

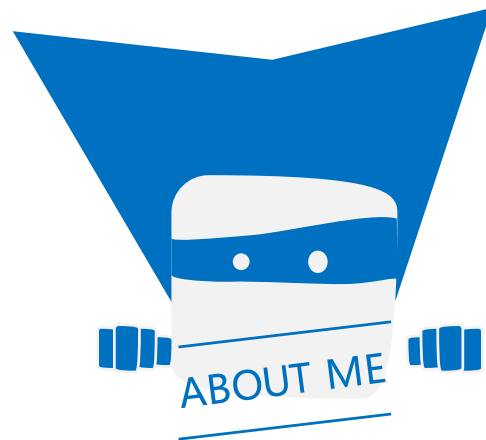
QCon 全球软件开发大会 【北京站】2016

Mangix: 美团云分布式对象存储系统 设计与实现

李凯 2016.4

自我介绍

- 李凯：美团云高级技术专家
 - 美团云计算 存储团队
 - 阿里巴巴 OceanBase
 - 百度 Pyramid
 - 北京邮电大学
 - 研究方向：分布式存储、高可用、数据库
 - Blog：<http://oceanbase.org.cn>



内容提要

- 云计算与存储
- 分布式存储系统设计
- 美团云对象存储系统设计
- 云存储系统生态体系

云计算与存储

- 云计算平台存储产品

- 主机本地存储
- 弹性块存储 (EBS)
- 对象存储 (S3)
- KV与数据库

	读写延迟	可用性	可扩展性	按量付费	多点读写	长度适配
主机本地存储	低	低	无	✗	✗	通用
弹性块存储 (EBS)	低	高	中	✗	✗	通用
对象存储 (S3)	高	高	高	✓	✓	>百KB
KV与数据库	中	高	高	✓	✓	<百KB

云计算与存储

- 云计算平台存储产品应用场景
 - 主机本地存储，弹性块存储（EBS）：主机本地磁盘
 - KV与数据库：结构化数据存储与缓存，数据库事务
 - 对象存储（S3）：
 - 内容存储和分发（图片、视频、网站静态资源）
 - 数据分析的存储
 - 备份、归档和灾难恢复
 - 静态网站托管
 - 主机镜像

云计算与存储

- 对象存储的特点

- Key-Value：用户指定Key
- 一次写入多次读取，少量更新
- 小对象与大对象共存（万亿个，几十KB~几TB）
- HTTP接口
- AWS-S3事实标准
- 账户-桶管理结构
- 权限体系

内容提要

- 云计算与存储
- 分布式存储系统设计
- 美团云对象存储系统设计
- 云存储系统生态体系

分布式存储系统设计

- 数据分布与元数据
 - 运维、迁移恢复、元数据可扩展性
 - 一致性Hash：Swift，Ceph
 - 元数据集中存储：GFS，HDFS，Haystack
- 高可用
 - Swift：计算在Hash Ring上的新位置
 - GFS：更换一组新的副本
 - Spanner：多数派
 - 跨机房副本：实时 vs. 非实时

分布式存储系统设计

- 物理存储形式
 - 直接存储
 - 聚合存储
 - Update/Delete vs. Append Only
 - Replica vs. Erasure Code , EC恢复
- 硬件环境
 - 万兆网卡
 - 低成本：低端CPU、高密度存储

