Subject: Re: Re: [PATCH] allow a task to join a pid namespace
Posted by Daniel Lezcano on Tue, 05 Jun 2012 13:18:59 GMT
View Forum Message <> Reply to Message

On 06/05/2012 02:53 PM, Glauber Costa wrote:
> On 06/05/2012 04:52 PM, Daniel Lezcano wrote:
>> On 06/05/2012 12:00 PM, Glauber Costa wrote:
>>> On 06/05/2012 01:37 PM, Glauber Costa wrote:
>>>> On 06/05/2012 01:36 PM, Daniel Lezcano wrote:
>>>>> On 06/04/2012 03:33 PM, Glauber Costa wrote:
>>>>>> Currently, it is possible for a process to join existing
>>>>>> net, uts and ipc namespaces. This patch allows a process to join an
>>>>>> existing pid namespace as well.
>>>>>>
>>>>>> For that to remain sane, some restrictions are made in the calling
>>>>>> process:
>>>>>>
>>>>>> * It needs to be in the parent namespace of the namespace it
>>>>>> wants to
>>>>>> jump to
>>>>>> * It needs to sit in its own session and group as a leader.
>>>>>>
>>>>>> The rationale for that, is that people want to trigger actions in a
>>>>>> Container
>>>>>> from the outside. For instance, mainstream linux recently gained the
>>>>>> ability
>>>>>> to safely reboot a container. It would be desirable, however, that
>>>>>> this
>>>>>> action is triggered from an admin in the outside world, very much
>>>>>> like a
>>>>>> power switch in a physical box.
>>>>>>
>>>>>> This would also allow us to connect a console to the container,
>>>>>> provide a
>>>>>> repair mode for setups without networking (or with a broken one),
>>>>>> etc.
>>>>>
>>>>> Hi Glauber,
>>>>>
>>>>> I am in favor of this patch but I think the pidns support won't be
>>>>> complete and some corner-cases are not handled.
>>>>>
>>>>> May be you can look at Eric's patchset [1] where, IMO, everything is
>>>>> taken into account. Some of the patches may be already upstream.
>>>>>
>>>>> Thanks
>>>>> -- Daniel
>>>>

>>>> I don't remember seeing such patchset in the mailing lists, but that
>>>> might be my fault, due to traffic...
>>>>
>>>> I'll take a look. If it does what I need, I can just drop this.
>>>>
>>>
>>> Ok. In a quick look, it does not seem to go all the way. This is just
>>> by reading, but your reboot patch, for instance, is unlikely to work
>>> with that, since if it doesn't alter pid->level, things like task
>>> ns_of_pid won't work.
>>>
>>> Running the test scripts I wrote for my testing of that patch also
>>> doesn't seem to produce the expected result:
>>>
>>> after doing setns, the pid won't show up in that namespace.
>>
>> Yes, AFAIR, pid won't show up, you have to do fork-exec.
>
> Ah, so you mean the kid will show up... Well, ok.
>
> That's acceptable, but how about the behavior I am proposing ? (in the
> patch I sent as a reply to this thread).

Let me look at the patch closely.
>
> I believe it to be saner, even though there is a price tag attached to
> it. None of the other setns calls require you to do any such trickery...

Yeah, but the pidns is different from the other namespace, it is not
supposed to be unshared.
I remember we had a discussion about this and Eric had some good reasons
to do it this way. One of them of the pid cached by the glibc. Also, we
don't want to have our pid changing in our application.

You may find more informations in there [1]

Thanks
  -- Daniel

[1] http://thread.gmane.org/gmane.linux.network/153200