
KubeCon
CloudNativeCon
North America 2023

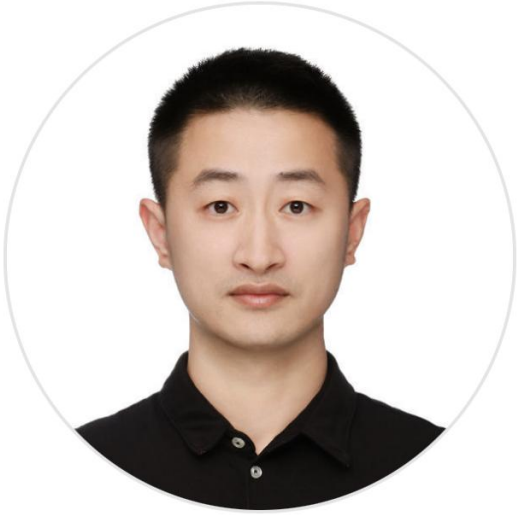

KubeCon | CloudNativeCon

North America 2023

Make Underlay CNI to Be Powerful and Simple

Weizhou Lan (蓝维洲)
Qiuping Dai (戴秋萍)

Speakers



Weizhou Lan (蓝维洲)
Senior Technical Leader, Daocloud
Cloud native experience in Network, eBPF, Chaos, Observability, Mesh

Qiuping Dai (戴秋萍)
Product manager, Daocloud
Cloud native experience in Storage, Network, Scheduling

Kubernetes CNI

Overlay CNI

Calico

cilium

flannel

VS

Underlay CNI

Macvlan

IPVLAN

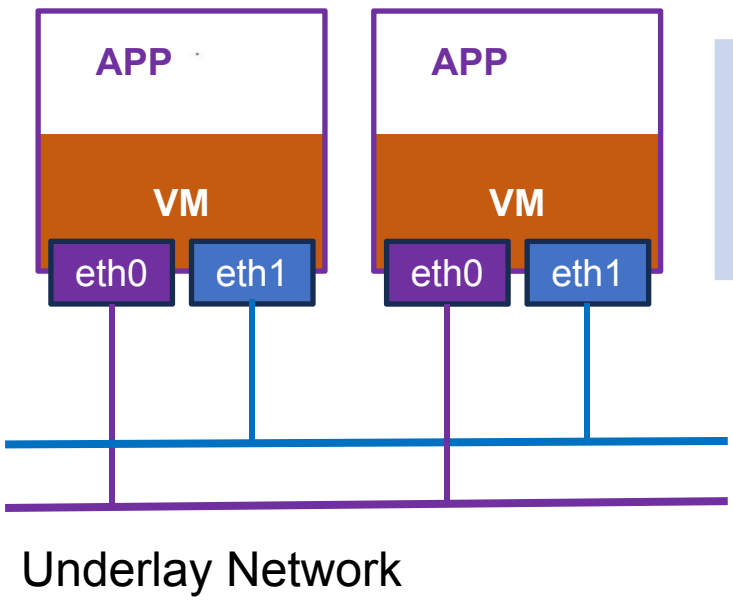
SRI-OV

Overlay CNI is popular

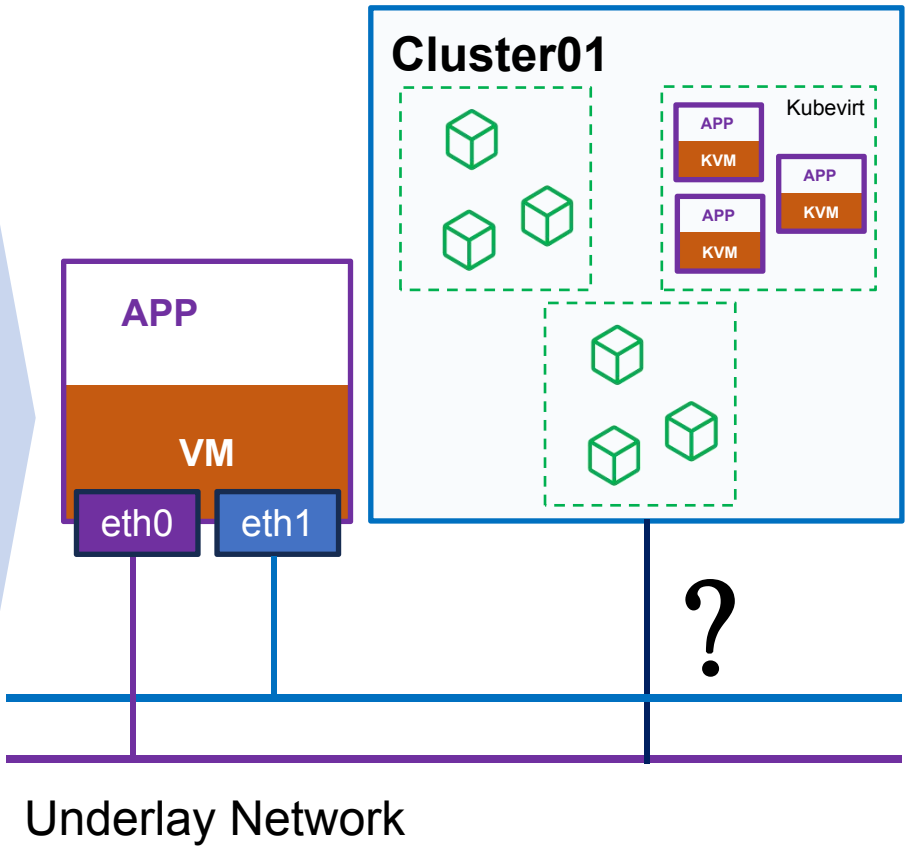
But underlay CNI can not be replaced at all

Traditional APP Containerized

Traditional Apps on Host



Traditional Apps on Kubernetes



Underlay Network Access

-- Multicast / Groupcast
-- ARP

Fixed IP

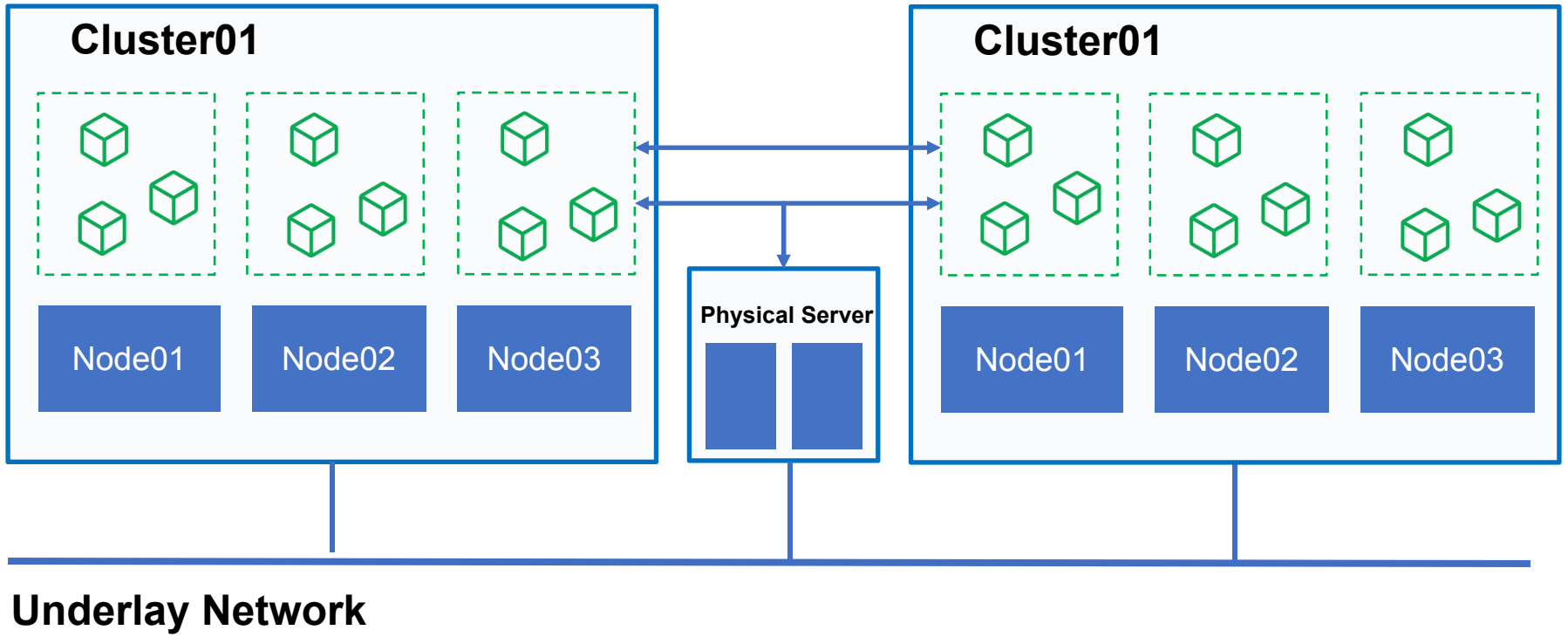
-- Expose Service by IP
-- Firewall policy with no NAT

