数据库高可用架构的最新进展

网易 郭忆

目录

01 数据库高可用发展历程

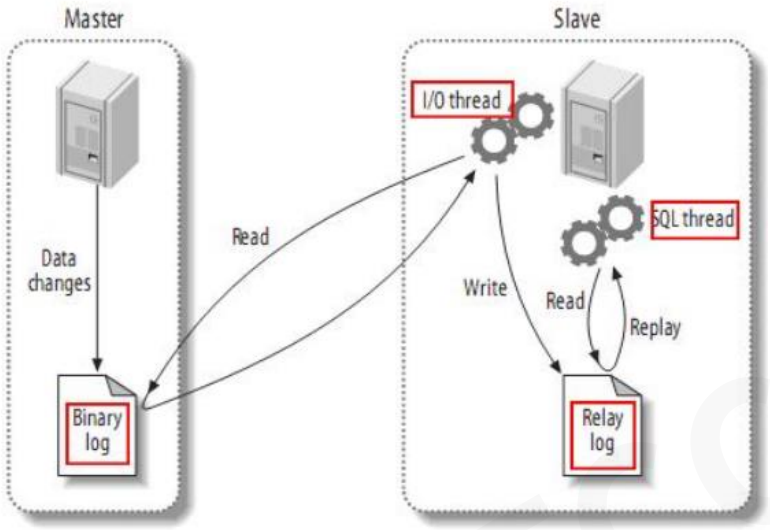
02 Aurora 高可用架构设计

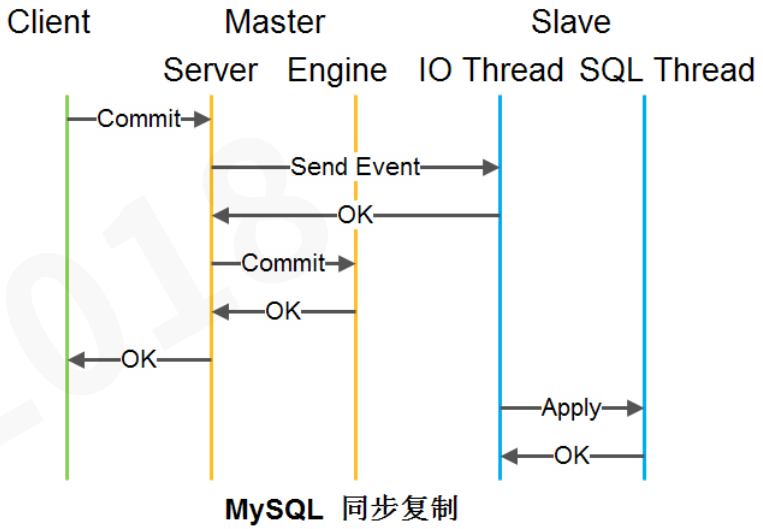
03 MGR 高可用架构设计

04 网易多副本数据一致高可用架构设计

PART 01 数据库高可用发展历程

基于复制的数据库高可用架构





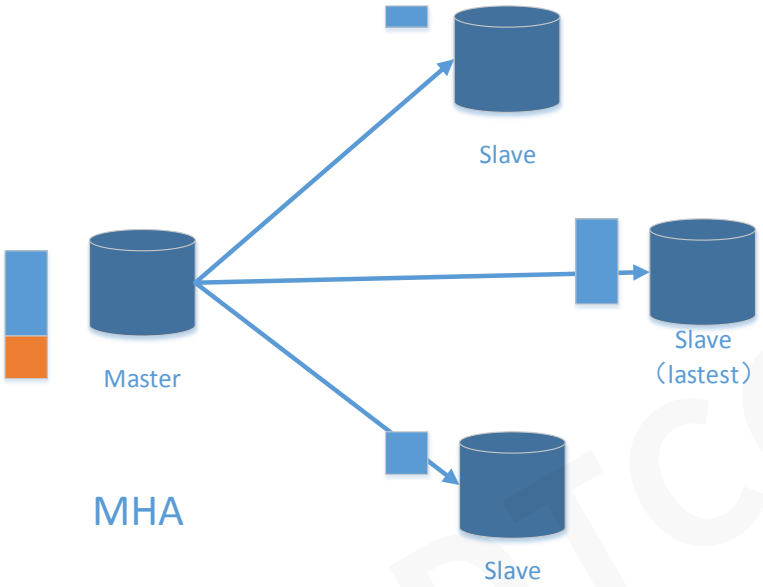
Pros

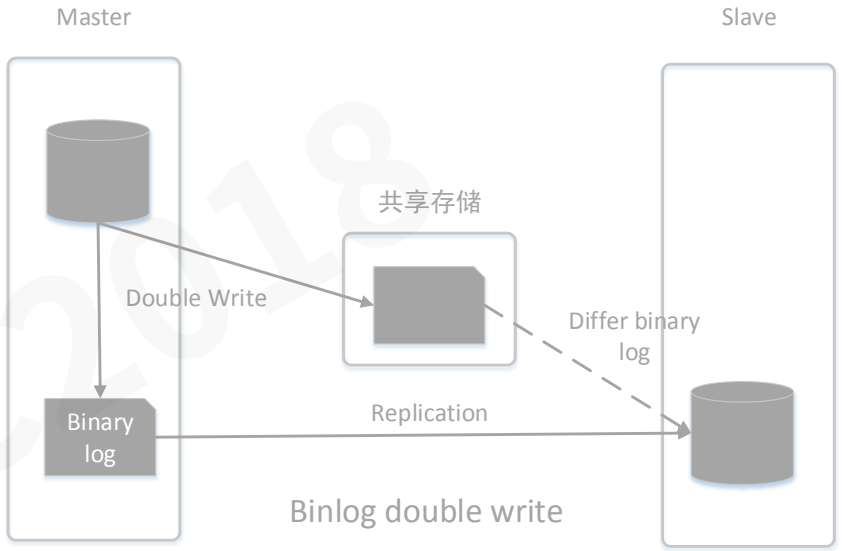
- 原生支持
- 快速部署，易维护
- 同步复制可以确保数据一致

Cons

- 异步复制丢数据
- 同步复制无法支撑写密集业务
- 复制延迟

基于日志的数据库高可用架构





