KUBESPHERE x CLOUD NATIVE COMPUTING FOUNDATION

Exposing Your Services in Bare Metal Environment Using PorterLB and KubeSphere

Jiong Duan – KubeSphere Team

KUBESPHERE

Agenda

- What is KubeSphere
- Introduction to PorterLB, and its cloud native architecture
- How to install PorterLB on Kubernetes using KubeSphere
- How to use Porter to expose LoadBalancer type of Service from bare metal Kubernetes

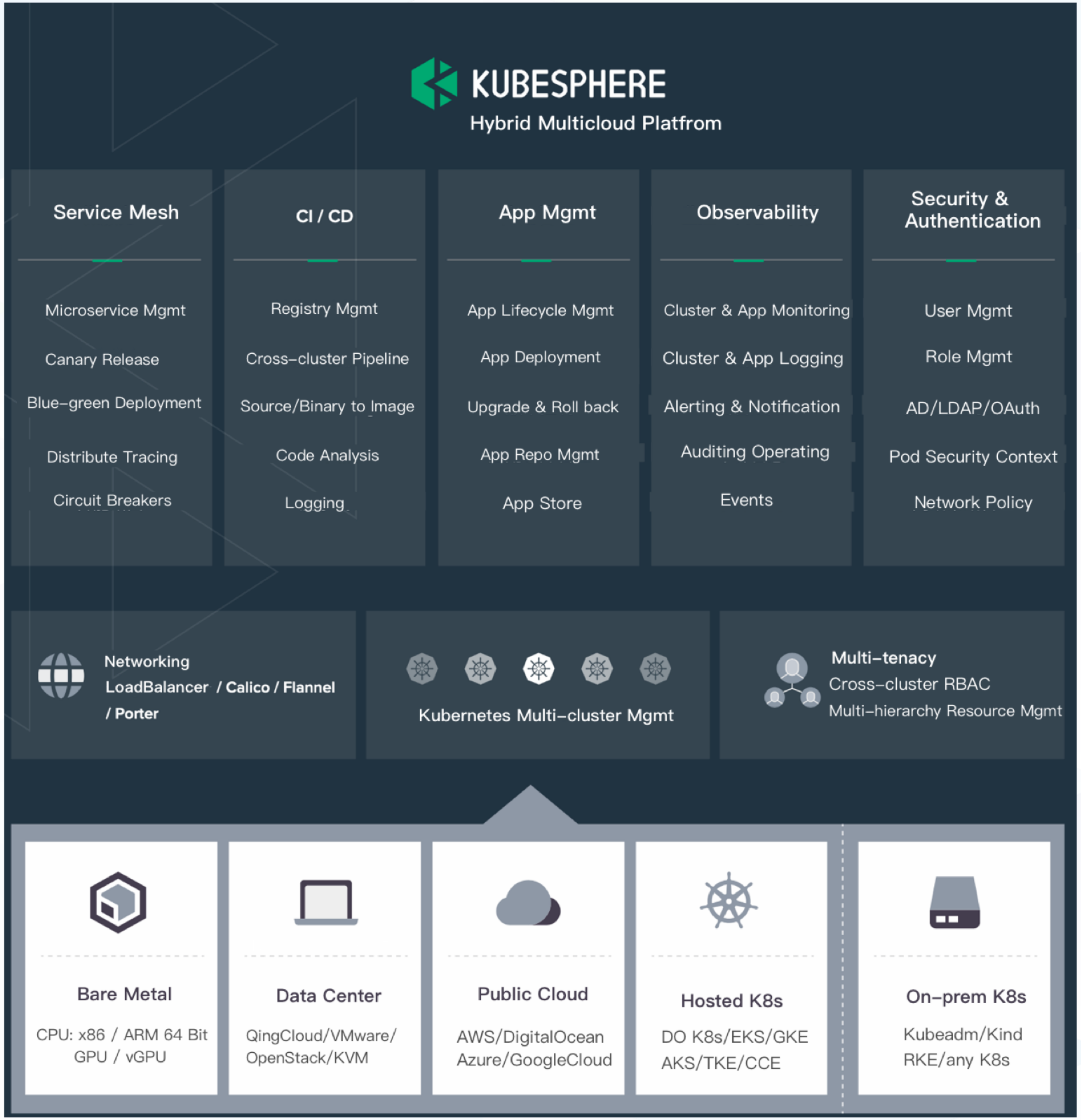
KUBESPHERE

What is KubeSphere

01

KubeSphere (https://kubesphere.io) is a **distributed operating system managing cloud native applications** with Kubernetes as its kernel, and provides a plug-and-play open architecture for third-party applications seamless integration to boost its ecosystem.

Architecture



KUBESPHERE

Hybrid Multicloud Platfrom

Service Mesh

- Microservice Mgmt
- Canary Release
- Blue-green Deployment
- Distribute Tracing
- Circuit Breakers

CI / CD

- Registry Mgmt
- Cross-cluster Pipeline
- Source/Binary to Image
- Code Analysis
- Logging

App Mgmt

- App Lifecycle Mgmt
- App Deployment
- Upgrade & Roll back
- App Repo Mgmt
- App Store

Observability

- Cluster & App Monitoring
- Cluster & App Logging
- Alerting & Notification
- Auditing Operating
- Events

Security & Authentication

- User Mgmt
- Role Mgmt
- AD/LDAP/OAuth
- Pod Security Context
- Network Policy

Networking
LoadBalancer / Calico / Flannel
/ Porter

Kubernetes Multi-cluster Mgmt

Multi-tenacy
Cross-cluster RBAC
Multi-hierarchy Resource Mgmt

Bare Metal
CPU: x86 / ARM 64 Bit
GPU / vGPU

Data Center
QingCloud/VMware/
OpenStack/KVM

Public Cloud
AWS/DigitalOcean
Azure/GoogleCloud

Hosted K8s
DO K8s/EKS/GKE
AKS/TKE/CCE

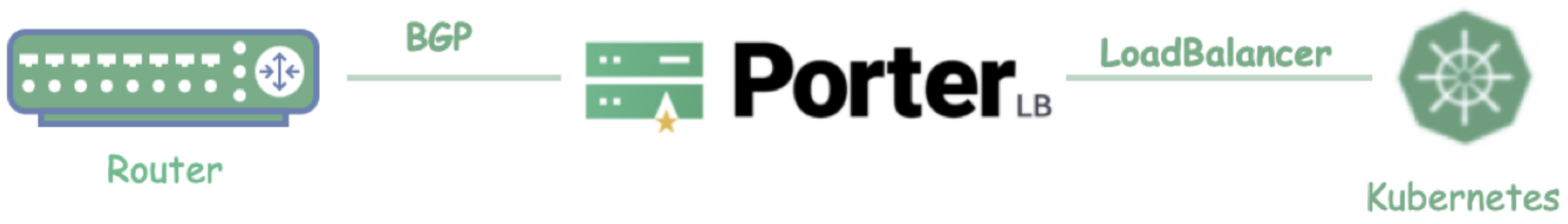
On-prem K8s
Kubeadm/Kind
RKE/any K8s

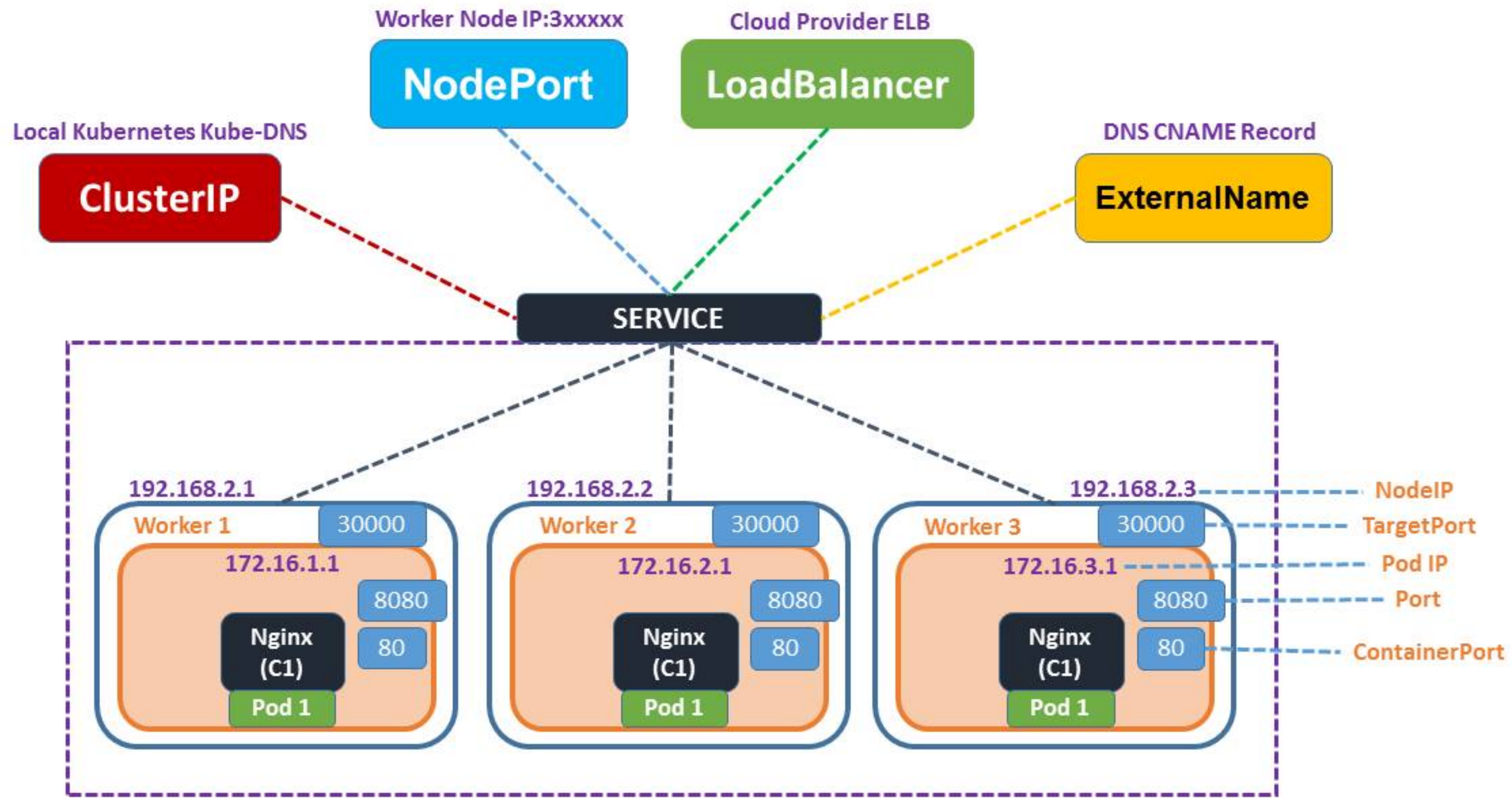
02

Introduction to PorterLB, and its cloud native architecture

What is PorterLB

- A community-driven open source load balancer
- Designed for Bare Metal Kubernetes clusters
- Load balancing via BGP and ECMP





Why PorterLB

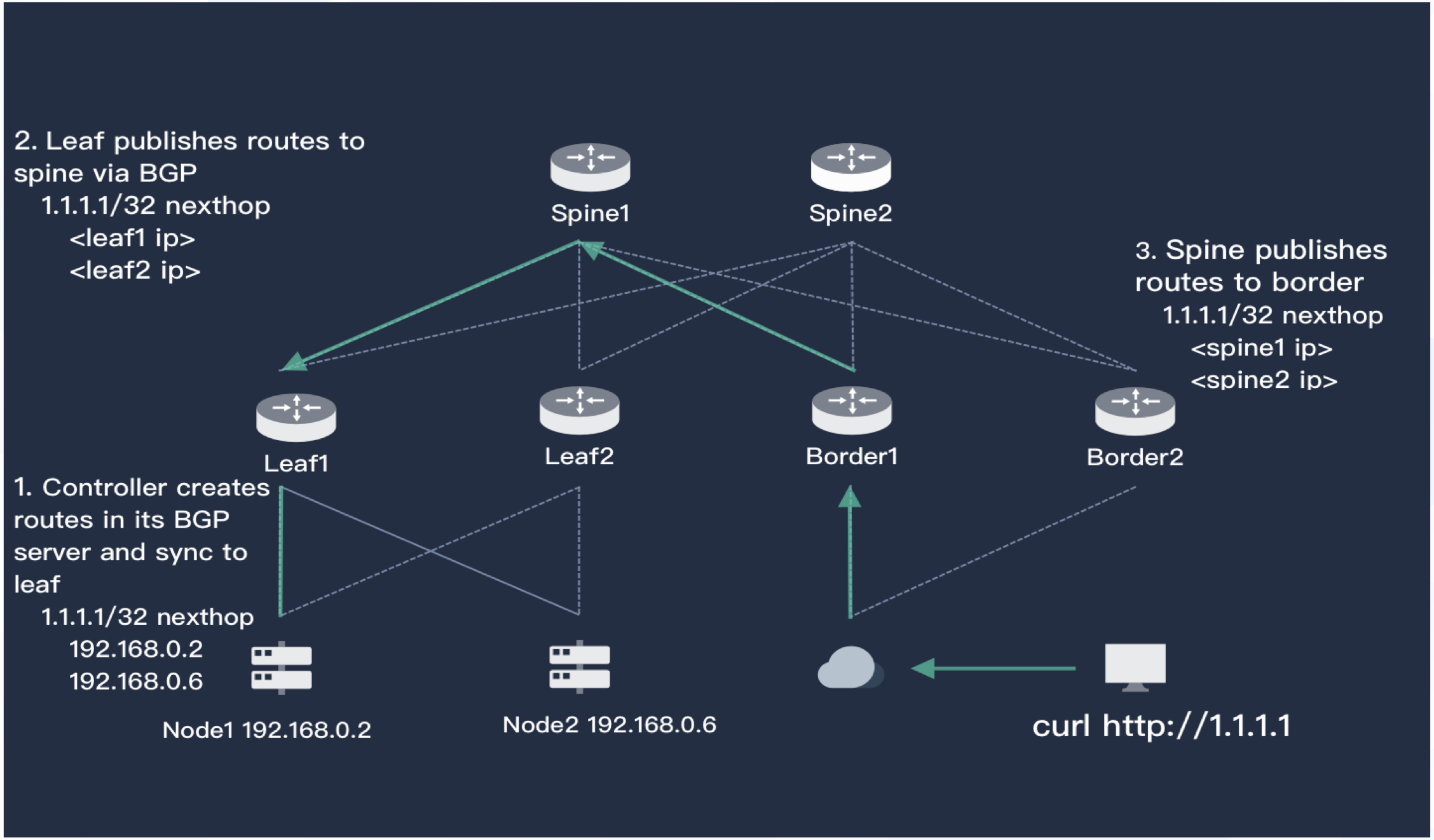
- Cloud Providers
  - QingCloud
  - Openstack
  - GCE
  - …
- SDN
  - Cisco ACI
  - …

✗

- Common Switch
- Bare Metal Environment
- No SDN Capability

In the cloud-hosted Kubernetes cluster, the cloud providers usually provide the Load-Balancer to assign IPs and bring traffic into Kubernetes cluster. However, Kubernetes does not provide a load-balancer for bare metal cluster.

