ODF 2017 开源数据库论坛(北京)
OPEN-SOURCE DATABASE FORUM(BEIJING)

开源数据库正在改变世界

2017年8月24日-25日　北京－京仪大酒店

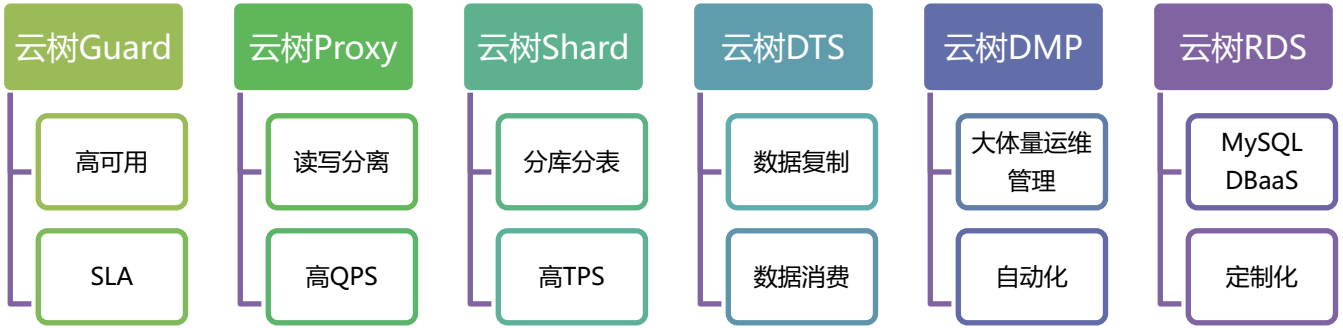
金融级MySQL高可用方案选型

上海爱可生信息技术股份有限公司

资深DBA 张沈波

我们干了啥

爱可生产品全家福——云树系列



www.actionsky.com

客户案例

客户概览

- 金融机构
- 银行业
- 保险业
- 互联网金融业
- 通信行业
- 广电行业
- 能源电力行业
- 航空航天行业
- 政府教育行业
- 高科技行业
- 互联网行业
- 零售行业

金融机构

中国人民银行 THE PEOPLE'S BANK OF CHINA　中国外汇交易中心　中国银联 China UnionPay　全国中小企业股份转让系统 NATIONAL EQUITIES EXCHANGE AND QUOTATIONS　CICC 中国国际金融有限公司 CHINA INTERNATIONAL CAPITAL CORPORATION LIMITED

银行业

中国银行 BANK OF CHINA　招商银行 CHINA MERCHANTS BANK　兴业银行 INDUSTRIAL BANK CO.,LTD.　中国民生银行 CHINA MINSHENG BANK　KEB

保险业

PICC 中国人保　太平洋保险 CPIC　阳光保险 Sunshine Insurance　百年人寿 AEON LIFE　前海人寿 FORESEA LIFE INSURANCE

