阿里云
aliyun.com

开放存储服务架构设计

吴锦波

2012-07-03

ADC
Alibaba Developer Conference

Agenda

- 什么是开放存储服务
- 开放存储整体设计
- 取舍和教训


海量的
对象数目
任意大小
数据对象
公共服务
高可用性
高可靠性
什么是开放
存储服务

设计原则

- 硬件故障透明
- 数据的多份拷贝分布在不同机架/机房
- 易扩展
  - 容量扩容
  - 自动应对爆发式访问


阿里云
aliyun.com
云服务引擎
ACE
开放存储服务
OSS
开放结构化数据服务
OTS
开放数据处理处理服务
ODPS
弹性计算服务
ECS
关系型数据库服务
RDS
集群部署
Deployment
集群监控
Monitoring
分布式文件系统
Distributed File System
任务调度
Job Scheduling
远程过程调用
Remote Procedure Call
安全管理
Security Management
分布协同服务
Distributed Coordination Service
资源管理
Resource Management
Linux
数据中心
Data Center

Agenda

- 什么是开放存储服务
- 开放存储整体设计
- 取舍和教训

开放存储服务架构

```
http://<bucket>.oss.aliyuncs.com/pathname/to/object
```

负载均衡

协议处理、访问控制

分区、索引

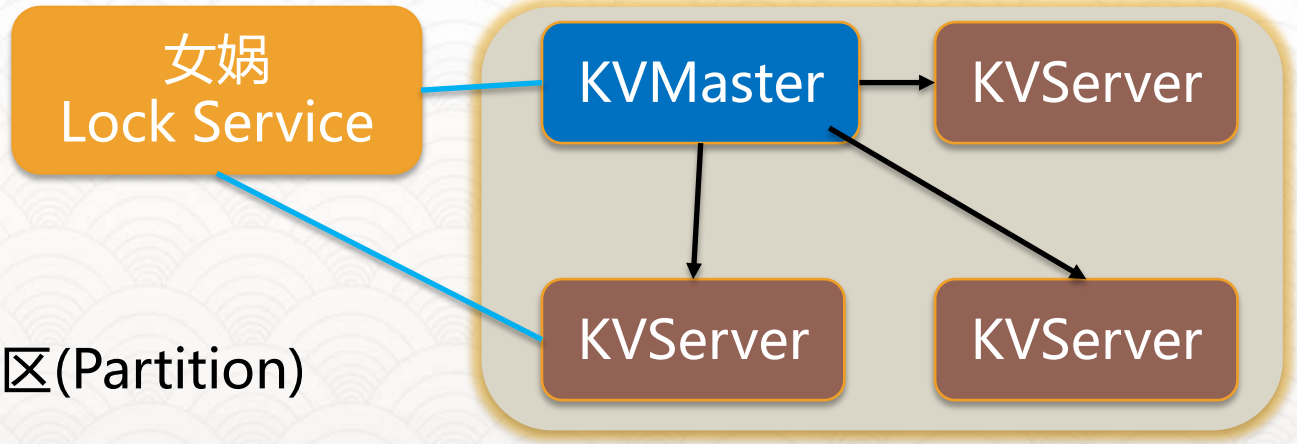
持久、冗余、容错

协议接入层

Web Server + Protocol Module

- ✓ 无状态接入层Server
- ✓ 协议解析
- ✓ 授权/认证
- ✓ 请求路由

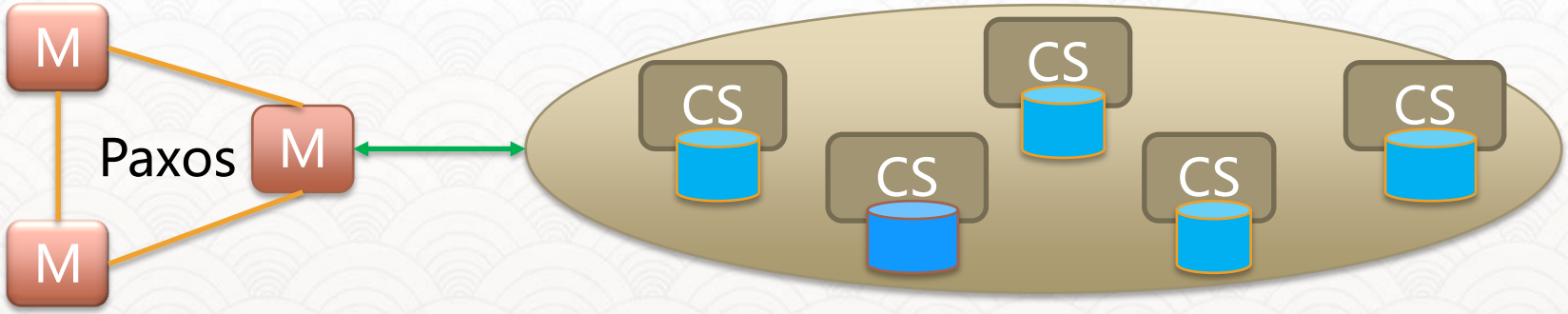
数据访问层KV引擎--管理分区/索引



- KVMaster
  - 管理全局分区(Partition)元信息
  - 数据分区控制
  - 调度分区到KVServer
- KVServer
  - 负责若干个分区
  - 通过Pangu存取数据
  - 建立索引

