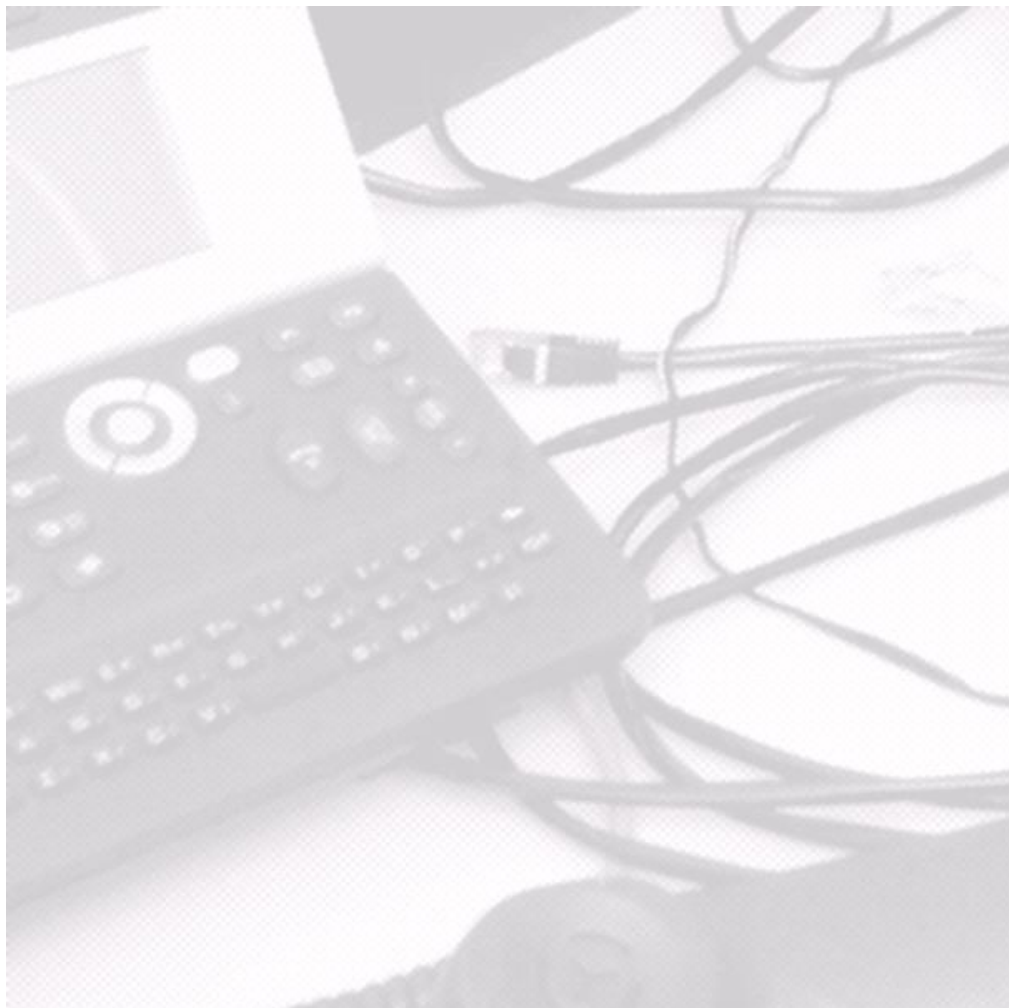


# ZSUT

Zakład Sieci i Usług Teleinformatycznych

## Kubernetes laboratory



## Lab part 2: Kubernetes networking

Prepared by: dr inż. Dariusz Bursztynowski

**ZSUT.** Zakład Sieci i Usług Teleinformatycznych  
Instytut Telekomunikacji  
Wydział Elektroniki i Technik Informacyjnych  
Politechnika Warszawska

Last update: March 2023

## Table of contents

1. Introduction .....	3
2. Kubernetes networking with flannel CNI .....	3
2.1. Checking the correctness of flannel installation .....	3
2.2. Basic pod communication checks (subjective selection).....	3
2.3. Checking routing settings in cluster nodes.....	5
2.4. Checking inter-container connectivity.....	11
2.4.1. Between containers inside the same Pod .....	11
2.4.2. Between containers in different Pods inside the same node.....	14
2.4.3. Between containers in different nodes .....	14
3. Installing MetalLB from manifest files .....	21
3.1. About MetalLB.....	21
3.2. Installation of MetalLB.....	22
3.3. Checking if MetalLB has been installed .....	23
3.4. Configuring IP address pool for MetalLB .....	23
4. Checking Traefik with external IP and basic load balancing.....	24
4.1. Checking the access to Traefik dashboard .....	26
4.2. Checking basic (in-cluster) load balancing.....	27
4.3. Closing remarks - analysing the taint in the master for metallb speakers .....	29
5. Additional readings/videos and hints.....	31

## 1. Introduction

In this part of the lab we focus on selected aspects of Kubernetes networking. In particular, we will investigate the operation of flannel CNI, install and configure MetalLB load balancer, check its operation based on Traefik<sup>1</sup> ingress controller and enabling external access to its dashboard, and analyse the role of kube-proxy by observing the operation of in-cluster load balancing (i.e., balancing among pod replicas of a given service).

## 2. Kubernetes networking with flannel CNI

In this section, we investigate the details of k3s networking using flannel as the CNI in use. Flannel Intro on Kubernetes networking can be found here: <https://www.tigera.io/learn/guides/kubernetes-networking/>.

In the following, we first quickly check if flannel has been installed correctly and then run a series of experiments to see the role of flannel in typical scenarios of pod communication.

### 2.1. Checking the correctness of flannel installation

Kubernetes defines a network model called the container network interface (CNI), but the actual implementation relies on network plugins. The network plugin is responsible for allocating internet protocol (IP) addresses to pods and enabling pods to communicate with each other within the Kubernetes cluster. There are a variety of network plugins for Kubernetes, but this section will use Flannel which is the default CNI in Kubernetes and for k3s in particular. Flannel is very simple and uses a Virtual Extensible LAN (VXLAN) overlay by default.

#### a) Checking k3s and flannel versions

```
ubuntu@kpi091:~$ k3s -v
k3s version v1.24.3+k3s1 (990ba0e8)
go version go1.18.1

ubuntu@kpi091:~$ /var/lib/rancher/k3s/data/*/bin/flannel
CNI Plugin flannel version v0.18.1 (linux/arm64) commit
990ba0e88c90f8ed8b50e0ccd375937b841b176e built on 2022-07-19T01:08:03Z
```

The above confirms flannel has been installed. Note: flannel version check shown above works from k3s version v1.23.7-rc1+k3s1.

#### b) Check if DNS is working

Use this link: <https://kubernetes.io/docs/tasks/administer-cluster/dns-debugging-resolution/>

### 2.2. Basic pod communication checks (subjective selection)

#### a) K3s/Flannel networking models:

- <https://www.sysspace.net/post/kubernetes-networking-explained-flannel-network-model>
- <https://mvallim.github.io/kubernetes-under-the-hood/documentation/kube-flannel.html>
- <https://www.henrydu.com/2020/11/16/k3s-cni-flannel/>
  - about container-shim: <https://iximiuz.com/en/posts/implementing-container-runtime-shim/>
  - about CRI, containerd, dockerd, ctr and crictl: <https://iximiuz.com/en/posts/containerd-command-line-clients/>

---

<sup>1</sup> Traefik is the default ingress controller in K3s.

Notice that flannld in k3s is a daemon, not “bridge” or “ovs bridge” and thus is not shown with brctl commands and alike. It connects to eth0 of the host and creates flannel.1 veth.

b) List bridges on a node: `brctl show`

c) Useful commands. Listing containers in a pod of a given Kubernetes namespace and inspecting pod/container internals:

- list containers in a pod  
`kubect1 get pod -n kube-system -o="custom-columns=NAME:.metadata.name,INIT-CONTAINERS:.spec.initContainers[*].name,CONTAINERS:.spec.containers[*].name"`
- list both running and stopped containers on a given cluster node (you have to ssh):
  - `sudo ctr container list`
  - or similar (but not equal) effect
    - `sudo crictl ps -a`

Note: *ctr* is a command-line client shipped as part of the containerd project, while *crictl* is a command-line client for [Kubernetes] [CRI-compatible container runtimes](#)).

- list *network-namespaces* on a given cluster node (each pod corresponds to a distinct net-namespace). Note: this works with containerd, but may return no results (an empty list) with Docker as the container engine.  
`ip netns list`
- list the IP addresses of the interfaces in a given network-namespace on a given k3s node (ssh-ed):  
`sudo ip netns exec cni-061c7b42-d791-f458-1abf-15b29536d510 ip addr show`
  - # alternatively use the command (no address info will be shown) ... `ip link show`
- For CRI-compatible container managers (as containerd): display help info, list containers, list pods on a given cluster node, run commands in a container to check current the status of the container  
`sudo crictl help`  
`sudo crictl ps`  
`sudo crictl pods`

```
# run a command in a container
ubuntu@kpi091:~$ sudo crictl exec a60432cef5826 ip addr show
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever
2: eth0@if8: <BROADCAST,MULTICAST,UP,LOWER_UP,M-DOWN> mtu 1450 qdisc noqueue state UP
    link/ether e2:ce:32:e2:00:15 brd ff:ff:ff:ff:ff:ff
    inet 10.42.0.34/24 brd 10.42.0.255 scope global eth0
        valid_lft forever preferred_lft forever
    inet6 fe80::e0ce:32ff:fee2:15/64 scope link
        valid_lft forever preferred_lft forever
# check/inspect current status of a container (below, we grep for the pid of the container)
ubuntu@kpi091:~$ sudo crictl inspect <contained-id> | grep pid
"pid": 3052,
      "pid": 1
      "type": "pid"
```

d) List *veth* pairs in a cluster node.

- To list *veth* pairs on a server, install (git clone + make) this application:  
<https://github.com/t1anz0ng/iftree> (installing *make* and *go-lang* will be necessary if these missing).

The following link shows how to make the exe file and then copy it to the remaining cluster nodes: <https://askubuntu.com/questions/903038/how-to-copy-or-move-files-from-remote-machine-to-local-machine>. In the simplest case, just run the command `$ sudo ./iftree` in the cloned directory (where the `make` command was run).

The veth pairs on the master can look like in the frame shown below. Notice: this tool cancels out the suffix part `@ifX` in the interface names (cf. example below where the meaning of the suffix is explained). Notice the tool also shows the name of the network namespace (prefixed with `cni0-`) where the `eth0` interface resides.

```
ubuntu@kpi091:~/tools/iftree$ sudo ./iftree

      Bridge <----> veth pair
— cni0 up
  └─ /var/run/netns/cni-061c7b42-d791-f458-1abf-15b29536d510
     └─ vethaecbf8cb eth0
  └─ /var/run/netns/cni-8f8f3060-086f-a2cb-cef5-d1521d118a4d
     └─ veth0cfce568 eth0
  └─ /var/run/netns/cni-a3f385e2-6118-8b1f-e40f-eec831646e68
     └─ vethcefe6b58 eth0
```

- Alternatively, combining `veth` pairs is possible based on `ip link show` in respective network namespaces:
  - Check IPs in one net-namespace, e.g. do `ip link show` in the root ns; then you will see something like this for one of the interfaces (notice the suffix `@ifX` in the interface name):
 

```
6: vethaecbf8cb@if2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1450 qdisc noqueue master cni0 state UP mode DEFAULT group default
```
  - Do the same in some other net-namespace; then for one of the displayed interfaces we can see something like this (below, `cni-061c7b42-d791-f458-1abf-15b29536d510` is net-namespace name; it can be found in the `iftree` output shown in the frame above):
 

```
sudo ip netns exec cni-061c7b42-d791-f458-1abf-15b29536d510 ip link show
2: eth0@if6: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1450 qdisc noqueue state UP mode DEFAULT group default link/ether 6a:b8:44:71:ac:d3 brd ff:ff:ff:ff:ff:ff link-netnsid 0
```

We can infer a given `veth` pair based on the correspondence of link indices X in the suffixes `@ifX` as bolded in the example.
  - Note: `ethtool` can also be used.

## 2.3. Checking routing settings in cluster nodes

We will inspect the routing starting from reviewing the settings of pods and containers and then move on to the settings of routing on node and flannel level.

- Checking the pods running on a given cluster node (`kpi091` in this case) in `kubectl` and checking pods' interfaces on that node

### Checking node pods using `kubectl`

```
xubuntu@xubulab:~$ kubectl get pods --all-namespaces -o wide --field-selector spec.nodeName=kpi091
NAMESPACE   NAME                                     READY   STATUS    RESTARTS
AGE  IP           NODE      NOMINATED NODE   READINESS GATES
kube-system  helm-install-traefik-crd-ddqhm         0/1     Completed  0
50d   10.42.0.4    kpi091    <none>           <none>
```

kube-system	helm-install-traefik-bxjnb	0/1	Completed	1
50d	10.42.0.2 kpi091 <none> <none>			
kube-system	svclb-traefik-d3b37a72-d5pgj	2/2	Running	12 (16h ago)
50d	10.42.0.26 kpi091 <none> <none>			
kube-system	local-path-provisioner-7b7dc8d6f5-zpl8h	1/1	Running	7 (16h ago)
50d	10.42.0.28 kpi091 <none> <none>			
kube-system	metrics-server-668d979685-g6d5s	1/1	Running	6 (16h ago)
50d	10.42.0.27 kpi091 <none> <none>			

*Note: similar results can be obtained using `sudo crictl pods` run directly on the cluster node.*

***Node: mapping from container itf => container and pod name.***

- Here we display pod interfaces (suffix -ethX or -lo for veth and loopback port, respectively)

```
ubuntu@kpi091:~$ sudo ls /var/lib/cni/results -al
total 32
drwx----- 2 root root 4096 Sep 17 12:57 .
drwx----- 5 root root 4096 Jul 28 15:49 ..
-rw----- 1 root root 1543 Sep 17 12:57 cbr0-
08fd8a402c5182a78fbb324d4fc997ec8175alecf338187d801429de4fc4ccb1-eth0
-rw----- 1 root root 1679 Sep 17 12:57 cbr0-
271c1cd19eca72d4c4c8884e1eff1a3346d4c4c0c0d9d58ca82d733921ac9e22-eth0
-rw----- 1 root root 1535 Sep 17 12:57 cbr0-
7dd8b441a1527173820a9e53b93e077dfc2ed4d633ded37ac224f7391efbelba-eth0
-rw----- 1 root root 1126 Sep 17 12:57 cni-loopback-
08fd8a402c5182a78fbb324d4fc997ec8175alecf338187d801429de4fc4ccb1-lo
-rw----- 1 root root 1262 Sep 17 12:57 cni-loopback-
271c1cd19eca72d4c4c8884e1eff1a3346d4c4c0c0d9d58ca82d733921ac9e22-lo
-rw----- 1 root root 1118 Sep 17 12:57 cni-loopback-
7dd8b441a1527173820a9e53b93e077dfc2ed4d633ded37ac224f7391efbelba-lo
```

- Here, we retrieve and check network configuration data for one pod to see its final network settings. We use the pod's id (cbr0-prefixed and eth0-suffixed) we retrieved by using the command shown above

```
ubuntu@kpi091:~$ sudo cat /var/lib/cni/results/cbr0-08fd8a402c5182a78fbb324d4fc997ec8175a1ecf338187d801429de4fc4ccb1-eth0 | jq
{
  "kind": "cniCacheV1",
  "containerId": "08fd8a402c5182a78fbb324d4fc997ec8175a1ecf338187d801429de4fc4ccb1",
  "config": REDACTED,
  "ifName": "eth0",
  "networkName": "cbr0",
  "cniArgs": [
    [
      "IgnoreUnknown",
      "1"
    ],
    [
      "K8S_POD_NAMESPACE",
      "kube-system"
    ],
    [
      "K8S_POD_NAME",
      "local-path-provisioner-7b7dc8d6f5-zpl8h"
    ],
    [
      # first part of this id is shown as POD ID in "sudo crictl pods" and "sudo crictl ps"
      "K8S_POD_INFRA_CONTAINER_ID",
      "08fd8a402c5182a78fbb324d4fc997ec8175a1ecf338187d801429de4fc4ccb1"
    ],
    [
      "K8S_POD_UID",
      "0125992b-e574-4354-a765-6ce9b87243fd"
    ]
  ],
  "capabilityArgs": {
    "dns": {
      "Servers": [
        "10.43.0.10"
      ],
      "Searches": [
        "kube-system.svc.cluster.local",
        "svc.cluster.local",
        "cluster.local",
        ""
      ],
      "Options": [
        "ndots:5"
      ]
    },
    "io.kubernetes.cri.pod-annotations": {
      "kubernetes.io/config.seen": "2022-09-17T12:57:49.040662889Z",
      "kubernetes.io/config.source": "api"
    }
  },
  "result": {
    "cniVersion": "1.0.0",
    "dns": {},
    "interfaces": [
      { # this is the cni0 bridge itself (check by running ip link show master cni)
        "mac": "c6:76:6d:5c:2c:43",
        "name": "cni0"
      },
      { # this is veth interface in cni0 bridge (check by running ip link show master cni)
        "mac": "da:42:40:a4:2f:77",
        "name": "vetha7541b74"
      },
      { # this is veth interface in the pod namespace
        "mac": "36:e6:d4:9d:18:57",

```

```

    "name": "eth0",
    "sandbox": "/var/run/netns/cni-1984eda9-38ae-66cb-c817-34f8ec6a3ea8"
  },
],
"ips": [
  { # this is pod's cluster-level IP address (all its containers will share this address)
    "address": "10.42.0.28/24",
    "gateway": "10.42.0.1",
    "interface": 2
  }
],
"routes": [
  {
    "dst": "10.42.0.0/16"
  },
  {
    "dst": "0.0.0.0/0",
    "gw": "10.42.0.1"
  }
]
}
}
ubuntu@kpi091:~$

```

- Also, cni0 bridge configuration for its particular veth interfaces (per **pod** network-namespace) is stored in folder `/var/lib/cni/flannel/` and this information is used to create cni0. In particular, it instructs that cni0 should be of type *bridge* and that cni0 should use the *host-local* IPAM for assigning IP addresses to pods (the range of IP addresses for use by IPAM is given by the attribute *subnet*).

```

# list cni0 information for veth interfaces in cni0
ubuntu@kpi091:~$ sudo ls /var/lib/cni/flannel
08fd8a402c5182a78fbb324d4fc997ec8175a1ecf338187d801429de4fc4ccb1
271c1cd19eca72d4c4c8884e1eff1a3346d4c4c0c0d9d58ca82d733921ac9e22
7dd8b441a1527173820a9e53b93e077dfc2ed4d633ded37ac224f7391efbe1ba

# list cni0 information for a selected veth
kpi091:~$ sudo cat
/var/lib/cni/flannel/08fd8a402c5182a78fbb324d4fc997ec8175a1ecf338187d801429de4fc4ccb1 | jq
{
  "cniVersion": "1.0.0",
  "forceAddress": true,
  "hairpinMode": true,
  "ipMasq": false,
  "ipam": {
    "ranges": [
      [
        {
          "subnet": "10.42.0.0/24"
        }
      ]
    ],
    "routes": [
      {
        "dst": "10.42.0.0/16"
      }
    ]
  },
  "type": "host-local"
},
"isDefaultGateway": true,
"isGateway": true,
"mtu": 1450,
"name": "cbr0",
"type": "bridge"
}
ubuntu@kpi091:~$

```



- b) Check routing settings in a given pod. Notice default gateway is set to 10.42.0.1 which is consistent with what we can find in the table above.

```
xubuntu@xubulab:~$ kubectl exec -it -n kube-system local-path-provisioner-7b7dc8d6f5-zpl8h -- /bin/sh
/ # ip addr show
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever
2: eth0@if7: <BROADCAST,MULTICAST,UP,LOWER_UP,M-DOWN> mtu 1450 qdisc noqueue state UP
    link/ether 46:8d:d4:ef:84:7c brd ff:ff:ff:ff:ff:ff
    inet 10.42.0.48/24 brd 10.42.0.255 scope global eth0
        valid_lft forever preferred_lft forever
    inet6 fe80::448d:d4ff:feef:847c/64 scope link
        valid_lft forever preferred_lft forever

/ # ip route sh
default via 10.42.0.1 dev eth0
10.42.0.0/24 dev eth0 scope link src 10.42.0.48
10.42.0.0/16 via 10.42.0.1 dev eth0
/ #
```

- c) Check flannel settings – run the following commands in a selected cluster node (here, we used our master)

```
# show vxlan devices
ubuntu@kpi091:~$ ip link show type vxlan
4: flannel.1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1450 qdisc noqueue state UNKNOWN mode DEFAULT
group default
    link/ether 5e:14:d0:84:69:81 brd ff:ff:ff:ff:ff:ff

# show flannel.1 VXLAN details;
# notice the IP of the cluster node and assigning flannel.1 device to eth0 of the cluster node
ubuntu@kpi091:~$ ip -d a show flannel.1
4: flannel.1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1450 qdisc noqueue state UNKNOWN group default
    link/ether be:b1:aa:6b:74:85 brd ff:ff:ff:ff:ff:ff promiscuity 0 minmtu 68 maxmtu 65535
    vxlan id 1 local 192.168.1.38 dev eth0 srcport 0 0 dstport 8472 nolearning ttl auto ageing
300 udpcsum noudp6zerocsumtx noudp6zerocsumrx numtxqueues 1 numrxqueues 1 gso_max_size 65536
gso_max_segs 65535
    inet 10.42.0.0/32 scope global flannel.1
        valid_lft forever preferred_lft forever
    inet6 fe80::bcb1:aaff:fe6b:7485/64 scope link
        valid_lft forever preferred_lft forever

# check the ARP table on kpi091 created by flannel (mapping IP/MAC of flannel interfaces on
remaining nodes)
# PERMANENT flag denotes that the entry is permanent and flannel.1 device needs not broadcast ARP
queries "who has"
ubuntu@kpi091:~$ ip neigh show dev flannel.1
10.42.2.0 lladdr 2e:cb:94:23:86:79 PERMANENT
10.42.1.0 lladdr 36:84:0d:94:d6:34 PERMANENT
10.42.3.0 lladdr 9a:29:f6:23:6f:d9 PERMANENT

# check the forwarding data base for flannel.1 vxlan device; mind we are on node 192.168.1.38
ubuntu@kpi091:~$ bridge fdb show dev flannel.1
6a:36:f5:77:50:0f dst 192.168.1.40 self permanent
e6:b9:9b:1b:ee:63 dst 192.168.1.41 self permanent
36:84:0d:94:d6:34 dst 192.168.1.41 self permanent
2e:cb:94:23:86:79 dst 192.168.1.39 self permanent
ae:d5:67:1d:c4:2e dst 192.168.1.39 self permanent
9a:29:f6:23:6f:d9 dst 192.168.1.40 self permanent

# Nitty-gritty detail: flannel.1 does SNAT for packets from the outside of the flannel network to
cni0 (10.42.0.0/16). To disable SNAT edit 10-flannel.conflist and add '"ipMasq":false' after the
line '"isDefaultGateway":true,' and reboot. Below, SNAT is enabled.
```

```
# Check iptables on the node for flanneld
ubuntu@kpi091:~$ sudo iptables-save | grep flanneld
-A FORWARD -s 10.42.0.0/16 -m comment --comment "flanneld forward" -j ACCEPT
-A FORWARD -d 10.42.0.0/16 -m comment --comment "flanneld forward" -j ACCEPT
-A POSTROUTING -s 10.42.0.0/16 -d 10.42.0.0/16 -m comment --comment "flanneld masq" -j RETURN
-A POSTROUTING -s 10.42.0.0/16 ! -d 224.0.0.0/4 -m comment --comment "flanneld masq" -j MASQUERADE
--random-fully
-A POSTROUTING ! -s 10.42.0.0/16 -d 10.42.0.0/24 -m comment --comment "flanneld masq" -j RETURN
-A POSTROUTING ! -s 10.42.0.0/16 -d 10.42.0.0/16 -m comment --comment "flanneld masq" -j MASQUERADE
--random-fully
# Check current version of 10-flannel.conflist
ubuntu@kpi091:~$ sudo cat /var/lib/rancher/k3s/agent/etc/cni/net.d/10-flannel.conflist
{
  "name": "cbr0",
  "cniVersion": "1.0.0",
  "plugins": [
    {
      "type": "flannel",
      "delegate": {
        "hairpinMode": true,
        "forceAddress": true,
        "isDefaultGateway": true
      }
    },
    {
      "type": "portmap",
      "capabilities": {
        "portMappings": true
      }
    }
  ]
}
ubuntu@kpi091:~$
```

