SACC 2014中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2014
发现架构之美

百度个人云存储架构与实践

周伟 2014/9/18
zhouwei04@baidu.com

内容纲要

- 百度个人云存储简介
- 百度ObjectStore系统
- 集群部署实践

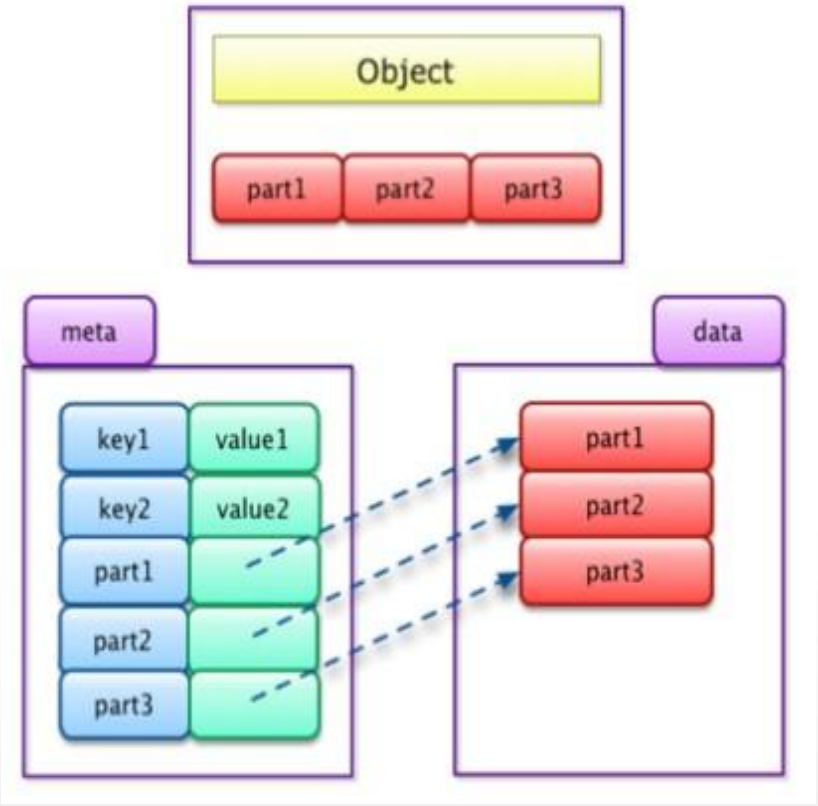
百度个人云存储

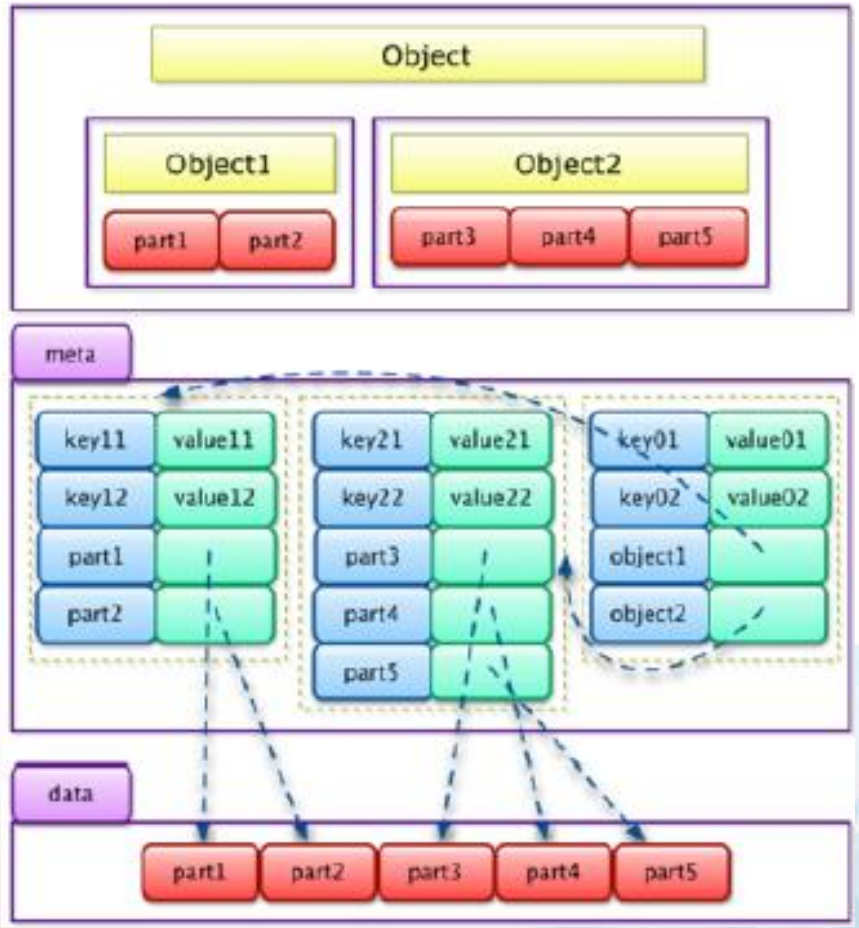
- 免费大空间、海量存储。
- 支持文件、目录及缩略图管理。
- 多终端自动同步。
- 轻松的文件分享功能。
- 基于多种平台提供丰富的SDK。
- 专业的技术支持团队。