PorterLB Principle

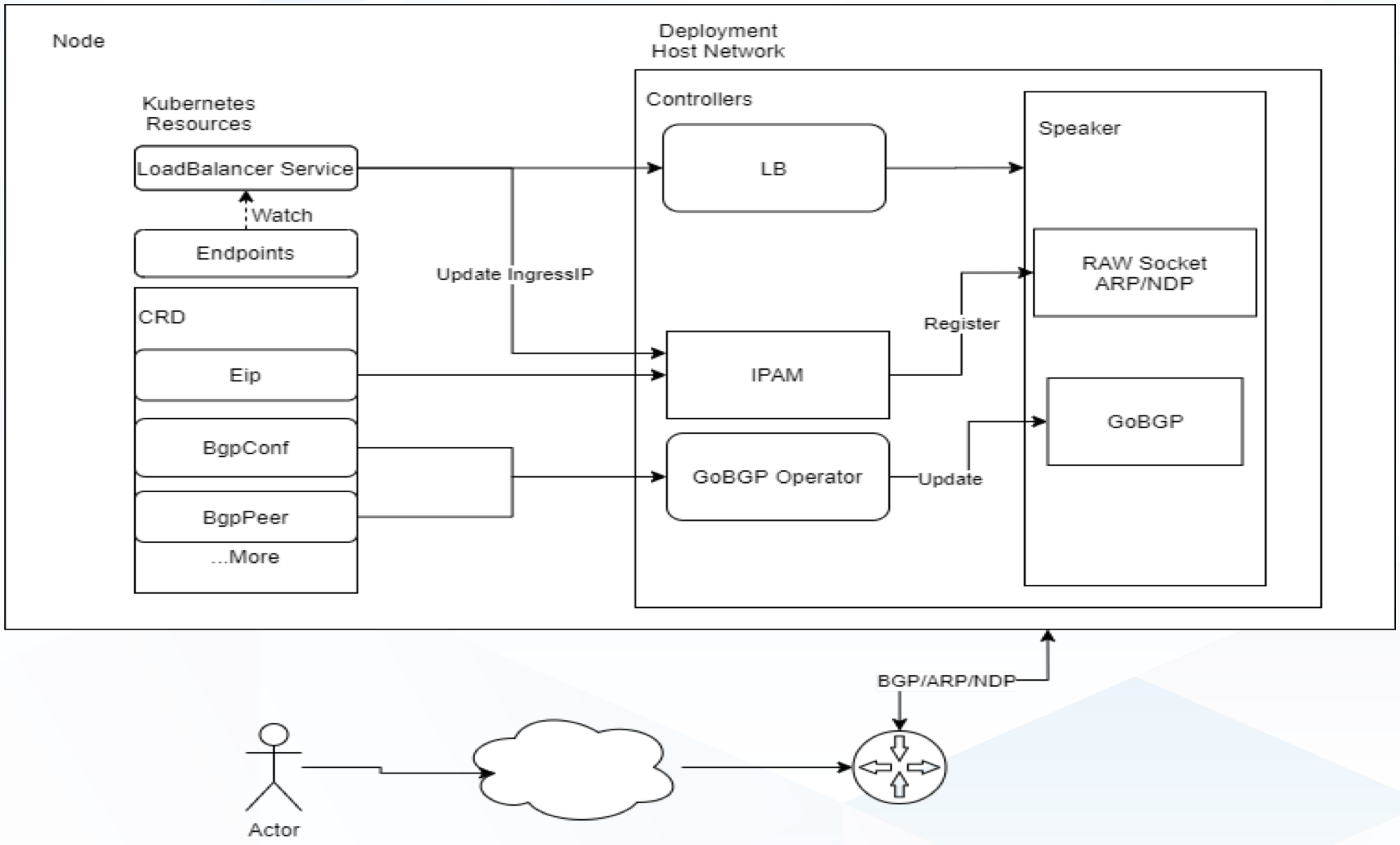


2. Leaf publishes routes to spine via BGP
   1.1.1.1/32 nexthop
      <leaf1 ip>
      <leaf2 ip>

3. Spine publishes routes to border
   1.1.1.1/32 nexthop
      <spine1 ip>
      <spine2 ip>

1. Controller creates routes in its BGP server and sync to leaf
   1.1.1.1/32 nexthop
      192.168.0.2
      192.168.0.6

Cloud Native Architecture of PorterLB



Why GoBGP

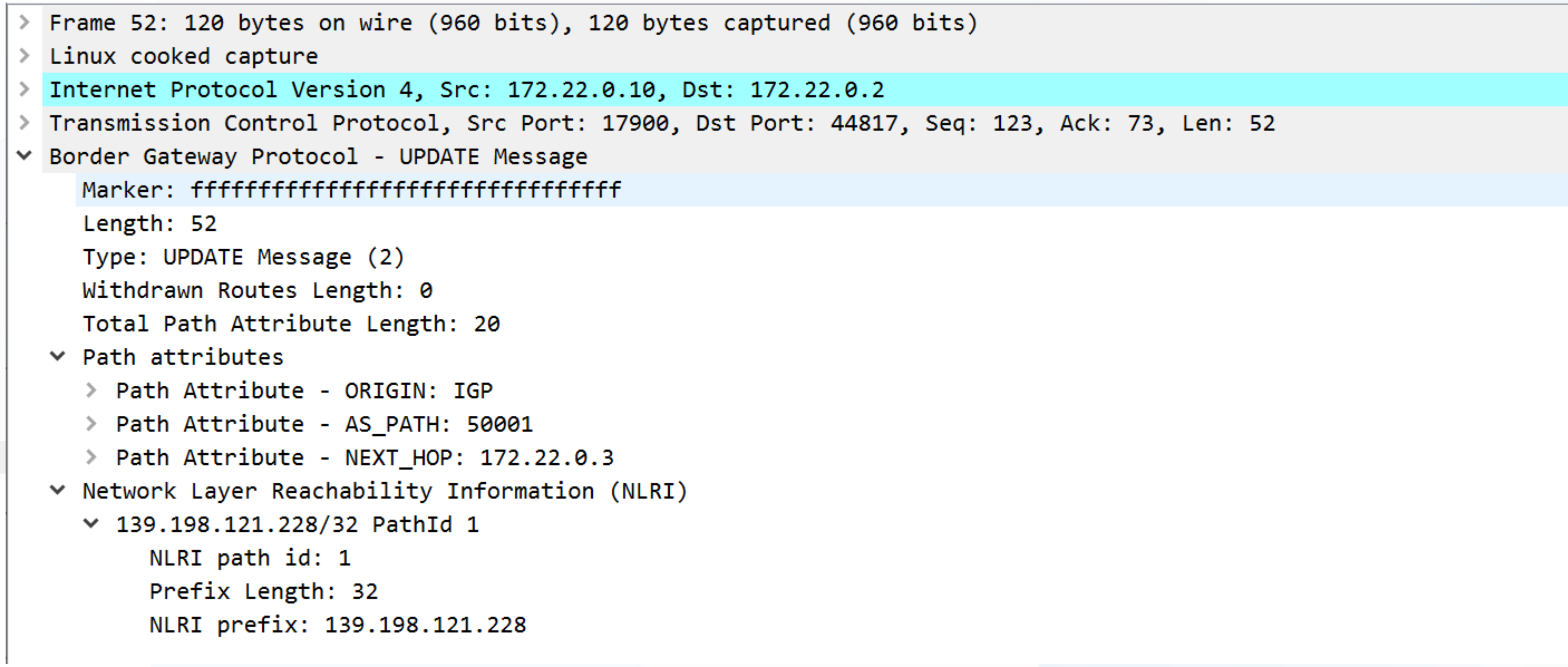
- GoBGP as a Go Native BGP library
- Rich Features
- Low development costs, community support
- Automation Friendly
  - GoBGP is designed to be easily integrated with other software with its RPC APIs instead of manually changing its config via CLI. GoBGP also supports its CLI though.

Who uses GoBGP in production?

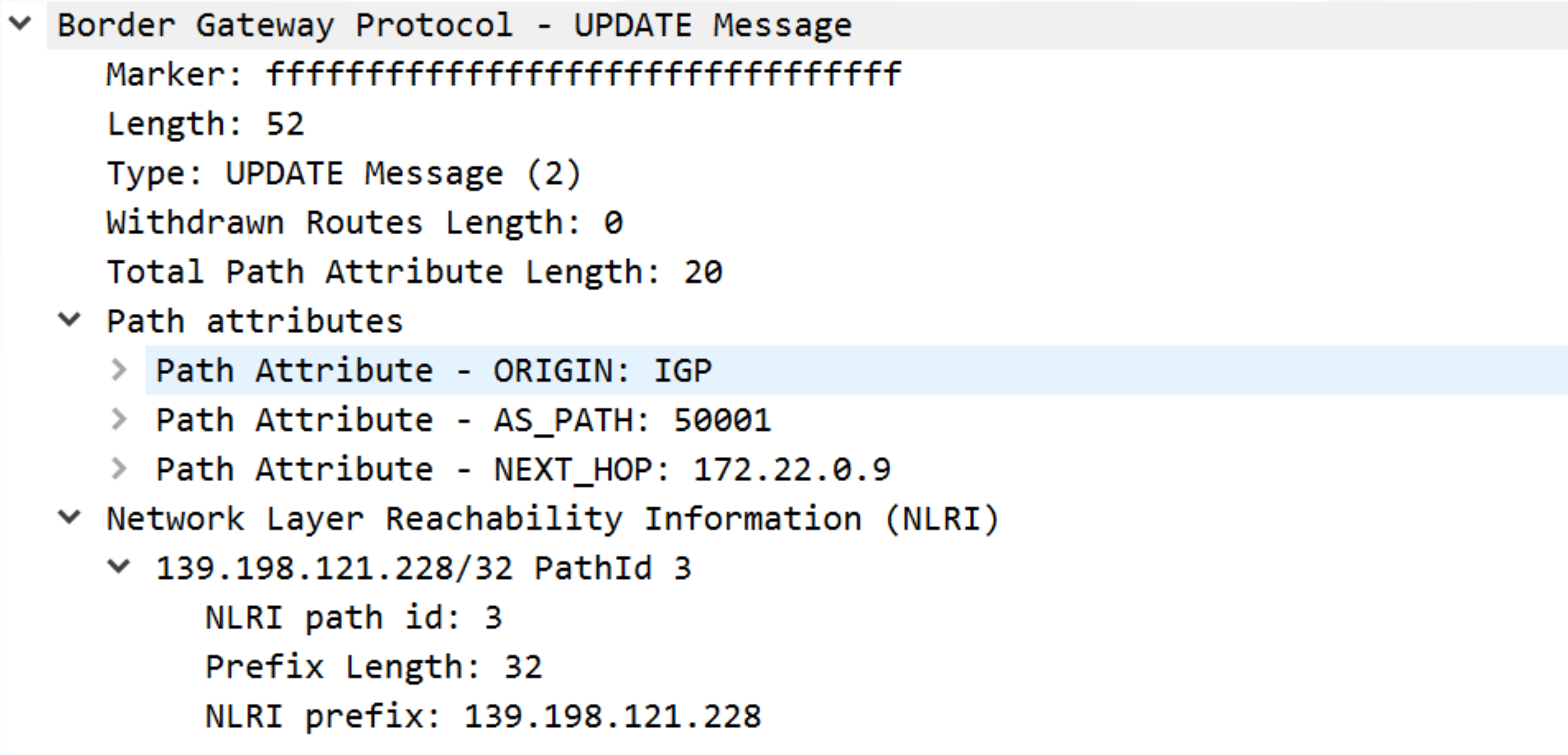
JPNAP
CLOUDWAN
nic.br
COM HEM
SAKURA internet
NEXUSGUARD
FASTNETMON
DigitalOcean

BGP Additional Paths(1/2)

