OSPF-LVS 测试性能报告 v1

为什么要使用 OSPF？

目前主流的 LVS+keepalived 瓶颈在于调度器的压力。使用 NAT 模式，进出包受到 LVS 最大转发限制。使用 DR 模式调度器收包受到 LVS 最大转发限制。

为了解决上面这些问题，所以我们开始尝试 LVS（DR）通过 OSPF，做 lvs 集群，实现一个 VIP，多台 LVS 同时工作提供服务，不存在热备机器。提高至少相当于 LVS 的 5 倍以上转发性能。

测试交换机型号：

cisco WS-C3750G-24TS 24 口　全千兆

简介

OSPF(Open Shortest Path First 开放式最短路径优先）是一个内部网关协议(Interior Gateway Protocol，简称 IGP)，用于在单一自治系统（autonomous system,AS）内决策路由。是对链路状态路由协议的一种实现，隶属内部网关协议（IGP)，故运作于自治系统内部。著名的迪克斯加算法(Dijkstra)被用来计算最短路径树。OSPF 分为 OSPFv2 和 OSPFv3 两个版本,其中 OSPFv2 用在 IPv4 网络，OSPFv3 用在 IPv6 网络。OSPFv2 是由 RFC 2328 定义的，OSPFv3 是由 RFC 5340 定义的。与 RIP 相比，OSPF 是链路状态协议，而 RIP 是距离矢量协议。

ECMP（Equal-CostMultipathRouting）等价多路径，存在多条不同链路到达同一目的地址的网络环境中，如果使用传统的路由技术，发往该目的地址的数据包只能利用其中的一条链路，其它链路处于备份状态或无效状态，并且在动态路由环境下相互的切换需要一定时间，而等值多路径路由协议可以在该网络环境下同时使用多条链路，不仅增加了传输带宽，并且可以无时延无丢包地备份失效链路的数据传输。

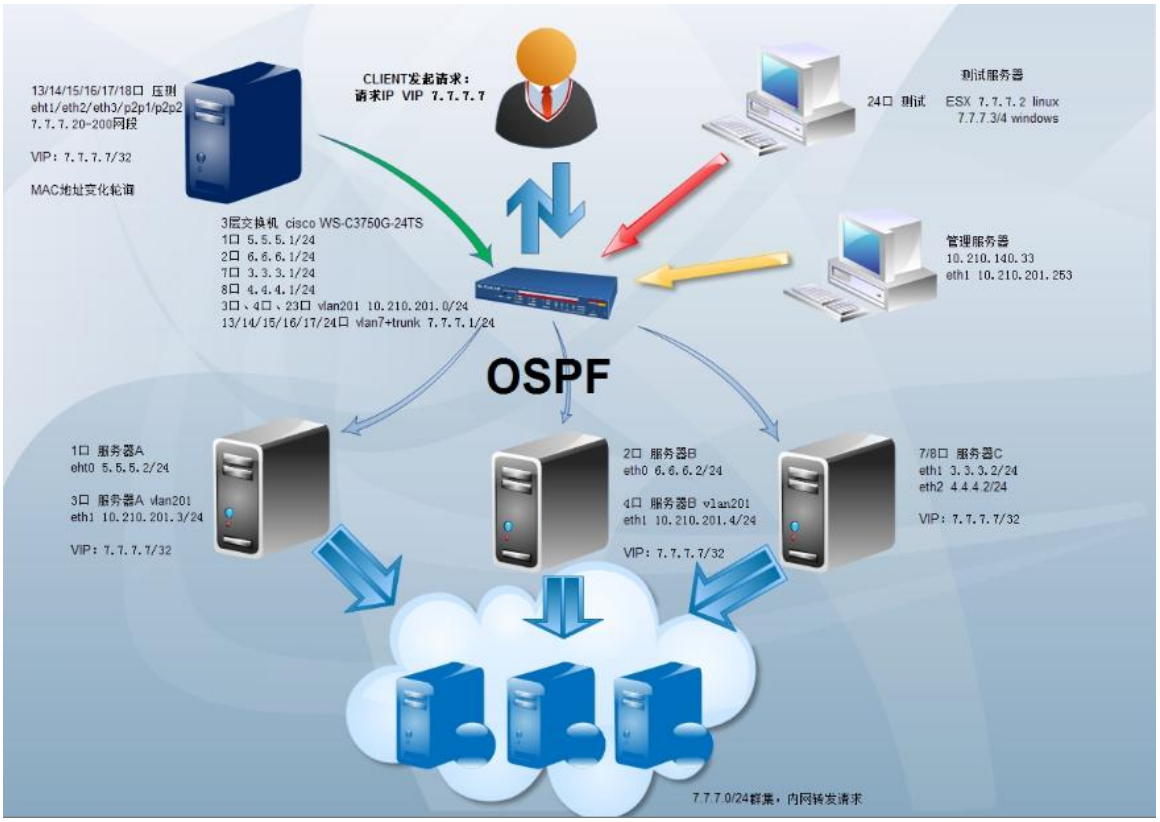
访问过程

用户请求（VIP：7.7.7.7）到达三层交换机之后，通过对原地址、端口和目的地址、端口的 hash，将链接分配到集群中的某一台 LVS 上，LVS 通过内网向后端转发请求，后端再将数据返回给用户，整个会话完成。

测试环境

网段	交换机 IP	交换端口	交换属性	服务器 ip	服务器网口	VIP	属性	COST
3.3.3.0/24	3.3.3.1	7	route	3.3.3.2	eth1	7.7.7.7	OSPF	2
4.4.4.0/24	4.4.4.1	8	route	4.4.4.2	eth2	7.7.7.7	OSPF	2
5.5.5.0/24	5.5.5.1	1	route	5.5.5.2	eth0	7.7.7.7	OSPF	2
6.6.6.0/24	6.6.6.1	2	route	6.6.6.2	eth0	7.7.7.7	OSPF	2
7.7.7.0/32	7.7.7.1	13/14/15/16/17	vlan7 trunk	7.7.7.20-200	eth1/eth2/eth3/p2p1/p2p2			

