



2017第八届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2017

新浪Redis运维实践

赵景波

新浪网高级DBA

自我介绍

- Oracle/MySQL/NoSQL DBA
- 2015年加入新浪数据库平台
- 热爱开源DB内部原理探究
- 微博：@zbdba



大纲

- 新浪数据库平台概览
- 精细化运维
- Redis Cluster Proxy

大纲

- 新浪数据库平台概览
- 精细化运维
- Redis Cluster Proxy

新浪数据库平台概览



2008

2010

2012

2014

2015

2016

2017

2018

平台规模：

- 15个IDC数据中心
- 1200+ 物理机器
- 7000+ 实例
- 1000亿+ hits/天

重要业务：



大纲

- 新浪数据库平台概览
- 精细化运维
- Redis Cluster Proxy

精细化运维

基础服务：

- 服务高可用
- 监控报警
- 服务化

细化服务：

- 成本优化
- 数据支撑

精细化运维

基础服务：

- 服务高可用
- 监控报警
- 服务化

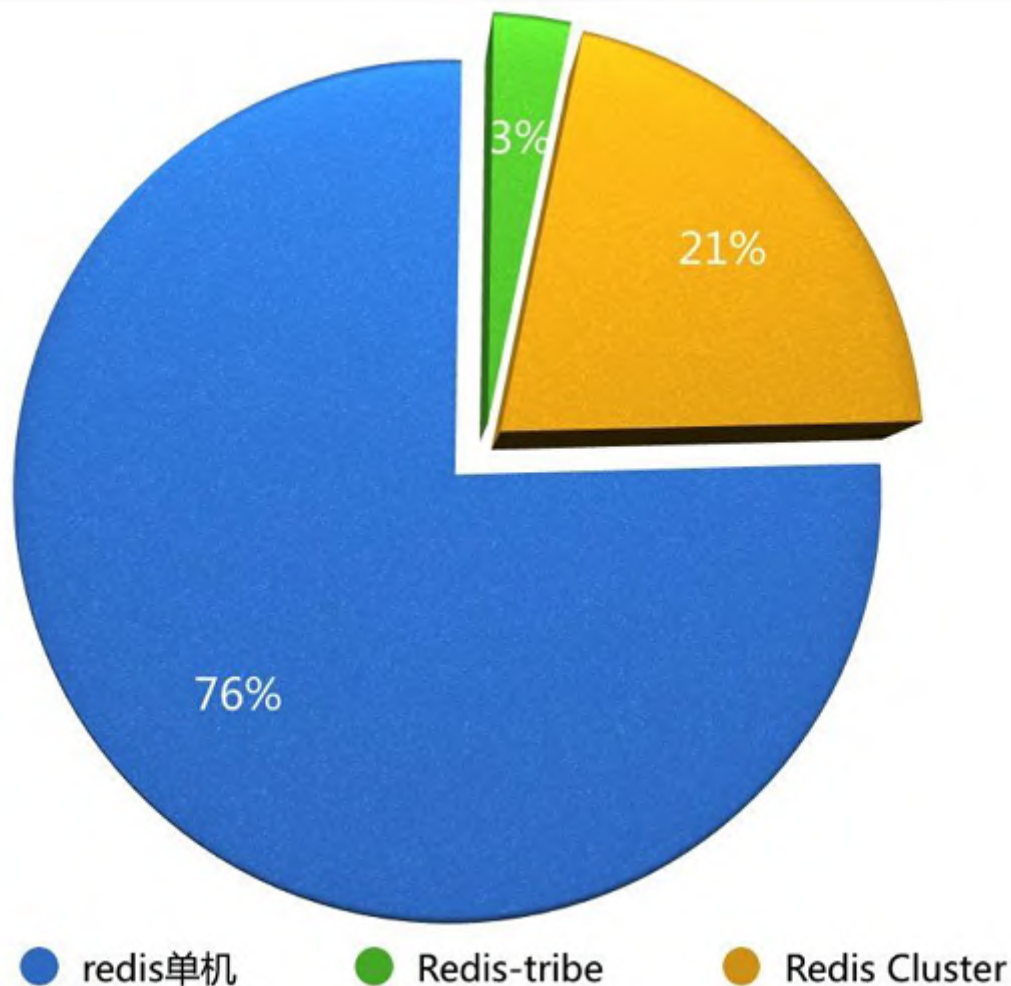
细化服务：

- 成本优化
- 数据支撑

精细化运维

部署架构

- Redis单机
- Redis Cluster
- Redis Tribe



精细化运维

高可用-Sentinel

一个sentinel集群，17sentinel节点，分布于南北共9个数据中心，监控500+个端口，1000+实例。

IDC	数量
北方IDC1	2
北方IDC2	1
北方IDC3	2
北方IDC4	3
北方IDC5	3
北方IDC6	1
北方IDC7	1
南方IDC1	2
南方IDC2	2



- 并发切换40+端口
- 切换成功率 98%
- 单次切换时间<5s

精细化运维

高可用-Sentinel踩的坑

- 单个IDC的数量不建议超过(Sentinel数量-quorum)
- 客户端长连接问题
- 可以通过设置slave-priority 控制选举 (跨异地机房部署)
- 防止误切, 切换灵敏度控制 (quorum、 down-after-milliseconds、 failover-timeout)

