Cloud Native Qumulo: An Architectural Overview

Architecture White Paper

January 2025

This document describes how Cloud Native Qumulo's versatile architecture can accommodate a wide range of use cases, from the most demanding high-performance computing and artificial intelligence applications to cost-efficient cold archive storage. Its design reduces the total cost of ownership for cloud file services.

Introducing Qumulo's New Cloud Architecture

Data centers are reaching their limits. With little room for expansion, modern workloads require high-speed access to petabytes of data and immense computational power, capabilities most often found in the cloud. To mitigate risk and optimize budgets, many enterprises are embracing hybrid or multi-cloud operating models. In today's data-driven world, the efficiency and scalability of file storage solutions have become more crucial than ever.

One of the biggest challenges to large-scale cloud adoption by enterprises has been the absence of cost-effective, scalable file services among public cloud providers. File services in the cloud have lacked the rich data management features that enterprises need, and the few enterprise-class file service options in the cloud offer only limited scalability at a high cost.

Most cloud-based file services still lack an elastic consumption model, forcing customers to provision storage with mixed levels of performance and often overprovision capacity simply to meet requirements. Even with the pay-as-you-go model that cloud computing inherently suggests, file services customers are billed for the capacity they provision, not the capacity they actually use.

Cloud Native Qumulo (CNQ) integrates natively with core cloud components by utilizing low-level object storage calls deep within the file system, which has eliminated the need for performance tiering altogether. Unlike other cloud file storage solutions, CNQ customers pay only for the capacity and performance they actually use, thanks to its fully elastic design. It delivers low-latency file access due to architectural components like its Intelligent Cache Manager, and allows customers to elastically change the file system's performance capabilities real-time and without disruption.

Qumulo completely re-engineered the data persistence layer in the file system for the cloud. We leverage the cloud provider's native object storage services for cost-effective storage, their disk services for low latency durable write-back caching, and local NVMe for high performance read caching. This approach provides built-in performance, cost-effectiveness, high availability, resiliency, elasticity, and security. Combined with CNQ's architecture, performance can now be dynamically adjusted up or down by adding, changing, or removing nodes without disruption. This process takes only minutes, as there is no need to restripe or reprotect data. CNQ can deliver any performance at any capacity.

The key to understanding how CNQ's architecture stands apart starts with recognizing the differences between scale-up and scale-out architectures. Scale-up architectures, in use for over 30 years, rely on centralized controllers. When capacity or performance needs to change in a scale-up architecture, customers need to add storage capacity and/or compute capability to the controllers. This often leads to limitations in scalability, efficiency, and increased cost. In

contrast, scale-out architectures distribute workloads across multiple nodes, allowing for linear scaling of performance. The scale-out approach eliminates bottlenecks, enhances availability, and enables cloud-like elasticity.

CNQ is a scale-out architecture built on top of an elastic cloud infrastructure. Its architecture inherently allows traditional scale-up capability as well. This combination enables customers to scale in any direction: in and out, up and down, and even adjust the size of the caches as needed. Furthermore, these changes can be made non-disruptively and very quickly.