访问图解



部署说明

1、OSPF 下属服务器安装 quagga 软件，用于学习路由。

```
router ospf
 ospf router-id 3.3.3.2
 network 3.3.3.2/24 area 0.0.0.0
 network 4.4.4.2/24 area 0.0.0.0
 network 5.5.5.2/24 area 0.0.0.0
 network 6.6.6.2/24 area 0.0.0.0
 network 7.7.7.7/32 area 0.0.0.0
```

2、设置交换机需要 OSPF 端口启动 route 功能，并配置 IP 和 OSPF 学习网段。

```
router ospf 100
 log-adjacency-changes
 network 3.3.3.0 0.0.0.255 area 0
 network 4.4.4.0 0.0.0.255 area 0
 network 5.5.5.0 0.0.0.255 area 0
 network 6.6.6.0 0.0.0.255 area 0
 network 7.7.7.0 0.0.0.255 area 0
```

```
r0#sh ip ospf neighbor

Neighbor ID     Pri   State           Dead Time   Address         Interface
3.3.3.2           1   FULL/BDR        00:00:39    4.4.4.2         GigabitEthernet3/0/8
3.3.3.2           1   FULL/BDR        00:00:34    3.3.3.2         GigabitEthernet3/0/7
6.6.6.2           1   FULL/BDR        00:00:30    6.6.6.2         GigabitEthernet3/0/2
5.5.5.2           1   FULL/BDR        00:00:34    5.5.5.2         GigabitEthernet3/0/1
```

```
r0#sh ip route ospf
     7.0.0.0/8 is variably subnetted, 2 subnets, 2 masks
O       7.7.7.7/32 [110/12] via 6.6.6.2, 00:08:47, GigabitEthernet3/0/2
                   [110/12] via 5.5.5.2, 00:08:47, GigabitEthernet3/0/1
                   [110/12] via 4.4.4.2, 00:08:47, GigabitEthernet3/0/8
                   [110/12] via 3.3.3.2, 00:08:47, GigabitEthernet3/0/7
```

测试工具

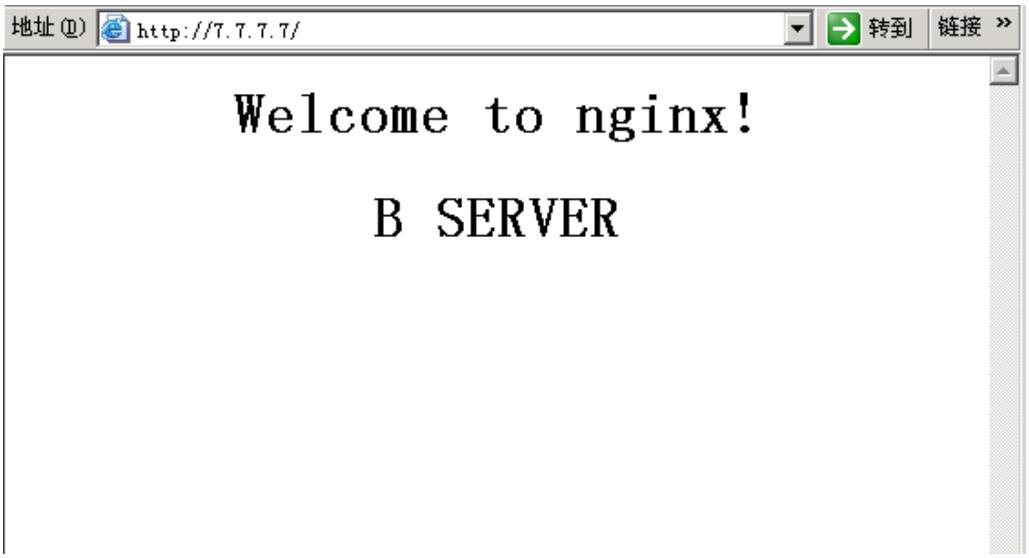
1、 Pktgen 多线程内核级压测工具

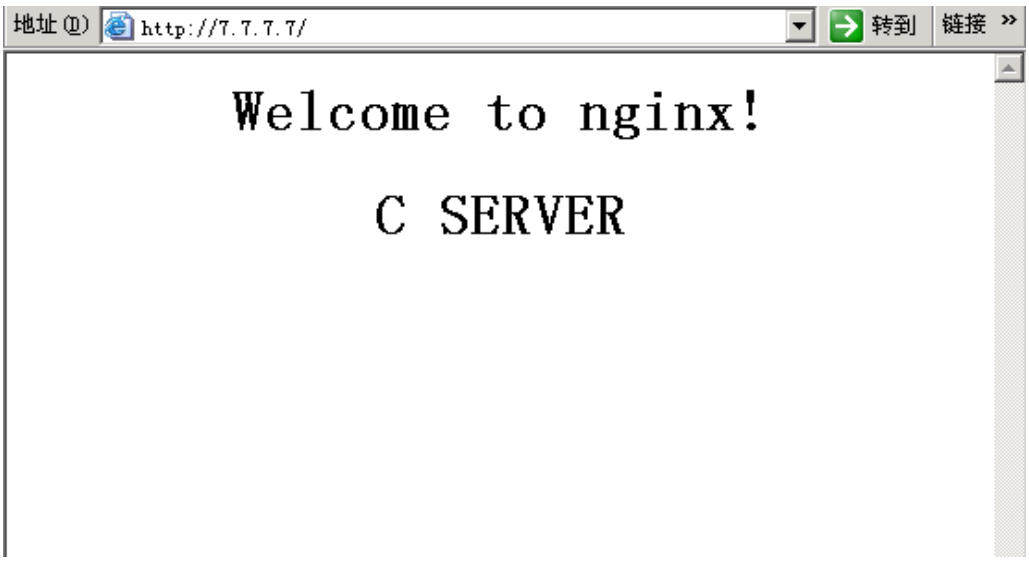
```
top - 15:12:19 up 7 days,  1:19,  2 users,  load average: 3.68, 2.35, 2.03
Tasks: 240 total,   4 running, 236 sleeping,   0 stopped,   0 zombie
Cpu0  :  0.0%us,  0.0%sy,  0.0%ni, 96.2%id,  0.0%wa,  0.0%hi,  3.8%si,  0.0%st
Cpu1  :  0.0%us,100.0%sy,  0.0%ni,  0.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu2  :  0.0%us, 99.7%sy,  0.0%ni,  0.0%id,  0.0%wa,  0.0%hi,  0.3%si,  0.0%st
Cpu3  :  0.0%us,  0.3%sy,  0.0%ni, 99.7%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu4  :  0.0%us, 99.7%sy,  0.0%ni,  0.0%id,  0.0%wa,  0.0%hi,  0.3%si,  0.0%st
Cpu5  :  0.0%us,100.0%sy,  0.0%ni,  0.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu6  :  0.0%us,  0.0%sy,  0.0%ni, 96.1%id,  0.0%wa,  0.0%hi,  3.9%si,  0.0%st
Cpu7  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu8  :  0.0%us,  0.0%sy,  0.0%ni, 99.6%id,  0.0%wa,  0.0%hi,  0.4%si,  0.0%st
Cpu9  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu10 :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu11 :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  16280412k total,   893976k used, 15386436k free,   288156k buffers
Swap:  8388600k total,        0k used,  8388600k free,   302444k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
49124 root      20   0     0    0    0 R 100.0  0.0   1:42.01 kpktgend_1
49127 root      20   0     0    0    0 S 100.0  0.0   1:42.01 kpktgend_4
49125 root      20   0     0    0    0 R 99.7  0.0   1:41.91 kpktgend_2
49128 root      20   0     0    0    0 R 99.7  0.0   1:42.01 kpktgend_5
```

2、 expect 自动登录交换机采集数据命令

测试数据展示

1、部署 nginx 展示，负载模式正常，关闭其中一台 VIP，刷新 2 秒内调转到存活路由的对应的主机

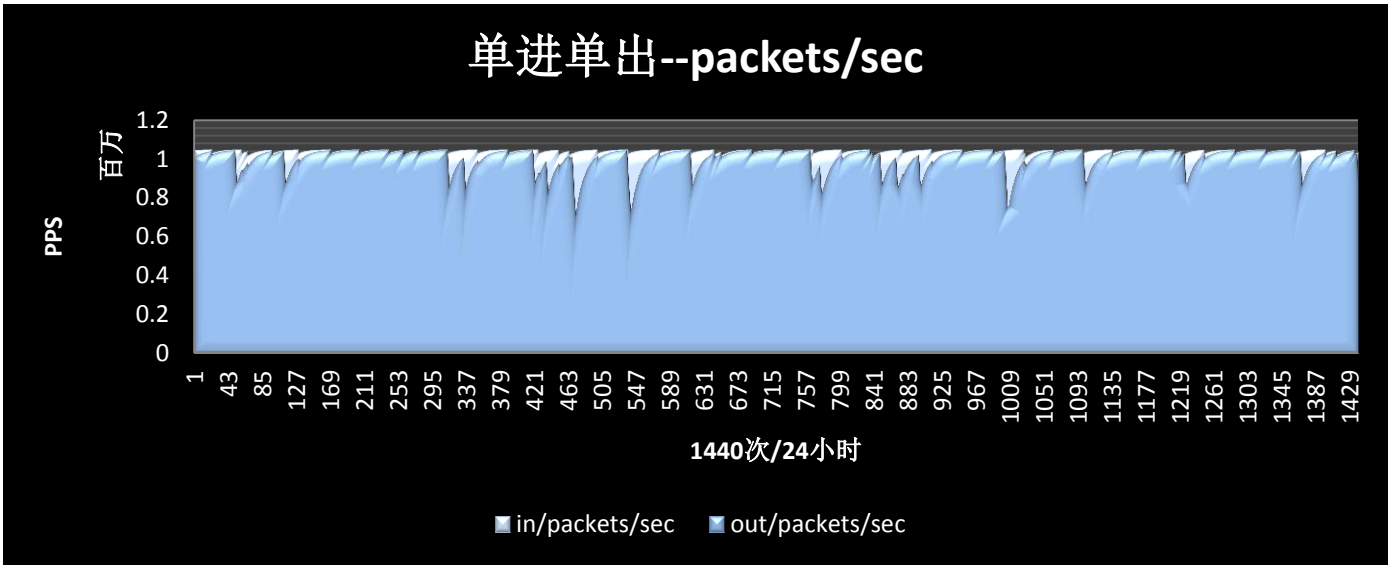




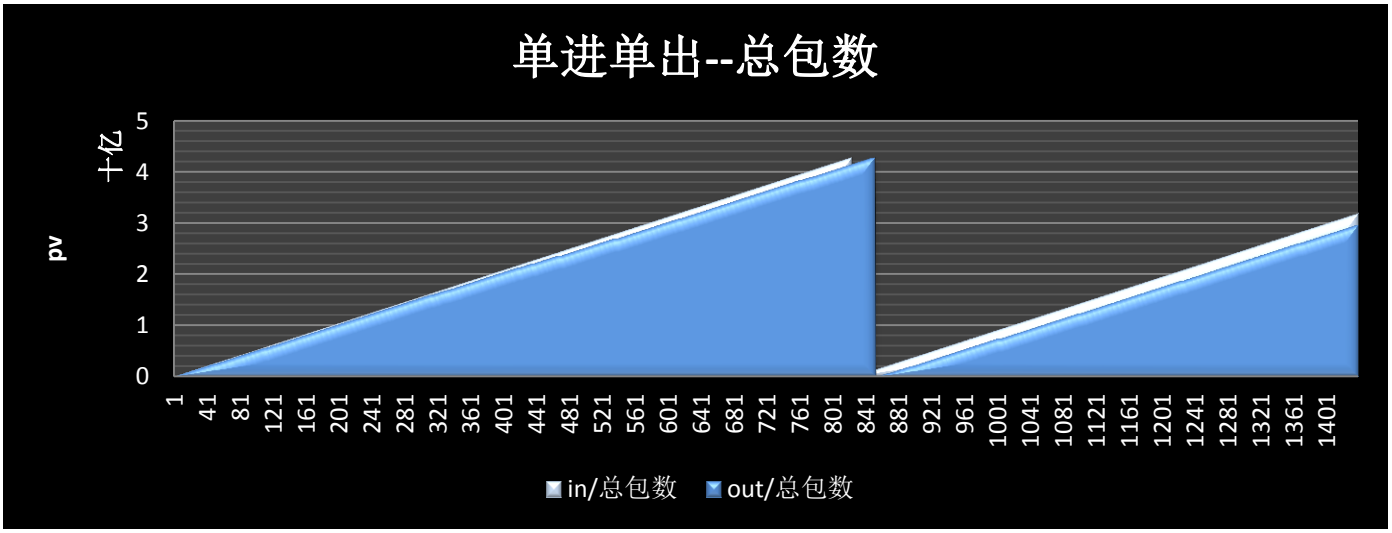
3、 数据（数据采集 24 小时，1 次/60s）

看图说明：in/out 进/出　packets/sec 秒包数　in/总包数 out/总包数 进出总包数　mem/used/b 用内存 M

