新浪微博
weibo.com

微博REDIS定制化之Tribe系统介绍

微博研发中心 张冬洪

2016.11

Agenda

- 个人介绍
- 微博REDIS的使用介绍
- Tribe系统设计的考量
- Tribe系统性能测试对比
- Tribe系统运维点滴

个人介绍

新浪微博 weibo.com

- Redis中国用户组发起人
- 微博研发中心DBA，专注于MySQL和NoSQL架构设计与运维以及自动化平台的开发；2016年3月加入微博，目前在微博主要负责Feed系统相关业务的数据库运维和业务保障工作

张冬洪 Lv.14

419	154	86
关注	粉丝	微博

redis 中国用户组

Redis2016 Lv.4

3	435	15
关注	粉丝	微博

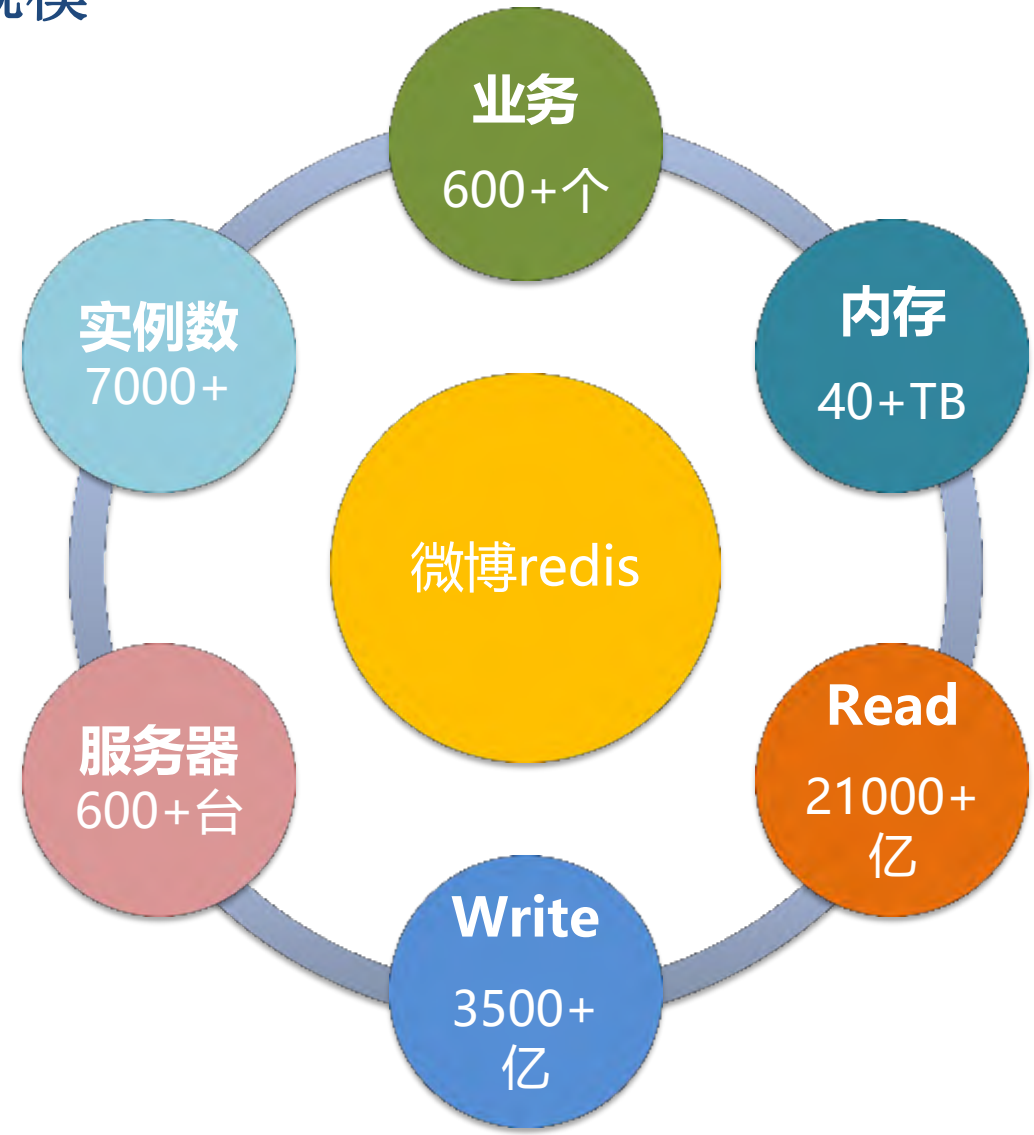
Agenda

- 个人介绍
- 微博REDIS的使用介绍
- Tribe系统设计的考量
- Tribe系统性能测试对比
- Tribe系统运维点滴

微博REDIS的使用介绍

新浪微博 weibo.com

- 使用规模



微博REDIS的使用介绍

新浪微博 weibo.com

定制化演进

版本定制化

- 2010年引入2.0 ~ 2016年定制版2.4.12
- 无阻塞落地，实现数据持久化
- 借鉴MySQL复制机制，实现redis主从同步，支持增量复制
- 在线热升级、config dump、主从延迟检查、危险命令认证等

业务定制化

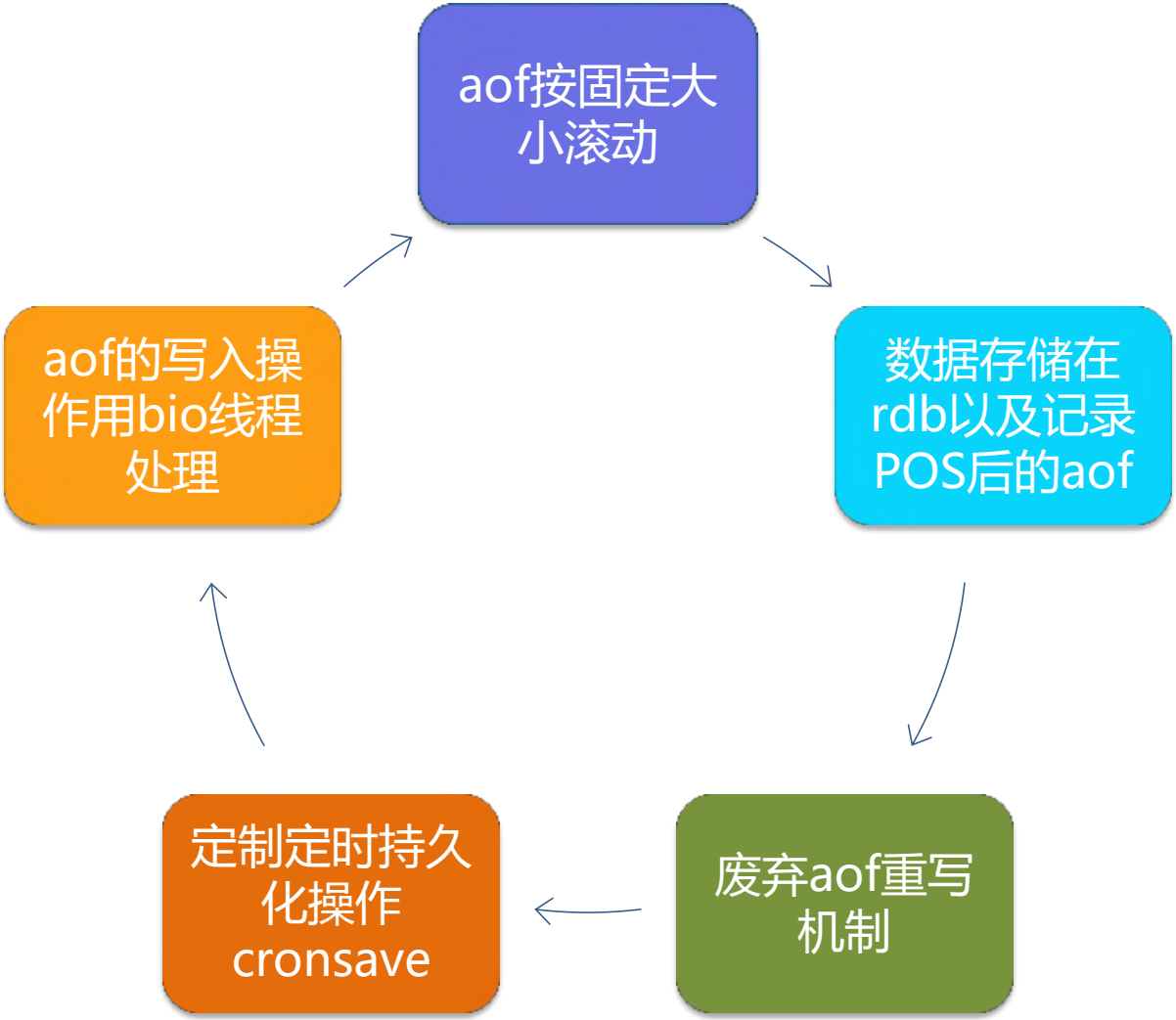
- 关系类，定制longset数据结构
- 计数类，rediscounter服务、CounterService服务、SSDCache服务
- 判断类，BloomFilter服务
- 聚合类，VerctorService服务

服务定制化

- 主从结构场景下redis HA高可用实现，成功率99%
- 实例容量检测与自动扩容实现
- 异构服务间，数据平滑迁移redisTransfer
- CacheService服务框架实现，Tribe系统

