Scalability and Observability

with

Isovalent Cilium Enterprise

Filip Nikolic
Kubernetes Engineer at PostFinance

Some Numbers

- 104 Billion CHF Customer Assets (~116 Billion USD)
- 2.5 Million Users
  - 1.9 Million Users E-Finance (Online)
- 1.3 Billion Transactions each year
  - 3.5 Million Transactions a day

https://www.postfinance.ch/content/dam/pfch/doc/440_459/450_01_2022_de.pdf

Kube-Proxy

- Network proxy
- Runs on every node
- Two modes
  - iptables (default)
  - ipvs
- Lookup is linear (iptables)

```yaml
apiVersion: v1
kind: Service
metadata:
  name: nginx
spec:
  selector:
    app: nginx
  ports:
  - port: 80
---
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx
spec:
  replicas: 2
  selector:
    matchLabels:
      app: nginx
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx
```

```yaml
apiVersion: v1
kind: Service
metadata:
  name: nginx
spec:
  selector:
    app: nginx
  ports:
  - port: 80
---
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx
spec:
  replicas: 2
  selector:
    matchLabels:
      app: nginx
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx
```

```yaml
apiVersion: v1
kind: Service
metadata:
  name: nginx
spec:
  selector:
    app: nginx
  ports:
  - port: 80
---
apiVersion: apps/v1
kind: Deployment
metadata:
  name: nginx
spec:
  replicas: 2
  selector:
    matchLabels:
      app: nginx
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx
```

```
$ kubectl get svc
NAME    TYPE        CLUSTER-IP       EXTERNAL-IP   PORT(S)   AGE
nginx   ClusterIP   10.108.169.159   <none>        80/TCP    4s

$ kubectl get pods -o wide
NAME                     READY   STATUS    RESTARTS   AGE   IP               NODE    NOMINATED NODE   READINESS GATES
nginx-b9d76c89c-cv7sg    1/1     Running   0          8s    10.244.166.161   node1   <none>           <none>
nginx-b9d76c89c-gd5ld    1/1     Running   0          8s    10.244.104.9     node2   <none>           <none>

$ kubectl get ep
NAME    ENDPOINTS                           AGE
nginx   10.244.104.9:80,10.244.166.161:80   12s
```

```
$ kubectl get svc
NAME    TYPE        CLUSTER-IP       EXTERNAL-IP   PORT(S)   AGE
nginx   ClusterIP   10.108.169.159   <none>        80/TCP    4s

$ kubectl get pods -o wide
NAME                     READY   STATUS    RESTARTS   AGE   IP               NODE    NOMINATED NODE   READINESS GATES
nginx-b9d76c89c-cv7sg    1/1     Running   0          8s    10.244.166.161   node1   <none>           <none>
nginx-b9d76c89c-gd5ld    1/1     Running   0          8s    10.244.104.9     node2   <none>           <none>

$ kubectl get ep
NAME    ENDPOINTS                           AGE
nginx   10.244.104.9:80,10.244.166.161:80   12s
```

```
$ kubectl get svc
NAME    TYPE        CLUSTER-IP       EXTERNAL-IP   PORT(S)   AGE
nginx   ClusterIP   10.108.169.159   <none>        80/TCP    4s

$ kubectl get pods -o wide
NAME                     READY   STATUS    RESTARTS   AGE   IP               NODE    NOMINATED NODE   READINESS GATES
nginx-b9d76c89c-cv7sg    1/1     Running   0          8s    10.244.166.161   node1   <none>           <none>
nginx-b9d76c89c-gd5ld    1/1     Running   0          8s    10.244.104.9     node2   <none>           <none>

$ kubectl get ep