内容提要

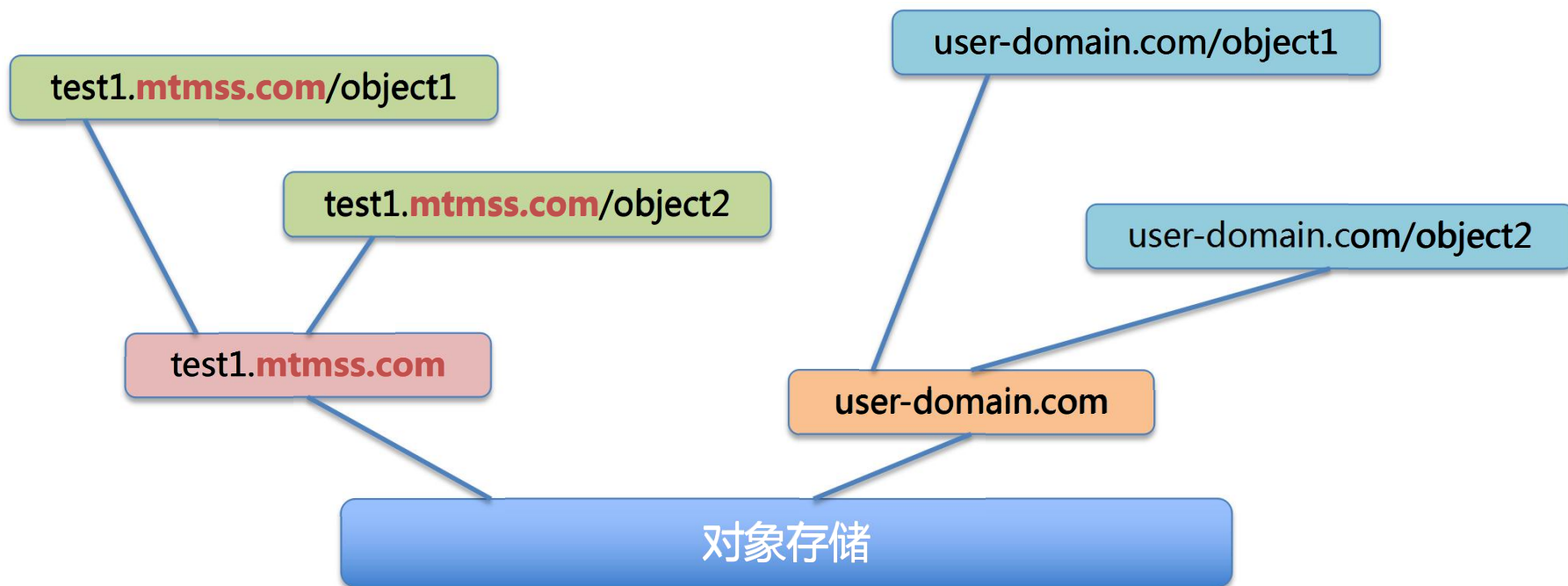
- 云计算与存储
- 分布式存储系统设计
- 美团云对象存储系统设计
- 云存储系统生态体系