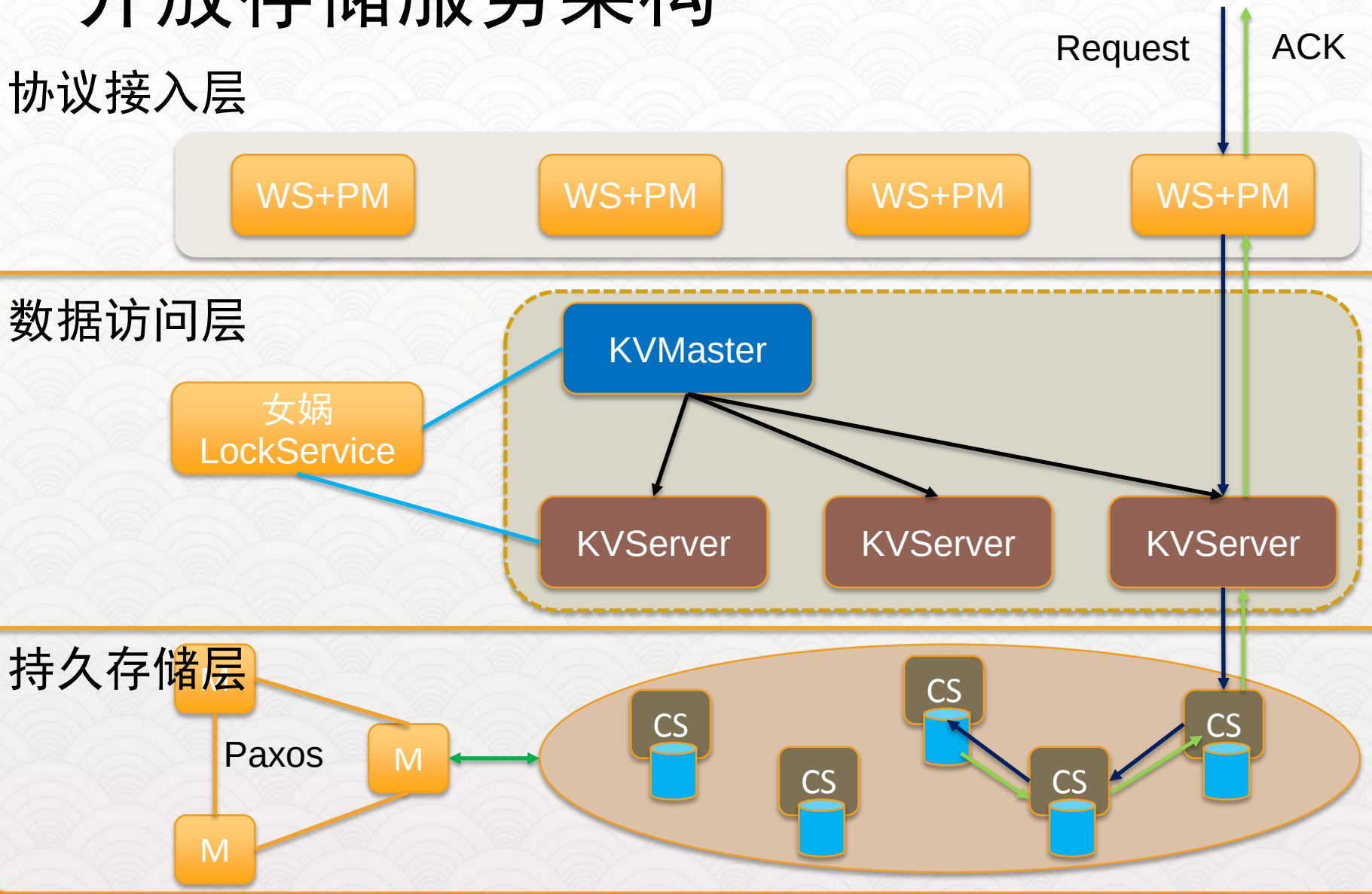
- 女娲
  - 命名服务
  - 分布式锁服务

持久存储层Pangu--大规模分布式文件系统



- Master-Slave结构
  - Master管理元数据，Chunk Server负责Data读写
- 基于Paxos的多Master，故障恢复小于一分钟
- 文件分块(Chunk)，每块存三份，分布在不同机架
- End2End的checksum
- 磁盘、机器、机架及checksum fail时数据自动复制

开放存储服务架构


Request
ACK
协议接入层
WS+PM
WS+PM
WS+PM
WS+PM
数据访问层
KVMaster
女娲
LockService
KVServer
KVServer
KVServer
持久存储层
M
Paxos
M
M
CS
CS
CS
CS
CS

Agenda

什么是开放存储服务

开放存储整体设计

- 协议接入层
- 数据访问层
- 持久存储层

取舍和教训

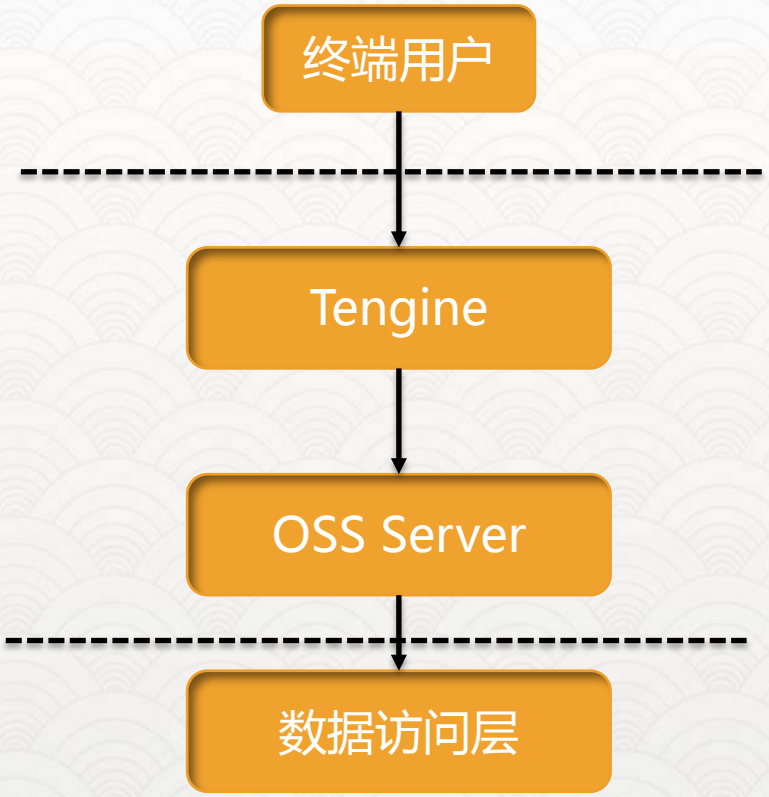
协议接入层--路由与协议处理

- RESTful格式协议处理
  - HTTP协议 - PUT、GET、HEAD和DELETE四类操作
  - 用户请求抽象成数据访问层操作
  - 业务管理逻辑
  - 资源消耗的计量
  - 多种防攻击策略

协议接入层--关键词

- **Service**
  - OSS提供给用户的虚拟存储空间
  - 在这个虚拟空间中，每个用户可拥有一个到多个Bucket
- **Bucket**
  - Bucket是OSS的命名空间
  - Bucket Name在整个OSS中具有全局唯一性
- **Object**
  - 在OSS中，每个文件都是一个Object
- **AccessKeyID、AccessKeySecret**
  - 安全标识，为访问OSS做签名验证

