Subject: Re: [RFC] network namespaces
Posted by Herbert Poetzl on Fri, 08 Sep 2006 06:02:00 GMT
View Forum Message <> Reply to Message

On Thu, Sep 07, 2006 at 12:29:21PM -0600, Eric W. Biederman wrote:
> Daniel Lezcano <dlezcano@fr.ibm.com> writes:
> >
> > IHMO, I think there is one reason. The unsharing mechanism is
> > not only for containers, its aim other kind of isolation like a
> > "bsdjail" for example. The unshare syscall is flexible, shall the
> > network unsharing be one-block solution ? For example, we want to
> > launch an application using TCP/IP and we want to have
> > an IP address only used by the application, nothing more.
> > With a layer 2, we must after unsharing:
> > 1) create a virtual device into the application namespace
> > 2) assign an IP address
> > 3) create a virtual device pass-through in the root namespace
> > 4) set the virtual device IP
> >
> > All this stuff, need a lot of administration (check mac addresses
> > conflicts, check interface names collision in root namespace, ...)
> > for a simple network isolation.
>
> Yes, and even more it is hard to show that it will perform as well.
> Although by dropping CAP_NET_ADMIN the actual runtime administration
> is about the same.
>
> > With a layer 3:
> > 1) assign an IP address
> >
> > In the other hand, a layer 3 isolation is not sufficient to reach
> > the level of isolation/virtualization needed for the system
> > containers.
>
> Agreed.
>
> > Very soon, I will commit more info at:
> >
> > http://wiki.openvz.org/Containers/Networking
> >
> > So the consensus is based on the fact that there is a lot of common
> > code for the layer 2 and layer 3 isolation/virtualization and we can
> > find a way to merge the 2 implementation in order to have a flexible
> > network virtualization/isolation.
>
> NACK In a real level 3 implementation there is very little common
> code with a layer 2 implementation. You don't need to muck with the
> socket handling code as you are not allowed to dup addresses between

> containers. Look at what Serge did that is layer 3.
>
> A layer 3 isolation implementation should either be a new security
> module or a new form of iptables. The problem with using the lsm is
> that it seems to be an all or nothing mechanism so is a very coarse
> grained tool for this job.

IMHO LSM was never an option for that, because it is
a) very complicated to use it for that purpose
b) missing many hooks you definitely need to make this work
c) is not really efficient and/or performant

with something 'like' iptables, this could be done, but
I'm not sure that is the best approach either ...

best,
Herbert

> A layer 2 implementation (where you have network devices isolated and
> not sockets) should be a namespace.
>
> Eric
> ____
> Containers mailing list
> Containers@lists.osdl.org
> https://lists.osdl.org/mailman/listinfo/containers

____
Containers mailing list
Containers@lists.osdl.org
https://lists.osdl.org/mailman/listinfo/containers