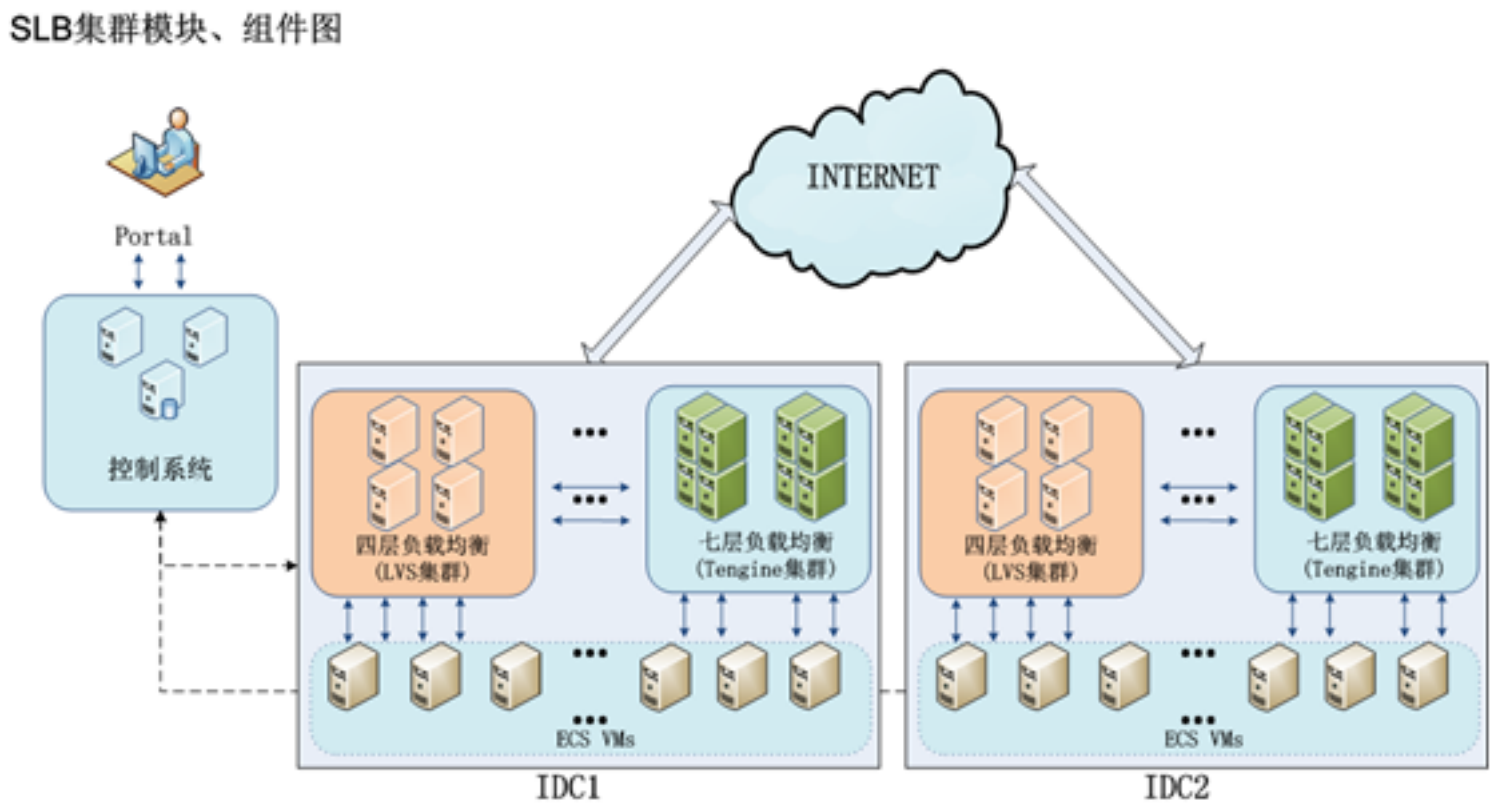
吴佳明(普空)：LVS在大规模网络环境中的应用

9月13日，阿里云课堂第一期在北京准时开课，到场与会人员众多，现场气氛非常热烈。阿里云三位讲师为大家献上了精彩的演讲，参会者也纷纷积极参与现场互动，通过问答交流，收获颇丰。应广大用户要求，我们将云课堂讲师现场分享内容全文整理出来，供大家参考。阿里云课堂会继续在全国各地陆续开课，欢迎大家继续支持！

1、 SLB总体架构

LVS本身是开源的，我们对它进行了多方面的改进，并且也已开源-https://github.com/alibaba/LVS。



接下来我们看一下LVS在整个SLB中的位置在哪里，整个图是SLB的架构图。SLB功能比较简单，主要是做负载均衡，最主要两个模块，一个是四层的负载均衡-LVS,还有七层的负载均-tengine，两个软件都是开源的；后端挂的是ECS。

一般来说，一个业务部署在两台或者两台以上的ECS-VM上面，建议大家选用SLB做负载均衡。

无论是LVS-四层也好，Tengine七层也好，我们负载均衡都是集群，都会有

冗余的，一台宕机了对用户来说没有影响。SLB在杭州region内也有很多IDC-数据中心，同一个VIP可以在IDC1和IDC2，一旦IDC1宕了就切换到IDC2，即实现IDC间的冗余。对可靠性要求特别高的业务，建议在ECS两个可用区里部署VM，这样一个IDC宕了也会有冗余。