美团云对象存储系统设计

- S3存储模型
- 系统架构
- 存储设计
- 元数据设计
- 工程实践经验
- 分布式系统质量控制

美团云对象存储系统设计

- S3存储模型



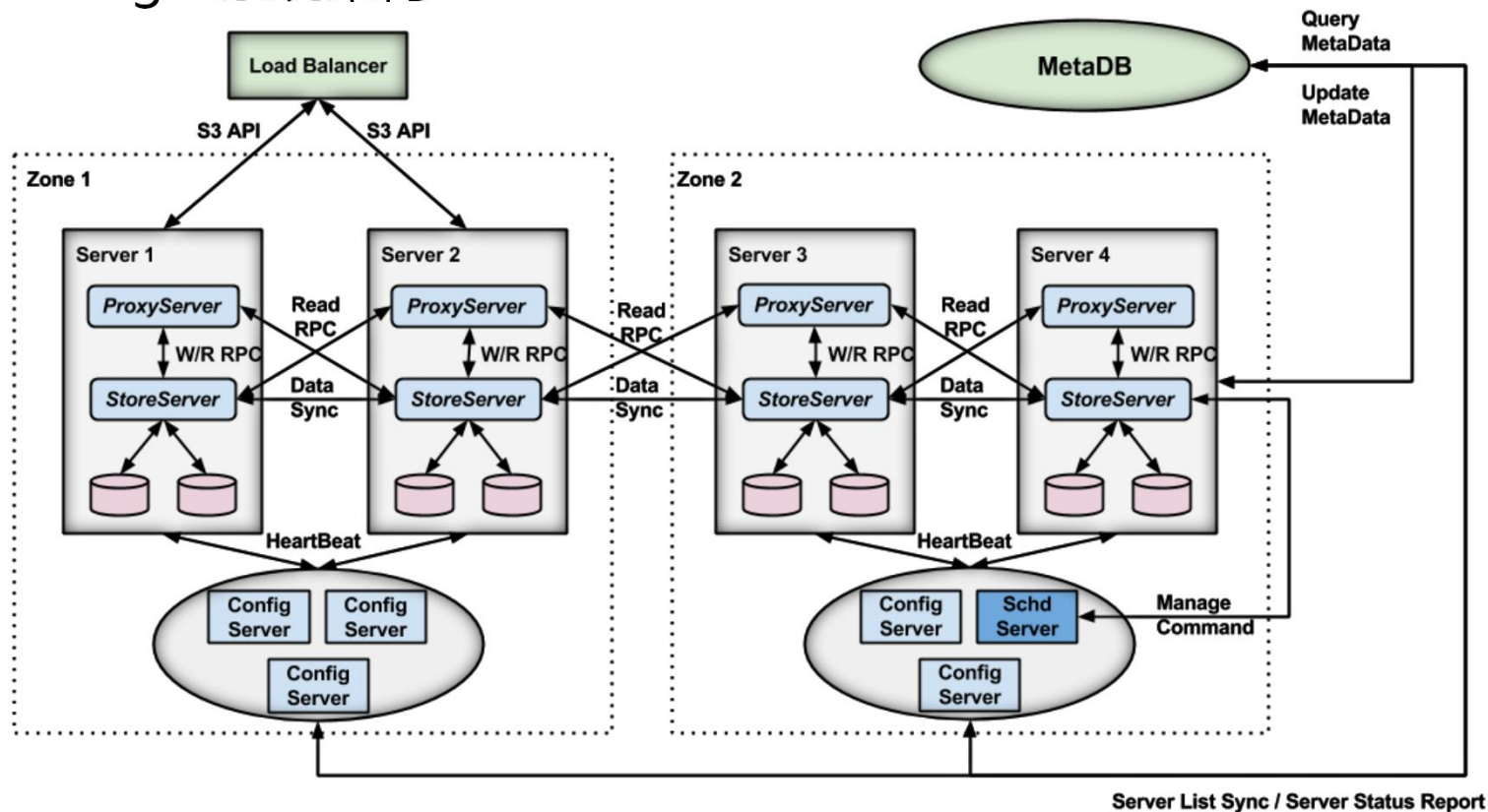
美团云对象存储系统设计

- Mangix 系统模块
 - StoreServer
 - 存储节点：C，并发网络框架
 - ProxyServer
 - Http访问入口：GoLang，AWS-S3/Swift协议兼容
 - SchedServer
 - 调度与监控：Golang，进程监控，ErasureCode/GarbageCollection/数据恢复/负载均衡
 - MetaDB
 - 中心化元数据存储：定制Oceanbase开源版本，跨机房高可用