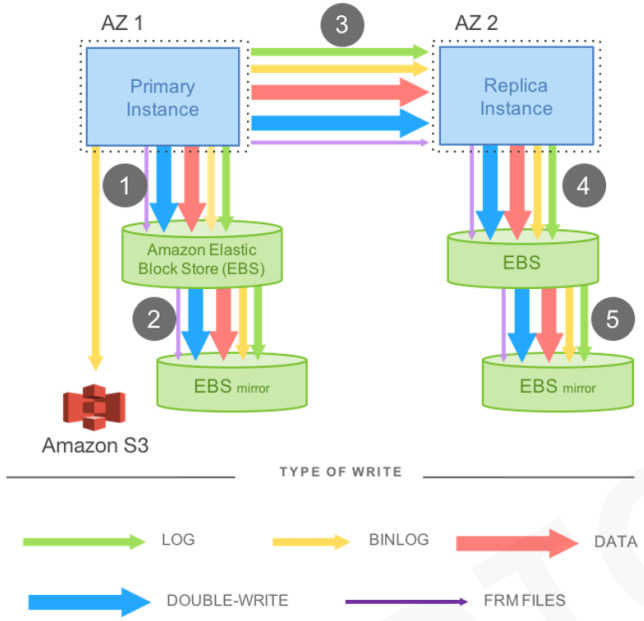
Pros

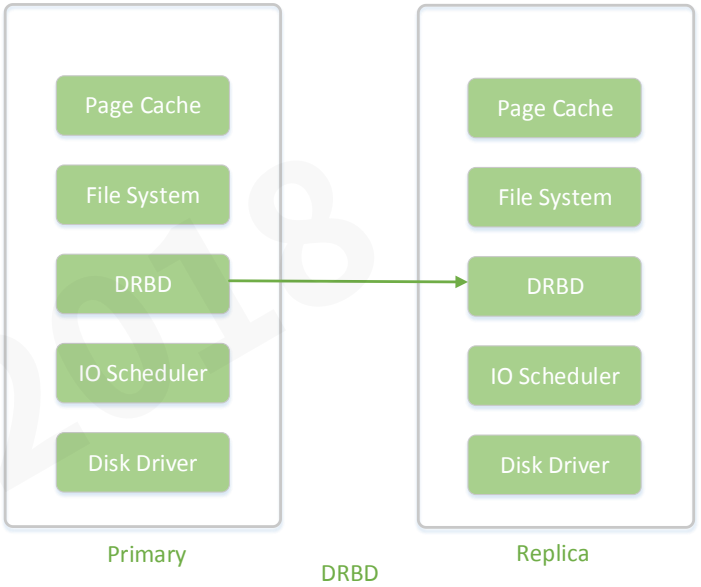
- 限定条件下能够保证数据的一致性

Cons

- MHA： 主机服务器SSH无法访问丢数据
- 日志双写：异步状态下丢数据

基于块设备镜像的数据库高可用





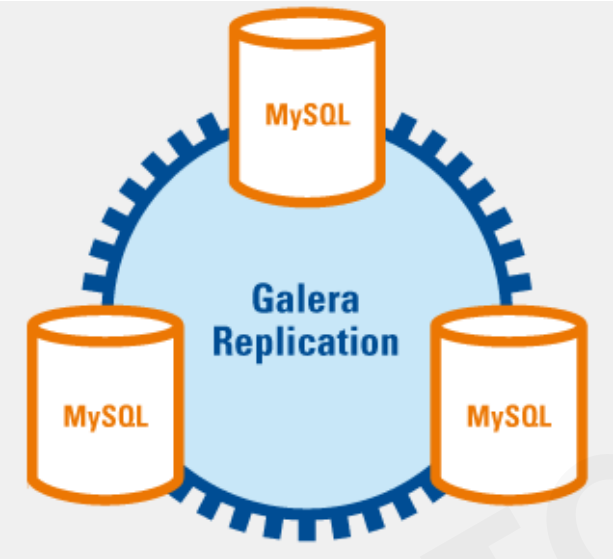
Pros

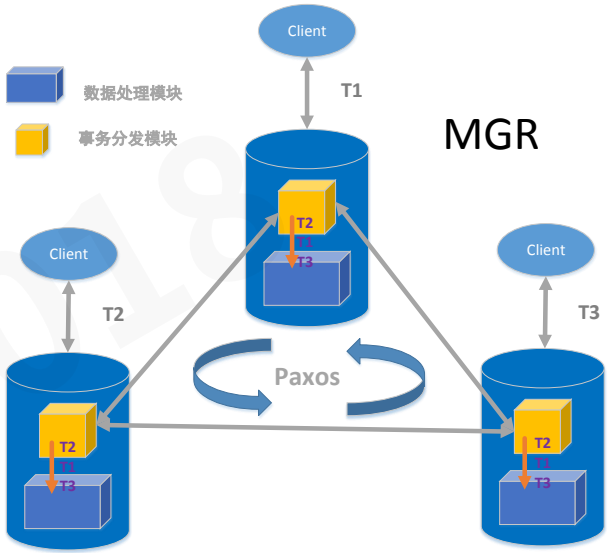
- 对数据库透明
- 通用高可用解决方案

Cons

- 性能无法满足
- 跨网络流量

Shared nothing 多副本高可用架构





Pros

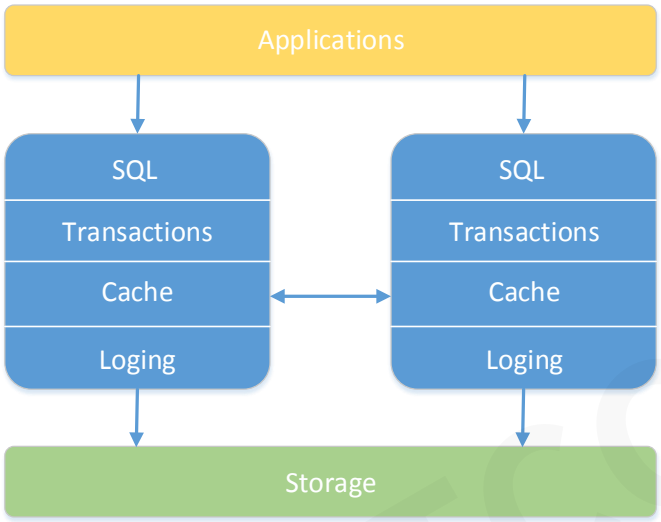
- Shared nothing
- 多副本，金融级可靠性
- 支持多写，解决写扩展问题

