淘宝网 Taobao.com 淘我喜欢！

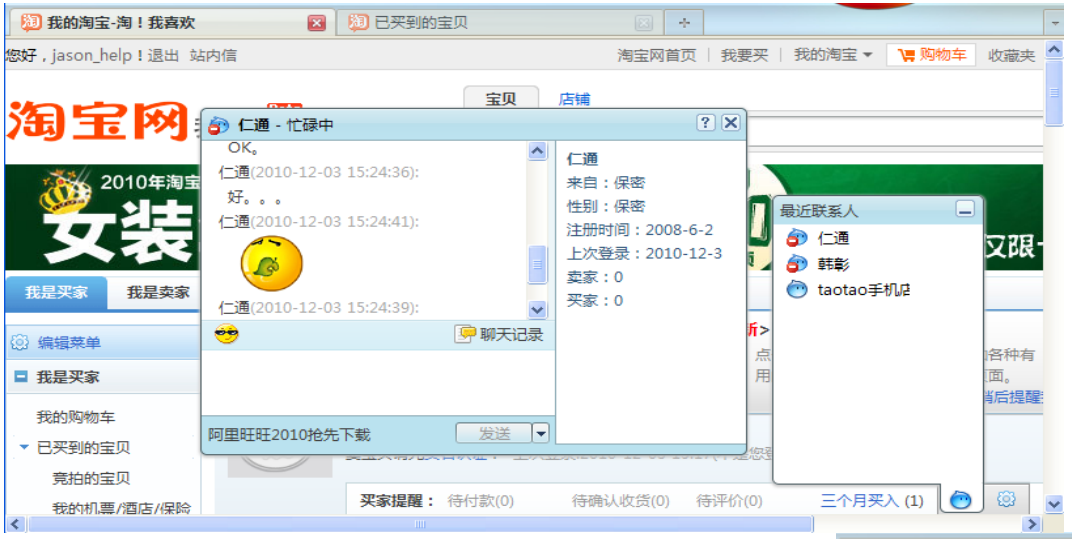
面向浏览器的消息推送

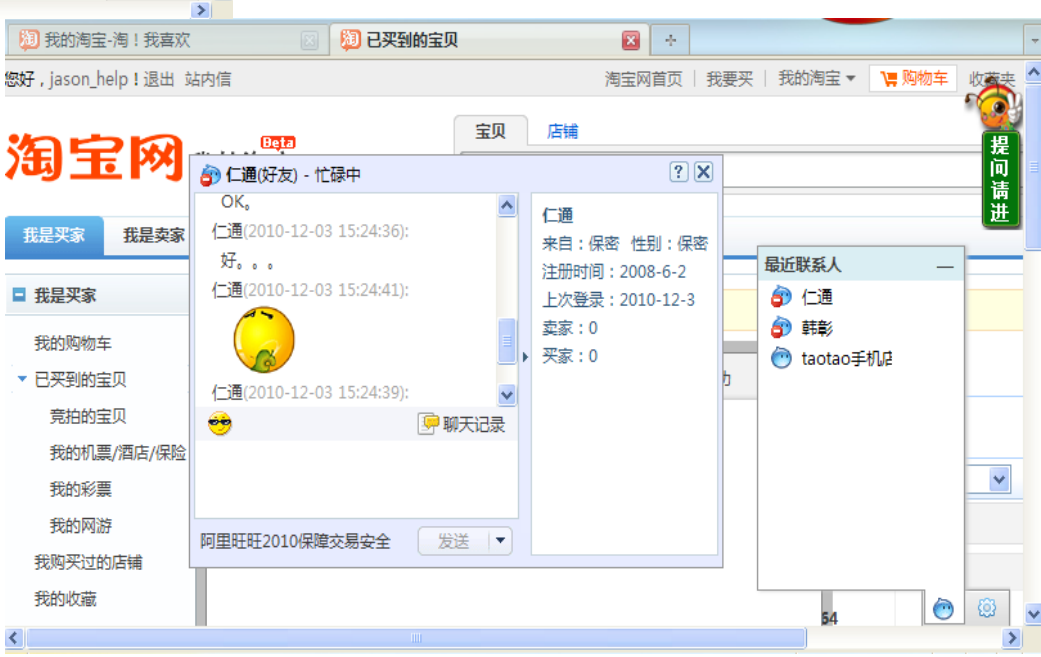
张乐伟（韩彰）

议程

- **http 长连接**
- 多页面多域名
- 消息路由
- 相关调优
- webIM
- 问题交流

Web旺旺概述


我的淘宝-淘！我喜欢
已买到的宝贝
您好，jason_help！退出 站内信
淘宝网首页 | 我要买 | 我的淘宝 购物车 收藏夹
宝贝 店铺
淘宝网
仁通 - 忙碌中
OK。
仁通(2010-12-03 15:24:36):
好。。。
仁通(2010-12-03 15:24:41):
仁通(2010-12-03 15:24:39):
聊天记录
阿里旺旺2010抢先下载
发送
仁通
来自：保密
性别：保密
注册时间：2008-6-2
上次登录：2010-12-3
卖家：0
买家：0
最近联系人
仁通
韩彰
taotao手机店
我是买家
我是卖家
编辑菜单
我的购物车
已买到的宝贝
竞拍的宝贝
买家提醒： 待付款(0) 待确认收货(0) 待评价(0) 三个月买入 (1)

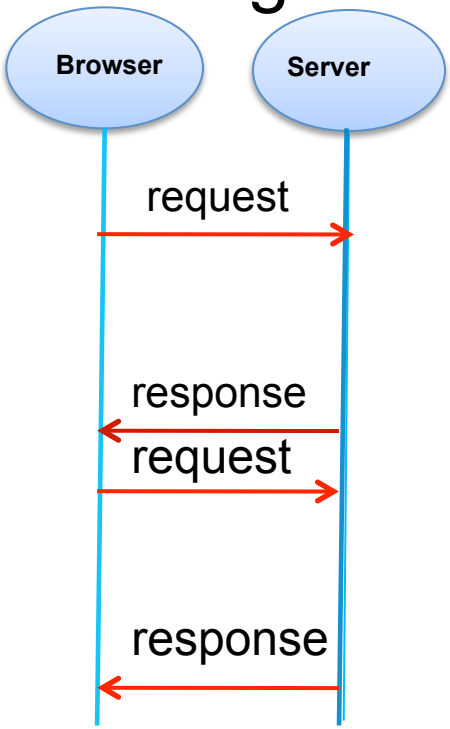

我的淘宝-淘！我喜欢
已买到的宝贝
您好，jason_help！退出 站内信
淘宝网首页 | 我要买 | 我的淘宝 购物车 收藏夹
宝贝 店铺
淘宝网
Beta
仁通(好友) - 忙碌中
OK。
仁通(2010-12-03 15:24:36):
好。。。
仁通(2010-12-03 15:24:41):
仁通(2010-12-03 15:24:39):
聊天记录
阿里旺旺2010保障交易安全
发送
仁通
来自：保密 性别：保密
注册时间：2008-6-2
上次登录：2010-12-3
卖家：0
买家：0
最近联系人
仁通
韩彰
taotao手机店
提问请进
我是买家
我是卖家
我的购物车
已买到的宝贝
竞拍的宝贝
我的机票/酒店/保险
我的彩票
我的网游
我购买过的店铺
我的收藏

特点

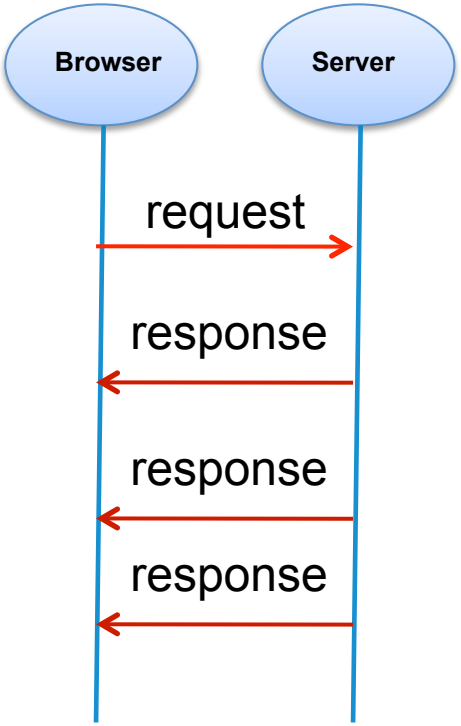
- 浏览器
  - IE6，IE7，IE8，Chrome，Firefox。。。。
- 实时性
  - http长连接
- 用户量
  - 大量用户（大量连接）
  - 多个页面
  - 多个域名

