

CEPH WORKSHOP

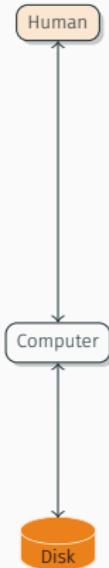
GridKA School 2015

Diana Gudu

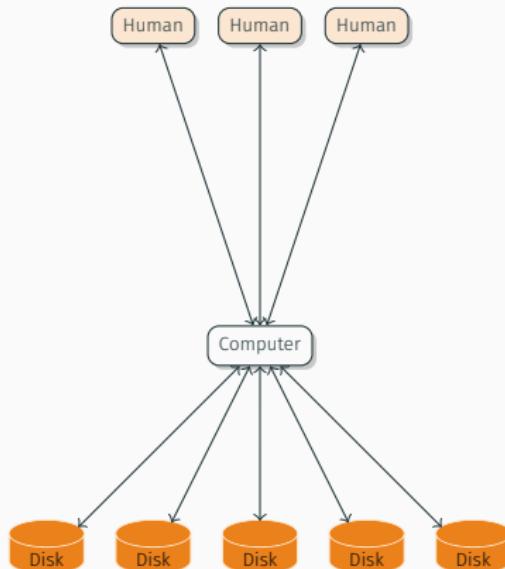
July 30, 2015

Karlsruhe Institute of Technology

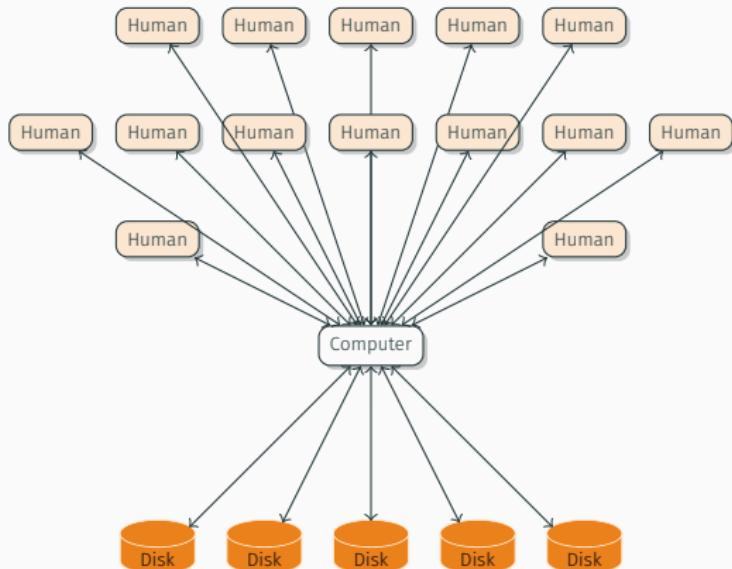
EVOLUTION OF STORAGE



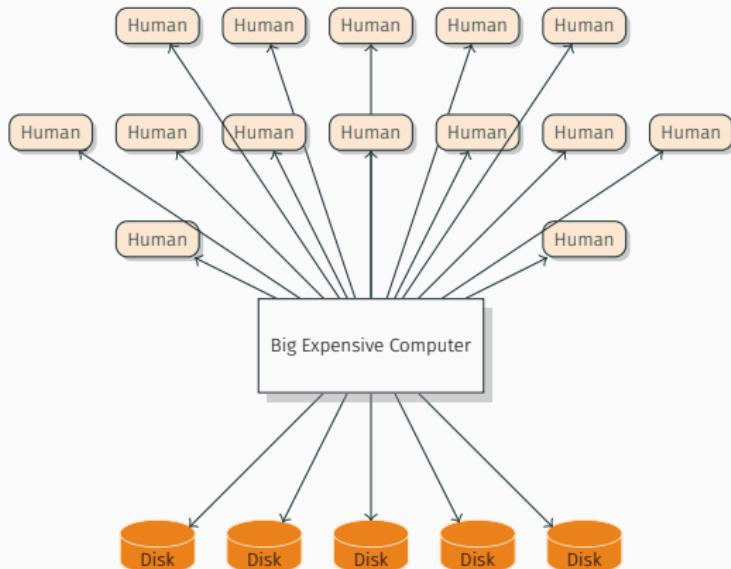
EVOLUTION OF STORAGE



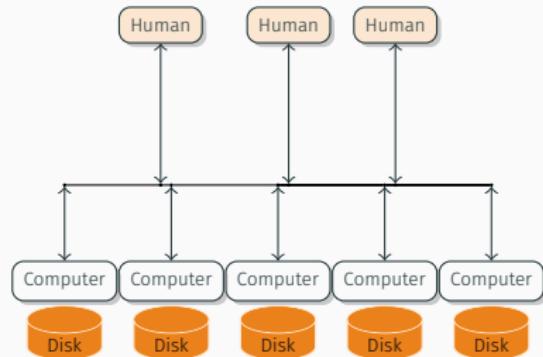