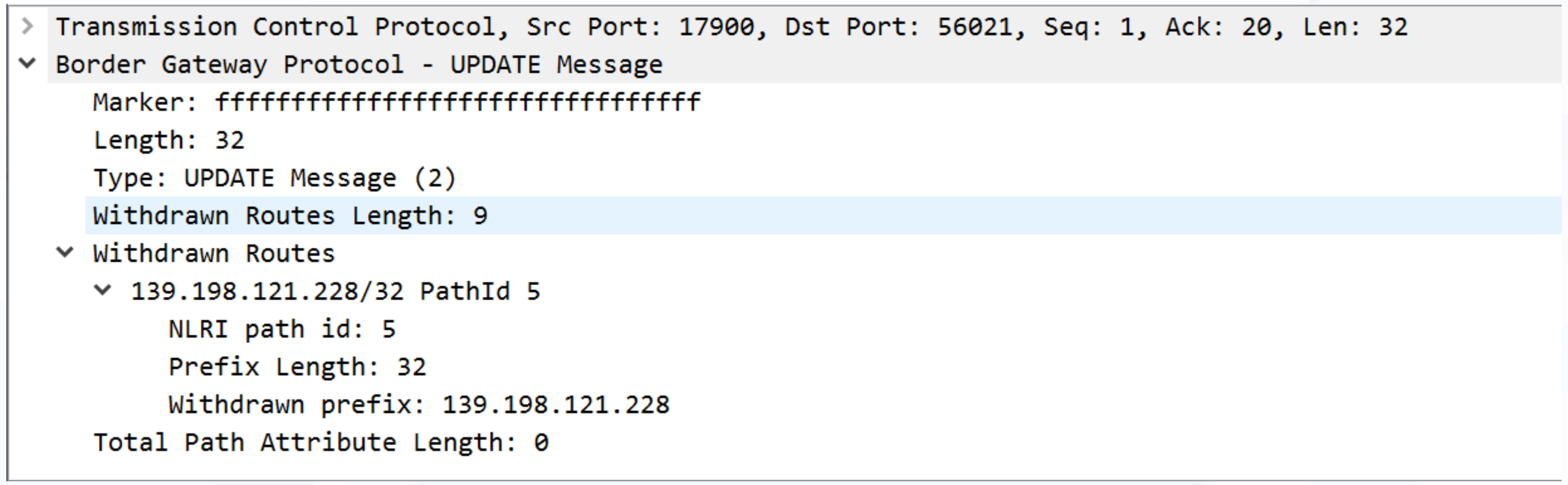
- Advertisement of Multiple Paths in BGP
- Best Practices for Advertisement of Multiple Paths in IBGP
- A Border Gateway Protocol 4 (BGP-4)



```
> Frame 52: 120 bytes on wire (960 bits), 120 bytes captured (960 bits)
> Linux cooked capture
> Internet Protocol Version 4, Src: 172.22.0.10, Dst: 172.22.0.2
> Transmission Control Protocol, Src Port: 17900, Dst Port: 44817, Seq: 123, Ack: 73, Len: 52
v Border Gateway Protocol - UPDATE Message
    Marker: ffffffffffffffffffffffffffffffff
    Length: 52
    Type: UPDATE Message (2)
    Withdrawn Routes Length: 0
    Total Path Attribute Length: 20
  v Path attributes
    > Path Attribute - ORIGIN: IGP
    > Path Attribute - AS_PATH: 50001
    > Path Attribute - NEXT_HOP: 172.22.0.3
  v Network Layer Reachability Information (NLRI)
    v 139.198.121.228/32 PathId 1
        NLRI path id: 1
        Prefix Length: 32
        NLRI prefix: 139.198.121.228
```



```
v Border Gateway Protocol - UPDATE Message
    Marker: ffffffffffffffffffffffffffffffff
    Length: 52
    Type: UPDATE Message (2)
    Withdrawn Routes Length: 0
    Total Path Attribute Length: 20
  v Path attributes
    > Path Attribute - ORIGIN: IGP
    > Path Attribute - AS_PATH: 50001
    > Path Attribute - NEXT_HOP: 172.22.0.9
  v Network Layer Reachability Information (NLRI)
    v 139.198.121.228/32 PathId 3
        NLRI path id: 3
        Prefix Length: 32
        NLRI prefix: 139.198.121.228
```



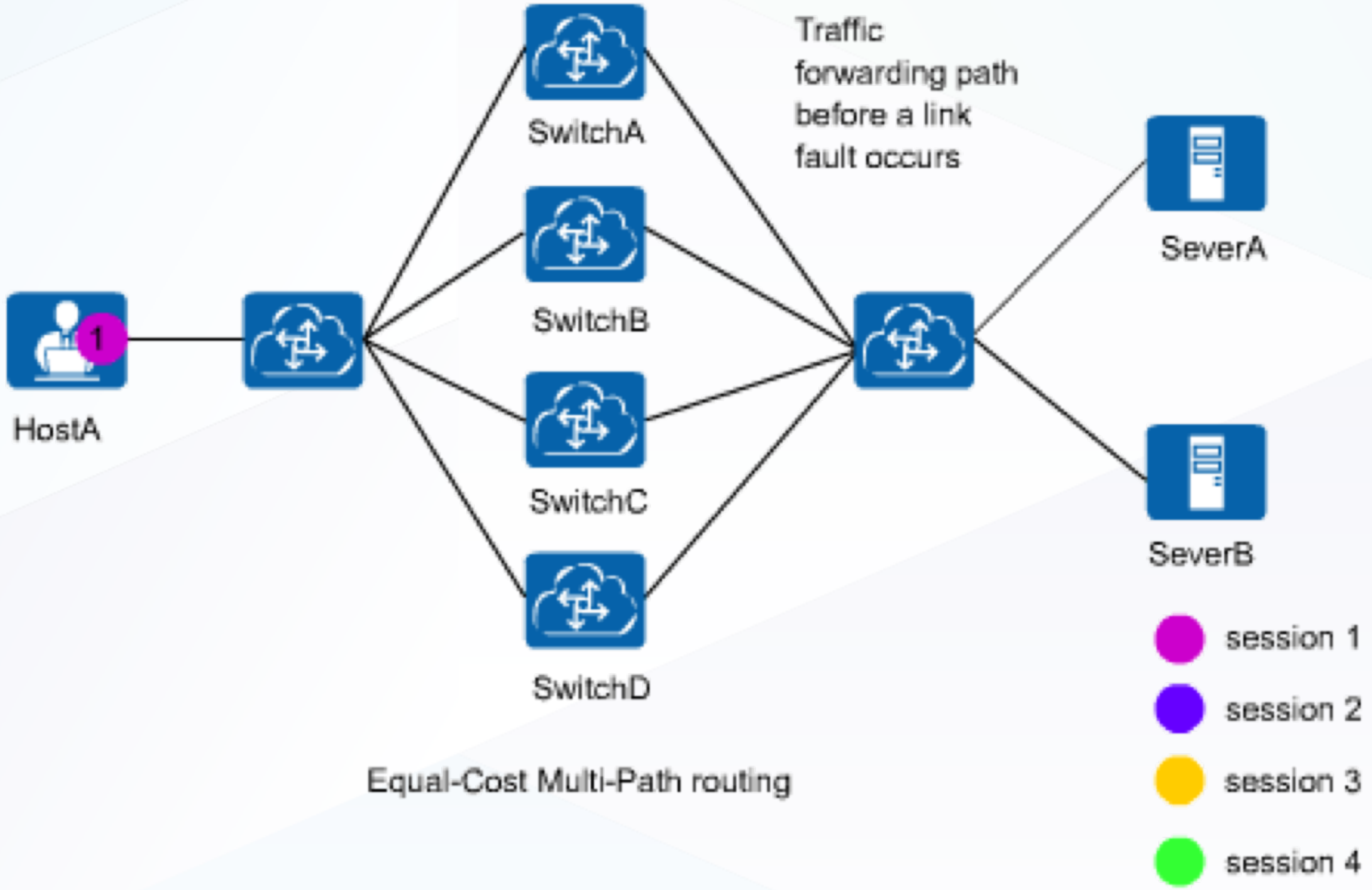
```
> Transmission Control Protocol, Src Port: 17900, Dst Port: 56021, Seq: 1, Ack: 20, Len: 32
v Border Gateway Protocol - UPDATE Message
    Marker: ffffffffffffffffffffffffffffffff
    Length: 32
    Type: UPDATE Message (2)
    Withdrawn Routes Length: 9
  v Withdrawn Routes
    v 139.198.121.228/32 PathId 5
        NLRI path id: 5
        Prefix Length: 32
        Withdrawn prefix: 139.198.121.228
    Total Path Attribute Length: 0
```

BGP Additional Paths(2/2)

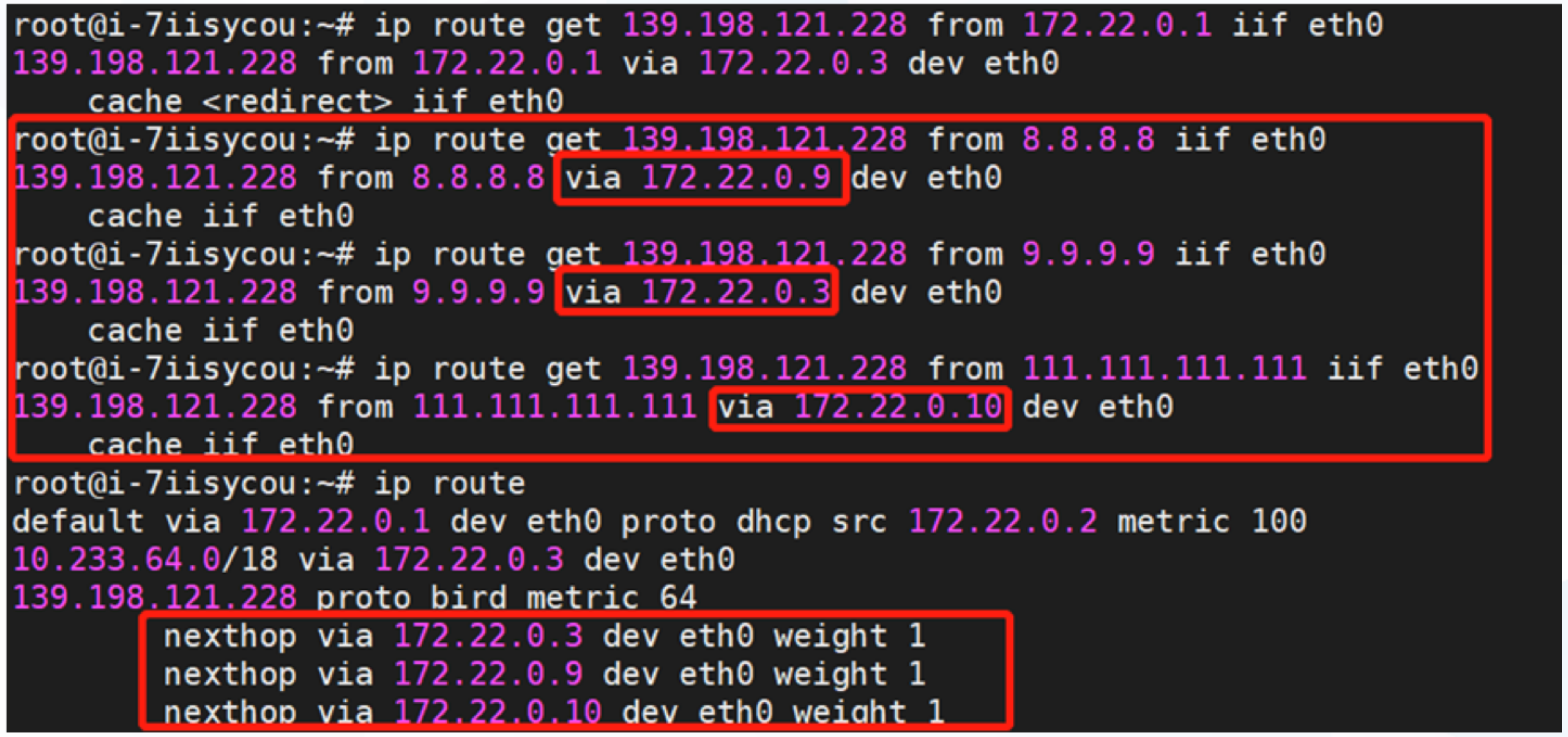
- externalTrafficPolicy
  - Cluster
    - Equivalent route Nexthop will be all nodes
  - Local
    - Equivalent route Nexthop will be the node where the endpoints are located

ECMP

- Equal-cost multi-path routing
  - Per-packet hash
  - L3 hash
  - L4 hash (aka Layer3 + Layer4)
  - More…
- Multipath Routing in Linux