Http长连接

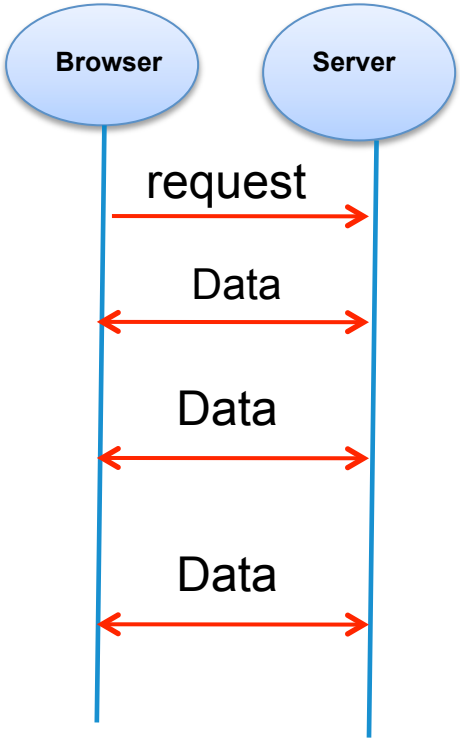
Comet Long Pooling



Comet Streaming



Websocket



httpServer

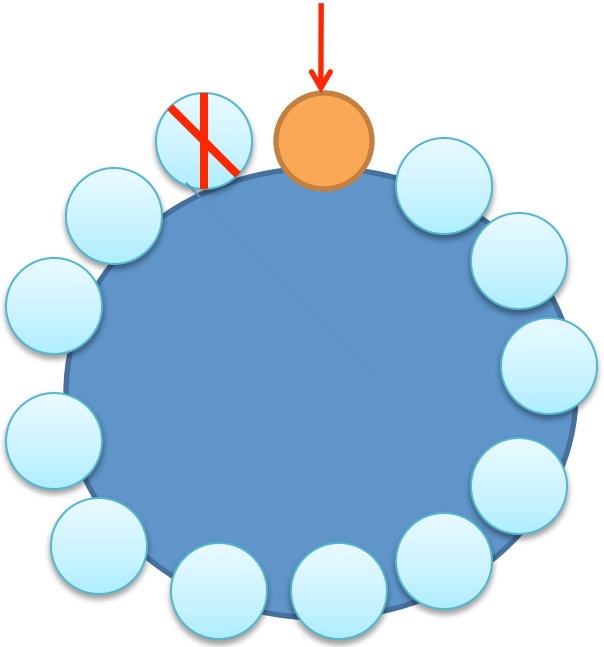
jetty

自己写（netty，c，erlang）

淘宝网 Taobao.com 淘我喜欢！

连接管理

- 特性
  - 连接量（20万）
  - 实时性（每秒需要调度超时的连接）
  - 同一个用户不允许多个连接（查找是否已经存在）
- 方法
  - 一个连接一个定时器（不现实）
  - 一个定时器遍历Map（耗时）
  - TimerWheel+Map



议程

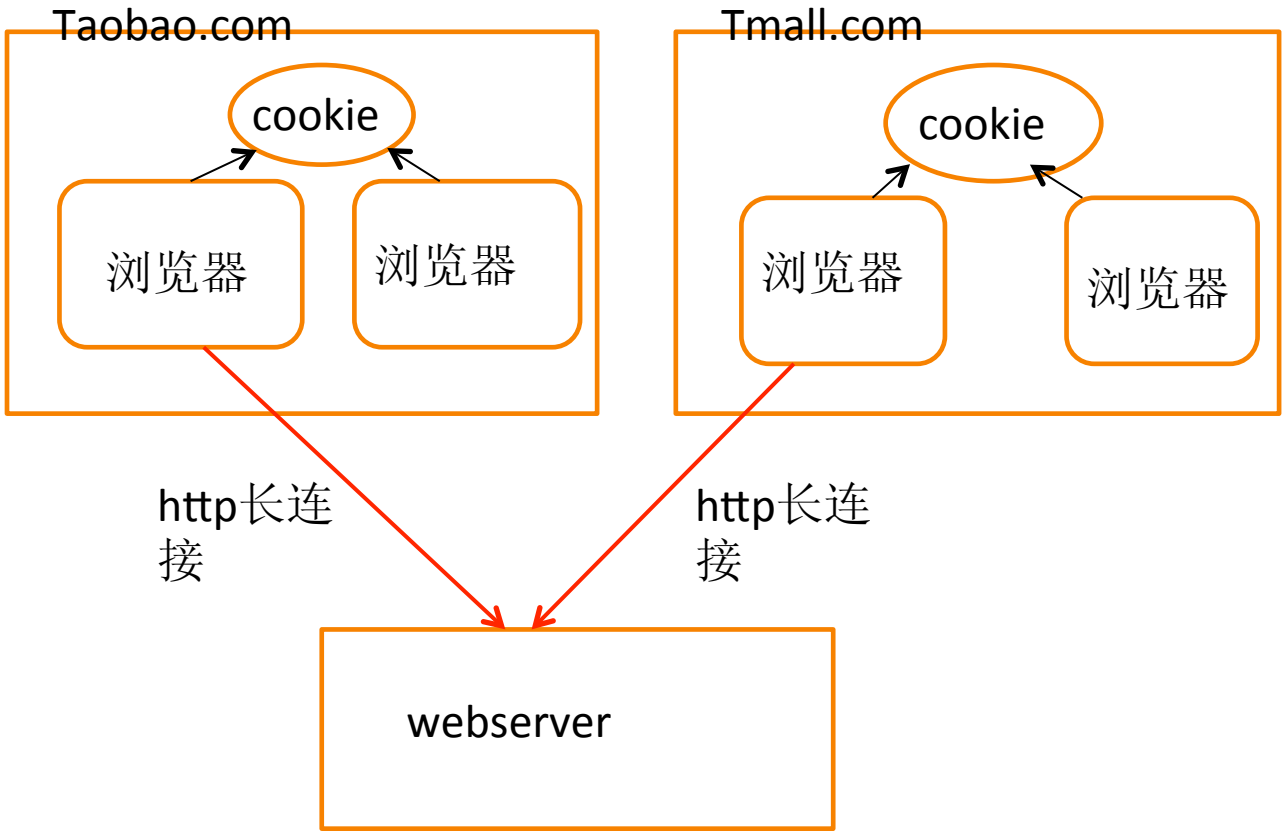
- http 长连接
- 多页面多域名
- 消息路由
- 相关调优
- webIM
- 问题交流

