



十年架构 成长之路

SACC 第十届中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2018

2018年10月17-10月21日 北京海淀永泰福朋喜来登酒店



京东KV存储产品演进之路



第十届中国系统架构师大会
SYSTEM ARCHITECT CONFERENCE CHINA 2018



KV存储体系简介

内存存储-jimdb

- 高吞吐、低延迟
- 在线伸缩、自动故障恢复
- 广域复制

持久化存储-sharkstore

- 高可靠
- 强一致
- 大容量

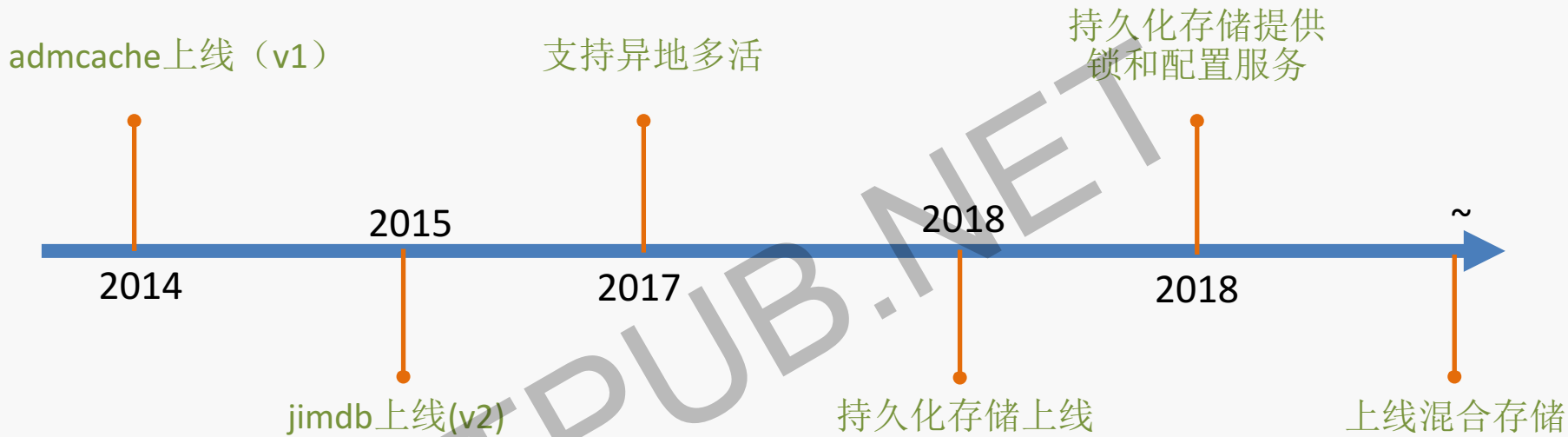
混合存储

- 冷热数据存储分离
- 内存级别的性能体验，磁盘级别的存储容量



十年架构 成长之路







1

内存存储

2

持久化存储

3

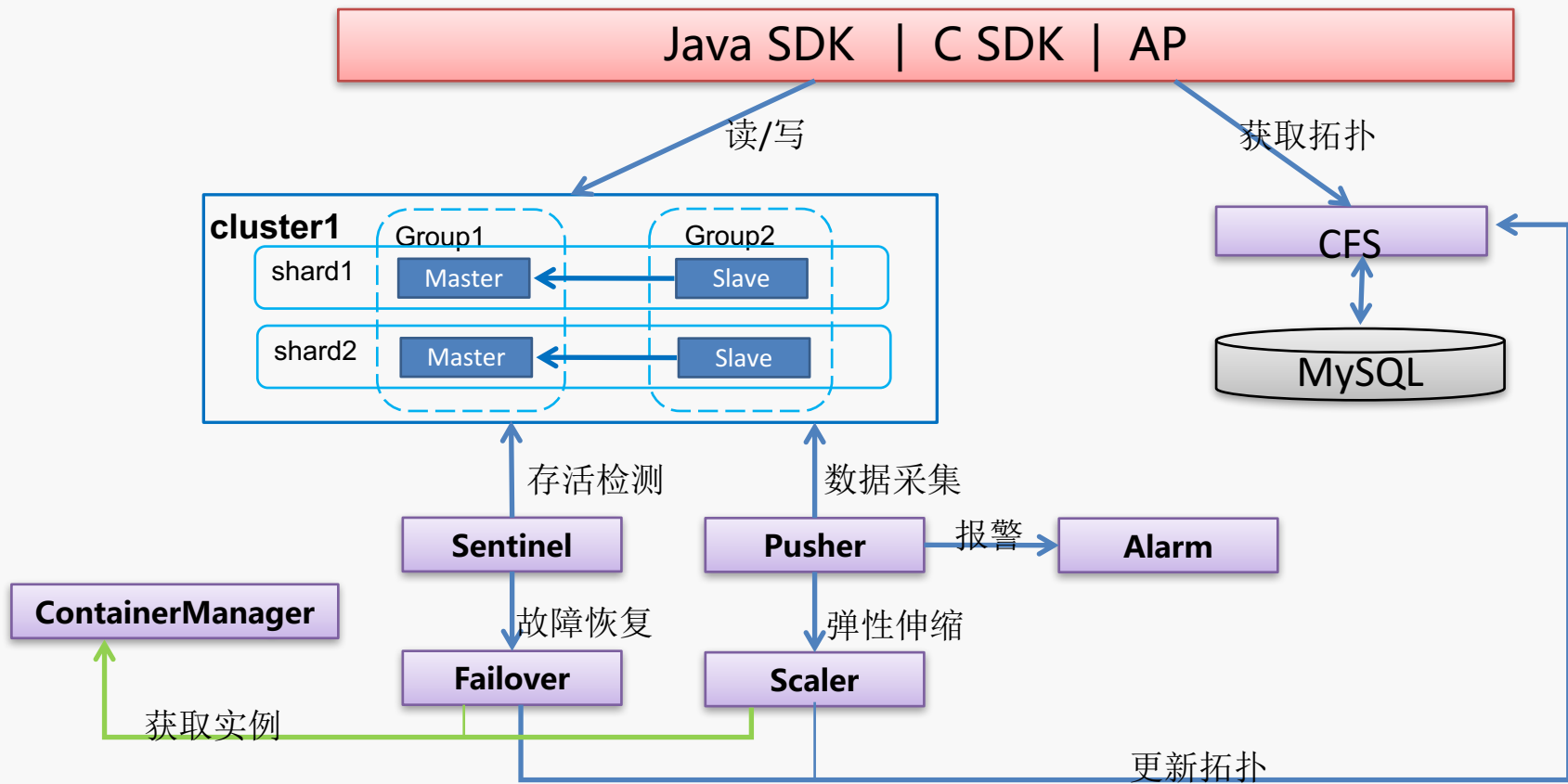
混合存储

内存存储-运营数据



十年架构 成长之路





内存存储-核心特性

- ❖ 高吞吐、低延迟
- ❖ 自动故障恢复
- ❖ 在线伸缩
- ❖ 广域复制



十年架构 成长之路

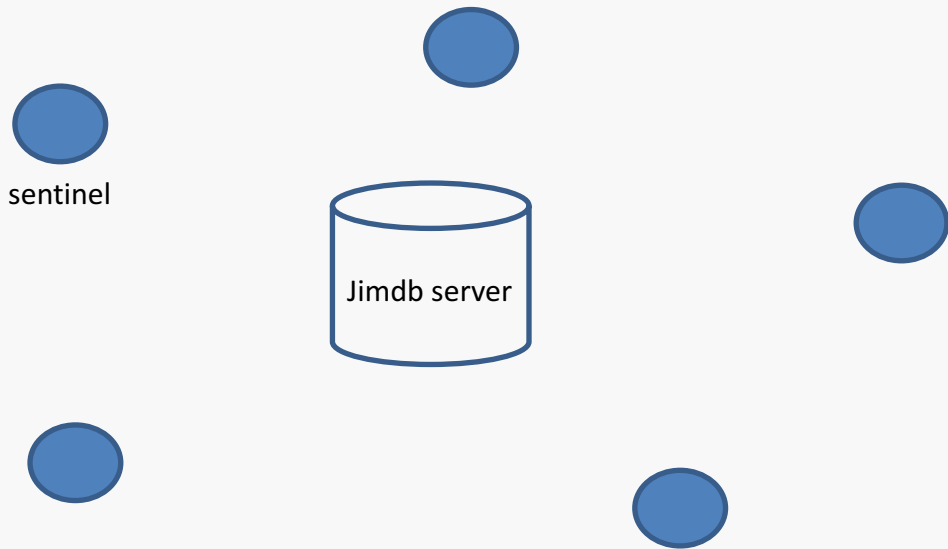


内存存储-故障检测

误判后果

误判原因

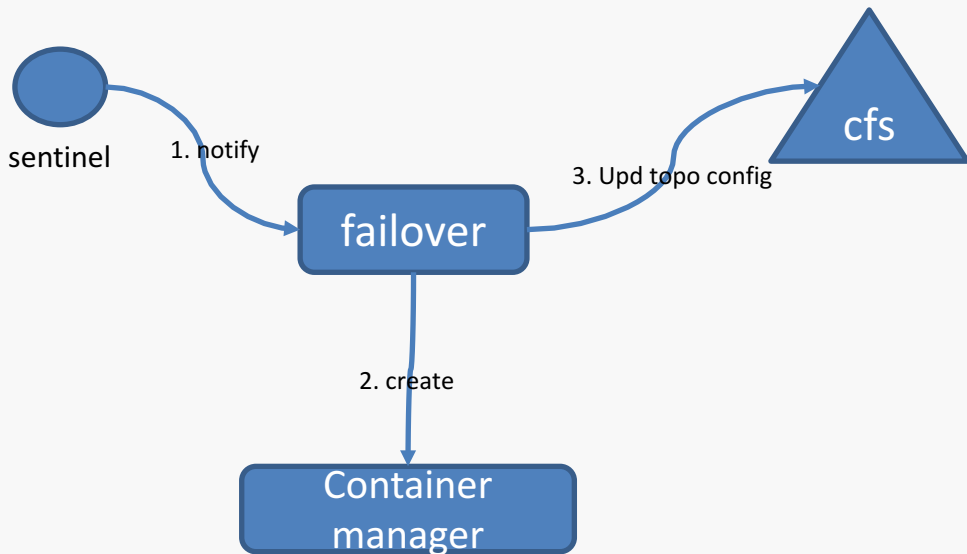
防止误判



十年架构 成长之路



内存存储-故障自动恢复



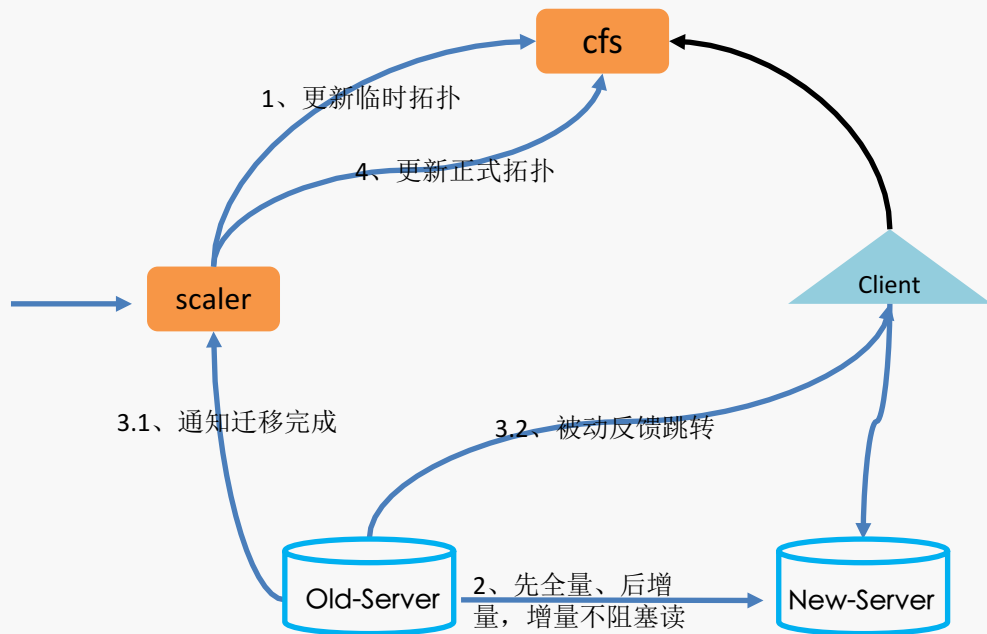
客户端容灾
大规模断电：熔断、限流



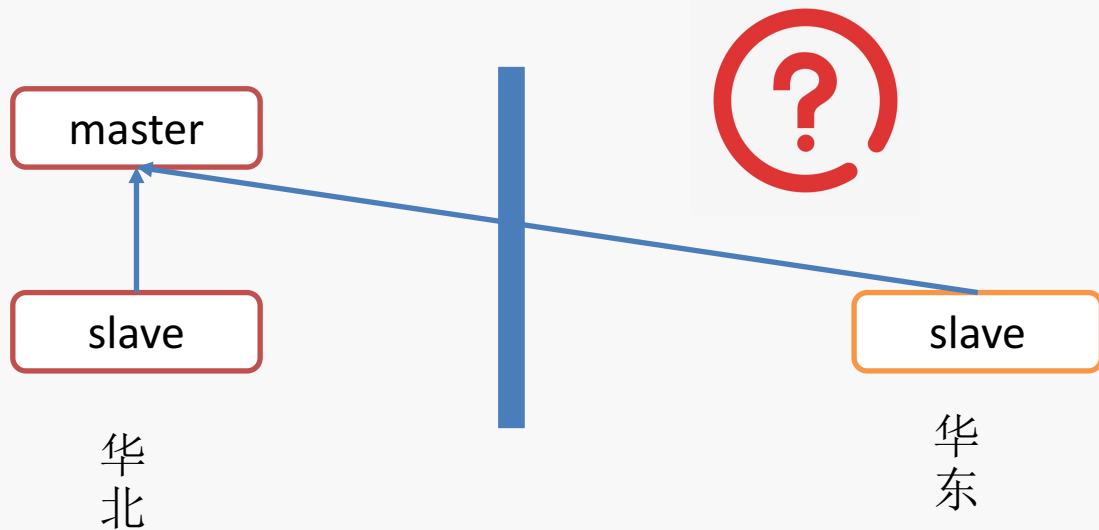
十年架构 成长之路



内存存储-在线迁移流程



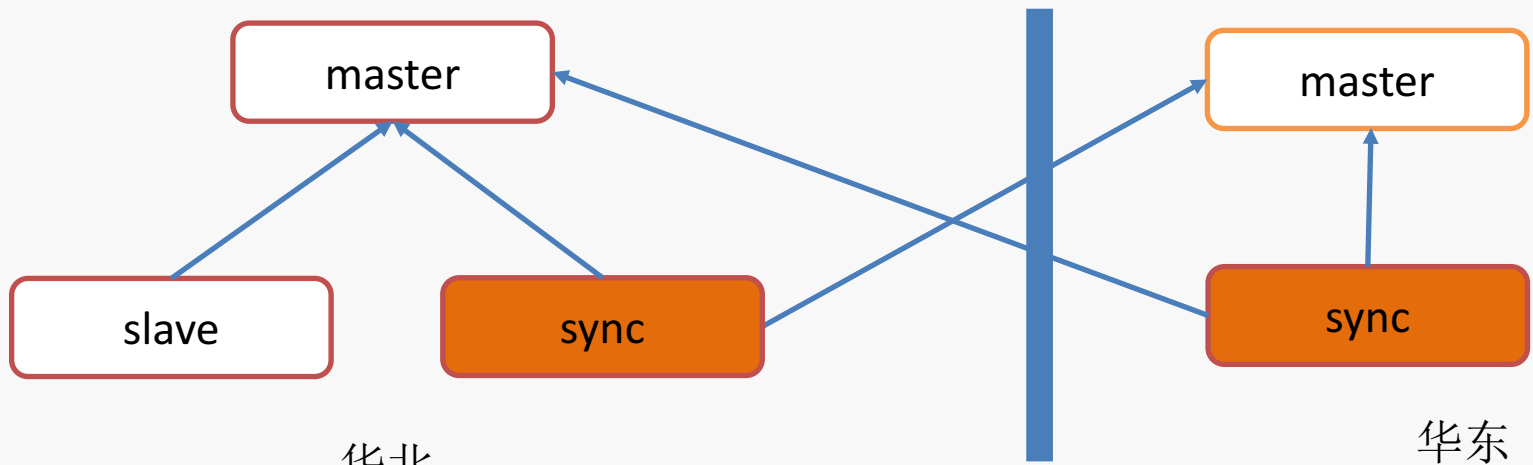
内存存储-广域复制的问题



- ❖ 专线带宽有限，质量无法保证，延迟高
- ❖ 循环缓冲区大小有限，网络质量不可靠的情况下极易造成频繁全量同步



内存存储-广域复制



- ❖ sync增量数据落盘，^{华北}保证网络抖动、延时高等广域网络因素不会触发复制
- ❖ 支持双向复制
- ❖ 源端和目标端能够异构，能够各自进行故障切换



十年架构 成长之路



1 内存存储



2 持久化存储

3 混合存储

持久化存储-核心特性

- ❖ 分布式强一致
- ❖ 支持在线分裂、自动故障恢复
- ❖ 支持schema，海量数据
- ❖ 支持范围查询，单表操作

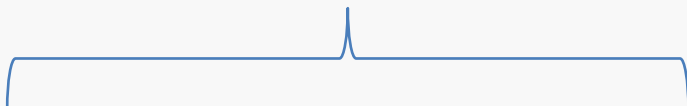


十年架构 成长之路



持久化存储-逻辑视图

Primary Key



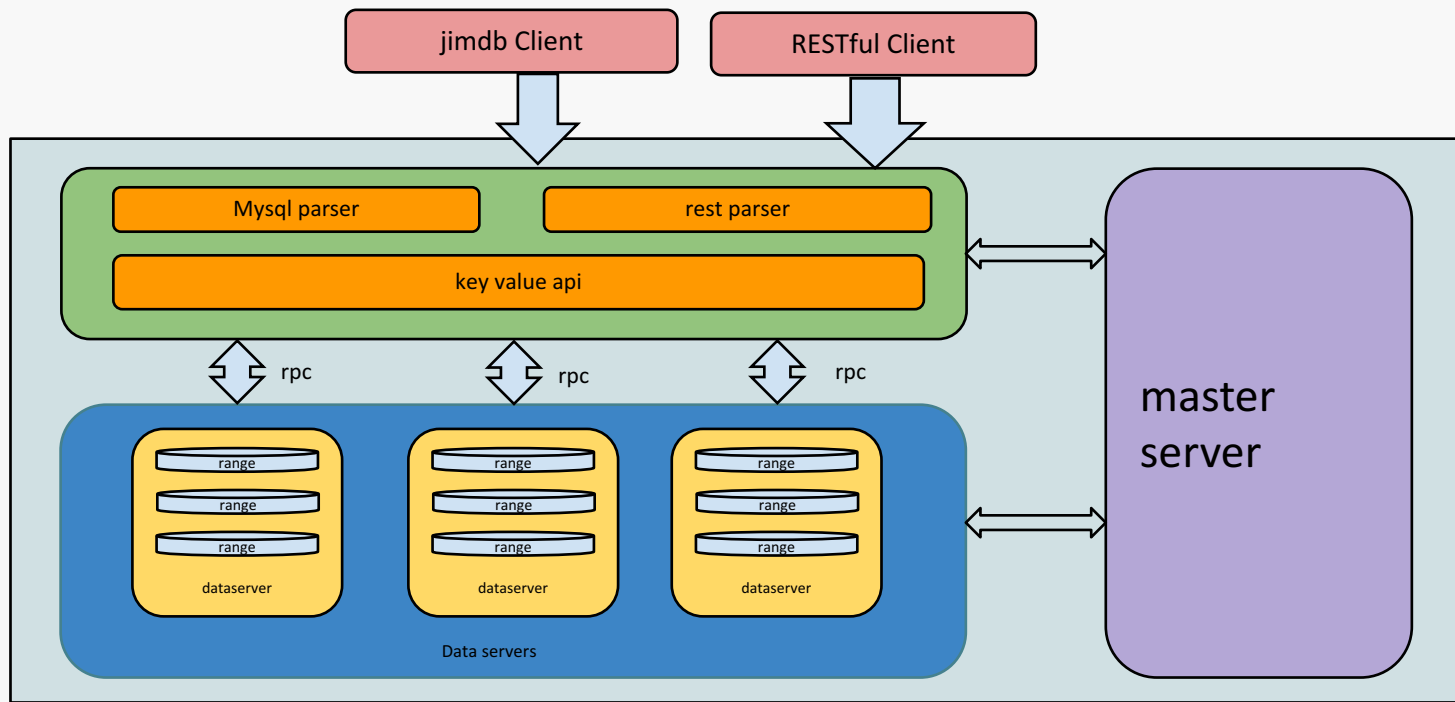
Column 1	Column2	Column3	Colum4	Column5
value1-1	value2-1	value3-1	value4-2	
value1-2	value2-2	value3-2		value5-2
value1-3	value2-3		value4-3	value5-3



十年架构 成长之路



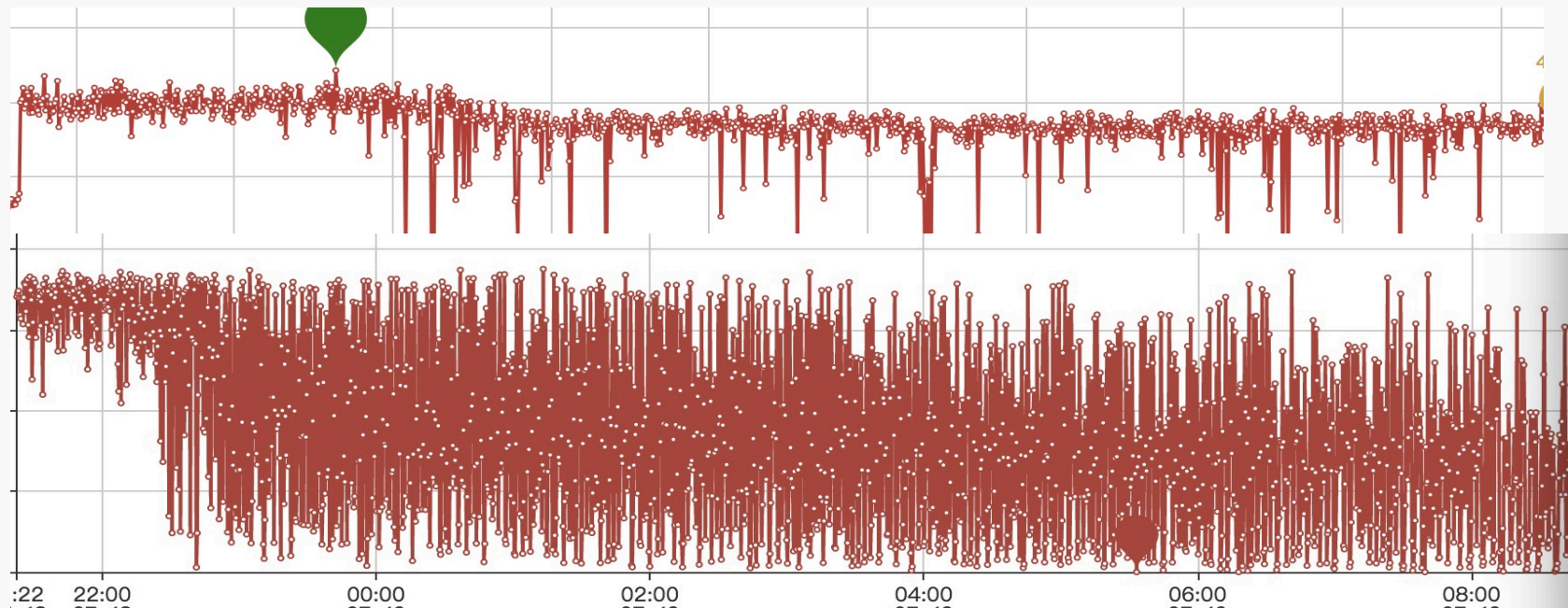
持久化存储-系统结构



十年架构 成长之路



Compact影响写入

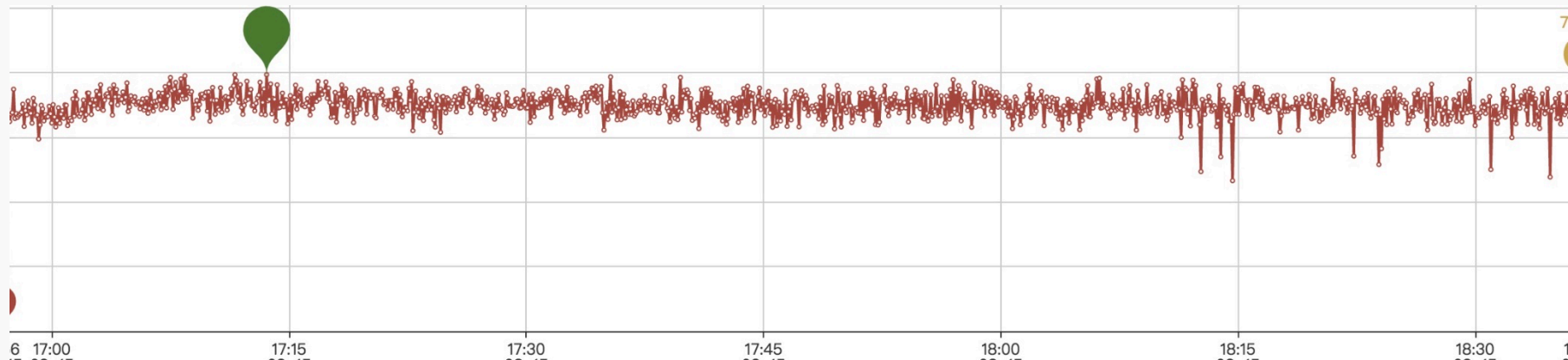


十年架构 成长之路



Key-value分离

基于rocksdb的blobdb功能完善和改造

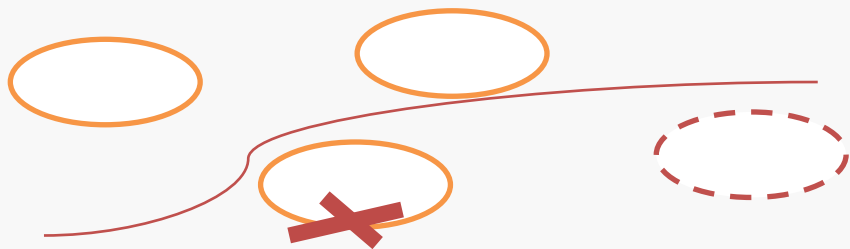


十年架构 成长之路



持久化存储-Raft成员变更

数据迁移、负载均衡等涉及到成员变更



- 1、只复制
- 2、不发起LEADER选举
- 3、不投票

sharkstore raft 会在leader端监测learner的日志进度，
由leader自动发起一次raft成员变更，提升leaner成员为正常成员



十年架构 成长之路



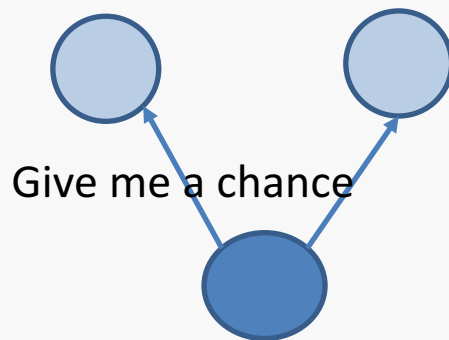
持久化存储-RAFT PREVOTE

follower重新加入复制组会发起选举，可能干扰正常业务

- 1、发起选举的follower本身就是落后节点
- 2、现有的leader可以正常工作，不必要重新选举

Raft的作者提出了prevote算法：

在发起选举前，先进行一次预选举Pre-Candidate，
如果预选举时能得到大多数的投票，再增加term，进行正常的选举。



十年架构 成长之路



持久化存储-延伸

分布式锁

配置服务

sharkstore



十年架构 成长之路



分布式锁

数据库

条件更新->防止死锁->加时间戳检查，数据库高可用，检查死锁任务的高可用，性能和扩展性

Redis/Memcached

服务本身的高可用、redlock算法、合理的超时时间

Etcd/zk

性能、水平扩容、网络异常后任务能否中止



十年架构 成长之路



