Our Storage Journey:

Migrating VMWare onto SUSE Enterprise Storage (SES)

Joe Comeau
Information Systems Manager
James Hogg Research Center
joe.comeau@hli.ubc.ca

Sean Richerd
Senior Sales Engineer - Canada
SUSE
srickerd@suse.com
Agenda

• Who We Are
• SUSE Enterprise Storage
• Hardware Configuration
• Installation & Testing
• Problems & Issues
• Our VMWare Setup
• Future
Who Are We?
World Leaders in Heart & Lung Disease

Dr. Bruce McManus joins Order of Canada

JULY 4, 2018

Congratulations to Dr. Bruce McManus, who has been appointed a Member of the Order of Canada.
Providence Healthcare Research Institute

Providence Health Care
• St Paul’s Hospital
  - Built in 1894

Research Centre
• Started by James Hogg in 1976
New St Paul’s!

• New Hospital just approved Feb 15, 2019

• Build to be completed by 2026

• Will include a dedicated research building
University of British Columbia

University of BC
• Enrollment ~ 60,000 students
• Medical School for Province of BC

(1) Vancouver
• ~ 10 km
Why SUSE Enterprise Storage?
Much of Our Research Is Image Based

• Research data growth
  - More than 25% per year growth

• New

• Examples of data producers:
  - Micro-CT (~ 20GB / scan)
  - Digital slide (~ 2 GB/slide)
  - Gene sequencer (~ 80GB / Scan)
  - Confocal Microscope
  - MRI imaging
Storage Issues

Array Challenges
• Capital cost
• Maintenance
• Upgrade
• Expansion
• Forklift replacement

Needed Something Flexible
• More storage
• Lower purchase price
• Affordable maintenance
• Supported
• Expandable

Common Limitations of Traditional Enterprise Storage

- Unable to scale and manage data growth
- Expensive
- Won’t extend to software-defined data center
Replace Current Array with SES

- Primarily VMware Block
- Oracle and Microsoft Database
- Web Services
- Windows Applications
- Network File Shares
- Backup and Archive
Configuration Choices
What We Chose

• Dell Servers
• SUSE Enterprise Storage v4
• Intel p3700 NVMe plus Intel CAS
## Hardware

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
<th>Qty</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSD Server</td>
<td>Dell R730xd – 2 x E5-2650v3 cpu – 8 x 32 GB DDR4 – 2 x 120GB Boot SSD – 12 x 8 TB NL-SAS HDD – 2 x Intel P3700 NVMe - 2 x 40GBe</td>
<td>6</td>
</tr>
<tr>
<td>OSD Server</td>
<td>Dell MD1400 exp tray – 12 x 8TB NL-SAS HDD</td>
<td>6</td>
</tr>
<tr>
<td>Monitor Node</td>
<td>Dell R630 – 2 x E5-2623v3 – 8 x 8GB DDR4 – 1 x 120 Boot SSD – 2 40GB-E</td>
<td>3</td>
</tr>
<tr>
<td>Network</td>
<td>Dell S6000-ON – 2 x 32 port 40GB-E QSFP switch</td>
<td>2</td>
</tr>
</tbody>
</table>
Network Architecture
Intel CAS

SUSE Solid Driver Program
https://drivers.suse.com/intel/Intel-CAS/sle-12-sp2-x86_64/1.0/

Enterprise Version v3.8
• We worked with Intel & SUSE on the

Intel CAS is now available through SUSE Solid Driver Program

Intel in partnership with SUSE is providing Intel CAS through the SUSE Solid Driver Program which removes the complexities and risks associated with deploying required kernel drivers, and gives customers an easy way to identify fully compatible and supported software code. Intel CAS combined with Intel DC SSDs can improve performance and user experience when used with SUSE Enterprise Storage.

In addition, Intel and SUSE are working on interoperability between Intel CAS with SUSE Enterprise Storage and further enhancing the performance of the new Ceph BlueStore functionality. BlueStore is a new storage backend for Ceph that operates close to the hardware, thus eliminating the abstraction layers associated with POSIX filesystem operations. Early testing indicated that BlueStore will improve write speeds by up to a factor of two over previous versions, although the actual number could vary depending on your hardware and your Ceph configuration.
Increased Storage at a Reduce Price

- Cost per Gbyte
- Maintenance
- Expansion & Scalability
- Performance
- Capacity
- Latency
Installation & Testing
## Base Network and Disk Configurations

**Network**
- Bonded 40 Gb network
- Combined cluster and public vlans
- Network Speed tests (iperf)

**Server**
- Turned off all power-saving features
- Set cpu speeds
- Numa Settings
- Baseline disk tests (dd)

<table>
<thead>
<tr>
<th>Test</th>
<th>Command</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>iperf</td>
<td>iperf -s on MON01 and iperf -c on OSD01</td>
<td>25.6 Gb/s</td>
</tr>
<tr>
<td>iperf</td>
<td>iperf -s on OSD04 &amp; iperf -P 10 -c on MON02 so 10 streams</td>
<td>36.9 Gb/s</td>
</tr>
<tr>
<td>dd</td>
<td>dd if=/dev/zero of=/dev/sde1 bs=1G count=1 oflag=direct</td>
<td>218 MB/s</td>
</tr>
<tr>
<td>dd</td>
<td>dd if=/dev/zero of=/dev/nvme1n1p15 bs=1G count=1 oflag=direct</td>
<td>1433 MB/s</td>
</tr>
</tbody>
</table>
Storage Tests

• **fio storage tests**
  - Ran fio on our old storage array from a VM
  - Ran fio on the SES storage from a VM
  - Goal was to have the SES + Intel CAS meet or beat our array

• **We also ran fio tests against the SES cluster with and without Intel iCAS enabled**

• **Fio command**
  ```
fio --filename=/dev/rbd0:/dev/rbd1:/dev/rbd2:/dev/rbd3:/dev/rbd4 --direct=1 --ioengine=libaio --bs=4k --rw=randwrite --iodepth=16 --numjobs=4 --ramp_time=120 --runtime=900 --group_reporting --name=4kRandWriteI016J4-ceph-vm-run1 >> /root/IntelTests/4kRandWriteI016J4-ceph-vm-run1
  ```
Array Comparison Results

Guest machines

- VMware - 4 Gb – 2 vcpu
- 8Gb FC (Current) vs 40 Gb iSCSI (SES + Intel CAS)
- 4k Random Write
- Tests performed with Fio
  - 3 runs each and then averaged

<table>
<thead>
<tr>
<th>Test</th>
<th>Current Array</th>
<th>SES + Intel CAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth (MB/s)</td>
<td>7.48</td>
<td>120.33</td>
</tr>
<tr>
<td>IOPs</td>
<td>1915</td>
<td>30,771</td>
</tr>
<tr>
<td>Avg Latency (ms)</td>
<td>34</td>
<td>2.08</td>
</tr>
<tr>
<td>Latency – 99.95%</td>
<td>647.8</td>
<td>158</td>
</tr>
</tbody>
</table>
Fio Test Comparisons

Array Comparison

- BW (MB/s)
- IOPS
- Avg Lat (ms)
- Lat 99.95%

SES with & without iCAS

- BW (MB/s)
- IOPS
- Avg Lat (ms)
- Lat 99.95%
Problems & Workarounds
Some Growing Pains

- Weekly crash
- Migration
- Operator error
- Scrubbing
- Placement groups
- Network card failure
Ghost in the Machine
Weekly Crash

Before going into production

• We had a weekly crash
  - Took out one or more OSD servers
  - Different servers each week – always sometime Monday

• 24 journals on NVMe with iCAS looked like 24 SSDs XFS attempted to trim
  - This kicked off after midnight on Sundays

• Limiting XFS to just the sys volume fixed this issue
• Bluestore does not need trimming
Network Switch Failure

• We have two switches
  - Redundant paths

• Intermittent slow network connections
  - Always different issues

• One of the switches was faulty
  - It eventually failed after 4 months
Upgrades
Filestore to Bluestore

• Upgraded from SES4 to SES5
  - Maintained filestore
  - One node at a time
  - As we were running Intel CAS

• Upgraded Filestore to Bluestore
  - Removed OSD server from cluster
  - Added server back in with Bluestore
  - One node at a time

