IBM Storage Scale System 6000

Accelerated infrastructure for artificial intelligence, high-performance computing, analytics, and hybrid cloud

Highlights
Integrated hardware and software for building a global data platform

NVIDIA GPUdirect support to accelerate AI training and inferencing performance

Provides distributed file and object storage for the most demanding AI, HPC, analytics, and hybrid cloud workloads

Up to 340 GB/s throughput and 28 million IOPS; up to 4.7 PB raw capacity in a compact 4U rack

The rapid rise of artificial intelligence (AI), high-performance computing (HPC), analytics, and hybrid cloud is transforming the business world. Unlike traditional applications that rely on structured databases, these modern workloads and data lakes process vast amounts of unstructured data, including documents, audio, images, and videos.

To stay competitive, organizations must rethink their storage strategies and leverage AI to unlock the value of their data—wherever it resides. IT leaders face key challenges, including:
- Accessing and analyzing data and workloads scattered across the globe.
- Managing the growing AI infrastructure and ensuring scalability for evolving workloads.
- The increasing time needed by AI training and inferencing workloads.

Addressing these challenges requires specialized software and hardware:
- **IBM Storage Scale** is software-defined file and object storage for both structured and unstructured data.
- **IBM Storage Scale System 6000** is a hardware implementation of Storage Scale software and is optimized for the most demanding AI, HPC, analytics, and hybrid cloud workloads.
- **IBM Scale System All-Flash Expansion Enclosure** greatly expands the maximum capacity of Storage Scale 6000 configurations.
- **IBM Storage Scale System 3500** is for customers requiring an enterprise-ready entry-level or mid-level system.

IBM Storage Scale
The unstructured and semi-structured data from AI workloads, advanced analytics, data lakes, and other data-intensive apps must be stored in distributed file and object systems to make it accessible to geographically dispersed applications, services, and devices.

"We fundamentally believe that core to the competitiveness of every company going forward will be their ability to use AI to unlock real-time value from their data wherever the data resides."

Arvind Krishna – CEO, IBM

IBM Storage Scale software is designed to address these requirements with global data abstraction services that provide connectivity from multiple data sources and multiple locations to bring together data wherever it lives, including non-IBM storage environments.

Storage Scale achieves this with active file management (AFM), which abstracts other storage systems, including customers' existing on-prem or public cloud storage, while providing a transparent caching layer that improves performance by unifying data into a single global namespace, accessible simultaneously via multiple protocols.

Built on a massively parallel file system, Storage Scale supports deployment across x86, IBM Power, IBM Z, ARM-based POSIX clients, virtual machines, and Kubernetes. It delivers high-performance access to unstructured data across any protocol, location, or format—helping customers accelerate AI pipelines and unlock the value of their data estate.

Unlocking AI Potential with Content-Aware Storage

Very little enterprise data has been indexed for generative AI applications, which prevents AI assistants from providing accurate, up-to-date answers. The content-aware storage capabilities in Storage Scale address this challenge by extracting the semantic meaning hidden inside unstructured data so that AI assistants can automatically generate smarter answers.

Storage Scale enriches data using embedded compute and data pipelines that minimize data movement and latency to help reduce costs and improve performance. It now introduces asynchronous notifications, enabling faster, event-driven data ingestion into AI inferencing workflows, reducing latency and improving operational efficiency.

