NM05 OSPF 结合 LVS 实现跨网段多活负载均衡集群

使用 LVS 做负载均衡方案时，都会讨论 LVS 的单点问题，一般使用 keepavlied 维持两个节点组成的 LVS 冷备网关组，当一个 LVS 网关失效时，另外一个 LVS 网关会迅速接管连接状态与负载信息，并漂移 VIP，以此来实现 LVS 网关的高可用。

虽然实现了高可用，单毕竟只有一个节点作为网关，如果选择 NAT 模式这样比较消耗资源的方式，可能对 LVS 机器的负载能力构成挑战，怎么办呢？

Keepalived 实现高可用的方式，是使用了 VRRP 协议进行选举与 VIP 的宣告，VRRP 是啥呢，华三有篇技术文章，是我看过的关于 VRRP 最通俗易懂的文章了，链接如下：
VRRP 技术白皮书
http://www.h3c.com/cn/d_200802/335873_30003_0.htm

VRRP 协议本是路由器和网络设备上用来保证高可用时使用的，因为 LINUX 太强大了，模拟路由器的网络操作很方便，所以有了 LINUX 上的 keepalived，将 linux 服务器模拟成一个网络设备的样子，对外发 VRRP 通告。

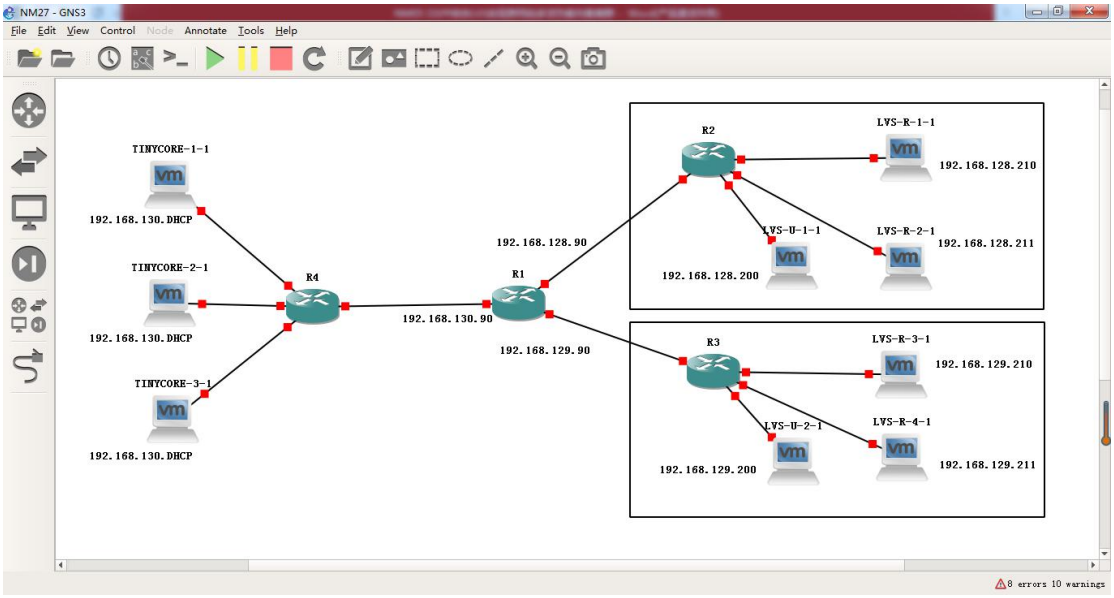
VRRP 协议本身是不能跨 VLAN 的，所以 LVS 集群也没法跨 VLAN。

既然本身 keepalived 就是学习的网络设备的高可用技术，那有没有其他更高级的负载均衡能力呢？有，OSPF 与 ECMP。

简单来说，就是将 VRRP 这样的在一个 VLAN 中进行的虚 IP 宣告仪式，转而使用 OSPF 在更高层次和更高级的设备上进行宣告，在更大范围上实现一个虚 IP 对应多个子网，进而提供更强的服务能力的事情。

外网上关于 OSPF+LVS 的文章很多，我这里不进行直接转述，只将自己的实验过程，以及最终的一些收获和坑总结下。

最终拓扑图如下：



本次实验共涉及到 9+1 台虚拟机（VMWare 管理），三台交换机，一台路由器，组成一个私有的虚拟网络。

网段划分：

192.168.128.0 第一个 LVS 集群

192.168.129.0 第二个 LVS 集群

192.168.130.0 客户端浏览器集群

路由器的三个口分别设置 IP 地址为

192.168.128.90 接 R2 交换机 连接第一个 LVS 集群

192.168.129.90 接 R3 交换机 连接第二个 LVS 集群

192.168.130.90 接 R4 交换机 连接客户端浏览器集群

两个 LVS 集群共同宣告的虚拟地址为 1.1.1.1

主要配置：

交换机 R2，R3，R4（因为是用路由器模拟的交换机，没改默认的名字和图标）配置，见另外一篇文章：

NM04 GNS3 网络拓扑结构中模拟交换机节点

路由器 R1 配置

```
#关闭 CEF
no ip cef

#接口配置
interface FastEthernet0/0
 ip address 192.168.128.90 255.255.255.0
 ip ospf hello-interval 1
 duplex auto
```

```
 speed auto

#接口配置
interface FastEthernet0/1
 ip address 192.168.129.90 255.255.255.0
 ip ospf hello-interval 1
 duplex auto
 speed auto

#接口配置
interface FastEthernet1/0
 no switchport
 ip address 192.168.130.90 255.255.255.0

#ospf 配置
router ospf 100
 log-adjacency-changes
 network 192.168.128.0 0.0.0.255 area 0
 network 192.168.129.0 0.0.0.255 area 0
 network 192.168.130.0 0.0.0.255 area 0
```

实验效果：

```
R1#show ip route
Codes: C - connected, S - static, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2
       i - IS-IS, su - IS-IS summary, L1 - IS-IS level-1, L2 - IS-IS level-2
       ia - IS-IS inter area, * - candidate default, U - per-user static route
       o - ODR, P - periodic downloaded static route

Gateway of last resort is not set

     1.0.0.0/32 is subnetted, 1 subnets
O       1.1.1.1 [110/110] via 192.168.129.200, 00:01:18, FastEthernet0/1
                [110/110] via 192.168.128.200, 00:01:18, FastEthernet0/0
C    192.168.128.0/24 is directly connected, FastEthernet0/0
C    192.168.129.0/24 is directly connected, FastEthernet0/1
C    192.168.130.0/24 is directly connected, FastEthernet1/0
R1#
```

对于到 1.1.1.1 地址的路由出口选择，通过一次 ping 测试。来查看路由切换的情况，发现可以在两个出口之间轮询

```
R1#show ip route 1.1.1.1
Routing entry for 1.1.1.1/32
  Known via "ospf 100", distance 110, metric 110, type intra area
  Last update from 192.168.128.200 on FastEthernet0/0, 00:01:43 ago
  Routing Descriptor Blocks:
    192.168.129.200, from 192.168.129.200, 00:01:43 ago, via FastEthernet0/1
      Route metric is 110, traffic share count is 1
  * 192.168.128.200, from 192.168.128.200, 00:01:43 ago, via FastEthernet0/0
      Route metric is 110, traffic share count is 1

R1#ping 1.1.1.1 re
R1#ping 1.1.1.1 repeat 1

Type escape sequence to abort.
Sending 1, 100-byte ICMP Echos to 1.1.1.1, timeout is 2 seconds:
!
Success rate is 100 percent (1/1), round-trip min/avg/max = 8/8/8 ms
R1#show ip route 1.1.1.1
Routing entry for 1.1.1.1/32
  Known via "ospf 100", distance 110, metric 110, type intra area
  Last update from 192.168.128.200 on FastEthernet0/0, 00:01:58 ago
  Routing Descriptor Blocks:
  * 192.168.129.200, from 192.168.129.200, 00:01:58 ago, via FastEthernet0/1
      Route metric is 110, traffic share count is 1
    192.168.128.200, from 192.168.128.200, 00:01:58 ago, via FastEthernet0/0
      Route metric is 110, traffic share count is 1

R1#
```

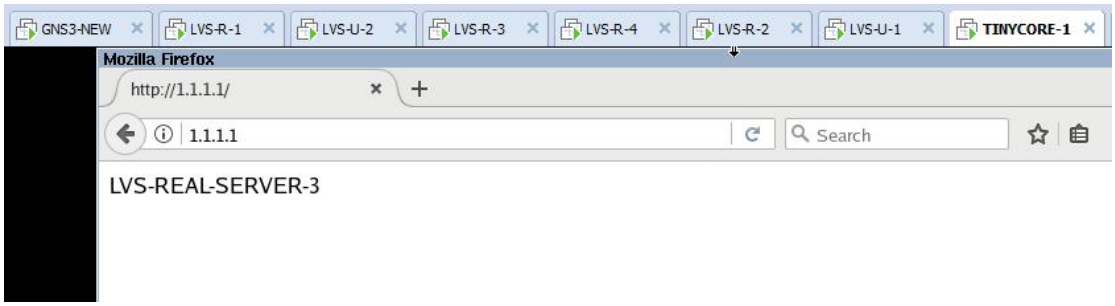
第一次，　星花*在 128 地址的出口上，第二次，星花*在 129 地址的出口上。

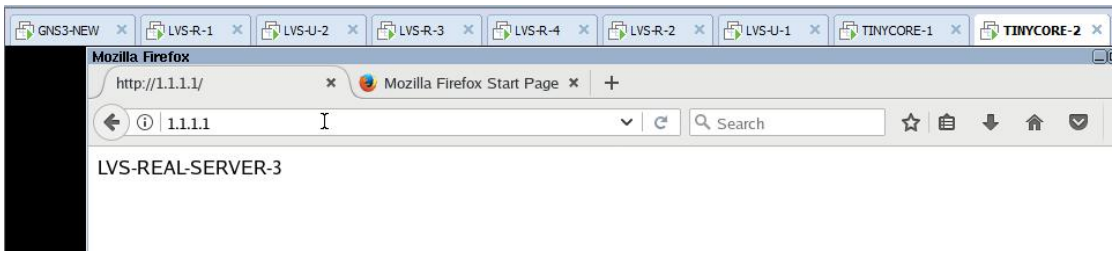
演示一下客户端访问 1.1.1.1 的 web 页面吧。

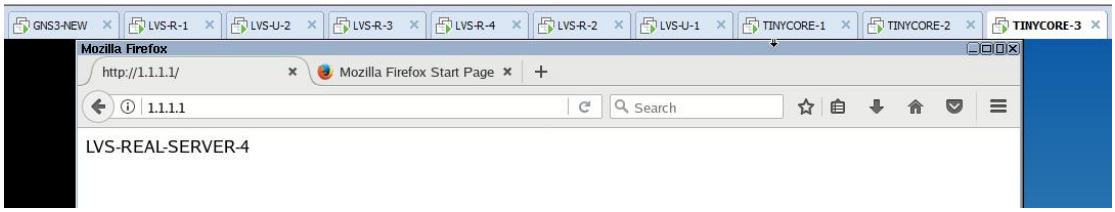
打开浏览器，http://1.1.1.1 演示失败，连接不上，咋回事呢？先重启一次 R1 路由器吧，在其他机器和交换机都已经就绪的情况下。GNS3 这个模拟器，我发现重启 R1 路由器，然后全网就都正常了，重新学习一次路由。