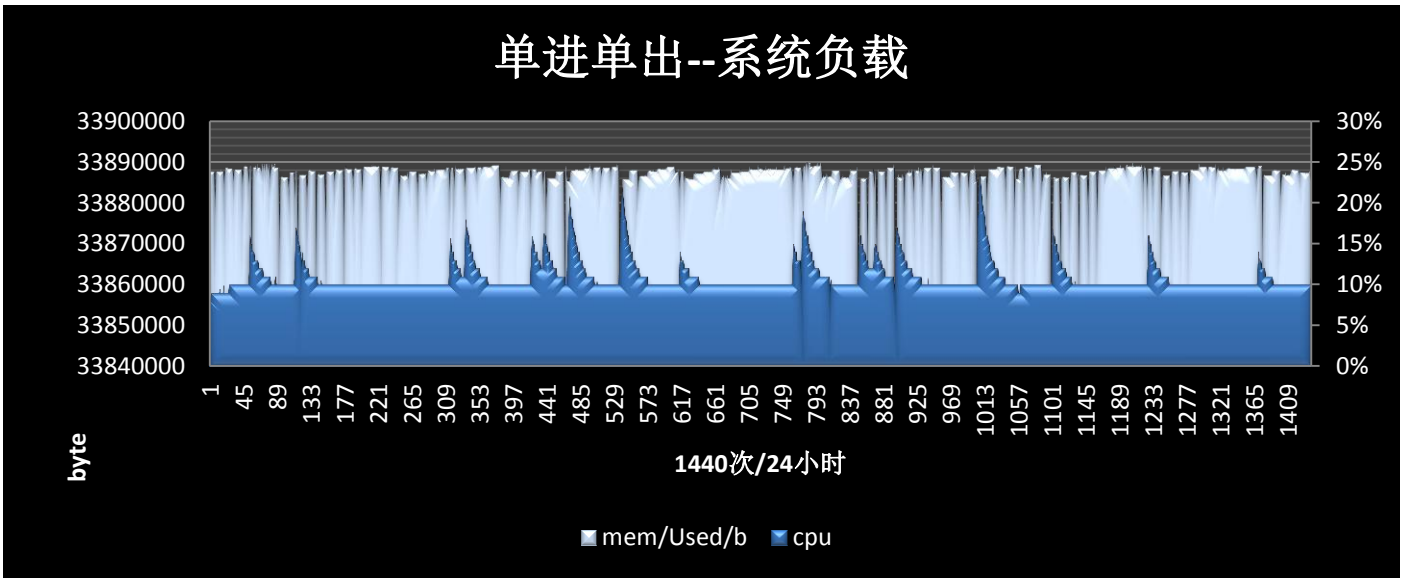
交换机端口说明：1/2/7/8 为压力进入口　13/14/16/17 为 ospf 压力输出口



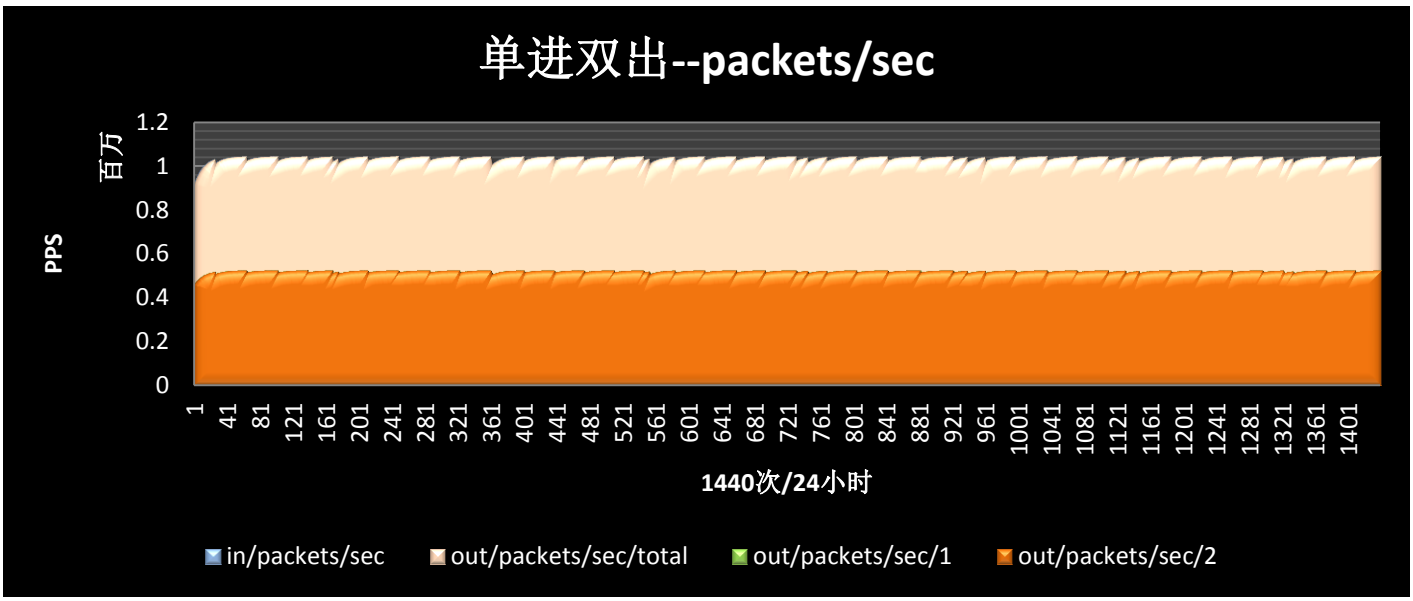
阐述:最高差值：瞬间最大差值 20 万。13 口进入 1 口流出，进口包数高于出口包数。原因：第一、千兆网口满载进入，通过计算分配，哈希到达出口。第二、交换机自身性能，自动调整进出负载平衡。第三、接收端口服务器网卡性能问题。



