“Big Storage, Little Budget”

Kyle Hutson

Adam Tygart

Dan Andresen

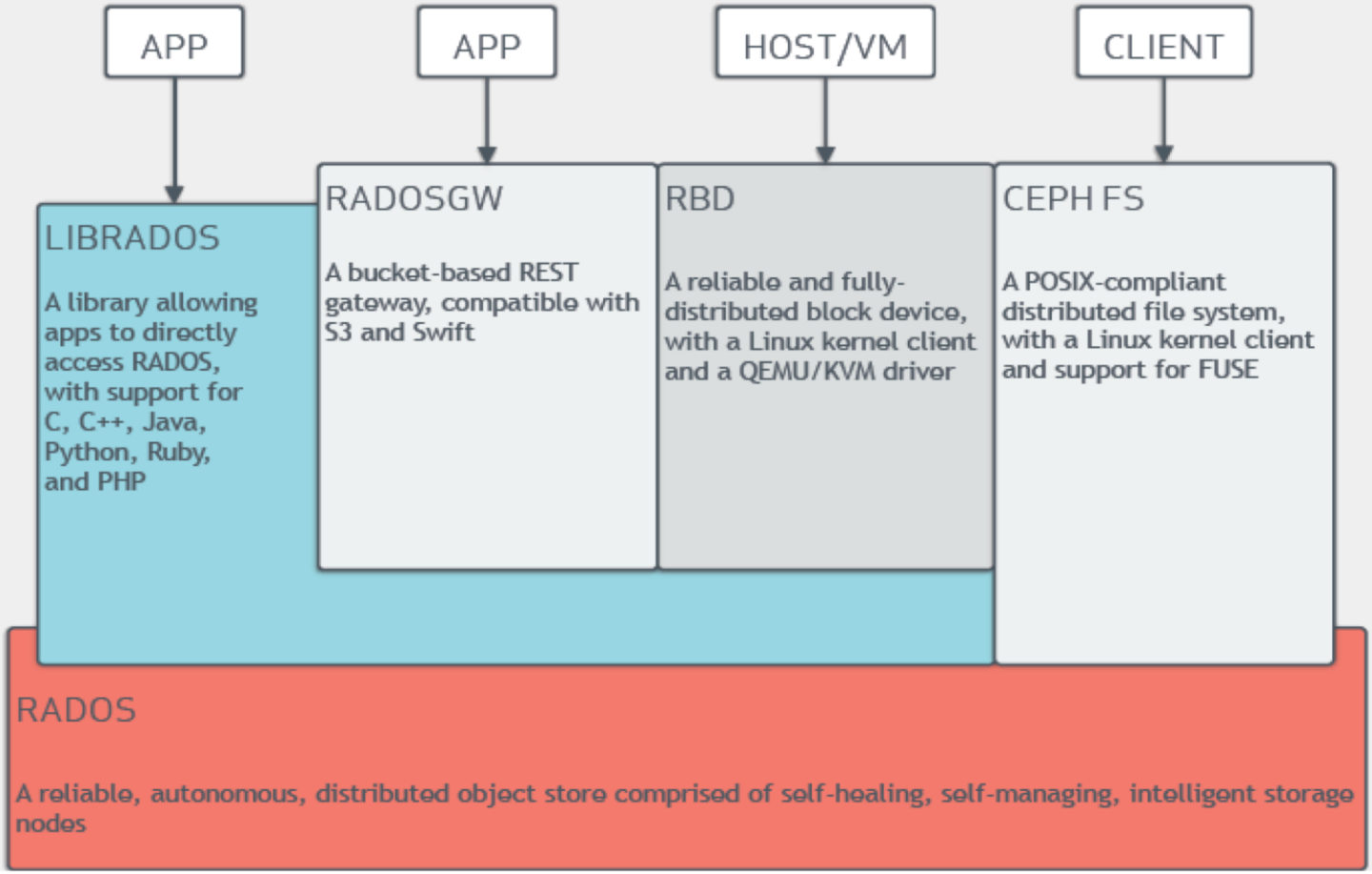
Background

- Previously using glusterfs
- Running out of disk space
- Used Ceph for ~1.5 years for backup space

About Ceph

- Primarily Object Storage
- Grown from Sage Weil's PhD thesis on storage systems
- Baseline:
  - RADOS ("reliable, autonomous, distributed object store")
  - CRUSH maps ("controlled replication under scalable hashing")

About Ceph (continued)



