What’s new in Kraken

XSKY Haomai Wang

2016.12.09

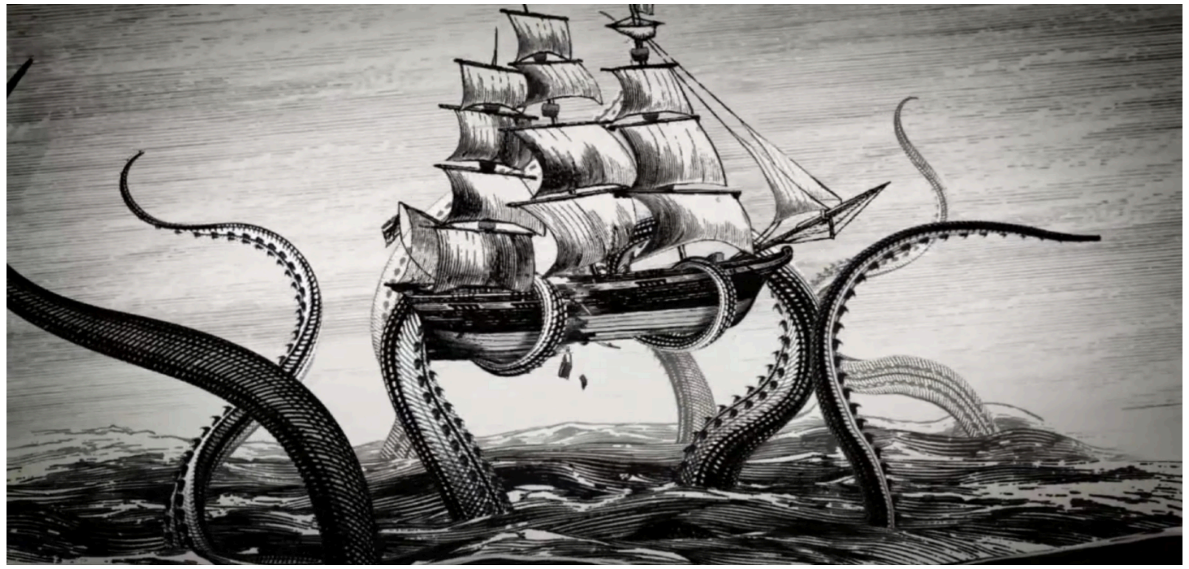
X·SKY
www.xsky.com

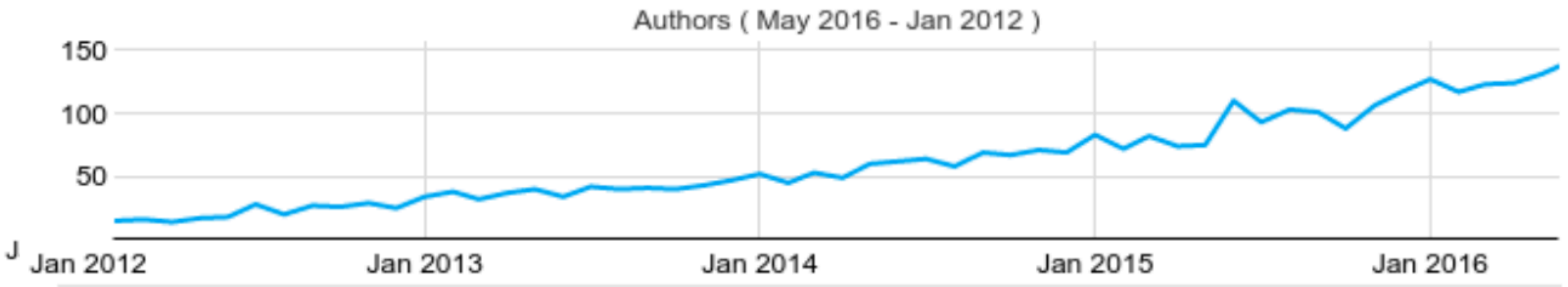
CREATE THE **ECOSYSTEM**
TO BECOME THE **LINUX**
OF DISTRIBUTED **STORAGE**

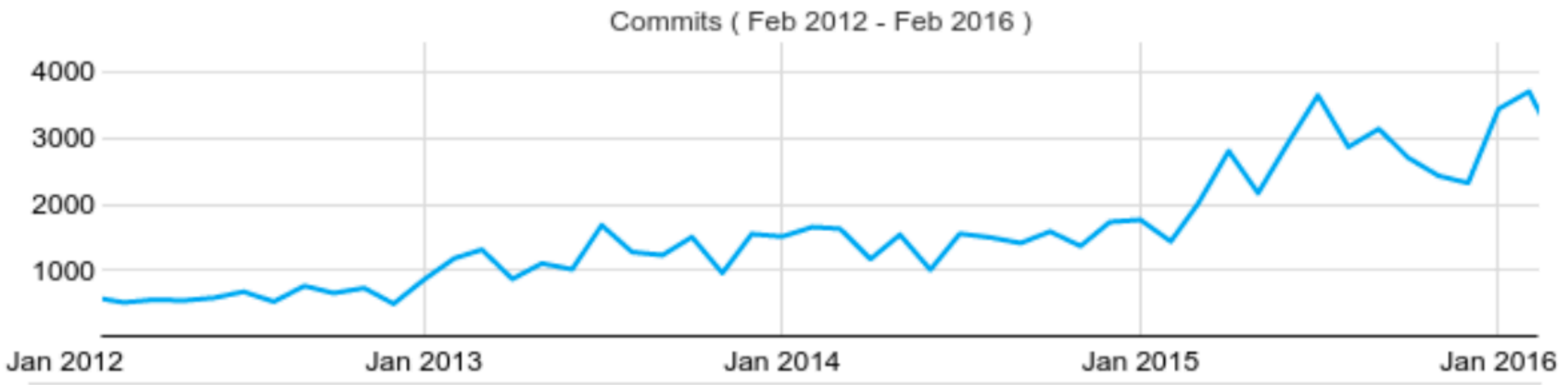
OPEN COLLABORATIVE GENERAL PURPOSE

Releases

- Hammer v0.94.x (LTS) – March '15
- Infernalis v9.2.x – November '15
