NVIDIA DGX GH200 AI Supercomputer

AI Supercomputer for the Generative AI Era

White Paper

WP-11400-001_v01 | June 2023

Table of Contents

List of Figures

List of Tables

Introduction

The NVIDIA® DGX™ platform provides the most powerful and comprehensive AI infrastructure for enterprises to adopt the rapid innovations of Large Language Models (LLM), data analytics, graph analytics, and generative AI to transform their entire business model. In this whitepaper, we introduce the latest NVIDIA DGX system, walk through the technological innovation at the system hardware level, system software level, and data center full-stack level to showcase how adopting such a system is as easy as traditional IT infrastructure.

Today's mainstream AI models contain billions of parameters, requiring tremendous amounts of compute, memory storage, and system connectivity for model training and deployment. Moore's Law has not kept pace with these exponential increases in computing demand. To overcome Moore's law, enterprises adopt AI the NVIDIA way, scaling up and scaling out their diverse AI and data analytics applications with NVIDIA DGX systems. The core of every DGX system is NVLink-connected GPUs that access each other's memory at NVLink speed. Many DGX systems are interconnected with high-speed networking to form supercomputers such as NVIDIA Selene.

NVIDIA DGX GH200 is the newest supercomputer in the DGX family. It provides an ExaFLOP of FP8 (with sparsity) compute performance and up to 144 Terabytes of fast GPU-accessible memory with 128 TB/s bisection bandwidth. DGX GH200 empowers AI scientists and practitioners to train giant and emerging models and solve the world's largest challenges.

NVIDIA DGX GH200 Overview

NVIDIA DGX GH200 (see Figure 1) brings together the class-leading performance of the NVIDIA GH200 Grace Hopper™ Superchip, NVIDIA NVLink-C2C, NVIDIA NVLink-C2C, NVIDIA NVLink™ Switch System, NVIDIA Quantum-2 InfiniBand networking, and a comprehensive and optimized NVIDIA DGX software stack that includes NVIDIA Base Command™ and NVIDIA AI Enterprise.

The NVIDIA GH200 Grace Hopper Superchip (see Table 1 and Figure 2) is the first true heterogeneous accelerated platform for AI workloads that accelerates applications with the strengths of both GPUs and CPUs. It improves developer productivity by accelerating their programming model with features such as bulk data transfers, NVIDIA Magnum IO™ for NCCL, MPI, NVSHMEM, and direct memory accesses. It combines the NVIDIA Grace CPU and NVIDIA Hopper GPU with NVIDIA NVLink-C2C into a single package with up to 576 GB of fast accessible GPU memory, 5x more memory than what is available in a standalone NVIDIA Hopper GPU's HBM memory. NVIDIA NVLink-C2C is a memory coherent, high-bandwidth, low-latency, and energy-efficient superchip interconnect that delivers up to 900 GB/s bi-directional bandwidth, 7x higher bandwidth than x16 PCIe Gen5 lanes commonly used in traditional x86 systems.

The NVIDIA NVLink Switch System connects up to 256 NVIDIA Grace Hopper Superchips. The resulting NVIDIA DGX GH200 system delivers up to an ExaFLOP of FP8 (with Sparsity) AI performance and up to 144 TB of GPU memory with 115.2 TB/s NVLink bisection bandwidth or 128 TB/s bisection bandwidth considering both NVLink and InfiniBand. This tremendous increase of more than 460 times higher capacity of fast GPU-accessible memory over the original DGX A100 320GB system enables DGX GH200 to address the most demanding giant AI and data analytics workloads and makes it the first supercomputer to break the 100-terabyte barrier of fast GPU-accessible memory over NVLink.

The NVIDIA NVLink Switch System is a two-level NVLink network. In the first level, groups of 8x NVIDIA Grace Hopper Superchip modules are connected with three NVLink switches that deliver 3.6 TB/s of bisection bandwidth. Thirty-two 8-GPU building blocks are interconnected in the second level by 36 NVLink switches to form the 256-GPU DGX GH200 system.

Figure 1. NVIDIA DGX GH200 AI Supercomputer

NVIDIA GH200 Grace Hopper Superchip

Figure 2 shows a logical overview of the NVIDIA Grace-Hopper Superchip. Table 1 lists the NVIDIA Grace-Hopper Superchip specifications.

Figure 2. NVIDIA Grace Hopper Superchip Logical Overview

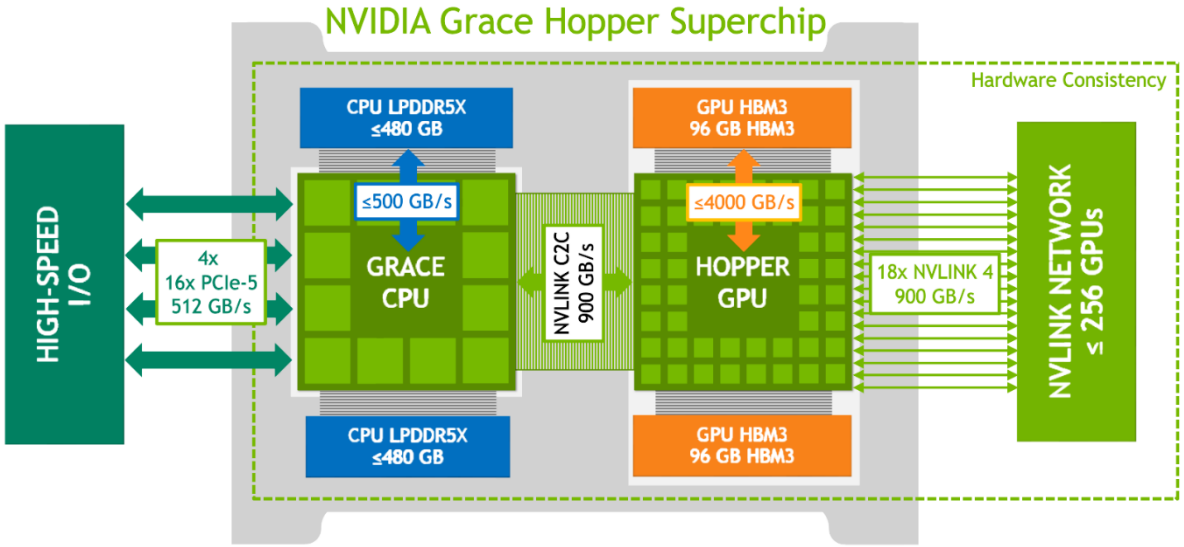


Table 1. NVIDIA GH200 Grace Hopper Superchip Specifications

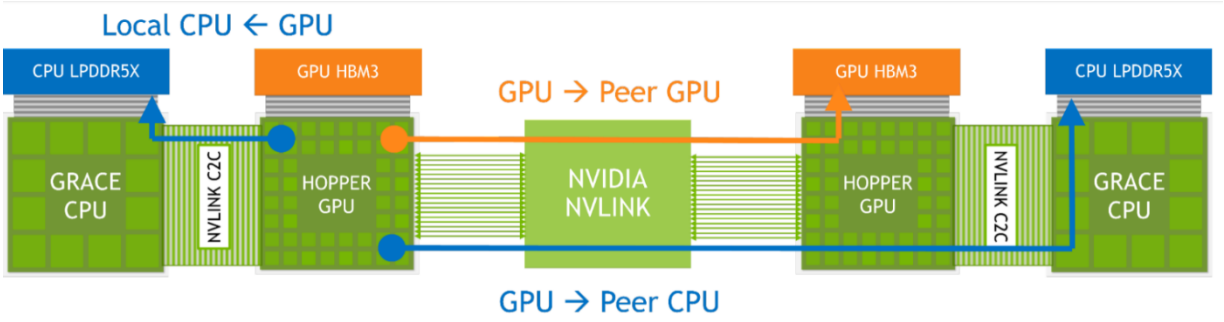
GPU	Hopper 96 GB HBM3, 4 TB/s
CPU	72 Core Arm Neoverse V2
CPU Memory	Up to LPDDR5 480 GB at up to 500 GB/s, 4x more energy efficient than DDR5
CPU-to-GPU NVLink-C2C	900 GB/s bi-directional coherent link 5x more energy efficient than PCIe Gen5.
GPU-to-GPU NVLink	900 GB/s bi-directional
High-speed I/O	4x PCIe Gen5 x16 at up to 512 GB/s
TDP	Configurable from 450W to 1000W

Extended GPU Memory

The NVIDIA Grace Hopper Superchip, with its combined GPU and CPU memory subsystems, is designed to accelerate applications with exceptionally large memory footprints, significantly larger than the capacity of the HBM3 and LPDDR5X memory subsystems of each Grace Hopper Superchip (see Figure 3).

