NVIDIA DGX SuperPOD: Next Generation Scalable Infrastructure for AI Leadership Reference Architecture Featuring NVDIA DGX GB200

Release latest

NVIDIA Corporation

Jun 16, 2025

Reference Architecture

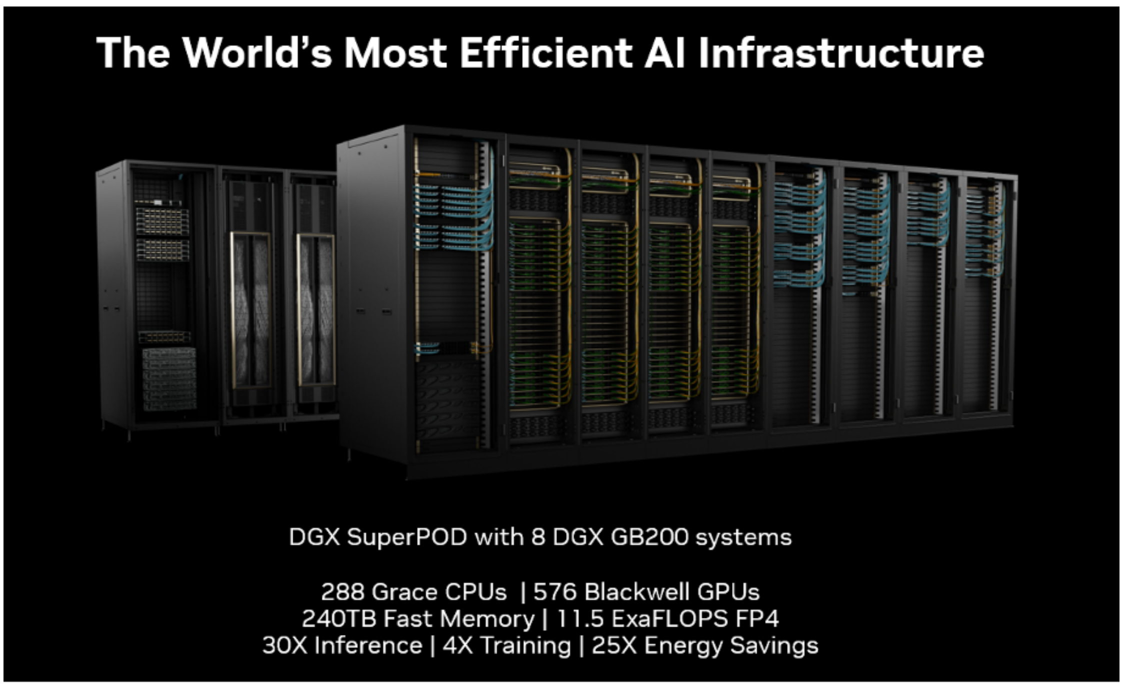
Document Number	RA-11338-001 V01
Publication Date	2025-06-16

Tip

The NVIDIA DGX SuperPOD: Next Generation Scalable Infrastructure for AI Factories Reference Architecture Featuring NVIDIA DGX GB200 is also available as a PDF.

Chapter 1. Abstract

The NVIDIA DGX SuperPOD architecture has been designed to power the next-generation AI factories with unparalleled performance, scalability, and innovation that supports all customers in the enterprise, higher education, research, and the public sector. It is a physical twin of the main NVIDIA research and development system, meaning the company's infrastructure software, applications, and support are first tested and vetted on the same architecture.



This DGX SuperPOD Reference Architecture (RA) is based on DGX GB200 systems powered by Grace CPUs and Blackwell GPUs. The RA discusses the components that define the scalable and modular architecture of DGX SuperPOD. DGX SuperPOD is built on the concept of scalable units (SU); each SU contains 8 DGX GB200 systems, which enables rapid deployment of DGX SuperPOD of anysize. The RA also includes details regarding the SU design and specifics of InfiniBand, NVLink network, Ethernet fabric topologies, storage system specifications, recommended rack layouts, and wiring guidelines.

This RA combines the latest NVIDIA technologies to help companies and industries develop their own AI factories. To achieve the most scalability, DGX SuperPOD is powered by several key NVIDIA technologies and solutions, including:

- NVIDIA DGX GB200 systems: Powerful computational building block for AI and HPC.
- NVIDIA NDR (400 Gbps) InfiniBand: High performant, low latency, and scalable network interconnect.

- NVIDIA Spectrum-4 (800 Gbps) Ethernet: High performant, low latency, and scalable ethernet connectivity for storage.
- NVIDIA fifth-generation NVLink® technology: a high-speed interconnect for CPU and GPU processors in accelerated compute systems to provide unprecedented performance for most demanding communication patterns.
- NVIDIA Mission Control: a unified operations and orchestration software stack for managing AI factories.

Chapter 2. Key Components of the DGX SuperPOD

The DGX SuperPOD architecture is designed to meet the high demands of AI factories for the era of reasoning. AI factories require specialized components like high-performance GPUs and CPUs, advanced networking, and cooling systems to support the intensive computational needs of AI workloads. These factories excel at AI reasoning—enabling faster, more accurate decision-making across industries. And using NVIDIA's end-to-end accelerated computing platform, they're optimized for energy efficiency while accelerating AI inference performance, helping enterprises deploy secure, future-ready AI infrastructure.

DGX SuperPOD is the integration of key NVIDIA components, as well as storage solutions from partners certified to work in a DGX SuperPOD environment. By leveraging the concept of Scalable Units (SUs, as defined below in this document), DGX SuperPOD reduces AI factory deployment times from months to weeks, which in turn reduces time-to-solution and time-to-market of next generation models and applications.

DGX SuperPOD is to be deployed on-premises, meaning the customer owns and manages the hardware. This can be within a customer's data center or co-located at a commercial data center, but the customer owns the hardware, the service it provides, and is responsible for their cluster infrastructure.

The key components of DGX GB200 SuperPOD are shown below, each component will be discussed in detail:

- NVIDIA DGX GB200 NVL72 rack system
- NVIDIA InfiniBand
- Mission Control software platform

2.1. NVIDIA DGX GB200 Rack System

The NVIDIA DGX GB200 system (Figure 2.1) is an AI powerhouse that enables enterprises to expand the frontiers of business innovation and optimization. Each DGX GB200 delivers breakthrough AI performance in a rack-scale, 72 GPU configuration. The NVIDIA Blackwell GPU architecture provides the latest technologies that brings months of computational effort down to days and hours, on some of the largest AI/ML workloads.

To accommodate the high compute density per rack and at the same time to more efficiently use datacenter space, the DGX GB200 rack system employs a sophisticated hybrid cooling solution to manage the substantial heat generated by the powerful GPUs and other components. The most power-intensive components, such as GPUs and CPUs, are liquid cooled. Other components are air cooled.

NVIDIA's Infrastructure Specialist team can provide end-to-end assistance for datacenter planning, system deployment, and bring-up services.

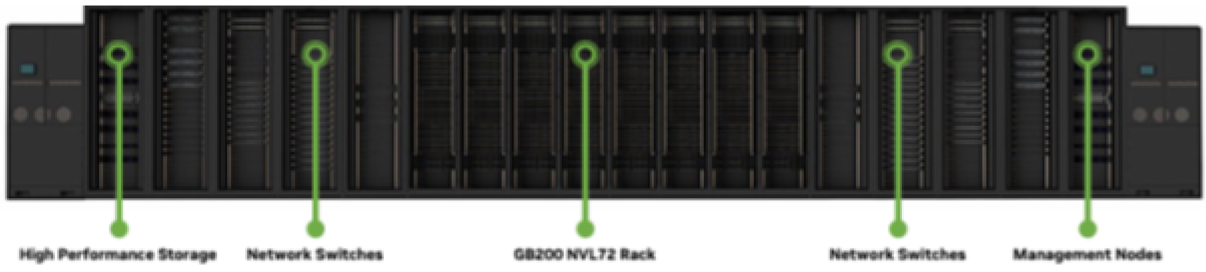


Figure 2.1: DGX GB200 (4 racks are shown here)

2.1.1. DGX GB200 Compute Tray

The compute nodes for DGX GB200 rack system are using the 72x1 NVLink topology, which contains 72 GPUs in a single NVLink domain. There are 18 compute nodes (AKA trays) in a DGX GB200 rack system. Each compute tray contains two GB200 Superchips, and each Superchip has two B200 GPUs and one Grace CPU. A coherent chip-to-chip interconnect link, called NVLink-C2C, bridges the two Superchips and enables them to act as a single logical unit for a single OS instance to operate.

The compute tray integrates four ConnectX-7 (CX-7) NICs to support InfiniBand NDR (400Gbps) connectivity for the cross racks Compute network, and two BlueFiled-3 (BF3) NICs to support 2x200Gbps connectivity for the In-band Management and Storage networks. All network ports are located at the front-side of the rack, intentionally for the cold aisle.

Each compute tray also provides a total of 4x 3.84TB E1.S NVMe as RAID0 for local storage and 1x 1.92TB M.2 NVMe for the OS image.

Figure 2.2 and Figure 2.3 show the front and rear of a DGX GB200 compute tray respectively.

Figure 2.4 shows the block diagram of the GB200 compute tray with two GB200 Superchips, each chip combines two NVIDIA B200 Tensor Core GPUs and one NVIDIA Grace CPU, connected over a 900GB/s ultra-low-power NVLink-C2C interconnect.

