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

NVIDIA GB200 NVL72 Systems
High-Performance Storage Reference Architecture

July 2025

Table of Content

Introduction

The NVIDIA® Cloud Partner Reference Architecture (NCP RA) for AI training and inference, featuring NVIDIA® GB200 NVL72 systems, is a comprehensive recipe for building a high-performance solution for AI infrastructure by leveraging NVIDIA's expertise in GPU servers, storage, networking, and software, providing the next generation of data center architecture for artificial intelligence (AI) for NVIDIA Cloud Partners (NCPs). Customers working with NCPs who have implemented this architecture benefit from optimal performance, proven reliability, support from NVIDIA's subject-matter experts, and the flexibility to scale as demand grows, ensuring they stay competitive in the rapidly evolving AI market. This document covers the high-performance storage (HPS) reference architecture for IBM Storage Scale System 6000.

1. IBM Storage Scale for NVIDIA Cloud Partners

The NVIDIA Cloud Partners (NCPs) with NVIDIA GB200 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 fuel future innovation.

In this reference architecture (RA), we will describe how to use the IBM Storage Scale System 6000 to support DL workloads for NCPs. IBM Storage Scale System 6000 is a storage appliance offering low-latency NVMe physical storage, advanced erasure coding, and supports NVIDIA Spectrum™ 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 storage system that makes it easy to deploy, manage, and grow fast storage for AI with NVIDIA GB200 systems .

As configured, tested, and deployed in the NCP's environment, the IBM Storage Scale System 6000 can be used for all 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 GB200 systems

IBM Storage Scale is the perfect solution for NCP because of its great scalability and high performance.
Key advantages of IBM Storage Scale for NCPs are:

- Meets and exceeds NVIDIA standard performance guidelines
- RDMA communication with NVIDIA GPUDirect® Storage (GDS) support
- Great scalability up to 16k nodes and Exascale filesystems
- Comprehensive System monitoring
- Seamless integration with third-party tools (Prometheus, Grafana, etc.)

3. IBM Storage Scale System

3.1 Hardware components

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

The 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 and 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.

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 Gb/s Ethernet connectivity. The management server requires one connection from the ConnectX-7 to the storage fabric.

Front View of the Scale System Utility Node

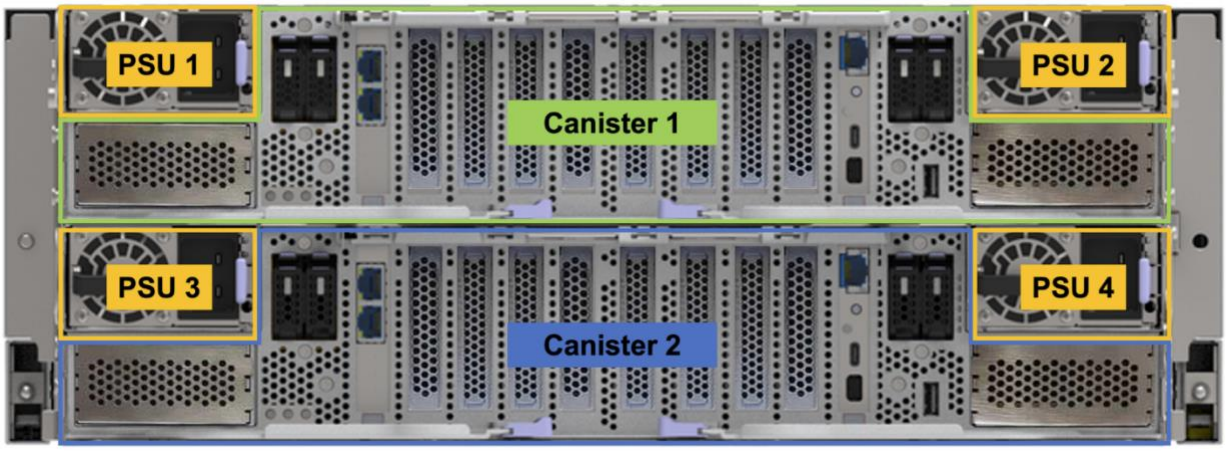
In addition, a dedicated management switch from the management server is recommended, to provision and manage the system. This switch may be customer-owned; however, an IBM switch is recommended as it is pre-configured with the appropriate configuration to manage the Scale System 6000.

