Subject: Re: Pid namespaces problems
Posted by ebiederm on Sun, 04 Nov 2007 04:06:48 GMT
View Forum Message <> Reply to Message

Pavel Emelyanov <xemul@openvz.org> writes:

> Hi, Eric, Suka.
>
> Eric, you and Ulrich claim that pid namespaces are full of BUGs.
> Can you please share you BUG list with me, so I could correct
> mine.

To be clear. I think the current pid namespace work is incomplete.
I do not think the pid namespaces is fundamentally buggy the way
Ingo and Ulrich were suggesting (my apologies for the delayed
reply I have been away from my computer).

I think I shared just about everything I know of off the top
of my head in earlier threads. But I haven't tried to
find an exhaustive list of uncorrected code as they pop
up fairly easily when I audit various pid users. Which
lead me to conclude that the pid namespace is not complete.

> Things as I see them now are the following:
> 1. signals delivery is not perfect in the namespace
> 2. fs/lock.c will report wrong ids in the namespace
> 3. some kernel threads (nfs) still use old api (relevant
> for a namespace only)
> 4. tsk->pid and tsk->tgid should not be explicitly used
A wrapper that gets those values for use in printk.

As a general principle I am opposed to using global pid values
(except in kernel print statements). Using them we continue
to have pid wrap around issues if we store them, and mixing
global pid_t values and non-global pid_t values is all too possible.

For example:
fs/autofs/inode.c: line 83 *pgrp = task_pgrp_nr(current);
fs/autofs/inode.c: line 117: *pgrp = option;

We are simultaneously assigning a global pid and a pid from
the current pid namespace to the same variable. Ouch!

Used in anything but the init_pid namespace that code is wrong.

So with stupid things like that I would very much like to
convert everything to storing and comparing struct pid pointers
which have essentially the same cost as pid_t values, can

be used the same way, but cannot be accidentally mixed with
with pid_t operations.  So they are just less error prone.

> I'd appreciate some specific information, like "the ttys
> in drivers/char/tty_io.c may break the pid refcounting"
> rather than abstract "this is not clear whether the
> refcounting is good in fs/locks.c"

You are picking on the one instance that I figured I would need
further review to see if there was work that needed to be done.  That
is what I meant by unclear.  I didn't know if the code was safe and
the code wasn't using one of the idioms that would have made me
certain that the code was safe.

- We need to store a struct pid reference on the sysvipc semaphores (and
  probably the other sysvipc objects) so that if they are used across
  namespace boundaries we can convert and give processes the pid for
  their local namespace.

- There are several architectures with their own signal functions for
  other OS compatibility that have are using _pid and not _vpid
  variants of functions. (irix and solaris)
  arch/mips/kernel/irixsig.c:irix_waitsys
  arch/mips/kernel/sysirix.c:irix_setpgrp
  arch/sparc64/solaris/misc.c:solaris_procids

Eric

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

---

Subject: Re: Pid namespaces problems
Posted by Pavel Emelianov on Tue, 06 Nov 2007 07:39:00 GMT
View Forum Message <> Reply to Message

Eric W. Biederman wrote:
> Pavel Emelyanov <xemul@openvz.org> writes:
>
>> Hi, Eric, Suka.
>>
>> Eric, you and Ulrich claim that pid namespaces are full of BUGs.
>> Can you please share you BUG list with me, so I could correct
>> mine.
>
> To be clear.  I think the current pid namespace work is incomplete.
> I do not think the pid namespaces is fundamentally buggy the way

> Ingo and Ulrich were suggesting (my apologies for the delayed
> reply I have been away from my computer).

Thanks :)

> I think I shared just about everything I know of off the top
> of my head in earlier threads.  But I haven't tried to
> find an exhaustive list of uncorrected code as they pop
> up fairly easily when I audit various pid users.  Which
> lead me to conclude that the pid namespace is not complete.
>
>> Things as I see them now are the following:
>> 1. signals delivery is not perfect in the namespace
>> 2. fs/lock.c will report wrong ids in the namespace
>> 3. some kernel threads (nfs) still use old api (relevant
>>    for a namespace only)
>> 4. tsk->pid and tsk->tgid should not be explicitly used
>      A wrapper that gets those values for use in printk.
>
> As a general principle I am opposed to using global pid values
> (except in kernel print statements).  Using them we continue
> to have pid wrap around issues if we store them, and mixing
> global pid_t values and non-global pid_t values is all too possible.
>
> For example:
>   fs/autofs/inode.c: line 83 *pgrp = task_pgrp_nr(current);
>   fs/autofs/inode.c: line 117: *pgrp = option;
>
> We are simultaneously assigning a global pid and a pid from
> the current pid namespace to the same variable.  Ouch!

Yup. I agree with this too. I'm about to deprecate the pid and tgid
fields in the task_struct (but not remove to make printk-s faster
and smaller).

The work is in progress here.

> Used in anything but the init_pid namespace that code is wrong.
>
> So with stupid things like that I would very much like to
> convert everything to storing and comparing struct pid pointers
> which have essentially the same cost as pid_t values, can
> be used the same way, but cannot be accidentally mixed with
> with pid_t operations.  So they are just less error prone.

Agree. I'm working on this as well.

>> I'd appreciate some specific information, like "the ttys

>> in drivers/char/tty_io.c may break the pid refcounting"
>> rather than abstract "this is not clear whether the
>> refcounting is good in fs/locks.c"
>
> You are picking on the one instance that I figured I would need
> further review to see if there was work that needed to be done.  That
> is what I meant by unclear.  I didn't know if the code was safe and
> the code wasn't using one of the idioms that would have made me
> certain that the code was safe.
>
> - We need to store a struct pid reference on the sysvipc semaphores (and
>   probably the other sysvipc objects) so that if they are used across
>   namespace boundaries we can convert and give processes the pid for
>   their local namespace.

Hm.. What if they are used across two not-connected namespaces? E.g.
two different children of init namespace?

> - There are several architectures with their own signal functions for
>   other OS compatibility that have are using _pid and not _vpid
>   variants of functions. (irix and solaris)
>   arch/mips/kernel/irixsig.c:irix_waitsys
>   arch/mips/kernel/sysirix.c:irix_setpgrp
>   arch/sparc64/solaris/misc.c:solaris_procids

Ok. Looks like your list is the same as mine. That's good to hear
that I haven't missed anything important.

So, I see that you're about to take a closer look at the pid
namespaces. If so, then what time can we expect the net namespace
activity to go on? Or (if you don't mind) can we start merging
the patches to David as soon as he opens his 2.6.25 merge window?

> Eric
>

Thanks,
Pavel



---

Subject: Re: Pid namespaces problems
Posted by ebiederm on Tue, 06 Nov 2007 16:28:19 GMT
View Forum Message <> Reply to Message

Pavel Emelyanov <xemul@openvz.org> writes:

> Yup. I agree with this too. I'm about to deprecate the pid and tgid
> fields in the task_struct (but not remove to make printk-s faster
> and smaller).
>
> The work is in progress here.

Last time I looked at the printks, it looked to me like we wanted
to print both the global pid and tsk->comm.  When we printed which
process did what, and helpers to do that would be nice.

printk of any of this information is rare enough that size is
the important thing not speed.

I'm hoping we can come up with a nice wrapper to  use with
printk that we can bury all of the logic in.  Possibly something
like NIP_QUAD.

>>
>> - We need to store a struct pid reference on the sysvipc semaphores (and
>>   probably the other sysvipc objects) so that if they are used across
>>   namespace boundaries we can convert and give processes the pid for
>>   their local namespace.
>
> Hm.. What if they are used across two not-connected namespaces? E.g.
> two different children of init namespace?

Well we get a 0 pid value.  But at least we get a correct pid value when
they are mixed in a way that it possible to have one.

There are also the credentials for AF_UNIX sockets that need this same
treatment.

>> - There are several architectures with their own signal functions for
>>   other OS compatibility that have are using _pid and not _vpid
>>   variants of functions. (irix and solaris)
>>   arch/mips/kernel/irixsig.c:irix_waitsys
>>   arch/mips/kernel/sysirix.c:irix_setpgrp
>>   arch/sparc64/solaris/misc.c:solaris_procids
>
> Ok. Looks like your list is the same as mine. That's good to hear
> that I haven't missed anything important.

I guess there are also all of the things we have been discussing
with respect to signals, and clone semantics.

In particular after having looked at the issue it looks like

we can send signals between namespaces, and it doesn't look
like it is going to hard to support setting si_pid properly.

So because we can do that I don't see any reason for the
extra special case in clone for a new pid namespace.

I expect the only places we really differ is how to handle
pids in the cross namespace cases.

> So, I see that you're about to take a closer look at the pid
> namespaces. If so, then what time can we expect the net namespace
> activity to go on? Or (if you don't mind) can we start merging
> the patches to David as soon as he opens his 2.6.25 merge window?

My current plan is roughly:
- Take a trip and be offline for about a week (sorry folks my timing sucks)
- Poke on pid namespaces until I have a good feeling about where
  things are going (especially the cross namespace work).
- Get back to merging the network namespaces.
  sysctl, sysfs, af_packet, af_unix, and eventually ipv4.

I try and work where I can do the most good.

Eric

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

---

Subject: Re: Pid namespaces problems
Posted by Daniel Lezcano on Wed, 07 Nov 2007 08:18:30 GMT
View Forum Message <> Reply to Message

Cedric Le Goater wrote:
>>> - There are several architectures with their own signal functions for
>>>   other OS compatibility that have are using _pid and not _vpid
>>>   variants of functions. (irix and solaris)
>>>   arch/mips/kernel/irixsig.c:irix_waitsys
>>>   arch/mips/kernel/sysirix.c:irix_setpgrp
>>>   arch/sparc64/solaris/misc.c:solaris_procids
>> Ok. Looks like your list is the same as mine. That's good to hear
>> that I haven't missed anything important.
>
> We've also talked about af_unix credentials.
>
>> So, I see that you're about to take a closer look at the pid
>> namespaces. If so, then what time can we expect the net namespace

>> activity to go on? Or (if you don't mind) can we start merging
>> the patches to David as soon as he opens his 2.6.25 merge window?
>
> I think daniel and benjamin are also getting ready for the 2.6.25
> merge window.

Yes, right.
We are identifying the ipv4 subset patches to send to David.



---

Subject: Re: Pid namespaces problems
Posted by Cedric Le Goater on Wed, 07 Nov 2007 08:28:29 GMT
View Forum Message <> Reply to Message

>
>> - There are several architectures with their own signal functions for
>>  other OS compatibility that have are using _pid and not _vpid
>>  variants of functions. (irix and solaris)
>>  arch/mips/kernel/irixsig.c:irix_waitsys
>>  arch/mips/kernel/sysirix.c:irix_setpgrp
>>  arch/sparc64/solaris/misc.c:solaris_procids
>
> Ok. Looks like your list is the same as mine. That's good to hear
> that I haven't missed anything important.

We've also talked about af_unix credentials.

> So, I see that you're about to take a closer look at the pid
> namespaces. If so, then what time can we expect the net namespace
> activity to go on? Or (if you don't mind) can we start merging
> the patches to David as soon as he opens his 2.6.25 merge window?

I think daniel and benjamin are also getting ready for the 2.6.25
merge window.

C.