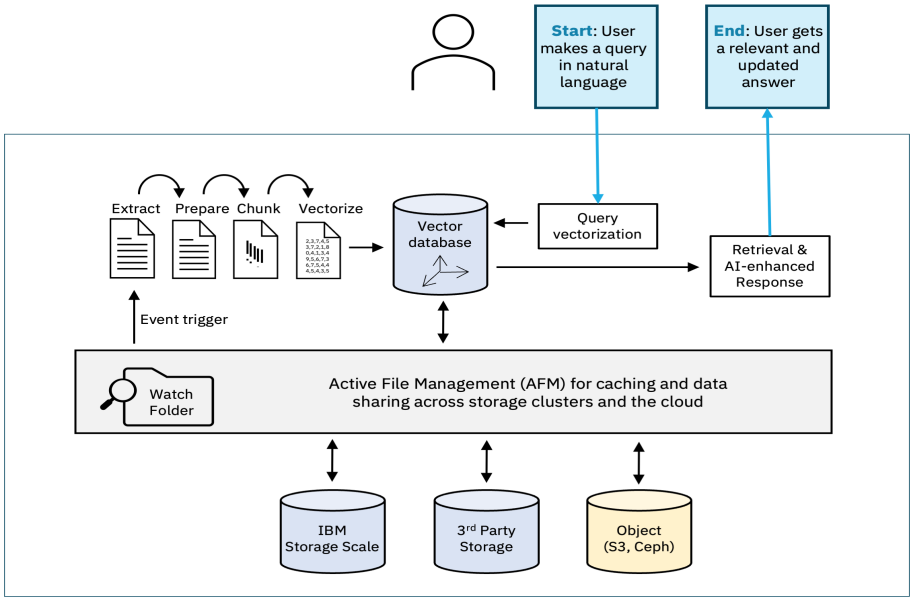

Figure 1. Storage Scale automates data extraction, vectorization, and storage updates, enabling seamless retrieval via a vector database. When users submit natural language queries, AI enhances search results for optimized responses.

IBM Storage Scale System 6000

Figure 2 – The IBM Storage Scale System 6000 delivers up to 340 GB/s throughput, 28 million IOPS, and up to 2.2 PB of raw capacity in a compact 4U rack.

IBM Storage Scale System 6000 is a high-performance hardware platform that combines IBM Storage Scale software with NVMe flash technology to deliver extreme throughput and massive scalability for AI, data analytics, and storage-intensive workloads. It supports up to nine SAS hard disk drive expansion enclosures for flexible capacity growth and leverages IBM Storage Scale RAID erasure coding for enhanced data efficiency, hardware failure mitigation, and intelligent monitoring with dynamic data tuning.

Delivered as containerized software, it simplifies deployment and maintenance, while Safeguarded Copy provides cyber-resilient, point-in-time snapshots that create isolated, immutable copies for rapid and secure data recovery in the event of a cyberattack.

Data Acceleration Tier (DAT)
Built on an NVMe over Fabrics (NVMe-oF) architecture, DAT boosts the performance and efficiency of IBM Storage Scale System 6000 by eliminating I/O bottlenecks that limit real-time AI workloads. Engineered for small, random I/O patterns typical of AI inferencing, it delivers consistent, predictable performance across large-scale environments. Working with IBM Storage Scale software, DAT provides extreme IOPS and ultra-low latency, enabling faster data access and decision-making across hybrid environments.

DAT offers the following key enhancements:

- High-performance design delivering up to 28 million IOPS and 340 GB/s throughput in NVMeoF configurations.
- Optimized for small and random I/O patterns to improve responsiveness in AI inferencing and real-time analytics.
- Seamless hybrid integration with existing IBM Storage Scale environments, supporting both dedicated acceleration tiers and mixed clusters.
- NVMe-optimized data path that enhances data ingestion and retrieval efficiency at scale.

DAT supports GPU-accelerated AI, data lakehouses, high-performance computing, IT modernization, and backup and archiving—workloads that benefit from automated tiering, advanced data placement, and secure long-term retention within the platform’s global data architecture.

Together with Content-Aware Storage, DAT provides a unified foundation that combines performance acceleration with intelligent data management for AI-driven workloads.

Accelerating AI Workloads
As AI and analytics datasets grow, slow data loading can bottleneck performance, limiting the full potential of high-speed GPUs. IBM Storage Scale System 6000 eliminates I/O constraints and optimizes processing for AI-scale workloads through NVIDIA GPUDirect Storage, enabling a direct data path between GPU memory and NVMe storage, including NVMe over Fabrics (NVMe-oF). By bypassing the host CPU and DRAM, this architecture reduces latency, maximizes GPU utilization, and accelerates AI-driven insights across distributed environments.

340 GB/s

High-performance throughput for demanding workloads

4.7 PB

Massive raw capacity in a compact 4U rack space

42.7 PB

Raw capacity per standard 42U rack

Storage Scale System 6000 is the simplest and fastest way to deploy a global data platform for AI and NVIDIA GPU infrastructure, delivering exceptional performance with CNSA support, enhanced Base Command Manager, NVIDIA Nsight, and certifications for NVIDIA BasePOD, SuperPOD, and Grace Blackwell platforms.

IBM Scale System All-Flash Expansion Enclosure

The IBM Scale System All-Flash Expansion Enclosure is designed to address workloads that operate on massive data sets, providing up to 100GB/s of throughput per enclosure, up to 200GB/s per rack, and more than 3PB of raw flash capacity in a 2U form factor.

The all-flash expansion enclosure integrates up to four NVIDIA BlueField-3 DPUs for maximum performance at scale, with up to 400Gb/s of high-performance network throughput and NVMeoF offload capability for the Storage Scale System.

The expansion enclosure is optimized for high performance AI training, data inferencing, HPC, and data intensive workloads. It supports up to 26 dual-port industry standard QLC flash drives in a 2U form factor. When configured with 128TB QLC NVMe SSDs, Storage Scale System 6000 can now deliver more than 42PB of cost-effective high-density flash capacity in a single 42U rack.

The scalability of the new All-Flash Expansion enclosure also supports bigger caches and multitenant service provider requirements. With IBM Storage Scale System 6000, multitenancy is highly configurable at the cluster, file system, or fileset level. This gives service providers and large-scale AI factories the ability to isolate workloads, enforce security policies, and optimize resource allocation without sacrificing performance. Whether supporting supercomputers, HPC environments, or AI-as-a-service offerings, this architecture ensures secure, scalable, and cost-efficient multitenant operations across diverse data-intensive workloads.

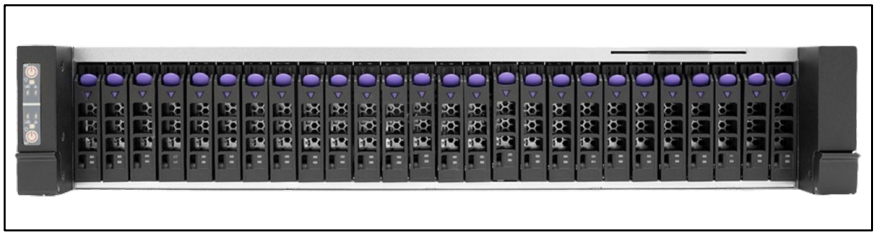

Figure 3 – The IBM Scale System All-Flash Expansion Enclosure enables organizations to cost-effectively deploy workloads operating on massive data sets.

For workloads with lower performance requirements, there's the disk-based IBM Storage Scale Expansion Enclosure, which economically extends the capacity of the Storage Scale System 6000. Each unit supports up to 91 self-encrypting SAS HDDs with 20TB or 22TB of storage, ensuring robust security and performance.

With support for up to nine enclosures per system, it scales to 18PB per rack over 24Gb SAS, allowing organizations to cost-effectively manage multi-petabyte workloads.