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.

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)	4,800*	200V to 240V single-phase 50 Hz or 60 Hz 13A (x4)

* This value represents the absolute maximum power draw. System load, ambient temperature, and several other factors affect actual usage. 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.

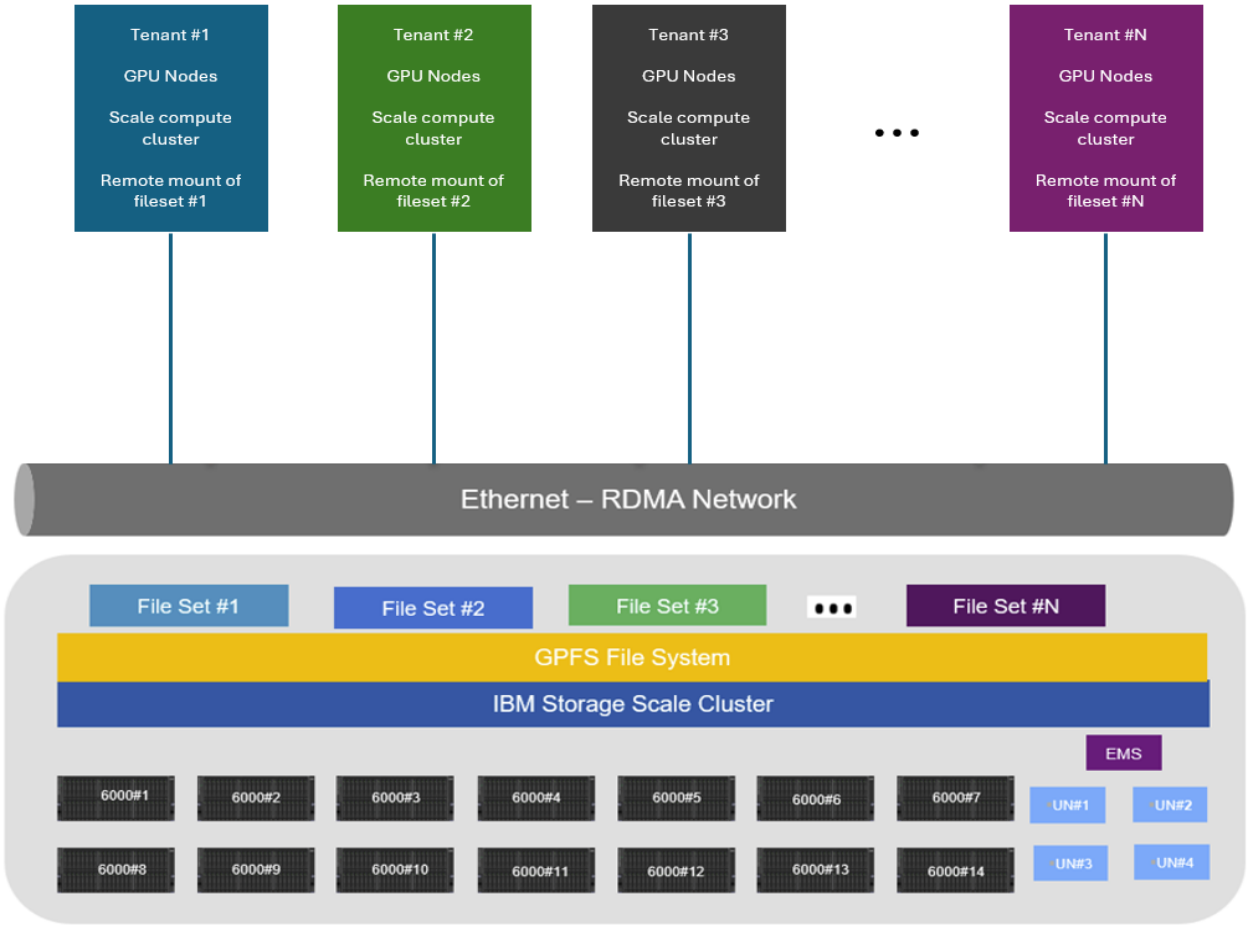
3.2 IBM Storage Scale System – Software Architecture and Features

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 GPUDirect® Storage (GDS), which enables a direct data path between storage and GPU memory, bypassing the CPU and system memory for both read and write operations.

