FUTURIOM
FUTURE OF CLOUD TECH

Special Report

# Object Storage in the AI Era: Emerging Trends and Players

February 2025

**Sponsored by:**

wasabi
hot cloud storage

# Storing the World's Data™

Wasabi Hot Cloud Storage gives you the freedom to store more and do more with your data thanks to unbeatable pricing, rock-solid protection, and blazing-fast performance.

**Learn more at Wasabi.com**

## Price

Up to 80% less than the hyperscalers

No fees for egress or API requests

## Performance

Faster than the competition

Quick uploads & downloads

Private network options

## Protection

Data center redundancy

11x9s data durability

Immutability & Multi-User Authentication

wasabi®
hot cloud storage

# Highlights

- **AI is changing the game plan for object storage providers.** As inferencing takes hold in enterprise environments, vendors with longstanding expertise in object storage are expanding their platforms to address AI workloads.
- **The rise of object storage has invigorated global file systems.** Vendors providing systems that furnish a common namespace for files and objects stored in various media and distributed geographic locations have also reached a turning point toward AI.
- **Unstructured data is emerging as key to AI applications.** By most accounts, unstructured data—including text file, videos, audio clips, images, and other unformatted data—accounts for 80% of corporate data, comprising a rich source of input to AI models.
- **Unstructured data is held captive in enterprise silos.** Unformatted data has often been housed in servers and on network attached storage (NAS), making it hard to retrieve for use in AI.
- **Object storage is ideal for unstructured data.** By storing massive amounts of data in a flat file format requiring no hierarchical directory system, object storage is an ideal medium for storing unstructured data.
- **Object storage services are on the rise.** Providers of content delivery networks (CDNs) as well as independent managed service providers are offering object storage services designed to compete against offerings from the leading public cloud hyperscalers.
- **AWS S3 is pervasive.** AWS set the bar for cloud-native object storage nearly two decades ago, and its Simple Storage Service (S3) continues to govern the format of object storage for providers throughout the market, which universally offer S3-compatible REST APIs for use with their platforms.
- **Object storage must be adapted for use in AI**. While object storage offers unlimited scalability, issues of networking, throughput, load balancing, and adaptability to AI workloads must be addressed in order for market needs to be adequately met.
- **Companies covered in this report:** Akamai, Alibaba, AWS, Backblaze, Cloudflare, Cloudian, CTERA, DataCore, DDN, Deft, Dell, DigitalOcean, Fastly, Fujitsu, Gcore, Google Cloud, Hammerspace, Hetzner, Hewlett Packard Enterprise, Hitachi Vantara, IBM, Komprise, Leaseweb, Lenovo, MinIO, Microsoft, Nasuni, NetApp, Nutanix, Oracle, OSNexus, Panzura, Pure Storage, Quantum, Qumulo, Scality, Seagate, Storj, Supermicro, VAST Data, Veeam, Wasabi, WEKA, Zadara

# Table of Contents

# 1. Introduction: AI and Object Storage

The artificial intelligence (AI) revolution changing how we store enterprise data. As firms move to AI inferencing, the process of running corporate data against pretrained models, data storage is being redesigned to accommodate unprecedented performance and scaling requirements.

Inferencing requires good data. Without it, language models large and small remain irrelevant for enterprise purposes. This need for data extends not only to data formatted for use in databases, but also to unstructured data, which by most accounts comprises about 80% (or more) of enterprise data. This is data that's increasingly valuable to enterprises because it holds unique information in the form of email, photos, video, audio clips, messages, social media posts, and the like. The issue is getting it all into a single location, ready and secured for analytics, business intelligence, and use in AI applications.

Here's the challenge: Much unstructured data has been held as files in servers and in network attached storage (NAS) devices that often are siloed within widely scattered departments or divisions. In these environments, data remains stranded and difficult to incorporate into inferencing or analytics workloads. The need to organize data for these workloads has led to demand for distributed storage that provides easy and secure access to specific data items within largescale data lakes.

Key to this approach is object storage, a format that breaks unstructured data into units called objects, each of which contains metadata and a unique identifier. The metadata allows for easy management and the unique identifiers aid information retrieval, typically via REST (Representational State Transfer) APIs, which support HTTP or HTTPS requests.

Object storage allows massive amounts of data to exist in a flat file format wherein all units are equal and no hierarchical directory system is needed. It can be deployed on premises with solid-state drives or flash array hardware or in a cloud-native format that works with public or private clouds. Hybrid environments are increasingly popular.

Cloud-native object storage is not new: AWS first introduced it with its Simple Storage Service (S3) in 2006. Over time, other clouds, including Microsoft Azure and Google Cloud, released their own cloud object storage services. That said, hyperscaler charges to move data, deploy APIs, and add resources can be expensive. Alternative object storage products and services compatible with S3 and other formats are available from a growing number of companies.

