京东分布式K-V存储设计与挑战

SACC2017

京东商城-基础架构部-丁俊

2017-10

产品介绍

非持久化存储—JIMDB
持久化存储— FBASE

JIMDB：兼容REDIS协议，在线弹性伸缩的，数据全部保存在内存的K-V存储系统
FBASE：支持多协议，支持范围查找的持久化K-V存储系统

SACC2017

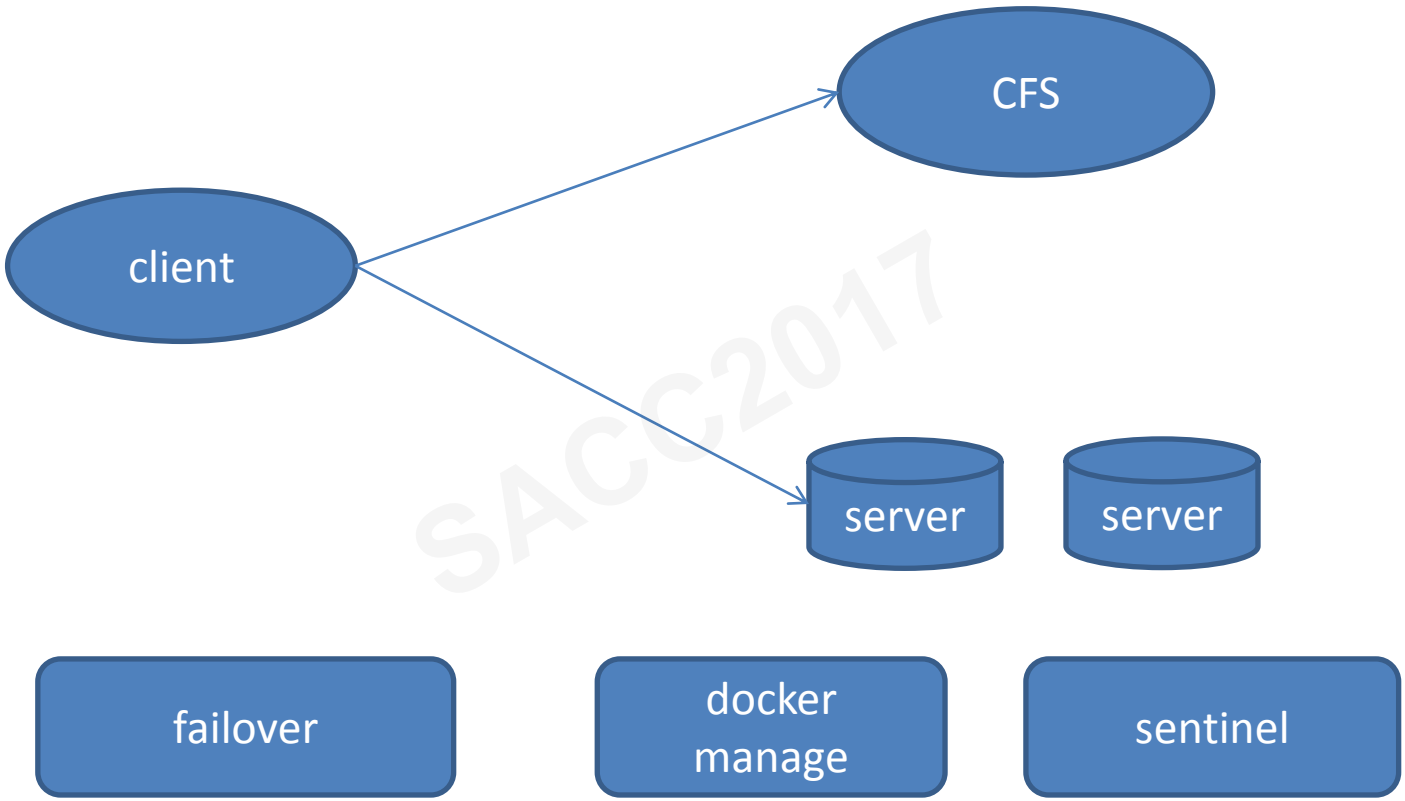
应用场景

JIMDB：读写性能要求高，性能要求优先于数据可靠性
FBASE：对数据可靠性要求高，数据量大，数据冷热分布明显

	JIMDB	FBASE
持久化	异步或无	同步
读写性能	高	低
顺序访问	不支持	支持
数据分区	哈希	范围或哈希
复制	主从异步	同步/异步
访问协议	redis	mysql/http/redis

