CoreWeave

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

How usage-aware storage cuts cost and accelerates innovation

Introduction

In dynamic AI development, storage is no longer passive or predictable. As models are trained on petabytes of ephemeral data, teams are left storing it "just in case" and paying hot-tier prices for files they may or may not need again.

Efforts to wrangle storage sprawl with tiering rules, lifecycle policies, or cleanup scripts only add potential friction, operational debt, and risk. In fast-moving AI workflows, yesterday's discarded file may become tomorrow's critical training input, or a cached model artifact powering a production endpoint.

Even for the most strategic teams, there's a fundamental challenge: you don't always know which data will be valuable later. That uncertainty is why many teams default to keeping nearly everything accessible, even as hot-tier pricing costs pile up. And while no public benchmarks quantify inactive data in AI storage, CoreWeave's internal tracking shows that 60% to 80% of data typically sits inactive, with some customer environments exceeding 90%.[1]

Given the cost and performance stakes, it's time for storage to adapt to how AI teams actually work—not the other way around. This eBook explores why traditional tiering strategies fall short,and what a usage-aware, automated AI-native approach unlocks.

More than 80% of stored AI data sits inactive.

Why tiering and “intelligent” archive tactics fail AI teams

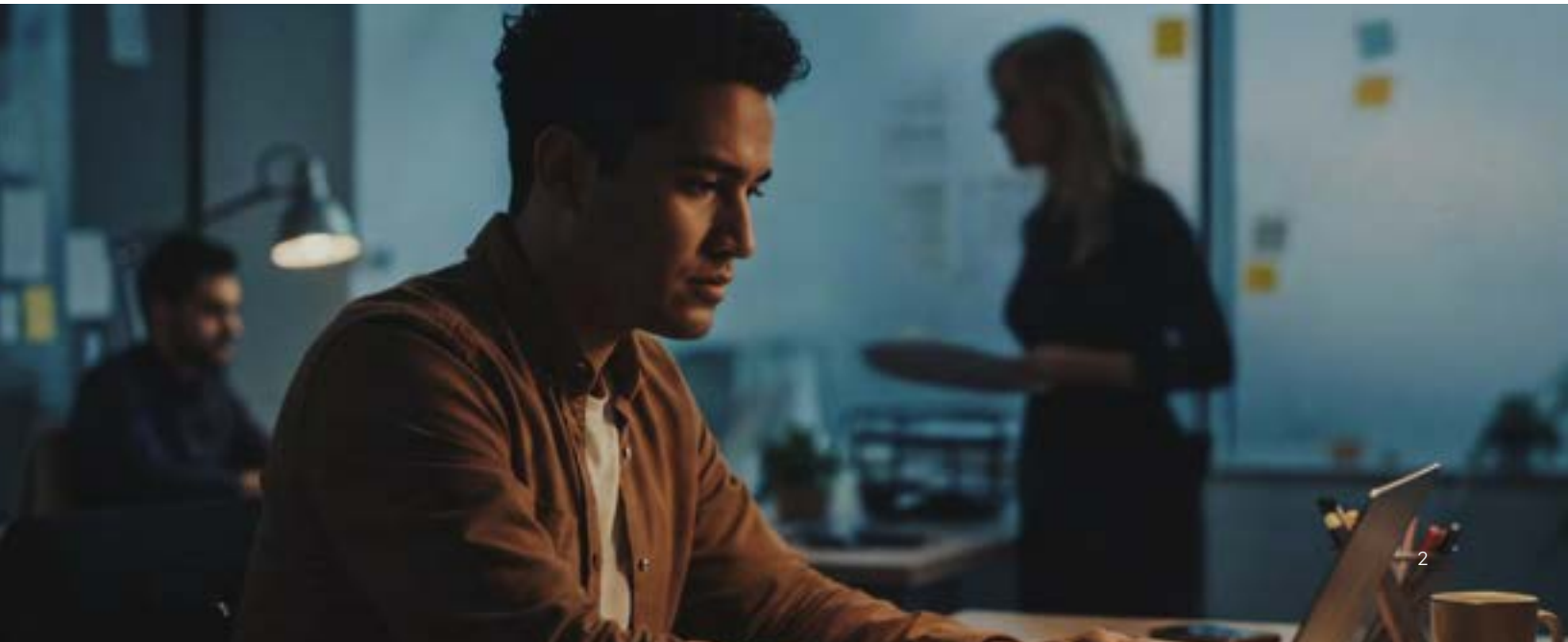
Traditional cloud storage tiering was designed for backup, compliance,and long-term historical analysis, not for the velocity and scale of modern AI pipelines. To retrofit their platforms, hyperscalers have bolted on archive tiers and “intelligent” automation. These systems apply policies that demote older files based on recent usage patterns, promising lower hot-storage bills.

