K8S klaszterek összekötése adatközponti hálózati környezetben Isovalent – Cilium CNI

Nagy Tibor
Technical Solutions Architect
tinagy@cisco.com

Agenda

01 What is Cilium/Isovalent Enterprise

02 Nexus Fabric Design – Overlays and Underlays

- Software Overlay
- Native IP routing
- EVPN/VXLAN

03 Observability / Operations

- Hubble

What is Cilium/Isovalent Enterprise

Kubernetes – The Emerging Standard To Deploy All Applications

A Hybrid and Multi-Cloud Infrastructure Layer Supporting Both Modern & Legacy Workloads

Traditional VM Workloads

Modern Applications

AI / ML Workloads

kubernetes

ISOVALENT

Networking / Routing

Firewall / Microseg

Load Balancing

Network Visibility

VPN / Encryption

Runtime Security

BARE METAL

vmware

aws

Azure

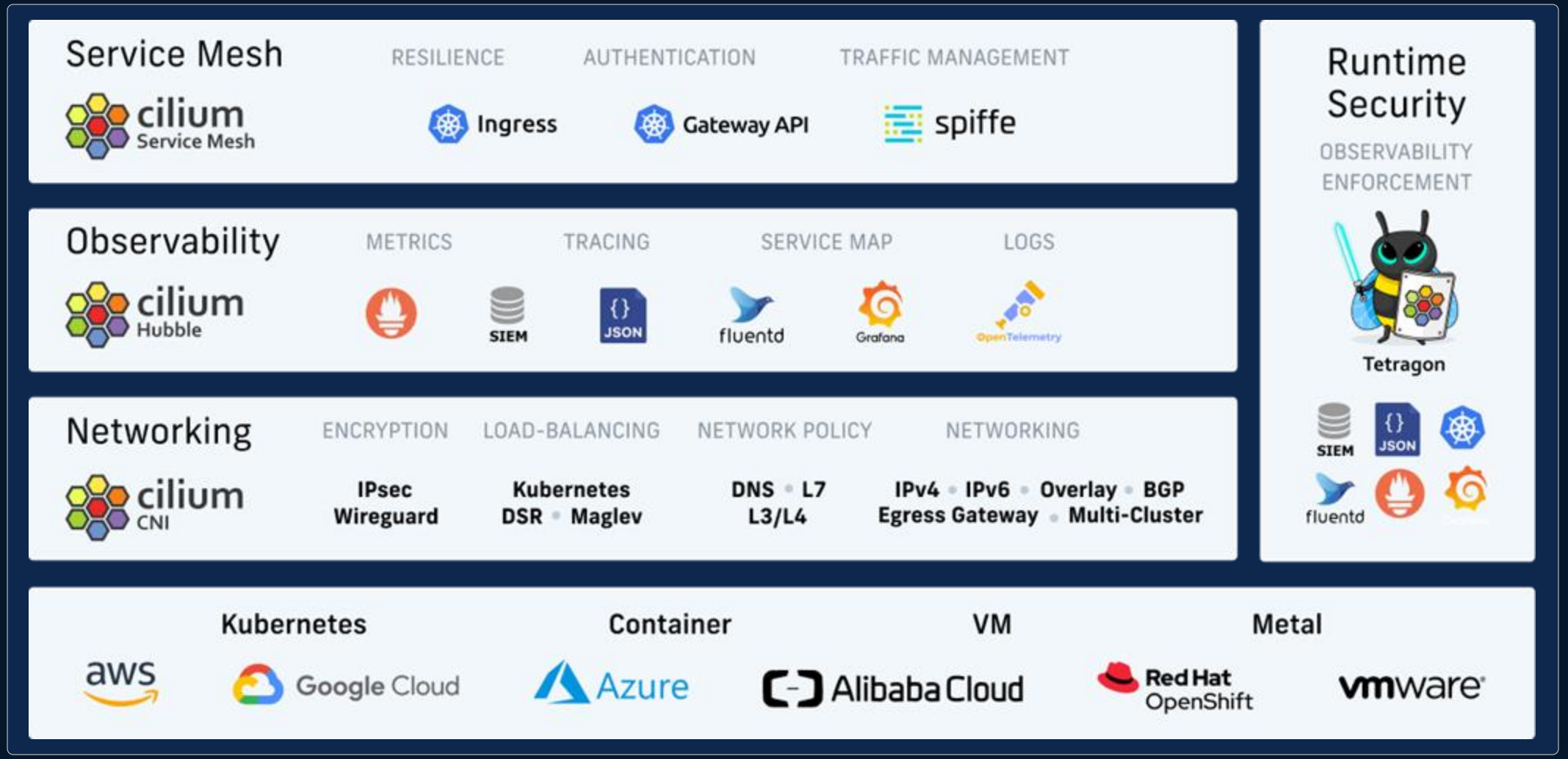
Google Cloud

ORACLE Cloud Infrastructure

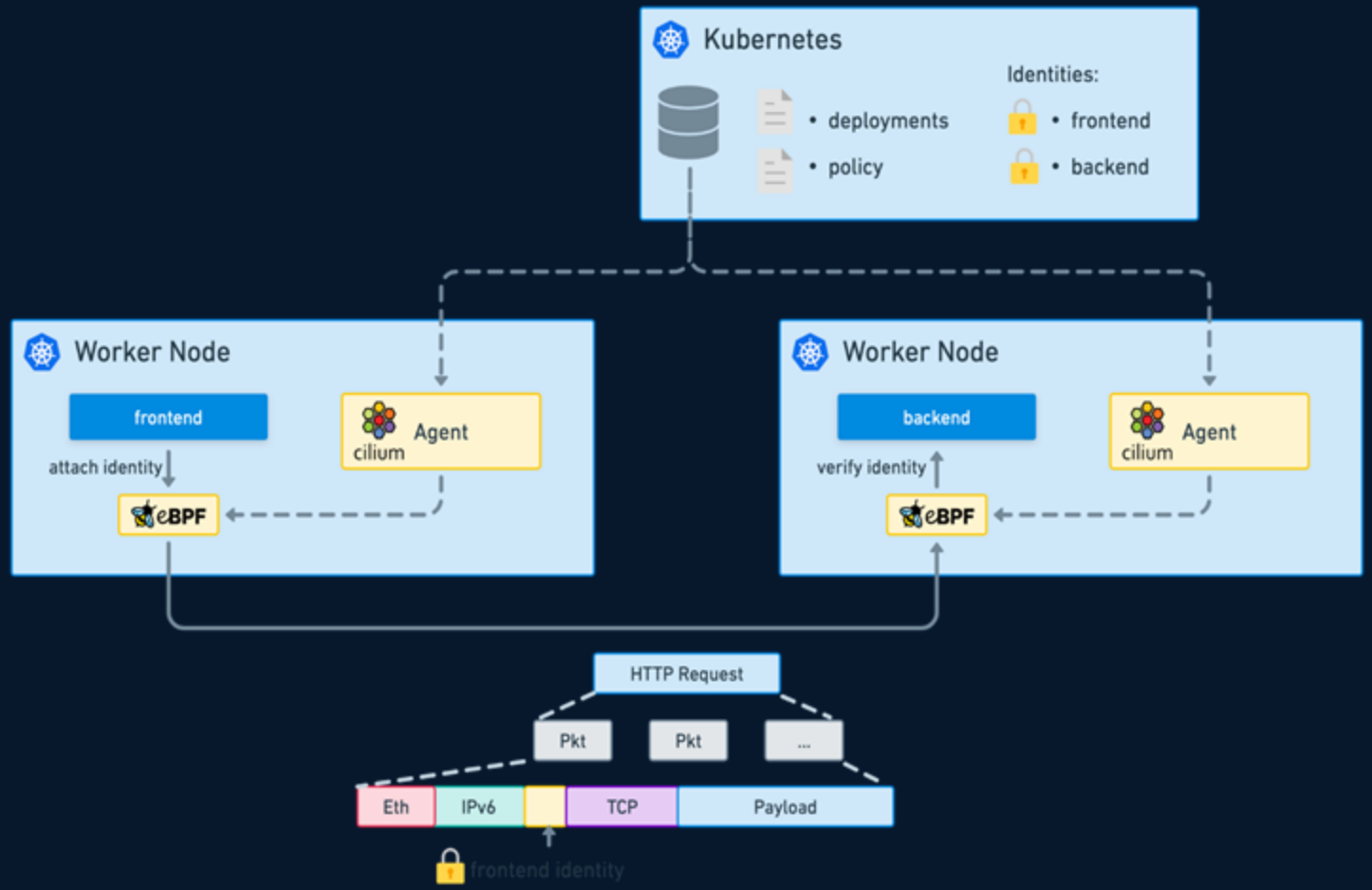
Alibaba Cloud

cisco

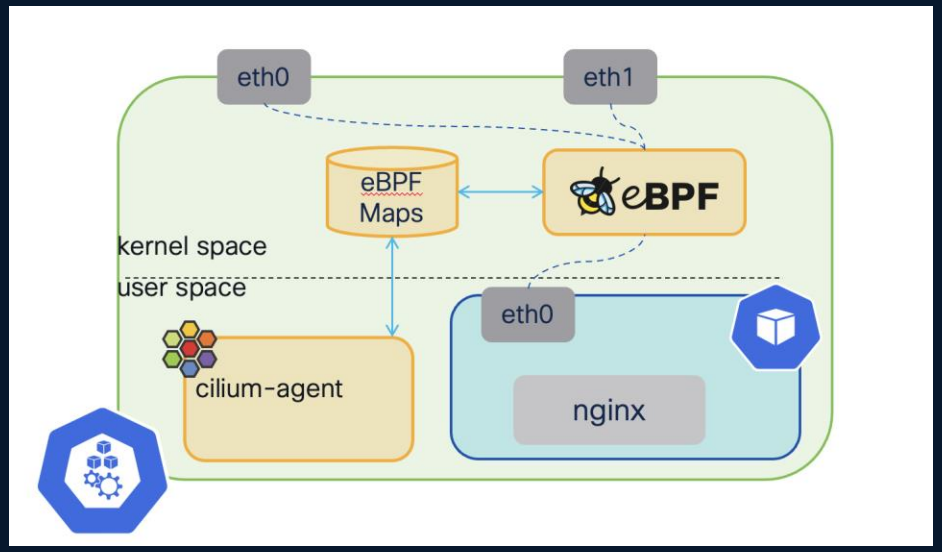
Isovalent Enterprise Platform



Identity-based Security for Network Policy

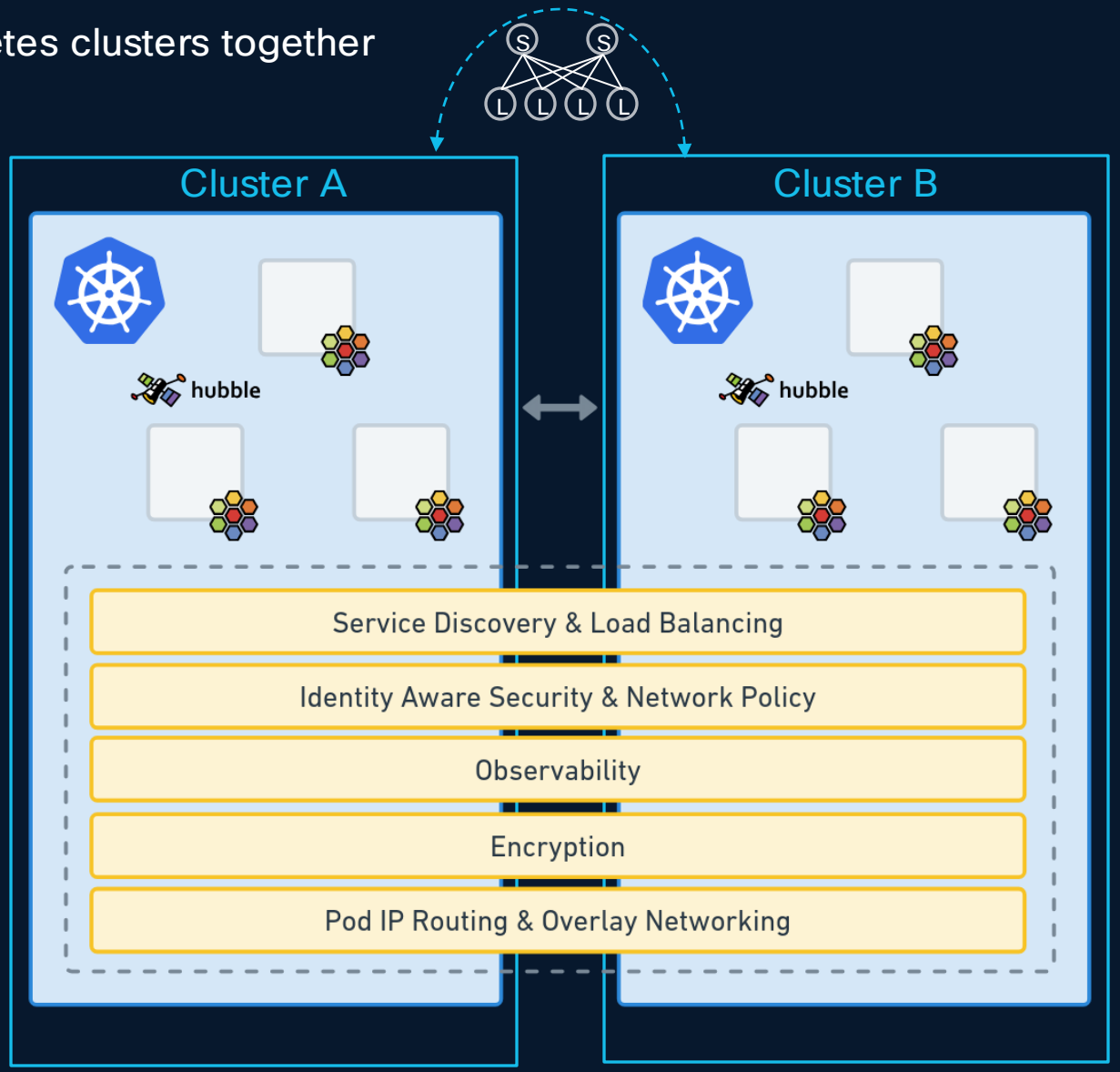


Frontend identity



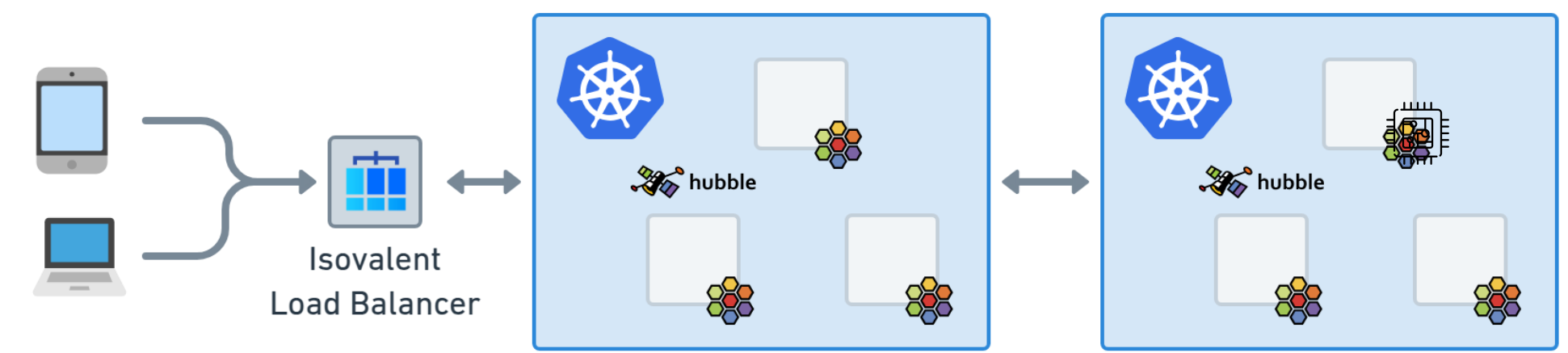
