Red Hat

Ceph BlueStore and OpenStack

Evaluating Red Hat Ceph Storage and MySQL

Deploy Red Hat Ceph Storage for critical OpenStack applications like MySQL.

Explore hyperconverged compute and storage solutions that combine Red Hat OpenStack Platform and Red Hat Ceph Storage on a single platform.

Employ Ceph BlueStore technology to accelerate storage performance and reduce latency on all-flash storage systems.

Table of contents

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

Introduction

Ceph Storage remains a popular software-defined storage solution for applications based on OpenStack.[1] Organizations worldwide run hyperscale production workloads on Red Hat® Ceph® Storage and Red Hat OpenStack® Platform, and benefit from advanced integration between the two platforms. With each release, the level of integration has grown and performance and automation have increased.

There is growing interest in running compute and storage as a single unit, providing a hyperconverged infrastructure (HCI) layer based on OpenStack and Ceph. Ideal for infrastructure core and edge workloads, HCI offers colocated, software-defined compute and storage resources and a unified management interface, all running on industry-standard hardware. Red Hat OpenStack Platform and Red Hat Ceph Storage can be combined on single servers to offer common life cycle and support (Figure 1).

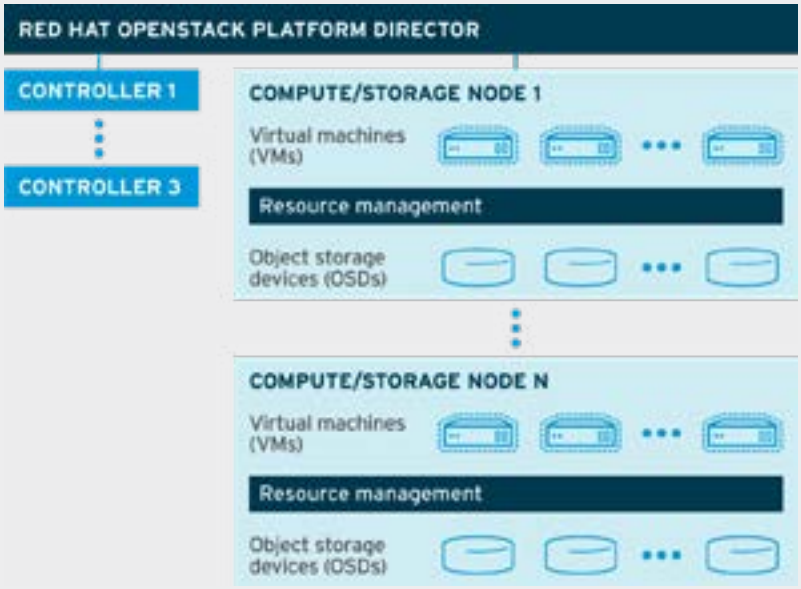

Figure 1. A hyperconverged architecture featuring Red Hat OpenShift® Platform and Red Hat Ceph Storage

As a part of evaluating potential solutions, the Red Hat Storage Solution Architectures team wanted to explore the performance impact of moving from decoupled infrastructure (i.e., Red Hat Ceph Storage and Red Hat OpenStack Platform running as two independent pools of servers) to hyperconverged infrastructure. They also evaluated the performance implications of Ceph BlueStore technology along with all-flash media.

1 *"2018 OpenStack User Survey Report."* OpenStack, 2018.

Engineers configured a cluster to serve a MySQL workload with options that included:

- Two different deployment strategies (decoupled or hyperconverged)
- Two different media types (mixed/hard disk drive (HDD) media vs. all-flash NVMe cluster)
- Two different Ceph object storage daemon (OSD) backends (FileStore vs. BlueStore)

Configuration and performance summary

The Red Hat testing methodology consisted of performing multiple iterations of MySQL database queries with 100% read, 100% write, and a 70%/30% read/write mixture. The results were evaluated in terms of database transactions per second (TPS), queries per second (QPS), average latency, and 99th percentile tail latency.

Three configurations were assembled on a mixed/HDD cluster, including:

- A decoupled architecture with a standalone Red Hat Ceph Storage FileStore OSD backend.
- A hyperconverged architecture with Red Hat Ceph Storage FileStore OSD backend.
- A hyperconverged architecture with Red Hat Ceph Storage BlueStore backend.

Subsystem metrics captured during the hyperconverged BlueStore write and read tests showed that limitations for spinning media resulted in increased latencies for both write and read MySQL database workloads. Low latency is critical for database workloads and media latency can adversely affect application performance. Another set of hyperconverged tests were run on a separate set of Ceph OSD servers equipped with all-flash storage, again using the BlueStore OSD backend.

Ceph FileStore and BlueStore OSD backends

Within Ceph, OSDs store most of the data. The Ceph FileStore OSD backend is the original Ceph object store. As shown in Figure 2, FileStore relies on a standard file system (normally XFS) to read and write from raw block devices. It works in combination with a key/value database (traditionally LevelDB). Ceph FileStore is well tested and widely used, but its layered design can cause performance deficiencies.

