cisco Live!

Let’s go

#CiscoLive

The bridge to possible

Kubernetes (K8s) Infrastructure Connectivity

Network Designs for the Modern Data Center

Shangxin Du
Technical Marketing Engineer, Datacenter Switching

BRKDCN-2662

Cisco Webex App

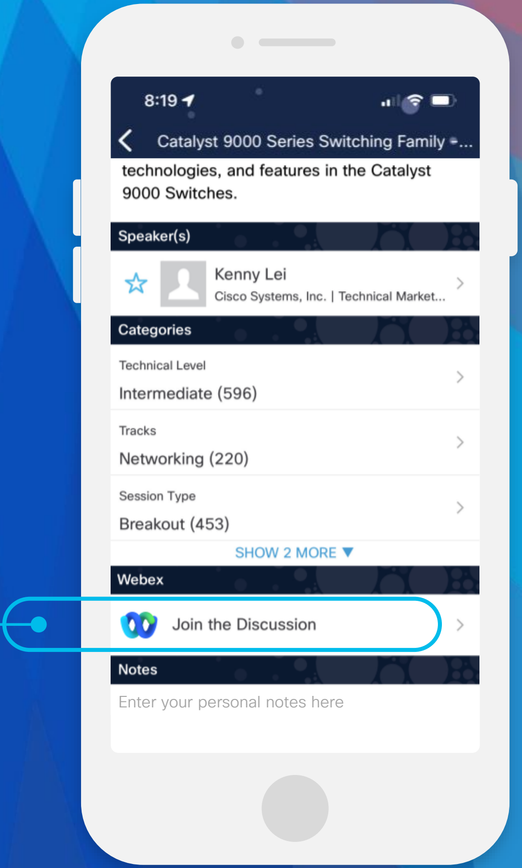
Questions?

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 9, 2023.



8:19

Catalyst 9000 Series Switching Family •...

technologies, and features in the Catalyst 9000 Switches.

Speaker(s)

☆ Kenny Lei
Cisco Systems, Inc. | Technical Market...

Categories

Technical Level
Intermediate (596)

Tracks
Networking (220)

Session Type
Breakout (453)

SHOW 2 MORE ▼

Webex

Join the Discussion

Notes

Enter your personal notes here

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

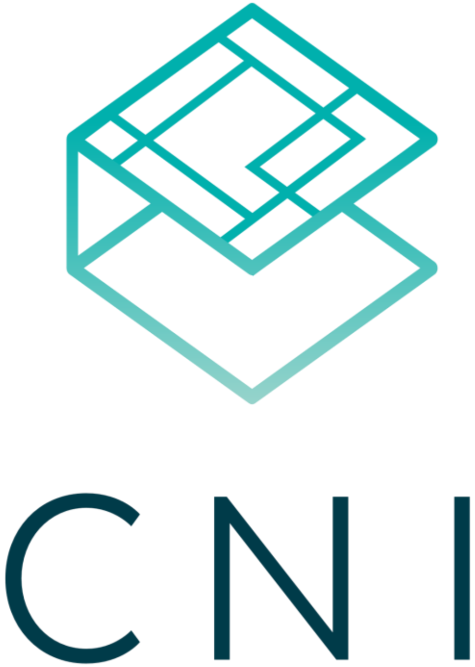
Agenda

- What is Container Network Interface(CNI) Plugin
- Design the Kubernetes network on IP Fabric
- Design the Kubernetes network on VXLAN EVPN Fabric
- Integration with Nexus Dashboard Fabric Controller(NDFC)

Agenda

- **What is Container Network Interface(CNI) Plugin**
- Design the Kubernetes network on IP Fabric
- Design the Kubernetes network on VXLAN EVPN Fabric
- Integration with Nexus Dashboard Fabric Controller(NDFC)

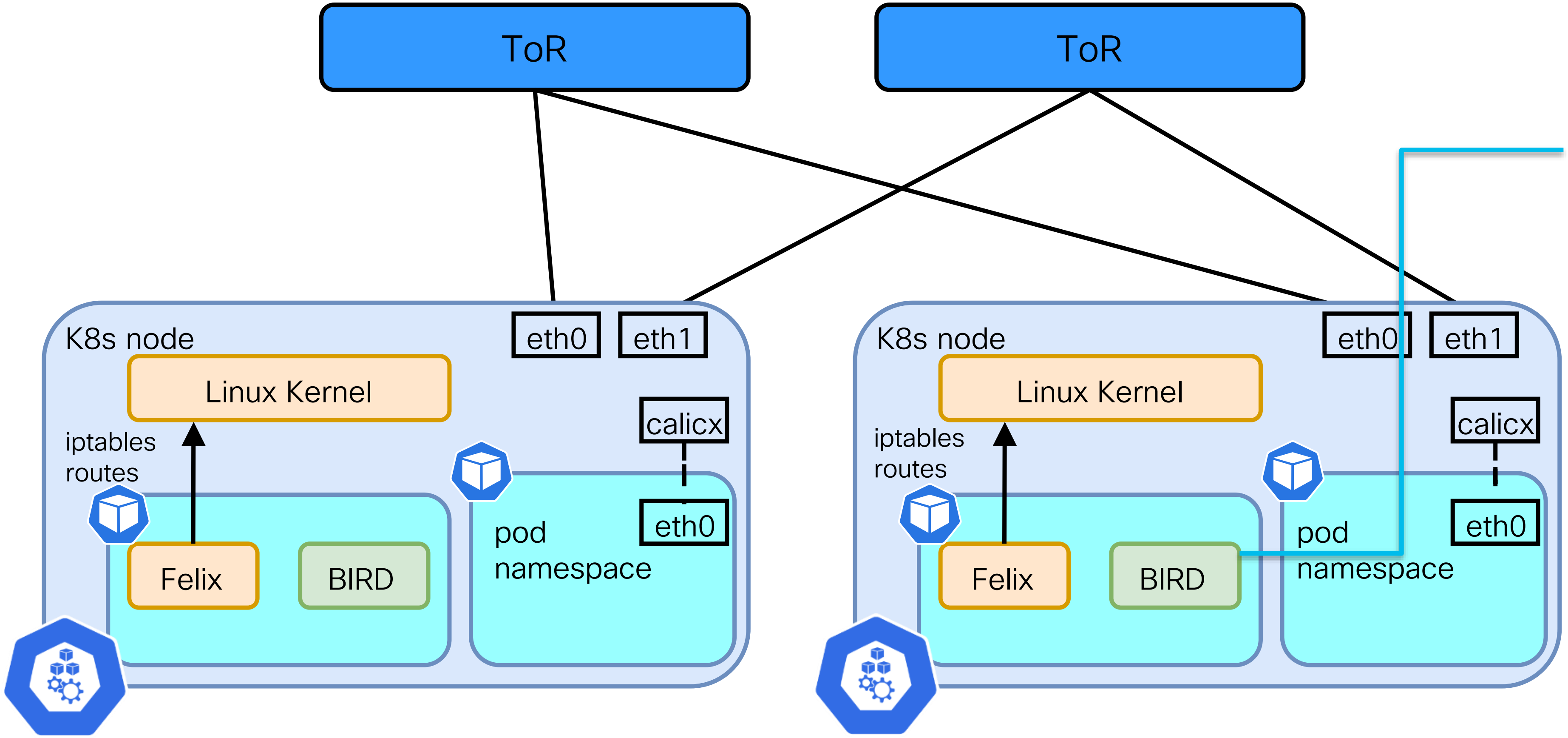
“Outsourcing the issue” - Container Networking Interface



- 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.

Project Calico

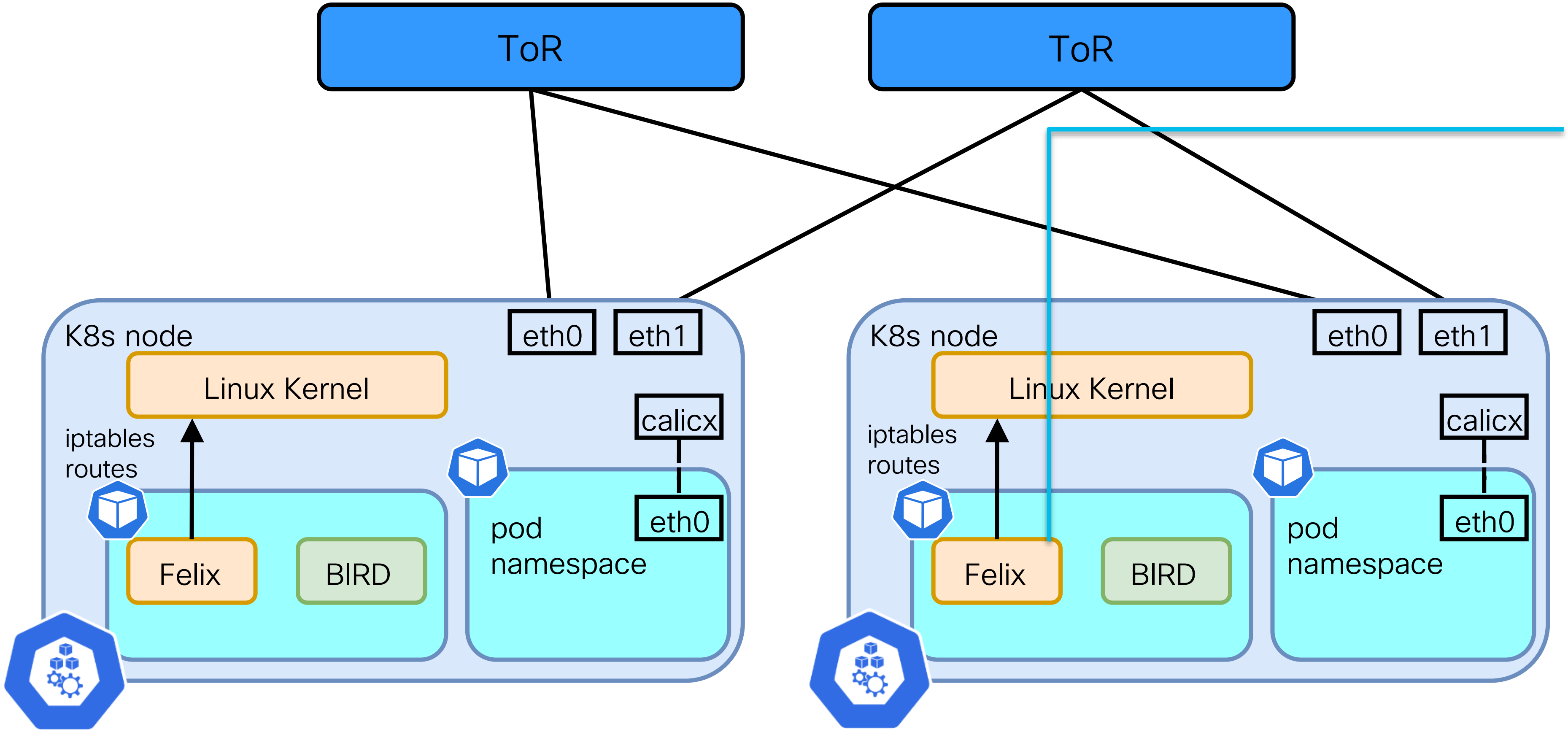
A CNI plugin of Kubernetes



BIRD: It is a routing daemon responsible for peering with other K8s nodes and exchanging routes of pod network and service network for inter-node communication.

Project Calico

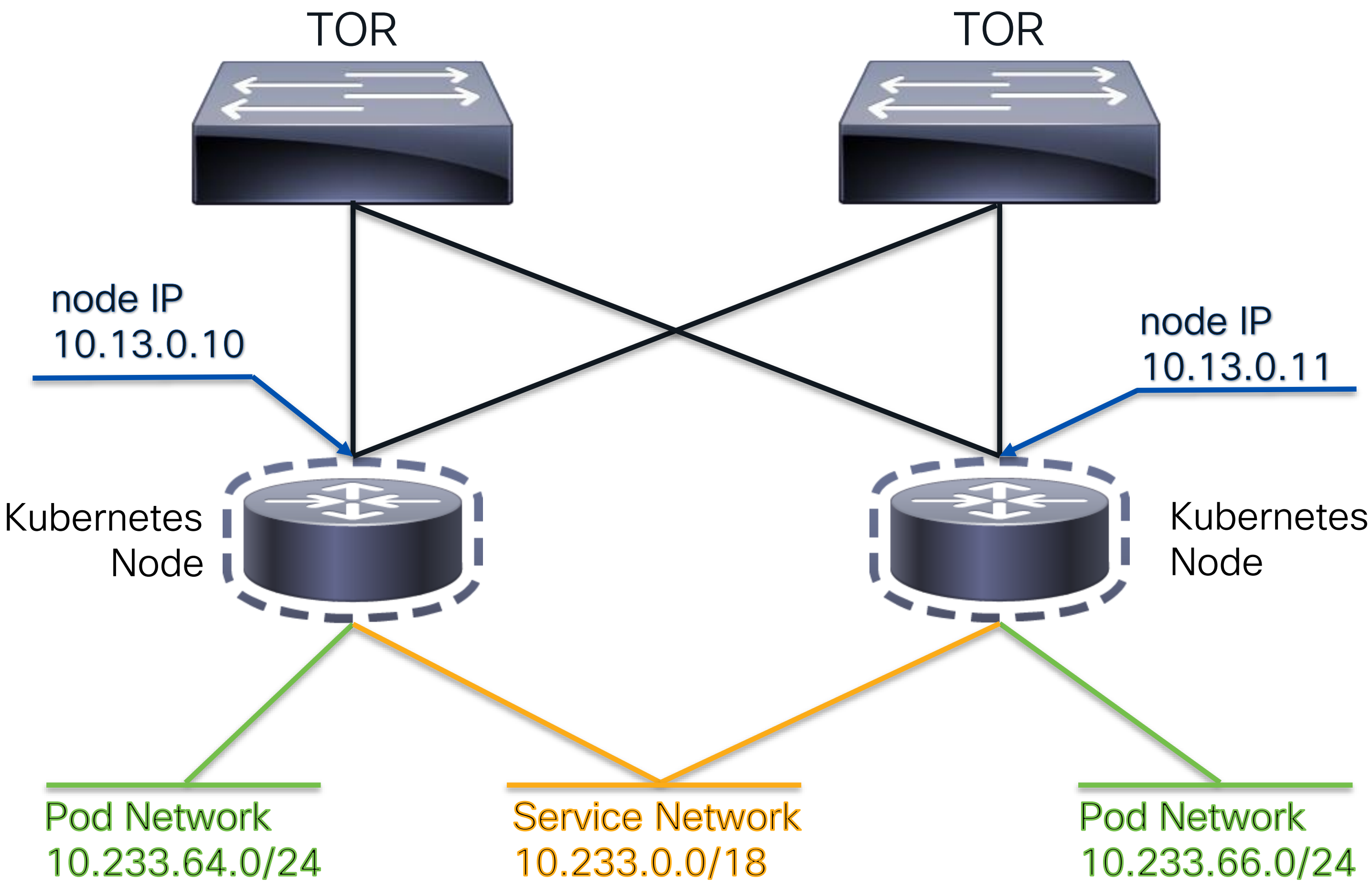
A CNI plugin of Kubernetes



Felix: Running in same pod as BIRD, programs routes and ACLs (iptables) and anything required on Calico node to provide connectivity for the pods scheduled on that node

