Subject: Re: [PATCH] pidns: Place under CONFIG_EXPERIMENTAL (take 2)
Posted by Cedric Le Goater on Mon, 29 Oct 2007 18:05:12 GMT
View Forum Message <> Reply to Message

Eric W. Biederman wrote:
> Andrew Morton <akpm@linux-foundation.org> writes:
>
>>> On Sat, 27 Oct 2007 04:04:08 +0200 Adrian Bunk <bunk@kernel.org> wrote:
>>>> be happy to hear if someone has a better idea.
>>> There is a difference between "complete the feature" and "early adopters
>>> to start playing with the feature" on the one side, and making something
>>> available in a released kernel on the other side.
>>>
>>> For development and playing with it it can depend on BROKEN (perhaps
>>> with the dependency removed through the first -rc kernels), but as soon
>>> as it's available in a -final kernel the ABI is fixed.
>>>
>> Yes, if we're not 100% certain that the interfaces are corrent and unchanging
>> and that the implementation is solid, we should disable the feature at Kconfig
>> time.
>
> Reasonable. So far things look good for a single pid namespace. Multiple
> pid namespaces look iffy.
>
>> The best option would be to fix things asap. But assuming that option isn't
>> reasonable and/or safe, we can slip a `depends on BROKEN' into -rc6 then
>> resume development for 2.6.25.
>
> I think we can make a lot of progress but there is enough development
> yet to do to reach the target of correct and unchanging interfaces,
> with a solid interface. That unless we achieve a breakthrough I
> don't see us achieving that target for 2.6.24.
>
> The outstanding issues I can think of off the top of my head:
> - signal handling for init on secondary pid namespaces.
> - Properly setting si_pid on signals that cross namespaces.

these are being addressed by suka patches, and also you with the latest patch
you sent. right ?

> - The kthread API conversion so we don't get kernel threads
>   trapped in pid namespaces and make them unfreeable.

a lot of work has been done on that part. take a look at it. the clean up
is really impressive !

NFS still uses the kernel_thread() API. the first thing to do on the kthread
topic is to improve the kthread API.

I think we can discard the remaining drivers for the moment.

> - At fork time I think we are doing a little bit too much work
>   in setting the session and the pgrp, and removing the controlling
>   tty.

yes probably. this needs to be sorted out. it makes a container init
process different from the system init process.

> - AF_unix domain credential passing.

see commit b488893a390edfe027bae7a46e9af8083e740668 which is covering
UNIX socket credentials and more. Are you thinking we should do more for
credentials and use a struct pid* ? This looks scary.

> - misc pid vs vpid sorting out (autofs autofs4, coda, arch specific
>   syscalls, others?)

autofs* is fixed. netlink ?

> - Removal of task->pid, task->tgid, task->signal->__pgrp,
>   tsk->signal->__session or some other way to ensure that we have
>   touched and converted all of the kernel pid handling.

well, __pgrp and __session are pretty well covered with the __deprecated
attribute. I don't see what else we could to do on these. we can't remove
the task_{session,pgrp}_* routines.

we could apply the same __deprecated technique to task->pid, task->tgid.
This is going to be a challenge :)

> - flock pid handling.

Pavel again.

> It hurts me to even ponder what thinking makes it that
> CONFIG_EXPERIMENTAL isn't enough to keep a stable distro
> from shipping the code in their stable kernel, and locking us into
> trouble.
>
> With that said. I think I should just respin the patchset now and add
> the "depends on BROKEN".
>
> The user namespace appears to need that treatment as well.

The kernel will be protected by a CONFIG_NAMESPACES option as soon as it
gets in. Unfortunately, it didn't make 2.6.24 so this will be 2.6.25

material.

Cheers,

C.

> The network namespace has so little there and it already depends
> on !SYSFS so I don't think we are going to run into any trouble
> with it. Happily I managed to parse that problem differently,
> so I could slice of the parts of the networking stack that
> had not been converted.
>
> Eric
> -
> To unsubscribe from this list: send the line "unsubscribe linux-kernel" in
> the body of a message to majordomo@vger.kernel.org
> More majordomo info at http://vger.kernel.org/majordomo-info.html
> Please read the FAQ at http://www.tux.org/lkml/
>