Subject: Re: Pid namespaces problems
Posted by Daniel Lezcano on Wed, 07 Nov 2007 09:00:23 GMT
View Forum Message <> Reply to Message

Cedric Le Goater wrote:
>>> - There are several architectures with their own signal functions for
>>>  other OS compatibility that have are using _pid and not _vpid
>>>  variants of functions. (irix and solaris)
>>>  arch/mips/kernel/irixsig.c:irix_waitsys
>>>  arch/mips/kernel/sysirix.c:irix_setpgrp
>>>  arch/sparc64/solaris/misc.c:solaris_procids
>> Ok. Looks like your list is the same as mine. That's good to hear
>> that I haven't missed anything important.
>
> We've also talked about af_unix credentials.
>
>> So, I see that you're about to take a closer look at the pid
>> namespaces. If so, then what time can we expect the net namespace
>> activity to go on? Or (if you don't mind) can we start merging
>> the patches to David as soon as he opens his 2.6.25 merge window?
>
> I think daniel and benjamin are also getting ready for the 2.6.25
> merge window.

Yes. It can be cool if we can sync up Benjamin, Pavel, Denis, Eric and I with the different parts to be posted. Benjamin and I we began to look at ipv4. This is a big part, perhaps we can split that into several subset and dispatch them, except if Pavel and Denis already rebase ipv4 for net-2.6, in this case feel free to send them out.

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

Subject: Re: Pid namespaces problems
Posted by Pavel Emelianov on Wed, 07 Nov 2007 16:12:14 GMT
View Forum Message <> Reply to Message

Daniel Lezcano wrote:
> Cedric Le Goater wrote:
>>>> - There are several architectures with their own signal functions for
>>>>  other OS compatibility that have are using _pid and not _vpid
>>>>  variants of functions. (irix and solaris)

>>>>  arch/mips/kernel/irixsig.c:irix_waitsys
>>>>  arch/mips/kernel/sysirix.c:irix_setpgrp
>>>>  arch/sparc64/solaris/misc.c:solaris_procids
>>> Ok. Looks like your list is the same as mine. That's good to hear
>>> that I haven't missed anything important.
>> We've also talked about af_unix credentials.
>>
>>> So, I see that you're about to take a closer look at the pid
>>> namespaces. If so, then what time can we expect the net namespace
>>> activity to go on? Or (if you don't mind) can we start merging
>>> the patches to David as soon as he opens his 2.6.25 merge window?
>> I think daniel and benjamin are also getting ready for the 2.6.25
>> merge window.
>
> Yes. It can be cool if we can sync up Benjamin, Pavel, Denis, Eric and I
> with the different parts to be posted.

Yup. Team work will give us a chance to get in to the 2.6.25 with
the core virtualization. By core I mean unix, netlinux, ipv4 and ipv6.

Netfilters virtualization is a complex task :)

> Benjamin and I we began to look
> at ipv4. This is a big part, perhaps we can split that into several
> subset and dispatch them, except if Pavel and Denis already rebase ipv4
> for net-2.6, in this case feel free to send them out.

Well, actually we have almost moved to the net-2.6 with the ipv4
set. There are only some minor (I hope they are minor ;)) things.
So we would be glad to go on with ipv4 further. What's up with the
ipv6 patches, Daniel? You said that you and Benjamin make some
big progress in this area, no?

Thanks,
Pavel



---

Subject: Re: Pid namespaces problems
Posted by Daniel Lezcano on Wed, 07 Nov 2007 17:24:12 GMT
View Forum Message <> Reply to Message

Pavel Emelyanov wrote:
> Daniel Lezcano wrote:

>> Cedric Le Goater wrote:
>>>>> - There are several architectures with their own signal functions for
>>>>>  other OS compatibility that have are using _pid and not _vpid
>>>>>  variants of functions. (irix and solaris)
>>>>>  arch/mips/kernel/irixsig.c:irix_waitsys
>>>>>  arch/mips/kernel/sysirix.c:irix_setpgrp
>>>>>  arch/sparc64/solaris/misc.c:solaris_procids
>>>> Ok. Looks like your list is the same as mine. That's good to hear
>>>> that I haven't missed anything important.
>>> We've also talked about af_unix credentials.
>>>
>>>> So, I see that you're about to take a closer look at the pid
>>>> namespaces. If so, then what time can we expect the net namespace
>>>> activity to go on? Or (if you don't mind) can we start merging
>>>> the patches to David as soon as he opens his 2.6.25 merge window?
>>> I think daniel and benjamin are also getting ready for the 2.6.25
>>> merge window.
>> Yes. It can be cool if we can sync up Benjamin, Pavel, Denis, Eric and I
>> with the different parts to be posted.
>
> Yup. Team work will give us a chance to get in to the 2.6.25 with
> the core virtualization. By core I mean unix, netlinux, ipv4 and ipv6.

Yeah, if we can push these protocols in time, that will be *very very*
cool :)

When you talk about ipv4/ipv6 do you include tcp/udp ?

> Netfilters virtualization is a complex task :)
>
>> Benjamin and I we began to look
>> at ipv4. This is a big part, perhaps we can split that into several
>> subset and dispatch them, except if Pavel and Denis already rebase ipv4
>> for net-2.6, in this case feel free to send them out.
>
> Well, actually we have almost moved to the net-2.6 with the ipv4
> set.

Excellent !
Did you took the different patches I sent for udplite and multicast ?

If you rebased netns49 to net-2.6 and you plan to keep synced with the
Dave Miller tree, perhaps it is time to switch the git tree.
It can be cool if you can put a git tree at openvz.

> There are only some minor (I hope they are minor ;)) things.

Perhaps, we can help here.

> So we would be glad to go on with ipv4 further. What's up with the
> ipv6 patches, Daniel? You said that you and Benjamin make some
> big progress in this area, no?

Yes, for the moment we reach the addrconf stuff, so we have routing
table, ip6_fib, fib6_rules, ndisc and addrconf per namespaces.
The patches sent by Alexey Dobriyan making /proc/net/ipv6_route to the
seq_file interface has made our life easier.

IMHO this part of ipv6 is the most difficult, the different protocols
relying on it will be much more easy to implement.

