淘宝网
Taobao.com

阿里LVS优化实践

阿里云事业部-核心系统研发-网络组
陈家军/莫凡

about me

- 花名莫凡
- 2011.7 毕业后入职淘宝，做过一段时间 taobao kernel维护工作，然后一直专注于系统网络
- 现在主要从事lvs定制化、性能优化工作

主要内容

- LVS简介
- LVS性能优化
- LVS功能增强
- 技术难点和体会

LVS简介

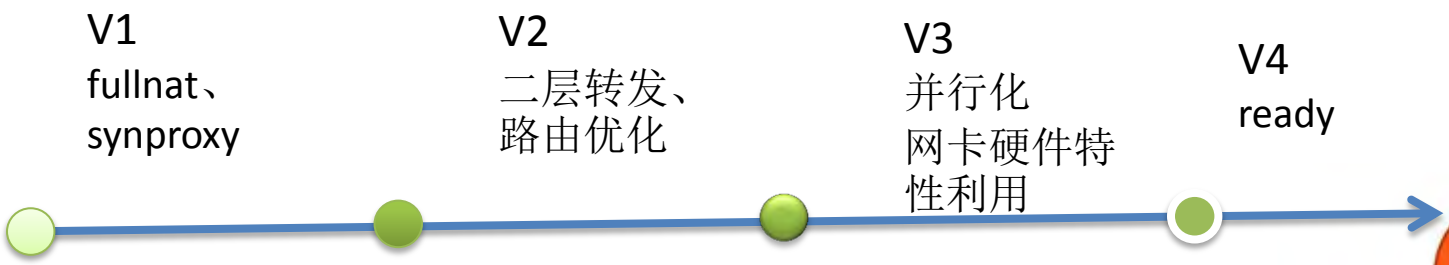
- LVS(Linux Virtual Server)，由章文嵩博士于1998年5月创立，是中国国内最早出现的自由软件项目之一
- lvs是内核标准组件，以内核模块的形式放在内核源码树net/netfilter/ipvs路径下

阿里LVS简介

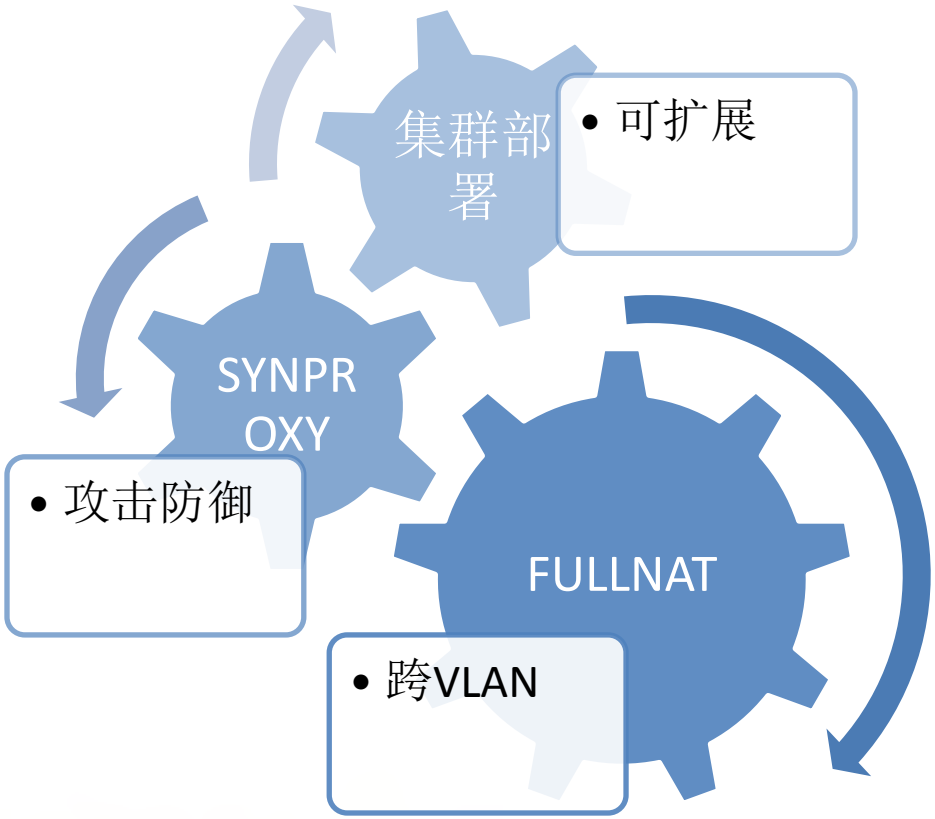
- 根据业务需求，定制淘宝自己的高性能LVS，节省成本，简化网络架构
- 现在发展成四层统一接入平台

