OLTP-Level Performance Using Seagate NVMe SSDs with MySQL and Ceph

Technology Paper

Authored by:
Rick Stehno

Introduction

For quite some time, the Ceph community has wanted to apply Ceph as the back-end storage system for the MySQL database in a manner that ensures optimum performance in database online transactional processing (OLTP). This has been a challenge, however, as Ceph was not initially designed to provide that level of performance for database applications. Ceph has the capability to provide block and file access from applications, but the underlying Ceph cluster is an object store. Ceph also provides data replication between cluster nodes, and to ensure this replication process is successful, it performs the process synchronously—rather than semi-synchronously or asynchronously—which can introduce performance inefficiencies.

This document explores the possibility of using a Ceph cluster to provide OLTP-level performance for database applications. In doing so, it describes a Ceph-cluster implementation that incorporates Seagate NVMe solid state drives (SSDs) in existing stand-alone MySQL databases and compares the performances of all of them. As the paper will show, taking advantage of the Ceph cluster's scalability, the resulting configurations continue to realize the OLTP performance that the MySQL databases were delivering prior to the implementation.

Major Components

This section describes the following major components of the system:

- MySQL relational database
- Ceph software defined storage (SDS)
- Ceph monitors
- Seagate NVMe SSD storage technology
- Block storage

MySQL Relational Database

MySQL is the most popular and widely used open-source database in the world. Feature-rich in performance, scalability and reliability, MySQL also sports a small footprint, which contributes to its success in the embedded database market.

Ceph Software Defined Storage

Ceph is an open source software defined storage (SDS) application designed to provide scalable object, block and file system storage to clients. Ceph introduced new methods for

storing data over the cluster. It is massively scalable and distributed, and it runs on commodity hardware. When configured properly, the Ceph cluster has no single point of failure. The Ceph cluster includes the following components:

- Object Storage Daemon (OSD):

  The OSD stores the data and handles data replication, recovery, and rebalancing. Ceph requires at least two OSDs to achieve a healthy state when replication is enabled, for example when setting replica to 2.

- Controlled Replication Under Scalable Hashing (CRUSH):

  CRUSH describes the storage map of the cluster, including device locations within the cluster and rule sets that determine how Ceph will store the data. The CRUSH map can be configured to tier data, for example, creating a pool of data on NVMe devices for peak performance and/or for creating HDD pools for performance and/or archiving. In the Ceph cluster used for this paper, multiple pools were defined over various hard disk drives (HDDs) and NVMe SSDs, with one pool created using NVMe for the MySQL database server.

- Ceph Monitors:

  The Ceph monitors maintain maps of the cluster's state, the OSD, the Placement Group (PG) and the CRUSH.

- Seagate NVMe SSD Storage Technology:

  Seagate provides the storage components that an enterprise data center requires, including HDDs with a range of different speeds and capacities as well as various SAS, SATA and NVMe SSDs when peak performance is required. The benchmarks for this study used both Seagate HDDs and Seagate NVMe SSDs.

- Block Storage:

  The MySQL database uses block storage as well as other storage-access alternatives. The benchmarks for this study used block storage on both the MySQL host server and in Ceph using RADOS Block Devices (RBD).

Why MySQL with Ceph?

MySQL and Ceph have many traits in common. Both are leaders in their fields of technology, and both are open source, scalable and fault tolerant. To provide storage fault tolerance and enhanced performance, most MySQL stand-alone databases incorporate some sort of storage-based mirroring or striping, mainly through some type of RAID level using external arrays, attached RAID cards or some form of software RAID on the host. In contrast, Ceph provides replication between the nodes or OSDs in the cluster, eliminating the need for any RAID configuration. In Ceph, when using Seagate NVMe SSDs, the pool replica should be set to 2.

Both MySQL and Ceph are also highly configurable and are able to handle any type of workload that the user's application requires. Because Ceph is highly scalable, it means that a Ceph cluster can have hundreds, if not thousands, of nodes to provide the needed storage capacity and fault tolerance that a MySQL database—or even a whole enterprise data center—would need. Adding more storage devices (HDDs or SSDs) to a Ceph node is painless, as is adding a Ceph node to a cluster to increase storage without impacting the end-user. Also, because a Ceph cluster can grow so easily, with each node containing one to many storage devices, you can easily carve out chunks of storage on specific storage devices (HDDs and/or SSDs) to assign them to specific applications and databases. Taking performance and capacity requirements into consideration, for example, it isn't difficult to allocate SSD storage to a MySQL database while allocating HDD storage to an application that performs archiving.

Benchmark Overview

This section includes the following overviews:

- MySQL Database with 18 HDDs
- MySQL Database with 4 NVMe SSDs
- Ceph Cluster with 4 NVMe SSDs
- Sysbench OLTP Benchmark with 100 threads
- RBD Setup on MySQL Database Server

OLTP-Level Performance Using Seagate NVMe SSDs with MySQL and Ceph

SEAGATE

MySQL Database with 18 HDDs

The following table describes the MySQL Database:

Component	Description
OS	CentOS kernel 3.10.0-229.el7.x86_64
RAM	256GB
CPU	Intel(R) Xeon(R) E5-2680 0 @ 2.70GHz (32 cores)
Network	40G Mellanox ethernet
Server Version	MySQL 5.7.15
Storage	18 HDDs, RAID 0 for performance (mdadm --create /dev/md0 --level=0 --raid-devices=18 /dev/ sd[b-s])
InnoDB Buffer Pool	• innodb_buffer_pool_size = 6G • innodb_log_group_home_dir = '/raid10/logs' (local storage) • innodb_log_file_size = 4G • innodb_log_files_in_group = 6
DB File(s) Location	On local disk or mapped to Ceph RBD

MySQL Database with Four NVMe SSDs

This configuration is similar to the configuration above, but it incorporates four NVMe SSDs set up with RAID 0 and using MDADM as the storage. The following table describes the NVMe device tuning component of the MySQL database with NVMe SSDs:

Component	Description

NVMe Device Tuning	Entries in /etc/rc.local:
	`setenforce 0`
	`echo deadline > /sys/block/nvme0n1/queue/scheduler`
	`echo "1" > /sys/block/nvme0n1/queue/rq_affinity`
	`echo "0" > /sys/block/nvme0n1/queue/rotational`
	`echo "0" > /sys/block/nvme0n1/queue/add_random`
	`echo "0" > /sys/block/nvme0n1/queue/nomerges`
	`echo 1 > /sys/block/nvme0n1/queue/iosched/fifo_batch`
	`echo 0 > /sys/block/nvme0n1/queue/iosched/front_merges`
	`echo 5 > /sys/block/nvme0n1/queue/iosched/writes_starved`
	`blockdev --setra 4096 /dev/nvme0n1`
	`echo deadline > /sys/block/nvme1n1/queue/scheduler`
	`echo "1" > /sys/block/nvme1n1/queue/rq_affinity`
	`echo "0" > /sys/block/nvme1n1/queue/rotational`
	`echo "0" > /sys/block/nvme1n1/queue/add_random`
	`echo "0" > /sys/block/nvme1n1/queue/nomerges`
	`echo 1 > /sys/block/nvme1n1/queue/iosched/fifo_batch`
	`echo 0 > /sys/block/nvme1n1/queue/iosched/front_merges`
	`echo 5 > /sys/block/nvme1n1/queue/iosched/writes_starved`
	`blockdev --setra 4096 /dev/nvme1n1`
	`echo deadline > /sys/block/nvme2n1/queue/scheduler`
	`echo "1" > /sys/block/nvme2n1/queue/rq_affinity`
	`echo "0" > /sys/block/nvme2n1/queue/rotational`
	`echo "0" > /sys/block/nvme2n1/queue/add_random`
	`echo "0" > /sys/block/nvme2n1/queue/nomerges`
	`echo 1 > /sys/block/nvme2n1/queue/iosched/fifo_batch`
	`echo 0 > /sys/block/nvme2n1/queue/iosched/front_merges`
	`echo 5 > /sys/block/nvme2n1/queue/iosched/writes_starved`
	`blockdev --setra 4096 /dev/nvme2n1`
	`echo deadline > /sys/block/nvme3n1/queue/scheduler`
	`echo "1" > /sys/block/nvme3n1/queue/rq_affinity`
	`echo "0" > /sys/block/nvme3n1/queue/rotational`
	`echo "0" > /sys/block/nvme3n1/queue/add_random`
	`echo "0" > /sys/block/nvme3n1/queue/nomerges`
	`echo 1 > /sys/block/nvme3n1/queue/iosched/fifo_batch`
	`echo 0 > /sys/block/nvme3n1/queue/iosched/front_merges`
	`echo 5 > /sys/block/nvme3n1/queue/iosched/writes_starved`
	`blockdev --setra 4096 /dev/nvme3n1`

