .....	<b>15</b>
NVIDIA Enterprise Reference Architecture: NVIDIA L40S and NVIDIA Spectrum Platforms	15
Use Cases.....	15
NVIDIA OVX L40S Reference Configurations.....	15
<b>Appendix: B</b> .....	<b>17</b>
NVIDIA Enterprise Reference Architecture: NVIDIA H100 NVL and NVIDIA Spectrum	
Platforms.....	17
Use Cases.....	17
NVIDIA H100 NVL Reference Configurations.....	17
<b>Appendix: C</b> .....	<b>19</b>
NVIDIA Enterprise Reference Architecture: NVIDIA HGX H100/H200/B200 and NVIDIA	
Spectrum-X Networking Platform.....	19
Use Cases.....	19
NVIDIA HGX H100/H200/B200 Reference Configurations.....	19
<b>Appendix: D</b> .....	<b>21</b>
NVIDIA Enterprise Reference Architecture: NVIDIA H200 NVL or RTX™ PRO 6000	
Blackwell Server Edition and NVIDIA Spectrum-X Networking Platform.....	21
Use Cases.....	21
NVIDIA H200 NVL Systems.....	21
NVIDIA RTX PRO 6000 Blackwell Server Edition Systems.....	22

---

# Building AI Factories for the Enterprise

Traditional IT infrastructures are straining under the demanding requirements of Gen AI. Data size and model complexity have increased exponentially, while CPUs processing power has only doubled, according to Moore's Law. Despite the rapid advancement in computing capabilities, it's outpaced by the computing demands of enterprises in this new AI era. General-purpose computing has reached its limits, highlighting the tipping point in favor of accelerated computing platforms. A new approach is necessary—one that's specifically designed to harness the power of AI. This is where the AI Factory comes in. Like physical factories that powered the industrial revolution, AI Factories will drive the AI revolution, but instead of producing physical goods from raw materials, AI Factories will transform data and electricity into intelligence and tokens with great scale and efficiency. Every business will need an AI Factory to deliver fast, repeatable, flexible, and efficient outcomes.

The shift to AI Factories isn't without its challenges. As traditional data centers evolve, businesses must effectively modernize and build their own mini-supercomputers, which can be quite complex, time-consuming, and resource-intensive. Enterprises face significant challenges in designing, scaling, and optimizing these systems, often requiring specialized expertise and substantial investment. Integrating accelerated computing platforms with energy-efficient designs to manage the increased heat and power consumption plus planning for future advanced cooling solutions adds to the complexity. Building these systems from scratch can take years, especially when coupled with interoperability challenges.

Ultimately, enterprises are looking for simplified guidance to get to value faster and maximize the ROI of their AI investments. By adopting AI Factories and leveraging expert guidance, businesses can streamline their AI infrastructure, reduce complexity, and drive significant business value efficiently, transforming operations and driving a new era of innovation and competitive advantage.

---

# Introducing the NVIDIA Enterprise Reference Architectures

NVIDIA developed Enterprise Reference Architectures (Enterprise RAs) to provide clear and consolidated recommendations for NVIDIA system partners and our joint enterprise customers building AI Factories. By bringing the same technical components from the supercomputing world and packaging them with design recommendations based on decades of experience, NVIDIA's goal is to eliminate the burden of building these systems from scratch with a streamlined approach for flexible and cost-effective configurations, taking the guesswork and risk out of deployment. This ensures customers have the best experience in terms of performance, utilization, uptime, total cost of ownership (TCO), and supportability. Ultimately, this helps our partners and joint customers to achieve value sooner and maximize the return on their investments.

NVIDIA architects have gained extensive knowledge from the many hours of testing to determine best practices for configuring systems to maximize performance and create a baseline of performance standards. NVIDIA Enterprise RAs represent the shared learnings of common design patterns to help customers avoid the pitfalls NVIDIA experts have encountered with guidance on delivering well-balanced systems in which bottlenecks caused by individual components are minimized. This enables partners and enterprise customers to confidently deliver AI solutions faster, allowing them to focus on running their business rather than fighting deployments. Whether you choose to implement a full-fledged data center using our guidelines or adapt the node configurations with your own networking, these NVIDIA Enterprise RAs provide an invaluable starting point.

**Example of shared learning:** Transceivers and cabling are paramount in large-scale implementations, and getting these wrong can have a major impact on delivery times and customer experience. An NVIDIA partner deviated from our design recommendations despite our advice to follow the reference architecture. To save costs, they opted for copper cables instead of the recommended transceivers. While copper cables are suitable for many installations, our engineers had previously encountered a heat issue with this configuration at scale. Unfortunately, the partner did not heed our warning and subsequently faced the same heat wall, leading to unnecessary struggles and a longer, more expensive deployment for the customer. Sharing these types of learnings and insights is crucial so that customers understand the potential impact of their design choices as they scale. This helps our partners deliver reliable solutions faster and ensures customers are happy and successful.

## Methodology for Bringing Reference Architectures to Market

NVIDIA has a highly structured approach to introducing reference architectures for new technologies, such as NVIDIA Hopper GPUs, Blackwell GPUs, Grace CPUs, Spectrum-X networking platform, and BlueField architecture-based offerings. For each new technology, NVIDIA provides configuration guides

to assist partners in designing, building and deploying optimized system configurations. Partners then can submit these systems for certification through the [NVIDIA-Certified systems](#) program. This program involves rigorous testing, including thermal analysis, mechanical stress tests, power consumption evaluations, and signal integrity assessments to ensure the components function optimally within the server design. Furthermore, an NVIDIA-Certified server must pass a comprehensive suite of performance tests covering various workload categories, networking capabilities, security features, and management functionalities across a wide range of applications and use cases.

NVIDIA offers two key reference architecture programs that leverage NVIDIA-Certified servers: NVIDIA Enterprise Reference Architectures and the [NVIDIA Cloud Partner \(NCP\) Reference Architecture](#).

## NVIDIA Enterprise Reference Architecture

NVIDIA Enterprise Reference Architectures are tailored for enterprise-class deployments, ranging from 32 to 256 GPUs. Depending on the base technology, they include configurations for 4 up to 32 nodes, complete with the appropriate networking topology, switching, and allocations for storage and control plane nodes. Derived from the NCP Reference Architecture but right-sized for enterprise-scale deployments, it provides deployment guides, cluster characterization, provisioning automation using BCME, and sizing guides for common enterprise AI implementations. NVIDIA Enterprise RAs are designed to support a diverse range of workloads, including fine-tuning, Retrieval-Augmented Generation (RAG), model training, inference, and small-scale High-Performance Computing (HPC) tasks. These designs provide a versatile foundation for enterprise AI with a focus on single-tenant, Ethernet-based environments.

