Advanced Tuning and Operation guide for Block Storage using Ceph

netmarble

Who’s Here

John Han (sjhan@netmarble.com)

Netmarble

Jaesang Lee (jaesang_lee@sk.com)

Byungsu Park (bspark8@sk.com)

SK Telecom

Network IT Convergence R&D Center

Open system Lab.

- Mission is change all legacy infra to OpenStack cloud infrastructure.
- OREO(Open Reliable Elastic On OpenStack)
  - Openstack on k8s
  - Openstack with helm
- SONA(Self-Operating Networking Architecure)
  - Optimized tenant network virtualization
  - Neutron ML2 driver and L3 service plugin


Playing Starcraft 2
파괴자 has joined the game
(Other Team)
ALT-F4
you just lost The Game by 9GAG.COM

NETMARBLE IS GLOBAL FIRST-CLASS PUBLISHER

GLOBAL TOP GROSSING GAME PUBLISHERS (CONSOLIDATED BASIS, 2015 - FEB 2017)

RANK	2015	2016	FEB 2017
1	SUPERCELL	Tencent 腾讯	Tencent 腾讯
2	king	SUPERCELL	NetEase Games
3	mixi	NetEase Games	netmarble
4	GungHo	MACHINE ZONE	SUPERCELL
5	Tencent 腾讯	ACTIVISION BLIZZARD	MACHINE ZONE
6	MACHINE ZONE	mixi	mixi
7	LINE	netmarble	ACTIVISION BLIZZARD
8	netmarble	BANDAI NAMCO	BANDAI NAMCO

RANK	2015	2016	FEB 2017
1	SUPERCELL	SUPERCELL	netmarble
2	mixi	ACTIVISION BLIZZARD	SUPERCELL
3	king	netmarble	MACHINE ZONE
4	netmarble	BANDAI NAMCO	ACTIVISION BLIZZARD
5	GungHo	LINE	mixi
6	LINE		BANDAI NAMCO
7	colopl	GungHo	LINE
8	BANDAI NAMCO	mixi	SONY

NOTE: Netmarble's revenue for 2016 includes that of Jam City, but not of Kabam

SOURCE: App Annie

OpenStack at Netmarble

10K +

running instances

40 +

Game Services

Ceph at Netmarble

8

Clusters

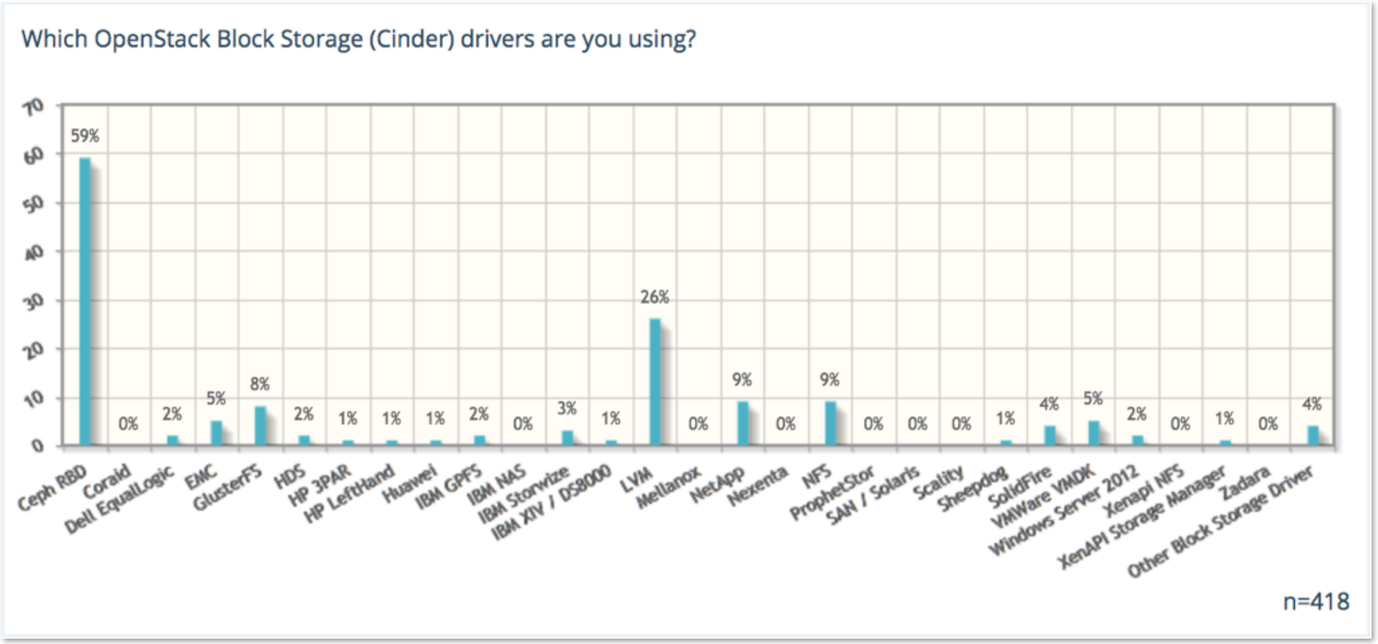
2.2PB +

Total Usage

1,900 +

OSDs

Background



OpenStack Survey 2017.(https://www.openstack.org/user-survey/survey-2017/)

- Strength of Ceph
  - Unified Storage System,
  - software defined storage
  - supports ephemeral, image, volume and backup backend
  - copy on write → fast provisioning

However, it's not easy to operate OpenStack with Ceph in production.

Background

We have to Think a lot of Things to Do

Background

- Several Tips for operating OpenStack with Ceph
- Here's our journey
  - Performance Tuning
    - CRUSH Tunables
    - Bucket Type
    - Journal Tuning
  - Operation
    - High Availability
    - Volume Migration
    - Volume Replication
    - Tips & Tricks


Performance
Tuning
• IOPS
• Throughput
• Latency


High
Availability


Volume
Migration


Volume
Replication

Performance Tuning

- Performance of Ceph
  - numerical performance
    - read/write performance performance etc..
  - rebalancing performance
    - minimalize the impact of recovery/rebalance
- Focusing on the rebalance performance → Advanced tuning points

Performance Tuning

- Tunables
  - improvements to the CRUSH algorithm used to calculate the placement of data
  - a series of tunable options that control whether the legacy or improved variation of the algorithm is used
- CRUSH Profile
  - ceph sets tunables “profiles” named by the release
  - legacy, argonaut, bobtail, firefly, optimal, default

```
user@ubuntu:~$ ceph osd crush show-tunables
{
    "choose_local_tries": 0,
    "choose_local_fallback_tries": 0,
    "choose_total_tries": 50,
    "chooseleaf_descend_once": 1,
    "chooseleaf_vary_r": 1,
    "chooseleaf_stable": 0,
    "straw_calc_version": 1,
    "allowed_bucket_algs": 22,
    "profile": "firefly",
    "optimal_tunables": 0,
    "legacy_tunables": 0,
    "minimum_required_version": "firefly",
    "require_feature_tunables": 1,
    "require_feature_tunables2": 1,
    "has_v2_rules": 0,
    "require_feature_tunables3": 1,
    "has_v3_rules": 0,
    "has_v4_buckets": 0,
    "require_feature_tunables5": 0,
    "has_v5_rules": 0
}
```

Performance Tuning

- CRUSH Tunables Version
  - ARGONAUT (LEGACY) : original legacy behavior
  - BOBTAIL(CRUSH_TUNABLES2)
    - some PGs map to fewer than the desired number of replicas
    - choose_local_tries = 0, choose_local_fallback_tries = 0, choose_total_tries = 50, chooseleaf_descend_once=1
  - FIREFLY(CRUSH_TUNABLES3)
    - chooseleaf_vary_r = 1
    - STRAW_CALC_VERSION TUNABLE : fix internal weight calculated algorithm for straw bucket
  - HAMMER(CRUSH_V4) : new bucket type straw2 supported
  - JEWEL(CRUSH_TUNABLES5)
    - improves the overall behavior of CRUSH
    - chooseleaf_stable

Performance Tuning

- CAUTION!
  - ceph client kernel must support the feature of tunables when you use not librbd but KRBD.

TUNABLE	RELEASE	CEPH VERSION	KERNEL
CRUSH_TUNABLES	argonaut	v0.48.1 ↑	v3.6 ↑
CRUSH_TUNABLES2	bobtail	v0.55 ↑	v3.9 ↑
CRUSH_TUNABLES3	firefly	v0.78 ↑	v3.15 ↑
CRUSH_V4	hammer	v0.94 ↑	v4.1 ↑
CRUSH_TUNABLES5	Jewel	V10.0.2 ↑	v4.5 ↑

Performance Tuning

- Bucket Type
  - ceph supports 4 bucket types, each representing a tradeoff between performance
  - straw
  - straw2
    - hammer tunable profile(CRUSH_V4 feature) straw2 support
    - straw2 bucket type fixed several limitations in the original straw bucket
    - the old straw buckets would change some mapping that should have changed when a weight was adjusted
    - straw2 achieves the original goal of only chaging mappings to or from the bucket item whose weight has changed
    - default set to straw2 after optimal of tunables

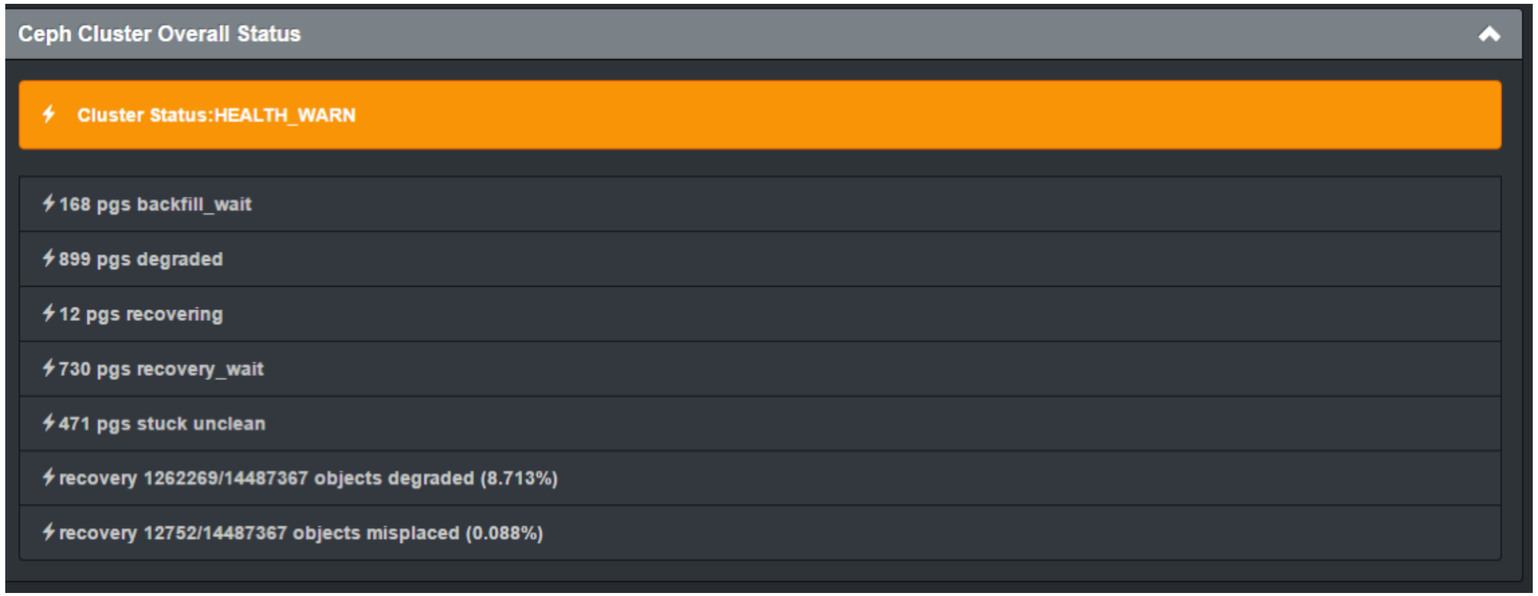
Performance Tuning

- Object Movement Test
  - Environment
    - 84 ODDs / 6 Hosts
    - Pick OSDs randomly {0, 14, 28, 42, 56, 70}

```
root rack4 {
        id -10          # do not change unnecessarily
        # weight 0.000
        alg straw
        hash 0  # rjenkins1
}
```