Expanded to Multi-Cluster Networking & Security

Securely Connect Kubernetes clusters together



Isovalent Load Balancer

A modern L3-L7 Load Balancer with Native Kubernetes Capabilities



aws Google Cloud Azure OPENSHIFT

Networking

- Container Networking
- L3-L4 Load Balancing
- Standalone LB
- Multi-Cluster

splunk>

Network Security

- Network Policy L3-L7
- Encryption
- SIEM Integration

Grafana Prometheus

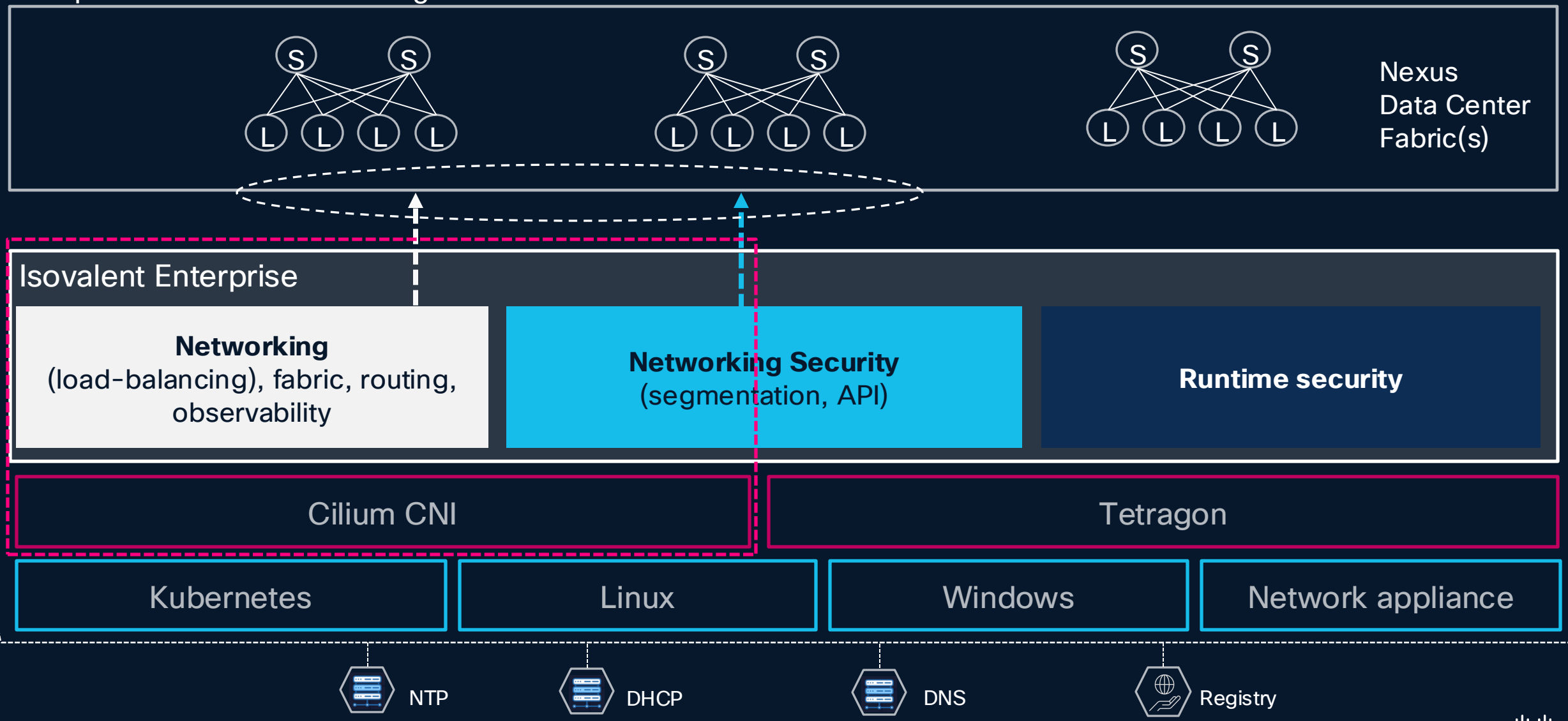
Network Observability

- Flow Logs
- Metrics
- Troubleshooting

Nexus NXOS Fabric Design with Cilium Enterprise

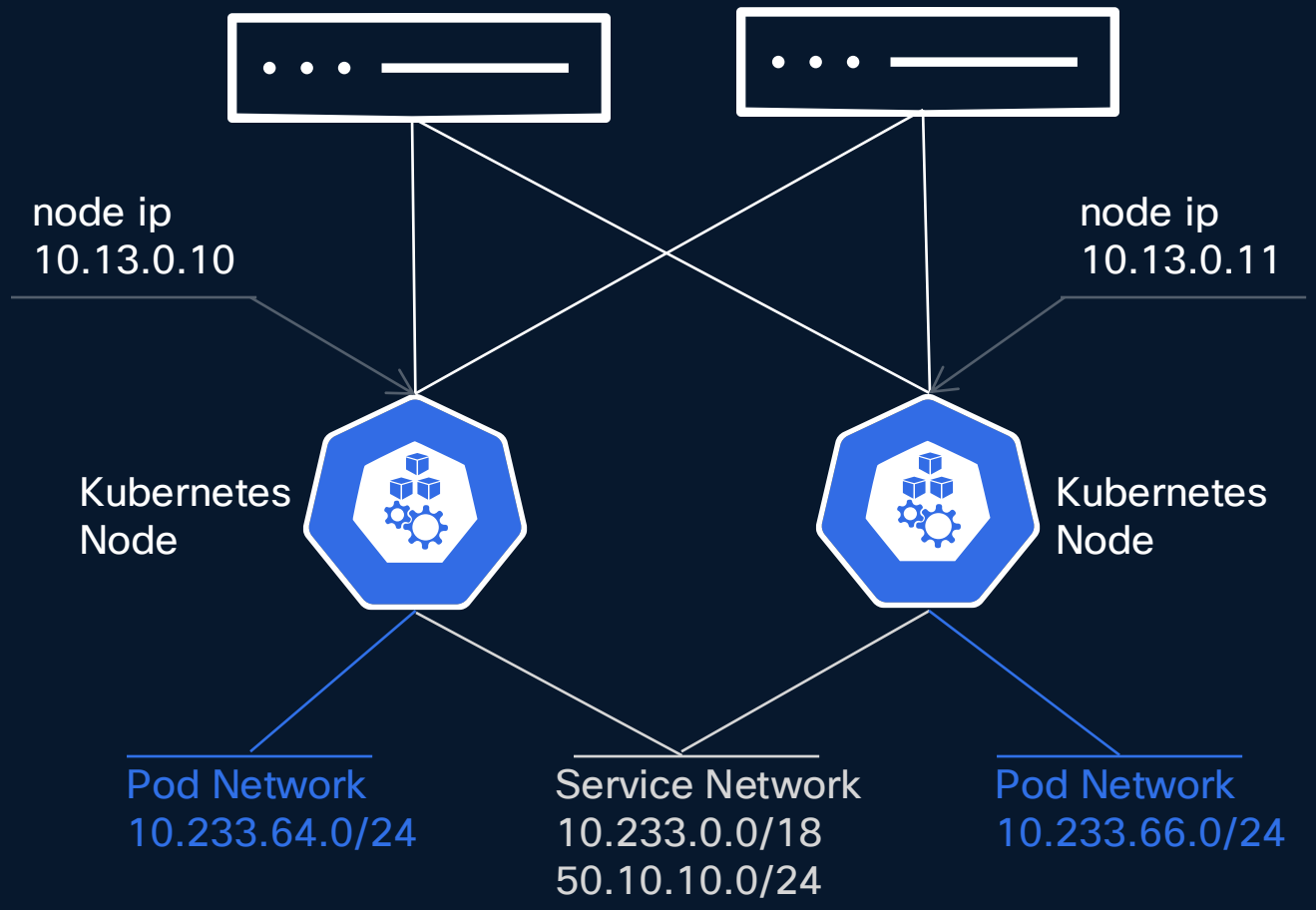
Isovalent

Enterprise Platform – Networking with CNI



Container Network Interface

Simplified



- Each Kubernetes node has one node IP
- One or more ranges of IP addresses (CIDRs) for pod networks
- One or more service networks shared by the cluster

Fabric Underlay to Cilium CNI Deployment Models

1

Software Overlay using Cilium's CNI

2

Native IP Routing – Cilium BGP Peering to Fabric

No EVPN/VXLAN

3

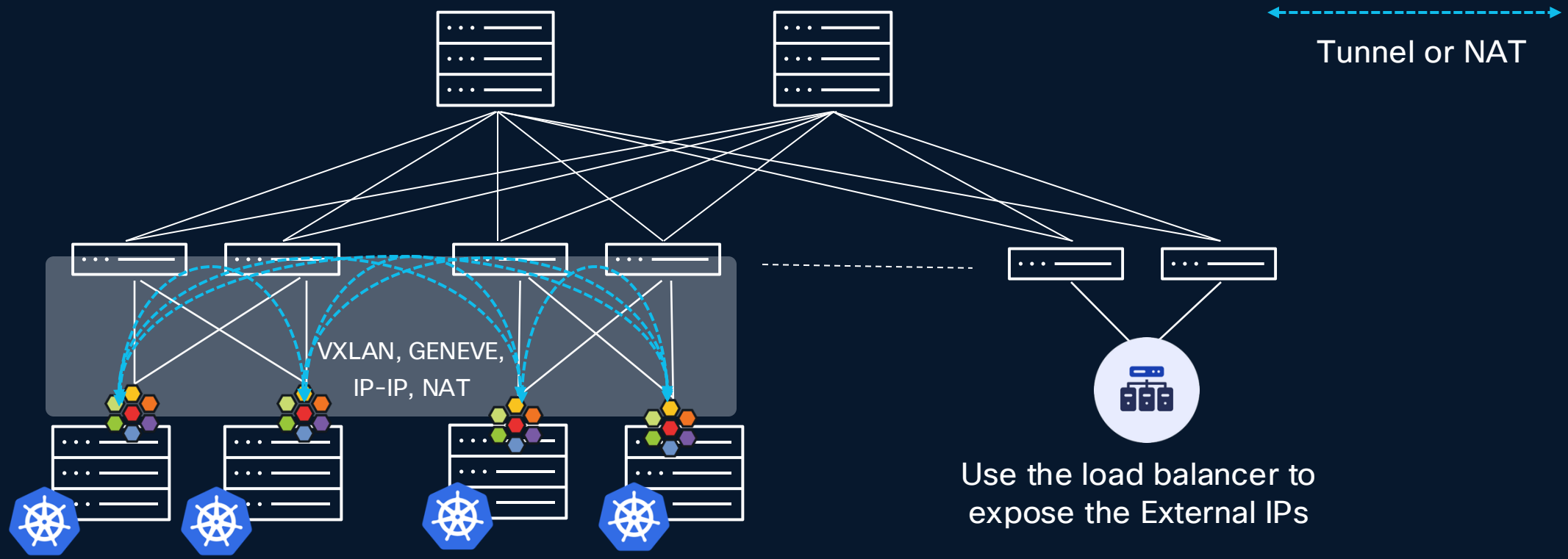
Cilium CNI BGP peers with EVPN/VXLAN Fabric

