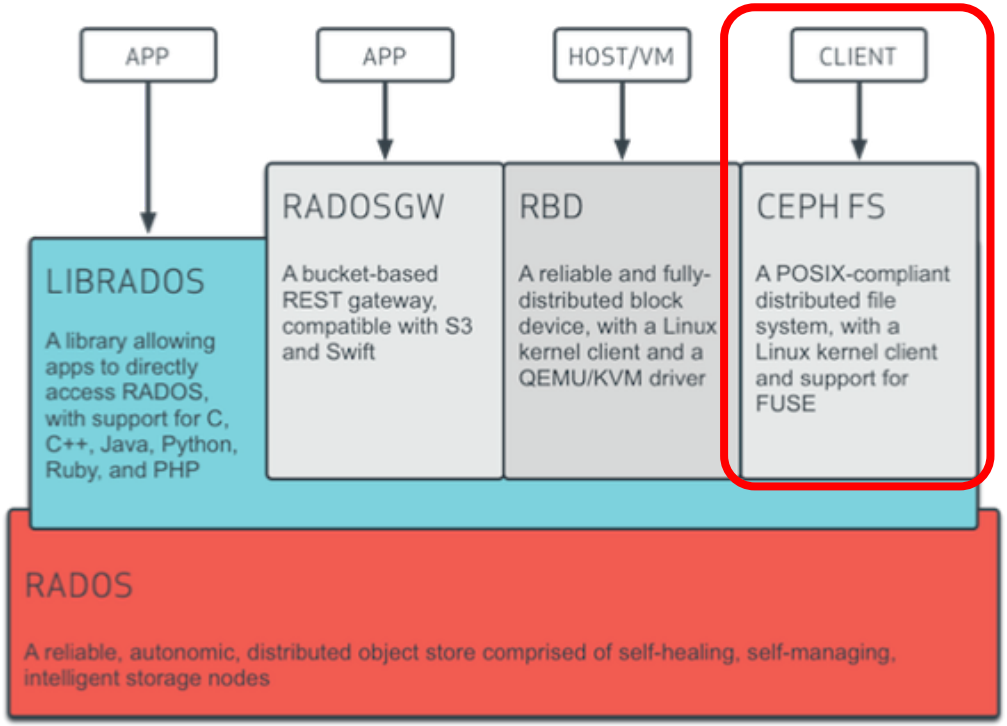
CephFS with OpenStack Manila based on Bluestore and Erasure Code

NAVER 유장선

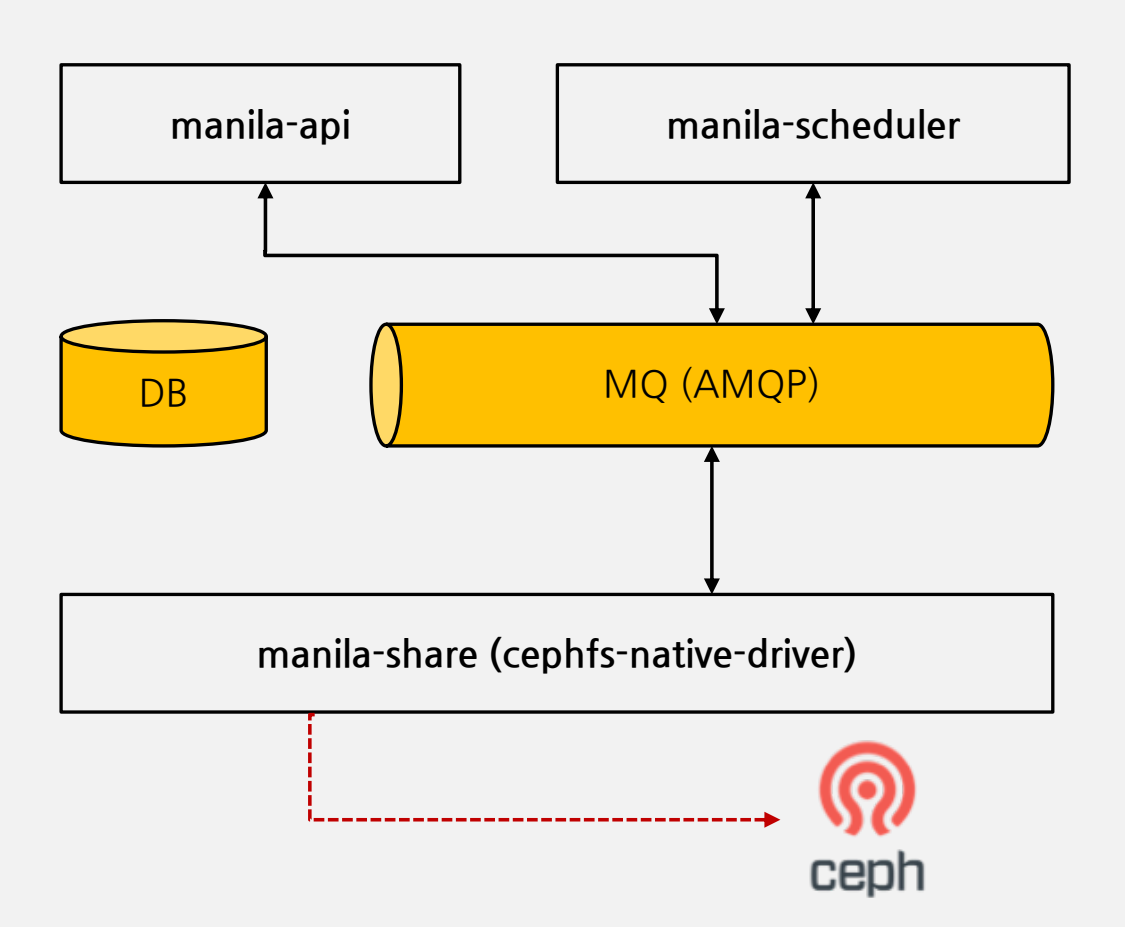
CephFS with OpenStack Manila based on Bluestore and Erasure Code



http://docs.ceph.com/docs/master/architecture/

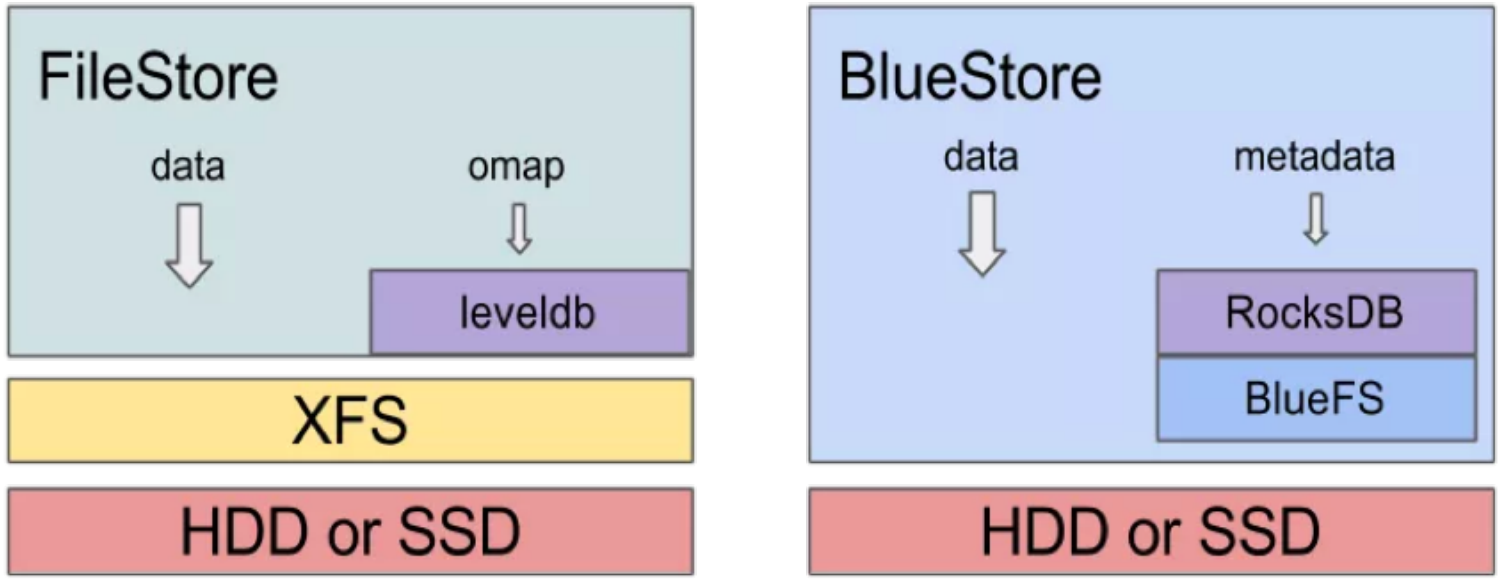
- POSIX-compliant shared FS
- Support kernel client and FUSE
- First Stable Release : Jewel (in April 2016)
- Multiple Active MDS : Luminous
- Directory Fragmentation
- Subtree Pinning
- Experimental Features
  - INLINE DATA
  - MANTLE : Programmable metadata LB
  - Snapshot
  - Multiple FileSystem

CephFS with *OpenStack Manila* based on Bluestore and Erasure Code



- OpenStack File Share Service
- Incubated project in Juno
- Core Service in Kilo
- share drivers : 20
- Create/Delete share
- Access Allow/Deny
- Quota
- Consistency Group
- Snapshot
- Share Replication

