LVS 工作模式以及工作原理

LVS 简介

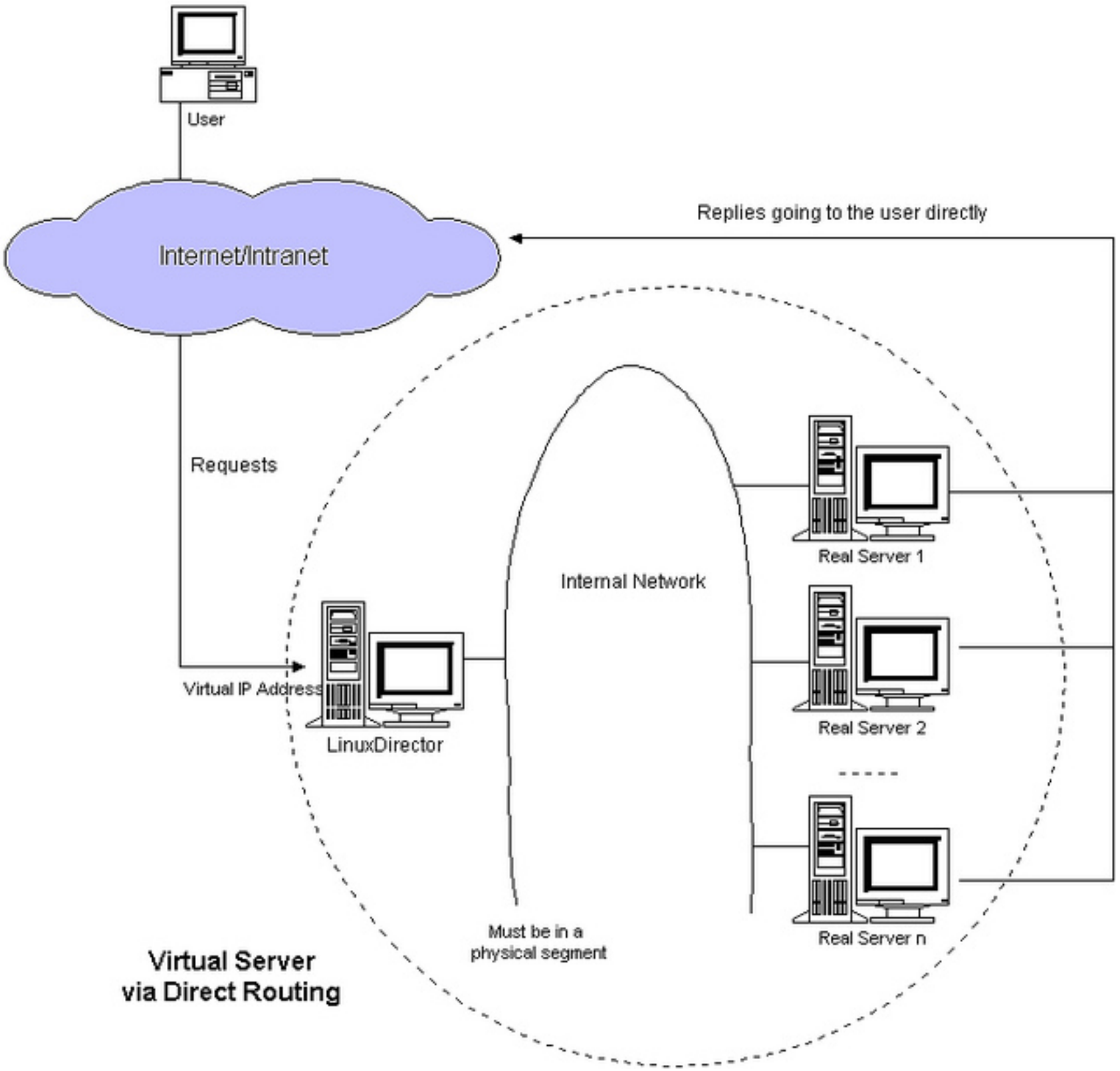
LVS 是 Linux Virtual Server ，Linux 虚拟服务器；是一个虚拟的服务器集群【多台机器 LB IP】。LVS 集群分为三层结构:

- 负载调度器(load balancer)：它是整个LVS 集群对外的前端机器，负责将client请求发送到一组服务器[多台LB IP]上执行，而client端认为是返回来一个同一个IP【通常把这个IP 称为虚拟IP/VIP】
- 服务器池(server pool)：一组真正执行client 请求的服务器，一般是我们的web服务器；除了web，还有FTP，MAIL，DNS
- 共享存储(shared stored)：它为 server pool 提供了一个共享的存储区，很容易让服务器池拥有相同的内容，提供相同的服务[不是很理解]

LVS 有4中常用的模式，分别讲一下4中模式的区别:

LVS DR 模式

1. DR(Direct Routing)模式的网络结构:



2. 工作的基本原理：

(1). client 发送一个pv请求给VIP；VIP 收到这请求后会跟LVS设置的LB算法选择一个LB 比较合理的realserver，然后把此请求的package 的MAC地址修改为realserver的MAC地址；下面是我们通信的package的基本格式：

Src mac	Dst mac	type	...	source ip	src port	dst ip	dst port	...	CRC
...	...	...	...	192.168.57.135	55014	192.168.57.126	80	...	...

source MAC	dest MAC
00:18:82:3c:e8:96	00:0c:29:6a:8d:5d

在这个通信的Package 有六个主要的字段：**src mac 、 dst mac 、 src ip 、 src prot 、 dst ip 、 dst ip ；** 现在这个包里面的 dst mac 是LVS VIP的网卡MAC [在TCP 三次握手完成时就只知道dsp ip 和dsp mac了]

- DR 模式会把packet 里面的dst mac 改成 realserver的MAC 地址；然后VIP会把这个包广播到当前的这个LAN里面；所以，要提前保证VIP 和所有的realserver 在同一个网段，也就是在用过LAN里面。

同一个网段：用子网掩码来实现的，我们知道我们的网络中有局域网，一个局域网有很多台机器，这些LAN里面的所有机器都公用一个外网IP；我们是怎样界定这个LAN的呢？用的就是网段号；**IP只是是32位二进制数表示，这32位分为：网络位 + 主机位；表现在子网掩码是就是：网络位是1，主机位是0；这样网络位 = IP 按位 与 子网掩码**；所以，我们在把realserver 挂到LVS上前，需要确认DR模式，且IP 在同一个网段内。

- ARP协议会把这个包发送给真正的realserver【根据MAC 找到机器】
- 把这个src ip----->realserver 的mac 地址建立一个hash表；这此次连接未断开前，同一个client发送的请求通过查询hash表，在次发送到这台realserver上面；
- realserver 收到这个pachet后，**首先判断dst ip 是否是自己的IP地址；**如果不是就丢掉包；如果是就处理这个包。所以，DR模式还要在所有的realserver 的机器上面绑定VIP的ip地址：

```
/sbin/ifconfig lo:0 inet VIP netmask 255.255.255.255   -----> 这个要注意!
```

- 这样realserver 发现package 的dst 自己能识别【绑定了2个IP】，会处理这个包，处理完后把package的src mac dst mac src ip dst ip 都修改后再通过ARP 发送给VIP，通过VIP 发送给client。 realserver 发送给 VIP 的package的格式：

Src mac	Dst mac	type	...	source ip	src port	dst ip	dst port	...	CRC
...	...	...	...	192.168.57.126	80	192.168.57.135	55014	...	...

source MAC	dest MAC
00:0c:29:b1:97:82	00:18:82:3c:e8:96

51CTO.com
技术成就梦想

- realserver 处理这个包后，会跟dst 为client ip 直接发送给 client ip；不经过lvs；这样虽然效率比较高，但是有安全漏洞。

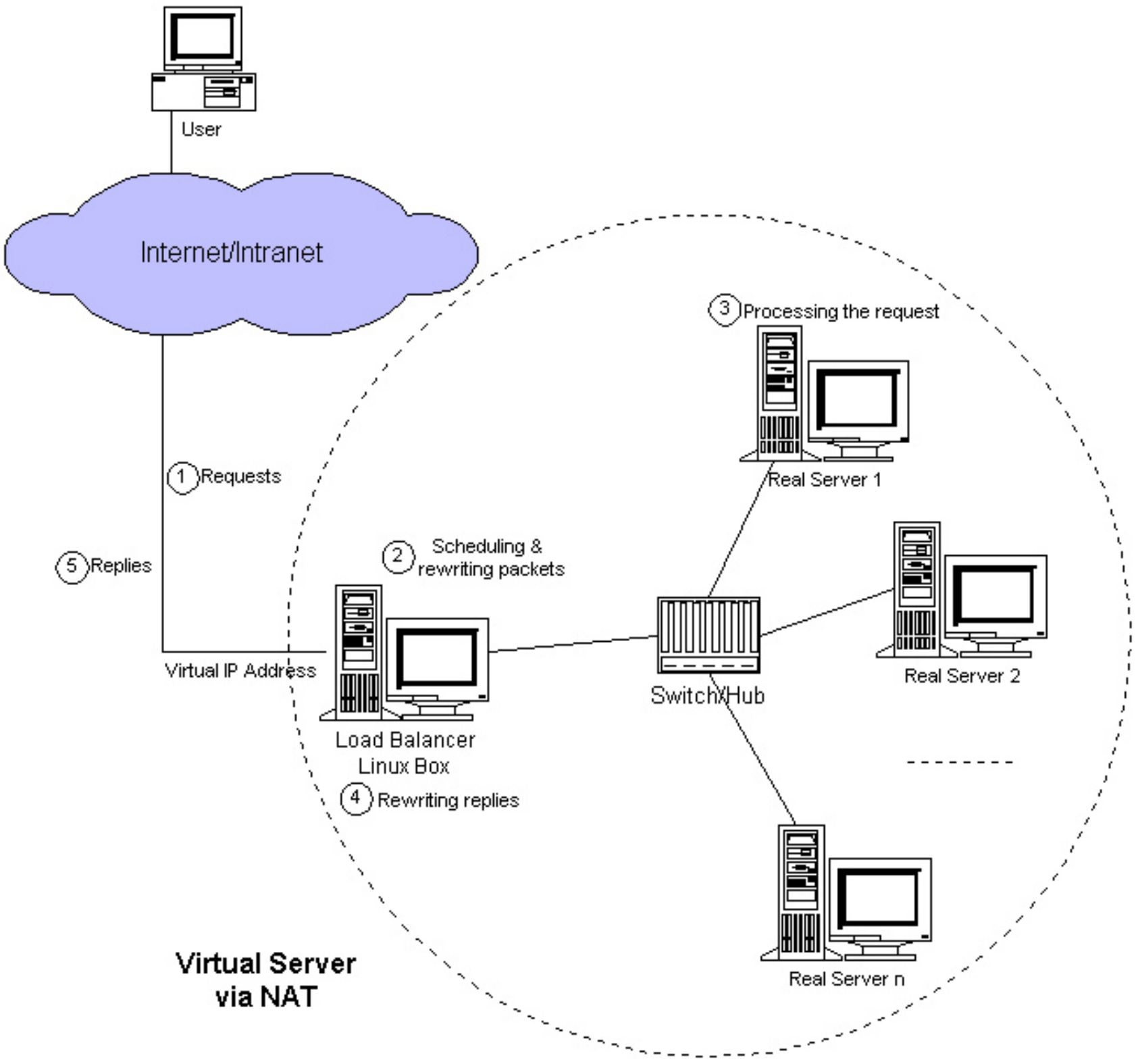
LVS DR 工作的基本原理package 的详细信息：http://os.51cto.com/art/201105/264303.htm

3. LVS DR模式的注意情况：

- LVS 的VIP 和 realserver 必须在同一个网段，不然广播后所有的包都会丢掉： 提前确认LVS/硬件LB 是什么模式，是否需要在同一个网段
- 所有的realserver 都必须绑定VIP的IP地址，否则realserver 收到package后发现dst 不是自己的I怕，所有包都会丢掉。
- realserver · 处理完包后直接把package 通过dst IP 发送给 client ， 不通过LVS/迎接IP 了这样的LVS /VIP 效率会更高一点。【通过把realserver的ip暴漏给外界，不是很安全】