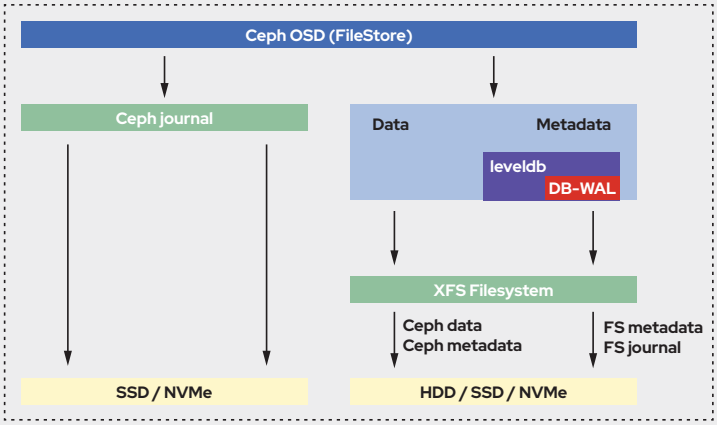

Figure 2. FileStore implementation

BlueStore (Figure 3) is a special-purpose storage backend designed specifically for managing data on disk for Ceph OSD workloads. Unlike the FileStore backend, BlueStore stores objects directly to block devices without any file system interference. BlueStore can improve performance for the cluster, with features that include:

- Direct management of storage devices
- Metadata management with RocksDB
- Full data and metadata checksumming
- Inline compression
- Multidevice metadata tiering
- Efficient copy-on-write

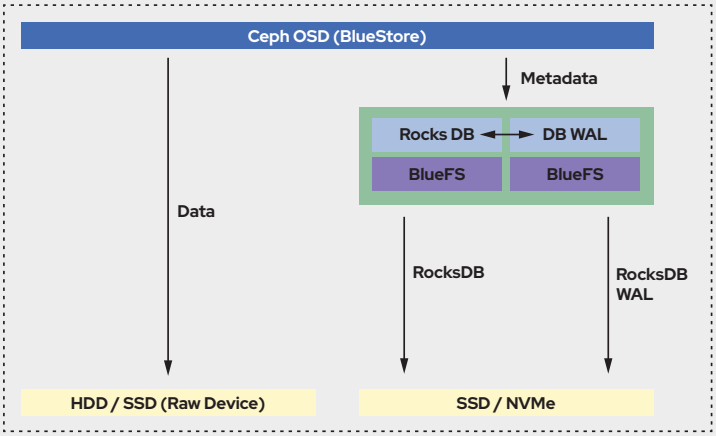

Figure 3. BlueStore implementation

Hardware components

Table 1 lists the hardware configuration used in the mixed/HDD Red Hat testing. The same systems were used for both decoupled and hyperconverged testing though software component configurations differed for the two approaches.

Table 1. Hardware and software configuration for Red Hat testing

Hardware detail	Decoupled (FileStore)	Hyperconverged (FileStore and BlueStore)
5x Dell EMC PowerEdge R730xd servers: ▸ 2x Intel Xeon E5-2630 v3 (8 cores @2.4GHz) ▸ 256GB memory ▸ 2x 10GbE networking ▸ 12x 6TB SATA HDDs 7.2K RPM ▸ 1x 800GB Intel DC P3799 NVMe	▸ Red Hat Ceph Storage with FileStore OSDs	▸ Red Hat Ceph Storage FileStore or BlueStore OSDs ▸ Red Hat OpenStack Platform compute nodes
3x Dell EMC PowerEdge R630 servers: ▸ 2x Intel Xeon E5-2650 v3 (10 cores @2.30GHz) ▸ 128GB memory ▸ 2x 10GbE networking	▸ Red Hat Ceph Storage MON ▸ Red Hat OpenStack Platform compute ▸ Red Hat OpenStack Platform Controller	▸ Red Hat Ceph Storage MON ▸ Red Hat OpenStack Platform controller
8x Dell EMC PowerEdge R220 servers: ▸ 1x Intel Celeron G1820 (2 cores @2.7GHz) ▸ 16GB memory ▸ 2x 1GbE networking	▸ Red Hat OpenStack Platform Director (1) ▸ RADOSbench (7)	
10x VM instances: ▸ 4x vCPU ▸ 8GB memory	▸ Red Hat OpenStack Platform Instance (VM) ▸ 40GB Ceph RBD (OS) ▸ 1x 100GB Ceph RBD (Cinder volume)	
2x Dell EMC Force10 switches (10GbE)	▸ Networking	

All-flash testing substituted Cisco unified computing system (UCS) servers as OSD servers, as configured in Table 2.

Table 2. Ceph OSD servers for all-flash testing

6x Cisco UCS C220-M5 servers	▸ 2x Intel Xeon Platinum 8180 (28 cores @2.5GHz) ▸ 192GB memory ▸ 7x 4TB Intel SSD DC P4500 NVMe ▸ 1x Intel Optane SSD DC P4800 375GB ▸ 2x 40GbE networking

Software components

Software components used in Red Hat testing included the following:

- Red Hat Enterprise Linux® 7.5 (7.6 for all-flash configurations)
- Red Hat Ceph Storage 3.2
- Red Hat OpenStack Platform 13.0
- FIO 2.41
- Sysbench 1.0.1
- MySQL 5.7

Overall performance summary

Red Hat MySQL testing consisted of 10 MySQL instances provisioned on top of 10 OpenStack VM instances, keeping the ratio of MySQL databases to VM instances at 1:1. Each MySQL database instance was configured to use a 100GB block device provisioned by OpenStack Cinder from Red Hat Ceph Storage. The Sysbench online transaction processing (OLTP) application benchmark was used to exercise the MySQL database. MySQL used the InnoDB storage engine with data and log files stored on Red Hat Ceph Storage block storage.

The MySQL InnoDB engine maintains an in-memory cache called the buffer pool. Performance can be directly affected by the ratio of the dataset size to the size of the buffer pool. For example, if the buffer pool is large enough to hold the entire dataset, most read operations will be served through the in-memory cache and never generate input/output (I/O) traffic.

Because the intent was to benchmark the underlying storage, the ratio of the InnoDB buffer pool to the dataset was kept to 1:25 (buffer pool size of 5GB and dataset size of 75GB). This ratio ensured that the database could not fit in system memory, leading to increased I/O traffic from the InnodB storage engine to Red Hat Ceph Storage block storage.

