ceph File System

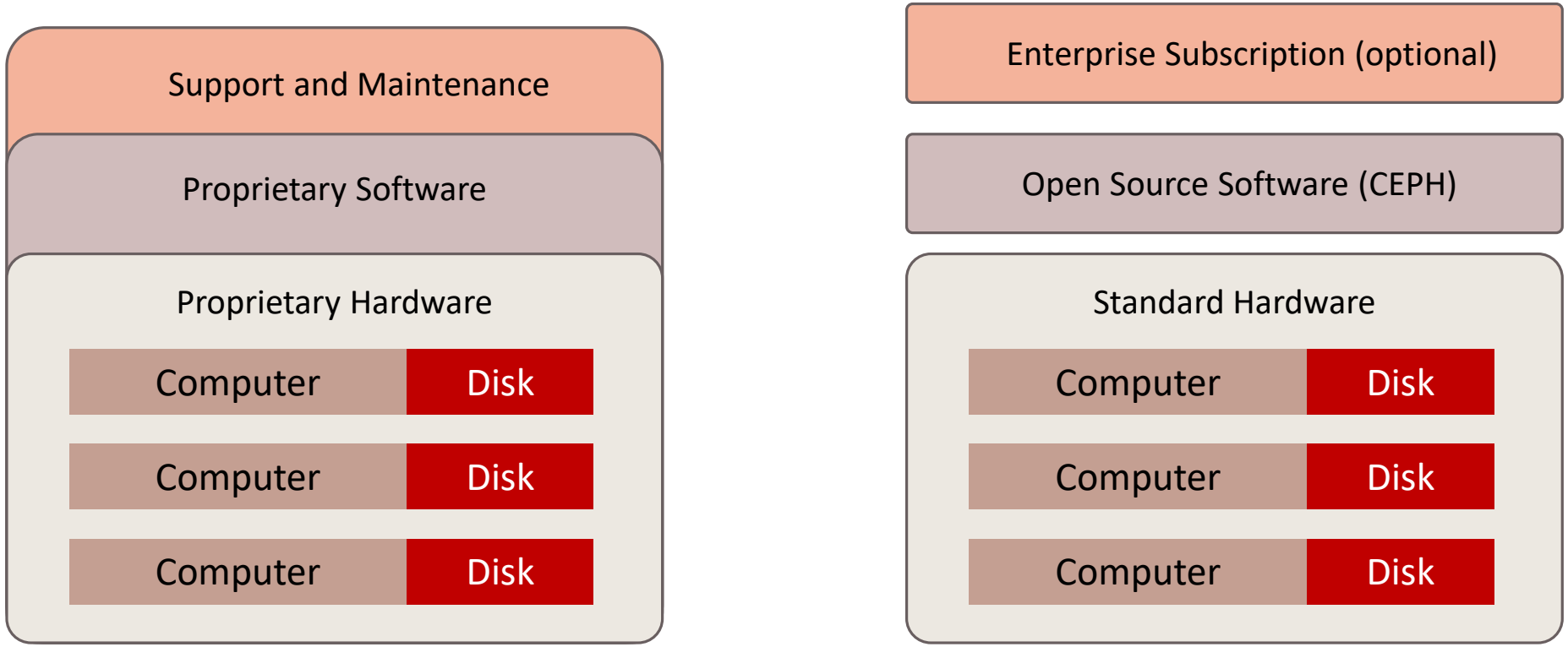
MINA RADY, MINARADY@STUDENT.LUT.FI

COMPUTING CLUSTERS, COMPUTING GRIDS, COMPUTING CLOUDS COURSE

PROFESSOR ANDREY SHEVEL, ITMO UNIVERSITY

5 JUNE 2017, ST. PETERSBURG, RUSSIA

Proprietary vs Open Cluster Storage



What is CEPH?

- Software Defined Storage Cluster solution that is:
  - Self-healing.
  - Self-managed.
  - “No” bottlenecks.
- Three Interfaces:
  - Object Access
  - Block Access
  - Distributed File System Access.

