中国数据库联盟 All China Database Union

墨天轮

2023 10/14

ACDU·中国行

数据库前沿技术探索及应用之路

中国·成都

ACDU·中国行

中国数据库联盟 All China Database Union

墨天轮

构建可靠MySQL服务：RPO=0实现方案分析

演讲人：冯光普

2023.10

冯光普

中国数据库联盟 All China Database Union

墨天轮

- 多点 DMALL 数据库负责人
  - MySQL、TiDB、OB、Redis、MongoDB
  - 数据库平台
  - DB中间件，双活架构
- 更早，阿里巴巴数据库AliSQL团队

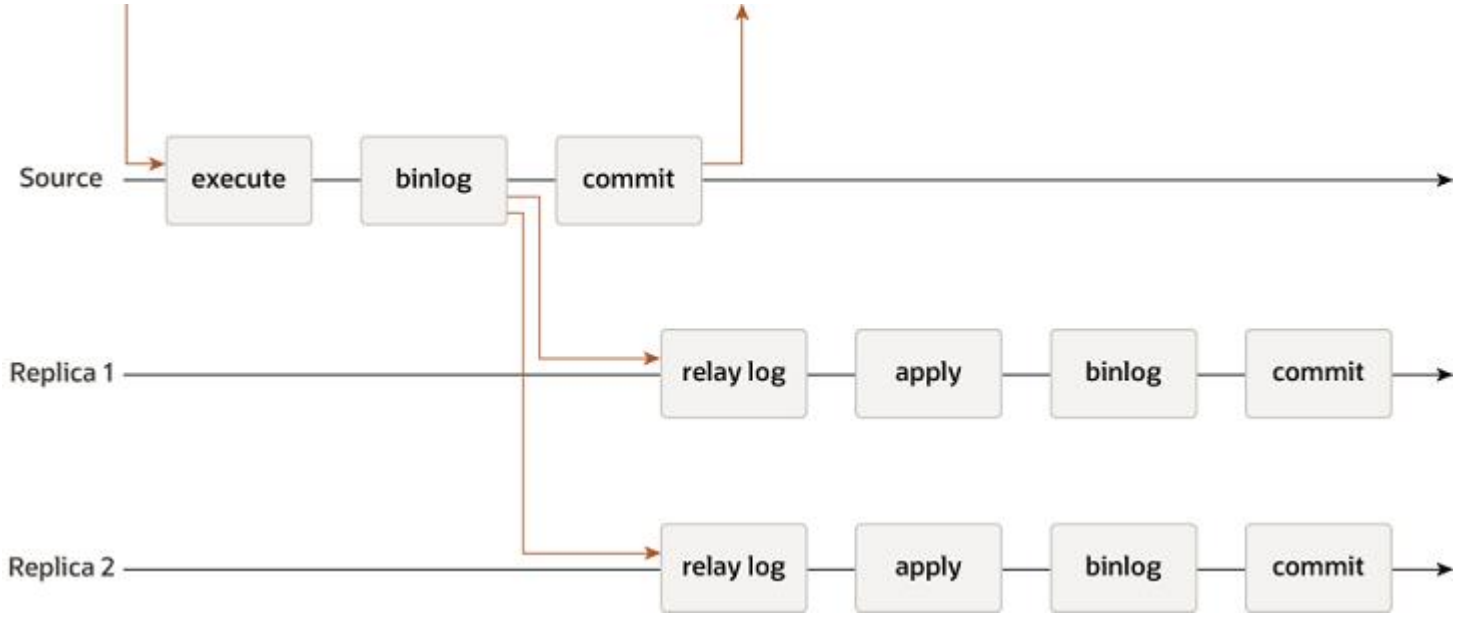
目录 CONTENTS

RTO & RPO

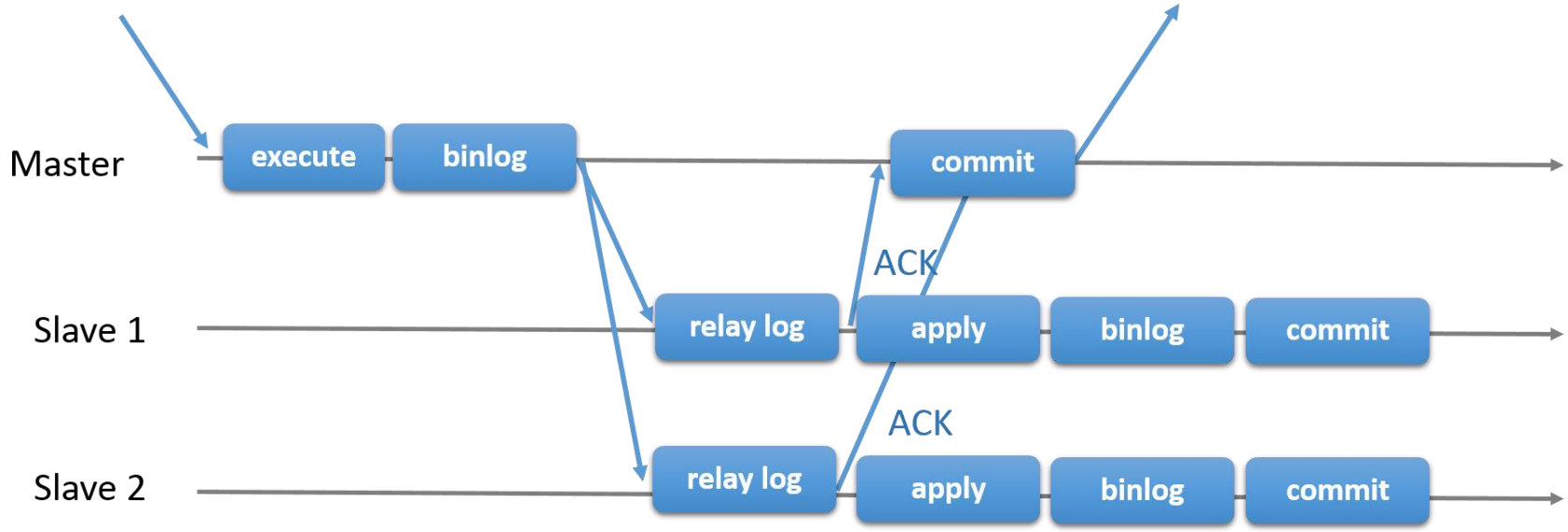
- **RTO（恢复时间目标）**
  - 单位是时间，恢复可用需要的时长
- **RPO（恢复点目标）**
  - 单位是时间，最多可能丢失的数据