The Extended GPU Memory (EGM) feature over NVIDIA NVLink, and NVIDIA NVLink-C2C enables GPUs to access up to 144 TBs of memory from all CPUs and GPUs in the system at the minimum of NVLink or LPDDR5X or HBM3 bandwidth. The Extended GPU Memory (EGM) can be accessed for loads, stores, and atomic operations by the connected CPU at LPDDR5X speeds. It can also be accessed by the GPU in the Grace Hopper Superchip over the NVIDIA NVLink-C2C and by other GPUs in DGX GH200 over NVIDIA NVLink, thanks to Magnum IO GPUDirect. EGM can also be used with standard MAGNUM IO libraries with NCCL, UCX, MPI, and NVSHMEM memory programming (see Figure 3).

Figure 3. Memory Accesses Across NVLink Connected Grace Hopper Superchips

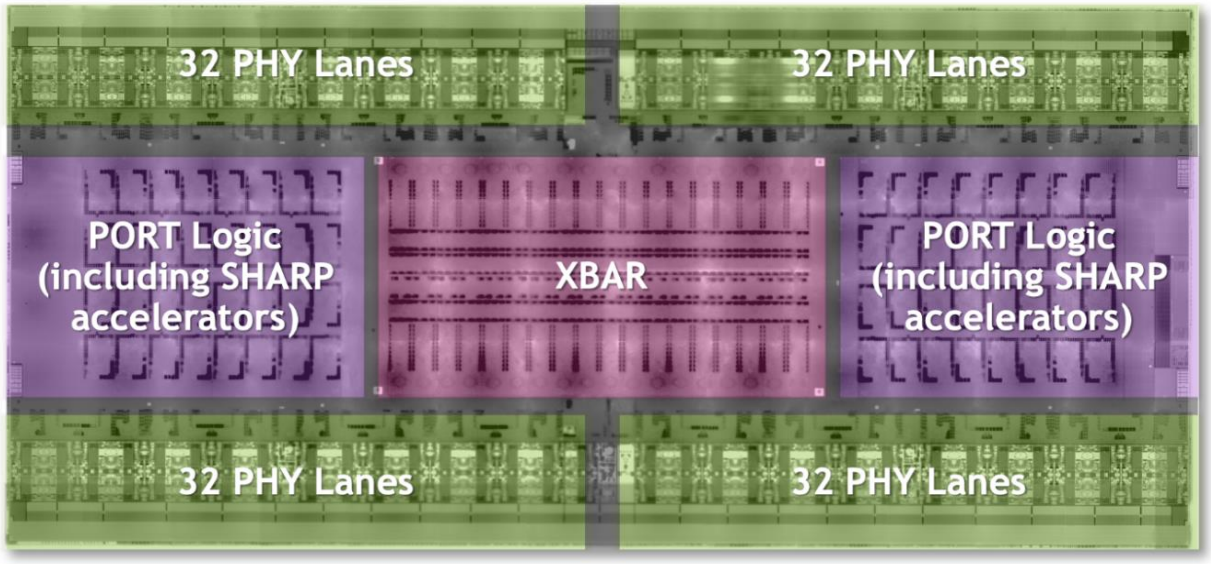


NVIDIA NVLink Switch System

While the NVIDIA NVlink-C2C interconnect provides super-fast and low-latency connectivity between the Grace CPU and the Hopper GPU, the NVIDIA NVLink Switch System extends high-bandwidth, low-latency connectivity using the fourth-generation NVLink technology and third-generation NVIDIA sNVSwitch™ architecture (see Figure 4) to all 256 NVIDIA Grace Hopper Superchips in a DGX GH200 system.

The NVIDIA NVLink Switch is a rackmount switch that delivers an unprecedented 25.6 Terabits per second (Tbps) of full duplex bandwidth for the fourth-generation NVLink in a 1U chassis design. Each NVLink switch includes two third-generation NVIDIA NVSwitch™ ASICs and exposes 128 NVLink fourth-generation ports. Each NVLink port consists of two lanes (fixed-width) with a per-link raw bandwidth of 26.6 GB/s. Together, the two-level NVLink switches form the NVLink Switch System.

Figure 4. NVIDIA Third-Generation NVSwitch using Fourth-Generation NVLink Technology Logical Overview

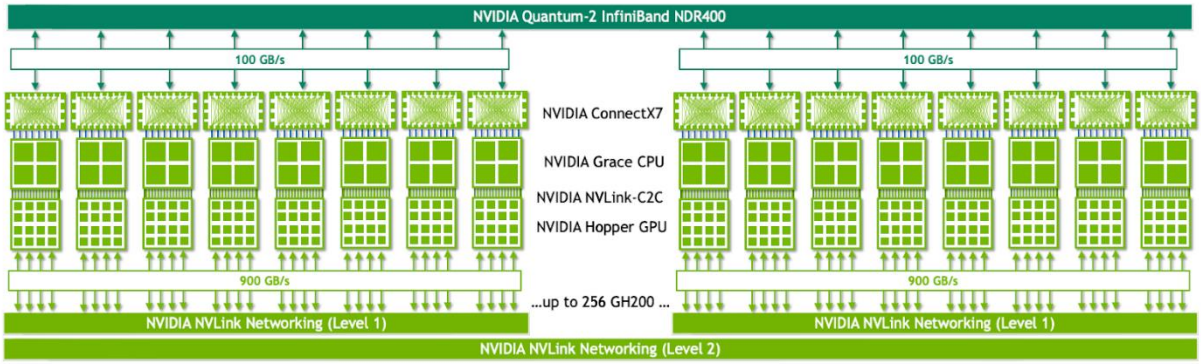


Compute trays hosting Grace Hopper Superchips are connected to the NVLink Switch System using a custom cable harness for the first layer of NVLink fabric. Each second-level NVLink switch exposes 128 NVLink fourth-generation ports via 32 OSFP cages. NVIDIA LinkX® cables extend the connectivity in the second layer of NVLink Switches.

The NVLink Switch System forms a two-level, non-blocking, fat tree NVLink fabric to fully connect 256 NVIDIA Grace Hopper Superchips in a DGX GH200 system with 115.2 TB/s of bisection bandwidth, 9x the all-to-all bandwidth of the NVIDIA Quantum-2 InfiniBand switch with NDR400. The NVIDIA NVLink Switch System allows all GPUs running on up to 256 NVLink-connected Grace Hopper Superchips to access up to 144 TB of memory at low latency and high bandwidth.

In the DGX GH200 system, every GPU can address peer HBM3 and LPDDR5X memory from other Grace Hopper Superchips in the NVLink network. NVIDIA Magnum IO™ acceleration libraries optimize GPU communications for efficient application scaling with all 256 GPUs. The DGX GH200 delivers 230.4 TFLOPS of NVIDIA SHARP™ in-network computing to accelerate AI collective operations and doubles the effective bandwidth of the NVLink Switch System network (see Figure 5) by reducing communication overheads.

Figure 5. NVIDIA NVLink Switch System connects upto 256 NVIDIA Grace Hopper Superchips



Management and Networking Subsystem

Each compute baseboard hosting an NVIDIA Grace Hopper Superchip in the DGX GH200 has its own management and I/O subsystem, accessible through a GbE management port. Each NVLink switch also has its management subsystem accessible through a GbE port.

For networking, each Grace Hopper Superchip is paired with an NVIDIA ConnectX®-7 network adapter and an NVIDIA BlueField®-3 DPU, both directly connected to the Grace CPU root complex.

In addition to supporting protocols like MPI, gRPC, and so on for DGX GH200, NVIDIA ConnectX-7 adapters can interconnect multiple DGX GH200 systems for scaling beyond 256 GPUs.

BlueField-3 enables in-band and storage network connectivity to every Grace Hopper Superchip in the DGX GH200 system and transforms enterprise computing environments into secure, accelerated virtual private clouds for running application workloads in multi-tenant environments.

.