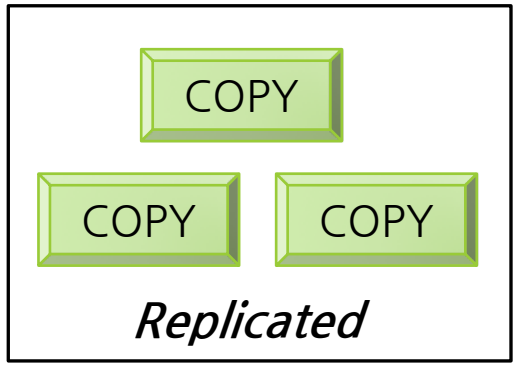
CephFS with OpenStack Manila based on *Bluestore* and Erasure Code



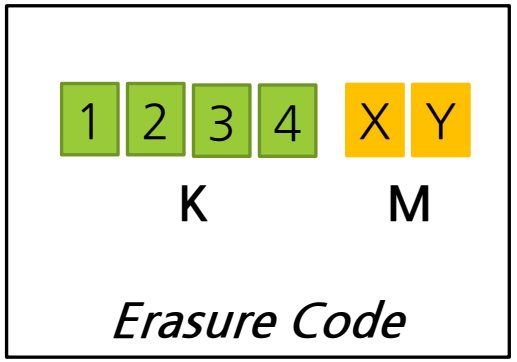
- BlueStore = Block + NewStore
- Consume raw block device
- RocksDB for metadata
- Luminous Default Data Store

https://ceph.com/community/new-luminous-bluestore/

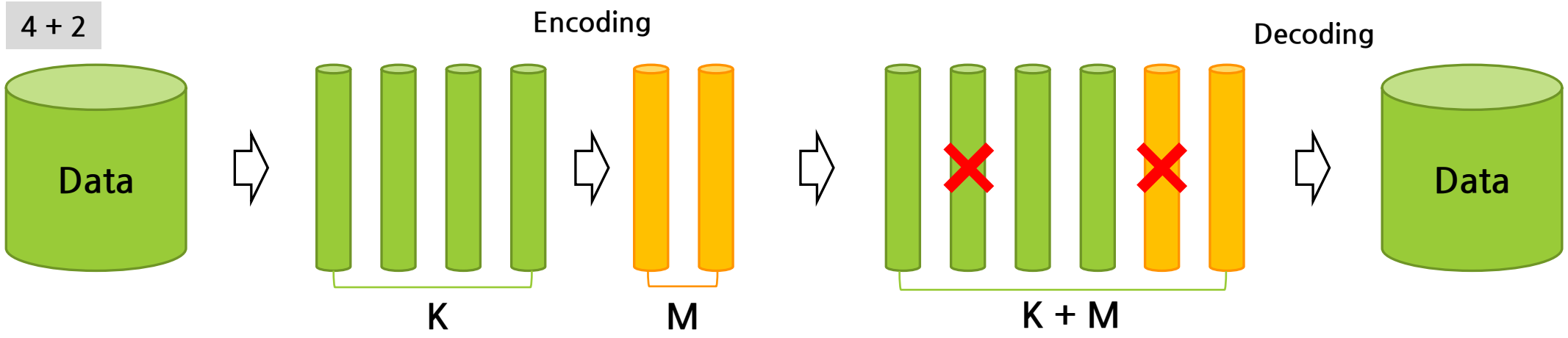
CephFS with OpenStack Manila based on Bluestore and *Erasure Code*



- Very high durability
- 200 % overhead
- Quick recovery



- Cost-Effective
- 50 % overhead
- Expensive Recovery



Erasure Code : allow_ec_overwrites

- Luminous Support
- RBD and CephFS on EC Pools
- Only be enabled on **BlueStore OSDs**
- Erasure coded pools do not support omap
- Needs **Metadata Pool with Replicated**

```
$ ceph osd pool set ec_pool allow_ec_overwrites true
```

```
$ rbd create --size 1G --data-pool ec_pool replicated_pool/image_name`
```

```
$ ceph fs new <fs_name> <metadata> <data>
```

```
$ setfattr -n ceph.file.layout.pool -v cephfs_data file2
```

Erasure Code : Profile

- K : Data-Chunks (4)
- M : Coding Chunks (2)
- Plugin : Jerasure / ISA / Locally repairable
- Technique : reed_sol_van / cauchy
- ruleset-failure-domain : rack / host / osd

“ISA only runs on Intel processors”

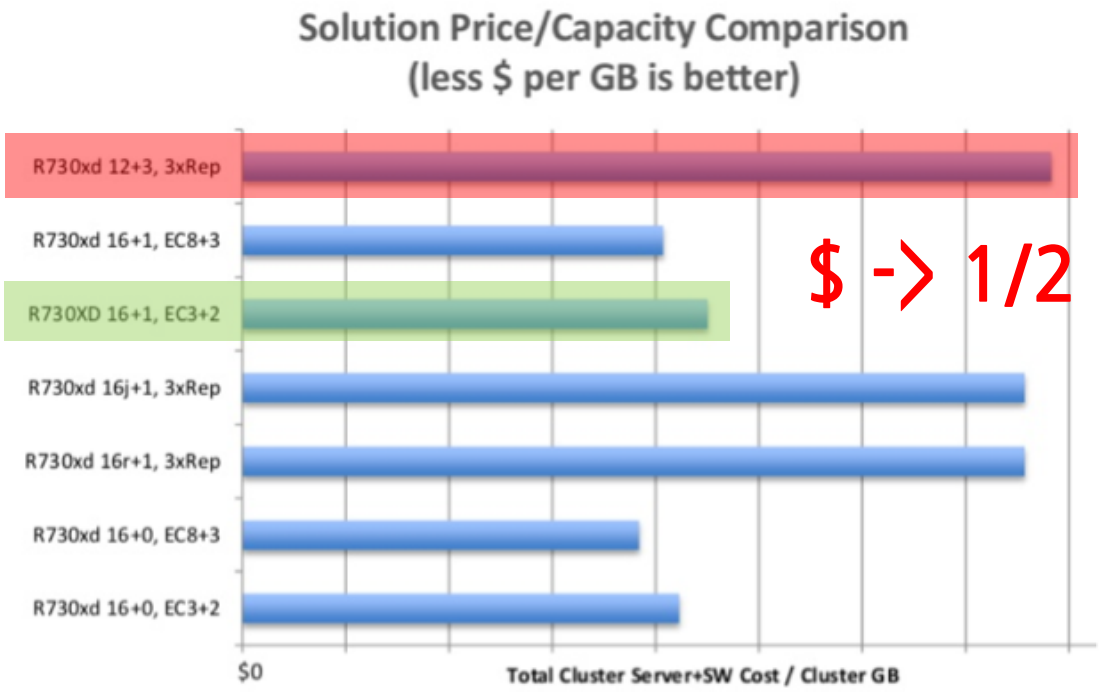
Plugin		**Jerasure**		**ISA**	
techniques		**reed_sol_van**	**cauchy_good**	**reed_sol_van**	**cauchy**
Encode Times(s)		1.140	1.039	0.574	**0.561**
Decode Times(s)	**1 OSD LOST**	0.521	0.522	**0.333**	0.404
	2 OSD LOST	1.416	1.113	0.557	**0.547**

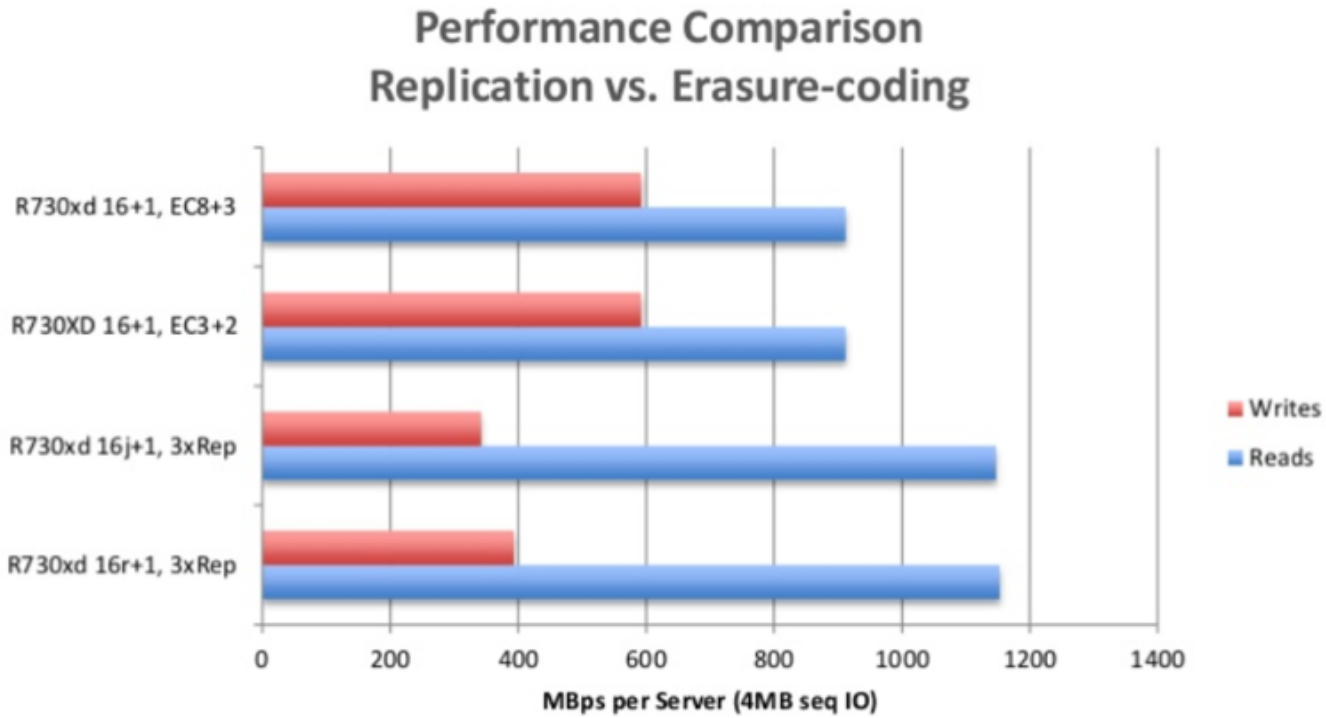
https://ceph.com/geen-categorie/benchmarking-ceph-erasure-code-plugins/

Replication vs. Erasure Coding

- EC is only half the cost ($ per GB) of rep.

