OPENSTACK DAYS
CHINA

UMCLOUD

MIRANTIS

Topic：深入理解 Ceph RGW 对象存储

Speaker：朱荣泽 & 任家英

议程

- 对象存储对 OpenStack 的意义
- Ceph 对象存储 RGW 原理解析
- 基于 Ceph 的对象存储的架构设计
- Ceph RGW 发展现状
- Ceph RGW 案例分享

Copyright © 2015 Mirantis, Inc. All rights reserved.

对象存储对 OpenStack 的意义

Copyright © 2015 Mirantis, Inc. All rights reserved.

云平台的常见存储需求

- 镜像存储(虚拟机镜像，容器镜像)
- 视频，音频，图片的存储
- 归档/备份数据的存储
- 大数据分析平台的存储支持
- CDN

Copyright © 2015 Mirantis, Inc. All rights reserved.

海量非结构化数据的存储

- 访问特征
  - 大块的顺序读写(blob)
  - 单次写入，多次读取(WORM)
  - 几乎不会修改
  - 文件的上传者和访问者并不是同一个人

Copyright © 2015 Mirantis, Inc. All rights reserved.

海量非结构化数据的存储

- 存储特征
  - 强调吞吐，而不是延迟
  - 容量都很大，PB 甚至 EB 级别
  - 存储生命周期长，有些甚至是永久归档
  - 被存储的数据之间是没有关系或者是弱关系的

OPENSTACK DAYS CHINA
Copyright © 2015 Mirantis, Inc. All rights reserved.
OPENSTACK DAYS CHINA

海量非结构化数据的存储

- 为复杂存储问题提供简单的使用接口
- 典型实现
  - AWS S3
  - OpenStack Swift
- 采用 HTTP 协议，RESTfull 风格的 API
- 3 个核心概念
  - 用户 - 对象存储的使用者，存储桶的拥有者
  - 存储桶 - 作为存放对象的容器
  - 对象 - 用户实际上传的文件

Copyright © 2015 Mirantis, Inc. All rights reserved.

UMCLOUD MIRANTIS

Ceph 对象存储 RGW 原理解析

Copyright © 2015 Mirantis, Inc. All rights reserved.

Ceph 对象存储 RGW 原理解析

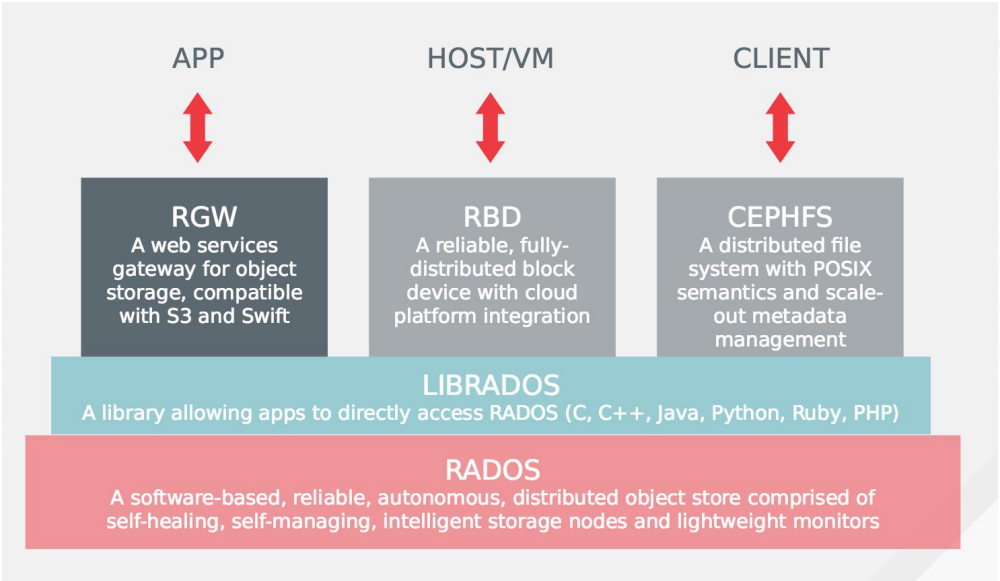
- Ceph 软件架构
- RGW 数据组织
- RGW IO 路径

OPENSTACK DAYS CHINA
Copyright © 2015 Mirantis, Inc. All rights reserved.
OPENSTACK DAYS CHINA