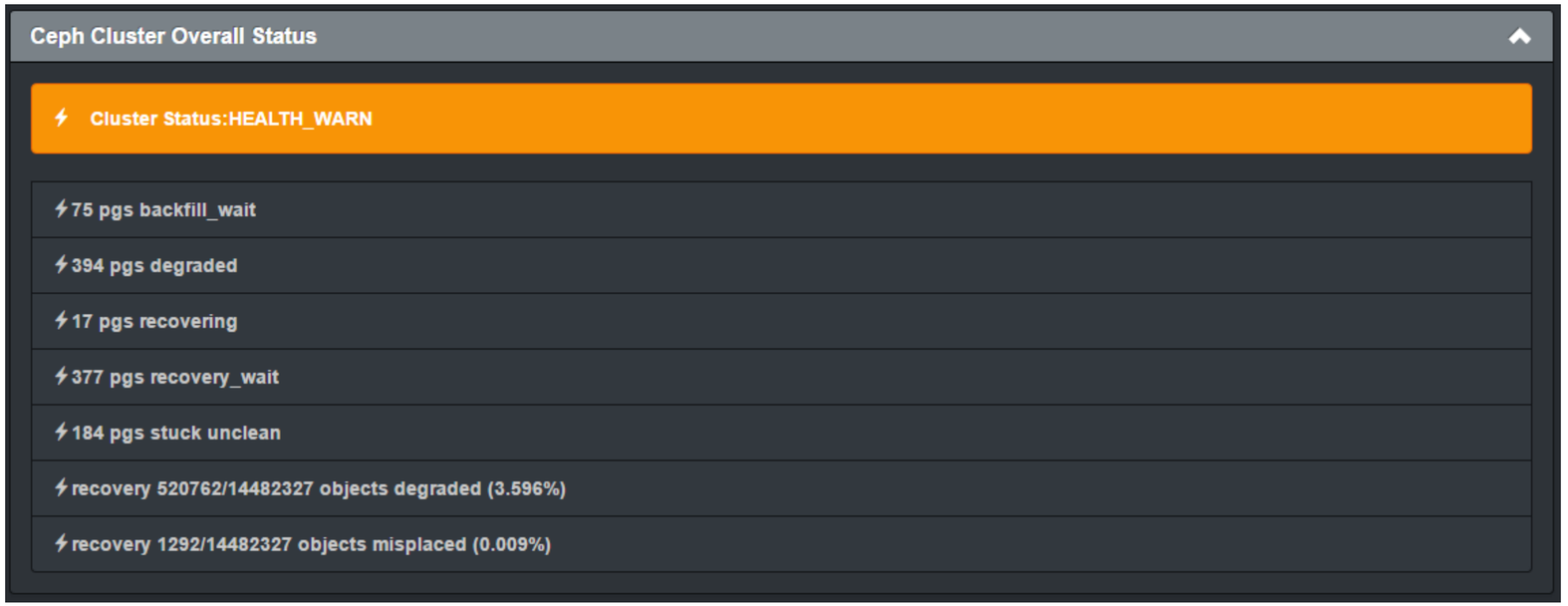
Performance Tuning

- in straw bucket → change weight to 0.3
  - 8.713 % degraded



Performance Tuning

- in straw2 bucket → change weight to 0.3
  - 3.596 % degraded

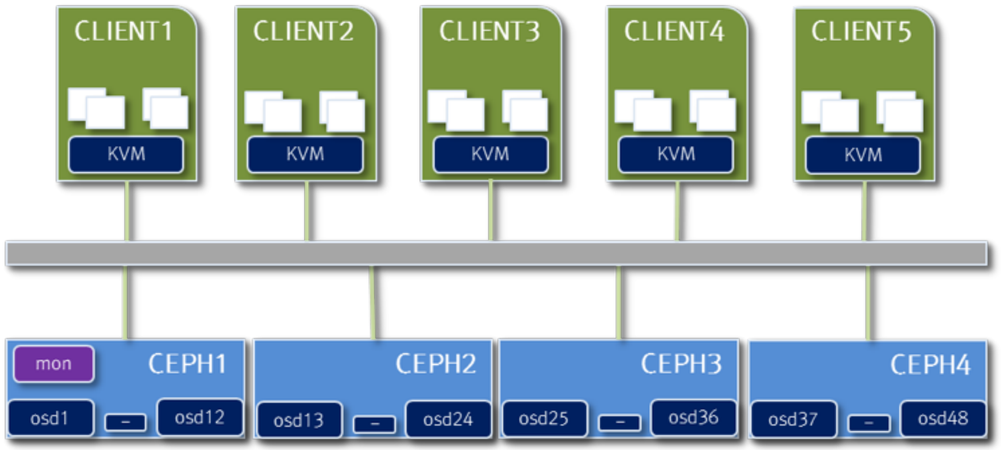


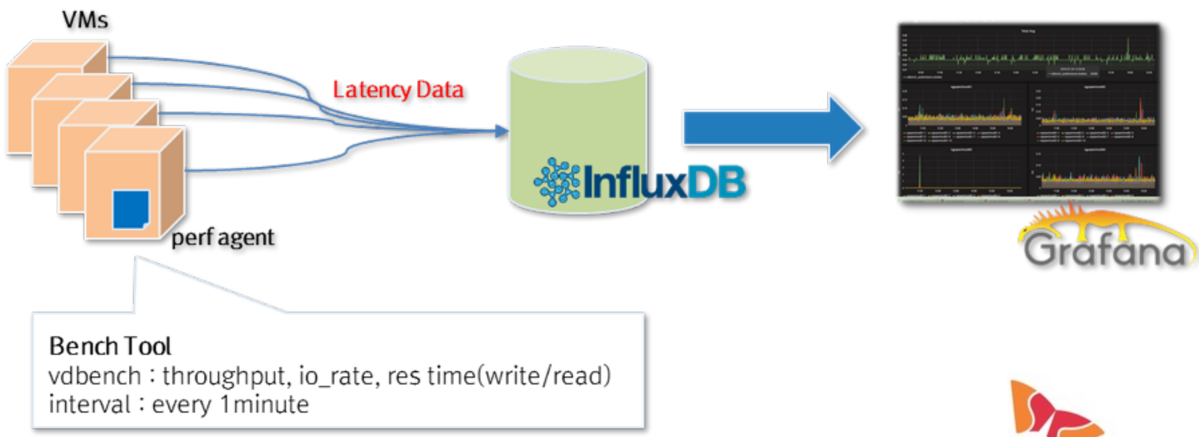
Performance Tuning

- SSD types
  - read intensive
    - solid cost-to-performance benefits for applications that demand low latency read speeds and greater bandwidth.
  - mixed use
    - based on a parallel processing architecture to deliver tested and proven reliability.
  - write intensive
    - featuring an I/O pattern designed to support applications with heavy write workloads.
- Most Cloud Environments
  - write I/O is more than read I/O (our case – 9:1)
  - rebalancing: SSD Journal can be bottleneck of IO

Performance Tuning

- Recovery Test Environment
  - total 100 VMs : Windows(50) + Linux(50)
  - bench Tool : python agent(vdbench)
  - recorded latency data every a minute through benchmark tool during recovery
  - overall stress of storage system : 400 ~ 700 MB/s



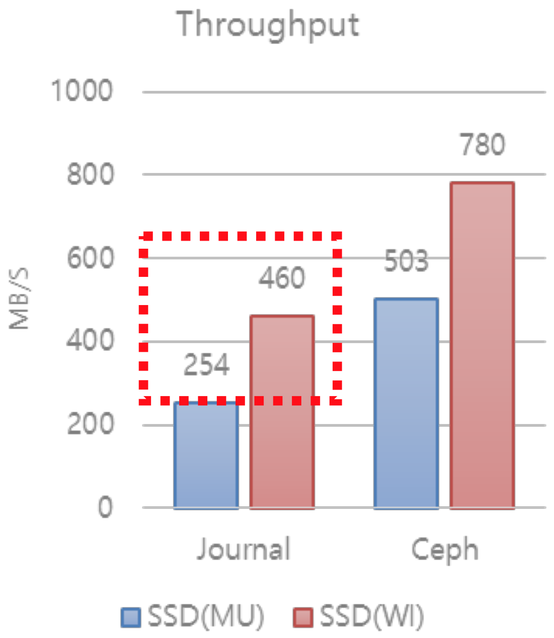


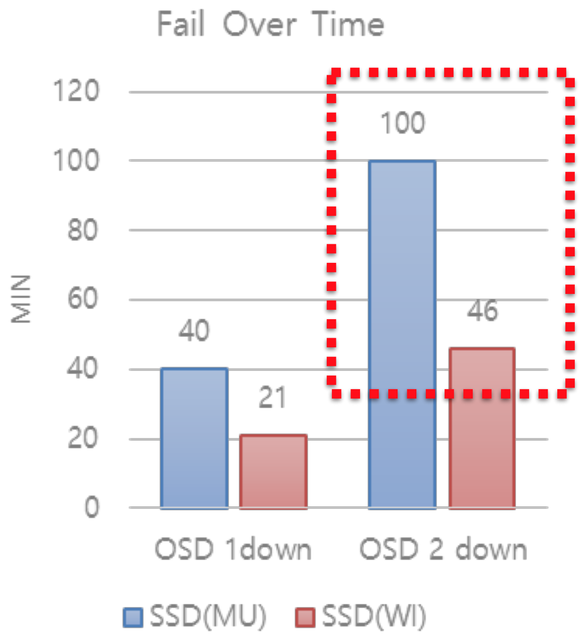
netmarble

SK telecom

Performance Tuning

- SSD Journal can be a bottleneck during recovery
- mixed unit
  - throughput : 254 MB/S
  - failover time
    - OSD 1 down : 40 min
    - OSD 2 down : **100** min
- write intensive
  - throughput : 460 MB/s
  - failover time
    - OSD 1 down : 21 min
    - OSD 2 down : **46** min





MU : Mixed Unit / WI : Write Intensive

netmarble

SK telecom

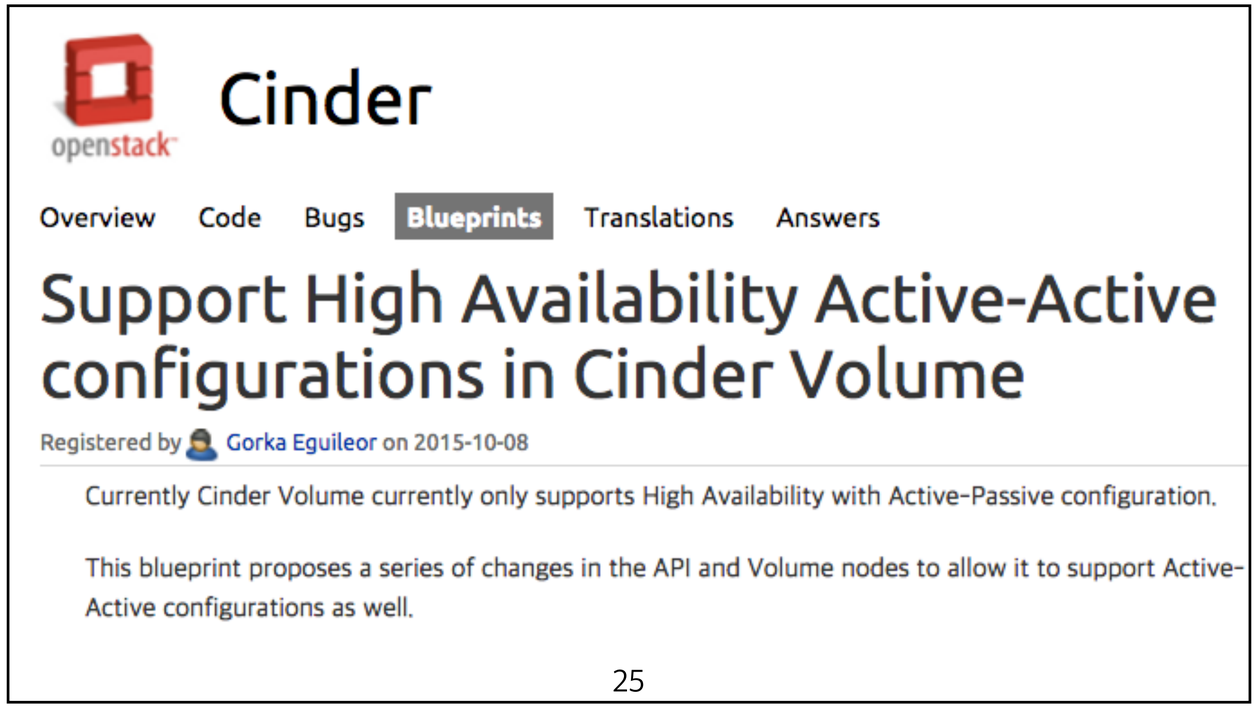

Performance
Tuning
High
Availability
Volume
Migration
Volume
Replication

High Availability

- Cinder Service
  - Cinder-API
  - Cinder-Scheduler
  - **Cinder-Volume**
  - **Cinder-Backup**

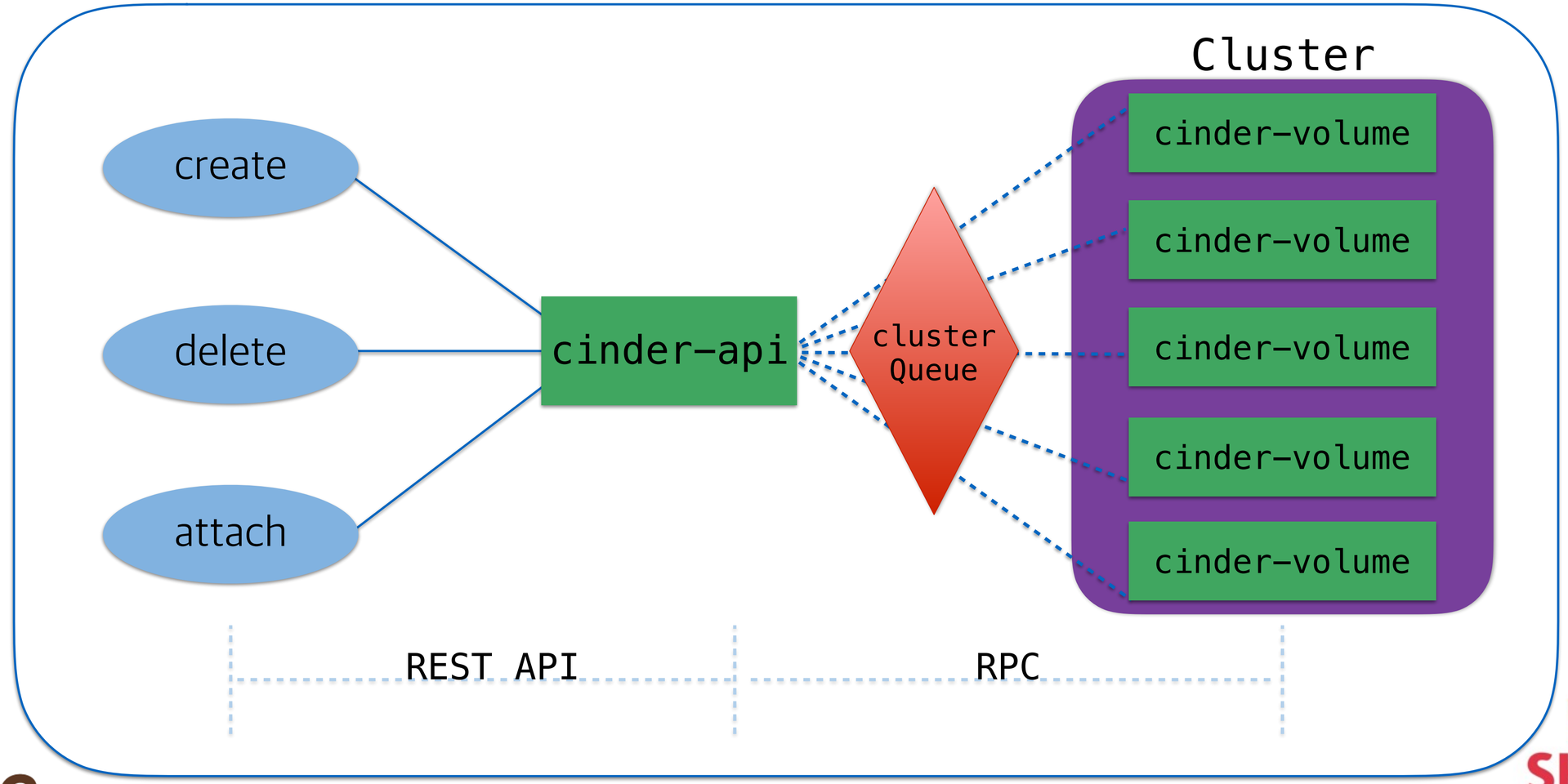
High Availability

- Cinder-Volume
- Status:
  - Traditionally, Active-Standby is recommended.
  - Active-Active is under construction but valuable to try.



High Availability

- Cinder-Volume Workflow



High Availability

- PoC: Cinder-Volume Active/Active
- Cinder Release: Master
- Some Volume Nodes
- Add “SUPPORTS_ACTIVE_ACTIVE” option to ceph volume driver

```
@interface.volumedriver
class RBDDriver(driver.CloneableImageVD,
                driver.MigrateVD, driver.ManageableVD, driver.BaseVD):
    """Implements RADOS block device (RBD) volume commands."""

    VERSION = '1.2.0'

    # ThirdPartySystems wiki page
    CI_WIKI_NAME = "Cinder_Jenkins"

    SYSCONFDIR = '/etc/ceph/'

    # NOTE(geguileo): This is not true, but we need it for our manual tests.
    SUPPORTS_ACTIVE_ACTIVE = True
```

High Availability

- PoC: Cinder-Volume Active/Active
- Add cluster option to cinder configuration file

```
[DEFAULT]
cluster = <YOUR_CLUSTER_NAME>
host = <HOSTNAME>
```

- Example

```
[DEFAULT]
cluster = cluster1
host = host1
```

```
[DEFAULT]
cluster = cluster1
host = host2
```

High Availability

- Cluster list

```
$ echo $OS_VOLUME_API_VERSION
3.29
$ cinder cluster-list --detail
+------+--------+-------+--------+-----------+---------------+----------------+-----------------+------------+------------+
| Name | Binary | State | Status | Num Hosts | Num Down Hosts | Last Heartbeat | Disabled Reason | Created At | Updated at |
+------+--------+-------+--------+-----------+---------------+----------------+-----------------+------------+------------+
+------+--------+-------+--------+-----------+---------------+----------------+-----------------+------------+------------+
```

- Cinder-Volume host1, host2 start

```
$ cinder cluster-list --detail
+-----------------------+---------------+-------+---------+-----------+----------------+-----------------------------+-----------------+-----------------------------+------------+
| Name                  | Binary        | State | Status  | Num Hosts | Num Down Hosts | Last Heartbeat              | Disabled Reason | Created At                  | Updated at |
+-----------------------+---------------+-------+---------+-----------+----------------+-----------------------------+-----------------+-----------------------------+------------+
| mycluster@lvmdriver-1 | cinder-volume | up    | enabled | 2         | 0              | 2017-04-25T12:11:37.000000 | -               | 2017-04-25T12:10:31.000000 |            |
| mycluster@rbd1        | cinder-volume | up    | enabled | 2         | 0              | 2017-04-25T12:11:43.000000 | -               | 2017-04-25T12:10:31.000000 |            |
+-----------------------+---------------+-------+---------+-----------+----------------+-----------------------------+-----------------+-----------------------------+------------+
$ cinder service-list
+------------------+------------------+------+---------+-------+----------------------------+----------------+-----------------+
| Binary           | Host             | Zone | Status  | State | Updated_at                 | Cluster        | Disabled Reason |
+------------------+------------------+------+---------+-------+----------------------------+----------------+-----------------+
| cinder-backup    | host1            | nova | enabled | up    | 2017-05-05T08:31:17.000000 | -              | -               |
| cinder-scheduler | host1            | nova | enabled | up    | 2017-05-05T08:31:16.000000 | -              | -               |
| cinder-volume    | host1@rbd1       | nova | enabled | up    | 2017-05-05T08:31:13.000000 | myCluster@rbd1 | -               |
| cinder-volume    | host2@rbd1       | nova | enabled | up    | 2017-05-05T08:31:15.000000 | myCluster@rbd1 | -               |
+------------------+------------------+------+---------+-------+----------------------------+----------------+-----------------+
```

High Availability

- Rally Test Scenario: Create-and-delete-volume.json

```
{
    "CinderVolumes.create_and_delete_volume": [
        {
            "args": {
                "size": 1
            },
            "runner": {
                "type": "constant",
                "times": 30,
                "concurrency": 2
            },
            "context": {
                "users": {
                    "tenants": 10,
                    "users_per_tenant": 2
                }
            }
        },
```

```
            "args": {
                "size": {
                    "min": 1,
                    "max": 5
                }
            },
            "runner": {
                "type": "constant",
                "times": 30,
                "concurrency": 2
            },
            "context": {
                "users": {
                    "tenants": 10,
                    "users_per_tenant": 2
                }
            }
        }
    ]
}
```

High Availability

- Rally Test Scenario: Create-and-delete-volume.json
- 2 host is running

```
--------------------------------------------------------------------------------
Task 88a9e631-4071-43c4-8502-a26d0075ee4c has 0 error(s)
--------------------------------------------------------------------------------

+-----------------------------------------------------------------------------------------------------------------------+
|                                            Response Times (sec)                                                       |
+------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
| Action                 | Min (sec) | Median (sec) | 90%ile (sec) | 95%ile (sec) | Max (sec) | Avg (sec) | Success | Count |
+------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
| cinder_v2.create_volume | 2.622     | 2.765        | 2.852        | 2.861        | 2.892     | 2.757     | 100.0%  | 30    |
| cinder_v2.delete_volume | 0.424     | 2.35         | 2.487        | 2.56         | 2.617     | 2.251     | 100.0%  | 30    |
| total                  | 3.176     | 5.116        | 5.287        | 5.342        | 5.469     | 5.009     | 100.0%  | 30    |
+------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
```

High Availability

- Rally Test Scenario: Create-and-delete-volume.json
- 1 host is running, 1 host is down

```
--------------------------------------------------------------------------------
Task 5cd94f22-d75b-4d3d-bbbb-3b7d7f58a6df has 0 error(s)
--------------------------------------------------------------------------------

+-------------------------------------------------------------------------------------------------------------------------+
|                                            Response Times (sec)                                                         |
+-------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
| Action                  | Min (sec) | Median (sec) | 90%ile (sec) | 95%ile (sec) | Max (sec) | Avg (sec) | Success | Count |
+-------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
| cinder_v2.create_volume | 2.585     | 2.725        | 2.874        | 2.9          | 2.961     | 2.74      | 100.0%  | 30    |
| cinder_v2.delete_volume | 2.293     | 2.338        | 2.452        | 2.494        | 2.529     | 2.357     | 100.0%  | 30    |
| total                   | 4.921     | 5.082        | 5.249        | 5.317        | 5.457     | 5.097     | 100.0%  | 30    |
+-------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
```

High Availability

- Rally Test Scenario: Create-and-attach-volume.json

```
{% set flavor_name = flavor_name or "m1.tiny" %}
{% set availability_zone = availability_zone or "nova" %}
{
    "CinderVolumes.create_and_attach_volume": [
        {
            "args": {
                "size": 10,
                "image": {
                    "name": "^cirros.*-disk$"
                },
                "flavor": {
                    "name": "{{flavor_name}}"
                },
                "create_volume_params": {
                    "availability_zone": "{{availability_zone}}"
                }
            },
            "runner": {
                "type": "constant",
                "times": 5,
                "concurrency": 1
            },
            "context": {
                "users": {
                    "tenants": 2,
                    "users_per_tenant": 2
                }
            }
        },
```

```
        {
            "args": {
                "size": {
                    "min": 1,
                    "max": 5
                },
                "flavor": {
                    "name": "{{flavor_name}}"
                },
                "image": {
                    "name": "^cirros.*-disk$"
                },
                "create_volume_params": {
                    "availability_zone": "{{availability_zone}}"
                }
            },
            "runner": {
                "type": "constant",
                "times": 5,
                "concurrency": 1
            },
            "context": {
                "users": {
                    "tenants": 2,
                    "users_per_tenant": 2
                }
            }
        }
    ]
}
```

High Availability

- Rally Test Scenario: Create-and-delete-volume.json
- 2 host is running

```
--------------------------------------------------------------------------------
Task aa444409-7863-4605-878e-a9c50f6392bb has 0 error(s)
--------------------------------------------------------------------------------

+-----------------------------------------------------------------------------------------------------------------------+
|                                            Response Times (sec)                                                       |
+------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
| Action                 | Min (sec) | Median (sec) | 90%ile (sec) | 95%ile (sec) | Max (sec) | Avg (sec) | Success | Count |
+------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
| nova.boot_server       | 4.588     | 4.917        | 5.004        | 5.008        | 5.012     | 4.867     | 100.0%  | 5     |
| cinder_v2.create_volume| 2.481     | 2.548        | 2.584        | 2.595        | 2.605     | 2.54      | 100.0%  | 5     |
| nova.attach_volume     | 2.803     | 2.961        | 3.01         | 3.024        | 3.038     | 2.935     | 100.0%  | 5     |
| nova.detach_volume     | 9.551     | 9.645        | 9.757        | 9.776        | 9.794     | 9.65      | 100.0%  | 5     |
| cinder_v2.delete_volume| 2.321     | 2.34         | 2.356        | 2.36         | 2.364     | 2.341     | 100.0%  | 5     |
| nova.delete_server     | 2.622     | 2.784        | 2.899        | 2.905        | 2.911     | 2.78      | 100.0%  | 5     |
| total                  | 23.809    | 24.097       | 24.451       | 24.563       | 24.675    | 24.113    | 100.0%  | 5     |
+------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
```

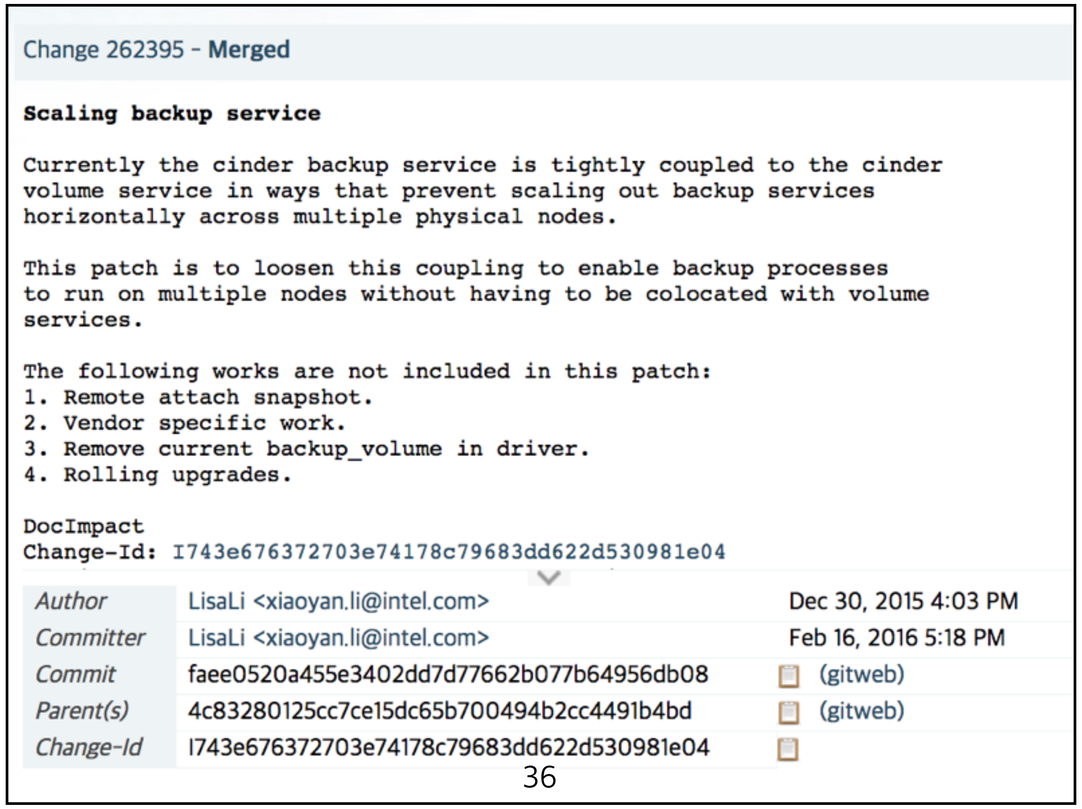
High Availability

- Rally Test Scenario: Create-and-attach-volume.json
- 1 host is running, 1 host is down

```
--------------------------------------------------------------------------------
Task bf6ab722-2fad-4830-8e38-5947d9f7a8f0 has 0 error(s)
--------------------------------------------------------------------------------


+-------------------------------------------------------------------------------------------------------------------------+
|                                            Response Times (sec)                                                         |
+-------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
| Action                  | Min (sec) | Median (sec) | 90%ile (sec) | 95%ile (sec) | Max (sec) | Avg (sec) | Success | Count |
+-------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
| nova.boot_server        | 4.617     | 4.853        | 4.921        | 4.941        | 4.962     | 4.814     | 100.0%  | 5     |
| cinder_v2.create_volume | 2.522     | 2.561        | 2.615        | 2.618        | 2.621     | 2.573     | 100.0%  | 5     |
| nova.attach_volume      | 2.898     | 2.955        | 3.054        | 3.056        | 3.058     | 2.98      | 100.0%  | 5     |
| nova.detach_volume      | 9.487     | 9.633        | 9.762        | 9.774        | 9.786     | 9.636     | 100.0%  | 5     |
| cinder_v2.delete_volume | 2.322     | 2.361        | 2.375        | 2.379        | 2.383     | 2.358     | 100.0%  | 5     |
| nova.delete_server      | 2.685     | 2.887        | 2.919        | 2.926        | 2.934     | 2.838     | 100.0%  | 5     |
| total                   | 23.854    | 24.162       | 24.457       | 24.523       | 24.589    | 24.198    | 100.0%  | 5     |
+-------------------------+-----------+--------------+--------------+--------------+-----------+-----------+---------+-------+
```

High Availability

- Cinder-Backup
- Status:
  - It support scale-out, not HA strictly speaking.