美团云对象存储系统设计

- Mangix 系统架构



美团云对象存储系统设计

- 存储设计

- Partition

- 负载均衡、迁移复制的最小单位
 - 变长，最大长度限制256MB
 - Primary / Secondary Partition

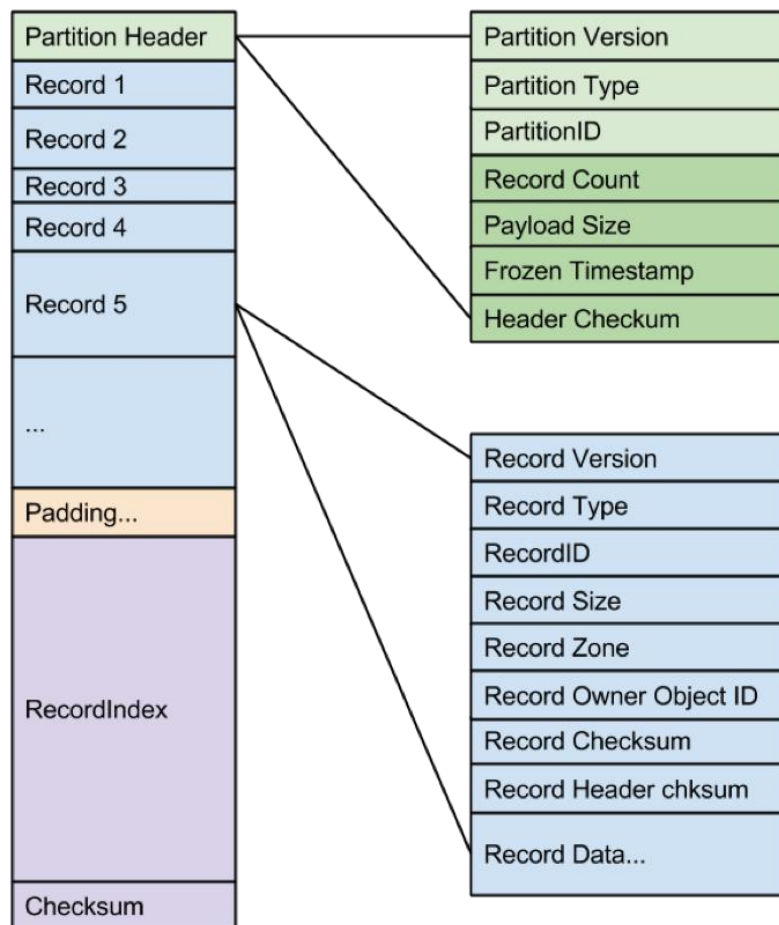
- Record

- 最大长度限制2MB
 - 大文件拆分，流式写入
 - Group Commit

- Record Index

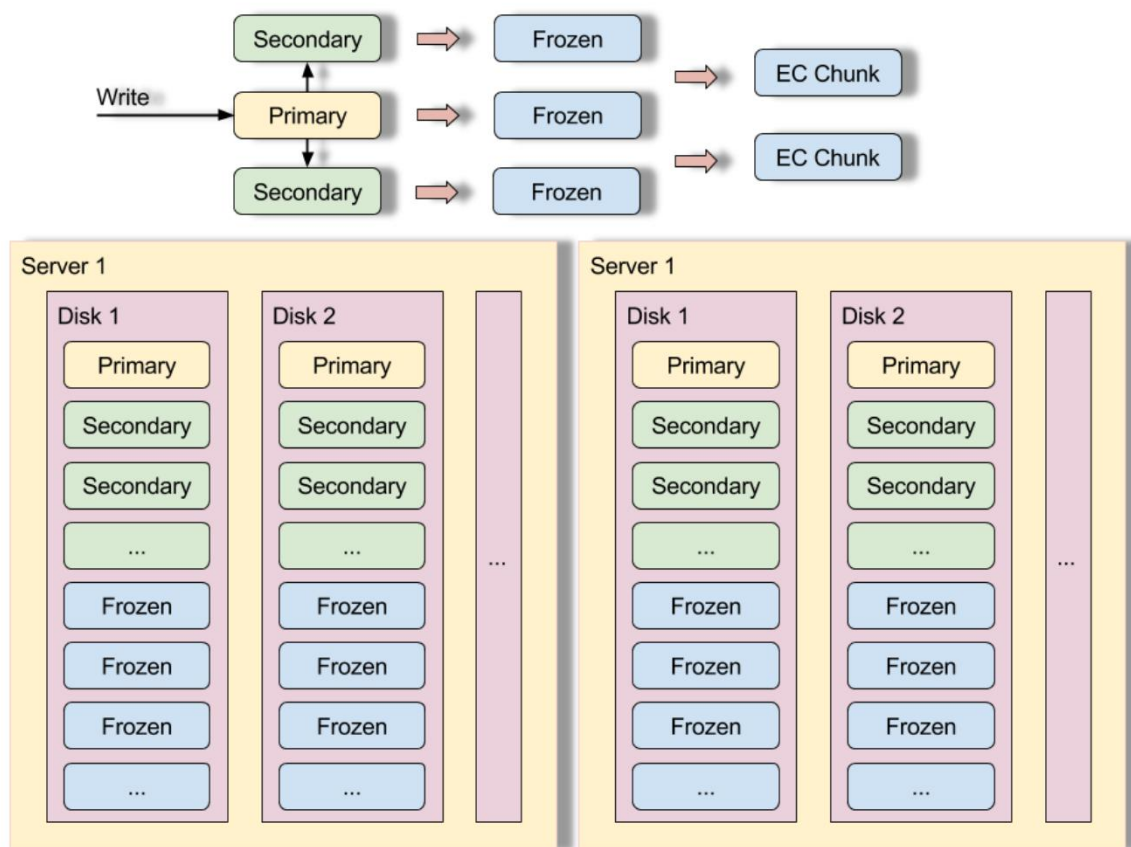