一. Ceph 软件架构

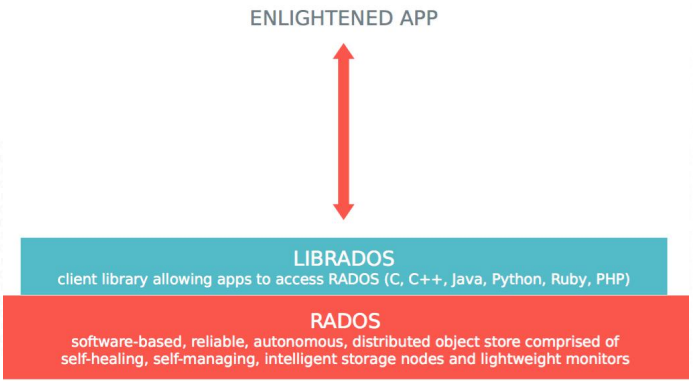
Copyright © 2015 Mirantis, Inc. All rights reserved

Ceph 软件架构



Copyright © 2015 Mirantis, Inc. All rights reserved.

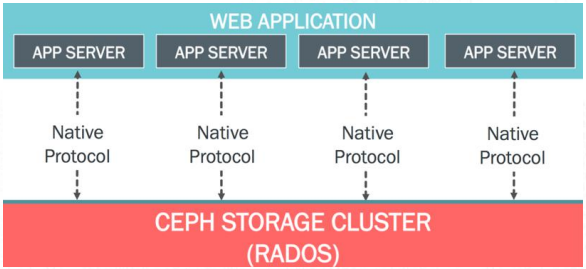
RADOS 客户端编程接口



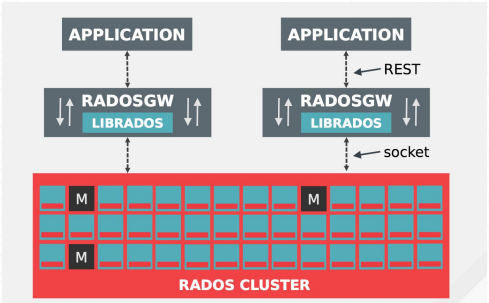
- 命名空间 -> pool
- 存储单元 -> rados-object
- 键值对 -> omap/xattr

OPENSTACK DAYS CHINA
Copyright © 2015 Mirantis, Inc. All rights reserved.
OPENSTACK DAYS CHINA

RGW - RADOS 的 HTTP 协议转发层



直接通过 librados 访问 RADOS 集群



通过 RGW 以 HTTP 请求访问 RADOS 集群

Copyright © 2015 Mirantis, Inc. All rights reserved

RADOS “对象” 和 RGW “对象” 比较

对比项	RADOS 中的对象	RGW 中的对象
简称	rados-object	rgw-object
访问协议	Librados 协议	HTTP 协议
大小	固定大小(为了条带化,通常比较小,默认 4MB)	没有限制(S3下默认5TB)
可变性	可变，支持覆盖写	不可变
是否索引	无	有
命名空间的划分	通过存储池划分	通过存储桶划分
对象级别的ACL	不支持	支持
版本控制	不支持	支持
对象生命周期管理	不支持	不支持

Copyright © 2015 Mirantis, Inc. All rights reserved.

二. RGW 数据组织

Copyright © 2015 Mirantis, Inc. All rights reserved

数据组织的逻辑层级

- 元数据
  - 每个用户创建的存储桶(bucket per user)
  - 存储桶索引 - 每个存储桶中的对象列表(rgw-object per bucket)
- 数据
  - 每个对象拆分的 RADOS 层对象(rados-object per rgw-object)

OPENSTACK DAYS CHINA
Copyright © 2015 Mirantis, Inc. All rights reserved.
OPENSTACK DAYS CHINA

