QUMULO ON AWS / AZURE / GCP / OCI

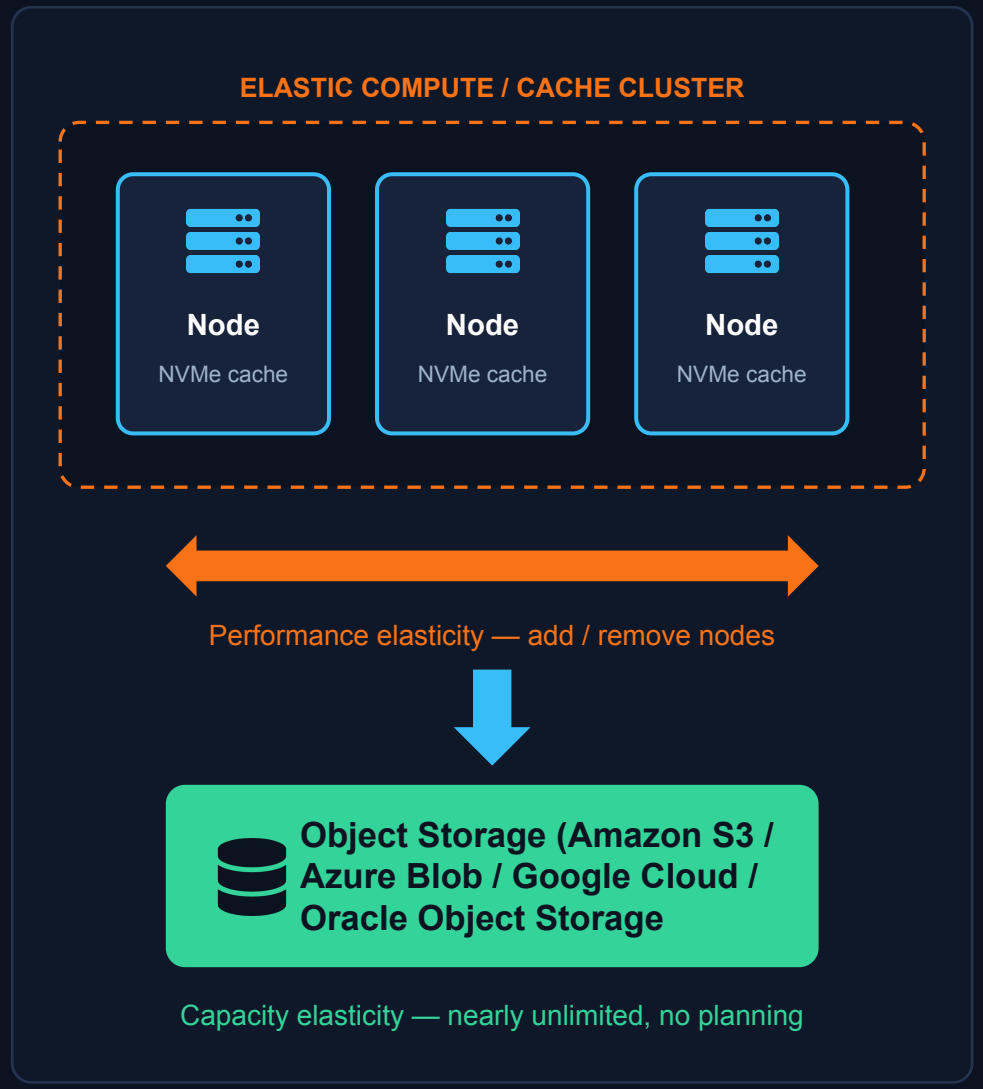
Cloud Native Qumulo

File storage re-architected for the cloud: compute fully disaggregated from storage, accelerated by an intelligent cache tier, persisted on object-storage economics.

Up to 99%
fewer object API calls

Used
capacity billing, per minute

Minutes
to scale performance



Legacy cloud file storage scales the wrong way

Every other file solution in the cloud is built on managed block disks — so compute and storage are welded together.