- Jewel v10.2.x (LTS) – April '16
- Kraken v11.2.x – December '16
- Luminous v12.2.x (LTS) – April '17

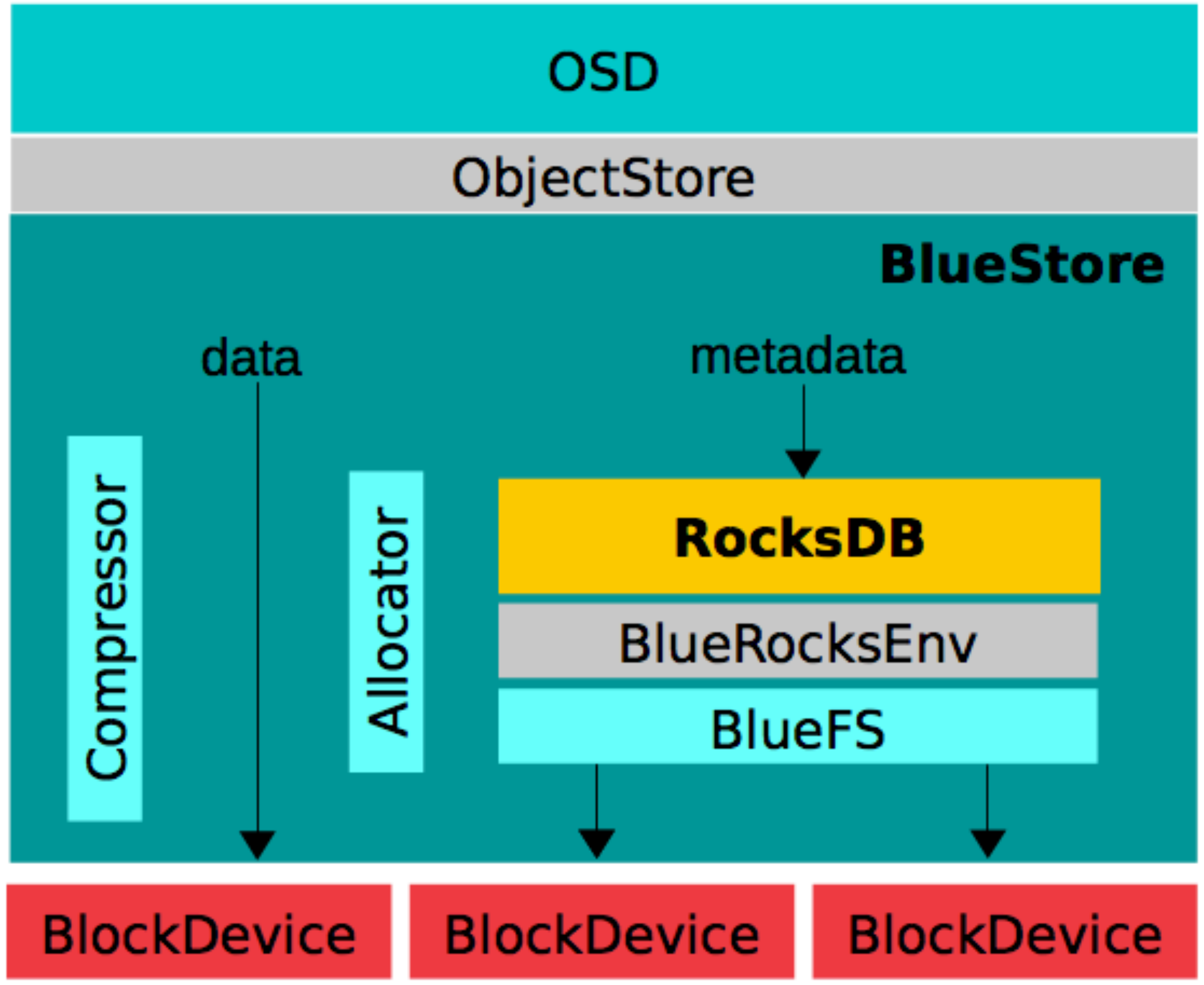






RADOS -- BlueStore

- BlueStore = Block + NewStore
- Key/value database(RocksDB) for metadata
- All data written directly to raw block device(s)
- Inline compression(zlib, snappy, zstd)
- ~2x faster than FileStore



