阿里云 奥运会全球指定云服务商

阿里云redis异地多活及冲突解决

耿倞

Agenda

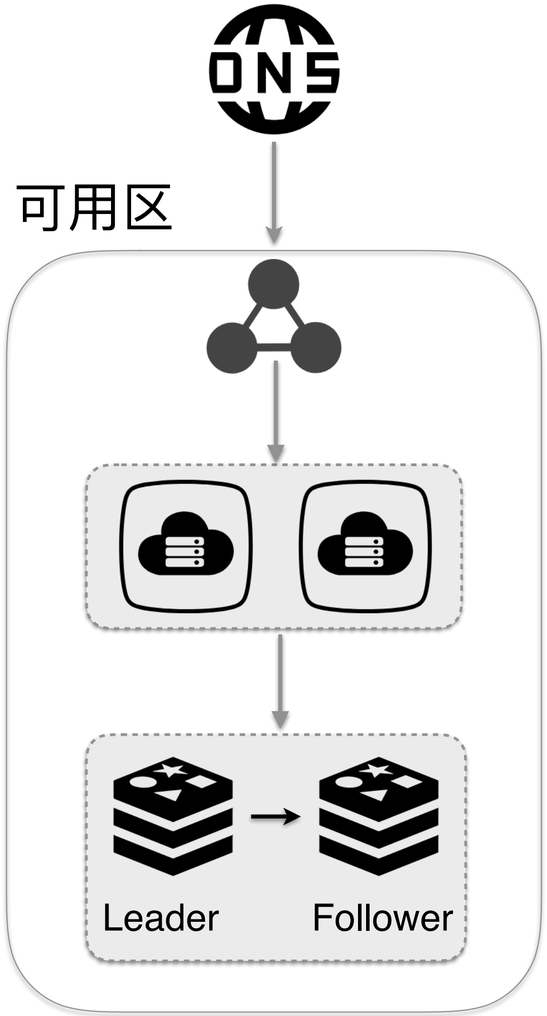
阿里云 奥运会全球指定云服务商

- 为什么需要异地多活
- 怎样实现异地多活
  - redis怎样做复制, 有哪些不足, 我们做了什么改进
  - 改进之后几个关键问题的解决
    - 断点续传, 不丢不重, 无环保证, 冲突解决
- 产品介绍 & 未来工作