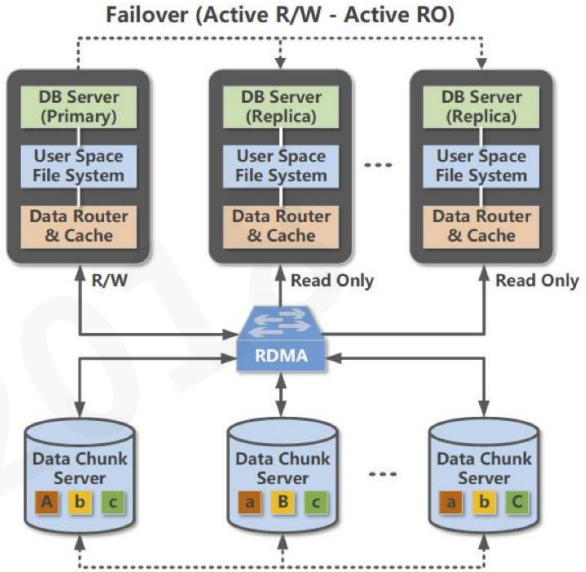
Cons

- 基于binlog的数据同步，复制延迟难以避免
- 多写模式下所有的事务提交都必须要冲突检测，即使没有冲突

Cloud-Native Database

Aurora





PolarDB

Pros

- Shared disk cluster
- 计算和存储分离，解决扩展性
- 方案具有通用性
- 获得更好的性能

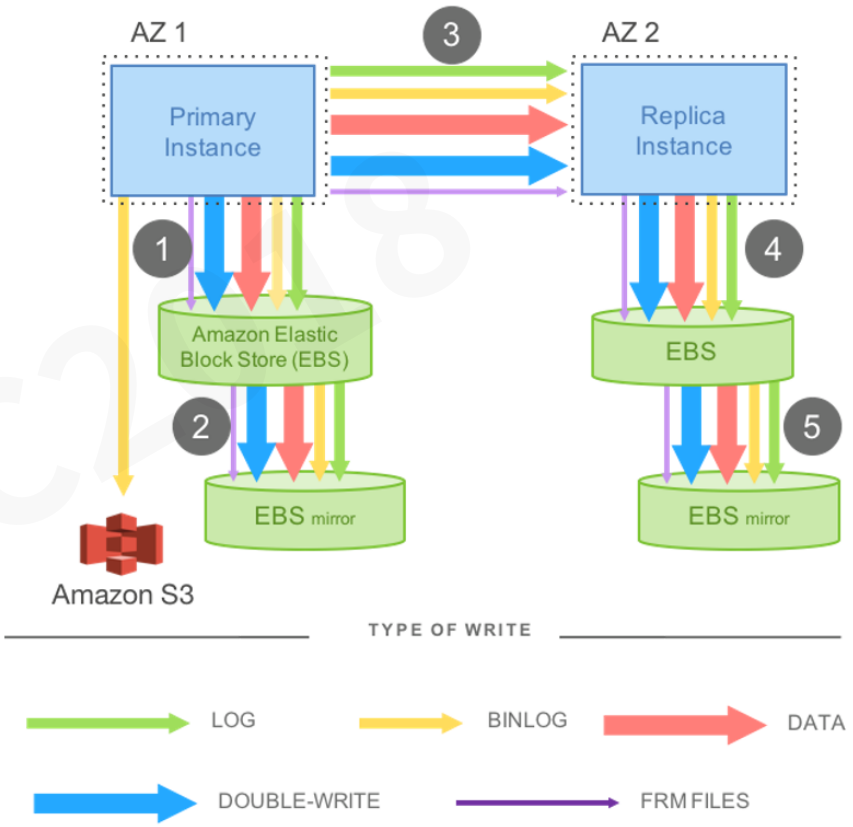
Cons

- 技术门槛高
- 强依赖于底层基础设施

PART 02 Aurora 高可用设计

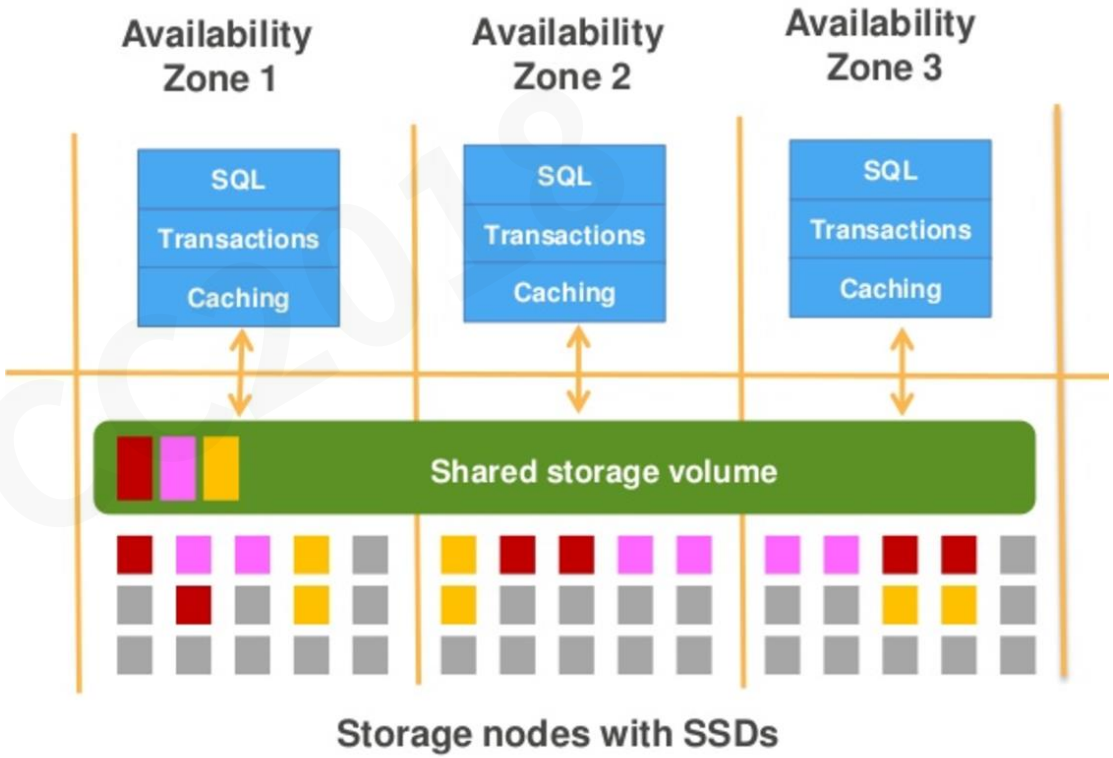
为什么要有Aurora?

- 跨网络传输的数据
  - Redo log
  - Binlog
  - Data Page
  - Double Write
  - FRM
- 所有数据要跨网络传输5次，
  其中3次传输还是串行的



Aurora

- 统一的跨可用区、多副本、高可用的共享存储系统
- 数据库实例和存储系统仅通过redo同步数据
- Cache 之间通过redo 同步减少时延
- Recovery 异步执行，秒级恢复

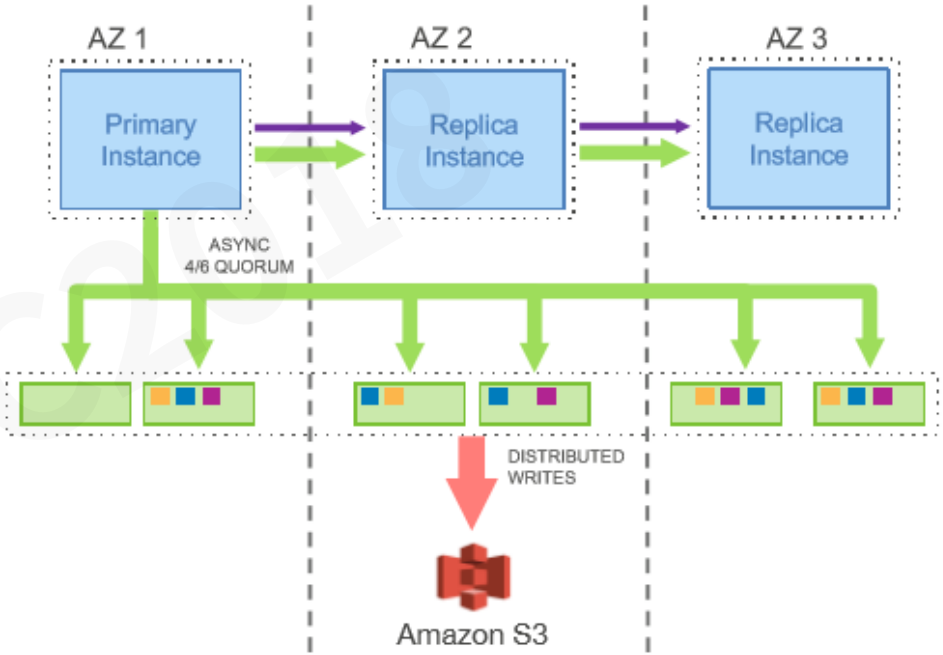


Storage Design

- 数据库实例的存储由10G 大小的Segment组成
- 每个Segment 有6个副本，分布在3个可用域
- 存储节点是由挂载了本地SSD的EC2组成
- 单个数据库最大支持64 TB
- Segment是存储系统数据修复的最小单元
- 通过标记Segment不可用完成计划内的迁移操作（热点不均衡）

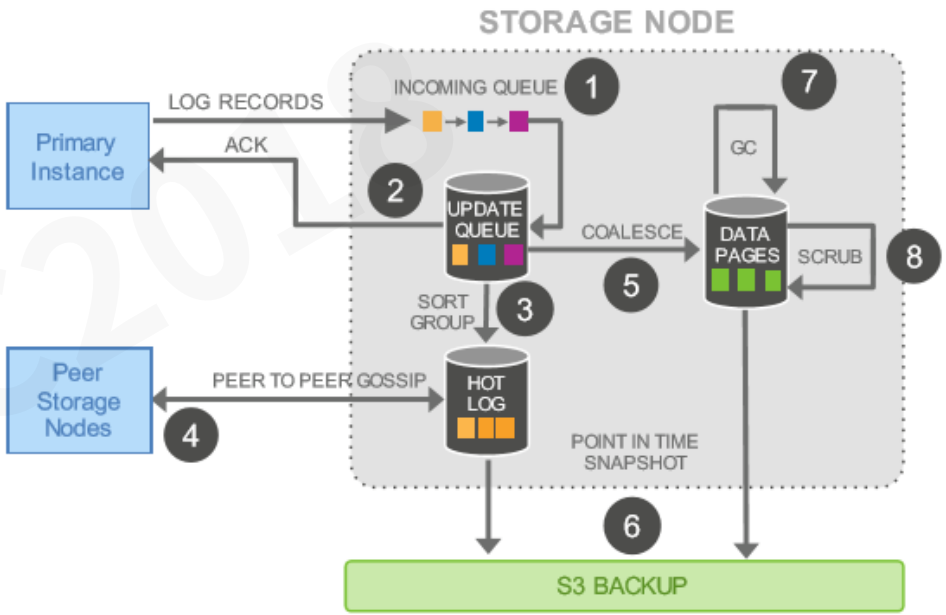
Quorum Design

- Quorum
  - a write quorum of 4/6 (Vw = 4)
  - a read quorum of 3/6 (Vr = 3)
- 容错性
  - 失去整个可用域和另外的一个存储节点，不影响整体系统的读可用性
  - 失去任意两个节点，包括同一个可用域或者不同可用域，不影响写可用



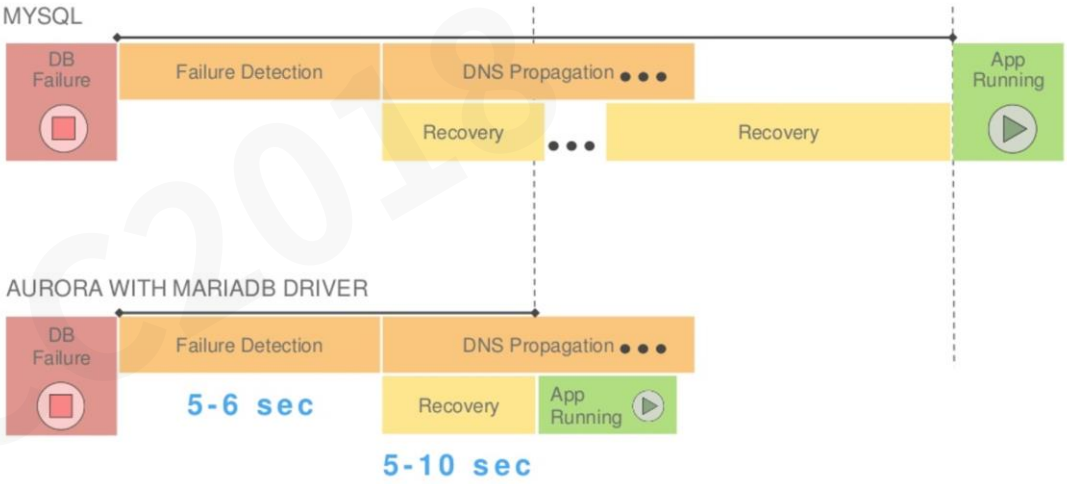
事务提交

- redo log在mtr提交时copy log buffer
- Log record 根据page所在的存储节点分成多个batch，然后发送到对应的存储节点
- 如果一个PG内的6个Segment中4个写入成功，则数据库返回客户端事务提交成功，同时推动VDL
- 每个Page 根据待更新的redo log record长度决定page materialization
- 同一个PG内的不同Segment 通过Gossip协议补全日志