Scale-up and Scale-out architectural differences

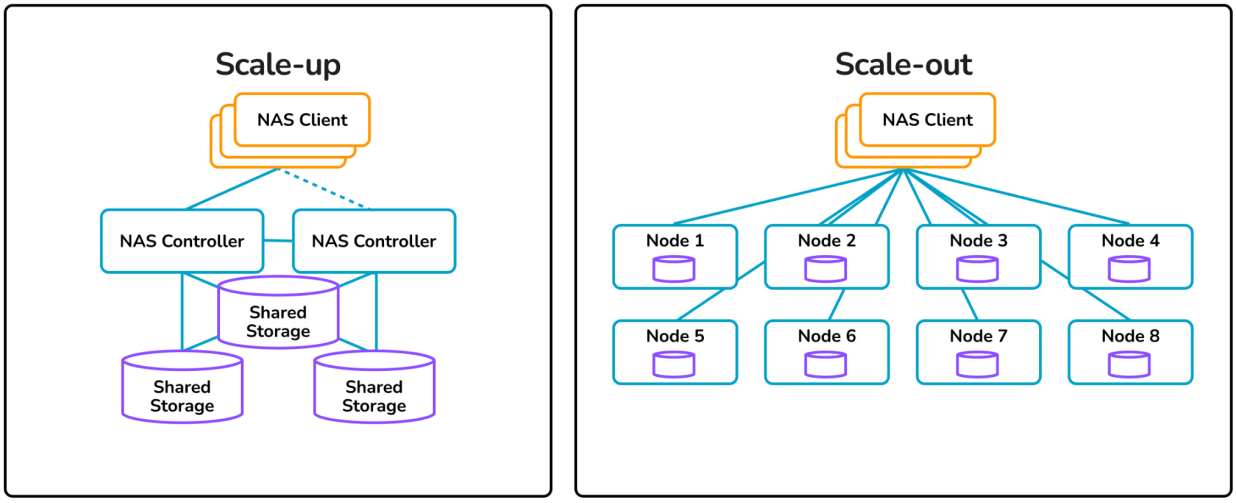

Figure 1: Scale-Up and Scale-Out Architectural Differences

For additional details on scale-up versus scale-out storage architectures, see the Appendix at the end of this document.

Redefining Scale-Out with Cloud Native Architecture

Qumulo's cloud native architecture redefines cloud storage by completely decoupling capacity from performance. Unlike other cloud storage file systems that are architecturally locked into a "performance per capacity" model, CNQ allows these to be adjusted independently. This provides the flexibility to change underlying components, such as the compute instance type, compute instance count, and cache disk capacity, allowing for rapid and non-disruptive performance adjustments. This architecture, which includes the innovative Predictive Cache, enables exceptional elasticity and virtually unlimited capacity. It provides an adaptive storage

platform, ensuring that businesses can efficiently manage and scale their data storage as their needs evolve, without compromising on performance or reliability.

CNQ retains all the core Qumulo functionalities, including real-time analytics, robust data protection, security, and global collaboration. Its architecture is fully integrated into the cloud's elastic resource model, providing exceptional flexibility and efficiency. This makes CNQ ideal for both hybrid-cloud and multi-cloud enterprises. Let's explore further.

