Simple and scalable software-defined storage with Seagate and Red Hat

Highlights:

Modernize your storage as you transform your applications and move to the cloud.

Dramatically simplify software-defined storage deployments with Red Hat® Ceph® Storage 5 and Seagate® Exos® AP 4U100.

Deploy high-performance object, block, and file storage for your most demanding applications.

Start small and scale to store hundreds of petabytes without having to test, qualify, and procure third-party servers for Red Hat Ceph Storage operations.

Table of contents

facebook.com/redhatinc
@RedHat
linkedin.com/company/red-hat

Executive summary

Digital transformations, application modernization, and cloud-native development all compel an ever-escalating need for storage that can perform at scale. New container-based application models come with storage demands that are as diverse as the applications themselves. Modern enterprises are collecting more data than ever before from digital customer interactions, operations, and post-sale support.[2] Much of this data is unstructured, making it ideal for object storage. Analytics, artificial intelligence (AI), machine learning (ML), and complex data pipelines are helping organizations derive more value from their data. While software-defined storage has clear advantages for these and other cloud-based applications, deployment requires expertise that some organizations lack.

Research sponsored by Seagate and conducted by IDC found that between 2020 and 2022, enterprises will see a 42% annual increase in generated data. However, only 32% of the available data will actually be stored and used.[1]

Seagate and Red Hat have worked together to deliver high-performance storage technology that dramatically simplifies deploying software-defined Red Hat® Ceph® Storage clusters based on the Seagate Exos AP 4U100 integrated storage system. Red Hat Ceph Storage 5 includes management support for the entire Ceph cluster life cycle, starting with the bootstrapping process. The Seagate Exos AP 4U100 provides extremely high-density compute and storage resources in a single system. A high-performance Red Hat Ceph Storage cluster can be deployed using only Seagate Exos AP 4U100 enclosures—eliminating the need to qualify third-party servers and further simplifying deployment and system configuration tasks.

Testing conducted by Seagate has shown that the Seagate Exos AP 4U100 performs well as a Red Hat Ceph Storage cluster building block when equipped with Seagate Exos X16 Enterprise nearline SAS hard disk drives (HDDs) and Seagate Nytro Enterprise solid state drives (SSDs).[3] Organizations can build a high-performance starter cluster with as few as three half-populated Seagate Exos AP 4U100, scaling capacity with additional drives and enclosures as their needs escalate. Red Hat Ceph Storage and Seagate Exos systems have demonstrated scalability to over 10 billion objects.

Combining Red Hat and Seagate technology for simplicity and scale

Red Hat Data Services provides multiple ways to deploy software-defined storage technology with the Seagate Exos AP 4U100 storage system.

Red Hat Ceph Storage 5

Red Hat Ceph Storage provides an open and robust data storage solution for on-premise or cloud deployments. As a self-healing and self-managing platform with no single point of failure, Red Hat Ceph Storage significantly lowers the cost of storing enterprise data and helps organizations automate management for exponential data growth. Red Hat Ceph Storage is optimized for large installations—efficiently scaling to support hundreds of petabytes of data. Powered by industry-standard x86 servers, the platform delivers solid reliability and data durability with either standard three-way (3x) replication or erasure coding. Red Hat Ceph Storage is also multisite aware and enabled for disaster recovery.

[1] *"Rethink Data: Put More of Your Business Data to Work—From Edge to Cloud." Seagate, July 2020.*

[2] *"Rethink Data: Put More of Your Business Data to Work—From Edge to Cloud." Seagate, July 2020.*

[3] *Evaluator Group, "The 10 Billion Object Challenge with Red Hat Ceph 4.1," Nov. 2020.*

For more information on deploying container storage on external mode, see the Red Hat OpenShift Data Foundation documentation.

A single Red Hat Ceph Storage cluster can support object, block, and file access methods. The cluster's scale-out capabilities can be configured for capacity or input/output (I/O) performance as needed to match intended workloads. Clusters can be expanded or shrunk on demand. Updates can be applied without interrupting vital data services, with hardware added or removed while the system is online and under load.

Red Hat Ceph Storage 5 adds new features that let organizations maximize the value of their software-defined storage installations, including:

- **Functionality.** A new integrated control plane includes `cephadm` (a tool for deploying and managing Ceph clusters), a stable management API, new Ceph filesystem capabilities, and new Reliable Autonomic Distributed Object Store (RADOS) block device (RBD) capabilities such as RBD snapshot-based data migration across clusters.
- **Performance.** Red Hat Ceph Storage 5 includes dramatic performance improvements, including a new cache architecture that offers 80% improved block performance for virtual machines and an object aggregate throughput of greater than 80 GB/s. Continued object-store scalability improvements have demonstrated over 10 billion objects through the RADOS object gateway (RGW).
- **Security.** Improved security features include an object lock that supports "write once, read many' (WORM) governance, FIPS 140-2 cryptographic libraries, enhanced access control, external key manager integration, and granular object encryption.
- **Efficiency.** The release includes improved space utilization for small objects with a 4K allocation size replacing the previous Bluestore 64K allocation size for HDDs, significantly reducing the overhead for storing small objects. Red Hat Ceph Storage 5 also provides significantly faster recovery for erasure coded volumes. The RADOS object gateway can operate across sites, providing massive scalability.

Red Hat OpenShift Data Foundation

Red Hat OpenShift® Data Foundation—previously Red Hat OpenShift Container Storage—is software-defined storage for containers. Based on Ceph, Rook, and Noobaa technology, OpenShift Data Foundation is the data and storage services platform for Red Hat OpenShift. OpenShift Data Foundation helps teams develop and deploy applications quickly across public, private, and hybrid cloud environments.

OpenShift Data Foundation gives solution architects multiple deployment options to support their diverse workloads. The platform offers converged and disaggregated internal modes that deploy application and storage pods within the Red Hat OpenShift cluster. OpenShift Data Foundation also supports an external mode that decouples storage from Red Hat OpenShift clusters entirely.

With OpenShift Data Foundation external mode, multiple Red Hat OpenShift clusters can consume storage from an external Red Hat Ceph Storage cluster. Decoupling storage in this manner provides maximum flexibility, allowing compute and storage resources to be scaled independently. A Red Hat Ceph Storage cluster built on Seagate Exos AP 4U100 storage systems could be shared by multiple Red Hat OpenShift clusters using OpenShift Data Foundation external mode.

Seagate Exos AP 4U100

Pictured in Figure 1, the Exos AP 4U100 combines 96 3.5-inch drive slots and four 2.5-inch drive slots in a four rack-unit (4U) enclosure. Configured in a split-chassis, shared nothing configuration, each Exos AP 4U100 provides two entirely separate dual-socket compute nodes (also referred to as application controllers), with each accessing 50 drives within the chassis. Each application controller supports two Intel Xeon Scalable processors, up to 256GB of RAM, and optional 100 Gigabyte Ethernet (GbE) network links.

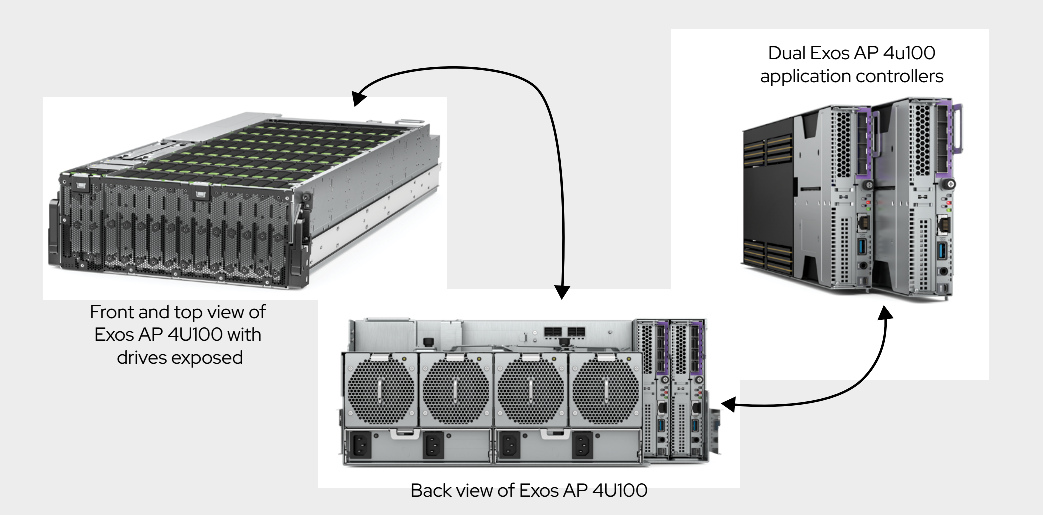

Figure 1. The Seagate Exos AP 4U100 provides extensive storage and two integrated application processors.