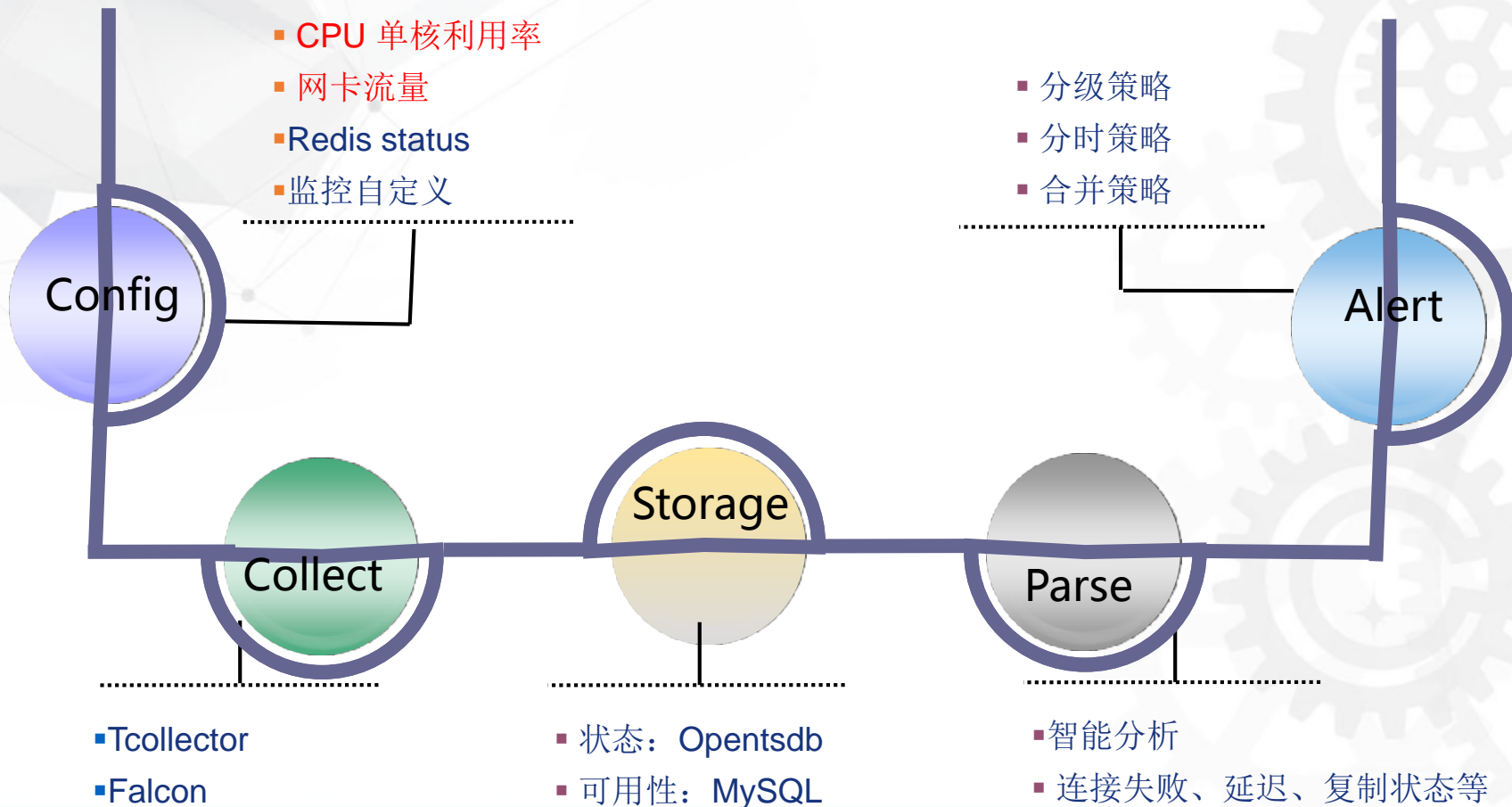
- 设置maxclient、 timeout参数
- 建议采用Sentinel 3.2.8 及以上版本

quorum 12
down-after-milliseconds 20000
failover-timeout 120000

Error registering fd event for the new client: Numerical result out of range (fd=10135)

精细化运维

监控-Redis监控总览



精细化运维

监控-Redis监控总览

可用性监控		状态监控	
连接检测	连接失败检测	访问量	read
	插入检测		write
变量检测	readonly	内存	maxmemory
	maxmemory		used maxmemory
	maxmemory-policy		内存碎片率
	连接数监测(20000)	key	evicted_keys
	内存比监测(80%)		key 数量
主从复制监测	角色监测	命中率	
	复制状态检测	连接数	
	延迟检测	其他状态	从库延迟
			慢查数量

精细化运维

监控-Redis监控总览

RedisMaster监控配置

监控设置	报警设置
连接监测 <input checked="" type="radio"/> 开 <input type="radio"/> 关	连接失败 <input checked="" type="radio"/> 开 <input type="radio"/> 关
插入监测 <input type="radio"/> 开 <input checked="" type="radio"/> 关	插入失败 <input checked="" type="radio"/> 开 <input type="radio"/> 关
变量监测 <input checked="" type="radio"/> 开 <input type="radio"/> 关	变量异常
readonly <input type="radio"/> readonly <input type="radio"/> no	readonly <input checked="" type="radio"/> 开 <input type="radio"/> 关
状态监测 <input checked="" type="radio"/> 开 <input type="radio"/> 关	状态异常
连接数监测 < <input type="text" value="10000"/>	连接数 <input checked="" type="radio"/> 开 <input type="radio"/> 关
内存使用比监测 < <input type="text" value="80"/> %	内存使用比 <input checked="" type="radio"/> 开 <input type="radio"/> 关
主从复制监测 <input checked="" type="radio"/> 开 <input type="radio"/> 关	主从复制异常 <input checked="" type="radio"/> 开 <input type="radio"/> 关
角色监测 Role == Master	

精细化运维

监控-Redis监控总览



精细化运维

服务化-服务自助



精细化运维

服务化-服务自助

编写Redis资源信息

产品线

族群名称 +

标题

用途描述

Redis类型 存储 缓存

数据类型

机房选择 北京 master 北京 slave
 北京 master 广州 slave
 广州 master 广州 slave
 广州 master 北京 slave

最大内存

内存是否过期 否 是

当前峰值(rps读)

当前峰值(rps写)

预估最大峰值(rps读)

预估最大峰值(rps写)

备份策略要求

存储的数据是什么

宕机影响范围

后场是否有数据库 无 有

精细化运维

服务化-服务自助

概要信息

标题: 北美发布系统申请redis
提案类型: 服务 > 运维中心 > 平台部 > 数据云平台 > Redis资源申请

提案ID: 195294 TA的轨迹

服务申请人

服务申请人: [模糊]
申请人手机: [模糊]
申请人分机: 2336

基本信息

提案状态: 已关闭
是否紧急: 非紧急
提案结果: 成功
提案创建时间: 2016-11-29 17:48:18
提案关闭时间: 2016-12-06 10:15:42
最后更新时间: 2016-12-06 10:15:42
最后更新人: [模糊]

