IBM Storage Scale and IBM Storage Scale System 6000 for NVIDIA Enterprise Partners

NVIDIA HGX H100, H200, and B200 Based Servers High-Performance Storage Reference Architecture

August 2025

Table of Contents

1. IBM Storage Scale for NVIDIA Enterprise Partners

The NVIDIA HGX™ platform provides a fully optimized AI and HPC hardware and software stack capable of solving some of the most challenging computational problems today. The IBM Storage Scale System has been designed to provide the necessary performance, bandwidth, and reliability to support these demanding deep learning, artificial intelligence, and high-performance computing workloads.

I This document describes how to use the IBM Storage Scale System 6000 to support the NVIDIA Enterprise Reference Architecture with NVIDIA HGX H100, H200, or B200 systems using the 2-8-9-400 (8 GPU) node configuration.

IBM Storage Scale System 6000 is a storage appliance offering low-latency NVMe physical storage, advanced erasure coding, and support for NVIDIA Spectrum™ Ethernet networking. Multiple IBM Storage Scale System 6000's can be aggregated to create a high-performance clustered 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 storage system that makes it easy to deploy, manage, and grow fast storage for AI with NVIDIA HGX servers.

As configured, tested, and deployed in the NVIDIA Enterprise environment, the IBM Storage Scale System 6000 can be used for all AI/ML workloads and includes:

- Efficient training and checkpointing of AI models with data directly accessed from IBM Storage Scale.
- Automatic caching of local resources to minimize rereading of 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.

2. IBM Storage Scale Solutions with NVIDIA HGX Servers

IBM Storage Scale is the perfect solution for NVIDIA Enterprise deployments because of great scalability and high performance.

Key advantages of IBM Storage Scale for NVIDIA Enterprise workloads are:

- Meets and exceeds NVIDIA performance guidelines
- RDMA communication with NVIDIA GPU Direct® Storage (GDS) support
- Great scalability with up to 16k nodes and Exascale filesystems

- Comprehensive System monitoring. seamless integration with third party tools (NVIDIA Base Command™ Manager, Prometheus, Grafana, etc.)

3. IBM Storage Scale System

3.1 Hardware components

3.1.1 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, 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 connect to compute clusters supporting analytics or AI workloads.

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 excellent performance scalability. Scale System 6000 solutions may also be used as an all-flash NVMe performance tier combined with higher latency, 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 IBM Storage Scale's 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 (SIEM) platforms.
IBM Storage Scale Systems integrate with NVIDIA software such as NVIDIA Base Command Manager to streamline administration and configuration of the entire solution.

3.1.2 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 one NVIDIA ConnectX®-7 dual-port adapter, providing up to 400 GbE connectivity. The management server requires one connection from the ConnectX-7 to the storage fabric.

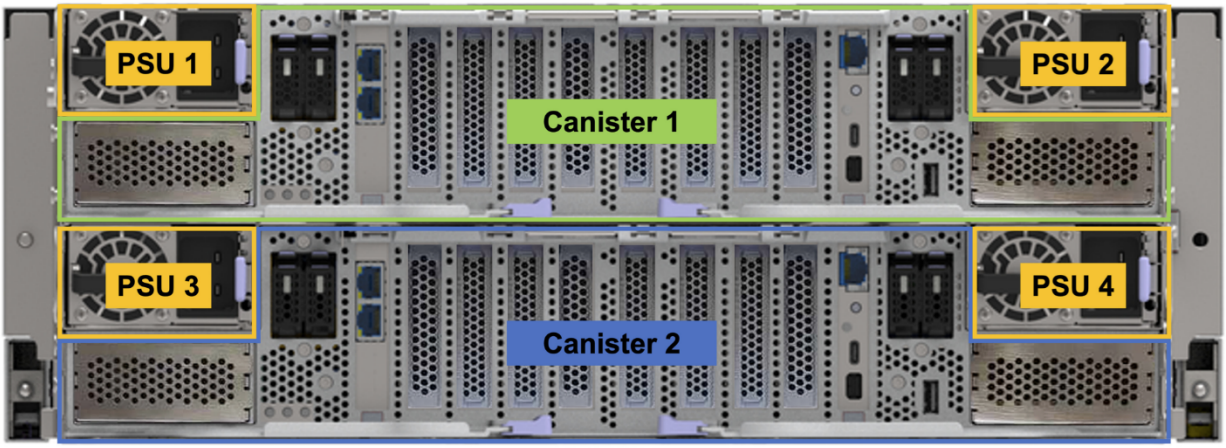
Front View of the Scale System Utility Node

In addition, a dedicated management switch from the management server is recommended, 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.