Storage Scale Multitenancy

IBM Storage Scale provides powerful multi-cluster functionality, which makes it easy to build a secure multi-tenant environment. Each tenant forms its own Scale cluster with a set of NVIDIA GB200 systems, which consumes the storage from the central storage cluster, provided by the IBM Storage Scale System 6000.
Simply share the same storage with all your NVIDIA GB200 systems and segregate it for multi-tenancy, either on filesystem or fileset (directory tree) level.
Tenant data is secured by IBM Storage Scale encryption, where each tenant uses its own encryption keys and own key server.
IBM Storage Scale QoS functionality ensures that tenants cannot overconsume the available resources (bandwidth, IOPS) and Quotas efficiently limit the available capacity and inodes they can consume.
Tenant networks can further be isolated by using per-tenant VLANs. With that, NVIDIA Cloud Partners get a powerful, secure, and flexible solution to implement multi-tenant infrastructures.



4. Scale Architectures for NCP Deployments

4.1 Storage

The IBM Storage Scale system offers flexible deployment, supporting NVIDIA Spectrum Ethernet networks. In addition to gaining capacity, performance scales when adding 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 NCP's out-of-band management network if they are configured appropriately. This configuration may require configuring VLANs on the switches to separate traffic as needed.

When configuring the storage, it is recommended to refer to Table 1 below for sizing guidelines. Some workloads may benefit from additional throughput or additional capacity. These additional requirements can be met by adding IBM Scale System 6000 building blocks.

The storage performance target for training or inference can vary depending on the type of model and dataset. The guidelines in Table 1 provide standard throughput for the various Scalable Unit (SU) system sizes. The final HPS requirements for throughput and capacity will be specified for each NCP opportunity.

Table 1. Guidance for Standard HPS aggregate storage performance

Compute Components			Performance (GB/s)	
SUs	NVL72 Racks	GPUs	Read	Write
1	16	1,152	192	96
4	64	4,608	768	384
8	128	9,216	1,536	768
16	256	18,432	3,072	1,536

4.2 Compute

IBM and NVIDIA have worked to ensure that the IBM Storage Scale System 6000 meets the performance requirements for NVIDIA GB200 systems.

