Deploying Ceph clusters with Salt

FOSDEM 17 – Brussels – UA2.114 (Baudoux)

Jan Fajerski
Software Engineer
jfajerski@suse.com

Saltstack

“Software to automate the management and configuration of any infrastructure or application at scale.”

- **Configuration management and remote execution**
- **Based on Python, Jinja and ZeroMQ**
- **Master applies state to Minions**
- **States define dependencies as DAG**

Ceph

“Ceph is a unified, distributed storage system designed for excellent performance, reliability and scalability.”

- **Provides block, object and file system storage**
- **Scalable, fault-tolerant and self healing**
- **Designed to run on commodity hardware**

DeepSea

- **Collection of Salt files for Ceph cluster creation and management**
- **Goals:**
  - Start after OS installed and salt setup
  - Automate hardware discovery
  - Find problems before they are deployed
  - Manage complete cluster life cycle
- **Open source – GPLv3**
- **Status: discovery, deployment and basic management works**

Bug reports and contributions welcome
https://github.com/SUSE/DeepSea

DeepSea – basic workflow

- **Install OS, salt, accept minion key, install DeepSea**
- **Run DeepSea stages:**
  - **0 – Preparation**: sync salt, update kernel
  - **1 – Discovery**: query minions hardware & network, write config fragments
  - Manual step: create your policy.cfg which governs your cluster topology
  - **2 – Configuration**: assemble configuration and push to minions
  - **3 – Deployment**: install ceph, deploy configuration, start Ceph
  - **4 – Services**: start extra ceph services: MDS, rgw, iscsi
- **Can be much more complex**
- **Stage 5 implements removal of components**

DeepSea

- **Stages are orchestration files**
  salt-run state.orch ceph.stage.*n*
- **These call salt states with correct targeting based on role assignments**
- **States can be called manually**

- **Common pattern – init.sls redirection:**

```
include:
    - .{{ salt['pillar.get']('mon_init', 'default') }}
```

- **Requires a minion on the master node**

Let’s try it

Demo Cluster

- **Cluster of 10 kvm machines**
- **1 GB RAM, 1 CPU, 2 network interfaces**
- **OSD nodes in two flavours**
  - 4 OSD nodes with 5 x 5 GB drives
  - 2 OSD nodes with 1 x 1 GB drive + 5 x 5 GB drives
  - 32 drives overall
- **Convieniently named:**
  - mon[1,2,3]
  - data[1 – 6]
  - admin

Stage 0 - Preparation

- make sure all minions are in the same state
- is still rather SUSE specific – being worked on
- can be skipped
- Sync salt, add repos to zypper, install a few packages, updates
- Might reboot your minions...including the master

Stage 1 - Discovery

- **Query minions for storage hardware and network connections**
- **Write config fragments to /src/pillar/ceph/proposals/**
  - Cluster assignment
  - Role assignment
  - Some ceph configuration
  - Storage profiles
- **~ per fragment and minion one file**

```
root@admin /srv/pillar/ceph/proposals
# l
total 0
drwxr-xr-x 1 salt salt 14 Feb  1 12:10 cluster-ceph
drwxr-xr-x 1 salt salt 14 Feb  1 12:10 cluster-unassigned
drwxr-xr-x 1 salt salt 10 Feb  1 12:10 config
drwxr-xr-x 1 salt salt 24 Feb  1 12:10 profile-1Disk1GB-5Disk5
GB-1
drwxr-xr-x 1 salt salt 24 Feb  1 12:10 profile-5Disk5GB-1
drwxr-xr-x 1 salt salt 14 Feb  1 12:10 role-admin
drwxr-xr-x 1 salt salt 14 Feb  1 12:10 role-client-cephfs
drwxr-xr-x 1 salt salt 14 Feb  1 12:10 role-client-iscsi
drwxr-xr-x 1 salt salt 14 Feb  1 12:10 role-client-nfs
drwxr-xr-x 1 salt salt 14 Feb  1 12:10 role-client-radosgw
drwxr-xr-x 1 salt salt 24 Feb  1 12:10 role-ganesha
drwxr-xr-x 1 salt salt 24 Feb  1 12:10 role-igw
drwxr-xr-x 1 salt salt 14 Feb  1 12:10 role-master
drwxr-xr-x 1 salt salt 14 Feb  1 12:10 role-mds
drwxr-xr-x 1 salt salt 24 Feb  1 12:10 role-mon
drwxr-xr-x 1 salt salt 14 Feb  1 12:10 role-rgw
drwxr-xr-x 1 salt salt  0 Feb  1 12:10 role-storage
root@admin /srv/pillar/ceph/proposals
#
```

```
root@admin /srv/pillar/ceph/proposals
# cat role-mon/cluster/mon1.sls
roles:
- mon
root@admin /srv/pillar/ceph/proposals
# cat role-mon/stack/default/ceph/minions/mon1.yml
public_address: 192.168.100.53
root@admin /srv/pillar/ceph/proposals
#
```

```
root@admin /srv/pillar/ceph/proposals
# cat profile-5Disk5GB-1/cluster/data3.sls
roles:
- storage
root@admin /srv/pillar/ceph/proposals
# cat profile-5Disk5GB-1/stack/default/ceph/minions/data3.yml
storage:
  data+journals: []
  osds:
  - /dev/vdf
  - /dev/vdd
  - /dev/vdb
  - /dev/vde
  - /dev/vdc
root@admin /srv/pillar/ceph/proposals
#
```

