

LVS在淘宝环境中的应用

吴佳明_普空 核心系统部
关注网络技术

个人简介



吴佳明_普空——核心系统研发

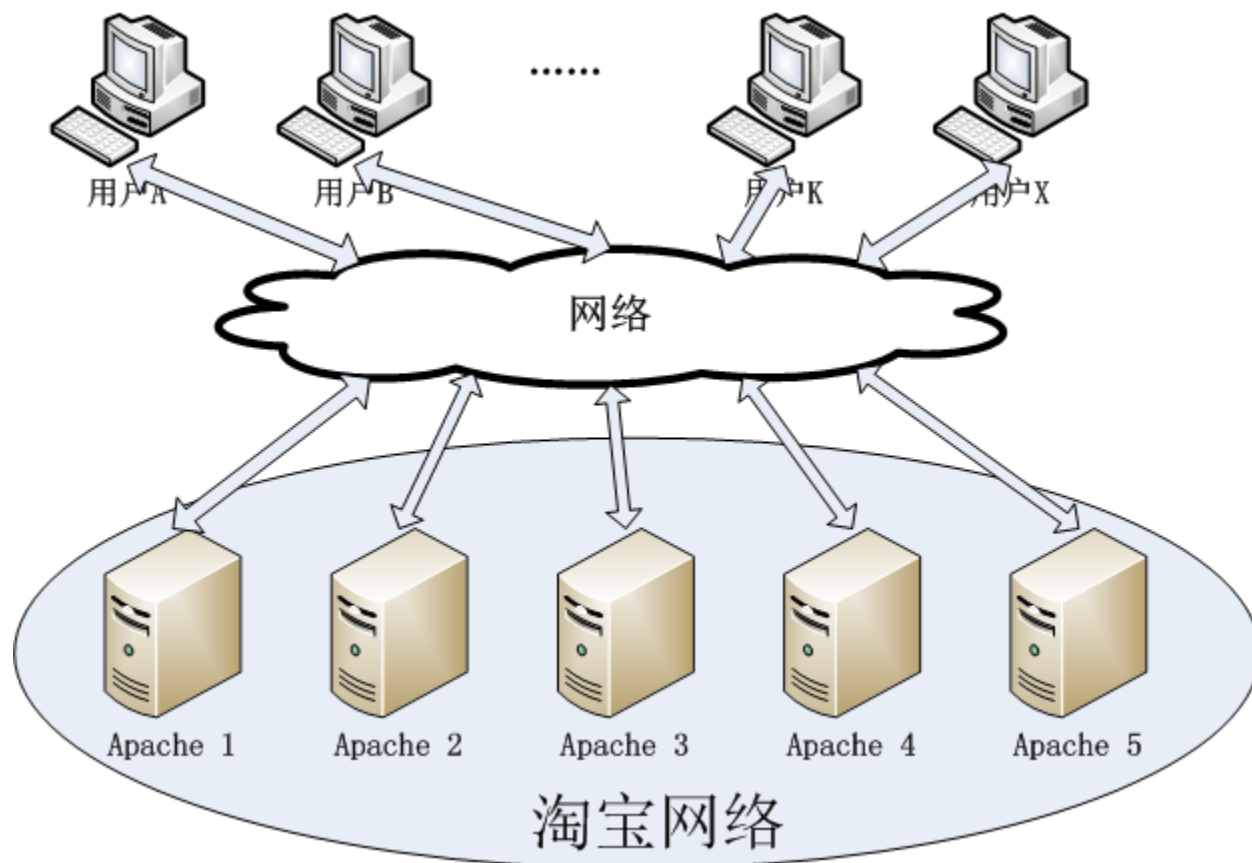
一直从事系统网络相关技术研发，包括
IDC网络、内核TCP/IP协议、4/7层负载均衡、CDN、DDOS攻击防御等；

- 2007.4~2011.5 就职于 百度，资深系统工程师，完成 百度网络4层统一接入和接出；
- 2011.5~至今 就职于 淘宝，技术专家，从事 LVS 等网络技术研发；

主要内容

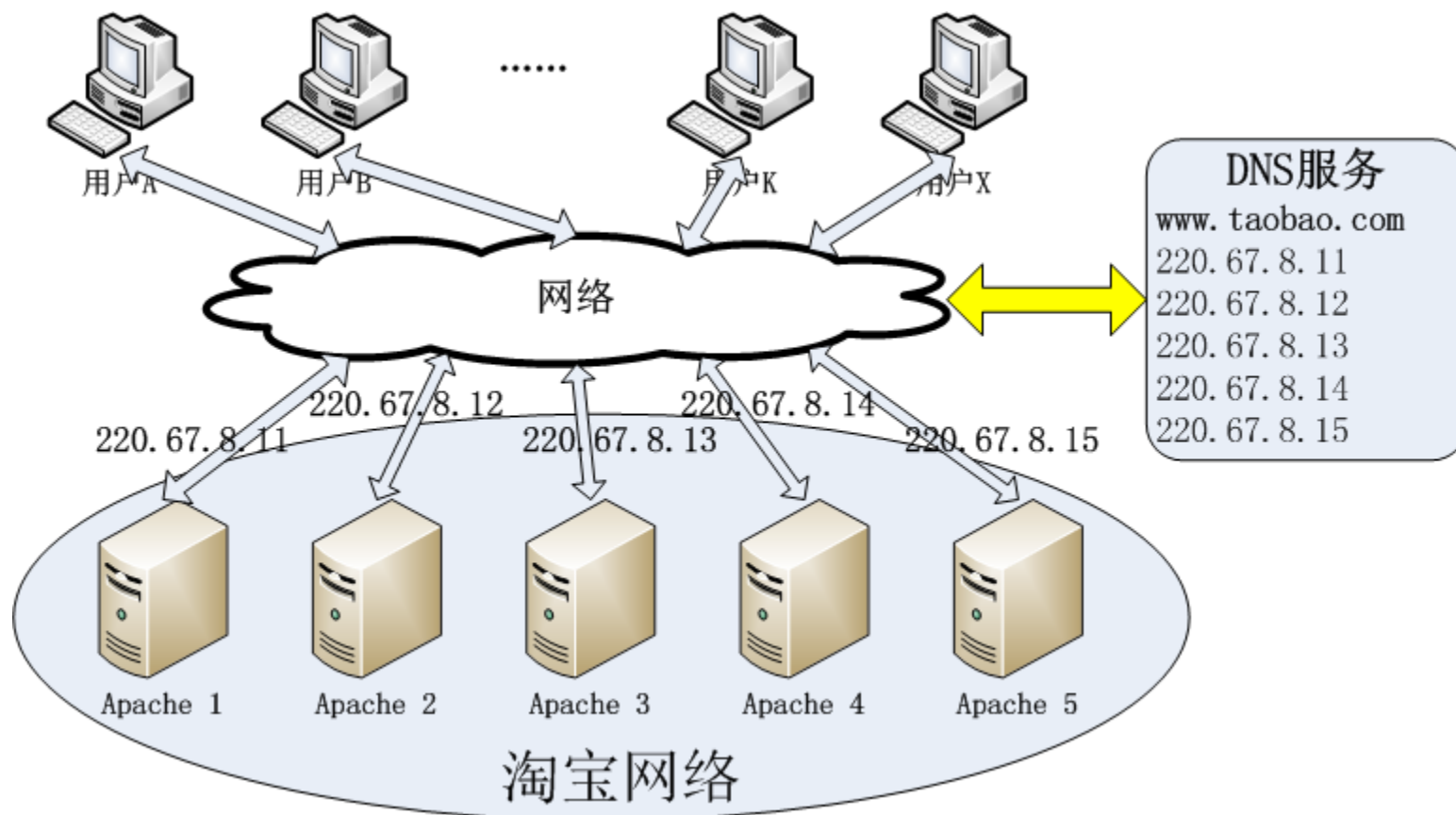
1. LVS-简介
2. LVS-问题
3. LVS-fullnat
4. LVS-synproxy
5. LVS-cluster
6. LVS-performance
7. LVS-todo list

简介-why



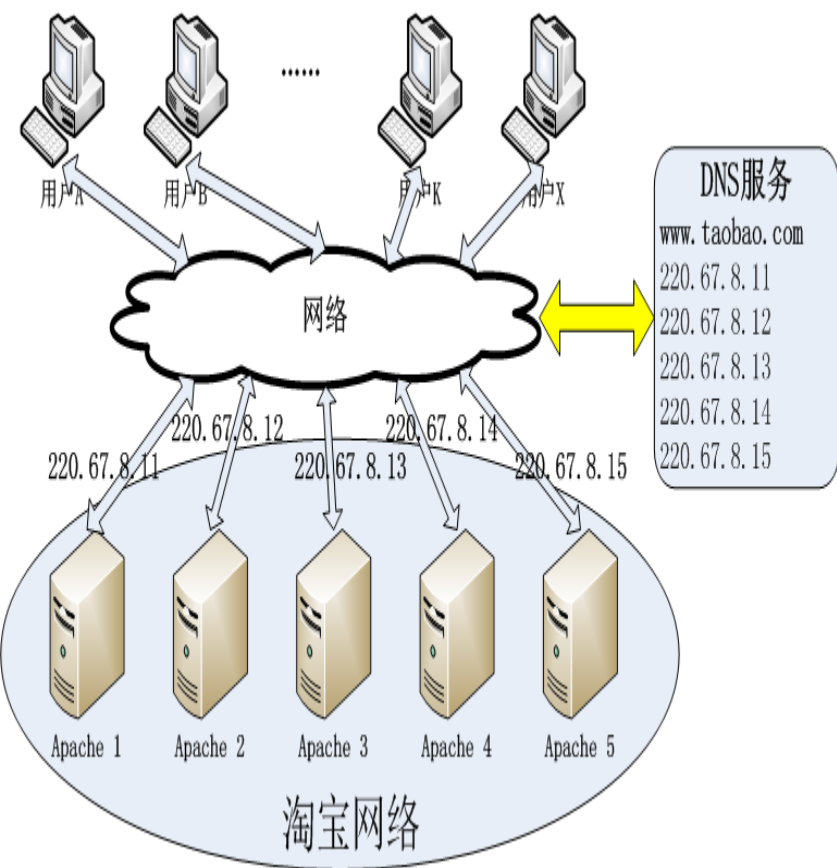
Q: 用户访问淘宝，如何决定访问哪一台Apache?