DGX GH200 System Architecture

The DGX GH200 supercomputer has 256 GH200 Grace Hopper compute trays and an NVLink Switch System, forming a two-level NVLink fat tree. Each compute tray contains an NVIDIA GH200 Grace Hopper Superchip, networking components (ConnectX-7 InfiniBand/Ethernet, BlueField-3), a management system (BMC), and SSDs for data and the operating system. Eight compute trays connect to three first-level NVLink NVSwitch trays to form a single 8-GPU chassis. Each NVLink Switch tray contains two NVSwitch ASICs that connect to the compute trays with a custom blind mate cable cartridge and second-level NVLink Switches with LinkX cables. 36 second-level NVLink Switches connect 32 chassis to form the NVIDIA DGX GH200 supercomputer. Table 2 contains the specifications of compute tray with Grace Hopper superchip. Table 3 lists the NVLink Switch specifications.

Table 2. NVIDIA Grace Hopper Compute Tray

CPU/GPU	1x NVIDIA Grace Hopper Superchip with NVLink-C2C
GPU/GPU	18x NVLink fourth-generation ports
Networking	1x NVIDIA ConnectX-7 with OSFP: > NDR400 InfiniBand Compute Network 1x Dual port NVIDIA BlueField-3 with 2x QSFP112 or 1x Dual port NVDIA ConnectX-7 with 2x QSFP112: > 200 GbE In-band Ethernet network > NDR200 IB storage network Out of Band Network: > 1 GbE RJ45
Storage	Data Drive: 2x 4 TB (U.2 NVMe SSDs) SW RAID 0 OS Drive: 2x 2 TB (M.2 NVMe SSDs) SW RAID1

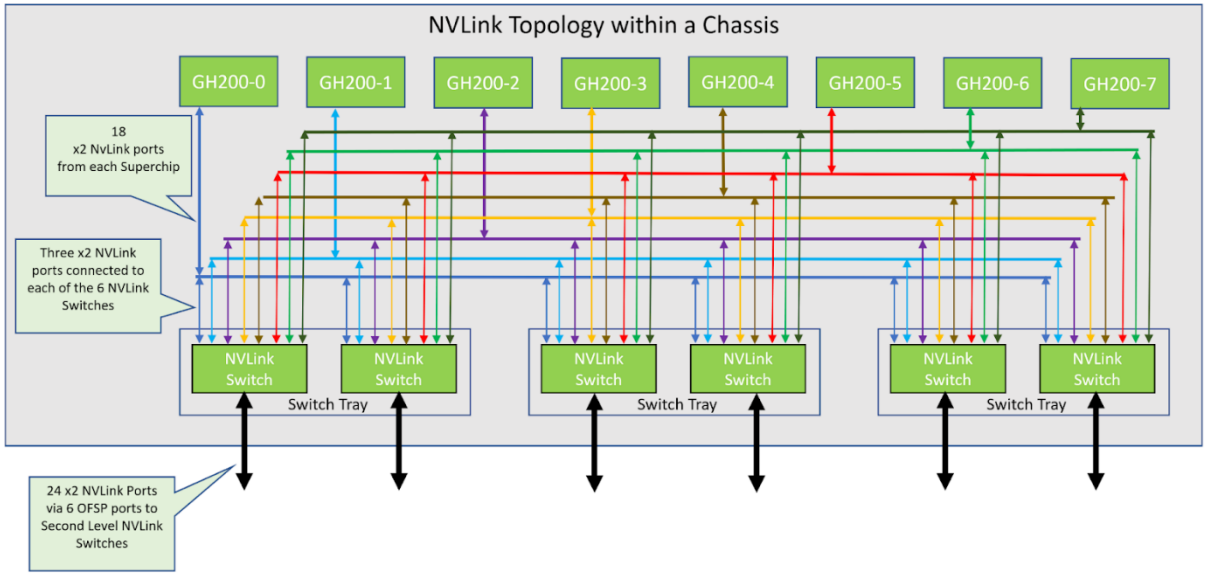
Table 3. NVLink Switch Specifications

NVSwitch	2x Third-Generation NVSwitch ASIC supporting NVLink fourth-generation
NVLink Ports	48x NVLink to Compute trays through passive cable cartridge. Inside Chassis > 6x NVLink per Compute tray > 12x OSFP (48x NVLink) to connect to second-level Switches

Each NVIDIA GH200 Grace Hopper Superchip has eighteen 53.1 GB p/s bi-directional bandwidth NVLink ports. Three ports connect to each of the six NVLink ASICs in the Switch trays through a custom cable cartridge, fully connecting all eight Grace Hopper Superchips within a chassis.

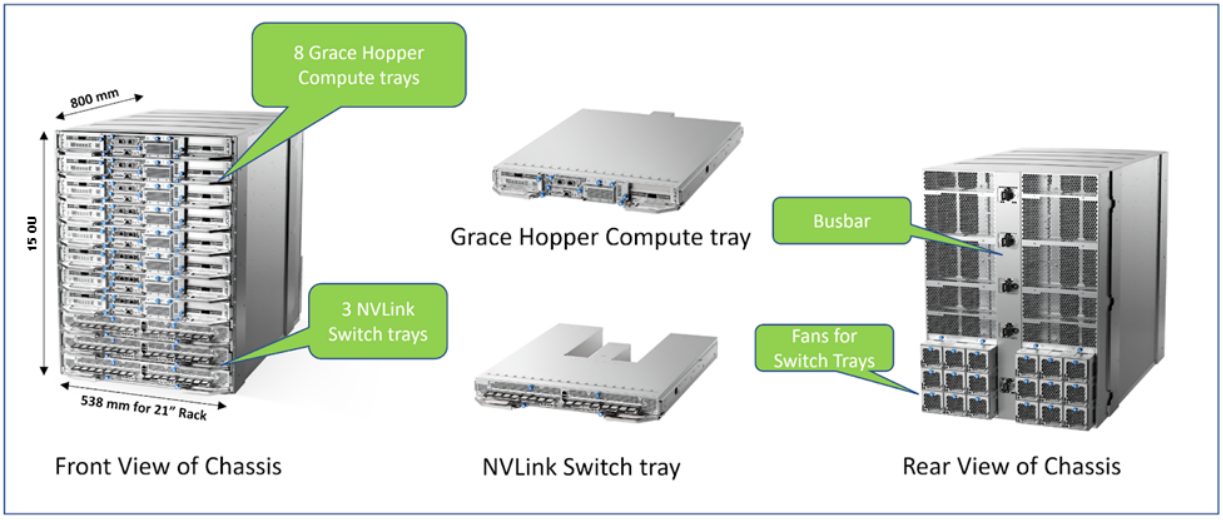
Each chassis's three NVLink switch trays provide 36 Octal Small Form Factor Pluggable (OSFP) ports for Switch-to-Switch connectivity between the 32 chassis (see Figure 6). Twelve additional OSFP ports are available for redundancy or to support alternate topologies.

Figure 6. NVLink topology within an 8-GraceHopper Superchip Chassis



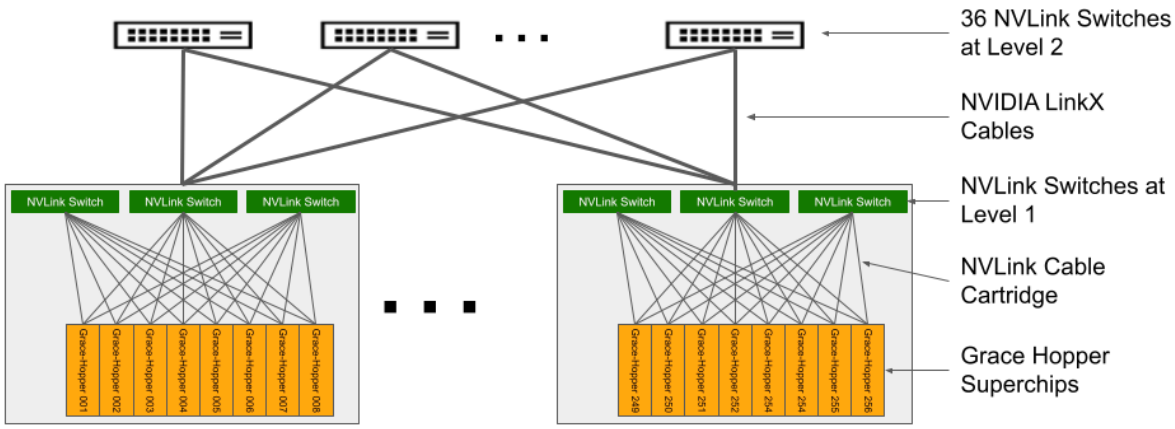
A 15 OU chassis houses eight compute trays and three NVSwitch trays, with airflow designed to flow from front to back (see Figure 7).

