Subject: VE IP address stops working after several hours
Posted by fatbrother on Wed, 17 Sep 2008 17:55:31 GMT
View Forum Message <> Reply to Message

I have Debian Etch system with FZA OpenVZ kernel:
root@olympic:~# uname -a
Linux olympic 2.6.18-fza-028stab053.5-amd64 #1 SMP Sat Mar 1 19:50:43 UTC 2008 x86_64 GNU/Linux
root@olympic:~# vzctl --version
vzctl version 3.0.22-1dso1

HN connected to two VLANs:
eth0 Link encap:Ethernet HWaddr xxx
inet addr:10.4.0.97 Bcast:10.4.0.127 Mask:255.255.255.224

eth0.425 Link encap:Ethernet HWaddr xxx
inet addr:real_ip1 Bcast:real_ip_network Mask:255.255.255.240

Default router is eth0, source routing is set for real_ip_network to eth0.425.

VE4 has two IP addresses added by vzctl set 4 --ipadd. 10.4.0.104 and real_ip2

After I boot VE4 both IP addresses are accessible, so the routing and ARP proxy seem to be correct. But when VE4 runs for several hours, real_ip2 stops working. VE4 is still accessible via 10.4.0.104 and real_ip2 is pingable from HN, but not from outside.
When real_ip2 fails, record
real_ip2 * <from_interface> MP eth0.425
is present in ARP table, and HN real_ip1 is still accessible, but tcpdump shows that HN stops answering ARP requests for real_ip2.

VE4 restart or vzctl set 4 --ipdel real_ip2; vzctl set 4 --ipadd real_ip2; clear the condition: real_ip2 starts responding again, but after several hours it stops responding again, etc.
It runs for at least 3 hours, but no longer than 5 hours. Sorry, cannot get more precise estimate yet, and I'm not even sure that time is constant.
I found no specific external events and no log records that could be related to this. It just runs and then stops running for no visible reason.

Subject: Re: VE IP address stops working after several hours
Posted by maratrus on Thu, 18 Sep 2008 08:32:54 GMT
View Forum Message <> Reply to Message

Hello,

it's a strange situation.
Could you possibly show a little bit more information.

1. # ip rule list

2. # ip route list table all
3. # sysctl -a | grep arp_filter
4. # sysctl -a | grep proxy_arp

And does the following helps:
ip neigh del proxy real_ip2 dev eth0.425
ip neigh add proxy real_ip2 dev eth0.425
?

Does "arp -n" output differs before and after the real_ip2 is broken down? (I mean only that records that concerns with VE ip addresses).

Subject: Re: VE IP address stops working after several hours
Posted by fatbrother on Thu, 18 Sep 2008 10:05:16 GMT
View Forum Message <> Reply to Message

maratrus wrote on Thu, 18 September 2008 15:32Hello,
it's a strange situation.

maratrusCould you possibly show a little bit more information.

1. # ip rule list
0:	from all lookup 255
32764:	from real_ip_network/28 lookup Real
32765:	from real_ip1 lookup Real
32766:	from all lookup main
32767:	from all lookup default
maratrus2. # ip route list table all
default via real_ip_router dev eth0.425 table Real
real_ip2 dev venet0 scope link
10.4.0.106 dev venet0 scope link
10.4.0.104 dev venet0 scope link
10.4.0.105 dev venet0 scope link
10.4.0.102 dev venet0 scope link
10.4.0.103 dev venet0 scope link
10.4.0.101 dev venet0 scope link
real_network/28 dev eth0.425 proto kernel scope link src real_ip1
10.4.0.96/27 dev eth0 proto kernel scope link src 10.4.0.97
default via 10.4.0.126 dev eth0
broadcast real_ip_network dev eth0.425 table 255 proto kernel scope link src real_ip1
broadcast 10.4.0.127 dev eth0 table 255 proto kernel scope link src 10.4.0.97
broadcast 127.255.255.255 dev lo table 255 proto kernel scope link src 127.0.0.1
local real_ip1 dev eth0.425 table 255 proto kernel scope host src real_ip1
broadcast real_broadcast dev eth0.425 table 255 proto kernel scope link src real_ip1
broadcast 10.4.0.96 dev eth0 table 255 proto kernel scope link src 10.4.0.97
broadcast 127.0.0.0 dev lo table 255 proto kernel scope link src 127.0.0.1
local 10.4.0.97 dev eth0 table 255 proto kernel scope host src 10.4.0.97