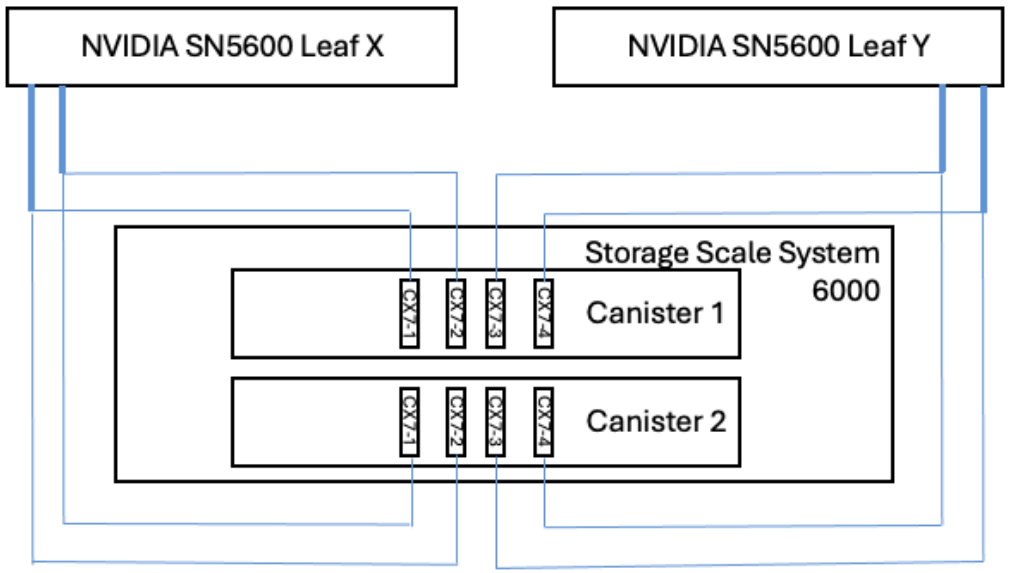
4.3 Storage Networking Recommendation

The IBM Storage Scale 6000 RA utilizes the recommended converged network in the NCP Reference Architecture (NCP RA). 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 using functions such as NVIDIA GPUDirect®. While the IBM Storage Scale 6000 can operate with or without RoCE, RoCE is recommended for optimal performance and functionality.

The NCP RA uses a spine-and-leaf configuration of SN5600 leaf switches to provide optimal speed and redundancy. Each IBM Storage Scale 6000 appliance in the RA should be configured with eight NVIDIA single-port ConnectX-7 400 GbE adapters, four per canister.

The NVIDIA SN5600 contains 64 800 GbE OSFP ports. To maximize port count and density, using splitter cables is recommended. IBM recommends utilizing an 800 GbE to 2x400 GbE cable. For short connections, copper cables can be used, or optical connections of varying length may be used for longer cable runs.

