淘宝网
Taobao.com

# Tair存储系统在淘宝大规模应用实践

淘宝-核心系统-存储
杨成虎


# 目录

淘宝网
Taobao.com

# 1. 初看tair

Tair在淘宝的现状

# Tair 诞生

1. 请求太慢了，要读IO，那就做个Apache Module，数据放内存里

2. 需要提高命中率，集中存放，弄个简单的分布式--Tdbm

3. 业务太多了，需要管理机器和资源--Tair

# 2011年Tair部署规模

- 500台机器
- 40个集群
- 80%的是cache，承载了90%请求
- 百亿级别的记录


双11，双12是平时流量的3～4倍

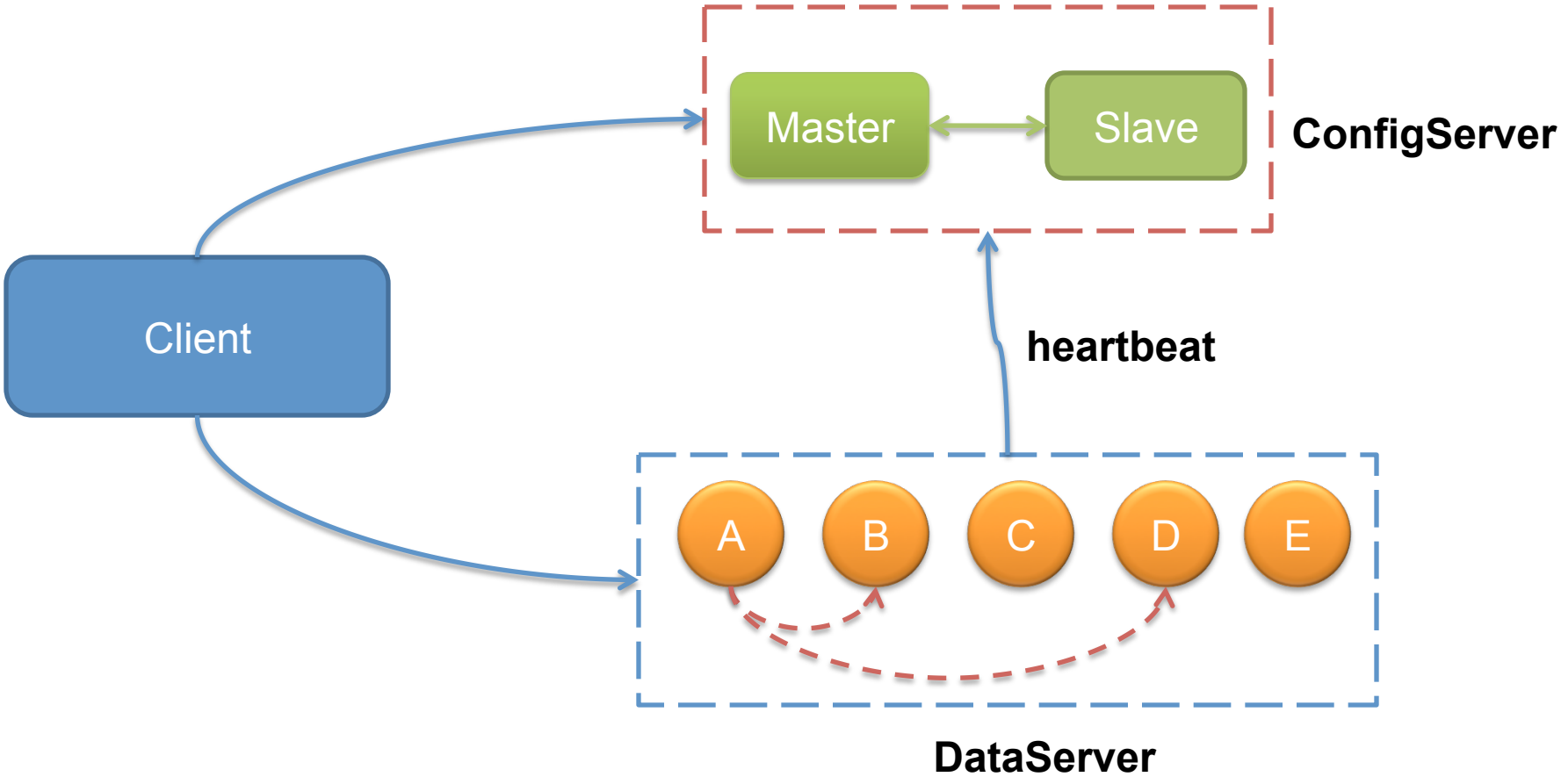
淘宝网
Taobao.com

# 2. 原理解析

## 数据路由与节点管理

# 2.1 模块介绍

- ConfigServer
  - 控制节点，路由信息
- DataServer
  - 数据处理与管理
- Storage
  - 存储接口
- Client
  - API C++/JAVA/Restful
  - localcache


Master
Slave
ConfigServer
Client
heartbeat
A
B
C
D
E
DataServer

# 2.2 ConfigServer

- 管理数据路由信息
- 主备
- 轻量级，足够简单
- 短时间的故障不会影响服务

# 2.2.1 对照表

- 数据划分成Bucket
  - bucketid = hash(key) % n
- 每个Bucket 只属于一台Server
- 一台Server包含n个Bucket

# 2.2.1 对照表

2个节点

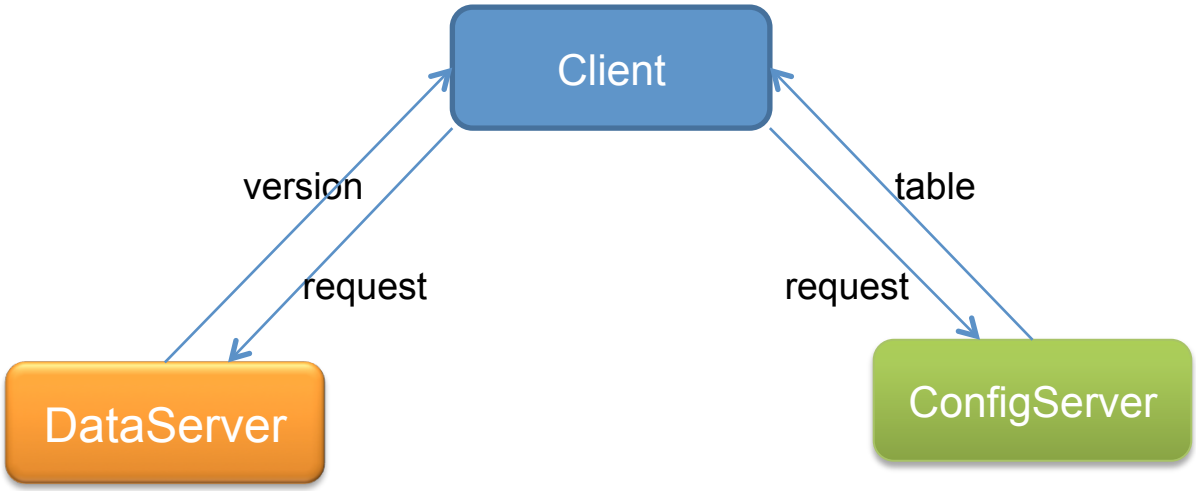
| Bucket number | datanode |
|---|---|
| 1 | 192.168.100.1 |
| 2 | 192.168.100.2 |
| 3 | 192.168.100.1 |
| 4 | 192.168.100.2 |
| 5 | 192.168.100.1 |
| 6 | 192.168.100.2 |

2个节点扩展到3个节点

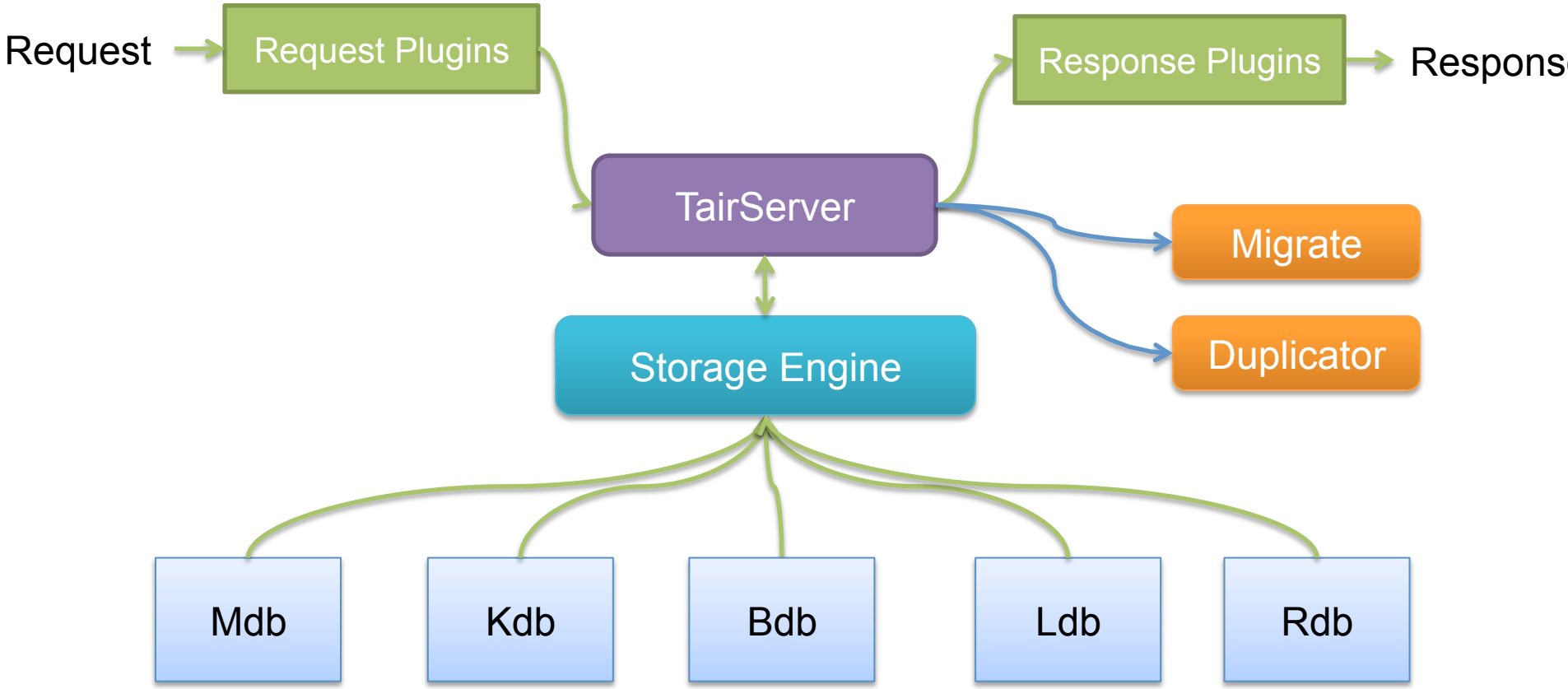
| Bucket number | datanode |
|---|---|
| 1 | 192.168.100.1 |
| 2 | 192.168.100.2 |
| 3 | 192.168.100.1 |
| 4 | 192.168.100.2 |
| **5** | **192.168.100.3** |
| **6** | **192.168.100.3** |

# 2.2.2 对照表的同步

- 客户端缓存
- 版本号机制来同步



# 2.3 DataServer



# 2.4 Storage

- MDB
  - 共享内存，高性能，Cache系统
- LDB
  - 持久化，Leveldb改造，定制化修改
- RDB
  - 数据结构丰富，使用了Redis的数据结构

淘宝网
Taobao.com

# 3. 实施案例分享

Tair在淘宝应用

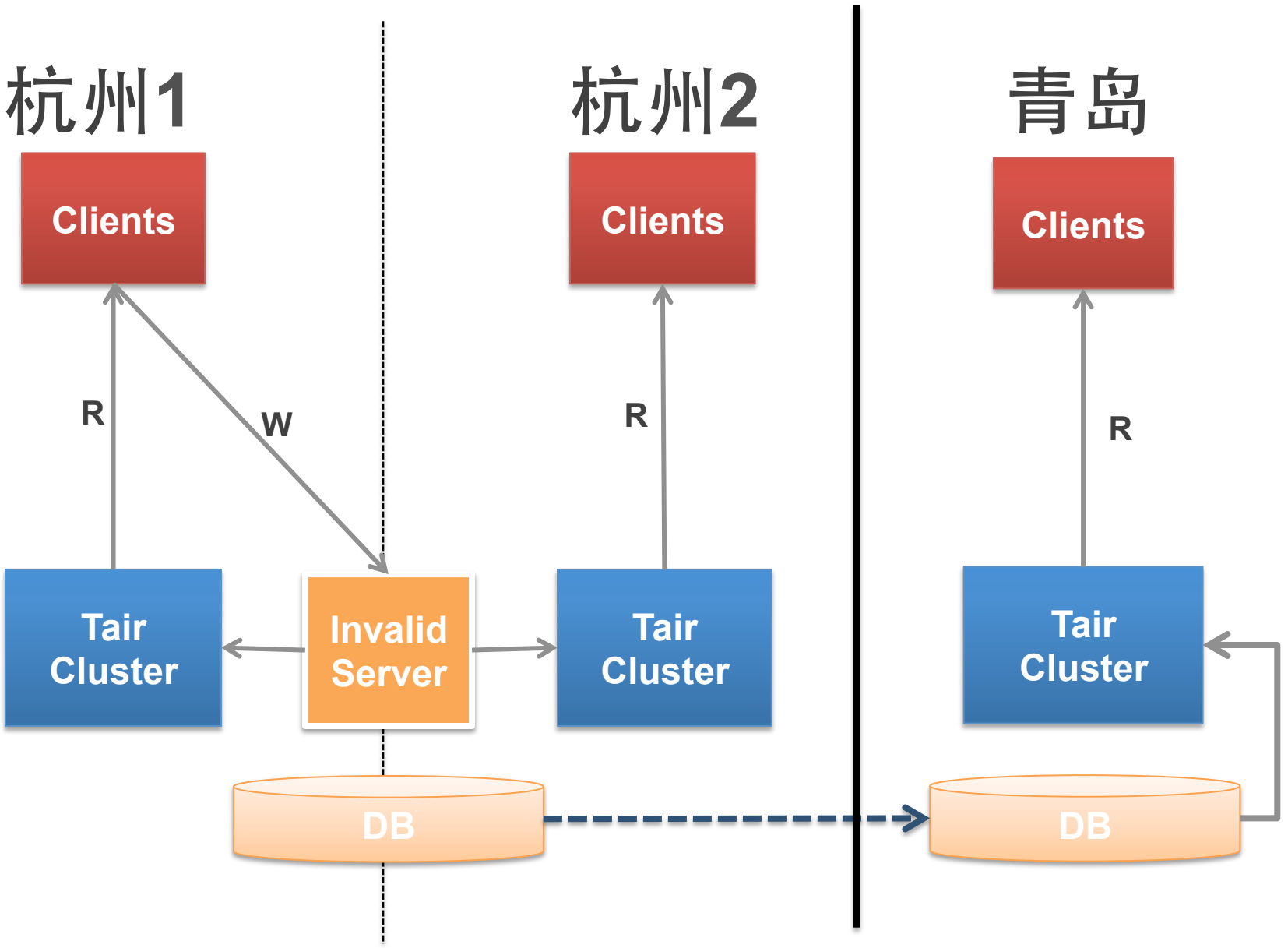
淘宝网
Taobao.com

# 3.1 产品信息/用户信息缓存

商品/用户

# 3.1.1 产品信息Cache需求

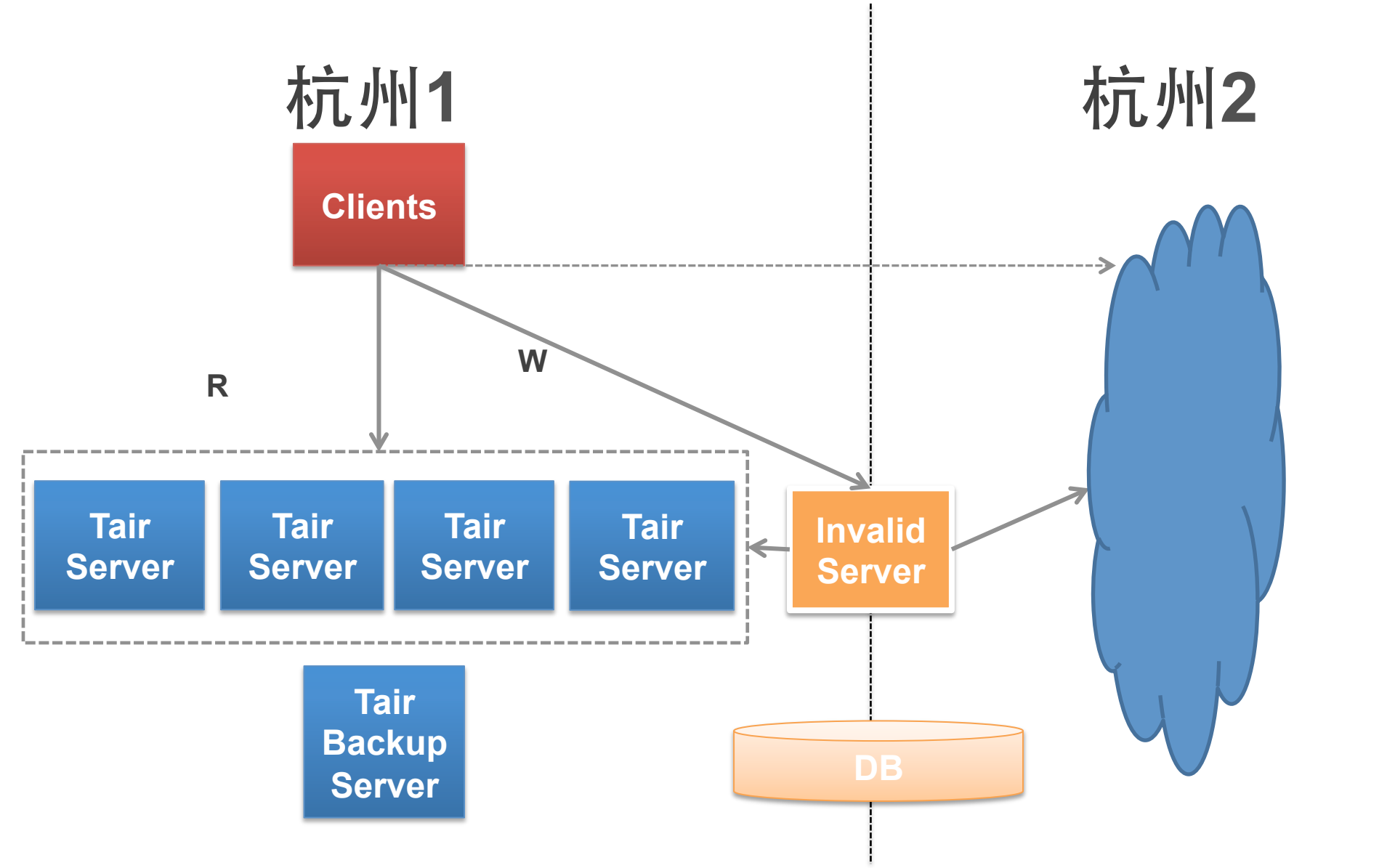
- 高性能，且稳定
  - 内存，MDB
- 容灾，机房容灾，地域容灾
  - 多集群，多地域部署


杭州1
杭州2
青岛
Clients
Clients
Clients
R
W
R
R
Tair Cluster
Invalid Server
Tair Cluster
Tair Cluster
DB
DB

但是，Tair 宕机时，请求还是会穿透到DB上！

# 3.1.2 用户信息Cache-需求

- 高性能，且稳定
  - 压缩数据可以调高性能
- 访问量相对商品信息大一个数量级
  - Tair机器不能宕机
- 可靠性要求高
  - 双备份？数据同步代价？成本问题，不能每个机房都2份
- Tair 宕机给数据库带来巨大压力
  - 需要快速恢复