互联网金融业

借贷宝　京东金融 JD Finance　宜人贷 www.yirendai.com　欧治云商 Ouyeel　海尔消费金融 haierxiaofei.com

个人介绍

- 拥有丰富的一二线MySQL运维经验；
- 先后在阿里云，爱可生担任数据库运维；
- 目前为爱可生数据库产品负责人，MySQL技术专家；

目录

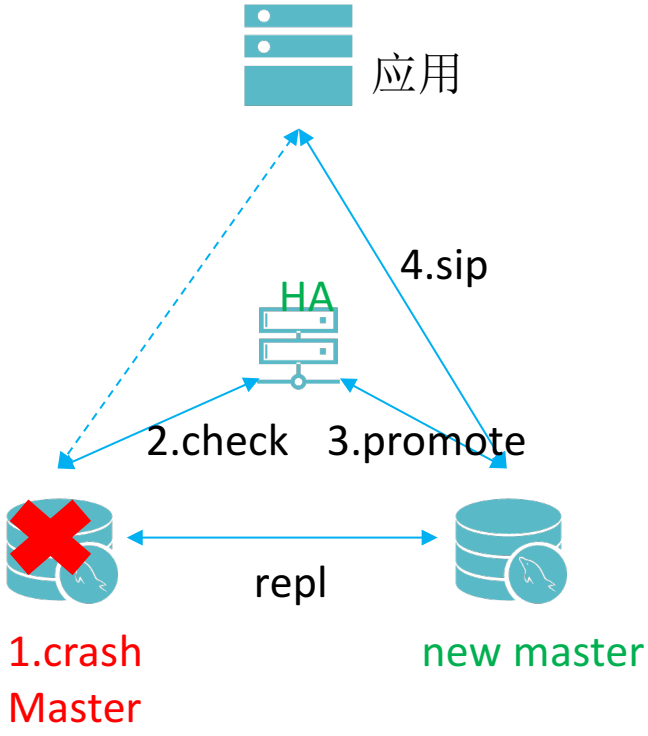
CONTENTS

一. 高可用的考量

二.高可用覆盖的故障

三.高可用方案选型

高可用考量-背景



高可用考量

1. 数据一致性
   - 通过不同的高可用保障方式（异步 / 同步复制）实现主库发生切换时，数据零丢失
2. 业务连续性
   - 业务连续性，是指数据库的消费者（业务），是否可以一直访问和使用数据库．在发生主备切换时，数据库的业务连续性会受到多长时间的影响
3. 数据库性能
   - 由于主备上至少存有两份数据，与只有一份数据相比，主数据库承担业务压力的能力可能会受到影响，需要做好性能平衡

高可用考量

探讨

- 一致性vs连续性
- 一致性VS性能

目录

CONTENTS

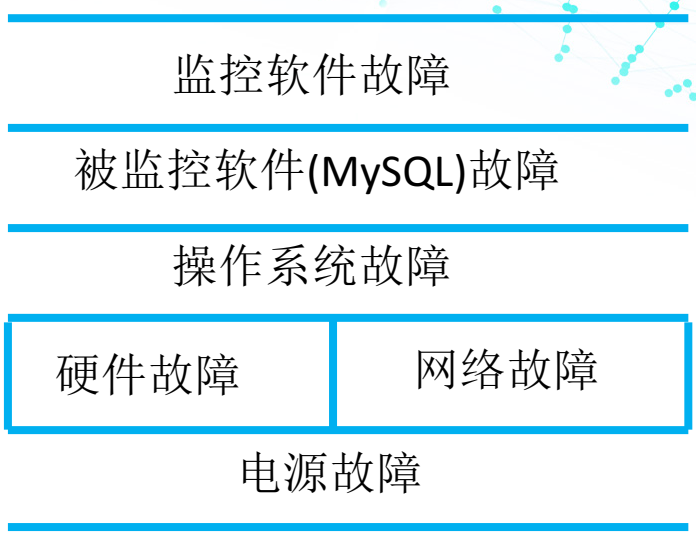
一. 高可用的考量

二.高可用覆盖的故障

三.高可用方案选型

高可用覆盖的故障

- 硬件故障
- 网络故障
- 系统故障
- 被监控软件（MySQL）故障
- 监控软件故障
- 脑裂



高可用覆盖的故障

硬件故障

硬件故障可能是可恢复的, 也可能不可恢复. 由于无法全面且准确地预测硬件故障的发生时间/发生种类/产生的影响

一般对硬件故障的检测方法是:

- 对可预测的硬件故障进行故障检测/预防性检测 (bbu,raid,bond,...)
- 检测由高层应用抛出的错误 (disk read only, mysql abort_server ,...)

高可用覆盖的故障

网络故障

网络不可用

- 网络不可用指较长时间网络通路不可用，可通过节点间心跳来检测.

网络闪断

- 网络闪断指较短时间内网络在可用和不可用状态间震荡. 可将心跳检测的超时时间设为能容忍的网络闪断的最长时间 t，即容忍最长 t 时间的网络中断，超时则认为网络中断.

网络稳定性和延迟

- 网络稳定性和延迟可以由以下特征量进行描述，节点间的网络通讯协议需能正常工作于由以下特征量描述的某网络上
  - 网络包丢失率
  - 网络包损坏率
  - 网络包乱序率
  - 网络包重复率
  - 网络包延时时间