We are actually facing two problems:

  * the first one is the locking of the network namespace list by
rtnl_lock, so from the timer callback we can not browse the network
namespace list to check the age of the routes. It is a problem I would
like to talk with Denis if he has time

  * the loopback refcounting is not correctly handled in ipv6. This
protocol do not expect to have the loopback to be unregistered, so there
is some problem with the addr_ifdown function when exiting the network
namespace

Regards.
   -- Daniel

_____



---

Subject: Re: Pid namespaces problems
Posted by Pavel Emelianov on Thu, 08 Nov 2007 10:40:37 GMT
View Forum Message <> Reply to Message

Daniel Lezcano wrote:
> Pavel Emelyanov wrote:
>> Daniel Lezcano wrote:
>>> Cedric Le Goater wrote:
>>>> - There are several architectures with their own signal functions for
>>>>  other OS compatibility that have are using _pid and not _vpid
>>>>  variants of functions. (irix and solaris)
>>>>  arch/mips/kernel/irixsig.c:irix_waitsys
>>>>  arch/mips/kernel/sysirix.c:irix_setpgrp
>>>>  arch/sparc64/solaris/misc.c:solaris_procids
>>> Ok. Looks like your list is the same as mine. That's good to hear

>>>>> that I haven't missed anything important.
>>>> We've also talked about af_unix credentials.
>>>>
>>>>> So, I see that you're about to take a closer look at the pid
>>>>> namespaces. If so, then what time can we expect the net namespace
>>>>> activity to go on? Or (if you don't mind) can we start merging
>>>>> the patches to David as soon as he opens his 2.6.25 merge window?
>>>> I think daniel and benjamin are also getting ready for the 2.6.25
>>>> merge window.
>>> Yes. It can be cool if we can sync up Benjamin, Pavel, Denis, Eric and I
>>> with the different parts to be posted.
>> Yup. Team work will give us a chance to get in to the 2.6.25 with
>> the core virtualization. By core I mean unix, netlinux, ipv4 and ipv6.
>
> Yeah, if we can push these protocols in time, that will be *very very*
> cool :)
>
> When you talk about ipv4/ipv6 do you include tcp/udp ?
>
>> Netfilters virtualization is a complex task :)
>>
>>> Benjamin and I we began to look
>>> at ipv4. This is a big part, perhaps we can split that into several
>>> subset and dispatch them, except if Pavel and Denis already rebase ipv4
>>> for net-2.6, in this case feel free to send them out.
>> Well, actually we have almost moved to the net-2.6 with the ipv4
>> set.
>
> Excellent !
> Did you took the different patches I sent for udplite and multicast ?

Not yet, sorry. But we will look at them as soon as we finish with
the existing netns tree.

> If you rebased netns49 to net-2.6 and you plan to keep synced with the
> Dave Miller tree, perhaps it is time to switch the git tree.
> It can be cool if you can put a git tree at openvz.

Sure, but right now we don't have a git repo with it :( We
just have a series of patches, some of them are fixes, some
not. We are going to re-split them and publish to git. I think
that the next week the git.openvz.org will have that repo.

What about your ipv6? Do you have a git repo with it?

>> There are only some minor (I hope they are minor ;)) things.
>
> Perhaps, we can help here.

Yes, of course. When we publish the git repo we'll try to
provide a TODO list so that everyone can participate.

>> So we would be glad to go on with ipv4 further. What's up with the
>> ipv6 patches, Daniel? You said that you and Benjamin make some
>> big progress in this area, no?
>
> Yes, for the moment we reach the addrconf stuff, so we have routing
> table, ip6_fib, fib6_rules, ndisc and addrconf per namespaces.
> The patches sent by Alexey Dobriyan making /proc/net/ipv6_route to the
> seq_file interface has made our life easier.

:) We plan to make some cleanups in the networking code that
make sense even now, but are useful for namespaces.

> IMHO this part of ipv6 is the most difficult, the different protocols
> relying on it will be much more easy to implement.
>
> We are actually facing two problems:
>
>  * the first one is the locking of the network namespace list by
> rtnl_lock, so from the timer callback we can not browse the network
> namespace list to check the age of the routes. It is a problem I would
> like to talk with Denis if he has time

Sure. I will kick him in case he missed it accidentally :)

>  * the loopback refcounting is not correctly handled in ipv6. This
> protocol do not expect to have the loopback to be unregistered, so there
> is some problem with the addr_ifdown function when exiting the network
> namespace

Hm! This is one of the issues we can send to David right now. Do
you have any patches? If not, I can take care of it if you don't mind.

> Regards.
>  -- Daniel
>

Thanks,
Pavel

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

Subject: Re: Pid namespaces problems
Posted by den on Thu, 08 Nov 2007 10:51:15 GMT
View Forum Message <> Reply to Message

Daniel Lezcano wrote:

> * the first one is the locking of the network namespace list by
> rtnl_lock, so from the timer callback we can not browse the network
> namespace list to check the age of the routes. It is a problem I would
> like to talk with Denis if he has time

>From my point of view, the situation is clear. The timer should be per/namespace. The situation is completely different as one in IPv4.

> * the loopback refcounting is not correctly handled in ipv6. This
> protocol do not expect to have the loopback to be unregistered, so there
> is some problem with the addr_ifdown function when exiting the network
> namespace

I think that default routing targets and similar staff should be dynamically allocated as a start and submitted now. No need to wait NETNS infrastructure. I have sent similar cleanup for fib rules recently. No answer from David yet.

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

Subject: Re: Pid namespaces problems
Posted by Daniel Lezcano on Thu, 08 Nov 2007 11:07:48 GMT
View Forum Message <> Reply to Message

Pavel Emelyanov wrote:
> Daniel Lezcano wrote:
>> Pavel Emelyanov wrote:
>>> Daniel Lezcano wrote:
>>>> Cedric Le Goater wrote:
>>>>> - There are several architectures with their own signal functions for
>>>>> other OS compatibility that have are using _pid and not _vpid
>>>>> variants of functions. (irix and solaris)
>>>>> arch/mips/kernel/irixsig.c:irix_waitsys
>>>>> arch/mips/kernel/sysirix.c:irix_setpgrp
>>>>> arch/sparc64/solaris/misc.c:solaris_procids
>>>> Ok. Looks like your list is the same as mine. That's good to hear
>>>> that I haven't missed anything important.
>>> We've also talked about af_unix credentials.
>>>

>>>>>> So, I see that you're about to take a closer look at the pid
>>>>>> namespaces. If so, then what time can we expect the net namespace
>>>>>> activity to go on? Or (if you don't mind) can we start merging
>>>>>> the patches to David as soon as he opens his 2.6.25 merge window?
>>>>> I think daniel and benjamin are also getting ready for the 2.6.25
>>>>> merge window.
>>>> Yes. It can be cool if we can sync up Benjamin, Pavel, Denis, Eric and I
>>>> with the different parts to be posted.
>>> Yup. Team work will give us a chance to get in to the 2.6.25 with
>>> the core virtualization. By core I mean unix, netlinux, ipv4 and ipv6.
>> Yeah, if we can push these protocols in time, that will be *very very*
>> cool :)
>>
>> When you talk about ipv4/ipv6 do you include tcp/udp ?
>>
>>> Netfilters virtualization is a complex task :)
>>>
>>>> Benjamin and I we began to look
>>>> at ipv4. This is a big part, perhaps we can split that into several
>>>> subset and dispatch them, except if Pavel and Denis already rebase ipv4
>>>> for net-2.6, in this case feel free to send them out.
>>> Well, actually we have almost moved to the net-2.6 with the ipv4
>>> set.
>> Excellent !
>> Did you took the different patches I sent for udplite and multicast ?
>
> Not yet, sorry. But we will look at them as soon as we finish with
> the existing netns tree.
>
>> If you rebased netns49 to net-2.6 and you plan to keep synced with the
>> Dave Miller tree, perhaps it is time to switch the git tree.
>> It can be cool if you can put a git tree at openvz.
>
> Sure, but right now we don't have a git repo with it :( We
> just have a series of patches, some of them are fixes, some
> not. We are going to re-split them and publish to git. I think
> that the next week the git.openvz.org will have that repo.
>
> What about your ipv6? Do you have a git repo with it?

Unfortunatly no, but we have a patchset we can port to your future git
repo and send them to you to be integrated within the git repo.

>>> There are only some minor (I hope they are minor ;)) things.
>> Perhaps, we can help here.
>
> Yes, of course. When we publish the git repo we'll try to
> provide a TODO list so that everyone can participate.

Good idea.

>>> So we would be glad to go on with ipv4 further. What's up with the
>>> ipv6 patches, Daniel? You said that you and Benjamin make some
>>> big progress in this area, no?
>> Yes, for the moment we reach the addrconf stuff, so we have routing
>> table, ip6_fib, fib6_rules, ndisc and addrconf per namespaces.
>> The patches sent by Alexey Dobriyan making /proc/net/ipv6_route to the
>> seq_file interface has made our life easier.
>
> :) We plan to make some cleanups in the networking code that
> make sense even now, but are useful for namespaces.
>
>> IMHO this part of ipv6 is the most difficult, the different protocols
>> relying on it will be much more easy to implement.
>>
>> We are actually facing two problems:
>>
>>  * the first one is the locking of the network namespace list by
>> rtnl_lock, so from the timer callback we can not browse the network
>> namespace list to check the age of the routes. It is a problem I would
>> like to talk with Denis if he has time
>
> Sure. I will kick him in case he missed it accidentally :)
>
>>  * the loopback refcounting is not correctly handled in ipv6. This
>> protocol do not expect to have the loopback to be unregistered, so there
>> is some problem with the addr_ifdown function when exiting the network
>> namespace
>
> Hm! This is one of the issues we can send to David right now. Do
> you have any patches? If not, I can take care of it if you don't mind.

I was working on this problem since yesterday with the patchset for
ipv6. I didn't manage to reproduce it with the initial network
namespace. Benjamin is looking this problem right now. I think he will
be glad to be helped.

I have some suspicions on the loopback unregistering and the notifier
call chain. I think when the initial network namespace is initialized,
the notifier call chain for ipv6 is initialized after the loopback is
registered. When the system goes down, the notifier call chain is
disabled before the loopback is unregistered. So the ipv6 protocol works
well for the init netns and does not receive event for the loopback
register/unregister. But when we create a new netns, a new instance of
the loopback is done and the NETDEV_REGISTER event is raised to ipv6
(notifier call chain are not per namespace), and this protocol is not

expecting such event for the loopback.

I did a quick fix, when we receive a NETDEV_REGISTER/NETDEV_UNREGISTER
event and the device is a loopback, just ignore the event, in the code
of addrconf_notify (NETDEV_DOWN and NETDEV_UNREGISTER must be splited in
the switch). If that makes sense to protect ipv6 from such events, I
think this patch can be sent to Dave Miller.

There some patches to fix that but nothing definitive for multiple
network namespace, we still have a problem when the loopback is up when
exiting the network namespace. These patches are in the hands of Benjamin.

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

---

Subject: Re: Pid namespaces problems
Posted by Daniel Lezcano on Thu, 08 Nov 2007 13:29:03 GMT
View Forum Message <> Reply to Message

Denis V. Lunev wrote:
> Daniel Lezcano wrote:
>
>> * the first one is the locking of the network namespace list by
>> rtnl_lock, so from the timer callback we can not browse the network
>> namespace list to check the age of the routes. It is a problem I would
>> like to talk with Denis if he has time
>
> From my point of view, the situation is clear. The timer should be
> per/namespace. The situation is completely different as one in IPv4.

We thought to make a timer per namespace for ipv6, but we are a little
afraid for the performances when there will be a lot of containers.
Anyway, we can do a timer per namespace and optimize that later. I will
cook a new patch to take into account that for the next week.

