REFERENCE ARCHITECTURE

DEPLOYING MYSQL DATABASES ON RED HAT CEPH STORAGE

Combining Percona MySQL Server, Supermicro storage servers, and Ceph storage for IOPS-intensive workloads

ABSTRACT

As the sheer number of deployed MySQL databases has grown, database administrators (DBAs) are increasingly seeking public or private cloud-based storage solutions that complement their successful non-cloud deployments. Identifying capable, reliable, and flexible cloud storage that can provide the required performance and latency is essential for these efforts. As the most popular OpenStack storage solution[1], Ceph can provide resilient elastic storage pools that have similar operational approaches with the public cloud. To evaluate performance and suitability, the high-performance Percona MySQL Server was tested with Red Hat Ceph Storage on Ceph-optimized storage server configurations from Supermicro.

TABLE OF CONTENTS

Coupled with flash media and the high-performance Percona MySQL Server, Red Hat Ceph Storage provides a high-performance solution for IOPS-intensive MySQL workloads.

Based on extensive Red Hat and Percona testing, the solution demonstrates performance and operational fidelity comparable with public cloud offerings, while offering a cost-effective storage solution for MySQL deployments on Red Hat Ceph Storage clouds.

Supermicro offers a range of Ceph-optimized platforms that can accelerate throughput-optimized, cost/capacty-optimized, and IOPS-optimized Ceph workloads..

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

redhat.com

[1] *Ceph is and has been the leading storage for OpenStack according to several semi-annual OpenStack user surveys.*

INTRODUCTION

Both legacy and cloud native applications are benefiting from design patterns that emerged from the public cloud. Diverse organizations are now looking to model the very successful public cloud Database-as-a-Service (DaaS) experiences with private or hybrid clouds — using their own hardware and resources. The combination of OpenStack and Red Hat Ceph Storage can provide an environment that is familiar to anyone who has built applications atop services like Amazon Elastic Compute Cloud and Elastic Block Store (EBS) or Google Compute Engine and Persistent Disk. Providing a high level of fidelity with public cloud gives developers and operators alike a familiar environment in which to work — increasing application agility all while offering better price performance characteristics.

OpenStack cloud deployments are growing, and organizations are increasingly selecting both MySQL databases and Ceph storage in their OpenStack environments. The sheer number of databases is growing dramatically as well, with many database administrators (DBAs) reporting that they manage hundreds to thousands of separate databases. While virtualized server environments have helped to enable this growth, inflexible siloed storage infrastructure has made limited progress. In fact, rigid and monolithic storage infrastructure is often an impediment to either effective database scalability or effective management of multi-database environments.

As an open software-defined storage platform, Red Hat Ceph Storage provides a compelling solution. Ceph storage clusters can serve diverse types of workloads with carefully chosen and configured hardware. Though throughput-intensive and cost/capacity-focused workloads are common, using Ceph to serve IOPS-intensive MySQL workloads is is increasingly seen in production environments. Cloud-like MySQL storage solutions are made possible by the availability of flash storage media, high-performance MySQL implementations from Percona, and Ceph-optimized hardware platforms from companies like Supermicro.

Any storage solution for MySQL must provide sufficient low latency IOPS throughput to support the needs of key databases and applications, at a cost point that is comparable to public cloud infrastructure. Working closely with Percona and Supermicro, Red Hat has conducted extensive evaluation and testing of MySQL workloads on Red Hat Ceph Storage clusters. Results clearly demonstrate that Percona Server for MySQL running on Red Hat Ceph Storage and Supermicro servers compares favorably with common public cloud solutions in terms of both cost and performance. Specifically, organizations that have chosen a private or hybrid cloud model can provide SSD-backed Ceph RADOS Block Device (RBD) storage at a price point that is even more favorable than public cloud offerings while retaining essential performance characteristics.

DEPLOYING MYSQL ON CEPH CLUSTERS

Modern DBAs have many deployment options for MySQL and similar database technologies. To evaluate Ceph storage solutions, it is important to understand both the technical synergies and the trends that are shaping MySQL and OpenStack environments.

CEPH AND MYSQL: A COMPELLING TECHNOLOGY COMBINATION

Many organizations have become comfortable with deploying MySQL for their applications in the public cloud. As a result, storage technology for hybrid or private cloud MySQL deployments should emulate public cloud methods. Ceph and MySQL represent highly complementary technologies, providing:

- **Strong synergies**. MySQL, OpenStack, and Ceph are often chosen to work together. Ceph is the leading open source software-defined storage solution. MySQL is the leading open source relational database management system (RDBMS).[1] Moreover, Ceph is the number-one block storage for OpenStack clouds[2], with MySQL-based applications figuring as the number-four OpenStack workload.
- **Operational efficiencies**. Ceph storage contributes directly to operational efficiencies for MySQL databases. Ceph provides a shared, elastic storage pool with flexible volume resizing and dynamic database placement. Live instance migration is supported. With Ceph, operators can back up to an object pool and read replicas are easily created via copy-on-write snapshots.
- **Public cloud fidelity**. Developers want platform consistency, and effective private or hybrid clouds require familiar patterns to those established by existing public clouds. Ceph provides block and object storage like public cloud solutions with consistent storage features (Table 1), while letting organizations use their own hardware, data centers, and staff.

TABLE 1. COMPARING CEPH WITH PUBLIC CLOUD STORAGE SOLUTIONS.

FEATURE	CEPH (RBD)	GOOGLE PERSISTENT DISK	AMAZON ELASTIC BLOCK STORE (EBS)
FAULT TOLERANCE	Yes	Yes	Yes
SNAPSHOTS	Yes	Yes	Yes
VOLUME RESIZING	Live	Live	Detached
VOLUME MIGRATION	Yes	Yes	Yes
VOLUME LIVE MIGRATION	Yes	Yes	No
ZONE MIGRATION	Detached	Live	Detached
READ CHECKSUMS	No	Yes	No

OPENSTACK DATABASE AND STORAGE TRENDS

OpenStack has been adopted by organizations of all shapes and sizes to provide cloud infrastructure services for internal developers and operators. Results of the OpenStack user survey are published at their bi-annual OpenStack Summit. Starting with the November 2014 survey and continuing in the most recent survey, Ceph RBD has dominated the polls as the top Cinder driver for providing block storage services in both production and development environments.

[1] *gartner.com/doc/3033819/state-opensource-rdbmss*

[2] *Ceph is and has been the leading storage for OpenStack according to several semi-annual OpenStack user surveys.*

Figure 1 contrasts the number of survey respondents choosing Ceph RBD in production environments, compared to the second most common Cinder driver, Logical Volume Management (LVM). This trend is particularly impressive given that LVM is the default reference driver. At the April 2016 OpenStack Summit, the Cinder development team announced that Ceph RBD would be elevated to the status of reference driver, alongside LVM.