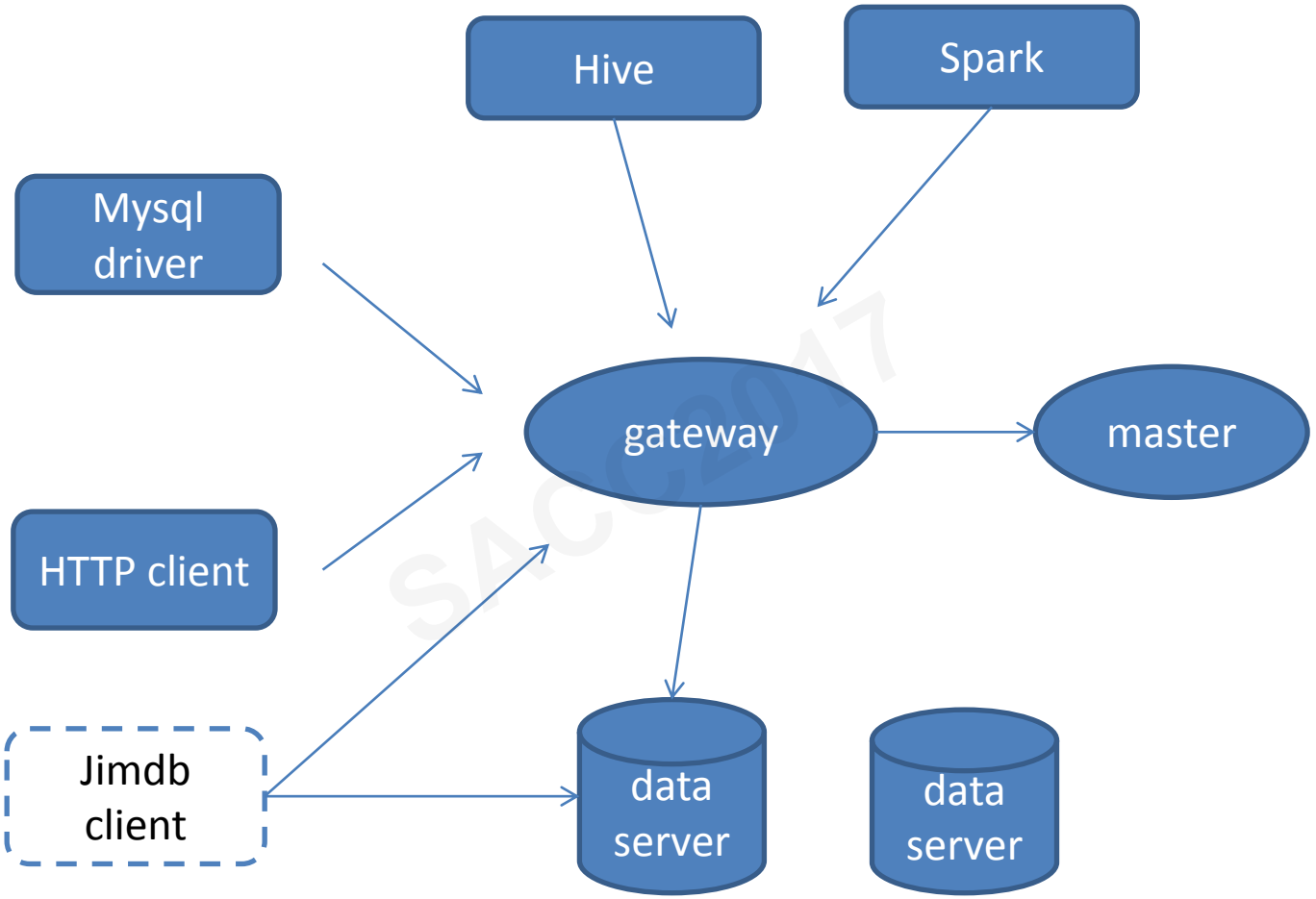
SACC2017

JIMDB架构图



SACC2017

FBASE架构图



SACC2017

面临的挑战与设计方案

- 故障检测与恢复
- 在线扩容
- 高可用
- 升级

SACC2017

JIMDB的故障检测与恢复

问题：

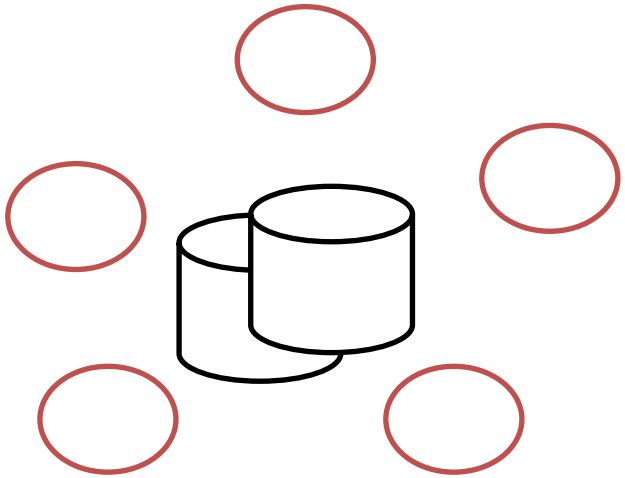
基数大，故障次数多

人工响应慢

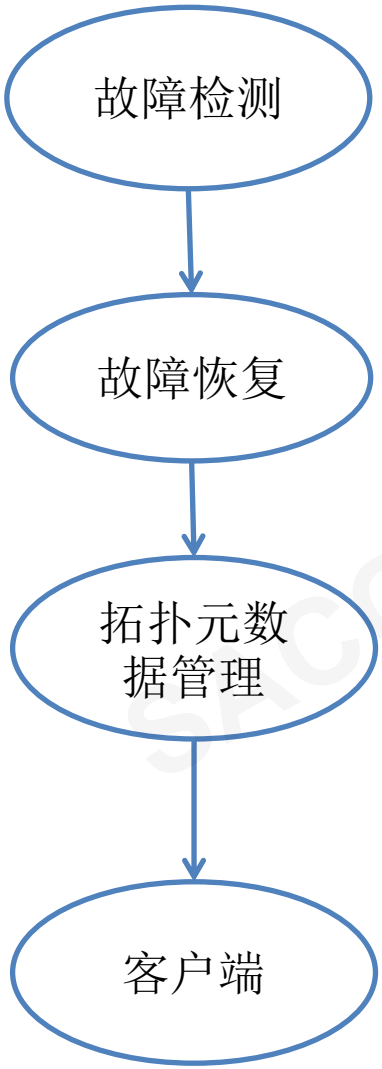
误判可能的问题：1、短暂的多master；2、频繁切换

如何避免误判：1、部分网络故障；2、服务程序繁忙响应慢

1、故障检测程序独立部署，分散在不同机架上
2、投票决定，存活状态一票否决
3、一个机房部署多组，每组负责部分实例
4、宿主机agent辅助检测确认



SACC2017



SACC2017

JIMDB在线扩容

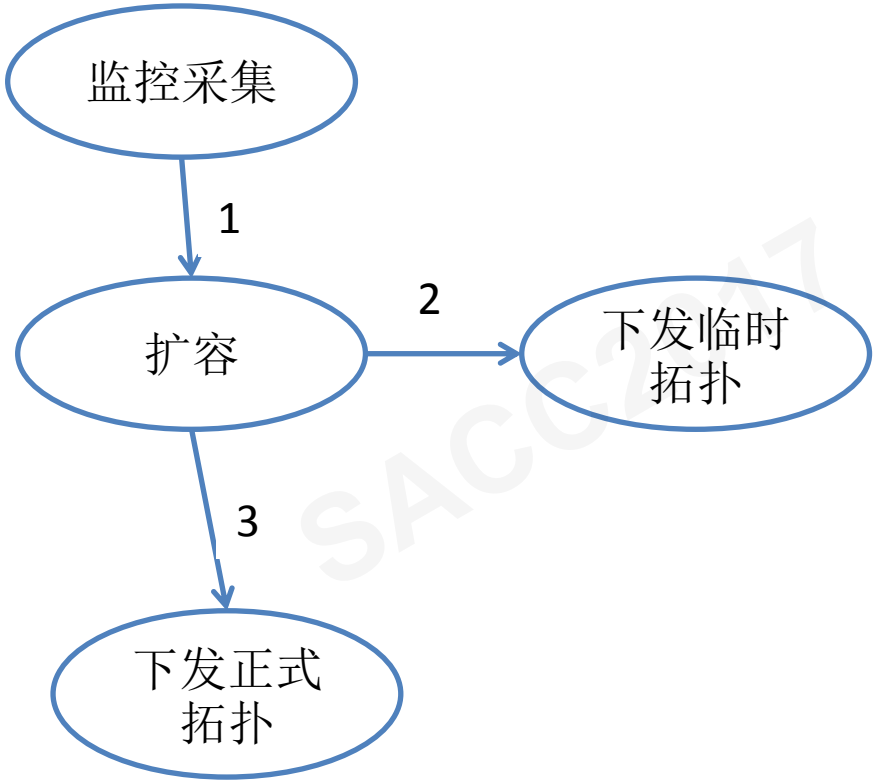
为什么要在线扩容：
1、业务增长超预期，预估不准
2、避免资源闲置
3、业务快速成长，资源紧缺是一种常态