2.1.2. NVLink Switch Tray

Each DGX GB200 NVL72system rack also features nine NVLink switches (AKA NVLink switch tray). All NVLink interconnections are done through the rear-side, blind-mate connectors to the cable cartridges within the rack. Each switch tray features two COMe RJ45 ports for NVOS and one BMC module for out-of-band management with one RJ45 port. Figure 2.5 showcases the front design of a NVLink Switch Tray.

2.1.3. DGX Power Shelves

The power shelf used for DGX DGX GB200 SuperPOD has six 5.5kW PSUs configured as N redundancy and can deliver up to 33kW of power. There are eight total power shelves in a single DGX GB200 NVL72 rack system. At the rear of the power shelf is a set of RJ45 ports used for power brake and current sharing feature. The power shelves are daisy chained to each other using these RJ45 ports. At the front of the power shelf is the BMC port. Figure 2.6 shows the front of the power shelf.

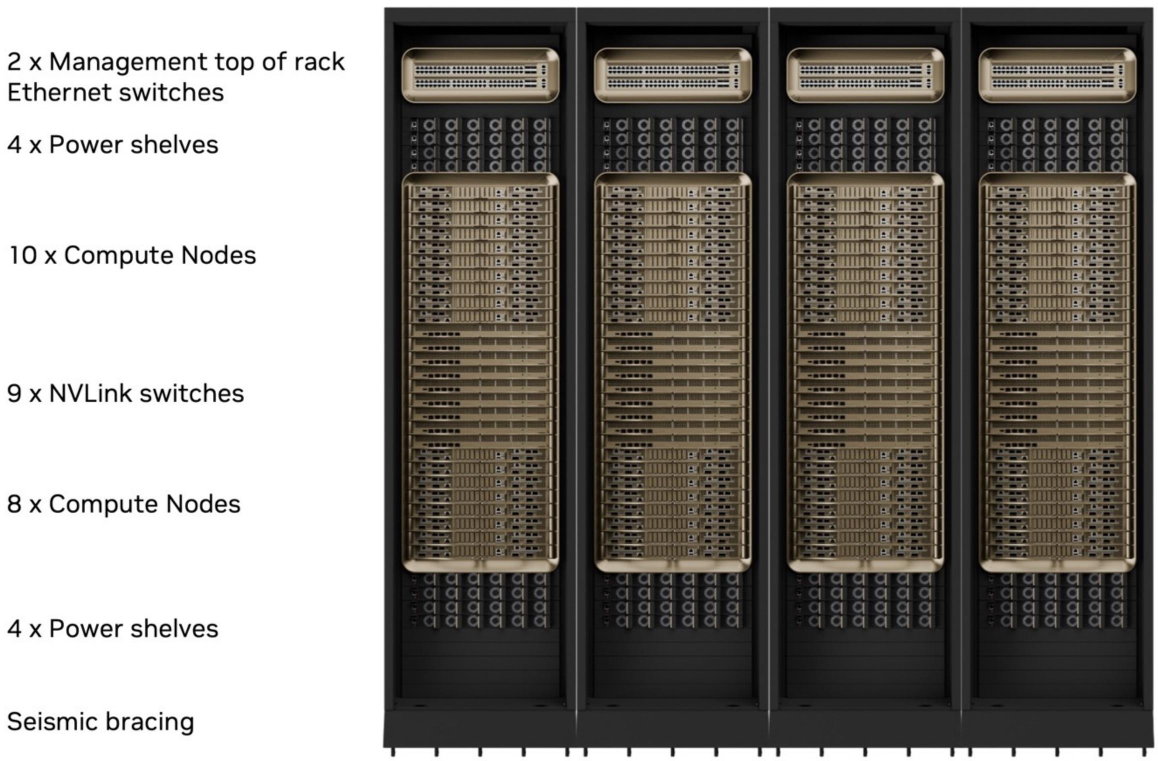


Figure 2.2: GB200 Compute Tray Front

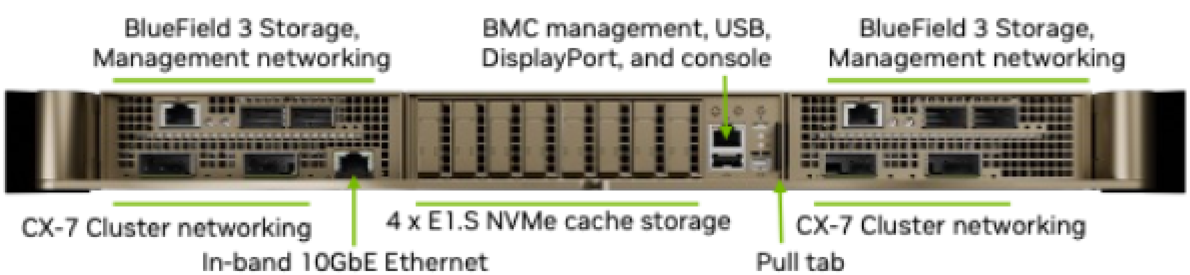


Figure 2.3: GB200 Compute Rear Front

Figure 2.4: GB200 Compute Tray Block Diagram

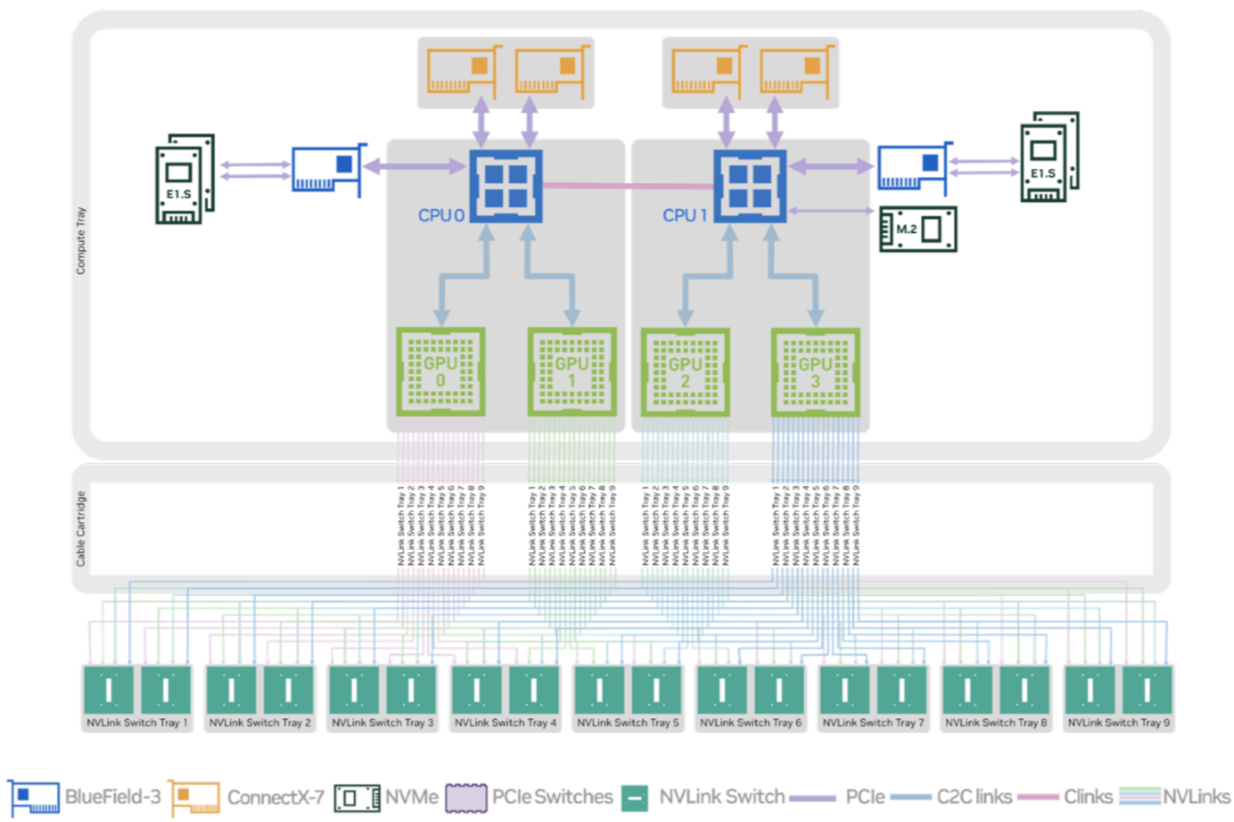


Figure 2.5: GB200 Compute Tray Block Diagram

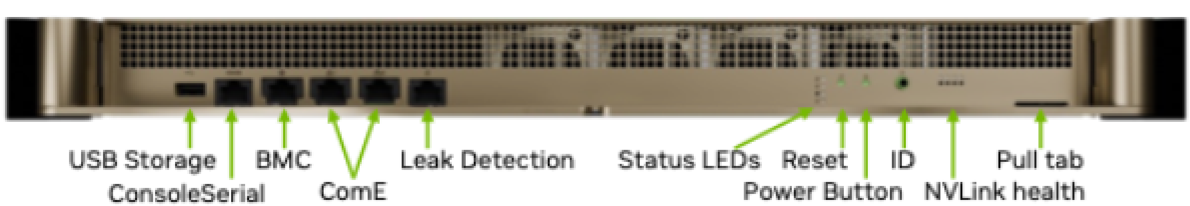


Figure 2.6: Power Shelf

2.2. NVIDIA InfiniBand Technology

InfiniBand is a high-performance, low latency, RDMA capable networking technology, proven over 20 years in the harshest compute environments to provide the best inter-node network performance. InfiniBand continues to evolve and lead data center network performance.