回复正常后的表现

三台客户机的表现：

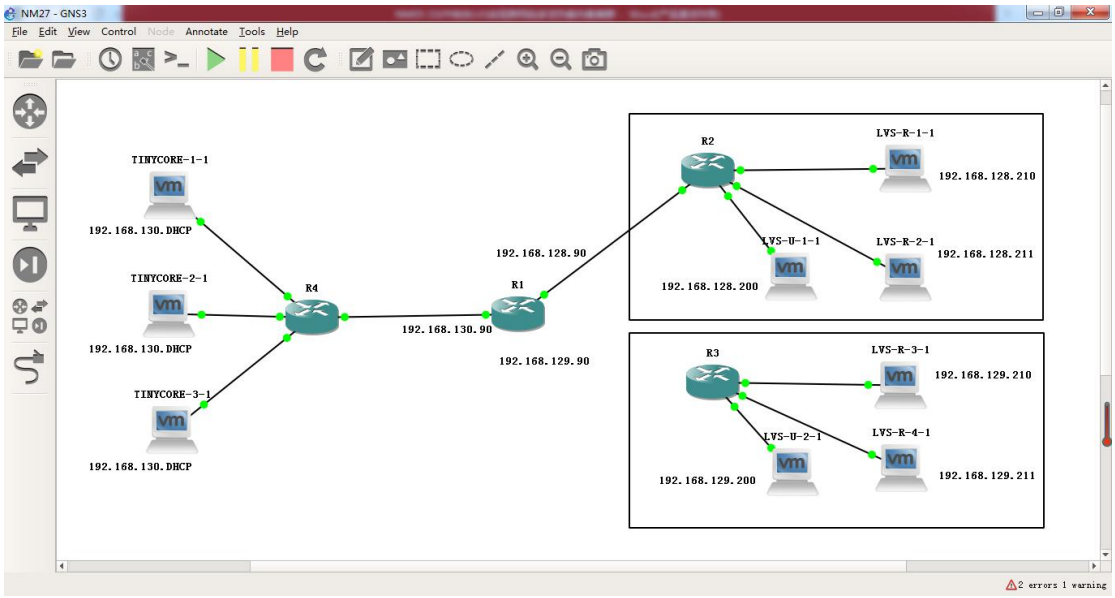






然后我们拔掉 129 子网网段的交换机网线。

拔掉后的拓扑图是这样的。



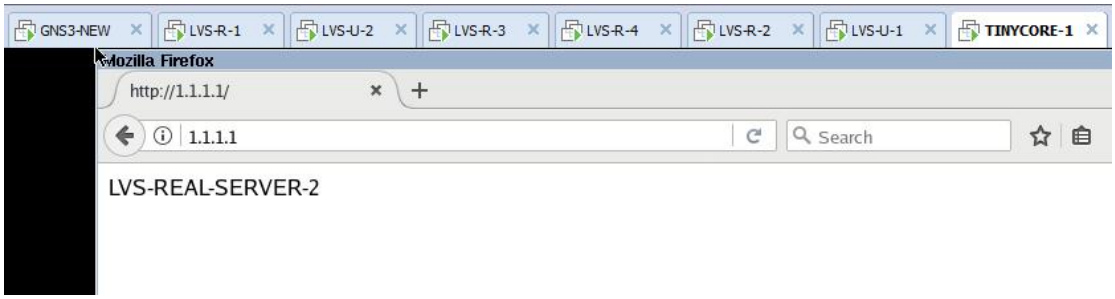
R1 路由器检测到 OSPF 的链路断开

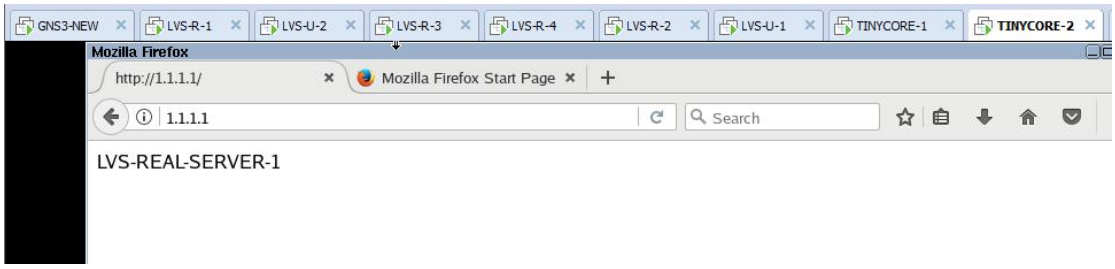
```
*Mar  1 00:04:44.307: %OSPF-5-ADJCHG: Process 100, Nbr 192.168.129.200 on FastEthernet0/1 from FULL to DOWN, N
eighbor Down: Dead timer expired
R1#
```

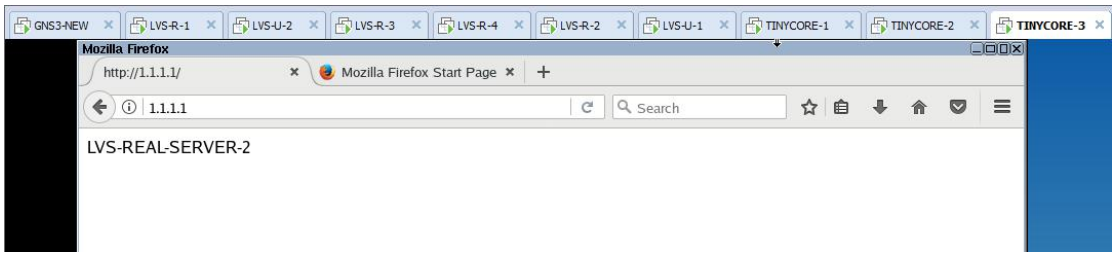
查看路由信息

```
R1#show ip route
Codes: C - connected, S - static, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2
       i - IS-IS, su - IS-IS summary, L1 - IS-IS level-1, L2 - IS-IS level-2
       ia - IS-IS inter area, * - candidate default, U - per-user static route
       o - ODR, P - periodic downloaded static route

Gateway of last resort is not set

     1.0.0.0/32 is subnetted, 1 subnets
O       1.1.1.1 [110/110] via 192.168.128.200, 00:01:10, FastEthernet0/0
C    192.168.128.0/24 is directly connected, FastEthernet0/0
C    192.168.129.0/24 is directly connected, FastEthernet0/1
C    192.168.130.0/24 is directly connected, FastEthernet1/0
R1#
```

再看客户机的表现：







总结一点：
OSPF 的跨网段应用层负载均衡，没有很好的实现，可能我的理解比较初级，不知道掉到哪个坑里去了，或者我在客户端的测试比较简单？另外还有 web 服务器那边的有效期问题，以及浏览器缓存问题，在这个实验中没有处理彻底。

实现的是：
第一， 在路由器上查看包的路由选择，是实现了负载均衡的。
第二， 当一条线路出现问题时，可以完成故障线路的隔离，并切换到可用线路上。

另外，在这个过程中可能遇到的一些坑，在另外的几篇连续的文章中，已经全部展开，不在这里重复。

GNS3 这个模拟器还是很强大的，非常适合作为测试与练手环境使用。尤其是可以让开发人员对于完整的网络结构以及网络设备有更加直接的认识，辅助理解真实的网络结构。

警告：以上配置与解读为非专业网络人员的业余活动，仅做理论层面的探讨，不能直接作为生产活动的依据。