



数据驱动 存储创新

2017企业存储技术峰会·北京站

2017 ENTERPRISE STORAGE TECHNOLOGY SUMMIT BEIJING

2017.01.18 · 北京东方美爵酒店

主办方

日知录技术社区



Ceph 的现在与将来

- 王豪迈 XSKY CTO

大纲

- 现在
 - 什么是 Ceph ?
 - 为什么是开源的 ?
 - 为什么是统一存储 ?
 - OpenStack
- 未来
 - 路线图
 - 后续
 - 社区

关于 Ceph

- 分布式存储
- 所有组件必须能够横向扩展
- 没有单点故障
- 完全软件实现
- 只需要普通商用服务器
- 对象，文件，块都可以在同一个集群
- 尽可能自我管理
- 开源



为什么 Ceph 是开源的？

- 避免软件厂商锁定
- 避免硬件厂商锁定
 - 并且允许异构部署
- 降低 TCO
- 透明
- 自助支持
- 增加，提供或者扩展特性



为什么 Ceph 是统一存储？

- 小集群:
 - 简化部署
 - 单一集群提供所有 APIs
 - 高效利用存储能力
- 大集群:
 - 简化管理
 - 单一工具集
 - 分布式存储引擎提供必要功能
 - 擦除码，分层，性能，监控