简介-why



A: 传统做法, DNS服务

简介-why



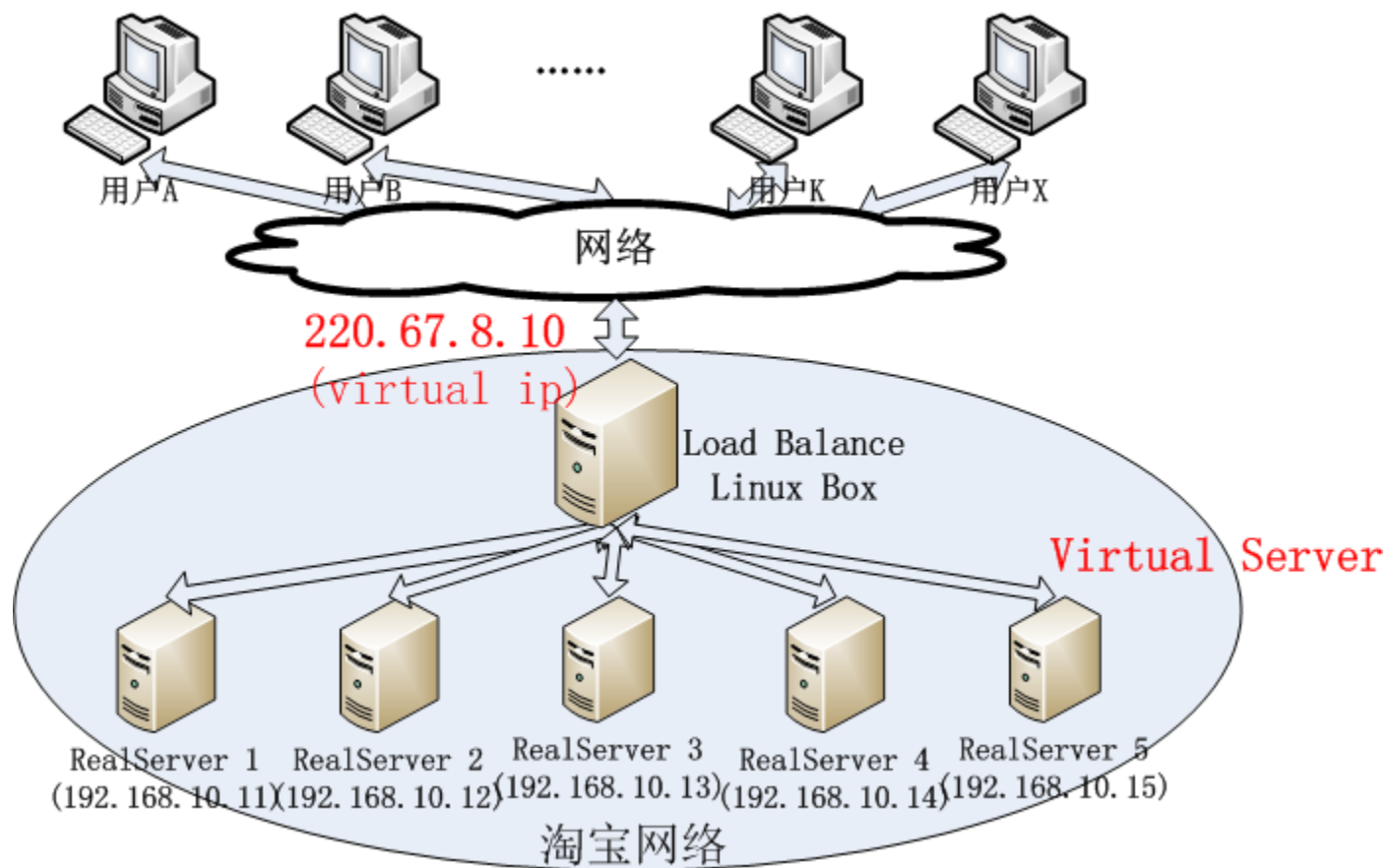
Q1: apache2 down, remove生效时间不可控

Q2: 只支持WRR的调度策略

Q3: apache间负载不均匀

Q4: 攻击防御能力弱

简介-why



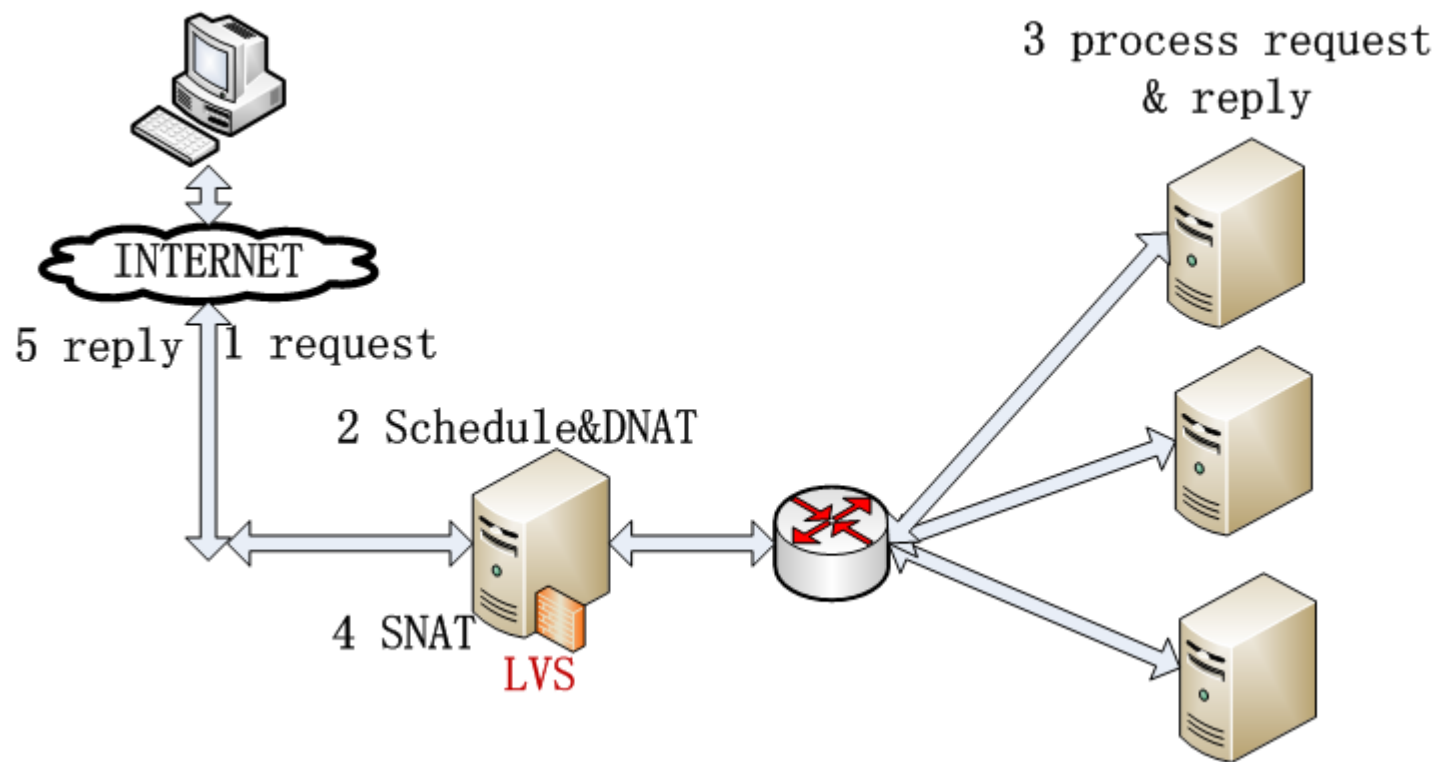
A: 引入Virtual Server

简介-what

- ▶ 4层Load Balance
 - 基于传输层信息进行 调度
 - 调度算法：WRR/WLC 等
 - 工作模式：NAT/DR/TUNNEL
 - 传输协议：TCP/UDP

