Red Hat Data Services solutions

Scaling to 10 billion objects

Highlights

Demonstrated scalability to more than 10 billion objects in a six-node, 4.5 PB Red Hat Ceph Storage cluster

Intel solid state drive (SSD) data center (DC) technology played a key role in achieving deterministic performance at scale.

Massively scalable Seagate Exos E 4U106 storage enclosures with nearline enterprise drives delivered spatial capacity to store 10 billion objects or more.

Resilient software-defined Red Hat Data Services solutions for availability and business continuity

Introduction

With the growing embrace of containers and Kubernetes orchestration, container-native object storage is now particularly vital technology for diverse applications. As applications move to the cloud, persistent data storage is an important factor in cloud-based application scalability, performance, and reliability. Solutions must be able to scale rapidly to support enormous numbers of objects with predictable performance, and they must be able to keep vital applications and services running during planned and unforeseen failure events.

With these challenges in mind, Red Hat recently engaged Evaluator Group to test the scale and performance of a system designed to store over 10 billion objects running on Red Hat® Ceph® Storage. Scaling to 10 billion objects is a monumental undertaking that requires highly capable hardware, a balanced configuration, and a software environment that can scale and recover from failure without placing arbitrary limitations on performance. The tested configuration included Intel and Seagate technology:

- Servers equipped with Intel Xeon processors.
- A 4.5 petabyte (PB) Red Hat Ceph Storage cluster using Seagate's Exos E 4U106 high-density storage enclosures.
- Seagate Exos X16 16TB SAS nearline enterprise drives for capacity.
- Intel solid state drive (SSD) data center (DC) P4610 Series for Ceph Bluestore metadata.

The successful configuration demonstrated deterministic, nearly linear performance as system capacity grew to more than 10 billion objects and 80% of usable capacity. Engineers observed the following results using the Amazon Simple Storage Service (S3) protocol driven (in parallel) by running the COSBench object storage workload generation tool.

- Small 64KB objects:
  - An average of more than 28,000 objects/second for GET (read) operations
  - An average of more than 17,000 objects/second for PUT (write) operations
- Large 128MB objects:
  - An average of more than 11.6GB/second GET (read) bandwidth.
  - An average of more than 10.6 GB/second PUT (write) bandwidth

Scaling to 10 billion objects with Intel, Seagate, and Red Hat Data Services solutions

Cloud-based applications depend on cloud-native and container-native data services. Red Hat offers a comprehensive portfolio of technologies designed for cloud-native applications.

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

- Red Hat OpenShift® Container Platform offers a consistent Kubernetes-based hybrid cloud foundation for building and scaling containerized applications. Trusted by more than 2,000 customers, OpenShift Container Platform helps deliver business-critical applications for migrating existing workloads or building new cutting-edge experiences.
- Red Hat OpenShift Container Storage is software-defined storage for containers. Engineered as the data and storage services platform for Red Hat OpenShift, it offers both internal mode and external mode storage options, allowing organizations to choose the storage platform that makes sense for their applications and environment.
- Red Hat Ceph Storage is an open, massively scalable, simplified storage solution for modern data pipelines. Delivering software-defined storage on a choice of industry-standard hardware, Red Hat Ceph Storage clusters like the one under test can offer independent storage scalability, performance optimization, and simultaneous access from multiple Red Hat OpenShift clusters through OpenShift Container Storage external mode.

Test configuration

One of the principal advantages of software-defined Red Hat Ceph Storage is that it can be deployed on industry-standard servers with storage tailored to provide appropriate capacity and performance profiles for a given workload. The Evaluator Group project sought to configure and test a Red Hat Ceph Storage cluster capable of ingesting, storing, and serving 10 billion small objects with deterministic performance. Testing was conducted in Evaluator Group labs using equipment from Intel and Seagate. Table 1 details the test configuration used by Evaluator Group, with key Intel and Seagate technologies described in the sections that follow.