The Exos AP 4U100 simplifies the deployment of clustered storage by bringing storage and compute resources together into a single, solution-certified building block. With as little as three enclosures, the six application controllers can host all the containerized services of a Red Hat Ceph Storage cluster. No external third-party servers are required to realize a high-performance, ultra-dense, and cost-effective Red Hat Ceph Storage cluster. For system administrators, this simplification eliminates the need to source, configure, maintain, and rack separate servers to host Ceph services, reducing operational cost and complexity.

Lab configuration overview

To evaluate the performance of Red Hat Ceph Storage 5, Seagate engineers built a cluster using three Exos AP 4U100 enclosures, with each hosting two application controllers (compute nodes) to provide six Red Hat Ceph Storage nodes. As shown in Figure 2, each two-socket application controller operates as a separate server. Storage on each Exos AP 4U100 is configured with split zoning so that nothing is shared between the application controllers.

Each application controller has 12 Gb/s SAS links to each of the 50 drive slots. As configured, the drive slots for each application controller were populated with 44 Seagate Exos X16 16TB[4] nearline SAS HDDs and six Seagate Nytro 3.84TB[4] SSDs. This storage configuration allowed each application controller to control 42 HDDs (as well as two hot spares). The six SSDs served as metadata caching for each of the six application processor nodes.

Each Exos AP 4U100 application controller and the four test clients featured an NVIDIA ConnectX 100GbE network interface controller (NIC), with all NICs routed via QSFP28 links to an NVIDIA Spectrum SN2100 100GbE switch. No separate public or private cluster networks were created.

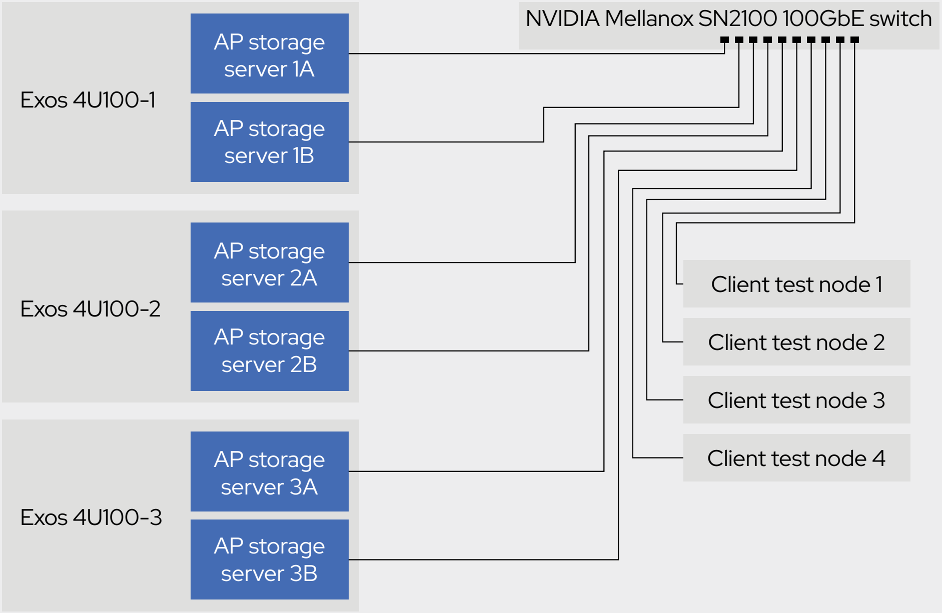

Figure 2. Lab configuration deployment architecture

High-density storage array with two integrated servers

Each application controller in the Seagate Exos AP 4U100 was configured as a storage server with two Intel Xeon Silver 4110 processors, 256GB of RAM, and an NVIDIA Mellanox ConnectX 5 100GbE (ENx 5) network interface controller (NIC). Table 1 lists details for both the storage servers and the client test nodes.

[4] *When referring to drive capacity, one gigabyte, or GB, equals one billion bytes and one terabyte, or TB, equals one trillion bytes. Your computer's operating system may use a different standard of measurement and report a lower capacity. In addition, some of the listed capacity is used for formatting and other functions, and thus will not be available for data storage. Seagate reserves the right to change, without notice, product offerings or specifications.*

Table 1. Storage server configuration details

Hardware Element	Description
Enclosures	Three Seagate Exos AP 4U110 enclosures, each with two dual-socket application controllers, 96x 3.5-inch drive slots, and 4x 2.5-inch drive slots per enclosure
Application controllers (storage servers)	Processors: Two Intel Xeon Silver 4110 @2.1GHz processors per application controller with eight cores and 16 threads per socket
	RAM: 256GB per application controller
SAS HDD drives	88 Seagate Exos X16 16TB[4] ST16000NM002G (44 per application controller), 264 total HDDs used
SAS SDD drives	12 Seagate Nytro XS3840SE10103 3.84TB[4] (six per application controller), 36 total SSDs used
Network	Two NVIDIA Mellanox CX516-A ConnectX-5 Dual-100GbE NICs per enclosure (one per application controller)
	NVIDIA Spectrum SN2100 16-Port 100GbE Data Switch

Table 2 lists the software elements used on storage servers in Seagate testing.

Table 2. Software configuration details

Software Element	Description
Host operating system	Red Hat Enterprise Linux® release 8.4 (Ootpa)
Containers	Docker, deployed and managed by podma
Ceph	Red Hat Ceph Storage 5.0 Beta (Pacific Stable 16.2.0-4.el8cp)
Data protection	3x replication and 4+2 erasure coding
Deployment method	`cephadm`
Ceph object storage daemon (OSD) count	264
Ceph monitor (MON) count	Five of six nodes
Ceph manager (MGR) count	Five of six nodes
RADOS gateways	12 (two per application controller)

Software Element	Description
HDDs	BlueStore with all permutations of four disk schedulers
SSDs	Seven partitions of 512GB, with each serving seven OSDs

Client test nodes

For performance testing, the Red Hat Ceph Storage cluster was addressed as an Amazon Web Services (AWS) Simple Storage Service (S3) target for performance testing. Four client test nodes exercised the storage cluster using the Minio Warp S3 benchmarking tool (version 0.4.3 - 76801ae). Table 3 lists the configuration of the individual client test nodes.

Table 3. Client test node configuration

Hardware Element	Description
Server	Intel Server System R1208WFTYS (1U)
Processor	Intel Xeon E5-2640
RAM	128GB DDR3 D3-68SA104SV-13
Network	Dual-port 100GbE NVIDIA Mellanox MCX516A-CCAT ConnectX-5 EN NICs

Provisioning the cluster

The Seagate engineers wanted to evaluate the ease of installing a Ceph cluster as a focus for testing. The team used iPXE network boot to perform bare-metal provisioning of Red Hat Enterprise Linux 8.4. (The kickstart script that clears all cluster storage disks and automatically selects the correct boot device in the Exos AP 4U100 can be found in Appendix B: Scripts.) The engineers were able to bare-metal provision any arbitrary number of Exos AP 4U100 application controllers with Red Hat Enterprise Linux in less than 10 minutes.

Red Hat Ceph Storage deployment used the powerful **`cephadm bootstrap`** utility. When fed an appropriate YAML configuration file, **`cephadm`** can deploy a fully containerized storage cluster. (Full details on the Red Hat Ceph Storage deployment method are provided in Appendix A: Cluster provisioning.)

Testing the cluster

Engineers configured the four test clients to run Warp in client mode while a fifth node ran Warp in test orchestrator mode. The bandwidth measurement tool iPerf3 was used to verify overall network operation and performance between the application controllers and the S3 test client nodes. The basic test methodology focused on either “GET” (read) or “PUT” (write) operations while iterating across object sizes of 128KB, 512KB, 1MB, 4MB, 16MB, and 64MB. Thread concurrency was incremented across each of the four test clients at eight (32 total), 16 (64 total), 64 (256 total), 96 (384 total), and 128 (256 total) threads. (The script used for testing is provided in Appendix B.)

Each test was granular in that it only tested one type and size of object transaction at a time. The team got stable and repeatable test results by using the same test process and method for every test and quickly reprovisioning the cluster.

Performance results

The same performance tests were conducted across a 3x replicated cluster and 4+2 erasure-coded cluster. Testing also evaluated the importance of using SSDs for Ceph metadata caching.

Performance for three-way (3x) replication

Initial testing configured the Red Hat Ceph Storage cluster for 3x replication, implying that the cluster stored three full data replicas for data protection and resilience.

GET/read performance

Figure 3 shows the result of running all Red Hat Ceph Storage services exclusively on the Exos AP 4U100 application controllers. For GET operations, the peak throughput was 12.7GB/s for objects larger than 512KB, indicating that the 100GbE network connection was fully utilized.

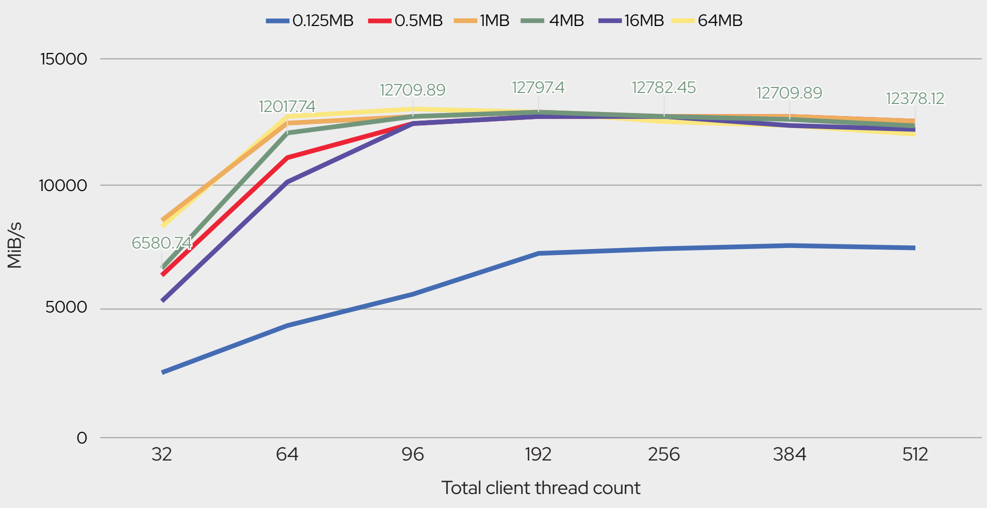

Figure 3. GET/read performance with 3x replication for various object sizes (MQ-deadline disk scheduler, 4MB results annotated)

PUT/write performance

Figure 4 shows PUT (write) performance under 3x replication. The S3 PUT performance peaked at 2.3GB/s throughput for 16MB objects. Analysis of the results indicated that the use of larger capacity or additional SSDs would improve write performance.

MQ-deadline is the default disk scheduler, but engineers have observed some gains with the other options in the past.

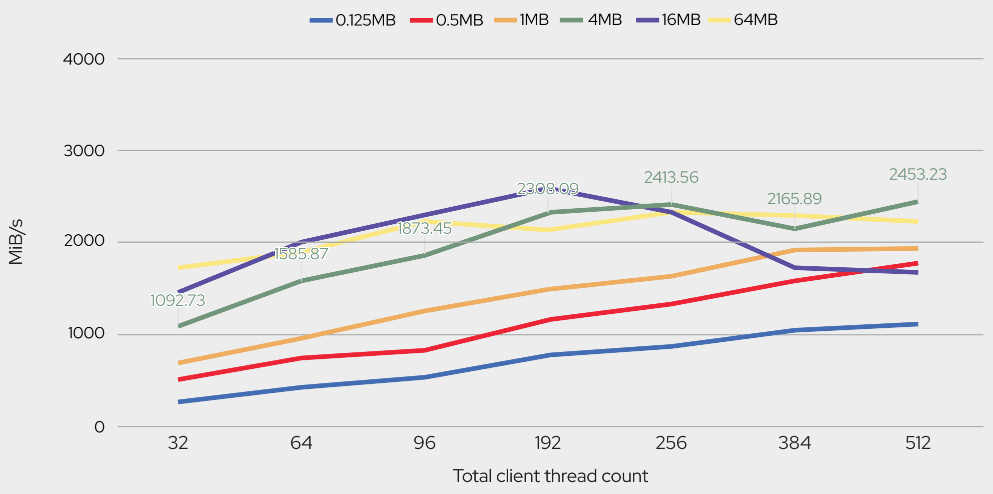

Figure 4. PUT/write performance with 3x replication for various object sizes (MQ-deadline disk scheduler, 4MB results annotated)

The effect of HDD disk schedulers

Red Hat Enterprise Linux 8 offers the choice of four disk schedulers: bfq, kyber, MQ-deadline, and "none". As a part of performance testing, Seagate engineers wanted to understand the effect of HDD disk schedulers on both read and write performance. In a pure HDD Ceph solution, engineers in the Seagate Reference Architecture lab had previously observed performance variations as a function of the chosen disk scheduler.

However, in the current testing, engineers found that the different HDD disk schedulers had a minimal overall effect (Figure 5 and Figure 6). The metadata services that high-performance Seagate Nytro 12Gb/s SAS SSDs provided likely negated any beneficial effects on performance from using different disk schedulers.

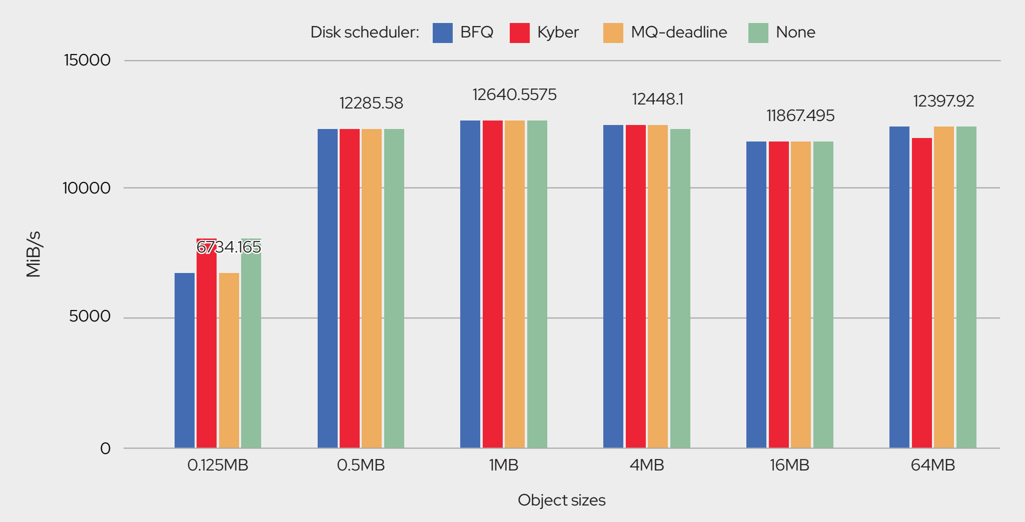

Figure 5. Disk scheduler performance for GETs/reads across multiple object sizes (MQ-deadline results annotated)

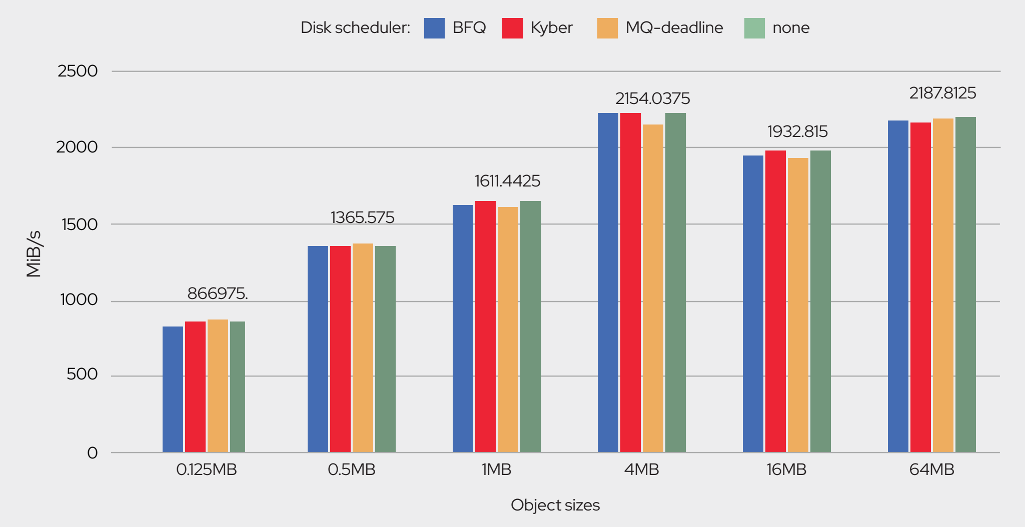

Figure 6. Disk scheduler performance for PUTs/writes across multiple object sizes (MQ-deadline results annotated)

Performance for 4+2 erasure coding

Three-way replication (or triple redundancy) provides reliable data protection but yields only a third (33%) of raw cluster capacity as usable storage. In contrast, 4+2 erasure coding increases the usable storage capacity to 66% while allowing for any two devices in a six-drive placement group to fail without data loss. Erasure-coded deployments use two parity blocks for every four data chunks, resulting in 4+2 erasure coding.

GET/read performance

As with 3x replication, with 4+2 erasure coding the peak GET throughput was 10.6GB/s for objects larger than 512KB in size, which indicated nearly full utilization of the 100GbE network. For small 128KB objects, GET throughput peaked at around 4GB/s.

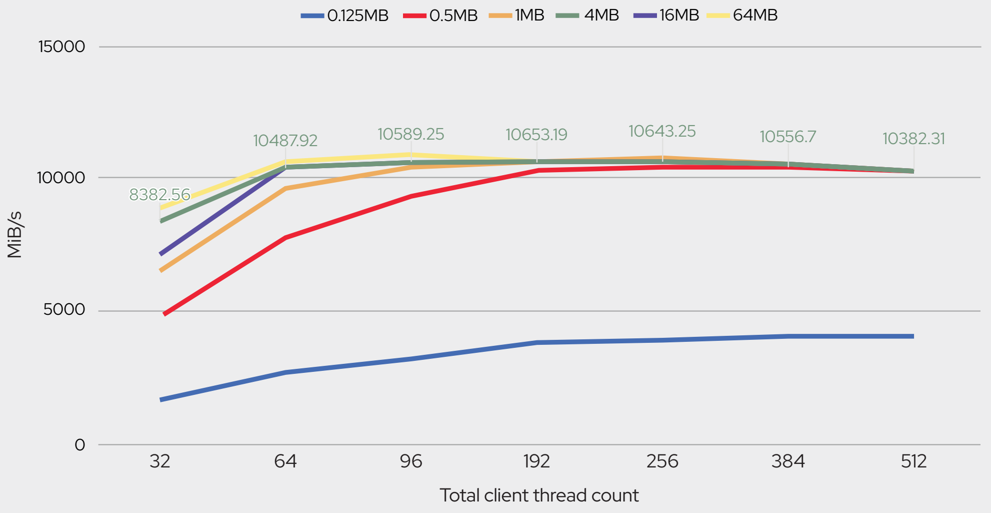

Figure 7. GET performance with 4+2 erasure coding, MQ-Deadline disk scheduling, and 128KB stripe unit (4MB results highlighted)

PUT/write performance

Using 4+2 erasure coding for data protection, the maximum PUT throughput of 3.4GB/s was greater than the 3x replication maximum PUT through of 2.3GB/s for 16MB objects, PUT performance was limited to the number of I/O operations per second (IOPS) available through the six SSDs connected to each storage node (Figure 8). As configured, each SSD hosted seven 512GB partitions, each used for metadata placement for a 16TB HDD Ceph OSD. With 42 HDDs per application controller, 42 SSD partitions in total were required (six SSDs with seven partitions each).

Though not tested, our cluster instrumentation indicated that better PUT performance could potentially be realized by using 7.36TB[4] Seagate Nytro SSDs and increasing the SSD count per application controller from six to eight. This reconfiguration would have the effect of reducing the number of OSDs being serviced by a single SSD from seven to five, while more than doubling the size of the partitions to host the Ceph metadata.

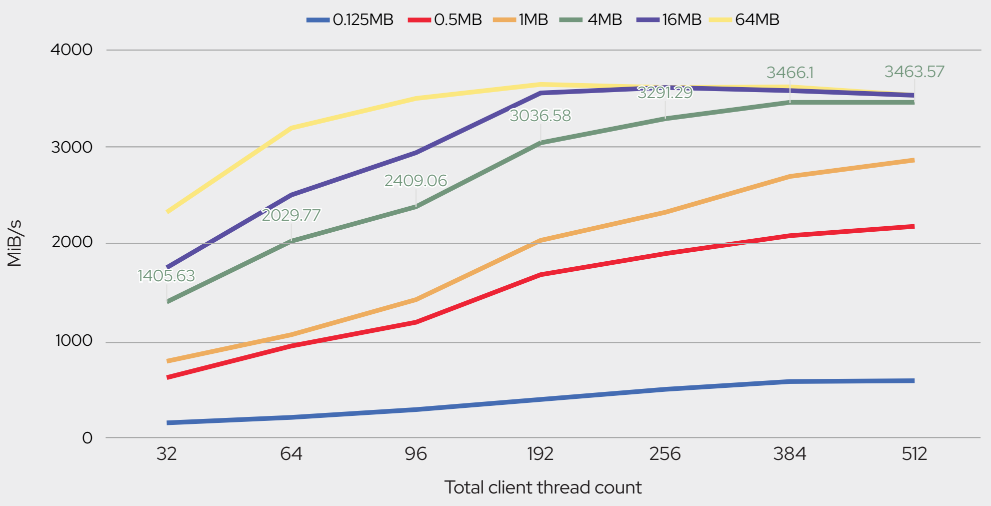

Figure 8. PUT performance with 4+2 erasure coding, MQ-Deadline disk scheduling, and 128KB stripe unit (4MB results highlighted)

The effect of placement group stripe size

With erasure coding, data and parity are "striped" across a set of OSDs. What the optimal stripe width size should be for a given workload is a question that naturally arises. To answer that question, the team tested four different configurations of stripe unit widths (128KB, 256KB, 512KB, and 1MB) for both GET and PUT performance for object sizes of 128KB, 512KB, 1MB, 4MB, 16MB, and 64MB.

Testing showed that the stripe unit width had a surprisingly small overall effect on performance other than storage efficiency. This finding is partly because of the gain provided by the twelve RADOS gateways (two containerized instances per application controller), each acting as a distinct S3 endpoint.

Figure 9 shows GET tests for stripe unit byte sizes of 0.125MB, .5MB, 1MB, 4MB, 16MB, and 64MB. For 128KB objects, engineers observed a relatively constant transfer rate of about 3.5MB/s. For objects of all other sizes, transfer rates were consistently around 10.4 GB/s.

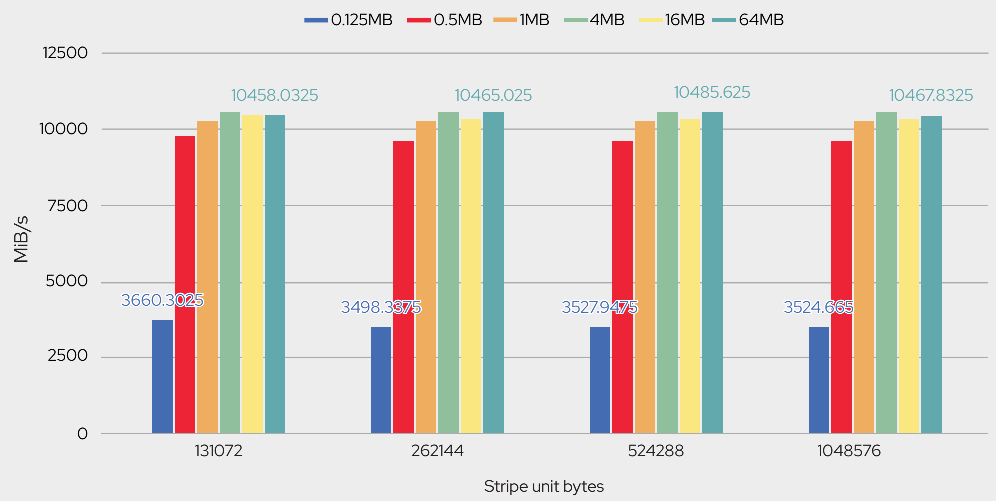

Figure 9. Stripe unit bytes compared with. GET transfer rates

Figure 10 shows stripe unit bytes compared to PUT transfer rates. In this case, 128KB objects have a reasonably consistent transfer rate of about 470MB/s. On the other hand, 64MB objects had transfer rates that were consistently around 3.4GB/s, irrespective of the placement group stripe width.

