Low-Latency NVMe™ SSDs Unlock High-Performance, Fault-Tolerant Ceph® Object Stores

Micron® 7450 NVMe SSDs Enable High Performance With Erasure Coding for Optimal TCO

Finding the best data protection and performance using Micron SSDs

With the introduction of fast, secure SSDs and modern software, object storage solutions are no longer considered "too slow" for most applications. All-flash object storage solutions now offer the scalability, enhanced security, and performance[1] that applications need.

One popular storage solution for object storage is Red Hat Ceph Storage (RHCS), a scalable, simplified, open storage solution for modern data-centric applications—from artificial intelligence and machine learning to data analytics and emerging cloud solutions. RHCS offers a broad suite of access protocols, including block, file, and object interfaces. As an object storage solution, RHCS is a core component of Red Hat OpenStack Platform and Red Hat OpenShift® Data Foundation, enabling cloud integration through OpenStack® Swift™ and Amazon Simple Storage Service™ (Amazon S3).

Micron designs SSDs for almost all use cases with a broad range of form factors and capacities—and we designed the 7450 NVMe SSD with a low-latency architecture for workloads like RHCS. Micron has worked extensively with Red Hat to build a deep understanding of SSD performance within RHCS.

Based on our experience, one of the key configuration steps in deploying RHCS for object storage is to identify the right balance of data protection, cost, capacity, and performance for your environment. This is much simpler when pairing new RHCS capabilities, such as erasure coding, with low-latency, high-performance NVMe™ SSDs from Micron.

Key Benefits

Low-latency, high-performance NVMe SSDs enable data protection and performance in the same object store.

Erasure coding (EC) works great with NVMe SSDs, optimizing both performance and total cost of ownership (TCO). EC stores data differently than replication. EC breaks an object into chunks—data chunks and coding chunks. Data chunks and coding chunks are then stored on different physical storage devices.

If a failure occurs, the EC algorithm can use the surviving chunks to recreate the missing information.[1]

Compared to 3x replication, 4+2 erasure coding offers:

- 2x usable capacity reduces the number of servers needed by half[2]
- Same level of data protection (failures to tolerate)
- 80% of the read performance

Your data protection configuration can significantly affect cluster performance and capacity. In addition, reducing the number of servers by half cuts capital expenses and power consumption by half as well.

Red Hat Ceph Storage with Micron 7450 NVMe SSDs can build a key storage foundation for cloud-ready object storage solutions in the modern data center.

1. In this document, "performance" means throughput (GB/s), Input Output Operations Per Second (IOPS), response time, or any combination of these.
2. Example: Using 3x replication with 6 storage devices, 2 store data and 4 store replicas. Using EC 4+2, 4 store data and 2 store encoding chunks.

Test configuration

The test was conducted on a RHCS cluster consisting of six data nodes that host Ceph object storage deamons (OSD) and three monitor nodes, as illustrated in Figure 1. Load generation was created using six servers (not shown). For specific server configuration information, please review the "How we tested" section at the end of this brief.

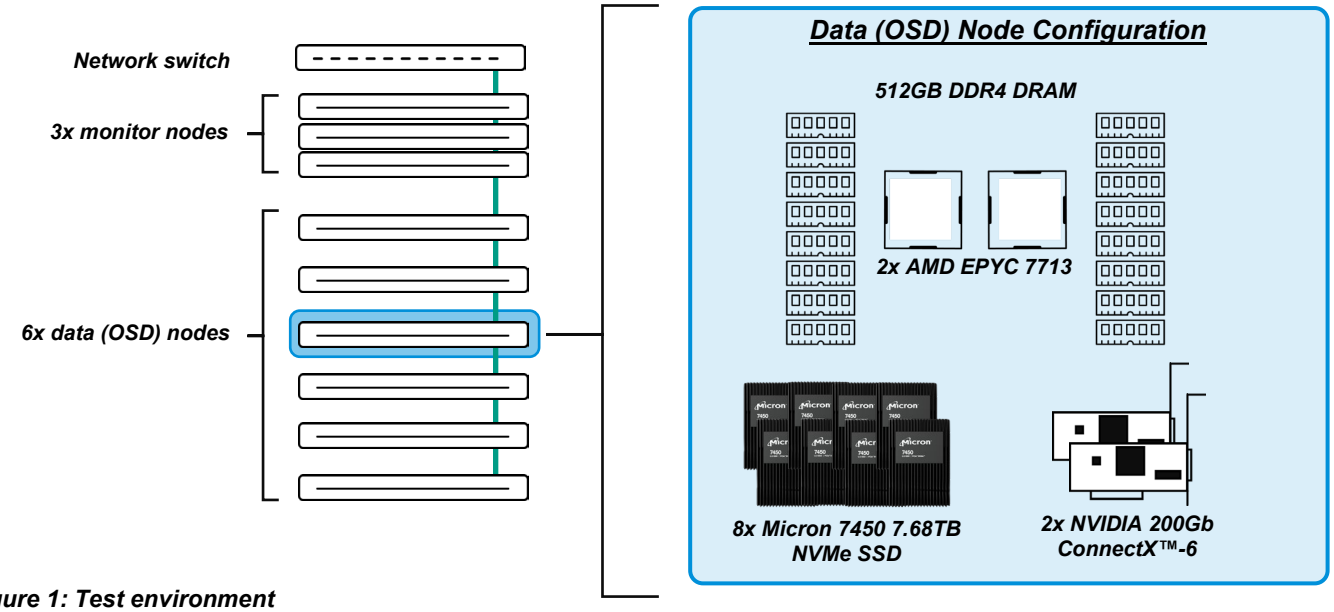


Figure 1: Test environment

Comparing data protection configurations

RHCS data protection is focused on continuous operation after individual data node failure (failures to tolerate, or FTT). RHCS supports two modes of data protection: replication and erasure coding. Replication can be configured to support multiple node failures by configuring the number of data replicas (copies of original data) to store within the RHCS cluster. The default for RHCS is 3x replication. The recommended number of replicas should be between 2 (FTT = 1) and 3 (FTT = 2).

Erasure coding (EC), first introduced in RHCS version 1.2, provides data protection through parity calculation and storage (similar to the process used for RAID 5/6 in disk arrays). The EC configuration is defined as (N+K) where *N* is the number of nodes to store the actual data, and *K* is the number of node failure to be tolerated.[3]

Write operations within RHCS pools always take place on a "primary" OSD for a given client session for each of these data protection mechanisms. Once the data is written to the primary OSD, the configured data protection algorithm is executed, and the data is distributed to the other OSD nodes for that storage pool. RHCS intelligently distributes client connections throughout the OSD nodes within the pool to help ensure that no single OSD node is overloaded.

Finding an optimum data protection method has a direct impact on the success of the project. Each data protection configuration achieves the goal of protecting data but can make dramatic differences in cost and performance as well, as shown in Table 1 below.

3. Additional details on EC are available here: https://access.redhat.com/node/1500653/chapter-31-erasure-code-profiles

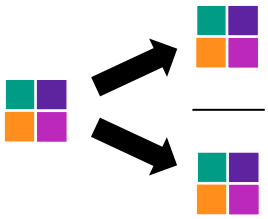
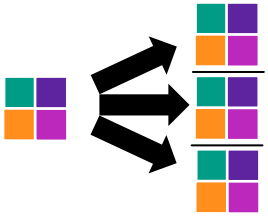
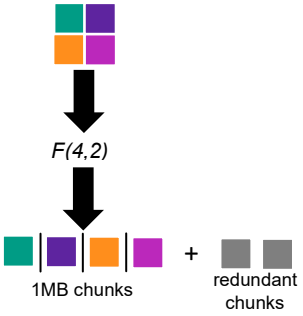
2x Replication		Each 4MB object is stored on two different OSD nodes within the RHCS cluster. 1TB of usable storage capacity requires 2TB of raw capacity.
3x Replication		Each 4MB object is stored on three different OSD nodes within the RHCS cluster. 1TB of usable storage capacity requires 3TB of raw capacity.
4+2 **Erasure Coding**		Each 4MB object is subdivided into four 1MB data chunks and distributed to four different OSD nodes. Two extra 1MB chunks, called redundant chunks, are generated using an erasure coding process and written to two different OSD nodes. 1TB of usable storage capacity requires 1.5TB of raw capacity.

Table 1: RHCS data protection configurations tested

Table 2 shows the characteristics of each type of data protection, including how many faults a storage pool can sustain while keeping data available, and the ratio of the total amount of storage in the configuration divided by the amount of storage available for use. For example, 2x means the total amount of storage in the configuration is twice what is available for use (as seen in 2x replication, which stores two copies of the data).

Failures to Tolerate (FTT)	Data Protection Method	Raw Storage Needed (lower is better)
1 Failure	2x replication	2x
	4+1 erasure coding	1.25x
2 Failures	3x replication	3x
	4+2 erasure coding	1.5x
3 Failures	8+3 erasure coding	1.375x

Table 2: Raw storage requirements for 1PB of "usable storage" using various data protection configurations

Recommendations

Which data protection should be used for your data? We found that 4+2 erasure coding offers the best utilization of storage capacity and the same fault tolerance as a 3x replication configuration (and better fault tolerance than 2x replication). EC 4+2 random read performance was slightly lower than both the 2x and 3x replication configurations. EC 4+2 random write performance falls between the 2x and 3x replication configurations.

Micron used the Ceph RADOS bench[4] benchmarking tool to measure performance of the test cluster and 4MB objects to generate bidirectional data I/O (to and from the RHCS cluster) using six OSD (data) nodes, each hosting eight Micron 7450 PRO 7.68TB[4] SSDs. Complete configuration details are shown in the "How we tested" section at the end of this brief.

Based on the analysis of the test results, several recommendations can be made.

Category	Micron Recommendation
Data Protection	Erasure coding offers the best utilization of storage capacity and the same fault tolerance level as 3x replication.
SSD Selection	The Micron 7450 NVMe SSD offers the throughput and excellent quality of service needed for performance-focused Ceph clusters using erasure coding. This combination yields fast read and write speeds for storage responsiveness, while its wide variety of form factors (U.3, 7mm, and 15mm, as well as E1.S 5.9mm, 15mm, and 25mm) and high capacity (up to 15.36TB) enable OSD node design flexibility, capacity, and growth.
CPU Selection	While the test environment used 2x AMD 7713 64-core CPUs, test results show the optimal OSD node confirmation would use a single-socket system with lower CPU core count. Micron recommends AMD Milan class at 2.0 GHz base/3.6 GHz boost, or faster with: • 16-core CPUs for data replication configurations (AMD 7313P or 7343) • 24-core CPUs for erasure coding configurations (AMD 7443P or 74F3)
Server Selection	One-rack unit servers provide efficient use of data center space, allowing each OSD server to host up to 10 U.2/U.3 SSDs per server.
Network Selection	Red Hat recommends separate client and storage networks for Ceph. The test environment used 200 GbE network links to ensure that network bandwidth was not saturated during testing. For up to 10 SSDs per server (see "Server Selection" above), using a separate 100 GbE link for client and storage networks is optimal.
DRAM Selection	The OSD servers under test had 512GB of DDR4 3200 SDRAM in a one-DIMM per channel configuration, using 32GB DIMMS. While testing, Ceph utilized a maximum of 33% of available DRAM per OSD server. Using a single-socket 1U server and eight 32GB DIMMs at one DIMM per channel (256GB total) should not impact performance.

Table 3: RHCS server configuration recommendations

Since both EC 4+2 and 3x replication provide the same level of data protection — 2 OSD node failures — comparisons between these two configurations will be more useful than comparing to 2x replication (which offers lower FTT). As the data will show, erasure-coded storage pools will outperform 3x replication for write operations, though not read operations.

Throughput and latency analysis

Data protection configuration throughput and latency were evaluated by executing multiple test runs using varying scaling parameters. Object testing utilized the RADOS bench tool to measure object I/O performance (provided as part of the Ceph package). This benchmark reports throughput performance in GB/s and represents the best-case object performance. Object I/O uses a RADOS gateway service operating on each load generation server. (The configuration of RADOS gateway is beyond the scope of this document.)

To measure object read throughput, 60 RADOS bench instances executed 4MB object reads against the storage pool while scaling RADOS bench thread count between 2 threads and 32 threads in base-2 increments. Three test iterations executed for 10 minutes. Before each iteration, the test script cleared all Linux filesystem caches. The results reported are the mathematical average across all test runs.

To measure object write throughput, each test executed RDOS bench[4] with a "threads" value of 16 on a load generation server writing directly to a Ceph storage pool using 4MB objects. The number of RADOS bench instances scaled from 2 to 60. Objects were purged from the pool between each test.

For each data protection configuration, data points marked by the square marker (n) are the recommended scaling maximum based on the RADOS Bench results. Optimal point selections (shown as squares) are based on: 1) maintaining latency below 50ms while 2) ensuring there is not a significant increase in the I/O latency with only marginal throughput improvement. Note that the optimal point may depend on the workload and design imperatives. Plotted lines in Figure 2 are approximations of a best-fit curve and are not representative of actual test results between each tested data point. These lines are provided as visual aids. The 2x and 3x data overlap in the 100% random read chart.

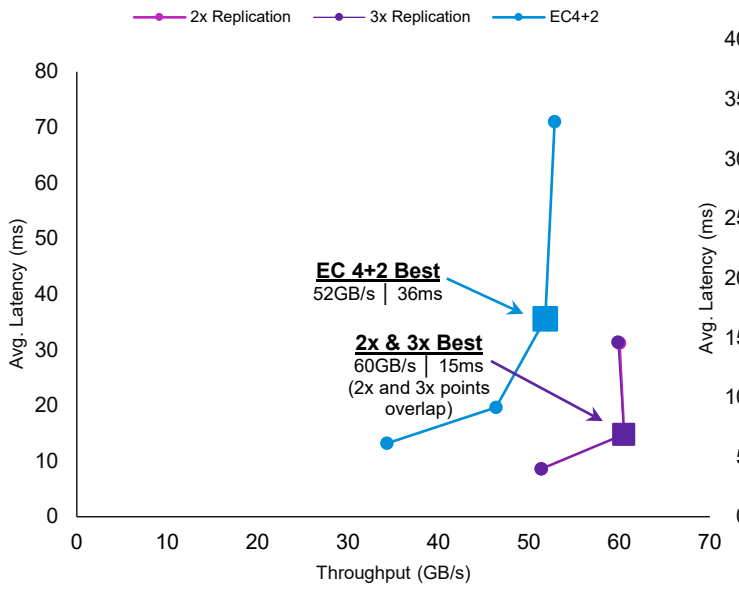


Figure 2a: RHCS performance vs. latency for 4MB object 100% random read

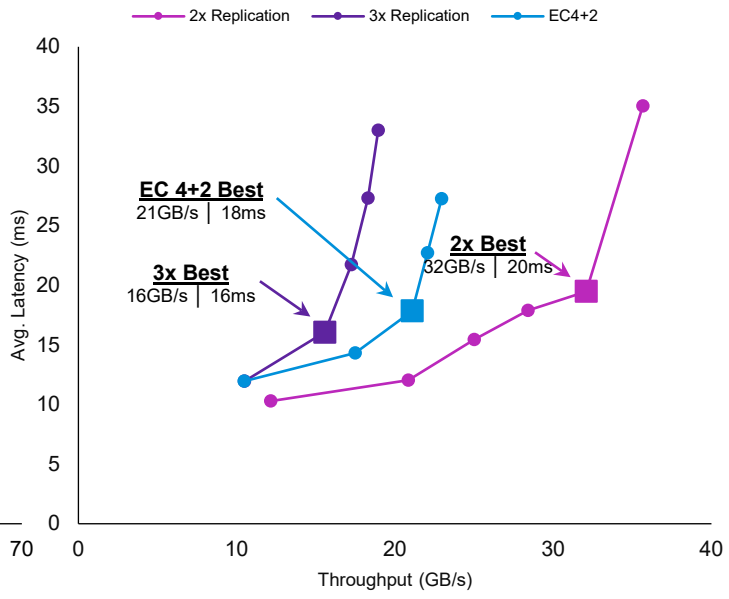


Figure 2b: RHCS performance vs. latency for 4MB object 100% random write

4. Additional details on performance benchmarking is available here: https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/1.3/html/administration_guide/benchmarking_performance

Read Performance: Read performance for replicated data is the same for both 2x and 3x replication (the data points and lines overlap in Figure 2). Beyond eight threads, latency increased significantly, and performance decreased. The EC 4+2 configuration offered a more predictable performance curve, providing approximately 85% of the performance of the replication configurations. At this throughput, EC 4+2 reads had a latency of 36ms.

Write Performance: The 2x replication configuration showed the best object write performance and latency profile, reaching an optimal throughput of 32 GB/s at an average latency of 20ms. The 3x replication had lower object write performance, reaching an optimal throughput of 16 GB/s at 16ms average latency. The EC 4+2 configuration reached 21 GB/s, beating 3x replication by 27% at an 18ms average latency.

Individual application performance can vary based on unique I/O profiles, but this analysis illustrates how important it is to fully understand the application's impact on storage and identify the point where latency becomes too high for smaller increases in data throughput.

SSD performance

Individual SSD performance in the OSD servers when operating at the optimal performance points indicated in Figure 2 was analyzed for each data protection configuration. All I/O operations during the tests were greater than 128KB request size and are considered large-block operations. As shown in Table 4 below, each SSD in the system is responsible for over 1 GB/s of throughput for read operations with sub-millisecond response time. For writes, erasure coding uses a smaller request size, which translates to higher required IOPS than replication configurations.

	Object Reads			Object Writes		
	Disk Throughput	Average Request Size	Average Read Latency	Disk Throughput	Average Request Size	Average Write Latency
3x Replication	1.28 GB/s	752KB	0.87ms	1.27 GB/s	540KB	1.2ms
2x Replication	1.29 GB/s	791KB	0.87ms	1.57 GB/s	496KB	0.83ms
EC 4+2	1.13 GB/s	534KB	0.55ms	816 MB/s	188KB	0.74ms

Table 4: Per SSD comparisons for throughput, latency, and average I/O request size

The Micron 7450 NVMe SSD is an advanced data center SSD. It delivers exceptionally low, consistent latency[5] and extensive deployment options, making it an ideal choice for EC 4+2 data protection in RHCS.

CPU sizing analysis

CPU sizing is expressed by the number of CPU cores per OSD node that were consumed during testing and can be used to help size CPU configurations.

Figure 3 shows that CPU needs for EC 4+2-based data protection are higher than replication configurations. In our tests, replication used four cores for reads and up to 13 cores for writes, while EC 4+2 used 10 cores for reads and up to 18 cores for writes.

The EC 4+2 has higher CPU requirements on the data nodes; since they are dedicated appliances, this is a good trade-off for the performance benefits. Each data node tested had 128 cores available. Based on this data, we recommend using a server with fewer cores than the systems that were evaluated.

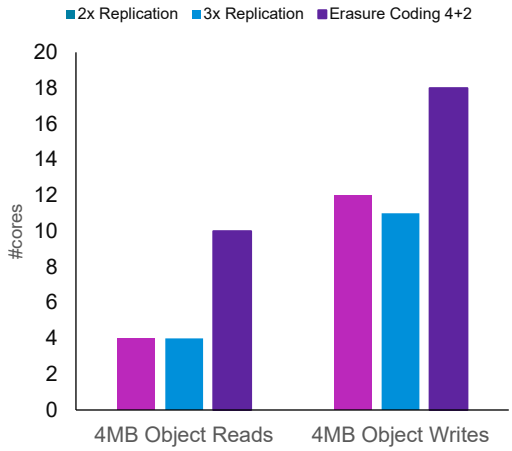


Figure 3: CPU cores consumed for each data protection configuration

5. The Micron 7450 SSD consistently delivers 2ms and lower latency for 99.9999% Quality of Service 1Up to queue depth - 32 for 4KB, 100% random, 90% read workload; up to queue depth = 32 for 4KB, 100% random, 70% read workload

Power efficiency

System power efficiency highlights how much work can be accomplished for the power consumed. For a storage-centric solution such as RHCS, the work accomplished is measured in throughput (GB/s).

Figure 4 shows a comparison of power consumed (in watts) for each unit of throughput performance (GB/s) for the cluster. The lower the value, the more power-efficient the configuration.

This data shows even with erasure coding using more CPU resources, the associated power draw versus the throughput performance may be a positive trade-off depending on design goals and limitations, with EC being close on reads and more efficient on writes.

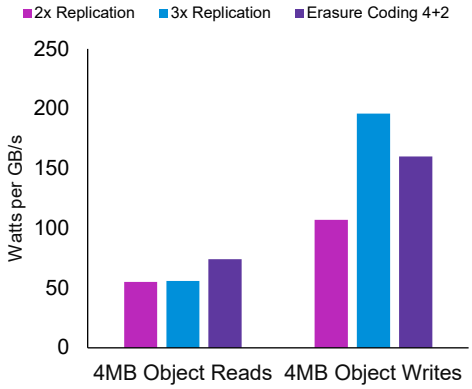


Figure 4: Power consumed for each data protection configuration

Fast networking is required

Network bandwidth analysis is useful when planning what speed of network to use. Each OSD server connects to two separate networks:

Client network: Connects the OSD nodes to the application servers consuming storage on the RHCS cluster.

Storage network (private): Provides communications and data movement among OSD servers and monitor nodes.

During read operations for 2x and 3x replication configurations, performance is 10 GB/s (or 80 Gb/s), which is nearly full bandwidth for a 100 Gb/s link. Data is not moved between OSD servers, so there is no storage network traffic. For erasure coding, data related to the calculated redundancy chunks must be read to support data validation before sending the data to the client application server. Hence, for EC 4+2, this results in a lower throughput to the client when compared to 2x and 3x replication configurations and traffic on the storage network. Tables 5 and 6 show results in both GB/s (to facilitate comparison to other metrics in this document) and typical network speed units of Gb/s.

Read Operations	Client Network	Storage Network	
Data Protection Method	Bandwidth Consumed Out	Bandwidth Consumed In	Bandwidth Consumed Out
3x Replication	10 GB/s (80 Gb/s)	0 GB/s (0 Gb/s)	0 GB/s (0 Gb/s)
2x Replication	10 GB/s (8 0Gb/s)	0 GB/s (0 Gb/s)	0 GB/s (0 Gb/s)
EC 4+2	8.6 GB/s (68.8 Gb/s)	5.4 GB/s (43.2 Gb/s)	5.4 GB/s (43.2 Gb/s)

Table 5: OSD server network analysis for read operations

Write operations require data movement on the storage network and the client network. The 2x and 3x replication configurations require that data written to the primary OSD server be replicated across the storage network. EC 4+2 configurations must first calculate the redundancy chunks and then distribute the data chunks and redundancy chunks to the target OSD servers.

Write Operations	Client Network	Storage Network	
Data Protection Method	Bandwidth Consumed Out	Bandwidth Consumed In	Bandwidth Consumed Out
3x Replication	3.2 GB/s (25.6 Gb/s)	6.4 GB/s (51.2 Gb/s)	6.4 GB/s (51.2 Gb/s)
2x Replication	5.9 GB/s (47.2 Gb/s)	5.9 GB/s (47.2 Gb/s)	6 GB/s (48 Gb/s)
EC 4+2	4 GB/s (32 Gb/s)	4.2 GB/s (33.6 Gb/s)	4.2 GB/s (33.6 Gb/s)

Table 6: OSD server network analysis for write operations

Conclusion

As software-defined storage solutions embrace the unique features and performance of SSDs, all-flash object storage solutions such as RHCS enable high-performance solutions for data analytics, artificial intelligence, and machine learning. With this expansion into more active data read and write applications, it is extremely important that data is protected from loss — loss of storage devices, server nodes, or an entire rack.

This tech brief illustrates how Micron 7450 SSDs enable fault-tolerant RHCS clusters with the same protection level as 3x-replicated clusters (2 FTT), twice the usable capacity, and 80% of read performance. While 2x replication offered the best overall write performance, it required more raw storage capacity and it reduced data protection. The 3x replication offers similar read performance to 2x replication, with better data protection.

The default data protection method in RHCS is 3x replication, but as we have seen — erasure coding offers the best utilization of server capacity and provides better throughput in write-intensive object storage environments than 3x replication configurations.

Actual application performance may vary from the results shown here. For solutions where maximum performance is required and single-failure data protection is enough, 2x replication is recommended. Where data protection is a higher priority, 3x replication provides the best write performance. The 4+2 erasure coding offers a balanced option with good relative performance and more efficient storage utilization than 3x replication, while offering the same double-failure protection.

Learn More

Learn about all of our Ceph reference architectures by visiting our Micron Accelerated Ceph solutions page.

