Subject: Re: Grr sysfs networking changes...
Posted by ebiederm on Tue, 03 Apr 2007 15:57:30 GMT
View Forum Message <> Reply to Message

Daniel Lezcano <dlezcano@fr.ibm.com> writes:

> Kirill Korotaev wrote:
>> Eric,
>>
>> can you please describe a bit what sysfs problems you experience and
>> what approach you want to you to sysfs virtualization?

Yes. Although I refuse to think of it as virtualization. :)

> Great !
>
> Can you describe what you are experiencing ?
> I did a try to port the netns patchset to the latest -mm kernel and I fall into
> some sysfs problems. It can be related to sysfs changes or I could have made
> some mistakes while doing the port. Having the sysfs problems you experience can
> help to find them :)
> Can be the problems related to the usage of struct device instead of struct
> class_device in some cases ?

My problem starts with the usage of struct device instead of struct
class_device.

The implementation of sysfs with respect to network devices seems very
brittle and redundant.

When everything was struct class_device how the network devices showed
up in sysfs was very simple. They were all in a directory under
/sys/class/net/. So all I did was to implement a magic follow link
method so I could have multiple copies of that directory.

Now if you disable CONFIG_SYSFS_DEPRECATED they don't just show up
under /sys/class/net but also under:

> # find /sys -name 'net' | xargs ls -l
> /sys/class/net:
> total 0
> lrwxrwxrwx 1 root root 0 Apr 3 09:27 eth0 ->
../../devices/pci0000:00/0000:00:1c.5/0000:04:00.0/net/eth0
> lrwxrwxrwx 1 root root 0 Apr 3 09:20 lo -> ../../devices/virtual/net/lo
> lrwxrwxrwx 1 root root 0 Apr 3 09:20 sit0 -> ../../devices/virtual/net/sit0
>
> /sys/devices/pci0000:00/0000:00:1c.5/0000:04:00.0/net:
> total 0

> drwxr-xr-x 4 root root 0 Apr 3 09:27 eth0
>
> /sys/devices/virtual/net:
> total 0
> drwxr-xr-x 4 root root 0 Apr 3 09:27 lo
> drwxr-xr-x 4 root root 0 Apr 3 09:27 sit0

So the network devices that have real parents are all over the place,
in lots of different directories all over sysfs, ugh. Plus we have
to cope with directories like
"/sys/devices/pci0000:00/0000:00:1c.5/0000:04:00.0/net" coming and
going (hot plugged devices you know) while the network namespace
is mounted. Which can easily lead to all kinds of interesting
races.

Another piece of the problem is that sysfs just seems inherently to
have a lot of duplicate information and to be very fragile. Which
makes modifying the sysfs code hard and more work than it should be.

I really don't care what the solution is so long as it is reasonably
simple, efficient and maintainable.

I'm open to suggestions.

Currently adopting the technique of mounting the sysfs multiple times
(which was the solution for proc) looks hard as the sysfs
representation in memory is partly dentry based.

Having a filter method on directory contents sounds good but it fails
because the dcache can't cope with two directory entries with the same
name. Hmm.. If I can replace d_compare and have that also compare
by the network namespace I might be able to make filtering work.

So mostly I am looking at fixing up the techniques that I used for
one sysfs directory. But the races and corner cases are nasty.

Eric

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