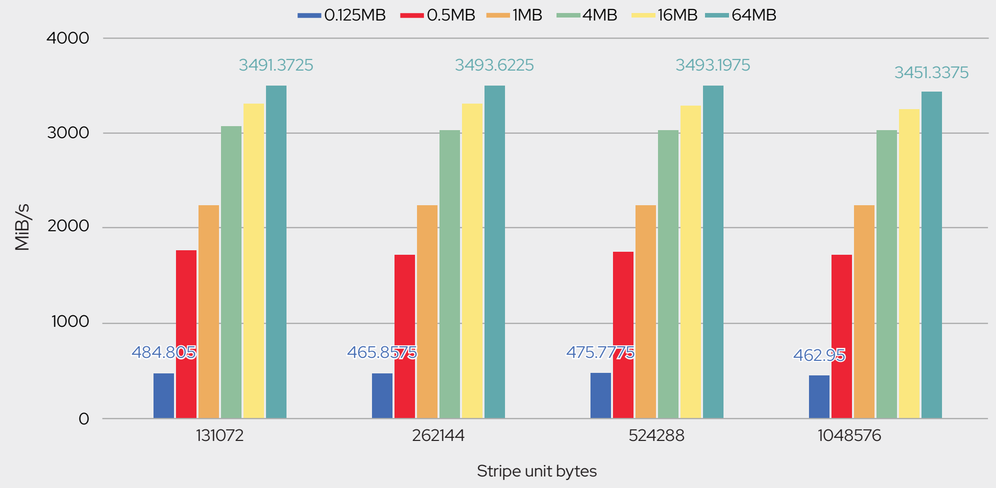

Figure 10. Stripe unit bytes compared with PUT transfer rates

Performance effects of SSDs for metadata caching

Because enterprise-class 12GB SAS SSDs can significantly impact the total cost to run a Red Hat Ceph Storage cluster, engineers wanted to evaluate performance with and without SSDs for metadata caching.

GET/read performance without SSDs

Figure 11 shows GET/read performance without the benefit of any SSDs configured to offload the Ceph OSD metadata. Figure 12 shows performance with the standard six SSDs and 42 HDDs configured. The relatively small performance difference indicates that workloads centered on GET/read operations don't require SSDs to accelerate OSD metadata.

Seagate Exos X16 16TB Enterprise HDDs effectively reorder reads to serialize them as much as possible. This factor, combined with an abundance of RADOS gateways via a containerized Red Hat Ceph Storage deployment, demonstrates that SSDs have little effect on the network performance ceiling of about 11GB/s for GET operations.

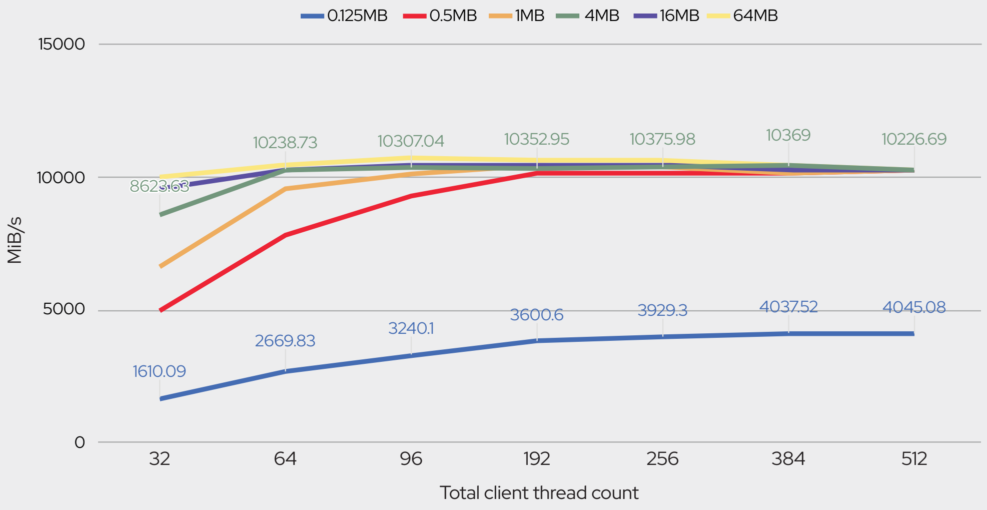

Figure 11. GET/read operations with 4+2 erasure coding and no SSDs configured for Ceph metadata

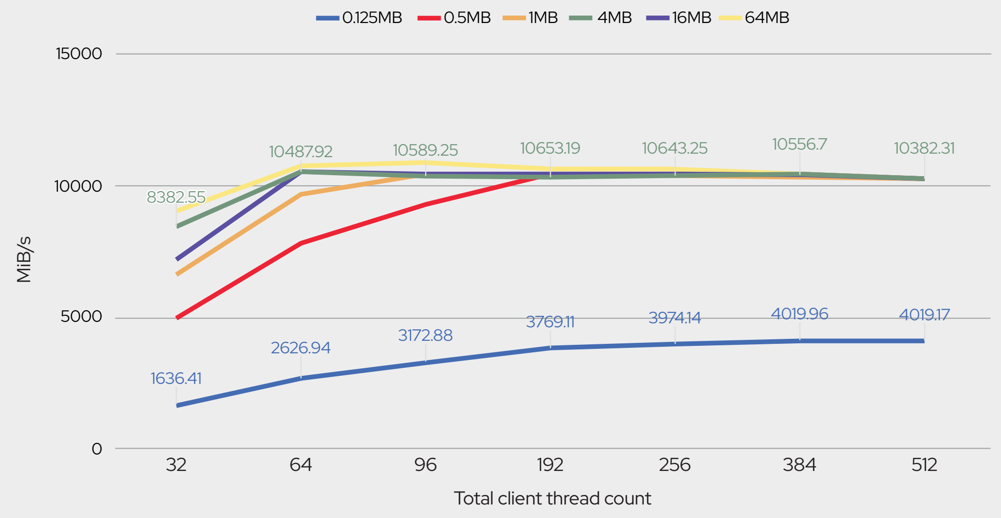

Figure 12. GET/read operations with 4+2 erasure coding and six SSDs configured for Ceph metadata

PUT/write performance without SSDs for metadata caching

Figure 13 shows PUT/write operations without SSDs for Ceph metadata, while Figure 14 shows the standard configuration of six SSDs and 42 HDDs. Unlike read operations, write operations demonstrate that the positive performance effect of SSDs for metadata caching is far more profound. The performance trendline for 4MB objects is nearly double for the cluster with the SSD cache enabled. The effect on 128KB object writes is even more compelling.

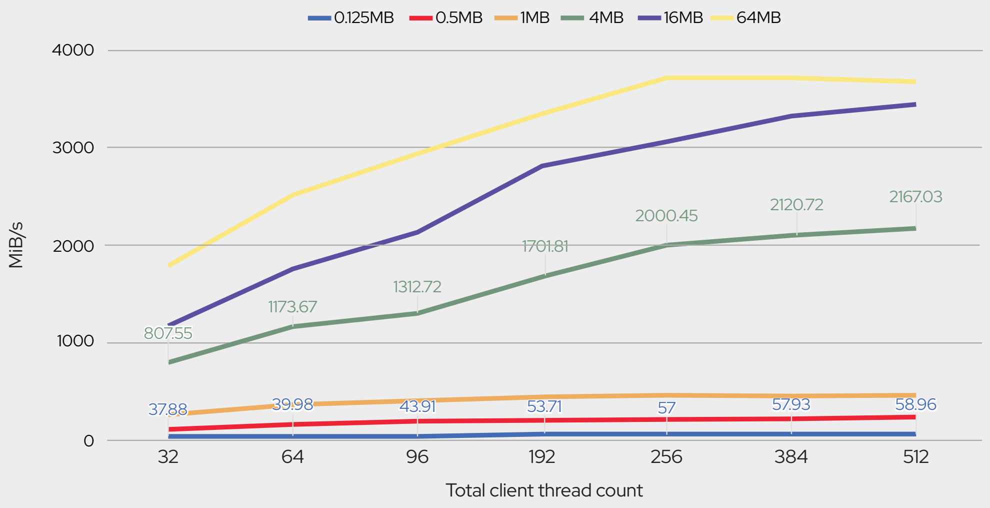

Figure 13. PUT/write operations with 4+2 erasure coding and no SSDs for Ceph metadata

