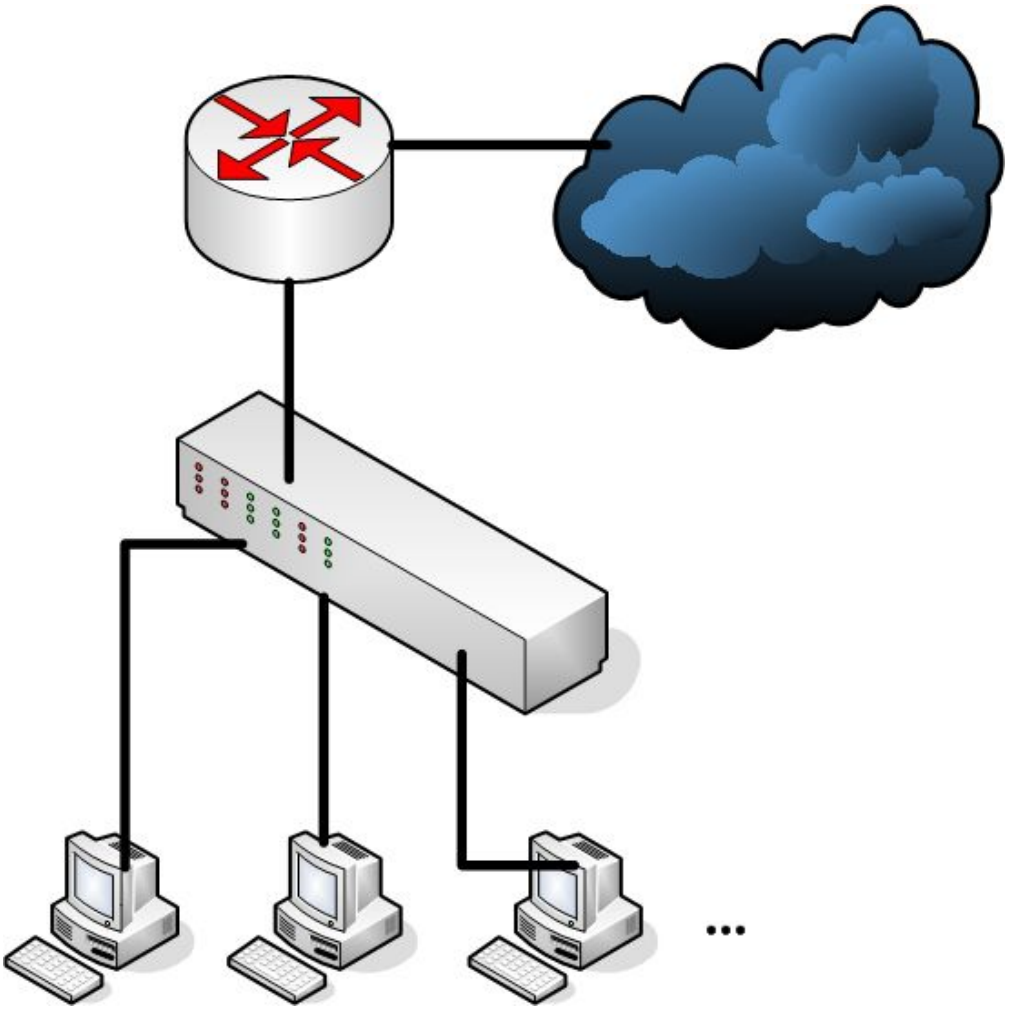
1. VRRP 产生背景及应用环境

1.1 为什么要用 VRRP

VRRP（Virtual Router Redundancy Protocol）------虚拟路由器冗余协议，其最新技术标准是 RFC3768。

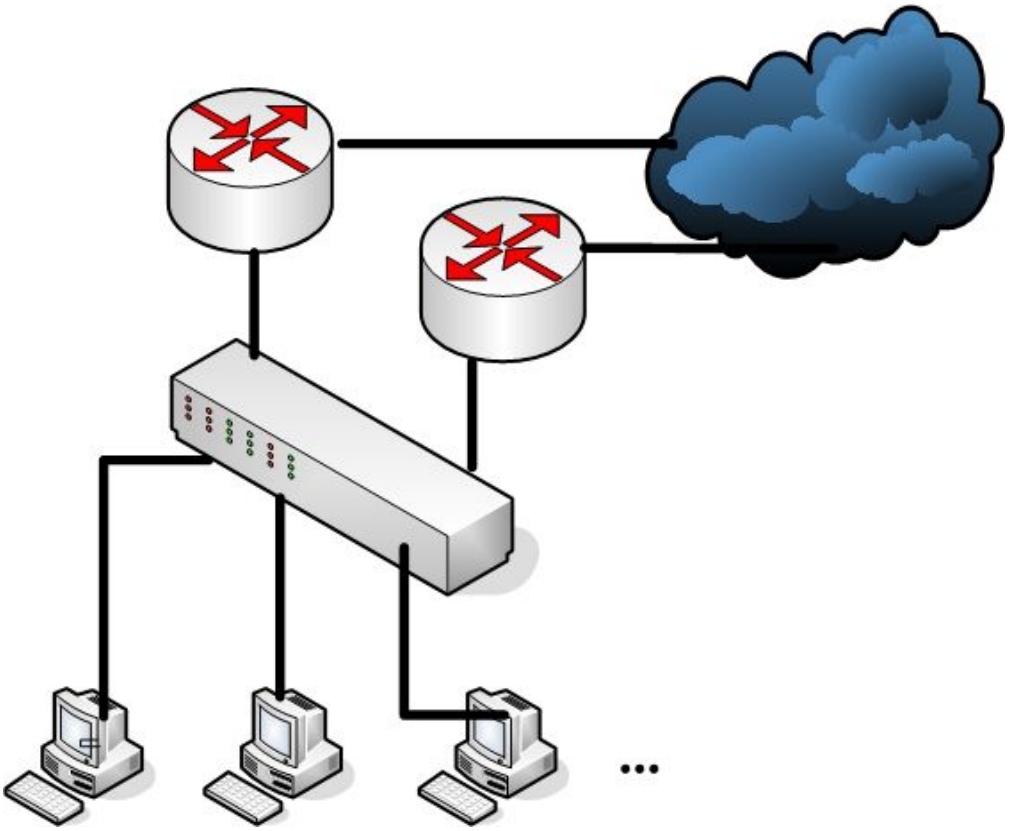
为什么要用 VRRP 呢，主要是为了实现数据链路层互通设备的冗余备份功能，我们来看图一：



图一（常规网络架构）

通过上图可以看到，常规的局域网一般都是多个终端接到交换机上，然后通过单独的出口路由器连接到 Internet，这时候问题来了，如果这个出口路由器坏掉了，那么整个上行的流量就会全部断掉，这就是传说中的单点故障。

