Subject: Re: [DRAFT] Container mini-summit notes v0.01
Posted by ebiederm on Sat, 08 Sep 2007 20:26:08 GMT
View Forum Message <> Reply to Message

"Serge E. Hallyn" <serue@us.ibm.com> writes:

>> > then you should have taken CAP_SYS_MKNOD away from the container.
>>
>> no serge,
>>
>> we want the container to be able to mknod()
>
> Someone give me one good reason why this is needed.

The picture that I see is still fuzzy, so I cannot say exactly what for a device namespace needs to take. The practical issues is that we have virtual devices that when we migrate people will want to continue using. ptys are the common case here, but there are loop devices and other virtual devices.

Doing things like changing the major/minor numbers on a device we currently have open during migration could be painful.

For non-virtual devices we can treat it as a device hot plug event, because we really cannot continue with the device open. For the virtual devices we can do better and so it is quite likely that we want to.

This isn't an important issue until we get to the point of dealing with migration however.

>> >> Or mounts it from somewhere outside.
>> >
>> > and CAP_SYS_MOUNT
>>
>> and that also.
>
> Same here. Restricting containers to user mounts - which include
> a great deal of things including fuse loopback etc - should be fine.

The last I looked at user mounts they implied nosuid and nodev.

Which leads to an interesting implication. sys_mknod support in a container does not appear to be fundamental, while device namespaces so we can keep virtual devices at their same major/minor numbers looks fundamental.

> But again, if everyone but me agrees on this, we can try to focus on
> this instead of devpts this year. Cedric, was this mentioned at the
> kernel summit? Was there any reaction to this idea?

We didn't go into much technical detail a kernel summit. The goal
was to stick to topic that were of general interest to most of the
group. Which was mostly kernel process related. We did talk about
our high level objectives and the biggest question was when will the
container work be done? No real objections were answered.

So for technical details we still need to discuss them on the appropriate
mailing lists.

> This of course is also something that could be implemented pretty simply
> as a container subsys defining the security_mknod hook, with the
> whitelist defined through the task container interface.

Something to mention. I keep thinking for the isolation aspects of this
it may make sense to refactor the code behind the security hooks to
be a table based implementation like netfilter. Allowing code from
multiple parties to be used together instead of the current all or
nothing paradigm.

>> > Anyway if people really all agree on a per-container device whitelist,
>> > I won't object. Just seems like overkill to me.
>> >>> Whereas devpts you do need namespaces for.
>> >>> -serge

The practical question is what do we need to do to migrate applications
that are using virtual devices.

>> let's get back on the mailing list !

Back.

Eric

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