申请内容

项目名称	北美发布系统
所属产品线	SINA / Portal / COM / 发布系统
用途描述	北美发布系统申请redis
redis 类型	存储
机房选择	北京master / 广州slave
最大内存	16G
内存能否过期	否
当前峰值(rps读)	1000/s
当前峰值(rps写)	1000/s
预估最大峰值(rps读)	15000/s
预估最大峰值(rps写)	10000/s
备份策略要求	实时备份
宕机影响范围	北美发布系统, 网站
后端是否有数据库	有

审批记录

step 1 审批结果: Redis资源申请设备组
[绿色进度条] 已审批通过

执行记录

step 1 执行结果: 数据云平台机器人
[绿色进度条] 操作成功

注释

数据云平台机器人 2016-12-02 12:20:50 [12-02] 添加了注释
执行成功

数据云平台机器人 2016-12-02 12:21:02 [12-02] 添加了注释
域名信息: 北京主库 sina.com.cn 广州从库 ta.com.cn 端口 22400

精细化运维

服务化-服务自助

自动扩容

自动扩容 扩容内容 扩容统计

管理员 不限 yan32 su1

机型 全选 sas sas

角色 全选 m s mb rs

IDC 全选 上海 北京 北京 北京 北京 北京 北京 北京 北京 北京 北京 天津 天津 广州 广州 香港

搜索空闲机器

限制条件 流量限制 开 关 磁盘利用率 ≤ 60 % 模板数 ≤ 20

新增磁盘

共22条 | 每页10条 | 当前1/3页 | 1 2 3

同产品单元(22) 其他推荐(22)

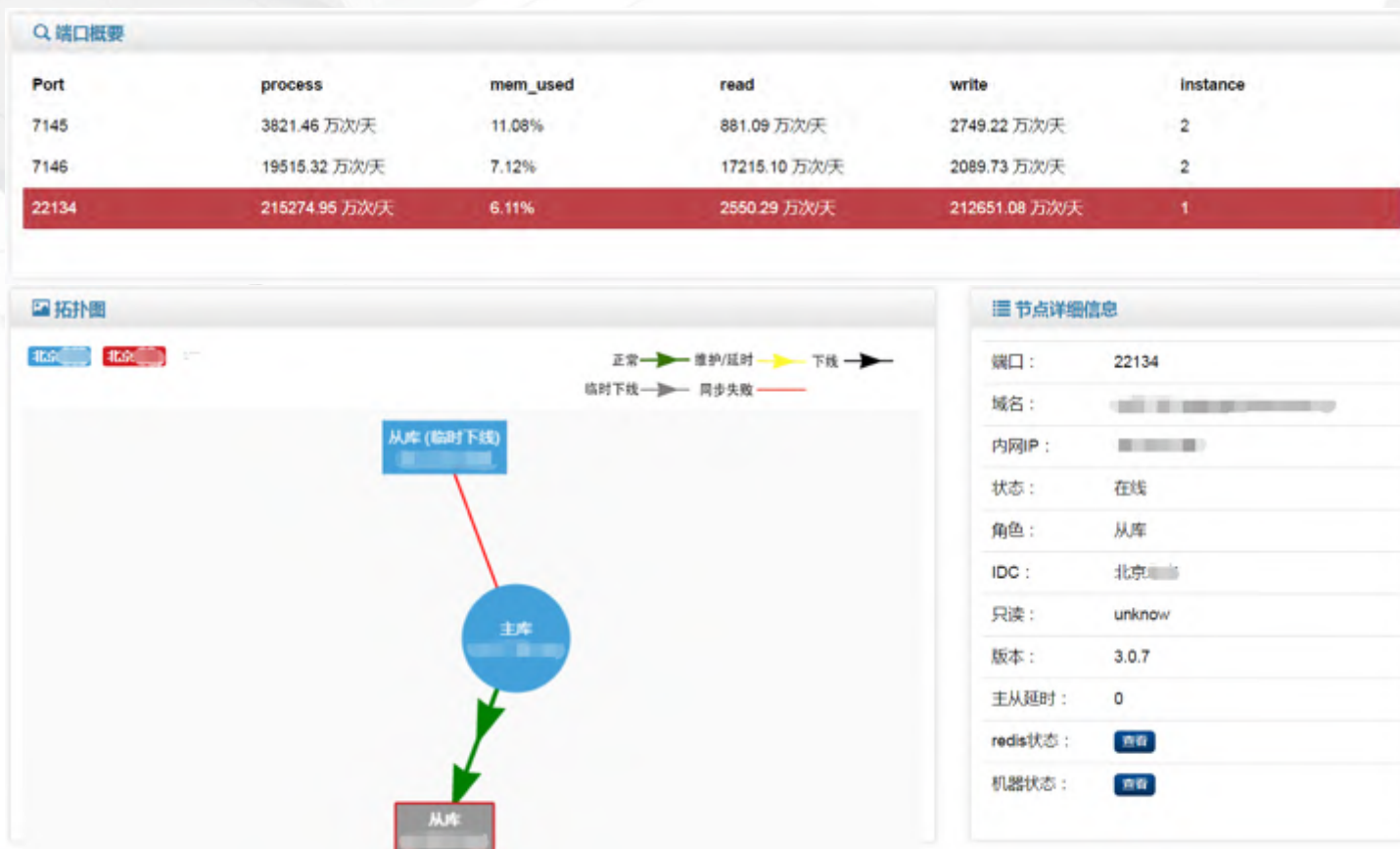
	IP	管理员	磁盘利用率	磁盘剩余/总量	内存剩余/总量	机型类型	IDC	服务	状态	MySQL/Redis (端口/产品单元/读/写/状态/混路)	系统负载	CPU空间	IO利用率	更新时间
<input type="checkbox"/>		yanan32	4%/7.69%	410/423	86/125	sas	北京	redis	正常	redis_s11924/新闻推荐产品/0万/0万/1/可以混路 redis_s22414/天乙项目-用户兴趣/0万/0万/1/可以混路 redis_s22426/天乙项目-用户兴趣/0万/0万/1/可以混路 redis_s22462/天乙项目-用户曝光历史/0万/0万/1/可以混路 dummyredis/MALL/0万/0万/1/可以混路 redis_c22693/新闻推荐测试端口/0万/0万/1/可以混路 redis_c22760/新闻推荐测试端口/0万/0万/1/可以混路	0.26	98.99%	0.02%	2016-11
<input type="checkbox"/>		yanan32	5%/8.65%	406/423	78/125	sas	北京	redis	正常	redis_m22454/天乙项目-用户曝光历史/0万/0万/1/可以混路 redis_m22465/天乙项目-用户曝光历史/0万/0万/1/可以混路 redis_s22454/天乙项目-用户曝光历史/0万/0万/1/可以混路 redis_s22465/天乙项目-用户曝光历史/0万/0万/1/可以混路 redis_s22922/新闻推荐时用户兴趣存储/0万/0万/1/可以混路 dummyredis/MALL/0万/0万/1/可以混路	0.35	98.47%	0.12%	2016-11
<input type="checkbox"/>		yanan32	8%/12.03%	392/423	74/125	sas	北京	redis	正常	redis_m11433/新闻推荐/0万/0万/1/可以混路 redis_m22489/新闻推荐的用户兴趣存储/0万/0万/1/可以混路				