异地多活: 容灾

阿里云 奥运会全球指定云服务商

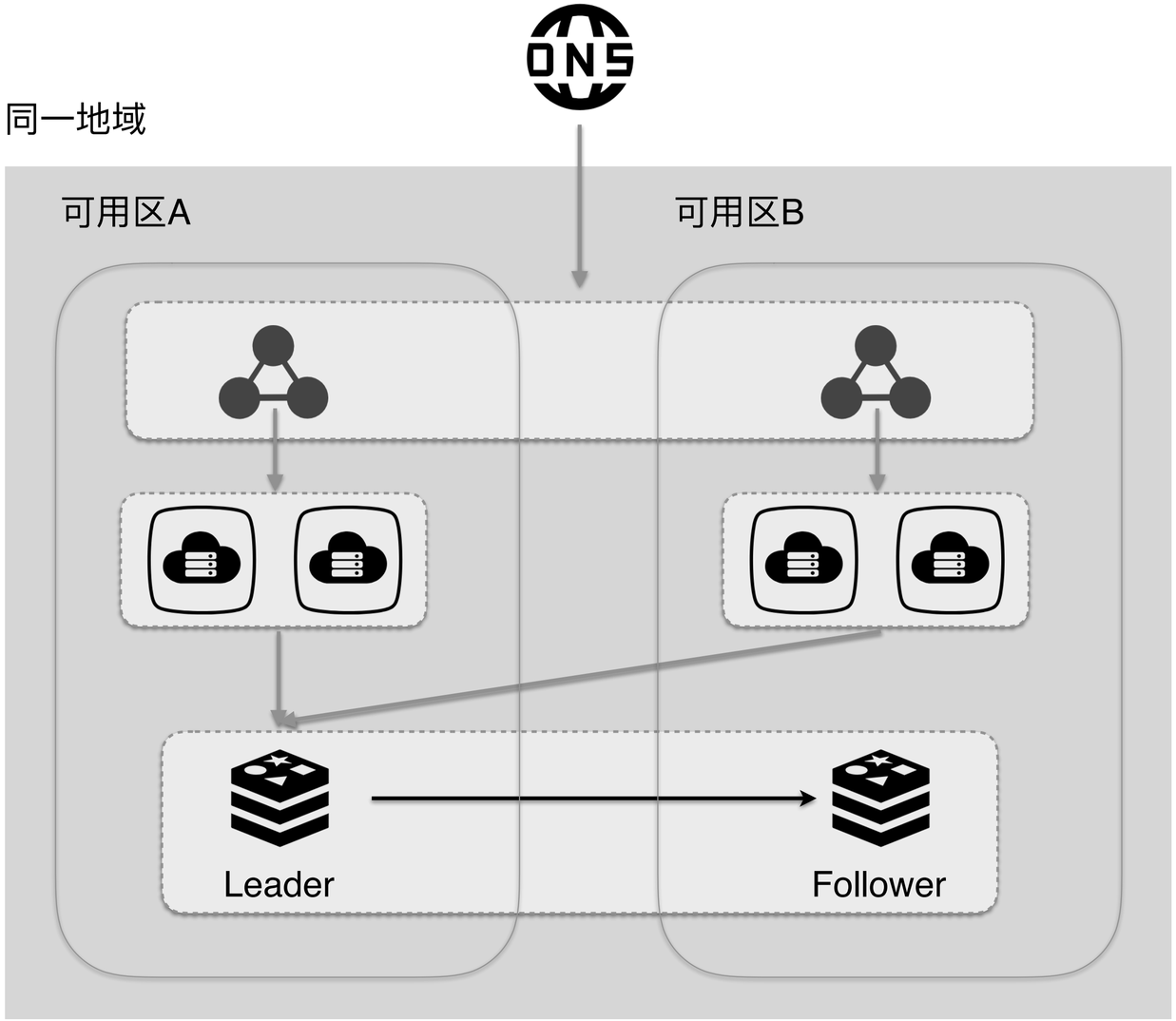
- 单机房



异地多活: 容灾

阿里云 奥运会全球指定云服务商

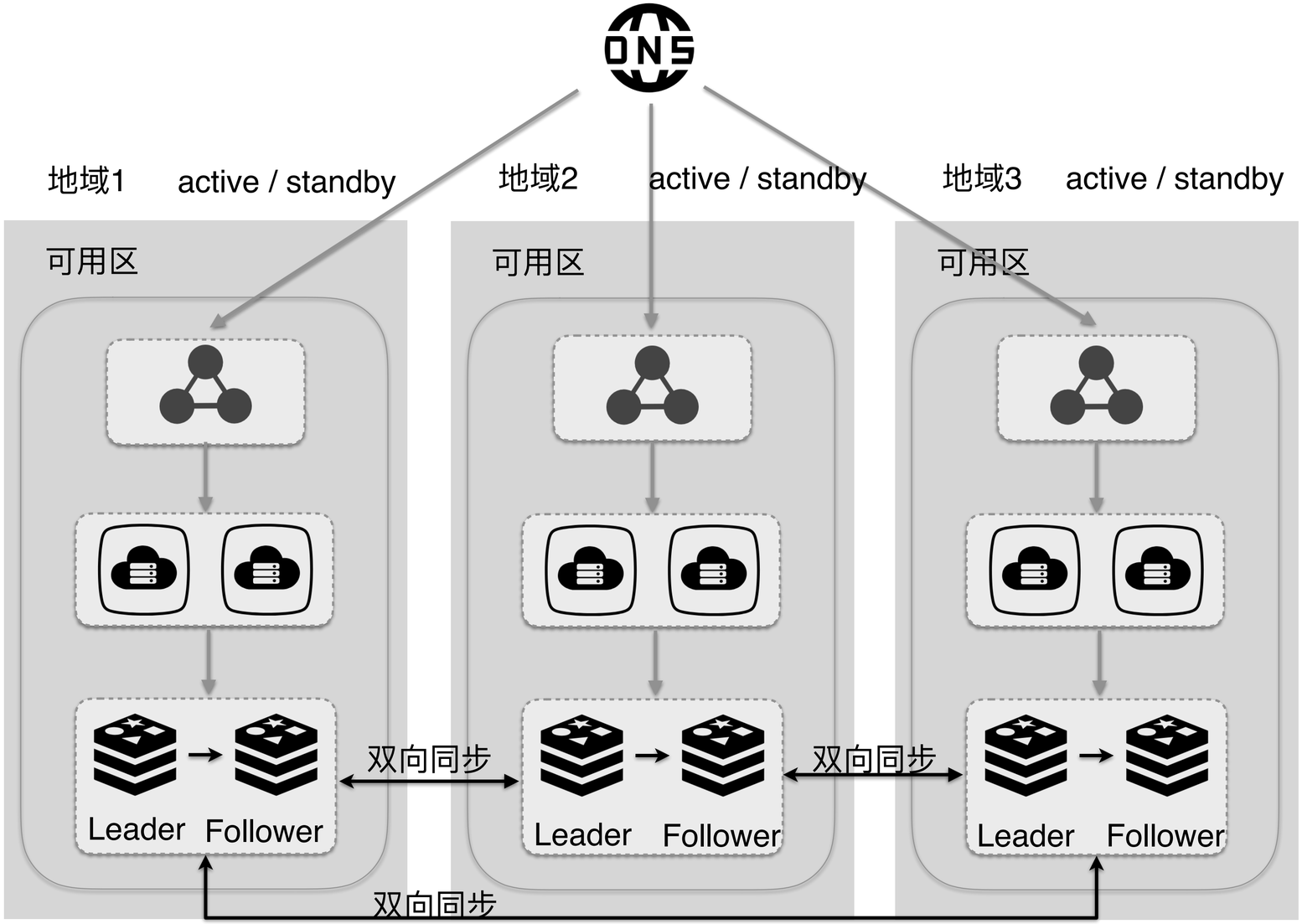
- 跨可用区



异地多活: 容灾

阿里云 奥运会全球指定云服务商

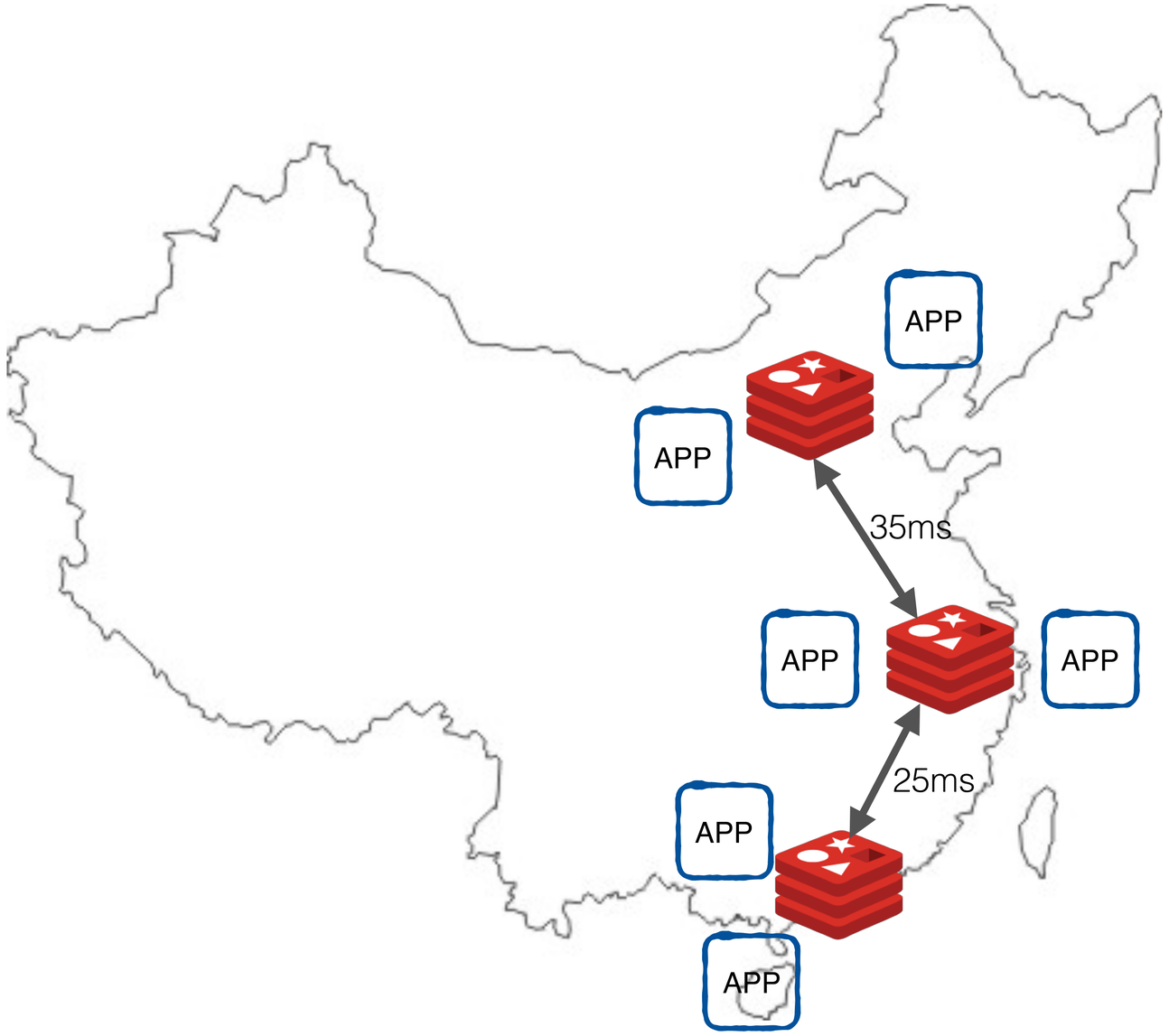
- 跨地域



异地多活: 性能

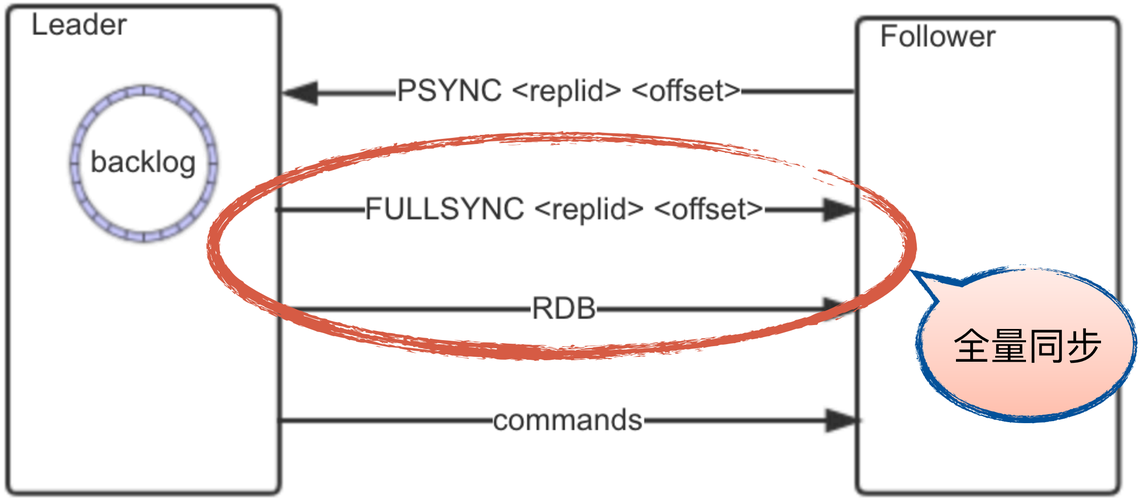
阿里云 奥运会全球指定云服务商

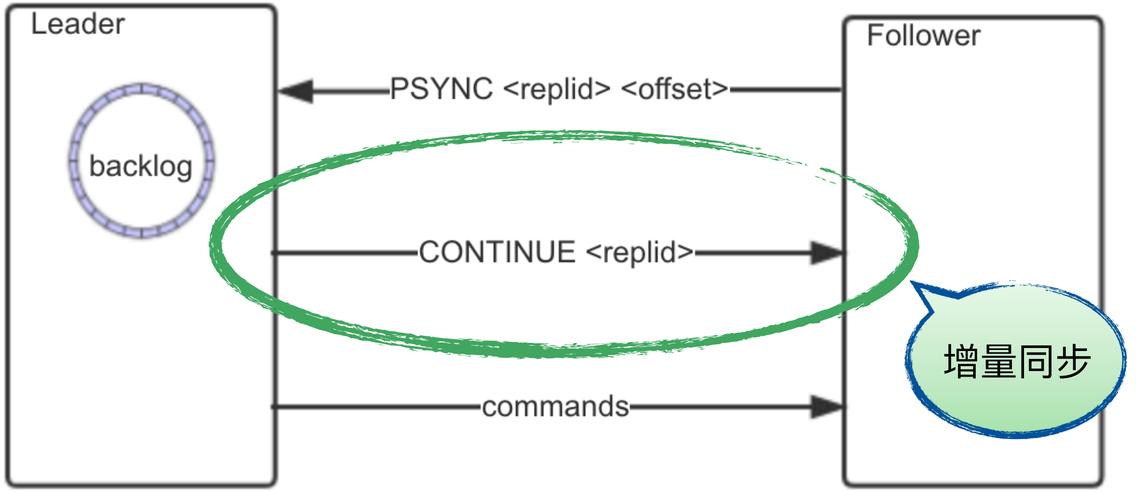
- 就近访问



Redis Replication回顾

- 正常同步: 略
- 断点续传
  - 2.8以前: SYNC
    - 断点后重新全量
  - 2.8到4.0间: PSYNC
    - slave持有replid和offset
