MOG-HA

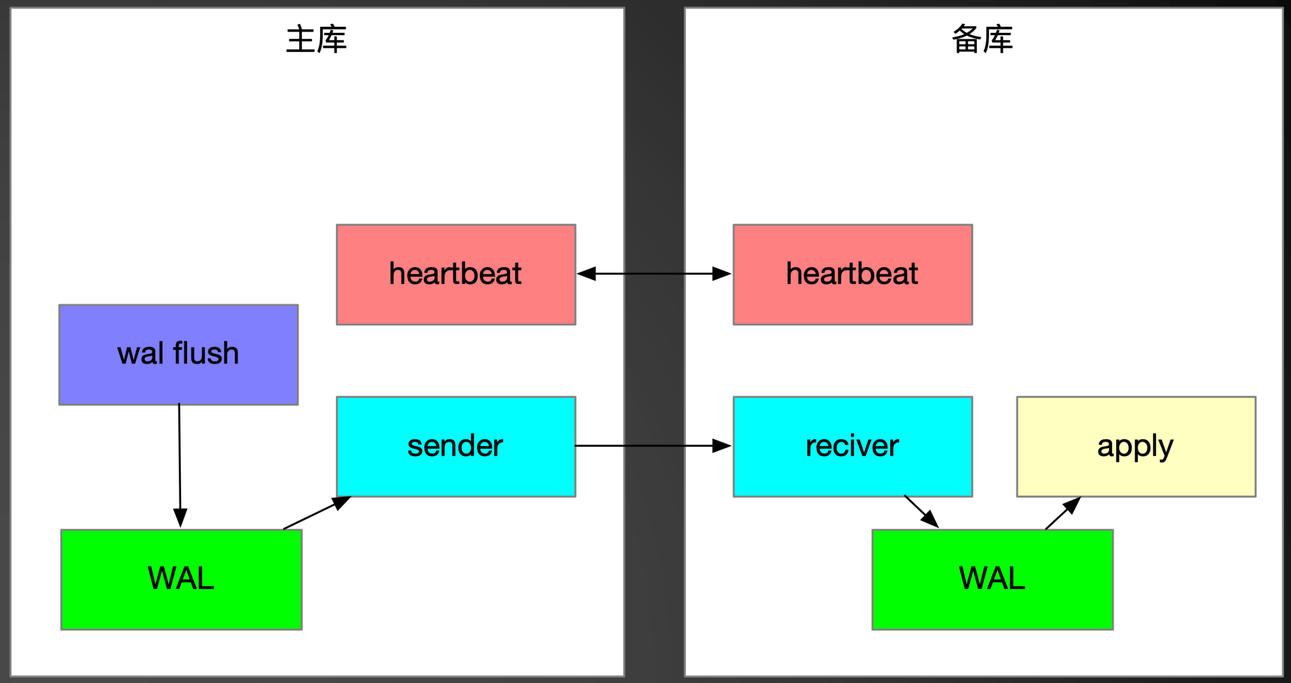
刘伟@云和恩墨

OPENGAUSS

- 基于PostgreSQL改造
- 物理流复制
- 快捷切换命令
- 支持级联

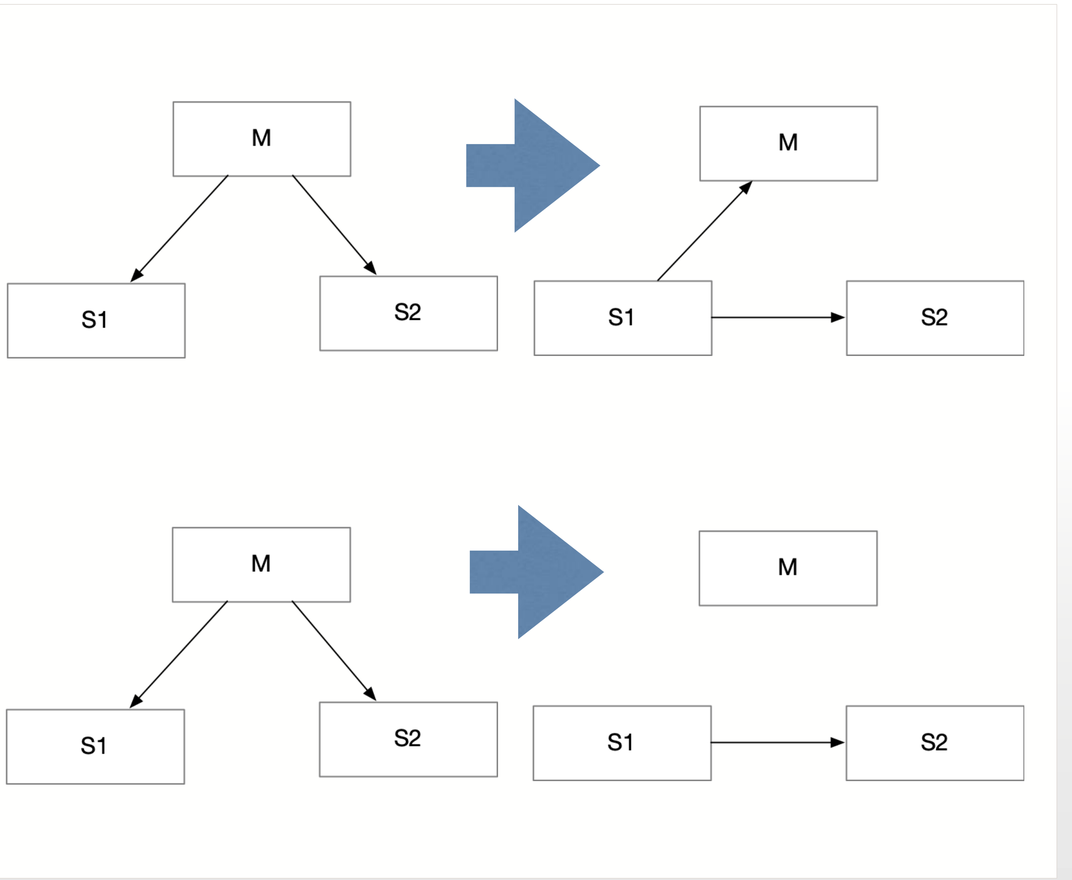
OPENGAUSS物理流复制

- 基于wal日志
- 多线程
- replconninfo[n]
  - 监听端口
  - 心跳端口
  - 安全认证
- 性能
  - 多路重放



OPENGAUSS切换命令 FAILOVER 与 SWITCHOVER

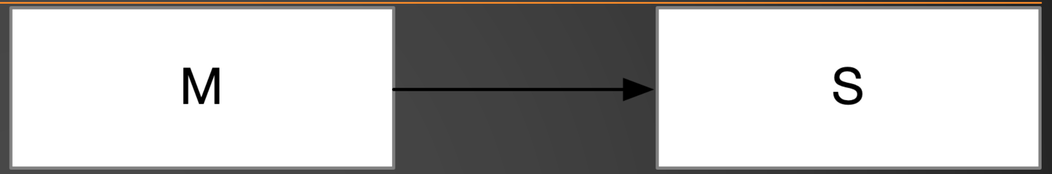
- failover
  - 备库执行后成为新主库
  - 老主库不处理（脑裂危机）
- switchover
  - 备库执行后成为新主库
  - 老主库成为新主库的备库



