
Subject: Re: Network virtualization/isolation

Posted by [Herbert Poetzl](#) on Tue, 28 Nov 2006 17:37:19 GMT

[View Forum Message](#) <> [Reply to Message](#)

On Tue, Nov 28, 2006 at 09:51:57AM -0700, Eric W. Biederman wrote:

- >
- > I do not want to get into a big debate on the merits of various
- > techniques at this time. We seem to be in basic agreement
- > about what we are talking about.
- >
- > There is one thing I think we can all agree upon.
- > - Everything except isolation at the network device/L2 layer, does not
- > allow guests to have the full power of the linux networking stack.
- >
- > - There has been a demonstrated use for the full power of the linux
- > networking stack in containers..
-
- There has been a demonstrated use for the full performance
- IP layer isolation too, both in BSD and Linux for several
- years now ...
-
- > - There are a set of techniques which look as though they will give
- > us full speed when we do isolation of the network stack at the
- > network device/L2 layer.
- >
- > Is there any reason why we don't want to implement network namespaces
- > without the full power of the linux network stack?

duplicate negation ->

"Is there any reason why we want to implement network namespaces
with the full power of the linux network stack?"

yes, I think you have some reasons for doing so, especially
the migration part seems to depend on it

OTOH, we also want IP isolation, as it allows to separate
services (and even handle overlapping sets) in a very natural
(linux) way, without adding interfaces and virtual switches
and bridges at a potentially high overhead just to do simple
layer 3 isolation

- > If there is a case where we clearly don't want the full power of the
- > linux network stack in a guest but we still need a namespace we can
- > start looking at the merits of the alternatives.

see above, of course, all cases can be 'simulated' by a
fully blown layer 2 virtualization, so that's not an argument

but OTOH, all this can also be achieved with Xen, so we could as well bring the argument, why have network namespaces at all, if you can get the same functionality (including the migration) with a Xen domU ...

> > What is this new paradigm you are talking about ?

>

> The basic point is this. The less like stock linux the inside of a
> container looks, and the more of a special case it is the more
> confusing it is. The classic example is that for a system container
> routing packets between containers over the loopback interface is
> completely unexpected.

I disagree here, from the point of isolation that would be the same as saying:

"having a chroot(), it is completely unexpected that the files reside on the same filesystem and even will be cached in the same inode cache"

the thing is, once you depart from the 'container' = 'box' idea, and accept that certain resources are shared (btw, one of the major benefits of 'containers' over things like Xen or UML) you can easily accept that:

- host local traffic uses loopback
- non local traffic uses the appropriate interfaces
- guests are local on the host, so
- guest - guest and guest - host traffic is local
an therefore will be more performant than remote traffic (unless you add various virtual switches and bridges and stacks to the pathes)

> > There is not extra networking data structure instantiation in the
> > Daniel's L3.

> Nope just an extra field which serves the same purpose.

>

> >> - Bind/Connect/Accept filtering. There are so few places in
> >> the code this is easy to maintain without sharing code with
> >> everyone else.

> >

> > For isolation too ? Can we build network migration on top of that ?

>

> As long as you can take your globally visible network address
> with you when you migrate you can build network migration on
> top of it. So yes bind/accept filtering is sufficient to
> implement migration, if you are only using IP based protocols.

correct, don't get me wrong, I'm absolutely not against layer 2 virtualization, but not at the expense of light-weight layer 3 isolation, which is the traditional way 'containers' are built (see BSD, solaris ...)

HTC,
Herbert

> 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>
