CoreWeave

CoreWeave AI Object Storage: AI-Native Storage Without Limits

Accelerate training, fine-tuning, and inference workloads while eliminating costly data transfer barriers

Set your data free

Move your data with CoreWeave Zero Egress Migration and eliminate egress fees—no matter where your data is located.

Learn more

Why general-purpose storage throttles AI innovation

Modern AI workloads place continuous, simultaneous pressure on storage in ways general-purpose systems were never designed to support. When storage becomes the constraint, GPUs wait on data, pipelines break under scale, and infrastructure costs outpace performance gains across training, fine-tuning, and inference.

Conventional infrastructure breaks under AI pressure

Modern AI workloads overwhelm general-purpose infrastructure through:

- Extreme parallelism
- Massive read throughput to keep GPUs busy
- High write throughput for checkpointing

The impact: idle GPUs, brittle pipelines, and rising costs at scale

AI-native storage must align tightly with compute

AI-native storage should meet the demands of modern AI workloads by delivering:

- Tight alignment with compute
- Sustained, predictable throughput at GPU scale
- Low-latency object access across training and inference
- Native support for massively parallel data pipelines

The impact: productive GPUs, stable pipelines at scale, and infrastructure spend that directly maps to performance

COREWEAVE AI OBJECT STORAGE ADVANTAGES

Up to 7 GB/s
per GPU

99%
uptime

Zero egress fees
migration program

ENGINEERED FOR MAXIMUM PERFORMANCE

Object storage purpose-built for AI at scale

CoreWeave AI Object Storage is an AI-native object storage platform designed to align tightly with GPU compute, delivering predictable, high-throughput data access for AI workloads.

With native S3 compatibility, CoreWeave AI Object Storage integrates seamlessly into existing AI workflows, allowing teams to use familiar tools and pipelines without modification.

Built to scale with GPU demand, it delivers predictable, high-throughput performance across training, fine-tuning, and inference—without the replication overhead or performance bottlenecks common to general-purpose storage.

CoreWeave AI Object Storage, which is designed for enterprise and research environments, combines global-scale data access, transparent pricing, and enterprise-grade reliability and security. It delivers a scalable foundation that keeps GPUs fed, simplifies data management, and supports everything from low-latency inference to massively parallel training pipelines.

Conventional storage	CoreWeave AI Object Storage
Tiered architecture	**Flat, unified design**
Replication lag	**Consistent global access**
High egress fees	**Zero egress fees**
Hot-tier dependency for performance	**Instant access to all data**
Single-cloud lock-in	**Architected for multi-cloud deployments**

Throughput that scales with GPU demand

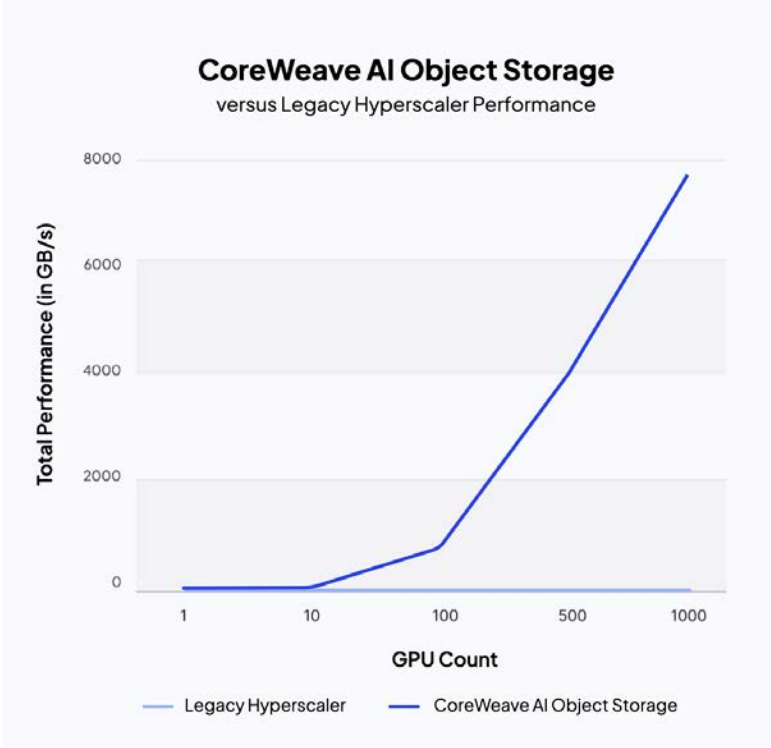
Under sustained, high-concurrency training workloads, CoreWeave AI Object Storage scales throughput as GPU clusters expand. General-purpose object storage architectures flatten under extreme parallel load.