简介-what

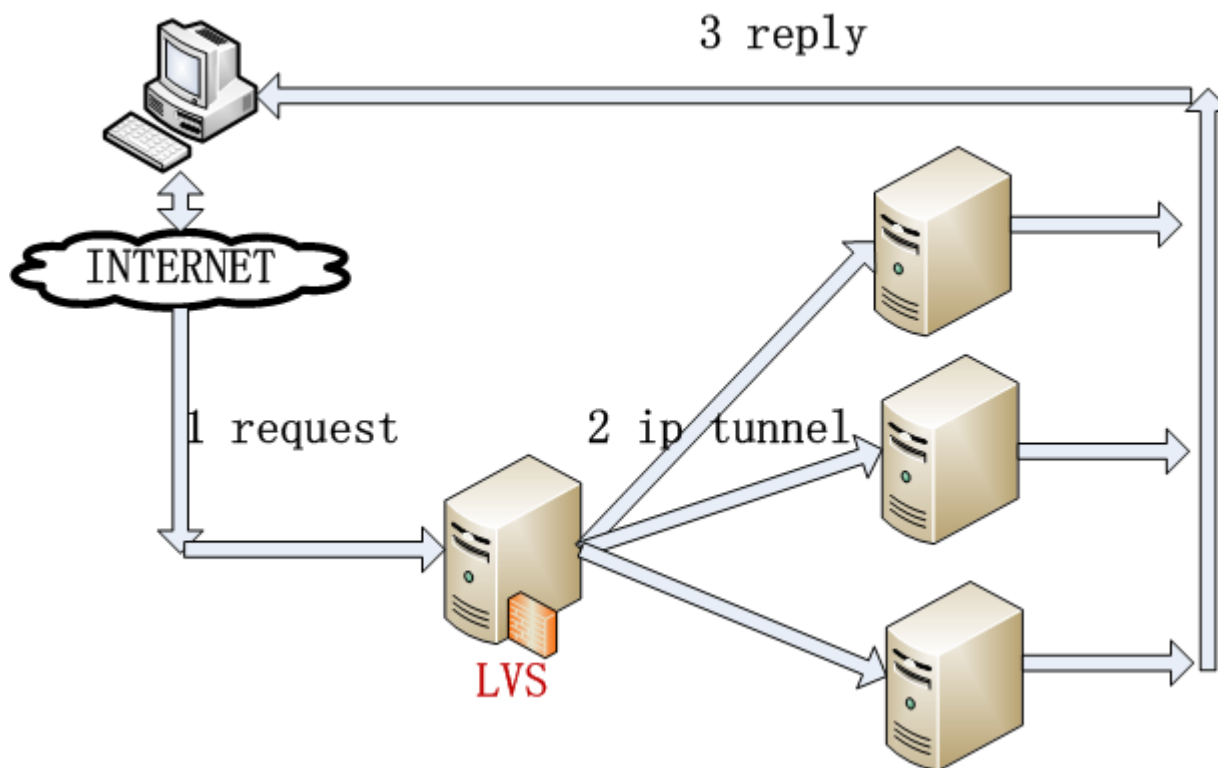
▶ NAT



IN(2): DNAT
OUT(4): SNAT

简介-what

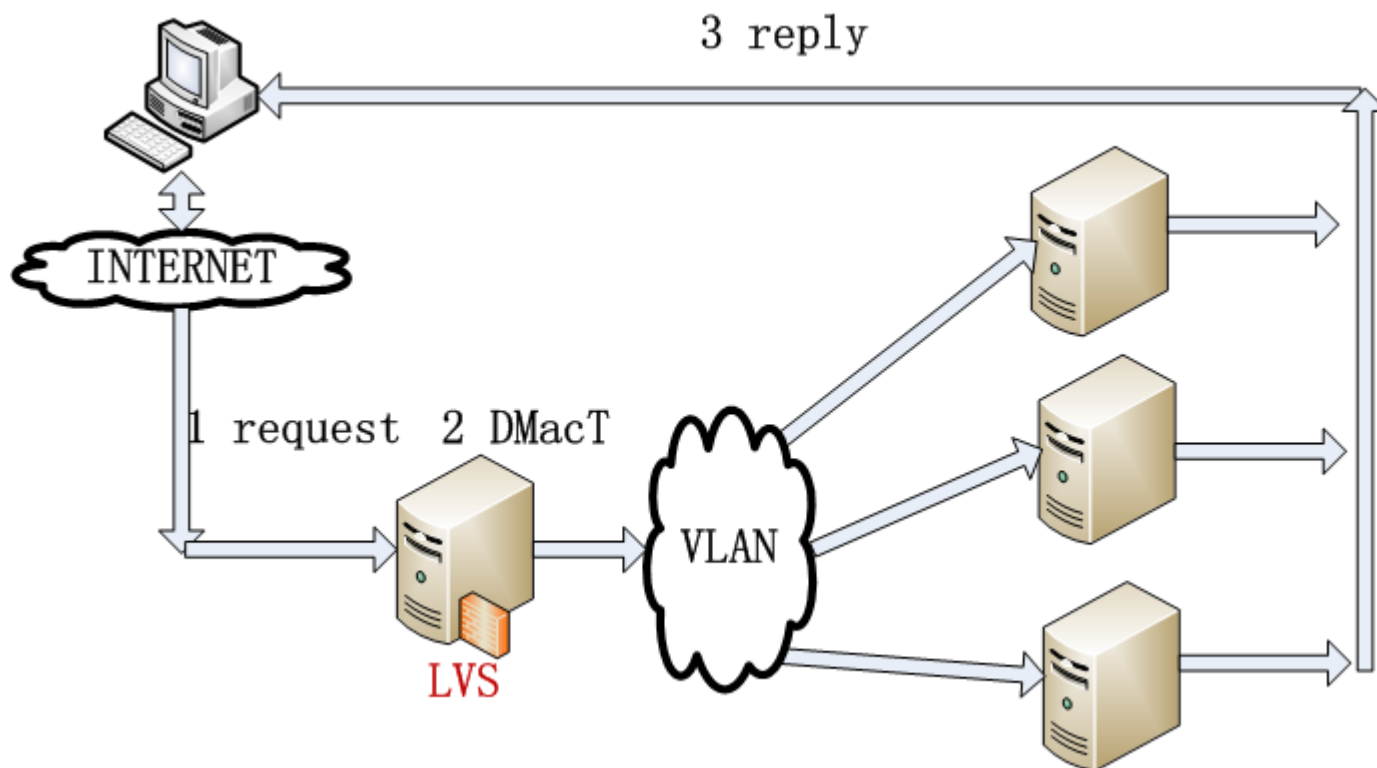
▶ TUNNEL



IN: 增加1个IP头
OUT: NULL

简介-what

► DR



IN: 更改目的MAC
OUT: NULL

简介-应用

▶ LVS

- 内核模块：ip_vs
- 实现了负载均衡

▶ Q

- 某台RealServer down了，怎么办？
- LVS本身down了，怎么办？

简介-应用

▶ A

- 某台RealServer down了，怎么办？ --- 健康检测
- LVS本身down了，怎么办？ ---LVS冗余

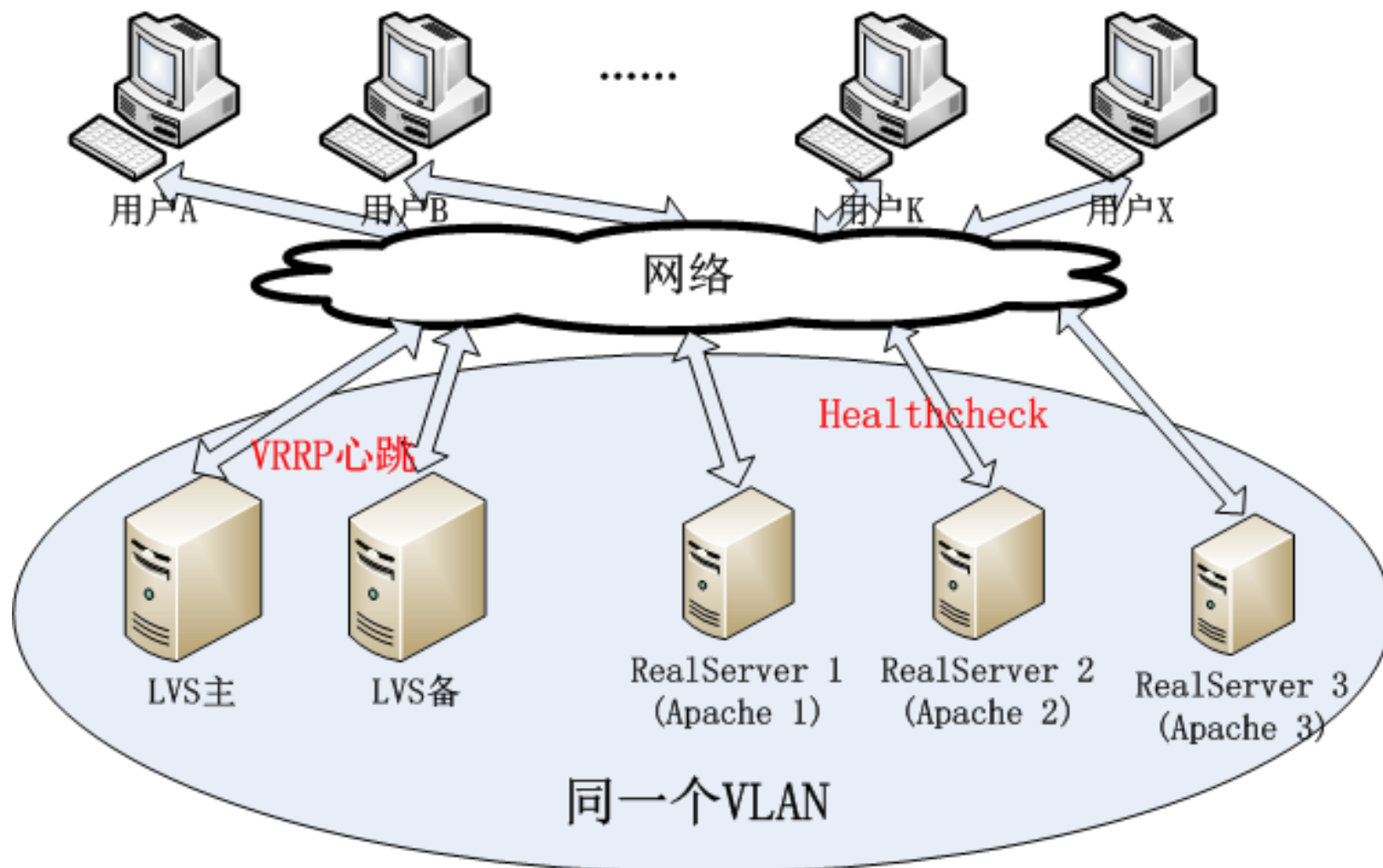
▶ Keepalived - LVS管理软件

- 健康检测：支持4/7监测；
- 主备冗余：采用VRRP协议的HeartBeat；
- 如何配置？ --- 配置文件

Keepalived -f /etc/keepalived/keepalived.conf

Q：缺少监控系统？ LVS具有开源SNMP Patch

简介-应用



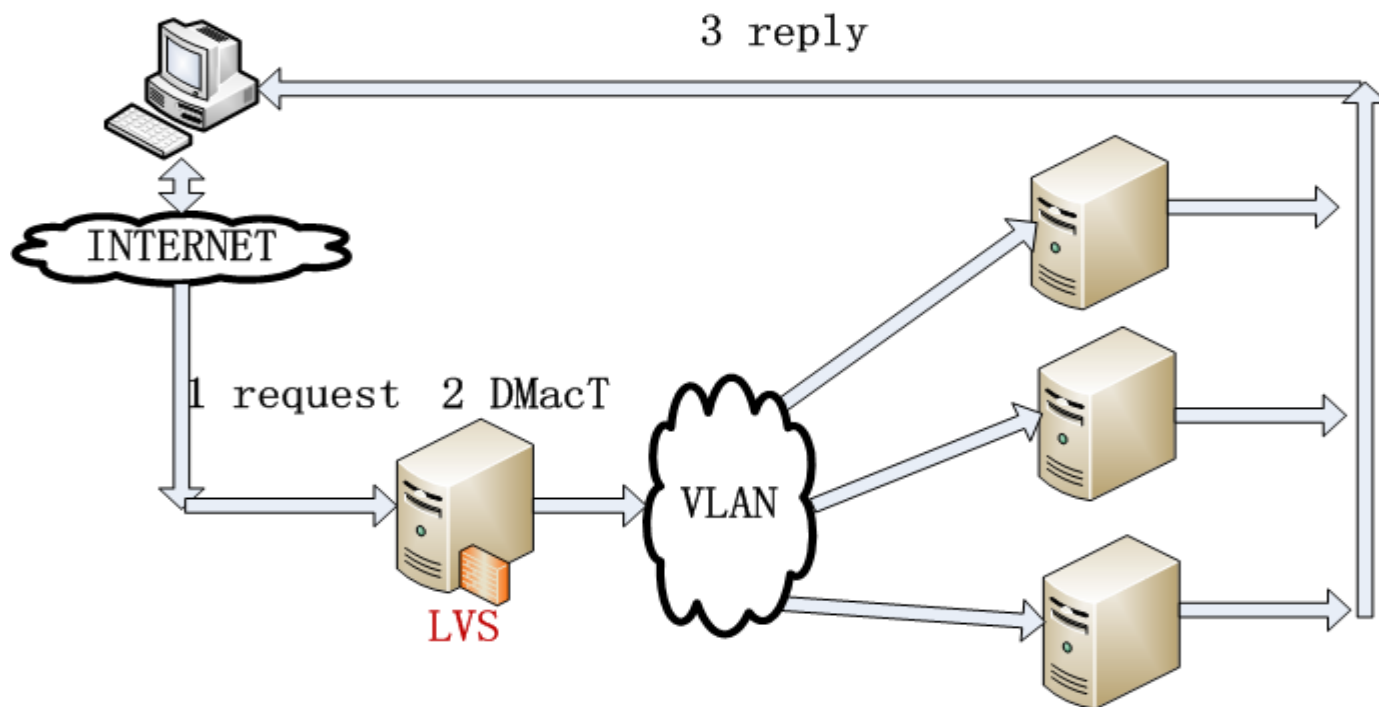
淘宝CDN LVS DR网络拓扑

问题

- ▶ LVS在大规模网络中应用存在不足
 - 各转发模式，网络拓扑复杂，运维成本高
- ▶ 和商用LB设备相比
 - 缺少TCP标志位DDOS攻击防御
- ▶ 主备部署方式不足
 - 性能无法线性扩展