The latest generation InfiniBand, NDR, has a peak speed of 400 Gbps per direction with an extremely low port-to-port latency. It is backwards compatible with the previous generations of InfiniBand specifications. InfiniBand is more than just peak bandwidth and low latency. It provides additional features to optimize performance including adaptive routing (AR), collective communication with SHARP™, dynamic network healing with SHIELD™, and supports several network topologies including fat-tree, Dragonfly, and multi-dimensional Torus to build the largest network fabrics and computer systems possible.

2.3. NVIDIA Mission Control

The DGX GB200 SuperPOD Reference Architecture represents the best practices for building high-performance AI factories. There is flexibility in how these systems can be presented to customers and users. NVIDIA Mission Control software is used to manage all DGX GB200 SuperPOD deployments.

NVIDIA Mission Control is a sophisticated full-stack software solution. As an essential part of the DGX SuperPOD experience, it optimizes developer workload performance and resiliency, ensures continuous uptime with automated failure handling, and provides unified cluster-scale telemetry and manageability. Key features include full-stack resiliency, predictive maintenance, unified error reporting, data center optimizations, cluster health checks, and automated node management.

Mission Control software incorporates the same technology that NVIDIA uses to manage thousands of systems for our ad-winning data scientists and provides an immediate path to a TOP500 supercomputer for organizations that need the best of the best.

DGX SuperPOD is to be deployed on-premises, meaning the customer owns and manages the hardware. This can be within a customer's data center or co-located at a commercial data center, but the customer owns the hardware, the service it provides, and is responsible for their cluster infrastructure as well as providing the building management system for integration.

2.4. Components

The hardware components of DGX SuperPOD are described in Table 2.1. The software components are shown in Table 2.2.

Table 2.1: DGX SuperPOD hardware components by NVIDIA

Component	Technology	Description
8x Compute Racks	NVIDIA DGX GB200 NVL72	The world's premier rack-scale, purpose-built AI systems featuring NVIDIA Grace-Blackwell GB200 modules. GB200 Superchips. NVL72 scale-out GPU interconnect and integrated NVLink switch trays.
NVLink Fabric	NVIDIA NVLink 5	NVLink Switches supports fast, direct memory access between GPUs on the same compute rack.
Compute Fabric	NVIDIA Quantum QM9700 InfiniBand switches	Rail-optimized, non-blocking, full fat-tree network with eight NDR400 connections per system for cross rack GPU communications.
Storage and In-band Management Fabric	NVIDIA Spectrum 4 SN5600 Ethernet switches	The fabric is optimized to match peak performance of the configured storage array, built with 64 port 800 Gbps Ethernet switch providing high port density with high performance.
InfiniBand management	NVIDIA Unified Fabric Manager Appliance, Enterprise Edition	NVIDIA UFM combines enhanced, real-time network telemetry with AI powered cyber intelligence and analytics to manage scale-out InfiniBand data centers
NVLink Management	NVIDIA Network Manager eXperience (NMX)	NVIDIA NMX manages, operates the NVLink switches and provides real-time network telemetry to manage all NVLink infrastructure
Out-of-band (OOB) management network	NVIDIA SN2201 switch	48 x 1 Gbps Ethernet switch leveraging copper ports to minimize complexity

Table 2.2: DGX SuperPOD software components

Component	Description
Mission Control Software	NVIDIA Mission Control software delivers a full stack data center solution engineered for NVIDIA DGX SuperPOD with DGX GB200 or DGX B200 infrastructure deployments. It integrates essential management and operational capabilities into a unified platform, thereby providing enterprise customers with simplified control over their NVIDIA DGX SuperPOD infrastructure deployments at scale. NVIDIA Mission Control also leverages NVIDIA Run;ai functionality, providing seamless workload orchestration.
NVIDIA AI Enterprise	NVIDIA AI Enterprise is an end-to-end, cloud-native software platform that accelerates data science pipelines and streamlines development and deployment of production-grade LLMs and other generative AI applications.
Magnum IO	Enables increased performance for AI and HPC
NVIDIA NGC	The NGC catalog provides a collection of GPU-optimized containers for AI and HPC
Slurm	A classic workload manager used to manage complex workloads in a multi-node, batch-style, compute environment

2.5. Design Requirements

DGX SuperPOD is designed to minimize system bottlenecks throughout the tightly coupled configuration to provide the best performance and application scalability. Each subsystem has been thoughtfully designed to meet this goal. In addition, the overall design remains flexible so that SuperPOD can be tailored to better integrate into existing data centers.

2.5.1. System Design

DGX SuperPOD is optimized for a customers' particular workload of multi-node AI and HPC applications:

- A modular architecture based on Scalable Units (SUs) of 8 DGX GB200 systems.NVL72 rack systems.
- A fully tested system scales to two SUs, but larger deployments can be built based on customer requirements
- Rack-A fully tested system scales for a single SU, seamlessly.
- level integrated design which allows rapid installation and deployment of liquid cooled, high density compute racks.
- Storage partner equipment that has been certified to work in DGX SuperPOD environment.
- Storage partner equipment that has been certified to work in DGX SuperPOD environments.
- Full system support—including compute, storage, network, and software—is provided by NVIDIA Enterprise Support (NVEX).

2.5.2. Compute Fabric

- The compute fabric is rail-optimized to the top layer of the fabric.
- The compute fabric is a balanced, full-fat tree.
- The compute fabric is designed with current state-of-the-art, top performing, high-performance, low-latency network switches, and supports future generation of networking hardware.
- Managed NDR switches are used throughout the design to provide better management of the fabric.
- The fabric is designed to support the latest SHARPv3 features.

2.5.3. Storage Fabric (High Speed Storage)

The storage fabric provides high bandwidth to shared storage. It also has the following characteristics:

- It is independent of the compute fabric to maximize performance of both storage and application performance.
- Provides single-node line-rate of 2x 200Gbps to each DGX GB200 compute tray.
- Storage is provided over RDMA over Converged Ethernet to provide maximum performance and minimize CPU overhead.
- User-accessible management nodes provide access to shared storage.

2.5.4. In-Band Management Network

- The in-band management network fabric is Ethernet-based and is used for node provisioning, data movement, Internet access, and other services that must be accessible by the users.
- The in-band management network connections for compute and management nodes operate at 200 Gbps and are bonded for resiliency.

2.5.5. Out-of-Band Management Network

The OOB management network connects all the baseboard management controllers (BMC), BlueField baseboard management controllers, NVSwitch Management Interfaces (COMe), as well as other devices that should be physically isolated from system users.

2.5.6. Storage Requirements

The DGX SuperPOD compute architecture must be paired with a high-performance, balanced, storage system to maximize overall system performance. DGX SuperPOD is designed to use two separate storage systems, high-performance storage (HPS) and user storage, optimized for key operations of throughput, parallel I/O, as well as higher IOPS and metadata workloads.

2.5.6.1 High-Performance Storage

High-Performance Storage is provided via RDMA over Converged Ethernet v2 (RoCEv2) connected storage from a DGX SuperPOD certified storage partner, and is engineered and tested with the following attributes in mind:

- High-performance, resilient, POSIX-style file system optimized for multi-threaded read and write operations across multiple nodes.
- Certified for Grace-based systems.
- Native RoCE support.
- Local system RAM for transparent caching of data.
- Leverage local flash device transparently for read and write caching.

The specific storage fabric topology, capacity, and components are determined by the DGX SuperPOD certified storage partner as part of the DGX SuperPOD design process.

2.5.6.2 User Storage

User Storage differs from High-Performance storage in that it exposes an NFS share on the in-band management fabric for multiple uses. It is typically used for "home directory" type usage (especially with clusters deployed with SLURM), administrative scratch space, and shared storage as needed by DGX SuperPOD components in a High Availability configuration (e.g., Mission Control), the requirement for log files collection, and system configuration files.

With that in mind, User Storage has the following requirements:

- Designed for high metadata performance, IOPS, and key enterprise features such as log collection, data capacity. This is different than the HPS, which is optimized for parallel I/O and large capacity.
- Communicate over Ethernet, using NFS.
- 100GbE DR1 Connectivity.

User storage in a DGX SuperPOD is often satisfied with existing NFS servers already deployed, such that a new export is created and made accessible to the DGX SuperPOD's in-band management network. User Storage is therefore not described in detail in this DGX SuperPOD reference architecture.

Chapter 3. DGX SuperPOD Architecture

The DGX SuperPOD architecture is a combination of DGX systems, InfiniBand and Ethernet networking, management nodes, and storage. Figure 3.1 shows the rack layout of a single SU (Scalable Unit). Each SU requires a Thermal Design Power (TDP) of 1.2 Megawatts (MW). Generally, the data center should meet or exceed Uptime Institute Tier 3 design standards, or alternatively the TIA942-B Rated 3 or EN50600 Availability Class 3 design standards, including concurrent maintainability and no single point of failure.