Equal-Cost Multi-Path routing



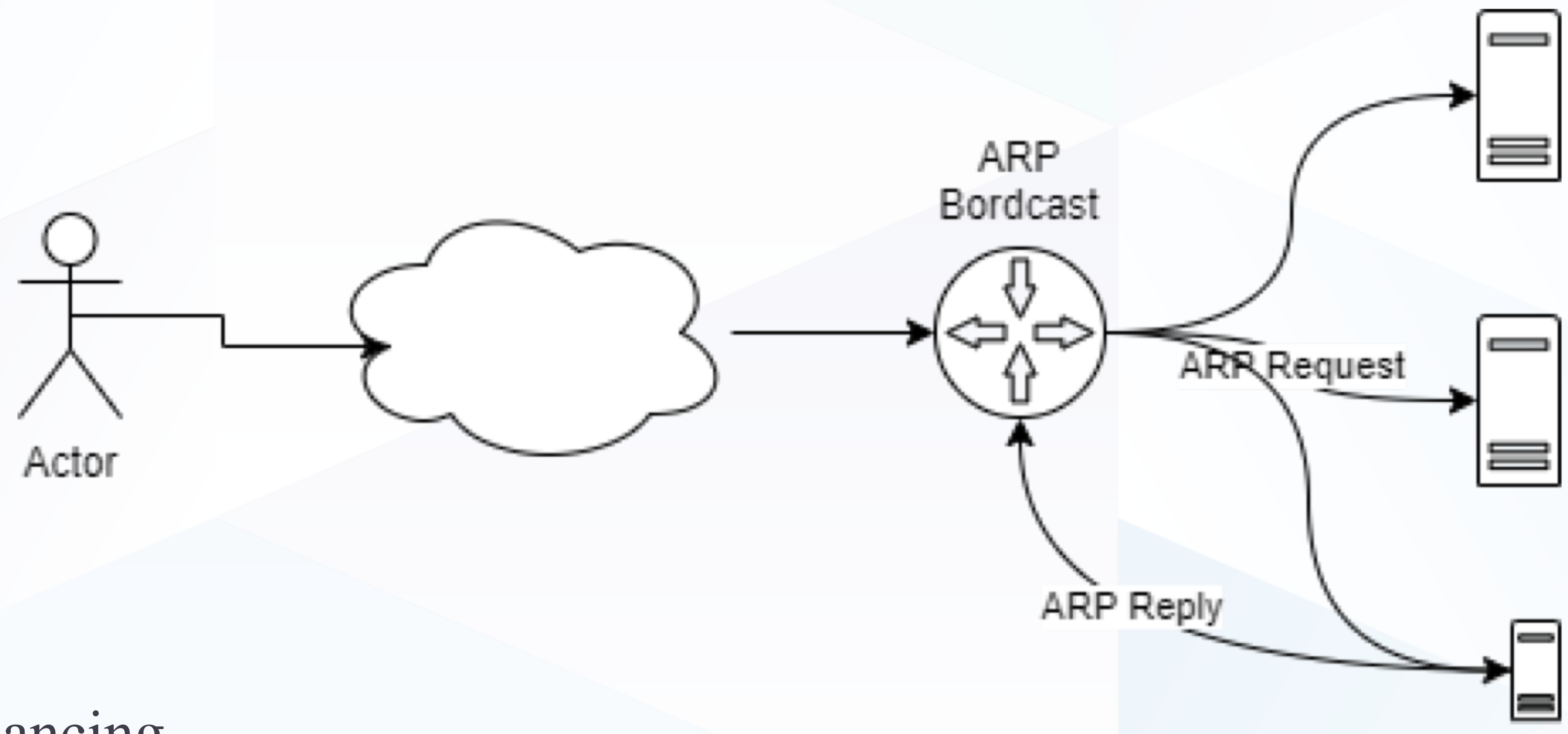
```
root@i-7iisycou:~# ip route get 139.198.121.228 from 172.22.0.1 iif eth0
139.198.121.228 from 172.22.0.1 via 172.22.0.3 dev eth0
    cache <redirect> iif eth0
root@i-7iisycou:~# ip route get 139.198.121.228 from 8.8.8.8 iif eth0
139.198.121.228 from 8.8.8.8 via 172.22.0.9 dev eth0
    cache iif eth0
root@i-7iisycou:~# ip route get 139.198.121.228 from 9.9.9.9 iif eth0
139.198.121.228 from 9.9.9.9 via 172.22.0.3 dev eth0
    cache iif eth0
root@i-7iisycou:~# ip route get 139.198.121.228 from 111.111.111.111 iif eth0
139.198.121.228 from 111.111.111.111 via 172.22.0.10 dev eth0
    cache iif eth0
root@i-7iisycou:~# ip route
default via 172.22.0.1 dev eth0 proto dhcp src 172.22.0.2 metric 100
10.233.64.0/18 via 172.22.0.3 dev eth0
139.198.121.228 proto bird metric 64
        nexthop via 172.22.0.3 dev eth0 weight 1
        nexthop via 172.22.0.9 dev eth0 weight 1
        nexthop via 172.22.0.10 dev eth0 weight 1
```

HA Configuration of PorterLB

- HA Configuration of PorterLB
  - Porter Manager Multicopy
  - EIP Address Management Stateless
  - Speaker Stateless
- Routing Table HA
  - BGP Graceful Restart
  - Multiple BGP Sessions with multiple copies

Layer2

- Why not BGP
  - Security Compliance
  - Hardware too old to support BGP
- How it works
  - Kubernetes Leader Election
  - Single point of bottleneck and no load balancing
  - Save nexthop to the annotation of the service
  - Enable strictARP in kube-proxy configmap
  - Gratuitous ARP
- The Problem
  - rp spoofing



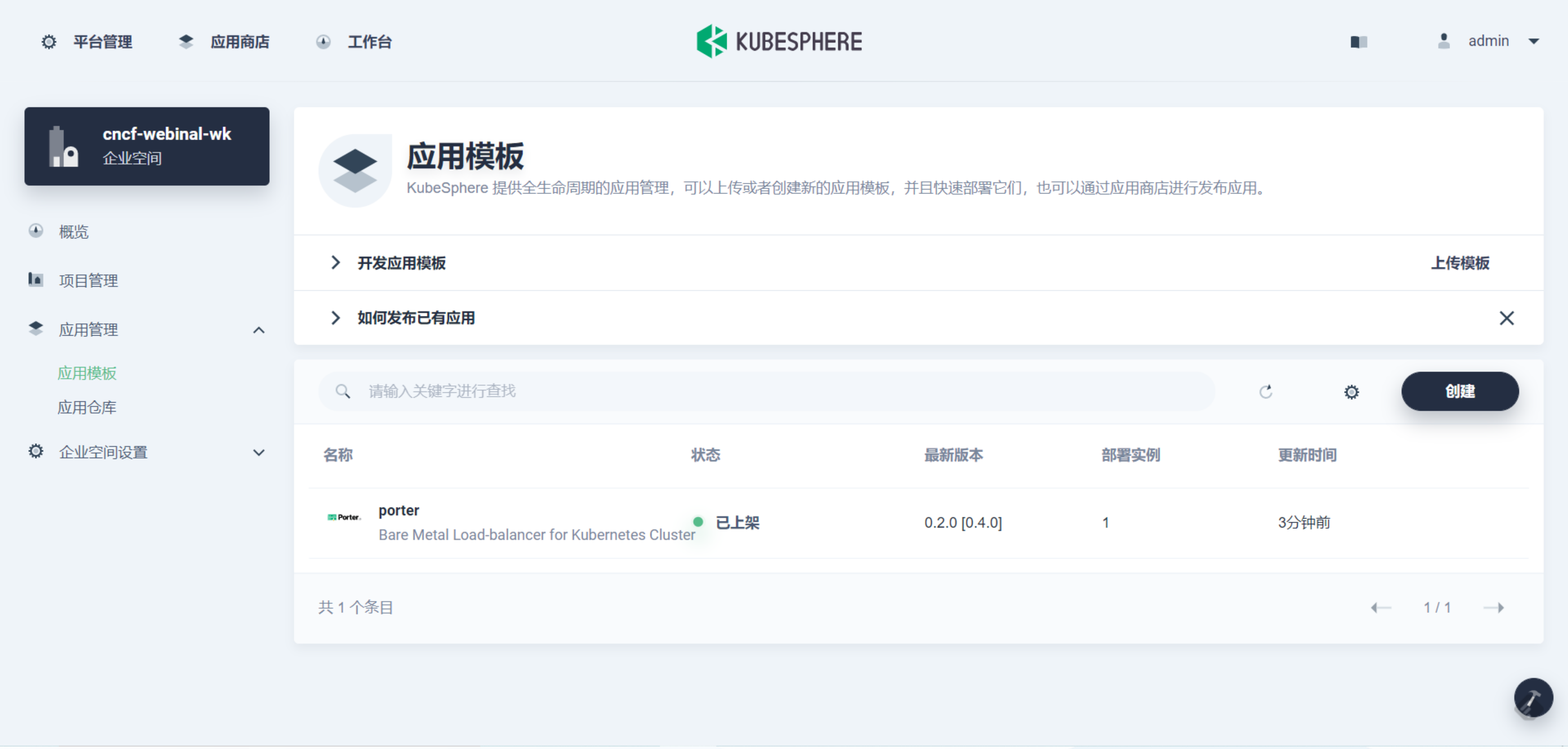
03

Install PorterLB on KubeSphere

Install PorterLB

- Install Porter in one click
  - kubectl apply -f https://raw.githubusercontent.com/kubesphere/porter/master/deploy/porter.yaml
- Install via chart package
  - helm repo add test https://charts.kubesphere.io/test
  - helm repo update
  - helm install porter test/porter
- Install via KubeSphere Console

Install PorterLB on KubeSphere (1/3)



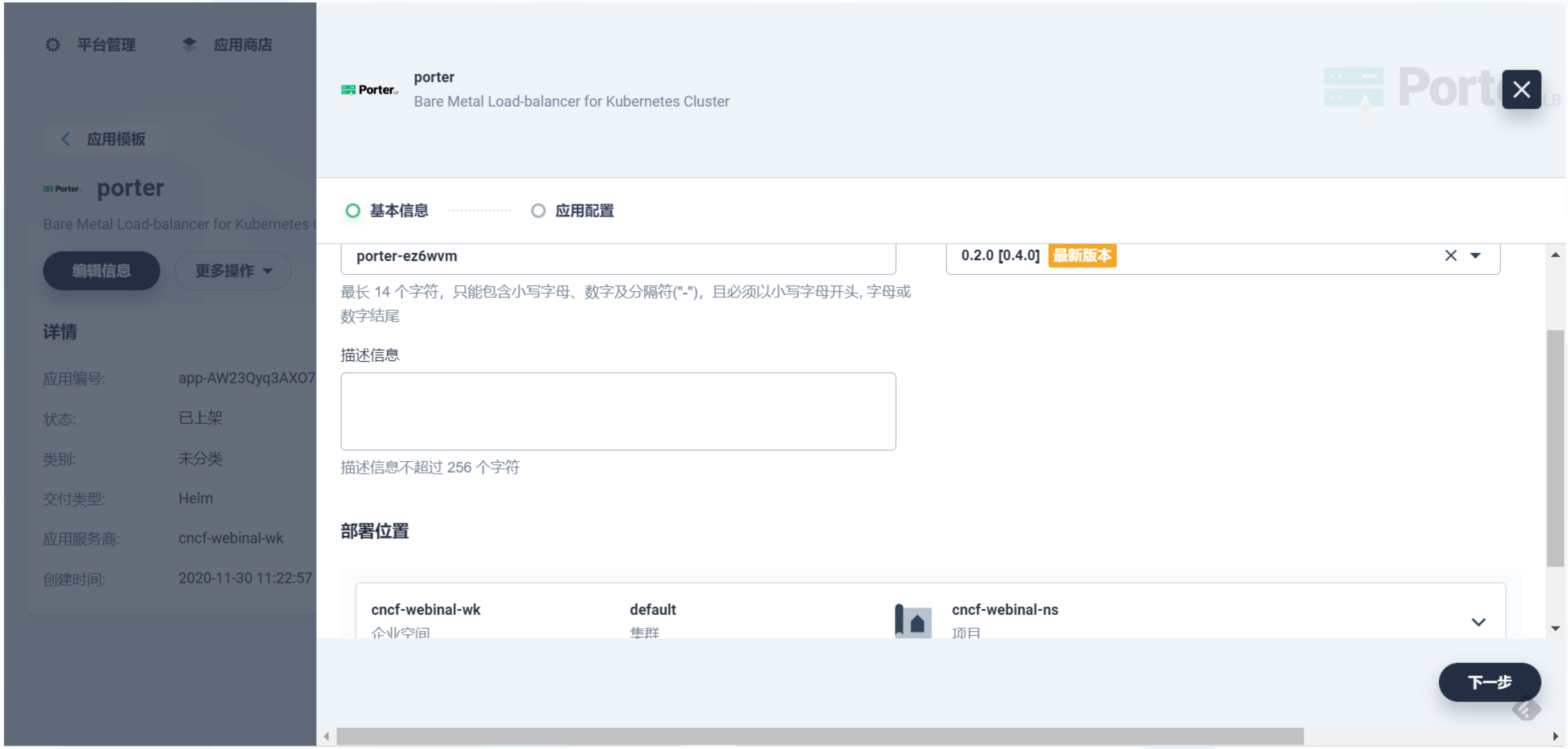
应用模板

KubeSphere 提供全生命周期的应用管理，可以上传或者创建新的应用模板，并且快速部署它们，也可以通过应用商店进行发布应用。

名称	状态	最新版本	部署实例	更新时间
porter Bare Metal Load-balancer for Kubernetes Cluster	● 已上架	0.2.0 [0.4.0]	1	3分钟前

共 1 个条目

Install PorterLB on KubeSphere (2/3)