NAME    ENDPOINTS                           AGE
nginx   10.244.104.9:80,10.244.166.161:80   12s
```

```
$ kubectl get svc
NAME    TYPE        CLUSTER-IP       EXTERNAL-IP   PORT(S)   AGE
nginx   ClusterIP   10.108.169.159   <none>        80/TCP    4s

$ kubectl get pods -o wide
NAME                     READY   STATUS    RESTARTS   AGE   IP               NODE    NOMINATED NODE   READINESS GATES
nginx-b9d76c89c-cv7sg    1/1     Running   0          8s    10.244.166.161   node1   <none>           <none>
nginx-b9d76c89c-gd5ld    1/1     Running   0          8s    10.244.104.9     node2   <none>           <none>

$ kubectl get ep
NAME    ENDPOINTS                           AGE
nginx   10.244.104.9:80,10.244.166.161:80   12s
```

```
$ iptables -n -t nat -L PREROUTING
Chain PREROUTING (policy ACCEPT)
target         prot opt source               destination
KUBE-SERVICES  all  --  0.0.0.0/0            0.0.0.0/0            /* kubernetes service portals */

$ iptables -n -t nat -L KUBE-SERVICES
Chain KUBE-SERVICES (2 references)
target                     prot opt source               destination
KUBE-SVC-4Q6ZM72RGSBA3SUJ  tcp  --  0.0.0.0/0            10.108.169.159       /* kcd/nginx cluster IP */ tcp dpt:80
KUBE-SVC-NPX46M4PTMTKRN6Y  tcp  --  0.0.0.0/0            10.96.0.1            /* default/kubernetes:https cluster IP */ tcp
KUBE-SVC-TCOU7JCQXEZGVUNU  udp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns cluster IP */ udp
KUBE-SVC-ERIFXISQEP7F7OF4  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns-tcp cluster IP */
KUBE-SVC-JD5MR3NA4I4DYORP  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:metrics cluster IP */
```

```
$ iptables -n -t nat -L PREROUTING
Chain PREROUTING (policy ACCEPT)
target         prot opt source               destination
KUBE-SERVICES  all  --  0.0.0.0/0            0.0.0.0/0            /* kubernetes service portals */

$ iptables -n -t nat -L KUBE-SERVICES
Chain KUBE-SERVICES (2 references)
target                     prot opt source               destination
KUBE-SVC-4Q6ZM72RGSBA3SUJ  tcp  --  0.0.0.0/0            10.108.169.159       /* kcd/nginx cluster IP */ tcp dpt:80
KUBE-SVC-NPX46M4PTMTKRN6Y  tcp  --  0.0.0.0/0            10.96.0.1            /* default/kubernetes:https cluster IP */ tcp
KUBE-SVC-TCOU7JCQXEZGVUNU  udp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns cluster IP */ udp
KUBE-SVC-ERIFXISQEP7F7OF4  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns-tcp cluster IP */
KUBE-SVC-JD5MR3NA4I4DYORP  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:metrics cluster IP */
```

```
$ iptables -n -t nat -L PREROUTING
Chain PREROUTING (policy ACCEPT)
target         prot opt source               destination
KUBE-SERVICES  all  --  0.0.0.0/0            0.0.0.0/0            /* kubernetes service portals */

$ iptables -n -t nat -L KUBE-SERVICES
Chain KUBE-SERVICES (2 references)
target                     prot opt source               destination
KUBE-SVC-4Q6ZM72RGSBA3SUJ  tcp  --  0.0.0.0/0            10.108.169.159       /* kcd/nginx cluster IP */ tcp dpt:80
KUBE-SVC-NPX46M4PTMTKRN6Y  tcp  --  0.0.0.0/0            10.96.0.1            /* default/kubernetes:https cluster IP */ tcp