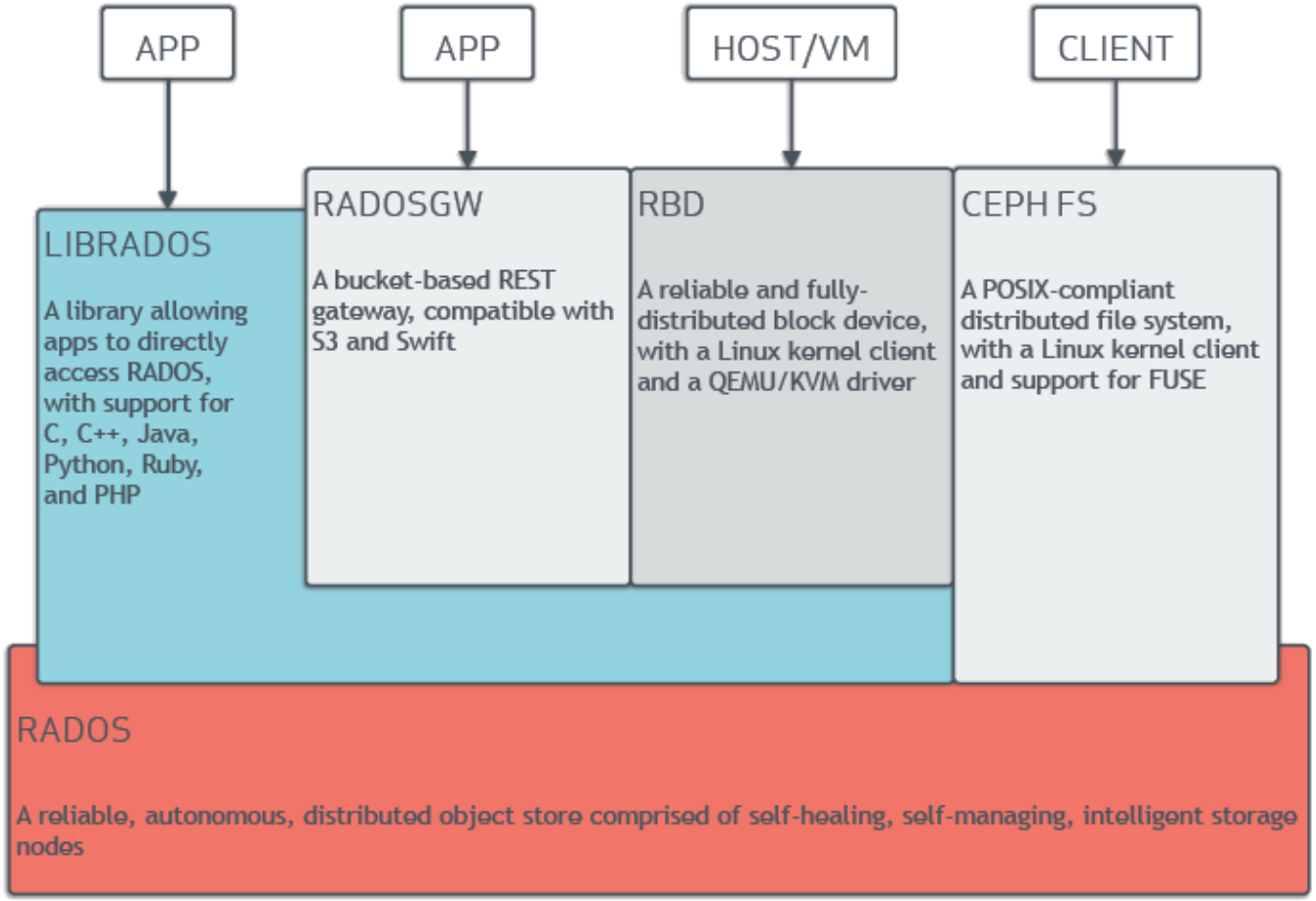
Ceph Features

- CephFS Offers a traditional file system with interface with POSIX compatibility.
- Importance of POSIX (Portable Operating System Interface):
  - It implements system calls usually used across multiple operating systems (Unix, mainly).
  - Allows safe transition of applications deploying POSIX calls such as: `read`, `write`, `seekdir`.
- Provides stronger data safety for mission-critical applications.
- RADOS: Ceph core is Reliabale Autonomic Distributed Object Store
- No single point of failure
- Software Based