扩容触发条件：
单个分片内存占用大小
进出流量（CPU使用率）

单个分片的大小主要考虑：
1、扩容过程的持续时间
2、CPU与内存的使用率

SACC2017

JIMDB在线扩容流程



SACC2017

JIMDB在线扩容

怎么平滑扩容：

提前把将要变更的拓扑信息下发给客户端
客户端捕捉到特定异常后使用临时拓扑
扩容完成后临时拓扑变更为正式拓扑

碰到的问题：

单个热KEY导致流量高
单个大KEY导致存储占用高
大促前的扩容需要提前规划

注意事项：

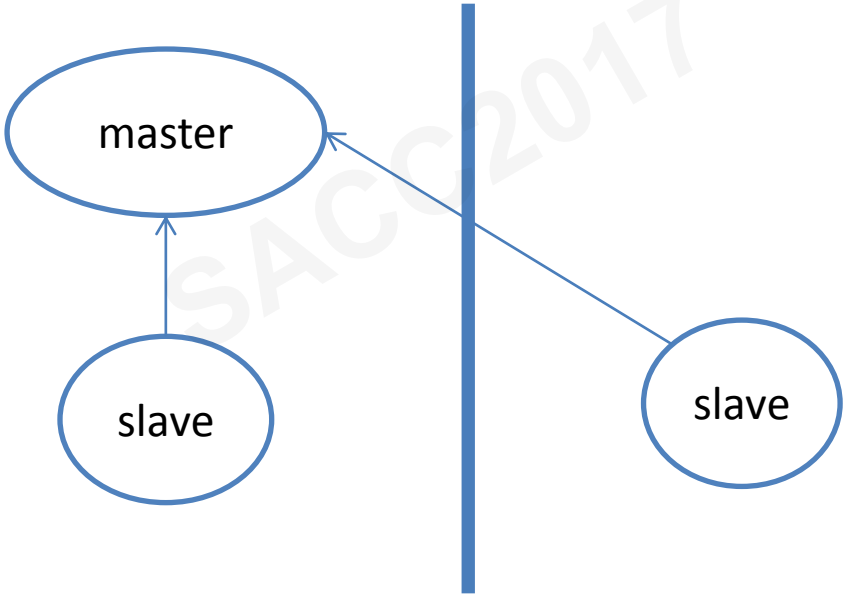
数据迁移最小单位为槽
单个shard需要控制大小，避免迁移数据多时间长

SACC2017

JIMDB复制

多副本异步复制

副本部署要求：
1、跨物理机
2、跨机架
3、同城跨机房
4、异地数据中心



SACC2017

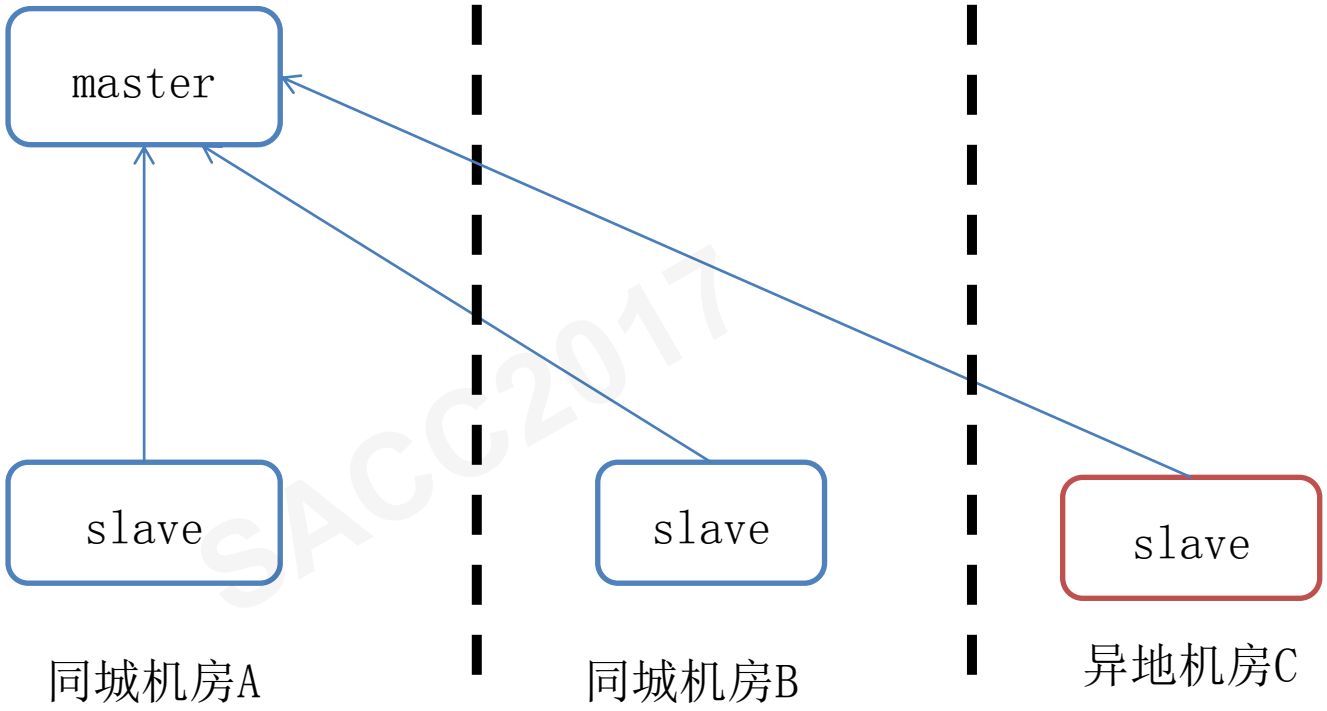
JIMDB异地灾备

1、直接部署slave，内存缓冲区