RADOS -- BlueStore

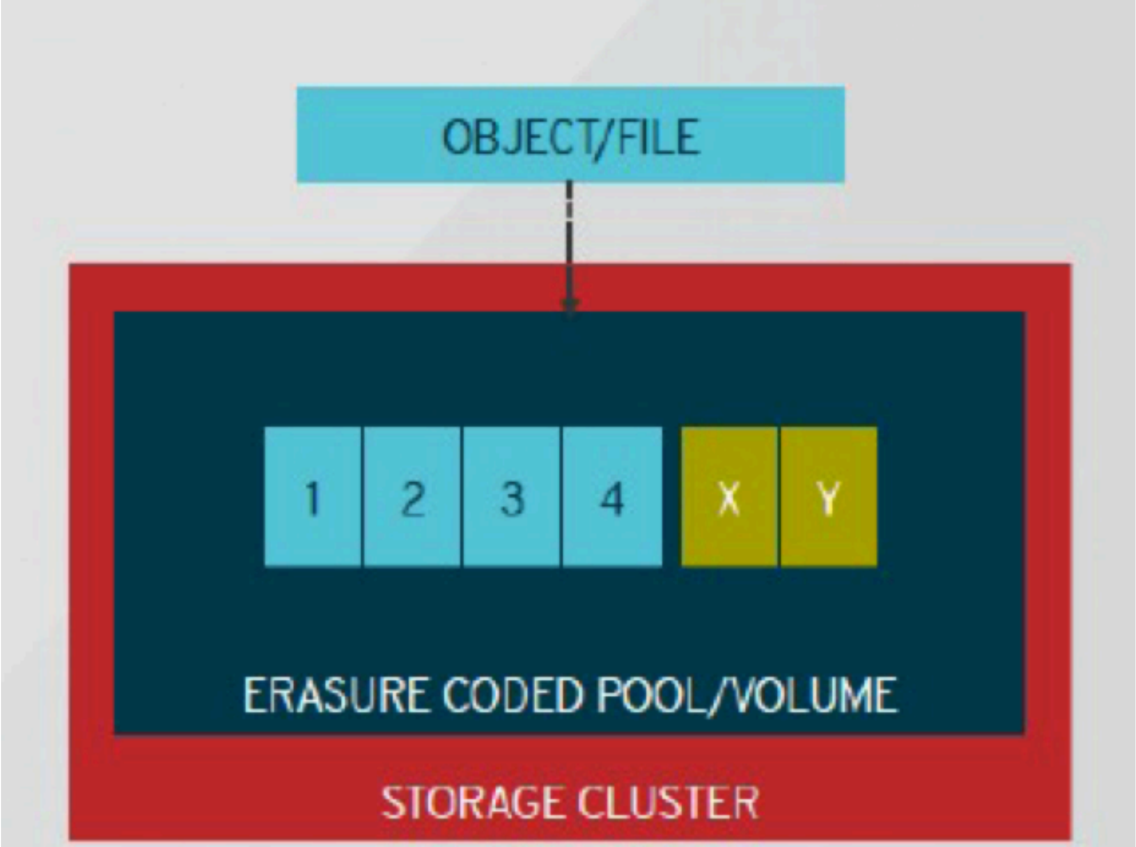
- A stable disk format
- Passing failure and stress testing
- Still flagged as experiment feature
- Non-production and non-critical env

- Luminou
- Remove experiment feature
- Full stable and ready for broad usage

RADOS – Erasure Code Overwrite

- Experiment feature
- Disk format and implementation are not stable

- Initial RBD with EC Overwrite testing
- Sequential write performance looks good!
- reads suffer vs 3x replication as expected
- small random writes also suffer as expected

