淘宝网 Taobao.com 淘我喜欢！

LVS在大规模网络环境中的应用

吴佳明_普空 阿里巴巴
关注网络技术

LVS历史

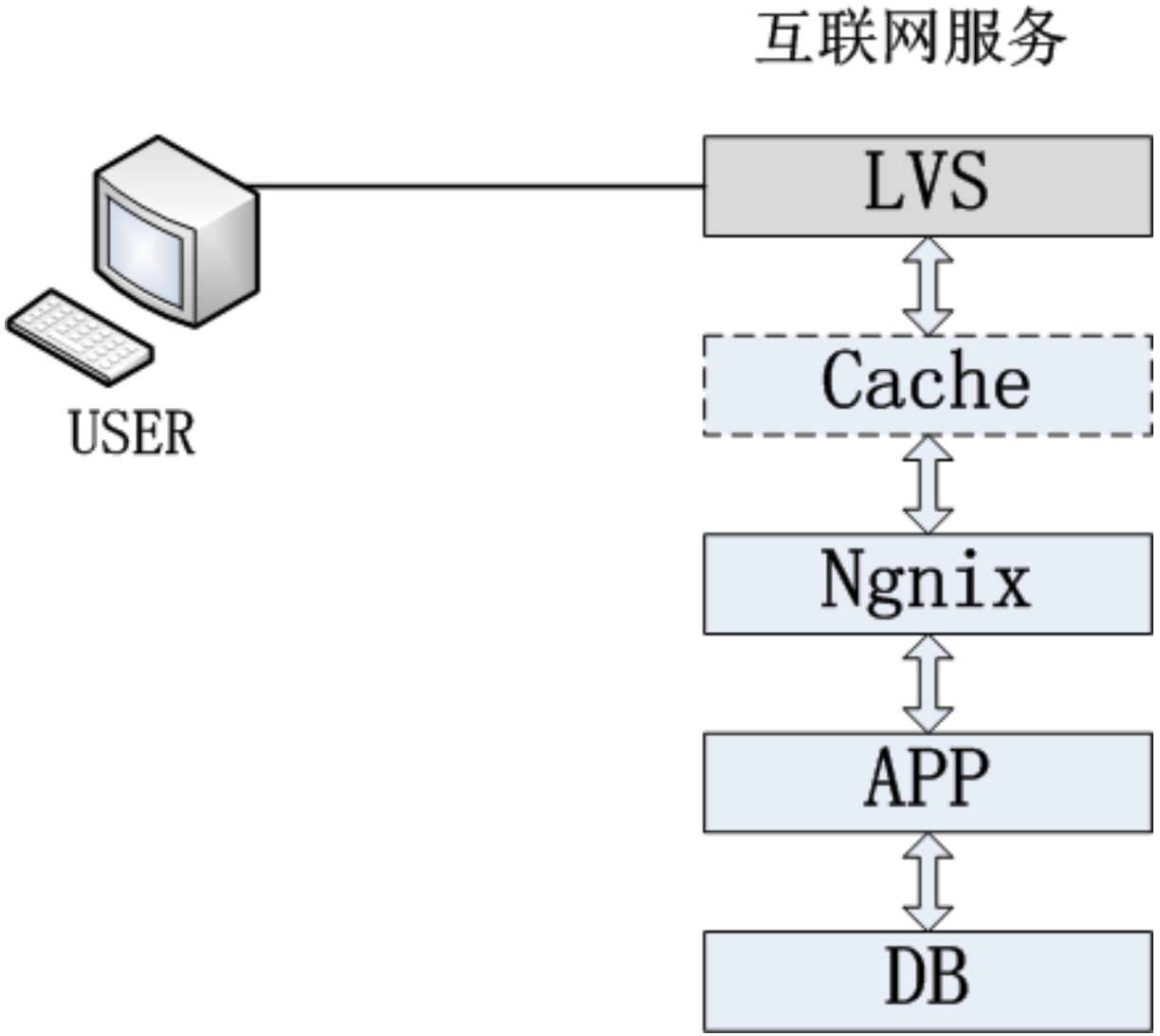
LVS是一个开源的软件，由毕业于国防科技大学的章文嵩博士于1998年5月创立，可以实现LINUX平台下的负载均衡。