一直在进化

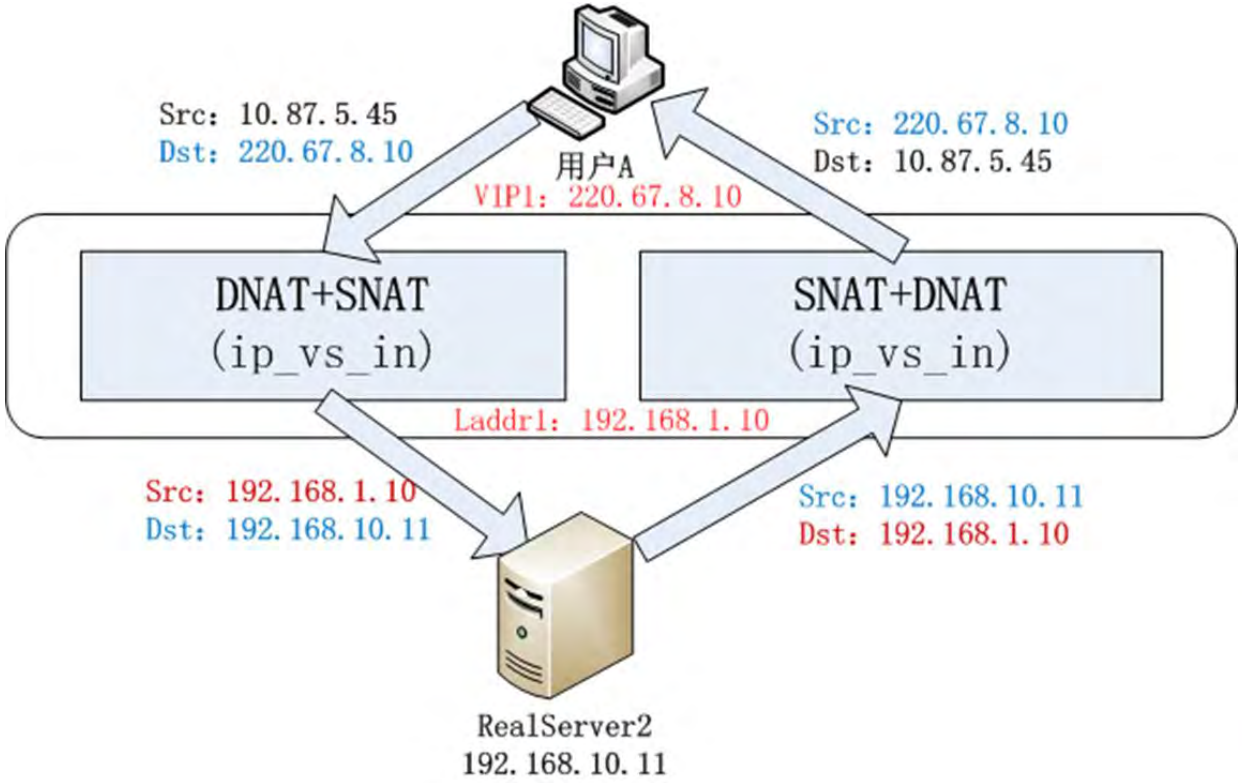
- v1实现基本功能fullnat、synproxy
- v2运用二层转发，解决了route cache问题。攻击防御性能提升一倍，正常转发提升20%+
- v3并行化改造，综合性能进一步提升一倍
- v4进一步提升性能近线速