数据的实际存储

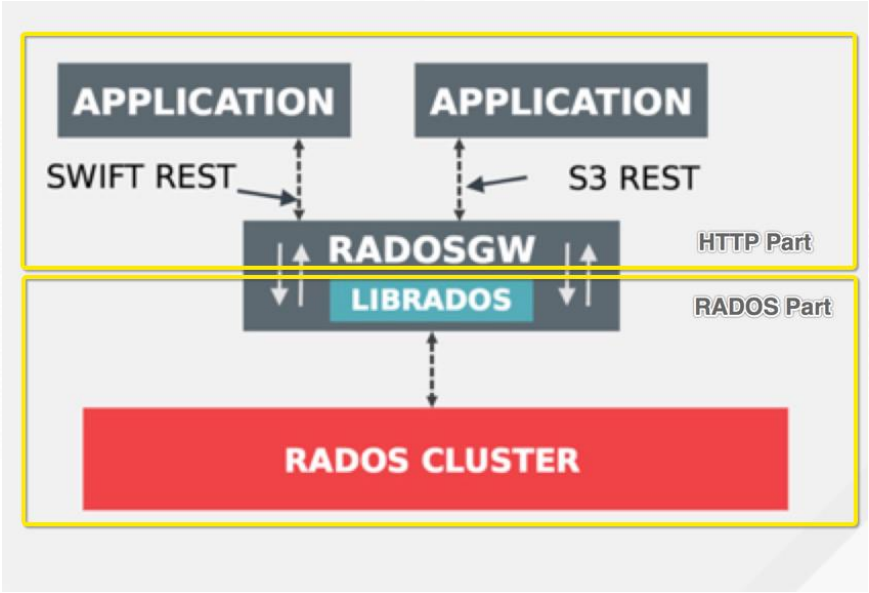
- 数据和元数据均保存在 RADOS 集群的存储池中
- 可扩展性为王，避免引入额外的元数据管理方案
- 性能问题通过存储池的存储策略去改善

Copyright © 2015 Mirantis, Inc. All rights reserved.

3. RGW IO 路径

Copyright © 2015 Mirantis, Inc. All rights reserved

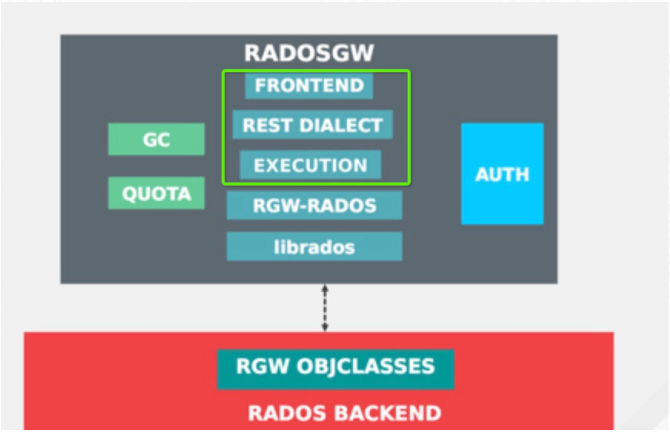
RGW IO 路径



OPENSTACK DAYS CHINA
Copyright © 2015 Mirantis, Inc. All rights reserved.
OPENSTACK DAYS CHINA

RGW IO 路径 -> HTTP 前端

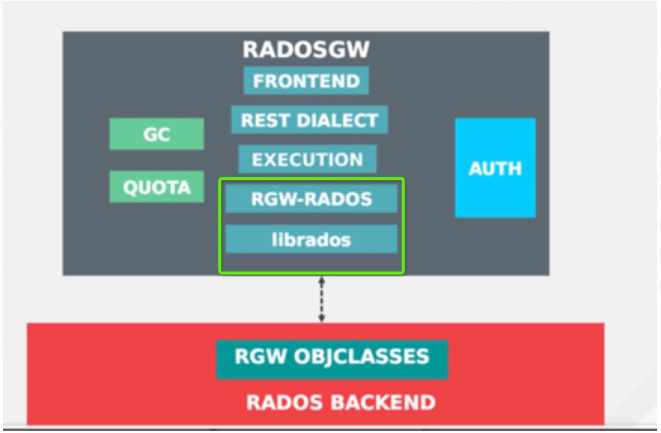
- Civetweb(可嵌入的 C++ 实现的 HTTP 服务端库)
- Loadgen(测试专用，并不处理数据 IO)
- FCGI(作为 Apache 模块，支持 CGI 协议)
- 新的 HTTP 前端



Copyright © 2015 Mirantis, Inc. All rights reserved.

