Subject: Re: Network virtualization/isolation
Posted by Herbert Poetzl on Sun, 03 Dec 2006 16:37:19 GMT
View Forum Message <> Reply to Message

On Sun, Dec 03, 2006 at 07:26:02AM -0500, jamal wrote:
> On Wed, 2006-14-11 at 16:17 +0100, Daniel Lezcano wrote:
> > The attached document describes the network isolation at the layer 2
> > and at the layer 3 ..
>
> Daniel,
>
> I apologize for taking this long to get back to you. The document (I
> hope) made it clear to me at least the difference between the two
> approaches. So thanks for taking the time to put it together.
>
> So here are my thoughts ...
> I havent read the rest of the thread so i may be repeating some of the
> discussion; i have time today, I will try to catchup with the
> discussion.
>
> * i think the L2 approach is the more complete of the two approaches:
>
> It caters to more applications: eg i can have network elements such as
> virtual bridges and routers. It doesnt seem like i can do that with the
> L3 approach. I think this in itself is a powerful enough reason to
> disqualify the L3 approach.
>
> Leading from the above, I dont have to make _a single line of code
> change_ to any of the network element management tools inside the
> container. i.e i can just run quagga and OSPF and BGP will work as is or
> the bridge daemon and STP will work as is or tc to control "real"
> devices or ip to control "real" ip addresses. Virtual routers and
> bridges are real world applications (if you want more info ask me or ask
> google, she knows).
>
> **** This wasnt clear to me from the doc on the L3 side of things, so
> please correct me:
> because of the pid virtualization in the L2 approach(openvz?) I can run
> all applications as is. They just dont know they are running on a
> virtual environment. To use an extreme example: if i picked apache as a
> binary compiled 10 years ago, it will run on the L2 approach but not on
> the L3 approach. Is this understanding correct? I find it hard to
> believe that the L3 approach wouldnt work this way - it may be just my
> reading into the doc.

the 10 year old apache will run with layer 3 isolation
as well as with layer 2 virtualization (probably a little
faster though, we do not know yet :), because what it

does is IP (layer 3) traffic ...

> So lets say the approach taken is that of L2 (I am biased towards this
> because i want to be able to do virtual bridges and routers).
> The disadvantage of the L2 approach (or is it just the openvz?)
> approach is:
>
> - it is clear theres a lot more code needed to allow for the two level
> multiplexing every where. i.e first you mux to select the namespace then
> you do other things like find a pid, netdevice, ip address etc. I am
> also not sure how complete that code is; you clearly get everything
> attached to netdevices for free (eg networkc scheduler block) - which is
> nice in itself; but you may have to do the muxing code for other blocks.
> If my understanding is correct everything in the net subsystem has this
> mux levels already in place (at least with openvz)? I think each
> subsystem may have its own merit discussed (eg the L3 tables with the
> recent changes from Patrick allow up to 2^32 -1 tables, so a muxing
> layer at L3 maybe unnecessary)

> ---> To me this 2 level muxing looks like a clean design in that there
> is consistency (i.e no hack thats specific to just one sub-subsystem but
> not others). With this approach one could imagine hardware support that
> does the first level of muxing (selecting a namespace for you). This is
> clearly already happening with NICs supporting several unicast MAC
> addresses.

> I think the litmus test for this approach is the answer to the question:
> If i compiled in the containers in and do not use the namespaces, how
> much more overhead is there for the host path? I would hope that it is
> as close to 0 as possible. It should certainly be 0 if i dont compile in
> containers.

IMHO there are three cases to consider, to get valid
'performance' numbers:

 - host system with and without containers enabled
 - single guest (container) compared to host system _without_
 - bunch of guests (e.g. 10) compared to 10 apps/threads on host

one proven feature of the L3 isolation is that those
all end up with the same or even better performance

> - The desire for many MAC addresses. I dont think this is a killer
> issue. NICs are begining to show up which capabilities for many unicast
> MACs; many current have multicast hardware tables that can be used for
> stashing unicast MAC addresses; it has also been shown you can use
> multicast MAC addresses and get away with it if there is no conflict
> (protocols such as VRRP, CARP etc do this).

>
> - Manageability from the host side. It seems to be more complex with the
> L2 than with L3. But so what? These tools are written from scratch and
> there is no "backward compatibility" baggage.

well, no, actually the 'tools' to manage layer 3 isolation
are already there, and except for the 'setup' there is
nothing special to configure, as networking still lives
on the host

> Ok, I am out of coffee for the last 10 minutes;-> But above sit my views
> worth about $0.02 Canadian (which is about $0.02 US these days).
>
> I will try later to catch up with the discussion that started from
> Daniels original posting.

I would be interested in a config layout for a typical
L3 isolation setup when you 'only' have L2 virtualization

- typical host system with apache, mysql, postfix, ssh
  and ftp is broken down into security contexts to
  allow for increased security
- as part of that process, the services are isolated,
  while apache and ftp share the same ip [ip0], mysql
  will be using a local one [ip1], and postfix/ssh a
  second public one [ip2]

the L3 isolation approach is straight forward:

- assign the two public ips to eth0, the local one
  to lo or dummy0
- create five isolation areas where 0 and 1 share ip0,
  2 uses ip1 and 3,4 uses ip2

that's it, all will work as expected ... let's see with
what L2 isolation example you come up with, which is
able to 'mimic' this setup ...

note: no question it is possible to do that with L2

best,
Herbert

> cheers,
> jamal
>
> _____
> 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