2、经过synclog模块，异地机房只是一个远程副本
3、集群间有复制关系

SACC2017

JIMDB异地灾备

直接部署slave，内存缓冲区



1、网络故障，发生全量同步的次数会增多
2、跨机房写
3、控制管理跨地域访问影响性能
4、如果要跨地域添加多个副本，同一份数据多次传输

SACC2017

JIMDB异地灾备

经过synclog模块，异地机房只是一个远程副本

避免全量同步的主要场景：

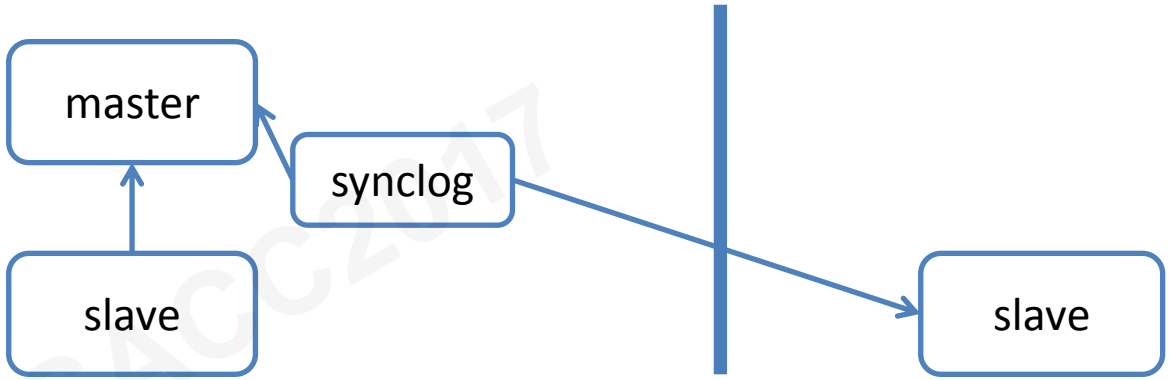
源端主从切换；
网络中断；
批量导数；
synclog模块宕机。

数据堆积：

短暂的网络不通
批量写入数据

异常行为：

Client get → null
Client get → Exception



SACC2017

JIMDB异地灾备

集群间有复制关系

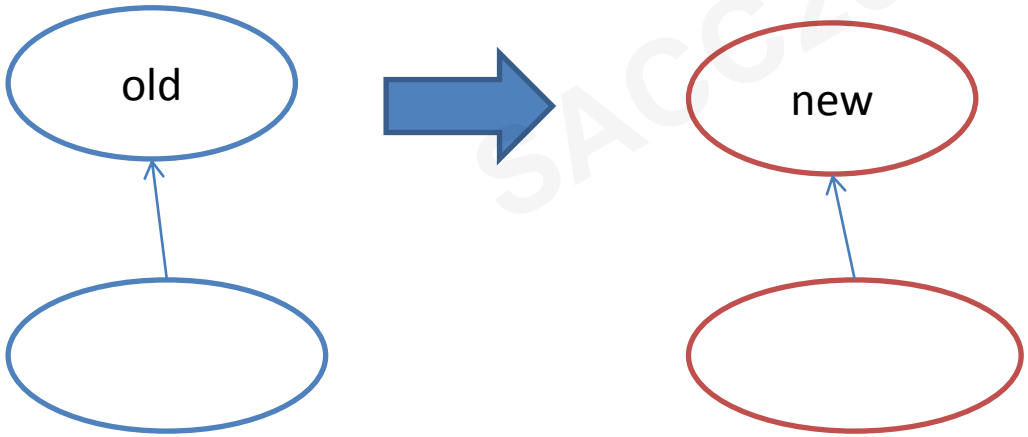
优点：
多机房同步可写
可以双向复制
形成复制关系的集群间，shard数量可以不一致

