KIOXIA

Performance Brief

KumoScale™ Software Improves Flash Performance at Data Center Scale compared to Ceph™ Software

In combination with the NVMe-oF™ Protocol, Delivers Disaggregated Network Storage with Performance Comparable to Locally-Attached Storage

The NVMe-oF specification makes it possible to provision networked storage with performance comparable to locally-attached[1] NVMe™ SSDs, and paves the way for true workload mobility while enabling 'run anywhere' scheduling of stateful applications. It delivers low-latency access across the network, as well as data replication across multiple availability zones, which has typically been a capability that was only available from the largest of public cloud providers. The performance benefits that the NVMe-oF protocol delivers take advantage of flash-based storage versus protocols designed for hard drives.

Many data centers have and continue to deploy Ceph storage, an open-source software storage platform that implements object storage on a distributed computer cluster using commodity hardware. It is based on RADOS (Reliable Autonomic Distributed Object Store) which converts data into objects and stores them in a distributed manner across multiple servers. The distributed nature of the Ceph architecture tends to make it relatively slower in a flash storage environment because every I/O requires multiple touches by system CPUs calculating data locations, multiple network hops, and many messages in order to achieve consensus across multiple servers on the correctness of data. And, since Ceph resiliency resides at the target, it requires several synchronous write operations to other targets in order to achieve a resilient state, which can also slow performance.

The NVMe-oF protocol improves on these Ceph architectural limitations as it is designed for today's flash-based storage versus yesterday's hard drive storage. It is fast and parallelized, and takes advantage of fast 100 to 200 gigabit Ethernet (GbE) network paths. One main benefit of the NVMe-oF protocol is that I/O access requires absolutely no CPU processing because it bypasses the Linux® operating system's I/O stack entirely. Once a client connects to a volume on a target, the data path performs at full NVMe-oF protocol speed, enabling a native NVMe-oF software-defined storage (SDS) solution, such as KumoScale software, to provide next generation data center storage performance.

Developed by KIOXIA Corporation (formerly Toshiba Memory Corporation), KumoScale software is a high-performance storage platform. It includes a virtualization layer of disaggregated storage delivering performance comparable to locally-attached[1] NVMe drives and the flexibility of virtualized resource management designed for data center scale. The software automates the provisioning and management of storage resources to dynamically connect the right amount of shared, resilient storage to individual workloads, and uses analytics to optimize storage utilization across the entire data center.

KumoScale software is an alternative to Ceph software and it is representative of the performance gains that can be achieved by moving to an SDS platform built natively on top of the NVMe-oF protocol. The test comparisons[2] utilized an identical benchmark process, server clusters and SSDs.

Test Environment

Both KumoScale software (version 3.13) and Ceph software (version 14.2.11) were tested in a networked environment that included the NVMe-oF specification (version 1.0a) and a TCP/IP transport. The test environment consisted of three storage nodes - each containing five KIOXIA enterprise CM5 Series PCIe® Gen3 SSDs, with 3.84 terabyte[3] (TB) capacities (Figure 1), ~20TB of storage per node.

The test stimulus and associated measurements were provided by four test clients. All storage nodes utilized 100 gigabit per second (Gb/s) Ethernet speed via a single 100GbE

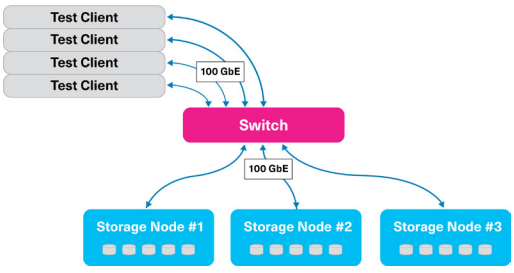

Figure 1: the software test environment

network switch. An identical hardware configuration was used for both software tests, and four logical 200GiB volumes were created - one assigned to each of the test clients.

As part of the testing process, data replication for each was enabled as follows:

For KumoScale software, volumes were triply replicated (i.e., a single logical volume with one replica mapped to each of three storage nodes).

For Ceph software, volumes were sharded (horizontally partitioned) using the 'CRUSH' hashing scheme, with a replication factor set to three.

Test Description

The test process used for both KumoScale and Ceph software included the following workload description and expected behavior:

Workload Metrics:

- *100% random, 4kB reads and 100% random, 4kB writes*
- *Load is varied by controlling the total number of outstanding read commands (queue depth)*
- *Record latency (both average and 99th percentile) and input/output operations per second (IOPS) tests against queue depth*

Expected Behavior:

- *As the queue depth is increased initially, the IOPS rate grows linearly while the latency remains constant or grows slowly. The result is that the load is absorbing inherent parallelism in the storage system.*
- *At some queue depth, the latency begins to increase while the IOPS rate stops increasing. This is an indication that the storage system has reached its maximum performance, so queuing up more read commands will simply increase the wait time.*

*Disclaimer: Unlike the performance test conducted, data center loads typically are not limited by their outstanding queue depth, and issue commands at an arbitrary rate. If this rate exceeds the maximum IOPS capability of the storage system, issued commands will rapidly begin to fail. Data center operators **MUST** ensure that storage systems are operated below their maximum IOPS limit, but should be loaded as close to that capability in order to achieve maximum capital efficiency.*

