



DTCC

2016中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2016

数据定义未来

SequeMedia
盛拓传媒

IT168.com

ChinaUnix

ITPUB

Baidu Distriected Redis Platform

百度分布式Redis平台

百度DBA架构师 张东阳



DTCC

2016年中国数据库技术大会
DATABASE TECHNOLOGY CONFERENCE CHINA 2015

SequeMedia
数据库

IT168

ChinaUnix

ITPUB

自我介绍

- 张东阳 2010年硕士毕业于中科院计算所
- 百度Redis方向技术负责人
- 专注于底层技术（FS、CACHE、Linux kernel）



大纲

- BDRP在百度的应用
- 架构介绍
- 挑战与解决方案
- 总结



大纲

- **BDRP**在百度的应用
- 架构介绍
- 挑战与解决方案
- 总结



BDRP在百度的应用

 Baidu 推广 百度凤巢	 度秘 提供秘书化 搜索服务	 百度外卖 在线外卖订餐产品
 Baidu 推广 网盟推广	 直达号	 百度糯米 省钱更省心，全场随便退！
 Baidu 信誉		 百度钱包 随身随付有 优惠的钱包
 百度私有云 bpc.baidu.com	 百度开放云	
Redis 实例 5w+, 服务数2k+, 物理机5k台+		



典型业务

- 凤巢（商户物料信息）
- 诚信（首页加V信息）
- 糯米（促销信息）
- 度秘（用户聊天信息）
- 百度钱包（红包库存信息）



大纲

- BDRP在百度的应用
- 架构介绍
- 挑战与解决方案
- 总结

系统框架

对外接口

名字服务

客户端

Dash Board

基础组件

Proxy

Redis

元数据管理

集群管理

上线部署

容量管理

集群信息管理

权限管理

版本管理/升级

虚拟化

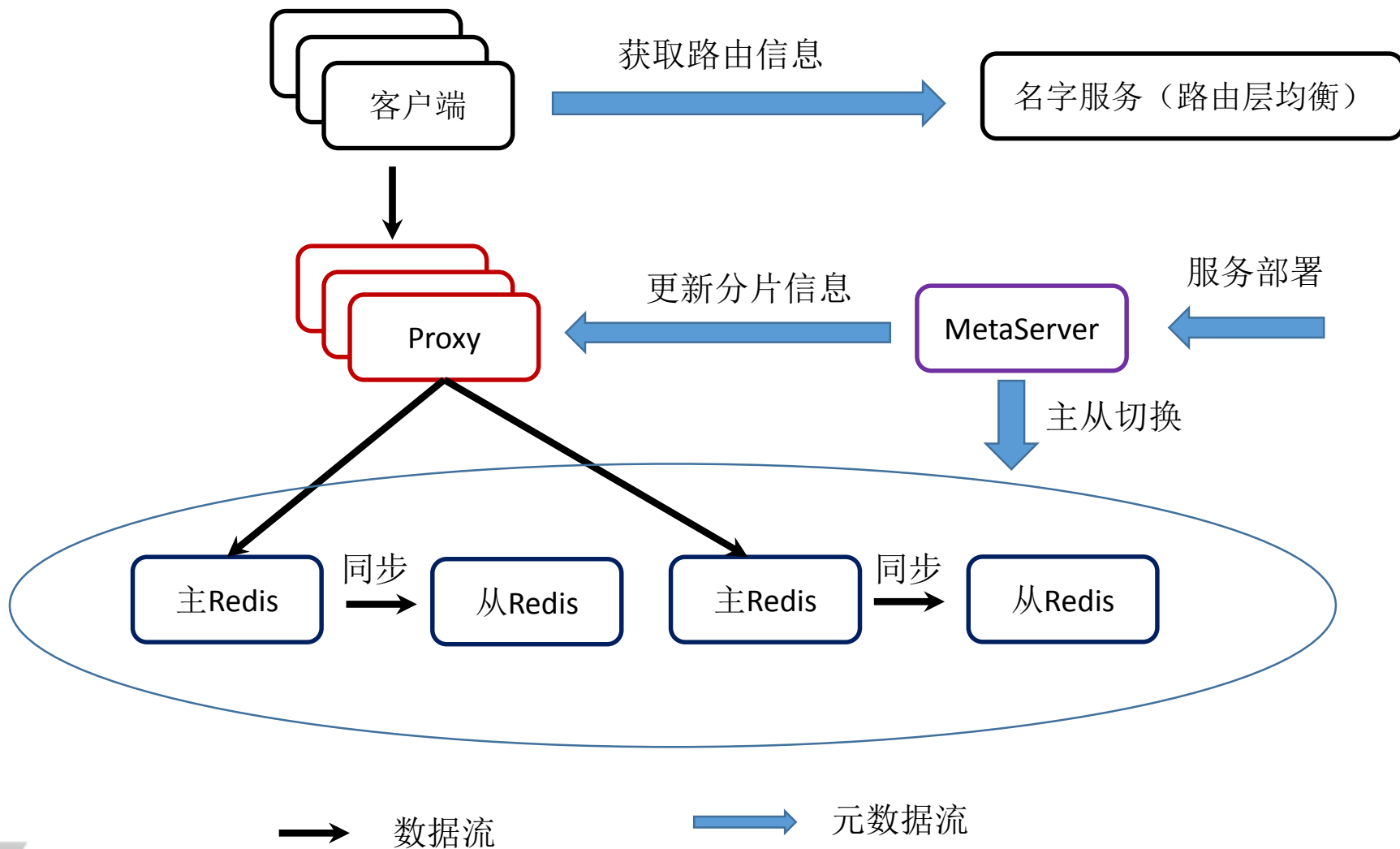
Container

Container

Container



数据流图

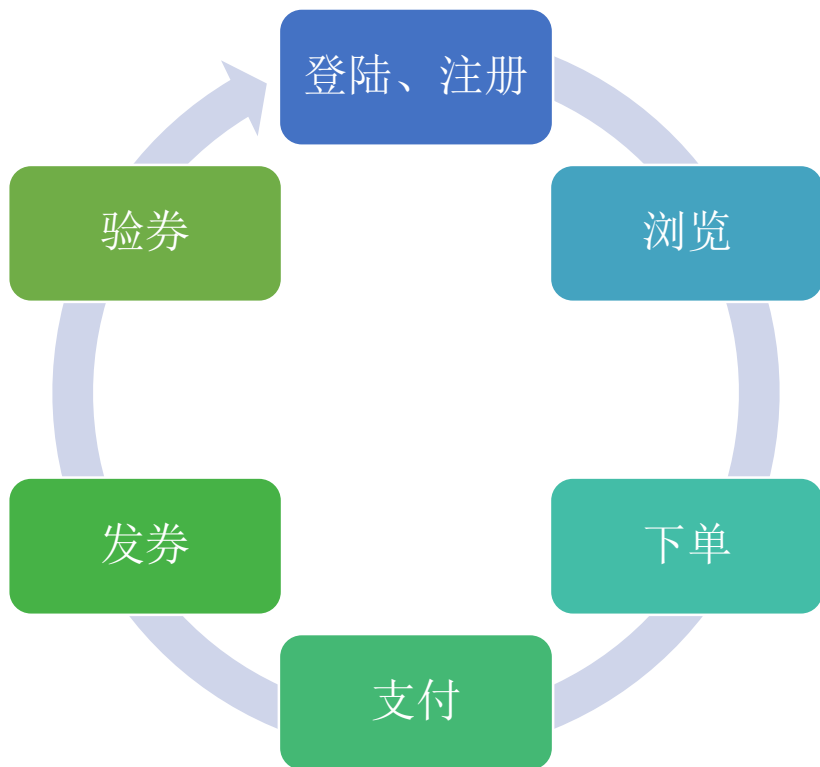


大纲

- BDRP在百度的应用
- 架构介绍
- 挑战与解决方案
- 总结



挑战-以糯米为例

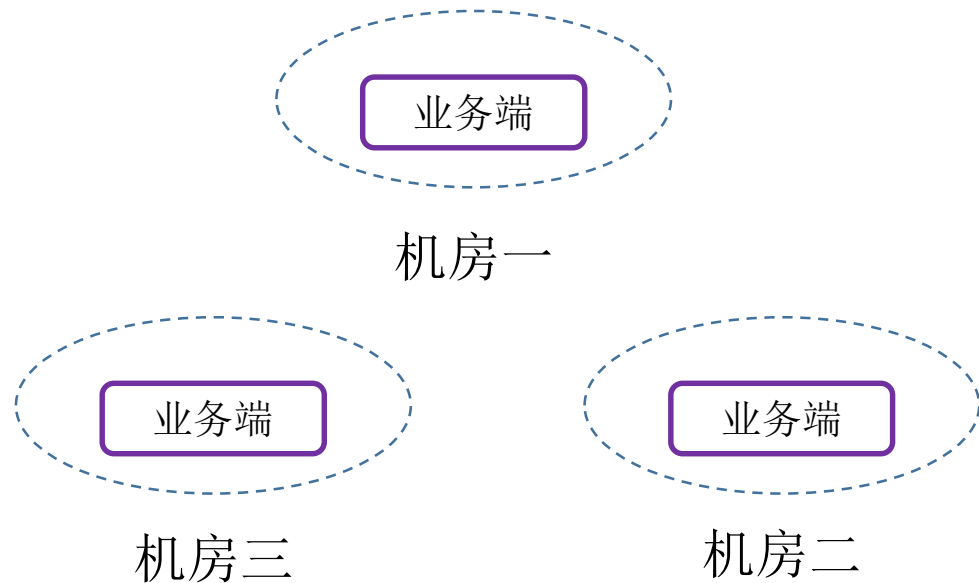


- 全交易闭环重度依赖，如登录90%以上流量由Redis承担
 - 可用性（机器、机房）
- 访问量巨大，大促时促销服务一天访问量数千亿
 - 性能（时延、吞吐）
 - 扩展性
- 存储（MySQL、Redis）资源使用巨大，数千台物理机
 - 存储使用配比