高可用覆盖的故障

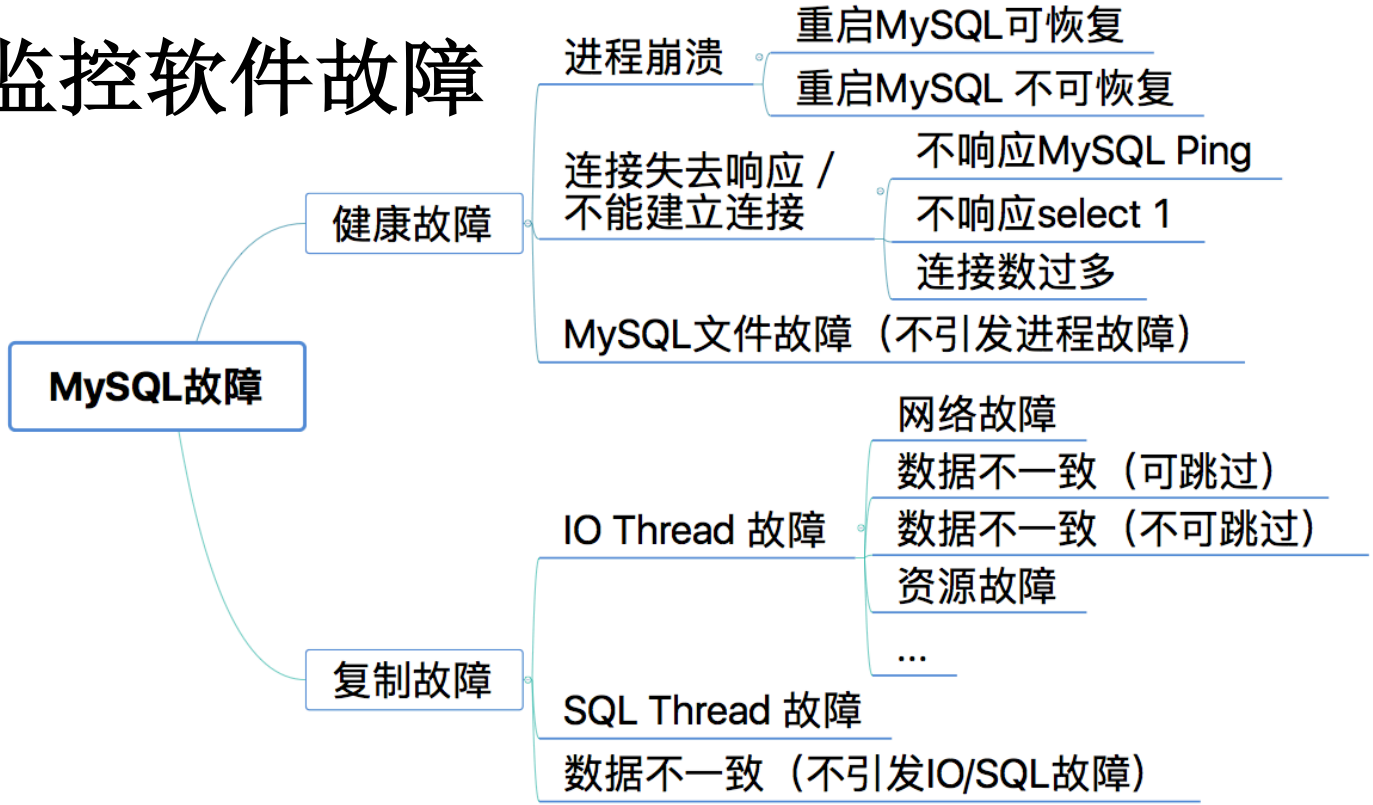
操作系统故障

类似于硬件故障, 无法全面且准确预测其发生时间/发生种类/产生的影响. 类似于硬件故障的检测方法

- 对可预测的操作系统故障进行故障检测/预防性检测
- 对资源的使用, 如磁盘/内存等, 进行监控和阈值告警
- 监控软件在申请资源开销前, 进行预估
- 检测由高层应用抛出的错误

高可用覆盖的故障

被监控软件故障



高可用覆盖的故障
监控软件故障

监控软件的故障常见的有:

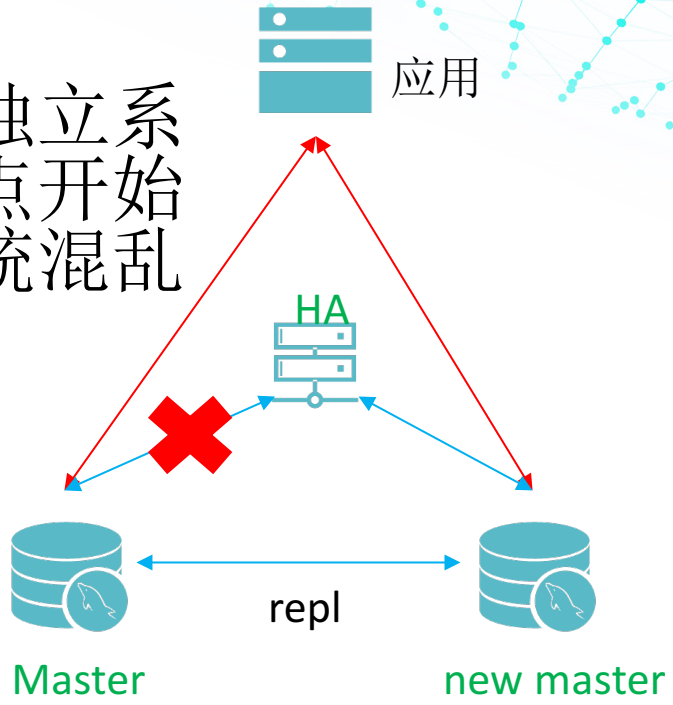
- 意外退出
- 失去响应
- 由于资源被其它应用抢占, 发生崩溃
- 由于资源被其它应用抢占, 响应速度变慢, 引发时序错误
- 长期运行时, 资源被缓慢占用得不到释放
- 一些时序服务, 如日志回收/分布式锁/等, 在监控软件发生故障后无法回收现场,产生后效性

高可用覆盖的故障

脑裂

一个整体的系统，分裂为两个独立系统，这时两个系统各自的主节点开始争抢共享资源，结果会导致系统混乱数据损坏

- 添加冗余的心跳线
- **fence**
- **一致性选举(**paxos/raft**)**
- **…**



高可用覆盖的故障

探讨

集群可用性

- HA自身可用性
- 脑裂难题

目录

CONTENTS

一. 高可用的考量

二.高可用覆盖的故障

三.高可用方案选型

高可用方案选型

一致性 & 性能

一致性 & 连续性

集群可用性

高可用方案选型

一致性 & 性能

- 数据零丢失
- 最大性能

选型

- 半同步
- 异步

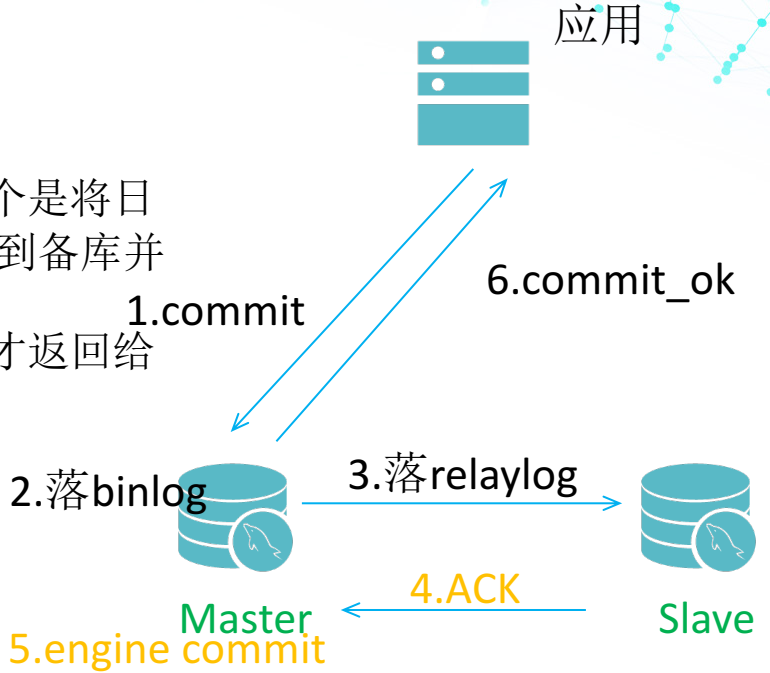
高可用方案选型

前提

单机MySQL 通过Binlog来做一致性协调

实现

1.事务提交的时候，发起两个写日志操作，一个是将日志写到本地磁盘的操作，另一个是将日志同步到备库并且确保落盘的操作；
2.主库此时等待两个操作全部成功返回之后，才返回给应用方，事务提交成功；



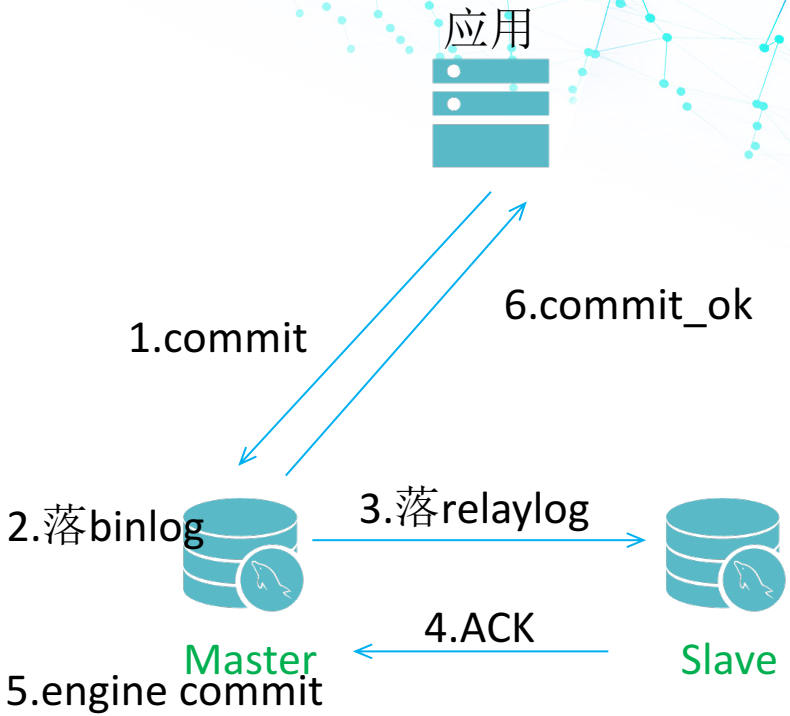
限制

MySQL 5.7
半同步降级？(丢数据，master hang)