异步复制
主从有延迟
无法保障RPO

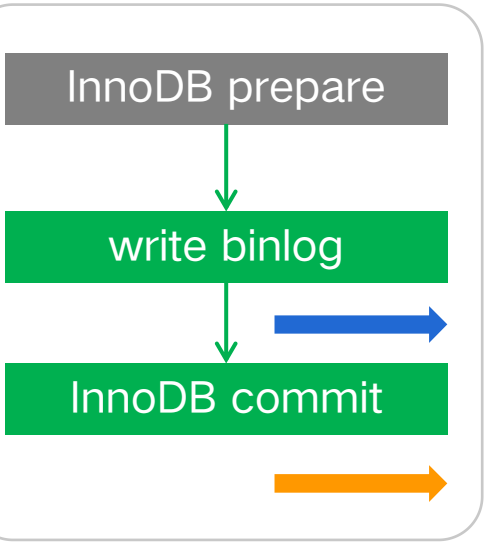
主库性能最佳



半同步复制丢数据问题（1/3）



相对异步复制
极大降低了丢数据可能
但不是100%

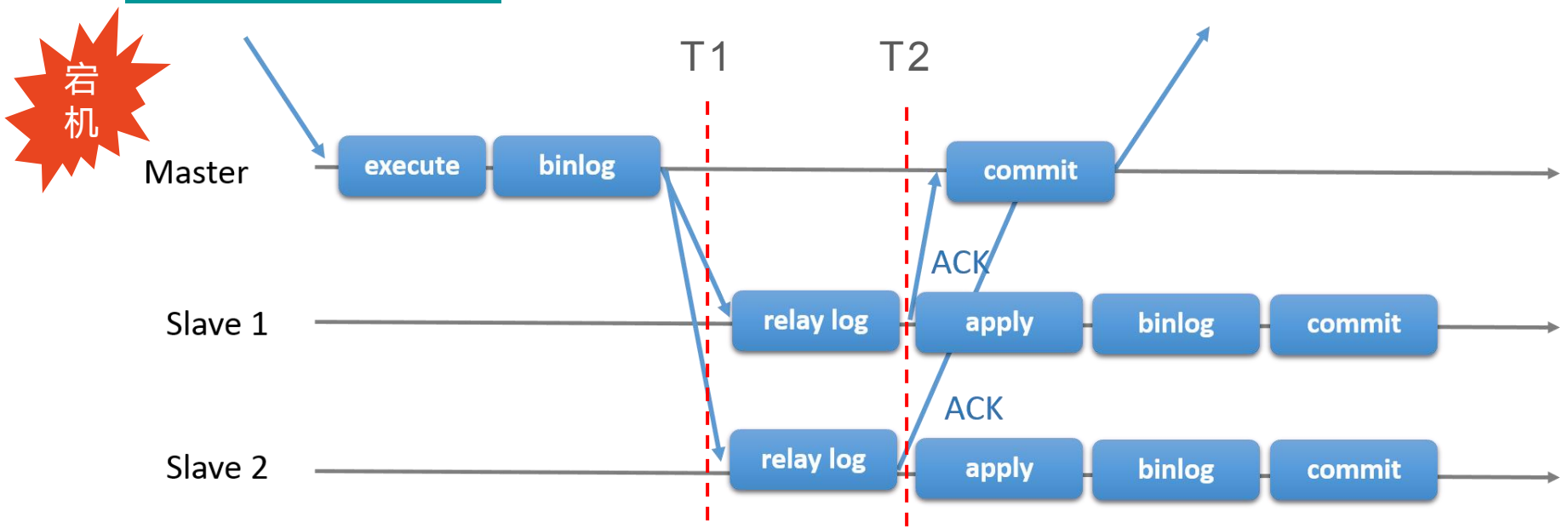


InnoDB Crash Recovery：一旦write binlog完成，事务会被commit

AFTER_SYNC：write binlog后，备库未写relay，主库crash recovery后会多数据

AFTER_COMMIT：主库上数据实际已提交（数据可见），只是不返回客户端

半同步复制丢数据问题（2/3）



从客户端视角来看
可能【多了数据】

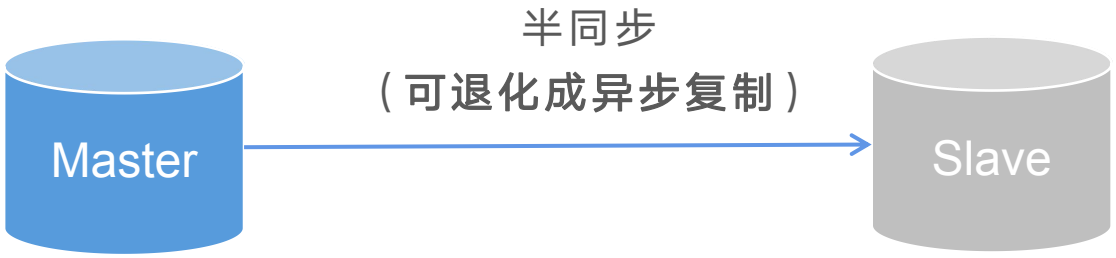
若Master不可恢复

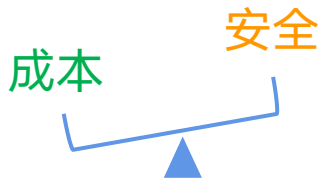
- T1时刻宕机：在切换前后，客户端看到了一致的数据
- T2时刻宕机：在切换后，客户端看到了更多的数据