Each reference architecture is designed around an NVIDIA-Certified server that follows a prescriptive design pattern, called a Reference Configuration, to ensure optimal performance when deployed in a cluster. The Reference Configurations standardize the description of compute nodes based on their CPU, GPU, Network, and Bandwidth configurations. The C-G-N-B nomenclature simplifies system selection by clearly defining compute power, networking capabilities, and bandwidth performance where each digit (ex: 2-8-5-200) refers to the ratio of # of sockets (CPUs) - # of GPUs - # of network adaptors (NICs)- average East-West network bandwidth per GPU (GbE) respectively.

With GPU and networking advancements on the horizon, these architectures ensure scalability and future-proofing for enterprise applications.

Table 1 Examples of Reference Configuration Node and Networking Patterns

C-G-N-B Configuration	Description
2-8-9-400	2 CPUs, 8 GPUs, 9 NICs (1 North/South, 8 East/West), 400 GbE per GPU
2-2-3-400	2 CPUs, 2 GPUs, 3 NICs (1 North/South, 2 East/West), 400 GbE per GPU
2-8-5-200	2 CPUs, 8 GPUs, 5 NICs (1 North/South, 4 East/West), 200 GbE per GPU

## NVIDIA Cloud Partner (NCP) Reference Architecture

The existing NCP Reference Architecture is designed for larger-scale foundational model training, starting from 128 nodes and scaling up to over 16,000 GPUs. These architectures are derived from larger-scale HPC and superclusters, which can range up to 100,000 GPUs. The NCP Reference Architecture predominantly relies on NVIDIA's professional services for deployment. It is intended for use cases such as large language model foundational training and GPU-as-a-service, where a large number of GPUs are made available for customers to rent. This architecture supports both single and multi-tenant environments and uses both InfiniBand and Ethernet. The design is more rigid to ensure that deployments by NVIDIA's professional services do not require additional engineering time for reconfiguration at customer sites. By adhering to reference architecture, NVIDIA can ensure a consistent and efficient deployment process.

These two reference architecture programs provide a robust starting point for NVIDIA partners and our joint customers, helping them avoid common pitfalls and achieve faster, more efficient AI deployments.

*Note: The rest of this paper will be focused on the NVIDIA Enterprise RA program. For larger-scale solutions, please reference the NCP program.*

---

# Key building blocks of Enterprise Reference Architectures

Each NVIDIA Enterprise Reference Architecture includes deployment guidelines for multiple workloads, supporting flexible cluster sizing, networking, and expansion needs. Our Enterprise RAs encompass infrastructure and optimized server networking configurations based on common design patterns for CPU, GPU, and networking. The key components of an Enterprise RA include the accelerated computing clusters, East-West networking, North-South networking, and switches.

## Accelerated Computing Clusters

NVIDIA Enterprise RAs include design recommendations for building accelerated computing clusters using NVIDIA-Certified servers with balanced CPU to GPU to NIC patterns to avoid potential bottlenecks and sub-optimal performance. NVIDIA delivers design patterns for both GPU scale-up and scale-out configurations leveraging NVIDIA® NVLink® and for high-speed, multi-GPU communication facilitating all-to-all GPU communication.

**PCIe-Optimized Reference Configurations:** Enterprise RAs based upon PCIe-Optimized Reference Configurations utilize NVLink technology within individual servers (when supported, ex: NVIDIA H100/H200 NVL) while providing guidelines for scaling out the cluster with additional servers to enhance overall capacity and performance. This approach is ideal for workloads demanding high performance from each server within the cluster.

These scale-out systems employ optimized PCIe technology, allowing for flexible expansion of GPUs and networking capabilities as needed for each node. The NVIDIA Enterprise RAs follow Reference Configuration recommendations for balancing CPU, GPU, and Network Interface Card (NIC) configurations to prevent potential bottlenecks and ensure optimal performance. This balanced approach enables organizations to scale their computational resources efficiently, meeting the demands of complex, distributed workloads while maintaining high performance across the entire cluster.

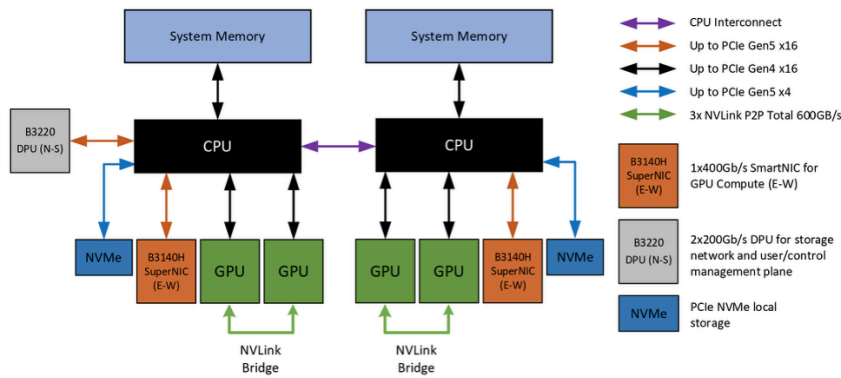
- PCIe-Optimized 2-4-3-200 (CPU-GPU-NIC-Bandwidth) Reference Configurations are for 2U NVIDIA-Certified compute nodes using PCIe allowing you to deploy 2 CPUs balanced with up to 4 GPUs plus 3 NICs with East-West traffic of 200 GbE per GPU. This pattern can scale from 8 up to 32 nodes in a cluster.
  - Eligible GPU: NVIDIA H100 NVL and L40S
  - Eligible CPU: AMD EPYC™ Processors: Milan, Genoa, Turin; Intel® Xeon® Scalable Processors: Sapphire Rapids, Emerald Rapids, Granite Rapids
  - East-West Networking: NVIDIA BlueField-3 B3140H or NVIDIA ConnectX-7

- North-South Networking: NVIDIA BlueField-3 B3220

Figure 1. Below is an example of pattern 2-4-3-200 from the NVIDIA H100 NVL and NVIDIA Spectrum Platforms Enterprise RA which is designed for a variety of use cases, including visual computing for 3D graphics and rendering, AI inference for medium model parameter workloads, and AI training for small model training and fine-tuning. Refer to Appendix A for more details.

