Red Hat

Shienna Kim <shkim@redhat.com>

bz 2107083 status

Doug Smith <dosmith@redhat.com> Fri, Feb 17, 2023 at 11:50 PM
To: Sehee Oh <soh@redhat.com>, Tomofumi Hayashi <tohayash@redhat.com>
Cc: Andreas Karis <akaris@redhat.com>, Andres Guzman <aguzman@redhat.com>, Edu Alcaniz <ealcaniz@redhat.com>, Einat Glottmann <eglottma@redhat.com>, Franck Baudin <fbaudin@redhat.com>, Glenn West <gwest@redhat.com>, Inho Cho <incho@redhat.com>, Marc Curry <mcurry@redhat.com>, Robert Welling <rwelling@redhat.com>, Shienna Kim <shkim@redhat.com>

Andreas, big thanks for your help and your lab instructions, thank you.

Marc is this the impetus behind RFE-3503?

cc: +Tomofumi Hayashi

On Thu, Feb 16, 2023 at 9:50 PM Sehee Oh <soh@redhat.com> wrote:

> Dear Andreas.
> I hope you are doing well.
>
> Samsung tested it again with OCP 4.10.37, but the issue is still existing. Could you please look at the configs below?
> The default gateway part is different from yours. I'm not sure if it's a configuration issue or a bug fix version issue.
>
> *** networkattachmentdefinition**

```
[root@bastion01 ~]# oc get net-attach-def -n ss-smf test-nad -o yaml
apiVersion: k8s.cni.cncf.io/v1
kind: NetworkAttachmentDefinition
metadata:
annotations:
  kubectl.kubernetes.io/last-applied-configuration: |
    {"apiVersion":"k8s.cni.cncf.io/v1","kind":"NetworkAttachmentDefinition","metadata":{"annotations":{},"name":"test-
nad","namespace":"ss-smf"},"spec":{"config":"{ \"cniVersion\": \"0.3.0\", \"name\": \"test-nad\", \"plugins\": [{ \"type\":
\"bridge\", \"bridge\": \"br-ext\", \"vlan\": 0, \"ipMasq\": true, \"ipam\": { \"datastore\": \"kubernetes\", \"kubernetes\": {
\"kubeconfig\": \"/etc/kubernetes/cni/net.d/whereabouts.d/whereabouts.kubeconfig\" }, \"type\": \"whereabouts\",
\"range\": \"172.19.56.0/24\", \"range_start\": \"172.19.56.50\", \"range_end\": \"172.19.56.59\", \"routes\": [], \"log_file\":
\"/var/log/whereabouts.log\", \"log_level\": \"debug\" } }, { \"type\": \"tuning\", \"sysctl\": { \"net.ipv4.conf.net1.proxy_arp\":
\"0\" } }] }"}}
creationTimestamp: "2023-02-16T06:22:45Z"
generation: 3
name: test-nad
namespace: ss-smf
resourceVersion: "21080040"
uid: 6082b6af-9e5a-4f9e-85d6-d9b07653f948
spec:
config: '{ "cniVersion": "0.3.0", "name": "test-nad", "plugins": [{ "type": "bridge",
  "bridge": "br-ext", "vlan": 0, "ipMasq": true, "ipam": { "datastore": "kubernetes",
  "kubernetes": { "kubeconfig": "/etc/kubernetes/cni/net.d/whereabouts.d/whereabouts.kubeconfig"
  }, "type": "whereabouts", "range": "172.19.56.0/24", "range_start": "172.19.56.50",
  "range_end": "172.19.56.59", "routes": [], "log_file": "/var/log/whereabouts.log",
  "log_level": "debug" } }, { "type": "tuning", "sysctl": { "net.ipv4.conf.net1.proxy_arp":
  "0" } }] }'
```

> *** create br-ext at the worker node1 using nmtui**

```
[core@worker01 ~]$ ifconfig br-ext
br-ext: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
      inet 172.19.56.99  netmask 255.255.255.0  broadcast 172.19.56.255
      inet6 fe80::92db:d6:318:a150  prefixlen 64  scopeid 0x20<link>
      ether 48:df:37:75:49:28  txqueuelen 1000  (Ethernet)
      RX packets 22  bytes 1666 (1.6 KiB)
      RX errors 0  dropped 0  overruns 0  frame 0
```

```
      TX packets 14  bytes 1188 (1.1 KiB)
      TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

*** configure default gw and NAD at the deployment**

```
spec:
template:
  metadata:
    annotations:
      k8s.v1.cni.cncf.io/networks: '[{ "name" : "test-nad", "namespace": "ss-smf",
        "default-route": ["172.19.56.254"]}]'
```

*** verify the pod net1**