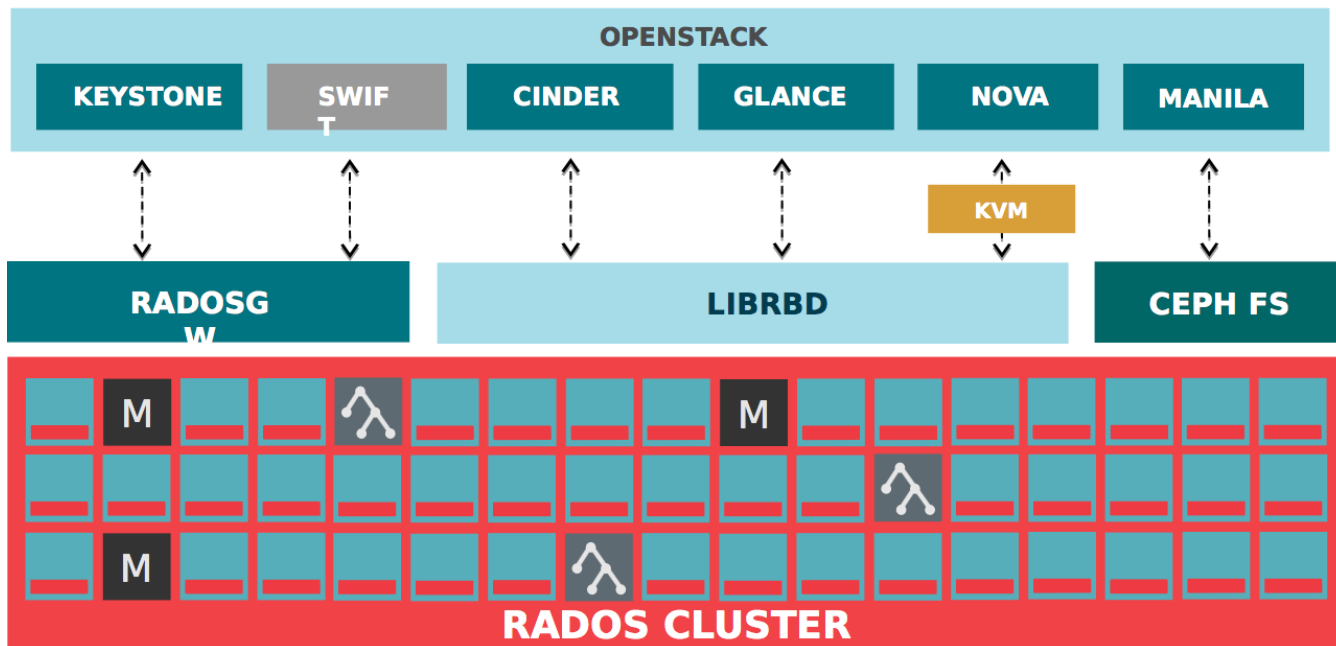
CREATE THE **ECOSYSTEM** TO BECOME THE **LINUX** OF DISTRIBUTED **STORAGE**

OPEN

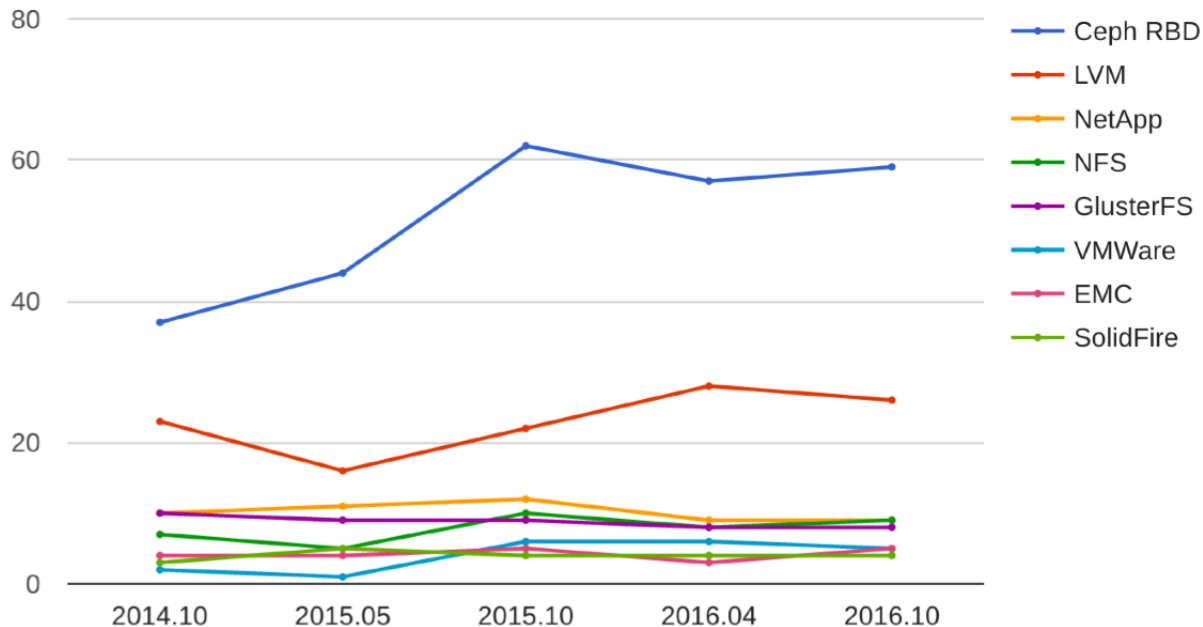
COLLABORATIVE

GENERAL PURPOSE

OpenStack



Cinder 使用后端



未来要做什么？

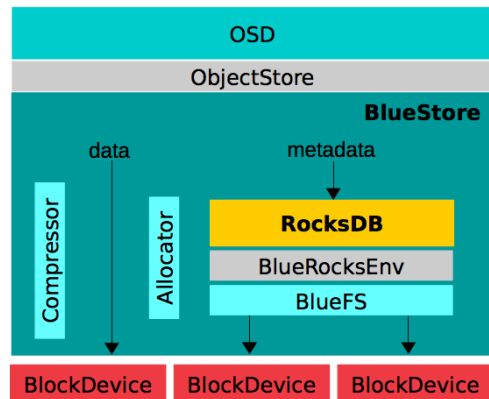
版本进展

- 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



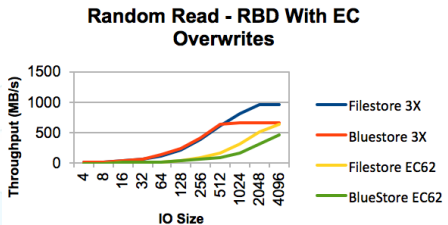
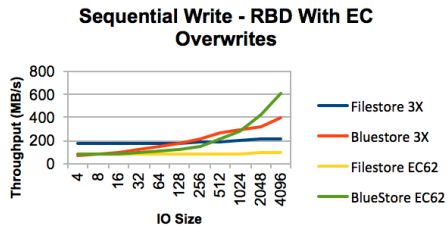
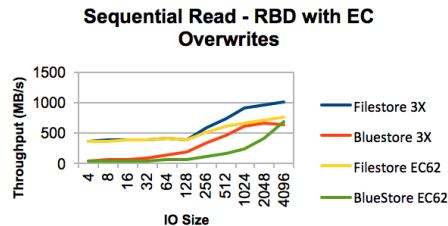
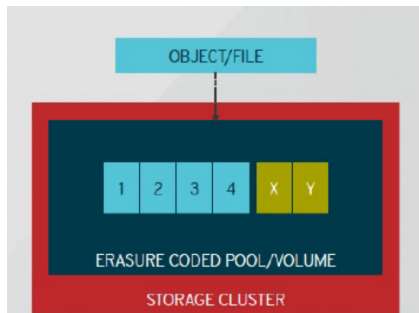
BlueStore

