Subject: Re: strict isolation of net interfaces
Posted by ebiederm on Fri, 30 Jun 2006 14:20:24 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.
>
> I have a few questions about all the network isolation stuff:
```

So far I have seen two viable possibilities on the table,
neither of them involve multiple names for a network device.

layer 3 (filtering the allowed ip addresses at bind time roughly the current vserver).
  - implementable as a security hook.
  - Benefit no measurable performance impact.
  - Downside not many things we can do.

layer 2 (What appears to applications a separate instance of the network stack).
  - Implementable as a namespace.
  - Each network namespace would have dedicated network devices.
  - Benefit extremely flexible.
  - Downside since at least the slow path must examine the packet
    it has the possibility of slowing down the networking stack.

For me the important characteristics.
- Allows for application migration, when we take our ip address with us.
  In particular it allows for importation of addresses assignments
  mad on other machines.

- No measurable impact on the existing networking when the code
  is compiled in.

- Clean predictable semantics.

This whole debate on network devices show up in multiple network namespaces is just silly. The only reason for wanting that appears to be better management. We have deeper issues like can we do a reasonable implementation without a network device showing up in multiple namespaces.

I think the reason the debate exists at all is that it is a very approachable topic, as opposed to the fundamentals here.

If we can get layer 2 level isolation working without measurable overhead with one namespace per device it may be worth revisiting things. Until then it is a side issue at best.

Eric