Subject: Even worse thing when migrating online
Posted by divB on Sat, 09 May 2009 22:04:42 GMT
View Forum Message <> Reply to Message

Hi,

Now I have a problem even much more worse (does this exist?). Sometimes when I do online-migration (from HN1 to HN2) the network connection between the two hosts drops (the funny thing: ONLY between the two hardware nodes!) and this leads to a fatal situation:

* The ssh commands from vzmigrate are not executed any more
* The VE is still up on HN1
* But it is also up on HN2 as "zombie VE"

```
HN1:~# vzlist
      VEID      NPROC STATUS  IP_ADDR         HOSTNAME
       201          6 running -


HN2:~# vzlist
      VEID      NPROC STATUS  IP_ADDR         HOSTNAME
       201          9 running -
HN2:~# vzctl enter 201
enter into VE 201 failed
HN2:~#
```

This is where the migration looks like:

```
NH1:~# vzmigrate2 -r no --keep-dst --online -v 192.168.200.1 201
OPT:-r
OPT:--keep-dst
OPT:--online
OPT:-v
OPT:192.168.200.1
Starting online migration of VE 201 on 192.168.200.1
OpenVZ is running...
   Loading /etc/vz/vz.conf and /etc/vz/conf/201.conf files
   Check IPs on destination node:
Preparing remote node
   Copying config file
201.conf                                                      100% 1756
1.7KB/s   00:00
Saved parameters for VE 201
   Creating remote VE root dir
   Creating remote VE private dir
```

```
   VZ disk quota disabled -- skipping quota migration
Syncing private
Live migrating VE
Stop apache2 if it is installed
Stopping web server: apache2 ... waiting .
   Suspending VE
Setting up checkpoint...
        suspend...
        get context...
Checkpointing completed succesfully
   Dumping VE
Setting up checkpoint...
        join context..
        dump...
Checkpointing completed succesfully
   Copying dumpfile
dump.201                                                      100% 1492KB
1.5MB/s   00:01
   Syncing private (2nd pass)
   VZ disk quota disabled -- skipping quota migration
   Undumping VE
Restoring VE ...
Starting VE ...
VE is mounted
        undump...
Setting CPU units: 1000
Configure meminfo: 2147483647
Configure veth devices: veth201.0
        get context...
VE start in progress...
Restoring completed succesfully
Adding interface veth201.0 to bridge br-lan on CT0 for CT201
```

After that, the script hangs. Clearly, as said, pinging HN2 is not possible any more. This leads to a hang of the SSH commands:

```
HN1:~# ps aux
[...]
root      3914  0.2  0.1   3928  1320 pts/1    S+   01:43   0:00 /bin/sh /usr/local/sbin/vzmigrate2 -r no
--keep-dst --online -v 192.168.200.1 201
root      3974  0.2  0.2   5124  2288 pts/1    S+   01:43   0:00 ssh root@192.168.200.1 vzctl restore
201 --undump --dumpfile /var/tmp/dump.201 --skip_arpdet
```

After killing PID 3974, the next ssh command from the vzmigrate script is spawned:

```
HN1:~# ps aux
[...]
root     3914  0.1  0.1   3928  1320 pts/1    S+   01:43   0:00 /bin/sh /usr/local/sbin/vzmigrate2 -r no
--keep-dst --online -v 192.168.200.1 201
root     3975  0.0  0.1   4248  1676 pts/2    Ss   01:43   0:00 /bin/bash
root     3978  6.0  0.1   5124  1828 pts/1    S+   01:44   0:00 ssh root@192.168.200.1 rm -f
/var/tmp/quotadump.201
```

As mentioned above, both hardware nodes are now inconsistent and "buggy". Just deleting /etc/vz/conf/201.conf and then rebooting BOTH hardware nodes resolves the problem

Well, but what exactly happens when starting my machines? First I have to mention that I only use vzeth and no vznet. So I have to make sure to bridge the veth-Device together with the bridges on the hardware node.

Additionally I have to big problem that Debian lenny does not yet support the EXTERNAL_SCRIPT functionality. So I hacked the wurgaround I found in [1].

So in common, my /etc/vz/conf/vps.mount looks like [2].

