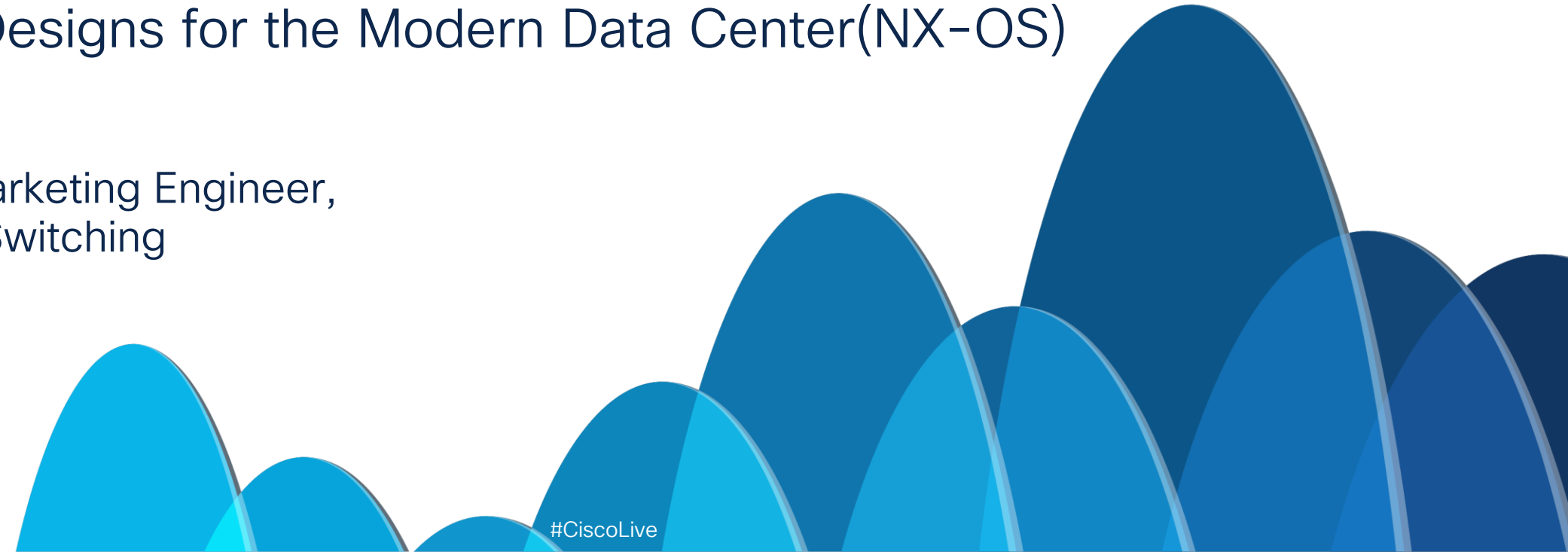
The bridge to possible

# Kubernetes (K8s) Infrastructure Connectivity

## Network Designs for the Modern Data Center(NX-OS)

Shangxin Du
Technical Marketing Engineer,
Datacenter Switching

BRKDCN-2662



cisco Live!

# Cisco Webex App

## Questions?

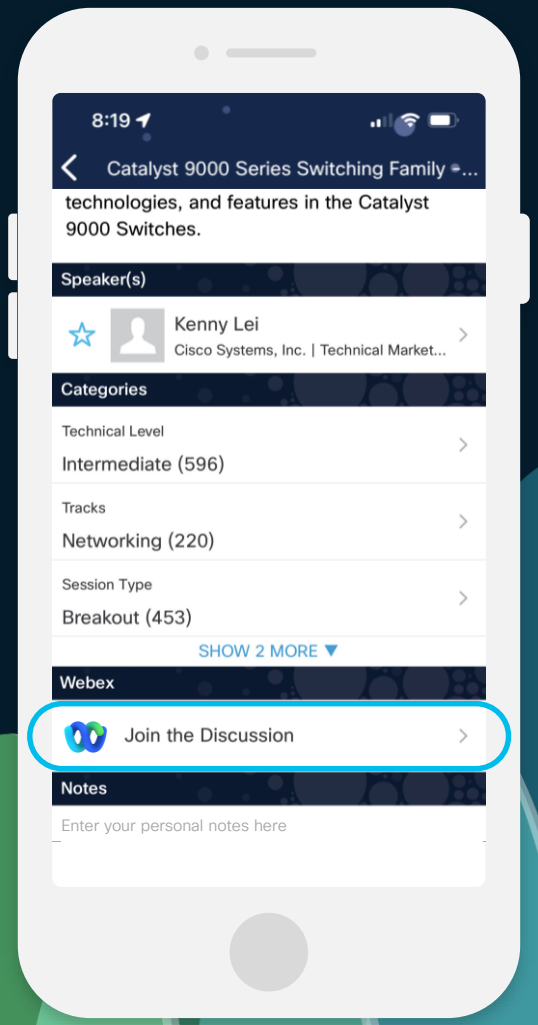
Use Cisco Webex App to chat
with the speaker after the session

## How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

## Webex spaces will be moderated by the speaker until June 7, 2024.

https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-2662



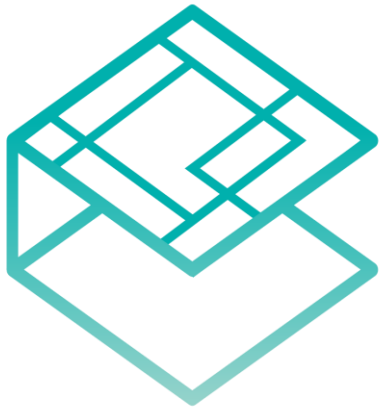
# Agenda

- What is Container Network Interface(CNI) Plugin
- A simple network design – software overlay
- A more scalable design – native routing
  - Design BGP network on IP Fabric
  - Design BGP network on VXLAN EVPN Fabric
- Integration with Nexus Dashboard Fabric Controller(NDFC)

# Agenda

- **What is Container Network Interface(CNI) Plugin**
- A simple network design – software overlay
- A more scalable design – native routing
  - Design BGP network on IP Fabric
  - Design BGP network on VXLAN EVPN Fabric
- Integration with Nexus Dashboard Fabric Controller(NDFC)

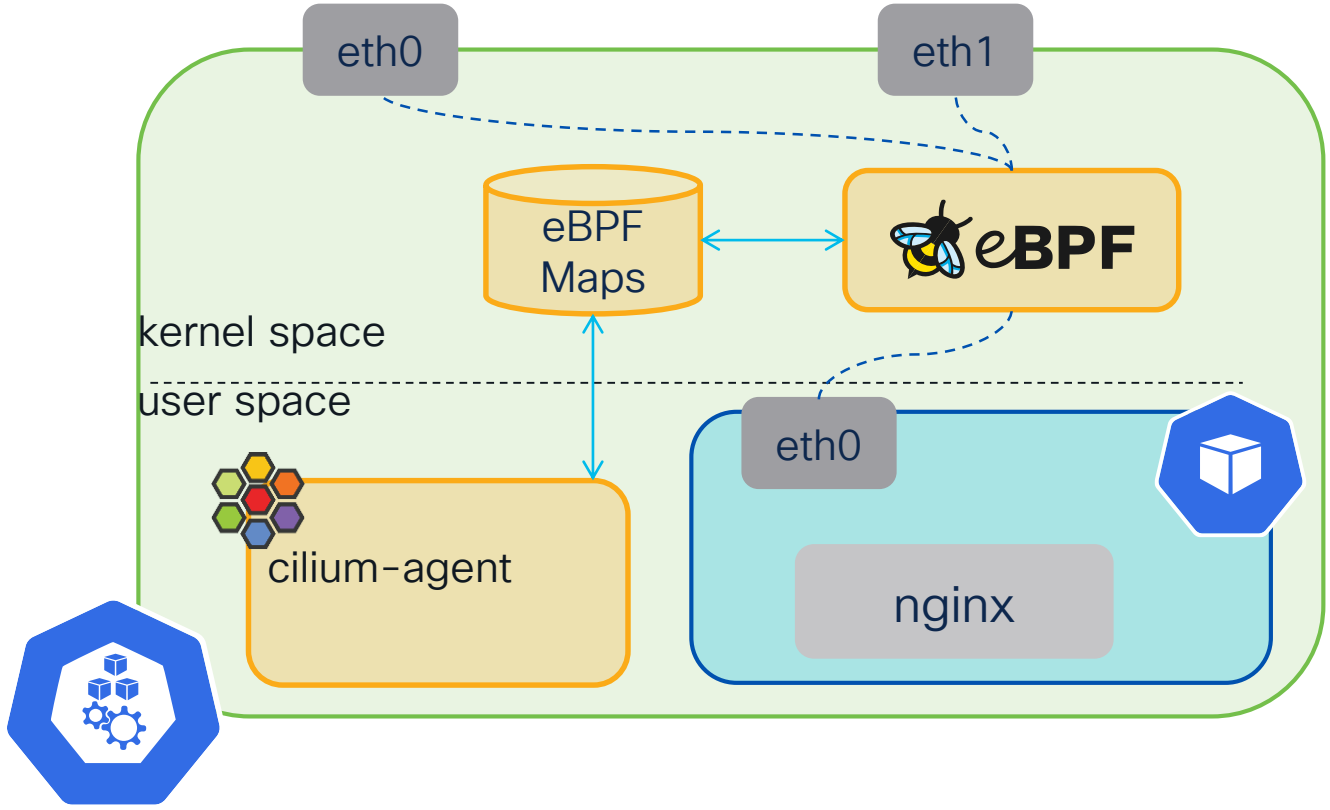
# “Outsourcing the issue” – Container Networking Interface



CNI

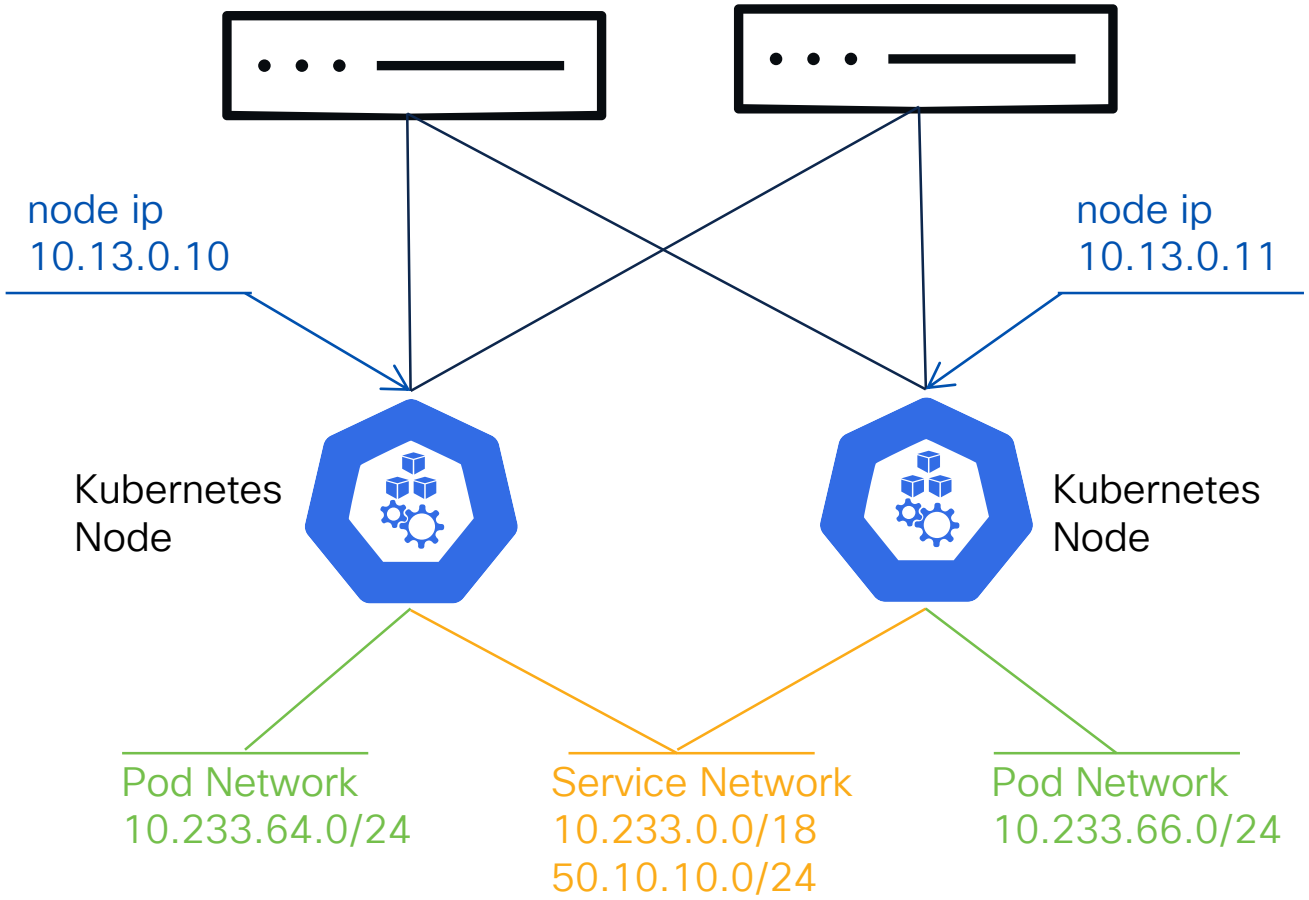
- A generic plugin-based networking solution for application containers on Linux
- The spec defines a container as being a Linux network namespace
- The plugin must connect containers to networks and is responsible for IPAM and DNS configurations.