Under the hood, however, these approaches potentially introduce as many tradeoffs as benefits:

- Manual tagging and policy configuration
- Charges to track access or move files between tiers
- Rehydration delays for “cold” data, even when urgency is high
- Unpredictable costs for retrieval, requests, and egress

For teams running model-scale workloads, data access patterns are irregular and hard to predict. Yesterday’s checkpoint or intermediate artifact might be critical tomorrow for a regression fix, an audit trail, or a fine-tuning step. As a result, most teams either keep everything hot and absorb the cost or rely on tiering rules that may delay access when speed matters most.

On the surface, tiering looks like cost optimization. In practice, it introduces operational overhead, pipeline fragility, and enough performance drag to offset much of the intended savings. That impact only compounds at scale.



How traditional tiering adds friction

Claims	Reality	Provider(s)
Intelligent tiering automatically lowers cost	Requires tagging, policy setup, and per-object monitoring fees[2]	AWS
Archive tiers offer rapid access	Restores can still take minutes to hours depending on volume/class[3]	AWS, GCP
Flexible lifecycle management	Requires access tracking and multi-step configuration[4]	Azure
Millisecond archive retrieval	Applies only to certain classes; not guaranteed for pipeline readiness[5]	GCP
Lower-cost storage tiers	Retrieval fees, retention windows, and rehydration lags reduce actual savings[6]	All
Built-in automation	Still requires manual config; introduces version sprawl and surprise line items[7]	All

Ultimately, most strategies treat storage optimization as a configuration problem. But that mindset has proven incapable of fully supporting the speed, scale, and unpredictability of AI.

AI teams don't need more lifecycle logic. They need storage that adapts to real usage rather than assumptions.

The real cost of maintaining the status quo

Even when well-intentioned, most storage strategies for AI workloads force teams into tradeoffs between cost, speed, and complexity. Here's what that looks like in practice.

The recurring cost and impacts of inactive or tiered data at scale:

Scenario 1: Hot tier everything	Scenario 2: Rehydration delay	Scenario 3: Manual policies
A large AI platform team stores 20 PB of data, including model checkpoints, logs, datasets, intermediate artifacts, and inference outputs. To ensure everything is always available, whether for retraining, rollback, audit, or model comparison, they leave it all in the hot tier.	An ML engineer is investigating a model performance regression in production. To diagnose the issue, the engineer needs to compare current inference behavior with outputs from a previous training run. However, those files were automatically moved to a lower-cost storage class under lifecycle rules designed to save on hot-tier costs.	To avoid spiraling hot-storage bills, the infra team maintains a patchwork of tagging logic, lifecycle rules, and cleanup scripts across multiple buckets. As workloads scale, so does the operational burden.
Impacts: • 20 PB (billed as 20 PiB) at hot-tier pricing • Monthly cost at $0.023/GB (e.g. AWS S3 Standard[8]) $482,344)/month • Annual cost: $5.8 million • No rehydration delays, no cost controls	Impacts: • 3 –5 hour restore delay to access archived artifacts • Drift diagnosis is stalled while the team waits or reruns experiments • GPU clusters sit idle: even 512 GPUs × 4 hrs = $30K–$50K wasted capacity • Request, restore, and egress fees pile up	Impacts: • 10–15 hours/month spent on policy tuning, script maintenance, and access auditing • At a blended cost of $150/hour, that's $18,000–$27,000/year in engineering time • Risk of demoting or deleting critical data too soon • Inconsistent object access behavior leads to pipeline fragility and on-call escalations
→ Fast access, but expensive and unmanaged	→ Policy-driven storage transitions block root-cause analysis and burn valuable GPU time	→ Changing workloads break brittle scripts and steal engineering focus, especially at scale

Rethinking storage and archiving for speed, scale, and surprises

CoreWeave offers an AI-native approach to storage. CoreWeave AI Object Storage leverages automated usage-based billing to deliver high performance while continuously optimizing cost based on real access patterns.

