Subject: Re: L2 network namespace benchmarking
Posted by ebiederm on Thu, 29 Mar 2007 13:01:20 GMT
View Forum Message <> Reply to Message

Benjamin Thery <benjamin.thery@bull.net> writes:

> Eric W. Biederman wrote:
>> Daniel Lezcano <daniel.lezcano@free.fr> writes:
>
> [...]
>
>>> * When do you expect to have the network namespace into mainline ?
>> My current goal is to finish my rebase against 2.6.linus_lastest in
>> the next couple of days after having figured out how to deal with sysfs.
>
> Great news!
> I also have some questions about this updated version:
>
> - Have you integrated the bug fixes and cleanups(*) Daniel wrote for
> your previous netns patchset (and the few glitches I reported too)?

About half of them so far. It is my intention to incorporate all of them.
They weren't all trivial to include. A couple of Daniel's patches
address a real issue in the wrong way so I have to give them some more
thought.

> (*) available in LXC8 patchset
>
> - Do you already have a public git repository set up for the rebase?
> - If it is private, any plan to make it public soon? (That would be great)
Yes. Where the current one is now.

>> I have been doing reviewing in more code then I know what to do with,
>> and fighting some very strange bugs during the stabilization window.
>> Which has kept me from doing additional development. Plus I have
>> had a cold.
>
> I hope you're getting better... and you'll be able to provide us the
> updated patchset very soon :)

Hopefully. I think I have fixed my last non network regression I know
about for 2.6.21-rcX. Which means I can begin to focus again.

> [...]
>
>> If I read the results right it took a 32bit machine from AMD with
>> a gigabit interface before you could measure a throughput difference.
>> That isn't shabby for a non-optimized code path.

>
> Indeed the throughput difference is not significant.
> This is very good to see that it stays constant when using the container.
> What I'm more worried about is the CPU load increase. But it seems
> we've identified some of the culprits.

Yes, and the good news is that they all seem to be in getting the
packets to the network namespace.

> This afternoon I had a look at why the bridge setup isn't better than
> the route setup (section 2.3 and 2.4 of Daniel's report).
>
> In the bridge case, we encounter the same problems as the routes case.
> The oprofile profile is the same: the most demanding routines are
> pskb_expand_head and csum_partial_copy_generic.
> pskb_expand_head() is also called by skb_cow(), but this time
> skb_cow() is called by netfilter's nf_bridge_copy_header().
>
> We can avoid this copy by removing option CONFIG_BRIDGE_NETFILTER.
> This copy is made even if netfilter is not used on the host.
> Maybe some optimizations can be made in netfilter's code to prevent this.

Sounds reasonable. I guess the next step is to get some numbers with
CONFIG_BRIDGE_NETFILTER disabled. (So we don't hit that case and just
in case there are more). I suspect the bridging code has a small
enough user base right now it just hasn't been optimized much.

Eric

_____
Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers