Subject: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by Pavel Emelianov on Tue, 09 Oct 2007 12:19:25 GMT
View Forum Message <> Reply to Message

Currently indexes for netdevices come sequentially one by
one, and the same stays true even for devices that are
created for namespaces.

Side effects of this are:
 * lo device has not 1 index in a namespace. This may break
   some userspace that relies on it (and AFAIR something
   really broke in OpenVZ VEs without this);
 * after some time namespaces will have devices with indexes
   like 1000000 os similar. This might be confusing for a
   human (tools will not mind).

So move the (currently "global" and static) ifindex variable
on the struct net, making the indexes allocation look more
like on a standalone machine.

Moreover - when we have indexes intersect between namespaces,
we may catch more BUGs in the future related to "wrong device
was found for a given index".

Signed-off-by: Pavel Emelyanov <xemul@openvz.org>

```
diff --git a/include/net/net_namespace.h b/include/net/net_namespace.h
index 93aa87d..83a18d0 100644
--- a/include/net/net_namespace.h
+++ b/include/net/net_namespace.h
@@ -29,6 +29,8 @@ struct net {
  struct list_head  dev_base_head;
  struct hlist_head  *dev_name_head;
  struct hlist_head *dev_index_head;
+
+ int   ifindex;
 };

 #ifdef CONFIG_NET
diff --git a/net/core/dev.c b/net/core/dev.c
index e7e728a..a08ed8c 100644
--- a/net/core/dev.c
+++ b/net/core/dev.c
@@ -3443,12 +3443,11 @@ int dev_ioctl(struct net *net, unsigned
  */
 static int dev_new_index(struct net *net)
 {
- static int ifindex;
  for (;;) {
-  if (++ifindex <= 0)
-   ifindex = 1;
-  if (!__dev_get_by_index(net, ifindex))
-   return ifindex;
+  if (++net->ifindex <= 0)
+   net->ifindex = 1;
+  if (!__dev_get_by_index(net, net->ifindex))
+   return net->ifindex;
  }
 }
```

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by Daniel Lezcano on Tue, 09 Oct 2007 14:48:16 GMT
View Forum Message <> Reply to Message

Pavel Emelyanov wrote:

> Currently indexes for netdevices come sequentially one by
> one, and the same stays true even for devices that are
> created for namespaces.
>
> Side effects of this are:
> * lo device has not 1 index in a namespace. This may break
>   some userspace that relies on it (and AFAIR something
>   really broke in OpenVZ VEs without this);
> * after some time namespaces will have devices with indexes
>   like 1000000 os similar. This might be confusing for a
>   human (tools will not mind).
>
> So move the (currently "global" and static) ifindex variable
> on the struct net, making the indexes allocation look more
> like on a standalone machine.
>
> Moreover - when we have indexes intersect between namespaces,
> we may catch more BUGs in the future related to "wrong device
> was found for a given index".
>
> Signed-off-by: Pavel Emelyanov <xemul@openvz.org>

Applied and tested against netns49. Works fine.

Acked-by: Daniel Lezcano <dlezcano@fr.ibm.com>

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by David Stevens on Tue, 09 Oct 2007 16:18:25 GMT

View Forum Message <> Reply to Message

Sorry if this is a dumb question, but what is the model you intend for
SNMP? Do you want each namespace to be its own virtual machine with
its own, separate MIB?

Ifindex's have to uniquely identify the interface (virtual or otherwise)
to remote
queriers (not just local applications), so unless you pay the price of
separating
all the SNMP MIBs per namespace too, it seems you'll need some way to
remap these for SNMP queries, right?

+-DLS

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by ebiederm on Tue, 09 Oct 2007 17:41:16 GMT

View Forum Message <> Reply to Message

Pavel Emelyanov <xemul@openvz.org> writes:

> Currently indexes for netdevices come sequentially one by
> one, and the same stays true even for devices that are
> created for namespaces.
>
> Side effects of this are:
> * lo device has not 1 index in a namespace. This may break
>   some userspace that relies on it (and AFAIR something
>   really broke in OpenVZ VEs without this);

As it happens lo hasn't been registered first for some time
so it hasn't had ifindex of 1 in the normal kernel.

> * after some time namespaces will have devices with indexes
>   like 1000000 os similar. This might be confusing for a
>   human (tools will not mind).

Only if we wind up creating that many devices.

> So move the (currently "global" and static) ifindex variable
> on the struct net, making the indexes allocation look more
> like on a standalone machine.
>
> Moreover - when we have indexes intersect between namespaces,
> we may catch more BUGs in the future related to "wrong device

> was found for a given index".

Not yet.

I know there are several data structures internal to the kernel that
are indexed by ifindex, and not struct net_device *. There is the
iflink field in struct net_device. We need a way to refer to network
devices in other namespaces in rtnetlink in an unambiguous way.   I
don't see any real problems with a global ifindex assignment until
we start migrating applications.

So please hold off on this until the kernel has been audited and
we have removed all of the uses of ifindex that assume ifindex is
global, that we can find.