3.1.3 Scale System Power

Each IBM Scale System 6000 enclosure contains four 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. For example, for a fully populated IBM Storage Scale System 6000 (48 NVMe drives) at a theoretical 100% system utilization, the estimated power consumption is ~3.2kW.

Refer to the Scale System 6000 Hardware Planning Guide for additional information regarding system power, cooling, and other considerations.

3.2 IBM Storage Scale System – Software Architecture and Features

3.2.1 Native High-Performance IBM Storage Scale Client

IBM Storage Scale is optimized for high-performance parallel I/O and AI workloads. Due to the native Scale client, I/O is fully parallelized and distributed across the storage nodes and volumes, to get the best performance out of the attached storage systems. The native client takes care of optimizing I/O patterns and prefetching data as needed.
The IBM Storage Scale client provides the appearance of a mountable POSIX

file system to the applications and users on the workstation where the IBM Storage Scale client is installed.
The preferred method of accessing IBM Storage Scale data is to install the IBM Storage Scale client on every workstation or server that accesses IBM Storage Scale data. The IBM Storage Scale client establishes multiple connections (TCP/RDMA) to the data servers to provide high-performance parallel throughput. While doing so, IBM Storage Scale also manages full read/write data integrity between multiple users who are working with the data in the file system.
IBM Storage Scale supports NVIDIA GPU Direct Storage (GDS) that allows to directly read data into and write data from GPU memory.

3.2.2 Multi-Protocol Access

An IBM Scale System 6000 utilizes a high-speed, proprietary protocol to provide access to data. This protocol provides parallel, consistent, and redundant access to data concurrently from multiple systems. To access data using this protocol, clients require special software to be installed to provide access to the data. In the reference architecture, NVIDIA HGX systems require the Storage Scale client to be installed for high-speed access.

For external access to data stored on the Scale System 6000 by other users, the Storage Scale client can be used, or the solution can be configured with optional protocol nodes to support NFS, SMB, HDFS, and low-latency S3 object access to data. This allows external users to access the data using standard protocols, and to read, write, or view data directly.

To take advantage of multi-protocol access, from 2 to 32 protocol nodes can be installed depending on the number of users, speed of access required, and protocols used. These nodes can either be Storage Scale Utility nodes, or any standard x86 or IBM Power system running RHEL or Ubuntu Linux. See the IBM Storage Scale FAQ for the latest OS's and releases supported.

3.2.3 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, IBM 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. This functionality allows cloud users to store data on external storage systems until processing is required, freeing local storage for multiple tenants.

3.2.4 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, the least recently used data is automatically migrated to lower cost, higher latency storage.

To extend capacity with lower-cost storage, IBM Storage Scale 6000 systems can be configured with optional SAS adapters and spinning disks. In addition, products such as IBM Storage Archive Enterprise Edition can be used to connect to external tape enclosures using additional nodes. Storage Archive Enterprise Edition can be used to connect to external tape enclosures using additional nodes.