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

NVIDIA H100 NVL & L40S Accelerated Computing Platforms



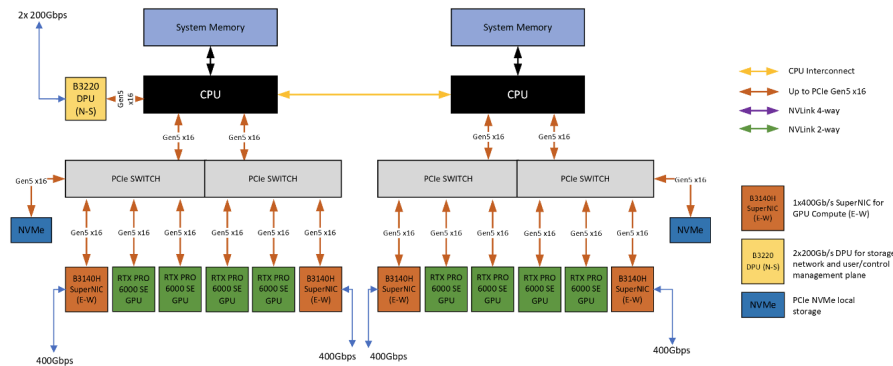
- **PCIe-Optimized 2-8-5-200 (CPU-GPU-NIC-Bandwidth) Reference Configuration** are for 4U NVIDIA-Certified compute nodes using PCIe allowing you to deploy 2 CPUs balanced with up to 8 GPUs plus 5 NICs with East-West traffic of 200 GbE per GPU. This pattern scales from 4 up to 32 nodes in a cluster.
  - Eligible GPU: NVIDIA Blackwell RTX™ PRO 6000 Server Edition and H200 NVL
  - Eligible CPU: AMD EPYC™ Processors: Milan, Genoa, Turin; Intel® Xeon® Scalable Processors: Sapphire Rapids, Emerald Rapids, Granite Rapids
  - East-West Networking: NVIDIA BlueField-3 B3140H or NVIDIA ConnectX-7
  - North-South Networking: NVIDIA BlueField-3 B3220

Figure 2. Below is an example of a PCIe-Optimized 2-8-5-200 Reference Configuration based upon the NVIDIA H200 NVL or RTX™ PRO 6000 Blackwell Server Edition and NVIDIA Spectrum Platforms Enterprise RA which is designed for AI inference for large to medium model parameter workloads and

AI training and fine-tuning. Refer to Appendices B and D for more details.

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

NVIDIA H200 NVL and RTX™ PRO 6000 Blackwell Server Edition



**HGX Reference Configurations:** With Enterprise RAs based upon HGX Reference Configurations, NVLink connections can be extended to create seamless, high-bandwidth, multi-node GPU clusters. Deploying compute clusters with these scale-up systems effectively transforms a data center into a single, massive GPU, capable of handling workloads distributed across multiple servers. The fifth-generation NVLink technology addresses the escalating need for high-speed scale-up interconnects in GPU clusters, supporting up to 576 GPUs. This advanced interconnect is essential for rapidly feeding extensive datasets into models and facilitating swift data exchange between GPUs.

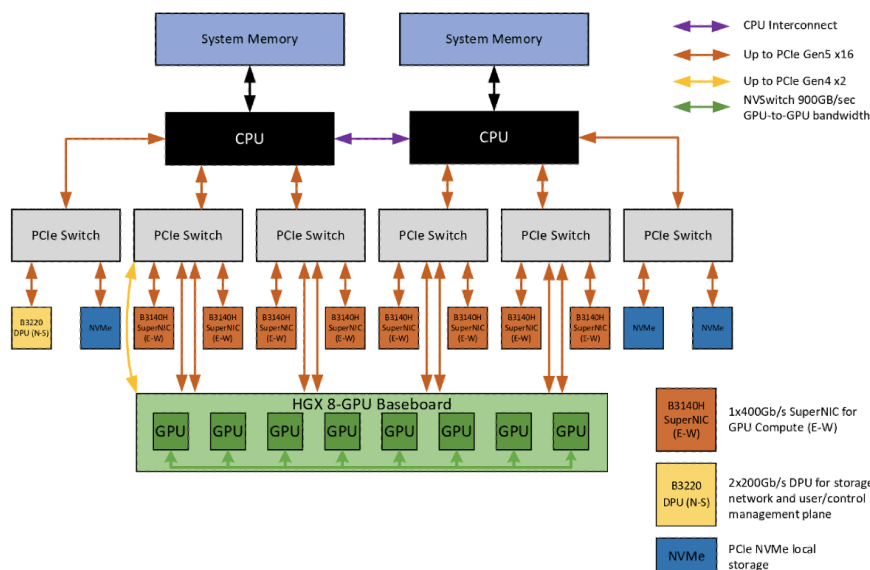
These NVIDIA Enterprise RAs are specifically designed for NVIDIA SXM baseboard architectures, offering predefined configurations for 4-GPU and 8-GPU systems, such as HGX and Grace class offerings. These systems can scale up to their maximum corresponding NVLink capability. For example, the GB200 NVL72 can scale up to 72 GPUs, functioning as a single, cohesive unit, providing unprecedented computational power for complex AI and HPC workloads.

- **HGX 2-8-9-400 (CPU-GPU-NIC-Bandwidth) Reference Configuration** is for scale-up NVIDIA 8-GPU HGX NVIDIA Certified servers with 2 CPUs balanced with 8 GPUs plus 9 NICs with East-West traffic of 400 GbE per GPU. This pattern scales from 4 to 32 nodes in a cluster.
  - Eligible GPU: NVIDIA HGX™ H100, H200, or B200
  - Eligible CPU: AMD EPYC™ Processors: Milan, Genoa, Turin; Intel® Xeon® Scalable Processors: Sapphire Rapids, Emerald Rapids, Granite Rapids
  - East-West Networking: NVIDIA BlueField-3 B3140H/L or NVIDIA ConnectX-7
  - North-South Networking: NVIDIA BlueField-3 B3220

Figure 3. Below is an example of the HGX 2-8-9-400 Reference Configuration from the NVIDIA HGX H100/H200/B200 and NVIDIA Spectrum-X Networking Platform Enterprise RA which is designed for AI inference with large (per node) and medium (per GPU) model parameter workloads, as well as AI training for large-to-small model training and fine-tuning based on cluster sizing. Refer to Appendix C for more details.