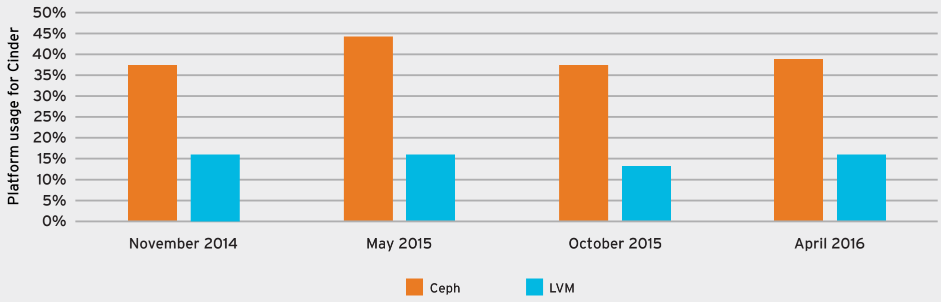

Figure 1. Ceph continues to dominate polls as the number one Cinder driver for production block storage services.

The OpenStack User Survey also collected information about which application stacks were the most popular in production OpenStack environments. Figure 2 shows the percentage of respondents that reported using a particular application framework in their OpenStack environment for the October 2015 and April 2016 surveys. The results illustrate the popularity of MySQL for both custom and off-the-shelf applications running on OpenStack clouds.

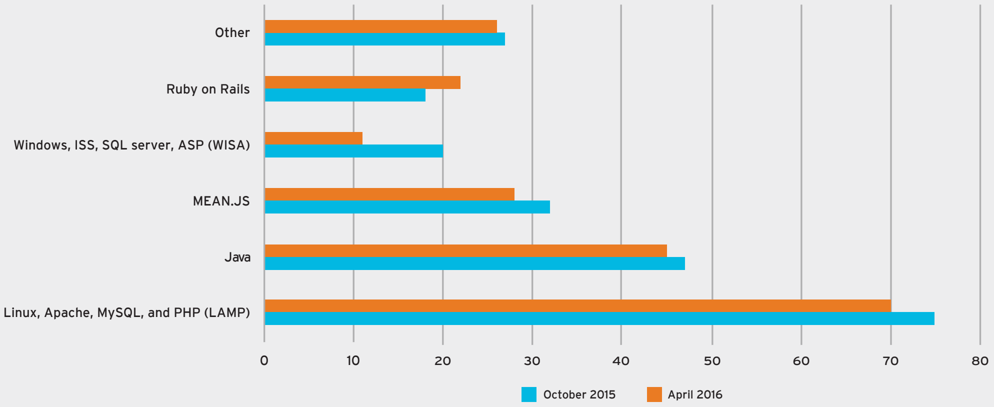

Figure 2. OpenStack application framework trends.

Recent surveys have asked respondents which database they use to underpin the OpenStack control plane. MySQL derived databases dominate the standings (Figure 3). The results also show an increased adoption of Galera clusters with either MySQL or MariaDB. This is likely a sign that OpenStack environments are adopting business-critical applications, making a highly available control plane increasingly attractive. The chart below depicts OpenStack database trends.

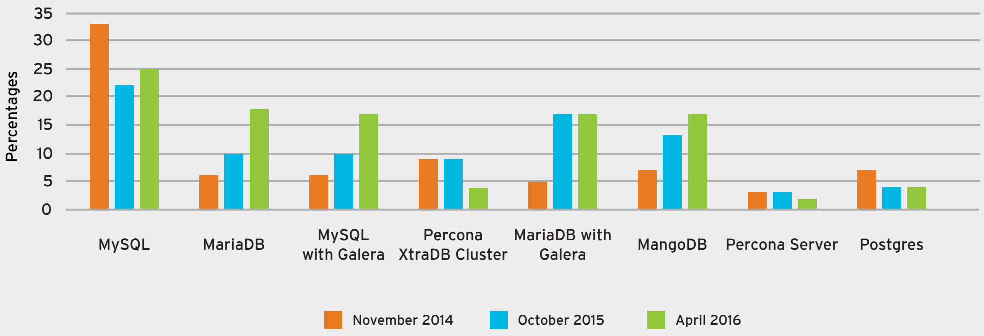

Figure 3. MySQL continues to lead Openstack database trends.

The results of these surveys echo conversations with Red Hat customers. The common thread is that application migration to OpenStack environments is accelerating, and that databases are a critical component of many of these applications. As more applications are being migrated, the demand for IOPS-optimized storage volumes is growing. Simply, our customers have been asking how to deliver IOPS optimized storage volumes using their favorite open source software defined storage technology – Ceph.

SUPPLEMENTING OTHER ARCHITECTURAL CHOICES

DBAs have many architectural options available to them when designing database services to support an organization's applications. A number of objectives can influence the deployment architecture, including database size, schema design, table size, number of databases, availability, write and read scalability, and consistency model. Table 2 and the sections that follow provide a summary of relevant architectural choices and objectives, along with implications for Ceph storage.

TABLE 2. FACTORS DRIVING MYSQL ARCHITECTURE DECISIONS.

FACTORS	MYSQL ON CEPH RBD	MYSQL CLUSTER (NBD)	GALERA CLUSTER	MYSQL REPLICATION	MYSQL ON DRBD	MYSQL SHARDING
DATASET LARGER THAN SINGLE HOST	Yes	Yes	No	No	No	Yes
TABLE LARGER THAN SINGLE HOST	Yes	Yes	No	No	No	No
RELIANCE ON DATASET PARALLELISM?	Yes	No	No	No	No	No
FAULT-TOLERANT STORAGE	Yes	Yes	Yes	Yes	Yes	No
READ SCALING	N nodes	48 nodes	Yes	Yes	No	N nodes
WRITE SCALING	N nodes	24 nodes	1 node	1 node	1 node	N nodes
CONSISTENT WRITES	Synch-ronous	Synch-ronous	Synch-ronous	Asynchronous	Synch-ronous	N/A
MULTI-MASTER?	No	Yes	Yes	No	No	No

MySQL on Ceph RBD

When managing a large collection of MySQL databases across a large number of servers, database operators must balance the load of the collection by consolidating databases onto servers with other databases that each present unique resource constraints. Storage is typically a hard constraint, due to its spatial nature. A target server either does or does not have sufficient remaining space. When databases have a storage footprint approaching the size of a single server, but without commensurate demands on the rest of the server's resources, those other resources become trapped capacity. Lower utilization is the result.