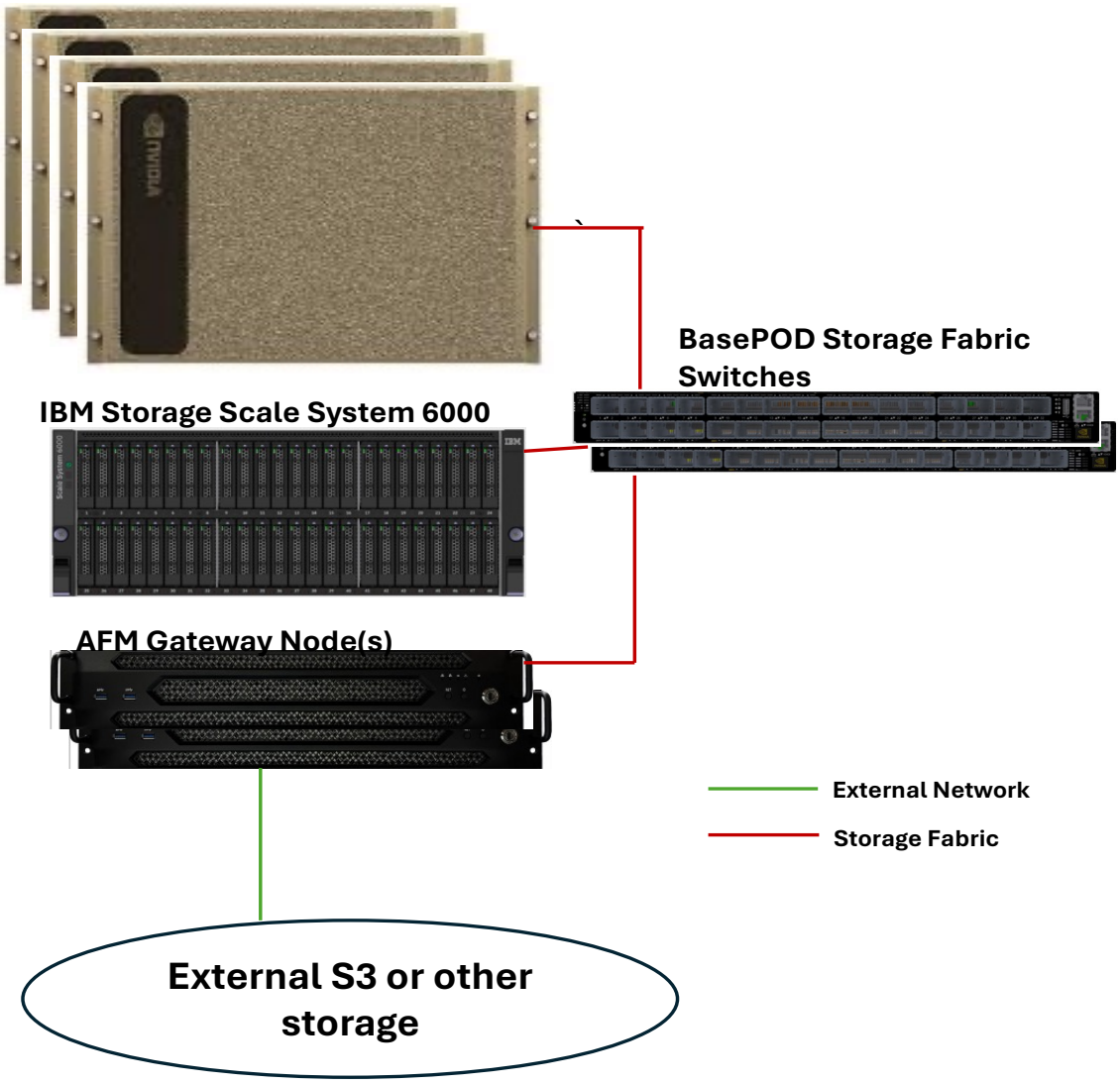
ILM can be used to extend the capacity in a single namespace. However, any data that is actively being used by HGX server or GPUs should reside on NVMe storage for optimal performance.

3.2.5 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. This capability allows for the file system to be extended beyond the Scale System 6000 to public or private clouds using the S3 protocol, to additional IBM Storage Scale clusters, or to other NFS-compliant storage systems.

The ability to tier data to S3-compliant clouds allows applications to create and modify data directly. That data is then cached to the Scale System 6000, which provides high-speed access to the HGX servers for high-speed training or other jobs. DL or AI algorithms then have the benefit of training with the most recent and up-to-date data automatically.

To provide connectivity, a Storage Scale cluster requires a minimum of one AFM gateway node, however at least two 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. The following diagram contains an example of how the system is connected.



Each AFM gateway node requires two connections to the storage fabric, with NVIDIA ConnectX-7 400 Gb/s Ethernet connectivity recommended. In addition, each gateway node requires at least one connection to a high-speed external 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, three factors should be considered:

- The size of the active data set being cached
- The number of files/objects being cached
- The change rate of the data that is cached

Depending on these factors, additional gateway nodes may be needed to meet the given workload. The IBM Storage Scale FAQ and the planning section of the knowledge center provide additional guidance on gateway node configuration.

4. Scale Architectures for NVIDIA Enterprise Deployments

4.1 Storage

The IBM Storage Scale system offers flexible deployment, fully supporting NVIDIA Spectrum Ethernet networks. In addition to gaining capacity, performance scales when adding additional Storage Scale 6000 building blocks to the system. In addition to the IBM Storage Scale 6000 building blocks, one Scale System utility node is required to manage the solution. A single utility node can manage multiple Storage Scale 6000 systems.

A dedicated management switch for the Storage Scale 6000 and management node can be ordered from IBM and is preconfigured for ease of use and deployment. However, the system can use existing 1 GbE switches such as the NVIDIA Spectrum SN2201 in the NVIDIA Enterprise out-of-band management network if they are configured appropriately. This configuration may require configuring VLANs on the switches to separate traffic as needed.