### Observations:

- The ~~striketrough~~ entries in *flannel.1* vxlan device forwarding data base in the table above indicate the entries that appeared in the list but which I could not associate with any existing object despite a careful inspection. Ideally they should not appear at all.
- Flannel interface *flannel* provides VXLAN named *flannel.1* and is attached to eth0 interface of the cluster node
- Flannel daemon flanneld has also created PERMANENT (non expiring) entries in the ARP table of the node. For example, the entry 10.42.2.0 lladdr 2e:cb:94:23:86:79 denotes IP/MAC addresses mapping for *flannel.1* interface in a remote cluster node (in this case this is *flannel.1* sitting on node kpi092). In other words, local ARP table informs that *flannel.1* interface with IP 10.42.2.0 is reachable within the flannel VXLAN on the MAC address 2e:cb:94:23:86:79. So, flannel.1 device will use this address as the MAC address in the inner Ethernet frame of the VXLAN. Also, it will use it for mapping onto the IP of the remote cluster node containing the VTEP to which VXLAN tunnel extends – see the following bullet point.
- Based on the above, flannel populates the forwarding data base of the node for flannel.1 vxlan device. For example, entry 2e:cb:94:23:86:79 dst 192.168.1.39 tells flannel.1 in node kpi091 that flannel.1 VXLAN port with MAC address 2e:cb:94:23:86:7 is available in the network on IP address 192.168.1.39.
- Notice that the two tables described above together provide a two-step mapping from the IP address of the destination pod subnetwork to the IP address of the cluster node where that pod resides.
- Check ip n or ip r
- Check for ARP: arp -a or arp -n

## 2.4. Checking inter-container connectivity

In this section we analyse the routing of packets between containers for three following scenarios: when the containers belong to same Pod, belong to different Pods running in same node, and are located in different nodes. Follow the instructions given below.

Based on this:

- <https://msazure.club/flannel-networking-demystify/>
- <https://www.sysspace.net/post/kubernetes-networking-explained-flannel-network-model>
- How containers inside a pod communicate: <https://www.redhat.com/sysadmin/kubernetes-pod-network-communications>
- How pods on different nodes communicate: <https://www.redhat.com/sysadmin/kubernetes-pods-communicate-nodes>
- Kubernetes network stack fundamentals: How pods on different nodes communicate: <https://www.redhat.com/sysadmin/kubernetes-pods-communicate-nodes>
- Capture packets in Kubernetes: <https://www.redhat.com/sysadmin/capture-packets-kubernetes-ksniff>
- Kubernetes troubleshooting: <https://www.redhat.com/sysadmin/kubernetes-troubleshooting>
- Check flannel forwarding base (on cluster node): `$ bridge fdb show dev flannel.1`
- <https://serverfault.com/questions/988736/possible-to-list-members-of-a-network-bridge>
- Digging into linux namespaces:  
<https://blog.quarkslab.com/digging-into-linux-namespaces-part-1.html>  
<https://blog.quarkslab.com/digging-into-linux-namespaces-part-2.html>

### 2.4.1. Between containers inside the same Pod

For this scenario, we create a pod with two containers. The nginx container serves a default web page, and the busybox container sleeps indefinitely ready to ssh to it and run shell commands, e.g.. curl to nginx. These two images are often used in Kubernetes courses for learning Kubernetes networking. Some of the steps in this section are repetitions of the steps from section 2.3. However, while the description in section 2.3 was assumed to be general, currently we analyze the network settings in the context of a particular instance of communication between pods.

a) Create a manifest file of a pod, say nginx-busybox.yaml, with the following contents and run it:

```
# manifest file contents
apiVersion: v1
kind: Pod
metadata:
  name: nginx-busybox
spec:
  containers:
  - command:
    - sleep
    - infinity
    image: busybox
    name: busybox
  - image: nginx
    name: nginx

xubuntu@xubulab:~/cluster-pi/manifests/lab$ kubectl apply -f nginx-busybox.yaml
pod/nginx-busybox created
xubuntu@xubulab:~/cluster-pi/manifests/lab$ kubectl get pods -o wide
```

NAME	READY	STATUS	RESTARTS	AGE	IP	NODE	NOMINATED NODE	READINESS GATES
dnsutils	1/1	Running	41 (38m ago)	8d	10.42.3.37	kpi093	<none>	<none>
nginx-busybox	2/2	Running	0	3m23s	10.42.1.30	kpi094	<none>	<none>

- b) Launch a shell for the busybox container and check if busybox can communicate over *localhost* with nginx (wget should be able to retrieve a document in HTML format).

```
xubuntu@xubulab:~/cluster-pi/manifests/lab$ kubectl exec -it -c busybox nginx-busybox -- /bin/sh

# check the connectivity over local host (both containers use port 80 for HTTP)
/ # wget localhost -O - 2>/dev/null
<!DOCTYPE html>
<html>
<head>
<title>Welcome to nginx!</title>
<style>
html { color-scheme: light dark; }
body { width: 35em; margin: 0 auto;
font-family: Tahoma, Verdana, Arial, sans-serif; }
</style>
</head>
<body>
<h1>Welcome to nginx!</h1>
<p>If you see this page, the nginx web server is successfully installed and
working. Further configuration is required.</p>

<p>For online documentation and support please refer to
<a href="http://nginx.org/">nginx.org</a>.<br/>
Commercial support is available at
<a href="http://nginx.com/">nginx.com</a>.</p>

<p><em>Thank you for using nginx.</em></p>
</body>
</html>
```

- c) You can do the following extra checks to see both containers run in the same network namespace but in different process namespaces. They can communicate (same network namespace), but do not see each other's process identifier (PID). The output shows busybox can see that "something" is listening on localhost:80 (0.0.0.0:80), but the PID of the listener (nginx in this case) is not visible to busybox.

On the other hand, if you do similar check when logged (kubectl exec) on nginx then you will notice nginx can see its own PID (equal 1). The latter is visible in the second part of the log in the table below.

```
# still on the busybox container
/ # netstat -tlpn
Active Internet connections (only servers)
Proto Recv-Q Send-Q Local Address           Foreign Address         State       PID/Program name
tcp        0      0 0.0.0.0:80              0.0.0.0:*               LISTEN      -
tcp        0      0 :::80                   :::*                     LISTEN      -

# do the same for nginx (exit, exec to nginx and install net-tools first)
/ # exit
xubuntu@xubulab:~/cluster-pi/manifests/lab$ kubectl exec -it -c nginx nginx-busybox -- /bin/bash
root@nginx-busybox:/# apt update
root@nginx-busybox:/# apt install net-tools

root@nginx-busybox:/# netstat -tlpn # or set -elnopt
Active Internet connections (only servers)
Proto Recv-Q Send-Q Local Address           Foreign Address         State       PID/Program name
tcp        0      0 0.0.0.0:80              0.0.0.0:*               LISTEN      1/nginx: master
pro
tcp6       0      0 :::80                   :::*                     LISTEN      1/nginx: master
pro
```

- d) Now we can take a look at the containers from the pod/network namespace perspective. To do that, exit the container and then `exec` our pod namespace following the scenario from the table below (part 1 in the table is optional and can be skipped).

```
# 1. optional part
# get pods (we are in default namespace of k3s)
xubuntu@xubulab:~/cluster-pi/manifests/lab$ kubectl get pod -o json nginx-busybox | jq

# find the following part of the description of the pod where (long) container IDs are given
...
  "containerStatuses": [
    {
      "containerID":
"containerd://24a0a0e95f8dd20bd60aa8bbe075e9ad0518217976f0243f1c16243dcec35458",
      "image": "docker.io/library/busybox:latest",
      "imageID":
"docker.io/library/busybox@sha256:ad9bd57a3a57cc95515c537b89aaa69d83a6df54c4050fcf2b41ad367bec0c
d5",
      "lastState": {},
      "name": "busybox",
      "ready": true,
      "restartCount": 0,
      "started": true,
      "state": {
        "running": {
          "startedAt": "2022-09-22T20:11:01Z"
        }
      }
    },
    {
      "containerID":
"containerd://67336c7aa0c399d106947e846651019b9adec5e980f14242761836ff0d1b9a4c",
      "image": "docker.io/library/nginx:latest",
      "imageID":
"docker.io/library/nginx@sha256:0b970013351304af46f322da1263516b188318682b2ab1091862497591189ff1
",
      "lastState": {},
      "name": "nginx",
      "ready": true,
      "restartCount": 0,
      "started": true,
      "state": {
        "running": {
          "startedAt": "2022-09-22T20:11:24Z"
        }
      }
    }
  ]
...

# 2. sufficient part
# identify the node where our nginx-busybox pod runs
xubuntu@xubulab:~/cluster-pi/manifests/lab$ kubectl get pods -o wide
NAME                READY   STATUS    RESTARTS   AGE   IP            NODE       NOMINATED NODE
READINESS GATES
nginx-busybox        2/2     Running   0           62m   10.42.1.30    kpi094     <none>
<none>
dnsutils             1/1     Running   42 (37m ago)  8d    10.42.3.37    kpi093     <none>
<none>

# ssh to the node where the pod runs and find container process IDs one by one
xubuntu@xubulab:~$ ssh kpi094
ubuntu@kpi094:~$ sudo crictl ps
CONTAINER          IMAGE                                     CREATED           STATE             NAME
ATTEMPT           POD ID                                POD
67336c7aa0c39      0c404972e1305                         About an hour ago Running           nginx
0                   d5635f85512eb                         nginx-busybox
24a0a0e95f8dd      410fde8b14eed                         About an hour ago Running           busybox
0                   d5635f85512eb                         nginx-busybox
```

90c00c4c3d34d	b12bbeclf4615	5 hours ago	Running	lb-tcp-443
13	fa440c15de16f	svclb-traefik-d3b37a72-mxqk4		
44c1dca6f0402	2ef507d0470ec	5 hours ago	Running	traefik
13	de5ac56b7f02a	traefik-7cd4fcff68-tg94h		
f636764728bbf	b12bbeclf4615	5 hours ago	Running	lb-tcp-80
13	fa440c15de16f	svclb-traefik-d3b37a72-mxqk4		