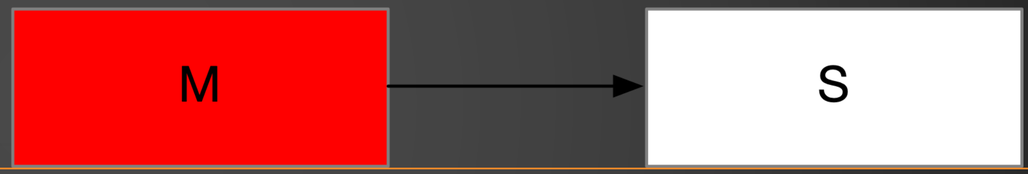
双机HA

双机HA

一主一备

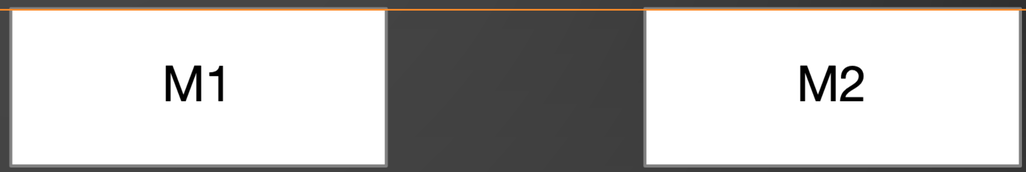


主库挂掉找备库



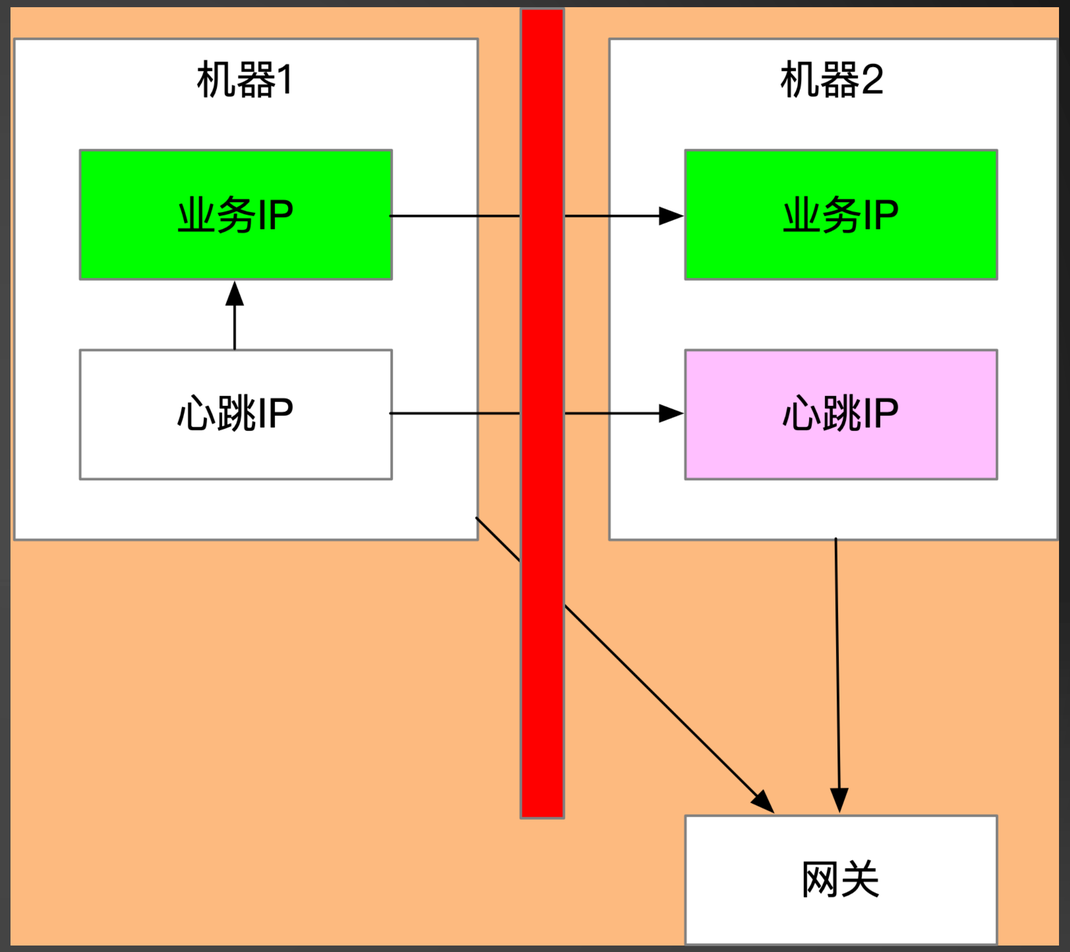
基于VIP

脑裂风险



双机HA—防脑裂

- 独立仲裁（可选）
  - 全局共用
  - Etcd保障高可用
  - 多数派决策？
- 网关检查
- 心跳网络
  - 主库孤单检查：到所有其他机器均ping不通
  - 备库丢主检查：除了主库的ip与心跳ip之外，所有ip均ping通



双机HA—结构自动发现与防脑裂

- 主备角色互换
  - 外部触发/内部触发
  - VIP跟随变更
- 双主
  - 谁是最新设置的主（时间戳）
  - 关闭老主

主库心跳

- 孤单检查
  - 超时自杀
- 双主检查
  - 非新主自杀
- 备库检查
  - 确认有且仅有一个备库

备库心跳

- 下线处理
  - 备库永远不挂VIP
- 主库丢失判断
- 复制检查

FALIOVER/SWITCH操作

- 询问确认（显示主备lsn信息）
- 写入本地primaryinfo
- 切换主备角色
- （维护性failover情况下）关闭老主库（强制切换）
- （维护性switchover情况下）下线老主库（平滑切换）
- 上线VIP

