Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by dev on Thu, 07 Sep 2006 16:01:46 GMT
View Forum Message <> Reply to Message

BTW...

```
> --- 2.6.18-rc4-mm3.orig/include/linux/sched.h
> +++ 2.6.18-rc4-mm3/include/linux/sched.h
> @@ -26,6 +26,7 @@
>  #define CLONE_STOPPED  0x02000000 /* Start in stopped state */
>  #define CLONE_NEWUTS  0x04000000 /* New utsname group? */
>  #define CLONE_NEWIPC  0x08000000 /* New ipcs */
> +#define CLONE_NEWUSER  0x10000000 /* New user */
```

we have place for 3 namespaces more only.
Does anyone have a plan what to do then?
I warned about this at the beginning when we were discussing the interfaces
and this flags soon going to be exhausted, so probably it is time to
do something in advance...

Thanks,
Kirill

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by Herbert Poetzl on Thu, 07 Sep 2006 17:15:08 GMT
View Forum Message <> Reply to Message

On Thu, Sep 07, 2006 at 08:05:30PM +0400, Kirill Korotaev wrote:

```
> BTW...
>
> > --- 2.6.18-rc4-mm3.orig/include/linux/sched.h
> > +++ 2.6.18-rc4-mm3/include/linux/sched.h
> > @@ -26,6 +26,7 @@
> >  #define CLONE_STOPPED  0x02000000 /* Start in stopped state */
> >  #define CLONE_NEWUTS  0x04000000 /* New utsname group? */
> >  #define CLONE_NEWIPC  0x08000000 /* New ipcs */
> > +#define CLONE_NEWUSER  0x10000000 /* New user */
> we have place for 3 namespaces more only.
> Does anyone have a plan what to do then?
```

what about having a new clone syscall with 32 or
better 64 bits reserved for namespace stuff, and
only put basic/generic namespaces or even aggregate
flags into the existing clone interface?

something like: uts+ipc+user -> CLONE_NEWXYZ
but CLONE2_NEWUTS, CLONE2_NEWIPC, CLONE2_NEWUSER

best,
Herbert

PS: what happened to the idea of forwarding
this whole stuff to _both_ mailing lists?
as far as I can tell we are adding those lists
every now and then to the CC, could that be
addressed soon?

> I warned about this at the beginning when we were discussing the
> interfaces and this flags soon going to be exhausted, so probably it
> is time to do something in advance...
>
> Thanks,
> Kirill
> ____
> Containers mailing list
> Containers@lists.osdl.org
> https://lists.osdl.org/mailman/listinfo/containers

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by serue on Thu, 07 Sep 2006 17:29:05 GMT

View Forum Message <> Reply to Message

Quoting Herbert Poetzl (herbert@13thfloor.at):
> On Thu, Sep 07, 2006 at 08:05:30PM +0400, Kirill Korotaev wrote:
> > BTW...
> >
> > > --- 2.6.18-rc4-mm3.orig/include/linux/sched.h
> > > +++ 2.6.18-rc4-mm3/include/linux/sched.h
> > > @@ -26,6 +26,7 @@
> > > #define CLONE_STOPPED 0x02000000 /* Start in stopped state */
> > > #define CLONE_NEWUTS 0x04000000 /* New utsname group? */
> > > #define CLONE_NEWIPC 0x08000000 /* New ipcs */
> > > +#define CLONE_NEWUSER 0x10000000 /* New user */
> > we have place for 3 namespaces more only.
> > Does anyone have a plan what to do then?
>
> what about having a new clone syscall with 32 or
> better 64 bits reserved for namespace stuff, and
> only put basic/generic namespaces or even aggregate
> flags into the existing clone interface?
>
> something like: uts+ipc+user -> CLONE_NEWXYZ
> but CLONE2_NEWUTS, CLONE2_NEWIPC, CLONE2_NEWUSER
>
> best,

> Herbert
>
> PS: what happened to the idea of forwarding
> this whole stuff to _both_ mailing lists?
> as far as I can tell we are adding those lists
> every now and then to the CC, could that be
> addressed soon?