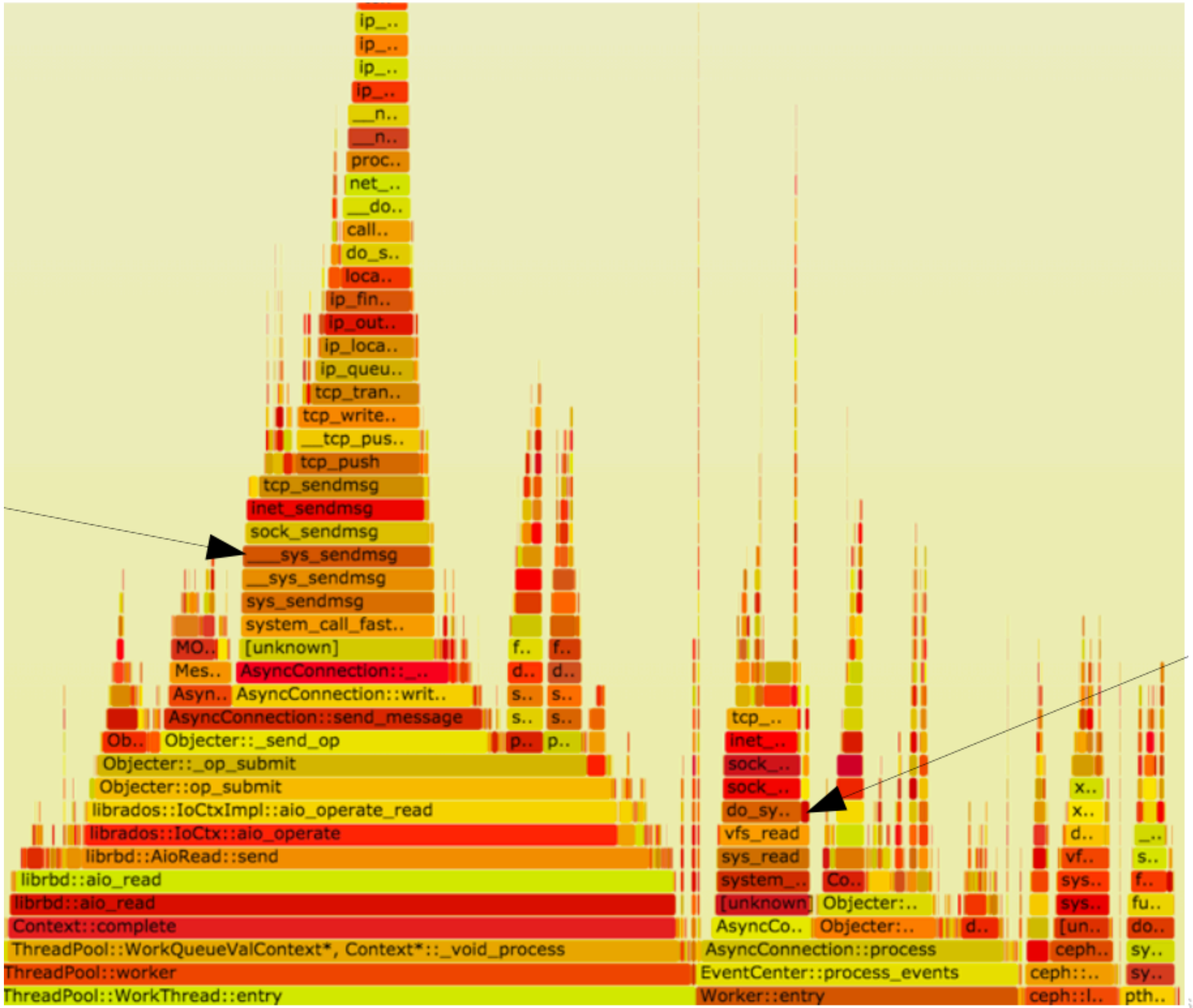


RADOS – AsyncMessenger

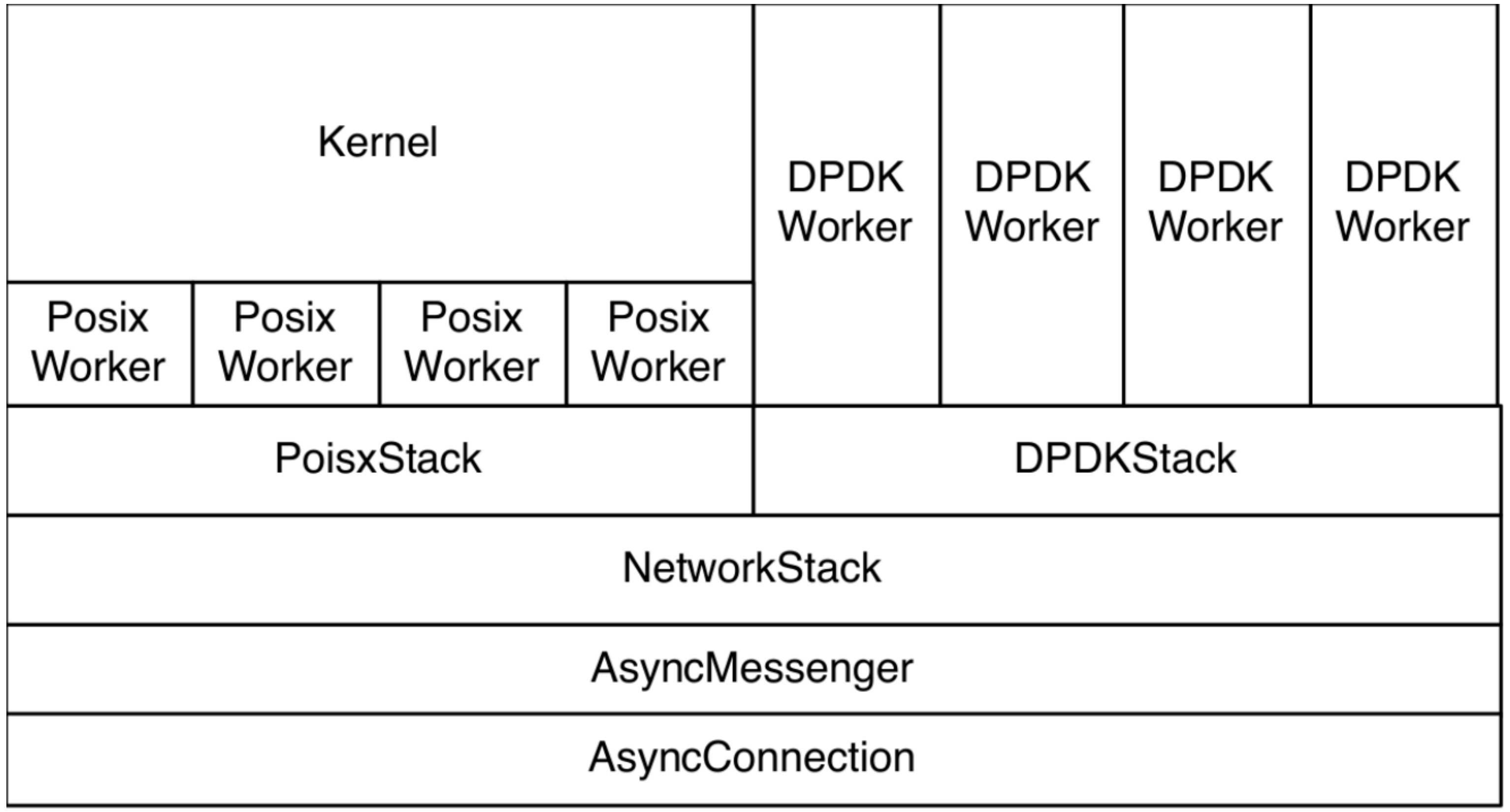
- New implementation of network layer
- replaces aging SimpleMessenger
- fixed size thread pool (vs 2 threads per socket)
- scales better to larger clusters
- more healthy relationship with tcmalloc
- now the default!