# Container Network Interface –Cilium



# 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

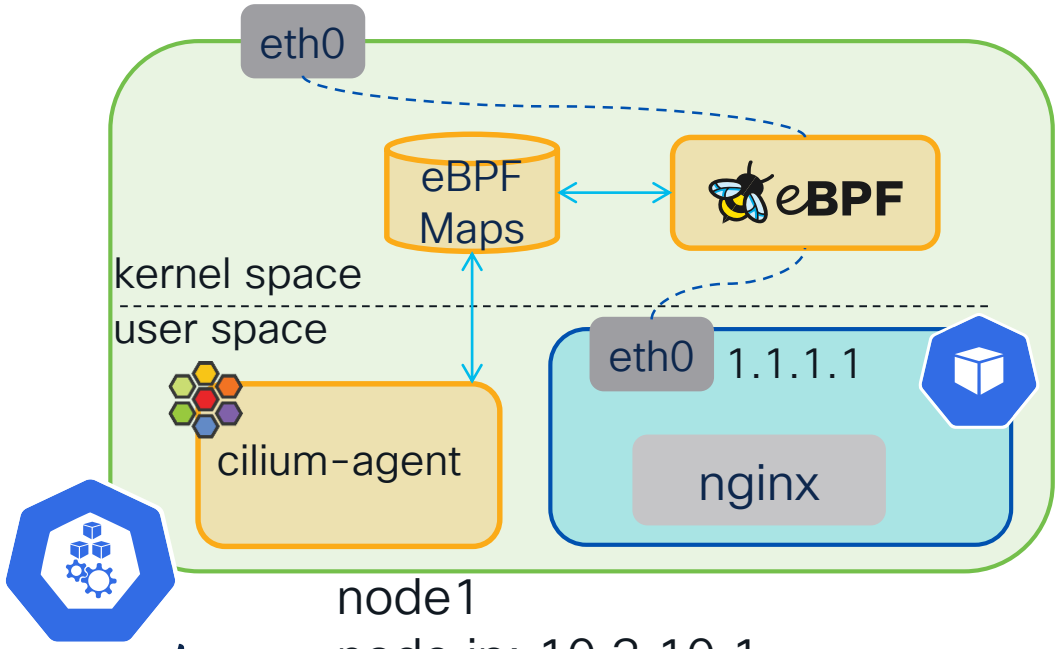
# Kubernetes Service

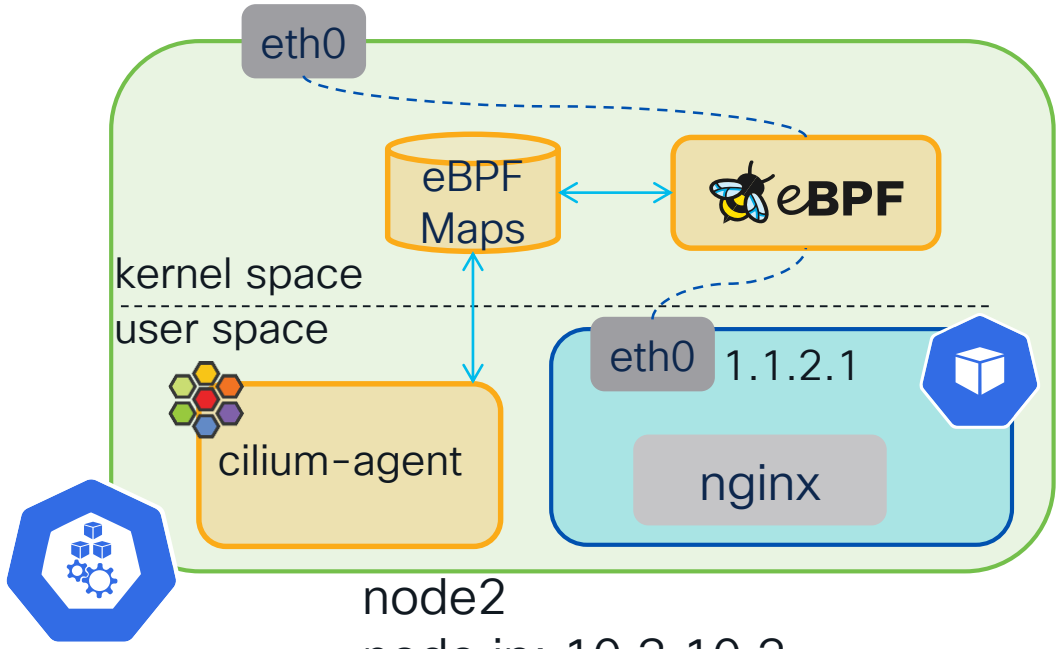
## Internal Service

http://10.233.10.10

type: ClusterIP
service ip: 10.233.10.10

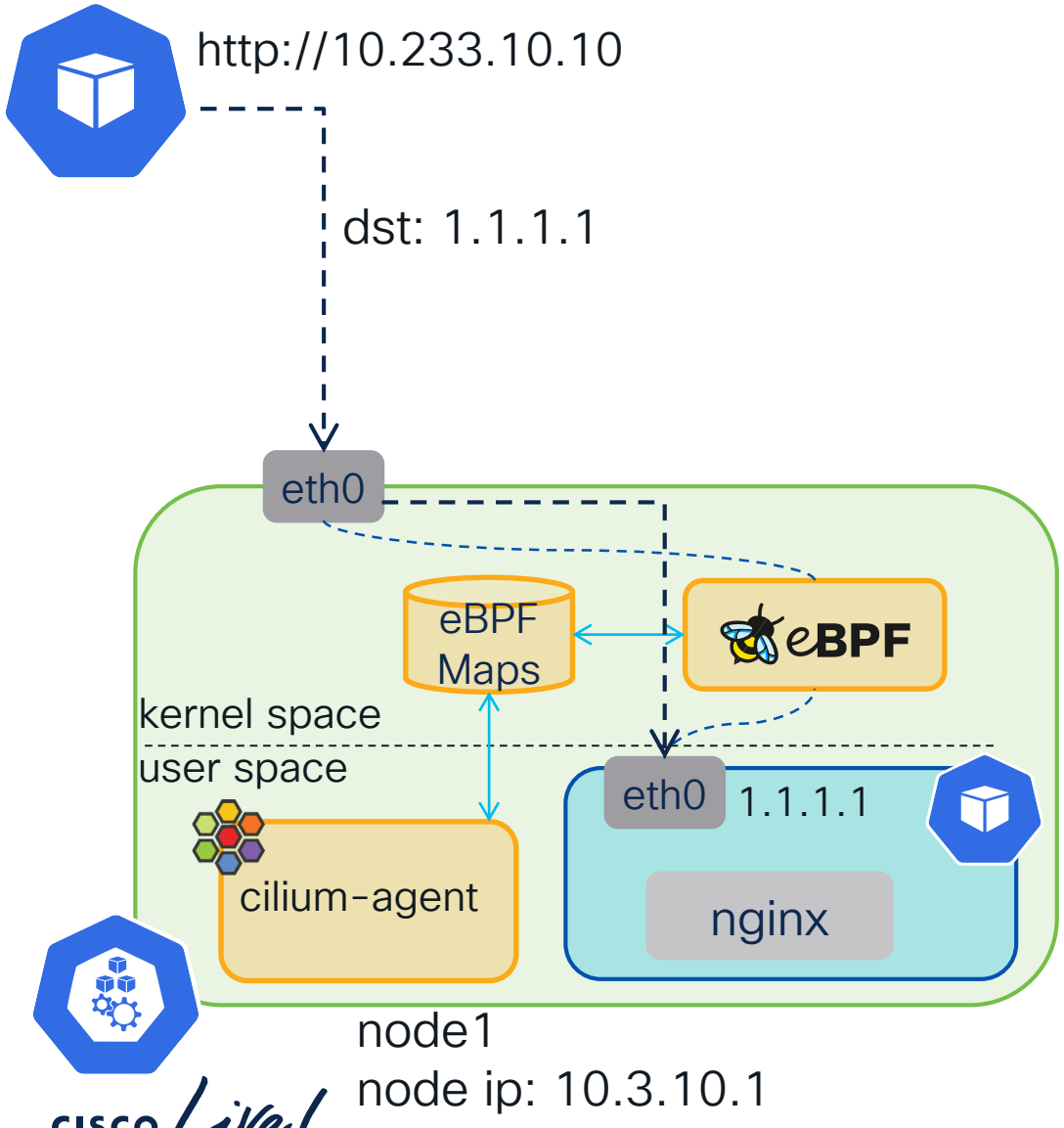
1.1.1.1
1.1.2.1



# Kubernetes Service

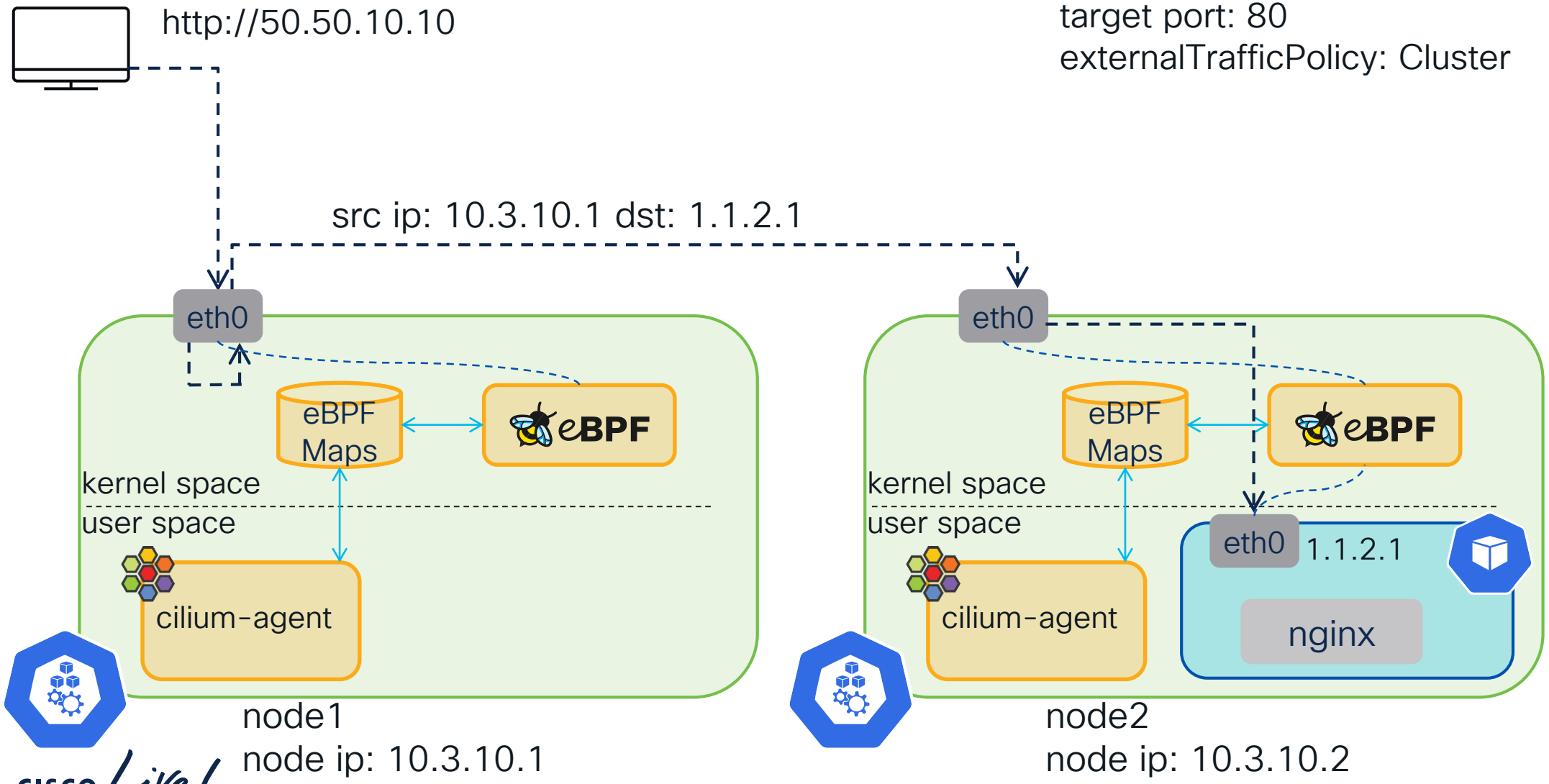
## Internal Service



# Kubernetes Service

## External Service

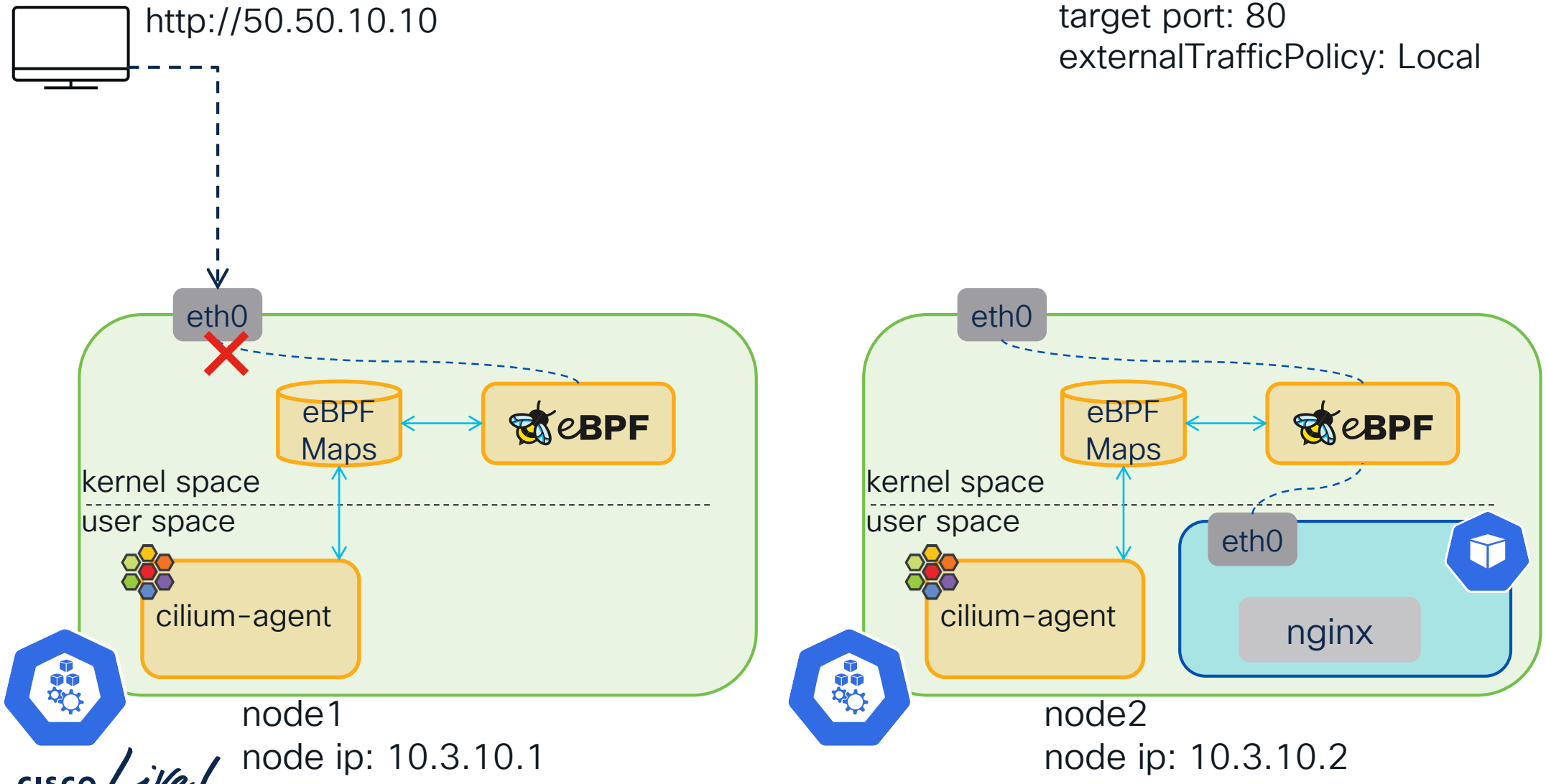
type: NodePort/LoadBalancer
external ip: 50.50.10.10
target port: 80
externalTrafficPolicy: Cluster



# Kubernetes Service

## External Service

type: NodePort/LoadBalancer
external ip: 50.50.10.10
target port: 80
externalTrafficPolicy: Local