My understanding is that both openvz and vserver should
be subscribed to the containers list.

Is that not the case?

-serge

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by Herbert Poetzl on Thu, 07 Sep 2006 17:57:04 GMT

View Forum Message <> Reply to Message

On Thu, Sep 07, 2006 at 12:29:05PM -0500, Serge E. Hallyn wrote:
> Quoting Herbert Poetzl (herbert@13thfloor.at):
> > On Thu, Sep 07, 2006 at 08:05:30PM +0400, Kirill Korotaev wrote:
> > > BTW...
> > >
> > > > --- 2.6.18-rc4-mm3.orig/include/linux/sched.h
> > > > +++ 2.6.18-rc4-mm3/include/linux/sched.h
> > > > @@ -26,6 +26,7 @@
> > > > #define CLONE_STOPPED 0x02000000 /* Start in stopped state */
> > > > #define CLONE_NEWUTS 0x04000000 /* New utsname group? */
> > > > #define CLONE_NEWIPC 0x08000000 /* New ipcs */
> > > > +#define CLONE_NEWUSER 0x10000000 /* New user */
> > > we have place for 3 namespaces more only.
> > > Does anyone have a plan what to do then?
> >
> > what about having a new clone syscall with 32 or
> > better 64 bits reserved for namespace stuff, and
> > only put basic/generic namespaces or even aggregate
> > flags into the existing clone interface?
> >
> > something like: uts+ipc+user -> CLONE_NEWXYZ
> > but CLONE2_NEWUTS, CLONE2_NEWIPC, CLONE2_NEWUSER
> >
> > best,

> > Herbert
> >
> > PS: what happened to the idea of forwarding
> > this whole stuff to _both_ mailing lists?
> > as far as I can tell we are adding those lists
> > every now and then to the CC, could that be
> > addressed soon?
>
> My understanding is that both openvz and vserver should
> be subscribed to the containers list.
>
> Is that not the case?

not that I would know of .. but maybe our list is
misconfigured, you'll never know, please double
check on your side too

TIA,
Herbert

> -serge
____
Containers mailing list
Containers@lists.osdl.org
https://lists.osdl.org/mailman/listinfo/containers

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by ebiederm on Thu, 07 Sep 2006 20:01:00 GMT

View Forum Message <> Reply to Message

Kirill Korotaev <dev@sw.ru> writes:

> BTW...
>
>> --- 2.6.18-rc4-mm3.orig/include/linux/sched.h
>> +++ 2.6.18-rc4-mm3/include/linux/sched.h
>> @@ -26,6 +26,7 @@
>> #define CLONE_STOPPED 0x02000000 /* Start in stopped state */
>> #define CLONE_NEWUTS 0x04000000 /* New utsname group? */
>> #define CLONE_NEWIPC 0x08000000 /* New ipcs */
>> +#define CLONE_NEWUSER 0x10000000 /* New user */
> we have place for 3 namespaces more only.
> Does anyone have a plan what to do then?
> I warned about this at the beginning when we were discussing the interfaces
> and this flags soon going to be exhausted, so probably it is time to
> do something in advance...

Actually there is another unused bit in the middle :)
Plus there are a bunch of bits that unshare can use but clone can't.
Plus what other namespaces are on the todo list?
We have network, and pid, and time.
What else?

Eric

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by Herbert Poetzl on Fri, 08 Sep 2006 05:57:53 GMT

View Forum Message <> Reply to Message

On Thu, Sep 07, 2006 at 02:01:00PM -0600, Eric W. Biederman wrote:
> Kirill Korotaev <dev@sw.ru> writes:
>
> > BTW...
> >
> >> --- 2.6.18-rc4-mm3.orig/include/linux/sched.h
> >> +++ 2.6.18-rc4-mm3/include/linux/sched.h
> >> @@ -26,6 +26,7 @@
> >> #define CLONE_STOPPED 0x02000000 /* Start in stopped state */
> >> #define CLONE_NEWUTS 0x04000000 /* New utsname group? */
> >> #define CLONE_NEWIPC 0x08000000 /* New ipcs */
> >> +#define CLONE_NEWUSER 0x10000000 /* New user */
> > we have place for 3 namespaces more only.
> > Does anyone have a plan what to do then?
> > I warned about this at the beginning when we were discussing the interfaces
> > and this flags soon going to be exhausted, so probably it is time to
> > do something in advance...
>
> Actually there is another unused bit in the middle :)
> Plus there are a bunch of bits that unshare can use but clone can't.
> Plus what other namespaces are on the todo list?
> We have network, and pid, and time.
> What else?