You can also learn more about the Micron 7450 NVMe SSD by visiting the Micron 7450 SSD page.

Learn more about Red Hat Ceph Storage by visiting their website.

How We Tested

Object testing utilizes the RADOS bench benchmarking tool, provided as part of the RHCS package, to measure object I/O performance. This benchmark reports throughput performance in GiB/s (2^{30}). Object I/O uses a RADOS gateway service operating on each load generation server.

To measure object write throughput performance, each test executed RADOS bench with a "threads" value of 16 on a load generation server writing directly to a RHCS storage pool using 4MB objects. The number of RADOS bench instances was scaled from 2 to 60 to determine the maximum throughput value. Objects were purged from the pool between each test.

To measure object read throughput, 60 RADOS bench instances execute 4MB object reads against the storage pool while scaling RADOS bench client thread count between 2 threads and 32 threads in base-2 increments.

In all test cases, three test iterations were executed for 10 minutes each. Before each iteration, all Linux filesystem caches were cleared. The results reported are the mathematical mean across all test runs.

Server configuration

Table 7 describes the hardware and software configuration for each of the server types used in the test configuration. The test environment consisted of six OSD (data) nodes, three monitor nodes, and six load-generation servers.

The six OSD nodes used a single 200 Gb/s port to communicate with the load generation servers and a single 200 Gb/s port to connect to each other and to the RHCS monitor nodes.

	Data (OSD) Nodes	Monitor Nodes	Load Generation Servers
CPU Architecture	AMD EPYC® 7713 (64-cores) Dual Socket NUMA per socket: 4 SMT: enabled IOMMU: enabled	AMD EPYC® 7713 (64-cores) Single Socket	AMD EPYC® 7713 (64-cores) Dual Socket
CPU Cores per Server	128	64	128
Memory	Micron 512GB DDR4 DRAM	Micron 256GB DDR4 DRAM	Micron 512GB DDR4 DRAM
Network	2x NVIDIA® 200Gb ConnectX™-6 (MCX623105AN-VDAT)	1x NVIDIA® 200Gb ConnectX™-6 (MCX623105AN-VDAT)	1x NVIDIA® 200Gb ConnectX™-6 (MCX623105AN-VDAT)
Operating System	Red Hat® Linux® 8.4	Red Hat® Linux® 8.4	Red Hat® Linux® 8.4
Boot Device	Micron data center SATA SSD (240GB)	Micron data center SATA SSD (240GB)	Micron data center SATA SSD (240GB)
Data Storage	8x Micron 7450 SSD (7.68TB)	NA	NA

Table 7: Server configurations

Ceph configuration parameters

Eight OSDs per SSD were configured, totaling 64 OSDs per server and 384 OSDs for the entire storage cluster. Each OSD storage node was configured as a failure domain within the RHCS infrastructure to ensure that data chunks from a protected object were stored on different server nodes.

Raw storage for the RHCS cluster was approximately 365TB. Storage pools were configured for each data protection type, as shown in Table 8.

Pool Data Protection Type	Placement Groups	Usable Capacity
2x Replication	16,384	184TB
3x Replication	16,384	122TB
4+2 Erasure Coding	8,192	244TB

Table 8: Usable capacity for RHCS storage cluster used in testing

Network configuration

A single NVIDIA® SN4700 400 GbE switch was used for test purposes only. It is recommended that at least two switches be used for production environments.

Storage nodes were configured with the following:

- Each NIC was installed in PCIe slots assigned to different CPU sockets.
- IRQ affinity enabled to lock each network interface to the NUMA node assigned during system boot and preventing IRQs being assigned across CPU socket interconnect.
- Transmission queue length was configured to 20,000 and large receive offload (LRO) was enabled by setting the following values in the file: /etc/udev/rules.d/60-mlx-txqueuelen.rules

```
SUBSYSTEM=="net", ACTION=="add", KERNEL=="ens*", ATTR{tx_queue_len}="20000"
SUBSYSTEM=="net", ACTION=="add", KERNEL=="ens3", RUN+="/sbin/ethtool -K ens3 lro on"
SUBSYSTEM=="net", ACTION=="add", KERNEL=="ens6", RUN+="/sbin/ethtool -K ens6 lro on"
```

micron.com