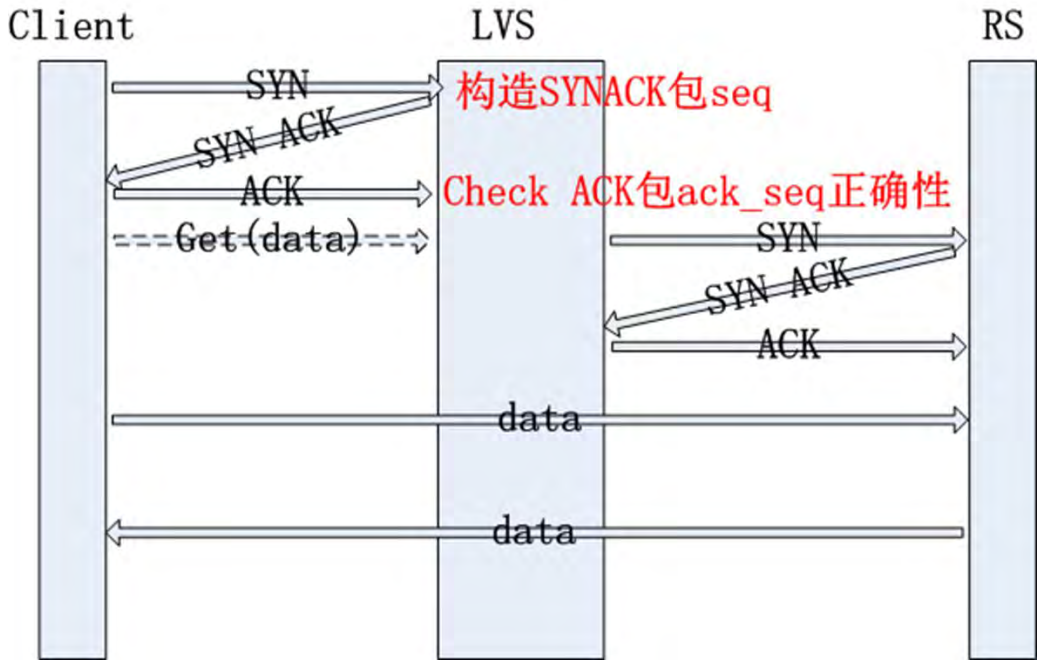
三个基本特点



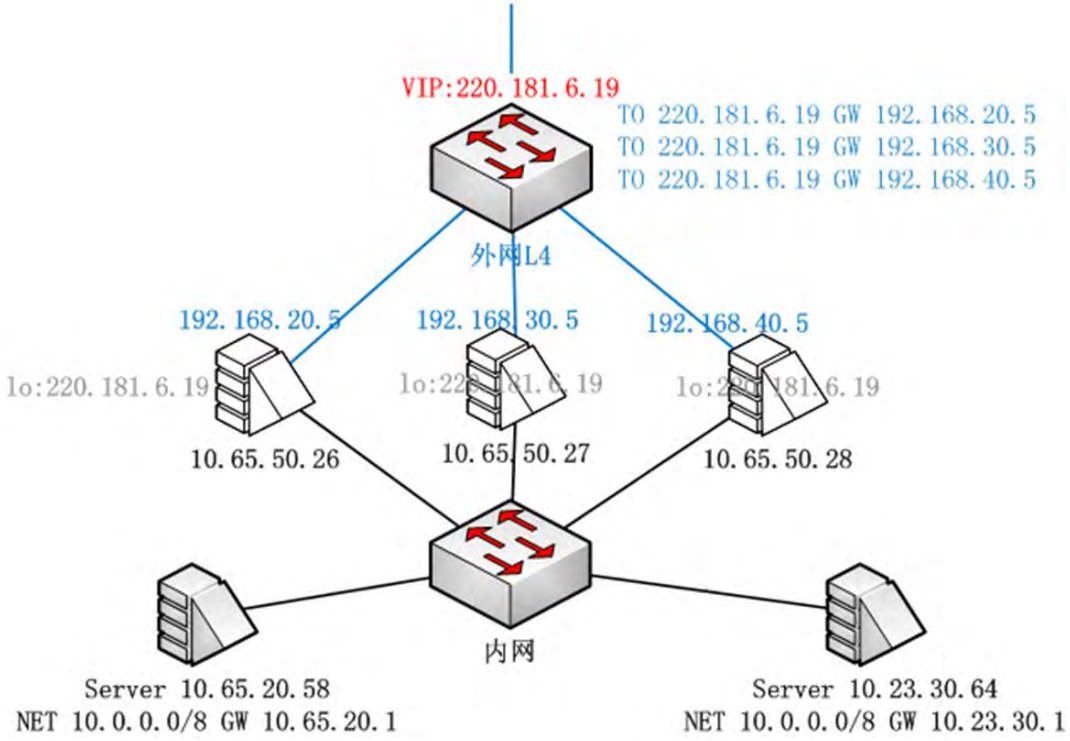
fullnat


Src：10.87.5.45
Dst：220.67.8.10
用户A
VIP1：220.67.8.10
Src：220.67.8.10
Dst：10.87.5.45
DNAT+SNAT
(ip_vs_in)
SNAT+DNAT
(ip_vs_in)
Laddr1：192.168.1.10
Src：192.168.1.10
Dst：192.168.10.11
Src：192.168.10.11
Dst：192.168.1.10
RealServer2
192.168.10.11

synproxy


Client
LVS
RS
SYN
构造SYNACK包seq
SYN ACK
ACK
Check ACK包ack_seq正确性
Get(data)
SYN
SYN ACK
ACK
data
data