多页面

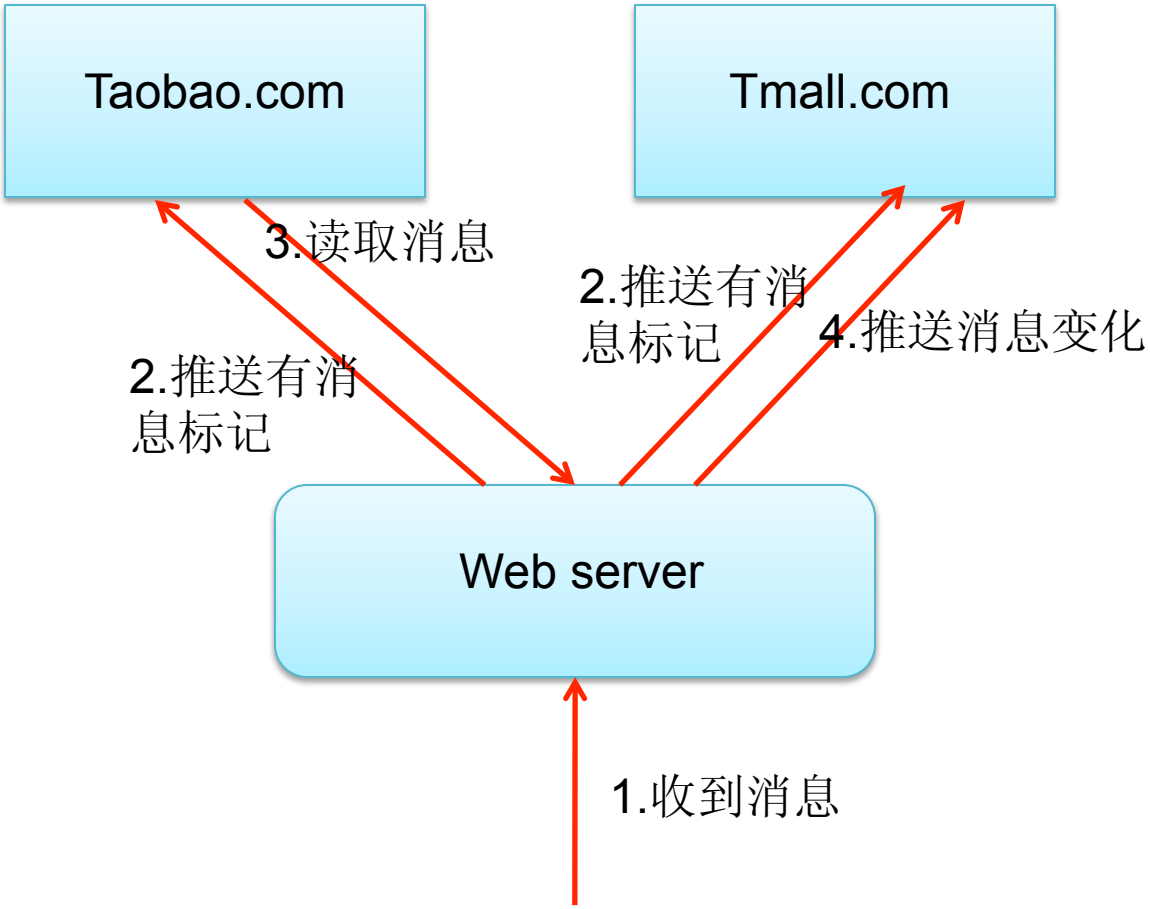
- 一个页面一个连接
  - 大量连接
- 多个页面共享一个连接
  - 页面之间如何同步消息
    - Flash 的localconnection
    - 轮询cookie
    - html5
  - 消息存储
    - Flashdb
    - Server端存储
    - html5

多域名

一个域名一个连接



多域名下消息推送


Taobao.com
Tmall.com
3.读取消息
2.推送有消息标记
2.推送有消息标记
4.推送消息变化
Web server
1.收到消息

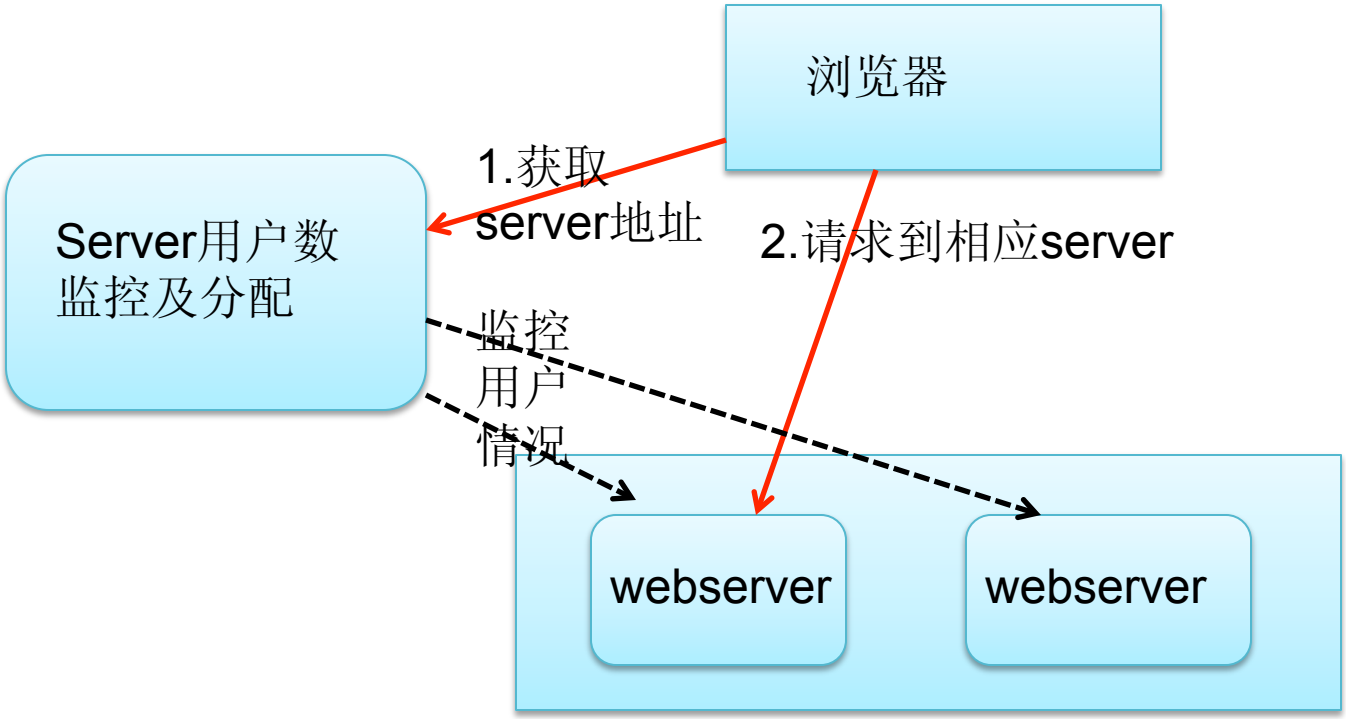
议程

- http 长连接
- 多页面多域名
- **消息路由**
- 相关调优
- webIM
- 问题交流

负载均衡

- 利用现有负载均衡
  - 4层
    - 硬件负载（F5,A10…)
    - 软件负载(LVS)
  - 7层
    - Haproxy
- 自己控制均衡

自己控制均衡


浏览器
1.获取
server地址
2.请求到相应server
Server用户数
监控及分配
监控
用户
情况
webserver
webserver

淘宝网 Taobao.com 淘我喜欢！

负载均衡

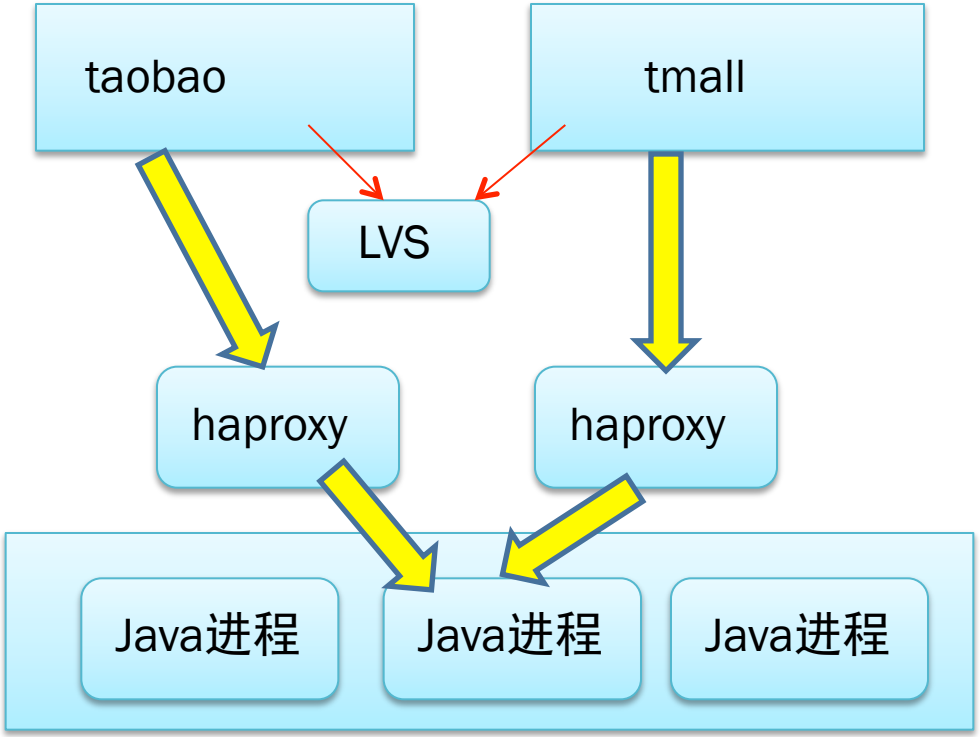
- **4层**
  - 硬件负载
    - 大量连接
  - LVS
    - NAT(IP地址转换）
    - TUN(IP隧道)
    - DR
- **7层**
  - haproxy
    - 单线程模型
    - 一致性hash的均衡性
    - 内存瓶颈

Haproxy

- 引入原因
  - Session sticky
- 特点
  - 负载均衡
    - RR　轮询
    - Least Connection　最少连接
    - Ip hash
    - Uri hash
    - url_param hash
  - Session sticky（注意多台haproxy情况）
  - 虚拟主机
  - 一致性hash支持
  - Connection keep alive

负载均衡

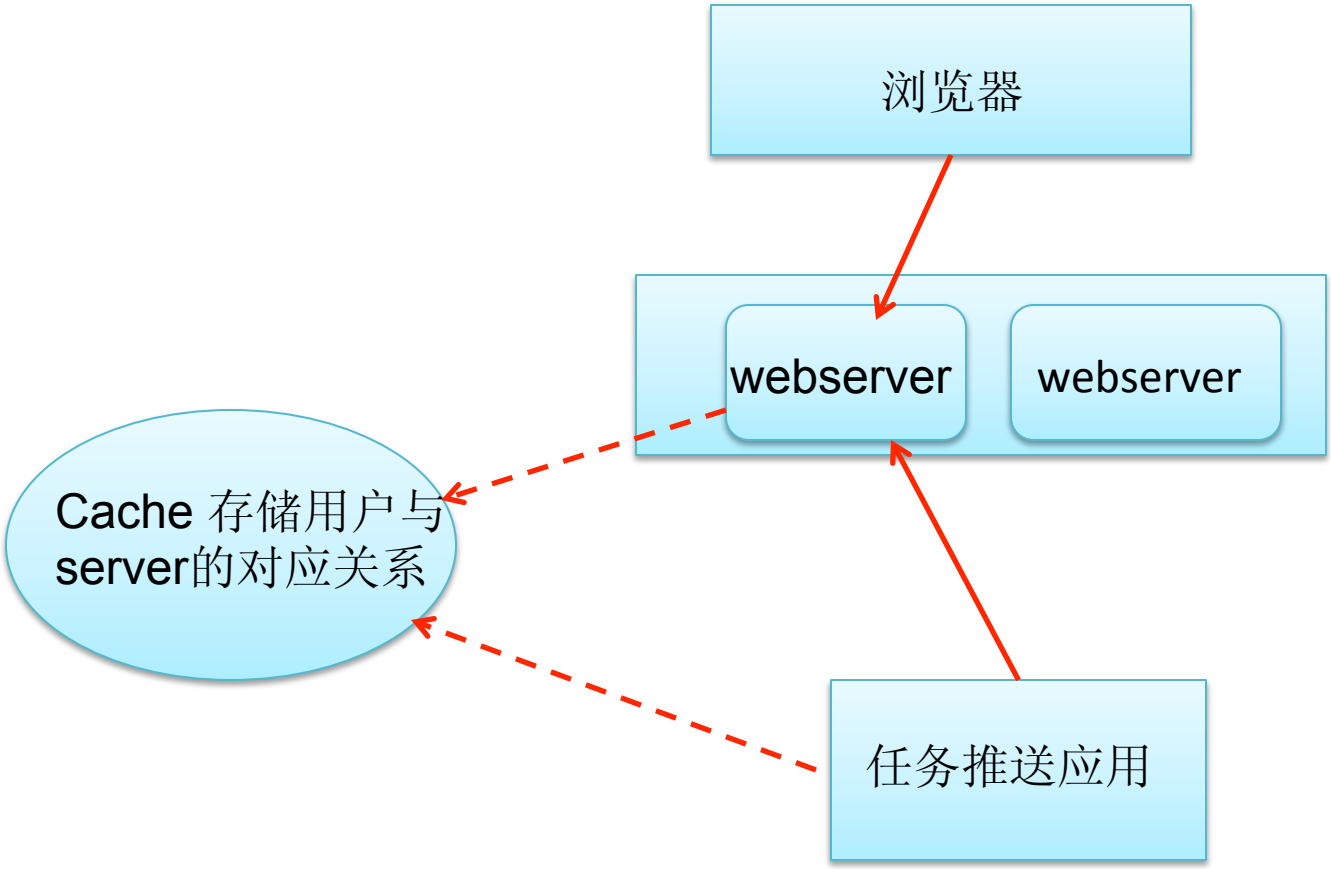
- **LVS+haproxy**



淘宝网 Taobao.com 淘我喜欢！

消息路由

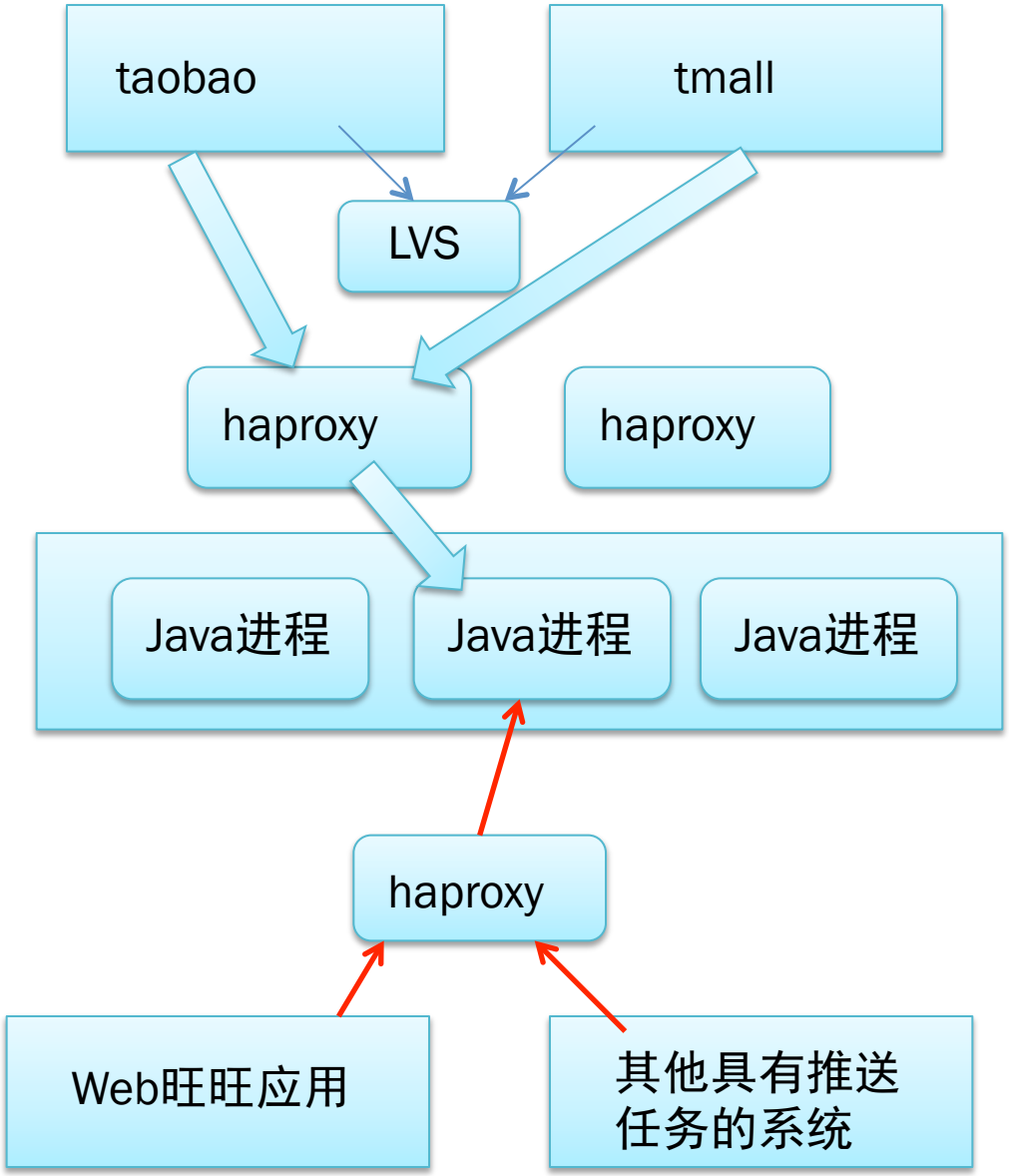
- 集中式存储用户与机器对应关系



消息路由

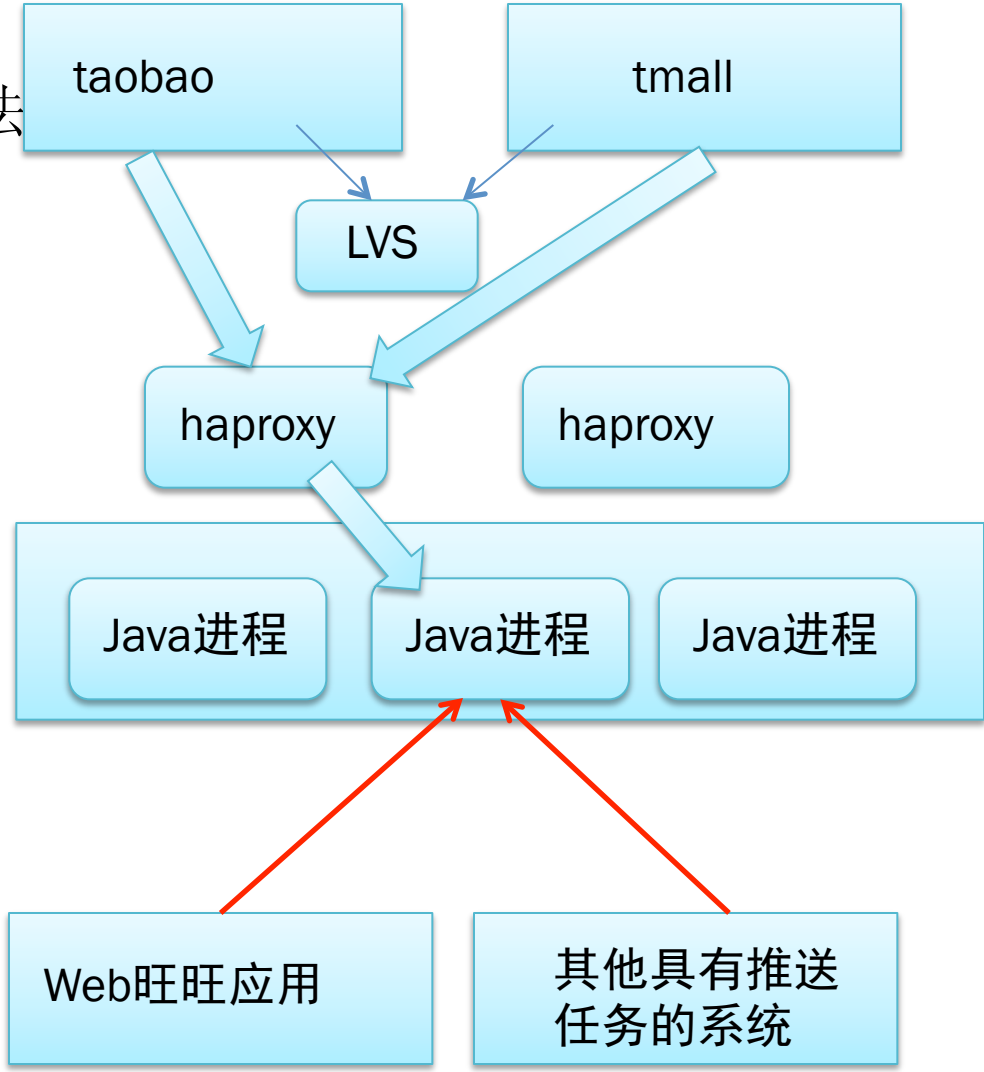
淘宝网 Taobao.com 淘我喜欢！

利用haproxy



消息路由

- 提取haproxy的一致性hash算法
- Haproxy 问题
  - 不支持长连接
  - 7层只支持http协议
- TCP hash版
  - 用tcp支持7层hash
  - 实现长连接
  - 同haproxy同样的hash
  - java实现，去掉proxy点
  - 简化系统结构



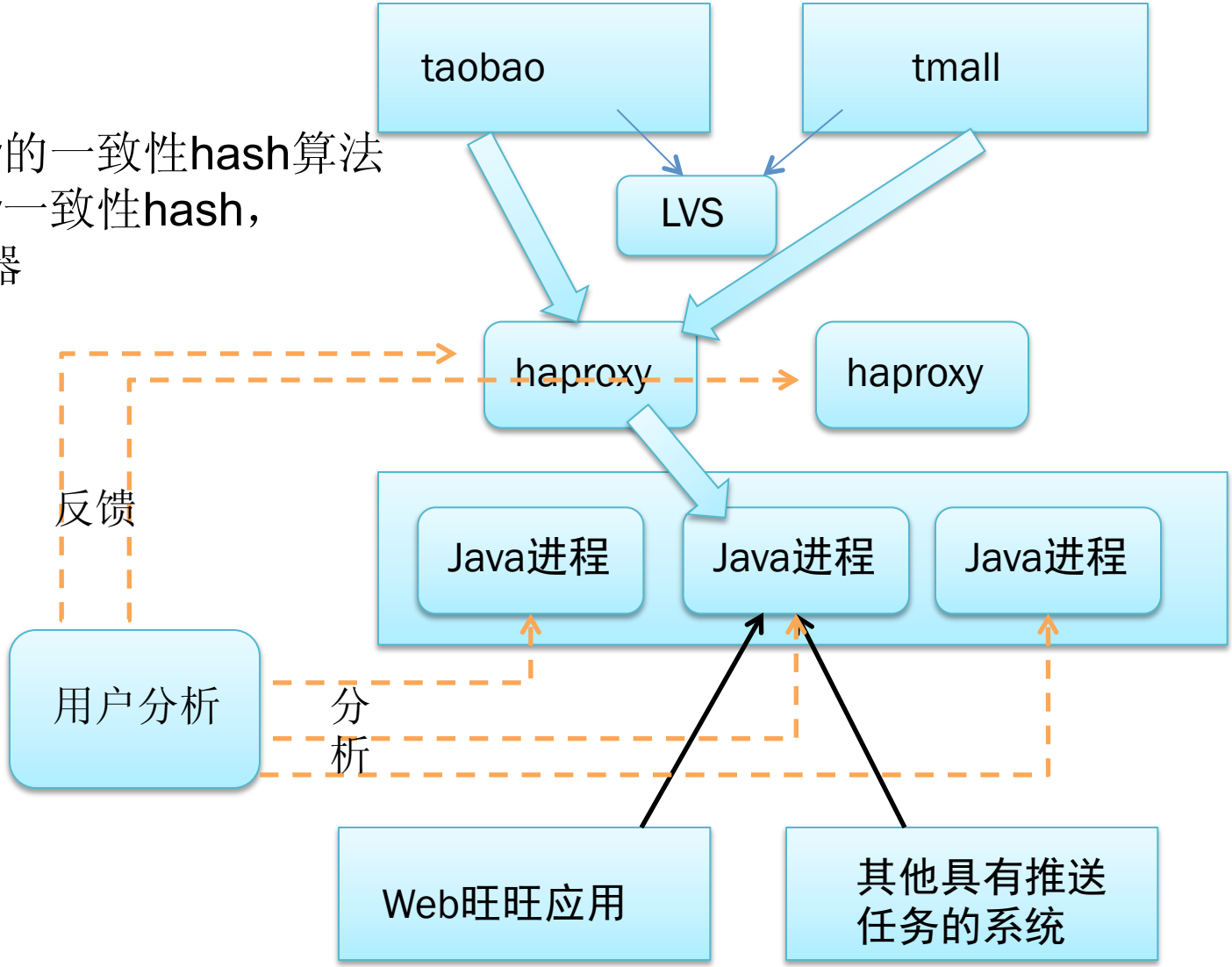
Server情况

- 用户不均衡
  - why ?
- **haproxy一致性hash**

```
#server list
server mpp177063.cm3 172.23.177.63:7001 id 1001 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177065.cm3 172.23.177.65:7001 id 1002 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177066.cm3 172.23.177.66:7001 id 1003 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177067.cm3 172.23.177.67:7001 id 1004 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177069.cm3 172.23.177.69:7001 id 1005 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177070.cm3 172.23.177.70:7001 id 1006 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177071.cm3 172.23.177.71:7001 id 1007 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177072.cm3 172.23.177.72:7001 id 1008 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177087.cm3 172.23.177.87:7001 id 1009 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177088.cm3 172.23.177.88:7001 id 1010 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177090.cm3 172.23.177.90:7001 id 1011 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177091.cm3 172.23.177.91:7001 id 1012 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177092.cm3 172.23.177.92:7001 id 1013 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177093.cm3 172.23.177.93:7001 id 1014 maxconn 200000 check port 7001 inter 5000 fall 8
server mpp177094.cm3 172.23.177.94:7001 id 1015 maxconn 200000 check port 7001 inter 5000 fall 8
```