Benchmark context

Performance was measured across 16× NVIDIA Blackwell Ultra nodes using the Warp S3 benchmarking tool for distributed object reads. Testing included both Ethernet (TCP) and RDMA (NVIDIA Quantum InfiniBand), with peak sustained throughput achieved over RDMA with LOTA caching enabled.

Result: Sustained 7+ GB/s per GPU with linear scaling as GPU count increased.

Results vary by workload and configuration.



Data works harder with LOTA

Modern AI workloads demand sustained, predictable data access at GPU scale. CoreWeave AI Object Storage delivers this performance through Local Object Transport Accelerator (LOTA), an AI-native object transport and caching layer embedded directly into the storage service.

LOTA runs on GPU cluster nodes as a distributed proxy that acts as an S3 endpoint, intelligently caching frequently accessed objects on local NVMe disks.

With LOTA, proximity to GPUs streamlines performance.

By staging data close to compute, LOTA eliminates the latency and throughput constraints that force GPUs to idle in traditional object storage architectures. This approach enables highly parallelized reads and writes across GPU nodes, keeping training and fine-tuning workloads continuously fed with data.

Unlike general-purpose storage, which often requires teams to deploy and manage separate caching layers, LOTA is embedded into CoreWeave AI Object Storage by design.

Powered by a Quantum InfiniBand backbone and local NVMe caching, it **delivers up to 7 GB/s per GPU** and scales linearly as GPU clusters grow, sustaining peak performance across regions and clouds without adding operational complexity or new points of failure.

4 reasons caching layers fail at AI scale

- They're bolted onto general-purpose storage, introducing coordination overhead and new failure points
- They collapse under extreme parallel access, becoming throughput bottlenecks as GPU counts rise
- They shift performance into an operational problem that teams must size, tune, and constantly manage
- They fragment data visibility and consistency across cache tiers and regions

Learn more

Beyond the Hot Tier: Cut AI Storage Costs While Accelerating What Comes Next

Read the eBook

Automated, usage-based billing for predictable storage costs

In large-scale AI pipelines, data volumes grow fast. Every experiment, checkpoint, and branch adds artifacts that accumulate quickly—and most of that data becomes inactive but potentially useful. Keeping everything "hot" inflates costs, adds operational drag, and limits headroom for new data. Moving data to traditional low cost storage makes it unavailable to production workloads when needed.

Why AI storage costs spiral

- AI workloads skew heavily toward inactive data over time
- General-purpose object storage prices all data as if it's active—or relies on cumbersome tiering to offset costs
- As data sprawl grows, costs rise faster than model performance or iteration speed

Impact: Teams overpay for data they rarely access

Why deletion and archiving fail

- Deleting checkpoints breaks auditability, rollback, and reproducibility
- Manual cleanup scripts add operational risk and on-call burden
- Traditional archive tiers introduce policy complexity, retrieval delays, and extra fees

Impact: Teams keep more data hot than necessary

How CoreWeave is different

Automated hot, warm, and cold usage-based billing:

- Recognizes inactive data based on real access patterns—without moving data or changing how it's accessed
- Adjusts automatically based on usage, while every object remains instantly accessible at full performance

Impact: No rehydration, tier transitions, or performance cliffs

The advantage

CoreWeave recognizes data inactivity and bills it accordingly without changing where data lives or how it's accessed, resulting in:

- **Up to 75% lower storage costs** for typical AI workloads
- One performance profile across all data
- No egress, request, retrieval, or tiering fees

Impact: Storage economics align with AI workloads

Rely on one global dataset across regions and clouds

CoreWeave AI Object Storage enables a single global dataset that can be accessed consistently across regions, clouds, and environments, eliminating data silos and simplifying AI at global scale.