Change 262395 - **Merged**

Scaling backup service

Currently the cinder backup service is tightly coupled to the cinder volume service in ways that prevent scaling out backup services horizontally across multiple physical nodes.

This patch is to loosen this coupling to enable backup processes to run on multiple nodes without having to be colocated with volume services.

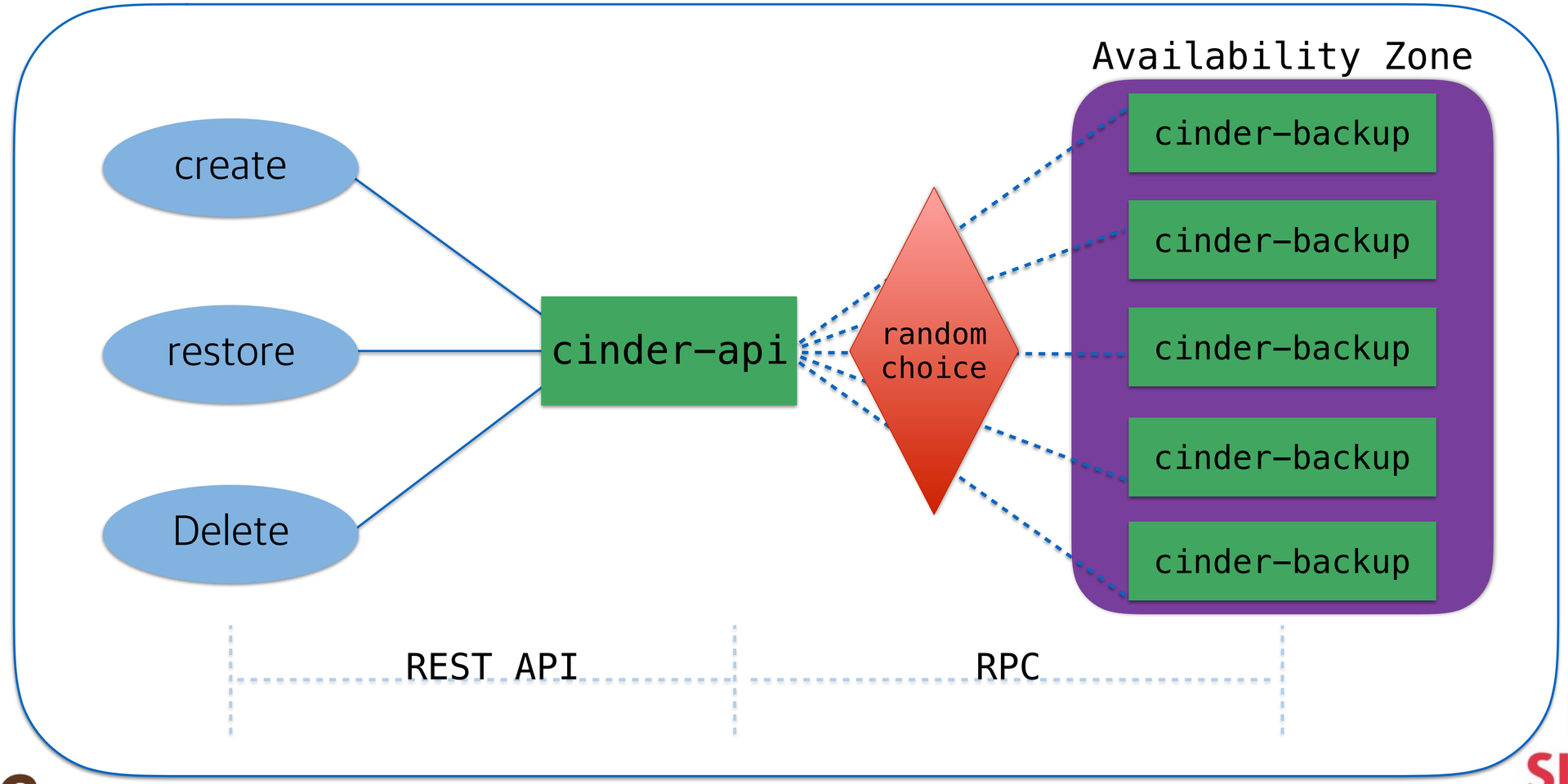
The following works are not included in this patch:
1. Remote attach snapshot.
2. Vendor specific work.
3. Remove current backup_volume in driver.
4. Rolling upgrades.

DocImpact
Change-Id: I743e676372703e74178c79683dd622d530981e04

Author	LisaLi <xiaoyan.li@intel.com>	Dec 30, 2015 4:03 PM
Committer	LisaLi <xiaoyan.li@intel.com>	Feb 16, 2016 5:18 PM
Commit	faee0520a455e3402dd7d77662b077b64956db08	(gitweb)
Parent(s)	4c83280125cc7ce15dc65b700494b2cc4491b4bd	(gitweb)
Change-Id	I743e676372703e74178c79683dd622d530981e04	

High Availability

- Cinder-Backup Workflow



High Availability

- PoC: Cinder-Backup Active/Active
- Cinder Release: Master
- Add backup option to cinder configuration file
- Options

```
[DEFAULT]
host=[hostname]
backup_driver=cinder.backup.drivers.ceph
backup_ceph_conf=[ceph config file]
backup_ceph_user=[ceph user for cinder-backup]
backup_ceph_pool=[ceph pool for cinder-backup]
```

- Example

```
[DEFAULT]
host = host1
backup_driver=cinder.backup.drivers.ceph
backup_ceph_conf=/etc/ceph/ceph.conf
backup_ceph_user=cinder-backup
backup_ceph_pool=backups
```

High Availability

- cinder service-list after running cinder-backup at 2 hosts.

```
$ cinder service-list
+------------------+------------------+------+---------+-------+----------------------------+-----------------+
| Binary           | Host             | Zone | Status  | State | Updated_at                 | Disabled Reason |
+------------------+------------------+------+---------+-------+----------------------------+-----------------+
| cinder-backup    | host1            | nova | enabled | up    | 2017-05-02T08:47:15.000000 | -               |
| cinder-backup    | host2            | nova | enabled | up    | 2017-05-02T08:47:21.000000 | -               |
| cinder-scheduler | host1            | nova | enabled | up    | 2017-05-02T08:47:16.000000 | -               |
| cinder-volume    | host1@lvmdriver-1 | nova | enabled | up    | 2017-05-02T08:47:17.000000 | -               |
| cinder-volume    | host1@lvmdriver-2 | nova | enabled | up    | 2017-05-02T08:47:17.000000 | -               |
| cinder-volume    | host1@rbd1       | nova | enabled | up    | 2017-05-02T08:47:14.000000 | -               |
+------------------+------------------+------+---------+-------+----------------------------+-----------------+
```

High Availability

- Rally Test Scenario: Create-volume-backup.json

```
{
    "CinderVolumes.create_volume_backup": [
        {
            "args": {
                "size": 1,
                "do_delete": true,
                "create_volume_kwargs": {},
                "create_backup_kwargs": {}
            },
            "runner": {
                "type": "constant",
                "times": 30,
                "concurrency": 2
            },
            "context": {
                "users": {
                    "tenants": 1,
                    "users_per_tenant": 1
                },
                "roles": ["Member"]
            }
        }
    ]
}
```

High Availability

- Rally Test Scenario: Create-volume-backup.json
- 2 hosts is running

```
--------------------------------------------------------------------------------
Task ab0d730f-e1a6-42a3-8595-1de2e81c4c06 has 0 error(s)
--------------------------------------------------------------------------------

+--------------------------------------------------------------------------------------------------------------------------------+
|                                              Response Times (sec)                                                              |
+--------------------------+-----------+---------------+---------------+---------------+-----------+-----------+---------+-------+
| Action                   | Min (sec) | Median (sec)  | 90%ile (sec)  | 95%ile (sec)  | Max (sec) | Avg (sec) | Success | Count |
+--------------------------+-----------+---------------+---------------+---------------+-----------+-----------+---------+-------+
| cinder_v2.create_volume  | 2.483     | 2.663         | 2.911         | 2.952         | 2.985     | 2.704     | 100.0%  | 30    |
| cinder_v2.create_backup  | 10.713    | 12.924        | 17.113        | 17.188        | 17.224    | 13.818    | 100.0%  | 30    |
| cinder_v2.delete_volume  | 0.37      | 2.346         | 2.491         | 2.529         | 2.549     | 2.306     | 100.0%  | 30    |
| cinder_v2.delete_backup  | 2.21      | 2.237         | 2.282         | 2.289         | 2.368     | 2.245     | 100.0%  | 30    |
| total                    | 17.881    | 20.35         | 24.347        | 24.531        | 24.7      | 21.072    | 100.0%  | 30    |
+--------------------------+-----------+---------------+---------------+---------------+-----------+-----------+---------+-------+

Load duration: 316.327989
Full duration: 319.620925
```

High Availability

- Rally Test Scenario: Create-volume-backup.json
- 1 host is running, 1 host is down

Response Times (sec)								
Action	Min (sec)	Median (sec)	90%ile (sec)	95%ile (sec)	Max (sec)	Avg (sec)	Success	Count
cinder_v2.create_volume	2.563	2.712	2.88	2.901	2.943	2.733	93.3%	30
cinder_v2.create_backup	8.582	8.804	13.078	13.271	15.019	10.31	93.3%	30
cinder_v2.delete_volume	0.402	2.37	2.44	2.474	2.498	2.31	96.6%	29
cinder_v2.delete_backup	2.204	4.328	8.538	8.549	8.583	4.736	96.6%	29
total	15.931	20.192	24.49	26.158	28.725	20.089	93.3%	30

Load duration: 896.267822
Full duration: 901.213922

High Availability

- Rally Test Scenario: Create-volume-backup.json
- 1 host is running, 1 host is down

```
--------------------------------------------------------------------------------
Task 03485d5b-cef1-4356-99ee-9c3e84e9e473 has 2 error(s)
--------------------------------------------------------------------------------

TimeoutException: Rally tired waiting for VolumeBackup s_rally_ba5f2042_F02V0nth:3b81449e-bda3-4065-b560-1d9ccc44b33c to become ('AVAILABLE') current status CREATING

Traceback (most recent call last):
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/task/runner.py", line 72, in _run_scenario_once
    getattr(scenario_inst, method_name)(**scenario_kwargs)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/plugins/openstack/scenarios/cinder/volumes.py", line 629, in run
    backup = self.cinder.create_backup(volume.id, **create_backup_kwargs)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/task/service.py", line 116, in wrapper
    return func(instance, *args, **kwargs)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/plugins/openstack/services/storage/block.py", line 270, in create_backup
    snapshot_id=snapshot_id)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/task/service.py", line 116, in wrapper
    return func(instance, *args, **kwargs)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/plugins/openstack/services/storage/cinder_v2.py", line 321, in create_backup
    incremental=incremental, force=force, snapshot_id=snapshot_id))
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/task/service.py", line 116, in wrapper
    return func(instance, *args, **kwargs)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/task/atomic.py", line 82, in func_atomic_actions
    f = func(self, *args, **kwargs)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/plugins/openstack/services/storage/cinder_v2.py", line 161, in create_backup
    return self._wait_available_volume(backup)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/plugins/openstack/services/storage/cinder_common.py", line 56, in _wait_available_volume
    check_interval=CONF.benchmark.cinder_volume_create_poll_interval
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/task/utils.py", line 252, in wait_for_status
    resource_status=get_status(resource, status_attr))
TimeoutException: Rally tired waiting for VolumeBackup s_rally_ba5f2042_F02V0nth:3b81449e-bda3-4065-b560-1d9ccc44b33c to become ('AVAILABLE') current status CREATING

--------------------------------------------------------------------------------
TimeoutException: Rally tired waiting for VolumeBackup s_rally_ba5f2042_ptYW7qjo:00d1bd1c-315c-440d-9624-2effcc69849e to become ('DELETED') current status DELETING

Traceback (most recent call last):
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/task/runner.py", line 72, in _run_scenario_once
    getattr(scenario_inst, method_name)(**scenario_kwargs)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/plugins/openstack/scenarios/cinder/volumes.py", line 633, in run
    self.cinder.delete_backup(backup)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/task/service.py", line 116, in wrapper
    return func(instance, *args, **kwargs)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/plugins/openstack/services/storage/block.py", line 275, in delete_backup
    self._impl.delete_backup(backup)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/plugins/openstack/services/storage/cinder_common.py", line 524, in delete_backup
    self._impl.delete_backup(backup)
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/plugins/openstack/services/storage/cinder_common.py", line 274, in delete_backup
    check_interval=(CONF.benchmark
  File "/opt/stack/rally/local/lib/python2.7/site-packages/rally/task/utils.py", line 252, in wait_for_status
    resource_status=get_status(resource, status_attr))
TimeoutException: Rally tired waiting for VolumeBackup s_rally_ba5f2042_ptYW7qjo:00d1bd1c-315c-440d-9624-2effcc69849e to become ('DELETED') current status DELETING

--------------------------------------------------------------------------------
```