Right now a namespace local ifindex seems to be just asking for
trouble.

Eric

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by ebiederm on Tue, 09 Oct 2007 17:43:58 GMT

View Forum Message <> Reply to Message

David Stevens <dlstevens@us.ibm.com> writes:

> Sorry if this is a dumb question, but what is the model you intend for
> SNMP? Do you want each namespace to be its own virtual machine with
> its own, separate MIB?

Each network namespace appears to user space as a completely separate
network stack. So yes a separate instance of the MIB is appropriate.

Eric

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by davem on Tue, 09 Oct 2007 20:11:13 GMT

View Forum Message <> Reply to Message

From: ebiederm@xmission.com (Eric W. Biederman)
Date: Tue, 09 Oct 2007 11:43:58 -0600

> David Stevens <dlstevens@us.ibm.com> writes:
>
> > Sorry if this is a dumb question, but what is the model you intend for

> > SNMP? Do you want each namespace to be its own virtual machine with
> > its own, separate MIB?
>
> Each network namespace appears to user space as a completely separate
> network stack. So yes a separate instance of the MIB is appropriate.

We don't think you can validly do that, as David tried to explain.

The interface indexes are visible remotely to remote SNMP querying
applications. They have to be unique within the physical system.

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by davem on Tue, 09 Oct 2007 20:12:11 GMT

View Forum Message <> Reply to Message

From: David Stevens <dlstevens@us.ibm.com>
Date: Tue, 9 Oct 2007 09:18:25 -0700

> Ifindex's have to uniquely identify the interface (virtual or
> otherwise) to remote queriers (not just local applications), so
> unless you pay the price of separating all the SNMP MIBs per
> namespace too, it seems you'll need some way to remap these for SNMP
> queries, right?

I don't see how it can work even with per-namespace MIBs,
the interface indexes have to be unique per "system".

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by davem on Tue, 09 Oct 2007 20:12:32 GMT

View Forum Message <> Reply to Message

From: Pavel Emelyanov <xemul@openvz.org>
Date: Tue, 09 Oct 2007 16:19:25 +0400

> Currently indexes for netdevices come sequentially one by
> one, and the same stays true even for devices that are
> created for namespaces.
>
> Side effects of this are:
> * lo device has not 1 index in a namespace. This may break
>   some userspace that relies on it (and AFAIR something
>   really broke in OpenVZ VEs without this);
> * after some time namespaces will have devices with indexes
>   like 1000000 os similar. This might be confusing for a
>   human (tools will not mind).

>
> So move the (currently "global" and static) ifindex variable
> on the struct net, making the indexes allocation look more
> like on a standalone machine.
>
> Moreover - when we have indexes intersect between namespaces,
> we may catch more BUGs in the future related to "wrong device
> was found for a given index".
>
> Signed-off-by: Pavel Emelyanov <xemul@openvz.org>

Based upon Eric's and other's comments, I'm holding off on
this for now.

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by ebiederm on Tue, 09 Oct 2007 21:00:10 GMT

View Forum Message <> Reply to Message

David Miller <davem@davemloft.net> writes:

> From: ebiederm@xmission.com (Eric W. Biederman)
> Date: Tue, 09 Oct 2007 11:43:58 -0600
>
>> David Stevens <dlstevens@us.ibm.com> writes:
>>
>> > Sorry if this is a dumb question, but what is the model you intend for
>> > SNMP? Do you want each namespace to be its own virtual machine with
>> > its own, separate MIB?
>>
>> Each network namespace appears to user space as a completely separate
>> network stack. So yes a separate instance of the MIB is appropriate.
>
> We don't think you can validly do that, as David tried to explain.
>
> The interface indexes are visible remotely to remote SNMP querying
> applications. They have to be unique within the physical system.

I think figuring out what we are doing with SNMP is not any harder
or easier then any other user space interface, and like I said I
don't think we are ready yet.

>From the perspective of monitoring network namespaces make the entire
system looks more like a cluster then it does a single machine, and
that is how I would look at portraying the system to SNMP if I had to
do that work today. A switch with a bunch of different machines
behind it. Especially in the context of container migration this
becomes an attractive model.

Regardless it is early yet and there is plenty of time to revisit this
after we solved the easier and less controversial problems.

Eric

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by davem on Tue, 09 Oct 2007 21:17:31 GMT

View Forum Message <> Reply to Message

From: ebiederm@xmission.com (Eric W. Biederman)
Date: Tue, 09 Oct 2007 15:00:10 -0600

> Regardless it is early yet and there is plenty of time to revisit this
> after we solved the easier and less controversial problems.

Ok.

I would encourage you to learn how the SNMP mibs work, and whether
they associate things with interfaces and/or unique MAC addresses.
The semantics may have conflicts with your envisioned cluster
abstraction.

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by Pavel Emelianov on Wed, 10 Oct 2007 08:55:32 GMT

View Forum Message <> Reply to Message

Eric W. Biederman wrote:

```
> Pavel Emelyanov <xemul@openvz.org> writes:
>
>> Currently indexes for netdevices come sequentially one by
>> one, and the same stays true even for devices that are
>> created for namespaces.
>>
>> Side effects of this are:
>>  * lo device has not 1 index in a namespace. This may break
>>    some userspace that relies on it (and AFAIR something
>>    really broke in OpenVZ VEs without this);
>
> As it happens lo hasn't been registered first for some time
> so it hasn't had ifindex of 1 in the normal kernel.
>
>>  * after some time namespaces will have devices with indexes
>>    like 1000000 os similar. This might be confusing for a
>>    human (tools will not mind).
```

>
> Only if we wind up creating that many devices.

Nope. Create and destroy new net ns for 10000 times and you'll get it.

>> So move the (currently "global" and static) ifindex variable
>> on the struct net, making the indexes allocation look more
>> like on a standalone machine.
>>
>> Moreover - when we have indexes intersect between namespaces,
>> we may catch more BUGs in the future related to "wrong device
>> was found for a given index".
>
> Not yet.
>
> I know there are several data structures internal to the kernel that
> are indexed by ifindex, and not struct net_device *. There is the
> iflink field in struct net_device. We need a way to refer to network
> devices in other namespaces in rtnetlink in an unambiguous way.   I
> don't see any real problems with a global ifindex assignment until
> we start migrating applications.
>
> So please hold off on this until the kernel has been audited and
> we have removed all of the uses of ifindex that assume ifindex is
> global, that we can find.

Ok.

> Right now a namespace local ifindex seems to be just asking for
> trouble.

You said the same about caching the global pid on the task_struct,
but looks like you were wrong ;) Just kidding.

> Eric
>
>

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by ebiederm on Wed, 10 Oct 2007 18:15:33 GMT

View Forum Message <> Reply to Message

Pavel Emelyanov <xemul@openvz.org> writes:

>> I know there are several data structures internal to the kernel that
>> are indexed by ifindex, and not struct net_device *. There is the
>> iflink field in struct net_device. We need a way to refer to network

>> devices in other namespaces in rtnetlink in an unambiguous way.   I
>> don't see any real problems with a global ifindex assignment until
>> we start migrating applications.
>>
>> So please hold off on this until the kernel has been audited and
>> we have removed all of the uses of ifindex that assume ifindex is
>> global, that we can find.
>
> Ok.

Thanks.

Eric

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by Johannes Berg on Wed, 10 Oct 2007 18:34:03 GMT

View Forum Message <> Reply to Message

On Tue, 2007-10-09 at 11:41 -0600, Eric W. Biederman wrote:

> So please hold off on this until the kernel has been audited and
> we have removed all of the uses of ifindex that assume ifindex is
> global, that we can find.

I certainly have this assumption in the wireless code (cfg80211). How
would I go about removing it? Are netlink sockets per-namespace so I can
use the namespace of the netlink socket to look up a netdev?

johannes

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by ebiederm on Wed, 10 Oct 2007 19:51:50 GMT

View Forum Message <> Reply to Message

Johannes Berg <johannes@sipsolutions.net> writes:

> On Tue, 2007-10-09 at 11:41 -0600, Eric W. Biederman wrote:
>
>> So please hold off on this until the kernel has been audited and
>> we have removed all of the uses of ifindex that assume ifindex is
>> global, that we can find.
>
> I certainly have this assumption in the wireless code (cfg80211). How
> would I go about removing it? Are netlink sockets per-namespace so I can
> use the namespace of the netlink socket to look up a netdev?

Yes. Netlink sockets are per-namespace and you can use the namespace
of a netlink socket to look up a netdev.

Eric

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by Johannes Berg on Thu, 11 Oct 2007 09:32:30 GMT

View Forum Message <> Reply to Message

On Wed, 2007-10-10 at 13:51 -0600, Eric W. Biederman wrote:

> Yes. Netlink sockets are per-namespace and you can use the namespace
> of a netlink socket to look up a netdev.

Ok, thanks. I still haven't really looked into the wireless vs. net
namespaces problem but this will probably help.

johannes

Subject: Re: [PATCH][NETNS] Make ifindex generation per-namespace
Posted by ebiederm on Thu, 11 Oct 2007 17:22:57 GMT

View Forum Message <> Reply to Message

Johannes Berg <johannes@sipsolutions.net> writes:

> On Wed, 2007-10-10 at 13:51 -0600, Eric W. Biederman wrote:
>
>> Yes. Netlink sockets are per-namespace and you can use the namespace
>> of a netlink socket to look up a netdev.
>
> Ok, thanks. I still haven't really looked into the wireless vs. net
> namespaces problem but this will probably help.

I think I may even have some patches in my proof of concept tree that
address some of the wireless issues. Especially rtnetlink ones.
Generally those cases haven't been hard to spot.

Having hash tables and the like that hash and do key compares
on an ifindex instead of a net_device * are the in kernel places that
make it very hard to have duplicate ifindexes.

Thinking about it probably the biggest challenge to deal with
is iff in struct sk_buff.

Eric