Table 1. Evaluator Group test configuration

Component	Configuration details	Total capacities
6x Red Hat Ceph Storage servers	Intel 2U, 2-socket server - 2x Intel Xeon Gold 6154 processors - 18 cores - 36 cores per server, 384GB RAM - 6x Intel SSD DC P4610 7.6TB - 45TB of Ceph BlueStore metadata - LSI/Broadcom SAS 3108 HBA (single 4x 12Gb/s port for 48Gb/s) - 2x 25Gb/s Ethernet network adaptors	- 216 processor cores - 2.3TB RAM - 273.6TB metadata cache - 12 rack units

Component	Configuration details	Total capacities
3x storage subsystems	Seagate Exos E 4U106 enclosure ▸ 106x Seagate Exos X16 enterprise nearline SAS 16TB drives ▸ (53 per Ceph storage server) ▸ Split between two Ceph storage servers, shared nothing configuration	▸ 318x 16TB HDDs = 4.8PB of raw storage capacity ▸ 12 rack units
6x workload clients	Intel 2U, 2-socket server ▸ 2x Intel Xeon E5-2699 v4, ▸ 22 core ▸ 44 cores per server, 256GB RAM ▸ 8x NVMe drive support ▸ 2x 25Gb/s Ethernet network adaptors	▸ 264 processor cores ▸ 1.5TB RAM ▸ 12 rack units
Network	Mellanox SN2100, 24x 25Gb/s Ethernet	▸ 50Gb/s per node, bonded
Object data protocol	S3 (GET, PUT, LIST, DELETE)	
Data protection	Red Hat Ceph Storage erasure coding	▸ 4+2 erasure coded pool

Intel Xeon Scalable Processors and Intel SSDs

Intel Xeon Gold processors powered the Red Hat Ceph Storage cluster in the large-scale object testing. With support for higher memory bandwidth, enhanced memory capacity per core, and 48 lanes of PCIe, Intel Xeon Gold processors are optimized for demanding mainstream datacenter, multi-cloud compute, network, and storage workloads. Intel Xeon processors were used for load-generating workload clients.

Red Hat Ceph Storage performance benefits from fast storage for metadata caching for Ceph object storage daemons (OSDs). As a part of testing, engineers configured Intel SSD DC P4600 series on Red Hat Ceph Storage servers to serve as the Ceph BlueStore metadata device. The Intel SSD DC P4610 is architected with 64-layer TLC Intel 3D NAND technology, offering performance, quality of service, and capacity improvements over previous generations. Each Red Hat Ceph Storage node was configured with six Intel SSD DC P4610 for a total of 45TB of Ceph BlueStore metadata capacity.

Seagate Exos E 4U106

With support for 106 high-capacity hard disk drives (HDDs) in a single enclosure, the Seagate Exos E 4U106 represents one of the industry's largest storage building blocks, delivering significant capacity and density without sacrificing data access speed. By storing up to 1.9 petabytes in a single 4 rack unit (RU) enclosure (using 18TB HDDs) and providing overall maximum bandwidth of 36GB/s, organizations can minimize datacenter footprint and power consumption while maximizing storage space.

In the configuration tested, each Seagate Exos E 4U106 enclosure was shared between two Red Hat Ceph Storage nodes, in a split chassis shared-nothing configuration. Each of the six Red Hat Ceph Storage nodes connected to the Exos enclosures via a 48Gb/s host bus adapter (HBA) with a single 4x 12Gbps port (48Gbps). Connected to 53 16TB HDDs, each Red Hat Ceph Storage node had access to 848TB of dedicated storage.

Performance summary

Red Hat Ceph Storage was configured using an erasure coded object pool, offering fault tolerance while maximizing storage capacity. Erasure coding (4+2) represented a particularly good fit for the six-node configuration as it supported the loss of up to two nodes or multiple devices across multiple nodes.[1]

PUT operations (writes) are necessarily more expensive in terms of input/output (I/O) operations than GET operations (reads). With 4+2 erasure coding, every 64KB object was split into four chunks of 16KB each and stored along with two 16KB of parity. As a result, every object stored required six I/O operations for every PUT operation.

