Subject: TCP Packets get lost under moderate network load
Posted by klathor on Thu, 22 Jan 2009 22:22:56 GMT
View Forum Message <> Reply to Message

Hello!

I've filed this as a bug (http://bugzilla.openvz.org/show_bug.cgi?id=1156),
but I was hoping to get some feedback/suggestions to make sure I'm not doing
something idiotic.
I apologize in advance for the length of this note...

Background:
I'm converting a fairly basic legacy webserver to an OpenVZ VPS. I've been using
OpenVZ for a few months on some infrastructure stuff and it's been working fine,
but this is the first production server I've tried converting. It's running
apache and php, gets about 2000 hits/min (mostly images, php is just for some
templates). The new OpenVZ machine is a Dell, dual Xeon 3G, 4G ram, this new VPS
is the only thing running on it.

Issue:
The morning after I converted, http monitors started to throw alerts just as
traffic was ramping up. Upon logging in, I saw 500+ sockets in SYN_RECV state.
Restarting the webserver would clear them out, but within a minute or two they
would reach 100+. Also, there were many "Orphaned socket dropped" messages in
the kernel logs, but no barriers were reached. After many false leads, I
discovered that the problem was very similar to the problem reported by Max
Deineko in January of 2008:
http://forum.openvz.org/index.php?t=msg&goto=25678
Unfortunately no resolution was posted.

I have reverted back to the original server and done some testing, but
unfortunately I have to break our live site to do so because the issue only
occurs when the VPS is hit with many different IPs. I can't replicate the
behavior even by running ab from ~7 machines. (Doing that only makes load climb
and the net link saturate, but SYN_RECV states do not pile up.)

Possible clues:
-Out of desperation I turned off iptables entirely (no rules, no modules loaded)
on both the HN & VPS, which appeared to clear up everything.
-The issue reappears with iptables turned ON on the HN but turned OFF on the VPS
(IPTABLES="" in /etc/vz/vz.conf).
-I've tried logging all INVALID state packets in case it was some sort of
conntrack issue, but I saw no hits for the problem connections captured.
-Happens whether syncookies are turned on or off.
-I've looked over the links referenced in Mr. Deineko's thread, but I don't
believe this is a buffer size or window size issue as the problem happens while
initiating the connection.

Does anyone have any hints? Anyone out there running very busy webservers? I've seen some random google hits for people complaining about DDOS problems with their OpenVZ sites which may be related since the issue looks like a DDOS at first glance (many many SYN_RECV states). Any info at all would be greatly appreciated.

Thanks,
-JayKim

Sample connection:
netstat -anp:
tcp 0 0 SERVER:80 CLIENT:41842 SYN_RECV -

tcpdump (same thing seen on HN and VPS):
11:09:38.296692 IP CLIENT.41842 > SERVER.80: S 4100993638:4100993638(0) win 64512 <mss 1460,nop,nop,sackOK>
11:09:41.275425 IP CLIENT.41842 > SERVER.80: S 4100993638:4100993638(0) win 64512 <mss 1460,nop,nop,sackOK>
11:09:41.275468 IP SERVER.80 > CLIENT.41842: S 1902995022:1902995022(0) ack 4100993639 win 5840 <mss 1460,nop,nop,sackOK>
11:09:41.319954 IP CLIENT.41842 > SERVER.80: . ack 1 win 64512
11:09:41.346429 IP CLIENT.41842 > SERVER.80: P 1:462(461) ack 1 win 64512
11:09:44.982597 IP SERVER.80 > CLIENT.41842: S 1902995022:1902995022(0) ack 4100993639 win 5840 <mss 1460,nop,nop,sackOK>
11:09:45.025814 IP CLIENT.41842 > SERVER.80: . ack 1 win 64512
11:09:47.295413 IP CLIENT.41842 > SERVER.80: P 1:462(461) ack 1 win 64512
11:09:50.983478 IP SERVER.80 > CLIENT.41842: S 1902995022:1902995022(0) ack 4100993639 win 5840 <mss 1460,nop,nop,sackOK>
11:09:51.030111 IP CLIENT.41842 > SERVER.80: . ack 1 win 64512
11:09:59.334340 IP CLIENT.41842 > SERVER.80: P 1:462(461) ack 1 win 64512
11:10:03.185371 IP SERVER.80 > CLIENT.41842: S 1902995022:1902995022(0) ack 4100993639 win 5840 <mss 1460,nop,nop,sackOK>
11:10:03.226847 IP CLIENT.41842 > SERVER.80: . ack 1 win 64512
11:10:27.195543 IP SERVER.80 > CLIENT.41842: S 1902995022:1902995022(0) ack 4100993639 win 5840 <mss 1460,nop,nop,sackOK>

uname -a:
Linux SERVER 2.6.18-92.1.13.el5.028stab059.6PAE #1 SMP Fri Nov 14 20:46:53 MSK 2008 i686 i686 i386 GNU/Linux

rpm -qa | grep vz:
vzquota-3.0.12-1
vzctl-3.0.23-1
vztmpl-centos-5-2.0-3
ovzkernel-PAE-2.6.18-92.1.13.el5.028stab059.6
vzyum-2.4.0-11
vzctl-lib-3.0.23-1
vzpkg-2.7.0-18

vzrpm43-python-4.3.3-7_nonptl.6
vzrpm43-4.3.3-7_nonptl.6
vzrpm44-4.4.1-22.5
vzrpm44-python-4.4.1-22.5

/etc/redhat-release (both HN and VPS):
CentOS release 5.2 (Final)

various sysctls:
net.core.rmem_default = 113664
net.core.wmem_default = 113664
net.core.rmem_max = 131071
net.core.wmem_max = 131071
net.ipv4.tcp_mem = 16384 20480 24576
net.ipv4.tcp_rmem = 4096 87380 655360
net.ipv4.tcp_wmem = 4096 16384 655360
net.ipv4.tcp_max_syn_backlog = 1024
net.ipv4.netfilter.ip_conntrack_tcp_timeout_syn_recv = 60
net.ipv4.netfilter.ip_conntrack_tcp_timeout_syn_sent = 120
net.ipv4.tcp_synack_retries = 5
net.ipv4.tcp_syn_retries = 5

VPS limits (generated from vzsplit):

uid	resource	held	maxheld	barrier	limit	failcnt
3000:	kmemsize	5098959	26116859	29720985	32693083	0
	numtcpsock	9	359	1333	1333	0
	tcpsndbuf	71552	2126436	4447027	9906995	0
	tcprcvbuf	132912	999560	4447027	9906995	0