IBM Storage Scale and IBM Storage Scale System 6000 for NVIDIA DGX SuperPOD

NVIDIA DGX GB200 and GB300 High-Performance Storage Reference Design

February 2026

NVIDIA DGX SuperPOD with IBM Storage Scale and IBM Storage Scale System 6000

NVIDIA DGX GB200 and GB300 Storage Reference Architecture

Abstract

The NVIDIA DGX SuperPOD™ with NVIDIA DGX™ GB200 and NVIDIA DGX™ GB300 systems is an artificial intelligence (AI) supercomputing infrastructure, which provides the computational power necessary to train today's state-of-the-art deep learning (DL) models and to fuel future innovation. The DGX SuperPOD delivers groundbreaking performance, deploys as a fully integrated system, and is designed to solve the world's most challenging computational problems.

This DGX SuperPOD storage reference architecture (RA) is the result of collaboration between DL scientists, application performance engineers, and system architects to build a system capable of supporting the widest range of DL workloads. The performance delivered by the DGX SuperPOD with DGX systems enables the rapid training of DL models at great scale. The integrated approach of provisioning, management, compute, networking, and fast storage enables a diverse, multi-tenant system that can span data analytics, model development, and AI inference.

In this paper, we will describe how to use the IBM Storage Scale System 6000 to supporting DL workloads when connected to the DGX SuperPOD. IBM Storage Scale System 6000 is a storage appliance offering low latency NVMe physical storage, advanced erasure coding, and supports either NVIDIA Quantum InfiniBand or Ethernet networking. Multiple IBM Storage Scale System 6000's can be aggregated to create a high-performance cluster filesystem, or connected to multiple clusters for geographic and cross platform data sharing in a single global data platform. The Scale System 6000 is a 4U building block that makes it easy to deploy, manage, and grow fast storage for AI with NVIDIA DGX systems.

The DGX SuperPOD is a turnkey solution validated at scale with scale-out Scale System 6000s. Joint testing and integration ensure the NVIDIA DGX SuperPOD is a rapidly deployed and a robust solution for scalable AI development. NVIDIA and IBM jointly test, plan, and install the system, with the storage backed by IBM global deployment and support services.

As configured, tested, and deployed in the NVIDIA SuperPOD, the IBM Storage Scale System 6000 can be used for all DL workloads including:

- Training models efficiently with directly from IBM Storage Scale.
- Automatically leverage local resources as cache to minimizing rereading data across

the network.
- Workspace for long-term storage (LTS) of datasets.
- A centralized repository for the acquisition, manipulation and sharing of results using standard protocols like NFS, SMB, and S3.

Table of Contents

Storage Overview

Training performance can be limited by the rate at which data can be read and reread from storage. The key to performance is the ability to read data multiple times, ideally from local storage. The closer the data is cached to the GPU, the faster it can be read. Storage must be designed considering the hierarchy of different storage technologies, either persistent or nonpersistent, to balance the needs of performance, capacity, and cost.

Storage Caching Hierarchy

The storage caching hierarchy of the DGX system is shown in Table 1. Depending on data size and performance needs, each tier of the hierarchy can be leveraged to maximize application performance.

Table 1. DGX system storage and caching hierarchy

Storage Hierarchy Level	Technology	Total Capacity[1]	Performance[1]
RAM	DDR4	2 TB per system[2]	> 200 GB/s
Internal Storage	NVMe	30 TB per system[3]	> 50 GB/s

1. Total capacity and performance values are per system.
2. Shared between the operating system, application, and other system processes
3. PCIe NVMe SSD storage

Caching data in local RAM provides the best performance for reads. This caching is transparent after the data is read from the filesystem.

While local storage is fast, it is not practical to manage a dynamic environment with local disk alone in multi-node environment. Functionally, centralized storage can be as quick as local storage on many workloads.

Storage Performance Requirements

Performance requirements for high-speed storage greatly depend on the types of AI models and data formats to be used. The DGX SuperPOD has been designed as a capability-class system that can manage any workload both today and in the future. However, if systems are going to focus on a specific workload, such as natural language processing (NLP), it may be possible to better estimate performance needs of the storage system.