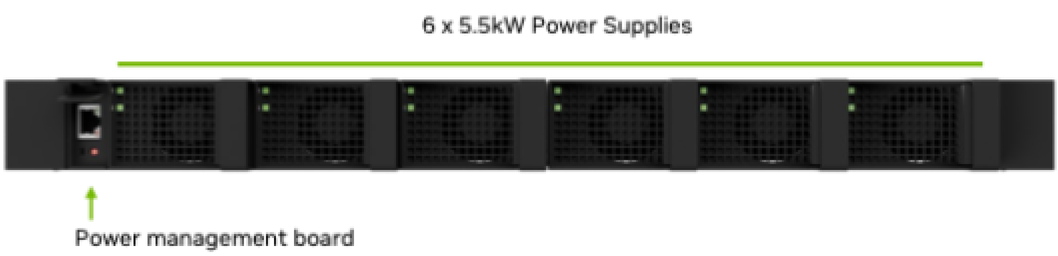


Figure 3.1: Complete Rack Layout in One Row

This reference architecture is focused on the design of a single SU (8 DGX GB200 rack systems). DGX SuperPOD can scale to much larger configurations up to and beyond 128 racks with 9216 GPUs.

Contact your NVIDIA representative for information regarding DGX SuperPOD solutions of four SUs or more.

Chapter 4. Network Fabrics

Building systems by SU provides the most efficient designs. However, if a different node count is required due to budgetary constraints, data center constraints, or other needs, the fabric should be designed to support the full SU, including leaf switches and leaf-spine cables, and leave the portion of the fabric unused where these nodes would be located. This will ensure optimal traffic routing and ensure that performance is consistent across all portions of the fabric.

DGX SuperPOD configurations utilize five network fabrics:

- NVLink 5
- Compute Fabric
- Storage Fabric
- In-Band Management Network
- Out-of-Band Management Network

These network segments are carried by four different physical fabrics:

- Multi-node NVLink Fabric (MN-NVL)
- Compute InfiniBand Fabric
- Storage and In-band Ethernet Fabric
- Out-of-Band network
- Each of these fabrics are discussed in this section.

4.1. Multi-node NVLink Fabric (NVL5)

Each DGX GB200 rack is built with 18 compute trays and 9 NVLink switch trays. Each NVLink switch tray is equipped with 2 NVLink switch chips and are responsible for the full-mesh connectivity between all 72 GPUs within the same DGX GB200 rack. Each B200 GPU features 18 NVL5 links and has one dedicated NVL5 link connectivity to each one of the 18 switch chips, delivering a total bandwidth of 1.8 TB/s low latency bandwidth.

> **Note**
>
> Each network is detailed in this section.

Figure 4.1 shows the ports on the back of the DGX B200 CPU tray and the connectivity provided. The compute fabric ports in the middle use a two-port transceiver to access all eight GPUs. Each pair of in-band management and storage ports provide parallel pathways into the DGX B200 ystem for increased performance. The OOB port is used for BMC access. (The LAN port next to the BMC port is not used in DGX SuperPOD configurations.)

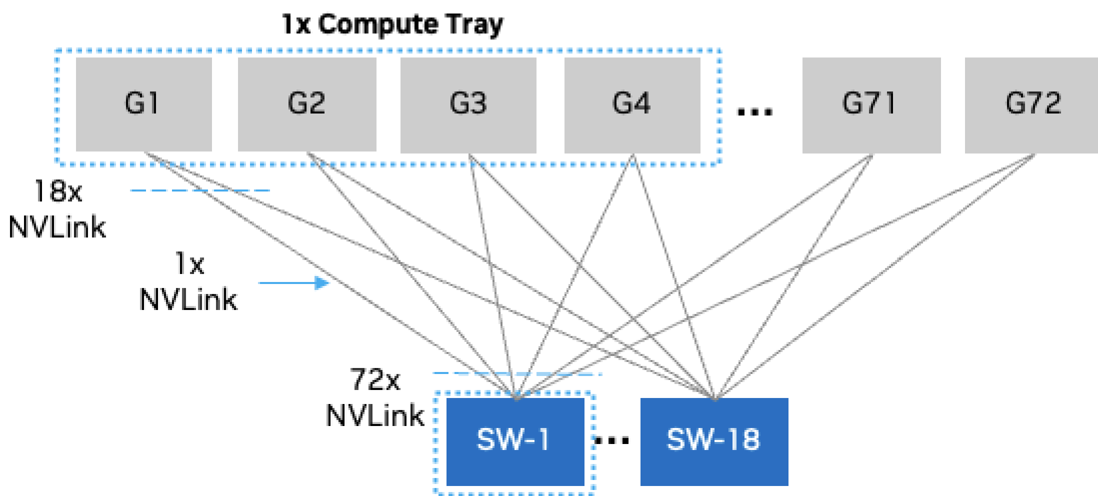


Figure 4.1: Multi-node NVLink Topology

4.2. Compute Fabric

Figure 4.2 shows the compute fabric layout for the full GB200 SuperPOD Scalable Unit (8 DGX GB200 systems). Each compute rack is rail-aligned. Traffic per rail of each compute tray is always one hop away from other compute trays in the same Scalable Unit. Traffic between different compute racks, or between different rails, traverses the spine layer.

For designs larger than one SU, we provide spine-leaf-group (SLG) based scalable design with scalability for up to and including 16 SUs. Each SU contains 4 SLGs to match with the number of IB rails (which equals the number of GPUs per compute tray). There are 8 leaf switches (one for each compute rack) and 6 spine switches in each SLG – allowing a fully non-blocking fat tree topology for each SU to be attached to 6 core groups. The details of this design for scale-out is presented in Figure 4.3. Table 4.1 shows an outline for the switches required for scale out build of DGX SuperPOD.

Table 4.1: Larger SuperPOD component counts

# GPUs	# SUs	# Core Group	Switch per Core Group	Core Switch	IB Leaf Switch		IB Spine Switch	
					Per SU	Total	Per SU	Total
1152	2	6	3	18	32	64	24	48
2304	4	6	6	36	32	128	24	96
4608	8	6	12	72	32	256	24	192
9216	16	6	24	144	32	512	24	384