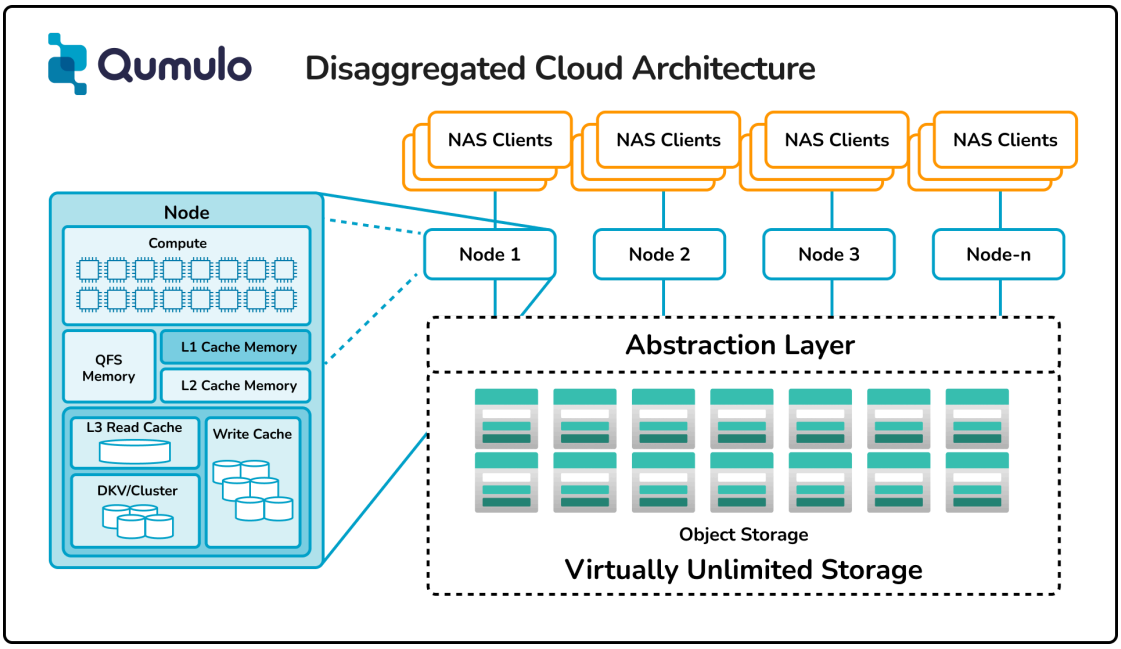

Figure 2: Cloud Native Qumulo Architecture

Dynamic Performance Scalability

Cloud Native Qumulo's architecture expands the scale-out model by allowing adjustments in multiple dimensions without service disruption. By dynamically changing underlying cloud components, such as the compute instance or the configuration of the write-cache volumes, a CNQ cloud instance can scale up or down in real-time to meet changing workload demands. The traditional scale-out method of adding nodes still exists, but now compute instances can also be removed or resized to fine-tune performance in minutes. Beyond adjusting the compute layer, additional high-performance write cache disks can be added or increased in size to further improve read and write performance.

CNQ adapts readily to accommodate various workloads and respond to changing demands. This adaptability, combined with the ability to make adjustments without interrupting service, ensures continuous availability and efficiency for enterprise applications, even under rapidly changing conditions.

Data resiliency and encryption in the cloud

CNQ leverages the inherent data resiliency of cloud hyperscalers' object storage services, eliminating the need for additional data-layer protection (such as erasure coding) and data rebalancing across nodes. The data layer operates independently of individual CNQ nodes, further enhancing efficiency.

By eliminating data restriping and rebalancing processes, CNQ's modular architecture improves storage efficiency and performance, allowing enterprises to scale resources up or down rapidly.

For example, customers can add performance for a nightly batch job and then reduce it before morning, minimizing costs and providing a true cloud experience: pay only for the performance you use, only when you need it.

Not only does Cloud Native Qumulo's architecture provide seamless integration with existing workflows and environments, it also supports all of the same core features as an on-prem Qumulo deployment. Each CNQ instance is designed to handle massive unstructured data workloads and support exabyte-scale data within a single namespace. With real-time analytics, high performance, data protection, simplified management, multi-protocol support, security, global collaboration, and API-driven capabilities, users can dynamically scale storage capacity and performance to meet their evolving needs.

This scalability is particularly valuable for industries with large datasets and specialized workflows, such as media and entertainment, healthcare, and scientific research. With file data in the cloud, these organizations can leverage cloud-based resources like artificial intelligence, machine learning, big-data analytics, and managed container services.

Beyond that, with the cost–effective file services that Cloud Native Qumulo offers, all enterprises can use the cloud's unlimited scalability to relieve the pressure on their on-premises data centers and colocation facilities by moving massive amounts of cold data to long-term archive services on Azure or AWS.

Decoupling Performance from Capacity

Many cloud-based file services have a fixed relationship between IOPS and storage capacity. This often forces users to overprovision capacity to reach specific performance targets, increasing costs and leading to wasted resources. These services often have fixed limits on the size of a single file system, typically ranging from 100-500 TiB. Workloads exceeding these limits must be split across multiple instances, complicating management and requiring ongoing rebalancing. CNQ overcomes these limitations by leveraging object storage directly instead of dedicated virtual disks. This approach decouples capacity from performance, allowing users to tailor each to their specific needs.

With CNQ, customers can meet IOPS and throughput targets without overprovisioning capacity. This flexibility ensures that both performance-intensive and capacity-intensive applications can operate efficiently. Qumulo's advanced data management features, such as real-time analytics and comprehensive data protection, further enhance its value for modern data-driven enterprises.

