Patroni

介绍

一、 介绍

1 Patroni 简介

patroni 是一款运用 etcd 集群来检测、存储数据库节点的主备状态与配置,并且通过 patroni 来实现自动切换的软件。运用 haproxy+keepalived 保持在主备切换或者节点故障后，访问地址、端口对上层不变。使用一套模板化的配置文件来自动搭建初始化数据库流复制集群以及配置数据库。patroni 高可用集群由 postgresql,patroni,etcd,haproxy,keepalived 组成。

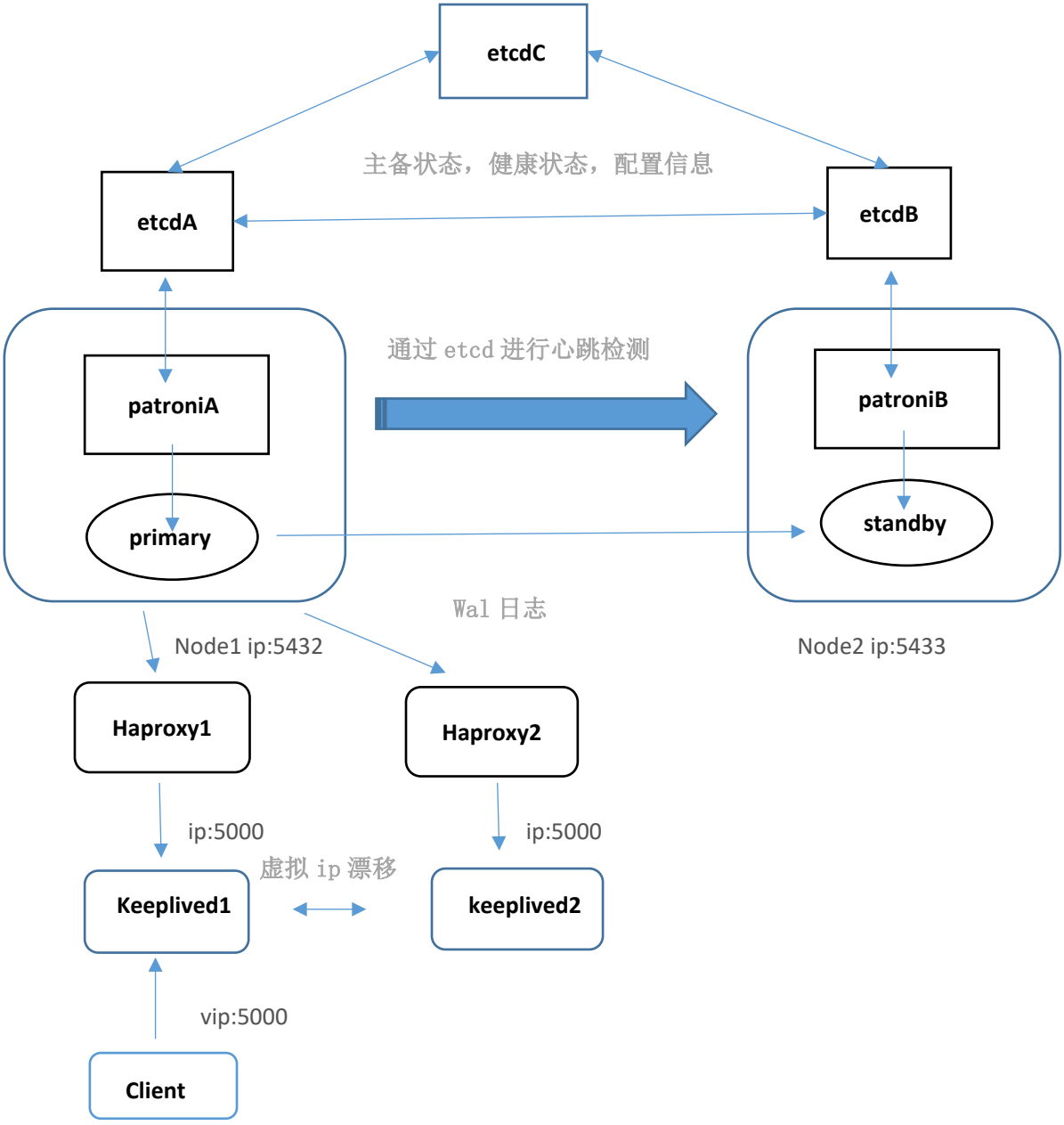
组件分别的作用(不包括 pg 数据库):