resource (could be limits and/or accounting),
lightweight-net, (maybe fs in contrast to vfs)

best,
Herbert

> Eric
> ____
> Containers mailing list
> Containers@lists.osdl.org
> https://lists.osdl.org/mailman/listinfo/containers

Containers mailing list
Containers@lists.osdl.org
https://lists.osdl.org/mailman/listinfo/containers

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by dev on Fri, 08 Sep 2006 15:45:02 GMT
View Forum Message <> Reply to Message

> On Thu, Sep 07, 2006 at 08:05:30PM +0400, Kirill Korotaev wrote:
>
>>BTW...
>>
>>
>>>--- 2.6.18-rc4-mm3.orig/include/linux/sched.h
>>>+++ 2.6.18-rc4-mm3/include/linux/sched.h
>>>@@ -26,6 +26,7 @@
>>> #define CLONE_STOPPED 0x02000000 /* Start in stopped state */
>>> #define CLONE_NEWUTS 0x04000000 /* New utsname group? */
>>> #define CLONE_NEWIPC 0x08000000 /* New ipcs */
>>>+#define CLONE_NEWUSER 0x10000000 /* New user */
>>
>>we have place for 3 namespaces more only.
>>Does anyone have a plan what to do then?
>
>
> what about having a new clone syscall with 32 or
> better 64 bits reserved for namespace stuff, and
> only put basic/generic namespaces or even aggregate
> flags into the existing clone interface?
>
> something like: uts+ipc+user -> CLONE_NEWXYZ
> but CLONE2_NEWUTS, CLONE2_NEWIPC, CLONE2_NEWUSER
I would suggest to do it another way then:
remove CLONES_NEWXXXNS from clone() at all (except for MNT NS for compatibility)
and introduce sys_clone_ns() with totatally new 64bit flags like
CLONE_NS_UTS
CLONE_NS_IPC
CLONE_NS_USER
CLONE_NS_NET
etc.

Thanks,
Kirill

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by Cedric Le Goater on Mon, 11 Sep 2006 08:46:52 GMT
View Forum Message <> Reply to Message

Kirill Korotaev wrote:
>> On Thu, Sep 07, 2006 at 08:05:30PM +0400, Kirill Korotaev wrote:
>>
>>> BTW...
>>>
>>>
>>>> --- 2.6.18-rc4-mm3.orig/include/linux/sched.h
>>>> +++ 2.6.18-rc4-mm3/include/linux/sched.h
>>>> @@ -26,6 +26,7 @@
>>>> #define CLONE_STOPPED 0x02000000 /* Start in stopped state */
>>>> #define CLONE_NEWUTS 0x04000000 /* New utsname group? */
>>>> #define CLONE_NEWIPC 0x08000000 /* New ipcs */
>>>> +#define CLONE_NEWUSER 0x10000000 /* New user */
>>> we have place for 3 namespaces more only.
>>> Does anyone have a plan what to do then?
>>
>> what about having a new clone syscall with 32 or
>> better 64 bits reserved for namespace stuff, and
>> only put basic/generic namespaces or even aggregate
>> flags into the existing clone interface?
>>
>> something like: uts+ipc+user -> CLONE_NEWXYZ
>> but CLONE2_NEWUTS, CLONE2_NEWIPC, CLONE2_NEWUSER
> I would suggest to do it another way then:
> remove CLONES_NEWXXXNS from clone() at all (except for MNT NS for compatibility)
> and introduce sys_clone_ns() with totatally new 64bit flags like
> CLONE_NS_UTS
> CLONE_NS_IPC
> CLONE_NS_USER
> CLONE_NS_NET

yep. I like the idea of a specific syscall. It would certainly help us to
handle some corner cases in the namespaces.