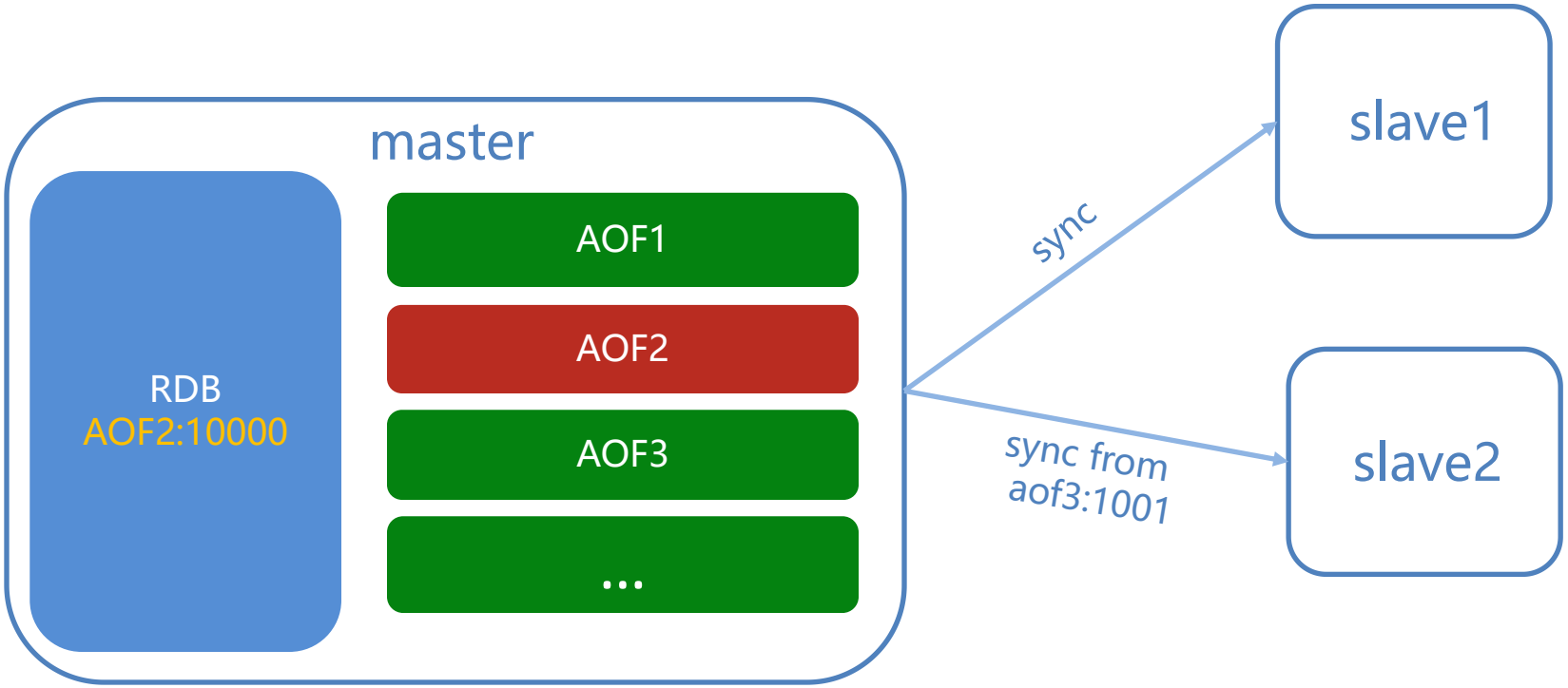
微博REDIS的使用介绍

- 无阻塞落地、数据持久化



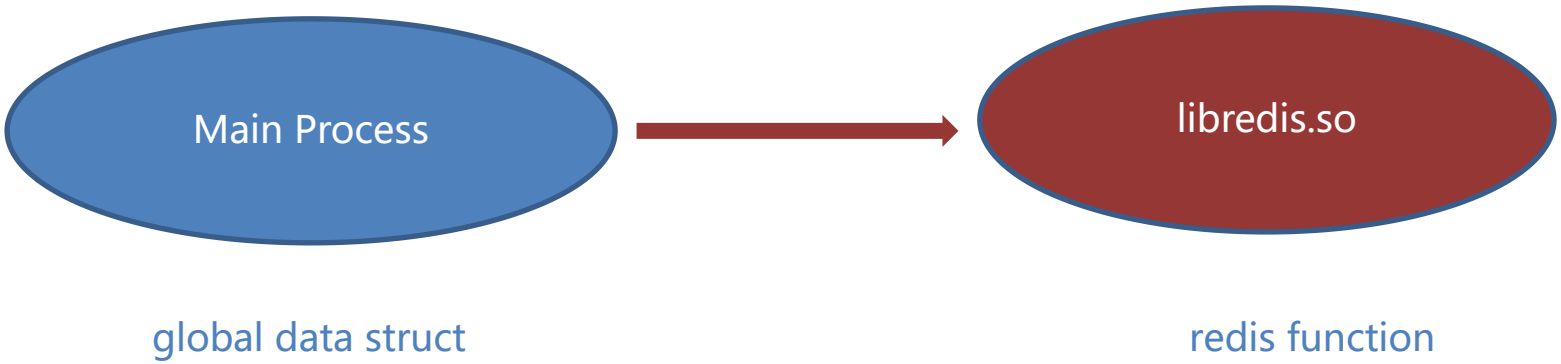
❖ 主从增量复制的实现

- 使用rdb+aof+aof position方式实现
- 主从只需进行一次全量复制，其余增量，如：主从断开、切主等



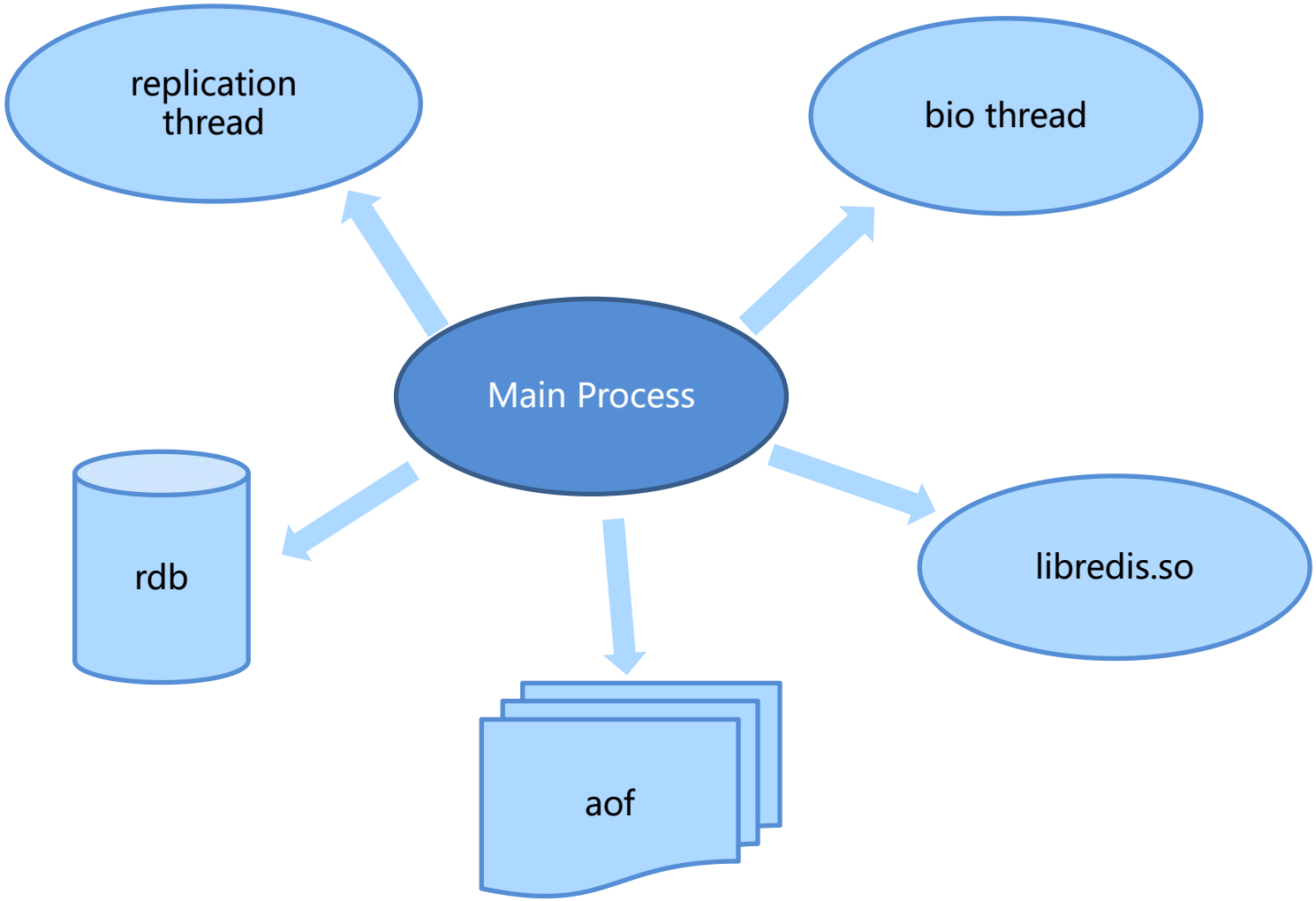
在线热升级

- 主程序存全局数据结构；内存中的数据保存在全局变量里
- 其他核心处理逻辑封装到动态库
- 外部程序通过调用动态库里相应函数来读写数据
- 升级只需替换动态库，无须重新载入数据
- 毫秒级完成升级，无业务影响