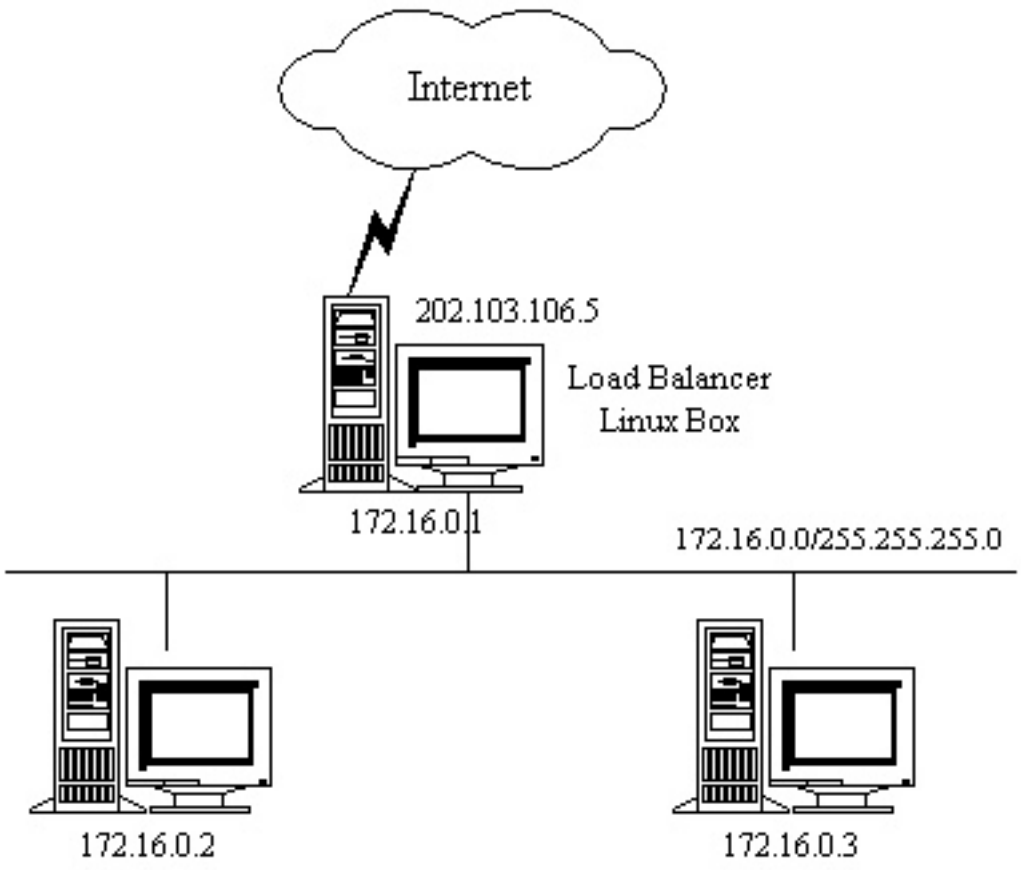
LVS NAT 模式

1. LVS NAT 模式的网络结构：



2. NAT 模式的基本原理:

- NAT 模式工作原理的模拟图:



- client：202.100.1.2
  VIP:　202.103.106.5

realserver : 172.16.0.2 172.16.0.3 分别提供http 和ftp服务

(1). 首先client 发送请求[package] 给VIP；

```
#client 发送给VIP的package：

SOURCE 202.100.1.2:3478  EDST   202.103.106.5:80
```

(2). VIP 收到package后，会根据LVS设置的LB算法选择一个合适的realserver，然后把package 的DST IP 修改为realserver：

```
# VIP 发送给realserver的package：

 SOURCE   202.100.1.2:3478  EDST   172.16.0.3：8000
```

(3). realserver 收到这个package后判断dst ip 是自己，就处理这个package，处理完后把这个包发送给LVS VIP：

```
# realserver 处理完成后发送给VIP的package：

 SOURCE   172.16.0.3：8000  EDST     202.100.1.2:3478    # lvs 收到这个package 后发现dst ip 不是自己的会不会丢掉？感觉有错误
```

(4). LVS 收到这个package 后把sorce ip改成VIP的IP，dst ip改成 client ip 然后发送给client：

```
#VIP收到package 后修改sourceip 发送给client的包：
SOURCE   202.103.106.5.80：80  EDST   202.100.1.2:3478
```

3. NAT 模式的注意事项：

- NAT 模式修改的是dst IP，直接走 switch 或pub 不需要修改MAC 所以，不需要VIP 和realserver 同在一个网段内。
- NAT 模式 package in 和package out 都需要经过LVS ； 因此LVS 的可能会成为一个系统瓶颈问题。

LVS FULL NAT 模式

1. FULL NATT的基本原理：

FULL NAT 在client请求VIP 时，不仅替换了package 的dst ip，还替换了package的 src ip；但VIP 返回给client时也替换了src ip；还是通过上面 NAT 模式的工作原因的图进行分析 FULL NAT 的工作原理：

(1). 首先client 发送请求[package] 给VIP；

```
#client 发送给VIP的package：
SOURCE 202.100.1.2:3478  EDST   202.103.106.5:80
```

(2). VIP 收到package后，会根据LVS设置的LB算法选择一个合适的 realserver，然后把package 的DST IP 修改为realserver；把sorce ip 改成 lvs 集群的LB IP

```
# VIP 发送给realserver的package：
 SOURCE   172.24.101.135[lb ip]  EDST   172.16.0.3：8000
```

(3). realserver 收到这个package后判断dst ip 是自己，就处理这个package ，处理完后把这个包发送给LVS VIP：

```
# realserver 处理完成后发送给VIP的package:

 SOURCE   172.16.0.3:8000  EDST     172.24.101.135[这个ip是 LVS VIP
集群的一太机器]
```

(4). LVS 收到这个package 后把sorce ip改成VIP的IP，dst ip改成 client ip 然后发送给client:

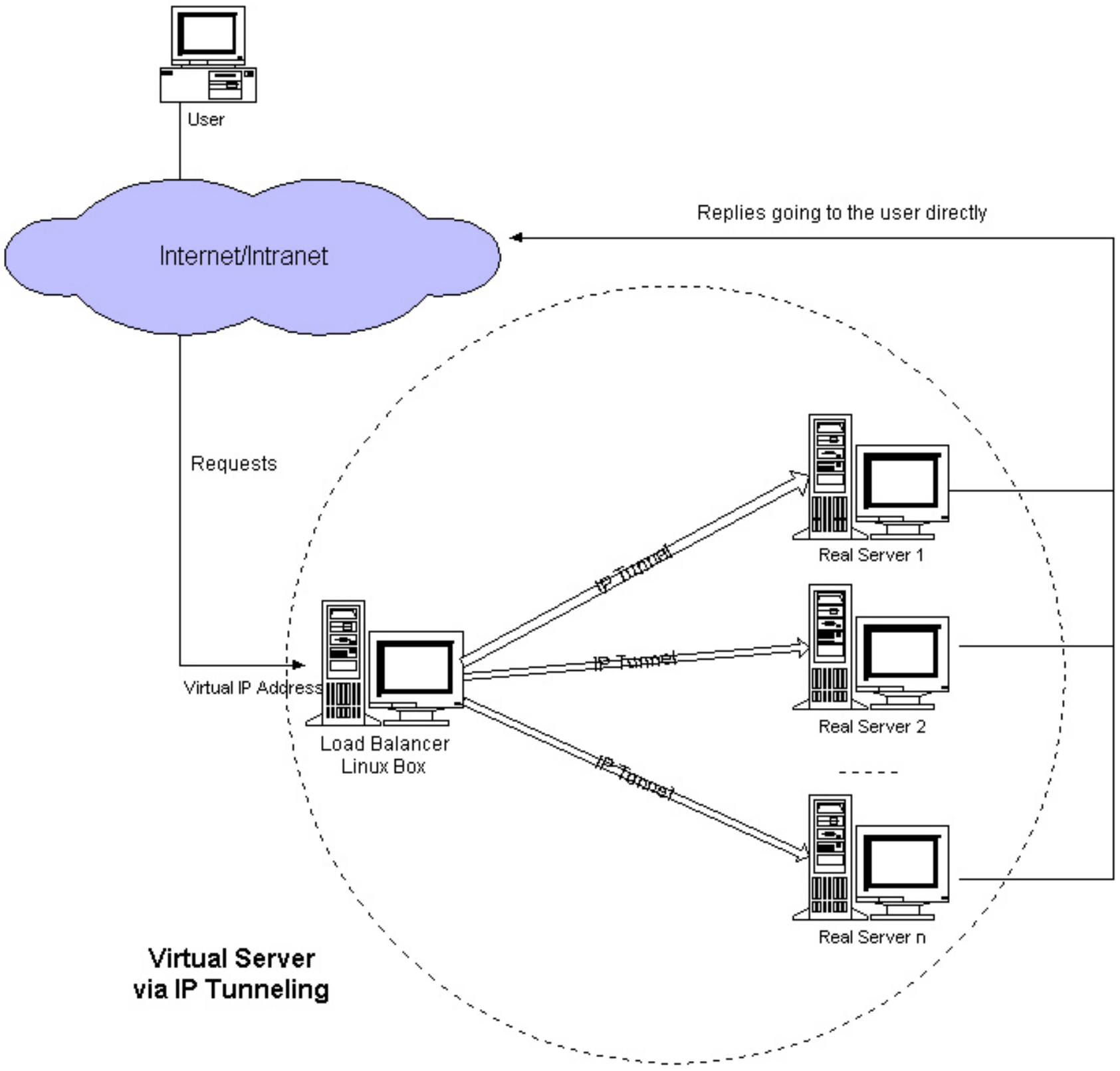
```
#VIP收到package 后修改sourceip 发送给client的包:

SOURCE   202.103.106.5.80:80  EDST   202.100.1.2:3478
```