# get the PIDs of the containers in the root

```
ubuntu@kpi094:~$ sudo crictl inspect 67336c7aa0c39 | jq '["info"].pid'
```

```
6677
```

```
ubuntu@kpi094:~$ sudo crictl inspect 24a0a0e95f8dd | jq '["info"].pid'
```

```
6587
```

# obtain namespaces for each container process; it can be seen that both containers/processes \ # share network/ipc/uts namespaces (column NPROSC) while other namespaces like pid are different

```
ubuntu@kpi094:~$ sudo lsns -p 6677
```

NS	TYPE	NPROCS	PID	USER	COMMAND
4026531834	time	164	1	root	/sbin/init fixrtc splash
4026531837	user	164	1	root	/sbin/init fixrtc splash
4026532368	net	7	6549	65535	/pause
4026532602	uts	7	6549	65535	/pause
4026532603	ipc	7	6549	65535	/pause
4026532608	mnt	5	6677	root	nginx: master process nginx -g daemon off;
4026532609	pid	5	6677	root	nginx: master process nginx -g daemon off;
4026532610	cgroup	5	6677	root	nginx: master process nginx -g daemon off;

```
ubuntu@kpi094:~$ sudo lsns -p 6587
```

NS	TYPE	NPROCS	PID	USER	COMMAND
4026531834	time	163	1	root	/sbin/init fixrtc splash
4026531837	user	163	1	root	/sbin/init fixrtc splash
4026532368	net	7	6549	65535	/pause
4026532602	uts	7	6549	65535	/pause
4026532603	ipc	7	6549	65535	/pause
4026532605	mnt	1	6587	root	sleep infinity
4026532606	pid	1	6587	root	sleep infinity
4026532607	cgroup	1	6587	root	sleep infinity

### e) Conclusion

A single pod may seem like a very simple construct, but there is more going on under the hood of the Linux network stack that enables pods to function. Additionally, the ability for containers within a pod to communicate over the localhost address facilitates the use of **patterns as sidecar and init containers**.

### 2.4.2. Between containers in different Pods inside the same node

This case is left to the students for investigation as a **micro-project**. Please, elaborate on the communication to show relevant details by adopting selected mechanisms used in points 2.4.1 and 2.4.3.

### 2.4.3. Between containers in different nodes

This is the most complex case where one can observe flannel at work to a full extent. We will use two pods again, this time instantiated in different nodes, to check their routing settings and inspect protocol encapsulation during the communication of those pods.

#### a) Setting the environment

To set the environment we will create two pods, nginx and busybox<sup>2</sup>, this time placed in different nodes. To control pod placement we will use Kubernetes *nodeSelector* attribute that allows to specify the allowed set of cluster nodes where a pod can be instantiated (refer to Kubernetes documentation for a detailed description of *nodeSelector*). Follow the instruction from the box given below.

<sup>2</sup> It is recommended to set a unique name to each pod to make it distinguishable from other pods during experiments.

```
# Create manifest file with the following contents (adjust node names to your environment)

xubuntu@xubulab:~/cluster-pi/manifests/lab$ cat nginx-busybox-diffnodes.yaml
# Two pods will be created in predefined (different) cluster nodes

apiVersion: v1
kind: Pod
metadata:
  name: nginx
spec:
  nodeSelector:
    kubernetes.io/hostname: kpi092
  containers:
  - image: nginx
    name: nginx

---
apiVersion: v1
kind: Pod
metadata:
  name: busybox
spec:
  nodeSelector:
    kubernetes.io/hostname: kpi093
  containers:
  - command:
    - sleep
    - infinity
    image: busybox
    name: busybox

# Follow the commands to get the pods running and verify they have been created as expected

xubuntu@xubulab:~/cluster-pi/manifests/lab$ ls
nginx-busybox-diffnodes.yaml  nginx-busybox.yaml
xubuntu@xubulab:~/cluster-pi/manifests/lab$ kubectl get pods
NAME                READY   STATUS    RESTARTS   AGE
dnsutils            1/1     Running   49 (37h ago)    17d
nginx-busybox        2/2     Running   4 (37h ago)     8d
xubuntu@xubulab:~/cluster-pi/manifests/lab$ kubectl apply -f nginx-busybox-diffnodes.yaml
pod/nginx created
pod/busybox created
xubuntu@xubulab:~/cluster-pi/manifests/lab$ kubectl get pods -o wide
NAME                READY   STATUS    RESTARTS   AGE   IP            NODE       NOMINATED NODE
READINESS GATES
dnsutils            1/1     Running   49 (37h ago)    17d   10.42.3.43    kpi093    <none>         <none>
nginx-busybox        2/2     Running   4 (37h ago)     8d   10.42.1.36    kpi094    <none>         <none>
busybox             1/1     Running   0           64s   10.42.3.44    kpi093    <none>         <none>
nginx               1/1     Running   0           64s   10.42.2.19    kpi092    <none>         <none>
```

## b) Layer 2 settings

Later on in this exercise we will log to busybox and wget a file from nginx pod. First, let's check layer 2 configuration relevant to the communication between those pods.

### Settings for nginx pod

```
# Enter the nginx pod
xubuntu@xubulab:~$ kubectl exec -it nginx -- /bin/bash

# Install missing packages
root@nginx:/# apt update
Get:1 http://deb.debian.org/debian bullseye InRelease [116 kB]
...
root@nginx:/# apt install -y iproute2
Reading package lists...
...
root@nginx:/# apt install ethtool
```

```
Reading package lists...
```

```
...
```

#### # Check the network configurations on nginx pod

```
root@nginx:/# ip addr show
```

```
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever
2: eth0@if7: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1450 qdisc noqueue state UP group default
    link/ether 4a:ec:0e:72:6e:45 brd ff:ff:ff:ff:ff:ff link-netnsid 0
    inet 10.42.2.21/24 brd 10.42.2.255 scope global eth0
        valid_lft forever preferred_lft forever
    inet6 fe80::48ec:eff:fe72:6e45/64 scope link
        valid_lft forever preferred_lft forever
```

#### # eth0@if7 means that the remote end of the veth pair eth0 belongst to (on the host) is 7.

# it can be confirmed as follows with ethtool or from the root network namespace on the host

```
root@nginx:/# ethtool -S eth0
```

```
NIC statistics:
    peer_ifindex: 7
```

```
...
```

# confirming by checking network configuration on kpi092 (host for our nginx pod) (we skip remaining interfaces for brevity) - @if2 and peer\_ifindex: 2 relate to eth0 in our nginx pod

```
root@nginx:/# exit
```

```
exit
```

```
xubuntu@xubulab:~$ ssh kpi092
```

```
ubuntu@kpi092:~$ ip link show
```

```
...
```

```
7: veth3a7939d6@if2: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1450 qdisc noqueue master cni0 state UP
mode DEFAULT group default
    link/ether 72:12:d0:d0:b5:60 brd ff:ff:ff:ff:ff:ff link-netns cni-034fea82-9ee2-5bf7-eaba-3aaeebdfcbf7
```

```
ubuntu@kpi092:~$ ethtool -S veth3a7939d6
```

```
NIC statistics:
    peer_ifindex: 2
```

```
...
```

# The eth0-veth3a7939d6 virtual Ethernet pair connects to a bridge on the host. Once you accomplish the exercise in section 2.3 you will see this setting allows pods on the same host # to communicate directly with each other over the bridge. The cni0 bridge interface has an IP # address assigned to it, which will be important in the Layer 3 routing process:

```
ubuntu@kpi092:~$ brctl show
```

bridge name	bridge id	STP enabled	interfaces
cni0	8000.868c45b22963	no	vethlad1bd98

veth3a7939d6

```
ubuntu@kpi092:~$ ip -br addr show cni0
```

```
cni0      UP                10.42.2.1/24 fe80::848c:45ff:feb2:2963/64
```

the same

## Settings for busybox pod

They can be checked in an analogous way as above, except that entering the busybox pod has to be done using shell, .e.g., `kubectl exec -it busybox -- /bin/sh`.

### c) Layer 3 settings

L3 packet has to follow a path going from the source pod via cni0, then flannel.1 and eth0 on the source node, all the way down across the cluster network towards the eth0, flannel.1, cni0 and the target pod on the destination node. In the following we will check the routing settings of respective network elements in the direction from busybox to nginx.

First, the busybox pod must decide where to send traffic for the remote network (mind busybox is in 10.42.3./24 network while nginx address 10.42.2.21 is derived from 10.42.2.0/24 network) . The routing table in busybox pod includes a route matching 10.42.2.21 that has the mask 10.42.0.0/16, so busybox will send the traffic to cni0



bridge through the `eth0` interface of the pod (via `10.42.3.1 dev eth0`). Notice the default gateway for `busybox` is at `10.42.3.1`, which is the IP address of the `cni0` bridge on the host. All these can be seen below.

```
# routing table on the busybox pod
xubuntu@xubulab:~$ kubectl exec -it busybox -- /bin/sh
/ # ip route show
default via 10.42.3.1 dev eth0
10.42.0.0/16 via 10.42.3.1 dev eth0
10.42.3.0/24 dev eth0 scope link src 10.42.3.47
```

Once the traffic has reached `cni0` bridge, the next routing decision will determine how to forward that traffic to the desired network (`10.42.2.0/24`). The host's routing table shows that traffic destined to the `10.42.2.0/24` network will be sent via `10.42.2.0` on the `flannel.1` interface:

```
# routing table on node kpi093 (busybox host)
ubuntu@kpi093:~$ ip route sh
default via 192.168.1.1 dev eth0 proto dhcp src 192.168.1.40 metric 100
10.42.0.0/24 via 10.42.0.0 dev flannel.1 onlink
10.42.1.0/24 via 10.42.1.0 dev flannel.1 onlink
10.42.2.0/24 via 10.42.2.0 dev flannel.1 onlink
10.42.3.0/24 dev cni0 proto kernel scope link src 10.42.3.1
192.168.1.0/24 dev eth0 proto kernel scope link src 192.168.1.40 metric 100
192.168.1.1 dev eth0 proto dhcp scope link src 192.168.1.40 metric 100
```

Now, `flannel.1` interface is responsible for providing VXLAN service between `cni0` and peer `cni0` bridges in remaining cluster nodes. `flannel.1` interface description contains the basic parameters needed to create VXLAN tunnels to the remaining cluster nodes, i.e., `VXLAN Id = 1`, source (local) IP address for the outer IP packets with the tunneled traffic, and the destination port (`dstport`) of UDP packets that encapsulate VXLAN frames of the tunnels (same VXLAN id and UDP `dstport` number for all tunnels in the cluster):

```
# flannel.1 interface on node kpi093
ubuntu@kpi093:~$ ip -d a show flannel.1
4: flannel.1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1450 qdisc noqueue state UNKNOWN group
default
    link/ether 66:16:cf:19:e8:2c brd ff:ff:ff:ff:ff:ff promiscuity 0 minmtu 68 maxmtu 65535
    vxlan id 1 local 192.168.1.40 dev eth0 srcport 0 0 dstport 8472 nolearning ttl auto ageing
300 udpchecksum noudp6zerocsumtx noudp6zerocsumrx numtxqueues 1 numrxqueues 1 gso_max_size 65536
gso_max_segs 65535
    inet 10.42.3.0/32 scope global flannel.1
        valid_lft forever preferred_lft forever
    inet6 fe80::6416:cfff:fe19:e82c/64 scope link
        valid_lft forever preferred_lft forever
```

We know from section 2.3 that having received a packet from `cni0` bridge, `flannel.1` uses the ARP table on the host and the internal `flannel.1` routing table to map the destination IP address of the packet received from `cni0` (directed to a pod in another node).

