链滴

LVS 原理简介

作者：qq1072023998
原文链接：https://ld246.com/article/1605611802305
来源网站：链滴
许可协议：署名-相同方式共享 4.0 国际 (CC BY-SA 4.0)

lvs

1.简介

LVS是一个开源软件，由章文嵩博士于1998年5月创立，可以实现Linux平台下的简单负载均衡。LVS Linux Virtual Server的简写，是一个虚拟的服务器集群系统。在后续的维护过程中，LVS添加了各种的支持，包括FreeBSD的支持、IPV6的支持、TCPHA的支持等等。

LVS为传输层协议，基于linux内核的netfilter模块，现在一般都集成与linux内核模块，无需单独安装使用命令：lsmod | grep ip_vs，可以查看lvs是否已在内核模块中。

lvs分为ipvs和ipvsadmin，前者为已集成与内核模块的ipvs核心功能模块（内核如没有可以编译进内），后者为lvs的命令行工具，其命令格式也类似与iptables。

lvs工作模式：ipvs是工作在内核中netfilter的INPUT的钩子函数上，对进入的报文在没有进入用户空前，对这些报文进行操作。优点是性能高，支持并发大，无socket限制和最大文件句柄数限制。缺点调度方式简单，仅能通过ip+port进行调度。

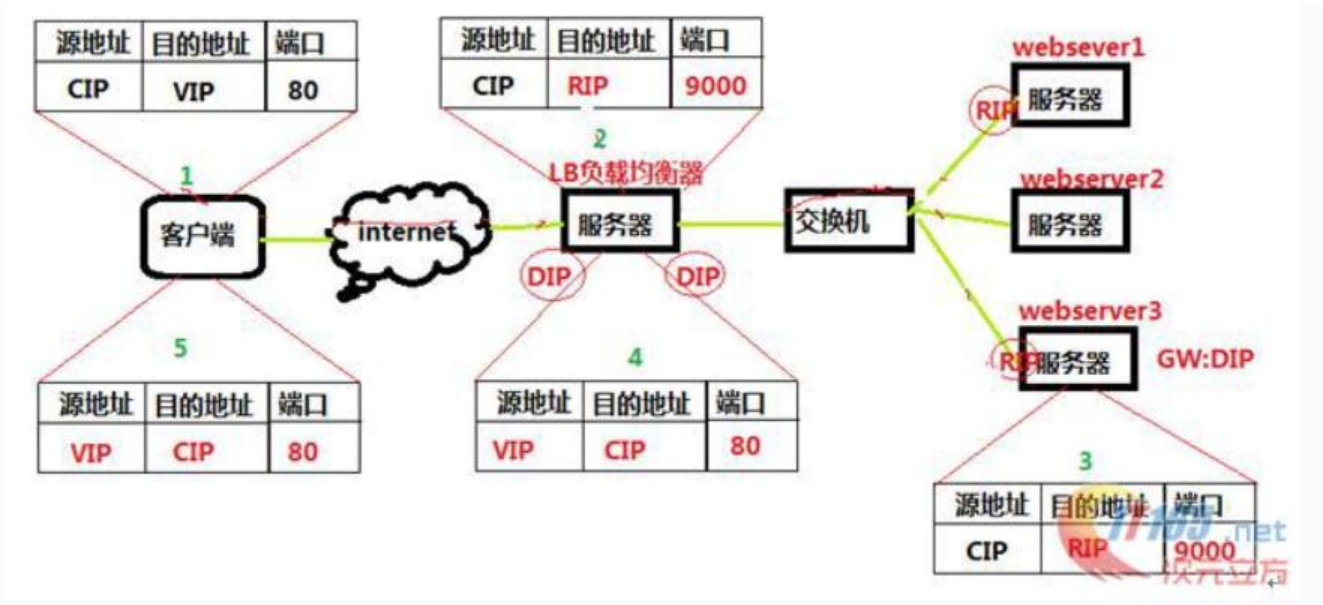
2.调度模式

lvs有三种调度模式，NAT,DR,TUN

备注：CIP(客户端IP)，VIP（lvs调度器的外网监听IP），DIP（lvs调度器和后端realserver联系的IP）RIP（后端realserver的IP）