DR模式-不足

- ▶ 不足
 1. LVS-RS间必须在同一个VLAN
 2. RS上绑定VIP，风险大；

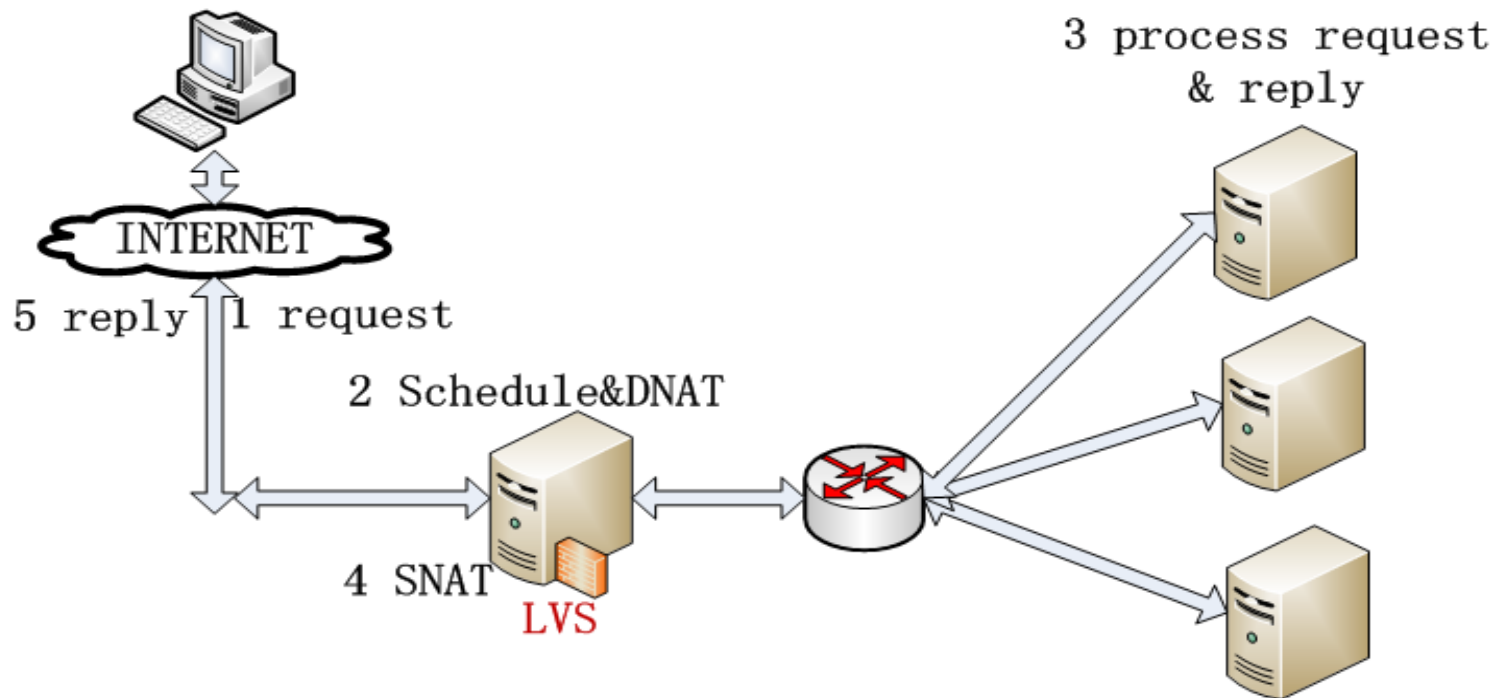


IN: 更改目的MAC
OUT: NULL

NAT模式-不足

▶ 不足

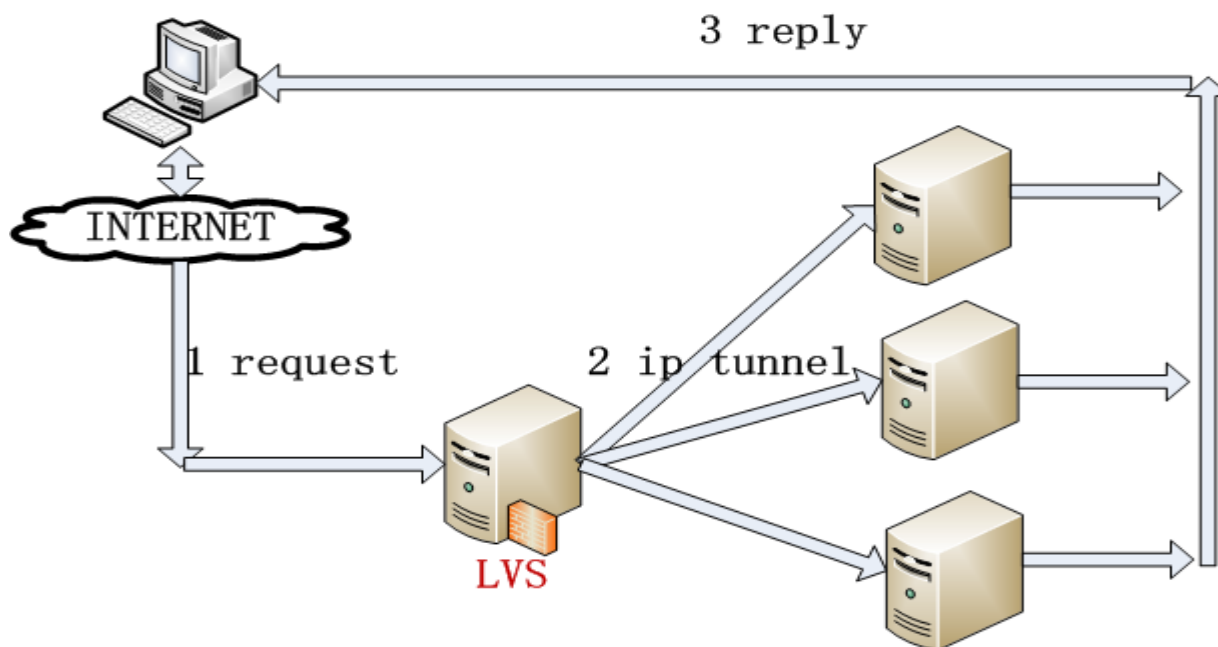
1. RS/ROUTER配置策略路由
由



IN(2): DNAT
OUT(4): SNAT

TUNNEL-不足

- ▶ 不足
 1. RS配置复杂（IPIP模块等）
 2. RS上绑定VIP，风险大；



IN: 增加1个IP头
OUT: NULL

解决方法

- ▶ 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

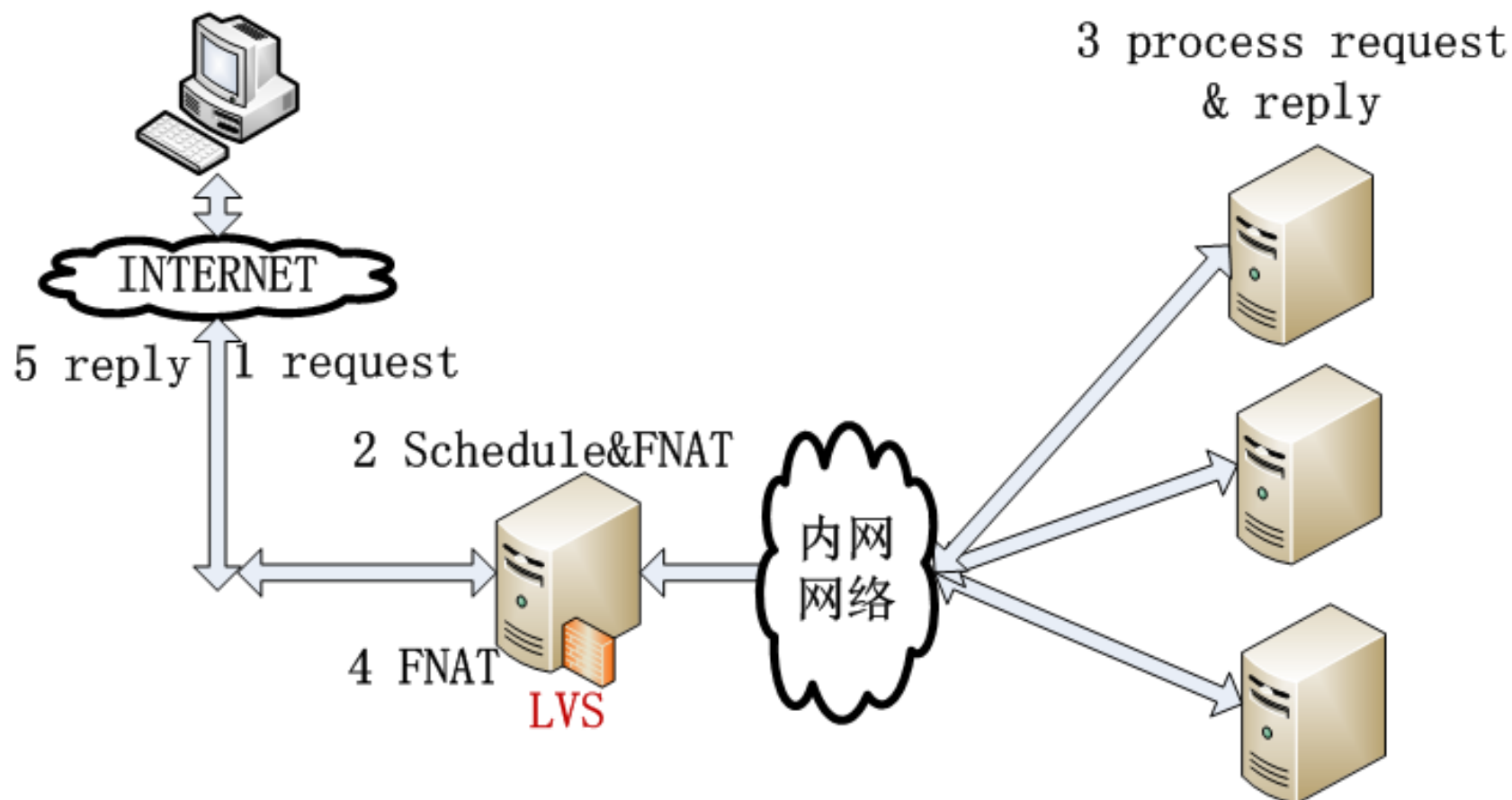
FULLNAT

- ▶ FULLNAT是一种新的转发模式
 - 主要思想：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；
 - keepalived配置方式：

```
virtual_server 1.1.1.1 {  
    lb_kind FNAT/DR/NAT/TUNNEL  
    local_address {  
        192.168.1.1  
    }  
}
```