In this report, we take a close look at these offerings, exploring the technologies, trends, and major players and looking at the relationship of object storage to AI application deployment. Read on and enjoy!

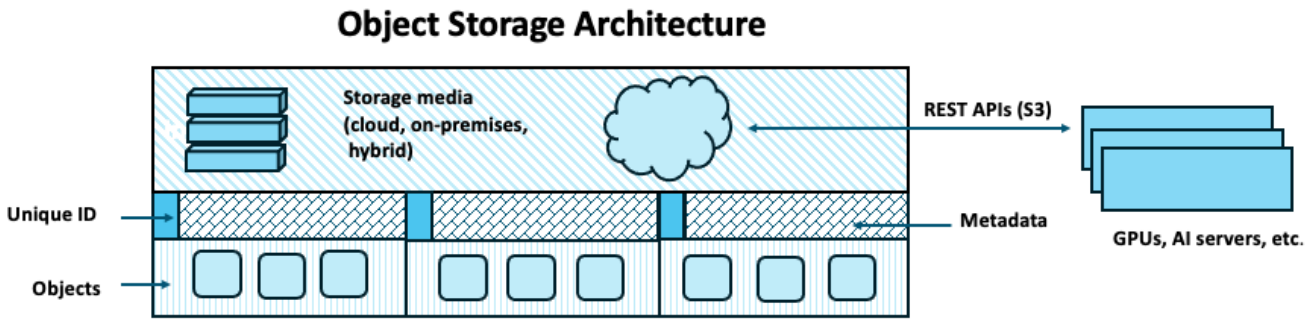
# 2. Object Storage: Definition and Overview

Let's start our overview of object storage for AI with some basic definitions. First off, object storage is part of a spectrum of storage formats for data, which includes block and file storage. Below is a comparison of the various choices:

**Block storage.** Primarily used in transaction processing that requires low latency, block storage divides data into blocks or units of a specific size that carry unique identifiers and can be stored on separate systems. When data is accessed, those identifiers are gathered for delivery of the data. Since block storage uses fixed-sized blocks, it isn't suited to handling unstructured data that comes in many sizes. Block storage also doesn't have too much metadata (on purpose, to boost performance), which limits its ability to search unstructured data. Block storage is associated with storage area networking (SAN), which is used in high-end transaction processing of database and virtual machine (VM) file systems.

**File storage.** This type of storage, as the name implies, relies on a hierarchical structure of files, folders, and directories to store and manage unstructured data, such as .doc and .jpeg files. It uses protocols such as Server Message Block (SMS) and Network File System (NFS), typically implemented on network-attached storage (NAS) hardware. Many object storage systems also store files.

**Object storage.** As noted earlier, object storage divides data into units with unique identifiers and metadata information for each, making it ideal for retrieving data that lacks hierarchical classification. It is also a good fit for distributed architectures and can unify data from multiple systems, on-premises or in clouds.




## Why Object Storage Now

As enterprises move to incorporate AI into their organizations, the advantages of object storage become plain. First, object storage is scalable, thanks to its flat architecture, which moves data in small

units without the potential holdup of a file or directory hierarchy. This means large amounts of data, even to the exabyte scale, can be easily moved and accessed—which is vital to AI model training. While object storage still faces issues of distribution and load balancing, as well as networking, memory, and other functions, the nature of object-based data makes storing it impervious to size. Cloud-native object storage based on microservices is also highly reliable.

Object storage is key to using unstructured data in AI workloads—something that has become essential as enterprises shift to inferencing. Often an organization's text files, audio and video clips, receipts, documents, medical images, and even social media posts all contain valuable information for use in Retrieval Augmented Generation (RAG) or model fine tuning. What's more, object storage's structure allows data to be immediately searchable, making it useful for ingestion in AI workloads requiring current data. The metadata associated with object storage also allows data to be organized according to specific characteristics or informational items. This is useful in AI as well as in analytics.

Security is another factor favoring object storage. Encryption at the object level allows for tighter control of specific data items. Objects storage can be made immutable. Compliance with geographical regulations is possible thanks to the distributed nature of object storage.

Object storage is also more resilient because it can replicate data across multiple distributed devices, making disaster recovery faster and easier. Further, object storage is adaptable to the parallel processing infrastructure of accelerated computing GPUs because it breaks data into small increments that can be fed individually into GPUs operating in parallel. Still, the object storage system must be architected for this kind of optimal throughput, a topic we deal with in the next section.

## The Role of S3

As noted earlier, all object storage systems on the market today are compatible with AWS's Simple Storage Service (S3) API, which set the bar for cloud-native object storage nearly two decades ago. One estimate is that of approximately trillions of objects stored across AWS, Azure, and GCP, roughly 60% are on AWS. Via REST APIs, companies can link to AWS S3 data in object storage, making a system compatible with a vast ecosystem of applications, APIs, and other object storage systems outside of AWS.

Vendors vary in their support of S3. Some are compatible with all elements of the S3 core API, while others select only what they deem useful to their particular platform, because some S3 functionality is pertinent only to instances in AWS. Some vendors base their S3 compatibility on the need to interact with applications such as Dremio, Trino, Hive, and MapReduce, which store and/or query "big data."

# 3. Object Storage Systems and Services

The object storage market is complex and multilayered. There are many vehicles for delivering object storage, and the technical details can be perplexing for customers. There is plenty of room for vague claims and doublespeak. Further, there is so much variety among providers that any attempt to break the market down will inevitably leave out one or more viable alternatives. But let's try nevertheless to generalize about the various segments.

## Object Storage Services

All the major public cloud hyperscalers offer object storage services: **AWS** (S3), **Microsoft** (Azure Blob Storage), **Google Cloud** (Google Cloud Storage), **Oracle** (Oracle Cloud Infrastructure Object Storage), **IBM** (IBM Cloud Object Storage), and **Alibaba** (Object Storage Service). Most of these cloud vendors also provide object storage for the content delivery networks (CDNs) associated with their services.

Independent CDN companies with extensive international networks are using object storage to deliver video and other unstructured content to enterprise customers. **Akamai,** for instance, can retrieve data from the object storage service it acquired with Linode in 2022, then cache it on Akamai Connected Cloud service and deliver it to the end customer at an edge location in the network. Other CDNs, including **Cloudflare, DigitalOcean, Fastly, Gcore, Leaseweb**, and others also offer object storage services to serve customers that have high volumes of unstructured data to deliver over their networks.

Often, these object storage services come from integration with the likes of **Wasabi**, an object storage provider that supplies managed service providers (MSPs) with object storage from its own network of multiple datacenters worldwide. Wasabi's object storage infrastructure-as-a-service (IaaS) serves 90,000 customers and tens of thousands of channel partners. The vendor maintains that its purpose-built storage network is not just a solid origin server for CDNs but a viable alternative to the services offered by the public cloud hyperscalers due to reduced cost and added security features (more on that in a moment). Other service providers in this space, including **Backblaze, Deft, Hetzner, Seagate, Storj,** and **Zadara,** also offer object storage at considerably lower cost than the major cloud hyperscalers.

There are a couple of caveats. Users have noted that storage services often perform only as well as the WAN linking the customer site to the Internet. That can be a bottleneck. Also, independent storage services don't have compute instances on offer like the hyperscalers do. This means the cloud vendors can offer some options in the cloud for customers using their object storage services, which the independent service providers may offer only via partnerships.

## Global File Systems

A separate product category in object storage includes file systems that are designed to act as front ends for object storage systems, including those from the leading hyperscalers as well as from independent vendors (listed in the next section). In this group are a series of companies with longstanding data management credentials, which are now adapting their products to fit the AI era. Following is a quick look at three leading providers:

**CTERA.** Headquartered in Israel with offices in Europe, the U.S., and Singapore, CTERA, which has over 150 employees, was founded in 2008 to deliver a software-defined file system that delivers a single namespace for all files and objects, regardless of geographical distribution or storage environments. CTERA's platform includes CTERA Portal, which gathers audit logs and file system activity data from CTERA Edge Filers and CTERA Drive, products that extend the vendor's file system to remote edge locations and mobile devices, respectively.

The vendor recently extended its platform to link unstructured enterprise data to specific AI models and applications. Called CTERA Data Intelligence, the new addition, still in preview as of this writing, continuously retrieves metadata from a range of object file stores, indexes it, and links it to third-party AI models from the likes of OpenAI, Anthropic, or Microsoft. As a bonus, CTERA has added specialized AI agents for information retrieval from those models. Another new product, CTERA Insight, allows storage administrators to dig into details of how stored files are being used. This product draws on data from CTERA Portal, CTERA Edge Filers, and CTERA Drive.

On July 17, 2024, CTERA announced a growth investment of $80 million from PSG Equity. Besides PSG, investors include Bessemer, Benchmark, and Red Dot Capital. Customers include The Carlyle Group, McDonalds, Humana, Santander, and the U.S. Veterans Administration.

**Nasuni.** The company was founded in 2009 by Andres Rodriguez (now CTO) and offers the Nasuni File Data Platform, a system that purports to offer unlimited object-based file storage in a cloud environment, freeing data from NAS silos. The system supports anomaly detection and ransomware protection. A Nasuni IQ product gathers usage statistics and system information for chargeback reporting, capacity planning, data resource optimization, and other functions, including preparation for AI workloads.

Nasuni has 550 employees and operates in the U.S. (with headquarters in Boston) and Europe, with presence in Australia, New Zealand, and Singapore. On July 9, 2024, Nasuni revealed a majority equity stake led by Vista Equity Partners that valued the company at $1.2 billion. Customers include Perkins+Will, Western Digital, GeoVerra, and TBWA.

**Panzura.** Founded in 2008 by Randy Chou (ex-Aruba, now CEO of Nubeva) and John Taylor (ex-Aruba), Panzura specializes in global file management for hybrid cloud. It is headquartered in Dallas and has about 200 employees. The company has 37 patents and has raised over $80 million from investors that include Profile Capital Management, Kayne Partners, and CIBC Innovation Banking. Customers include Disney, Milwaukee Tool, BBC America, and AmeriHome Mortgage.

Panzura's hybrid cloud file services platform, CloudFS, provides a single global namespace, enabling delivery of data without making file copies. It consolidates data from multiple data sources, eliminating data silos. Deduplicating and compressing data enable savings on data volumes of up to 80%, the vendor says. Disaster recovery is facilitated with immutable data and immutable global snapshots that enable a global recovery point objective (RPO) of less than a minute.

In September 2024, Panzura released a platform called Symphony that simplifies data management, including establishing AI pipelines, for a range of third-party cloud and on-premises file and object storage systems, including Cloud FS as well as ones **from AWS, Microsoft Azure, Google Cloud Platform, NetApp,** and **Wasabi**. The user interface is also integrated with **IBM's** Fusion turnkey hyperconverged infrastructure (HCI) system, which supports a range of storage and data management options. Symphony is the data services platform Panzura acquired with its purchase of Moonwalk Universal earlier this year (terms not disclosed).

## Global File Index

One company claiming to have an alternative to global file systems is **Komprise**, a ten-year-old firm based in Campbell, Calif., that deploys a metadata catalog called a Global File Index to manage unstructured object and file data regardless of the type of underlying storage and its location.

"We add a layer to the side of storage," said Krishna Subramanian, president, COO, and cofounder of Komprise, in an interview with Futuriom. She explained that this distinguishes Komprise from global file systems, which create a new data access layer. She also added: "We extend any file system namespace, thus preserving the performance of your file systems and enabling vendor choice."

## Object Storage Systems

Object storage systems are available from a growing array of vendors, including network attached storage (NAS) providers such as **Dell, Fujitsu, Hewlett Packard Enterprise (HPE), Hitachi Vantara, IBM,** and **NetApp,** which now offer software- and service-based object storage, along with options for various hardware configurations—and combinations of all these choices. **HPE**, for instance, offers the HPE X10000 software-defined object storage solution as part of its Alletra platform, which deploys flash-based hardware appliances to streamline operations.

Object storage solutions are often touted as replacements for NAS, though vendors such as **NetApp,** the leading provider of NAS, offer object storage software for their appliances. NetApp customers, for instance, can adopt the vendor's ONTAP object storage software to work with smaller object stores on NetApp NAS products, graduating to large-scale NetApp StorageGRID software as storage volumes grow. StorageGRID can be implemented with hardware or within virtual machines or containers in Kubernetes environments.

Overall, object storage system suppliers are proliferating, as demands for scalable data storage increase with the role of AI. To meet demand, traditional storage vendors offer their own takes on object storage, older suppliers are updating their wares, and startups are entering the market. The features and functions of the systems offered are also varied, though there are several "table stakes" features that are offered across the board. These include the following:

**S3 compatibility.** All of today's leading object storage systems and services support the AWS S3. All of them typically offer Representational State Transfer (REST) APIs to interact with S3 objects. A few systems, including **Cloudian** HyperStore, **DataCore** Swarm, **MinIO** AIStor, **Nutanix** Objects Storage, **Qumulo** Cloud Native Qumulo, and **Scality** RING also support Azure Blob storage. **Cloudian**, **MinIO**, and **Scality** also support Google Cloud Storage.

**On-premises orientation.** Most object storage systems operate on premises in tandem with proprietary or standard hardware and storage media. An exception is **Qumulo's** Cloud Native Qumulo, which operates within instances on AWS or Azure. Several other vendors, including **Cloudian, HPE, Hitachi Vantara, IBM, MinIO, NetApp, Scality, VAST Data,** and **WEKA,** support hybrid cloud implementations, meaning they work with a cloud-based version of their software or within public cloud environments as well as with on-premises equipment. Public cloud storage is universally supported by object storage providers for backup or resiliency, with AWS and Microsoft Azure the leading preferred environments.

**Kubernetes support.** Support for Kubernetes microservices management has become common across object storage systems, with rivals **Cloudian** and **MinIO** supporting particularly close integration, though most others provide object storage within containerized environments.

**Ransomware protection.** Object storage systems universally claim ransomware protection, though details vary. Most vendors claim to offer data immutability, or the blocking of any changes to data once it is stored. This helps to prevent ransomware attacks, in which attackers encrypt data and extort ransom to unlock it. WORM (write once, read many) technology is related, allowing data to be frozen against any changes once it's written to a backup device. Other security measures, including encryption for data at rest and in transit are offered by several leading players, including **Cloudian, DataCore, MinIO,** and service providers such as **Cloudflare, Backblaze,** and **Wasabi.**

A feature calledMulti-User Authentication (MUA) is also supported by **Wasabi**, which requires that up to three individuals must sign off on deleting an account in the object store.

**Compliance and regulatory support.** To support the regulations governing data in various jurisdications, object storage providers offer a range of data management functions. These include versioning of data, restricted access, automated retention timeframes, general compliance reports, and the like.

**Veeam integration.** Related to ransomware protection is integration with the data backup, protection, and recovery capabilities of data management vendor **Veeam**. This provider leads the pack of many object storage system partnerships because it provides unusually strong methods of backing up data and restoring it in the event of a disaster. Nearly all major object storage service and system providers offer Veeam integration.

**Multi-tenancy.** Storage shared across multiple departments, divisions, or locations can be more efficient and cheaper to operate. Nearly all leading object storage vendors support this feature, though they differ in methods of implementation.

**Public cloud integration.** A few object storage providers, such as **OSNexus, Qumulo, VAST Data,** and **WEKA,** function within public cloud services such as AWS, Azure, and GCP. Nearly all vendors, while not running within the public cloud, support the ability to back up data with these services or exchange data with same.

**Flexible pricing.** The leading cloud providers typically charge for moving data into and out of their services, a tack that the growing roster of independent object storage providers eschews with vigor. Pricing is usually based on usage of storage increments, though subscriptions and other options are available. Some vendors, such as **Qumulo,** offer economies of scale based on volume usage: The more you use, the cheaper the increments. The result of these pricing options can be dramatic: **Wasabi** claims its service is up to 80% cheaper than services offered by the hyperscalers.

## Other Features

There are other features offered by object storage providers that aren't universally common. Among these is the ability to offer a global or universal namespace. This feature, supported by **Cloudian, DDN, DataCore, Hammerspace, MinIO, Nutanix, Quantum, Qumulo, VAST Data,** and **WEKA,** ensures that data appears to be stored in one place, though actual object storage systems can be distributed across multiple geographical locations. This streamlines the retrieval and management of data items.

Another feature not offered by all is simultaneous support of file and object data in a single system, although this is increasingly becoming another table stake. Support for files and objects typically

entails technology capable of parsing file directories as well as unstructured object data. **Cloudian, DDN, Hammerspace, OSNexus, Pure Storage, Quantum, Qumulo, Scality, VAST Data,** and **WEKA** (to name just a few) support the ubiquitous Server Message Block (SMB) for objects along with Network File System (NFS) and/or File Transfer Protocol (FTP).

**VAST Data**, interestingly, normalizes data and disaggregates it from protocols, allowing data to be stored and retrieved regardless of whether it was originally in file, table, or object format. Each data element, VAST says, is associated with multiprotocol support within a global namespace and stored in a lakehouse format termed an Element Store by VAST. This makes data management and retrieval faster and more efficient.

Data lake/data lakehouse capability allows for analytics to be run on stored objects gathered into a common framework. These data lakes may also support input to AI workloads. Besides **VAST Data,** vendors offering data lake functionality include **Cloudian, Dell, Hitachi Vantara, IBM, MinIO,** and **Wasabi.**

Notably, analytics on stored objects in the past has been difficult due to the lack of formatting inherent in unstructured data. So companies offering data lake/lakehouse and analytics have spent time to overcome bottlenecks formerly associated with this step. **MinIO**, for instance, claims to gather extensive metrics on a range of aspects of object storage, including access times, usage, replication status, and other variables.

In this vein, **VAST Data** announced in October 2024 a product called VAST InsightEngine with NVIDIA, which leverages VAST's ability to store and retrieve large amounts of file, object, and table data in a unified fashion. Data is then combined with NVIDIA's Inference Microservices (NIM) technology to create vector embeddings. The results are stored in VAST's DataBase, available instantly—within milliseconds, VAST says—for use in AI workloads.

# 4. Hardware Choices

By now, it's clear that the object storage market bristles with variety. There are all kinds of software choices available, but hardware options proliferate too, and any discussion of object storage isn't complete without reference to hardware.

