Subject: Re: [PATCH 06/16] net: Add a network namespace parameter to struct sock
Posted by Daniel Lezcano on Fri, 21 Sep 2007 07:30:22 GMT
View Forum Message <> Reply to Message

Eric W. Biederman wrote:
> "Denis V. Lunev" <den@sw.ru> writes:
>
>> Daniel Lezcano wrote:
>>>> This place is a very tricky, indeed. If we keep the namespace until
>>>> timewait bucket death - we'll keep the namespace alive at least 5
>>>> _minutes_ after all process death.
>>> Yes, that's right. And for me that makes totally sense. The namespace
>>> should not be destroyed until it is referenced somewhere.
>> If all incoming interfaces are stopped, sure they do, no incoming
>> packets will be. So, it is completely pointless to keep TW bucket for 5
>> minutes. This is a resources wastage.
>
> Agreed, at least in principle.
>>>> If we stop a VE (in terms of OpenVz) and restart it, we'll 100% have an
>>>> _OLD_ namespace with all buckets shown :(So, in OpenVz we use a number
>>>> of VE instead of pointer to a VE. Additionally, on VE death we can wipe
>>>> all TW buckets. VE start stop from outside world looks very much like a
>>>> computer power on/off.
>>> That makes sense too. But if you wipe out the sockets when stopping the
>>> VE where is the problem with the restart ?
>>>
>>>
>> classical egg/chicken problem. If TW bucket holds namespace, how to
>> decide when to destroy it? :(
>
> TW bucket must have a reference to a namespace because otherwise
> we cannot interpret them.
>
> However if need be we can just do hold_net, release_net style reference
> counting, if we know that when the namespace exits we will flush all
> of those sockets.
>
> I looked and it doesn't appear that I am actually initializing
> this field in my current patchset. :(
> - So either my skim through my code is wrong.
> - Something got dropped in keeping the patches up to date.
> - This was never addressed :(
> I would be a good idea to see if we can make certain that we are
> initializing the field right now (at least to &init_net). That
> way we won't get into a subtle problem later when we try and use it.

With Denis's remark I looked at the code and I noticed that too.

I am currently doing some testing to check that. I will provide a patchset to hold a network namespace reference for the timewait socket and to wipe out timewait socket for the network namespace in a few hours.

BTW, the orphan sockets will lead to a similar problem ...

-- Daniel