Performance Testing of Object-Based Block Storage "Ceph" For Use in Cloud Services

Testing with the real-world usage scenario using storage servers fitted with Intel® Xeon® processor E5 family CPUs

"Cloud services have a need for greater storage capacity. With object storage having earned recognition for its high scalability and cost savings, we were able to demonstrate that Ceph is an effective distributed storage system."

– Takashi Kanai
R&D Office
Technology Development Department
IDC Frontier Inc.

Joint Testing Team
Takashi Kanai
R&D Office
Technology Development Division
IDC Frontier Inc.

Kiichi Yamato
Senior Researcher
R&D Office
Technology Development Division
IDC Frontier Inc.

Naoyuki Mori
Applications Engineer
Software Applications
Intel Architecture Technology Group
Intel K.K.

Executive Summary

A scalable block storage system was implemented using the Ceph open-source software for distributed storage.

As a provider of cloud services, IDC Frontier Inc. became interested in highly scalable object storage as a way to increase the capacity and reduce the cost of the storage used by these services. Ceph is a proven distributed storage software that supports block access, for which there is strong demand from users. IDC Frontier deployed a Ceph storage system and conducted tests of its basic data read and write performance. The tests also looked at the behavior when a disk fault occurs. The results demonstrated that this configuration provided adequate performance for use as a storage platform for cloud services, and that Ceph storage system was suitable for deployment.

1. Superior Object-Based Block Storage with Cost-Advantage of Being Vendor-Free

1.1 IDC Frontier

A member of the Yahoo! JAPAN group, IDC Frontier is a strategic IT infrastructure provider with a business based around cloud services and data centers. In terms of cloud services, the company launched a public cloud in 2009, followed in September 2011 by the release of a self-managed cloud service that uses the CloudStack* cloud platform software to allow users to perform tasks such as setting up virtual machines from a portal screen. The company currently offers public and private clouds through both self-managed and managed services. In April 2014, they released an object storage service that uses Riak* CS from Basho Technologies which has a high degree of compatibility with Amazon S3*

1.2 Scalable Storage in Growing Demand

Advances in IT are making the forms of data handled by businesses more diverse, extending beyond business files to encompass photographs, audio, video, and logs. Along with the associated transition to big data, it has been projected that the total quantity of data in the world will reach 50 ZB (40 trillion MB) by 2020. Meanwhile, the trend toward keeping data for longer to make it available for compliance and other uses is also contributing to the need for the efficient storage of both infrequently accessed "cold data" and frequently used "hot data". As for cloud services, demand for scalable storage is growing steadily, with a single virtual machine sometimes requiring terabytes of storage capacity.

The different types of storage can be broadly divided into block, file, and object storage. Block storage is able to provide high-speed access at a similar level to a server's local disk, making it suitable for database applications that require a high number of input/output operations per second (IOPS). However, as expensive storage area networks (SANs) are typically used for block storage, there are issues with cost.

Object storage, on the other hand, provides excellent extensibility by managing objects using unique identifiers. Because it has no restrictions on where data is stored, allowing the use of standard commercially available server hardware, highly scalable storage platforms can be implemented at low cost without being tied to a particular vendor. This led IDC

Frontier to evaluate Ceph as an economical alternative to conventional block storage solutions.

1.3 Ceph Selected for Testing: A Proven OSS Storage Software System

What led IDC Frontier to consider Ceph was its highly scalable architecture with the unlimited ability to scale out from nominal parameters. Because it is available as open-source software (OSS), it also reduces costs. Another feature is that it provides detailed settings for data redundancy, level of consistency, and where to store redundant data. Also, while a number of distributed file systems exist, another key feature of Ceph is its proven track record, having been included as standard in Linux* kernel version 2.6.34. It has also been the subject of performance tests by Intel. Having been impressed by the technical information this testing provided, IDC Frontier decided to conduct their own evaluation as a preliminary step before adopting the software themselves.

2. Assessment of Storage Performance in Cloud Services

2.1 Objectives

The objective of the testing was to determine whether a distributed storage system using Ceph could be utilized in IDC Frontier's cloud services. Accordingly, they tested its basic performance and reliability under load conditions likely to be encountered in real situations.