A large deployment of servers represents a significant financial investment, both in capital and operational expenditure. Recapturing trapped capacity is one of the major drivers for implementing distributed storage technologies like Ceph. Ceph lets groups of servers pool their storage resources, and allows any server in the group to access storage over the network. This in turn allows more flexible and efficient placement of databases across the cluster, without requiring tools like rsync or MySQL replication to move data. Ceph also provides a wide range of features including: live volume resize, volume mirroring, volume migration, virtual machine live migration, copy-on-write cloning, and snapshots — both full and differential. Ceph is thus an ideal fit for managing large collections of MySQL database instances.

MySQL Cluster (NDB)

MySQL Cluster is a multi-master MySQL database technology that is shared nothing with no single point of failure. There are three types of nodes/processes in a MySQL Cluster:

- **Management nodes** are responsible for the configuration, monitoring, and starting or stopping the cluster.
- **Application nodes** provide a familiar SQL interface and execute queries on behalf of application.
- **Data nodes** store the actual data of the cluster, and are grouped into pairs that each manage partitions, or shards, and transparently handle replication, self-healing, and load sharing.

MySQL Cluster was originally designed for the telecommunications market to store subscriber data where high write performance and high availability were paramount. MySQL Cluster shards at the application level, by hashing a table's primary key, or hashing a hidden primary key if a primary key is absent from a table. This ability lets a single table grow well beyond the storage capacity of a single host, because a table can be partitioned into up to 192 shards, or eight partitions per node group. The entire MySQL cluster can support up to 24 node groups or 48 data nodes. Since there is no straightforward way to identify which database is consuming CPU resources, MySQL Cluster is relegated to single tenant applications.

Galera Cluster for MySQL

Galera Cluster for MySQL provides nearly synchronous multi-master replication, is highly available, and provides read scalability without slave lag or error-prone master failovers. A Galera Cluster is known as a primary component in Galera parlance. Galera Cluster does not provide its own load balancing component, so it is often deployed in conjunction with load balancing software such as HAproxy to spread load across members of the cluster. All data is near synchronously written to every member of the cluster. Having data on every node means write performance is bound by the round trip time to the farthest node. Galera Cluster's replication strategy constrains the size of a database to the capacity of the smallest node.

Galera Cluster for MySQL can use Ceph block devices to avoid storage constraints and allow rapid reconstitution of a failed member by moving the volume to a new node. Ceph snapshots also allow very fast State Snapshot Transfer (SST) when a new node joins or when a node is found to be corrupted. This ability is particularly helpful in cases where the Galera Cluster spans multiple sites, each with a local Ceph storage pool. Galera Cluster for MySQL comes from a variety of sources, namely Percona XtraDB Cluster and MariaDB Galera Cluster, based on the original library from Codership.

MySQL replication

MySQL asynchronous replication is the traditional way of achieving read scaling. In a traditional replication cluster, each node has a full copy of the dataset. This approach can become an issue when the dataset is larger or when the number of IOPS required by the application is large. Ceph RBD can help ease this burden.

When MySQL replication is used in conjunction with Ceph RBD volumes instead, only the leader would have a full copy of the dataset, and the followers would operate on a thin-provisioned clone of the leader's dataset. Furthermore, while the leader's dataset would be fully replicated to all storage nodes, the follower clones would have only one copy. A leader failover simply moves the leader server and its associated Ceph block device to another node. Since the follower clones are copy-on-write and ephemeral, disk capacity can be saved by reprovisioning followers at a semi-regular interval.

Distributed Replicated Block Device (DRBD)-based shared storage cluster

The network block device replication tool, distributed replicated block device (DRBD), is still a popular MySQL high-availability solution. In such a cluster, the primary node mounts the DRBD device and runs MySQL. All the disk updates are copied by DRBD to the second member of the cluster – the secondary – which is in standby with no service running. The disk read operations on the primary are only served by the primary local storage. In case of a failover event, the secondary can be promoted to primary by mounting the DRBD device and starting MySQL. Operators of DRDB configurations need to decide between failing safe – loss of availability – or risk running with two masters – loss of integrity.

MySQL sharding

Sharding is a way of dividing a dataset into more manageable pieces. The simplest form of sharding is to divide a dataset by tables and put collections of tables or single tables on their own distinct nodes. With a specially designed schema, and application level intelligence, tables can also be partitioned across nodes. Sharding can be difficult as each shard can grow at different rates, or present different demands on node resources. Typically, sharding is used in conjunction with other techniques. Using Ceph RBD with MySQL sharding can help avoid these issues by providing pooled storage resources that can grow as needed – both in terms of size and IOPS.

CEPH ARCHITECTURE OVERVIEW

A Ceph storage cluster is built from large numbers of Ceph nodes for scalability, fault tolerance, and performance. Each node is based on commodity server hardware and uses intelligent Ceph daemons that communicate with each other to:

- Store and retrieve data.
- Replicate data.
- Monitor and report on cluster health.
- Redistribute data dynamically upon cluster expansion or hardware failure (remap and backfill).
- Ensure data integrity (scrubbing).
- Detect and recover from faults and failures.

To the Ceph client interface that reads and writes data, a Ceph storage cluster looks like a simple pool where data is stored. However, the storage cluster performs many complex operations in a manner that is completely transparent to the client interface. Ceph clients and Ceph object storage daemons (Ceph OSD daemons, or OSDs) both use the CRUSH (controlled replication under scalable hashing) algorithm for storage and retrieval of objects.

When a Ceph client reads or writes data – referred to as an I/O context – it connects to a logical storage pool in the Ceph cluster. Figure 4 illustrates the overall Ceph architecture, with concepts that are described in the sections that follow.

- **Pools**. A Ceph storage cluster stores data objects in logical dynamic partitions called pools. Pools can be created for particular data types, such as for block devices, object gateways, or simply to separate user groups. The Ceph pool configuration dictates the number of object replicas and the number of placement groups (PGs) in the pool. For data protection, Ceph storage pools can be either replicated or erasure coded, as appropriate for the application and cost model. Additionally, pools can take root at any position in the CRUSH hierarchy, allowing placement on groups of servers with differing performance characteristics to optimize storage for different workloads.

- **Placement groups.** Ceph maps objects to PGs. PGs are shards or fragments of a logical object pool that are composed of a group of Ceph OSD daemons that are in a peering relationship. Peer OSDs each receive an object replica (or erasure-coded chunk) upon a write. Fault-domain policies within the CRUSH ruleset can force OSD peers to be selected on different servers, racks, or rows. PGs provide a means of creating replication or erasure coding groups of coarser granularity than on a per-object basis. A larger number of placement groups (e.g., 200 per OSD or more) leads to better balancing.
- **CRUSH ruleset.** The CRUSH algorithm provides controlled, scalable, and declustered placement of replicated or erasure-coded data within Ceph and determines how to store and retrieve data by computing data storage locations. CRUSH empowers Ceph clients to communicate with OSDs directly, rather than through a centralized server or broker. By determining a method of storing and retrieving data by algorithm, Ceph avoids a single point of failure, a performance bottleneck, and a physical limit to scalability.

