CEPH and Mass Storage

how to use CEPH in CASTOR

CERN IT Department
CH-1211 Genève 23
Switzerland

Goal and philosophy

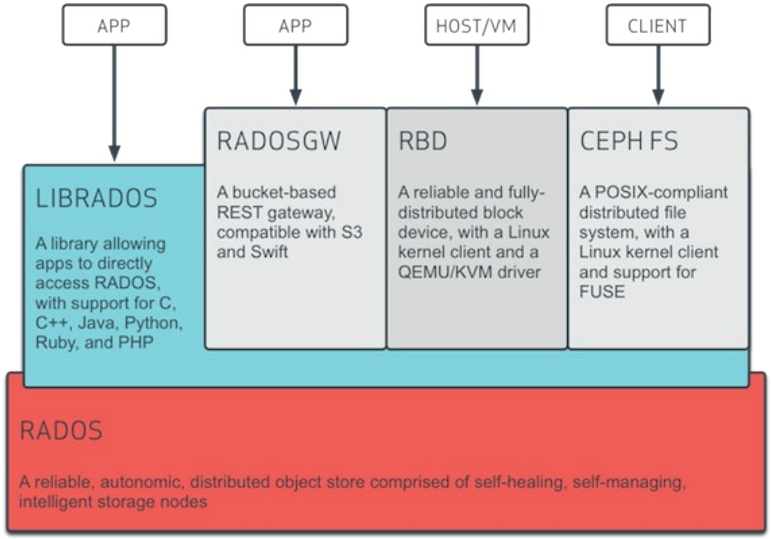
- Disk management layer is not specific to any Mass Storage solution
- Dedicated tools do it better (& cheaper) than us
  - Striping, replication, draining, rebalancing, …
- Sharing this layer between different DSS products would have major operations gains
  - Common tools, easy machine moves
- Ceph is already tested in DSS

→ Let's try CEPH in CASTOR

CERN IT Department
CH-1211 Genève 23
Switzerland

What does CEPH provide ?

- LibRados
  - Object store
  - No striping
- RadosGW
  - S3 compatible
- RBD
  - Block device
- Ceph FS
  - Filesystem (thus kernel code)
  - Beta state



CERN IT Department
CH-1211 Genève 23
Switzerland

What do we need ?

- Storing files
  - Any size (objects should be small)
  - Having external attributes
- (tunable) reliability
  - Number of replicas
  - Erasure coding
- Performance
  - Striping of files
  - Scalability
- Easiness of operation
  - Rebalancing, draining, easy management

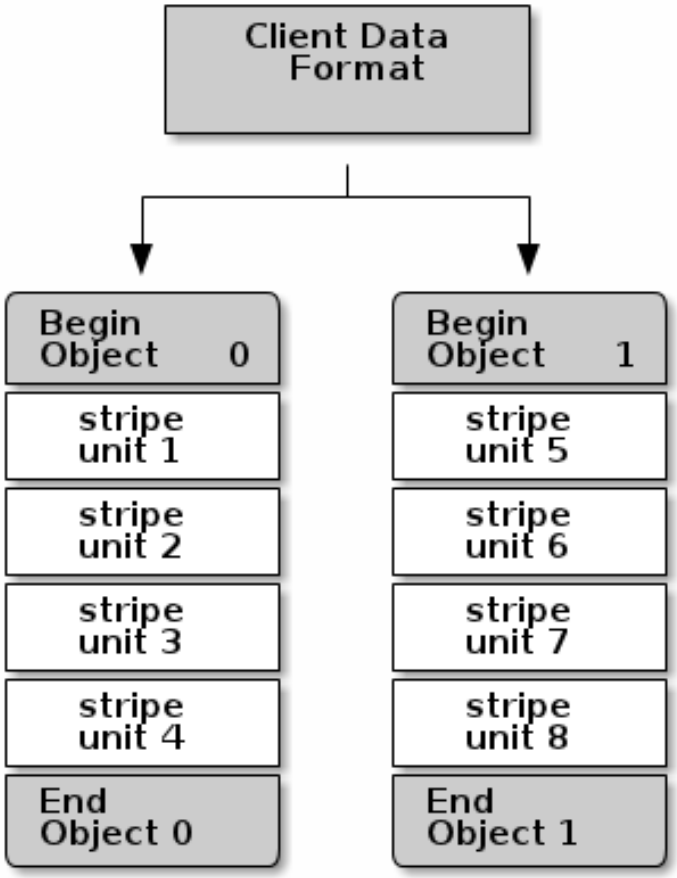
CERN IT Department
CH-1211 Genève 23
Switzerland

Best candidate is librados

- Has most things we need :
  - Scalability
  - External attributes
  - Number of replicas tunable per object
  - Erasure coding is under work
    - Andreas involved
  - Rebalancing, draining, easy management
- Is missing two points only
  - Striping
  - Support for big files
    - Objects should be rather small (1-100 MB max)