半同步复制丢数据问题（3/3）

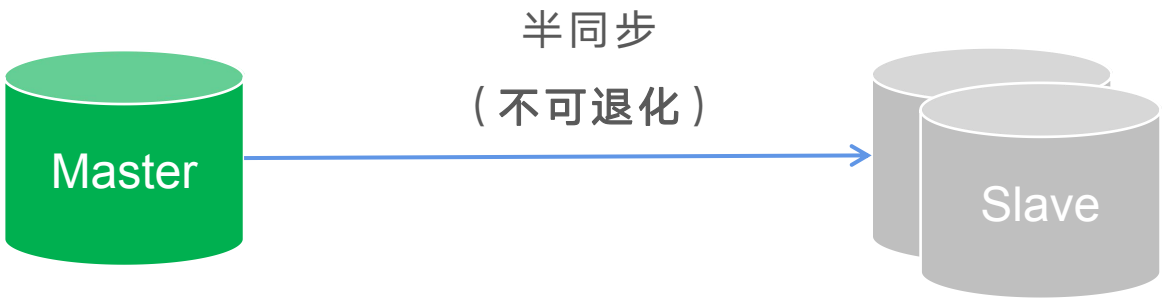
中国数据库联盟 All China Database Union　墨天轮

rpl_semi_sync_master_timeout 参数控制





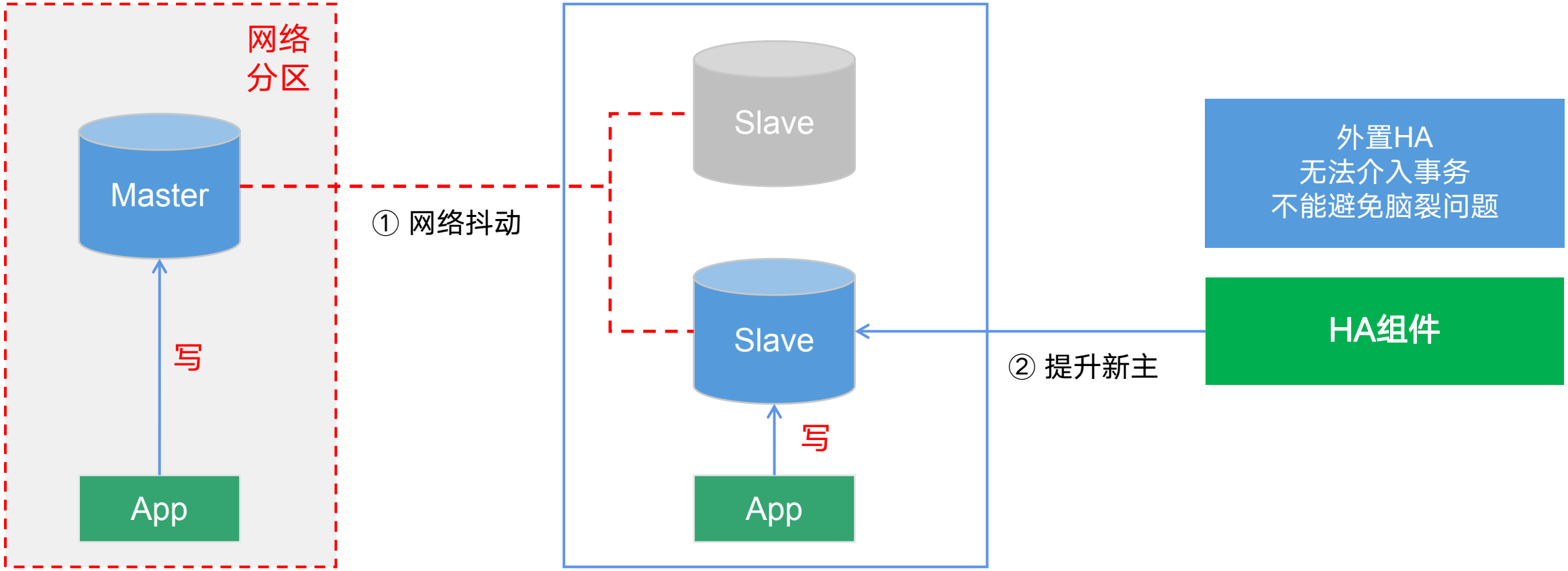
云RDS-主从两节点
允许退化成异步
可能丢数据



3节点
不允许退化成异步
单个ACK即可

实现RPO=0：至少3节点集群（2+ slave）：可允许1个slave不可用，不影响主库事务提交

外置HA脑裂问题



- 网络抖动超过HA探测阈值，选出了新主库
- 各网络分区内部，App与数据库仍保持连接，并可写入

RPO=0之MySQL方案

- 基于分布式一致性协议（Paxos / Raft）同步数据
  - **收益**：数据强一致性、避免脑裂双写、内置HA自动切换
  - **代价**：至少3副本、事务延迟增加、复杂度
  - 是未来的趋势

