Subject: Re: [PATCH 0/16] Pid namespaces
Posted by Sukadev Bhattiprolu on Tue, 17 Jul 2007 04:23:39 GMT
View Forum Message <> Reply to Message

Pavel Emelianov [xemul@openvz.org] wrote:
| >My x86_64 system boots fine but crashes as below, when I run my
| >'pidns_exec' test with a simple program that prints getpid(), getppid()
| >etc of the process in the child pid ns.
| >
| >Pls see
| >
| >http://www.geocities.com/sukadevb/Pidspace/2.6.22-rc6-mm1-pavel-1.tgz
| >
| >for the patches I currently have applied and let me know if I need more
| >on top.
| >
| >And see
| >
| >http://www.geocities.com/sukadevb/Pidspace/Test1/
| >
| >for the test programs. You may need to run the 'mypid-loop.x' script
| >to repro the crash. The pidns_exec.c program calls clone() with
| >CLONE_NEWPID
| >and execs the given program (it was included in Patch 0 of the patchset I
| >posted to Containers).
| >
| >Suka
| >
| >login: Unable to handle kernel NULL pointer dereference at
| >00000000000002fc RIP:
| > [<ffffffff802b9e5e>] proc_get_sb+0xfb/0x138
| >PGD 104d53067 PUD 104d4d067 PMD 0
| >Oops: 0002 [1] SMP
| >CPU 2
| >Modules linked in:
| >Pid: 3279, comm: pidns_exec Not tainted 2.6.22-rc6-mm1-ovz1 #10
| >RIP: 0010:[<ffffffff802b9e5e>] [<ffffffff802b9e5e>] proc_get_sb+0xfb/0x138
| >RSP: 0018:ffff8101029d7d28 EFLAGS: 00010202
| >RAX: ffff810100651840 RBX: ffff810104461400 RCX: ffff810100651878
| >RDX: 0000000000000000 RSI: ffffffff806e5460 RDI: 0000000000000238
| >RBP: ffff810102d886f8 R08: ffff810104461400 R09: ffff810100026000
| >R10: 0000000000000000 R11: 0000000000000002 R12: ffff8101029c6000
| >R13: 0000000002000000 R14: ffffffff806ee920 R15: ffff810102088cc0
| >FS: 00002b0b499ec6f0(0000) GS:ffff81010069c3c0(0000)
| >knlGS:0000000000000000
| >CS: 0010 DS: 0000 ES: 0000 CR0: 000000008005003b
| >CR2: 00000000000002fc CR3: 000000010381b000 CR4: 00000000000006e0
| >DR0: 0000000000000000 DR1: 0000000000000000 DR2: 0000000000000000

```
| >DR3: 0000000000000000 DR6: 00000000ffff0ff0 DR7: 0000000000000400
| >Process pidns_exec (pid: 3279, threadinfo ffff8101029d6000, task
| >ffff81010269e7f0)
| >Stack:  ffff810102088cc0 ffff810102088cc0 00000000fffffff4 ffffffff806ee920
| > ffffffff8065f9d9 ffff8101029c6000 0000000002000000 ffffffff80287164
| > 00000000000000d0 ffff8101029c6000 ffffffff806e5460 ffff8101029c6000
| >Call Trace:
| > [<ffffffff80287164>] vfs_kern_mount+0x4f/0x8b
| > [<ffffffff802b9cf4>] pid_ns_prepare_proc+0x13/0x2e
| > [<ffffffff80245be3>] copy_pid_ns+0xd7/0x164
| > [<ffffffff8024af34>] create_new_namespaces+0xde/0x192
| > [<ffffffff8024b0aa>] copy_namespaces+0x4b/0x85
| > [<ffffffff802347e2>] copy_process+0xcb4/0x1439
| > [<ffffffff8020bbee>] system_call+0x7e/0x83
| > [<ffffffff8023556a>] do_fork+0x6c/0x1e7
| > [<ffffffff8020bf07>] ptregscall_common+0x67/0xb0
| >
|
| Here's the patch fixing the problem.
```

Yes, I think it fixes the problem I was seeing before.
However, my next test failed. This time I run:

```
 $ cat ./lxc-wrap.sh
 #!/bin/sh

 if [ $$ -eq 1 ]; then
  echo "$$ (`basename $0`): Executing in nested pid ns..."
 fi

 port_num=$1;
 if [ $# -lt 1 ]; then
     port_num=2709
 fi

 mount -t proc lxcproc /proc
 /usr/sbin/sshd -D -p $port_num
 umount /proc


 $ ./pidns_exec ./lxc-wrap.sh
```

This creates a new pid ns and an sshd in that ns. From another window
I ssh to the system with port number 2709 and I am in the new pid ns.
(a simple prog to print getpid(), getppid() etc seems to be fine).

But "ps -e" fails:

```
$ ps -e
Error: /proc must be mounted
  To mount /proc at boot you need an /etc/fstab line like:
      /proc   /proc   proc    defaults
  In the meantime, mount /proc /proc -t proc

$ mount -t proc none /proc
mount: /proc is busy
```

If I comment out the "mount" command in the lxc-wrap.sh script above,
this mount seems to work, but "ps -e" still reports the same error.

| So, Suka, I propose that you review my patches, point out things that you
| don't like and would like to see your code instead.

I am reviewing them and I thought we wanted to start with my patchset. I
started porting my patches to rc6-mm1 plus patches we already sent to -mm.
Well, I can hold off on that and review/test your patches.

| After all I will re-split the set with your fixes, mark
| some patches with your From: and send them to Andrew. What do you think?

I don't mind re-splitting the patches, but would that be more work for us
and would Andrew and other reviewers prefer patches split logically. For
instance if you and I contributed to a patch, can we just put both our
Signed-off on one patch rather than splitting it ?

```
|
| ---
|
| --- ./fs/proc/root.c.procpidnsfix 2007-07-16 10:32:00.000000000 +0400
| +++ ./fs/proc/root.c 2007-07-16 12:34:35.000000000 +0400
| @@ -79,8 +79,6 @@ static int proc_get_sb(struct file_syste
|   if (!ei->pid)
|    ei->pid = find_get_pid(1);
|   sb->s_flags |= MS_ACTIVE;
| -
| -  mntput(ns->proc_mnt);
|   ns->proc_mnt = mnt;
|  }
```

Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers