淘宝网
Taobao.com

OceanBase
淘宝云存储实践

2011.9

rizhao.ych@taobao.com

SACC2011

淘宝网 Taobao.com

在线存储的挑战

- 海量数据：PB级别，每天处理数据量TB级别
- 时效性：秒级，有时可折衷到分钟级
- 性能高
  - OLTP：几十万QPS，几万TPS
  - OLAP：支持千万行记录实时计算
- 易用性：支持类SQL使用方式

淘宝网 Taobao.com

已有方案分析

- 关系型数据库：功能好，但可扩展性不够
- 分布式方案
  - 类Dynamo方案（Cassandra）：弱一致性，关注度越来越低
  - 类Bigtable方案（HBase）
    - 功能支持弱，缺乏事务，跨机房，宕机恢复等各种问题；
    - 线下增量计算及分析型应用
- NOSQL与SQL融合方案(Google Megastore)：不开源
- “拿来主义”不能完全满足真实需求

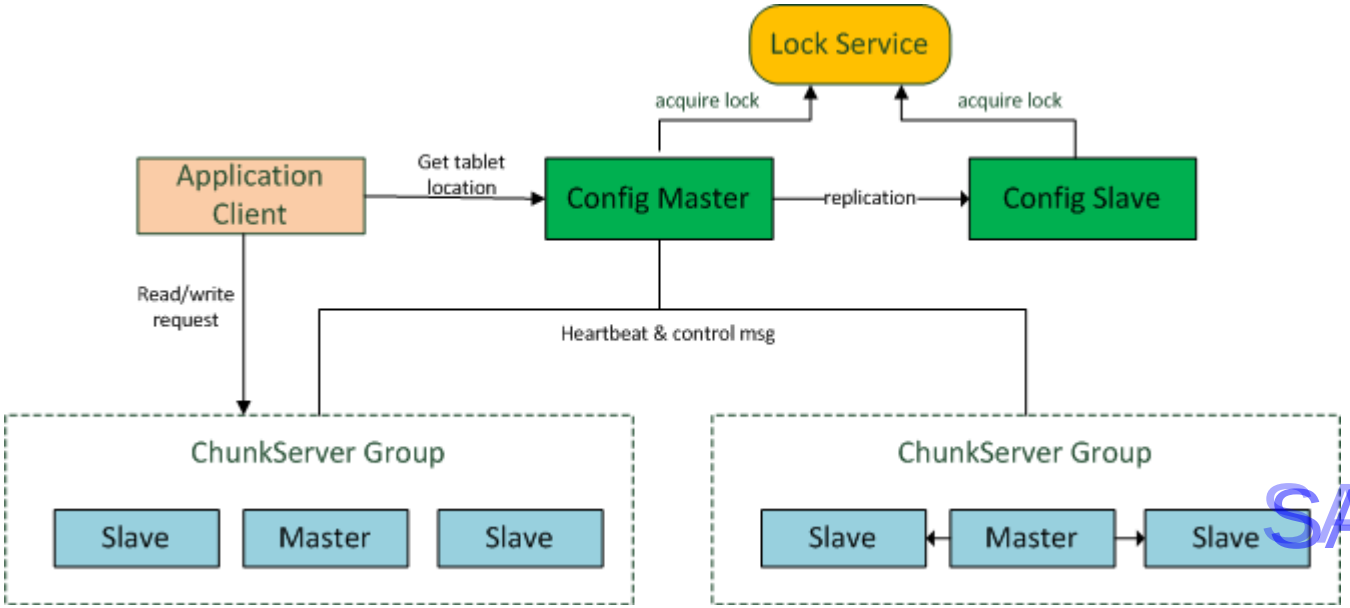
淘宝网 Taobao.com

自主研发通用存储

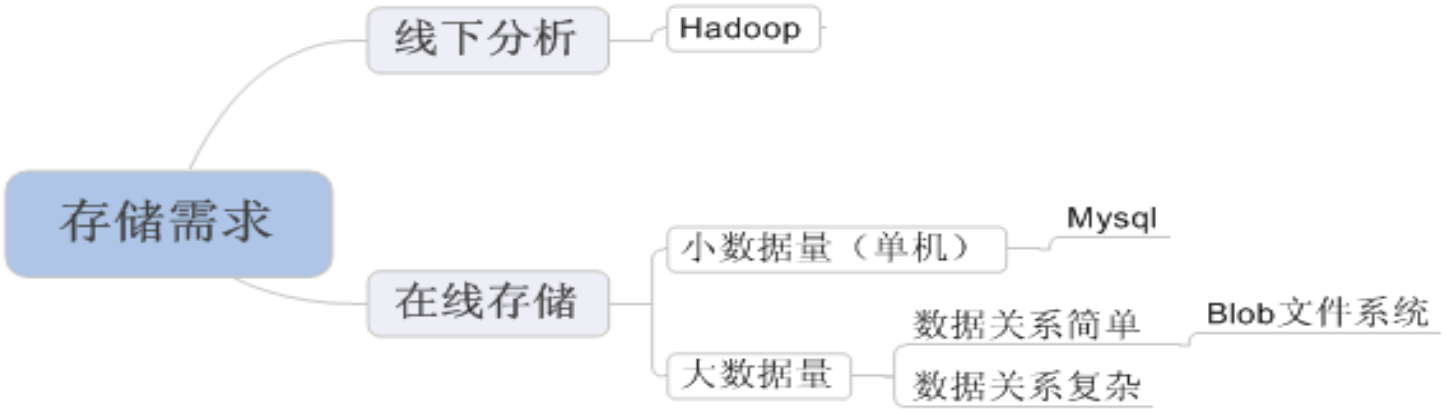
- 解决特定存储难题
  - 海量数据的事务；
  - 大表Join问题；
- 更好的大数据量OLTP解决方案
  - 低成本，高性能
  - 可扩展，舍弃无用的SQL功能；
- 支持千万级海量数据的实时OLAP分析
- 存储服务化

可扩展性挑战

- 同构的分布式系统
  - 将机器分成group，每个group内的机器存放的数据完全相同
  - 问题：数据迁移量太大，group内部增加副本做不到自动化；
    - 假设服务数据量1TB，内部传输带宽限制20MB/s，增加副本的时间为1TB / 20MB/s = 50000s，大约10几个小时；
    - 迁移过程中机器再次出现故障怎么办？