Performance
Tuning
High
Availability
Volume
Migration
Volume
Replication

Volume Migration

- What kind of feature?
  - migrating a volume transparently moves its data from the current backend for the volume to a new one
- How can we take advantage of this feature?
  - storage evacuation for maintaing or decommissioning
  - Manual optimization
- Driver Specific Migration
  - storage driver internal commands to migrate
  - calling the method migrate_volume in the driver
  - ex> EMC(vmax, vnx), HP(3PAR, Lefthand), Nexenta
  - volume migrate in rbd driver
    - Not Implemented

```
def migrate_volume(self, context, volume, host):
    return (False, None)
```

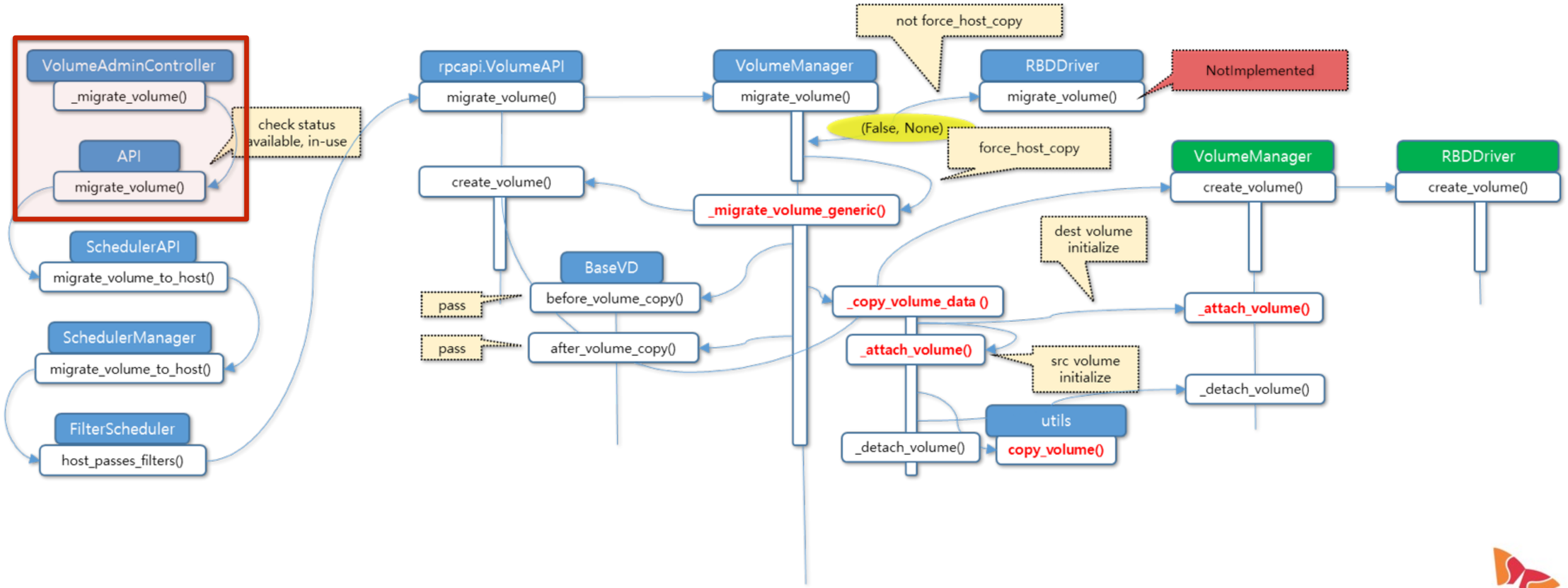
Volume Migration

- Host Assisted Migration
  - based on the volume attachment to the node of cinder volume service.
  - block-based : iSCSI supported storage, transfer mode Is done by ‘dd’ command
  - file-based : transfer mode is done by file copy (ex> ceph)
- Use Cases

NO	attachment	src driver	dest driver	volume type	migrate type	Result
1	available	rbd	rbd	same	host assisted	??
2	available	rbd	rbd	different	host assisted	??
3	available	block based	ceph	different	host assisted	??
		ceph	block based			
4	in-use	rbd	rbd	same	host assisted	??

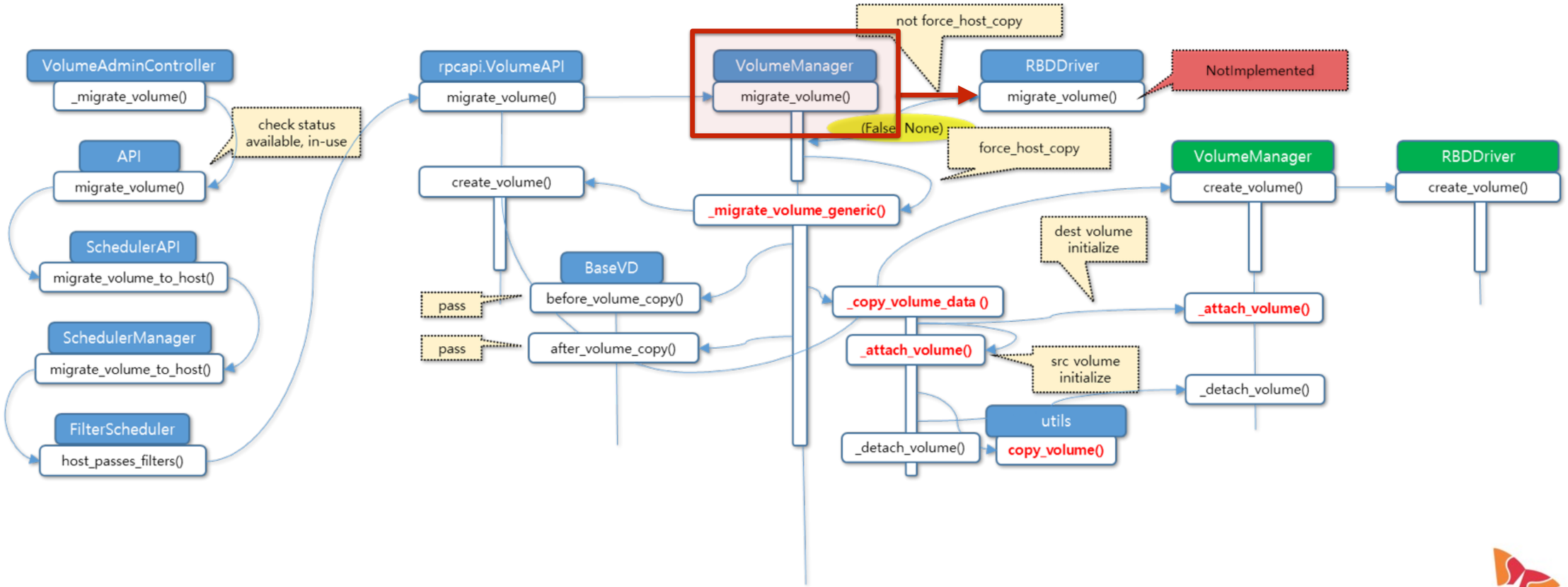
Volume Migration

- volume migration flow



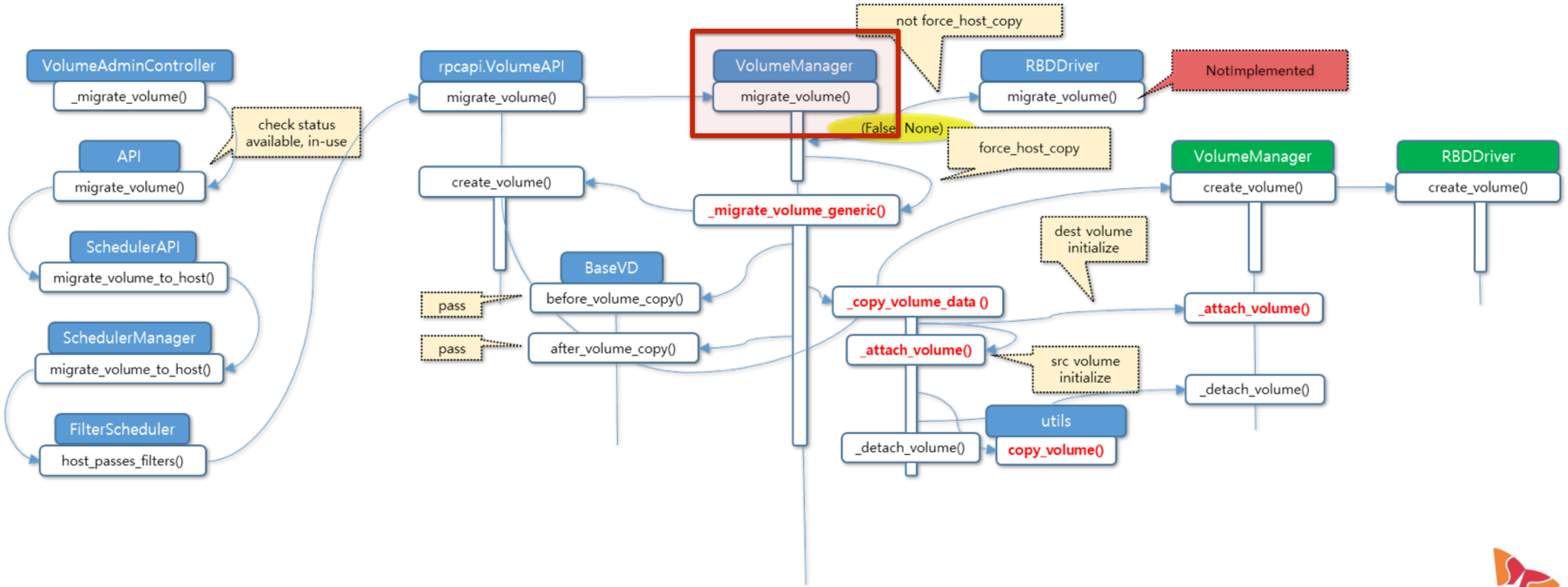
Volume Migration

- volume migration flow



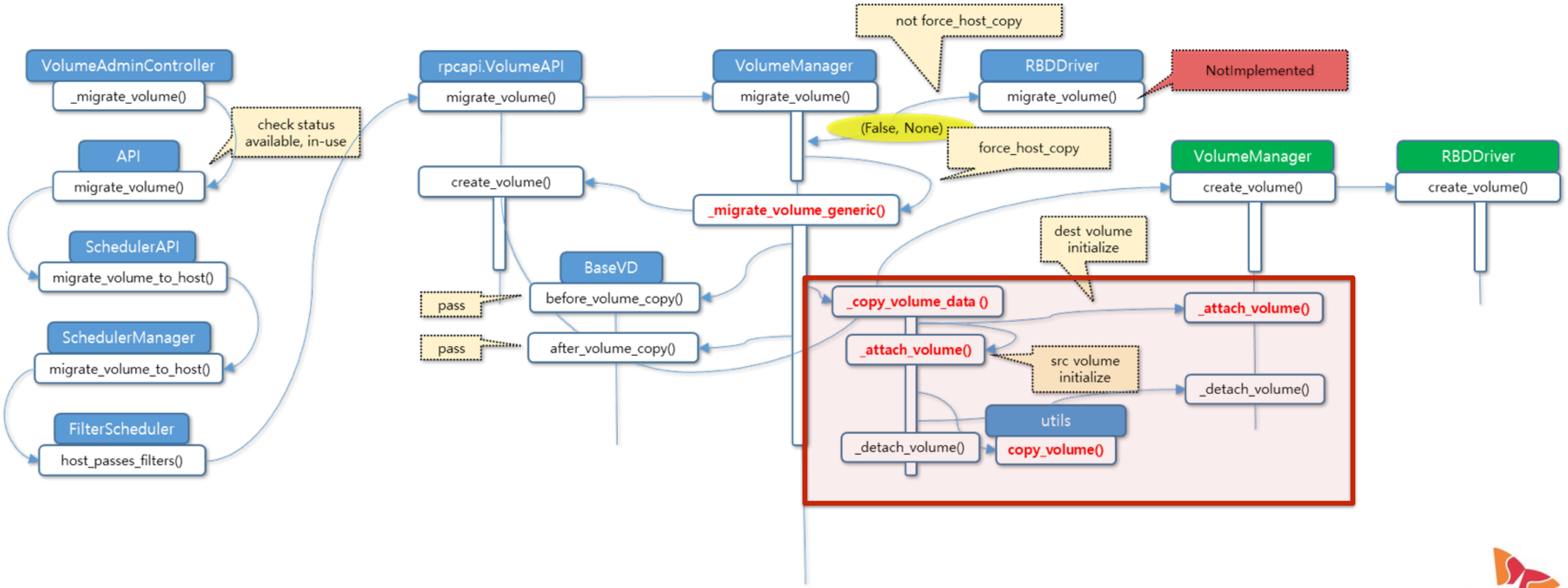
Volume Migration

- volume migration flow



Volume Migration

- volume migration flow



Volume Migration

- Use Case #1 (available, between rbd driver, same type)
  - show volume type

```
# cinder type-list
+--------------------------------------+--------+--------------+-----------+
| ID                                   |  Name  |  Description | Is_Public |
+--------------------------------------+--------+--------------+-----------+
| 318f62d1-c76a-4a59-9b3a-53bc69ce8cd0 |  ceph  |              | True      |
+--------------------------------------+--------+--------------+-----------+
```

  - show host list

```
# cinder get-pools
+----------+----------------------------------+
| Property |              Value               |
+----------+----------------------------------+
|   name   | ngopenctrl01@ceph-1#CEPH         |
+----------+----------------------------------+
+----------+----------------------------------+
| Property |              Value               |
+----------+----------------------------------+
|   name   | ngopenctrl01@ceph-2#CEPH         |
+----------+----------------------------------+
```

Volume Migration

- migrate volume command

```
# cinder migrate 705d635b-54ef-4ff9-9d88-d150a9e5ace8 --host ngopenctrl01@ceph-2#CEPH --force-host-copy True
Request to migrate volume 705d635b-54ef-4ff9-9d88-d150a9e5ace8 has been accepted.
```

- migrating volume

```
# cinder list
+--------------------------------------+-----------+-------+------+-------------+----------+-------------+
| ID                                   | Status    | Name  | Size | Volume Type | Bootable | Attached to |
+--------------------------------------+-----------+-------+------+-------------+----------+-------------+
| 705d635b-54ef-4ff9-9d88-d150a9e5ace8 | available | vol-1 | 10   | ceph        | true     |             |
| 9343e778-8acc-4550-a0fc-ee9e32e58112 | available | vol-1 | 10   | ceph        | true     |             |
+--------------------------------------+-----------+-------+------+-------------+----------+-------------+
```

Volume Migration

- Show Migration Result

```
# cinder show 705d635b-54ef-4ff9-9d88-d150a9e5ace8
+--------------------------------+--------------------------------------+
| Property                       | Value                                |
+--------------------------------+--------------------------------------+
| id                             | 705d635b-54ef-4ff9-9d88-d150a9e5ace8 |
| migration_status               | success                              |
| name                           | migrate-1                            |
| os-vol-host-attr:host          | f1osctrl01@ceph-2#CEPH               |
| os-vol-mig-status-attr:migstat | success                              |
| os-vol-mig-status-attr:name_id | None                                 |
| volume_type                    | ceph                                 |
+--------------------------------+--------------------------------------+
```

Volume Migration

- Use Case #2 (available, between rbd driver, different type)
  - show volume type

```
# cinder type-list
+--------------------------------------+--------+-------------+-----------+
| ID                                   |  Name  | Description | Is_Public |
+--------------------------------------+--------+-------------+-----------+
| 318f62d1-c76a-4a59-9b3a-53bc69ce8cd0 | ceph-1 | -           | True      |
| a2abb593-6331-4899-9030-2d5872ffbdeb | ceph-2 | -           | True      |
+--------------------------------------+--------+-------------+-----------+
```

  - show host list

```
# cinder get-pools
+----------+----------------------------------+
| Property |              Value               |
+----------+----------------------------------+
|   name   | ngopenctrl01@ceph-1#CEP          |
+----------+----------------------------------+
+----------+----------------------------------+
| Property |              Value               |
+----------+----------------------------------+
|   name   | ngopenctrl01@ceph-2#CEPH         |
+----------+----------------------------------+
```

Volume Migration

- migrate volume completed! However, volume type not changed
- volume show

```
# cinder show 9343e778-8acc-4550-a0fc-ee9e32e58112
+--------------------------------+--------------------------------------+
| Property                       | Value                                |
+--------------------------------+--------------------------------------+
| id                             | 9343e778-8acc-4550-a0fc-ee9e32e58112 |
| migration_status               | success                              |
| name                           | migrate-1                            |
| os-vol-host-attr:host          | f1osctrl01@ceph-2#CEPH               |
| os-vol-mig-status-attr:migstat | success                              |
| os-vol-mig-status-attr:name_id | None                                 |
| volume_type                    | ceph-1                               |
+--------------------------------+--------------------------------------+
```

Volume Migration

- volume retype

```
# cinder retype 9343e778-8acc-4550-a0fc-ee9e32e58112 ceph-2
# cinder list
+--------------------------------------+-----------+--------+------+-------------+----------+--------------------------------------+
| ID                                   | Status    | Name   | Size | Volume Type | Bootable | Attached to                          |
+--------------------------------------+-----------+--------+------+-------------+----------+--------------------------------------+
| 9343e778-8acc-4550-a0fc-ee9e32e58112 | available | vol-1  | 1    | ceph-2      | false    |                                      |
+--------------------------------------+-----------+--------+------+-------------+----------+--------------------------------------+
```

Volume Migration

- Use Case #3 (available, different driver, different type)
  - At first, we should remove CapabilitiesFilter in scheduler_filter
  - volume list

```
stack@devstack01:~$ cinder list
+--------------------------------------+--------+--------+------+-------------+----------+--------------------------------------+
| ID                                   | Status | Name   | Size | Volume Type | Bootable | Attached to                          |
+--------------------------------------+--------+--------+------+-------------+----------+--------------------------------------+
| 9343e778-8acc-4550-a0fc-ee9e32e58112 | in-use | vol-1  | 10   | ceph        | true     | 02444a9a-f6a9-4f25-a55a-4718f6944c32 |
| e852542e-d4df-47cf-bc4a-a5974a2af330 | in-use | eqlx-1 | 1    | eqlx        | false    | 02444a9a-f6a9-4f25-a55a-4718f6944c32 |
+--------------------------------------+--------+--------+------+-------------+----------+--------------------------------------+
```

  - Create a text file in the vm

```
root@sjtest-3:/DATA# echo "Hello World" > test.txt
root@sjtest-3:/DATA# cat test.txt
Hello World
```

Volume Migration

- migrating the volume

```
# cinder migrate e852542e-d4df-47cf-bc4a-a5974a2af330 devstack01@ceph-1#CEPH --force-host-copy
Request to migrate volume e852542e-d4df-47cf-bc4a-a5974a2af330 has been accepted.
# cinder list
+--------------------------------------+-----------+--------+------+-------------+----------+--------------------------------------+
| ID                                   | Status    | Name   | Size | Volume Type | Bootable | Attached to                          |
+--------------------------------------+-----------+--------+------+-------------+----------+--------------------------------------+
| 9343e778-8acc-4550-a0fc-ee9e32e58112 | in-use    | vol-1  | 10   | ceph        | true     | 02444a9a-f6a9-4f25-a55a-4718f6944c32 |
| e852542e-d4df-47cf-bc4a-a5974a2af330 | available | eqlx-1 | 1    | eqlx        | false    |                                      |
| eae08411-7555-4160-8975-b9105131a733 | available | eqlx-1 | 1    | eqlx        | false    |                                      |
+--------------------------------------+-----------+--------+------+-------------+----------+--------------------------------------+
```

- retype command
- re-attach the volume to the vm

Volume Migration

- mounting the volume

```
root@sjtest-3:~# mount /dev/vdb1 /DATA/
mount: wrong fs type, bad option, bad superblock on /dev/vdc1,
       missing codepage or helper program, or other error

       In some cases useful info is found in syslog - try
       dmesg | tail or so.
```

- we can fix

```
root@sjtest-3:~# e2fsck -f /dev/vdb1
e2fsck 1.42.13 (17-May-2015)
The filesystem size (according to the superblock) is 264704 blocks
The physical size of the device is 261888 blocks
Either the superblock or the partition table is likely to be corrupt!
Abort<y>? no
Pass 1: Checking inodes, blocks, and sizes
Pass 2: Checking directory structure
Pass 3: Checking directory connectivity
Pass 4: Checking reference counts
Pass 5: Checking group summary information
Free blocks count wrong (252018, counted=252017).
Fix<y>? yes
Free inodes count wrong (66229, counted=66228).
Fix<y>? yes

/dev/vdb1: ***** FILE SYSTEM WAS MODIFIED *****
/dev/vdb1: 12/66240 files (0.0% non-contiguous), 12687/264704 blocks
```

```
root@sjtest-3:~# resize2fs /dev/vdb1
resize2fs 1.42.13 (17-May-2015)
Resizing the filesystem on /dev/vdb1 to 261888 (4k) blocks.
The filesystem on /dev/vdb1 is now 261888 (4k) blocks long.
```

Volume Migration

- Use Case #4 (in-use, between rbd driver, same type)
  - volume list

```
# cinder list
+--------------------------------------+--------+-------+------+-------------+----------+--------------------------------------+
| ID                                   | Status | Name  | Size | Volume Type | Bootable | Attached to                          |
+--------------------------------------+--------+-------+------+-------------+----------+--------------------------------------+
| ef682166-fe19-4947-bfdb-ef7bc366765d | in-use | mig-1 | 1    | ceph        | false    | efbde317-32e7-4b4a-bf80-f38b056578c9 |
+--------------------------------------+--------+-------+------+-------------+----------+--------------------------------------+
```

  - However, an error raised in nova compute

```
[instance: efbde317-32e7-4b4a-bf80-f38b056578c9]   File "/opt/stack/nova/nova/virt/libvirt/driver.py", line 1311, in swap_volume
[instance: efbde317-32e7-4b4a-bf80-f38b056578c9]     raise NotImplementedError(_("Swap only supports host devices"))
[instance: efbde317-32e7-4b4a-bf80-f38b056578c9] NotImplementedError: Swap only supports host devices
```

  - related blueprint
    - https://blueprints.launchpad.net/nova/+spec/in-use-volume-migrate-rbd

Volume Migration

- Summary
- Use Cases

NO	attachment	src driver	dest driver	volume type	migrate type	Result
1	available	rbd	rbd	same	host assisted	Perfect!
2	available	rbd	rbd	different	host assisted	Good! (need retype)
3	available	block based	ceph	different	host assisted	Possible but no recommend
		ceph	block based			
4	in-use	rbd	rbd	same	host assisted	Impossbile!

netmarble

SK telecom


Performance
Tuning
High
Availability
Volume
Migration
Volume
Replication

Volume Replication

- RBD Mirroring
  - asynchronously mirrored between two ceph clusters
  - uses the RBD journaling image feature to ensure crash-consistent replication
  - 2 ways
    - per pool : automatically mirrored
    - image : specific subset of images
  - Use cases
    - disaster recovery
    - global block device distribution

Volume Replication

- Replication in Cinder
  - depends on the driver's implementation
  - There is no automatic failover since the use case is Disaster Recovery, and it must be done manually when the primary backend is out.
- Ceph Replication in Cinder

Step	To Do
1	prepare diffrent ceph clusters
2	configure ceph clusters in mirror mode and to mirror the pool used by cinder
3	copy cluster key to the cinder volume node
4	configure ceph driver in cinder to use replication

Volume Replication

- Install ceph-mirror
- enable ceph mirror services on all clusters

```
# apt / yum install ceph-mirror
# systemctl enable ceph-rbd-mirror@admin
# systemctl start ceph-rbd-mirror@admin
```

- enable volumes pool on all clusters

```
# rbd mirror pool enable volumes image
```

  - pool: When configured in pool mode, all images in the pool with the journaling feature enabled are mirrored.
  - image: When configured in image mode, mirroring needs to be explicitly enabled on each image.