Project Calico

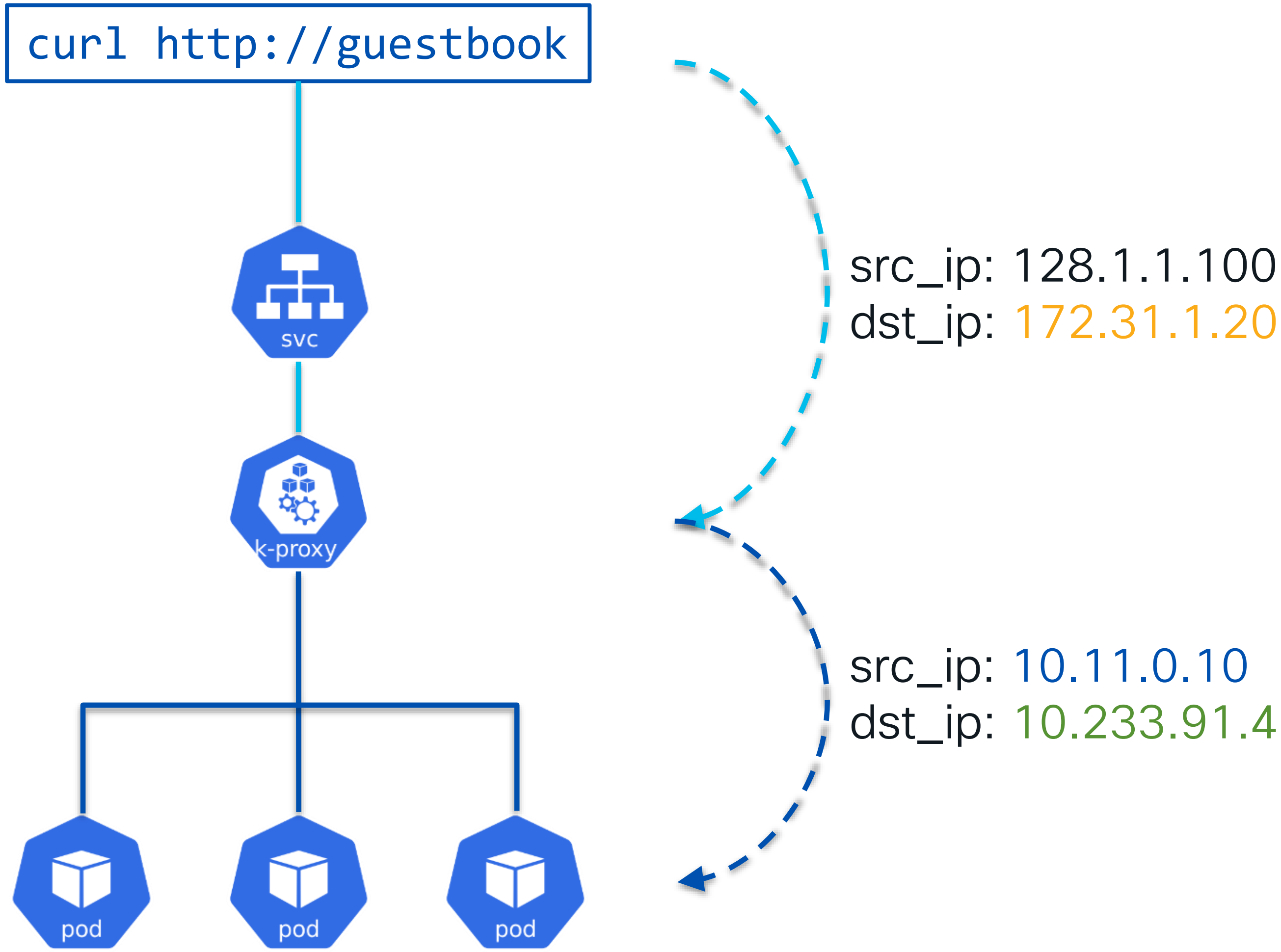
Simplified



- Each Calico node has one node IP
- one or more ranges of IP addresses (CIDRs) for Pod Networks
- a shared network for the whole Kubernetes cluster which is called the Service Network.

Kubernetes Service

A life of packet



```
curl http://guestbook
```

src_ip: 128.1.1.100
dst_ip: 172.31.1.20

src_ip: 10.11.0.10
dst_ip: 10.233.91.4

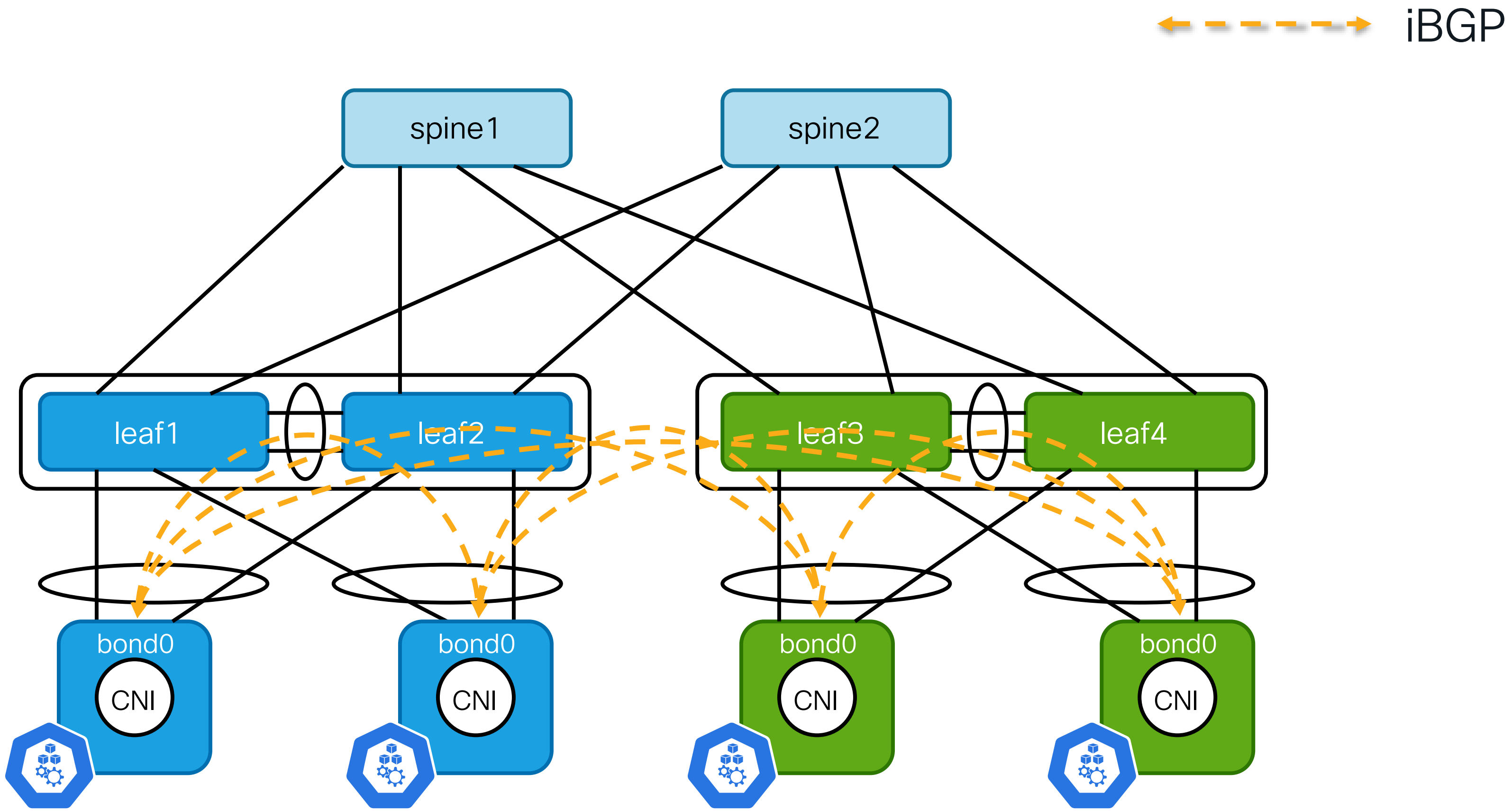
- The HTTP request is sent to service ip
- One of the Kubernetes nodes will first receive the request
- Source IP is rewritten to node ip and destination ip is rewritten to one of the pod ips

Agenda

- What is Container Network Interface(CNI) Plugin
- **Design the Kubernetes network on IP Fabric**
- Design the Kubernetes network on VXLAN EVPN Fabric
- Integration with Nexus Dashboard Fabric Controller(NDFC)