Software Overlay Using Cilium’s CNI

Software Overlay using Cilium’s CNI

Software Overlay using Cilium's CNI

- The Network is just an underlay, carries only Host/Node IP information (looks like any other server)
- Abstraction of network visibility (No network awareness of software overlays or NAT)
- Need an External Load Balancer for the service exposure



Native IP Routing – Cilium BGP Peering to Fabric

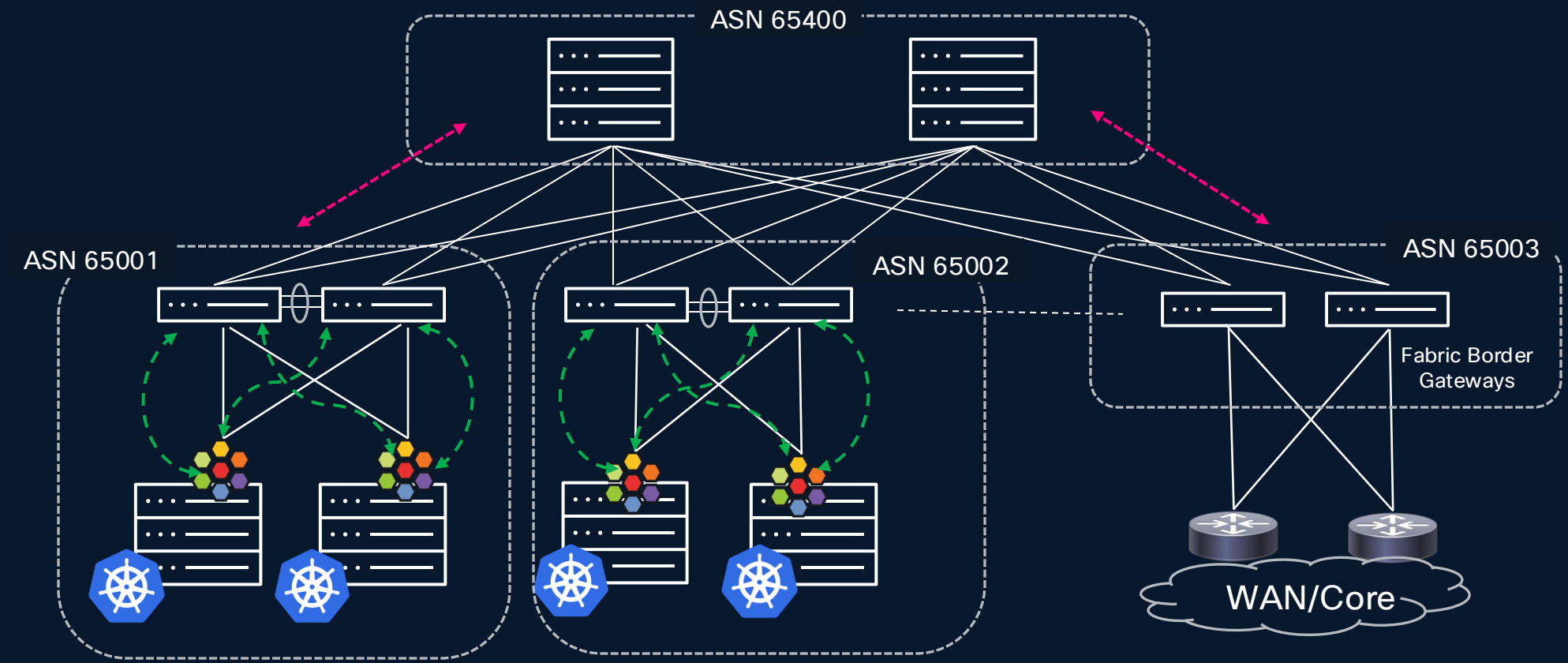
RFC 7938 - BGP for Routing in Large-Scale Data Centers

Native IP Routing – Cilium BGP Peering to Fabric

No EVPN/VXLAN

BGP Network on IP Fabric

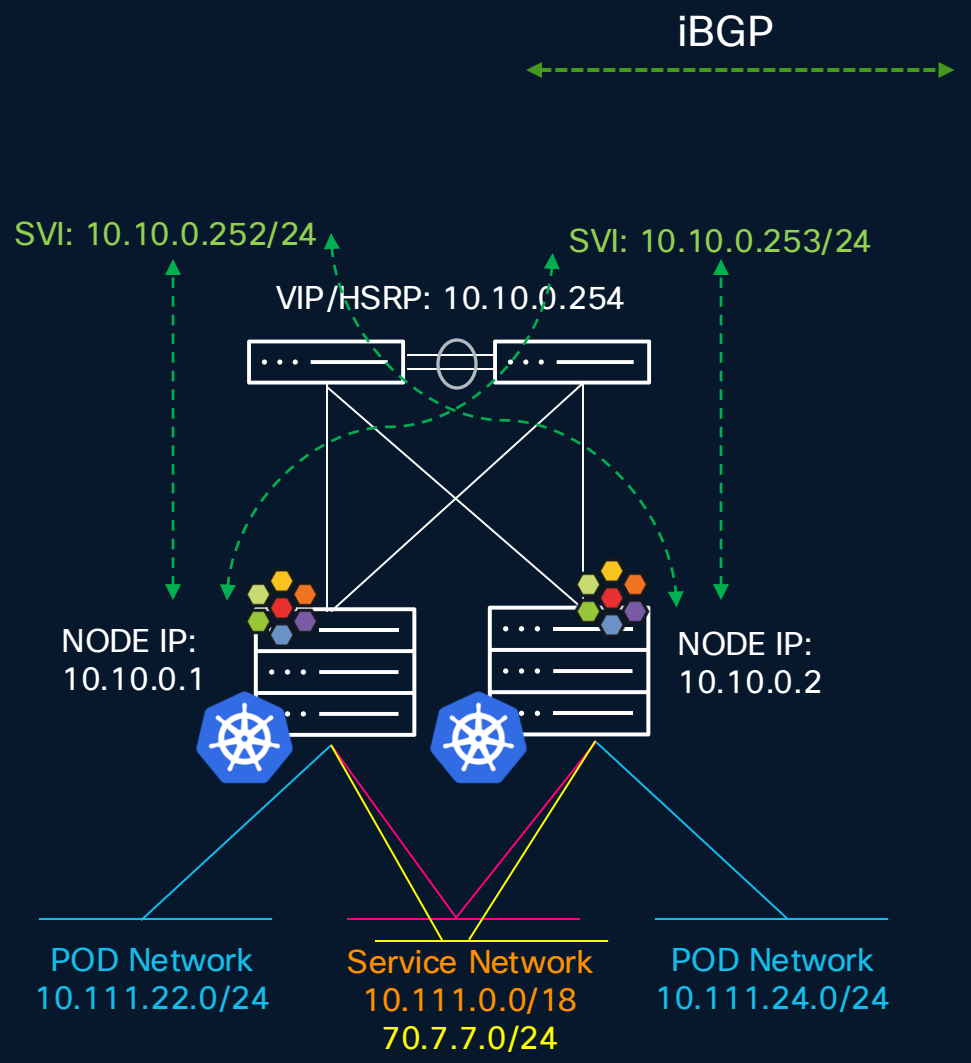
- Scalable approach, the leaf switches become Route-Reflector Route-Server
- Data is transported with the original headers
- AS-per-Rack Design
- Exclusively for IP Fabric (RFC 7938 - BGP for Routing in Large-Scale Data Centers)



BGP Network on IP Fabric

Cilium CNI peers with switches

- HSRP/VRRP is used for gateway redundancy
- Each Kubernetes Node peers with the Primary SVI IP of the redundant switches
- The POD Networks and External IPs (e.g., load balancer, egress gateway, api gateway or node IPs) are advertised into the fabric
- Only specific Service Networks are advertised to the Fabric – in this example we show 70.7.7.0/24 being advertised

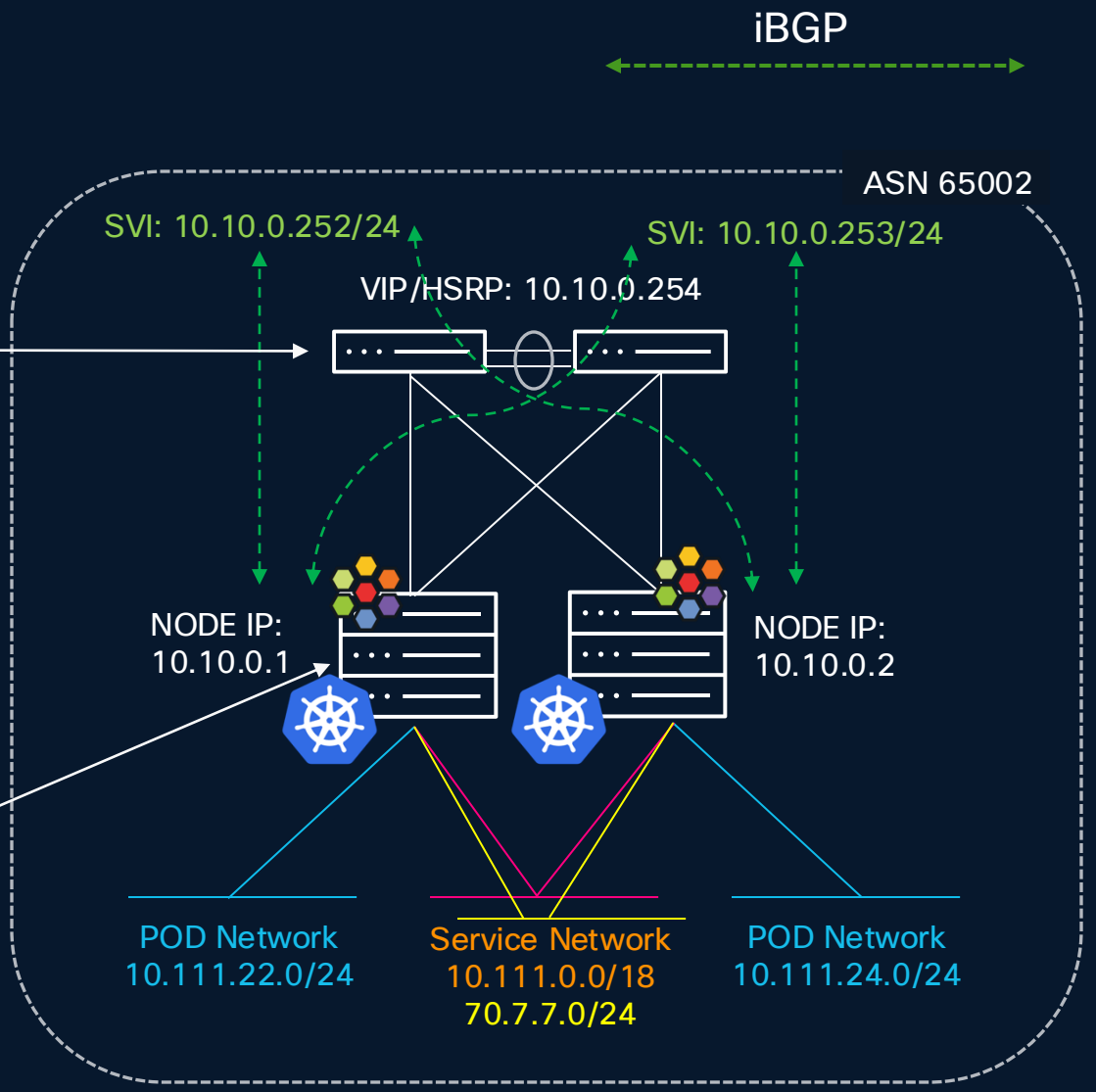


BGP Network on IP Fabric

Cilium CNI peers with switches

```
ToR Configs
interface Vlan1001
ip address 10.10.0.252/24
  no shutdown
  vrf member vrf-TenantA
  hsrp 10
    ip 10.10.0.254
    priority 101
  end
!
Router bgp 65002
template peer Cilium_K8s_TenantA
    remote-as 65002
    update-source loopback1001
    ebgp-multihop 5
    address-family ipv4 unicast
    as-override
vrf vrf-tenantA
    neighbor 10.10.0.1
      inherit peer Cilium_K8s_TenantA
      address-family ipv4 unicast
    neighbor 10.10.0.2
      inherit peer Cilium_K8s_TenantA
      address-family ipv4 unicast
```

```yaml
apiVersion: isovalent.com/v1alpha1
kind: IsovalentBGPClusterConfig
metadata:
  name: cilium-bgp
spec:
  bgpInstances:
  - name: "instance-65002"
    localASN: 65002
    peers:
    - name: "leaf101"
      peerASN: 65002
      peerAddress: 10.10.0.252
      peerConfigRef:
        name: "cilium-peer"
    - name: "leaf102"
      peerASN: 65002
      peerAddress: 10.10.0.253
      peerConfigRef:
        name: "cilium-peer"
```



Exposing Services in Kubernetes with Cilium

Note on ”externalTrafficPolicy”

External Traffic Policy Denotes if this Service desires to route external traffic to node-local or cluster-wide endpoints.

- externalTrafficPolicy == Cluster
  - Pros: Overall good load-balance between pods
  - Cons: Potential second hop which will bring additional latency
- externalTrafficPolicy == Local
  - Pros: Avoid the second hop, source IP is preserved
  - Cons: Potentially imbalanced workload spreading
  - Use nexthop unchanged to overcome
  - Pods can be spread evenly with topologySpreadConstraints

*Note: Ingress Controller (e.g., Load balancer, API gateway, etc.) must be External Traffic Policy aware

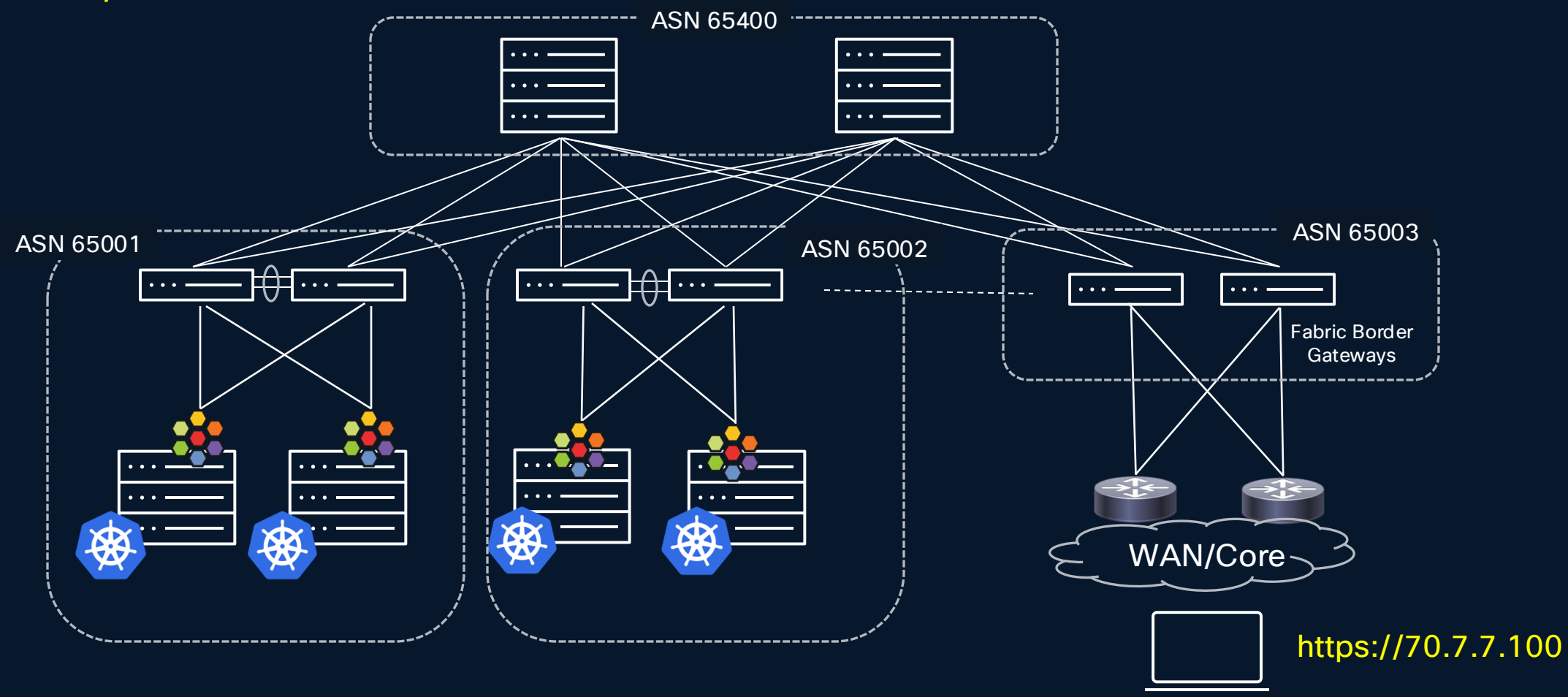
BGP Network on IP Fabric

Service Traffic w/out bestpath command

```
router bgp 65400
  bestpath as-path multipath-relax
```

```
70.7.7.100, ubest/mbest: 2/0
    *via 10.5.0.21, [20/0], 2d10h, bgp-65400, external, tag 65001
    *via 10.5.0.29, [20/0], 2d10h, bgp-65400, external, tag 65001
```

External Service IP:
70.7.7.100/32



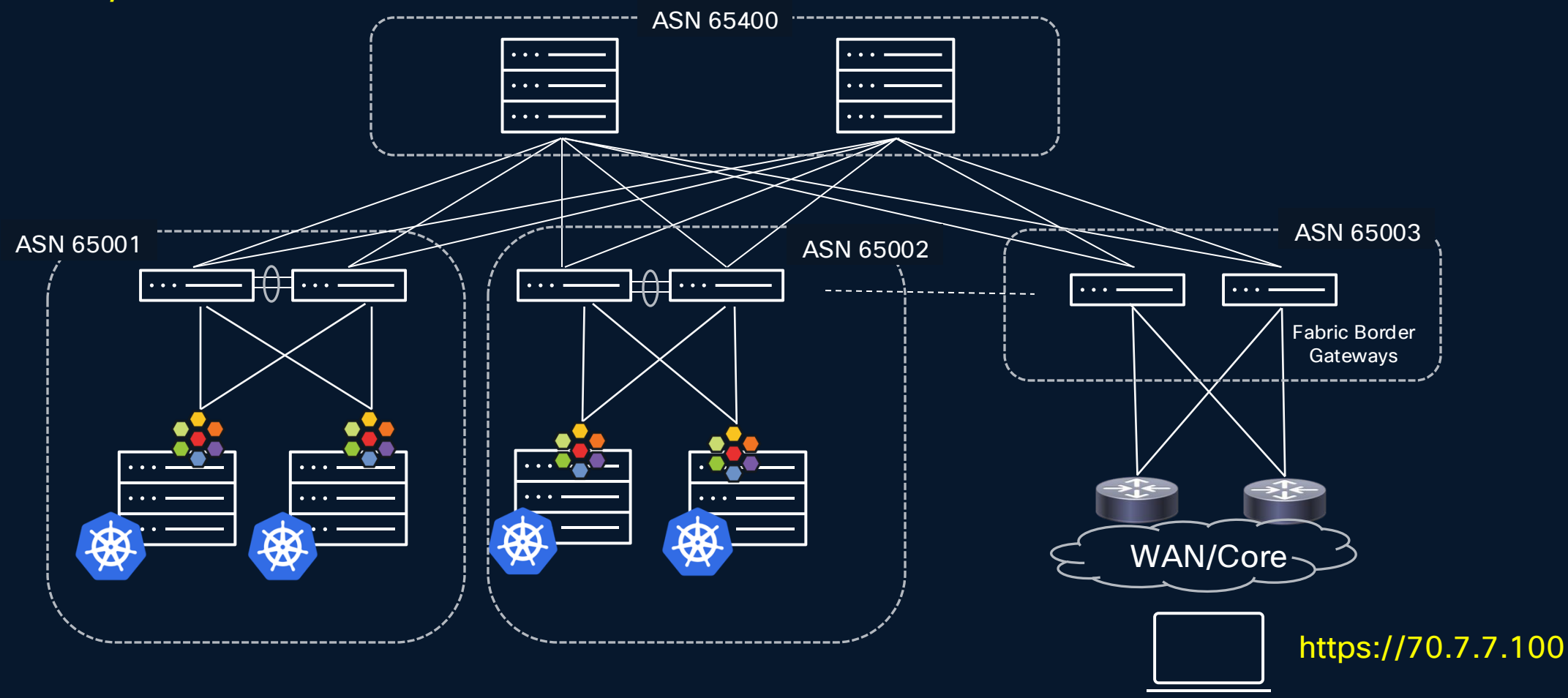
BGP Network on IP Fabric

Service Traffic with bestpath command

```
router bgp 65400
  bestpath as-path multipath-relax
```

```
70.7.7.100, ubest/mbest: 4/0
   *via 10.5.0.21, [20/0], 2d10h, bgp-65400, external, tag 65001
   *via 10.5.0.29, [20/0], 2d10h, bgp-65400, external, tag 65001
   *via 10.5.0.37, [20/0], 2d10h, bgp-65400, external, tag 65002
   *via 10.5.0.45, [20/0], 2d10h, bgp-65400, external, tag 65002
```

External Service IP:
70.7.7.100/32



BGP Network on IP Fabric

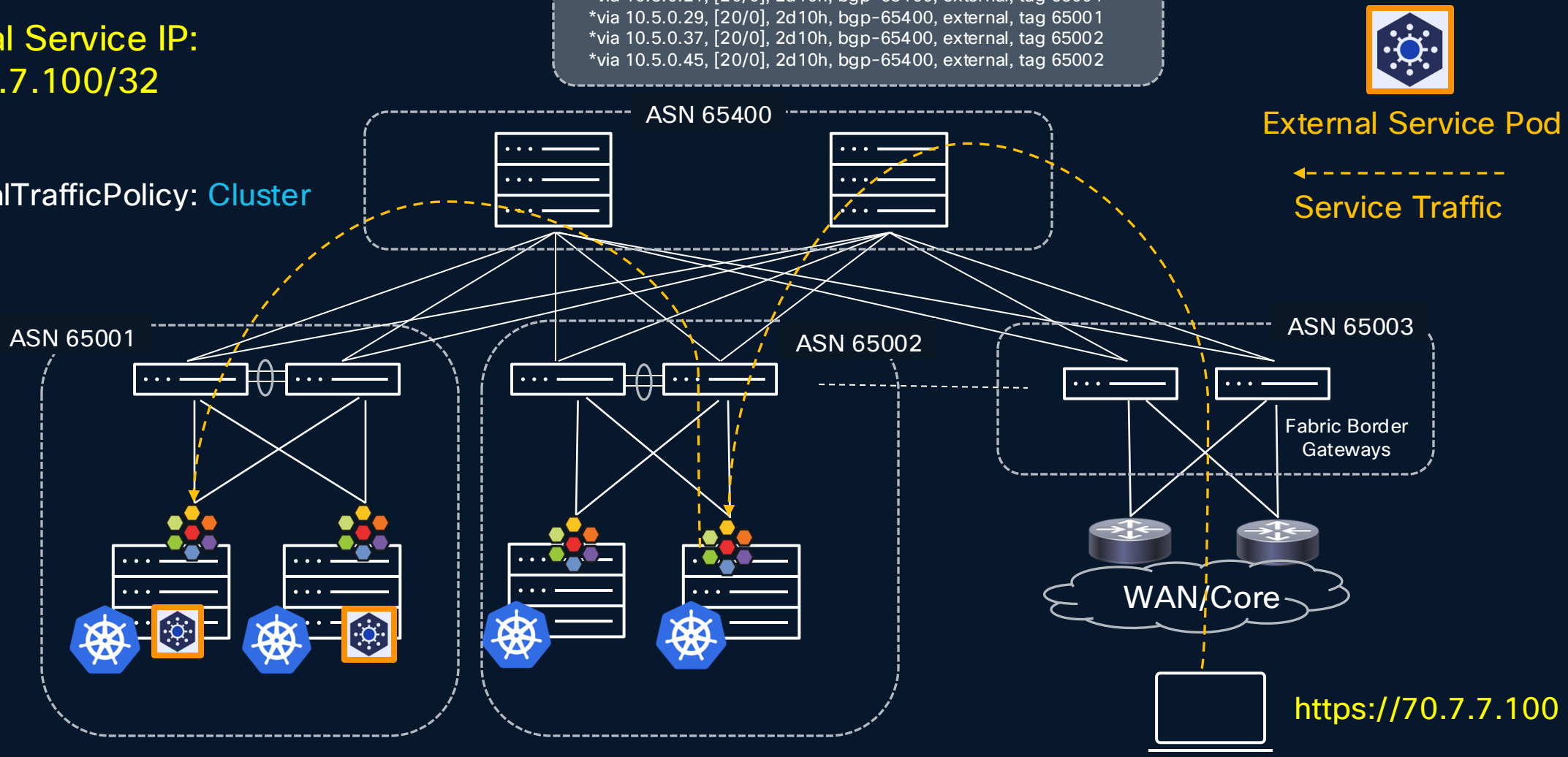
Service Traffic with bestpath command

```
router bgp 65400
  bestpath as-path multipath-relax
```

```
70.7.7.100, ubest/mbest: 4/0
    *via 10.5.0.21, [20/0], 2d10h, bgp-65400, external, tag 65001
    *via 10.5.0.29, [20/0], 2d10h, bgp-65400, external, tag 65001
    *via 10.5.0.37, [20/0], 2d10h, bgp-65400, external, tag 65002
    *via 10.5.0.45, [20/0], 2d10h, bgp-65400, external, tag 65002
```

External Service IP:
70.7.7.100/32

externalTrafficPolicy: Cluster



External Service Pod

Service Traffic

BGP Network on IP Fabric

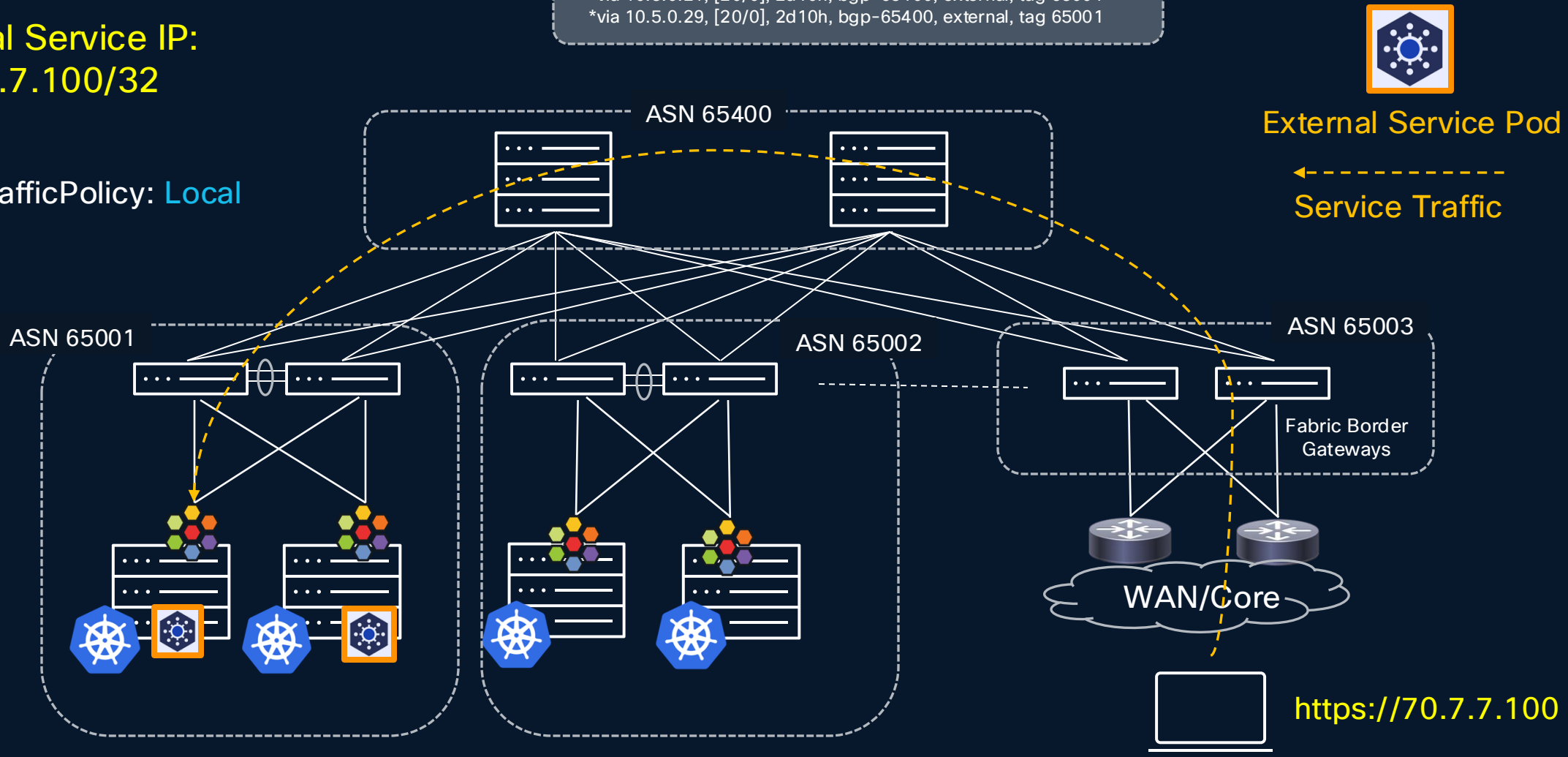
Service Traffic with bestpath command

```
router bgp 65400
  bestpath as-path multipath-relax
```

```
70.7.7.100, ubest/mbest: 2/0
    *via 10.5.0.21, [20/0], 2d10h, bgp-65400, external, tag 65001
    *via 10.5.0.29, [20/0], 2d10h, bgp-65400, external, tag 65001
```

External Service IP:
70.7.7.100/32

externalTrafficPolicy: Local



External Service Pod

Service Traffic

BGP Network on IP Fabric

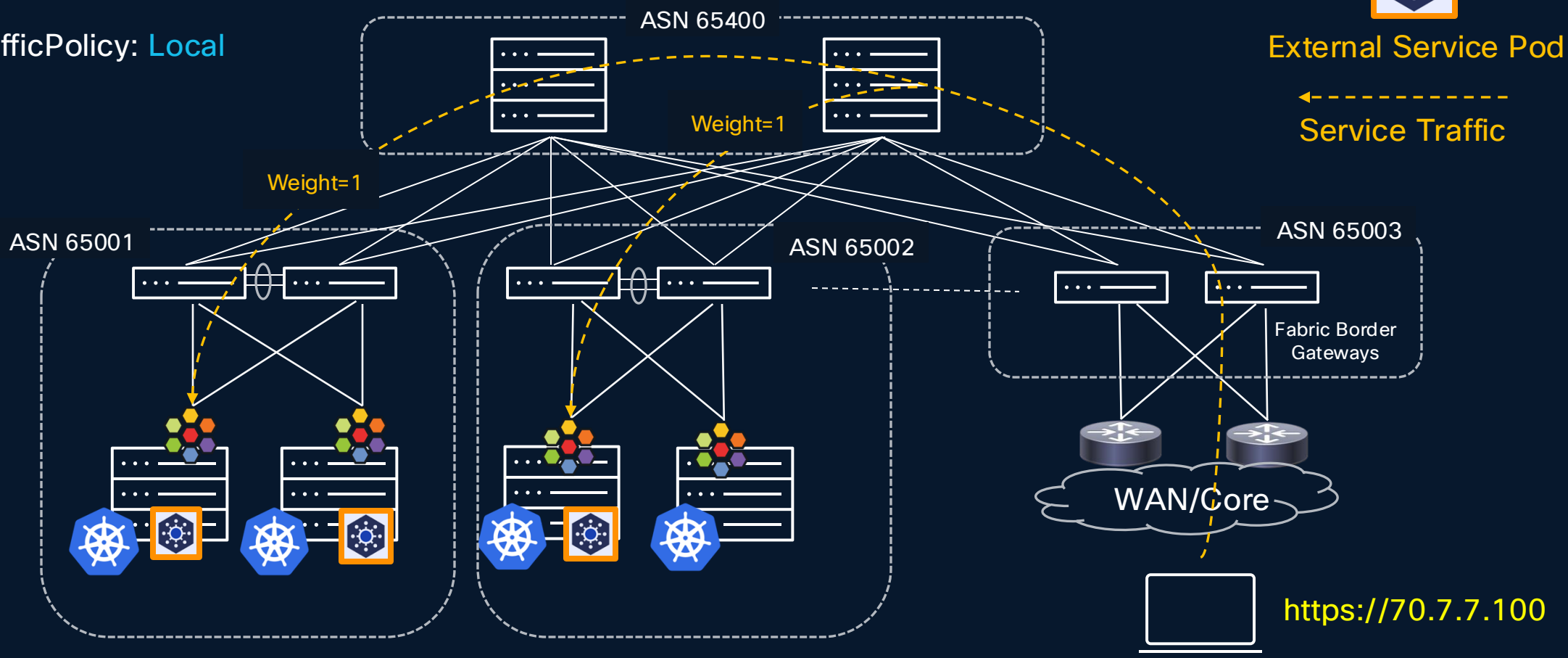
Service Traffic with bestpath command

```
router bgp 65400
  bestpath as-path multipath-relax
```

```
70.7.7.100, ubest/mbest: 4/0
    *via 10.5.0.21, [20/0], 2d10h, bgp-65400, external, tag 65001
    *via 10.5.0.29, [20/0], 2d10h, bgp-65400, external, tag 65001
    *via 10.5.0.37, [20/0], 2d10h, bgp-65400, external, tag 65002
    *via 10.5.0.45, [20/0], 2d10h, bgp-65400, external, tag 65002
```

External Service IP:
70.7.7.100/32

externalTrafficPolicy: Local



External Service Pod

Service Traffic

BGP Network on IP Fabric

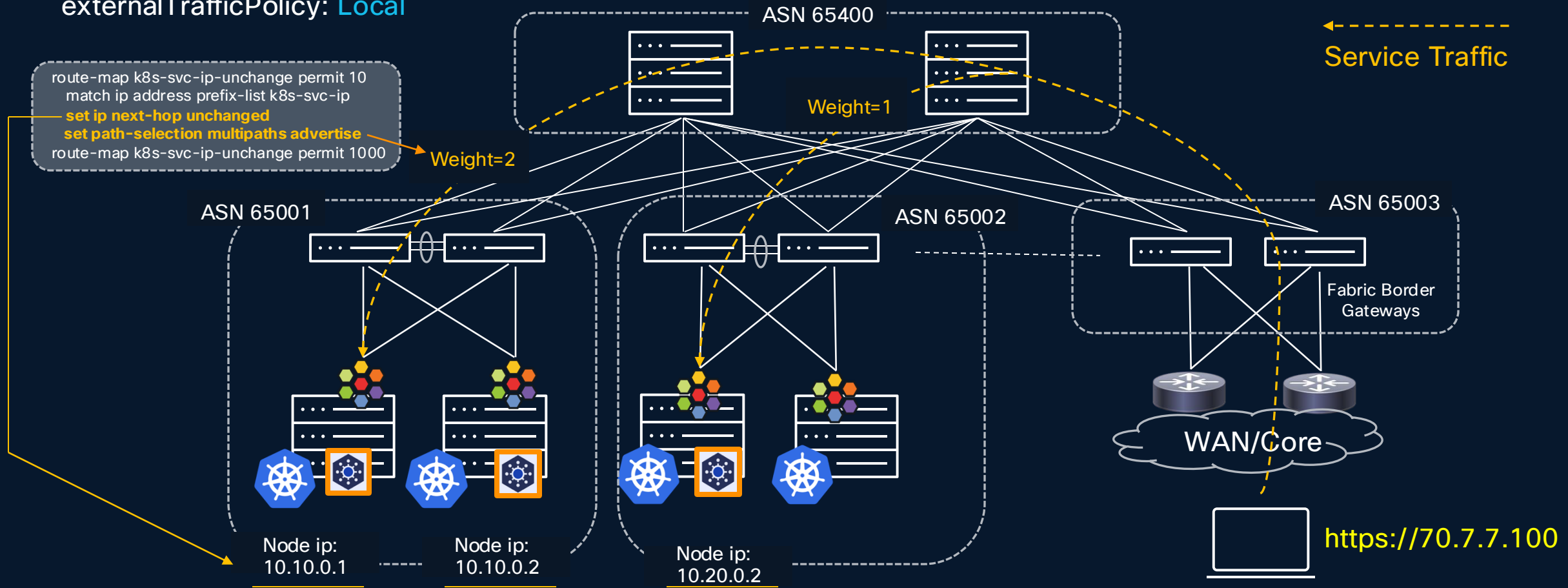
```
router bgp 65400
  bestpath as-path multipath-relax
```

Service Traffic with bestpath command

```
70.7.7.100, ubest/mbest: 3/0
   *via 10.10.0.1, [20/0], 2d10h, bgp-65400, external, tag 65001
   *via 10.10.0.2, [20/0], 2d10h, bgp-65400, external, tag 65001
   *via 10.20.0.2, [20/0], 2d10h, bgp-65400, external, tag 65002
```