微博REDIS的使用介绍

- 定制化版本总结



Agenda

- 个人介绍
- 微博REDIS的使用介绍
- Tribe系统设计的考量
- Tribe系统性能测试对比
- Tribe系统运维点滴

Tribe系统设计的考量

当前结构及问题

业务端访问逻辑较重，每个业务需要实现一套

1. 数据迁移通过脚本完成，繁琐
2. 扩容/缩容（拆分）需要业务端更改配置，高耦合

健康检查，流量切换需要人工干预； HA已比较成熟，自动扩容系统也较完善

Tribe系统设计的考量

新浪微博 weibo.com

Tribe系统设计背景

开发痛点

1、重复开发拆库，读写分离、负载均衡、跨机房部署调配、复杂度高

2、短连接访问redis，吞吐量=1/3长连接，RT=2倍吞吐量

3、拆库扩容需要业务进行双写，工作量大

运维痛点

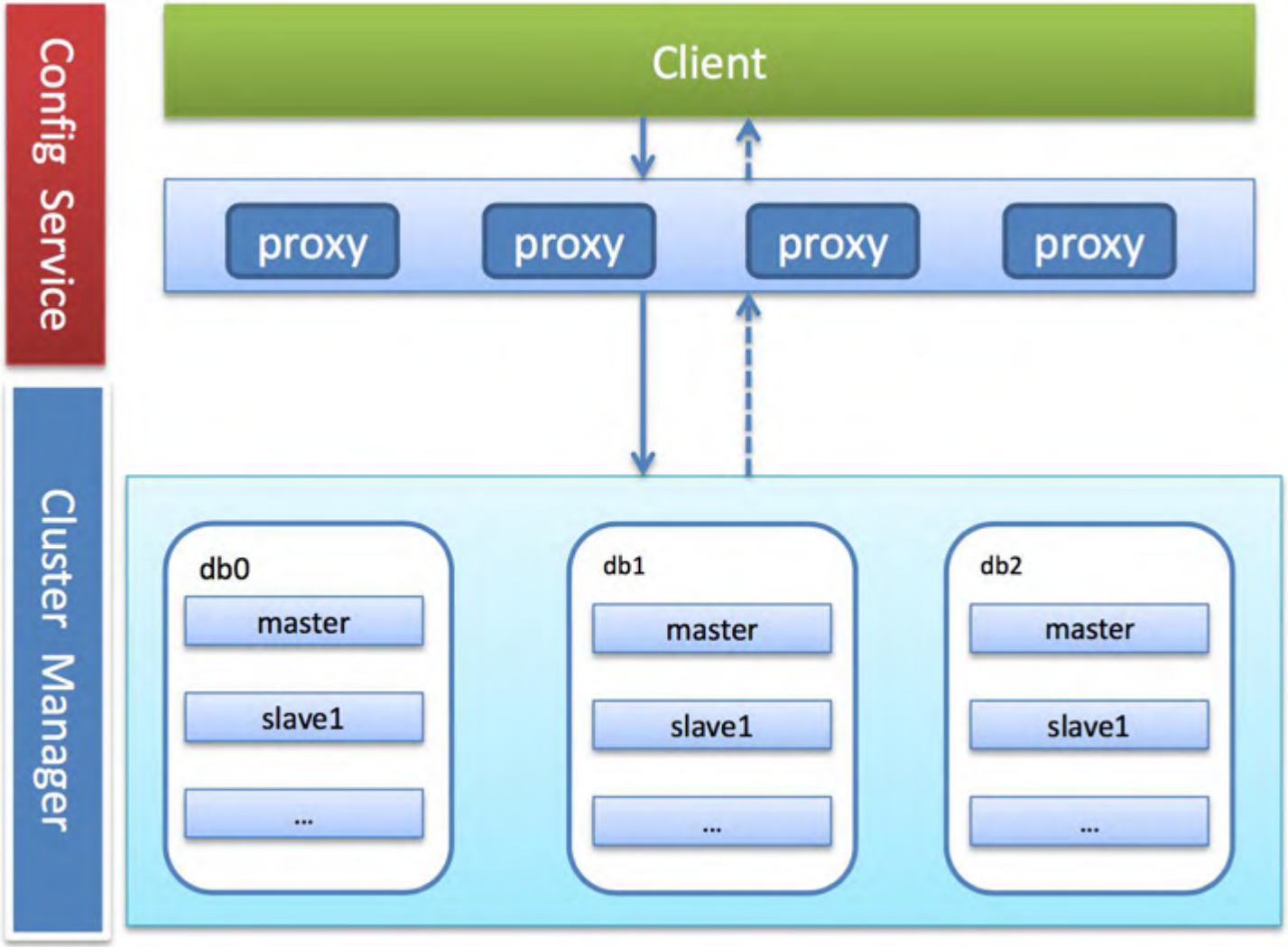
1、扩容迁移需业务配合完成，沟通成本高

2、拆分、导入、校验、双写，每一步都需人工完成，出错率大

3、拆库困难、导致单端口数据量大，难于备份和恢复

Tribe系统设计的考量

❖ Tribe系统架构



- 请求路由
- 读写分离
- 负载均衡
- 配置更新
- 数据聚集
- 动态扩容
- Failover
- 数据迁移
- 服务成本
- 集群管理

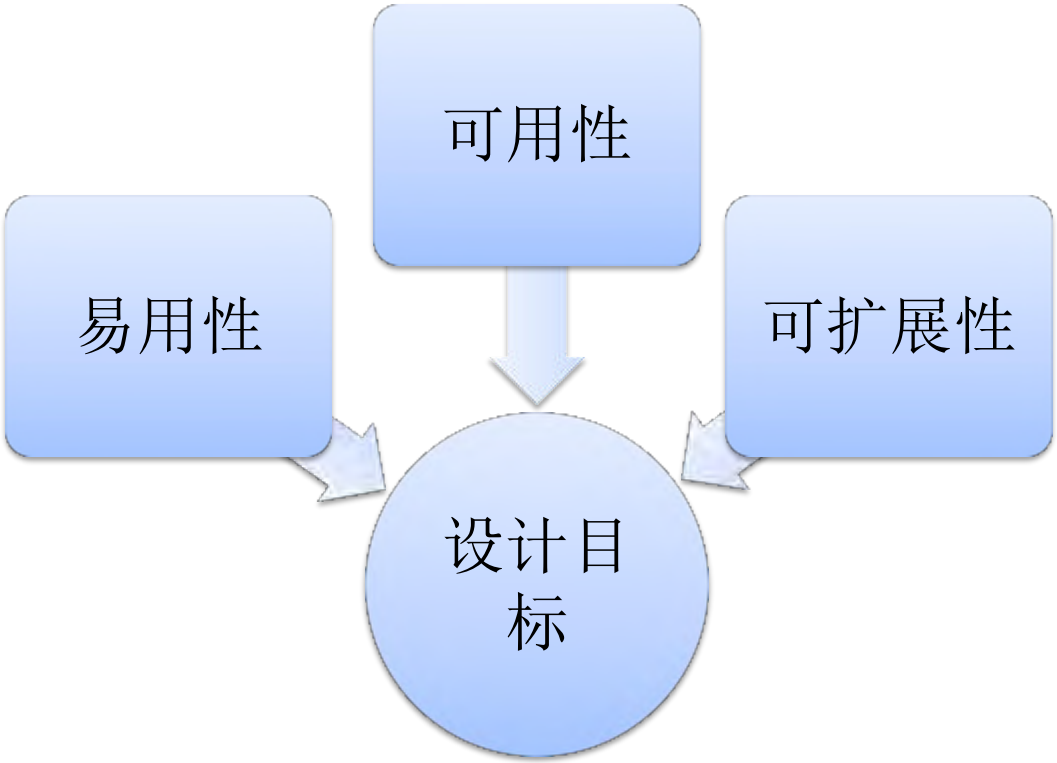
Tribe系统设计的考量

❖ Tribe系统功能

- 请求路由
  - 数据根据配置路由表，自动进行切分
  - 读写请求根据配置路由表，自动路由
- 读写分离
  - 支持配置读写分离以及只读主库
- 负载均衡
  - 支持配置机房优先策略
  - 支持round-robin负载均衡策略
- 配置更新
  - Proxy自动从configservice抓取配置、并在线更新并保存本地
- 数据聚集
  - 支持配置前缀，将相同前缀的数据落到同一个slot中
- 动态扩容
  - 支持在线拆分端口
  - 支持自动切换流量