## 8 GPU System Configuration (2-8-9-400)

NVIDIA HGX™ H100, H200, or B200 Accelerated Computing Platforms



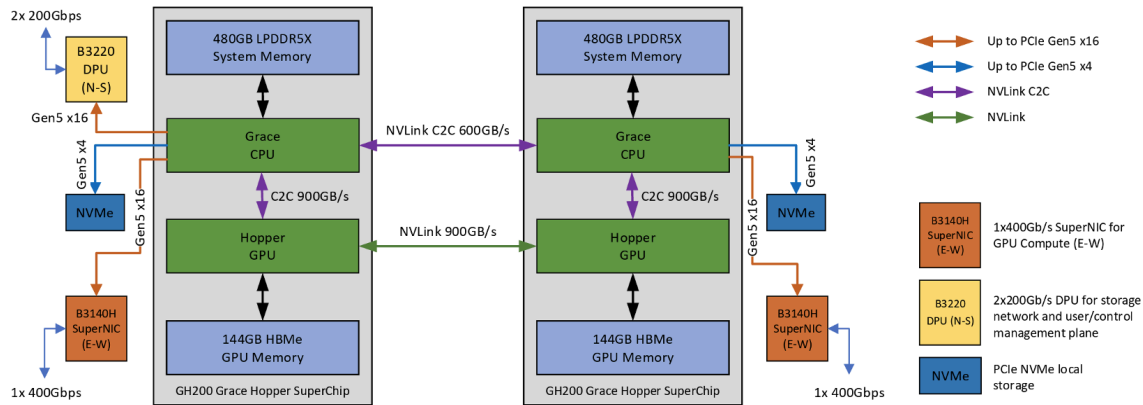
Scale-out NVIDIA Grace Configurations: These scale-out Enterprise RAs leverage NVIDIA's Grace CPU technology instead of the x86 variants above, bringing together the groundbreaking performance of NVIDIA GPUs with the versatility of the NVIDIA Grace™ CPU. At the heart of the Grace Hopper Superchip is NVIDIA's memory-coherent NVLink-C2C interconnect, enabling applications to oversubscribe the GPU's memory and directly utilize the Grace CPU's memory at high bandwidth. This makes it ideal for compute- and memory-intensive workloads like single-node LLM inference, retrieval-augmented generation (RAG), recommenders, graph neural networks (GNNs), high-performance computing (HPC), and data processing.

- Scale-out Grace 2-2-3-400 reference configuration is for scale-out NVIDIA superchips with 2 CPUs balanced with 2 GPUs plus 3 NICs with East-West traffic of 400 GbE per GPU. This pattern scales from 4 to 32 nodes in a cluster.
  - Eligible GPU: NVIDIA GH200 NVL2 Superchip
  - Eligible CPU: NVIDIA Grace
  - East-West Networking: NVIDIA BlueField-3 B3140H or NVIDIA ConnectX-7
  - North-South Networking: NVIDIA BlueField-3 B3220

Figure 4. Below is an example from the NVIDIA GH200 NVL2 and NVIDIA Spectrum-X Networking Platform Enterprise RA and is optimized for workloads of multi-node AI, HPC, or hybrid applications.

## Dual Superchip System Configuration (2-2-3-400)

NVIDIA GH200 NVL2 Accelerated Computing Platforms



## Networking

Enterprise RA networking leverages NVIDIA expertise in AI cloud data centers and optimizes network traffic flow, enabling the highest AI performance and scale while ensuring cloud manageability and security. Each Enterprise RA includes design recommendations based on the NVIDIA Spectrum-X Ethernet platform — combining Spectrum-4 Ethernet switches and NVIDIA BlueField-3 SuperNICs for optimal performance. Additionally, each Enterprise RA includes guidance on the appropriate network topology at multiple design points based on scale and use case.

- East-West Networking:** These recommendations are critical for AI processing, handling internal data transfers that affect model training and scaling, requiring high bandwidth and low latency solutions. These are tailored for AI clusters to improve communication between GPUs and other components, ensuring seamless data flow within the data center. This is critical for scaling as data is processed and passed between various layers in AI models (across GPUs, CPUs, and storage). Poorly managed east-west traffic can lead to bottlenecks, slowing down training times and reducing the overall efficiency of the AI pipeline.
- North-South Networking:** Supports external communication and is especially important for storage connectivity for data ingestion and result delivery. Presently, NVIDIA recommends

NVIDIA BlueField Data Processing Units (DPUs) for all North-South traffic to offload and ensure secure, efficient handling of requests from outside the network.

- **Switching:** For all Enterprise RAs, NVIDIA provides configuration recommendations for Ethernet, which is the preferred switching for enterprise workloads.

