GPU Liquidity: Wherever You Find GPUs, Your Data Is Already There

ANY DATA
ANY LOCATION
TOTAL CONTROL

EXECUTIVE SUMMARY

Over 200 exabytes of enterprise data still sit in data centers, while 70% of AI workflows run in the cloud. Qumulo Cloud AI Accelerator deploys in minutes and projects live data from wherever it lives to GPUs wherever you find them — no prestaging, no waiting on results.

THE CHALLENGE

Most of the world's GPU's are hosted by public-cloud providers, and generalized demand far outpaces localized supply. A pool of available GPUs may appear in a different Availability Zone (AZ) from where your data sits, or even a completely different region; and then disappear again just as quickly. The data is rarely near where the GPUs are. Organizations looking to leverage GPU cycles are left choosing between two costly options:

OPTION 1: RESERVE & WAIT

Secure scarce GPU capacity at high cost as soon as it becomes available, then pay for hours or days of idle time while the needed data copies into the right zone. This the option most choose: Reserved GPUs sit idle up to 95% of the time waiting for data to be staged.

OPTION 2 : PRE-STAGE & HOPE

Replicate the data to multiple regions and/or Availability Zones in advance, hoping that GPU resources will materialize in one of those destinations before the data goes stale — multiplying network, storage, and transaction costs with a low probability of return.

Either way, organizations spend money before work begins — a hidden "GPU-Hunting Tax" that shows up in the cloud bills, delayed projects, and missed windows.

KEY BENEFITS

With Qumulo, enterprises can capitalize on GPU liquidity anywhere:

- **On-Prem GPUs:** Maximum performance for sensitive AI workloads.
- **Cloud GPUs:** Real-time data projection without duplicating datasets.
- **Multi-Zone, Multi-Region:** 3x more opportunities to access GPU capacity
- **Burst Capacity, Instantly:** Attach data to compute in minutes with no pre-staging or replication.

Enterprises must rethink their dependence on large capital GPU purchases and build an architecture that takes advantage of compute wherever capacity becomes available.

THE SOLUTION: GPU LIQUIDITY WITH QUMULO CLOUD ACCELERATORS

Qumulo Cloud AI Accelerator eliminates the need to move data to wherever GPU capacity becomes available. Instead, it projects live enterprise data directly to cloud GPUs through Qumulo Cloud Data Fabric.

Cloud AI Accelerators deploy in minutes alongside available GPUs, presenting data locally while maintaining a single, globally consistent system of record. Intelligent local caching delivers high-performance access, while Cloud Data Fabric synchronizes changes in real time, ensuring every GPU, user, and application always works from the same authoritative dataset.

The result is true GPU liquidity. Compute moves to available capacity—not data—eliminating pre-staging, replicated datasets, and idle GPUs waiting for data transfers. Whether data resides on-premises or in the cloud, organizations can attach GPUs to live enterprise data within minutes.

KEY CAPABILITIES

- **Cloud AI Accelerator**
  - Deploy beside available GPUs in minutes.
  - Project live enterprise data directly to compute.
  - No pre-staging or migration.
- **Cloud Data Fabric**
  - One globally consistent dataset.
  - Real-time synchronization.
  - No stale copies or reconciliation.
- **NeuralCache**
  - Intelligent local caching for low-latency access.
  - Data available before it's requested.
  - Reduce cloud API and egress costs.

AI starts where GPUs are, not where data lives - enabling access without replication, migration, or waiting.

BUSINESS OUTCOMES

- **Accelerate AI time-to-insight** by eliminating data staging delays.
- **Increase GPU utilization** by attaching compute immediately to live datasets.
- **Reduce cloud costs** by eliminating replicated storage, unnecessary egress, and object transaction charges.
- **Maintain one authoritative copy** of enterprise data across on-premises infrastructure and every public cloud.
- **Increase opportunities to secure scarce GPU capacity** by treating multiple Availability Zones and Regions as a single pool of available compute.
- **Simplify AI operations** with one architecture that spans on-premises storage, cloud-native infrastructure, and every major cloud provider.

CONCLUSION

As AI adoption accelerates, organizations need to move compute—not data. Qumulo Cloud AI Accelerator enables GPU liquidity by projecting live enterprise data to available GPU resources in minutes, eliminating data movement while reducing cost, complexity, and time-to-insight.

ABOUT QUMULO

Qumulo is the leading provider of cloud file data platforms, offering unrivaled performance, scale, and data management solutions. Qumulo's platform is trusted by Fortune 500 companies and global enterprises to manage petabytes of data, enabling them to unlock the value of their data and drive innovation. For more information, visit www.qumulo.com.