所以说我们要避免出现这种情况，本着冗余备份的思想，我们对上面的网络进行物理改造，如下图：



图二（消除了单点故障的常规网络架构）

现在，这个网络一共有两个 Internet 出口，这样任何一个出口路由器出现故障都不会导致终端用户的上行流量断掉。

另外一个问题出现了，我们怎么让终端 PC 知道局域网中有两个出口路由器，并在其中一个出现故障后自动选择另外一个呢？可采用的方案包括让终端 PC 运行动态路由协议，比如 RIP、OSPF，或者 ICMP router Discovery client（DISC），或者指定一条静态缺省路由。

但是这三种实现方法都有其劣势及不可行之处，我们来具体分析一下。首先对于在每一个终端 PC 上运行动态路由协议来讲，几乎是不可能的，这其中牵涉到网管的技术能力和日常维护、安全性问题、以及某些终端平台不支持动态路由协议，比如我们常用的 XP、Windows7 都不支持，而 windows Server 系列 OS 支持。

假定我们在终端 PC 上部署了动态路由协议，那么每一个终端用户都会遇到下面这种情况：
10086：尊敬的用户您好，申报 RIP 故障请按 1，申报 OSPF 故障请按 2，申报 ISIS 故障请按 3…….
用户：(⊙ o ⊙)啊！我家是 OSPF，按 2。

10086：您好，您申报的故障是 OSPF，请进一步选择，OSPF 邻居无法建立请按 1，OSPF 密钥不对请按 2，链路状态数据库异常请按 3，路由表错误请按 4……..
用户：(⊙ o ⊙)啊！……然后吐血身亡……

所以说，N 多现实问题和困难导致在终端 PC 上部署动态路由协议具有不可行性。

那么对于在终端 PC 上部署 DISC 等邻居或路由器发现协议呢？也存在种种问题，例如在网络内存在大量主机，每一台都需要运行 DISC，除了增加主机的处理负担外，也会导致协议收敛缓慢，从而不能及时发现不可用邻居路由器，产生路由黑洞，这是不可接受的。

现在只剩下在终端 PC 上配置静态缺省路由了（其具体表现形式一般是设置网关），这是几乎每一个 IP 平台都支持的配置功能，即使是一部 IP 电话机，根据这个思路，我们在终端上配置多个默认网关即可实现路由备份了，但是存在以下两个问题：

1. 对于下行设备是 PC 来讲，配置了多个默认网关之后，其中一个会作为活动默认网关，其它的作为备份默认网关，其按照下列过程执行流量转发和失效网关检测：

当 TCP/IP 在通过活动默认网关向某个目标 IP 地址进行 TCP 通信时，如果失败的尝试次数达到 TcpMaxDataRetransmissions 注册表值(默认为 5)的一半(即 3 次)还没有收到响应，TCP/IP 将到达该目标 IP 地址的通信改为使用列表中的下一默认网关，这一步是通过更改该远程 IP 地址的路由缓存项（Route Cache Entry，RCE）来实现的，从而使用列表中的下一个默认网关来作为下一跳地址。其中 RCE 是路由表中的一个条目，用于存储目的地的下一跳 IP 地址。当超过 25%的 TCP 连接转向下一默认网关时，TCP/IP 将活动默认网关修改为这些连接当前使用的默认网关。

如果此时原始默认网关从故障中回复，TCP/IP 将继续使用当前的活动默认网关，而不会转移到原始默认网关，除非重启计算机。如果当前的活动默认网关也出现故障，那么 TCP/IP 就会继续尝试使用列表中的下一个默认网关，在尝试完整个列表后将返回到列表的开始，又从第一个默认网关开始进行尝试。

死网关检测仅监视 TCP 流量，如果其他类型的流量连接失败，不会切换默认网关。另外 TCP 是端到端的协议，因此即使当前默认网关完全正常，本地计算机的 TCP 通信失败也可能会导致切换默认网关。

当不同网络接口所连接的网络之间没有连接性时(如一个网络接口连接到 Internet，而一个网络接口连接到内部网络)，如果在多个网络接口上同时配置默认网关，在活动默认网关出现故障导致切换默认网关时，就可能会引起连接性故障。比如活动默认网关为 Internet 连接，当它出现问题时，此时默认网关切换为内部连接，此时，本地计算机将无法再访问位于 Internet 连接上的主机。对于这种情况，微软建议使用 route add 来添加对应目的网络的匹配路由，而不是设置多个默认网关，这其实就是最长匹配原则，精确路由优先于缺省路由。

2．对于下行设备是路由器的情况，其不会切换默认路由，只会按照配置好的缺省路由优先级进行流量转发，从而导致路由黑洞。

结合上面两个原因，在网络出口路由器的下行设备上配置缺省路由的方法也不可行。

综上所述，要想消除单点故障，又同时实现下行设备在故障发生时的流量无障碍转发，以上的三个方法均不可行，所以人们开发出了一种全新的协议：VRRP，这种协议无需下行设备与出口路由器进行交互性操作，却完全实现了网络出口的冗余备份，下一节，我们就来详细讨论下 VRRP 的基本原理及实现过程。

2．VRRP 基本原理及实现过程

2.1 VRRP 基本概念

VRRP 路由器：运行 VRRP 协议一个或多个实例的路由器

虚拟路由器：由一个 Master 路由器和多个 Backup 路由器组成。其中，无论 Master 路由器还是 Backup 路由器都是一台 VRRP 路由器，下行设备将虚拟路由器当做默认网关。

VRID：虚拟路由器标识，在同一个 VRRP 组内的路由器必须有相同的 VRID，其实 VRID 就相当于一个公司的名称，每个员工介绍自己时都要包含公司名称，表明自己是公司的一员，同样的道理，VRID 表明了这个路由器属于这个 VRRP 组。

Master 路由器：虚拟路由器中承担流量转发任务的路由器

Backup 路由器：当一个虚拟路由器中的 Master 路由器出现故障时，能够代替 Master 路由器工作的路由器

虚拟 IP 地址：虚拟路由器的 IP 地址，一个虚拟路由器可以拥有一个或多个虚拟 IP 地址。

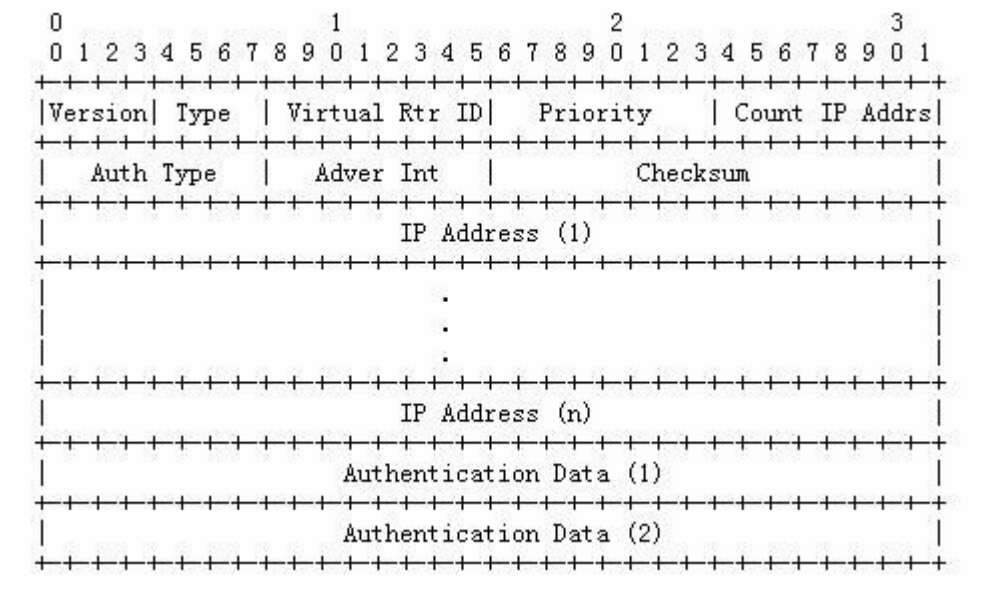
IP 地址拥有者：接口 IP 和虚拟路由器 IP 地址相同的路由器就叫做 IP 地址拥有者。

主 IP 地址：从物理接口设置的 IP 地址中选择，一个选择规则是总是选用第一个 IP 地址，VRRP 通告报文总是用主 IP 地址作为该报文 IP 包头的源 IP。

虚拟 MAC 地址：组成方式是 00-00-5E-00-01-{VRID}，前三个字节 00-00-5E 是 IANA 组织分配的，接下来的两个字节 00-01 是为 VRRP 协议指定的，最后的 VRID 是虚拟路由器标识，取值范围[1，255]

2.2 VRRP 报文组成

下面我们来看 VRRP 报文的具体组成：



图三（VRRP 报文格式，取自 RFC3768）

具体字段含义：

Version：VRRP 协议版本号，RFC3768 定义了版本 2.

Type：该字段指明了 VRRP 报文的类型，RFC3768 只定义了一种 VRRP 报文，那就是 VRRP 通告报文，所以该字段总是置为 1，若收到的 VRRP 通告报文拥有非 1 的类型值，那么会被丢弃。

Virtual Rtr ID：也就是我们上面介绍过的 VRID，一个 VRID 唯一地标识了一个虚拟路由器，取值范围是[1,255]，所以一台路由器的接口可以同时运行最多 255 个 VRRP 实例，此字段没有缺省值，必须人为设定。

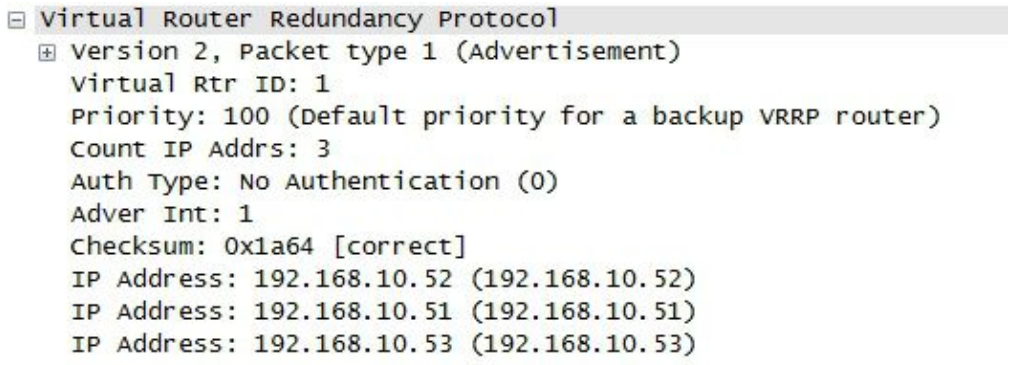
Priority：优先级，在一个虚拟路由器中用来选取 Master 路由器和 Backup 路由器，值越大表明优先级越高，此字段共有 8 个 bit，取值范围[1,254]，若没有人为指定，缺省值是 100。其中，VRRP 协议会将 IP 地址拥有者路由器的该字段永远设置为 255，若人为指定为其它值，也不会影响 VRRP 协议的默认行为，即 IP 地址拥有者路由器的该字段总是 255。另外，此字段设置为 0 会出现在下面这种情形中，当 Master 路由器出现故障后，它会立刻发送一个 Priority 置 0 的 VRRP 通告报文，当 Backup 路由器收到此通告报文后，会等待 Skew time 时间，然后将自己切换为 Master 路由器，其中 Skew time=(256-Backup 路由器的优先级)/256，单位为秒，例如若 Backup 路由器的优先级为 100，那么 Skew time=156/256=0.609 秒，对于主路由器来说，Skew time 并没有实际意义，虽然 cisco 的路由器也会计算并显示出来。

Count IP Addrs：VRRP 通告报文中包含的 IP 地址数量，这个字段其实就是为一个 VRRP 虚拟路由器所分配的 IP 地址的数量，我们来看一个 cisco 的实际例子：配置如下：

```
interface Ethernet1/0
```

```
 ip address 192.168.10.102 255.255.255.0
 duplex half
 vrrp 1 ip 192.168.10.52
 vrrp 1 ip 192.168.10.51 secondary
 vrrp 1 ip 192.168.10.53 secondary
end
```

我们来看一下上面的配置在封装成 VRRP 通告报文的时候，是如何进行的，如下图：



图四（VRRP 报文的抓包分析）

大家可以看到，VRRP 通告报文中的 Count IP Addrs 字段的值为 3，这是因为我们配置了 3 个虚拟 IP 地址，另外，下面的 IP Address 字段也按照我们配置虚拟 IP 的顺序进行了封装。

Auth Type：认证类型字段，是一个 8 位的无符号整数，一个虚拟路由器只能使用一种认证类型，如果 Backup 路由器收到的通告报文中认证类型字段是未知的或和本地配置的不匹配，那么它将丢弃该数据包。

值得注意的是，在 RFC2338 中为 VRRP 定义了 3 种认证类型：无认证、明文认证、MD5 认证，但是在后续的实践中发现，这些手段无法提供行之有效的安全性，并且还会导致多个 Master 路由器的问题，所以在最新的 VRRP 标准：RFC3768 中已经去掉了所有的认证类型。

目前认证类型字段的定义如下：

0 - 无认证，此时下面的 Authentication Data 字段将会被置为全 0，接收到的路由器也会忽略此字段。

1 - 保留，是为了向前一个版本的 RFC2338 提供兼容性

2 - 保留，是为了向前一个版本的 RFC2338 提供兼容性

Adver Int:：此字段规定了 Mater 路由器向外发送 VRRP 通告报文的时间间隔，以秒为单位，取值范围是[1,255]，若没有人工配置，缺省为 1 秒。

Checksum：整个 VRRP 报文的校验和，计算过程中，将 Checksum 字段置为 0，计算完成后将结果填入此字段。若希望进一步了解 Checksum 的计算，可以查看 RFC1071(CKSM)。

IP Address：此字段存放 3 个 VRRP 虚拟路由器的虚拟 IP 地址，配置了几个就封装几个，在上面的 cisco 实例中我们配置了三个，那么 VRRP 通告报文就会封装 3 个。

Authentication Data：RFC3768 中规定，此字段只是为了向 RFC2338 兼容，在实际的封装时，全置为 0，接收方也会忽略此字段。

2.3 VRRP 协议状态机

对一个 VRRP 虚拟路由器来讲，参与它的每一台 VRRP 路由器，都只有 3 种 VRRP 状态：Initialize，Master，Backup，在讲述这三种状态时会碰到一些新的概念，我们会在第一次遇到时做详细解释。

2.3.1 初始状态（Initialize）

这是配置好 VRRP 后，VRRP 等待一个开始事件时的状态，当本地 VRRP 进程切换到此状态后，会依次执行下列操作：

2.3.1.1 如果本地优先级为 255，也就是说自己是 IP 拥有者路由器，那么接下来它会：

1.发送 VRRP 通告报文

2.广播免费 ARP 请求报文，内部封装是虚拟 MAC 和虚拟 IP 的对应，有几个虚拟 IP 地址，那么就发送几个免费 ARP 请求报文。

3.启动一个 Adver_Timer 计时器，初始值为 Advertisement_Interval（缺省是 1 秒），当该计时器超时后，会发送下一个 VRRP 通告报文

4.本地 VRRP 进程将自己切换为 Master 路由器

2.3.1.2 如果，本地优先级不是 255，那么接下来它会：

1.设置 Master_Down_Timer 计时器等于 Master_Down_Interval，也就是主路由器死亡时间间隔，如果此计时器超时，那么 Backup 路由器就会宣布主路由器死亡。其中 Master_Down_Interval =（3*Advertisement_Interval）+ Skew_time 举例来说，一个 VRRP 实例（也就是一个 VRRP 虚拟器）的优先级是 100，报文发送间隔是 1 秒，那么 Master_Down_Interval = 3*1s + （256-100）/256s = 3.609 秒。

2．本地 VRRP 进程将自己切换为 Backup 路由器

2.3.2 备份路由器状态（Backup）

2.3.2.1

备份路由器是为了监控 Master 路由器的状态，如果一个 VRRP 路由器处于此状态，那么它会：

1. 不响应对虚拟 IP 地址的 ARP 请求报文
2. 丢弃帧头目的 MAC 地址是虚拟 MAC 的帧
3. 丢弃 IP 头中目的 IP 地址是虚拟 IP 的 IP 包

2.3.2.2

如果此时该 VRRP 路由器收到了一个 shutdown 事件，那么它会：
1. 取消 Master_Down_Timer
2. 转换为初始状态（Initialize state）

2.3.2.3

如果 Master_Down_Timer 超时，那么该 VRRP 路由器会执行：
1. 发送一个 VRRP 通告报文，
2. 广播免费 ARP 请求报文，内部封装是虚拟 MAC 和虚拟 IP 的对应，有几个虚拟 IP 地址，那么就发送几个免费 ARP 请求报文。
3. 设置 Adver_Timer 计时器为 Advertisement_Interval（缺省为 1 秒）
4. 切换到 Master 状态

2.3.2.4

如果该 Backup 状态的 VRRP 路由器收到了一个 VRRP 通告报文; 当该 VRRP 通告报文的优先级字段为 0 时，那么路由器会将当前的 Master_Down_Timer 设置为 Skew_Time；如果优先级不为 0，并且大于或等于本地优先级，那么本地路由器会重置 Master_Down_Timer 计时器并保持 Backup 状态；如果优先级不为 0，并且小于本地优先级，如果开启了抢占模式（Preempt mode），那么该 Backup 路由器等待指定的抢占延迟时间后将自己切换为 Master 路由器；并执行 Master 路由器的所有动作；例如：vrrp 1 preempt delay minimum 10，表示等待 10 秒后切换自己为 Master；如果优先级不为 0，并且小于本地优先级，如果没有开启抢占模式（Preempt mode），那么本地路由器保持 Backup 状态。

2.3.3 Master 路由器（Master state）

处于 Master 状态的路由器会执行目的 MAC 为虚拟 MAC 数据帧的转发，这里要清楚的是对于下行设备的 arp 表里，该虚拟 MAC 是和虚拟 IP 地址相对应的。

2.3.3.1

当路由器处于 Master 状态时，会进行下面的动作：
1. 响应对虚拟 IP 地址的 ARP 请求
2. 转发目的 MAC 地址是虚拟 MAC 的数据帧
3. 拒绝目的 IP 地址是虚拟 IP 的数据包，除非它是 IP 地址拥有者（也就是优先级是 255 的那个路由器）。

2.3.3.2

如果处于 Master 状态的 VRRP 进程收到了一个 shutdown 事件，那么它会：
1. 取消 Adver_Timer 计时器
2. 发送一个优先级字段置零的 VRRP 通告报文
3. 切换为初始状态（Intialize state）

2.3.3.3

如果 Adver_Timer 计时器超时，那么：

1. 发送一个 VRRP 通告报文
2. 重置 Adver_Timer 计时器

2.3.3.4

如果收到了一个 VRRP 报文且其优先级为 0，那么：

1. 发送一个 VRRP 通告报文
2. 重置 Adver_Timer 计时器

2.3.3.5

如果收到了一个 VRRP 报文且其优先级高于本地优先级，或者收到的 VRRP 报文优先级等于本地优先级但是主 IP 地址高于本地的主 IP 地址，那么：

1. 取消 Adver_Timer 计时器
2. 设置 Master_Down_Timer 计时器为 Master_Down_Interval
3. 切换为 Backup 状态

2.4 VRRP 通告报文的发送与接收处理流程

2.4.1 当收到一个 VRRP 通告报文时，执行以下操作：

1. 检查 IP 包头的 TTL 是否为 255
2. 检查 VRRP 报文的 version 字段是否为 2
3. 检查 VRRP 报文的完整性，即是否包含 RFC 所定义的各字段
4. 检查 checksum 字段
5. 检查 VRID 字段是否和本地配置的 VRID 一致
6. 确保本地路由器不是 IP 地址拥有者，即优先级不是 255（如果自己是 IP 地址拥有者，那么自己永远都是 Master 路由器，所以丢弃任何收到的 VRRP 通告报文）
7. 检查认证类型和认证数据字段，确保和本地一致

