

CISCO *Live!*

ALL IN

#CiscoLive



The bridge to possible

# Kubernetes (K8s) Infrastructure Connectivity

Network Designs for the Modern Data Center

Shangxin Du, Technical Marketing Engineer, Cloud Networking  
BRKDCN-2410



#CiscoLive

# Cisco Webex App

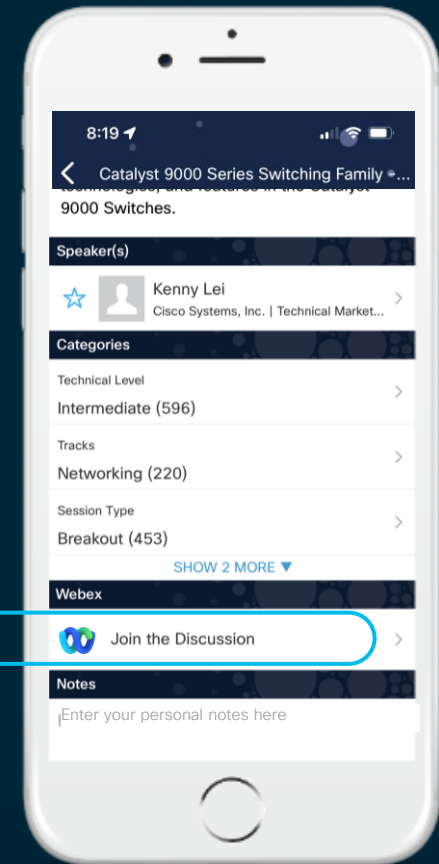
## Questions?

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 17, 2022.



<https://ciscolive.ciscoevents.com/ciscolivebot/#BRKDCN-2410>



# Agenda

- What is Calico
- Design Calico network on IP Fabric
- Design Calico network on VXLAN EVPN Fabric

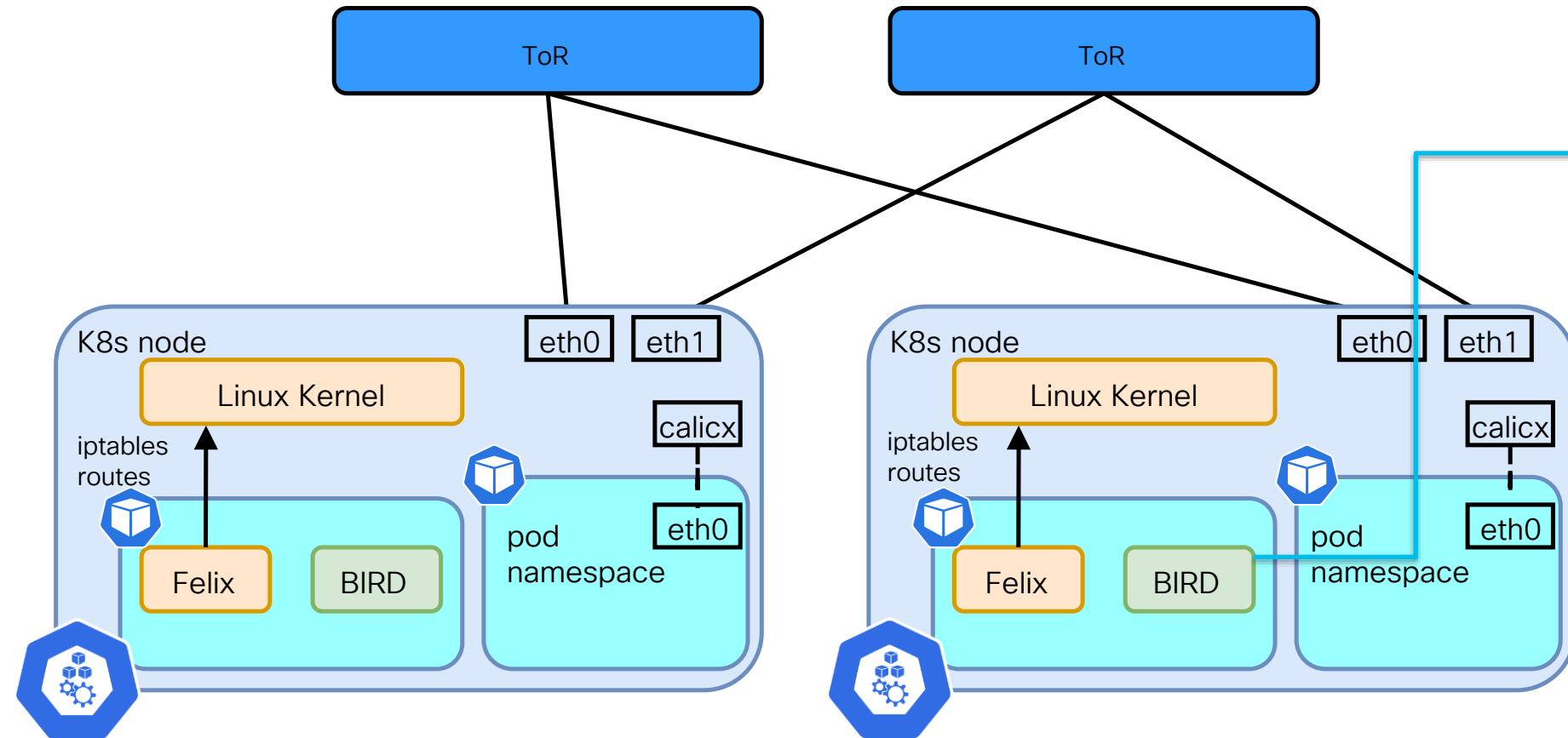
# What is Calico

A Kubernetes CNI plugin



# Calico

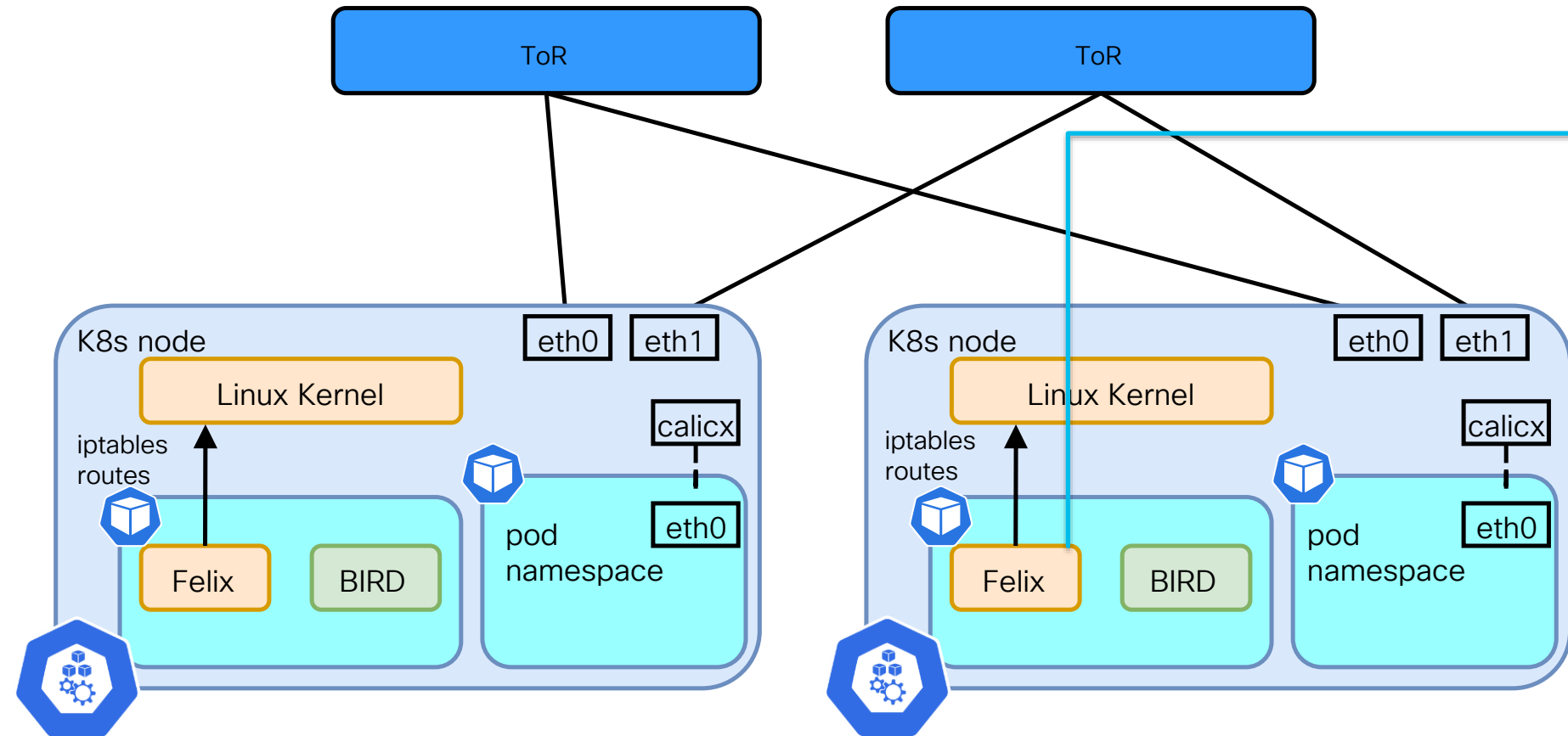
A CNI plugin of Kubernetes



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.