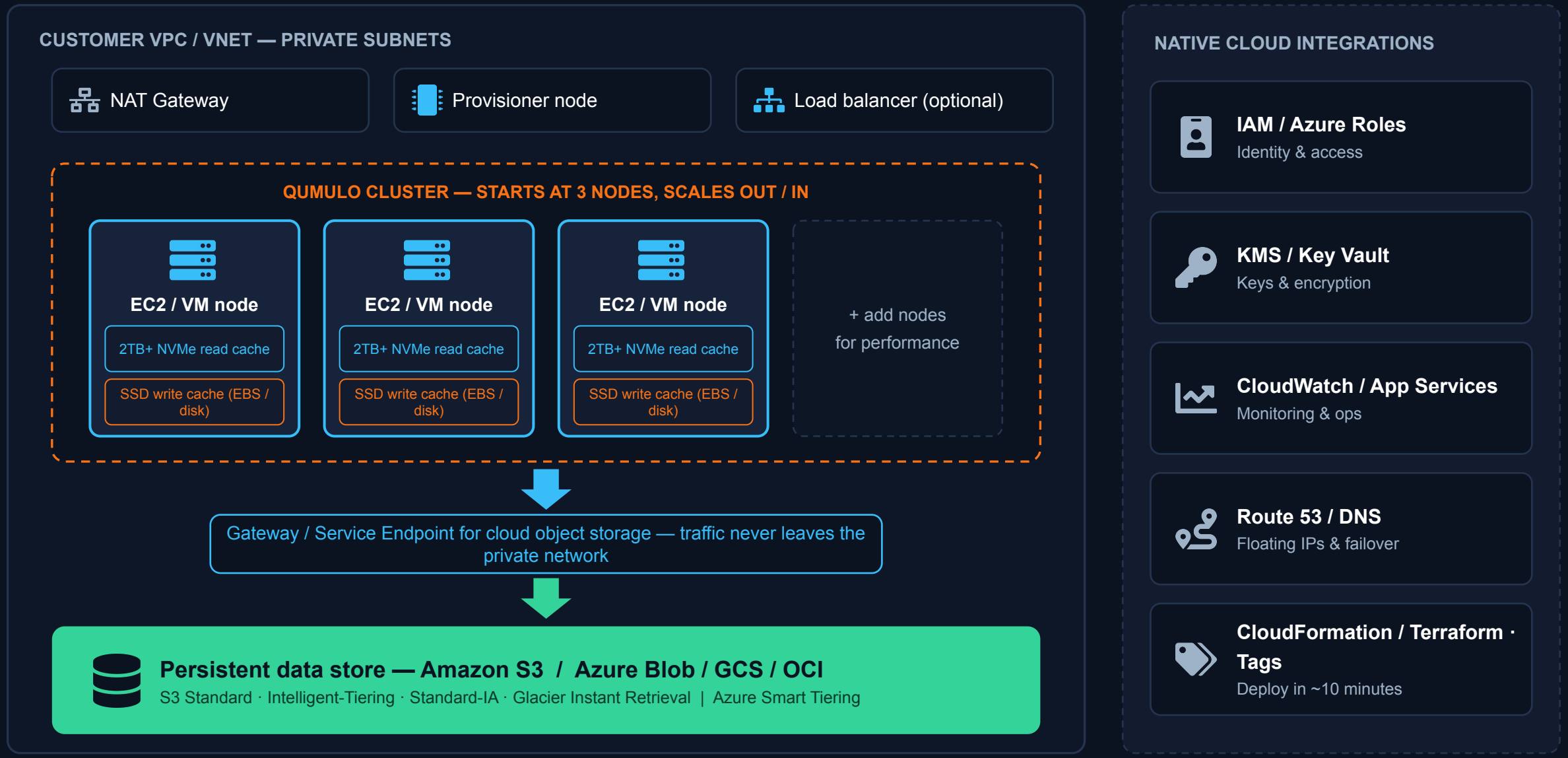
Managed-disk file solutions

- Capacity lives on provisioned block disks attached to each node, does not provide 11x9s of Availability or Durability vs Object storage
- Need more TB? Buy more nodes and disks — even when performance is already fine
- You pay for provisioned capacity, not what you actually use
- Millions of small writes hammer disks and drive up IOPS cost

Cloud Native Qumulo (CNQ)

- Compute and storage fully disaggregated — capacity lives in S3 / Azure Blob
- Performance = node count, re configure nodes with increased CPU/Memory/NVME. Add or remove Nodes in minutes, storage untouched
- Pay for used capacity, Billed per minute — no over-provisioning
- NeuralCache Coalesces many small writes into a few large object PUTs

One architecture for every public cloud



The write path — where the transformation happens

NeuralCache sits between your applications and object storage, turning file-system writes into cloud-efficient objects.