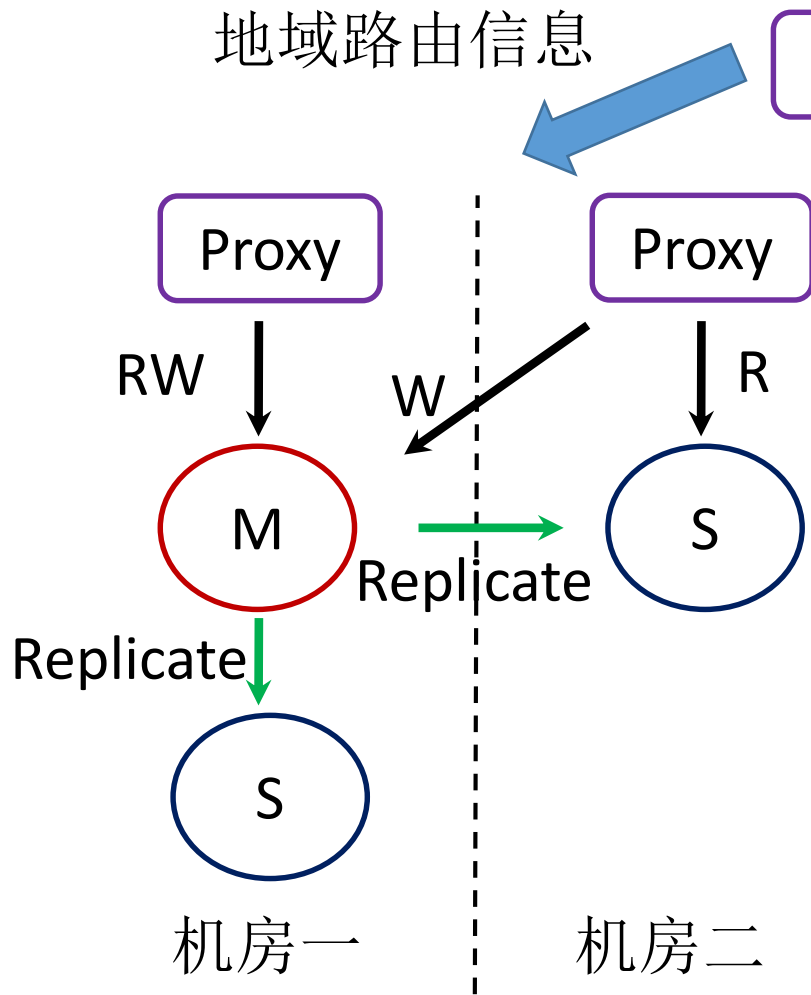


挑战-性能

- 近地域多机房（十几ms网络时延）
 - 一次业务访问几十次读取Redis（读时延放大）
 - 读写比高（> 100: 1）
 - 读吞吐高（单集群数百万QPS）