详情

应用编号: app-AW23Qyq3AXO7

状态: 已上架

类别: 未分类

交付类型: Helm

应用服务商: cncf-webinal-wk

创建时间: 2020-11-30 11:22:57

porter

Bare Metal Load-balancer for Kubernetes Cluster

porter-ez6wvm

0.2.0 [0.4.0] 最新版本

最长 14 个字符，只能包含小写字母、数字及分隔符("-")，且必须以小写字母开头, 字母或数字结尾

描述信息

描述信息不超过 256 个字符

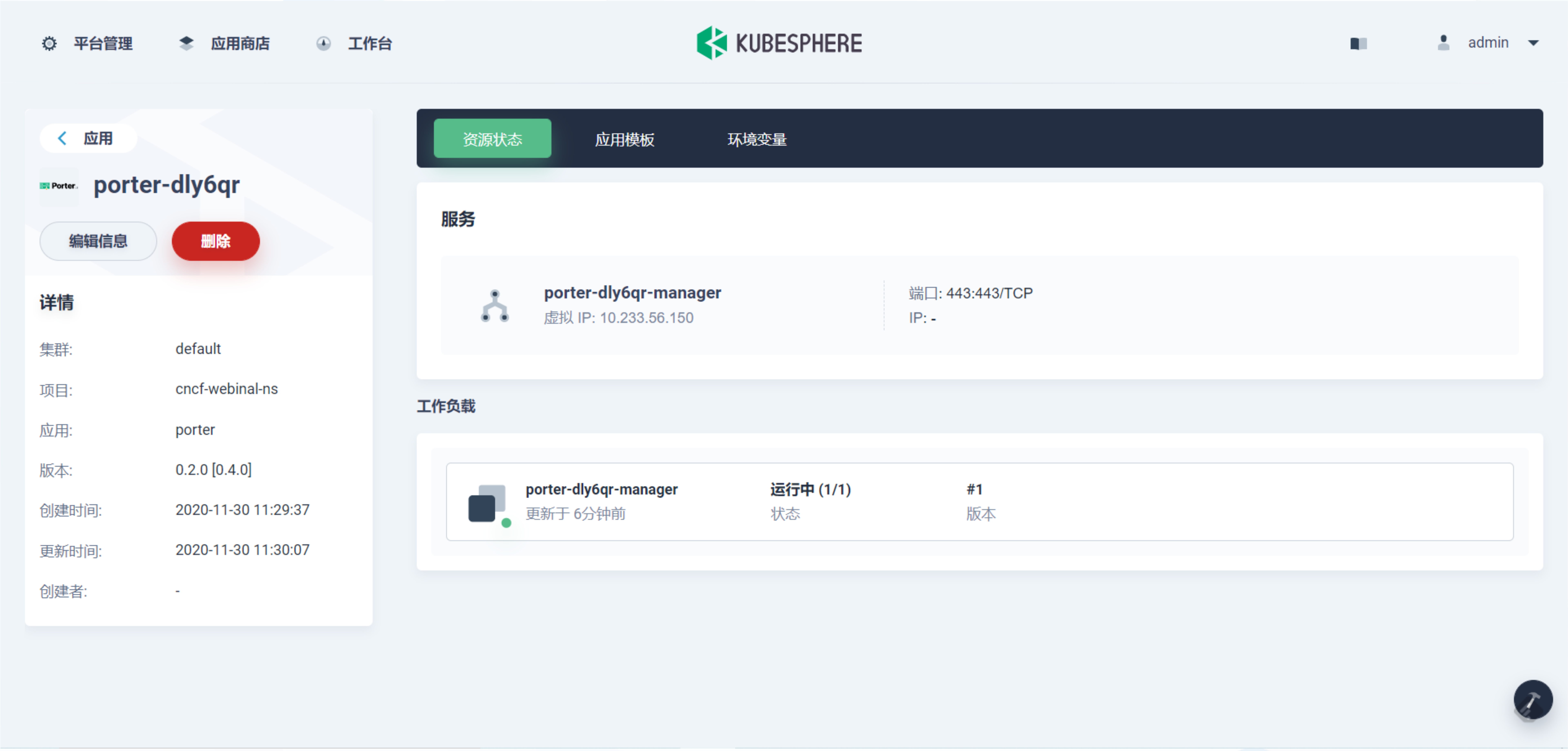
部署位置

cncf-webinal-wk 企业空间

default 集群

cncf-webinal-ns 项目

Install PorterLB on KubeSphere (3/3)



porter-dly6qr

详情

集群:	default
项目:	cncf-webinal-ns
应用:	porter
版本:	0.2.0 [0.4.0]
创建时间:	2020-11-30 11:29:37
更新时间:	2020-11-30 11:30:07
创建者:	-

服务

porter-dly6qr-manager
虚拟 IP: 10.233.56.150
端口: 443:443/TCP
IP: -

工作负载

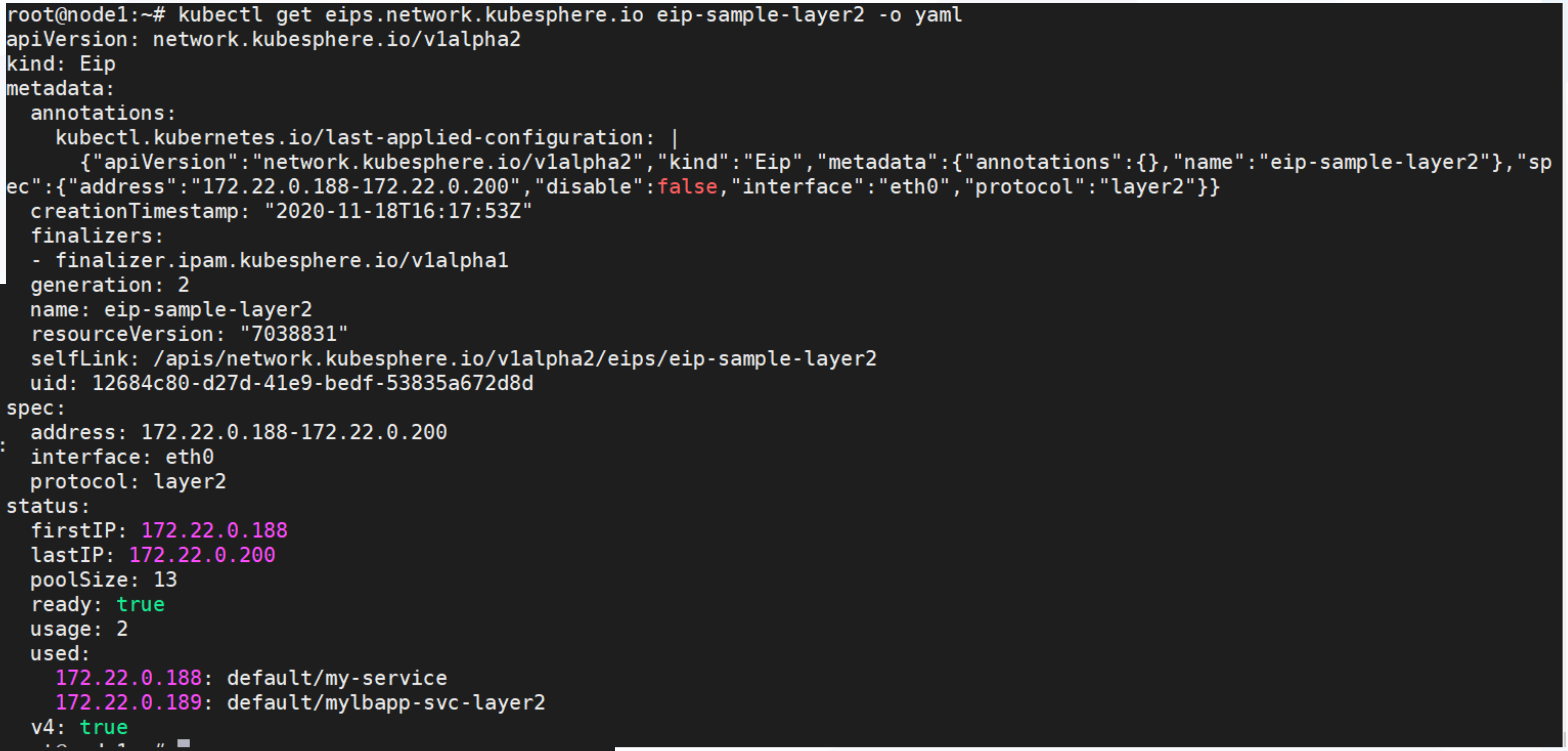
porter-dly6qr-manager
更新于 6分钟前
运行中 (1/1)
状态
#1
版本

04

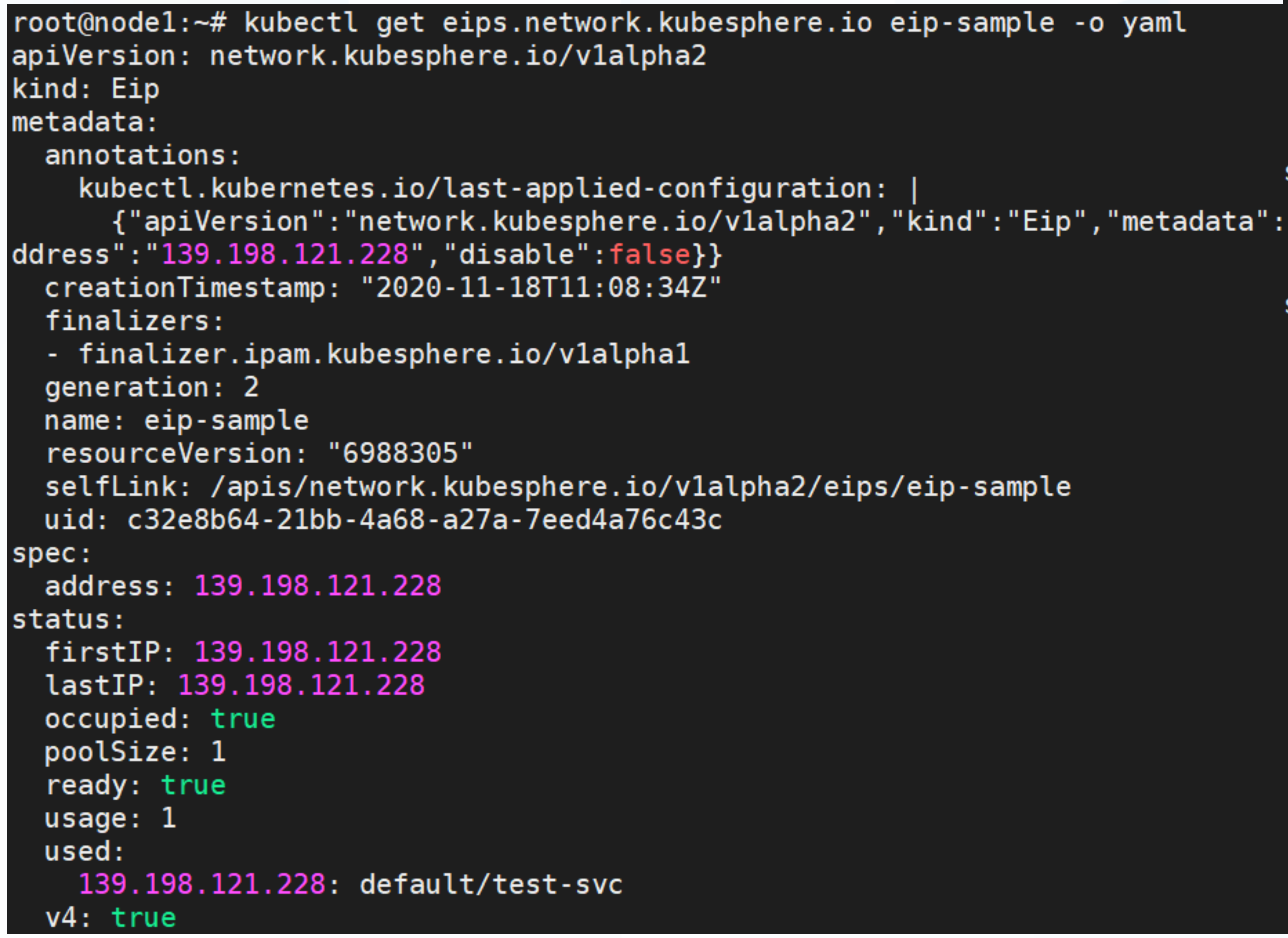
Expose your LoadBalancer Type of Services from Bare Metal Kubernetes

Config EIP

- Support IPv4 now, support for IPv6 will be completed soon
- Support Protocol BGP and Layer2
- View EIP allocation status via kubectl
- EIP Config Guide