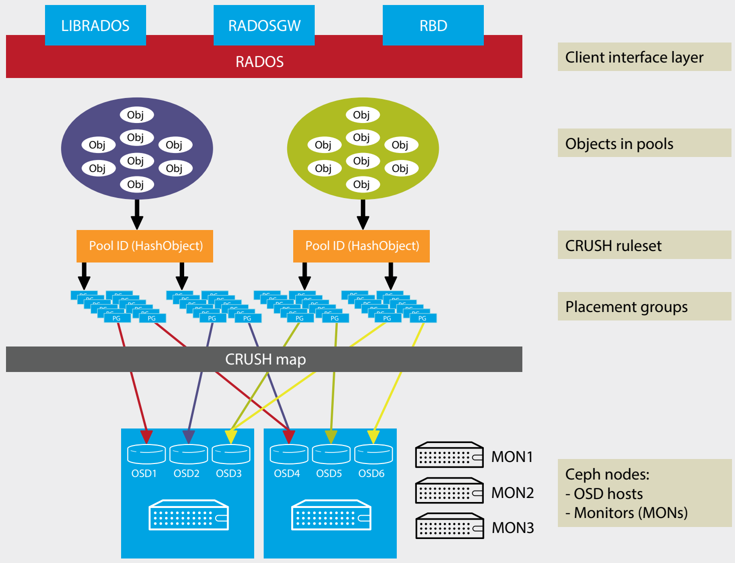

Figure 4. Clients write to Ceph storage pools while the CRUSH ruleset determines how placement groups are distributed across object storage daemons (OSDs).

- **Ceph monitors (MONs).** Before Ceph clients can read or write data, they must contact a Ceph MON to obtain the current cluster map. A Ceph storage cluster can operate with a single monitor, but this introduces a single point of failure. For added reliability and fault tolerance, Ceph supports an odd number of monitors in a quorum – typically three or five for small to mid-sized clusters. Consensus among various monitor instances ensures consistent knowledge about the state of the cluster.

- **Ceph OSD daemons**. In a Ceph cluster, Ceph OSD daemons (Ceph OSDs, or simply OSDs) store data and handle data replication, recovery, backfilling, and rebalancing. They also provide some cluster state information to Ceph monitors by checking other Ceph OSD daemons with a heartbeat mechanism. A Ceph storage cluster configured to keep three replicas of every object requires a minimum of three Ceph OSD daemons, two of which need to be operational to successfully process write requests. Ceph OSD daemons roughly correspond to a file system on a physical hard disk drive.

REFERENCE ARCHITECTURE ELEMENTS

Red Hat's tested solution architecture included Red Hat Ceph Storage and Percona Server for MySQL paired with IOPS-optimized Supermicro storage servers. MySQL instances were running either in Linux containers or in KVM virtual machines (VMs).

RED HAT CEPH STORAGE

Red Hat Ceph Storage significantly lowers the cost of storing enterprise data and helps organizations manage exponential data growth. The software is a robust, petabyte-scale storage platform for public, private, or hybrid clouds. As a modern storage system for cloud deployments, Red Hat Ceph Storage offers mature interfaces for enterprise block and object storage, making it well-suited for cloud infrastructure workloads such as OpenStack. Delivered in a unified self-healing and self-managing platform with no single point of failure, Red Hat Ceph Storage handles data management so businesses can focus on improving application availability, with properties that include:

- Scaling to exabytes.
- No single point of failure in the cluster.
- Lower capital expenses (CapEx) by running on commodity server hardware.
- Lower operational expenses (OpEx) by self-managing and self-healing.

PERCONA SERVER FOR MYSQL

Percona Server for MySQL is a free, fully compatible, enhanced, open source drop-in replacement for MySQL that provides superior performance, scalability and instrumentation. With over 1,900,000 downloads[3], Percona Server's self-tuning algorithms and support for extremely high-performance hardware delivers high performance and reliability. Percona Server is optimized for cloud computing, NoSQL access, and modern hardware — such as solid-state drive (SSD) and flash storage. It delays or completely avoided sharding and offers faster and more consistently-run queries, improved efficiency through server consolidation, and better return on investment (ROI) through lower hosting fees and power usage.

In addition, Percona Server for MySQL offers:

- Cloud readiness by dramatically reducing downtime on servers with slow disks and large memory.
- Software-as-a-Service (SaaS) deployability by increasing flexibility for architectures such as co-located databases with hundreds of thousands of tables and heterogeneous backup and retention policies.
- Vertical scalability and server consolidation, saling to over 48 CPU cores, with the ability to achieve hundreds of thousands of I/O operations per second on high-end solid-state hardware.

[3] *Percona estimate.*

- Query-, object-, and user-level instrumentation for detailed query logging with per-query statistics about locking, I/O, and query plan, as well as performance and access counters per-table, per-index, per-user, and per-host.
- Enterprise readiness with advanced, fully-enabled external authentication, audit logging, and threadpool scalability features that are typically only available in Oracle's commercial MySQL Enterprise Edition.

SUPERMICRO STORAGE SERVERS

With the advancement of large scale cloud computing platforms like OpenStack, the business of data storage has been forever changed. Scale-out software defined storage solutions are rapidly becoming pervasive as the implementation model of choice for large scale deployments. Agile business models require equally agile platforms to build upon. By replacing the proprietary data silos of the past with flexible server hardware, cloud providers have more control over their environments, resulting in better value for their customers. Supermicro has embraced this change, and offers a wide selection of server hardware, optimized for different workloads.

Supermicro storage servers employ advanced components available in workload-optimized form factors. These solutions offer high storage density coupled with up to 96% power efficiency, as well as advantages in procurement and operational costs for deployments of all sizes. Supermicro storage servers optimized for Red Hat Ceph Storage infrastructure feature the latest Intel Xeon CPUs, and offer:

- **Role-specific cluster configurations.** Supermicro offers turn-key cluster configurations with performance, capacity, and density to fit popular application workloads. Memory and networking options are easily customized to meet specific requirements.
- **Optimized network configurations.** Cluster- and rack-level integration offers streamlined deployment of Red Hat Ceph Storage and infrastructure with consistency not attainable using improvised methods.
- **Storage-to-media ratios to fit user applications.** Specific combinations of flash and rotating magnetic media let Supermicro provide diverse solutions that meet workload-tuned performance and density targets.