近地域多机房方案

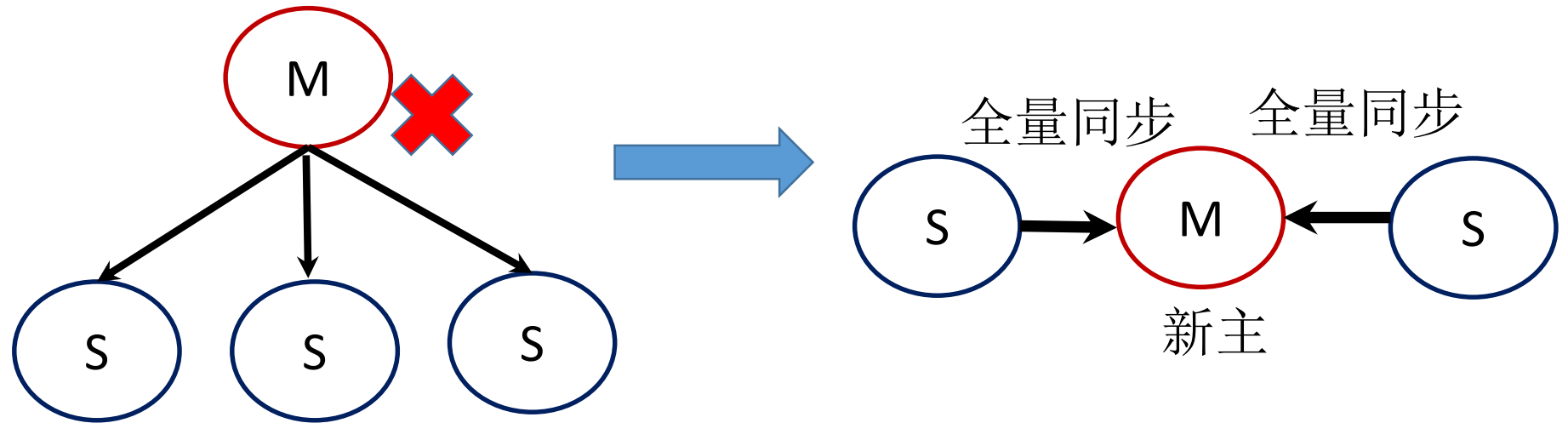


- Proxy读写分离
- MetaServer传播地域路由信息

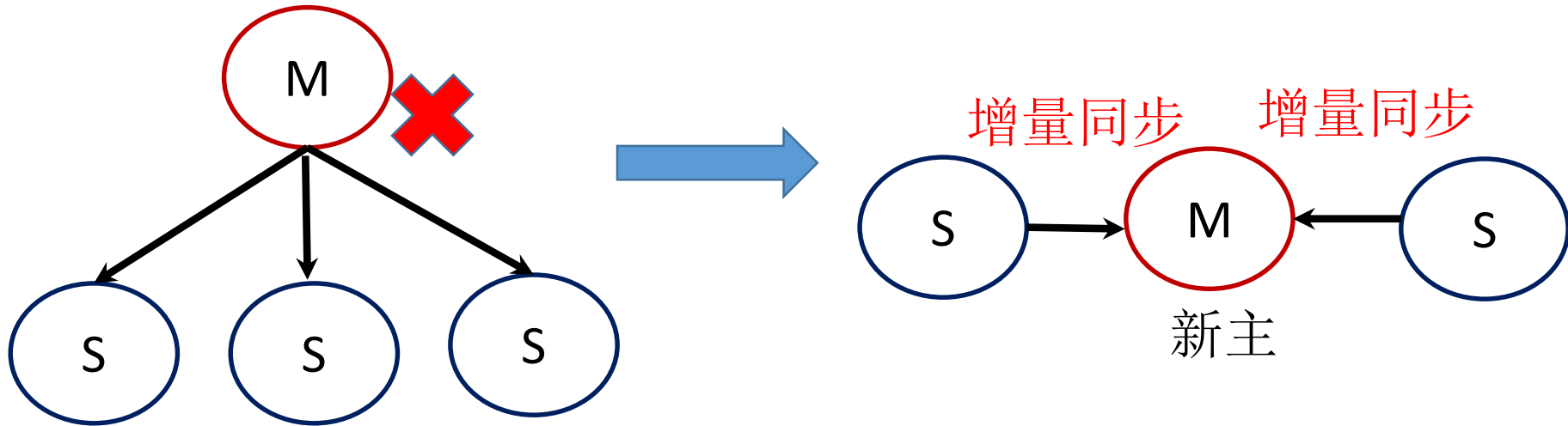


挑战-可用性

- 社区版切主时从库向新主全量同步
 - 主库网卡压力陡增 - 服务不可写
 - 从库同步时间几十分钟 - 服务不可读



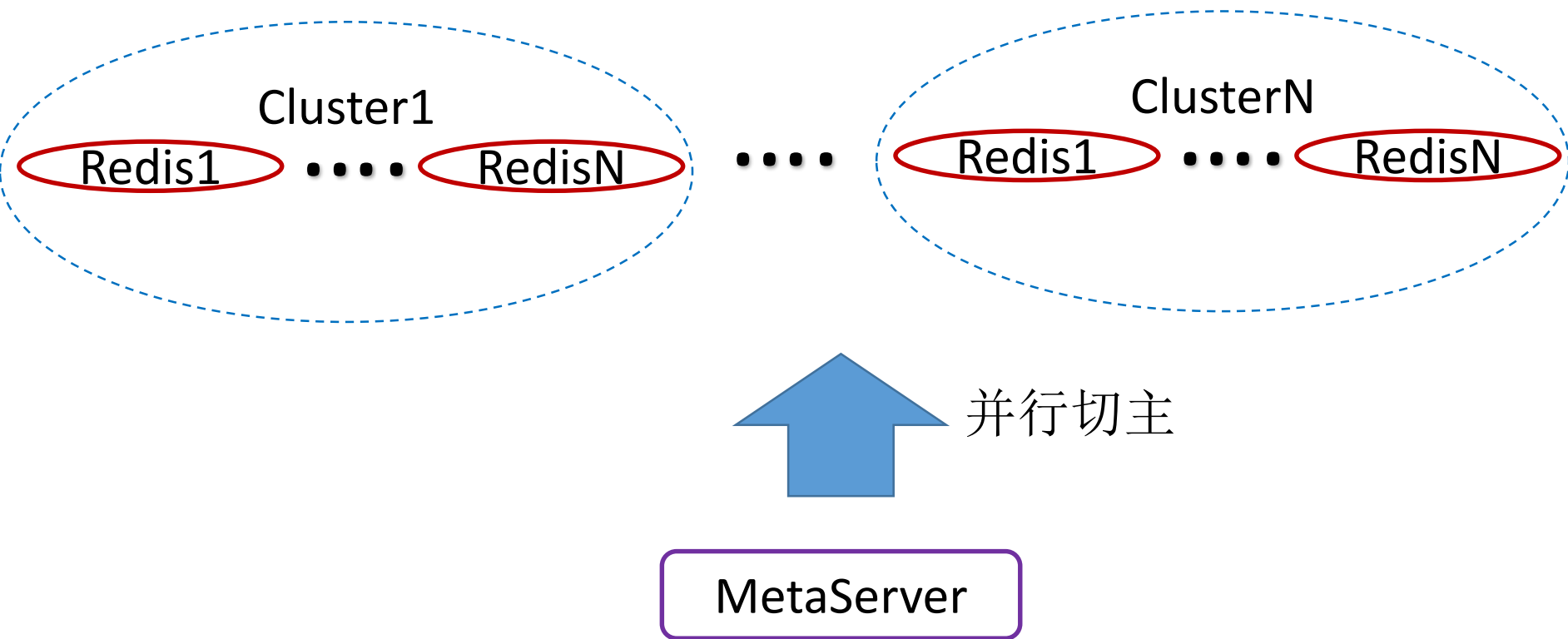
机器故障秒级切主



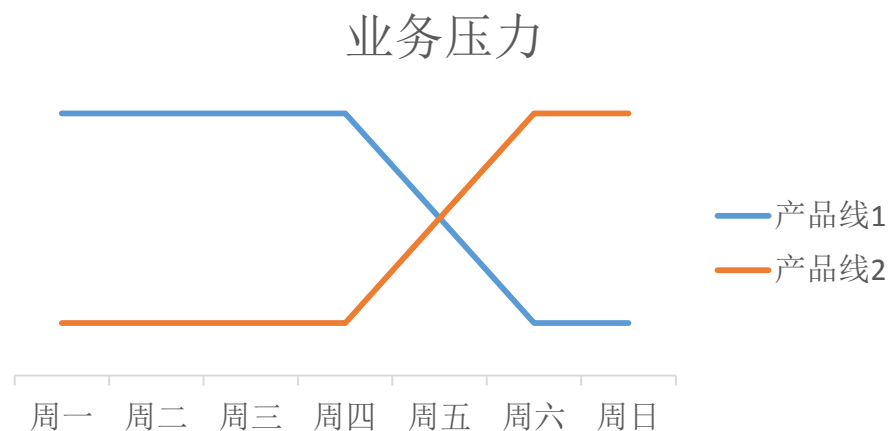
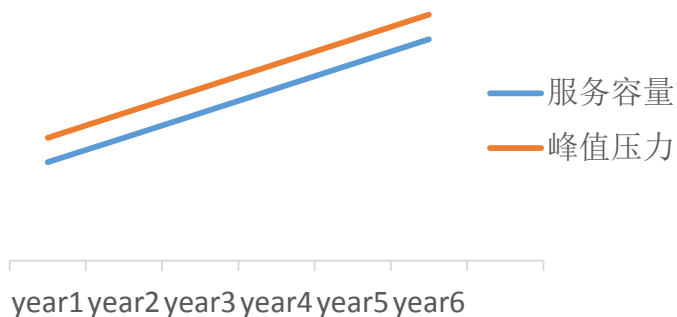
- 同源增量同步
 - 网卡压力**突增**->**平稳**
 - 10G内存同步时间**10min**->**5s**