2. FULL NAT 模式的注意事项:

- FULL NAT 模式也不需要 LBIP 和realserver ip 在同一个网段;
- full nat 跟nat 相比的优点是：保证RS回包一定能够回到LVS；因为源地址就是LVS--> 不确定
- full nat 因为要更新sorce ip 所以性能正常比nat 模式下降 10%

LVS IP TUNNEL 模式

1. IP TUNNEL 模式的网络结构图:



2. IP TUNNEL 模式的基本原理：

还是按NAT 模式的基本框架来说明TUNNEL 模式的基本原理：

(1). 首先client 发送请求[package] 给VIP；

```
#client 发送给VIP的package：

SOURCE 202.100.1.2:3478  DST   202.103.106.5:80
```

(2). VIP 收到package后，会根据LVS设置的LB算法选择一个合适的realserver；并把client发送的package 包装到一个新的IP包里面；新的IP包的dst是realserver的IP

```
# VIP 发送给realserver的package:

 client 发送的包   DST   172.16.0.3: 8000
```

(3). realserver 收到这个package后判断dst ip 是自己，然后解析出来的package的dst是VIP；会检测我们的网卡上是否帮了VIP的ip地址；如果帮了就会处理这个包，如果没有直接丢掉。 我们一般在realserver上面 lo:0 绑定了VIP的ip地址，就可以处理

```
# realserver 处理完成后直接发送给client响应包:

 SOURCE   172.16.0.3: 8000  DST      202.100.1.2:3478   【client ip】
```

3. IP TUNNEL 模式的注意:

- TUNNEL 模式必须在所有的realserver 机器上面绑定VIP的IP地址
- TUNNEL 模式的vip ------>realserver 的包通信通过TUNNEL 模式，不管是内网和外网都能通信，所以不需要lvs vip跟realserver 在同一个网段内
- TUNNEL 模式 realserver会把packet 直接发给client 不会给lvs了
- TUNNEL 模式走的隧道模式，所以运维起来比较难，所以一般不用

LVS DR、NAT、FULL NAT、IP TUNNEL 模式的区别:

1. 是否需要lvs vip跟realserver 在同一个网段:

DR 模式因为只修改 package的 MAC地址通过ARP广播的形势找到realserver，所以 要求LVS 的VIP 和realserver的IP 必须在同一个网段内，也就是在挂载VIP 时先确认LVS的工作模式，如果是DR模式需要先确认这个IP只是否能挂在这个LVS下面。

其他模式因为都会修改DST ip 为 realserver的IP 地址，所以不需要在同一个网段内

2. 是否需要在realserver 绑定LVS vip 的IP 地址:

realserver 收到package后会判断dst ip 是否是自己的ip，如果不是就直接丢掉包；因为DR模式dst 没有修改还是LVS的VIP；所以需要在

realserver上面绑定VIP；IP TUNNEL 模式只是对package 重新包装一层，realserver解析后的IP包的DST 仍然是 LVS的VIP ；也需要在realserver上面绑定VIP；其他的都不需要

3. 四种模式的性能比较：

因为DR模式 TP TUNELL 模式都是在package in 时经过LVS ；在package out是直接返回给client；所以二者的性能比NAT 模式高；但IP TUNNEL 因为是TUNNEL 模式比较复杂，其性能不如DR模式；FULL NAT 模式因为不仅要更换 DST IP 还更换 SOURCE IP 所以性能比NAT 下降10%

所以，4中模式的性能如下：DR --> IP TUNNEL --->NAT ----->FULL NAT

LVS 实践中的积累

1. lvs 不会主动断开连接

比如 client 通过LVS VIP 采用长链接方式访问server，即使我们把LVS下面的realserver的status.html文件删除了；本来通过LVS 跟这台realserver 链接的请求也不会被LVS

强制断开；要等到client自己断开连接；【在client主动断开期间；client可以跟这台realserver 正常通信】；这样有个好处时在网络抖动时；LVS不会频繁的流量截断，到不同的RS上面