另外还有我们SLB整个可用性为5个9，为什么我们做IDC冗余，据说国外最好数据中心的可用性是5个9，SLB位于数据中心中，必须靠数据中心之间的冗余做到5个9。

2、 LVS历史

LVS是章文嵩博士1998年做的，LVS是Linux虚拟服务器的简称；

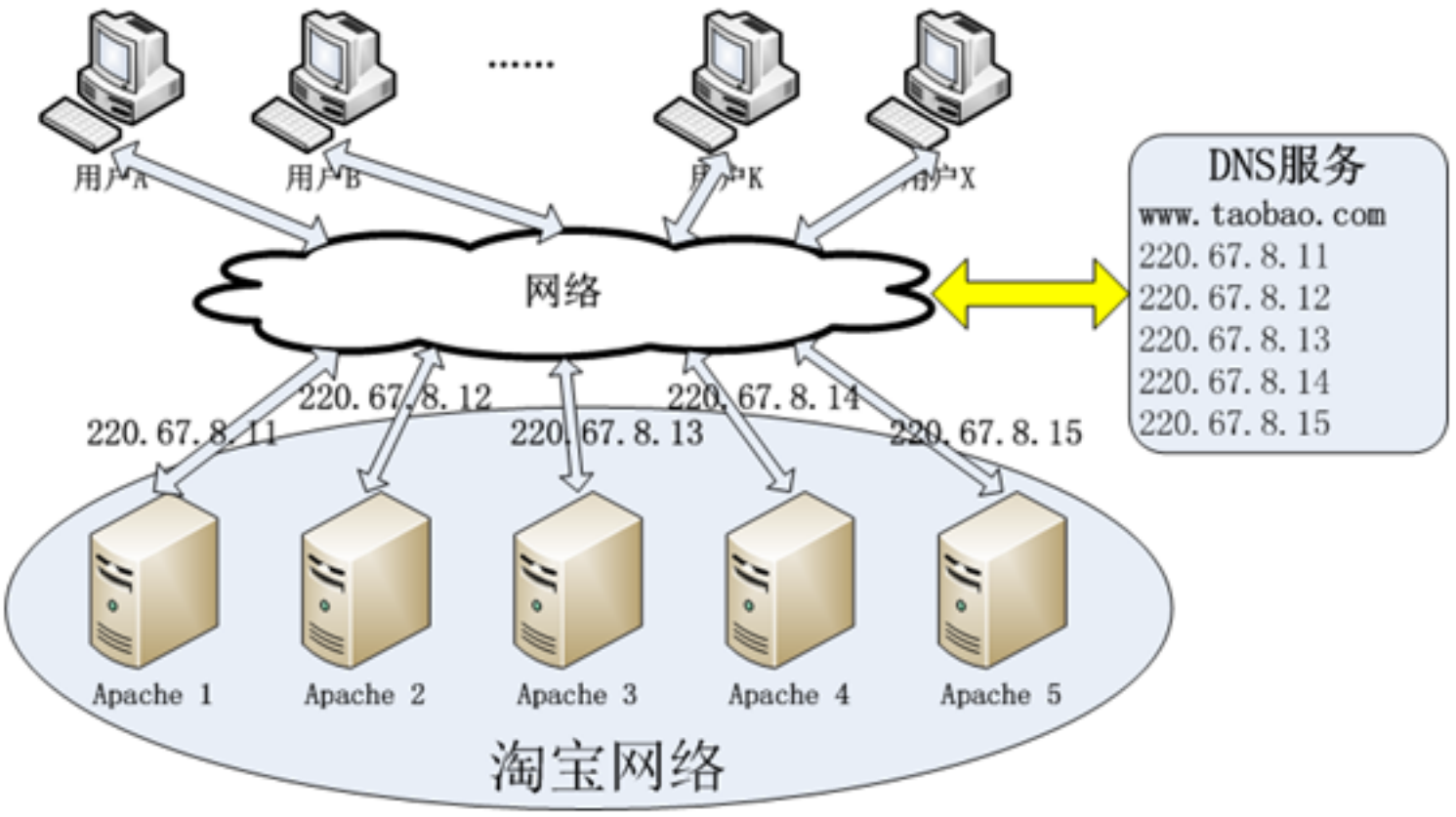
章文嵩当前是阿里云技术负责人。

3、本次LVS分享主要内容

本次分享的内容如下：为什么引入LVS？在大规模网络下用的时候存在哪些问题？针对这些问题，我们做了一些改进：FULLNAT，SYNPROXY，集群部署；接下来，介绍LVS性能优化的一些技术，这些技术不仅仅用在LVS，大家可以用在你们自己网络业务里面；最后介绍一下我们接下来LVS做哪些事情。

4、 LVS-why

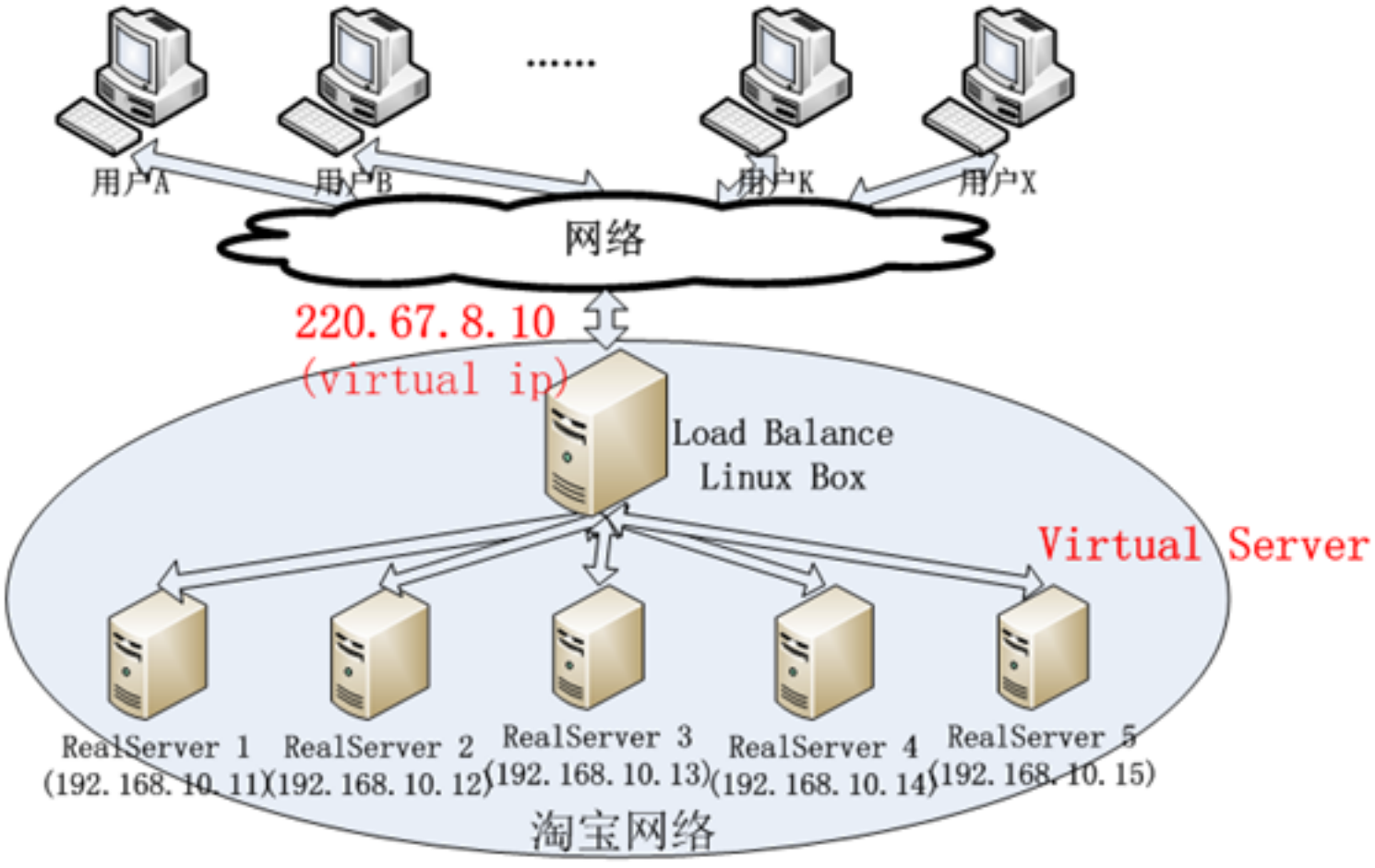
比如说，一个用户访问淘宝网站，淘宝网前端共有5台apache服务器，如何决定访问哪一台apache？常用的方式是用DNS做负载均衡，将5台apache服务器的ip地址添加到域名www.taobao.com中。



但DNS有一些缺点，第一个缺点：例如第二台apache宕了，运维赶紧把DNS中该apache的ip地址删除掉，但 很多地方的local DNS不一定遵守TTL协议，这样删除操作什么时候生效，你根本不可控的；尤其移动网络中，这个问题更突出，我记得10年时移动网络部分地区local DNS一天才更新。

第二个缺点：服务调度算法只支持WRR。如果你用户范围很有限，就会有负载不均衡的问题。第三个缺点：攻击防御能力很弱，每次有攻击靠一台机器抗。

针对DNS的不足，引入了Virtual Server的概念，即最前端有一个入口设备把流量均衡到后端的apache上去；无论是LVS软负载 还是 F5硬负载均衡 也都是这种概念。



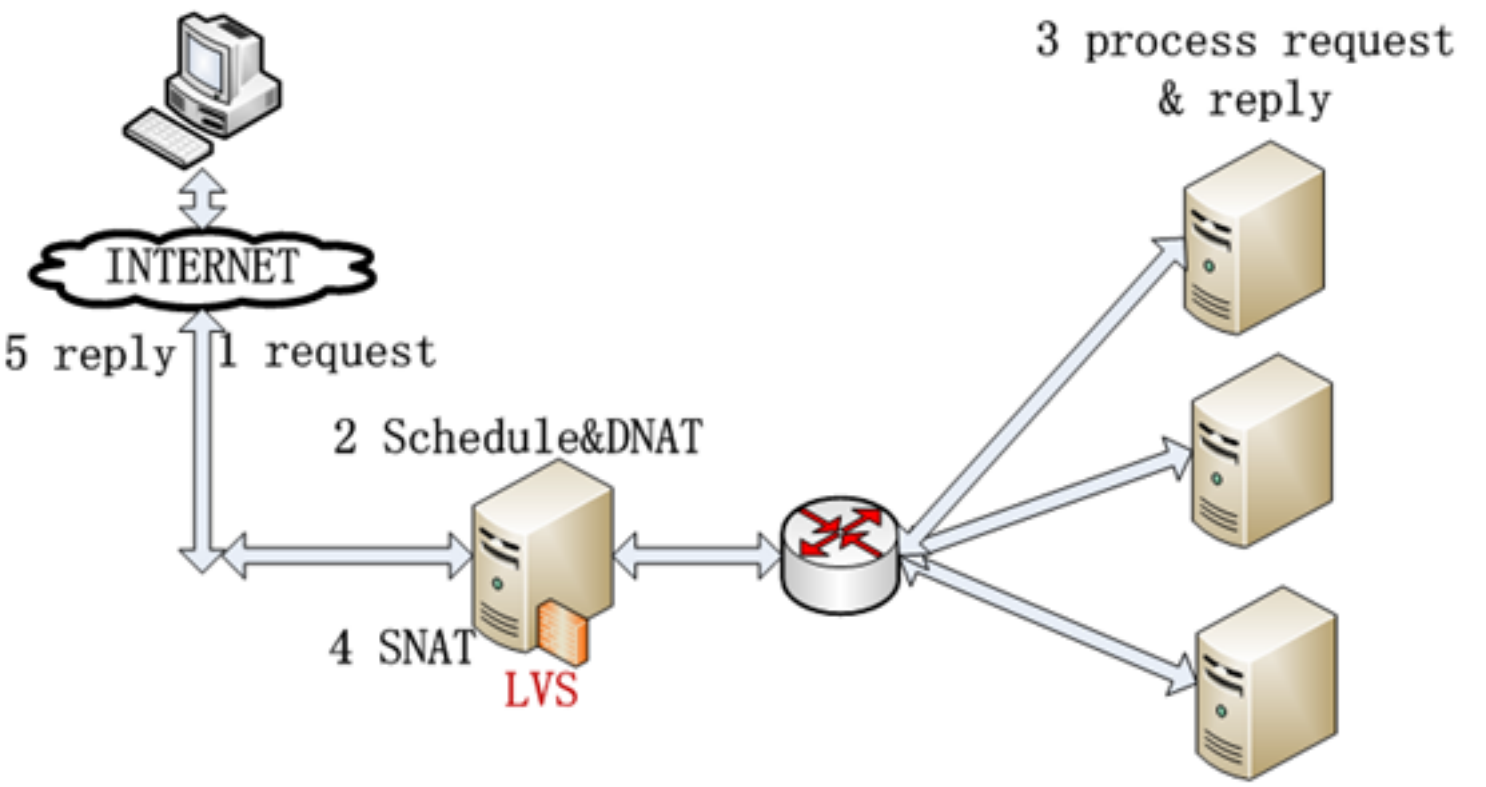
5、LVS-what

LVS的基本概念，是4层load balance，这个4层对于着OSI网络模型中的传输层，需要用到端口信息。

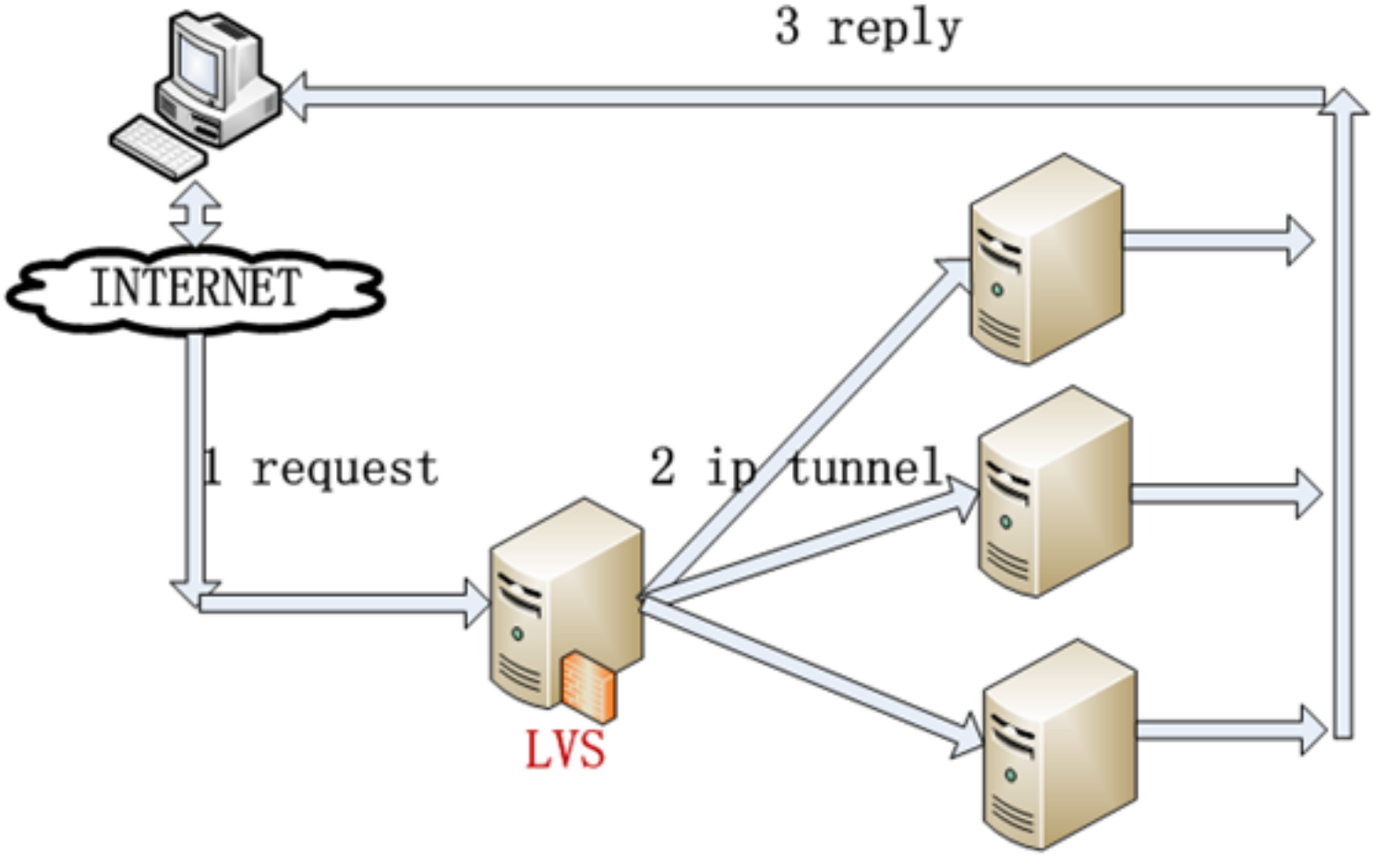
LVS支持WRR、WLC等调度算法；WRR是带权重的轮询；WLC是带权重的最小连接调度策略，即将请求往最少连接的后端服务器上调。