集群部署

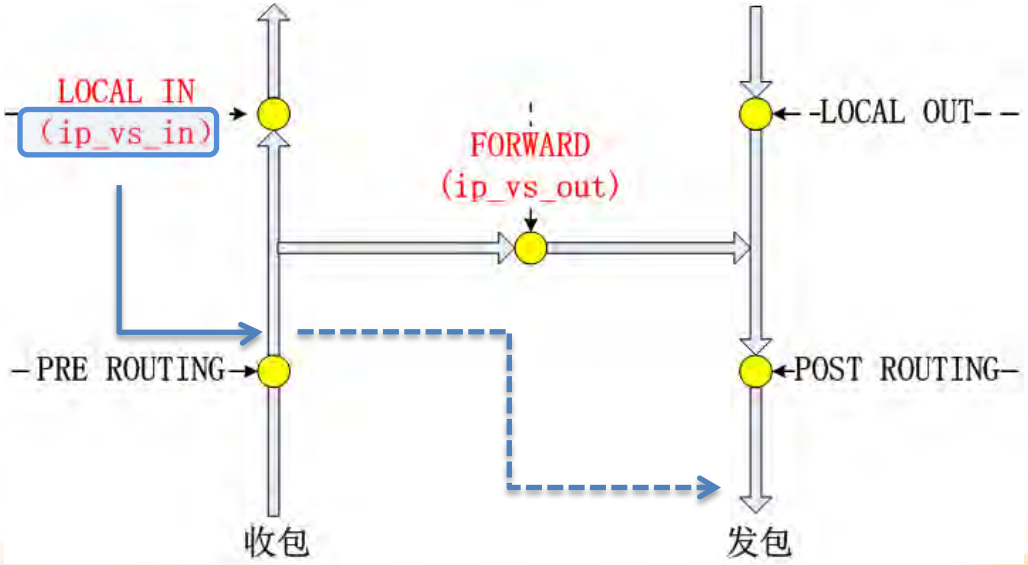

VIP:220.181.6.19
TO 220.181.6.19 GW 192.168.20.5
TO 220.181.6.19 GW 192.168.30.5
TO 220.181.6.19 GW 192.168.40.5
外网L4
192.168.20.5
192.168.30.5
192.168.40.5
lo:220.181.6.19
10.65.50.26
10.65.50.27
10.65.50.28
内网
Server 10.65.20.58
NET 10.0.0.0/8 GW 10.65.20.1
Server 10.23.30.64
NET 10.0.0.0/8 GW 10.23.30.1

LVS性能优化

- 路由优化
- 并行化改造
- synproxy
- 内存优化
- 系统参数

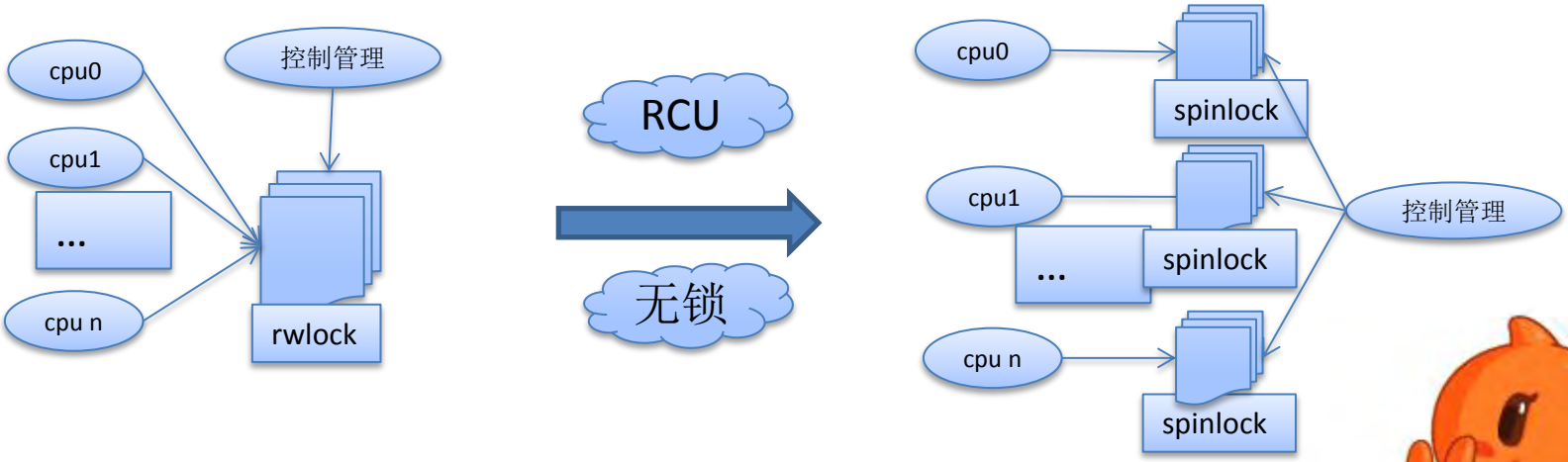
性能优化- 路由优化

- Hook点前移到prerouting
- 回包二层转发



性能优化- 并行化改造

- 上层业务逻辑改造
  - 消除CPU间资源竞争，每个核都有一份本地的数据

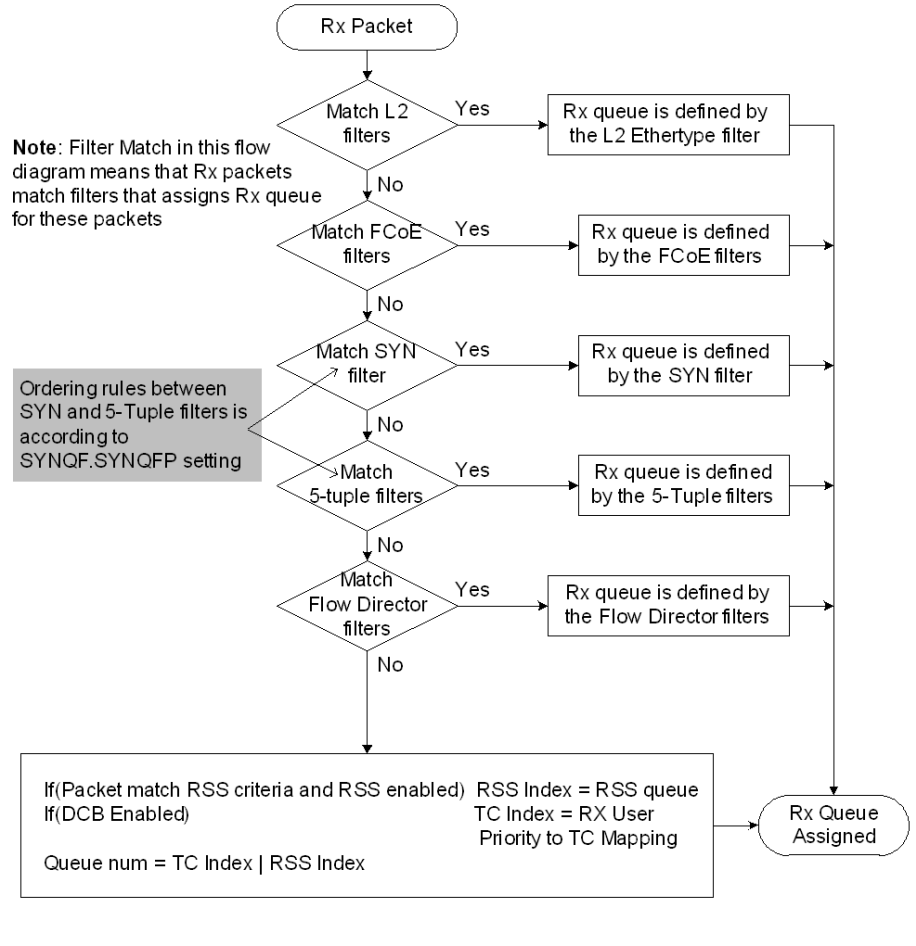


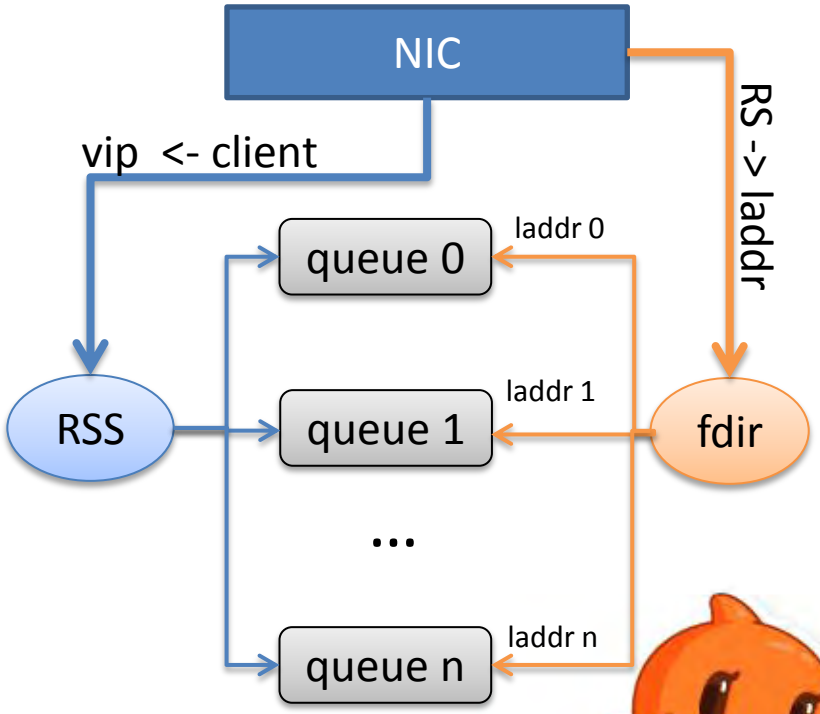
- 底层收发的并行

多队列网卡一个队列绑定一个核，哪个核收的包就从哪个核发出去

网卡特性Receive-Side Scaling (RSS)、Flow Director Filters

local addr (内网snat地址)划分到各个核上