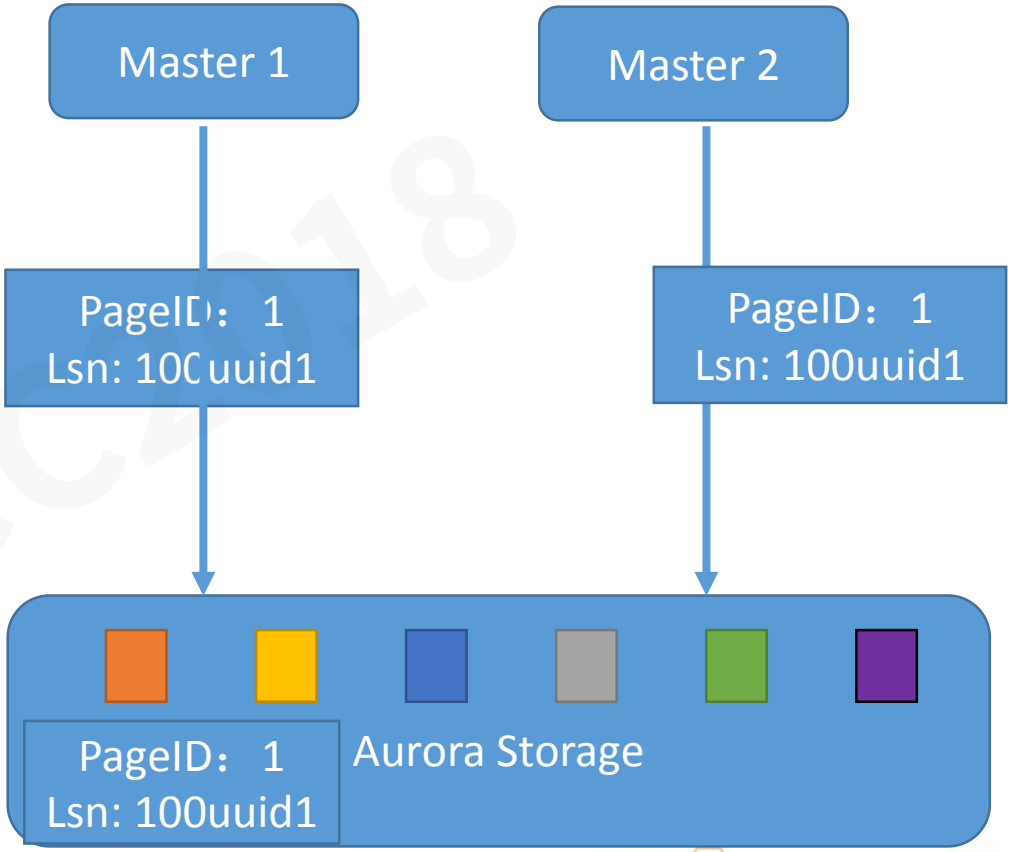
故障恢复

- CPL ： 每个mtr 最后一个log record对应的lsn
- VCL：持久化的最大的log record lsn
- VDL：小于VCL 的最大的CPL
- Recovry的过程只需要建立VDL即可，与MySQL 回放redo 不同
- 未提交事务异步回滚



Muti-Master

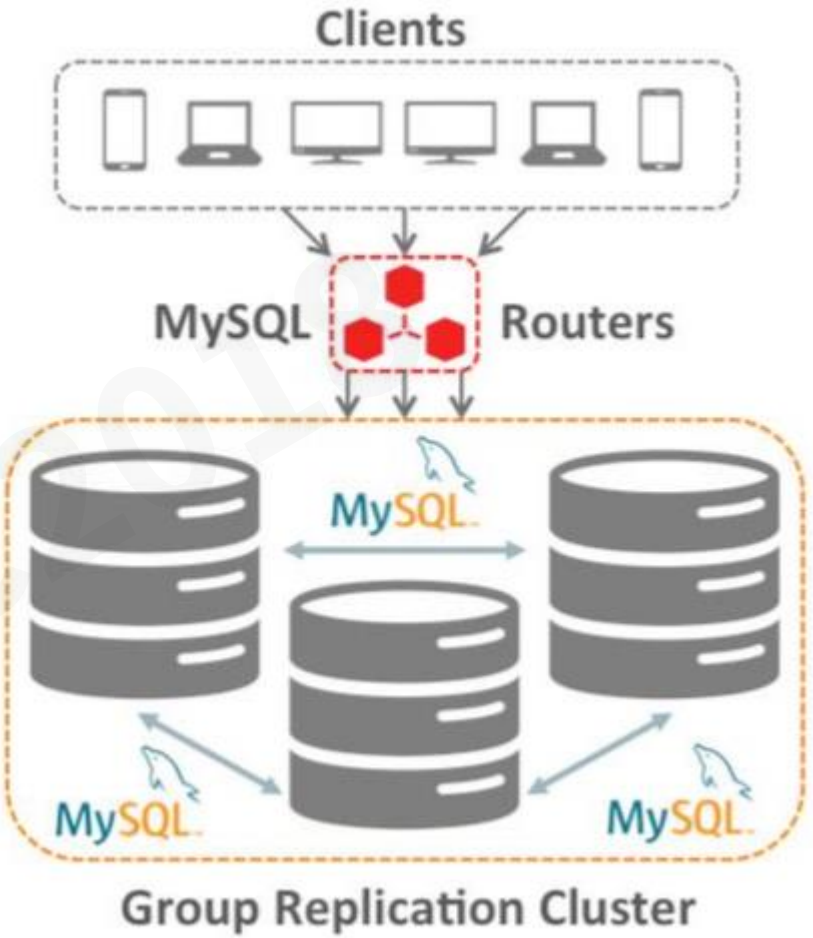
- 冲突检测
  - 以Page为粒度进行冲突检测
  - 基于lsn 实现版本管理和冲突检测
  - 利用逻辑时钟（Lamport Clock）解决因果关系的事务顺序执行问题
  - 基于Quorum原则，最先写成功4个的事务提交，冲突事务回滚



PART 03 MySQL Group Replication

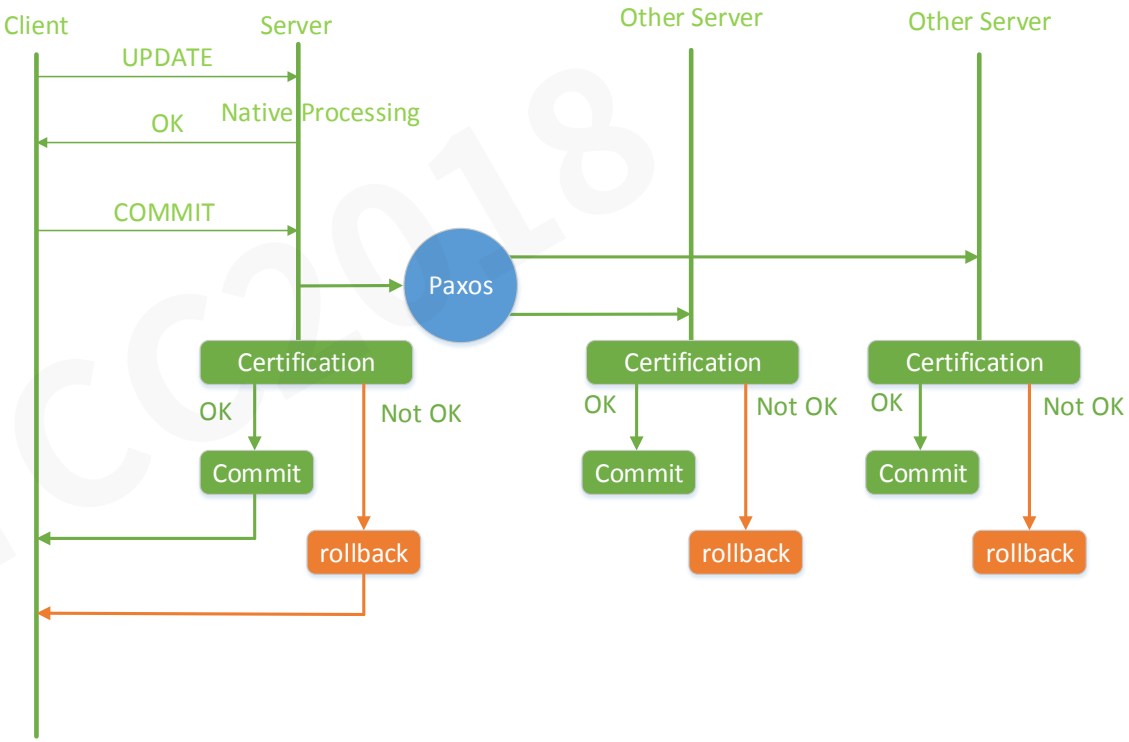
MGR

- MySQL 5.7.17 发布
- 基于Paxos 协议实现的多副本数据一致性集群
- 支持single-master 模式和multi-master模式
- 作为MySQL Innodb cluster 解决方案一部分