EVOLUTION OF STORAGE



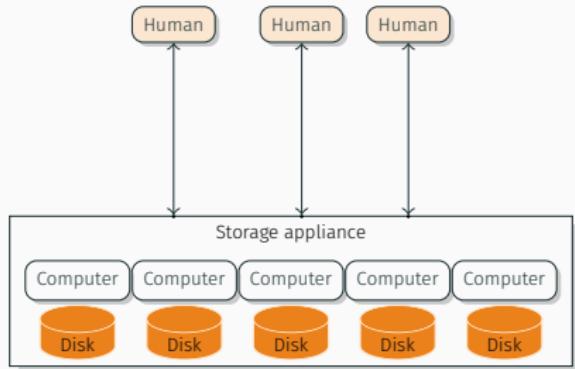
EVOLUTION OF STORAGE



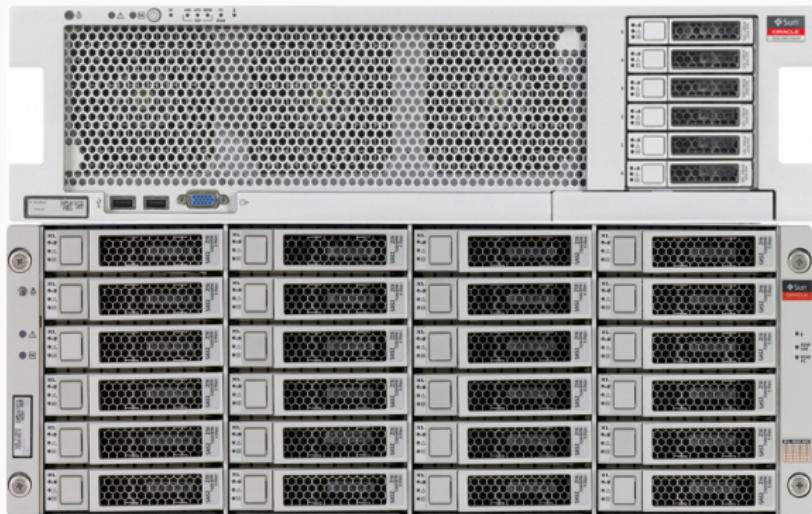
EVOLUTION OF STORAGE



EVOLUTION OF STORAGE

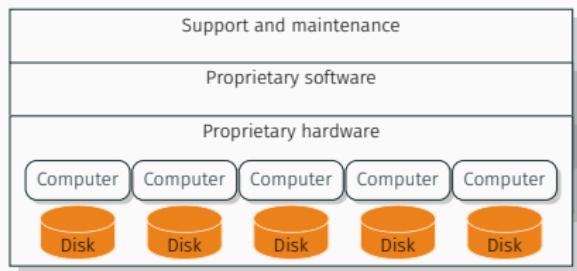


STORAGE APPLIANCE

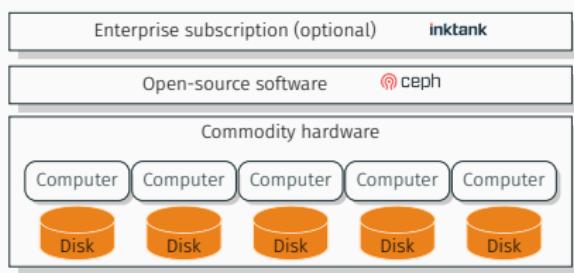
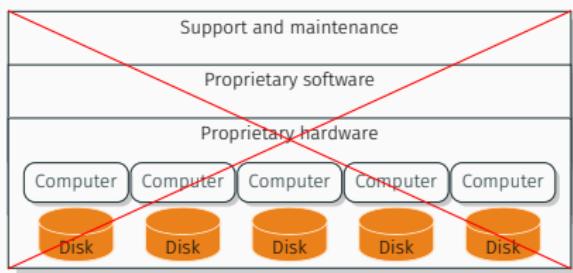