KUBE-SVC-TCOU7JCQXEZGVUNU  udp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns cluster IP */ udp
KUBE-SVC-ERIFXISQEP7F7OF4  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns-tcp cluster IP */
KUBE-SVC-JD5MR3NA4I4DYORP  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:metrics cluster IP */
```

```
$ iptables -n -t nat -L PREROUTING
Chain PREROUTING (policy ACCEPT)
target         prot opt source               destination
KUBE-SERVICES  all  --  0.0.0.0/0            0.0.0.0/0            /* kubernetes service portals */

$ iptables -n -t nat -L KUBE-SERVICES
Chain KUBE-SERVICES (2 references)
target                     prot opt source               destination
KUBE-SVC-4Q6ZM72RGSBA3SUJ  tcp  --  0.0.0.0/0            10.108.169.159       /* kcd/nginx cluster IP */ tcp dpt:80
KUBE-SVC-NPX46M4PTMTKRN6Y  tcp  --  0.0.0.0/0            10.96.0.1            /* default/kubernetes:https cluster IP */ tcp
KUBE-SVC-TCOU7JCQXEZGVUNU  udp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns cluster IP */ udp
KUBE-SVC-ERIFXISQEP7F7OF4  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns-tcp cluster IP */
KUBE-SVC-JD5MR3NA4I4DYORP  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:metrics cluster IP */
```

```
$ iptables -n -t nat -L PREROUTING
Chain PREROUTING (policy ACCEPT)
target         prot opt source               destination
KUBE-SERVICES  all  --  0.0.0.0/0            0.0.0.0/0            /* kubernetes service portals */

$ iptables -n -t nat -L KUBE-SERVICES
Chain KUBE-SERVICES (2 references)
target                     prot opt source               destination
KUBE-SVC-4Q6ZM72RGSBA3SUJ  tcp  --  0.0.0.0/0            10.108.169.159       /* kcd/nginx cluster IP */ tcp dpt:80
KUBE-SVC-NPX46M4PTMTKRN6Y  tcp  --  0.0.0.0/0            10.96.0.1            /* default/kubernetes:https cluster IP */ tcp
KUBE-SVC-TCOU7JCQXEZGVUNU  udp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns cluster IP */ udp
KUBE-SVC-ERIFXISQEP7F7OF4  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns-tcp cluster IP */
KUBE-SVC-JD5MR3NA4I4DYORP  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:metrics cluster IP */
```

```
$ iptables -n -t nat -L PREROUTING
Chain PREROUTING (policy ACCEPT)
target         prot opt source               destination
KUBE-SERVICES  all  --  0.0.0.0/0            0.0.0.0/0            /* kubernetes service portals */

$ iptables -n -t nat -L KUBE-SERVICES
Chain KUBE-SERVICES (2 references)
target                     prot opt source               destination
KUBE-SVC-4Q6ZM72RGSBA3SUJ  tcp  --  0.0.0.0/0            10.108.169.159       /* kcd/nginx cluster IP */ tcp dpt:80
KUBE-SVC-NPX46M4PTMTKRN6Y  tcp  --  0.0.0.0/0            10.96.0.1            /* default/kubernetes:https cluster IP */ tcp
KUBE-SVC-TCOU7JCQXEZGVUNU  udp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns cluster IP */ udp
KUBE-SVC-ERIFXISQEP7F7OF4  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:dns-tcp cluster IP */
KUBE-SVC-JD5MR3NA4I4DYORP  tcp  --  0.0.0.0/0            10.96.0.10           /* kube-system/kube-dns:metrics cluster IP */
```

```
$ kubectl get svc -n kcd
NAME    TYPE        CLUSTER-IP       EXTERNAL-IP   PORT(S)   AGE
nginx   ClusterIP   10.108.169.159   <none>        80/TCP    4s
```