RGW IO 路径 -> 与 RADOS 集群的交互

- 统一的执行层
- 与 RADOS 交互的有两种方式
  - 调用 librados 接口函数
  - 定义 object class(在 RADOS集群端进行计算的机制，避免额外的数据传输)



OPENSTACK DAYS CHINA
Copyright © 2015 Mirantis, Inc. All rights reserved.
OPENSTACK DAYS CHINA

UMCLOUD MIRANTIS

基于 Ceph 的对象存储的架构设计

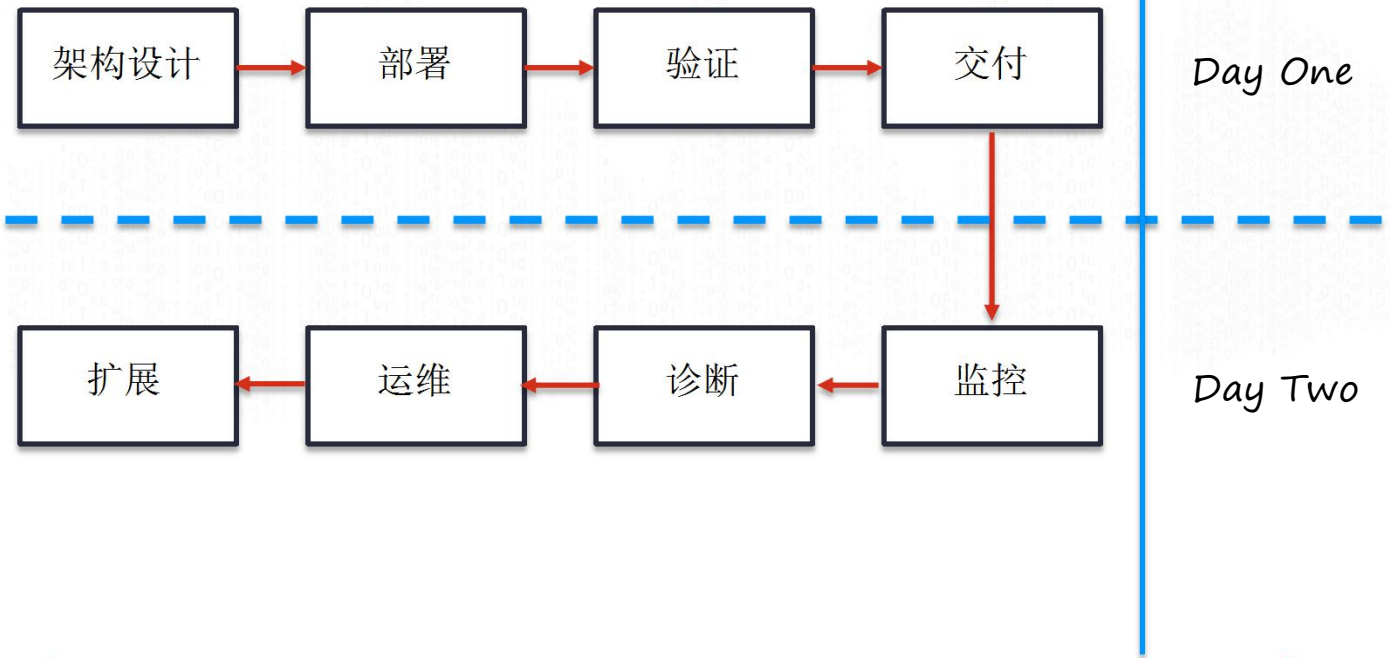
OPENSTACK DAYS CHINA
Copyright © 2015 Mirantis, Inc. All rights reserved.
OPENSTACK DAYS CHINA

基于 Ceph 的对象存储架构设计

- 架构设计的重要性？
- SDX
- 逻辑架构/角色划分
- 大规模部署
- 对象存储的性能优化
- 对象存储系统与 OpenStack 的集成

Copyright © 2015 Mirantis, Inc. All rights reserved.

架构设计的重要性？ 70%



Copyright © 2015 Mirantis, Inc. All rights reserved.

