Table of Contents

CEPH for Windows

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

CEPH list pool

```
ceph osd lspools
```

CEPH list 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;
```

ISCSI

```
sudo apt install ceph-iscsi
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*

```
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
```

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
```

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

Show pool stats

```
rados -p ceph df
```

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.