如果上面的七项检查有一项不通过，那么就会丢弃该报文，如果配置了网管程序，就会自动记录 log 并报告错误。

如果以上检查通过，那么可能会检查 VRRP 通告报文中的 Count IP Addrs 字段和 IP 地址字段，确保和本地一致，如果检查结果不一致并且该报文不是由 IP 地址拥有者产生的，那么丢弃该报文。**注意：**这里用了可能，所以各厂家在实现 VRRP 时，可能会检查也可能不会检查该字段，Cisco 就不会检查。

如果以上检查均通过，那么接下来会查看 Advertisement_Interval 是否和本地一致，若不一致，会丢弃该报文。如果配置了网管程序，就会自动记录 log 并报告错误。

2.4.2 发送一个 VRRP 报文的时候，需要执行下列动作：

1. 将当前手动或默认的 VRRP 配置封装进报文的相关字段
2. 计算 VRRP 校验和，将结果放入 checksum 字段
3. 设置帧的源 MAC 地址为虚拟 MAC 地址（这样就确保了下联交换机可以建立正确的 MAC 表，即实际端口和虚拟 MAC 的对应关系，这个对应关系会因为 Master 路由器的切换而发生变化，所以这个处理对于上行数据帧的正确转发至关重要）
4. 设置 IP 包头的源 IP 地址为该物理接口的主 IP 地址

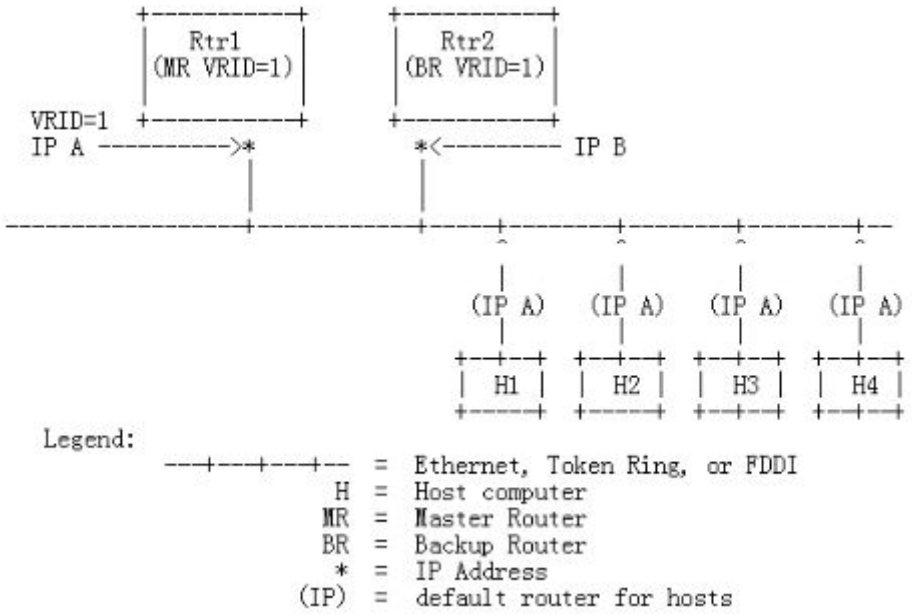
5． 设置 IP 头中的 protocol 字段为 0x70，也就是十进制的 112，表示上层封住协议是 VRRP
6． 将封装好的 VRRP 通告报文发送出去，目标 IP 为组播地址 224.0.0.18，目的 MAC 地址为组播 MAC ：01:00:5e:00:00:12

好了，以上就是关于 VRRP 全部的协议实现细节，下面我们来看一下 VRRP 的应用需求。

3. VRRP 的两种应用需求：

冗余备份、负载均衡

3.1 我们来看一个具体冗余备份的图例：



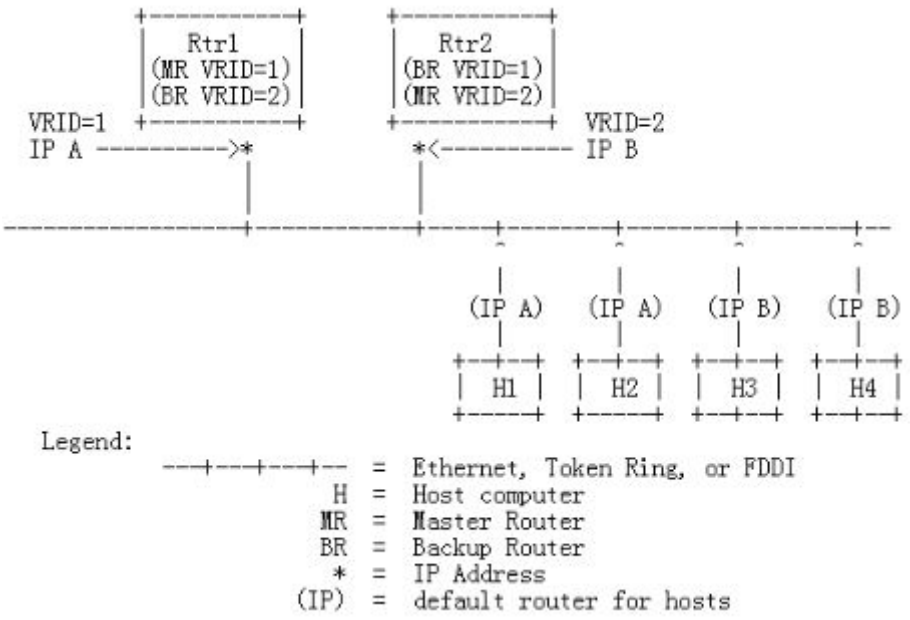
图五（VRRP 为下行设备提供冗余备份功能）

上图的实例中我们定义了一个虚拟路由器，标识为 1（即 VRID=1）；其虚拟 IP 设置为 IP A，是 Router1 的接口 IP，所以 Router1 就是一个 IP 地址拥有者，系统会将其优先级设置为 255，并作为 VRID1 的 Master 路由器，而 Router2 的优先级默认为 100，Router2 的 VRRP 实例 VRID1 将自己设置为 Backup 路由器；下行的各主机将网关设置为 VRID1 虚拟路由器的虚拟 IP，即 IP A。

通过以上设置，即可消除出口路由器的单点故障，假如此时 Router1 挂掉，那么 Router2 的 VRRP 进程经过 Master_Down_Interval = （3*Advertisement_Interval）+ Skew_time，时间之后，就会宣布 Master 挂掉，将自己设置为 Master 路由器，并广播免费 ARP 请求报文，这样下行交换机就能更新虚拟 MAC 到与 Router2 接口的对应表项，从而实现不间断转发用户的流量。