Engineers tested the following workload patterns:

- 100% write transactions
- 100% read transactions
- 70%/30% read/write transactions

They gathered the following metrics:

- Transactions per second (TPS)
- Queries per second (QPS)
- Average latency (ms)
- 99th percentile tail latency

Observations led to the following performance highlights:

- The HCI BlueStore configuration showed both ~50% lower average latency and lower 99th percentile tail latency than the HCI Ceph FileStore configuration.
- Red Hat OpenStack Platform and Red Hat Ceph Storage deployed in the HCI configuration showed performance parity compared to the decoupled (classic standalone) configuration. Almost no performance drop was observed while moving from the decoupled architecture to the HCI architecture.
- For MySQL write workloads, Ceph BlueStore deployed on an all-flash (NVMe) cluster showed 16x higher TPS, 14x higher QPS, ~200x lower average latency, and ~750x lower 99th percentile tail latency than when deployed on mixed/HDD media.
- For MySQL read workloads, Ceph BlueStore deployed on an all-flash (NVMe) storage cluster delivered up to ~15x lower average and 99th percentile tail latency, than on a mixed/HDD media-based cluster.
- Similarly, for a 70%/30% read/write mix workload, Ceph BlueStore OSD deployed on an all-flash (NVMe) cluster showed up to 5x higher TPS and QPS compared to a mixed/HDD media-based cluster.

MySQL 100% write performance

Write performance was compared between HDD-based configurations and also with an all-flash Red Hat Ceph Storage cluster.

HDD-based configurations

Figures 4 and 5 show write performance comparisons in terms of TPS and QPS, respectively – using the same mixed/HDD media-based cluster configuration. These figures show performance of decoupled FileStore and HCI FileStore to be nearly identical. As such, the performance loss in moving from the decoupled to the HCI architecture was minimal.

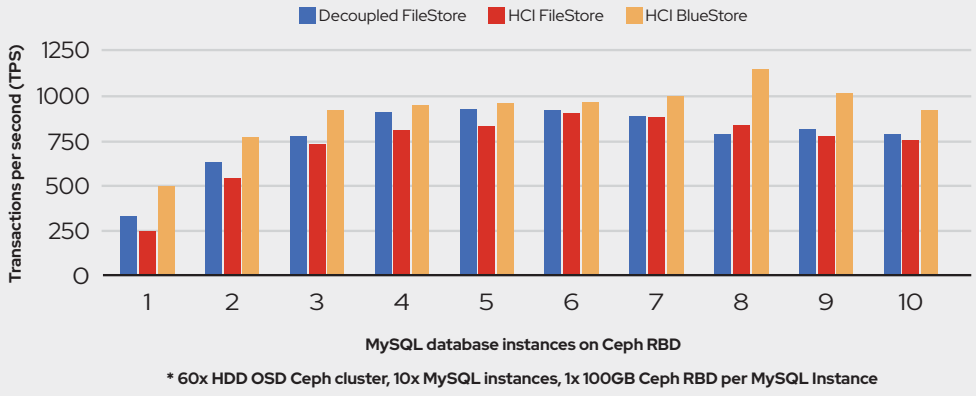

Figure 4. Write transactions per second across three cluster types (higher is better)

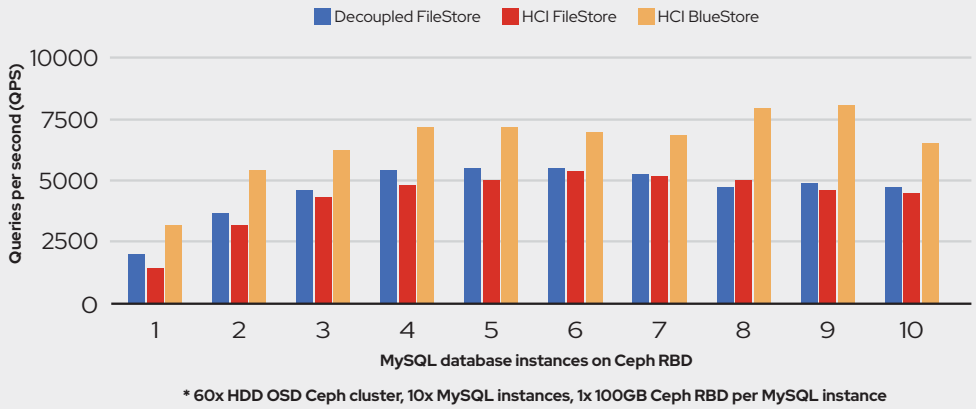

Figure 5. Write queries per second across three cluster types (higher is better)

As expected, the hyperconverged Ceph BlueStore configuration showed higher write TPS and QPS compared to the HCI FileStore configuration. The Ceph BlueStore OSD backend stores data directly on the block devices without any file system interface, which improves cluster performance.

Figures 6 and 7 illustrate average write latency and write 99% tail latency, respectively. These data show that Ceph BlueStore delivers ~50% lower average latency and better 99th percentile tail latency than the Ceph FileStore backend on the same hardware.

