
DigitalOcean

Containers and Anycast IPs

Andrew Sy Kim - Software Engineer at DigitalOcean

@a_sykim

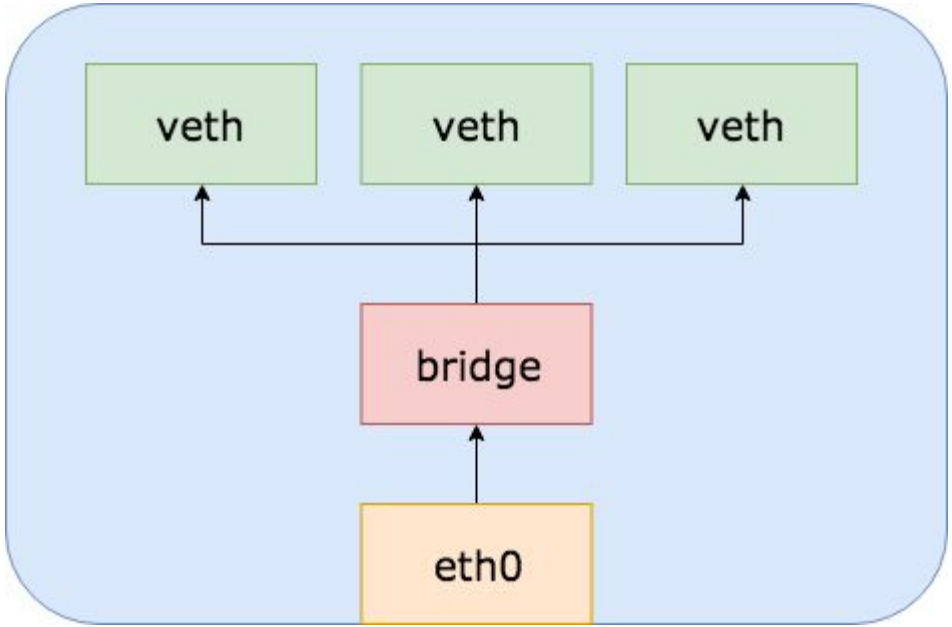
@andrewsykim

Agenda

- Container Networking
- Kubernetes Networking
- Data center Networking
- Anycast Routing/IPs with Kubernetes

Container Networking

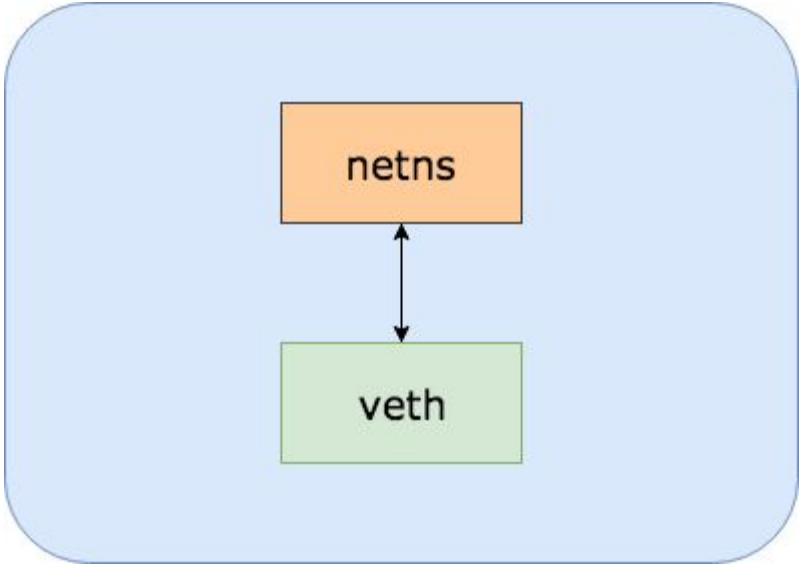
Container Networking



Container Networking

```
$ ip route
default via 10.0.0.1 dev eth0
172.20.100.0/25 dev kube-bridge proto kernel scope
link src 172.20.100.1
```

Container Networking

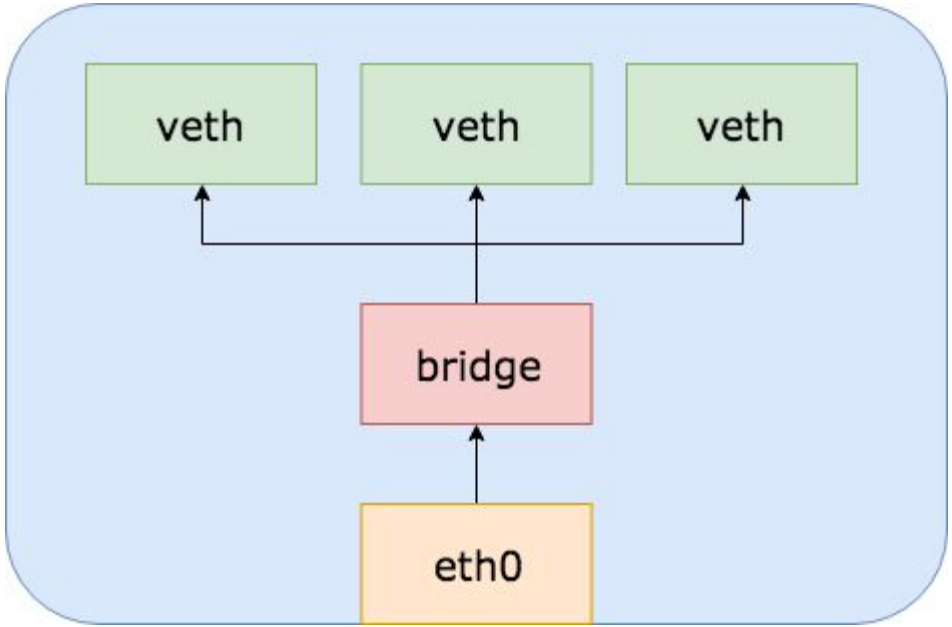


Container Networking

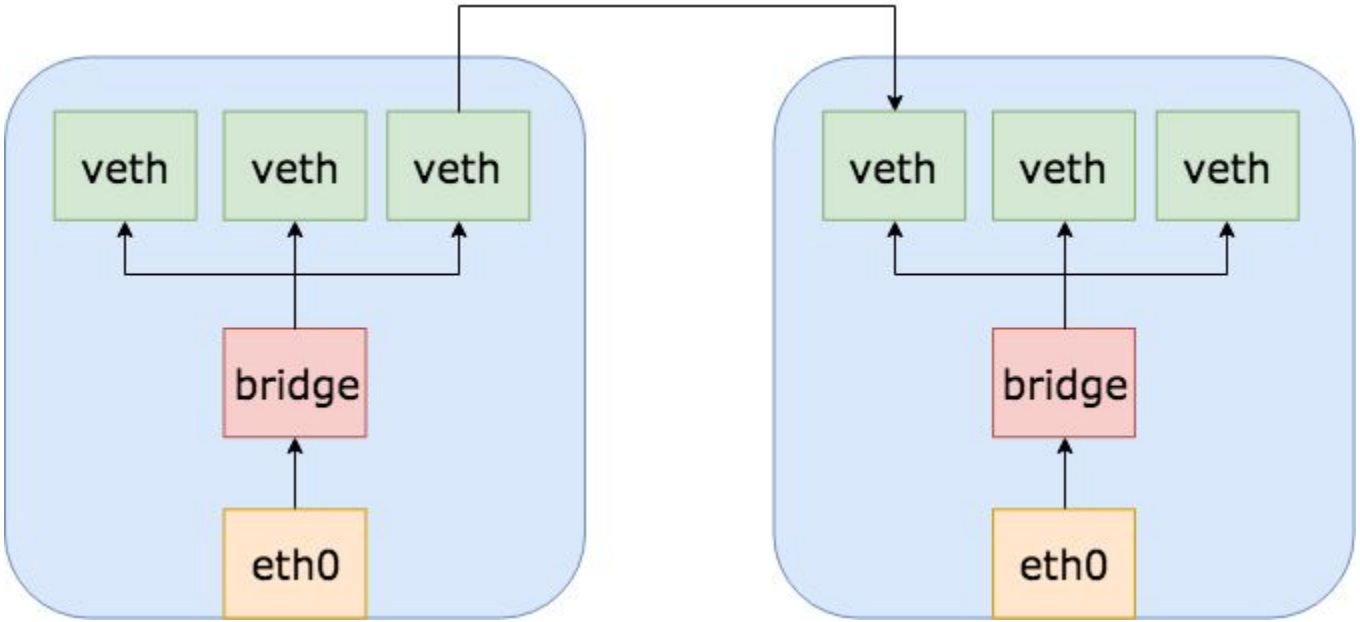
veth

```
$ ip link add vethA type veth peer
name vethB
$ ip netns add mynetns
$ ip link set vethA netns mynetns
```