但是，实际情况中我们很少这么设置 VRRP，因为始终是一台路由器在承担所有流量，不符合物尽其用的原则。

3.2 我们来看一个冗余备份和负担均衡相结合的图例：



图六（VRRP 同时实现冗余备份和负载均衡）

上图中的 VRID1 的设置和图五一样；VRID2 的设置将 Router2 作为 Master 路由器，Router1 作为 Backup 路由器，然后将一半下行主机的网关设置为 VRID1 的虚拟 IP，即 IP A，将另一半主机的网关设置为 VRID2 的虚拟 IP，即 IP B。

这样，在两台设备都正常运行的情况下，终端流量一半走 Router1，一半走 Router2，实现了负载均衡；而当其中一台路由器挂掉时，依靠 VRRP 的功能，会将另一台路由器设置为 Master 路由器，继续流量转发，从而实现了冗余备份功能，此时，这台路由器会同时作为 VRID1 和 VRID2 的 Master 路由器。

4. 实例研究

实际组网中，绝大多数情况都是双交换机或双路由器上行，所以图六的 VRRP 实现具有普遍意义，下面，我们就来看看 VRRP 在各厂家设备上的配置实现（都以图六作为试验拓扑）：

4.1 VRRP 在 Cisco 路由器上的实现

下面两图是 Router1 和 Router2 的 VRRP 配置及状态：

```
Router1#show runn int e1/0
Building configuration...

Current configuration : 220 bytes
!
interface Ethernet1/0
 ip address 192.168.10.101 255.255.255.0
 duplex full
 vrrp 1 ip 192.168.10.101
 vrrp 1 preempt delay minimum 5
 vrrp 1 priority 120
 vrrp 2 ip 192.168.10.102
 vrrp 2 preempt delay minimum 5
end

Router1#show vrrp
Ethernet1/0 - Group 1
  State is Master
  Virtual IP address is 192.168.10.101
  Virtual MAC address is 0000.5e00.0101
  Advertisement interval is 1.000 sec
  Preemption enabled, delay min 5 secs
  Priority is 255 (cfgd 120)
  Master Router is 192.168.10.101 (local), priority is 255
  Master Advertisement interval is 1.000 sec
  Master Down interval is 3.003 sec

Ethernet1/0 - Group 2
  State is Backup
  Virtual IP address is 192.168.10.102
  Virtual MAC address is 0000.5e00.0102
  Advertisement interval is 1.000 sec
  Preemption enabled, delay min 5 secs
  Priority is 100
  Master Router is 192.168.10.102, priority is 255
  Master Advertisement interval is 1.000 sec
  Master Down interval is 3.609 sec (expires in 3.137 sec)
```

```
Router2#show runn in e1/0
Building configuration...

Current configuration : 220 bytes
!
interface Ethernet1/0
 ip address 192.168.10.102 255.255.255.0
 duplex full
 vrrp 1 ip 192.168.10.101
 vrrp 1 preempt delay minimum 5
 vrrp 2 ip 192.168.10.102
 vrrp 2 preempt delay minimum 5
 vrrp 2 priority 120
end

Router2#show vrrp
Ethernet1/0 - Group 1
  State is Backup
  Virtual IP address is 192.168.10.101
  Virtual MAC address is 0000.5e00.0101
  Advertisement interval is 1.000 sec
  Preemption enabled, delay min 5 secs
  Priority is 100
  Master Router is 192.168.10.101, priority is 255
  Master Advertisement interval is 1.000 sec
  Master Down interval is 3.609 sec (expires in 1.505 sec)

Ethernet1/0 - Group 2
  State is Master
  Virtual IP address is 192.168.10.102
  Virtual MAC address is 0000.5e00.0102
  Advertisement interval is 1.000 sec
  Preemption enabled, delay min 5 secs
  Priority is 255 (cfgd 120)
  Master Router is 192.168.10.102 (local), priority is 255
  Master Advertisement interval is 1.000 sec
  Master Down interval is 3.003 sec
```

通过 VRRP 的运行状态，我们可以知道：

1. 即使为 IP 地址拥有者配置了优先级，系统也会使用 255
2. 若不指定优先级，系统缺省认为是 100.
3. 缺省通告时间是 1 秒

4.2 VRRP 在 Redback 路由器上的实现

下面就是两个 VRRP 实例在 Redback 路由器上的配置：

context vrrp

interface downlink

ip address 192.168.10.201/24

vrrp 1 owner----------这表明这个 VRRP 实例的 IP 是路由器的接口 IP

virtual-address 192.168.10.201—必须是真实 interface 的 IP，否则系统报错

advertise-interval millisecond 100----VRRP 通告报文的发送间隔，这里为 100 毫秒

authentication redback-md5 vrrp-auth—采用 MD5 方式认证

vrrp 2 backup----------这表明这个 VRRP 实例的 IP 地址一定不是 interface 接口的 IP，但是和 interface 接口 IP 一定在同一个网段内。

virtual-address 192.168.10.202

advertise-interval millisecond 100-----VRRP 通告报文的发送间隔，这里为 100 毫秒
priority 120----------本地 VRRP 优先级为 120
init-wait 1----------VRRP 实例 2 启动后，等待接收 VRRP 通告报文的时间
track interface uplink vrrp decrement 100---VRRP 实例监控上行链路的状态，若为 down，则将本地 VRRP 优先级降低 100
authentication simple vrrp-authentication----VRRP 实例 2 采用明文认证

!
interface uplink----------上行端口，连接骨干网
ip address 10.1.1.1/30
!
key-chain vrrp-auth key-id 1----------VRRP1 采用 MD5 方式认证
key-string encrypted DC2C32F335F4113A949FEF2997659B974A70687575316125
!

