Subject: Re: [PATCH 1/3] change clone_flags type to u64
Posted by Cedric Le Goater on Thu, 10 Apr 2008 13:18:20 GMT
View Forum Message <> Reply to Message

Andi Kleen wrote:
>> I guess that was a development rationale.
>
> But what rationale? It just doesn't make much sense to me.

Let's add Eric in Cc:

>> Most of the namespaces are in
>> use in the container projects like openvz, vserver and probably others
>> and we needed a way to activate the code.
>
> You could just have added it to feature groups over time.

Yes if the feature group had existed, that would have been a good
option.

Don't take me wrong. I agree with this group direction. Most
namespaces can't be safely decoupled from each other with a clone
flag.

>> Not perfect I agree.
>>
>>> With your current strategy are you sure that even 64bit will
>>> be enough in the end? For me it rather looks like you'll
>>> go through those quickly too as more and more of the kernel
>>> is namespaced.
>> well, we're reaching the end. I hope ! devpts is in progress and
>> mq is just waiting for a clone flag.
>
> Are you sure?

I'm never sure ! :) That's what we have in plan for the moment.

>>> Also I think the user interface is very unfriendly. How
>>> is a non kernel hacker supposed to make sense of these
>>> myriads of flags? You'll be creating another
>>> CreateProcess123_extra_args_extended()
>>> in the end I fear.
>> well, the clone interface is a not friendly interface anyway. glibc wraps
>> it
>
> But only for the stack setup which is just a minor detail.
>
> The basic clone() flags interface used to be pretty sane and usable

> before it could overloaded with so many tiny features.
>
> I especially worry on how user land should keep track of changing kernel
> here. If you add new feature flag for lots of kernel features it is
> reasonable to expect that in the future there will be often new features.
>
> Does this mean user land needs to be updated all the time? Will this
> end up like another udev?
>
>> We will need a user library, like we have a libphtread or a libaio, to
>
> That doesn't make sense. The basic kernel syscalls should be usable,
> not require some magic library that would likely need intimate
> knowledge of specific kernel versions to do any good.

No magic there. but running a container will require some userland code
to be set up properly.

>> but we still need a way to extend the clone flags because none are left.
>
> Can we just take out some again that were added in the .25 cycle and
> readd them once there is a properly thought out interface? That would
> leave at least one.

well, CLONE_STOPPED is being recycle in 2.6.26. so we could use that one
to group namespaces.

and CLONE_NEWPID would probably be a good candidate to group namespaces.

That would be fine for me but it would still leave clone with one to zero
flags left.

Thanks,

C.

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