2.2 Test Items

IDC Frontier performed standard benchmark tests for cloud services on a triply redundant configuration commonly used in cloud services. However, whereas conventional practice in cloud services is to treat a read or write as valid if it is successful on two out of three nodes, IDC Frontier selected the more rigorous standard for data consistency of success on all three nodes.

(1) Operational testing under normal conditions

Five test patterns were selected. These consisted of four patterns for providing base data (256kB sequential read and write, and 4kB random read and write) and one mixed pattern (16kB random 10% read/90% write) to simulate a production environment. For each pattern, the total IOPS, throughput, and latency were measured as the number of virtual machines was progressively increased from one to 100.

Test patterns

1) 256kB sequential read

2) 256kB sequential write

3) 4kB random read

4) 4kB random write

5) 16kB 10% random read/90% random write

(2) Operational testing with HDD fault

During pattern 5 (16kB 10% random read/90% random write), a fault was triggered on one HDD and the latency was compared to when operating under normal conditions.

2.3 Test Schedule

December 2013 to January 2014: Configuration of test system, preliminary testing

February 2014: Determine tests to perform, reconfigure

March to April 2014: Formal testing

2.4 Test System

The test system was configured using IDC Frontier's own server, switch, rack, and other hardware at its Shirakawa data center. It consisted of nine storage node servers, each fitted with two Intel® Xeon® processor E5620 CPUs (2.40GHz, 4 cores, 8 threads) and 32GB of memory. The disks consisted of six 146GB SAS HDDs for data, one 146GB SAS HDD for the OS, and one 400GB SATA SSD (Intel® Solid-State Drive DC S3700 series) for journal data. For high-speed network access between servers, the Intel® Ethernet Converged Network Adapter X520 series (10Gbit Ethernet) was selected for the network interface cards.

The five client nodes node servers were each fitted with two Intel® Xeon® processor E5620 CPUs (2.40GHz, 4 cores, 8 threads), 64GB of memory, 146GB SAS HDD, and an Intel® Ethernet Converged Network Adapter X520 series (10Gbit Ethernet) network interface card.

The following software was used:

- Storage node software
  - OS: CentOS* 6.5 64bit
  - Ceph server: 0.72.2 (Emperor)
- Client node software
  - OS: CentOS* 6.5 64bit
  - Hypervisor: KVM
  - Cloud OS: OpenStack* Havana
  - Ceph client: 0.72.2 (Emperor)
  - Benchmark software: fio-2.0.13, libaio-0.3.107-10

Fig. 1 shows the test system configuration.

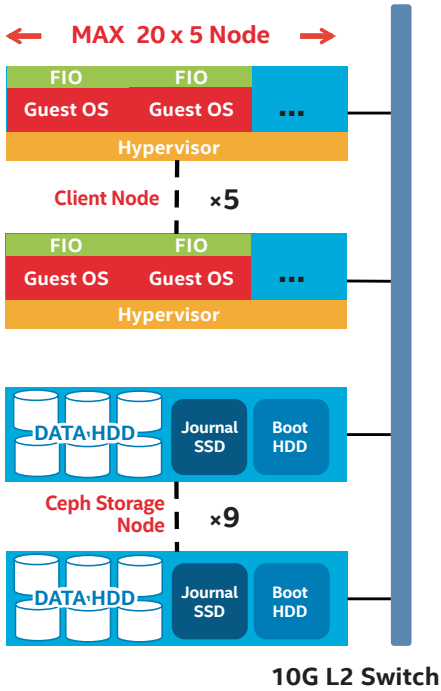

Fig. 1. Test System Configuration

3. Performance Matches Predictions

3.1 Test Results

3.1.1 Read/Write under Normal Conditions

The results of the 256kB sequential read and 256kB sequential write indicated no performance problems, with a total throughput of 1,467MB/s for reading and 767MB/s for writing (maximums in both cases).

Fig. 2 shows a graph of total IOPS for random reading and writing on the triply redundant configuration under normal conditions. Fig. 3 plots the mean latency measured under the same conditions as Fig. 2. As testing used HDDs running at 15,000 rpm, the estimated performance of each disk was 160 IOPS. A simple calculation assuming nine storage node servers each with six data storage HDDs gives a total of 8,640 IOPS (160 IOPS × 54 HDDs). Allowing for the triply redundant configuration, this gives a predicted performance of 2,880 IOPS (8,640 IOPS / 3).

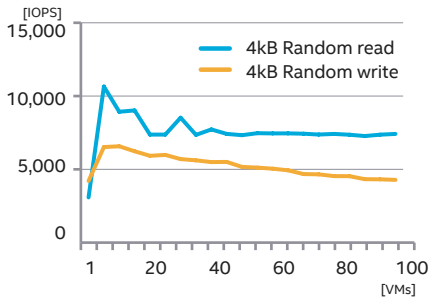

Fig. 2. Total IOPS for 4kB Random Read and Random Write on Triply Redundant Configuration under Normal Conditions

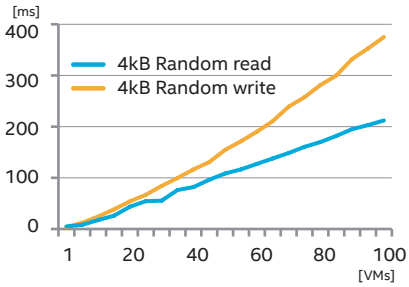

Fig. 3. Mean Latency for 4kB Random Read and Random Write on Triply Redundant Configuration under Normal Conditions

The IO performance measured in testing was roughly equal to the predictions, demonstrating that a Ceph distributed storage system can achieve performance close to the HDD capabilities. When a similar test using a doubly redundant configuration was performed during preliminary testing, it found that IO performance scaled with the number of disks.

3.1.2 16kB 10% Random Read/90% Random Write under Normal Conditions

Fig. 4 shows a graph of total IOPS measured for a test pattern that replicates an actual cloud service (10% random read/90% random write of 16kB) on the triply redundant configuration under normal conditions. As the number of virtual machines was progressively increased from one to 100, the total read and write IOPS values peaked around 5,000 for 20 virtual machines before falling away. If additional virtual machines led to IO performance degradation in a real service environment, additional nodes would be added to augment resources before the degradation occurred. That is, this testing was also able to measure what would happen if resources were not added.

54 HDDs delivering about 5,000 IOPS corresponds to about 92 IOPS per HDD, indicating that overheads are no more than expected. As the testing involved higher loads that occur in practice, even higher IO performance can be anticipated when used in a typical service environment.

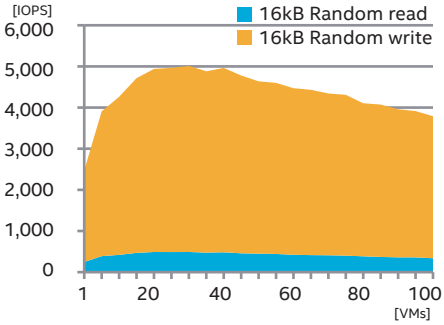

Fig. 4. Total IOPS for 16kB 10% Random Read/90% Random Write on Triply Redundant Configuration under Normal Conditions

3.1.3 Comparison of IO Performance under Normal and HDD Fault Conditions

Fig. 5 shows a graph of latency for 16kB 10% random read/90% random write and Fig. 6 shows the maximum and mean latency for a triply redundant configuration of ten virtual machines when a fault is present on one of the HDDs alongside the maximum and mean latency under normal conditions.

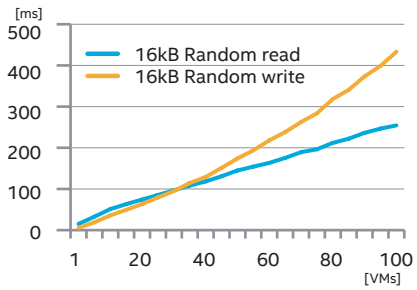

Fig. 5. Latency for 16kB 10% Random Read/90% Random Write on Triply Redundant Configuration under Normal Conditions

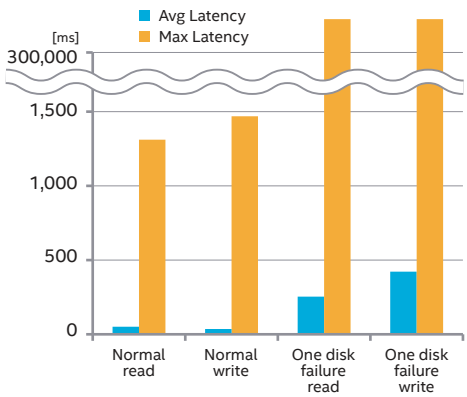

Fig. 6. Latency under Normal and HDD Fault Conditions (10 virtual machines)