怎么避免循环复制：
KEY的属性中编码机房标识，同步模块复制的KEY机房标识不做更改

SACC2017

升级

内存中的数据做迁移
按照shard滚动升级，
新版本的容器创建在同一台宿主机上
迁移完成后客户端捕捉到数据已迁移的异常，会使用新的拓扑

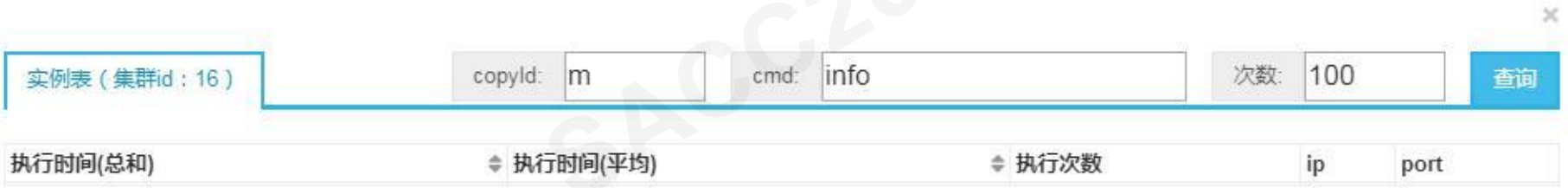


SACC2017

性能监控与排查

1、监控指标，曲线图

2、模拟用户发送命令进行检测，按性能排序



SACC2017

Fbase介绍

KEY全局有序排列

支持多种复制模式

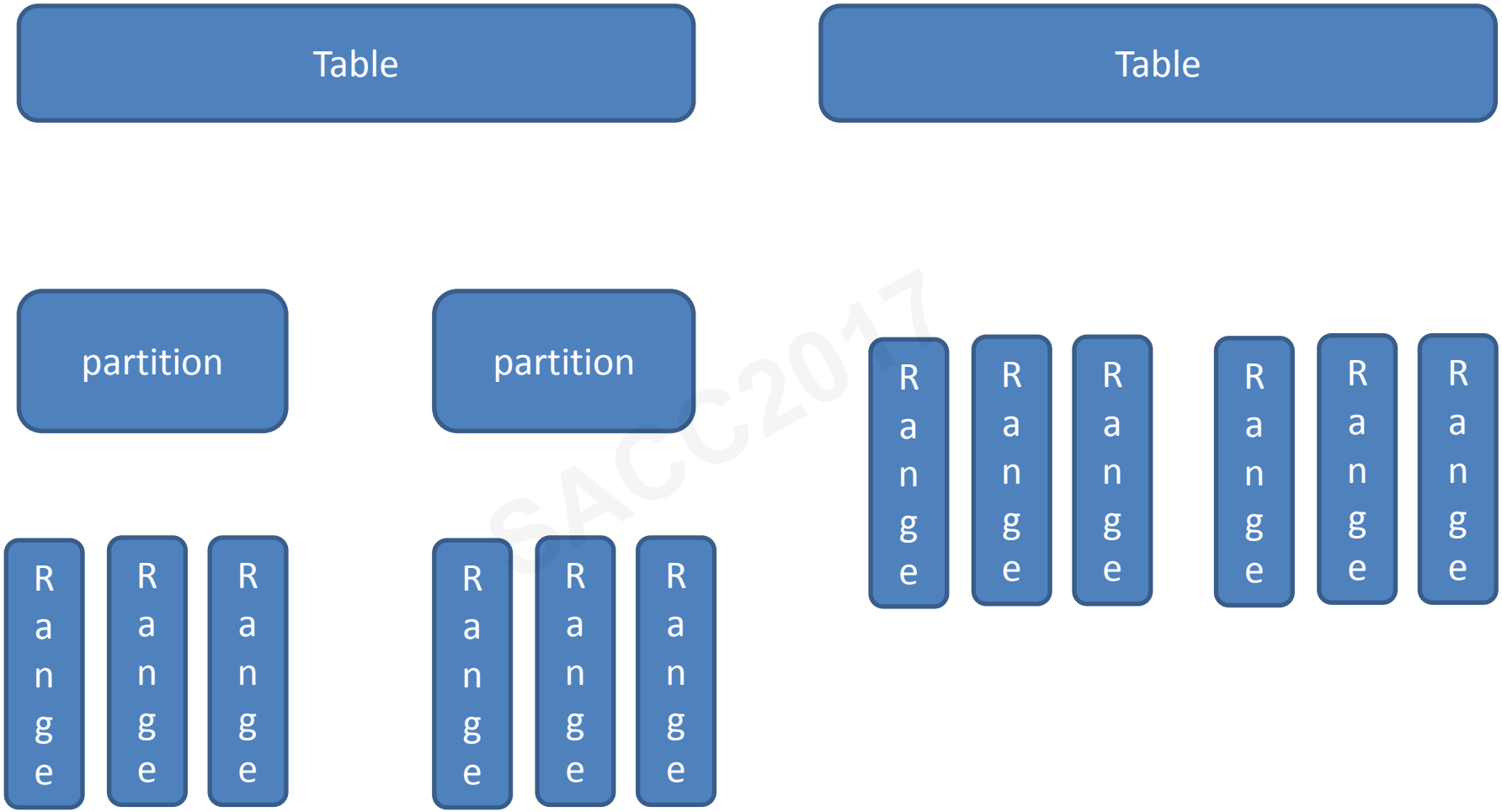
支持schema

支持模板列，插入时可以自动添加列

存储层： LSM-Tree(Log-Structured Merge Tree)

SACC2017

数据组织方式



SACC2017

Partition table

规则： Hash/list

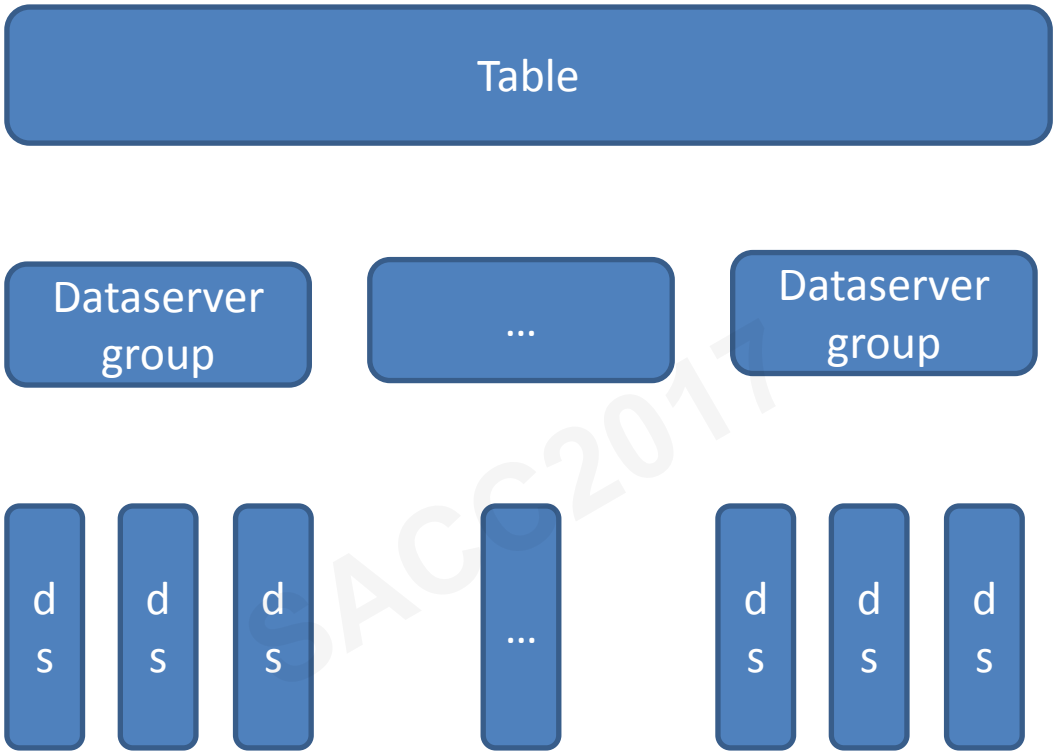
场景： 按key访问，或者单个partition内范围扫描

优点： 写入流量可以相对均匀分布

缺点：
不能全局范围扫描，
读取必须带有partition key，
兼容redis协议、partition second index等特性的Table，一个partition对应一个dataserver，
有容量限制，需要提前规划。

