

Non disruptive migration of Ceph storage from community Hammer version to SUSE enterprise Storage 5.5

April 4, 2019

Hironobu Ishii, Fujitsu Limited,
ishii.hironobu@fujitsu.com

Michiko Nogi, Sr. Storage Technologist, SUSE
michiko.nogi@suse.com

- About us
- Introduction of Fujitsu
- Non disruptive migration
 - Project background
 - Migration overview
 - Migration steps
 - Conclusion

Hironobu Ishii, Linux Development Div., Fujitsu Japan

■ These days

- Director,
 - Customer support of SUSE products like SLES, SOC, SES (2016-now)
 - Helping hardware development team solve issues with SUSE products (2016-now)
 - Internal support of FUJITSU public cloud infrastructure (2015-2018)

■ In the past

- Director/Manager, customer support of Red Hat products (2004-2014)
- Linux device driver development (2002-2003)
- UNIX device driver development & customer support (1991-2002)
 - FC driver development / FC-HBA architect
 - NIC drivers, Protocol drivers development

Michiko Nogi, SUSE based in Japan

■ These days

■ Sr. Storage Technologist for SUSE APAC

- Work on with APAC customers & partners to support & implement SES solutions & systems

■ In the past

■ Principal Engineer for System Engineering in SanDisk (HGST, Western Digital)

- Work on Ceph & SDS solution and system to integrate All Flash and High density storage

■ Senior Systems Engineer for EMC

- Work as technical solution consultant to implement Unified storage for hundreds of customer sites

■ Systems Engineer for NetApp

- Work with largest partner to implement many telecom, manufacturing & other customers sites

■ Others

Fujitsu at a glance



■ Headquarters:
Tokyo, Japan

■ Established:
1935

■ President:
Tatsuya Tanaka

■ Principal Business Areas:
Technology Solutions
Ubiquitous Solutions
Device Solutions

■ Employees:
140,000 worldwide

■ Revenue:
4,098.3 billion yen

■ Operating profit:
182.4 billion yen

■ R&D Expenses:
158.6 billion yen
(Approx. 3.9% of Revenue)