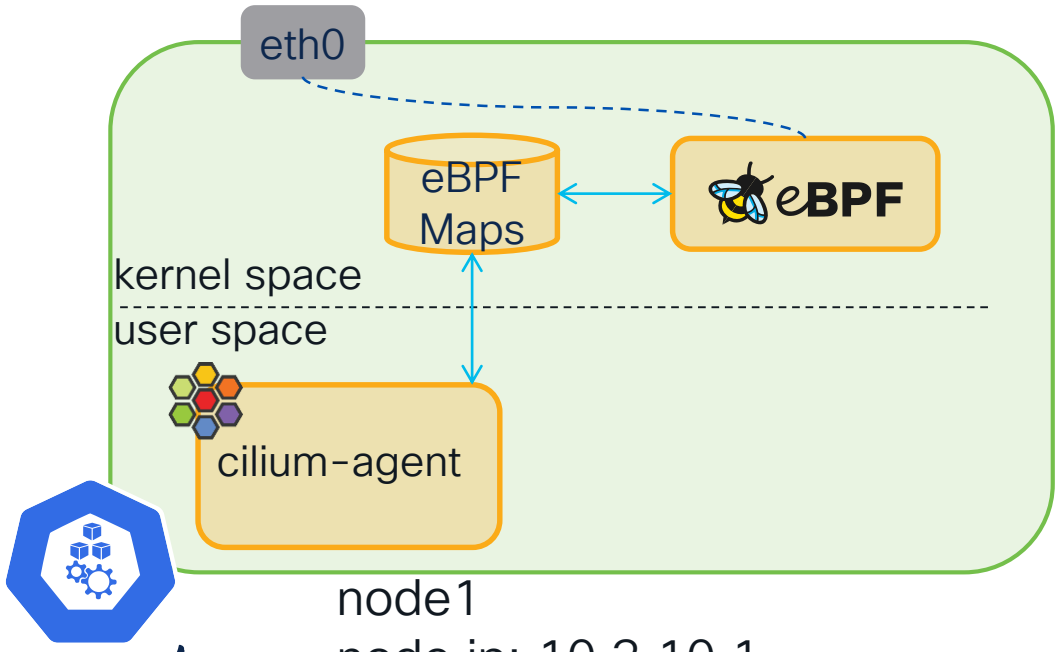


# Kubernetes Service

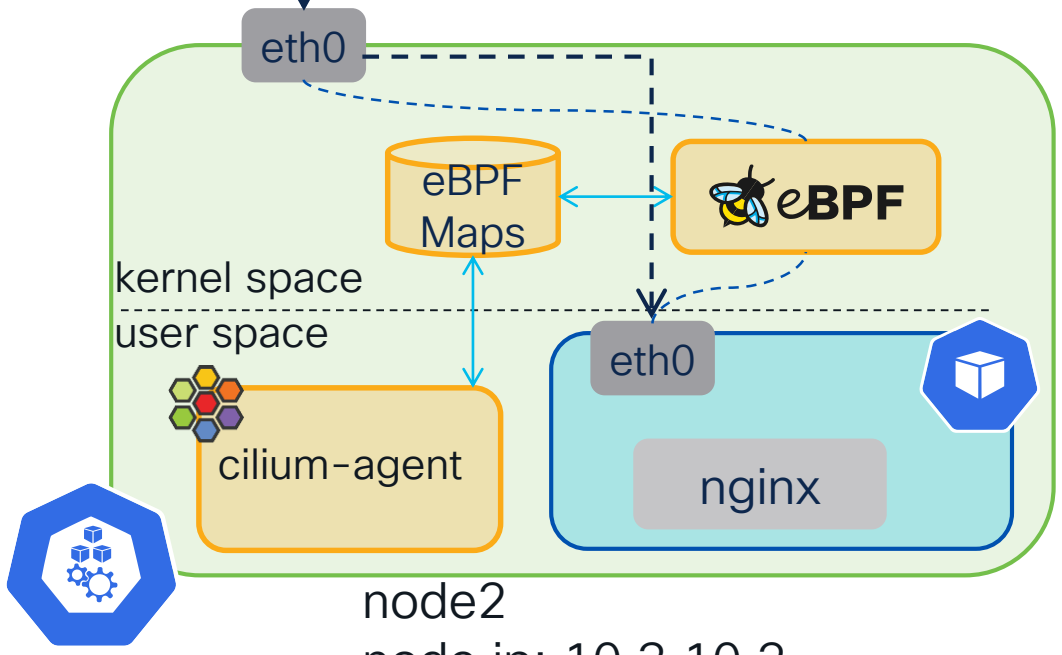
## External Service

type: NodePort/LoadBalancer
external ip: 50.50.10.10
target port: 80
externalTrafficPolicy: Local

http://50.50.10.10




node1
node ip: 10.3.10.1




node2
node ip: 10.3.10.2

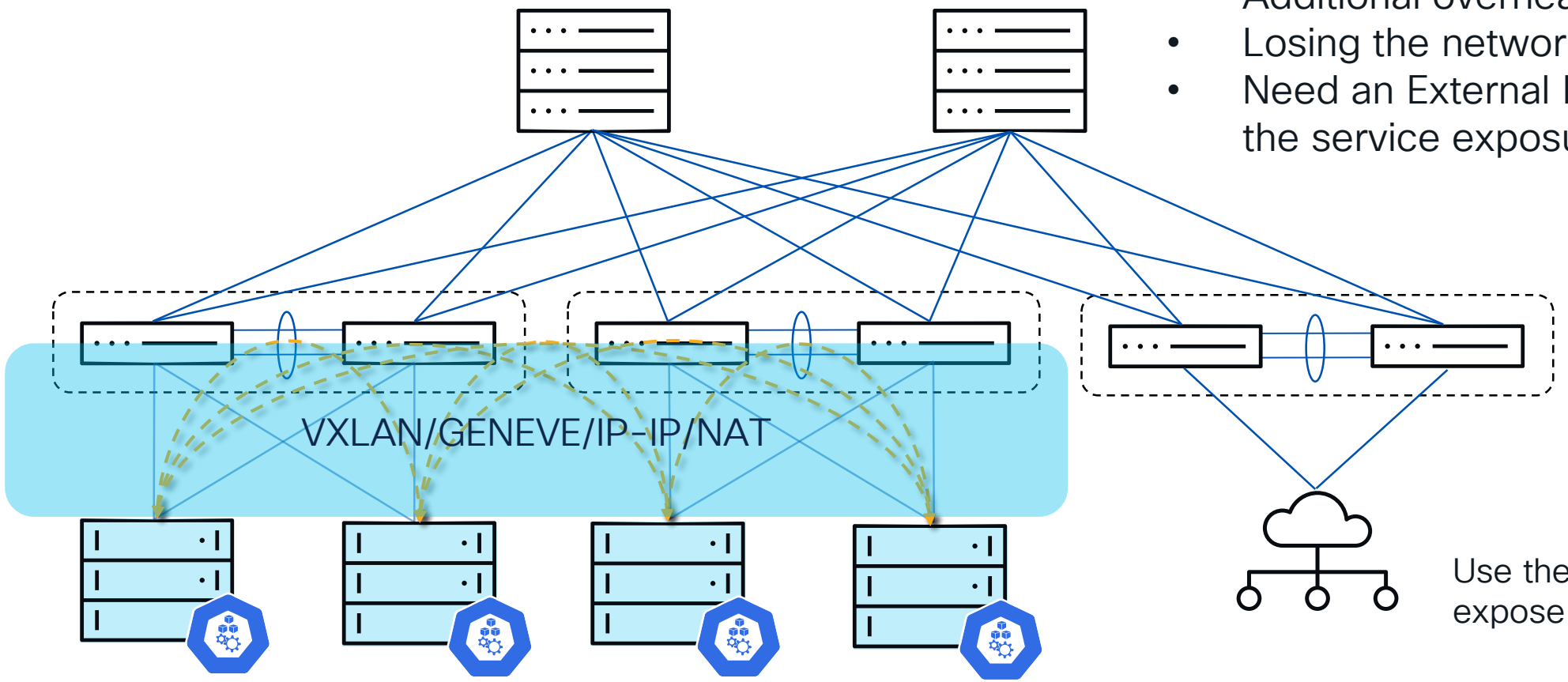
# Agenda

- What is Container Network Interface(CNI) Plugin
- **A simple network design – software overlay**
- A more scalable design – native routing
  - Design BGP network on IP Fabric
  - Design BGP network on VXLAN EVPN Fabric
- Integration with Nexus Dashboard Fabric Controller(NDFC)

# Software overlay

Tunnel or NAT

- Additional overhead
- Losing the network visibility
- Need an External Load Balancer for the service exposure




Use the load balancer to expose the External IPs

# Agenda

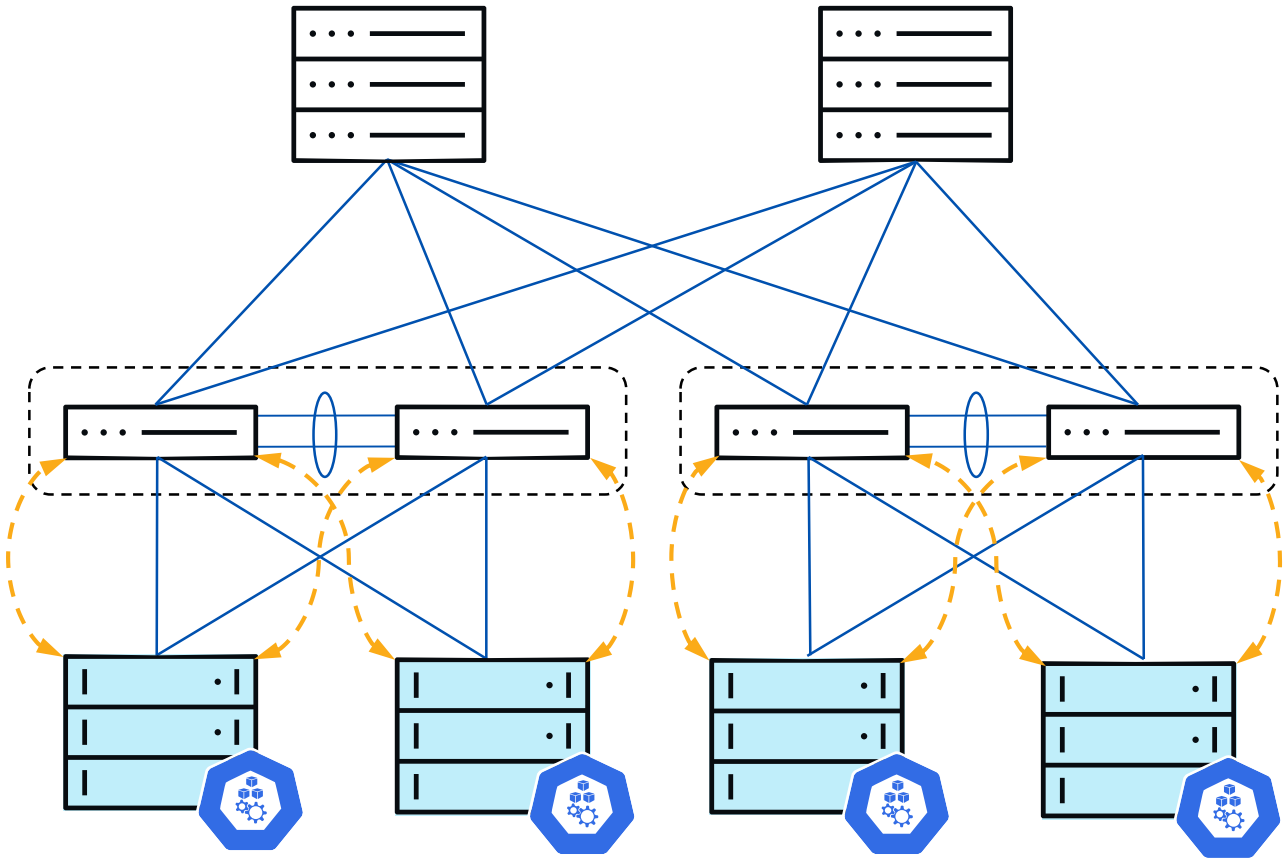
- What is Container Network Interface(CNI) Plugin
- A simple network design – software overlay
- **A more scalable design – native routing**
  - Design BGP network on IP Fabric
  - Design BGP network on VXLAN EVPN Fabric
- Integration with Nexus Dashboard Fabric Controller(NDFC)

# Agenda

- What is Container Network Interface(CNI) Plugin
- A simple network design – software overlay
- **A more scalable design – native routing**
  - **Design BGP network on IP Fabric**
  - Design BGP network on VXLAN EVPN Fabric
- Integration with Nexus Dashboard Fabric Controller(NDFC)