杭州1
杭州2
Clients
R
W
Tair Server
Tair Server
Tair Server
Tair Server
Invalid Server
Tair Backup Server
DB

# 3.1.2 用户信息Cache-运维

- 流量恢复前，要预热cache
  - 保证命中率
- 分批恢复流量
- 异常发生时，自动切换流量，恢复时需要人工检查。

# 其他问题

- 个别Key过热访问
  - 活动
  - 恶意攻击
- 消息体过大，容易撑满网卡

# 3.1.3 LocalCache

- Client 缓存访问量特别多的请求
- 本地缓冲极高的热点
- 拦截掉部分攻击
- 脏数据
  - 极短的过期时间
  - 级别可配置(PUT GET)

淘宝网
Taobao.com

# 3.2 定制化的改造

LDB针对性优化

# LevelDB 1分钟

特性

- KV Database
- Bigtable Chrome
- Memtable -> SSTable
- LSM(log-structured-merge)

场景

- Key基本有序
- 更新少
- 小数据
- 读取随机（SSD）

# Tair存储引擎之LDB

- 基于LevelDB定制化开发
  - Namespace
  - 多实例
  - Range
  - Cache
  - FastDump
  - BloomFilter

# 3.2.1 BloomFilter

- 查找不可怕，可怕的是每次都找不到
  - 检索大量文件，无效的随机IO

- BloomFilter Patch
  - 需要足够大的内存，将BloomFilter放在内存中

# 3.2.2 FastDump

- 广告，推荐系统，数据仓库
  - 定点批量导入
  - 老数据无效
- 现有系统，导入时间过长
  - 24小时都不够?
  - 消耗大量机器资源

# 3.2.3 FastDump 优化数据结构

- LDB优化
  - Kernel调优，最大化SSD性能
  - 多实例，每实例一块SSD
  - 数据源提前排序
  - 既能并发，写入又有序
    - 并发量大于集群磁盘数量
    - Key 按Range划分
    - 多memtable，每memtable对应一Range
  - 去oplog