Supermicro offers a range of storage servers, optimized for different types of workloads.[4] For IOPS-intensive MySQL database workloads, Supermicro offers the SYS-5038MR-OSD006P (Figure 5), a three rack-unit (3U) system comprised of eight individual OSD nodes and 3.2TB of usable capacity with 2x data replication. Each OSD node is comprised of:

- **CPU:** A single Intel Xeon Processor E5-2650 v4
- **Memory:** 32GB
- **OSD storage:** A single Intel SSD Data Center (DC) P3700 800GB NVM Express (NVMe)
- **Boot device:** Mirrored hot-swap Intel SSD DC S3510 80GB, SATA 6Gb/s MLC 2.5-inch
- **Networking:** Single-port 10Gb Ethernet (GbE) SFP+ port

[4] *supermicro.com/solutions/storage_ceph.cfm.*

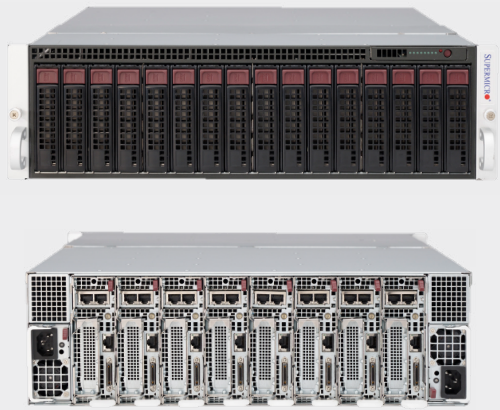

Figure 5. Supermicro SYS-5038MR-OSD006P offers a Ceph-optimized platform with eight IOPS-focused nodes in only three rack units.

PERSISTENT STORAGE FOR CONTAINERS

Containers are revolutionizing the ways that organizations develop, test, and deploy applications — with the potential to affect almost every process and person within the data center. Containers require less overhead than virtualized environments, and instantiate quickly, simplifying deployment and maintenance of applications by bundling the application and its required libraries into a single entity. However, while run-time containers are intended to be disposable, their data is definitely not. Despite their light weight nature, containers still require reliable and available storage so that data is persistent in the event of failed containers, failed disks, or crashed servers.

Red Hat is building an extensive storage ecosystem around containers to bring stability, security, and simplicity to this critical area. Rather than expecting organizations to cobble together container environments — or hire significant container expertise — Red Hat's full technology stack approach provides an end-to-end containerized ecosystem. With this approach, containerized applications get access to the highly available persistent block, file, or object storage that they need without compromise.

With its inherent hardware independence, Red Hat Ceph Storage is specially designed to address enterprise storage challenges. Software-defined scale-out storage is uniquely capable of being managed under a single control plane — a key benefit of containers and a challenge for traditional storage technology.

ARCHITECTURAL DESIGN CONSIDERATIONS: CONFIGURING CEPH FOR IOPS-INTENSIVE WORKLOADS

One of the key benefits of Ceph is its ability to serve storage pools to workloads with diverse needs, depending on chosen underlying hardware configurations. Historically, Ceph has performed very well with high-throughput workloads and has been widely deployed for these use cases. These scenarios are frequently characterized by large-block, asynchronous, sequential I/O (e.g., digital media performance nodes). In contrast, high-IOPS workloads are frequently characterized by small-block synchronous random I/O (e.g., 8/16KB random I/O). Moreover, when Ceph is deployed as Cinder block storage for OpenStack VM instances, it typically serves a mix of IOPs- and throughput-intensive I/O patterns. Table 3 contrasts the characteristics of traditional Ceph workloads with the needs of MySQL workloads.

TABLE 3. ARCHITECTURAL CONSIDERATIONS FOR CEPH CLUSTERS.

	TRADITIONAL CEPH WORKLOAD	MYSQL CEPH WORKLOAD
Performance focus	MB/second	IOPS
Cost focus	Cost/GB	Cost/IOPS
Capacity	Petabytes	Terabytes
Data	Unstructured data	Structured data
Per-server capacity	50-300TB per server	Less than 10TB per server
Media type	Magnetic media (hard disk drives, HDDS)	Flash (SSDs moving to NVMe)
CPU core to OSD ratio	Low CPU-core to OSD ratio	High CPU-core to OSD ratio
Networking	10GbE moving to 25 or 40 GbE	10 GbE

Configuring Ceph for IOPS-intensive, software-defined storage requires a careful balance of computing, storage, and network resources. Supporting performance-intensive workloads in the public cloud was a challenge before the advent of SSD-based block storage. The same is true in a private cloud based on Ceph storage. Providing high performance SSD-based block storage is critical for private clouds that support applications that place a heavy demand on their database.

PUBLIC CLOUD PERFORMANCE

To succeed as a cloud storage solution for MySQL, Ceph should provide performance parity with popular public cloud storage options. Table 4 shows advertised IOPS/GB targets for public cloud solutions from Google, Amazon, and Microsoft. Faster SSD-based options such as Google Persistent Disk or Amazon Provisioned IOPS are typically recommended for hosting IOPS-intensive workloads such as MySQL-based applications.

TABLE 4. ADVERTISED PERFORMANCE LEVELS FOR PUBLIC CLOUD SSD-BASED BLOCK STORAGE.

PUBLIC CLOUD OPTION	PERFORMANCE LEVEL*
Google Persistent Disk (SSD)	30 IOPS/GB
Amazon EBS Provisioned IOPS	30 IOPS/GB in EC2 3-10 IOPS/GB in RDS
Amazon EBS General Purpose SSD	3 IOPS/GB
Microsoft Azure Premium Storage Disk	4-5 IOPS/GB

* At time of writing

To validate these numbers, Red Hat measured the performance of the General Purpose SSD and Provisioned IOPS classes of Amazon Elastic Block Store (EBS). The sysbench testing methodology used on the EBS instances was identical to the methodology described later in this document for evaluating Red Hat Ceph Storage performance. As shown in Figure 6, achieved results were in line with expectations set by product literature.

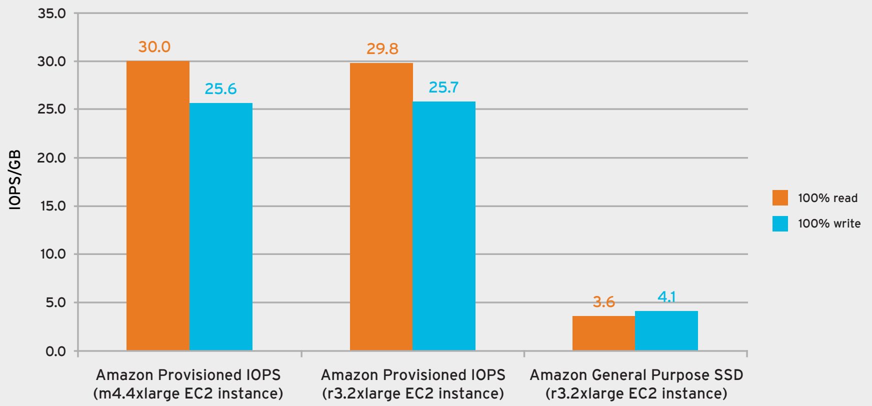

Figure 6. Red Hat testing confirmed Amazon's performance targets for both Provisioned IOPS and General Purpose SSD storage options for their public cloud.