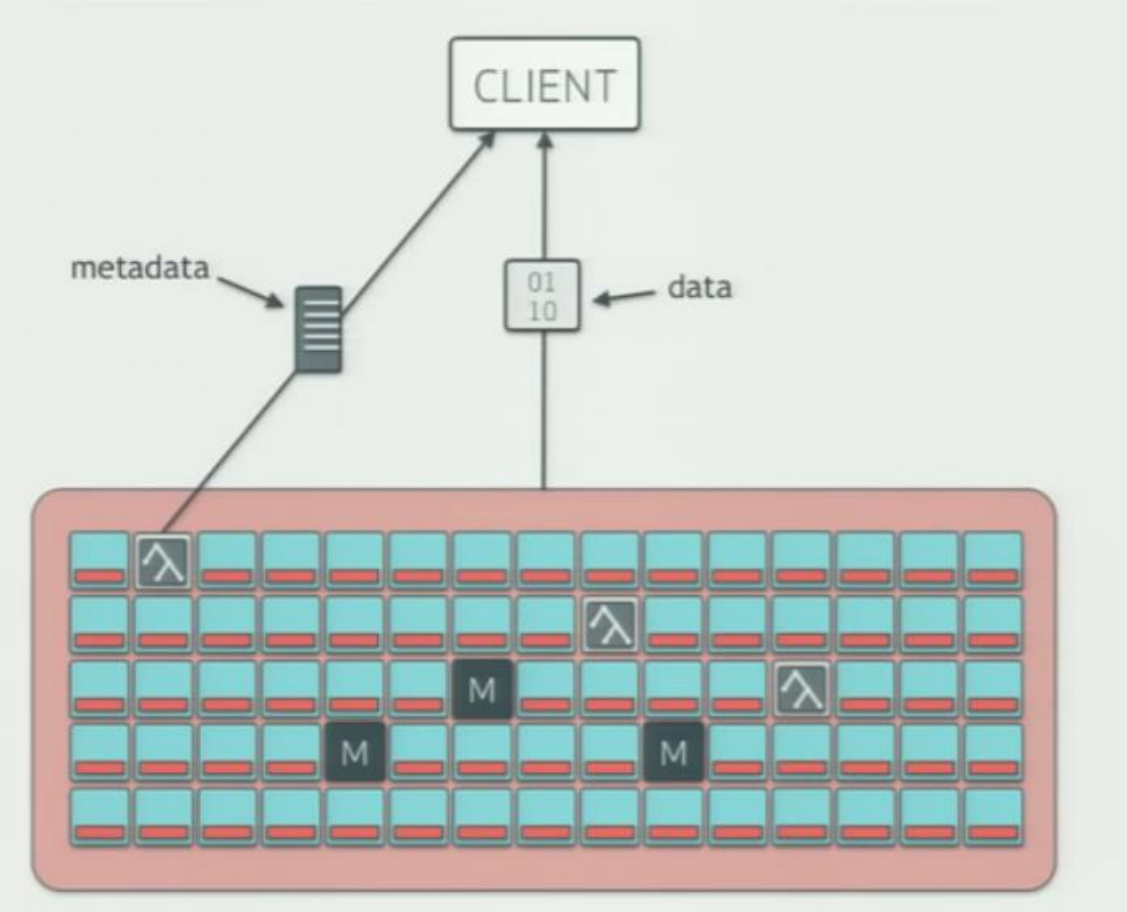
What is Ceph?: Main Concepts

- Any Ceph deployment consists of:
  - OSD:
    - Object Storage Daemon to run on top of each storage node.
    - Each OSD points out to a path in the storage disk and makes it a storage location for Ceph.
    - The path can even point out to a RAID.
    - OSDs responsible for peering, replication, and recovery tasks.
  - Monitor Nodes:
    - First point of contact to clients, after that, client talks to OSDs.
    - Odd number of nodes are needed for consensus and distributed decision making.
  - Librados:
    - Library for accessing RADOS.
    - Uses native sockets to talk to RADOS or to receive commands from C, C++ (no HTTP overload).
    - Offers REST interface.
    - Offers Support for C, C++, Java, PHP, and Python.