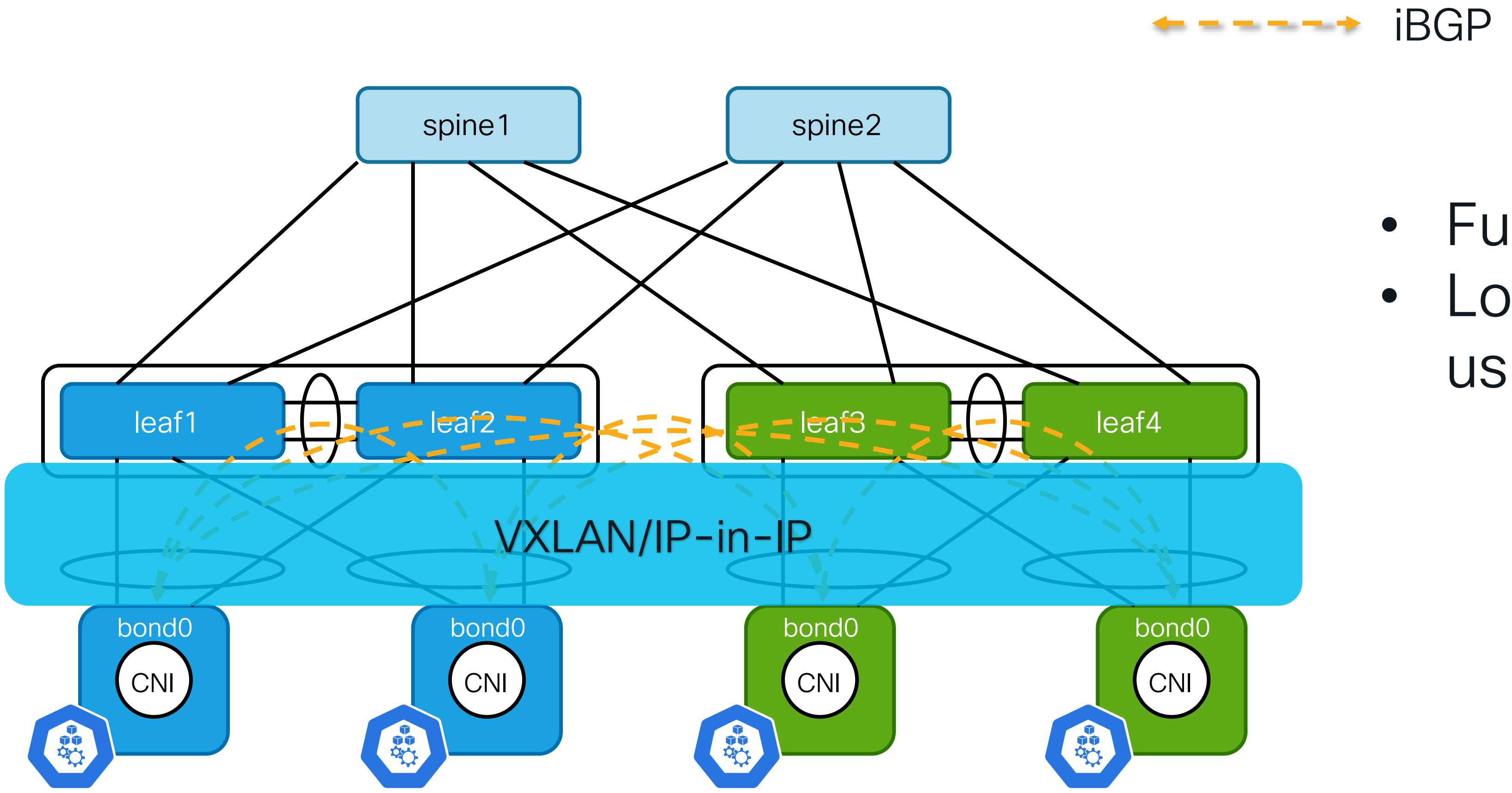
Network Architecture

Full mesh



Network Architecture

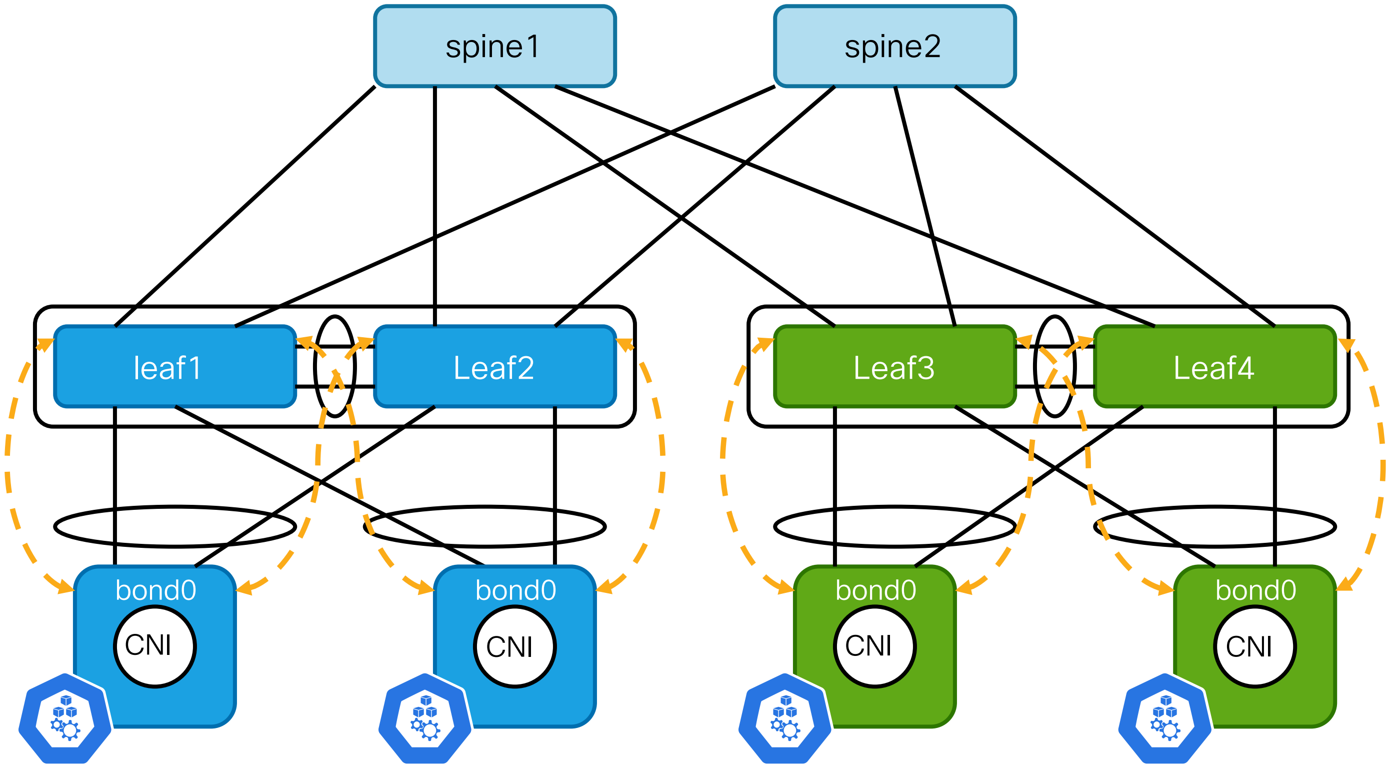
Full mesh data plane



- Full mesh does not scale!
- Losing visibility when using software overlay

Network Architecture

Peer with Switch



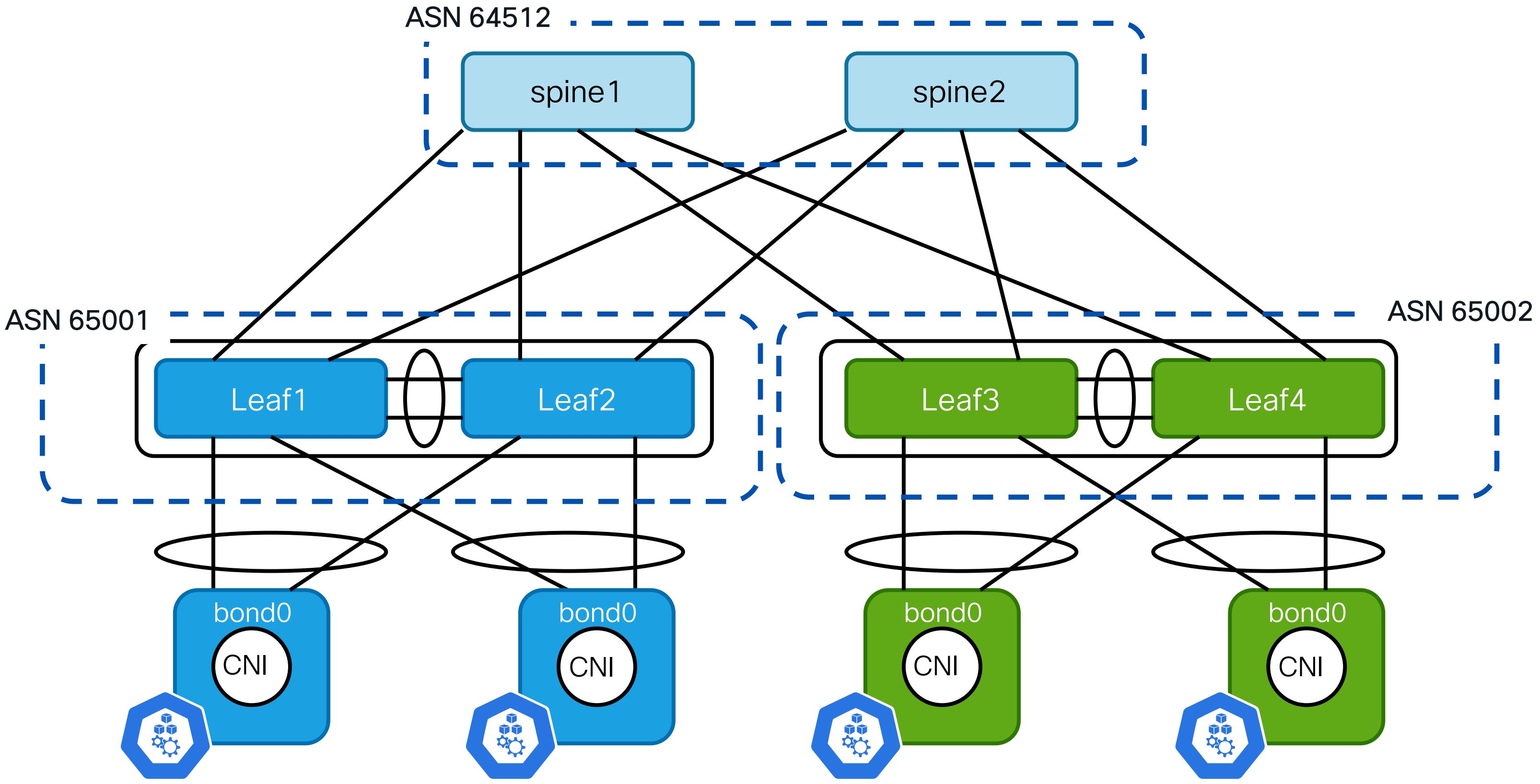
⇠ ⇢ iBGP

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

```
apiVersion: projectcalico.org/v3
kind: IPPool
metadata:
  name: default-pool
spec:
  blockSize: 24
  cidr: 10.233.64.0/20
  ipipMode: Never
  nodeSelector: all()
  vxlanMode: Never
```