Oracle <http://www.e-business.com/zfs-7420-storage-appliance>

FUTURE OF STORAGE



FUTURE OF STORAGE



CEPH

Philosophy

- open-source

Philosophy

- open-source
- community focused

Philosophy

- open-source
- community focused
- software-defined

Philosophy

- open-source
- community focused
- software-defined
- scale-out hardware

Philosophy

- open-source
- community focused
- software-defined
- scale-out hardware
- self-managing

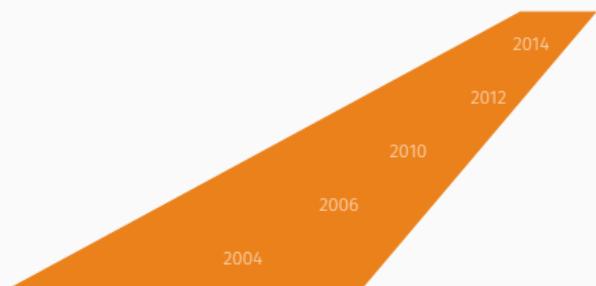
Philosophy

- open-source
- community focused
- software-defined
- scale-out hardware
- self-managing
- failure is normal

Philosophy

- open-source
- community focused
- software-defined
- scale-out hardware
- self-managing
- failure is normal

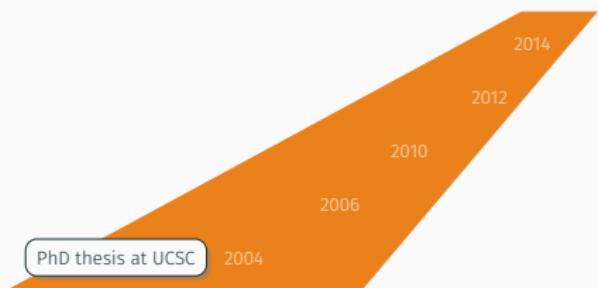
History



Philosophy

- open-source
- community focused
- software-defined
- scale-out hardware
- self-managing
- failure is normal

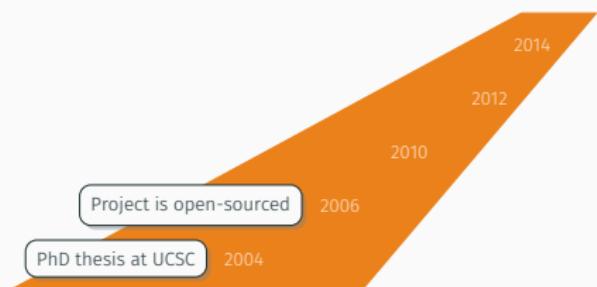
History



Philosophy

- open-source
- community focused
- software-defined
- scale-out hardware
- self-managing
- failure is normal

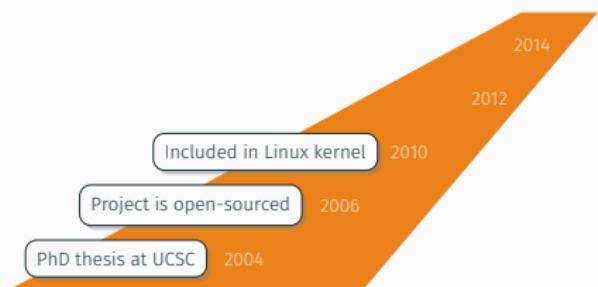
History



Philosophy

- open-source
- community focused
- software-defined
- scale-out hardware
- self-managing
- failure is normal

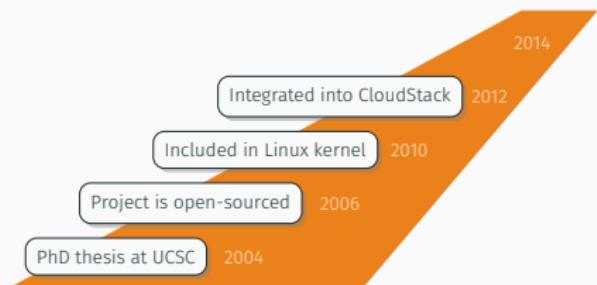
History



Philosophy

- open-source
- community focused
- software-defined
- scale-out hardware
- self-managing
- failure is normal