>> * the loopback refcounting is not correctly handled in ipv6. This
>> protocol do not expect to have the loopback to be unregistered, so there
>> is some problem with the addr_ifdown function when exiting the network
>> namespace
>
> I think that default routing targets and similar staff should be
> dynamically allocated as a start and submitted now. No need to wait
> NETNS infrastructure. I have sent similar cleanup for fib rules
> recently. No answer from David yet.

Interesting.

Subject: Re: Pid namespaces problems
Posted by Pavel Emelianov on Thu, 08 Nov 2007 13:37:38 GMT
View Forum Message <> Reply to Message

Daniel Lezcano wrote:
> Denis V. Lunev wrote:
>> Daniel Lezcano wrote:
>>
>>>  * the first one is the locking of the network namespace list by
>>> rtnl_lock, so from the timer callback we can not browse the network
>>> namespace list to check the age of the routes. It is a problem I would
>>> like to talk with Denis if he has time
>> From my point of view, the situation is clear. The timer should be
>> per/namespace. The situation is completely different as one in IPv4.
>
> We thought to make a timer per namespace for ipv6, but we are a little
> afraid for the performances when there will be a lot of containers.
> Anyway, we can do a timer per namespace and optimize that later. I will
> cook a new patch to take into account that for the next week.

I propose to start a new mailing thread for net namespaces discussions
or at least change this one's subject ;)

>>>  * the loopback refcounting is not correctly handled in ipv6. This
>>> protocol do not expect to have the loopback to be unregistered, so there
>>> is some problem with the addr_ifdown function when exiting the network
>>> namespace
>> I think that default routing targets and similar staff should be
>> dynamically allocated as a start and submitted now. No need to wait
>> NETNS infrastructure. I have sent similar cleanup for fib rules
>> recently. No answer from David yet.
>
> Interesting.
>

Subject: Re: Pid namespaces problems
Posted by den on Thu, 08 Nov 2007 13:41:52 GMT
View Forum Message <> Reply to Message

Daniel Lezcano wrote:
> Denis V. Lunev wrote:
>> Daniel Lezcano wrote:
>>
>>>  * the first one is the locking of the network namespace list by
>>> rtnl_lock, so from the timer callback we can not browse the network
>>> namespace list to check the age of the routes. It is a problem I would
>>> like to talk with Denis if he has time
>>
>> From my point of view, the situation is clear. The timer should be
>> per/namespace. The situation is completely different as one in IPv4.
>
> We thought to make a timer per namespace for ipv6, but we are a little
> afraid for the performances when there will be a lot of containers.
> Anyway, we can do a timer per namespace and optimize that later. I will
> cook a new patch to take into account that for the next week.

IMHO not a problem. tcp_write_timer is per/socket timer. If this works
efficiently, per/namespace one will work also.

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

Subject: net namespace plans for 2.6.25 (was Re: Pid namespaces problems)
Posted by Cedric Le Goater on Thu, 08 Nov 2007 13:42:25 GMT
View Forum Message <> Reply to Message

Pavel Emelyanov wrote:
> Daniel Lezcano wrote:
>> Denis V. Lunev wrote:
>>> Daniel Lezcano wrote:
>>>
>>>>  * the first one is the locking of the network namespace list by
>>>> rtnl_lock, so from the timer callback we can not browse the network
>>>> namespace list to check the age of the routes. It is a problem I would
>>>> like to talk with Denis if he has time
>>> From my point of view, the situation is clear. The timer should be
>>> per/namespace. The situation is completely different as one in IPv4.
>> We thought to make a timer per namespace for ipv6, but we are a little
>> afraid for the performances when there will be a lot of containers.
>> Anyway, we can do a timer per namespace and optimize that later. I will
>> cook a new patch to take into account that for the next week.

>
> I propose to start a new mailing thread for net namespaces discussions
> or at least change this one's subject ;)

done.

C.



---

Subject: Re: net namespace plans for 2.6.25 (was Re: Pid namespaces problems)
Posted by Daniel Lezcano on Thu, 08 Nov 2007 13:45:33 GMT
View Forum Message <> Reply to Message

Denis V. Lunev wrote:
> Daniel Lezcano wrote:
>> Denis V. Lunev wrote:
>>> Daniel Lezcano wrote:
>>>
>>>>  * the first one is the locking of the network namespace list by
>>>> rtnl_lock, so from the timer callback we can not browse the network
>>>> namespace list to check the age of the routes. It is a problem I would
>>>> like to talk with Denis if he has time
>>> From my point of view, the situation is clear. The timer should be
>>> per/namespace. The situation is completely different as one in IPv4.
>> We thought to make a timer per namespace for ipv6, but we are a little
>> afraid for the performances when there will be a lot of containers.
>> Anyway, we can do a timer per namespace and optimize that later. I will
>> cook a new patch to take into account that for the next week.
>
> IMHO not a problem. tcp_write_timer is per/socket timer. If this works
> efficiently, per/namespace one will work also.

That's right, this is a good argument. By the way, the amount of work to
be done in the tcp_write_timer is perhaps smaller than the one done in
the ipv6 routing age check, no ? Anyway, I'm not against a timer per
namespace in this case, I already did a try before rolling back to a
for_each_net in the gc timer, that changes a little the API, but nothing
we can handle easily.