# Calico

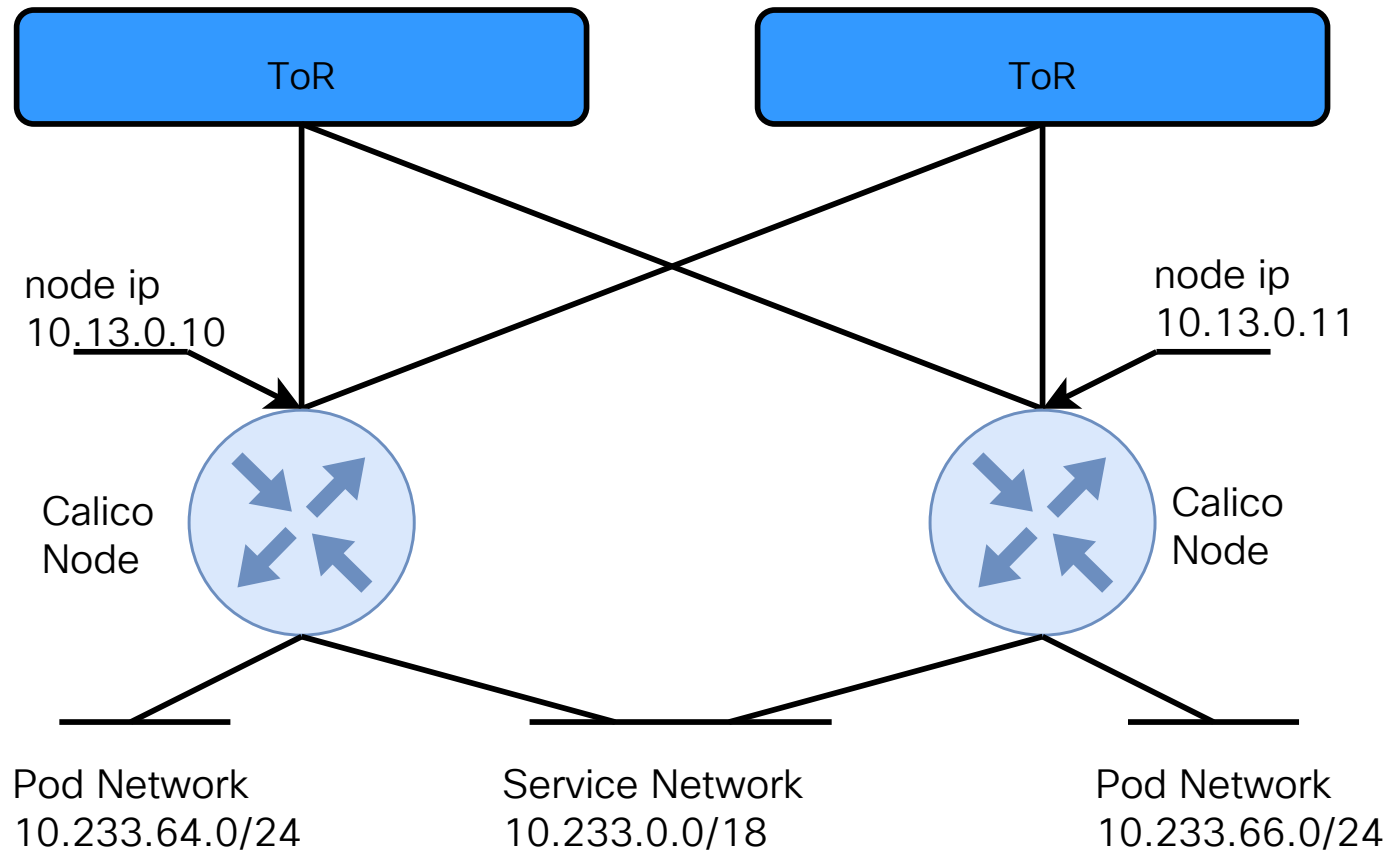
A CNI plugin of Kubernetes



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

# Calico

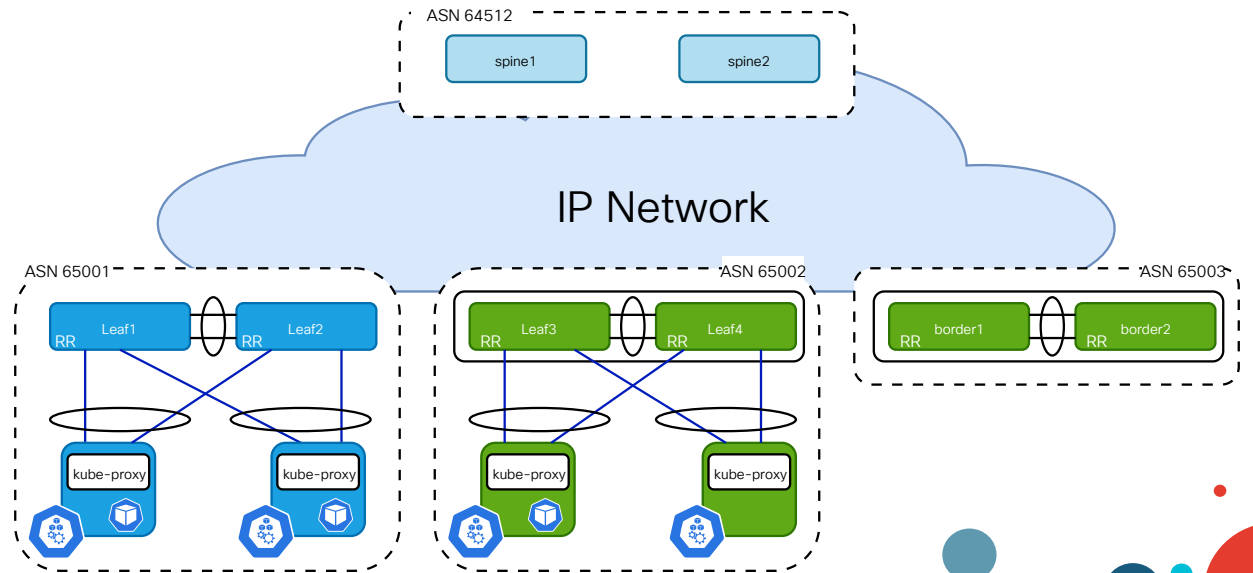
## Simplified



- 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.