高可用方案选型

更多实现

- 定义配置标准
- 调度半同步起停
- 检测半同步状态
- 调度slave count数
- 量化回放延时和日志延时



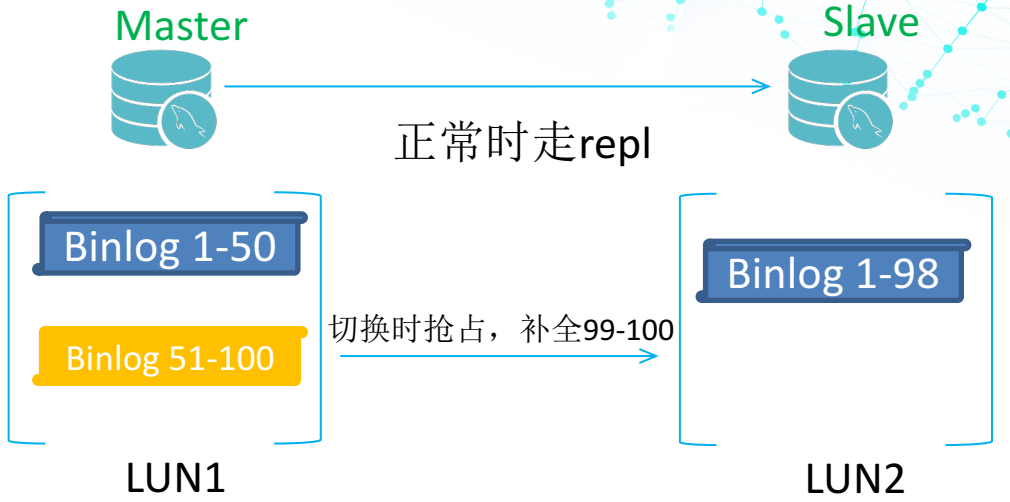
高可用方案选型

前提

- Binlog做一致性协调
- 引入了一个外部技术会降低可用性
- 保证外部技术自身可用性够高

实现

通过binlog落共享存储盘，切换时争抢主机端binlog盘来补偿数据，保障数据不丢失



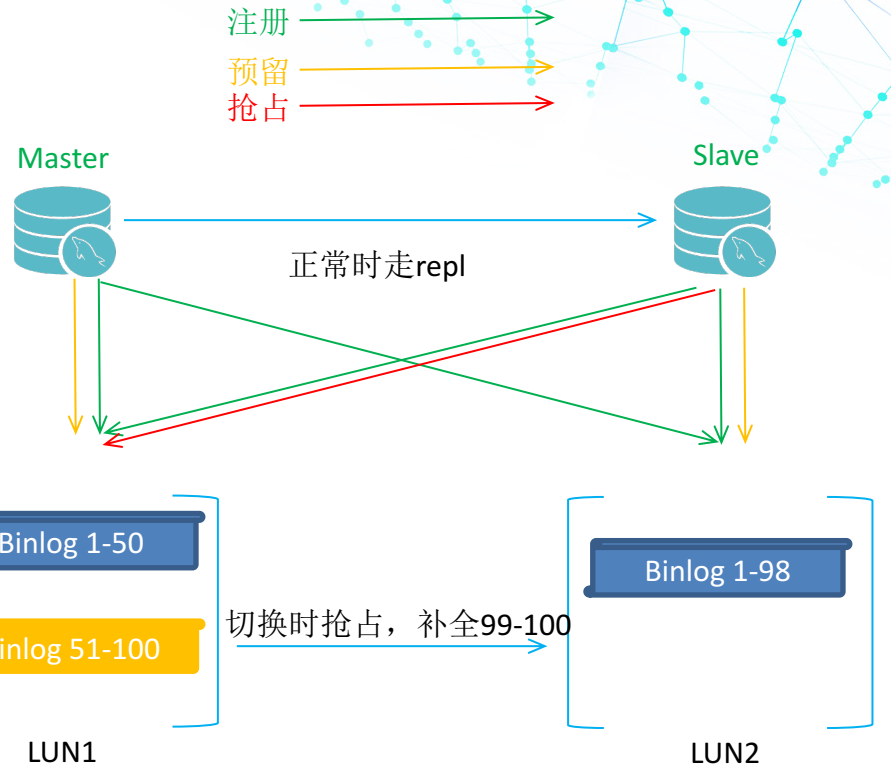
限制

成本较高

高可用方案选型

更多实现

通过国际标准的Scsi PR协议
完成磁盘注册，预留，抢占



高可用方案选型

一致性 & 连续性

- 一致性优先：尽一切手段补全
- 连续性优先：规定时间内必须切换
- SLA的定义：听我的

SLA的定义

- SLA服务等级协议（简称：SLA，全称：service level agreement），是业务根据需求与SLA组件签订的提供数据库数据一致性或数据库服务连续性等级协议；