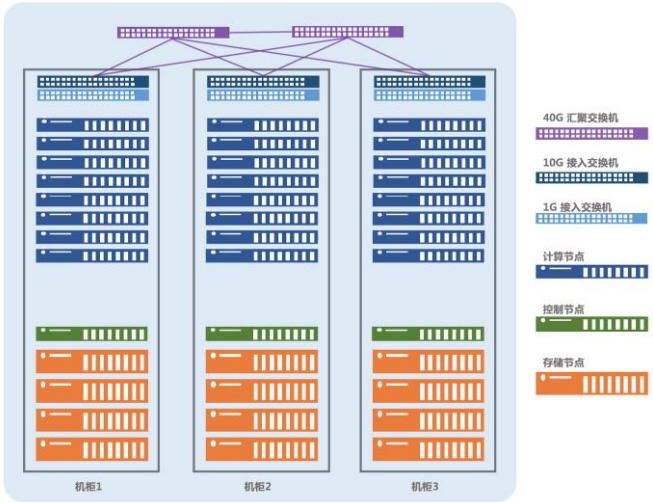
架构设计的主要内容

- 最终设计
  - 设计原则
  - 内容介绍
- 当前服务器规划
  - 服务器配置
  - 交换机配置
  - 服务器柜列分布
  - 网络连接
  - 节点分组
  - 网络配置
  - 系统配置
- 整体架构设计
  - 软件规划
  - 容量规划
  - 性能规划
  - 角色划分
  - 逻辑架构
- 部署方案
  - 工具
  - 部署安装节点
  - 硬件验证
  - 硬件性能基准测试(发现慢盘)
  - Puppet部署
  - 硬盘分区
  - Ceph集群部署
  - 负载均衡部署
  - 部署步骤
  - 网络监测
- Ceph集群配置
  - 故障域设计
- 节点角色分配
- 机架位置
- CRUSH MAP设计
  - 设置crush_tunables
  - 副本数设置
  - 层级设计
  - CRUSH RULESET设计
  - CEPH OSD TREE设计
  - CEPH OSD TREE规划
- PG 数目规划
- Ceph 参数配置
  - Global参数优化
  - OSD参数优化
  - MON参数优化
  - RGW参数优化
  - Pool 配置
- 告警的对接
- 测试验证方案
- 系统集成
  - 运营平台集成
  - 监控告警集成
- 运维方案
  - 常见运维操作
  - 升级操作
- 运维练习
- 其他
  - 扩容方案
  - 数据持久性计算

Copyright © 2015 Mirantis, Inc. All rights reserved.