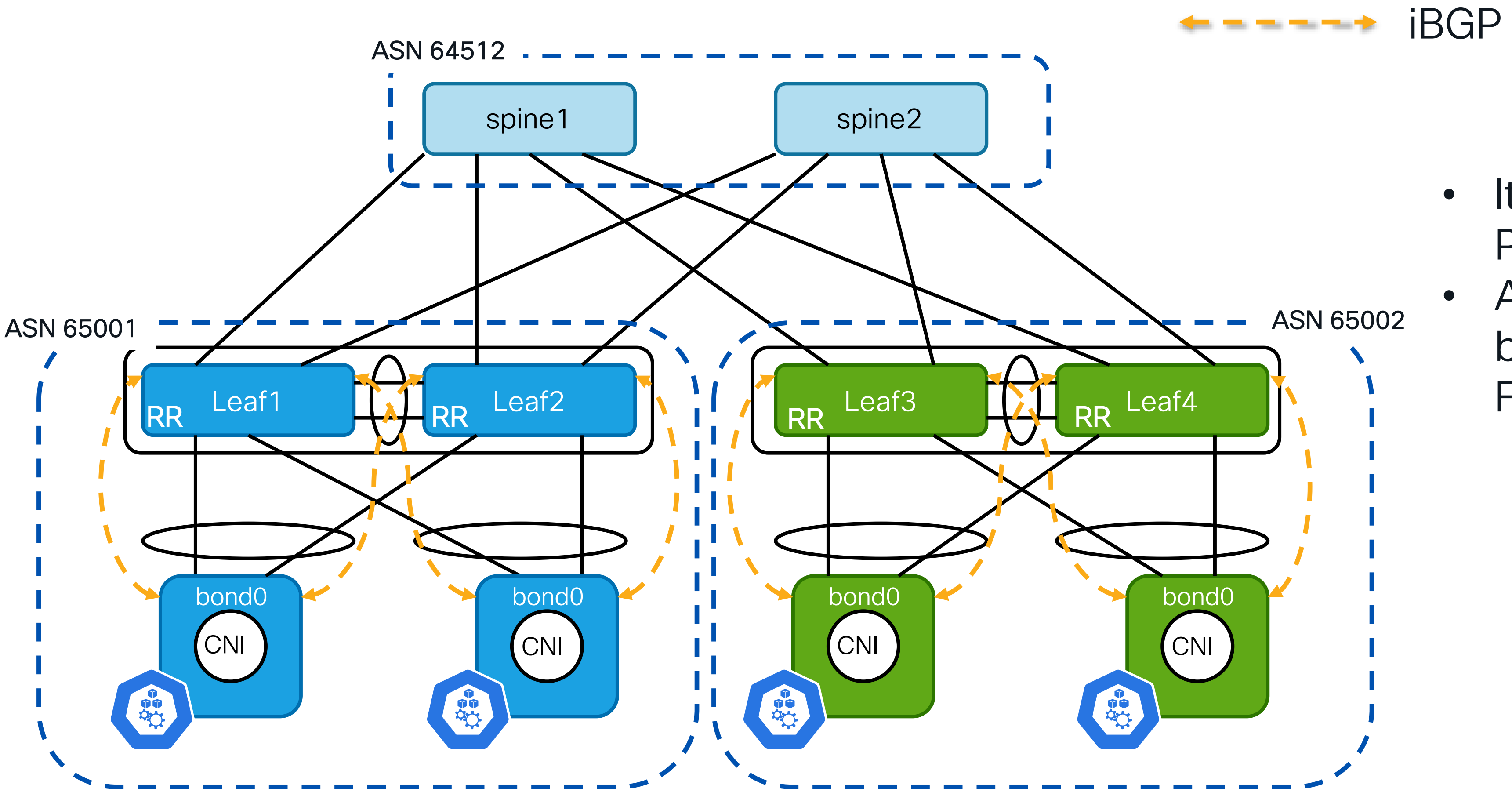
Network Architecture

Deploy Over IP Fabric



Network Architecture

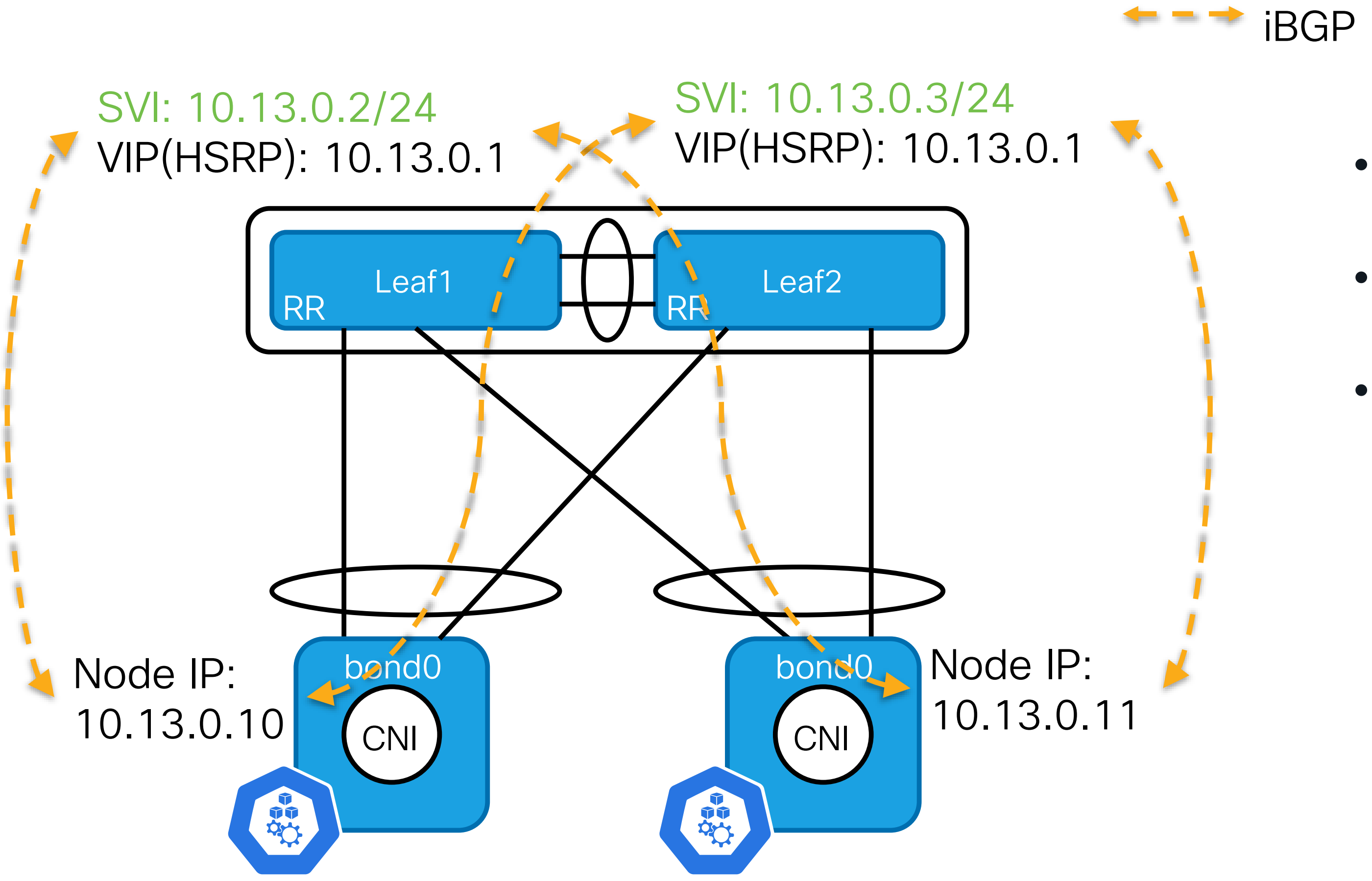
Deploy Over IP Fabric



- It is usually referred to as AS-Per-Rack design.
- AS-Per-Rack is recommended by Calico, but exclusively for IP Fabric(RFC 7938)

Network Architecture

Deploy Over IP Fabric



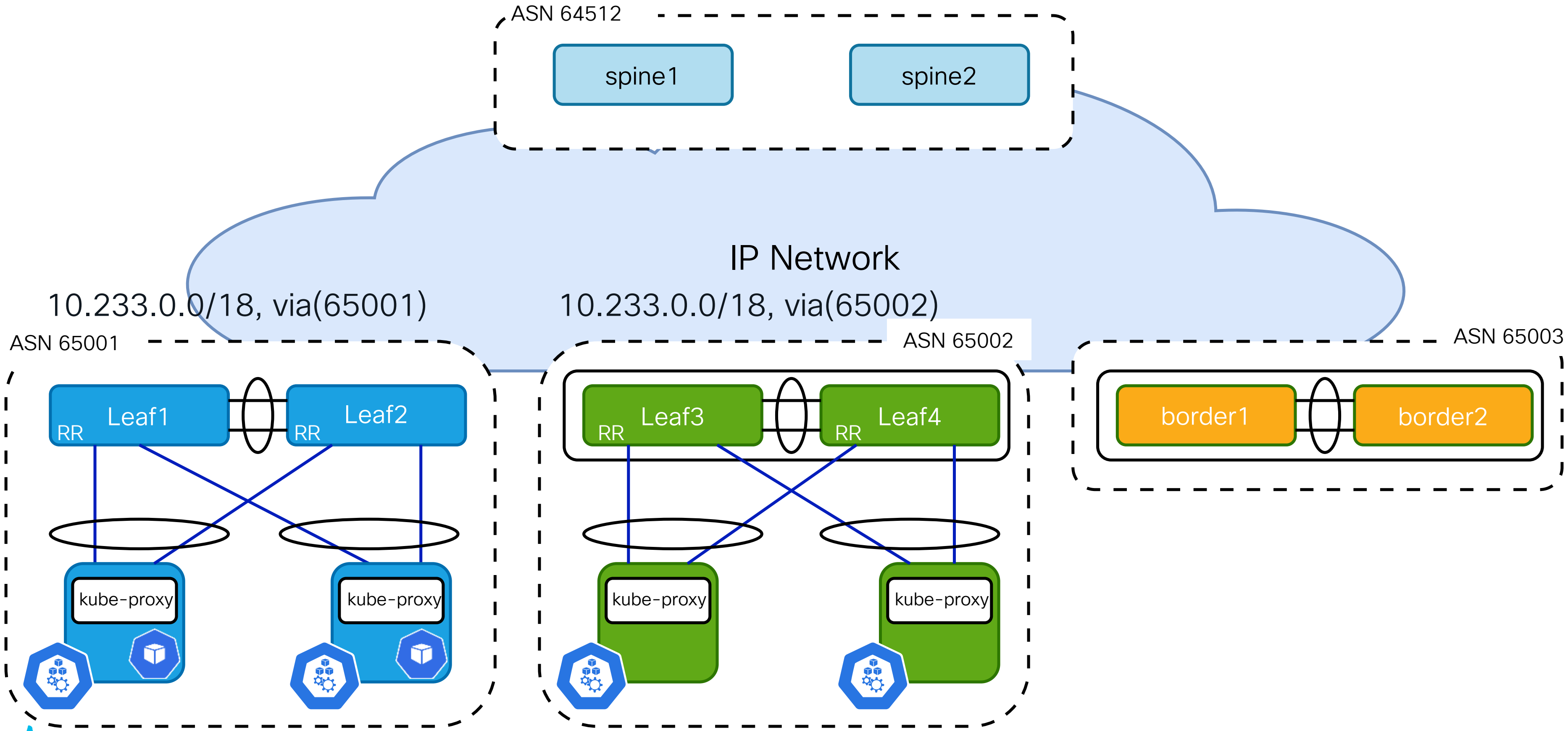
- HSRP/VRRP is used for gateway redundancy
- Kubernetes nodes peer with the primary IP address of SVI
- The node subnets are advertised into BGP to provide nodes reachability

Deploy over IP Fabric

Service Traffic

Service Subnet:
10.233.0.0/18

```
10.233.0.0/18, ubest/mbest: 4/0
    *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
```



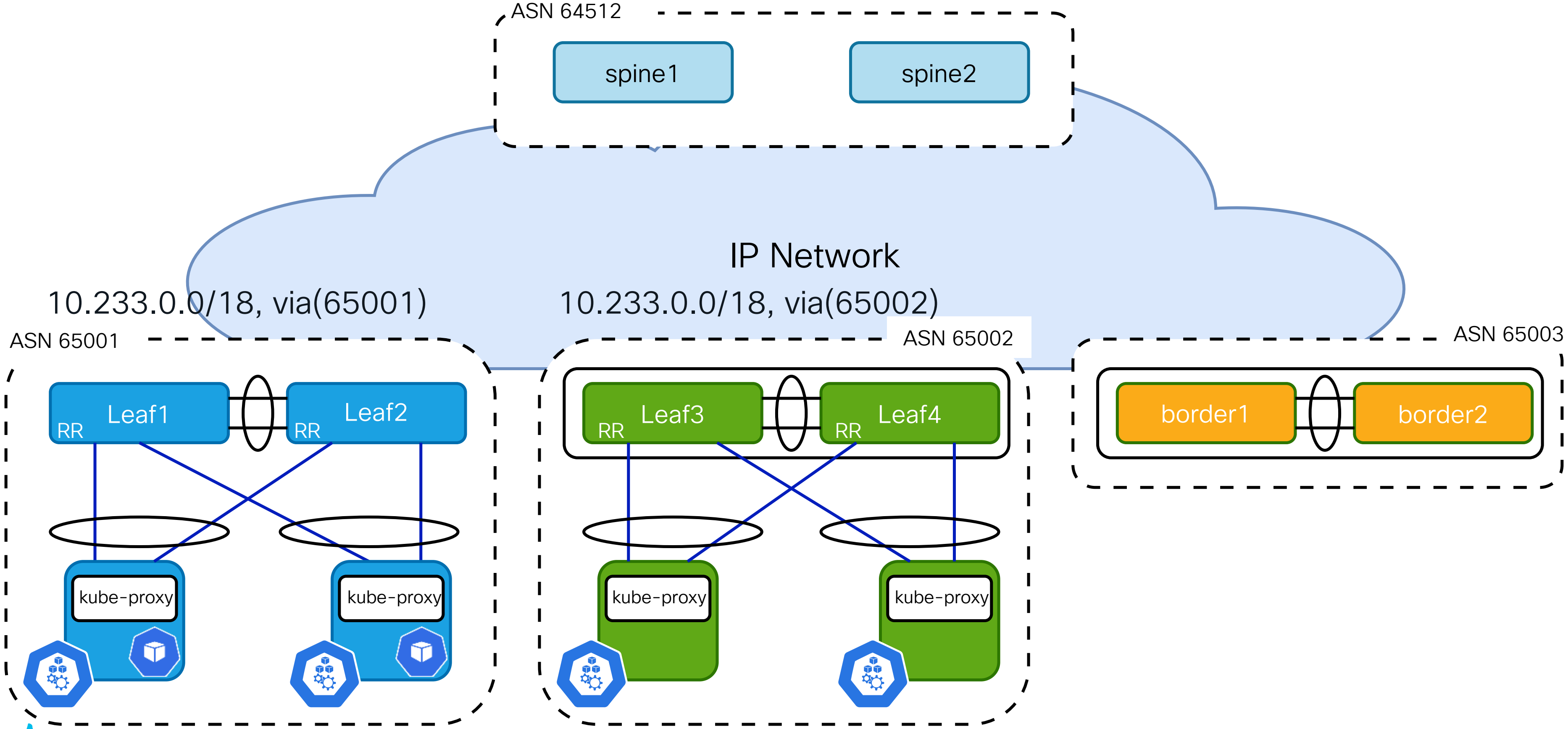
Deploy over IP Fabric

Service Traffic

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

Service Subnet:
10.233.0.0/18

```
10.233.0.0/18, 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
```



Deploy over IP Fabric

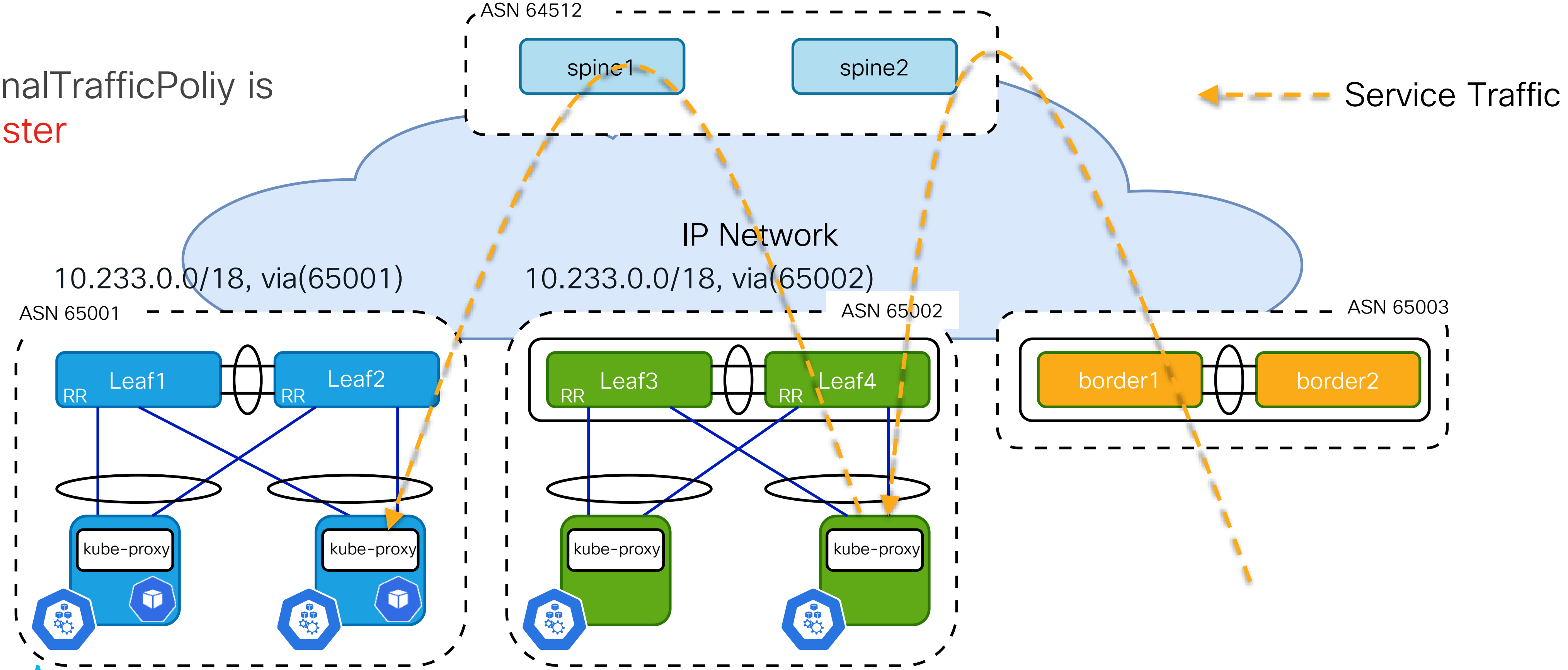
Sub-optimal service traffic

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

Service Subnet:
10.233.0.0/18

```
10.233.0.0/18, 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
```