To enable customers to characterize their own performance requirements, some general guidance on common workloads and datasets is shown Table 2.

Table 2. Characterizing different I/O workloads

Level	Work Description	Dataset 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.

Performance estimates for the storage system necessary to meet the guidelines in Table 2 are in:

- Table 5 of the *NVIDIA DGX SuperPOD Reference Architecture—DGX GB200 Systems*.
- Table 4 of the *NVIDIA DGX SuperPOD Reference Architecture—DGX GB300 Systems*.

Achieving these performance characteristics may require the use of optimized file formats such as TFRecord, RecordIO, or HDF5.

The IBM Storage Scale System 6000 can be used as shared storage to provide a single view of an organization's entire data set. It provides high peak system performance and high aggregate performance and meets the 'best' category requirements for the H100, B200, and B300 architectures.

IBM Storage Scale System 6000 Overview

The IBM Storage Scale System 6000 (Figure 1) combines the performance of NVMe storage technologies with the reliability and the rich features of IBM Storage Scale, along with several high-speed attachment options such as 400 Gb/s Ethernet and NVIDIA Quantum InfiniBand (IB) — all in a powerful 4U storage system that scales out for performance and capacity.

Figure 1: IBM Storage Scale System 6000

IBM Storage Scale System on NVMe is designed to be the market leader in all-flash performance, and scalability with a bandwidth of 330 GB/s per NVMe all-flash appliance with low latency. Providing data-driven multicloud storage capacity, the NVMe all-flash appliance is deeply integrated with the software defined capabilities of IBM Storage Scale to seamlessly plug it into an analytics, scalable cluster, or AI workload.

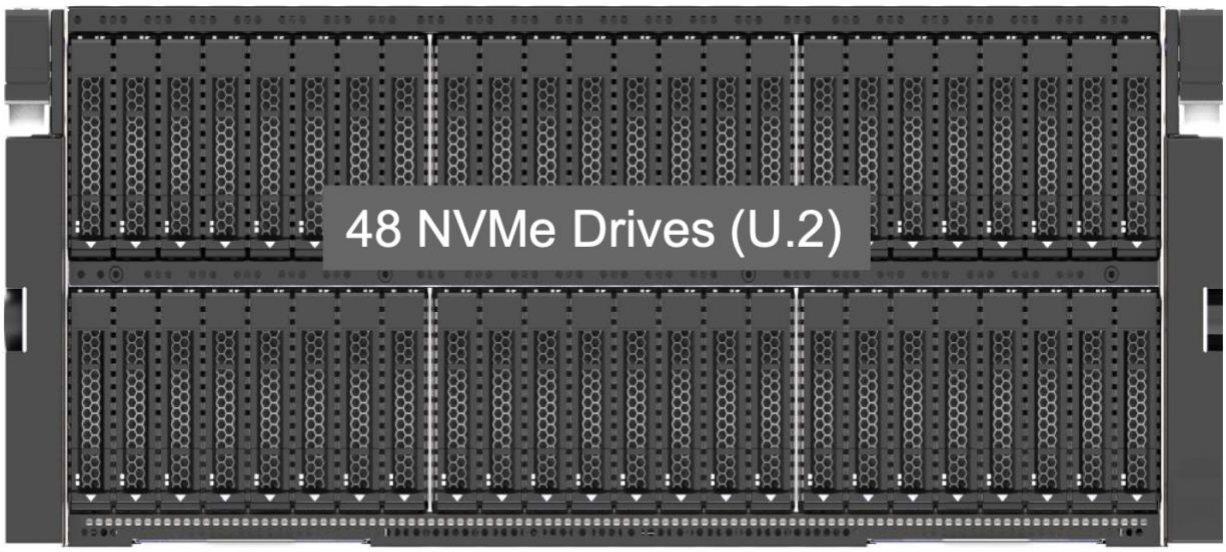
Available with multiple drive options and advanced erasure coding, the Scale System 6000 provides options to optimize costs for different installation sizes. As with all IBM Storage Scale solutions, capacity and performance can be scaled. Combining Scale System 6000 systems provides nearly linear performance scalability. Scale System 6000 solutions may also be used as an all-flash NVMe performance tier combined with slower, more cost effective storage, including tape or object storage.

IBM Storage Scale is an industry leader in high-performance file systems. The underlying general parallel file system (GPFS) provides scalable throughput and low-latency data access, as well as superior metadata performance. Unlike other systems that can easily bottleneck, the distributed architecture of a parallel filesystem provides reliable performance for multi-user sequential and random read or write. This is particularly important in AI clusters where multiple compute nodes may need to read or write to the same file. IBM Storage Scale provides Container Native Access and Operators to support Kubernetes driven DevOps and Data Ops practices. In addition, IBM Storage Scale provides enterprise features such as call-home proactive support, encryption, and audit file logging that works with enterprise security information and event management (SEIM) platforms. IBM Storage Scale Systems integrate well with NVIDIA Base Command Manager to streamline administration and configuration of the solution.

Planning

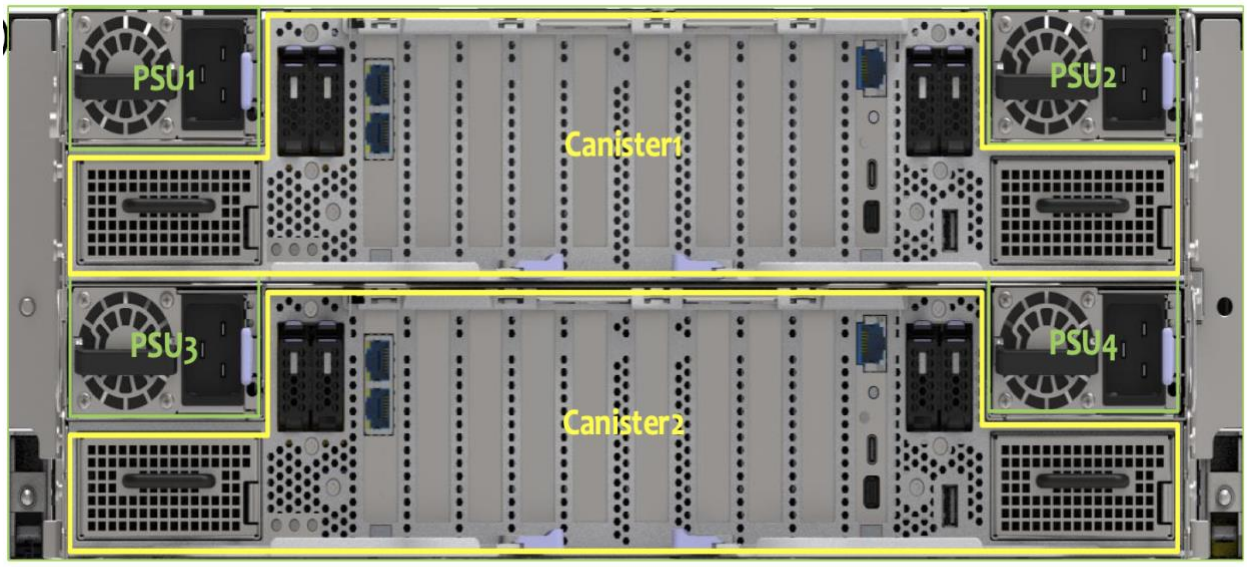
Hardware Overview

The IBM Scale System 6000 is a fully redundant, centrally managed solution. The system is deployed in building blocks, with up to 48 U.2 form factor NVMe drives per building block.



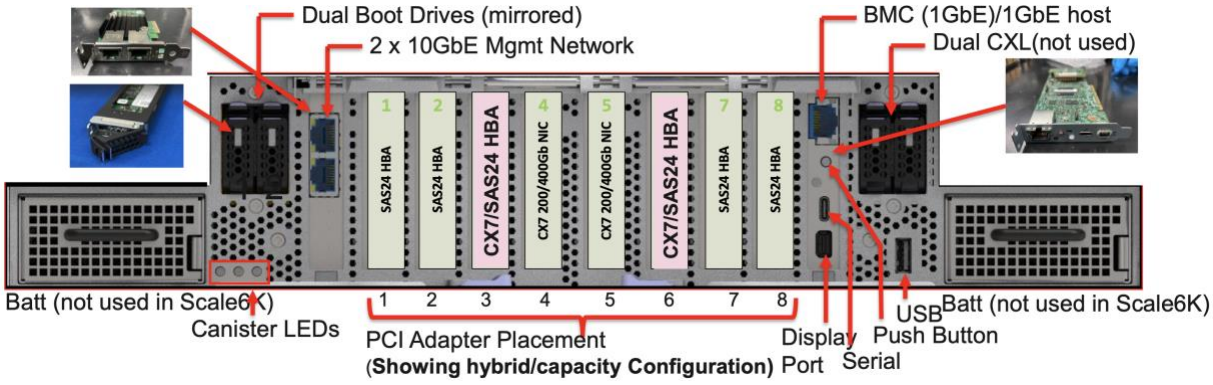
Front View

The rear of the system contains 4 power supplies and 2 redundant compute canisters



Rear View

The Scale System 6000 is a flexible solution that can be configured for either high capacity, using external SAS drive enclosures, or for performance. To meet the performance requirements for the NVIDIA SuperPOD, it is recommended to use the 'performance' configuration, populating 4 PCI gen5 slots per canister (slots 3-6 in the figure below) with NVIDIA ConnectX-7 adapters. The Scale System 6000 should also be configured with NVMe drives as outlined in the sizing section below.



The Scale System 6000 can be configured as a part of a solution to meet performance, capacity, and cost requirements for any need.

Multi-Protocol Access

An IBM Scale System 6000 utilizes a high-speed, proprietary protocol to provide access to data. This protocol provides high speed, consistent, redundant access to data. To access data, a client requires special software to provide access to the data. NVIDIA DGX systems will have the Storage Scale client installed for high speed access.

For user home directories and some cluster specific data, the NVIDIA DGX SuperPOD architecture recommends NFS for data access on the management servers. To meet this need, a minimum of 2 storage scale protocol nodes are required. Each protocol node provides access to the file system via NFS, and optionally supports SMB, HDFS, and Object storage. Two nodes provide redundancy of access, and up to 32 protocol nodes can be added to the system. Additional protocol nodes can provide higher speeds, or can support clients external to the SuperPOD for easier data access and movement.

Each protocol node can run either on a Storage Scale Utility node or on standard x86 or IBM Power system running a standard OS such as RHEL or Ubuntu. See the Storage Scale FAQ for the latest OS's and releases supported.

Each protocol node is connected to the Storage Fabric with 2 200 Gb NDR adapters. These adapters can be connected via a splitter cable to the 400 Gb Storage Fabric for greater port density and to reduce the number of InfiniBand ports required for the solution. In addition, protocol nodes will connect to the

Management Network for NFS access and can optionally connect to additional networks to support clients outside of the DGX SuperPOD.

Data Tiering and Caching

IBM Storage Scale offers both the ability to tier data within the file system, or to cache data from external systems.

The Integrated Lifecycle Management (ILM) functionality of Storage Scale moves data seamlessly between various storage mediums such as NVMe, hard drives, and tape drives. By placing data on the appropriate storage type, Storage Scale allows for high-speed access to data while offering cost-effective capacity expansion.

The Active File Management (AFM) function caches storage from external sources such as Object, NFS, or other Storage Scale file systems. By caching storage on local storage, users are given high-speed local access to data even if the source copy resides on external storage.

Integrated Lifecycle Management (ILM)

The IBM Storage Scale ILM functionality combines multiple storage tiers, or pools, such as NVMe, disk, or tape, into a single namespace. Data can be moved between the storage tiers seamlessly to an end user at any time. In addition, a robust policy syntax allows for automatic movement of data in certain conditions – for example once the hard disk pool reaches a certain capacity, automatically migrate the least recently used data to tape.