# Native Routing

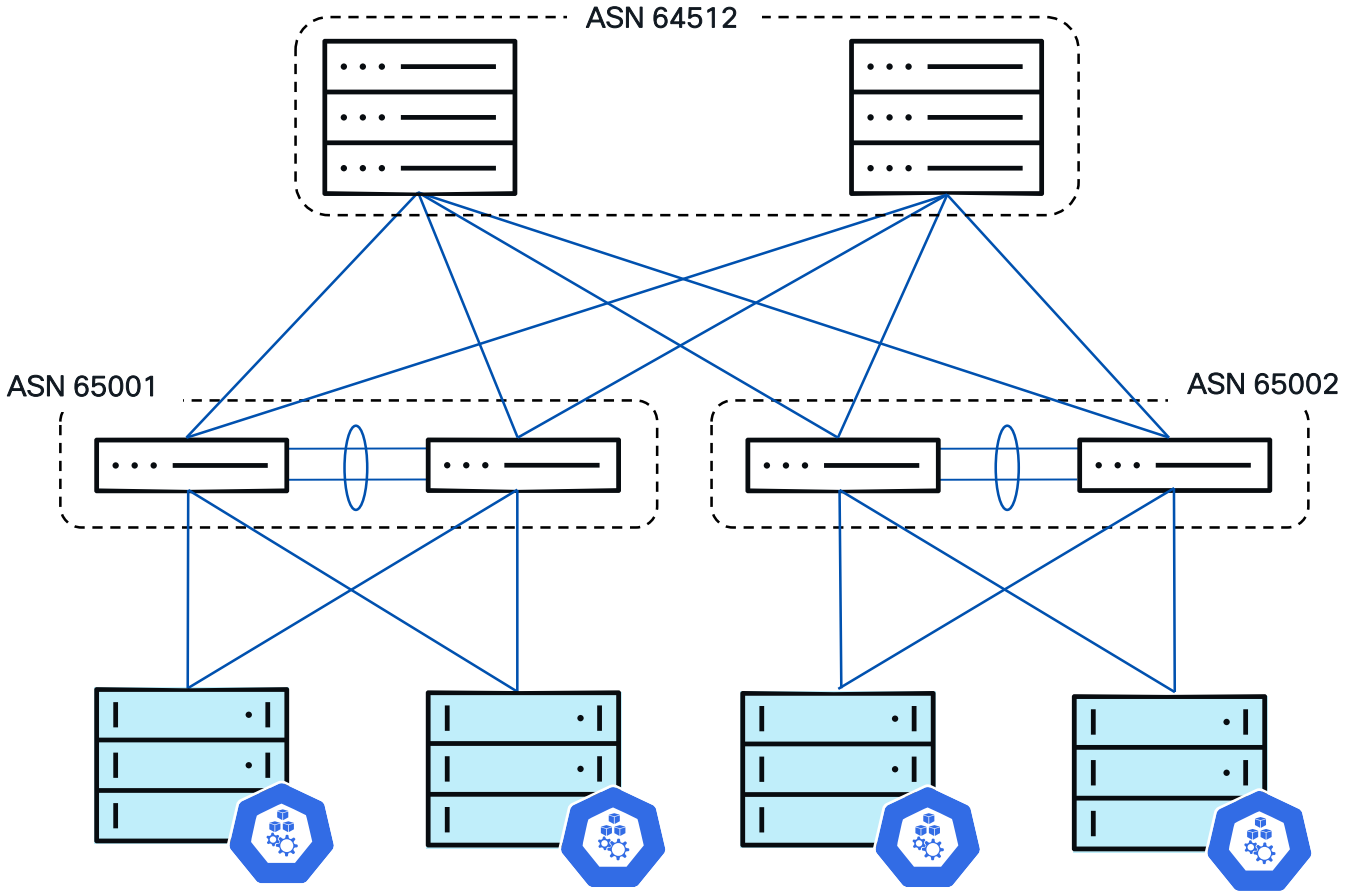
## Peer with switches



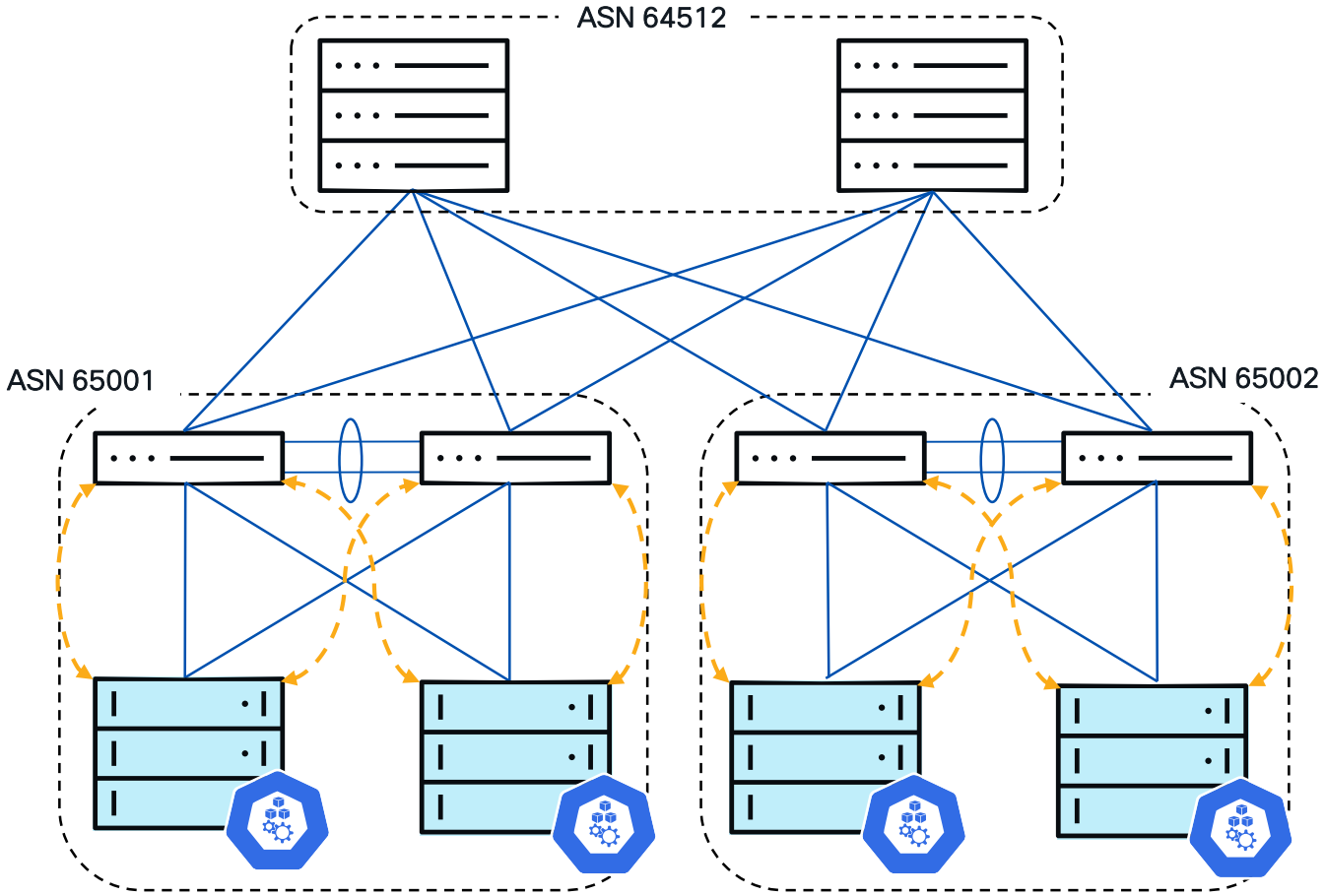
BGP

- Scalable approach, the leaf switches become Route-Reflector Route-Server
- Data is transported with the original headers

# Design BGP network on IP Fabric



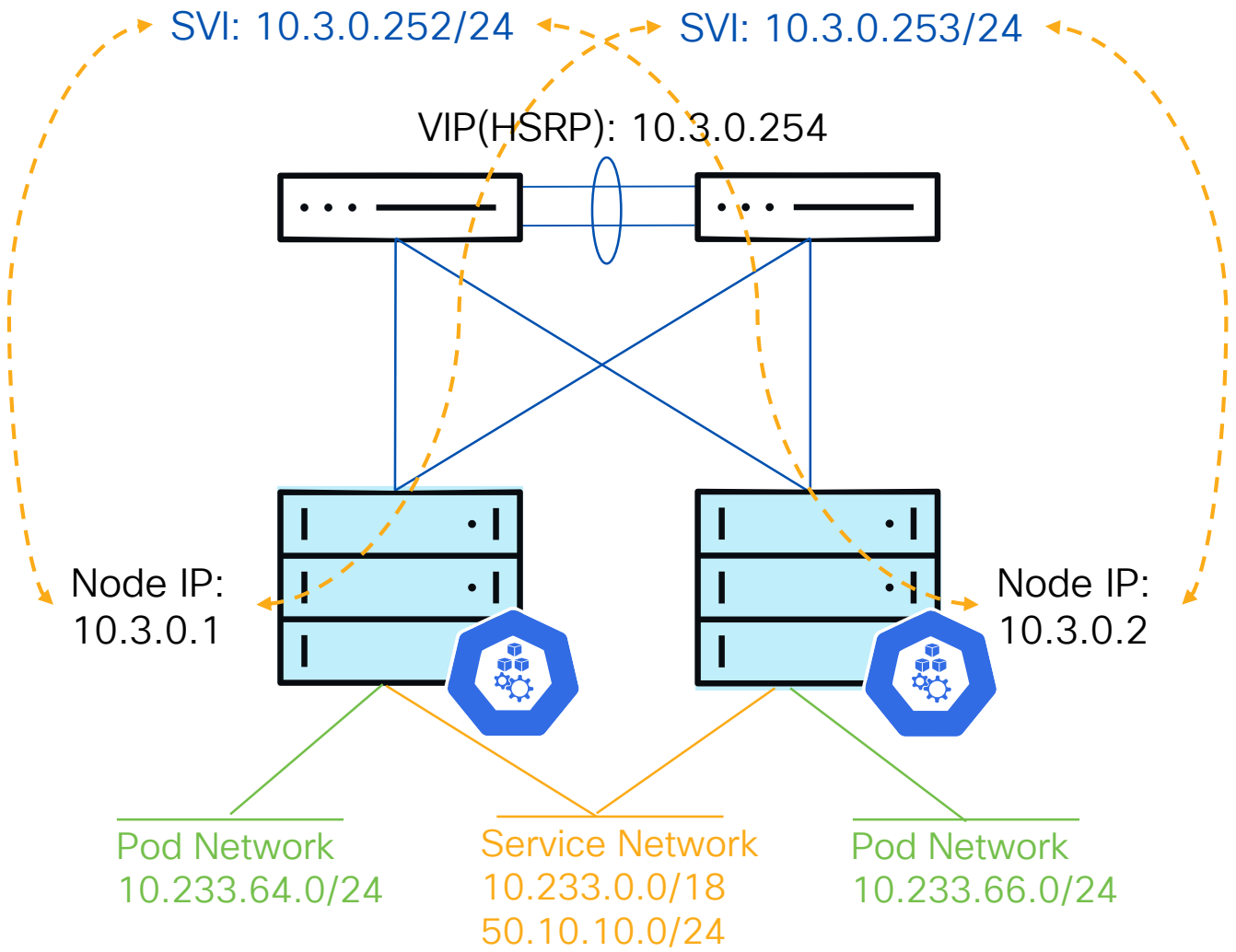
# Design BGP network on IP Fabric




iBGP

- It is usually referred to as AS-per-Rack design.
- Exclusively for IP Fabric(RFC 7938)

# Design BGP network on IP Fabric




iBGP

- HSRP/VRRP is used for gateway redundancy
- Kubernetes nodes peer with the primary IP addresses of the SVIs
- The pod networks and External IPs are advertised into BGP

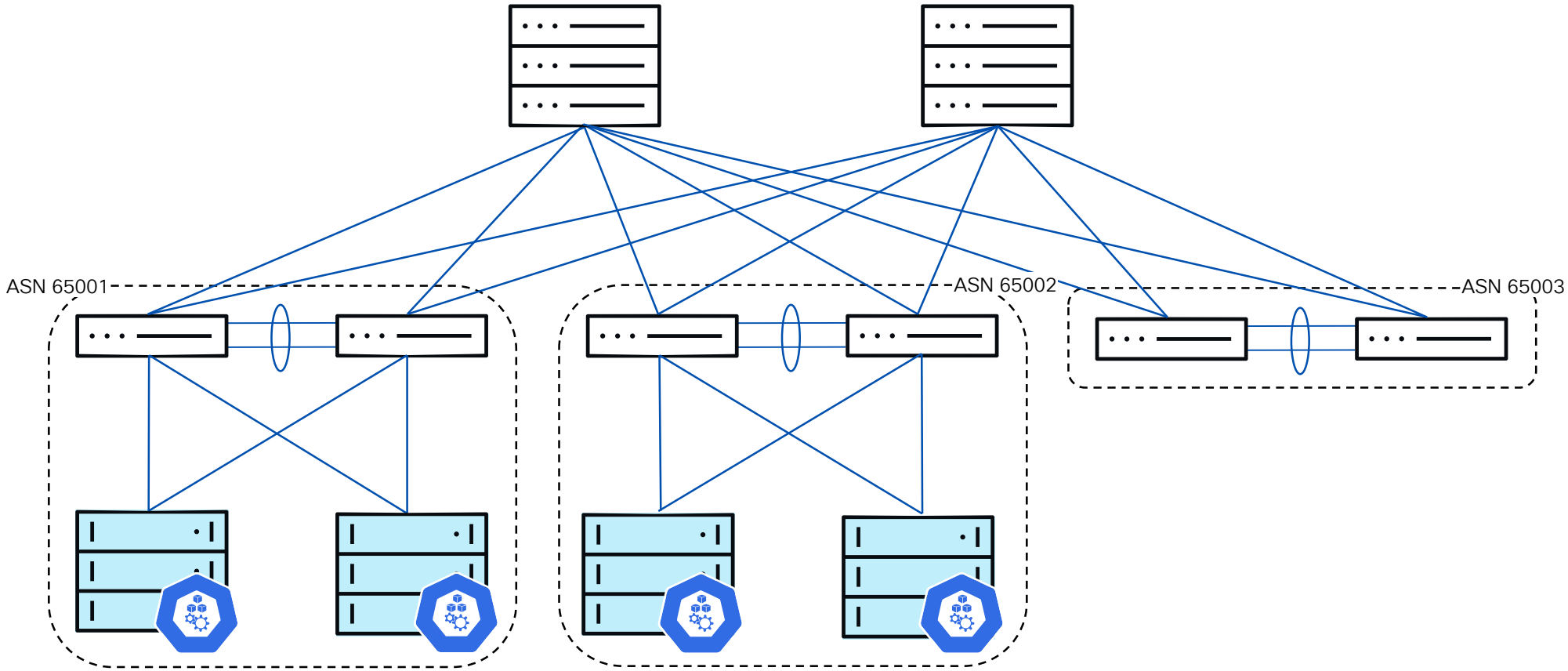
# Design BGP network on IP Fabric

## Service Traffic

```
router bgp 64512
  bestpath as-path multipath-relax
```

external service ip:
50.30.30.35/32

```
50.30.30.35/32, ubest/mbest: 2/0
    *via 10.4.0.37, [20/0], 2d10h, bgp-64512, external, tag 65001
    *via 10.4.0.45, [20/0], 2d10h, bgp-64512, external, tag 65001
```



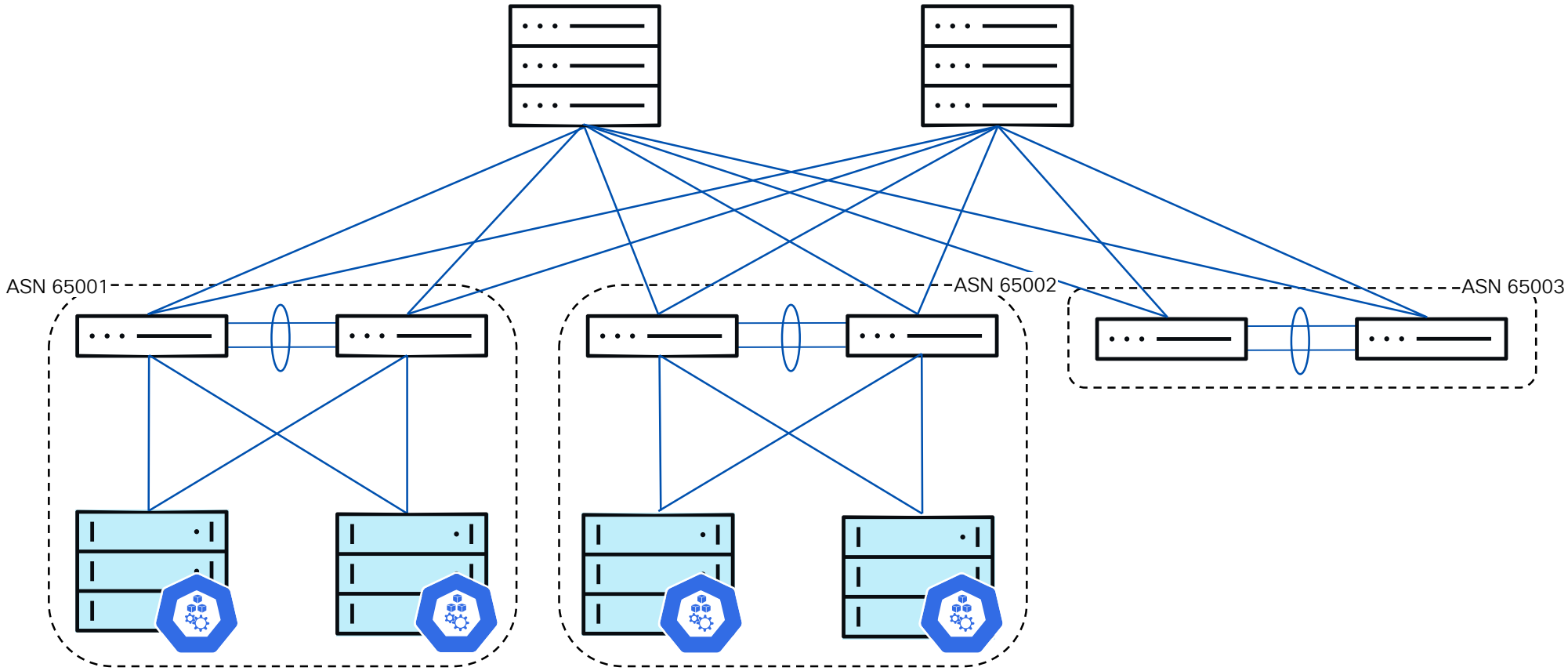
# Design BGP network on IP Fabric

## Service Traffic

```
router bgp 64512
  bestpath as-path multipath-relax
```

external service ip:
50.30.30.35/32

```
50.30.30.35/32, ubest/mbest: 4/0
    *via 10.4.0.21, [20/0], 2d10h, bgp-64512, external, tag 65002
    *via 10.4.0.29, [20/0], 2d10h, bgp-64512, external, tag 65002
    *via 10.4.0.37, [20/0], 2d10h, bgp-64512, external, tag 65001
    *via 10.4.0.45, [20/0], 2d10h, bgp-64512, external, tag 65001
```