Inside the Architecture

Cloud Native Qumulo’s architecture enables unparalleled elasticity, and a virtually unlimited capacity. Qumulo’s Predictive Cache provides an adaptive storage platform that ensures businesses can efficiently manage and scale their data storage as their needs evolve without compromising on performance or reliability. Continuous availability and operational efficiency are supported across a wide range of scenarios, such as adding compute instances for scale-out expansion, reconfiguring nodes with more compute resources, or dynamically increasing or decreasing their read-cache size. All these adjustments can be performed without disrupting existing clients, and without requiring additional storage capacity or data rebalancing.

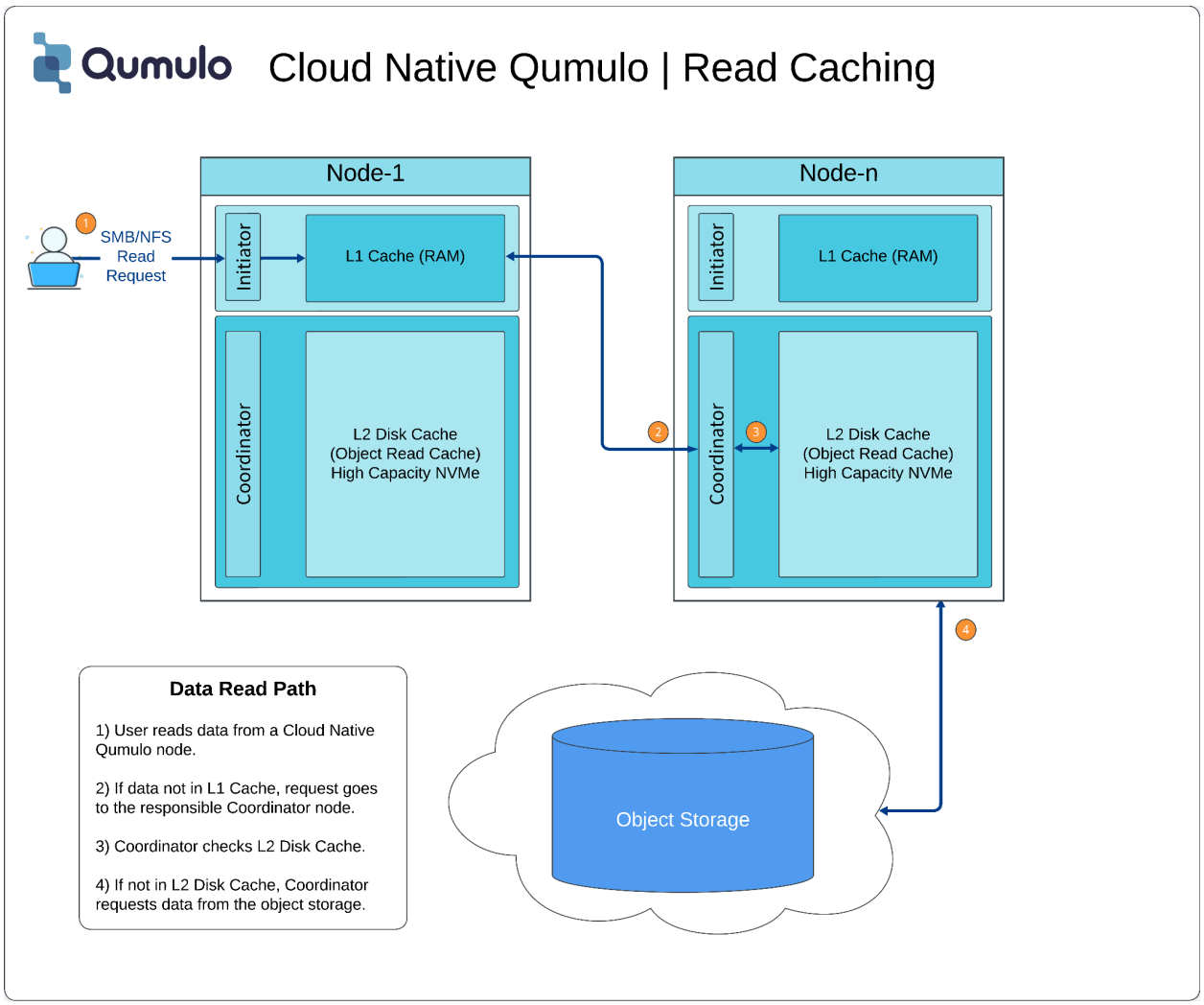

Figure 3: Cloud Native Qumulo | Read Caching

Initiator

Every node in a CNQ cluster functions as the initiator for its connected clients. When a request is received, the initiator first checks the L1 RAM cache, which operates at nanosecond speeds. This cache stores both metadata and application data and satisfies most read requests, minimizing I/O latency and improving performance. Qumulo's Predictive Cache may pre-emptively copy data that has a high probability of being imminently read from the L2 Disk Cache into the L1 RAM cache.

Since multiple clients on the same node often request the same data, the L1 cache acts as an efficiency multiplier.

Coordinator

The Coordinator and Predictive Cache work together to manage the L2 Disk Cache and all related caching operations. The Predictive Cache analyzes information from real-time activities, such as data access time, source, and category (e.g., file table, directories, file names), to make informed caching and pre-caching decisions. After the Predictive Cache makes decisions, the coordinator for that data prioritizes promoting that data to the L2 cache against all of the activity reported by other nodes. By prioritizing caching predictions in real-time, the Coordinator enables predictive caching. This minimizes latency and improves read performance.

L2 Disk Cache

The L2 Disk Cache, managed by the Predictive Cache, significantly increases caching capacity by utilizing high-speed, high-capacity NVMe disks. It offers over 100 times the capacity of the RAM cache. The Predictive Cache analyzes various factors, including access patterns and historical data, to determine which blocks are best suited for caching in the L2 cache. This analysis goes beyond basic sequential I/O patterns and can even