LVS支持3种工作模式：NAT、DR、TUNNEL，这几种模式跟你IDC网络部署方式有关系的。

传输协议支持TCP、UDP两种。



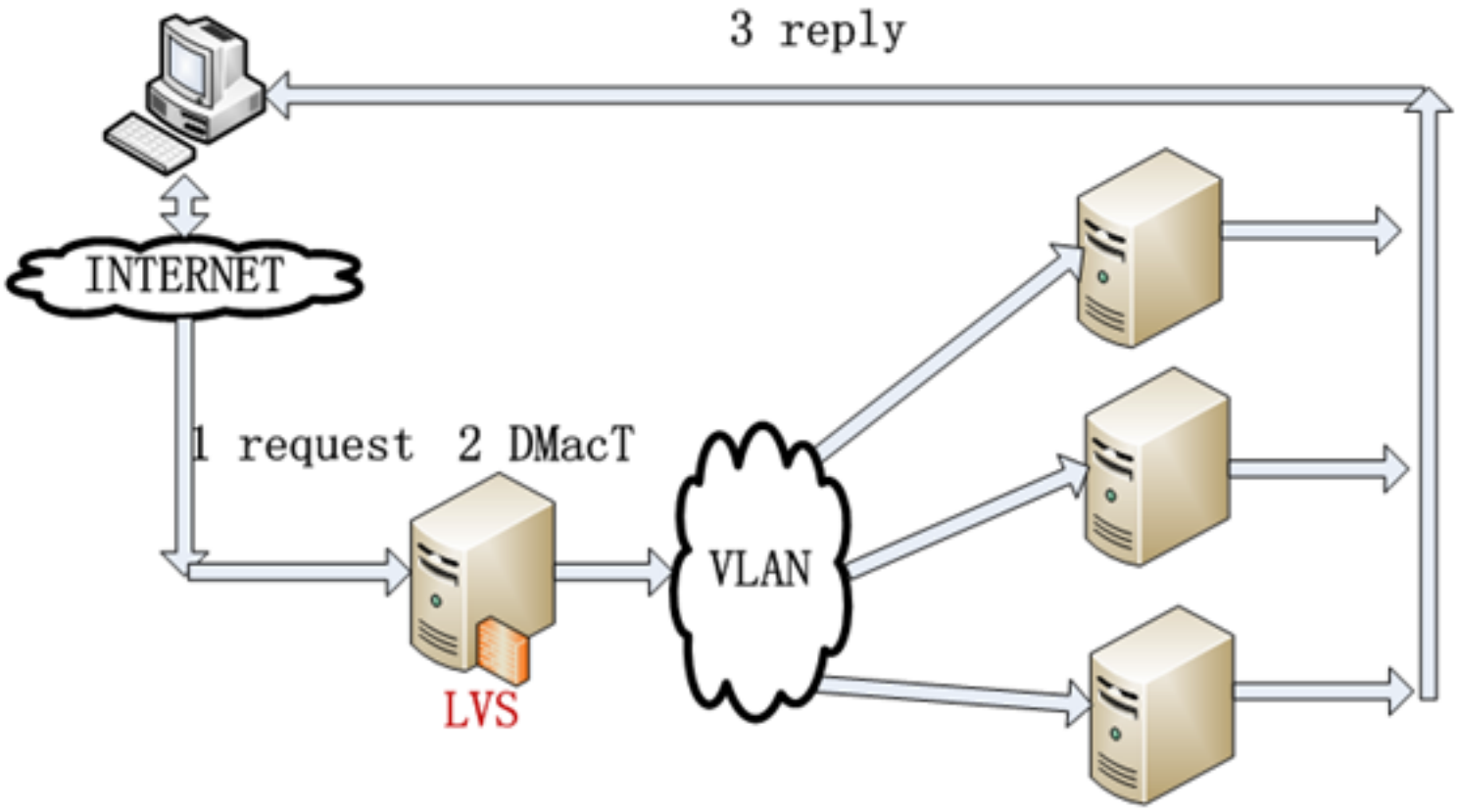
第一种是NAT模式，进来和出去的数据流都是经过LVS设备。进来的时候把目的IP改成实际后端服务器的IP-DNAT，出去的时候则做SNAT。一般买的F5等商用设备，都采用NAT模式，因为NAT模式可以防DDOS攻击，该攻击防御功能依赖于进出数据都通过设备。



第二种是TUNNEL，这个是进的流量经过LVS，出去的时候不经过了。TUNNEL是在原来IP头部再新增封装一个IP。据说，腾讯采用IP隧道的模式；TUNNEL模式的最大问题是 每个数据包都需要增加一个IP头部，如果收到的数据包是已经达到以太网帧最大长度1.5K，IP头就添不进去。这时候常用做法是会回一个因MTU导致目的不可达的ICMP包给客户端，客户端如

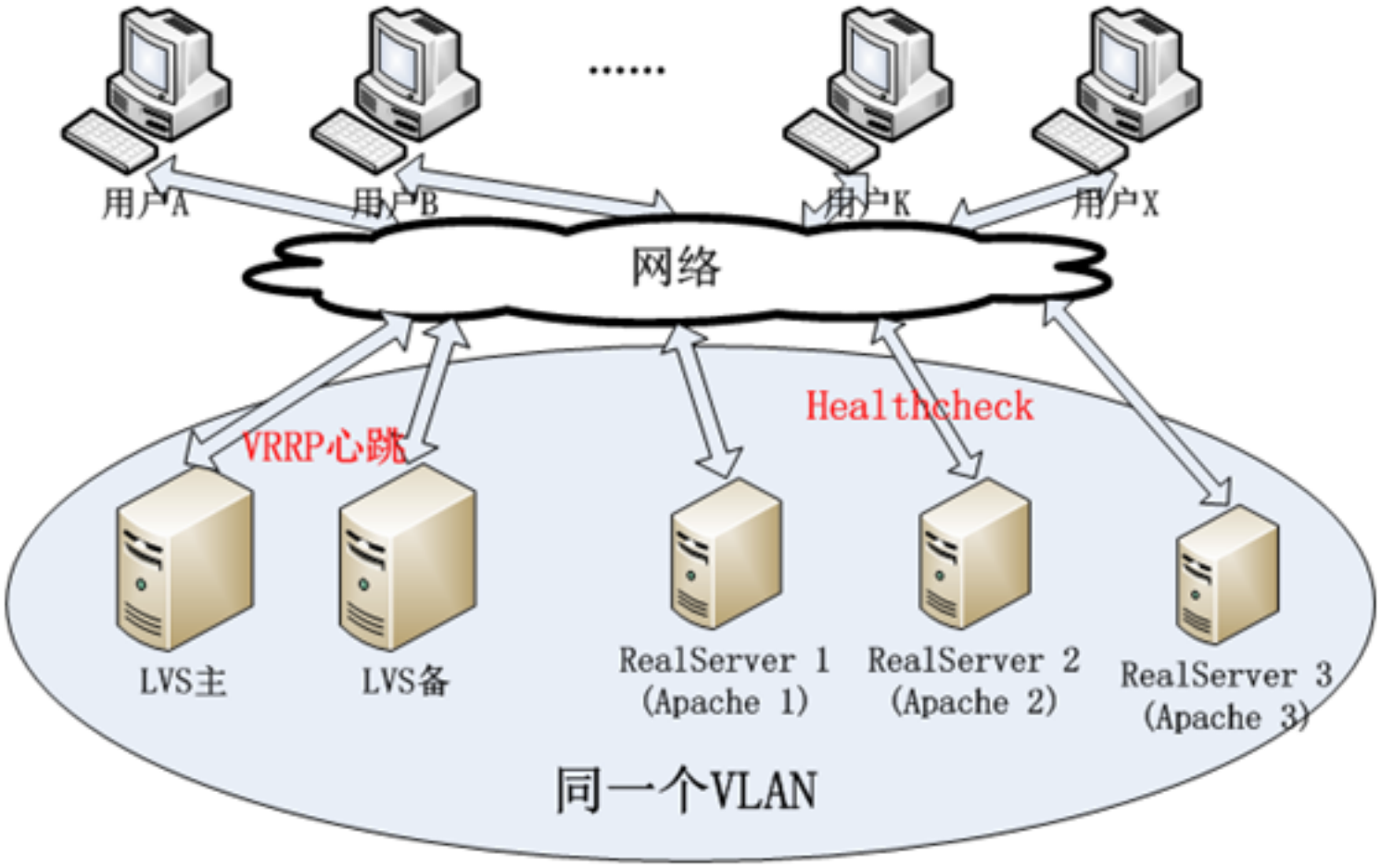
果支持PMTU的话，就会把大包分成小包发过来。

解决上述问题的一个办法是 交换机开启巨帧。另一个方法是将后端服务器上的MSS改小，我一个IP头是20个字节，默认MSS是1460，将其改成1440可不可以？可以，大部分用户可以正常支持，但是总会存在百万份之几，它不支持的标准MSS协商协议，你即使将MSS调的很小，但是客户端还是会发一个大包出来。