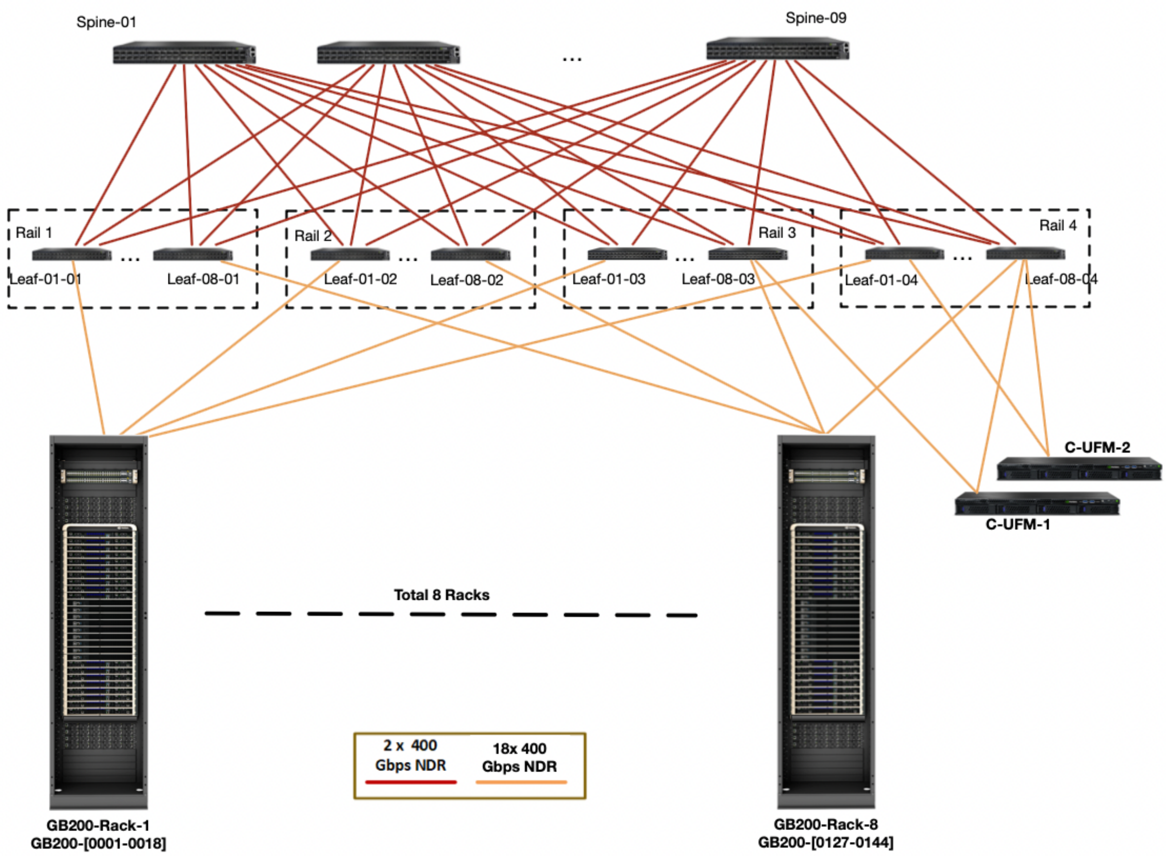


Figure 4.2: Compute fabric for full 576 GPUs DGX SuperPOD

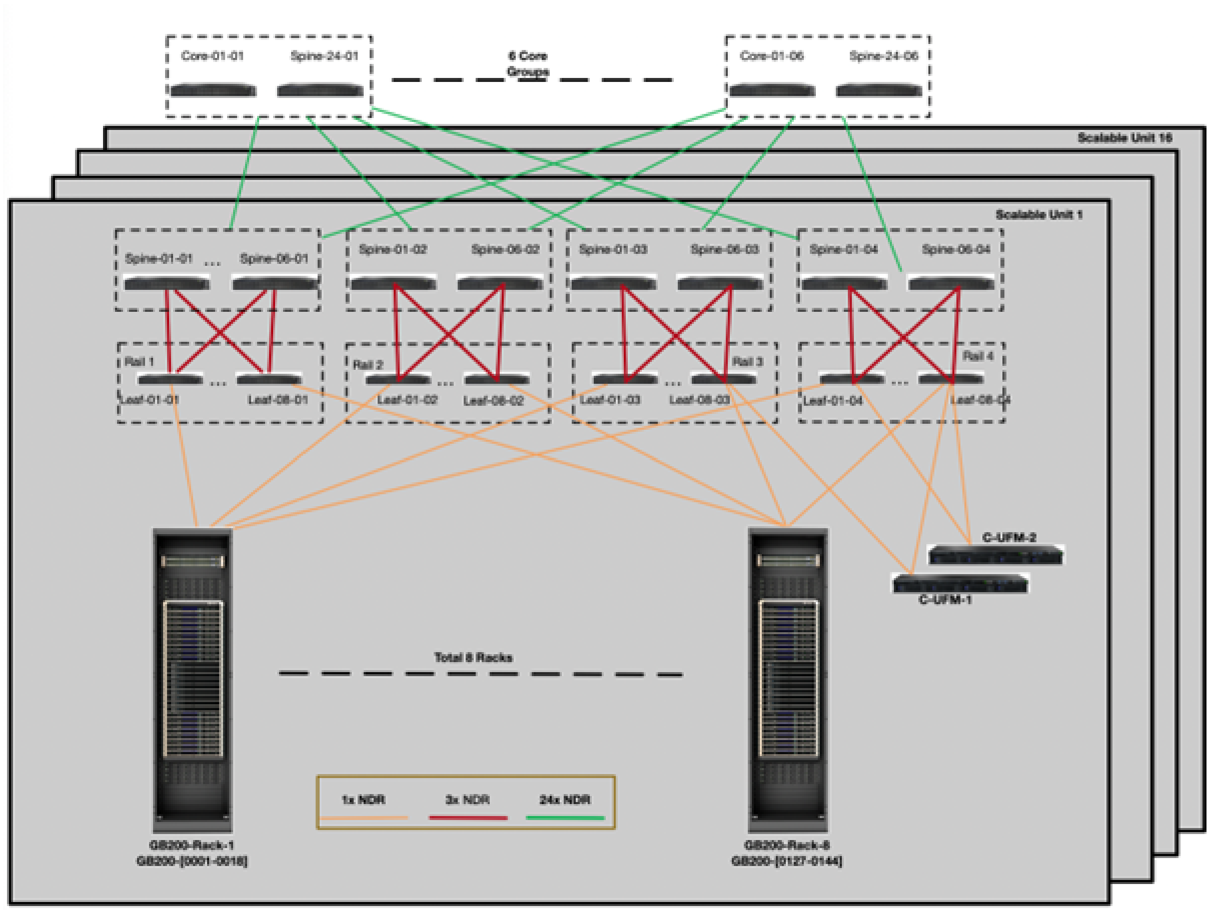


Figure 4.3: Compute Fabric for Scale Out of up to 16 SUs

4.3. Storage and In-band Ethernet Fabric

With DGX GB200, we introduced a new generation of ethernet-based fabric for storage and in-band network to enhance cost-efficiency while maintaining the high level of performance required by the storage network in a large-scale AI training cluster.

The storage and in-band fabric use SN5600 switches and SN2201 switches shown in Figure 4.4 and Figure 4.5 respectively.

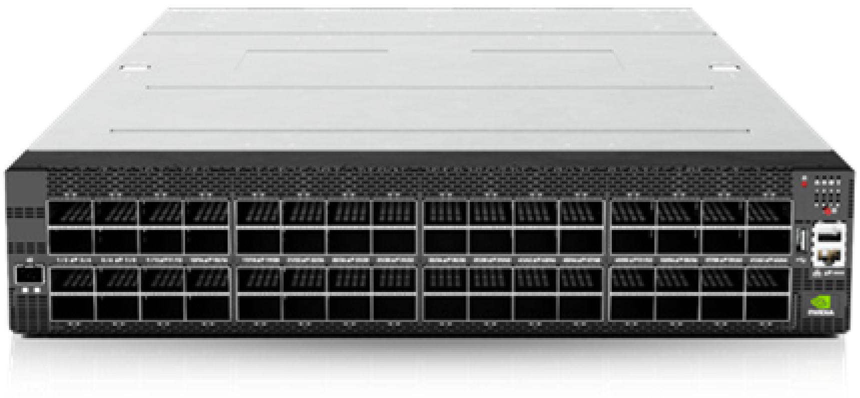


Figure 4.4: SN5600 Switch

Figure 4.5: SN2201 Switch

Figure 4.6 shows the physical layout of a single SU. Figure 12 shows the physical layout of a single SU. Each SU features two SN5600 switches as the aggregation layer or spine layer for the physical network. Here – all the leaf level switches facing DGX, Storage, and the out-of-band connection are aggregated on these pair of switches. On the leaf layer, DGX compute trays are connected at 4x 200GbE on their BlueField 3 DPUs. One additional pair of SN5600 switches serves as the ingestion point for Storage Appliances and control plane nodes. One additional pair of SN2201 switches are to connect all legacy devices requiring RJ45 connections or QSFP based uplink connectivity.

To achieve desired scale-out to up to and including 16 SUs, a third layer of switches – known as super spine – are added. Figure 4.7 shows the scale-out design for SuperPOD. Similar to the compute fabric, each SU in the SuperPOD can be implemented incrementally – given the super spine layer is built to support the maximum number of spine switches in the SuperPOD. The super spine is designed with 2 groups. Each Spine is expected to have 28x 800GbE connection uplinks to maintain non-blocking characteristics for the disaggregated storage design in the scalable SuperPOD reference architecture.

Figure 4.7 shows this design for scale out and Table 4.2 summarize the required number of switches for example size of SUs.

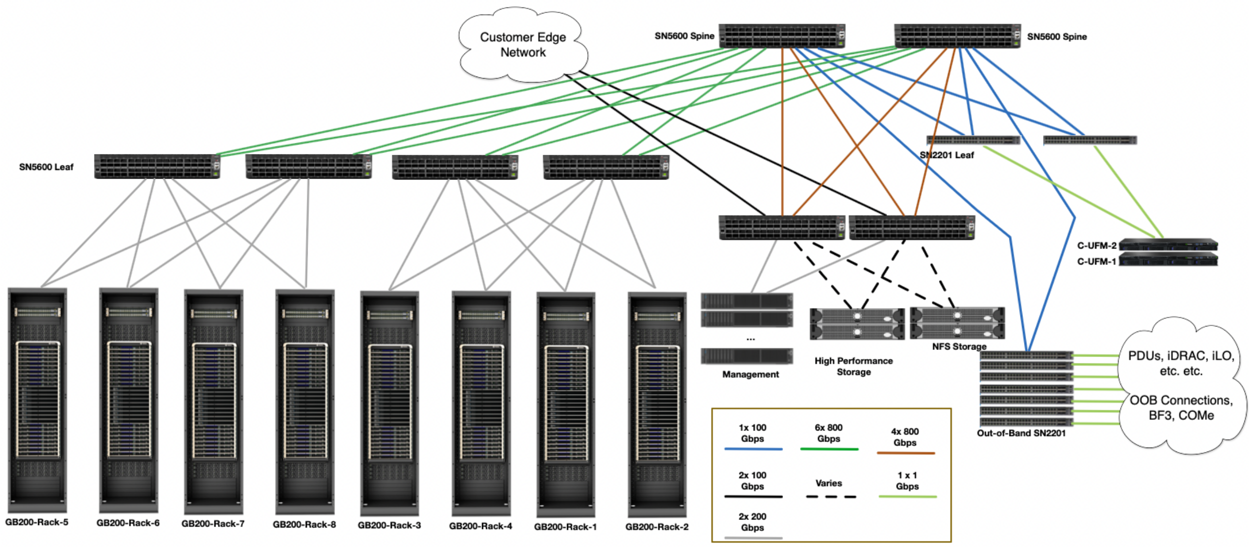


Figure 4.6: Storage and In-band Ethernet fabric logical design

Table 4.2: Spine and Super Spine Switch Requirements for Scale Out

# GPUs	# SUs	# Super Spine Groups	Super Spines per Group	# Super Spine Switches	# Spine Switches		IB Spine Switch	
					4	Total	Per SU	Total
2304	4	2	2	4	8	64	24	48
4608	8	2	4	8	16	128	24	96
9216	16	2	7	14	32	256	24	192
9216	16	6	24	144	32	512	24	384

4.4. Network Segmentation of the Ethernet Fabric

The ethernet fabric is segmented into these segments on the SuperPOD:

- Storage Network
- In-band Network
- Out-of-Bnad Management Network

In the following, we introduce these networks in detail.