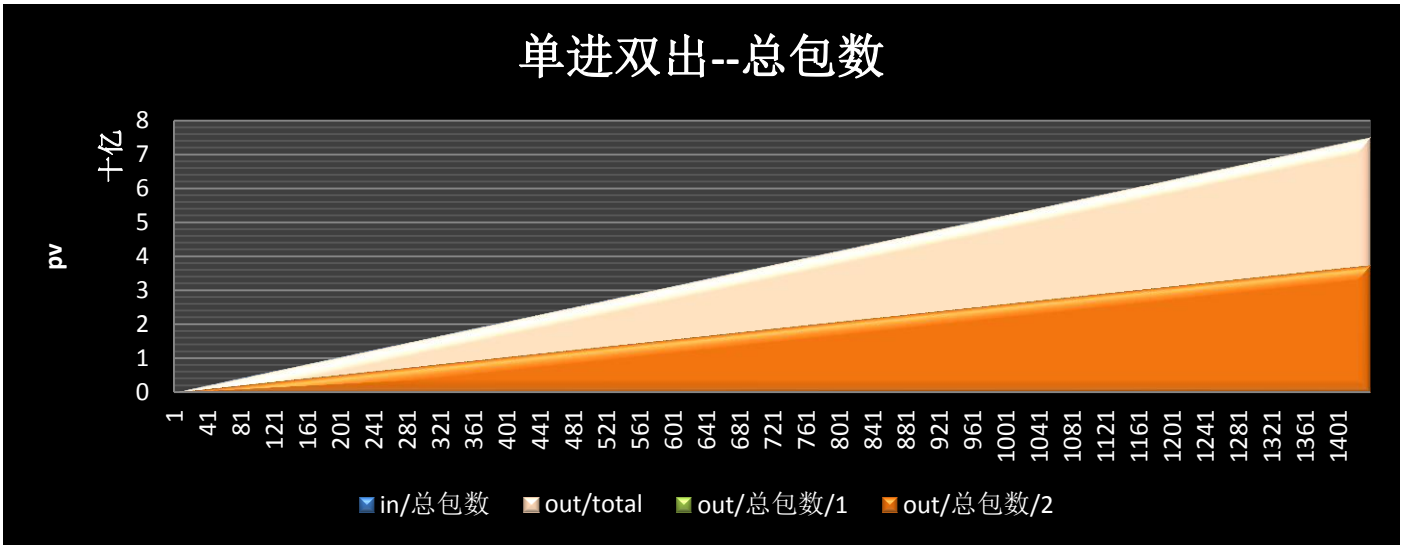
阐述：2 的 32 次方，32 位无符号整型来计数，4291293796 就会溢出。