分布式锁

基于心跳机制：任务中止

基于超时时间：时间设置是否合理、重启服务端等已知场景能否及时解锁

合理的超时时间+心跳续期

安全超时时间+报警检测

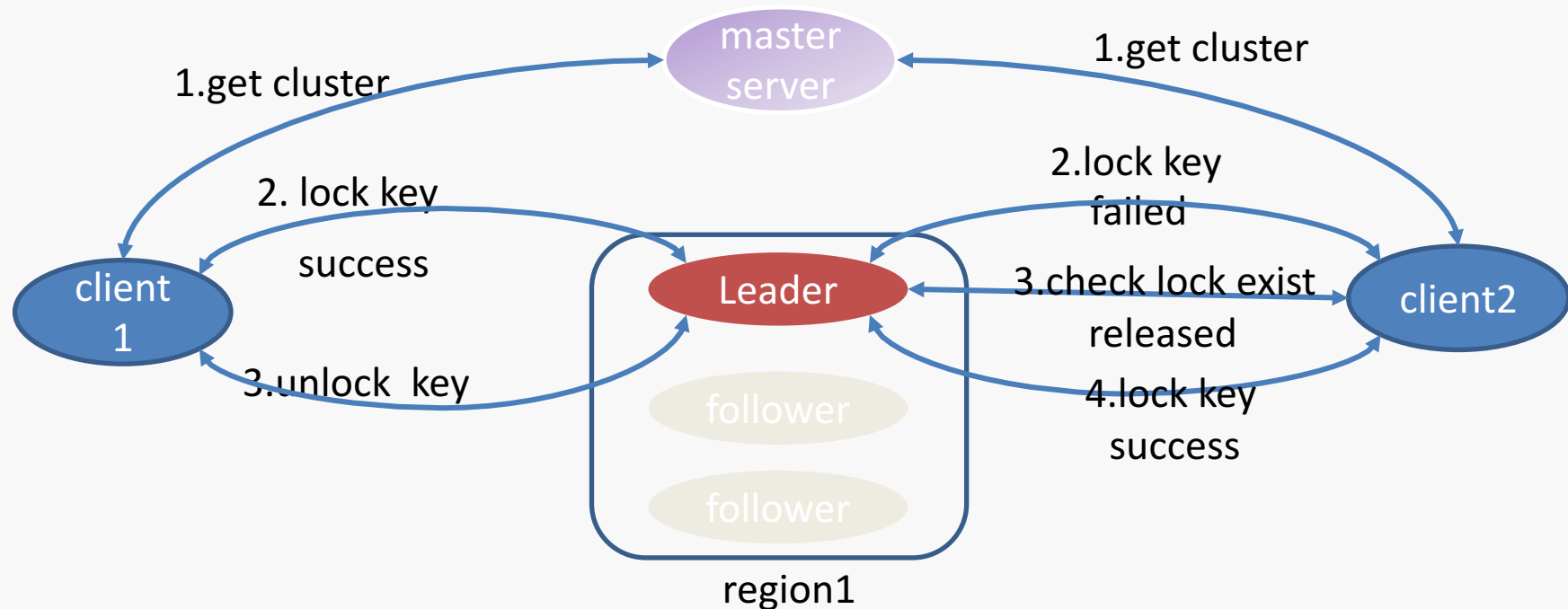
心跳过期+回调



十年架构 成长之路



分布式锁



配置服务

- 1、多range，多点写入
- 2、树型结构深度有限制
- 3、防止分离时group分组被割裂



十年架构 成长之路



1 内存存储

2 持久化存储



3 混合存储

混合存储



十年架构 成长之路



冷热分离引擎

核心问题：不常访问的数据淘汰到磁盘

算法：LRU、HIT DENSITY

问题：

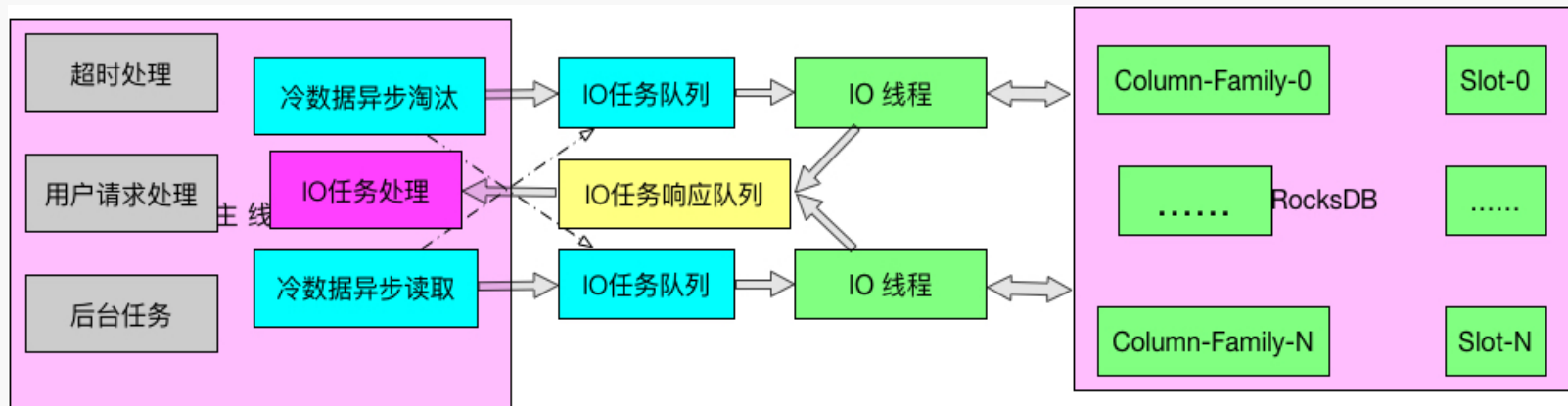
- 1、KEY要不要从内存里面删除
- 2、少量的冷数据访问，会不会拖慢整体性能
- 3、链表数据怎么处理



