Keepalived高可用概念篇

Keepalived介绍

Keepalived 软件起初是专为 LVS 负载均衡软件设计的，用来管理并监控 LVS 集群系统中各个服务节点的状态，后来又加入了可以实现高可用的 VRRP 功能。因此，Keepalived除了能够管理 LVS 软件外，还可以作为其他服务（例如：Nginx、Haproxy、MySQL等）的高可用解决方案软件。

Keepalived 软件主要通过 VRRP 协议实现高可用功能的，VRRP 是 Virtual Router Redundancy Protocol （虚拟路由器冗余协议）的缩写，VRRP 出现的目的就是为了解决动态路由单点故障问题的，它能够保证当个别节点宕机时，整个网络可以不间断的运行。所以，Keepalived 一方面具有配置管理 LVS 的功能，同时还具有对LVS 下面节点进行健康检查的功能，另一方面也可以实现系统网络服务的高可用功能。Keepalived 软件的官方站点： http://www.keepalived.org

Keepalived 服务的三个重要功能

1）管理 LVS 负载均衡软件

早期的 LVS 软件，需要通过命令行或脚本实现管理，并且没有针对 LVS 节点的健康检查功能。为了解决 LVS 的这些使用不便的问题，Keepalived就诞生了，可以说，Keepalived软件起初是专为了解决 LVS 的问题而诞生的。因此，Keepalived和LVS的感情很深，它们的关系如同夫妻一样，可以紧密的结合，愉快的工作。Keepalived 可以通过读取自身的配置文件实现通过更底层的接口直接管理 LVS 的配置以及控制服务的启动、停止等功能，这使得 LVS 的应用就更加简单方便了。

2）实现对 LVS 集群节点健康检查功能（healthcheck）

Keepalived 可以通过在自身的keepalived.conf文件里配置 LVS 的节点 IP 和相关参数实现对 LVS 的直接管理；除此之外，当 LVS 集群中的某一个甚至是几个节点服务器同时发生故障无法提供服务时，Keepalived 服务会自动将失效的节点服务器从 LVS 的正常转发队列中清

踢出去，并转换到别的正常节点服务器上，从而保证最终用户的访问不受影响；当故障的节点服务器被修复后，Keepalived 服务又会自动地把它们加入到正常转发队列中，对客户提供服务。

3）作为系统网络服务的高可用功能（failover）

Keepalived 可以实现任意两台主机之间，例如 Master 和 Backup 主机之间的故障转移和自动切换，这个主机可以是普通的不能停机的业务服务器，也可以是 LVS 负载均衡、Nginx 反向代理这样的服务器。

Keepalived 高可用功能实现的原理为：两台主机同时安装好 keepalived 软件并启动服务，开始正常工作时，由角色为 Master 的主机获得所有资源并对用户提供服务，角色 Backup 的主机作为 Master 主机的热备；当角色为 Master 的主机失效或出现故障时，角色为 Backup 的主机将自动接管 Master 主机的所有工作，包括接管 VIP 资源及相应资源服务；而当角色为 Master 的主机故障修复后，又会自动接管回它原来处理的工作，角色为 Backup 的主机则同时释放 Master 主机失效它接管的工作，此时，两台主机将恢复到最初启动时各自的原始角色及工作状态。

Keepalived 高可用故障切换转移原理（重点）

Keepalived 高可用服务对之间的故障切换转移，是通过 VRRP 协议（虚拟路由冗余协议）来实现的。

在 Keepalived 服务正常工作时，主 Master 节点会不断地向备节点发送（多播的方式）心跳消息，用以告诉备 Backup 节点自己还活着，当主 Master 节点发生故障时，就无法发送心跳消息了，备节点也就因此无法继续检测到来自Master 节点的心跳了，进而调用自身的接管程序，接管主 Master 节点的 IP 资源及服务。而当主 Master 节点恢复时，备 Backup 节点又会释放主节点故障时自身接管的 IP 资源及服务，恢复到原来备用角色。

VRRP协议

VRRP 协议，全称 Virtual Router Redundancy Protocol，中文名为虚拟路由冗余协议，VRRP 的出现就是为了解决静态路由的单点故障问题，VRRP 协议是通过一种竞选机制来将路由的任务交给某台 VRRP 路由器的。VRRP 协议早期是用来解决交换机、路由器等设备单点故障的。