故障场景—OS问题

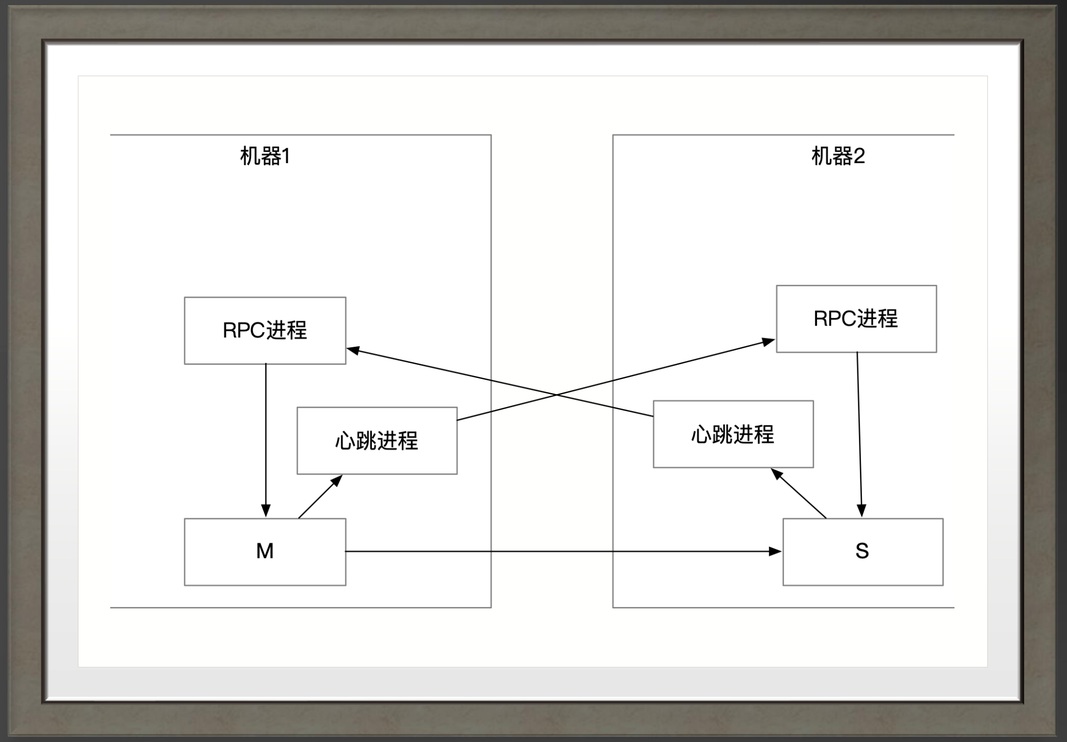
故障场景	现象	处理方式
主库进程挂掉	主库ping通，进程不在	不进行切换，等待人工介入
主库宕机	主库ping不通	超时（默认设置60s）进行自动化切换
主库机器hang死	主库（有可能）ping通，ssh不通	Ping不到的情况下，超时（默认60s）进行自动切换
主库机器过慢	主库ping通，ssh过慢	不进行切换，确实需要人工介入，需要调用手工failover

故障场景	现象	处理方式
主备业务网络隔离	主备业务网络相互不通	不切换
主备心跳网络隔离	主备心跳网络相互不通	不切换
主网络完全隔离	主双网络全部不通	主库超时（默认60s）后备shutdown，备超时（默认60s）后failover为主
备网络完全隔离	主双网络全部不通	不切换
备库与第三方隔离	备库与第三方网络不通	不切换
主库与第三方隔离	主库与第三方网络不通	不切换
全隔离	三者网络全部不通	主库超时（默认60s）后备shutdown，当网络恢复同时恢复，备无操作 当主库依然不通，备切换为主