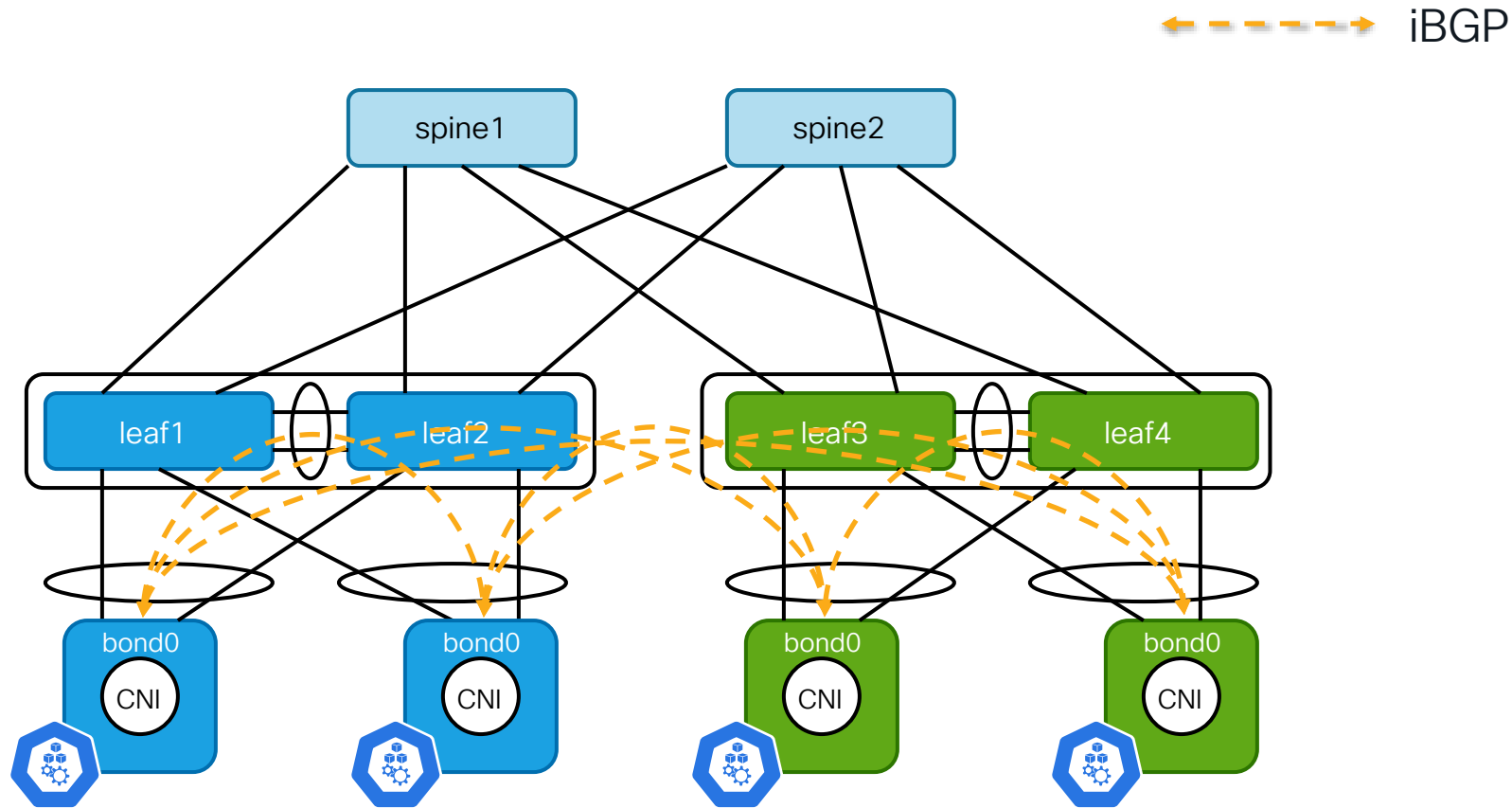


# Design Calico on IP Fabric



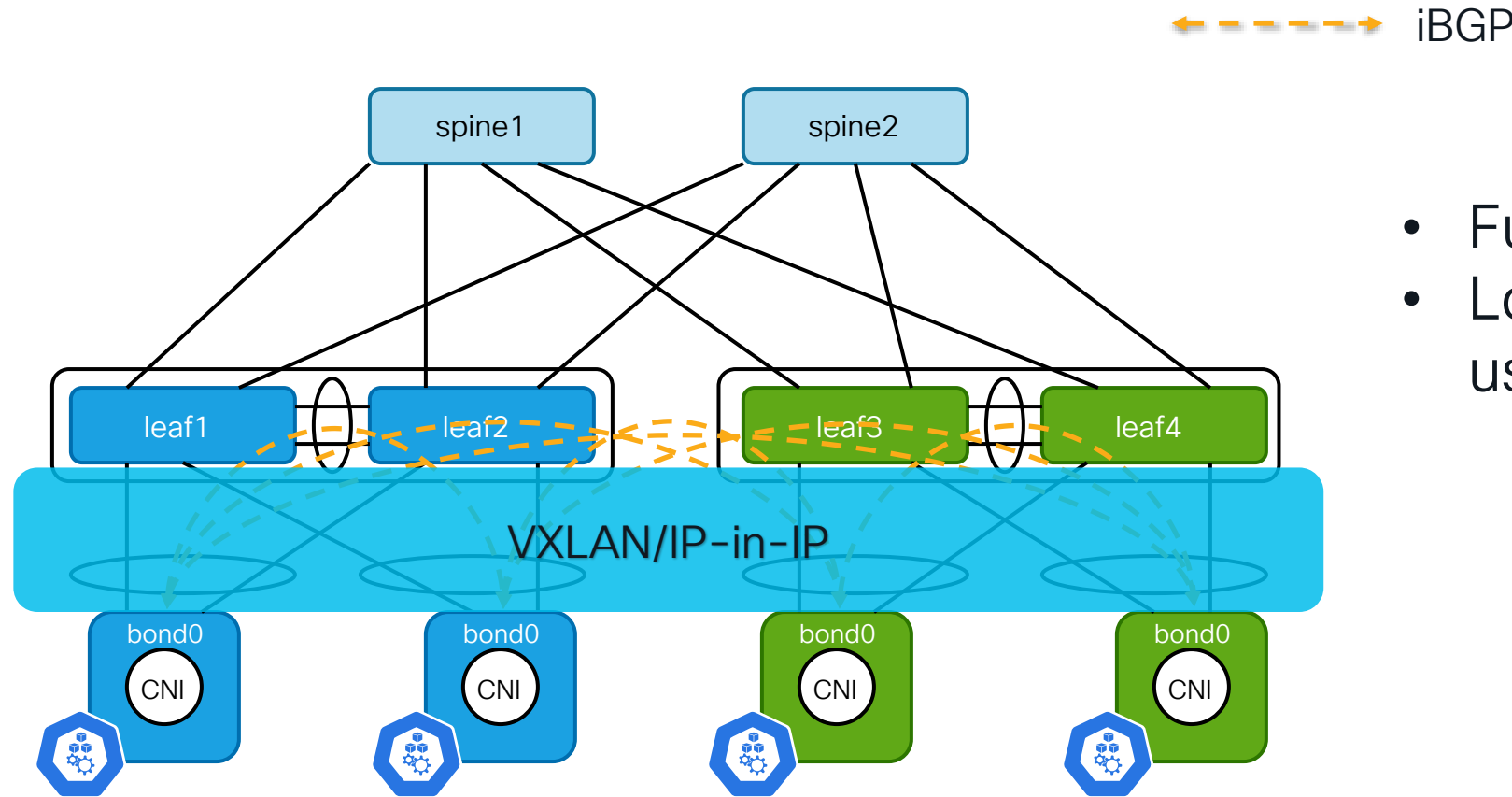
# Network Architecture

Full mesh



# Network Architecture

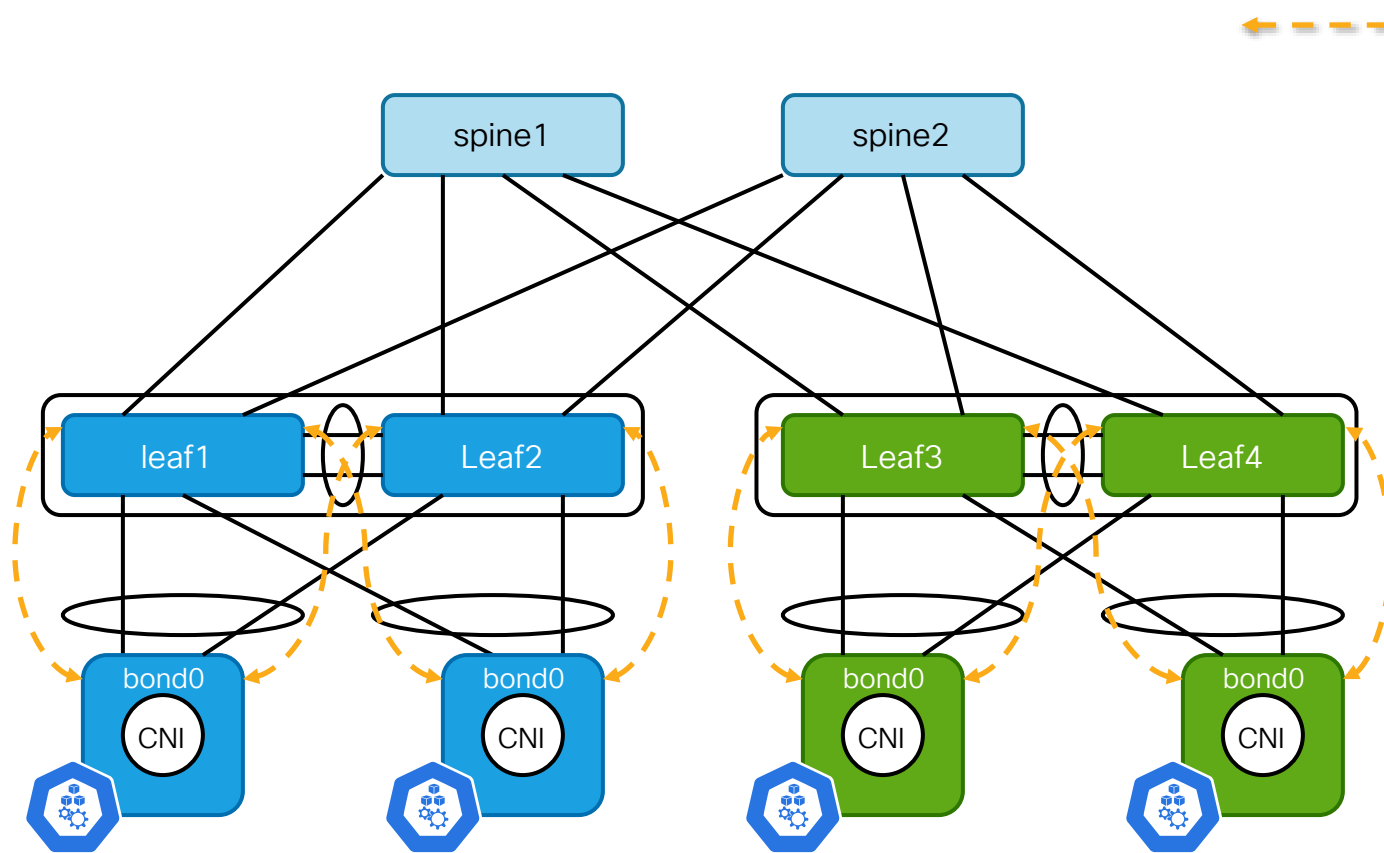
Full mesh data plane



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

# Network Architecture

## Peer with Switch



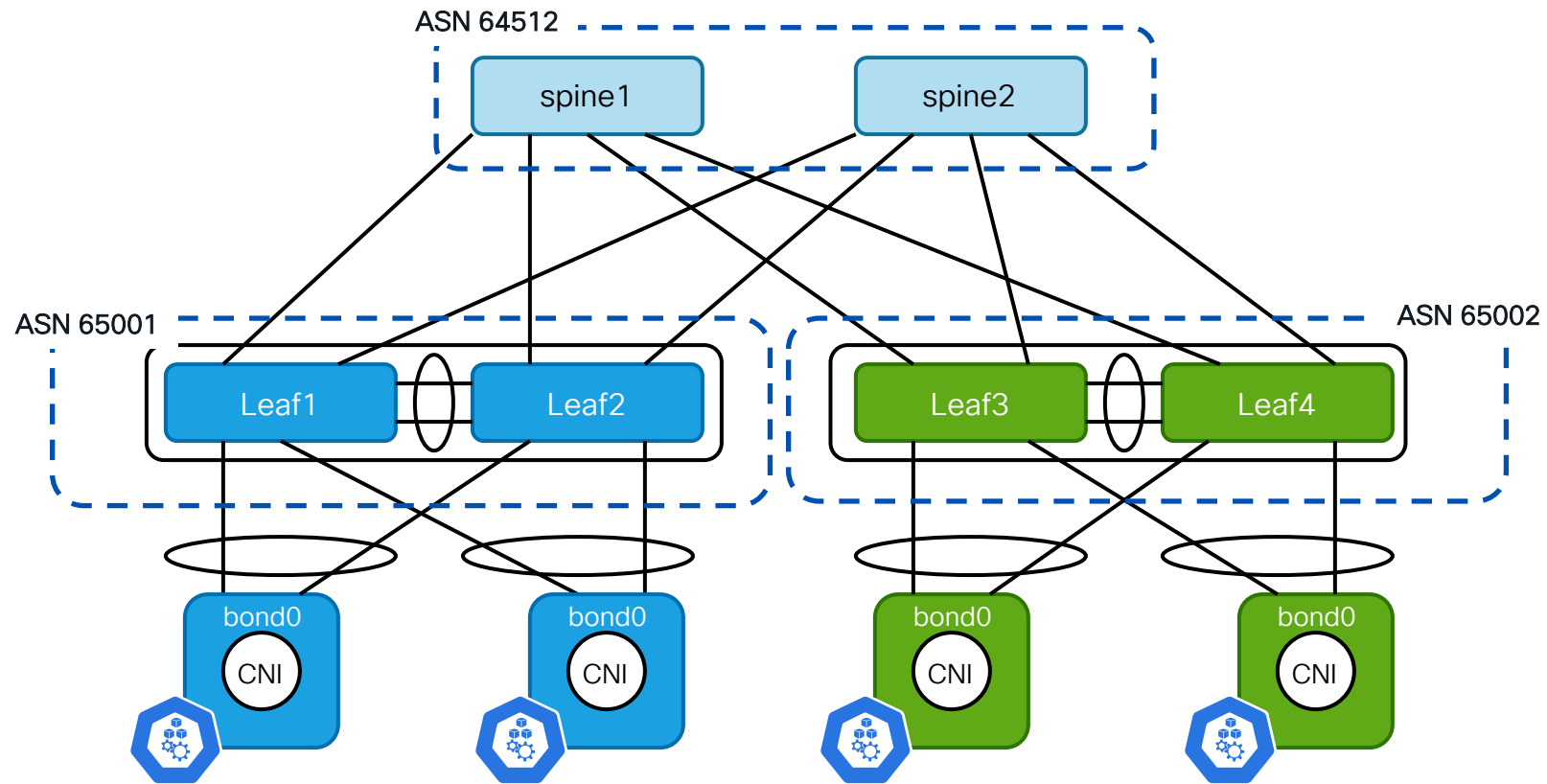