OTOH, the unshare/clone semantic is right in most cases.

How would the community feel about this ? would they say "fix
unshare/clone" or this is a new API, move it somewhere else ?

thanks,

C.

____
Containers mailing list
Containers@lists.osdl.org

https://lists.osdl.org/mailman/listinfo/containers

Subject: Re: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by Cedric Le Goater on Mon, 11 Sep 2006 08:59:04 GMT
View Forum Message <> Reply to Message

Herbert Poetzl wrote:
> On Thu, Sep 07, 2006 at 02:01:00PM -0600, Eric W. Biederman wrote:
>> Kirill Korotaev <dev@sw.ru> writes:
>>
>>> BTW...
>>>
>>>> --- 2.6.18-rc4-mm3.orig/include/linux/sched.h
>>>> +++ 2.6.18-rc4-mm3/include/linux/sched.h
>>>> @@ -26,6 +26,7 @@
>>>> #define CLONE_STOPPED 0x02000000 /* Start in stopped state */
>>>> #define CLONE_NEWUTS 0x04000000 /* New utsname group? */
>>>> #define CLONE_NEWIPC 0x08000000 /* New ipcs */
>>>> +#define CLONE_NEWUSER 0x10000000 /* New user */
>>> we have place for 3 namespaces more only.
>>> Does anyone have a plan what to do then?
>>> I warned about this at the beginning when we were discussing the interfaces
>>> and this flags soon going to be exhausted, so probably it is time to
>>> do something in advance...
>> Actually there is another unused bit in the middle :)
>> Plus there are a bunch of bits that unshare can use but clone can't.
>> Plus what other namespaces are on the todo list?
>> We have network, and pid, and time.
>> What else?
>
> resource (could be limits and/or accounting),
> lightweight-net, (maybe fs in contrast to vfs)

I guess we're reaching the limits anyway and it would not leave much room
in the clone flags for other features not related to containers.

It's not like we're adding one or two, we would take at least 6 : uts, ipc,
user, pid, net, time, etc. I'm sure ideas to extend the list will come when
this is in use ...

C.

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by ebiederm on Mon, 11 Sep 2006 11:16:51 GMT
View Forum Message <> Reply to Message

Cedric Le Goater <clg@fr.ibm.com> writes:

> Herbert Poetzl wrote:
>>
>> resource (could be limits and/or accounting),
>> lightweight-net, (maybe fs in contrast to vfs)
>
> I guess we're reaching the limits anyway and it would not leave much room
> in the clone flags for other features not related to containers.
>
> It's not like we're adding one or two, we would take at least 6 : uts, ipc,
> user, pid, net, time, etc. I'm sure ideas to extend the list will come when
> this is in use ...

I think the resource is possibly real, as at least ubc introduces
a new set of global names, and yet another global namespace sucks.
Something I now need to challenge the implementors on.

If we do a lightweight net I don't think it will be a namespace.
Because isolation does needs separate names, just some sort of filtering
mechanism.

I think being tight here is in some sense a virtue, as it forces
us to think very carefully about adding yet another namespace :)

Eric

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by serue on Mon, 11 Sep 2006 15:59:44 GMT

View Forum Message <> Reply to Message

Quoting Herbert Poetzl (herbert@13thfloor.at):
> On Thu, Sep 07, 2006 at 12:29:05PM -0500, Serge E. Hallyn wrote:
> > Quoting Herbert Poetzl (herbert@13thfloor.at):
> > > On Thu, Sep 07, 2006 at 08:05:30PM +0400, Kirill Korotaev wrote:
> > > > BTW...
> > > >
```
> > > > > --- 2.6.18-rc4-mm3.orig/include/linux/sched.h
> > > > > +++ 2.6.18-rc4-mm3/include/linux/sched.h
> > > > > @@ -26,6 +26,7 @@
> > > > >  #define CLONE_STOPPED  0x02000000 /* Start in stopped state */
> > > > >  #define CLONE_NEWUTS  0x04000000 /* New utsname group? */
> > > > >  #define CLONE_NEWIPC  0x08000000 /* New ipcs */
> > > > > +#define CLONE_NEWUSER  0x10000000 /* New user */
```
> > > > we have place for 3 namespaces more only.
> > > > Does anyone have a plan what to do then?
> > >