Ceph Daemons

- OSD ("object storage device") – usually an entire disk, but can use a partition. Where the data is saved. It is also a daemon which client computers can talk to.
- Monitor – a daemon that talks to clients and keeps track of the CRUSH map and where data is located
- MDS – Metadata server – a daemon which keeps track of CephFS metadata

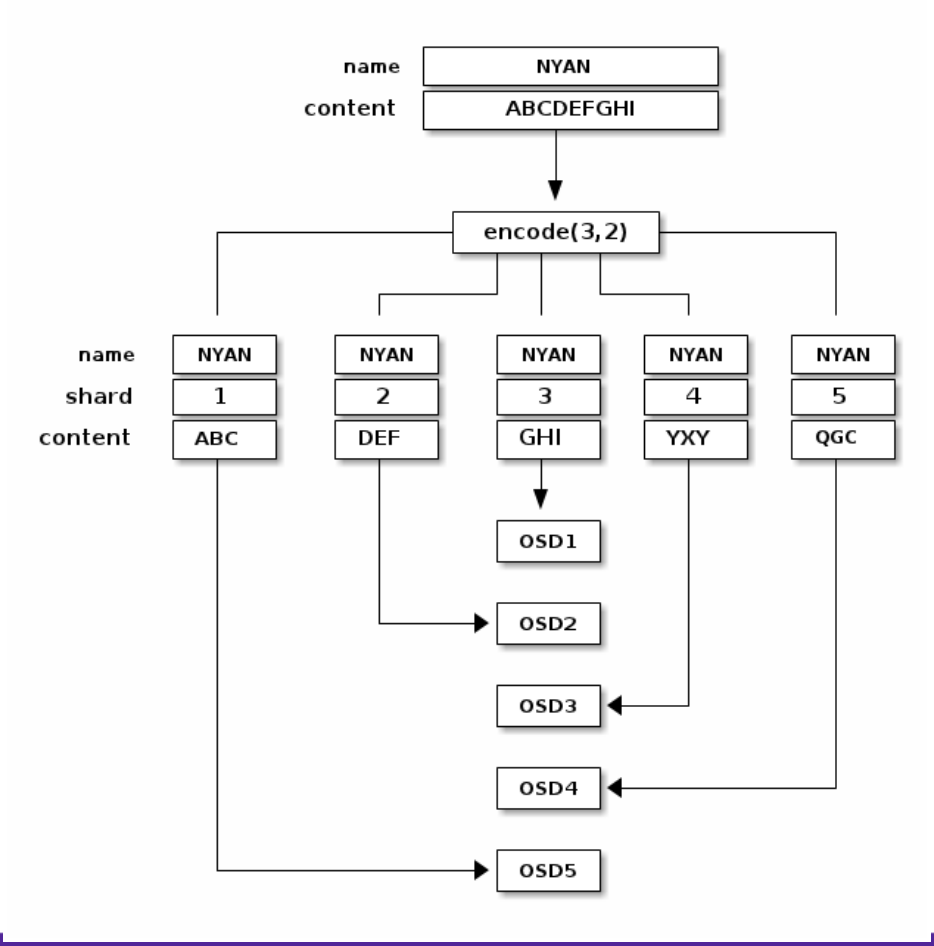
Our Hardware

- (24) Dell R730xd servers, each with
  - (2) E5-2650L (12-core) CPUs
  - (12) 6 TB HDDs
  - (4) 4 TB HDDs
  - (Total of 88 TB HDDs/server)
  - (2) 480 GB SSDs
  - 128 GB RAM
- Cost: $300,000
- Raw Storage: 2.1 PB

Cool Concepts

- Erasure Coded Pools
- Cache Tiering
- CephFS
- Scalability
- Very fast metadata – active/passive currently with active/active promised
- CRUSH hierarchy

Cool Concepts – Erasure coding



Timeline

- January: Hardware arrives, we begin playing
  - Erasure coding with cache tiers
- May: First attempt to cutover
- July: Actual cutover
- Present: Still fighting performance issues

Final configuration

- Erasure coding with m=8, k=4
- ~1.3 PB usable disk space
- Journals on OSD (HDDs)
- Metadata on SSDs
- Cache on HDDs

Future plans

- Multi-site archive
- Hadoop

Advice for the reader

- Get guidance on your proposed hardware
  - ‘ceph-users’ mailing list or on IRC
- Don’t skimp on quality of SSDs
- Cache tier sizing
- Testing, testing, testing