Subject: Re: [RFC] network namespaces
Posted by Daniel Lezcano on Thu, 07 Sep 2006 08:25:56 GMT

View Forum Message <> Reply to Message

Caitlin Bestler wrote:
> ebiederm@xmission.com wrote:
>
>
>>>Finally, as I understand both network isolation and network
>>>virtualization (both level2 and level3) can happily co-exist. We do
>>>have several filesystems in kernel. Let's have several network
>>>virtualization approaches, and let a user choose. Is that makes
>>>sense?
>>
>>If there are not compelling arguments for using both ways of
>>doing it is silly to merge both, as it is more maintenance overhead.
>>
>
>
> My reading is that full virtualization (Xen, etc.) calls for
> implementing
> L2 switching between the partitions and the physical NIC(s).
>
> The tradeoffs between L2 and L3 switching are indeed complex, but
> there are two implications of doing L2 switching between partitions:
>
> 1) Do we really want to ask device drivers to support L2 switching for
>   partitions and something *different* for containers?
>
> 2) Do we really want any single packet to traverse an L2 switch (for
>   the partition-style virtualization layer) and then an L3 switch
>   (for the container-style layer)?
>
> The full virtualization solution calls for virtual NICs with distinct
> MAC addresses. Is there any reason why this same solution cannot work
> for containers (just creating more than one VNIC for the partition,
> and then assigning each VNIC to a container?)

IHMO, I think there is one reason. The unsharing mechanism is not only
for containers, its aim other kind of isolation like a "bsdjail" for
example. The unshare syscall is flexible, shall the network unsharing be
one-block solution ? For example, we want to launch an application using
TCP/IP and we want to have an IP address only used by the application,
nothing more.
With a layer 2, we must after unsharing:
  1) create a virtual device into the application namespace
  2) assign an IP address
  3) create a virtual device pass-through in the root namespace

4) set the virtual device IP

All this stuff, need a lot of administration (check mac addresses conflicts, check interface names collision in root namespace, ...) for a simple network isolation.

With a layer 3:
1) assign an IP address

In the other hand, a layer 3 isolation is not sufficient to reach the level of isolation/virtualization needed for the system containers.

Very soon, I will commit more info at:

http://wiki.openvz.org/Containers/Networking

So the consensus is based on the fact that there is a lot of common code for the layer 2 and layer 3 isolation/virtualization and we can find a way to merge the 2 implementation in order to have a flexible network virtualization/isolation.

-- Regards

Daniel.