Western Digital.

SOLUTION BRIEF

Ceph™ Deployment on Ultrastar® DC HC520

Maximize performance and Capacity Minimize Power and Space

Enterprises and cloud providers are utilizing Ceph configurations as their preferred open-source, scale-out software-defined storage system. With commodity scale-out servers and locally attached storage, clusters from 50 terabytes (TB) to 10 petabytes (PB) are possible.

Unfortunately, the very nature of a scale-out system involves the possibility of server sprawl and the associated infrastructure headaches of increased power, cooling, rack space and management. The largest component of a Ceph cluster is the storage, so it is important to take care in selecting the type of storage used.

Optimizing Ceph Capacity and Density

In a Ceph deployment, the default method of ensuring data protection and availability is triple-replication, so for each usable byte of data there are two additional copies. This means that you must deploy at least three times as much raw disk space as compared to usable capacity, so the sizes rapidly add up. Western Digital provides high-density hard disk drives for Ceph deployment and because they're based on HelioSeal® technology, which seals helium inside the drive, they are also more power efficient. With the 12TB Ultrastar DC HC520 helium HDD, Ceph disk unit needs can be reduced by 50% as compared to 6TB air HDDs. That means you can halve the space and more than halve the power required without sacrificing capacity, or you can provide double the storage in the same space with a significantly smaller power footprint.

All writes to a Ceph cluster are double-buffered in a log drive before being committed to the Object Storage Device (OSD) drives. This allows for easier recovery in the case of a server or drive failure, but the resulting log can be a choke point for the entire system. As a rule of thumb, the write performance of the log device should match either the minimum of the network bandwidth or the sum of all the OSD drives' write performance. At 10 gigabit (Gb) speeds, the network can sustain around 1 gigabyte/second (GB/s). Most HDDs can write between 100 and 200 megabytes/second (MB/s), with a 12 drive system providing up to 2.4GB/s of raw drive bandwidth. To match these speeds, an NVMe™ compatible SSD is required. These can be installed in each OSD in either a front-loading U.2 format or a standard PCI Express add-in card.

Ceph Deployment Options

There are a multitude of options for Ceph deployment, ranging from an informal three-node cluster on repurposed hardware supporting a small office virtualization environment, to petabyte-scale deployments used in leading research institutions.

This paper focuses on two major deployment models: Enterprise-Scale Ceph Clusters and Rack-Scale Ceph Clusters. For Enterprise-Scale Ceph clusters, rollouts need hundreds of terabytes of storage, and the management, physical size of the array, and balance between storage and Ceph compute are crucial to success. For Rack-Scale Ceph clusters, the focus is more on storage density, as thousands of terabytes are required. In this case, storage characteristics such as power, cooling and interconnect all come into play in determining the proper strategy.

Enterprise-Scale Ceph Clusters

A compact, balanced rollout combining Ceph OSD compute and storage into multiple 1U high-density units is ideal for Ceph deployments at the hundreds of terabytes scale. A single, storage-optimized 1U server can be built to contain up to 12 Ultrastar DC HC520 HDDs in the unit, or a more general 2U configuration with all front-loading drives can be used. Industry-standard 10Gb networking can be used to access the storage, minimizing additional cost and the need for unique hardware.

Enterprise-Scale OSD Bill of Materials

- Dual-socket 1U/2U server, 3.5" drive slots
- 64GB DRAM
- 1 x NVMe Ultrastar DC SN200 add-in-card SSD for logs
- 12 × 12TB Ultrastar DC HC520 3.5" SAS or SATA HDDs for OSD data
- 10Gb network backbone

A four-node cluster of the configured OSD provides 576TB of raw storage, which is reduced to 576TB/3 = 192TB usable with replication. Additional nodes can be added to the cluster as storage needs grow.

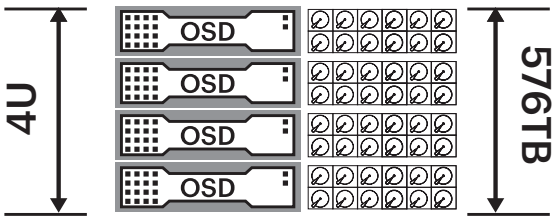

Figure 1: Enterprise-Scale Ceph OSD Architecture

Rack-Scale Ceph Clusters

Truly massive storage rollouts are what define Rack-Scale Ceph deployments, with petabytes of storage and more specialized hardware. Instead of in-server storage, external, high-density JBODs are required to meet capacity points. Higher CPU performance is also required in each of the OSD nodes, as they will be managing 60 to 90 12TB Ultrastar DC HC520 drives. Also needed is a large and fast SSD for the journals supporting these drives. Network bandwidth must also be increased to enable access to the stored data.

Rack-Scale OSD Bill of Materials

- Dual-socket 1U server, 3.5" drive slots
- 256GB DRAM
- 2 x NVMe Ultrastar DC SN200 add-in card SSD for logs
- 1 x SAS multiport RAID card and redundant cabling
- 1 × 4U 60-HDD JBOD
- 60 × 12TB Ultrastar DC HC520 3.5" SAS HDDs for OSD data
- 40Gb network backbone (or bonded 10Gb)

A full-rack (7-node) cluster of this configuration can provide over 5PB of raw storage, with 1.7PB usable capacity using three-way replication. At this scale, replication overhead often outweighs speed and simplicity benefits, so erasure-coded Ceph pools may be used. Using a "k=3, m=2" code divides each data element into five chunks, of which two may be lost without affecting availability (similar to three-way replication). Using this method requires less overhead than the replication method and would allow for nearly three petabytes usable per rack.