协议接入层架构

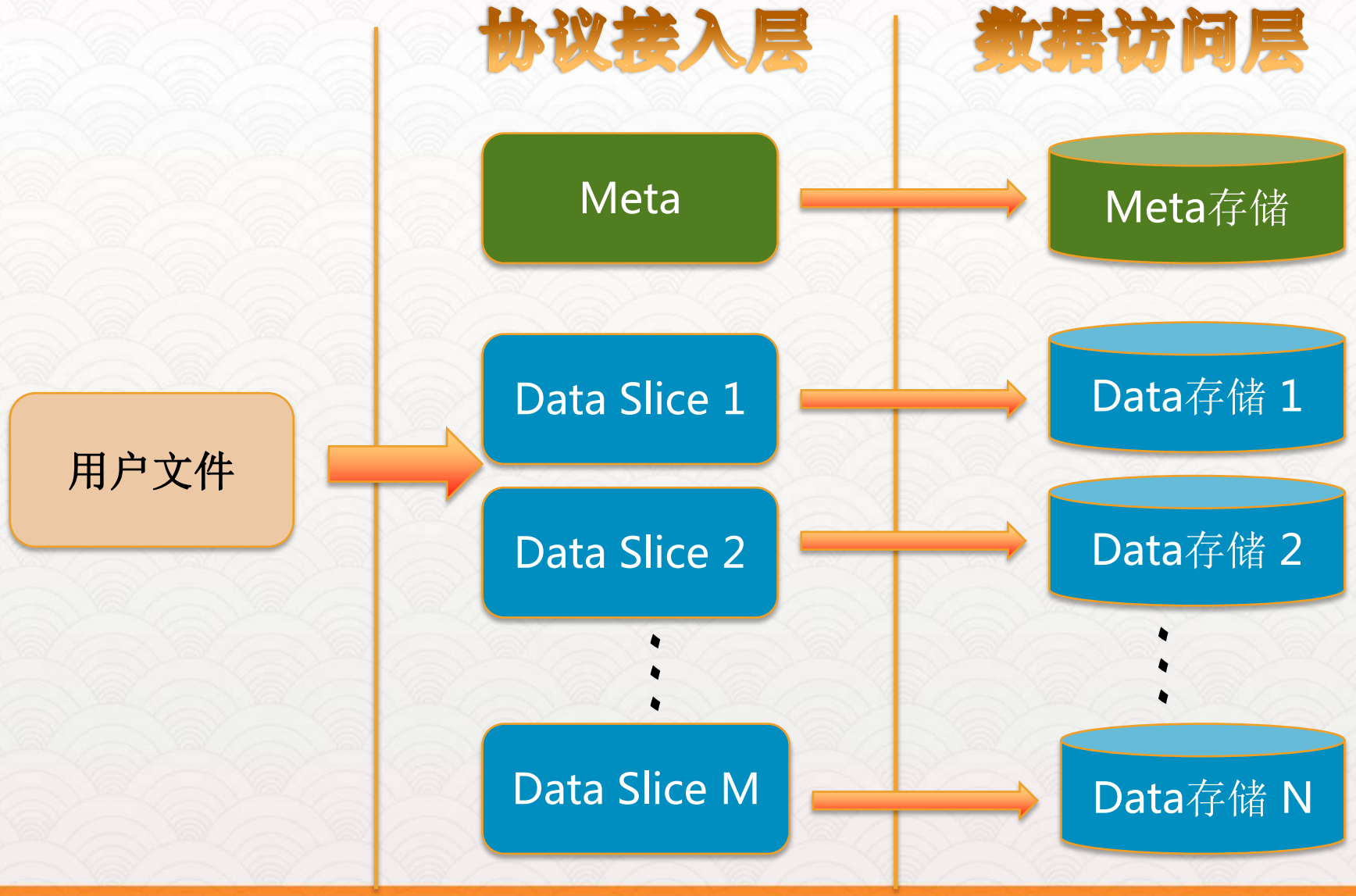


浏览器、SDK等

支持高并发，
管道式收发数据，
TMD防攻击模块

OSS业务逻辑
资源消耗计量
维护和数据访问层映射
逻辑

协议接入层—将用户文件映射成数据访问层存储



Agenda

什么是开放存储服务

开放存储整体设计

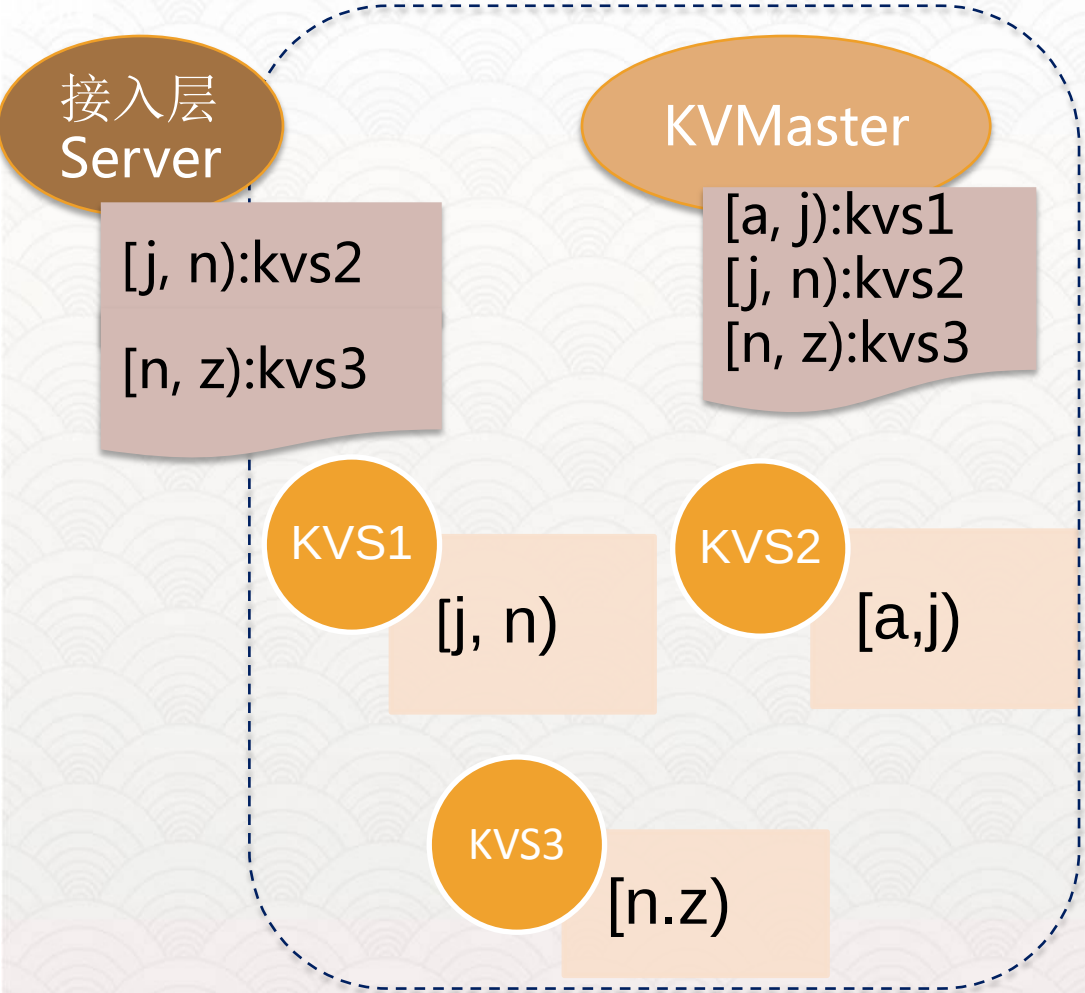
- 协议接入层
- 数据访问层
- 持久存储层

取舍和教训

数据访问层--海量对象索引

- 海量、分布式的Key-Value存储
  - 能扩展到成百上千台服务器
  - 对象能被快速查找、遍历及修改
  - 负载动态平衡

数据访问层--分区元信息

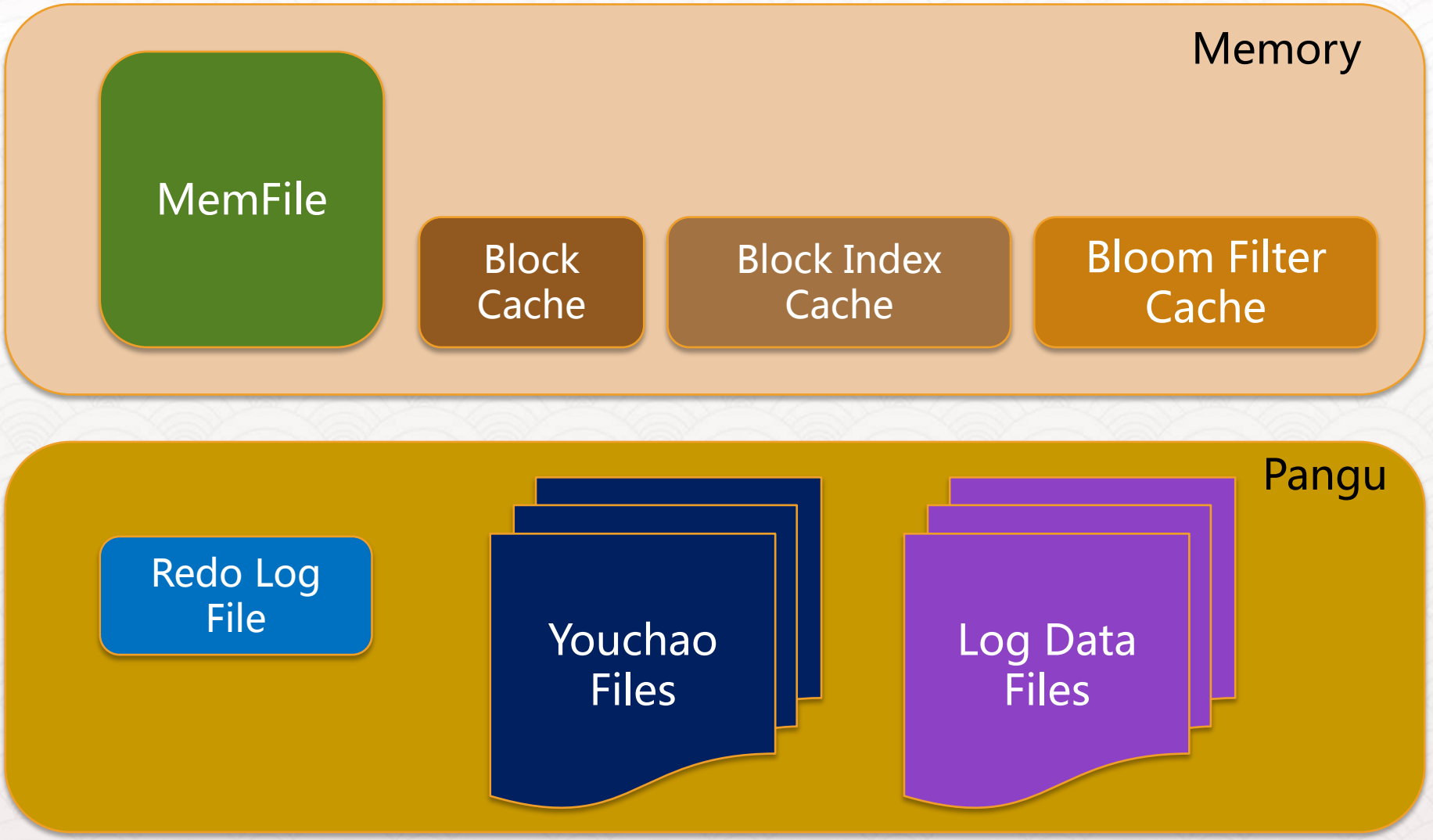


- 系统初始按字符串序预分成若干份
- 根据访问信息分裂/合并某些分区
- 一个分区只能在一个KVServer上被服务
- KVMaster维护分区到KVServer的映射关系
- Client缓存分区与KVServer映射关系