In this script, the vznetaddbr explained in [1] is called. The contents of this file is in [3].

The very big question now: Why does this happen? From a third computer I can ping both hardware nodes but they can't communicate anymore with each other! I am not sure if this problem is caused my bridging scripts...

Is there any hope to resolve this issue?

Thank you very much,
divB

[1] http://wiki.openvz.org/Veth#method_for_vzctl_version_.3C.3D_ 3.0.22
[2] http://pastebin.com/m33a4232a
[3] http://pastebin.com/m2136da98

Subject: Re: Even worse thing when migrating online
Posted by jeronimo on Wed, 05 Aug 2009 11:28:32 GMT
View Forum Message <> Reply to Message

Hi,

I'm experiencing the same problem -- have you managed to resolve it?

In fact, the bridge behaves quite mysteriously -- similarly, as soon as I stop a VE, the machine stops responsing for a while as well...

Thank you for any information.
Tom.

Subject: Re: Even worse thing when migrating online
Posted by divB on Wed, 05 Aug 2009 11:31:12 GMT
View Forum Message <> Reply to Message

No, sorry

In fact, I did not try it any more. But in future I will use offline migration. Very sad but online migration does not seem to work

---

Subject: Re: Even worse thing when migrating online
Posted by jeronimo on Wed, 05 Aug 2009 11:38:02 GMT
View Forum Message <> Reply to Message

That's a pity... ;-(

Well, I've checked that the online migration with the venet iface works fine -- I think that that is more a bridge problem that OpenVZ's problem... However, I don't know, why it behaves so strangely... ;-(

Tom.

Subject: Re: Even worse thing when migrating online
Posted by swindmill on Wed, 05 Aug 2009 21:34:42 GMT
View Forum Message <> Reply to Message

This sounds like an issue I've seen where a bridge's HWaddr changes when interfaces are added or removed from it.

Other machines on the network still have the old HWaddr in their arp cache and cannot communicate with the bridged machine for some period of time after the change occurs.

My workaround is to force the bridge to share it's HWaddr with the physical network interface on the host (ethX) and update the HWaddr to match upon adding or removing an interface.

Subject: Re: Even worse thing when migrating online
Posted by divB on Wed, 05 Aug 2009 23:58:59 GMT
View Forum Message <> Reply to Message

Hi,

Thank you for your ideas...

swindmill wrote on Wed, 05 August 2009 17:34This sounds like an issue I've seen where a bridge's HWaddr changes when interfaces are added or removed from it.

I am curious...why should this happen?

Quote:My workaround is to force the bridge to share it's HWaddr with the physical network interface on the host (ethX) and update the HWaddr to match upon adding or removing an interface.

How did you accomplish this?

Regards,
divB

Subject: Re: Even worse thing when migrating online
Posted by swindmill on Thu, 06 Aug 2009 00:11:17 GMT
View Forum Message <> Reply to Message

I believe it happens because of the way linux bridges assign themselves a MAC address.

IIRC, the bridge takes its MAC from the lowest MAC of
all the interfaces that form the bridge, and the bridge changes its MAC if that interface is detached from the bridge, taking then the next lowest one.

I use a command similar to this one in my /usr/sbin/vznetaddbr and /etc/vz/conf/CTID.umount scripts for any container using veth and bridging:

ifconfig br0 hw ether $(ifconfig eth0 | awk '{print $5; exit}')

Subject: Re: Even worse thing when migrating online
Posted by divB on Thu, 06 Aug 2009 00:38:32 GMT
View Forum Message <> Reply to Message

Thank you. I did not know this. Also, I can't reproduce this behaviour. All my br-* interfaces have

exactly the same MAC as their eth-interfaces - none of them takes the MAC from a veth*-Interface.

```
# brctl show
bridge name     bridge id               STP enabled     interfaces
br-lan          8000.000e0ca1b929       no              eth0.2
                                                        tap0
                                                        veth200.0
                                                        veth201.0
                                                        veth202.0
br-stfg         8000.000e0ca1b929       no              eth0.4
                                                        veth400.0
br-wan          8000.000e0ca1b929       no              eth0.3
                                                        veth300.0
                                                        veth301.0
                                                        veth302.0
```

So normally, the MAC can't change when removing/adding veth*-interfaces.

