IBM Storage for Data & AI
Solution Guide

Accelerating AI with IBM Storage Scale and NVIDIA

An integrated solution for developing and deploying enterprise AI applications at scale

Highlights

IBM Storage Scale integrates with NVIDIA technologies to help enterprises deploy AI applications at scale

Major use cases include generative AI, computer vision, financial modeling, healthcare, and simulation

Key integration areas include GPUs, networking, and an end-to-end software stack

Content-aware IBM Storage Scale leverages NVIDIA NeMo™ Retriever to extract the meaning inside unstructured data sources

Business leaders are challenged today by how to extract maximum value from AI and high-performance computing applications as they scale from initial implementation to widespread deployment. These resource-intensive workloads require maximum performance from every component – compute, networking, storage, and software.

IBM Storage and NVIDIA work together to help customers accelerate time-to-value as they deploy enterprise-scale AI workloads across their data centers, cloud platforms, and hybrid IT environments:

– As an innovator in graphics processing unit (GPU) technology, NVIDIA has emerged as a leader in AI because deep learning relies on matrix and tensor computations, which GPUs handle far more efficiently than traditional CPUs. The NVIDIA AI Data Platform is a customizable reference design that accelerates data preparation, processing, and analytics to power machine learning, deep learning, and real-time AI applications.

– IBM Storage Scale is software-defined file and object storage that provides a global data platform with unique data abstraction and content-aware storage capabilities. IBM Storage Scale System 6000 is a hardware implementation of Storage Scale that is optimized for the most data-intensive workloads.

Challenging Use Cases

Storage Scale and NVIDIA accelerated computing are often deployed together in high-value AI use cases that include:

- Generative AI & large language models (LLMs);
- Computer vision & autonomous systems;
- Financial modeling & risk analysis;
- Healthcare & drug discovery;
- Retail AI (fraud detection, recommendation systems).

Although they have very different applications, these AI use cases share many core requirements. All depend on high-performance computing – GPUs or specialized accelerators – to handle the demands of training and inference. These applications all depend on huge volumes of text, images, structured data, or sensor inputs.

To support this, other essentials include robust data storage systems and high-bandwidth, low-latency data pipelines. Scalability is another shared requirement, enabling systems to adapt to growing data volumes and model complexity. Finally, strong security, privacy, and compliance frameworks are necessary across the board – especially in regulated industries like finance and healthcare.

But these use cases also have unique requirements:

- Generative AI and LLMs require large-scale GPU clusters and advanced distributed training frameworks.
- Computer vision and autonomous systems typically require real-time inference at the edge, integrating high-bandwidth sensor data from cameras and LiDAR.
- Financial modeling and risk analysis emphasize low-latency execution, deterministic performance, and strict auditability.

Content-aware Storage

As enterprises scale their AI initiatives and workloads shift from training to inferencing, many are discovering that public LLMs often fall short when applied to internal use cases. These models were typically trained on publicly available datasets and lack access to the proprietary documents – such as PDFs, presentations, and reports – that hold critical organizational knowledge.

Storage Scale addresses this challenge with built-in content-aware capabilities that make enterprise data more accessible for AI by embedding vector search, metadata enrichment, and semantic understanding directly into the storage layer.

Graphics Processing Units

GPUs are crucial in AI and HPC applications because they accelerate parallel processing, efficiently handling thousands of simultaneous computations. In AI, they accelerate deep learning by performing matrix multiplications, convolutions, and tensor operations essential for training neural networks.

In high-performance computing, GPUs excel at numerical simulations, molecular dynamics, and scientific computing by rapidly executing floating-point operations and linear algebra calculations. Their architecture, optimized for massively parallel workloads, makes them indispensable for processing large datasets and complex algorithms far more efficiently than traditional CPUs.

Figure 1 – NVIDIA DGX SuperPOD™ is a scalable AI computing cluster built from interconnected DGX systems, optimized for training and deploying large AI models.