Tribe系统设计的考量

Tribe系统设计目标



易用性

- 简化业务访问逻辑，让业务更专注代码
- 减少前后端间耦合，减少变更成本

可用性

- 健康检查
- 自动流量切换
- 自动故障恢复

可扩展性

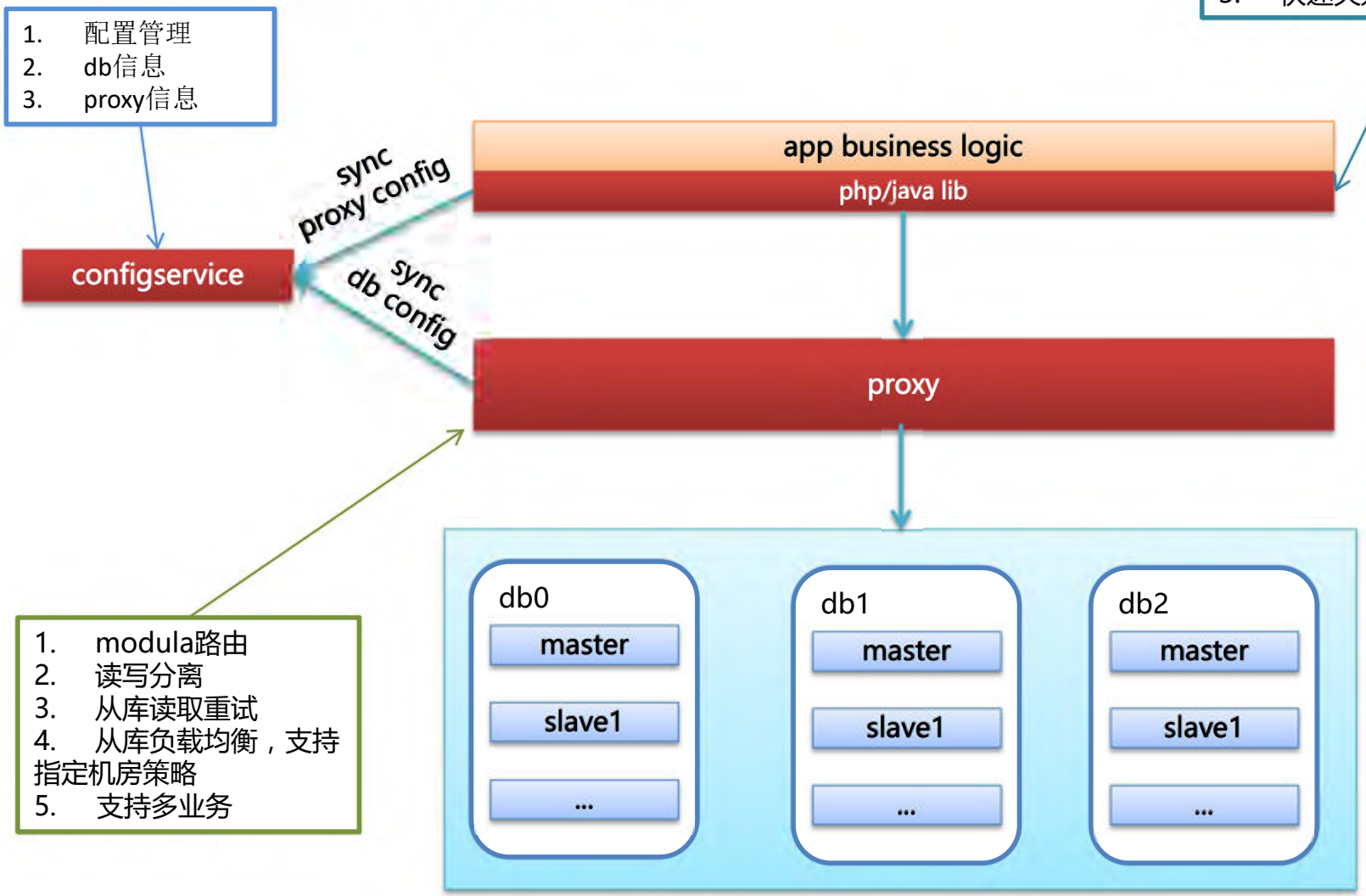
- 简单、扩展即加机器
- 扩展的策略，延迟分配
- 扩展后也能收缩

Tribe系统设计的考量

Tribe系统设计一期

1. 配置管理
2. db信息
3. proxy信息

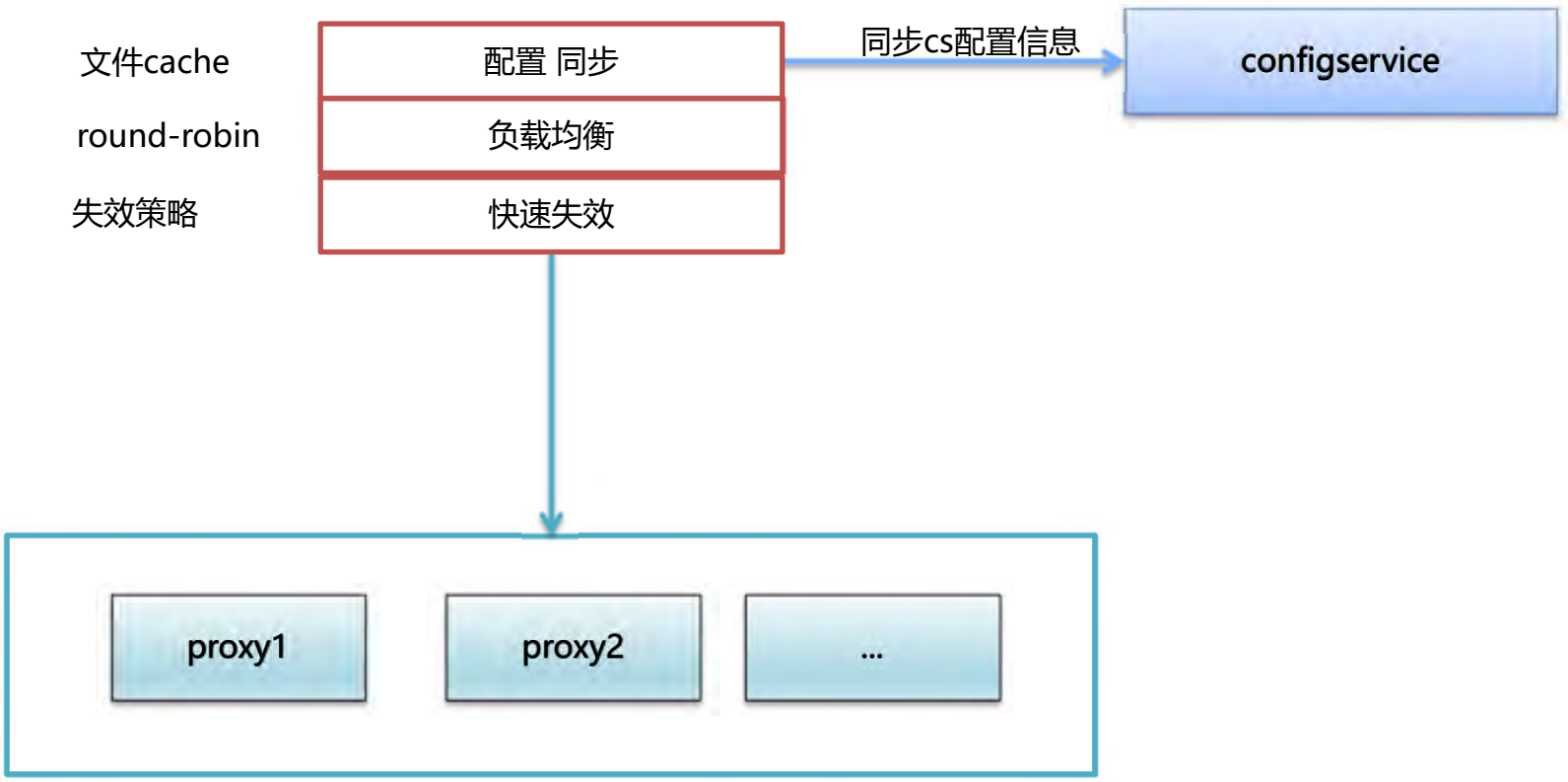
1. 配置信息同步
2. 负载均衡
3. 快速失效, 慢恢复



1. modula路由
2. 读写分离
3. 从库读取重试
4. 从库负载均衡，支持指定机房策略
5. 支持多业务

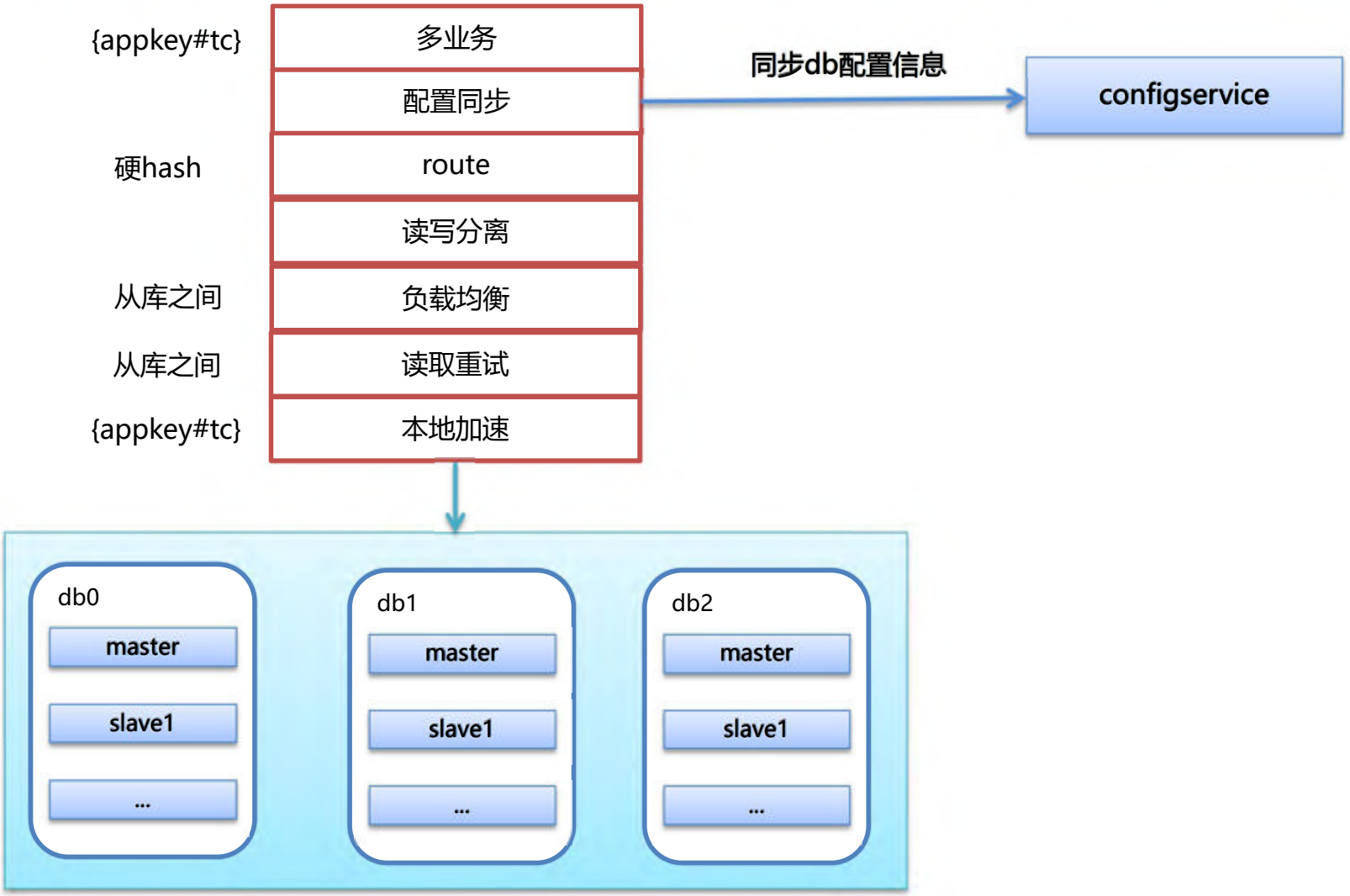
Tribe系统设计的考量

Java/php lib



Tribe系统设计的考量

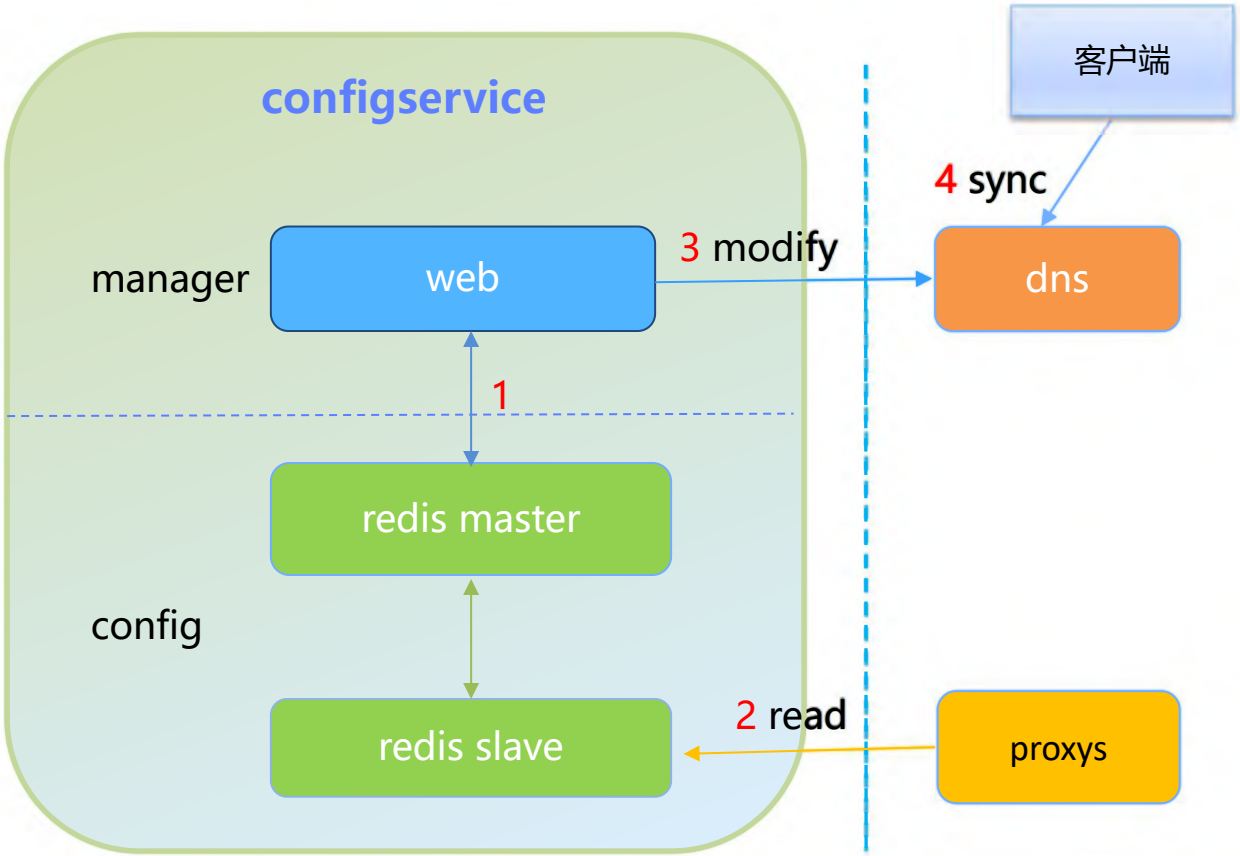
proxy功能



Tribe系统设计的考量

configService功能

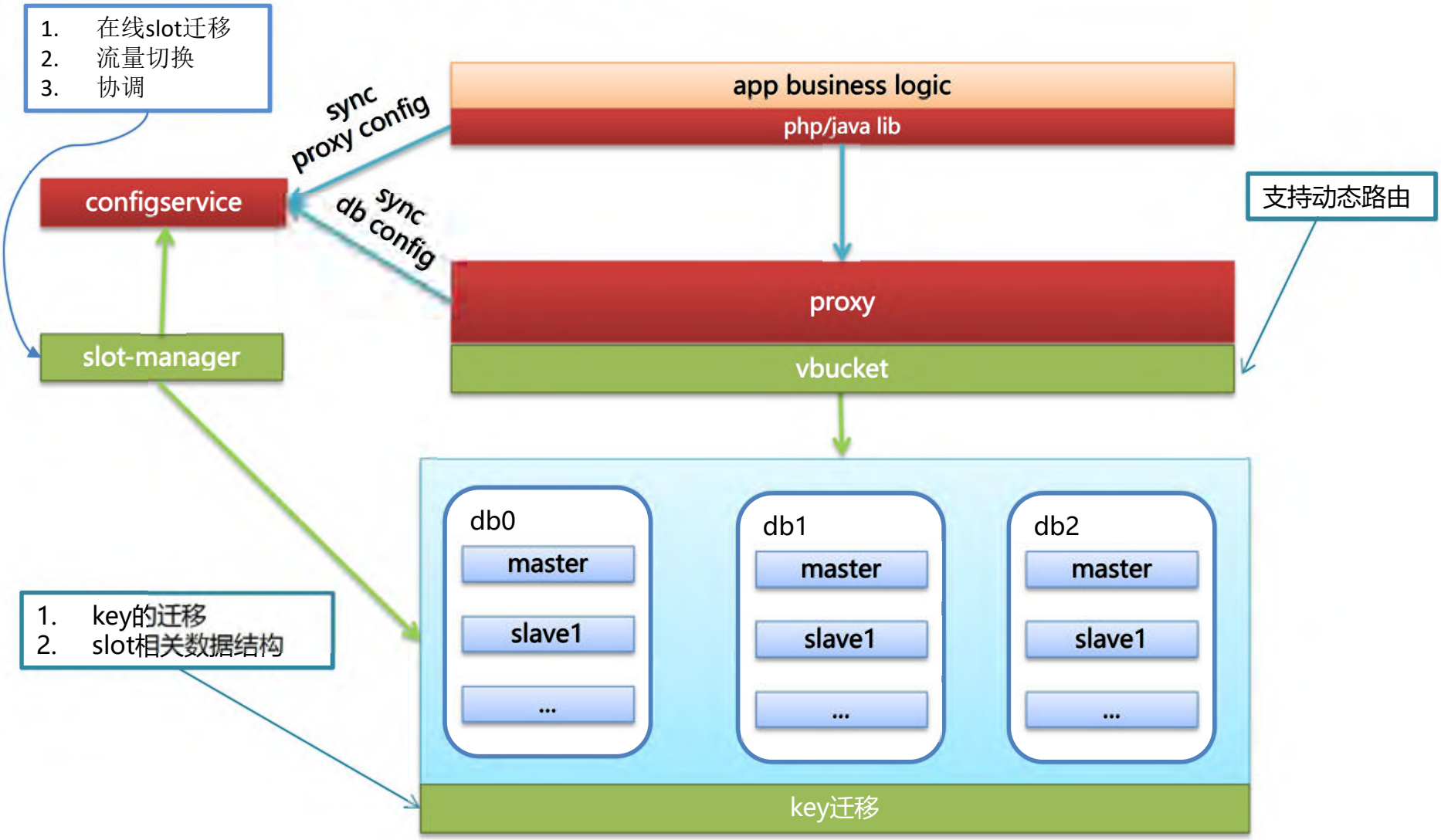
配置信息查询/修改
配置信息存储



1. 前端web将配置信息更新到redis
2. 每个proxy从redis中读取配置信息
3. web通过修改DNS来完成proxy的上下线
4. 客户端会通过DNS来sync proxy信息

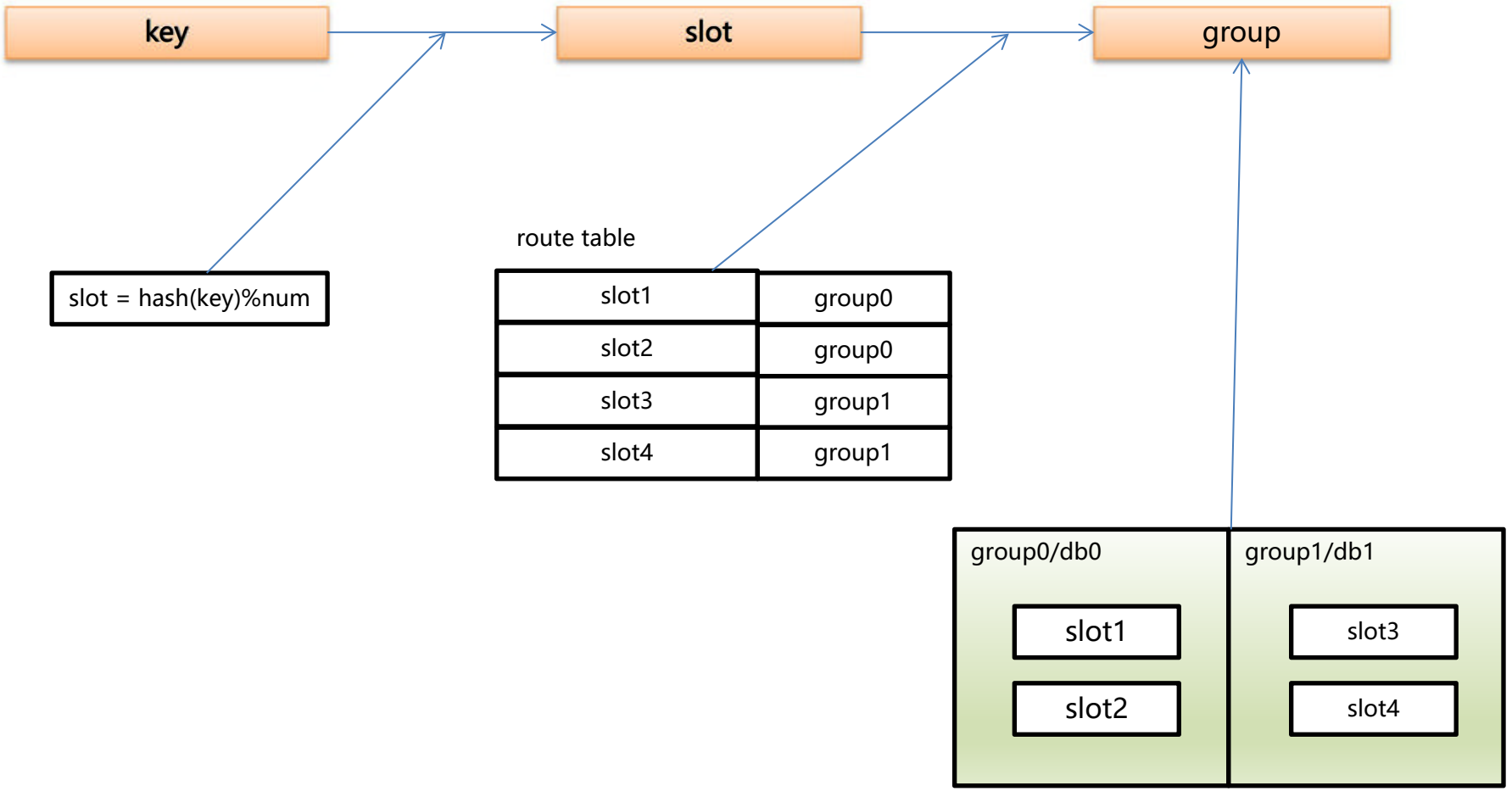
Tribe系统设计的考量

Tribe系统设计二期



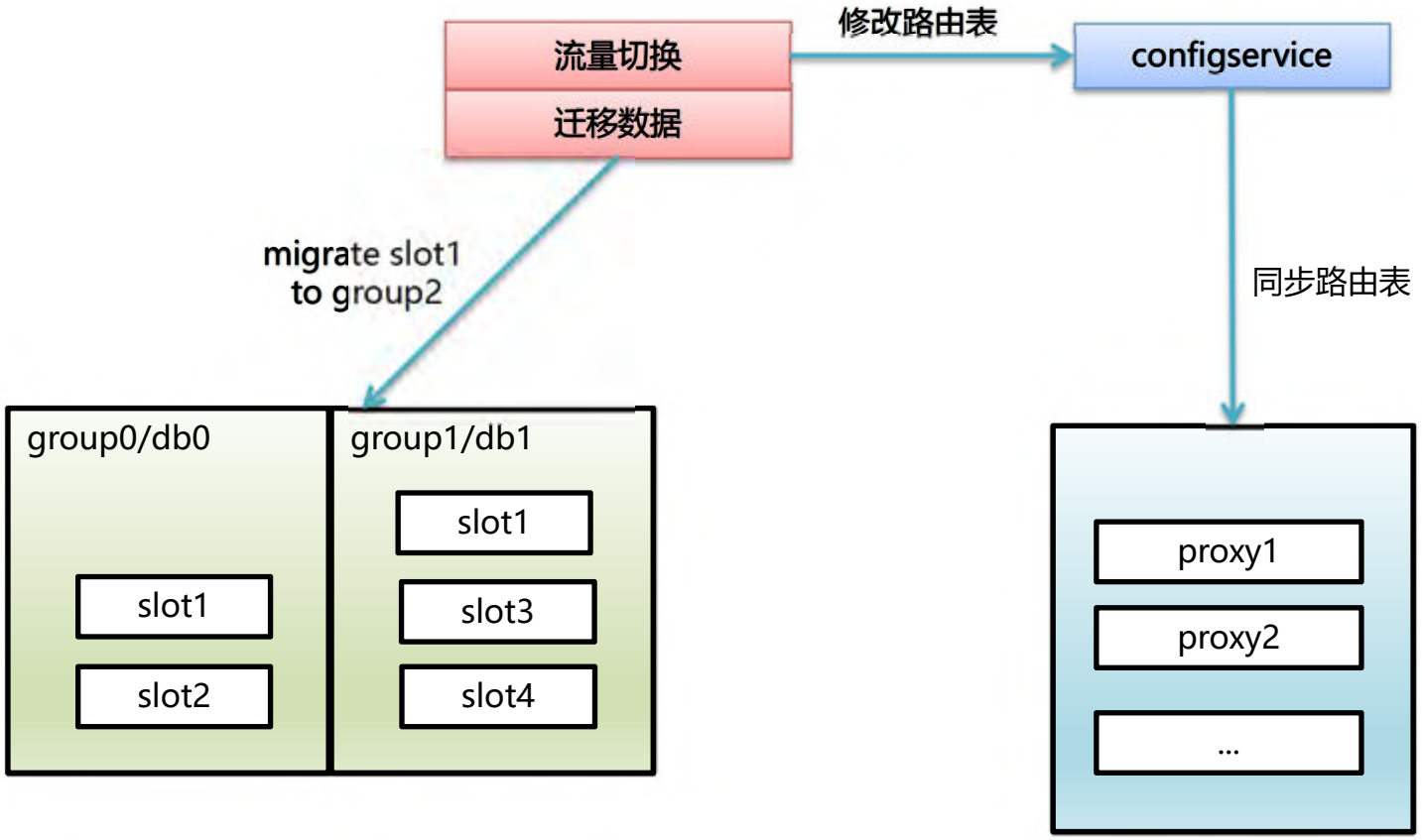
Tribe系统设计的考量

动态路由



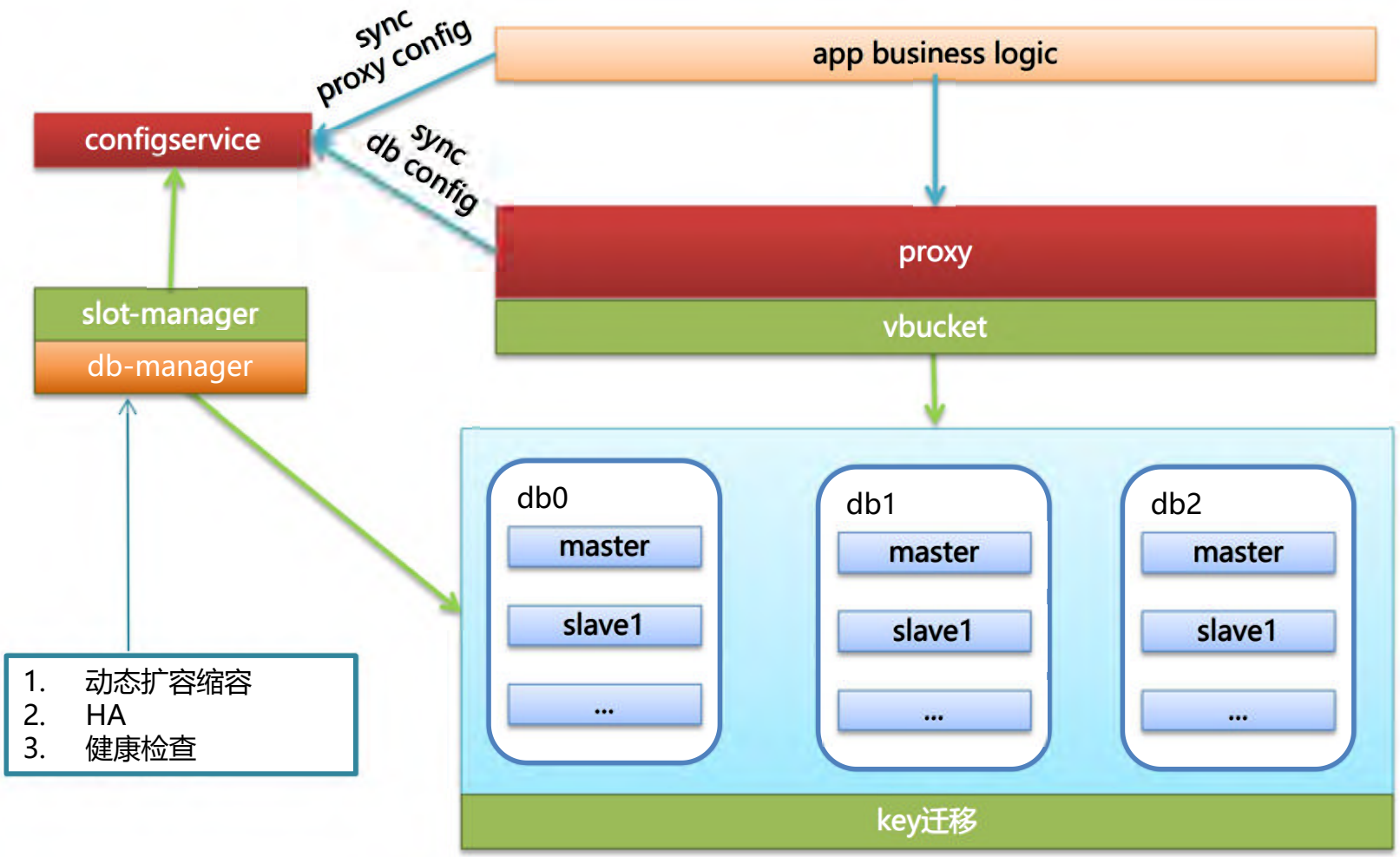
Tribe系统设计的考量

在线slot迁移



Tribe系统设计的考量

Tribe系统设计三期



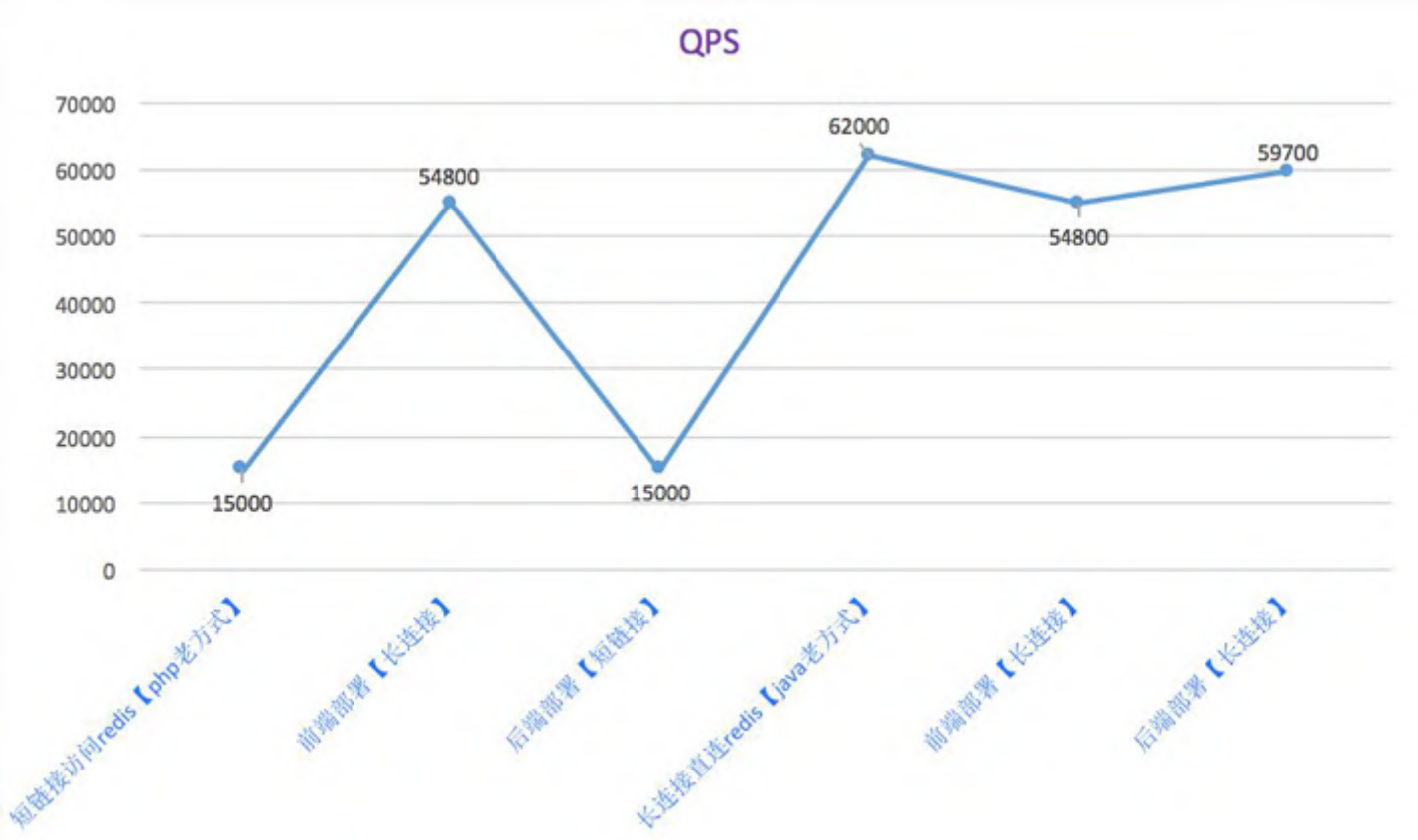
Agenda

- 个人介绍
- 微博REDIS的使用介绍
- Tribe系统设计的考量
- Tribe系统性能测试对比
- Tribe系统运维点滴

Tribe系统性能测试对比

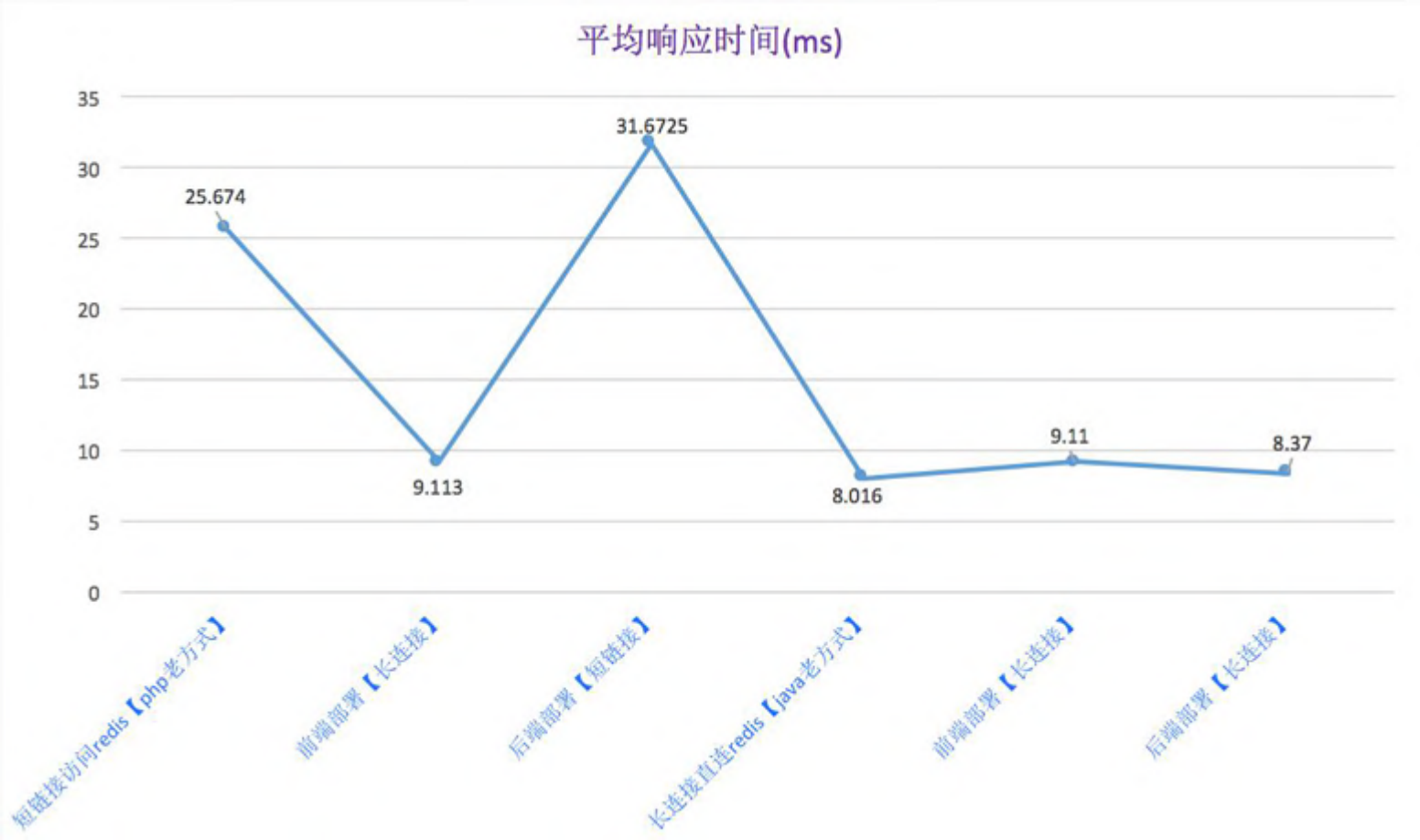
Tribe系统性能测试

- 300w/set/get
- 500线程并发
- Key:21B / value:10B



Tribe系统性能测试对比

Tribe系统性能测试



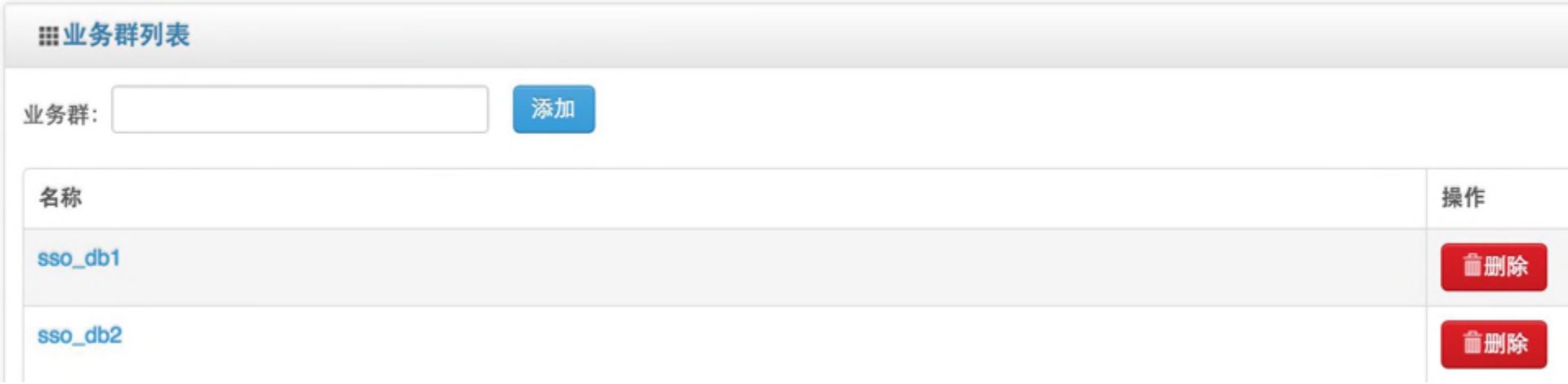
Agenda

- 个人介绍
- 微博REDIS的使用介绍
- Tribe系统设计的考量
- Tribe系统性能测试对比
- Tribe系统运维点滴

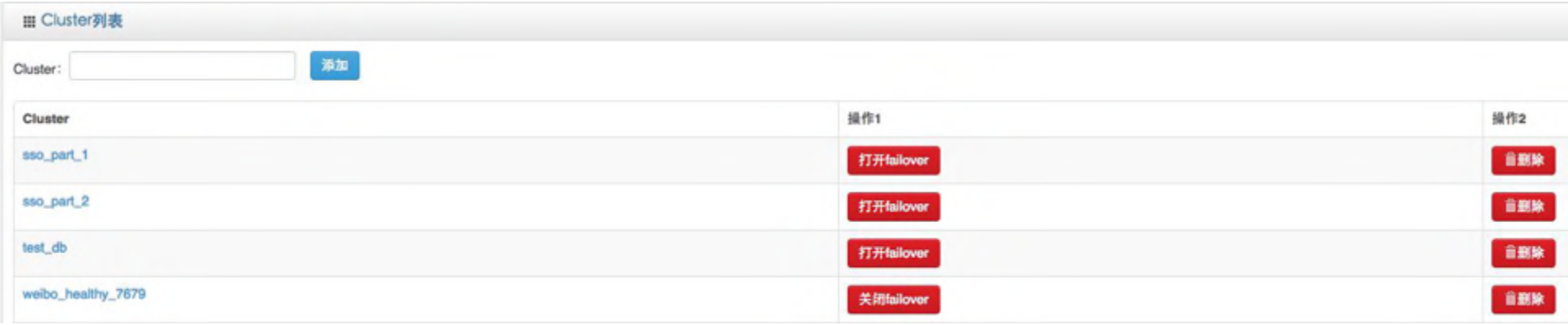
Tribe系统运维点滴

