Subject: Re: OpenVZ, Bind and stalling TCP connections.
Posted by agl241 on Fri, 26 Sep 2008 15:25:03 GMT
View Forum Message <> Reply to Message

Thanks for your response.

My previous tcpdump output was produced from inside of the container (tcpdumps on the hardware node itself produce similar results).

After further investigation and trying to trace Bind I've found that indeed named process listening on the TCP port does not receive any notification about incoming connections when it is locked.

At the same time I've found that number of listen (accept) queue overflows inside the VE is very high (for the 1 day uptime the appriopriate numbers in /proc/net/netstat was: ListenOverflows - 95584, ListenDrops - 95584) - this counter is increased each time incoming ACK packet (last from the 3-way handshake) is received, but accept queue for the listening socket is full, so connection is silenly dropped.

My hypothesis is that for some unknown reason accept queue of the
listening Bind socket is filled up and later it can't recover itself. No further incoming connections to this port can reach ESTABLISHED state , and on way to recover named is to restart it, so it allocates new listening socket. I think that locking wasn't the result of the sync flooding, because single syncs without later acknowledgment don't reach this fragment of code (tcp_v4_syn_recv_sock).

At the moment I can't reproduce error, so it makes further debugging difficult - but I'm sure that it will appear soon again.

Regards.

Andrzej Lemieszek

Below are little formated contents of my "/proc/net/netstat":

1: SyncookiesSent - 0
2: SyncookiesRecv - 0
3: SyncookiesFailed - 0
4: EmbryonicRsts - 60553
5: PruneCalled - 0
6: RcvPruned - 0
7: OfoPruned - 0
8: OutOfWindowIcmps - 1
9: LockDroppedIcmps - 0
10: ArpFilter - 0
11: TW - 0
12: TWRecycled - 0
13: TWKilled - 0
14: PAWSPassive - 0

15: PAWSActive - 0
16: PAWSEstab - 2
17: DelayedACKs - 22483
18: DelayedACKLocked - 306
19: DelayedACKLost - 125
20: ListenOverflows - 95584
21: ListenDrops - 95584
22: TCPPrequeued - 89
23: TCPDirectCopyFromBacklog - 0
24: TCPDirectCopyFromPrequeue - 40
25: TCPPrequeueDropped - 0
26: TCPHPHits - 290407
27: TCPHPHitsToUser - 0
28: TCPPureAcks - 9559422
29: TCPHPAcks - 6528630
30: TCPRenoRecovery - 1
31: TCPSackRecovery - 387
32: TCPSACKReneging - 0
33: TCPFACKReorder - 2
34: TCPSACKReorder - 5
35: TCPRenoReorder - 1
36: TCPTSReorder - 46
37: TCPFullUndo - 35
38: TCPPartialUndo - 1491
39: TCPDSACKUndo - 0
40: TCPLossUndo - 251
41: TCPLoss - 396
42: TCPLostRetransmit - 0
43: TCPRenoFailures - 0
44: TCPSackFailures - 5559
45: TCPLossFailures - 2396
46: TCPFastRetrans - 648
47: TCPForwardRetrans - 64
48: TCPSlowStartRetrans - 19419
49: TCPTimeouts - 14825
50: TCPRenoRecoveryFail - 0
51: TCPSackRecoveryFail - 82
52: TCPSchedulerFailed - 0
53: TCPRcvCollapsed - 0
54: TCPDSACKOldSent - 124
55: TCPDSACKOfoSent - 0
56: TCPDSACKRecv - 383
57: TCPDSACKOfoRecv - 0
58: TCPAbortOnSyn - 0
59: TCPAbortOnData - 44
60: TCPAbortOnClose - 32
61: TCPAbortOnMemory - 0
62: TCPAbortOnTimeout - 6259

63: TCPAbortOnLinger - 0
64: TCPAbortFailed - 0
65: TCPMemoryPressures - 0