```
[root@bastion01 sehun]# oc exec -it -n ss-smf nf-dnsintf-6bf5ddf4db-2rf5t -- ifconfig
Defaulted container "ndns" out of: ndns, ima
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1400
        inet 10.130.3.113  netmask 255.255.254.0  broadcast 10.130.3.255
        inet6 fe80::858:aff:fe82:371  prefixlen 64  scopeid 0x20<link>
        ether 0a:58:0a:82:03:71  txqueuelen 0  (Ethernet)
        RX packets 3861  bytes 2151440 (2.0 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 3753  bytes 405850 (396.3 KiB)
        TX errors 0  dropped 1 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 0  bytes 0 (0.0 B)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 0  bytes 0 (0.0 B)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

net1: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 172.19.56.50  netmask 255.255.255.0  broadcast 172.19.56.255
        inet6 fe80::acfc:dcff:fe3b:918  prefixlen 64  scopeid 0x20<link>
        ether ae:fc:dc:3b:09:18  txqueuelen 0  (Ethernet)
        RX packets 15  bytes 1180 (1.1 KiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 8  bytes 628 (628.0 B)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

*** verify the pod route**

```
[root@bastion01 sehun]# oc exec -it -n ss-smf nf-dnsintf-6bf5ddf4db-2rf5t -- route -n
Defaulted container "ndns" out of: ndns, ima
Kernel IP routing table
Destination     Gateway         Genmask         Flags Metric Ref    Use Iface
0.0.0.0         172.19.56.254   0.0.0.0         UG    0      0        0 net1
10.128.0.0      10.130.2.1      255.252.0.0     UG    0      0        0 eth0
10.130.2.0      0.0.0.0         255.255.254.0   U     0      0        0 eth0
172.19.56.0     0.0.0.0         255.255.255.0   U     0      0        0 net1
172.30.0.0      10.130.2.1      255.255.0.0     UG    0      0        0 eth0
```

*** curl test / check tcpdump at the worker01**

```
[root@bastion01 sehun]# oc exec -it -n ss-smf nf-dnsintf-6bf5ddf4db-2rf5t -- curl -v http://172.19.56.110:80/aaa
Defaulted container "ndns" out of: ndns, ima
Trying 172.19.56.110...
TCP_NODELAY set
Connected to 172.19.56.110 (172.19.56.110) port 80 (#0)
...
```

*** check tcpdump at the worker01**

```
[root@worker01 /]# tcpdump -nni any host 172.19.56.110
dropped privs to tcpdump
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on any, link-type LINUX_SLL (Linux cooked v1), capture size 262144 bytes
```