    - 网络闪断后可以续传
  - 4.0: PSYNC2
    - replid和offset持久化入RDB
    - 重启/主备切换后可以续传





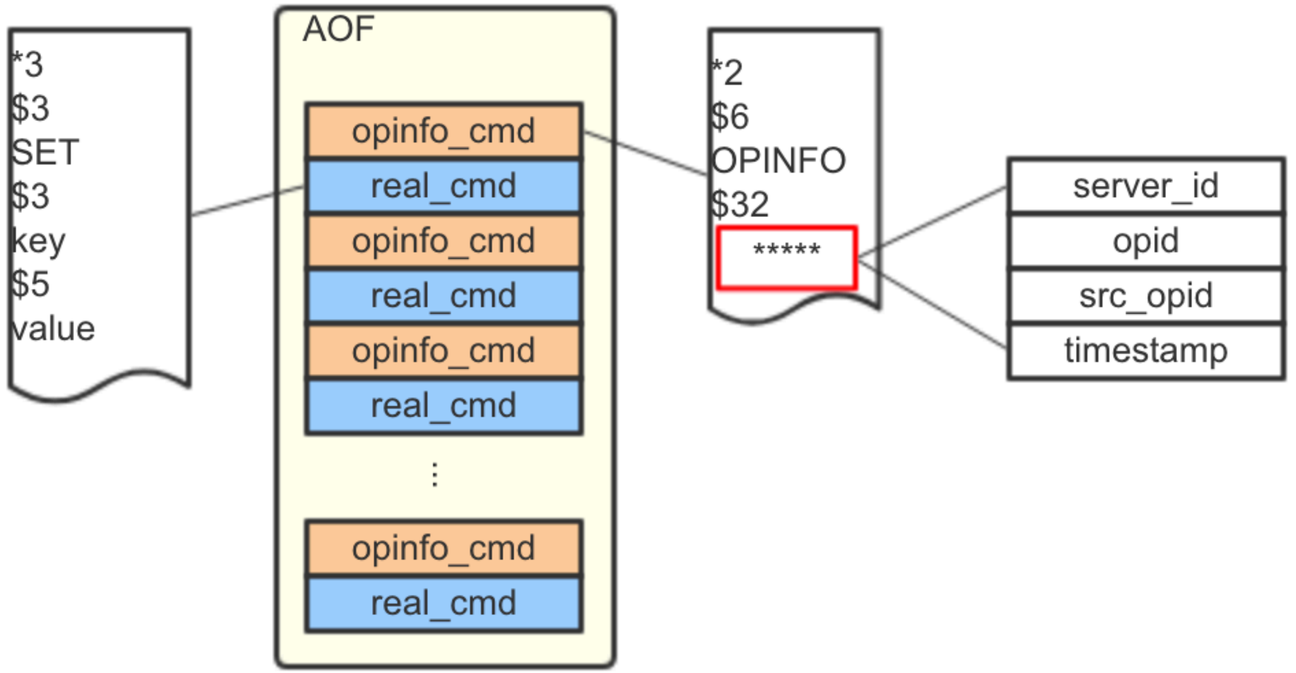
Redis Replication不足

阿里云 奥运会全球指定云服务商

- 断点续传能力有限
  - backlog大小有限
  - 全量重传不经济
  - 双活时无法进行全量同步
- 无法双向同步
  - 无法去环

改造: 扩展AOF

- 为每条执行命令扩展元信息
  - **新增server id:** 标识写入来源
  - **新增opid:** 单调递增表序
- 实现:
  - **real_cmd:** 与原有AOF一致
  - **opinfo_cmd:** 记录oplog元信息的命令
    - OPINFO <header>
    - <header>为元信息结构体的序列化
  - server_id及opid持久化入RDB