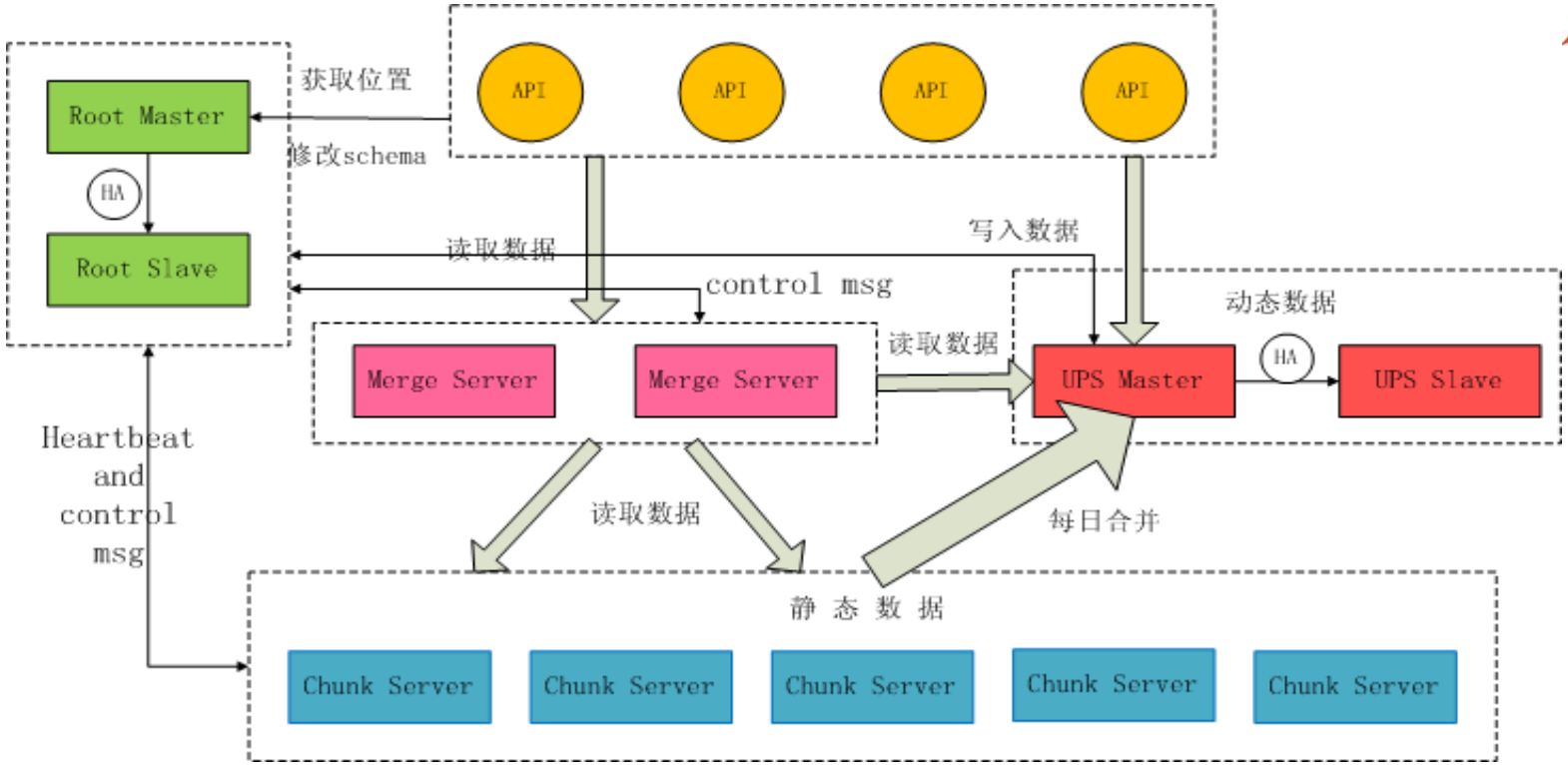


存储需求分解



- Oceanbase提供复杂关系的海量数据在线存储方案
- 在线存储特点：数据量大但最近一段时间修改数据量不大
  - 静态数据和动态数据分离
  - 动态数据不断地合并到静态数据
  - 静态数据：数据量大，一般采用SAS存储；
  - 动态数据：数据量小，一般采用内存或者SSD服务；

Oceanbase一期架构



- 主控服务器RootServer：主+备，Schema/B+树根节点/机器管理…
- 动态数据服务器UpdateServer：主+备，实时修改(内存+SSD)
- 静态数据服务器ChunkServer：多台，B+树数据节点(磁盘或SSD)
- 查询合并服务器MergeServer：多台，静态动态数据合并…

Oceanbase一期特点

- 功能
  - 数据模型类似关系型数据库
  - 强一致性
  - 支持跨行跨表事务
  - 高效Join
- 可扩展性：支持动态增减机器，无需分库分表；
- 可靠性高，机器宕机秒级恢复；
- 自动负载均衡；
- 采用Copy-on-write技术，单写多读不加锁，性能好；
- 支持Online schema change；

淘宝网 Taobao.com

Oceanbase一期落地

- 收藏夹需求
  - 收藏表保存收藏信息条目，40亿+
  - 商品表保存收藏的宝贝详细信息，4亿+
  - 收藏夹展示：收藏表和商品表两张大表join
- 收藏夹挑战
  - 一个用户可以收藏数千商品
  - 一件商品可被数十万用户收藏
  - 商品的属性实时变化
  - 单次查询响应时间<50ms
- 实验效果
  - Mysql 16 * 2减少为Oceanbase 12 + 2
  - Load值更低，短期无扩容需求；
  - 平均响应时间 <50ms

