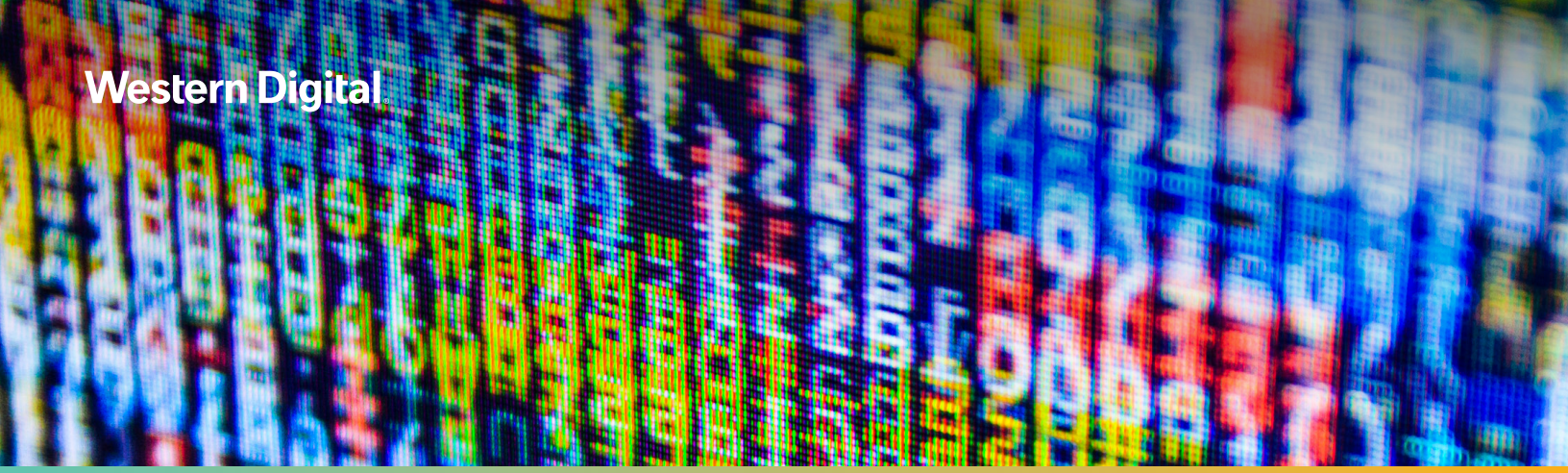

SOLUTION BRIEF

Optimizing Ceph™ Deployments

Ceph deployments are always scale-out, and in many cases these deployments can use cost-effective commodity hardware to provide a highly available, resilient data store for an enterprise. 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.

Highlights

Enterprises and cloud providers are using Ceph configurations as their preferred open-source, scale-out, software-defined storage systems. With commodity scale-out servers and locally attached storage, clusters from 50TB to 11.5PB are possible. Western Digital's enterprise-class hard drives and SSDs provide powerful ways to store more data and provide better performance than ever before.

Solution

- Store up to 672TB of raw data with a 4U storage cluster
- Power a rack-scale Ceph deployment with over 5.8PB of raw storage
- Provide faster writes for databases and critical operations with NVMe flash

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. For this reason, you must deploy at least three times as much raw disk space as usable capacity, so sizes rapidly increase.

Pain Point: Keeping Up with Storage Needs

Ultrastar® helium-filled HDDs provide the highest density available for Ceph deployment. Because they're based on HelioSeal® technology, which hermetically seals helium inside the drives, they are also the most power-efficient hard drives. By using 12TB Ultrastar DC HC520 helium HDDs, you can reduce Ceph disk unit needs by 50%, as compared to 6TB units and you can save even more with 14TB Ultrastar DC 530. This cuts the required space in half and the required power by more than half, without sacrificing capacity. Or, you can double the storage within the same space and realize an even smaller power footprint.

Pain Point: Low Performance of Databases and Write-Intensive Workloads

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 recovery in the case of a server or drive failure, but the resulting log can be a bottleneck 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 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 NVM Express™ (NVMe™) SSD with high write bandwidth, such as the Ultrastar DC SN200, is required. These SSDs 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 multiple options for Ceph deployment. These could range from informal three-node clusters on repurposed hardware for supporting a small office virtualization environment, to petabyte-scale deployments used in leading research institutions.

There are commonly 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 cloud-scale (or 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 play a part in determining the proper strategy.

Pain Point: Creating a Ceph Deployment for an Enterprise

For Ceph deployments of hundreds of terabytes, a compact and balanced rollout is ideal. This combines Ceph OSD compute and storage into multiple 1U high-density units. You can build a single storage-optimized 1U server with up to 12 Ultrastar DC HC530 HDDs, or you can use a more general 2U configuration with all front-loading drives. Also, you can use industry-standard 10Gb networking to access the storage, minimizing additional cost and the need for unique hardware. Finally, an NVMe Ultrastar DC SN630 SSD can power the write log, providing a good balance of performance and cost.

A four-node cluster in this configuration can use as little as a 4U rack space for the storage nodes while providing 672TB of raw capacity (=224TB usable with triple replication). Thanks to the scale-out nature of Ceph deployments, additional storage nodes can be added easily should storage requirements increase.

Pain Point: Providing a Ceph Deployment at Cloud Scale

Truly massive storage rollouts with petabytes of storage and more specialized hardware are also possible. Instead of in-server storage, external high-density 4U 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 14TB Ultrastar DC HC530 drives. For the journals, a large and fast NVMe SSD, like the Ultrastar DC SN620, is required to support the write performance of these massive amounts of hard disk drives. Higher speed networking, such as 25Gb or 100Gb Ethernet, is also a must to ensure that the data can be accessed by large numbers of users at high speeds.

A full-rack (7-node) cluster of this configuration can provide over 5.8PB of raw storage, with 1.96PB usable capacity using three-way replication. At this scale, replication overhead often outweighs speed and simplicity benefits, so erasure-coded Ceph pools might be used. Using a "k=3, m=2" code divides each data element into five chunks, two of which 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 3.5PB usable per rack.

Summary

The 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 tuned to meet your unique needs. For optimal cluster performance, combine the appropriate SSD—either SATA or NVMe-attached—with high-capacity hard drives.

	NVMe SSD	NVMe SSD	Helium HDD
Pain Point	**Ultrastar DC SN630**	**Ultrastar DC SN200**	**Ultrastar DC HC500 Series**
Keeping up with storage needs			● ● ●
Low Performance of databases and write-intensive workloads	● ● ●	● ● ●	
Creating a Ceph deployent for an enterprise	●		● ● ●
Providing a Ceph deployment at cloud scale	● ●	● ● ●	● ● ●

Legend: • = Good • • = Better • • • = Best

Western Digital.

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

www.westerndigital.com