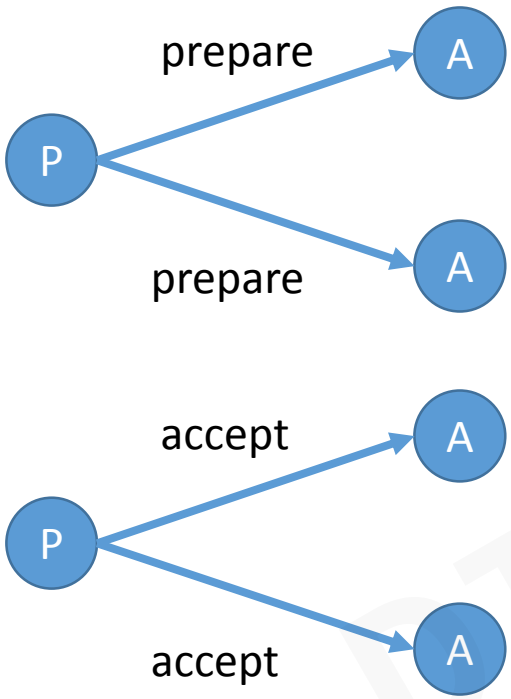


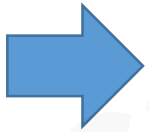
MGR 基本原理

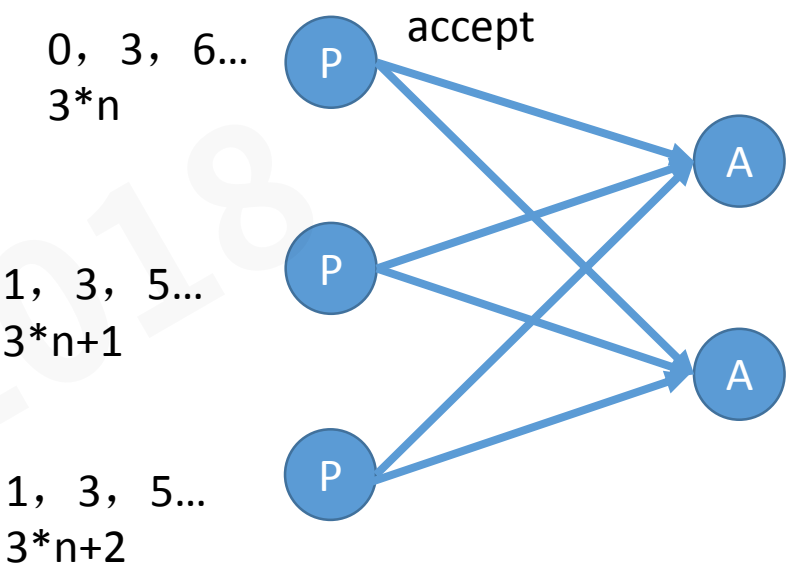
- 所有节点都有相同的数据副本
- 基于binlog实现数据的同步
- 本地事务提交时，进入全局排序
- 基于Paxos协议确保集群内所有节点按照相同的事务次序执行
- 所有事务基于write set 进行冲突检测



一致性协议







Basic Paxos

- 优势：任意节点都可以发起提议
- 劣势：
  - 每个value都需要至少2次网络开销，2次磁盘持久化
  - 容易产生活锁

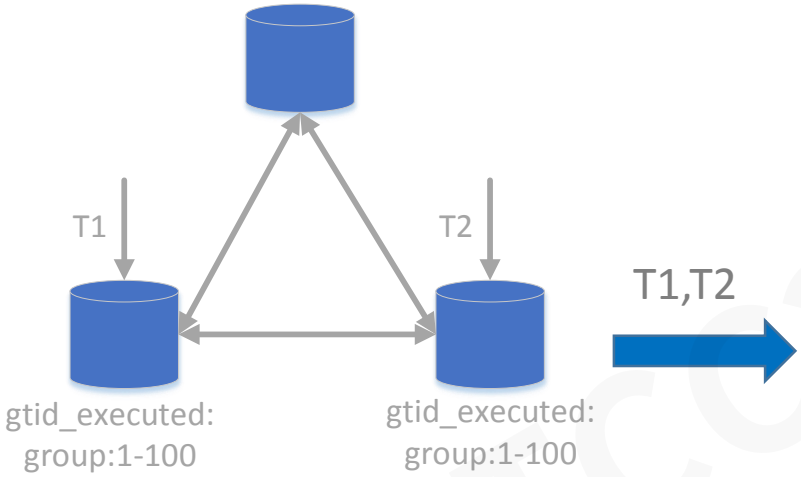
Mencius

- 优势：
  - 相比Basic Paxos 节约了Prepare阶段的性能开销
  - 相比Muti-Paxos 消除Leader 瓶颈，每个成员负责一部分的提议

Write set

- Write set：
  - 每个事务新增加一个Log Event（Transaction_context_log_event）
  - 包含信息
    - 事务更新的主键
    - 数据库快照版本(gtid_executed)
  - 只在内存中维护，不写入binlog 文件，保证兼容性
- 冲突检测：
  - 每个成员节点按照相同的次序（Paxox协议保障），分别进行冲突检测
  - 每个成员节点都维护了一个“冲突检测数据库”，所有待检测的事务对应的数据库版本必须大于冲突检测数据库中已经通过检测的记录的版本
  - 所有节点都已经执行的事务对应的记录会从冲突检测数据库中异步purge

冲突检测



T1: Update s1 set c2 = 5 where pk =1

T2: Update s1 set c2 = 6 where pk =1

T1' gtid_executed : group:1-100

主键	版本
1	group:1-100

T1通过冲突检测

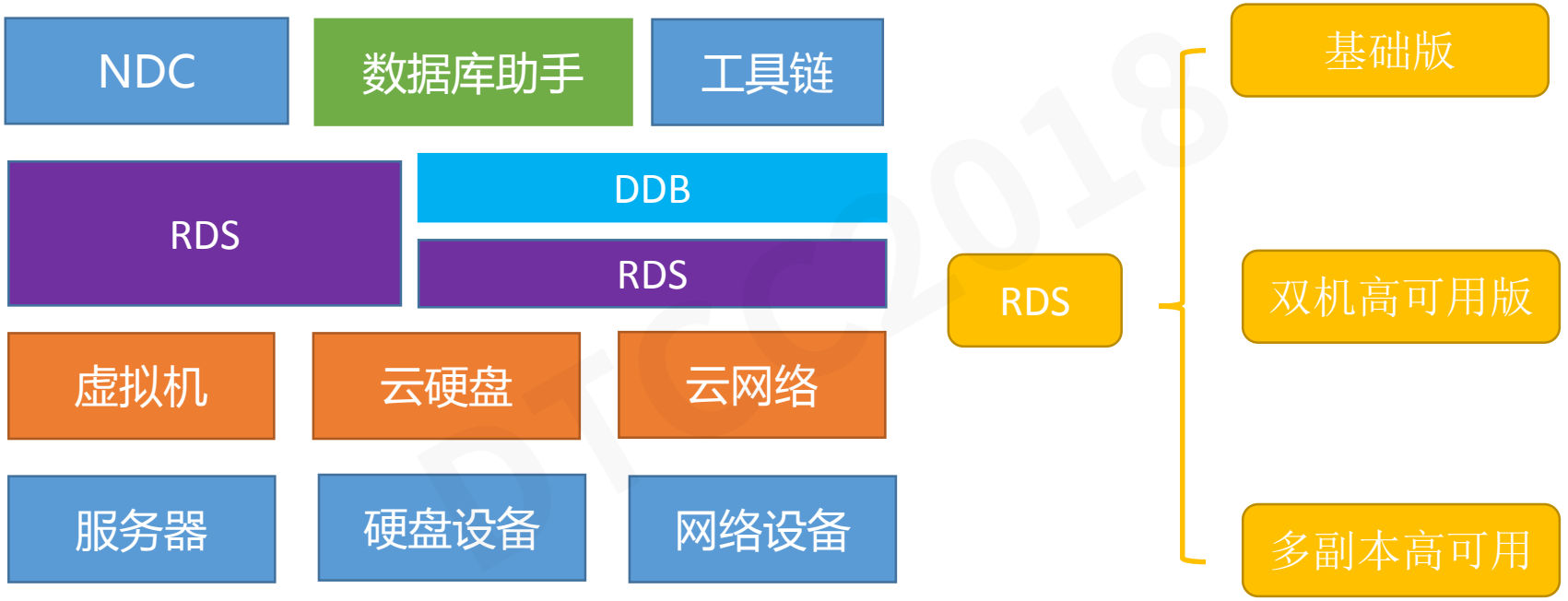
T2' gtid_executed : group:1-100

主键	版本
1	group:1-101

T2未通过冲突检测，回滚

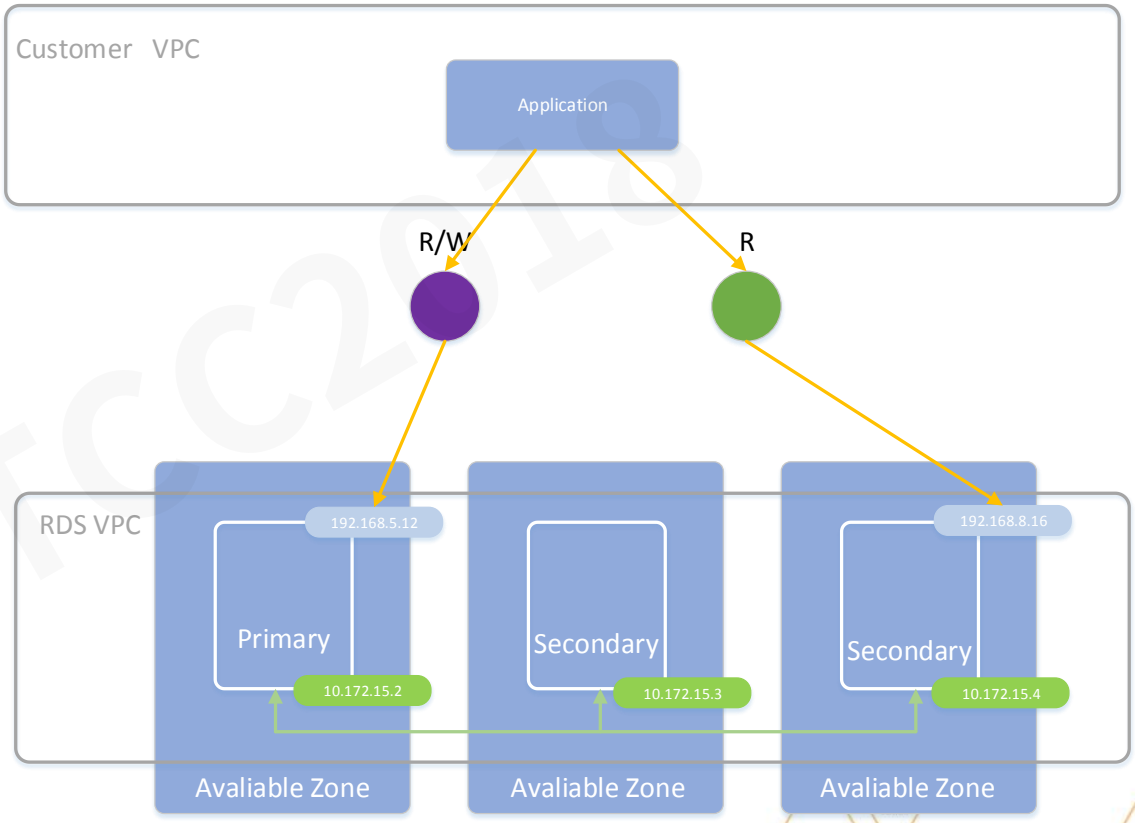
PART 04 网易数据库高可用解决方案

网易云数据库

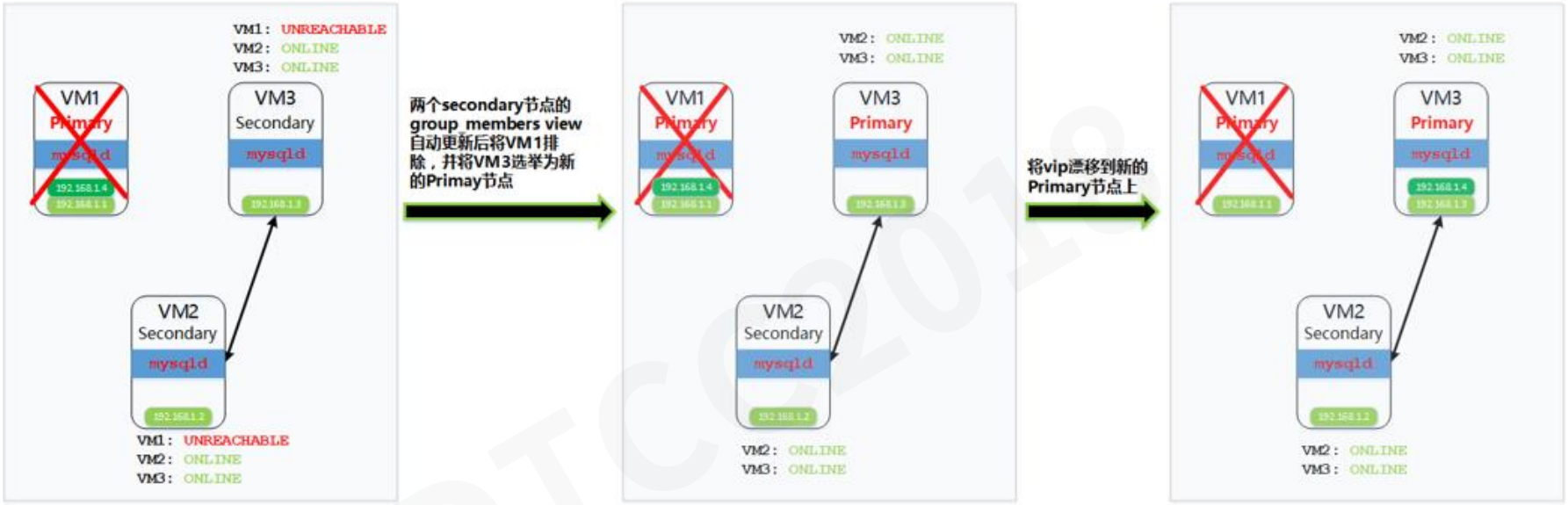


网易基于MGR 多副本数据库高可用架构

- 三个节点位于三个可用域内，物理隔离
- 基于RDS VPC 网络实现集群内的数据同步
- 提供读写和只读两个域名
- 三节点中设置一个只读节点
- 通过权重影响节点切换策略



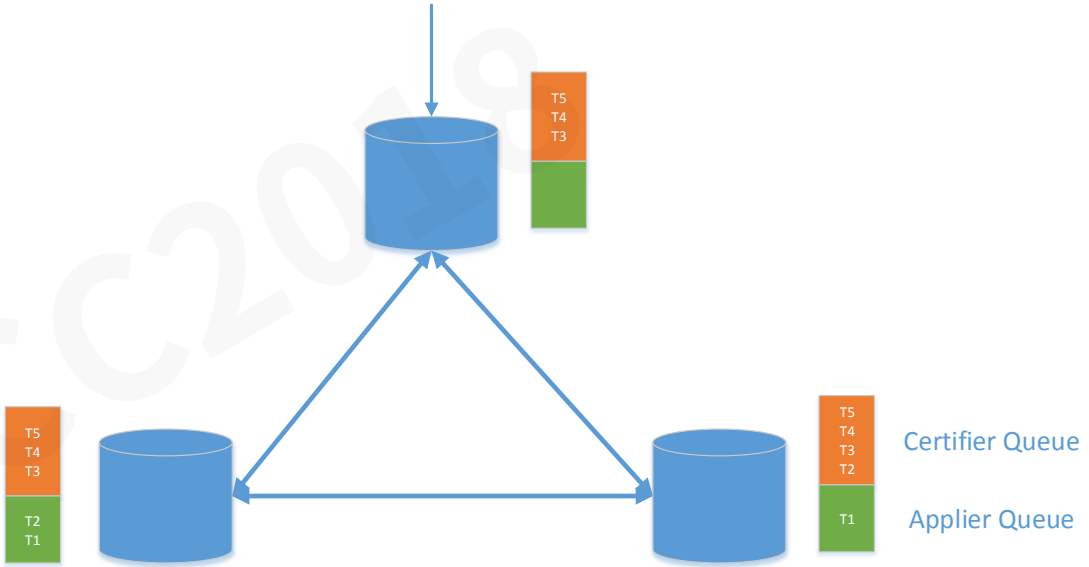
故障恢复



- Failover time = election time + apply time