> > > what about having a new clone syscall with 32 or
> > > better 64 bits reserved for namespace stuff, and
> > > only put basic/generic namespaces or even aggregate
> > > flags into the existing clone interface?
> > >
> > > something like: uts+ipc+user -> CLONE_NEWXYZ
> > > but CLONE2_NEWUTS, CLONE2_NEWIPC, CLONE2_NEWUSER
> > >
> > > best,
> > > Herbert
> > >
> > > PS: what happened to the idea of forwarding
> > > this whole stuff to _both_ mailing lists?
> > > as far as I can tell we are adding those lists
> > > every now and then to the CC, could that be
> > > addressed soon?
> >
> > My understanding is that both openvz and vserver should
> > be subscribed to the containers list.
> >
> > Is that not the case?
>
> not that I would know of .. but maybe our list is
> misconfigured, you'll never know, please double
> check on your side too

The vserver list is in fact subscribed. Do you have some logs you can
check to see whether container@lists.osdl.org messages are being dropped
for some reason?

thanks,
-serge
____
Containers mailing list
Containers@lists.osdl.org
https://lists.osdl.org/mailman/listinfo/containers

Subject: Re: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by Herbert Poetzl on Tue, 12 Sep 2006 10:54:51 GMT

View Forum Message <> Reply to Message

On Mon, Sep 11, 2006 at 10:59:04AM +0200, Cedric Le Goater wrote:
> Herbert Poetzl wrote:
> > On Thu, Sep 07, 2006 at 02:01:00PM -0600, Eric W. Biederman wrote:
> >> Kirill Korotaev <dev@sw.ru> writes:
> >>
> >>> BTW...

> >>>
> >>>> --- 2.6.18-rc4-mm3.orig/include/linux/sched.h
> >>>> +++ 2.6.18-rc4-mm3/include/linux/sched.h
> >>>> @@ -26,6 +26,7 @@
> >>>> #define CLONE_STOPPED 0x02000000 /* Start in stopped state */
> >>>> #define CLONE_NEWUTS 0x04000000 /* New utsname group? */
> >>>> #define CLONE_NEWIPC 0x08000000 /* New ipcs */
> >>>> +#define CLONE_NEWUSER 0x10000000 /* New user */

> >>> we have place for 3 namespaces more only. Does anyone have a plan
> >>> what to do then? I warned about this at the beginning when we
> >>> were discussing the interfaces and this flags soon going to be
> >>> exhausted, so probably it is time to do something in advance...

> >> Actually there is another unused bit in the middle :)
> >> Plus there are a bunch of bits that unshare can use but clone can't.
> >> Plus what other namespaces are on the todo list?
> >> We have network, and pid, and time.
> >> What else?
> >
> > resource (could be limits and/or accounting),
> > lightweight-net, (maybe fs in contrast to vfs)
>
> I guess we're reaching the limits anyway and it would not leave much
> room in the clone flags for other features not related to containers.
>
> It's not like we're adding one or two, we would take at least 6 : uts,
> ipc, user, pid, net, time, etc. I'm sure ideas to extend the list will
> come when this is in use ...

as I said, I'd opt for having a new clone() syscall in
addition to the existing one, with a separate 64bit
set of flags to decide what namespaces should be created
or cloned. there is no problem with putting 'important'
or generally 'useful' flags (like for example for pid,
uts or lightweight network isolation) into the existing
clone call (will require a simple mapping if done properly)
so that they can be used with 'older' libc interfaces too

I know, it would be 'nice' to keep the existing clone()
interface, but I think it already has become a complication
we should avoid (and we have not even used up all the
available flags :)

are there any strong arguments against having a new
clone() syscall, which I was missing so far?