1分钟整机房切换



挑战-扩展性

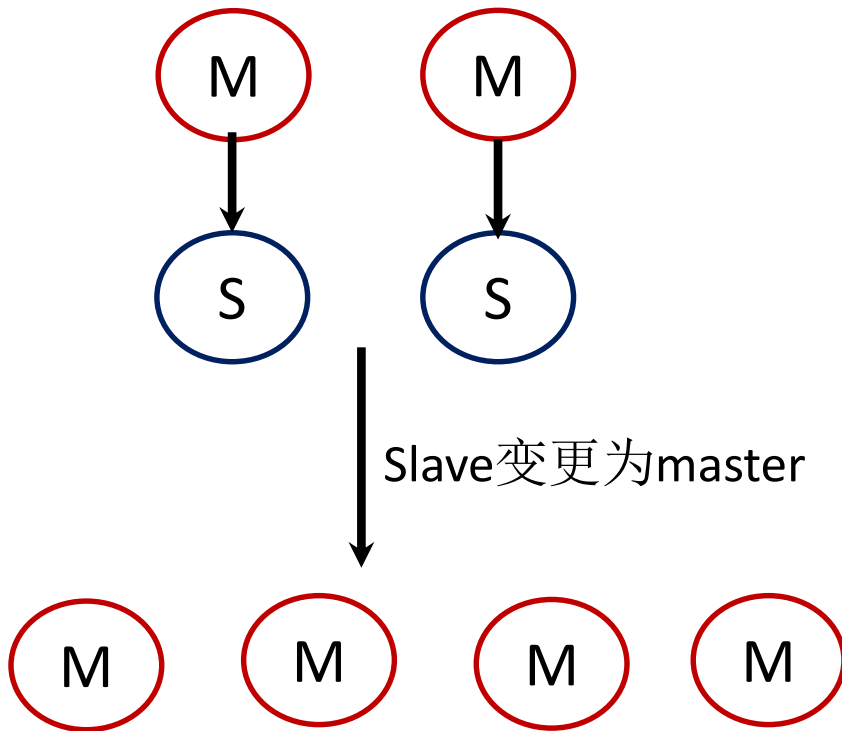


大促压力高于服务容量，频繁扩缩容->弹性扩缩容
不同业务不同时间段压力不同->服务混部

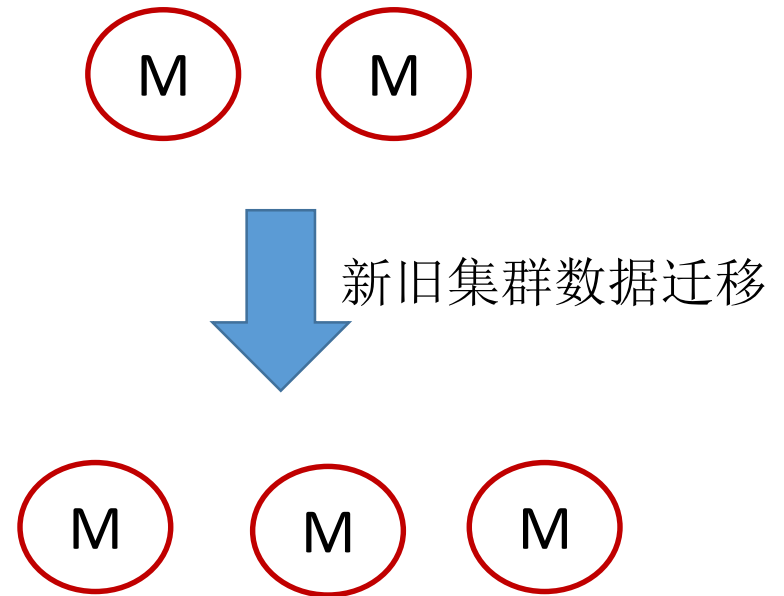


扩展性-redis V2.x 方案

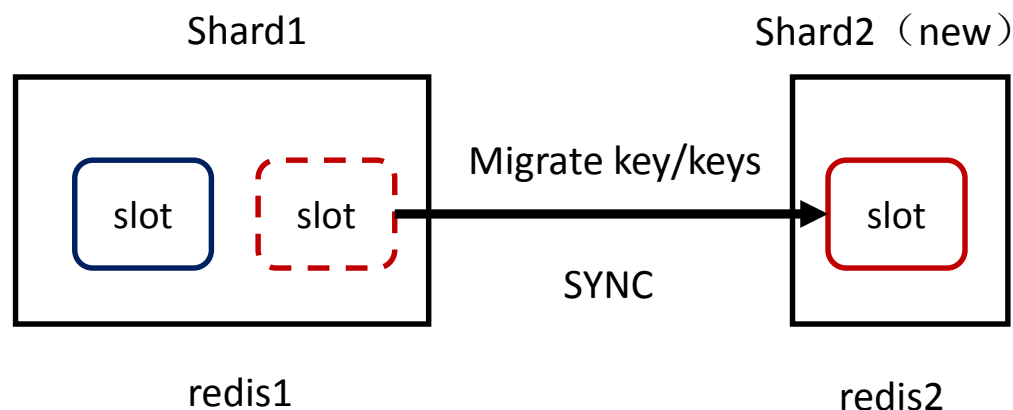
- 成倍扩容 (2->4)



- 服务迁移 (2->3)