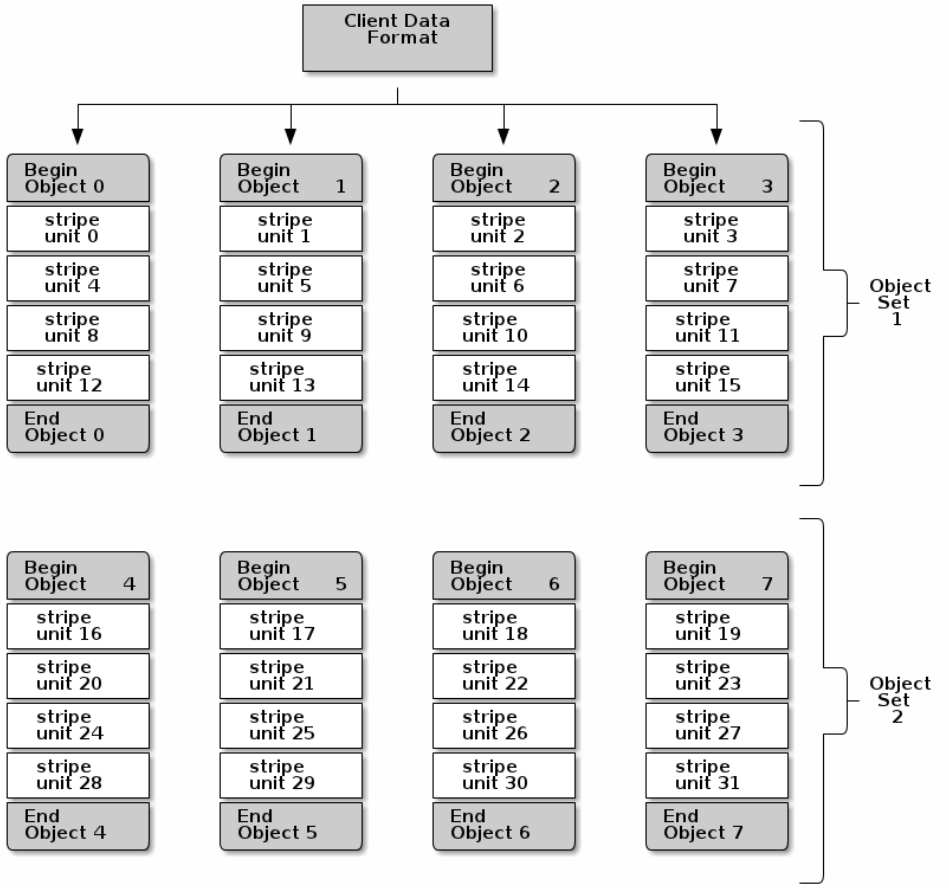
Striping is easy to add

- File is mapped to a set of objects
- Objects' names are <filename>###nb
- Demonstrated by Andreas in his cephcp
  - Excellent performances



Big files builds on striping

- Extension of striping with the notion of object sets
- Implemented in a very generic fashion in my libradosext
- To be backported to main ceph ?



CERN IT Department
CH-1211 Genève 23
Switzerland

CASTOR first prototype

- Goal :
  - replace CASTOR pools by CEPH cluster(s)
- Idea :
  - Hack rfiod to talk to libradosext rather than to local filesystem
    - Few system calls to change
  - Not touching anything from core CASTOR
    - Stager, NS, scheduling, tape
- Result :
  - 2 days were enough to have the test suite passing with ceph backend

Prototyping output

- Ceph can be used as CASTOR disk layer at very reasonable cost
  - And it can be mixed with regular pools
- It solves “for free” the tape performance problem
  - Tape can access CEPH directly and full benefit of the striping
- It eases and improves the disk scheduling
  - All “diskservers”/proxies see all files
  - Although this was not tested at this stage

From prototype to product

- CASTOR still believes that files are located on a given diskserver/now proxy
- Other protocols have to be modified
  - Including tape side
- Support for transition period is missing
  - Mixed pools
  - Replication between ceph and normal pools
- GC in ceph pools will have to be reviewed
- Global stress test / performance measurement is needed
- No validated CEPH service

CERN IT Department
CH-1211 Genève 23
Switzerland

Road map

- Provide generic file access to rados
  - Can be reused by other DSS projects
- First prototype in CASTOR
  - Proof of concept
- Complete CASTOR-ceph prototype
  - Including GC and tape rados access
- Build and test a CEPH service
  - Operation team involved
- Run heavy stress/performance tests
  - Including disk and tape
- Have ceph as an option in 2.1.15 (spring)
  - With support for mixed pools

CERN IT Department
CH-1211 Genève 23
Switzerland