Subject: TCP: time wait bucket table overflow
Posted by sspt on Wed, 26 Dec 2007 00:14:01 GMT
View Forum Message <> Reply to Message

I'm experiencing some network issues since i've moved from 2.6.18-8.1.8.el5.028stab039.1.

Quote:
Dec 25 14:17:05 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:17:10 hid01 kernel: printk: 5 messages suppressed.
Dec 25 14:17:10 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:17:15 hid01 kernel: printk: 4 messages suppressed.
Dec 25 14:17:15 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:17:20 hid01 kernel: printk: 9 messages suppressed.
Dec 25 14:17:20 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:17:26 hid01 kernel: printk: 2 messages suppressed.
Dec 25 14:17:26 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:17:32 hid01 kernel: printk: 11 messages suppressed.
Dec 25 14:17:32 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:17:36 hid01 kernel: printk: 1 messages suppressed.
Dec 25 14:17:36 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:17:44 hid01 kernel: printk: 4 messages suppressed.
Dec 25 14:17:44 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:17:46 hid01 kernel: printk: 3 messages suppressed.
Dec 25 14:17:46 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:17:50 hid01 kernel: printk: 5 messages suppressed.
Dec 25 14:17:50 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:17:55 hid01 kernel: printk: 1 messages suppressed.
Dec 25 14:17:55 hid01 kernel: TCP: time wait bucket table overflow
Dec 25 14:19:09 hid01 kernel: printk: 4 messages suppressed.
Dec 25 14:19:09 hid01 kernel: TCP: time wait bucket table overflow

Quote:
[root@hid01 ~]# sysctl -a | grep tw
net.ipv4.tcp_max_tw_buckets_ub = 165360
net.ipv4.tcp_max_tw_kmem_fraction = 384
net.ipv4.tcp_tw_reuse = 0
net.ipv4.tcp_tw_recycle = 0
net.ipv4.tcp_max_tw_buckets = 1800000

Quote:
[root@hid01 ~]# cat /proc/meminfo
MemTotal: 8075376 kB
MemFree: 32856 kB
Buffers: 267632 kB
Cached: 5551208 kB

SwapCached: 0 kB
Active: 3600512 kB
Inactive: 3416264 kB
HighTotal: 0 kB
HighFree: 0 kB
LowTotal: 8075376 kB
LowFree: 32856 kB
SwapTotal: 0 kB
SwapFree: 0 kB
Dirty: 4804 kB
Writeback: 0 kB
AnonPages: 1187560 kB
Mapped: 189116 kB
Slab: 912860 kB
PageTables: 56928 kB
NFS_Unstable: 0 kB
Bounce: 0 kB
CommitLimit: 4037688 kB
Committed_AS: 5340628 kB
VmallocTotal: 34359738364 kB
VmallocUsed: 16876 kB
VmallocChunk: 34359710628 kB
HugePages_Total: 0
HugePages_Free: 0
HugePages_Rsvd: 0
Hugepagesize: 2048 kB

Quote:
Linux hid01 2.6.18-53.el5.028stab051.1 #1 SMP Fri Nov 30 02:52:22 MSK 2007 x86_64 x86_64 x86_64 GNU/Linux

The node speeds (VE0) don't seems to be affected as i can move data to another box at more than 100Mbits, although speeds from each VE to the other box are like 100kbits.

Any ideas?

Subject: Re: TCP: time wait bucket table overflow
Posted by vaverin on Wed, 26 Dec 2007 09:07:23 GMT
View Forum Message <> Reply to Message

these messages are printed in tcp_time_wait() function
http://git.openvz.org/?p=linux-2.6.18-openvz;a=blob;f=net/ipv4/tcp_minisocks.c;h=e280537ef09aa804f79e733eaee1721a07b5aca c;hb=HEAD
in the follwong cases:

1) Per-system tw_count is greater than per-system max_tw_buckets limit:
tw_count < sysctl_max_tw_buckets
net.ipv4.tcp_max_tw_buckets = 1800000
IMHO very unlikely in your situation

2) Per-VE counter is greater than per-VE max_tw_buckets limit:
tw_count < sysctl_tcp_max_tw_buckets_ub
net.ipv4.tcp_max_tw_buckets_ub = 165360

IMHO unlikely too

3) inside VE tw_buckets eats too many memory (greater than allowed fraction of kmemsize)
net.ipv4.tcp_max_tw_kmem_fraction = 384 means 38.4% of kmemsize

IMHO very likely

4) kmemsize shortage inside VE -- in this case you should have new failcounters for kmemsize parameter.
IMHO unlikely

Resume:
You need to find VE that generates these messages and increase its kmemsize. Or You can increase setting for net.ipv4.tcp_max_tw_kmem_fraction sysctl

ToDo: make this message cleaner
http://bugzilla.openvz.org/show_bug.cgi?id=767

Subject: Re: TCP: time wait bucket table overflow
Posted by ittec on Fri, 24 Oct 2008 09:54:50 GMT
View Forum Message <> Reply to Message

Hi

very interesting post. Could someone explain two things?