We know from the routing table of node `kpi093` (see above) that the next hop for a packet from `busybox` to `nginx` is `10.42.2.0`. However, no directly reachable interface on the cluster network contains the `10.244.1.0` address. Instead, a static Address Resolution Protocol (ARP) entry exists on the ARP table on the host `kpi093`. This static entry indicates that the MAC address for `10.42.2.0` is `e6:4b:9c:85:29:63`. Then the `flannel.1` forwarding table (bridge forwarding database) directs traffic for this MAC address to a remote destination of `192.168.122.7`. This remote IP address, which is the physical interface on the `kpi092` node (hosting our `nginx` pod), is the other side of the VXLAN tunnel in the physical (cluster) network (actually, the VXLAN tunnel starts and terminates in `flannel.1` interfaces). One can check this as shown in the box presented below; notice the correspondence between the addresses in the bolded lines. Notice also the entries are permanent – they are set once by `flanneld` and used throughout the life of the cluster.

```
# Check kpi093 ARP table for flannel information (mapping form the IP address to the MAC address
# of flannel.1 VXLAN interfaces in the remote cluster nodes)
ubuntu@kpi093:~$ ip neigh show dev flannel.1
10.42.2.0 lladdr e6:4b:9c:85:29:63 PERMANENT # <= mapping for the remote flannel.1 from its IP to MAC
10.42.0.0 lladdr 46:32:55:ca:55:1f PERMANENT
10.42.1.0 lladdr 32:6d:8a:c8:c2:af PERMANENT

# Check flannel.1 forwarding table (maps form the MAC address of the remote flannel.1 VXLAN
# interface to the IP address of the remote cluster node)
ubuntu@kpi093:~$ bridge fdb show dev flannel.1
1a:b8:85:b9:77:bf dst 192.168.1.41 self permanent
e6:4b:9c:85:29:63 dst 192.168.1.39 self permanent # <= mapping from the remote flannel.1 MAC to the
# remote node IP
32:6d:8a:c8:c2:af dst 192.168.1.41 self permanent
46:32:55:ca:55:1f dst 192.168.1.38 self permanent
```

To see the entire picture of the mappings relevant for our example of communication between pods one should complete similar checks for the `nginx` side (node `kpi092` in our case; we omit this for sake of brevity). After consolidating all results, a complete network diagram as depicted in the following figure can be obtained.

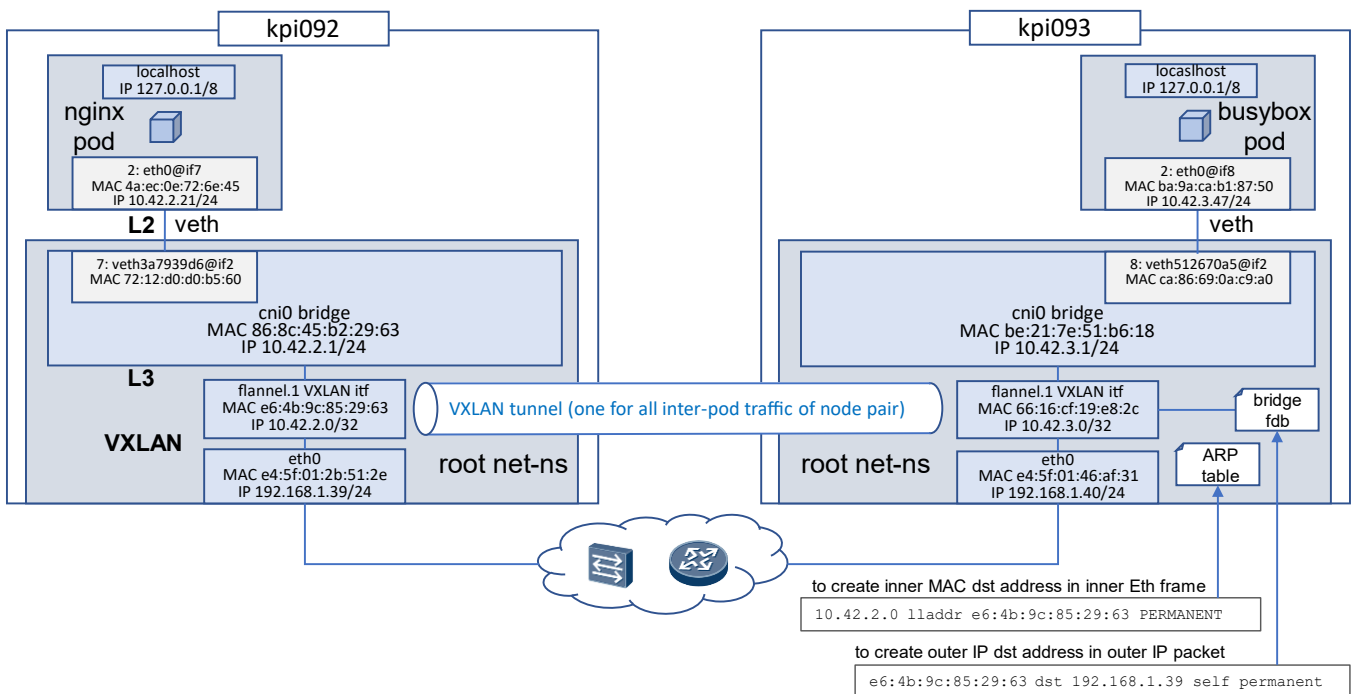


Figure 1 Complete picture of flannel network settings.

Worth of noting is the fact that flannel VXLAN extends between pairs of virtual flannel.1 VTEP's with packets from/to `cni0` bridges serving as the payload of the inner Eth frames, and the inner Eth frames being created by flannel.1 itself.

#### d) Communication on the wire

We will use remote packet capture with `tcpdump` and `Wireshark` to observe the communication on the wire. For the setup procedure and usage of both tools we recommend the link <https://www.comparitech.com/net-admin/tcpdump-capture-wireshark/>. One can find many useful tips for `tcpdump` therein.

Our scenario is documented below. We create three terminals: one terminal used to log to the cluster node where the `busybox` runs (`kpi093` in our case) and capture the packets from the physical network interface `eth0` of the node; the second terminal is used to log into the `busybox` container and send `wget` to `nginx`; third terminal is used to download `tcpdump.pcap` file from node `kpi093` for subsequent analysis in `Wireshark`. The commands entered in each terminal are

gathered in the boxes below, each box corresponding to one terminal. Use common sense to execute the commands in appropriate order. Notice that *xubulab* is my management host I use to *ssh* to cluster nodes and *kubectl* to Kube API.

**# Get to the terminal of cluster node where the busybox pod runs**

```
xubuntu@xubulab:~$ ssh kpi093

ubuntu@kpi093:~$ sudo tcpdump -s 0 -i eth0 -w tcpdump.pcap
tcpdump: listening on eth0, link-type EN10MB (Ethernet), snapshot length 262144 bytes
^C113 packets captured
123 packets received by filter
0 packets dropped by kernel
ubuntu@kpi093:~$ ls
tcpdump.pcap
ubuntu@kpi093:~$ sudo chmod 644 tcpdump.pcap
ubuntu@kpi093:~$
```

**# Get to the terminal of the busybox container in the busy box pod to wget nginx from there**

```
xubuntu@xubulab:~$ kubectl exec -it -c busybox busybox -- /bin/sh

/ # wget 10.42.2.23 -O - 2>/dev/null
<!DOCTYPE html>
<html>
<head>
<title>Welcome to nginx!</title>
<style>
html { color-scheme: light dark; }
body { width: 35em; margin: 0 auto;
font-family: Tahoma, Verdana, Arial, sans-serif; }
</style>
</head>
<body>
<h1>Welcome to nginx!</h1>
<p>If you see this page, the nginx web server is successfully installed and
working. Further configuration is required.</p>

<p>For online documentation and support please refer to
<a href="http://nginx.org/">nginx.org</a>.<br/>
Commercial support is available at
<a href="http://nginx.com/">nginx.com</a>.</p>

<p><em>Thank you for using nginx.</em></p>
</body>
</html>
/ #
```

**# Terminal on the host where Wireshark will be run**

```
xubuntu@xubulab:~/cluster-pi/lab$ kubectl get pods -o wide
NAME          READY   STATUS    RESTARTS   AGE   IP            NODE          NOMINATED NODE
READINESS GATES
dnsutils      1/1     Running   58 (19h ago)  18d   10.42.3.50    kpi093        <none>         <none>
nginx         1/1     Running   2 (19h ago)  29h   10.42.2.23    kpi092        <none>         <none>
busybox       1/1     Running   2 (19h ago)  29h   10.42.3.51    kpi093        <none>         <none>
nginx-busybox 2/2     Running   8 (19h ago)  9d    10.42.1.42    kpi094        <none>         <none>
xubuntu@xubulab:~/cluster-pi/lab$ scp ubuntu@192.168.1.40:/home/ubuntu/tcpdump.pcap ./
tcpdump.pcap                                100% 19KB  4.5MB/s   00:00
xubuntu@xubulab:~/cluster-pi/lab$
```

Having downloaded the *tcpdump.pcap* file we can analyze it in Wireshark. At the beginning, remember to configure Wireshark to decode VXLAN traffic. To this end, use the *Analyse/Decode as/* window to adjust the settings for UDP and HTTP protocols as shown in figure **Figure 2** (set the columns *Field*, *Value* and *Current*).

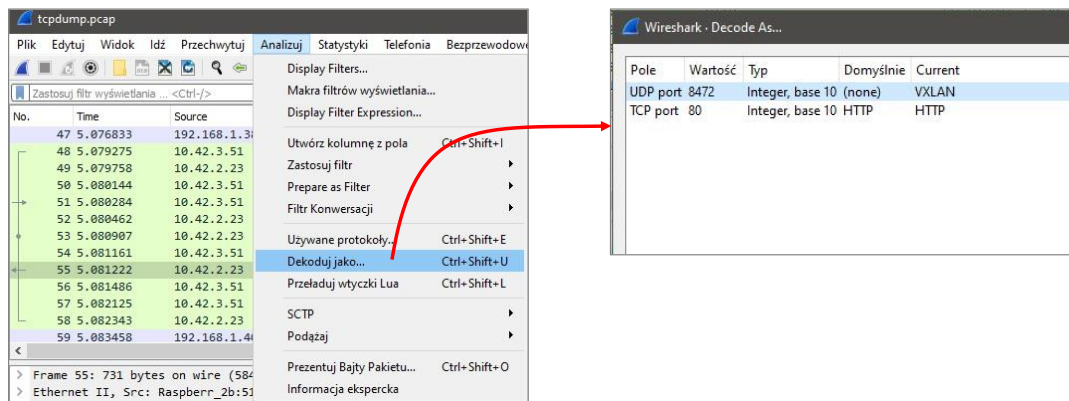


Figure 2 Setting Wireshark to decode UDP/VXLAN.

A decoded HTTP OK message received in response to wget/GET is shown below. One can easily identify the full VXLAN stack. Notice, however, that the IP and/or MAC addresses of many entities (pods, cni0, flannel.1) reported throughout this point (d) are different from the addresses presented in previous points (a), (b), (c). This is due to the fact that addresses (as well as the entities that own them, as pods) are ephemeral in Kubernetes and change between startups of the cluster, and in our example we did run the packet capture session after cluster restart. Fortunately, the assignment of new addresses to recreate the architecture like in Figure 1 can easily be inferred from the packet capture using the knowledge what we have learnt so far.

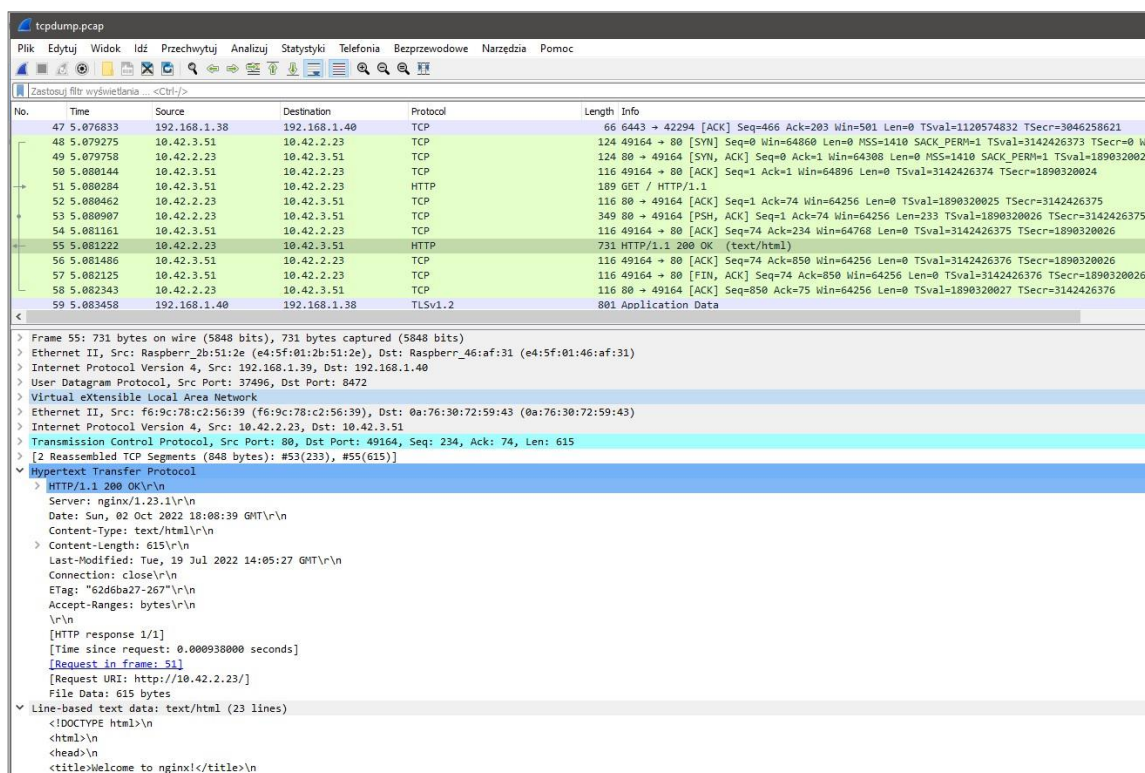


Figure 3 Decoded the communication over flannel VXLAN.

### 3. Installing MetalLB from manifest files

#### 3.1. About MetalLB

Kubernetes does not offer an implementation of network load balancers (Services of type `LoadBalancer`<sup>3</sup>) for bare-metal clusters. The implementations of network load balancers that Kubernetes does ship with are all glue code that calls out to various IaaS platforms (GCP, AWS, Azure...). If you're not running on a supported IaaS platform (GCP, AWS, Azure...), LoadBalancers will remain in the "pending" state indefinitely when created.

Bare-metal cluster operators are left with two lesser tools to bring user traffic into their clusters, `NodePort` and `ExternalIPs` services. Both of these options have significant downsides for production use, which makes bare-metal clusters second-class citizens in the Kubernetes ecosystem. MetalLB aims to suppress this imbalance by offering a network load balancer implementation that integrates with standard network equipment, so that external services on bare-metal clusters also "just work" as much as possible.

MetalLB hooks into your Kubernetes cluster, and provides a network load-balancer implementation. In short, it allows you to create Kubernetes services of type `LoadBalancer` in clusters that don't run on a cloud provider, and thus cannot simply hook into paid products to provide load balancers. It allows enabling `LoadBalancer` service addresses in any bare-metal Kubernetes installation using standard routing protocols. It has two features that work together to provide this service: **address allocation**, and **external announcement**.

- Address allocation

In a Kubernetes cluster on a cloud provider, you request a load balancer, and your cloud platform assigns an IP address to you. In a private bare-metal cluster, MetalLB is responsible for that allocation.

MetalLB cannot create IP addresses out of thin air, so you have to give it pools of IP addresses that it can use. MetalLB will take care of assigning and unassigning individual addresses as services come and go, but it will only ever hand out IPs that are part of its configured pools. Below in point 3.4, we will configure a pool of IP addresses for MetalLB in our cluster.

- External announcement

After MetalLB has assigned an external IP address to a service, it needs to make the network beyond the cluster aware that the IP "lives" in the cluster. MetalLB uses standard networking or routing protocols to achieve this, depending on which mode is used: layer 2 ARP/NDP, or layer 3 BGP.

In layer 2 mode, one machine (typically the master node) in the cluster takes ownership of the service, and uses standard address discovery protocols (ARP for IPv4, NDP for IPv6) to make those IPs reachable on the local network. From the LAN's point of view, the announcing machine simply has multiple IP addresses. Under the hood, MetalLB responds to ARP requests for IPv4 services, and NDP requests for IPv6. The major advantage of the layer 2 mode is its universality: it will work on any Ethernet network, with no special hardware required, not even fancy routers. This is the mode we are going to use in our cluster. More on layer 2 mode can be found here: <https://metallb.universe.tf/concepts/layer2/>.

In BGP mode, all machines in the cluster establish [BGP](#) peering sessions with nearby routers that you control, and tell those routers how to forward traffic to the service IPs. Using BGP allows for true load balancing across multiple nodes, and fine-grained traffic control thanks to BGP's policy mechanisms. Operating external router(s) is the additional cost to be paid for this flexibility. More on BGP mode can be found here: <https://metallb.universe.tf/concepts/bgp/>.

MetalLB speaker is in charge of IP advertisement. It is based on leader election on per-service-instance basis so that traffic incoming to the cluster can be load balanced evenly among cluster nodes. Leader election is deterministic based on hashes so no consensus protocol is needed for the speakers for consistent election decision. Current leader is kept track of by remaining speakers based on updating a dedicated resource. Actually, it is kube-proxy on the leader node of the service which load balances service's traffic (connections) among service's pods. However, there is a restriction that can limit the access to services through the ingress mechanism if nodes have taints that disallow MetalLB speakers on

---

<sup>3</sup> A quick introduction to service types in Kubernetes would be helpful here. For example this video can serve the purpose: <https://www.youtube.com/watch?v=T4Z7visMM4E>. Many other materials are available in the net.

them: the speaker pod that announces the external IP address for a service must be on the same node as an endpoint (a pod) for the service and the endpoint must be in the Ready condition.

More on MetalLB can be found here: <https://metallb.universe.tf/concepts/> and <https://docs.openshift.com/container-platform/4.9/networking/metallb/about-metallb.html>

### 3.2. Installation of MetalLB

#### a) Pre-installation check of services running in the cluster

Before installation, let's list all services in our cluster created as a result of cluster installation. Notice that the Traefik service of Kubernetes type LoadBalancer has not been assigned EXTERNAL-IP address (Traefik is the ingress controller in our configuration). That means that there were no service running to make this assignment – in fact, we installed k3s with the embedded k3s load balancer disabled `--disable servicelb` (in k3s, Klipper is the embedded load balancer). In fact, MetalLB will be used in our cluster instead of Klipper and once installed and configured with IP address ranges to assign from (sec. 3.4), it is expected to assign EXTERNAL-IP address to Traefik.

<code>xubuntu@xubulab:~/cluster-pi\$ kubectl get services -A</code>						
NAMESPACE	NAME	TYPE	CLUSTER-IP	EXTERNAL-IP	PORT(S)	AGE
default	kubernetes	ClusterIP	10.43.0.1	<none>	443/TCP	3d3h
kube-system	kube-dns	ClusterIP	10.43.0.10	<none>	53/UDP, 53/TCP, 9153/TCP	3d3h
kube-system	metrics-server	ClusterIP	10.43.172.124	<none>	443/TCP	3d3h
kube-system	traefik	LoadBalancer	10.43.42.129	<pending>	80:31446/TCP, 443:31051/TCP	3d3h

#### b) Labeling worker nodes – just for any case

For each worker node add label “worker” by issuing the following command. We do it just for any case - this label may turn out to be useful during future operations.

```
kubectl label node <node name> node-role.kubernetes.io/worker=true
```

#### c) Installing MetalLB

Execute the command in the frame below and observe the output similar to the one presented. **Before that, always check the current version of MetalLB** <https://metallb.universe.tf/installation/#installation-by-manifest> and use it in your `kubectl apply` command (in the example below, version v0.13.7 is used).

```
xubuntu@xubulab:~/cluster-pi/manifests$ kubectl apply -f \
https://raw.githubusercontent.com/metallb/metallb/v0.13.7/config/manifests/metallb-native.yaml
namespace/metallb-system created
customresourcedefinition.apiextensions.k8s.io/addresspools.metallb.io created
customresourcedefinition.apiextensions.k8s.io/bfdprofiles.metallb.io created
customresourcedefinition.apiextensions.k8s.io/bgppeeradvertisements.metallb.io created
customresourcedefinition.apiextensions.k8s.io/bgppeers.metallb.io created
customresourcedefinition.apiextensions.k8s.io/communities.metallb.io created
customresourcedefinition.apiextensions.k8s.io/ipaddresspools.metallb.io created
customresourcedefinition.apiextensions.k8s.io/l2advertisements.metallb.io created
serviceaccount/controller created
serviceaccount/speaker created
role.rbac.authorization.k8s.io/controller created
role.rbac.authorization.k8s.io/pod-lister created
clusterrole.rbac.authorization.k8s.io/metallb-system:controller created
clusterrole.rbac.authorization.k8s.io/metallb-system:speaker created
rolebinding.rbac.authorization.k8s.io/controller created
rolebinding.rbac.authorization.k8s.io/pod-lister created
clusterrolebinding.rbac.authorization.k8s.io/metallb-system:controller created
clusterrolebinding.rbac.authorization.k8s.io/metallb-system:speaker created
secret/webhook-server-cert created
service/webhook-service created
deployment.apps/controller created
daemonset.apps/speaker created
validatingwebhookconfiguration.admissionregistration.k8s.io/metallb-webhook-configuration
created
```



### 3.3. Checking if MetalLB has been installed

```
xubuntu@xubulab:~/cluster-pi/manifests/metalb$ kubectl get deployments --all-namespaces
NAMESPACE      NAME                                READY   UP-TO-DATE   AVAILABLE   AGE
kube-system    coredns                            1/1     1             1           1h
kube-system    local-path-provisioner             1/1     1             1           1h
kube-system    metrics-server                     1/1     1             1           1h
kube-system    traefik                            1/1     1             1           1h
metallb-system controller                        1/1     1             1           1h
xubuntu@xubulab:~/cluster-pi/manifests/metalb$ kubectl get pods -n metallb-system -o wide
NAME                                READY   STATUS    RESTARTS   AGE   IP            NODE    NOMINATED NODE
speaker-7n2lg                       1/1     Running   0           2h15m  192.168.1.40  kpi093  <none>
speaker-nwvpk                       1/1     Running   0           2h15m  192.168.1.41  kpi094  <none>
speaker-wc6r4                       1/1     Running   0           2h15m  192.168.1.39  kpi092  <none>
controller-6c58495cbb-fwzbn        1/1     Running   1 (2h14m ago)  2h15m  10.42.2.5     kpi093  <none>
```

As seen, one controller deployment and a daemon-set with three speakers have been created (in section 4.3 we discuss why three speaker are run and how to enable metallb speakers also on master/control nodes). For metallb working in L2 mode, in a given LAN segment, metallb speaker is in charge of ARP IP advertisements for those services for which it was assigned to be the leader.

### 3.4. Configuring IP address pool for MetalLB

Instructions according to: <https://metallb.universe.tf/configuration/>

Defining the IPs to assign to the Load Balancer services and announce the service ips to be used in Layer 2 configuration.

- d) On the configuration host, create directory `metallb`
- e) Create `ip-pool-config.yaml` in directory `metallb` with the following content (you can always modify the IP range in `ip-pool-config.yaml` if you want, and apply it again)

```
--
apiVersion: metallb.io/v1beta1
kind: IPAddressPool
metadata:
  name: first-pool
  namespace: metallb-system
spec:
  addresses:
  - 192.168.1.200-192.168.1.254 # adjust address ranges according to your environment
---
apiVersion: metallb.io/v1beta1
kind: L2Advertisement
metadata:
  name: l2-pool
  namespace: metallb-system
spec:
  ipAddressPools:
  - first-pool
```

- f) Apply the config and check the result

```
xubuntu@xubulab:~/cluster-pi/manifests/metalb$ nano ip-pool-config.yaml
xubuntu@xubulab:~/cluster-pi/manifests/metalb$ kubectl apply -f ip-pool-config.yaml
ipaddresspool.metallb.io/first-pool created
l2advertisement.metallb.io/l2-pool created
xubuntu@xubulab:~/cluster-pi/manifests/metalb$ kubectl get -n metallb-system ipaddresspools
NAME      AGE
first-pool 6m45s
xubuntu@xubulab:~/cluster-pi/manifests/metalb$ kubectl describe -n metallb-system ipaddresspool
first-pool
Name:      first-pool
Namespace: metallb-system
Labels:    <none>
```

```

Annotations:  <none>
API Version:  metallb.io/v1beta1
Kind:         IPAddressPool
Metadata:
  Creation Timestamp:  2022-10-10T20:12:33Z
  Generation:         1
  Managed Fields:
    API Version:  metallb.io/v1beta1
    Fields Type:  FieldsV1
    fieldsV1:
      f:metadata:
        f:annotations:
          .:
            f:kubectl.kubernetes.io/last-applied-configuration:
      f:spec:
        .:
          f:addresses:
          f:autoAssign:
          f:avoidBuggyIPs:
    Manager:      kubectl-client-side-apply
    Operation:    Update
    Time:         2022-10-10T20:12:33Z
    Resource Version:  115222
    UID:          080a5c65-6810-46ba-b341-28f020c54be0
Spec:
  Addresses:
    192.168.1.200-192.168.1.254
  Auto Assign:      true
  Avoid Buggy I Ps:  false
Events:            <none>

```

g) Check all resources (including custom resources) in a given namespace (here metallb-system):

```

xubuntu@xubulab:~$ kubectl api-resources --verbs=list --namespaced -o name \
| xargs -n 1 kubectl get --show-kind --ignore-not-found -n metallb-system

```

(Note: kubectl get command get does not get custom resources. All resource types in Kubernetes cluster are displayed using kubectl api-resources command. See, e.g., <https://www.studytonight.com/post/how-to-list-all-resources-in-a-kubernetes-namespace>)

h) Check Traefik EXTERNAL-IP address after having configured MetalLB. Observe an assigned value – address that should be reachable on the HTTP level from the outside of the cluster.

```

xubuntu@xubulab:~/cluster-pi/manifests/metallb$ kubectl get svc -n kube-system

```

NAME	TYPE	CLUSTER-IP	EXTERNAL-IP	PORT(S)
kube-dns	ClusterIP	10.43.0.10	<none>	53/UDP,53/TCP,9153/TCP
metrics-server	ClusterIP	10.43.172.124	<none>	443/TCP
traefik	LoadBalancer	10.43.42.129	192.168.1.200	80:31446/TCP,443:31051/TCP

Kubernetes Services, Load Balancing, and Networking: [https://kubernetes.io/docs/concepts/services-networking/\\_print/](https://kubernetes.io/docs/concepts/services-networking/_print/)

#### 4. Checking Traefik with external IP and basic load balancing

a) ??? Check the logs again, this time for errors (below, we use a command slightly different from the one used previously but producing similar output)

```

kubectl logs $(kubectl get pods --namespace kube-system | grep "^traefik" | awk '{print $1}') --namespace kube-system

```

if you see errors like:



```

time="2022-09-10T16:52:23Z" level=info msg="Configuration loaded from flags."
E0910 19:04:51.935590      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.TraefikService: the server is
currently unable to handle the request (get traefikservices.traefik.containo.us)
E0910 19:04:53.136361      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.TLSStore: the server is
currently unable to handle the request (get tlsstores.traefik.containo.us)
E0910 19:04:53.530293      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.TraefikService: failed to list
*vlalpha1.TraefikService: the server is currently unable to handle the request (get
traefikservices.traefik.containo.us)
E0910 19:04:54.021591      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.TLSStore: failed to list
*vlalpha1.TLSStore: apiserver not ready
E0910 19:04:54.690477      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vl.Secret: the server is currently unable
to handle the request (get secrets)
E0910 19:04:55.098232      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.ServersTransport: the server is
currently unable to handle the request (get serverstransports.traefik.containo.us)
E0910 19:04:55.625322      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vl.Secret: failed to list *vl.Secret:
apiserver not ready
E0910 19:04:55.930556      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.ServersTransport: failed to list
*vlalpha1.ServersTransport: apiserver not ready
E0910 19:04:56.009639      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.MiddlewareTCP: the server is
currently unable to handle the request (get middlewaretcps.traefik.containo.us)
E0910 19:04:56.018650      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.TLSStore: failed to list
*vlalpha1.TLSStore: apiserver not ready
E0910 19:04:56.421826      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.TraefikService: failed to list
*vlalpha1.TraefikService: apiserver not ready
E0910 19:04:57.217422      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.MiddlewareTCP: failed to list
*vlalpha1.MiddlewareTCP: apiserver not ready
E0910 19:04:57.525423      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vl.Secret: failed to list *vl.Secret:
apiserver not ready
E0910 19:04:58.358102      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vlalpha1.ServersTransport: failed to list
*vlalpha1.ServersTransport: apiserver not ready
E0910 19:04:58.533464      1 reflector.go:138] k8s.io/client-
go@v0.22.1/tools/cache/reflector.go:167: Failed to watch *vl.Service: the server is currently
unable to handle the request (get services)
I0910 19:05:11.668610      1 trace.go:205] Trace[1699467881]: "Reflector ListAndWatch"
name:k8s.io/client-go@v0.22.1/tools/cache/reflector.go:167 (10-Sep-2022 19:04:59.621) (total time:
12046ms):
Trace[1699467881]: ---"Objects listed" 12046ms (19:05:11.668)
Trace[1699467881]: [12.046814764s] [12.046814764s] END
I0910 19:05:11.941403      1 trace.go:205] Trace[1736545453]: "Reflector ListAndWatch"
name:k8s.io/client-go@v0.22.1/tools/cache/reflector.go:167 (10-Sep-2022 19:04:59.869) (total time:
12072ms):
Trace[1736545453]: ---"Objects listed" 12072ms (19:05:11.941)
Trace[1736545453]: [12.072132709s] [12.072132709s] END

```

**Then take a look here and pray for finding good hints (or maybe you'd better reinstall the cluster):**

<https://stackoverflow.com/questions/69619850/traefik-the-server-could-not-find-the-requested-resource>

- o <https://docs.k0sproject.io/v1.23.6+k0s.2/examples/traefik-ingress/>
- o <https://community.traefik.io/t/traefik-on-kubernetes-stuck-to-access-dashboard/14999/4>
- o <https://blog.zachinachshon.com/traefik-ingress/>
- o <https://doc.traefik.io/traefik/routing/entrypoints/>
- o <https://blog.knoldus.com/how-to-install-traefik-ingress-controller-in-kubernetes/>
- o <https://www.codementor.io/@slavko/unobtrusive-local-development-with-kubernetes-k3s-traefik2-15uq596oja>
- o <https://doc.traefik.io/traefik/v2.3/reference/dynamic-configuration/kubernetes-crd/#definitions>

- o <https://github.com/traefik/traefik-helm-chart/issues/77>

b) In case of getting `Error` status for pods, you can try restarting the failing pod:

<https://www.containiq.com/post/using-kubectl-to-restart-a-kubernetes-pod>

c) **Finally, when everything works fine**, just out of curiosity we can also check the MAC address of MetalLB speaker that has taken leadership for Traefik – run the following on the management host and catch ARP responses to “who has 192.168.1.200” (adjust the IP address of Traefik and the interface name of the management host according to your environment if needed):

```
sudo tcpdump -ennqti enp0s3 # this tcpdump is optional, arping will suffice
arping 192.168.1.200 -c 5
```

#### 4.1. Checking the access to Traefik dashboard

Here, we check the access to Traefik dashboard in the context of LoadBalancer service type used to expose it externally. We will not focus specifically on detailed Traefik configuration during this lab, but its also good to know Traefik can be managed from a browser.

- Check using a browser accessing from the outside of the cluster: <http://192.168.1.200/dashboard/>
- You will most probably get the error 404** and will need to update the IngressRoute CRD for Traefik manually:

- go to your manifests/traefik directory and retrieve the ingress route for Traefik by running

```
kubectl get ingressroutes -n kube-system traefik-dashboard -o yaml >
traefik-dashboard.yaml
```

- in file `traefik-dashboard.yaml` find the `spec` section and check whether it has the value of `entryPoints` set to `'- traefik'`. If so, change `'- traefik'` to `'- web'` and the IngressRoute will allow routing of external queries to the Traefik dashboard (ignore warnings):

```
spec:
  entryPoints:
    - web
```

- and apply: `kubectl apply -f traefik-dashboard.yaml`
- check again: <http://192.168.1.200/dashboard/> (remember to put the closing `"/` sign as on some OS-es you get HTTP 404 if closing `"/` sign is missing)

Note: the adjustment of the ingress route can also be done using `kubectl edit` command as shown in c) below (in this case, saving changes with the `vi` command `:w` automatically triggers updating the resource by Kubernetes). In either case, remember that this change will persist as long as you do not reinstall Traefik.

- you can confirm the IngressRoute CRD for Traefik has been changed using `kubectl edit` (you'd better not save it, i.e., quit the `vi` editor by entering `:q!`)

```
kubectl edit ingressroute -n kube-system traefik-dashboard
```

- NOTE: if Traefik dashboard persists unreachable** it is most probably because of the fact that we introduced additional taint `CriticalAddonsOnly=true:NoExecute` for the master node during k3s installation in Ansible (refer to the Ansible play for the master role in your Ansible files). This results in that metallb speakers can't be executed on the master node which seems to be the source of problems (in fact, it should not be so, but the experience suggests it may happen). All in all, if you face this problem you need to manually update DaemonSet resource/manifest as described in section 4.3. This situation is odd, but as it happened once to me it is worth commenting on at least a basic level. Hopefully you will not face it.

## 4.2. Checking basic (in-cluster) load balancing

This section relates to internal cluster load balancing that load balances the load among multiple pods (replicas) of a given Service instance. It is beyond MetalLB load balancing (another level) and kube-proxy responsible for the configuration of ip-tables in cluster nodes is the main actor in this case.

Create deployment.yaml and service.yaml manifests for nginx as given in the frame below and run the set of commands from the following frame. Notice that our deployment requests 3 replicas of nginx to load balance between them. They can be seen as three Endpoints: 10.42.1.24:80,10.42.2.17:80,10.42.3.17:80 displayed with the command `kubectl describe` below. You can check on which nodes they are instantiated by running `kubectl get pods -o wide`.