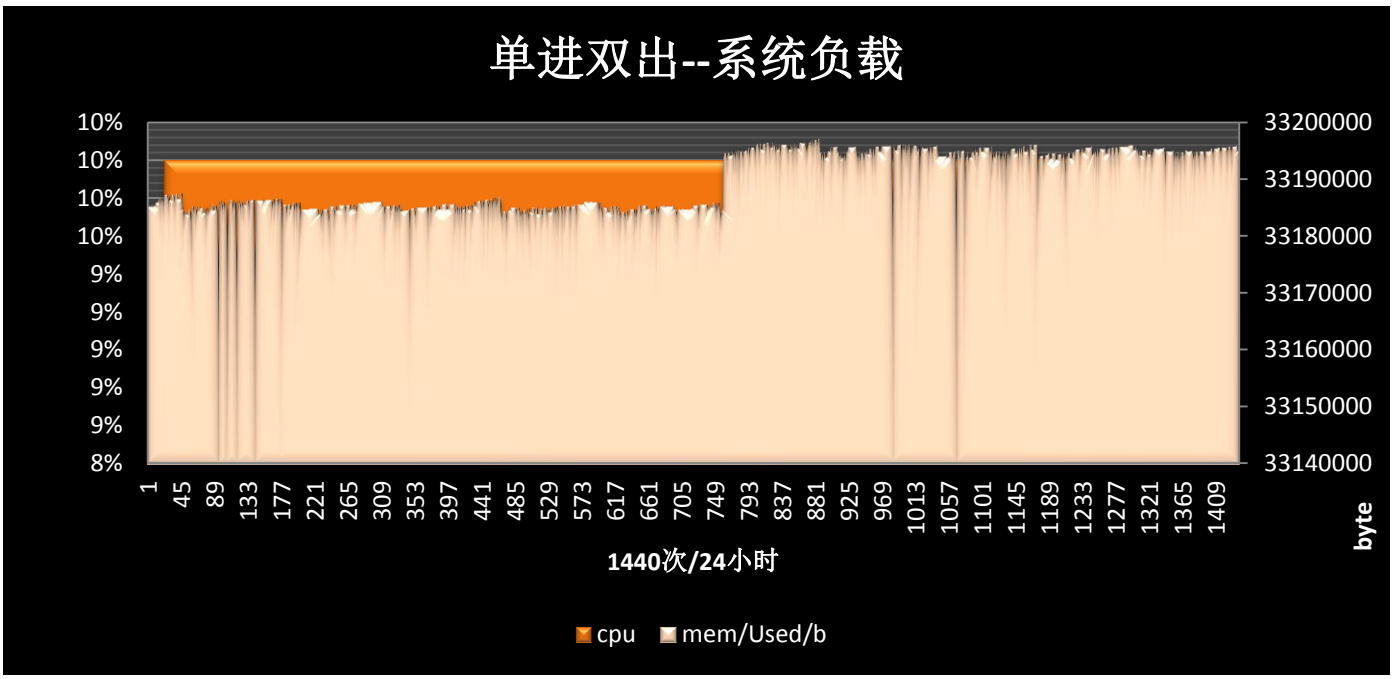
阐述：正常



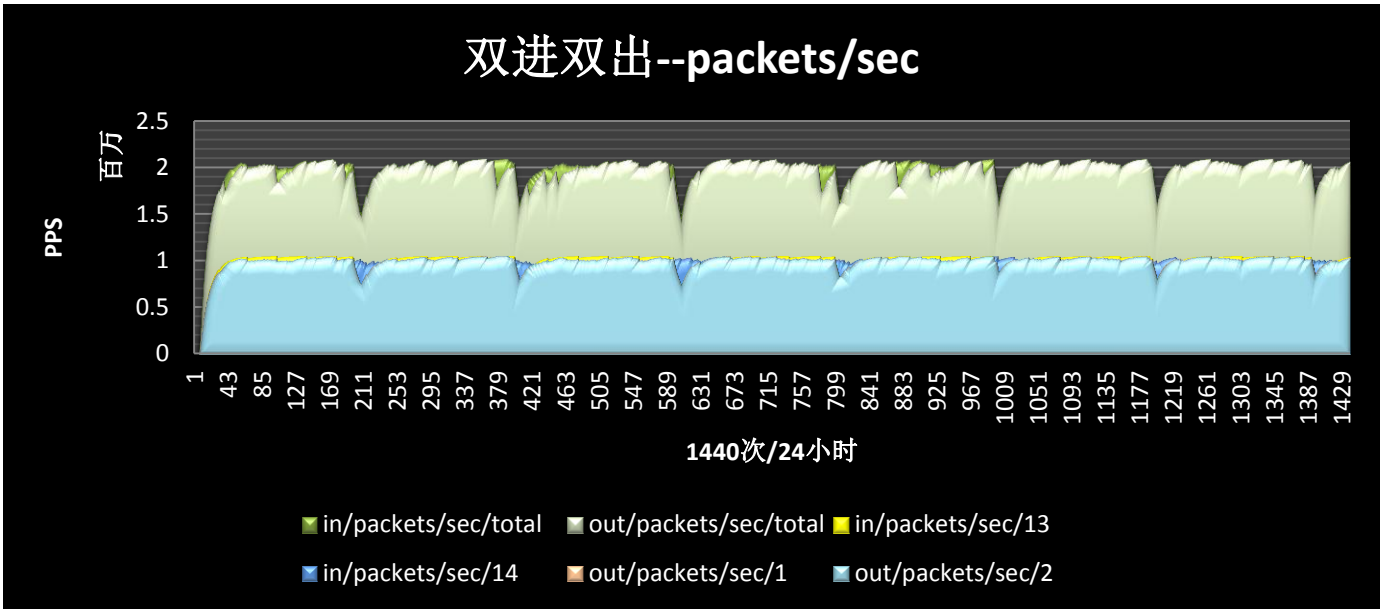
阐述：单进双出 100 万 pps，就不存在一些高峰低谷的问题了。



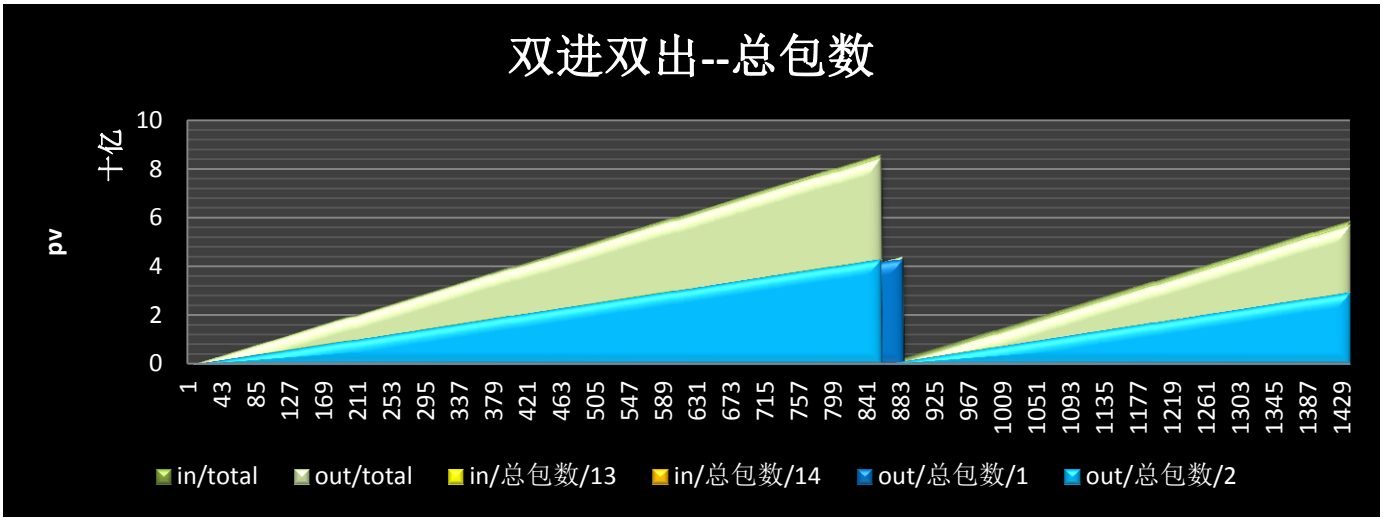
阐述：很稳定，因为双口计算 outpps，所以到达 4291293796*2 就会溢出清零。