Small object performance

Small (64KB) workload testing was conducted using the COSbench object storage workload tool employing 12 COSbench clients, each with a dedicated Ceph storage target URL (Ceph RADOS Gateway, RGW, using the S3 interface). Each run consisted of three separate workloads with the following parameters:

- A write-only portion that generated a total of 76.8 million 64KB objects
- A read-only portion that read as many 64KB objects as possible in 10 minutes
- A mixed 10-minute workload (70% read, 20% write, 5% list, and 5% delete)

Figure 1 shows the small-object workload performance for PUTs (writes) and GETs (reads) as the object count scaled to 10 billion objects. As shown in the chart, there was a brief, significant drop in both PUT and GET performance after approximately 500 million objects. This was attributed to Ceph initiating a "deep scrubbing" operation due to the very high PUT rates. Red Hat Ceph Storage was reconfigured to decrease the rate of deep scrubbing and performance quickly returned to normal. GET operation performance began to degrade only after the metadata cache capacity was reached and the cluster's usable capacity surpassed 80%. At this point metadata increasingly spilled over from fast Intel SSD DC P4610s to slower HDDs.

1 For production environments seven nodes are recommended for an erasure-coded 4+2 profile.

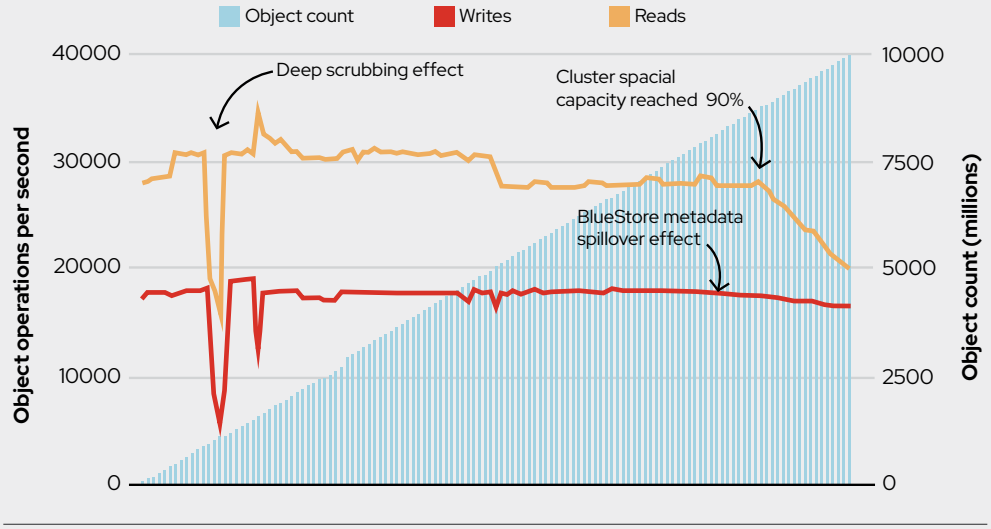

Figure 1. Object count vs. operations per second for up to 10 billion small objects

Large object performance

The primary consideration for this testing was read/write performance for small 64KB objects in terms of objects per second, and the system was optimized accordingly. However, engineers also wanted to evaluate throughput using larger (128MB) objects that are representative of big data workloads such as MapReduce, Spark, and Presto. Testing, again, utilized workloads with multiple operation types that were run repeatedly as the system capacity increased to over 80% utilization.

As with the small-object testing, 12 COSbench clients were employed, with each using a dedicated Ceph RGW interface, including:

- A write-only (PUT) portion that generated a total of 288 thousand 128MB objects.
- A read-only (GET) portion that ran for 30 minutes and read as many objects as possible.
- A mixed workload that ran for 30 minutes (70% read, 20% write, 5% list, and 5% delete).

As shown in Figure 2, Red Hat Ceph Storage demonstrated deterministic performance of more than 10GB/s for both GET and PUT operations – with almost no variation as the system filled to 80% of its capacity. Because the testing involved fewer large objects, it generated less metadata and there was no noticeable spill-over from the Intel SSD DC P4610s to HDDs. Importantly, the system was optimized for small objects, not large objects, with network bandwidth limited to 50Gb/s per node and a server-to-storage bandwidth limitation of 48Gb/s over a single SAS connector. Engineers anticipate overall object throughput rates could have been higher without these limitations.