• Cluster is encrypted

• Took approximately 1 week for each upgrade
We Noticed

- **Encryption**
  - Slows down performance

- **Bluestore vs Filestore**
  - A slight decrease in performance with Bluestore
  - Filestore used to provide cache via the xfs filesystem

- **Boot times**
  - 45 minutes for a boot due to encryption
  - Hopefully addressed with SES6
Operator Error
Operator Error

During migration or upgrades
- Cluster is in a noout config
- Let the cluster then rebuild the data on the new node
- Make sure it finishes!
  - Regardless of health

iSCSI
- Adding and removing RBDs for iSCSI
- OpenAttic features added with SES 5.5
Placement Groups
Scrubbing

- Scrubbing limits performance
- Set for times not heavily used
- Disable when performing maintenance

Our Settings

Scrub sleep = 1
Scrub begin hour = 19
Scrub end hour = 7
Scrub chunk min = 5
Scrub chunk max = 20

Osd scrub min interval = 172800 (2 days)
Osd scrub max interval = 1209600 (14 days)

Osd deep scrub interval = 2419200 (28 days)
Placement Groups

Understand placement groups
• Global placement groups vs pool placement groups
• Effectively assign storage to your cluster

We had to remove RBDs and pools
• Increase placement groups per pool
• We didn’t know what we would use the storage pools for
Ceph Balancer

New to us

ceph mgr module enable balancer

ceph balancer eval
ceph balancer optimize mar15
ceph balancer show mar15
ceph balancer eval mar15
ceph balancer execute mar15 (*rebalances)

ceph mgr module disable balancer

The lower the setting, the better
Host Failure
Network Card Failure in Host

Planned datacenter work
• Cooling tower work was being performed
• Temporary chillers in place

Soon after starting temp chillers
• Lost a ceph node
• But still connected to server
• 2 hours to diagnose
• 1 hour to replace and test
• 45 minute boot
Recovery During Nic Failure
Our VMWare Config
VMware on SUSE Enterprise Storage

Servers
• Approximately 150 VMs
• Occupying ~120 TB data
• Located on SES

Backup
• On SES ~100 TB
• Remote Backup ~200 TB
  - 350 KM away
Virtualization Config

Hardware

Hosts
• 256 GB RAM
• 24 Core x 2 cpu
• Local ssd for boot
• 40 GB Nic
• 1.6 TB NVMe for caching

Network
• 40 Gb network

Internet config
• 1Gb network (Remote backup)

Software

Vcenter
• Ver 6.7 Update 1

Virtucache
• Host cache

Veeam Backup and Replication
• Ver 9.5 Update 4

Backup Repository
• Local = 2 x Windows 2019 ReFS
• Remote = Dedupe appliance
Virtucache

Using NVMe in host server
We have write/read cache to reduce latency of VMs
iSCSI config

2 iSCSI portals configs
1 pool assigned for
5 RBD images for datastores

Use openattic to create RBD images
iSCSI Tuning

iSCSI Gateway performance tuning for VMware environment (70230530)

iSCSI Gateway performance tuning for Veeam environments (7023150)

Additional settings we have implemented

1. CPU settings
   - CPU performance: we permanently increased speed

2. Round Robin
   - Set round robin to 1
Backup Repository

• Windows servers as backup repositories
• Using Resilient File System (ReFS)
  • Major feature is block cloning

• Allows us to maintain 4 weeks of full backups locally
• Only using slightly more than 1 week’s backup space
Backup Repository

Showing repository usage & backup file sizes

Pre-Backup

During Backup

Post-Backup
Backup Loads
Backup Performance – Jobs View
Backup Performance – Cluster View

iSCSI Portal & Client

Intel Cas – Dirty Cache
OpenAttic View of Backups

Cluster Read & Write Performance
Future
Our Future

AMD Epyc Servers
• High cpu and memory configs
• Many 2.5 U.2 connections for NVMe

VSan or Storage Spaces Direct
• For High transaction servers and applications
• USE NVMe drives for storage

SUSE Enterprise Storage
• Additional storage servers
• NFS Ganesha or Ceph FS for HPC mount

Openstack?
• Need a backup solution
Thanks to

SUSE
Cameron Seader
Craig Liddle
Scott Warner
Boyd Memmott **

SUSE Canada
Sean Rickerd
Mathew Himelfarb
Maxx Piette
Unpublished Work of SUSE LLC. All Rights Reserved.