1

Ingest

Clients write over NFS, SMB or S3 to any node in the cluster.

2

Ack in sub milliseconds

The write lands in the SSD write cache (Mirrored EBS / Managed Persistent Disk) and is acknowledged immediately.

3

Compact + compress

NeuralCache aggregates, merges and compresses many small writes into fewer, larger objects.

4

Encrypt + protect

Data is encrypted with AES-256 and written in Qumulo's protected block format — never as raw files.

5

Commit to object

A few large PUTs go to the persistent object storage layer — cutting API call volume by up to 99%+.

Many small writes → few large object PUTs. API requests to object storage are a major cloud cost lever — CNQ eliminates up to 99%+ of them.

The read path — predictive by design

NeuralCache doesn't just cache what you've read — it models what your application will read next.

1

Cache hit

Hot data is served straight from the 2TB+ local NVMe or directly from system memory— flash-fast, with zero api calls to object storage.

2

Digital twin

Qumulo builds a digital twin of each application's I/O based on 14 years of telemetry and over 250 trillion data points — simulating how it accesses files and blocks.

3

Predict + prefetch

The model predicts future reads before they occur; blocks are prefetched from object storage into NVMe ahead of demand.

4

Cache miss

A rare miss triggers one large, efficient GET — the data is then promoted into cache so the next read is local. Analysis from Qumulo install base, show cache hit rates of above 99%

