

OPEN SOURCE STORAGE FOR DIGITAL PRESERVATION

SAGE WEIL LOC - 2015.09.10

WHAT IS CEPH



- Scale-out distributed storage
- Self manage whenever possible
- Fault tolerant no single points of failure
- Storage hardware agnostic
- Commodity components
- Single cluster, multiple protocols
 - Object, Block, File
- Free and open source

VALUE OF OPEN SOURCE FOR ARCHIVES



- Cost at scale
- Hardware vendor independence
 - Drives down cost
 - Price vs performance vs robustness
- Software vendor independence
 - Data lifetime far exceeds vendor lifetime
- Transparency
 - How do you read your data in 10, 20, 50 years?
 - Data is not hostage to proprietary platform source code is open
- Efficient investment of tax dollars
 - Technology investment benefits all users, not a single vendor

CEPH COMPONENTS



APP



RGW

A web services gateway for object storage, compatible with S3 and Swift HOST/VM



RBD

A reliable, fully-distributed block device with cloud platform integration

CLIENT



CEPHFS

A distributed file system with POSIX semantics and scale-out metadata management

LIBRADOS

A library allowing apps to directly access RADOS (C, C++, Java, Python, Ruby, PHP)

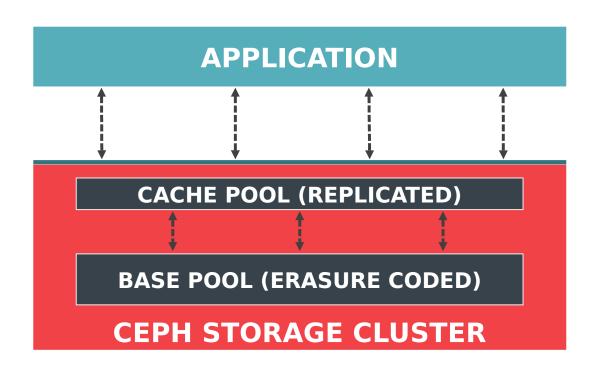
RADOS

A software-based, reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes and lightweight monitors

RADOS CACHE TIERING



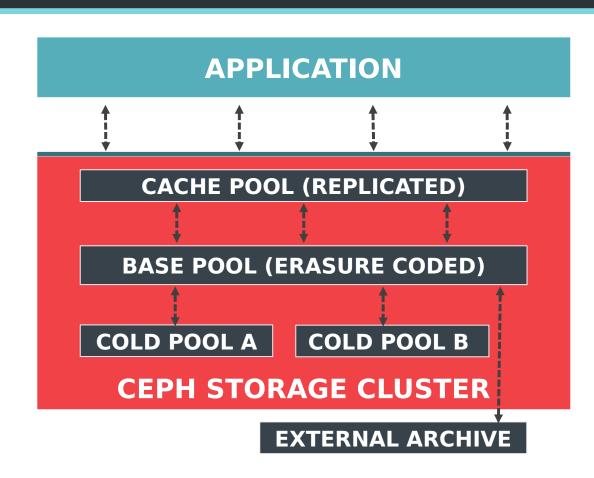
- Cache slow pool with fast pool
 - Replication vs erasure coding
 - SSDs vs HDDs
 - Fast vs slow servers
- Widely deployed today
- Range of erasure codes available
 - Pluggable algorithms
 - LRC (local recovery codes)



ADDITIONAL RADOS COLD TIERS



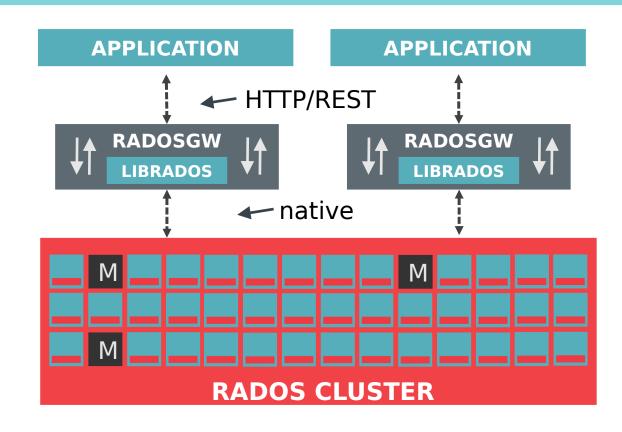
- Planned
- Push objects to slow tiers
 - Pluggable backend
 - RADOS pools
 - External archives (e.g., S3)
 - Tape
 - ?
- Access schedules for cold tiers
 - Power down; block or fail reads
 - Absorb writes in cache



RGW (RADOS GATEWAY)



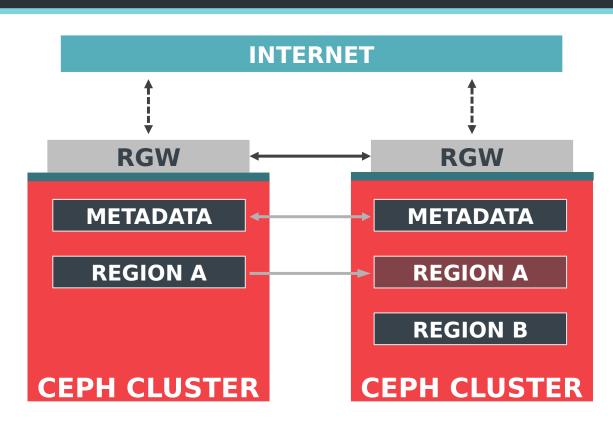
- User-facing object storage API
 - S3 and Swift compatible
 - Flexible security
- Stateless proxy
 - Scale gateways horizontally
 - Combine with load balancer, caching proxy, etc.
- Consumes RADOS pools
 - Replicated or erasure coded



RGW FEDERATION



- Multi-site
 - Federate multiple "zones"
- Global user/bucket namespace
- Asynchronous replication
 - Zone to zone
 - Eventually consistent
 - Master/slave (today)
 - Active/Active (coming soon)
- Disaster recovery across DCs
 - Dynamic DNS, redirects, etc.



QUESTIONS



- Do archive maintainers want to combine active and cold archives?
- Powered down HDDs or tape?
- Granularity of storage policy?
 - Per-object? bucket? pool?
- What APIs do users want beyond vanilla S3 (get/put/remove)?

THANK YOU!

Sage Weil
CEPH PRINCIPAL ARCHITECT



sage@redhat.com



@liewegas