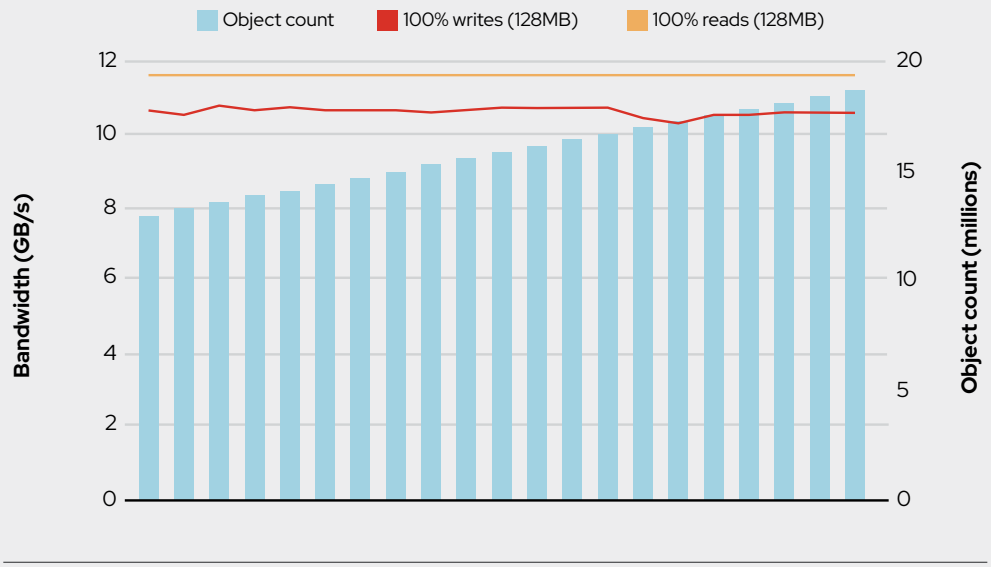

Figure 2. Bandwidth versus capacity for large objects

Performance under failure conditions

Any storage cluster designed to ingest and serve billions of objects is a vital resource that must be highly available. While the Seagate Exos E 4U106 enclosures feature hot-swap disk drives, power supplies, fans, and SAS expander modules, Red Hat engineers sought to evaluate performance under failover conditions, using the erasure coded (4+2) data protection inherent to the tested configuration. As shown in Figure 3, under the small object workload, operations per second continued at a high level even when six individual HDDs in a single enclosure were artificially failed within the cluster.