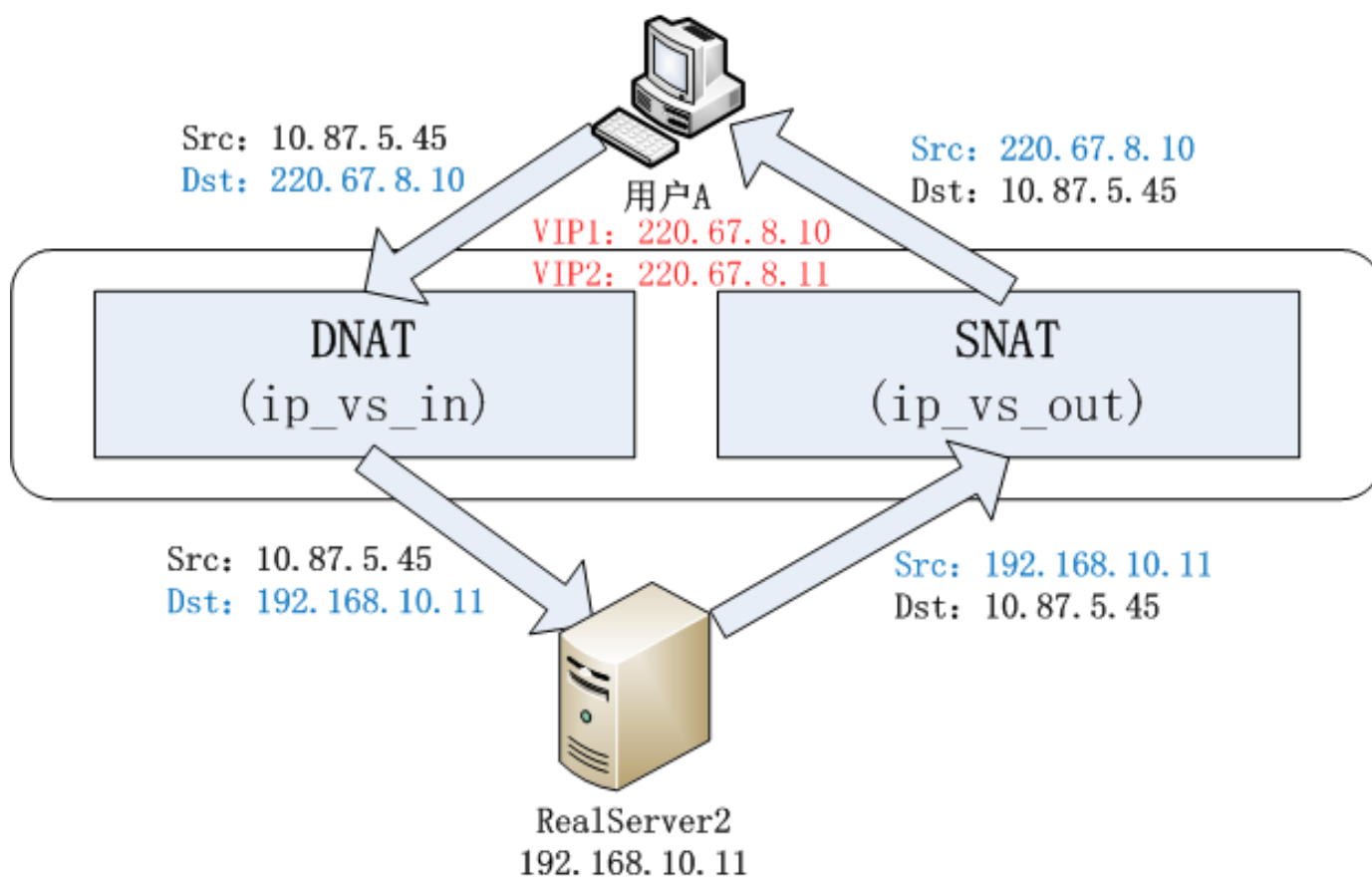
FULLNAT

▶ FULLNAT转发模式



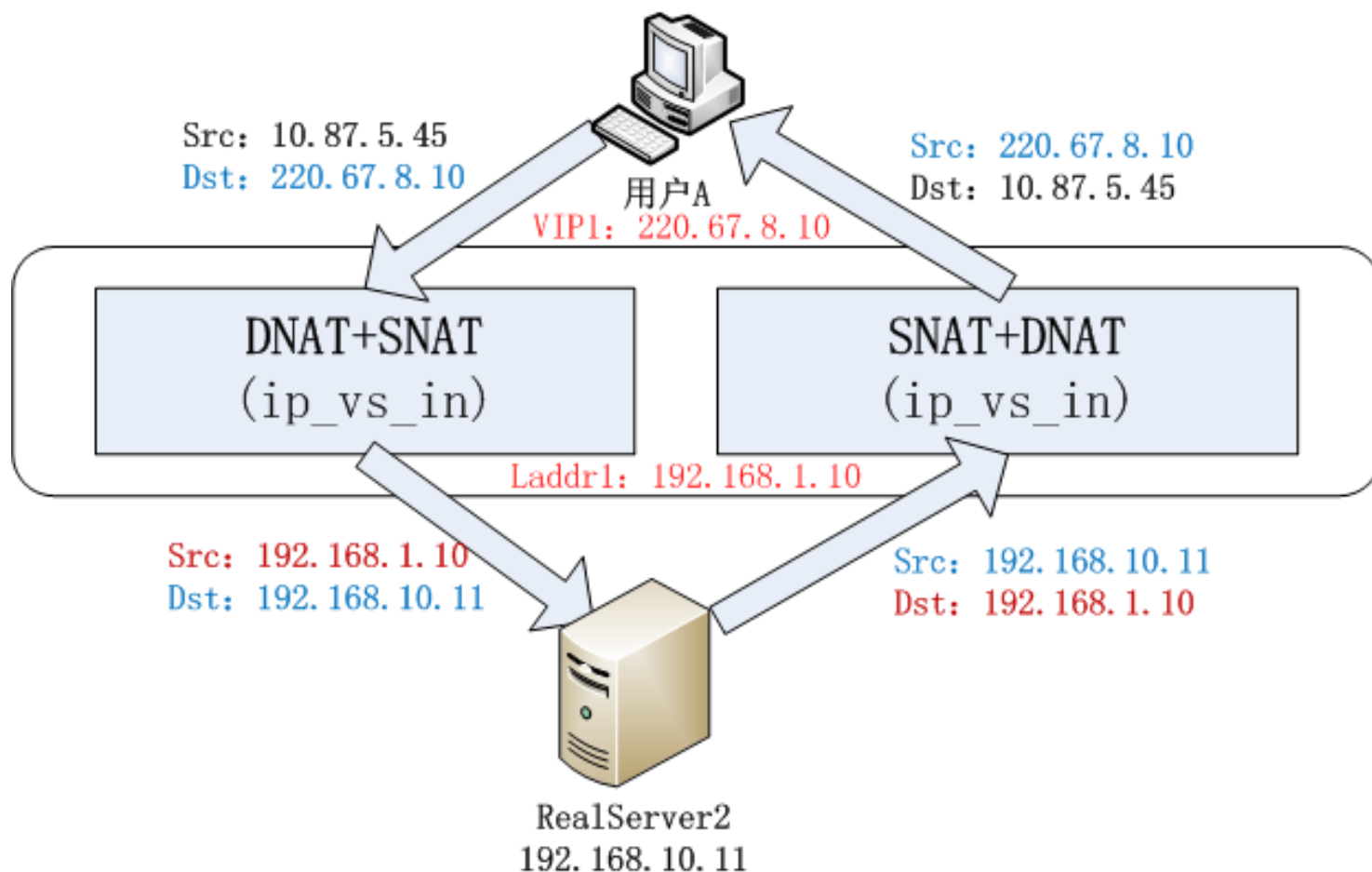
FULLNAT

▶ NAT实现原理



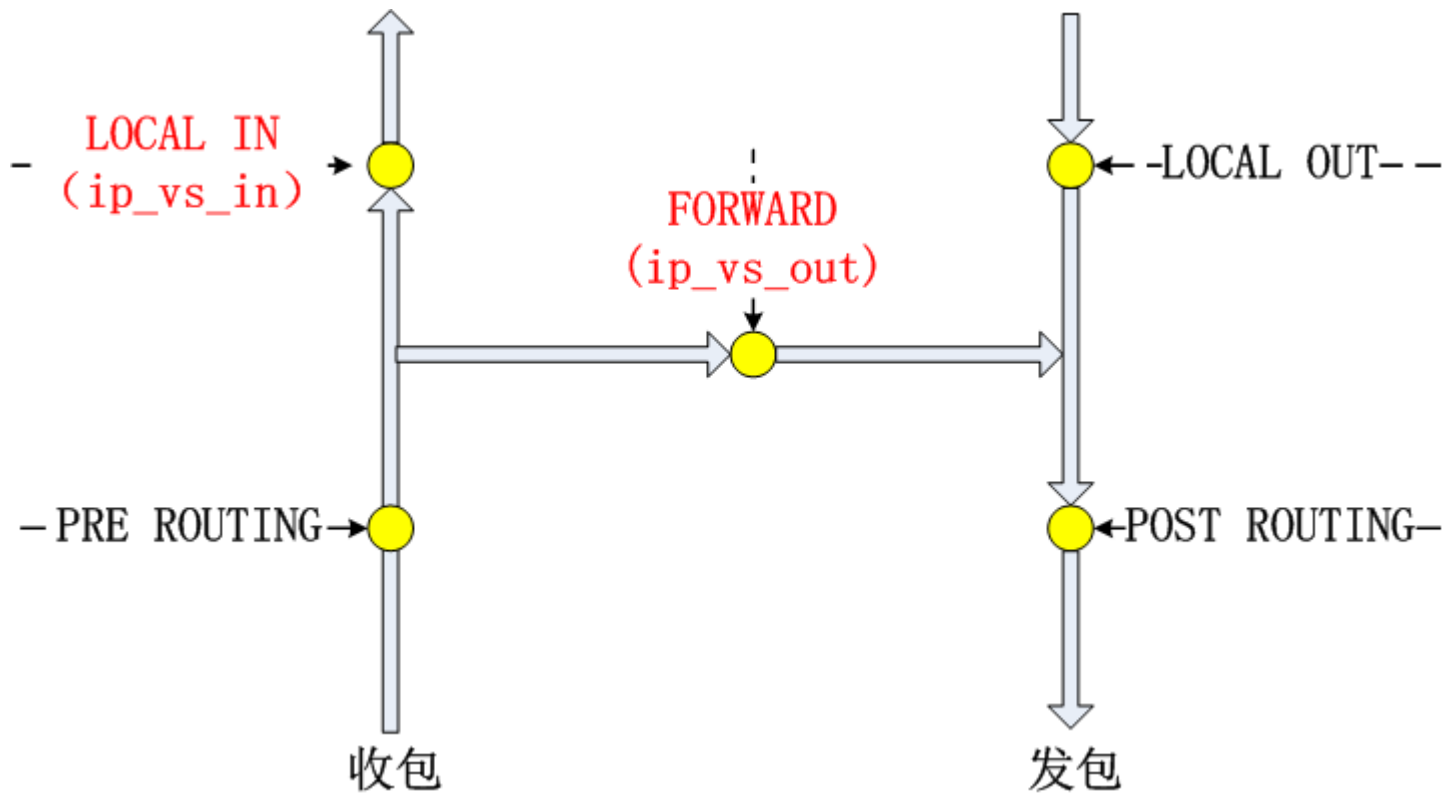
FULLNAT

▶ FULLNAT实现原理



FULLNAT

▶ NAT-HOOK点

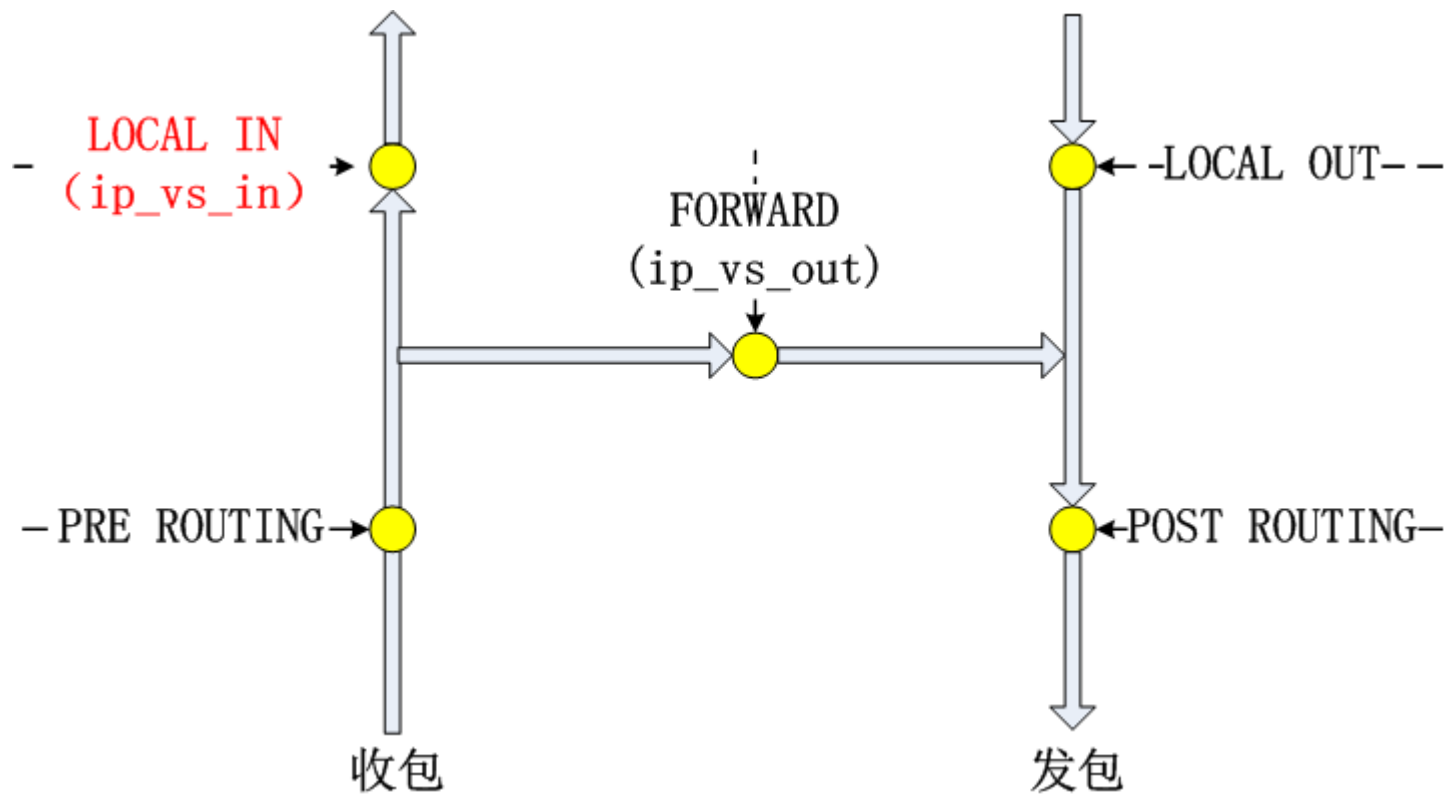


NETFILTER HOOK点, 同iptables

为什么是这2个HOOK点?