Most object storage vendors support servers from the likes of **HPE, Lenovo**, and **Supermicro** that are specifically architected to support flash storage on solid-state drives (SSDs). Several, including **Cloudian** and **MinIO**, support these options along with their own appliances.

System vendors such as **HPE** and **IBM** rely on their own hardware, which is available as part of a managed services option. Longstanding storage providers such as **Pure Storage** offer similar service options based on proprietary hardware.

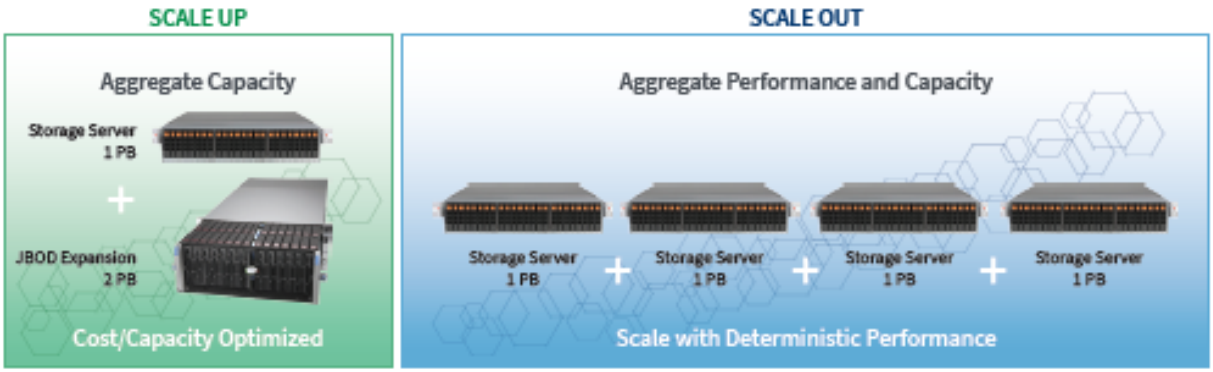
An essential element of any object storage implementation is Non-volatile Memory Express (NVMe), a protocol that speeds up the transfer of data within a storage system by using Peripheral Component Interconnect Express (PCIe), the multi-lane, serial high-speed bus that links a computer's motherboard to peripheral devices. NVMe is an essential element in improving the performance of storage networks and linking storage to GPUs in AI environments (more on that momentarily). Most of the vendors mentioned in this report support NVMe, with the exception of **Pure Storage,** which has created its own flash interface.

Flash storage isn't the only option for object storage systems. **Quantum** maintains that efficient scalability requires a tiered approach that incorporates flash, disk, and tape to support data at all points in the lifecycle—including data that is "hot," or active, and "cold," or archived.

## Software-Defined Storage and the Rise of AI

As AI workloads continue to proliferate, storage hardware providers have created new solutions. Datacenter system supplier **Supermicro**, for example, offers a comprehensive lineup of storage servers catering to object storage as well as block and file storage. Configurations range from small (under 500 terabytes) to cloud-scale (more than 10 petabytes), comprising computing units, or nodes, based on AMD or x86 Intel CPUs, attached to hard disk drives (HDDs) or flash storage in solid state drives (SSDs).

Supermicro also specializes in software-defined storage (SDS), a technique which allows for the massive scale typically required for object storage. By decoupling the software choices from hardware, SDS, which is also supported by other suppliers, including **DataCore, Dell, Nutanix**, and **Scality,** allows customers the option to *scale out* storage via clusters of multiple servers, with storage and controller capacity increasing in unison. This approach contrasts with *scaling up* storage, in which storage media is added as capacity is required. In the scale-out approach, SDS allows for the distribution of data across multiple servers, deploying data replication or erasure coding to ensure that if a server fails, data isn't lost.




*Scale-up storage aggregates capacity into one system, while scale-out storage aggregates capacity from multiple servers. Source: Supermicro. Used with permission.*

Supermicro has added designs to its storage fleet to accommodate AI workloads. In October 2024, the vendor announced an addition to its Petascale storage server series that features a JBOF (just a bunch of flash) system supporting SDS and configured with up to four NVIDIA BlueField-3 data processing units (DPUs) in a 2U form factor. By adding the NVIDIA DPUs, Supermicro eliminates the need for CPU, network card, PCIe switch, memory, and baseboard management controller, making the server faster and more power-efficient for AI workloads. The BlueField DPU supports 400-Gb/s Ethernet or InfiniBand networking and accelerates the NVMe over Fabrics (NVMe-oF) storage protocol and other Remote Direct Memory Access (RDMA)-based storage traffic. The JBOF is handled in scale-out fashion.