TIA,

Herbert

> C.
> ____
> Containers mailing list
> Containers@lists.osdl.org
> https://lists.osdl.org/mailman/listinfo/containers

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by dev on Tue, 12 Sep 2006 13:53:50 GMT

View Forum Message <> Reply to Message

Eric W. Biederman wrote:
> Kirill Korotaev <dev@sw.ru> writes:
>
>
>>BTW...
>>
>>
>>>--- 2.6.18-rc4-mm3.orig/include/linux/sched.h
>>>+++ 2.6.18-rc4-mm3/include/linux/sched.h
>>>@@ -26,6 +26,7 @@
>>>#define CLONE_STOPPED 0x02000000 /* Start in stopped state */
>>> #define CLONE_NEWUTS 0x04000000 /* New utsname group? */
>>> #define CLONE_NEWIPC 0x08000000 /* New ipcs */
>>>+#define CLONE_NEWUSER 0x10000000 /* New user */
>>
>>we have place for 3 namespaces more only.
>>Does anyone have a plan what to do then?
>>I warned about this at the beginning when we were discussing the interfaces
>>and this flags soon going to be exhausted, so probably it is time to
>>do something in advance...
>
>
> Actually there is another unused bit in the middle :)
> Plus there are a bunch of bits that unshare can use but clone can't.
:))) I suggest to write HOWTO-select-unused-bits in CodingStyle :))

> Plus what other namespaces are on the todo list?
> We have network, and pid, and time.
I think more.

proc-ns,
sysfs-ns,
printk-ns or syslog-ns?: syslog should be virtualized
and more...

semi-namespaces:
fs-ns (should regulate which filesystems are accessiable from container, but
probably this is not exact name space... need to think over...),
dev-ns (should regulate which devices are accessiable from container)

Thanks,
Kirill

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by ebiederm on Tue, 12 Sep 2006 15:06:28 GMT
View Forum Message <> Reply to Message

Kirill Korotaev <dev@sw.ru> writes:

> Eric W. Biederman wrote:
>> Kirill Korotaev <dev@sw.ru> writes:
>>
>>
>>>BTW...
>>>
>>>
>>>>--- 2.6.18-rc4-mm3.orig/include/linux/sched.h
>>>>+++ 2.6.18-rc4-mm3/include/linux/sched.h
>>>>@@ -26,6 +26,7 @@
>>>>#define CLONE_STOPPED 0x02000000 /* Start in stopped state */
>>>> #define CLONE_NEWUTS 0x04000000 /* New utsname group? */
>>>> #define CLONE_NEWIPC 0x08000000 /* New ipcs */
>>>>+#define CLONE_NEWUSER 0x10000000 /* New user */
>>>
>>>we have place for 3 namespaces more only.
>>>Does anyone have a plan what to do then?
>>>I warned about this at the beginning when we were discussing the interfaces
>>>and this flags soon going to be exhausted, so probably it is time to
>>>do something in advance...
>>
>>
>> Actually there is another unused bit in the middle :)
>> Plus there are a bunch of bits that unshare can use but clone can't.
> :))) I suggest to write HOWTO-select-unused-bits in CodingStyle :))
>
>> Plus what other namespaces are on the todo list?
>> We have network, and pid, and time.
> I think more.
>
> proc-ns,
> sysfs-ns,
> printk-ns or syslog-ns?: syslog should be virtualized

> and more...

I don't think those meet the criteria for namespaces.
But certainly there is work we need to do there.

> semi-namespaces:
> fs-ns (should regulate which filesystems are accessiable from container, but
> probably this is not exact name space... need to think over...),

I think the problem there is the same as allowing untrusted users the ability
to mount filesystems, in which case we just tag filesystems that are safe
for untrusted users to use.

> dev-ns (should regulate which devices are accessiable from container)

Yes. Devices certainly have global names that we need to bring under
control. The easy solution is just to limit CAP_SYS_MKNOD but we
may need something more.

One of the pieces that needs consideration when it comes to permissions
is the plan9 style of permission control.   Where file have an initial
owner, and if someone else needs access to them you chmod, chown them
so that everyone who needs to has access. I think that is an simpler
model to get right than to have a bunch of special cases.

