Qumulo

How Does Qumulo NeuralCache Help in HPC and AI Workloads?

Supercomputing 2025

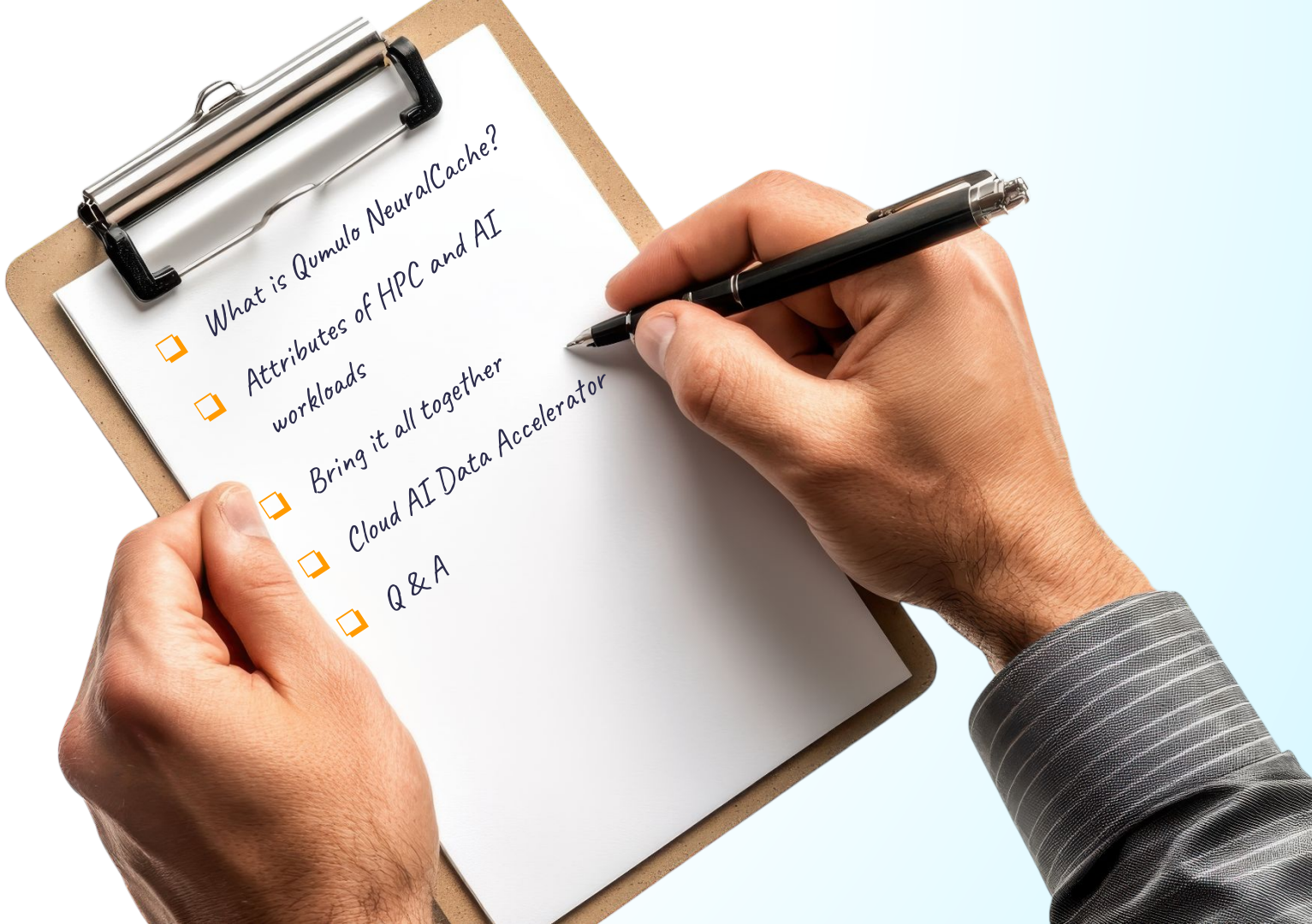
Before we get started…

Some housekeeping details

- Please feel free to ask questions along the way. We have plenty of time to answer.
- We are recording today's presentation and you can get to all the show content using the QR code here or on the booth flyer.
- Thanks for dropping by, please make sure to get scanned and enter to win the LOOI Robot- AI Desktop Companion

Agenda

What we will be covering today



Qumulo

NeuralCache

predictive caching layer that dramatically accelerates workload performance by anticipating data needs ***before*** requested