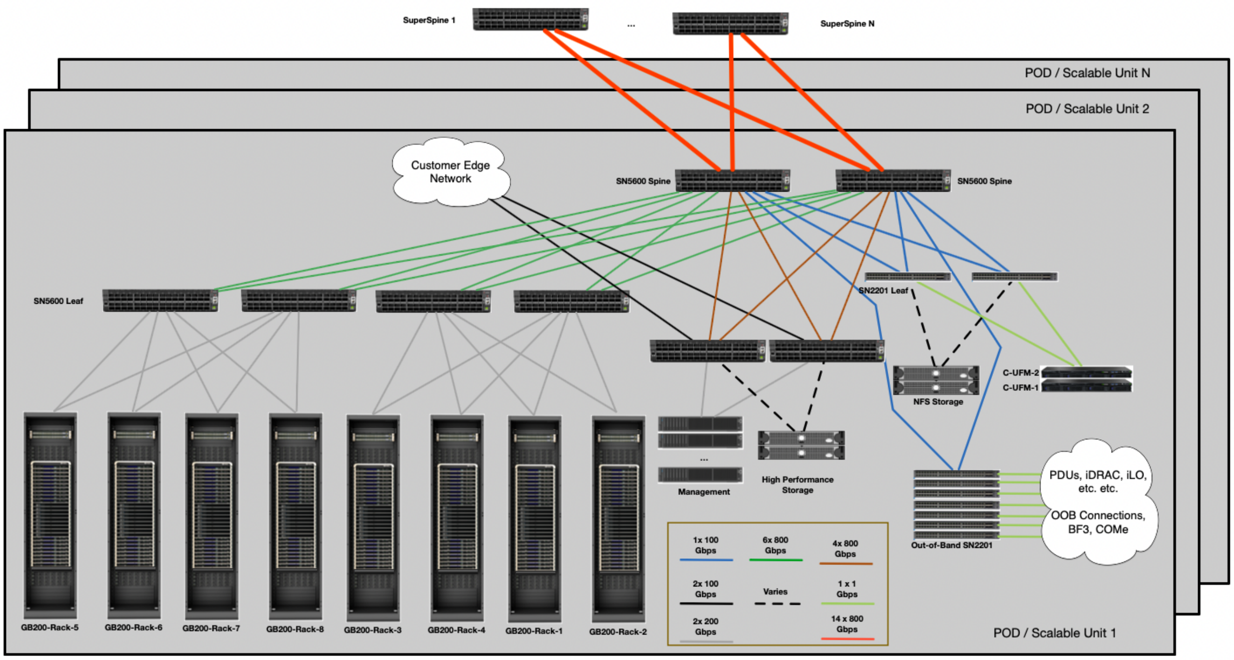


Figure 4.7: Storage and In-band Ethernet fabric scale out

4.4.1. Storage Network

The storage network embodies the performance for the high-speed storage while keeping the support for high availability. To achieve this, 2 out of the 4 available ports on each BlueField-3 DPU are dedicated for storage access.

The physical ethernet fabric carries a dedicated VXLAN with termination points on the leaf switches for the DGX nodes' storage NIC ends on each SN5600 Leaf switches. In addition, one pair of SN5600 leaf in each SU provides the connectivity to the storage appliances. RoCE is a basic requirement for the storage appliances, which benefits from advanced fabric management features, such as congestion control and AR (Adaptive Routing), provides lower latency while maintaining the high bandwidth requirement.

Each scalable unit is designed to carry 16x 800Gbps non-blocking bandwidth to the storage appliances. On the DGX node side, each scalable unit carries a slightly blocking fabric with a blocking factor of 5:3. Figure 4.8 shows the logical view of the storage fabric.

4.4.2. In-Band Management Network

The in-band management network provides several key functions:

- Connects all the services that manage the cluster.
- Enables access to the lower-speed, NFS tier of storage.
- Provides uplink (border) connectivity for the in-cluster services such as Mission Control, Base Command Manager, Slurm, and Kubernetes to other services outside of the cluster such as the NGC registry, code repositories, and data sources.
- Provides end user access to the Slurm head nodes and Kubernetes services.

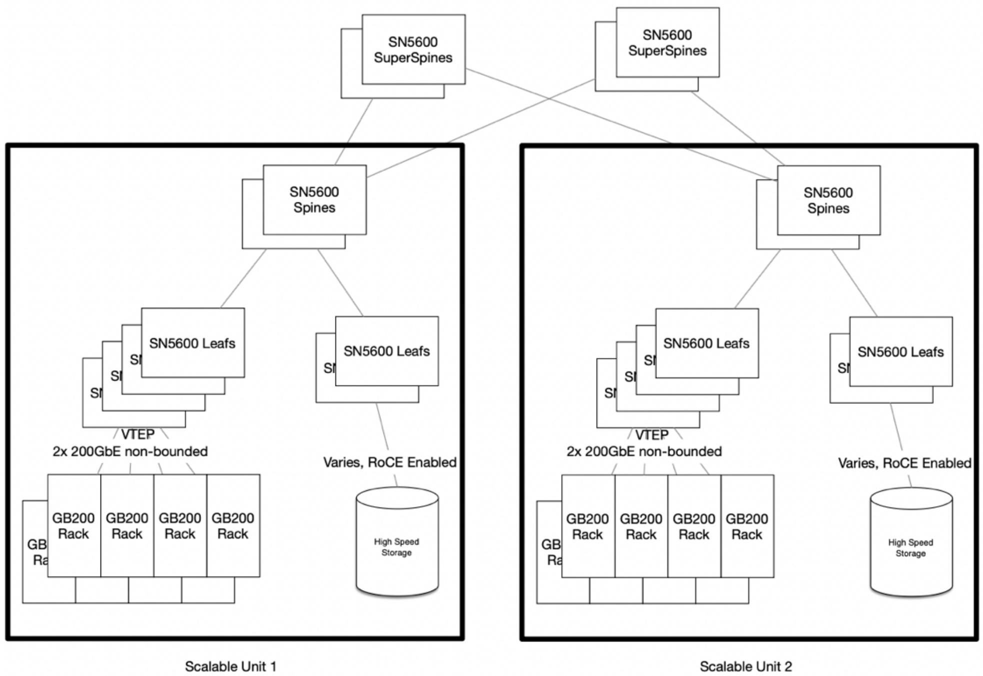


Figure 4.8: Storage Fabric Underlay Network

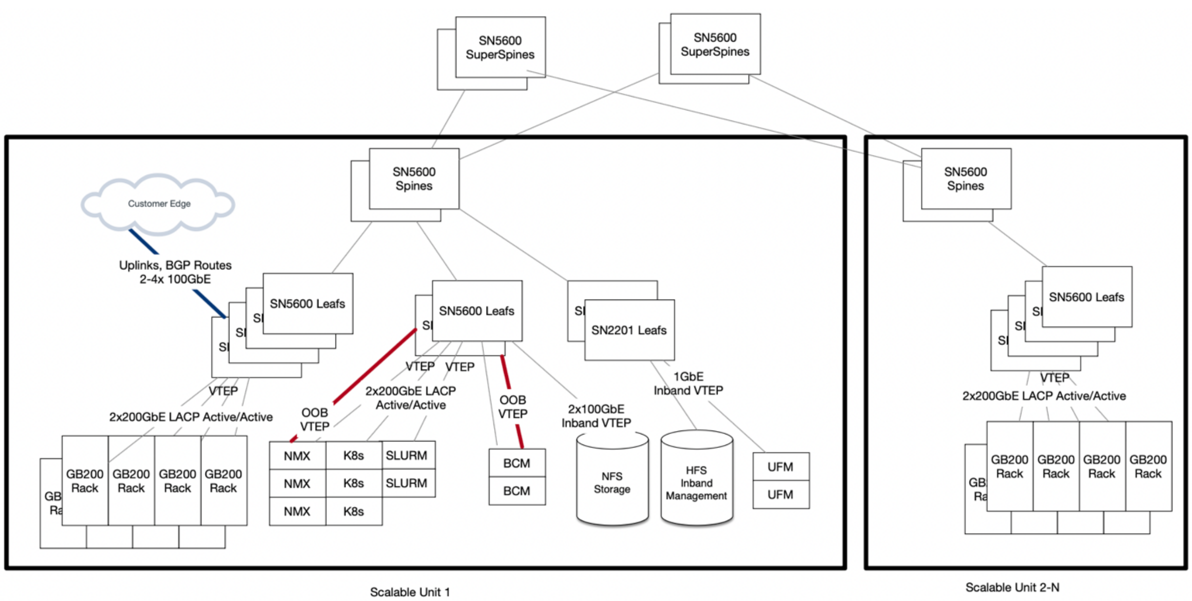


Figure 4.9: In-band Fabric Underlay Network

The in-band network itself is split into 3 different segments:

- A dedicated VTEP for uplink, the default hand over to customer edge is based on BGP peering and providing routes from in-band (and leaked routes from OOB) to customers' edge.
- Customer is also expected to provide link to the building management system network (BMS) as part of their edge connectivity.
- A dedicated VTEP for out-of-band network for the network management devices (NMX-Manager) and for BCM to access the telemetry and perform management functions to the OOB devices.
- The in-band VTEP, which carries the network for user access, home-directory storage access via NFS, service delivery, and general control traffic.

4.4.3. Out-of-Band Management Network

Figure 4.10 shows the OOB Ethernet fabric. It connects the management ports of all devices including DGX GB200 compute trays, switch trays, and management servers, storage, networking gear, rack PDUs, and all other devices. These are separated onto their own network. There is no use-case where a non-privileged user needs direct access to these ports and are secured using logical network separation.

The OOB network carries all IPMI related control traffic and serves as the network for fabric management of the compute InfiniBand fabric and compute NVLink fabric.

The OOB network is physically rolled up into the aggregation layer (spine layer) of each SU as a dedicated VXLAN. The OOB management network use SN2201 switches, shown in Figure 4.11

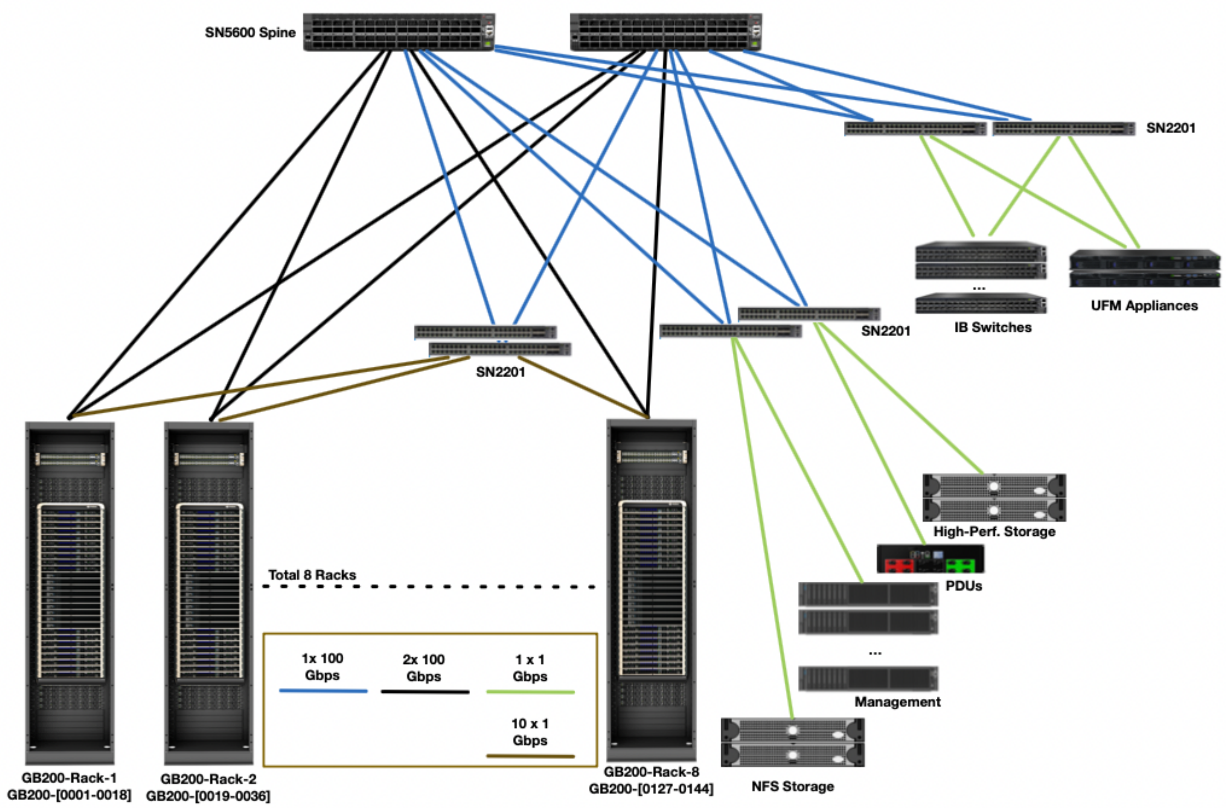


Figure 4.10: Logical OOB management network layout

Figure 4.11: SN2201 switch

4.5. Customer Edge Connectivity

For connecting DGX SuperPOD to customer edge for uplink and customer corporate network access, we recommend at least 2x 100GbE links with DR1 single-mode connectivity to cope with the growing demand on high-speed data transfer into and from DGX SuperPOD.

A new introduction to DGX SuperPOD with DGX GB200 systems because of its complex cooling and power requirement is the connection to the Building Management System (BMS). The BMS serves is the management plane of the operational technology (OT) side of the data center infrastructure.

For route handover, we prepare eBGP protocol to peer with customer's network. Routes to/from in-band, out-of-band, and building management system are announced. Figure 4.12 shows an example for customer edge connectivity.

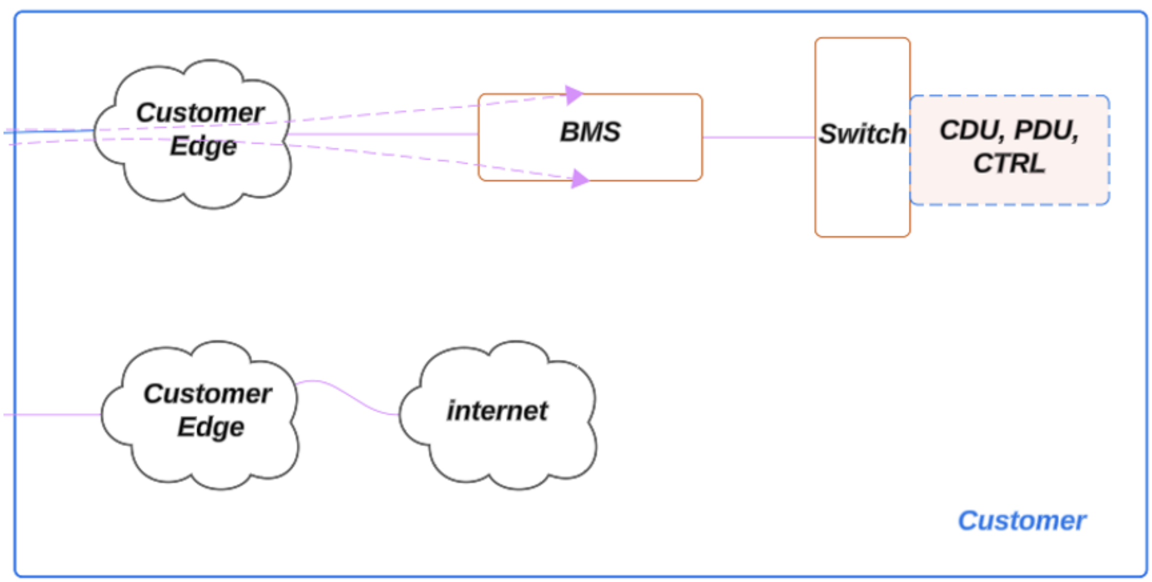


Figure 4.12: Example Customer Edge Connectivity

Chapter 5. Storage Architecture

Data, lots of data, is the key to development of accurate deep learning (DL) models. Data volume continues to grow exponentially, and data used to train individual models continues to grow as well. Data format, not just volume can play a key factor in the rate at which data is accessed, so storage system performance must scale commensurately.

The key I/O operation in DL training is re-read. It is not just that data is read, but it must be reused again and again due to the iterative nature of DL training. Pure read performance still is important as some model types can train in a fraction of an epoch (ex: some recommender models) and inference of existing can be highly I/O intensive, much more so than training. Write performance can also be important. As DL models grow and require more time to train, writing checkpoints is necessary for fault tolerance. The size of checkpoint files can be terabytes in size, and while not written frequently, are typically written in a synchronously way that blocks forward progress of training process.

Ideally, data is cached during the first read of the dataset, so data does not have to be repeatedly retrieved across the network. Shared filesystems typically use RAM as the first layer of cache. Reading files from cache can be an order of magnitude faster than from remote storage. The DGX GB200 system provides local NVMe storage that can also be used for caching or staging data.DGX GB200 systems provide local NVMe storage that can also be used for caching or staging data.

DGX SuperPOD is designed to support all workloads, but the storage performance required to maximize training performance can vary depending on the type of model and dataset.

High-speed storage provides a shared view of an organization's data to all nodes. It must be optimized for small, random I/O patterns, and provide high peak node performance and high aggregate filesystem performance to meet the variety of workloads an organization may encounter. High-speed storage should support both efficient multi-threaded reads and writes from a single system, but most DL workloads will be read-dominant.Use cases in automotive and other computer vision-related tasks, where high-resolution images are used for training (and in some cases are uncompressed) involve datasets that easily exceed 30 TB in size. In these cases, 4 GBps per GPU for read performance is needed.

While NLP and LLM cases often do not require as much read performance for training, peak performance for reads and writes are needed for creating and reading checkpoint files. This is a synchronous operation, and training stops during this phase. If you are looking for best end-to-end training performance, do not ignore I/O operations for checkpoints. Consider at least ½ of the read performance as recommended write performance for LLM and large model use cases.

The preceding metrics assume a variety of workloads, datasets, and need for training locally and directly from the high-speed storage system. It is best to characterize workloads and organizational needs before finalizing performance and capacity requirements. Table 5.1 and Table 5.2 are provided to help determine the I/O levels required for different types of models.