# Design BGP network on IP Fabric

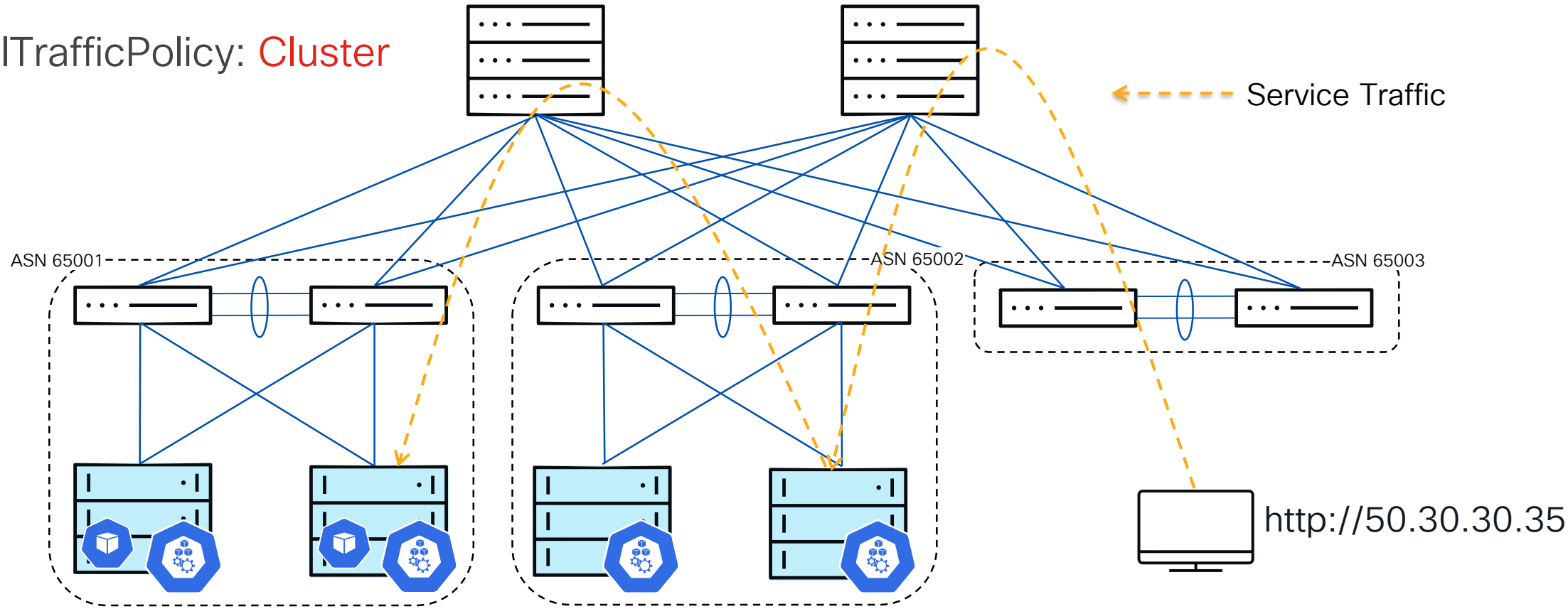
## Service Traffic

```
router bgp 64512
  bestpath as-path multipath-relax
```

external service ip:
50.30.30.35/32

```
50.30.30.35/32, ubest/mbest: 4/0
    *via 10.4.0.21, [20/0], 2d10h, bgp-64512, external, tag 65002
    *via 10.4.0.29, [20/0], 2d10h, bgp-64512, external, tag 65002
    *via 10.4.0.37, [20/0], 2d10h, bgp-64512, external, tag 65001
    *via 10.4.0.45, [20/0], 2d10h, bgp-64512, external, tag 65001
```

externalTrafficPolicy: Cluster



# Design BGP network on IP Fabric

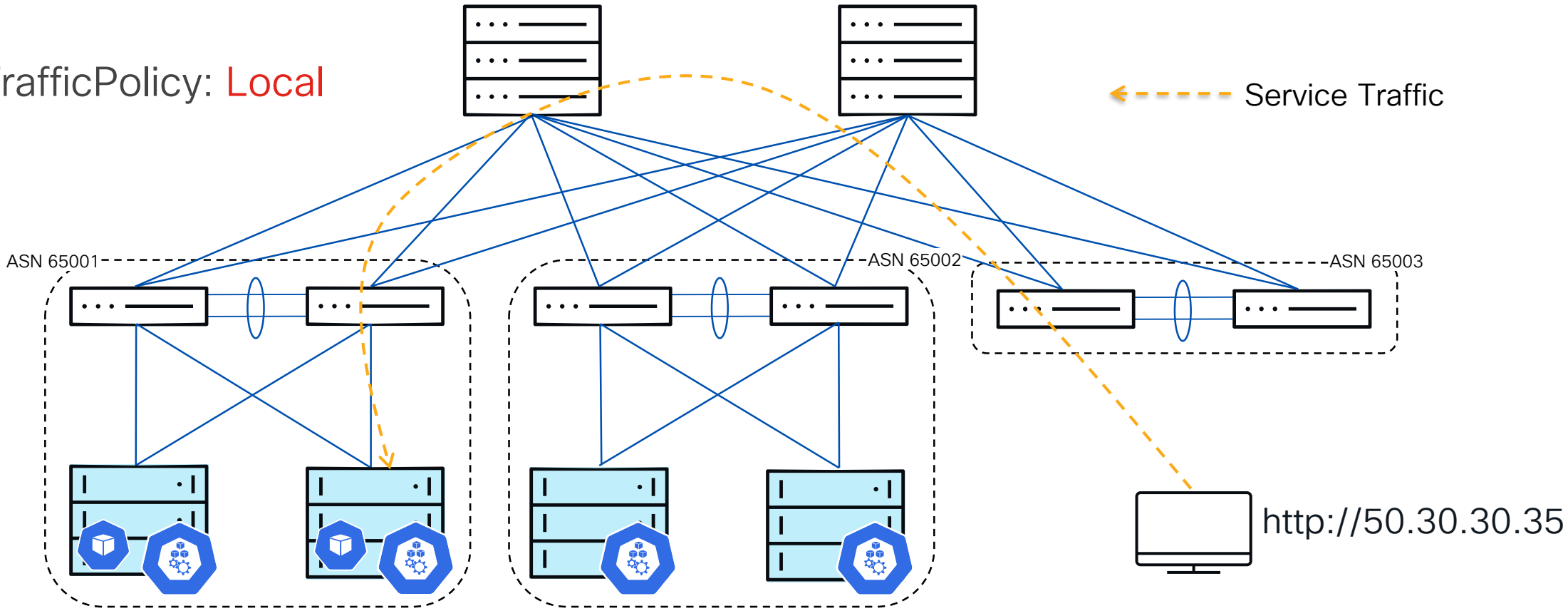
## Service Traffic

```
router bgp 64512
  bestpath as-path multipath-relax
```

external service ip:
50.30.30.35/32

```
50.30.30.35/32, ubest/mbest: 2/0
    *via 10.4.0.37, [20/0], 2d10h, bgp-64512, external, tag 65001
    *via 10.4.0.45, [20/0], 2d10h, bgp-64512, external, tag 65001
```

externalTrafficPolicy: Local



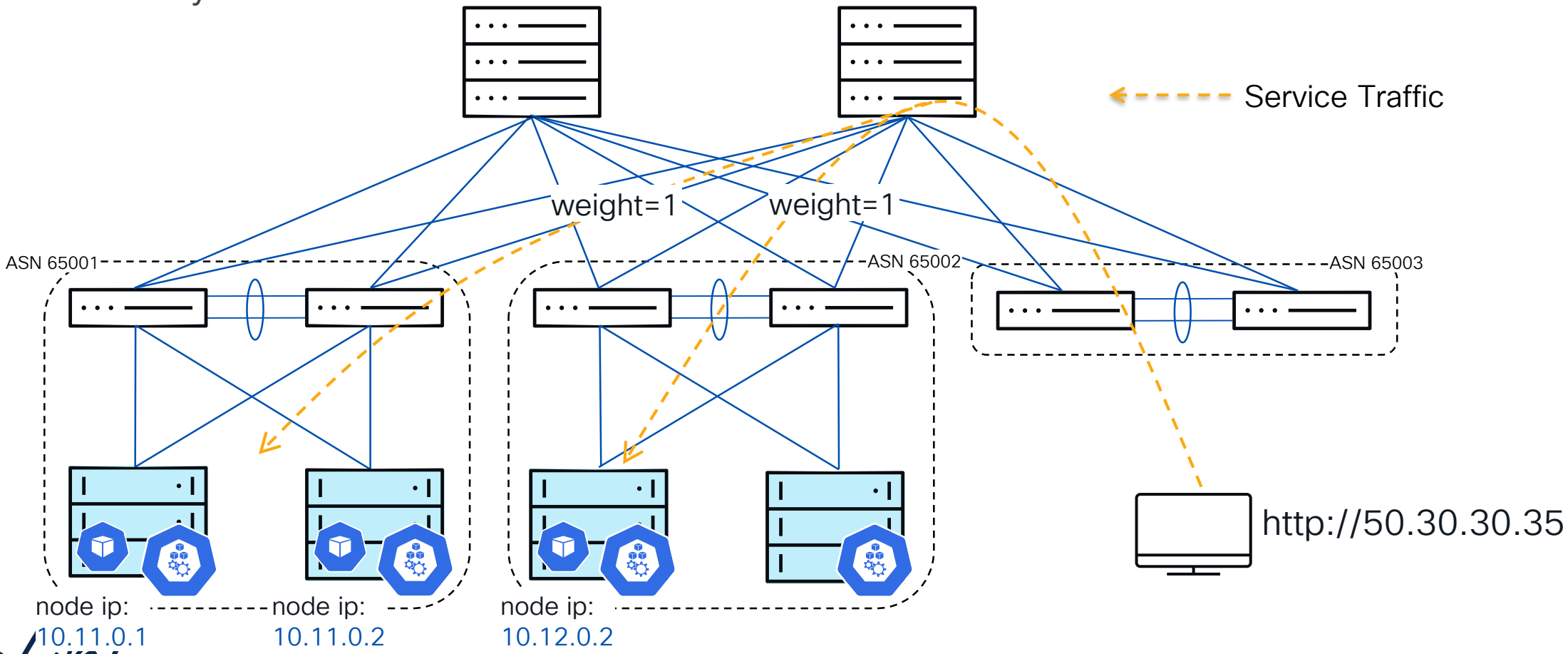
# Design BGP network on IP Fabric

## Service Traffic

```
router bgp 64512
  bestpath as-path multipath-relax
```

external service ip:
50.30.30.35/32
externalTrafficPolicy: Local

```
50.30.30.35/32, ubest/mbest: 4/0
    *via 10.4.0.21, [20/0], 2d10h, bgp-64512, external, tag 65002
    *via 10.4.0.29, [20/0], 2d10h, bgp-64512, external, tag 65002
    *via 10.4.0.37, [20/0], 2d10h, bgp-64512, external, tag 65001
    *via 10.4.0.45, [20/0], 2d10h, bgp-64512, external, tag 65001
```



# Design BGP network on IP Fabric

## Service Traffic

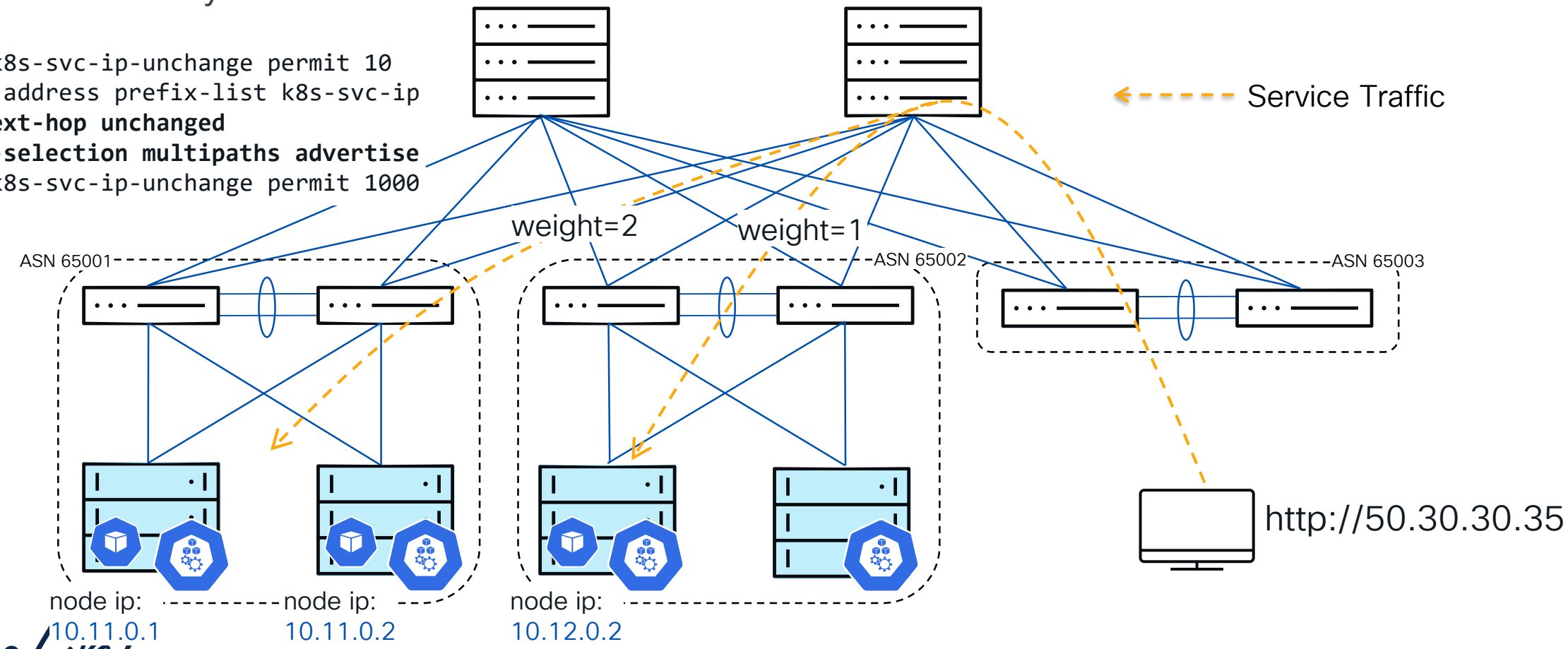
```
router bgp 64512
  bestpath as-path multipath-relax
```

external service ip:
50.30.30.35/32

externalTrafficPolicy: Local

```
50.30.30.35/32, ubest/mbest: 4/0
    *via 10.11.0.2, [20/0], 05:34:09, bgp-64512, external, tag 65001
    *via 10.11.0.1, [20/0], 07:24:07, bgp-64512, external, tag 65001
    *via 10.12.0.2, [20/0], 07:24:07, bgp-64512, external, tag 65002
```

```
route-map k8s-svc-ip-unchange permit 10
  match ip address prefix-list k8s-svc-ip
  set ip next-hop unchanged
  set path-selection multipaths advertise
route-map k8s-svc-ip-unchange permit 1000
```



# Exposing Services

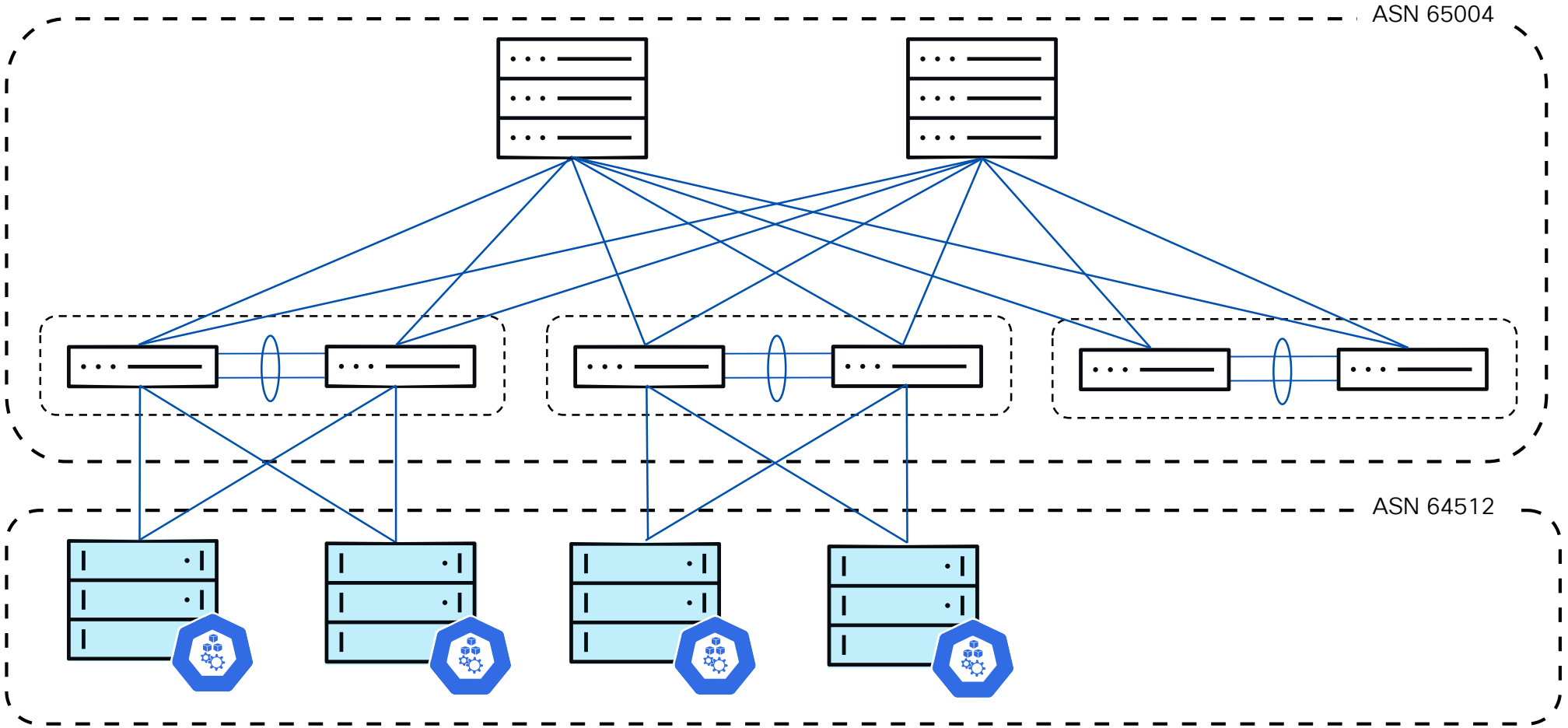
## A note on “externalTrafficPolicy”