- Replication better performance for read.
- Erasure Coding better performance for write.

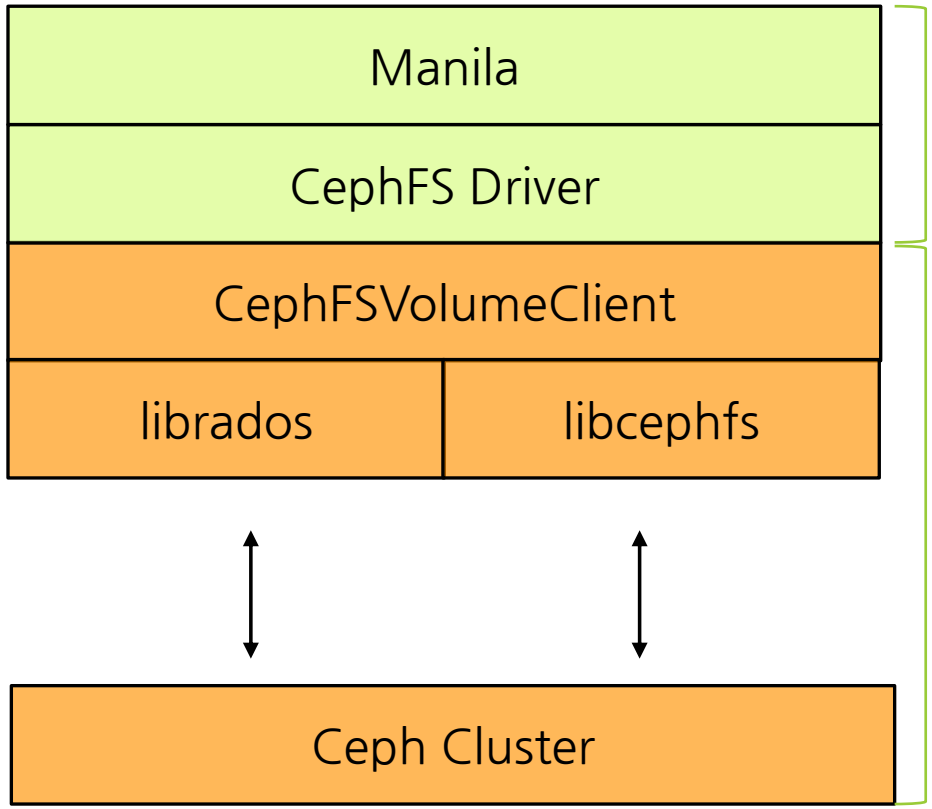




https://www.slideshare.net/JoseDeLaRosa7/ceph-perfsizingguide

In-Depth : OpenStack Manila

Manila - cephfs driver



github.com/opensteack/manila

github.com/ceph/ceph

https://github.com/ceph/ceph/blob/master/src/pybind/ceph_volume_client.py

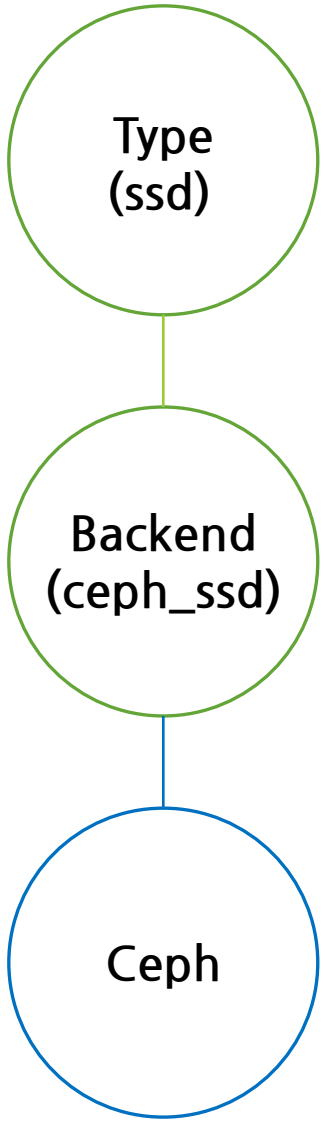
Create Type

$ manila type-create **{fs-name} {DHSS}**

$ manila type-set **{fs-name}** set share_backend_name=‘**{backend name}**'

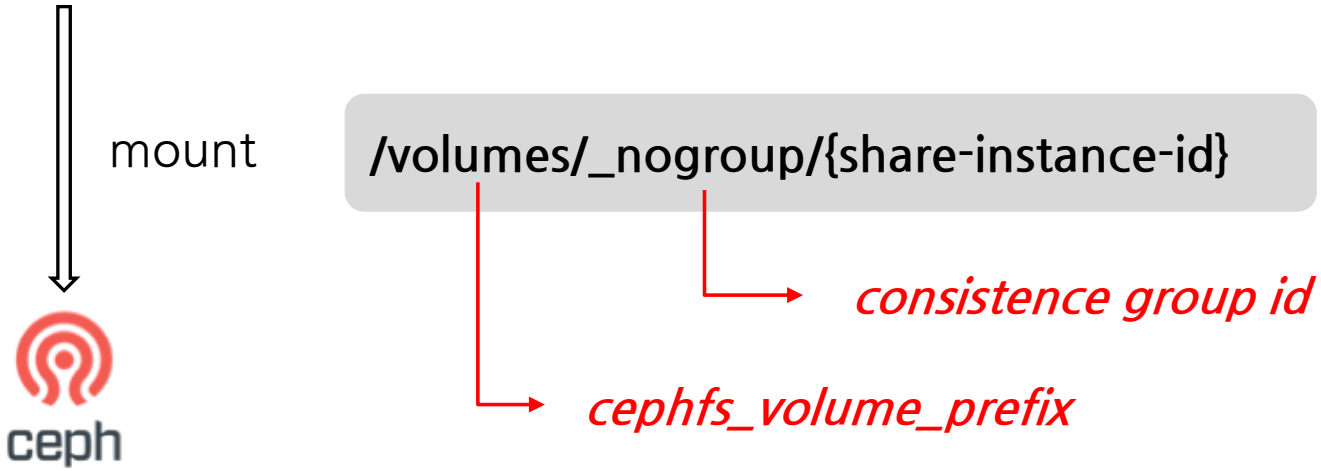
SAMPLE

[DEFAULT]
enabled_share_protocols = CEPHFS
enabled_share_backends = ceph_ssd
[ceph_ssd]
driver_handles_share_servers = False
share_backend_name = ceph_ssd
share_driver = **manila.share.drivers.cephfs.cephfs_native.CephFSNativeDriver**
cephfs_conf_path = /etc/ceph/ceph.conf
cephfs_auth_id = **manila**
cephfs_cluster_name = ceph
cephfs_enable_snapshots = **False**
cephfs_volume_prefix = /volumes
*cephfs_pool_namespace_prefix = fsvolumens_*



Create Share (Volume)

$ manila create --share-type {fs type} --name {fs name} **cephfs** {size:gb}

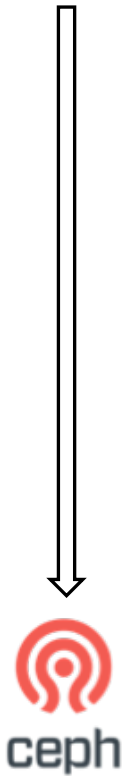


$ manila share-export-location-list {fs name}

ID	Path	Preferred
c1824b81-2eb2-47db-905c-103bf27e1978	10.10.10.10:6789:*/volumes/_nogroup/5e016f4f-2664-4afe-8f03-9d1afe065785*	False

Allowing access to shares

$ manila access-allow {fs name} **cephx** {user name}



ceph

[client.user-name]

key = AQA8+ANW/4ZWNRAAOtWJMFPEihBA1unFlmJczA==

caps: [mds] allow rw path=*/volumes/_nogroup/5e016f4f-2664-4afe-8f03-9d1afe065785*

caps: [mon] allow r

caps: [osd] allow rw pool=cephfs1-data ***namespace=****fsvolumes-fd43f701-e4d1-48c5-ad5a-2bb8f2daaaa0*

$ setfattr -n **ceph.file.layout.pool_namespace**

-v *fsvolumes-fd43f701-e4d1-48c5-ad5a-2bb8f2daaaa0*

/volumes/_nogroup/5e016f4f-2664-4afe-8f03-9d1afe065785

Mount

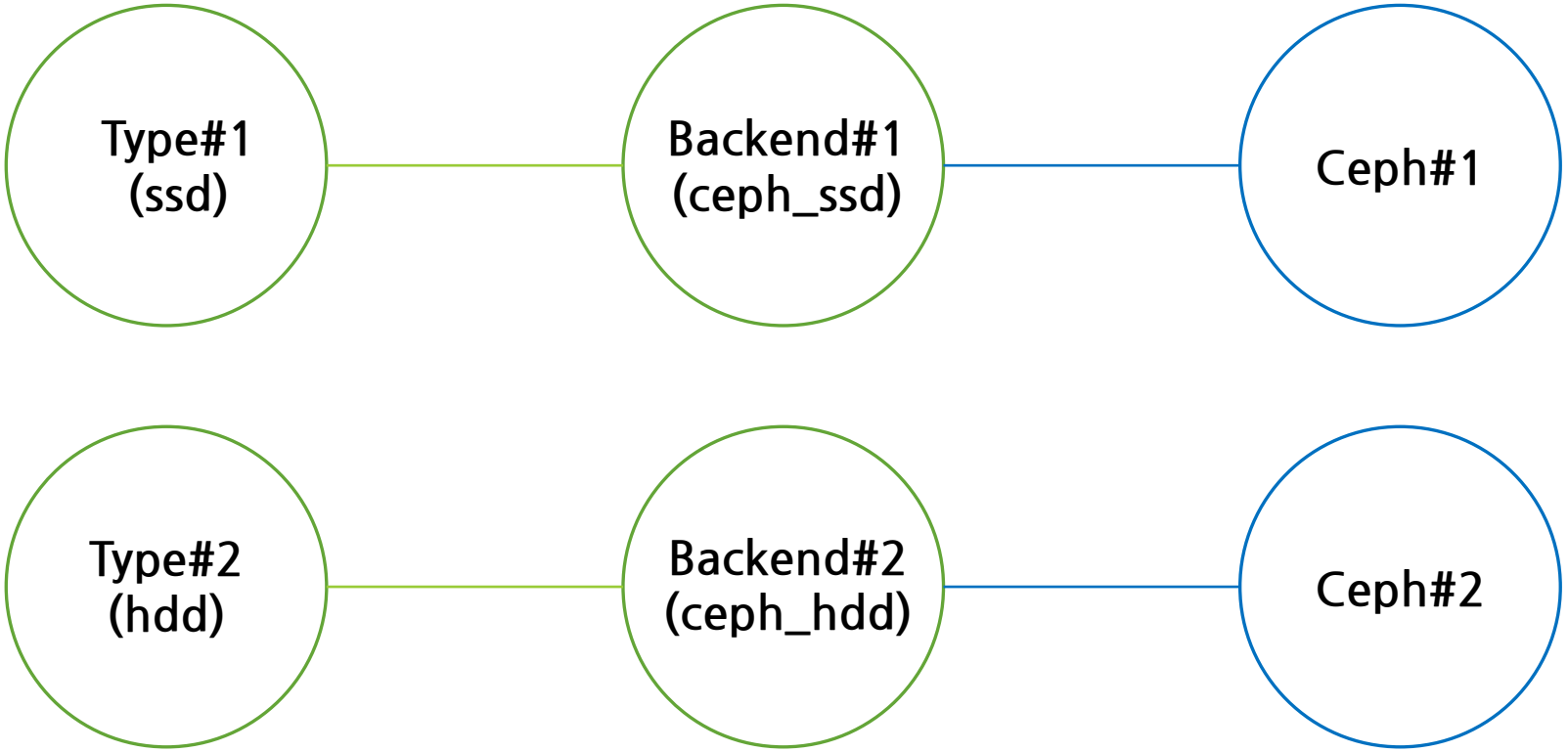
```
$ mount -t ceph {mon-ip}:{volume path} {mount path} -o name={user name},secret={secret}
```

```
$ manila share-export-location-list {fs name}
+--------------------------------------+--------------------------------------------------------------------------------------------+-----------+
| ID                                   | Path                                                                                       | Preferred |
+--------------------------------------+--------------------------------------------------------------------------------------------+-----------+
| c1824b81-2eb2-47db-905c-103bf27e1978 | 10.10.10.10:6789:/volumes/_nogroup/5e016f4f-2664-4afe-8f03-9d1afe065785                    | False     |
+--------------------------------------+--------------------------------------------------------------------------------------------+-----------+
```

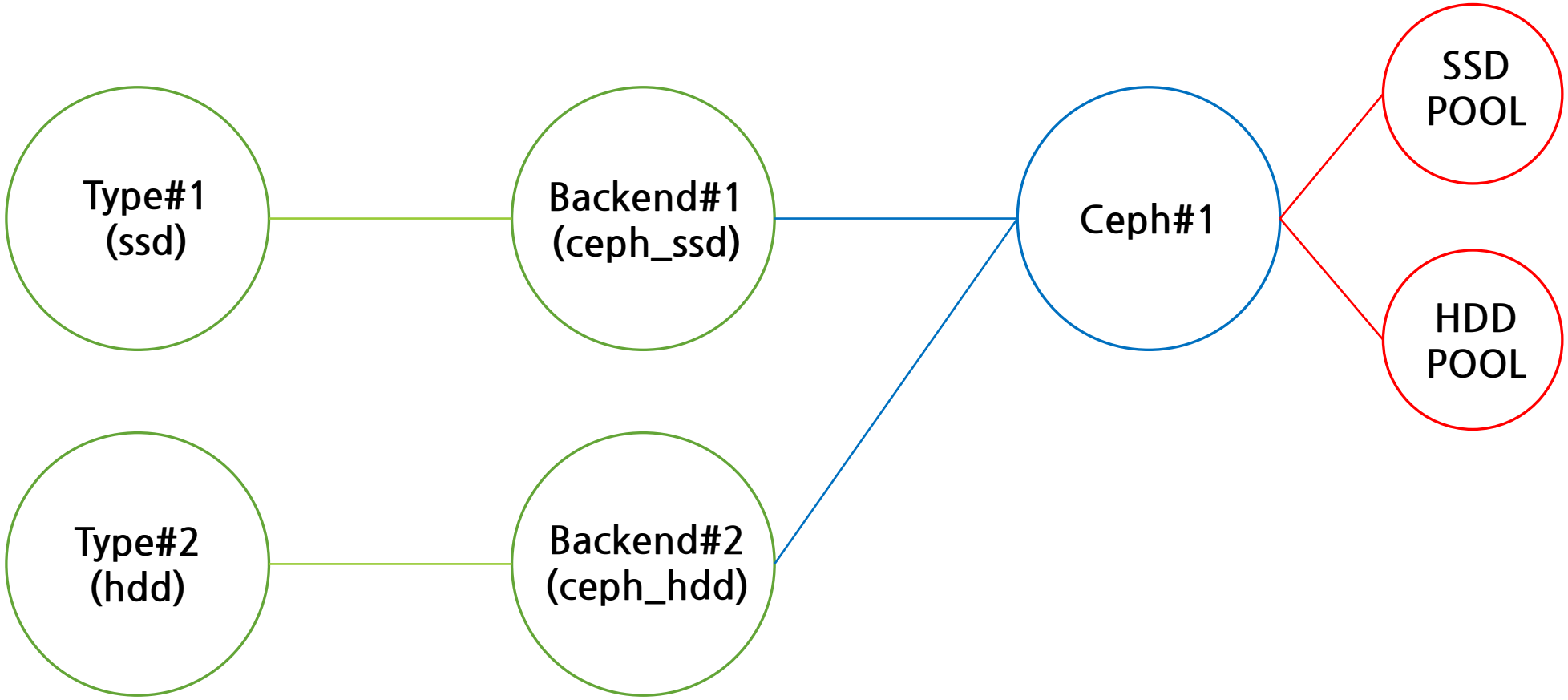
```
[client.user-name]
        key = AQA8+ANW/4ZWNRAAOtWJMFPEihBA1unFlmJczA==
        caps: [mds] allow rw path=/volumes/_nogroup/5e016f4f-2664-4afe-8f03-9d1afe065785
        caps: [mon] allow r
        caps: [osd] allow rw pool=cephfs1-data namespace=fsvolumes-fd43f701-e4d1-48c5-ad5a-2bb8f2daaaa0
```

```
$ mount -t ceph 10.10.10.10:6789:/volumes/_nogroup/5e016f4f-2664-4afe-8f03-9d1afe065785
/mnt -o name={user name},secret=AQA8+ANW/4ZWNRAAOtWJMFPEihBA1unFlmJczA==
```

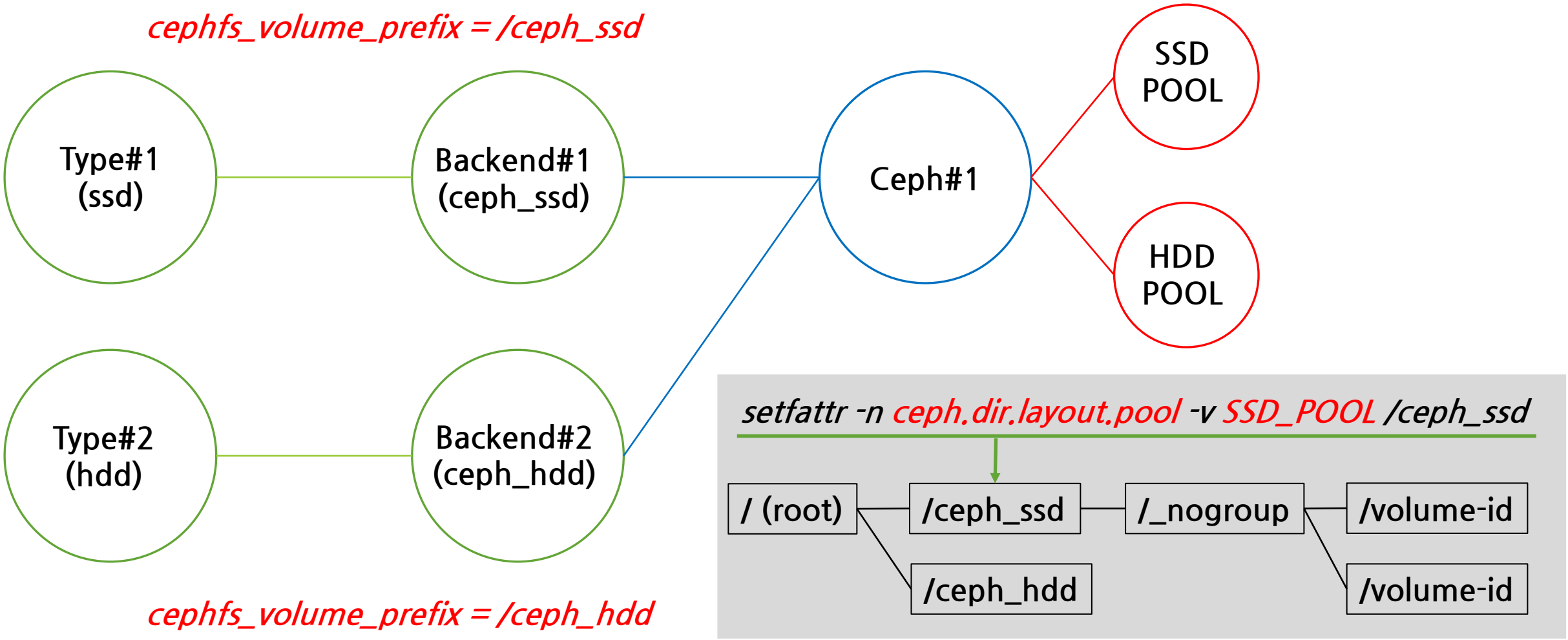
Multi-Backend (current)



Multi-Backend (our goal)



Multi-Backend (our goal)

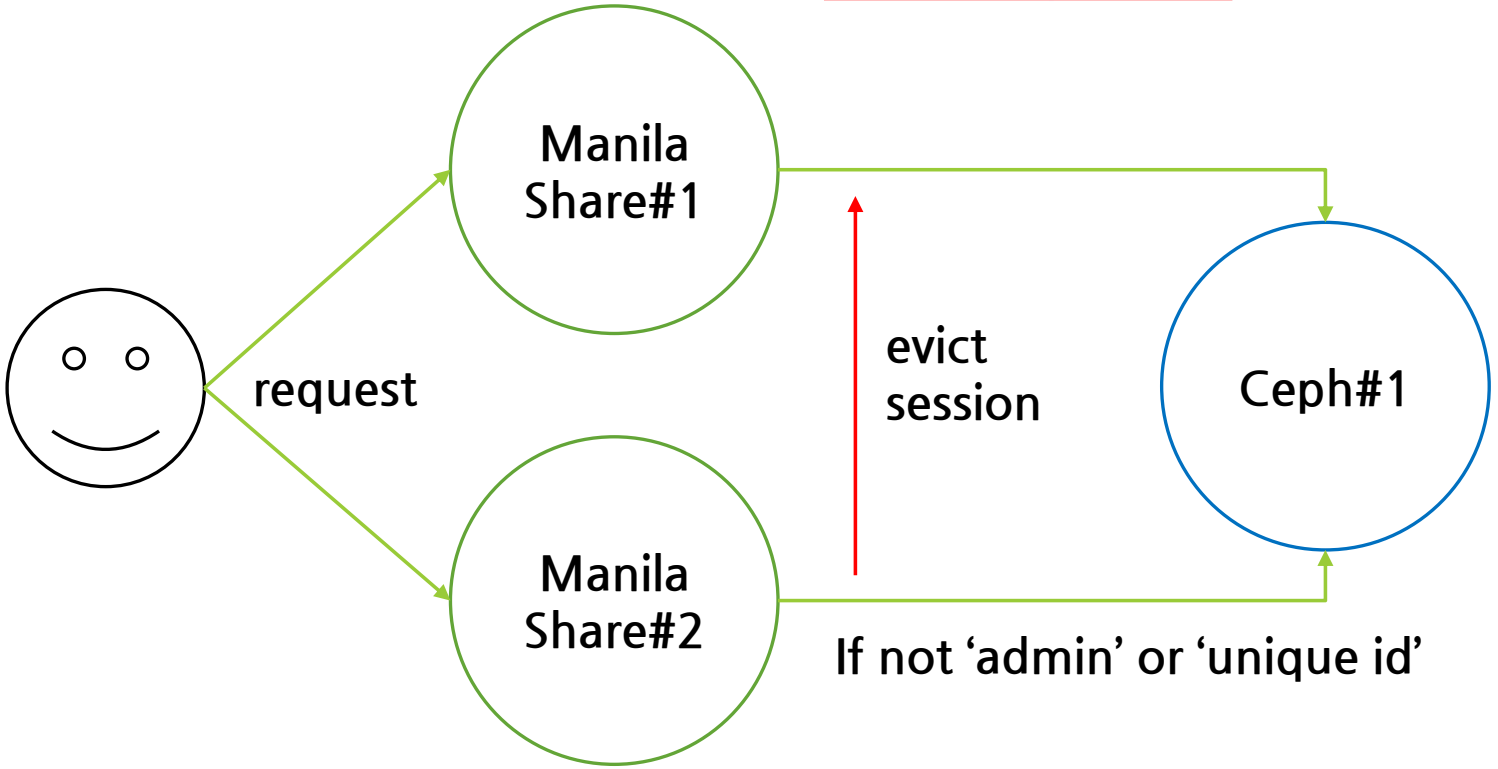


https://review.openstack.org/#/c/572022/

Eviction Issue

Known restrictions

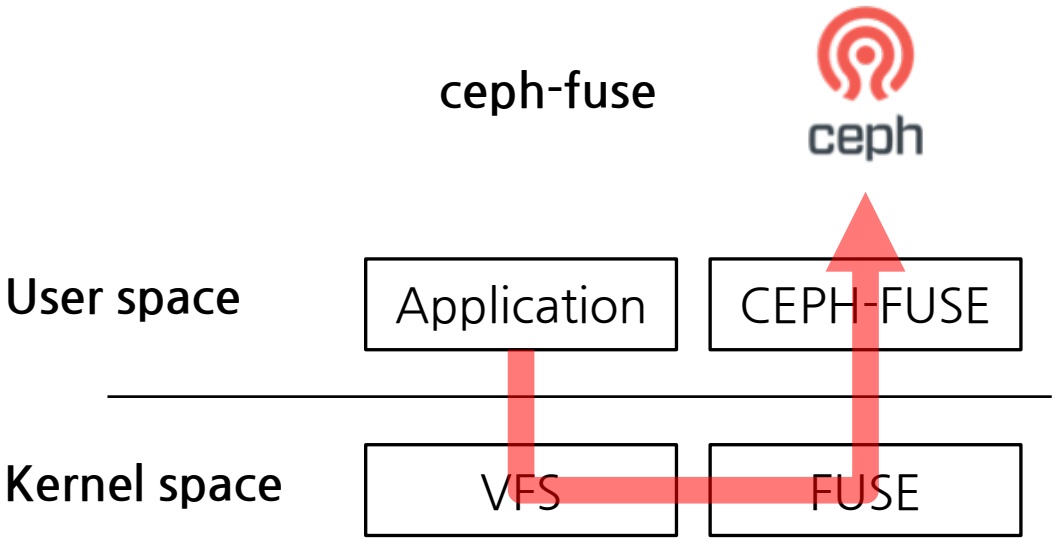
- A CephFS driver instance, represented as a backend driver section in manila.conf, requires a Ceph auth ID unique to the backend Ceph Filesystem. Using a non-unique Ceph auth ID will result in the driver unintentionally evicting other CephFS clients using the same Ceph auth ID to connect to the backend.

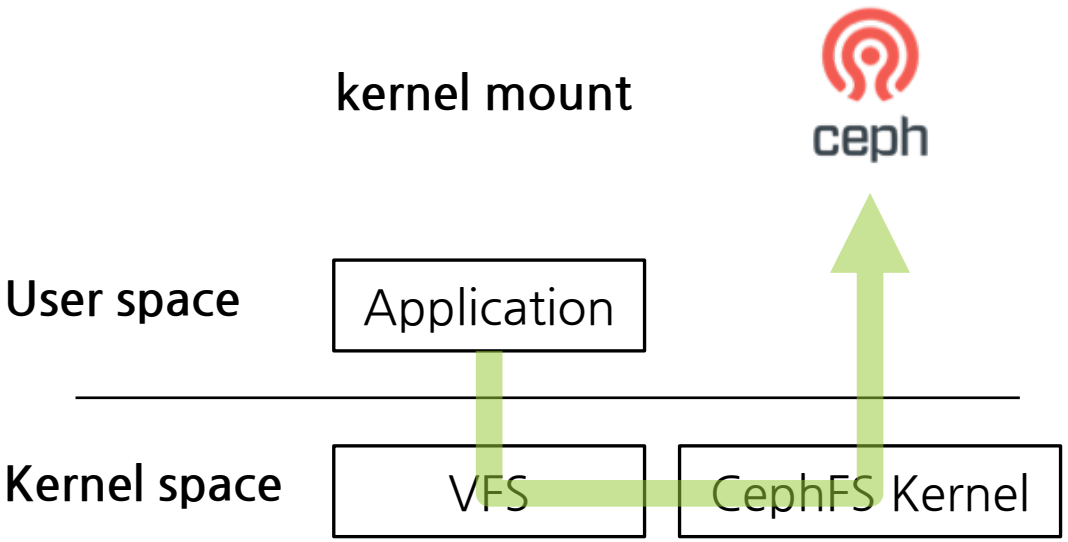


[backend name]
…
cephfs_auth_id = **manila**
…

https://docs.openstack.org/manila/queens/admin/cephfs_driver.html#known-restrictions

ceph-fuse vs kernel mount





Support Quotas

Fast

Quotas

- File Limit : ceph.quota.max_files
- Byte Limit : ceph.quota.max_bytes

```
$ setfattr -n ceph.quota.max_files -v {max files} {/path}
$ setfattr -n ceph.quota.max_bytes -v {max bytes} {/path}
```

ceph-fuse

```
$ df -h | grep fuse
ceph-fuse     100G    0 100G 0% /cephfs
```

kernel mount

```
$ df -h | grep cephfs
10.10.10.10:6789:/test 16T 7.9T 7.6T 52% /cephfs
```

3. *Quotas are implemented in the kernel client 4.17 and higher.* Quotas are supported by the userspace client (libcephfs, ceph-fuse). Linux kernel clients >= 4.17 support CephFS quotas but only on mimic+ clusters. Kernel clients (even recent versions) will fail to handle quotas on older clusters, even if they may be able to set the quotas extended attributes.

Fuse vs. Kernel

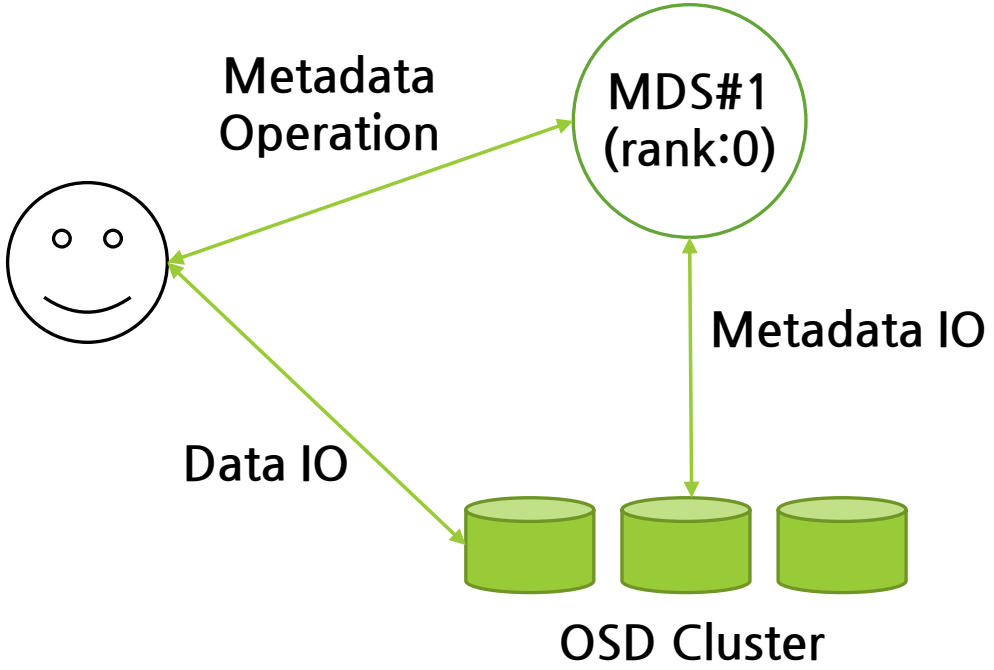
- Small File : for i in `seq 1000`; do echo hello > test${i}; done
- Large File : dd if=/dev/zero of=./1G bs=1M count=1000 oflag=direct
- Tar Extract : tar xf linux-4.15.14.tar.xz

Type	Ceph-Fuse (sec)	Kernel Mount (sec)
Small File Creation (1000 files)	17	4
Large File Creation (1gb)	56	23
Tar Extract (files:70k, 800mb)	147	32

In-Depth : MDS

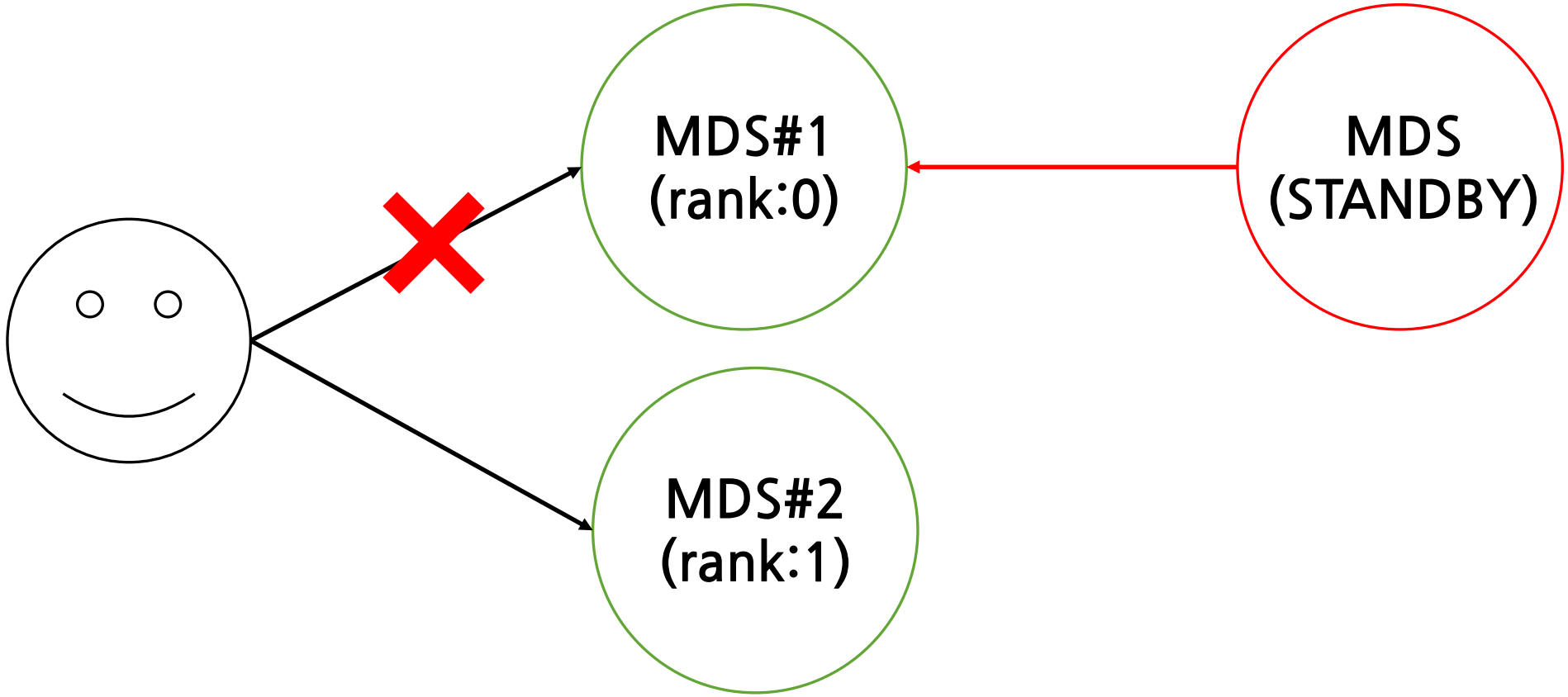
MDS : Metadata Server

- MDS Cluster is diskless (no local storage)
- All metadata stored at OSD Cluster (metadata pool)
- Just serve as an index for read and write
- Cache the metadata (lots of memory)



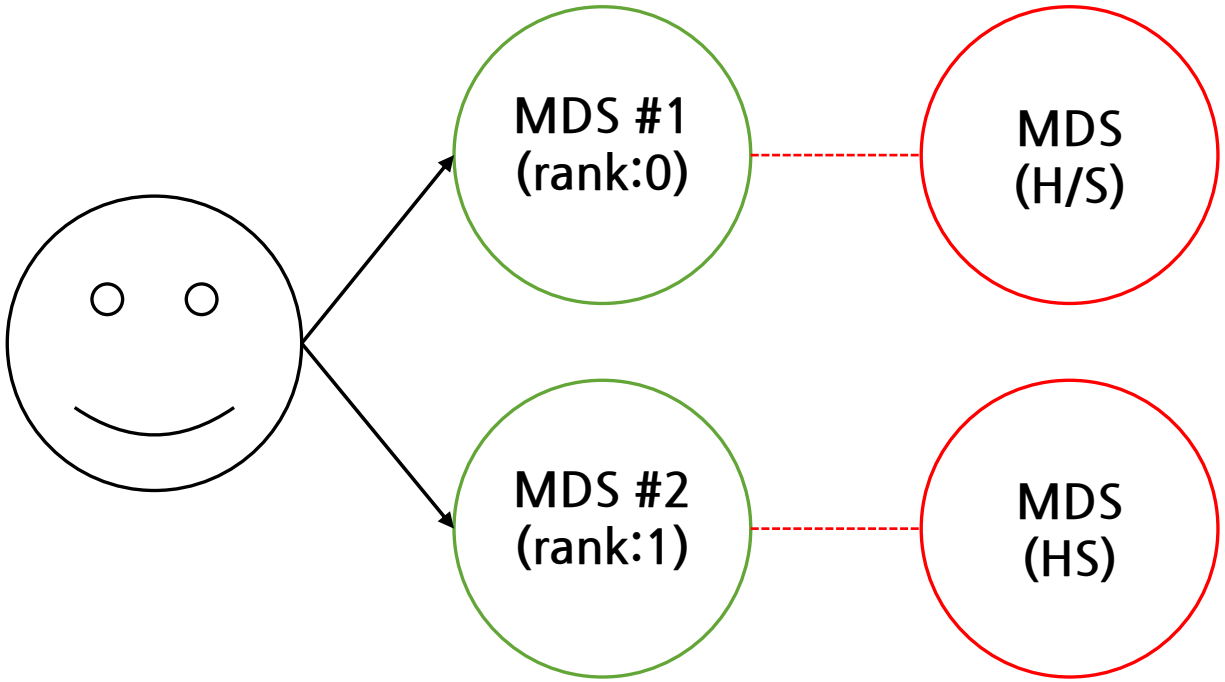
MDS High Availability : Floating Standby

- Floating Standby is not assigned a rank
- Take over for whichever other mds fails.