NVIDIA GPUs

NVIDIA's enterprise GPU lineup centers is built on the NVIDIA Blackwell architecture. The Blackwell GPU architecture introduces significant enhancements in performance and efficiency. Blackwell GPUs feature up to 208 billion transistors and incorporate fifth-generation Tensor Cores, supporting sub-8-bit data types for improved AI computation.

The architecture also includes NVIDIA NVLink™ 5.0 for high-speed GPU interconnects. NVIDIA H100 and B100 GPUs provide the compute backbone for NVIDIA DGX BasePOD™ and NVIDIA DGX SuperPOD™, which are scalable AI infrastructure solutions designed for AI training, inference, and HPC workloads:

- DGX BasePOD is a reference architecture for deploying clusters of DGX H100 systems, optimized for enterprises running AI workloads at scale.
- DGX SuperPOD is an AI supercomputer composed of 32 or more DGX H100 nodes, interconnected with NVIDIA Quantum-2 InfiniBand and NVIDIA AI Enterprise software, delivering exascale AI performance for training massive models.

The Blackwell-powered B200 and GB200 (a Grace Blackwell superchip combining the B200 GPU with a Grace CPU) now form the foundation of next-generation AI factories—massive-scale compute clusters optimized for trillion-parameter foundation models and generative AI workloads. These systems leverage the NVIDIA GB200 NVL72, which integrates 72 Blackwell GPUs and 36 Grace CPUs in a single liquid-cooled rack, interconnected by the NVLink switch system.

IBM Storage Scale

Storage Scale is software-defined storage that organizations deploy to build a global AI / HPC / analytics data platform. It provides distributed file and object storage for both structured data, like databases, and the unstructured data created in GenAI / AI / ML workloads, analytics, data lakes, IoT, cloud-native applications, and backup and archive applications.

Storage Scale is based on a massively parallel file system and can be deployed on multiple hardware platforms including x86, IBM Power, IBM zSystem mainframes, ARM-based POSIX client, virtual machines, and Kubernetes. It provides global data abstraction services to connect seamlessly with multiple data sources and multiple locations, bringing together data from IBM and third-party storage environments.

This data storage architecture plays a critical role during AI model training, which can take days or even months. If a run is interrupted by a power outage, hardware failure, or other issue, the entire process might have to restart from the beginning. To avoid this, training periodically pauses to save a checkpoint – a full snapshot of the model's internal state, including weights, learning rate, and other variables.

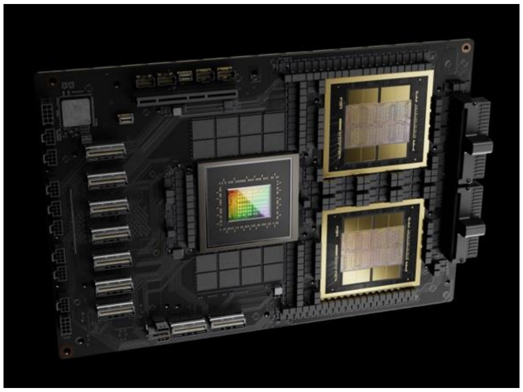

Figure 2 – The NVIDIA GB200 superchip is a multi-die processor that integrates the NVIDIA Grace CPU with the B200 GPU, connected using high-bandwidth NVLink.

Checkpointing provides fault tolerance, but it's a synchronous step that temporarily halts training. As models grow larger, so do their checkpoints – reaching 14TB for a trillion-parameter model. Storage Scale addresses this by using a POSIX-compliant file system optimized for high-throughput, multi-threaded I/O across many nodes. Acting as a cache between GPUs and object storage, Storage Scale's active file management (AFM) accelerates both model startup and checkpointing. Data loads into GPUs more quickly at job start or restart, and checkpoints are saved faster to the file system. AFM then pushes that data asynchronously to object storage, so training can continue without delay.

Content-aware Storage Scale

Storage Scale includes built-in content-aware capabilities that allow AI pipelines to extract meaning from enterprise data – such as PDFs, presentations, and other unstructured sources – without external indexing or data preparation. Using embedded natural language processing and vector embedding, Storage Scale helps AI systems interpret both the content and context of stored information.