11922 新闻推荐的用户兴趣存储 19G mars 可以混路

下一步(B) 返回 清空

精细化运维

服务化-服务Dashboard



精细化运维

基础服务：

- 服务高可用
- 监控报警
- 服务化

细化服务：

- 成本优化
- 数据支撑

精细化运维

成本优化-第一阶段

业务：

- 业务存储类型
- 响应时间要求
- 存储容量
- QPS
-

DBA：

- 业务场景
- 资源成本
- 运维成本



精细化运维

成本优化-第二阶段

Redis 低读写量比例端口

port	write	read	rw_ratio	cpunit	time
7896	123253119	22242	5541.4585	猜你喜欢新闻客户端新闻推荐	2017/5/9 17:21
7906	101591130	19834	5122.0697	猜你喜欢新闻客户端新闻推荐	2017/5/9 17:21
22803	340604862	104237	3267.6004	媒体平台Push系统下发队列	2017/5/9 17:21

Redis无读写端口（连续一周）

port	内存使用(单位M)	产品单元
8193	6	直播聊天室消息存储
8067	5	体育竞猜项目(非微博)
8068	5	舆情监控系统(非微博)

Redis低内存利用率端口

端口	内存使用(单位M)	分配内存(单位M)	使用比	产品单元
22126	2	19072	0.0001	直播平台
22128	9	57220	0.0002	战争项目
7991	20	95364	0.0002	媒体缓存RSS抓取队列

精细化运维

成本优化-第三阶段

开启超线程

- 配比不合理机器开启超线程
- 共享池机器全部开启超线程

划分资源池

- 重点业务划分资源池
- 小业务放共享池
- 机器以资源池划分

制定部署规范

- 选择资源池
- 剩余内存 \geq 服务器内存的20%
- 服务器剩余内存 $>$ 最大端口的内存
- 总实例数 $<$ CPU的核数
- 服务器负载评分 $>$ 80