消息路由

提取haproxy的一致性hash算法
调整haproxy一致性hash，
均衡每台机器



一些故事

- 硬件负载
  - 断开连接
  - 大量连接
- **Haproxy**
  - 内存
  - 连接

议程

- http 长连接
- 多页面多域名
- 消息路由
- **相关调优**
- webIM
- 问题交流

Haproxy进程

两个haproxy进程

```
top - 11:47:45 up 325 days, 20:33,  2 users,  load average: 0.94, 0.98, 0.99
Tasks: 120 total,   2 running, 118 sleeping,   0 stopped,   0 zombie
Cpu0  : 13.4%us, 25.5%sy,  0.0%ni, 29.5%id,  0.0%wa,  1.0%hi, 30.5%si,  0.0%st
Cpu1  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu2  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu3  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu4  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu5  :  0.0%us,  0.0%sy,  0.0%ni, 99.3%id,  0.0%wa,  0.0%hi,  0.7%si,  0.0%st
Cpu6  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%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
Mem:   8175012k total,  3828812k used,  4346200k free,   565572k buffers
Swap:  2096472k total,        0k used,  2096472k free,   355100k cached
```

```
top - 11:50:53 up 325 days, 20:36,  2 users,  load average: 1.07, 0.99, 0.99
Tasks: 120 total,   2 running, 118 sleeping,   0 stopped,   0 zombie
Cpu0  :  0.0%us,  0.0%sy,  0.0%ni, 60.9%id,  0.0%wa,  1.0%hi, 38.1%si,  0.0%st
Cpu1  :  8.0%us, 17.9%sy,  0.0%ni, 71.4%id,  0.0%wa,  0.0%hi,  2.7%si,  0.0%st
Cpu2  :  7.0%us, 21.9%sy,  0.0%ni, 66.8%id,  0.0%wa,  0.0%hi,  4.3%si,  0.0%st
Cpu3  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu4  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu5  :  0.0%us,  0.0%sy,  0.0%ni, 99.3%id,  0.0%wa,  0.0%hi,  0.7%si,  0.0%st
Cpu6  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%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
Mem:   8175012k total,  3880504k used,  4294508k free,   565588k buffers
Swap:  2096472k total,        0k used,  2096472k free,   355120k cached
```

淘宝网 Taobao.com 淘我喜欢！

网络

- 网卡软中断
  - 1w包，4%左右cpu

```
Cpu0  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu1  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu2  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu3  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu4  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu5  :  0.0%us,  0.0%sy,  0.0%ni, 42.0%id,  0.0%wa,  0.0%hi, 58.0%si,  0.0%st
Cpu6  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu7  :  0.0%us,  0.0%sy,  0.0%ni, 44.0%id,  0.0%wa,  0.0%hi, 56.0%si,  0.0%st
```

mpstat

```
04:08:20 PM  CPU   %user   %nice    %sys %iowait    %irq   %soft  %steal   %idle    intr/s
04:08:21 PM  all    0.00    0.00    0.00    0.00    0.00   14.89    0.00   85.11  27802.00
04:08:21 PM    0    0.00    0.00    0.00    0.00    0.00    0.00    0.00  100.00   1001.00
04:08:21 PM    1    0.00    0.00    0.00    0.00    0.00    0.00    0.00  100.00      0.00
04:08:21 PM    2    0.00    0.00    0.00    0.00    0.00    0.00    0.00  100.00      0.00
04:08:21 PM    3    0.00    0.00    0.00    0.00    0.00    0.00    0.00  100.00      0.00
04:08:21 PM    4    0.00    0.00    0.00    0.00    0.00    0.00    0.00  100.00      0.00
04:08:21 PM    5    0.00    0.00    0.00    0.00    0.00   59.60    0.00   40.40  13348.00
04:08:21 PM    6    0.00    0.00    0.00    0.00    0.00    0.00    0.00  100.00      0.00
04:08:21 PM    7    0.00    0.00    0.99    0.00    0.00   58.42    0.00   40.59  13449.00
```

网络

- 网卡软中断
  - 多队列网卡
  - 单队列网卡 +receive packet steering（2.6.32）

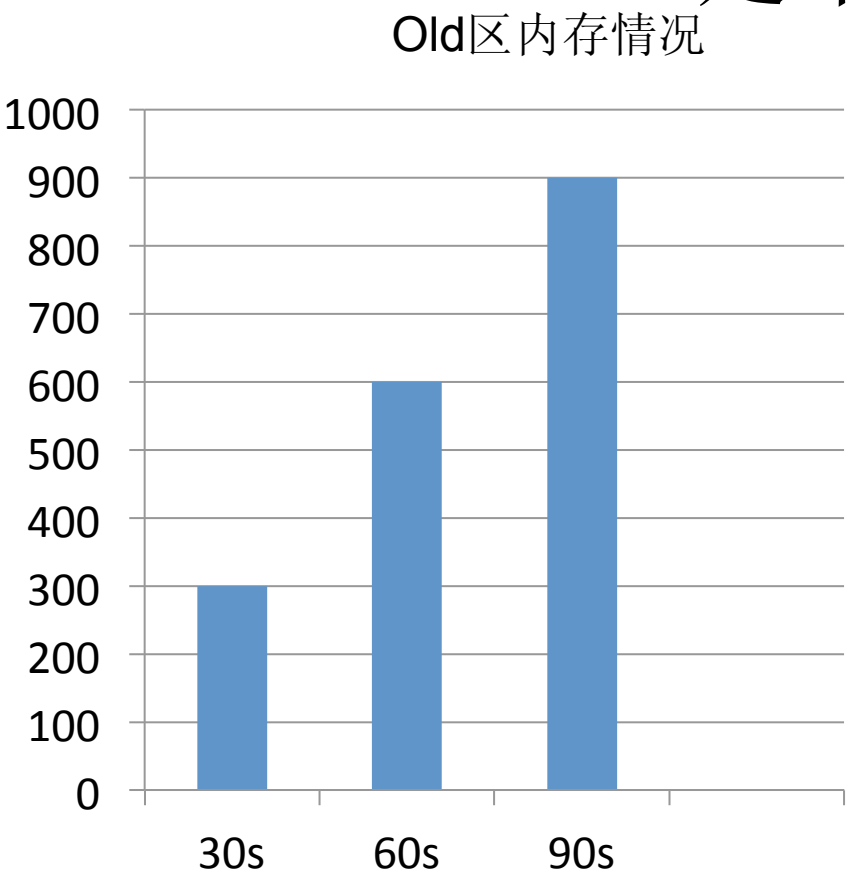
```
top - 15:52:25 up 11 days,  5:37,  6 users,  load average: 0.00, 0.00, 0.00
Tasks: 200 total,   1 running, 199 sleeping,   0 stopped,   0 zombie
Cpu0  :  0.0%us,  0.0%sy,  0.0%ni, 87.3%id,  0.0%wa,  0.0%hi, 12.7%si,  0.0%st
Cpu1  :  0.0%us,  0.0%sy,  0.0%ni, 88.1%id,  0.0%wa,  0.0%hi, 11.9%si,  0.0%st
Cpu2  :  0.0%us,  0.0%sy,  0.0%ni, 87.1%id,  0.0%wa,  0.0%hi, 12.9%si,  0.0%st
Cpu3  :  0.0%us,  1.0%sy,  0.0%ni, 88.1%id,  0.0%wa,  0.0%hi, 10.9%si,  0.0%st
Cpu4  :  0.0%us,  0.0%sy,  0.0%ni, 88.9%id,  0.0%wa,  0.0%hi, 11.1%si,  0.0%st
Cpu5  :  0.0%us,  0.0%sy,  0.0%ni, 87.5%id,  0.0%wa,  0.0%hi, 12.5%si,  0.0%st
Cpu6  :  0.0%us,  0.0%sy,  0.0%ni, 87.9%id,  0.0%wa,  0.0%hi, 12.1%si,  0.0%st
Cpu7  :  0.0%us,  0.0%sy,  0.0%ni, 85.7%id,  0.0%wa,  0.0%hi, 14.3%si,  0.0%st
Mem:  16468316k total,  1581932k used, 14886384k free,   194736k buffers
Swap:  2096472k total,        0k used,  2096472k free,   174640k cached
```