Figure 3 – The IBM Storage Scale 6000 is a certified storage solution for NVIDIA DGX BasePOD and DGX SuperPOD.

These capabilities are implemented directly within the storage layer, allowing organizations to build intelligent pipelines without rearchitecting how or where data is stored. Using AFM, Storage Scale enables high-performance, content-aware access to remote or distributed data sources without the need to copy, migrate, or duplicate the data. AI workloads can operate on live datasets in place – regardless of location – reducing both latency and operational overhead.

IBM Storage Scale System 6000

IBM Storage Scale System 6000 is a hardware implementation of Storage Scale software that is optimized for the most data-intensive workloads and is a certified storage solution for NVIDIA DGX BasePOD and DGX SuperPOD.

It is available in all-flash and hybrid configurations providing:

- Up to 330 gigabytes per second (GB/S) throughput with low latency;
- Up to 13 million IOPS using NVMeoF;
- Each Scale System 6000 supports up to 2.2PB raw capacity in a 4U rack space.

Storage Scale System 6000 supports the NVIDIA GPUdirect Storage protocol, which enables a direct data path between GPU memory and local or remote storage, such as NVMe or NVMe over Fabric (NVMe-oF). The GPUDirect storage optimization removes the host server CPU and DRAM from the data path, so the IO path between storage and the GPU is shorter and faster, helping ensure that fast GPUs aren't being starved by slow IO.

To help customers accelerate their AI deployments, IBM and NVIDIA provide certified reference architectures for the IBM Scale System 6000 with NVIDIA DGX BasePOD and NVIDIA DGX SuperPOD AI infrastructure, including the NVIDIA DGX A100, H100, H200, and the B200 that incorporates NVIDIA's new Blackwell GPUs. NVIDIA and IBM jointly test, plan, and install the systems, with the storage backed by IBM global deployment and support services.

Networking

NVIDIA provides high-performance networking platforms that include NVIDIA Quantum InfiniBand, NVIDIA Spectrum-X Ethernet, and NVIDIA BlueField DPUs and SuperNICs. Together these enhance Storage Scale and Scale System 6000 by improving data transfer efficiency, reducing CPU bottlenecks, and enabling high-speed, scalable AI and HPC workloads.

NVIDIA Quantum Infiniband is a high-bandwidth, low-latency networking solution designed for AI, supercomputing, and large-scale storage environments. The latest platform, Quantum-X800 InfiniBand, features speeds up to 800Gb/s per link, adaptive routing to avoid congestion and improve data flow efficiency, and Remote Direct Memory Access (RDMA) to bypass CPU overhead in storage and computing tasks.

NVIDIA Spectrum-X is the world's first Ethernet platform purpose-built for AI training and inference – it combines NVIDIA Spectrum-4 Ethernet switches with NVIDIA BlueField™-3 and NVIDIA ConnectX-8 SuperNICs for AI, machine learning, and natural language processing applications.

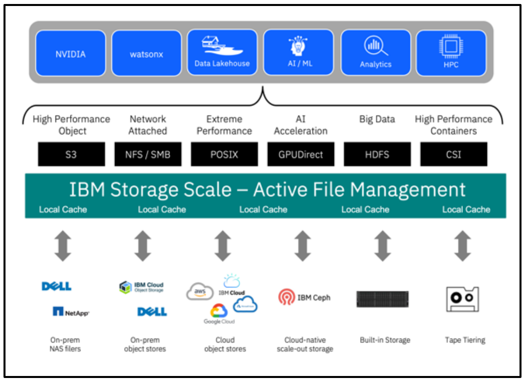

Figure 4 – IBM Storage Scale has unique virtualization capabilities that make it an excellent global data platform for structured and unstructured data, whether on-premises, at the edge, or in the cloud.

IBM Storage Scale and Scale 6000 use NVIDIA Spectrum-X to boost data access speed, storage performance, and GPU efficiency:

- Fast AI data access: Spectrum-X's 400GbE fabric enables high-throughput retrieval from Storage Scale nodes to GPUs.
- Fewer training bottlenecks: Optimizes Ethernet storage access, narrowing the gap with InfiniBand.
- Low-latency NVMe access: RoCEv2 enables direct, fast NVMe storage access in Scale 6000, cutting CPU use and speeding data pipelines.
- Efficient multi-GPU training: Delivers scalable, smooth data flow to large GPU clusters over Ethernet.
- BlueField-3 DPU acceleration: Offloads storage and security tasks, reducing CPU load and enabling hardware-accelerated encryption and access control.

Data Processing Units

NVIDIA has also innovated in the development of DPUs (data processing units), which are designed to offload and accelerate data-intensive tasks like networking, security, and storage, freeing up CPU and GPU resources for other tasks. Storage Scale supports NVIDIA BlueField-3 DPUs.

NVIDIA BlueField-3 DPUs are capable of delivering:

- Lower CPU load: Offloads packet processing, storage, and security tasks.
- Faster data flow: Optimizes transfers between storage, CPU, and GPU.
- GPUDirect Storage: Enables direct storage-to-GPU data movement.
- Built-in security: Zero-trust encryption, access control, and edge firewalling.
- Cloud-ready: Works with VMware, Kubernetes, and OpenStack.

Integrated Software Stack

One reason for NVIDIA's success is a software ecosystem that ensures seamless integration across development, training, and deployment stages. The core of this strategy is the NVIDIA AI Data Platform.

NVIDIA AI Data Platform

The NVIDIA AI Data Platform is a customizable reference design for a new class of AI infrastructure that brings enterprise storage into the era of AI. The NVIDIA AI Data Platform complements IBM Storage Scale by ensuring that data flows seamlessly from distributed sources to the compute layer. As AI models grow larger and more complex, this platform helps enterprises orchestrate data-intensive pipelines like RAG and multi-modal inferencing, without sacrificing performance or scalability.

Storage Scale's integration with key NVIDIA software tools, including GPUDirect Storage and Magnum™ IO, enables faster I/O, checkpointing, and training cycles, while NVIDIA AI Enterprise ensures compatibility and performance optimization.

Other key components of the NVIDIA software stack include:

- NVIDIA NeMo Retriever – a a collection of extraction, embedding, and reranking microservices for building RAG pipelines, central to Storage Scale's content-aware capabilities;
- NVIDIA AI Enterprise – a cloud-native suite of software tools, libraries, and frameworks, including NVIDIA NIM and NeMo microservices, that accelerate and simplify the development, deployment, and scaling of AI applications;
- GPUDirect® – tools for high-speed communication between GPUs, CPUs, and storage systems;
- Magnum IO – software to accelerate I/O in multi-GPU and multi-node computing environments.

NVIDIA NeMo Retriever and Content-Aware Storage

NVIDIA NeMo Retriever is a GPU-accelerated microservice that performs fast, semantic search across vector embeddings to retrieve contextually relevant data for RAG and other generative AI applications. The content-aware capabilities in Storage Scale are enhanced by direct integration with NeMo Retriever, which it leverages to help ingest and store unstructured enterprise data. This data is then processed into vector embeddings (representations of semantic meaning), which are stored alongside the original content in an adjacent vector database.

When a generative AI application sends a query, NeMo Retriever performs a fast similarity search against these embeddings to identify the most relevant content. Storage Scale then serves that data – via its high-throughput, parallel architecture – directly to the inference engine, typically a large language model. By embedding this retrieval mechanism into the storage layer, Storage Scale reduces latency, offloads preprocessing from the application layer, and ensures that AI models receive high-quality, context-aware data – improving accuracy, scalability, and response time for enterprise RAG applications.

NVIDIA AI Enterprise

NVIDIA AI Enterprise is a cloud-native software suite designed to optimize, deploy, and manage AI, ML, and deep learning workloads on NVIDIA GPU-accelerated infrastructure. It provides enterprise-grade AI frameworks, tools, and support to run AI applications efficiently in on-prem, cloud, and hybrid environments. AI Enterprise integrates with content-aware Storage Scale to support continuous, fine-grained vector updates using NVIDIA NeMo Retriever and NVIDIA NIM, improving inferencing precision and enabling near real-time RAG pipelines.

Key features include optimized AI frameworks for TensorFlow, PyTorch, RAPIDS, and other AI/ML/DL libraries, and access to NVIDIA® NGC™ with pre-trained models for computer vision, NLP, speech AI, and recommender systems. AI Enterprise is certified for VMware, Red Hat OpenShift, Kubernetes, and major cloud providers (AWS, Azure, and GCP), and includes NVIDIA Triton Inference Server for scalable AI model deployment and NVIDIA Fleet Command for edge AI model management.

Storage Scale software leverages AI Enterprise to efficiently feed large AI datasets into NVIDIA accelerated computing as part of end-to-end AI data pipelines that encompass model training, validation, and inferencing.

NVIDIA Magnum IO™ GPUDirect® Storage

GPUDirect is a family of NVIDIA technologies that enable direct, high-speed data transfer between GPUs and other system components such as network adapters, storage devices, and other GPUs, bypassing the CPU to reduce overhead and latency. It includes GPUDirect Peer-to-Peer for fast communication between GPUs in a system, NVIDIA Magnum IO™ GPUDirect® RDMA for direct GPU-to-GPU communication across nodes, and GPUDirect Storage for efficient data movement between GPUs and storage devices like NVMe and parallel file systems.

In content-aware storage workflows, GPUDirect Storage enables GPUs to directly access up-to-date vector embeddings stored adjacent to original data, eliminating the need for intermediate I/O steps and improving model responsiveness. By minimizing CPU involvement, GPUDirect significantly boosts performance in AI training, HPC simulations, and large-scale data processing, making CUDA® applications more efficient.

NVIDIA Magnum IO is the architecture for parallel, intelligent data center IO. It maximizes storage, network, and multi-node, multi-GPU communications for mission-critical applications, using large language models, recommender systems, imaging, simulation, and scientific research.

NVIDIA GPUDirect Storage (GDS) facilitates IO transfers directly to the GPU memory, removing the expensive data path bottlenecks to and from the CPU/system memory. Avoids the latency overhead of an extra copy through system memory, which impacts smaller transfers and relieves the CPU utilization bottleneck by operating with greater independence.

NVIDIA GPUDirect RDMA (GDR) provides access for the network adapter to read or write memory data buffers directly in peer devices. It allows RDMA-based applications to use the peer device computing power without the need to copy data through the host memory.

IBM supports Magnum IO in Storage Scale, delivering increased performance for workloads including deep learning, AI model training, scientific simulations, and advanced analytics.

Conclusion

IBM Storage Scale and NVIDIA's full-stack AI platform combine to deliver a powerful, scalable, and efficient infrastructure for enterprise AI workloads. By integrating high-performance GPUs, intelligent storage, accelerated networking, and a robust software ecosystem, this joint solution addresses the end-to-end demands of modern AI – from data ingestion and training to inferencing and real-time analytics.

Whether you're building foundation models, deploying generative AI at scale, or optimizing high-performance computing applications, IBM and NVIDIA provide the performance, scalability, and integration needed to unlock the full potential of AI in the enterprise.

For more information

To learn more about IBM Storage Scale solutions, contact your IBM representative or IBM Business Partner, or visit go to:

- IBM Storage Scale
- IBM Storage Scale System 6000
- IBM NVIDIA AI Storage Solutions
- Solution Brief: Content-Aware IBM Storage Scale
- Solution Brief: Accelerating AI Training with IBM Storage Scale & Scale System 6000
- Reference architecture: NVIDIA DGX SuperPOD with IBM Storage Scale and IBM Storage Scale System 6000
- Reference architecture: NVIDIA DGX BasePOD with IBM Storage Scale and IBM Storage Scale System 6000
- IBM Redpaper: IBM Storage Scale System Introduction Guide

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
July 2025

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.