第三种是DR，DR的性能是所有模式中最高的，它只需要修改目的MAC；但部署上必须要求LVS和后端服务器在同一个VLAN中。

DR非常适合小规模网络，比如，阿里的CDN都是用的DR模式，几十台服务器的规模，特别适合DR这种高效的模式；因此，如果你业务规模比较小的话，建议采用DR。



6、LVS-应用

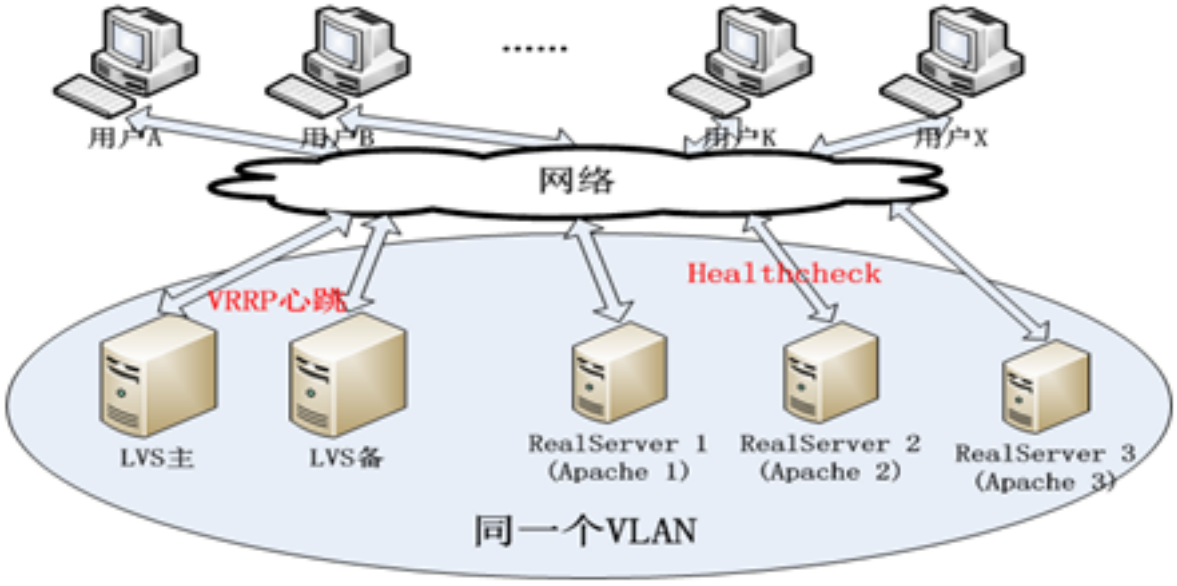
前面我们讲了LVS基本的特征。LVS本身只是一个内核模块：IP_VS，这个模块是做负载均衡，你只用这个模块来做工程应用是远远不够的。比如，一台RealServer宕机了怎么办？LVS本身宕机了着呢麽办?

针对上述问题，我们需要有辅助软件帮我们管理LVS，一般现在常用的是Keepalived；keepalived支持健康检测，4层和7层健康检检测，以解决RealServer宕机问题。

另外，keepalived支持VRRP心跳协议，可以实现LVS主备冗余，以解决LVS本身单点故障。

最后，keepalived支持配置文件的方式来管理LVS;

完成了上述工作，我们还缺少一个监控-服务运行怎么样，流量怎么样，CPU负载怎么样？大部分公司都有自己一套监控系统，LVS监控基本上都是集成到自己监控系统里面去。当然也可以用开源的组件，比如，SNMP Patch-可以跟传统网络一样接口获得LVS的信息。



该图是我讲到CDN网络拓扑，LVS两台实现主备冗余，同时对后端RealServer做Healthchech。

7、 LVS-问题 & 解决

前面介绍了官方LVS的一些基础知识；

但在大规模的网络下，在淘宝的业务中，官方LVS满足不了需求；原因有3点，

1) 刚才讲三种转发模式，部署成本比较高；

2) 和商用的负载均衡比，LVS没有DDOS防御攻击功能；

3) 主备部署模式，性能无法扩展；一个VIP下的流量特别大怎么办?

第一点- LVS转发模式的不足，下面来展开描述一下；

DR的不足：必须要求LVS跟后端所有的REPLY放在同一个VLAN里。当然有人会提出来分几个区，每个区布一个LVS，但一个区VM资源没有了，就只能用其它区的VM，而用户需要这些VM挂到同一个VIP下，这是无法实现的。

NAT的不足：NAT最主要问题就是你配置处理很复杂；阿里原来买的商业设备的时候，需要在交换机上配策略路由，OUT方向的策略路由；因为，冗余考虑会部署多套负载均衡，走默认路由只能到达一套负载均衡。