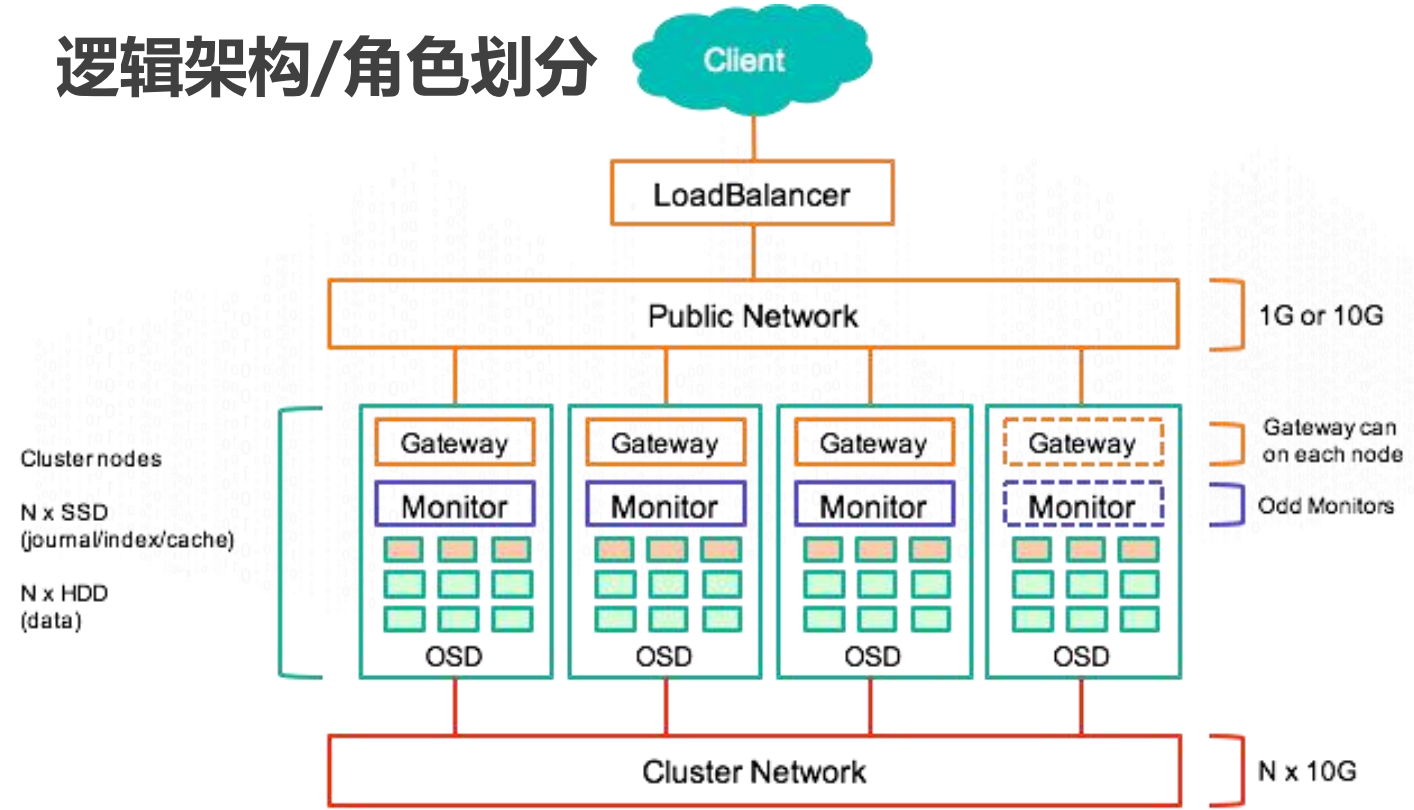
SDX - 如何满足客户的需求

Ceph是真正的SDS(软件定义存储)，通过灵活的配置和堆服务器硬件，可以让我们自定义存储集群的性能、容量、可用性、可靠性等指标，满足当前的需求和未来战略的需求。



Copyright © 2015 Mirantis, Inc. All rights reserved.

逻辑架构/角色划分



Copyright © 2015 Mirantis, Inc. All rights reserved.

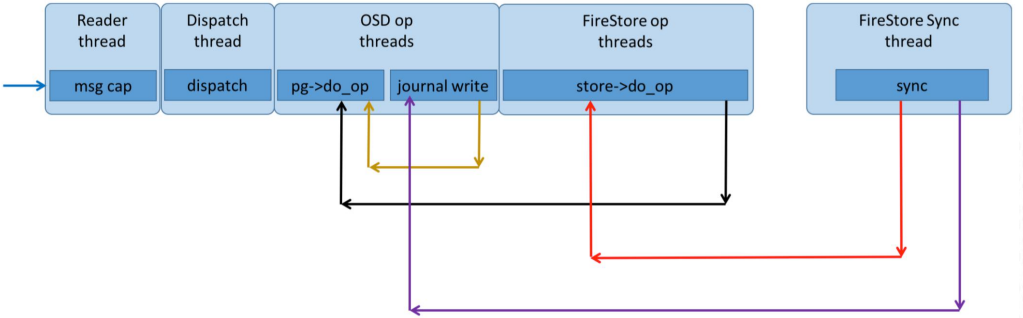
大规模部署

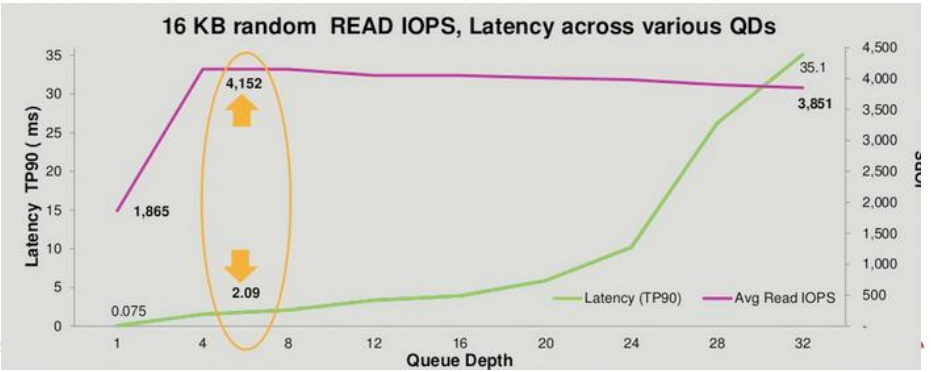
- 限定单个Cluster的规模，通过组合Cluster来扩展存储容量
- 重新设计CRUSH MAP，优化数据可靠性和持久性
- 自动化部署工具，提高部署效率，减少手工部署带来的错误
- 图形化管理平台
- 部署前的硬盘配置检测、硬盘性能基准测试、网络检测

Copyright © 2015 Mirantis, Inc. All rights reserved.