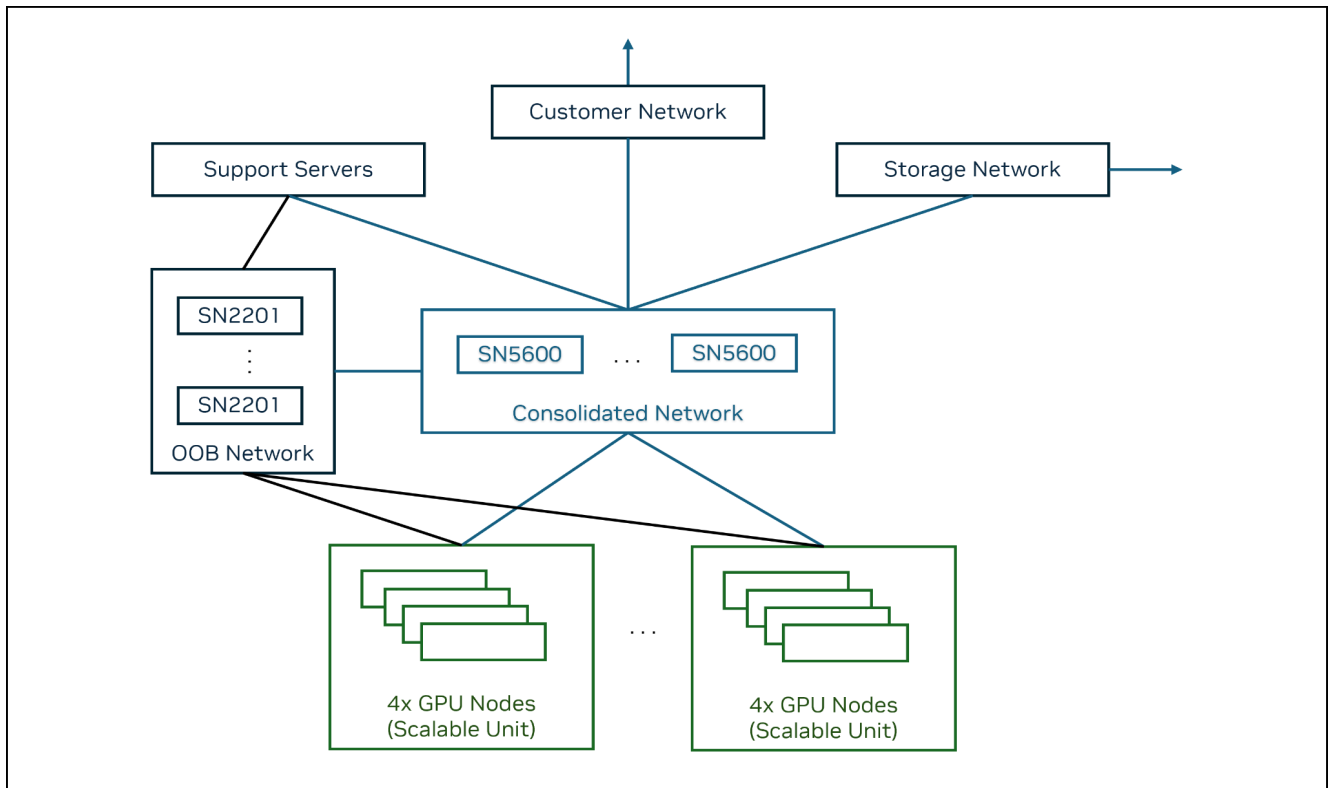


Figure 5: Shows the simplified networking topology for building clusters with 4 node scalable units.

## Storage

As enterprises build AI factories the importance of this data cannot be overstated: access to high quality data directly impacts the performance and reliability of AI models. Data is essential for developing and optimizing AI applications, and it must be fed across all stages of the AI pipeline, from model building to training, tuning, and inference, with varied storage requirements at each stage. It's the fuel for the AI factory.

The **NVIDIA-Certified Storage** program is designed to support the massive data demands of enterprise AI factories by offering a comprehensive storage certification that complements the Enterprise RA programs. This empowers partners and customers to deploy to build AI factories that efficiently

leverage massive amounts of data for faster, more accurate, and reliable AI models. NVIDIA Enterprise RAs have designated network end points to attach NVIDIA-Certified storage solutions.

The NVIDIA-Certified Storage program offers two levels of certification: Foundation and Enterprise. These storage certifications integrate seamlessly with corresponding NVIDIA Enterprise RAs to ensure the storage systems have the performance to support the needs of the North-South networking and feed the compute nodes with data. The Foundation level storage certification certifies storage partners for the 2-4-3-200 and 2-8-5-200 reference configurations. The larger scale Enterprise level storage certification certifies storage partners for 2-8-9-400 reference configuration.

This is important for organizations looking to build robust, high-performance systems that leverage NVIDIA accelerated computing with optimal storage performance. It's also important to ensure that the storage is fully compatible with NVIDIA's GPU and networking technologies, reducing integration risks and deployment complexity.

## Software

NVIDIA Enterprise RAs provide a bare-metal foundation optimized for performance and reliability across AI workloads. The RAs can be used with industry-standard orchestration tools, such as Kubernetes and Slurm. NVIDIA Enterprise RAs also provide the foundation for enterprise-grade AI software and models across a wide range of AI workloads. For example, NVIDIA AI Enterprise (including NVIDIA NIM microservices) and NVIDIA Omniverse software can be deployed on system configurations designed using an Enterprise RA to provide a full-stack solution.

Going forward, NVIDIA Enterprise RAs will provide specific software configuration recommendations to optimize performance for various AI workloads, including sizing and deployment guidelines. By following Enterprise RA recommendations, NVIDIA partners and our joint enterprise customers can simplify AI adoption and ensure optimal performance with faster time to value.

## Why This Matters

NVIDIA Enterprise Reference Architectures provide guidance for building high-performance, scalable data center infrastructure. The transformation of traditional data centers into AI Factories is revolutionizing how enterprises process and analyze data by integrating advanced computing and networking technologies to meet the substantial computational demands of AI applications.

To address these challenges, NVIDIA has introduced NVIDIA Enterprise Reference Architectures, offering clear and consolidated recommendations for our partners and joint enterprise customers building AI Factories based on NVIDIA-Certified systems with NVIDIA Certified storage partners. Informed by years of expertise in designing and building large-scale computing systems, these NVIDIA Enterprise RAs streamline the deployment process, providing flexible and cost-effective configurations that eliminate much of the guesswork and risk.

## Business Benefits

- **Accelerate Time to Token:** By leveraging NVIDIA's structured approach and recommended designs, enterprises can deploy AI solutions faster, reducing the time to achieve business value.
- **Resource and Cost Efficiency:** Optimized configurations ensure efficient use of resources, minimizing waste and reducing overall costs.
- **Risk Mitigation:** Recommending proven and tested design pattern increases customer confidence and helps mitigate deployment risks, ensuring consistent and predictable outcomes.

## IT Benefits

- **Performance:** High-performance computing capabilities meet the demanding requirements of AI workloads, ensuring optimal performance.
- **Scale and Manageability:** Flexible scaling options and manageable configurations allow enterprises to grow their AI infrastructure seamlessly.
- **Reduced Complexity and TCO:** Simplified deployment processes and efficient designs reduce complexity and total cost of ownership (TCO).
- **Supportability:** Following specific standardized design patterns allows for consistent operation from installation-to-installation, reduces the need for frequent support, and enables faster resolution times.

By following NVIDIA's structured approach to introducing new technologies and leveraging the NVIDIA Enterprise RA recommendations, our system partners and joint enterprise customers can confidently build and scale AI Factories. This enables them to focus on running their business and delivering innovative AI solutions, ultimately maximizing their return on investment and achieving significant business value.

For more information please visit the [Qualified System Catalog](#) page and filter the list for any of the NVIDIA-Certified System categories. View the [NVIDIA-Certified Systems documentation](#) for additional information.

---

# Appendix: A

## NVIDIA Enterprise Reference Architecture: NVIDIA L40S and NVIDIA Spectrum Platforms

The NVIDIA L40S and NVIDIA Spectrum Platforms Enterprise RA is optimized for multi-node AI or hybrid visual computing applications. This modular architecture is based on NVIDIA-Certified OVX L40S systems, each equipped with up to four L40S GPUs. Using a four-node scalable unit (SU), this can scale up to 32 NVIDIA-Certified OVX L40S systems, totaling 128 L40S GPUs. Fully tested systems can scale to twenty-four 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. Hardware support is provided through the fulfillment system partner, while software support from NVIDIA is available via a per GPU paid subscription to NVIDIA AI Enterprise.

### Use Cases

- **Visual Computing:** 3D Graphics, Rendering
- **AI Inference:** Medium model parameter inference workloads
- **AI Training:** Small model training and fine-tuning

### NVIDIA OVX L40S Reference Configurations

This Enterprise RA leverages NVIDIA OVX L40S systems, which deliver powerful AI and visual computing performance to accelerate the next generation of AI-enabled enterprise workloads in the data center. The building blocks of the OVX architecture are the performance-optimized NVIDIA L40S GPU server configurations with BlueField-3 DPUs and BlueField-3 SuperNICs.

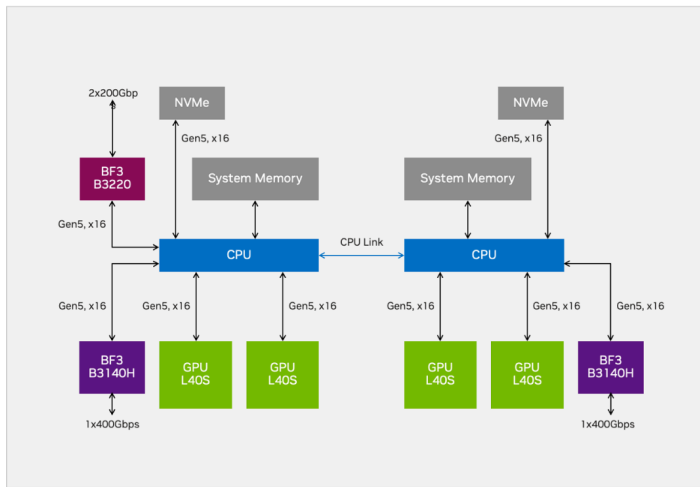
The NVIDIA L40S GPU, based on the Ada Lovelace architecture, is the most powerful universal GPU for the data center, delivering breakthrough multi-workload acceleration for AI, graphics, and video applications. With accelerated AI compute and best-in-class visual computing capabilities, the L40S GPU provides end-to-end performance to power the next generation of AI-enabled data center workloads.

NVIDIA-Certified OVX L40S systems are based on a common system design with flexibility for optimizing the configuration to match cluster requirements. Systems are available in 4-GPU and 8-GPU configurations. This was built using the 4-GPU pattern (2-4-3-200 (CPU-GPU-NIC-Bandwidth)), but the 8-GPU pattern (2-8-5-200 CPU-GPU-NIC-Bandwidth) can also be utilized based on specific needs.

Figure 6. 4-GPU OVX system configuration

## NVIDIA OVX L40S Enterprise Reference Architecture

### Optimized 2-4-3 Design



**2: CPU | 4: GPU | 3: DPU**

NVIDIA OVX L40S — SERVER	
CPU	2x 32c Intel Xeon Gold 6448Y 2x 32c AMD EPYC 9354
GPUs	4x NVIDIA L40S
Networking — E/W	2x BlueField-3, B3140H (1x400Gb)
Networking — N/S	1x BlueField-3, B3220 (2x200Gb)
Host Memory	Min 384GB DDR5 ECC (1 DIMM per slot)
Host Boot Drive	1x 1TB NVMe
Host Storage	2x 4TB NVMe



---

# Appendix: B

## NVIDIA Enterprise Reference Architecture: NVIDIA H100 NVL and NVIDIA Spectrum Platforms

The NVIDIA H100 NVL and NVIDIA Spectrum Platforms Enterprise RA is optimized for multi-node AI or hybrid applications. This modular architecture is based on NVIDIA-Certified H100 NVL systems, each equipped with four H100 NVL GPUs. Using a four-node scalable unit (SU), this can scale up to 32 NVIDIA-Certified H100 NVL systems, totaling 128 H100 NVL GPUs. Fully tested systems can scale to twenty-four 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. Hardware support is provided through the fulfillment system partner, while software support from NVIDIA is available via a per GPU paid subscription to NVIDIA AI Enterprise.

### Use Cases

- AI Inference: Medium model parameter inference workloads
- AI Training: Small model training and fine-tuning

### NVIDIA H100 NVL Reference Configurations

The NVIDIA H100 NVL Tensor Core GPU is the most optimized platform for LLM inference, offering high compute density, high memory bandwidth, high energy efficiency, and a unique NVLink architecture. It delivers unprecedented acceleration to power the world's highest-performing elastic data centers for AI, data analytics, and high-performance computing (HPC) applications. The NVIDIA H100 NVL card is a dual-slot 10.5-inch PCI Express Gen5 card based on the NVIDIA Hopper architecture.