百度ObjectStore系统

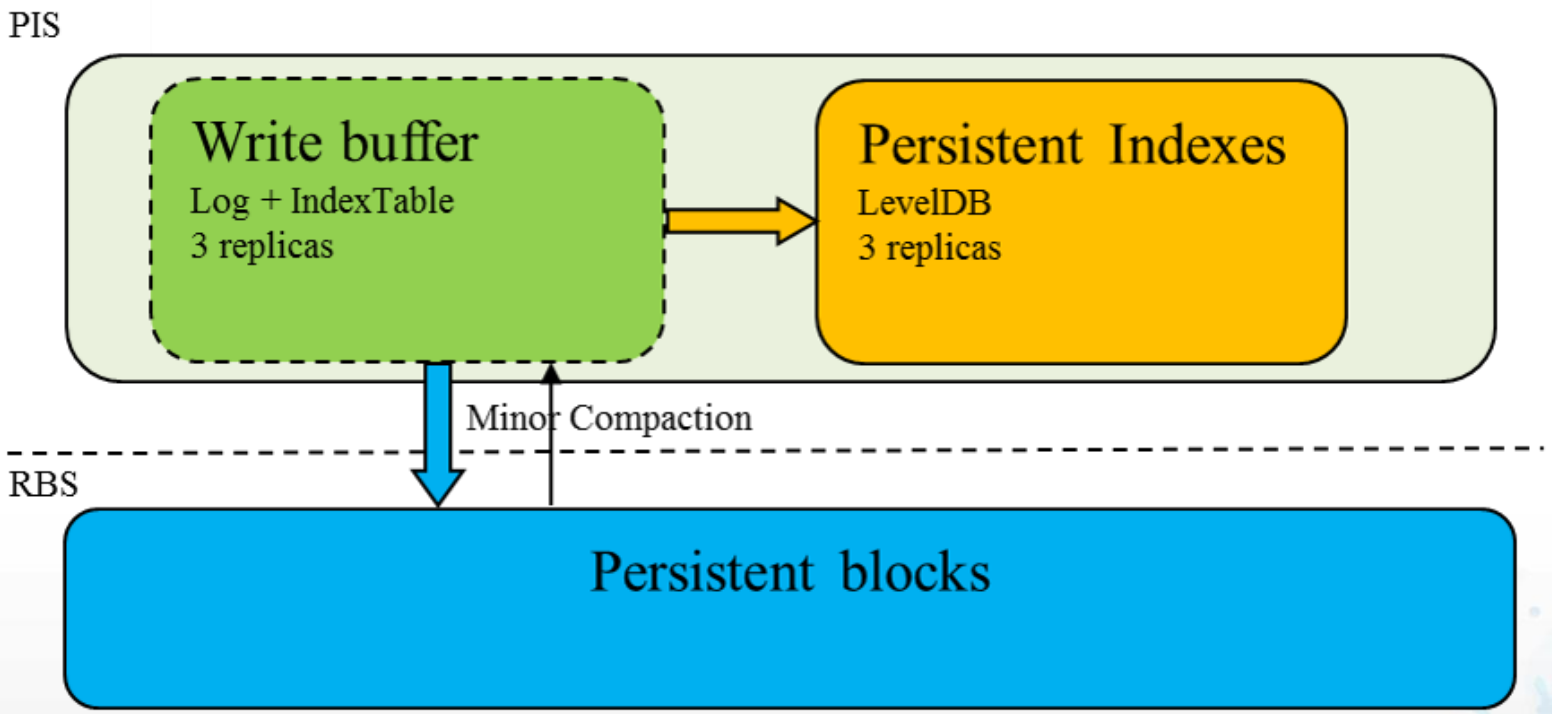
- 通用的object存储系统
- 一致性模型
  - 最终一致性，支持强一致性查询
- 灵活的存储方案
  - EC编码、多副本
- 对外接口
  - Put(UINT128 key, BYTE* value)
  - Get(UINT128 key, BYTE* value)
  - Delete(UINT128 key)