VLAN Subnet Isolation

-- Business traffic isolation
-- Log traffic isolation

Communication Outside the Cluster

- Cross-Cluster Redis / MySQL
- Application Registry Center out of the cluster



AI : Network is the bottleneck

OpenAI
ChatGPT3.5

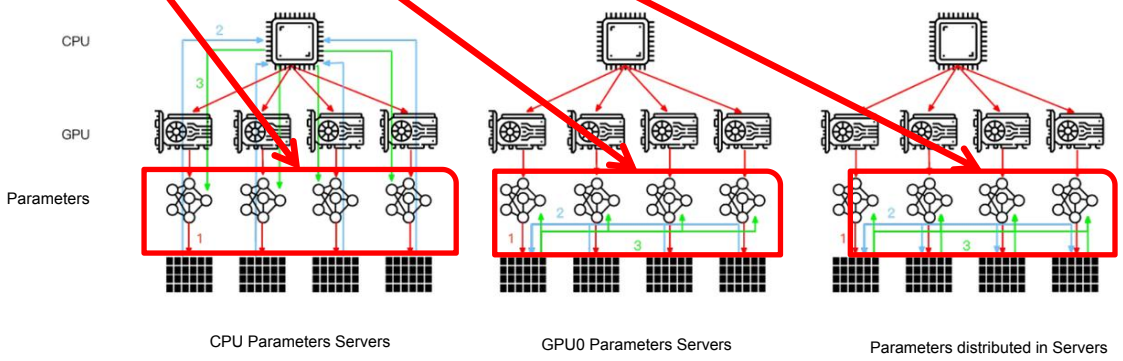
The communication between compute nodes:
hundreds of GB
Bandwidth:
≥ 800Gps+

175 Billion Parameters

10,000 GPU(V100)

2000+ Nodes

3640 PF − days Computing Power



RDMA :

- **Reduce Training Time**
- **Improve GPU Utilization**
- **Offload CPU**

Advantages of Underlay CNI

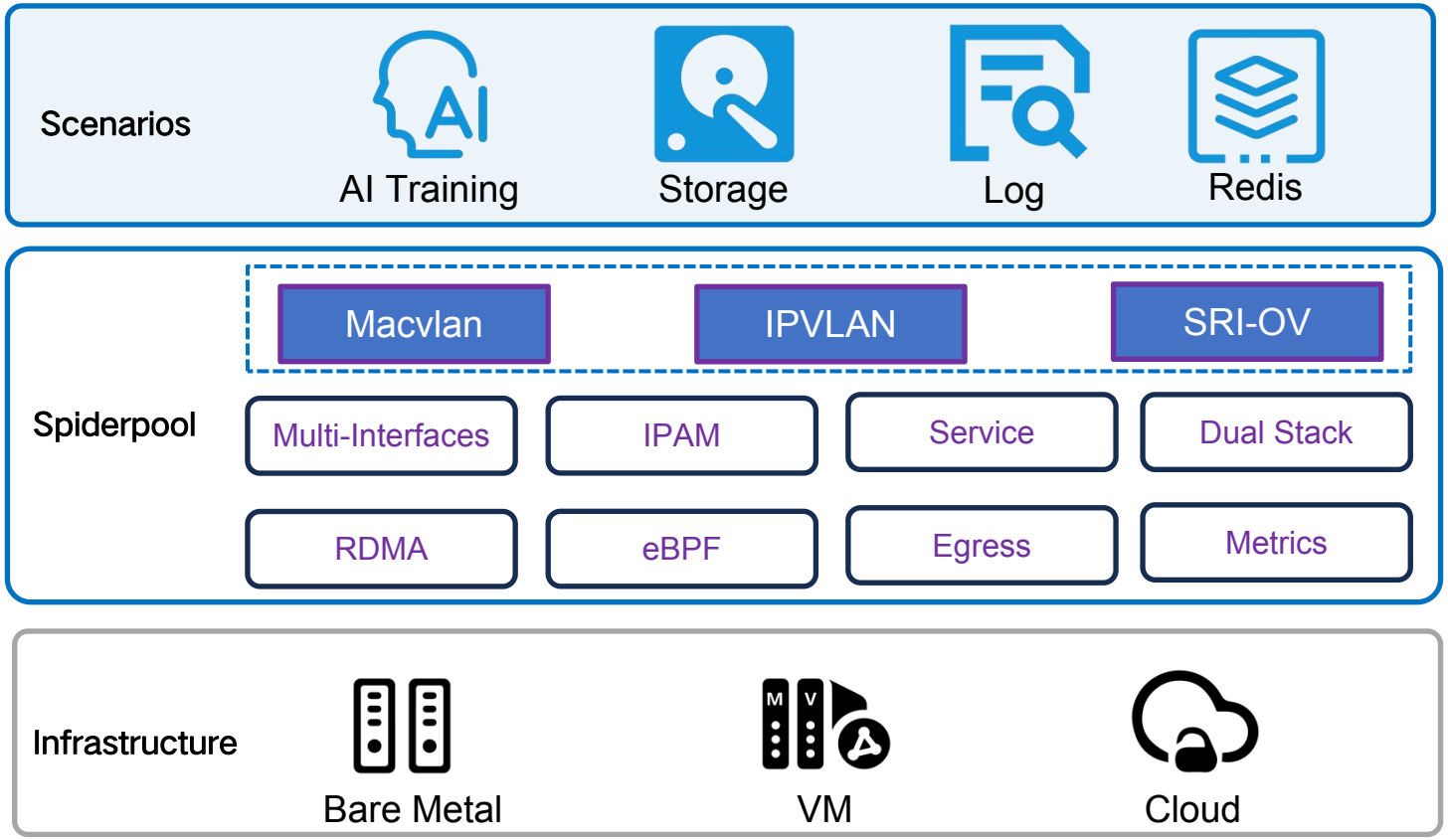
- ✓ High Performance
- ✓ Utilize RDMA
- ✓ Reduce Apps Migration Cost
- ✓ Bandwidth Isolation
- ✓ Strong VLAN network isolation and Firewall
- ✓ Multi Cluster Communication

Open Source Solutions

- Underlay CNI has some Communication Problems for Health Check/Cluster IP
- No efficient IPAM allocation mechanism in large scale case
- No solution to solve the communication between Multiple CNIs

Spiderpool

Spiderpool is an underlay network solution that enhances the capabilities of Macvlan CNI, ipvlan CNI, SR-IOV CNI. It meets various networking requirements and run on **bare metal, virtual machine, and public cloud**.



Milestone

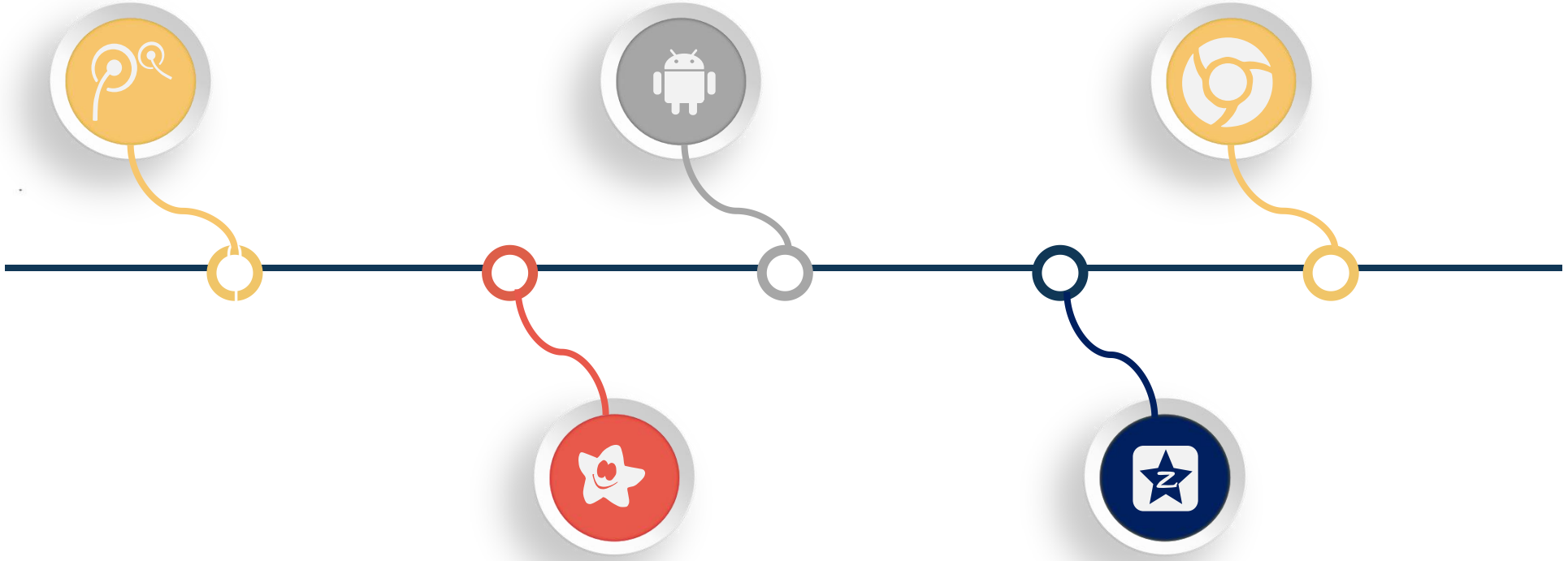
2015
Private project named ‘Dswitch’ in Daocloud