Regards,
divB

Subject: Re: Even worse thing when migrating online
Posted by jeronimo on Thu, 06 Aug 2009 15:50:40 GMT
View Forum Message <> Reply to Message

Hi Swindmill,

swindmill wrote on Wed, 05 August 2009 23:34This sounds like an issue I've seen where a bridge's HWaddr changes when interfaces are added or removed from it.

Other machines on the network still have the old HWaddr in their arp cache and cannot communicate with the bridged machine for some period of time after the change occurs.

My workaround is to force the bridge to share it's HWaddr with the physical network interface on the host (ethX) and update the HWaddr to match upon adding or removing an interface.

thank you very much! That was exactly the problem -- now, the online migration succeeds (previously, the ssh connection became lost and thus it has failed due to the MAC change) and everything is working fine.

More precisely, almost fine, since now I have to somehow persuade the switches to automatically and quickly learn the location of the migrated VE (currently, even if the VE has been migrated, all the pings coming outside of the VE's segment are "routed" to the source HW node -- as soon as I

ping anything from the migrated VE, the switches learn this new location and "route" all the packets as suspected.

Once again, thank you for solving the issue.
Tom.

Subject: Re: Even worse thing when migrating online
Posted by divB on Thu, 06 Aug 2009 17:54:52 GMT
View Forum Message <> Reply to Message

I do not understand the issue, however, thank you too. I will also try it

---

Subject: Re: Even worse thing when migrating online
Posted by divB on Sun, 06 Sep 2009 20:35:29 GMT
View Forum Message <> Reply to Message

Hi,

I have built in the hack too and now the migration also succeeds! I have tried a few times, no problems!

But actually I do NOT understand why because the MAC of the bridge never changes!!

Quote:More precisely, almost fine, since now I have to somehow persuade the switches to automatically and quickly learn the location of the migrated VE (currently, even if the VE has been migrated, all the pings coming outside of the VE's segment are "routed" to the source HW node -- as soon as I ping anything from the migrated VE, the switches learn this new location and "route" all the packets as suspected.

My hack is to read out the ARP table just before migration and after migration, I ping all those IPs. For this, I modified vzmigrate:

```
--- /usr/sbin/vzmigrate 2009-05-08 14:40:23.000000000 +0200
+++ /usr/local/sbin/vzmigrate2  2009-09-05 13:07:07.000000000 +0200
@@ -446,6 +446,13 @@
 if [ $online -eq 1 ]; then
        log 1 "Live migrating VE"

+       log 1 "Saving arp table"
+       # vzctl exec2 $VEID "arp -n | awk '{print $1}' | egrep '^[0-9\.]+$' >/var/run/arp_table"
+       vzctl exec $VEID 'arp -n | awk "{print \$1}" | egrep "^[0-9\.]+\$" >/var/run/arp_table'
+
        log 2 "Suspending VE"
```

```
        time_suspend=$(date +%s.%N)
        if ! logexec 2 vzctl chkpnt $VEID --suspend ; then
@@ -560,6 +567,12 @@
        log 2 "Removing dumpfiles"
        rm -f "$VE_DUMPFILE"
        $SSH "root@$host" "rm -f $VE_DUMPFILE"
+
+       log 1 "Pinging hosts in arp cache"
+       $SSH root@$host vzctl exec $VEID "'[ -x /usr/sbin/fping ] && fping -i 1 -r 1 -f
/var/run/arp_table'"
 else
        if [ "$state" = "running" ]; then
                log 1 "Starting VE"
```

Regards,
divB

Subject: Re: Even worse thing when migrating online with veth(ernet) devices and bridges
Posted by curx on Sun, 06 Sep 2009 23:21:35 GMT
View Forum Message <> Reply to Message

Hi,

>[...] But it is also up on HN2 as "zombie VE"

^_ have you checked the restore process, see /proc/rst on HN2 ?

Bye,
Thorsten

Subject: Re: Even worse thing when migrating online with veth(ernet) devices and bridges
Posted by divB on Mon, 07 Sep 2009 09:01:56 GMT
View Forum Message <> Reply to Message

No, but with the patch in this thread it seems to work. I hope that this is not by accident...