■ Stock Exchange Listings:
Tokyo (Code:6702), Nagoya

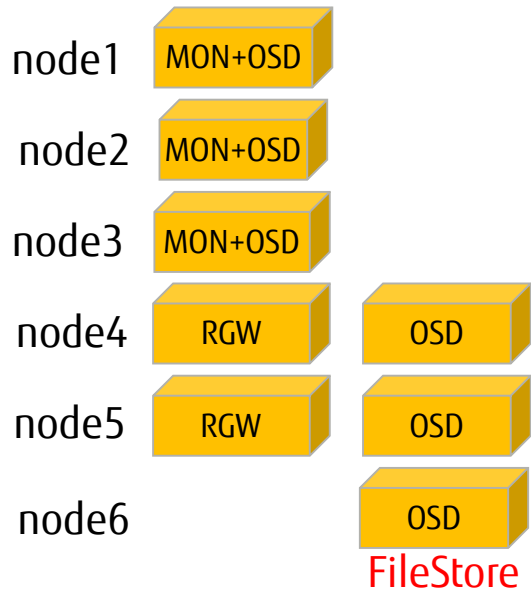
Non disruptive migration

Project back ground

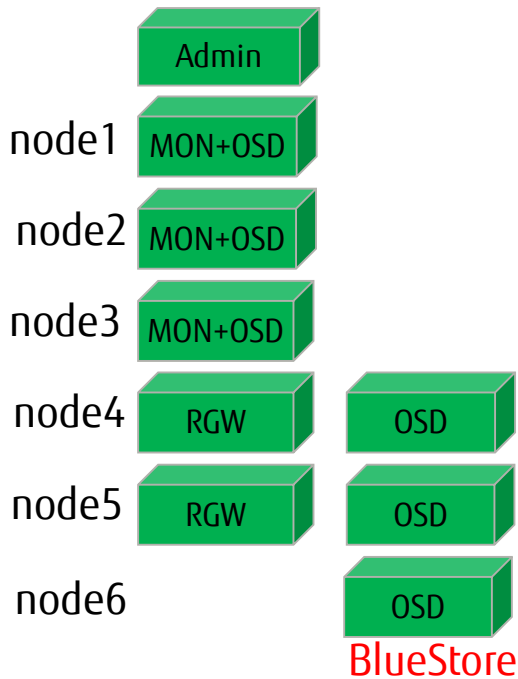
- Customer introduced community ceph with Firefly at first. Now using Hammer version.
- In initial phase, customer purchased InkTank pre-production service. Then InkTank was acquired by Red Hat. Then customer purchased support service from a vendor. But they felt that support quality of the vendor was not good. They stopped buying support service.
 - Customer uses various OSS, and various OSes as Ceph client.
 - The vendor's support was formalistic and did not respond seriously to customer issues.
- Some years later, customer had a hope to upgrade Ceph version, and to get support services from a distributor. Then customer contacted SUSE.
- SUSE and FUJITSU started POC for this customer.

Migration overview

Starting point



Goal



OSD : Object Storage Device
MON : Monitor daemon
RGW : RADOS Gateway

of OSD : 6
of MON : 3
of RGW : 2

Ubuntu 14.04 (Trusty)
Ceph Hammer

SLES12 SP3
SES 5.5 (Ceph Luminous)

■ Stated Goals

- Convert existing Community based Ceph installations to SES5
- Minimize disruption potential

■ Pre-requisites

- Each cluster is running the most current RHCS or community Ceph
- There is enough spare capacity on the cluster to accommodate a failed node

1. Drain the node and rebuild with BlueStore

1. Slower process
2. Ensures the OSDs are using BlueStore
3. Least risk because of NOT trying to import existing OSD

2. Maintain existing OSDs and do not migrate to BlueStore

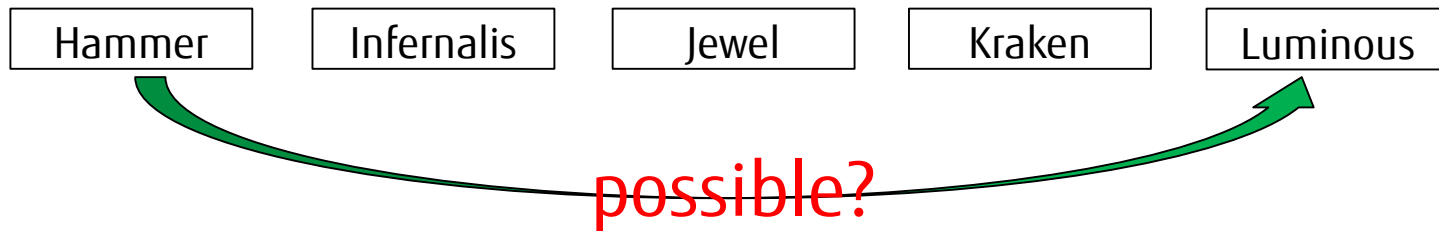
1. No data migration
2. Uncertainty about existing OSDs importing cleanly, which requires additional testing/validation

1. Upgrade Ceph Hammer to Jewel on Ubuntu 14.04 (FileStore)
2. Upgrade Ceph Jewel to Luminous on Ubuntu 14.04 (FileStore)
3. Replace the OS and install SES5.5 (FileStore)
4. Configure DeepSea/Salt and openATTIC
5. Convert OSD from FileStore to BlueStore using Salt.

Is it possible to jump from Hammer to Luminous?

■ Do we need two steps to upgrade from Hammer to Luminous?

1. Ubuntu Ceph **Hammer to Jewel**(FileStore)
2. Ubuntu Ceph **Jewel to Luminous**(FileStore)



■ We tried it anyway.