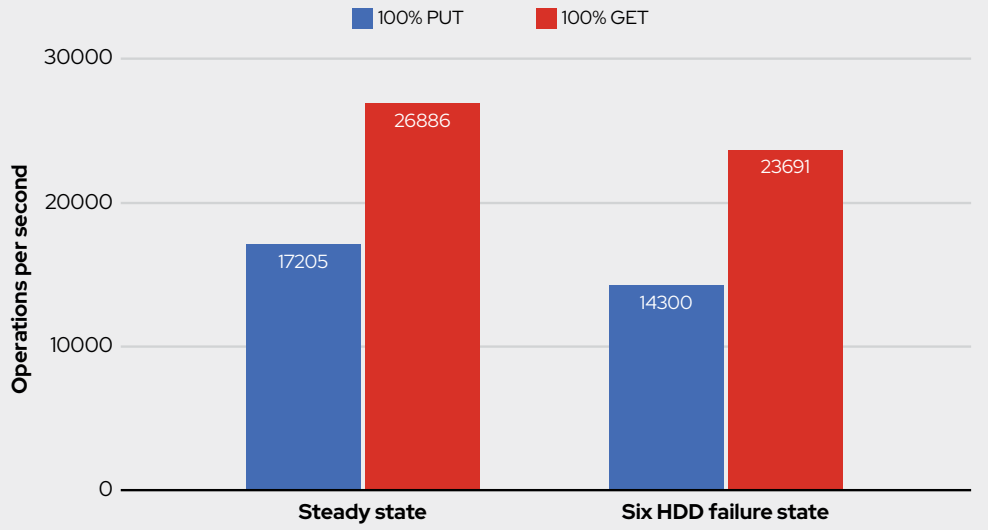

Figure 3. Small object (64KB) steady state versus failure state performance

In addition to individual HDD failures, engineers also recorded bandwidth during degraded state operation with failure of a full Red Hat Ceph Storage node, taking 53 HDDs out of the cluster. Figure 4 shows throughput for the large-object workload, considering steady state operation as well as failure of six individual disks and an entire Red Hat Ceph Storage node. The results demonstrated that the cluster was resilient to both device and node failure, allowing processing to continue at a high level, even under failure conditions.

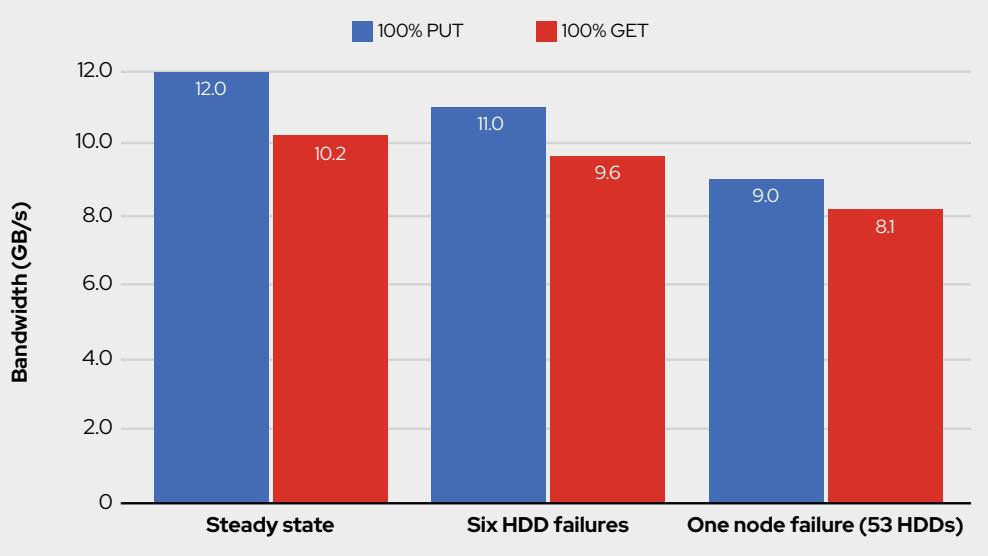

Figure 4. Large object (128MB) steady state vs. failure state performance

Conclusion

Evaluator Group testing validated that Red Hat Ceph Storage can scale effectively to serve more than 10 billion objects with only six storage servers, expanding the tested object scale by an order of magnitude over previous Red Hat object storage tests. With support for industry-standard servers and a choice of storage solutions, this software-defined solution facilitates custom solutions to serve specific needs.

For large object-count workloads, high performance Intel Xeon Gold processors provide Red Hat Ceph Storage with considerable computational power, while Intel SSD DC P4610 Series provide high speed and high capacity metadata caching for performance scalability. Seagate Exos E 4U106 enclosures and Seagate Exos X16 HDDs provide the storage capacity needed to serve massive object storage workloads without compromising performance. Red Hat Ceph Storage unifies these elements to offer massive scalability and resilience in an open software-defined storage solution that can be used as an external mode storage cluster for OpenShift Container Platform to serve applications running on OpenShift Container Platform.

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 integrate new and existing IT applications, develop cloud-native applications, standardize on our industry-leading operating system, and automate, secure, and manage complex environments. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500. As a strategic partner to cloud providers, system integrators, application vendors, customers, and open source communities, Red Hat can help organizations prepare for the digital future.

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

North America
1 888 REDHAT1
www.redhat.com

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
#F25947_1120