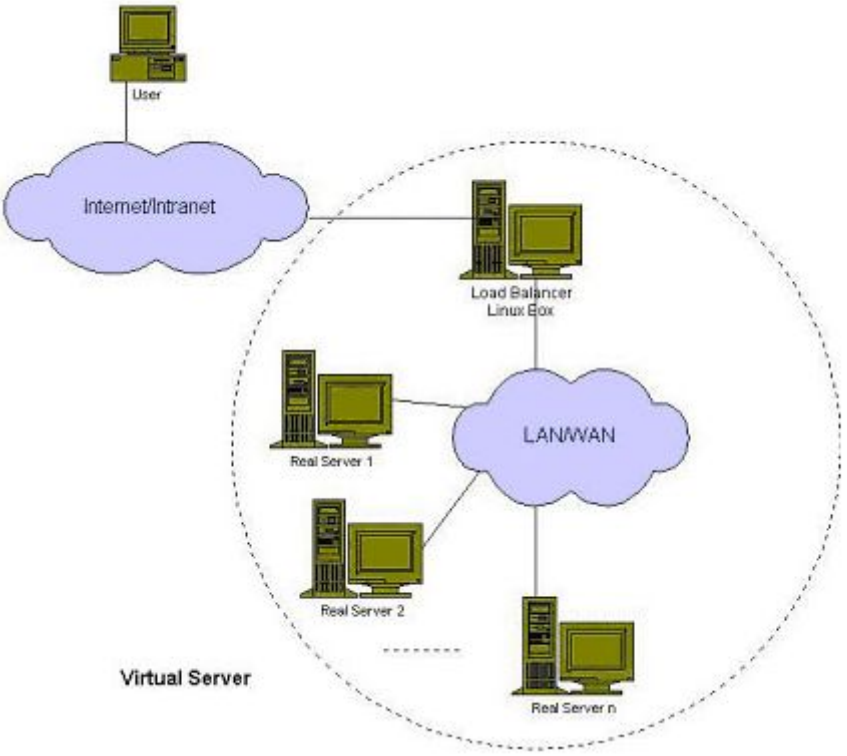
LVS配置初步

1. **简介**

针对高可伸缩、高可用网络服务的需求，出现基于IP层（四成）和基于内容请求分发（七层）的负载平衡调度解决方法，并在Linux内核中实现了这些方法，将一组服务器构成一个实现可伸缩的、高可用网络服务的虚拟服务器。虚拟服务器的体系结构如图所示，**一组服务器通过高速的局域网或者地理分布的广域网相互连接，在它们的前端有一个负载调度器（Load Balancer）。负载调度器能无缝地将网络请求调度到真实服务器上，从而使得服务器集群的结构对客户是透明的**，客户访问集群系统提供的网络服务就像访 问一台高性能、高可用的服务器一样。客户程序不受服务器集群的影响不需作任何修改。系统的伸缩性通过在服务机群中透明地加入和删除一个节点来达到，通过检 测节点或服务进程故障和正确地重置系统达到高可用性。由于我们的负载调度技术是在Linux内核中实现的，我们称之为Linux虚拟服务器（Linux Virtual Server）。