Eric

Subject: Re: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by Cedric Le Goater on Tue, 12 Sep 2006 15:16:47 GMT

View Forum Message <> Reply to Message

Herbert Poetzl wrote:

[...]

> as I said, I'd opt for having a new clone() syscall in
> addition to the existing one, with a separate 64bit
> set of flags to decide what namespaces should be created
> or cloned. there is no problem with putting 'important'
> or generally 'useful' flags (like for example for pid,
> uts or lightweight network isolation) into the existing
> clone call (will require a simple mapping if done properly)
> so that they can be used with 'older' libc interfaces too
>
> I know, it would be 'nice' to keep the existing clone()
> interface, but I think it already has become a complication
> we should avoid (and we have not even used up all the

> available flags :)

agree and so does Kirill.

> are there any strong arguments against having a new
> clone() syscall, which I was missing so far?

I don't see any.

I'm going to revive execns() syscall into a clone_ns() syscall as suggested
by Kirill and you. Then, others will be free to nack ;)

Thanks,

C.

Subject: Re: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by ebiederm on Tue, 12 Sep 2006 15:44:51 GMT

View Forum Message <> Reply to Message

Cedric Le Goater <clg@fr.ibm.com> writes:

> Herbert Poetzl wrote:
>
> [...]
>
>> as I said, I'd opt for having a new clone() syscall in
>> addition to the existing one, with a separate 64bit
>> set of flags to decide what namespaces should be created
>> or cloned. there is no problem with putting 'important'
>> or generally 'useful' flags (like for example for pid,
>> uts or lightweight network isolation) into the existing
>> clone call (will require a simple mapping if done properly)
>> so that they can be used with 'older' libc interfaces too
>>
>> I know, it would be 'nice' to keep the existing clone()
>> interface, but I think it already has become a complication
>> we should avoid (and we have not even used up all the
>> available flags :)
>
> agree and so does Kirill.
>
>> are there any strong arguments against having a new
>> clone() syscall, which I was missing so far?
>
> I don't see any.
>

> I'm going to revive execns() syscall into a clone_ns() syscall as suggested
> by Kirill and you. Then, others will be free to nack ;)

I think it is silly, but I see not real problems with the idea.

Eric

Subject: Re: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by Cedric Le Goater on Tue, 12 Sep 2006 16:07:03 GMT
View Forum Message <> Reply to Message

Eric W. Biederman wrote:

>>> as I said, I'd opt for having a new clone() syscall in
>>> addition to the existing one, with a separate 64bit
>>> set of flags to decide what namespaces should be created
>>> or cloned. there is no problem with putting 'important'
>>> or generally 'useful' flags (like for example for pid,
>>> uts or lightweight network isolation) into the existing
>>> clone call (will require a simple mapping if done properly)
>>> so that they can be used with 'older' libc interfaces too
>>>
>>> I know, it would be 'nice' to keep the existing clone()
>>> interface, but I think it already has become a complication
>>> we should avoid (and we have not even used up all the
>>> available flags :)
>> agree and so does Kirill.
>>
>>> are there any strong arguments against having a new
>>> clone() syscall, which I was missing so far?
>> I don't see any.
>>
>> I'm going to revive execns() syscall into a clone_ns() syscall as suggested
>> by Kirill and you. Then, others will be free to nack ;)
>
> I think it is silly, but I see not real problems with the idea.

that's not a violent agreement :)

i'll work on it.

thanks,

C.

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by dev on Sat, 16 Sep 2006 12:05:09 GMT
View Forum Message <> Reply to Message

>>>Plus what other namespaces are on the todo list?
>>>We have network, and pid, and time.
>>
>>I think more.
>>
>>proc-ns,
>>sysfs-ns,
>>printk-ns or syslog-ns?: syslog should be virtualized
>>and more...
>
>
> I don't think those meet the criteria for namespaces.
> But certainly there is work we need to do there.
Well, it is hard to say what is the criteria...