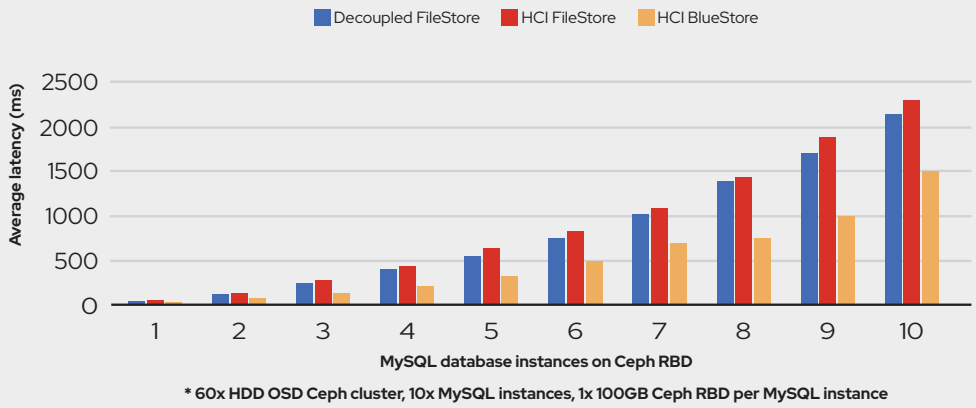

Figure 6. Average write latency across three cluster types (lower is better)

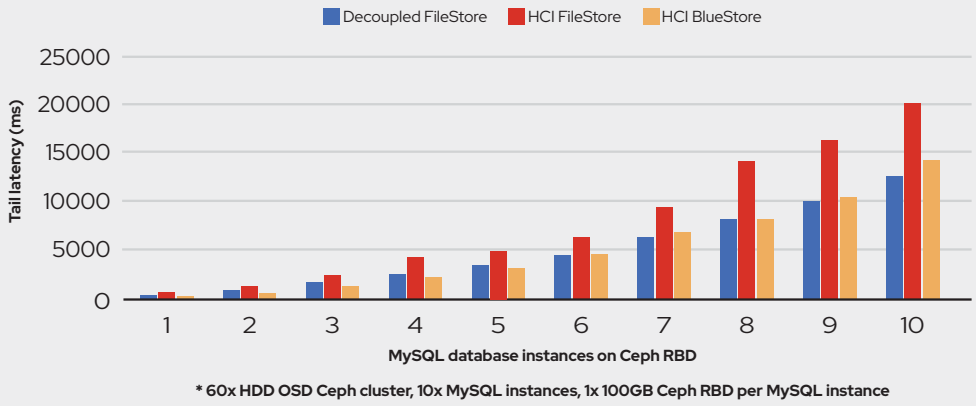

Figure 7. Write 99% tail latency across three cluster types (lower is better)

All-flash configuration

When running the HCI BlueStore tests on a mixed/HDD cluster, engineers observed relatively high transactional latencies due to bottlenecking by the higher latencies of spinning drives. As such, 100% write TPS and QPS performance with the BlueStore OSD backend on HDDs did not scale as expected.

To explore this issue, engineers ran a similar test on an all-flash configuration based on Cisco UCS OSD servers. These additional tests showed the lower latencies and higher TPS performance Ceph BlueStore can deliver (Figures 8 and 9). Running Ceph BlueStore on an all-flash (NVMe) configuration delivered improved performance compared to Ceph BlueStore on mixed HDDs, with substantially lower latency.

When compared to Ceph BlueStore on mixed/HDD media, Ceph BlueStore on all-flash delivered:

- 16x higher TPS
- 14x higher QPS
- ~200x lower average latency
- ~750x lower 99th percentile latency

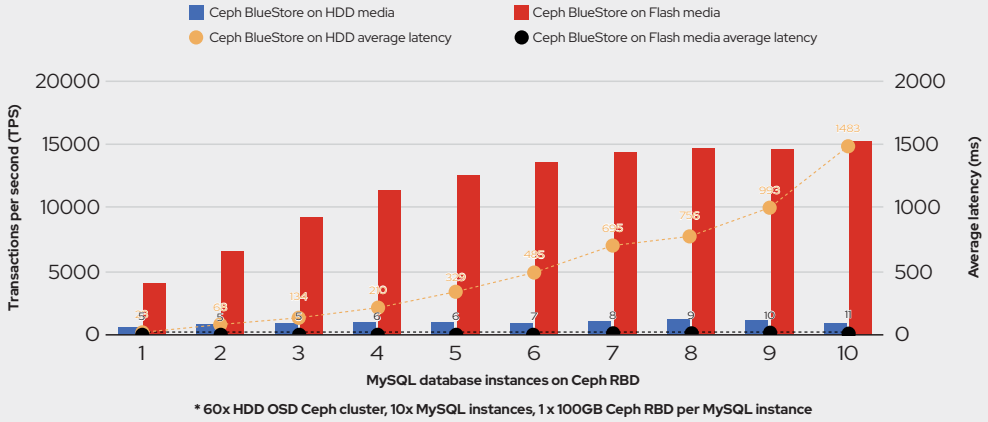

Figure 8. BlueStore and flash media provide higher transactions per second and lower latency than mixed/HDD configuration.

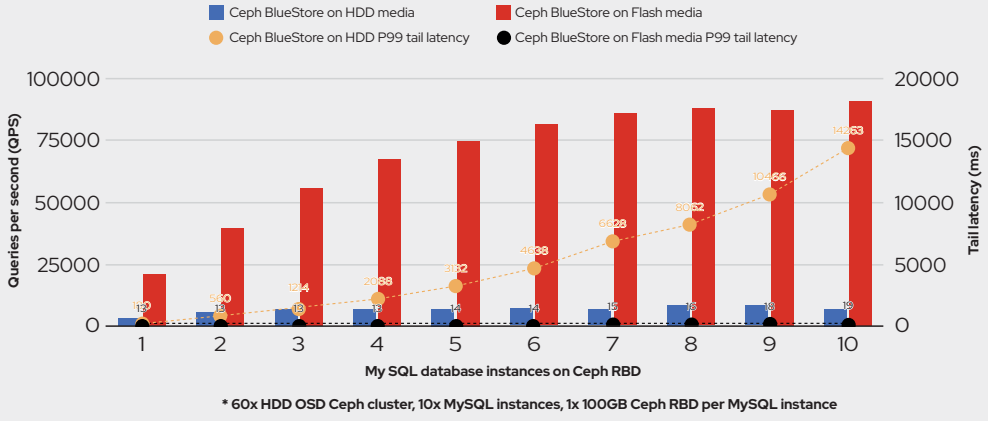

Figure 9. BlueStore and flash media provide higher queries per second and lower latency than a mixed/HDD configuration.