2.1 NAT模式

nat模式是通过地址转换来实现的调度方式，当客户端访问服务器时，访问的地址应是lvs监听的外网址，也就是VIP，此时请求报文的地址应为CIP->VIP（请求地址->目标地址），到达lvs调度器以后，ls通过修改该报文的目标地址，来进行转发到后端server，此时该报文为CIP->RIP（通过DIP转发）。端realserver在处理完请求之后再将处理完的报文发送到lvs，此时报文为RIP->CIP（因为此时realser er指定的出口网关为DIP，所以即使该报文目标地址为CIP，也会发送到DIP所在的lvs调度服务器上面。lvs调度器收到报文后将该报文改为VIP->CIP，再发送给客户端，完成处理本次请求。



原文链接：LVS 原理简介

NAT模式优缺点：

1. NAT技术将请求的报文和响应的报文都需要通过LB进行地址改写，因此网站访问量比较大的时候L负载均衡调度器有比较大的瓶颈，一般要求最多之能10-20台节点

2. 只需要在LB上配置一个公网IP地址就可以了。

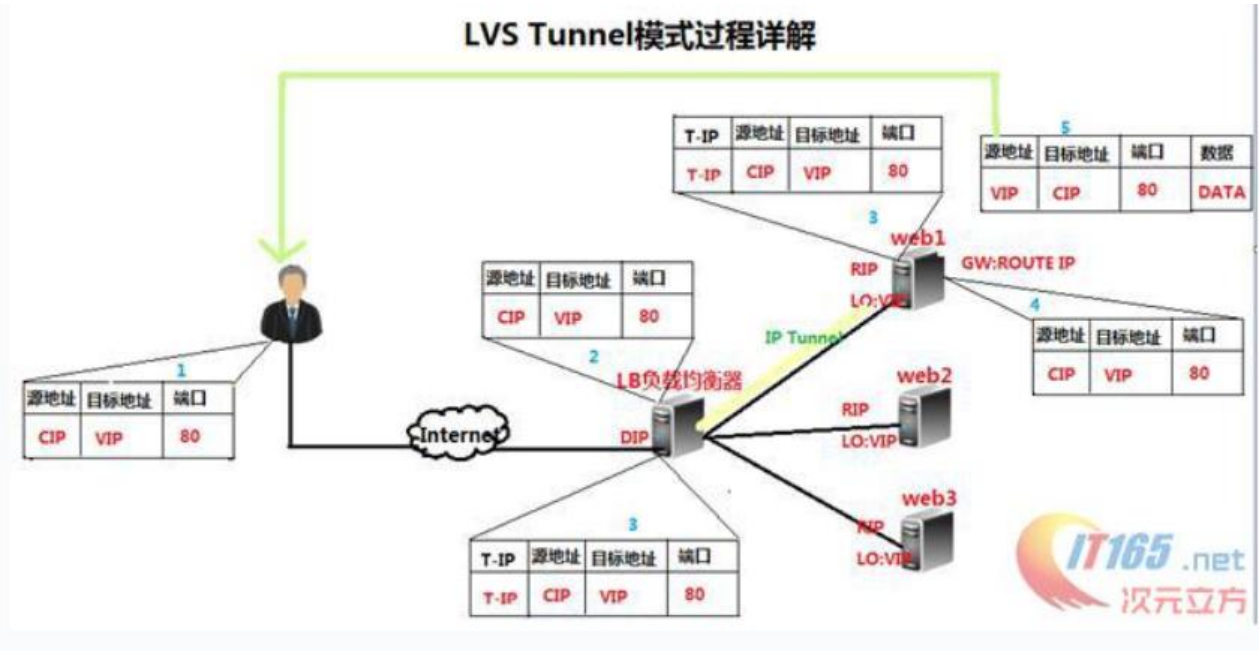
3. 每台内部的节点服务器的网关地址必须是调度器LB的内网地址。

4. NAT模式支持对IP地址和端口进行转换。即用户请求的端口和真实服务器的端口可以不一致。

2.2 TUN模式

virtual server via ip tunneling模式:采用NAT模式时，由于请求和响应的报文必须通过调度器地址重，当客户请求越来越多时，调度器处理能力将成为瓶颈。为了解决这个问题，调度器把请求的报文通过P隧道转发到真实的服务器。真实的服务器将响应处理后的数据直接返回给客户端。这样调度器就只理请求入站报文，由于一般网络服务应答数据比请求报文大很多，采用VS/TUN模式后，集群系统的大吞吐量可以提高10倍。