故障场景—网络问题

- 最简化配置
- 双机一致配置
- 进程命名
  - 心跳进程
  - （web）RPC接口进程
    - 权限控制

一主一备



事后恢复与检查

切换时记录LSN

当老主库机器还能回来

- waldump工具
- 数据变更检查
- 数据重要性检查
  - 回写？

一主多备

一主多备与一主一备的区别

- 更复杂的脑裂场景
- 更复杂的切换决策
  - 选主
- 网络条件差别
  - 心跳网络
- 跨机房
  - **VIP**方式兼容
- 读写分离

一主多备下脑裂处理

自动化

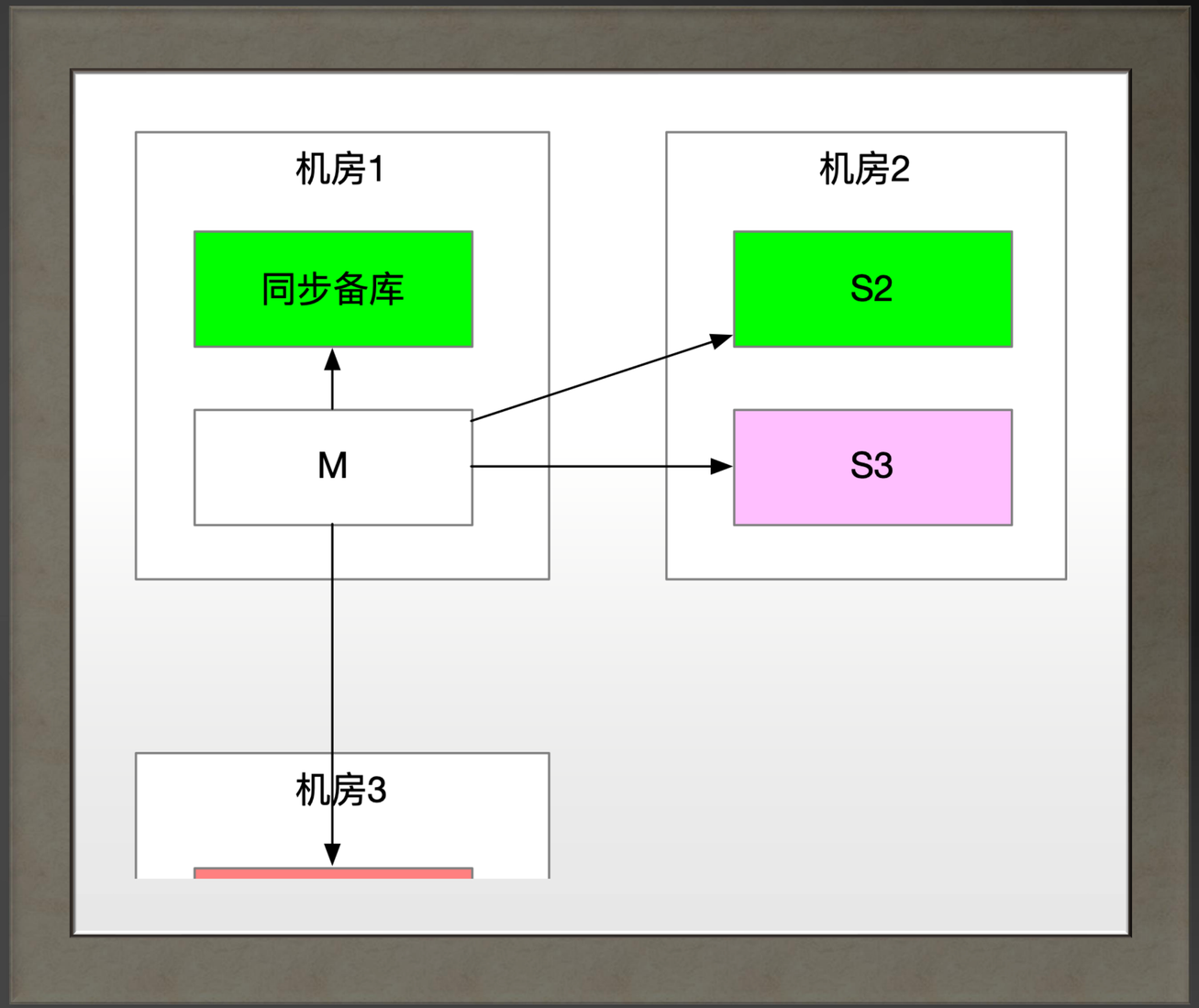
- 主机房内脑裂

人工处理

- 跨机房出现脑裂

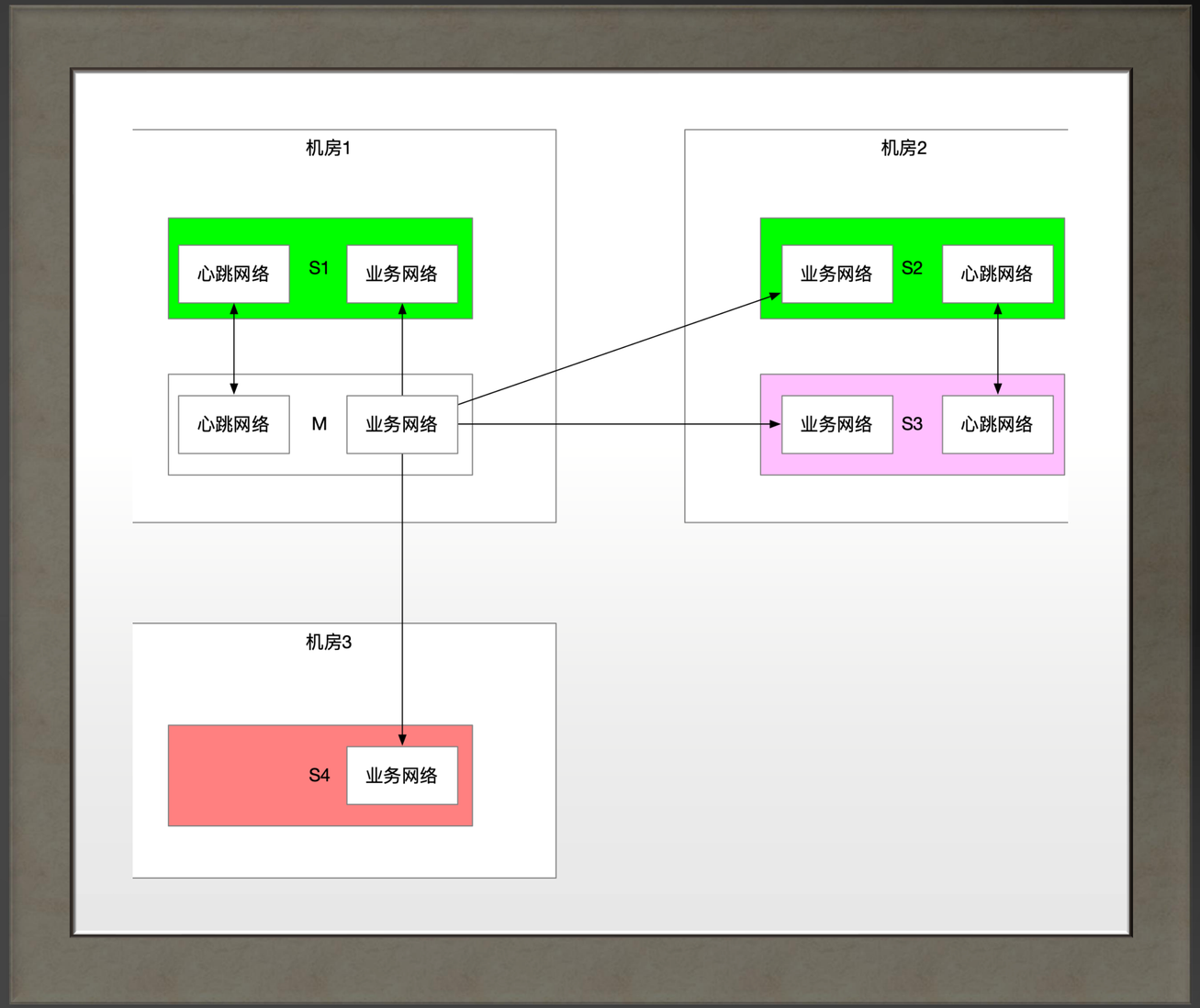
多备之间选主

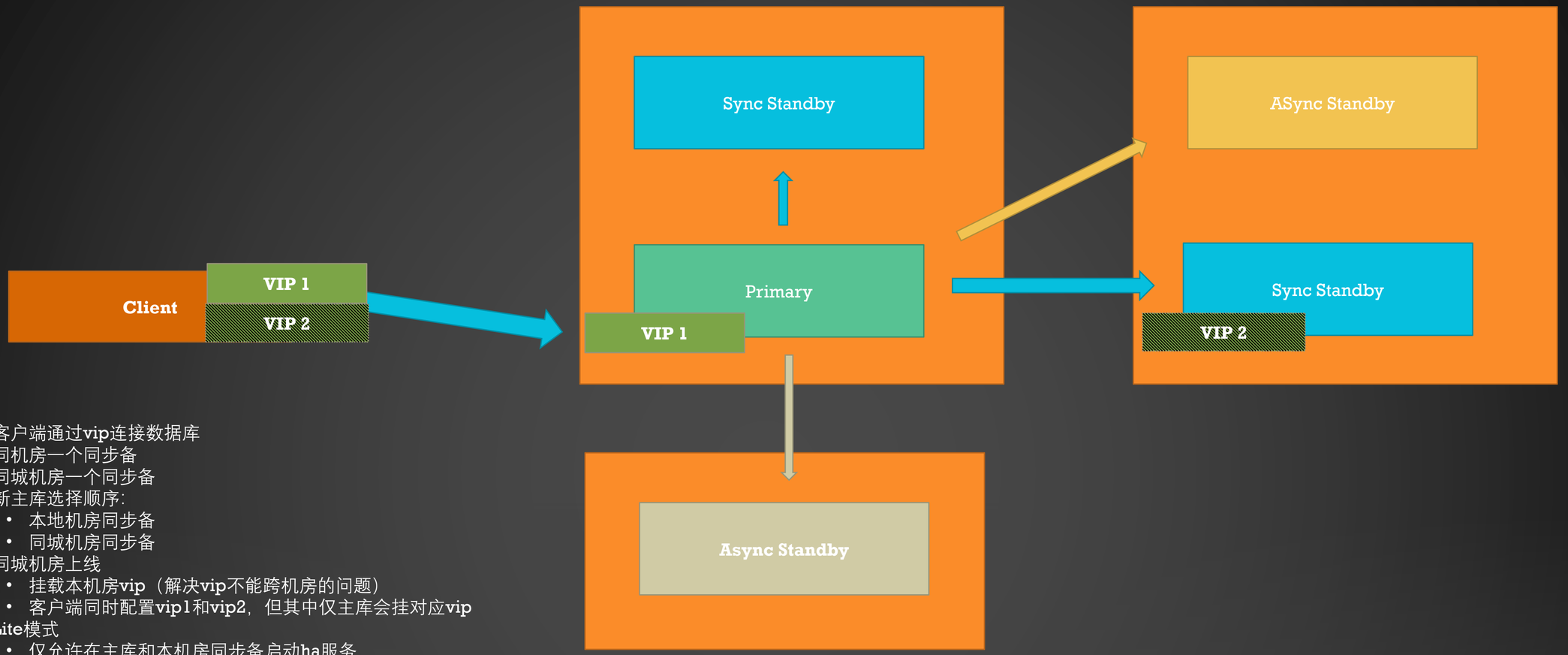
- 同机房自动
  - 固定切换到同机房备库
- 跨机房手动
- **sync names**保证同机房备必然是同步的



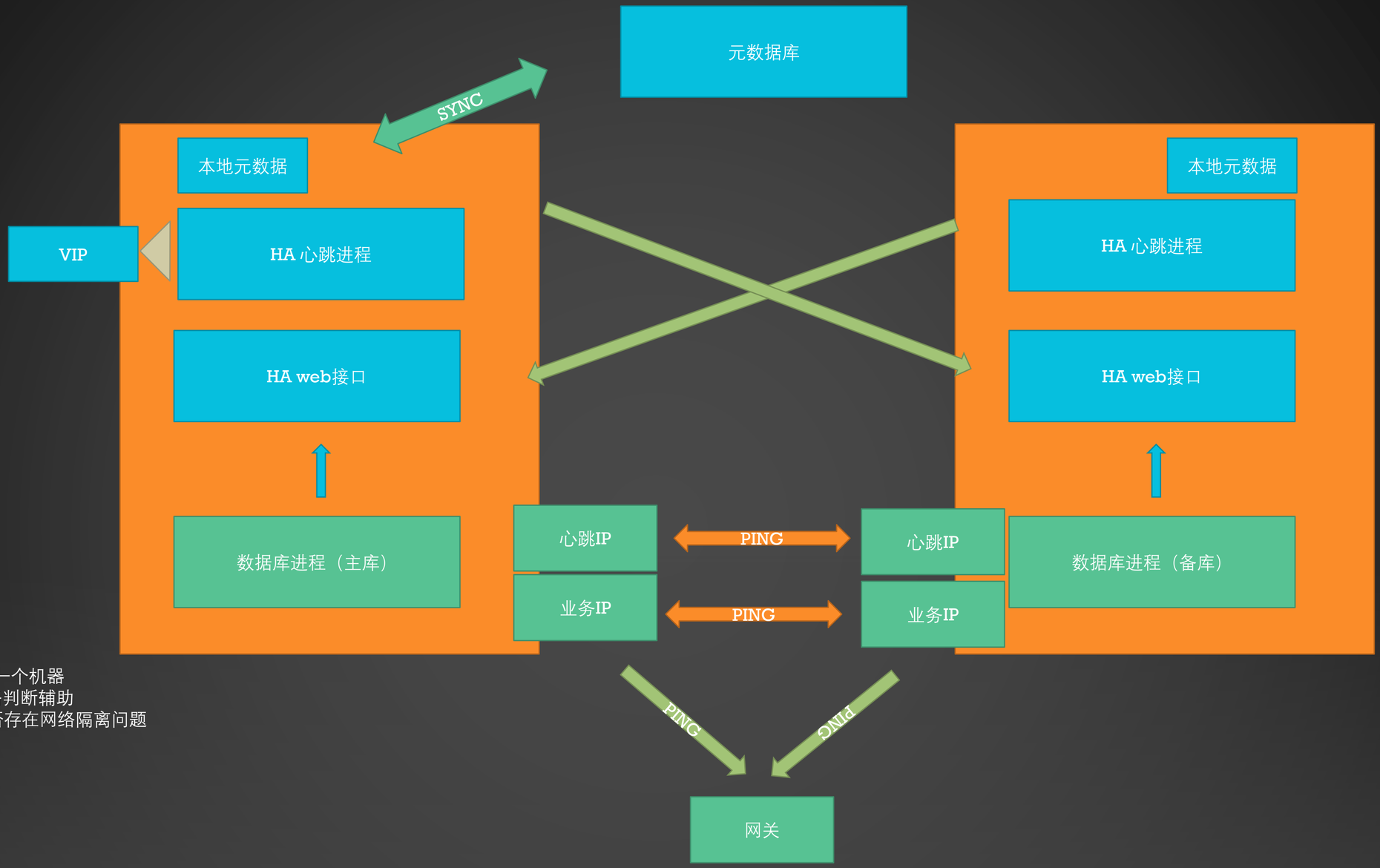
一主多备网络

- 多网卡
  - 机房内心跳
- 辅助判断
  - 多机房仲裁节点
    - 非交互式，仅需要ping
  - 分布式网关
    - 有时候通，有时候不通





- 客户端通过vip连接数据库
- 同机房一个同步备
- 同城机房一个同步备
- 新主库选择顺序：
  - 本地机房同步备
  - 同城机房同步备
- 同城机房上线
  - 挂载本机房vip（解决vip不能跨机房的问题）
  - 客户端同时配置vip1和vip2，但其中仅主库会挂对应vip
- Lite模式
  - 仅允许在主库和本机房同步备启动ha服务
  - 支持单次切换，完成切换后需要人工参与事后处理
  - 不会对数据库配置做任何变更
- Full模式
  - 需要在所有实例都启动ha服务
  - 同步备宕机后，自动提升异步备为同步备
  - 允许人工不参与情况下的连续切换
  - 需要变更数据库高可用相关配置