OLTP-Level Performance Using Seagate NVMe SSDs with MySQL and Ceph

SEAGATE

MySQL Database Using Ceph Cluster with Four NVMe SSDs

Component	Description
OS	Ubuntu 16.04 – kernel 4.4.0-36-generic #55-Ubuntu
RAM	256GB
CPU	Intel(R) Xeon(R) E5-2660 v3 @ 2.60GHz (40 cores)
Network	40G Mellanox public and private networks
Ceph Release	Jewel 10.2.2 with Filestore
Storage Pools	Multiple Storage Pools: HDD, SSD, and NVMe SSDs NVMe pool used for MySQL storage: Four Seagate Nytro® XF1440 1.92TB 2.5-inch form factor NVMe SSDs
Replicas	2
OSD Mount Options	osd_mount_options_xfs = inode64,noatime,logbsize=256k
OSD settings	[osd] osd _ enable _ op _ tracker = false osd _ op _ num _ threads _ per _ shard = 2 osd _ op _ num _ shards = 12 filestore _ min _ sync _ interval = 1 filestore _ max _ sync _ interval = 10 filestore _ odsync _ write = true filestore _ max _ inline _ xattr _ size = 254 filestore _ max _ inline _ xattrs = 6 filestore _ queue _ committing _ max _ bytes = 1048576000 filestore _ queue _ committing _ max _ ops = 5000 filestore _ queue _ max _ bytes = 1048576000 filestore _ queue _ max _ ops = 4000 journal _ max _ write _ bytes = 1048576000 journal _ max _ write _ entries = 1000 journal _ queue _ max _ bytes = 1048576000 journal _ queue _ max _ ops = 3000 filestore _ fd _ cache _ shards = 32 filestore _ fd _ cache _ size = 128 ms _ dispatch _ throttle _ bytes = 0 osd _ client _ message _ size _ cap = 0 osd _ client _ message _ cap = 0
Ceph Monitor	Maintains maps of the following statuses: • Cluster state • OSD • Placement group • CRUSH
NVMe Device Tuning	Entries in /etc/rc.local: echo Y > /sys/module/scsi _ mod/parameters/use _ blk _ mq echo op ons ib _ srp ch _ count=$n > /etc/modprobe.d/ib _ srp.conf setenforce 0 echo “1” > /sys/block/nvme0n1/queue/rq _ affinity echo “0” > /sys/block/nvme0n1/queue/rotational echo “0” > /sys/block/nvme0n1/queue/add _ random echo “0” > /sys/block/nvme0n1/queue/nomerges blockdev --setra 4096 /dev/nvme0n1 echo “1” > /sys/block/nvme1n1/queue/rq _ affinity echo “0” > /sys/block/nvme1n1/queue/rotational echo “0” > /sys/block/nvme1n1/queue/add _ random echo “0” > /sys/block/nvme1n1/queue/nomerges blockdev --setra 4096 /dev/nvme1n1 echo “1” > /sys/block/nvme2n1/queue/rq _ affinity echo “0” > /sys/block/nvme2n1/queue/rotational echo “0” > /sys/block/nvme2n1/queue/add _ random echo “0” > /sys/block/nvme2n1/queue/nomerges blockdev --setra 4096 /dev/nvme2n1 echo “1” > /sys/block/nvme3n1/queue/rq _ affinity echo “0” > /sys/block/nvme3n1/queue/rotational echo “0” > /sys/block/nvme3n1/queue/add _ random echo “0” > /sys/block/nvme3n1/queue/nomerges blockdev --setra 4096 /dev/nvme3n1