With splitter cables used, a pair of leaf switches can be connected to a Storage Scale 6000 system with two OSFP ports per switch, as shown in the following diagram.



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 200 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 NCP deployment.

Recommended Cables and Transceivers

As noted in the previous section, using splitters is recommended in any deployment, to reduce the number of ports required by the deployment. The following table lists options for optical transceivers, as well as copper that can be used to connect the NVIDIA SN5600 to the IBM Storage Scale 6000. It is recommended to tailor the cable types and lengths for a particular deployment. Again, note that four splitter cables, consuming four OSFP ports on the NVIDIA SN5600 will be required per IBM Storage Scale 6000 building block that is deployed.

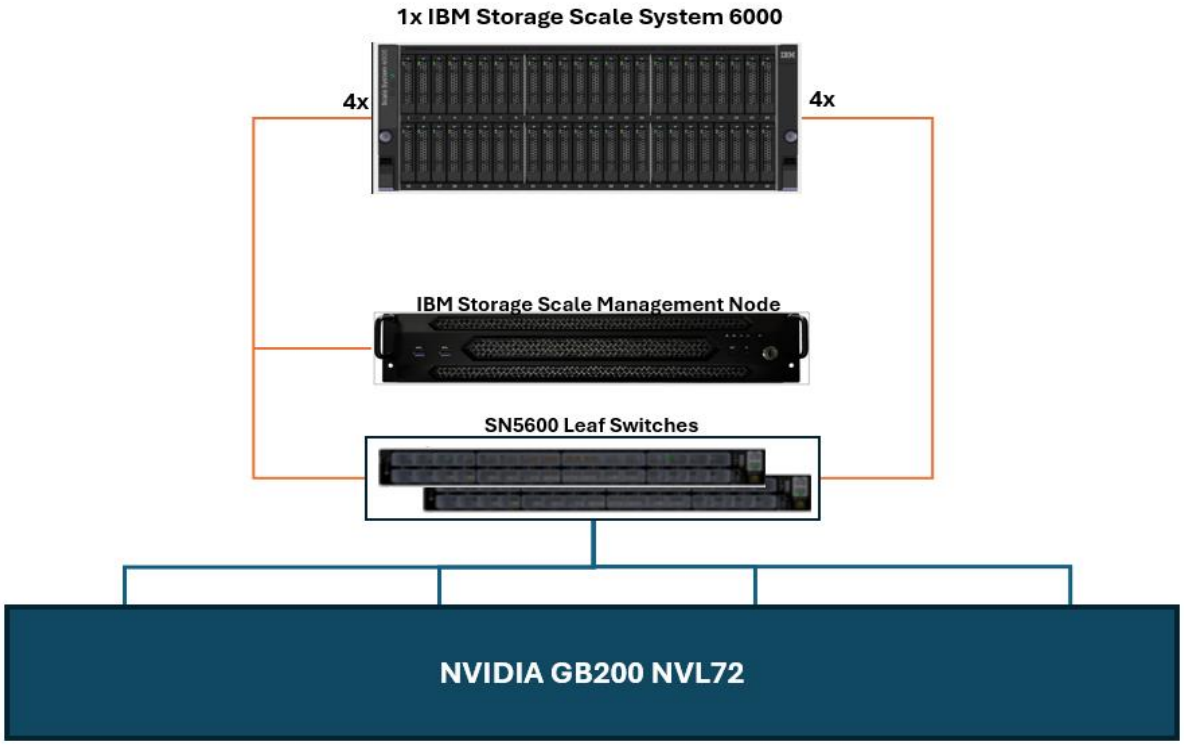
Cable Type	Switch Transceiver	Storage Scale 6000 Transceiver	Cable
Active Copper	MCA7J60-N00*X*	N/A	N/A
Optical Multimode	MMA4Z00-NS-T	MMA4Z00-NS400-T	MFP7E10-N0*XX*
Optical Single Mode	MMS4X00-NS-T	MMS4X00-NS400	MFP7E30-N*XXX*