Storage performance requirements may vary depending on the workload, model size, parameters, cluster usage, and a great number of other factors. The deployments in this reference architecture provide at least 12.5 Gb/sec of read bandwidth per GPU, which is recommended by the NVIDIA Enterprise Reference Architecture. Due to the scalable performance the Storage Scale 6000 system provides, additional systems and hardware may always be used to provide additional performance if required.

4.2 Compute

IBM and NVIDIA have worked to ensure that the IBM Storage Scale System 6000 meets the performance requirements for NVIDIA HGX servers. The IBM Storage Scale System 6000 has been qualified for the following models.

Supported NVIDIA HGX baseboards
HGX H100, H200, B200

4.3 Storage Networking Recommendation

The IBM Storage Scale 6000 reference architecture utilizes the recommended converged network in the NVIDIA Enterprise Reference Architecture. The network utilizes NVIDIA SN5600 Ethernet leaf switches for scalable high-speed connectivity between compute nodes and the Storage Scale 6000 system. The SN5600 architecture also allows for the use of RDMA over Converged Ethernet (RoCE) for lower latency connectivity and the use of functions such as NVIDIA GPU Direct and NVIDIA GPUDirect Storage. While the IBM Storage Scale 6000 can operate with or without RoCE, RoCE is recommended for optimal performance and functionality.

The NVIDIA Enterprise Reference Architecture uses a spine-and-leaf configuration of SN5600 leaf switches to provide optimal speed and redundancy. Each IBM Storage Scale 6000 appliance in the reference architecture can be configured with up to eight NVIDIA dual-port ConnectX-7 200 GbE adapters, four per canister. As a result, each building block contains up to 16 200 GbE ports.

The NVIDIA Enterprise Reference Architecture provides connectivity to an external storage network. The IBM Storage Scale System, when running with 200 GbE, the Storage Scale System 6000 can meet the performance requirements of 12.5 Gb/sec per GPU without requiring additional switch infrastructure. If AFM, CES, or additional performance is required, additional NVIDIA SN5600 or similar switches can be used to uplink to the NVIDIA consolidated network and provide additional ports and functionality.

Finally, at least one EMS node, with an optional second for redundancy must also be attached to the converged network. The EMS node connectivity does not require high bandwidth, so can be connected at speeds as low as 100 GbE using shared splitter cables with other infrastructure or can be connected at higher speeds. The EMS node uses NVIDIA ConnectX-7 adapters for flexibility in connectivity. A single EMS node can manage all the Storage Scale 6000 systems in the deployment.

4.3.1 Recommended Cables and Transceivers

The NVIDIA Enterprise Reference Architecture allows either 100 GbE or 200 GbE for the Converged Network. For optimal performance with the Storage Scale System 6000, 200 GbE is recommended, with RoCE enabled. The following transceivers should be ordered for the Storage Scale System 6000 to provide the 200 GbE connectivity.

Cable Type	Switch Transceiver	Storage Scale 6000 Transceiver	Cable
Active Copper	MCA7J60-N00*X*	N/A	N/A
Optical Multimode	MMA4Z00-NS-T	MMA4Z00-NS400	MFP7E20-N*XXX*
Optical Single Mode	MMS4X00-NS-T	MMS4X00-NS400	MFP7E40-N*XXX*

*Note *'X'* indicates cable distance

Also, as noted, the EMS management node does require one to two connections per deployment regardless of the number of building blocks used. To maximize port counts, a splitter mentioned above may be used for the EMS, or a slower-speed 100 GbE connection can be used and shared with other systems that require connectivity.

4.4 NVIDIA Enterprise Deployments

The NVIDIA Enterprise Reference Architecture describes several different cluster sizes and GPU counts. This section details the recommended IBM Storage Scale 6000 deployments for these cluster sizes. Due to the flexibility that IBM SSS 6000 offers, the size of these deployments may easily scale up or down to meet nearly any workload. Additional functionality such as CES or AFM can also be added to these architectures but may require additional switches and hardware to provide the necessary connectivity. The requirement for these functions is described in their respective sections above.

4.4.1 Deployments with 4-12 Nodes with 32-96 GPUs

With up to 12 nodes, the NVIDIA Enterprise Reference Architecture provides for 12 200 GbE ports for the Storage Network.