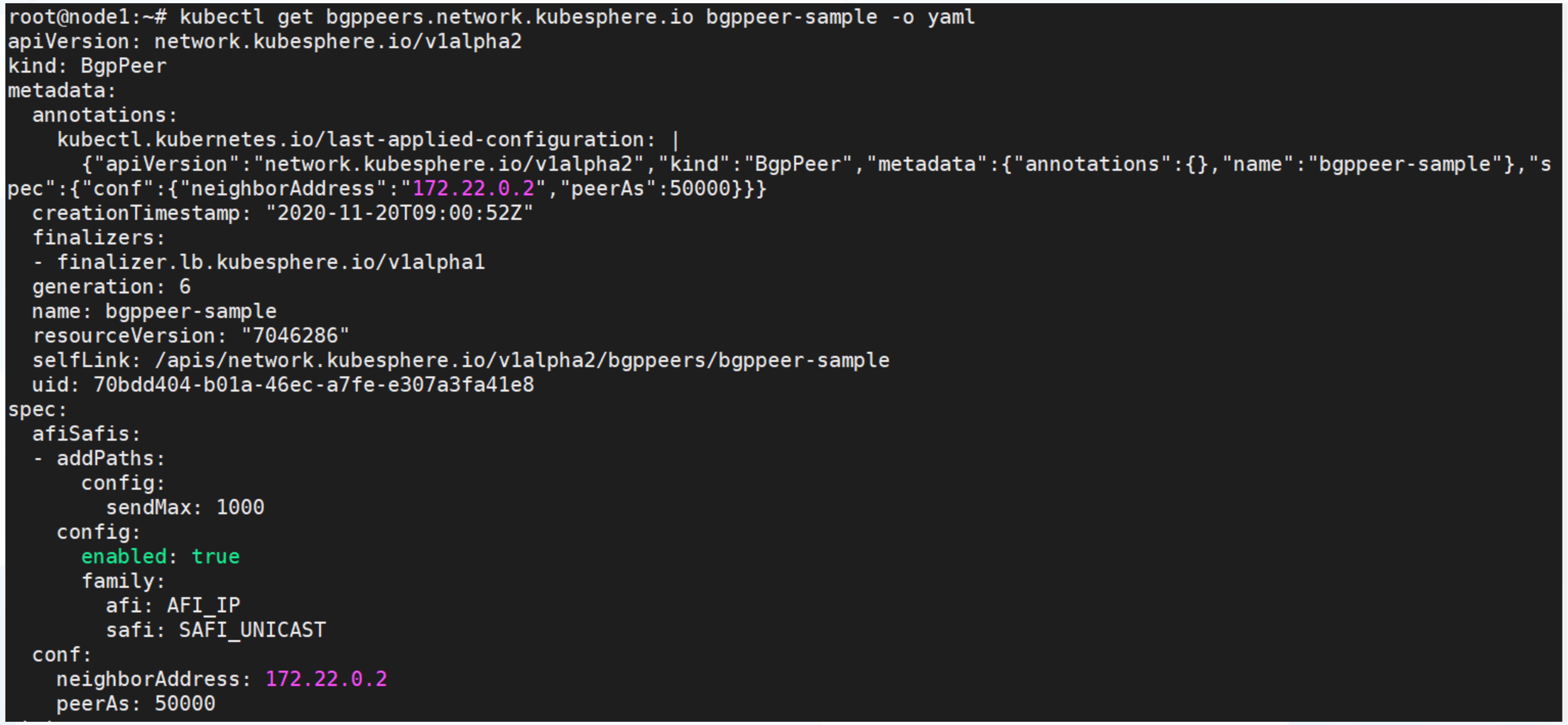
```
root@node1:~# kubectl get eips.network.kubesphere.io eip-sample-layer2 -o yaml
apiVersion: network.kubesphere.io/v1alpha2
kind: Eip
metadata:
  annotations:
    kubectl.kubernetes.io/last-applied-configuration: |
      {"apiVersion":"network.kubesphere.io/v1alpha2","kind":"Eip","metadata":{"annotations":{},"name":"eip-sample-layer2"},"sp
ec":{"address":"172.22.0.188-172.22.0.200","disable":false,"interface":"eth0","protocol":"layer2"}}
  creationTimestamp: "2020-11-18T16:17:53Z"
  finalizers:
  - finalizer.ipam.kubesphere.io/v1alpha1
  generation: 2
  name: eip-sample-layer2
  resourceVersion: "7038831"
  selfLink: /apis/network.kubesphere.io/v1alpha2/eips/eip-sample-layer2
  uid: 12684c80-d27d-41e9-bedf-53835a672d8d
spec:
  address: 172.22.0.188-172.22.0.200
  interface: eth0
  protocol: layer2
status:
  firstIP: 172.22.0.188
  lastIP: 172.22.0.200
  poolSize: 13
  ready: true
  usage: 2
  used:
    172.22.0.188: default/my-service
    172.22.0.189: default/mylbapp-svc-layer2
  v4: true
```



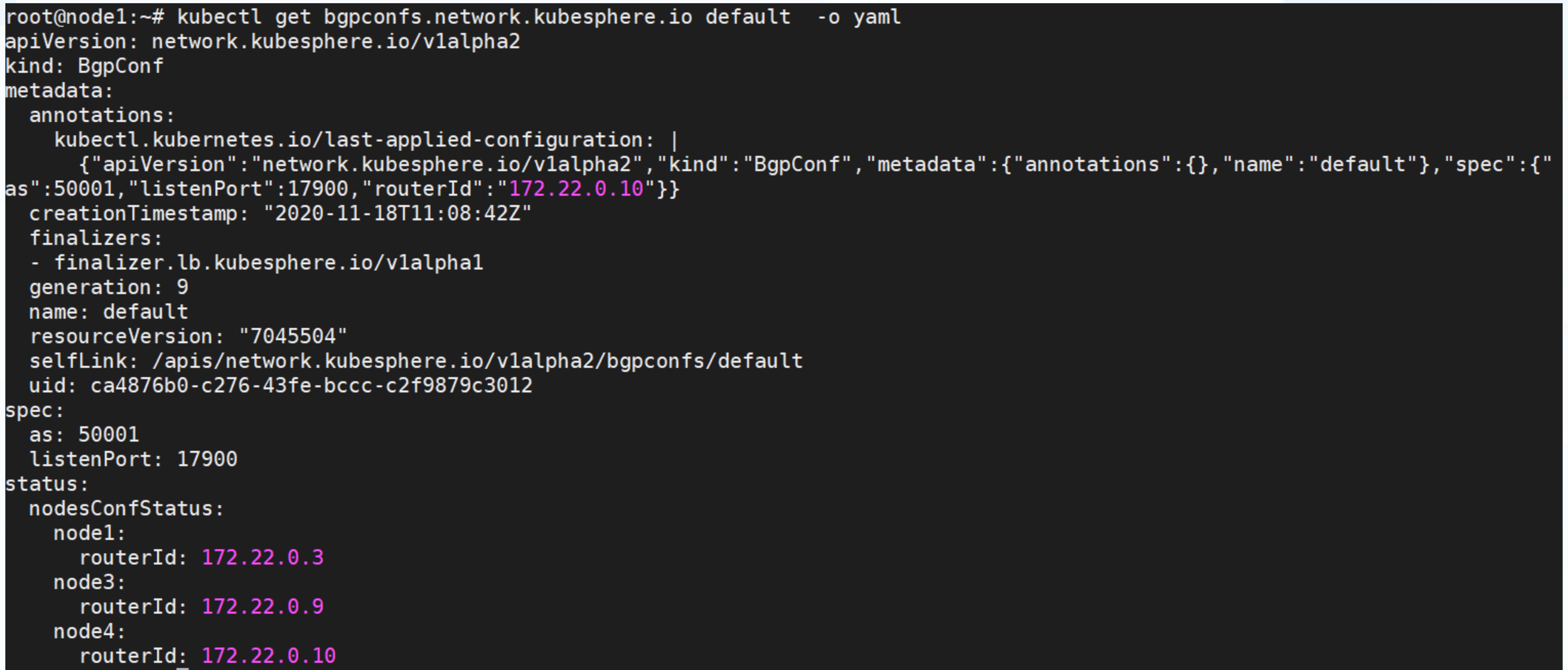
```
root@node1:~# kubectl get eips.network.kubesphere.io eip-sample -o yaml
apiVersion: network.kubesphere.io/v1alpha2
kind: Eip
metadata:
  annotations:
    kubectl.kubernetes.io/last-applied-configuration: |
      {"apiVersion":"network.kubesphere.io/v1alpha2","kind":"Eip","metadata":
ddress":"139.198.121.228","disable":false}}
  creationTimestamp: "2020-11-18T11:08:34Z"
  finalizers:
  - finalizer.ipam.kubesphere.io/v1alpha1
  generation: 2
  name: eip-sample
  resourceVersion: "6988305"
  selfLink: /apis/network.kubesphere.io/v1alpha2/eips/eip-sample
  uid: c32e8b64-21bb-4a68-a27a-7eed4a76c43c
spec:
  address: 139.198.121.228
status:
  firstIP: 139.198.121.228
  lastIP: 139.198.121.228
  occupied: true
  poolSize: 1
  ready: true
  usage: 1
  used:
    139.198.121.228: default/test-svc
  v4: true
```

Config BGP(1/2)

- BgpConf
  - as
  - routerId
- BgpPeer
  - peerAs
  - neighborAddress
- More
  - Config BGP Guide

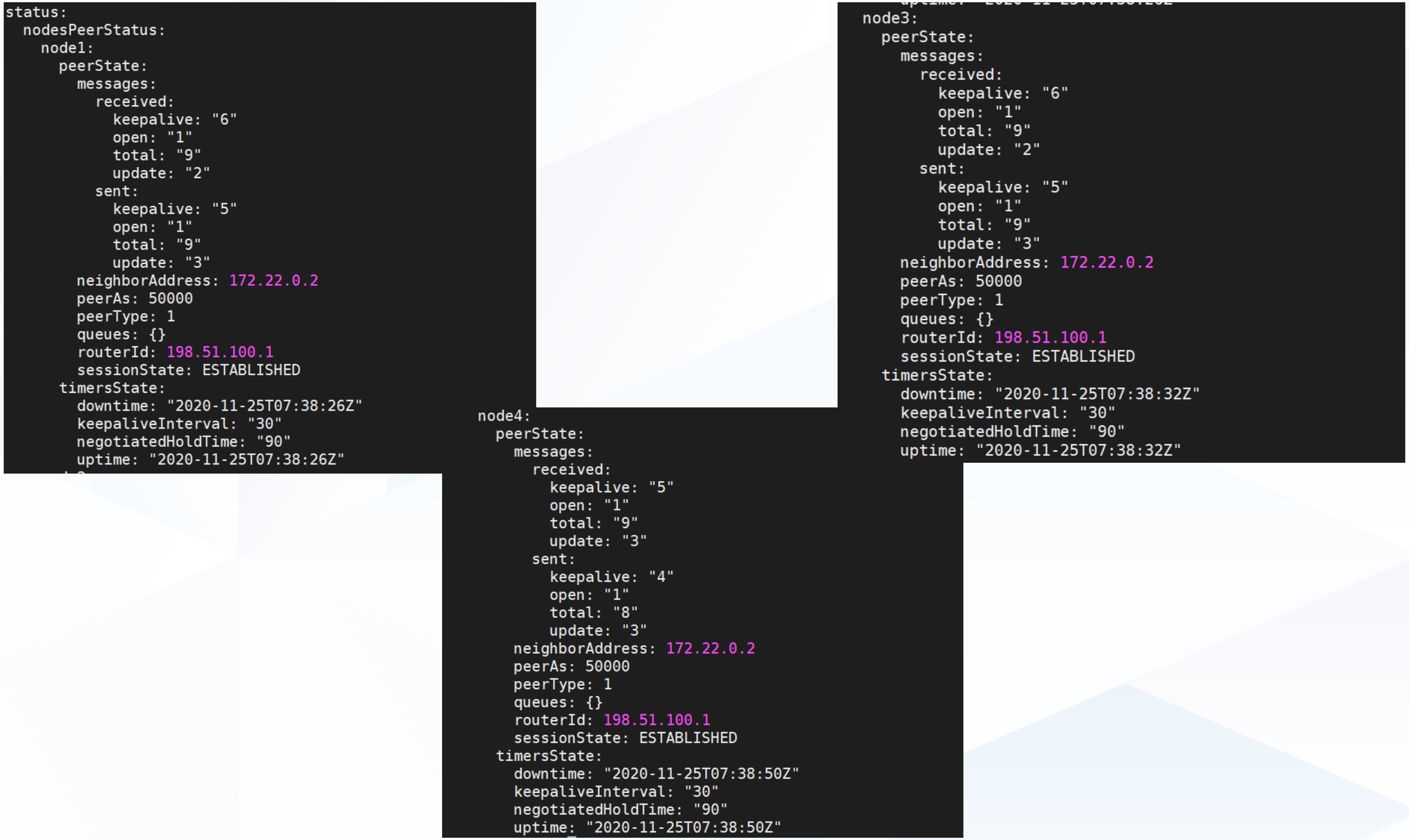


```
root@node1:~# kubectl get bgppeers.network.kubesphere.io bgppeer-sample -o yaml
apiVersion: network.kubesphere.io/v1alpha2
kind: BgpPeer
metadata:
  annotations:
    kubectl.kubernetes.io/last-applied-configuration: |
      {"apiVersion":"network.kubesphere.io/v1alpha2","kind":"BgpPeer","metadata":{"annotations":{},"name":"bgppeer-sample"},"s
pec":{"conf":{"neighborAddress":"172.22.0.2","peerAs":50000}}}
  creationTimestamp: "2020-11-20T09:00:52Z"
  finalizers:
  - finalizer.lb.kubesphere.io/v1alpha1
  generation: 6
  name: bgppeer-sample
  resourceVersion: "7046286"
  selfLink: /apis/network.kubesphere.io/v1alpha2/bgppeers/bgppeer-sample
  uid: 70bdd404-b01a-46ec-a7fe-e307a3fa41e8
spec:
  afiSafis:
  - addPaths:
      config:
        sendMax: 1000
    config:
      enabled: true
      family:
        afi: AFI_IP
        safi: SAFI_UNICAST
  conf:
    neighborAddress: 172.22.0.2
    peerAs: 50000
```



```
root@node1:~# kubectl get bgpconfs.network.kubesphere.io default  -o yaml
apiVersion: network.kubesphere.io/v1alpha2
kind: BgpConf
metadata:
  annotations:
    kubectl.kubernetes.io/last-applied-configuration: |
      {"apiVersion":"network.kubesphere.io/v1alpha2","kind":"BgpConf","metadata":{"annotations":{},"name":"default"},"spec":{"
as":50001,"listenPort":17900,"routerId":"172.22.0.10"}}
  creationTimestamp: "2020-11-18T11:08:42Z"
  finalizers:
  - finalizer.lb.kubesphere.io/v1alpha1
  generation: 9
  name: default
  resourceVersion: "7045504"
  selfLink: /apis/network.kubesphere.io/v1alpha2/bgpconfs/default
  uid: ca4876b0-c276-43fe-bccc-c2f9879c3012
spec:
  as: 50001
  listenPort: 17900
status:
  nodesConfStatus:
    node1:
      routerId: 172.22.0.3
    node3:
      routerId: 172.22.0.9
    node4:
      routerId: 172.22.0.10
```