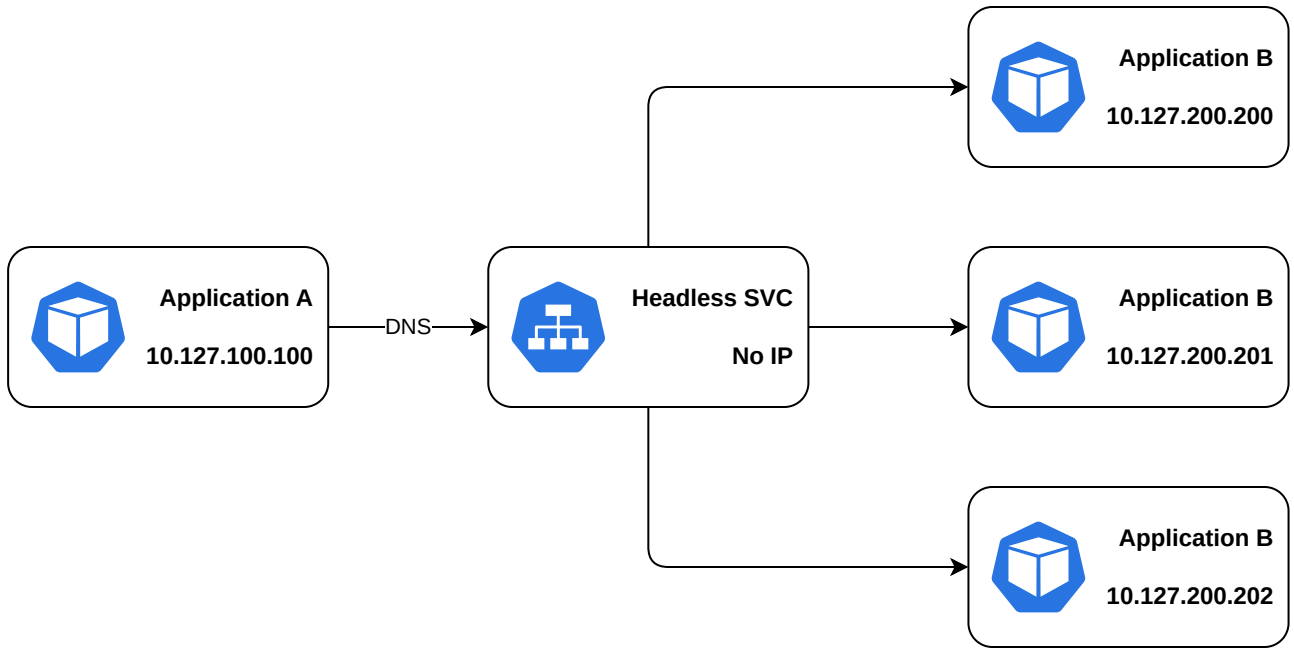
Cilium

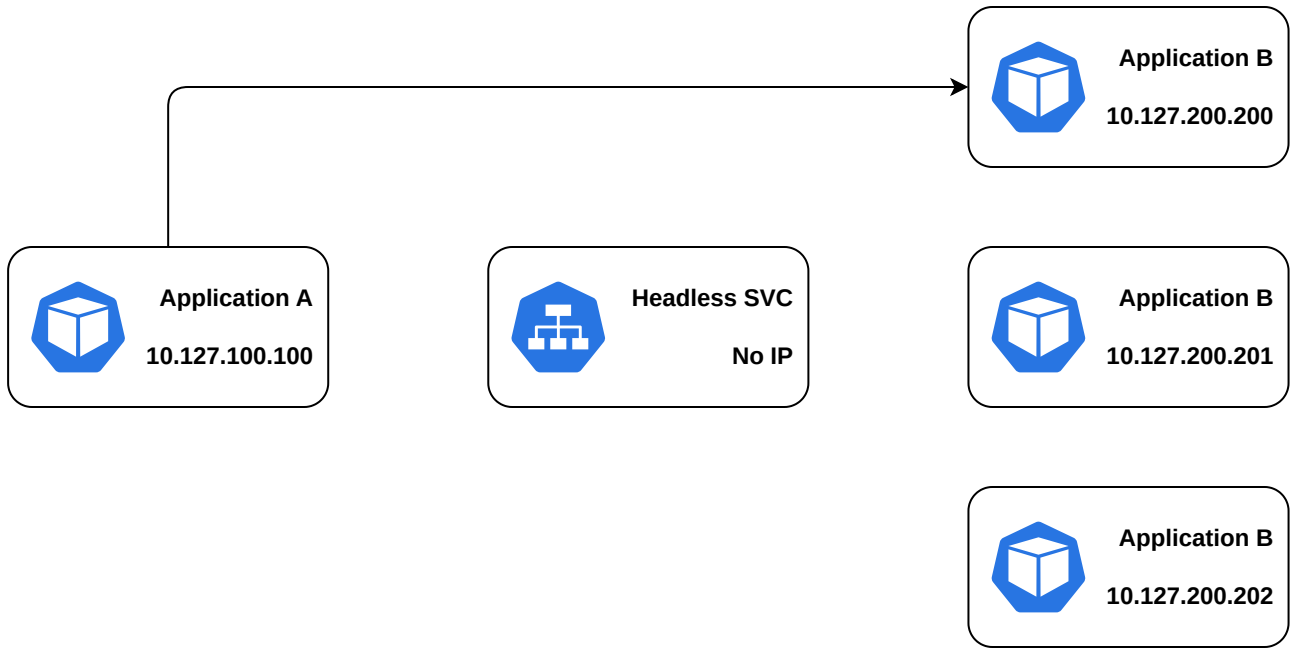
- **C**ontainer **N**etwork **I**nterface (CNI)
- Runs on every node
- Uses eBPF
- Can replace kube-proxy
- Lookup is linear (iptables)