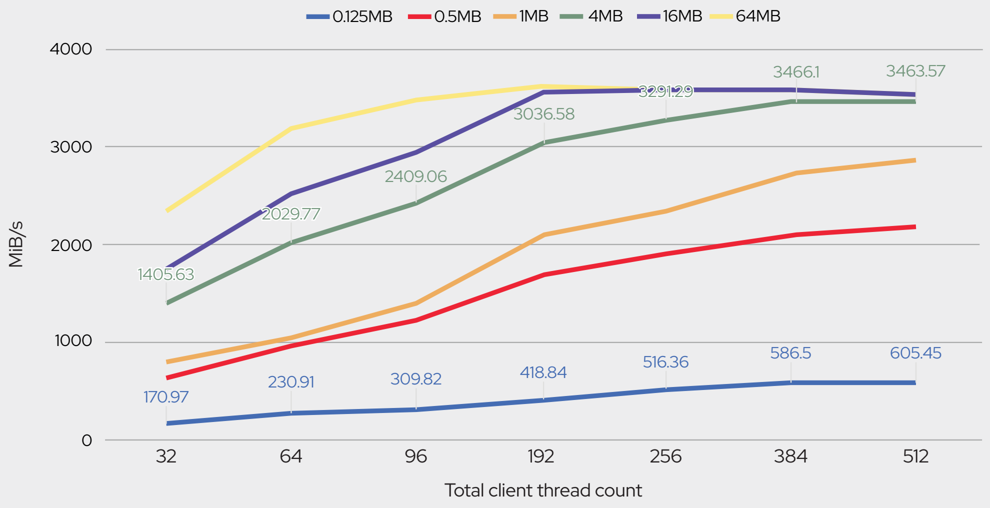

Figure 14. PUT/write operations with 4+2 erasure coding and six SSDs for Ceph metadata

To better illustrate the performance delta for PUT/Write operations, Figure 14 shows the comparative performance improvement from adding just six 3.84TB Seagate Nytro SSDs for every 42 HDDs in the Ceph cluster. Even for 4MB objects, engineers saw a performance improvement averaging over 70%. For 128K objects, they saw a 685% performance improvement with just 192 S3 client threads.

Conveniently, engineers were able to use the `cephadm bootstrap` command to rapidly reconfigure the existing cluster to use only 24 HDDs. See the Appendix for full details.

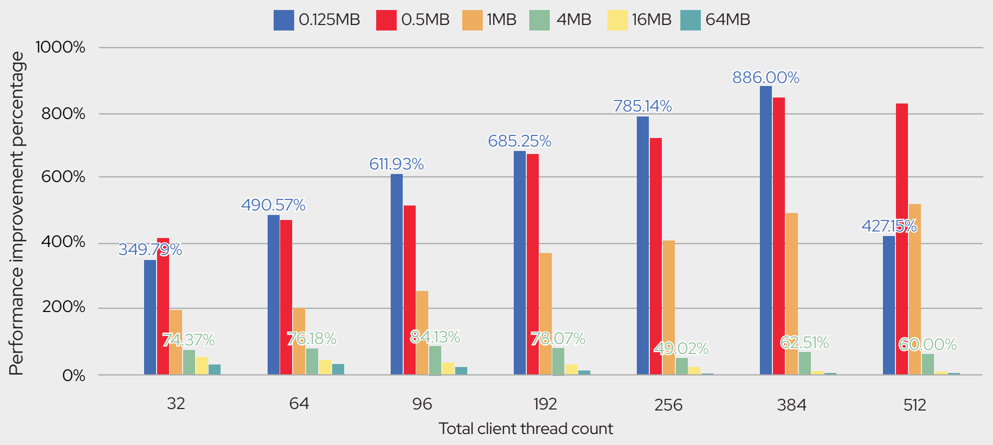

Figure 15. Percentage performance improvement for PUT/write operations with six SSDs for Ceph metadata caching

Performance with a half-populated enclosure

Red Hat Ceph Storage provides extensive storage scalability, allowing organizations to start small and scale to hundreds of petabytes. As a part of their testing, Seagate engineers wanted to evaluate the performance of a smaller starter cluster. Using **`cephadm`**, they could rapidly configure a Red Hat Ceph Storage cluster with only 24 HDDs per application controller (See Appendix B for YAML script details).

The minimum viable erasure-coded Red Hat Ceph Storage cluster still requires three Seagate EXOS AP 4U100 enclosures (or a total of six application controllers). The team wanted to understand what kind of performance to expect from half-populated enclosures. This configuration approach lets organizations benefit from the simplicity and performance of Red Hat Ceph Storage 5 on the Exos AP 4U100 while deferring full capacity expansion.

GET/read performance with only 24 HDDs

Figure 16 shows GET/read performance with only 24 HDDs configured per application controller within the Red Hat Ceph Storage cluster. As before, six SSDs were configured per application controller for Ceph metadata caching. Once again, the Seagate Exos 16TB SAS nearline HDDs saturated network read performance—this time with just 24 drives. The ability of the Exos X16 16TB drive to transform pseudo-random read requests into semi-sequential data bursts contributes directly to this performance.

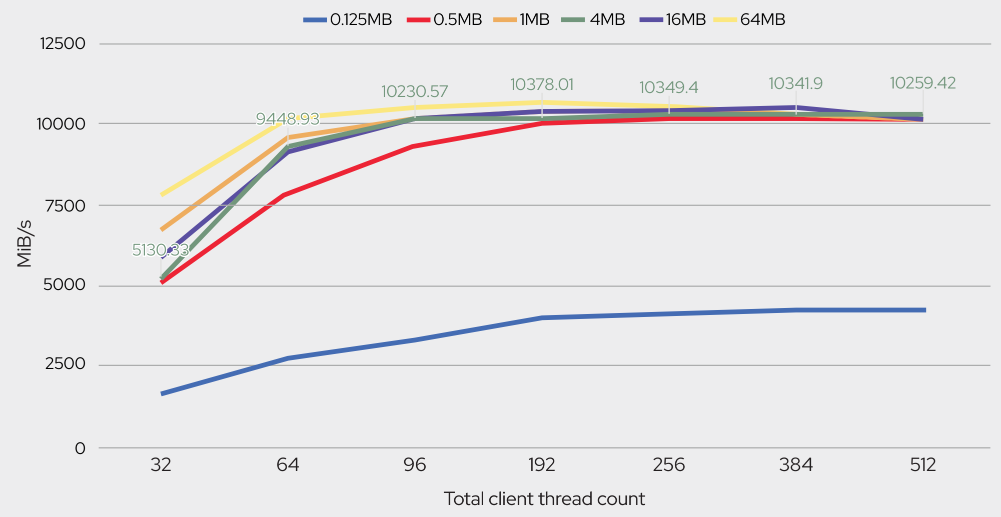

Figure 16. GET/read performance with only 24 HDDs configured per application controller

PUT/write performance with only 24 HDDs

Figure 17 and Figure 18 compare the performance of a cluster with 24 and 42 HDDs per application controller, respectively. Remarkably, a 24-spindle solution can perform within a few percentage points of a 42 HDD cluster, even though it has fewer HDD spindles. That small performance differential is a testament to the performance of the Exos X16 Enterprise HDDs, and the Seagate Nytro SSDs used to augment them as metadata caches on S3 writes to the Red Hat Ceph Storage cluster.

This comparison also demonstrates that write speed was limited by the size and number of Nytro SSDs integrated into the cluster deployment recipes. These results show that systems administrators can start with smaller storage configurations and expand both horizontally and internally with no performance degradation.

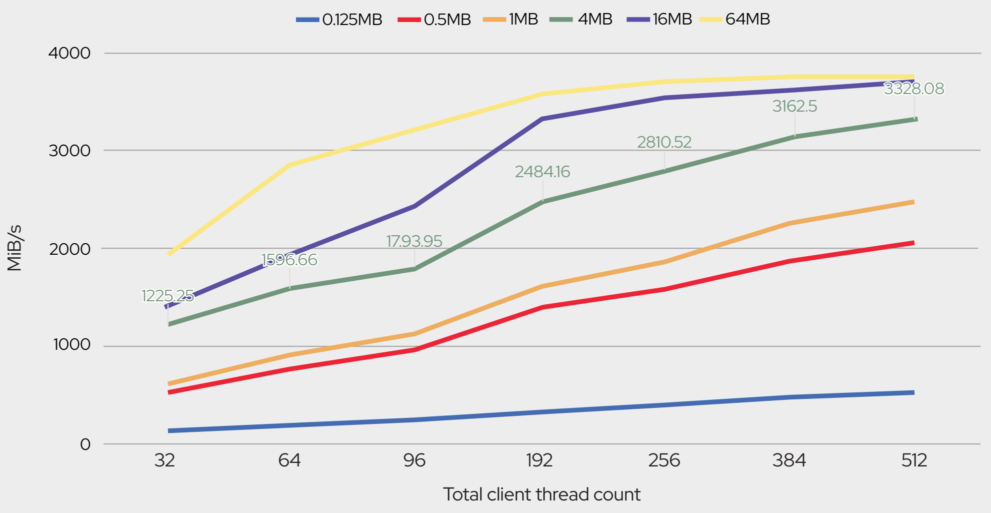

Figure 17. PUT/write performance with only 24 HDDs configured per application controller

