Subject: Re: [PATCH] namespaces: introduce sys_hijack (v7)
Posted by serue on Fri, 02 Nov 2007 13:43:50 GMT
View Forum Message <> Reply to Message

Quoting Paul Menage (menage@google.com):
> On Oct 31, 2007 4:13 PM, Serge E. Hallyn <serue@us.ibm.com> wrote:
> >
> > Paul would like to be able to 'enter a cgroup', even if it is empty.
> > Hijack takes more than just the nsproxy from the hijacked task, so
> > this would result in different behavior between hijacking a populated
> > cgroup and an empty cgroup. So we might want to introduce a third
> > type of hijacking, so we have HIJACK_PID, HIJACK_CGROUP, and
> > HIJACK_EMPTY_CGROUP.
>
> Do we need all three distinctions? If there was a process in the
> cgroup, you could just use HIJACK_PID to hijack it. So HIJACK_CGROUP

No, then you might worry about pid wraparound. Sure it may seem silly
in most use cases, but if you're requesting the attach through some web
interface, it could happen.

> could just do what you're currently calling HIJACK_EMPTY_CGROUP.
>
> >
> > It also then acts like the nsproxy cgroup patchset I sent out months
> > ago for simply entering namespaces. In fact this would need to be
> > restricted to ns cgroups, and ns cgroups would need to grab a reference
> > to the nsproxy.
>
> Doesn't the nsproxy cgroup already grab an nsproxy reference?

No. That was done in my patches implementing namespace enter.
But that can easily be added.

> > So do we want to allow hijacking/entering an empty cgroup?
>
> In general, entering an emtpy cgroup is a perfectly fine thing to do -
> it's only the ns_proxy case where this is complicated, since some
> namespaces aren't safe against third-party changes to the task's
> ns_proxy.
>
> There really should be some way to enter such a set of namespaces, and
> doing it at fork time pretty much has to be safe since that's when
> nsproxy changes normally occur.

Well the main difference is that on fork, most namespaces give you a
copy of the original namespace, so open resources remain valid. Whereas
with hijack/enter, you get a "random" populated namespace.

Still I don't think it should be a problem since the resources generally
retain references to the namespaces in which they are valid.

> Being able to do it at other times
> (maybe only operating on current?) would be nice too.
>
> Paul

-serge