LVS是Linux Virtual Server的缩写，意思是Linux虚拟服务器。

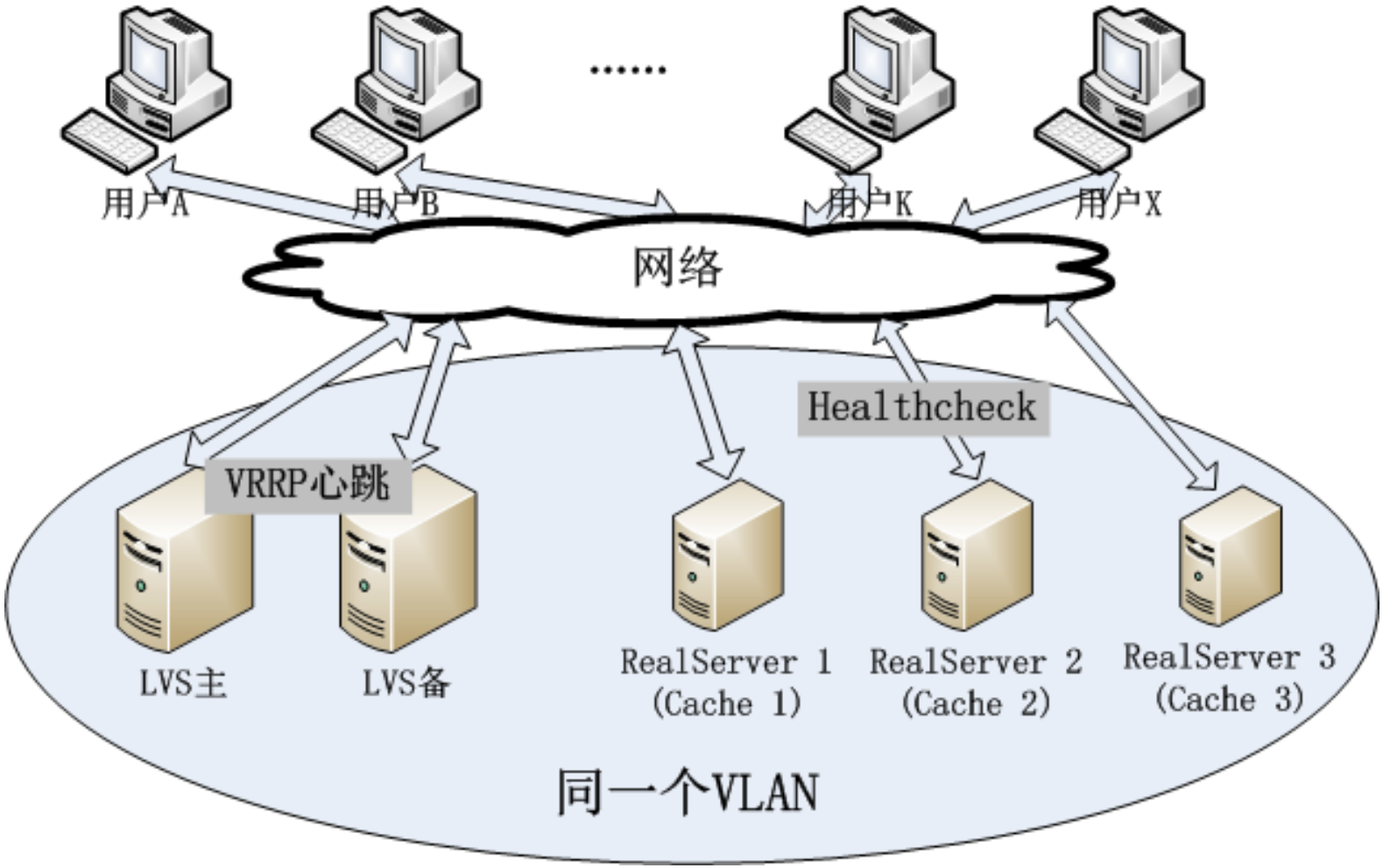
主要内容

1. LVS-问题
2. LVS功能-fullnat
3. LVS功能-synproxy
4. LVS性能-cluster
5. LVS性能-IPVS改进
6. LVS性能-系统改进
7. LVS-todo list

LVS在互联网应用中的位置



LVS网络拓扑



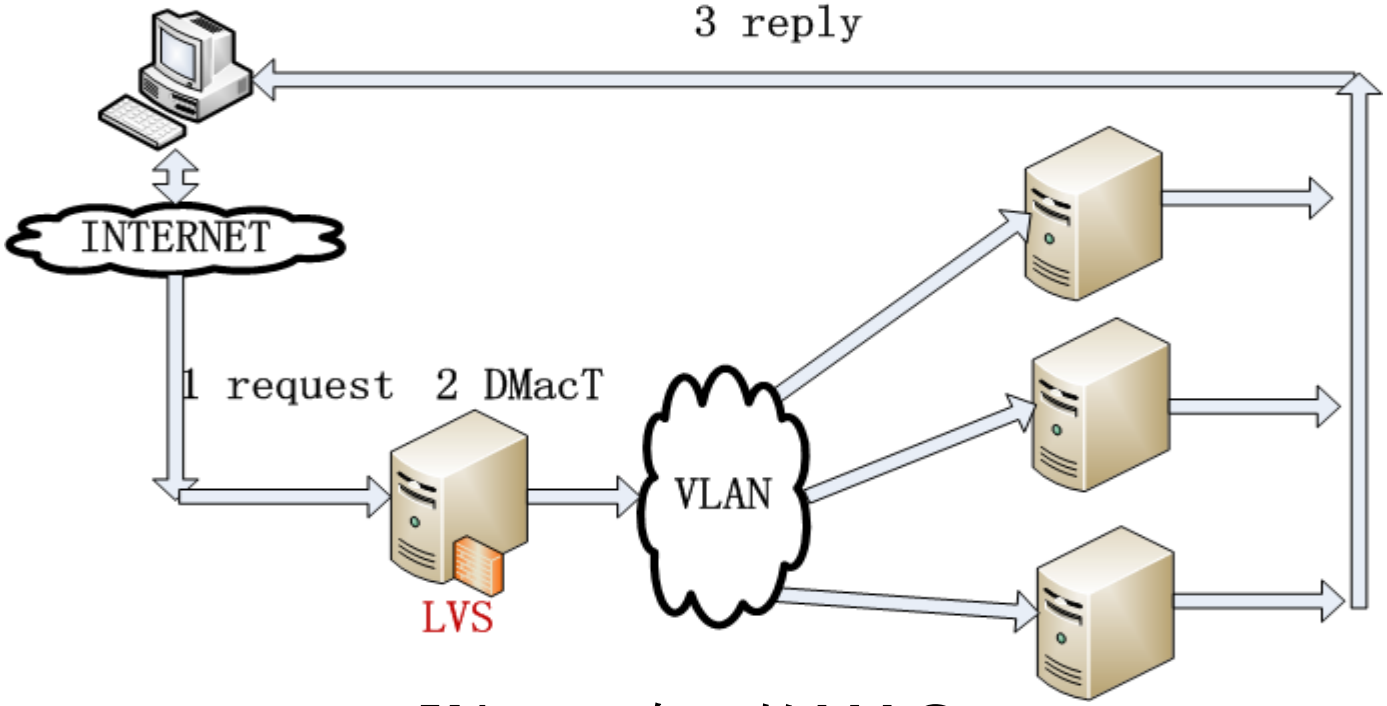
淘宝CDN LVS DR网络拓扑

问题

- LVS部署方式存在不足
  - 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
- 性能不足
  - 10G+ bps的HTTP流量
  - 1000w+ pps的synflood攻击

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；

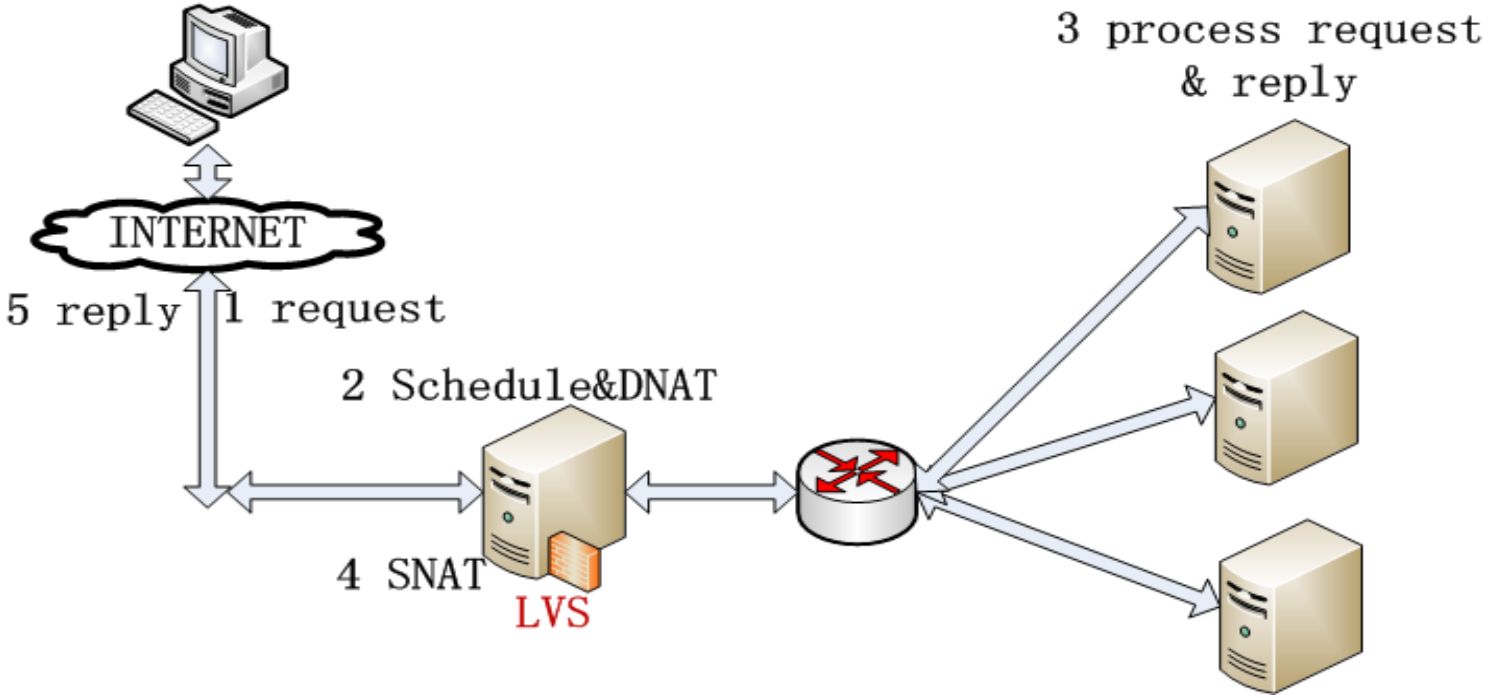


IN：更改目的MAC
OUT：NULL

NAT模式-不足

- 不足

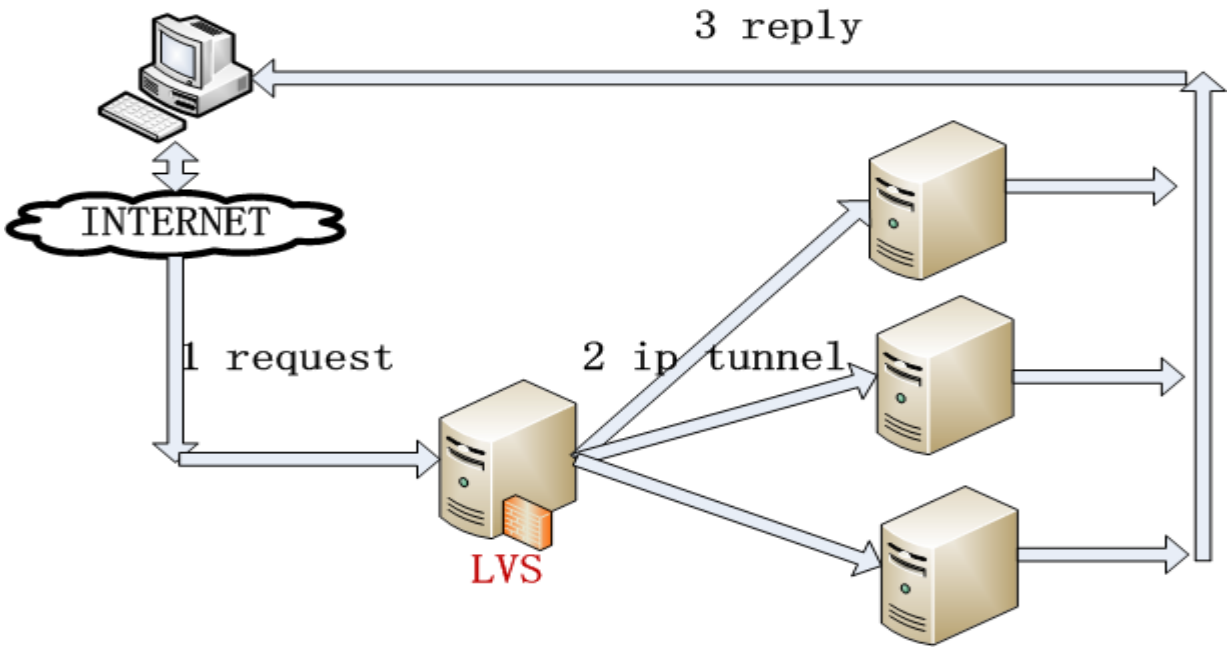
1. RS/ROUTER配置策略路由



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
  - **FULLNAT**：一种新的转发模式，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
- 性能不足
  - 硬件/软件/部署方式 多个层面的优化

FULLNAT

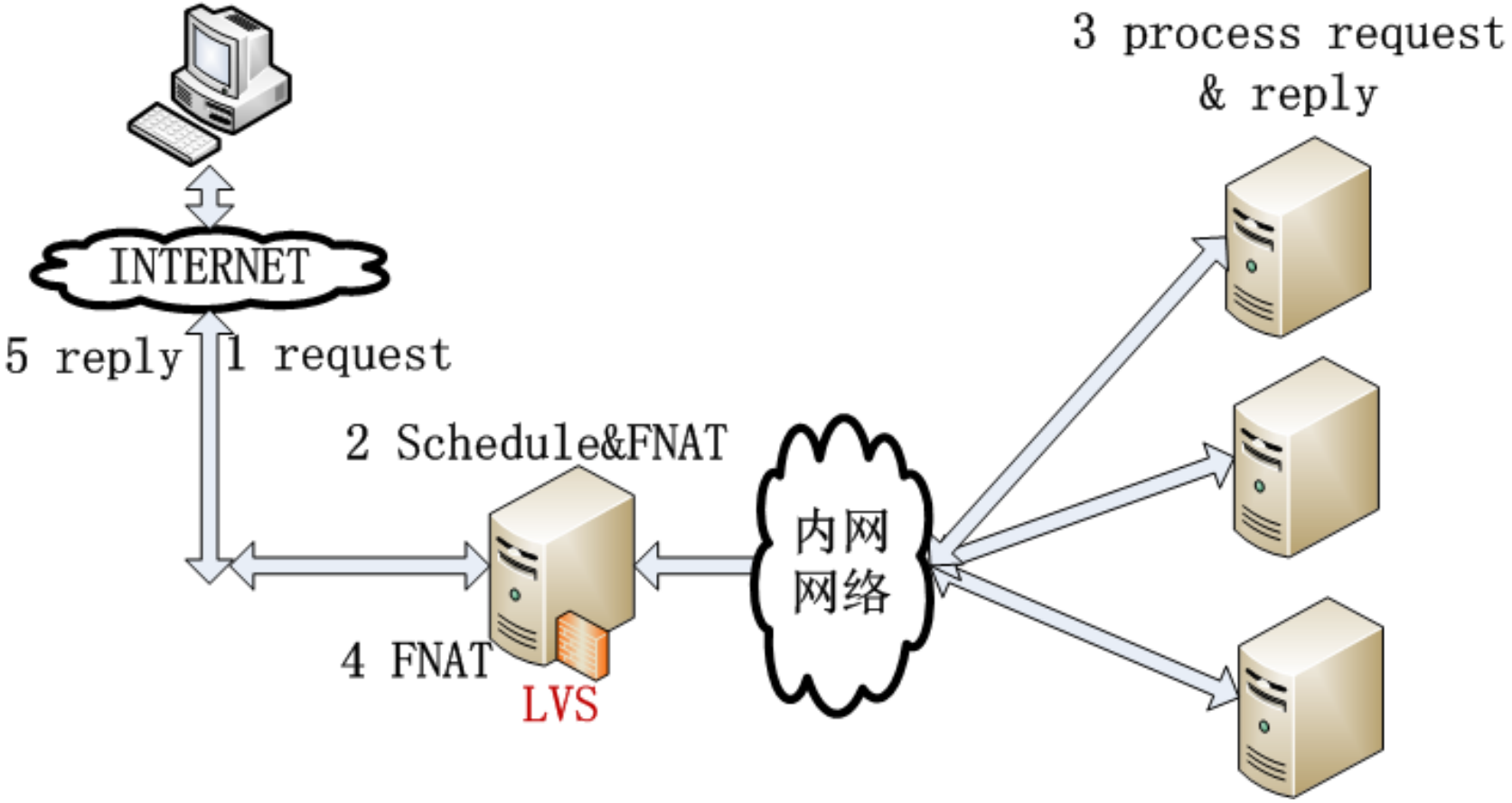
- FULLNAT是一种新的转发模式
  - 主要思想

    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：

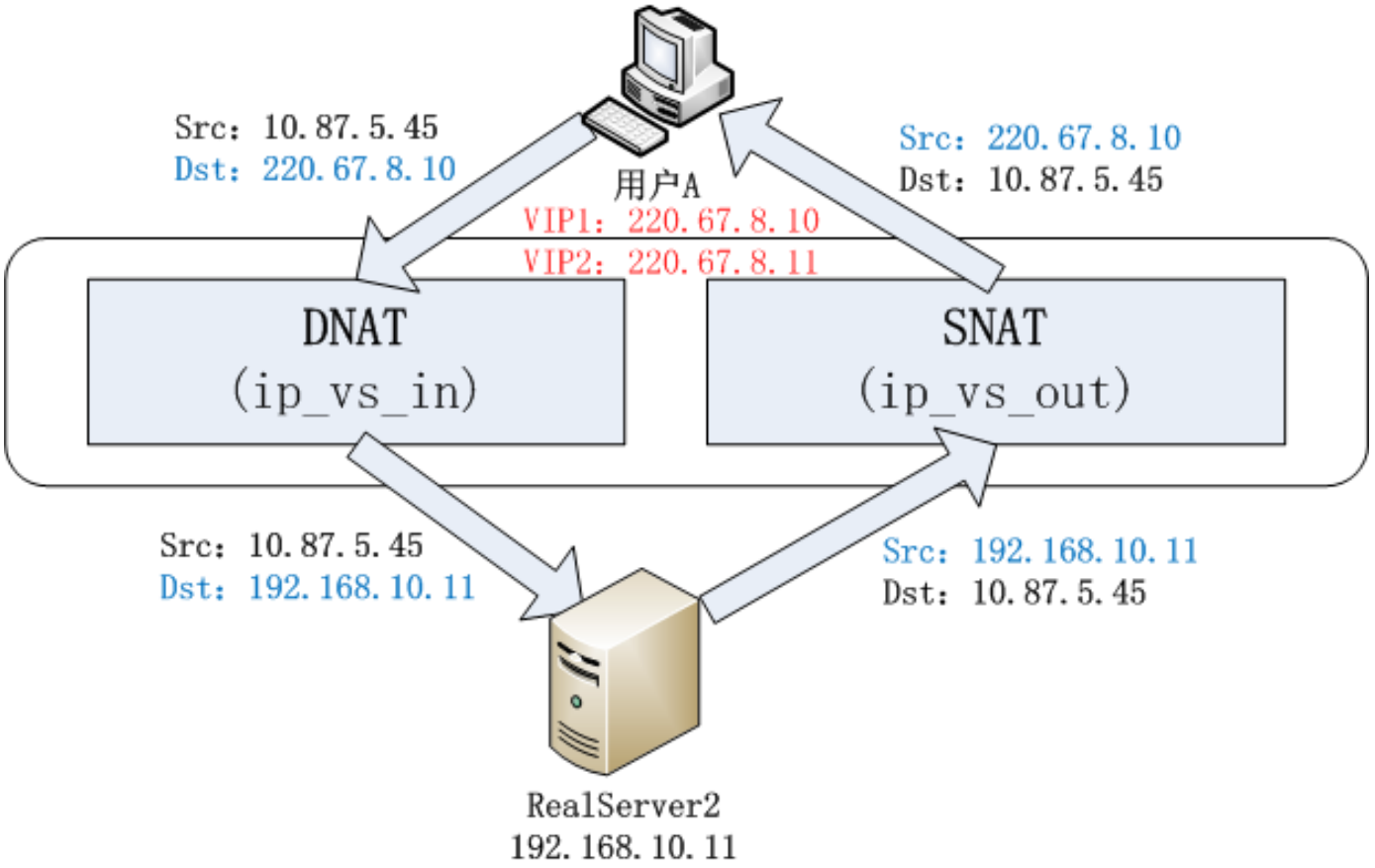
```
virtual_server 125.76.224.240 {
    lb_kind FNAT/DR/NAT/TUNNEL
    local_address {
        192.168.1.1
    }
```