- 故障节点修复后重新加入集群，选择只读节点作为Seed
- 控制只读能力，尽快完成数据的修复

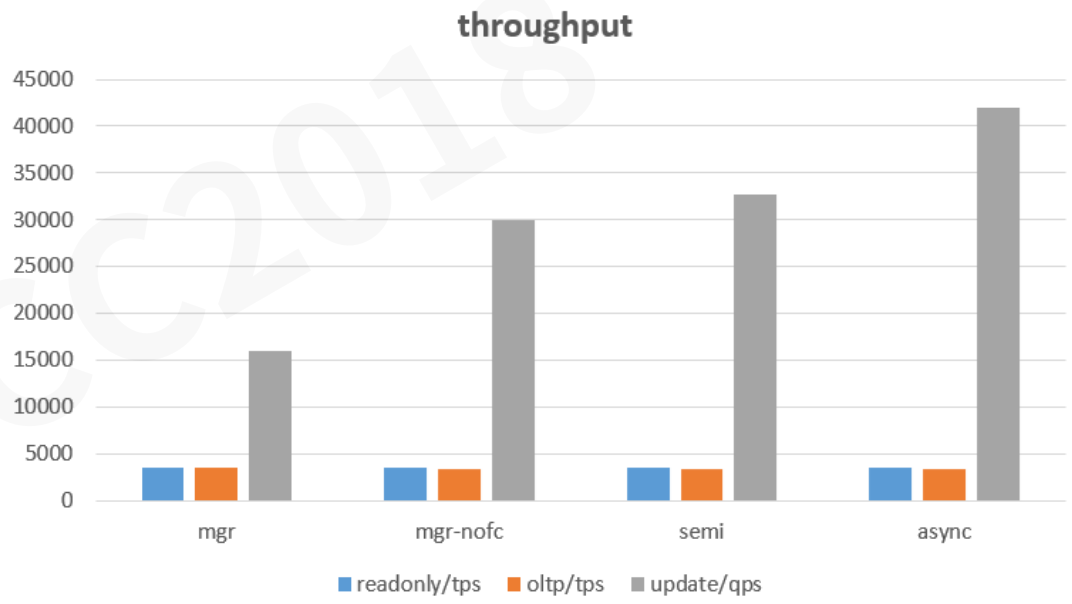
流控设计

- 参数
  - group_replication_flow_control_mode
  - group_replication_flow_control_certifier_threshold
  - group_replication_flow_control_applier_threshold



性能表现

ssd/4g/256t	readonly/tps	oltp/tps	update/qps
mgr	3585.60	3470.86	15903.40
mgr-nofc	3544.15	3429.83	29992.78
semi	3480.51	3330.07	32649.29
async	3505.40	3411.60	42039.33



THANKS


HANA
Hadoop
SQL
MySQL
DB2
BigData
Spark
SQL Server
NewSQL
Oracle
MPP
NoSQL
OLAP

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168