Policy.cfg

- Central configuration file
- Choose which config fragments to use
- Supports globs, list slicing and regex
- Order is important – options can be overwritten

```
root@admin /srv/pillar/ceph/proposals
# cat policy.cfg
# Cluster assignment
cluster-ceph/cluster/*.sls
# Hardware Profile
profile-1Disk1GB-5Disk5GB-1/cluster/data*.sls
profile-1Disk1GB-5Disk5GB-1/stack/default/ceph/minions/data*.y
ml
profile-5Disk5GB-1/cluster/data*.sls
profile-5Disk5GB-1/stack/default/ceph/minions/data*.yml
# Common configuration
config/stack/default/global.yml
config/stack/default/ceph/cluster.yml
# Role assignment
role-master/cluster/admin*.sls
role-admin/cluster/mon*.sls
role-mon/cluster/mon*.sls
role-mon/stack/default/ceph/minions/mon*.yml
root@admin /srv/pillar/ceph/proposals
#
```

```
root@admin /srv/pillar/ceph/proposals
# cat policy.cfg
# Cluster assignment
cluster-ceph/cluster/*.sls
# Hardware Profile
profile-1Disk1GB-5Disk5GB-1/cluster/data*.sls
profile-1Disk1GB-5Disk5GB-1/stack/default/ceph/minions/data*.y
ml
profile-5Disk5GB-1/cluster/data*.sls
profile-5Disk5GB-1/stack/default/ceph/minions/data*.yml
# Common configuration
config/stack/default/global.yml
config/stack/default/ceph/cluster.yml
# Role assignment
role-master/cluster/admin*.sls
role-admin/cluster/mon*.sls
role-mon/cluster/mon*.sls
role-mon/stack/default/ceph/minions/mon*.yml
root@admin /srv/pillar/ceph/proposals
#
```

```
root@admin /srv/pillar/ceph/proposals
# cat policy.cfg
# Cluster assignment
cluster-ceph/cluster/*.sls
# Hardware Profile
profile-1Disk1GB-5Disk5GB-1/cluster/data*.sls
profile-1Disk1GB-5Disk5GB-1/stack/default/ceph/minions/data*.y
ml
profile-5Disk5GB-1/cluster/data*.sls
profile-5Disk5GB-1/stack/default/ceph/minions/data*.yml
# Common configuration
config/stack/default/global.yml
config/stack/default/ceph/cluster.yml
# Role assignment
role-master/cluster/admin*.sls
role-admin/cluster/mon*.sls
role-mon/cluster/mon*.sls
role-mon/stack/default/ceph/minions/mon*.yml
root@admin /srv/pillar/ceph/proposals
#
```

```
root@admin /srv/pillar/ceph/proposals
# cat policy.cfg
# Cluster assignment
cluster-ceph/cluster/*.sls
# Hardware Profile
profile-1Disk1GB-5Disk5GB-1/cluster/data*.sls
profile-1Disk1GB-5Disk5GB-1/stack/default/ceph/minions/data*.y
ml
profile-5Disk5GB-1/cluster/data*.sls
profile-5Disk5GB-1/stack/default/ceph/minions/data*.yml
# Common configuration
config/stack/default/global.yml
config/stack/default/ceph/cluster.yml
# Role assignment
role-master/cluster/admin*.sls
