Subject: Re: [RFC] network namespaces
Posted by ebiederm on Tue, 05 Sep 2006 14:45:39 GMT
View Forum Message <> Reply to Message

Daniel Lezcano <dlezcano@fr.ibm.com> writes:

>>>2. People expressed concerns that complete separation of namespaces
>>>   may introduce an undesired overhead in certain usage scenarios.
>>>   The overhead comes from packets traversing input path, then output path,
>>>   then input path again in the destination namespace if root namespace
>>>   acts as a router.
>
> Yes, performance is probably one issue.
>
> My concerns was for layer 2 / layer 3 virtualization. I agree a layer 2
> isolation/virtualization is the best for the "system container".
> But there is another family of container called "application container", it is
> not a system which is run inside a container but only the application. If you
> want to run a oracle database inside a container, you can run it inside an
> application container without launching <init> and all the services.
>
> This family of containers are used too for HPC (high performance computing) and
> for distributed checkpoint/restart. The cluster runs hundred of jobs, spawning
> them on different hosts inside an application container. Usually the jobs
> communicates with broadcast and multicast.
> Application containers does not care of having different MAC address and rely on
> a layer 3 approach.
>
> Are application containers comfortable with a layer 2 virtualization ? I don't
> think so, because several jobs running inside the same host communicate via
> broadcast/multicast between them and between other jobs running on different
> hosts. The IP consumption is a problem too: 1 container == 2 IP (one for the
> root namespace/ one for the container), multiplicated with the number of
> jobs. Furthermore, lot of jobs == lot of virtual devices.
>
> However, after a discussion with Kirill at the OLS, it appears we can merge the
> layer 2 and 3 approaches if the level of network virtualization is tunable and
> we can choose layer 2 or layer 3 when doing the "unshare". The determination of
> the namespace for the incoming traffic can be done with an specific iptable
> module as a first step. While looking at the network namespace patches, it
> appears that the TCP/UDP part is **very** similar at what is needed for a layer
> 3 approach.
>
> Any thoughts ?

For HPC if you are interested in migration you need a separate IP per
container. If you can take you IP address with you migration of
networking state is simple. If you can't take your IP address with

you a network container is nearly pointless from a migration perspective.

Beyond that from everything I have seen layer 2 is just much cleaner than any layer 3 approach short of Serge's bind filtering.

Beyond that I have yet to see a clean semantics for anything resembling your layer 2 layer 3 hybrid approach. If we can't have clear semantics it is by definition impossible to implement correctly because no one understands what it is supposed to do.

Note. A true layer 3 approach has no impact on TCP/UDP filtering because it filters at bind time not at packet reception time. Once you start inspecting packets I don't see what the gain is from not going all of the way to layer 2.

Eric