Object数据组织

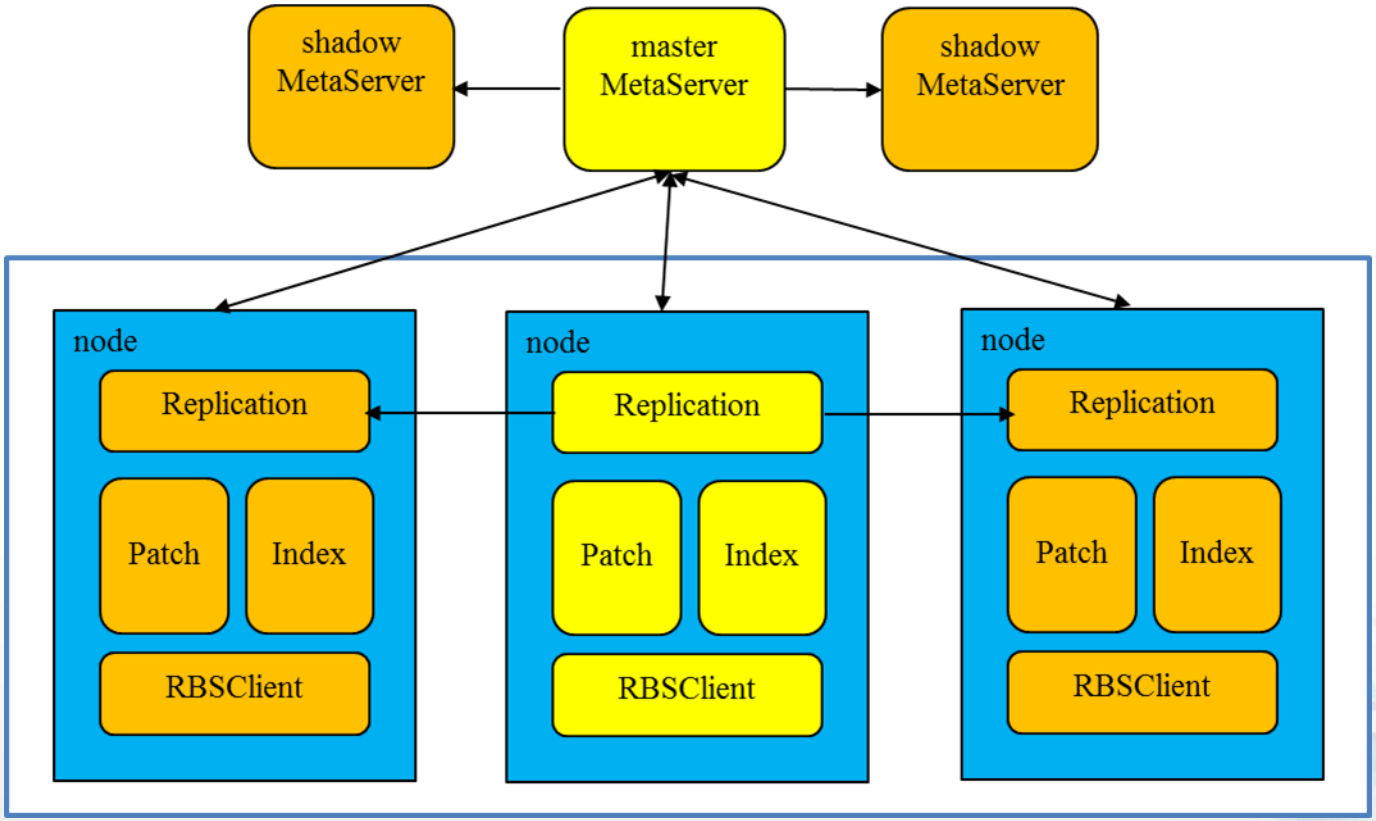

Object
part1
part2
part3
meta
data
key1
value1
key2
value2
part1
part2
part3
part1
part2
part3


Object
Object1
Object2
part1
part2
part3
part4
part5
meta
key11
value11
key12
value12
part1
part2
key21
value21
key22
value22
part3
part4
part5
key01
value01
key02
value02
object1
object2
data
part1
part2
part3
part4
part5

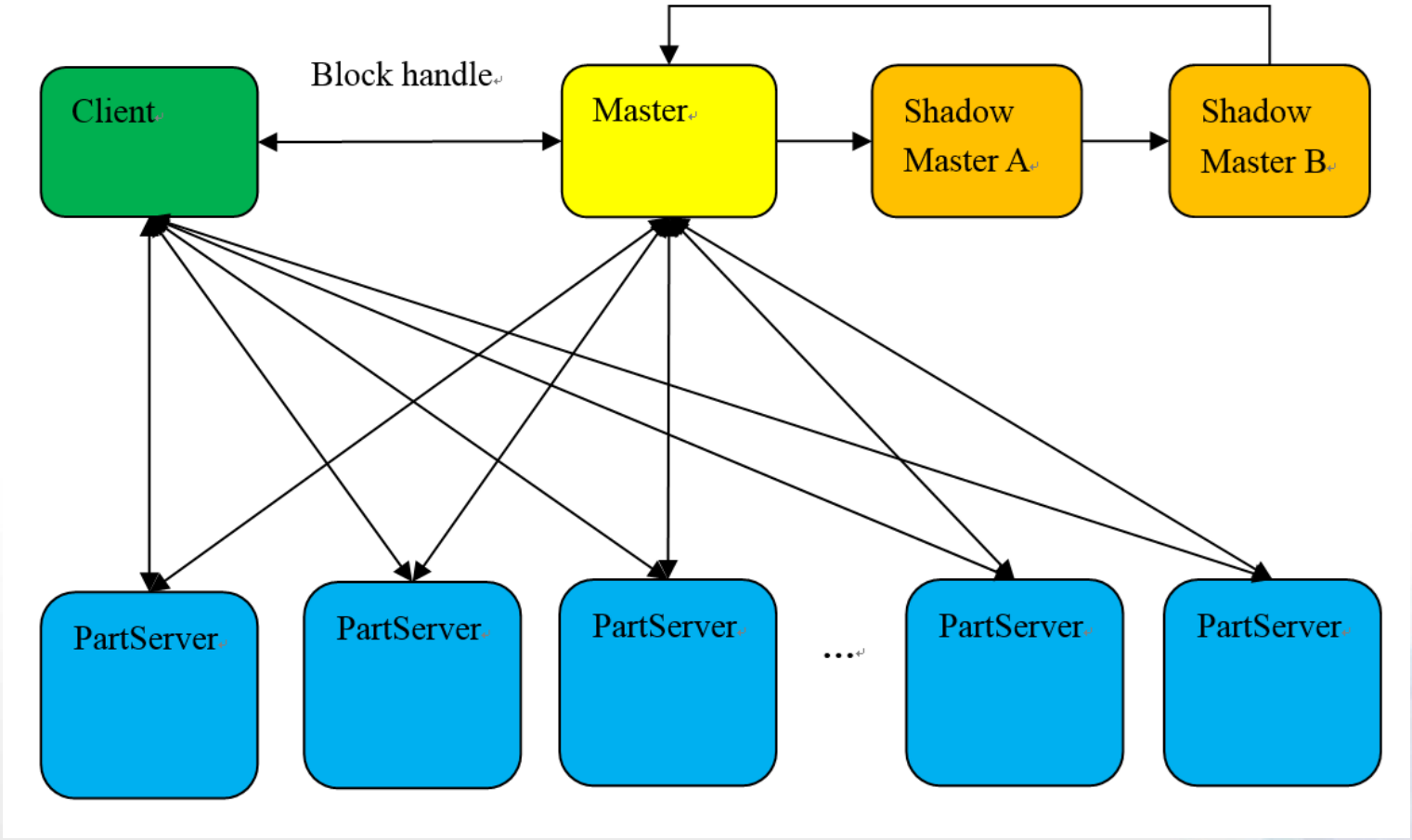
ObjectStore系统架构



PIS架构


shadow
MetaServer
master
MetaServer
shadow
MetaServer
node
Replication
Patch
Index
RBSClient
node
Replication
Patch
Index
RBSClient
node
Replication
Patch
Index
RBSClient

RBS架构


Client
Block handle
Master
Shadow Master A
Shadow Master B
PartServer
PartServer
PartServer
...
PartServer
PartServer

EC编码

- Reed-Solomon Coding : RS(k,m)
  - 基于伽罗华域的运算
  - k个data块,生成m个parity块
  - k+m块中任意丢失至多m个都可修复
- k,m的选择
  - 数据可靠性不能低于3副本模型
  - k/m尽可能小，降低数据冗余度
  - k+m尽可能小，降低元数据存储
  - k越大，网络和磁盘IO越大

系统部署环境

- PIS和RBS混合部署
- Intel X86服务器
  - (12*3T SATA, 64GB Mem)

系统部署环境

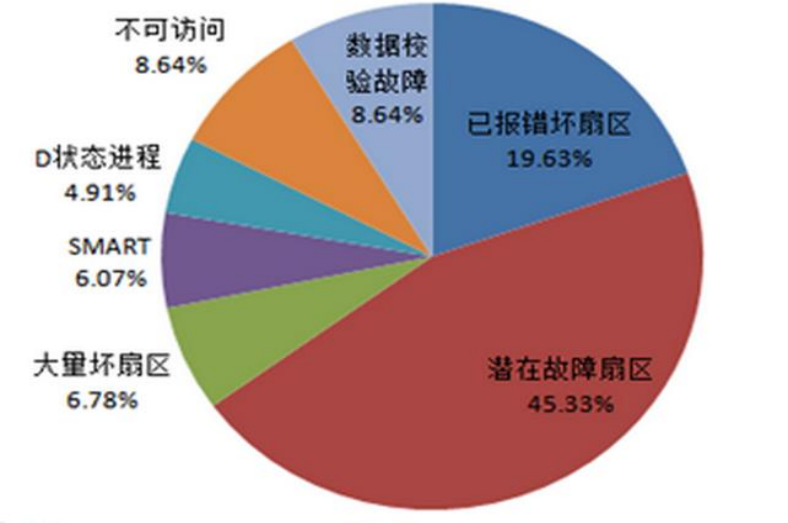
- ARM服务器
  - 2U6(4*3T SATA, 8GB Mem, 4 Core) 10Gbps
  - 存储密度高、低功耗
  - 存储密度提升**75%**，TCO降低**25%**

系统部署环境

- Intel Avaton服务器
  - 1U(18 * 4T SATA, 自主整机柜架构)
  - 存储密度高、低功耗、低成本
  - 存储密度提升**50%**，TCO降低**40%**

RBS实践优化

- 磁盘预警和数据预修复
  - 周期性采集硬盘运行时S.M.A.R.T数据
  - 对S.M.A.R.T数据预测磁盘故障
  - 根据磁盘故障预警，提前迁移part数据，降低EC降级读和数据修复代价



主要故障召回率 *80.84%*

IDC集群部署方案

- 镜像集群
  - 双集群镜像冗余存储，提高数据可靠性
- 主副集群
  - 主集群全量数据存储，副集群缓存最新写入，数据过期失效
  - 降低存储成本，保证数据可靠性，快速验证新集群
- 混合集群
  - object元数据和object数据合并存储
  - object元数据和object数据混部

跨地域多IDC

- CDN
- 外部流量调度
- 内部流量调度
- 协议栈优化
- 运营商友好
- 就近存取

百度云网址：yun.baidu.com

Q&A

THANKS

SequeMedia 盛拓传媒 | IT168 www.it168.com | ChinaUnix | ITPUB