server_id	opid	src_opid	Op
A	101	-1	Cmd1

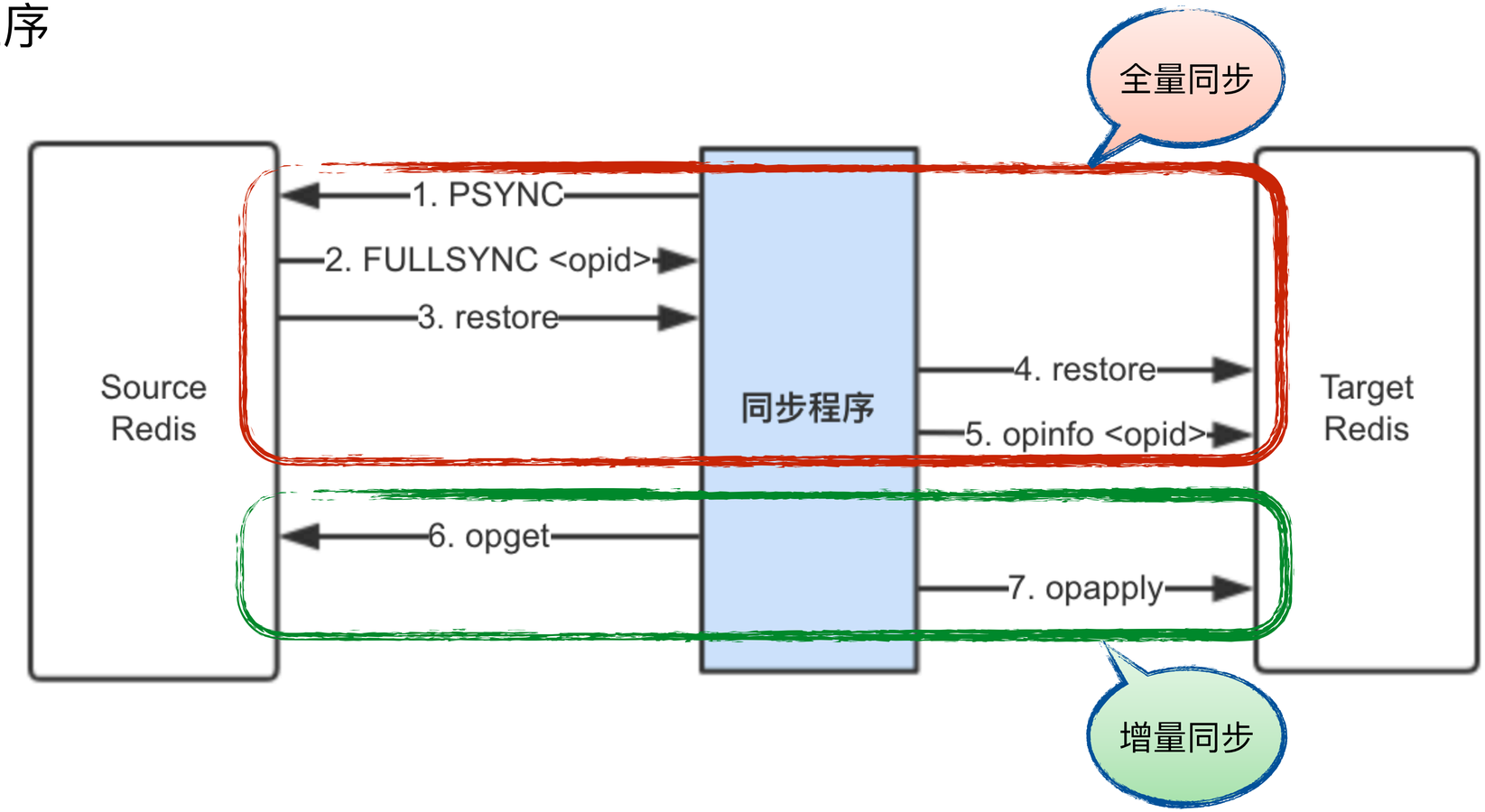
server_id	opid	src_opid	Op
B	20	-1	Cmd2

server_id	opid	src_opid	Op
A	101	-1	Cmd1
B	102	20	Cmd2

改造: 扩展同步模式

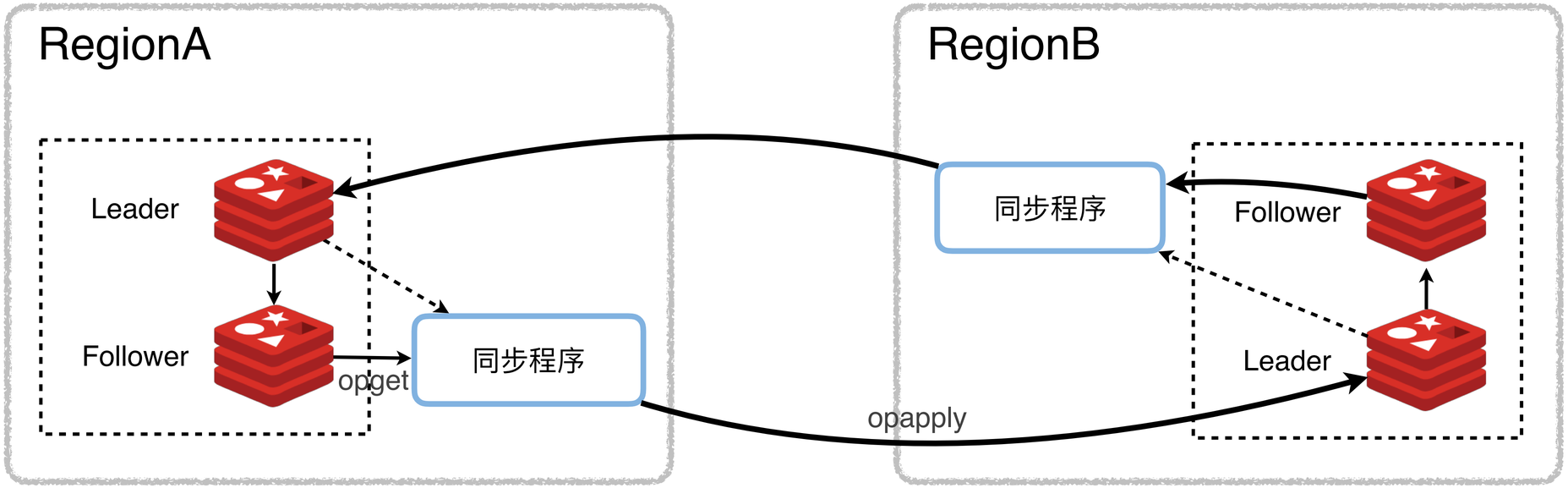
阿里云 奥运会全球指定云服务商

- 主从之间: 原生模式
- 实例之间: 新增命令 & 同步程序
  - OPGET <server_id> <count>
    - 批量抓取
  - OPAPPLY
    - 开启pipeline模式应用



整体架构

- 同步程序:
  - 解耦内核逻辑
    - 自适应主备切换, 备库重搭
    - 限流, 加密, 压缩, 监控等
  - 集群模式下可随节点个数做横向扩展
  - 无状态



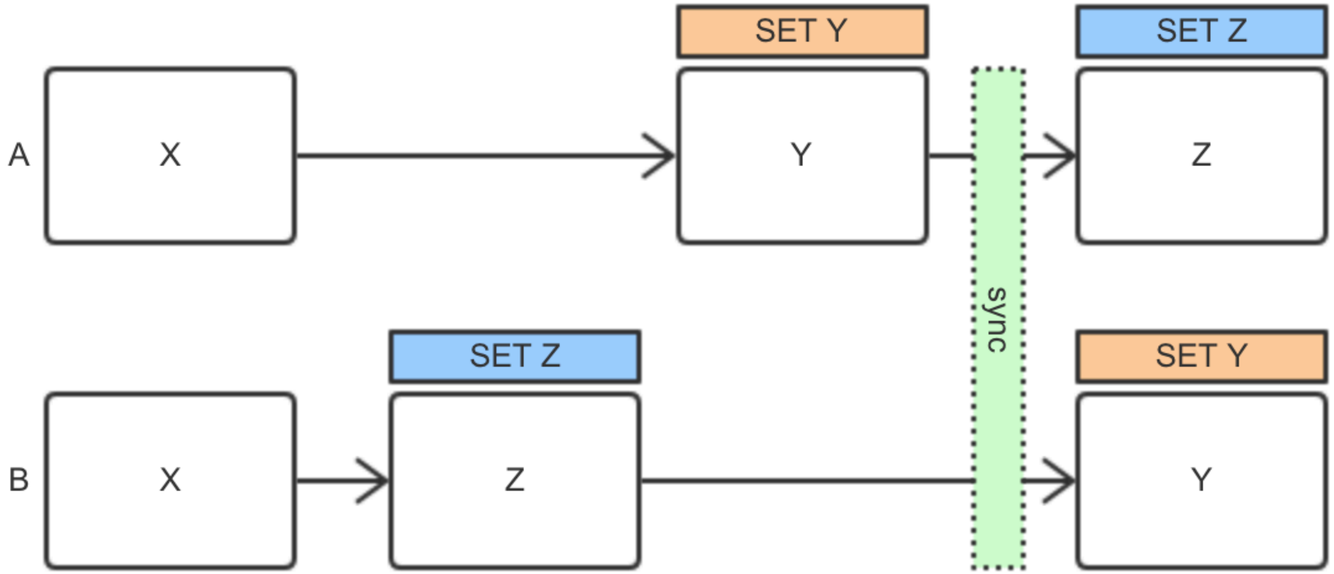
回顾几个关键问题

- 断点续传
  - server_id + opid, 记录于redis
- 不丢不重 (Exactly Once)
  - 根据opid顺序同步
  - server_id + opid确保不做重复应用
- 无环 (Prevent Loop)
  - 同步程序只过滤指定server_id的log

多活真正痛点: 冲突解决 (Conflict Resolution)

