Subject: Re: strict isolation of net interfaces
Posted by ebiederm on Fri, 30 Jun 2006 18:09:44 GMT
View Forum Message <> Reply to Message

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

> Serge E. Hallyn wrote:
>> Quoting Cedric Le Goater (clg@fr.ibm.com):
>>
>>>we could work on virtualizing the net interfaces in the host, map them to
>>>eth0 or something in the guest and let the guest handle upper network layers ?
>>>
>>>lo0 would just be exposed relying on skbuff tagging to discriminate traffic
>>>between guests.
>> This seems to me the preferable way. We create a full virtual net
>> device for each new container, and fully virtualize the device
>> namespace.
>

Answers with respect to how I see layer 2 isolation,
with network devices and sockets as well as the associated routing
information given per namespace.

> I have a few questions about all the network isolation stuff:
>
>   * What level of isolation is wanted for the network ? network devices ?
> IPv4/IPv6 ? TCP/UDP ?
>
>   * How is handled the incoming packets from the network ? I mean what will be
> mecanism to dispatch the packet to the right virtual device ?

Wrong question. A better question is to ask how do you know which namespace
a packet is in.
Answer: By looking at which device or socket it just came from.

How do you get a packet into a non-default namespace?
Either you move a real network interface into that namespace.
Or you use a tunnel device that shows up as two network interfaces in
two different namespaces.

Then you route, or bridge packets between the two. Trivial.

>   * How to handle the SO_BINDTODEVICE socket option ?

Just like we do now.

>   * Has the virtual device a different MAC address ?

All network devices are abstractions of the hardware so they are all
sort of virtual. My implementation of a tunnel device has a mac
address so I can use it with ethernet bridging but that isn't a hard
requirement. And yes the mac address is different because you can't
do layer 2 switching if everyone has the same mac address.

But there is no special ``virtual'' device.

> How to manage it with the real MAC address on the system ?
Manage?

> How to manage ARP, ICMP, multicasting and IP ?

Like you always do. It would be a terrible implementation if
we had to change that logic. There is a little bit of that
where we need to detect which network namespace we are going to because
the answers can differ but that is pretty straight forward.

> It seems for me, IMHO that will require a lot of translation and browsing
> table. It will probably add a very significant overhead.

Then look at:
 git://git.kernel.org/pub/scm/linux/kernel/git/ebiederm/linux -2.6-ns.git#proof-of-concept
or the OpenVZ implementation.

It isn't serious overhead.

>   * How to handle NFS access mounted outside of the container ?

The socket should remember it's network namespace.
It works fine.

>   * How to handle ICMP_REDIRECT ?

Just like we always do?

Eric