- 该SLA协议可以从数据一致性，数据连续性两个维度来签订协议，满足灵活的业务场景

SLA的定义

数据一致性优先

级别	服务等级描述	日志差异	主从延时
P1	数据零丢失，秒级切换	0	<60s
P2	数据零丢失，分钟级切换	0	<10分钟
P3	数据零丢失，大于10分钟切换	0	>10分钟
PE1	数据丢失，不切	>0	<60s
PE2	数据丢失，不切	>0	<10分钟
PE3	数据丢失，不切	>0	>10分钟

SLA的定义

业务连续性优先

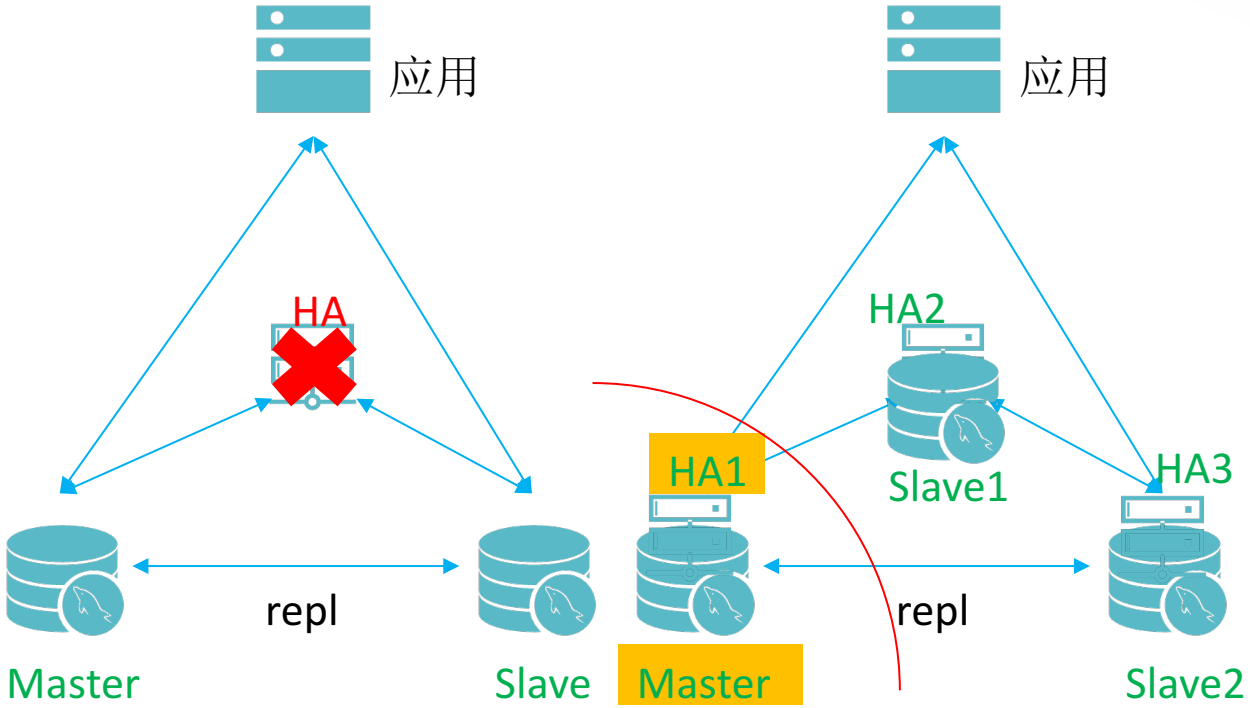
级别	服务等级描述	切换时间	主从延时
T1	数据最多补全 10 分钟，数据最多丢失 0	10分钟	<10分钟
T2	数据最多补全 10 分钟，数据最多丢失 60s	10分钟	<11分钟
T3	数据最多补全 10 分钟，数据最多丢失 15 分钟	10分钟	<25分钟
TE	数据最多补全 10 分钟，数据丢失超过极限 15 分钟, 不切换	-	>25分钟