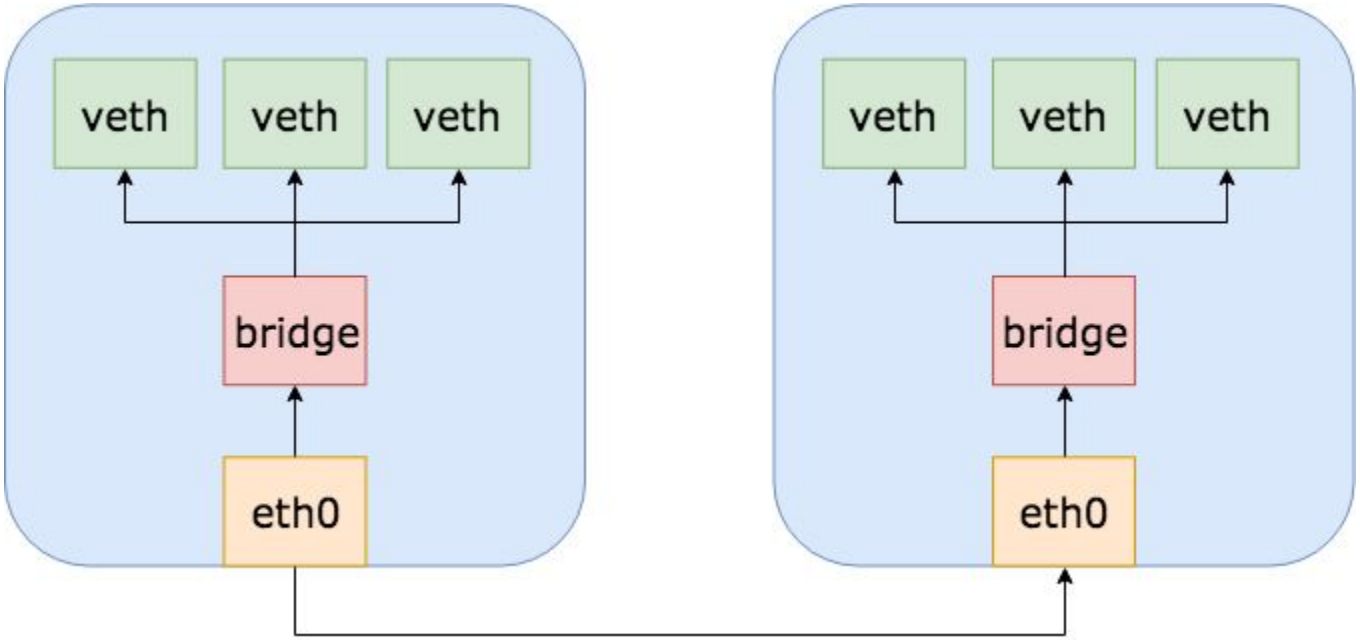
Container Networking



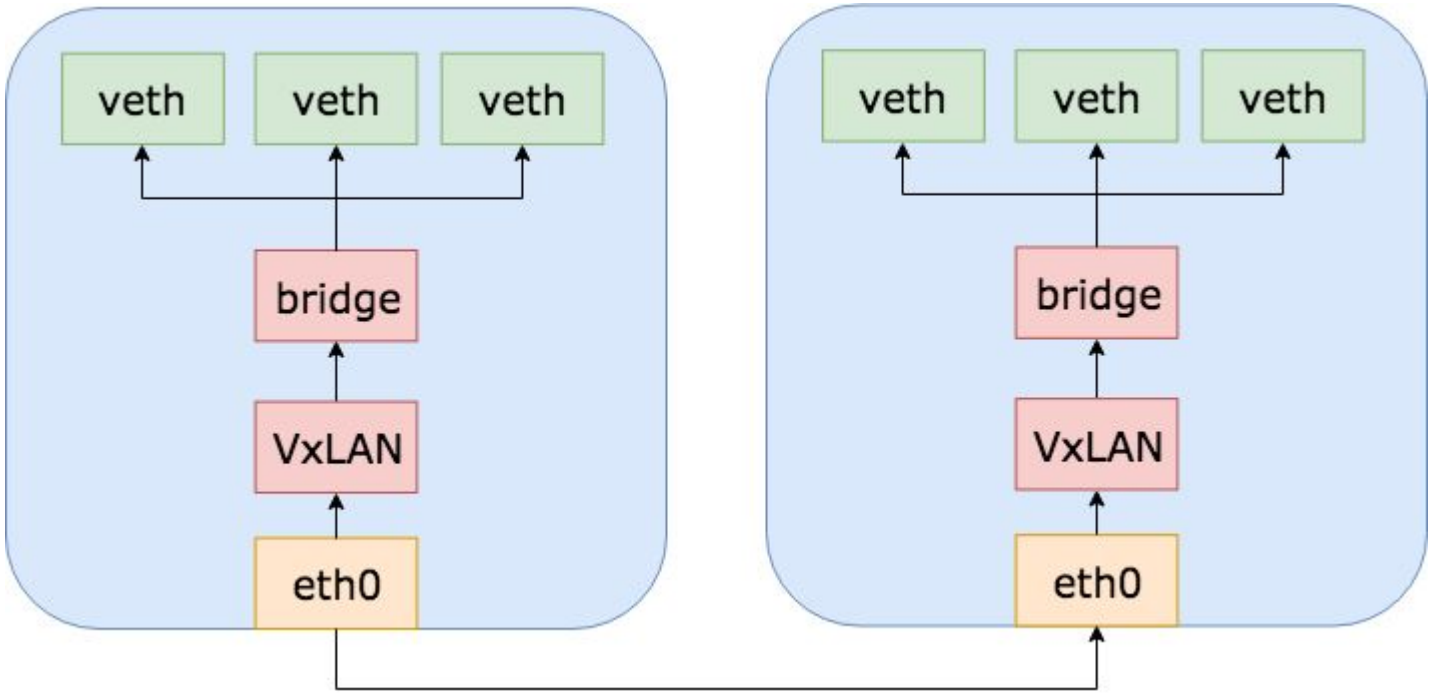
Container Networking



Container Networking



Container Networking



Kubernetes Primitives

Pods

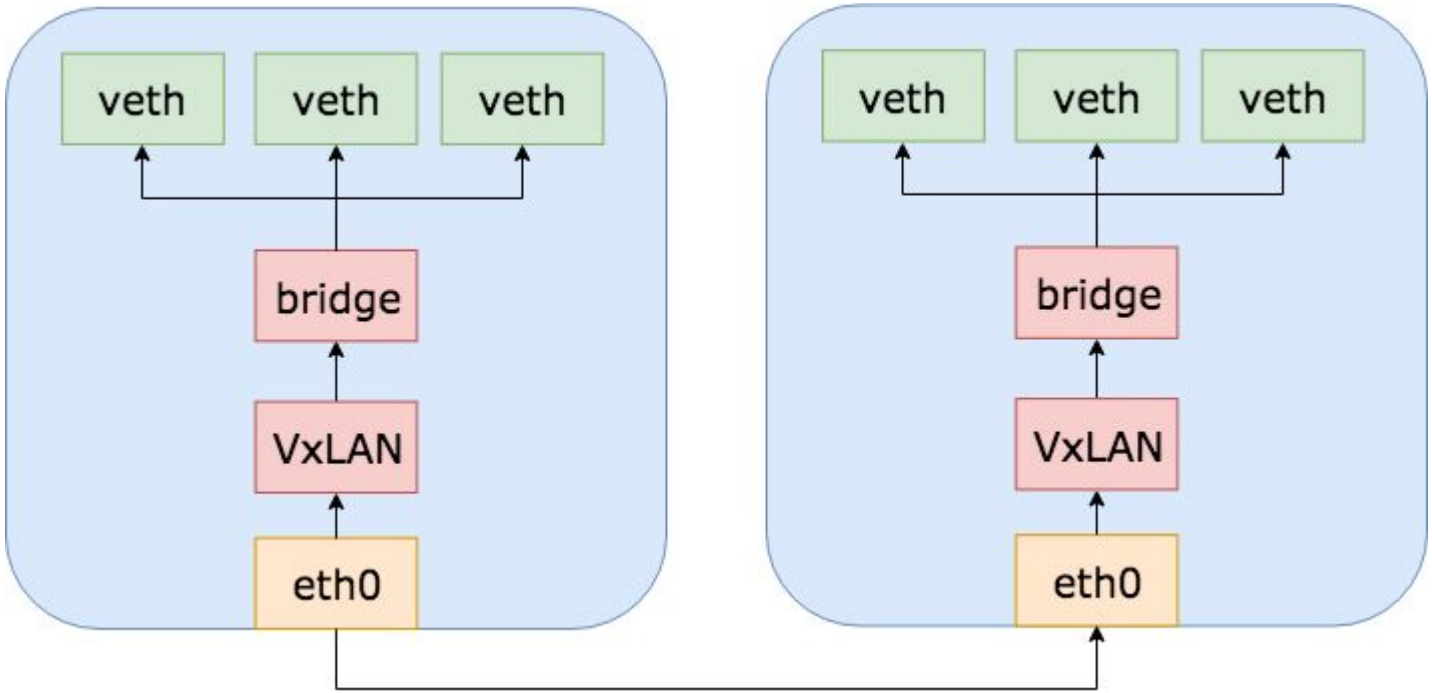
A pod is a group of one or more containers, with shared storage/network, and a specification for how to run the containers.

Services

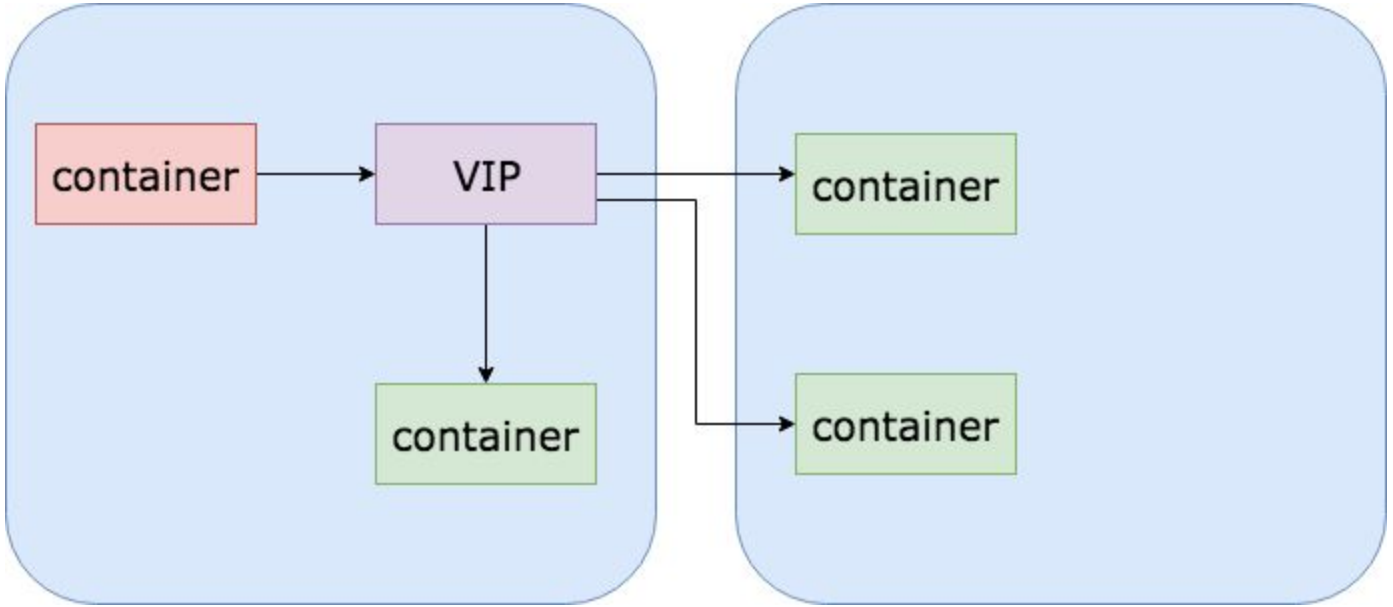
A Kubernetes Service is an abstraction which defines a logical set of Pods and a policy by which to access them.

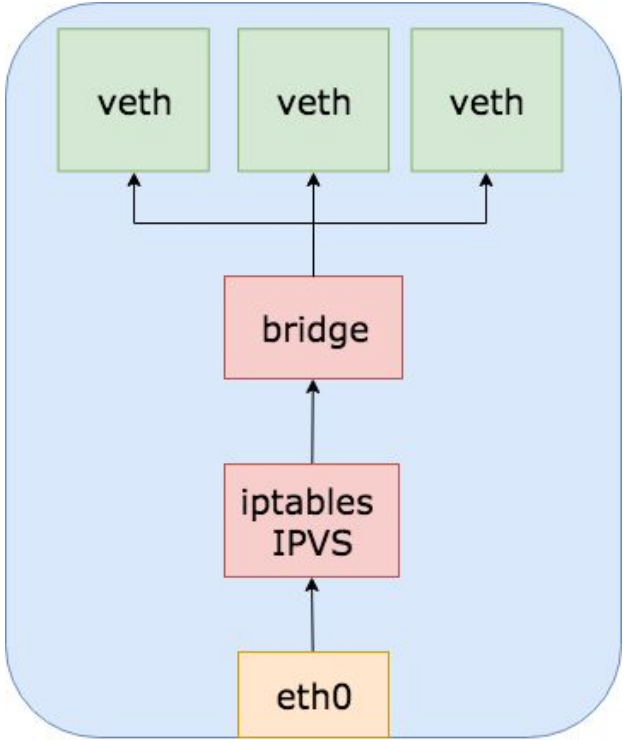
Kubernetes Networking

Pod to Pod


veth
veth
veth
bridge
VxLAN
eth0
veth
veth
veth
bridge
VxLAN
eth0

Services





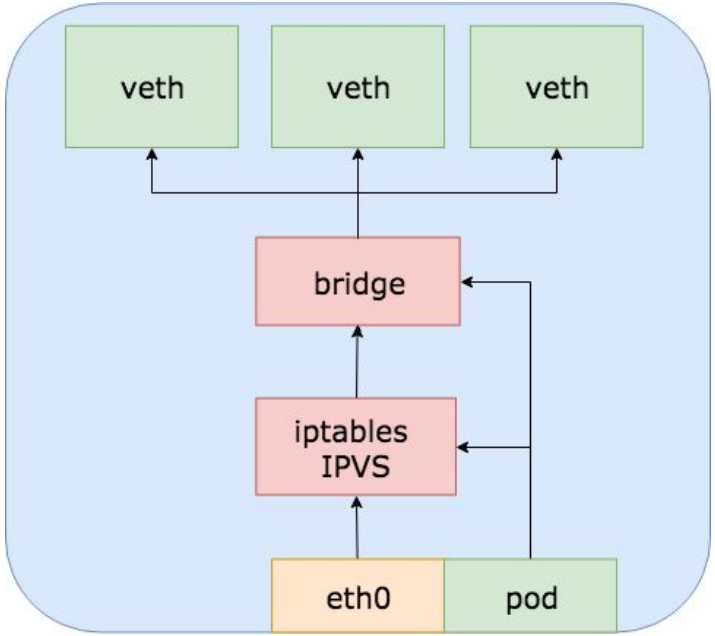
Iptables:

```
$ iptables-save
...
-A KUBE-SERVICES -d 10.38.3.1/32 -p tcp -m comment
--comment "default/kubernetes:https cluster IP" -m
tcp --dport 443 -j KUBE-SVC-NPX46M4PTMTKRN6Y
...
```

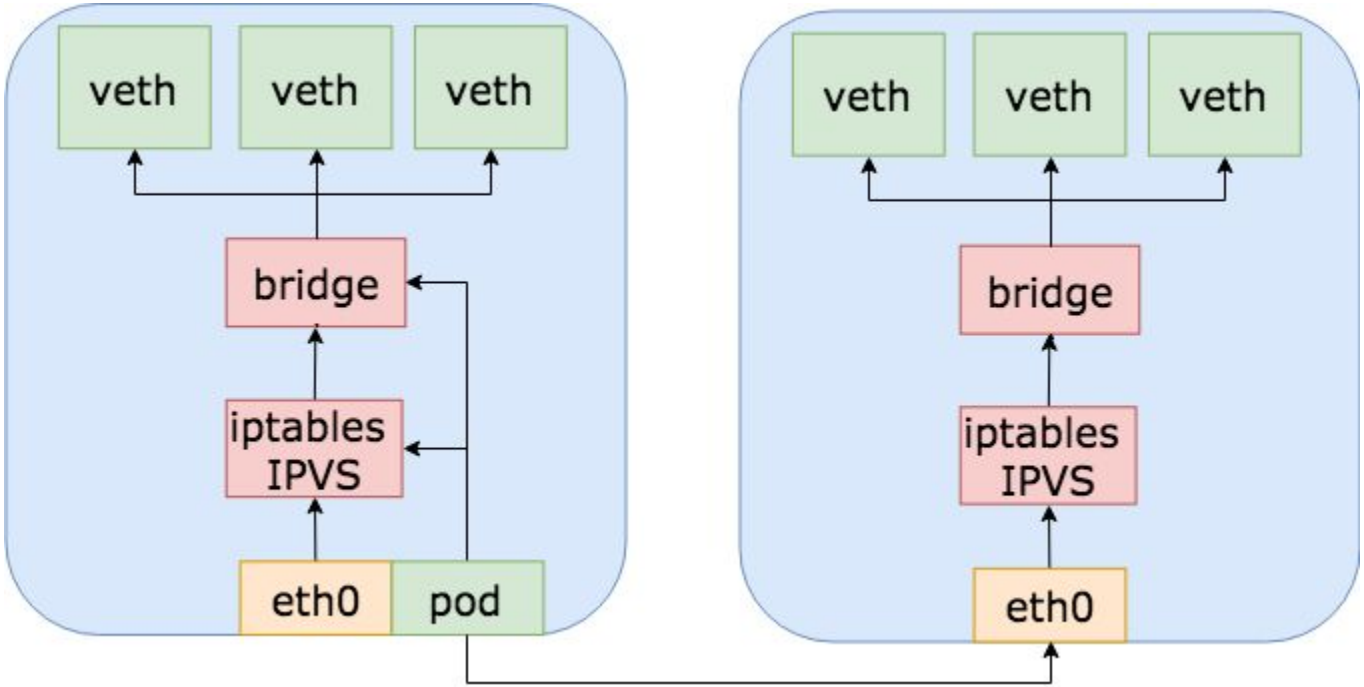
IPVS:

```
$ sudo ipvsadm -L -t 10.38.3.1:443
Prot LocalAddress:Port Scheduler Flags
  -> RemoteAddress:Port           Forward Weight
ActiveConn InActConn
TCP  worker01                     rr
  -> 100.96.1.3:domain            Masq    1      0
0
  -> 100.96.1.4:domain            Masq    1      0
0```
```