- Pluggable backends
- PosixStack – Linux sockets, TCP (default, supported)
- Two experimental backends!

RADOS – AsyncMessenger



RADOS – AsyncMessenger Plugin



Kernel

Posix Worker | Posix Worker | Posix Worker | Posix Worker

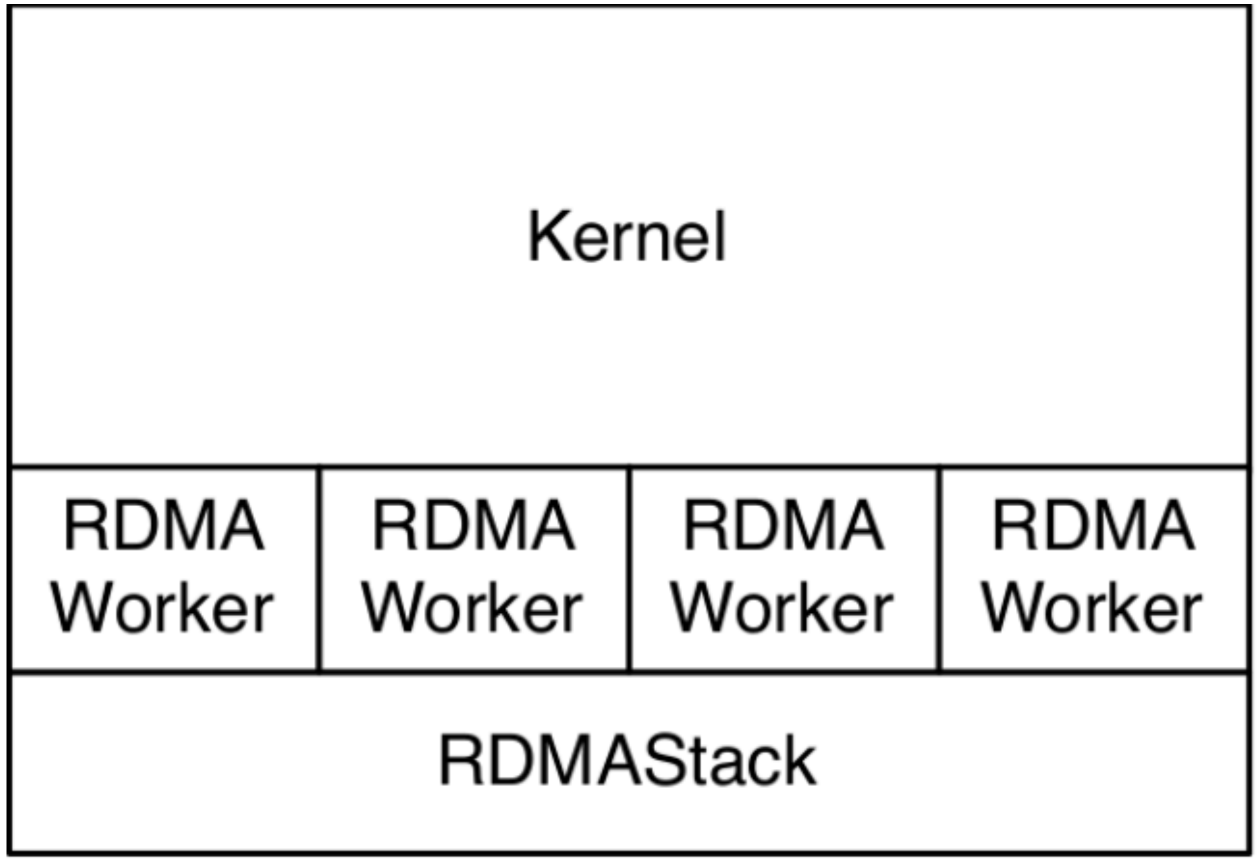
DPDK Worker | DPDK Worker | DPDK Worker | DPDK Worker

