

---

Subject: [patch 0/1][NETNS49] Make af\_unix autobind per namespace

Posted by [Daniel Lezcano](#) on Tue, 02 Oct 2007 15:18:46 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

The following patch change autobind fonction to use the ordernum from the network namespace instead of using the local static variable.

--

---

Containers mailing list

Containers@lists.linux-foundation.org

<https://lists.linux-foundation.org/mailman/listinfo/containers>

---

---

Subject: [patch 1/1][NETNS49] Make af\_unix autobind per network namespace

Posted by [Daniel Lezcano](#) on Tue, 02 Oct 2007 15:18:47 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

This patch change the static ordernum variable to be relative to the network namespace.

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

---

```
include/net/net_namespace.h | 1 +
net/unix/af_unix.c          | 5 ++++-
2 files changed, 5 insertions(+), 1 deletion(-)
```

Index: linux-2.6-netns/include/net/net\_namespace.h

```
=====
--- linux-2.6-netns.orig/include/net/net_namespace.h
+++ linux-2.6-netns/include/net/net_namespace.h
@@ -60,6 +60,7 @@ struct net {
 /* AF_UNIX */
 int sysctl_unix_max_dgram_qlen;
 void *unix_sysctl;
+ u32 ordernum;
```

```
/* XFRM */
```

```
u32 sysctl_xfrm_aevent_etime;
```

Index: linux-2.6-netns/net/unix/af\_unix.c

```
=====
--- linux-2.6-netns.orig/net/unix/af_unix.c
+++ linux-2.6-netns/net/unix/af_unix.c
@@ -681,8 +681,8 @@ static int unix_autobind(struct socket *
 struct sock *sk = sock->sk;
 struct net *net = sk->sk_net;
 struct unix_sock *u = unix_sk(sk);
- static u32 ordernum = 1;
```

```
struct unix_address * addr;
+ u32 ordernum = net->ordernum;
int err;

mutex_lock(&u->readlock);
@@ -720,6 +720,7 @@ retry:
u->addr = addr;
__unix_insert_socket(&unix_socket_table[addr->hash], sk);
spin_unlock(&unix_table_lock);
+ net->ordernum = ordernum;
err = 0;

out: mutex_unlock(&u->readlock);
@@ -2164,6 +2165,8 @@ static int unix_net_init(struct net *net
int error = -ENOMEM;

net->sysctl_unix_max_dgram_qlen = 10;
+ net->ordernum = 1;
+
#ifdef CONFIG_PROC_FS
if (!proc_net_fops_create(net, "unix", 0, &unix_seq_fops))
goto out;

--
```

---

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

---

---

Subject: Re: [patch 0/1][NETNS49] Make af\_unix autobind per namespace  
Posted by [Daniel Lezcano](#) on Tue, 02 Oct 2007 15:31:52 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Daniel Lezcano wrote:  
> The following patch change autobind fonction to use the ordernum  
> from the network namespace instead of using the local static variable.  
>

I forgot to add this trivial program which does autobind.

---

```
#include <sys/socket.h>
#include <sys/un.h>
#include <unistd.h>
#include <errno.h>
```

```
#include <string.h>
#include <stdio.h>

int main(int argc, char* argv[])
{
    int fd;
    struct sockaddr_un addr;

    fd = socket(PF_UNIX, SOCK_DGRAM, 0);
    if (fd == -1) {
        perror("socket");
        return 1;
    }

    memset(&addr, 0, sizeof(addr));

    addr.sun_family = AF_UNIX;
    strcpy(addr.sun_path, "");

    if (bind(fd, &addr, sizeof(short))) {
        perror("bind");
        return 1;
    }

    return 0;
}
```

---

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

---

---

Subject: Re: [patch 0/1][NETNS49] Make af\_unix autobind per namespace  
Posted by [ebiederm](#) on Tue, 02 Oct 2007 17:18:03 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

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

> The following patch change autobind fonction to use the ordernum  
> from the network namespace instead of using the local static variable.