FULLNAT

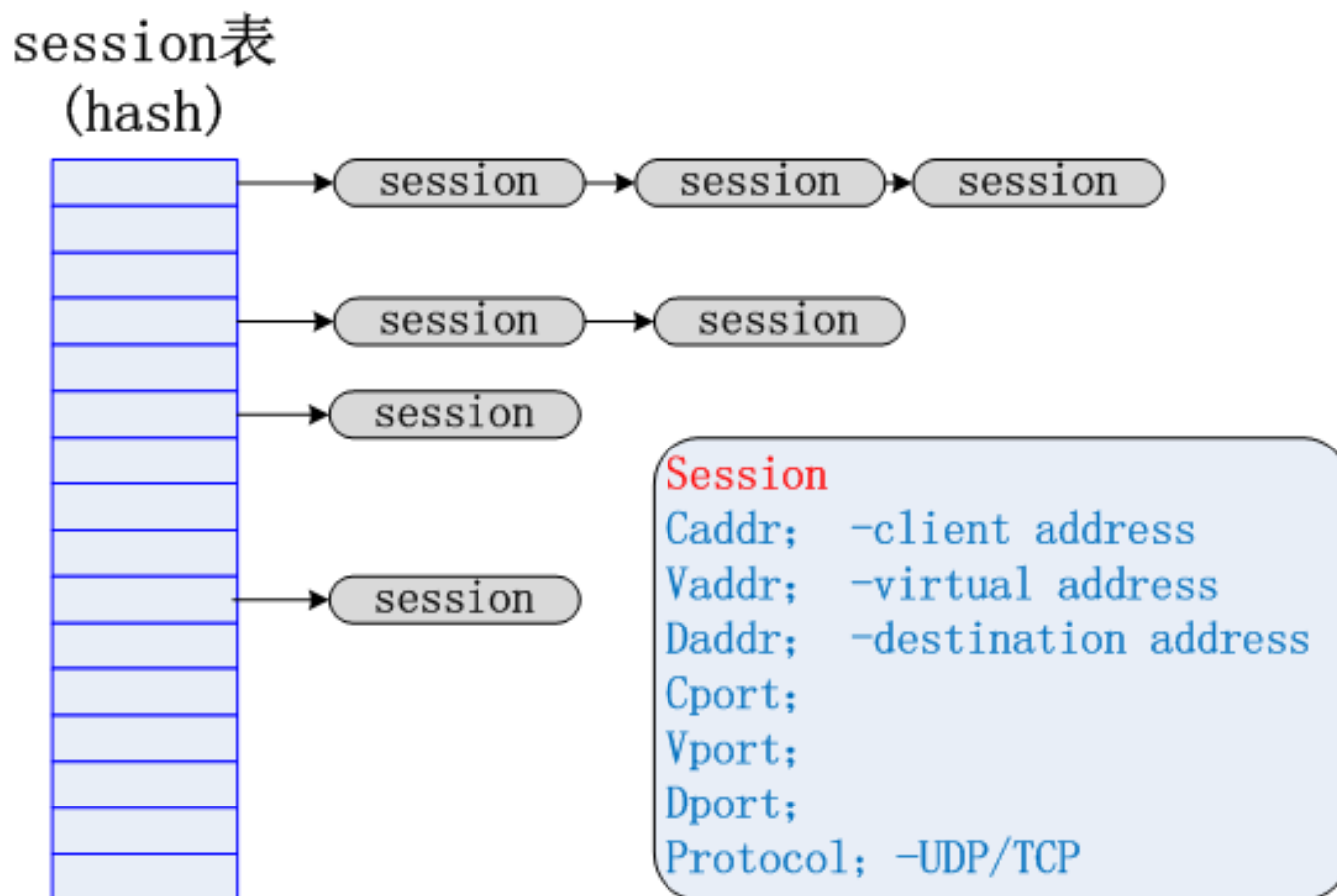
▶ FULLNAT-HOOK点



区分 IN/OUT 流

FULLNAT

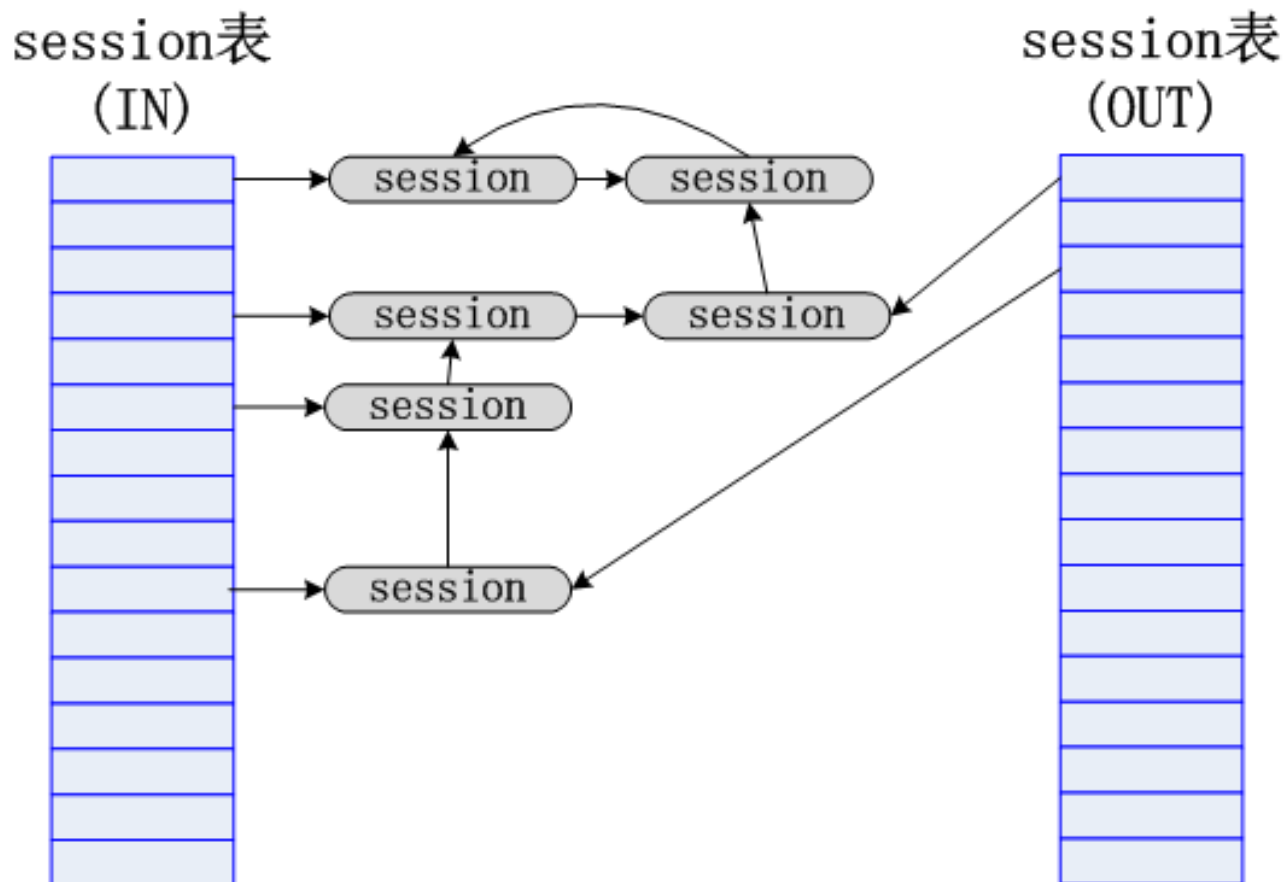
▶ NAT-session表



用client address作为hash key

FULLNAT

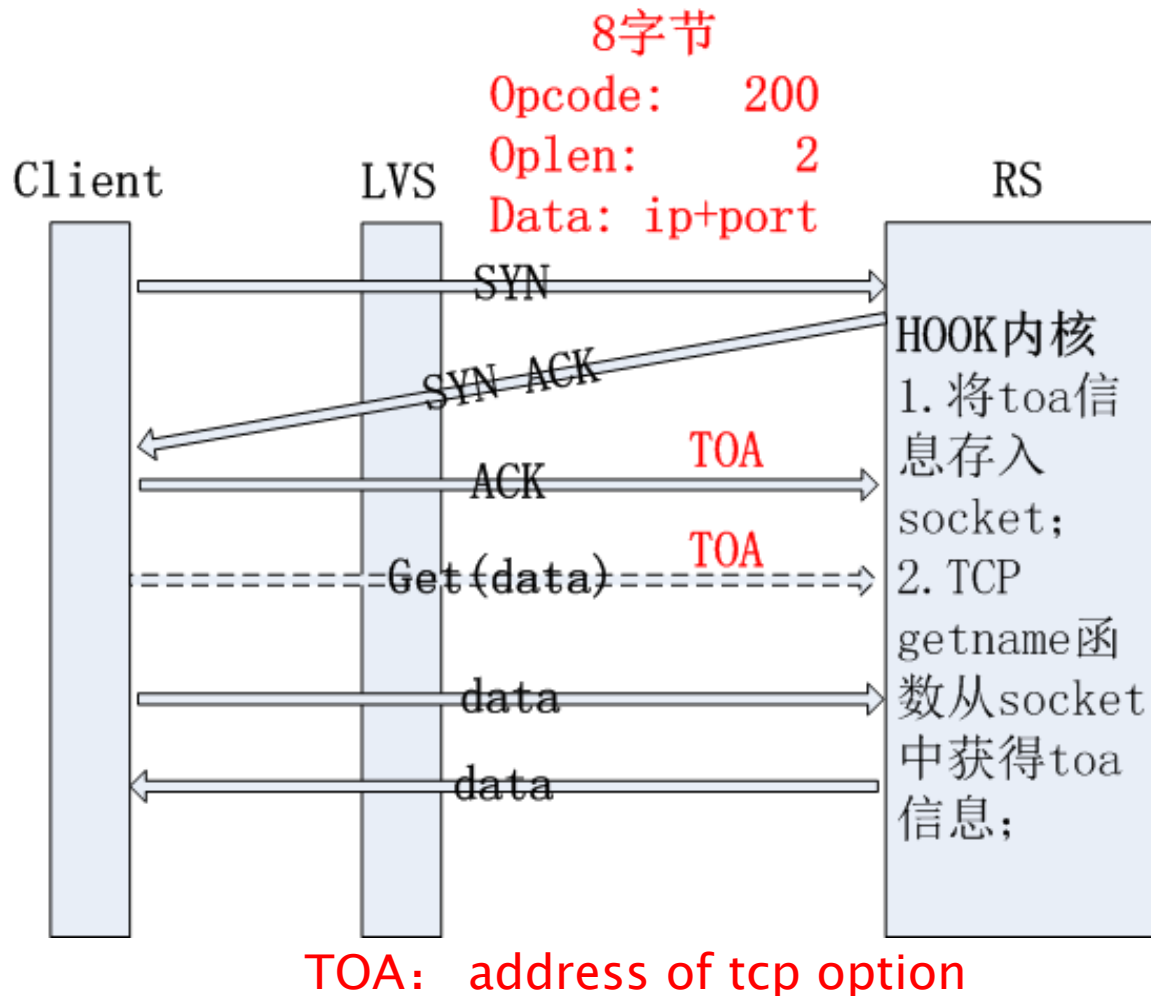
▶ FULLNAT-session表



双向hash, 用五元组作为hash key

FULLNAT

▶ FULLNAT-获取client address (TOA)



FULLNAT

▶ FULLNAT-设计考虑

◦ TCP OPT-TIMESTAMP

- RealServer kernel开启tcp_tw_recycle
- 用户A和B, timestamp大的访问成功, timestamp小的访问失败

◦ TCP OPT-MSS

- TCP三次握手最后一个ack包为GET请求
- GET请求>1个数据包, toa无法插入

◦ TCP - Sequence