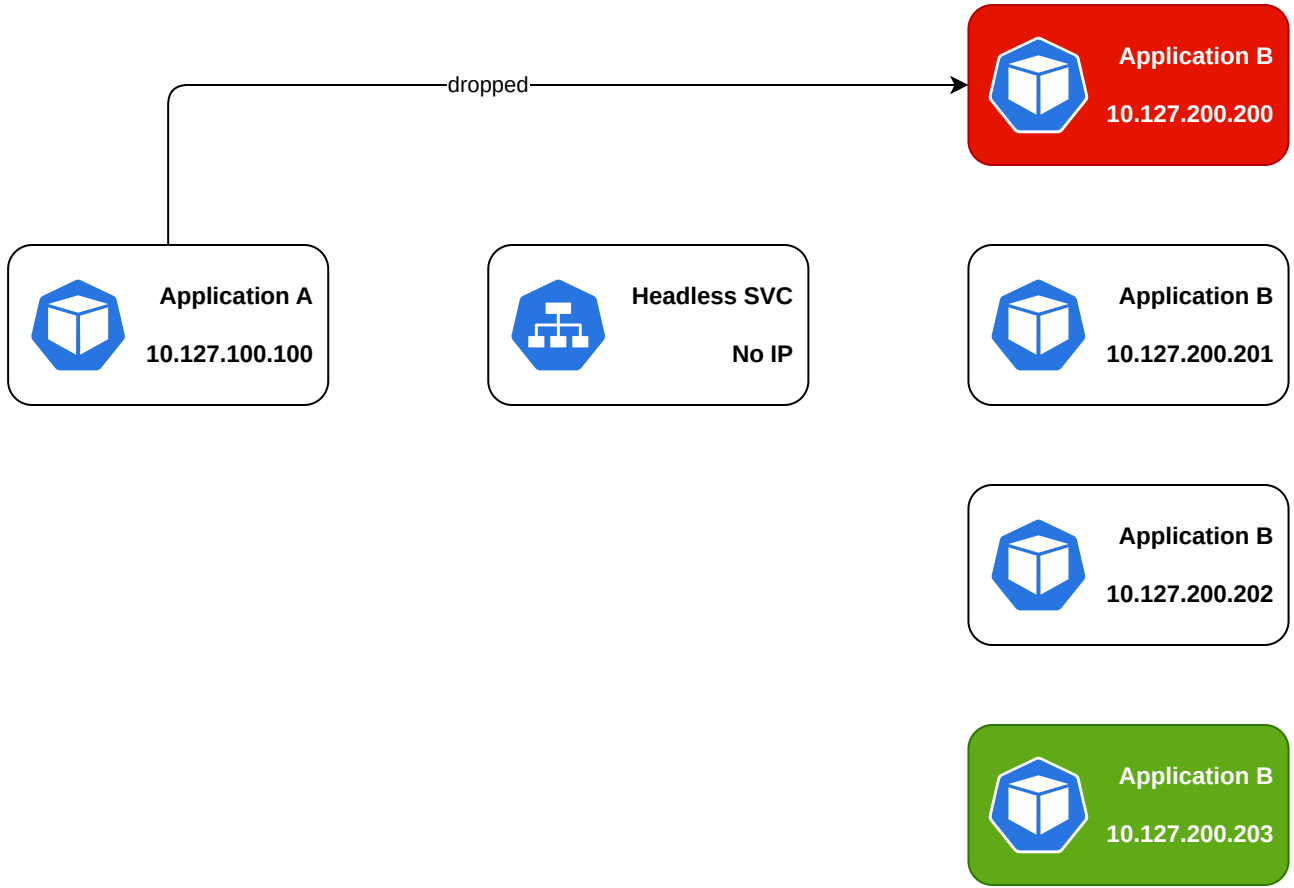
TYPE	IPTABLES BASED CNI (MAX)	CILIUM (MAX)
Outgoing connection	75ms	48ms
Connection to k8s service	12s	12ms
Connection to pod IP	60s	5ms