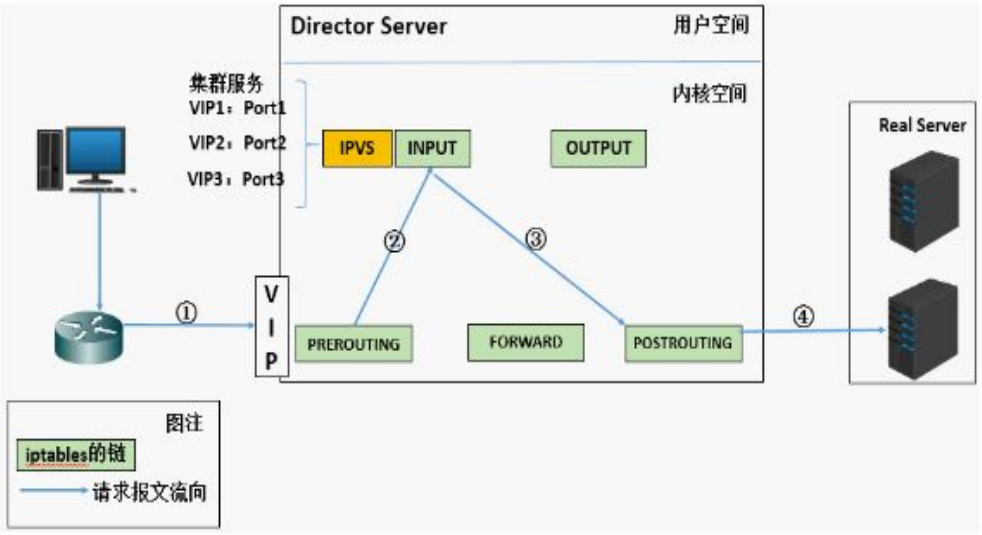
2. 基本工作原理

a) 当用户向负载均衡调度器（Director Server）发起请求，调度器将请求发往至内核空间

b) PREROUTING链首先会接收到用户请求，判断目标IP确定是本机IP，将数据包发往INPUT链

c) IPVS是工作在INPUT链上的，当用户请求到达INPUT时，IPVS会将用户请求和自己已定义好的集群服务进行比对，如果用户请求的就是定义的集群服务，那么此时IPVS会强行修改数据包里的目标IP地址及端口，并将新的数据包发往POSTROUTING链

d) POSTROUTING链接收数据包后发现目标IP地址刚好是自己的后端服务器，那么此时通过选路，将数据包最终发送给后端的服务器



3. LVS组成和相关术语

a) LVS由两部分组成（ipvs，ipvsadm）

i. ipvs(ip virtual server)：一段代码工作在内核空间，叫ipvs，是真正生效实现调度的代码。

ii. ipvsadm：另外一段是工作在用户空间，叫ipvsadm，负责为ipvs内核框架编写规则，定义谁是集群服务，而谁是后端真实的服务器(Real Server)，是面向用户的，管理ipvs的工具

b) 基本术语如下

i. DS（director server）：前端负载均衡节点

ii. RS（real server）：真实服务器

iii. VIP（virtual IP）：面向internet的前端负载均衡节点的IP

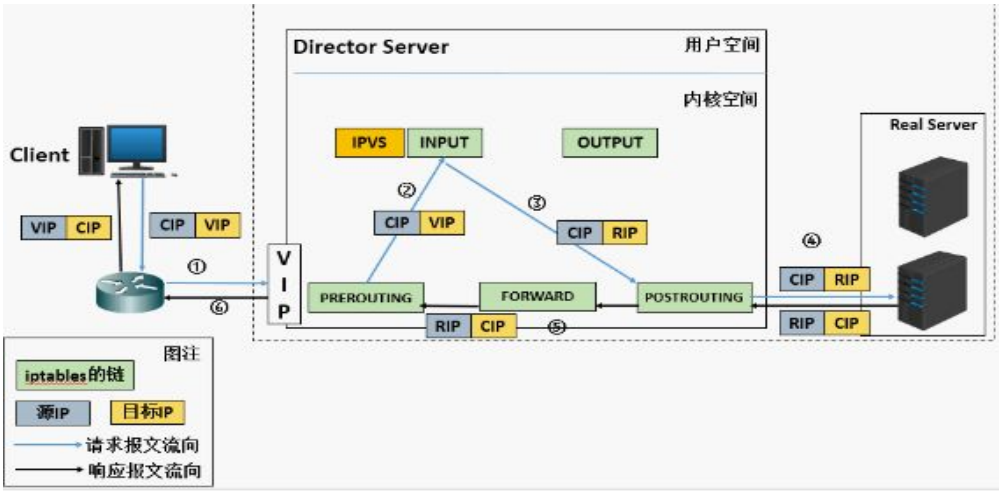
iv. DIP（director server IP）：DS用于和内部主机通讯的IP

v. RIP（real server IP）：后端真实服务器IP地址

vi. CIP（client IP）:客户端IP地址

4. 三种工作模式

a) NAT



NAT模式需要：RS使用私有IP地址，并且RS网关指向DIP，DIP和RIP处于同一网段（支持端口映射）

i. 当用户请求到达Director Server，此时请求的数据报文会先到内核空间的PREROUTING链。 此时报文的源IP为CIP，目标IP为VIP

ii. PREROUTING检查发现数据包的目标IP是本机，将数据包送至INPUT链

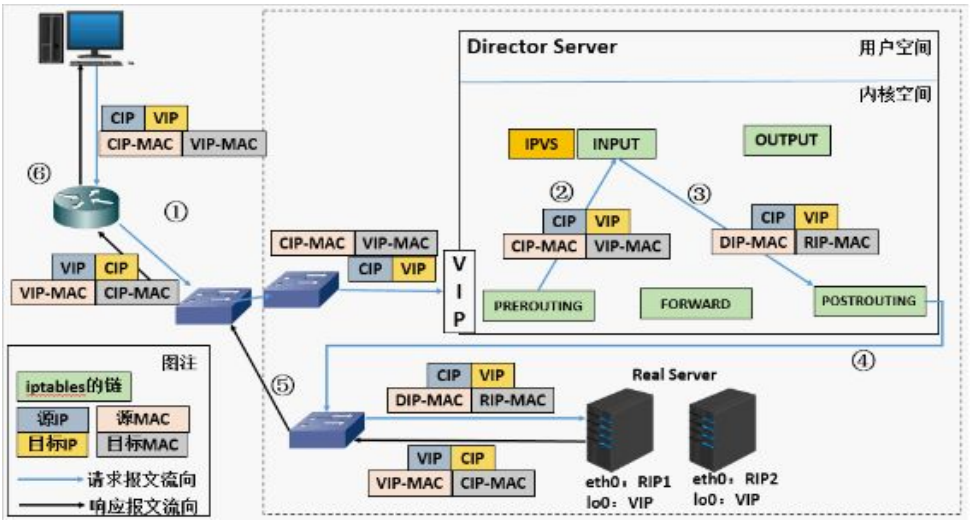
iii. IPVS比对数据包请求的服务是否为集群服务，若是，修改数据包的目标IP地址为后端服务器IP，然后将数据包发至POSTROUTING链。 此时报文的源IP为CIP，目标IP为RIP

iv. POSTROUTING链通过选路，将数据包发送给Real Server

v. Real Server比对发现目标为自己的IP，开始构建响应报文发回给Director Server。 此时报文的源IP为RIP，目标IP为CIP

vi. Director Server在响应客户端前，此时会将源IP地址修改为自己的VIP地址，然后响应给客户端。 此时报文的源IP为VIP，目标IP为CIP

b) DR（将请求目标的MAC地址改为挑选出的RS的MAC地址）



i. 当用户请求到达Director Server，此时请求的数据报文会先到内核空间的PREROUTING链。 此时报文的源IP为CIP，目标IP为VIP

ii. PREROUTING检查发现数据包的目标IP是本机，将数据包送至INPUT链

iii. IPVS比对数据包请求的服务是否为集群服务，若是，将请求报文中的源MAC地址修改为DIP的MAC地址，将目标MAC地址修改RIP的MAC地址，然后将数据包发至POSTROUTING链。 此时的源IP和目的IP均未修改，仅修改了源MAC地址为DIP的MAC地址，目标MAC地址为RIP的MAC地址

iv. 由于DS和RS在同一个网络中，所以是通过二层来传输。POSTROUTING链检查目标MAC地址为RIP的MAC地址，那么此时数据包将会发至Real Server。

v. RS发现请求报文的MAC地址是自己的MAC地址，就接收此报文。处理完成之后，将响应报文通过lo接口传送给eth0网卡然后向外发出。 此时的源IP地址为VIP，目标IP为CIP

vi. 响应报文最终送达至客户端

vii. tips：保证前端路由将目标地址为VIP报文统统发给Director Server，而不是RS；RS可以使用私有地址，也可以是公网地址；RS和DS在同一物理网络；请求报文经过DS，但响应报文不经过DS；不能进行端口映射，不能进行地址转换；RS的lo接口回环地址设置为VIP的地址；RS网关不允许指向VIP