External Service IP:
70.7.7.100/32

externalTrafficPolicy: Local



Nexus EVPN/VXLAN Fabric with Cilium SDN leveraging Fabric Overlay

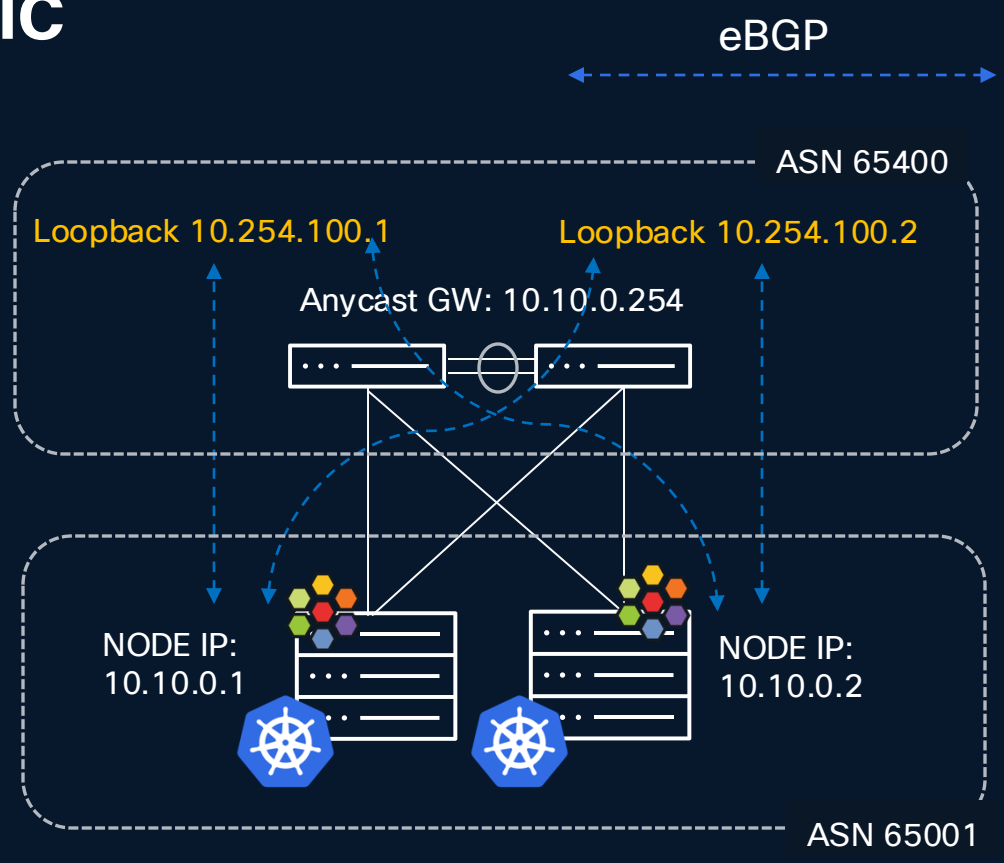
CNI peers with Fabric & Cilium SDN leverages

Network EVPN/VXLAN Overlay

Deploy over EVPN VXLAN Fabric

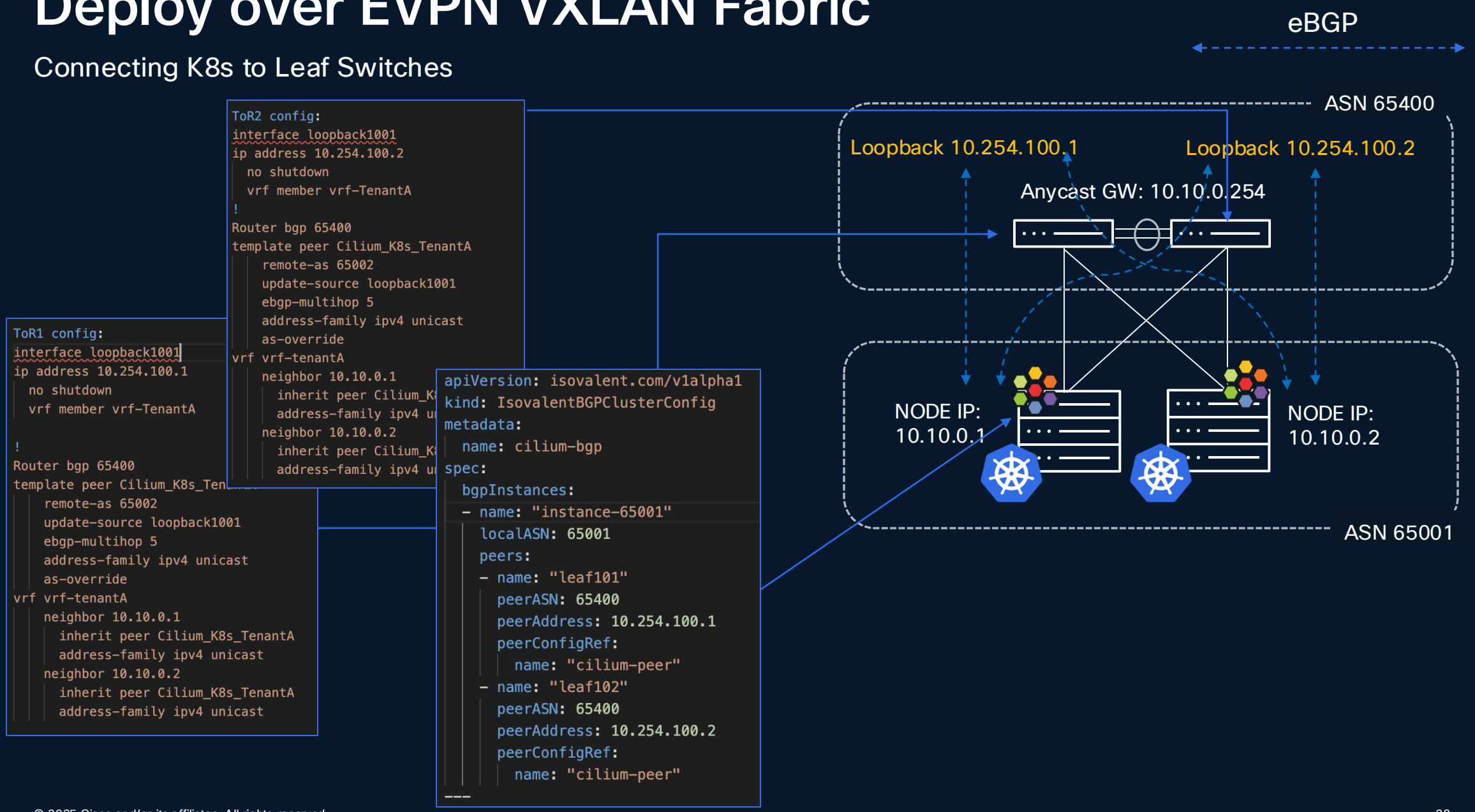
Connecting K8s to Leaf Switches

- K8s nodes connect to Leaf switches using VPC or active-standby
- Nodes are learned as type-2 routes in EVPN
- Peering eBGP between K8s nodes and leaf switches using node IP and localized loopback addresses on each leaf (we don't advertise leaf loopbacks into fabric – only locally significant)
- BGP configuration requires '`disable-peer-as-check`' and '`as-override`' commands
- Loopback Address must be reachable between VPC Peer members



Deploy over EVPN VXLAN Fabric

Connecting K8s to Leaf Switches



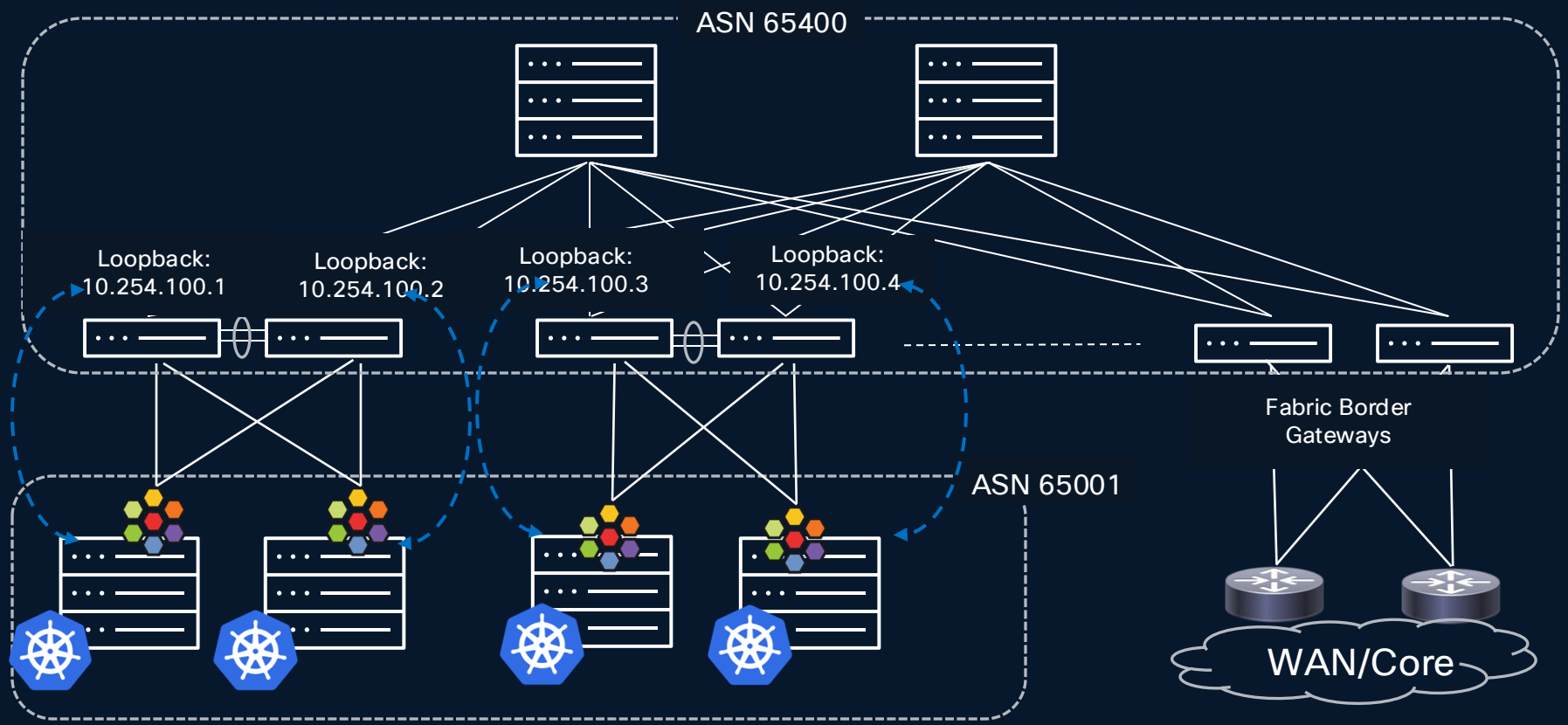

```
ToR1 config:
interface loopback1001
ip address 10.254.100.1
  no shutdown
  vrf member vrf-TenantA

!
Router bgp 65400
template peer Cilium_K8s_Ten
    remote-as 65002
    update-source loopback1001
    ebgp-multihop 5
    address-family ipv4 unicast
    as-override
vrf vrf-tenantA
    neighbor 10.10.0.1
      inherit peer Cilium_K8s_TenantA
      address-family ipv4 unicast
    neighbor 10.10.0.2
      inherit peer Cilium_K8s_TenantA
      address-family ipv4 unicast
```

```
ToR2 config:
interface loopback1001
ip address 10.254.100.2
  no shutdown
  vrf member vrf-TenantA
!
Router bgp 65400
template peer Cilium_K8s_TenantA
    remote-as 65002
    update-source loopback1001
    ebgp-multihop 5
    address-family ipv4 unicast
    as-override
vrf vrf-tenantA
    neighbor 10.10.0.1
      inherit peer Cilium_K
      address-family ipv4 u
    neighbor 10.10.0.2
      inherit peer Cilium_K
      address-family ipv4 u
```

```
apiVersion: isovalent.com/v1alpha1
kind: IsovalentBGPClusterConfig
metadata:
  name: cilium-bgp
spec:
  bgpInstances:
  - name: "instance-65001"
    localASN: 65001
    peers:
    - name: "leaf101"
      peerASN: 65400
      peerAddress: 10.254.100.1
      peerConfigRef:
        name: "cilium-peer"
    - name: "leaf102"
      peerASN: 65400
      peerAddress: 10.254.100.2
      peerConfigRef:
        name: "cilium-peer"
---
```

