Subject: Re: [PATCH 10/12] L2 network namespace: playing with pass-through device
Posted by Mishin Dmitry on Thu, 14 Dec 2006 11:17:44 GMT
View Forum Message <> Reply to Message

On Thursday 14 December 2006 13:48, Daniel Lezcano wrote:
> Dmitry Mishin wrote:
> > On Wednesday 13 December 2006 23:27, Daniel Lezcano wrote:
> >> Vlad Yasevich wrote:
> >>> Hi Daniel
> >>>
> >>> Daniel Lezcano wrote:
> >>>> Herbert Poetzl wrote:
> >>>>> On Tue, Dec 12, 2006 at 04:50:50PM +0100, Daniel Lezcano wrote:
> >>>>>> Dmitry Mishin wrote:
> >>>>>>> On Tuesday 12 December 2006 17:19, Daniel Lezcano wrote:
> >>>>>>>> Dmitry Mishin wrote:
> >>>>>>>>
> >>>>>>>>>> Why do yo need to have a child list and sibling list ?
> >>>>>>>>> Because of the level2<->level3 hierarchy, for example.
> >>>>>>>> This hierarchy doesn't exist with ns->parent ? Do you have an example
> >>>>>>>> when the hierarchy should be used ? I mean when we need to browse from
> >>>>>>>> l2 -> l3 ?
> >>>>>>> For example, to check that new ifaddr is already used by child l3 namespace.
> >>>>>>> The devinet isolation does already do that, you can not add a new ifaddr
> >>>>>>> if it already exists. Do you have another example ?
> >>>>>>> Could devinet isolation provide ifaddrs list with namespaces?
> >>>>>>> What will be with child namespaces if you decide to destroy parent namespace?
> >>>>>>> If we decide to destroy them, than how we could get their list?
> >>>>>>> It is a question of flexibility and easy management.
> >>>>>>> Why do you want to remove this code?
> >>>>>> I don't want to especially remove this code, I just want to understand
> >>>>>> what it does and why. If it appears to be useless, let's remove it, if
> >>>>>> it appears to be useful, let's keep it.
> >>>>>>
> >>>>>> By the way, what is the meaning on destroying the namespaces directly,
> >>>>>> is it not the kref mechanism which needs to do that ? For example, if
> >>>>>> you create a l2 namespace and after you create l3 namespaces. You want
> >>>>>> to destroy the l2 namespace, the l2 namespace should stay "zombie" until
> >>>>>> all the l3 namespaces exit. If you need to wipe out all the namespaces,
> >>>>>> you should destroy all the related namespaces' ressources, like killing
> >>>>>> all processes inside it. The namespaces will "put" their respective kref
> >>>>>> and will trigger the freeing of the ressources.
> >>>>> networking (mostly sockets) will probably require
> >>>>> some mechanism to 'zap' them, ignoring the defined
> >>>>> timeouts. otherwise the spaces could hang around
> >>>>> for quite a while waiting for some response, which
> >>>>> might never come ...

> >>>> Yes, exact. We will need a specific socket cleanup by namespace in order
> >>>> to do network migration. This is the only case I see to 'zap' the sockets.
> >>>> The sockets should never be flushed in other cases. For example, you
> >>>> launch an application into a network namespace, it sends 10MB to a peer
> >>>> and exits. The network namespace should stay "alive" until all orphans
> >>>> sockets have flushed their buffers to the peer. This behavior is
> >>>> perfectly handled by the kref mechanism because sock_release will "put"
> >>>> the network namespace and that will trigger the network namespace
> >>>> destruction.
> >>>>
> >>> Are you saying that you can't see the reason to be able to shutdown/destroy a
> >>> given container. What if it's misbehaving or has been compromised???
> >>>
> >>> I would think an administrator, should be able to shutdown/destroy a
> >>> given container or namespace from above or outside of such container/namespace
> >>> if it's warranted. If this case, if we destroy an L2 namespace, L3 children
> >>> should probably be cleaned up as well.
> >> Yes, I agree, you should want to destroy a specific container. IMHO,
> >> freeing it directly is not the right way, you should destroy the
> >> namespaces' ressources and the namespace will be released. For example,
> >> you create a l3 container and into it you spawn 10 processes and each of
> >> them creates 10 connections, roughly you have 10 + 100 reference to the
> >> namespace. You want to destroy this container for any reason, you kill
> >> the 10 processes and you destroy the 100 connections => the ref count
> >> reach zero and the namespace is released.
> >> In the case of the l2 namespace having l3 childs, I think it is up to
> >> the administrator to know what he does. In other words, he should kill
> >> all l3 namespaces and after kill the l2 parent.
> > So, you suppose, that administrator remembers which l3 namespaces are over
> > this l2 namespace.
>
> I guess it remembers which IP address has been assigned to the
> namespaces, what name he gave to them, how many they are, what are the
> virtual hostname, etc ... The administrator can be a human or a
> container system manager.
>
> > List of childs gives an ability to track such relations easily.
> > If administrator decides to kill l2, he could decide is it force kill, when all
> > underlied l3 namespaces and their processes should be killed, or soft kill,
> > where only l2 refcount decremented.
>
> If you want the list of childs to be used to browse the namespace
> hierarchy in order to destroy them, what about the pid namespace,
> ipc_namespace, utsname namespace, ... ? You should have the same list
> for each namespace's ressources, no ? Perhaps having a child list into
> the network namespace is not the right place ...
L3 namespace is dependant on L2 resources. This is not a true for above
namespaces, which are independent. That's why hierarchy is required.

>
> >> Imagine you have a shell and you are into the foo directory, in another
> >> shell you remove the foo directory, the shell in the foo directory will
> >> not be pushed outside of the directory. The foo dir will stay in a
> >> "transient" state until the shell exits or change directory, I think
> >> that should be same with the namespaces or better to say "containers".
> > It is unacceptable in some cases. As Eric already said, network connections
> > could take refcount for a long time and you will wait container stop indefinitely,
> > because somebody spoofs your namespace. It is necessary to have an ability of
> > force kill.
>
> Don't me wrong, I don't say we should not have a force kill for a
> container, I am just saying the ressources must be killed for the
> container and that will refdown the ressources to reach the namespaces
> destruction.
The easiest way to kill such resources is to know which namespace resources
should be killed.

You propose to do the same from userspace. Userspace tools should know, which
namespaces are underlied, which resources are used by them, etc. I vote, that
easier and more nature to do this from kernel space. Just because there is no
need to design different user-space handles for force flush different resources
instead of one to flush them all.

--
Thanks,
Dmitry.