- patroni:通过参数文件来配置自动初始化数据库搭建流复制（配置 pg 参数文件、创建用户、可以配置预加脚本），指定 etcd 节点等。负责通过一个 api 接口连接到 dcs(分布式存储系统:etcd 集群)，向其插入键值记录 patroni 参数、数据库参数、主备信息以及连接信息。平常通过 etcd 对其它节点做心跳检测。与数据库的主备切换或者做恢复时通过向 etcd 拿取键值中储存的主备信息来判断各节点的状态进行切换。

- etcd:最少需要三个节点且为奇数来进行 leader 选举（脑裂发生时会 etcd 集群会僵死等待恢复，不会发生都认为自己是主的情况）。在各个节点上同步健康状态信息以及数据库节点的主备状态与连接、配置信息。平常会对其余节点做心跳检测。

- haproxy+keepalived: haproxy 可以代理主节点，并统一由其所在节点的 5000 端口发出。Keepalived 负责产生虚拟 ip 和虚拟 ip 漂移,数据库发生主备切换或者节点故障后，访问地址对上层不变。

2 patroni 流程及介绍



i. 基本流程

- Patroni 自动创建主备流复制集群通过 api 接口往 etcd 记录键值来储存主备信息与连接信息以及配置信息
- Etcd 进行心跳检测（etcd 之间的心态检测）与存储键值信息
- patroni 通过连接 etcd 对其它节点做心跳检测，每 loop_wait 秒一次
- patroni 通过连接到 etcd 集群，向其插入键值记录 patroni 参数、数据库参数、主备信息以及连接信息。进行数据库的主备切换时通过向 etcd 拿取键值中储存的主备信息来判断各节点的状态来切换。各节点会在 data 目录下生成 recovery.done(与 recovery.conf 一样，里面的 primary_conninfo 记录是上一次主节点的连接信息)，原主节点发生切换时自动改变后缀为 recovery.conf，原备节点会删除掉自身的 recovery.conf 文件，再通过 pg_rewind 来快速恢复节点，不需要做基础备份。
- 异步流复制时主从之间延时：主从之间 wal 日志延时超过 maximum_lag_on_failover(byte)的大小，主备有可能会重启但不会发生切换。
  数据丢失量通过 maximum_lag_on_failover，ttl,loop_wait 三个参数控制。
  最坏的情况下的丢失量：
  maximum_lag_on_failover 字节+最后的 TTL 秒时间内写入的日志量（loop_wait /2 在平均情况下）。
- haproxy+keeplived 保持对外的访问 ip 端口不变

ii. 优势

- 自动检测主备状态进行切换
- 统一模板配置
- 在上图最基本的架构中，任意 down 一个 etcd 节点或者任意一个 patroni 节点、数据节点，通过转换都能使集群继续运行下去。测试中有针对于三个 etcd 网络互不通做了脑裂测试，故障发生后 etcd 集群会僵死等待恢复，不会发生都认为自己是主的情况。主库会变成只读状态，恢复网络后，主、备继续用 etcd 的数据信息恢复到故障前的状态，etcd 也恢复正常。
- 在线添加 etcd、patroni 节点以及数据节点
- 支持同步异步流复制，级联流复制
- 异步流复制可设置最小丢失数据量
- 使用 pg_rewind 进行恢复，缩短恢复时间。
- haproxy+keeplived 可以保持在主备切换或者节点故障后，实现 ip 漂移。对外的 ip+端口不变。

iii. 限制

- patroni 对数据库操作需要普通用户
- 需要至少三个以上且为奇数的 etcd 节点
- 底层基于的是流复制

- 大部分参数都需要通过更改 etcd 中键值来修改
- 因故障发生的连接会回滚，但是需要客户端重新发起连接