*Notice: in my case, it once happened that one replica of nginx was instantiated on the master node which is odd considering the fact that master was installed with flag node-taint set to `--node-taint CriticalAddonsOnly=true:NoExecute`. The latter taint is expected to disable workload execution on the master. My conjecture is that setting the taint may not work properly when done using `curl -sFL https://get.k3s.io`. I've left this behaviour to investigate it in more detail for the future, and a provisional workaround for this that seems to work well is to run manually the following:*

- o `kubectl taint nodes <master-node-name> CriticalAddonsOnly=true:NoExecute`
- o `kubectl describe node <master-node-name> # to check the change is in effect`

### Manifests

```
# deployment.yaml content
---
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx
spec:
  selector:
    matchLabels:
      app: nginx
  replicas: 3
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
        - name: nginx
          image: nginx:alpine
          ports:
            - containerPort: 80

# service.yaml content
---
apiVersion: v1
kind: Service
metadata:
  name: nginx
spec:
  selector:
    app: nginx
  ports:
    - port: 80
      targetPort: 80
  type: LoadBalancer
```

### Commands

```
xubuntu@xubulab:~/cluster-pi/manifests$ kubectl apply -f example/deployment.yaml
deployment.apps/nginx created
xubuntu@xubulab:~/cluster-pi/manifests$ kubectl apply -f example/service.yaml
service/nginx created

# display the endpoints (replica Pods)

xubuntu@xubulab:~/cluster-pi/manifests$ kubectl describe service nginx
Name: nginx
Namespace: default
Labels: <none>
Annotations: <none>
Selector: app=nginx
Type: LoadBalancer
IP Family Policy: SingleStack
IP Families: IPv4
IP: 10.43.71.110
IPs: 10.43.71.110
LoadBalancer Ingress: 192.168.1.201
Port: <unset> 80/TCP
TargetPort: 80/TCP
NodePort: <unset> 32456/TCP
Endpoints: 10.42.1.24:80,10.42.2.17:80,10.42.3.17:80
Session Affinity: None
External Traffic Policy: Cluster
Events:
  Type     Reason            Age   From                      Message
  ----     -
  Normal   IPAllocated       8s    metallb-controller        Assigned IP ["192.168.1.201"]
  Normal   nodeAssigned      8s    metallb-speaker           announcing from node "worker-2" with protocol
"layer2"

# reachability check
# here, you can also check the nodes where replicas are running (see the description above)

xubuntu@xubulab:~/cluster-pi/manifests$ curl 192.168.1.201
<!DOCTYPE html>
<html>
<head>
<title>Welcome to nginx!</title>
<style>
html { color-scheme: light dark; }
body { width: 35em; margin: 0 auto;
font-family: Tahoma, Verdana, Arial, sans-serif; }
</style>
</head>
<body>
<h1>Welcome to nginx!</h1>
<p>If you see this page, the nginx web server is successfully installed and
working. Further configuration is required.</p>

<p>For online documentation and support please refer to
<a href="http://nginx.org/">nginx.org</a>.<br/>
Commercial support is available at
<a href="http://nginx.com/">nginx.com</a>.</p>

<p><em>Thank you for using nginx.</em></p>
</body>
</html>

# clean after checking

xubuntu@xubulab:~/cluster-pi/manifests$ kubectl delete -f example/deployment.yaml
deployment.apps "nginx" deleted
xubuntu@xubulab:~/cluster-pi/manifests$ kubectl delete -f example/service.yaml
service "nginx" deleted
```

### 4.3. Closing remarks - analysing the taint in the master for metallb speakers

In this additional section, we analyse why metallb speaker is not run on the master node (if that's the case in your cluster). The reason for that is the interplay between node *taints* and pod *tolerations* that express the willingness of a node to accept specific type of workloads, and explicit permission for a given workload (pod) to be accepted by otherwise tainted node, respectively. To see this we simply have to check and compare taints in the master node and tolerances of the speaker pods.

*Notice: we introduced the taint to master node in the k3 install command run in Ansible playbook for the master role (one can check it); by doing that we wanted to avoid running unnecessary workloads on the master. As metallb speakers do not have sufficient tolerances defined they are blocked from running on the master.*

Follow the commands in the frame and analyse the taints and tolerations.

```
# get metallb pods to see there is no speaker running on master node (we saw it already during metallb
installation, but did not pay attention to)
xubuntu@xubulab:~/cluster-pi/manifests/metallb$ kubectl get pods -n metallb-system
NAME                                READY   STATUS    RESTARTS   AGE
speaker-7n2lg                       1/1     Running   0           4h47m
speaker-nwvpk                       1/1     Running   0           4h47m
speaker-wc6r4                       1/1     Running   0           4h47m
controller-6c58495cbb-fwzbn        1/1     Running   1 (4h46m ago)  4h47m

# check the master to see it has not a speaker running on it; check the taints defined in the master
xubuntu@xubulab:~/cluster-pi/manifests/metallb$ kubectl describe node kpi091
Name:                               kpi091
Roles:                              control-plane,etcd,master
(. . .)
Taints:                             CriticalAddonsOnly=true:NoExecute
Unschedulable:                      false
Lease:
  HolderIdentity:  kpi091
(. . .)

# describe any of the speakers to see it does not contain toleration against CriticalAddonsOnly (is
unexecutable on the master kpi091); speakers run as DaemonSet so they all share this restriction and it is
the reason why no speaker was scheduled on the master kpi091
xubuntu@xubulab:~/cluster-pi/manifests/metallb$ kubectl describe -n metallb-system pod speaker-7n2lg
Name:                               speaker-7h8jj
Namespace:                          metallb-system
Priority:                             0
Node:                               kpi093/192.168.1.40
(. . .)
Node-Selectors:                      kubernetes.io/os=linux
Tolerations:                         node-role.kubernetes.io/control-plane:NoSchedule op=Exists
                                     node-role.kubernetes.io/master:NoSchedule op=Exists
                                     node.kubernetes.io/disk-pressure:NoSchedule op=Exists
                                     node.kubernetes.io/memory-pressure:NoSchedule op=Exists
                                     node.kubernetes.io/network-unavailable:NoSchedule op=Exists
                                     node.kubernetes.io/not-ready:NoExecute op=Exists
                                     node.kubernetes.io/pid-pressure:NoSchedule op=Exists
                                     node.kubernetes.io/unreachable:NoExecute op=Exists
                                     node.kubernetes.io/unschedulable:NoSchedule op=Exists

# One can allow speakers run on master nodes either directly from kubectl edit command; refer to the edited
# DaemonSet shown below in this frame to see how the addition should look like.
xubuntu@xubulab:~/cluster-pi/manifests$ kubectl edit -n metallb-system daemonset speaker
daemonset.apps/speaker edited
xubuntu@xubulab:~/cluster-pi/manifests$ kubectl get pods -n metallb-system
NAME                                READY   STATUS    RESTARTS   AGE
controller-6c58495cbb-fwzbn        1/1     Running   2 (16h ago)  23h
speaker-vcq4k                      1/1     Running   0           67s
speaker-2wf45                      1/1     Running   0           35s
speaker-16tz8                      1/1     Running   0           14s
speaker-vp2xw                      0/1     ContainerCreating   0           2s
xubuntu@xubulab:~/cluster-pi/manifests$ kubectl get pods -n metallb-system
NAME                                READY   STATUS    RESTARTS   AGE
controller-6c58495cbb-fwzbn        1/1     Running   2 (16h ago)  23h
speaker-vcq4k                      1/1     Running   0           92s
speaker-2wf45                      1/1     Running   0           60s
```

```

speaker-l6tz8          1/1      Running  0          39s
speaker-vp2xw          1/1      Running  0          27s

# or by exporting the manifest to a file, modifying it and applying again updated speaker daemonset as shown
below in this frame. The overall procedure: delete speaker daemonset, add toleration
CriticalAddonsOnly=true:NoExecute in exported daemonset manifest, re-apply daemonset.

xubuntu@xubulab:~/cluster-pi/manifests/metallb$ kubectl get daemonsets --all-namespaces
NAMESPACE      NAME      DESIRED  CURRENT  READY  UP-TO-DATE  AVAILABLE  NODE SELECTOR
metallb-system  speaker   4         4         4       4           4          kubernetes.io/os=linux

# delete daemonset
xubuntu@xubulab:~/cluster-pi/manifests/metallb$ kubectl delete -n metallb-system daemonset speaker
daemonset.apps "speaker" deleted

# Edit speaker daemonset to add the toleration; steps 1, 2, 3:

# 1. copy complete Daemonset manifest (search Daemonset in the downloaded file) to a separate yaml file, say
speaker-daemonset.yaml
xubuntu@xubulab:~/cluster-pi/manifests/metallb$ kubectl get -n metallb-system daemonset speaker -o yaml >
speaker-daemonset.yaml

# 2. and edit this new file to add the missing toleration (see below to find the update)
xubuntu@xubulab:~/cluster-pi/manifests/metallb$ nano speaker-daemonset.yaml

# edited daemonset manifest - relevant part
---
apiVersion: apps/v1
kind: DaemonSet
metadata:
  labels:
    app: metallb
    component: speaker
  name: speaker
  namespace: metallb-system
spec:
  selector:
    matchLabels:
      app: metallb
      component: speaker
  (. . .)
  tolerations:
    - effect: NoSchedule
      key: node-role.kubernetes.io/master
      operator: Exists
    - effect: NoSchedule
      key: node-role.kubernetes.io/control-plane
      operator: Exists
# this is to be added: toleration to enable runing a speaker on the master node
    - effect: NoExecute
      key: CriticalAddonsOnly
      operator: Equal
      value: "true"
  ---

# 3. apply modified daemonset manifest to re-instantiate the speakers
xubuntu@xubulab:~/cluster-pi/manifests/metallb$ kubectl apply -f speaker-daemonset.yaml
daemonset.apps/speaker created

# 4. get pods to check if there is a speaker running on every node
xubuntu@xubulab:~/cluster-pi/manifests/metallb$ kubectl get pods -n metallb-system -o wide
NAME                                READY  STATUS   RESTARTS   AGE   IP              NODE      NOMINATED NODE
controller-6d5cb87f6-5hkh1          1/1    Running  0           52s   10.42.3.75      kpi093   <none>
speaker-2qs5m                       1/1    Running  0           52s   192.168.1.38    kpi091   <none>
speaker-66b5m                       1/1    Running  0           52s   192.168.1.39    kpi092   <none>
speaker-l7njq                       1/1    Running  0           52s   192.168.1.41    kpi094   <none>
speaker-n7mps                       1/1    Running  0           52s   192.168.1.40    kpi093   <none>

```

## 5. Additional readings/videos and hints

- Fully Automated K3s etcd High Availability Install and more:  
<https://docs.technotim.live/posts/k3s-etcd-ansible/>  
<https://github.com/techno-tim/k3s-ansible>
- In case of problems with a given namespace, it is sometimes best to delete the whole namespace and recreate it. <https://stackoverflow.com/questions/47128586/how-to-delete-all-resources-from-kubernetes-one-time>
- Pod troubleshooting <https://able8.medium.com/automated-troubleshooting-of-kubernetes-pods-issues-c6463bed2f29>
- Flannel video, detailed, good:  
<https://www.youtube.com/watch?v=U35C0EPSwoY&list=PLSAko72nKb8QWsfPpBlsw-kOdMBD7sra-&index=2>
- What is VXLAN and How It is Used as an Overlay Network in Kubernetes?  
<https://www.youtube.com/watch?v=WMLSD2y2Ig4>
- If you want to experiment with Kubernetes certificates check this You can use our cluster of use kind or minikube to start safely):
  - <https://www.fullstaq.com/knowledge-hub/blogs/setting-up-your-own-k3s-home-cluster>
  - <https://www.youtube.com/watch?v=hoLUigg4V18>