Jump from Hammer to Luminous was not possible

- When we start mon, osd daemons, mon daemon disappeared.

```
root@ceph-node1:~# ceph -v
ceph version 12.2.10 (177915764b752804194937482a39e95e0ca3de94) luminous (stable)
root@ceph-node1:~/ceph-luminous/download.ceph.com/debian-luminous/pool/main/c/ceph# ps aux | grep ceph
avahi    713  0.0  0.0 32360 1680 ?    S   Mar22  0:00 avahi-daemon: running [ceph-node1.local]
root     1284  0.0  0.8 308380 69764 ?    Ssl Mar22  7:55 /usr/bin/ceph-mon --cluster=ceph -i ceph-node1 -f
root     1534  0.2  1.2 826420 101132 ?    Ssl Mar22  28:59 /usr/bin/ceph-osd --cluster=ceph -i 0 -f
root     16430 0.0  0.0  8876  648 pts/4  S+  15:23  0:00 grep --color=auto ceph
root@ceph-node1:~# kill -15 1284 1534
root@ceph-node1:~# ps aux | grep ceph
avahi    713  0.0  0.0 32360 1680 ?    S   Mar22  0:00 avahi-daemon: running [ceph-node1.local]
root     1534  0.2  1.2 826420 101132 ?    Ssl Mar22  29:00 /usr/bin/ceph-osd --cluster=ceph -i 0 -f
root     16437 4.5  0.2 290652 19944 ?    Ssl 15:26  0:00 /usr/bin/ceph-mon --cluster=ceph -i ceph-node1 -f
root     16439 1.5  0.0 34164  7524 ?    Ss  15:26  0:00 /usr/bin/python /usr/sbin/ceph-create-keys --cluster=ceph -i ceph-node1
root     16444 2.5  0.2 248692 17372 ?    S   15:26  0:00 /usr/bin/python2.7 /usr/bin/ceph --cluster=ceph --admin-daemon=/var/run/ceph/ceph-
mon.ceph-node1.asok mon_status
root     16454 0.0  0.0  8876  648 pts/4  S+  15:26  0:00 grep --color=auto ceph
root@ceph-node1:~# ps aux | grep ceph
avahi    713  0.0  0.0 32360 1680 ?    S   Mar22  0:00 avahi-daemon: running [ceph-node1.local]
root     16651 0.0  0.0  8876  644 pts/4  S+  16:01  0:00 grep --color=auto ceph
root@ceph-node1:~#
```

Luminous mon cannot coexist with Hammer mons (1/2)

```
root@ceph-node1:~# /usr/bin/ceph-mon --id ceph-node1 -c /etc/ceph/ceph.conf
terminate called after throwing an instance of 'ceph::buffer::malformed_input'
  what(): buffer::malformed_input: void object_stat_sum_t::decode(ceph::buffer::list::iterator&) decode past end of struct encoding
*** Caught signal (Aborted) **
in thread 7f410afeef40 thread_name:ceph-mon
ceph version 12.2.10 (177915764b752804194937482a39e95e0ca3de94) luminous (stable)
1: ((()+0x931669) [0x7f410b945669]
2: ((()+0x10330) [0x7f410a3ac330]
3: (gsignal()+0x37) [0x7f4108997c37]
4: (abort()+0x148) [0x7f410899b028]
5: (_gnu_cxx::_verbose_terminate_handler()+0x155) [0x7f41092a6535]
:
11: (pg_stat_t::decode(ceph::buffer::list::iterator&)+0x1d5) [0x7f410b7716f5]
12: (PGMap::update_pg(pg_t, ceph::buffer::list&)+0xf4) [0x7f410b4c1d84]
13: (PGMonitor::read_pgmap_full()+0x161) [0x7f410b4913a1]
14: (PGMonitor::update_from_paxos(bool*)+0x699) [0x7f410b498d99]
15: (PaxosService::refresh(bool*)+0x1a3) [0x7f410b529593]
16: (Monitor::refresh_from_paxos(bool*)+0x183) [0x7f410b3f6ca3]
17: (Monitor::init_paxos()+0xfd) [0x7f410b3f707d]
18: (Monitor::preinit()+0xa7e) [0x7f410b3f7b3e]
19: (main()+0x3bfa) [0x7f410b3265ea]
20: (__libc_start_main()+0xf5) [0x7f4108982f45]
21: ((()+0x3b933e) [0x7f410b3cd33e]
2019-03-29 16:02:35.555503 7f410afeef40 -1 *** Caught signal (Aborted) **
in thread 7f410afeef40 thread_name:ceph-mon
```

Luminous mon cannot coexist with Hammer mons (2/2)



```
root@ceph-node1:~# ceph -s
cluster f721b1f2-9489-4053-b06e-79e6f3ded466
health HEALTH_WARN
  128 pgs stale
  128 pgs stuck stale
  1 mons down, quorum 1,2 ceph-node2,ceph-node3
monmap e1: 3 mons at {ceph-node1=10.19.3.128:6789/0,ceph-node2=10.19.3.129:6789/0,ceph-node3=10.19.3.130:6789/0}
election epoch 290, quorum 1,2 ceph-node2,ceph-node3
osdmap e91: 6 osds: 5 up, 5 in
pgmap v29468: 496 pgs, 10 pools, 4947 MB data, 1291 objects
  15061 MB used, 61683 MB / 76744 MB avail
    368 active+clean
    128 stale+active+clean
root@ceph-node1:~#
root@ceph-node1:~# ceph osd tree
ID WEIGHT TYPE NAME          UP/DOWN REWEIGHT PRIMARY-AFFINITY
-1 0.05997 root default
-2 0.00999 host ceph-node1
  0 0.00999  osd.0  down   0      1.00000
-3 0.00999 host ceph-node2
  1 0.00999  osd.1  up    1.00000  1.00000
```

- Engineer wondered which upgrade source should be used?
 - Ubuntu trusty provides only Ceph hammer version.
 - Ceph community provides source and prebuilt packages.

- Engineer decided following:
 - For Jewel upgrade, he will build the Ceph from source code.
 - For Luminous upgrade, he will use prebuilt packages which Ceph community provides.

1. Upgrade Ceph Hammer to Jewel on Ubuntu 14.04 (FileStore)
2. Upgrade Ceph Jewel to Luminous on Ubuntu 14.04 (FileStore)
3. Replace the OS and install SES5.5 (FileStore)
4. Configure DeepSea/Salt and openATTIC
5. Convert OSD from FileStore to BlueStore using Salt.

Hammer to Jewel (1)

- In this step, we tried to build Jewel from community latest source code.

1. Clone to Jewel source repository

```
root@ceph-node1:~# git clone --recursive https://github.com/ceph/ceph.git -b jewel
```

2. Install additional packages to build ceph

```
root@ceph-node1:~# cd ceph
```

```
root@ceph-node1:~/ceph# ./install-deps.sh
```

3. Install packaging tools

```
root@ceph-node1:~/ceph# apt-get install dpkg-dev
```

```
root@ceph-node1:~/ceph# dpkg-checkbuilddeps
```

Hammer to Jewel (2)

4. Create packages

```
root@ceph-node1:~/ceph# dpkg-buildpackage  
root@ceph-node1:~/ceph#
```

As a result, we find packages in parent directory.

```
root@ceph-node1:~/ceph# ls /root | grep deb  
ceph_10.2.11-1_amd64.deb          ceph-mon_10.2.11-1_amd64.deb  
libcephfs-java_10.2.11-1_all.deb  librbd1_10.2.11-1_amd64.deb  
python-rados_10.2.11-1_amd64.deb.....
```

Hammer to Jewel (3)

- Install Jewel packages on all servers. Here, I described ceph-node2 case as an example.

1. Copy Jewel packages into /root/jewel-deb of ceph-node2

```
root@ceph-node2:~# scp -r root@ceph-node1:/root/*deb ./jewel-deb
```

2. Install Jewel packages

```
root@ceph-node2:~/jewel-deb# dpkg -i --force-overwrite *deb  
root@ceph-node2:~/jewel-deb# apt-get -f install
```

3. Restart daemons (Restarting order is mon, osd then rgw)

- Restarting mon daemons on all nodes which mons are running.

```
# kill -15 <mon process id>  
# /usr/bin/ceph-mon --id <mon node name> -c /etc/ceph/ceph.conf
```

- Restarting osd daemons on all nodes which osd resides.

```
# kill -15 <osd process id>  
# /usr/bin/ceph-osd --id <osd instance id>
```

- Restarting rgw service on all nodes which rgw resides.

```
# kill -15 <rgw process id>  
# /usr/bin/radosgw -f --cluster ceph --name client.rgw.<rgw node name>
```

Hammer to Jewel (5)

- Finishing migration

1. Check the cluster status

```
root@ceph-node1:~# /usr/bin/ceph -s
  cluster 5d4c5d20-a0c5-483e-9648-e8799af940fd
  health HEALTH_WARN
:
```

In case of migration from Hammer to Jewel, you will see always HEALTH_WARN. Therefore, please perform 3 steps on the next slide.

Hammer to Jewel (6)

2. Finishing tasks

Update legacy tunable parameters to optimal value for Jewel.

```
root@ceph-node1:~# /usr/bin/ceph osd crush tunables optimal  
adjusted tunables profile to optimal
```

This flag tells mons about end of migration to Jewel.

```
root@ceph-node1:~# /usr/bin/ceph osd set require_jewel_osds  
set require_jewel_osds
```

Hammer to Jewel (7)

Set the “sortbitwise” flag to enable the new internal object sort order.

Enabling “sortbitwise” changes in the internal sorting algorithm. And it is exposed to end-user only because of legacy (pre- jewel) compatibility.

```
root@ceph-node1:~# /usr/bin/ceph osd set sortbitwise  
set sortbitwise
```


Hammer to Jewel (8)

3. Check the cluster version

```
root@ceph-node1:~# /usr/bin/ceph -v  
ceph version 10.2.11-7-g3b165d0 (3b165d04be802df246f40f9042168897be279929)
```

4. Confirm you can access objects

```
root@ceph-node7:~# s3cmd la  
2018-12-07 04:21    5145  s3://bucket_test/file1  
2018-12-07 04:27    5145  s3://bucket_test/file2  
2018-12-07 04:29    5145  s3://bucket_test/file3
```

Migration from Hammer to Jewel has finished.

1. Upgrade Ceph Hammer to Jewel on Ubuntu 14.04 (FileStore)
2. Upgrade Ceph Jewel to Luminous on Ubuntu 14.04 (FileStore)
3. Replace the OS and install SES5.5 (FileStore)
4. Configure DeepSea/Salt and openATTIC
5. Convert OSD from FileStore to BlueStore using Salt.

Jewel to Luminous (1)

- Download the Luminous packages for Debian on node1. Then distribute them to other nodes. Then install them on all nodes.
 - <http://download.ceph.com/debian-luminous/pool/main/c/ceph/>

```
# wget -r --no-parent -A "*_12.2.10*trusty*.deb" -R "*-dbg_*" http://download.ceph.com  
/debian-luminous/pool/main/c/ceph/  
# dpkg -i download.ceph.com/debian-luminous/pool/main/c/ceph/*
```

3. Restart daemons (Restarting order is mon, osd then rgw)

- Restarting mon daemons on all nodes which mons are running.

```
# kill -15 <mon process id>  
# /usr/bin/ceph-mon --id <mon node name> -c /etc/ceph/ceph.conf
```

- Restarting osd daemons on all nodes which osd resides.

```
# kill -15 <osd process id>  
# /usr/bin/ceph-osd --id <osd instance id>
```

- Restarting rgw service on all nodes which rgw resides.

```
# kill -15 <rgw process id>  
# /usr/bin/radosgw -f --cluster ceph --name client.rgw.<rgw node name>
```

Jewel to Luminous (3)

■ Finishing migration

1. Check the cluster status

```
root@ceph-node1:~# /usr/bin/ceph -s
cluster:
  id:   5d4c5d20-a0c5-483e-9648-e8799af940fd
health: HEALTH_WARN
all OSDs are running luminous or later but require_osd_release < luminous
      no active mgr
:
```

In case of migration from Jewel to Luminous, you will see always HEALTH_WARN. Therefore, please perform steps on the next slide.