VS/TUN的工作流程图如下所示，它和NAT模式不同的是，它在LB和RS之间的传输不用改写IP地址。是把客户请求包封装在一个IP tunnel里面，然后发送给RS节点服务器，节点服务器接收到之后解开IP unnel后，进行响应处理。并且直接把包通过自己的外网地址发送给客户不用经过LB服务器。



原理图过程简述：

1. 客户请求数据包，目标地址VIP发送到LB上。

2. LB接收到客户请求包，进行IP Tunnel封装。即在原有的包头加上IP Tunnel的包头。然后发送出去。

3. RS节点服务器根据IP Tunnel包头信息（此时就又一种逻辑上的隐形隧道，只有LB和RS之间懂）收请求包，然后解开IP Tunnel包头信息，得到客户的请求包并进行响应处理。

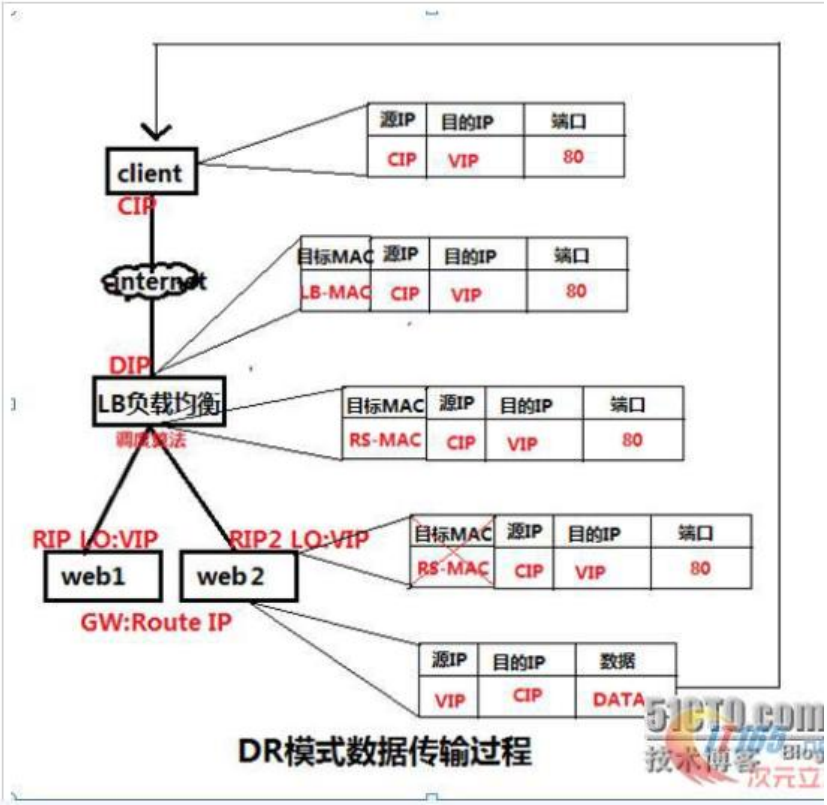
4. 响应处理完毕之后，RS服务器使用自己的出公网的线路，将这个响应数据包发送给客户端。源IP地还是VIP地址。（RS节点服务器需要在本地回环接口配置VIP，后续会讲到）

2.3 DR模式

DR模式是通过改写请求报文的目标MAC地址，将请求发给真实服务器的，而真实服务器响应后的处结果直接返回给客户端用户。同TUN模式一样，DR模式可以极大的提高集群系统的伸缩性。而且DR式没有IP隧道的开销，对集群中的真实服务器也没有必要必须支持IP隧道协议的要求。但是要求调度器B与真实服务器RS都有一块网卡连接到同一物理网段上，必须在同一个局域网环境。