optimize I/O patterns that initially appear random. In real-world workloads, many seemingly random I/O patterns become predictable over time. Others have true randomness within working sets small enough to be loaded into the L2 disk cache but too large for the L1 RAM cache. This predictive capability reduces read-request latency and allows CNQ to deliver data stored in objects at near solid-state speeds.

Object Storage

Cloud Native Qumulo persists data in a cloud-native manner by writing it to objects in the cloud provider's native object storage service. This pay-for-use model aligns storage costs with actual consumption, eliminating the use of fixed volumes or pre-provisioned capacity common to other cloud file services. The use of object storage for the persistent data layer ensures high availability and data durability, making it a highly reliable and scalable solution for enterprises with dynamic or large-capacity storage needs. CNQ also inherits the security characteristics of cloud object stores, including built-in server-side encryption (SSE) using FIPS 140-2 compliant AES encryption.

Write Cache

The Write Cache coalesces small write operations into a contiguous object, resulting in a single large write to the object store. This results in fewer PUT transactions, which increases performance and lowers cost.

Write operations are journaled in a write-ahead log residing on mirrored, cloud-protected SSDs to ensure durability, recovery, and fault tolerance. This approach prevents data loss during platform disruptions, providing protection for recently written data.

Distributed Key-Value Store

The Distributed Key-Value Store (DKV) is a critical component of every CNQ instance. It resides on a small disk (EBS/ManagedDisk) attached to each node (see Figure 2). The DKV stores cluster-specific configuration information essential for platform operation and management. This ensures that all nodes in the cluster can access the necessary configuration data, allowing for efficient coordination and management across the distributed system.

How CNQ Uses Object Storage

Unlike many other cloud storage file systems that use a "tier-to-object" approach, Cloud Native Qumulo persists all data directly to object storage. The file system is designed to anticipate which data segments to promote to the L2 cache. Each object contains file and metadata blocks from potentially many files, similar to how a sector on a hard disk drive stores data. Internal indexes and data maps are also stored within objects, ensuring structural integrity.