Unlike traditional platforms that rely on multiple storage classes and archive tiers, CoreWeave AI Object Storage employs an automated usage-based model that delivers both consistent performance and adaptive and transparent pricing. All data resides in a high-throughput environment, with every object remaining fully accessible at line-rate performance—up to 7 GB/s per GPU—regardless of how frequently it's accessed.

Instead of relying on manual tagging or static storage classes, pricing automatically adjusts based on actual access patterns. This enables teams to keep everything available without worrying about what tier it belongs to or when it was last used.

Storage that automatically adapts to your workflow

AI pipelines generate irregular, long-tail access patterns that defy age-based logic. CoreWeave AI Object Storage adapts dynamically without guesswork, tagging, or cleanup. Whether a file was touched yesterday or a month ago, performance remains consistent.

CoreWeave automatically classifies data into three billing levels—Hot ($0.06/GB/mo), Warm ($0.03), and Cold ($0.015)—based on how frequently it's accessed. Unlike traditional platforms that move data between separate storage classes, all billing levels in CoreWeave AI Object Storage share the same high-performance storage environment.

This is storage that self-optimizes based on how your data is actually used. Your team can focus on building, not managing while ensuring your GPUs stay fully utilized without pipeline stalls or restore lag. For existing CoreWeave customers, this simplified, usage-based model has reduced storage costs by more than 75% while maintaining the same high performance and ease of use.

This is storage that self-optimizes based on how your data is actually used. Your team can focus on building, not managing, while ensuring your GPUs stay fully utilized without pipeline stalls or restore lag. And because performance stays consistent across all data, there's no need to choose between saving money and keeping everything ready to use.

From lifecycle rules to workflow-aware storage

How CoreWeave reimagined storage for AI workloads:

From…	To…
Age-based lifecycle policies	Usage-based billing logic
Tagged files and scripted rules	Hands-off, automated classification
Cold storage with restore delays	Instant access with no performance penalty
Manual tradeoffs between cost and speed	Unified tier with consistent throughput
Tracking and retrieval fees	No access or egress charges
“What if we need it later?” anxiety	Keep everything without cost anxiety or complexity

In the end, high-performance object storage doesn’t just cut costs, it removes infrastructure drag on your model velocity.

How the benefits of usage-aware storage add up

Let's revisit the earlier scenarios to see how usage-aware storage improves performance and cost before diving into a full comparison table.

The advantages of usage-aware storage:

Scenario 1: Blended pricing, same performance	Scenario 2: Always-on access	Scenario 3: Zero engineering overhead
For the 20 PB workload (billed as 20 PiB), CoreWeave AI Object Storage dynamically adjusts pricing based on actual access patterns. Objects unused for 7 days automatically shift from Hot ($0.06/GB/mo) to Warm ($0.03), and after 30 days to Cold ($0.015) rates—all without affecting performance or requiring manual tiering.	With CoreWeave AI Object Storage, every object stays immediately accessible at full line-rate performance no matter how long it's been since last use . Pricing adapts automatically to access frequency, so there's no need to demote or reclassify data manually.	With CoreWeave AI Object Storage, infra teams focus on system performance and ML enablement instead of spending hours per month tuning lifecycle scripts. Usage-aware pricing just works.
Impacts: • Cost ranges from just $307,000 to $409,000/ month vs. $482,244/month with traditional hot storage • All data remains instantly accessible • Predictable billing aligned to real data usage	Impacts: • 0 hrs lost to restore lag • No GPU idle time • No retrieval or egress fees • Root cause analysis happens in real time	Impacts: • $18K–$27K/year in engineering time reclaimed • No more versioning surprises or accidental data loss • Reduced cognitive load and operational debt

Now the question is: What kind of cost savings and simplicity can a usage-aware model actually deliver versus legacy approaches? The comparison table below offers a side-by-side look. It models a 20 PB workload typical of large-scale AI pipelines.

Assumptions include:

- Average file size: 1 MB
- Monthly uploads: 20% of total data
- Monthly reads: 30% of total data
- Monthly egress to another cloud: 30%

These figures reflect real-world usage across training, inference, and artifact reuse. CoreWeave's pricing reflects usage-aware billing with no added access or retrieval fees. Hyperscaler costs reflect public hot/standard storage classes, including typical operational and data-transfer charges based on their public pricing structure.

Why lower rates don't always equal lower bills

While the raw storage rate for hyperscaler storage may look lower on paper, it's only part of the picture. Once you factor in common per-request charges, rehydration lag, and egress fees, the savings simply don't hold up.

