Subject: Re: containers and cgroups mini-summit @ Linux Plumbers
Posted by ebiederm on Thu, 26 Jul 2012 10:57:09 GMT
View Forum Message <> Reply to Message

Glauber Costa <glommer@parallels.com> writes:

> On 07/26/2012 07:57 AM, Eric W. Biederman wrote:
>> Glauber Costa <glommer@parallels.com> writes:
>>
>>> I just came up with the following preliminary list of sessions:
>>>
>>> http://wiki.linuxplumbersconf.org/2012:containers
>>>
>>> Since people mostly said what they wanted to talk about, but without
>>> extensive descriptions, I took the liberty of coming up with a small
>>> text for each in the blueprints. If you believe this is inaccurate, or
>>> would like to see it extended (although I personally don't see the point
>>> about going into very formal and deep details here), just let me know
>>> and I will edit it.
>>>
>>> This is all still subject to change.
>>
>> Something that just came up recently and worth looking at if it hasn't
>> already be resolved.
>>
>> The network namespace, the user namespace, and the memory control group
>> are not meshing well.
>>
>> In particular we need some additional checks for an unprivileged user
>> who can set tcp_mem. If you are the creator of a network namespace you
>> should at least be able to set the values down. I don't know at all
>> about increasing the amount of memory consumed by the tcp stack.
>
> This is between the user namespace and net namespace only, right ?
>
> To be quite honest, I haven't looked thoroughly at UNS after your last
> work. How do you yourself believe this should be?

I looked a little deeper and there are a few more places in the
networking stack besides tcp_mem, that are setting memory limits
that unprivileged users should not be able to touch.

What I would expect one of.
- A global limit accross network namespaces that the per netns limit
  can not allow you to escape.
- Something like rlimits where the limit can be reduced but not
  increased.
- capability checks that prevent anyone except the global root from

changing the value (of course this has the problem that creating
a fresh network namespace allows memory limit escaping).

The driving factor is that in the 3.7 time frame it will be possible to
create user namespaces and network namespaces as unprivileged users. So
we have to be careful that we setup the limits so that the global root
on the host can set limits that are not trivially overridden.

>> The non-nesting nature of memory control groups with respect to the
>> network stack also seems very bizarre.
>
> Correction:
>
> The non-nesting nature of memory control groups is very bizarre. No need
> for modifiers. It does support nesting, though. Just that it is
> selectable, and not the default. But there is work in progress to change
> that.

Which leads to another bit of fun. It is possible to create containers
in containers in containers, which has interesting implications for
control groups in general, and especially interesting implications for
control groups that don't nest.

Eric