- RealServer上timewait的socket复用条件: seq递增

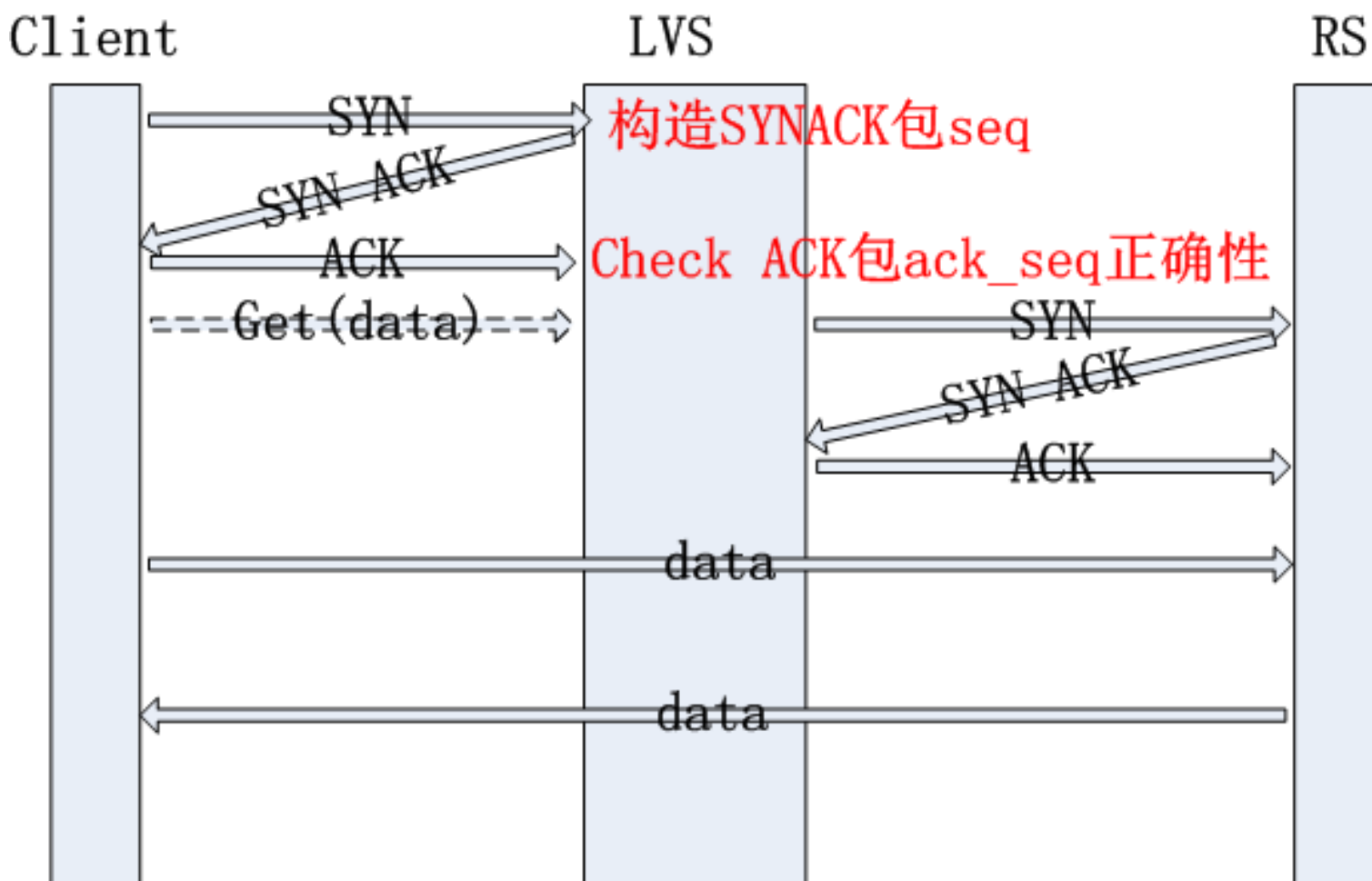
SYNPROXY

- ▶ SYNPROXY用于防御synflood攻击
 - 主要思想：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；
 - 配置方式

```
virtual_server 1.1.1.1 {  
    syn_proxy
```

SYNPROXY

▶ SYNPROXY实现原理

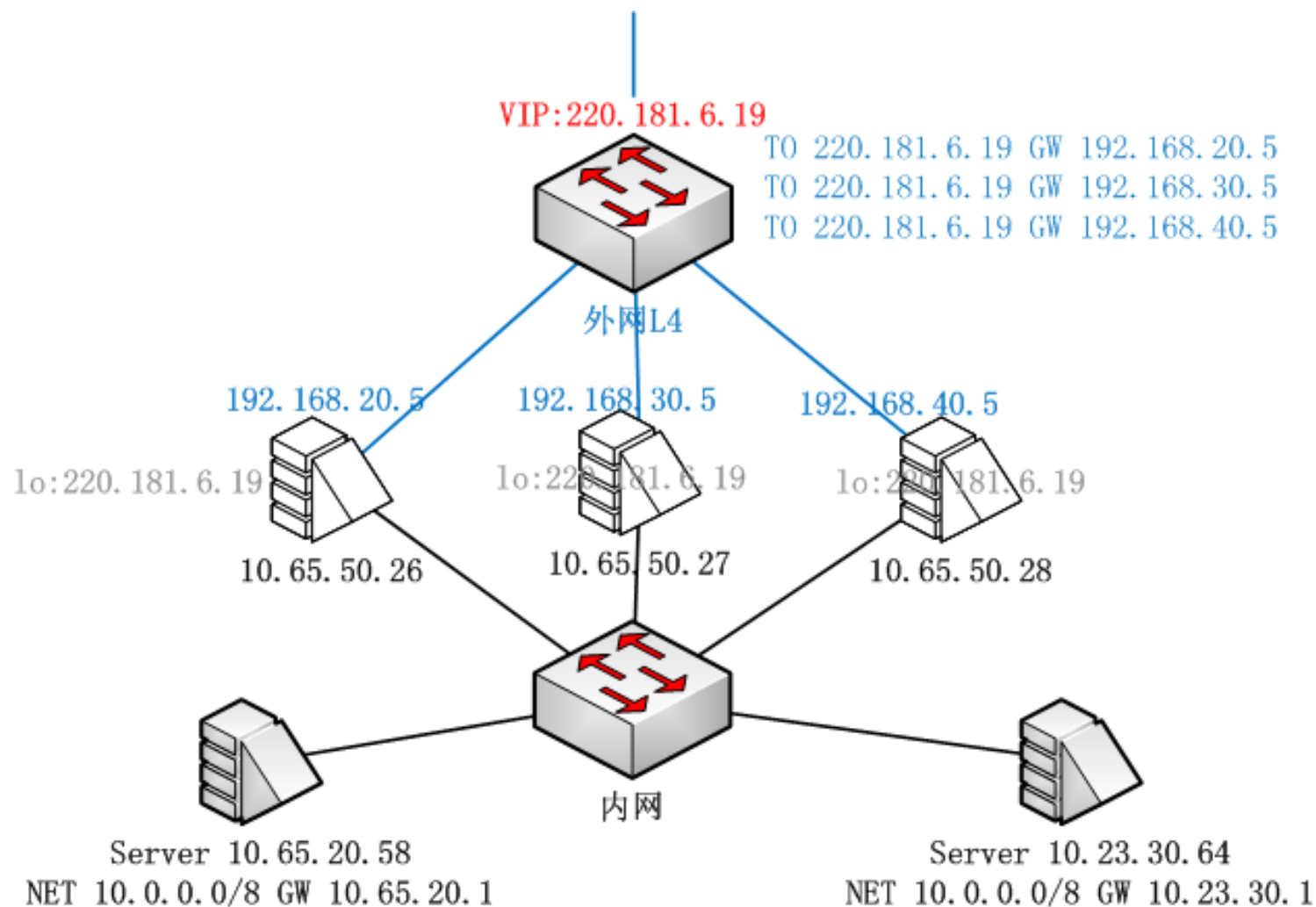


SYNPROXY

▶ SYNPROXY-设计考虑

- TCP – Sequence
 - Lvs->client 和 apache->lvs的syn_ack包中seq不相同
- TCP OPT
 - Lvs->client syn_ack包中tcp opt支持mss/wsack/sack
- Session reused
 - 多个用户通过NAT网关用同一个ip/port访问LVS
- Ack Storm
 - Tcp seq转换导致ack storm