BlueStore was developed to utilize the available performance of flash media more fully. As such, small-block, high-transactional workloads (e.g., databases) are appropriate for flash media instead of HDDs.

100% write performance summary

When serving 100% write workloads, Red Hat OpenStack Platform and Red Hat Ceph Storage showed performance parity in an HCI configuration compared to the decoupled (classic standalone) configuration. The HCI Ceph BlueStore configuration showed both ~50% lower average latency and 99th percentile tail latency when compared to the HCI Ceph FileStore configuration. Similar tests showed a higher write transaction rate and higher queries per second on HCI Ceph in BlueStore configuration.

When deployed on the all-flash (NVMe) based cluster, Ceph BlueStore showed 16x higher TPS, 14x higher QPS, ~200x lower average latency, and ~750x lower 99th percentile tail latency. Ceph BlueStore, together with all-flash (NVMe) hardware is recommended when performance is paramount for high transactional workloads.

MySQL 100% read performance

As with the write performance comparison, read performance was compared across two deployment strategies (standalone clusters vs. HCI), as well as two Ceph OSD backends (FileStore vs. BlueStore) and two media types (mixed/HDD vs. all-flash).

HDD-based configurations

TPS and QPS for the 100% read workload are shown in Figures 10 and 11, respectively. The hyperconverged configuration with a Ceph BlueStore backend showed strong performance until six parallel MySQL instances were present. As the number of workload generators increased, a minor performance difference emerged, likely from the colocation of compute and storage on the hyperconverged configuration. Ceph FileStore and BlueStore OSD backends showed similar read performance for both TPS and QPS for the HCI configuration.

Importantly, Ceph FileStore uses XFS, which relies on kernel memory management that consumes all the free memory for caching. Ceph BlueStore, on the other hand, is implemented in user space where the BlueStore OSD daemon manages its own cache, resulting in a lower memory footprint and a smaller cache.[2]

2 *This testing was conducted using Red Hat Storage 3.0. Later versions (e.g., Red Hat Ceph Storage 3.2) introduce a new osd_memory_target option that tells the BlueStore backend to adjust its cache size in an attempt to cache more. Though not tested, we expect later versions of Red Hat Ceph Storage to deliver better performance compared to Ceph FileStore.*

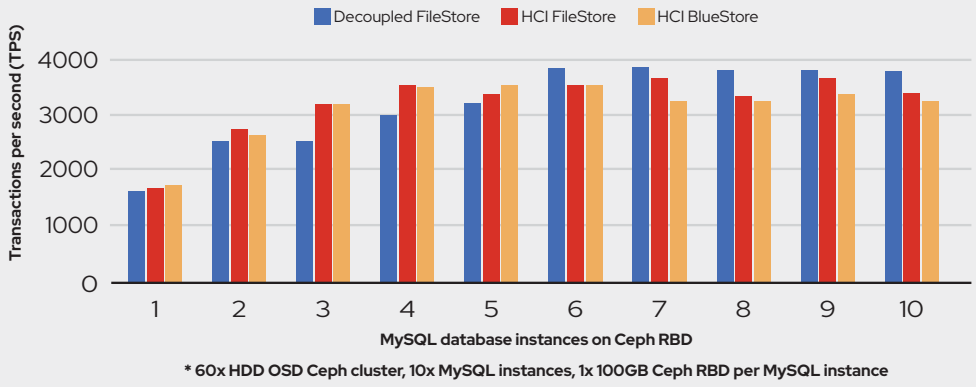

Figure 10. 100% read transactions per second across three cluster types

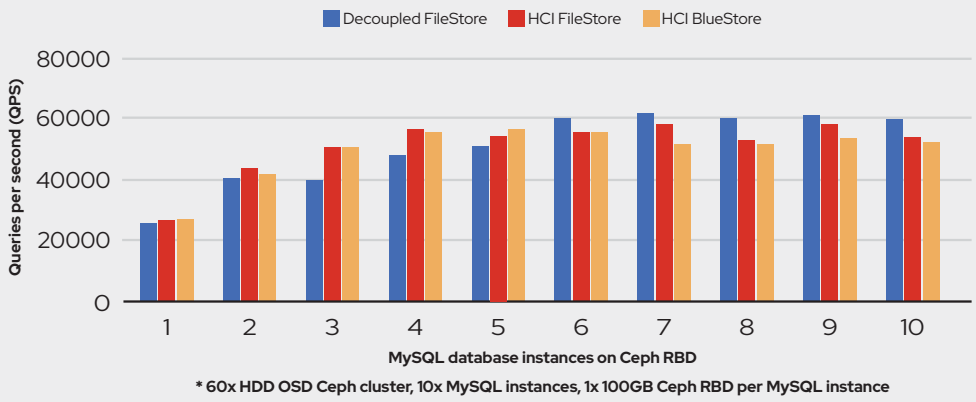

Figure 11. 100% read queries per second across three cluster types

Figures 12 and 13 show that average latency and 99th percentile tail latencies were similar across all three cluster configurations. This result is in line with expectations, as the different cluster configurations ran on the same hardware but with different software deployments (e.g., decoupled vs. HCI, and Ceph FileStore vs. Ceph BlueStore backends).