Subject: Re: net namespace plans for 2.6.25 (was Re: Pid namespaces problems)
Posted by Pavel Emelianov on Thu, 08 Nov 2007 13:58:13 GMT
View Forum Message <> Reply to Message

Daniel Lezcano wrote:
> Denis V. Lunev wrote:
> > Daniel Lezcano wrote:
> >> Denis V. Lunev wrote:
> >>> Daniel Lezcano wrote:
> >>>
> >>>> * the first one is the locking of the network namespace list by
> >>>> rtnl_lock, so from the timer callback we can not browse the network
> >>>> namespace list to check the age of the routes. It is a problem I would
> >>>> like to talk with Denis if he has time
> >>> From my point of view, the situation is clear. The timer should be
> >>> per/namespace. The situation is completely different as one in IPv4.
> >> We thought to make a timer per namespace for ipv6, but we are a little
> >> afraid for the performances when there will be a lot of containers.
> >> Anyway, we can do a timer per namespace and optimize that later. I will
> >> cook a new patch to take into account that for the next week.
> > >
> > IMHO not a problem. tcp_write_timer is per/socket timer. If this works
> > efficiently, per/namespace one will work also.
>
> That's right, this is a good argument. By the way, the amount of work to
> be done in the tcp_write_timer is perhaps smaller than the one done in
> the ipv6 routing age check, no ? Anyway, I'm not against a timer per
> namespace in this case, I already did a try before rolling back to a
> for_each_net in the gc timer, that changes a little the API, but nothing

We can easily make the netns list rcu protected to address this issue.
If you're interested, I can prepare a patch tomorrow.

> we can handle easily.
>
>

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

Subject: Re: net namespace plans for 2.6.25 (was Re: Pid namespaces problems)
Posted by Daniel Lezcano on Thu, 08 Nov 2007 14:00:03 GMT
View Forum Message <> Reply to Message

Denis V. Lunev wrote:

> Daniel Lezcano wrote:
>> Denis V. Lunev wrote:
>>> Daniel Lezcano wrote:
>>>> Denis V. Lunev wrote:
>>>>> Daniel Lezcano wrote:
>>>>>
>>>>>>  * the first one is the locking of the network namespace list by
>>>>>> rtnl_lock, so from the timer callback we can not browse the network
>>>>>> namespace list to check the age of the routes. It is a problem I would
>>>>>> like to talk with Denis if he has time
>>>>> From my point of view, the situation is clear. The timer should be
>>>>> per/namespace. The situation is completely different as one in IPv4.
>>>> We thought to make a timer per namespace for ipv6, but we are a little
>>>> afraid for the performances when there will be a lot of containers.
>>>> Anyway, we can do a timer per namespace and optimize that later. I will
>>>> cook a new patch to take into account that for the next week.
>>> IMHO not a problem. tcp_write_timer is per/socket timer. If this works
>>> efficiently, per/namespace one will work also.
>> That's right, this is a good argument. By the way, the amount of work to
>> be done in the tcp_write_timer is perhaps smaller than the one done in
>> the ipv6 routing age check, no ? Anyway, I'm not against a timer per
>> namespace in this case, I already did a try before rolling back to a
>> for_each_net in the gc timer, that changes a little the API, but nothing
>> we can handle easily.
>>
>>
> I think you are wrong. The amount of work to "purge" all namespaces is a
> constant in the IPv6 case, where we'll have per/namespace cache. So, for
> a multiple timers model only multiple timer overhead counts and this
> overhead is small, as timer list is efficient.
>
> This argument does not for for IPv4 case, where there is a one big cache
> for all namespaces.

Interesting, thanks for the precision.



---

Subject: Re: net namespace plans for 2.6.25 (was Re: Pid namespaces problems)
Posted by den on Thu, 08 Nov 2007 14:04:41 GMT
View Forum Message <> Reply to Message

Daniel Lezcano wrote:
> Denis V. Lunev wrote:
>> Daniel Lezcano wrote:

>>> Denis V. Lunev wrote:
>>>> Daniel Lezcano wrote:
>>>>
>>>>>  * the first one is the locking of the network namespace list by
>>>>> rtnl_lock, so from the timer callback we can not browse the network
>>>>> namespace list to check the age of the routes. It is a problem I would
>>>>> like to talk with Denis if he has time
>>>> From my point of view, the situation is clear. The timer should be
>>>> per/namespace. The situation is completely different as one in IPv4.
>>> We thought to make a timer per namespace for ipv6, but we are a little
>>> afraid for the performances when there will be a lot of containers.
>>> Anyway, we can do a timer per namespace and optimize that later. I will
>>> cook a new patch to take into account that for the next week.
>>
>> IMHO not a problem. tcp_write_timer is per/socket timer. If this works
>> efficiently, per/namespace one will work also.
>
> That's right, this is a good argument. By the way, the amount of work to
> be done in the tcp_write_timer is perhaps smaller than the one done in
> the ipv6 routing age check, no ? Anyway, I'm not against a timer per
> namespace in this case, I already did a try before rolling back to a
> for_each_net in the gc timer, that changes a little the API, but nothing
> we can handle easily.
>
>
I think you are wrong. The amount of work to "purge" all namespaces is a
constant in the IPv6 case, where we'll have per/namespace cache. So, for
a multiple timers model only multiple timer overhead counts and this
overhead is small, as timer list is efficient.

This argument does not for for IPv4 case, where there is a one big cache
for all namespaces.

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

---

Subject: Re: net namespace plans for 2.6.25 (was Re: Pid namespaces problems)
Posted by Benjamin Thery on Thu, 08 Nov 2007 14:08:56 GMT
View Forum Message <> Reply to Message

Pavel Emelyanov wrote:
> Daniel Lezcano wrote:
>> Denis V. Lunev wrote:
>>  > Daniel Lezcano wrote:
>>  >> Denis V. Lunev wrote:
>>  >>> Daniel Lezcano wrote:

>> >>>
>> >>>> * the first one is the locking of the network namespace list by
>> >>>> rtnl_lock, so from the timer callback we can not browse the network
>> >>>> namespace list to check the age of the routes. It is a problem I would
>> >>>> like to talk with Denis if he has time
>> >>> From my point of view, the situation is clear. The timer should be
>> >>> per/namespace. The situation is completely different as one in IPv4.
>> >> We thought to make a timer per namespace for ipv6, but we are a little
>> >> afraid for the performances when there will be a lot of containers.
>> >> Anyway, we can do a timer per namespace and optimize that later. I will
>> >> cook a new patch to take into account that for the next week.
>> >
>> > IMHO not a problem. tcp_write_timer is per/socket timer. If this works
>> > efficiently, per/namespace one will work also.
>>
>> That's right, this is a good argument. By the way, the amount of work to
>> be done in the tcp_write_timer is perhaps smaller than the one done in
>> the ipv6 routing age check, no ? Anyway, I'm not against a timer per
>> namespace in this case, I already did a try before rolling back to a
>> for_each_net in the gc timer, that changes a little the API, but nothing
>
> We can easily make the netns list rcu protected to address this issue.
> If you're interested, I can prepare a patch tomorrow.