Figure 7. 8-Grace Hopper Superchip Chassis



Thirty-six NVLink Rack Switches in the second level of the NVLink Switch system connect 32 chassis for full connectivity, as shown in Figure 8.

Figure 8. NVLink Topology



Network Architecture

A DGX GH200 system also includes four other networks:

Compute InfiniBand Fabric

NVIDIA ConnectX-7 and NVIDIA Quantum-2 switches form a full-fat tree NDR400 InfiniBand fabric. The InfiniBand compute fabric is a rail-optimized, balanced, full-fat tree. Managed NDR switches provide better fabric management and support the latest SHARPv3 features. Compute InfiniBand network is designed to support protocols like MPI and gRPC. It also allows connectivity between multiple DGX GH200 systems to form even larger and more capable systems.

Storage Fabric

The NVIDIA BlueField-3 DPU (Data Processing Unit) powers the high-performance storage InfiniBand fabric for DGX GH200 over one QSFP112 port on each DPU, capable of NDR200 data rate. The storage traffic is separated from the compute traffic to prevent network congestion and maintain peak application performance. Storage network architecture is flexible, and the capacity and capability of the attached storage solution are configurable to meet the specific needs of the workloads.

In-band Management Fabric

The GPU in-band Ethernet fabric serves several essential functions. It connects all services managing the system and provides access to the home filesystem and storage pool through NFS. It also enables connectivity for in-system services like Slurm and Kubernetes and external services such as the NVIDIA NGC™ (NVIDIA GPU Cloud) registry, code repositories, and data sources. The in-band network connects the Grace Hopper superchip baseboards and management nodes. In addition, the out-of-band (OOB) network is connected to the in-band network to provide high-speed interfaces from the management nodes to support parallel operations to devices connected to the OOB storage fabric, such as storage.

Out-of-band Management Fabric

Operating at 1GbE, the out-of-band fabric manages the Grace Hopper superchips and BlueField-3 DPU through their Baseboard Management Controller (BMC), and NVLink switches through ComEx and connects all networking equipment to the control interfaces. It provides critical out-of-band management, such as telemetry and firmware upgrades, to prevent conflicts between management traffic and other services.

Storage Requirements

The DGX GH200 system should be paired with a high-performance and balanced storage system to maximize overall system performance. Each GH200 system can read or write data at up to 25 GB/s across the NDR200 interface. For a 256 Grace Hopper DGX GH200, NVIDIA recommends an aggregate storage performance of 450 GB/s for overall read-throughput. Depending on application needs, the storage solution can be expanded to support multiple TB/s of throughput.

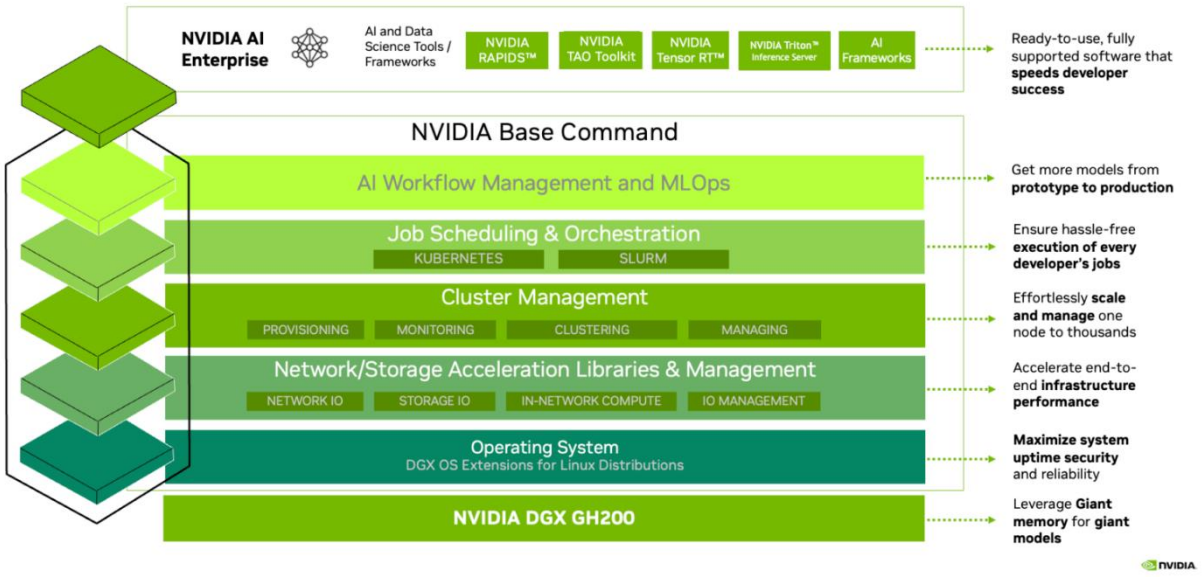
DGX GH200 Software

The NVIDIA DGX GH200 supercomputer has a comprehensive software suite because a robust AI infrastructure requires more than great hardware. The DGX GH200 supports the entire CUDA platform ecosystem, including tools, libraries, and frameworks, which run on Arm CPUs “out of the box.” NVIDIA optimized the ecosystem for the DGX GH200. This section will examine the DGX software stack and highlight enhancements to help customers get the most out of their DGX GH200.

NVIDIA Base Command is the operating system of a DGX data center, helping organizations speed the return on investment of AI. With the help of Base Command, Enterprises can tap into the full potential of their DGX infrastructure with a proven platform built and hardened by NVIDIA’s practitioners. NVIDIA Base Command for DGX GH200 includes enhancements to OS, enterprise-grade cluster management, libraries that accelerate compute, storage and network infrastructure, and system software optimized for running AI workloads.

DGX™ OS provides a customized installation of Ubuntu Linux with system-specific optimizations and configurations, additional drivers, and diagnostic and monitoring tools. It provides a stable, fully tested, and supported OS to run AI, machine learning, and analytics applications on DGX Supercomputers. Grace Hopper Superchips in DGX GH200 are shipped preinstalled with DGX OS. Figure 9 shows the NVIDIA Base Command for DGX GH200.

Figure 9. NVIDIA Base Command for DGX GH200



The NVIDIA Base Command Manager is enhanced to support DGX GH200 system management. We have added support for parallel and automated provisioning, monitoring, and managing Grace Hopper Superchips. The base command manager also supports the firmware lifecycle management of the DGX GH200 system. Base Command Manager also simplifies the deployment and management of multi-GPU scheduling, including the administration of users and jobs with a topology-aware scheduler.

Using Base Command Manager, data center administrators can manage instances of DGX GH200 alongside other systems (DGX and non-DGX) from the same management control plane, allowing users to place AI workflows on the appropriate hardware platform.

DGX GH200 continues to use the best-in-class capabilities of NVIDIA Unified Fabric Manager to provision, monitor, and manage the InfiniBand fabric. We also integrated NVIDIA Base Command Manager with NetQ for a unified monitoring and health-checking view of the NVLink switch system.

NVIDIA Unified Fabric Manager platforms combine enhanced, real-time network telemetry with AI-powered cyber intelligence and analytics to support scale-out InfiniBand data centers. The UFM platforms empower IT managers to discover operation anomalies and predict network failures for preventive maintenance. The UFM platform consists of three tiers, a basic UFM Telemetry platform that provides network validation and network performance monitoring, a mid-tier UFM Enterprise platform that adds enhanced monitoring, management and workload optimization, and the top-tier UFM Cyber AI platform that uses deep learning algorithms to monitor, predict, and correct performance degradation and system failures.

NVIDIA NetQ™ is a modern network operations tool that provides actionable visibility for the NVIDIA NVLink Switch System, Spectrum™ Ethernet switches, and NVIDIA DPUs. Using NetQ GUI/CLI/APIs, a user can manage NVLink logical grouping (domains) and devices, and data center admins can perform NVLink device upgrades and monitor events. A Domain is a group of Grace Hopper Superchips with NVLink switch routing boundary. We will cover the details of domain management in a subsequent section. NetQ uses fabric-wide telemetry data to provide real-time visibility and troubleshoot the overlay and underlay network. NVIDIA NetQ offers hosting support for the Global Fabric Manager (GFM), allowing single-click NVLink domain control and parallelized management of all NVLink switches in the NVLink Switch System.