MATCHING PERFORMANCE WITH RED HAT CEPH STORAGE

Red Hat testing revealed that it is possible to match, and in some cases surpass, the performance of the Provisioned IOPS class of Amazon EBS (Figure 7), dependent on provisioned capacity. It is important to note that AWS EBS throttles I/O to provide deterministic IOPS/GB performance in a multitenant environment. Test results infer that throttles reduce IOPS in 100% read and mixed read/write workloads to the same IOPS/GB level as unthrottled, 100% write workloads. As a result, any read/write mix – including the most adversarial 100% write pattern – can meet IOPS/GB targets. However, the Ceph clients in this study are not throttling I/O. For this reason, unthrottled Ceph read IOPS/GB far exceed throttled AWS read IOPS/GB. Therefore, the most pertinent comparison is between write IOPS/GB. Finally, in the Ceph tests, clients were configured to match advertised hardware characteristics of corresponding AWS EC2 instances.

At 14% provisioned capacity, Ceph RBD is well above the performance target. At 87% capacity, however, Ceph RBD is slightly below the performance target of 30 IOPS/GB for writes (Figure 7). A typical Ceph cluster operates at 70% capacity (utilized percentage of usable capacity), which provided a performance level closer to the highest performing public cloud target. This capacity target also gives the cluster headroom for the expected normal distribution of objects across OSDs, and the ability to self-heal after a host failure in a cluster with at least eight nodes.

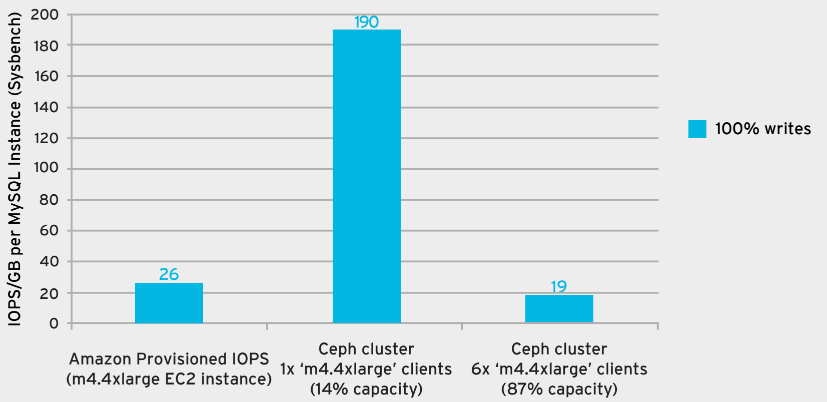

Figure 7. Ceph meets or exceeds AWS Provisioned IOPS performance per MySQL instance, depending on Ceph cluster capacity provisioning level.

FINE-TUNING HARDWARE SELECTION

To achieve the performance levels of the public cloud, storage servers must be carefully balanced in terms of the number and capacity of SSDs, processing power, networking, and memory. For high-IOPS workloads, CPU and media latency are commonly the system bottleneck. To evaluate these parameters, Red Hat varied the number of Intel Xeon 2.3 GHz physical cores and Intel DC P3700 800GB NVMe SSDs per server to see how performance levels were affected (Figure 8). Results showed that one NVMe SSD per 10 server cores provided an ideal level of performance. In fact, using too many NVMe SSDs was counterproductive, and brought the IOPS/GB target down a level closer to that of the General Purpose SSD and Amazon RDS.

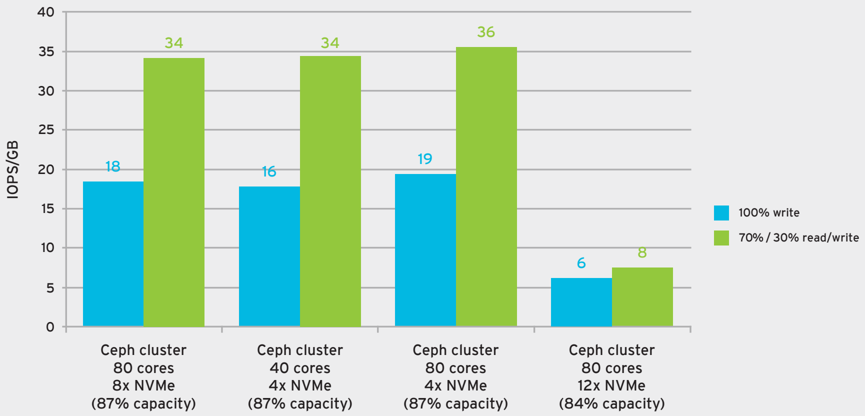

Figure 8. OSD nodes with one NVMe SSD per 10 server cores provided ideal performance characteristics in Red Hat testing.

COMPARING ECONOMIC VIABILITY

High-performance storage systems can be constructed by assembling the newest and most exotic off-the-shelf components that can be sourced. For a fair comparison with the public cloud, however, configurations must be analyzed to ensure that they are also economically comparable. Figure 9 compares the three-year amortized capital expenses (CapEx) of various Supermicro configurations alongside the operational expenses (OpEx) of the Provisioned IOPS class of Amazon EBS.

A more rigorous comparison would factor labor, power, and cooling into the amortized cost of the Supermicro configurations. With this consideration, costs for Red Hat Ceph Storage on Supermicro configurations would need to be three-fold higher to equal the costs of the Amazon Provisioned IOPS configuration. The goal of this analysis is to show that organizations using a private or hybrid cloud model can provide SSD-backed Ceph RBD storage at a price that is comparable, or even more favorable, than public cloud offerings while retaining similar performance characteristics.

