Subject: [patch 0/2][NETNS45][V3] remove timewait sockets at namespace exit
Posted by Daniel Lezcano on Fri, 28 Sep 2007 09:51:08 GMT
View Forum Message <> Reply to Message

Denis Lunev spotted that using a reference to the network namespace
with the timewait sockets will be a waste of time because they
are pointless while we will remove the network stack at network
namespace exit.

The following patches do the following:
 - fix missing network namespace reference in timewait
 socket
 - do the effective timewait socket cleanup at network
 namespace exit.

The following code is a test program which creates timewait
sockets.

```
#include <stdio.h>
#include <string.h>
#include <sys/types.h>
#include <sys/socket.h>
#include <sys/poll.h>
#include <netinet/in.h>
#include <netinet/tcp.h>
#include <arpa/inet.h>

#include <unistd.h>

#define MAXCONN 100

int client(int *fds)
{
	int i;
	struct sockaddr_in addr;

	close(fds[1]);

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

	addr.sin_family = AF_INET;
	addr.sin_port =  htons(10000);
	addr.sin_addr.s_addr = inet_addr("127.0.0.1");

	if (read(fds[0], &i, sizeof(i)) == -1) {
		perror("read");
		return 1;
	}
```

```
	for (i = 0; i < MAXCONN; i++) {
		int fd = socket(PF_INET, SOCK_STREAM, 0);
		if (fd == -1) {
			perror("socket");
			return 1;
		}

		if (connect(fd, (const struct sockaddr *)&addr, sizeof(addr))) {
			perror("connect");
			return 1;
		}
	}

	return 0;
}

int server(int *fds)
{
	int i, fd, fdpoll[MAXCONN];
	struct sockaddr_in addr;
	socklen_t socklen = sizeof(addr);

	close(fds[0]);

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

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

	addr.sin_family = AF_INET;
	addr.sin_port = htons(10000);
	addr.sin_addr.s_addr = inet_addr("127.0.0.1");

	if (bind(fd, (const struct sockaddr *)&addr, sizeof(addr))) {
		perror("bind");
		return 1;
	}

	if (listen(fd, MAXCONN)) {
		perror("listen");
		return 1;
	}

	if (write(fds[1], &i, sizeof(i)) == -1) {
```

```
                perror("write");
                return 1;
        }

        for (i = 0; i < MAXCONN; i++) {
                int f = accept(fd, (struct sockaddr *)&addr, &socklen);
                if (f == -1) {
                        perror("accept");
                        return 1;
                }
                fdpoll[i] = f;
        }

        return 0;
}

int main(int argc, char *argv[])
{
        int fds[2];
        int pid;

        if (pipe(fds)) {
                perror("pipe");
                return 1;
        }

        pid = fork();
        if (pid == -1) {
                perror("fork");
                return 1;
        }

        if (!pid)
                return client(fds);
        else
                return server(fds);
}

--
```


Subject: [patch 1/2][NETNS45][V3] add a reference to the netns for timewait
Posted by Daniel Lezcano on Fri, 28 Sep 2007 09:51:09 GMT

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

When a socket changes to a timewait socket, the network namespace
is not copied from the original socket.

Here we hold a usage reference, not the ref count on the network
namespace, so the network namespace will be freed either the usage
reference is not 0. The network namespace cleanup function will
fail if there is any usage of it. In this case, we should ensure
there is no usage of the network namespace.

Signed-off-by: Daniel Lezcano <dlezcano@fr.ibm.com>
Acked-by: Denis V. Lunev <den@openvz.org>
---
 include/net/inet_timewait_sock.h | 2 ++
 net/ipv4/inet_timewait_sock.c | 1 +
 2 files changed, 3 insertions(+)

Index: linux-2.6-netns/include/net/inet_timewait_sock.h
===
--- linux-2.6-netns.orig/include/net/inet_timewait_sock.h
+++ linux-2.6-netns/include/net/inet_timewait_sock.h
@@ -197,12 +197,14 @@ static inline void inet_twsk_put(struct
 {
  if (atomic_dec_and_test(&tw->tw_refcnt)) {
   struct module *owner = tw->tw_prot->owner;
+ struct net *net = tw->tw_net;
   twsk_destructor((struct sock *)tw);
 #ifdef SOCK_REFCNT_DEBUG
   printk(KERN_DEBUG "%s timewait_sock %p released\n",
          tw->tw_prot->name, tw);
 #endif
   kmem_cache_free(tw->tw_prot->twsk_prot->twsk_slab, tw);
+ release_net(net);
   module_put(owner);
  }
 }
Index: linux-2.6-netns/net/ipv4/inet_timewait_sock.c
===
--- linux-2.6-netns.orig/net/ipv4/inet_timewait_sock.c
+++ linux-2.6-netns/net/ipv4/inet_timewait_sock.c
@@ -108,6 +108,7 @@ struct inet_timewait_sock *inet_twsk_all
   tw->tw_hash   = sk->sk_hash;
   tw->tw_ipv6only   = 0;
   tw->tw_prot   = sk->sk_prot_creator;
+ tw->tw_net   = hold_net(sk->sk_net);
   atomic_set(&tw->tw_refcnt, 1);

```
  inet_twsk_dead_node_init(tw);
  __module_get(tw->tw_prot->owner);

--
```

