Subject: Re: Network virtualization/isolation
Posted by jamal on Mon, 04 Dec 2006 13:22:37 GMT
View Forum Message <> Reply to Message

Daniel,

On Mon, 2006-04-12 at 11:18 +0100, Daniel Lezcano wrote:
> Hi Jamal,

> Currently, there are some resources moved to a namespace relative
> access, the IPC and the utsname and this is into the 2.6.19 kernel.
> The work on the pid namespace is still in progress.
>
> The idea is to use a "clone" approach relying on the "unshare_ns"
> syscall. The syscall is called with a set of flags for pids, ipcs,
> utsname, network ... You can then "unshare" only the network and have an
> application into its own network environment.
>

Ok, so i take it this call is used by the setup manager on the host
side?

> For a l3 approach, like a l2, you can run an apache server into a
> unshared network environment. Better, you can run several apaches server
> into several network namespaces without modifying the server's network
> configuration.
>

ok - as i understand it now, this will be the case for all the
approaches taken?

> Some of us, consider l2 as perfectly adapted for some kind of containers
>   like system containers and some kind of application containers running
> big servers, but find the l2 too much (seems to be a hammer to crush a
> beetle) for simple network requirements like for network migration,
> jails or containers which does not take care of such virtualization. For
> example, you want to create thousands of containers for a cluster of HPC
> jobs and just to have migration for these jobs. Does it make sense to
> have l2 approach ?
>

Perhaps not for the specific app you mentioned above.
But it makes sense for what i described as virtual routers/bridges.
I would say that the solution has to cater for a variety of
applications, no?

> Dmitry Mishin and I, we thought about a l2/l3 solution and we thing we

> found a solution to have the 2 at runtime. Roughly, it is a l3 based on
> bind filtering and socket isolation, very similar to what vserver
> provides. I did a prototype, and it works well for IPV4/unicast.
>

ok - so you guys seem to be reaching at least some consensus then.

> So, considering, we have a l2 isolation/virtualization, and having a l3
> relying on the l2 network isolation resources subset. Is it an
> acceptable solution ?

As long as you can be generic enough so that a wide array of apps can be met, it should be fine. For a test app, consider the virtual bridges/routers i mentioned.
The other requirement i would see is that apps that would run on a host would run unchanged. The migration of containers you folks seem to be having under control - my only input into that thought since it is early enough, you may want to build your structuring in such a way that this is easy to do.

cheers,
jamal