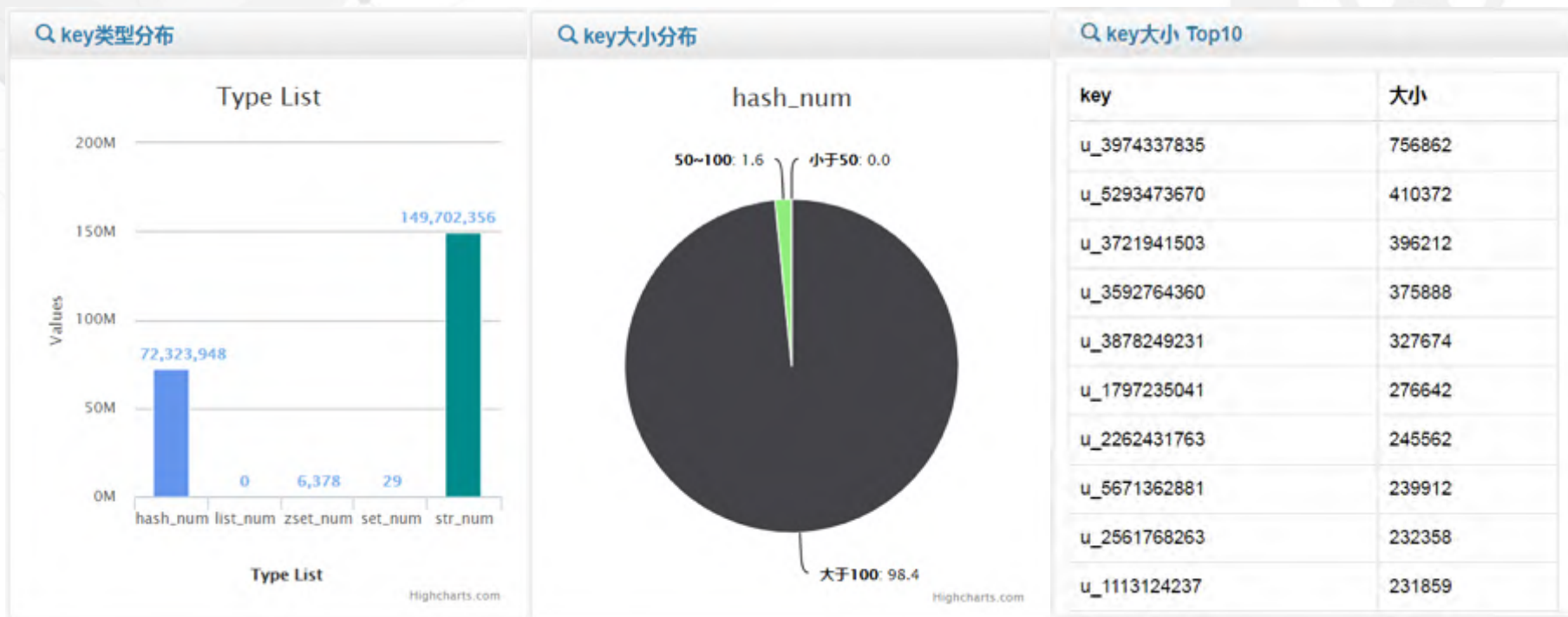
精细化运维

数据支撑-背景

- 业务使用的什么数据类型？
- 分别有多少个key？
- 有没有大key？
- key都活跃吗？
- 响应时间是多少？

精细化运维

数据支撑-Redis key 分析



精细化运维

数据支撑- cold/dead Key/hot key

- 空闲时间超过15天则为cold key
- 空闲时间超过30天则为dead key
- Hot key 待完善

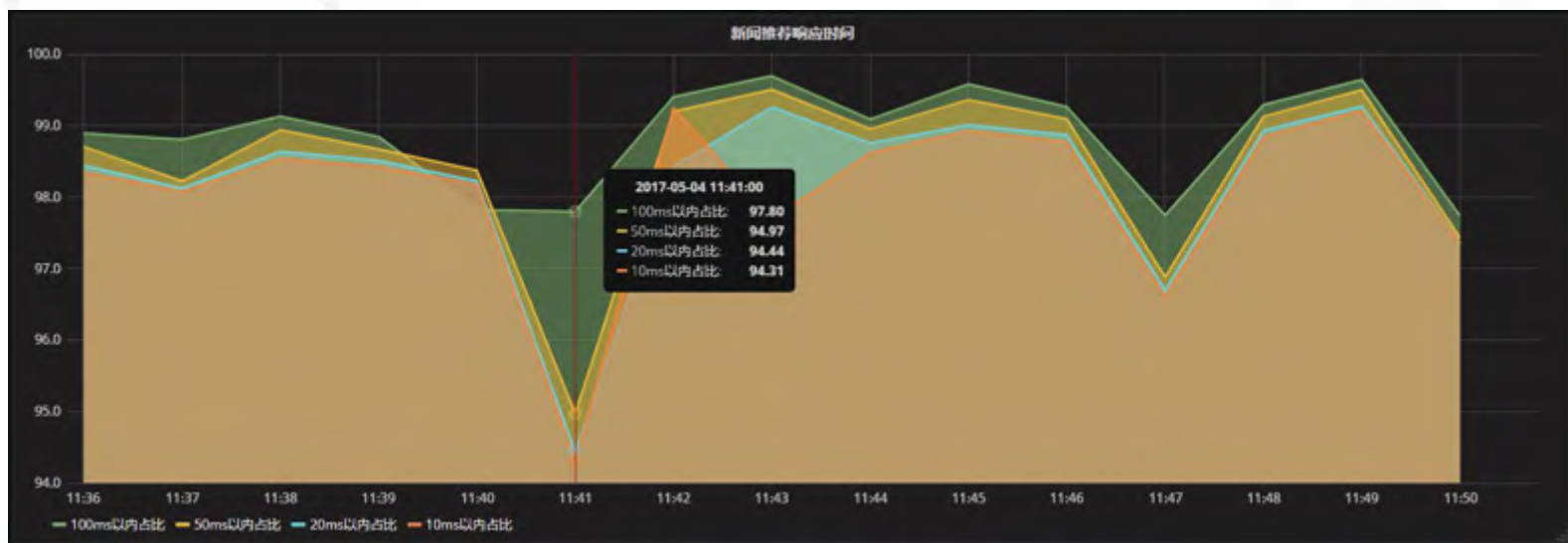
port	product_name	key	idle_time(day)	parse_time
8002	猜你喜欢-微博好友亲密度数据	stat_read	118.7	2017/1/4 11:25:07
8002	猜你喜欢-微博好友亲密度数据	new_stat_read	118.7	2017/1/4 11:25:07
8002	猜你喜欢-微博好友亲密度数据	stat_vote	118.7	2017/1/4 11:25:07
20211	测试端口-报警请忽略	vpid_124991065	44.1	2017/1/4 18:28:14

精细化运维

数据支撑- 响应时间

定制tcprstat+Grafana

timestamp	count	50ms	20ms	5ms	3ms	stddev	50ms	5ms	95_std	50ms	5ms	99_std
1494310766	159	159	159	157	156	1387	151	151	1003	157	157	1153
1494310767	161	161	161	160	153	1326	152	152	945	159	159	1208
1494310768	136	136	136	131	127	1701	129	129	1052	134	131	1450
1494310769	149	149	149	149	146	1149	141	141	934	147	147	1076
1494310770	173	173	173	172	168	1292	164	164	1031	171	171	1192
1494310771	158	157	157	157	152	5277	150	150	927	156	156	1127