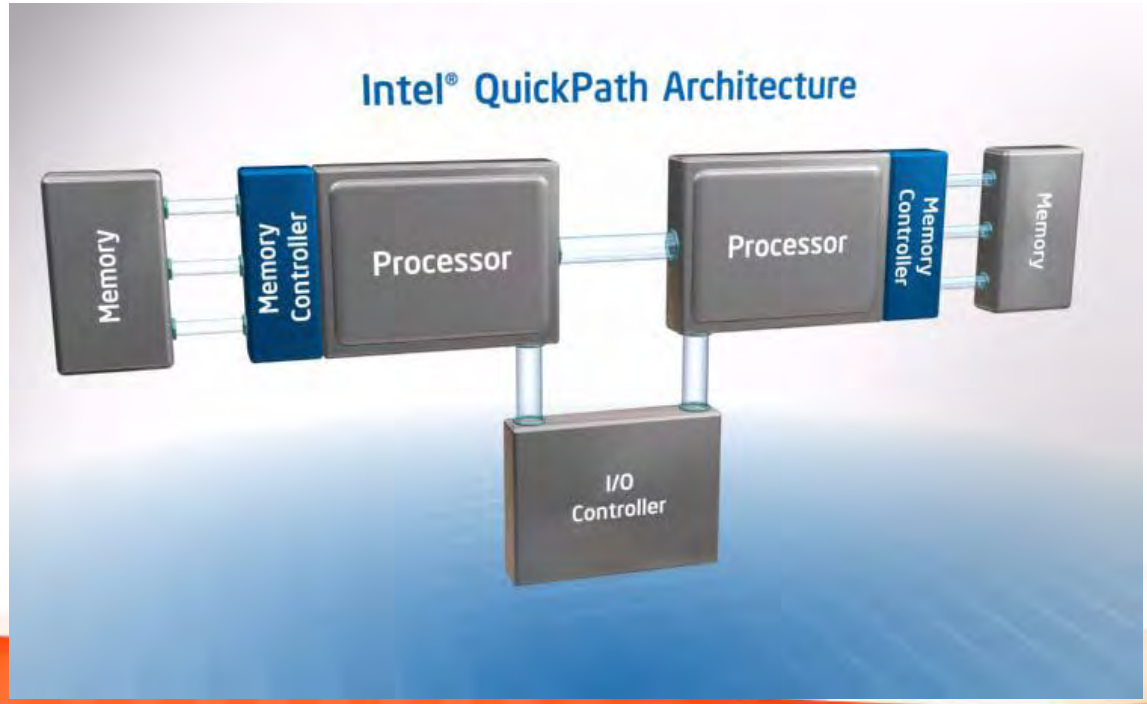
性能优化- synproxy

- 简化syncookie算法
- 更底层的收发接口

md5 vs crc32c(31s vs 4s round 200,000,000)
SND 2.3GHz 单核100% 190～210w pps)

性能优化- 内存优化

- numa，local node
- 数据结构cacheline



性能优化- 系统参数

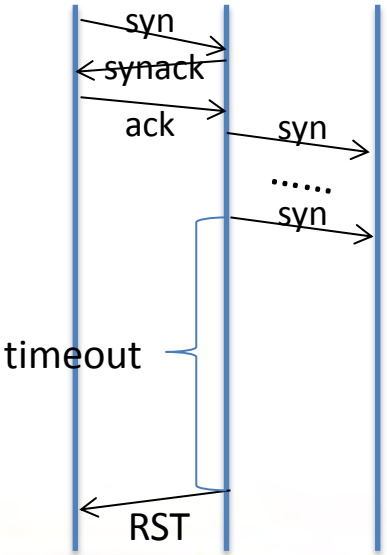
- Timer
  - kernel.timer_migration = 0
- Session同步
  - net.core.wmem_default=16777216
  - net.core.rmem_default=16777216
- 限制网卡中断速率
  - 例如ixgbe驱动模块参数 InterruptThrottleRate