local 127.0.0.1 dev lo table 255 proto kernel scope host src 127.0.0.1
local 127.0.0.0/8 dev lo table 255 proto kernel scope host src 127.0.0.1
(IPv6 stuff skipped)
unreachable default dev lo proto none metric -1 error -101 hoplimit 255
maratrus3. # sysctl -a | grep arp_filter
error: "Operation not permitted" reading key "net.ipv6.route.flush"
net.ipv4.conf.venet0.arp_filter = 0
net.ipv4.conf.eth0/425.arp_filter = 0
net.ipv4.conf.eth0.arp_filter = 0
net.ipv4.conf.lo.arp_filter = 0
net.ipv4.conf.default.arp_filter = 0
net.ipv4.conf.all.arp_filter = 0
error: "Operation not permitted" reading key "net.ipv4.route.flush"
maratrus4. # sysctl -a | grep proxy_arp
error: "Operation not permitted" reading key "net.ipv6.route.flush"
net.ipv4.conf.venet0.proxy_arp = 0
net.ipv4.conf.eth0/425.proxy_arp = 0
net.ipv4.conf.eth0.proxy_arp = 0
net.ipv4.conf.lo.proxy_arp = 0
net.ipv4.conf.default.proxy_arp = 0
net.ipv4.conf.all.proxy_arp = 0
error: "Operation not permitted" reading key "net.ipv4.route.flush"
maratrusAnd does the following helps:
ip neigh del proxy real_ip2 dev eth0.425
ip neigh add proxy real_ip2 dev eth0.425
?

I'll try it next time it fails again.

maratrusDoes "arp -n" output differs before and after the real_ip2 is broken down? (I mean only that records that concerns with VE ip addresses).
No. That's most confusing. I tried manipulating real_ip2 record via arp -i eth0.425 ... pub, but that does not seem to work.

Today it failed three times: at 06:20, at 10:31 and between 14:22 and 14:55 (i wrote a script that pings the real_ip2 every 100 secunds and does vzctl set --ipdel, vzctl set --ipadd when ping fails). Looks like it fails every 4 hours 11 minutes +-100 seconds (however I need several more datapoints to tell for sure). Does this time interval mean something to you?

Subject: Re: VE IP address stops working after several hours
Posted by maratrus on Thu, 18 Sep 2008 11:38:55 GMT
View Forum Message <> Reply to Message

Hi,

1. I'm not absolutely certain of my facts by I'm inclined to think that your HN never sends arp responce even everything works great.

Other machines just have your VEs arp-record in their arp-table which occur there during the VE strartup or changing ip address.
Could possibly check this fact by using tcpdump when everything works fine.

2. So, you have Real named table which contains only one record "default via real_ip_router dev eth0.425 table Real", don't you?
If yes, I don't understand how this configuration works at all. If something comes to VE it comes to HN first then the packet should be routed to your VE but Real table doesn't contain any mention of your VE which is on the venet interface. But your rule says that everything that comes "from real_ip_network/28" should be passed through Real table.

3. If the Real table would contain the record like "real_ip2 dev venet0 scope link" I can explain the first and the second points.
Could you possibly put the record like "real_ip2 dev venet0 scope link" to the Real table and check if something change.

4.
Quote:
32765: from real_ip1 lookup Real

By the way why do you have this rule?

Subject: Re: VE IP address stops working after several hours
Posted by fatbrother on Thu, 18 Sep 2008 13:31:12 GMT
View Forum Message <> Reply to Message

maratrus wrote on Thu, 18 September 2008 18:38Hi,

1. I'm not absolutely certain of my facts by I'm inclined to think that your HN never sends arp responce even everything works great.
Other machines just have your VEs arp-record in their arp-table which occur there during the VE strartup or changing ip address.
Could possibly check this fact by using tcpdump when everything works fine.
I though I've seen it responding to ARPs. But now I did clean check from another host and yes it does not respond, even when real_ip2 is pingable.
After IP change it sends out "who has real_ip2 tell real_ip2", may be that's what Cisco router uses for creating it's APR record.
So I'm ready to accept that you're right.

BTW, I tried ip neigh del/add sequence you suggested and it does not help.

Quote:2. So, you have Real named table which contains only one record "default via real_ip_router dev eth0.425 table Real", don't you?

Yes. That's first line of my "ip route list table all":
>default via real_ip_router dev eth0.425 table Real

Quote:If yes, I don't understand how this configuration works at all.
If something comes to VE it comes to HN first then the packet should be routed to your VE but Real table doesn't contain any mention of your VE which is on the venet interface. But your rule says that everything that comes "from real_ip_network/28" should be passed through Real table.
Oops. Well, I can explain how it works. I never tried to access real_ip2 from other hosts on real_network. So all packets I've sent to real_ip2 weren't "from real_ip_network/28".

Quote:3. If the Real table would contain the record like "real_ip2 dev venet0 scope link" I can explain the first and the second points.
Could you possibly put the record like "real_ip2 dev venet0 scope link" to the Real table and check if something change.
I added this record and the real_ip2 become pingable from other hosts on real_network!!

And it probably responded to ARP. I wasn't running tcpdump at that moment, but other hosts now have normal ARP records for real_ip2. I need to wait for 4 hours to tell whether that solves main problem. I have no access to the router (at least not before our Cisco admin will come to the work tomorrow), so I cannot flush it's ARP cache and cannot do a clean check.