role-admin/cluster/mon*.sls
role-mon/cluster/mon*.sls
role-mon/stack/default/ceph/minions/mon*.yml
root@admin /srv/pillar/ceph/proposals
#
```

```
root@admin /srv/pillar/ceph/proposals
# cat policy.cfg
# Cluster assignment
cluster-ceph/cluster/*.sls
# Hardware Profile
profile-1Disk1GB-5Disk5GB-1/cluster/data*.sls
profile-1Disk1GB-5Disk5GB-1/stack/default/ceph/minions/data*.y
ml
profile-5Disk5GB-1/cluster/data*.sls
profile-5Disk5GB-1/stack/default/ceph/minions/data*.yml
# Common configuration
config/stack/default/global.yml
config/stack/default/ceph/cluster.yml
# Role assignment
role-master/cluster/admin*.sls
role-admin/cluster/mon*.sls
role-mon/cluster/mon*.sls
role-mon/stack/default/ceph/minions/mon*.yml
root@admin /srv/pillar/ceph/proposals
#
```

```
root@admin /srv/pillar/ceph/proposals
# cat policy.cfg
# Cluster assignment
cluster-ceph/cluster/*.sls
# Hardware Profile
profile-1Disk1GB-5Disk5GB-2/cluster/data*.sls
profile-1Disk1GB-5Disk5GB-2/stack/default/ceph/minions/data*.y
ml
profile-5Disk5GB-1/cluster/data[3,4,5].sls
profile-5Disk5GB-1/stack/default/ceph/minions/data[3,4,5].yml
# Common configuration
config/stack/default/global.yml
config/stack/default/ceph/cluster.yml
# Role assignment
role-master/cluster/admin*.sls slice=[2:5]
role-admin/cluster/mon*.sls re=.*[^N].*X1.*
role-mon/cluster/mon*.sls
role-mon/stack/default/ceph/minions/mon*.yml
root@admin /srv/pillar/ceph/proposals
#
```

```
root@admin /srv/pillar/ceph/proposals
# cat policy.cfg
# Cluster assignment
cluster-ceph/cluster/*.sls
# Hardware Profile
profile-1Disk1GB-5Disk5GB-2/cluster/data*.sls
profile-1Disk1GB-5Disk5GB-2/stack/default/ceph/minions/data*.y
ml
profile-5Disk5GB-1/cluster/data[3,4,5].sls
profile-5Disk5GB-1/stack/default/ceph/minions/data[3,4,5].yml
# Common configuration
config/stack/default/global.yml
config/stack/default/ceph/cluster.yml
# Role assignment
role-master/cluster/admin*.sls slice=[2:5]
role-admin/cluster/mon*.sls re=.*[^N].*X1.*
role-mon/cluster/mon*.sls
role-mon/stack/default/ceph/minions/mon*.yml
root@admin /srv/pillar/ceph/proposals
#
```

Stage 2 - Configuration

- **Pulls in config fragments as specified in policy.cfg**
- **Based on stack.py – merges yaml files** (included since 2016.3)
- **Option to customize specific options**
  - /srv/pillar/ceph/stack/default – default created by DeepSea
  - /srv/pillar/ceph/stack – custom options for specific minions
- **Check config with** salt *$minion* pillar.items