硬件利用

- SSD
  - 随机读性能好 (Intel SSD：3~4万IOPS vs SAS磁盘：180 IOPS)
  - 顺序写性较很好(~100MB/s)，随机写性能差
  - SSD ￥20/GB vs SAS ￥3/GB
  - OceanBase
    1. 数据先进入内存，超过阀值时一次性dump到SSD
    2. 很多随机读，大块顺序写，**没有随机写**
    3. 读操作基本不加锁，充分发挥SSD；
- 万兆网卡
  - 多线程收发包
  - 千兆网卡：50万/s
  - 万兆网卡：500万/s，780MB/s

淘宝网 Taobao.com

跨机房支持

- 一期方案
  - 主机房写节点将操作日志同步到备机房
  - 客户端配置多个机房地址，发生切换后自动轮询
  - 机房内自动切换，跨机房人工脚本切换
  - 暂不依赖类Paxos锁服务
- 二期方案
  - 支持按比例将流量切分到多个机房

淘宝网 Taobao.com

其它应用

- CTU：每天写入25亿条，写入数据量2.5TB
  - MongoDB => Oceanbase
  - 5个集群，单UPS一天写入500G
  - 逐步上线中，目前单UPS写入约为200GB；
- 量子统计：千万条数据的实时统计
  - 统计三个月数据，最大用户每天数据量10~20W条
  - OLAP分布式统计功能开发中，预计年底部分上线
- SNS feed index：Cassandra => Oceanbase
- 店铺装修。。。

淘宝网 Taobao.com

后续功能-SQL

- 索引支持
  - 如何支持动态创建/删除分布式索引？
  - 查询操作如何智能地选择索引？
- SQL支持
  - 支持where, having, group by, order by, limit, offset
  - 支持IF语句；
  - 支持IN, 比较运算，四则运算
  - 不支持嵌套查询，外键约束等；

淘宝网 Taobao.com

后续发展 - OLTP

- 千亿级海量数据库
  - 每天几百GB到TB写数据量，总数据量百TB级别
  - 单UPS，支持跨行跨表事务
  - UPS单机性能持续优化，探索更好的UPS硬件
  - 更多SQL功能支持，支持简单的SQL优化
  - 相当于廉价的Oracle + 小型机 + EMC共享存储方案
  - 存储服务化：支持多UPS
    - 将数据划分为多个entity group，entity group内部强一致性，entity group之间最终一致性

淘宝网 Taobao.com

后续发展 - OLAP

- 分析型应用
  - 支持千万级记录实时计算（类似一次MapReduce）（Doing）
    1. 请求拆分
    2. 请求分发到多个ChunkServer
    3. 每个ChunkServer计算局部结果
    4. 合并汇总，计算top N等
  - 支持多UPS
  - 支持按列存储；（Done）
  - 压缩算法研究；
  - OLAP算法优化探索，如top N，distinct近似算法

Oceanbase开源

- OB开源
  - 只有心态开放，才能发展得更好
  - 原版，无删减
  - 系统发展中，很简陋但将持续改进
  - 开源地址：http://code.taobao.org/project/view/587/

经验总结

- 客户第一，应用为王
  - 平台需求来源于业务需求提取
  - 避免过度设计
- 重视测试
  - 测试资源严重不足
  - 单元测试 + 代码Review
  - 开发执行模块级压力测试，测试执行系统级测试
  - 开发编写并执行测试用例，测试专注测试工具自动化
- 成败在于细节
  - 数据拷贝：Direct IO，权衡接口模块化与性能
  - 内存分配：内存池，线程缓存
  - 锁：线程缓存，减少Cache锁冲突，copy-on-write数据结构
  - 上下文切换：替换传统的任务队列模型

淘宝网
Taobao.com

谢谢

- 杨传辉（日照）
- rizhao.ych@taobao.com
- 新浪微博：淘宝日照
- 个人博客：http://nosqlnotes.net