阿里云 奥运会全球指定云服务商

- 产生原因:
  - 多点写入情况下同时对同一数据进行操作



- 解决方案:
  - 从业务, 系统, 产品三个层面展开

冲突解决——业务层面

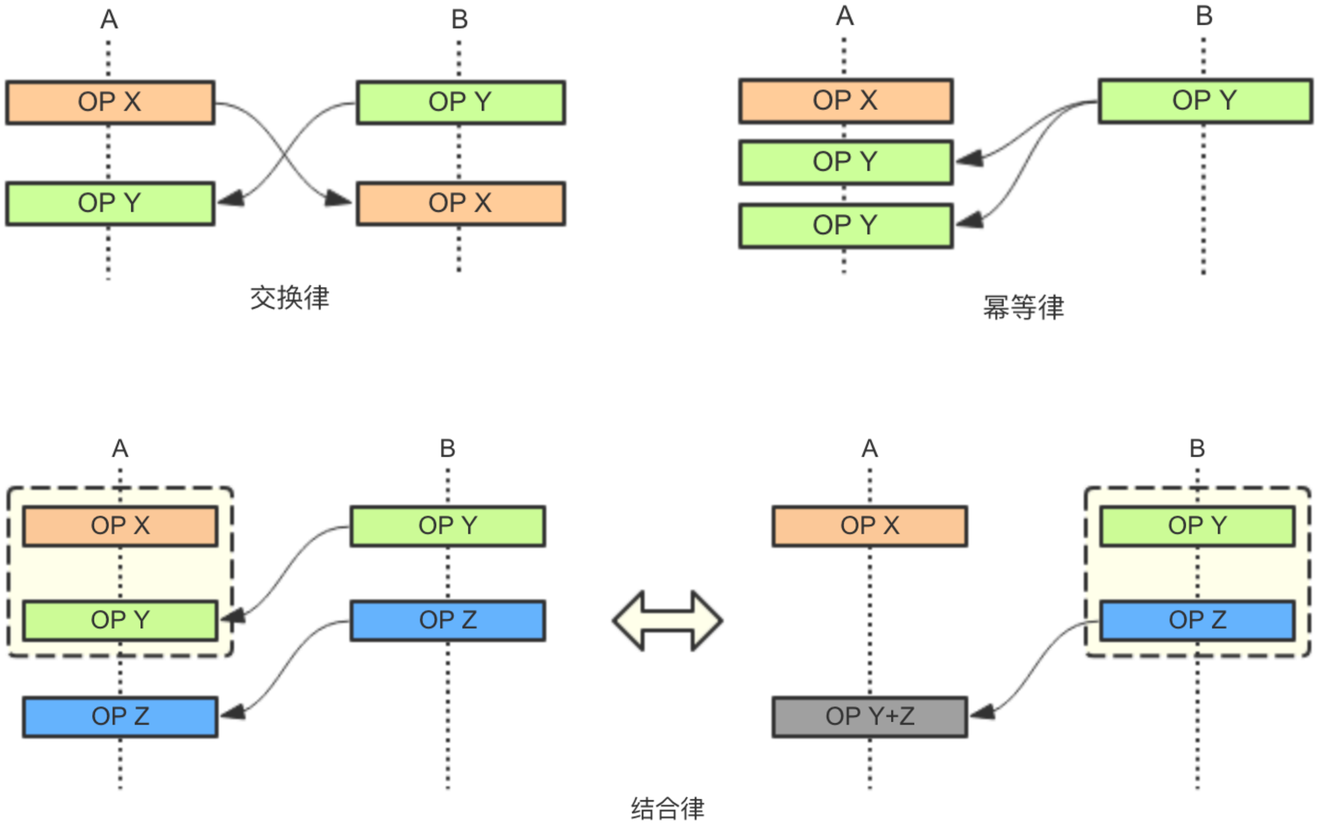
阿里云 奥运会全球指定云服务商

- 能否避免?
  - 数据同步为异步过程: 无法第一时间进行冲突探测
  - 冲突解决方案与业务预期存在差异
- 业务层面解决:
  - 业务层面可以容忍冲突
  - 一写多读(Read local, Write global)
  - 数据分离(Read local, Write partitioned)
    - eg: 给不同Region的user id带上不同前缀

冲突解决——系统层面

阿里云 奥运会全球指定云服务商

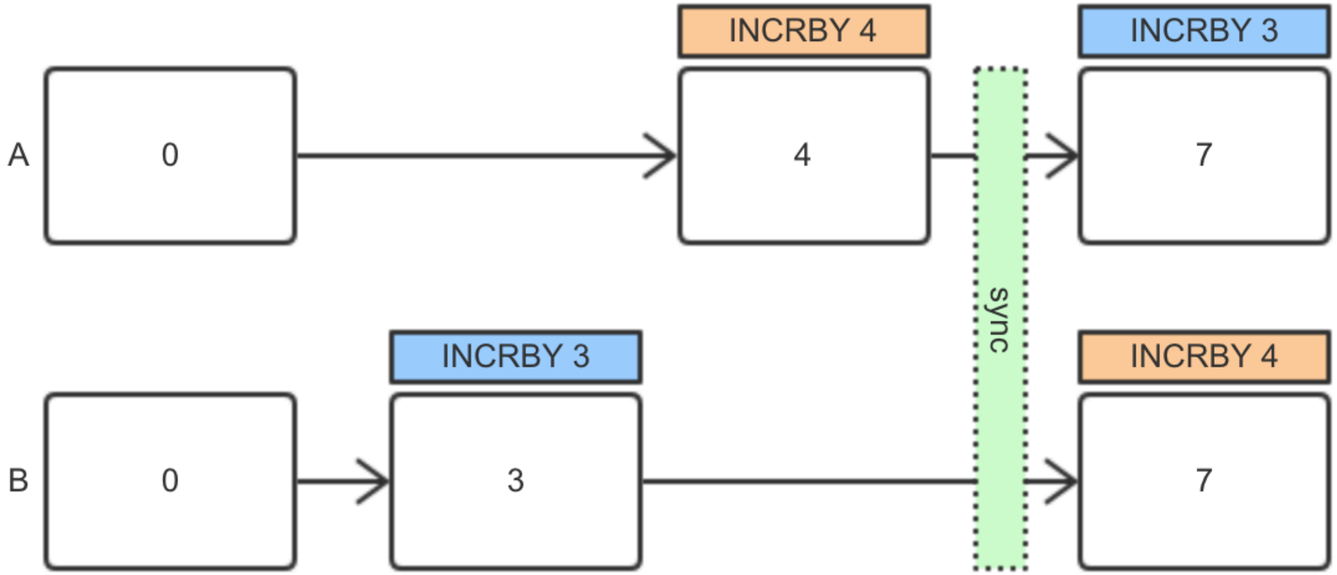
- CRDT (Conflict-Free Replicated Data Type)
  - 原理:
    - 交换律, 结合律, 幂等律
    - 或附带额外元信息使其满足
  - 支持数据结构:
    - 计数器 (Counter)
    - Memory Cell (Register)
    - 集合(Set)