FULLNAT



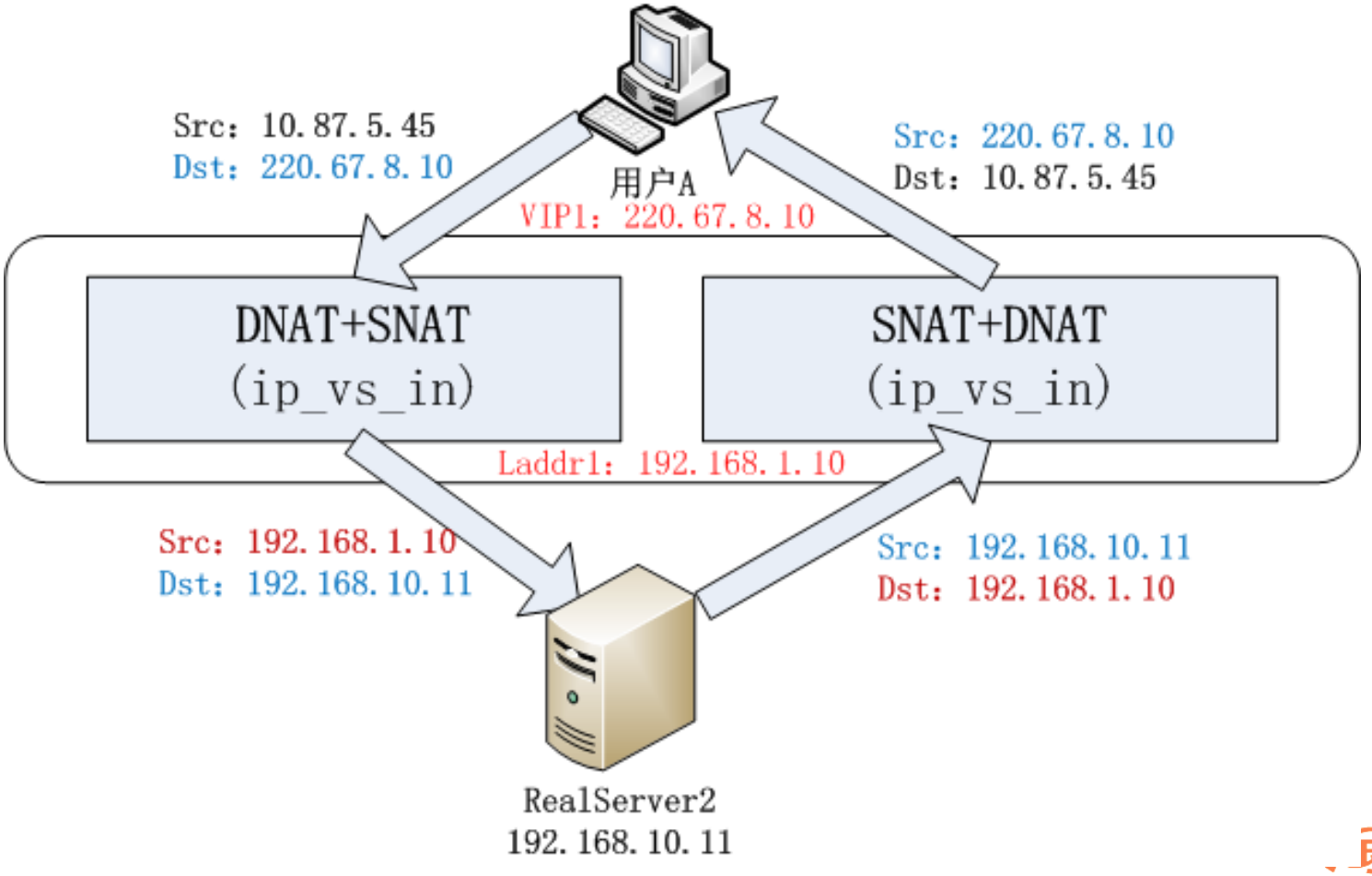
FULLNAT

- NAT实现原理



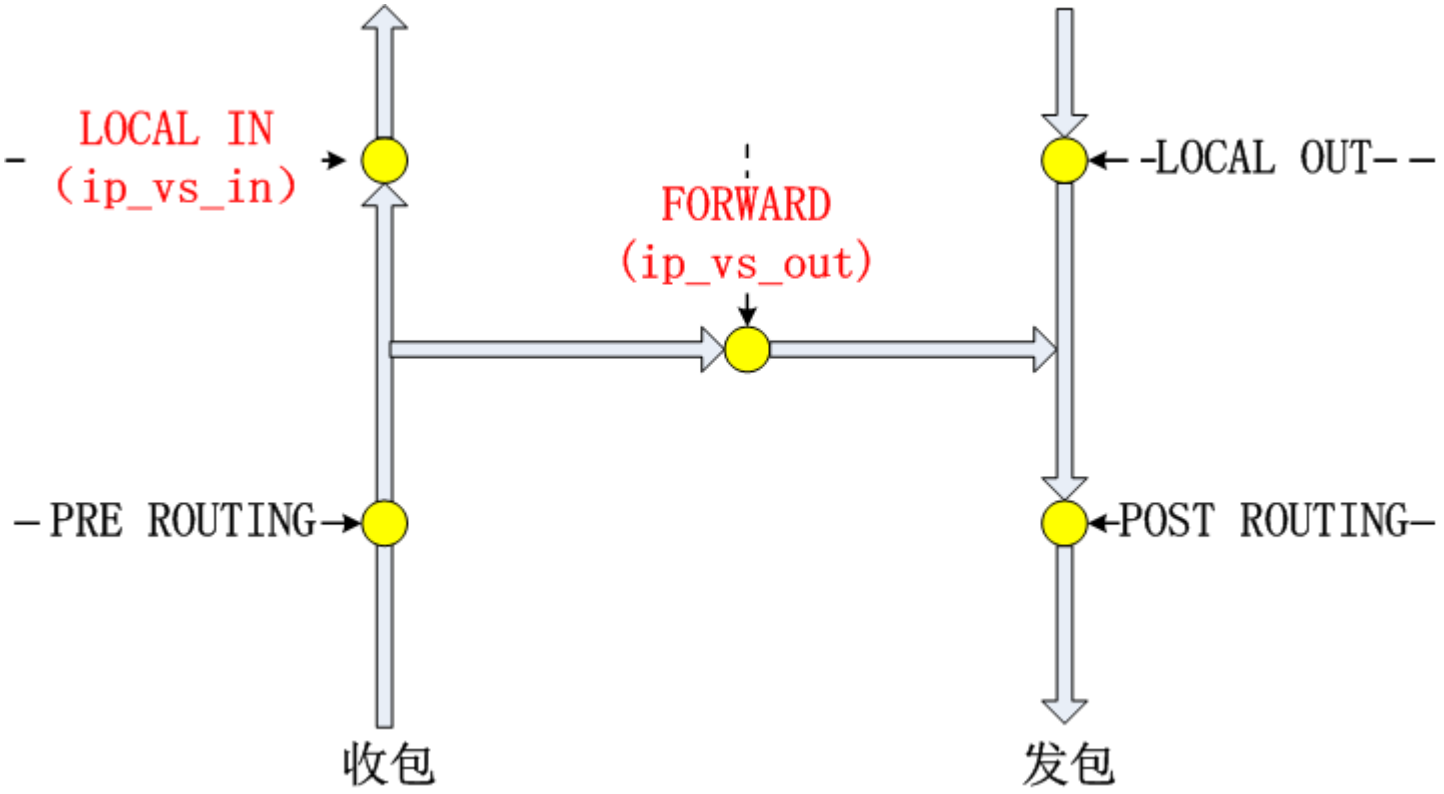
FULLNAT

- FULLNAT实现原理



FULLNAT

- NAT-HOOK点

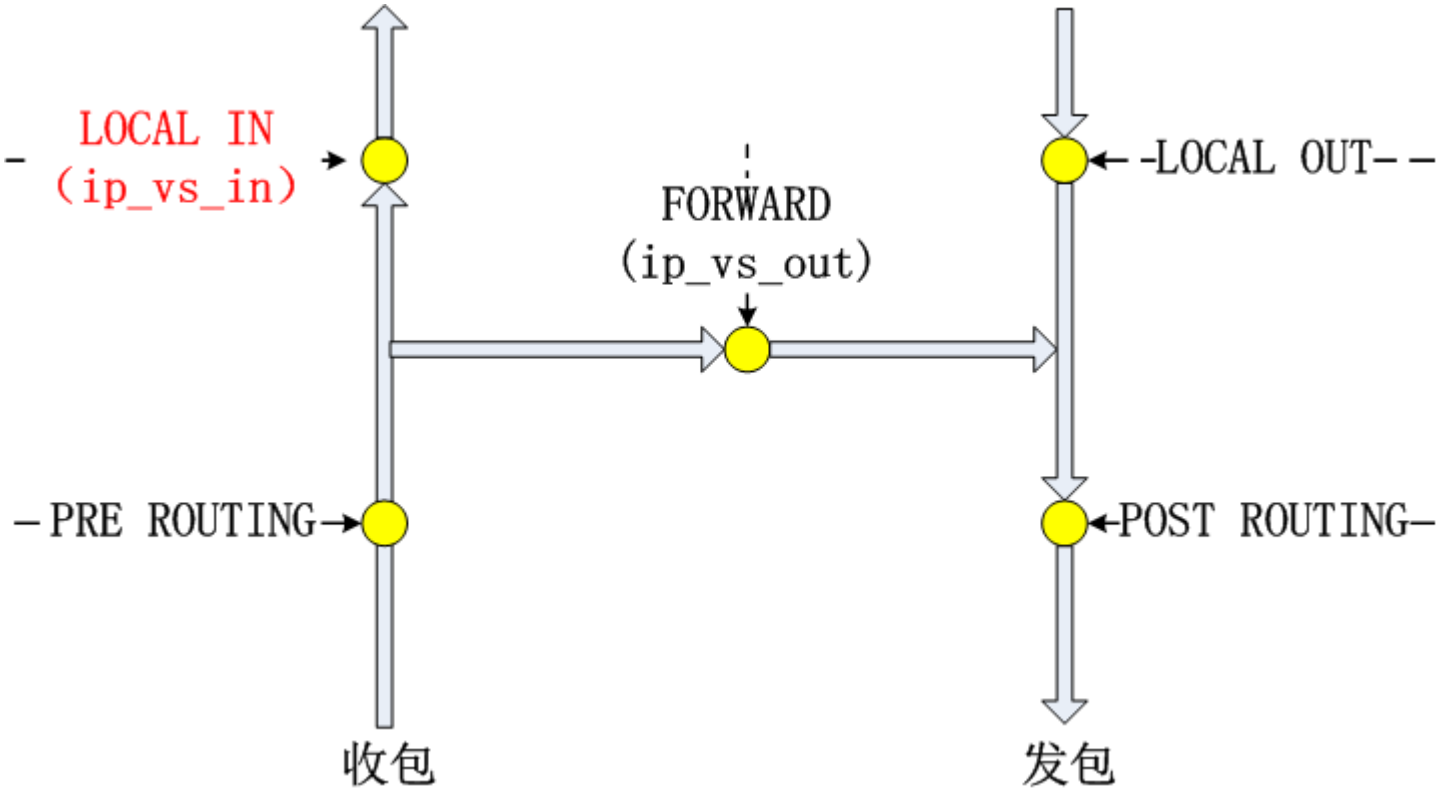


NETFILTER HOOK点，同iptables

为什么是这2个HOOK点?