Sysbench OLTP Benchmark

The Sysbench OLTP Benchmark with 100 threads setup is as follows:

```
sysbench --test=oltp --oltp-table-size=45000000 --mysql-db=test --mysql-user=userid --mysql-password=pwd
--db-driver=mysql --mysql-table-engine=innodb prepare

sysbench --test=oltp --oltp-table-size=45000000 --max-requests=0 --mysql-db=test --mysql-user=userid --mysql-
password=pwd --max-time=600 --oltp-read-only=off --mysql-engine-trx=yes --oltp-test-mode=complex --num-
threads=100 --db-driver=mysql --init-rng=on run
```

RDB Setup

The RDB setup on the MySQL Database Server is as follows:

```
rbd map bdi-180nvme --name client.admin -p nvme -m xxx.xxx.xxx.xxx -k
/etc/ceph/ceph.client.admin.keyring

mkfs.xfs /dev/rbd0
mkdir /rbd0
mount -o noatime,attr2,delaylog,inode64,noquota  /dev/rbd0  /rbd0

df
/dev/rbd0       314419200    /rbd0
```

Benchmark Results

This section provides the results of the following benchmark tests:

- MySQL with Local HDDs
- MySQL with Local NVMe SSDs
- MySQL Using Ceph Cluster Storage with Replica 1
- MySQL Using Ceph Cluster Storage with Replica 2

MySQL with Local HDDs

```
sysbench 0.4.12:  multi-threaded system evaluation benchmark

Running the test with following options:
Number of threads: 100
Initializing random number generator from timer.

Doing OLTP test.
Running mixed OLTP test
Using Special distribution (12 iterations,  1 pct of values are returned in 75 pct cases)
Using "BEGIN" for starting transactions
Using auto_inc on the id column
Threads started!

OLTP test statistics:
transactions:                           1866277 (6220.64 per sec.)
read/write requests:                    35477314 (118252.29 per sec.)
other operations:                       3733686 (12445.05 per sec.)

Test execution summary:
per-request statistics:
avg:                                    16.07ms
```

MySQL with Local NVMe SSD

```
sysbench 0.4.12:  multi-threaded system evaluation benchmark

Running the test with following options:
Number of threads: 100
Initializing random number generator from timer.

Doing OLTP test.
Running mixed OLTP test
Using Special distribution (12 iterations,  1 pct of values are returned in 75 pct cases)
Using "BEGIN" for starting transactions
Using auto_inc on the id column
Threads started!

OLTP test statistics:
    transactions:                        3293190 (10977.13 per sec.)
    read/write requests:                 62594516 (208645.08 per sec.)
    other operations:                    6587881 (21959.26 per sec.)

Test execution summary:
  per-request statistics:
  avg:                                   9.10ms
```