大纲

- 新浪数据库平台概览
- 精细化运维
- **Redis Cluster Proxy**

Redis Cluster Proxy

平台Redis集群的演化

Redis Tribe



Redis Cluster



Redis Cluster + Proxy

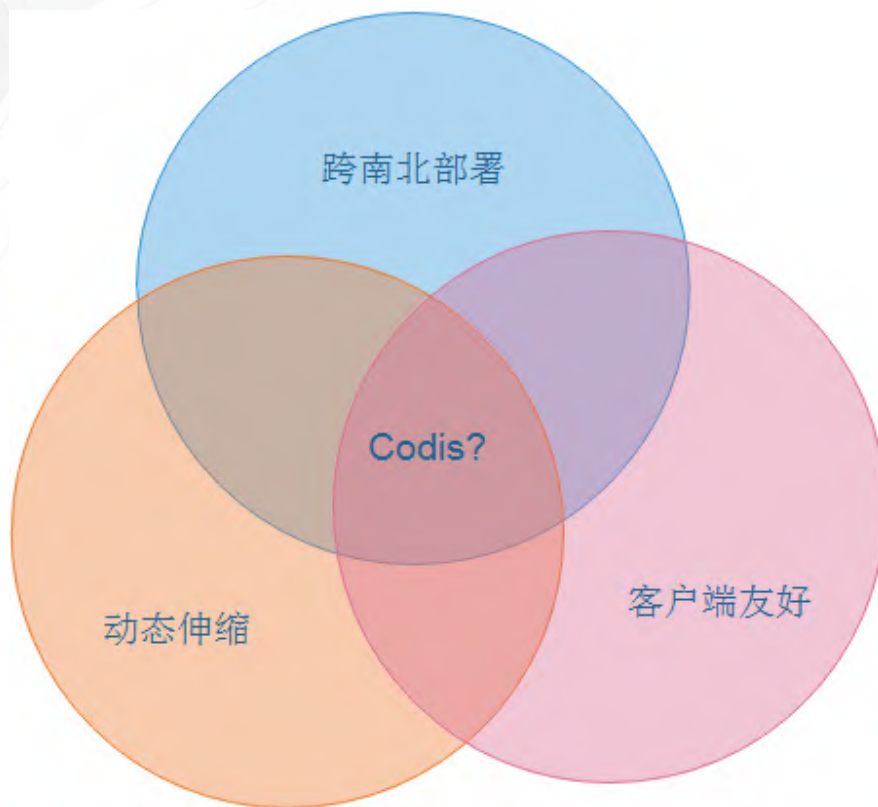
- 类Codis产品
- 需要前后端部署
- 无人维护

- 客户端不友好
- 无法读取从库
- 无法跨异地IDC部署

- 支持动态伸缩
- 跨南北部署
- client无需支持smart client
- 支持本地化读从库

Redis Cluster Proxy

面临的问题



Redis Cluster Proxy

产品选择

功能特性	Codis	Corvus	Redis Tribe
是否支持Cluster	No	Yes	No
是否支持动态扩容	Yes	Yes	Yes
是否修改Redis	Yes	No	Yes
是否依赖外部组件	Zookeeper	No	MySQL
是否支持HA	Sentinel	Cluster	Yes
项目是否活跃	Seldom	Normal	Never
是否有管理界面	Yes	No	Yes
是否支持节点之间数据迁移	Yes	Yes	Yes