When a non-cached data segment is requested, the coordinator can leverage parallel object reads, accessing hundreds of objects simultaneously. The file system intelligently retrieves portions of an object or the entire object based on the request. The Predictive Cache guides these decisions using deterministic mechanisms built into the file system, allowing for parallelized reads and hiding object latency even when the workload itself provides limited parallelism. This approach optimizes data access and retrieval, improving performance across diverse workloads

Dynamic Performance Scaling

Cloud Native Qumulo's performance scales up or down by adding or removing compute instances, providing typical benefits of a scale-out architecture. This approach enhances both IOPS and throughput capabilities incrementally and reliably, bolstering the cluster's performance in a modular fashion. By distributing the workload across multiple compute instances, it can efficiently handle increased demand without the risk of overloading a single node. This method also allows for flexible scaling, where resources can be added as needed, minimizing downtime and ensuring continuous operation. Resources may be removed as

needed, allowing for additional cost savings. Easy removal of resources reduces operational burden by making all decisions to add more resources reversible.

CNQ's architecture also offers the ability to increase (or decrease) the size of each compute instance. Customers can add CPU cores, increase memory, add higher-capacity SSD or NVMe L2 cache disks, and enable higher network limits without disrupting services to existing clients or workloads. The ability to right-size the compute instances dynamically means CNQ can scale up quickly to accommodate more demanding workloads, and then easily scale back down again during slow periods or as demands on the service decrease. Compute instances can be initially sized based on upfront performance estimates, and then adjusted up or down as needs evolve.

Predictive Cache

A key component of Cloud Native Qumulo's read performance includes the Predictive Cache, an intelligent caching algorithm that monitors client requests and overall data-access patterns, anticipating which data blocks are likely to be requested and prefetching that data into the cache layer before clients actually request it. This feature results in significant performance improvements for both sequential and non-sequential workloads.

Object-Level Caching

The Predictive Cache's direct interaction with object data results in a substantial read performance gain. Key technical benefits include:

Compression: Compression allows for a larger working set using a very efficient on-disk representation.

Concurrency: Prefetch and promotion occurs outside of global locking, maximizing concurrency.

Comprehensive Cache: A single unified cache efficiently stores both data, metadata, and meta-metadata (internal indexes that support the object layer).

Rapid Rehydration: Data can be rehydrated quickly in the case of failure, upgrade, routine maintenance, or scaling events.

Prefetching

The system has been enhanced with more efficient prefetching capabilities. It currently identifies access patterns based on file offset locality (ex: streaming or random access within a section), or file-in-directory access order (ex: directory copy). These patterns allow the

Predictive Cache to anticipate and prefetch data the user will likely need soon. This aggressive prefetching prioritizes data based on predicted distance from the next read, ensuring the most relevant data is readily available. The system is designed to easily incorporate new patterns in the future, allowing it to adapt to evolving workloads.

Heat Scoring

A multi-dimensional heat scoring mechanism ensures that frequently accessed data remains in the cache. This "heat" tracking system considers several factors, including:

- Activity generation (roughly time-based) of last access or cache entry.
- Source of the data (e.g., write cache, commit, hydration).
- Data category (e.g., system metadata, directories, file metadata, and file data).

Cache Allowances

The Predictive Cache utilizes internal allowances for various data categories. If a category exceeds its allowance, the coldest data from that category is compared against data from other categories for potential eviction. Most customers' metadata should fit within the metadata allowance, ensuring it stays cached. For customers with many billions of files, the coldest metadata may be evicted, depending on how large of a cache is provisioned.

Benefits

Increased performance

- CNQ delivers faster data access and improved overall performance for both sequential and non-sequential workloads.
- Prefetch and promotion of data from object storage to the cache can happen outside of any global locks, enabling maximum data concurrency.

Greater efficiency

- CNQ can manage data, metadata, and meta-metadata (internal indexes of object data) in a single unified cache layer.
- Data in the cache is compressed, allowing for a larger working dataset to reside in the available cache capacity. .

Improved scalability

- CNQ can support larger datasets and higher workloads without performance degradation. This is partly due to the overhead being proportional to the cache capacity, not the cluster capacity.