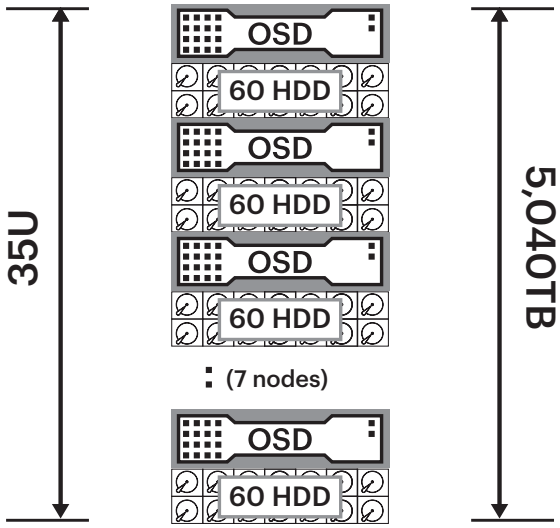

Figure 2: Rack-Scale Ceph OSD Architecture

Enterprise-Scale Ceph Cluster Proof of Concept

A four-node Enterprise-Scale Ceph cluster was rolled out in Western Digital labs to determine real-world power, latency and bandwidth. Standard three-way replication was used across the 2U JBOD systems, each with 12 SAS Ultrastar DC HC520 HDDs installed, along with an Ultrastar SSD in each OSD server for logs. In total, 48 Ultrastar DC HC520 HDDs and four Ultrastar NVMe SSDs were used for this test. Networking was provided by an FDR fabric.

The standard RADOS bench test was run against the cluster. The 4MB read and write results are summarized in the following graphs:

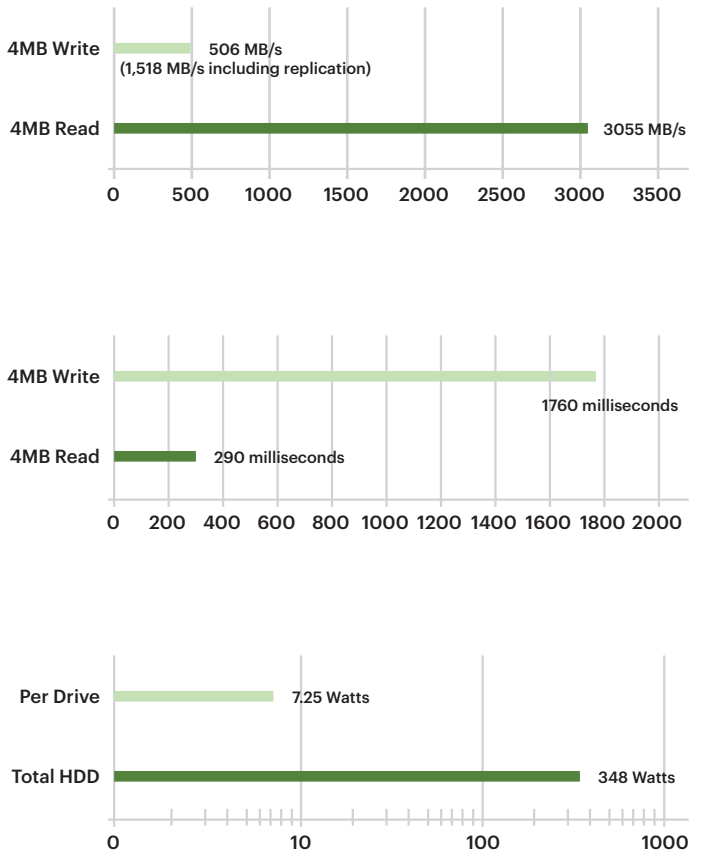

Figure 3: Instantaneous individual and total HDD power

As seen in Figure 3, client reads of over 3GB/s were achieved. Figure 4 shows latencies of under 300 milliseconds (ms) for the 4MB read tests. Client writes were somewhat lower due to the replication overhead (approximately 3x the bandwidth seen by the client was actually in use for the 3x replication).

During the 4MB read tests, the JBOD internal power monitoring was used to determine the instantaneous power of a single drive and of all 48 drives in the cluster. These results are reported in Figure 5. The 348 watts corresponds to a total watt/terabyte of 348/576 = 0.6W/TB.

Western Digital Delivers Unique Value for Ceph Deployments

The four-node Enterprise-Scale Ceph deployment proof of concept produced the following performance and power results:

- Over 3GB/s 4MB reads
- Over 500MB/s triplicated 4MB write
- Under 300ms 4MB read latency
- Only 0.6 W/TB power usage

A full Western Digital product portfolio, with everything from the world's highest capacity hard disk drives to cutting-edge solid state drives, makes it easy to roll out Ceph clusters. With the introduction of the 12TB Ultrastar DC HC520 hard drive, Western Digital has made it simpler than ever to build performant, low-power, and high-reliability OSD storage nodes for both enterprises and cloud organizations.

Learn more about Ceph storage by visiting the Ceph website at ceph.com. To learn more about the benefits of Ultrastar DC HC520 helium hard drives, visit www.westerndigital.com.

Western Digital.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
US (Toll-Free): 800.801.4618
International: 408.717.6000

www.westerndigital.com