Subject: Re: [patch 2/6] [Network namespace] Network device sharing by view
Posted by Herbert Poetzl on Mon, 26 Jun 2006 22:54:40 GMT
View Forum Message <> Reply to Message

On Mon, Jun 26, 2006 at 03:13:17PM -0700, Ben Greear wrote:
> Eric W. Biederman wrote:
>
> >Basically it is just a matter of:
> >if (dest_mac == my_mac1) it is for device 1.
> >If (dest_mac == my_mac2) it is for device 2.
> >etc.
> >
> >At a small count of macs it is trivial to understand it will go
> >fast for a larger count of macs it only works with a good data
> >structure. We don't hit any extra cache lines of the packet,
> >and the above test can be collapsed with other routing lookup tests.
>
> I think you should do this at the layer-2 level, well before you get
> to routing. That will make the virtual mac-vlan work with arbitrary
> protocols and appear very much like a regular ethernet interface.
> This approach worked well with .1q vlans, and with my version of the
> mac-vlan module.

yes, that sounds good to me, any numbers how that
affects networking in general (performance wise and
memory wise, i.e. caches and hashes) ...

> Using the mac-vlan and source-based routing tables, I can give a
> unique 'interface' to each process and have each process able to bind
> to the same IP port, for instance. Using source-based routing (by
> binding to a local IP explicitly and adding a route table for that
> source IP), I can give unique default routes to each interface as
> well. Since we cannot have more than 256 routing tables, this approach
> is currently limitted to around 250 virtual interfaces, but that is
> still a substantial amount.

an typically that would be sufficient IMHO, but
of course, a more 'general' hash tag would be
better in the long run ...

> My mac-vlan patch, redirect-device patch, and other hackings are
> consolidated in this patch:
>
> http://www.candelatech.com/oss/candela_2.6.16.patch

great! thanks!

best,

Herbert

> Thanks,
> Ben
>
> --
> Ben Greear <greearb@candelatech.com>
> Candela Technologies Inc http://www.candelatech.com

Subject: Re: [patch 2/6] [Network namespace] Network device sharing by view
Posted by Ben Greear on Mon, 26 Jun 2006 23:08:23 GMT

View Forum Message <> Reply to Message

Herbert Poetzl wrote:
> On Mon, Jun 26, 2006 at 03:13:17PM -0700, Ben Greear wrote:

> yes, that sounds good to me, any numbers how that
> affects networking in general (performance wise and
> memory wise, i.e. caches and hashes) ...

I'll run some tests later today. Based on my previous tests,
I don't remember any significant overhead.

>>Using the mac-vlan and source-based routing tables, I can give a
>>unique 'interface' to each process and have each process able to bind
>>to the same IP port, for instance. Using source-based routing (by
>>binding to a local IP explicitly and adding a route table for that
>>source IP), I can give unique default routes to each interface as
>>well. Since we cannot have more than 256 routing tables, this approach
>>is currently limitted to around 250 virtual interfaces, but that is
>>still a substantial amount.
>
>
> an typically that would be sufficient IMHO, but
> of course, a more 'general' hash tag would be
> better in the long run ...

I'm willing to offer a bounty (hardware, beer, money, ...)
if someone will 'fix' this so we can have 1000 or more routes....

Being able to select these routes at a more global level (without
having to specifically bind to a local IP would be nice as well.)

Ben

--
Ben Greear <greearb@candelatech.com>

Subject: Re: [patch 2/6] [Network namespace] Network device sharing by view
Posted by Ben Greear on Tue, 27 Jun 2006 16:07:38 GMT
View Forum Message <> Reply to Message

Ben Greear wrote:
> Herbert Poetzl wrote:
>
>> On Mon, Jun 26, 2006 at 03:13:17PM -0700, Ben Greear wrote:
>
>
>> yes, that sounds good to me, any numbers how that
>> affects networking in general (performance wise and
>> memory wise, i.e. caches and hashes) ...
>
>
> I'll run some tests later today. Based on my previous tests,
> I don't remember any significant overhead.

Here's a quick benchmark using my redirect devices (RDD). Each
RDD comes in a pair...when you tx on one, the pkt is rx'd on the peer.
The idea is that it is exactly like two physical ethernet interfaces
connected by a cross-over cable.

My test system is a 64-bit dual-core Intel system, 3.013 Ghz processor with 1GB RAM.
Fairly standard stuff..it's one of the Shuttle XPC systems.
Kernel is 2.6.16.16 (64-bit).

Test setup is: rdd1 -- rdd2   [bridge]   rdd3 -- rdd4

I am using my proprietary module for the bridge logic...and the default
bridge should be at least this fast. I am injecting 1514 byte packets
on rdd1 and rdd4 with pktgen (bi-directional flow). My pktgen is also
receiving the pkts and gathering stats.

This setup sustains 1.7Gbps of generated and received traffic between
rdd1 and rdd4.

Running only the [bridge] between two 10/100/1000 ports on an Intel PCI-E
NIC will sustain about 870Mbps (bi-directional) on this system, so the
virtual devices are quite efficient, as suspected.

I have not yet had time to benchmark the mac-vlans...hopefully later today.

Thanks,

Ben

--
Ben Greear <greearb@candelatech.com>
Candela Technologies Inc http://www.candelatech.com

Subject: Re: [patch 2/6] [Network namespace] Network device sharing by view
Posted by Herbert Poetzl on Tue, 27 Jun 2006 22:48:37 GMT

View Forum Message <> Reply to Message

On Tue, Jun 27, 2006 at 09:07:38AM -0700, Ben Greear wrote:
> Ben Greear wrote:
> >Herbert Poetzl wrote:
> >
> >>On Mon, Jun 26, 2006 at 03:13:17PM -0700, Ben Greear wrote:
> >
> >>yes, that sounds good to me, any numbers how that
> >>affects networking in general (performance wise and
> >>memory wise, i.e. caches and hashes) ...
> >
> >I'll run some tests later today. Based on my previous tests,
> >I don't remember any significant overhead.
>
> Here's a quick benchmark using my redirect devices (RDD). Each RDD
> comes in a pair...when you tx on one, the pkt is rx'd on the peer.
> The idea is that it is exactly like two physical ethernet interfaces
> connected by a cross-over cable.
>
> My test system is a 64-bit dual-core Intel system, 3.013 Ghz processor
> with 1GB RAM. Fairly standard stuff..it's one of the Shuttle XPC
> systems. Kernel is 2.6.16.16 (64-bit).
>
>
> Test setup is: rdd1 -- rdd2   [bridge]   rdd3 -- rdd4
>
> I am using my proprietary module for the bridge logic...and the
> default bridge should be at least this fast. I am injecting 1514 byte
> packets on rdd1 and rdd4 with pktgen (bi-directional flow). My pktgen
> is also receiving the pkts and gathering stats.
>
> This setup sustains 1.7Gbps of generated and received traffic between
> rdd1 and rdd4.
>
> Running only the [bridge] between two 10/100/1000 ports on an Intel
> PCI-E NIC will sustain about 870Mbps (bi-directional) on this system,
> so the virtual devices are quite efficient, as suspected.
>

> I have not yet had time to benchmark the mac-vlans...hopefully later
> today.

hmm, maybe you could also benchmark loopback connections
(and their throughput) on your system?

my (not so fancy) PIII, 32bit, 2.6.17.1 seems to do
roughly 2Gbs on the loopback device (tested with dd
and netcat)

best,
Herbert

> Thanks,
> Ben
>
> --
> Ben Greear <greearb@candelatech.com>
> Candela Technologies Inc http://www.candelatech.com