MDS High Availability : Hot Standby

- Standby daemon will continuously read the metadata of up rank
- Give a warm metadata cache
- Speed up the process of failing over



[mds.a]
mds standby replay = true
mds standby for rank = <rank>

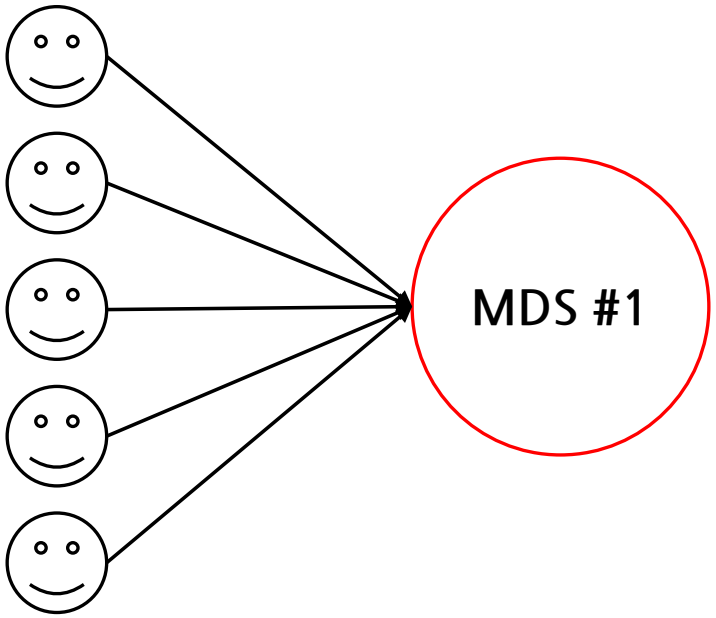
MDS
(STANDBY)

--hot-standby <rank>

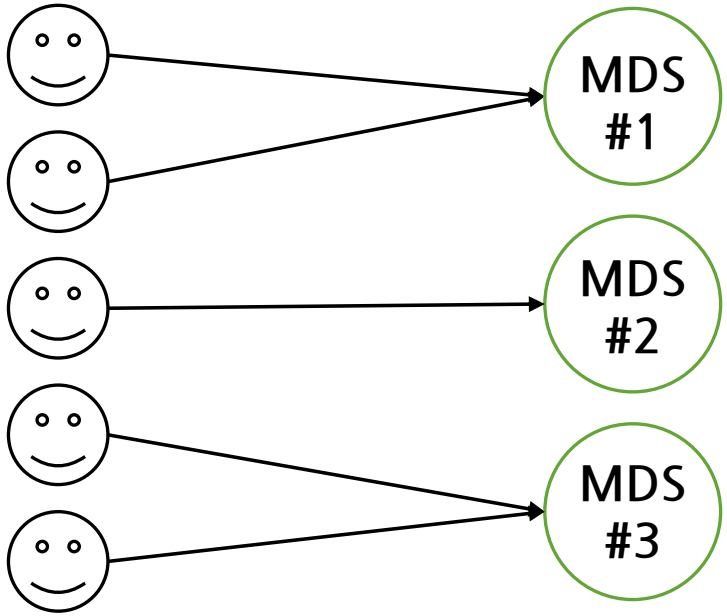
Multiple MDS

ceph fs set <fs_name> max_mds 3

- Single MDS has bottleneck
- Multiple MDS may not increase performance on all workloads
- Benefit from many clients working on many separate directories



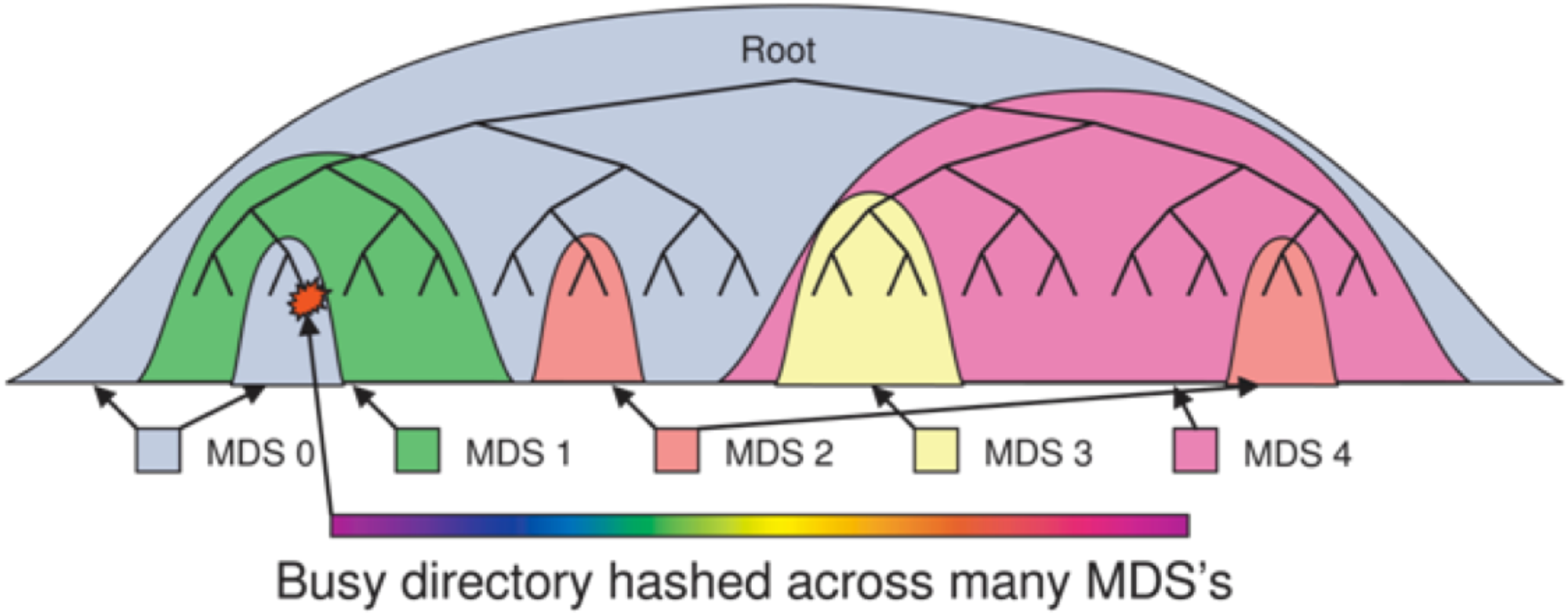
Single MDS



Multiple MDS

Dynamic Subtree Partitioning

- Avoid high management overhead : static subtree partitioning
- Avoid destroying locality : hash-based partitioning



https://www.usenix.org/legacy/event/osdi06/tech/full_papers/weil/weil_html/index.html

Load Balancer

`mds bal mode`

Description: The method for calculating MDS load.

- 0 = Hybrid.
- 1 = Request rate and latency.
- 2 = CPU load.

Type: 32-bit Integer

Default: 0

mds_bal_split_size (default 10000)

mds_bal_split_wr (default 10000)

mds_bal_split_rd (default 25000)

http://docs.ceph.com/docs/mimic/cephfs/mds-config-ref/

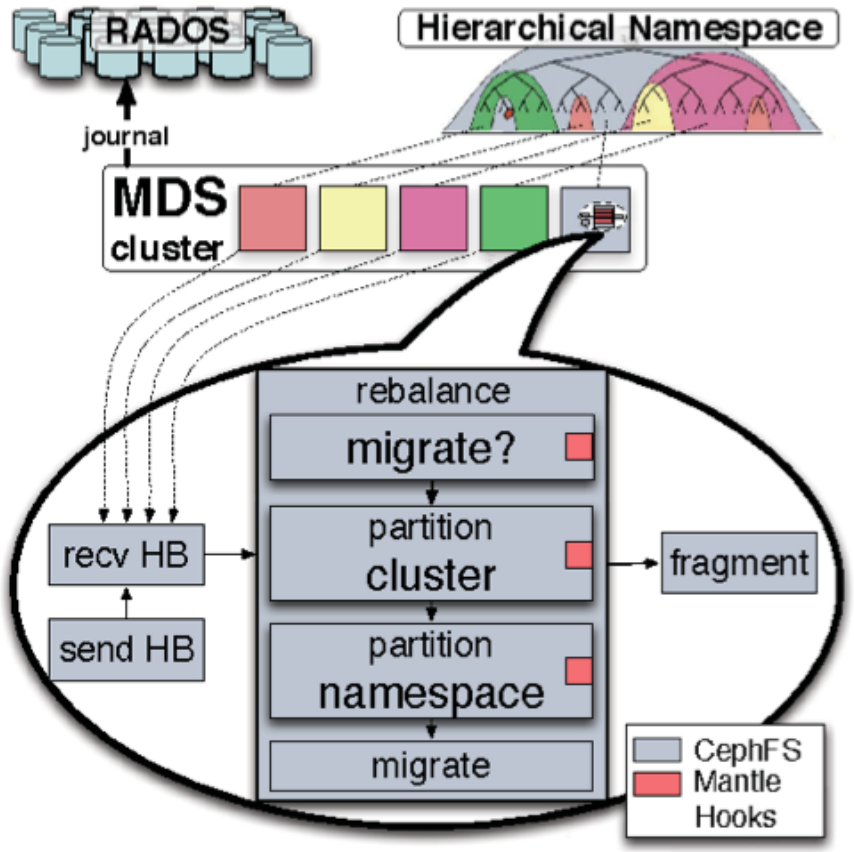
```
double mds_load_t::mds_load() const
{
  switch(g_conf->mds_bal_mode) {
  case 0:
    return
      .8 * auth.meta_load() +
      .2 * all.meta_load() +
      req_rate +
      10.0 * queue_len;

  case 1:
    return req_rate + 10.0*queue_len;

  case 2:
    return cpu_load_avg;

  }
  ceph_abort();
  return 0;
}
```