PoisxStack | DPDKStack

NetworkStack

AsyncMessenger

AsyncConnection



Kernel

RDMA Worker | RDMA Worker | RDMA Worker | RDMA Worker

RDMAStack

Ceph-MGR

- ceph-mon monitor daemons currently do a lot
- more than they need to (PG stats to support things like 'df')
- this limits cluster scalability

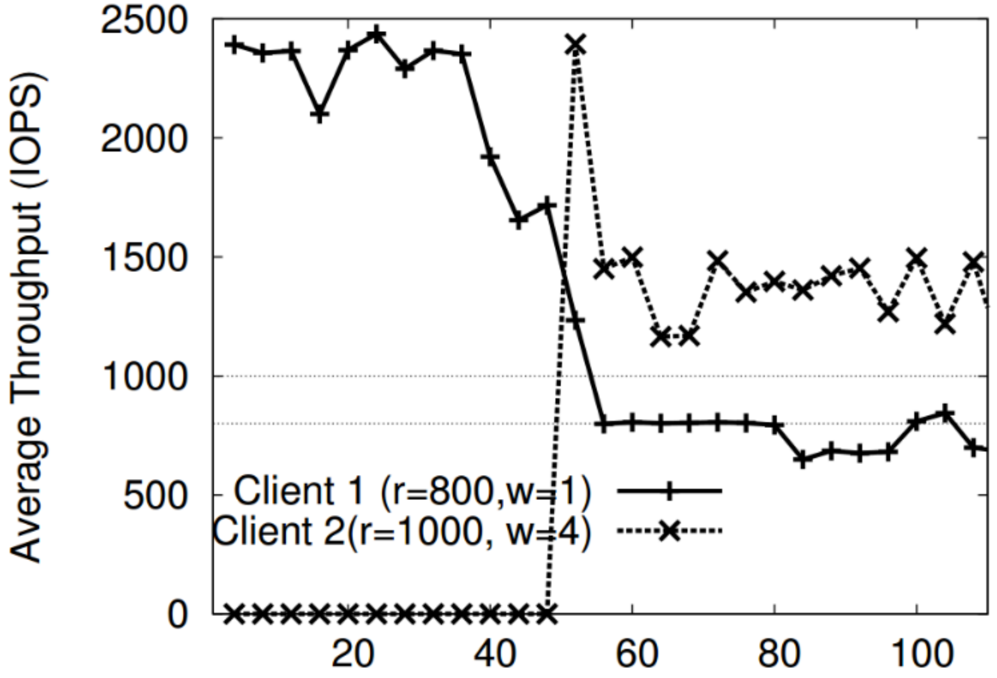
- ceph-mgr moves non-critical metrics into a separate daemon
- that is more efficient
- that can stream to graphite, influxdb
- that can efficiently integrate with external modules (even Python!)

???
(time for new iconography)

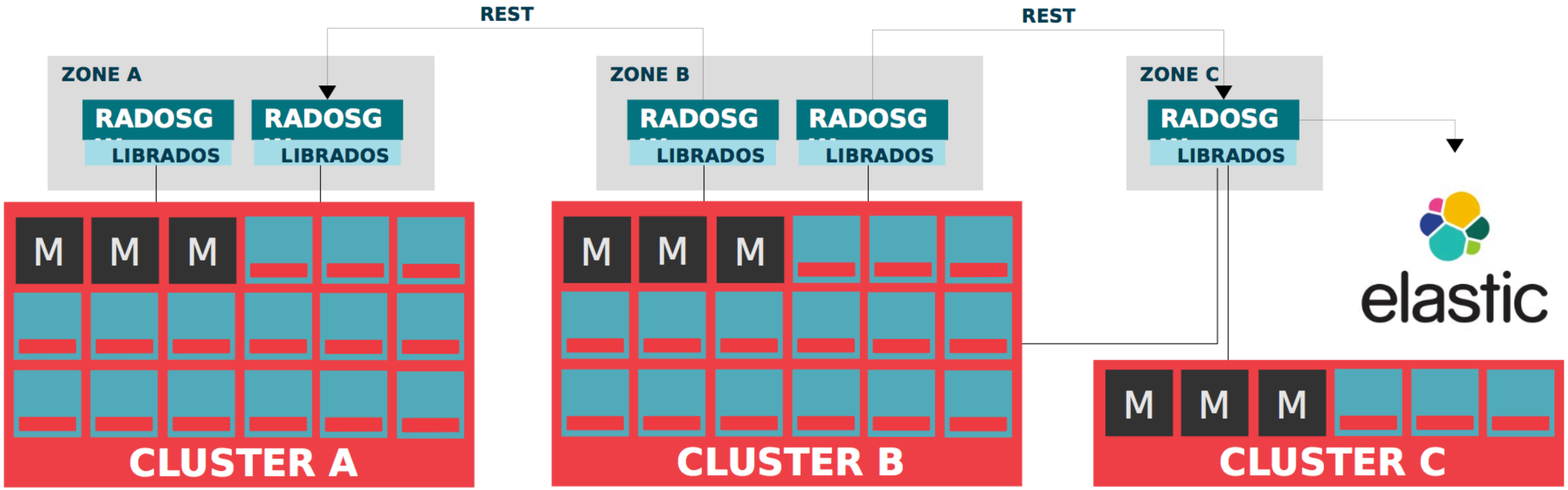
- Good host for
- integrations, like Calamari REST API endpoint
- coming features like 'ceph top' or 'rbd top’
- high-level management functions and policy