**Cloudian** and **Hammerspace** have integrated their object management software systems with the Supermicro Petascale JBOF with NVIDIA BlueField-3 DPUs. According to Cloudian, the solution streamlines AI workloads, particularly since it works with NVIDIA's GPUDirect storage interface, which allows storage to directly interact with NVIDIA GPUs. (More on that in a moment.)

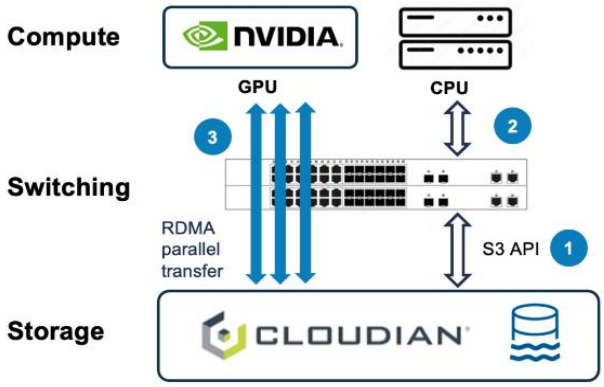
# 5. How Object Storage Is Changing for AI

Despite object storage's inherent advantages for AI workloads, there are challenges. One of these has been throughput, or the rate at which data can be transmitted from one device to another—an issue for any type of storage. Unless throughput between storage and GPU is fast enough, the GPU can stop processing, eating into ROI.

Object storage providers address the throughput issue in a couple of ways. Several, including **MinIO** and **Cloudian**, have adopted RDMA and the associated RDMA over Converged Ethernet (RoCE) protocol to allow faster transmission of data between servers. As noted above, Cloudian's HyperStore also supports the Supermicro Petascale JBOF storage server with integrated DPUs.

Cloudian's Hyperstore also supports GPUDirect, NVIDIA's interface that links a GPU directly to storage, bypassing the CPU. **Quantum, VAST Data,** and **WEKA** also claim GPUDirect Storage certification from NVIDIA. Other vendors, including **Nutanix**, say they're in the process of implementing GPUDirect Storage.

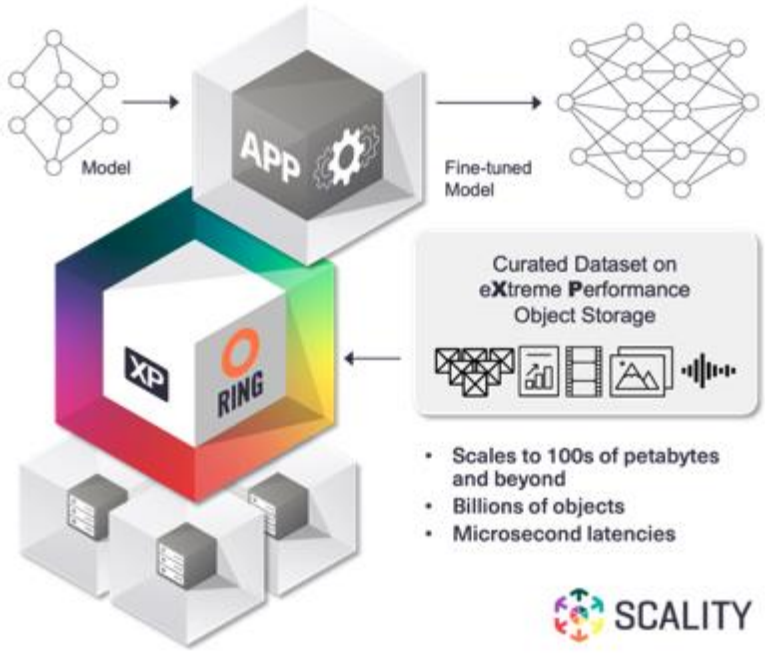
The changes wrought by RDMA and GPUDirect are significant. **Cloudian** claims its system reduces CPU utilization by 45% during data transfers and allows for throughput of over 200 Gbytes/second from object storage to GPU. Without RDMA, throughput would be just a fraction of that rate, perhaps in the tens of Gbytes/second, depending on various aspects of the storage configuration. Cloudian also says that RDMA allows parallel transfer of data from multiple nodes, further improving performance, as illustrated in the diagram below:




1. **Data requests are initiated via the S3 API**
2. **Instead of routing through system memory and CPU, data moves directly to GPU memory**
3. **RDMA enables parallel transfer from multiple Cloudian nodes**

*Source: Cloudian, used with permission.*

Another approach to throughput is taken by **Scality**, which provides RING XP, an AI add-on to its flagship RING system for object and file storage that delivers microsecond response times for feeding AI modeling and fine-tuning workflows. Scality's achieved this by streamlining the object storage API of RING to support a limited set of operations, optimizing small file I/O acceleration, and running RING XP on all-flash, x64-based NVMe storage servers from Dell, HPE, and Supermicro. The result, Scality claims, is that RING XP runs "an order of magnitude" faster than AWS's S3 Express One Zone, Amazon's storage class designed to accelerate the performance of S3 object storage by 10X (according to AWS), achieving a claimed "single-digit millisecond latency." In contrast, Scality says that RING XP can achieve rates in the "hundreds of microseconds" when handling data items of 4KB or less.