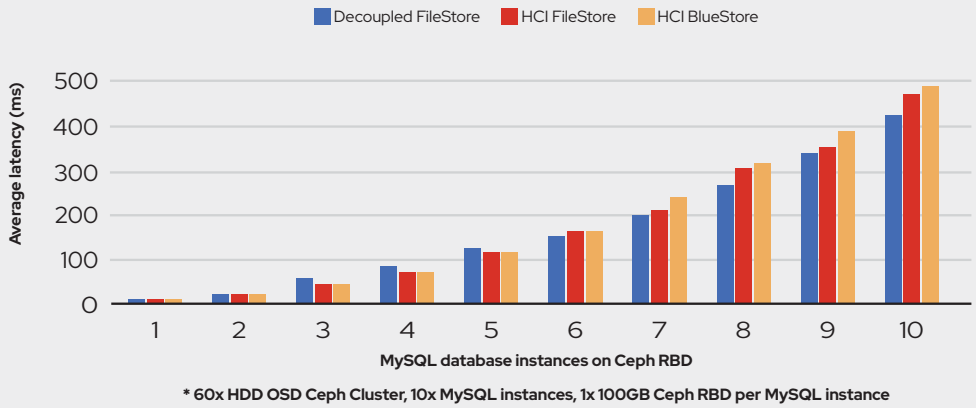

Figure 12. 100% read latency across three cluster types

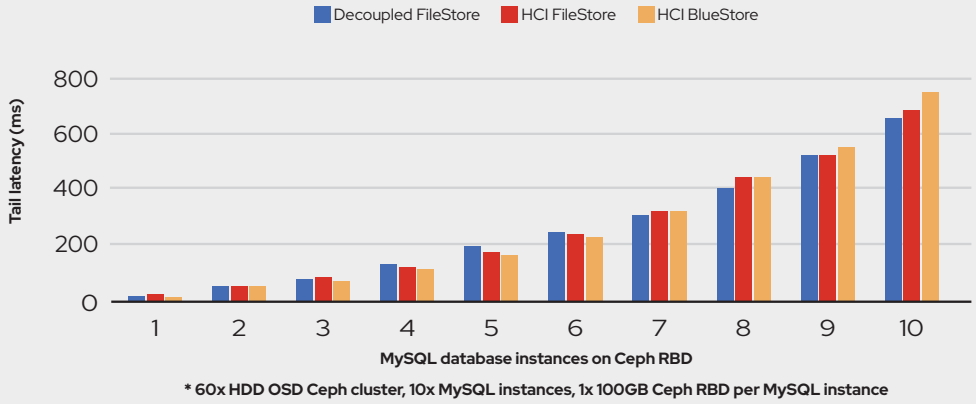

Figure 13. 100% read 99th percentile tail latency across three cluster types

All-flash configuration

As with 100% write workloads, engineers also wanted to gain insights into 100% read workloads on an all-flash cluster. As expected, latencies on spinning media were found to be higher (Figure 14). Ceph BlueStore on an all-flash (NVMe) configuration delivered significantly lower latencies compared to Ceph BlueStore running on a mixed/HDD cluster. Ceph BlueStore on an all-flash cluster delivered ~15x lower average latency and better 99th percentile tail latency.

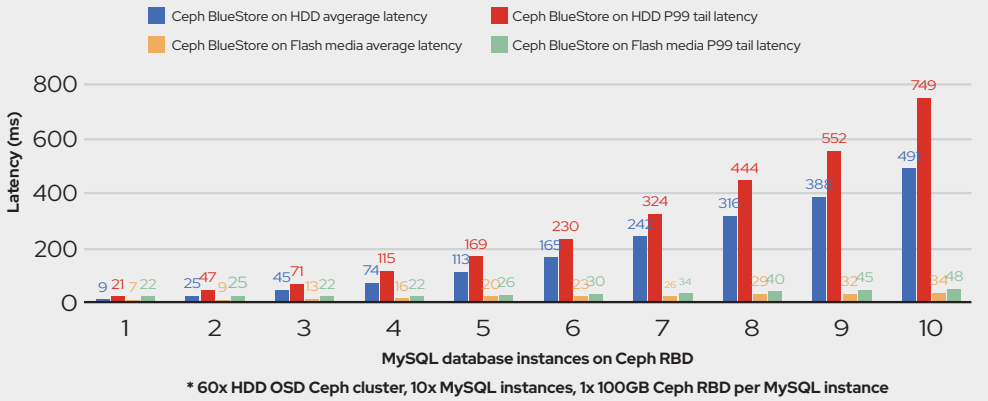

Figure 14. 100% read average latency and 99th percentile tail latency for all-flash media compared to mixed/HDD configurations

100% read performance summary

When the Ceph BlueStore backend was deployed on an all-flash cluster, it delivered up to 15x lower average latency and lower 99th percentile tail latency when compared with the mixed/HDD media cluster. When designing infrastructure for latency-sensitive applications like MySQL, all-flash media should be considered.

MySQL 70%/30% read/write mix performance

Finally, engineers ran a mixture of 70%/30% read/write MySQL database workloads across the same two deployment strategies (decoupled vs. HCI), two Ceph OSD backends (FileStore vs. BlueStore) and two media types (HDD vs. all-flash storage).

HDD-based configurations

Figures 15 and 16 show that the HCI BlueStore configuration delivered slightly better TPS and QPS than the decoupled FileStore and FileStore HCI configurations. Performance of the 70%/30% read/write mix was in line with the previous 100% write and 100% read tests. Again, higher latencies were induced by spinning media.

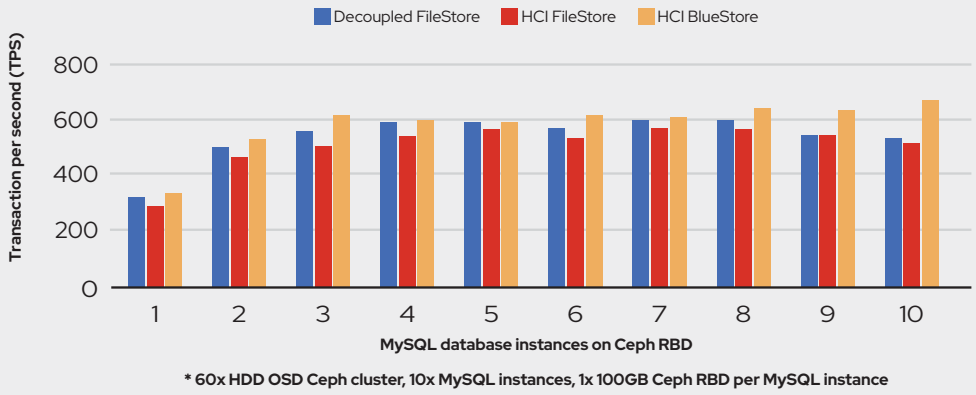

Figure 15. 70%/30% read/write transactions per second across three cluster types (higher is better)

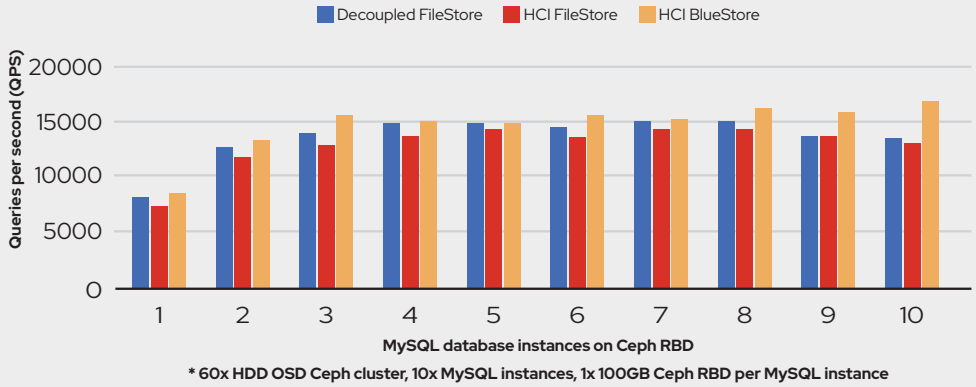

Figure 16. 70%/30% read/write queries per second across three cluster types (higher is better)

As shown in Figure 17, the HCI BlueStore configuration showed slightly better average latency compared to both Ceph FileStore configurations. On the other hand, 99th percentile tail latency on HCI BlueStore was somewhat higher than in Ceph FileStore configurations (Figure 18). Again, spinning media were identified as the primary bottleneck causing higher latencies. Based on testing and observations, Red Hat engineers expect that the configuration could deliver even better latencies without HDDs as the bottleneck.

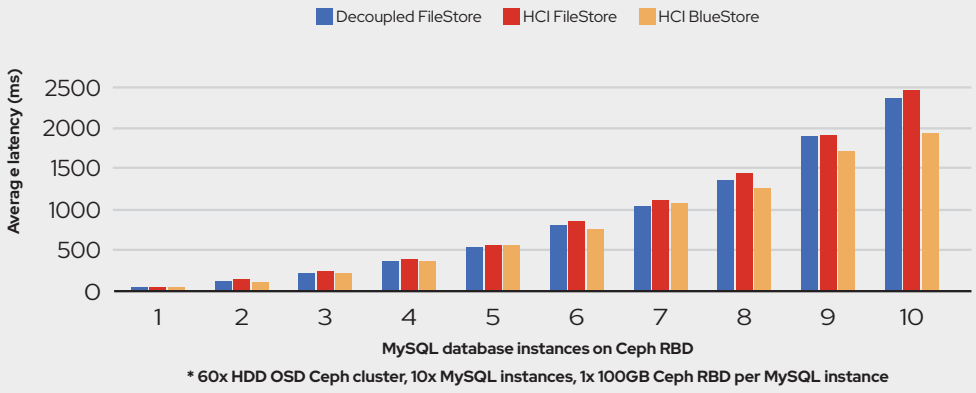

Figure 17. 70%/30% read/write average latency across three cluster types (lower is better)

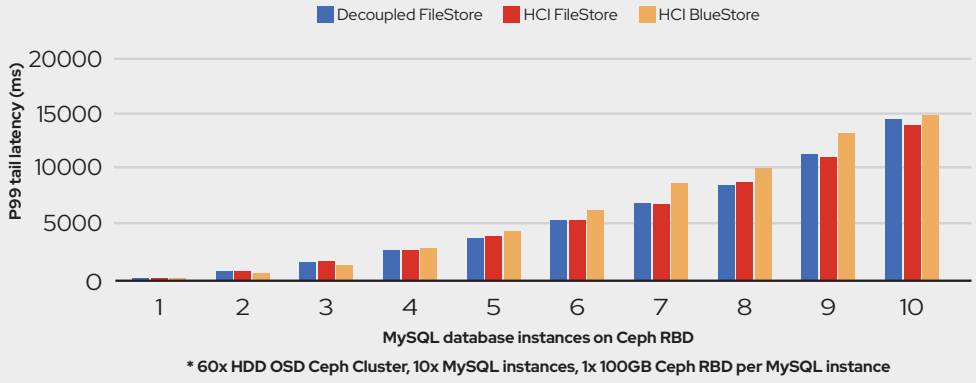

Figure 18. 70%/30% read/write 99th percentile tail latency across three cluster types (lower is better)

All-flash configuration

To isolate the latency of spinning media as the primary bottleneck, engineers ran another test on an all-flash (NVMe) configuration. As with other read and write tests, this configuration delivered performance improvements compared to Ceph BlueStore on mixed/HDDs. As shown in Figures 19 and 20, these tests showed that the Ceph BlueStore backend could deliver lower latencies and both higher TPS and QPS.