Ceph Technical Outline

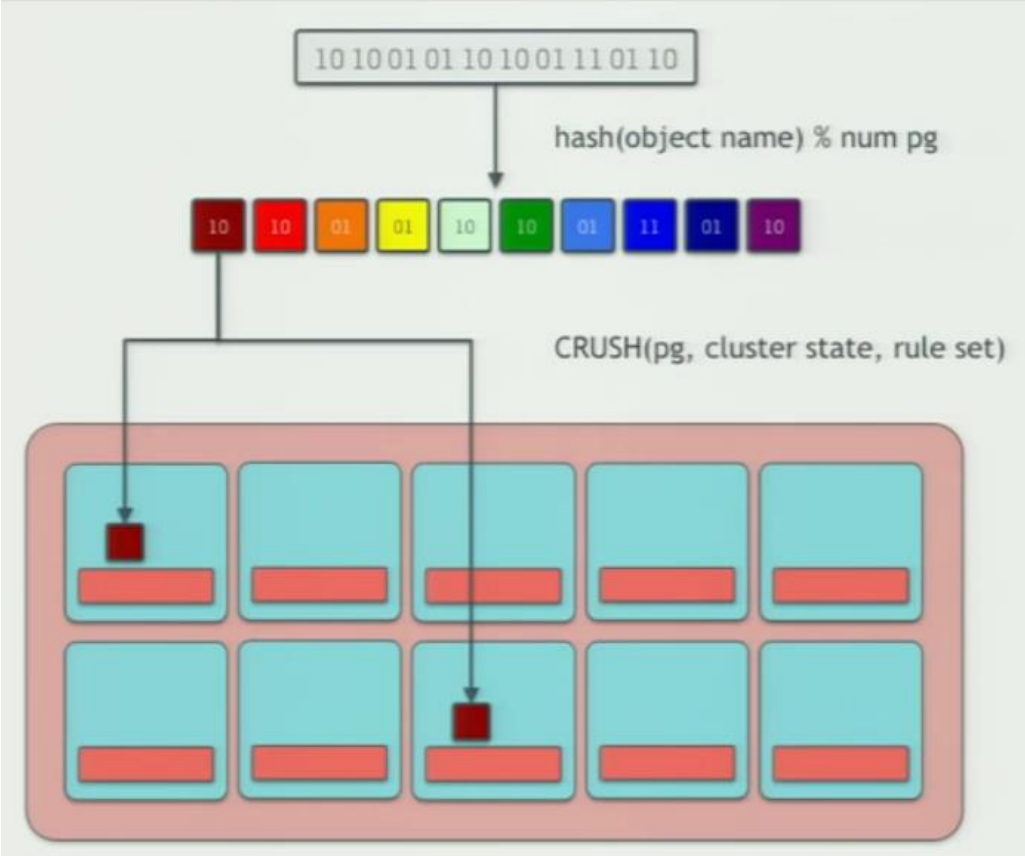


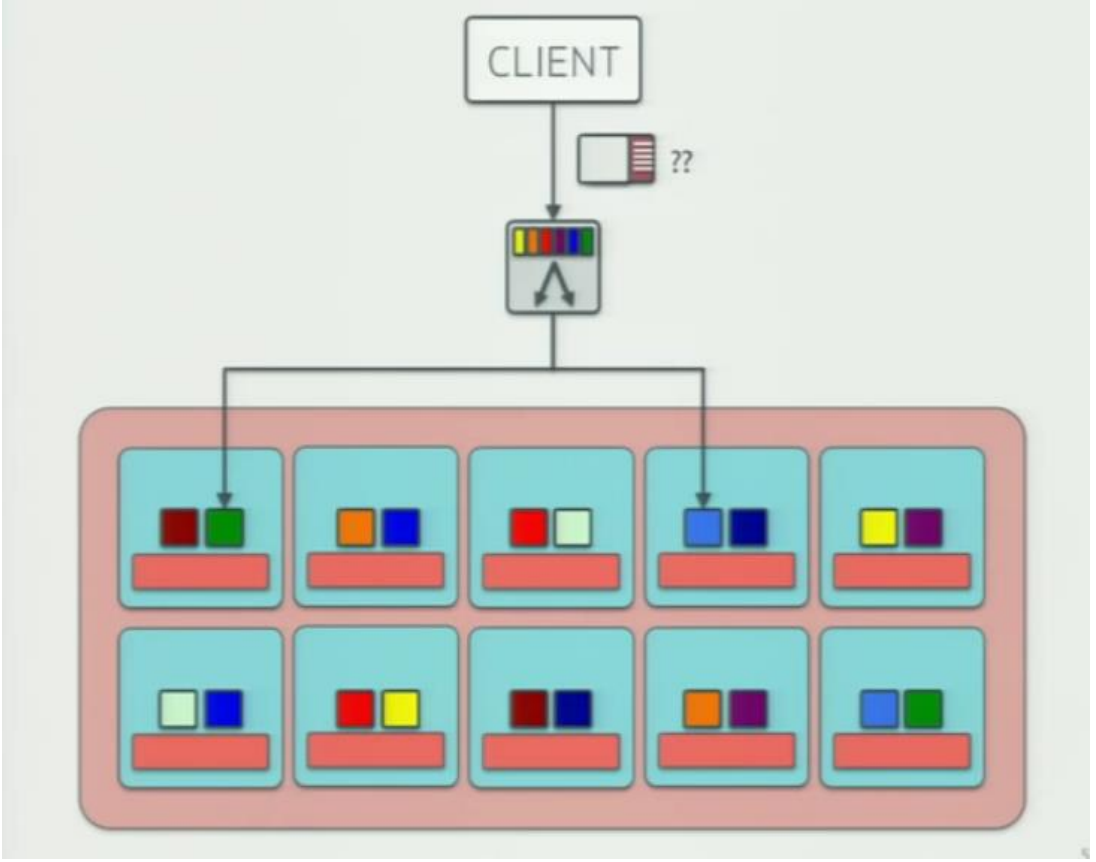
Ceph Technical Outline



Ceph Uniqueness: CRUSH Algorithm

CRUSH: Controlled Replication Under Scalable Hashing





Source: https://www.youtube.com/watch?v=OyH1C0C4HzM&t=626s

CephFS Main Concepts

