Table of Contents

PetaSAN 1
Replication 1
Clean S3 pool 1
Ansible 1
List all pools 1
OSD disk free 1
CEPH rebalance 1
Check OSD Blocklist 1
Set minimum version 1
Remove OSD hard 2
Insert object into RADOS 2
Copy Pool 2
Where are data? 2
CEPH Print key 2
CEPH for Windows 2
CEPH list pool 2
CEPH delete pool 2
CEPH Create Erasure pool 3
RADOSGW 3
ceph.conf 3
ISCSI 3
Insert it into dashboard 4
Set OSD configs 4
OSD dump info 4
CEPH Repair 4
Edit crush-map 4
Turn cache on 5
Change device class 5
Partitions 5
CEPH LVM List 5
OSD Weight 5
Benchmark 6
Remove benchmark data 6
Show pool stats 6
Enable dashboard 6
Add new MON 6
CEPH List Auth 6
Show clock-skew 7
Ceph Evict Client 7
Replication 7
FIO 3xNodes 7
rbd ... object map is invalid 8

PetaSAN

Ceph for dummies http://www.petasan.org/

Replication

2 buildings (4/4 replication)
3 rooms (3/3 replication)

Clean S3 pool

```
radosgw-admin gc process --include-all
```

Ansible

https://docs.ceph.com/projects/ceph-ansible/en/latest/

List all pools

```
ceph osd pool ls detail
```

OSD disk free

```
ceph osd df tree
```

CEPH rebalance

```
ceph osd reweight-by-utilization
```

Check OSD Blocklist

```
ceph osd blocklist ls
ceph osd blocklist rm 127.0.0.1:0/3710147553
```

Set minimum version

```
ceph osd require-osd-release octopus
```

Remove OSD hard

```
dd if=/dev/zero of=/dev/sd{X} bs=1M count=10 conv=fsync
```

Insert object into RADOS

```
rados -p pool put {object} filename
```

rados -p pool ls

Copy Pool

```
pool={poolname}
ceph osd pool create $pool.new 128 128 erasure EC_RGW
rados cppool $pool $pool.new
ceph osd pool rename $pool $pool.old
ceph osd pool rename $pool.new $pool
```

Where are data?

```
ceph osd map {pool} object {object} -f json-pretty
```

CEPH Print key

```
ceph -k /etc/ceph/ceph.client.admin.keyring auth print-key entity
```

CEPH for Windows

https://github.com/dokan-dev/dokany/releases - required
https://cloudbase.it/ceph-for-windows/

CEPH list pool

```
ceph osd lspools
```

CEPH delete pool

```
ceph osd pool delete <pool-name> <pool-name> --yes-i-really-really-mean-it
```

CEPH Create Erasure pool

```
ceph osd pool create {name} {pgsize} erasure
ceph osd pool set {name} allow_ec_overwrites true;
ceph osd pool application enable {name} rbd;
```

RADOSGW

```
ceph-authtool --create-keyring /etc/ceph/ceph.client.radosgw.keyring
```

```
ceph-authtool /etc/ceph/ceph.client.radosgw.keyring -n client.radosgw.node01 --gen-key
```

```
ceph-authtool -n client.radosgw.node01 --cap osd 'allow rwx' --cap mon 'allow rwx'
/etc/ceph/ceph.client.radosgw.keyring
```

```
ceph -k /etc/ceph/ceph.client.admin.keyring auth add client.radosgw.node01 -i
/etc/ceph/ceph.client.radosgw.keyring
```

ceph.conf

```
[client.radosgw.node01]
       host = node01
       keyring = /etc/ceph/ceph.client.radosgw.keyring
       log file = /var/log/ceph/client.radosgw.$host.log
```

apt install radosgw
systemctl restart radosgw

http://node01:7480

ISCSI

```
sudo apt install ceph-iscsi targetcli-fb
systemctl daemon-reload
systemctl enable rbd-target-gw
systemctl start rbd-target-gw
systemctl enable rbd-target-api
systemctl start rbd-target-api
```

start *gwcli*

```
cd /iscsi-targets
create iqn.2003-01.com.janforman.iscsi-gw:iscsi-igw
cd /iscsi-targets/iqn.2003-01.com.janforman.iscsi-gw:iscsi-igw/gateways
create {nodename} {IP}
cd /disks
create pool=rbd image=disk_1 size=90G
cd /iscsi-targets/iqn.2003-01.com.janforman.iscsi-gw:iscsi-igw/hosts
create iqn.1994-05.com.janforman:client
cd /iscsi-targets/iqn.2003-01.com.janforman.iscsi-gw:iscsi-
igw/hosts/iqn.1994-05.com.janforman:client
auth username=myiscsiusername password=myiscsipassword
```

```
disk add rbd/disk_1
```

Insert it into dashboard

file: http://admin:admin@10.160.1.15:5001

```
ceph dashboard iscsi-gateway-add -i file
```

Set OSD configs

```
ceph tell osd.* config set osd_heartbeat_grace 20
ceph tell osd.* config set osd_heartbeat_interval 5
```

OSD dump info

```
ceph osd dump
```

CEPH Repair

```
ceph health detail
```

```
HEALTH_ERR 1 scrub errors; Possible data damage: 1 pg inconsistent
OSD_SCRUB_ERRORS 1 scrub errors
PG_DAMAGED Possible data damage: 1 pg inconsistent
    pg 3.31 is active+clean+inconsistent, acting [5,2,0]
```

Corrupted PG on OSD 5,2,0

```
ceph pg repair 3.31
```

```
2019-07-29 10:01:54.975649 mon.cloud-gis00 (mon.0) 21584 : cluster [INF] Health check
cleared: OSD_SCRUB_ERRORS (was: 1 scrub errors)
2019-07-29 10:01:54.975690 mon.cloud-gis00 (mon.0) 21585 : cluster [INF] Health check
cleared: PG_DAMAGED (was: Possible data damage: 1 pg inconsistent)
2019-07-29 10:01:54.975709 mon.cloud-gis00 (mon.0) 21586 : cluster [INF] Cluster is now
healthy
2019-07-29 10:01:52.358272 osd.5 (osd.5) 428 : cluster [ERR] 3.31 shard 0 soid
3:8df0528b:::rbd_data.9f8f474b0dc51.0000000000002485:head : candidate had a read error
2019-07-29 10:01:52.358608 osd.5 (osd.5) 429 : cluster [ERR] 3.31 repair 0 missing, 1
inconsistent objects
2019-07-29 10:01:52.358616 osd.5 (osd.5) 430 : cluster [ERR] 3.31 repair 1 errors, 1 fixed
```

Edit crush-map

```
ceph osd getcrushmap -o /tmp/crushmap
crushtool -d /tmp/crushmap -o crush_map

crushtool -c crush_map -o /tmp/crushmap
```

```
ceph osd setcrushmap -i /tmp/crushmap
```

Turn cache on

```
[client]
rbd_cache = true
```

May improve performance

```
osd_enable_op_tracker = false
throttler perf counter = false
```

Change device class

If the automatic device class detection gets something wrong (e.g., because the device driver is not properly exposing information about the device via /sys/block), you can also adjust device classes from the command line:

```
$ ceph osd crush rm-device-class osd.2 osd.3
done removing class of osd(s): 2,3
$ ceph osd crush set-device-class ssd osd.2 osd.3
set osd(s) 2,3 to class 'ssd'
```

Partitions

```
# types
type 0 osd
type 1 host
type 2 chassis
type 3 rack
type 4 row
type 5 pdu
type 6 pod
type 7 room
type 8 datacenter
type 9 region
type 10 root
```

CEPH LVM List

```
ceph-volume lvm list
```

OSD Weight

```
ceph osd crush set 0 0.5 pool=default host=proxmox01
ceph osd crush set 1 0.5 pool=default host=proxmox02
ceph osd crush set 2 0.5 pool=default host=proxmox03
```

Benchmark

```
rados -p ceph bench 60 write --no-cleanup
```

Default object size is 4 MB, and the default number of simulated threads (parallel writes) is 16.
-t (threads)
write / seq / read

Remove benchmark data

```
rados -p pool cleanup --prefix benchmark_data
```

Show pool stats

```
rados -p ceph df
```

Enable dashboard

```
ceph mgr module enable dashboard
```

Generate selfsigned certificate

```
ceph dashboard create-self-signed-cert
```

Disable TLS

```
ceph config set mgr mgr/dashboard/ssl false
```

```
ceph dashboard ac-user-create <username> -i <file-containing-password> administrator
```

Add new MON

```
ceph auth get mon. -o /tmp/keyring
ceph mon getmap -o /tmp/map
sudo ceph-mon -i {HOSTNAME} --mkfs --monmap /tmp/map --keyring /tmp/keyring
chown -R ceph:ceph /var/lib/ceph/mon
```

manual run

```
ceph-mon -f -i {HOSTNAME} --public-addr {IP}
```

CEPH List Auth

```
ceph auth list
```

Show clock-skew

```
ceph time-sync-status
```

Ceph Evict Client

https://docs.ceph.com/en/latest/cephfs/eviction/

Replication

```
ceph osd pool set data size 3
ceph osd pool set data min_size 2
```

For n = 4 nodes each with 1 osd and 1 mon and settings of replica min_size 1 and size 4 three osd can fail, only one mon can fail (the monitor quorum means more than half will survive). 4 + 1 number of monitors is required for two failed monitors (at least one should be external without osd). For 8 monitors (four external monitors) three mon can fail, so even three nodes each with 1 osd and 1 mon can fail. I am not sure that setting of 8 monitors is possible.

For three nodes each with one monitor and osd the only reasonable settings are replica min_size 2 and size 3 or 2. Only one node can fail. If you have an external monitors, if you set min_size to 1 (this is very dangerous) and size to 2 or 1 the 2 nodes can be down. But with one replica (no copy, only original data) you can loose your job very soon.

- Ensure you have a realistic number of placement groups. We recommend
- approximately 100 per OSD. E.g., total number of OSDs multiplied by 100
- divided by the number of replicas (i.e., osd pool default size). So for
- 10 OSDs and osd pool default size = 4, we'd recommend approximately
- (100 * 10) / 4 = 250.

FIO 3xNodes

```
ceph_test: (groupid=0, jobs=16): err= 0: pid=1884829: Tue Jan 21 14:52:38 2025
  read: IOPS=16.5k, BW=2066MiB/s (2166MB/s)(1211GiB/600014msec)
    slat (usec): min=5, max=9989, avg=15.24, stdev= 9.73
    clat (usec): min=2, max=833395, avg=3854.45, stdev=6681.18
     lat (usec): min=225, max=833407, avg=3869.91, stdev=6681.22
    clat percentiles (usec):
     |  1.00th=[  553],  5.00th=[  611], 10.00th=[  676], 20.00th=[  799],
     | 30.00th=[  922], 40.00th=[ 1090], 50.00th=[ 1434], 60.00th=[ 1991],
     | 70.00th=[ 2933], 80.00th=[ 4686], 90.00th=[ 9765], 95.00th=[16712],
     | 99.00th=[31589], 99.50th=[39584], 99.90th=[57410], 99.95th=[63701],
     | 99.99th=[79168]
   bw (  MiB/s): min=  125, max= 3034, per=100.00%, avg=2069.56, stdev=18.99, samples=19162
   iops        : min= 1005, max=24278, avg=16556.49, stdev=151.89, samples=19162
  lat (usec)   : 4=0.01%, 10=0.01%, 20=0.01%, 50=0.01%, 250=0.01%
  lat (usec)   : 500=0.19%, 750=15.82%, 1000=19.21%
  lat (msec)   : 2=24.95%, 4=16.83%, 10=13.27%, 20=6.21%, 50=3.30%
  lat (msec)   : 100=0.21%, 250=0.01%, 500=0.01%, 750=0.01%, 1000=0.01%
  cpu          : usr=0.53%, sys=1.66%, ctx=9602473, majf=0, minf=2249
  IO depths    : 1=0.1%, 2=0.1%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued rwts: total=9917559,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=4
```

```
Run status group 0 (all jobs):
   READ: bw=2066MiB/s (2166MB/s), 2066MiB/s-2066MiB/s (2166MB/s-2166MB/s), io=1211GiB
(1300GB), run=600014-600014msec

Disk stats (read/write):
    dm-0: ios=9918511/2691, merge=0/0, ticks=38120430/45168, in_queue=38165598, util=100.00%,
aggrios=9918511/2638, aggrmerge=0/140, aggrticks=38166176/39876, aggrin_queue=38206052,
aggrutil=73.52%
  vda: ios=9918511/2638, merge=0/140, ticks=38166176/39876, in_queue=38206052, util=73.52%
```

rbd ... object map is invalid

```
rbd -p ceph-pool object-map rebuild vm-100-disk-0
```

From:
https://wiki.janforman.com/ - **wiki.janforman.com**

Permanent link:
https://wiki.janforman.com/storage:ceph

Last update: **2026/07/03 10:27**