CRDT: Counter

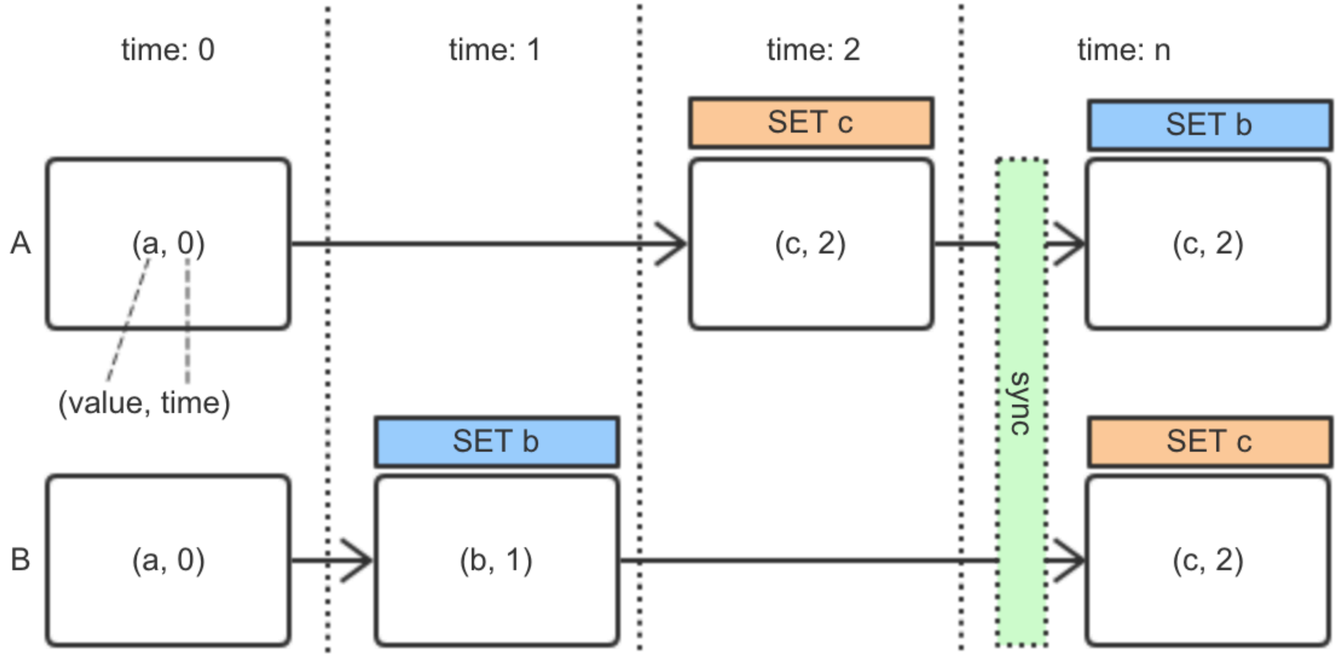
阿里云 奥运会全球指定云服务商

- 操作: 加减法
  - 天然满足交换律, 结合律
- Redis对应操作: INCR, DECR, INCRBY, DECRBY



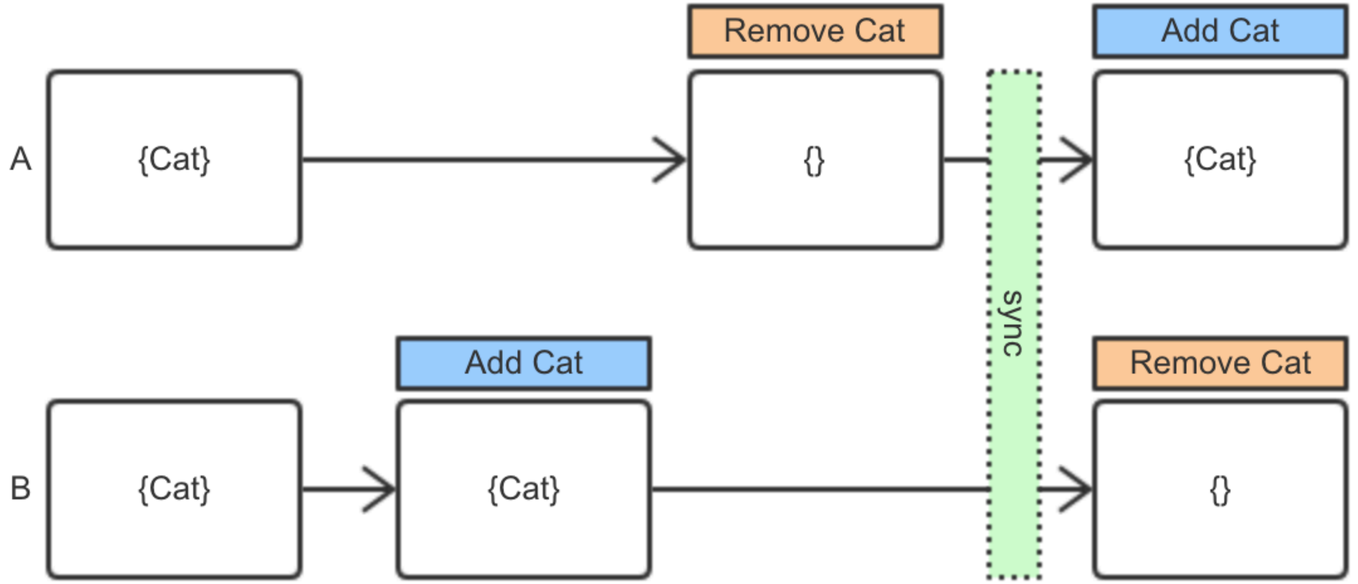
CRDT: Register

- 操作: Set
  - 不满足交换律, 附加时间戳(Last Write Wins)
  - 时间戳单调递增
- Redis对应操作: SET



CRDT: Set

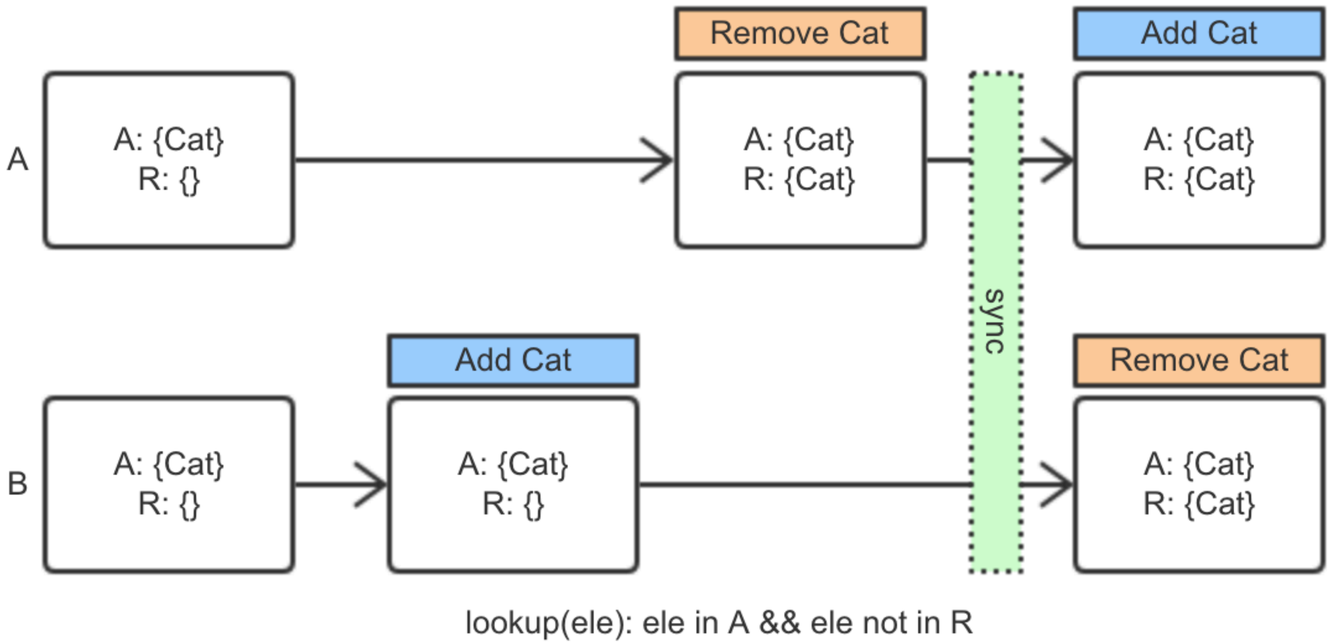
- 操作: Add, Remove
  - Add本质: 求并集(Union), 满足交换律、结合律
  - Add & Remove不满足交换律



CRDT: Set

阿里云 奥运会全球指定云服务商

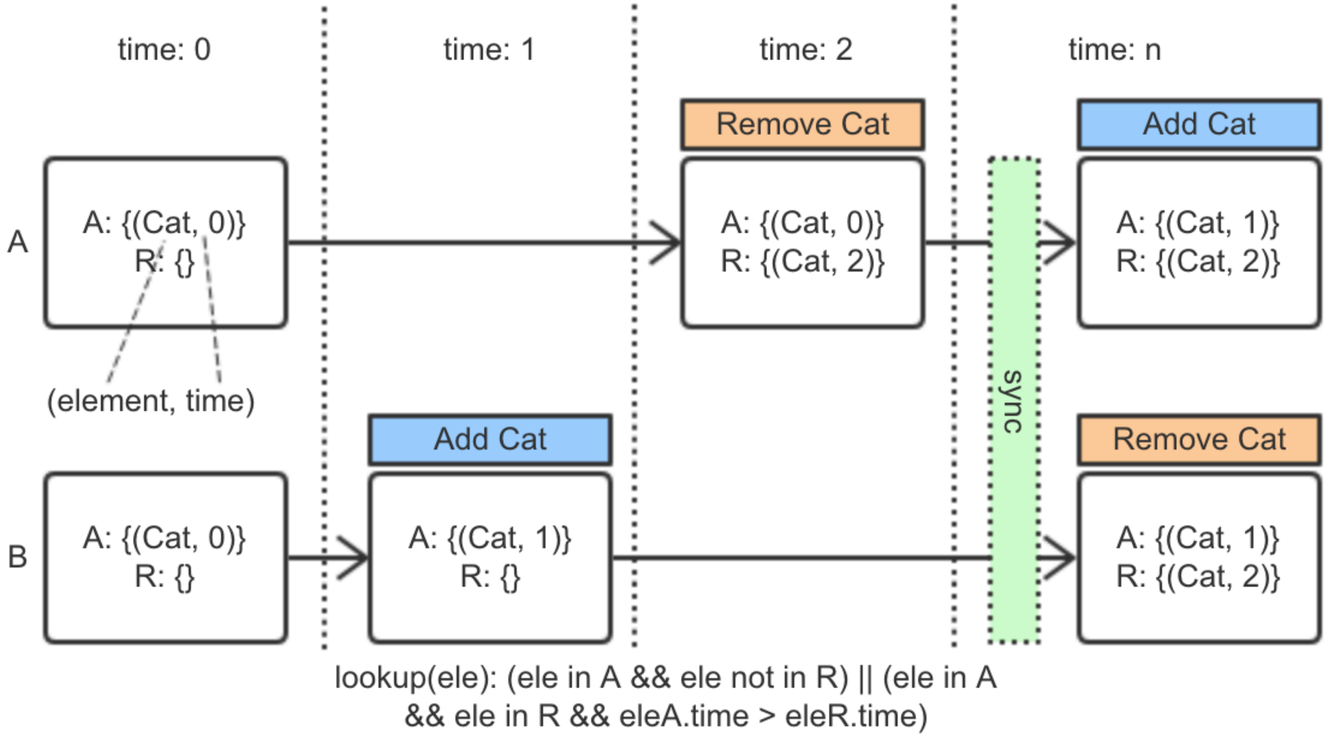
- 2P-Set
  - 将Remove转化为Add: Add和Remove都是单独的Set
- 问题
  - Remove的元素不能再Add
  - 两个Set只会不断膨胀



