Tame Tomorrow's Data Growth Today with Ceph® Storage and Micron 9100® MAX NVMe SSDs

Better Building Blocks for Better Ceph Solutions

Private clouds, big data, real-time sensors, self-monitoring and self-reporting devices combined with ever-changing archival requirements all add up: we are generating and capturing new data at unprecedented rates.

Virtualized environments, media streaming, cloud-based infrastructures and a more distributed workforce need continuous access to that data — at the speed of now.

In this technical brief, we see how a configuration consisting of Micron® 9100 MAX NVMe™ SSDs and Red Hat® Ceph Storage enables phenomenal performance (across demanding workloads), delivering the small, random IOPS (read, write and mixed I/O loading), throughput and streaming capabilities needed for massive, complex data sets.

We will also see how its granular scaling (based on 1U standard servers) lets you easily scale a cluster to meet your needs.

Figure 1 shows an overview of the tested Ceph cluster. Performance, configuration and testing details are also provided.

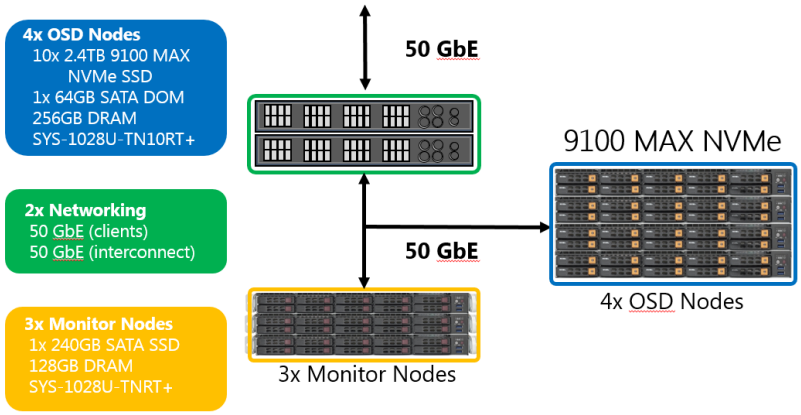


Figure 1: Ceph Configuration

Impressive 4KB Random IOPS with Low Average Latency

Virtualized environments are very demanding, and their highly random, small I/O size storage profile can be very difficult — legacy storage platforms have a hard time keeping up.

When gauging virtualization's I/O performance requirements for any storage platform[1], 4KB random IOPS are an important metric. We measured IOPS performance across three workloads on each of three 4-OSD node configurations.

Workloads:

- 100% read
- 70% read and 30% write
- 100% write

OSD Node Configurations:

- Four OSD node configurations each containing two Micron® 9100 MAX NVMe SSDs in each OSD node.
- Four OSD node configurations each containing four Micron 9100 MAX NVMe SSDs in each OSD node.
- Four OSD node configurations each containing 10 Micron 9100 MAX NVMe SSDs in each OSD node.

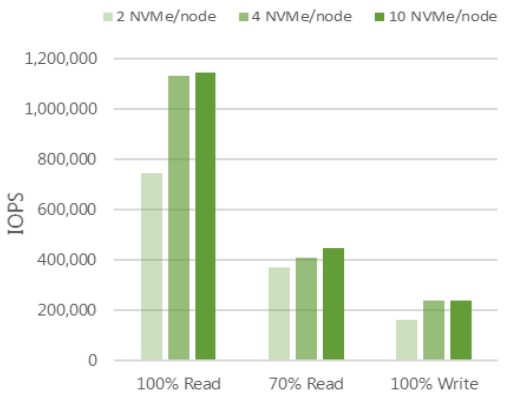


Figure 2: 4KB IOPS

4KB Average Latency	100% Read	70% Read	100% Write
2 NVMe/node	2.1ms	1.3ms/73µs	60µs
4 NVMe/node	1.4ms	1.1ms/87µs	50µs
10 NVMe/node	1.1ms	1.0ms/104µs	50µs

Table 1: 4KB Average Latency

These three configurations were used for all tests in this document and are referred to as the 2 NVMe/node, 4 NVMe/node and 10 NVMe/node configurations, respectively.