- Share Everything
  - Aurora
  - PolarDB
  - TDSQL-C
  - GaussDB

- 云厂商的解决方案

- Share Nothing
  - TiDB
  - OceanBase
  - MySQL Group Replication

- 数据库厂商的解决方案

RPO=0之MySQL方案

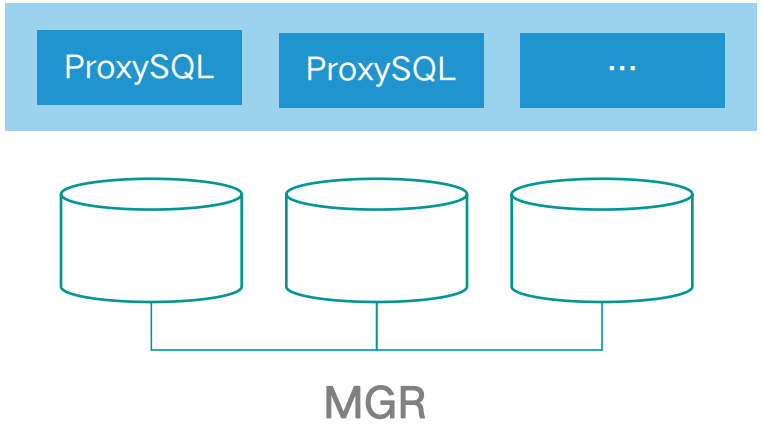
中国数据库联盟 All China Database Union　墨天轮

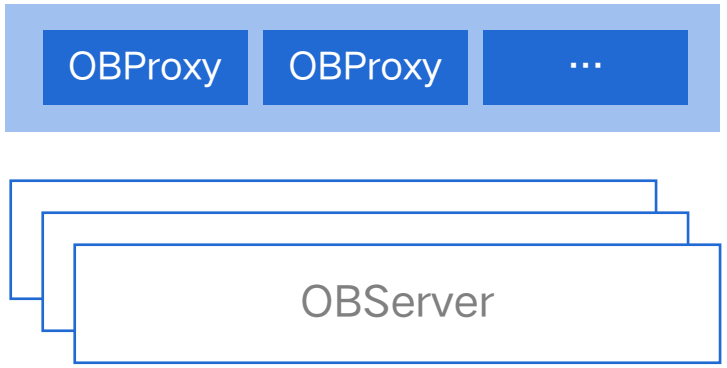
- 文件系统层数据同步
  - 分布式存储技术
  - RDMA
  - **部署难度大，仅可购买云服务**
- Share Everything
  - Aurora
  - PolarDB
  - TDSQL-C
  - GaussDB
- 云厂商的解决方案

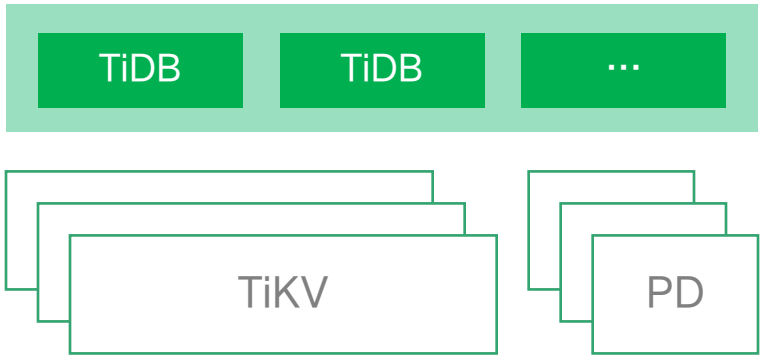
- 数据库日志层数据同步
  - 通用技术
  - 可自建部署
  - **需实现接入层高可用**
- Share Nothing
  - TiDB
  - OceanBase
  - MySQL Group Replication
- 数据库厂商的解决方案