弹性扩缩容-Redis Cluster方案

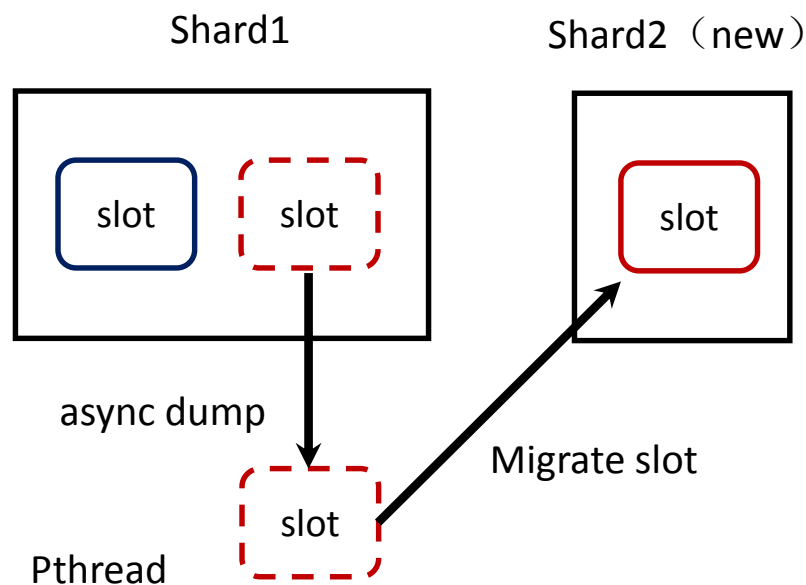


核心：数据迁移方案

- 迁移效率：单个key迁移效率低下（V3.2 keys）
- 迁移方式：同步迁移，迁移过程中影响服务读写



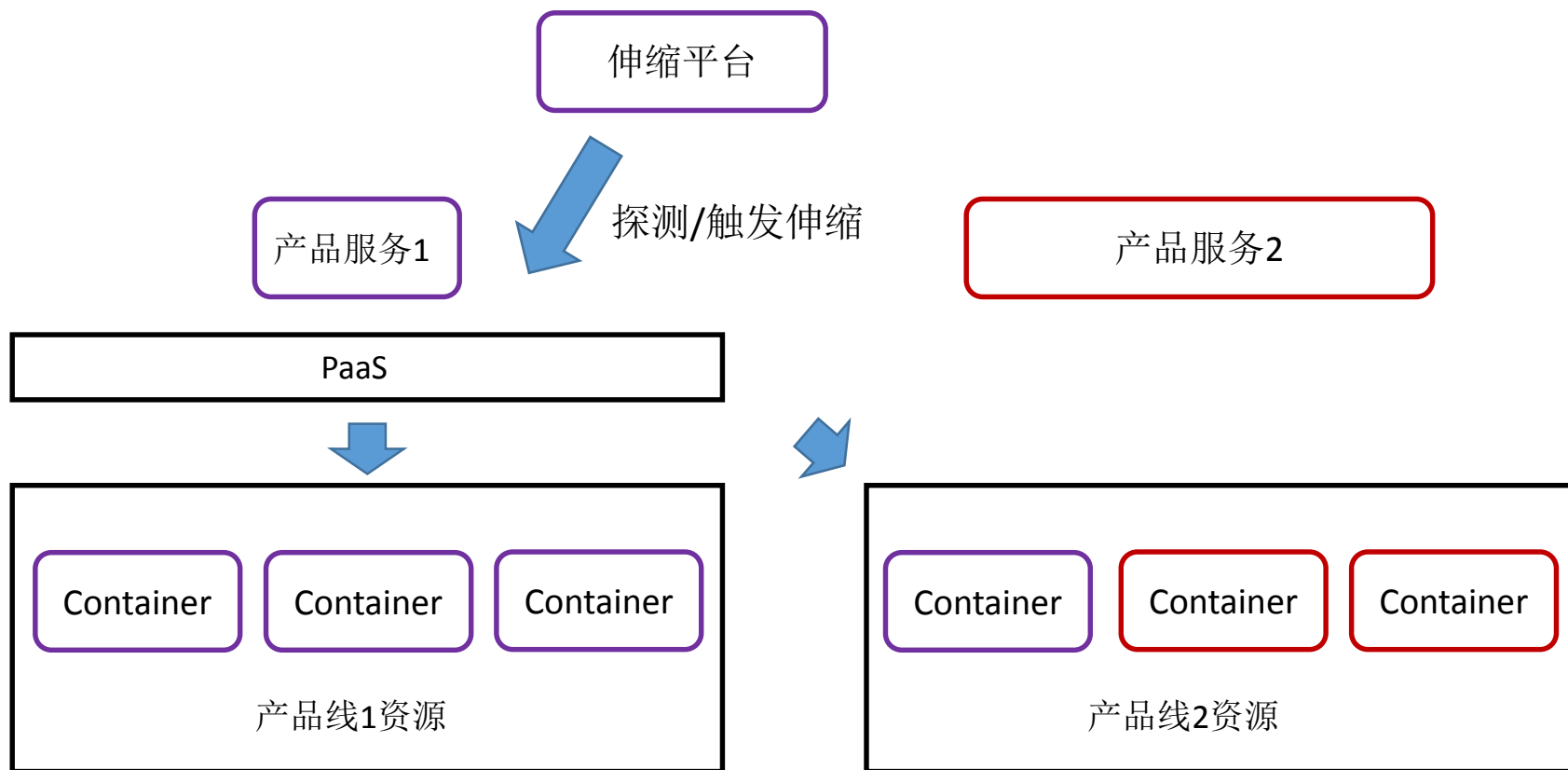
弹性扩缩容-BDRP方案



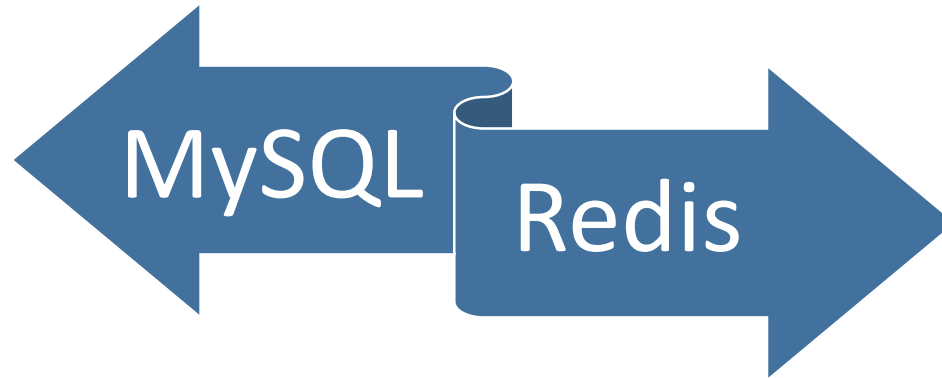
- 迁移效率：数据块整体聚合迁移，效率高
- 迁移方式：子线程异步dump和迁移数据，不影响读服务