IBM Storage Scale 6000 systems can be configured with optional SAS adapters and external spinning disk enclosures in order to extend capacity. Storage Archive Enterprise Edition (https://www.ibm.com/docs/en/storage-archive-ee/1.3.4?topic=overview-introduction-storage-archive-enterprise-edition) can be used to connect to external tape enclosures using additional nodes.

Active File Management (AFM)

IBM Storage Scale AFM seamlessly caches data from external data stores. These data stores can be remote Storage Scale, NFS, or Object stores.

To provide connectivity, a Storage Scale cluster requires a minimum of 1 AFM gateway node, however at least 2 are needed for redundancy. The hardware for the gateway nodes can be a Storage Scale Utility node, or standard x86/Power hardware. Please see the IBM Storage Scale FAQ for the latest guidelines.

AFM gateway nodes require 2 connections to the SuperPOD data network, with NVIDIA ConnectX-7 400 Gb IB connectivity being recommended. In addition, each gateway node requires at least one connection to a high speed network that can connect to the storage to cache. The external network may vary depending on the speed and latency to the external storage source.

When sizing the number of gateway nodes and network connectivity, 3 factors should be considered:

1. The size of the active data set being cached
2. The number of files/objects being cached
3. The change rate of the data that is cached

Depending on these factors, additional gateway nodes may be needed in to meet the given workload. The IBM Storage Scale FAQ and the planning section of the knowledge center (https://www.ibm.com/docs/en/storage-scale/5.2.1?topic=planning-afm) provide additional guidance on gateway node configuration.

Scale System Management Server

The IBM Scale System 6000 requires a management server (EMS) to provision and manage the storage system. Each EMS can manage multiple Scale System 6000 systems in a single cluster. Typically, the management server is deployed on a dedicated 2U Scale System utility node, with 512 GB of RAM and 1 ConnectX-7 dual-port InfiniBand or Ethernet adapter. The management server requires one connection from the ConnectX-7 to the SuperPOD storage fabric.

Front View of the Scale System Utility Node

In addition, a dedicated management switch from the management server is required, to provision and manage the system. This switch may be customer-owned; however an IBM switch is recommended as it is pre-configured with the appropriate configuration to manage the Scale System 6000.

For all configurations in the sizing section, a single management server and switch meeting these requirements will be necessary and is included in the rack units described.

Networking

The IBM Scale System 6000 uses multiple networks described below.

Storage Fabric

The Storage Fabric connects the Scale System 6000 systems to the DGX Superpod. In an NVIDIA SuperPOD solution, it is recommended to use 4 400 Gb InfiniBand adapters per canister, for a total of 8 adapters per Scale System 6000. These adapters should directly be connected to MQM9700 switches. The storage network can be scaled as required to meet any requirements for DGX compute systems, as well as storage capacity and performance. The standard SuperPOD spine-and-leaf configuration is recommended, as described in the NVIDIA SuperPOD reference architecture.

In addition to the InfiniBand connectivity for the Scale System 6000, the required EMS and a minimum of 2 protocol nodes require connectivity to the Storage Fabric. The EMS requires a single 200 Gbps NDR connection, while the protocol nodes should have 2 200 Gb connections each.

IPMI/BMC Network

The Scale System 6000 requires a dedicated management network to manage the systems, perform software upgrades, monitor, and configure the systems. While any network can be used for this task, it is recommended to use an IBM provided switch, which will be pre-configured and managed by the system.