Volume Replication

- Copy the configuration information of the other cluster to the primary and secondary nodes respectively.
  - in primary

```
# scp /etc/ceph/ceph.conf {secodary}:/etc/ceph/ceph-primary.conf
# scp /etc/ceph/ceph.client.admin.keyring {secondary}:/etc/ceph/ceph-primary.client.admin.keyring
```

  - in secondary

```
# scp /etc/ceph/ceph.conf {primary}:/etc/ceph/ceph-secondary.conf
# scp /etc/ceph/ceph.client.admin.keyring {primary}:/etc/ceph/ceph-secondary.client.admin.keyring
```

Volume Replication

- Peering
  - in primary

```
root@cluster001:~ rbd mirror pool peer add volumes client.admin@ceph-secondary
5d4fbcdb-c7e5-4966-9c24-fdfcf4413b28
root@cluster001:~# rbd mirror pool status volumes
health: OK
images: 0 total
```

  - in secondary

```
root@cluster002:~# rbd mirror pool peer add volumes client.admin@ceph-primary
d6ec5046-becd-4a06-9ad2-8f18cb396e08
root@cluster002:~# rbd mirror pool status volumes
health: OK
images: 0 total
```

  - cinder.conf

```
[ceph-1]
replication_device = backend_id:cluster002, conf:/etc/ceph/ceph2.conf, user:cinder2
volume_driver = cinder.volume.drivers.rbd.RBDDriver
volume_backend_name = CEPH
rbd_user = cinder1
rbd_secret_uuid = 0091c095-7417-4296-96c4-ce8343df92e9
rbd_pool = volumes
rbd_ceph_conf = /etc/ceph/ceph1.conf
```

Volume Replication

- Replication type

```
stack@openstack001:~$ cinder type-show replicated
+---------------------------------+--------------------------------------------------------------------------------+
| Property                        | Value                                                                          |
+---------------------------------+--------------------------------------------------------------------------------+
| description                     | None                                                                           |
| extra_specs                     | {'replication_enabled': '<is> True', 'volume_backend_name': 'CEPH'}            |
| id                              | 6ec681e3-7683-4b6b-bb1e-d460fe202fee                                           |
| is_public                       | True                                                                           |
| name                            | replicated                                                                     |
| os-volume-type-access:is_public | True                                                                           |
| qos_specs_id                    | None                                                                           |
+---------------------------------+--------------------------------------------------------------------------------+
```

- Create Replication Type Volume

```
stack@openstack001:~$ cinder create --volume-type replicated --name replicated-ceph4 1
+---------------------------------+--------------------------------------+
| Property                        | Value                                |
+---------------------------------+--------------------------------------+
| name                            | replicated-ceph4                     |
| replication_status              | None                                 |
| size                            | 1                                    |
| volume_type                     | replicated                           |
+---------------------------------+--------------------------------------+
```

Volume Replication

- cinder show

```
stack@openstack001:~$ cinder show replicated-ceph4
+---------------------------------+----------------------------------------+
| Property                        | Value                                  |
+---------------------------------+----------------------------------------+
| name                            | replicated-ceph4                       |
| os-vol-host-attr:host           | openstack001@ceph-1#CEPH               |
| os-vol-mig-status-attr:migstat  | None                                   |
| os-vol-mig-status-attr:name_id  | None                                   |
| os-vol-tenant-attr:tenant_id    | e6a2c4e409704845b73583856429de44       |
| replication_status              | enabled                                |
| size                            | 1                                      |
| snapshot_id                     | None                                   |
| source_volid                    | None                                   |
| status                          | available                              |
| updated_at                      | 2017-04-30T15:51:53.000000             |
| user_id                         | aee19624003a434ead55c4c5209854e5       |
| volume_type                     | replicated                             |
+---------------------------------+----------------------------------------+
```

Volume Replication

- Cluster 1

```
root@cluster001:~# rbd -p volumes info volume-aab62739-b544-454a-a219-12d9b4006372
rbd image 'volume-aab62739-b544-454a-a219-12d9b4006372':
size 1024 MB in 256 objects
order 22 (4096 kB objects)
block_name_prefix: rbd_data.3a9c3a810770
format: 2
features: layering, exclusive-lock, object-map, fast-diff, deep-flatten, journaling
journal: 3a9c3a810770
mirroring state: enabled
mirroring global id: 82eaa5f2-be3d-4954-a5fe-d14477fb5fed
mirroring primary: true
```

```
root@cluster001:~# rbd -p volumes mirror image status volume-aab62739-b544-454a-a219-12d9b4006372
  global_id:   82eaa5f2-be3d-4954-a5fe-d14477fb5fed
  state:       up+stopped
  description: remote image is non-primary or local image is primary
  last_update: 2017-05-01 00:57:19
```

netmarble

SK telecom

Volume Replication

- Cluster 2

```
root@cluster001:~# rbd -p volumes info volume-aab62739-b544-454a-a219-12d9b4006372
rbd image 'volume-aab62739-b544-454a-a219-12d9b4006372':
size 1024 MB in 256 objects
order 22 (4096 kB objects)
block_name_prefix: rbd_data.3a9c3a810770
format: 2
features: layering, exclusive-lock, object-map, fast-diff, deep-flatten, journaling
journal: 3a9c3a810770
mirroring state: enabled
mirroring global id: 82eaa5f2-be3d-4954-a5fe-d14477fb5fed
mirroring primary: true
```

```
root@cluster002:/etc/ceph# rbd -p volumes mirror image status volume-aab62739-b544-454a-a219-12d9b4006372
  global_id:   82eaa5f2-be3d-4954-a5fe-d14477fb5fed
  state:       up+replaying
  description: replaying, master_position=[object_number=3, tag_tid=1, entry_tid=3],
mirror_position=[object_number=3, tag_tid=1, entry_tid=3], entries_behind_master=0
```

Volume Replication

- Before Fail-Over

```
stack@openstack001:~$ cinder service-list --binary cinder-volume --withreplication
+---------------+-------------------------+------+---------+-------+----------------------------+-------------------+-------------------+--------+-----------------+
| Binary        | Host                    | Zone | Status  | State | Updated_at                 | Replication Status | Active Backend ID | Frozen | Disabled Reason |
+---------------+-------------------------+------+---------+-------+----------------------------+-------------------+-------------------+--------+-----------------+
| cinder-volume | openstack001@ceph-1     | nova | enabled | up    | 2017-04-30T16:12:52.000000 | enabled           | -                 | False  | -               |
+---------------+-------------------------+------+---------+-------+----------------------------+-------------------+-------------------+--------+-----------------+
```

- After Fail-Over

```
stack@openstack001:~$ cinder failover-host openstack001@ceph-1
stack@openstack001:~$ cinder service-list --binary cinder-volume --withreplication
+---------------+-------------------------+------+---------+-------+----------------------------+-------------------+-------------------+--------+-----------------+
| Binary        | Host                    | Zone | Status  | State | Updated_at                 | Replication Status | Active Backend ID | Frozen | Disabled Reason |
+---------------+-------------------------+------+---------+-------+----------------------------+-------------------+-------------------+--------+-----------------+
| cinder-volume | openstack001@ceph-1     | nova | enabled | up    | 2017-04-30T16:16:53.000000 | failing-over      | -                 | False  | -               |
+---------------+-------------------------+------+---------+-------+----------------------------+-------------------+-------------------+--------+-----------------+
```

netmarble

SK telecom

Volume Replication

- Cinder List

```
stack@openstack001:~$ cinder list
+--------------------------------------+-----------+------------------+------+-------------+----------+-----------
| ID                                   | Status    | Name             | Size | Volume Type | Bootable | Attached to |
+--------------------------------------+-----------+------------------+------+-------------+----------+-------------+
| 7d886f63-3d00-4919-aa40-c89ce78b76e2 | error     | normal-ceph      | 1    | ceph        | false    |             |
| aab62739-b544-454a-a219-12d9b4006372 | available | replicated-ceph4 | 1    | replicated  | false    |             |
+--------------------------------------+-----------+------------------+------+-------------+----------+------+
```

- Cluster 1 & Cluster 2

```
root@cluster001:~# rbd -p volumes mirror image status
volume-aab62739-b544-454a-a219-12d9b4006372
volume-aab62739-b544-454a-a219-12d9b4006372:
  global_id:   82eaa5f2-be3d-4954-a5fe-d14477fb5fed
  state:       up+replaying
  description: replaying, master_position=[],
mirror_position=[], entries_behind_master=0
  last_update: 2017-05-01 01:17:49
```

```
root@cluster002:~# rbd -p volumes mirror image status
volume-aab62739-b544-454a-a219-12d9b4006372
volume-aab62739-b544-454a-a219-12d9b4006372:
  global_id:   82eaa5f2-be3d-4954-a5fe-d14477fb5fed
  state:       up+stopped
  description: remote image is non-primary or local image
is primary
  last_update: 2017-05-01 01:17:41
```

netmarble

SK telecom


Performance
Tuning
TIPs
&
Tricks
Volume
Replication

Tips & Tricks

- **Maximum Volume in single VM**
  - Normally, alphabets order, 26 devices
  - nova/virt/libvirt/blockinfo.py

```
def get_dev_count_for_disk_bus(disk_bus):
    """Determine the number disks supported.

       Determine how many disks can be supported in
       a single VM for a particular disk bus.

       Returns the number of disks supported.
    """

    if disk_bus == "ide":
        return 4
    else:
        return 26
```

Tips & Tricks

- **Maximum Volume in single VM**
  - If you try to attach 27 volume to single VM, Error!
  - nova-compute.log

```
  File "/opt/stack/nova/nova/compute/manager.py", line 1493, in _get_device_name_for_instance
    instance, bdms, block_device_obj)
  File "/opt/stack/nova/nova/virt/libvirt/driver.py", line 7873, in get_device_name_for_instance
    block_device_obj, mapping=instance_info['mapping'])
  File "/opt/stack/nova/nova/virt/libvirt/blockinfo.py", line 395, in get_info_from_bdm
    device_name = find_disk_dev_for_disk_bus(padded_mapping, bdm_bus)
  File "/opt/stack/nova/nova/virt/libvirt/blockinfo.py", line 195, in find_disk_dev_for_disk_bus
    raise exception.InternalError(msg)
InternalError: No free disk device names for prefix 'vd'
```

Tips & Tricks

- **VM hangs during a number of volumes attaching**
  - issue when 10 volumes attached a VM
  - VM hang during mkfs to 8th volume in VM
  - cannot kill the process

Tips & Tricks

- VM Kernel Log

```
INFO: task beaver:1563 blocked for more than 120 seconds.
Not tainted 2.6.32-642.6.2.el6.x86_64 #1
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
beaver D 0000000000000000 0 1563 1557 0x00000080
ffff88043487fad8 0000000000000082 0000000000000000 0000000000000082
ffff88043487fa98 ffffffff8106c36e 00000007e69e1630 ffff880400000000
000000003487fb28 00000000fffbf0b1 ffff8804338cf068 ffff88043487ffd8
Call Trace:
[<ffffffff8106c36e>] ? try_to_wake_up+0x24e/0x3e0
[<ffffffffa006f09d>] do_get_write_access+0x29d/0x520 [jbd2]
[<ffffffff810a6920>] ? wake_bit_function+0x0/0x50
[<ffffffff8123bf10>] ? security_inode_alloc+0x40/0x60
[<ffffffffa006f471>] jbd2_journal_get_write_access+0x31/0x50 [jbd2]
[<ffffffffa00bcfa8>] __ext4_journal_get_write_access+0x38/0x80 [ext4]
[<ffffffffa0092724>] ext4_new_inode+0x414/0x11c0 [ext4]
[<ffffffffa006e3d5>] ? jbd2_journal_start+0xb5/0x100 [jbd2]
[<ffffffffa00a1540>] ext4_create+0xc0/0x150 [ext4]
[<ffffffff811a7153>] ? generic_permission+0x23/0xb0
[<ffffffff811a9456>] vfs_create+0xe6/0x110
[<ffffffff811ad26e>] do_filp_open+0xa8e/0xd20
[<ffffffff811ec0f3>] ? __posix_lock_file+0xa3/0x4e0
[<ffffffff811ec6c5>] ? fcntl_setlk+0x75/0x320
[<ffffffff812a885a>] ? strncpy_from_user+0x4a/0x90
[<ffffffff811ba072>] ? alloc_fd+0x92/0x160
[<ffffffff811969f7>] do_sys_open+0x67/0x130
[<ffffffff81196b00>] sys_open+0x20/0x30
[<ffffffff8100b0d2>] system_call_fastpath+0x16/0x1b
```

Tips & Tricks

- QEMU fd Limit
  - default value is 1024
  - After attaching the volume to the VM, mkfs make increase the number of connections between compute host and ceph
  - Number of FD is large as many disks
  - VM hangs when the connections over fd limits
    - [ceph-users] libvirt rbd issue: http://lists.ceph.com/pipermail/ceph-users-ceph.com/2015-September/004115.html
    - [ceph-users] mkfs.ext4 hang on RBD volume: http://lists.ceph.com/pipermail/ceph-users-ceph.com/2017-January/015775.html

Tips & Tricks

- How to Resolve
  - Edit /etc/libvirtd/qemu.conf

```
max_files = <your fd>
```

  - Estimate value
    - more than <default fd per VM> + <quota about how many volume to one vm> * <fd per Volume>

A. FD per new VM (approximately)	B. Volume per VM	C. Fd per Volume	A + (B * C)< max_files
150	10	100	150 + (10 * 100) = 1150 < max files

  - service restart

```
Systemctl restat libvirtd
```

Tips & Tricks

- rbd cache in OpenStack
  - In Ceph manual, It recommend add ‘rbd cache’ parameter in ceph.conf at every compute node.

CONFIGURING NOVA

In order to boot all the virtual machines directly into Ceph, you must configure the ephemeral backend for Nova.

It is recommended to enable the RBD cache in your Ceph configuration file (enabled by default since Giant). Moreover, enabling the admin socket brings a lot of benefits while troubleshooting. Having one socket per virtual machine using a Ceph block device will help investigating performance and/or wrong behaviors.