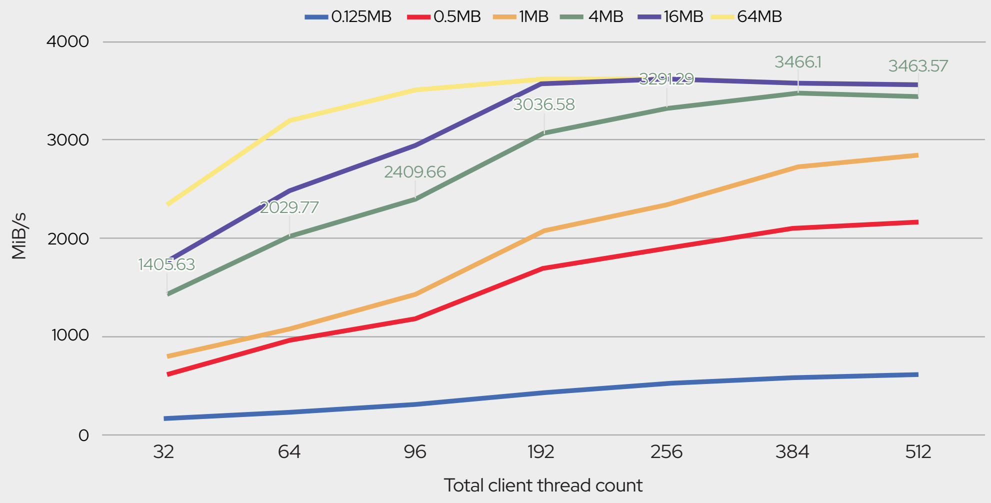

Figure 18. PUT/write performance with a full complement of 42 HDDs configured per application controller

Conclusion

Combined with a containerized Red Hat Ceph Storage 5 deployment, the Seagate Exos AP 4U100 platform offers unprecedented simplicity, state-of-the-art performance, and scalability. Ease of use mechanisms in Red Hat Ceph Storage 5 such as **`cephadm`** allow the rapid deployment or reconfiguration of software-defined storage clusters. Red Hat Ceph Storage clusters can scale independently from computing resources, be optimized for specific workloads, and be used directly by Red Hat OpenShift Data Foundation external mode.

The Seagate Exos AP 4U100 platform further simplifies the deployment of Red Hat Ceph Storage clusters. The high-density platform reduces supply chain decisions, eliminating the need for testing and qualifying third-party servers for Red Hat Ceph Storage operations. Organizations can deploy scalable, high-performance Red Hat Ceph Storage clusters while saving money and rack space.

Learn more about how your organization can build scalable and efficient object storage with Seagate Exos hardware and Red Hat Ceph Storage.

Appendix A: Cluster provisioning

The sections that follow describe the simplified cluster provisioning procedures afforded by Red Hat Ceph Storage 5.

Bare metal OS provisioning

iPXE network boot provided bare-metal provisioning of Red Hat Enterprise Linux 8.4. The kickstart script (Appendix B) clears all cluster storage disks and automatically selects the correct boot device in the Exos AP 4u100. Via iPXE, the team bare-metal provisioned any arbitrary number of Exos 4u100 AP nodes in less than 10 minutes.

Red Hat Ceph Storage cluster deployment

Following provisioning, the next step was to deploy Red Hat Ceph Storage 5 using the powerful **`cephadm bootstrap`** utility. When fed an appropriate YAML configuration file, **`cephadm`** can deploy a fully containerized storage cluster.

One of the six Exos AP application controller nodes is selected to be the first monitor and deployment orchestration point. In our case, that node had the hostname of "4u100-1a". After confirming that "4u100-1a" has passwordless SSH connectivity to all the other nodes, bootstrapping the cluster is as simple as:

```
 cephadm  bootstrap --mon-ip 172.20.2.12 --cluster-network ‘172.20.2.0/24’
--config initial-ceph.conf --registry-url registry.redhat.io --registry-
 username <OMITTED> --registry-password <OMITTED>
```

In the above command, ‘**`mon-ip`**’ was the IP address of our first monitor node, 4u100-1a. Notably, Ceph cluster settings can easily be injected into the bootstrap process. For the Exos AP 4U100 platform, we used an ‘**`initial-ceph.conf`**’ file with the following parameters:

init-ceph.conf

```
[global]

ceph config set mgr/dashboard/ssl=false

mon_allow_pool_delete=true

osd_pool_default_pg_autoscale_mode=off

osd_memory_target=4294967292

[client]

objecter_inflight_op_bytes=1073741824

objecter_inflight_ops=26240

rgw_max_concurrent_requests=10240
```

Once the first monitor and orchestrator node has been bootstrapped, the ‘`fullcluster.yml`’ file is applied to configure all other Ceph daemons on all the other nodes (see Appendix B for the fullcluster.yml script).

```
ceph orch apply -i fullcluster.yml
```

With containerized Red Hat Ceph Storage, running one or more RADOS Gateways (RGWs) on each participating Exos AP 4U100 platform is a simple process. This snippet shows the base configuration for the first RGW instance. Subsequent instances are given the following higher port numbers (e.g., 8000, 8001, etc.).

rgw.yml

```
service_type: rgw

service_id: lyve.seagate

service_name: rgw.lyve.seagate

placement:

 count_per_host: 2

 label: rgw

spec:

 rgw_frontend_port: 8000
```

Once the cluster is deployed, the last step is to create the RADOSGateway admin user as follows:

```
 radosgw-admin user create --uid=lyve --display-name=”Lyve User” --system
--access_key=<OMITTED> --secret_key=<OMITTED> --caps  “buckets=read,write;
 metadata=read,write; usage=read,write; zone=read,write”
```

Appendix B: Scripts

Kickstart script

The Anaconda Kickstart script for iPXE bare metal provisioning is as follows:

```
#version=RHEL8
# Use text mode install


# RHEL Auto boot disk selection for 4u100 AP
%pre --interpreter=/usr/bin/bash --log=/tmp/pre_ks.log
BOOTDEV=””
if [ -e /dev/disk/by-path/pci-0000:00:17.0-ata-1 ] ; then
   BOOTDEV=/dev/disk/by-path/pci-0000:00:17.0-ata-1
elif [ -e /dev/disk/by-path/pci-0000:00:17.0-ata-2 ] ; then
   BOOTDEV=/dev/disk/by-path/pci-0000:00:17.0-ata-2
elif [ -e /dev/disk/by-path/pci-0000:00:11.5-ata-1 ]; then
   BOOTDEV=/dev/disk/by-path/pci-0000:00:11.5-ata-1
elif [ -e /dev/disk/by-path/pci-0000:00:11.5-ata-2 ]; then
   BOOTDEV=/dev/disk/by-path/pci-0000:00:11.5-ata-2
elif [ -e /dev/disk/by-path/pci-0000:00:11.4-ata-1 ]; then
   BOOTDEV=/dev/disk/by-path/pci-0000:00:11.4-ata-1
elif [ -e /dev/disk/by-path/pci-0000:00:11.4-ata-2 ]; then
   BOOTDEV=/dev/disk/by-path/pci-0000:00:11.4-ata-2
elif [ -e /dev/disk/by-path/pci-0000:00:1f.2-ata-1 ]; then
   BOOTDEV=/dev/disk/by-path/pci-0000:00:1f.2-ata-1
elif [ -e /dev/disk/by-path/pci-0000:00:17.0-ata-5 ]; then
   BOOTDEV=/dev/disk/by-path/pci-0000:00:17.0-ata-5
fi
echo BOOTDEV=$BOOTDEV
(
cat <<DISKSELECT
ignoredisk --only-use=${BOOTDEV}
# System bootloader configuration
bootloader --append=”crashkernel=auto --location=partition --boot-
drive=${BOOTDEV}  console=ttyS0,115200n8”

autopart

# Partition clearing information

clearpart --all --initlabel --drives=${BOOTDEV}

DISKSELECT

)>/tmp/disk_selection.txt

### Fix up hostname from DHCP

hostnamectl set-hostname $(hostname | sed ‘s/-mgmt.*//g’)

hostnamectl set-hostname $(hostname | sed ‘s/-data.*//g’)

#wipe out EVANS drives

for X in $(ls /dev/disk/by-id/scsi-SSEAGATE_ST1600* | grep -v part)

do

     wipefs -af ${X} && dd if=/dev/zero bs=1M count=1024 of=${X} &

done

#Wipe out NYTRO drives

for X in $(ls /dev/disk/by-id/scsi-SSEAGATE_XS3840* | grep -v part)

do

     wipefs -af ${X} && dd if=/dev/zero bs=1M count=1024 of=${X} &

done

%end


%include /tmp/disk_selection.txt

sshpw --username=root root --plaintext

repo --name=”AppStream” --baseurl=http://mgmt.lyve.colo.seagate.com/repo/
rhel/pkg/8.4/rhel-8.4-x86_64-dvd/AppStream/

eula --agreed

reboot

selinux --disabled

firewall --disabled

firstboot --disable

skipx
```

```
%packages
@^minimal-environment
@development
@headless-management
@network-tools
@rpm-development-tools
@system-tools
iperf3
ipmitool
kexec-tools
nfs-utils
rsync
sg3_utils
tmux
traceroute
vim-enhanced
dmidecode
ipmitool
wget
efibootmgr
%end


# System language
lang en_US.UTF-8


# Network information - set by DHCP!

#network  --hostname=localhost.localdomain


# Use network installation
url --url="http://mgmt.lyve.colo.seagate.com/repo/rhel/pkg/8.3/
rhel-8.3-x86_64-dvd/"

# Run the Setup Agent on first boot
```

```
firstboot --enable

# System timezone
timezone America/Denver --isUtc
# Root password

rootpw --iscrypted <OMITTED>
user --groups=wheel --name=lyve --password=<OMITTED>

%addon com_redhat_kdump --enable --reserve-mb=’auto’

%end

%post
```