Why do we care?  
Information leak?  
Some application is expecting a predictable autobind value?

Just skimming the code it looks like it will work correctly without  
this.

Eric

---

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

---

---

Subject: Re: [patch 0/1][NETNS49] Make af\_unix autobind per namespace  
Posted by [Daniel Lezcano](#) on Tue, 02 Oct 2007 20:51:51 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Eric W. Biederman wrote:

> Daniel Lezcano <dlezcano@fr.ibm.com> writes:  
>  
>> The following patch change autobind fonction to use the ordernum  
>> from the network namespace instead of using the local static variable.  
>  
> Why do we care?  
> Information leak?  
> Some application is expecting a predictable autobind value?  
>  
> Just skimming the code it looks like it will work correctly without  
> this.

I think my summary is ... too short :)

I don't see any applications taking care of this. If they ask for an abstract socket, then they don't care about the bind result. So probably, the patchset is totally useless.

But from the POV of the checkpoint/restart, we should check if this value is somewhere visible from userspace and so storable by an application.

It appears this is the case with /proc/net/unix, where an abstract socket is symbolized by the path pattern "@". Example:

```
cat /proc/net/unix
```

```
Num    RefCount Protocol Flags   Type St Inode Path
c6a27710: 00000002 00000000 00000000 0002 01 4357 @00003
```

I agree by the fact that can be considered as a detail and the probability to have an application storing this informaton is very small ( eg. checkpointing while doing netstat in the container ). But IMHO, the paradigm "never seen from userspace" fails and that justifies to have the ordernum variable relative to a namespace.

-- Daniel

---

Subject: Re: [patch 0/1][NETNS49] Make af\_unix autobind per namespace  
Posted by [ebiederm](#) on Tue, 02 Oct 2007 22:43:38 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

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

> Eric W. Biederman wrote:  
>> Daniel Lezcano <dlezcano@fr.ibm.com> writes:  
>>  
>>> The following patch change autobind fonction to use the ordernum  
>>> from the network namespace instead of using the local static variable.  
>>  
>> Why do we care?  
>> Information leak?  
>> Some application is expecting a predictable autobind value?  
>>  
>> Just skimming the code it looks like it will work correctly without  
>> this.  
>  
> I think my summary is ... too short :)  
>  
> I don't see any applications taking care of this. If they ask for an abstract  
> socket, then they don't care about the bind result. So probably, the patchset is  
> totally useless.  
>  
> But from the POV of the checkpoint/restart, we should check if this value is  
> somewhere visible from userspace and so storable by an application.

Right. And we already can already specifically select this result.  
My point is that the semi random sequence generator logic does not  
need to be per namespace, because people don't care what the sequence.  
That sequence is not exported to user space.

> It appears this is the case with /proc/net/unix, where an abstract socket is  
> symbolized by the path pattern "@". Example:  
>  
> cat /proc/net/unix  
>  
> Num RefCount Protocol Flags Type St Inode Path  
> c6a27710: 00000002 00000000 00000000 0002 01 4357 @00003

Right, and that part we should definitely preserve for checkpoint/restart

purposes.

> I agree by the fact that can be considered as a detail and the probability to  
> have an application storing this informaton is very small ( eg. checkpointing  
> while doing netstat in the container ). But IMHO, the paradigm "never seen from  
> userspace" fails and that justifies to have the ordernum variable relative to a  
> namespace.

My point was that ordernum itself is not seen. It is just an arbitrary number  
and we are allowed to change the algorithm for selecting a new abstract  
namespace name at will.

If there is something in userspace that depends on the algorithm for selecting  
the abstract name then making ordernum per namespace make sense.

At the moment the code is simpler to use what is effectively a different  
algorithm for selecting the abstract namespace name of the socket.