1）VRRP 原理描述（同样适用于 Keepalived 的工作原理）

在一组 VRRP 路由器集群中，有多台物理 VRRP 路由器，但是这多台物理的机器并不是同时工作的，而是由一台称为 MASTER 的机器负责路由工作，其他的机器都是 BACKUP。MASTER 角色并非一成不变，VRRP 协议会让每个 VRRP 路由参与竞选，最终获胜的就是 MASTER。MASTER 拥有虚拟路由器的 IP 地址，我们把这个 IP 地址称为 VIP，MASTER 负责转发发送给网关地址的数据包和响应 ARP 请求。

2）VRRP 是如何工作的?

VRRP 协议通过竞选机制来实现虚拟路由器的功能，所有的协议报文都是通过 IP 多播（默认的多播地址：224.0.0.18）形式进行发送。虚拟路由器由 VRID （范围0-255）和一组 IP 地址组成，对外表现为一个周知的 MAC 地址：00-00-5E-00-01-{VRID}。所以，在一个虚拟路由器中，不管谁是 MASTER，对外都是相同的 MAC 地址和 IP 地址，如果其中一台虚拟路由器宕机，角色发生切换，那么客户端并不需要因为 MASTER 的变化修改自己的路由设置，可以做到透明的切换。这样就实现了如果一台机器宕机，那么备用的机器会拥有 MASTER 上的 IP 地址，实现高可用功能。

3）VRRP 是如何通信的?

在一组虚拟路由器中，只有作为 MASTER 的 VRRP 路由器会一直发送 VRRP 广播包，此时 BACKUP 不会抢占 MASTER 。当 MASTER 不可用时，这个时候 BACKUP 就收不到来自 MASTER 的广播包了，此时多台 BACKUP 中优先级最高的路由器会去抢占为 MASTER。这种抢占是非常快速的（可能只有1秒甚至更少），以保证服务的连续性。出于安全性考虑，VRRP 数据包使用了加密协议进行了加密。

Keepalived 高可用服务脑裂问题

什么是脑裂?

由于某些原因，导致两台高可用服务器在指定时间内，无法检测到对方的心跳消息，各自取得资源及服务的所有权，而此时的两台高可用服务器都还活着并在正常运行，这样就会导致同一个 IP 或服务在两端同时存在发生冲突，最严重的是两台主机占用同一个 VIP 地址，当用户写入数据时可能会分别写入到两端，这可能会导致服务器两端的数据不一致或造成数据丢失，这种情况就被称为脑裂。

导致脑裂发生的原因

一般来说，脑裂的发生，有以下几种原因：

1）高可用服务器之间心跳线链路故障，导致无法正常通信。

- 心跳线坏了（包括断了，老化）
- 网卡及相关驱动坏了，IP 配置及冲突问题（网卡直连）
- 心跳线连接的设备故障（网卡及交换机）

2）高可用服务器上开启了 iptables 防火墙阻挡了心跳消息传输。

3）高可用服务器上心跳网卡地址等信息配置不正确，导致发送心跳失败。

4）其他服务配置不当等原因，如心跳方式不同，心跳广播冲突、软件 BUG等。

注意：Keepalived 配置里同一 VRRP 实例如果 virtual_router_id 参数两端配置不一致，也会导致脑裂问题发生。

解决脑裂的具体方案

在实际生产环境中，可以从以下几个方面来防止脑裂问题的发生

1）同时使用串行电缆和以太网电缆连接，同时用两条心跳线路，这样一条线路坏了，另一个还是好的，依然能够传送心跳消息

2）当检测到脑裂时强行关闭一个心跳节点（这个功能需要特殊设备支持，如Stonith、fence）。相当于备节点接收不到心跳消息，发送关机命令通过单独的线路关闭主节点的电源。

3）做好对脑裂的监控报警（如邮件及手机短信等或值班），在问题发生时人为第一时间介入仲裁，降低损失。例如，百度的监控报警短信就有上行和下行的区别。报警信息报到管理员手机上，管理员可以通过手机回复对应数字或简单的字符串操作返回给服务器，让服务器根据指令自动处理相应故障，这样解决故障的时间更短。

4）如果开启防火墙，一定要让心跳消息通过，一般通过允许 IP 段的形式。