32 Trillion

parameters over 13+ years of workload I/O patterns across >**5** Exabytes

Digital Twin

Regression Training

96%

Inference Accuracy

Preemptive Caching

80%

Reduction in Cloud TCO

Remote as Local

64%

Reduction in AI GPU costs*

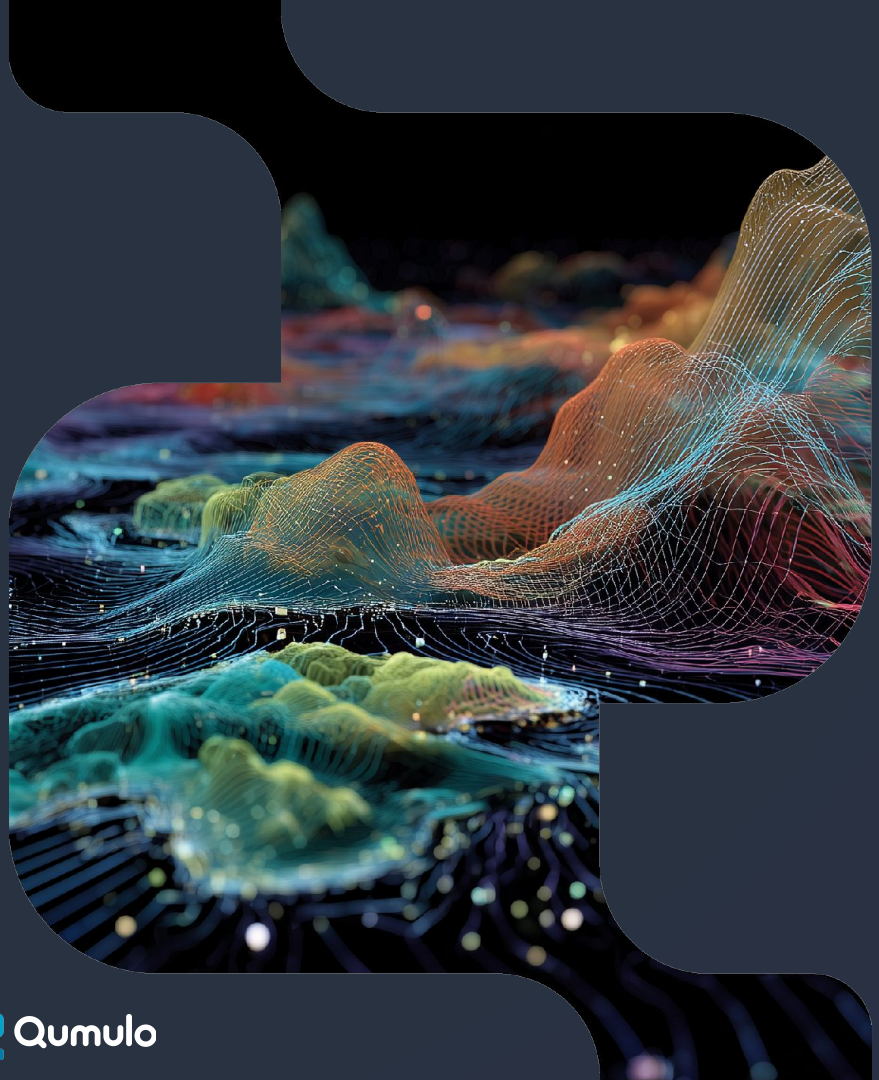
Data Center Space/Power Reduction

Qumulo

*Compared to Standard S3 AI pipelines

NeuralCache: Real World Acceleration

	NeuralCache Accelerated
M&E Autodesk Flame: Average 2.6ms / Read	Average 600µs / Read ***4x Faster*** **50% Lower TCO**
Healthcare DICOM Viewer: Average 82.6ms / Read	Average 6.4ms / Read ***12x Faster*** **80% Lower TCO**
Energy SLB Petrel: Average 44.1ms / Read	Average 1.7ms / Read ***25x Faster*** **64% Lower TCO**
PyTorch AI Reasoning Average 189.3ms / Read	Average 1.0ms / Read ***189x Faster*** **35% Lower TCO**



Making Caching Work

- Single namespace
- Decoupled performance and capacity
- Actionable visibility without expensive platform tree walks
- Optimized metadata storage and use of flash
- Strict data consistency
- Block-level streaming protocol that transfers only the exact portions of data needed

AI & HPC Workloads

- Compute intensive
- Huge datasets
- Low latency required
- High throughput
- Highly variable file size
- Large sequential reads and random access
- Long-running jobs
- Benefit from parallelism
- Benefit from prefetching

To unify the world's unstructured data, giving our customers the freedom to create, manage, and store ...any data, any location, with total control

Cloud AI Data Accelerator

- 189x faster performance for AI & HPC workloads compared to traditional cloud object storage
- Powered by NeuralCache and Qumulo's Cloud Data Platform
- Delivers a zero-copy, just-in-time data pipeline that bridges data centers, edge, and the cloud

Next Steps

Resources available to learn more

- Join us for a demonstration at our demo pod
- Reach out to our sales & solution engineering team (sales@qumulo.com)
- Learn more at www.qumulo.com
- Discovery Workshop

60-90 minute discovery workshop to review our engineering, operations, executive priorities, IT systems estate & landscape, and your current data strategy. Afterward, we will consult and provide an architectural design review, technology demonstration, and identify key workloads for your organization.

Q&A

Thank you!