Subject: Re: [PATCH 11/11] sysfs: user namespaces: fix bug with clone(CLONE_NEWUSER) with fairsched
Posted by serue on Wed, 25 Jun 2008 18:44:09 GMT
View Forum Message <> Reply to Message

Quoting Tejun Heo (htejun@gmail.com):
> Benjamin Thery wrote:
> > Mark the /sys/kernel/uids directory to be tagged so that processes in
> > different user namespaces can remount /sys and see their own uid
> > listings.
> >
> > Without this patch, having CONFIG_FAIR_SCHED=y makes user namespaces
> > unusable, because when you
> > clone(CLONE_NEWUSER)
> > it will auto-create the root userid and try to create
> > /sys/kernel/uids/0. Since that already exists from the parent user
> > namespace, the create fails, and the clone misleadingly ends up
> > returning -ENOMEM.
> >
> > This patch fixes the issue by allowing each user namespace to remount
> > /sys, and having /sys filter the /sys/kernel/uid/ entries by user
> > namespace.
> >
> > Signed-off-by: Serge Hallyn <serue@us.ibm.com>
> > Signed-off-by: Benjamin Thery <benjamin.thery@bull.net>
>
> Ditto as patch #10.

Except the sysfs mount holds no refcount on the userns. So as long as we do the ida tagging as you suggested in your response to patch 6, there should be no reference to the user_ns left in sysfs code.

The extra reference in patch #9 is for a light ref on the network namespace. I'm still not sure that needs to be there, since if the network namespace goes away, it will properly unregister its sysfs mounts. Eric, Benjamin, I really don't see any use for the hold_net() from sysfs. What is it doing?

-serge