Packet Drops

- Cilium stops forwarding packets
- Something doesn’t work as intended
  - Alert on metric --> `cilium_drop_count_total`
- Multiple causes
  - Unknown destination
  - Violating NetworkPolicies
  - …


Application A
10.127.100.100
DNS
Headless SVC
No IP
Application B
10.127.200.200
Application B
10.127.200.201
Application B
10.127.200.202

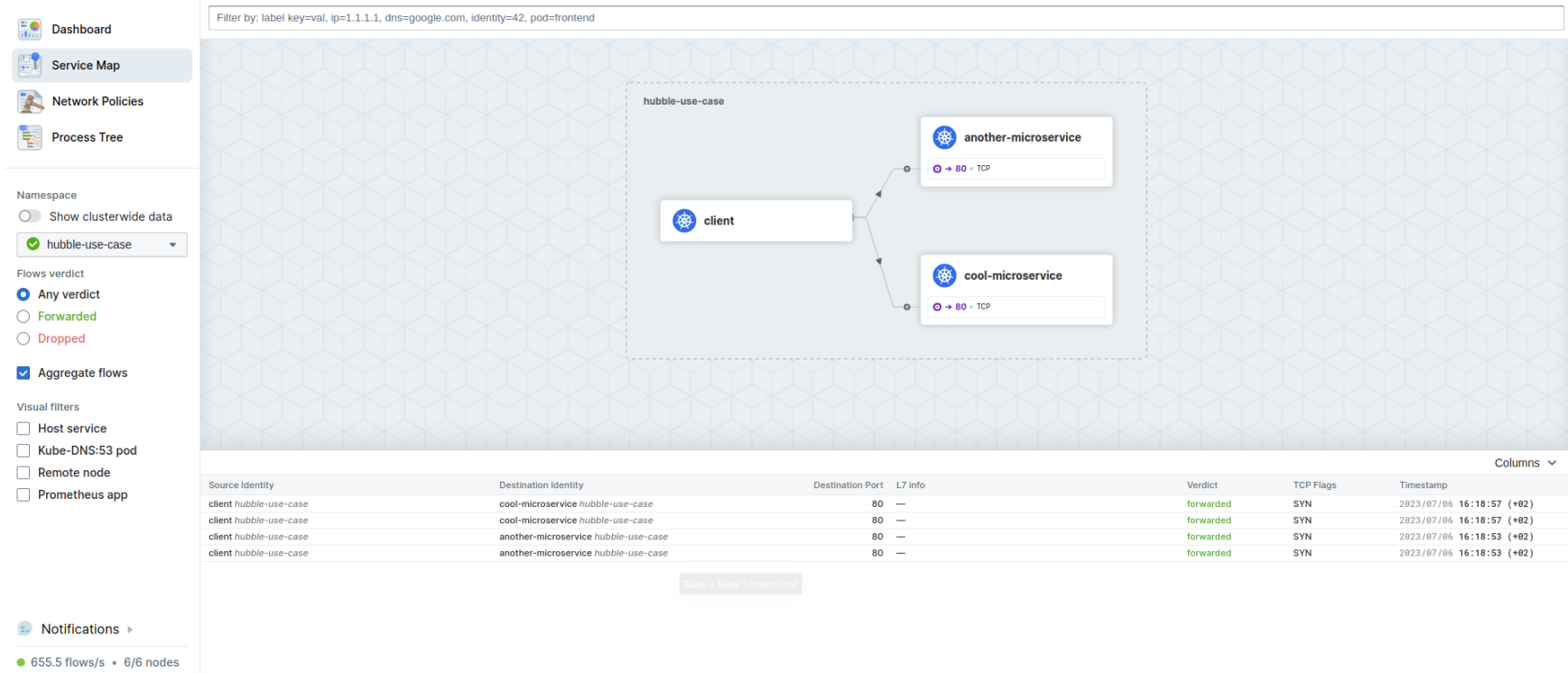

Application B
10.127.200.200
Application A
10.127.100.100
Headless SVC
No IP
Application B
10.127.200.201
Application B
10.127.200.202


dropped
Application B
10.127.200.200
Application A
10.127.100.100
Headless SVC
No IP
Application B
10.127.200.201
Application B
10.127.200.202
Application B
10.127.200.203