基于服务混部的Auto Scaling



挑战-DB&CACHE使用

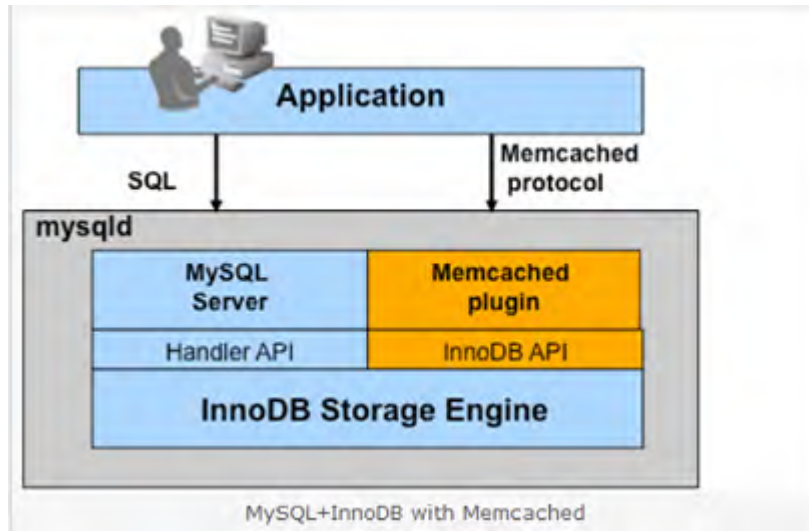


- Redis作为CACHE的场景下，业务无法判断MySQL和Redis的最佳配比
 - 资源
 - 性能
- 同时使用MySQL协议和Redis协议
- 维护MySQL和Redis数据一致性