2019
Upgrade to project ‘parcel’

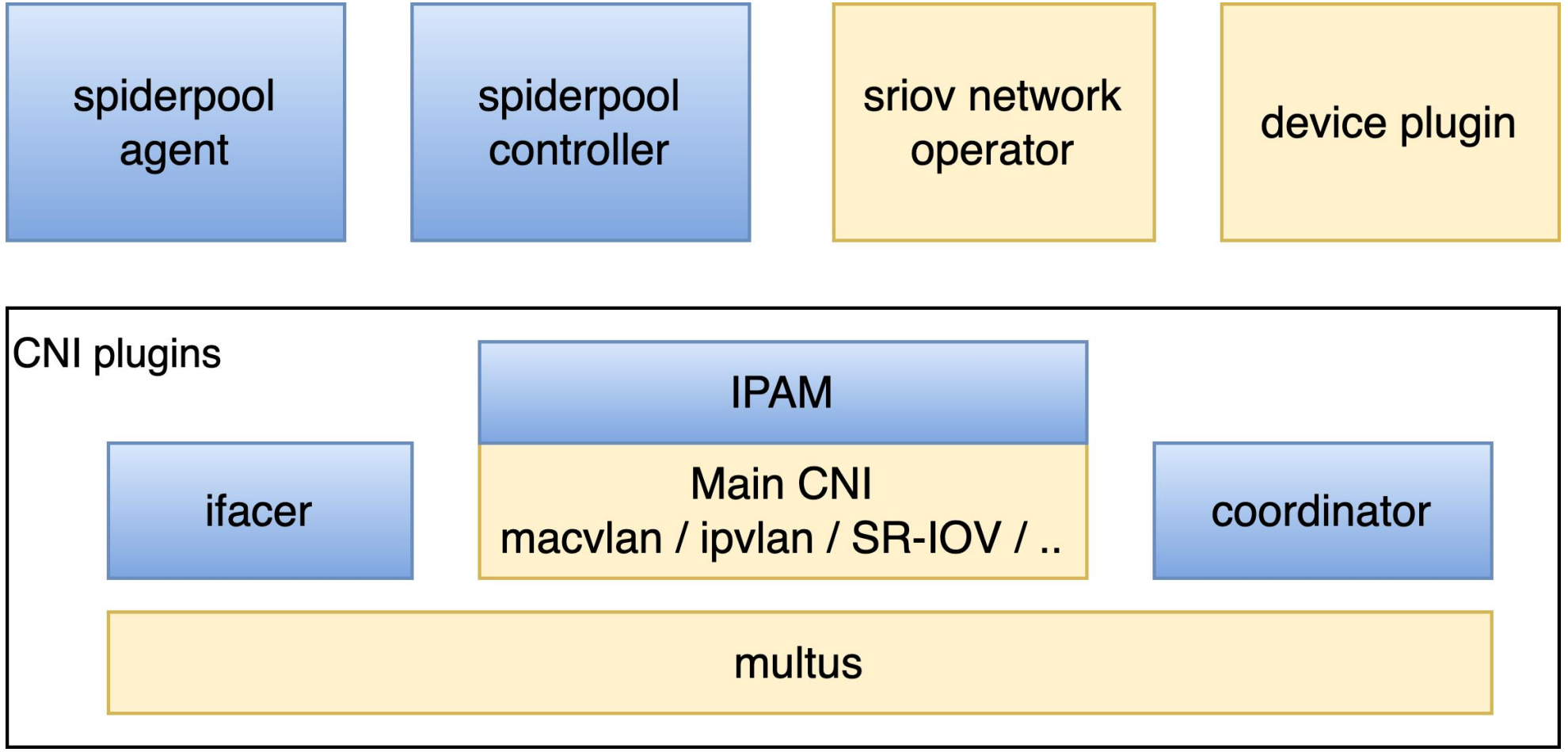
2022.4
Public at github, release spiderpool v0.0.1

2023.5
CNCF Landscape

2023.11
Spiderpool v0.8.0

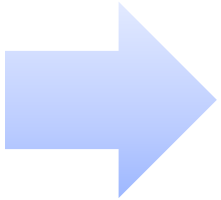


Architecture



Multus Enhancement


```yaml
apiVersion: spiderpool.spidernet.io/v2beta1
kind: SpiderMultusConfig
metadata:
  name: macvlan-eno1
spec:
  cniType: macvlan
  macvlan:
    master:
      - eno1
    vlanID: 100
    ippools:
      ipv4: ["eno1-net90"]
```



```yaml
apiVersion: k8s.cni.cncf.io/v1
kind: NetworkAttachmentDefinition
metadata:
  name: macvlan-eno1
spec:
  config: -|
    {
      "cniVersion": "0.3.1",
      "name": "macvlan-eno1",
      "plugins": [
        { "type": "ifacer", "vlanID": 100,"interfaces": ["eno1"] },
        { "type": "macvlan",
          "master": "eno1.100",
          "mode": "bridge",
          "ipam": { "type": "spiderpool",
                    "default_ipv4_ippool": [ "eno1-net90"]}
        },
        { "type": "coordinator" }
    ]}
```

User-friendly writing :

- Easy yaml format
- Integrate with multiple plugins
- Best practice value as the default value

IPAM: SpiderIPPool And Affinity

Based on CRD

strong verification when creating, updating and deleting

An IP address can be attempted to be assigned to a Pod on any node, which is different from overlay IPAM based on IP blocks.

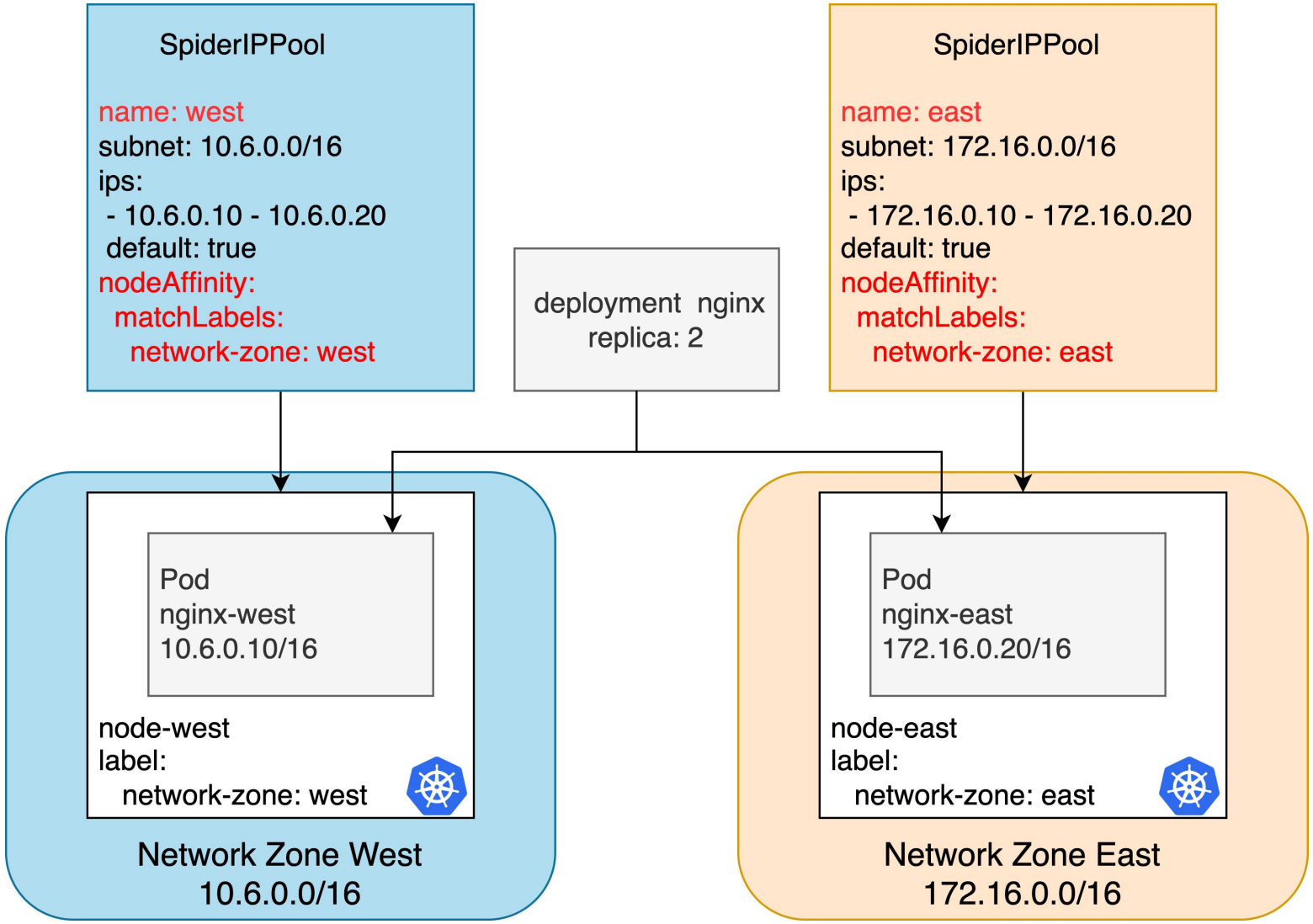
The affinity determines whether a Pod can successfully allocate an IP address from a specific IP pool :

- podAffinity, filter based Pod label.
- nodeAffinity, filter based on node where Pod is running.
- namespaceAffinity, filter based on Pod namespace.
- multusName, filter based on which NetworkAttachmentDefinition Pod is using.

```
kind: SpiderIPPool
spec:
  gateway: 172.81.0.1
  ips:
    - 172.81.0.100-172.81.0.120
  subnet: 172.81.0.0/16
  podAffinity :
    matchLabels :
      app: static
  nodeAffinity :
    matchLabels:
      network/zone: west
  namespaceAffinity :
    matchLabels:
      usergroup: tester
  multusName:
    - kube-system/ipvlan-eth0
```

IPAM: Replicas Across Subnets

For a cluster with nodes accross multiple network zones or subnets, how can you customize the IP addresses of replicas on different nodes ?