下面我们来看看两个 VRRP 实例的状态：

```
[vrrp]sx-szmfc-s1200-bas1#show vrrp

  --- VRRP Virtual Router downlink/1 (Owner) ---

State              : Master            Last Event          : Interface Up
Priority           : 255               Fast Adv Int (ms)   : 100
Last Adv Source    : 0.0.0.0           Up Time             : 00:31:03
Auth Type:         : Redback-MD5       Key Chain           : vrrp-auth
Auth Sequence      : 0
Address List:
192.168.10.201


  --- VRRP Virtual Router downlink/2 (Backup) ---

State              : Master            Last Event          : Master Timeout
Priority           : 120               Fast Adv Int (ms)   : 100
Last Adv Source    : 0.0.0.0           Up Time             : 00:31:02
Preempt            : Yes               Master Down  (ms)   : 300
Preempt HT (sec)   : 0                 Skew Time (u-sec)   : 0
Auth Type:         : Simple            Key Chain           : vrrp-authentication
Connected Route    : No                Init Wait (sec)     : 1
Address List:
192.168.10.202
Track Interface List:
Interface   Context     Priority
uplink      vrrp        100
```

Redback 路由器上 VRRP 实现特性：

1. 将 VRRP 的 owner 角色与 backup 角色从配置层次上区分开，这样做的好处就是非常清晰。而其他厂家没有关键字来区分这两种角色，只是依靠 virtual-address 是否和本地接口 IP 地址相同来判断本地 VRRP 实例是 owner 角色还是 backup 角色。
2. 若配置 VRRP 实例为 owner 角色，那么 virtual-address 必须是 interface 的 IP 地址，否则系统报错，并且此时不能配置 priority，因为系统为 owner 角色永远分配 255 的优先级；反之若配置 VRRP 实例为 backup 角色，那么 virtual-address 绝对不能是 interface 的 IP 地址，否则系统报错，并且此时不能配置 priority 为 255，因为 255 是系统为 owner 保留的优先级。
3. 只有 backup 状态的 VRRP 实例才能够配置 preempt，因为 master 无需抢占。
4. **Track：这是个非常有用的特性，它是用来快速检测上行链路的连通性并作出反应的命令。我们来假设这样一种情况，此时本地路由器作为 VRRP 实例 2 的 master 角色，假如路由器的上行接口 uplink 挂掉了，那么本地的所有上行流量都将被丢弃，对下行设**

备来讲形成了路由黑洞，这是不可接受的。所以，依靠这个 track 特性来监控指定链路，当检测到指定链路 down 掉后，立刻将本地 VRRP 实例的优先级降低 100.，从而让其他路由器抢占 master 角色。

我们把 uplink 认为 down 掉，来看看路由器的反应：

```
[vrrp]sx-szmfc-s1200-bas1#show vrrp

  --- VRRP Virtual Router downlink/1 (Owner) ---

State              : Master            Last Event          : Interface Up
Priority           : 255               Fast Adv Int (ms)   : 100
Last Adv Source    : 0.0.0.0           Up Time             : 00:18:03
Auth Type:         : Redback-MD5       Key Chain           : vrrp-auth
Auth Sequence      : 0
Address List:
192.168.10.201


  --- VRRP Virtual Router downlink/2 (Backup) ---

State              : Master            Last Event          : Master Timeout
Priority           : 20                Fast Adv Int (ms)   : 100
Last Adv Source    : 0.0.0.0           Up Time             : 00:18:02
Preempt            : Yes               Master Down  (ms)   : 300
Preempt HT (sec)   : 0                 Skew Time (u-sec)   : 0
Auth Type:         : Simple            Key Chain           : vrrp-authentication
Connected Route    : No                Init Wait (sec)     : 1
Address List:
192.168.10.202
Track Interface List:
Interface   Context     Priority
uplink      vrrp        100
```

大家应该可以看到，当 VRRP 实例 2，检测到 uplink 不通之后，将 VRRP2 的优先级降低 100,120-100=20，所以我们在上图中看到 VRRP2 的优先级为 20。

5． 对 show vrrp 的显示结果做一下简单说明，Last Adv Source 是指 VRRP 通告报文的发送者的 IP，所以此处总是置为 0，并且 VRRP 报文中也不包含此字段，只不过是 Redback 方便网管的一个处理方式。

6． Last Event：是指 VRRP 导致发生状态转化的事件，可能是配置了虚拟地址、端口 UP，主路由器超时，被别人抢占了 Master 角色等等，可以作为网管的一个辅助判断手段。

4.3 VRRP 在 Juniper 路由器上的实现

Juniper 路由器上 VRRP 的实现和 cisco 差不多，下面来看看具体配置：

```
mm# show interfaces ge-0/0/0    ----------在当前接口的 unit10 子接口上起了两个 VRRP 实例
unit 10 {
    family inet {
        address 192.168.10.201/24 {----------------------物理接口 IP 地址
            vrrp-group 1 {----------------------------------VRID = 1
                virtual-address 192.168.10.201;--------虚拟 IP 地址为物理接口 IP
                priority 120;-----------------------------虽然配置了 120，但是系统会改为 255
                fast-interval 100;-------------------------快速发布 VRRP 通告报文
                preempt {
                    hold-time 0;---------- ----实时抢占，因为这里是 owner，所以无意义
                }
```

```
                accept-data;----------------接受目的 IP 是虚拟 IP 的报文
                track {-----------------------监测上行端口
                    interface em1.0 {
                        bandwidth-threshold 100m priority-cost 100;--监测上行端口带宽
                    }
                    interface em2.0 {----------------监测上行端口 UP、down 状态
                        priority-cost 100;
                    }
                }当监测到 em1 的带宽不满 100M 或 em2  挂掉后，将本地优先级降低 100
            }
            vrrp-group 2 {
                virtual-address 192.168.10.202;
                priority 100;
                fast-interval 100;
                preempt {
                    hold-time 0;--------------------当发现 Master 挂掉后，执行实时抢占，而
                                                    不是等待 Master-down-interval
                }
            }
        }
    }
}
```

看一下 VRRP 运行状态：

```
ericsson@NMHH-OAM-ROUTER02-ERICJ2350> show vrrp
Interface     State       Group   VR state    Timer    Type   Address
ge-0/0/0.10   up              1   master      A  4.798 lcl    192.168.10.201
                                                    vip    192.168.10.201
              up              2   master      A  4.296 lcl    192.168.10.201
                                                    vip    192.168.10.202
```

5. VRRP 的安全性

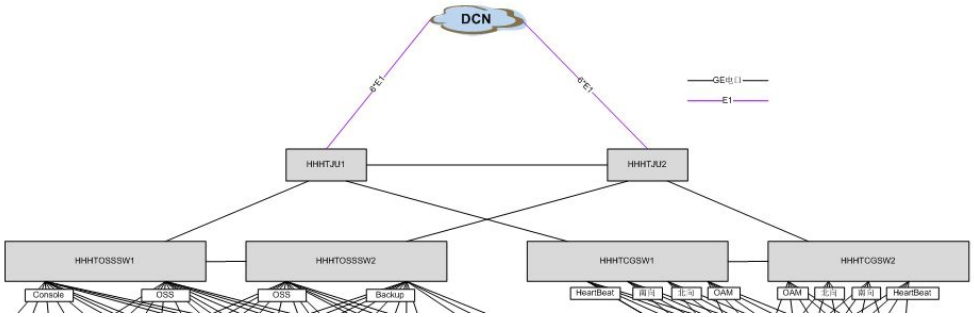
当前各厂家对 VRRP 的实现主要是依据 RFC2338，采用 sample、MD5、无认证，这三种方式，但最新的 RFC3768 已经取消了所有安全选项，这主要是因为在具体实施中遇到的问题，说白了就是上面的三种方式都无法提供实际的安全性，下面我们来具体分析一下：