Hubble

- Observe *current* traffic
- Runs on every node
- CLI and GUI
- Allows for filtering

```
$ hubble observe --from-namespace hubble-demo -t drop -f
Jun  4 14:03:38.365: hubble-demo/application-a:60886 <> 10.127.200.200:80 Stale or unroutable IP DROPPED TCP Flags: SYN
Jun  4 14:03:39.379: hubble-demo/application-a:60886 <> 10.127.200.200:80 Stale or unroutable IP DROPPED TCP Flags: SYN
Jun  4 14:03:41.395: hubble-demo/application-a:60886 <> 10.127.200.200:80 Stale or unroutable IP DROPPED TCP Flags: SYN
Jun  4 14:03:45.619: hubble-demo/application-a:60886 <> 10.127.200.200:80 Stale or unroutable IP DROPPED TCP Flags: SYN
Jun  4 14:03:53.811: hubble-demo/application-a:60886 <> 10.127.200.200:80 Stale or unroutable IP DROPPED TCP Flags: SYN
Jun  4 14:04:09.939: hubble-demo/application-a:60886 <> 10.127.200.200:80 Stale or unroutable IP DROPPED TCP Flags: SYN
```


Dashboard
Service Map
Network Policies
Process Tree
Namespace
Show clusterwide data
hubble-use-case
Flows verdict
Any verdict
Forwarded
Dropped
Aggregate flows
Visual filters
Host service
Kube-DNS:53 pod
Remote node
Prometheus app
Notifications
655.5 flows/s • 6/6 nodes
Filter by: label key=val, ip=1.1.1.1, dns=google.com, identity=42, pod=frontend
hubble-use-case
another-microservice
80 TCP
client
cool-microservice
80 TCP
Columns
Source Identity
Destination Identity
Destination Port
L7 info
Verdict
TCP Flags
Timestamp
client hubble-use-case
cool-microservice hubble-use-case
80
—
forwarded
SYN
2023/07/06 16:18:57 (+02)
client hubble-use-case
cool-microservice hubble-use-case
80
—
forwarded
SYN
2023/07/06 16:18:57 (+02)
client hubble-use-case
another-microservice hubble-use-case
80
—
forwarded
SYN
2023/07/06 16:18:53 (+02)
client hubble-use-case
another-microservice hubble-use-case
80
—
forwarded
SYN
2023/07/06 16:18:53 (+02)
Take a New Screenshot

What can you do?

What can you do?

For new clusters:

What can you do?

For new clusters:

```
$ kubeadm init --skip-phases=addon/kube-proxy
```

What can you do?

For new clusters:

```
$ kubeadm init --skip-phases=addon/kube-proxy
```

```
$ helm install cilium cilium/cilium --set kubeProxyReplacement=strict
```

What can you do?

For new clusters:

```
$ kubeadm init --skip-phases=addon/kube-proxy
```

```
$ helm install cilium cilium/cilium --set kubeProxyReplacement=strict
```

For existing clusters:

What can you do?

For new clusters:

```
$ kubeadm init --skip-phases=addon/kube-proxy
```

```
$ helm install cilium cilium/cilium --set kubeProxyReplacement=strict
```

For existing clusters:

```
$ kubectl -n kube-system delete ds kube-proxy
$ kubectl -n kube-system delete cm kube-proxy
$ iptables-save | grep -v KUBE | iptables-restore # *
```

* Warning: Service connections will not work until replacement is installed

What can you do?

For new clusters:

```
$ kubeadm init --skip-phases=addon/kube-proxy
```

```
$ helm install cilium cilium/cilium --set kubeProxyReplacement=strict
```

For existing clusters:

```
$ kubectl -n kube-system delete ds kube-proxy
$ kubectl -n kube-system delete cm kube-proxy
$ iptables-save | grep -v KUBE | iptables-restore # *
```

```
$ helm install cilium cilium/cilium --set kubeProxyReplacement=strict
```

* Warning: Service connections will not work until replacement is installed

Questions?