高可用方案选型

集群可用性

- HA自身可用性
- 脑裂

高可用方案选型



前提

HA集群化部署，角色对称
一致性选举出leader
超过众数节点才能选举

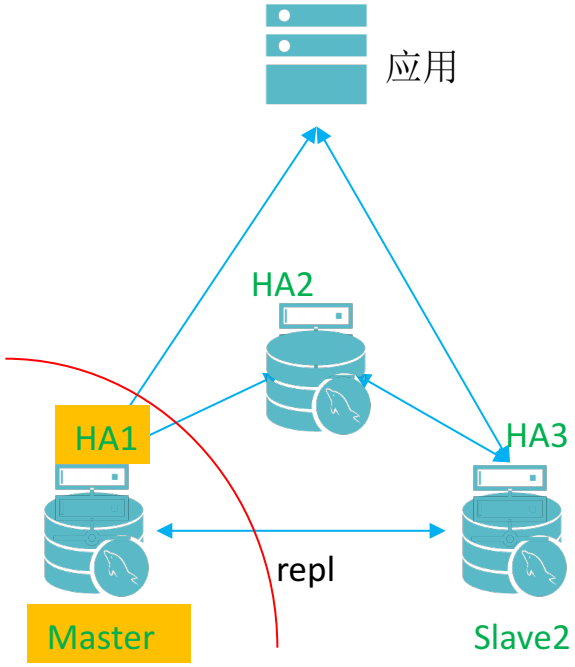
实现

解决Ha mgr单点问题
解决脑裂问题

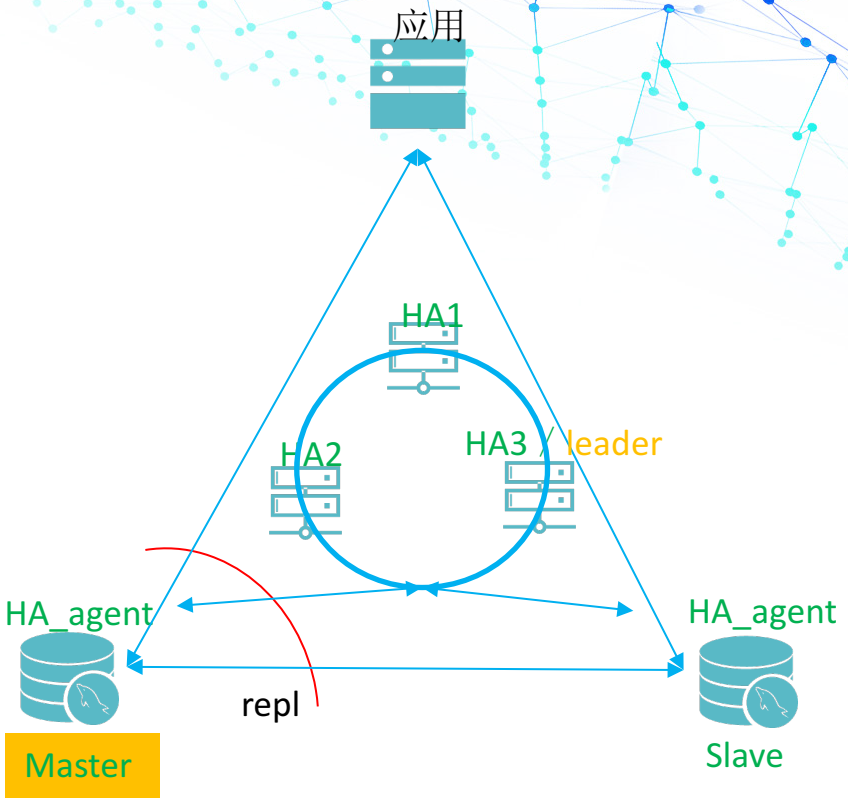
限制

至少需要三个节点
(非数据库)

高可用方案选型



图一：适合小集群



图二：适合大集群

高可用方案选型-小结

- 一致性&性能
  半同步
  共享存储
- 一致性&连续性
  SLA
- 集群可用性
  一致性选举

FAQ1

数据一致性延伸探讨

- **绝对一致性**

  绝对一致性：主备发生切换，主节点和备节点之间的数据完全一致. 如果主备间数据不一致，即发生数据丢失.

- **可见一致性**

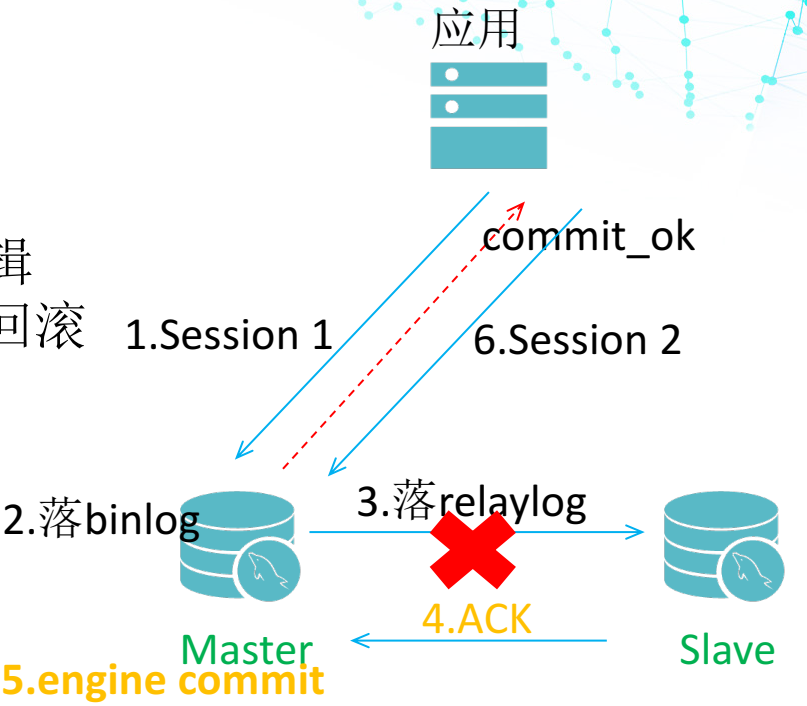
  可见一致性，主备发生切换

  – 对业务承诺的数据（其他session 可见），备节点一定落盘
  – 对业务未承诺的数据（其他session 不可见），备节点可能落盘

FAQ1

可见性一致

- 应用程序处理出错逻辑
- 事务可以选择提交或回滚

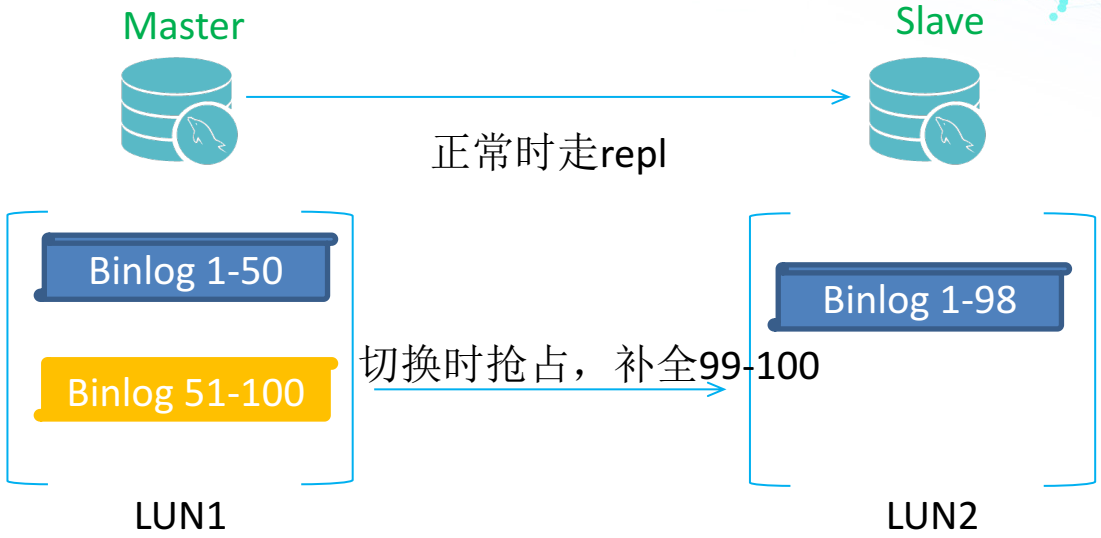


主从架构

FAQ1

绝对一致

- 应用程序处理出错逻辑
- 事务可以选择提交或回滚



FAQ2

数据回滚 or 数据补全的选型

- **数据回滚**

  只支持部分SQL，场景支持不全

- **数据补全**

  入binlog的数据都能回放

  半同步，主节点crash / 断网，可能导致老master出现“绝对一致性”问题

高可用我们还做了更多

1. 统一访问IP
2. 一机N实例
3. N个高可用组
4. 状态监测和故障切换
5. 复制延时检测
6. 复制状态监控
7. 复制自动修复
8. 故障告警
9. 策略化备份恢复
10. 半同步方式的数据零丢失
11. 共享存储方式的数据零丢失
12. 支持SLA协议
13. 故障MATER重启修复
14. 故障SLAVE自动重建

Thanks

关注开源数据库论坛


ACTION