IPAM: Fixed IP

Scenario :

- no need to update rules of firewall
- Services of stateful applications could be exposed by fixed Pod IP

Community :

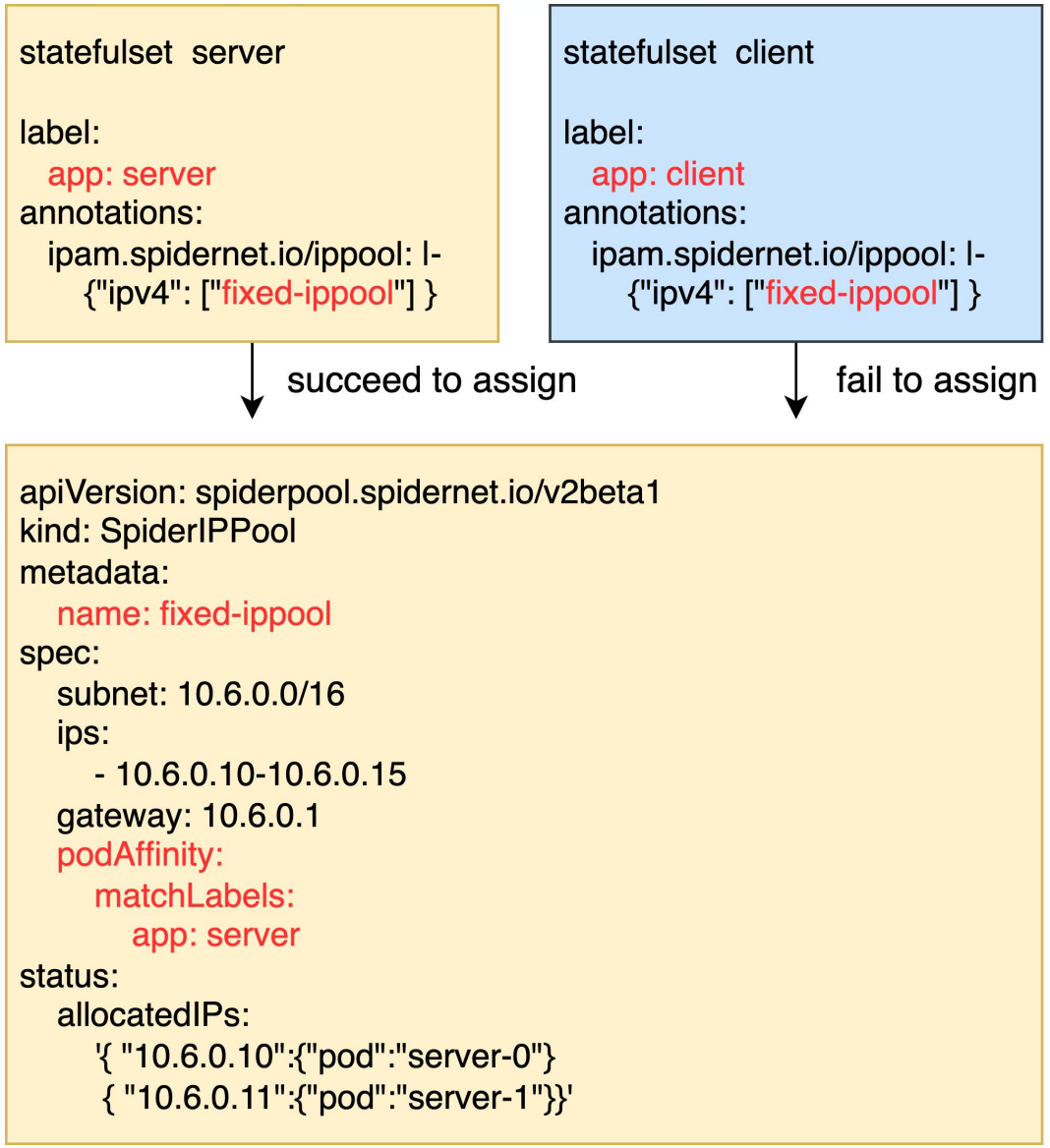
Pod annotation to fix the IP range:

xxxxx/ips: [“192.168.1.10”,”192.168.1.11”]

Prone to IP conflicts between applications, and can hardly observe the total IP usage.

Spiderpool :

- Strong verification of IP overlap between SpiderIPPools
- The stateful and stateless Pods could be assigned in a limited IP range.
- Each StatefulSet Pod and kubevirt VM could get a persistent IP

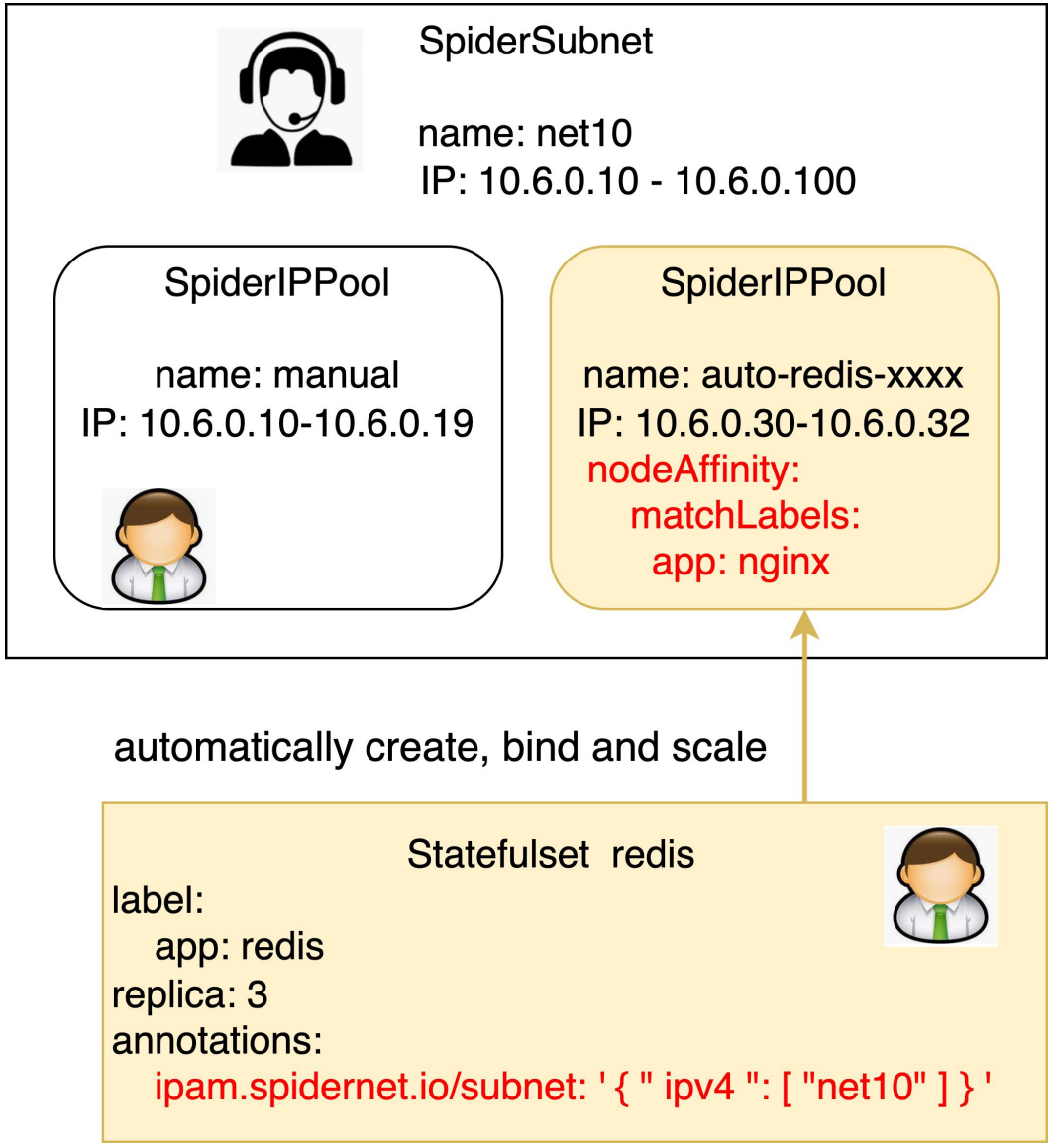


IPAM: SpiderSubnet (Experimental)

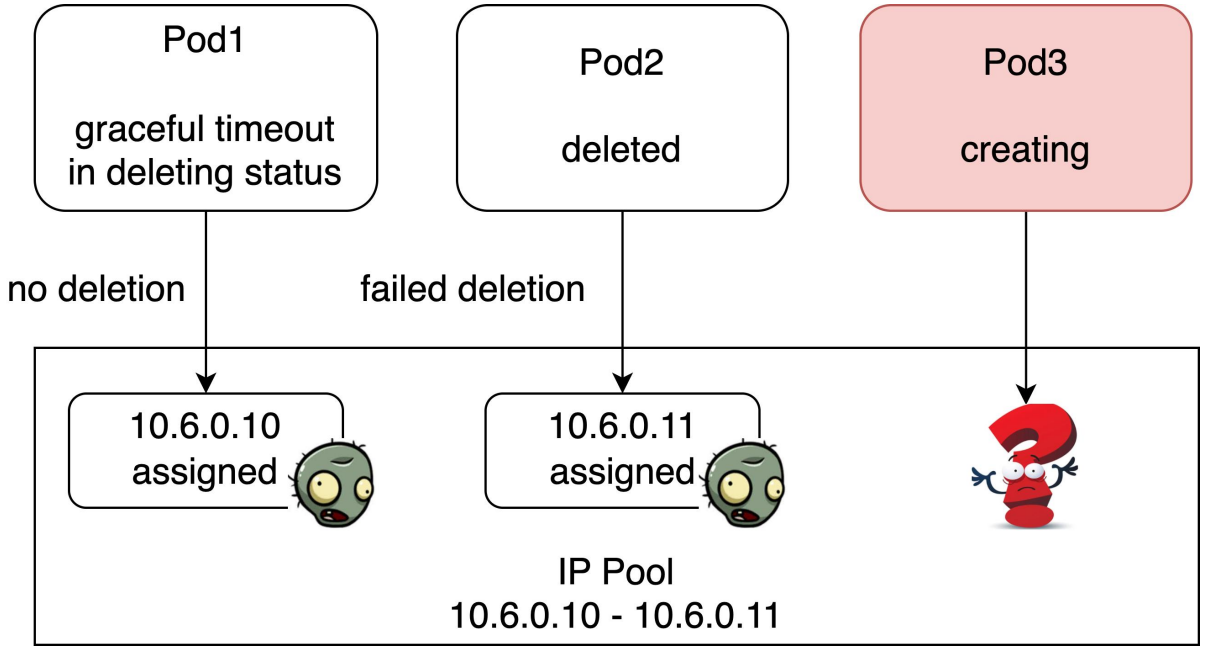
It is a burden for Application administrator to ask for help from platrform administrator, on figuring out which IP address is available,

CRD SpiderSubnet manages all child SpiderIPPool within the same subnet :

- Decoupling the responsibilities of the platrform and application department.
- Automate the processes of creating exclusive SpiderpIPPool for application, expand and delete IP addresses according to application replica number



IPAM: Reclaim Zombie IP



Spiderpool controller takes charge of reclaiming zombie IP :

- Basis
  Reclaim 'zombie IP' taken by deleted Pods.
- Enhancement
  Reclaim ‘zombie IP’ tabke by deleting Pod whose duration of deleting status is longer than graceful termination timeout.
  This is especially useful when a certain node break down, and ensure IP avalability for a new scheduled pod.

