高并发下Linux内核参数调整

2017年08月18日 14:55:22 奇葩也是花 阅读数：3385

用vim打开配置文件：#vim /etc/sysctl.conf

在这个文件中，加入下面的几行内容：

```
net.ipv4.tcp_syncookies = 1
net.ipv4.tcp_tw_reuse = 1
net.ipv4.tcp_tw_recycle = 1
net.ipv4.tcp_fin_timeout = 30
```

输入下面的命令，让内核参数生效：#sysctl -p

net.ipv4.tcp_syncookies = 1

#表示开启SYN Cookies。当出现SYN等待队列溢出时，启用cookies来处理，可防范少量SYN攻击，默认为0，表示关闭；
net.ipv4.tcp_tw_reuse = 1
#表示开启重用。允许将TIME-WAIT sockets重新用于新的TCP连接，默认为0，表示关闭；
net.ipv4.tcp_tw_recycle = 1
#表示开启TCP连接中TIME-WAIT sockets的快速回收，默认为0，表示关闭；
net.ipv4.tcp_fin_timeout
#修改系统默认的 TIMEOUT 时间。

在经过这样的调整之后，除了会进一步提升服务器的负载能力之外，还能够防御小流量程度的DoS、CC和SYN攻击。

此外，如果你的连接数本身就很多，我们可以再优化一下TCP的可使用端口范围，进一步提升服务器的并发能力。依然是往上面的参数文件中，加入下面这些配置：

```
net.ipv4.tcp_keepalive_time = 1200
net.ipv4.ip_local_port_range = 10000 65000
net.ipv4.tcp_max_syn_backlog = 8192
net.ipv4.tcp_max_tw_buckets = 5000
```

#这几个参数，建议只在流量非常大的服务器上开启，会有显著的效果。一般的流量小的服务器上，没有必要去设置这几个参数。

net.ipv4.tcp_keepalive_time = 1200
#表示当keepalive起用的时候，TCP发送keepalive消息的频度。缺省是2小时，改为20分钟。
net.ipv4.ip_local_port_range = 10000 65000
#表示用于向外连接的端口范围。缺省情况下很小：32768到61000，改为10000到65000。（注意：这里不要将最低值设的太低，否则可能会占用掉正常的端口！）
net.ipv4.tcp_max_syn_backlog = 8192
#表示SYN队列的长度，默认为1024，加大队列长度为8192，可以容纳更多等待连接的网络连接数。
net.ipv4.tcp_max_tw_buckets = 6000
#表示系统同时保持TIME_WAIT的最大数量，如果超过这个数字，TIME_WAIT将立刻被清除并打印警告信息。默 认为180000，改为6000。对于Apache、Nginx等服务器，上几行的参数可以很好地减少TIME_WAIT套接字数量，但是对于Squid，效果却不大。此项参数可以控制TIME_WAIT的最大数量，避免Squid服务器被大量的TIME_WAIT拖死。

内核其他TCP参数说明：
net.ipv4.tcp_max_syn_backlog = 65536
#记录的那些尚未收到客户端确认信息的连接请求的最大值。对于有128M内存的系统而言，缺省值是1024，小内存的系统则是128。
net.core.netdev_max_backlog = 32768
#每个网络接口接收数据包的速率比内核处理这些包的速率快时，允许送到队列的数据包的最大数目。
net.core.somaxconn = 32768
#web应用中listen函数的backlog默认会给我们内核参数的net.core.somaxconn限制到128，而nginx定义的NGX_LISTEN_BACKLOG默认为511，所以有必要调整这个值。

net.core.wmem_default = 8388608
net.core.rmem_default = 8388608
net.core.rmem_max = 16777216 #最大socket读buffer,可参考的优化值:873200
net.core.wmem_max = 16777216 #最大socket写buffer,可参考的优化值:873200