K8s externalTrafficPoliy is set to Cluster



Deploy over IP Fabric

Avoid Second Hop of Service Traffic

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

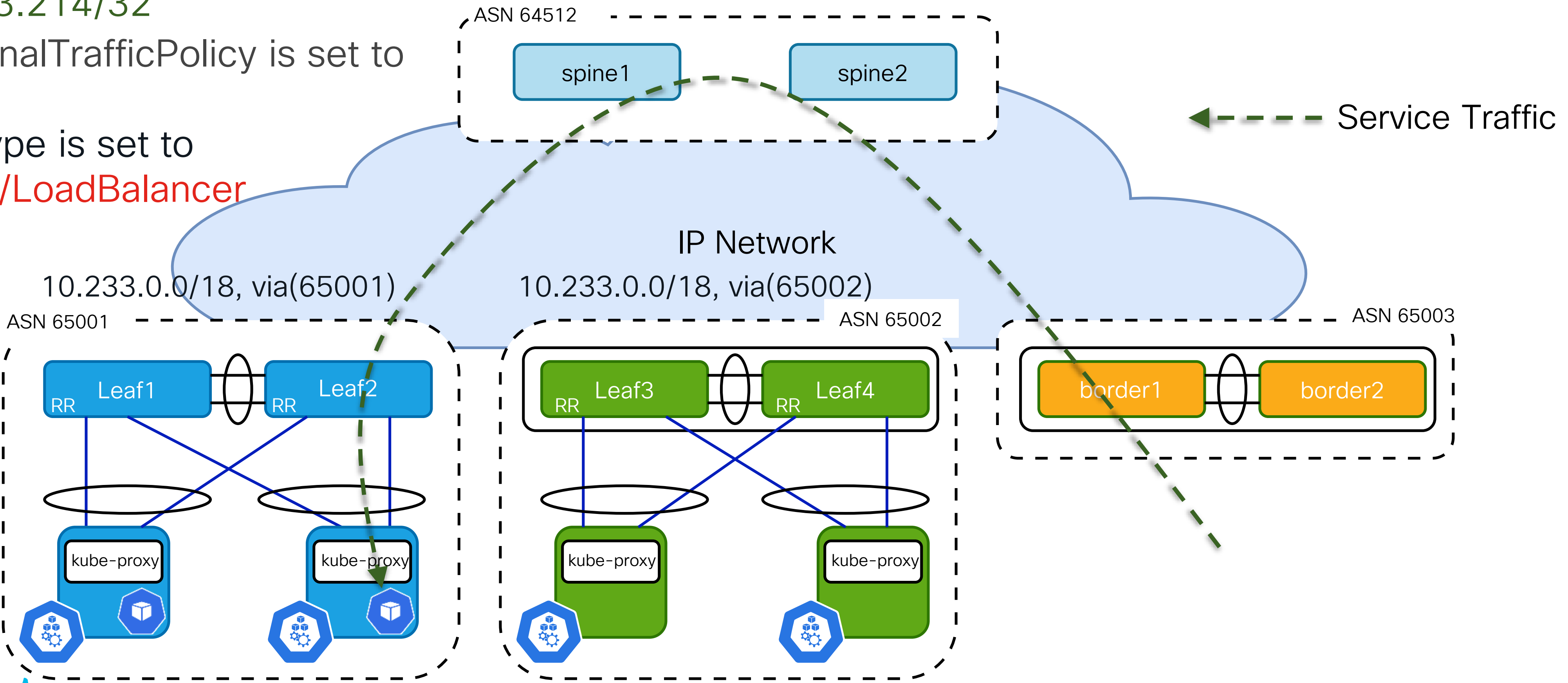
```
10.233.63.214/32, ubest/mbest: 2/0
    *via 10.4.0.37, [20/0], 2d10h, bgp-64512, external, tag 65001
    *via 10.4.0.45, [20/0], 2d10h, bgp-64512, external, tag 65001
```

Service Subnet:
10.233.0.0/18

Service ip:
10.233.63.214/32

K8s externalTrafficPolicy is set to
Local

Sevice Type is set to
NodePort/LoadBalancer



Exposing Services

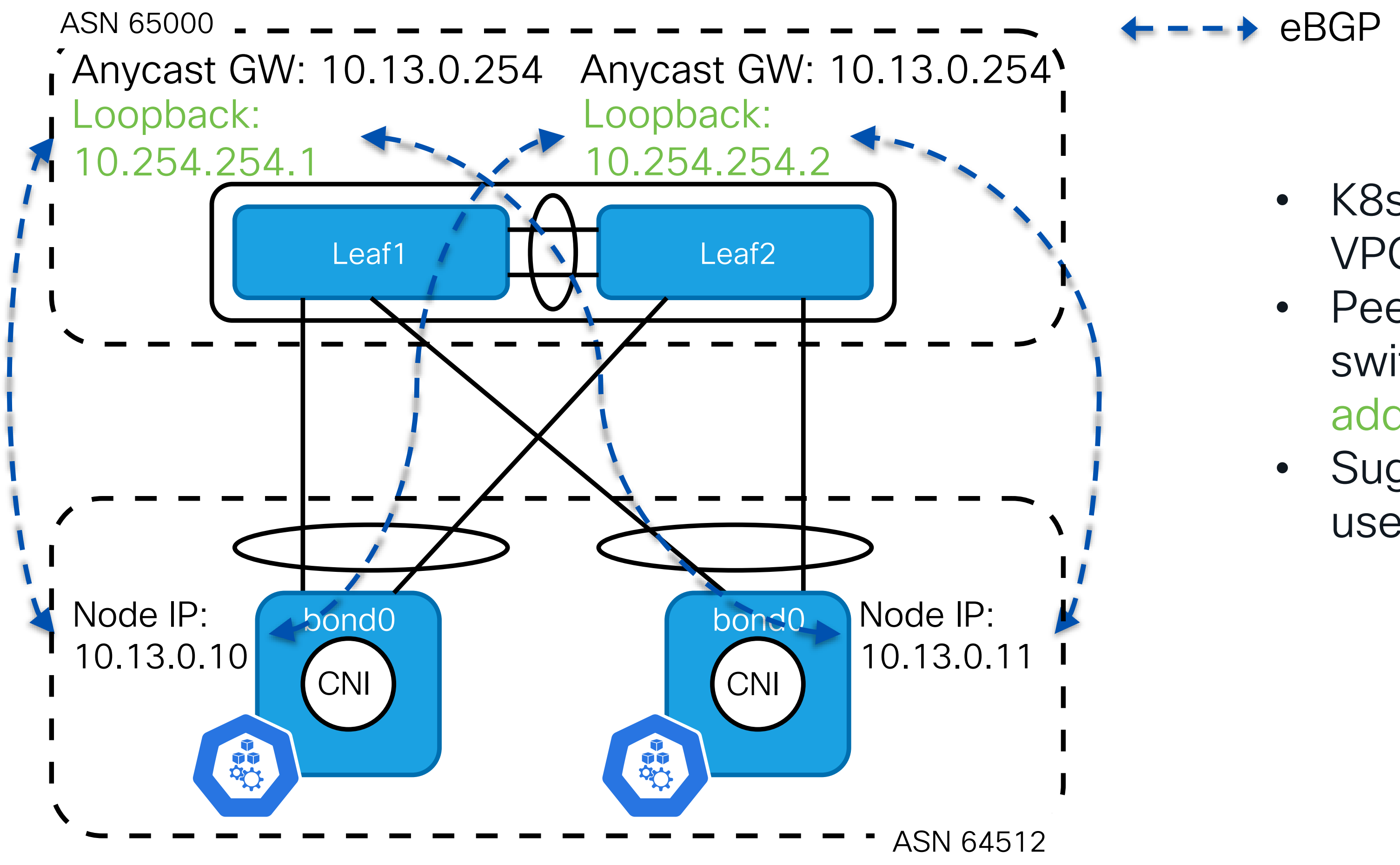
A note on "externalTrafficPolicy"

- 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
    - Pods can be spread evenly with topologySpreadConstraints

Agenda

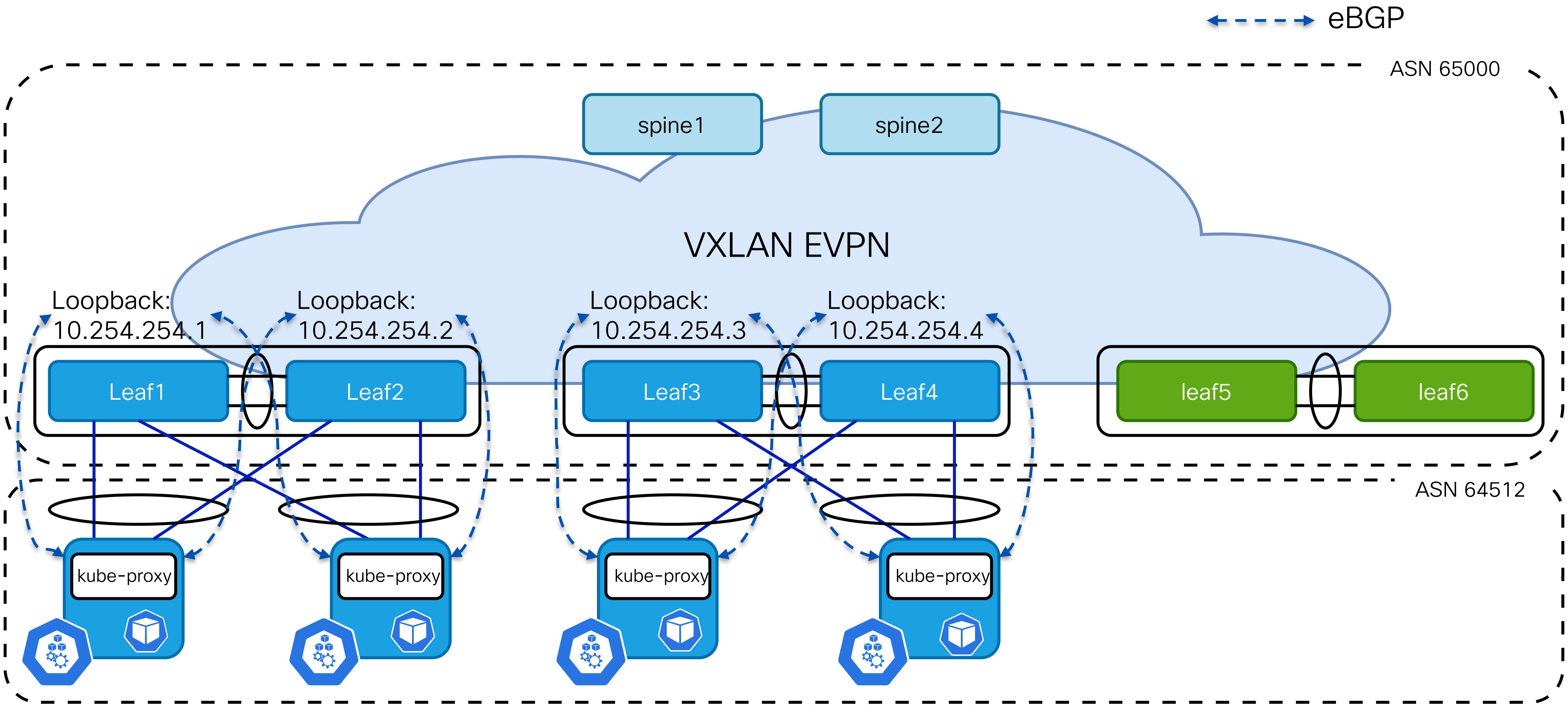
- What is Container Network Interface(CNI) Plugin
- Design the Kubernetes network on IP Fabric
- **Design the Kubernetes network on VXLAN EVPN Fabric**
- Integration with Nexus Dashboard Fabric Controller(NDFC)