Eric

---

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

---

---

Subject: Re: [patch 0/1][NETNS49] Make af\_unix autobind per namespace  
Posted by [den](#) on Wed, 03 Oct 2007 08:11:57 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Eric W. Biederman wrote:

> Daniel Lezcano <dlezcano@fr.ibm.com> writes:  
>  
>> The following patch change autobind fonction to use the ordernum  
>> from the network namespace instead of using the local static variable.  
>  
> Why do we care?  
> Information leak?  
> Some application is expecting a predictable autobind value?  
>  
> Just skimming the code it looks like it will work correctly without  
> this.

I also do not see a need for this :)

Regards,  
Den

---

---

Subject: Re: [patch 0/1][NETNS49] Make af\_unix autobind per namespace  
Posted by [den](#) on Wed, 03 Oct 2007 08:14:07 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Daniel Lezcano wrote:

> Eric W. Biederman wrote:

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

>>

>>> The following patch change autobind fonction to use the ordernum  
>>> from the network namespace instead of using the local static variable.

>>

>> Why do we care?

>> Information leak?

>> Some application is expecting a predictable autobind value?

>>

>> Just skimming the code it looks like it will work correctly without  
>> this.

>

> I think my summary is ... too short :)

>

> I don't see any applications taking care of this. If they ask for an  
> abstract socket, then they don't care about the bind result. So  
> probably, the patchset is totally useless.

>

> But from the POV of the checkpoint/restart, we should check if this  
> value is somewhere visible from userspace and so storable by an  
> application.

we do not care with this in checkpointing. One namespace socket does not  
see other namespace socket

Regards,  
Den

---

Subject: Re: [patch 0/1][NETNS49] Make af\_unix autobind per namespace  
Posted by [Daniel Lezcano](#) on Wed, 03 Oct 2007 08:35:31 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Eric W. Biederman wrote:  
> Daniel Lezcano <dlezcano@fr.ibm.com> writes:  
>  
>> Eric W. Biederman wrote:  
>>> Daniel Lezcano <dlezcano@fr.ibm.com> writes:  
>>>>  
>>>> The following patch change autobind fonction to use the ordernum  
>>>> from the network namespace instead of using the local static variable.  
>>> Why do we care?  
>>> Information leak?  
>>> Some application is expecting a predictable autobind value?  
>>>  
>>> Just skimming the code it looks like it will work correctly without  
>>> this.  
>> I think my summary is ... too short :)  
>>  
>> I don't see any applications taking care of this. If they ask for an abstract  
>> socket, then they don't care about the bind result. So probably, the patchset is  
>> totally useless.  
>>  
>> But from the POV of the checkpoint/restart, we should check if this value is  
>> somewhere visible from userspace and so storable by an application.  
>  
> Right. And we already can already specifically select this result.  
> My point is that the semi random sequence generator logic does not  
> need to be per namespace, because people don't care what the sequence.  
> That sequence is not exported to user space.  
>  
>> It appears this is the case with /proc/net/unix, where an abstract socket is  
>> symbolized by the path pattern "@". Example:  
>>  
>> cat /proc/net/unix  
>>  
>> Num RefCount Protocol Flags Type St Inode Path  
>> c6a27710: 00000002 00000000 00000000 0002 01 4357 @00003  
>  
> Right, and that part we should definitely preserve for checkpoint/restart  
> purposes.  
>  
>> I agree by the fact that can be considered as a detail and the probability to  
>> have an application storing this informaton is very small ( eg. checkpointing  
>> while doing netstat in the container ). But IMHO, the paradigm "never seen from  
>> userspace" fails and that justifies to have the ordernum variable relative to a  
>> namespace.  
>  
> My point was that ordernum itself is not seen. It is just an arbitrary number  
> and we are allowed to change the algorithm for selecting a new abstract  
> namespace name at will.



Hmm, right. That makes sense.

> If there is something in userspace that depends on the algorithm for selecting  
> the abstract name then making ordernum per namespace make sense.

