Subject: Re: [PATCH][RFC] Cleanup in namespaces unsharing
Posted by serue on Fri, 08 Jun 2007 14:07:58 GMT
View Forum Message <> Reply to Message

Quoting Pavel Emelianov (xemul@openvz.org):
> Cedric Le Goater wrote:
> > Pavel Emelianov wrote:
> >> Cedric Le Goater wrote:
> >>> Pavel Emelianov wrote:
>
> [snip]
>
> >>>> Did I miss something in the design or this patch worth merging?
> >>> I've sent a more brutal patch in the past removing CONFIG_IPC_NS
> >>> and CONFIG_UTS_NS. Might be a better idea ?
> >> In case namespaces do not produce performance loss - yes.
> >>
> >> By that patch I also wanted to note that we'd better make
> >> all the other namespaces check for flags themselves, not
> >> putting this in the generic code.
> >
> > yep. let's fix that in the coming ones if they have config option.
> >
> > a similar issue is the following check done in
> > unshare_nsproxy_namespaces() and copy_namespaces() :
> >
> > if (!capable(CAP_SYS_ADMIN))
> >   return -EPERM;
> >
> > it would be interesting to let the namespace handle the unshare
> > permissions. CAP_SYS_ADMIN shouldn't be required for all namespaces.
> > ipc is one example.
>
> Frankly, I think that some capability *is* required for
> cloning the namespaces.

We can
 1. start a long per-namespace discussion on which namespaces really
    need it
 2. add a new CAP_SYS_UNSHARE capability so at least we're not
    using CAP_SYS_ADMIN for this
 3. leave it as is

3 is really not that bad, though, since unshare activity can AFAICT
always be consolidated in small setuid helpers (or helpers with file
capabilities set :). Starting a vserver, starting a c-r job, and
unsharing mounts namespace on login using pam, can all be easily done
with privilege.

2 is unfortuntely a hassle since we have (last i looked) 1 free cap. Or
are we down to none?

I think had sent an email months ago starting a per-ns discussion on the
safety of not requiring a capability, but finding that coudl be a pain.
Off the bat, certain CLONE_NEWPID seems safe, right? CLONE_NEWNS could
be safe if we automatically made all the vfsmounts in the new ns slaves
of the original. CLONE_NEWNET would be pretty worthless since
presumably you'll always need CAP_NET_ADMIN to actually set up your
virtual net devices. CLONE_NEWIPC does seem safe. CLONE_NEWPTS should
be safe if we implement it the way Herbert suggested, with
/dev/pts/0 in a child ptsns showing up in /dev/pts/child_xyz/0 for the
parent.

thanks,
-serge