Table 5.1: Storage Performance Requirements

Level	Work Description	Data Set Size
Standard	Multiple concurrent LLM or fine-tuning training jobs and periodic checkpoints, where the compute requirements dominate the data I/O requirements significantly.	Most datasets can fit within the local compute systems' memory cache during training. The datasets are single modality, and models have millions of parameters.
Enhanced	Multiple concurrent multimodal training jobs and periodic checkpoints, where the data I/O performance is an important factor for end-to-end training time.	Datasets are too large to fit into local compute systems' memory cache requiring more I/O during training, not enough to obviate the need for frequent I/O. The datasets have multiple modalities and models have billions (or higher) of parameters.

Table 5.2: Guidelines for storage performance

Performance Characteristic	Standard (GBps)	Enhanced (GBps)
Single SU aggregate system read	40	125
Single SU aggregate system write	20	62
4 SU aggregate system read	160	500
4 SU aggregate system write	80	250

Chapter 6. DGX SuperPOD Software

NVIDIA Mission Control software delivers a full-stack data center solution engineered for enterprise infrastructure deployments, like NVIDIA DGX SuperPOD. It integrates essential management and operational capabilities into a unified platform, providing enterprise customers with seamless control over their infrastructure deployments at scale.

Figure 6.1 shows the detailed decomposition of DGX GB200 SuperPOD software stack.

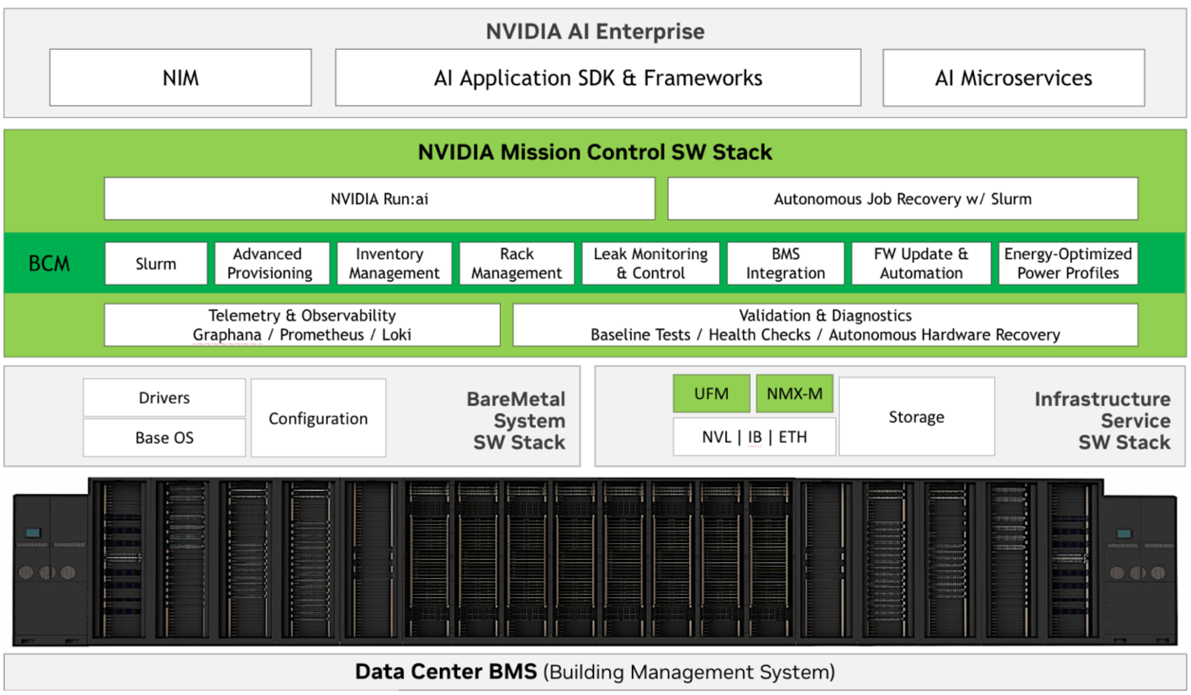


Figure 6.1: NVIDIA DGX GB200 SuperPOD Software Stack

NVIDIA Mission Control software stack comprises a five-layer architecture that builds from foundational system health and diagnostics to advanced NVIDIA DGX SuperPOD cluster management operations. It leverages NVIDIA Base Command Manager (BCM) technology and NVIDIA Run:ai functionality to provide scheduler access with SLURM and Kubernetes for AI and HPC workload orchestration. The Telemetry and Observability layer leverages proprietary diagnostic tools and health-checks for system insights, while the Validation and Diagnostics layer, supports the built-in autonomous recovery engine that ensures rapid failure recovery and system restoration with minimal time to recovery.

NVIDIA Mission Control also handles deployment optimization and system health monitoring, seamlessly cooperating NVIDIA Base Command Manager (BCM) technology for coordinating cluster provisioning & operations. Included in the software stack there is also Network Management eXpert (NMX) for monitoring & control of NVLINK Switch trays.

NVIDIA Mission Control delivers critical innovations that directly impact DL and HPC environments, improving efficiency, reducing downtime, and optimizing resource utilization. Here's how these advancements translate into tangible business value:

Automated Failure Detection & Self-Recovery:

NVIDIA Mission Control's automated failure detection and recovery mechanisms significantly minimize downtime compared to manual interventions. With faster recovery times, AI training continues without significant disruption, resulting in improved GPU utilization. This reduces the risk of resource wastage.

- Optimized Workload Migration & Resource Allocation: NVIDIA Mission Control ensures that workloads are continuously reassigned to healthy nodes, preventing idle GPU time. This leads to increased overall GPU efficiency, enabling the system to process more workloads within the same timeframe. By optimizing resource allocation, NVIDIA Mission Control helps organizations achieve higher throughput for AI training tasks, thereby accelerating time-to-results.
- Unified Diagnostics for Infrastructure & Applications: By combining infrastructure and application-level diagnostics, NVIDIA Mission Control helps significantly reduce troubleshooting time. Model builders can identify and resolve issues much faster compared to traditional methods. This considerably decreases the operational burden on SREs and DevOps, leading to savings in personnel costs and reducing the time to production.
- In-Memory Checkpointing for Seamless Job Recovery: NVIDIA Mission Control enables recovery of training workloads by automatically restarting failed processes/ranks from the most recent valid checkpoint, ensuring no data loss or rework. In environments with frequent failures, it significantly reduces retraining time, thereby eliminating downtime and preventing interruptions in the training loop. This results in an increase in model iteration speed and faster go-to-market timelines for AI products.

6.1. Run:ai

Included with NVIDIA Mission Control,he Run:ai platform employs a distributed architecture consisting of two primary components: the control plane (backend) and the cluster. The control plane serves as the central management system, capable of orchestrating multiple Run:ai clusters across an organization. For BCM-managed installations, the control plane is hosted on the control plane while cluster components are deployed directly onto the customer's Kubernetes infrastructure. This separation of responsibilities allows for centralized management while maintaining workload execution close to computational resources.

Figure 6.2 showcases the architecture of NVIDIA Run:ai

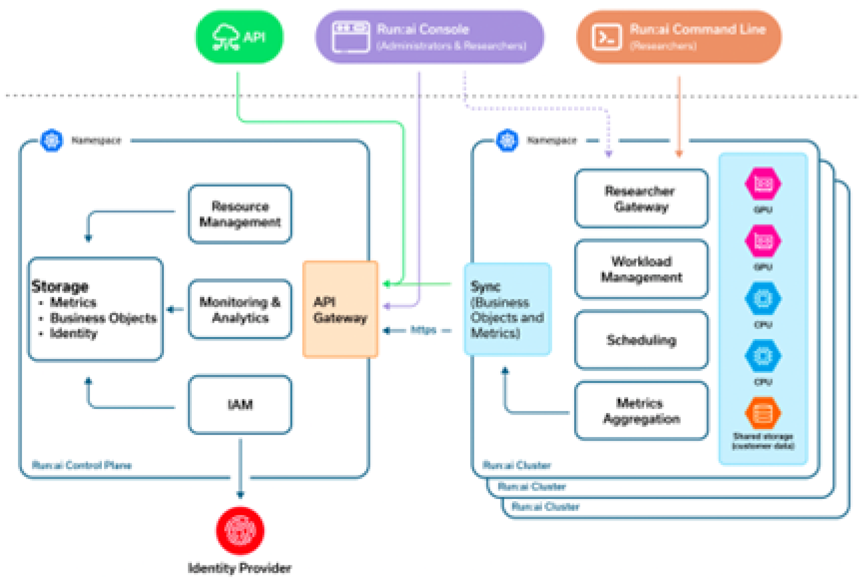


Figure 6.2: NVIDIA Run:ai

Chapter 7. Summary

NVIDIA DGX SuperPOD with NVIDIA DGX GB200 systems is the next generation of data center scale architecture to meet the demanding and growing needs of AI training and inferencing. This reference architecture document for DGX SuperPOD represents the architecture used by NVIDIA for our own AI model and HPC research and development. DGX SuperPOD continues to build upon its high-performance roots to enable training of the largest NLP models, support the expansive needs of training models for automotive applications, and scaling-up recommender models for greater accuracy and faster turn-around-time.

DGX SuperPOD represents a complete system of not just hardware but all the necessary software to accelerate time-to-deployment, streamline system management, proactively identify system issues. The 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 platform is designed to both support the workloads of today and grow to support tomorrow's applications. These are representatives of the configuration and must be finalized based on actual design. Your NVIDIA representative can work with you to finalize the exact components list.

Chapter 8. Notices

8.1. Notice

8.2. Trademarks

8.3. Copyright

Copyright