CUDA toolkit provides virtual memory management APIs that enable applications to import and export inter-process memory handles from other processes over the NVLink Switch System, extending the CUDA programming model from eight GPUs in DGX H100 to 256 GPUs and Extended GPU Memory in DGX GH200. CUDA extends the sharing mechanisms for virtual memory management APIs and the stream-ordered memory management APIs to enable memory sharing across Grace Hopper Superchips.

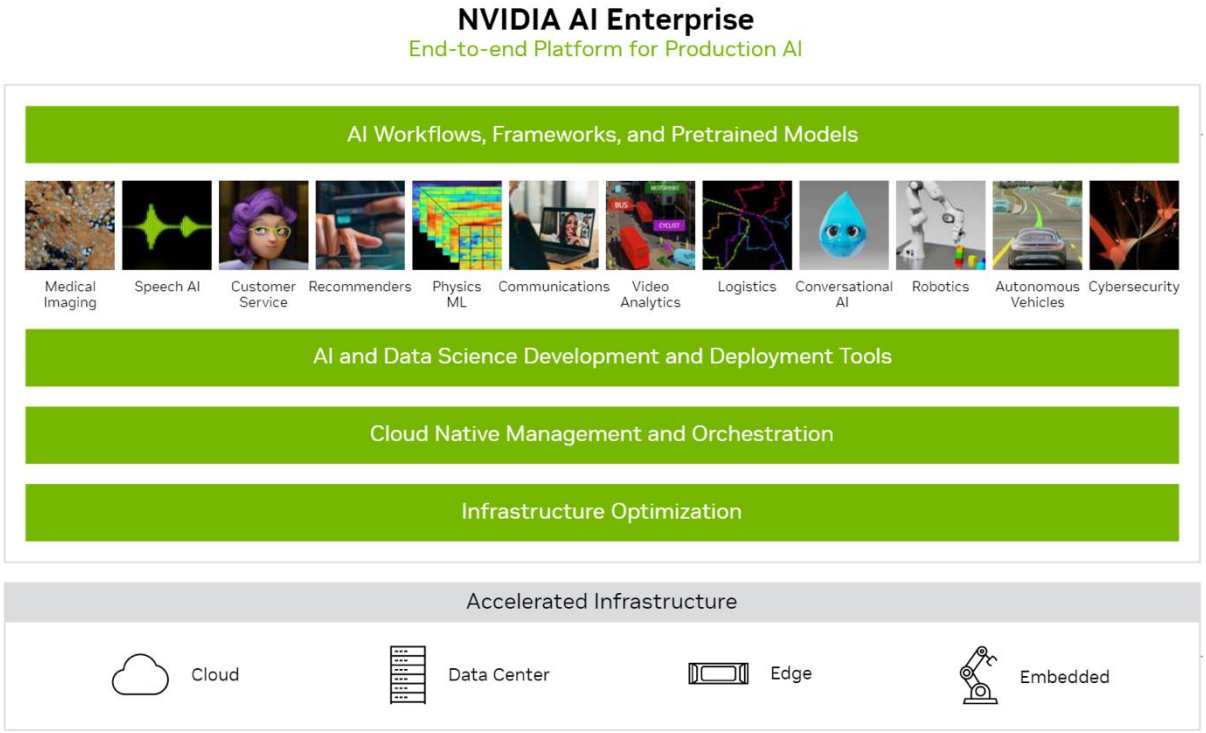
The CUDA programming model exposes peer memory as local memory: loads, stores, and atomic operations transparently work over the NVLink Switch System, enabling applications to use it with libraries. Multi-GPU programming models like NCCL and NVSHMEM abstract CUDA GPU peer memory and InfiniBand from applications and frameworks, optimally leveraging NVLink Switch System islands and InfiniBand.

Magnum IO Libraries speed up application software stack CUDA-XTM libraries (cuFFT, cuQuantum, and others). CUDA applications are optimized for the NVLink Switch System to harness the benefits of the fast, large memory in DGX GH200 and now work on the NVIDIA Grace CPUs using Arm instruction set architecture. The NVIDIA Collective Communication Library (NCCL) optimizes collective and point-to-point communication APIs using direct NVLink connectivity and NVLink SHARP among GPUs within an NVLink domain. It combines NVLink and network (such as InfiniBand) to allow the same APIs to work transparently across multiple NVLink domains. NCCL enables prominent Deep Learning Frameworks like PyTorch and TensorFlow to run on DGX GH200. NVSHMEM provides a partitioned global memory abstraction and one-sided communication APIs that work within and across multiple NVLink domains.

It allows applications to directly access distributed global memory through loads, stores, and atomics within an NVLink domain. NVSHMEM enables essential CUDA-X libraries like cuFFT. UCX is supported within and across NVLink domains to accelerate HPC applications written with MPI, Data Analytics workloads that use NVIDIA RAPIDS, and Python applications that use NVIDIA cuNumeric.

NVIDIA AI Enterprise is an end-to-end, secure, cloud-native suite of AI software that enables organizations to solve new challenges while increasing operational efficiency. It accelerates the data science pipeline and streamlines the development and deployment of predictive AI models to automate essential processes and gain rapid insights from data. Key frameworks (TensorFlow, PyTorch Nemo-Megatron, Merlin-CTR, and others) are optimized for GH200 to utilize NVLink-C2C and fast large memory across 256 Grace Hopper Superchips. With an extensive library of full-stack software, including AI solution workflows, frameworks, pre-trained models, and infrastructure optimization, the possibilities are endless. NVIDIA AI Enterprise lets organizations develop once and run anywhere, available in the cloud, in the data center, and at the edge. Global enterprise support and regular security reviews ensure business continuity and AI projects stay on track (see Figure 10).

Figure 10. NVIDIA AI Enterprise Software Suite for AI Development



NVLink Partitioning

Partitioning gives a system administrator the ability to isolate workloads from one another. Partitioning can be utilized to separate multiple user workloads and to contain the impact of system or hardware failures on applications.

The NVLink Switch System, which connects all the Grace Hopper superchips in the DGX GH200, can be set up in two ways:

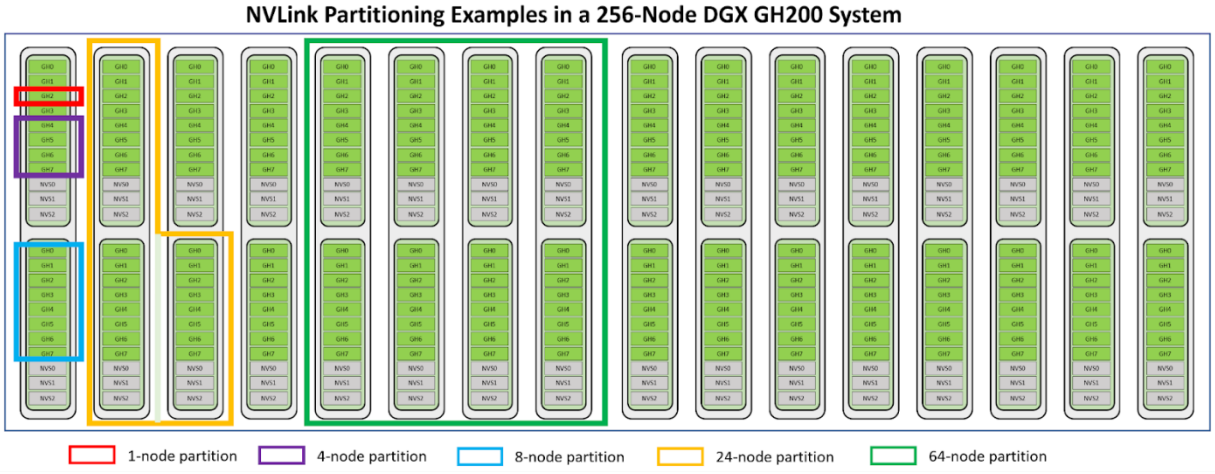
- All 256 Grace Hopper superchips can access each other's memory in non-partitioned mode.
- In partitioned mode, memory access is restricted at partition boundaries.

A partition can consist of one, two, or eight Grace Hopper superchips from within an eight Grace Hopper complex inside a chassis when partitions are created at the level 1 NVLink switch boundary or in groups of eight Grace Hopper superchips from one or more chassis when partitions are created at level 2 NVLink switches. The Global Fabric Manager (GFM) sets up the NVLink Switch routing, creates partitions, and isolates NVLink traffic and faults within a partition. GFM is hosted by NetQ, which provides a centralized health monitoring and partitioning API.