Performance Results: IOPS

The overall IOPS performance test results for both KumoScale and Ceph software (Figure 2) now follows:

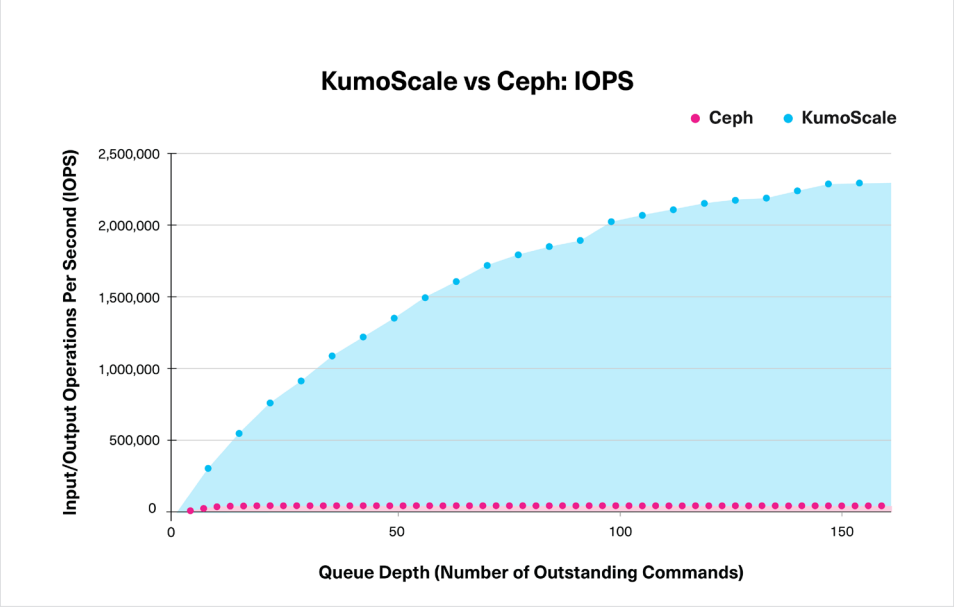

Figure 2: KumoScale vs Ceph IOPS test results

The KumoScale solution delivered more than 52x the IOPS performance of Ceph software on identical hardware (Table 1). For this performance metric, higher is better.

Test Metric	KumoScale	Ceph	KumoScale Advantage
IOPS at maximum load	2,293,860	43,852	>52x

Table 1: IOPS performance test results comparing KumoScale software to Ceph software

Performance Results: Read Latency

The read latency test results for KumoScale and Ceph software (Figure 3) were as follows:

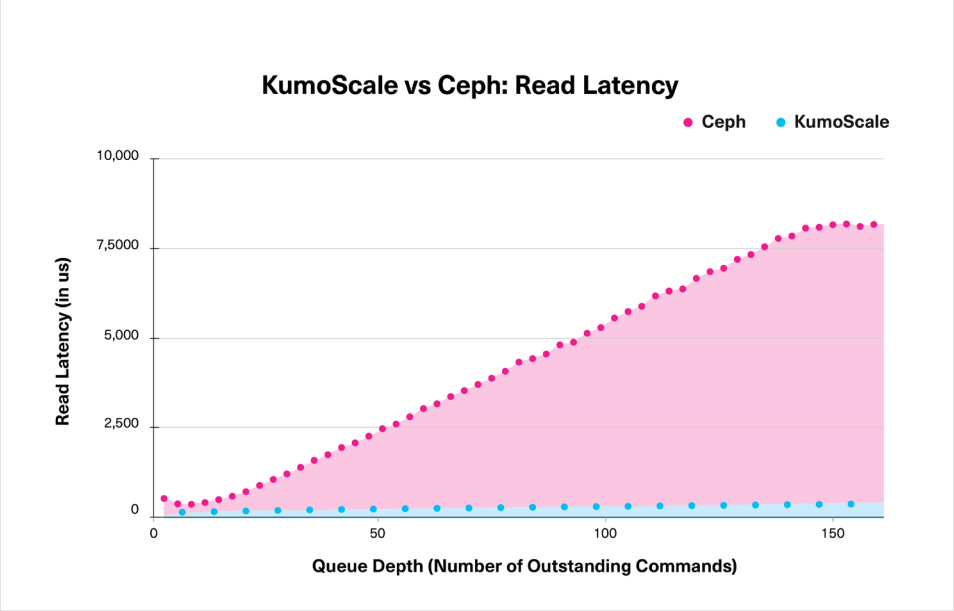

Figure 3: KumoScale vs Ceph read latency test results

At a queue depth of ~150, the read latency of the KumoScale solution was more than 22x faster than Ceph software (Table 2). For this latency metric, lower is better.