This socket can be accessed like this:

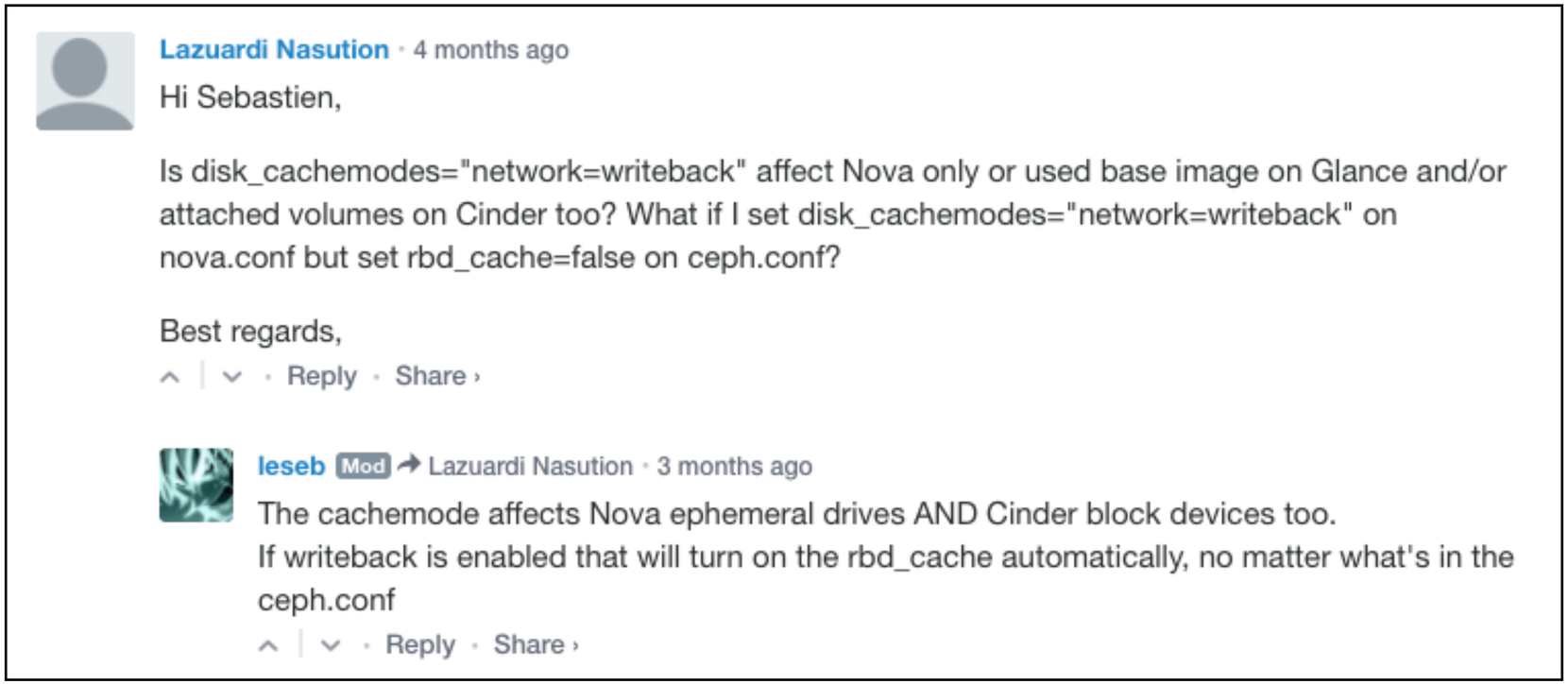
```
ceph daemon /var/run/ceph/ceph-client.cinder.19195.32310016.asok help
```

Now on every compute nodes edit your Ceph configuration file:

```
[client]
    rbd cache = true
    rbd cache writethrough until flush = true
    admin socket = /var/run/ceph/guests/$cluster-$type.$id.$pid.$cctid.asok
    log file = /var/log/qemu/qemu-guest-$pid.log
    rbd concurrent management ops = 20
```

Tips & Tricks

- rbd cache in OpenStack
  - But, It didn't relative, Just "disk_cachemodes=network:writeback" in nova.conf



Lazuardi Nasution · 4 months ago

Hi Sebastien,

Is disk_cachemodes="network=writeback" affect Nova only or used base image on Glance and/or attached volumes on Cinder too? What if I set disk_cachemodes="network=writeback" on nova.conf but set rbd_cache=false on ceph.conf?

Best regards,

Reply · Share ›

leseb Mod ➦ Lazuardi Nasution · 3 months ago

The cachemode affects Nova ephemeral drives AND Cinder block devices too.
If writeback is enabled that will turn on the rbd_cache automatically, no matter what's in the ceph.conf

Reply · Share ›

https://www.sebastien-han.fr/blog/2013/08/22/configure-rbd-caching-on-nova/

Tips & Tricks

- rbd cache in OpenStack
  - Just “disk_cachemodes=network:writeback” in nova.conf

```
[libvirt]
disk_cachemodes="network=writeback"
```

  - Restart nova-compute
  - Create VM, check the libvirt XML, nova-compute add cache parameter.

Tips & Tricks

- rbd cache in OpenStack
  - VM XML
    - before

```
  <devices>
    <emulator>/usr/bin/kvm</emulator>
    <disk type='network' device='disk'>
      <driver name='qemu' type='raw' cache='none'/>
```

    - after

```
<devices>
    <emulator>/usr/bin/kvm</emulator>
    <disk type='network' device='disk'>
      <driver name='qemu' type='raw' cache='writeback'/>
```

Tips & Tricks

- rbd cache in OpenStack
  - Performance Competition: It is as expected.
    - Before

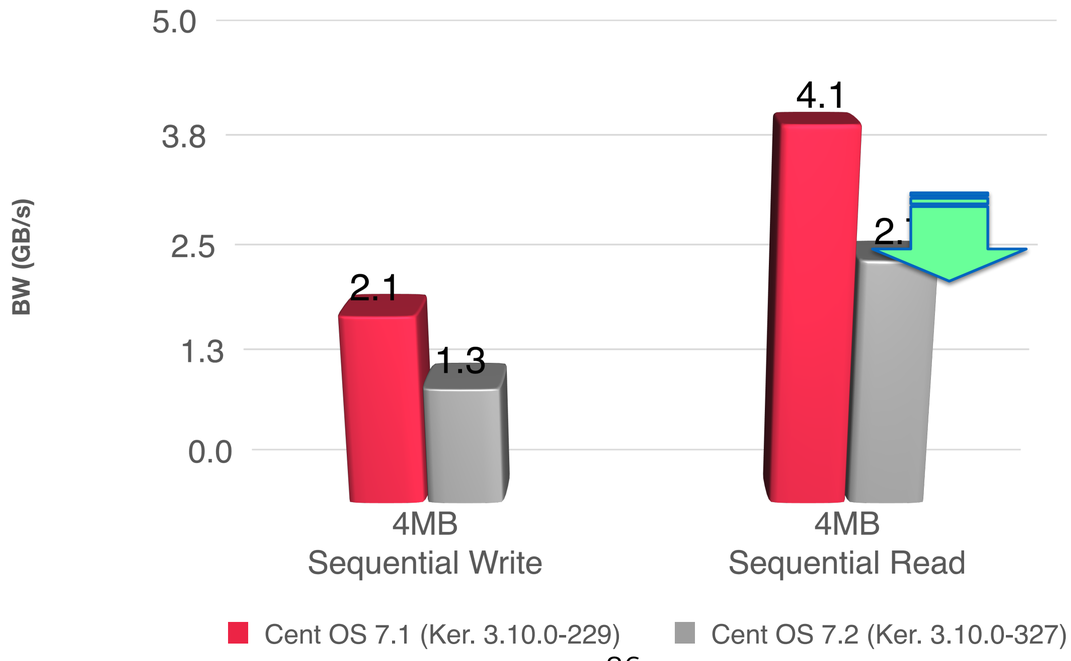
```
[centos@test-attach-vm03 here]$ dd if=/dev/zero of=here bs=32K count=100000
oflag=direct
100000+0 records in
100000+0 records out
3276800000 bytes (3.3 GB) copied, 746.158 s, 4.4 MB/s
```

    - After

```
[centos@test-attach-vm03 here]$ dd if=/dev/zero of=here bs=32K count=100000
oflag=direct
100000+0 records in
100000+0 records out
3276800000 bytes (3.3 GB) copied, 33.9688 s, 96.5 MB/s
```

Tips & Tricks

- **TCP Congestion Control Algorithm Issue**
  - Cent OS 7.1 (3.10.0-229 Kernel) VS Cent OS 7.2 (3.10.0-327 Kernel), total throughput is reduced by about 40%
  - Due to change in the TCP Congestion Control Algorithm in Linux Kernel



Tips & Tricks

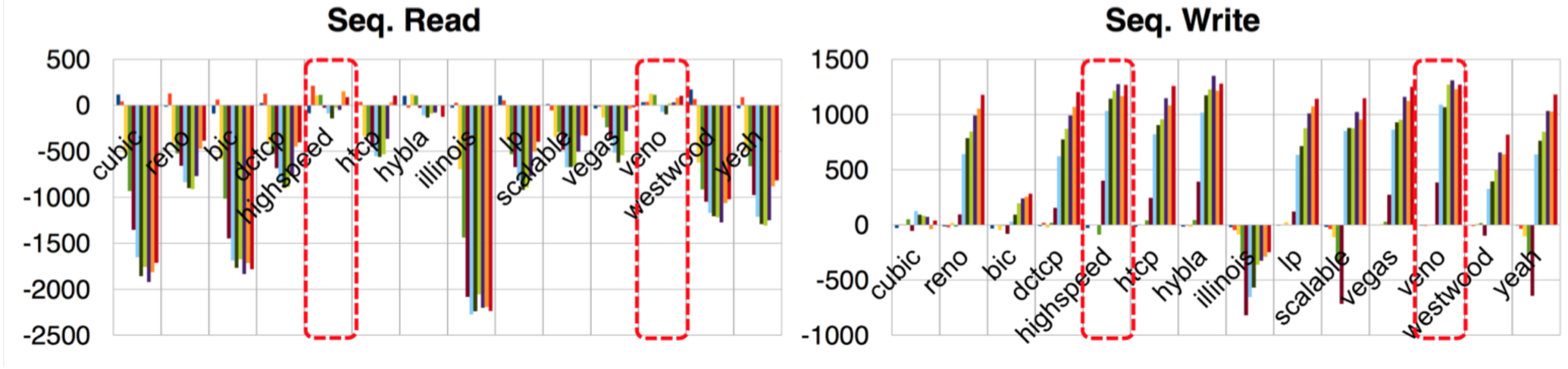
- **TCP Congestion Control Algorithm Issue**
  - Symptom: When TCP congestion control algorithm that contains patch, stretch ACK, is used in a situation where large packet workload is loaded and the network is busy
  - TCP Congestion Control Algorithm with Stretch ACK: Cubic (Default), Reno
  - Highspeed / Veno algorithm can help to relieve performance degradation phenomenon, also get better performance when running large block size write workload

Tips & Tricks

- TCP Congestion Control Algorithm Issue
  - Test bandwidth of TCP Congestion Control Algorithm

```
# modprobe tcp_highspeed.ko
# echo highspeed > /proc/sys/net/ipv4/tcp_congestion_algorithm
```

  - Best Result: High speed / Veno algorithm



Tips & Tricks

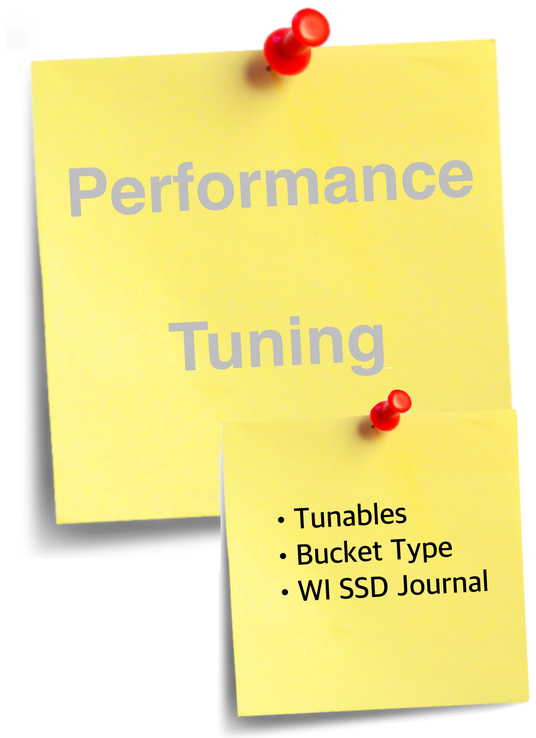
- **Host Failure**
  - OSD failure
    - reasons : disk dead, file system error
    - Solution
      - phisically remove the disk
      - Add new disk into the ceph cluster
  - Host failure
    - reasons : electric problem, network problem etc.. not because of disk failure
    - solution
      - power on the server

Tips & Tricks

- Subtree Limit Configuration
  - mon osd down out subtree limit
  - The smallest CRUSH unit type that ceph will not automatically mark out
  - default value is rack
- How to configure

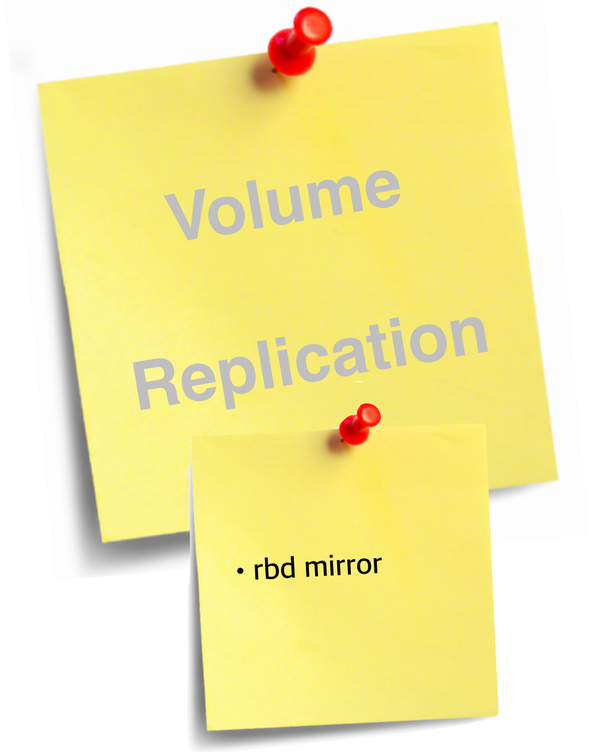
```
### in ceph.conf
[mon]
mon osd down out subtree limit = host

or

### runtime configuration

$ ceph tell mon.* injectargs ‘mon_osd_down_out_subtree_limit host’
```

Summary


Performance
Tuning
• Tunables
• Bucket Type
• WI SSD Journal


High
Availability
• Cinder-Volume
• Cinder-Backup


Volume
Migration
• Host-Assisted


Volume
Replication
• rbd mirror

Any Questions?

...

Thank you