Global Storage Consulting

MOC Ceph Storage Solution

Red Hat Global Storage Consulting

February 1st, 2016

v1

Confidentiality, Copyright, and Disclaimer

THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

Customer	**Boston University - MOC**
Location	**Boston, MA**
Duration	**2016-01-19 to 2016-01-29**
Report author	**Tyler Brekke**
Consultant(s)	**Tyler Brekke**
Ceph version	**Red Hat Ceph Storage (RHCS) 1.3.1 (Based on Ceph v0.94.4)**
Raw capacity	**432 TB (4 TB SATA disks, 12 disks/node, 10 nodes)**
Operating System	**RHEL 7.2**
Use Case	**Openstack**

Executive Summary

While onsite at Boston University I reviewed the current hardware and design for the Massachusetts Open Cloud (MOC). I made recommendations based on my findings, including adding SSDs for journals to the current machines. I configured and set up RHEL 7.2 on each of the storage nodes and installed and configured Red Hat Ceph Storage (RHCS).

Since there were multiple RAID configurations between the nodes, I ran performance benchmarks on each of the different configurations. Based on the results I was able to make recommendations for future hardware, which will yield greater and more consistent performance.

I also integrated RHCS to a Devstack install to show the complete functionality of Ceph with OpenStack Cinder, Glance, Nova, and Keystone. All hours in the Scope of Work were consumed and the engagement was successfully concluded.

Use Cases

The use case for Ceph at MOC is to provide block storage to an OpenStack environment. The Ceph Rados Gateway is used to provide Object Storage to the OpenStack users using Keystone for authentication.

Cluster Description

The Boston University Ceph cluster consists of 10 storage nodes and 3 monitor nodes. The Ceph Rados Gateway is currently collocated on one of the monitor nodes.

10 x OSD Nodes

Product	Lenovo x3650
CPU	2 x E5-2650 V2
Memory	128GB
Disk - OSD	12 x 4TB
Network	2 x 10GbE
Controller	M5110 - Raid, M5110 - JBOD, F5115 - RAID, or N2215 HBA (6-12GB/s)

Journal Relocation

SSDs were purchased while I was onsite. We were able to successfully add the SSDs to two of the OSD servers. ceph-lenovo01 and ceph-lenovo10.

The following script is an example of how the process works, a new partition is created on the SSD, the journal is flushed, the symlinks to the journal location are created, then the mkjournal command is ran.

```
#!/bin/bash

journal_device="/dev/sda"
osds=(0 1 2)
num_journal_part=${#osds[@]}
i=1
while [ $i -le $num_journal_part ]
do
    disk_id=`uuidgen`
    sgdisk --new=${i}:0:+5120M --change-name="${i}:ceph journal"
--partition-guid=${i}:${disk_id} --typecode=${i}:45b0969e-9b03-4f30-b4c6-b4b80ceff106
--mbrtogpt ${journal_device}
    partx -a $journal_device
    service ceph stop osd.${osds[$i-1]}
    ceph-osd -i ${osds[$i-1]} --flush-journal
    ln -sf /dev/disk/by-partuuid/${disk_id}
/var/lib/ceph/osd/ceph-${osds[$i-1]}/journal
    echo $disk_id > /var/lib/ceph/osd/ceph-${osds[$i-1]}/journal_uuid
    ceph-osd -i ${osds[$i-1]} --mkjournal
    service ceph start osd.${osds[$i-1]}
    i=$[$i+1]
done
```

Performance Testing

I expressed concerns over the different RAID configurations between the machines. My past experiences have shown me that the different controller configurations can have vast differences in performance.

Based on my findings I was surprised to find that the M5110 Controller with 1GB Flash performed the best on average for writes of different block sizes.

ceph-lenovo01 containing the M5110 Controller outperformed the other configurations significantly when performing writes. The reads were slower than the HBAs on ceph-lenovo09 and ceph-lenovo10, but only marginally. I chose this node to test with SSD journals as well as the HBA on ceph-lenovo10.

Node	4k Writes	4k Writes with SSD journal
ceph-lenovo01 (M5110)	8.63 MB/s (2.2K IOPS)	14.60 MB/s (3.7K IOPS)
ceph-lenovo10 (N2215 HBA)	7.26 MB/s (1.8K IOPS)	8.98 MB/s (2.3K IOPS)

What this shows is that ceph-lenovo01 benefited much more than ceph-lenovo10 when SSD journals were added. The controller is likely the limiting factor for ceph-lenovo10. The tests I ran did not encompass the amount of testing typically performance for the Ceph Performance Evaluation done by Red Hat, but it does show the SSDs helped significantly and shows that the M5110 card to be the correct choice in the future for new nodes.

Testing Notes

The test I used to benchmark the different machines is included with the Ceph packages.

```
$ rados bench <write|seq|rand> <time> -p <pool> -b <bytes> -t <threads> --no-cleanup
```

I setup different crush rules to allow the benchmark to only run on certain hosts. When creating a pool to bench, the crush rule is then set so all IO will only occur on a specific host. Running the test across the default crush rule will perform IO across the entire cluster.

A single run of rados bench will likely not be able max out the throughput of a single host. I performed this test concurrently on 6 different hosts using a set of scripts to run these in sync and aggregate the results.

Engagement Details

The work completed each day was:

Tuesday January 19

- Morning was spent reviewing and validated the current hardware. I recommended SSD journals, specifically Intel DC S3700s.
- We drove to the datacenter and spend the day configuring the hardware and installing the operating systems.

Wednesday January 20

- Started setup up the networking.
- Configured local RHEL repository
- Upgraded nodes to RHEL 7.2
- VM was setup to be used for Ceph Admin node.

Thursday January 21

- Networking configuration completed for public and cluster networks.
- Updated the monitor nodes to RHEL 7.2.
- Set up firewalls.
- Configured Ceph Admin VM as NTP Server.
- Created cephdeploy user across the systems, setup passwordless SSH, adjusted Kernel.
- Installed all the Ceph packages from the ISO.
- Deployed the monitors and OSDs.
- Ran into network issues on the Cluster Network.
- Deployed Calamari

Friday January 22

- Performed detailed profiling of the hardware while the cluster network issues were being resolved.
- Updated the PS testing suite as some of the RAID configurations did not work correctly with it.

Monday January 25

- Completely removed and redeployed the cluster. This time with Rado from Boston University driving. While I shadowed the deployment.
- Created custom Crush Roots for each of the hosts with different hardware configuration. This included the machines. ceph-lenovo01, ceph-lenovo03, ceph-lenovo07, ceph-lenovo09 and ceph-lenovo10.
- Began with rados bench testing using the PS team perf-testing toolkit.

Tuesday January 26

- Continued rados bench testing.
- Testing today showed which RAID configurations would be best to put the SSD journal into.
- SSD journals were installed in ceph-lenovo01 and ceph-lenovo10

Wednesday January 27

- SSD journals replaced the ondisk journals the machines ceph-lenovo01 and ceph-lenovo10 had.
- I performed the journal relocation on one of the nodes then walked Rado through how to do this on the other node.
- Continued running rados bench testing and the new SSD journal nodes.

Thursday January 28

- The latest Devstack was installed on a spare node.
- Ceph was integrated to cinder, glance and nova. All these functions were tested and appeared to be working as expected.

Friday January 29

- Gave presentation on performance findings.
- Integrated RadosGW with Keystone.
- UUID keys worked correctly, but fernet keys errored. Same issue as http://tracker.ceph.com/issues/12761

- Tested the patch as it is already included with the upstream version and the issue was fixed.
- Reported in the open BZ https://bugzilla.redhat.com/show_bug.cgi?id=1300855