新浪微博 weibo.com

❖ 添加业务



❖ 添加Cluster



新浪微博 weibo.com

基础配置


```
// 从库挂掉，是否自动剔除
    auto_eject_hosts: 'true'
// 数据分布方式，支持slot和modula
//  modula = hash(key) % [port number]
distribution: slot
// hash函数选择
hash: crc32a
// 优先访问本地从库
local_first: 'true'
// 如果从库全部挂掉，允许临时读取主库
master_emrg_read: 'true'
// 主库ip配置
masters:
- 192.168.1.11:6070:tc
- 192.168.1.12:6080:tc
// 机器连续失败多少次，被认定为failure
server_failure_limit: '6'
// 如果机器failure后，多少ms以后重试是否恢复
server_retry_timeout: '10000'
// 从库配置
slaves:
- - 192.168.1.21:6050:tc
- - ''
// proxy和后端redis的超时时间
timeout:   '10'
```

Tribe系统运维点滴

新浪微博 weibo.com

- 配置group

Group列表

ID	Master	Slaves	操作	操作1
4	ip:port:idc	ip:port:idc ip:port:idc ip:port:idc	添加	
17679	[redacted]	[redacted]	域名管理	删除

- 更新路由表

路由列表　同步路由表

vbucket start	vbucket end	group id	stat	keynum
0	254	17680	consistent	29964500
255	500	17681	consistent	28853395
501	768	17682	consistent	31609336
769	1023	17679	consistent	29969000

- 添加

DNS信息

IP: [redacted]　IDC: 请选择　添加Proxy

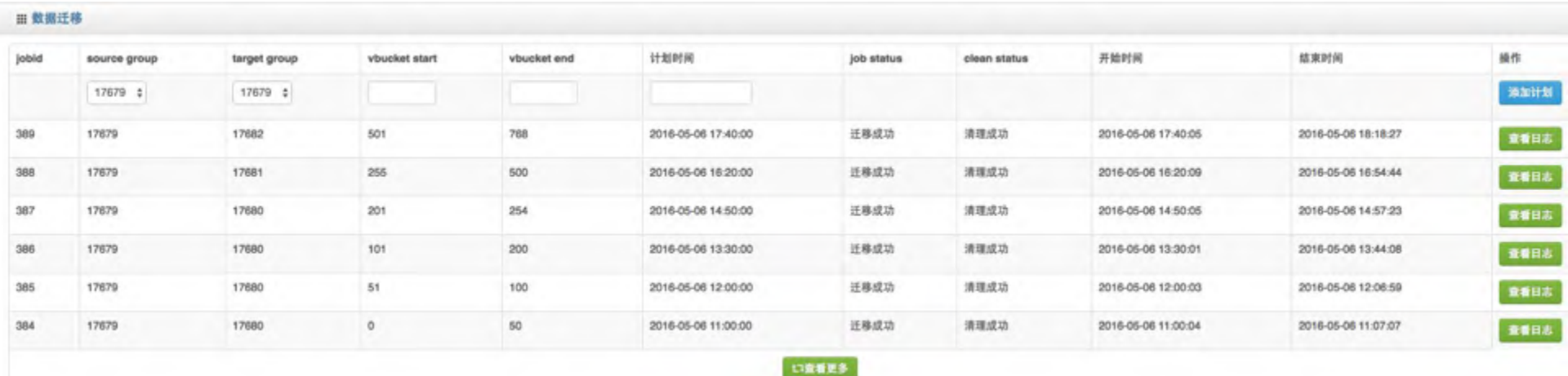
机房	域名	ip	版本	配置签名	路由签名	黑名单MD5	操作1	操作2
mars	[redacted]	[redacted]	2.0.0.35	[redacted]	[redacted]	[redacted]	下线 停止报警	
		[redacted]	2.0.0.35	[redacted]	[redacted]	[redacted]	下线 停止报警	
		[redacted]	2.0.0.35	[redacted]	[redacted]	[redacted]	下线 停止报警	

Tribe系统运维点滴

新浪微博 weibo.com

- 添加产品线信息

- 数据迁移



以微博之力 让世界更美！

weibo.com