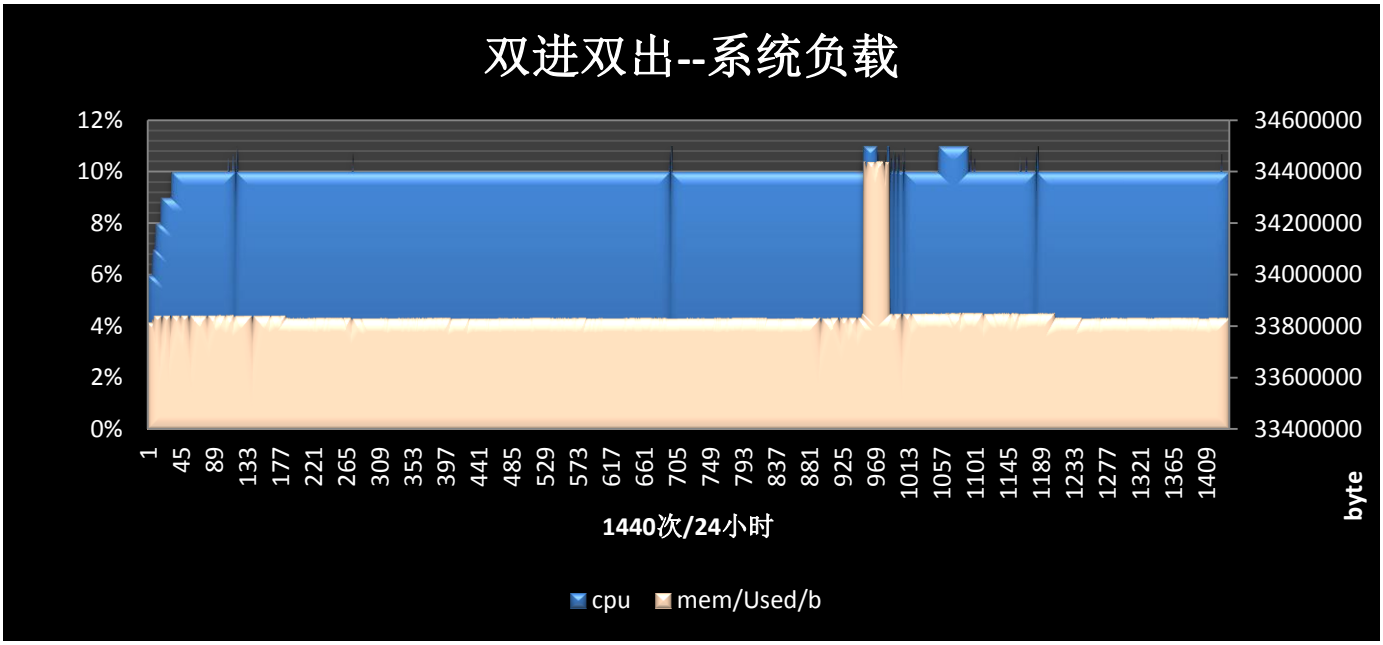
阐述：cpu 始终保持 9%-10%之间。内存峰值约 33M。内存总计 100M



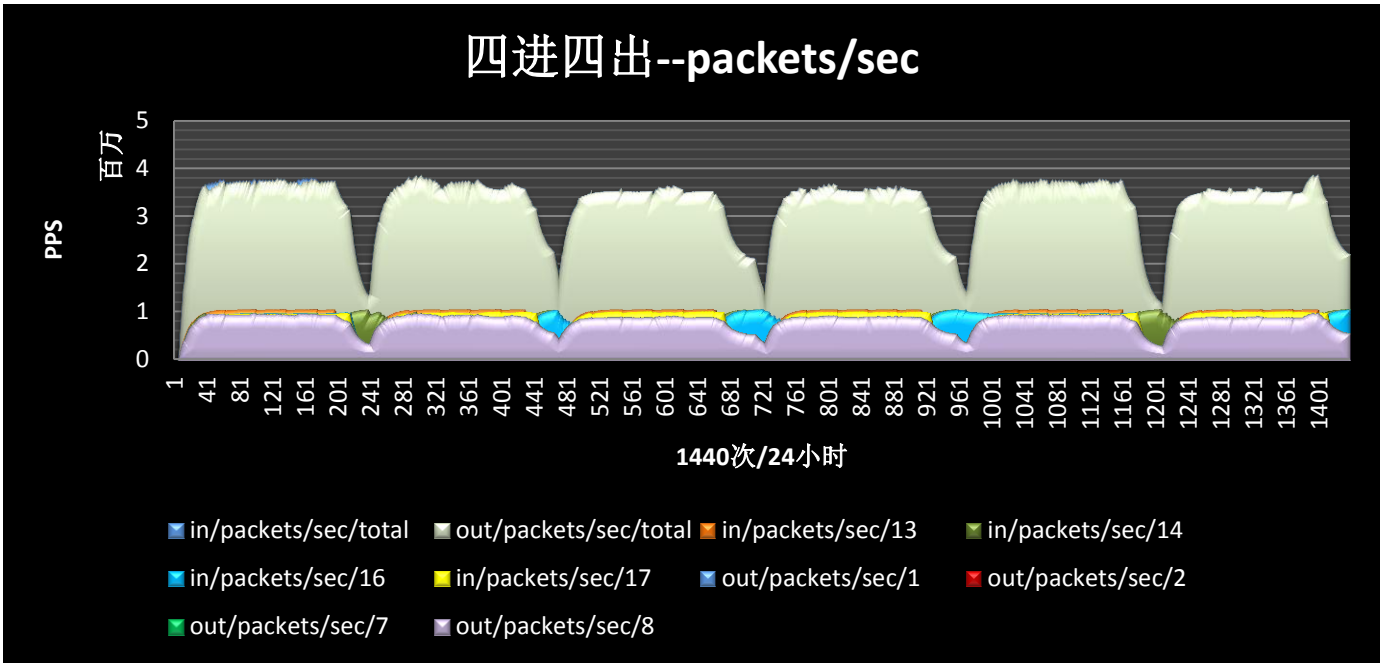
阐述：低谷为压力脚本重启波动，总输入输出基本持平。200 万 pps 进出口平稳。出包量略低于进包数量。