We ran each test for 15 minutes (repeated three times to ensure consistent performance), then averaged the results across all runs. Figure 2 shows the IOPS results. Table 1 shows the average latency results.

The 2 NVMe/node configuration reached 745,400 4KB random read IOPS at 2.1ms average latency. With the mixed 70% read/30% write workload, it reached 372,052 IOPS at 1.3ms (read) and 73µs (write) average latency. Its 100% write IOPS were also impressive at 163,300 IOPS at 60µs average latency.

The 4 NVMe/node configuration reached 1,134,700 4KB random read IOPS at 1.38ms average latency. With the mixed 70% read/30% write workload, it reached 418,070 IOPS at 1.1ms (read) and 87µs (write) average latency. Its 100% write IOPS were also impressive at 240,200 IOPS at 50µs average latency.

The 10 NVMe/node configuration was even more impressive. This configuration reached 1,148,000 4KB random read IOPS at 1.11ms average latency. When we measured its 70/30 mixed workload IOPS, we saw 447,823 at 1ms (read) and 104µs (write) average latency. Finally, its 100% write IOPS measured 241,500 at 50µs average latency.

1. For additional information on using 4KB random I/O as a demanding workload, see this paper on www.vmware.com.

Throughput Exceeding 21 GB/s Read and 4.5 GB/s Write

While small random I/O performance is critical to some Ceph deployments, for object workloads, a larger I/O size is far more important. Because Ceph excels at both, we measured both. This section highlights the 4MB object I/O capabilities of all three configurations.

We used RADOS Bench to test the object API performance of Ceph. This test simulates an application written to interface directly with Ceph (the test uses a 4MB object I/O).

We again ran each test for 15 minutes, then repeated the process three times and averaged the results. As noted earlier, this helps ensure consistent, more accurate results. We tested with the same configurations used in the 4KB tests. Figure 3 shows the throughput (in GB/s) of each configuration. Table 2 shows the average latency.

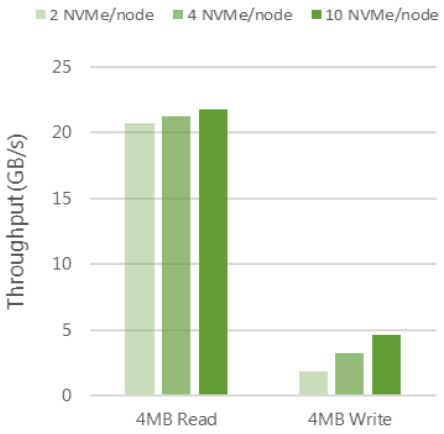


Figure 3: 4MB Object I/O

The 2 NVMe/node configuration measured 20.7 GB/s read (37ms average latency) and 1.8 GB/s write (140ms average latency).

The 4 NVMe/node configuration showed slightly improved read speed, measuring 21.2 GB/s at 36ms average latency. Its write capability was significantly improved over the 2 NVMe/node configuration, measuring 3.2GB/s at 81ms average latency.

The 10 NVMe/node configuration measured 21.8 GB/s read (35ms average latency) and 4.6 GB/s write (41ms average latency).

4MB Average Latency	100% Read	100% Write
2 NVMe/node	37ms	140ms
4 NVMe/node	36ms	81ms
10 NVMe/node	35ms	41ms

Table 2: 4MB Object Average Latency

Video Streaming: 7,100+ Ultra-High Definition Streams

We also calculated the video streaming capability of all configurations. We used 3, 5 and 25 megabits per second (Mb/s) bandwidth per standard definition (SD), high definition (HD) and ultra-high definition (UHD) streams to calculate the number of simultaneous SD, HD and UHD streams (respectively)[1] each configuration could support. Note that this is a read workload and that the results in Figure 4 are calculated based on measured throughput and documented stream requirements.

The 2 NVMe/node configuration supports 56,525 SD, 33,915 HD or 6,783 UHD streams. The 4 NVMe/node configuration offers more total bandwidth, so it supports more streams (as expected): 57,890 SD, 34,734 HD or 6,947 UHD streams. The 10 NVMe/node configuration (highest bandwidth) supports 59,529 SD, 35,717 HD or 7,143 UD streams.