Fullcluster.yml script

The fullcluster.yml script is provided.

```
service_type: host
addr: 4u100-1a
hostname: 4u100-1a
labels:
- mon
- mgr
- osd
- rgw
---
service_type: host
addr: 4u100-1b
hostname: 4u100-1b
labels:
- osd
- rgw
---
service_type: host
```

```
service_type: alertmanager
service_name: alertmanager
placement:
  count: 1
---
service_type: crash
service_name: crash
placement:
  host_pattern: ‘*’
---
service_type: grafana
service_name: grafana
placement:
  count: 1
---
service_type: mgr
```

```
service_type: osd
service_id: base_drivegroup
service_name: osd.base_drivegroup
placement:
  host_pattern: 4u100-[1-3][ab]
spec:
  block_db_size: 510GB
  data_devices:
      limit: 42
      rotational: 1
      size: ‘13TB:’
  db_devices:
      limit: 6
      rotational: 0
      size: 3TB:5TB
```

```
addr: 4u100-2a
hostname: 4u100-2a
labels:
- mon
- mgr
- osd
- rgw
---
service_type: host
addr: 4u100-2b
hostname: 4u100-2b
labels:
- osd
- rgw
- rgw
---
service_type: host
addr: 4u100-3a
hostname: 4u100-3a
labels:
- mon
- mgr
- osd
- rgw
---
service_type: host
addr: 4u100-3b
hostname: 4u100-3b
labels:
- osd
- rgw
```

```
service_name: mgr
placement:
  hosts:
  - 4u100-1a
  - 4u100-2a
  - 4u100-3a
---
service_type: mon
service_name: mon
placement:
  hosts:
  - 4u100-1a
  - 4u100-2a
  - 4u100-3a
---
service_type: node-exporter
service_name: node-exporter
placement:
  host_pattern: ‘*’
service_type: prometheus
service_name: prometheus
placement:
  count: 1
```

```
  db_slots: 7
  filter_logic: AND
  objectstore: bluestore
---
service_type: rgw
service_id: lyve.seagate
service_name: rgw.lyve.seagate
placement:
  count_per_host: 2
  label: rgw
spec:
  rgw_frontend_port: 8000
```

Cluster test script

The script used for cluster testing follows:

```
for OBJ_SIZE in $((128 * 1024)) \
             $((512 * 1024)) \
             $((1 *   1024 *1024)) \
             $((4 *   1024 *1024)) \
             $((16 *  1024 *1024)) \
             $((64 *  1024 *1024))
do
      echo “Object Size = ${OBJ_SIZE}”
      for CONC in 8 16 24 48 64 96 128
      do
             CONC_STR=$(printf “%03d” $CONC)

             for OP in put get
             do
                    for SCH in ${SCHEDULERS[*]};
                    do
                                   echo “=================================”
TEST_NAME=”${OP}-${CONC_STR}-${OBJ_SIZE}-${SCH}”
                                   DEL_OBJS=$(($CONC * 400))
                                   OPTIONS=” “
                                   OPTIONS=”${OPTIONS} --access-key=${ACCESS_
KEY} “
                                   OPTIONS=”${OPTIONS} --secret-key=${SECRET_
KEY} “
                                   OPTIONS=”${OPTIONS} --host=${SERVER_HOSTS}
“
                                   OPTIONS=”${OPTIONS} --warp-client=${CLIENT_
HOSTS} “
                                   OPTIONS=”${OPTIONS} --obj.size=${OBJ_SIZE}
“
                                   OPTIONS=”${OPTIONS} --benchdata=${TEST_
NAME} “
                                   OPTIONS=”${OPTIONS} --duration=7m30s “
                                   OPTIONS=”${OPTIONS} --concurrent=${CONC} “
                                   OPTIONS=”${OPTIONS} --autoterm --autoterm.
dur=20s “
                                   OPTIONS=”${OPTIONS} --quiet --noclear “
                                   echo “Starting ${TEST_NAME}”
                                   sysmonitor_start “${TEST_NAME}”
                                   echo “Launching Warp”
                                   echo “warp ${OPTIONS}”
                                   warp ${OPTIONS}
                                   sysmonitor_stop
                                   echo “Completed ${TEST_NAME}”
                                   echo “--------------------------------”
                        done
             done
      done
done
```

Configuring for a different sized Red Hat Ceph Storage cluster

Using the Red Hat Ceph Storage **`cephadm bootstrap`** command, engineers simply modified the **`osd.yml`** section of the **`fullcluster.yml`** file to configure a cluster to simulate half-populated Exos AP 4U100 enclosures.

Configuration for full 42 HDDs per application controller (two hot spares)

```
service_type: osd
service_id: base_drivegroup
service_name: osd.base_drivegroup
placement:
  host_pattern: 4u100-[1-3][ab]
spec:
  block_db_size: 510GB
  data_devices:
      limit: 42
      rotational: 1
      size: ‘13TB:’
  db_devices:
      limit: 6
      rotational: 0
      size: 3TB:5TB
  db_slots: 7
  filter_logic: AND
  objectstore: bluestore
```

Configuration for 24 HDDs Drive per application controller

```
service_type: osd
service_id: base_drivegroup
service_name: osd.base_drivegroup
placement:
  host_pattern: 4u100-[1-3][ab]
spec:
  data_devices:
      limit: 24
      rotational: 1
      size: ‘13TB:’
  db_devices:
      limit: 6
      rotational: 0
      size: 3TB:5TB
  filter_logic: AND
  objectstore: bluestor
```

About Seagate

Seagate Technology crafts the datasphere, helping to maximize humanity's potential by innovating world-class, precision-engineered mass-data storage and management solutions with a focus on sustainable partnerships. A global technology leader for more than 40 years, the company has shipped over three billion terabytes of data capacity. Learn more about Seagate by visiting www.seagate.com. Learn more about Seagate by visiting www.seagate.com or following us on Twitter, Facebook, LinkedIn, YouTube, and subscribing to our blog.

About Red Hat

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers develop cloud-native applications, integrate existing and new IT applications, and automate and manage complex environments. A trusted adviser to the Fortune 500, Red Hat provides award-winning support, training, and consulting services that bring the benefits of open innovation to any industry. Red Hat is a connective hub in a global network of enterprises, partners, and communities, helping organizations grow, transform, and prepare for the digital future.

facebook.com/redhatinc
@RedHat
linkedin.com/company/red-hat

NORTH AMERICA
1 888 REDHAT1

EUROPE, MIDDLE EAST, AND AFRICA
00800 7334 2835
europe@redhat.com

ASIA PACIFIC
+65 6490 4200
apac@redhat.com

LATIN AMERICA
+54 11 4329 7300
info-latam@redhat.com

redhat.com
#F29951_1021