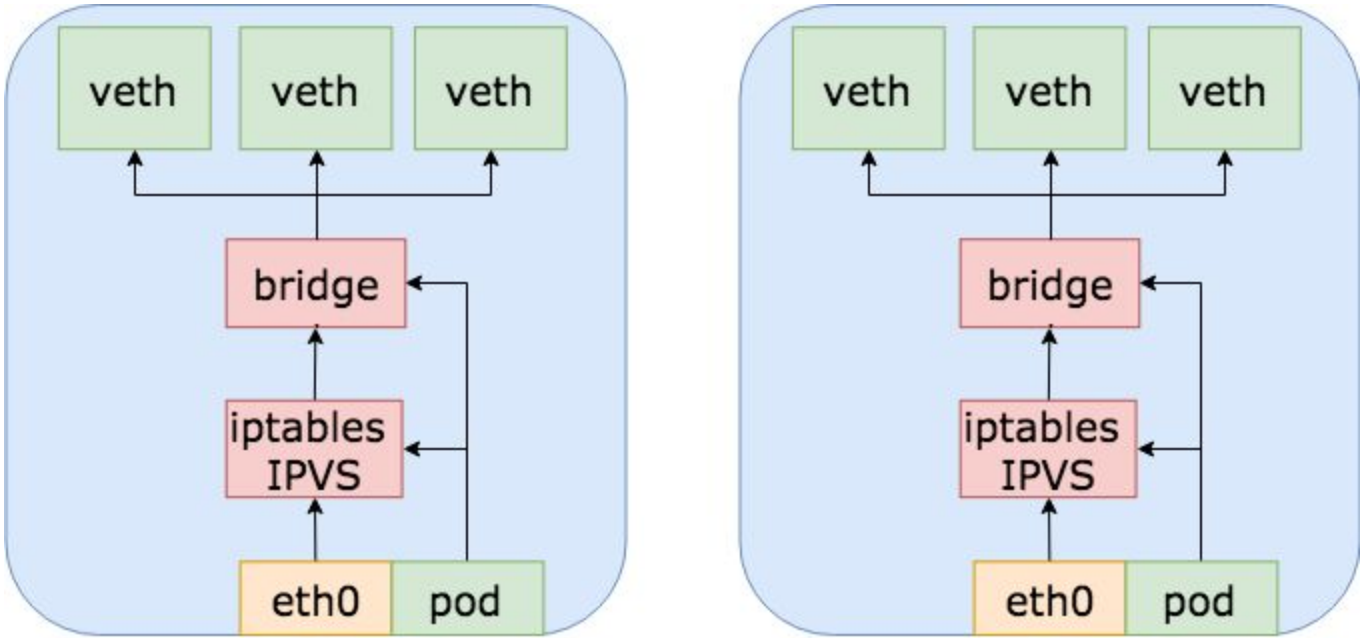
External Traffic



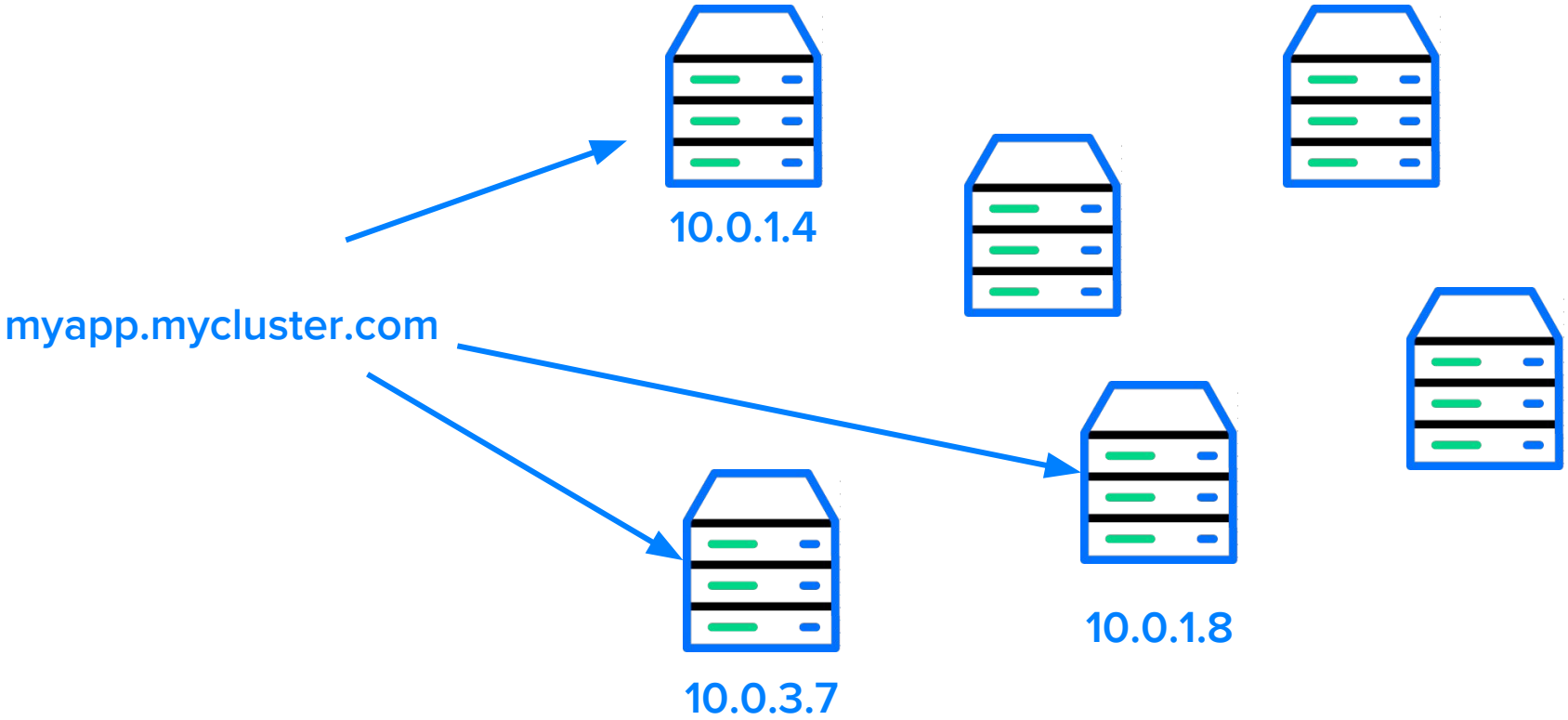
External Traffic



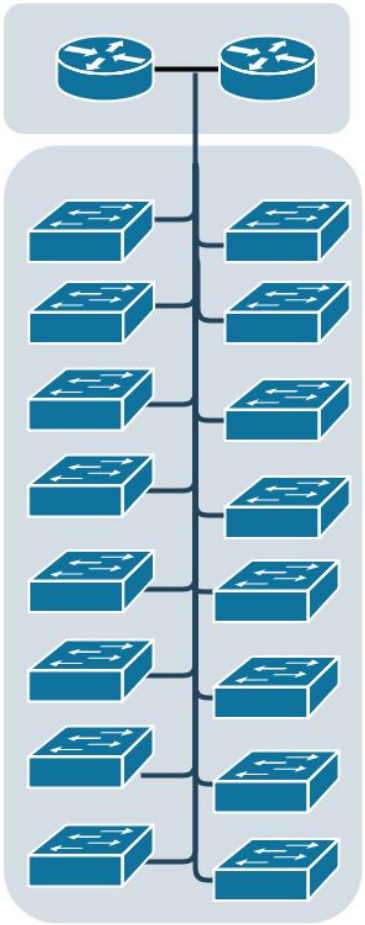
External Traffic

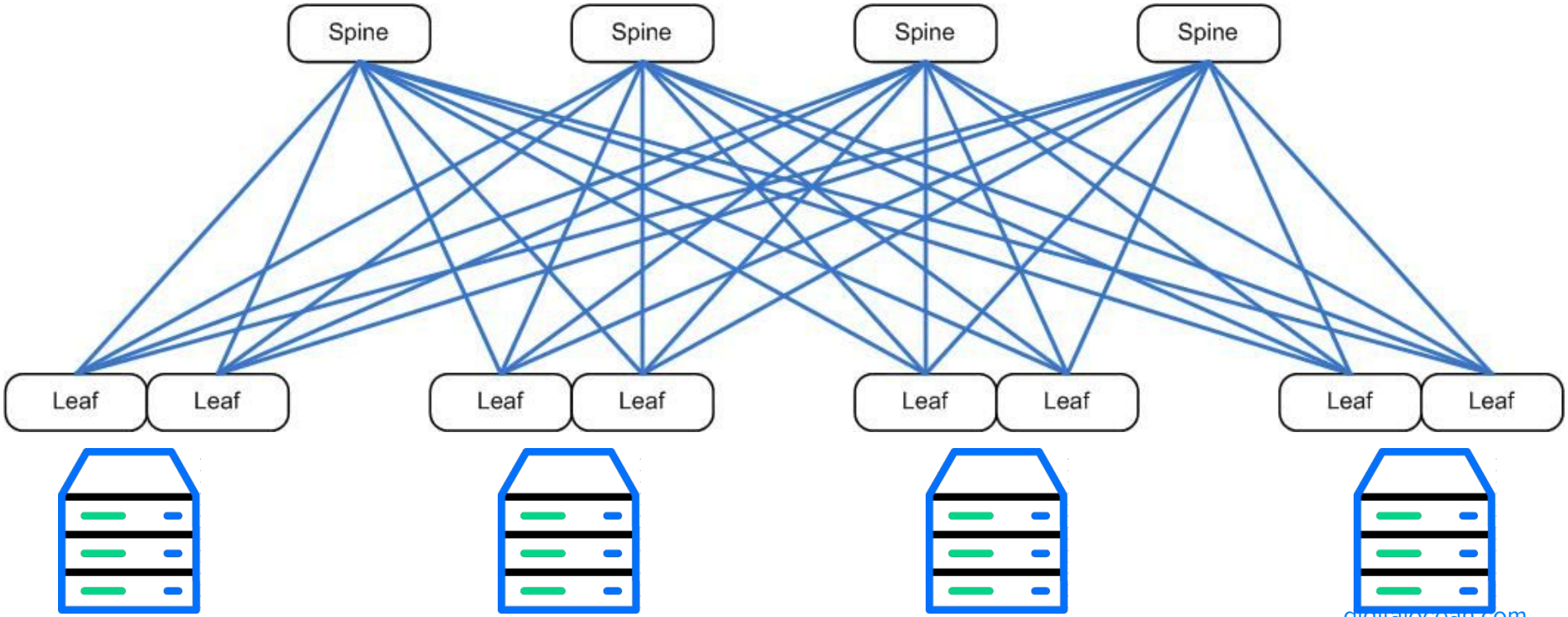


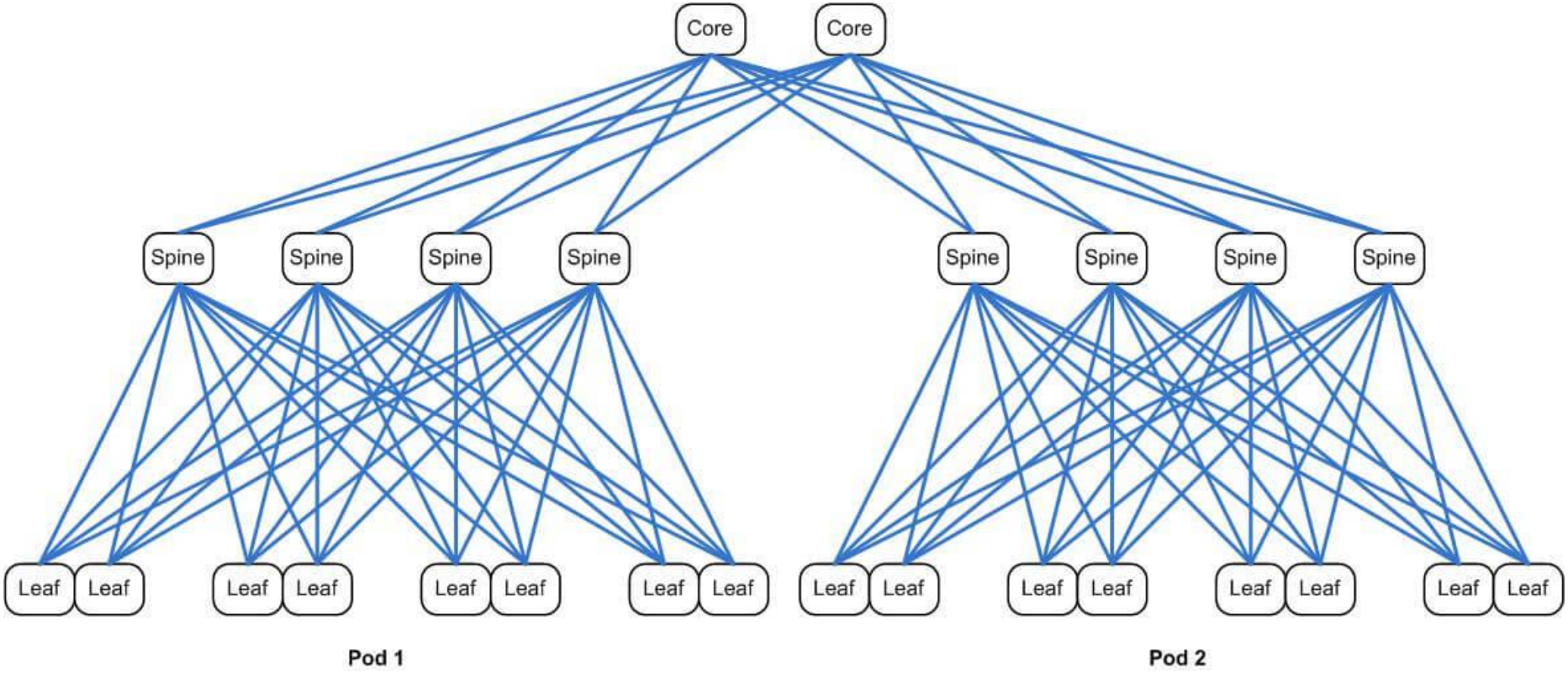
External Traffic into Kubernetes Cluster

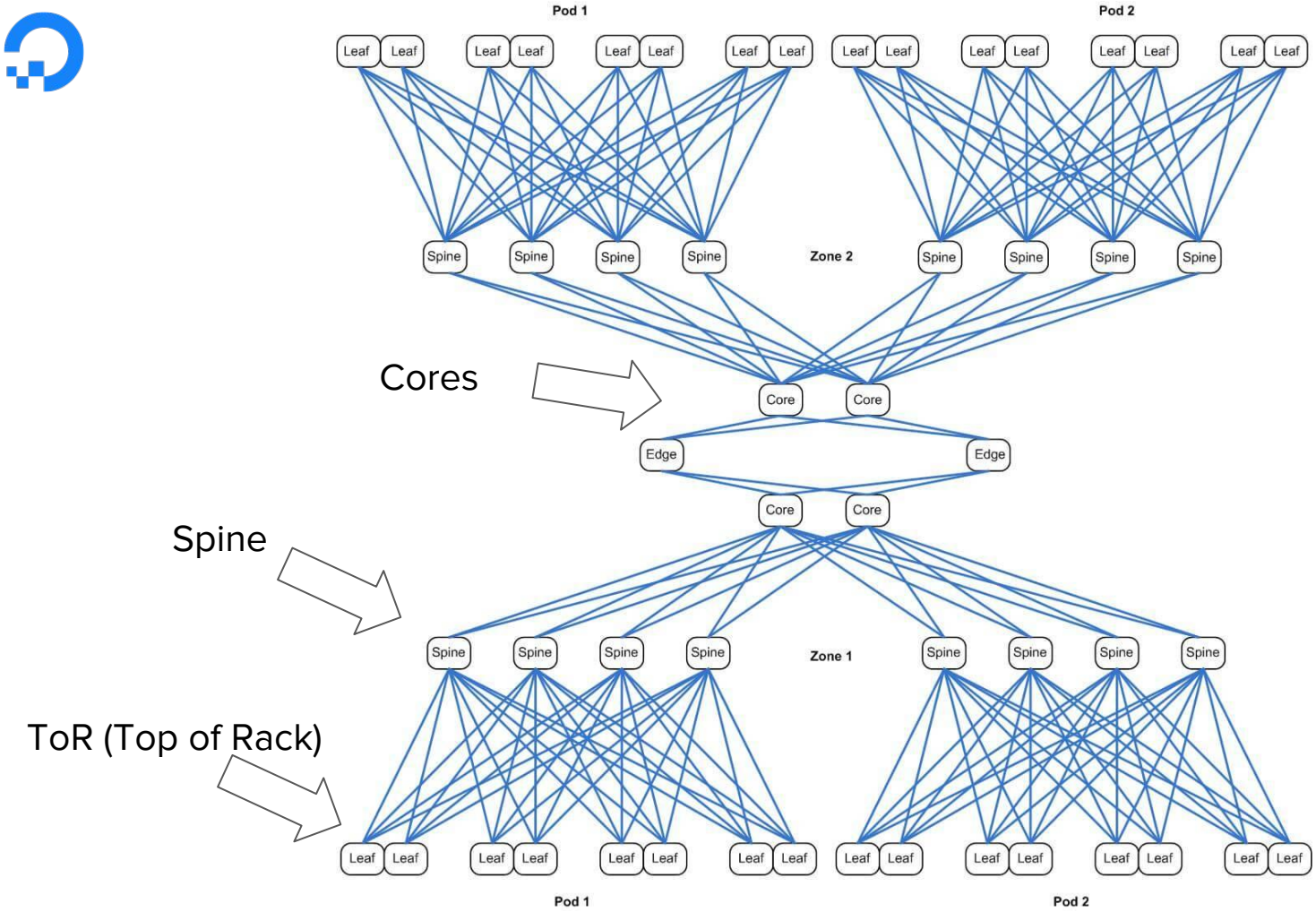


Data Center Networking




Spine
Spine
Spine
Spine
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf


Core
Core
Spine
Spine
Spine
Spine
Spine
Spine
Spine
Spine
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Leaf
Pod 1
Pod 2