MySQL Using Ceph Cluster Storage with Replica 1

```
sysbench 0.4.12:  multi-threaded system evaluation benchmark

Running the test with following options:
Number of threads: 100
Initializing random number generator from timer.

Doing OLTP test.
Running mixed OLTP test
Using Special distribution (12 iterations,  1 pct of values are returned in 75 pct cases)
Using "BEGIN" for starting transactions
Using auto_inc on the id column
Threads started!

OLTP test statistics:
  transactions:                        3312932 (11040.40 per sec.)
  read/write requests:                 62973258 (209859.48 per sec.)
  other operations:                    6627594 (22086.57 per sec.)

Test execution summary:
  per-request statistics:
  avg:                                   9.05ms
```

MySQL Using Ceph Cluster Storage with Replica 2

```
sysbench 0.4.12:  multi-threaded system evaluation benchmark

Running the test with following options:
Number of threads: 100
Initializing random number generator from timer.
```

OLTP-Level Performance Using Seagate NVMe SSDs with MySQL and Ceph

```
Doing OLTP test.
Running mixed OLTP test
Using Special distribution (12 iterations,  1 pct of values are returned in 75 pct cases)
Using "BEGIN" for starting transactions
Using auto_inc on the id column
Threads started!

OLTP test statistics:
  transactions:                          2546760 (8489.04 per sec.)
  read/write requests:                   48417674 (161389.19 per sec.)
  other operations:                      5095350 (16984.18 per sec.)

Test execution summary:
  per-request statistics:
  avg:                                   11.77ms
```

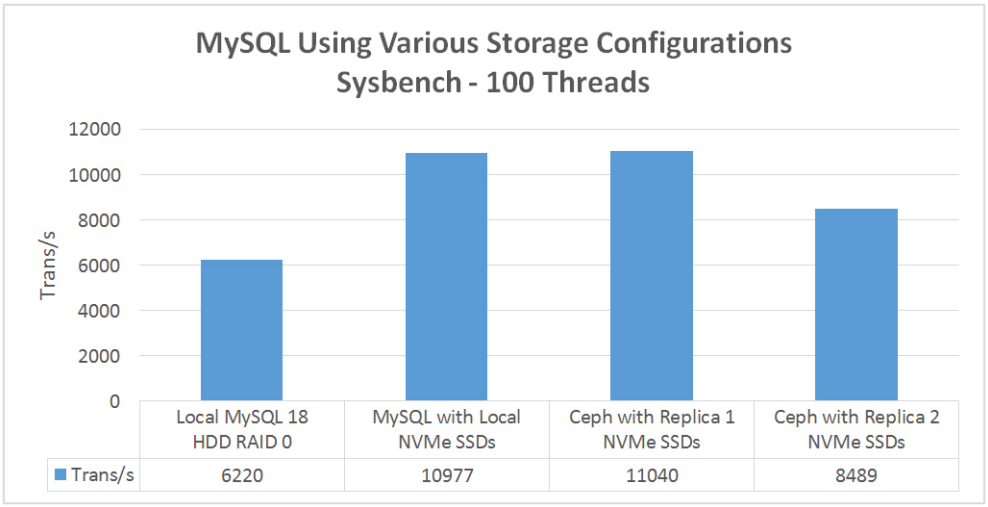

Figure 1: MySQL Using Various Storage Configurations (Sysbench - 100 Threads) - Transactions per second

