Subject: Re: [RFC] network namespaces
Posted by ebiederm on Tue, 05 Sep 2006 18:27:20 GMT
View Forum Message <> Reply to Message

Herbert Poetzl <herbert@13thfloor.at> writes:

> On Tue, Sep 05, 2006 at 08:45:39AM -0600, Eric W. Biederman wrote:
>> Daniel Lezcano <dlezcano@fr.ibm.com> writes:
>>
>> For HPC if you are interested in migration you need a separate IP
>> per container. If you can take you IP address with you migration of
>> networking state is simple. If you can't take your IP address with you
>> a network container is nearly pointless from a migration perspective.
>>
>> Beyond that from everything I have seen layer 2 is just much cleaner
>> than any layer 3 approach short of Serge's bind filtering.
>
> well, the 'ip subset' approach Linux-VServer and
> other Jail solutions use is very clean, it just does
> not match your expectations of a virtual interface
> (as there is none) and it does not cope well with
> all kinds of per context 'requirements', which IMHO
> do not really exist on the application layer (only
> on the whole system layer)

I probably expressed that wrong. There are currently three
basic approaches under discussion.
Layer 3 (Basically bind filtering) nothing at the packet level.
  The approach taken by Serge's version of bsdjails and Vserver.

Layer 2.5 What Daniel proposed.

Layer 2. (Trivially mapping each packet to a different interface)
        And then treating everything as multiple instances of the
        network stack.
      Roughly what OpenVZ and I have implemented.

You can get into some weird complications at layer 3 but because
it doesn't touch each packet the proof it is fast is trivial.

>> Beyond that I have yet to see a clean semantics for anything
>> resembling your layer 2 layer 3 hybrid approach. If we can't have
>> clear semantics it is by definition impossible to implement correctly
>> because no one understands what it is supposed to do.
>
> IMHO that would be quite simple, have a 'namespace'
> for limiting port binds to a subset of the available
> ips and another one which does complete network

> virtualization with all the whistles and bells, IMHO
> most of them are orthogonal and can easily be combined
>
> - full network virtualization
> - lightweight ip subset
> - both

Quite possibly. The LSM will stay for a while so we do have
a clean way to restrict port binds.

>> Note. A true layer 3 approach has no impact on TCP/UDP filtering
>> because it filters at bind time not at packet reception time. Once you
>> start inspecting packets I don't see what the gain is from not going
>> all of the way to layer 2.
>
> IMHO this requirement only arises from the full system
> virtualization approach, just look at the other jail
> solutions (solaris, bsd, ...) some of them do not even
> allow for more than a single ip but they work quite
> well when used properly ...

Yes they do. Currently I am strongly opposed to Daniel Layer 2.5 approach
as I see no redeeming value in it. A good clean layer 3 approach I
avoid only because I think we can do better.

Eric

_____