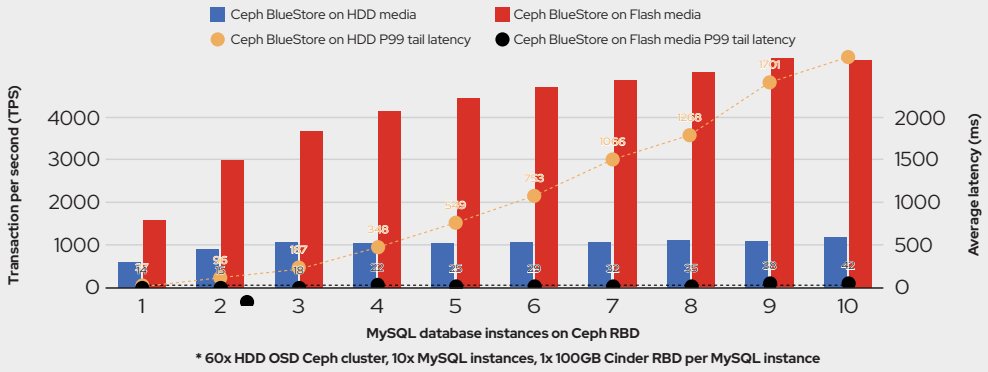

Figure 19. BlueStore and flash media offered higher transactions per second and lower average latency than a mixed/HDD configuration.

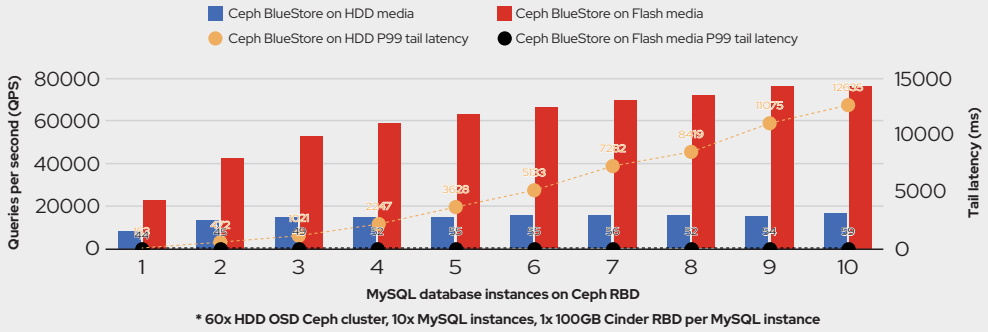

Figure 20. BlueStore and flash media provided higher queries per second and lower 99th percentile tail latency than a mixed/HDD configuration.

Figure 21 compares average and 99th percentile latencies between the HDD-based cluster and the all-flash storage cluster. Ceph BlueStore on an all-flash cluster delivered significantly lower latencies compared to Ceph BlueStore on HDDs. For 70%/30% read/write database workloads, Ceph BlueStore on an all-flash cluster delivered up to 46x lower average latency, and up to 216x lower 99th percentile tail latency for a MySQL database workload.

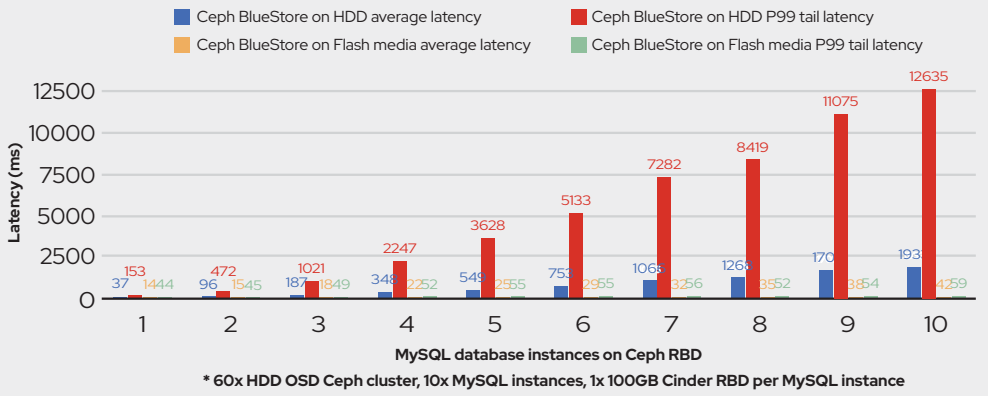

Figure 21. 70%/30% read/write average latency and 99th percentile tail latency media to mixed/HDD media (lower is better)

70%/30% read/write mix performance summary

Compared to Ceph BlueStore on mixed/HDD media, Ceph BlueStore on all-flash media delivered up to:

- 5x higher TPS and QPS
- ~46x lower average latency
- ~216x lower 99th percentile tail latency

Conclusion

Critical OpenStack applications based on MySQL need high-performance, low-latency storage. A hyperconverged approach that combines Red Hat OpenStack Platform and Red Hat Ceph Storage on a single platform can work well. The Ceph BlueStore OSD backend adds considerable performance over the traditional Ceph FileStore backend, especially when coupled with all-flash storage. When using all-flash storage, this approach can increase performance while dramatically reducing latency.

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

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
#F24578_0720