*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 200 GbE connection can be used.

4.1. NCP Deployments with 1,152 GPUs (1 SU)

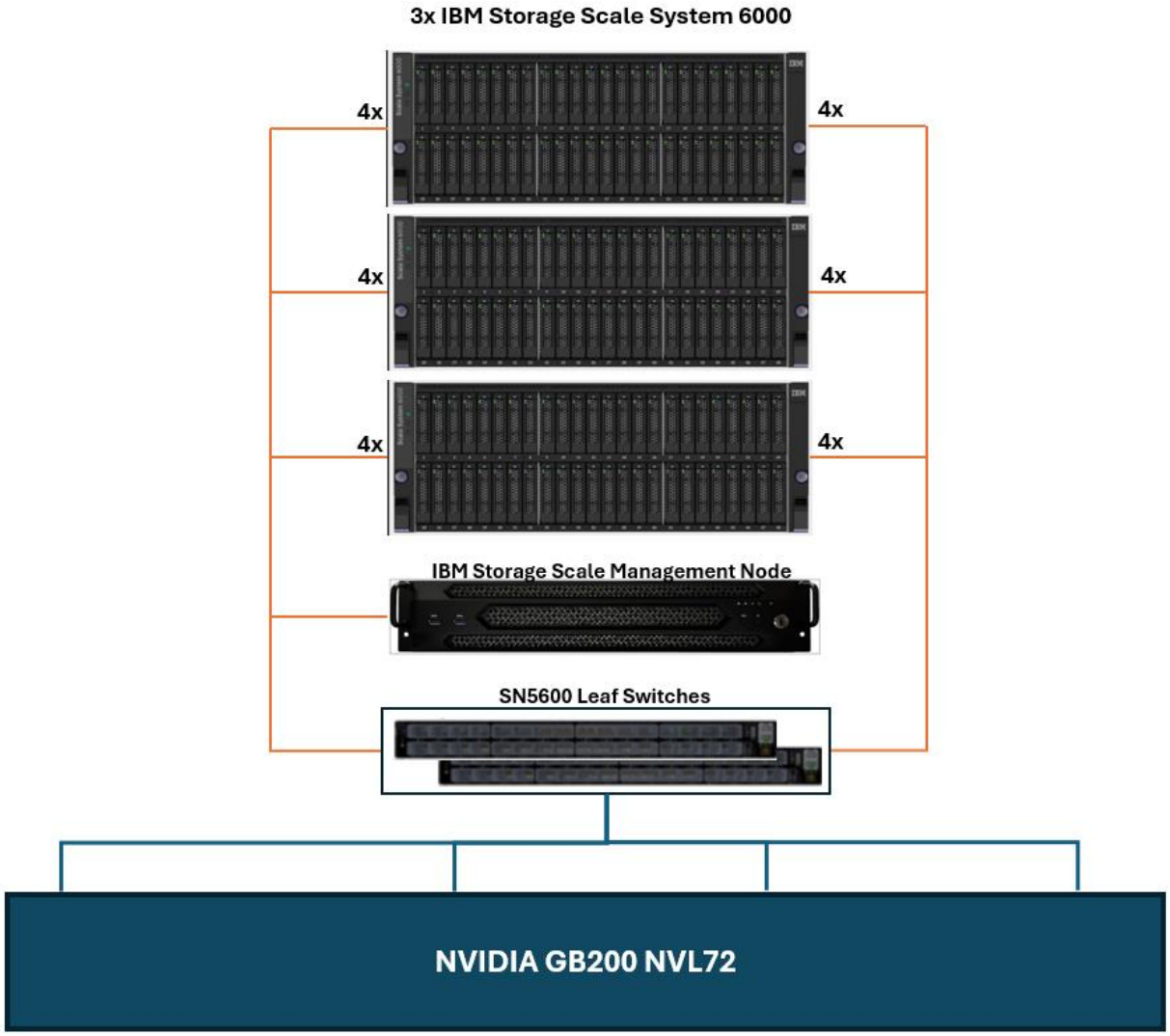
This section describes the IBM Storage Scale System deployment for 1 SU to achieve the standard performance target. Additional IBM Storage Scale Systems can be added to get better performance and exceed the performance target.



