CEPH

- OSD - (Object Storage Device) - Storage on a physical device or logical unit (LUN). Typically, data on an OSD is configured as a btrfs file system to take advantage of its snapshot features. However, other file systems such as XFS can also be used.
- MON - (Monitor) - A Ceph component used for tracking active and failed nodes in a Storage Cluster. The Ceph Monitor (MON) [5] maintains a master copy of the cluster map. For high availability, you need at least 3 monitors. One monitor will already be installed if you used the installation wizard. You won't need more than 3 monitors, as long as your cluster is small to medium-sized. Only really large clusters will require more than th
- MGR - (Manager) - The Ceph manager software, which collects all the state from the whole cluster in one place.
- MDS - (Meta Data Server)
- RBD - (RADOS Block Device) - A Ceph component that provides access to Ceph storage as a thinly provisioned block device. When an application writes to a Block Device, Ceph implements data redundancy and enhances I/O performance by replicating and striping data across the Storage Cluster.

Ceph RADOS Block Devices (RBD)

CEPH provides only a pools of object. To use it for VMs block devices additional layer (RBD) is needed. It can be created manually or during CEPH pool creation (option `Add as Storage`)

Ceph FS

It is implementation of POSIX compliant FS top of CEPH POOL. It requires one pool for data (block data) and to keep filesystem information (metadata). Performance strictly depends on metadata pool, so it is recommended to use for backups files.

Used ports:

- TCP 6789 - monitors
- TCP 6800-7300 - OSDs:
    1. One for talking to clients and monitors.
    2. One for sending data to other OSDs (replication, backfill and recovery).
    3. One for heartbeating.
- TCP 7480 - CEPH Object Gateway
- The MDS also uses a port above 6800.

Prepare

Read Proxmox CEPH requirements. It requires at least one spare hard drive on each node. Topic for later.

Installation

- On one of node:
  - `Datacenter` -> `Ceph` -> `Install Ceph-nautilus`
  - Configuration tab
  - First Ceph monitor - set to current node.
    - NOTE: Not possible to use other nodes now because there is no Ceph installed on it
- Repeat installation on each node. Configuration will be detected automatically.
- On each node - add additional monitors:
  - Select node -> `Ceph` -> `Monitor`
    - `Create` button in Monitor section, and select available nodes.

create OSD

Create Object Storage Daemon

On every node in cluster

- Select host node
- Go to menu `Ceph` → `OSD`
- `Create: OSD`
  - select spare hard disk
  - leave other defaults
  - press `Create`

If there is no `unused` disk to choose, erase content of disk with command:

```
ceph-volume lvm zap /dev/... --destroy
```

restart ceph services

```
systemctl stop ceph\*.service ceph\*.target
systemctl start ceph.target
```

create pool

- `Size` - number of replicas for pool
- `Min. Size` -
- `Crush Rule` - only possible to choose 'replicated_rule `* pg_num (Placement Groups) use` Ceph PGs per Pool Calculator `to calculate pg_num * NOTE: It's also important to know that the PG count can be increased, but NEVER decreased without destroying / recreating the pool. However, increasing the PG Count of a pool is one of the most impactful events in a Ceph Cluster, and should be avoided for production clusters if possible. *` Placement Groups `*` Add as Storage'' - automatically create Proxmox RBD storage (Disc

Image, Container)

pool benchmark

Benchmarks for pool name 'rbd' and 10 seconds duration

```
# Write benchmark
rados -p rbd bench 10 write --no-cleanup

# Read performance
rados -p rbd bench 10 seq
```