Deploy Cilium Over EVPN/VXLAN Fabric

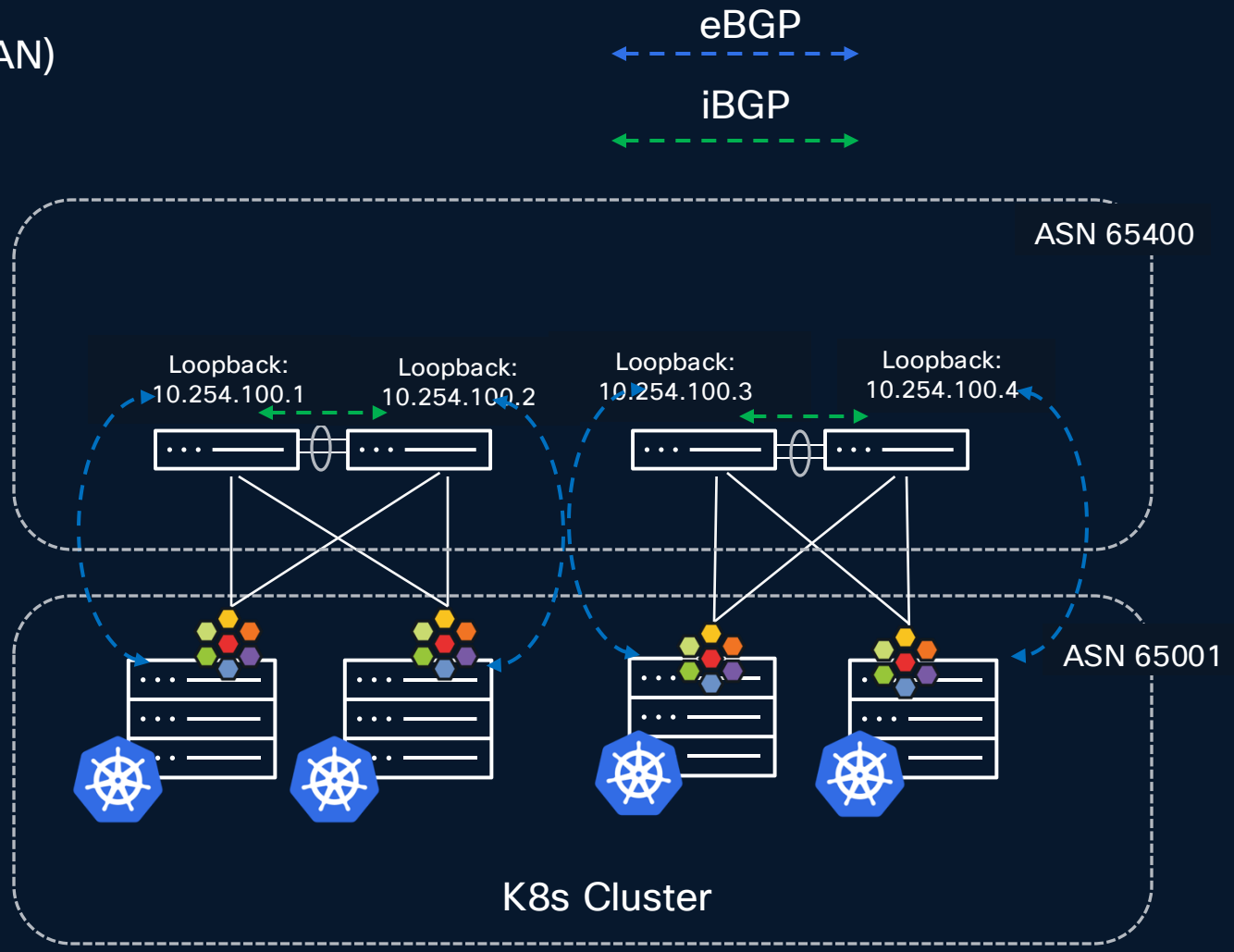
AS-per-Cluster (Recommended Option for EVPN VXLAN)



Deploy Cilium Over EVPN/VXLAN Fabric

AS-per-Cluster (Recommended Option for EVPN VXLAN)

- Using a single AS number per cluster reduces the complexity of bootstrapping K8s
- Loopback addresses are local to the leaf switches
  - No need to advertise Loopbacks to EVPN address family (Locally significant only)
    - Requires iBGP peering between vPC peer switches
- Minimum BGP configuration on Cilium
- Leaf switches BGP configuration requires '`disable-peer-as-check`' and '`as-override`' commands
- Loopback Address must be reachable between VPC Peer members
- Proportional Multipath can overcome unevenly distributed workload
  - '`evpn gateway-ip use-nexthop`' add next hop detail in type-5 advertisement



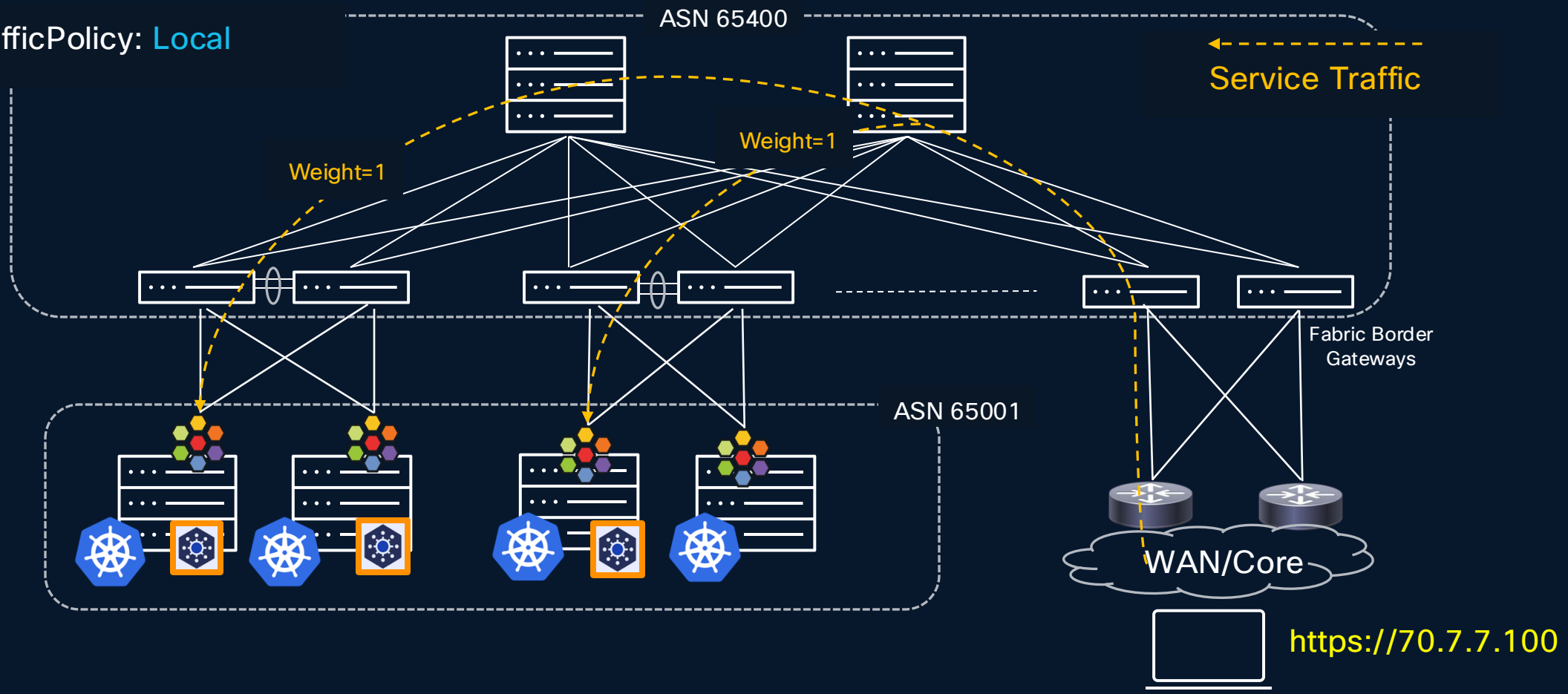
Deploy Cilium Over EVPN/VXLAN Fabric

Service Load Balancing

```
70.7.7.100, ubest/mbest: 2/0
    *via 10.5.0.21%default, [200/0], 01:24:29, bgp-65400, internal, tag 65001,
segid: 50001 tunnelid: 0xa030003 encap: VXLAN
    *via 10.5.0.29%default, [200/0], 01:24:29, bgp-65400, internal, tag 65001,
segid: 50001 tunnelid: 0xa030006 encap: VXLAN
```

External Service IP:
70.7.7.100

externalTrafficPolicy: Local



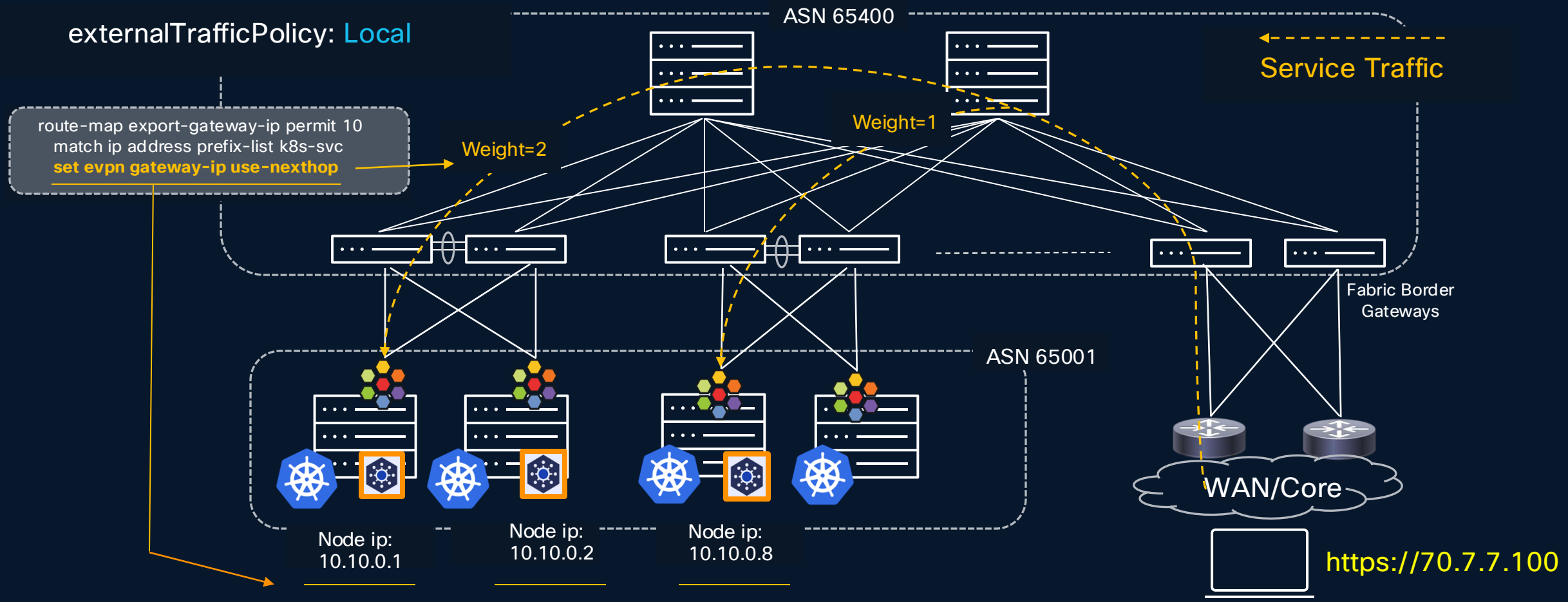
Deploy Cilium Over EVPN/VXLAN Fabric

Service Load Balancing with Proportional Multipath

```
70.7.7.100, ubest/mbest: 3/0
    *via 10.10.0.1%default, [200/0], 01:24:29, bgp-65400, internal, tag 65001
    *via 10.10.0.2%default, [200/0], 01:24:29, bgp-65400, internal, tag 65001
    *via 10.10.0.8%default, [200/0], 01:24:29, bgp-65400, internal, tag 65001
```

External Service IP:
70.7.7.100

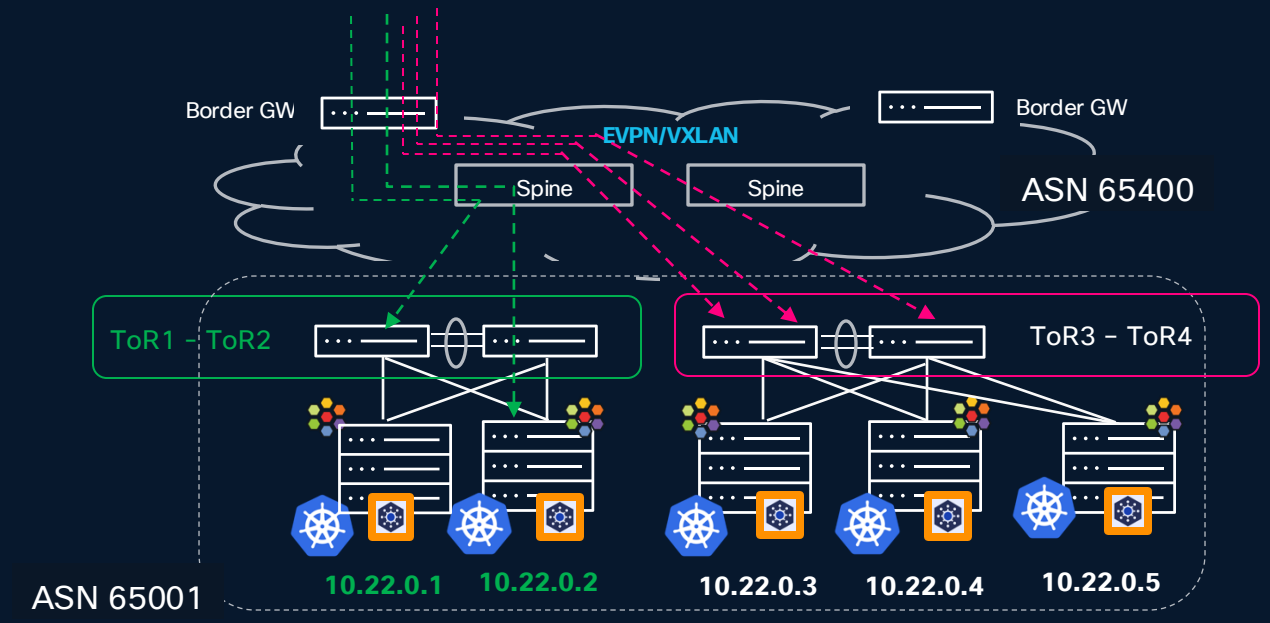
externalTrafficPolicy: Local



EVPN VXLAN

Proportional multipath details

```
172.18.0.0/18, ubest/mbest: 5/0
  *via 10.22.0.1%default, [200/0], 01:24:29, bgp-65400, internal, tag 65001
  *via 10.22.0.2%default, [200/0], 01:24:29, bgp-65400, internal, tag 65001

  *via 10.22.0.3%default, [200/0], 01:24:29, bgp-65400, internal, tag 65001
  *via 10.22.0.4%default, [200/0], 01:24:29, bgp-65400, internal, tag 65001
  *via 10.22.0.5%default, [200/0], 01:24:29, bgp-65400, internal, tag 65001
```




```
router bgp 65400
  !
vrf vrf_Tenant_A
    address family ipv4 unicast
    redistribute direct route-map permit-all
      maximum-paths 8
      maximum-paths ibgp 8
      export-gateway-ip
```