>>semi-namespaces:
>>fs-ns (should regulate which filesystems are accessiable from container, but
>>probably this is not exact name space... need to think over...),

> I think the problem there is the same as allowing untrusted users the ability
> to mount filesystems, in which case we just tag filesystems that are safe
> for untrusted users to use.
You need some groupping mechanisms, don't you?
Say, I need to allow isofs for containers 1,2,5,6
and ext3 for containers 2,3,4,5

>>dev-ns (should regulate which devices are accessiable from container)
> Yes. Devices certainly have global names that we need to bring under
> control. The easy solution is just to limit CAP_SYS_MKNOD but we
> may need something more.

CAP_SYS_MKNOD is not an option.

Can you please propose how to organize it?

You can check how it is implemented in OpenVZ in kernel/vecalls.c
devperms_struct
real_get_device_perms_ve()
real_setdevperms()

BTW, taking a look near this code, I found another bunch of interesting
functionality - statistics (e.g. real_update_load_avg_ve).

Though load avg statistics logically belong to pspace namespace there is a lot of other stats
which can not be associated so easily with the namespaces.

> One of the pieces that needs consideration when it comes to permissions
> is the plan9 style of permission control.   Where file have an initial
> owner, and if someone else needs access to them you chmod, chown them
> so that everyone who needs to has access. I think that is an simpler
> model to get right than to have a bunch of special cases.
it is Linux :)

Thanks,
Kirill

Subject: Re: [RFC][PATCH 1/2] add user namespace [try #2]
Posted by Herbert Poetzl on Sat, 16 Sep 2006 14:19:02 GMT
View Forum Message <> Reply to Message

On Sat, Sep 16, 2006 at 04:09:31PM +0400, Kirill Korotaev wrote:
> >>>Plus what other namespaces are on the todo list?
> >>>We have network, and pid, and time.
> >>
> >>I think more.
> >>
> >>proc-ns,
> >>sysfs-ns,
> >>printk-ns or syslog-ns?: syslog should be virtualized
> >>and more...
> >
> >
> > I don't think those meet the criteria for namespaces.
> > But certainly there is work we need to do there.
> Well, it is hard to say what is the criteria...
>
> >>semi-namespaces:
> >>fs-ns (should regulate which filesystems are accessiable from container, but
> >>probably this is not exact name space... need to think over...),
>
> > I think the problem there is the same as allowing untrusted users the ability
> > to mount filesystems, in which case we just tag filesystems that are safe
> > for untrusted users to use.
> You need some groupping mechanisms, don't you?
> Say, I need to allow isofs for containers 1,2,5,6
> and ext3 for containers 2,3,4,5
>
> >>dev-ns (should regulate which devices are accessiable from container)
> > Yes. Devices certainly have global names that we need to bring under
> > control. The easy solution is just to limit CAP_SYS_MKNOD but we
> > may need something more.
>

> CAP_SYS_MKNOD is not an option.

removing that is sufficient for Linux-VServer as is
but we have some plans for a better solution, in the
future ...

> Can you please propose how to organize it?
>
> You can check how it is implemented in OpenVZ in kernel/vecalls.c
> devperms_struct
> real_get_device_perms_ve()
> real_setdevperms()
>
> BTW, taking a look near this code, I found another bunch of
> interesting functionality - statistics (e.g. real_update_load_avg_ve).
>
> Though load avg statistics logically belong to pspace namespace there
> is a lot of other stats which can not be associated so easily with the
> namespaces.

most of them can be combined with the accounting or
limit namespace IMHO, at least that is true for
Linux-VServer

but don't get me wrong, I think we need a lot more
different namespaces in the future, very similar to
the cap requirements, which should get a lot more
fine grained than they are right now ...

best,
Herbert

> > One of the pieces that needs consideration when it comes to
> > permissions is the plan9 style of permission control. Where file
> > have an initial owner, and if someone else needs access to them
> > you chmod, chown them so that everyone who needs to has access. I
> > think that is an simpler model to get right than to have a bunch of
> > special cases.
> it is Linux :)
>
> Thanks,
> Kirill
> ____
> Containers mailing list
> Containers@lists.osdl.org
> https://lists.osdl.org/mailman/listinfo/containers

____
Containers mailing list