Redis Cluster Proxy

方案简介

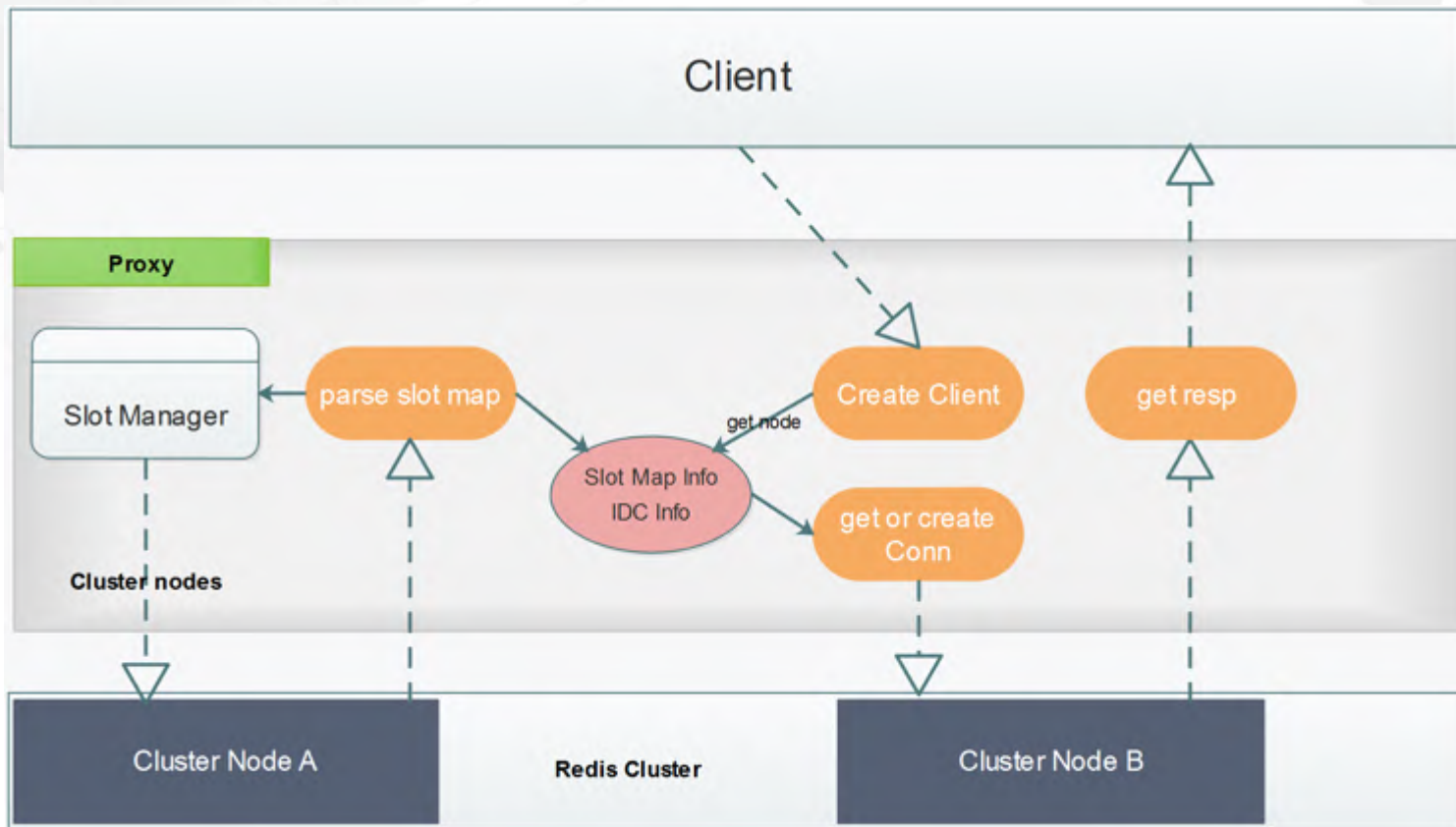
Corvus 

- 支持不同IDC本地化读
- 支持跨异地域部署
- 定制Redis Cluster Auto Failover

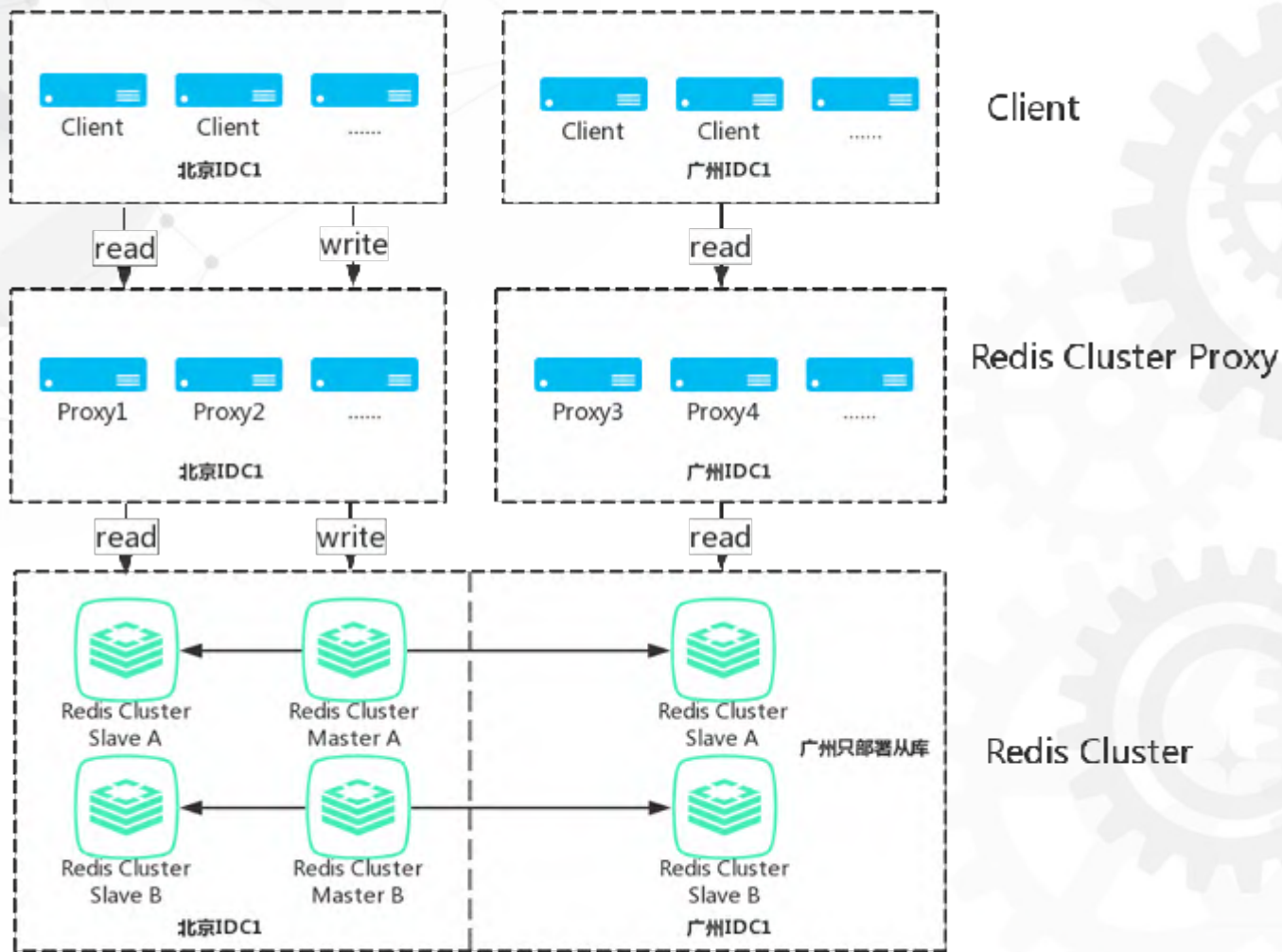
Corvus开源地址：<https://github.com/eleme/corvus>

Redis Cluster Proxy

Crovus具体实现



Redis Cluster Proxy



Redis Cluster Proxy

具体实现-元数据管理

Cluster Nodes Command



Get slot map info



单个Slot对应的节点：

- 集群主节点
- 与proxy相同IDC从节点
- 与Proxy相同地域从节点
- 其他从节点

IDC Config (动态链接库)