Mantle : programmable metadata load balancer for the ceph file system



gure 2: The MDS cluster journals to RADOS and

http://docs.ceph.com/docs/mimic/cephfs/mantle/

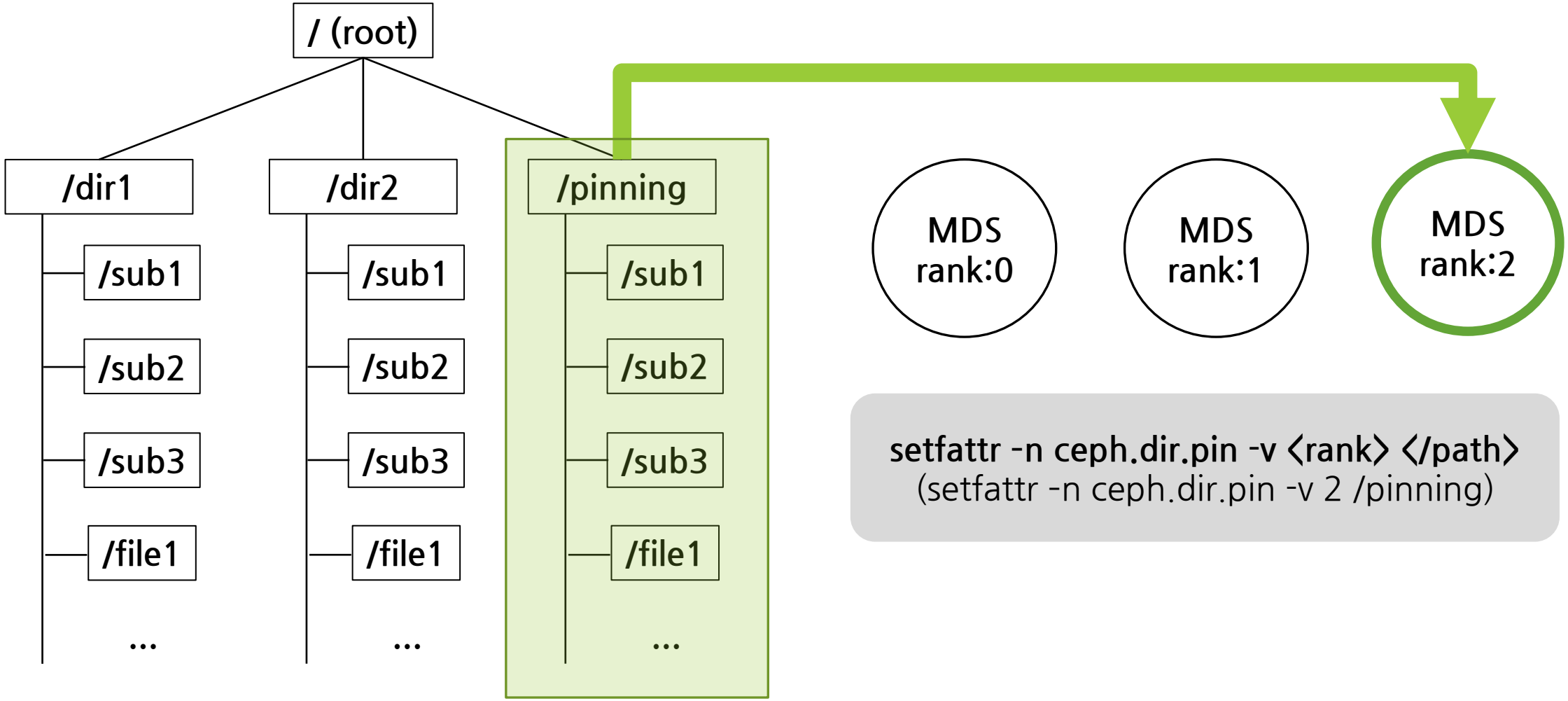
* Hook : when, where, how much, and load calculation policies

```
-- Shed load when you have load and your neighbor doesn't
local function when()
  if not mds[whoami+1] then
    -- i'm the last rank
    BAL_LOG(5, "when: not migrating! I am the last rank, nothing to spill to.");
    return false
  end
  my_load = mds[whoami]["load"]
  his_load = mds[whoami+1]["load"]
  if my_load > 0.01 and his_load < 0.01 then
    BAL_LOG(5, "when: migrating! my_load="..my_load.." hisload="..his_load)
    return true
  end
  BAL_LOG(5, "when: not migrating! my_load="..my_load.." hisload="..his_load)
  return false
end

-- Shed half your load to your neighbor
-- neighbor=whoami+2 because Lua tables are indexed starting at 1
local function where(targets)
  targets[whoami+1] = mds[whoami]["load"]/2
  return targets
end
```

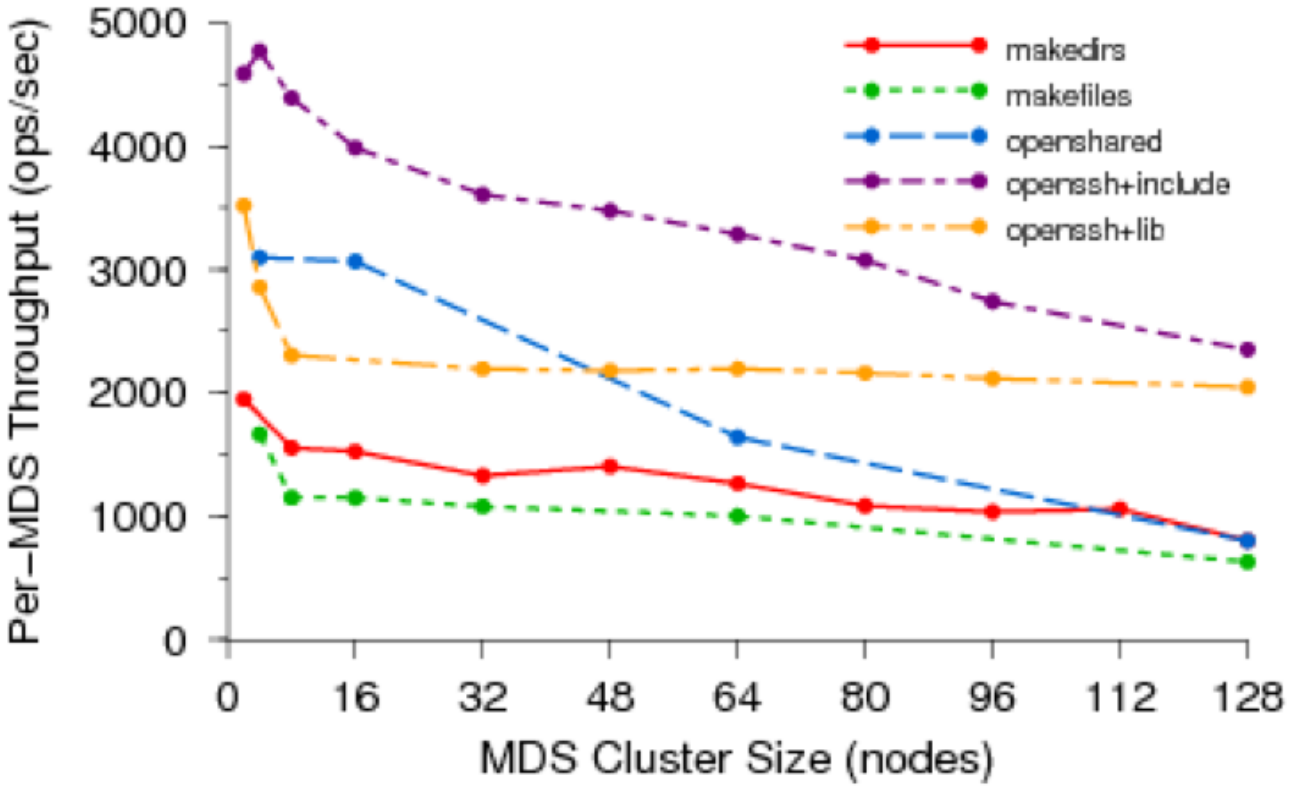
src/mds/balancers/greedyspill.lua

Subtree Pinning (static subtree partitioning)



setfattr -n ceph.dir.pin -v <rank> </path>
(setfattr -n ceph.dir.pin -v 2 /pinning)

MDS Scaling



MDS	reqs / sec
1	3,000
2	7,000
3	9,000

https://www.usenix.org/legacy/event/osdi06/tech/full_papers/weil/weil_html/index.html

QnA