But I probably should add rule for "to real_network" too, because now I get this:
>20:04:03.316468 IP real_ip3 > real_ip2: ICMP echo request, id 35397, seq 14, length 64
>20:04:03.316506 IP real_ip2 > real_ip3: ICMP echo reply, id 35397, seq 14, length 64
>20:04:03.316842 IP real_router > real_ip2: ICMP redirect real_ip3 to host real_ip3, length 36
and that redirect goes after every ping.
That's not a big issue, these hosts aren't supposed to communicate to each other over real_network, but that's annoying.

How do you suggest my ip rules should look like? I do not want to add a new explicit rule for every new real_ip for my VE, but probably that's the only correct way...

Quote:4.
Quote:
32765: from real_ip1 lookup Real

By the way why do you have this rule?

Err... I first added this rule, before I even added real_ip2. I just was testing HN connectivity to VLAN. Then I added rule for the network, and I haven't rebooted HN since then.

Subject: Re: VE IP address stops working after several hours
Posted by maratrus on Thu, 18 Sep 2008 14:09:47 GMT
View Forum Message <> Reply to Message

Hi,

Quote:
But I probably should add rule for "to real_network" too, because now I get this:
>20:04:03.316468 IP real_ip3 > real_ip2: ICMP echo request, id 35397, seq 14, length 64
>20:04:03.316506 IP real_ip2 > real_ip3: ICMP echo reply, id 35397, seq 14, length 64
>20:04:03.316842 IP real_router > real_ip2: ICMP redirect real_ip3 to host real_ip3, length 36
and that redirect goes after every ping.
That's not a big issue, these hosts aren't supposed to communicate to each other over real_network, but that's annoying.

I think the additional record "real_network/28 dev eth0.425" inside Real table should solve the problem. Redirect message in our case indicates that your HN send the packages through the router but might do it directly.

Quote:
How do you suggest my ip rules should look like? I do not want to add a new explicit rule for every new real_ip for my VE, but probably that's the only correct way...

You shouldn't create a new rule for each real_ip for your VE. You've created the single rule for the hole network segment. But I'm not quite understand why did you do that? Why don't you do without additional rules like
Quote:
32764: from real_ip_network/28 lookup Real
32765: from real_ip1 lookup Real
?
Why can't we delete them at all?

Subject: Re: VE IP address stops working after several hours
Posted by fatbrother on Thu, 18 Sep 2008 18:54:14 GMT
View Forum Message <> Reply to Message

maratrus wrote on Thu, 18 September 2008 21:09Hi,
Quote:
How do you suggest my ip rules should look like? I do not want to add a new explicit rule for every new real_ip for my VE, but probably that's the only correct way...

You shouldn't create a new rule for each real_ip for your VE.
Sorry, I was not clear. I do not want to create new rules and new explicit route records, but I had to create a routing table entry "ip route add real_ip2 dev venet0 scope link table Real". If I understand logic behind this, I must add similar record for every real_ip on my VEs.
Quote:
You've created the single rule for the hole network segment. But I'm not quite understand why did you do that? Why don't you do without additional rules like
Quote:

32764: from real_ip_network/28 lookup Real
32765: from real_ip1 lookup Real
?
Why can't we delete them at all?
Because HN default route points to 10.4.0.126 which is not the same host as real_ip_router, and 10.4.0.126 won't route real_ip_network packets to where I want them.
So as far as I understand I can delete real_ip1 rule ('cos it's redundant), but not real_ip_network rule.

Subject: Re: VE IP address stops working after several hours
Posted by maratrus on Fri, 19 Sep 2008 07:26:21 GMT
View Forum Message <> Reply to Message

Quote:
Sorry, I was not clear. I do not want to create new rules and new explicit route records, but I had to create a routing table entry "ip route add real_ip2 dev venet0 scope link table Real". If I understand logic behind this, I must add similar record for every real_ip on my VEs.

Yes, I'm afraid that you will have to do that. I'm going to specify what can we do with an inconvenience and if it is any news I'll let you now via this thread.

Quote:
So as far as I understand I can delete real_ip1 rule ('cos it's redundant), but not real_ip_network rule.

I think so.

Subject: Resolved (well, almost)
Posted by fatbrother on Fri, 19 Sep 2008 15:19:35 GMT
View Forum Message <> Reply to Message

It worked!!! At least it worked for full day and did not fail.
But one question remains (hopefully last).

Where should I add "ip route add real_ip2 dev venet0 ..." commands?

I run all other ip rule/ip route commands from /etc/network/if-up.d/sourcerouting script. But when that script runs, venet0 is not up yet, so the command "ip route add ... venet0 ..." fails. I had to enter it manually after HN reboot.

I am thinking about adding it to /etc/rc3.d/ with prefix 22 or 25, so it runs after /etc/rc3.d/S20vz, but it looks ugly for me.

Do you have any better idea? How is it supposed to be done?