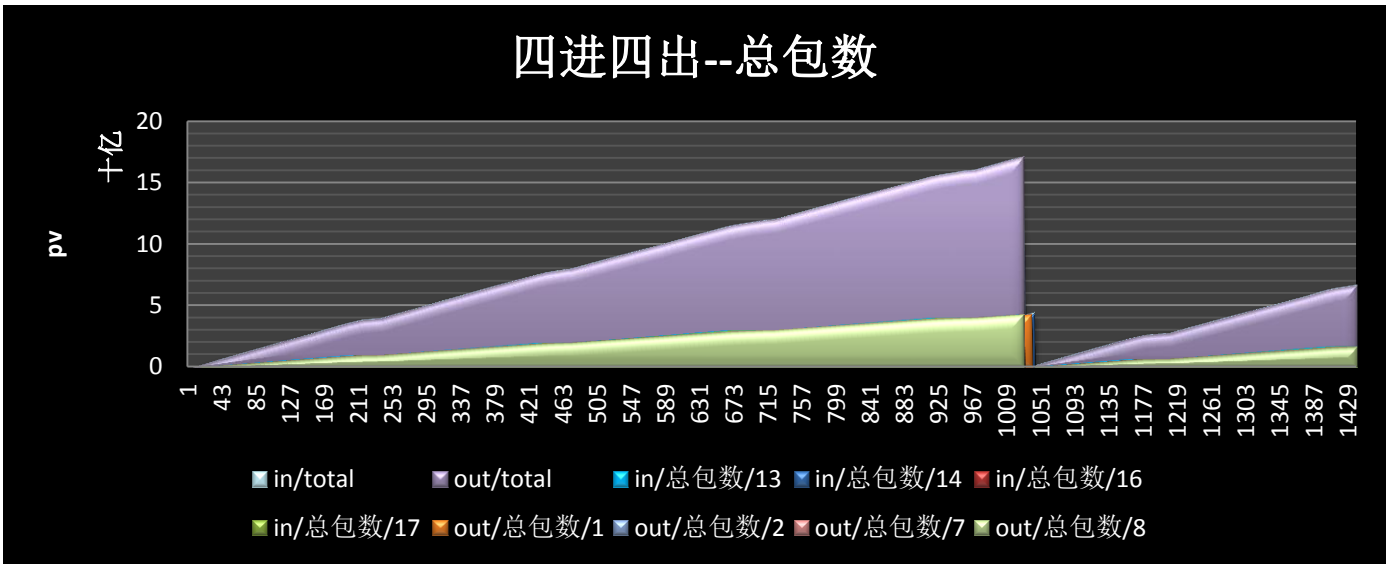
阐述：同样的溢出，因为双口计算 outpps，所以到达 4291293796*2 就会溢出清零。进出包数量依然很平稳。



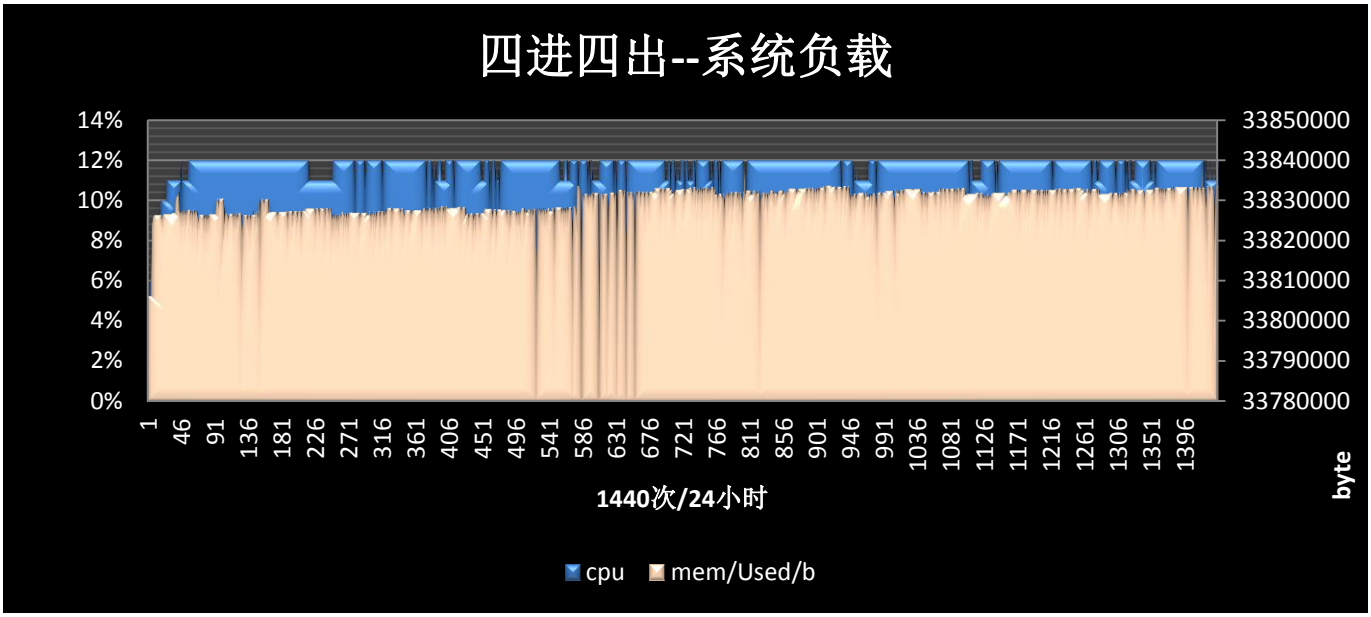
阐述：cpu 和内存峰值原因是由于包数溢出清 0，导致负载瞬间飙升。Cpu 峰值 11%，内存约 34M。



阐述：进出口总包数达到约 400 万 pps，进出包比较平稳。起伏由于压力源导致重启发包机制导致。单独端口转发基本维持在 90-110 万 pps 这个区间。但是 4 进还是要比 4 出包数多点。



阐述：因为是 4 个口 outpps，所以 4291293796*4 才会溢出清 0.



阐述：正常。

总结：以上 1-4 口压力数据取决于交换机性能、网络环境、测试服务器的影响，数据会有一定偏差。由于数据采用顺时采集，采集数据间隔会有数据变化。支持 4 条等价路由全千兆交换机，最大支持 400 万 pps。

优缺点

横向扩展：

LVS 调度机自由伸缩，横向线性扩展（最多机器数受限于三层设备允许的等价路由数目 maximum load-balancing ）

LVS 机器同时工作，不存在备机，提高利用率

Ospf 负载均衡取决于等价路由条数，默认 4 条，最大 16 条

切换灵活：

当某台服务器故障，直接摘掉即可，路由自动会断开此条链路

转发稳定：

硬件转发稳定性比较强

跨网工作：

使用 ospf 负载均衡可以跨网段，跨机房进行业务扩展

其他说明

华 3 设备 ospfd 调度算法的问题，一台宕机会使所有的长连接的断开重连，目前还无法解决;思科的设备已经支持一至性哈希算法，不会出现这个问题

本次测试使用千兆网络，线上可以使用万兆或做一些聚合来结合使用。

感谢 jialei3 对文档提出改正。