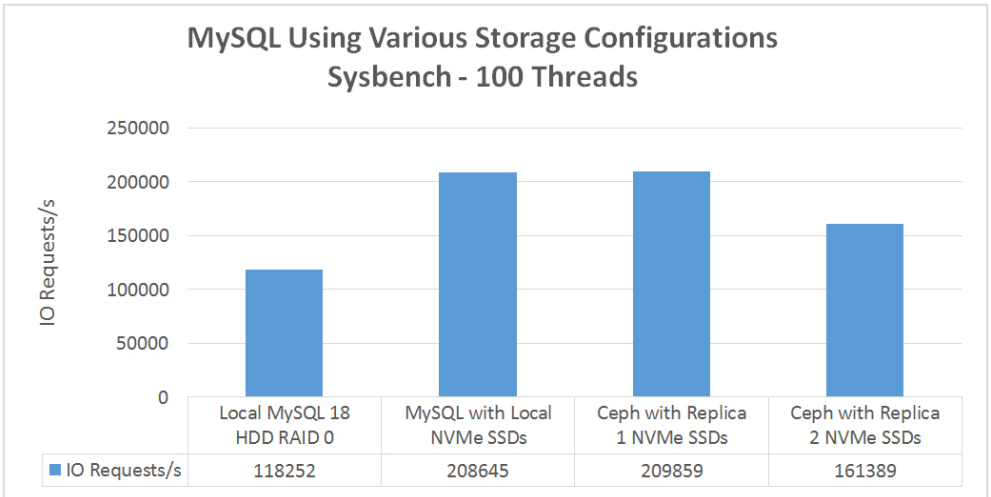

Figure 2: MySQL Using Various Storage Configurations (Sysbench - 100 Threads) - IO Requests per second

OLTP-Level Performance Using Seagate NVMe SSDs with MySQL and Ceph

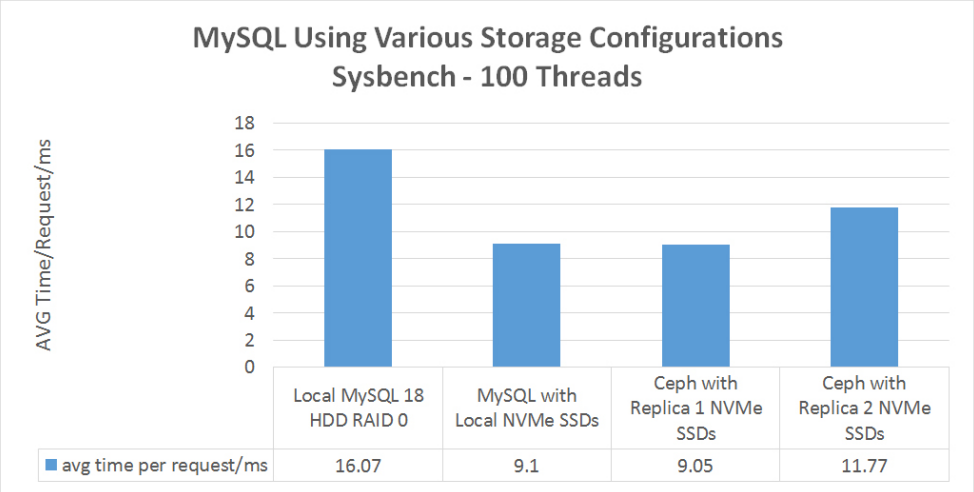

Figure 4: MySQL Using Various Storage Configurations (Sysbench - 100 Threads) - Average Time per Request

Conclusion

Software-defined storage (SDS) solutions, such as Ceph, have developed to a point where they can provide the flexibility required in current and future enterprise storage. Increasing numbers of enterprises are now considering deploying SDS in their data centers, and Ceph has become the distributed object store and file system of choice for SDS. Despite some of Ceph's inefficiencies in providing a scalable and reliable storage cluster, this paper's benchmarks and tests have shown that the combination of a MySQL database server using Ceph cluster storage for the database files and using NVMe SSDs with a 40G Mellanox network can provide the necessary IOPS to satisfy a database user's OLTP performance requirements.

seagate.com

AMERICAS Seagate Technology LLC 10200 South De Anza Boulevard, Cupertino, California 95014, United States, 408-658-1000
ASIA/PACIFIC Seagate Singapore International Headquarters Pte. Ltd. 7000 Ang Mo Kio Avenue 5, Singapore 569877, 65-6485-3888
EUROPE, MIDDLE EAST AND AFRICA Seagate Technology SAS 16–18, rue du Dôme, 92100 Boulogne-Billancourt, France, 33 1-4186 10 00

TP700-1702US February 2017