Subject: Re: [patch 2/6] [Network namespace] Network device sharing by view
Posted by ebiederm on Mon, 26 Jun 2006 15:21:19 GMT
View Forum Message <> Reply to Message

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

> Andrey Savochkin wrote:
>> Hi Daniel,
>
> Hi Andrey,
>
>> It's good that you kicked off network namespace discussion.
>> Although I wish you'd Cc'ed someone at OpenVZ so I could notice it earlier :).
>
> devel@openvz.org ?
>
>> When a device presents an skb to the protocol layer, it needs to know to which
>> namespace this skb belongs.
>> Otherwise you would never get rid of problems with bind: what to do if device
>> eth1 is visible in namespace1, namespace2, and root namespace, and each
>> namespace has a socket bound to 0.0.0.0:80?
>
> Exact. But, the idea was to retrieve the namespace from the routes.

The problem is that if you start at the routes you have to do things
at layer 3 and you can't do anything at layer 2.  (i.e. You can't use DHCP).
You loose a whole lot of flexibility and power when you make it a
layer 3 only mechanism.

> IMHO, I think there are roughly 2 network isolation implementation:
>
>  - make all network ressources private to the namespace
>
>  - keep a "flat" model where network ressources have a new identifier
> which is the network namespace pointer. The idea is to move only some network
> informations private to the namespace (eg port range, stats, ...)

The problem is that you have to add a lot of new logic which is very
hard to get right and has some really weird corner cases that are very
hard to understand.

- That makes the patches hard to review.
- It makes it hard for the implementors to get it right.
- It means that there will be corner cases that the users don't
  understand.
- It is less flexible/powerful in what you can express?

I've been down that route it sucks.  Anything more than the simple

filter at bind time is asking for real trouble until you do the whole
thing.

Eric