*Scality's architecture for microsecond latency. Source: Scality*

While Scality's claim isn't independently verified, it demonstrates the priority Scality's put on performance and speed, which are key to efficiency in delivering data to AI workloads running on expensive hardware.

## Incorporating AI Into Object Storage

Another trend in the object storage space is the addition of AI capabilities into the system itself. **MinIO's** AIStor, for example, includes a new API that extracts information from unstructured data using natural language prompts; a private repository for AI models that's compatible with Hugging Face; an overhauled user interface designed for easy access to MinIO functions by IT personnel as well as developers; and, as previously noted, S3 over RDMA for optimal delivery of data in AI workflows. AIStor effectively makes an AI tool out of the object storage platform. And no expertise with RAG models or vector databases is required; everything is taken care of under the covers by AIStor.

A RAG pipeline and vector database—two essentials for feeding data to AI workloads—are incorporated into **Pure Storage's** FlashBlade appliances. Pure developed these capabilities in partnership with NVIDIA, and the solution requires a range of NVIDIA prerequisite hardware and software. The goal is to use FlashBlade as the preparatory staging site for direct linkage to GPUs. To make this happen, FlashBlade//S packs a vector database named Milvus that preps data for use in RAG by turning data characteristics into vector embeddings.

### Selectivity Is Key

An essential benefit of object storage systems is the ability to select data for use in AI. Systems that provide capabilities to distinguish specific data sets enable AI workflows to avoid hallucinations. **Komprise's** president and COO Krishna Subramanian explains: "We hand off to AI data lakes but also to AI workflows. For example, if you are working on ChatGPT to get a chatbot about HR, you might have a bucket from Azure to feed prompts to the model. You want to be selective about the data you put in that bucket. Too much data can result in hallucinations. You might say just provide me the latest benefit documents and put that in the bucket."

It's worth recalling at this point that global file system provider **CTERA** retrieves metadata from a range of object file stores, indexes it, and links it to third-party AI models. CTERA also has added specialized AI agents for information retrieval from those models. Six agents are available, each with a specific persona, avatar, voice, and area of expertise. "Julian," for example, excels in documentation and can find answers to questions about file versions and the like. "Don Draper" creates marketing content from specific data. "Saul Goodman" is a legal expert, and so forth.

## 6. Conclusion: Object Storage Is on the Rise

Object storage is the only ideal technology for fueling AI workloads with unstructured data, which has become a rich source of information that enterprises are eager to incorporate into their inferencing environments. Indeed, the public cloud hyperscalers realized a while back that the *only* way to scale data delivery for AI was with object storage, which provides a flat file architecture for secure, easy distribution and access to exabyte-level volumes of data.

Still, there are challenges. Here's how Ugur Tigli, CTO of **MinIO** put it in a blog post:

*"MinIO made the choice to invest in these AI workloads because we knew MinIO could scale. We were still surprised at how challenges emerged at 1 EiB - and in places we didn't think about. Places like memory, networking, replication and load balancing."*

Vendors like MinIO have been in the object storage business long enough to figure out how to alleviate these roadblocks and respond to demand that is shifting toward AI. The result is that products are moving beyond merely coordinating storage in a detailed fashion for operational admins. "It's a whole new way of thinking," said **CTERA** CTO Aron Brand in a presentation recently. "Storage [is] not just a standalone piece anymore, it's part of a bigger integrated platform."

Growing demand for these integrated, intelligent platforms has formed a hot market. In replacing NAS with more efficient alternatives based on cloud technologies such as Kubernetes and serverless computing, object storage wares are growing faster than NAS and could by some estimates reach $20 billion annually by 2032. Indeed, considering all factors, some vendors say the cloud-native object storage market could conceivably reach over $110 billion within the next decade.

Futuriom's empirical research supports these forecasts. While many firms in this area are over ten years old, business has picked up considerably over the past two years. At least three of the privately held companies we interviewed for this report (including **MinIO, Qumulo**, and **Wasabi**) have valuations over $1 billion, putting them in a pre-IPO category. Double-digit growth rates are common, and customer lists increasingly include the world's largest firms. Growing demand for unstructured data in AI workloads should guarantee robust business for a long time to come.