Subject: Re: [RFC][PATCH] Do not set /proc inode->pid for non-pid-related inodes
Posted by ebiederm on Thu, 22 Mar 2007 14:16:41 GMT
View Forum Message <> Reply to Message

Cedric Le Goater <clg@fr.ibm.com> writes:

>>> So I suggested to have a kthread be pid == 1 for each new pid namespace.
>>> the kthread can do the killing of all tasks if needed and will die when
>>> the refcount on the pid namespace == 0.
>>>
>>> Would such a (rough) design be acceptable for mainline ?
>>
>> The case that preserves existing semantics requires us to be able to
>> run /sbin/init in a container. Therefore pid 1 should be a user space
>> process.
>
> /sbin/init can't run without being pid == 1. hmm ? need to check. When we
> have more of the pid namespace, it should be easier.

Correct.

```
>From sysvinit src/init.c:main
>       /*
>        *      Is this telinit or init ?
>        */
>       isinit = (getpid() == 1);
>       for (f = 1; f < argc; f++) {
>               if (!strcmp(argv[f], "-i") || !strcmp(argv[f], "--init"))
>                       isinit = 1;
>                       break;
>       }
>       if (!isinit) exit(telinit(p, argc, argv));
>
```

Plus there are the additional signal handling semantics of pid == 1
where signals are received unless pid == 1 has set up a signal
handler. This especially includes SIGKILL.

>> So I don't think a design that doesn't allow us to run /sbin/init as
>> in a container would be acceptable for mainline.
>
> I agree that user space is assuming that /sbin/init has pid == 1 but don't
> you think that's a strong assumption ?
>
> on the kernel side we have is_init() so it shouldn't be an issue.

Basically there are some of the semantics of pid == 1 that only apply to
the /sbin/init in the initial pid namespace. This is what is_init is for.

There are other semantics that should apply to every process that has pid == 1, like dropping signals from other processes in it's pid namespace or children of it's pid namespace that it doesn't have a handler for.

Back to the main subject I still don't understand the idea of running a kernel daemon as pid == 1. What would that buy us?

Eric

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