瓶颈

- 连接数 ？
  - Jetty 6万左右
  - 自己写server 20万
- QPS ？
  - 60000/30 = 2000
  - 20w/30 = 7000
- 内存 ？
  - 5K *60000 = 300M
  - 5K*200000 = 1G

淘宝网 Taobao.com 淘我喜欢！

GC是个问题



- 60000连接
  - 30s断开一次连接
  - 30s将建立60000个连接
  - 每30s产生300m内存
  - 全部进入old区
- full gc频繁，应用暂停时间长

GC调优

- 并发GC（CMS：Concurrent Mark-Sweep GC)
  - 1.Initial Marking 2.Concurrent Marking 3.Final Marking 4.Concurrent Sweeping
  - 第一步和第三步需要暂停应用
  - 仍将消耗较长时间

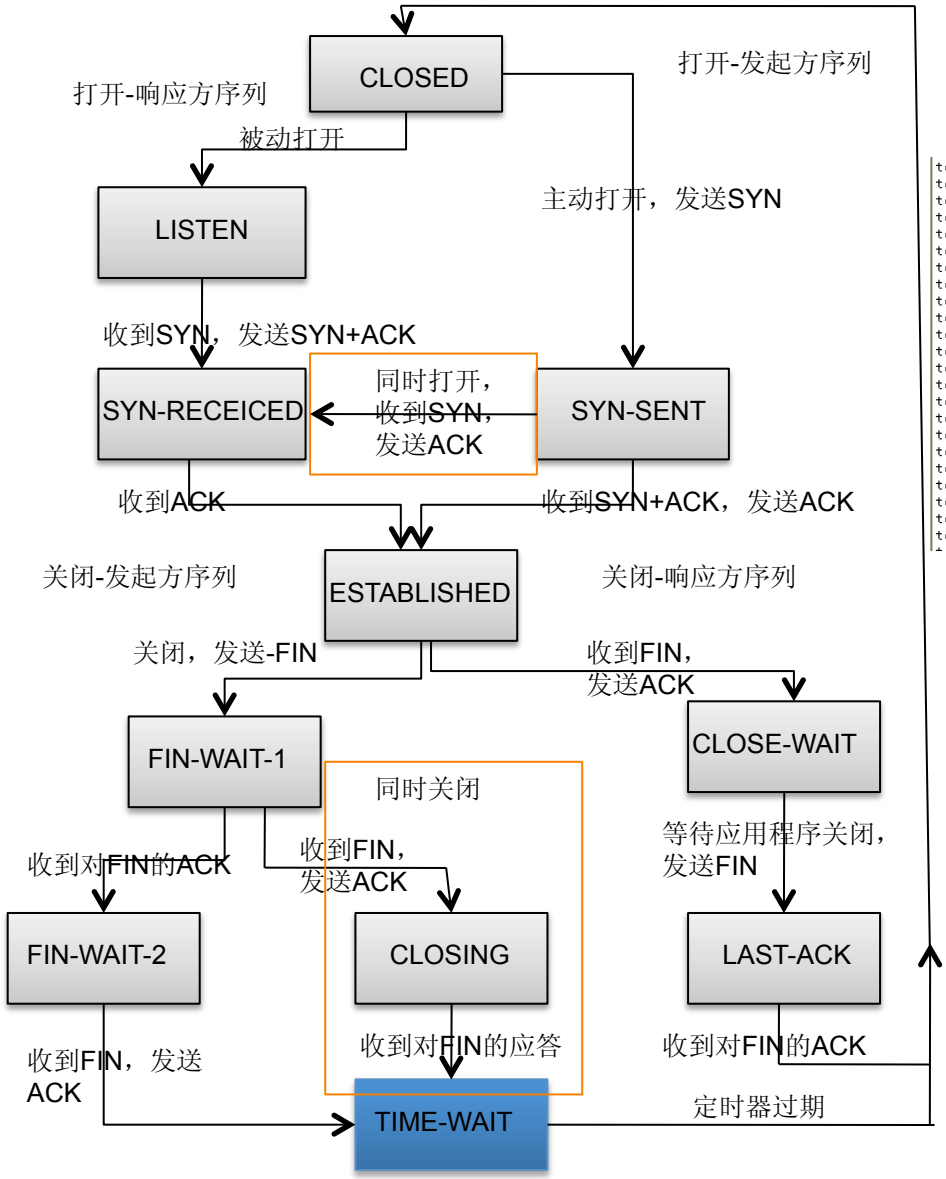
- 扩大S0,S1区， -XX:MaxTenuringThreshold=15 -XX:SurvivorRatio=1，确保不进入old区
  - Eden,S0,S1大小相等
  - Old区没有数据

- Minor GC
  - ParNewGC 并行GC，比较消耗CPU
  - ParallelGC 并行回收GC -XX:-UseAdaptiveSizePolicy

	Full GC	Minor GC
并发GC	-XX:+UseConcMarkSweepGC	-XX:+UseParNewGC
并行GC	-XX:+UseParallelGC	-XX:+UseParallelGC

淘宝网 Taobao.com 淘我喜欢！

TCP生命周期




```
tcp        0      0 172.23.20.67:80          120.207.144.56:1732      SYN_RECV    on (2.07/0/0)
tcp        0      0 172.23.20.67:80          115.215.145.204:51697    SYN_RECV    on (55.34/5/0)
tcp        0      0 172.23.20.67:80          218.26.48.11:52318       SYN_RECV    on (2.99/0/0)
tcp        0      0 172.23.20.67:80          220.161.47.91:15956      SYN_RECV    on (42949672.90/1/0)
tcp        0      0 172.23.20.67:80          211.138.31.7:4819        SYN_RECV    on (3.00/0/0)
tcp        0      0 0.0.0.0:8083             0.0.0.0:*                LISTEN      off (0.00/0/0)
tcp        0      0 0.0.0.0:7001             0.0.0.0:*                LISTEN      off (0.00/0/0)
tcp        0      0 0.0.0.0:4444             0.0.0.0:*                LISTEN      off (0.00/0/0)
tcp        0      0 0.0.0.0:4445             0.0.0.0:*                LISTEN      off (0.00/0/0)
tcp        0      0 0.0.0.0:4446             0.0.0.0:*                LISTEN      off (0.00/0/0)
tcp        0      0 172.23.20.67:37980       172.24.13.158:12200      ESTABLISHED keepalive (1620.24/0/0)
tcp        0      0 172.23.20.67:36244       172.24.24.105:12200      ESTABLISHED keepalive (1072.23/0/0)
tcp        0      0 172.23.20.67:54574       172.24.128.1:3100        ESTABLISHED keepalive (2508.13/0/0)
tcp        0      0 172.23.20.67:57440       172.23.112.61:3306       ESTABLISHED keepalive (4476.80/0/0)
tcp        0      0 172.23.20.67:36425       172.24.19.130:12200      ESTABLISHED keepalive (4733.05/0/0)
tcp        0      0 127.0.0.1:8009           127.0.0.1:36110          TIME_WAIT   timewait (12.02/0/0)
tcp        0      0 172.23.20.67:36654       172.24.19.134:12200      ESTABLISHED keepalive (4764.22/0/0)
tcp        0      0 172.23.20.67:36386       172.24.19.132:12200      ESTABLISHED keepalive (4861.16/0/0)
tcp        0      0 172.23.20.67:36822       172.24.13.154:12200      ESTABLISHED keepalive (982.20/0/0)
tcp        0      0 127.0.0.1:8009           127.0.0.1:36359          ESTABLISHED keepalive (7198.32/0/0)
tcp        0      0 127.0.0.1:8009           127.0.0.1:36358          ESTABLISHED keepalive (7198.32/0/0)
tcp        0      0 127.0.0.1:8009           127.0.0.1:36357          ESTABLISHED keepalive (7198.31/0/0)
tcp        0      0 172.23.20.67:36684       172.23.26.113:12200      ESTABLISHED keepalive (243.60/0/0)
```

```
tcp        0      0 172.23.20.67:34986       110.75.165.167:16000     ESTABLISHED keepalive (2242.06/0/
tcp        0      0 172.23.20.67:34980       110.75.165.167:16000     ESTABLISHED keepalive (2260.98/0/
tcp        0      0 172.23.20.67:34986       110.75.165.167:16000     ESTABLISHED keepalive (2242.05/0/
tcp        0      0 172.23.20.67:34980       110.75.165.167:16000     ESTABLISHED keepalive (2260.97/0/
tcp        0      0 172.23.20.67:34969       110.75.165.167:16000     ESTABLISHED keepalive (2241.15/0/
tcp        0    100 172.23.20.67:34958       110.75.165.167:16000     ESTABLISHED on (0.21/0/0)
tcp        0      0 172.23.20.67:34945       110.75.165.167:16000     ESTABLISHED keepalive (2237.85/0/
tcp        0      0 :::80                    :::*                     LISTEN      keepalive (0.00/0/0)
tcp        0      0 :::22                    :::*                     LISTEN      off (0.00/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:221.176.147.20:56140 TIME_WAIT timewait (33.51/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:211.138.105.20:51602 TIME_WAIT timewait (16.02/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:222.82.155.74:1219   TIME_WAIT timewait (38.35/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:222.130.163.92:2365  TIME_WAIT timewait (41.91/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:59.108.49.28:62721   TIME_WAIT timewait (35.85/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:60.17.237.238:3501   TIME_WAIT timewait (20.38/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:61.184.187.248:2628  TIME_WAIT timewait (42.06/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:58.17.81.232:1303    TIME_WAIT timewait (25.69/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:123.178.137.17:11822 TIME_WAIT timewait (9.35/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:114.238.28.187:4854  TIME_WAIT timewait (48.10/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:123.52.157.148:4482  TIME_WAIT timewait (58.06/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:117.24.27.208:1644   TIME_WAIT timewait (25.53/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:113.77.55.50:2056    TIME_WAIT timewait (16.02/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:115.225.53.56:1708   TIME_WAIT timewait (38.08/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:125.124.21.66:4459   TIME_WAIT timewait (28.20/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:180.208.66.163:9358  TIME_WAIT timewait (47.06/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:117.32.149.208:58074 TIME_WAIT timewait (42.67/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:120.65.252.75:4278   TIME_WAIT timewait (39.19/0/0)
tcp        0    326 ::ffff:172.23.20.67:80   ::ffff:222.93.2.81:32846    FIN_WAIT1 on (38.98/4/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:123.138.20.65:12191  TIME_WAIT timewait (46.17/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:118.119.215.23:50679 TIME_WAIT timewait (14.00/0/0)
tcp        0      0 ::ffff:172.23.20.67:80   ::ffff:180.99.148.137:1217  TIME_WAIT timewait (28.37/0/0)
```

淘宝网 Taobao.com 淘我喜欢！

Linux TCP调优

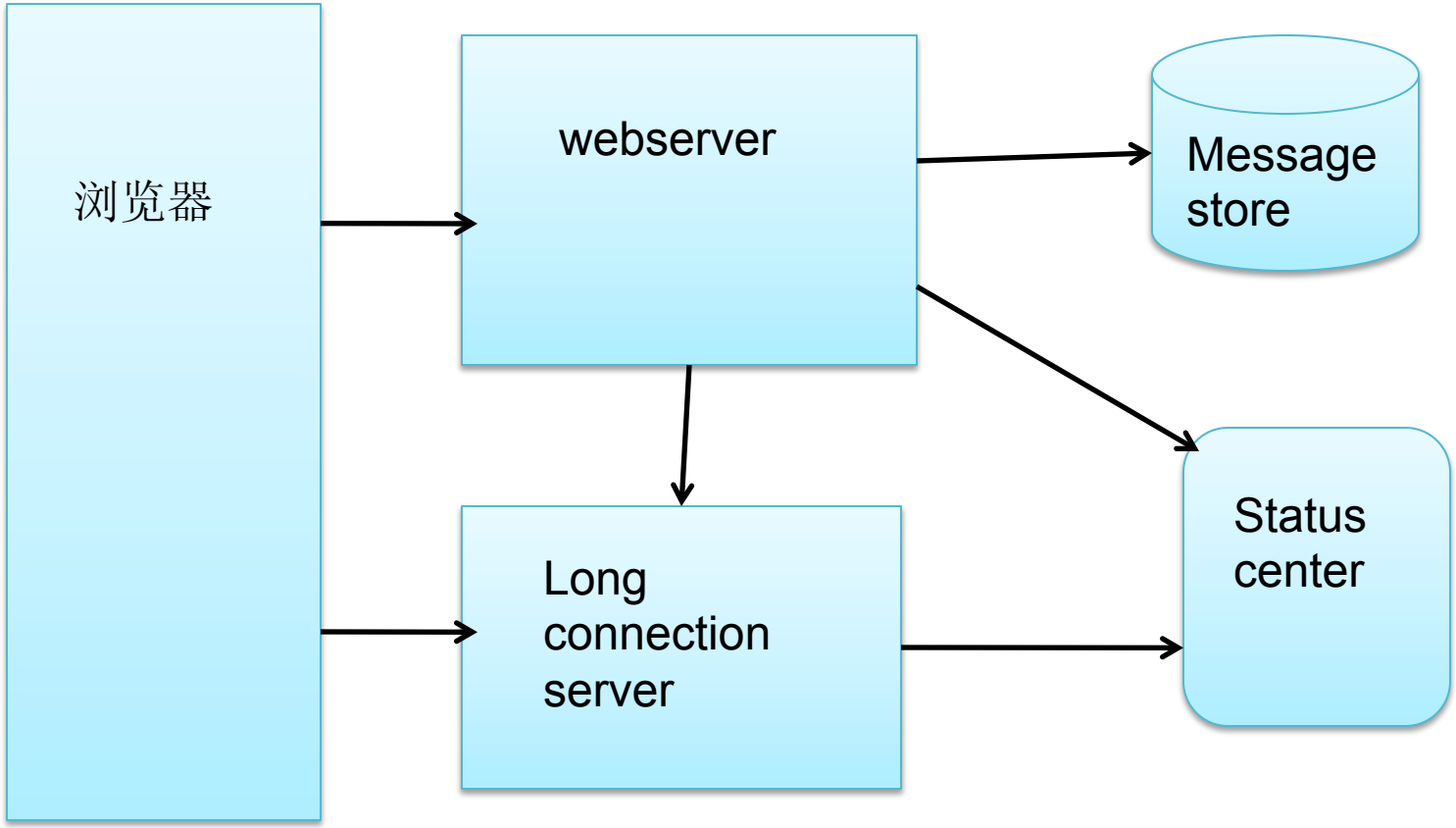
/proc/sys/net/ipv4/ip_local_port_range	向外连接可用端口范围
/proc/sys/net/ipv4/tcp_tw_reuse	time_wait连接重用
/proc/sys/net/ipv4/tcp_tw_recycle	快速回收time_wait连接
/proc/sys/net/ipv4/tcp_max_tw_buckets	最大time_wait连接长度
/proc/sys/net/ipv4/tcp_max_syn_backlog	最大等待处于客户端还没有应答回来的连接数（在三次握手中）
/proc/sys/net/core/somaxconn	每一个处于监听(Listen)状态端口的监听队列的长度(establisthed状态之前)
/proc/sys/net/core/netdev_max_backlog	最大等待cpu处理的包的数目
/proc/sys/fs/file-max	最大打开文体数
/proc/sys/net/ipv4/tcp_fin_timeout	FIN-WAIT-2状态等待回收时间

议程

- http 长连接
- 多页面多域名
- 消息路由
- 相关调优
- **webIM**
- 问题交流

淘宝网 Taobao.com 淘我喜欢！

WebIM体系结构



THANK YOU!