Test Metric	KumoScale	Ceph	KumoScale Advantage
Read latency at maximum load	369µs	8,169µs	>22x

Table 2: Read latency test results comparing KumoScale software to Ceph software

The KumoScale Solution:

- *Reached a peak of ~2.29 million IOPS at a queue depth of ~150*
- *At a queue depth of ~150, read latency held steady at 369us*

Ceph Software:

- *Reached a peak of ~43.8k IOPS at a queue depth of ~150*
- *At a queue depth of ~150, read latency hovered at 8,169us*

Performance Results: Write Latency

The write latency test results for KumoScale and Ceph software (Figure 4) were as follows:

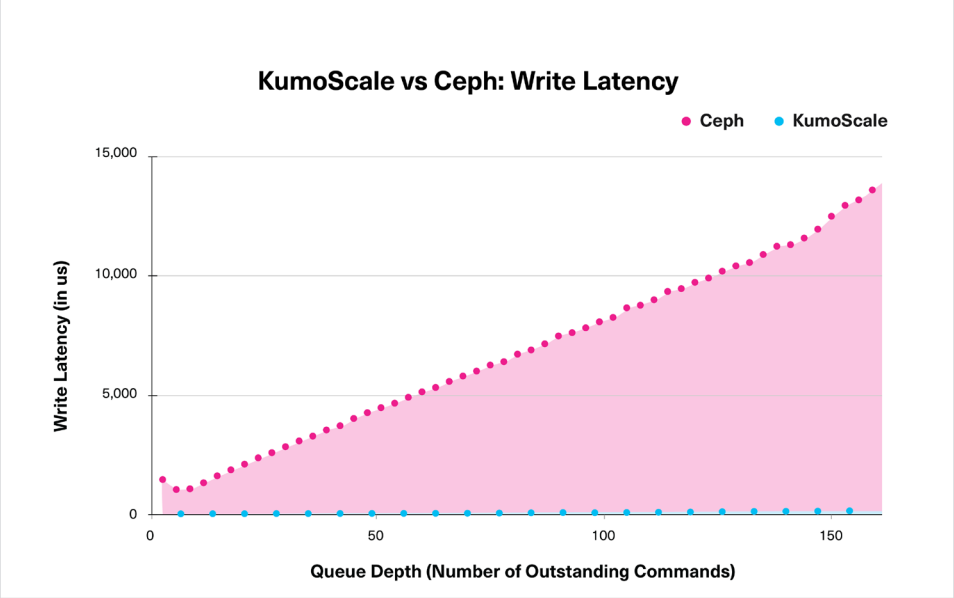

Figure 4: KumoScale vs Ceph write latency test results

KumoScale software write latency was more than 81x faster than Ceph software (Table 3). For this latency metric, lower is better.

Test Metric	KumoScale	Ceph	KumoScale Advantage
Write latency at maximum load	167µs	13,610µs	>81x

Table 3: Write latency test results comparing KumoScale software to Ceph software

The KumoScale Solution:

- *Reached a peak of ~2.29 million IOPS at a queue depth of ~150*
- *At a queue depth of ~150, write latency held steady at 167us*
- *Write performance was limited to test storage nodes containing only five SSDs each. These KIOXIA CM5 SSDs drives support a minimum write IOPS of 160k each, delivering an aggregate total of 800k for the five drives. At maximum load, 2.293k was achieved but expect that a full complement of 24 drives could reach 5 to 6 million write IOPS.*

Ceph Software:

- *Reached a peak of ~43.8k IOPS at a queue depth of ~150*
- *At a queue depth of ~150, write latency was 12,509us and growing*
- *The maximum write throughput was very low as a result of extremely long write-acknowledgements affecting latency*

Economic Impact

The read and write performance test results can be fully appreciated by calculating their respective economic impacts in a data center relating to the number of users that a single node can support (as depicted in Table 1). The ability to support more users per storage node directly translates into reduced capital expenditures per user. Based on the IOPS results presented above, KumoScale software can support about 52x more clients per storage node than Ceph software, at a much lower latency, and requires a lot fewer storage nodes to purchase resulting in a real economic return on investment in the data center.

Summary

This comparison highlights the raw performance capabilities of KumoScale software and showcases the economic benefits that this heightened performance can deliver to data center users. The solution takes full advantage of the NVMe-oF specification's architectural advantages for performance and throughput, and then adds data center scale virtualization, automation, provisioning, management and optimization of storage resources to deliver an almost 52x utilization advantage based on the IOPS results presented above, which may equate to lower operating costs and improved customer satisfaction.

For more information regarding KumoScale software, visit https://kumoscale.kioxia.com/.

Notes:

[1] Based on published performance specifications from NVMe SSD vendors as of September 2020.

[2] The performance comparison testing was conducted by KIOXIA Corporation in October 2019 and utilized an identical testing process, as well as identical server clusters and five identical KIOXIA enterprise CM5 Series PCIe Gen3 SSDs with 4TB capacities.

[3] Definition of capacity - KIOXIA Corporation defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1Gbit = 2^{30} bits = 1,073,741,824 bits, 1GB = 2^{30} bytes = 1,073,741,824 bytes and 1TB = 2^{40} bytes = 1,099,511,627,776 bytes and therefore shows less storage capacity. Available storage capacity (including examples of various media files) will vary based on file size, formatting, settings, software and operating system, and/or pre-installed software applications, or media content. Actual formatted capacity may vary.