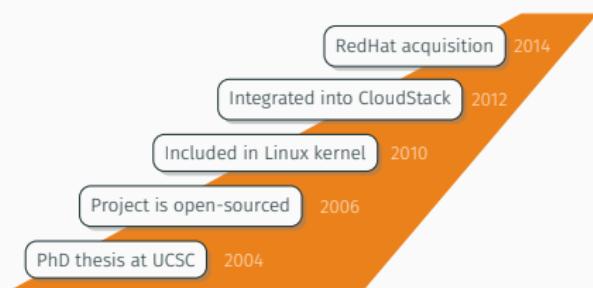
History



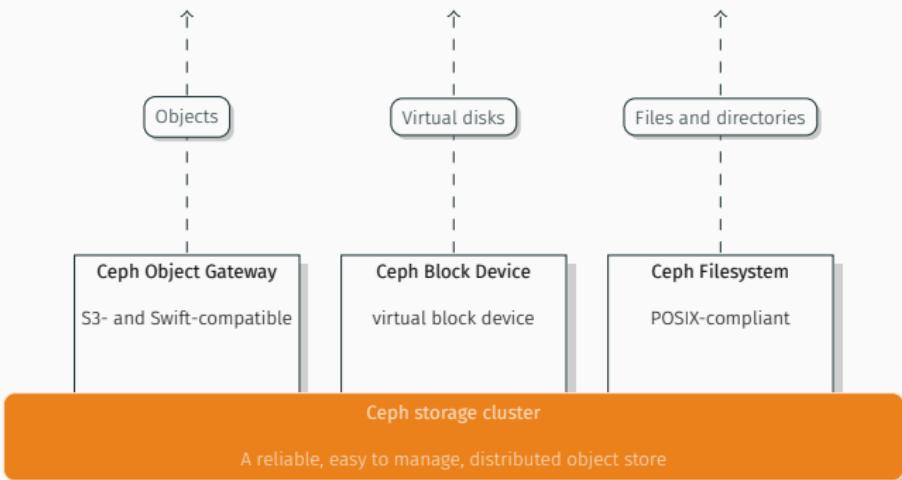
Philosophy

- open-source
- community focused
- software-defined
- scale-out hardware
- self-managing
- failure is normal

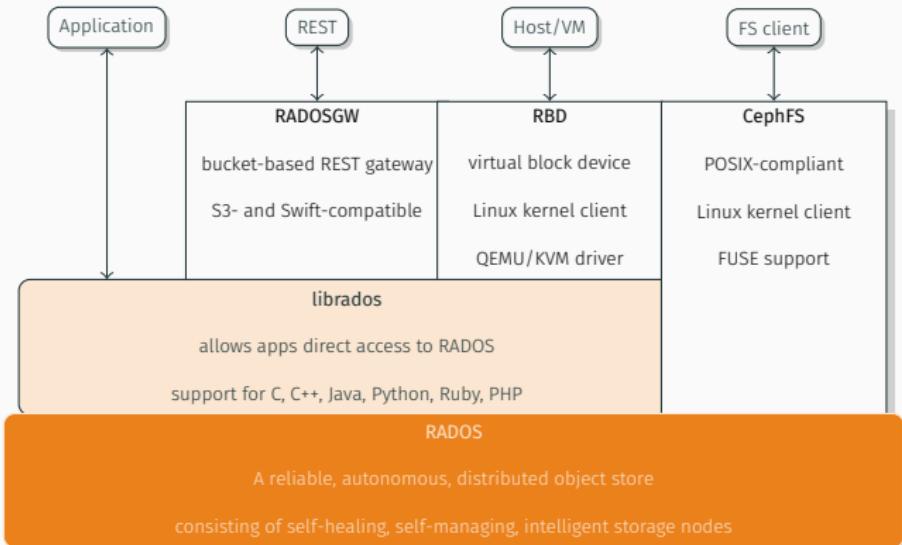
History



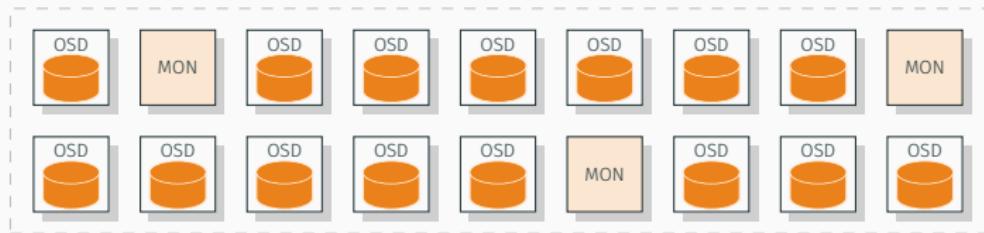
CEPH ARCHITECTURE



CEPH ARCHITECTURE



RADOS



CEPH DAEMONS

OSD

- serve objects to clients
- one per disk
- backend: btrfs, xfs, ext4
- peer-to-peer replication and recovery
- write-ahead journal

MON

- maintain cluster state and membership
- vote for distributed decision-making
- small, odd number

DATA PLACEMENT

HOTELS



<http://free-stock-illustration.com/hotel+key+card>



<http://2.bp.blogspot.com/-o-rlrv094E/TXxj8D-B2LI/AAAAAAAAGhb/VEbrbHpxVxo/s1600/DSC02213.JPG>

HOTELS

- What if the hotel had 1 billion rooms? Or ∞ ?

#13,565,983

What if the hotel changed constantly?



<http://waltonian.com/news/eastern-library-renovations-continue/>

Scale-up everything?



http://www.millenniumhotels.com/content/dam/global/en/the-heritage-hotel-manila/images/cons-photographics-lobby-reception-desk%2003062011_34-basicB-preview-2048.jpg

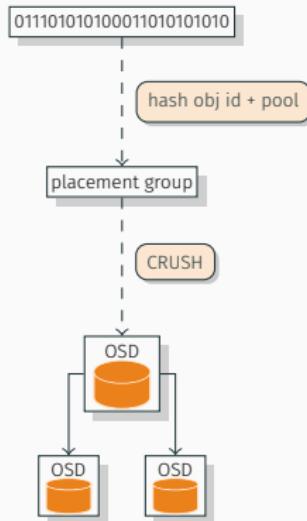
HOTELS

- The hotel itself must assign people to rooms instead of a centralized place
- The hotel should grow itself organically

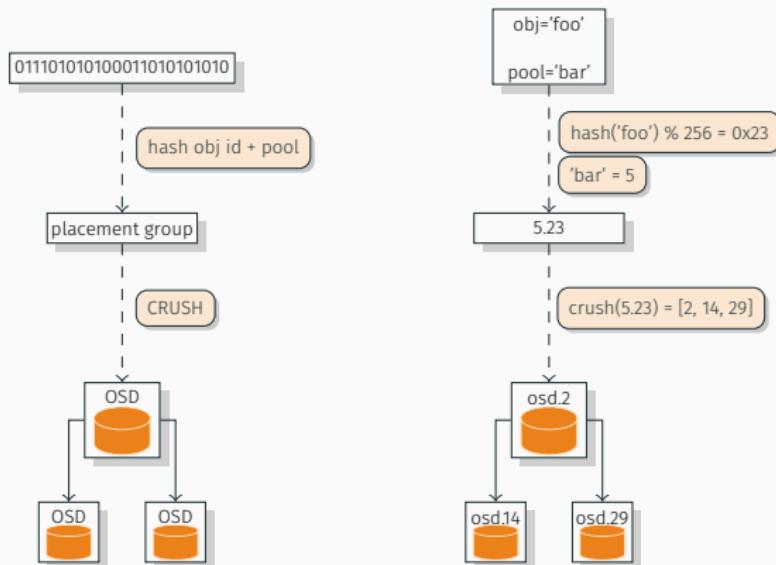
HOTELS

- The hotel itself must assign people to rooms instead of a centralized place
- The hotel should grow itself organically
- Deterministic placement algorithm
- Intelligent nodes

CRUSH



CRUSH

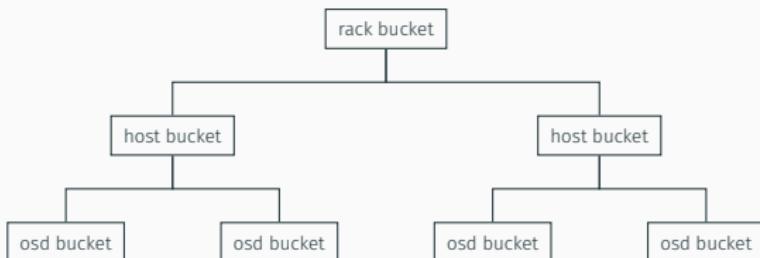


Controlled Replication Under Scalable Hashing

- Pseudo-random placement algorithm
- Repeatable, deterministic
- Statistically uniform distribution
- Stable mapping: minimal data migration
- Rule-based configuration, topology aware

Controlled Replication Under Scalable Hashing

- Pseudo-random placement algorithm
- Repeatable, deterministic
- Statistically uniform distribution
- Stable mapping: minimal data migration
- Rule-based configuration, topology aware