TUNNEL的不足：隧道的问题也是配置较复杂，RealServer需要加载一个IPIP模块，同时做一些配置。

针对上述问题，我们的解决方法如下：

| LVS各转发模式运维成本高

- 新转发模式FULLNAT：实现LVS-RealServer间跨vlan通讯，并且in/out流都经过LVS；

| 缺少攻击防御模块

- SYNPROXY：synflood攻击防御模块

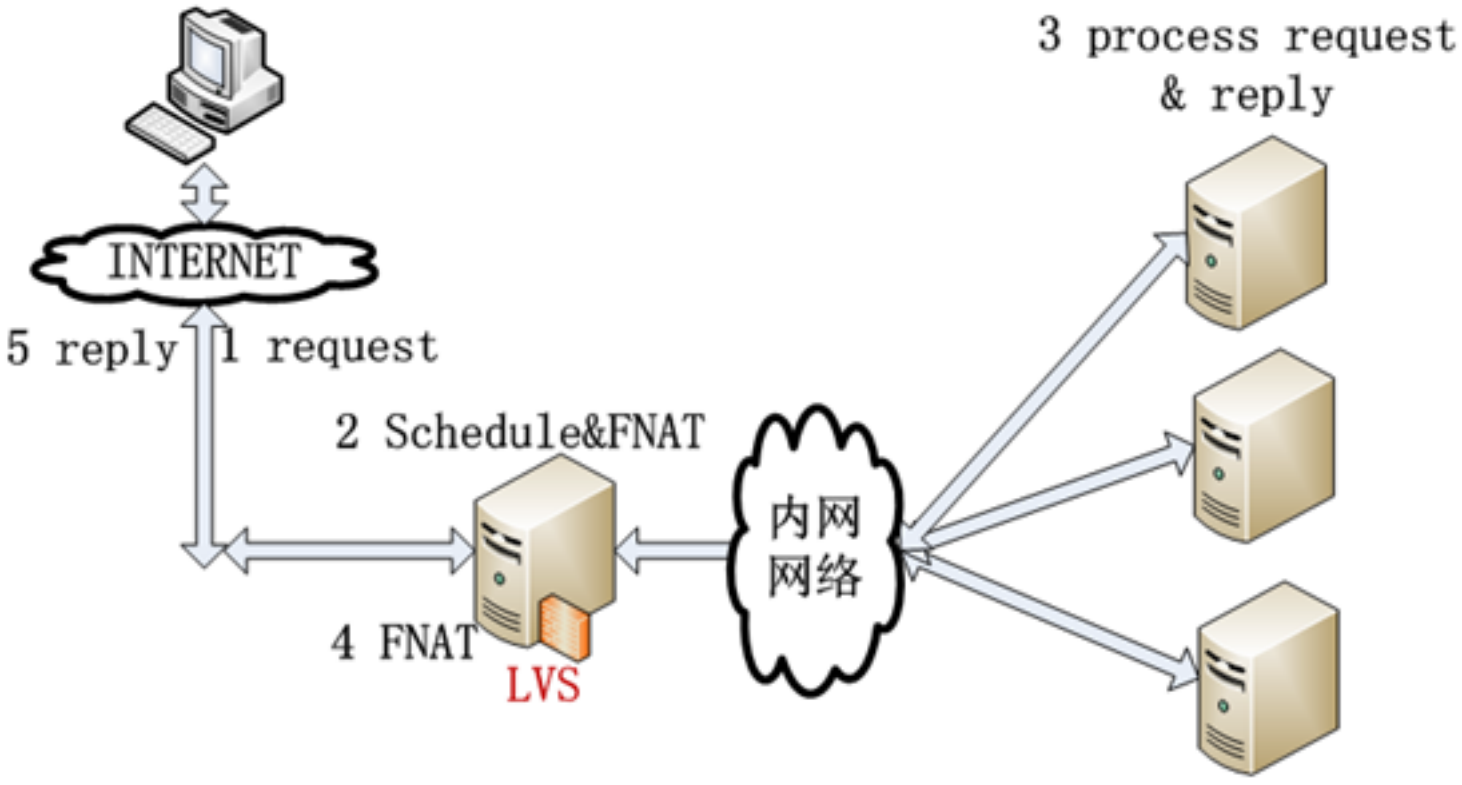
- 其它TCP FLAG DDOS攻击防御策略

| 性能无法线性扩展

- Cluster部署模式

下面我们分别介绍上述解决方法；

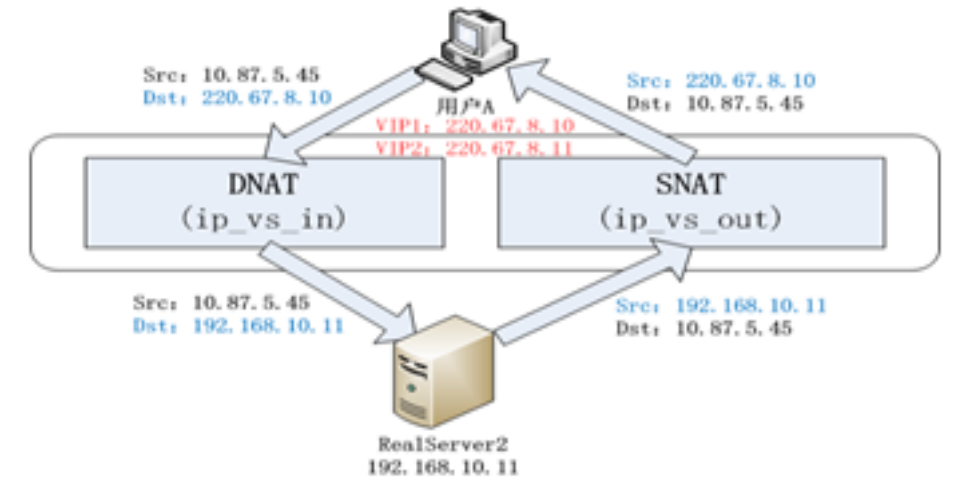
8、LVS-FULLNAT转发模式

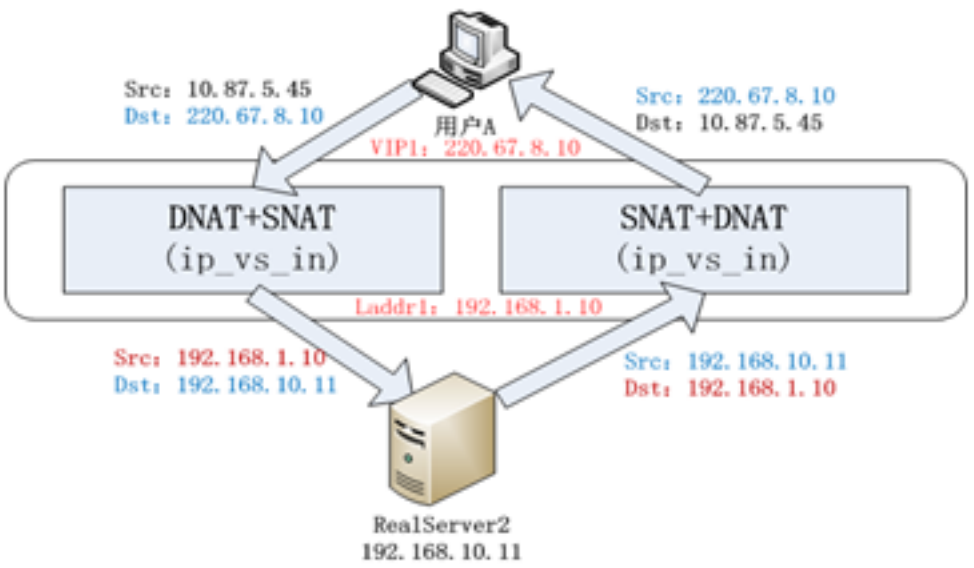


下面讲讲FullNAT，FULLNAT转发数据包是类似NAT模式，IN和OUT数据包都是经过LVS；唯一的区别：后端RealServer 或者 交换机 不需要做任何配置。

FULLNAT的主要原理是引入local address（内网ip地址），cip-vip转换为lip->rip，而 lip和rip均为IDC内网ip，可以跨vlan通讯；