Faster recoverability

- The read cache is ephemeral but easy to reconstruct, so it can be quickly rehydrated in the event of failures, Qumulo Core upgrades, addition of new nodes (e.g. performance scaling or adding nodes every morning to improve performance), etc.

Future-ready design

- The Predictive Cache is designed to adapt to evolving requirements and new technologies.

What Changed?

Qumulo re-engineered the read cache subsystem, resulting in the **Predictive Cache**, a system optimized for object storage and designed to meet the demands of sequential and non-sequential workloads. This redesign involved several key changes. First, the cache now operates directly on object data, reducing latency and improving efficiency. Second, prefetching capabilities have been enhanced, enabling the cache to anticipate and proactively retrieve data based on observed access patterns. A sophisticated heat scoring mechanism prioritizes frequently accessed data, ensuring it remains readily available. The system also incorporates adaptive cache allowances to manage the cache effectively, even for file systems containing billions of files. These changes collectively deliver a more performant, scalable, and efficient caching system, ready to adapt to future demands.

Minimizing Object Transaction Costs

While the data layer primarily resides on object storage, CNQ's architecture minimizes transaction costs by intelligently placing data in the most appropriate layer of the cluster's caching hierarchy. Writes are first committed to the Write-Ahead Log (WAL) of a mirrored write cache spanning at least two nodes. The write cache consists of 2-6 mirrored volumes (see Figure 2) per node. These writes are committed to disk before being acknowledged to the client. The file system consolidates writes of varying sizes into larger objects to minimize the number of PUT operations, which helps to control costs.

By leveraging cached data and efficiently managing write operations, CNQ significantly reduces the number of PUT requests required to satisfy front-end IOPS demand. This reduction

can range from 95% to as high as 99.5% by consolidating both large and small I/O operations into single PUT transactions. For example, in a VDI profiles scenario, a peak demand of 64,000 front-end write operations might be reduced to just 300 PUT requests to the underlying object storage layer. Fewer PUT operations ultimately leads to lower costs.

The data in Table 1 below was derived by analyzing workloads on existing CNQ instances, aggregating cluster performance data reported to Nexus.

Workload Type	Peak IOPS	PUT Calls/s	Reduction
VDI Workload	64,000	300	99.50%
M&E Shared Editorial	3,400	15	99.50%
Simulation/Testing	1,000	50	95%

Table 1: PUT Request Reduction (Writes)[1]

Qumulo instances have an average read hit ratio of ~93% globally, resulting in exceptional read performance. While the predictive prefetcher is primarily designed to enhance performance, workloads with strong data access locality can naturally reduce the number of GET requests to the object store, leading to cost savings. The actual savings, however, will vary based on the specific workload.

For example, as shown in Table 2, an M&E shared editorial workflow with a peak front-end requirement of 3,700 IOPS resulted in only 400 GET requests, representing an 89.19% reduction. In contrast, a VDI workload, characterized by more random I/O, saw a 12% reduction in GET requests.

Workload Type	Peak IOPS	GET Calls/s	Reduction
VDI Workload	25,000	22,000	12%
M&E Shared Editorial	3,700	400	89.19%
Simulation/Testing	10,000	190	98.10%

[1] Based on internal telemetry compiled from Cloud Native Qumulo instances running in customer environments. Qumulo does not have access to customer-facing data on any CNQ deployment.

Table 2: GET Request Reduction (Reads)

This is just one of the reasons Cloud Native Qumulo is more cost-effective than other options: it was designed for the cloud and engineered at every layer to maximize performance and flexibility while minimizing costs.

Conclusion

Cloud Native Qumulo is an advancement in cloud file storage, redefining scalability, performance, and flexibility to support modern workloads. By decoupling capacity from performance, CNQ enables exceptional adaptability, allowing for non-disruptive adjustments to compute instances and cache disk capacities. Designed to support industries with both massive datasets and high-performance requirements, CNQ's architecture retains Qumulo's core functionalities, including: real-time analytics, robust data protection, and global collaboration. This new approach, along with the innovative Predictive Cache capability, sets a new standard for cloud storage offerings. Cloud Native Qumulo offers a comprehensive solution designed to meet the evolving needs of data-driven enterprises, allowing them to confidently embrace the future of cloud technology.