CRDT: Set

- LWW-Set (Last Write Wins Set)
  - 给每一个元素带一个时间戳

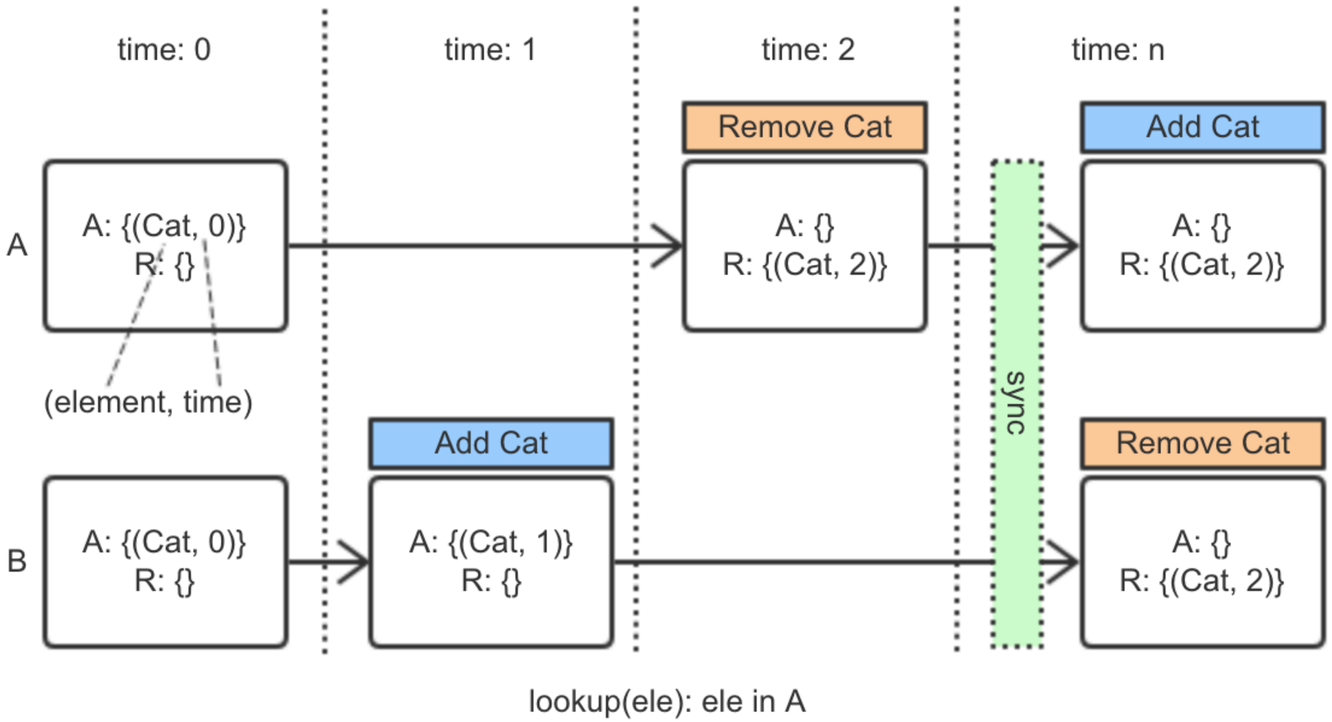
- 问题
  - Remove的元素不能再Add 解决
  - 两个Set只会不断膨胀 未解决



CRDT: Set

阿里云 奥运会全球指定云服务商

- 工程实践最佳
  - 保证A ∪ R = ∅
  - 定期对R进行回收



CRDT: Set

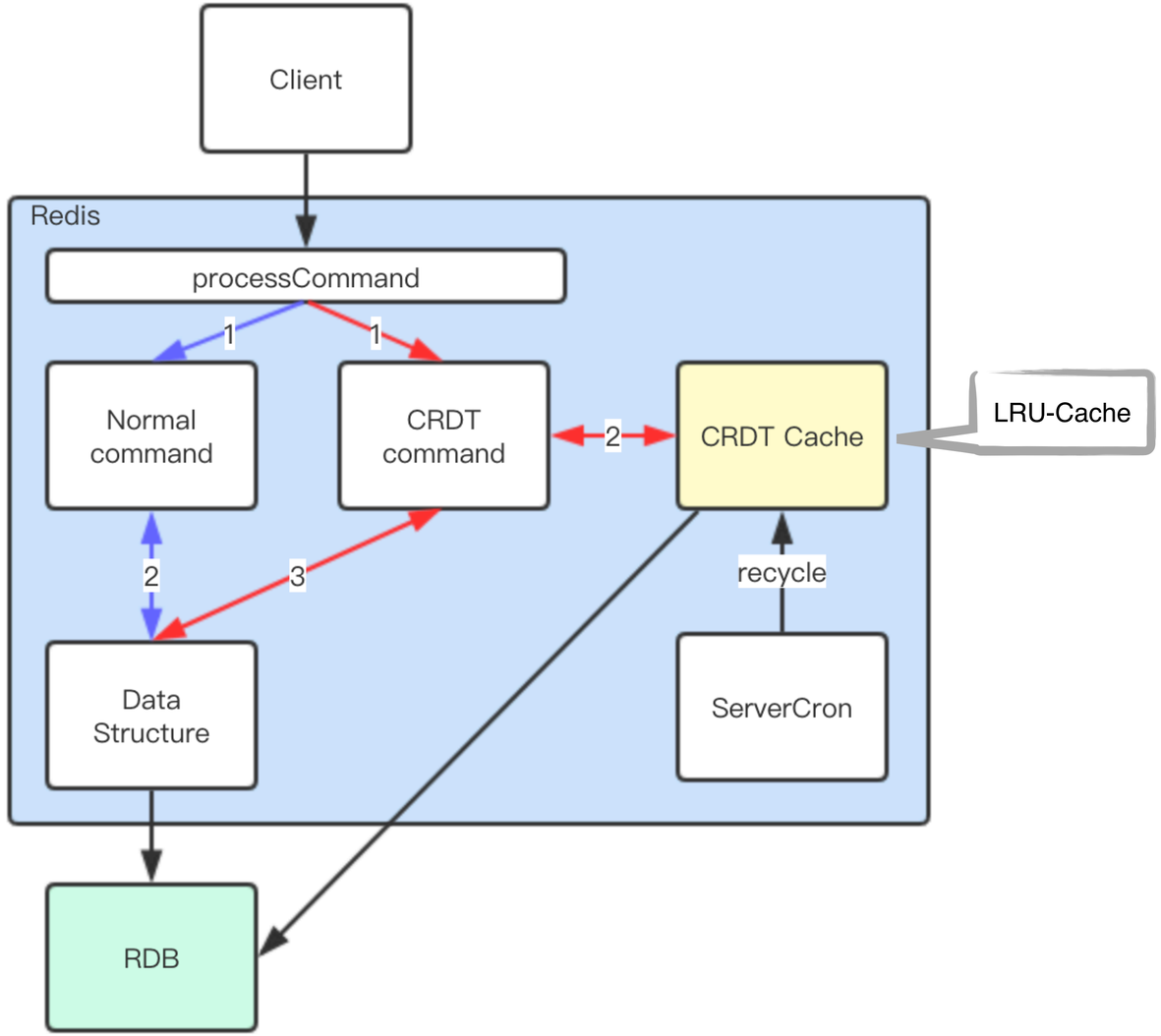
阿里云 奥运会全球指定云服务商

- Redis对应数据结构和操作:
  - Set: SADD, SREM, SPOP…
  - Hash: HSET, HDEL
  - SortedSet: ZADD, ZREM …
  - String: SET, DEL …