EVPN – Type 5

Prefix	Next-Hop	GW-IP
172.18.0.0/18	ToR1/ToR2	Node 1 (10.22.0.1)
172.18.0.0/18	ToR1/ToR2	Node 2 (10.22.0.2)
172.18.0.0/18	ToR3/ToR4	Node 3 (10.22.0.3)
172.18.0.0/18	ToR3/ToR4	Node 4 (10.22.0.4)

EVPN type5 updated

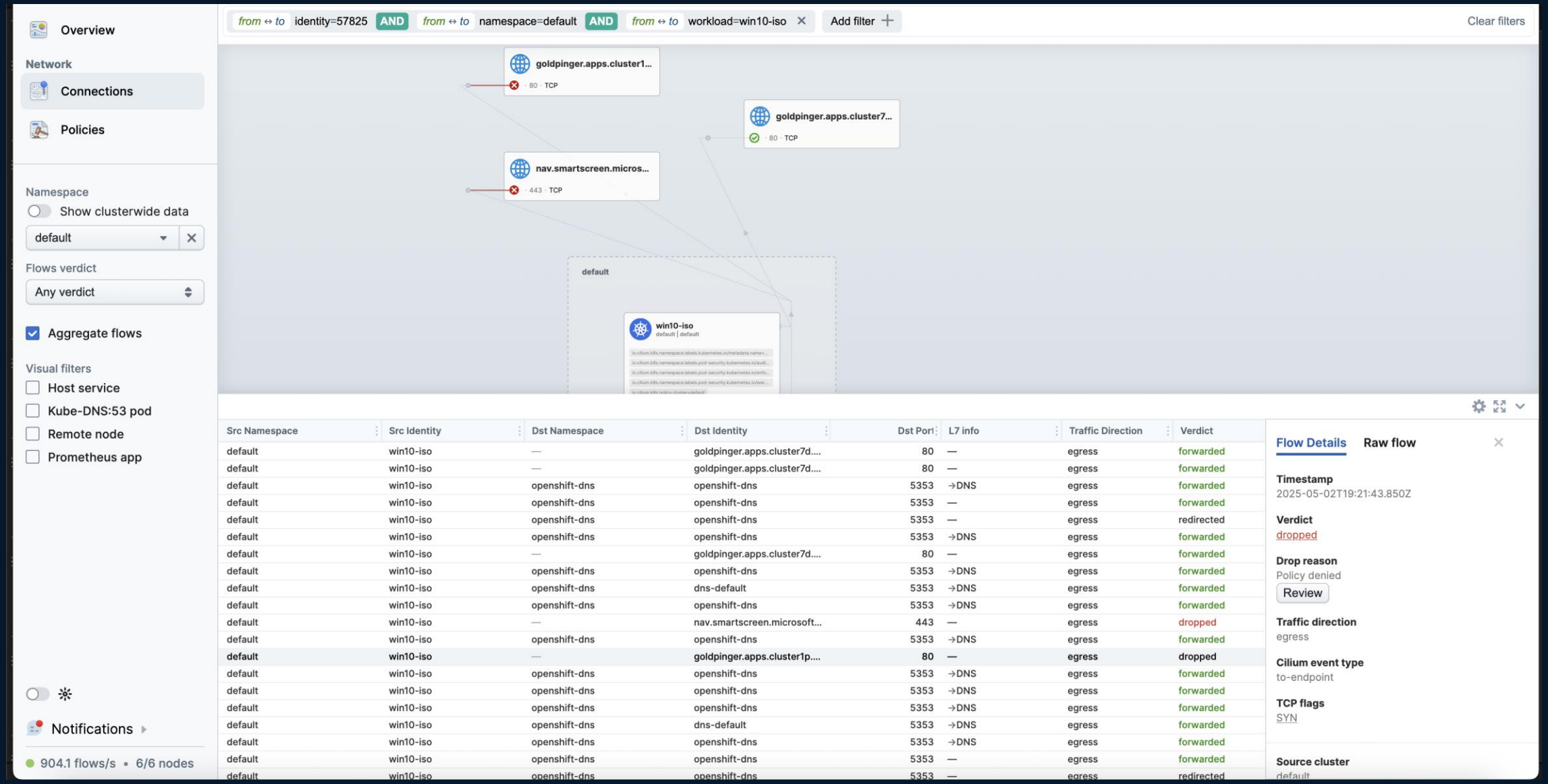
EVPN – Type 5 with GW

Cilium Peer	Next-Hop
Node 1	ToR1/ToR2
Node 2	ToR1/ToR2
Node 3	ToR3/ToR4
Node 4	ToR3/ToR4
Node 5	ToR3/ToR4

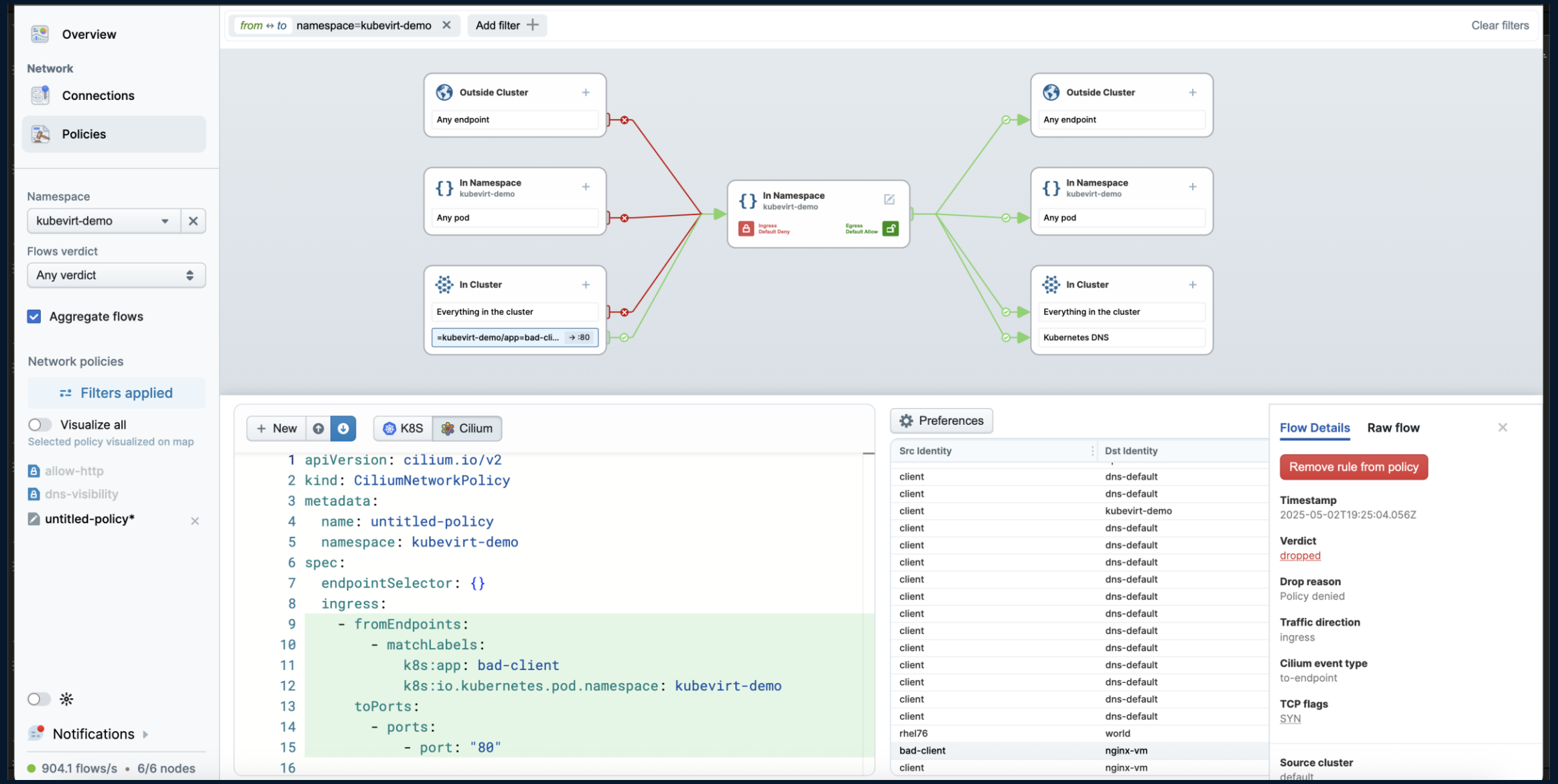
Observability with Isovalent

Workload Monitoring with Hubble

Hubble UI

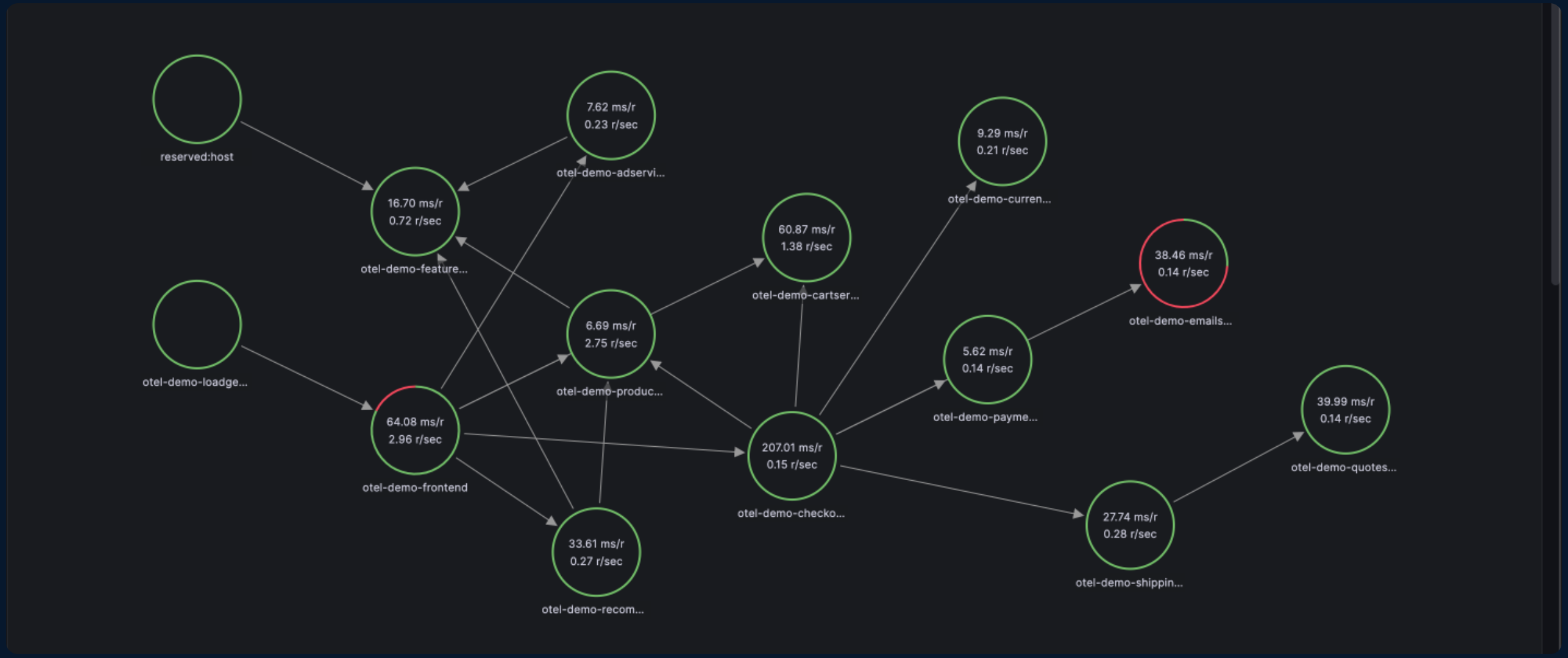


Policy Editor



Tetragon Networking Telemetry

Accurate Request / Response Between Services



Security Observability and Enforcement

Deep Visibility

- System, network, protocols, filesystem, applications, ...

Transparent

- App agonistic
- No changes to applications

Low-Overhead

- Minimal overhead
- Extensive filtering & aggregation

Integrations

- Prometheus, Grafana, SIEM, fluentd, OpenTelemetry, elasticsearch, Splunk

https://isovalent.com/labs/


CISCO