Contributors

This article is maintained by Qumulo. It was originally written by the following contributors.

Principal authors:

Kevin McDonald (KMac) | Principal Technical Marketing Engineer at Qumulo

James Walkenhorst | Sr. Technical Marketing Engineer at Qumulo

Related Resources

Qumulo Technical Overview

Qumulo Resource Library

Azure Native Qumulo Security - White Paper

Qumulo Exabyte-scale file storage on Azure

Azure Native Qumulo Cold

Qumulo Scale Anywhere

Qumulo Nexus

Appendix: Scale-up vs. Scale-Out Storage

Comparison of Scale-Up and Scale-Out Architectures

Understanding the differences between traditional scale-up and modern scale-out architectures is essential for appreciating the advancements in the Cloud Native Qumulo file system. In this section, we explore these fundamental differences. Scale-up architectures, characterized by a central NAS controller, often face scalability constraints, high initial costs, and potential single points of failure. In contrast, scale-out architectures distribute workloads across multiple nodes, enhancing reliability and scalability. This approach reduces costs, allows for seamless maintenance, and dynamically manages resources efficiently.

Legacy Scale-Up Architecture

In a scale-up NAS system, when customers need to add performance or capacity, the system's architecture forces difficult decisions. Additional storage capacity is added to NAS controllers, which are configured in pairs (either active/passive or active/active). This approach can lead to performance bottlenecks and limitations as the controller becomes both a single point of failure and a performance limiter.

The main drawback of scaling up is that all workload and data requests rely on a single server's resources. As demands on storage systems grow, this single server struggles to keep up with the increasing workload and will eventually reach the limit of its capabilities. This not only impacts performance but also introduces significant risks associated with hardware failures and system downtimes.

To support the level of scale at the level of performance that modern workflows need, scale-up storage customers must add more scale-up storage instances. These new instances add cost, while the need to continually rebalance data across the array of scale-up storage controllers adds complexity, and the inevitable stranding of unused capacity that results from such a fragmented architecture adds waste.

Legacy Scale-Out Architecture

In a scale-out NAS architecture, multiple nodes (in the form of either servers or virtual machines) are added to the system to increase both storage capacity and performance in building block increments. Each node functions as a storage device, a storage controller, and an access point, effectively distributing the workload evenly across the entire system. This approach eliminates the single-server bottleneck issue present in scale-up architectures, as the

workload is balanced across all nodes. Each node independently contributes to the system's overall performance and capacity, ensuring that as more nodes are added, both capacity and performance increase accordingly.

The linear scalability of scale-out architectures makes them ideal for environments with growing and unpredictable storage demands, offering flexibility and expandability. On the flip side, that also means that the amount of compute available to the system is directly related to the number of nodes in the system.

At first glance, porting a legacy scale-out storage architecture to the cloud is a much simpler process than deploying scale-up storage, because scale-out storage nodes can be deployed using cloud-based VMs, each with its own complement of compute and network resources, and with some number of dedicated virtual drives to provide the capacity. In scale-up architectures, there's still a direct correlation between compute and storage: if you need more performance, you have to add more capacity whether you use it or not. Additionally, the virtual drives in this type of solution must be deployed using block storage (EBS / ManagedDisk), which is much more expensive per GiB than object (S3/Blob)storage.

Another downside to using a legacy scale out storage in the cloud: new nodes are added to the array, the existing data needs to be re-protected and re-balanced across the expanded capacity. The rebalancing process consumes system resources and is time-consuming, often limiting other system capabilities during restriping and rebalancing operations.

Scale-up and scale-out storage architectures represent fundamentally different approaches to expanding storage performance and capacity. Certain vendors employ a hybrid architecture, combining scale-up and scale-out concepts by integrating multiple scale-up components into a single cluster. While this approach shares similarities with scale-out, it is not architecturally equivalent and still has a real performance and capacity ceiling.