CRDT实现

阿里云 奥运会全球指定云服务商

- redis内核改造



Redis CRDT

阿里云 奥运会全球指定云服务商

数据类型	当前支持操作	场景
Counter	INCR, DECR, INCRBY, DECRBY, INCRBYFLOAT	点赞数, 收藏数等
String	SET, DEL	全局基本信息: session等
Set	SADD, SREM, SPOP	购物车, 收藏夹等
Hash	HSET, HMSET, HDEL / HINCRBY, HINCRBYFLOAT	全局session信息
SortedSet	ZADD, ZDEL / ZINCRBY	和时间序列相关: 微博, timeline
HyperLogLog	PFADD	全局近似uv
GEO	GEOADD	全地域地理位置信息统计

冲突解决 ——产品层面

阿里云 奥运会全球指定云服务商

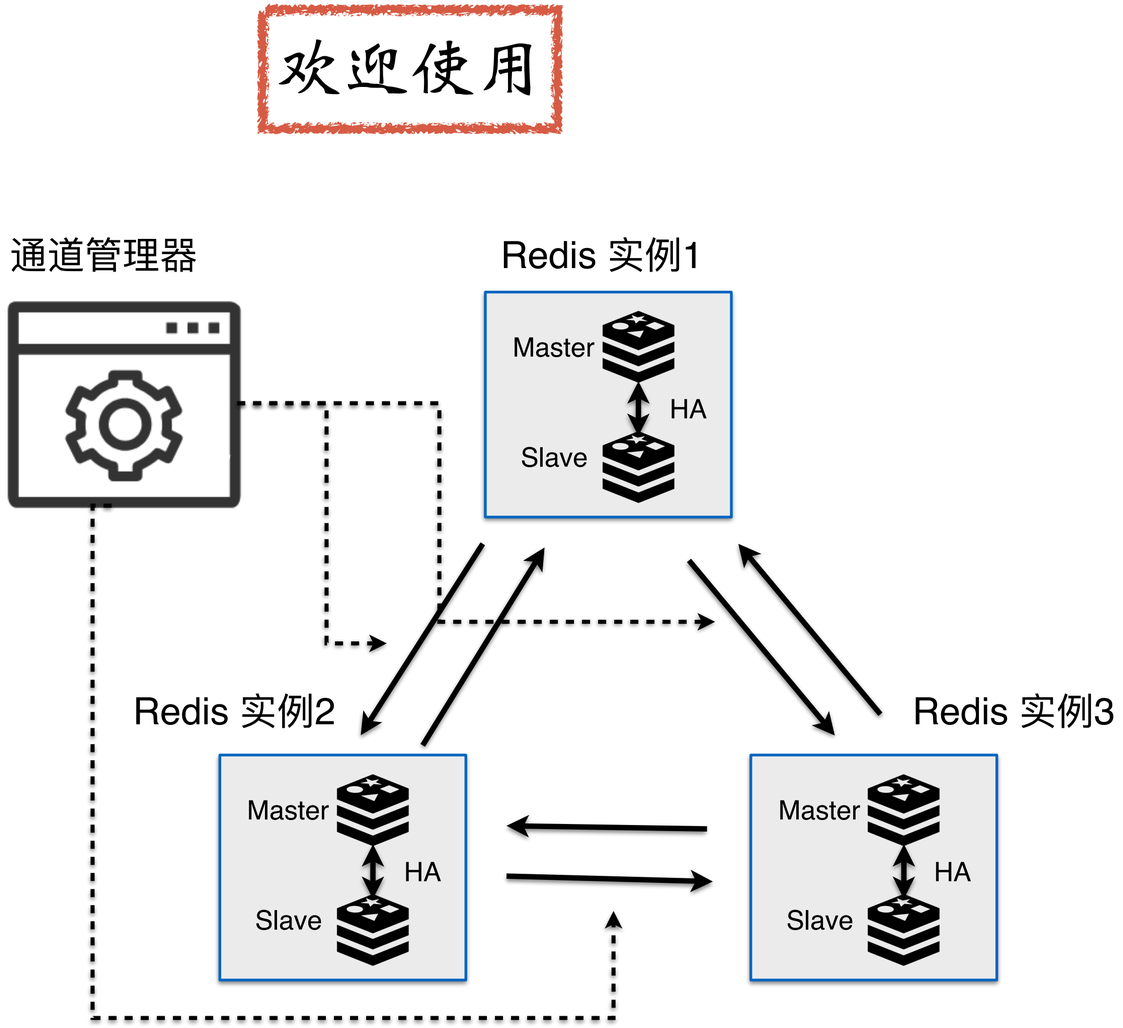
- CRDT非冲突解决的银弹
  - 非所有类型(List), 非所有写命令(SUNION)
  - 不同类型: HSET和SADD同一个key冲突无法解决
  - 同一类型不同命令: INCR和SET无法确保一致

- 产品级解决方案
  - **冲突**结果记录
  - 数据**一致性**校验
    - 定时分级全量校验
    - 实时增量校验

冲突	同步时非redis原因写失败
不一致	同步成功, 但最终数据不一致

redis全球多活产品

- 使用场景: 多活, 灾备, 迁移
- 支持规格
  - 标准版, 集群版, 读写分离版
- 特点
  - 高可用
    - 容忍超长断路
    - 自适应主备切换
  - 高吞吐, 低延迟
    - 标准版10万TPS, 集群版线性扩展
    - 国内 < 100ms, 跨洲际 < 1s

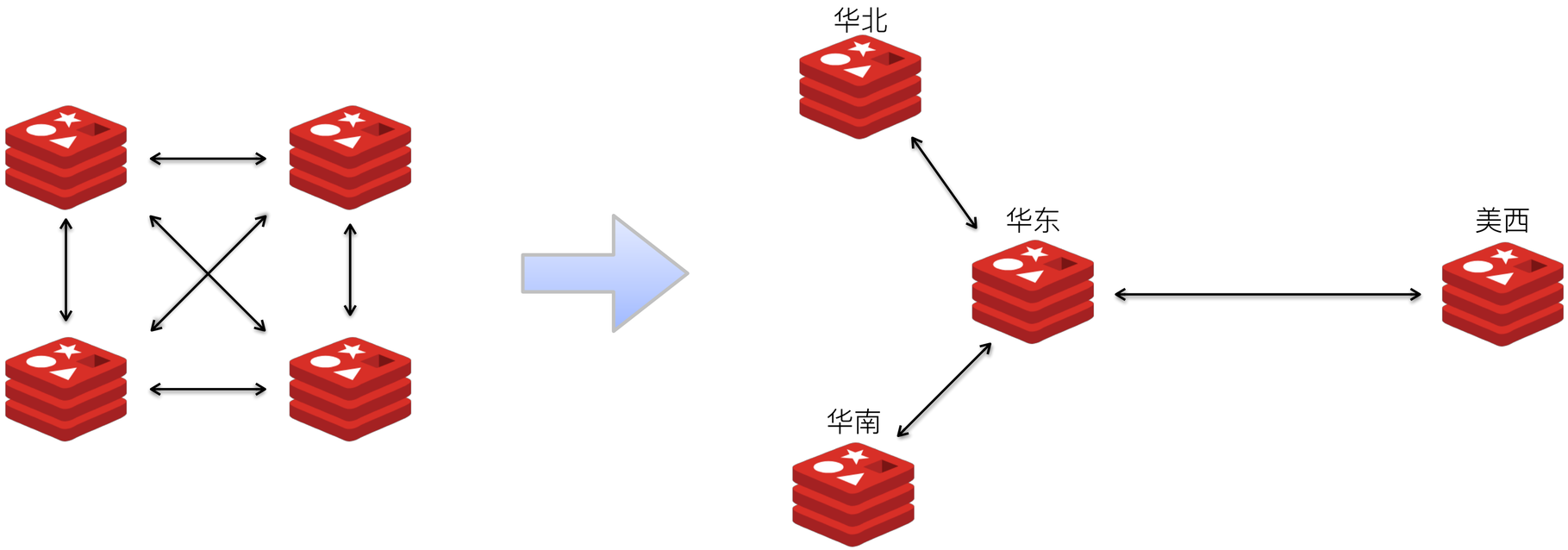


未来工作

阿里云 奥运会全球指定云服务商

- 更多数据库引擎
  - MySQL, PolarDB, MongoDB …
- 支持拓扑定制

敬请期待



本PPT来自Redis技术交流群2018年线下交流会

YouTube：https://youtu.be/lsX5qbosfE8

欢迎交流redis的开发和运维，群主每日精选一篇redis有关的文章在群里分享。所有每日分享的文章都记录在以下星球（免费）。希望入群的请扫描二维码加我的微信

Q & A