From per-GB price to true cost: how the numbers add up

Provider	Per-GB price	Storage cost only (20 PB)	Operational and transfer fees	Performance profile	Total cost (20 PB)
CoreWeave	$0.06 / $0.03 / $0.015 (usage-based; blended $0.026*)	$544,526	$0	Instant access at line-rate; no request or egress fees	$544,526
AWS S3 Express	$0.110/ GB/mo[9] (blended $0.126)	$2,306,867	$336,816	Immediate access; fees for requests, retrieval, and egress apply	$2,643,683
AWS S3 Standard	$0.021/ GB/mo[10] (blended $0.037)	$440,402	$338,625	Immediate access; request + egress fees apply	$779,027
GCP Cloud Storage Standard	$0.024/ GB/mo[11] (blended $0.048)	$482,345	$527,368	Immediate access; request + egress fees apply	$1,009,713
Azure Blob Hot LRS	$0.021/ GB/mo[12] (blended $0.037)	$440,402	$345,711	Immediate access; request + egress fees apply	$786,113

When you remove the layers of fees, friction, and delay, what you're left with is smarter infrastructure and a clear advantage for AI teams that need to move fast.

*CoreWeave AI Object Storage automated usage-based billing automatically manages data across three transparent pricing levels—Hot ($0.06/GB/mo), Warm ($0.03/GB/mo), and Cold ($0.015/GB/mo)—based on real access patterns. For this example, a representative data distribution (14% Hot / 32% Warm / 54% Cold) yields a blended effective rate of $0.026/GB/mo with no request, retrieval, or egress fees. Hyperscaler totals reflect publicly listed U.S. East region rates for hot/standard storage as of October 2025, including typical request, retrieval, and egress charges based on the modeled workload. CoreWeave pricing does not vary by region.

Note: Storage calculations use industry standard binary gigabytes (1 GB = 2^{30} bytes or 1,024 MB).

The compounding advantage of storage that just works

Storage that adapts to how your team actually works does more than cut costs — it accelerates everything downstream.

By eliminating the need to choose between performance, retention, and cost, CoreWeave turns storage into a strategic asset for AI teams. There's no need to tag, guess, hydrate, or justify. All of your data is available. All of it performs. And the pricing reflects how it's actually used.

For teams pushing to reduce iteration cycles, maximize GPU efficiency, and improve observability across experiments, this creates a flywheel:

- **Fewer stalls** → better GPU utilization
- **Less overhead** → faster iteration
- **More access** → richer model development and comparison
- **Lower cost** → reinvest in performance

You get to keep everything, move faster, and spend less so your team can stay focused on building, shipping, and scaling with confidence.

Six principles of optimal AI storage

1. Transparent pricing aligned to real usage
2. Automation without policy sprawl
3. Instant access, no rehydration
4. Predictable pricing, no hidden fees
5. Performance parity across all data
6. AI-native integration, not bolted-on tiers

