Subject: Re: [patch 2/6] [Network namespace] Network device sharing by view
Posted by Daniel Lezcano on Tue, 27 Jun 2006 11:21:02 GMT
View Forum Message <> Reply to Message

>>>My point is that if you make namespace tagging at routing time, and
>>>your packets are being routed only once, you lose the ability
>>>to have separate routing tables in each namespace.
>>
>>Right. What is the advantage of having separate the routing tables ?
>
>
> Routing is everything.
> For example, I want namespaces to have their private tunnel devices.
> It means that namespaces should be allowed have private routes of local type,
> private default routes, and so on...
>

Ok, we are talking about the same things. We do it only in a different way:

 * separate routing table :
   namespace
   |
   \--- route_tables
    |
    \---routes

 * tagged routing table :
  route_tables
   |
   \---routes
    |
    \---namespace

When using routes private to the namespace, globally the logic of the ip
stack is not changed, it manipulates only differents variables. It is
more clean than tagging the route for the reasons mentioned by Eric.

When using route tagging, the logic is changed because when doing lookup
on the routes table which is global, the namespace is used to match the
route and make it visible.

I use the second method, because I think it is more effecient and reduce
the overhead. But the isolation is minimalist and only aims to avoid the
application using ressources outside of the container (aka namespace)
without taking care of the system. For example, I didn't take care of
network devices, because as far as see I can't imagine an administrator
wanting to change the network device name while there are hundred of
containers running. Concerning tunnel devices for example, they should

be created inside the container.

I think, private network ressources method is more elegant and involves more network ressources, but there is probably a significant overhead and some difficulties to have __lightweight__ container (aka application container), make nfs working well, etc... I did some tests with tbench and the loopback with the private namespace and there is roughly an overhead of 4 % without the isolation since with the tagging method there is 1 % with the isolation.

The network namespace aims the isolation for now, but the container based on the namespaces will probably need checkpoint/restart and migration ability. The migration is needed not only for servers but for HPC jobs too.

So I don't know what level of isolation/virtualization is really needed by users, what should be acceptable (strong isolation and overhead / weak isolation and efficiency). I don't know if people wanting strong isolation will not prefer Xen (cleary with much more overhead than your patches ;))

Regards
-- Daniel