# 3.2.3 FastDump Service

- 两组集群互切换
  - 一组online，负责提供对外服务
  - 一组offline，接收dump数据
  - Offline提供数据容灾备份

- 100台->6台
- 每台TairServer提供 120w tps导入
- 24min

淘宝网
Taobao.com

# 3.3 快照信息存储

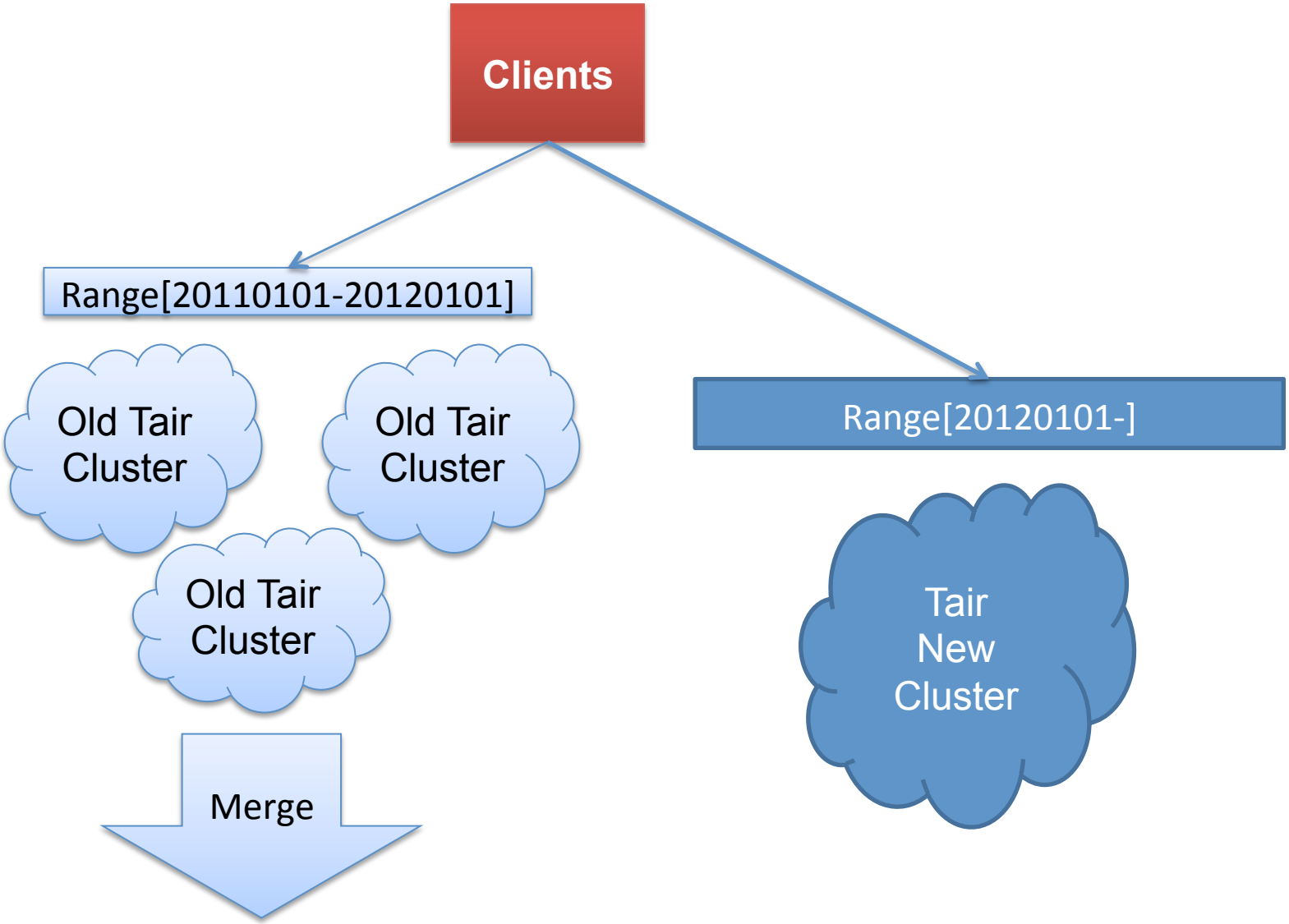
交易快照

# 3.3.1 淘宝交易快照

- 单条数据小，条数多，KV模型
  - 适合NoSQL
- 无覆盖，不能丢
  - 需要持久化
- 持续的增长
  - 扩容难题
- 只读取近期数据
  - 老数据价值降低

# 3.3.2 交易快照 on LDB

- 数据模型对LDB友好，写入性能高
  - 无修改，只增加
- 大内存，近期数据在内存中留存一份
  - 磁盘只有顺序写
- 数据量大，SSD不合适，SATA
  - 降低成本
- 扩容以集群为单位，老集群只读，Merge到更廉价的设备


Clients
Range[20110101-20120101]
Old Tair Cluster
Old Tair Cluster
Old Tair Cluster
Merge
Range[20120101-]
Tair New Cluster

淘宝网
Taobao.com

# 总结与未来


## 总结

- 分布式KV框架
- 支持LDB RDB MDB
- Client Java/C++/Restful

- 认清需求，根据业务特点部署
- 特性的解决方案扩展到通用方案

## 未来

- 继续开源，且所有的优化都开源

- 扩大交流与支持

- 服务化
- 流量权限控制完善
- 插件脚本
- 兼容memcached协议
- ……

淘宝网
Taobao.com

# 谢谢
# Q&A

# ArchSummit

## 中国·深圳 2012.08

## INTERNATIONAL ARCHITECT SUMMIT

## 全 球 架 构 师 峰 会

详情请访问：architectsummit.com

- 3天 • 6场主题演讲
- 3场圆桌论坛 • 9场专题会议
- 国内外30余家IT、互联网公司的50多位来自一线的讲师齐聚一堂

主办方：InfoQ 战略合作伙伴：Tencent 腾讯 特别支持：

http://architectsummit.com


QCon
杭州站 · 2012年10月25日~27日
www.qconhangzhou.com（6月启动）
QCon北京站官方网站和资料下载
www.qconbeijing.com