1) Per-system tw_count is greater than per-system max_tw_buckets limit:
tw_count < sysctl_max_tw_buckets

2) Per-VE counter is greater than per-VE max_tw_buckets limit:
tw_count < sysctl_tcp_max_tw_buckets_ub

I didn't understand when Vaverin talk about Per-System tw_count. What is tw_count?

And in the bottom of his post, he speaks about increase net.ipv4.tcp_max_tw_kmem_fraction or increase kmemsize of VE in conf. Really is a mistery for me how to solve the issue of Time Wait Bucket Overflow. Until now when issue appears I increase kmemsize of VE and restart the VE. But never Ive tried to increase kmem_fraction. What can I do in first time, increase kmemsize or

kmem_fraction?

In the other hand, vzcfgvalidate show "success" result in conf of VEs with "Time Wait Bucket Overlflow" issue.

Thanks!

Subject: Re: TCP: time wait bucket table overflow
Posted by locutius on Fri, 24 Oct 2008 11:10:52 GMT
View Forum Message <> Reply to Message

i have also seen vzcfgvalidate give success to a CT config when later in that CT's life it needs an increase to kmemsize

i read vaverin to say "increase kmemsize"

i would have expected any resource contraint to show in the beancounters

---

Subject: Re: TCP: time wait bucket table overflow
Posted by den on Fri, 24 Oct 2008 12:55:30 GMT
View Forum Message <> Reply to Message

tw_count is the amount of sockets in the TCP_TIMEWAIT state. They live in that state for a long time (~5 min) and can eat a lot of kernel memory.

So, it is natural that they are limited.

There are two type of limits:
* global (for all environments including VE0)
* per/container (calculated on the base of kmemsize)

Vaverin estimations are correct from my POW.

Pls contact me if the topic is still unclear.

Regards,
Den

Subject: Re: TCP: time wait bucket table overflow
Posted by ittec on Fri, 24 Oct 2008 14:11:18 GMT
View Forum Message <> Reply to Message

Hi Den, thanks by your help.

so Is there something shell command to see tw_count? per container or VEO?

Im really interested into kmemsize issues because I have some web applications that spend a lot kmemsize. For an example, I have a containers with 2GB of RAM(212644 privvmpages) but kmemsize barrier:limit is 50000000:60000000. I think is a enormous lot of memory but if I had not increase this value now application will be stopped.

Is interesting kmem_fraction too. Until now I have not changed this value. And it can be interesting perhaps to do some test about it?

My Kernel is up to date: 2.6.18-92.1.1.el5.028stab057.2PAE

Subject: Re: TCP: time wait bucket table overflow
Posted by den on Fri, 24 Oct 2008 17:41:23 GMT
View Forum Message <> Reply to Message

There is no way to obtain tw_count directly, but you can count it manually from an output of netstat.

Subject: Re: TCP: time wait bucket table overflow
Posted by locutius on Sun, 26 Oct 2008 07:21:18 GMT
View Forum Message <> Reply to Message

ittec wrote on Fri, 24 October 2008 10:11
I have a containers with 2GB of RAM(212644 privvmpages)

there is a problem

2GB RAM = 524288

did you also set the vmguarpages (2GB), oomguarpages (2GB) and shmpages (52428)?

Subject: Re: TCP: time wait bucket table overflow
Posted by ittec on Mon, 27 Oct 2008 12:06:26 GMT
View Forum Message <> Reply to Message

here is a problem

2GB RAM = 524288

did you also set the vmguarpages (2GB), oomguarpages (2GB) and shmpages (52428)?

Hi!

Until now, I setup VPS with vzsplit. The most of times I need to increase value of some parameters like privvmpages but not others. However, when I increase privvmpages value I don't do the same with vmguarpages, oomguarpges or shmpageS? Have I to do it? I take a results of vzcfgvalidate to guide me.

This is the .conf about one current VPS with a high KMEMSIZE failcount and a lot of troubles of performance. I have a lot of "Time Bucket Overflow" on logs about this VPS. Is a common problemm with our web applications.

```
KMEMSIZE="15000000:16000000"
LOCKEDPAGES="256:256"
PRIVVMPAGES="262144:262144"
SHMPAGES="21504:21504"
NUMPROC="240:240"
PHYSPAGES="0:2147483647"
VMGUARPAGES="33792:2147483647"
OOMGUARPAGES="26112:2147483647"
NUMTCPSOCK="720:720"
NUMFLOCK="188:206"
NUMPTY="16:16"
NUMSIGINFO="256:256"
TCPSNDBUF="1720320:3563520"
TCPRCVBUF="1720320:3563520"
OTHERSOCKBUF="1126080:2097152"
DGRAMRCVBUF="262144:262144"
NUMOTHERSOCK="360:360"
DCACHESIZE="3409920:3624960"
NUMFILE="9312:9312"
AVNUMPROC="180:180"
NUMIPTENT="128:128"

# Disk quota parameters (in form of softlimit:hardlimit)
DISKSPACE="50000000:600000000"
DISKINODES="500000:500000"
QUOTATIME="0"

# CPU fair sheduler parameter
CPUUNITS="50000"

VE_ROOT="/vz/root/$VEID"
VE_PRIVATE="/vz/private/$VEID"
OSTEMPLATE="centos-5-i386-minimal"
ORIGIN_SAMPLE="vps.basic"
IP_ADDRESS="one.ip.addre.ss"
```

CPULIMIT="90"
CPUS="3"

Do you see some error on setup? Im curious about it and how to resolve it forever.

Thanks!

Subject: Re: TCP: time wait bucket table overflow
Posted by locutius on Mon, 27 Oct 2008 12:40:55 GMT
View Forum Message <> Reply to Message

these are my settings for RAM (MB, pages):

256 65536
vzctl set 101 --privvmpages 256MB:256MB --save
vzctl set 101 --vmguarpages 256MB:256MB --save
vzctl set 101 --oomguarpages 256MB:256MB --save
vzctl set 101 --shmpages 6552:6552 --save

512 131072
vzctl set 123 --privvmpages 512m:512m --save
vzctl set 123 --vmguarpages 512m:512m --save
vzctl set 123 --oomguarpages 512m:512m --save
vzctl set 123 --shmpages 13106:13106 --save

768 196608
vzctl set 102 --privvmpages 768m:1024m --save
vzctl set 102 --vmguarpages 768m:768m --save
vzctl set 102 --oomguarpages 768m:768m --save
vzctl set 102 --shmpages 19660:19660 --save

1024 262144
vzctl set 129 --privvmpages 1024m:1024m --save
vzctl set 129 --vmguarpages 1024m:1024m --save
vzctl set 129 --oomguarpages 1024m:1024m --save
vzctl set 129 --shmpages 26214:26214 --save

vzctl set 129 --privvmpages 262144:262144 --save
vzctl set 129 --vmguarpages 262144:262144 --save
vzctl set 129 --oomguarpages 262144:262144 --save
vzctl set 129 --shmpages 26214:26214 --save

2048 524288
vzctl set 1001 --privvmpages 2048m:2048m --save

vzctl set 1001 --vmguarpages 2048m:2048m --save
vzctl set 1001 --oomguarpages 2048m:2048m --save
vzctl set 1001 --shmpages 52428:52428 --save

_____ __

increase your KMEMSIZE by a factor of 25

KMEMSIZE is not accounted in pages (4kb)

Subject: Re: TCP: time wait bucket table overflow
Posted by ittec on Mon, 27 Oct 2008 16:25:34 GMT
View Forum Message <> Reply to Message

Well, im starting to understand some things. Until now, for me, setup a VPS with 512 MB was job of privvmpages. But now I know that one thing is MB of RAM guaranteed(vmguarpages) and other is the theoretical limit of RAM that can be allocated(privvmpages). So now im going to check my differents setup to adjust this parameters.

But let me explain something. After change some configurations of some VPS I ran vzcfgvalidate to check some error, and now it says me that I need to change setups to a strange values

Error: limit should be = 2147483647 for vmguarpages (currently, 262144)
set to 2147483647
 (y/n) [y] n
Error: limit should be = 2147483647 for oomguarpages (currently, 262144)
set to 2147483647
 (y/n) [y] n
Warning: privvmpages.bar should be > 262144 (currently, 131072)
set to 262144
 (y/n) [y] n

I don't understand why I have to change some .bar to 2147483647 ??? Anybody nows why? Or I don't want to use vzcfgvalidate like a guide?

Thanks

Subject: Re: TCP: time wait bucket table overflow
Posted by kevinm on Sun, 17 May 2009 17:28:58 GMT
View Forum Message <> Reply to Message

den wrote on Fri, 24 October 2008 13:41There is no way to obtain tw_count directly, but you can count it manually from an output of netstat.

the command

Quote:ss -s

will output

Quote:
Total: 5221 (kernel 5444)
TCP: 1981 (estab 3, closed 1158, orphaned 0, synrecv 0, timewait 1104/0), ports 766

Transport	Total	IP	IPv6
*	5444	-	-
RAW	0	0	0
UDP	100	92	8
TCP	823	704	119
INET	923	796	127
FRAG	0	0	0

timewait == sockets in timewait state, ports == registered system ports allocated.

KEv