Management Network

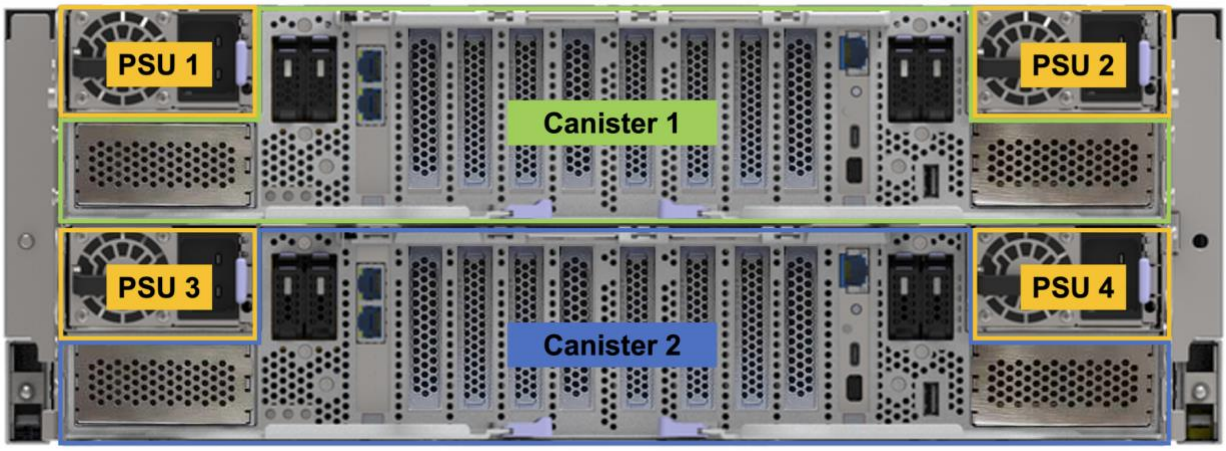
To manage the system, one connection from the management node to the management network is required. This connection provides management access and can optionally be configured for call-home.

Other Networking Considerations

Optional features described earlier in this document such as multi-protocol access or AFM may require additional networking capabilities to connect to external storage systems, object storage, or clients. The quantity and speed of these connections will vary depending on existing infrastructure, the size of the solution and other factors.

Power

Each Scale System 6000 enclosure contains 4 redundant power supplies. The power supplies are designed to provide redundant power to each canister in the system.



In a 48 NVMe drive configuration, the following power measurements represent the maximum draw for the system

Product	kVA	Amps	Inlet	Watts	Input Power
Scale System 6000 48 NVMe drive	4.50	22.5	C20 (x4)	4800*	200 V to 240 V single phase 50 Hz or 60 Hz

					13 A (x4)

* This value represents the absolute maximum power draw. Actual usage is affected by system load, ambient temperature, and several other factors

Refer to the Scale System 6000 Hardware Planning Guide (https://www.ibm.com/docs/en/5149-F48/pdf/ess6000_hwg.pdf) for additional information regarding system power, cooling, and other considerations.

Sizing

The IBM Storage Scale System 6000 has a flexible architecture and can scale to meet nearly any performance or capacity requirement. Each 4U Scale System 6000 enclosure can support up to 48 NVMe devices and offers up to 330 GB/sec seq. read speeds. Both capacity and performance can be scaled nearly linearly by adding more Scale System 6000 enclosures.

In addition to NVMe drives, each Scale System 6000 system can support up to 9 SAS-attached enclosures to expand capacity with low-cost storage. The Scale System 6000's Integrated Lifecycle Management (ILM) functionality can seamlessly move data between the high performing NVMe devices and slower drives to scale capacity in a cost-effective way. This reference only contains NVMe solutions, however mixed NVMe and SAS solutions can be deployed to meet additional capacity requirements.

The following tables contains guidance from NVIDIA regarding the storage performance required for many AI workloads.

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

Guidance Table for NVIDIA DGX GB200

(Source: https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-gb200/latest/storage-architecture.html)

Performance Characteristic	**Standard (GBps)**	**Enhanced (GBps)**
Single SU aggregate system read	90	280
Single SU aggregate system write	45	140
4 SU aggregate system read	360	1120
4 SU aggregate system write	180	560

Guidance Table for NVIDIA DGX GB300

(Source: https://docs.nvidia.com/pdf/dgx-spod-gb300-ra.pdf)

The following tables show how the Scale System 6000 can be scaled in order to meet the 'best' requirement depending on the number of scalable units deployed. The rack units and data network (400 Gb InfiniBand) connections required include the necessary management node, 2 protocol nodes, and internal switch for the configuration. These systems do not account for optional data movement via AFM, or additional protocol nodes.