Jewel to Luminous (4)

2. Finishing tasks

This flag tells mons about end of migration to Luminous.

```
root@ceph-node1:~# /usr/bin/ceph osd require-osd-release luminous
recovery_deletes is set
```

Start a MGR daemons on every node MON is running. MGR is newly introduced daemon in Luminous, and it is a mandatory daemon in Luminous or later.

```
# /usr/bin/ceph --cluster ceph auth get-or-create mgr.admin mon 'allow profile mgr' osd
'allow *' mds 'allow *'
# mkdir -p /var/lib/ceph/mgr/ceph-admin/
# /usr/bin/ceph auth get mgr.admin -o /var/lib/ceph/mgr/ceph-admin/keyring
# /usr/bin/ceph-mgr -i admin
```

3. Check the cluster version

```
root@ceph-node1:~# ceph versions
{
  "mon": {
    "ceph version 12.2.10 (177915764b752804194937482a39e95e0ca3de94) luminous (stable)": 3
  },
  "mgr": {
    "ceph version 12.2.10 (177915764b752804194937482a39e95e0ca3de94) luminous (stable)": 1
  },
  "osd": {
    "ceph version 12.2.10 (177915764b752804194937482a39e95e0ca3de94) luminous (stable)": 6
  },
  "mds": {},
  "rgw": {
    "ceph version 12.2.10 (177915764b752804194937482a39e95e0ca3de94) luminous (stable)": 2
  },
  "overall": {
    "ceph version 12.2.10 (177915764b752804194937482a39e95e0ca3de94) luminous (stable)": 12
  }
}
```

Jewel to Luminous (5)

4. Confirm you can access objects

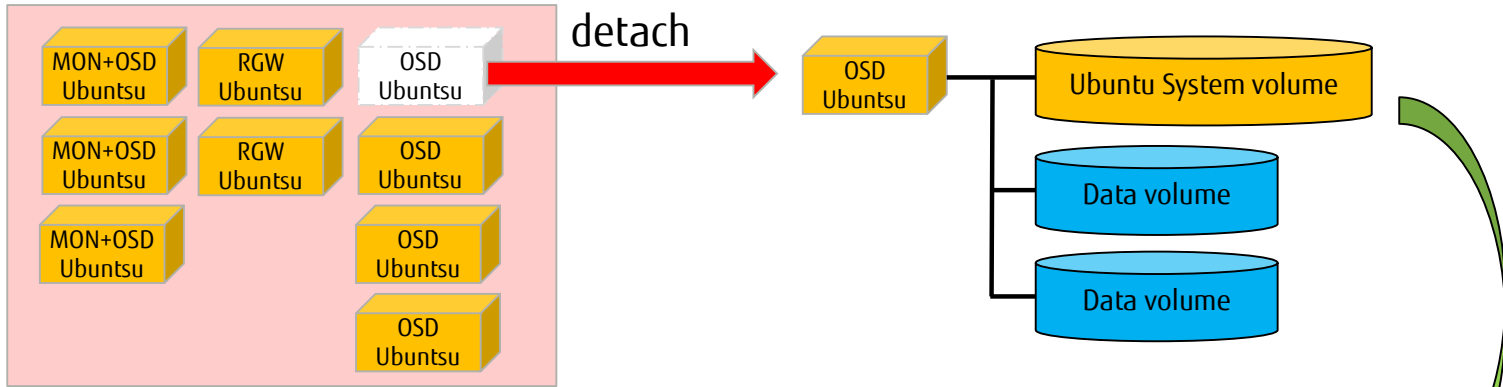
```
root@ceph-node7:~# s3cmd la
2018-12-07 04:21   5145  s3://bucket_test/file1
2018-12-07 04:27   5145  s3://bucket_test/file2
2018-12-07 04:29   5145  s3://bucket_test/file3
```

Migration from Jewel to Luminous has finished.