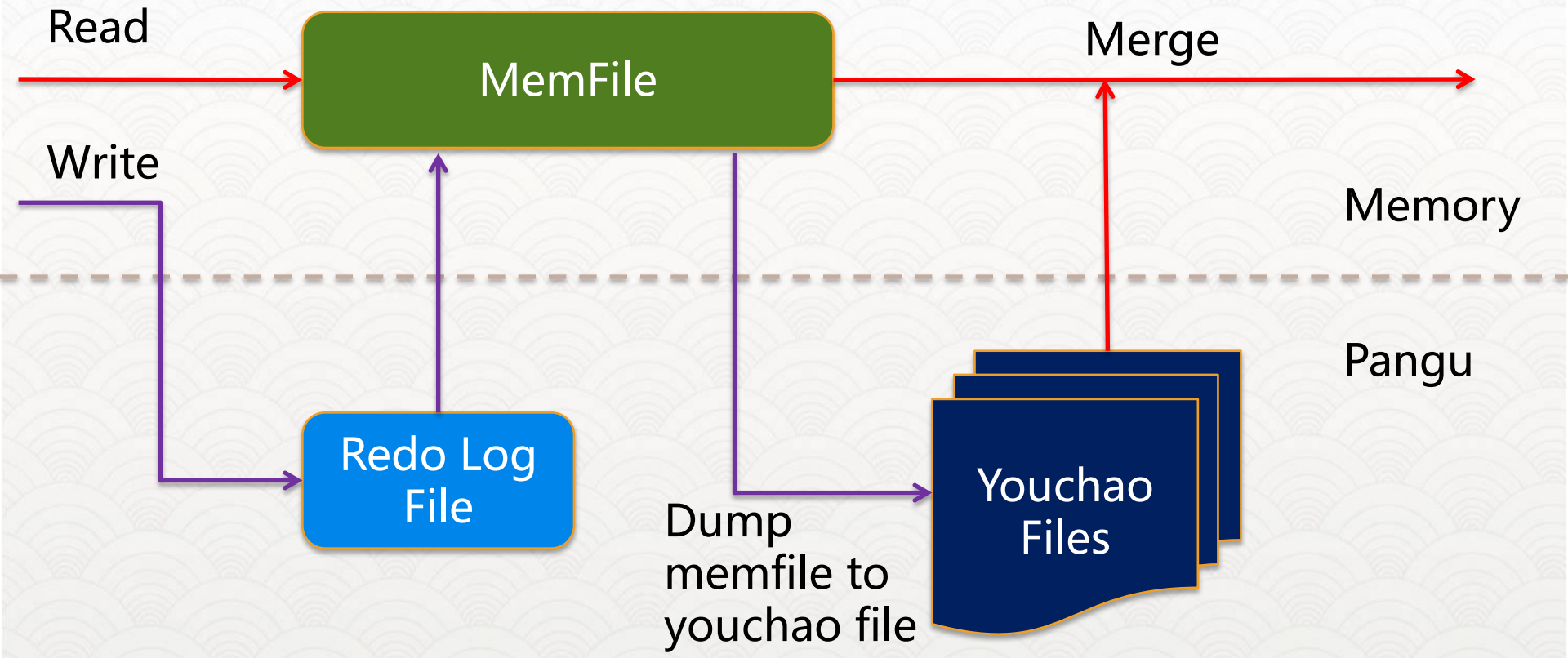
数据访问层--关键词

- **Cell**
  - Key-Value 对
- **Block**
  - 有序Cell集合
  - 读写Pangu层的最小单元
- **MemFile**
  - 内存中根据key排序的cell集合
- **YouchaoFile**
  - 存放在Pangu中的cell集合
  - 由Block/BlockIndex/BloomFilter组成
- **Partition**
  - 一个MemFile/若干YouchaoFile组成
  - 服务某一个范围的cell集合

数据访问层—Append/Dump/Merge



数据访问层--读写过程


Read
MemFile
Merge
Write
Memory
Pangu
Redo Log File
Dump memfile to youchao file
Youchao Files