- BlueStore = Block + NewStore
- RocksDB: 存储元数据
- 所有数据都直接写入裸设备
- 在线压缩(zlib,snappy,zstd)
- 支持在线校验和
- ~2x 性能高于 FileStore



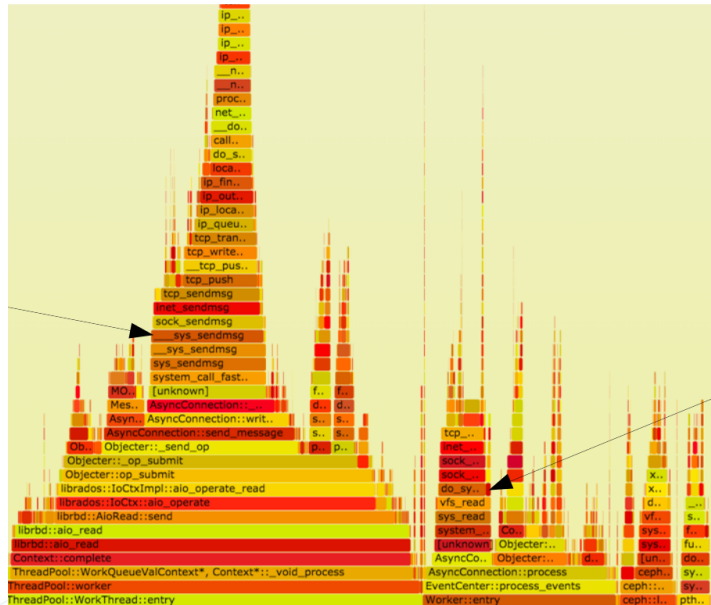
Erasure Code Overwrite

- 支持块设备使用擦除码
- 顺序读写性能跟副本差异不大
- 随机写性能比较差



支持 Infiniband/RDMA 和 DPDK

- Ceph 网络层的重构
- 支持多个网络引擎
 - 默认 Kernel
 - RDMA
 - DPDK



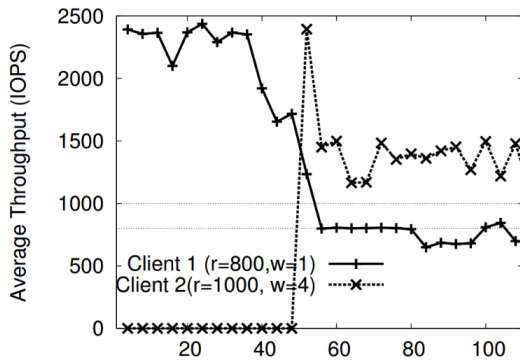
Ceph-MGR

- ceph-mon 目前承担了大量工作
 - 比如集群的 IO 统计, PG 历史信息管理等等
 - 限制集群的扩展能力
- ceph-mgr 将非关键指标移到里独立的进程
 - 使得管理功能更加高效
 - 可以将监控数据输出到 graphite, influxdb
 - 可以与其他外部模块进行整合
- 对于以下场景帮助
 - 管理面整合, 比如 Calamari REST API
 - 提供 “ceph top” 和 “rbd top ”
 - 集群整体管理功能和策略

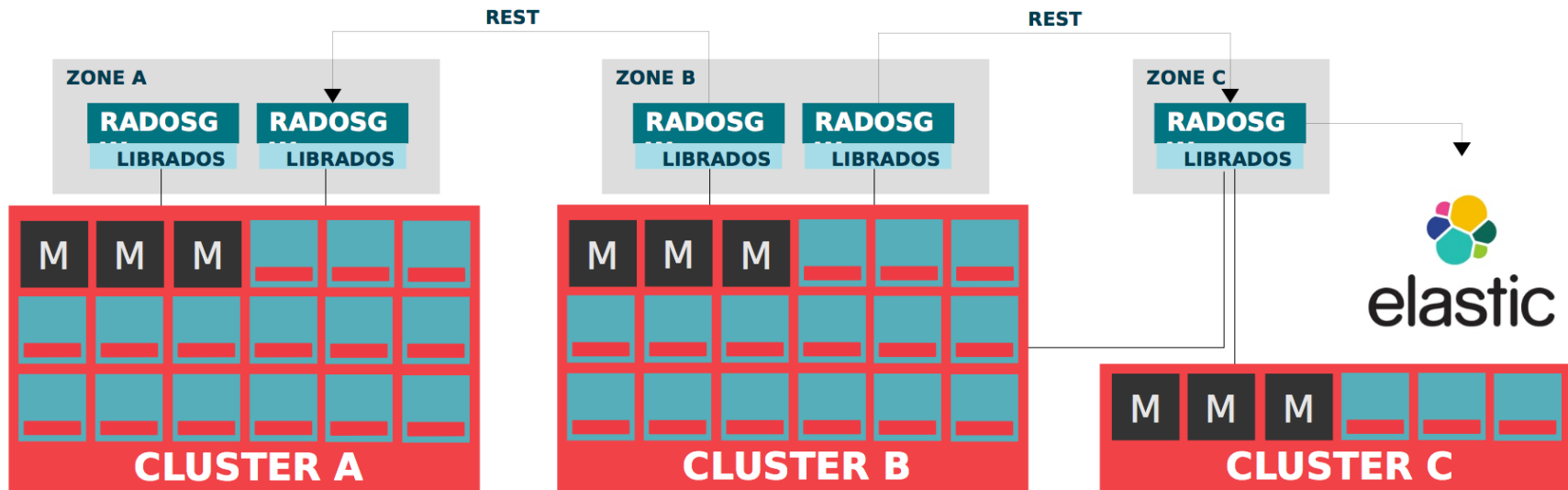


分布式 QoS(DMCClock)

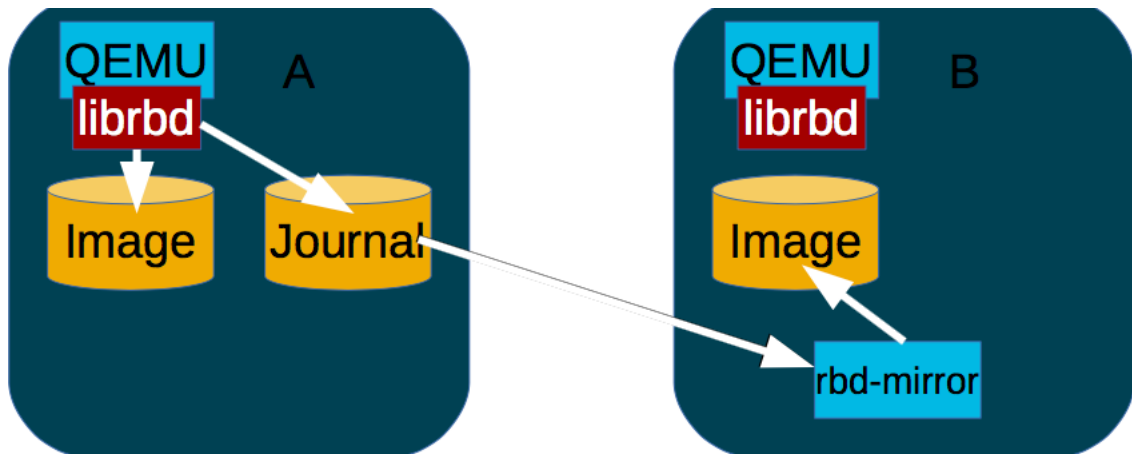
- 可以设置 QoS 策略
 - 预留/最小 IOPS
 - 根据容量提供对应比例的性能
- 识别不同的 IO 类型
 - IO 类型, Scrub/Recovery/Client
 - Pool
 - 客户端类型
- 基于 OSDI mClock 论文



RGW: Metadata Search



RBD 远程复制



感谢！



Ceph开发每周谈

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。