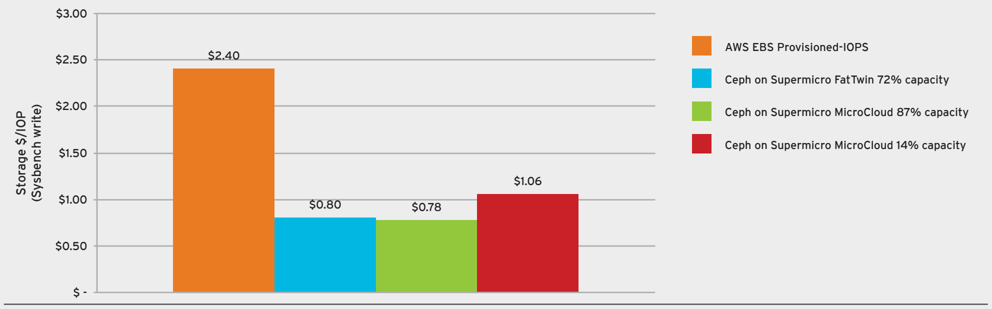

Figure 9. Three-year amortized CapEx of various Red Hat Ceph Storage configurations compares favorably to the OpEx of Amazon Provisioned IOPS. The target is one third cost per IOP, as power, cooling, and administrative costs must be added to Ceph CapEx costs for a complete comparison.

RED HAT TEST ENVIRONMENT

The sections that follow describe the environment used to test an evaluate MySQL running on Red Hat Ceph Storage on Supermicro storage servers.

LAB CONFIGURATION

To evaluate performance, Red Hat built a small Ceph storage cluster comprised entirely of Supermicro hardware (Figure 10). The server configurations used in Red Hat testing were intentionally broader than recommended, providing flexibility for testing a variety of different media and core combinations.

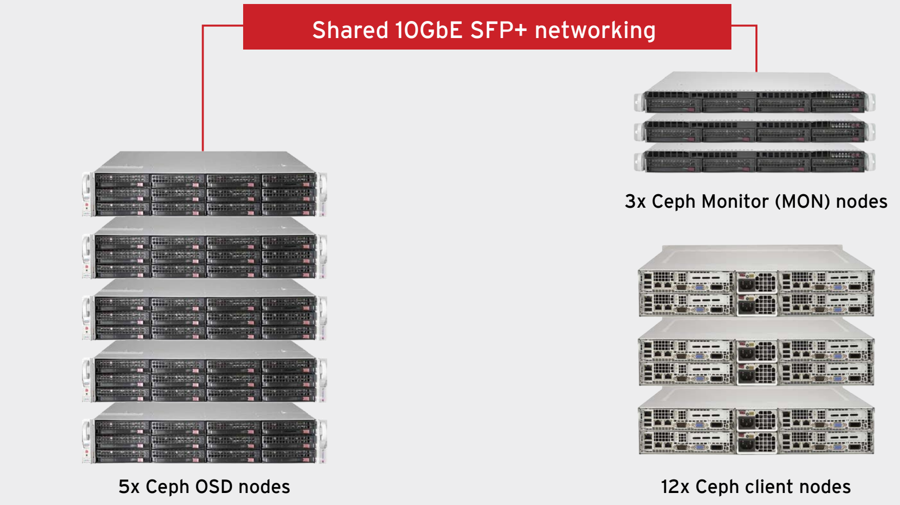

Figure 10. Sysbench was used to exercise 12x Ceph client nodes against 5x Supermicro Ceph OSD nodes.

Five Ceph OSD servers were provided by 2U Supermicro SuperStorage SSG-6028R-E1CF12L as the base servers, each with:

- Dual Intel Xeon Processor E5-2650 v3 (10 core)
- 32GB of 1866 MHz DDR3 ECC SDRAM DIMMs
- 2x 80GB boot drives, Intel SSD DC S3510 80GB, SATA 6Gb/s, MLC 2.5-inch
- 4x 800GB Intel DC P3700 NVMe SSDs
- Dual-port 10GbE network adapter (AOC-STGN-i2S)
- 8x Seagate 6TB 7200 RPM SAS (ST600MN00347200)
- Redundant hot-swap power supplies

Three Ceph MON nodes were provided by 1U Supermicro Superstorage SSG-6018R-MON2 servers, each with:

- Dual Intel Xeon Processor E5-2630 v3 (8-Core)
- 64GB memory per node
- 2x 2.5-inch 80GB SSD for boot/operating system
- Internal 800GB NVMe SSD
- Onboard dual-port 10GbE SFP+
- Redundant hot-swap power supplies

Twelve MySQL/Ceph clients were provided within three 2U Supermicro SuperServer Twin2 nodes, each with:

- Four nodes per server
- Dual Intel Xeon Processor E5-2670 v2 per node
- 64GB SDRAM DIMMs per node
- Redundant hot-swap power supplies

Software used in testing was:

- Red Hat Ceph Storage 1.3.2
- Red Hat Enterprise Linux 7.2
- Percona Server 5.7.10 (pre-general availability)

SYSBENCH

The Ceph cluster was exercised by SysBench.[5] SysBench is a modular, cross-platform and multi-threaded benchmark tool with different modes for testing, including:

- File I/O performance.
- Scheduler performance.
- Memory allocation and transfer speed.
- POSIX threads implementation performance.
- Database server performance.

SysBench runs a specified number of threads, all executing requests in parallel. The actual workload produced by requests depends on the specified test mode. Testers can limit the total number of requests, the total time for the benchmark, or both. In testing, Red Hat used sysbench to generate a pattern of MySQL database read, write, and mixed read/write requests.

[5] *github.com/akopytov/sysbench*

TUNING AND BEST PRACTICES

Red Hat, Percona, and Supermicro testing revealed a number of tuning opportunities and best practices for both MySQL and Ceph.

OPTIMIZING MYSQL FOR CEPH

Strategies for optimizing MySQL for other storage are often just as relevant when Ceph RBD is used for storage. Databases should use a journaled filesystem that supports barriers, such as XFS. InnoDB or Percona XtraDB are the engines of choice, and they should be configured to use O_DIRECT to bypass the system page cache. The optimal buffer pool size is a function of the size of the dataset, and its access distribution (e.g., uniform versus pareto). The goal of these configurations is to serve a substantial percentage of read requests directly from memory, and reduce the number of reads required for flushing InnoDB pages.

In addition, a number of measures can contribute to performance, when running MySQL on an elastic software-defined storage cloud such as Ceph, including:

- **Buffer pool**. If there is not enough room in the buffer pool for a minimal database page working-set, performance can suffer. The larger the buffer pool, the more InnoDB acts like an in-memory database by reading data from disk once, then accessing the data from memory during subsequent reads. Testing with a uniformly distributed workload showed an optimal buffer pool size of 20% of the dataset.
- **Flushing**. MySQL can be configured to flush each transaction separately or to flush all transactions occurring within a specified time period (e.g. 1 second). For workloads that can tolerate this batch flushing (e.g. storing social media clicks), performance improves dramatically. The MySQL parameter is `innodb_flush_log_at_trx_commit = 0`.
- **Percona parallelized double write buffer.** The MySQL double write buffer can become a bottleneck, as it has a finite number of slots. Percona has recently provided a modification to Percona Server that allows multiple double write buffers so that no single thread has to wait to flush pages. With this modification, the number of writes can be easily doubled.[6]

OPTIMIZING CEPH

Optimizing Ceph for MySQL workloads is generally fairly similar to any other workload that performs many small random reads and synchronous writes. Hardware selection remains the single most important factor for ensuring higher-performance block storage. This study shows Ceph performing very well with all-flash media driven by an appropriate number of CPU cores.