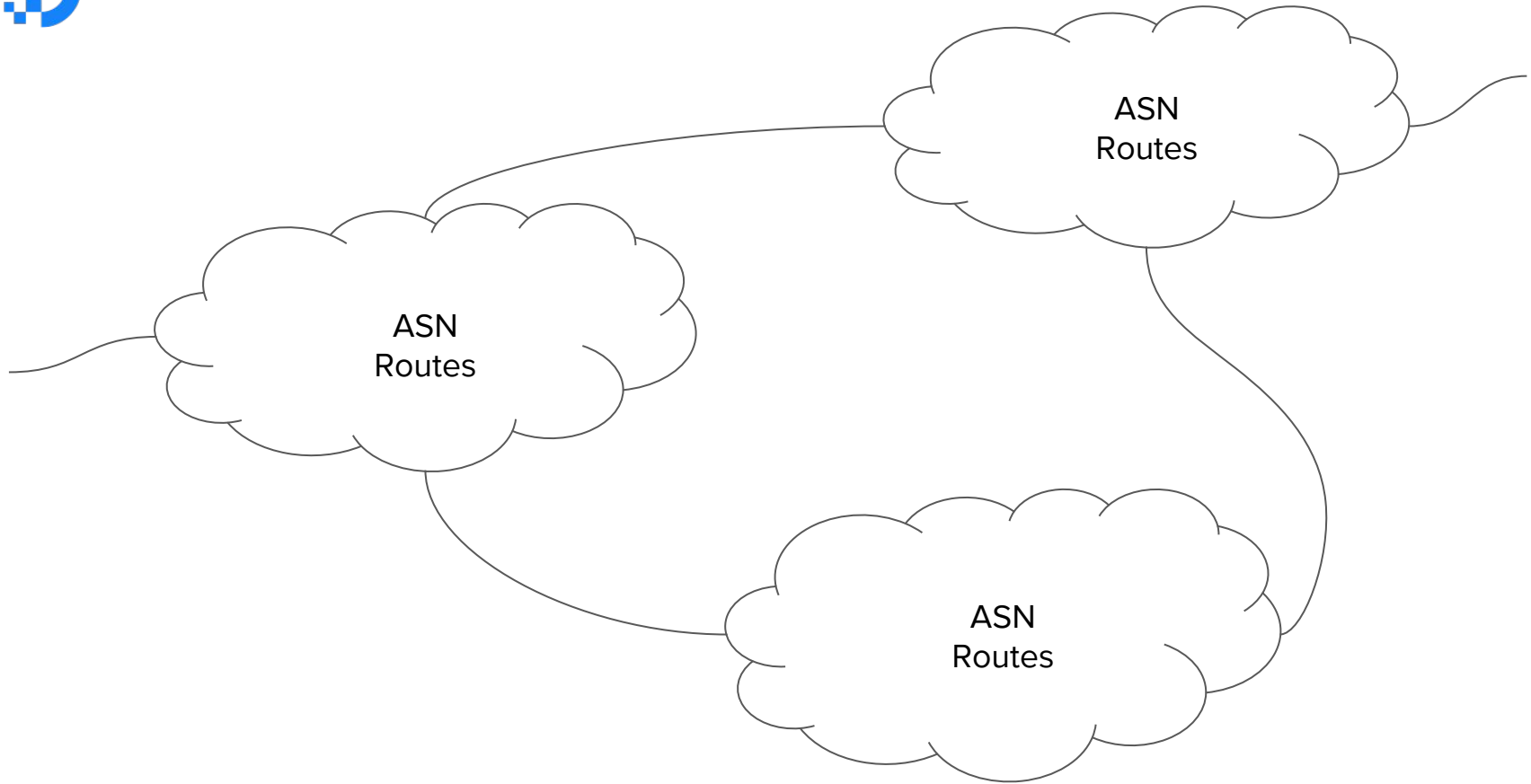


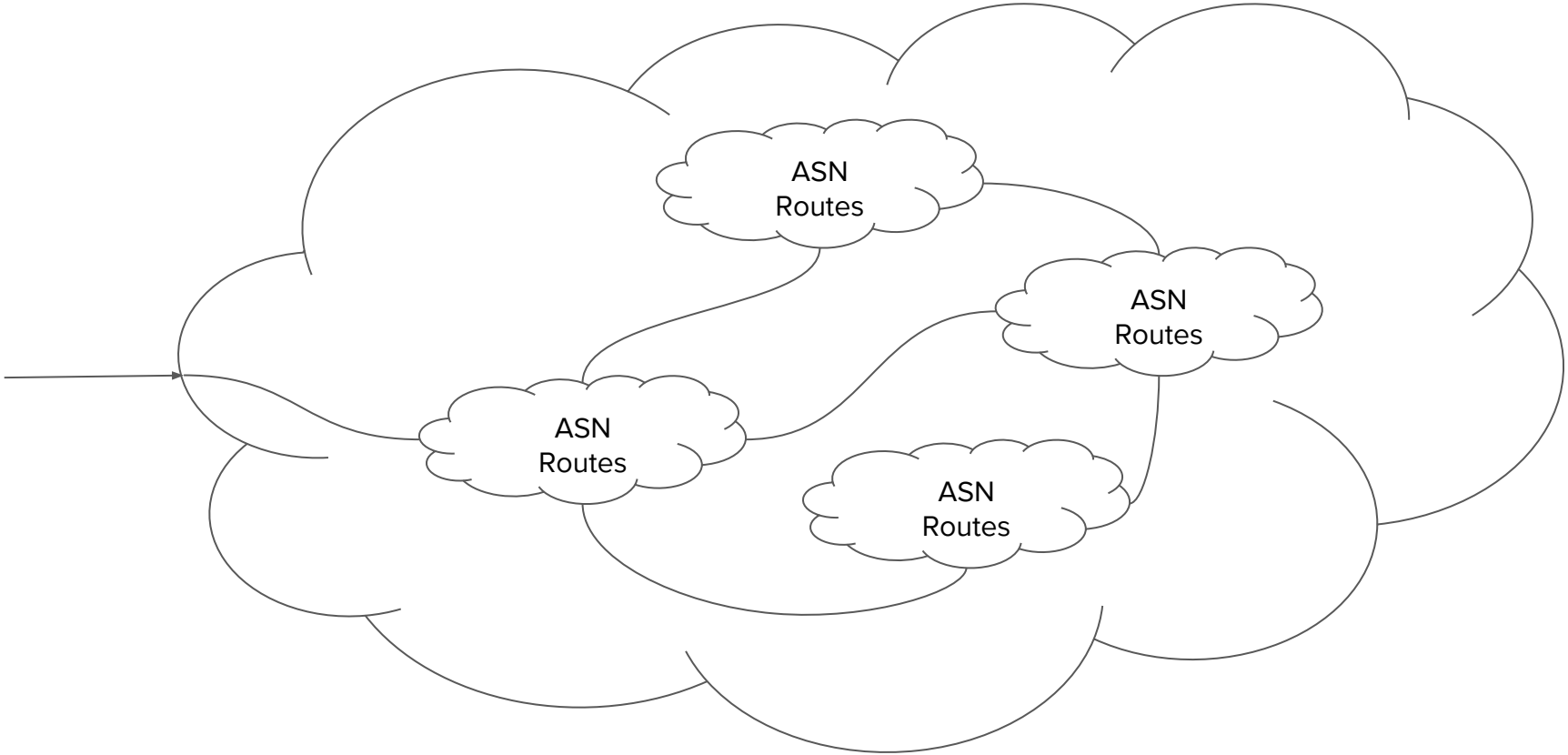
digitalocean.com

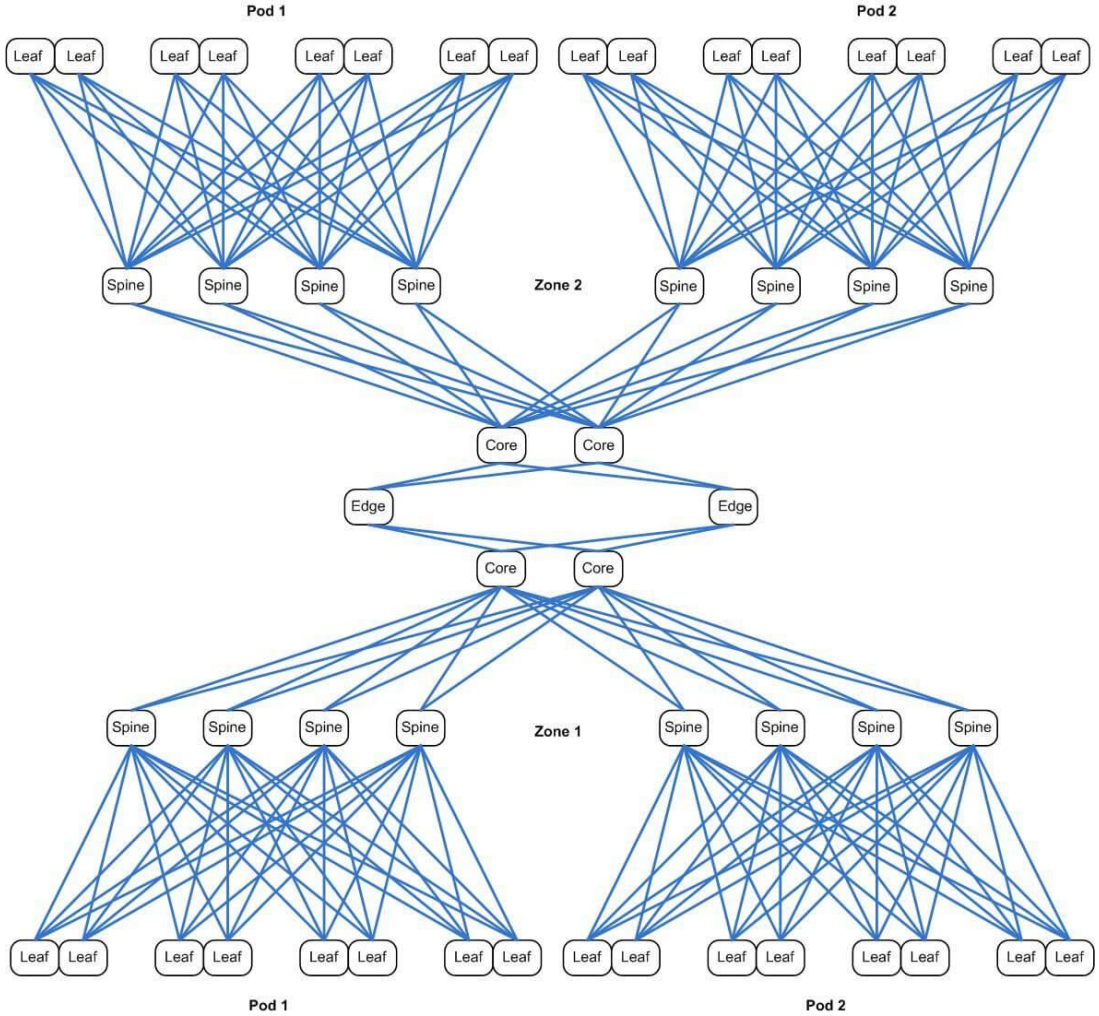
BGP

“The Border Gateway Protocol (BGP) is an inter Autonomous System routing protocol”