- 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

# Agenda

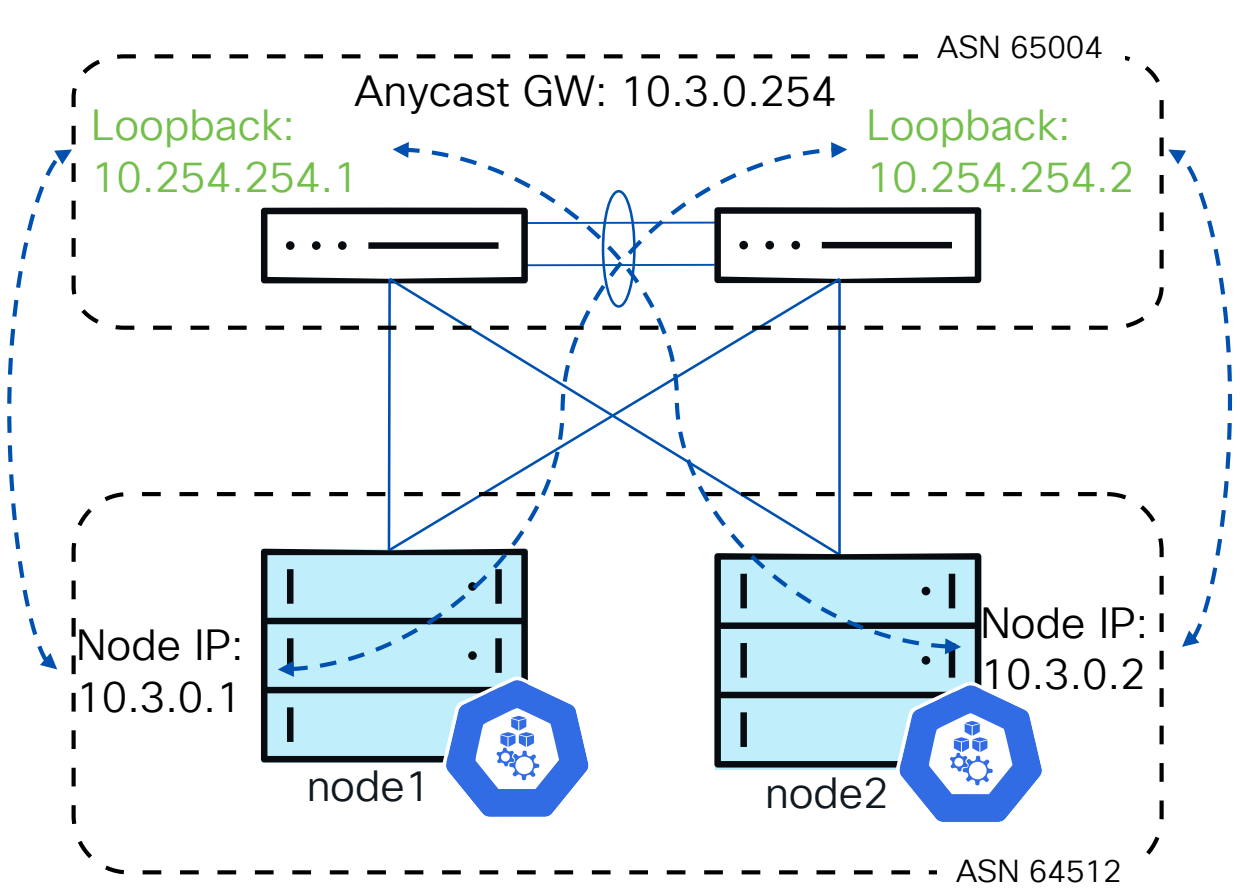
- What is Container Network Interface(CNI) Plugin
- A simple network design – software overlay
- **A more scalable design – native routing**
  - Design BGP network on IP Fabric
  - **Design BGP network on VXLAN EVPN Fabric**
- Integration with Nexus Dashboard Fabric Controller(NDFC)

# Deploy over VXLAN EVPN Fabric



# Deploy over VXLAN EVPN Fabric

## Connecting K8s nodes to Leaf Switches

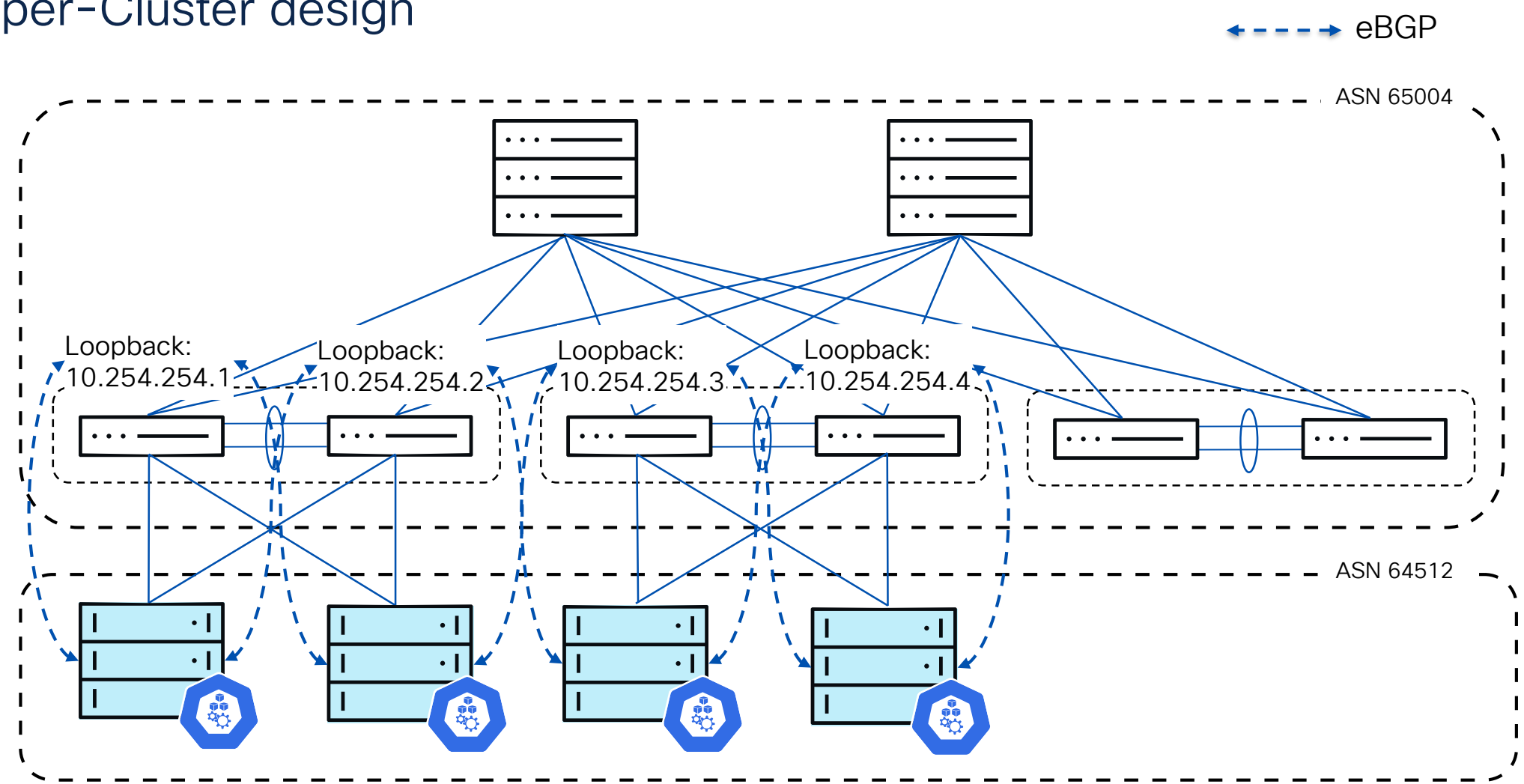



eBGP

- K8s nodes connect to Leaf switches using VPC or Active-Standby
- Node IPs are learned as type-2 routes
- Peering eBGP between K8s nodes and leaf switches using node IP and localized loopback addresses on each leaf switch
- `disable-peer-as-check` and `as-override` are needed
- loopback addresses must be reachable between the VPC peer members

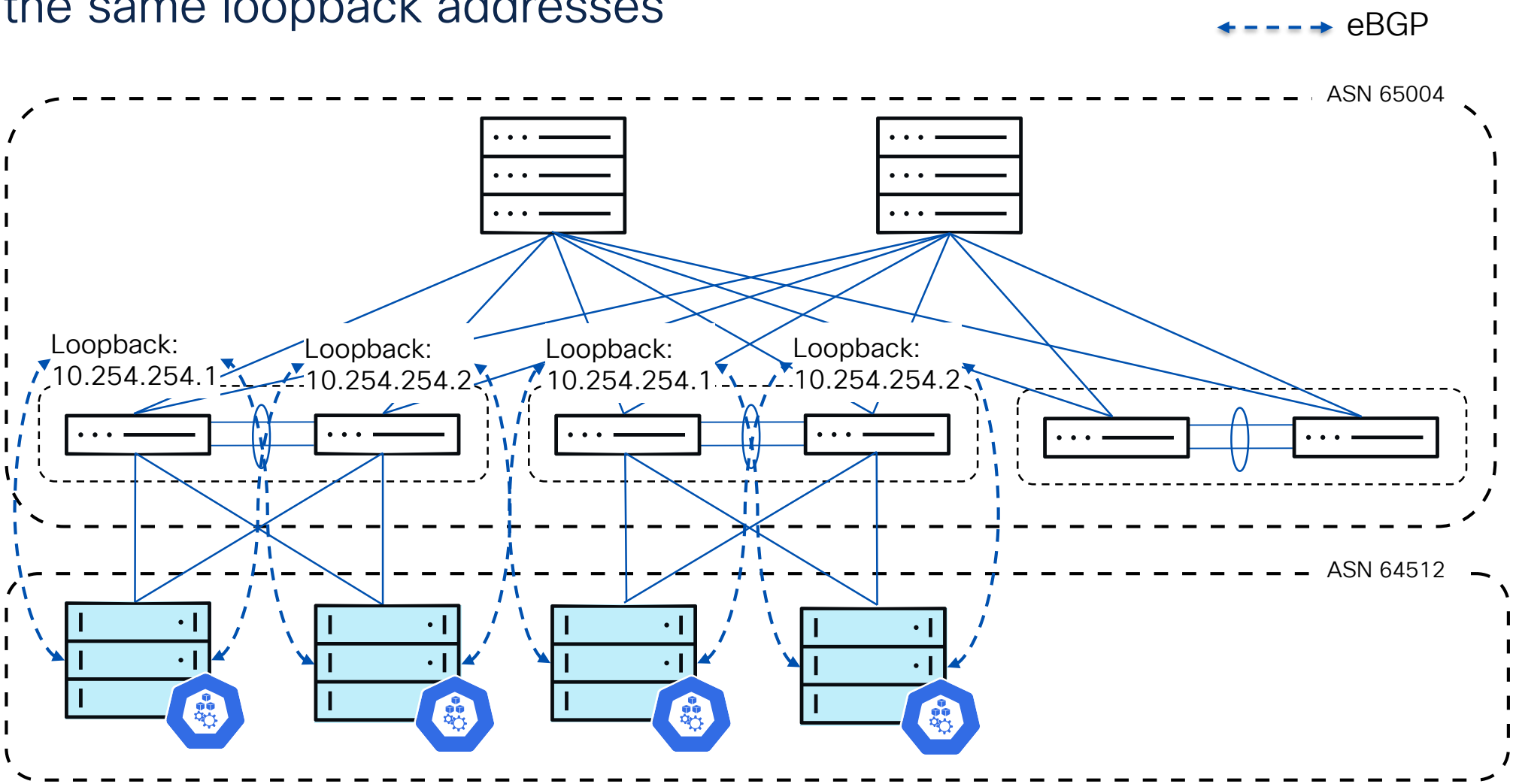
# Deploy over VXLAN EVPN Fabric

## As-per-Cluster design



# Deploy over VXLAN EVPN Fabric

## Use the same loopback addresses



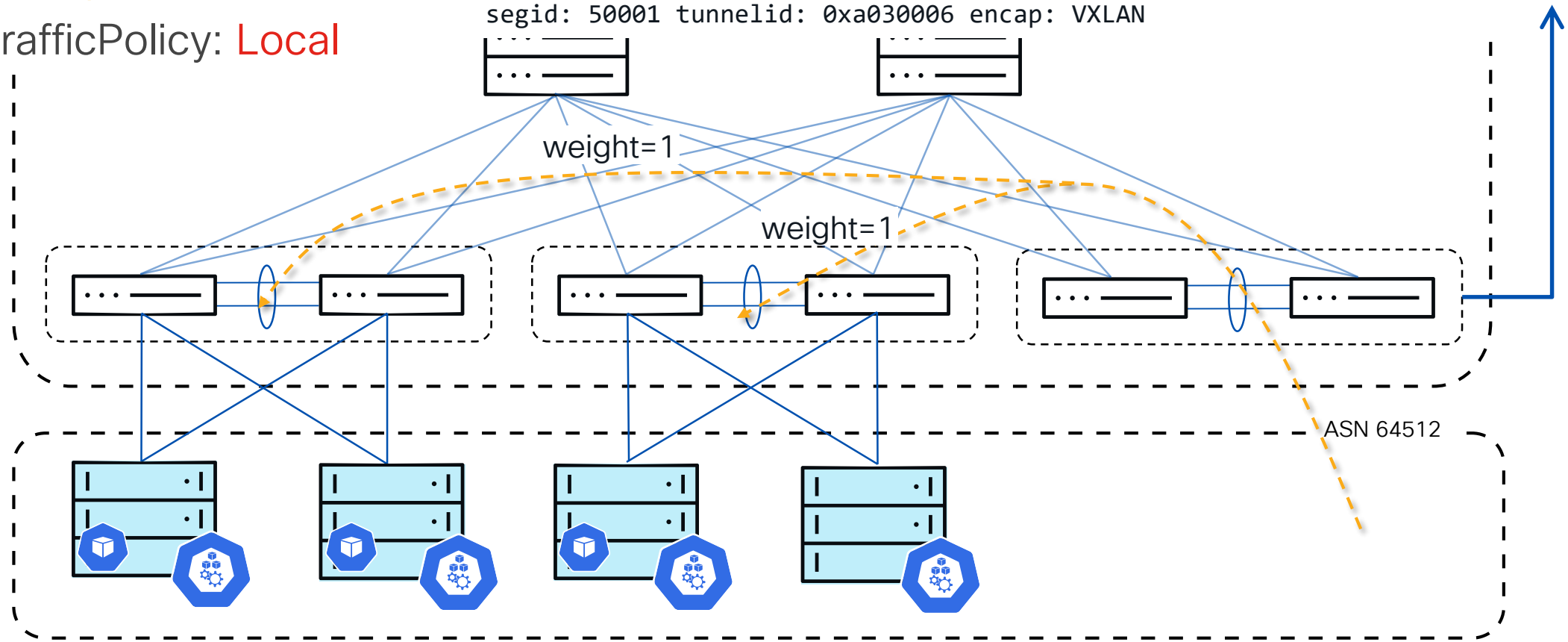
# Deploy over VXLAN EVPN Fabric

## Service Load Balancing

external service ip:
50.10.10.197/32

externalTrafficPolicy: Local

```
50.10.10.197/32, ubest/mbest: 2/0
    *via 10.3.0.3%default, [200/0], 01:24:29, bgp-65004, internal, tag 64512,
segid: 50001 tunnelid: 0xa030003 encap: VXLAN
    *via 10.3.0.6%default, [200/0], 01:24:26, bgp-65004, internal, tag 64512,
segid: 50001 tunnelid: 0xa030006 encap: VXLAN
```



# Deploy over VXLAN EVPN Fabric

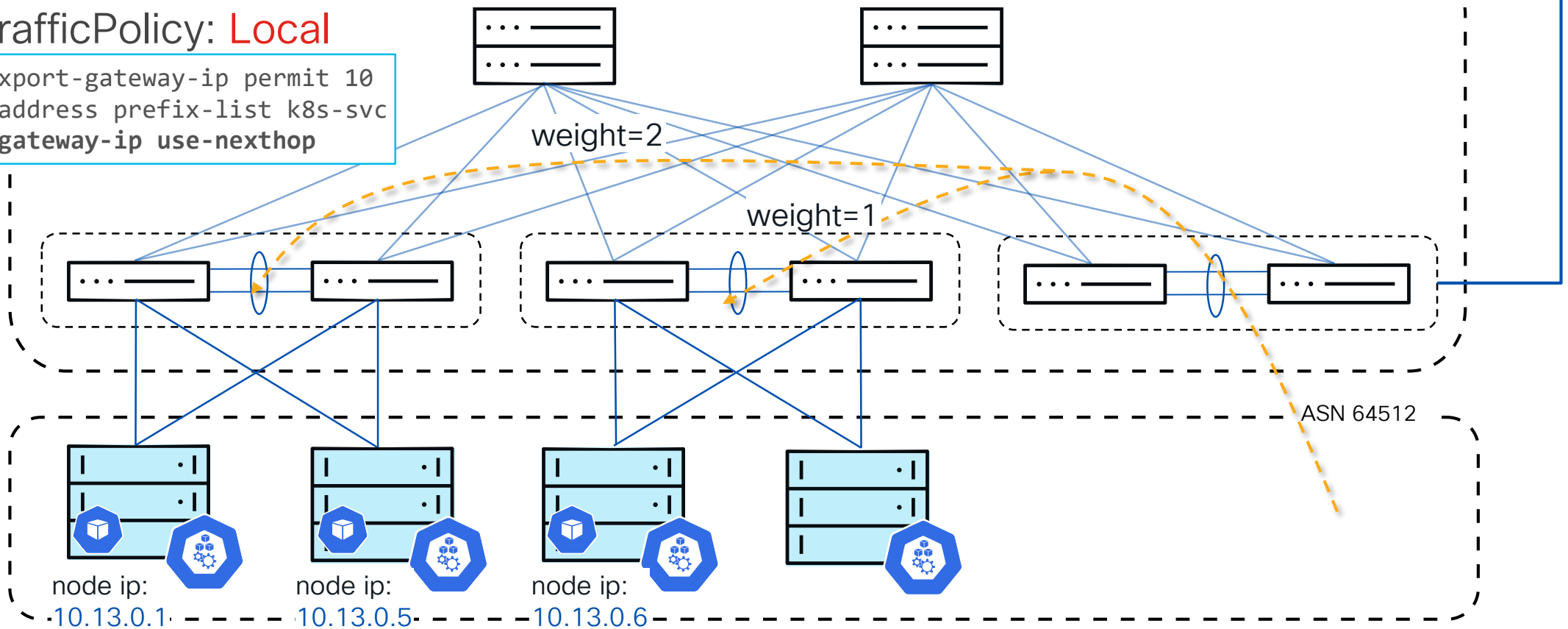
## Service Load Balancing

external service ip:
50.10.10.197/32

externalTrafficPolicy: Local

```
route-map export-gateway-ip permit 10
  match ip address prefix-list k8s-svc
  set evpn gateway-ip use-nexthop
```

```
50.10.10.197/32, ubest/mbest: 3/0
    *via 10.13.0.1, [200/0], 10:54:28, bgp-65004, internal, tag 64512
    *via 10.13.0.5, [200/0], 10:54:29, bgp-65004, internal, tag 64512
    *via 10.13.0.6, [200/0], 10:54:30, bgp-65004, internal, tag 64512
```



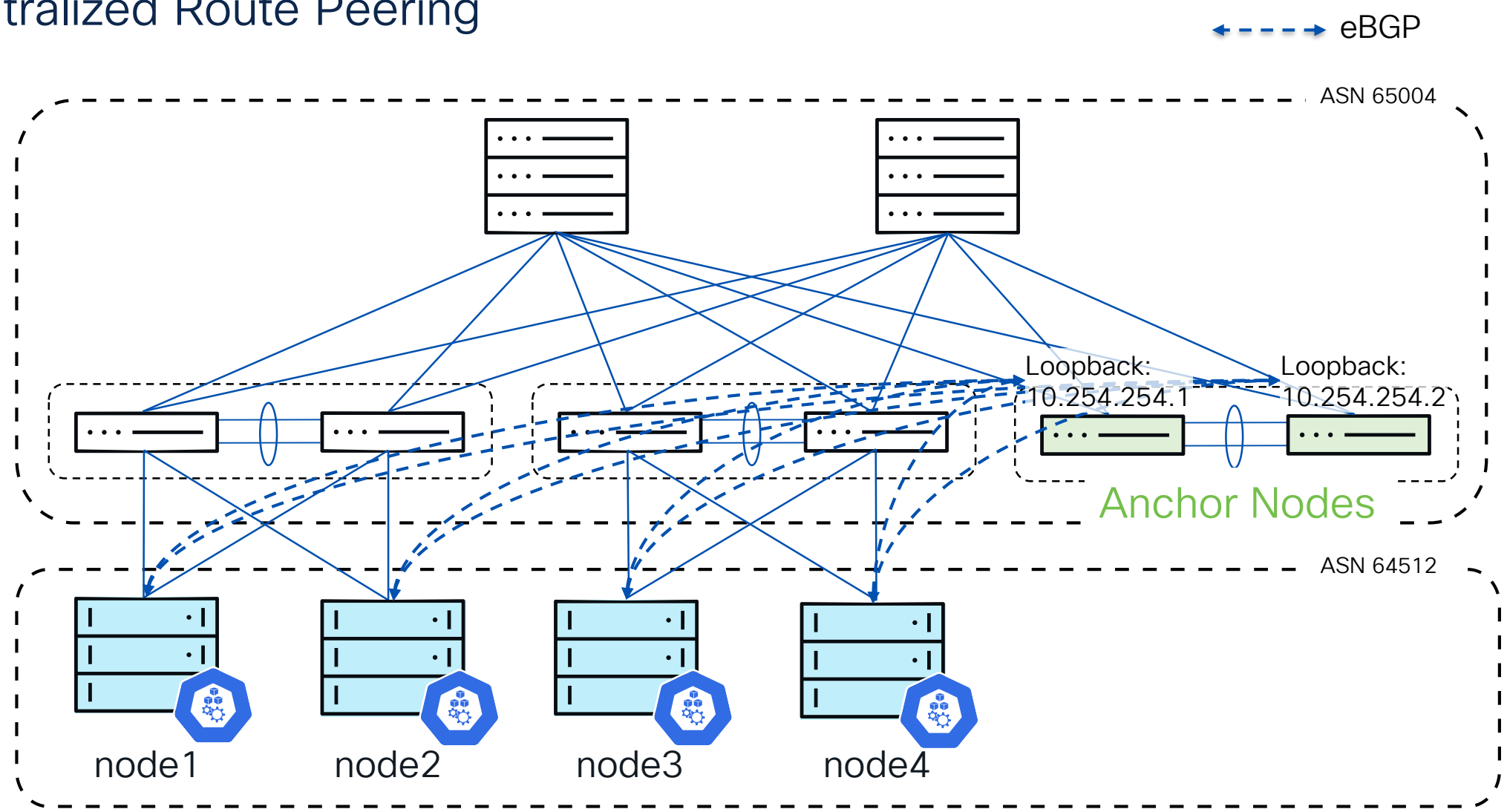

Proportional Multipath

# As-Per-Cluster design

- Using a single AS number per cluster reduces the complexity of bootstrapping K8s nodes
- Loopback addresses are local to the leaf switches
  - It does not need to be advertised to EVPN address family
    - But you will need iBGP peering between vPC peer switches
  - The same loopbacks can be used on all pairs of leaf switches
- Minimum BGP configuration can be tuned on CNI
  - `disable-peer-as-check` and `as-override` are needed on leaf switches
- Proportional Multipath can overcome the unevenly distributed workload

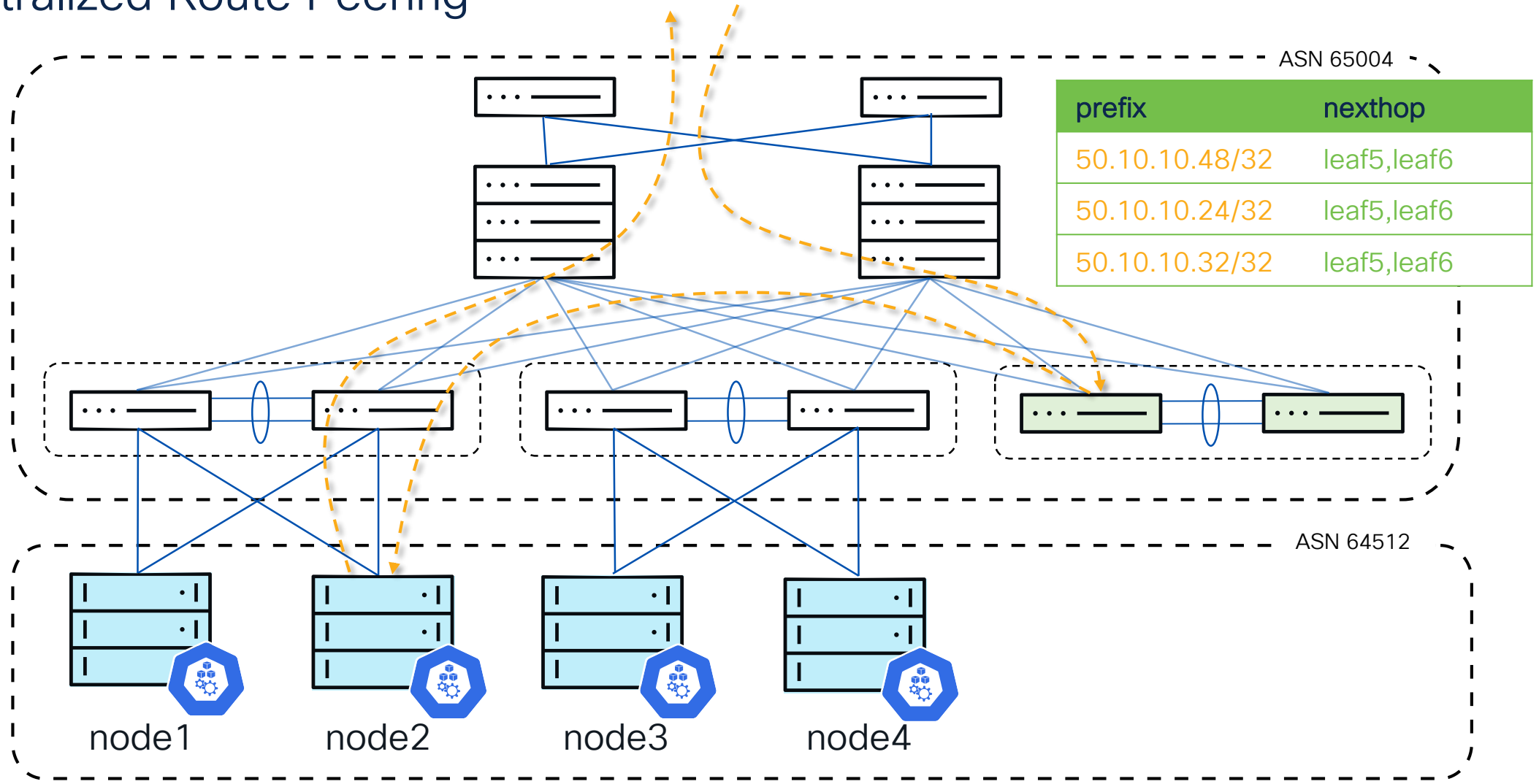
# Deploy over VXLAN EVPN Fabric

## Centralized Route Peering



# Deploy over VXLAN EVPN Fabric

## Centralized Route Peering

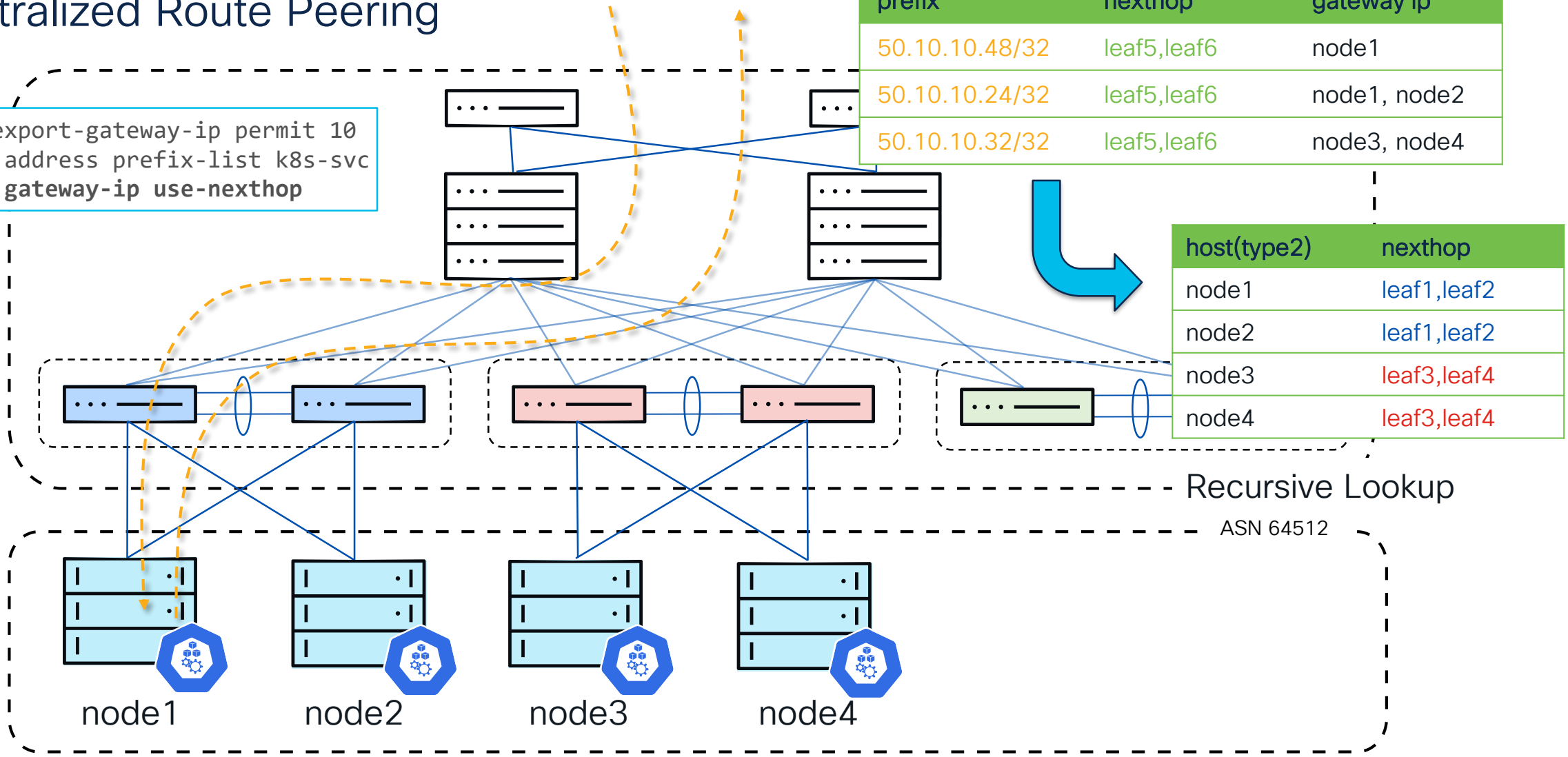


# Deploy over VXLAN EVPN Fabric

## Centralized Route Peering

```
route-map export-gateway-ip permit 10
  match ip address prefix-list k8s-svc
  set evpn gateway-ip use-nexthop
```

| prefix | nexthop | gateway ip |
|---|---|---|
| 50.10.10.48/32 | leaf5,leaf6 | node1 |
| 50.10.10.24/32 | leaf5,leaf6 | node1, node2 |
| 50.10.10.32/32 | leaf5,leaf6 | node3, node4 |

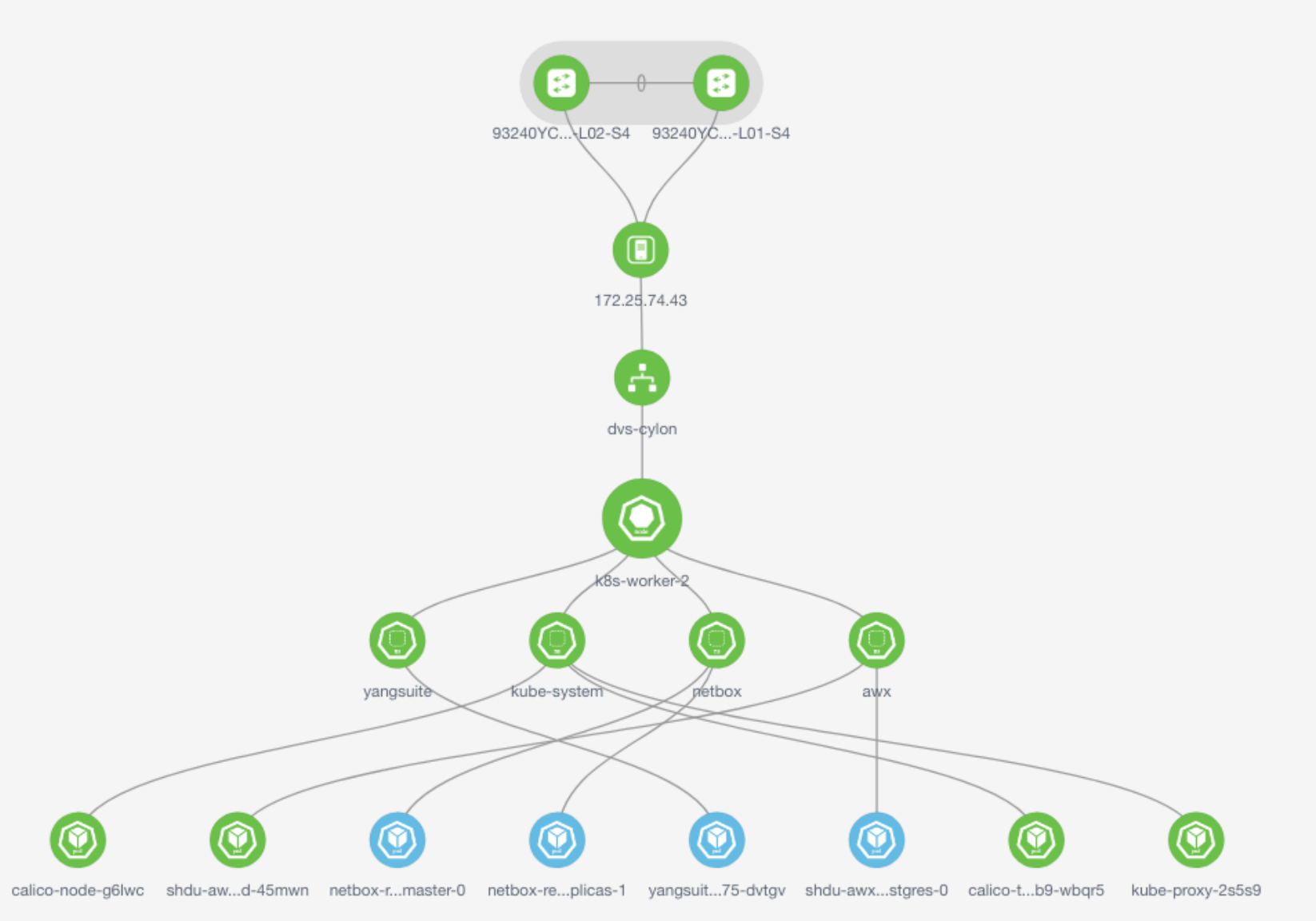



| host(type2) | nexthop |
|---|---|
| node1 | leaf1,leaf2 |
| node2 | leaf1,leaf2 |
| node3 | leaf3,leaf4 |
| node4 | leaf3,leaf4 |

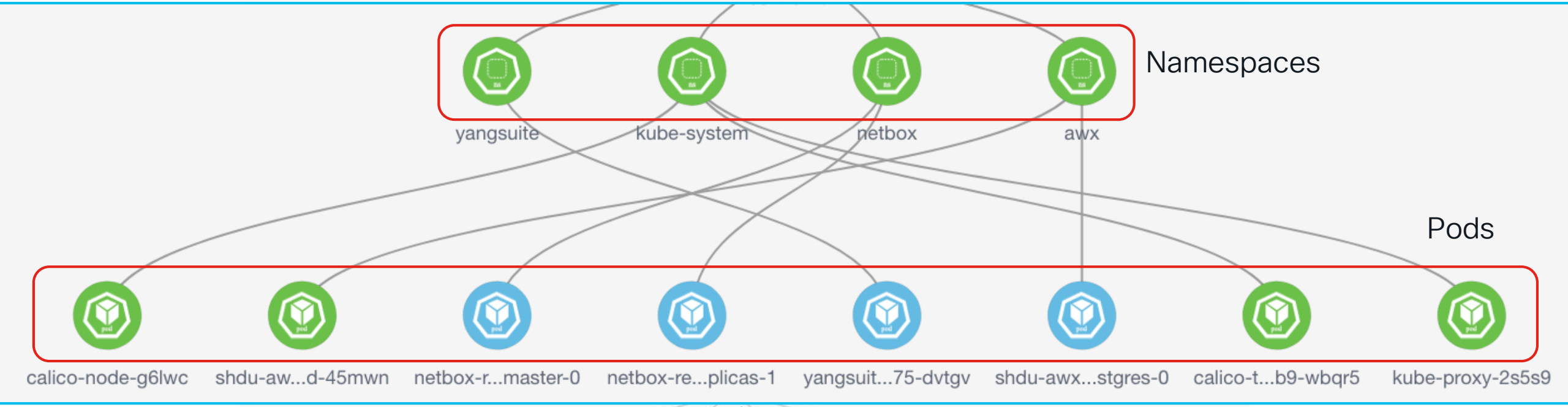
# Agenda

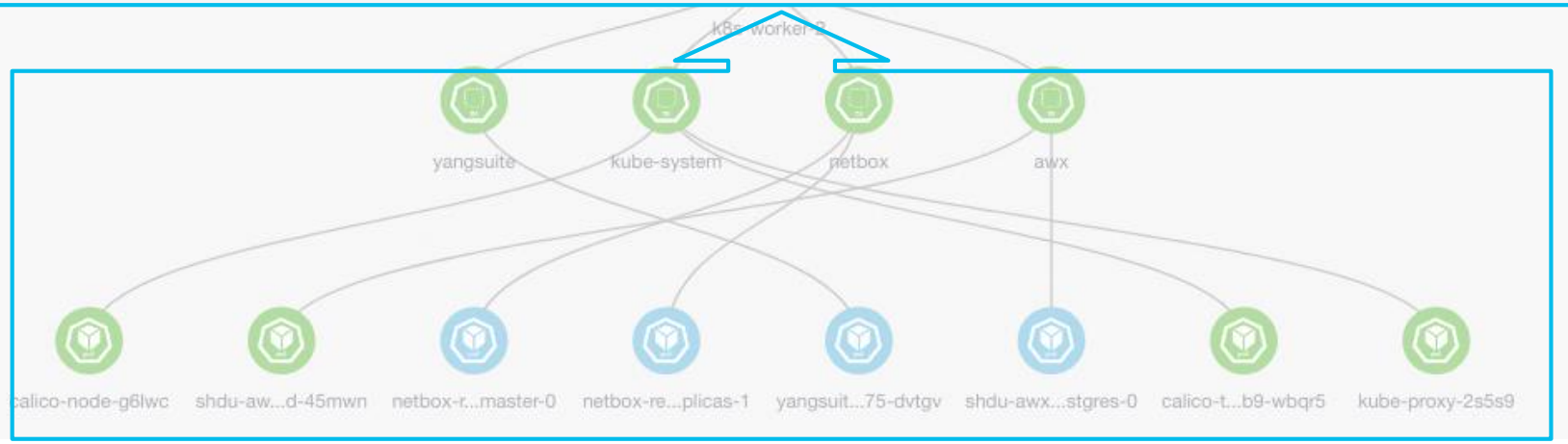
- What is Container Network Interface(CNI) Plugin
- A simple network design – software overlay
- A more scalable design – native routing
  - Design BGP network on IP Fabric
  - Design BGP network on VXLAN EVPN Fabric
- **Integration with Nexus Dashboard Fabric Controller(NDFC)**

# Kubernetes Visualization with NDFC

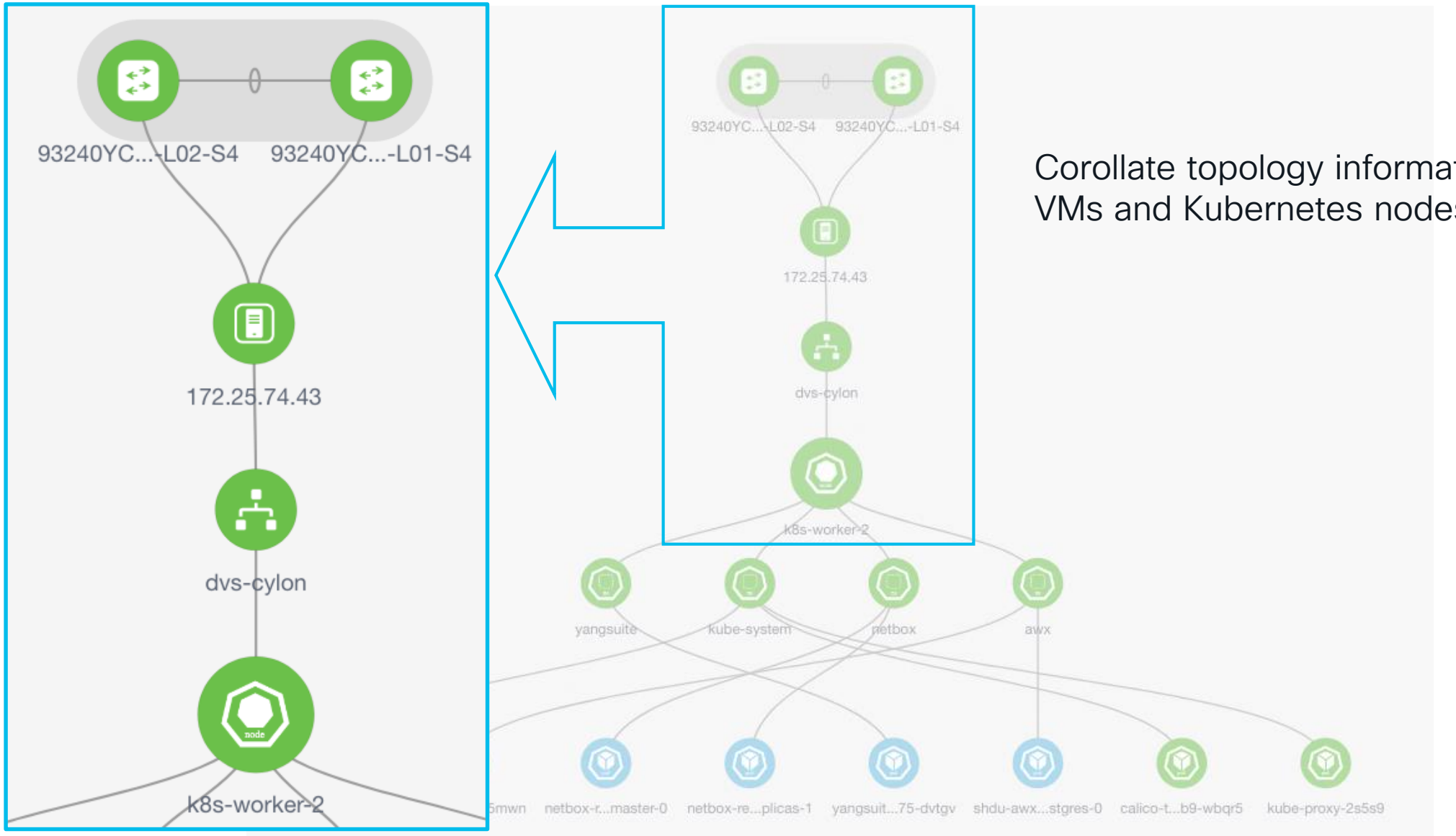


# Kubernetes Visualization with NDFC



# Kubernetes Visualization with NDFC




Corollate topology information with VMs and Kubernetes nodes

# Summary

- Software overlay or NAT is simple with some trade-offs
- Greenfield Kubernetes network does not require the L2 extension
- The best practice is peering BGP neighborship with local switches
- Centralized Route Peering can simplify the configuration of the leaf switches
- All the necessary features are shipped today on NX-OS

# Reference

- Cisco NX-OS Calico Network Design White Paper
  - https://www.cisco.com/c/en/us/td/docs/dcn/whitepapers/cisco-nx-os-calico-network-design.html
- Configuring Proportional Multipath for VNF
  - https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/93x/vxlan/configuration/guide/b-cisco-nexus-9000-series-nx-os-vxlan-configuration-guide-93x/b-cisco-nexus-9000-series-nx-os-vxlan-configuration-guide-93x_appendix_011010.html

# Complete Your Session Evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to **win 1 of 5 full conference passes** to Cisco Live 2025.


**Earn 100 points** per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn **exclusive prizes!**

Complete your surveys in the **Cisco Live mobile app.**

# Continue your education

- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Engineer meeting
- Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs
- Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand


The bridge to possible
Thank you
#CiscoLive