这和 VRRP 的实现原理有关，虽然 Master 的 VRRP 通告报文经过 MD5 方式加密，但是攻击者可以截获、复制并发送同样的加密报文，从而导致出现多个 Master，使下行交换机无法正确上传流量；另外一种情况，即使攻击者不复制发送 VRRP 通告报文，它也可以通过复制发送免费 ARP 报文，或者响应下行设备的 ARP 请求报文，从而误导下行交换机的流量选择，同样造成了路由黑洞。

基于以上情况，RFC4768 取消了这些在其位不谋其政的安全选项，只是为了向 RFC2338 兼容，仍然保留了这些安全字段。

6. 一个典型的 VRRP 故障分析

前期有同事从事 IMS 项目，其中需要在 Juniper 路由器上设置 VRRP，基本拓扑如下：



图（IMS 项目当中的 VRRP 部署拓扑）

两台 Juniper 路由器各连接两台交换机，然后两台路由器之间做 VRRP 备份组。遇到的问题是，所有配置完成以后，两台路由器的 VRRP 实例均显示自己是 Master，如下图：

```
ericsson@NMHH-OAM-ROUTER01-ERICJ2350> show vrrp
Interface     State       Group   VR state    Timer    Type   Address
ge-0/0/0.20   up              1   master      A  10.288lcl    10.218.168.66
                                                      vip    10.218.168.65
ge-0/0/1.110  up              2   master      A  50.848lcl    10.218.168.34
                                                      vip    10.218.168.33
ge-0/0/1.140  up              3   master      A  22.177lcl    10.218.168.130
                                                      vip    10.218.168.129
ge-0/0/1.150  up              4   master      A  83.347lcl    10.218.168.146
                                                      vip    10.218.168.145
ge-0/0/1.160  up              5   master      A  71.058lcl    10.218.168.162
                                                      vip    10.218.168.161
ge-0/0/1.170  up              6   master      A  26.634lcl    10.218.168.178
                                                      vip    10.218.168.177
```

图（Router-1 的 VRRP 状态）

```
ericsson@NMHH-OAM-ROUTER02-ERICJ2350> show vrrp
Interface     State       Group   VR state    Timer    Type   Address
ge-0/0/0.20   up              1   master      A  4.798 lcl    10.218.168.67
                                                      vip    10.218.168.65
ge-0/0/1.110  up              2   master      A  4.296 lcl    10.218.168.35
                                                      vip    10.218.168.33
ge-0/0/1.140  up              3   master      A  56.145lcl    10.218.168.131
                                                      vip    10.218.168.129
ge-0/0/1.150  up              4   master      A  39.664lcl    10.218.168.147
                                                      vip    10.218.168.145
ge-0/0/1.160  up              5   master      A  65.898lcl    10.218.168.163
                                                      vip    10.218.168.161
ge-0/0/1.170  up              6   master      A  11.063lcl    10.218.168.179
                                                      vip    10.218.168.177
```

图（Router-2 的 VRRP 状态）

通过上面两个图可以看到，对于 VRID 分别为 1、2、3、4、5、6 这六个 VRRP 备份组来说，两台路由器均认为自己是 Master，而正常情况是：对于每一个 VRRP 备份组，都会有一台路由器是 Master 路由器，其它路由器全部是 Backup 路由器。

通过比对 IMS 项目的 CND 设计，排除了配置错误的可能性，来看一下 VRRP 的统计信息：

```
ericsson@NMHH-OAM-ROUTER01-ERICJ2350> show vrrp extensive |no-more
Interface: ge-0/0/0.20, Interface index :68, Groups: 1, Active :1
  Interface VRRP PDU statistics
    Advertisement sent                          :197
    Advertisement received                      :0
    Packets received                            :0
    No group match received                     :0
  Interface VRRP PDU error statistics
    Invalid IPAH next type received             :0
    Invalid VRRP TTL value received             :0
    Invalid VRRP version received               :0
    Invalid VRRP PDU type received              :0
    Invalid VRRP authentication type received:0
    Invalid VRRP IP count received              :0
    Invalid VRRP checksum received              :0
```

图（Router-1 的 VRRP 统计信息）

```
ericsson@NMHH-OAM-ROUTER02-ERICJ2350> show vrrp extensive |no-more
Interface: ge-0/0/0.20, Interface index :69, Groups: 1, Active :1
  Interface VRRP PDU statistics
    Advertisement sent                          :10461
    Advertisement received                      :0
    Packets received                            :0
    No group match received                     :0
  Interface VRRP PDU error statistics
    Invalid IPAH next type received             :0
    Invalid VRRP TTL value received             :0
    Invalid VRRP version received               :0
    Invalid VRRP PDU type received              :0
    Invalid VRRP authentication type received:0
    Invalid VRRP IP count received              :0
    Invalid VRRP checksum received              :0
```

图（Router-2 的 VRRP 统计信息）

很明显可以看到各个 VRRP 备份组只有 VRRP 通告报文的发送，而完全没有接受，这是不正常的，而上面的故障现象就是没有收到 VRRP 通告报文所导致的。

两台 Juniper 路由器 J2350 的每一个 VRRP 进程都没有收到任何一个 Advertisement 数据包，所以均不知网络中其它 VRRP 路由器的优先级，所以始终认为自己是 master。

注：当一个 VRRP 进程启动以后，会自动将自己切换为 backup 状态，如果经过 3 个 Advertisement interval，这里是 3*100s=300s，没有收到更高优先级的 Advertisement 数据包，则将自己切换为 master 状态。

将关注投向了具体拓扑，这个拓扑和常规实现有所差异，两台路由器不是连接到同一台交换机上，而是分别连接到两台交换机，靠交换机之间的链路来交换 VRRP 通告报文，以及上行流量的冗余备份，所以断定 VRRP 通告报 c 文被阻断了。

综上所述：

1. 查看各端口及链路是否正常
2. 需要查看下行的互联交换机之间是否能够正常通信（即 J2350 与下行交换机的接口是否与两台交换机互联接口处于同一个 VLAN 之中）。
3. 查看是否是下行交换机的安全策略阻止了 VRRP 通告报文的转发

经过实际判断，最终确认是下行交换机的安全策略阻止了 VRRP 通告报文的转发，添加如下策略后，两台路由器的 VRRP 状态恢复正常：set security zone security-zone trust interfaces ge-0/0/0.10 host-inbound-traffic protocols vrrp

至此，故障解决。

7.后记

一年多没好好写篇文档了，其实写文档是一个对自己学习很好的总结手段，希望以后能够坚持下去，另外，本文基于个人对 RFC3768 的理解，可能存在疏漏，甚至错误，请大家不吝赐教。