17:06:08.599733 ARP, Request who-has 172.19.56.110 tell 172.19.56.50, length 28
17:06:08.599747 ARP, Request who-has 172.19.56.110 tell 172.19.56.50, length 28
17:06:08.599733 ARP, Request who-has 172.19.56.110 tell 172.19.56.50, length 28
17:06:08.599834 ARP, Reply 172.19.56.110 is-at 76:02:cc:b1:8e:0b, length 46
17:06:08.599839 ARP, Reply 172.19.56.110 is-at 76:02:cc:b1:8e:0b, length 46
17:06:08.599845 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [S], seq 3865280153, win 29200, options [mss 1460,sackOK,TS val 2283102418 ecr 0,nop,wscale 7], length 0
17:06:08.599855 IP **172.19.56.50**.59216 > 172.19.56.110.80: Flags [S], seq 3865280153, win 29200, options [mss 1460,sackOK,TS val 2283102418 ecr 0,nop,wscale 7], length 0 **<---------- 172.19.56.99 is not SNATed and the pod ip is seen**
17:06:08.599909 IP 172.19.56.110.80 > 172.19.56.50.59216: Flags [S.], seq 4225441355, ack 3865280154, win 28960, options [mss 1460,sackOK,TS val 1140493335 ecr 2283102418,nop,wscale 7], length 0
17:06:08.599913 IP 172.19.56.110.80 > 172.19.56.50.59216: Flags [S.], seq 4225441355, ack 3865280154, win 28960, options [mss 1460,sackOK,TS val 1140493335 ecr 2283102418,nop,wscale 7], length 0
17:06:08.599930 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [.], ack 1, win 229, options [nop,nop,TS val 2283102418 ecr 1140493335], length 0
17:06:08.599934 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [.], ack 1, win 229, options [nop,nop,TS val 2283102418 ecr 1140493335], length 0
17:06:08.600062 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [P.], seq 1:81, ack 1, win 229, options [nop,nop,TS val 2283102418 ecr 1140493335], length 80: HTTP: GET /aaa HTTP/1.1
17:06:08.600067 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [P.], seq 1:81, ack 1, win 229, options [nop,nop,TS val 2283102418 ecr 1140493335], length 80: HTTP: GET /aaa HTTP/1.1
17:06:08.600086 IP 172.19.56.110.80 > 172.19.56.50.59216: Flags [.], ack 81, win 227, options [nop,nop,TS val 1140493335 ecr 2283102418], length 0
17:06:08.600087 IP 172.19.56.110.80 > 172.19.56.50.59216: Flags [.], ack 81, win 227, options [nop,nop,TS val 1140493335 ecr 2283102418], length 0
17:06:08.601574 IP 172.19.56.110.80 > 172.19.56.50.59216: Flags [P.], seq 1:104, ack 81, win 227, options [nop,nop,TS val 1140493336 ecr 2283102418], length 103: HTTP: HTTP/1.1 404 Not Found
17:06:08.601577 IP 172.19.56.110.80 > 172.19.56.50.59216: Flags [P.], seq 1:104, ack 81, win 227, options [nop,nop,TS val 1140493336 ecr 2283102418], length 103: HTTP: HTTP/1.1 404 Not Found
17:06:08.601582 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [.], ack 104, win 229, options [nop,nop,TS val 2283102420 ecr 1140493336], length 0
17:06:08.601588 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [.], ack 104, win 229, options [nop,nop,TS val 2283102420 ecr 1140493336], length 0
17:06:08.601669 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [F.], seq 81, ack 104, win 229, options [nop,nop,TS val 2283102420 ecr 1140493336], length 0
17:06:08.601671 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [F.], seq 81, ack 104, win 229, options [nop,nop,TS val 2283102420 ecr 1140493336], length 0
17:06:08.601872 IP 172.19.56.110.80 > 172.19.56.50.59216: Flags [F.], seq 104, ack 82, win 227, options [nop,nop,TS val 1140493337 ecr 2283102420], length 0
17:06:08.601877 IP 172.19.56.110.80 > 172.19.56.50.59216: Flags [F.], seq 104, ack 82, win 227, options [nop,nop,TS val 1140493337 ecr 2283102420], length 0
17:06:08.601887 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [.], ack 105, win 229, options [nop,nop,TS val 2283102420 ecr 1140493337], length 0
17:06:08.601895 IP 172.19.56.50.59216 > 172.19.56.110.80: Flags [.], ack 105, win 229, options [nop,nop,TS val 2283102420 ecr 1140493337], length 0

Always appreciate your help.
Regards,
Sehee

On Wed, Feb 8, 2023 at 12:15 AM Andreas Karis <akaris@redhat.com> wrote:

> Hi Sehee,
>
> Do you and Samsung need any further help here, or does it actually work like I described? I'm happy to help if there are further issues with this :-)
>
> Thanks,
>
> Andreas
>
> On Thu, Feb 2, 2023 at 5:18 AM Sehee Oh <soh@redhat.com> wrote:
>
> > I really appreciate your work and it helps a lot! 🙏
> > Let me communicate with my CU.

Repectfully yours,
Sehee

2023년 2월 2일 (목) 오전 3:50, Andreas Karis <akaris@redhat.com>님이 작성:

> Follow-up: I just retested this with 4.10.45 and ipMasq works there, as well (the pod IP fc00:123::20 is correctly NATted to the bridge IP fc00:123::10). Both of my labs are SNO.
>
> I'd suggest the following next steps:
> a. Compare your setup to my lab, and provide more documentation in case the setups diverge. If the setups are similar, it should work, so look for configuration issues.
> b. If you have access to a reproducer lab, let's see if we can synchronize and have a quick look at this together
> c. There's still the option to use Step 2. from my earlier email, but given that I cannot reproduce your issue, I have a feeling that Step 2. is not needed.
> d. If after all of this we determine that this is a legit issue, let's see if we can work around it through configuration or what's actually missing to make it work
>
> - Andreas
>
> On Wed, Feb 1, 2023 at 7:13 PM Andreas Karis <akaris@redhat.com> wrote:
>
> > Hello,
> >
> > I read through the bugzilla and the jira and then I tried to reproduce your issue on OCP 4.12. I still have to test this on 4.10.
> >
> > My lab setup and test results are here:
> > https://github.com/andreaskaris/openshift-ipmasq-bridge
> > As you can see, on 4.12, this works out of the box, without loading any additional modules - I have a strong feeling that the result will be the same on 4.10, I will confirm this though. I also did the "negative" test and disabled ipMasq, and as expected, the pod IP was not masqueraded.
> >
> > So unless there is a surprising difference between 4.10 and 4.12, I suppose that my lab setup diverges from what you are trying, or you have another issue in your setup.
> >
> > Step 1: It would be great to get a comparable document including your lab layout and especially some kind of diagram (it can be an ASCII drawing) of what you want to achieve, in case that my reproducer is not correctly mimicking what you are looking for. Or maybe something is not configured correctly in your setup, and that's why it's failing for you.
> >
> > Step 2: Additionally, let's suppose that you needed the br_netfilter module (on 4.12, it seems that you don't). In the original bugzilla, you were asked to load the br_netfilter module and to test with it:
> > https://bugzilla.redhat.com/show_bug.cgi?id=2107083#c4 and https://bugzilla.redhat.com/show_bug.cgi?id=2107083#c10
> > You can easily load additional modules with the MachineConfigOperator, see: https://github.com/andreaskaris/openshift-ipmasq-bridge/blob/master/modprobe.yaml
> > This may or may not require a support exception, my gut feeling would tell me that this is shipped with CoreOS, and the change is small enough to state that this is a simple configuration, and thus it might not even require a support exception, etc. But either way, I would recommend loading the module with the MachineConfigOperator and checking if it fixes the issue for you.
> >
> > I hope this helps as a starting point. I will test this with 4.10 and if you want to I'm also available for having a look at this in your lab.
> >
> > - Andreas
> >
> >
> >
> > On Wed, Feb 1, 2023 at 2:33 PM Andres Guzman <aguzman@redhat.com> wrote:
> >
> > > + @Andreas Karis
> > >
> > > On Wed, Feb 1, 2023 at 7:56 AM Einat Glottmann <eglottma@redhat.com> wrote:
> > >
> > > > + @Franck Baudin @Marc Curry
> > > >
> > > > On Wed, Feb 1, 2023 at 7:50 AM Shienna Kim <shkim@redhat.com> wrote:

+Andres Guzman Gimenez +Inho Cho FYI only as it involves RFE. +Edu Alcaniz Just to seek your help when you are avail. Thank you.

On Wed, Feb 1, 2023 at 2:47 PM Sehee Oh <soh@redhat.com> wrote:

> Hello, team.
> While I was preparing for the regular TAM call tomorrow, I was wondering about bz 2107083 status.
>
> While I was preparing for the TAM call tomorrow, I remembered that bz2107083[1] had closed. Could you please tell me why this bz is closed? Will the bz be opened again once the RFE is done?
> [1] https://bugzilla.redhat.com/show_bug.cgi?id=2107083
>
> I opened the RFE[2] because the missing feature was blocking the bug. However, so far, there is no progress for the RFE, and closing the bz looks like giving up.
> [2] https://issues.redhat.com/browse/RFE-3503
>
> By the way, there is an official document stating the ipmasq flag can be used but in fact, can't. I think this is not just a simple RFE.
> [3] https://docs.openshift.com/container-platform/4.12//networking/multiple_networks/configuring-additional-network.html

Table 2. Bridge CNI plugin JSON configuration object

Field	Type	Description
`cniVersion`	`string`	The CNI specification version. The `0.3.1` value is required.
`name`	`string`	The value for the `name` parameter you provided previously for the CNO configuration.
`type`	`string`	
`bridge`	`string`	Specify the name of the virtual bridge to use. If the bridge interface does not exist on the host, it is created. The default value is `cni0` .
`ipam`	`object`	The configuration object for the IPAM CNI plugin. The plugin manages IP address assignment for the attachment definition.
`ipMasq`	`boolean`	Set to `true` to enable IP masquerading for traffic that leaves the virtual network. The source IP address for all traffic is rewritten to the bridge's IP address. If the bridge does not have an IP address, this setting has no effect. The default value is `false` .

> I can't give them the impression that RH is giving up this issue.
> Could you tell me the current status for this bz?
>
> Thank you.
> Regards,
>
> **오세희 | Sehee Oh**
> Technical Account Manager
> Red Hat
> **M:** +82 10 7374 3199
> **E:** soh@redhat.com

--
Thank you.
Regards,

오세희 | Sehee Oh
Technical Account Manager
Red Hat
M: +82 10 7374 3199

E: soh@redhat.com

--
Thank you.
Regards,

오세희 | Sehee Oh
Technical Account Manager
Red Hat
M: +82 10 7374 3199
E: soh@redhat.com

--

Douglas K Smith
Principal Software Engineer, OpenShift
Red Hat
@dougbtv (GitHub / Freenode)