Tame Tomorrow's Data Growth Today with Ceph Storage and NVMe SSDs

Better Building Blocks for Scalable Ceph Storage

Data is growing faster than ever. With more sources, more complexity and more Exabytes produced every day (in 2016 Northeastern University estimated we produce 2.5 Exabytes every day!),[1] we need better solutions to manage the data deluge, better ways to prepare for tomorrow's data today.

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, capturing, managing and extracting value from new data at an unprecedented rate.

Virtualized environments, media streaming, cloud-based infrastructures and a more distributed workforce need continuous access to that data — and they need it fast.

Ceph Storage can help you manage that growth. Scalable, high-performance platforms with NVMe SSDs can help you manage it better.

We generated more than 1.3 million 4K random reads and 250,000 random writes with a 4-nodes test cluster using standard 1U servers, six Micron® NVMe™ SSDs per node and Red Hat® Ceph v3.0 (beta) on Red Hat Enterprise Linux (RHEL) 7.4.

Figure 1 shows an overview of our tested Ceph cluster's performance. With just four 1U server nodes and six NVMe SSDs in each node, the cluster easily scales up and scales out, helping tame tomorrow's data growth today.

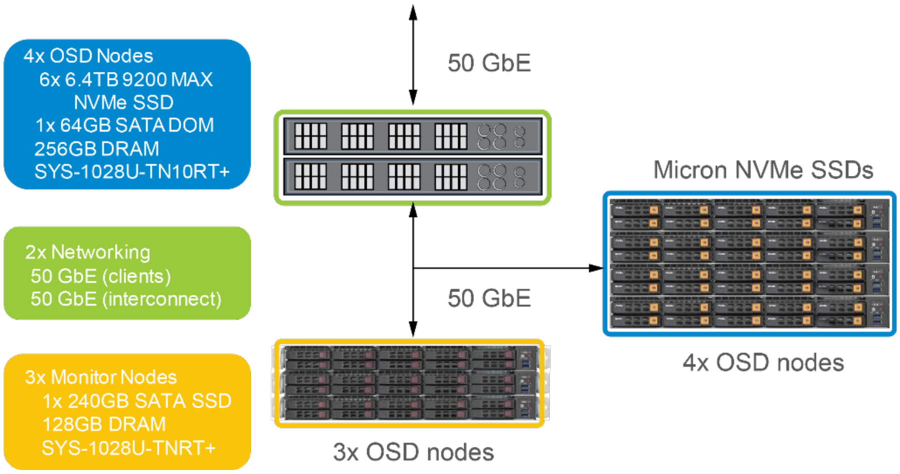

Figure 1: Ceph Storage Cluster Configuration

1.3 Million 4KB Random Read, 250,00 Random Write IOPS with Low, Consistent Latency

Virtualized environments tax storage; their highly random, small I/O size storage profile is extremely demanding. Legacy platforms have a hard time keeping up. When gauging virtualization's I/O performance requirements for any storage platform, 4K random IOPS is an important metric. [2]

We measured 4K random read IOPS performance average and 95% latencies against 100 RBD images to ensure the entire dataset was tested (see How We Tested), scaling the queue depth per FIO process from one to 32. We also measured 4K random write IOPS performance, average and 95% latencies by scaling the number of FIO clients writing to a unique RBD image per client (at queue depth 32). These FIO clients were distributed evenly across all ten load generating servers.

Figures 2 and 3 show these results.