- Record ID → File Position

- 副本多数派应答机制



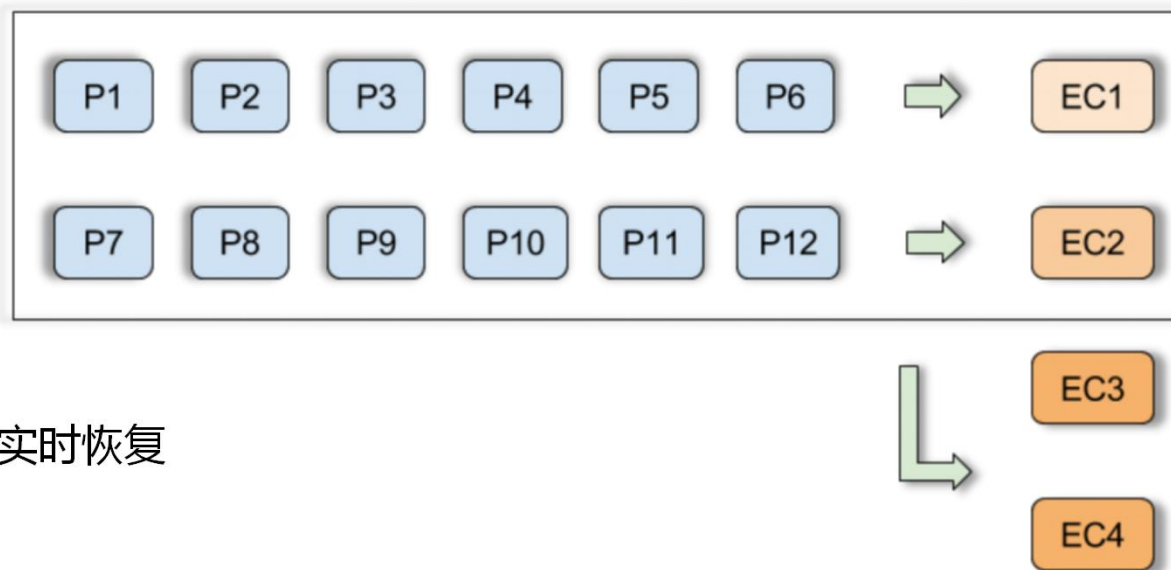
美团云对象存储系统设计

- Partition存储与状态转移



美团云对象存储系统设计

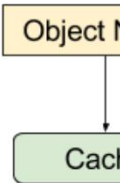
- 抛开恢复时间谈可靠性是要流氓 $P(t, R) = \frac{(\lambda t)^R e^{-\lambda t}}{R!}$
 - LRC : Local Reconstruction Codes