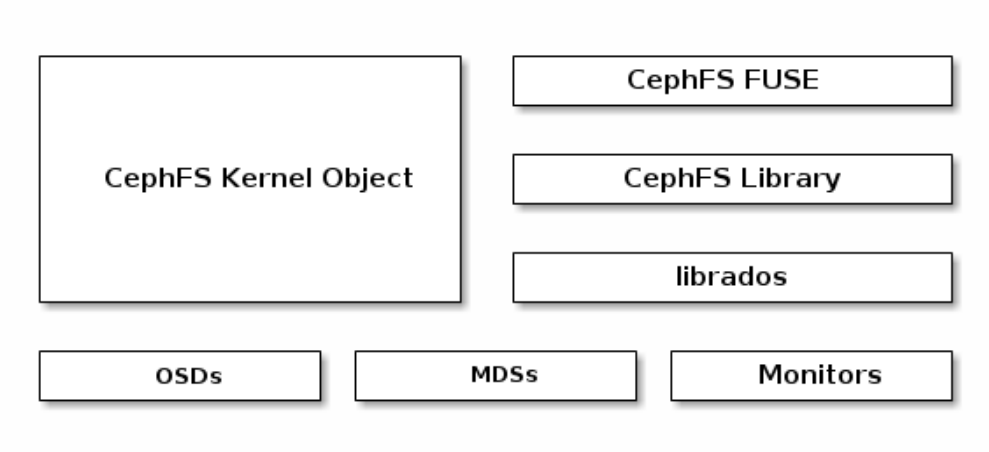
Pools: Logical Containers for Storage Objects

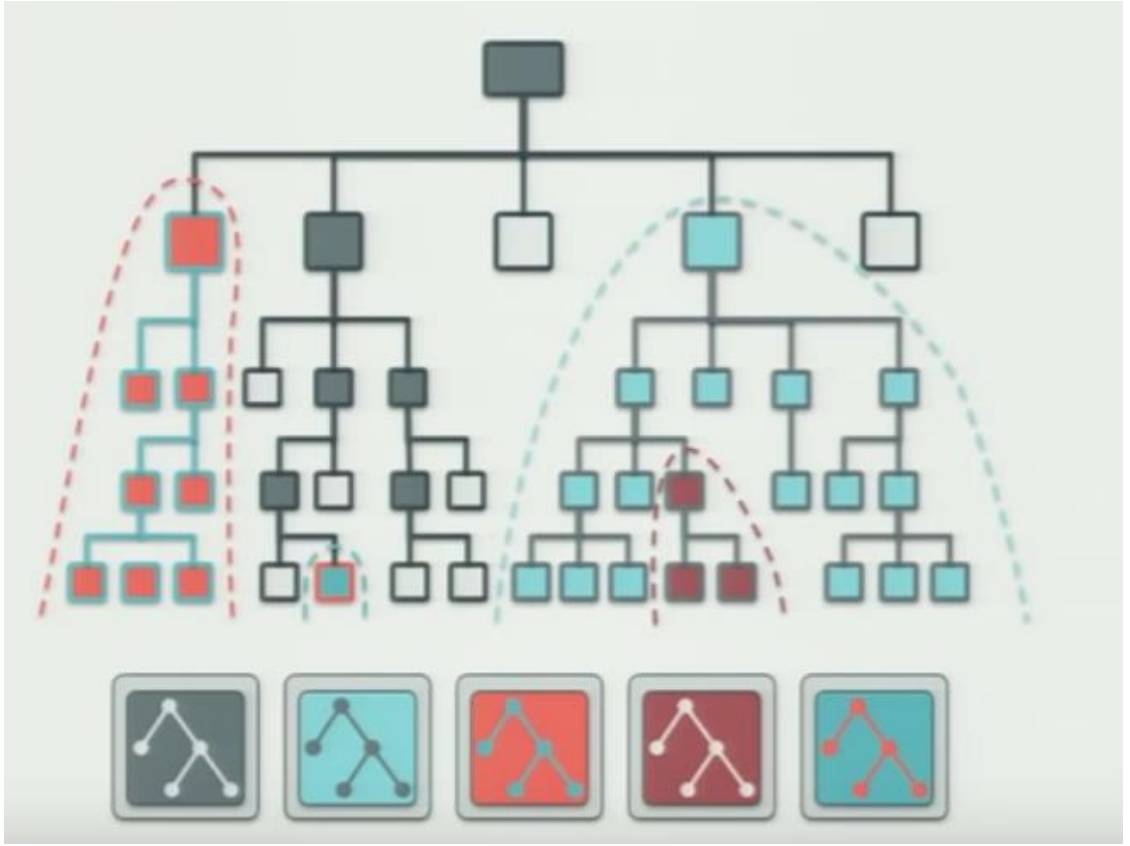
Placement Groups: You don't talk to them directly but you care how many there are, they are hash buckets. It is computationally intensive to map objects to OSDs, so you map them to a placement groups and then you get which OSD a pool is on

One PG spans several OSDs, one OSD serves meany PGs

CephFS Architecture

- Using CephFs require at least one Metadata Server.
- Metadata Servers perform replication using Dynamic Subtree Partitioning. It could even reach one server per file in case of peak demand.