LVS功能增强

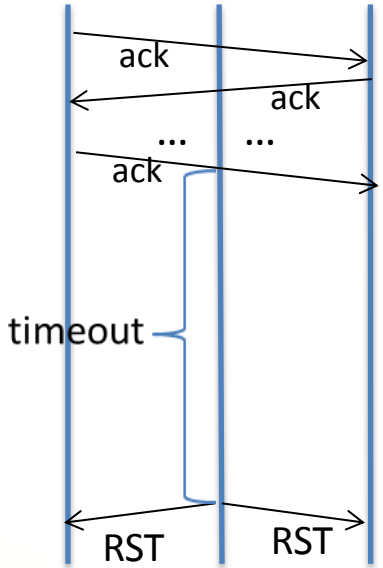
- reset功能
- timeout时间可调
- 数据流控制流分离

LVS功能增强-reset

- reset功能
  - 半连接和established



Synproxy半连接超时

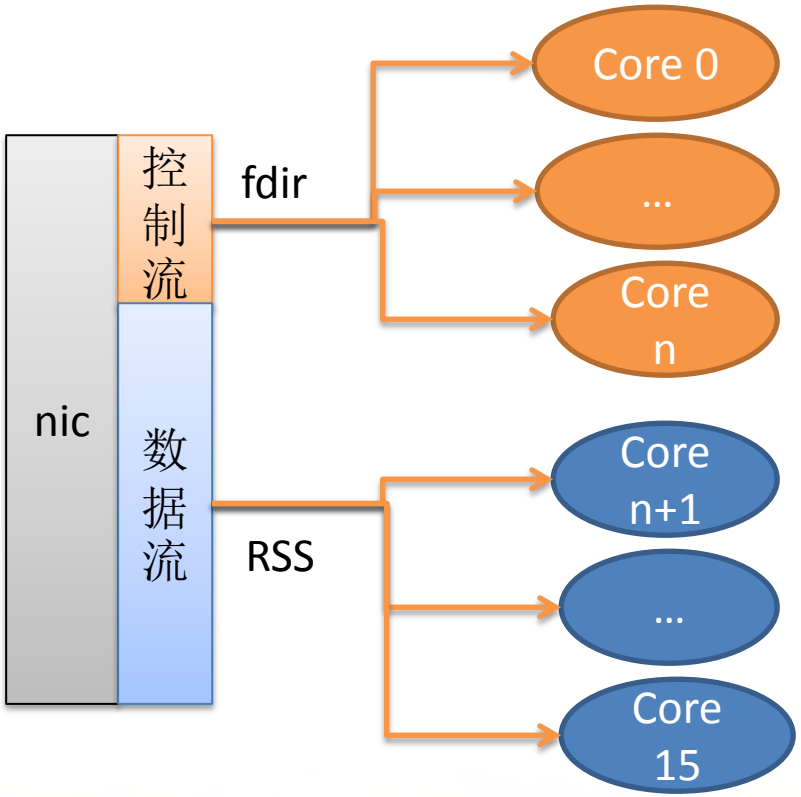


Established状态超时

LVS功能增强-timeout时间可调

- VS timeout时间可调
  - 全局的proc接口支持所有tcp状态超时时间调整
  - 每个vs都可设置自己不同的established timeout时间

LVS功能增强-数据流控制流分离



- 并行化的附产品
  - 空出部分硬件资源给控制管理使用
  - 例如将sshd绑在空出的核上

技术难点和体会

- 心得体会 
- 抛几个问题 

心得体会

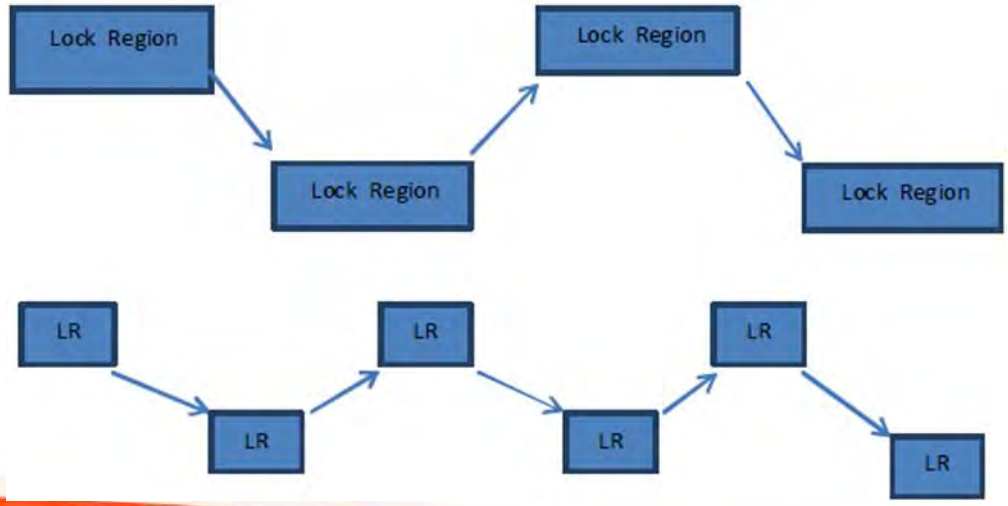
- 性能优化基本思路
  - profiling定量测试探寻瓶颈点
  - 软硬件结合优化
- 可靠稳定压倒一切
  - 功能模块有独立的开关控制启停
  - 统计计数反应运行状态
- 注意细节、异常情况的处理
  - 乱序、丢包、ICMP 、ipfrag的处理

大跌眼镜的性能退化

现象：同样的代码放到新的SND平台上性能下降大约25% (WSM vs SNB)