十年架构 成长之路



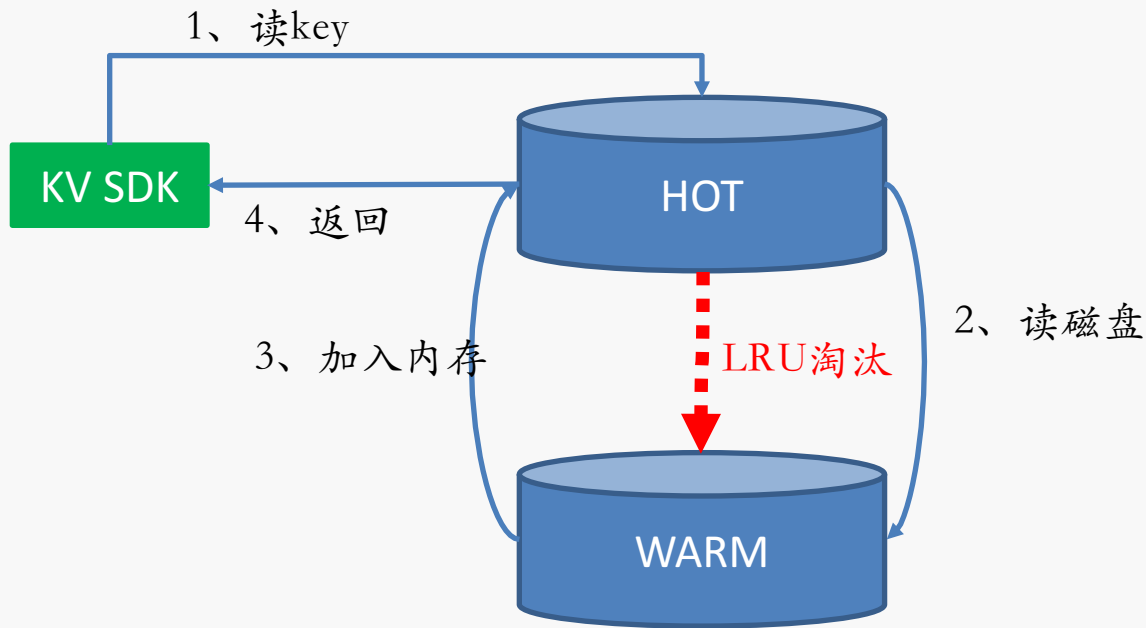
冷热分离引擎



十年架构 成长之路



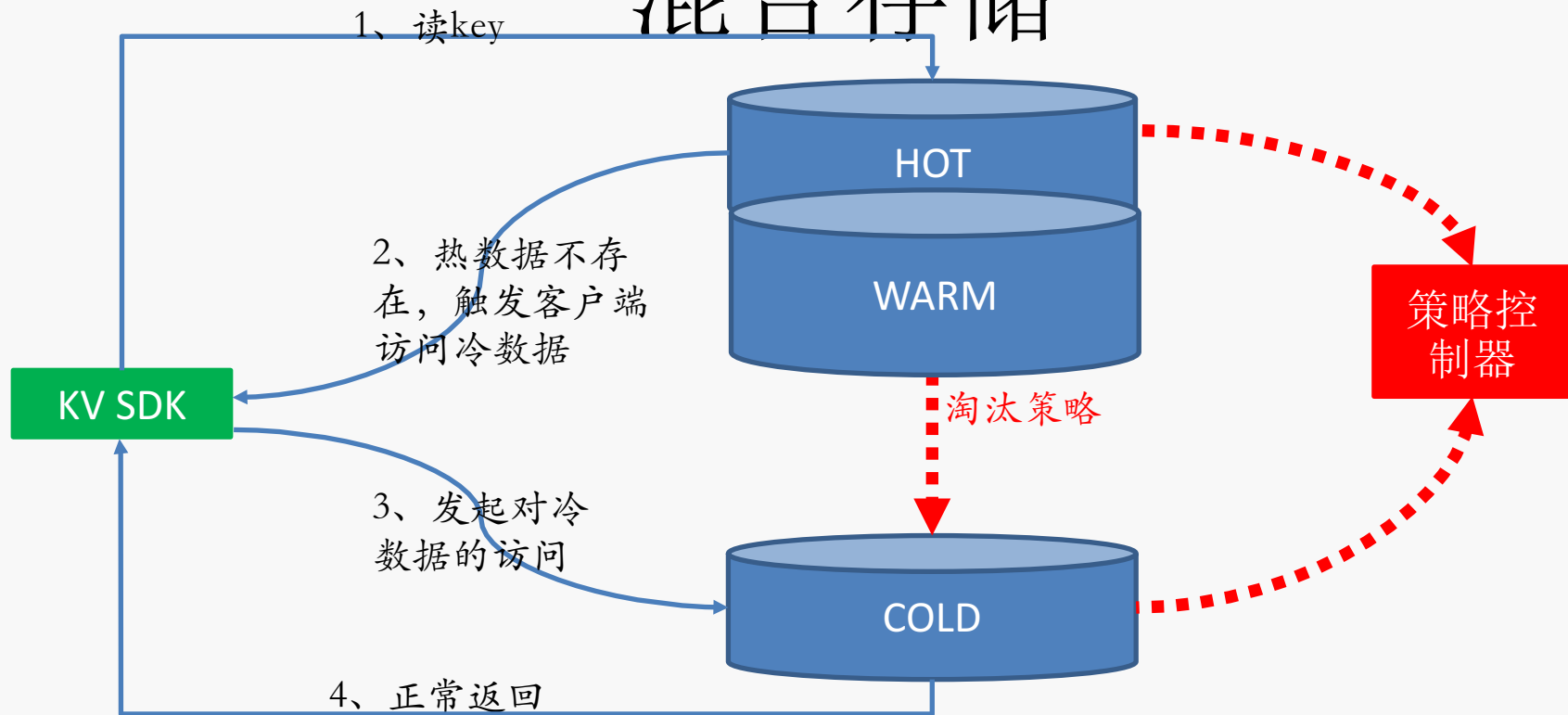
冷热分离



十年架构 成长之路



混合存储



十年架构 成长之路



THANKS