Ceph: QoS(DMClock)

- Set policy for both
- reserved/minimum IOPS
- proportional sharing of excess capacity
- by
- type of IO (client, scrub, recovery)
- pool
- client (e.g., VM)
- Based on mClock paper from OSDI'10
- IO scheduler
- distributed enforcement with cooperating clients



RGW: Metadata Search



RGW: MISC

- Compression
- Swift static website API
- S3 lifecycle API
- Custom search filter for LDAP auth
- Python binding for rgwfile

WORM

S3

Swift

Tiering

Erasure coding

Compression

Deduplication

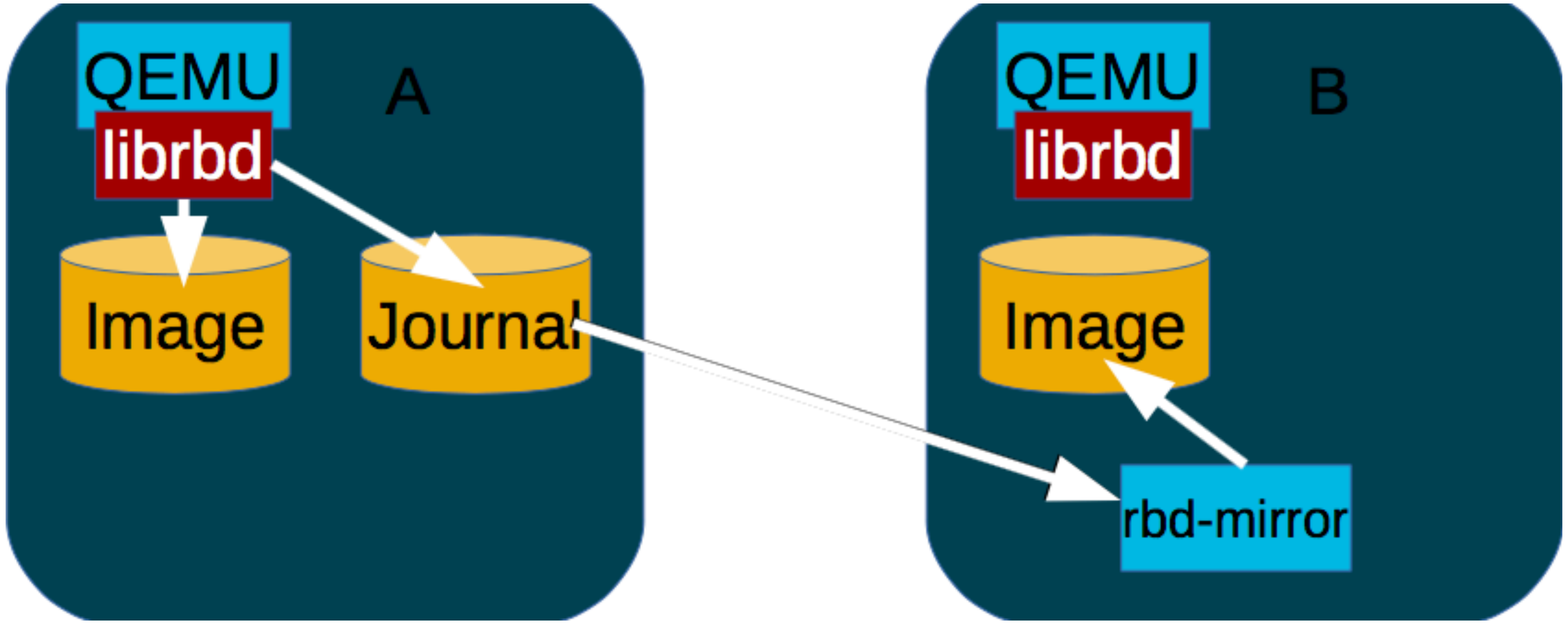
Multisite federation

Encryption

Multisite replication

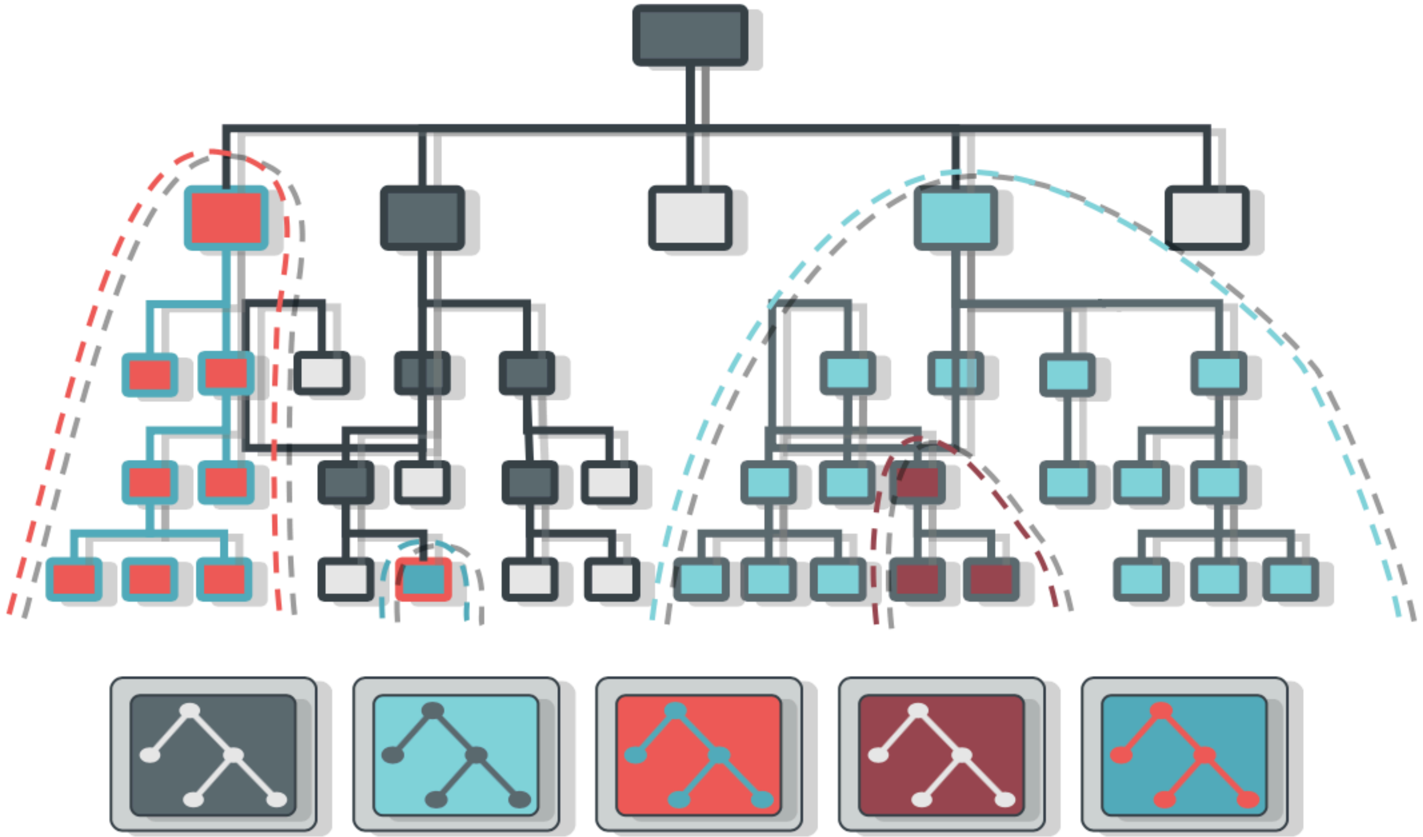
RBD: MISC

- Support erarsure coding pool
- RBD Mirror can support metadata replication
- Rbd python api supports async opeartions



CephFS: MISC

- Libcephfs enable proper uid/gid control
- A new `pg_files` subcommand to `cephfs-data-scan` can identify files affected a lost PG
- The false-positive "failing to respond to cache pressure" warnings have been fixed
- Standby replay MDS daemons now consume less memory on workloads doing deletions
- Scrub now repairs backtrace, and populates `damage ls` with discovered errors.



Upgrade

- All clusters must first be upgraded to Jewel 10.2.z before upgrading to Kraken 11.2.z (or, eventually, Luminous 12.2.z).
- New Mon will use rocksdb as backend

How To Help

Operator

- File bugs

  http://tracker.ceph.com/

- Document

  http://github.com/ceph/ceph

- Blog
- Build a relationship with the core team

Developer

- Fix bugs
- Help design missing functionality
- Implement missing functionality
- Integrate
- Help make it easier to use
- Participate in monthly Ceph Developer Meetings
  - video chat, EMEA- and APAC-friendly times

微信公众号

XSKY 微信公众号

了解XSKY 最新资讯，产品，信息，企业级解决方案，参加线上活动，请关注此公司官方微信公众号。

Ceph开发每周谈

豪迈面向Ceph社区与开源爱好者，总结Ceph社区每周开发进展的最新资讯，更加偏重于研发与方向。

福叔讲存储

为企业级存储解决方案量身打造，结合福叔在数据存储与管理方面多年的经验，推荐业内以及企业级存储运维人员关注。

Thank you

X·SKY

www.xsky.com