	Scalable Units (SU)	**1**
Compute	NVL72 Racks	16
	NVIDIA GPUs	1,152
Storage Fabric	Storage Ports 400 GbE	8
	Management Node Ports 400 GbE	1
	Switch Ports 800 GbE with 2y-splitter cable	4 + 1
IBM Storage	IBM Storage Scale System 6000	1
	Aggregate read throughput (GB/s)	330
	Aggregate write throughput (GB/s)	150

4.2. NCP Deployments with 4,608 GPUs (4 SUs)

This section describes the IBM Storage Scale System deployment for 4 SUs to achieve the standard performance target. Additional IBM Storage Scale Systems can be added to get better performance and exceed the performance target.

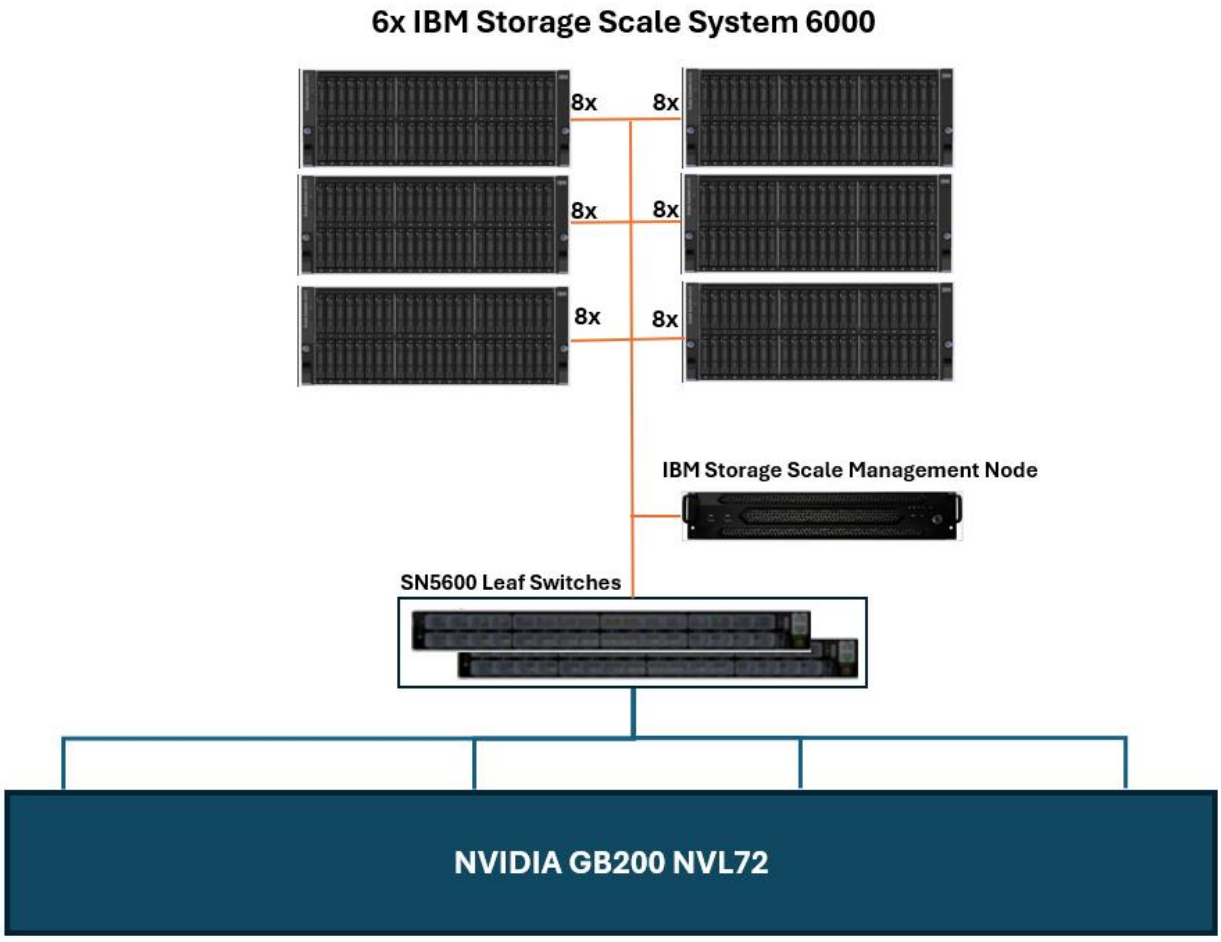


	Scalable Units (SU)	**4**
Compute	NVL72 Racks	64
	NVIDIA GPUs	4,608
Storage Fabric	Storage Ports 400 GbE	24
	Management Node Ports 400 GbE	1

	Switch Ports 800 GbE with 2y-splitter cable	8 + 1
IBM Storage	IBM Storage Scale System 6000	3
	Aggregate read throughput (GB/s)	990
	Aggregate write throughput (GB/s)	450