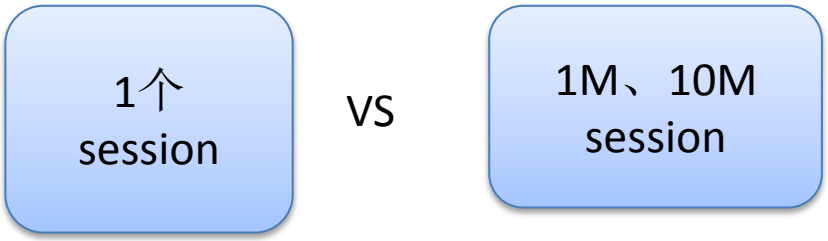
定位：通过一块块剥离代码定量测试排查最终锁定目标为rwlock的一把读锁上

收获：理论跟实际有一定的差异；通常都认为lock的粒度越小越好，这个场景却发现lock在各个核频繁切换导致性能下降



http://www.intel.com/content/www/us/en/processors/xeon/xeon-lock-scaling-analysis-paper.html

session查找

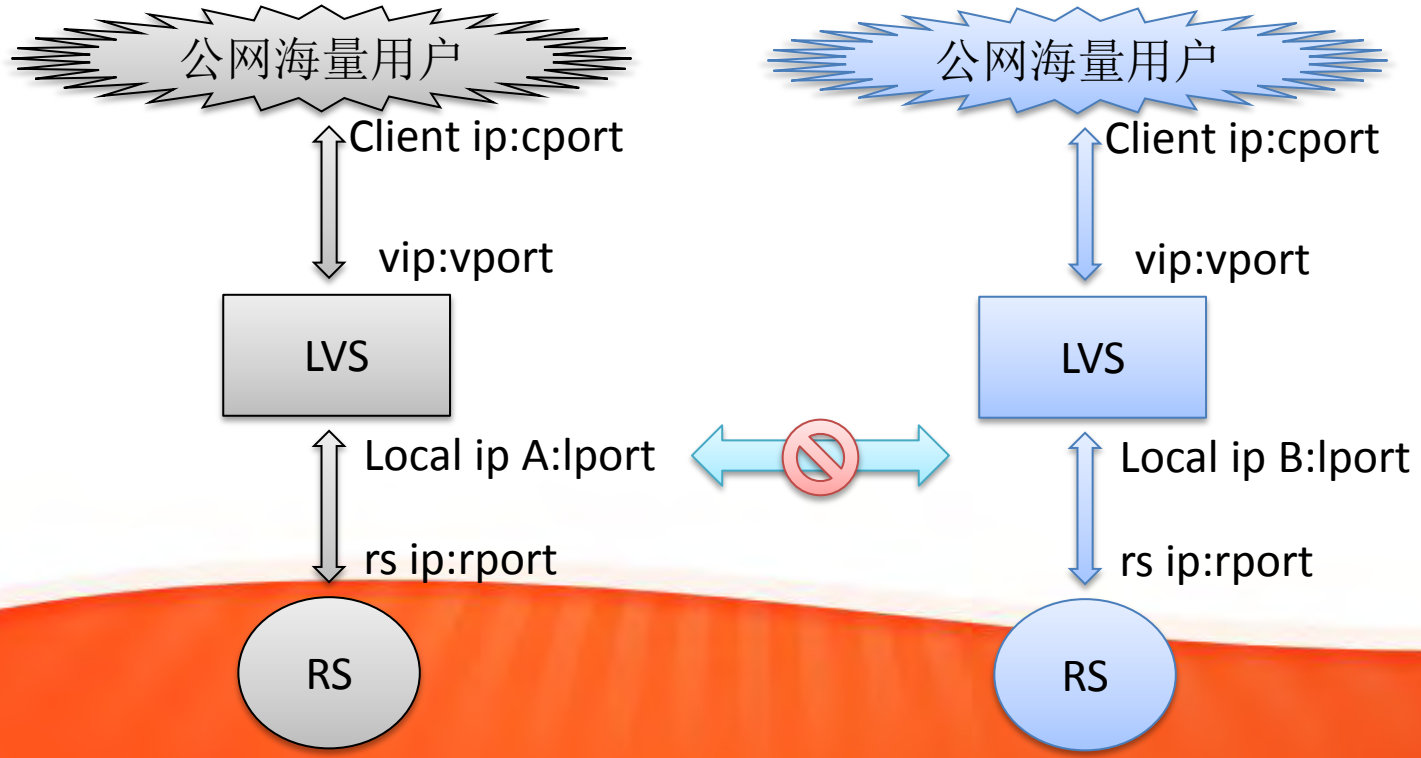


session查找消耗随并发连接数的增加而增加

- LVS流表是用hash table管理，这层面hash值越离散冲突越低越好
- 另一方面访存随机导致cpu cache misses增加

Fullnat集群session同步问题

- 集群中各台lvs配置的local ip都不同
- 有没有优美的解决方法



参考资料

- http://www.intel.com/content/dam/doc/white-paper/quick-path-interconnect-introduction-paper.pdf
- http://www.intel.com/content/dam/www/public/us/en/documents/datasheets/82599-10-gbe-controller-datasheet.pdf

淘宝网
Taobao.com

谢谢

Q&A