Backblaze

GPUs Are Only Half the Equation

Solving the AI data throughput bottleneck

Neocloud platforms have spent the last two years competing on GPU availability, interconnects, and cluster performance. But a GPU can only work as fast as the data pipeline supporting it.

GPUs read their data from high-performance flash storage when training models or running inference. To make that possible, datasets must first move from durable object storage into that flash tier. The upstream supply layer is responsible for delivering data there quickly and consistently. If it cannot keep up, the entire pipeline slows down, no matter how powerful the GPUs may be. That's why GPU availability is only half the equation. The other half is the data supply architecture that sustains it.

Solution in brief

A storage architecture designed to keep GPU clusters fully utilized by delivering sustained data throughput into high-performance flash tiers. As AI workloads scale, the upstream data pipeline becomes just as important as the compute layer. Platforms that ensure reliable data supply can translate GPU capacity into real-world performance.

Solution elements

- The challenge

  GPUs depend on a steady flow of data staged from durable object storage into high-performance flash storage. When the upstream data pipeline cannot keep pace with AI workloads, the GPUs wait for data and performance drops.

- The shift

  Treat the upstream object storage layer as performance infrastructure rather than simply durable storage. Instead of focusing only on capacity or archival use cases, design storage to deliver sustained throughput and predictable behavior under AI workloads.

- The approach

  Backblaze B2 Neo provides a high-throughput, white label object storage platform engineered to supply flash storage tiers with reliable data delivery. With sustained performance, private connectivity options, and managed infrastructure, Backblaze enables neocloud platforms to keep GPU clusters fed without building and operating complex storage systems in-house.

What Neocloud platforms gain with B2 Neo

- Higher GPU utilization

  Ensure GPUs spend more time processing workloads and less time waiting for data.

- Faster model iteration

  Support shorter training cycles and quicker experimentation for AI teams.

- Predictable platform performance

  Deliver consistent behavior as datasets grow and clusters scale.

- Simplified infrastructure

  Avoid the operational complexity of building and maintaining large-scale storage systems.

- Economic efficiency

  Convert GPU investment into usable compute performance rather than idle capacity.

Where the AI data bottleneck actually originates

AI workloads move large amounts of data at every stage of the lifecycle.

- Training datasets must be staged and prepared so jobs can access them quickly.
- Model checkpoints, artifacts, embeddings, and intermediate outputs must be written back for durability and reuse.
- Inference pipelines generate their own steady stream of reads and writes as models serve predictions and capture outputs.

All of this activity places constant pressure on the storage and networking layers that move data through the system. The challenge becomes more pronounced as AI workloads move from small experiments to large GPU clusters, growing from a handful of GPUs to dozens or hundreds. The object storage layer must sustain significantly higher aggregate throughput while also absorbing more frequent checkpoint and artifact writes.

If storage can't keep up, more GPUs just means more waiting. And when GPUs wait for data instead of performing useful work, the impact is not just technical—it's also financial.

The cost of idle GPUs

Idle GPUs create costs for three reasons.

- **First, they waste compute.** Organizations end up paying for GPU resources that aren't being used to process AI workloads.
- **Second, they slow development.** Training runs take longer than they should, clusters stay reserved for more time, and iteration cycles stretch out. This extends time-to-market for new models and features.
- **Finally, they become a platform risk.** Customers judge a platform by how quickly their workloads run and whether performance stays consistent as they scale. If results vary from run to run, customers assume the GPUs are the problem, even when the real issue is the data pipeline.

AI redefines what object storage must deliver

Traditional object storage systems are optimized for intermittent cloud workloads where traffic spikes occasionally rather than sustaining high throughput continuously. They are typically evaluated by peak throughput or average response time, not by how reliably they sustain heavy data movement over long periods.

AI infrastructure imposes a different set of demands. To reliably supply high-performance flash tiers, object storage must:

- Sustain high aggregate throughput, not just short bursts
- Deliver predictable performance under continuous data movement
- Absorb large checkpoint and artifact writes
- Remain isolated from shared network contention
- Operate consistently at production scale

Many traditional object storage architectures were not designed for this steady, high-volume supply model. They perform well for archival storage, backups, and general-purpose cloud workloads, but AI workloads introduce sustained pressure that exposes architectural limits.

Under sustained AI workloads, common constraints begin to surface:

- Shared network paths introduce contention
- Performance varies under heavy concurrency
- Request overhead accumulates at dataset scale
- Checkpoint write activity strains upstream throughput

These issues rarely cause outright system failures. But they do introduce variability. Performance fluctuates under load and scaling becomes less predictable, so the data pipeline struggles to keep pace with the demands of large AI workloads.

The competitive implication for neoclouds

GPU availability is becoming table stakes. What separates neocloud platforms is not how many GPUs they advertise, but how reliably those GPUs translate into real-world AI performance. And that reliability depends on the strength of the upstream data supply architecture.

Customers evaluate platforms based on outcomes: how quickly models train, how stable workloads remain under scale, and how rapidly teams can iterate. If scaling a cluster produces diminishing returns or inconsistent results, your platform's reputation could suffer, regardless of how advanced your GPU infrastructure may be.

Sustainable differentiation therefore requires treating the upstream data layer as performance infrastructure rather than simply durable storage.

Backblaze B2 Neo: Storage built for AI throughput

B2 Neo provides the high-throughput, white label object storage backbone required to sustain AI workloads at production scale.

S3-compatible platform engineered for sustained performance—including up to 1 Tbps aggregate throughput—supplies high-performance flash tiers under real-world load.

Private connectivity options create dedicated, low-latency data paths that reduce shared network contention.

Durability, availability, and scale are delivered as a managed service, eliminating the need to build and operate complex storage infrastructure in-house.

Branded revenue stream opportunities are built in so neoclouds can bill for storage as a native platform service.

For neocloud platforms, that means GPU investments translate into consistent, usable performance.

GPU availability is only half the equation. Backblaze delivers the other half.

Backblaze

Learn more about partnering with Backblaze, or reach out to start a technical and strategic conversation.