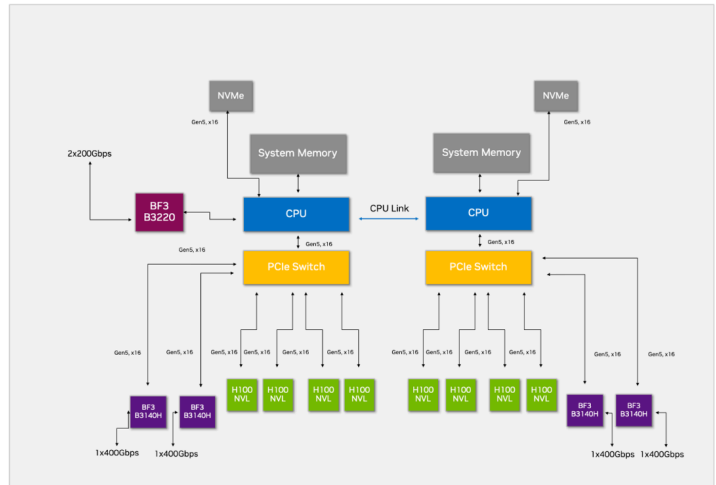
NVIDIA-Certified H100 NVL systems are based on a common system design with flexibility for optimizing the configuration to match cluster requirements. Systems are available in 2-GPU, 4-GPU, and 8-GPU configurations. This was built using the 4-GPU pattern (2-4-3-200 CPU-GPU-NIC-Bandwidth), but the 8-GPU pattern (2-8-5-200 CPU-GPU-NIC-Bandwidth) can also be utilized based on specific needs.

Figure 7. 8-GPU NVIDIA H100 NVL system configuration

## NVIDIA H100 NVL Enterprise Reference Architecture

Optimized 2-8-5 Design

NVIDIA H100 NVL— SERVER	
CPU	2x 56c Intel Xeon Gold 8480+ 2x 64c AMD EPYC 9554
GPUs	8x NVIDIA H100 NVL
Networking — E/W	4x BlueField-3, B3140H (1x400Gb)
Networking — N/S	1x BlueField-3, B3220 (2x200Gb)
Host Memory	Min 768GB DDR5 ECC (1 DIMM per slot)
Host Boot Drive	1x 1TB NVMe
Host Storage	4x 4TB NVMe



**2: CPU | 8: GPU | 5: DPU**

---

# Appendix: C

## NVIDIA Enterprise Reference Architecture: NVIDIA HGX H100/H200/B200 and NVIDIA Spectrum-X Networking Platform

The NVIDIA HGX H100/H200/B200 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 H100/H200/B200 systems, each equipped with eight H100 or H200 or B200 SXM GPUs. Using a four-node scalable unit (SU), this can scale up to 32 NVIDIA-Certified HGX 8-GPU systems, 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. Hardware support is provided through the fulfilling system partner, while software support from NVIDIA is available via a per GPU paid subscription to NVIDIA AI Enterprise.

### Use Cases

- **AI Inference:** Large (per node) and medium (per GPU) model parameter inference workloads
- **AI Training:** Large to small model training and fine-tuning based on cluster sizing

### NVIDIA HGX H100/H200/B200 Reference Configurations

The HGX 8-GPU baseboard is an AI powerhouse that enables enterprises to expand the frontiers of business innovation and optimization.

The HGX H100/H200 baseboard combines H100/H200 Tensor Core GPUs with high-speed interconnects to form the world's most powerful systems. With eight H100/H200 GPUs, the baseboard has up to 640 GB (1,128 GB for H200) of GPU memory for unprecedented acceleration.

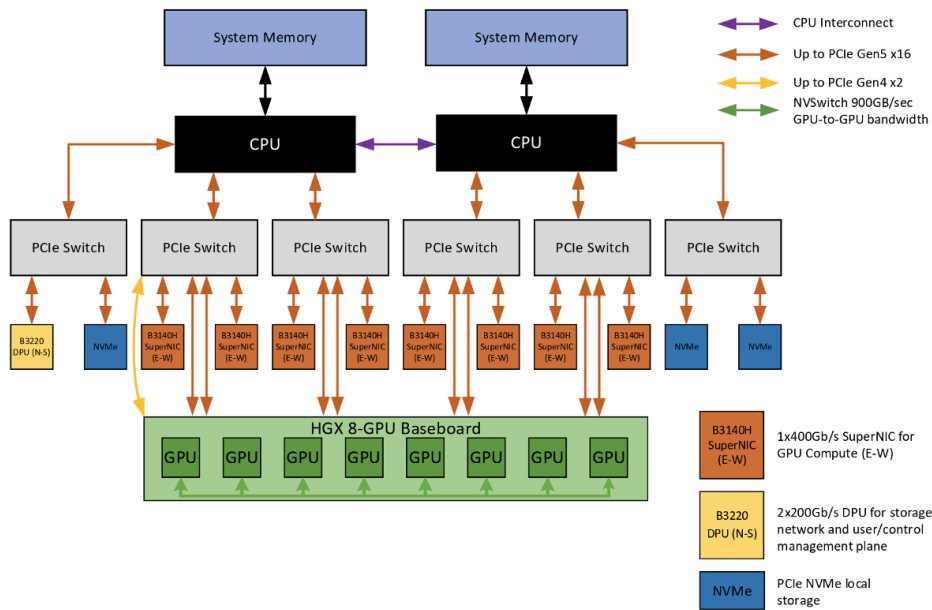
The HGX B200 baseboard is a Blackwell x86 platform, based on eight B200 GPUs, that has up to 1.44 TB of GPU memory and can deliver up to 144 petaFLOPs of AI performance. The HGX B200 baseboard delivers the best performance (15 times more than the HGX H100 baseboard) and TCO (12 times more than the HGX H100 baseboard) for x86 scale-up platforms and infrastructure. Each GPU is configurable up to 1 kW per GPU.

NVIDIA-Certified HGX H100/H200/B200 8-GPU systems are based on a common system design with flexibility for optimizing the configuration to match cluster requirements. This was built using the 8-GPU design pattern (2-8-9-400 CPU-GPU-NIC-Bandwidth), but the 4-GPU design can also be utilized based on specific needs. An example of this design is shown in Figure 7.

Figure 8. Example of a HGX H100, H200 or B200 8 GPU system configuration

## 8 GPU System Configuration (2-8-9-400)

NVIDIA HGX™ H100, H200, or B200 Accelerated Computing Platforms



---

# Appendix: D

## NVIDIA Enterprise Reference Architecture: NVIDIA H200 NVL or RTX™ PRO 6000 Blackwell Server Edition and NVIDIA Spectrum-X Networking Platform

This NVIDIA Enterprise Reference Architecture (Enterprise RA) is optimized for multi-node AI and hybrid applications, utilizing a 2-8-5-200 node architecture with either H200 NVL or RTX PRO 6000 Blackwell Server Edition GPUs. This setup is further enhanced by the NVIDIA Spectrum-X Networking Platform, which provides advanced networking capabilities tailored for AI workloads.