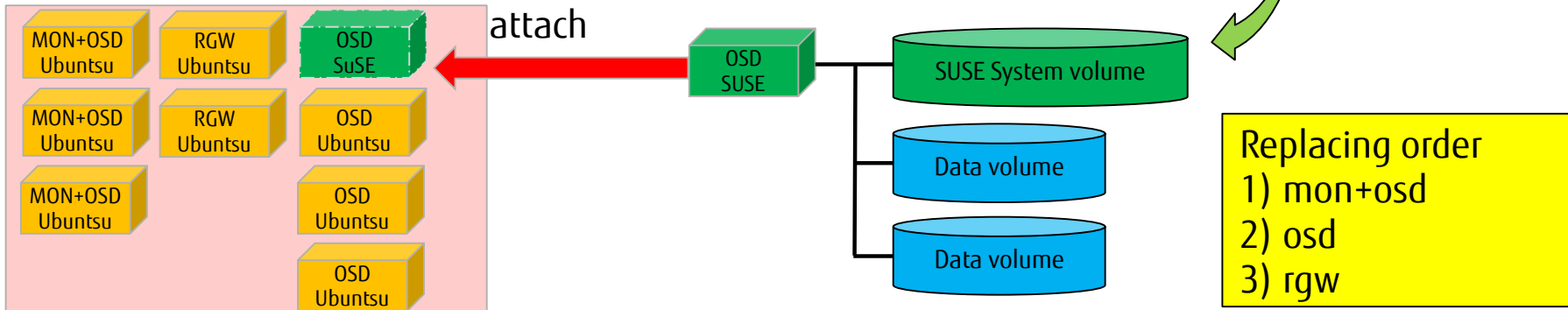
1. Upgrade Ceph Hammer to Jewel on Ubuntu 14.04 (FileStore)
2. Upgrade Ceph Jewel to Luminous on Ubuntu 14.04 (FileStore)
3. **Replace the OS and install SES5.5 (FileStore)**
4. Configure DeepSea/Salt and openATTIC
5. Convert OSD from FileStore to BlueStore using Salt.

Replacing Ubuntu with SES (1)

Detach



Attach



■ Replacing OS

- Backup the Ceph config files (/etc/ceph/*, /var/lib/ceph/*) on another server.
- Shutdown Ubuntu.
- Start installation of SLES12 SP3.
- Install SES5 packages.
- Restore back-upped Ceph config files. (/etc/ceph/*, /var/lib/ceph/*)
- Start Ceph daemons
systemctl start ceph.target

■ Tips

- Be careful not to use data volumes (OSD volumes) while installing SLES.
 - If you destroy the existing data on OSD volumes, you will need additional time to recover data of OSD volumes from other OSD nodes.
- Before starting shutdown of every node, check the cluster status is HEALTH_OK with “ceph -s” command.

1. Upgrade Ceph Hammer to Jewel on Ubuntu 14.04 (FileStore)
2. Upgrade Ceph Jewel to Luminous on Ubuntu 14.04 (FileStore)
3. Replace the OS and install SES5.5 (FileStore)
4. **Configure DeepSea/Salt and openATTIC**
5. Convert OSD from FileStore to BlueStore using Salt.

■ Deploy DeepSea

- Add DeepSea/Salt packages on admin node

```
# zypper install deepsea
```

- Execute the DeepSea engulf

- The engulf processes attempts to convert a non-DeepSea cluster into a DeepSea configured cluster without altering the existing cluster.

```
# salt-run populate.engulf_existing_cluster
```

- Run stage0 through stage4 with DeepSea command.

```
# deepsea stage run ceph.stage.0
```

```
:
```

```
# deepsea stage run ceph.stage.4
```

DeepSea and openATTIC (2)

- Stage0 upgrades to latest packages
- Stage1 collects various Ceph configuration information.
- Stage2 creates configuration file for Salt which defines Ceph cluster information.
- Stage3 deploys OSD and Mon daemons
- Stage4 deploys additional daemons like RADOS GW

DeepSea and openATTIC (3)

- openATTIC will be enabled in the process of stage0 through stage4.
- Check whether openATTIC is working or not with browser.

■ Tips

- DeepSea supports xfs FileStore but not ext4 FileStore. If the old cluster uses non-xfs FileStore, **DeepSea will destroy the data in the OSDs in stage3**. To workaround this you need to comment out the following if statement:

- /srv/salt/_modules/osd.py:

```
# if osdc.is_partition('osd', config.device, _partition) and _fsck(config.device, _partition):
```

- The step of "salt-run populate.engulf_existing_cluster" needs that Ceph services are running under system.

1. Upgrade Ceph Hammer to Jewel on Ubuntu 14.04 (FileStore)
2. Upgrade Ceph Jewel to Luminous on Ubuntu 14.04 (FileStore)
3. Replace the OS and install SES5.5 (FileStore)
4. Configure DeepSea/Salt and openATTIC
5. Convert OSD from FileStore to BlueStore using Salt.

Conversion from FileStore to BlueStore (1)

■ SUSE document is good starting point

- https://www.suse.com/documentation/suse-enterprise-storage-5/singlehtml/book_storage_deployment/book_storage_deployment.html#filestore2bluestore

■ Steps

We need to unlock the safety-lock to use Salt OSD conversion feature.

```
# salt-run disengage.safety
```

Migrate hardware profiles:

```
# salt-run state.orch ceph.migrate.policy
```

To migrate OSDs one at a time, run:

```
# salt-run state.orch ceph.migrate.osds
```

■ Tips

- You can confirm the conversion result with OSD meta data.

```
# ceph osd metadata 1 | grep objectstore  
"osd_objectstore": "bluestore"
```

- Duration of "ceph.migrate.osds" step depends on the amount of data in the OSD. Salt will timeout in 1 hour and command might end with error. But conversion of OSD continues. You can restart the same step after the cluster status become HEALTH_OK.

1. Upgrade Ceph Hammer to Jewel on Ubuntu 14.04 (FileStore)
2. Upgrade Ceph Jewel to Luminous on Ubuntu 14.04 (FileStore)
3. Replace the OS and install SES5.5 (FileStore)
4. Configure DeepSea/Salt and openATTIC
5. Convert OSD from FileStore to BlueStore using Salt.

Migration Finished

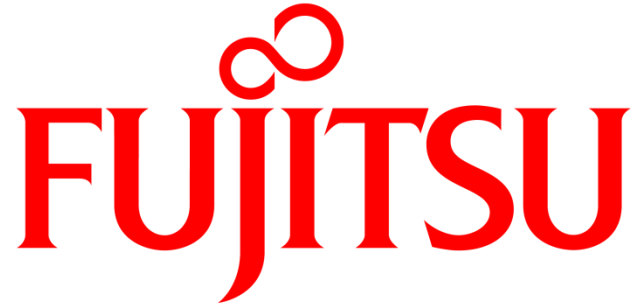
- We can migrate the community based Ceph cluster to SES with non disruptive way.
- We needed (2 months*1 person) to establish this migration procedure.
- When we follow the established procedure, it will take 2 days + X days
 - X depends on amount of existing data. We need further study to estimate this time.
- SES provides very good management tools (DeepSea/Salt) to ease the management of Ceph cluster.

- Please contact FUJITSU if you need help for Ceph migration.

Special Thanks



Kouya Shimura, Shunichi Sagawa,
Shingo Iwakura, Yuji Morita and
Mikio Ito



shaping tomorrow with you

Our products and services

Technology Solutions

Services

Solution / SI

- Systems Integration (system construction, business applications)
- Consulting
- Front-end Technologies (ATMs, POS systems etc.)

Infrastructure Services

- Outsourcing services (datacenters, ICT operation/management, SaaS, application operation/management, business process outsourcing, etc.)
- Cloud services (IaaS, PaaS, SaaS, etc.)
- Network services (business networks, distribution of Internet/mobile content)
- System support services (maintenance and surveillance services for information systems and networks)
- Security solutions (installation of information systems and networks)

Systems platform

System Products

- Full Range of Servers (mainframe, UNIX, mission-critical x86 servers and other x86 servers)
- Storage Systems
- Software (operating system, middleware)

Network Products

- Network Management Systems
- Optical Transmission Systems
- Mobile Phone Base Stations

Our products and services

Ubiquitous Product Solutions

- PCs

Device solutions

- LSI Devices
- Semiconductor Packages
- Batteries
- Electromechanical Components (relays, connectors, etc.)
- Optical Transceiver Modules
- Printed Circuit Boards

Revenue by sector – FY2017



(Billions of yen)

Technology Solutions	3,052.7		
Services	2,598.3	■ Solutions / SI	■ Infrastructure Services
System Platform	454.3	■ System Products	■ Network Products
Ubiquitous Solutions	663.9	■ PCs	
Device Solutions	560.0	■ LSI	■ Electronic Components
Others	-178.2		
Total	4,098.3		

Note: Consolidated Revenue by Business Segment, Including Intersegment Revenue