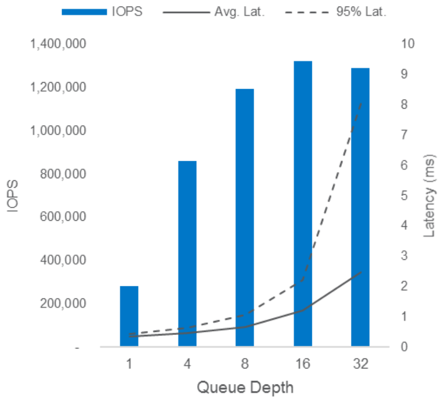

Figure 2: 4KB Random Read

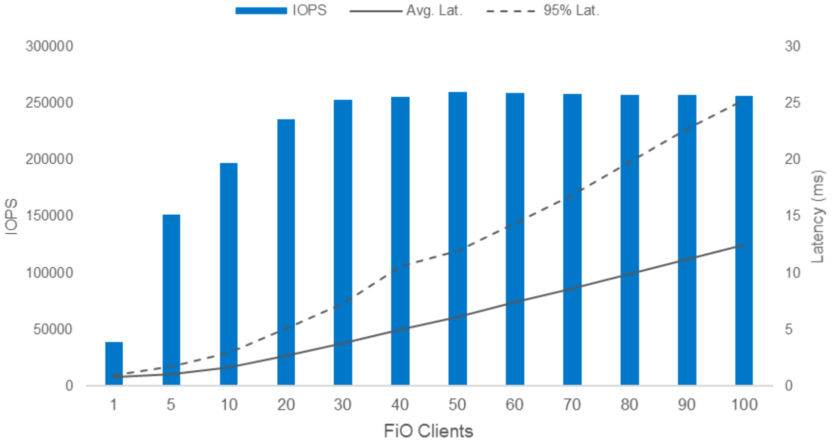

Figure 3: 4KB Random Write

4K Random Read IOPS (Figure 2): Random Read IOPS peaks when the queue depth is 16, after which there is no additional improvement. This is due to high CPU utilization, limitating results beyond queue depth 16 (at queue depth 16, CPU utilization reached 90%).

4K Random Write IOPS (Figure 3): These results are also limited by CPU utilization, reaching a peak (and plateau) at 30 FIO clients. Figure 3 shows that there is no performance gained above 30 FIO clients, while average and 95% latency continue to climb.

Despite high CPU utilization, the results are still very impressive: Read IOPS reached just over 1.3 milllion at 1.2ms average latency while Write IOPS peaked just under 254,000 IOPS at 3.8ms average latency.

Immense Throughput

While small random I/O performance is critical to some Ceph deployments, a larger I/O size is far more important for others. Because Ceph excels at both, we measured both.

We used RADOS bench to test the object API performance of Ceph. This test uses a 4MB I/O and simulates an application that interfaces directly with Ceph. Note that RADOS bench does not account for any RADOS Gateway overhead (RADOS bench represents best case results for Ceph object read and write performance. Depending on actual RADOS Gateway overhead, your results may differ).

We measured 4MB I/O with the same configuration we used for the 4KB random IOPS tests. Figures 4 and 5 show object IO performance (GB/s) and average latency.

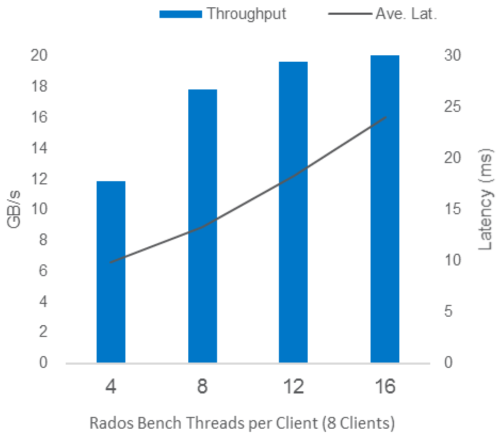

Figure 4: 4MB Object Read

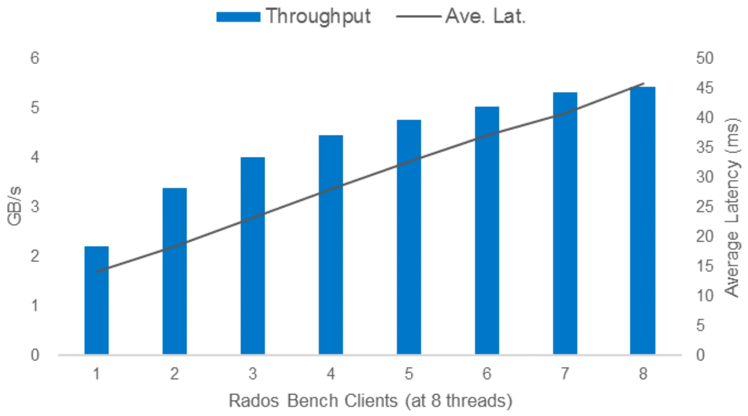

Figure 5: 4MB Object Write

4MB Object Read (Figure 4): 4MB object read performance is limited by network bandwidth. With 12 RADOS bench threads, the read through exceeds 20 GB/s (whichis very close to the theoretical limit of the Ethernet network interfaces in the storage nodes). Beyond 12 RADOS bench clients we saw no additional performance (while average latency continued to rise). When reading 20 GB/s, CPU utilization remained very low (about 20%).

4MB Object Write (Figure 5): Ceph filestore overhead limits 4MB object write performance. For each 4MB object write, the storage node issues two journal writes and two flushes to the OSD data partition — a total of 16MB written. CPU utillization was very low (about 30%).

Also very impressive was that the read throughput reached just over 20 GB/s at 18.3ms average latency, and write throughput peaked just under 5.5 GB/s at 41ms average latency.

Video Streaming: 6,600+ Ultra-High Definition Streams

We also calculated the platform's video streaming capability using 3, 5 and 25 Mb/s of bandwidth per standard definition (SD), high definition (HD) and ultra-high definition (UHD) streams.

Note that video streaming is a read workload and that these results are calculated based on measured throughput and documented stream requirements[3].

Table 1 shows the calculated number of streams.