net.ipv4.tcp_timestsmps = 0
#时间戳可以避免序列号的卷绕。一个1Gbps的链路肯定会遇到以前用过的序列号。时间戳能够让内核接受这种“异常”的数据包。这里需要将其关掉。
net.ipv4.tcp_synack_retries = 2
#为了打开对端的连接，内核需要发送一个SYN并附带一个回应前面一个SYN的ACK。也就是所谓三次握手中的第二次握手。这个设置决定了内核放弃连接之前发送SYN+ACK包的数量。
net.ipv4.tcp_syn_retries = 2
#在内核放弃建立连接之前发送SYN包的数量。
#net.ipv4.tcp_tw_len = 1
net.ipv4.tcp_tw_reuse = 1
# 开启重用。允许将TIME-WAIT sockets重新用于新的TCP连接。

net.ipv4.tcp_wmem = 8192 436600 873200
# TCP写buffer,可参考的优化值: 8192 436600 873200
net.ipv4.tcp_rmem = 32768 436600 873200
# TCP读buffer,可参考的优化值: 32768 436600 873200
net.ipv4.tcp_mem = 94500000 91500000 92700000
# 同样有3个值,意思是:
net.ipv4.tcp_mem[0]:低于此值，TCP没有内存压力。
net.ipv4.tcp_mem[1]:在此值下，进入内存压力阶段。
net.ipv4.tcp_mem[2]:高于此值，TCP拒绝分配socket。
上述内存单位是页，而不是字节。可参考的优化值是:786432 1048576 1572864

net.ipv4.tcp_max_orphans = 3276800
#系统中最多有多少个TCP套接字不被关联到任何一个用户文件句柄上。
如果超过这个数字，连接将即刻被复位并打印出警告信息。
这个限制仅仅是为了防止简单的DoS攻击，不能过分依靠它或者人为地减小这个值，
更应该增加这个值(如果增加了内存之后)。
net.ipv4.tcp_fin_timeout = 30
#如果套接字由本端要求关闭，这个参数决定了它保持在FIN-WAIT-2状态的时间。对端可以出错并永远不关闭连接，甚至意外当机。缺省值是60秒。2.2 内核的通常值是180秒，你可以按这个设置，但要记住的是，即使你的机器是一个轻载的WEB服务器，也有因为大量的死套接字而内存溢出的风险，FIN- WAIT-2的危险性比FIN-WAIT-1要小，因为它最多只能吃掉1.5K内存，但是它们的生存期长些。

其他：

清理内存：

可以定时去清理缓存

```
echo 1 > /proc/sys/vm/drop_caches

echo 3 > /proc/sys/vm/drop_caches  这个清理掉更彻底
```

----------我是分割线，以下为姓张大佬给的优化建议----------
初始化系统配置
cat /etc/redhat-release 看版本
uname -r 看内核
配置网络IP /etc/sysconfig/network-scripts/ifcfg-eth1
配置dns /etc/resolv.conf
配置hosts /etc/hosts
配置主机名 /etc/sysconfig/network
配置挂载文件 /etc/fstab
配置磁盘

fdisk /dev/xvdb -> n -> p -> 1 -> 回车 -> 回车 -> w
echo '/dev/xvdb1 /data ext4 defaults 0 0' >> /etc/fstab
mkfs.ext4 /dev/xvdb1
mount -a 挂载所有分区
df -hT 查看
配置超时时间
echo 'TMOUT=1800' >> /etc/profile && . /etc/profile
安装必要的软件
yum install -y sysstat lrzsz
yum groupinstall -y 'Development Tools'
#yum groupinstall -y 'x software development' 已弃

设定服务器中文

echo 'LANG="zh_CN.UTF-8"' > /etc/sysconfig/i18n
source /etc/sysconfig/i18n

同步时间
*/30 * * * * /usr/sbin/ntpdate pool.ntp.org > /dev/null 2>&1

系统优化细节
1.禁止root登录 改端口
PermiteMPTYpASSWORDS no #不允许空密码
Port 22022
PermitRootLogin no #不允许root登录
UseDNS no #不使用dns解析
2.创建共用账号,然后用key文件登录，不告诉其密码
TestBusinessUser
3.yum 源改为国内，推荐阿里YuM源，epel扩展源也用阿里 [epel.repo | CentOS-Base.repo]