Connecting K8s nodes to Leaf Switches



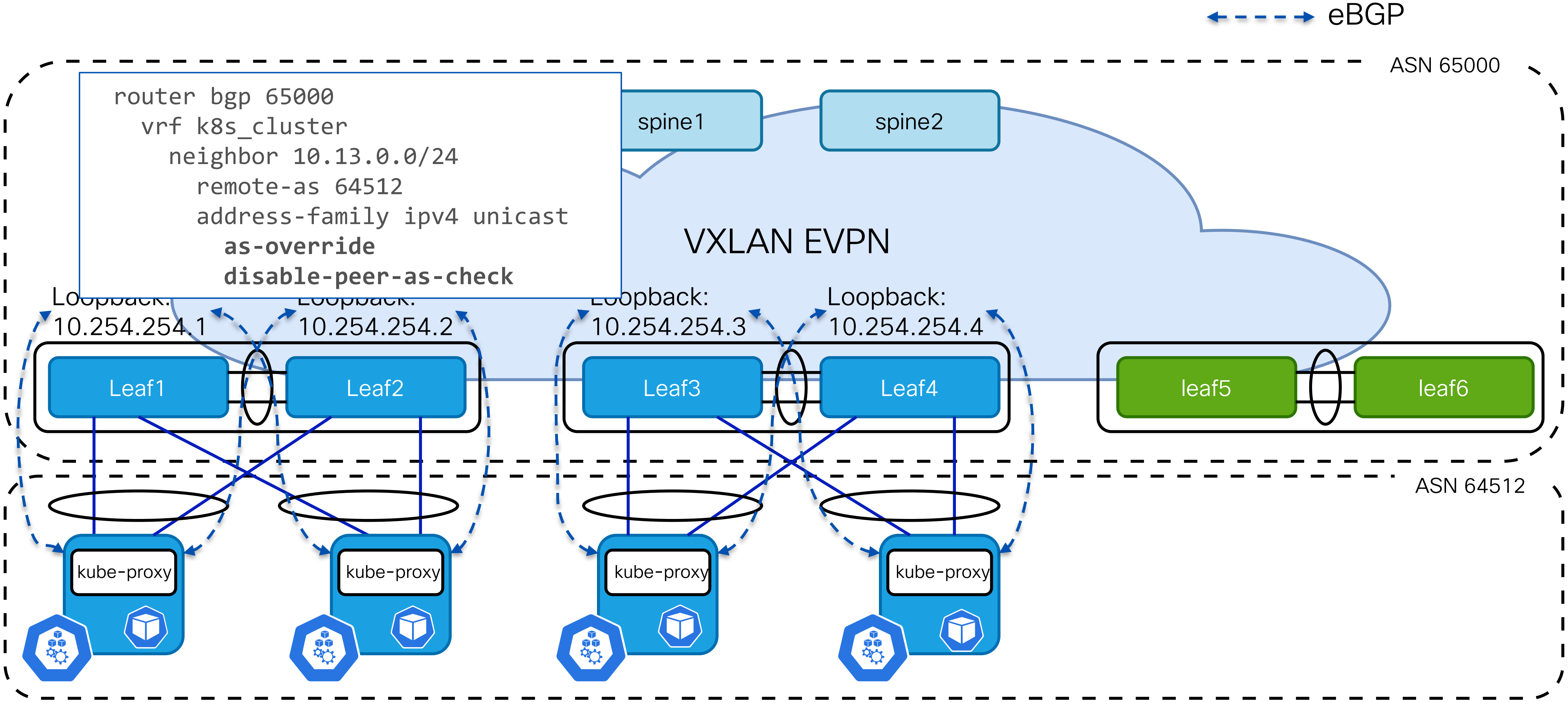
- K8s nodes connect to Leaf switches using VPC or Active-Standby
- Peering eBGP between K8s nodes and leaf switches using node IP and localized loopback addresses on each leaf switches
- Suggest peering iBGP between vPC pair in the user VRF

AS-Per-Cluster design



AS-Per-Cluster design

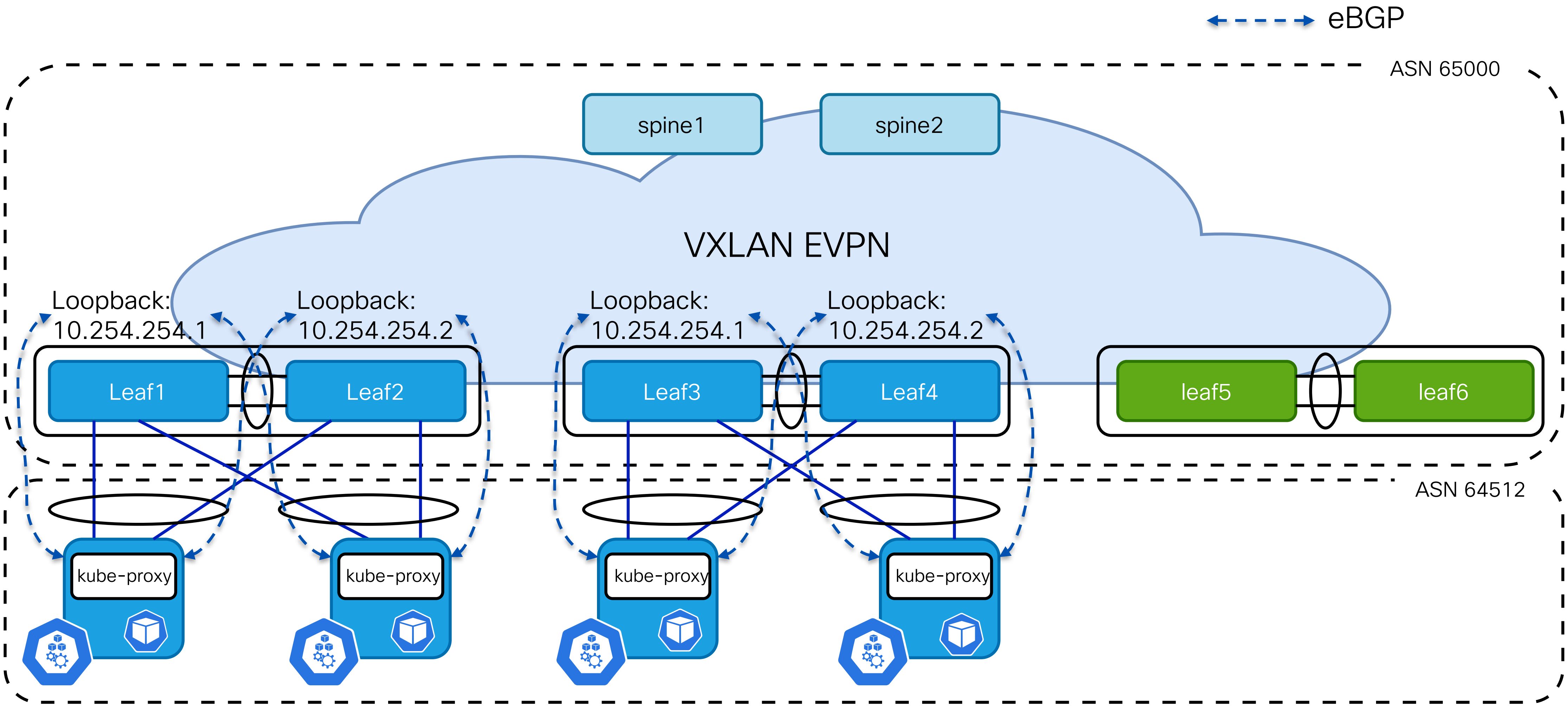
BGP tunning



```
router bgp 65000
  vrf k8s_cluster
    neighbor 10.13.0.0/24
      remote-as 64512
      address-family ipv4 unicast
        as-override
        disable-peer-as-check
```

AS-Per-Cluster design

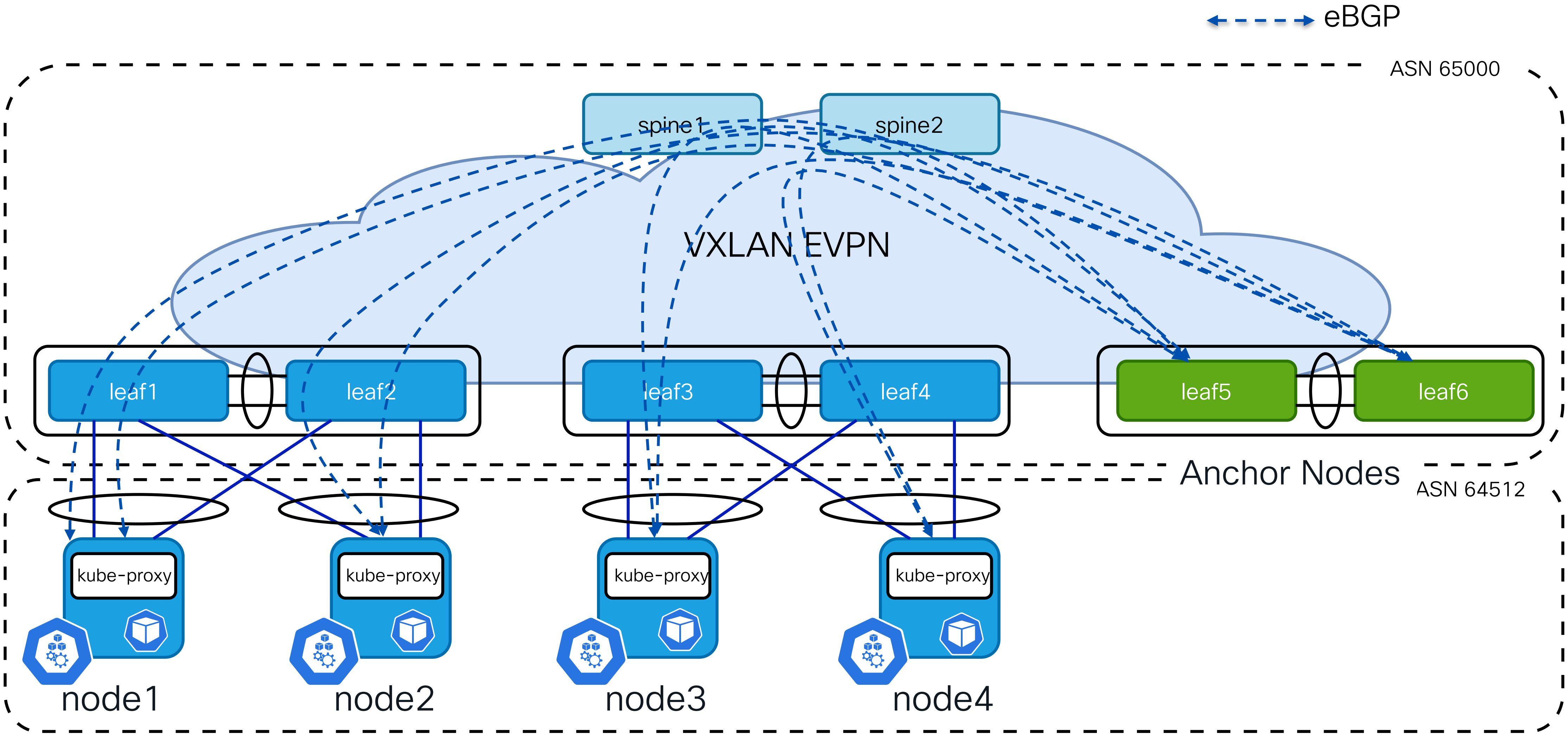
Use same loopback addresses



AS-Per-Cluster design

- Using single AS number per cluster reduces the complexity of bootstrap K8s node
- Loopback addresses are local to 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 Calico
  - `disable-peer-as-check` and `as-override` are needed on leaf switches

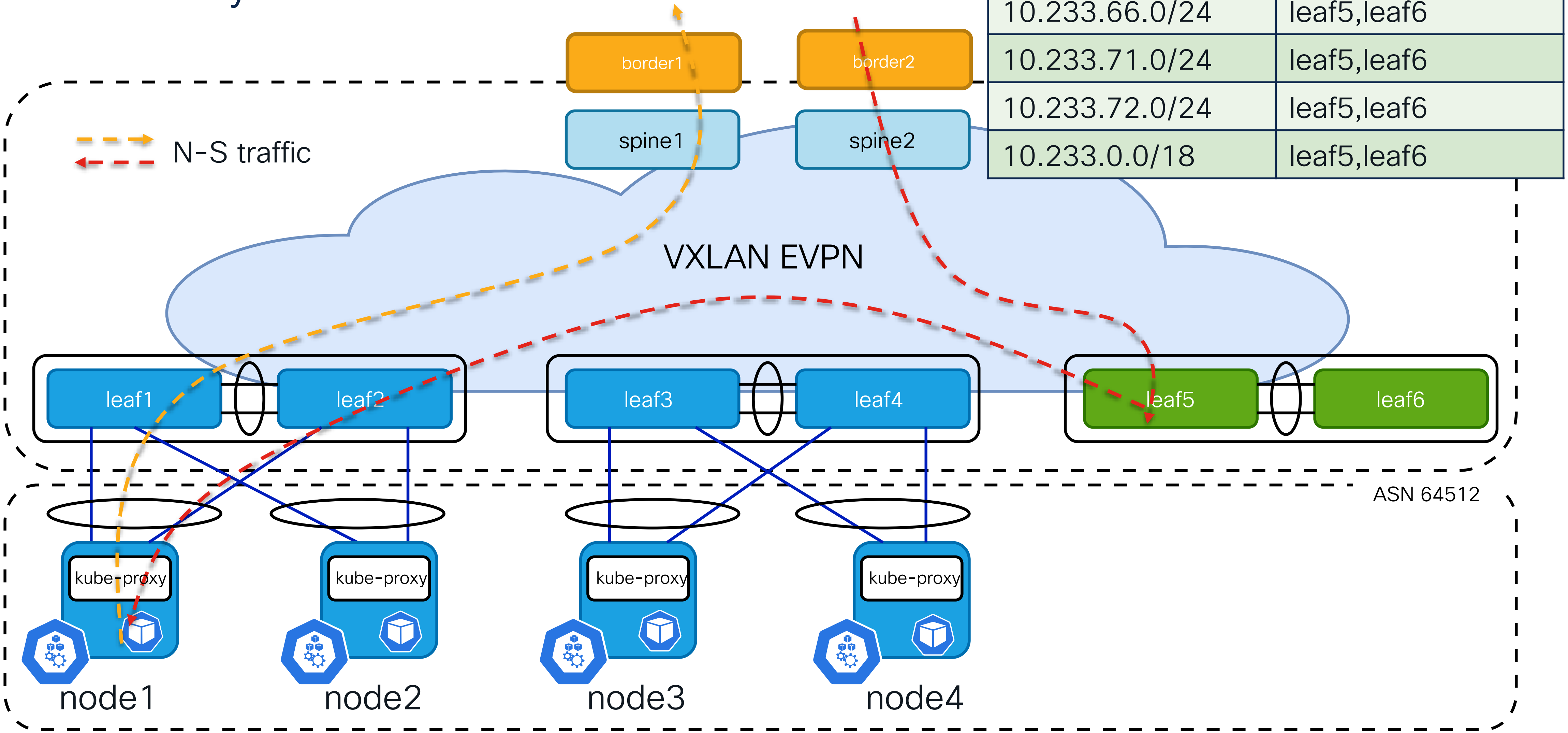
Centralized Routing Peering



Centralized Routing Peering

Problem: Asymmetric traffic

prefix	nexthop
10.233.64.0/24	leaf5,leaf6
10.233.66.0/24	leaf5,leaf6
10.233.71.0/24	leaf5,leaf6
10.233.72.0/24	leaf5,leaf6
10.233.0.0/18	leaf5,leaf6