FULLNAT

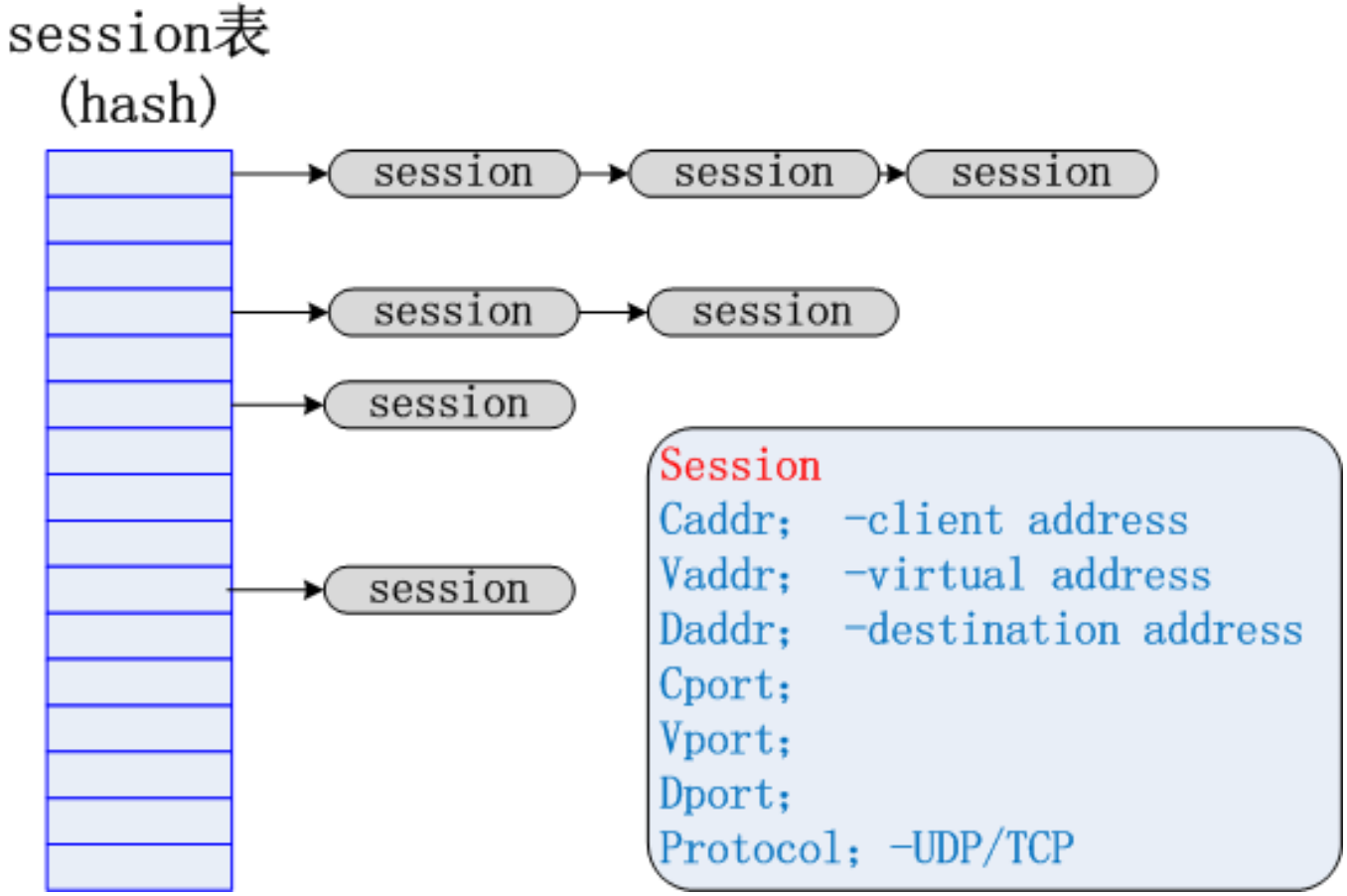
- FULLNAT-HOOK点



区分 IN/OUT 流

FULLNAT

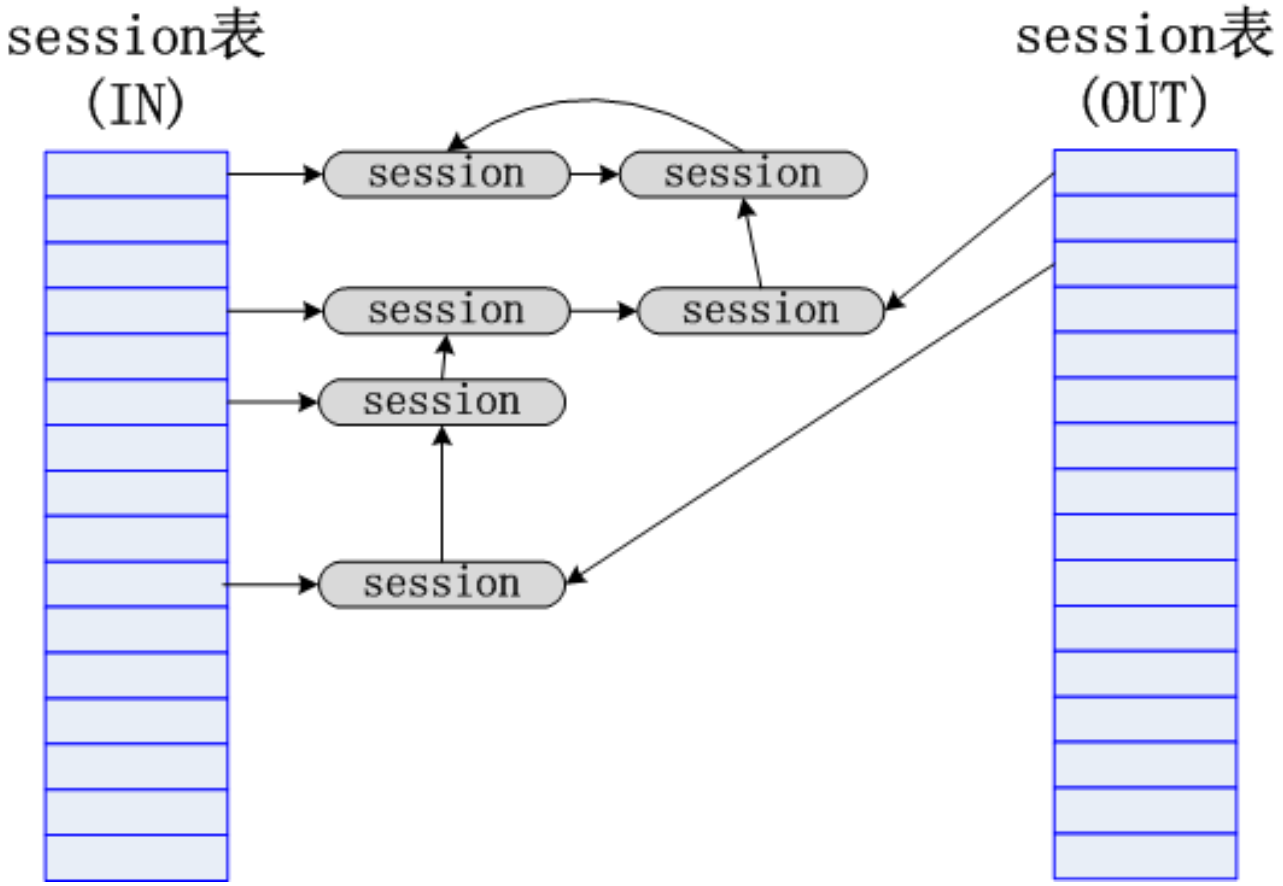
- NAT-session表



用client address作为hash key

FULLNAT

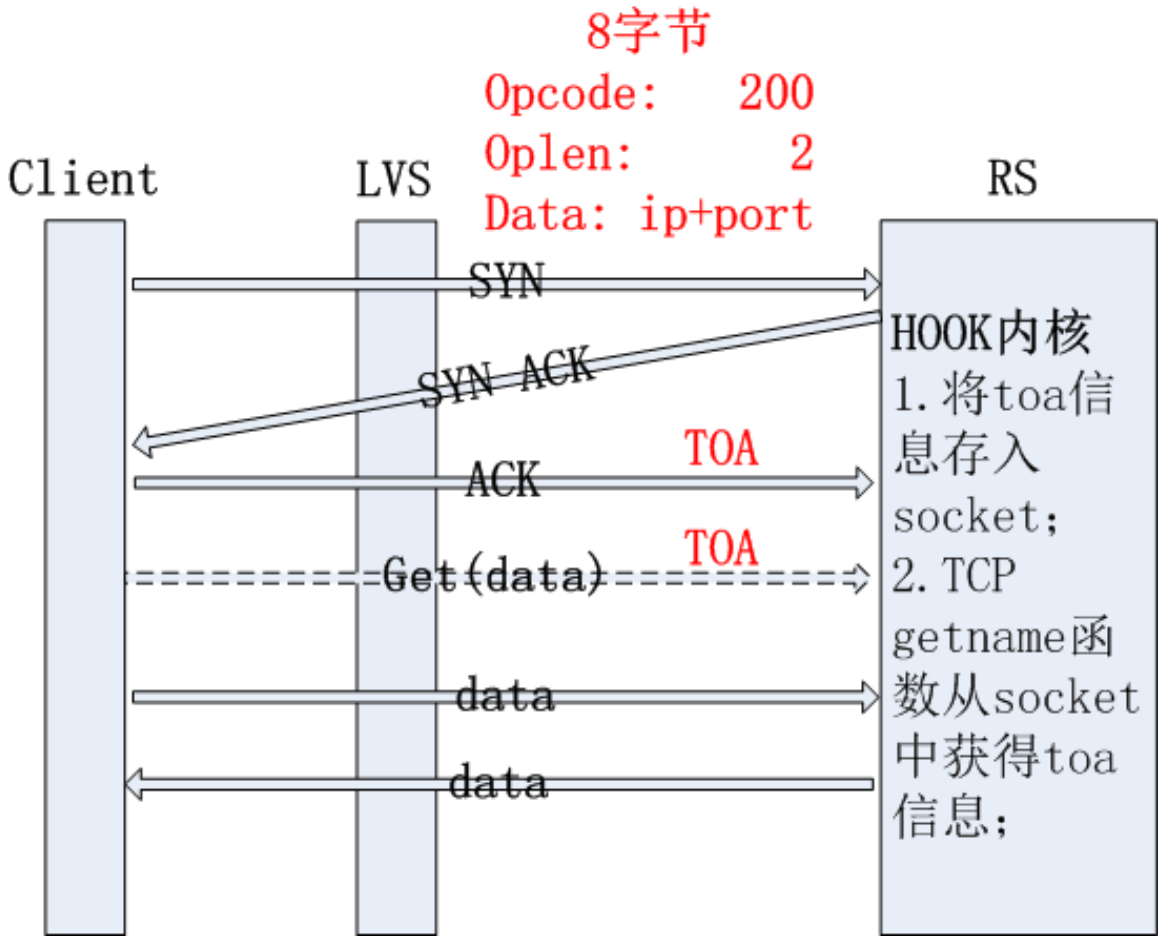
- FULLNAT-session表



