SOLUTION BRIEF

Qumulo NeuralCache for Accelerating Data Access

Accelerate Data Access with Predictive Caching for Hybrid & Distributed Workloads

Qumulo NeuralCache is an intelligent prefetch and caching engine built into the Qumulo Cloud Data Fabric that dramatically accelerates application performance and responsiveness — especially for distributed, data-intensive workflows that span on-premises and cloud environments. It uses machine learning and real-time behavior analysis to predict and pre-fetch the data workloads that will be needed next, reducing latency and bandwidth dependency while improving user and application performance.

The Challenge: Latency & Bandwidth Limitations Across Hybrid Environments

Modern data workflows often span edge locations, core data centers, and multiple cloud platforms and regions. In these hybrid scenarios, traditional caching and file access patterns frequently create bottlenecks:

- Remote data access over WAN induces latency
- Large files must often be fully transferred before use
- Cloud egress and I/O costs escalate with repeated access
- Static replication policies struggle with dynamic, unpredictable workflows
- New or changed data from remote sources may take hours to appear locally

These limitations can slow critical workloads — from media editing and HPC to AI/ML model training and inference workflows — and increase infrastructure costs.

What NeuralCache Does

NeuralCache radically transforms how data is accessed and delivered across the Qumulo Cloud Data Fabric by:

Predictive Prefetching

Instead of waiting for applications to request remote data, NeuralCache monitors client access patterns to anticipate which data blocks will be needed next and pre-emptively fetches them into a local cache tier. This behavioral intelligence boosts hit rates, often above 90%, translating into faster reads and reduced network latency.

Accelerated Distributed Workflows

By moving only the data actually needed at a block level, eliminating the need for full-file downloads over WAN links, NeuralCache delivers low-latency access even to remote datasets.

Adaptive, Self-Tuning Intelligence

NeuralCache continuously refines its models based on real usage patterns — across users, machines, applications, and network conditions — enabling automatic performance optimization without manual tuning.

Optimized Cost Efficiency

Reducing redundant data transfers over WAN and minimizing cloud I/O transactions translate into lower network costs and reduced compute consumption, especially in cloud environments.

Qumulo

A Modern File Data Experience

Qumulo Run Anywhere
Unlike legacy storage systems, Qumulo is 100% software-defined, meaning it runs consistently across on-prem, cloud, or hybrid environments—on any x86/ARM server. This lets modern enterprises avoid vendor lock-in, simplify infrastructure, and deploy anywhere while maintaining a single, unified file system.

Any Data, Any Location, Total Control
Qumulo's Cloud Data Fabric removes capacity and distance limits, making any data instantly available to any workload—anywhere—while ensuring sovereignty and security.

Intelligent Performance with NeuralCache
Qumulo NeuralCache uses machine learning to optimize reads and writes, improving performance and drastically reducing cloud I/O costs by up to 99%

Any Data, Any Location, Total Control
Qumulo's Cloud Data Fabric removes capacity and distance limits, making any data instantly available to any workload—anywhere—with strict consistency while ensuring sovereignty and security.

Customer Success
Qumulo's Zero Latency support model eliminates traditional ticketing and triage systems in favor of direct access to Customer Success Managers and Qumulo engineers via Slack and Microsoft Teams. Customers have responded, giving Qumulo an NPS score of 95 in the most recent survey.

Key Benefits

1. **Near-Local Performance Anywhere**
   Remote data feels local. NeuralCache brings requested data ahead of time into local caching tiers — eliminating waits and minimizing the performance gap between local and remote data access.
2. **Broad Workflow Acceleration**
   Whether for AI/ML training, media and entertainment playback/editing, healthcare imaging, or financial analytics, NeuralCache improves throughput and responsiveness without workflow rewrites.
3. **Reduced WAN Load & Costs**
   Predictive caching minimizes unnecessary data transfers, reducing WAN utilization and associated cloud bandwidth or egress charges.
4. **Intelligent Resource Utilization**
   By anticipating access needs and caching effectively, NeuralCache reduces reliance on high-cost cloud compute during peak demand.
5. **Seamless Integration with Cloud Data Fabric**
   NeuralCache is included with Qumulo's Cloud Data Fabric — there's no separate licensing requirement. It automatically complements globally distributed namespaces and unified cache strategies.

How NeuralCache Works

- Real-time monitoring of data access at the client and application level
- Prediction of next-needed data blocks based on historical access patterns
- Prefetch of predicted blocks into local memory or cache before actual request
- Adaptive re-prediction as access patterns evolve during workflow execution

Unlike traditional data acceleration solutions that move entire files only after they are requested, NeuralCache fetches only the relevant blocks ahead of time — dramatically reducing latency and minimizing transfer inefficiencies.

Who Benefits Most

NeuralCache delivers value across any organization with distributed, data-heavy applications, including:

- AI/ML research and inference pipelines
- Media & entertainment rendering and editing
- Healthcare imaging and diagnostics
- Engineering, simulation, and high-performance workloads
- Cloud burst and hybrid cloud compute models

It drives performance without compromising global data consistency or governance.

Conclusion

Qumulo NeuralCache is a breakthrough in intelligent caching for hybrid, distributed environments. By anticipating data needs and pre-positioning the right data blocks closer to the application, it eliminates traditional bottlenecks — delivering accelerated performance, reduced costs, and broader workflow efficiency while preserving a unified global namespace and real-time collaboration across locations.

Learn more or request a demo to see how NeuralCache can transform your organization's data-driven performance.

Real-World Acceleration Examples

Workload Type	Typical Read Latency	With NeuralCache	Performance Gain
Media Editing (Autodesk Flame)	~2.6 ms	~600 µs	~4× Faster
Healthcare DICOM Viewer	~82.6 ms	~6.4 ms	~12× Faster
Energy Workflows (Petrel)	~44.1 ms	~1.7 ms	~25× Faster
PyTorch AI Reasoning	~189.3 ms	~1.0 ms	~189× Faster