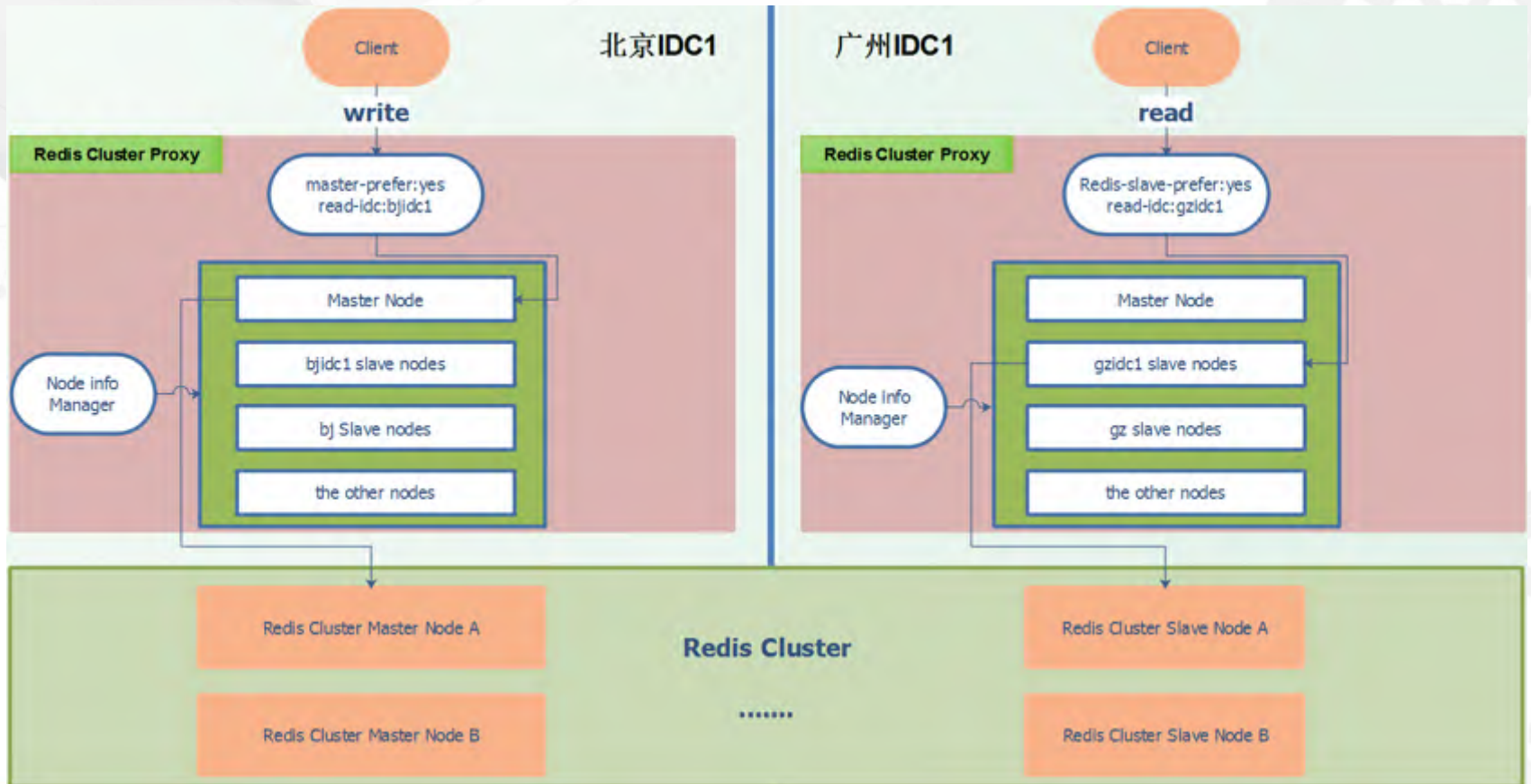


Get Node IDC Info



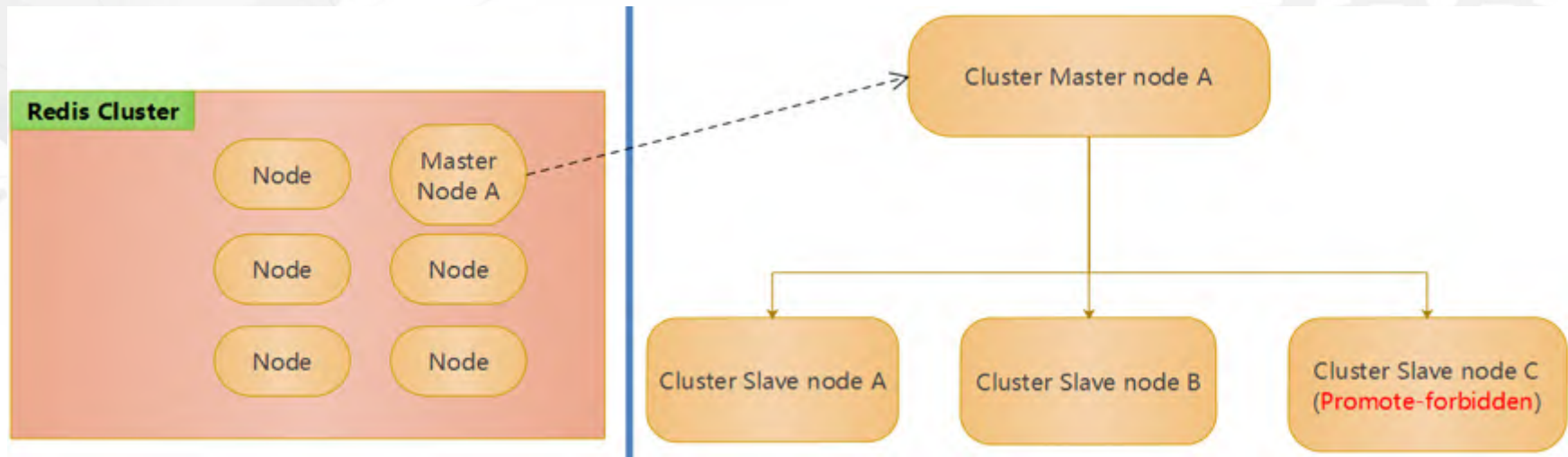
Redis Cluster Proxy

具体实现-跨异地域部署



Redis Cluster Proxy

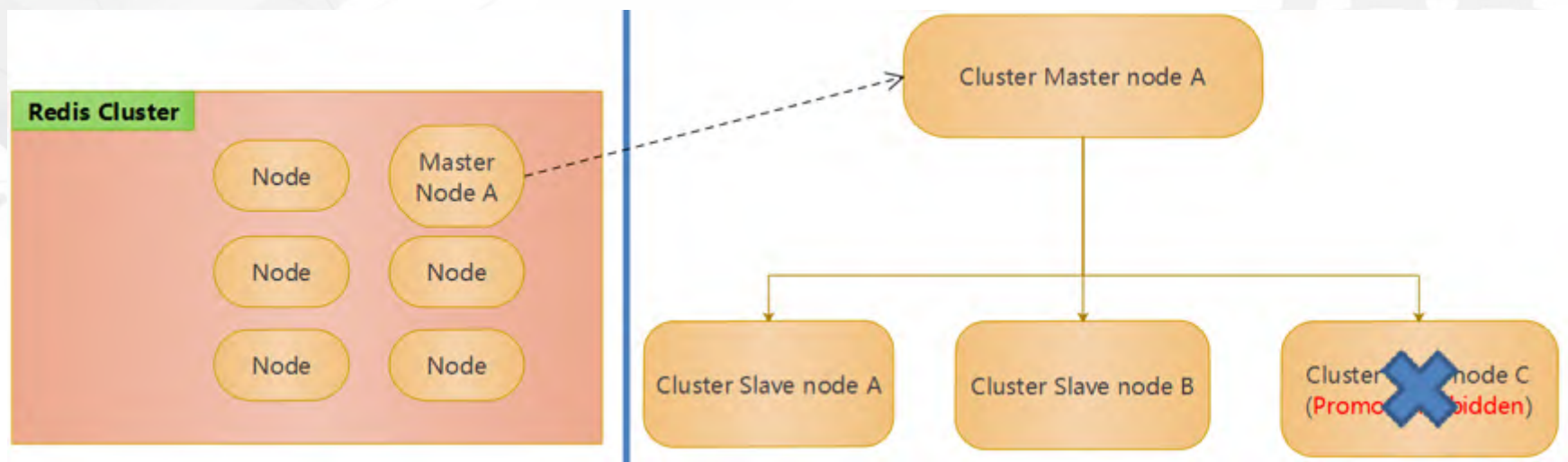
具体实现-定制Cluster Auto Failover



- 增加promote-forbidden参数禁止某些从库参与主从选举

Redis Cluster Proxy

具体实现-定制Cluster Auto Failover



- 在Master 故障时，Slave Node C 不能参与选举。
- 在无Slave可选择时，可采用CLUSTER FAILOVER 命令切换到Slave Node C上



招聘MySQL、NoSQL DBA



THANKS

SequeMedia
威拓传媒

IT168.com

ITPUB

ChinaUnix