– IETF RFC 4271


ASN
Routes
ASN
Routes
ASN
Routes

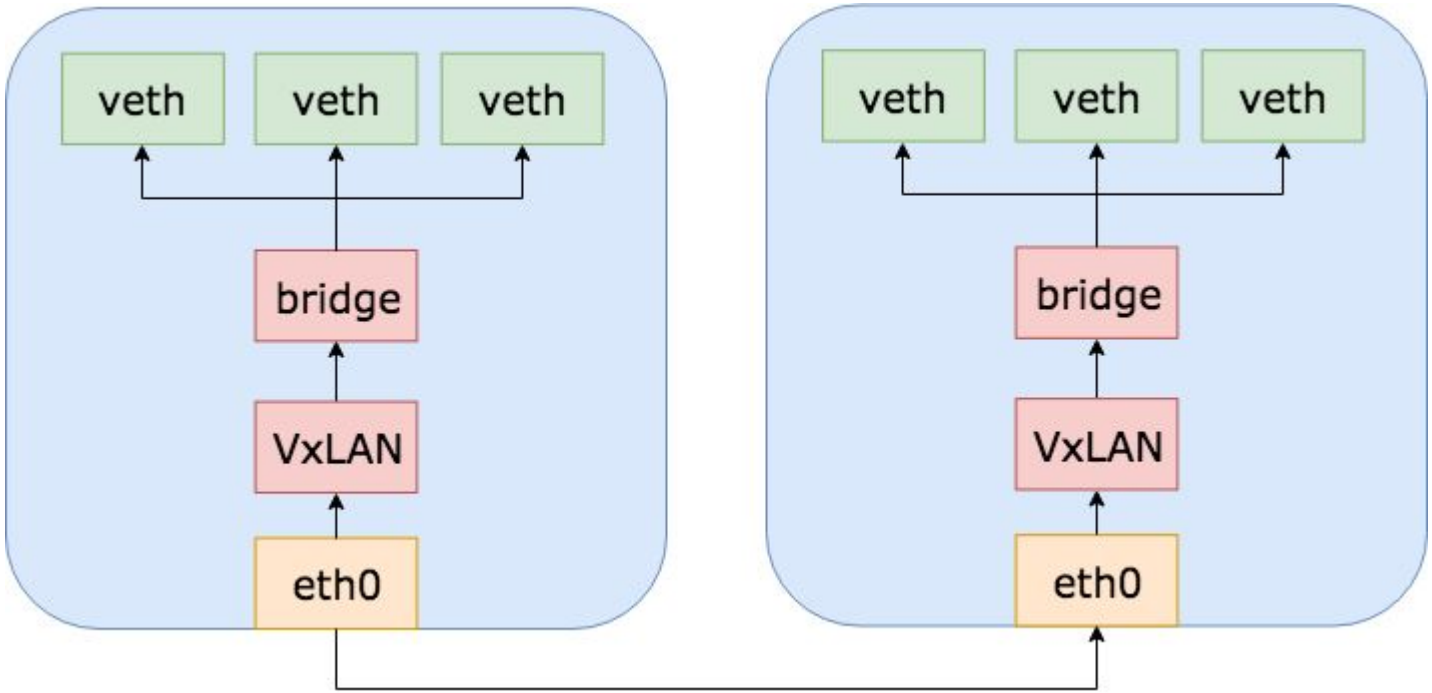

ASN
Routes
ASN
Routes
ASN
Routes
ASN
Routes


Pod 1
Pod 2
Leaf
Spine
Zone 2
Core
Edge
Zone 1
Pod 1
Pod 2

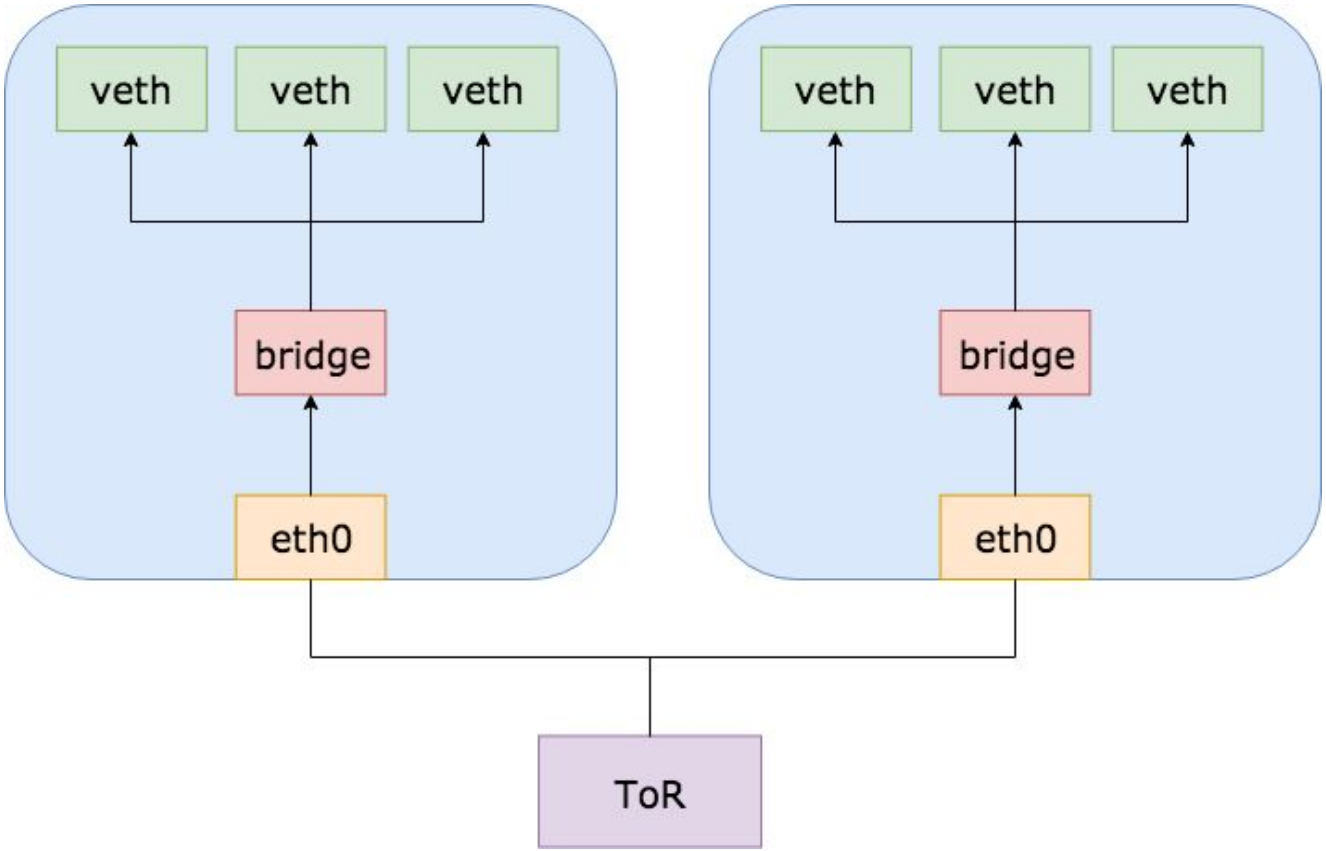
digitalocean.com

Bottlenecks in our Network

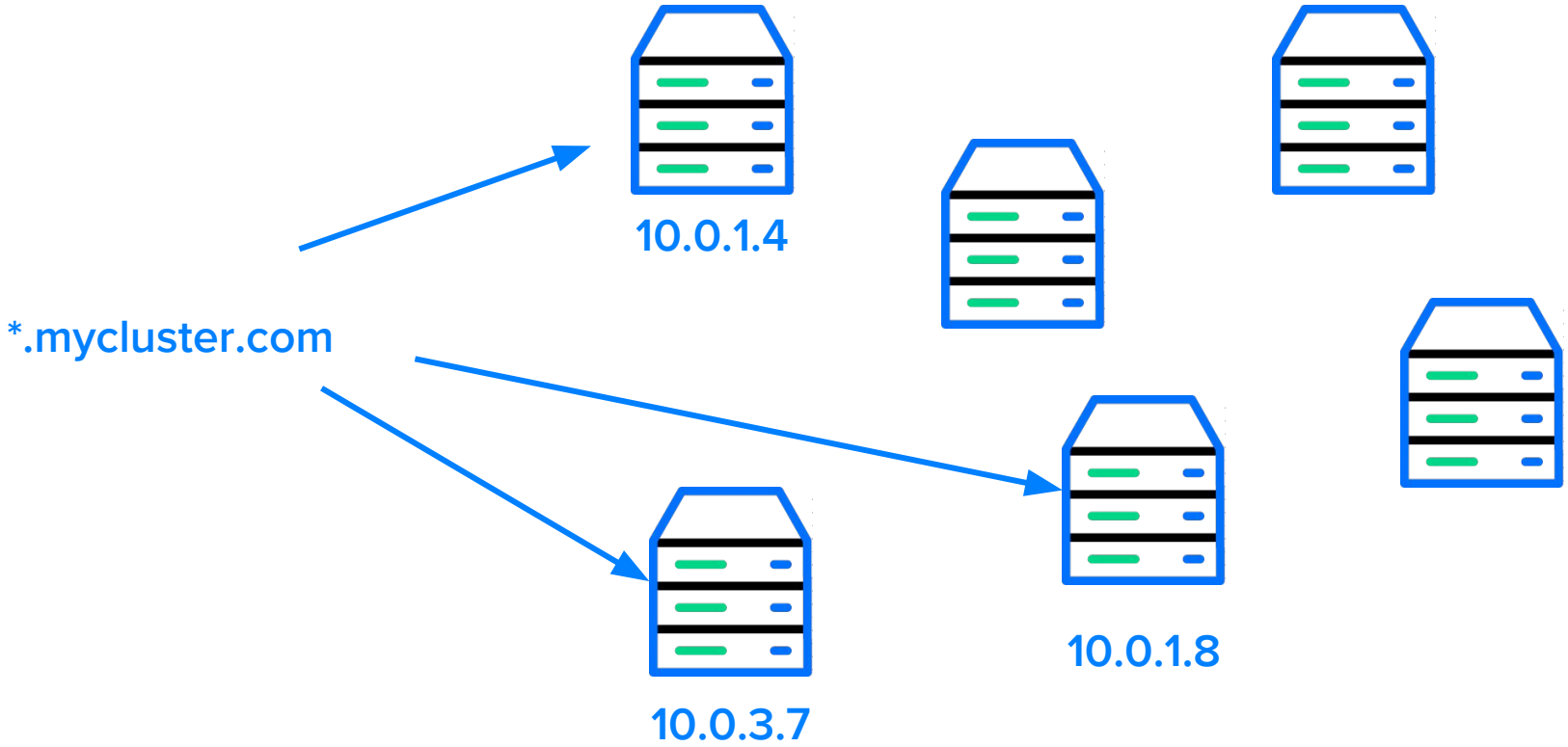
Pod/Container Networking



Pod/Container Networking



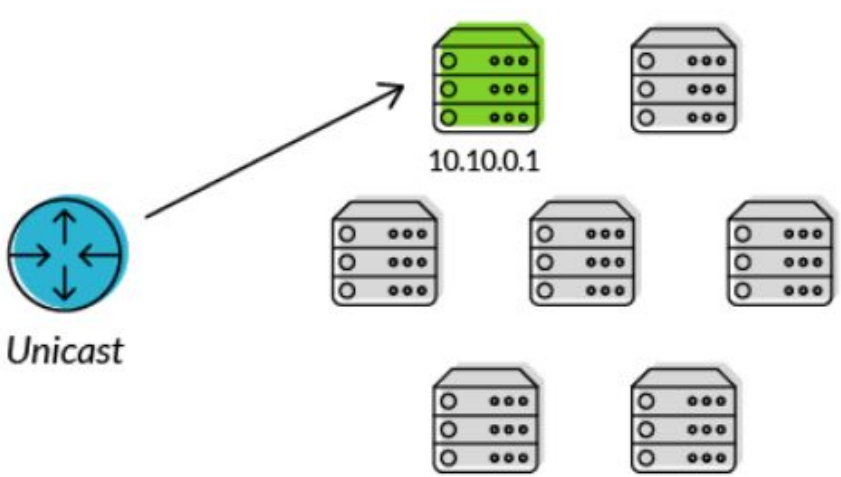
External Traffic into Kubernetes Cluster

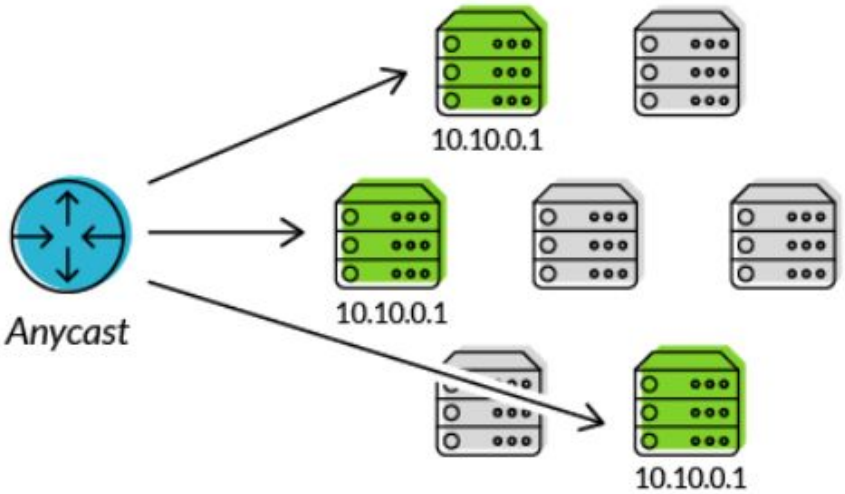


Anycast

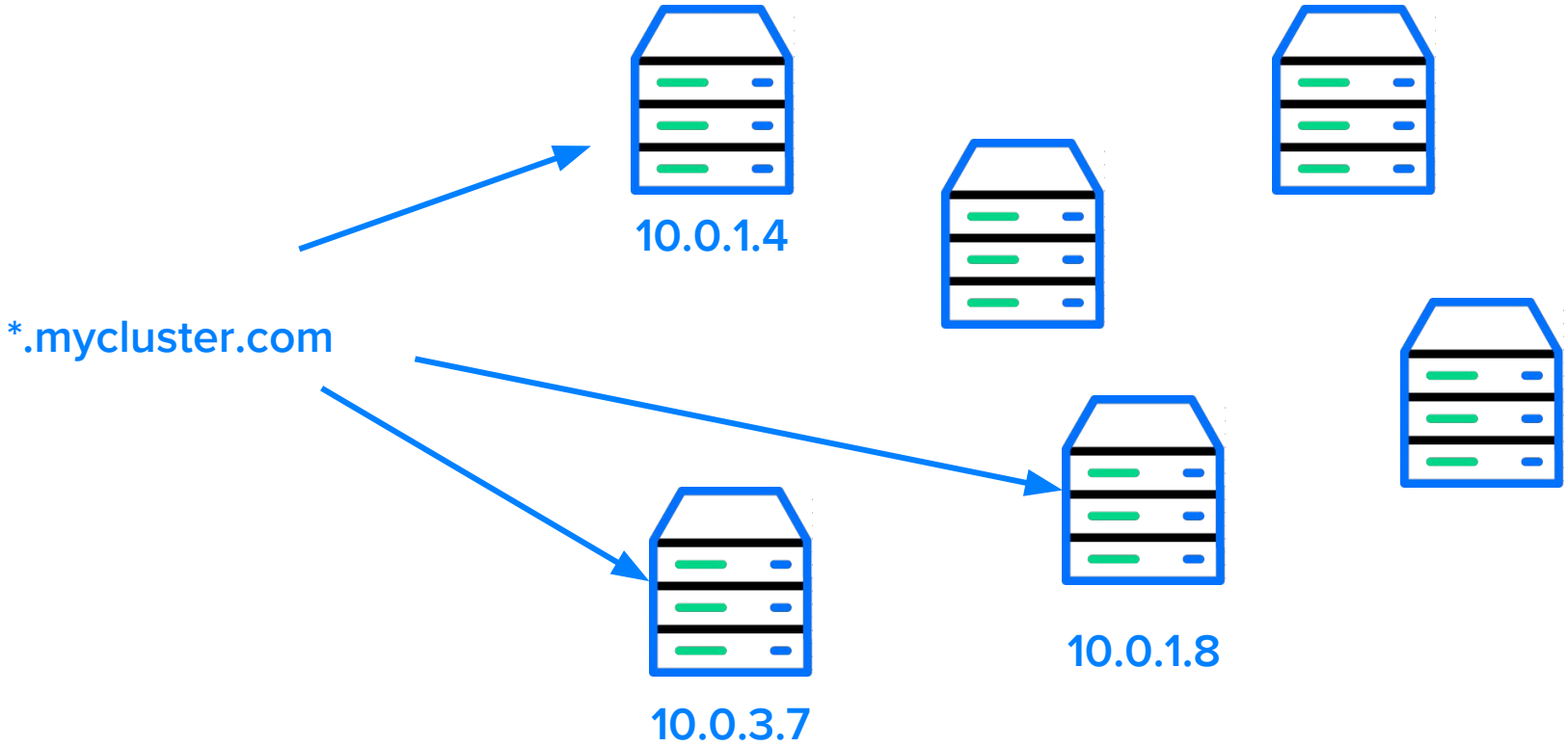
Anycast Routing

Anycast is a network addressing and routing methodology in which a single destination address has multiple routing paths to two or more endpoint destinations