Example 1: Within a chassis, one could have eight 1-Grace Hopper superchip partitions, four 2-Grace Hopper superchip partitions, two 4-Grace Hopper superchip partitions, or a single 8-Grace Hopper superchip partition.

Example 2: If a partition size is larger than an 8-Grace Hopper chassis, then they can be sized in multiples of 8 Grace Hopper superchips (8, 24, 32 Grace Hopper superchips, and so on). See Figure 11.

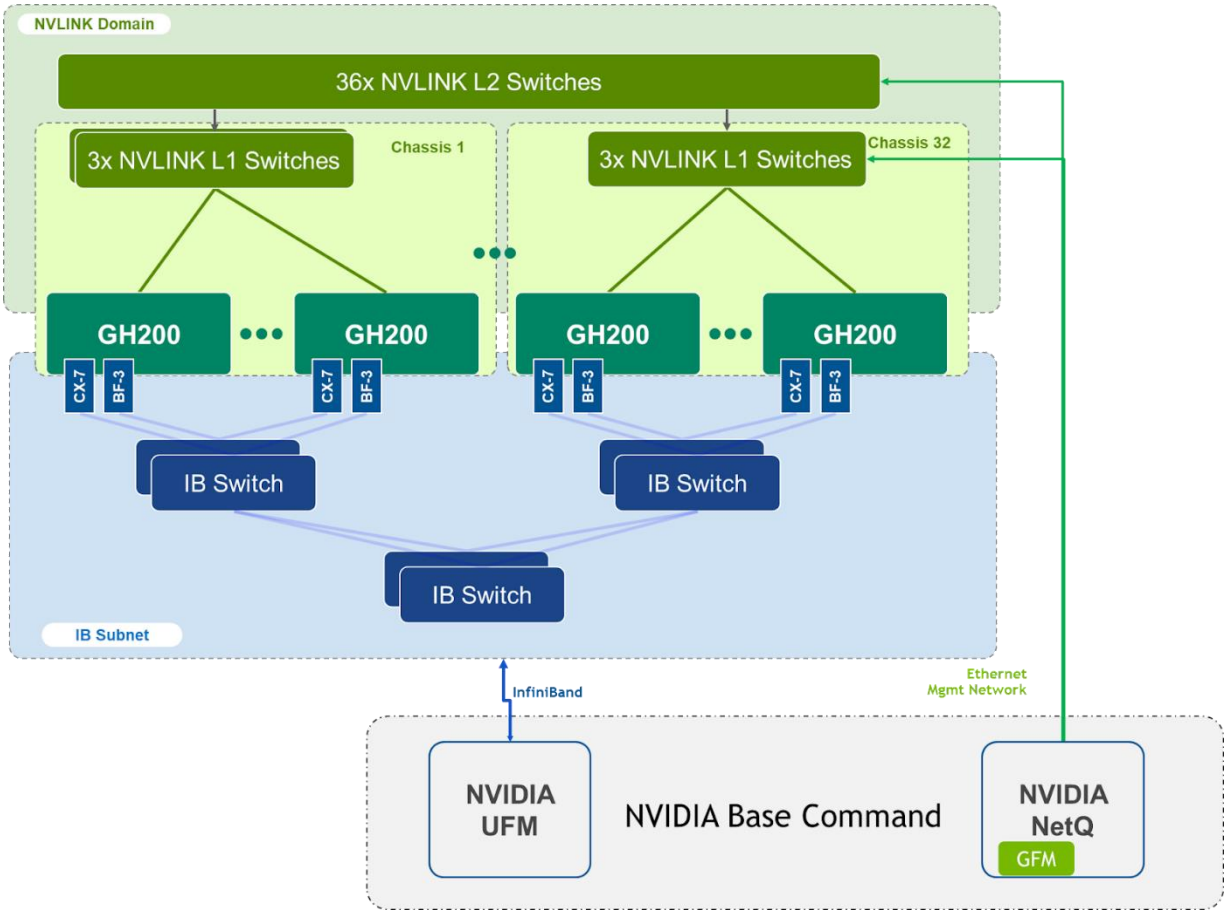
Figure 11. NVLink Partitioning Examples in a DGX GH200



Fabric Management

An instance of Local Fabric Manager (LFM) service runs on each L1 and L2 NVLink switch in the DGX GH200. LFM is the management interface agent for the NVLink switches. A Global Fabric Manager (GFM), which resides on the NetQ management server, is the primary control and management software layer that connects to all the LFMs over TCP/IP and provides APIs for NetQ to create and manage network partitions (see Figure 12).

Figure 12. Fabric Management Topology for DGX GH200



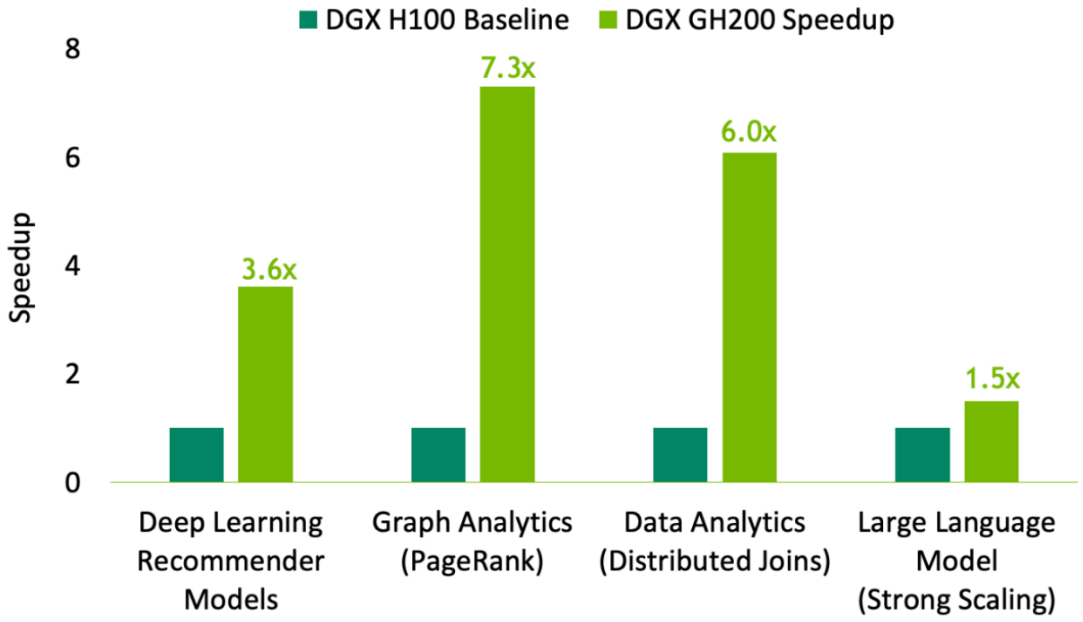
Initialization and Workload Deployment

Initializing the 256 Grace Hopper Superchips in the DGX GH200 supercomputer and deploying AI workloads across the Grace Hopper superchips is made easy by a comprehensive suite of cluster management tools. These include NVIDIA Base Command, which works seamlessly with partner software and scheduling agents like Slurm and Kubernetes.

DGX GH200 Performance

The NVIDIA DGX GH200 delivers 3x to 7x performance speedup compared to the DGX H100 on various giant AI models such as Recommender systems, graph analytics, and data analytics (see Figure 13).

Figure 13. NVIDIA DGX GH200 Delivers up to 6x Speedup Over DGX H100.



Many of today's mainstream AI models can entirely reside in the aggregate GPU memory of a single DGX H100. For such workloads, DGX H100 is the most performance-efficient solution. For giant workloads bottlenecked by GPU memory or networking, DGX GH200 is a more scalable solution. The following paragraphs discuss a few examples of network and memory-limited workloads.

Network Intense Workloads

Large Language Models (LLM) Training at Scale

Large language models (LLMs) continue to grow in size and complexity. To use a larger model, enterprises need to train on more data. Due to the quadratic relationship between training times and model size, large-scale multi-GPU training becomes essential to reduce training times. However, batch sizes usually do not increase with model sizes. Strong scaling with constant or lower batch sizes requires more aggressive use of tensor parallelism beyond the 8-GPU size of the DGX node. It also increases the amount of time that we spend in data-parallel communication (weight gradient reductions). These two factors impact scale-out efficiency and become dominant at large scales. The DGX GH200 addresses these two bottlenecks by extending the NVLink connectivity to up to 256 GPUs, compared to eight GPUs in the DGX H100 architecture.

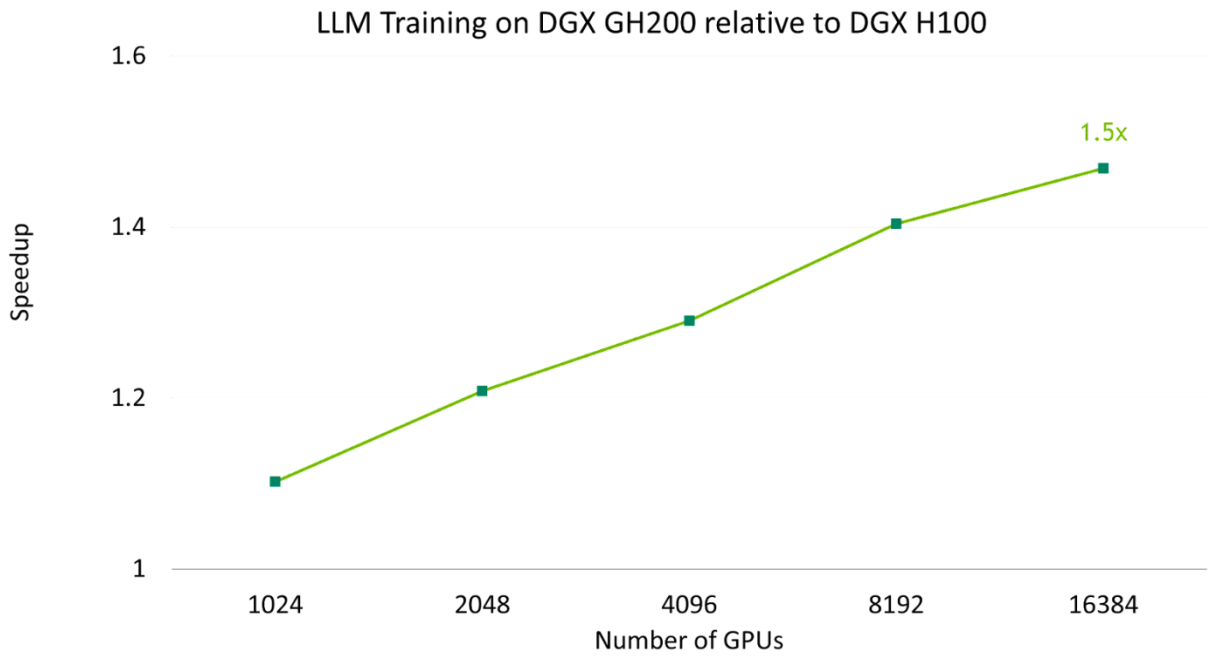
The larger NVLink domain helps speed training by allowing wider Tensor Parallelism and faster data parallel reduction.

Tensor Parallelism (TP) works by distributing Tensor to multiple GPUs. TP requires extremely high-bandwidth communication between GPUs. On DGX H100 architecture, TP is optimal with up to eight GPUs due to its maximum NVLink domain size. On DGX GH200, TP can go much wider with the extended NVLink domain without losing communication performance, enabling larger models with larger tensors to be processed most efficiently.

A larger NVLink domain enables faster data parallel reduction within the high-speed NVLink domain first, minimizing the amount of data that needs to go through slower networking like InfiniBand or Ethernet and speeding up the reduction time.

DGX GH200 enables more efficient parallel mapping and alleviates the networking communication bottleneck. As a result, up to 1.5x faster training time can be achieved over a DGX H100-based solution for LLM training at scale (see Figure 14).

Figure 14. Strong Scaling Benefits of the DGX GH200 Clusters vs DGX H100 Clusters for LLM Training with 175B Parameters.



Total GPU Memory Limited Workloads

The DGX GH200's large GPU memory and up to 115.2 TB/s bisection NVLink bandwidth eliminate bottlenecks for solving giant AI and data analytics problems, delivering unprecedented performance. Compared to a 256 GPU NVIDIA DGX H100 cluster interconnected with NDR InfiniBand, the 256 GPU DGX GH200 system offers significant performance improvements for these emerging workloads. The giant memory of up to 144 TB, the tremendous increase in I/O bandwidth for CPU to GPU communication as well as GPU to GPU communication, and the tremendous computing power of the Hopper GPU cores join forces to deliver 4x to 6x speedup over DGX H100 clusters connected over InfiniBand.

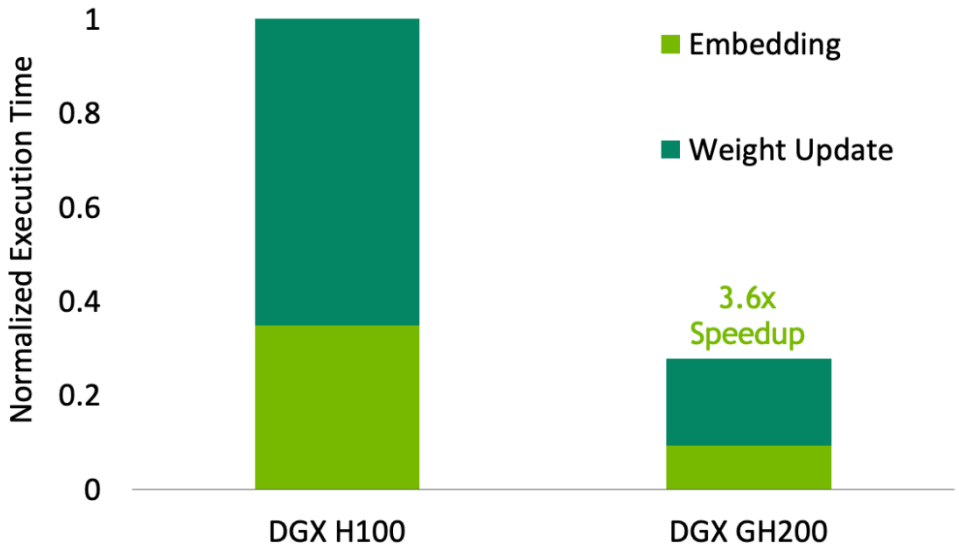
A few key AI models are discussed in the following sections.

Large Recommender Systems

For better recommendations, modern recommender systems need a lot of memory to store embedding tables containing semantic representations of items and user features. These embeddings follow a power-law distribution in the frequency of use, with some embeddings accessed more often than others.

NVIDIA Grace Hopper supports high-throughput recommender system pipelines by storing frequently used embedding vectors in HBM3 memory and less frequently used vectors in higher-capacity LPDDR5X memory. The NVLink-C2C gives Hopper GPUs high-bandwidth access to their local LPDDR5X memory, while the NVLink Switch System extends fast access to all LPDDR5X and HBM memory of all Grace Hopper Superchips in the NVLink network (see Figure 15).

Figure 15. Deep Learning Recommender System Training with Batch Size of 65k, 36 TB Embedding Tables, on 16x DGX H100 (128 GPUs) with 1x NDR400 NIC per GPU and on 16x DGX GH200 (128 Grace Hopper) with NVLink Switch System