← - - - - - → iBGP

- Scalable approach, the leaf switches become Route-Reflector or Route-Server
- 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
```

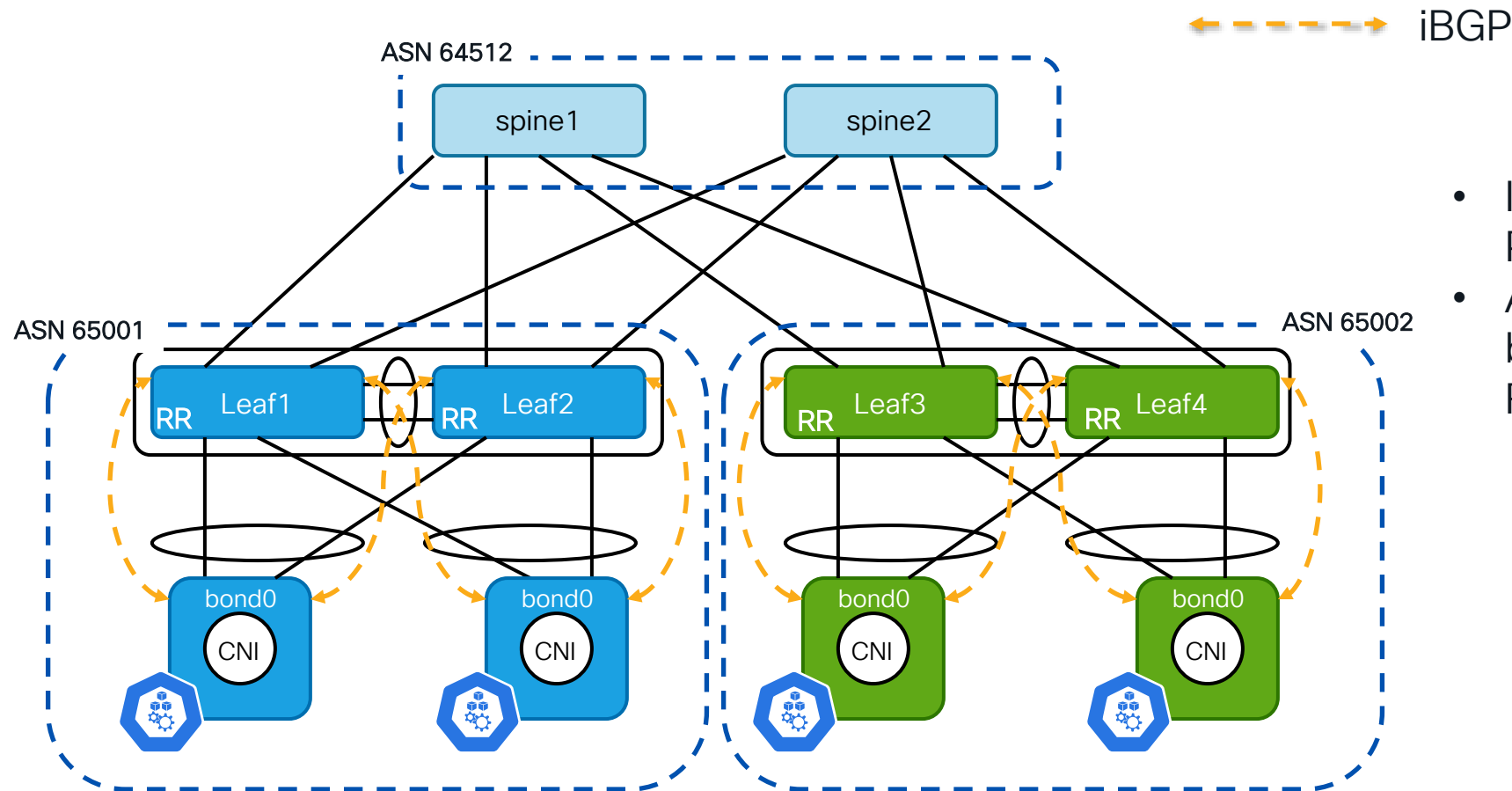
# Network Architecture

## Deploy Over IP Fabric



# Network Architecture

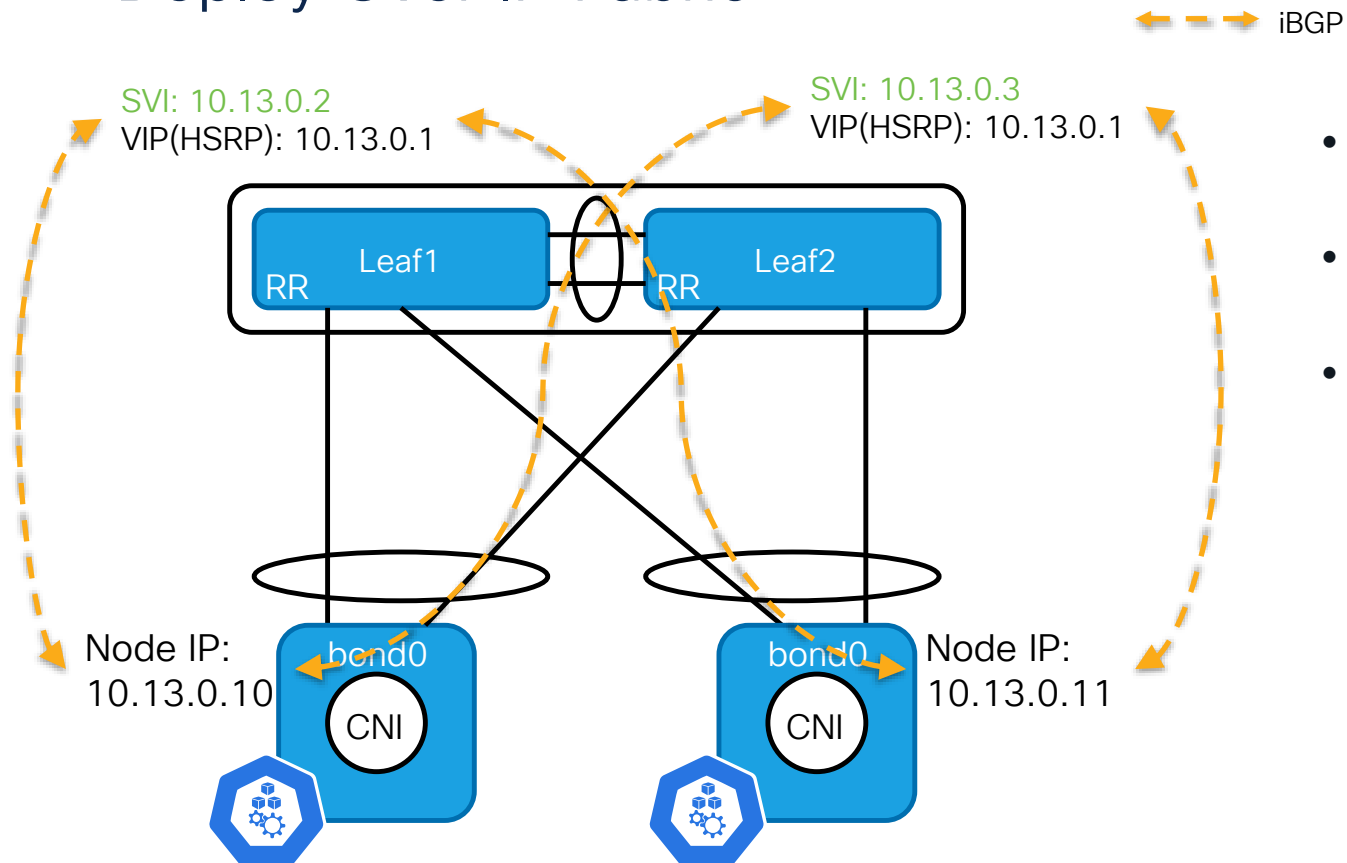