4.3. NCP Deployments with 9,216 GPUs (8 SUs)

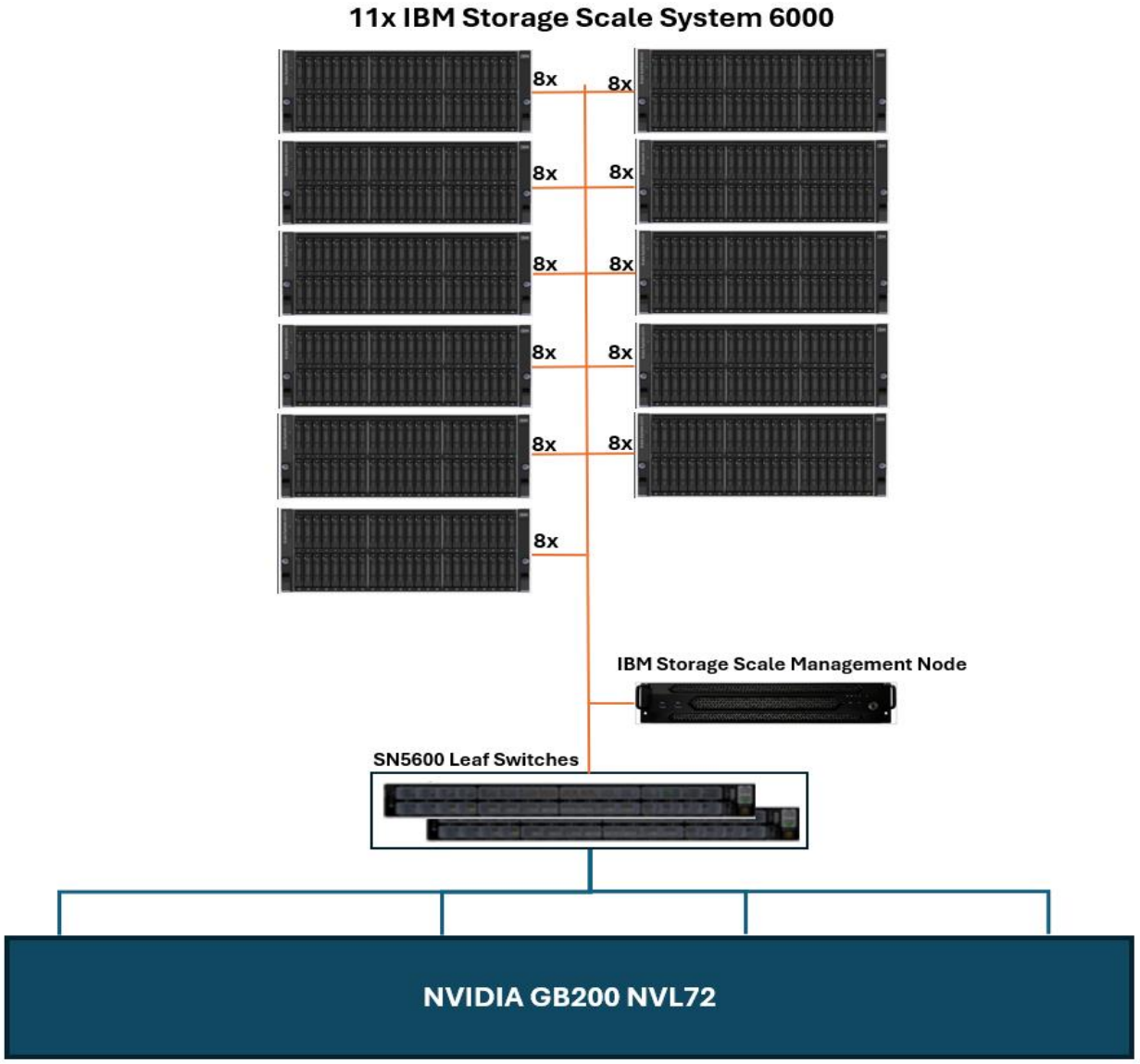
This section describes the IBM Storage Scale System deployment for 8 SUs to achieve the standard performance target. Additional IBM Storage Scale Systems can be added to get better performance and exceed the performance target.



	Scalable Units (SU)	**8**
Compute	NVL72 Racks	128
	NVIDIA GPUs	9,216
Storage Fabric	Storage Ports 400 GbE	48
	Management Node Ports 400 GbE	1
	Switch Ports 800 GbE with 2y-splitter cable	20 + 1
IBM Storage	IBM Storage Scale System 6000	6
	Aggregate read throughput (GB/s)	1,980
	Aggregate write throughput (GB/s)	900

4.4. NCP Deployments with 18,432 GPUs (16 SUs)

This section describes the IBM Storage Scale System deployment for 16 SUs to achieve the standard performance target. Additional IBM Storage Scale Systems can be added to get better performance and exceed the performance target.



	Scalable Units (SU)	**16**
Compute	NVL72 Racks	256
	NVIDIA GPUs	18,432
	Storage Ports 400 GbE	88

Storage Fabric	Management Node Ports 400 GbE	1
	Switch Ports 800 GbE with 2y-splitter cable	36 + 1
IBM Storage	IBM Storage Scale System 6000	11
	Aggregate read throughput (GB/s)	3,630
	Aggregate write throughput (GB/s)	1,650

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 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 GB200 systems deployments for NCPs. 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 GB200 systems .

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. Additionally, global file sharing using active file management (AFM) technologies allows for an organization to share data seamlessly 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 GB200 systems.

7. Appendix - Advanced Scale Features

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 RA, NVIDIA GB200 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, a minimum of 2, and up 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.

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.

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 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 NVIDIA GB200 systems or GPUs are actively using should reside on NVMe storage for optimal performance.

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 NVIDIA GB200 systems for high-speed training or other jobs. Deep learning 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. See the IBM Storage Scale FAQ for the latest guidelines.

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 of 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.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
July 2025