Description	Count
Total Storage Network Ports	12
Total Number of IBM Storage Scale 6000 Systems	1
ConnectX-7 adapters required per Storage Scale 6000	6 dual-ported
200 GbE Ports connected to Storage Scale 6000	10
100 GbE/200 GbE ports connected to EMS	1
Aggregate Read Throughput	225 GB/sec
Aggregate Write Throughput	150 GB/sec
Read Throughput per GPU (assuming 96 GPUs)	2.3 GB/sec

4.4.2 Deployments with 16-48 Nodes with 128-384 GPUs

For deployments with up to 48 nodes, the NVIDIA Enterprise Reference Architecture provides for 32 200GbE ports for the Storage Network. Due to the variation in number of GPUs, a single IBM Storage Scale System 6000 may provide enough bandwidth for 128-256 GPUs. However, for additional GPUs, a second system may be required. Therefore, in this section, we provide two separate sets guidance, the first for up to 256 GPUs, the second for up to 384 GPUs.

Up to 256 GPUs

Description	Count
Total Storage Network Ports	32
Total Number of IBM Storage Scale 6000 Systems	1
ConnectX-7 adapters required per Storage Scale 6000	8 dual-ported
200 GbE Ports connected to Storage Scale 6000	16
100 GbE/200 GbE ports connected to EMS	1
Aggregate Read Throughput	320 GB/sec
Aggregate Write Throughput	155 GB/sec
Read Throughput per GPU (assuming 256 GPUs)	1.25 GB/sec

Up to 384 GPUs

Description	Count
Total Storage Network Ports	32
Total Number of IBM Storage Scale 6000 Systems	2
ConnectX-7 adapters required per Storage Scale 6000	4 dual-ported
200 GbE Ports connected to Storage Scale 6000	28
100 GbE/200 GbE ports connected to EMS	1
Aggregate Read Throughput	640 GB/sec
Aggregate Write Throughput	310 GB/sec
Read Throughput per GPU (assuming 384 GPUs)	1.6 GB/sec

4.4.3 Deployments with 64-128 Nodes with 512-1024 GPUs

For deployments with up to 128 nodes, 3 or 4 IBM Storage Scale System 6000 systems can provide 1.2 GB/sec per GPU, which can fulfil several needs. Due to the flexible nature of the SSS 6000, up to 7 building blocks can fit into the infrastructure to provide extreme bandwidth for many of the most challenging workloads. This flexibility makes the Storage Scale 6000 an excellent choice, as performance can be easily tuned, and if necessary increased simply by adding more building blocks. In this section, we will show our recommended guidance of 4 systems.

Recommended configuration, up to 1024 GPUs

Description	Count
Total Storage Network Ports	128
Total Number of IBM Storage Scale 6000 Systems	4
ConnectX-7 adapters required per Storage Scale 6000	8 dual-ported
200 GbE Ports connected to Storage Scale 6000	64
100 GbE/200 GbE ports connected to EMS	1
Aggregate Read Throughput	1280 GB/sec
Aggregate Write Throughput	620 GB/sec
Read Throughput per GPU (assuming 1024 GPUs)	1.25 GB/sec

If greater performance is required, up to 7 IBM Storage Scale 6000 systems can be used, providing nearly 2.2 GB/sec of bandwidth per GPU. This balance of cost, performance, and capacity makes the IBM Storage Scale System 6000 an excellent choice for NVIDIA Enterprise deployments.

5. Solution Performance Validation

The IBM Storage Scale System 6000 provides excellent performance to meet AI training and inference needs. As models increase in complexity, the Scale System 6000's superior write bandwidth allows for efficient checkpointing, allowing GPUs to spend more time training and less time waiting for data.

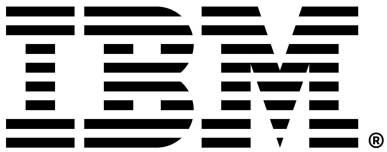
Due to the Scale System 6000's scalability, adding additional units can easily increase performance to meet nearly any requirement.

6. Summary

NVIDIA and IBM have jointly worked to ensure that the IBM Storage Scale System 6000 meets the requirements for NVIDIA HGX server deployments for NVIDIA Enterprise deployments. The IBM Storage Scale System 6000 has been rigorously tested and validated by NVIDIA and IBM to ensure a seamless experience when paired with NVIDIA HGX servers.

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 nearly any application performance requirement. The IBM Storage Scale parallel file system provides a platform that is fully supported with NVIDIA HGX servers.



IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
August 2025