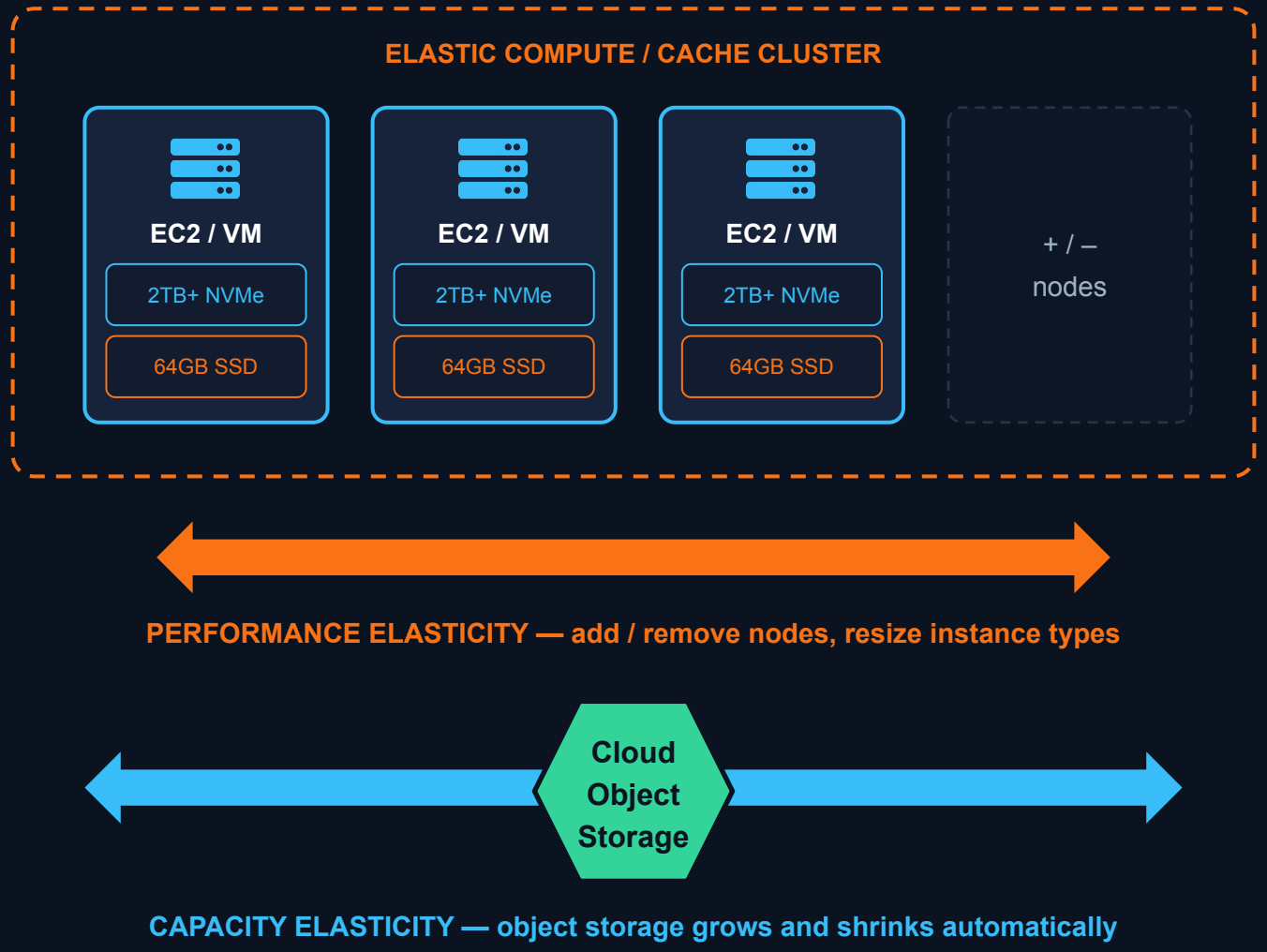
DIGITAL TWIN — ACCESS PATTERN

reading now

predicted next — already prefetched from S3 / Blob into NVMe

Reads stay at cache speed. Object storage is touched only for large, efficient GETs — latency stays low and API costs stay flat as you scale.

True elasticity — scale each dimension independently

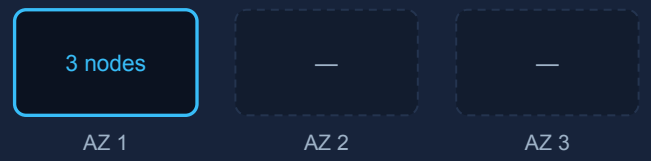


WHAT THIS MEANS FOR YOU

- Pay up front, pay only for USED capacity per minute
- Performance scales throughput and IOPS up and down as needed
- Storage is fully INDEPENDENT from compute
- Nearly unlimited storage — no capacity planning
- Single-AZ or TRUE Multi-AZ with one cluster

Start small, deploy anywhere

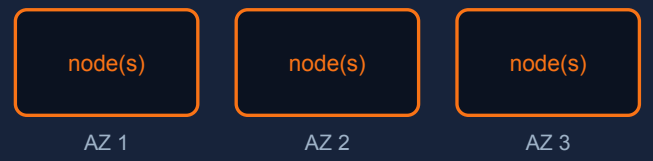
A CNQ cluster starts at just 3 EC2 / VM nodes and grows — or shrinks — from there. Choose the resilience model that fits each workload.



Single-AZ

Lowest-latency footprint for performance-critical workloads.

- 3+ nodes in one Availability Zone
- Data persisted to S3 / Blob with eleven 9s durability
- Rebuild or resize the compute layer at any time — data stays put



True Multi-AZ

One cluster spanning Availability Zones — not a mirrored second cluster.

- Nodes distributed across 3 AZs
- Survives a full AZ outage with continuous availability
- No replica cluster to buy, sync or fail over

Deployed from CloudFormation / Terraform in ~10 minutes — inside your own account, your own VPC / VNet.

Why the economics win

	Managed-disk solutions	Cloud Native Qumulo
Capacity layer	Provisioned block disks	S3 / Blob object storage
Data durability	RAID overhead you buy and manage, limited durability	Erasure coding and 11, 9s durability built into object storage automatically
Scaling	Compute + storage locked together	Fully independent
You pay for	Provisioned peak capacity	Used capacity, per minute
Small writes	Provisioned IOPS + disk sprawl	Compacted — 90%+ fewer PUTs
Growth	Forklift expansions, re-planning	Nearly unlimited, no planning

Up to 99%+

fewer object-storage API calls thanks to write-cache compaction

3 nodes

is all it takes to start — scale out / in as demand changes

Per-minute

billing on used capacity — never pay for empty disks

Object-storage economics + right-sized compute = a fraction of the TCO of any managed-disk file solution.