This work is an unpublished work and contains confidential, proprietary and trade secret information of SUSE LLC. Access to this work is restricted to SUSE employees who have a need to know to perform tasks within the scope of their assignments. No part of this work may be practiced, performed, copied, distributed, revised, modified, translated, abridged, condensed, expanded, collected, or adapted without the prior written consent of SUSE. Any use or exploitation of this work without authorization could subject the perpetrator to criminal and civil liability.

General Disclaimer

This document is not to be construed as a promise by any participating company to develop, deliver, or market a product. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. SUSE makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. The development, release, and timing of features or functionality described for SUSE products remains at the sole discretion of SUSE. Further, SUSE reserves the right to revise this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes. All SUSE marks referenced in this presentation are trademarks or registered trademarks of SUSE LLC. in the United States and other countries. All third-party trademarks are the property of their respective owners.
Backup Slides
Daily Traffic Through 1 iSCSI Portal
List the Most Used OSDs

Use this to see the 20 fullest OSDs

```bash
ceph osd df | (read -r; printf "%s\n" "$REPLY"; sort -nrk9) | head -n 20
```

<table>
<thead>
<tr>
<th>ID</th>
<th>CLASS</th>
<th>WEIGHT</th>
<th>REWEIGHT</th>
<th>SIZE</th>
<th>USE</th>
<th>AVAIL</th>
<th>%USE</th>
<th>VAR</th>
<th>PGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>114</td>
<td>hdd</td>
<td>7.32069</td>
<td>0.90000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>66.66</td>
<td>1.34</td>
<td>95</td>
</tr>
<tr>
<td>60</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>65.99</td>
<td>1.33</td>
<td>97</td>
</tr>
<tr>
<td>86</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>61.88</td>
<td>1.24</td>
<td>91</td>
</tr>
<tr>
<td>90</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>59.63</td>
<td>1.20</td>
<td>89</td>
</tr>
<tr>
<td>42</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>58.84</td>
<td>1.20</td>
<td>88</td>
</tr>
<tr>
<td>16</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>56.17</td>
<td>1.19</td>
<td>87</td>
</tr>
<tr>
<td>2</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>59.16</td>
<td>1.19</td>
<td>87</td>
</tr>
<tr>
<td>45</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>56.64</td>
<td>1.18</td>
<td>86</td>
</tr>
<tr>
<td>159</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>57.12</td>
<td>1.18</td>
<td>85</td>
</tr>
<tr>
<td>127</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>57.09</td>
<td>1.15</td>
<td>84</td>
</tr>
<tr>
<td>19</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>57.11</td>
<td>1.15</td>
<td>84</td>
</tr>
<tr>
<td>6</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>57.11</td>
<td>1.15</td>
<td>84</td>
</tr>
<tr>
<td>139</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>56.49</td>
<td>1.14</td>
<td>83</td>
</tr>
<tr>
<td>75</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>56.46</td>
<td>1.14</td>
<td>83</td>
</tr>
<tr>
<td>35</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>56.47</td>
<td>1.14</td>
<td>83</td>
</tr>
<tr>
<td>164</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>56.44</td>
<td>1.13</td>
<td>83</td>
</tr>
<tr>
<td>101</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>56.45</td>
<td>1.13</td>
<td>83</td>
</tr>
<tr>
<td>167</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>55.08</td>
<td>1.11</td>
<td>81</td>
</tr>
<tr>
<td>166</td>
<td>hdd</td>
<td>7.32069</td>
<td>1.00000</td>
<td>7.2e+09</td>
<td>4.8e+09</td>
<td>2.4e+10</td>
<td>55.09</td>
<td>1.11</td>
<td>81</td>
</tr>
</tbody>
</table>
Ceph Balancer