Ceph[2] Distributed Storage System

Ceph is a distributed storage system that runs on Linux*. It performs distributed management of object data using the RADOS method. Unlike conventional distributed file systems that have include centralized management, Ceph features a highly scalable architecture that eliminates single points of failure by using its CRUSH algorithm to store data in a distributed fashion. It is made up of three key components: Monitors, OSDs, and clients. The Monitors maintain a map of system resources which they adjust when faults are detected or additional storage is added. The OSDs manage object storage and replication, and using map information from the monitors, use the CRUSH algorithm to determine where replicated data is to be placed. The clients use map information from the monitors and use the CRUSH algorithm to determine which OSDs contain the primary copy of data. These components act as a system to achieve both high I/O performance and high reliability.

Another major feature of Ceph is that, through the client implementation, it can also be used for object storage, as a POSIX* compatible file system, or as block device storage as in this testing.

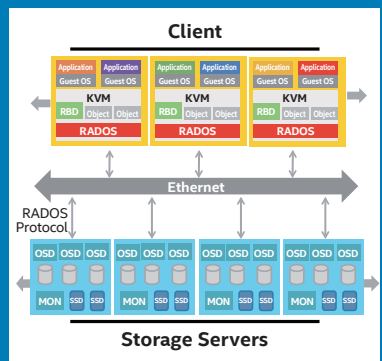

Fig. 7. Ceph Architecture

RADOS: Reliable Autonomic Distributed Object Store

CRUSH: Controlled Replication Under Scalable Hashing

RBD: RADOS Block Device

OSD: Object Storage Device

MON: Monitor

As shown in Fig. 6, the maximum latency when a disk fault occurs is 345s (344,800ms), an order of magnitude higher. In contrast, the maximum latencies under normal conditions are only 1.3s for reading and 1.5s for writing. On the other hand, the mean latencies are 254ms (read with fault), 423ms (write with fault), 51ms (normal read), and 36ms (normal write). This result indicates that, apart from the time when the actual fault occurs, the latency is less than the 1s used as a criterion for judging faults in practice, and therefore avoids a momentary outage in the cloud service.

3.2 Remarks

Results of the testing that simulated actual operation demonstrated that a Ceph distributed storage system is able to make effective use of HDD performance, and demonstrated that it is a viable storage system for cloud service. In terms of scalability, while there is a need for additional testing of performance when nodes are added, given that the testing described here was able to make full use of the HDD capabilities, it is anticipated that similar results would be obtained.

Meanwhile, while testing of HDD fault conditions adopted the strictest data consistency settings, whereby all three nodes holding distributed data must operate correctly for an operation to be deemed successful, it was recognized that, for reasons of usability, these redundancy settings need to be reviewed. The fault behavior tested here was consistent with the CAP Theorem which recognizes that requirements for Consistency, Availability and Partition tolerance cannot all be satisfied at once in a distributed system. That is, although meeting two of these requirements at once is possible, meeting all three is difficult. IDC Frontier intends to investigate redundancy methods further, including additional technical testing.

As these tests used the company's own hardware with which the Technology Development Department is already familiar from its routine work, including the processors and disks, one of the intangible results of the work was to provide an intuitive feel for everything from performance to cost, thereby helping with decision-making on whether to offer the service commercially.

3.3 Future Outlook

The Technology Development Department intends to continue evaluating the performance and scalability of Ceph distributed storage systems. While the testing described here used KVM as the hypervisor and OpenStack* as the cloud platform software, they intend to look at service introduction and will consider testing on CloudStack*, the standard cloud platform used by IDC Frontier. They also plan to look at establishing performance monitoring methods and fault response procedures, and at selecting or developing operational tools.

While this testing was focused on Ceph, it is not the only distributed storage system. IDC Frontier intends to continue its study and research, seeking actively to adopt technologies that benefit users in terms of both cost and functionality, and to provide high-quality services with overseas users in mind in the future.

For more information on the Intel® Xeon® processor E5 family, visit
http://www.intel.com/content/www/us/en/processors/xeon/xeon-processor-5000-sequence.html

[1] Source: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

[2] http://ceph.com/

JPN/1406/PDF/SE/DCG/ME 330676-001US