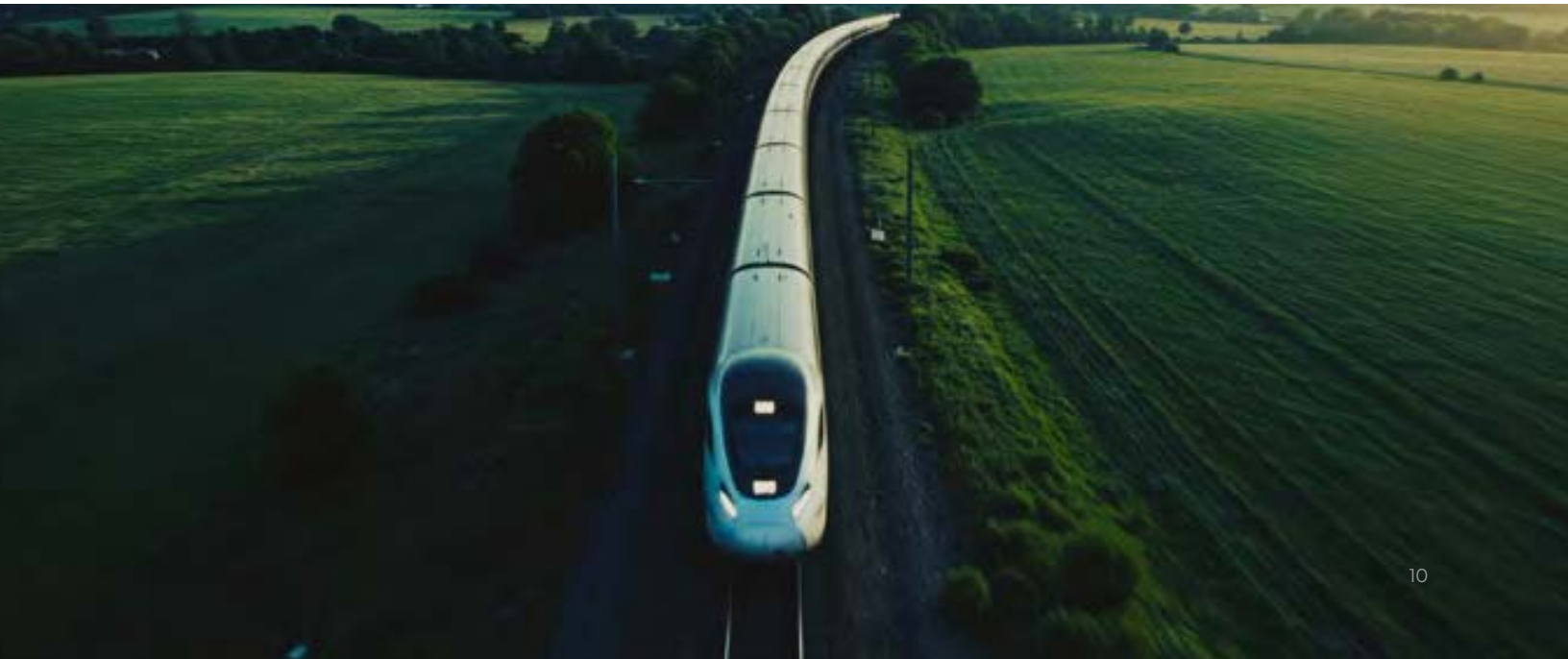
Take your AI ambitions further, faster

- Learn more about moving beyond tiers, tradeoffs, and runaway costs in AI storage. Watch the webinar.
- Discover how CoreWeave delivers performant, secure, and reliable storage for AI. Explore storage solutions .
- Learn how CoreWeave cuts storage costs by up to 75% for our customers with our new, automated usage-based billing levels. Read the blog.
- Talk directly with a CoreWeave expert. Book a time .

About CoreWeave

CoreWeave, the AI Hyperscaler™, powers AI innovations by bridging the gap between AI ambition and execution. Purpose-built for the demands of accelerated computing, CoreWeave delivers cutting-edge performance, scale, and expertise with the infrastructure solutions that AI needs today and in the future.

Learn more at www.coreweave.com



Sources

[1]*AWS S3 Pricing – Intelligent-Tiering monitoring and automation charges: aws.amazon.com/s3/pricing*

[2]*AWS Glacier retrieval classes (Instant, Flexible, Deep Archive): docs.aws.amazon.com – Glacier storage classes; GCP Archive retrieval design: cloud.google.com/storage/docs/storage-classes*

[3]*Azure Blob Storage lifecycle management configuration: learn.microsoft.com – lifecycle management policy*

[4]*Google Cloud Storage archive retrieval: cloud.google.com/storage/docs/storage-classes*

[5]*AWS archival restore times and retention: docs.aws.amazon.com – restoring objects; Azure archive rehydration details: learn.microsoft.com – rehydrate blob data; GCP archive lifecycle/pricing: cloud.google.com/storage/docs/lifecycle*

[6]*Lifecycle automation requirements across AWS, Azure, and GCP: docs.aws.amazon.com – Intelligent-Tiering overview; learn.microsoft.com – lifecycle performance; cloud.google.com/storage/docs/lifecycle*

[7]*AWS S3 Pricing – Standard Storage, $0.023/GB-month for first 50 TB: aws.amazon.com/s3/pricing*

[8]*AWS S3 Pricing – Standard-IA: aws.amazon.com/s3/pricing*

[9] AWS S3 Pricing – Express One Zone: https://aws.amazon.com/s3/pricing

[10] AWS S3 Pricing – Standard Storage.

[11] Google Cloud Storage Pricing – Standard: cloud.google.com/storage/pricing

[12] Azure Blob Storage Pricing – Hot LRS: https://azure.microsoft.com/en-us/pricing/details/storage/blobs