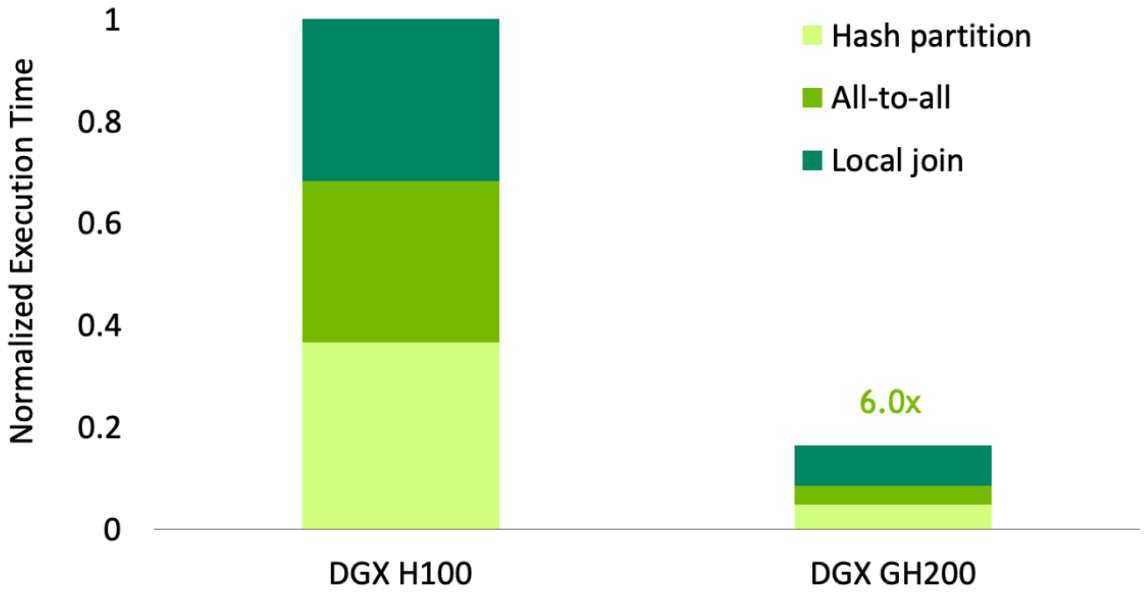
For example, when training the 36 TB large recommender model (Figure 16) running on the DGX H100, most network communication is hidden behind the computations accessing the embedding tables. However, when the embedding computation is accelerated with NVLink C2C, the communication must scale to prevent it from becoming the bottleneck. On the DGX GH200, the higher bandwidth NVLink across all 128 GPUs accelerates this communication, delivering a 3.6x speedup over DGX H100.

Distributed Hash Joins in Databases

The “join” operation is fundamental in relational databases and the ETL stage of machine learning to analyze large amounts of data from sources like social media, e-commerce websites, and IoT devices. Distributed hash-based join algorithms can scale to many GPUs but require a high-throughput, low-latency network like the NVLink Switch System. The algorithm has three steps: reordering input tables into N partitions based on hash values, performing an all-to-all exchange where each GPU sends its partitions to the corresponding GPU (partition 0 to GPU0, partition 1 to GPU1, and so on), and performing the local join independently on each GPU.

For a large synthetic dataset consisting of 25 TB of randomly distributed uniform integers in the CPU memory, the combination of NVLink Switch System, Magnum IO, and NVLink-C2C in a DGX GH200 (256 GPUs) results in a speedup of 6.0x over 32x DGX H100 (256 GPUs) interconnected by NVIDIA Quantum-2 switch and NDR400 (see Figure 16).

Figure 16. Distributed Hash Join of 25 TB Large Input Table using 32x DGX H100 (256 GPUs) with NDR400 and 32x DGX GH200 (256 GPUs) with NVLink Switch System



Graph Analytics (Page Rank)

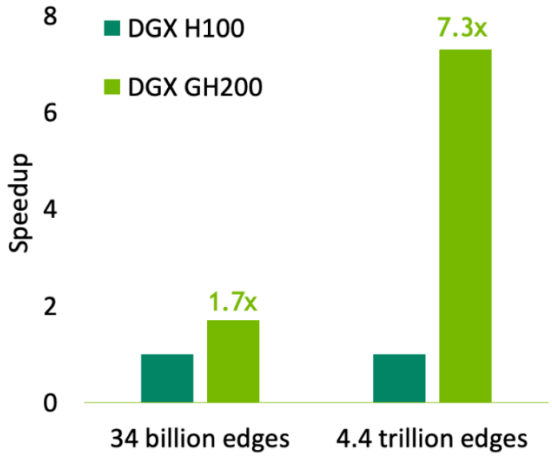
Graphs capture real-world interactions such as social networks and citation networks. Graph algorithms extract valuable information from graphs, and PageRank is one of the most representative graph algorithms. Storing a graph with hundreds of billions of vertices and trillions of edges requires over 100 TB of memory. Graph algorithms often require random accesses over a large array storing vertex properties. Graph algorithms running on multiple compute nodes often are bottlenecked by internode communication bandwidth.

DGX GH200 provides large memory capacity HBM and LPDDR5X memory; LPDDR5X memory can be accessed from GPU with fast NVLink-C2C and NVLink Switch System with larger system bisection bandwidth. DGX GH200 is a powerful platform to process large graph workloads with intensive communication.

The PageRank algorithm on a graph with 34 billion vertices (8-byte vertex ID) and 550 billion weighted edges (4-byte edge weight) requires more than 15 TB of memory or about 32 DGX H100 systems with 256 GPUs. GPU local computing takes 74% of the execution time, and communication takes the remaining 26%. Processing the 550 billion edge graph on DGX GH200 with 256 GPUs will be moderately faster (1.7x) than DGX H100 owing to higher HBM bandwidth of 4 TB/s and up to 9x faster communication speed due to NVLink Switch System.

However, analyzing an 8x larger graph with 4.4 trillion edges and 275 billion vertices requires more than 120 TB of memory (see Figure 17). A performance simulation of a DGX GH200 system with 256 GPUs reveals that the 480 GB of LPDDR5X capacity and the high bandwidth of the NVLink-C2C interconnect results in a 7.3x speedup over a 32 DGX H100 system with 256 GPUs, PCIe Gen5, and 512 GB of system memory per GPU.

Figure 17. PageRank Performance Simulation on Graphs with 34 Billion and 4.4 Trillion Edges of DGX H100 and DGX GH200 Systems with 256 GPUs.



NVIDIA GH200 Deployment Options

DGX GH200 is the culmination of NVIDIA's unmatched experience in designing and using AI supercomputers, driven by thousands of NVIDIA researchers and engineers who use this platform to bring innovations to market. NVIDIA DGX GH200 delivers the turnkey data center solution that can be installed on-premises or at a colocation provider's DGX-Ready Data Center for businesses focused on innovation instead of infrastructure. Stay tuned for more information as we work on making DGX GH200 available through NVIDIA DGX Cloud.

DGX GH200 can be scaled according to the number of Grace Hopper superchips in the system. Following are the supported configurations.

- 32 Grace Hopper superchips
- 64 Grace Hopper superchips
- 128 Grace Hopper superchips
- 256 Grace Hopper superchips

DGX GH200 Detailed Specifications

Table 4 lists the DGX GH200 Technical Specifications.

Table 4. DGX GH200 Technical Specification

CPU and GPU	256 NVIDIA Grace Hopper Superchips
CPU Cores	18,432 Arm' Neoverse V2 Cores with SVE2 128b
GPU Memory	144 TB
Performance	1 exaFLOPS FP8 with sparsity
Networking	> 256x OSFP single-port NVIDIA ConnectX-7 VPI with 400Gb/s InfiniBand > 256x dual-port NVIDIA BlueField-3 or 256 dual-port NVIDIA ConnectX-7 VPI with 200Gb p/s InfiniBand and Ethernet > 24x NVIDIA Quantum-2 QM9700 Infiniband Switches > 20x NVIDIA Spectrum SN2201 Ethernet Switches > 22x NVIDIA Spectrum SN3700 Ethernet Switches
NVIDIA NVLink Switch System	> 96x L1 NVIDIA NVLink Switches > 36x L2 NVIDIA NVLink Switches > 2304 LinkX Optical Cables
Management Network	> Host baseboard management controller (BMC) with RJ45 for Grace Hopper superchip baseboards > ComEx Controller RJ45 for NVLink switches
Software	> NVIDIA AI Enterprise (optimized AI software) > NVIDIA Base Command (orchestration, scheduling, and cluster management) > DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky (operating system
Support	Three-year business-standard hardware and software support

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. 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 document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information 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 (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform 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 document. NVIDIA accepts no 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 document 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 document. 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 document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND 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. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. 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 Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA, NVLink, NVPascal, NVIDIA Turing, DGX, NVIDIA Grace Hopper, NVIDIA GRACE, NVIDIA HOPPER, NVSwitch, NVIDIA LinkX, NVIDIA Magnum I/O, NVIDIA NGC, NVIDIA NetQ, NVIDIA Nsight, NVIDIA Volta, NVIDIA Jetson AGX Xavier, NVIDIA DGX POD, CUDA, CUDA-X, NVIDIA UFM, NVIDIA Base Command, NVIDIA ConnectX-7, NVIDIA BlueField-3, and NVIDIA OVX 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.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

Arm

Arm, AMBA, and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore, and Mali are trademarks of Arm Limited. All other brands or product names are the property of their respective holders. ”Arm” is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS, and Arm Sweden AB.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com