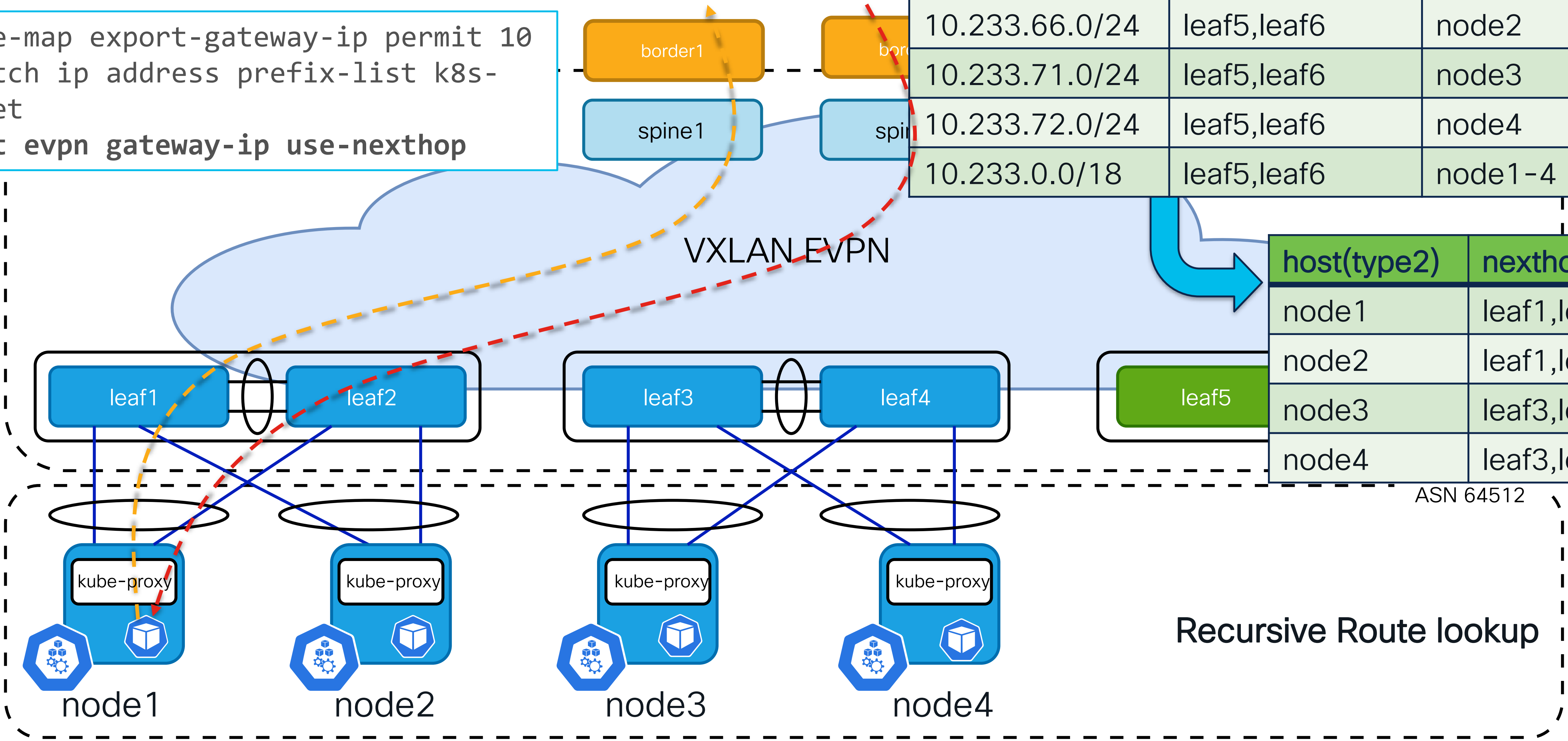
Centralized Routing Peering

Solution

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

prefix	nexthop	gateway
10.233.64.0/24	leaf5,leaf6	node1
10.233.66.0/24	leaf5,leaf6	node2
10.233.71.0/24	leaf5,leaf6	node3
10.233.72.0/24	leaf5,leaf6	node4
10.233.0.0/18	leaf5,leaf6	node1-4

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

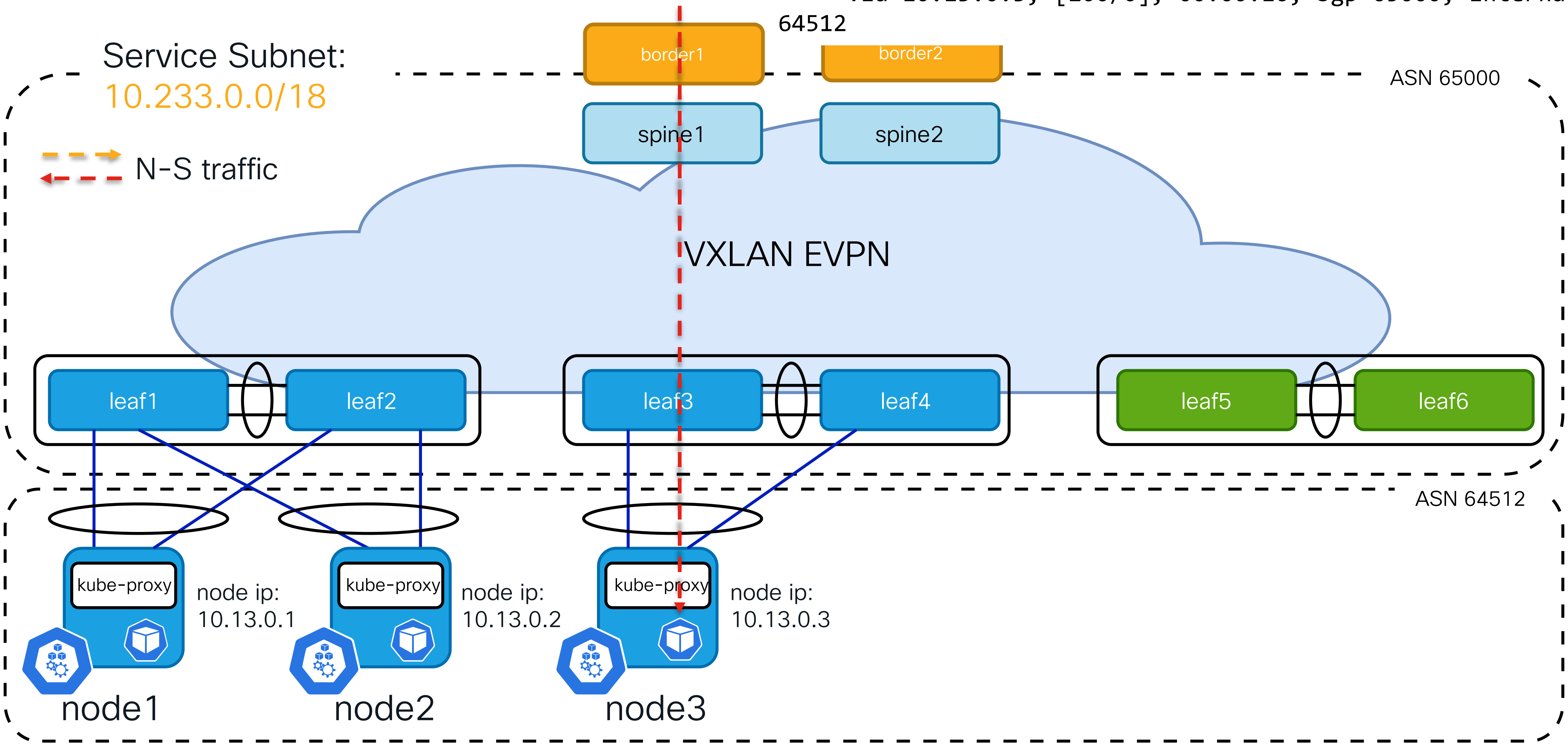


Recursive Route lookup

Centralized Routing Peering

Service Traffic

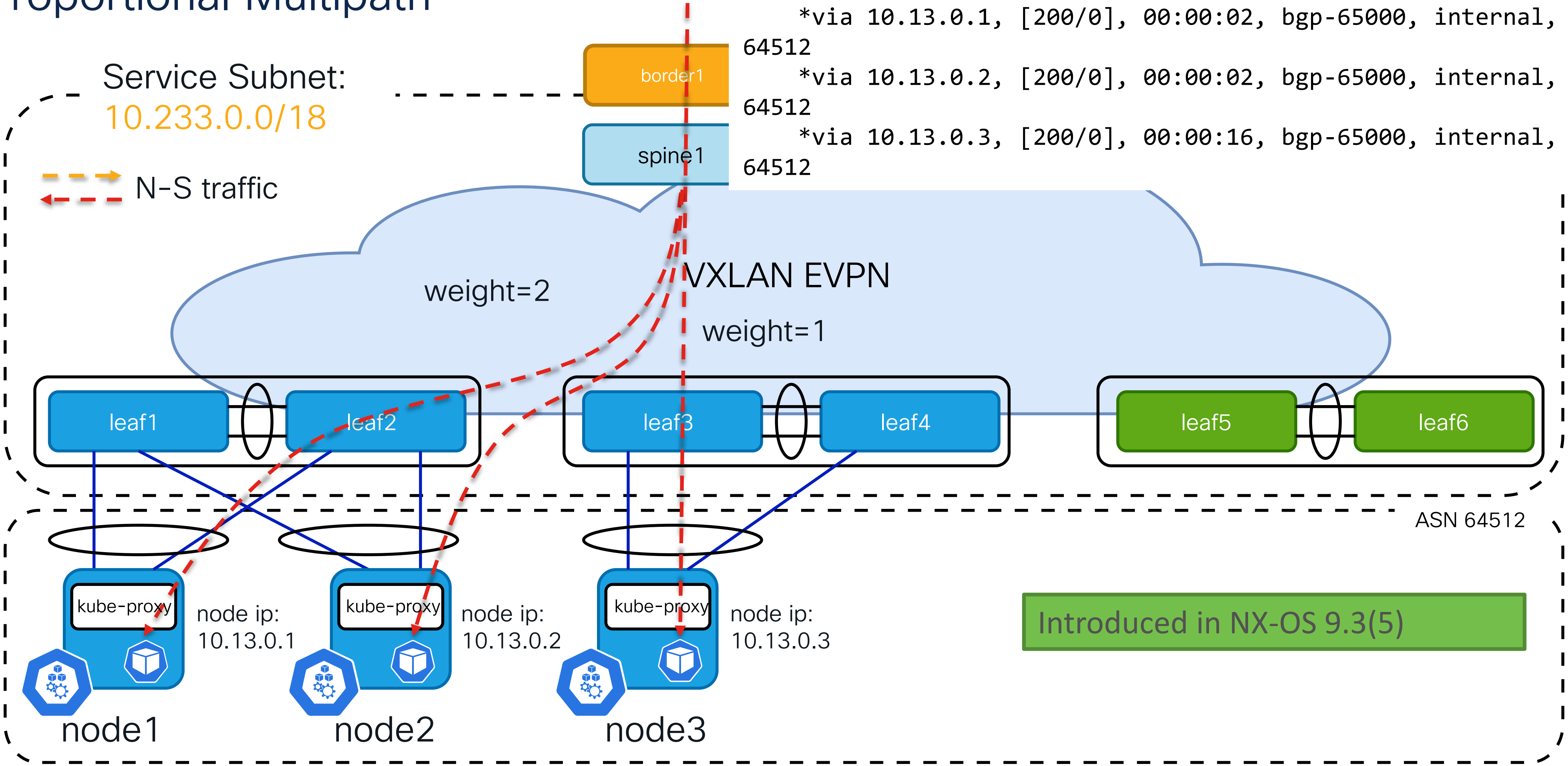
```
10.233.0.0/18, ubest/mbest: 4/0
    *via 10.13.0.3, [200/0], 00:00:16, bgp-65000, internal, tag
64512
```



Centralized Routing Peering

Proportional Multipath

```
10.233.0.0/18, ubest/mbest: 4/0
    *via 10.13.0.1, [200/0], 00:00:02, bgp-65000, internal, tag
64512
    *via 10.13.0.2, [200/0], 00:00:02, bgp-65000, internal, tag
64512
    *via 10.13.0.3, [200/0], 00:00:16, bgp-65000, internal, tag
64512
```