_____

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

---

Subject: [patch 2/2][NETNS45][V3] remove timewait sockets at cleanup
Posted by Daniel Lezcano on Fri, 28 Sep 2007 09:51:10 GMT
View Forum Message <> Reply to Message

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

Denis Lunev spotted that if we take a reference to the network namespace
with the timewait sockets, we will need to wait for their expiration to
have the network namespace freed. This is a waste of time, the timewait
sockets are for avoiding to receive a duplicate packet from the network,
if the network namespace is freed, the network stack is removed, so no
chance to receive any packets from the outside world.

This patchset remove/destroy the timewait sockets when the
network namespace is freed.

The exit method registered by netns_register_subsys is put in the tcp.c
file and not in inet_timewait_sock.c. The reasons are we browse the tcp
established hash table and I don't want to add references to tcp in inet
timewait sockets and, furthermore, dccp protocol uses the inet timewait
sock too. IMHO, if we status to cleanup dccp timewait too, we should add
a exit method in dccp file.

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

```
---
 net/ipv4/tcp.c |   41 ++++++++++++++++++++++++++++++++++++++++++++++
 1 file changed, 41 insertions(+)

Index: linux-2.6-netns/net/ipv4/tcp.c
===================================================================
--- linux-2.6-netns.orig/net/ipv4/tcp.c
+++ linux-2.6-netns/net/ipv4/tcp.c
@@ -2432,8 +2432,49 @@ static int tcp_net_init(struct net *net)
  return 0;
 }

+/*
+ * Wipeout tcp timewait sockets, they are no longer needed
```

```
+ * because we destroy the network namespace, so no risk to
+ * have duplicate packet coming from the network
+ */
+static void tcp_net_exit(struct net *net)
+{
+ struct inet_timewait_sock *tw;
+ struct sock *sk;
+ struct hlist_node *node;
+ int h;
+
+ local_bh_disable();
+
+        /* Browse the the established hash table */
+ for (h = 0; h < (tcp_hashinfo.ehash_size); h++) {
+  struct inet_ehash_bucket *head =
+   inet_ehash_bucket(&tcp_hashinfo, h);
+ restart:
+  write_lock(&head->lock);
+  sk_for_each(sk, node, &head->twchain) {
+
+   tw = inet_twsk(sk);
+   if (tw->tw_net != net)
+    continue;
+   sock_hold(sk);
+
+   write_unlock(&head->lock);
+
+   inet_twsk_deschedule(tw, &tcp_death_row);
+   inet_twsk_put(tw);
+
+   goto restart;
+  }
+  write_unlock(&head->lock);
+ }
+
+ local_bh_enable();
+}
+
 static struct pernet_operations tcp_net_ops = {
   .init = tcp_net_init,
+ .exit = tcp_net_exit,
 };

 void __init tcp_init(void)

--
```

_____

Containers mailing list

Subject: Re: [patch 2/2][NETNS45][V3] remove timewait sockets at cleanup
Posted by den on Fri, 28 Sep 2007 10:06:23 GMT
View Forum Message <> Reply to Message

Daniel Lezcano wrote:
> From: Daniel Lezcano <dlezcano@fr.ibm.com>
>
> Denis Lunev spotted that if we take a reference to the network namespace
> with the timewait sockets, we will need to wait for their expiration to
> have the network namespace freed. This is a waste of time, the timewait
> sockets are for avoiding to receive a duplicate packet from the network,
> if the network namespace is freed, the network stack is removed, so no
> chance to receive any packets from the outside world.
>
> This patchset remove/destroy the timewait sockets when the
> network namespace is freed.
>
> The exit method registered by netns_register_subsys is put in the tcp.c
> file and not in inet_timewait_sock.c. The reasons are we browse the tcp
> established hash table and I don't want to add references to tcp in inet
> timewait sockets and, furthermore, dccp protocol uses the inet timewait
> sock too. IMHO, if we status to cleanup dccp timewait too, we should add
> a exit method in dccp file.
>
> Signed-off-by: Daniel Lezcano <dlezcano@fr.ibm.com>
Signed-off-by: Denis V. Lunev <den@openvz.org>

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