## Deploy Over IP Fabric



- 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

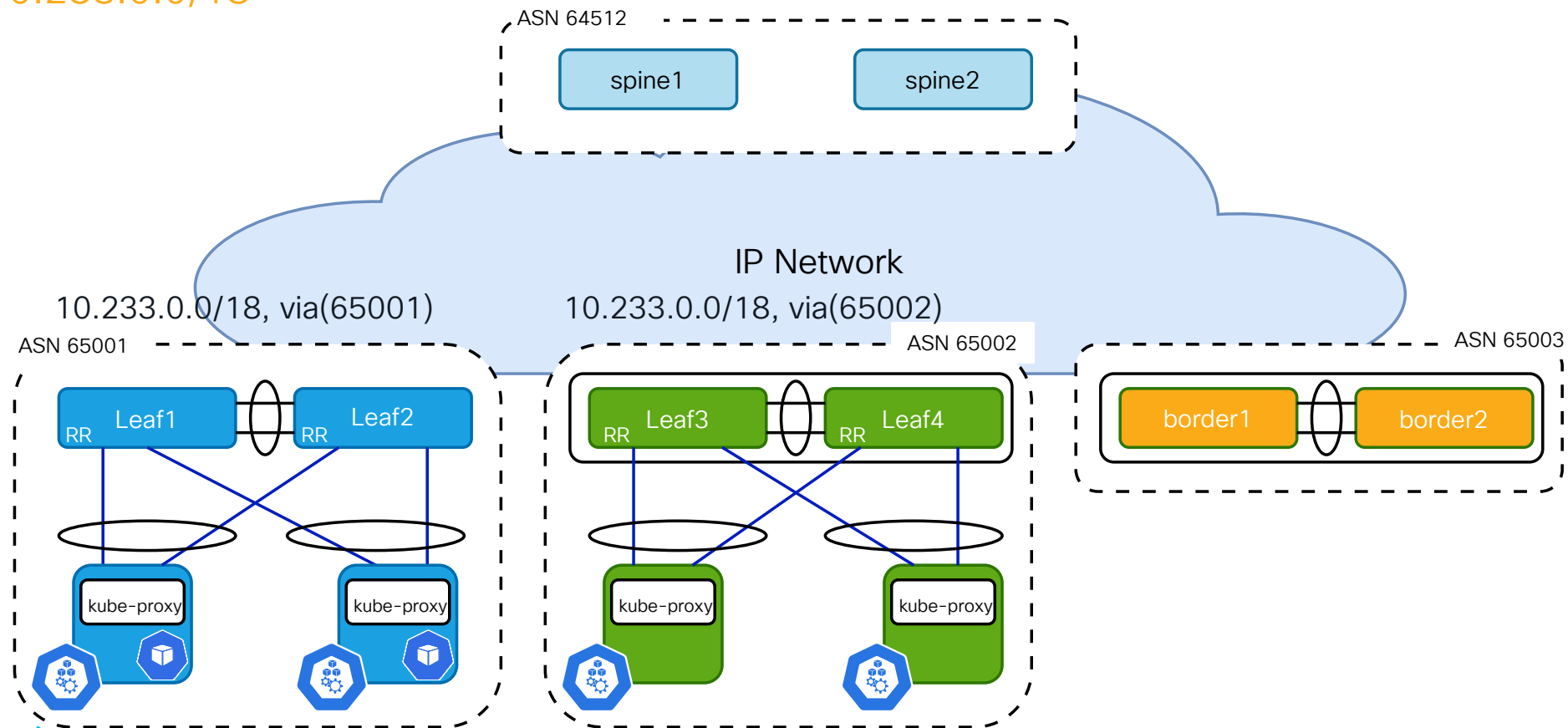
10.233.0.0/18, ubest/mbest: 4/0

\*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

Service Subnet:

10.233.0.0/18





# Deploy over IP Fabric

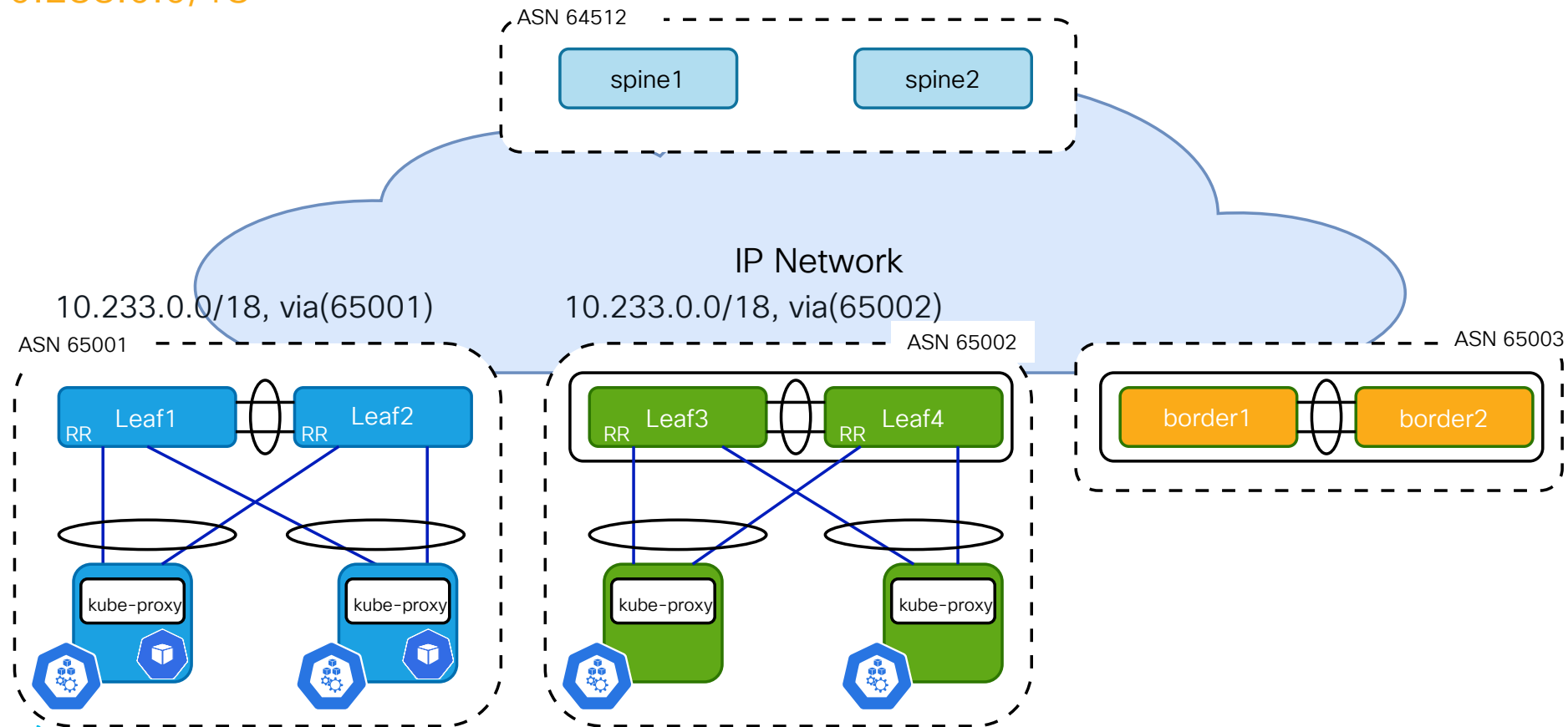
## Service Traffic

Service Subnet:  
10.233.0.0/18

10.233.0.0/18, ubest/mbest: 4/0

```
*via 10.4.0.21, [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
```

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



# Deploy over IP Fabric

## Sub-optimal service traffic

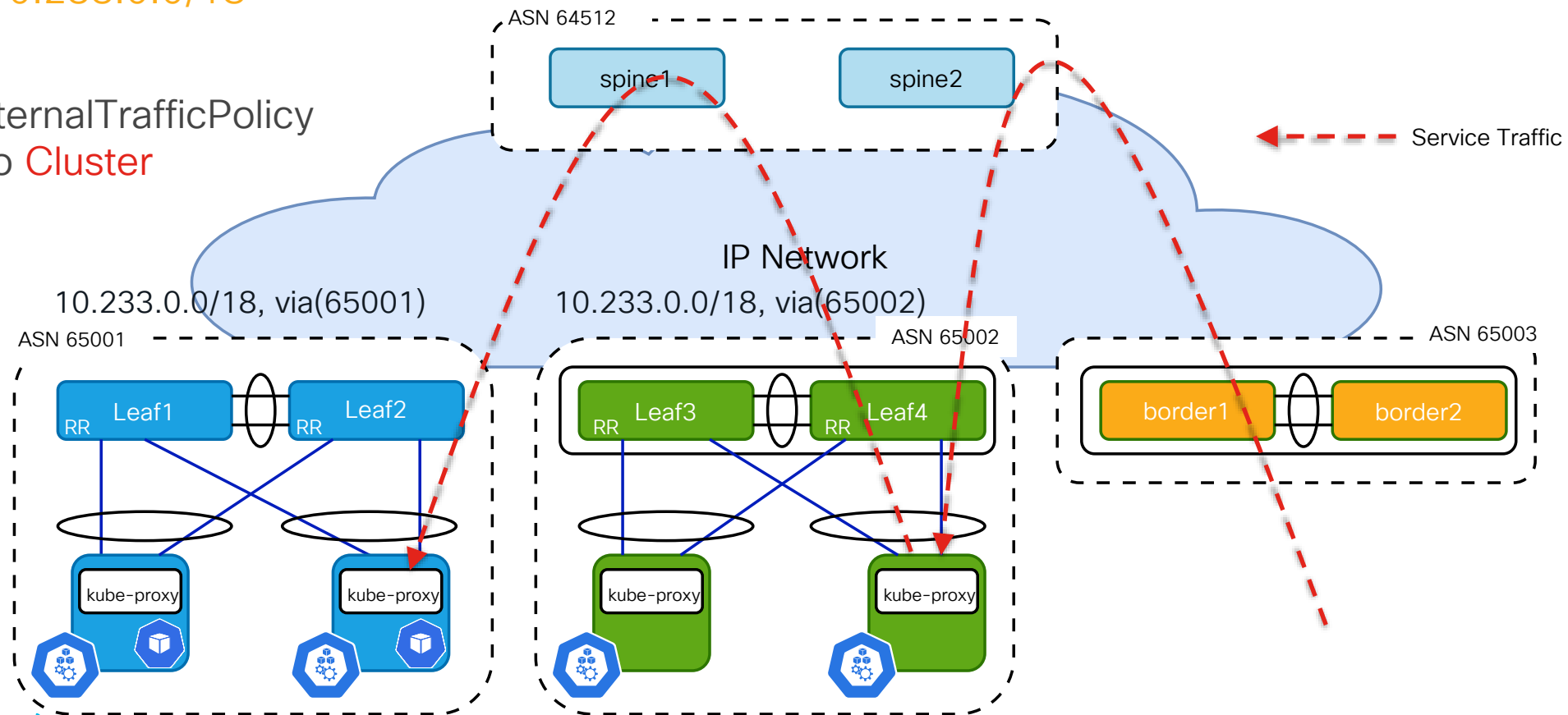
Service Subnet:  
10.233.0.0/18

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
```

router bgp 64512  
bestpath as-path multipath-relax

K8s externalTrafficPolicy  
is set to **Cluster**



# Deploy over IP Fabric

## Avoid Second Hop of Service Traffic

Service Subnet:

10.233.0.0/18

Service ip:

10.233.63.214/32

K8s externalTrafficPolicy is set to  
**Local**

Service Type is set to  
**NodePort/LoadBalancer**

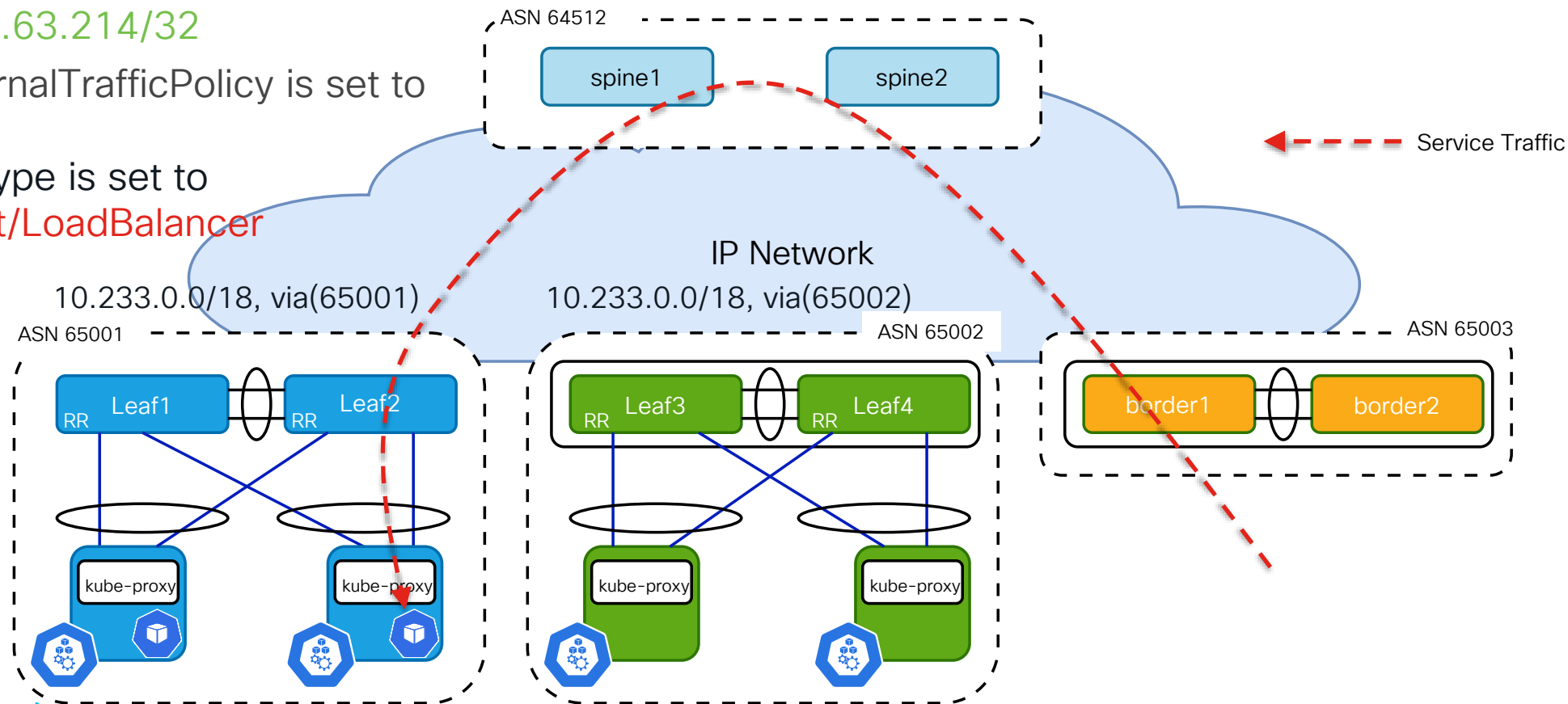
10.233.63.214/32, ubest/mbest: 2/0

\*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

router bgp 64512

bestpath as-path multipath-relax

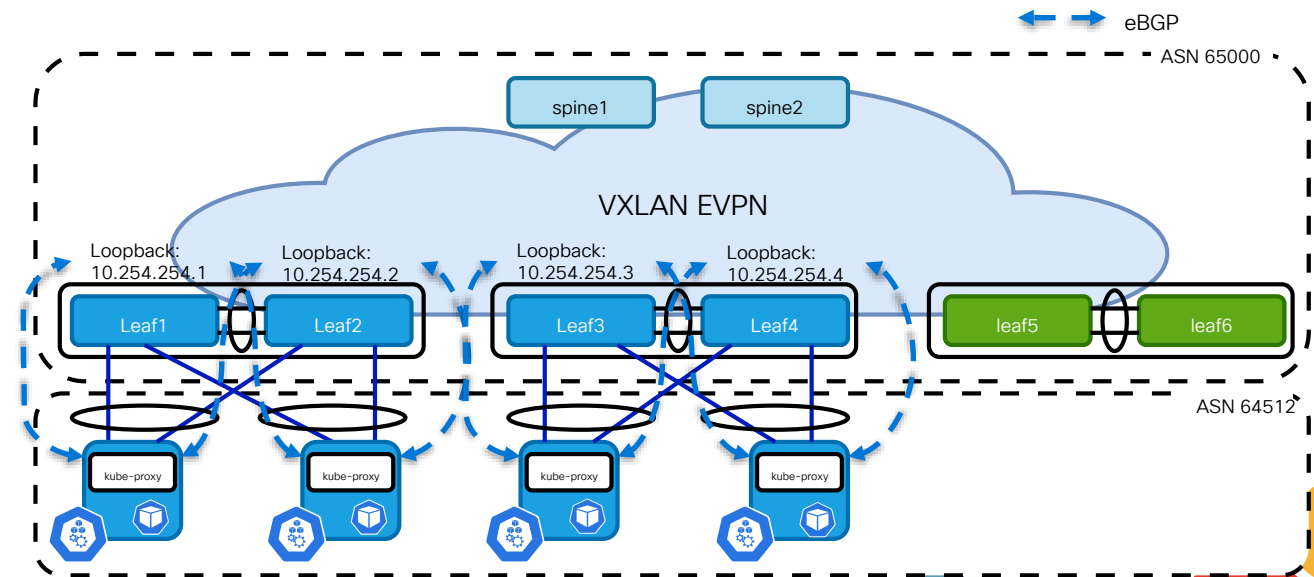


# externalTrafficPolicy

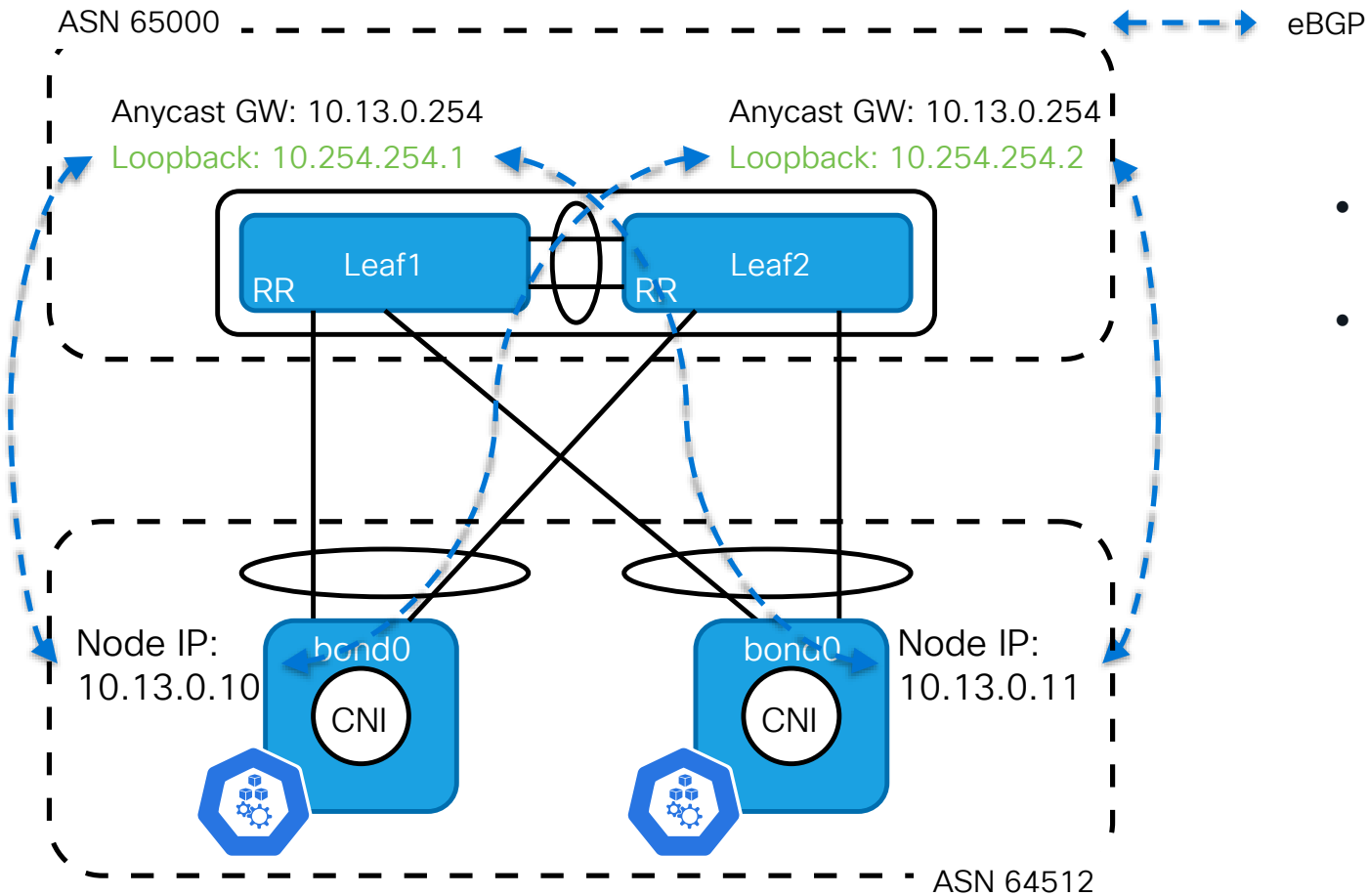
## Pros and Cons

- `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**

# Design Calico on VXLAN EVPN Fabric

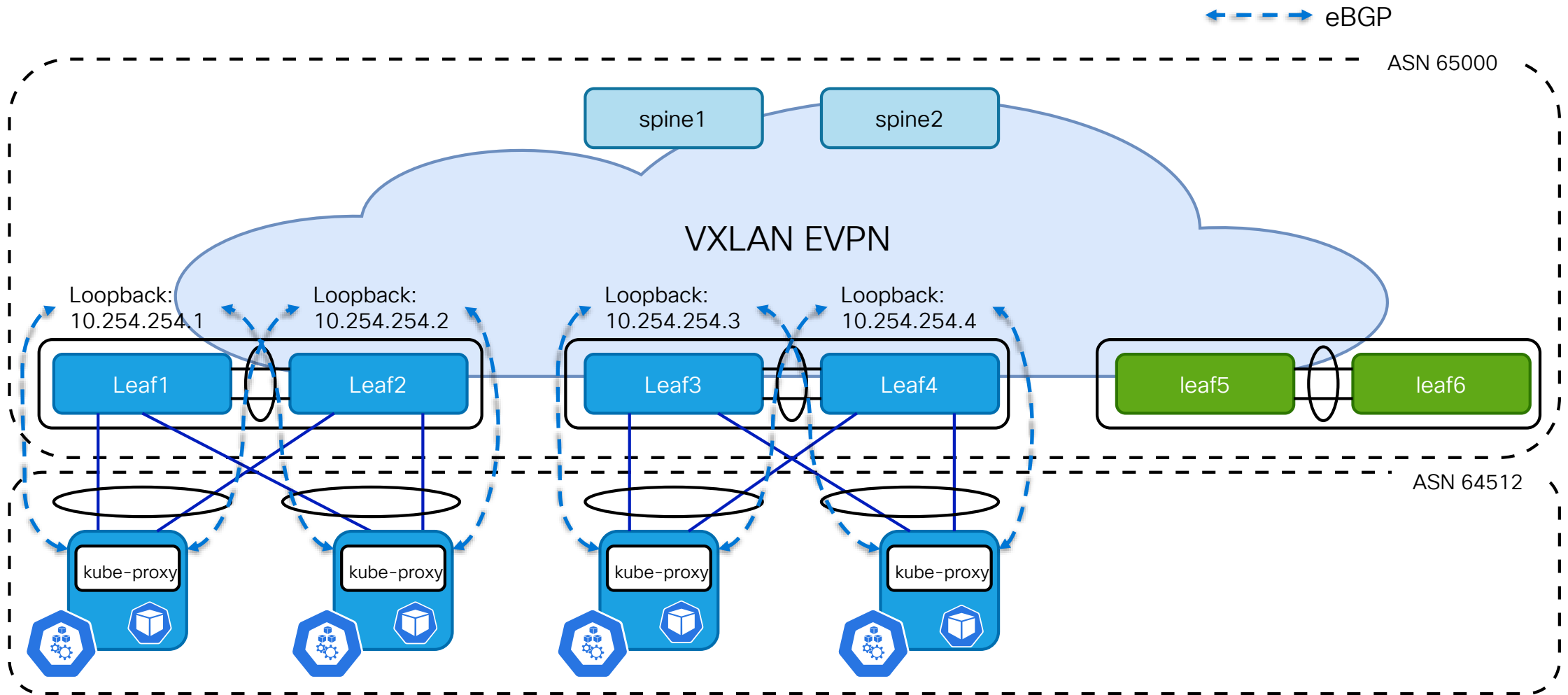


# Connecting K8s nodes to Leaf Switches



- 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

# As-Per-Cluster design

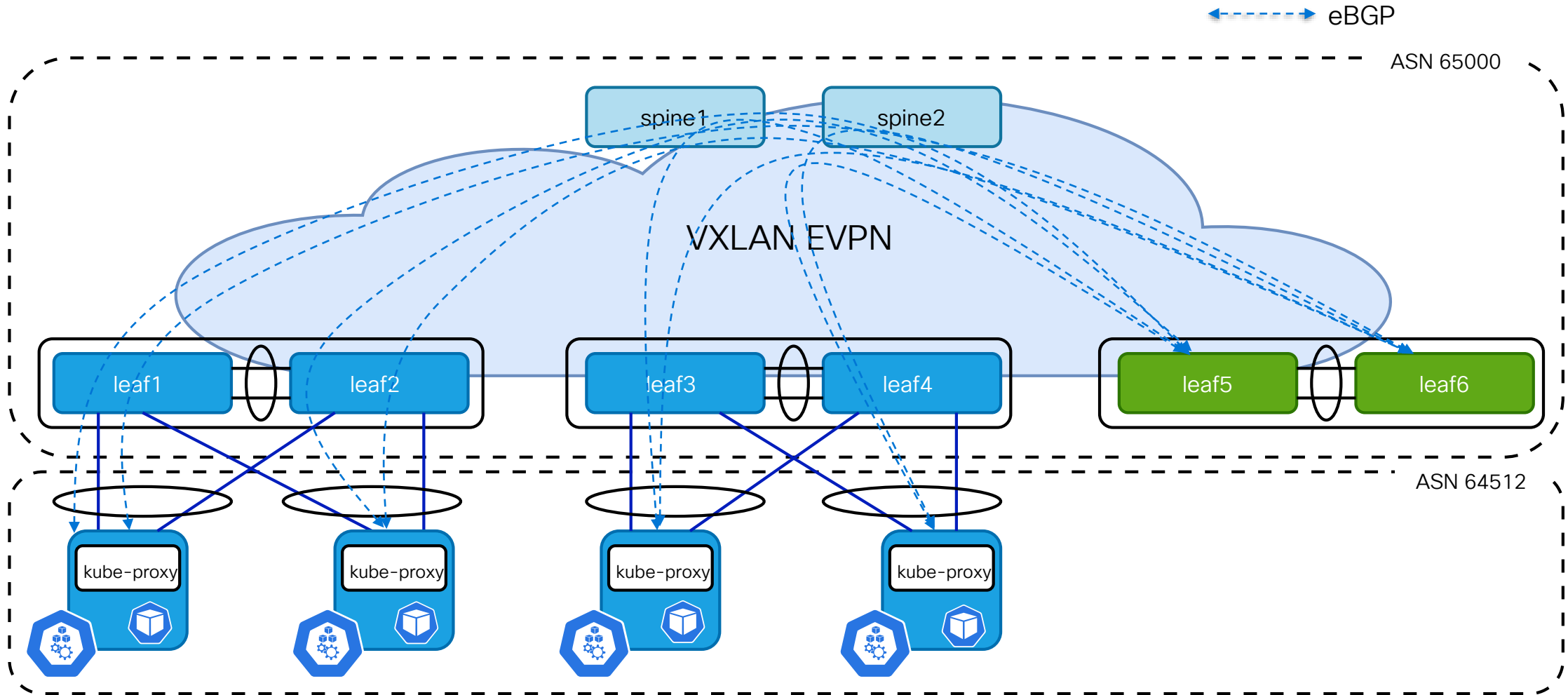


# As-Per-Cluster design

- 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
  - 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

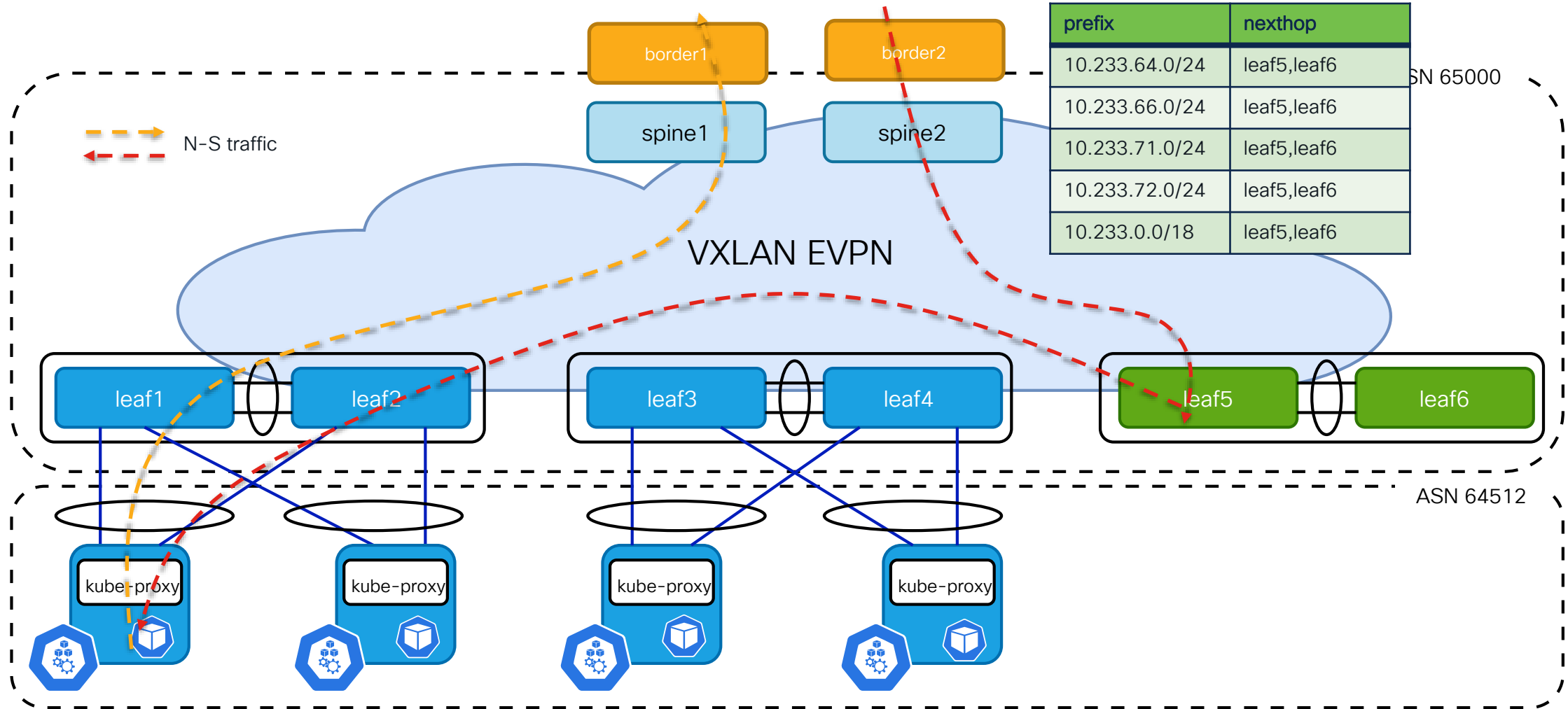


# Centralized Routing Peering



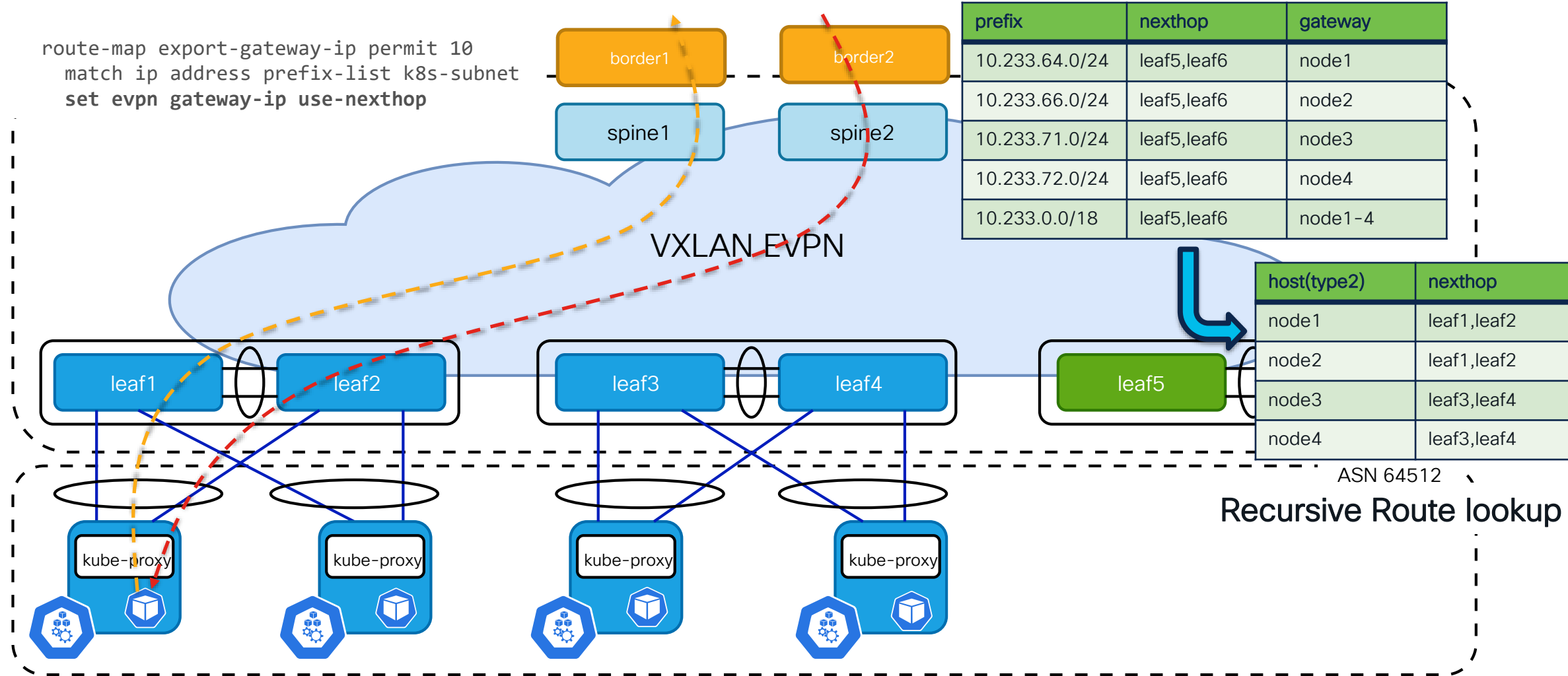
# Centralized Routing Peering

## Problem



# Centralized Routing Peering Solution

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

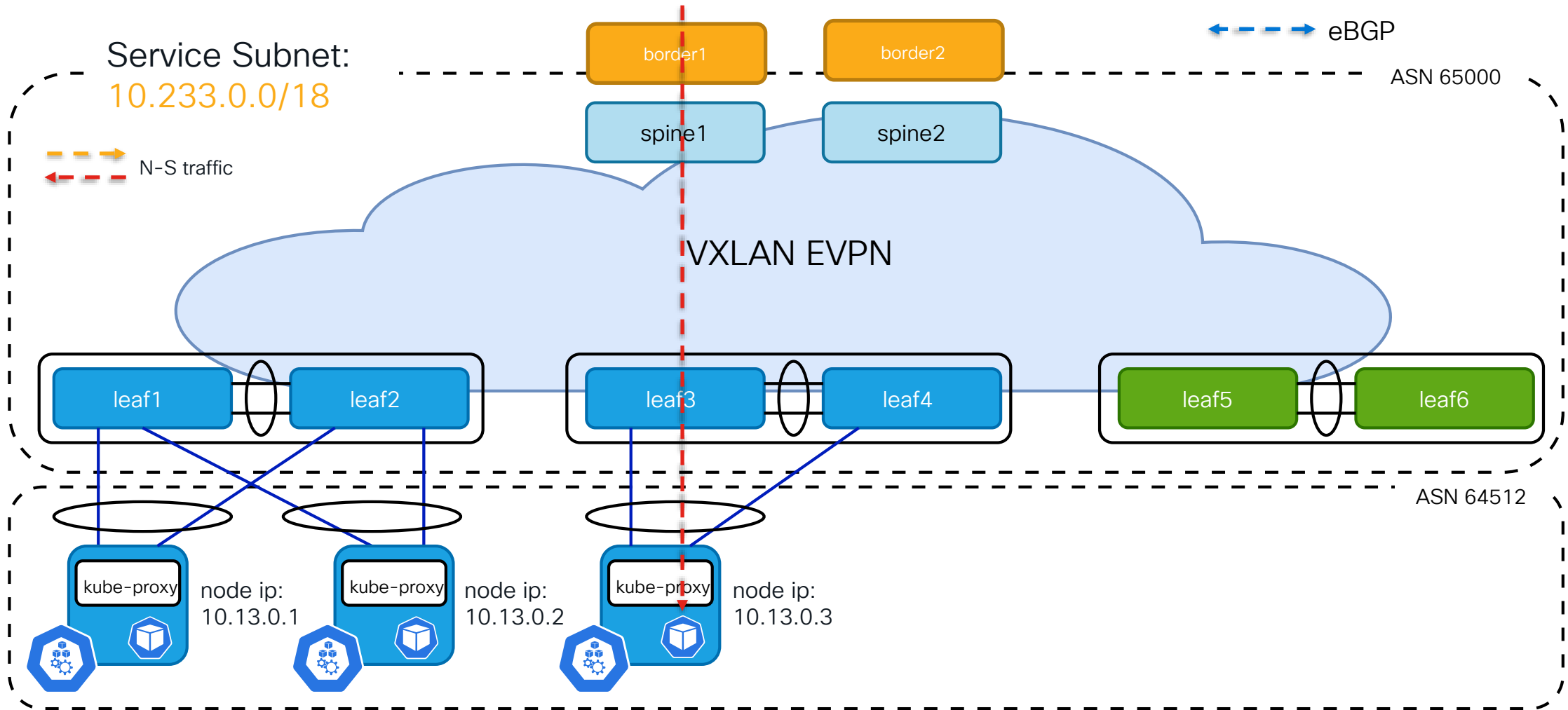


# Centralized Routing Peering

## 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



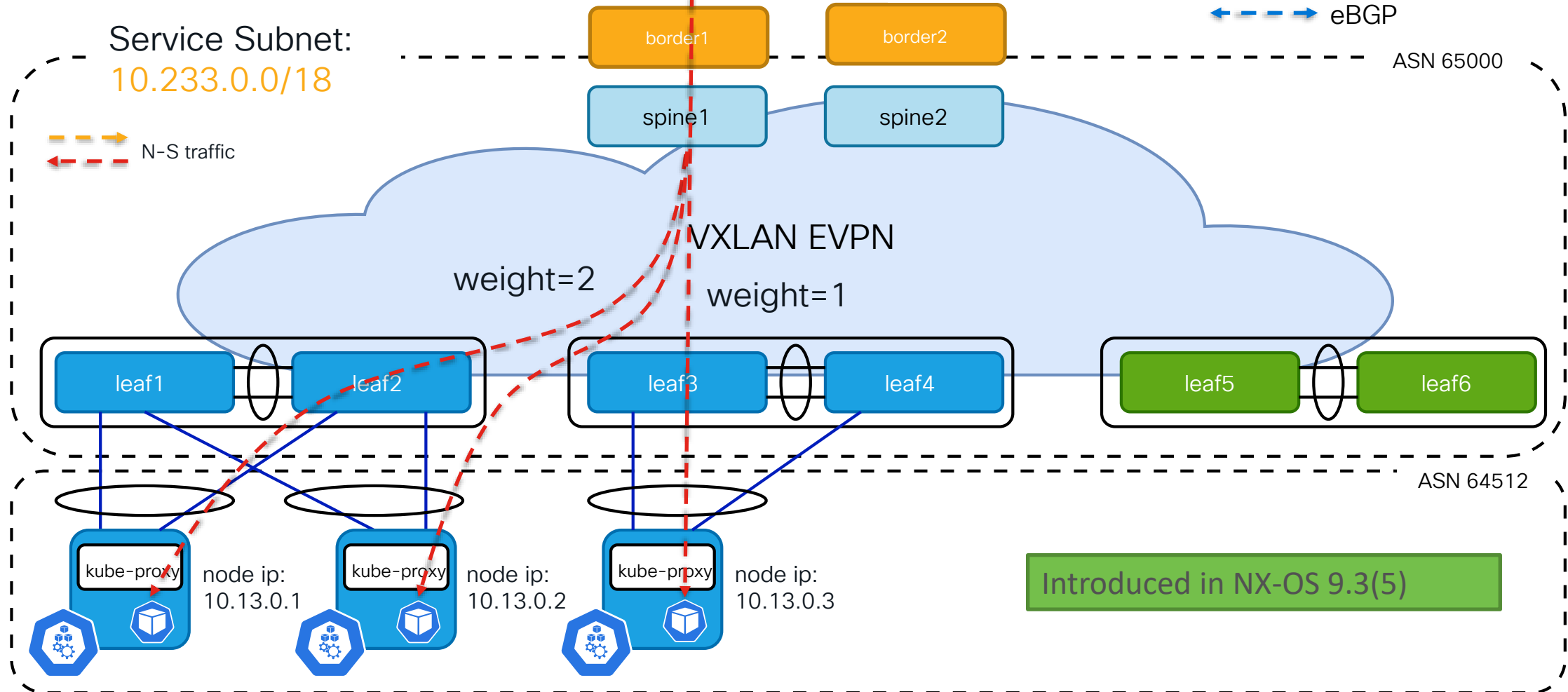
# Centralized Routing Peering

## 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

←--- eBGP ---→



# Summary

- Greenfield Calico network does not require L2 extension
- The best practice is peering BGP neighborhood 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](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)

# Technical Session Surveys

- Attendees who fill out a minimum of four session surveys and the overall event survey will get Cisco Live branded socks!
- Attendees will also earn 100 points in the Cisco Live Game for every survey completed.
- These points help you get on the leaderboard and increase your chances of winning daily and grand prizes.





# Cisco Learning and Certifications

From technology training and team development to Cisco certifications and learning plans, let us help you empower your business and career. [www.cisco.com/go/certs](https://www.cisco.com/go/certs)

## Pay for Learning with Cisco Learning Credits

(CLCs) are prepaid training vouchers redeemed directly with Cisco.



## Learn



### Cisco U.

IT learning hub that guides teams and learners toward their goals

### Cisco Digital Learning

Subscription-based product, technology, and certification training

### Cisco Modeling Labs

Network simulation platform for design, testing, and troubleshooting

### Cisco Learning Network

Resource community portal for certifications and learning



## Train



### Cisco Training Bootcamps

Intensive team & individual automation and technology training programs

### Cisco Learning Partner Program

Authorized training partners supporting Cisco technology and career certifications

### Cisco Instructor-led and Virtual Instructor-led training

Accelerated curriculum of product, technology, and certification courses



## Certify



### Cisco Certifications and Specialist Certifications

Award-winning certification program empowers students and IT Professionals to advance their technical careers

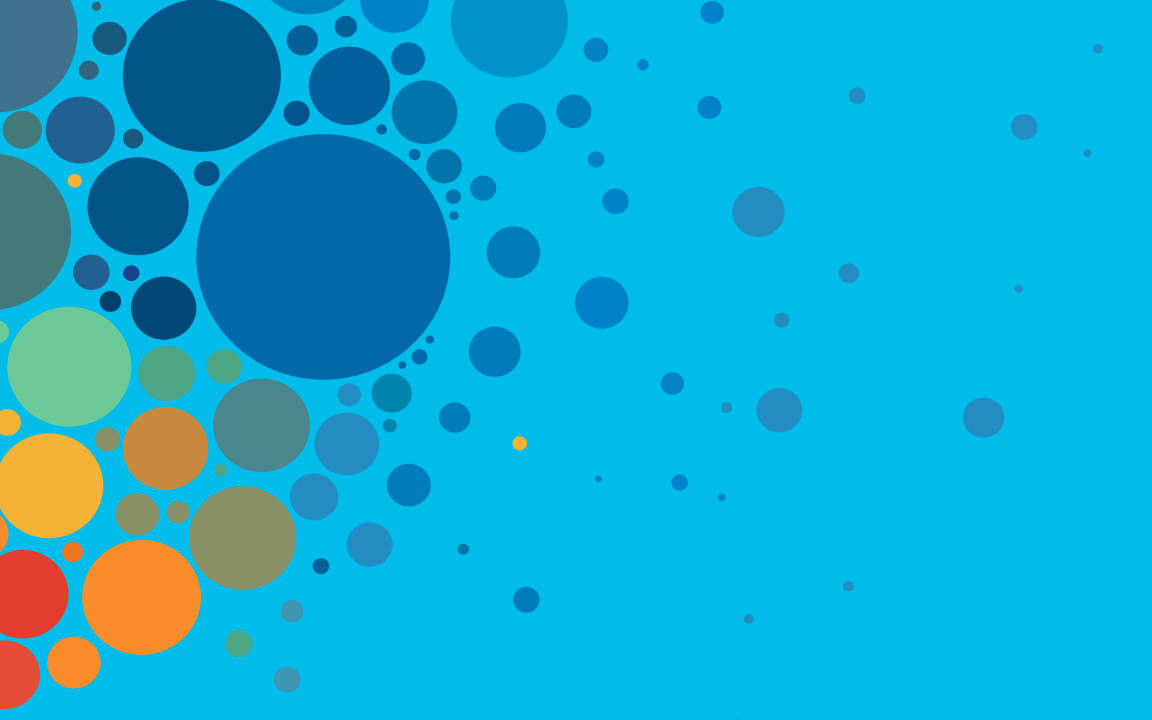
### Cisco Guided Study Groups

180-day certification prep program with learning and support

### Cisco Continuing Education Program

Recertification training options for Cisco certified individuals

Here at the event? Visit us at **The Learning and Certifications lounge at the World of Solutions**



# 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](https://www.CiscoLive.com/on-demand)



The bridge to possible

# Thank you

CISCO *Live!*

ALL IN

#CiscoLive