4.开启防火墙 service iptables start
cat /etc/sysconfig/iptables
Generated by iptables-save v1.4.7 on Fri Apr 22 10:57:48 2016
*filter
:INPUT DROP [0:0]
:FORWARD ACCEPT [0:0]
:OUTPUT ACCEPT [1:140]
:syn-flood - [0:0]
-A INPUT -m state --state RELATED,ESTABLISHED -j ACCEPT
-A INPUT -p tcp -m state --state NEW -m tcp --dport 22022 -j ACCEPT
-A INPUT -p tcp -m state --state NEW -m tcp --dport 80 -j ACCEPT
-A INPUT -p tcp -m state --state NEW -m tcp --dport 3306 -j ACCEPT
-A INPUT -p icmp -m limit --limit 100/sec --limit-burst 100 -j ACCEPT
-A INPUT -p icmp -m limit --limit 1/sec --limit-burst 10 -j ACCEPT
-A INPUT -p tcp -m tcp --tcp-flags FIN,SYN,RST,ACK SYN -j syn-flood
-A INPUT -j REJECT --reject-with icmp-host-prohibited
-A syn-flood -p tcp -m limit --limit 3/sec --limit-burst 6 -j RETURN
-A syn-flood -j REJECT --reject-with icmp-port-unreachable
-A INPUT -p tcp -j DROP
COMMIT

Completed on Fri Apr 22 10:57:48 2016

5.关闭sexlinux　临时关闭：setenforce 0
sed -i 's/SELINUX=enforcing/Selinux=disabled/' /etc/selinux/config

6.设定运行级别为3
sed -i 's/id:5:initdefault:/id:3:initdefault:' /etc/inittab

7.关闭不必要的随机启动项,保留必要的

aegis | agentwatch | iptables | crond | network | ntpd | rsyslog | sshd

8.visudo
sudo授权BusinessSystem用户,便于权限控制管理

9.sshd设置

11.文件描述符加大 查看 ulimit -n

echo '*　-　nofile　65535' > /etc/security/limits.conf
echo 'ulimit -HSn 65535' >> /etc/rc.local
echo 'ulimit -s 65535' >> /etc/rc.local

12.清理clientmqueue目录垃圾文件防止占满磁盘空间 var目录有大量的日志文件 尤其是邮件服务产生的大量没用日志
find /var/spool/clientmqueue -type f | xargs rm -f
可以设置每周六凌晨清理　echo '00 00 * * 6 /bin/bash /data/script/clientmqueue.sh > /dev/null 2>&1'

13.调整内核参数文件,web服务必须优化,提高并发

cat /etc/sysctl.conf
net.ipv4.icmp_echo_ignore_broadcasts = 1
net.ipv4.icmp_ignore_bogus_error_responses = 1
net.ipv4.tcp_syncookies = 1
net.ipv4.conf.all.log_martians = 1
net.ipv4.conf.default.log_martians = 1
net.ipv4.conf.all.accept_source_route = 0
net.ipv4.conf.default.accept_source_route = 0
net.ipv4.conf.all.rp_filter = 1
net.ipv4.conf.default.rp_filter = 1
net.ipv4.conf.all.accept_redirects = 0
net.ipv4.conf.default.accept_redirects = 0
net.ipv4.conf.all.secure_redirects = 0
net.ipv4.conf.default.secure_redirects = 0
net.ipv4.conf.all.send_redirects = 0
net.ipv4.conf.default.send_redirects = 0
kernel.exec-shield = 1
kernel.randomize_va_space = 1
fs.file-max = 65535
kernel.pid_max = 65536
net.core.netdev_max_backlog = 4096
net.ipv4.tcp_window_scaling = 1
net.ipv4.tcp_max_syn_backlog = 4096
net.ipv4.tcp_max_tw_buckets = 4096
net.ipv4.tcp_keepalive_time = 20
net.ipv4.ip_forward = 0