Config BGP(2/2)



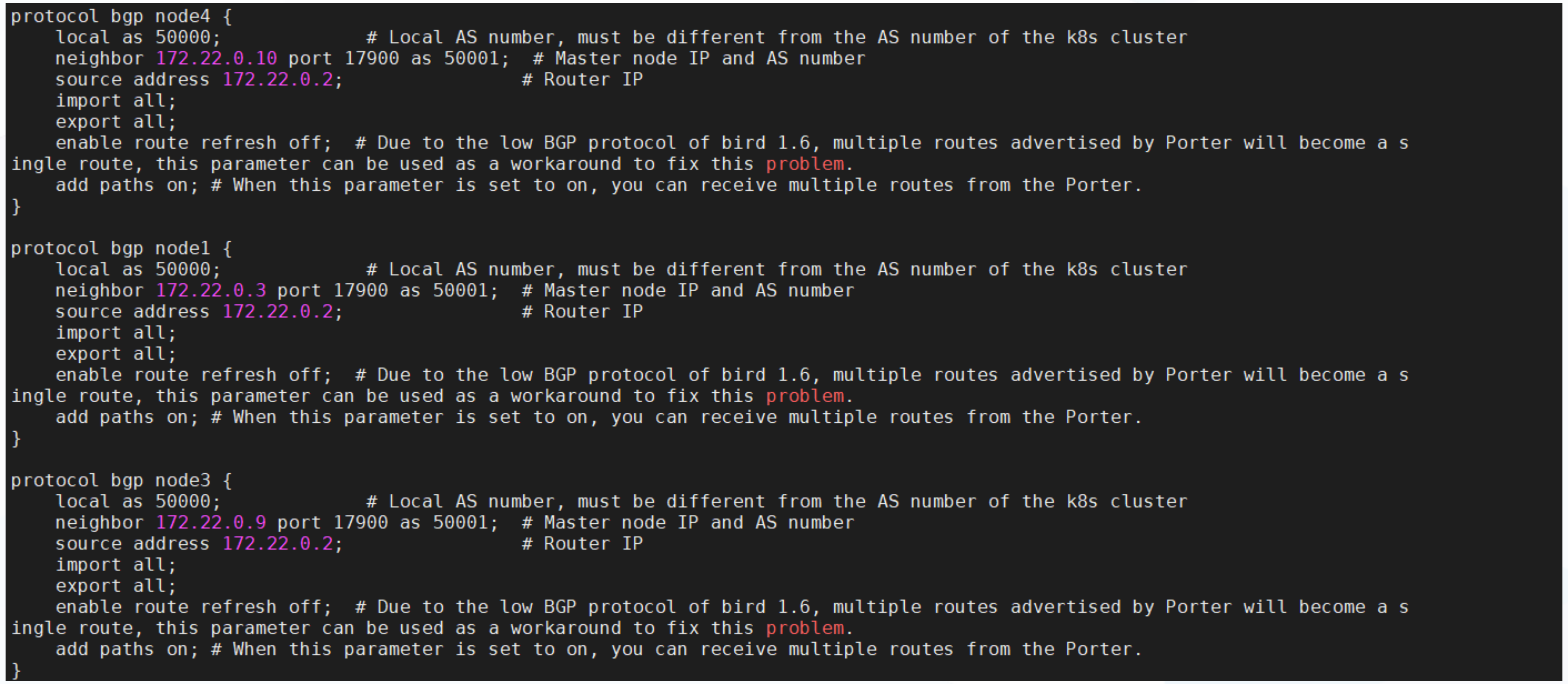
```
status:
  nodesPeerStatus:
    node1:
      peerState:
        messages:
          received:
            keepalive: "6"
            open: "1"
            total: "9"
            update: "2"
          sent:
            keepalive: "5"
            open: "1"
            total: "9"
            update: "3"
        neighborAddress: 172.22.0.2
        peerAs: 50000
        peerType: 1
        queues: {}
        routerId: 198.51.100.1
        sessionState: ESTABLISHED
      timersState:
        downtime: "2020-11-25T07:38:26Z"
        keepaliveInterval: "30"
        negotiatedHoldTime: "90"
        uptime: "2020-11-25T07:38:26Z"
```

```
node4:
  peerState:
    messages:
      received:
        keepalive: "5"
        open: "1"
        total: "9"
        update: "3"
      sent:
        keepalive: "4"
        open: "1"
        total: "8"
        update: "3"
    neighborAddress: 172.22.0.2
    peerAs: 50000
    peerType: 1
    queues: {}
    routerId: 198.51.100.1
    sessionState: ESTABLISHED
  timersState:
    downtime: "2020-11-25T07:38:50Z"
    keepaliveInterval: "30"
    negotiatedHoldTime: "90"
    uptime: "2020-11-25T07:38:50Z"
```

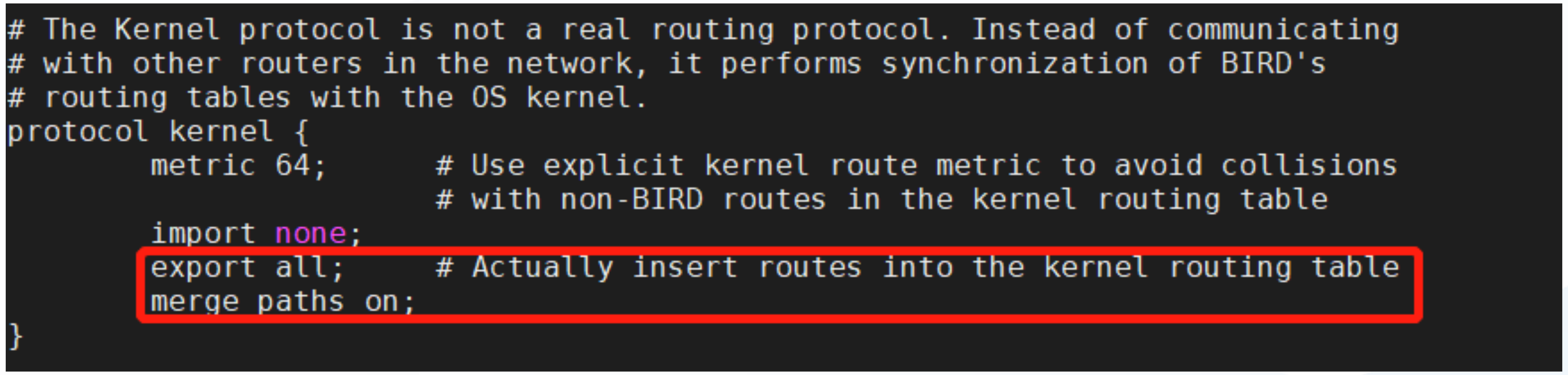
```
node3:
  peerState:
    messages:
      received:
        keepalive: "6"
        open: "1"
        total: "9"
        update: "2"
      sent:
        keepalive: "5"
        open: "1"
        total: "9"
        update: "3"
    neighborAddress: 172.22.0.2
    peerAs: 50000
    peerType: 1
    queues: {}
    routerId: 198.51.100.1
    sessionState: ESTABLISHED
  timersState:
    downtime: "2020-11-25T07:38:32Z"
    keepaliveInterval: "30"
    negotiatedHoldTime: "90"
    uptime: "2020-11-25T07:38:32Z"
```

Config Bird

- Install Bird
  - $sudo add-apt-repository ppa:cz.nic-labs/bird
  - $sudo apt-get update
  - $sudo apt-get install bird
  - $sudo systemctl enable bird
  - $sudo systemctl restart bird
- Config Protocol BGP
- Config Protocol kernel

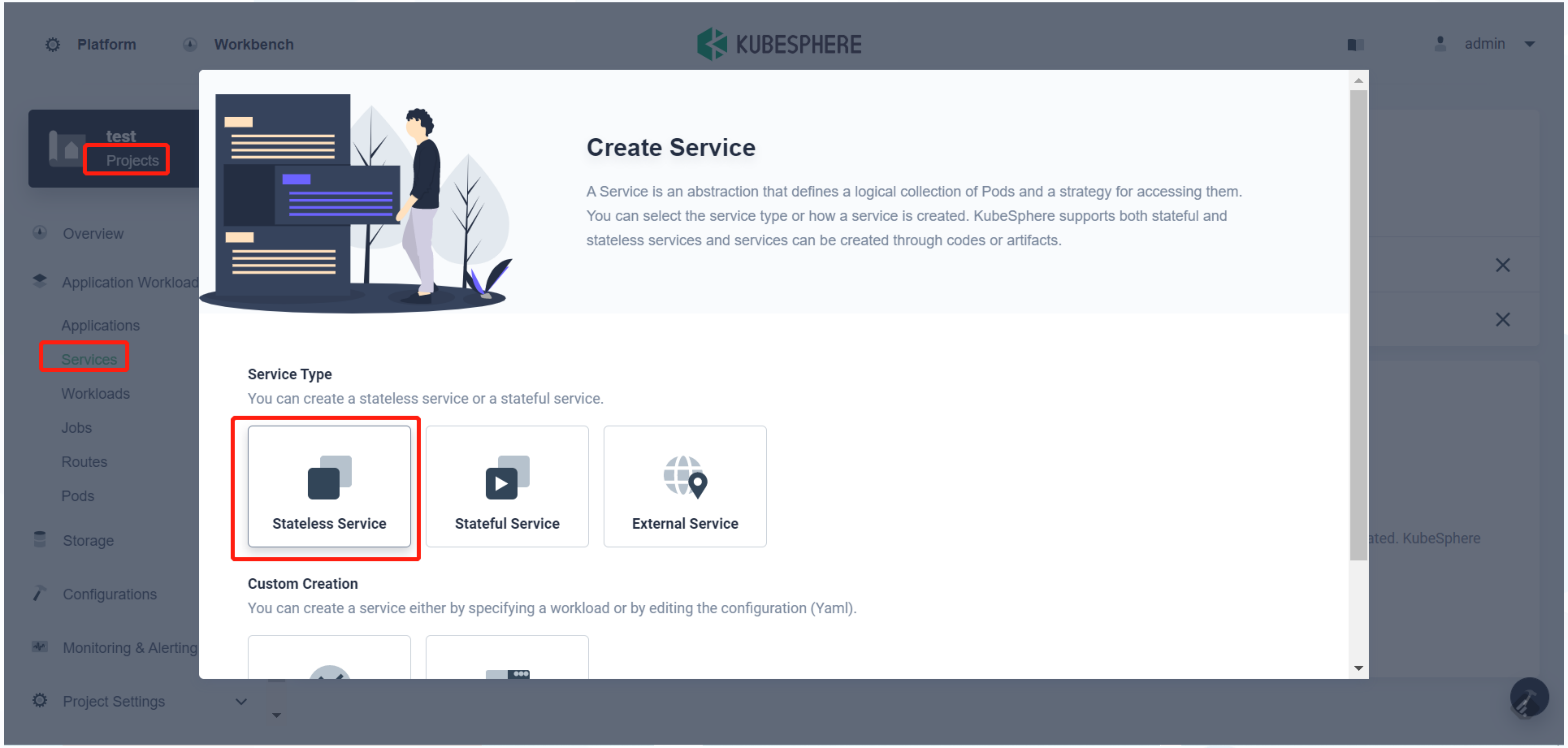


```
protocol bgp node4 {
    local as 50000;             # Local AS number, must be different from the AS number of the k8s cluster
    neighbor 172.22.0.10 port 17900 as 50001;  # Master node IP and AS number
    source address 172.22.0.2;              # Router IP
    import all;
    export all;
    enable route refresh off;  # Due to the low BGP protocol of bird 1.6, multiple routes advertised by Porter will become a s
ingle route, this parameter can be used as a workaround to fix this problem.
    add paths on; # When this parameter is set to on, you can receive multiple routes from the Porter.
}

protocol bgp node1 {
    local as 50000;             # Local AS number, must be different from the AS number of the k8s cluster
    neighbor 172.22.0.3 port 17900 as 50001;  # Master node IP and AS number
    source address 172.22.0.2;              # Router IP
    import all;
    export all;
    enable route refresh off;  # Due to the low BGP protocol of bird 1.6, multiple routes advertised by Porter will become a s
ingle route, this parameter can be used as a workaround to fix this problem.
    add paths on; # When this parameter is set to on, you can receive multiple routes from the Porter.
}

protocol bgp node3 {
    local as 50000;             # Local AS number, must be different from the AS number of the k8s cluster
    neighbor 172.22.0.9 port 17900 as 50001;  # Master node IP and AS number
    source address 172.22.0.2;              # Router IP
    import all;
    export all;
    enable route refresh off;  # Due to the low BGP protocol of bird 1.6, multiple routes advertised by Porter will become a s
ingle route, this parameter can be used as a workaround to fix this problem.
    add paths on; # When this parameter is set to on, you can receive multiple routes from the Porter.
}
```



```
# The Kernel protocol is not a real routing protocol. Instead of communicating
# with other routers in the network, it performs synchronization of BIRD's
# routing tables with the OS kernel.
protocol kernel {
        metric 64;      # Use explicit kernel route metric to avoid collisions
                        # with non-BIRD routes in the kernel routing table
        import none;
        export all;     # Actually insert routes into the kernel routing table
        merge paths on;
}
```

Create Service With KubeSphere(1/3)



Create Service

A Service is an abstraction that defines a logical collection of Pods and a strategy for accessing them. You can select the service type or how a service is created. KubeSphere supports both stateful and stateless services and services can be created through codes or artifacts.