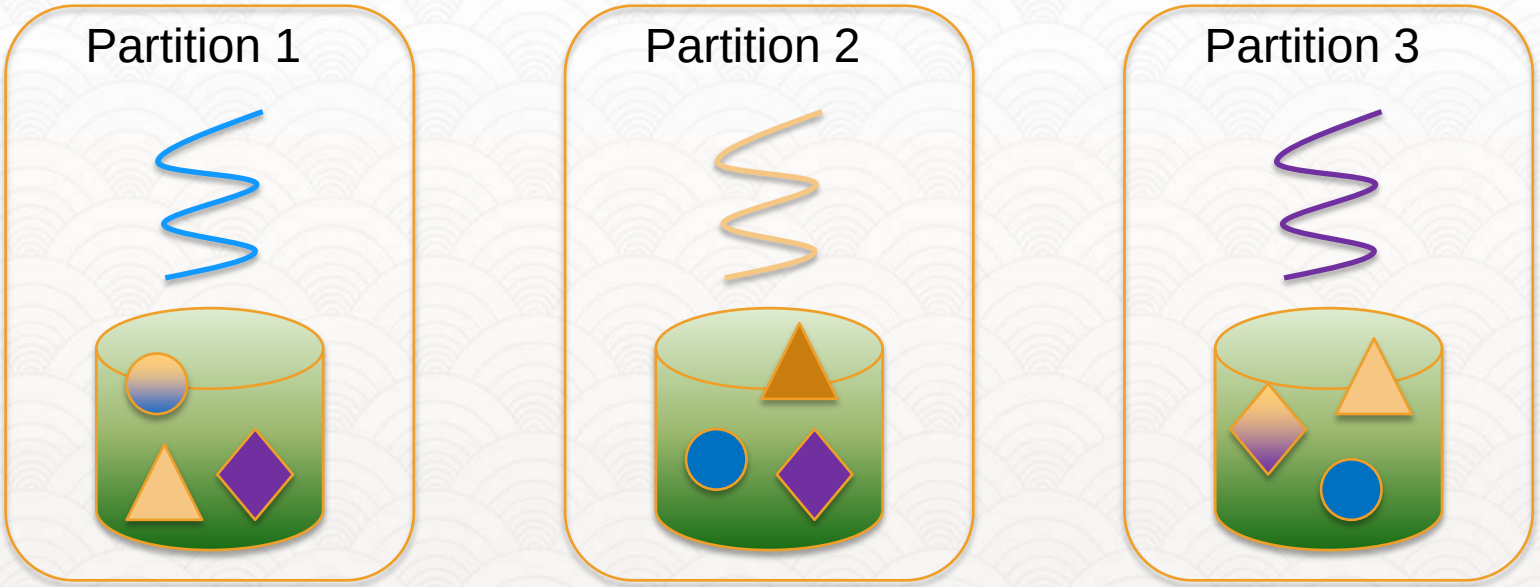
Agenda

- 什么是开放存储服务
- 开放存储整体设计
  - 协议接入层
  - 数据访问层
  - 持久存储层
- 取舍和教训

持久存储层Pangu--基本概念

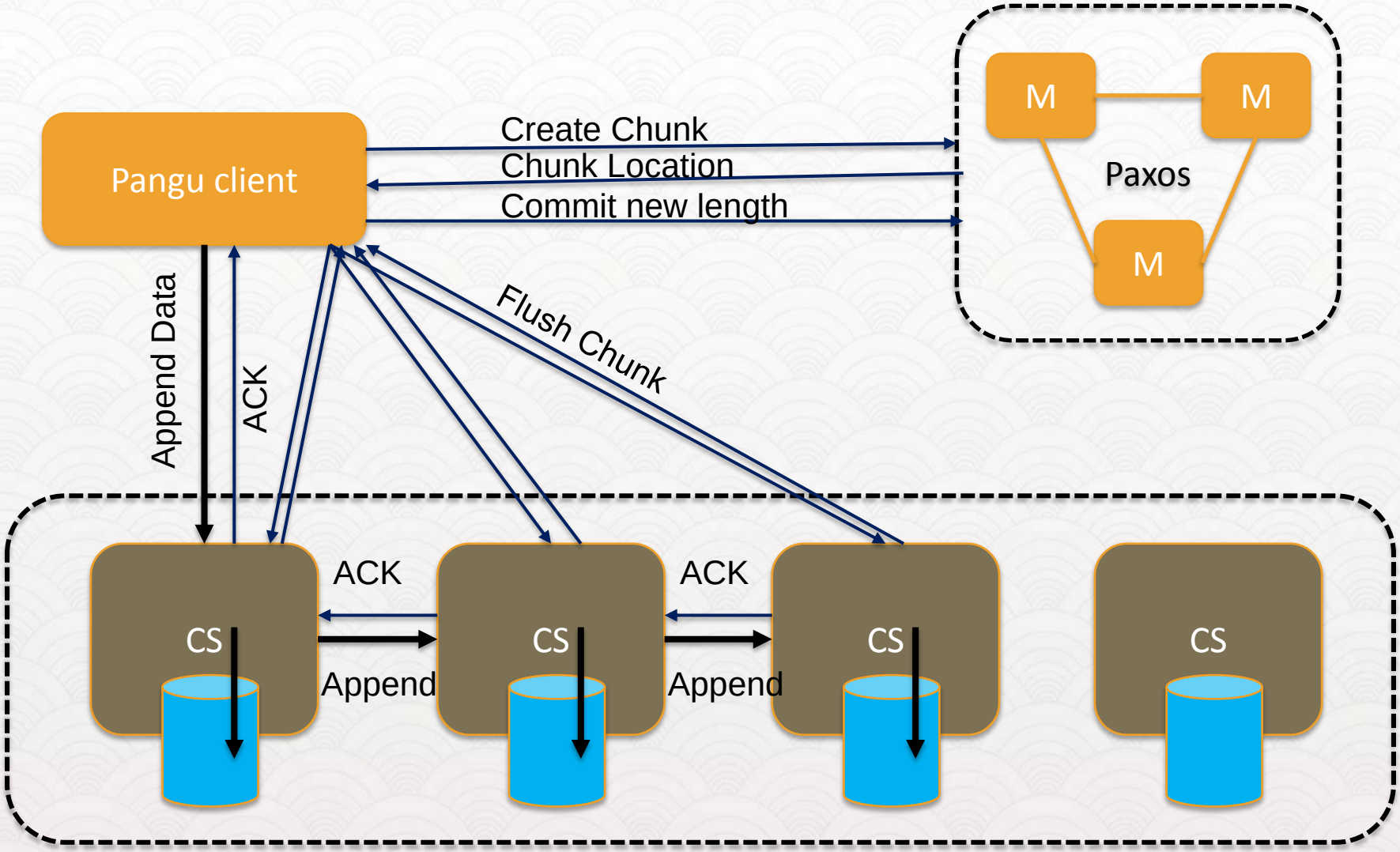
- Append only的分布式文件系统
  - 类似于文件系统的结构
  - 支持创建/打开/追加/关闭/删除/重命名等操作
- 偏重于存储大文件
  - 文件内容布局
    - App-Part
    - LocalFile
- Normal/Log两种文件类型