> “
>
> CoreWeave AI Object Storage delivers the throughput and reliability our research pipelines depend on, balancing speed and efficiency across active and inactive data. It’s allowed us to experiment faster while keeping costs under control.
>
> **Xander Dunn**
> *Technical staff member at Periodic Labs*

In general-purpose clouds, running AI workloads across regions or clouds has historically required duplicating massive datasets, managing synchronization drift, and absorbing significant egress costs. With CoreWeave AI Object Storage, that burden disappears.

One unified dataset

Work from a single source of truth across regions and environments—without duplicating data or managing synchronization drift.

No data divergence

Eliminate inconsistencies caused by maintaining multiple dataset copies. Every team, job, and model works from the same data, every time.

No egress, ingress, or request fees

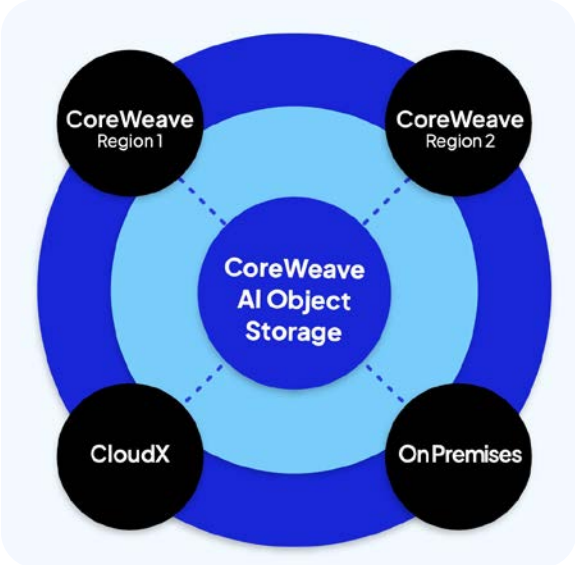
Move and access data without punitive transfer charges, keeping global AI deployments simple, predictable, and cost-efficient.

Consistent performance

Access data with predictable, high-throughput performance wherever workloads run, enabled by CoreWeave’s purpose-built networking backbone and LOTA acceleration.

Freedom to optimally place workloads

Run training, fine-tuning, and inference wherever performance, cost, or compliance requirements demand—without redesigning your data architecture. Run workloads wherever you have GPUs.



Move AI datasets into CoreWeave without duplication, disruption, or egress fees.

Cross-region access is available now. Multicloud access is under active development.

CoreWeave Zero Egress Migration Program

The CoreWeave Zero Egress Migration (0EM) program eliminates the cost, risk, and friction of moving large datasets into CoreWeave AI Object Storage.

Migrate data from other clouds or on-prem environments without paying egress fees, enabling fast, low-risk adoption at any scale.

01 **No egress fees, no ultimatums**

CoreWeave covers egress costs from your existing cloud without requiring account closures or one-time, all-or-nothing switches.

02 **Migrate at petabyte scale, on your terms**

Move datasets incrementally while retaining full access to existing environments—no forced cutovers or workflow disruption.

03 **Freedom to use data anywhere**

Once data is in CoreWeave AI Object Storage, you pay no CoreWeave egress fees, no matter where workloads run.

04 **Lower risk, faster time to value**

Remove the cost and friction that delay adoption, so teams can start training, fine-tuning, and deploying on CoreWeave immediately.

Talk to a CoreWeave storage expert

Discuss your data scale, performance requirements, and cost constraints—and see where CoreWeave AI Object Storage can remove friction.

Contact us

Zero Egress Migration

Move your data to CoreWeave and eliminate egress fees.

Discover how →

Move beyond AI storage tradeoffs

See how CoreWeave AI Cloud helps you simplify and accelerate storage for the AI era.

Watch on demand →

5 Costly AI Storage Traps

Uncover the hidden storage traps in general-purpose clouds that are holding your AI workloads back.

Explore the traps →