SU	Scale System 6000	Read GB/s	Write GB/s	Rack Units	Storage Fabric Ports	Usable Capacity[1]		
						7.68 TB	15.36 TB	30.74 TB
1	1[2]	330	155	11	13[2]	0.25 PB	0.5 PB	1 PB
2	1	330	155	11	13	0.25 PB	0.5 PB	1 PB
4	1	330	155	11	13	0.25 PB	0.5 PB	1 PB

Standard Performance table for NVIDIA DGX GB200

SU	Scale System 6000	Read GB/s	Write GB/s	Rack Units	Storage Fabric Ports	Usable Capacity[1]		
						7.68 TB	15.36 TB	30.74 TB
1	1[2]	330	155	11	13[2]	0.25 PB	0.5 PB	1 PB
2	1	330	155	11	13	0.25 PB	0.5 PB	1 PB
4	2	660	310	15	21	0.5 PB	1 PB	2 PB

Enhanced Performance table for NVIDIA DGX GB200

1. Usable capacity is an approximation assuming 8+2P and standard spare space. Actual capacity may vary
2. For the single SU configuration, it is possible to only ½ of the NVMe and ConnectX-7 adapters (with reduced performance and capacity) and still meet the standard category. Fully populating the system will allow for more flexible expandability to meet future needs. The guide assumes a fully populated system.

SU	Scale System 6000	Read GB/s	Write GB/s	Rack Units	Storage Fabric Ports	Usable Capacity[3]		
						7.68 TB	15.36 TB	30.74 TB
1	1[4]	330	155	11	13[4]	0.25 PB	0.5 PB	1 PB
2	1	330	155	11	13	0.25 PB	0.5 PB	1 PB
4	2	660	310	15	21	0.5 PB	1 PB	2 PB

Standard Performance table for NVIDIA DGX GB300

SU	Scale System 6000	Read GB/s	Write GB/s	Rack Units	Storage Fabric Ports	Usable Capacity[3]		
						7.68 TB	15.36 TB	30.74 TB

1	1[4]	330	155	11	13[4]	0.25 PB	0.5 PB	1 PB
2	2	660	310	15	21	0.5 PB	1 PB	2 PB
4	4	1320	620	23	37	1 PB	2 PB	4 PB

Enhanced Performance table for NVIDIA DGX GB300

3. Usable capacity is an approximation assuming 8+2P and standard spare space. Actual capacity may vary
4. For the single SU configuration, it is possible to only ½ of the NVMe and ConnectX-7 adapters (with reduced performance and capacity) and still meet the standard category. Fully populating the system will allow for more flexible expandability to meet future needs. The guide assumes a fully populated system.

Summary

NVIDIA evaluations show that the IBM Storage Scale System 6000 with Storage Scale file system meets the 'best' DGX SuperPOD performance and functionality requirements. It is a great choice to pair with a DGX SuperPOD to meet current and future storage needs.

The Scale System 6000 can tier data to hard disk, tape, and object storage to deliver a cost-effective solution. The robust integrated lifecycle management (ILM) engine automatically moves data to the appropriate storage type to deliver high performance while moving unused data to a more cost-effective form of storage. In addition, global file sharing using the active file management (AFM) technologies allows for an organization to seamlessly share data across the world.

As storage requirements grow, IBM Scale System 6000 building blocks can be added to seamlessly scale capacity, performance, and capability. The combination of NVMe hardware and IBM Storage Scale parallel file system architecture provides excellent random read performance, often just as fast as local storage for sequential read patterns. Testing has validated that each IBM Scale System 6000 can deliver the highest levels of per node performance and meet all our application performance requirements. The IBM Storage Scale parallel file system provides a platform that is fully supported with the NVIDIA DGX SuperPOD and has been deployed in production.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
February 2026