Deployment Considerations:

- Pricing: free source!
- Professional Support: By Kernel Software Inc, Wisconson, USA: $3700 storage architecture. $1700 over existing Open Stack Deployment.
- Intended for large enterprise operations. Not recommended for physically small storage deployments
- It automatically balances storage spaces. Therefore note recommended if special operations are required in the storage clusters.
- Preferable separate public and cluster networks
- Cluster network should be 2x public network bandwidth

Design Decisions

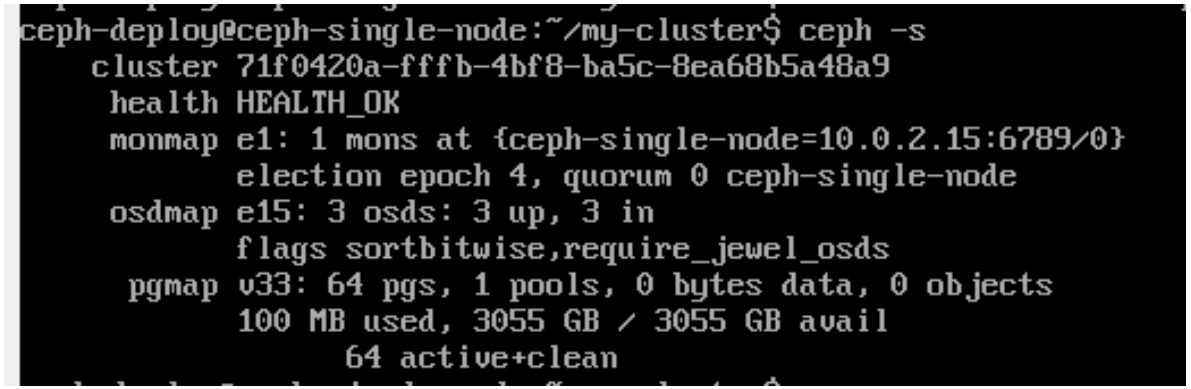
- Hard Disk Parameters:
  - Higher RPM, more energy cost but higher IOPS and throughput
  - Higher Density contribute positively to performance but negatively to maintainability
- Redundancy:
  - Replication: Increased read performances
  - Increases cluster utilization for writes
- Erasure Coding:
  - Customized data and parity bit distribution
  - Better Space and Network Efficiency
  - Higher CPU Overhead
- Scalability
  - Adding more nodes, Ceph will make sure data are evenly distributed across nodes until you hit network topology limitations.

Further Questions

- Smart Grid Metrics in load storage balancing
- Assessment of reliability of Ceph in Grid paradigm for large enterprise to alleviate the need for dedicated storage space.
- Optimization model for data distribution depending on access paterns.

Demo

- Deployment on virtual machine with one main storage disk of 8 GB (dynamic) and 3 storage disks of 1 TB each (dynamic).
- Installed 3 OSDs, 1 Monitor and 1 Metadata Server
- Networking Installation Caveat, not mentioned in installation guidelines: only real internal IP addresses need to be used to refer to hosts.
- Possible improvements?:
  - Adding an option of local host access to avoid load on network interface when possible.
  - Dashboard for analytics, visualization, and integration. JSON API for control?



Demo

- Health check: ceph –s
- Stats: ceph osd stat
- OSD List: ceph osd tree –f json
- Service Orientation, JSON API: Ceph osd dump –f json