SACC2017

部署



SACC2017

Dataserver group

物理隔离：

同一个集群中，不同的table数据分布在不同的dataserver group中。

减少raft心跳连接数：

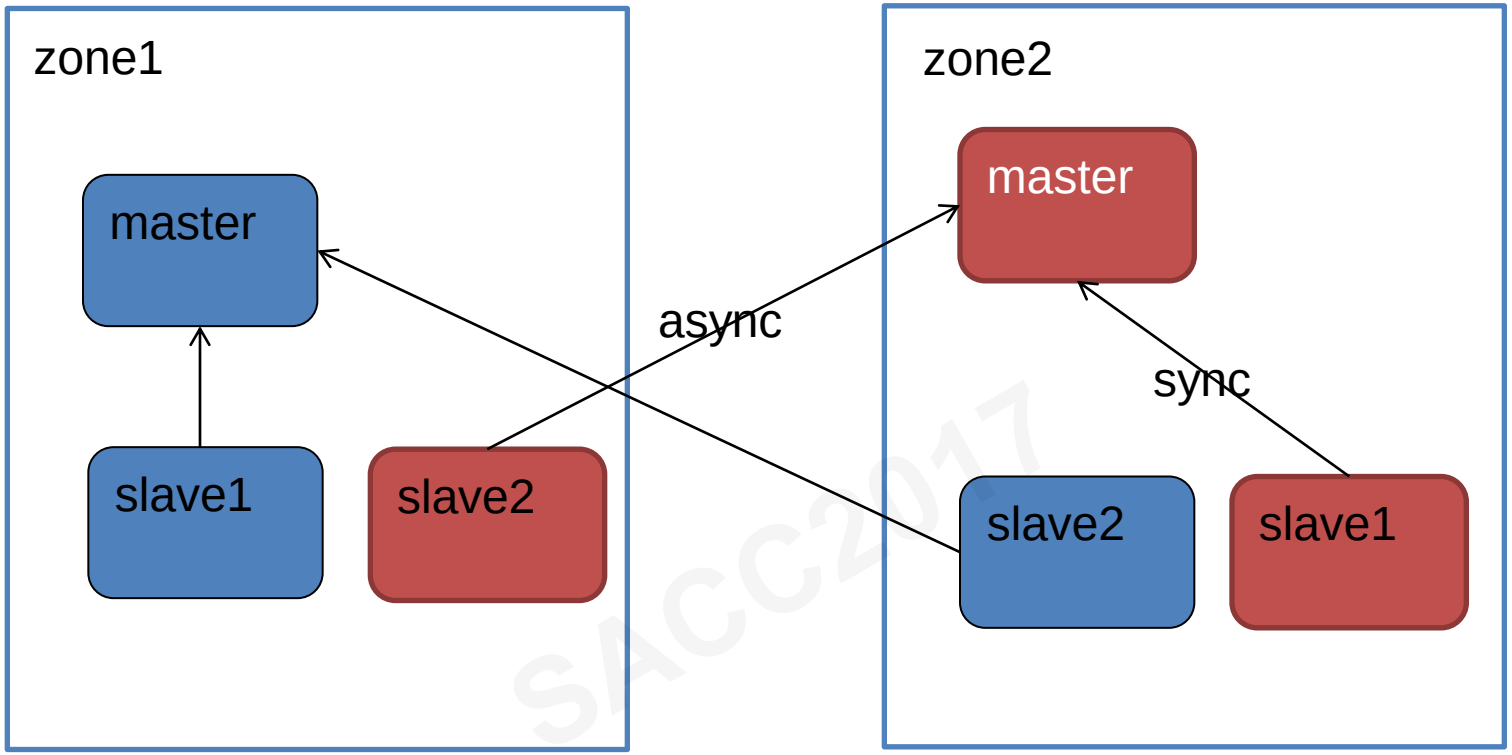
一个region的所有副本只能分布在一个dataserver group内，每个region是一个独立的raft复制组，raft心跳按照dataserver级别进行合并。