对象存储的性能优化 - 我们的思路

- 流水线/排队论
- 先找到IOPS峰值
  - Queue
  - Op threads
  - Msg cap
- 再优化Latency
  - Msg cap
  - Journal
  - WBThrottle
  - Cache
  - CPU





Ceph RGW 与 OpenStack 集成

- Glance
  - Image, Snapshot
- Cinder
  - Backup
- Sahara
  - Hadoop

Copyright © 2015 Mirantis, Inc. All rights reserved.

Ceph RGW 发展现状

Copyright © 2015 Mirantis, Inc. All rights reserved.

Ceph RGW 功能分类

- 接口功能
  - S3 兼容功能
  - Swift 兼容功能
- 用户不可见的功能

OPENSTACK DAYS CHINA
Copyright © 2015 Mirantis, Inc. All rights reserved.
OPENSTACK DAYS CHINA

接口功能 -> S3 接口

- 存储桶/对象操作
  - 分段上传/下载
  - get-by-range
- 数据管理
  - 对象多版本
  - 对象生命周期管理
    - 对象超时(开发中)
    - 对象归档/恢复(不支持)
- 访问管理
  - 强制访问控制
- 访问协议
  - 静态网站托管(支持)
  - BitTorrent 协议支持(开发中)
- 计费系统集成
  - tagging(不支持)
  - 请求者付费模式(支持)

OPENSTACK DAYS CHINA
Copyright © 2015 Mirantis, Inc. All rights reserved.
OPENSTACK DAYS CHINA

- 除了 CORS 之外，实现了

Copyright © 2015 Mirantis, Inc. All rights reserved.

用户不可见功能

- 动态的存储桶索引分片
- LDAP 认证集成
- 多数据中心数据方案 multisite v2
- 服务端加密(Mirantis 开发中)
- 服务端压缩(Mirantis 开发中)

Copyright © 2015 Mirantis, Inc. All rights reserved.

Ceph RGW 案例分享

Copyright © 2015 Mirantis, Inc. All rights reserved.

国外案例- AT&T

美国第二大电信运营商

2013年签订云平台技术运维服务合同，持续服务超过2年。

业务挑战

- 大规模多数据中心运维管理，
- 需要成熟的CI/CD解决方案
- 高性能需求，高级功能定制开发
- 5PB的分布式存储规模
- 超过10个数据中心的管理

为什么选择Mirantis

- Mirantis拥有大规模集群部署经验
- Mirantis提供分布式存储Ceph
- Mirantis可以派遣驻场工程师提供CI/CD设计实施
- Mirantis OpenStack提供OpenStack性能优化增强
- DPDK、SR-IOV、NUMA 和 vCPU绑定特性支持
- 为客户提供功能定制开发和长期维护

Copyright © 2015 Mirantis, Inc. All rights reserved.

国内案例 – 百联集团

百联集团是中国零售百强第1名，中国企业500强第16名。

业务挑战

- 转型商务电子化战略性项目
- 涵盖IT数据中心新建，云平台建设，全渠道各应用整合，全集团统一技术架构、数据交换平台，线上线下支付系统，百联E商场电子平台，消费大数据平台等
- 云平台初期400台规模
- 分布式存储未来需要存储5000万级别的高清图片和视频，大小为几M~几G

为什么使用OpenStack和Ceph

- 主流开源云平台技术方案，符合规模逐步扩容的长期发展目标
- OpenStack可提供灵活的网络架构，满足百联内部网络架构规划
- 分布式存储Ceph的容量和性能可以线性扩展

为什么选择我们

- Mirantis OpenStack产品将为百联提供强大的架构支撑
- Mirantis Ceph分布式存储满足百联对于海量图片和视频文件存储的需求
- 大数据分析Sahara、应用管理Murano符合百联的业务需求
- UMCloud帮助百联建设一支强有力的Openstack运维、研发技术团队

Copyright © 2015 Mirantis, Inc. All rights reserved.


UMCLOUD