Stage 3 - deployment

- Validates setup
- Authenticate keyrings
- Install ceph
- Creates MON cluster
- Creates OSDs
- Creates pool(s)

```
Summary for admin_master
-------------
Succeeded: 12 (changed=4)
Failed:     0
-------------
Total states run:     12
Total run time:    8.366 s
firewall                 : disabled
fsid                     : valid
public_network           : valid
public_interface         : valid
cluster_network          : valid
cluster_interface        : valid
monitors                 : valid
storage                  : valid
master_role              : valid
mon_host                 : valid
mon_initial_members      : valid
time_server              : valid
fqdn                     : valid
retcode:
    0
admin_master:
  Name: packages - Function: salt.state - Result: Clean Started: - 19:43:05.330075 Duration: 2207.573 ms
  Name: configuration check - Function: salt.state - Result: Clean Started: - 19:43:07.537886 Duration: 343.765 ms
  Name: configuration - Function: salt.state - Result: Clean Started: - 19:43:07.881878 Duration: 985.988 ms
  Name: admin - Function: salt.state - Result: Changed Started: - 19:43:08.868080 Duration: 613.934 ms
  Name: monitors - Function: salt.state - Result: Changed Started: - 19:43:09.482265 Duration: 3208.163 ms
  Name: osd auth - Function: salt.state - Result: Changed Started: - 19:43:12.690684 Duration: 12122.704 ms
  Name: storage - Function: salt.state - Result: Changed Started: - 19:43:24.813710 Duration: 225978.543 ms
  Name: pools - Function: salt.state - Result: Changed Started: - 19:47:10.792514 Duration: 25103.156 ms

Summary for admin_master
------------
Succeeded: 8 (changed=5)
Failed:    0
------------
Total states run:     8
Total run time: 270.564 s
salt-run state.orch ceph.stage.3  6.79s user 0.38s system 2% cpu 4:33.49 total
root@admin /srv/pillar/ceph/proposals
#
```

```
root@admin /srv/pillar/ceph/proposals
# ceph -s
    cluster 1a87e5a2-dff0-33a9-a50d-18f3299006a8
     health HEALTH_WARN
            too few PGs per OSD (6 < min 30)
     monmap e1: 3 mons at {mon1=192.168.100.53:6789/0,mon2=192
.168.100.232:6789/0,mon3=192.168.100.225:6789/0}
            election epoch 6, quorum 0,1,2 mon1,mon3,mon2
     osdmap e87: 32 osds: 32 up, 32 in
            flags sortbitwise,require_jewel_osds,require_krake
n_osds
      pgmap v152: 64 pgs, 1 pools, 16 bytes data, 3 objects
            1121 MB used, 135 GB / 136 GB avail
                  64 active+clean
root@admin /srv/pillar/ceph/proposals
#
```

Customize a deployment

Choose profile

- Choose profile with osd journal on separate partition
- DeepSea will generate this for SSD/HDD hardware
- Can also be hand-crafted

```
root@admin /srv/pillar/ceph/proposals
# cat profile-1Disk1GB-5Disk5GB-2/stack/default/ceph/mi
nions/data1.yml
storage:
  data+journals:
    - /dev/vdc: /dev/vdb
    - /dev/vdd: /dev/vdb
    - /dev/vde: /dev/vdb
    - /dev/vdf: /dev/vdb
    - /dev/vdg: /dev/vdb
  osds: []
root@admin /srv/pillar/ceph/proposals
#
```

On real hardware:

```
jan@jf_suse_laptop ~
% ssh -i ~/.ssh/id_rsa_blueshark root@blueshark-1.suse.de 'cat /s
rv/pillar/ceph/proposals/profile-2Intel745GB-6INTEL372GB-2/stack/
default/ceph/minions/blueshark-3.suse.de.yml'
storage:
  data+journals:
  - /dev/disk/by-id/ata-INTEL_SSDSC2BA400G4_BTHV608300VE400NGN: /
dev/disk/by-id/nvme-SNVMe_INTEL_SSDPEDMD80CVFT5505000D800CGN
  - /dev/disk/by-id/ata-INTEL_SSDSC2BA400G4_BTHV6082036S400NGN: /
dev/disk/by-id/nvme-SNVMe_INTEL_SSDPEDMD80CVFT5505000D800CGN
  - /dev/disk/by-id/ata-INTEL_SSDSC2BA400G4_BTHV608204YX400NGN: /
dev/disk/by-id/nvme-SNVMe_INTEL_SSDPEDMD80CVFT5505000D800CGN
  - /dev/disk/by-id/ata-INTEL_SSDSC2BA400G4_BTHV608203EX400NGN: /
dev/disk/by-id/nvme-SNVMe_INTEL_SSDPEDMD80CVFT5505000D800CGN
  - /dev/disk/by-id/ata-INTEL_SSDSC2BA400G4_BTHV608300WN400NGN: /
dev/disk/by-id/nvme-SNVMe_INTEL_SSDPEDMD80CVFT5505000D800CGN
  - /dev/disk/by-id/ata-INTEL_SSDSC2BA400G4_BTHV608203ET400NGN: /
dev/disk/by-id/nvme-SNVMe_INTEL_SSDPEDMD80CVFT5505000D800CGN
  osds:
  - /dev/disk/by-id/nvme-SNVMe_INTEL_SSDPEDMD80CVFT60750012800CGN
jan@jf_suse_laptop ~
%
```

Behavior customisation

- Demo VM setup is not a typical Ceph deployment
- Behavior is easily altered – redirection pattern
- Add custom method for OSD deployment
- Configure desired method in the pillar

```
root@admin /srv/salt/ceph/osd
# l
total 12
drwxr-xr-x 1 root root   38 Feb  1 11:08 auth
drwx------ 1 salt salt   34 Feb  1 18:24 cache
-rw-r--r-- 1 root root 1192 Jan 25 14:43 default.sls
drwxr-xr-x 1 root root   20 Feb  1 11:08 files
-rw-r--r-- 1 root root 1270 Feb  1 18:54 fosdem.sls
-rw-r--r-- 1 root root   65 Jan 25 14:43 init.sls
drwxr-xr-x 1 root root   38 Feb  1 11:08 key
drwxr-xr-x 1 root root   38 Feb  1 11:08 keyring
drwxr-xr-x 1 root root   58 Feb  2 12:46 partition
drwxr-xr-x 1 root root   38 Feb  1 11:08 restart
drwxr-xr-x 1 root root   38 Feb  1 11:08 scheduler
root@admin /srv/salt/ceph/osd
#
```

```
root@admin /srv/salt/ceph/osd
# cat init.sls


include:
  - .{{ salt['pillar.get']('osd_init', 'default') }}

root@admin /srv/salt/ceph/osd
#
```

```
root@admin /srv/salt/ceph/osd
# vim /srv/pillar/ceph/stack/global.yml
root@admin /srv/salt/ceph/osd
# cat /srv/pillar/ceph/stack/global.yml
# /srv/pillar/ceph/stack/global.yml
# Overwrites configuration in /srv/pillar/ceph/stack/defaul
t/global.yml

osd_init: fosdem
osd_partition: fosdem
root@admin /srv/salt/ceph/osd
#
```

```
root@admin /srv/salt/ceph/osd
# ceph -s
    cluster 1a87e5a2-dff0-33a9-a50d-18f3299006a8
     health HEALTH_WARN
            too few PGs per OSD (6 < min 30)
     monmap e1: 3 mons at {mon1=192.168.100.53:6789/0,mon2=192
.168.100.232:6789/0,mon3=192.168.100.225:6789/0}
            election epoch 4, quorum 0,1,2 mon1,mon3,mon2
     osdmap e68: 30 osds: 30 up, 30 in
            flags sortbitwise,require_jewel_osds,require_krake
n_osds
      pgmap v99: 64 pgs, 1 pools, 16 bytes data, 3 objects
            1039 MB used, 130 GB / 131 GB avail
                  64 active+clean
  client io 445 B/s rd, 0 B/s wr, 0 op/s rd, 0 op/s wr
root@admin /srv/salt/ceph/osd
#
```

Beyond deployment

Stage 4 -services

- **Add additional service**
  - MDS and cephfs
  - ISCSI
  - Rados gateway
  - NFS Ganesha
  - Client nodes

Stage 5 - removal

- Nodes will eventually be decommissioned
- Remove minion from policy.cfg
- Run stages 2, [3, 4] and 5

Thank you! Questions?