CLUSTER



PERFORMANCE

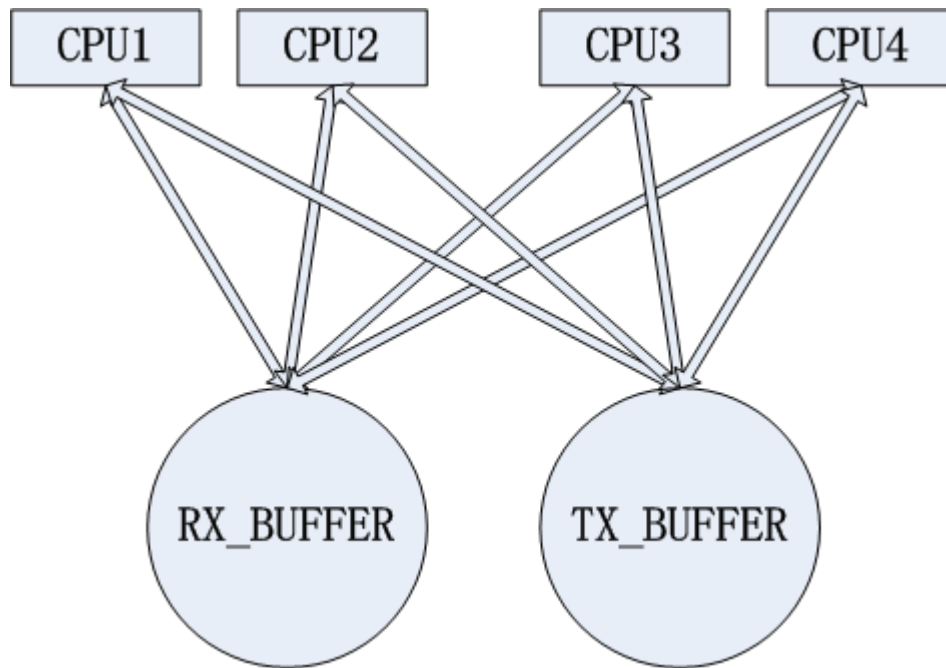
▶ IPVS优化

- 多队列网卡，1个队列绑定到1个cpu核上
- 增大session hash table
- 增大session hash bucket lock个数
- 避免路由cache条目过多
- LOCKLESS
- 硬件：Westmere(第二代nehalem)/bios配置

PERFORMANCE

▶ 单队列网卡

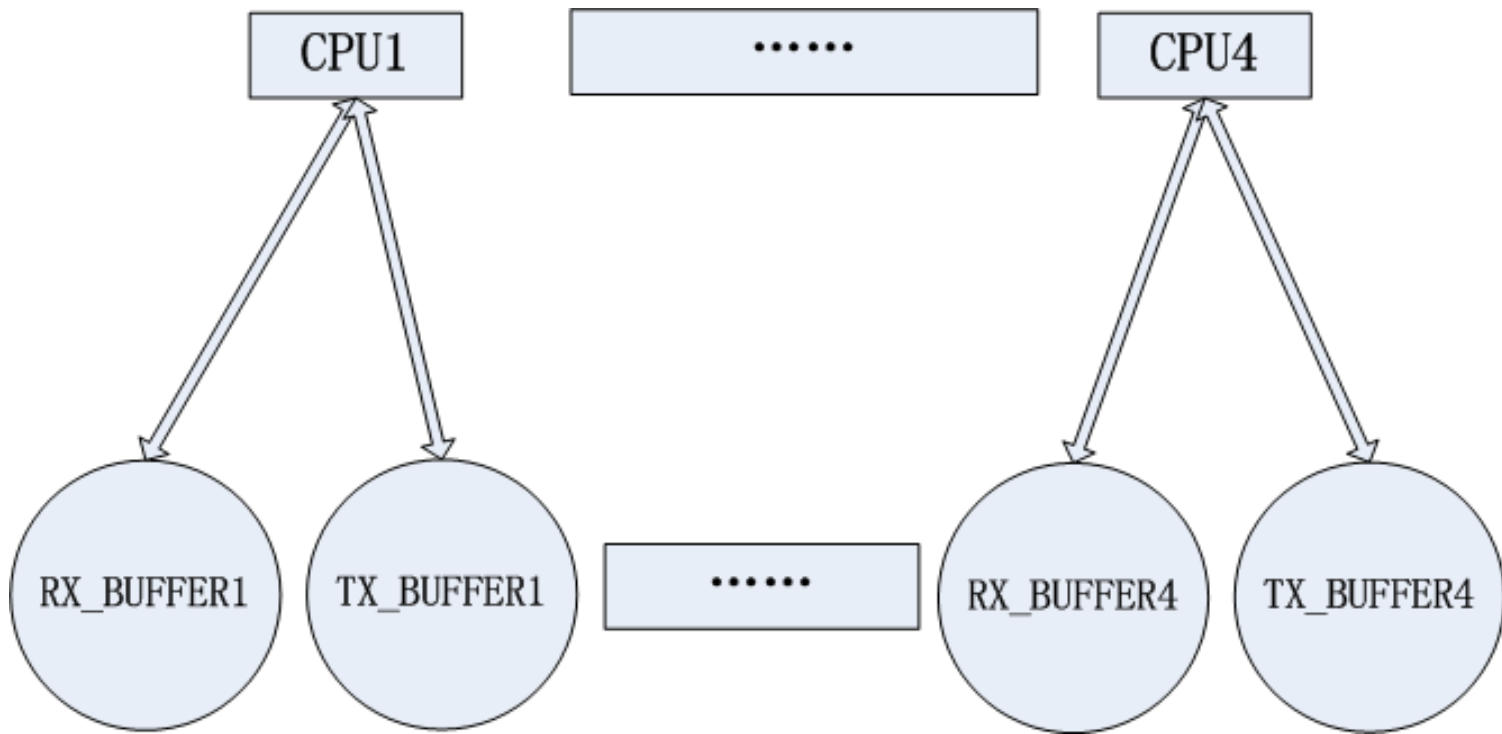
- 只有一个rx_buffer和一个tx_buffer;



PERFORMANCE

▶ 多队列网卡

- N个rx_buffer和N个tx_buffer, $N = \text{CPU核个数}$



PERFORMANCE

▶ 网卡中断- CPU核

◦ Cat /proc/interrupts

```
54: 188324418      0    IR-PCI-MSI-edge    eth0-  
TxRx-0
```

```
55: 167573416      0    IR-PCI-MSI-edge    eth0-  
TxRx-1
```

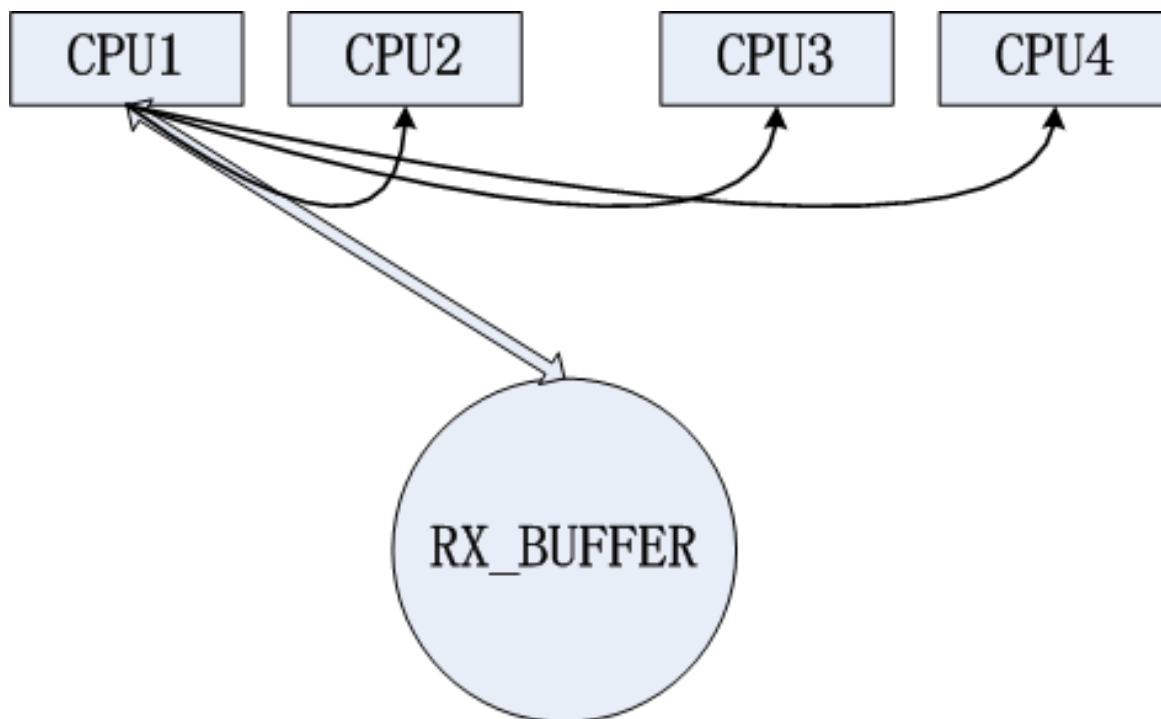
◦ 绑定

```
echo 01 > /proc/irq/54/smp_affinity
```

```
echo 02 > /proc/irq/55/smp_affinity
```

五 performance – 软多队列

- ▶ 软多队列：RPS(receive packet steering)



PERFORMANCE

▶ RPS配置

- `cat /sys/class/net/eth0/queues/rx-0/rps_cpus`
`cat /sys/class/net/eth0/queues/rx-1/rps_cpus`

- 绑定

`echo 01 > /sys/class/net/eth0/queues/rx-0/rps_cpus`
`echo 02 > /sys/class/net/eth0/queues/rx-1/rps_cpus`

PERFORMANCE

- ▶ KEEPALIVED优化
 - Select->epool
 - 减少reload时间和开销

PERFORMANCE

▶ 系统配置注意点

- 关闭网卡LRO/GRO
- 关闭irqbalance
- 增大proc参数: `net.core.netdev_max_backlog`

PERFORMANCE

▶ 性能指标

- Synflood: 350w pps
- Ack/rst/fin-flood: 800w pps
- HTTP: 150w pps
- New tcp connection: 30w
- MAX session: 4000w (24G memory)

机器: DELL R610(E5645 @ 2.40GHz), Intel 82599
NIC,

Todo List

- ▶ 提高性能
 - Ipv6: lockless
 - Keepalived: 多线程事件驱动
 - 新硬件: sandybridge – DDIO
- ▶ 完善功能
 - 攻击防御: ip黑白名单.....
 - 支持GRO(不支持LRO)

未来: 4/7层合一

谢谢

Q&A

新浪微博： 吴佳明_普空