Partition跨机房分布：

不同的dataserver group分布在不同的机房内，partition与dataserver关联，实现不同的partition分布在不同的机房中。

SACC2017

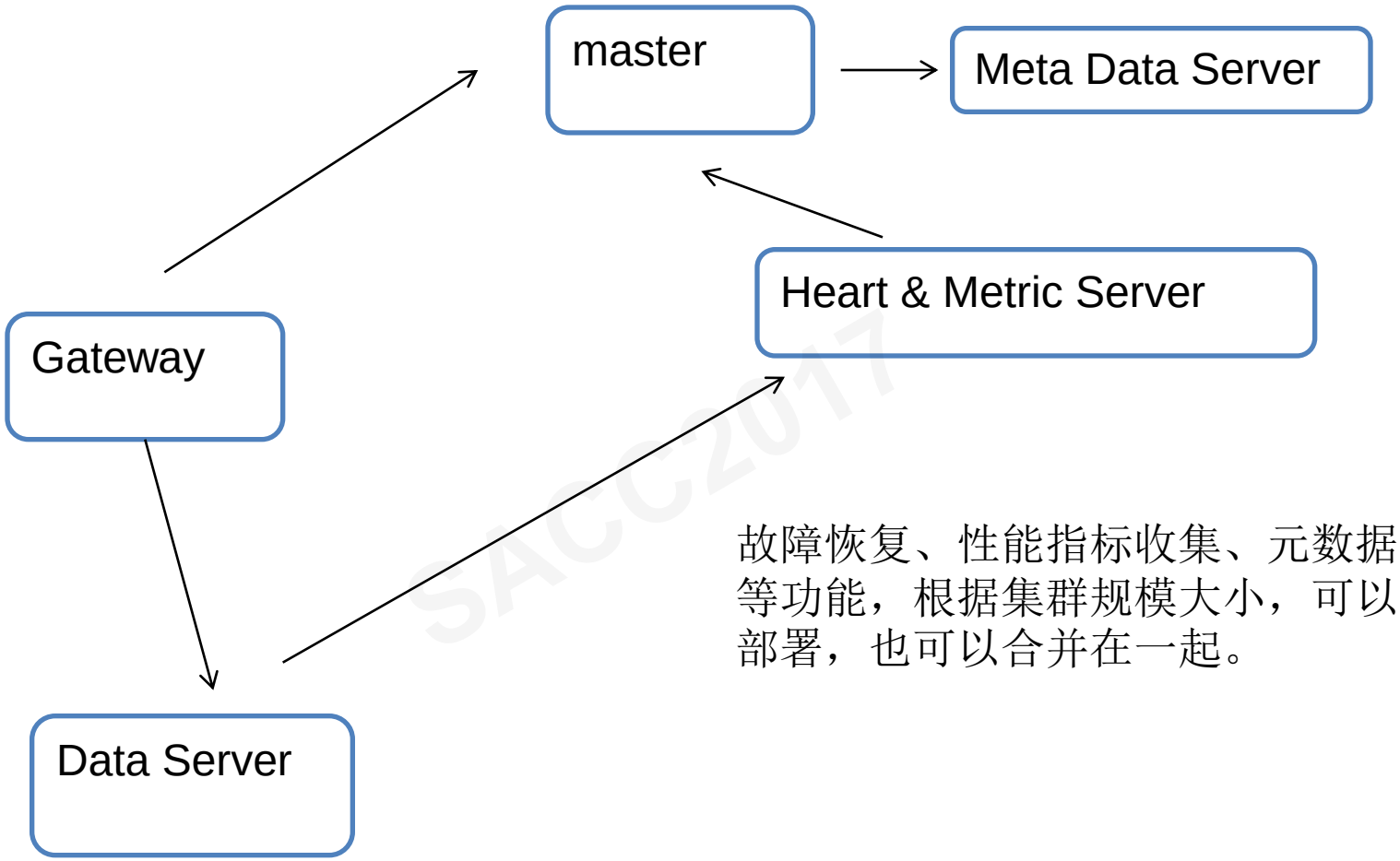
复制



复制方案的选择:
成本
性能
数据安全

Leader选举:
副本间选举
外部模块指定

心跳&元数据



故障恢复、性能指标收集、元数据管理等功能，根据集群规模大小，可以分开部署，也可以合并在一起。

SACC2017

缓存

块缓存

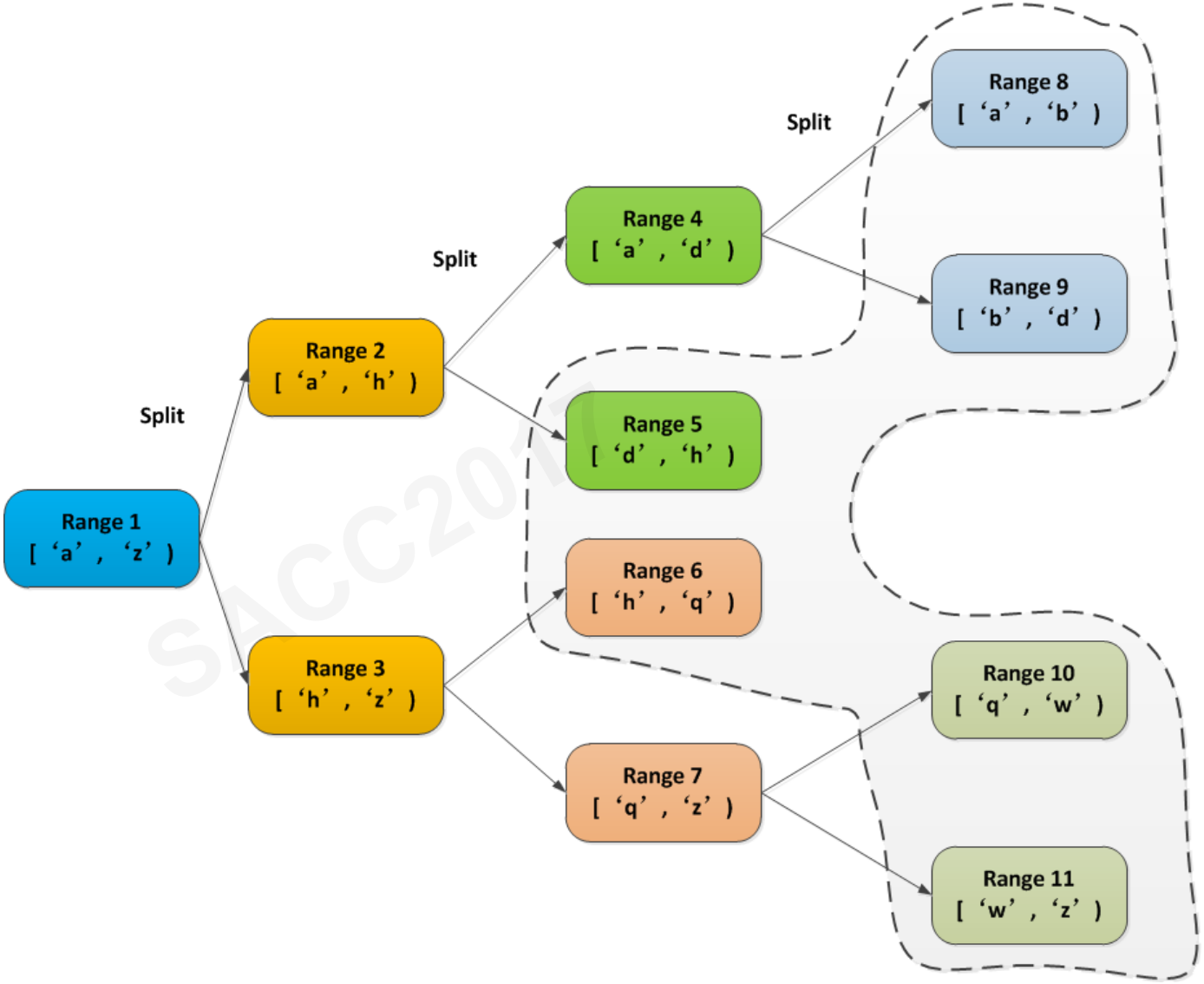
KEY缓存

按照hash规则进行分区的，需要开启KEY级别的缓存

SACC2017

分裂

本地分裂
文件引用
Compact阶段删除数据



SACC2017

TTL

存储层文件记录TTL的最近时间,避免不必要的扫描
读取时过滤

下一步优化：提供TTL分布数据

SACC2017

未来功能规划

- 支持redis数据结构
- 支持二级索引
- 支持事务

SACC2017

SACC 2017 云智未来 9th
IT168.com ChinaUnix IT PUB


THANKS