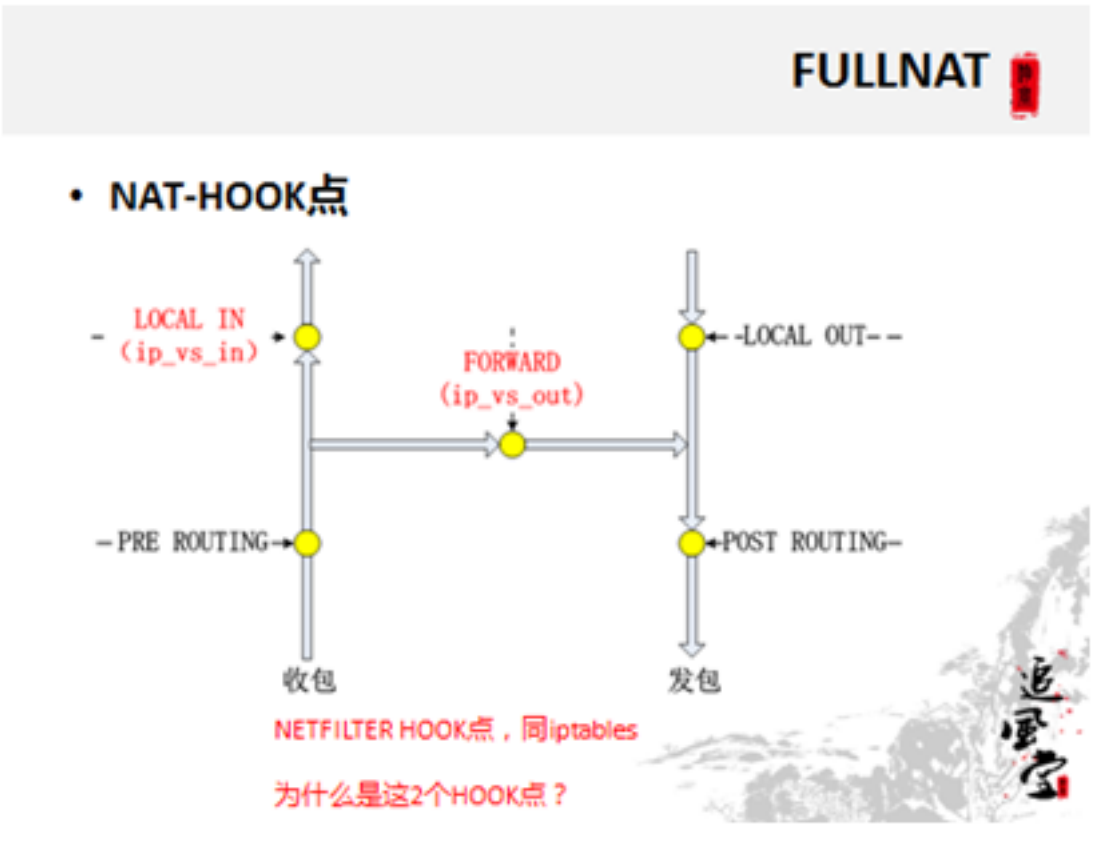
下面从IP地址转换的角度看一下，NAT和FULLNAT的区别；

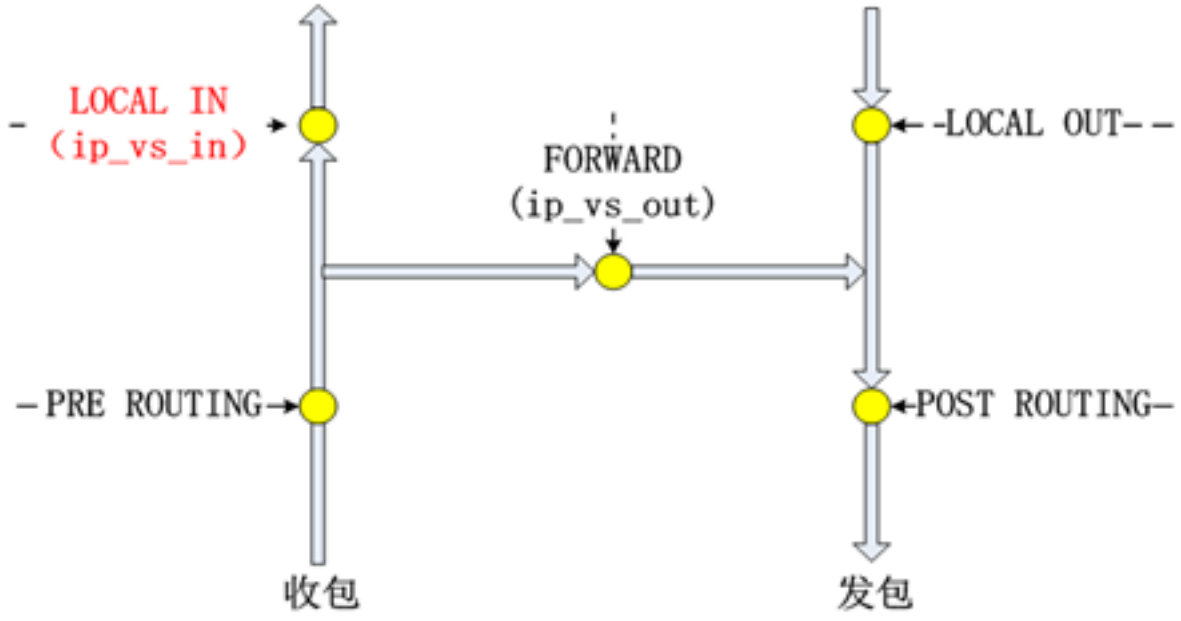




如图所示，相比NAT模式，FullNAT多了一个Local IP，IP地址转换时，源和目的IP都改了，即SNAT DNAT。

FULLNAT模式下，ipvs在NETFLITER框架的HOOK点也发生了变化；

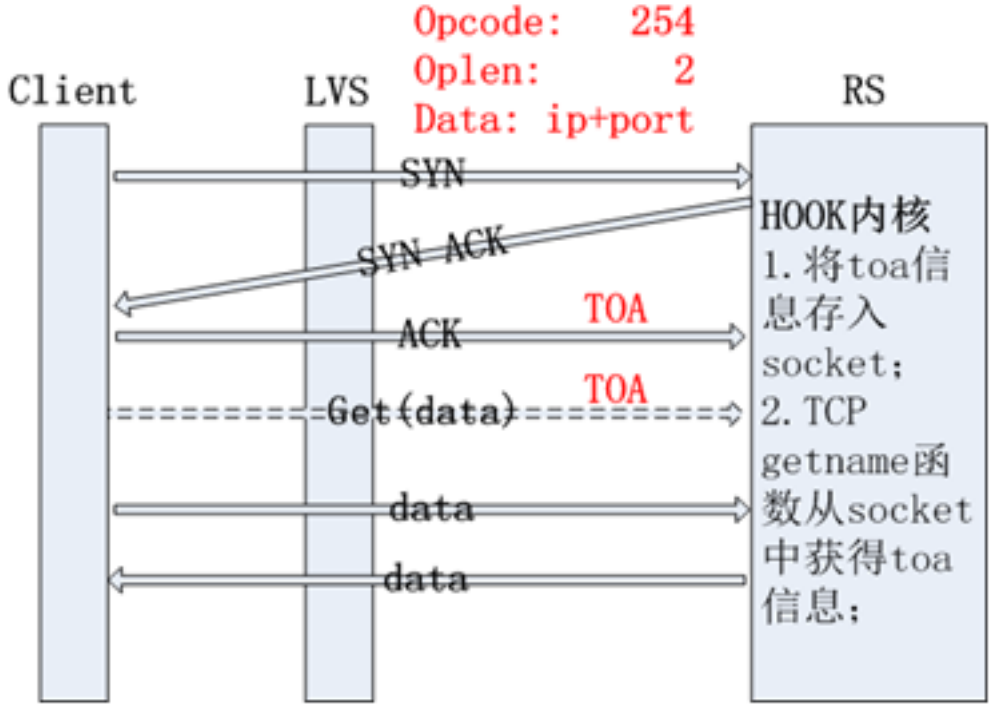




这个图就是内核NETFILTER的五个HOOK点；原来传统的NAT模式，在LOCAL_IN和FORWARD两个点，而FullNAT模式下，IN/OUT方向的目的IP都是LVS上的IP，因此，只能在LOCAL_IN这个点。

相比NAT，session表管理也发生了变化，有1个索引表，变成了IN和OUT 2个; 这是因为NAT模式只需要用client address作为hash key，而FULLNAT只能用5元组；

8字节



FULLNAT一个最大的问题是：RealServer无法获得用户IP；为了解决这个问题我们提出了TOA的概念，主要原理是：将client address放到了TCP Option里面带给后端RealServer，RealServer上通过toa内核模块hack了getname函数，给用户态返回TCP Option中的client ip。