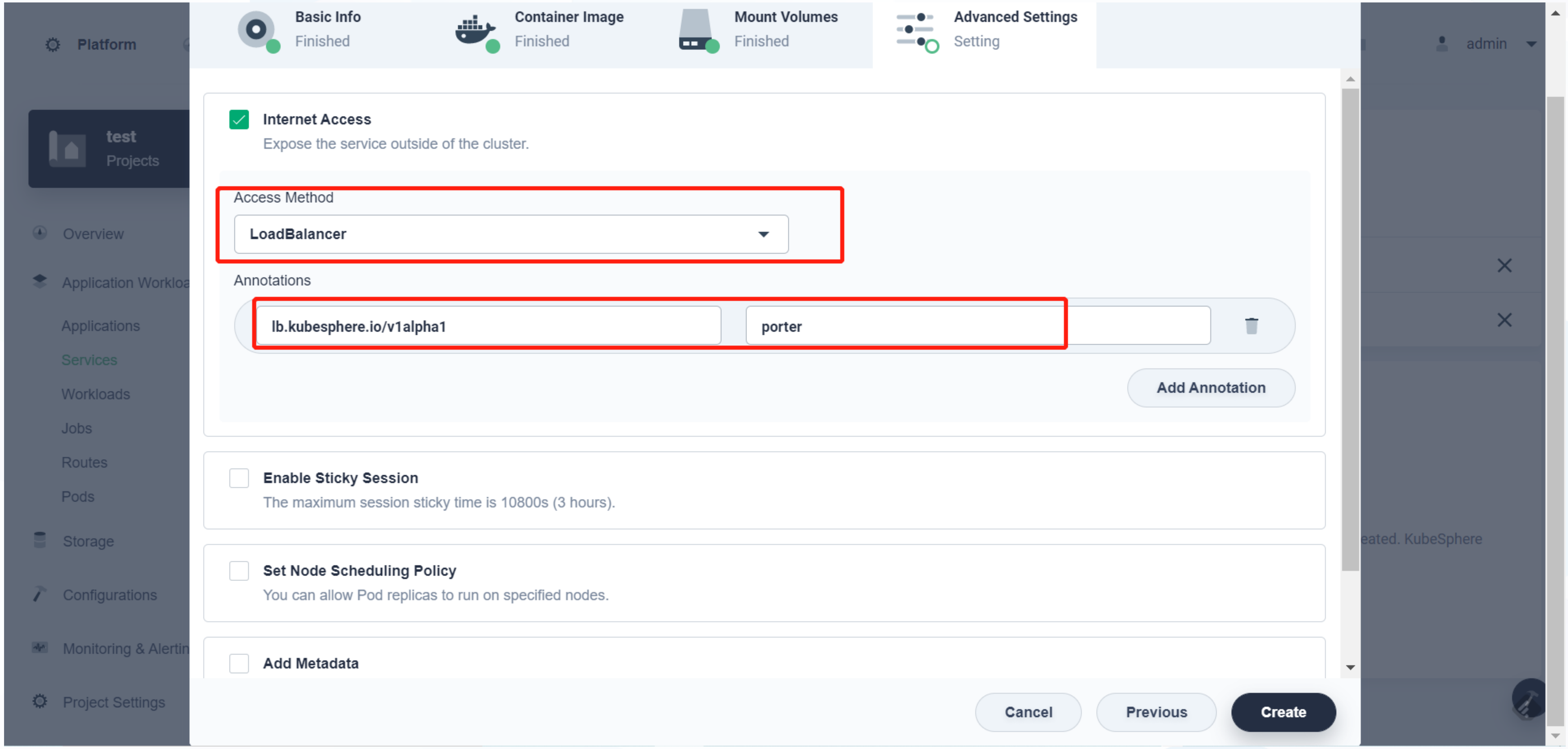
Service Type

You can create a stateless service or a stateful service.

Custom Creation

You can create a service either by specifying a workload or by editing the configuration (Yaml).

Create Service With KubeSphere(2/3)



☑ **Internet Access**

Expose the service outside of the cluster.

Access Method: LoadBalancer

Annotations: lb.kubesphere.io/v1alpha1 = porter

□ **Enable Sticky Session**

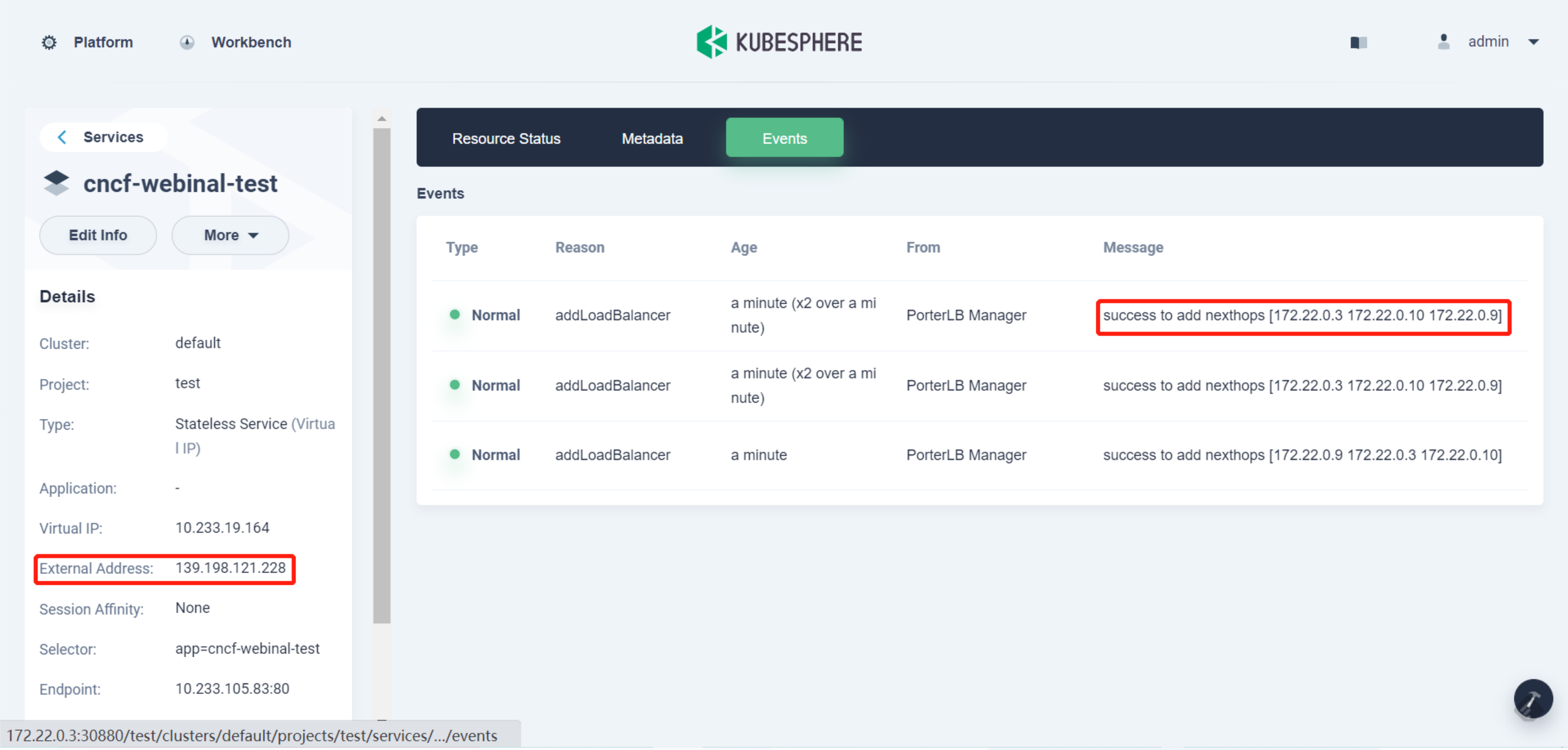
The maximum session sticky time is 10800s (3 hours).

□ **Set Node Scheduling Policy**

You can allow Pod replicas to run on specified nodes.

□ **Add Metadata**

Create Service With KubeSphere(3/3)



cncf-webinal-test

Details

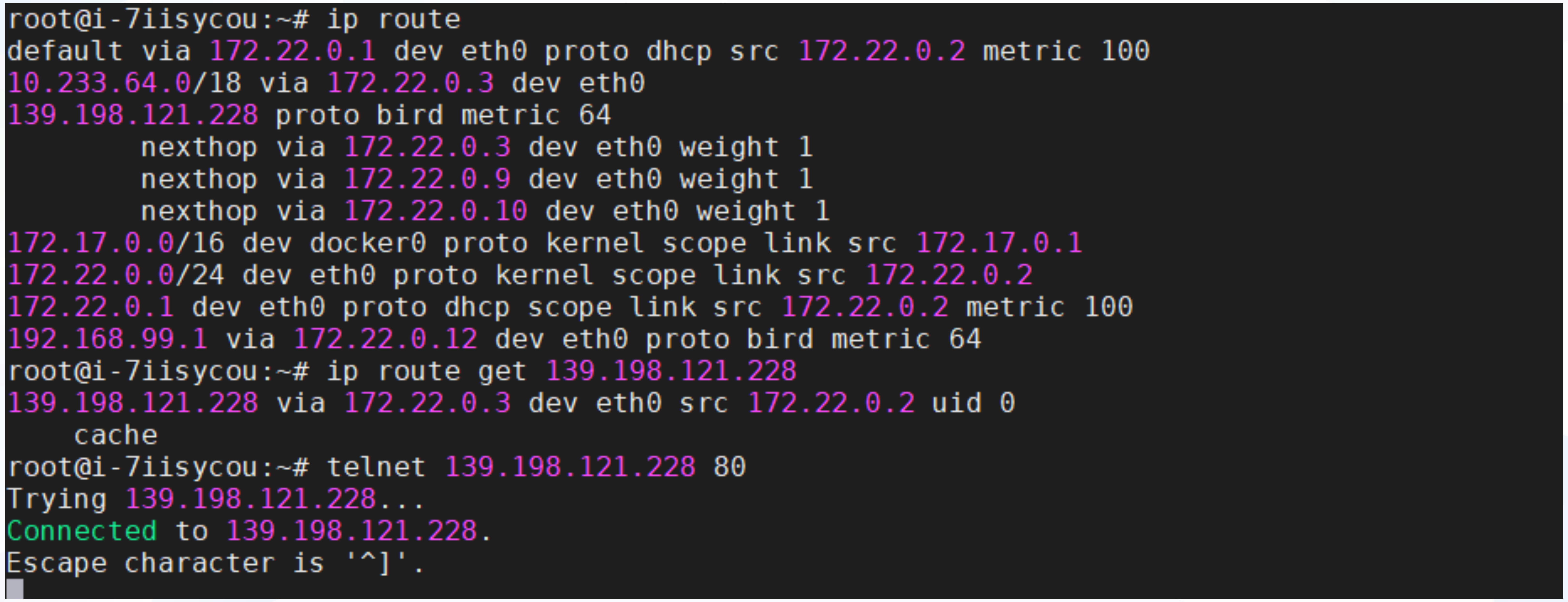
Cluster:	default
Project:	test
Type:	Stateless Service (Virtual IP)
Application:	-
Virtual IP:	10.233.19.164
External Address:	139.198.121.228
Session Affinity:	None
Selector:	app=cncf-webinal-test
Endpoint:	10.233.105.83:80

Events

Type	Reason	Age	From	Message
● Normal	addLoadBalancer	a minute (x2 over a minute)	PorterLB Manager	success to add nexthops [172.22.0.3 172.22.0.10 172.22.0.9]
● Normal	addLoadBalancer	a minute (x2 over a minute)	PorterLB Manager	success to add nexthops [172.22.0.3 172.22.0.10 172.22.0.9]
● Normal	addLoadBalancer	a minute	PorterLB Manager	success to add nexthops [172.22.0.9 172.22.0.3 172.22.0.10]

172.22.0.3:30880/test/clusters/default/projects/test/services/.../events

Verify Result



```
root@i-7iisycou:~# ip route
default via 172.22.0.1 dev eth0 proto dhcp src 172.22.0.2 metric 100
10.233.64.0/18 via 172.22.0.3 dev eth0
139.198.121.228 proto bird metric 64
        nexthop via 172.22.0.3 dev eth0 weight 1
        nexthop via 172.22.0.9 dev eth0 weight 1
        nexthop via 172.22.0.10 dev eth0 weight 1
172.17.0.0/16 dev docker0 proto kernel scope link src 172.17.0.1
172.22.0.0/24 dev eth0 proto kernel scope link src 172.22.0.2
172.22.0.1 dev eth0 proto dhcp scope link src 172.22.0.2 metric 100
192.168.99.1 via 172.22.0.12 dev eth0 proto bird metric 64
root@i-7iisycou:~# ip route get 139.198.121.228
139.198.121.228 via 172.22.0.3 dev eth0 src 172.22.0.2 uid 0
    cache
root@i-7iisycou:~# telnet 139.198.121.228 80
Trying 139.198.121.228...
Connected to 139.198.121.228.
Escape character is '^]'.
```

More

- Specify Protocol
  - protocol.porter.kubesphere.io/v1alpha1: bgp
  - protocol.porter.kubesphere.io/v1alpha1: layer2
- Specify EIP
- Share Eip
- Service Config Guide

References

- https://lwz322.github.io/2019/11/03/ECMP.html
- https://support.huawei.com/enterprise/it/doc/EDOC1100125816/822c6727/ecmp-load-balancing-consistency

KUBESPHERE x CLOUD NATIVE COMPUTING FOUNDATION

Follow us!

github.com/kubesphere

twitter.com/kubesphere

KubeSphere

https://kubesphere.com.cn