Specifications – IBM Storage Scale System 6000

System Features	• Active-active dual controller storage system: Dual socket AMD EPYC™ Genoa CPUs (48 cores) • 768 GB to 1536 GB memory per canister (system total is either 1536 GB or 3072 GB) • De-clustered RAID supporting erasure coding schemas: 3-way and 4-way replication, 4+2P, 4+3P, 8+2P, 8+3P • Up to 2.2 PB raw per 4U node or up to 5.4 PBe effective[1]
Software	• IBM Storage Scale for Storage Scale System software • Red Hat® Enterprise Linux® (RHEL)
Software Features	• Data access services with POSIX, NVIDIA GPUDirect, Container Native Storage Access (CNSA), CSI, HDFS, NFS v4, SMB, HTTP, S3 • Data abstraction services including cloud and non-IBM storage and muti-site asynchronous and synchronous replication • Data management services with integrated lifecycle management to optimize data from memory, NVMe flash, HDD, public cloud, external storage and tape • Data resilience services with FIPS 140-2, 256-bit encryption, cyber-protect and IBM Safeguarded Copy • Data Acceleration Tier (DAT) for NVMe-oF–based performance optimization in AI and analytics workloads
Performance	• Dual socket AMD EPYC™ Genoa • Sequential read performance up to 340 GB/s and 28 million IOPS[2](NVMe-oF configurations)[2]
Networking / Adapters	• NDR / HDR switch support • 16 x PCIe Gen5 slots per system (8 slots per canister) • Up to 12 24G SAS adapters (6 per canister) • NVIDIA CX7 supported adapters: o 400Gb single port (InfiniBand only) x16 Gen5 o 200Gb VPI dual port (InfiniBand/Ethernet) x16 Gen5 o 8 Max High Speed Network Adapters (CX7) o Up to 16 network ports o Certified for NVIDIA BasePOD, SuperPOD and Grace Blackwell platforms
Drive Support	• 24 or 48 U.2 NVMe SSDs (30.72TB) • 24 or 48 U.2 NVMe FCM (19.2TB or 38.4TB)
Environmental – Drives	• Nominal power: 20-25 watts per SSD • Nominal power: 30 watts per FCM
Environmental – Controller	• Input voltage: 200-240V 50/60 Hz • Nominal power: 1,500 W (empty); 4800 W (PSU max) • Power supplies: 4 hot swappable, redundant
Size – Controller	• 4RU; H:7” (175 mm), W: 19” rack (483 mm), D: 33.5” (850 mm) without bezel
Size - Enclosure	• 4RU; H:7” (175 mm), W: 17.5” (445 mm), W: 19.0” rack (483 mm), D: 39” (1000 mm), L: 36” (926 mm) chassis length

[1] Raw capacity per 4U node up to 2.2 PB; effective capacity based on 3:1 inline compression. Compressibility depends on the characteristics of the stored data.

[2] Disclaimer: Performance metrics were obtained by running sample programs in a controlled environment with standard hardware/software configurations and testing procedures. Since performance varies with configuration, program characteristics, and other installation and environment factors, results obtained in other operating environments may vary. IBM® does not represent, warrant, or guarantee that a user will achieve the same or similar results in the user's environment.

IBM Storage Scale System 3500

IBM Storage Scale System 3500 is for customers requiring an entry-level or mid-level system for less extreme product environments. It has many of the same capabilities as the Storage Scale System 6000 and is also capable of high capacity and performance.

To learn more, download the Storage Scale System 3500 data sheet.

Parameter	Storage Scale System 3500	Storage Scale System 6000
Size	2 rack units (2U)	4 rack units (4U)
Maximum capacity	24 x 30.72TB NVMe	48 x 30.72TB NVMe 48 x 38TB FCM4 modules
Maximum throughput	126GB/s	330GB/s
Expansion	Up to 4 direct-attached JBODs	Up to 9 direct-attached JBODs
Data transfer	12Gb SAS	24Gb SAS

IBM Scale System All-Flash Expansion Enclosure

Parameter	
Size	2 rack units (2U)
Maximum capacity	3.2PB per enclosure (using 122TB QLC)
Drive configurations	30TB, 60TB and 122TB QLC
Expansion	Up to 4 direct-attached expansion enclosures per rack
Networking	CX-7 200 GbE/NDR (Scale System 6000 server side)

For more information

To learn more about IBM Storage Scale System, contact your IBM representative or IBM Business Partner, or visit ibm.com/products/storage-scale-system.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
December 2025