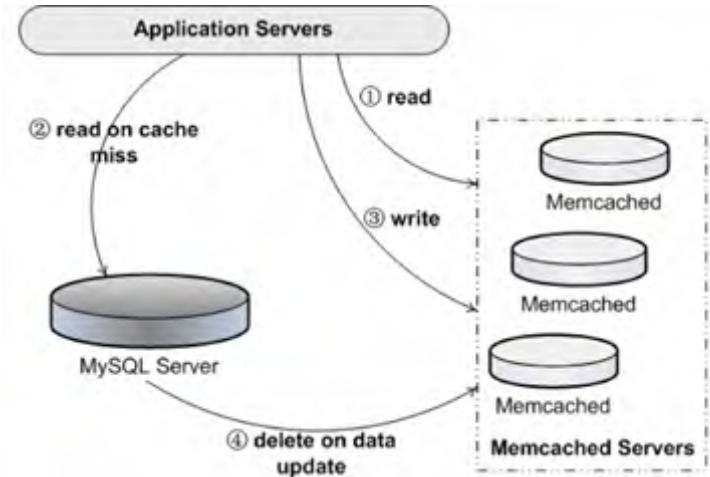
DB&CACHE-已有方案

InnoDB with Memcached



- 协议未统一
- 受限单机内存

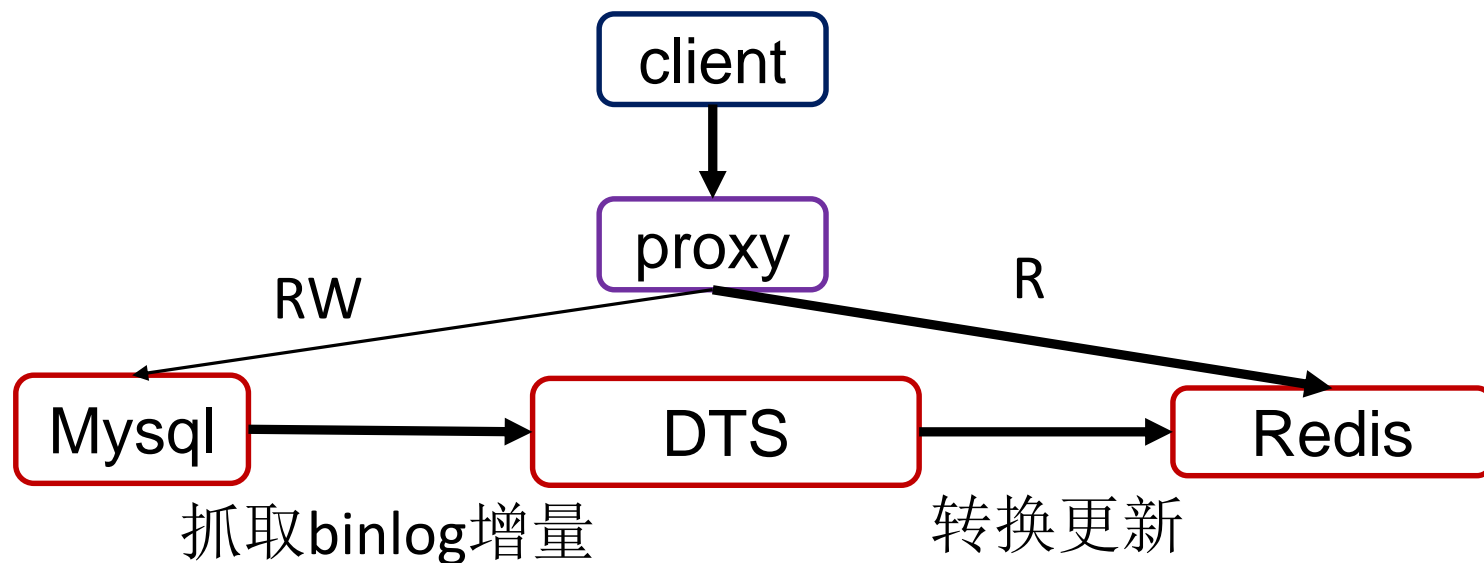
MySQL Memcached UDFs



- 协议未统一
- 需要用户处理数据一致性



MySQL&Redis融合



- Proxy: 分发不同存储
- DTS:完成MySQL->redis数据同步
- 访问协议为MySQL协议



MySQL&Redis融合-数据转化

- 库名:DB
- 表名: TABLE
- 表主键:
PRIMARY_KEY

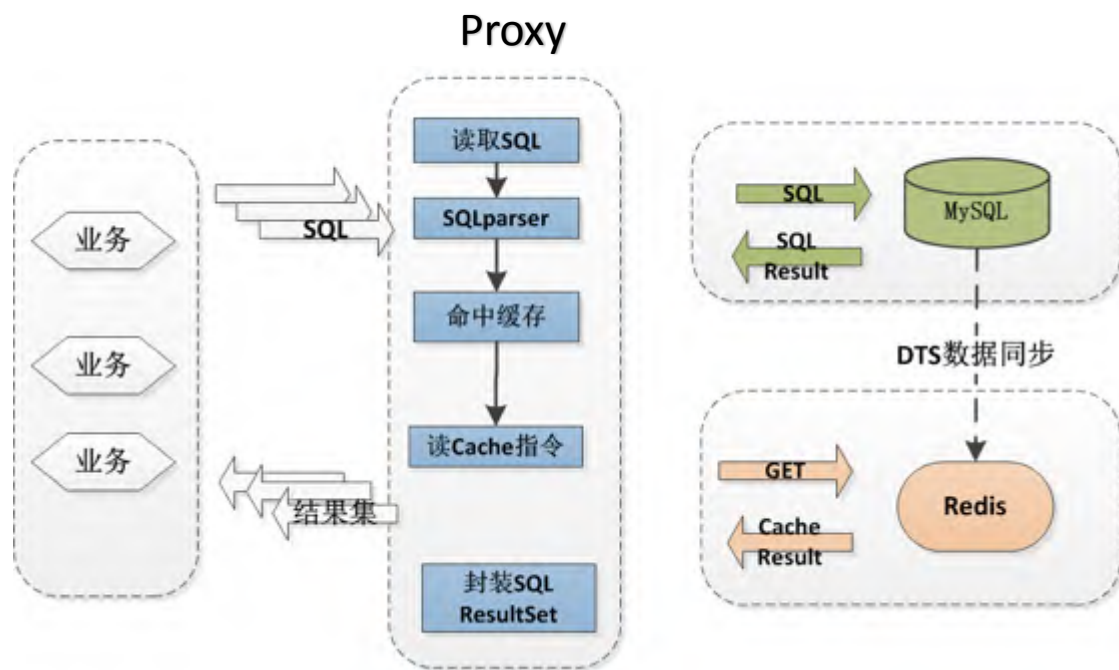


- key:DB.TABLE_11_PRIMARY_KEY
- value:PRIMARY_KEY
对应的组合内容



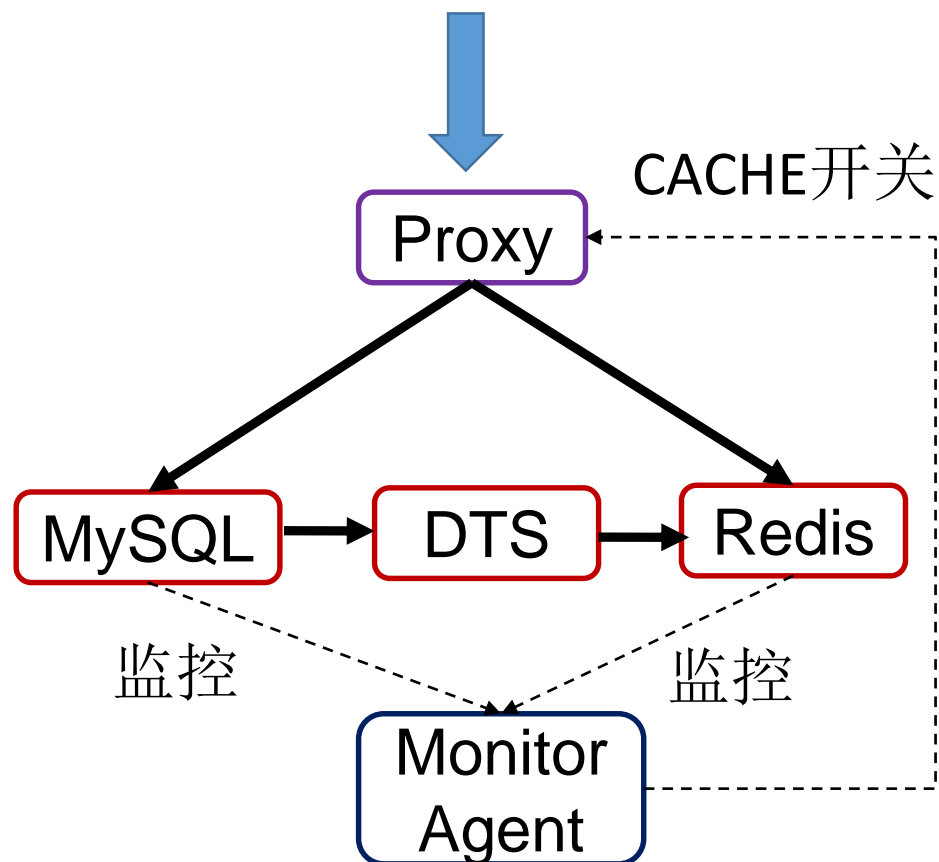
Mysql&Redis融合-查询逻辑

- Proxy自带SQL Parser
- 写入进入主库
- 非KV类查询进入从库
- KV类查询进入Redis
- 结果集为MySQL协议



MySQL&Redis融合-一致性保证

- 填充方案
 - 全量CACHE: DTS
 - 热点填充: Proxy + DTS
- 同步监控
 - Redis延时过大摘除Redis



Mysql&Redis融合-指标和收益

- Redis更新时延ms级
- 可靠性(99.99%)
- 对比纯MySQL集群收益（Redis命中率90%）
 - 集群吞吐提升10倍
 - 长尾降低70%，平响降低20%
 - 业务无需任何改动



大纲

- BDRP在百度的应用
- 架构介绍
- 挑战与解决方案
- 总结



总结

- bdrp应用情况
- 架构
- 挑战与方案（糯米）
 - 近地域多机房解决方案
 - 1分钟整机房切主
 - 基于服务混部的Auto Scaling
 - MySQL&Redis融合方案



其他工作

- 远地域多机房服务
- 基于磁盘引擎的KV存储
- SCS&RDS融合方案
- 运维管理平台



百度DBA诚招天下英豪

- MySQL方向研发和运维
- Redis方向研发和运维
- 邮箱: dba@baidu.com
- 微信/QQ: 174264744





THANKS TO YOU

SequeMedia
縱拓傳媒

IT168.com

ChinaUnix

ITPUB