Additionally, large quantities of small I/O operations places significant strain on the memory allocator used by Ceph (TCMalloc). Red Hat Ceph Storage 1.3.2 or later on Red Hat Enterprise Linux 7.2 or later has an optimized TCMalloc configuration, that can double MySQL on Ceph performance compared to using previous memory allocator settings.

NVMe drives can support very deep queues. To keep those queues full, Red Hat testing partitioned each drive into four journal partitions and four data partitions. This partitioning resulted in four OSDs operating on a single NVMe device, which proved to be optimal for the tested configuration.

[6] *https://www.percona.com/blog/2016/05/09/percona-server-5-7-parallel-doublewrite/*

Starting with the `performance-throughput` tuned profile, a number of parameters were adjusted, mostly in the networking stack.

```
net.ipv4.ip_forward=1
net.core.wmem_max=125829120
net.core.rmem_max=125829120
net.ipv4.tcp_rmem= 10240 87380 125829120
net.ipv4.tcp_wmem= 10240 87380 125829120
net.ipv4.tcp_window_scaling = 1
net.ipv4.tcp_timestamps = 1
net.ipv4.tcp_sack = 1
net.core.netdev_max_backlog = 10000
vm.swappiness=1
```

There are a number of `ceph.conf` parameters that were adjusted to keep the NVMe devices busy. The following were adjusted from the default:

```
# Filestore Tuning
filestore_xattr_use_omap = true
filestore_wbthrottle_enable = false
filestore_queue_max_byes = 1048576000
filestore_queue_committing_max_bytes = 1048576000
filestore_queue_max_ops = 5000
filestore_queue_committing_max_ops = 5000
filestore_max_sync_interval = 10
filestore_fd_cache_size = 64
filestore_fd_cache_shards = 32
filestore_op_threads = 6

# Filesystem Tuning
osd_mount_options_xfs = rw,noatime,inode64,logbsize=256k,delaylog
osd_mkfs_options_xfs = -f -i size=2048
```

```
# Journal Tuning
journal_max_write_entries = 1000
journal_queue_max_ops = 3000
journal_max_write_bytes = 1048576000
journal_queue_max_bytes = 1048576000

# Op tracker
osd_enable_op_tracker = false

# OSD Client
osd_client_message_size_cap = 0
osd_client_message_cap = 0

# Objector
objecter_inflight_ops = 102400
objector_inflight_op_bytes = 1048576000

# Throttles
ms_dispatch_throttle_bytes = 1048576000

# OSD Threads
osd_op_threads = 32
osd_op_num_shards = 5
osd_op_num_threads_per_shard = 2
```

DEPLOYING ON LINUX CONTAINERS VERSUS VIRTUAL MACHINES

While virtualization has become popular, light weight containerized environments demonstrate considerable promise. For example, Red Hat Ceph Storage can effectively provide persistent storage for containerized applications. Red Hat testing sought to quantify performance differences running MySQL in virtualized and containerized environments when using Ceph storage. Although the details of QEMU event loop optimization across multiple VMs are beyond the scope of this reference architecture, an overview comparison follows.

Testing compared performance of MySQL instances running on:

- Bare metal with RBD over kernel krdb.
- Linux containers with RBD over kernel krdb.
- KVM VMs with RBD over QEMU/KVM via librbd with various QEMU I/O settings.

This testing showed that containers can scale RBD performance better, due to data-level parallelization across InnoDB pages using a distinct thread for each operation (Figure 11).

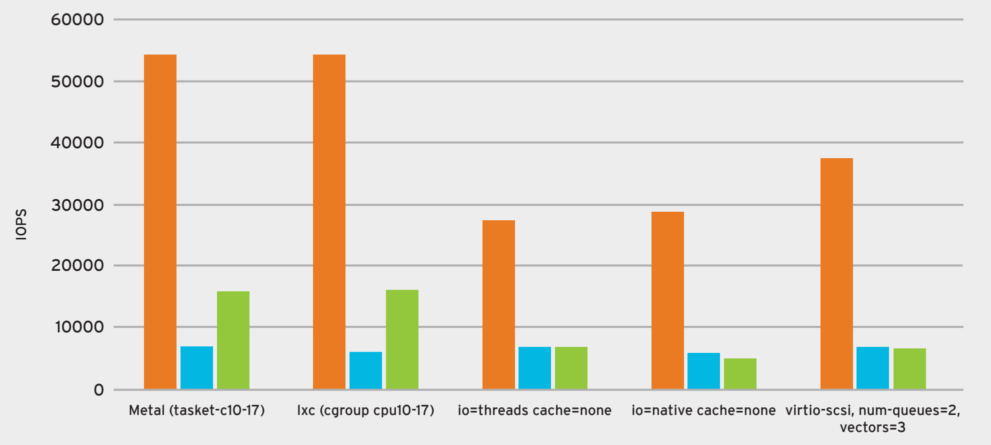

Figure 11. Containerized deployments offer comparable performance to bare-metal environments, and substantially better performance than virtualized environments.

SUMMARY

In addition to storage for traditional throughput-optimized and cost/capacity-optimized workloads, Red Hat Ceph Storage works well as IOPS-intensive cloud storage for MySQL and similar databases. Red Hat testing with Percona MySQL Server and Supermicro storage servers has shown that this flexible software-defined storage platform can provide the necessary performance and latency in a cost-effective solution.

With Ceph-optimized servers from Supermicro and the high-performance Percona MySQL Server implementation, OSDs can be easily configured to exploit the considerable throughput available with modern flash devices. Testing also revealed that Red Hat Ceph Storage support for Linux containers technology offers performance parity with bare-metal servers while delivering persistent storage for convenient, containerized MySQL applications. Together, these complementary technologies let organizations effectively deploy private and hybrid cloud storage for a range of MySQL applications that closely mirrors their successful experiences with the public cloud.

ABOUT RED HAT

Red Hat is the world's leading provider of open source software solutions, using a community-powered approach to provide reliable and high-performing cloud, Linux, middleware, storage, and virtualization technologies. Red Hat also offers award-winning support, training, and consulting services. As a connective hub in a global network of enterprises, partners, and open source communities, Red Hat helps create relevant, innovative technologies that liberate resources for growth and prepare customers for the future of IT.

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

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

redhat.com
#INC0448222_0016