接入层高可用方案（1/2）

- 无状态计算层 / Proxy层
  - Load balance
  - 智能DNS（自动摘除 + 较短TTL）
  - 技术实现相对容易



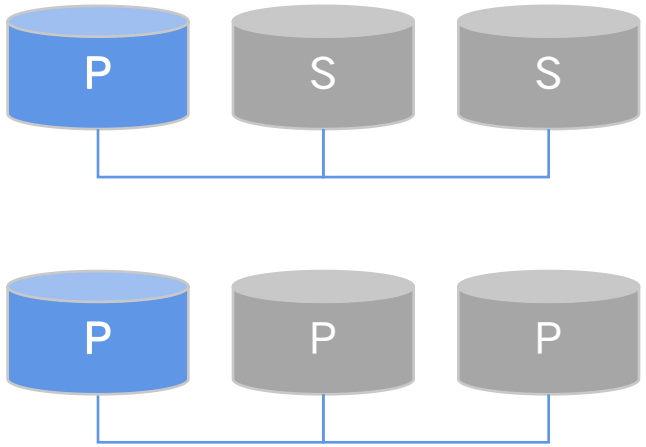


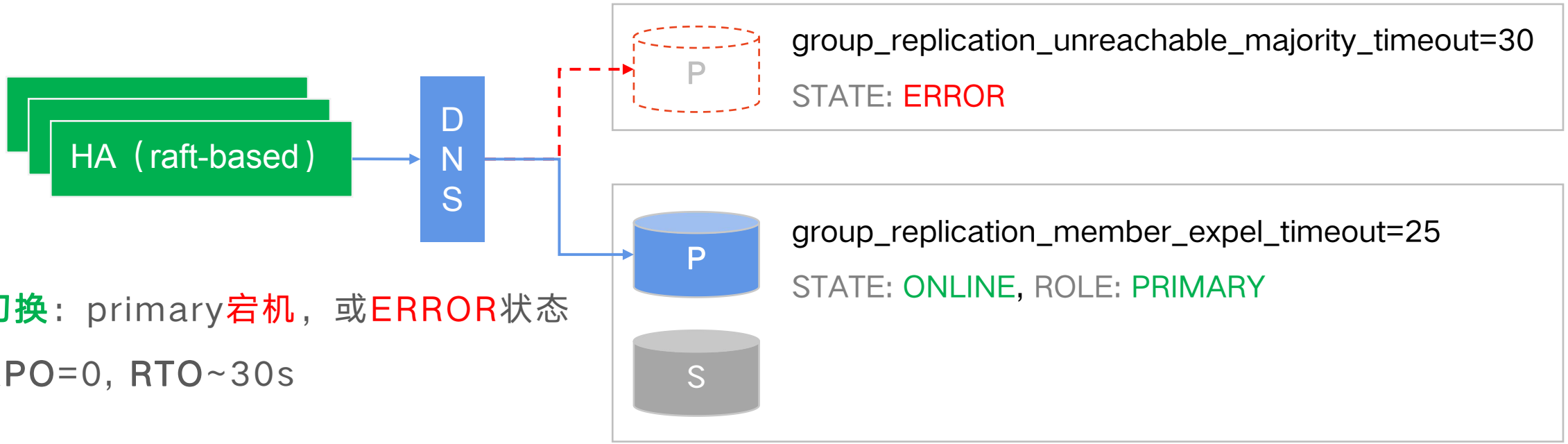


接入层高可用方案（2/2）

中国数据库联盟 All China Database Union　墨天轮

- 直连MGR节点（更高性能）
  - 方案1：single-primary
  - 方案2：multi-primary + 单写
  - 不建议：multi-primay + 多写





切换：primary宕机，或ERROR状态

RPO=0, RTO~30s

总结

- 基于半同步复制，RPO可接近于0，但不是100%
- 外置HA组件，不能避免脑裂问题
- RPO=0的方案，依赖基于Paxos / Raft的数据同步机制
- 趋势：云数据库（云厂商）、分布式数据库（数据库厂商）

ACDU·中国行

中国数据库联盟 All China Database Union 墨天轮

谢谢观看

THANKS FOR WATCHING

中国数据库联盟 All China Database Union 墨天轮