Heavy Load on Systems

Every 2.0s: ceph status

cluster:
  id: 1166bce1-b0b6-938e-8564-93b2b3208027
  health: HEALTH_WARN
    12629c15/745076337 objects misplaced (1.695%)
    Degraded data redundancy: 84/745076337 objects degraded (0.000%), 1 pg degraded

services:
  mon: 3 daemons, quorum mon01,mon02,mon03
  mvr: mon02 (active), standbys: mon01, mon00
  osd: 160 osds: 160 up, 160 in; 260 remapped pgs

data:
  pools: 2 pools, 8320 pgs
  objects: 298.36K objects, 227718
  usage: 711718 used, 5121E / 1.13P avail
  pgs: 84/745076337 objects degraded (0.000%)
    12629c15/745076337 objects misplaced (1.695%)
    7391 active+clean
    302 active+remapped+backfill_wait
    86 active+remapped+backfilling
    1 active+recovery_wait+decapped

io:
  client: 38.5GB/s rd, 10.5GiB/s wr, 354.81kop/s rd, 502.20kop/s wr
  recovery: 17.8GiB/s, 18.20kobjects/s
CPU Service Setting

Created a systemd service for cpu settings in /etc/system/system/cpu.service

[Unit]
Description=CPU Performance

[Service]
Type=oneshot
ExecStart=/usr/bin/cpupower frequency-set -g performance
ExecStart=/usr/bin/cpupower idle-set -D 0

[Install]
WantedBy=multi-user.target

See cpu settings

watch -n 1 grep MHz /proc/cpuinfo
CPU with vs CPU without Idle States

<table>
<thead>
<tr>
<th>Nehalem</th>
<th>SandyBridge</th>
<th>Mperf</th>
<th>RAFL</th>
<th>Idle States</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKG</td>
<td>CPU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>36.86</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>20</td>
<td>36.88</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>58.54</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>58.55</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>77.89</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>77.91</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>70.90</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>70.90</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>66.54</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>66.54</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>68.43</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>68.43</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Nehalem</th>
<th>SandyBridge</th>
<th>Mperf</th>
<th>RAFL</th>
<th>Idle States</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKG</td>
<td>CPU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>20</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>26</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>
Changing Scrub Settings

Located in:
/srv/salt/ceph/configuration/files/ceph.conf.d

- osd_scrub_sleep = 1
- osd_scrub_chunk_min = 5
- osd_scrub_chunk_max = 20
- osd_scrub_begin_hour = 19
- osd_scrub_end_hour = 7
- osd_scrub_min_interval = 172800
- osd_scrub_max_interval = 1209600
- osd_deep_scrub_interval = 2419200

salt-run state.orch ceph.stage.3

Completed successfully

Summary for salt.hli.ubc.ca_master
-------------
Succeeded: 59 (changed=45)
Failed: 0
-------------
Total states run: 59
Total run time: 1082.012 s
Monitor Change Config

Deepsea monitor - During

Deepsea completed
Watch as OSD’s Are Restarted

Watch ceph osd tree down
  • Progress through cluster 3-4 OSD’s at a time
Watch Ceph Status

Watch ceph status

|cluster:
| id: 11efbca1-b4b6-4305-8564-938abb320827
|Health: HEALTH_WARN
|3 osds down
|Reduced data availability: 2 pgs peering
|Degraded data redundancy: 10847331/732261990 objects degraded (1.81%), 182 pgs degraded
|services:
|mon: 3 daemons, quorum mon01,mon02,mon03
|mgr: mon02(active), stands: mon03, mon01
|osd: 168 osds: 165 up, 168 in
|data:
|pools: 2 pools, 4224 pgs
|objects: 244.00M objects, 228TiB
|usage: 690TiB used, 504TiB / 1.19TiB avail
|pgs: 0.816% pgs not active
|10847331/732261990 objects degraded (1.481%)
|3992 active+clean
|180 active+under sized+degraded
|23 peering
|20 stale+active+clean
|4 active+under sized
|2 activating+under sized
|2 active+recovery_wait+degraded
|1 activating
|io:
|client: 208MiB/s rd, 121MiB/s wr, 1.80kop/s rd, 3.44kop/s wr
|recovery: 7.59MiB/s, 7objects/s
Placement Group

- Pool from 4096 to 8192
- Steps of 1024
- Disable high load applications
- Disable backup