双向hash，用五元组作为hash key

FULLNAT

- FULLNAT-获取client address（TOA）



TOA： address of tcp option

SYNPROXY

- SYNPROXY用于防御synflood攻击
  - 主要思想

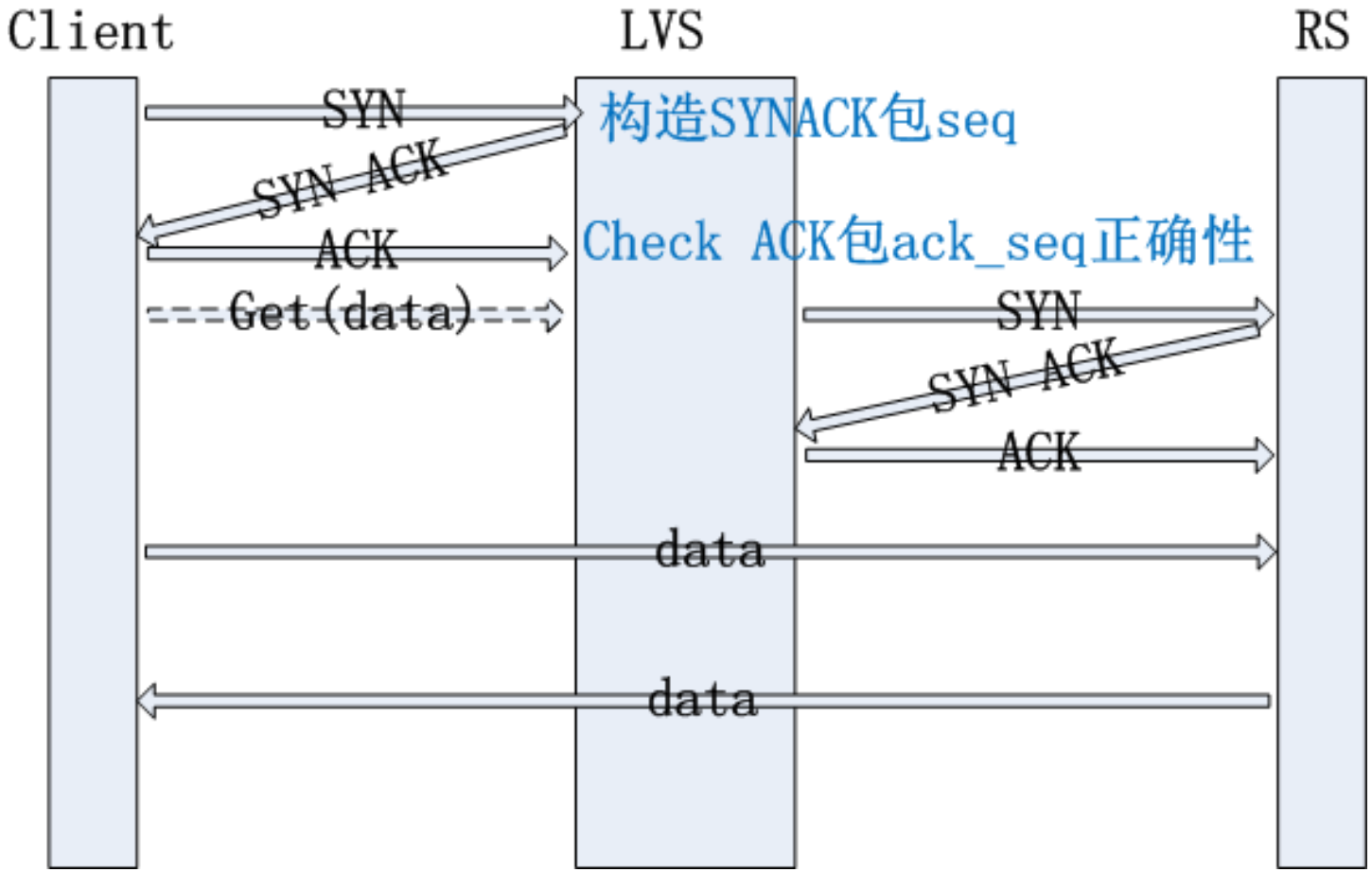
    参照linux tcp协议栈中syncookies的思想，LVS-构造特殊seq的synack包，验证ack包中ack_seq是否合法-实现了TCP三次握手代理；

  - 配置方式

    virtual_server 125.76.224.240 {

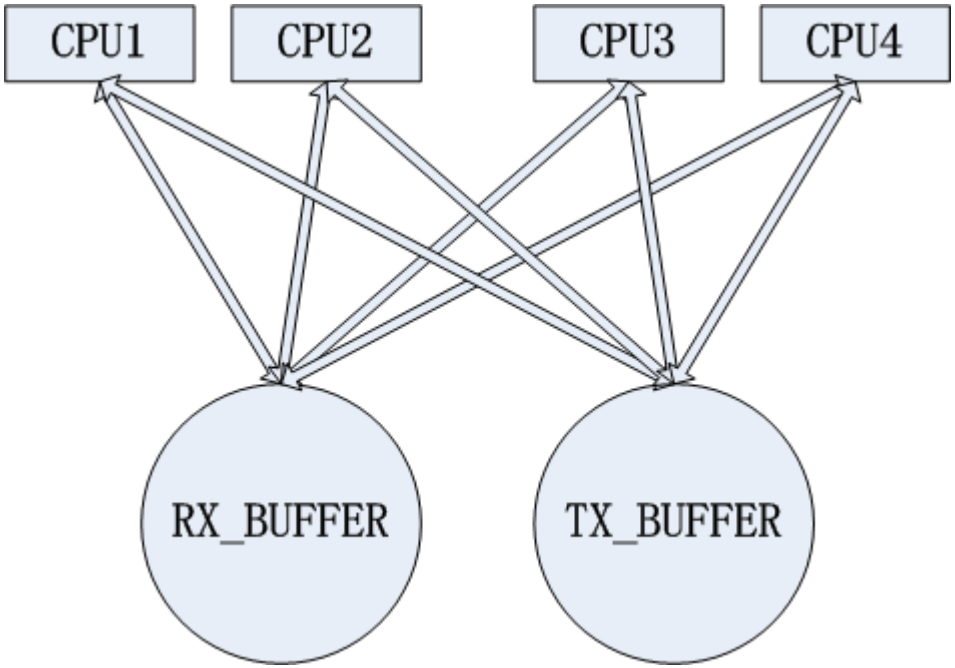
    syn_proxy

SYNPROXY



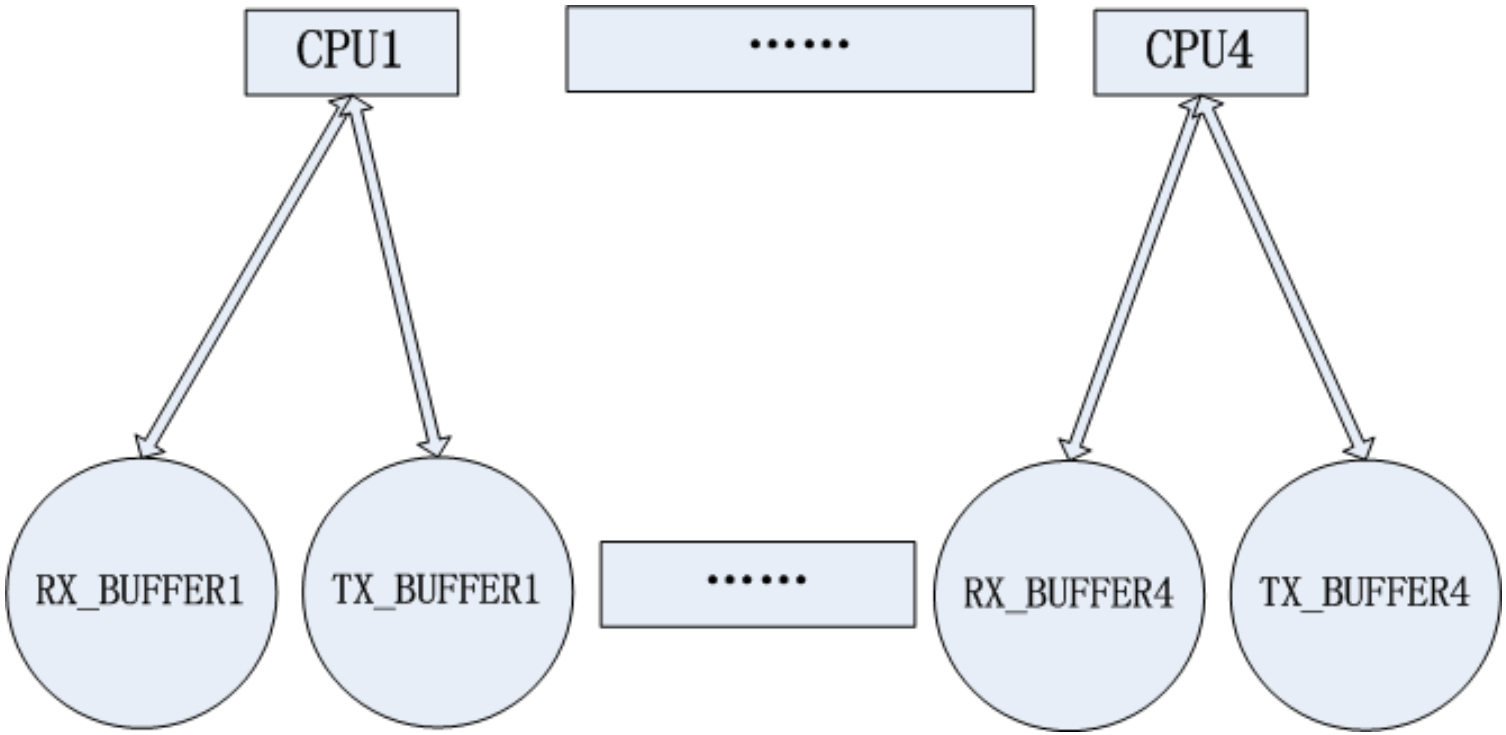
性能-多核多队列

- 单队列网卡
  - 只有一个rx_buffer和一个tx_buffer;

