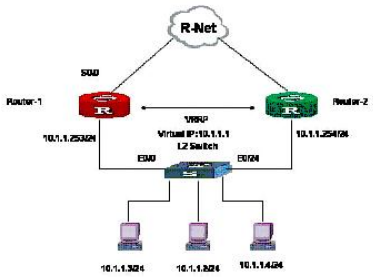
VRRP分析

郭晓翔 2006-02-15 发表

VRRP分析

VRRP（Virtual Router Redundancy Protocol）是一种容错协议。通常，一个网络内的主机设置一条缺省路由（下一跳为10.1.1.1），这样，主机发出的目的地址不在本网段的报文将通过缺省路由发往路由器Router，从而实现了主机与外部网络的通信。当路由器Router发生故障时，本网段内所有以Router为缺省路由下一跳的主机将断掉与外部的通信。

VRRP就是为解决上述问题而提出的，它为具有多播或广播能力的局域网（如：以太网）设计。VRRP将局域网的一组路由器（包括一个MASTER和若干个BACKUP）组织成一个虚拟的路由器，称之为一个备份组。



这个虚拟的路由器（即备份组）拥有自己的IP地址10.1.1.1（这个IP地址可以和备份组内的某个路由器的接口地址相同），备份组内的路由器也有自己的IP地址（MASTER的IP地址为10.1.1.253，BACKUP的IP地址为10.1.1.254）。局域网内的主机仅仅知道这个虚拟路由器的IP地址10.1.1.1，而并不知道具体的MASTER路由器的IP地址10.1.1.253以及BACKUP路由器的IP地址10.1.1.254，他们将自己的缺省路由设置为该虚拟路由器的IP地址10.1.1.1。于是，网络内的主机就通过这个虚拟的路由器来与其它网络进行通信。如果备份组内的MASTER路由器坏掉时，备份组内的其它BACKUP路由器将会接替成为新的MASTER，继续向网络内的主机提供路由服务。从而实现网络内的主机不间断地与外部网络进行通信。

在上图中，假定VRRP组号为10，VRRP的虚拟IP为10.1.1.1,主、从路由器IP为10.1.1.253和10.1.1.254,下接二层交换机，用户PC接入交换机，设定PC的网关为10.1.1.1。

VRRP工作原理：当Router-1和Router-2都启动VRRP后，他们会以224.0.0.18的组播地址发送VRRP协议报文，其中包含的主要信息为[组号,Virtual IP,hello time, priority,preempt-mode,authentication-mode],而我们主要关心的参数是组号和优先级。当Router-1和Router-2收到对方的VRRP报文后，经过比较，Router-1发现其它参数都一致（如组号，认证方式等），但它的优先级比Router-2高（假定它为120，而Router-2为100），这时，它就会认为它是Master,同时Router-2也收到了Router-1的参数时，因级别不够，100＜120，只好屈服在认为自己是Slave，双方相安无事，这种状态就保持了下来。而做为Master的一方，会发送一个免费的ARP报文（gratuitous arp）,此ARP报文是通告了IP和MAC的对应关系，不需要网络中的设备应答，而这个关系就是虚拟IP和虚拟MAC的对应关系，而和它在同一网络中的网络设备（包括PC），会在自己的ARP表中添加这样一条记录。本例中：这条ARP记录是[IP:10.1.1.1, MAC 00-00-5e-00-00-组号],而MAC的最后一位是组号的十六进制，如组号为10，那么最后一位就是0A。这样对于主机PC来看，当它发送的报文通过缺省网关时，就会发送给Master,因为它的ARP记录中记录是主设备公告的。

当主机接入的设备如图示，是L2交换机时，会发生什么？简单来看，当两个PC接入同一L2交换机时，PC1 PING PC2，最初没有ARP信息，对于交换机来说要依据MAC地址转发，这里PC1发送一个ARP请求报文，以广播形式发送，L2下的同一VLAN的所有主机都会收到这个报文，如果发现其中IP是自己的PC，就会填入自己的MAC，以ARP响应方式应答，别的PC就会丢弃这个报文。这时，交换机会根据这种ARP的请求和响应报文，获取PC的MAC，形成正常的转发。同样的原理，对于VRRP来说，因为Master发送了免费ARP，这个报文中的MAC地址也被L2学习到了，如图示，此MAC学习到了E0/0接口上，下接的主机一旦有报文转发给网关，L2同时也学习到了PC的MAC，所以形成了正常的转发表（L2主要根据MAC转发报文），而这个MAC其实是虚拟MAC，实际上Router-1和Router-2的实MAC，如果PC访问过，在交换机上也能学习到，同时PC的ARP表项中也有这样的记录。

风云突变：当Master - 1上的S0/0接口Down掉，而它又设置了监控S0/0，如果Down掉VRRP组优先级降低50，这样它的VRRP的优先级就从120 - 50＜100，当它和Slave交换VRRP的报文时，自己感到底气不足，这时Slave就名正言顺地跳了出来，因为它的优先级高，它就会发送免费ARP，通告虚拟IP＋虚拟MAC的关系，对于L2来说，当它监控到这种信息时，相当于更新了MAC，就会把这个虚拟MA

C学习到接Slave的E0/24端口，当PC转发报文给缺省网关时，IP没变，MAC没变，发送给L2，L2转发给新老板 - Router-2(其实这时它变成了Master),Router-2也会按正常流程转发（Router-1正在一旁哭泣这时，但也没忘记在约定的时间和Router-2交换VRRP信息，看能不能"复辟"）。当Router-1的S0/0接口恢复后，它的VRRP优先级就会复原，再和Router-2交换VRRP报文时，发现自己实力恢复了，优先级120 > 100，然后它发送了免费ARP，通告给主机，对于主机来说，反正ARP表项没变化，毕竟虚拟IP + MAC都没变，但L2却认为这是个更新，将虚拟MAC又学习到了接Router-1的E0/0接口，所以当PC转发报文时，就会被L2转发到Router-1(哈哈，Router-1狂笑中，我胡汉三又回来了！）

如果Router-1整个设备Down掉，Router-2在固定的时间没有收到VRRP报文，就会认为中原没有霸主，就会发送免费ARP，重现刚才夺权一幕。

这里的故障点有三：一是在Router-1和Router-2在交换VRRP信息时，可能因为Router-1或是Router-2本身性能问题或是CPU繁忙，没有很好地处理VRRP报文，导致没有及时发送免费ARP，使PC找不着明主。第二点是VRRP的hello时间太短，当内部L2中网络繁忙时，收不到VRRP报文，或是刚收到一个优先级高的，又收到一个优先级低的VRRP报文，主备频繁导换。第三点是L2交换机，它的CPU繁忙或是硬件表有问题，导致学习MAC时较慢或是学错，不能正常转发，有时会将VRRP的报文转发错误，导致网络中有两个Master，所以在排错时，首先查看两个VRRP路由器，看两者的Master和Slave状态，是否和实际情况一致，并调试VRRP报文。其次，查看L2的MAC地址表，查看虚拟MAC是否学习到了正确的端口上。这样就能清楚定位是哪个网络产品的问题。