The Enterprise RA is a modular architecture based on an NVIDIA-Certified system, each with eight H200 NVL or RTX™ PRO 6000 Blackwell Server Edition GPUs. Using a four-node scalable unit (SU), this Enterprise RA scales up to 32 NVIDIA-Certified systems for a total of 256 H200 NVL or RTX™ PRO 6000 Blackwell Server Edition GPUs.

### Use Cases

- **AI Inference:** Large (per node) and medium (per GPU) model parameter inference workloads
- **AI Training:** Large to small model training and fine-tuning based on cluster sizing

### NVIDIA H200 NVL Systems

NVIDIA H200 NVL is ideal for lower-power, air-cooled enterprise rack designs, delivering acceleration for every AI and HPC workload regardless of size. With up to four GPUs connected by NVIDIA NVLink™ and a 1.5X memory increase, large language model (LLM) inference can be accelerated up to 1.7X and HPC applications achieve up to 1.3X more performance over the H100 NVL. With four H200 NVL GPUs interconnected with NVLink 2 or 4 bridges, the combined H200 GPU memory is 564 GB for unprecedented acceleration.

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

NVIDIA H200 NVL

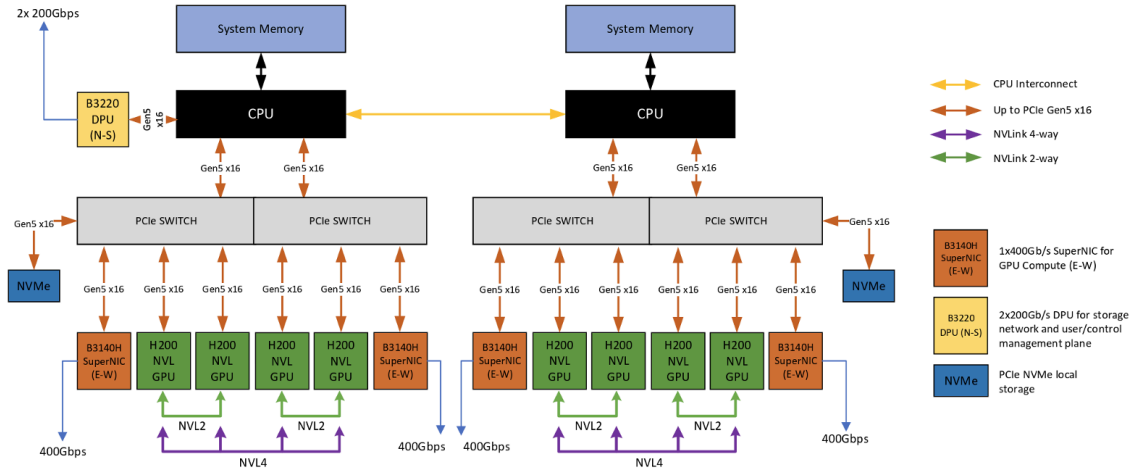


Figure 9. Example of a NVIDIA H200 NVL 8 GPU system configuration

## NVIDIA RTX PRO 6000 Blackwell Server Edition Systems

The RTX PRO 6000 Blackwell Server Edition GPU is a high-performance GPU designed for server environments. It features 96GB of GDDR7 memory per GPU, significantly surpassing the 48GB GDDR6 memory of its counterpart, the NVIDIA L40S. This results in a substantial increase in memory bandwidth, reaching up to 1.6TB/s per GPU, compared to 864GB/s for the L40S. When configured in an 8-GPU node, the RTX PRO 6000 offers 768GB of GDDR7 memory and achieves an aggregate memory bandwidth of up to 12.8TB/s, doubling the capacity and bandwidth of the L40S setup. This makes it particularly suited for demanding applications requiring high memory and bandwidth capabilities.

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

RTX™ PRO 6000 Blackwell Server Edition

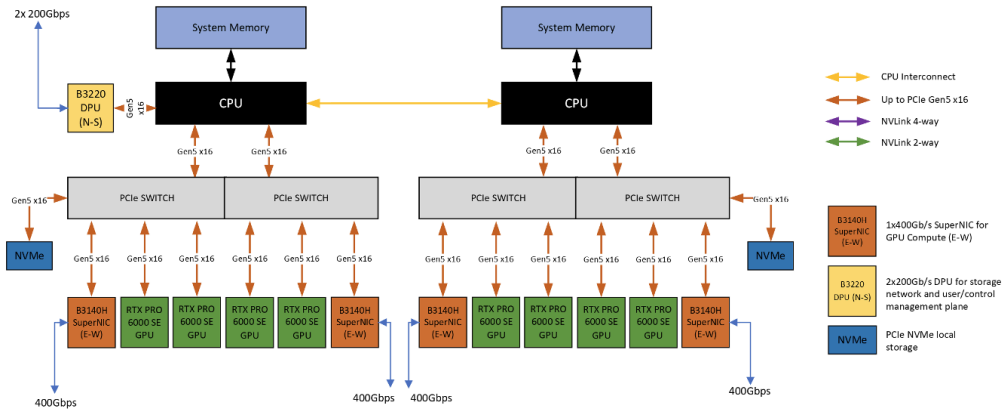


Figure 10. Example of a NVIDIA RTX PRO 6000 Blackwell Server Edition 8 GPU system configuration

## Notice

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation ("NVIDIA") does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regard to the purchase of the NVIDIA product referenced in this specification.

NVIDIA products are not designed, authorized or warranted to be suitable for use in medical, military, aircraft, space or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to ensure the product is suitable and fit for the application planned by the customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this specification. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this specification, or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this specification. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA. Reproduction of information in this specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

## Trademarks

NVIDIA, the NVIDIA logo, NVIDIA CUDA, NVIDIA Omniverse, NVIDIA RTX, NVIDIA Tesla, NVIDIA Turing, NVIDIA Volta, NVIDIA Jetson AGX Xavier, NVIDIA DGX, NVIDIA HGX, NVIDIA EGX, NVIDIA CUDA-X, NVIDIA GPU Cloud, GeForce, Quadro, CUDA, GeForce RTX, NVIDIA NVLink, NVIDIA NVSwitch, NVIDIA DGX POD, NVIDIA DGX SuperPOD, and NVIDIA TensorRT, are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

## Copyright

© 2025 NVIDIA Corporation. All rights reserved.