That would be great if you manage do it.
This was our initial idea with Daniel, but as I have a limited
knowledge of RCU, I didn't manage to obtain an acceptable patch.
One of the more problematic area is rtnl_unlock().

Benjamin

>
>> we can handle easily.
>>
>>
>
>

--
B e n j a m i n  T h e r y - BULL/DT/Open Software R&D

http://www.bull.com

Subject: Re: net namespace plans for 2.6.25 (was Re: Pid namespaces problems)
Posted by Daniel Lezcano on Thu, 08 Nov 2007 14:09:36 GMT
View Forum Message <> Reply to Message

Pavel Emelyanov wrote:
> Daniel Lezcano wrote:
>> Denis V. Lunev wrote:
>>  > Daniel Lezcano wrote:
>>  >> Denis V. Lunev wrote:
>>  >>> Daniel Lezcano wrote:
>>  >>>
>>  >>>>  * the first one is the locking of the network namespace list by
>>  >>>> rtnl_lock, so from the timer callback we can not browse the network
>>  >>>> namespace list to check the age of the routes. It is a problem I would
>>  >>>> like to talk with Denis if he has time
>>  >>> From my point of view, the situation is clear. The timer should be
>>  >>> per/namespace. The situation is completely different as one in IPv4.
>>  >> We thought to make a timer per namespace for ipv6, but we are a little
>>  >> afraid for the performances when there will be a lot of containers.
>>  >> Anyway, we can do a timer per namespace and optimize that later. I will
>>  >> cook a new patch to take into account that for the next week.
>>  >
>>  > IMHO not a problem. tcp_write_timer is per/socket timer. If this works
>>  > efficiently, per/namespace one will work also.
>>
>> That's right, this is a good argument. By the way, the amount of work to
>> be done in the tcp_write_timer is perhaps smaller than the one done in
>> the ipv6 routing age check, no ? Anyway, I'm not against a timer per
>> namespace in this case, I already did a try before rolling back to a
>> for_each_net in the gc timer, that changes a little the API, but nothing
>
> We can easily make the netns list rcu protected to address this issue.
> If you're interested, I can prepare a patch tomorrow.

Sure, I'm interested :)

Benjamin and I, we thought about using a rcu to avoid to use a timer per namespace in ipv6 but we faced to the problem with rtnl_unlock function when the network namespace is protected with the rtnl_lock/rtnl_unlock. In the function rtnl_unlock (not the one in net-2.6 but the one which is in netns49), there is loop, for_each_net, in this loop, we do rtnl_unlock, call sk_data_ready and take the lock again. If we are in rcu protected model, this loop will take a lock (one time just before sk_data_ready and one time in the sk_data_ready function). As far as I understand with rcu, we should not block inside a rcu_read_lock, right ?

_____
Containers mailing list

Subject: Re: net namespace plans for 2.6.25 (was Re: Pid namespaces problems)
Posted by Pavel Emelianov on Fri, 09 Nov 2007 10:14:32 GMT
View Forum Message <> Reply to Message

Daniel Lezcano wrote:
> Pavel Emelyanov wrote:
>> Daniel Lezcano wrote:
>>> Denis V. Lunev wrote:
>>>  > Daniel Lezcano wrote:
>>>  >> Denis V. Lunev wrote:
>>>  >>> Daniel Lezcano wrote:
>>>  >>>
>>>  >>>>  * the first one is the locking of the network namespace list by
>>>  >>>> rtnl_lock, so from the timer callback we can not browse the network
>>>  >>>> namespace list to check the age of the routes. It is a problem I would
>>>  >>>> like to talk with Denis if he has time
>>>  >>> From my point of view, the situation is clear. The timer should be
>>>  >>> per/namespace. The situation is completely different as one in IPv4.
>>>  >> We thought to make a timer per namespace for ipv6, but we are a little
>>>  >> afraid for the performances when there will be a lot of containers.
>>>  >> Anyway, we can do a timer per namespace and optimize that later. I will
>>>  >> cook a new patch to take into account that for the next week.
>>>  >
>>>  > IMHO not a problem. tcp_write_timer is per/socket timer. If this works
>>>  > efficiently, per/namespace one will work also.
>>>
>>> That's right, this is a good argument. By the way, the amount of work to
>>> be done in the tcp_write_timer is perhaps smaller than the one done in
>>> the ipv6 routing age check, no ? Anyway, I'm not against a timer per
>>> namespace in this case, I already did a try before rolling back to a
>>> for_each_net in the gc timer, that changes a little the API, but nothing
>> We can easily make the netns list rcu protected to address this issue.
>> If you're interested, I can prepare a patch tomorrow.
>
> Sure, I'm interested :)
>
> Benjamin and I, we thought about using a rcu to avoid to use a timer per
> namespace in ipv6 but we faced to the problem with rtnl_unlock function
> when the network namespace is protected with the rtnl_lock/rtnl_unlock.
> In the function rtnl_unlock (not the one in net-2.6 but the one which is
> in netns49), there is loop, for_each_net, in this loop, we do
> rtnl_unlock, call sk_data_ready and take the lock again. If we are in
> rcu protected model, this loop will take a lock (one time just before
> sk_data_ready and one time in the sk_data_ready function). As far as I

> understand with rcu, we should not block inside a rcu_read_lock, right ?

Right. I will look at it. I think that if we protect the list with RCU
the rtnl_lock() protection will be not needed any longer.

Thanks,
Pavel

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