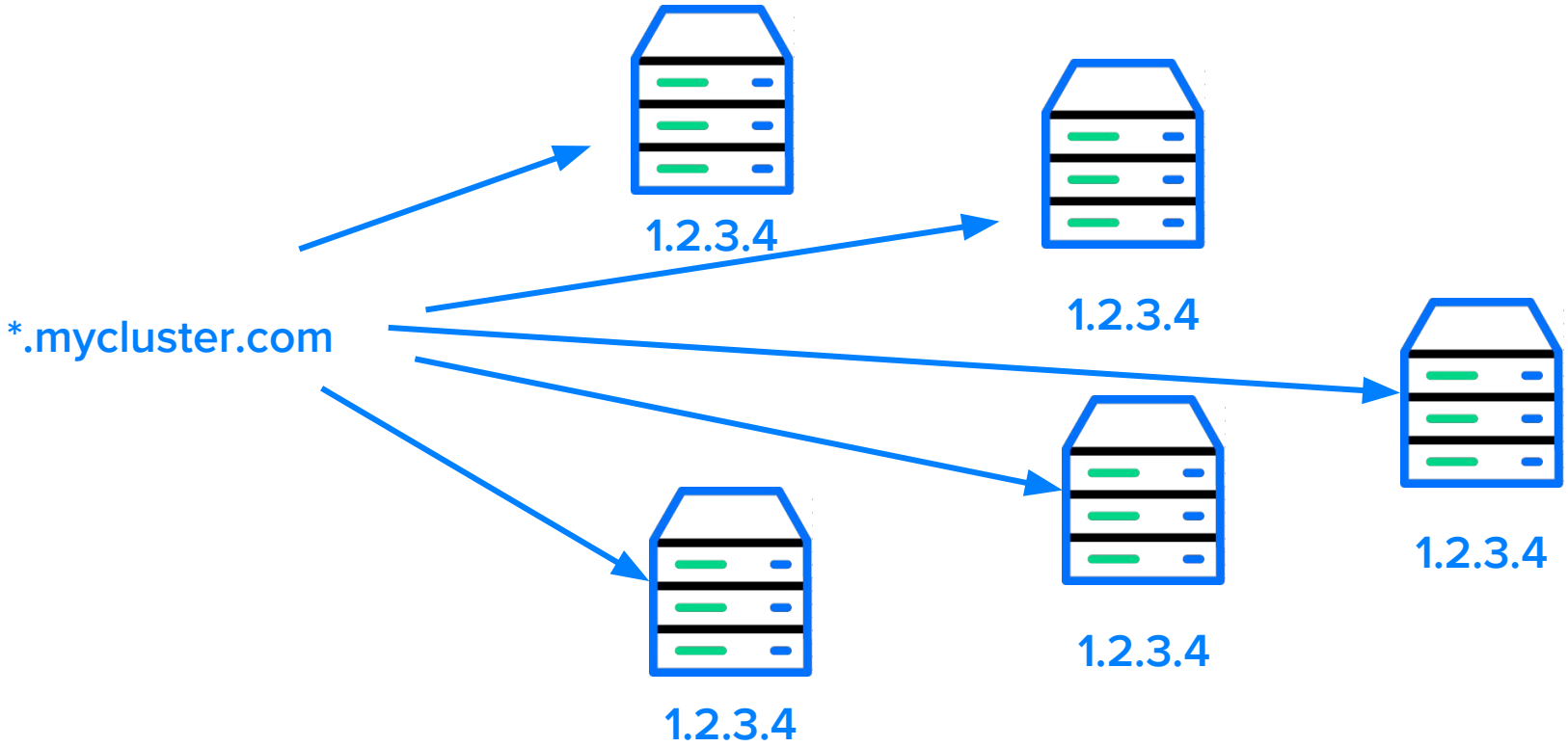

10.10.0.1
Unicast


10.10.0.1
10.10.0.1
Anycast
10.10.0.1

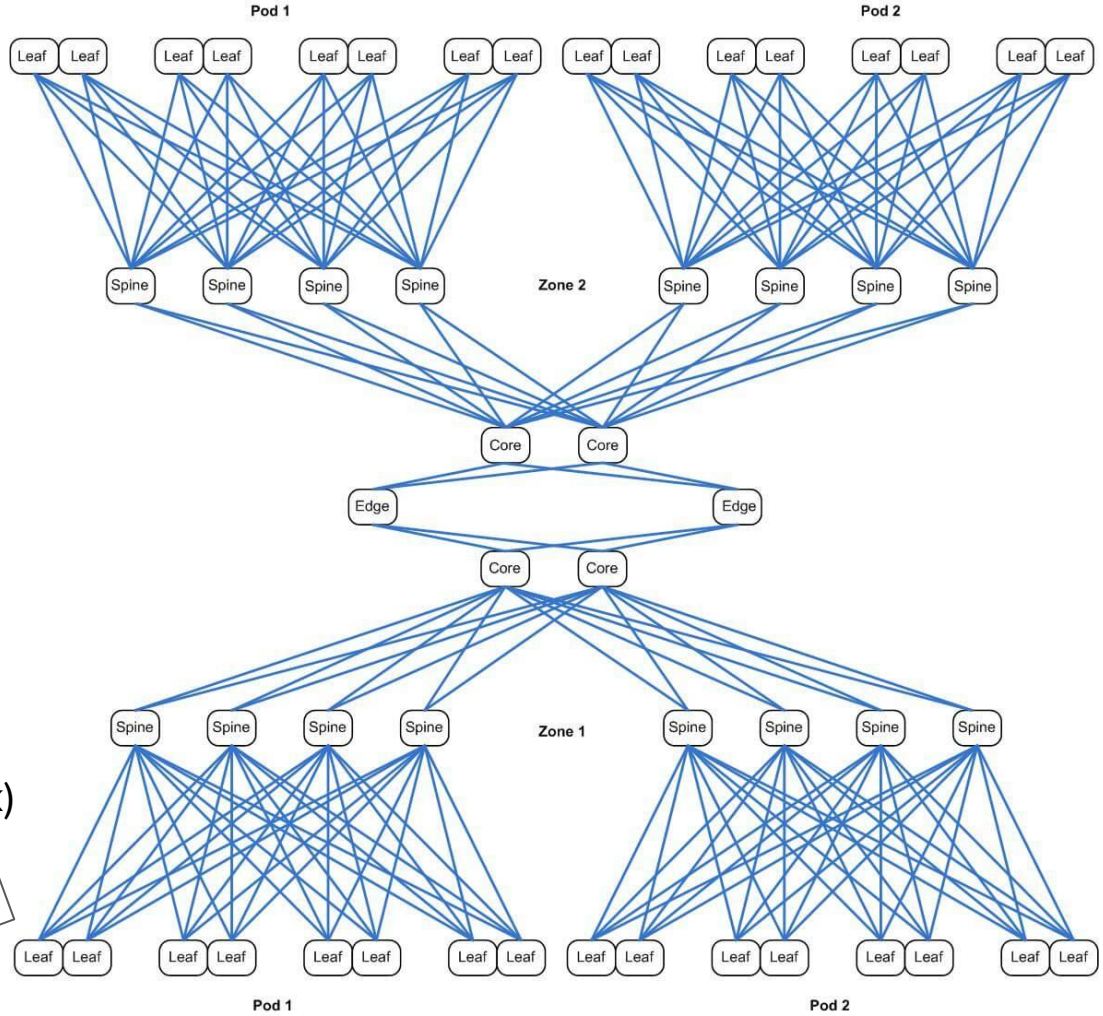
External Traffic into Kubernetes Cluster



External Traffic into Kubernetes Cluster



Kubernetes Clusters as BGP Autonomous Systems



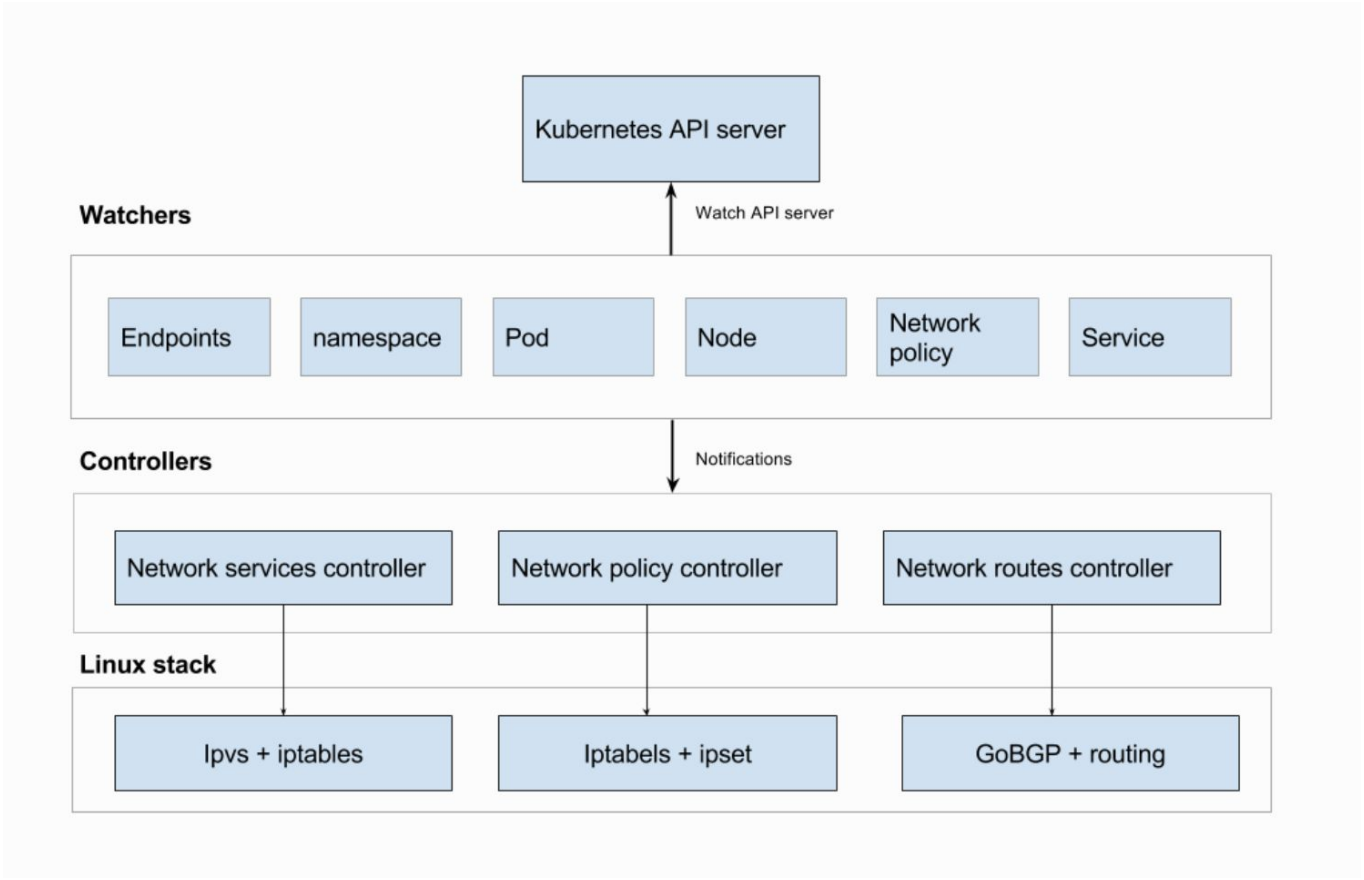
ToR (Top of Rack)

digitalocean.com

https://github.com/cloudnativelabs/kube-router

Maintained by **@murali-reddy**

- Supports iBGP / eBGP peering
- Automatic BGP peering of pod and service subnets
- Bonus: IPVS/DSR support, network policies, BGP route reflectors


Kubernetes API server
Watch API server
Watchers
Endpoints
namespace
Pod
Node
Network policy
Service
Notifications
Controllers
Network services controller
Network policy controller
Network routes controller
Linux stack
Ipvs + iptables
Iptabels + ipset
GoBGP + routing

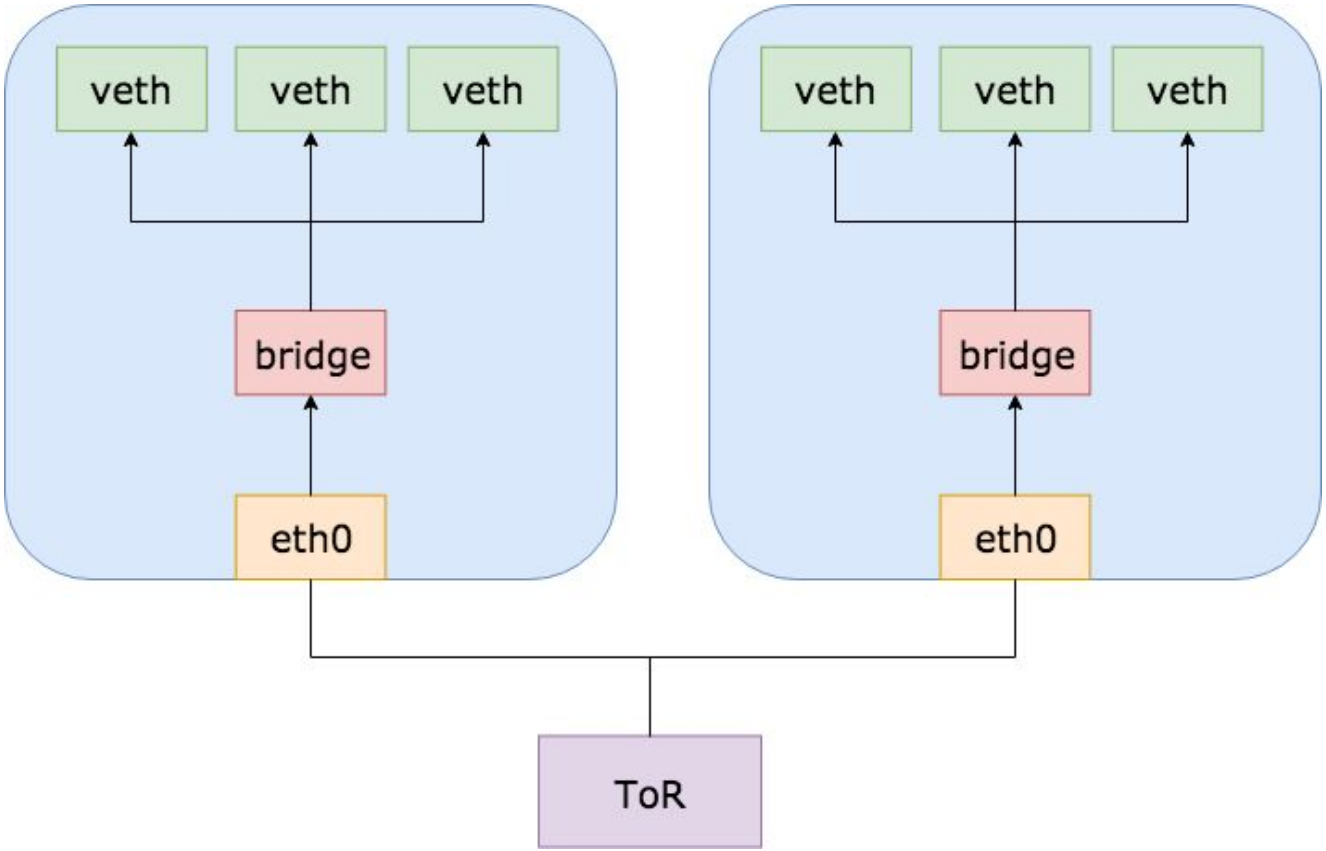
kube-router

- IP and ASN to peer with
- BGP peering required on all nodes

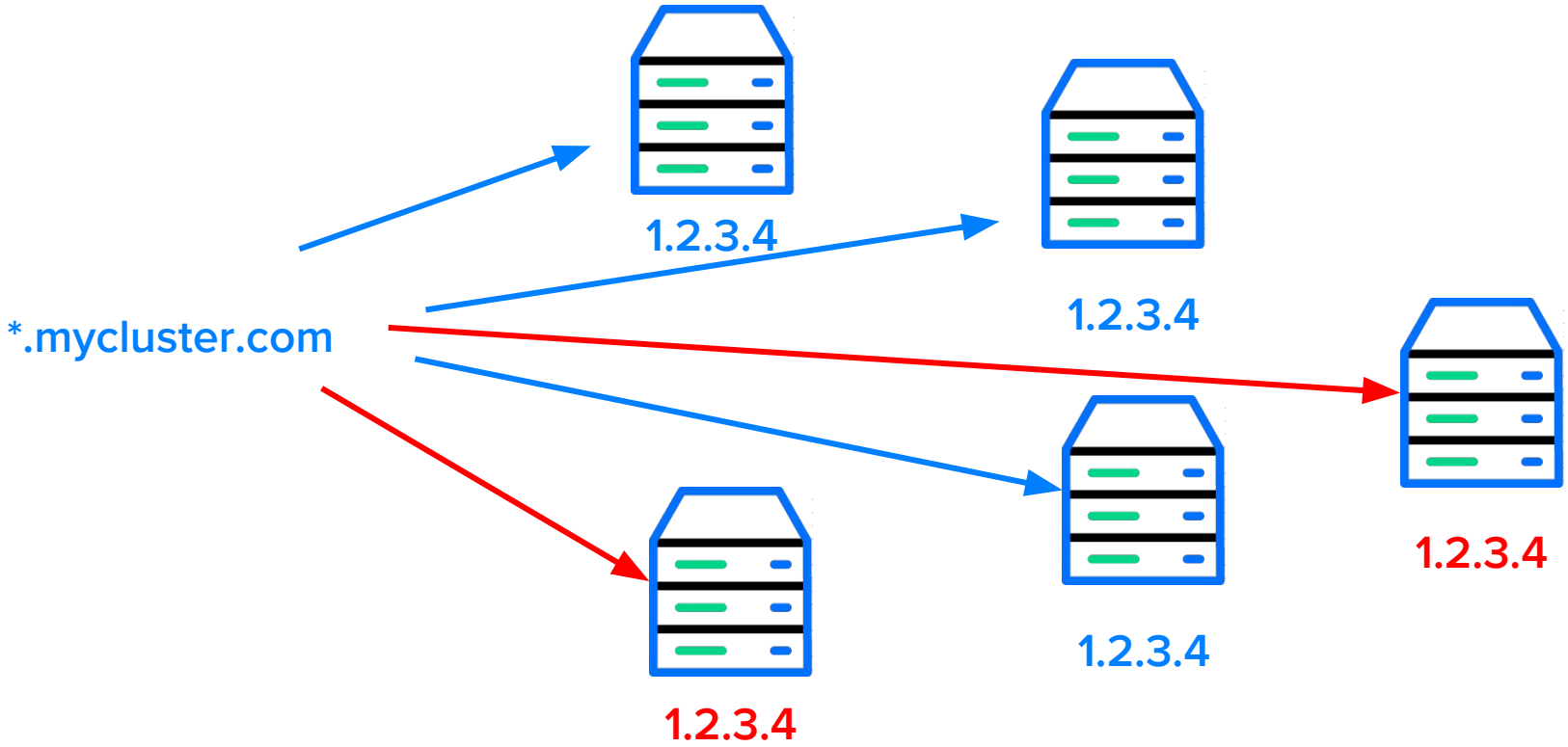
```
---
apiVersion: apps/v1
kind: DaemonSet
metadata:
  name: kube-router
  labels:
    k8s-app: kube-router
spec:
    ...
    containers:
      args:
        - --cluster-asn=<cluster-asn>
        - --peer-ips=<top-of-rack-ip>
        - --peer-asns=<top-of-rack-asn>
```

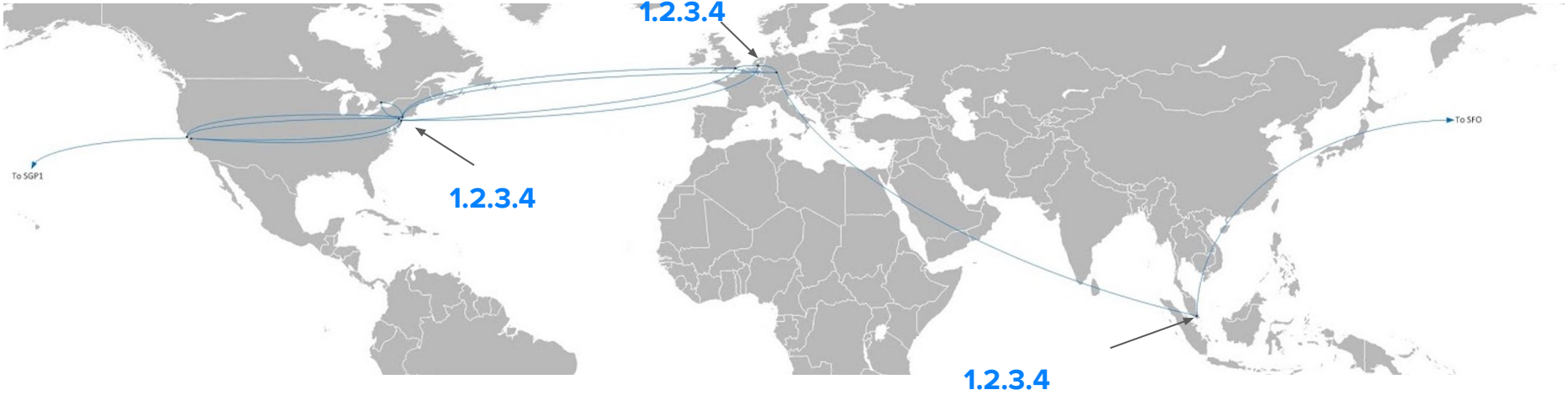
Benefits!

Faster Pod Network!



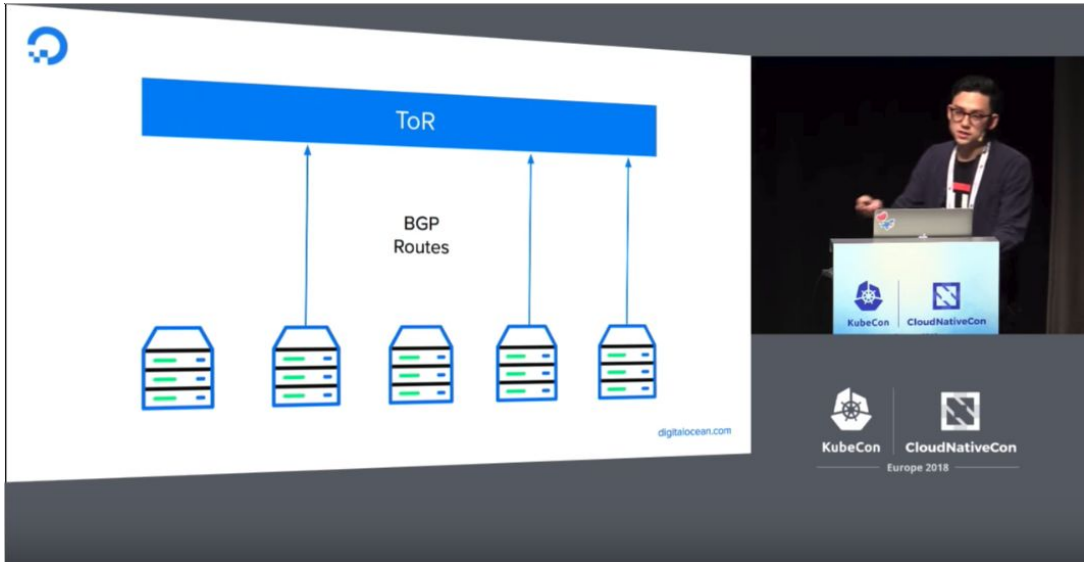
External Traffic into Kubernetes Cluster




1.2.3.4
To SFO
To SGP1
1.2.3.4
1.2.3.4

Summary

- BGP + Kubernetes works!!
- Anycast + Kubernetes Service IPs is a powerful combination!



Global Container Networks on Kubernetes at DigitalOcean
https://youtu.be/tHAkey-sZ9g

Thank you!

 @a_sykim

 @andrewsykim