DR模式是互联网使用比较多的一种模式。

DR模式原理图：



VS/DR模式的工作流程图如上图所示，它的连接调度和管理与NAT和TUN中的一样，它的报文转发方和前两种不同。DR模式将报文直接路由给目标真实服务器。在DR模式中，调度器根据各个真实服务的负载情况，连接数多少等，动态地选择一台服务器，不修改目标IP地址和目标端口，也不封装IP报，而是将请求报文的数据帧的目标MAC地址改为真实服务器的MAC地址。然后再将修改的数据帧在务器组的局域网上发送。因为数据帧的MAC地址是真实服务器的MAC地址，并且又在同一个局域网那么根据局域网的通讯原理，真实复位是一定能够收到由LB发出的数据包。真实服务器接收到请求数包的时候，解开IP包头查看到的目标IP是VIP。*（此时只有自己的IP符合目标IP才会接收进来，所以们需要在本地的回环借口上面配置VIP。另：由于网络接口都会进行ARP广播响应，但集群的其他机都有这个VIP的lo接口，都响应就会冲突。所以我们需要把真实服务器的lo接口的ARP响应关闭掉。）然后真实服务器做成请求响应，之后根据自己的路由信息将这个响应数据包发送回给客户，并且源IP址还是VIP。

DR模式小结：

1. 通过在调度器LB上修改数据包的目的MAC地址实现转发。注意源地址仍然是CIP，目的地址仍然是VP地址。

2. 请求的报文经过调度器，而RS响应处理后的报文无需经过调度器LB，因此并发访问量大时使用效很高（和NAT模式比）

3. 因为DR模式是通过MAC地址改写机制实现转发，因此所有RS节点和调度器LB只能在一个局域网里面

4. RS主机需要绑定VIP地址在LO接口上，并且需要配置ARP抑制。

原文链接：LVS 原理简介

5. RS节点的默认网关不需要配置成LB，而是直接配置为上级路由的网关，能让RS直接出网就可以。

6. 由于DR模式的调度器仅做MAC地址的改写，所以调度器LB就不能改写目标端口，那么RS服务器就使用和VIP相同的端口提供服务。

工作模式	VS/NAT	VS/TUN	VS/DR
Real server（节点服务器）	Config dr gw	Tunneling	Non-arp device/tie vip
Server Network	Private	LAN/WAN	LAN
Server number（节点数量）	Low 10-20	High 100	High 100
Real server gateway	Load balance	Own router	Own router
优点	地址和端口转换	Wan环境加密数据	性能最高
缺点	效率低	需要隧道支持	不能跨域LAN

lvs调度算法

Lvs的调度算法决定了如何在集群节点之间分布工作负荷。当director调度器收到来自客户端访问VIP上的集群服务的入站请求时，director调度器必须决定哪个集群节点应该处理请求。Director调度器的调度方法基本分为两类：

固定调度算法：rr，wrr，dh，sh

动态调度算法：wlc，lc，lblc，lblcr

算法	说明
rr	轮询算法，它将请求依次分配给不同的rs节点，也就是RS节点中均摊分 这种算法简单，但只适合于RS节点处理性能差不多的情况
wrr	加权轮训调度，它将依据不同RS的权值分配任务。权值较高的RS将优 任务，并且分配到的连接数将比权值低的RS更多。相同权值的RS得到相同 连接数。
Wlc	加权最小连接数调度，假设各台RS的全职依次为Wi，当前tcp连接数依 Ti，依次去Ti/Wi为最小的RS作为下一个分配的RS
Dh	目的地址哈希调度（destination hashing）以目的地址为关键字查找一 hash表来获得需要的RS
SH	源地址哈希调度（source hashing）以源地址为关键字查找一个静态ha 来获得需要的RS
Lc	最小连接数调度（least-connection）,IPVS表存储了所有活动的连接。 比较将连接请求发送到当前连接最少的RS.
Lblc	基于地址的最小连接数调度（locality-based least-connection）：将来 个目的地址的请求分配给同一台RS，此时这台服务器是尚未满负荷的。否则 这个请求分配给连接数最小的RS，并以它作为下一次分配的首先考虑。

原文链接：LVS 原理简介