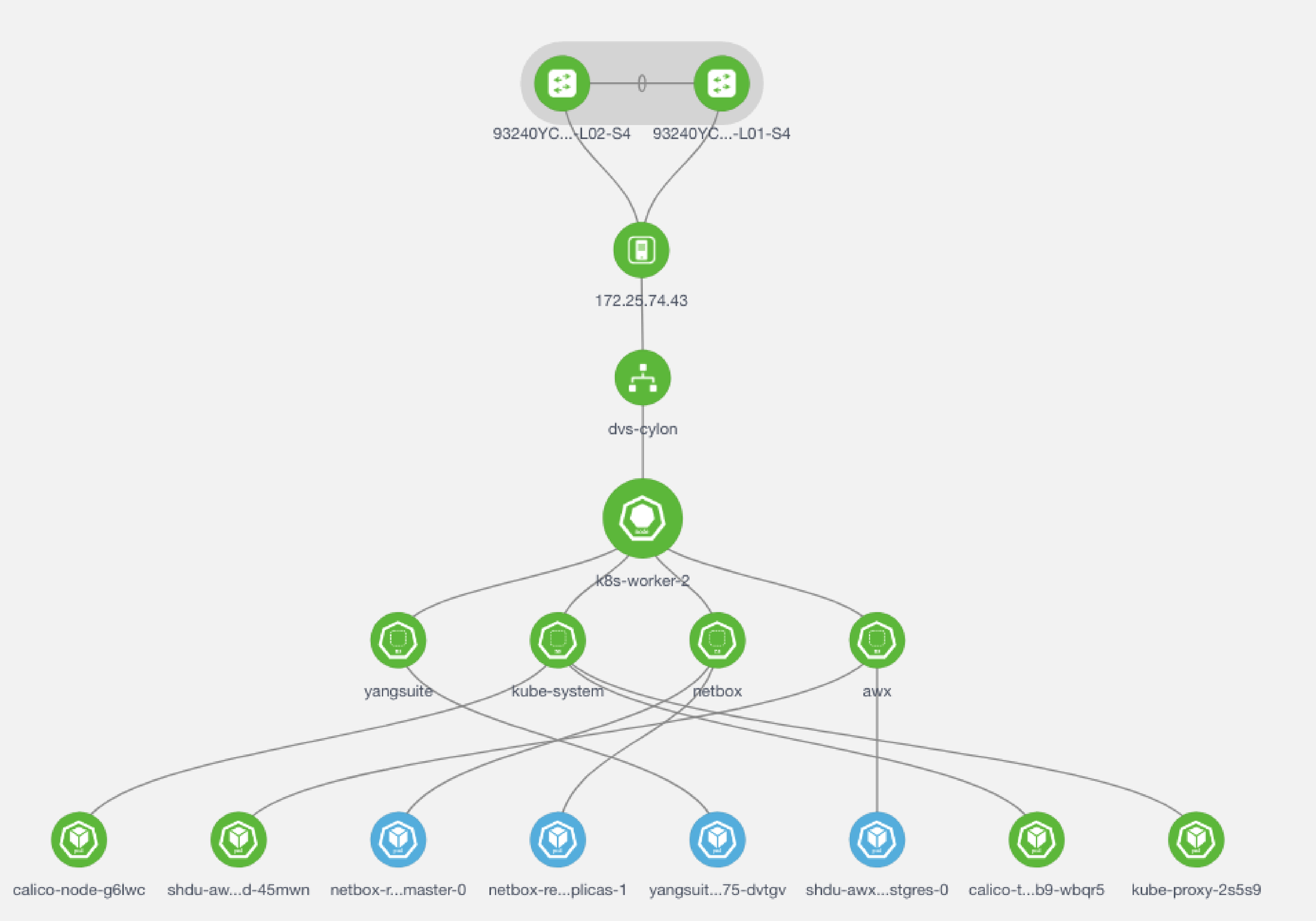


Introduced in NX-OS 9.3(5)

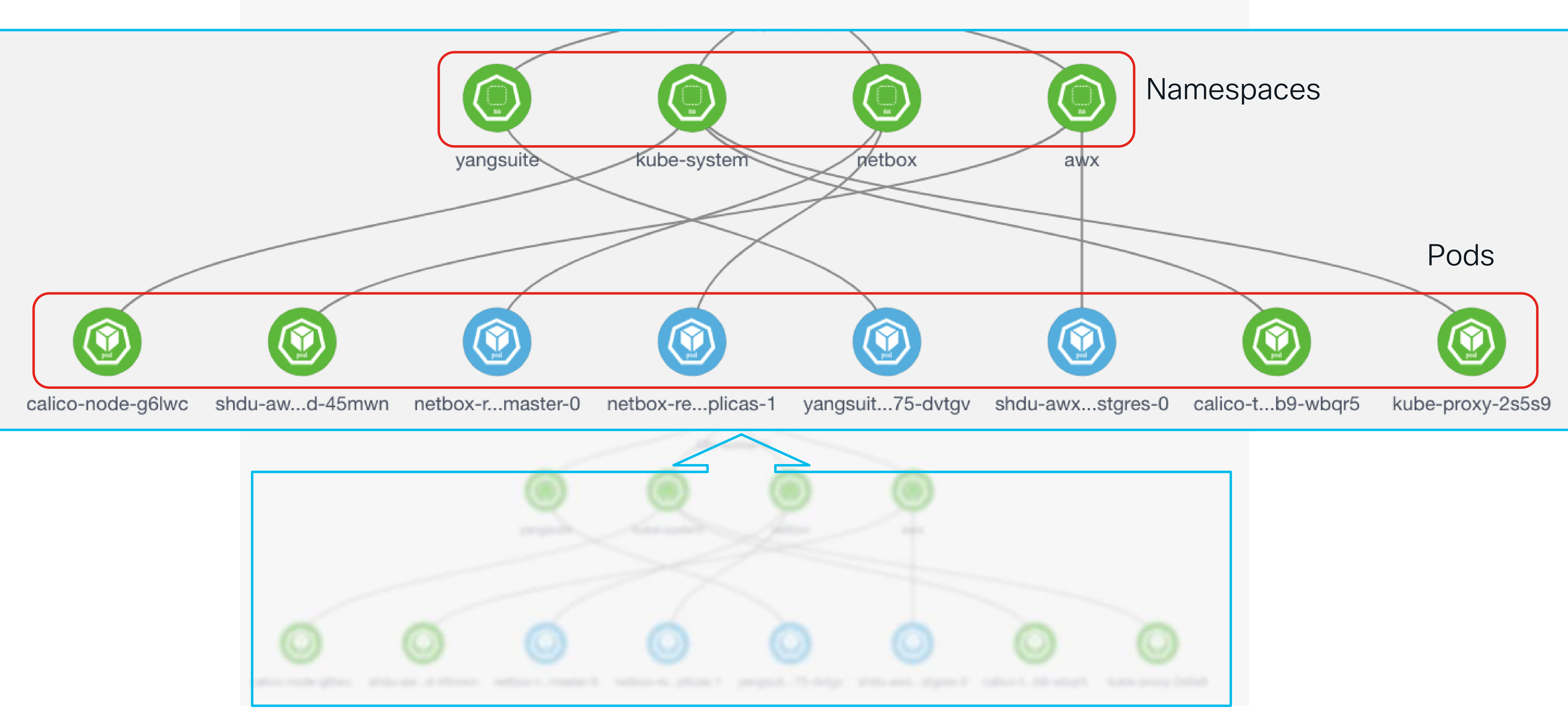
Agenda

- What is Container Network Interface(CNI) Plugin
- Design the Kubernetes network on IP Fabric
- Design the Kubernetes network on VXLAN EVPN Fabric
- **Integration with Nexus Dashboard Fabric Controller(NDFC)**

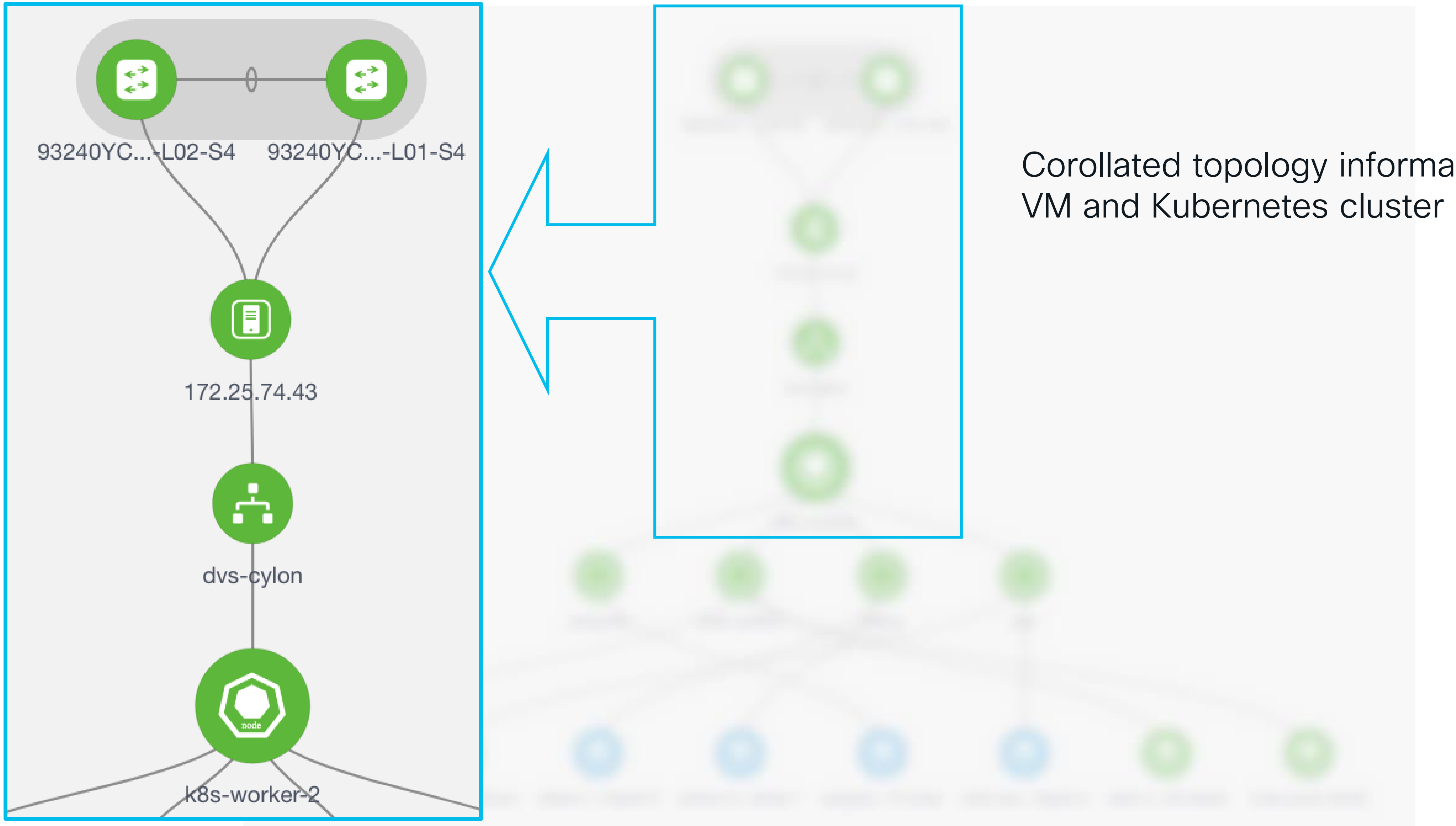
Kubernetes Visualization with NDFC



Kubernetes Visualization with NDFC



Kubernetes Visualization with NDFC



Corollated topology information with VM and Kubernetes cluster

Summary

- Greenfield Calico network does not require L2 extension
- The best practice is peering BGP neighborship with local switches
- Centralized Route Peering can simplify the configuration of Calico
  - But does require additional consideration to optimize traffic
- 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

Fill out your session surveys!

Attendees who fill out a minimum of four session surveys and the overall event survey will get **Cisco Live-branded socks** (while supplies last)!

Attendees will also earn 100 points in the **Cisco Live Challenge** for every survey completed.

These points help you get on the leaderboard and increase your chances of winning daily and grand prizes

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

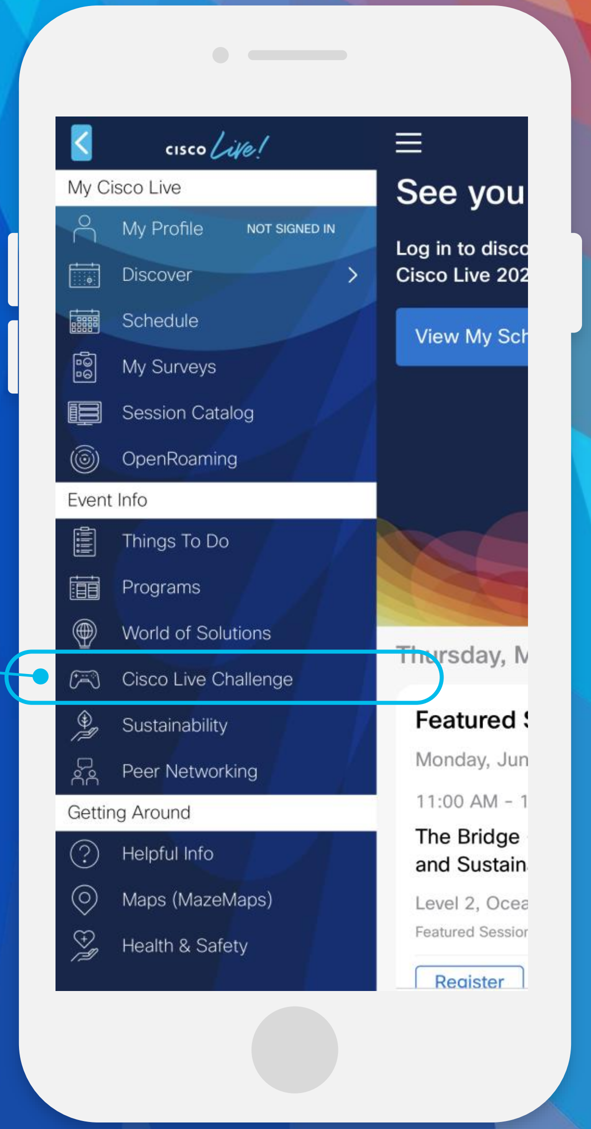
Cisco Live **Challenge**

Gamify your Cisco Live experience!

Get points for attending this session!

How:

1. Open the Cisco Events App.
2. Click on 'Cisco Live Challenge’ in the side menu.
3. Click on View Your Badges at the top.
4. Click the + at the bottom of the screen and scan the QR code:



cisco Live!

Let’s go

#CiscoLive