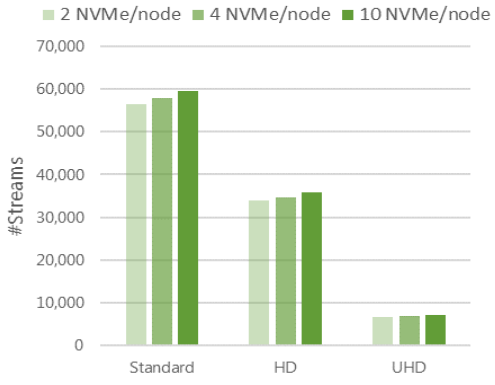


Figure 4: Video Streaming (Calculated)

1. Based on bandwidth requirements for streaming data from Netflix® Internet Connection Speed Recommendations.

Summary

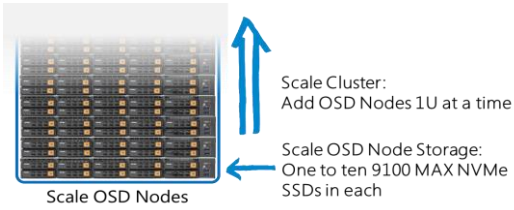
The rate of data growth is staggering. We see private clouds, big data, real-time sensors, self-monitoring and self-reporting devices gathering more and more data all the time. When we combine this data deluge with ever-changing archival requirements and users' demands for faster access, the scope of managing storage and access is daunting.

In this Technical Brief, we saw how an all-NVMe Ceph configuration built on standard 1U servers and our 9100 MAX U.2 NVMe SSDs enables phenomenal IOPS and throughput as well as granular scale out/scale up to help you deploy a Ceph cluster to meet your needs.

Standard 1U Building Blocks

We designed this Ceph cluster with standard 1U servers (supporting as many as 10 NVMe SSDs in each OSD node). This approach offers a key advantage compared with other options (like using a 2U chassis). With a 1U OSD node and the capability to use from 1-10 NVMe SSDs in each, the cluster can be easily scaled to match exact needs (capacity, IOPS or GB/s performance, available rack space, or some combination of these).

The figure below shows how we can do this. We can scale each node out by adding 9100 MAX NVMe SSDs to each node (up to 10 per node). We can also scale the entire cluster up by adding more OSD nodes – 1U at a time[1]. This is a very granular building block, enabling us to tune the cluster's capability efficiently.



1. Because the number of monitor nodes is fixed at 3, additional OSD nodes effectively scale out the cluster.

How We Tested

A Ceph storage cluster may be accessed by a variety of drivers and APIs. For this test, the focus was on the RADOS Block Driver (RBD) and RADOS Bench. RBD is used by applications to access Ceph as a block device. Images are defined within a pool in Ceph, then presented to specific clients to mount as block devices. This way, any standard Linux® server can use Ceph storage for any application.

FIO is a common tool for testing storage performance. Version 2.16-37 was used in testing along with the RBD plugin from Librbd-dev (0.1.9). This allows FIO clients to write to Ceph block images directly using the Linux RBD driver, similar to the method used by a cloud platform like OpenStack®.

4KB Testing Details

We ran each test for 15 minutes, repeating the test sequence three times to ensure consistent performance. We tested with two drives per storage node (eight drives total), four drives per storage node (16 drives total), and 10 drives per storage node (40 drives total).

We measured 4KB random block performance using FIO against the RADOS Block Driver.

4MB Testing Details

RADOS Bench tests the object API performance of Ceph. It simulates an application written to interface directly with Ceph. Each test was run for 15 minutes, three times to ensure consistent performance. Tests were run with two drives per storage node (eight drives total), four drives per storage node (16 drives total), and 10 drives per storage node (40 drives total).

We measured 4MB object performance using the RADOS Bench tool included in Ceph. We calculated video streaming capability based on the 4MB object read performance and the bandwidth per stream requirements from the cited Netflix requirements.