Pangu—数据聚簇模式

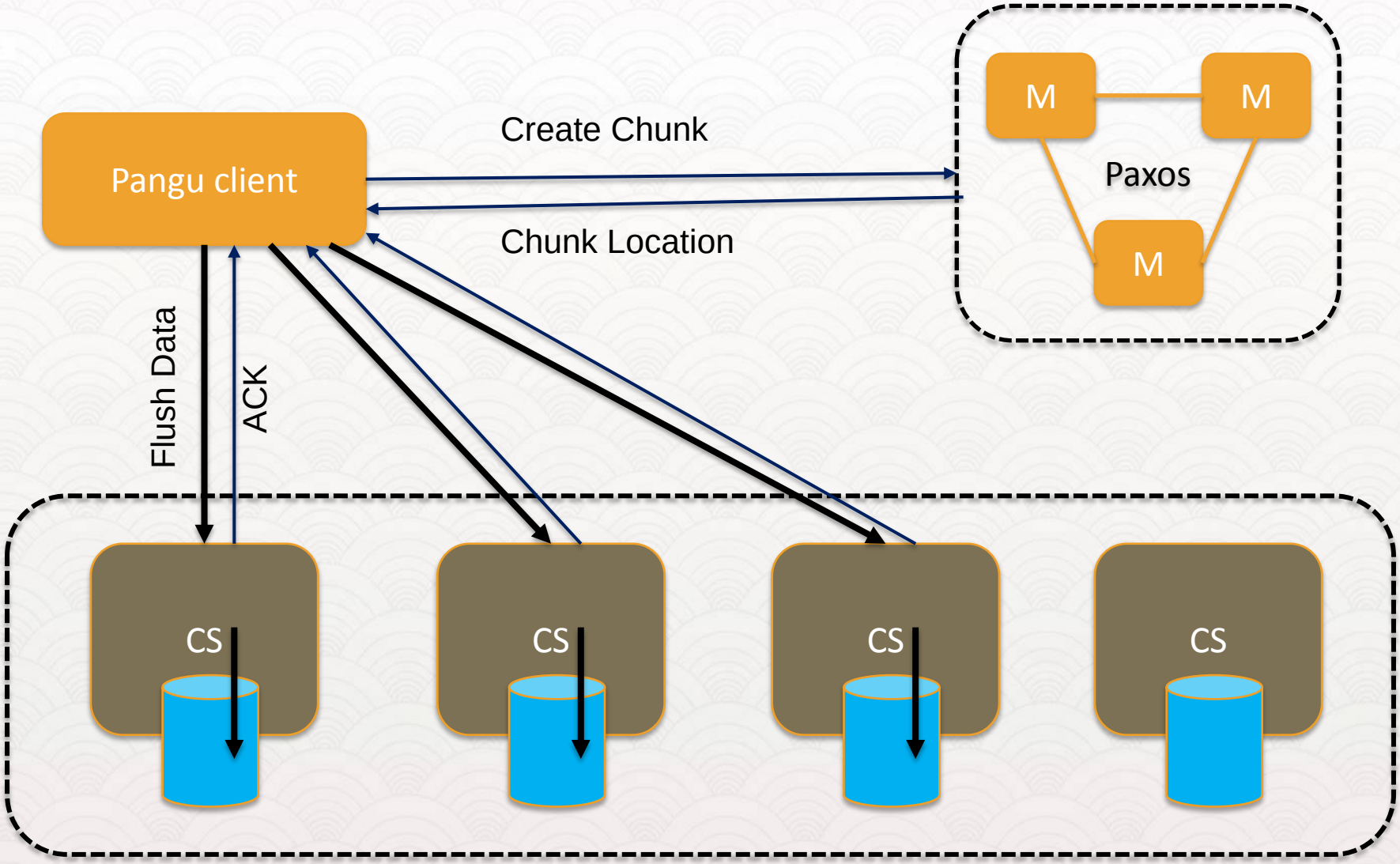


- 对服务访问的数据打标签<app, part>
- 相同<app, part>标签的文件在存储时聚簇

写Pangu Normal File



写Pangu Log File



Agenda

什么是开放存储服务

开放存储整体设计

- 协议接入层
- 数据访问层
- 持久存储层

取舍和教训

取舍

- **数据访问层索引**
  - 根据Value大小切分成多级索引，减少merge开销
  - BlockIndex/BloomFilter分多级，内存更可控
  - 成本：查询大对象需多级索引，不命中cache会变慢
- **只能追加的持久层**
  - 复制、错误恢复相对简单
  - 维护一致性、查错比较容易
  - 成本：需要回收垃圾数据
- **Normal/Log File**
  - Normal偏throughput优化
  - Log偏latency
  - 成本：工程复杂度增加

经验与教训

- **在线升级**
  - Online的互联网存储系统也需要升级
- **质量**
  - 错误模拟测试不能手软
  - 埋点提升错误发生概率并触发大量随机Failover
  - 小概率故障必然会发生
  - 无论多难触发的Bug也会在线上发生

阿里云 aliyun.com

谢谢

http://oss.aliyun.com

云之上的存储服务

简单易用，安全可靠

规模产生的成本效应