题外话，全球最大CDN厂商阿克曼也用了TCP Option携带附属信息；

下面来介绍一下FULLNAT开发时，遇到的几个坑，这几个坑对Linux网络应

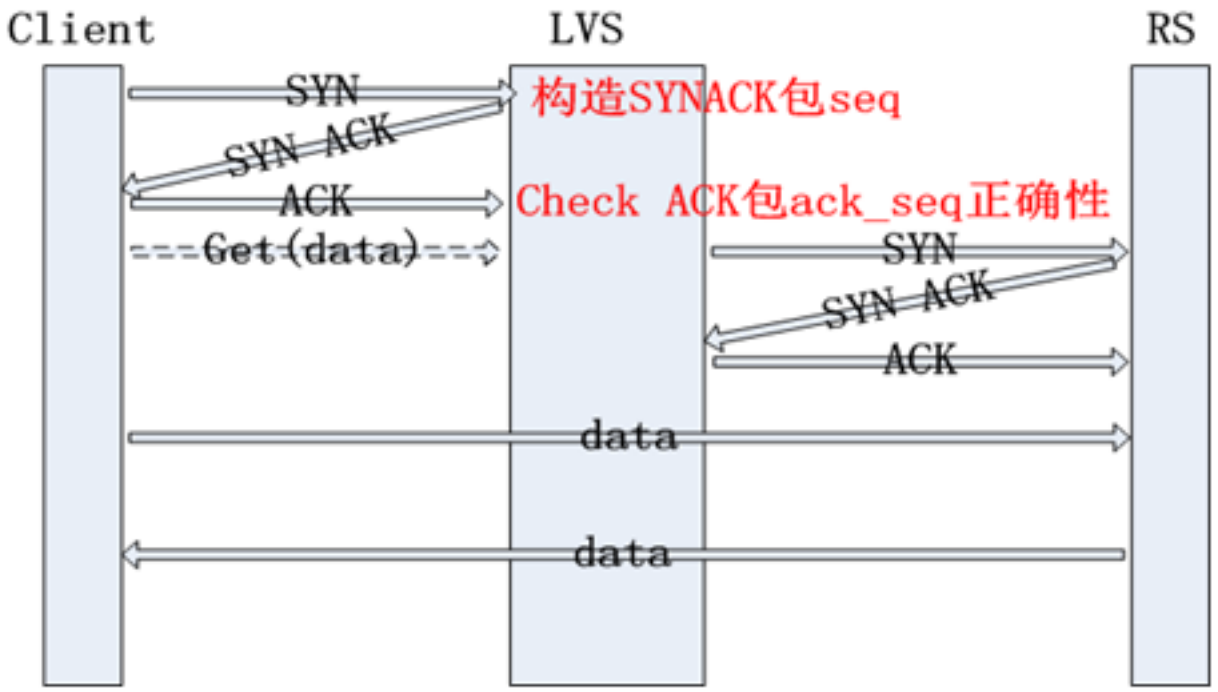
用开发也是有用的；比如，Realserver kernel开启tcp_tw_recycle，该参数开启会导致部分NAT网关出来的用户访问失败；

9、LVS-SYNPROXY

LVS可以防御DDOS 4层标志位攻击，其中，synproxy是用于防御synflood攻击的模块；

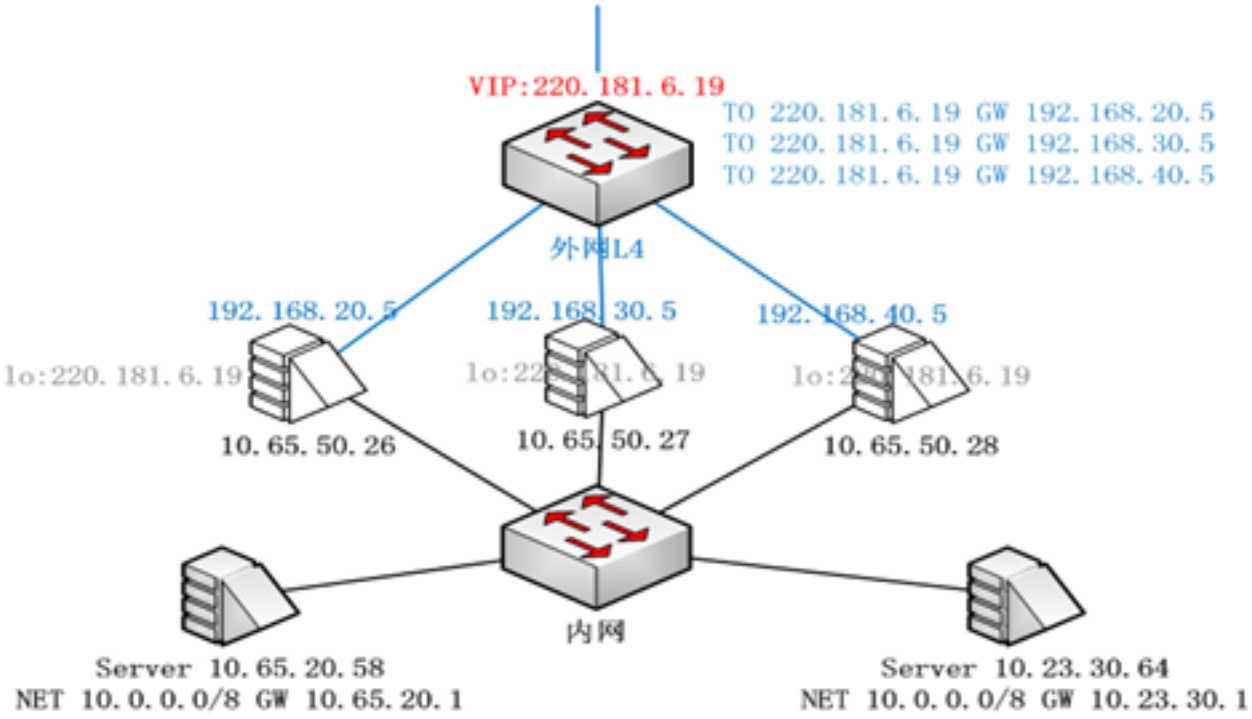
Synproxy实现的主要原理：参照linux tcp协议栈中syncookies的思想，LVS-构造特殊seq的synack包，验证ack包中ack_seq是否合法-实现了TCP三次握手代理；

简化一点说，就是client和LVS间建立3次握手，成功后，LVS再和RS建立3次握手；



10、LVS-集群

前面我们介绍了引入一种新的模式叫做FullNAT，方便大家部署，下面我们可以有集群部署模式做横向扩展。



谁把流量均衡的分到各台LVS上-交换机，LVS和交换机间运行OSPF协议，交换机上生成该VIP的等价路由-ECMP；

另外，这种部署方式很多地方可以用，比如说DNS，域名系统我们每个公司都需要，部署DNS系统的时候，不建议再去加一层LVS，而是DNS服务器直接和交换机跑OSPF协议。

交换机ECMP有一个问题：当前是不支持一致性哈希算法；比如，三台LVS有一台宕了，等价路由变成两条，你数据包全都乱掉了。像思科的交换机芯片它也支持了类似一致性hash的算法，但具有该功能的交换机还没有产品化出来；因此，我们不得不在各台LVS做session同步，即把连接表做成全局的，这样即使有一台LVS宕了也没有关系，因为表是全局的，这样请求过来其它LVS还可以正确地往后转发。

注：当前FullNAT模式没法支持同步的。

11、LVS-性能优化

这些性能优化方法对大家网络服务也是有用的。

第一点，多队列网卡，即一个队列绑定到一个CPU核上，让多核同时处理网络数据包。如果网卡不支持多队列，可以用google提供的软多队列-RPS，linux内核默认已经集成；

第二，对keepalived进行了优化，主要将网络模式从select改为了epool。

第三，大家如果自己买的服务器的话，建议把网卡LRO、GRO功能关掉，

尤其是broadcom的网卡，我们踩过很多坑。

12、 LVS-todo list

接下来我介绍一下我们下一步要做的事情。

我们接下来重点在：控制系统；LVS在五六月份的时候出现一系列的故障，很不稳定，其实不是LVS，也不是Tengine，而是控制系统的问题；我们第一步将控制系统做了精简，将用户操作逻辑和运维逻辑进行了分离；下一步重点是提高控制系统性能。

功能上，我们会支持UDP和HTTPS。

还有Session同步，FullNAT情况下很难支持Session同步的，这个问题我们也会解决。

后面性能我们也在尝试英特40G的网卡，我们也在评测看看。

我们未来如果做的可以的话，我们希望把4层7层做到一个里面去。

原文地址：http://blog.aliyun.com/1750 （有视频）