Stream Type	Supported Streams (calculated)
Standard Definition (SD)	55,000+
High Definition (HD)	33,000+
Ultra-high Definition (UHD)	6,600+

Table 1: Streaming

Advantages of Standard 1U Building Blocks

We designed this Ceph cluster with standard 1U servers (supporting as many as 10X NVMe SSDs in each storage 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).

Figure 6 shows how we can do this. We can scale each node out by adding 9200 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.

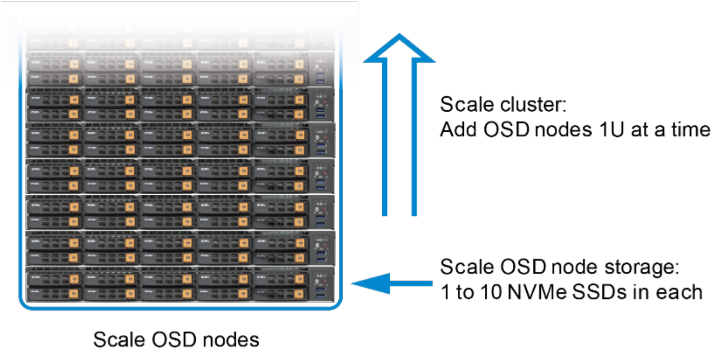

Figure 6: Building Blocks

How We Tested

We configured Ceph to use Filestore with 2 OSDs per Micron 9200MAX NVMe SSD and used a 20GB journal for each OSD. Each storage node had six 9200 MAX SSDs with NVMe. With two OSDs per SSD, Ceph had 48 total OSDs and 138TB of usable capacity. We created the Ceph pool and tested with 8192 placement groups and 2X replication.

- We measured 4KB random I/O performance using FIO against the RADOS block driver. We created 100 RBD images, each 50GB in size. This resulted in a dataset of 5TB (10TB with 2x replication). We ran each test for ten minutes.
- 4KB Random Read IOPS: using a client queue depth of 16, the Ceph storage nodes' average CPU utilization reached 90%+, limiting performance.
- 4KB Random Write IOPS: We saw an optimal mix of IOPS and latency with 30 FIO clients (measuring 254K IOPS and 3.8ms average latency). This combination generated an average CPU utilization on the Ceph storage nodes, which was over 90%, limiting performance.
- We measured the object API performance of Ceph using RADOS bench to simulate an application written to interface directly with Ceph. We noted that RADOS bench did not account for RADOS Gateway overhead and represented the best-case performance for a Ceph object workload. Each test was run for 10 minutes
- 4MB Object Read: We measured 4MB Object Read performance by reading from a 5TB dataset across eight clients, scaling the number of threads used by RADOS bench.
- 4MB Object Write: We measured 4MB object write performance by scaling the number of clients running RADOS bench (8 threads) against the Ceph Node Cluster.

A note on tuning options (threads, queue depth, image size and count, clients, and others): RADOS bench simulates traffic from a RADOS Object Gateway. There are multiple tuning options that are adjustable to enable optimal performance across a wide variety of deployments. Specific tuning settings may be deployment-specific and are typically adjusted for optimal results.

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.

We have shown how an all-NVMe Ceph configuration built using Micron® SSDs with NVMe™ and Red Hat® Ceph v3.0 (beta) on Red Hat Enterprise Linux (RHEL) 7.4 enables phenomenal IOPS and GB/s as well as granular scale out/scale up with standard, 1U servers.

Using 6 drives per storage node, this architecture has a usable storage capacity of 138TB that can be scaled by adding up to 4 drives per storage node, by adding storage nodes or both.

Ceph 3.0 is still beta, and there may be changes to the software that could affect your results, but the results we obtained are compelling. We encourage you to test this software package using our SSDs with NVMe to see the benefits you could reach with this combination and explore tuning options for best results in your deployment.

This capability and flexibility can help you deploy a Ceph cluster to meet your needs.

1. http://www.northeastern.edu/levelblog/2016/05/13/how-much-data-produced-every-day/.
2. For additional information on using 4KiB random I/O as a demanding workload, see VMware vSphere Virtual Machine Encryption Performance.
3. Based on bandwidth requirements for streaming data from this Netflix document.

Products are warranted only to meet Micron's production data sheet specifications. Products, programs and specifications are subject to change without notice. Dates are estimates only. This technical marketing brief is published by Micron and has not been authorized, sponsored, or otherwise approved by Red Hat or NVM Express. ©2017 Micron Technology, Inc. All rights reserved. All information herein is provided on an "AS IS" basis without warranties of any kind. Micron, the Micron logo, and all other Micron trademarks are the property of Micron Technology, Inc. NVMe is a trademark of NVM Express, Inc. Ceph and Red Hat are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries. All other trademarks are property of their respective owners. Rev. A 4/17 CCMMD-TBD