Ok, fair enough. Let's forget this patch. It is small enough to rewrite  
it if unexpectedly something bad happens with ordernum.

-- Daniel

---

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

---

---

Subject: Re: [patch 0/1][NETNS49] Make af\_unix autobind per namespace  
Posted by [Cedric Le Goater](#) on Wed, 03 Oct 2007 13:11:44 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Denis V. Lunev wrote:

> Daniel Lezcano wrote:

>> Eric W. Biederman wrote:

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

>>>>

>>>> The following patch change autobind fonction to use the ordernum

>>>> from the network namespace instead of using the local static variable.

>>> Why do we care?

>>> Information leak?

>>> Some application is expecting a predictable autobind value?

>>>>

>>>> Just skimming the code it looks like it will work correctly without

>>>> this.

>> I think my summary is ... too short :)

>>

>> I don't see any applications taking care of this. If they ask for an

>> abstract socket, then they don't care about the bind result. So

>> probably, the patchset is totally useless.

>>

>> But from the POV of the checkpoint/restart, we should check if this

>> value is somewhere visible from userspace and so storable by an

>> application.

>

> we do not care with this in checkpointing. One namespace socket does not

> see other namespace socket

my 2 cnts,

when 'restarting' a socket bound to an abstract name, we will have a EADDRINUSE if we try to rebind it to an abstract name which is already in use by a socket in a another namespace ?

it seems to me that this is an identifier and like any identifier it should be private to the namespace, which probably means having `unix_abstract_socket_table[]` per net namespace.

Cheers,

C.

---

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

---

---

Subject: Re: [patch 0/1][NETNS49] Make af\_unix autobind per namespace  
Posted by [ebiederm](#) on Wed, 03 Oct 2007 14:36:53 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Cedric Le Goater <[clg@fr.ibm.com](mailto:clg@fr.ibm.com)> writes:

> my 2 cnts,

>

> when 'restarting' a socket bound to an abstract name, we will have  
> a EADDRINUSE if we try to rebind it to an abstract name which is  
> already in use by a socket in a another namespace ?

No.

> it seems to me that this is an identifier and like any identifier  
> it should be private to the namespace, which probably means having  
> `unix_abstract_socket_table[]` per net namespace.

Yes it is. It is a hash table so we are filter the hash chain and not having two copies of the table. But effectively it's the same thing.

All this patch was suggesting was having a per network namespace copy of the data structure for the random number generator for generating the name.

The ``random number generator'' is just a 16bit counter that loops through all 64k values seeing if a name is available and if so using it. Sharing our place in the loop between different namespaces may be ineffeicient but it should work fine.

Eric

---

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

---

---

Subject: Re: [patch 0/1][NETNS49] Make af\_unix autobind per namespace  
Posted by [Cedric Le Goater](#) on Wed, 03 Oct 2007 15:34:10 GMT  
[View Forum Message](#) <> [Reply to Message](#)

---

Eric W. Biederman wrote:

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

>> my 2 cnts,

>>

>> when 'restarting' a socket bound to an abstract name, we will have

>> a EADDRINUSE if we try to rebind it to an abstract name which is

>> already in use by a socket in a another namespace ?

>

> No.

ok. i just went over your AF\_UNIX patch and saw that 'struct net' was being checked for abstract sockets.

C.

>> it seems to me that this is an identifier and like any identifier

>> it should be private to the namespace, which probably means having

>> unix\_abstract\_socket\_table[] per net namespace.

>

> Yes it is. It is a hash table so we are filter the hash chain

> and not having two copies of the table. But effectively it's

> the same thing.

>

> All this patch was suggesting was having a per network namespace

> copy of the data structure for the random number generator for

> generating the name.

>

> The ``random number generator'' is just a 16bit counter that loops

> through all 64k values seeing if a name is available and if so

> using it. Sharing our place in the loop between different namespaces

> may be ineffeicient but it should work fine.

>

> Eric

---

Containers mailing list  
Containers@lists.linux-foundation.org