- EC实时恢复

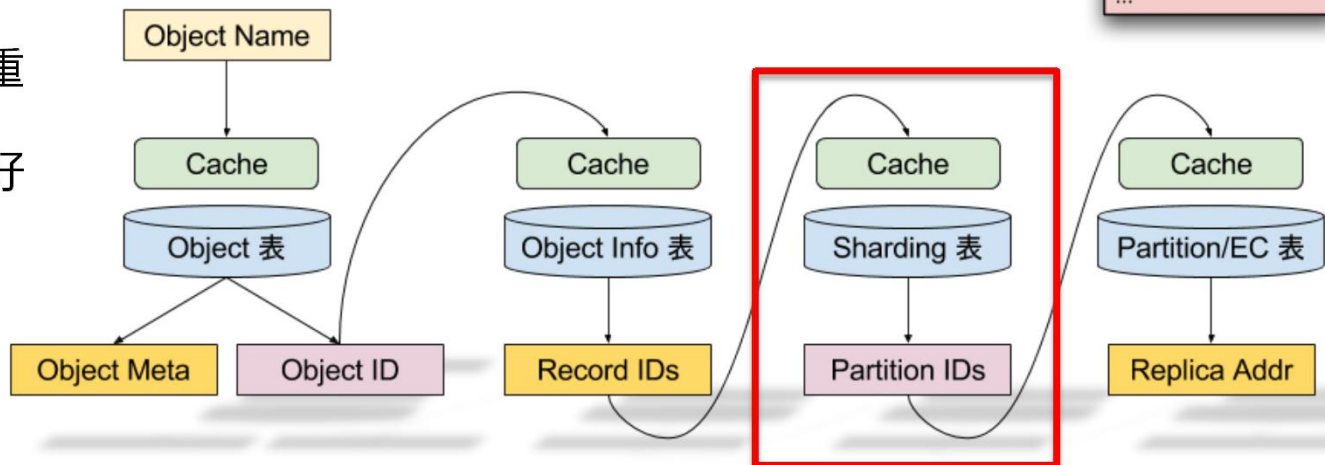
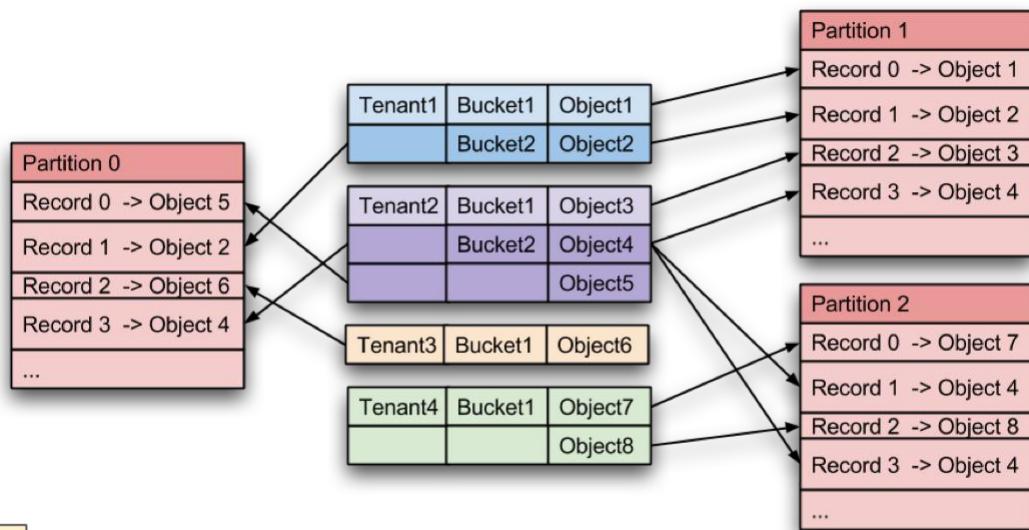
美团云对象存储系统设计

- 元数据设计

- Object Namespace
 - Object Meta数据
 - 对象覆盖写
 - 多版本存储
 - 数据去重
 - 缓存友好
- 
- ```

graph TD
 OM[Object Metadata] --> C[Cache]

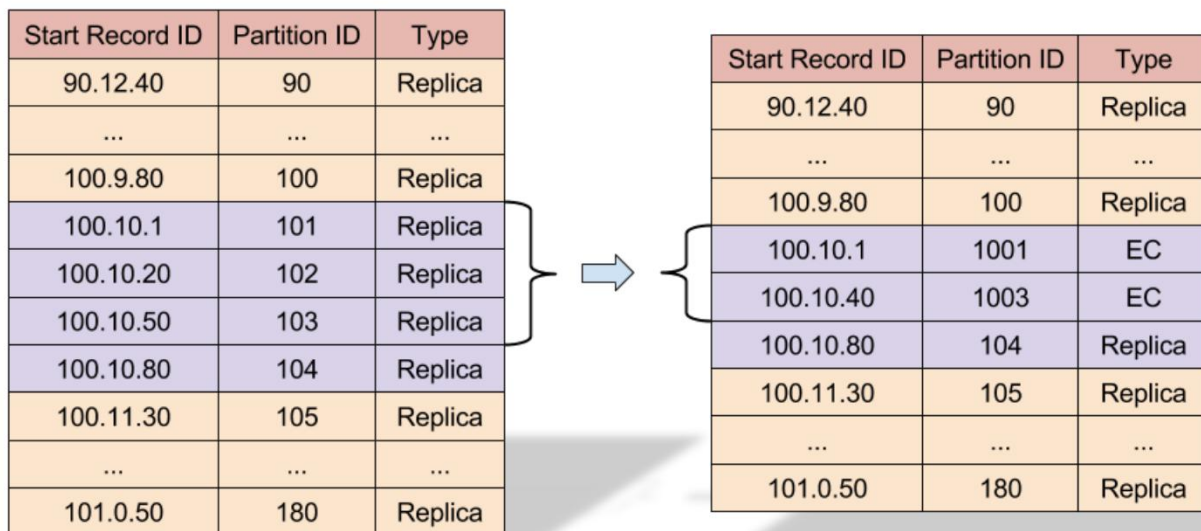
```



# 美团云对象存储系统设计

- Sharding表设计

- Record ID : [ServerIP + DiskID + Timestamp]
- Erasur Code / Garbage Collection : sharding重组

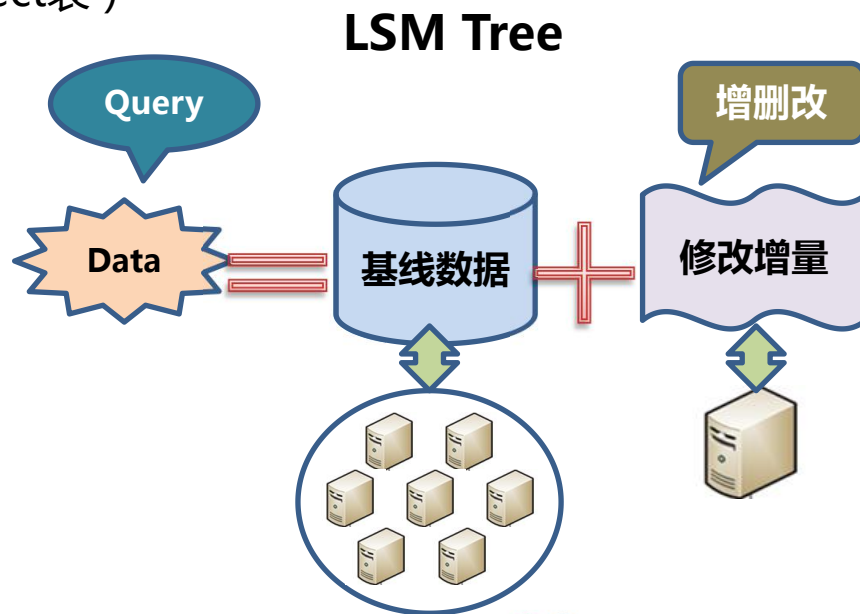


# 美团云对象存储系统设计

- Garbage Collection
  - 标记删除 → 物理删除
    - Bucket
    - Multi-part Index
    - Object
  - 孤儿数据
    - 上传Abort : Object , Multi-part
  - 碎片整理
    - Record碎片
    - Erasure Code重组

# 美团云对象存储系统设计

- MetaDB
  - 万亿级对象
  - 读多写少
  - 主键排序存储 ( Sharding表 / Object表 )
  - Sharding表多行事务
  - Paxos跨机房同步
  - HBase vs. Oceanbase



# 美团云对象存储系统设计

- 工程实践经验
  - 开发语言选择：Golang && C，CGO优化关键算法
  - 万兆网卡优势：并发异步网络框架
  - 快速恢复：StoreServer lazy加载
  - 多级数据完整性校验：RPC、Record、Partition、元数据
  - 进程独占机器，StoreServer / ProxyServer
  - 高可用：
    - ProxyServer/StoreServer黑名单机制
    - SchedServer监控机制

# 美团云对象存储系统设计

- 分布式系统质量控制
  - 0. 编码规范
  - 1. 单元测试
  - 2. 功能测试
    - 直接HTTP请求
    - S3-SDK
  - 3. 模拟异常测试
    - iptables
    - LD\_PRELOAD : 覆盖pread/pwrite , 构造数据错误 , io阻塞 , io过慢 , 磁盘错误
  - 4. 不间断 : 压力读写+数据校验+混合异常

# 内容提要

- 云计算与存储
- 分布式存储系统设计
- 美团云对象存储系统设计
- 云存储系统生态体系



# 云存储系统生态体系

- 云存储系统生态体系





# THANKS!