- HA进程和数据库运行在同一个机器
- 机器建议有心跳ip作为网络判断辅助
- 通过网关检查当前机器是否存在网络隔离问题
- HA心跳进程
  - 自动在主库挂载VIP
  - 检查是否存在脑裂
  - 自动在备库卸载VIP
  - Failover操作者
- HA web接口
  - 机器之间通讯用
- 元数据库
  - 一个openGuass库
  - 如果不存在（小规模ha部署，或者元数据库宕机的情况下），可以使用本地元数据完成切换

跨机房VIP

JDBC 多节点

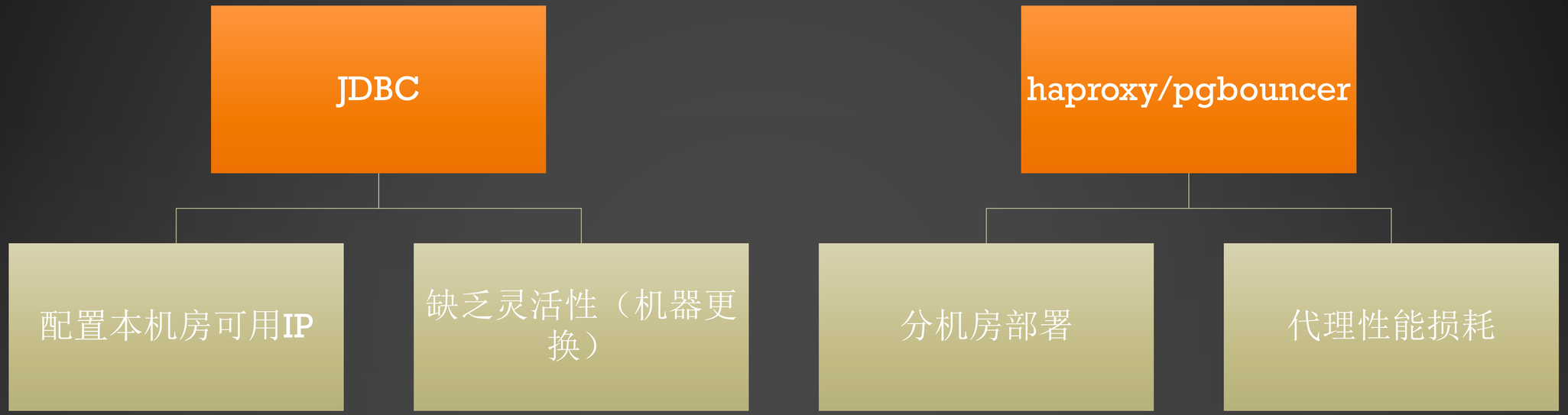
- 仅判断读写
- 连接失败轮询重试

多机房VIP

- 每个机房一个VIP
- Jdbc串中仅一个VIP可以通

读写分离

- 读写分离网关（pgpool）？
  - 不建议
  - 部署过重
  - 兼容性
- 开发控制读写
  - 开发工作较重
  - 读写分离场景区分
  - 运维读写分离关注



读负载均衡

Q&A