c) TUN（在原有的IP报文外再次封装多一层IP首部）

i. 当用户请求到达Director Server，此时请求的数据报文会先到内核空间的PREROUTING链。 此时报文的源IP为CIP，目标IP为VIP 。

ii. PREROUTING检查发现数据包的目标IP是本机，将数据包送至INPUT链

iii. IPVS比对数据包请求的服务是否为集群服务，若是，在请求报文的首部再次封装一层IP报文，封装源IP为为DIP，目标IP为RIP。然后发至POSTROUTING链。 此时源IP为DIP，目标IP为RIP

iv. POSTROUTING链根据最新封装的IP报文，将数据包发至RS（因为在外层封装多了一层IP首部，所以可以理解为此时通过隧道传输）。 此时源IP为DIP，目标IP为RIP

v. RS接收到报文后发现是自己的IP地址，就将报文接收下来，拆除掉最外层的IP后，会发现里面还有一层IP首部，而且目标是自己的lo接口VIP，那么此时RS开始处理此请求，处理完成之后，通过lo接口送给eth0网卡，然后向外传递。此时的源IP地址为VIP，目标IP为CIP

vi. 响应报文最终送达至客户端

vii. Tips：请求报文经过Ds，但响应报文不经过DS，不支持端口映射

5. 调度算法

a) 轮叫rr
这种算法是最简单的，就是按依次循环的方式将请求调度到不同的服务器上，该算法最大的特点就是简单。轮询算法假设所有的服务器处理请求的能力都是一样的，调度器会将所有的请求平均分配给每个真实服务器，不管后端 RS 配置和处理能力，非常均衡地分发下去。

b) 加权轮叫wrr
这种算法是最简单的，就是按依次循环的方式将请求调度到不同的服务器上，该算法最大的特点就是简单。轮询算法假设所有的服务器处理请求的能力都是一样的，调度器会将所有的请求平均分配给每个真实服务器，不管后端 RS 配置和处理能力，非常均衡地分发下去。

c) 最少连接lc
这种算法是最简单的，就是按依次循环的方式将请求调度到不同的服务器上，该算法最大的特点就是简单。轮询算法假设所有的服务器处理请求的能力都是一样的，调度器会将所有的请求平均分配给每个真实服务器，不管后端 RS 配置和处理能力，非常均衡地分发下去。

d) 基于局部性最少连接调度lblc
这种算法是最简单的，就是按依次循环的方式将请求调度到不同的服务器上，该算法最大的特点就是简单。轮询算法假设所有的服务器处理请求的能力都是一样的，调度器会将所有的请求平均分配给每个真实服务器，不管后端 RS 配置和处理能力，非常均衡地分发下去。

e) Lblcr

f) 目标地址散列调度dh
该算法是根据目标 IP 地址通过散列函数将目标 IP 与服务器建立映射关系，出现服务器不可用或负载过高的情况下，发往该目标 IP 的请求会固定发给该服务器。

g) 源地址散列调度sh
与目标地址散列调度算法类似，但它是根据源地址散列算法进行静态分配固定的服务器资源。

6. DR模式实践初步（virtual box 5.1.26 Centos7 minimal）

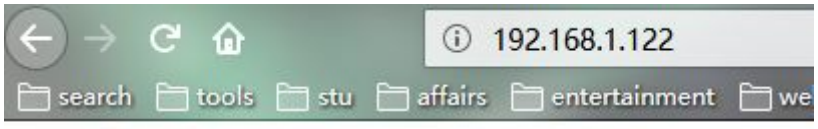
a) 配置虚拟机环境，创建centos虚拟机，并链接复制三台机器测试，网络连接选择桥接模式

b) 配置虚拟机静态IP地址，方法为编辑/etc/sysconfig/network-scripts/ifcfg-enp0s3
（enp0s3为网络接口的名称，不同的机器有不同的名称）
Loader Balance:192.168.1.120（VIP）

Real Server 1 :192.168.1.121
Real Server 2 :192.168.1.122
IP地址这样设置是因为宿主机IP段为192.168.1.0/255，默认网关为192.168.1.1
编辑该配置文件设置IP地址，默认网关，设置完成之后重启network服务，可访问外网，关闭windows防火墙或放通宿主机icmp数据包后虚拟机和宿主机可以相互访问

```
TYPE=Ethernet
#PROXY_METHOD=none
#BROWSER_ONLY=no
BOOTPROTO=static
IPADDR=192.168.1.122
NETMASK=255.255.255.0
DNS1=114.114.114.114
GATEWAY=192.168.1.1
#DEFROUTE=yes
#IPV4_FAILURE_FATAL=no
#IPV6INIT=yes
#IPV6_AUTOCONF=yes
#IPV6_DEFROUTE=yes
#IPV6_FAILURE_FATAL=no
#IPV6_ADDR_GEN_MODE=stable-privacy
NAME=enp0s3
UUID=4fe44a51-dd56-4398-8129-50c23405e627
DEVICE=enp0s3
ONBOOT=yes
ZONE=public
```

c) 安装httpd，编写测试页面，防火墙放通80端口，确认访问畅通



Real Server 2!

Real Server 1!

d) 创建RS机器的虚拟网卡
vim /etc/sysconfig/network-scripts/lo:0

```
DEVICE=lo:0
BOOTPROTO=static
ONBOOT=yes
TYPE=Ethernet
IPADDR=192.168.1.120
NETMASK=255.255.255.255
GATEWAY=192.168.1.1
```

重启network服务，使用ifconfig查看配置信息看到如下虚拟

```
lo:0: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 192.168.1.120  netmask 255.255.255.255
        loop  txqueuelen 1  (Local Loopback)
```

e) 抑制RS的arp协议响应
vim /etc/sysctl.conf

```
net.ipv4.conf.enp0s3.arp_ignore=1
net.ipv4.conf.enp0s3.arp_announce=2
net.ipv4.conf.all.arp_ignore=1
net.ipv4.conf.all.arp_announce=2
```

sysctl -p #load values

f) 配置Loader Balance的ipvsadm
yum install ipvsadm
ipvsadm -A -t 192.168.1.120 -s rr
ipvsadm -a -t 192.168.1.120:80 -r 192.168.1.121:80 -g
ipvsadm -a -t 192.168.1.120:80 -r 192.168.1.122:80 -g
ipvsadm -Sn > /etc/sysconfig/ipvsadm
-A选项添加一个虚拟服务器，-s指定负载均衡算法为rr（轮询调度），-a添加一台真实服务器，-t使用tcp连接，-r指定真实服务器的IP地址，-g指定ipvsadm调度模式为DR直连，-S将配置保存到标准输出

g) 配置Loader Balance允许重定向
vim /etc/sysctl.conf

```
net.ipv4.ip_forward=1
```

sysctl -p

h) 访问验证（两次访问需要间隔一段时间）
访问192.168.1.120

Real Server 1!

Real Server 2!

当前连接状态

```
[root@localhost ~]# ipvsadm -Lnc
IPVS connection entries
pro expire state       source             virtual            destination
TCP 06:16  ESTABLISHED 192.168.1.104:31865 192.168.1.120:80   192.168.1.122:80
TCP 08:01  ESTABLISHED 192.168.1.104:31975 192.168.1.120:80   192.168.1.122:80
TCP 13:59  ESTABLISHED 192.168.1.104:32135 192.168.1.120:80   192.168.1.122:80
TCP 07:57  ESTABLISHED 192.168.1.104:31971 192.168.1.120:80   192.168.1.121:80
TCP 05:58  ESTABLISHED 192.168.1.104:31858 192.168.1.120:80   192.168.1.121:80
TCP 07:12  ESTABLISHED 192.168.1.104:31876 192.168.1.120:80   192.168.1.121:80
TCP 07:30  ESTABLISHED 192.168.1.104:31957 192.168.1.120:80   192.168.1.122:80
TCP 13:53  ESTABLISHED 192.168.1.104:32048 192.168.1.120:80   192.168.1.121:80
[root@localhost ~]# _
```

i) DR模式设置时遇到的问题

i. 编辑RS的lo接口设置VIP时，编辑了ifcfg-lo配置文件，但是重启network后没有生效，重新编辑ifcfg-lo:0,使多个IP地址监听同一接口，使用ifconfig查看生效

ii. 配置RS静态IP地址时，指定GATEWAY才可以访问外网

iii. 理解sysctl的工作方式，控制IP转发的方法，了解抑制arp协议的方法