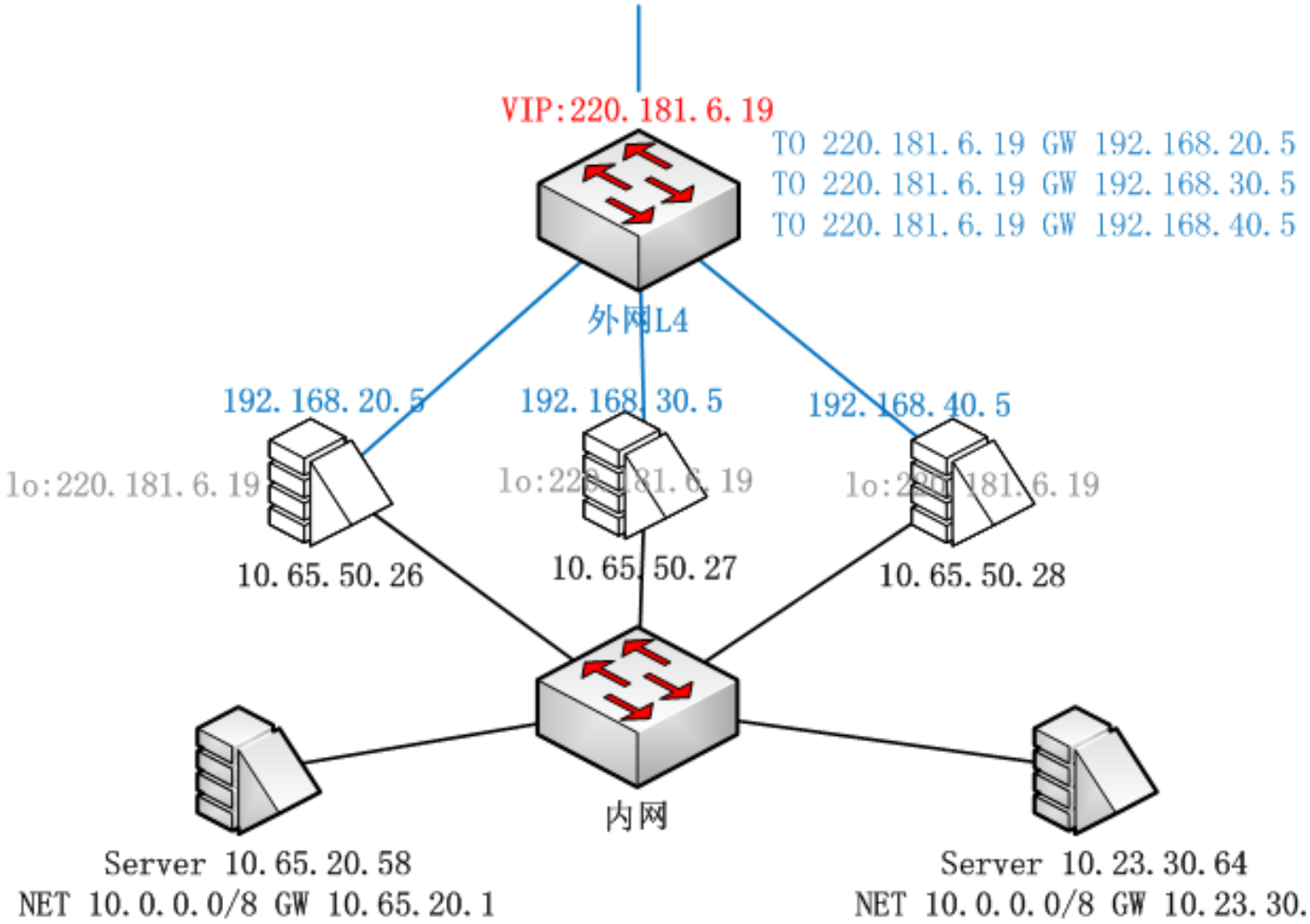


性能-多核多队列

- 多队列网卡
  - N个rx_buffer和N个tx_buffer, N=CPU核个数



性能-CLUSTER

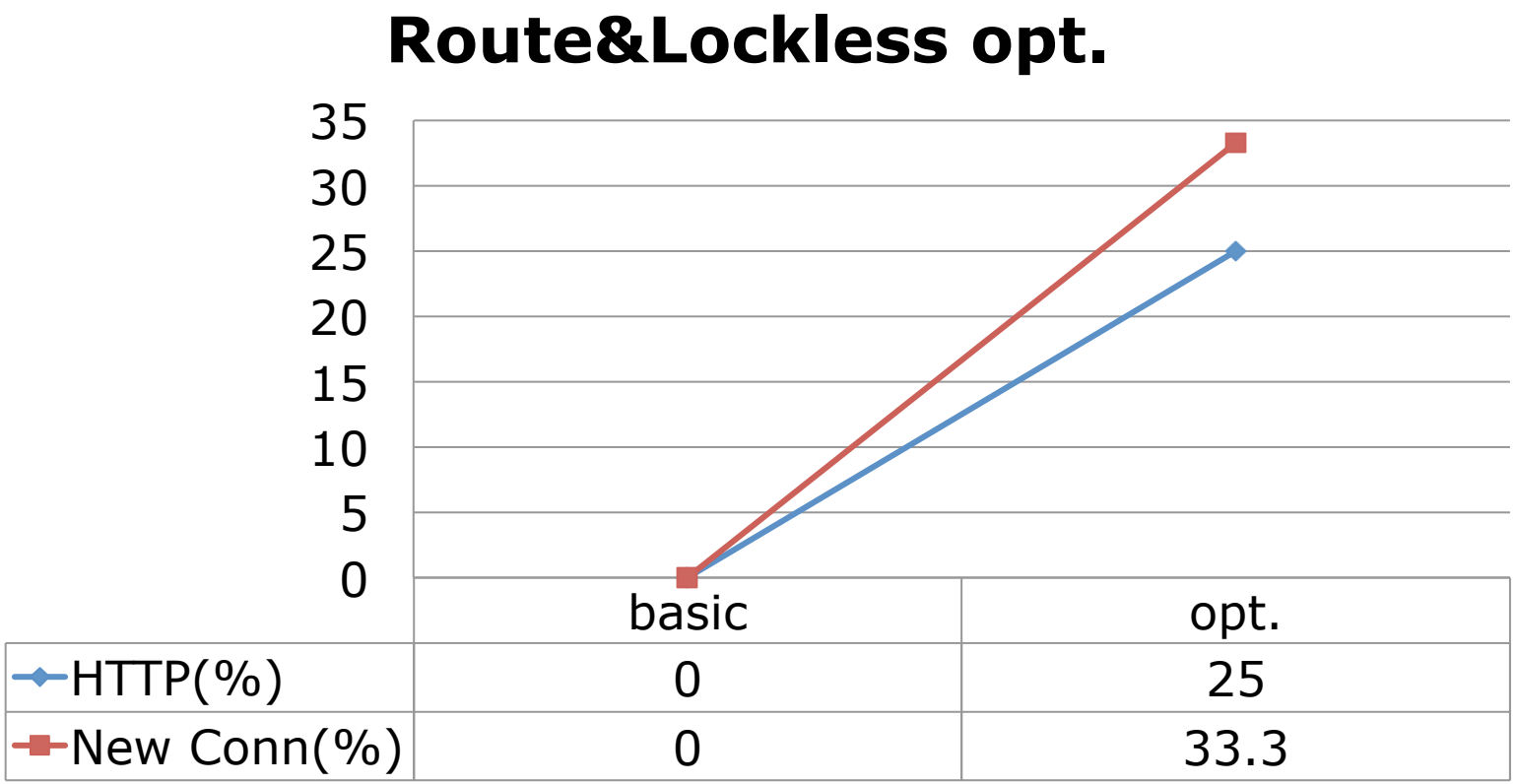


性能-软件优化

- LVS优化
  - 增大session hash table
  - 增大session hash bucket lock个数
  - Route opt.
  - Statistic Lockless

性能-软件优化

- 软件优化效果

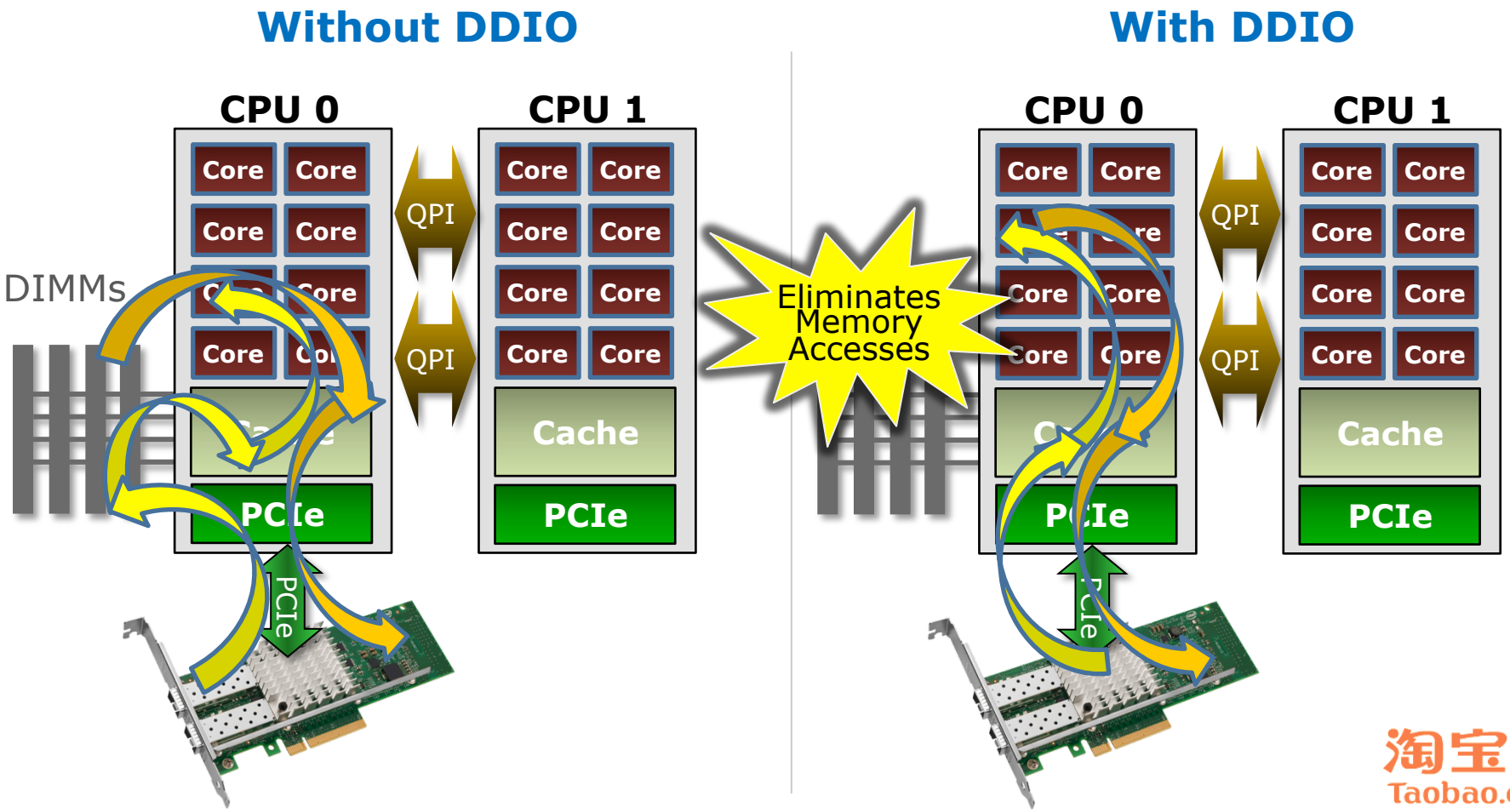


性能-硬件优化

- 硬件优化
  - Intel® E5-2600处理器（DDIO）
  - Synproxy cookies算法（AVX指令集）
  - BIOS设置优化

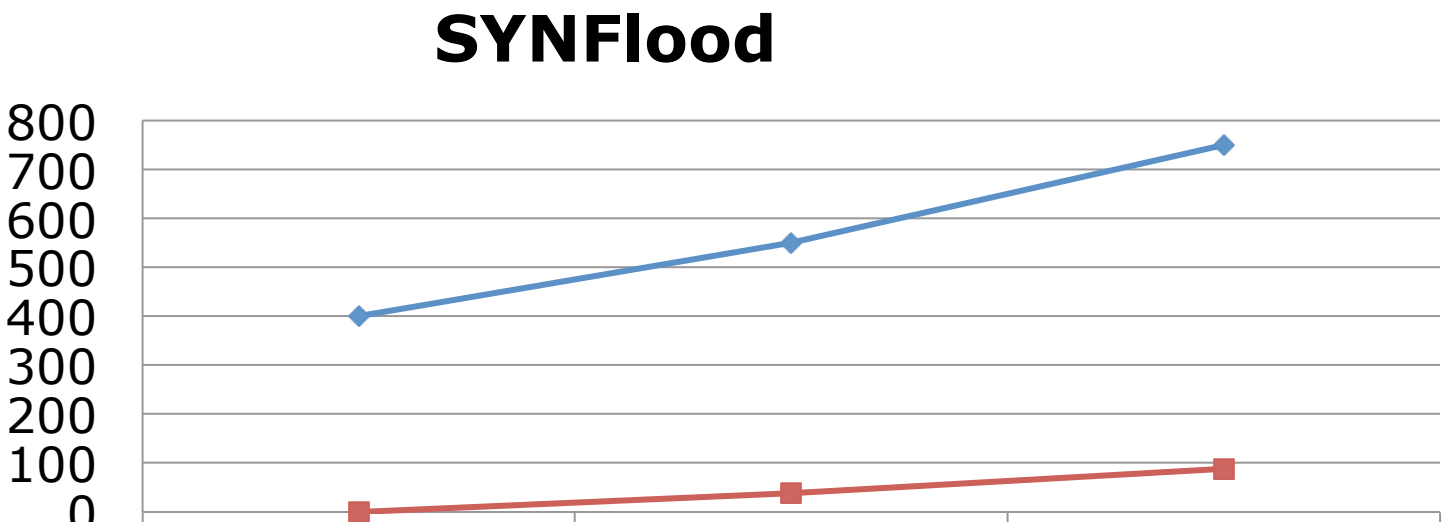
Intel ® E5-2600处理器(DDIO)介绍

- Intel E5 处理器平台 + Intel82599网卡
- 通过减少内存访问来提升系统IO性能



性能-硬件优化

- 硬件优化效果 – 性能提升87.5%



	westmere	sandybridge	sandybridge with AVX
数据包量(w)	400	550	750
百分比(%)	0	37.5	87.5

Todo List

- 性能：lockless
- 功能：7层；攻击防御；

- 开源： http://kb.linuxvirtualserver.org/wiki/IPVS_FULLNAT_and_SYNPROXY

谢 谢

Q&A

新浪微博：吴佳明_普空