net.ipv4.tcp_mem = 192000 300000 732000
net.ipv4.tcp_rmem = 51200 131072 204800
net.ipv4.tcp_wmem = 51200 131072 204800
net.ipv4.tcp_keepalive_intvl = 5
net.ipv4.tcp_keepalive_probes = 2
net.ipv4.tcp_orphan_retries = 3
net.ipv4.tcp_syn_retries = 3
net.ipv4.tcp_synack_retries = 3
net.ipv4.tcp_retries2 = 5
net.ipv4.tcp_fin_timeout = 30
net.ipv4.tcp_max_orphans = 2000
net.ipv4.tcp_tw_reuse = 1
net.ipv4.tcp_tw_recycle = 1
vm.min_free_kbytes=409600
vm.vfs_cache_pressure=200
vm.swappiness = 40
vm.dirty_expire_centisecs = 1500
vm.dirty_writeback_centisecs = 1000
vm.dirty_ratio = 2
vm.dirty_background_ratio = 100
运行sysctl -p 启用内核配置

14.目录结构整理，所有第三方文件，都安装到/data/ 目录下
data目录一定不要用系统盘，单独用数据盘

15.如果可能，把必要的系统文件锁定，及时黑进来也改不了关键文件
  但是要慎用，修改系统文件要记得 修改之前一定要备份
  chattr +i /etc/group 不允许添加用户
  chattr +i /etc/inittab
  chattr +i /etc/shadow 不允许修改密码
  chattr +i /etc/rc.local

16.隐藏服务器版本信息 cp /etc/issue /etc/issue.base
cat /dev/null > /etc/issue

17.全面中文支持
[xxx@iZ23cwcora9Z ~]$ locale
LANG=zh_CN.UTF-8
LC_CTYPE="zh_CN.UTF-8"
LC_NUMERIC="zh_CN.UTF-8"
LC_TIME="zh_CN.UTF-8"
LC_COLLATE="zh_CN.UTF-8"
LC_MONETARY="zh_CN.UTF-8"
LC_MESSAGES="zh_CN.UTF-8"
LC_PAPER="zh_CN.UTF-8"
LC_NAME="zh_CN.UTF-8"
LC_ADDRESS="zh_CN.UTF-8"
LC_TELEPHONE="zh_CN.UTF-8"
LC_MEASUREMENT="zh_CN.UTF-8"
LC_IDENTIFICATION="zh_CN.UTF-8"
LC_ALL=

18.参考资料：http://lovers.blog.51cto.com/5850489/1585178

innodb config
#innodb_data_file_path = ibdata1:50M;ibdata2:50M:autoextend:max:500M
Set buffer pool size to 50-80% of your computer's memory

```
innodb_buffer_pool_size = 2G
innodb_additional_mem_pool = 16M
# Set the log file size to about 25% of the buffer pool size
innodb_log_file_size = 512M
innodb_log_files_in_group = 2
# innodb_log_buffer_size set 2-8M
innodb_log_buffer_size = 3M
innodb_flush_log_at_trx_commit = 2
innodb_lock_wait_timeout = 50
innodb_file_per_table = 1
innodb_open_files = 800
innodb_flush_method = O_DIRECT
innodb_max_dirty_pages_pct = 90
lower_case_table_names=1


skip-external-locking
key_buffer_size = 16M
max_allowed_packet = 16M
table_open_cache = 64
sort_buffer_size = 1M
net_buffer_length = 8K
read_buffer_size = 1M
read_rnd_buffer_size = 512K
slave_skip_errors = all


wait_timeout = 240
interactive_timeout = 20
net_read_timeout = 20
net_write_timeout = 30
skip-name-resolve


max_connections = 2000
max_user_connections = 1000
```