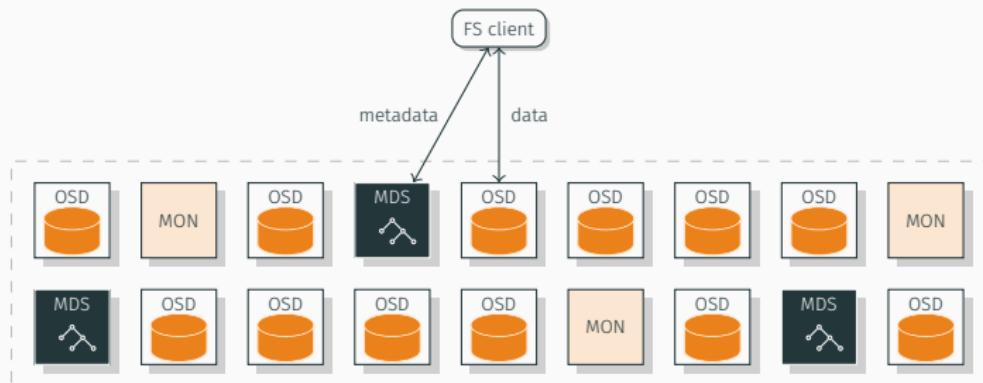
CEPH CLIENTS

- direct access to RADOS for applications
- C, C++, Python, Java, Erlang, PHP
- native socket access, no HTTP overhead

- RESTful API
- unified object namespace
- S3 and Swift compatible
- user database and access control
- usage accounting, billing

- Storage of disk images in RADOS
- Images are striped across the cluster
- Decoupling of VMs from host
- Thin provisioning
 - physical storage only used once you begin writing
- Snapshots, copy-on-write clones
- Support in Qemu, KVM

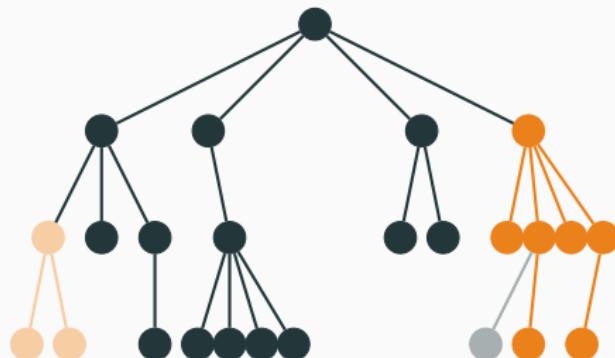
CEPHFS



Metadata Server

- Manages metadata for POSIX-compliant filesystem
 - directory hierarchy
 - file metadata: owner, timestamps, mode etc
- Stores metadata in RADOS
- Multiple MDS for HA and load balancing

DYNAMIC SUBTREE PARTITIONING

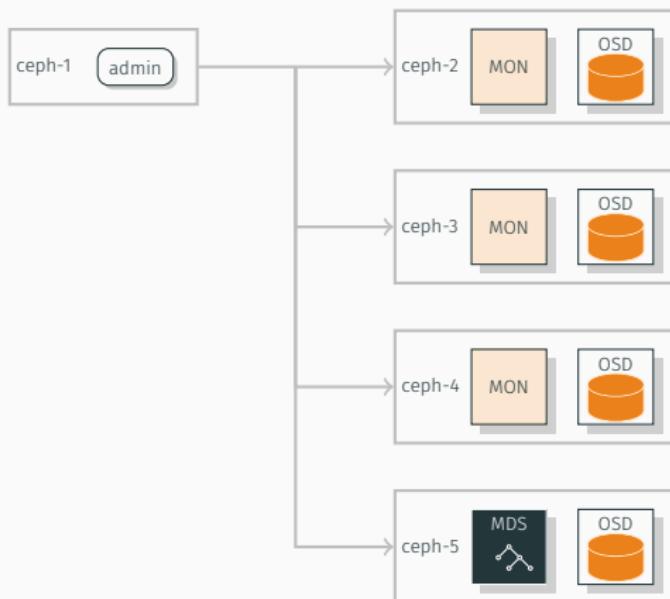


TUTORIAL

OVERVIEW

- Deploy a Ceph cluster
- Simple operations with the storage cluster
- CRUSH
- RBD
- CephFS
- Advanced topics: erasure coding, cache tiering

CLUSTER SET-UP



QUESTIONS?