Subject: Re: Re: Network virtualization/isolation
Posted by Daniel Lezcano on Wed, 29 Nov 2006 22:10:39 GMT
View Forum Message <> Reply to Message

Brian Haley wrote:
> Eric W. Biederman wrote:
>> I think for cases across network socket namespaces it should
>> be a matter for the rules, to decide if the connection should
>> happen and what error code to return if the connection does not
>> happen.
>>
>> There is a potential in this to have an ambiguous case where two
>> applications can be listening for connections on the same socket
>> on the same port and both will allow the connection. If that
>> is the case I believe the proper definition is the first socket
>> that we find that will accept the connection gets the connection.
No. If you try to connect, the destination IP address is assigned to a
network namespace. This network namespace is used to leave the listening
socket ambiguity.
>
> Wouldn't you want to catch this at bind() and/or configuration time and
> fail? Having overlapping namespaces/rules seems undesirable, since as
> Herbert said, can get you "unexpected behaviour".

Overlapping is not a problem, you can have several sockets binded on the
same INADDR_ANY/port without ambiguity because the network namespace
pointer is added as a new key for sockets lookup, (src addr, src port,
dst addr, dst port, net ns pointer). The bind should not be forced to a
specific address because you will not be able to connect via 127.0.0.1.

>
>> I think with the appropriate set of rules it provides what is needed
>> for application migration. I.e. 127.0.0.1 can be filtered so that
>> you can only connect to sockets in your current container.
>>
>> It does get a little odd because it does allow for the possibility
>> that you can have multiple connected sockets with same source ip,
>> source port, destination ip, destination port. If the rules are
>> setup appropriately. I don't see that peculiarity being visible on
>> the outside network so it shouldn't be a problem.
>
> So if they're using the same protocol (eg TCP), how is it decided which
> one gets an incoming packet? Maybe I'm missing something as I don't
> understand your inside/outside network reference - is that to the
> loopback address comment in the previous paragraph?

The sockets for l3 isolation are isolated like the l2 (this is common
code). The difference is where the network namespace is found and used.

At the layer 2, it is at the network device level where the namespace is found. At the layer 3, from the IP destination. So when you arrive to sockets level, you have the network namespace packet destination information and you search for sockets related to the specific namespace.

-- Daniel