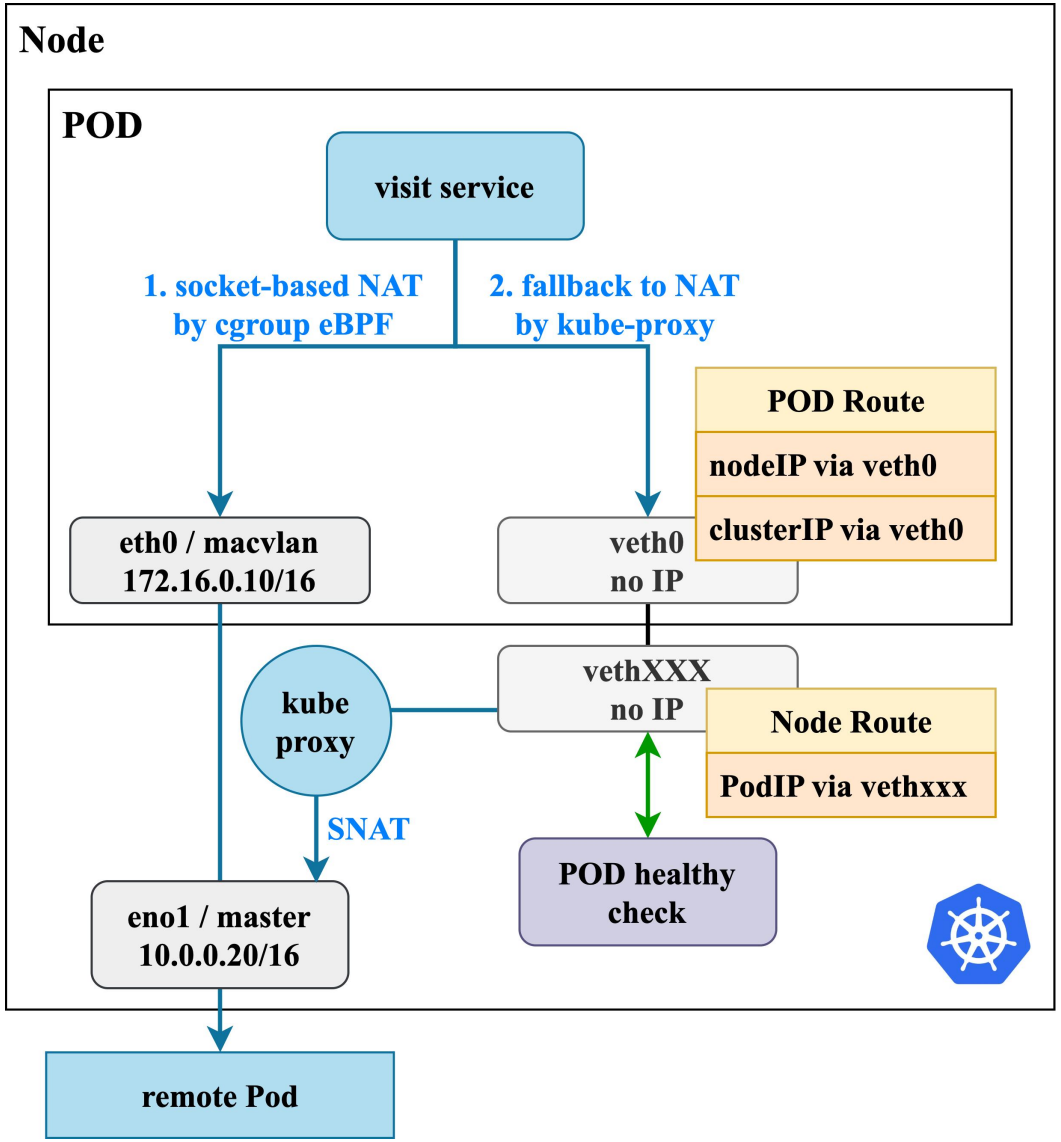
Accessibility Enhancement

Enhance the network accessibility of macvlan, ipvlan and SR-IOV :

- Connetivity between Pod and local node
  The Pod healthy check works even when the Pod and the local node join different subnets.

- Access service
  1. NAT by kube-proxy
  2. NAT by cgroup eBPF
     Up to 25% improvement on network delay , up to 50% improvement on network throughput

- Accessibility check based on probing ARP when launching the Pod
  1. IP conflict check
  2. Gateway reachability check



Multiple Interfaces: Underlay CNI

Scenario :

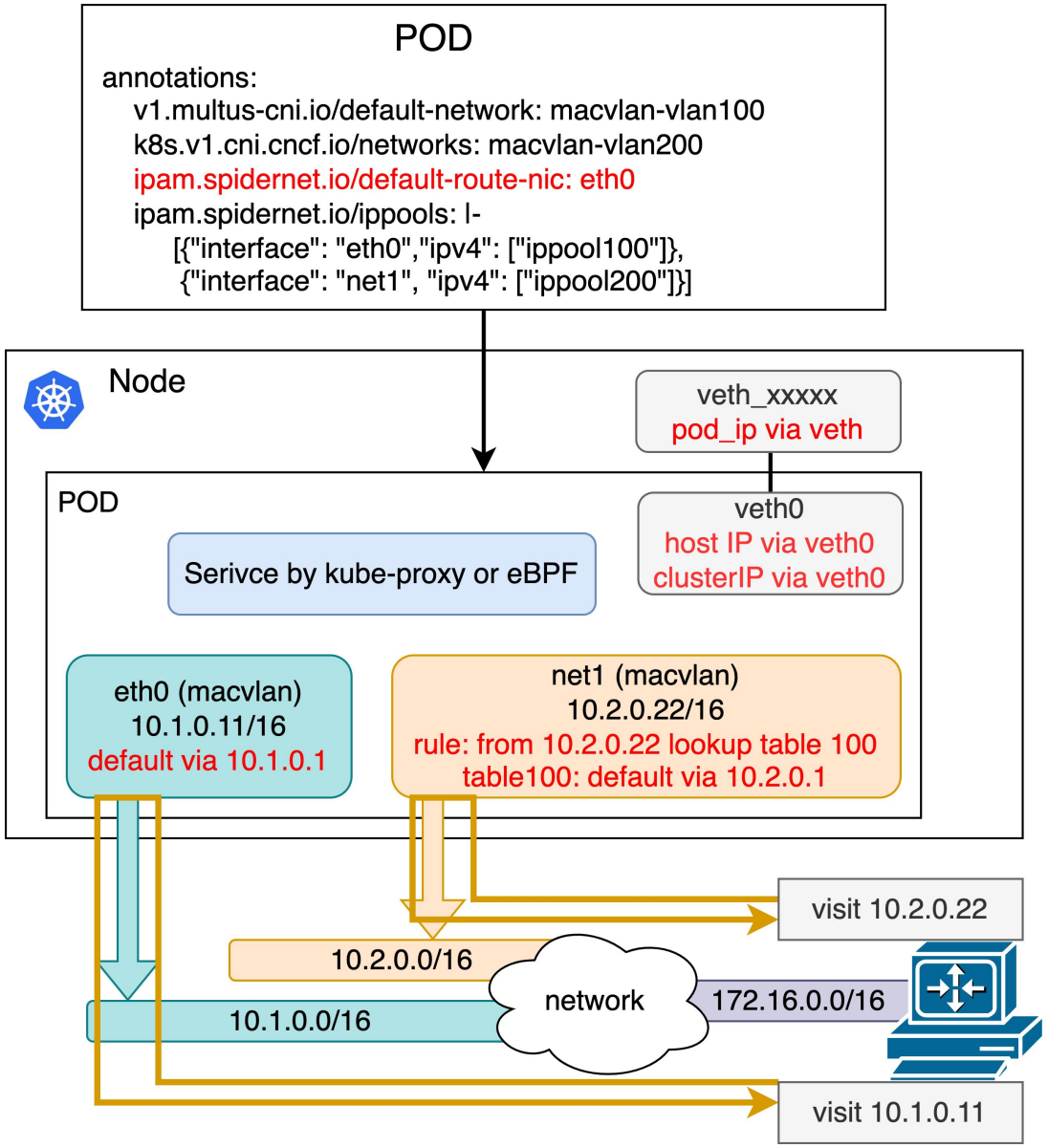
- Pod access subnets isolated from each other

Enhance mutiple underlay interfaces by multus :

- IPAM

  Support to specify IP addresses for each interface.
- Tune the route for all interfaces

  Generate the policy rule and move related gateway route to non-main route table. This guarantees the consistent data path of request and reply packets, avoiding packet loss.



Multipe Interfaces: Underlay and Overlay CNI

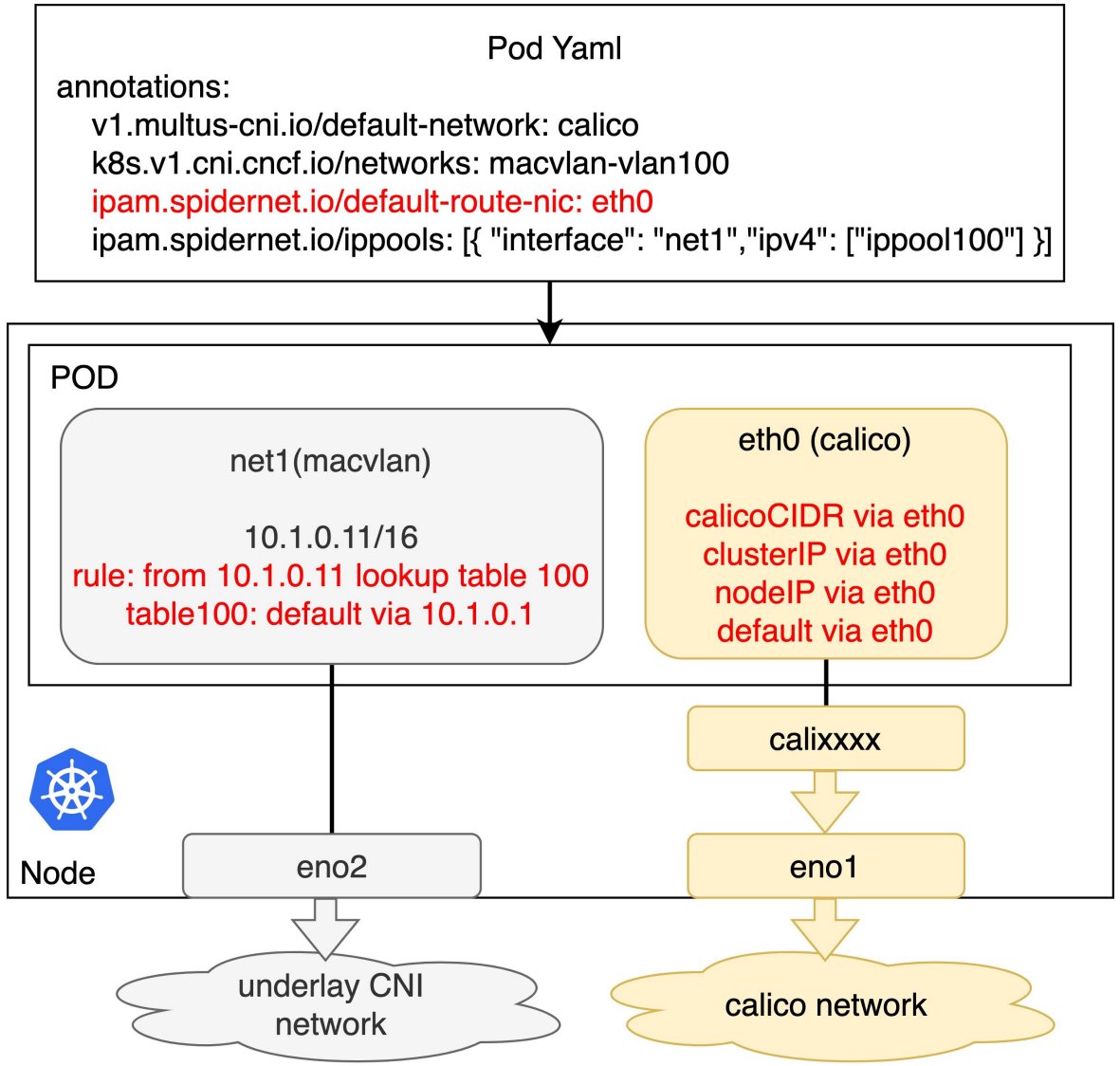
Scenario :

- Overlay Pods need a secondary underlay interface to separately transmit plenty of data on an bandwidth-independent subnet. For example, VM migration of kubevirt
- Overlay interface for TCP communication, underlay interface for RDMA communication

Spiderpool :

Enhance the combination of one overlay interface and mutiple underlay interfaces by multus:

- IPAM for underlay interfaces
- Tune the route for all interfaces



Underlay CNI On Any Public Cloud

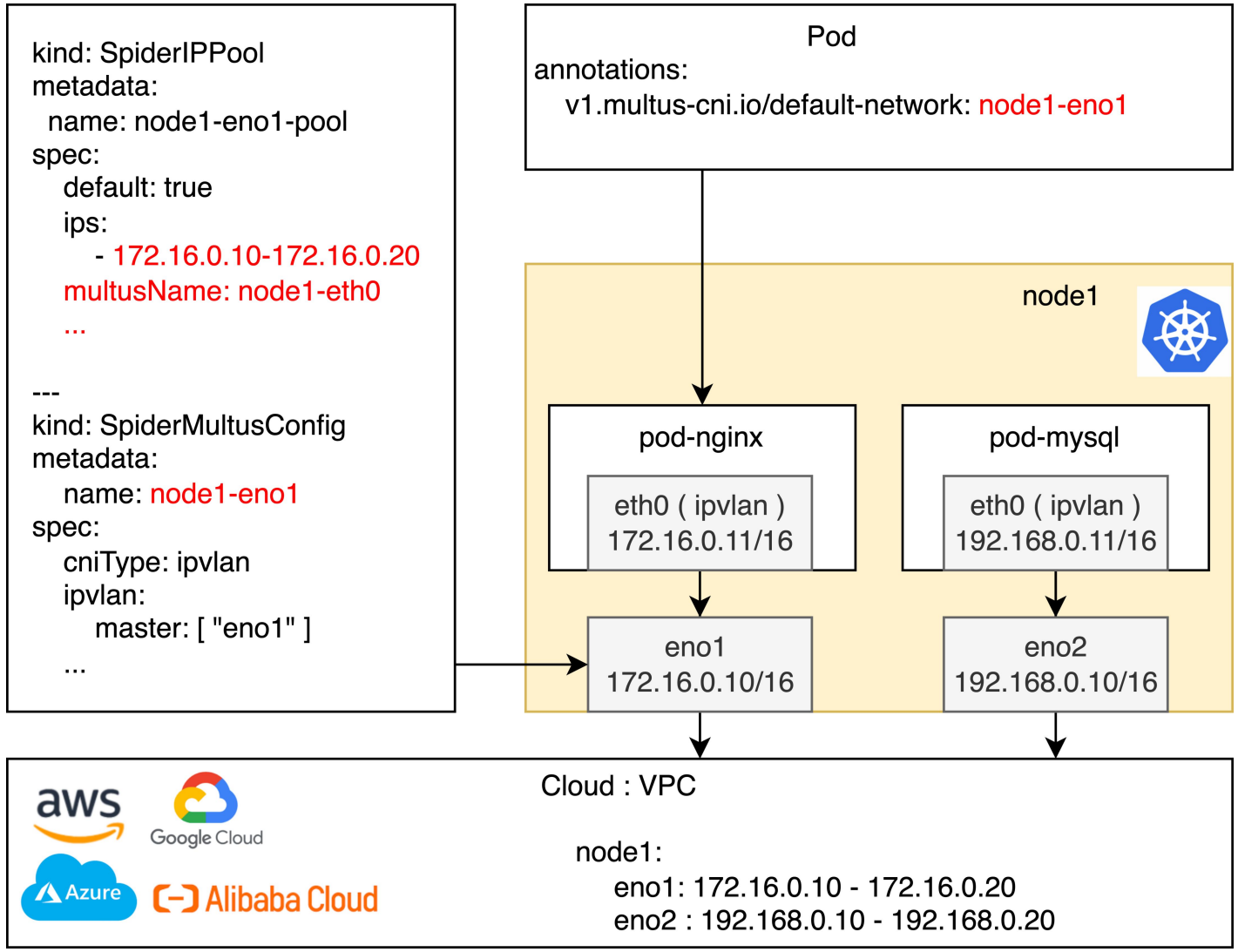
Owing to IP and Mac address reason in the VPC network, few options of Underlay CNI on Cloud :

- Cilium on AWS, ACK, Azure, GKE
- Underlay CNI from Cloud vendor

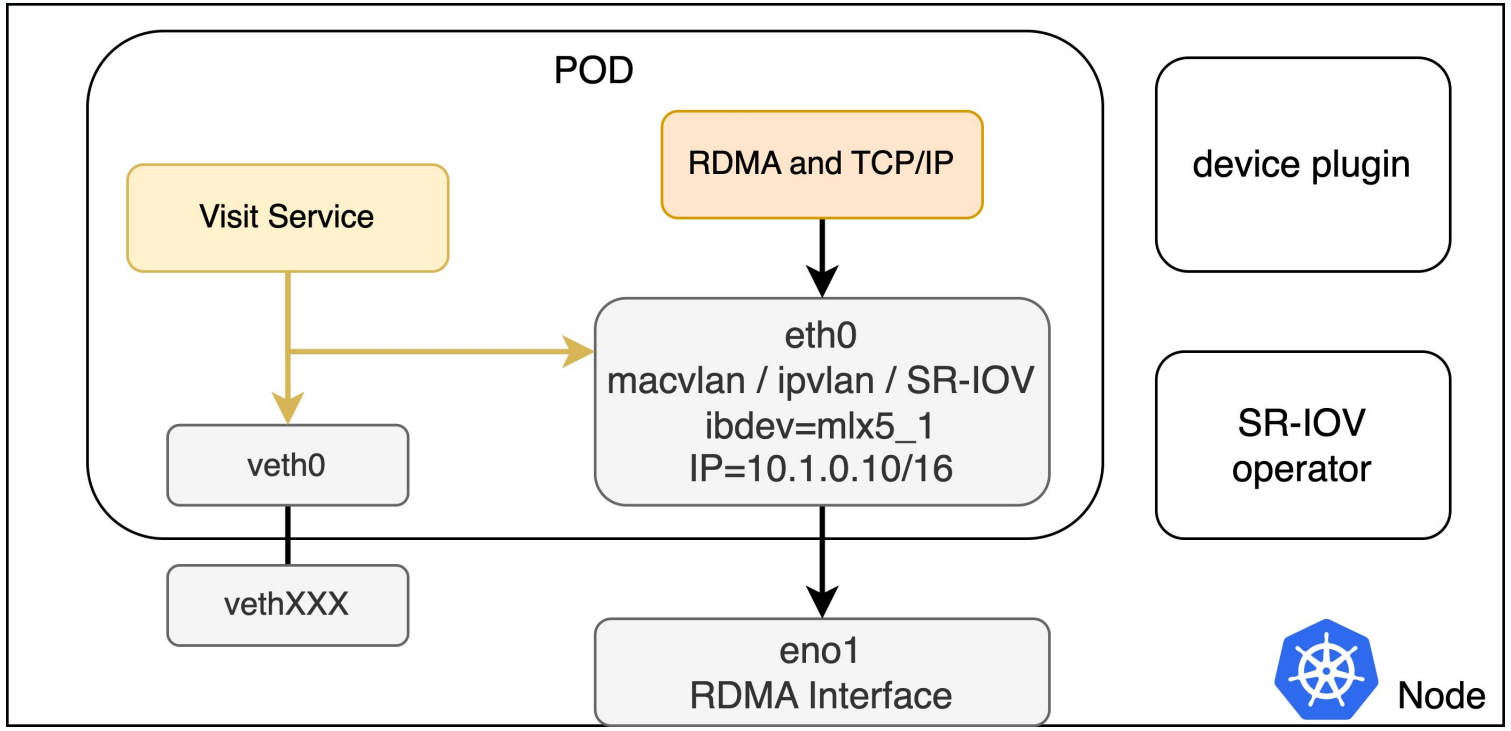
Spiderpool :

Combining ipvlan CNI with node-topology based IPAM, Spiderpool provides an unified solution of underlay CNI on any cloud platforms, especially for hybrid cloud.

Verified: AWS, ACK



RDMA



RDMA provides good performance :

- Low network delay
- High throughput
- Free CPU to serve more applications

Spiderpool supports solutions :

- RoCE with macvlan/ipvlan (shared mode of RDMA)
- RoCE with SR-IOV (exclusive mode of RDMA)
- Infiniband with IB-SR-IOV (exclusive mode of RDMA, coming soon)
- Infiniband with IPoIB (coming soon)

Egress Policy (Experimental)

Scenario :

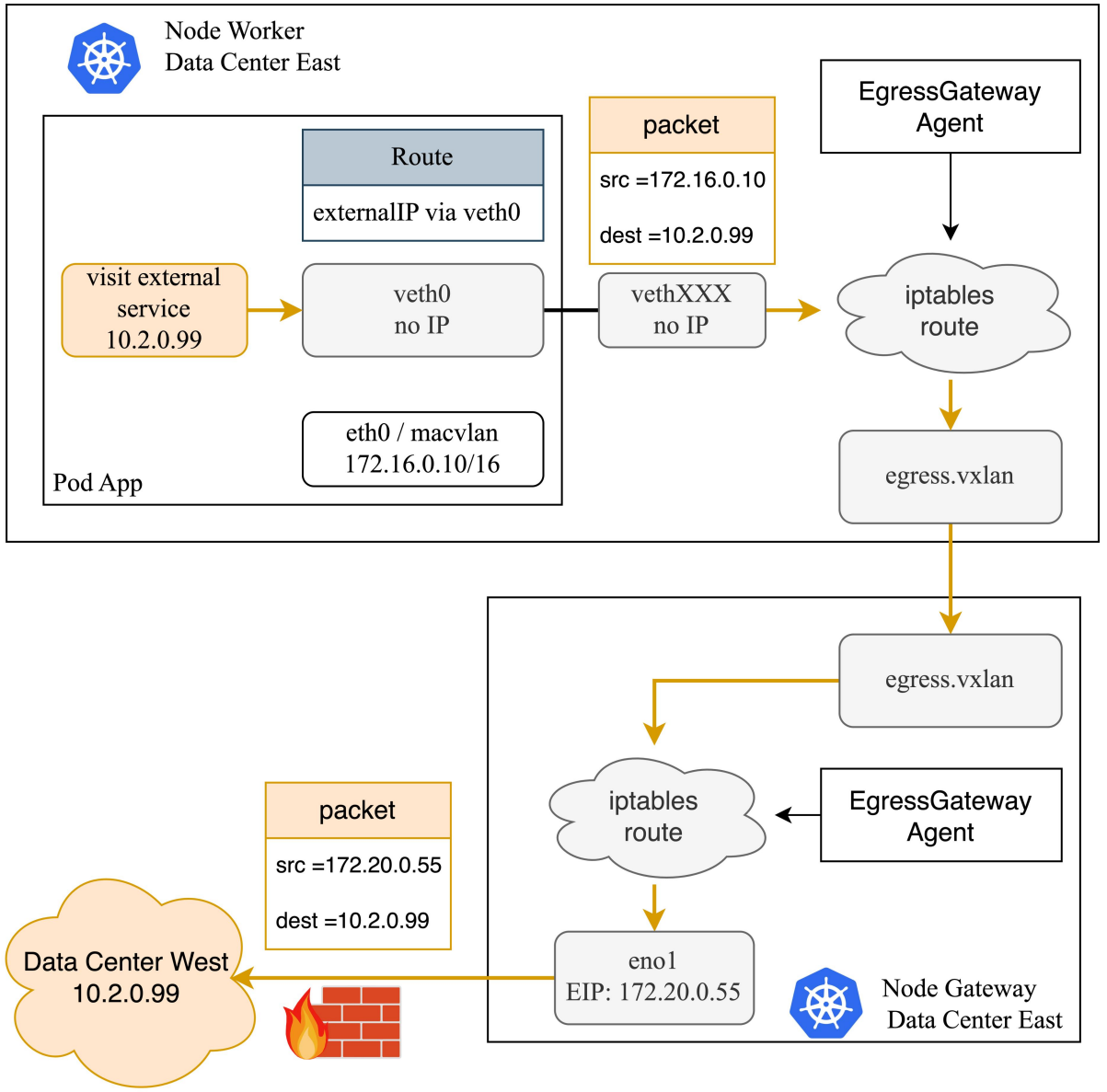
- For visits across data centers, the firewall implements source IP filter

Introduce a new project EgressGateway to implement EgressPolicy feature for Spiderpool.

Features :

- Support Spiderpool, Calico, Flannel, Weave
- Shared or exclusive EIP
- Multiple GatewayClass instances
- Support active-active gateway node
- Support TCP and UDP
- Dual-stack support

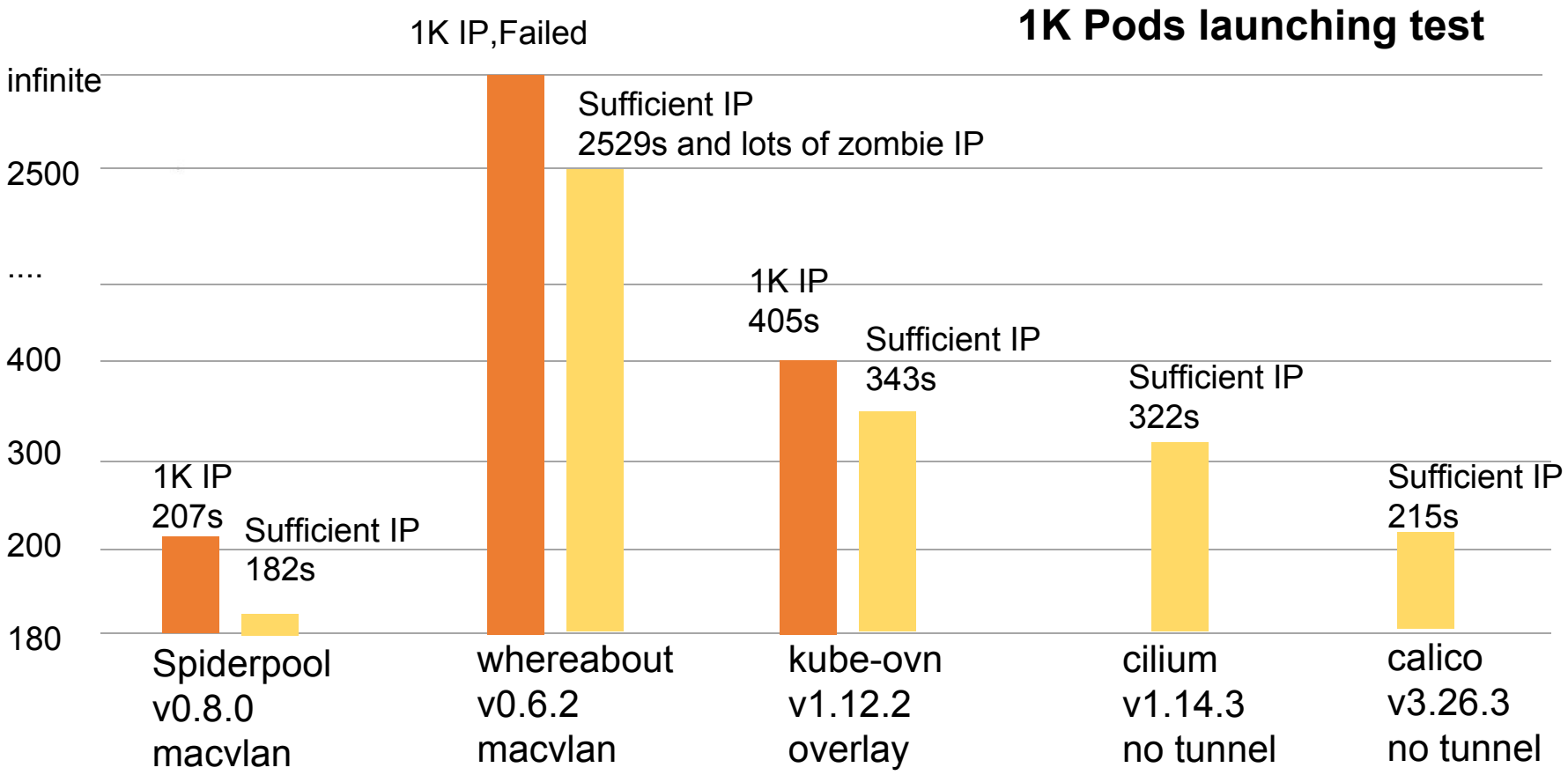
BTW: EgressGateway is not production-ready



Test : 1000 IPs and Pods

A test calculates the time cost to launch 1K Pods with the 1K IP or sufficient IP, which evaluates IPAM efficiency and stability.

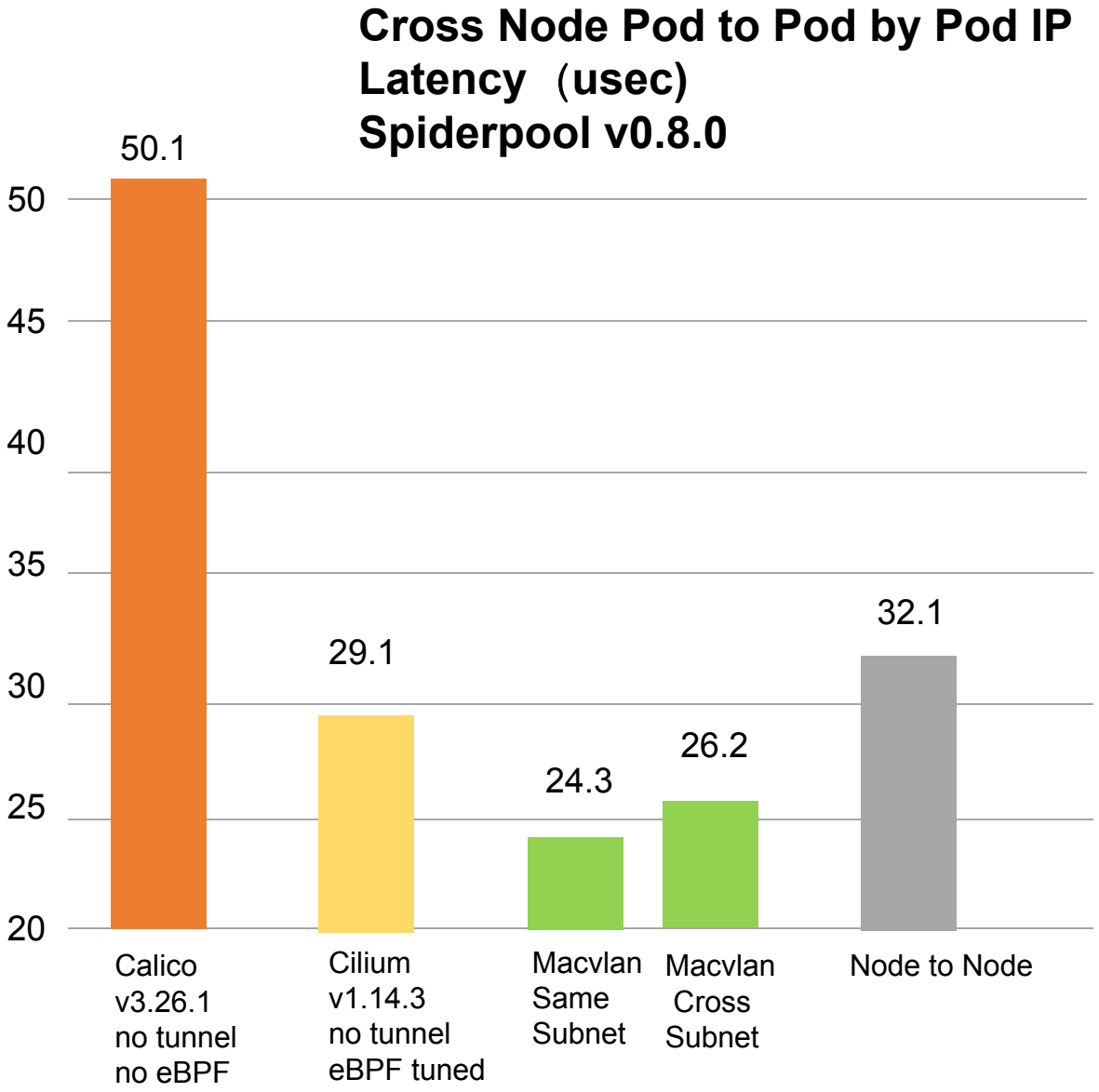
- no IP conflict and wastage
- good efficiency of assigning and relasing IP

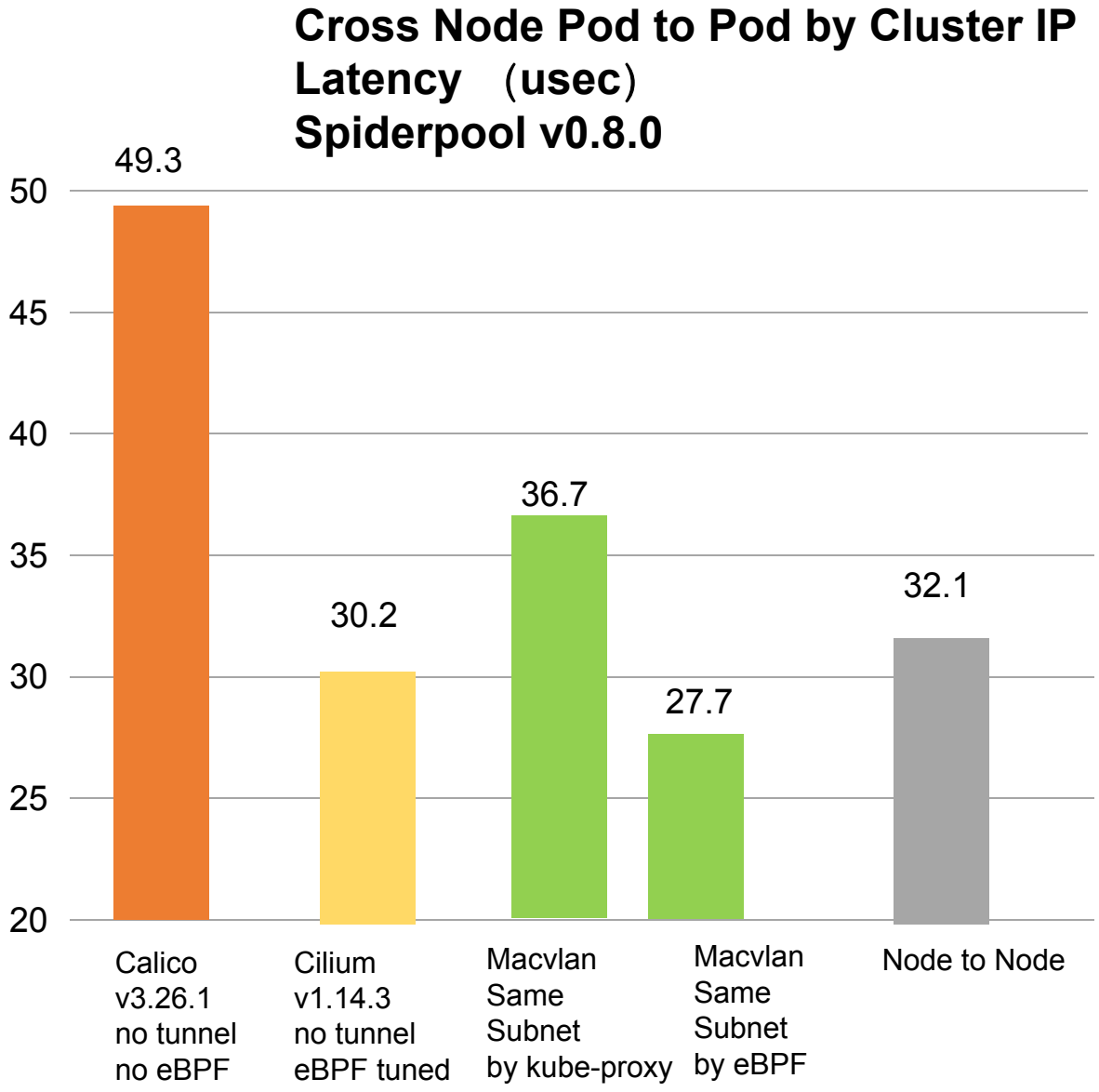


Environment :

- VM with 3 C / 8 G
- 10 VMs including 3 masters
- Kubernetes v1.26.7
- Assign IPv4 and IPv6 IP to Pod
- Launch 100 deployments with 10 replicas each

Sockperf Latency Performance

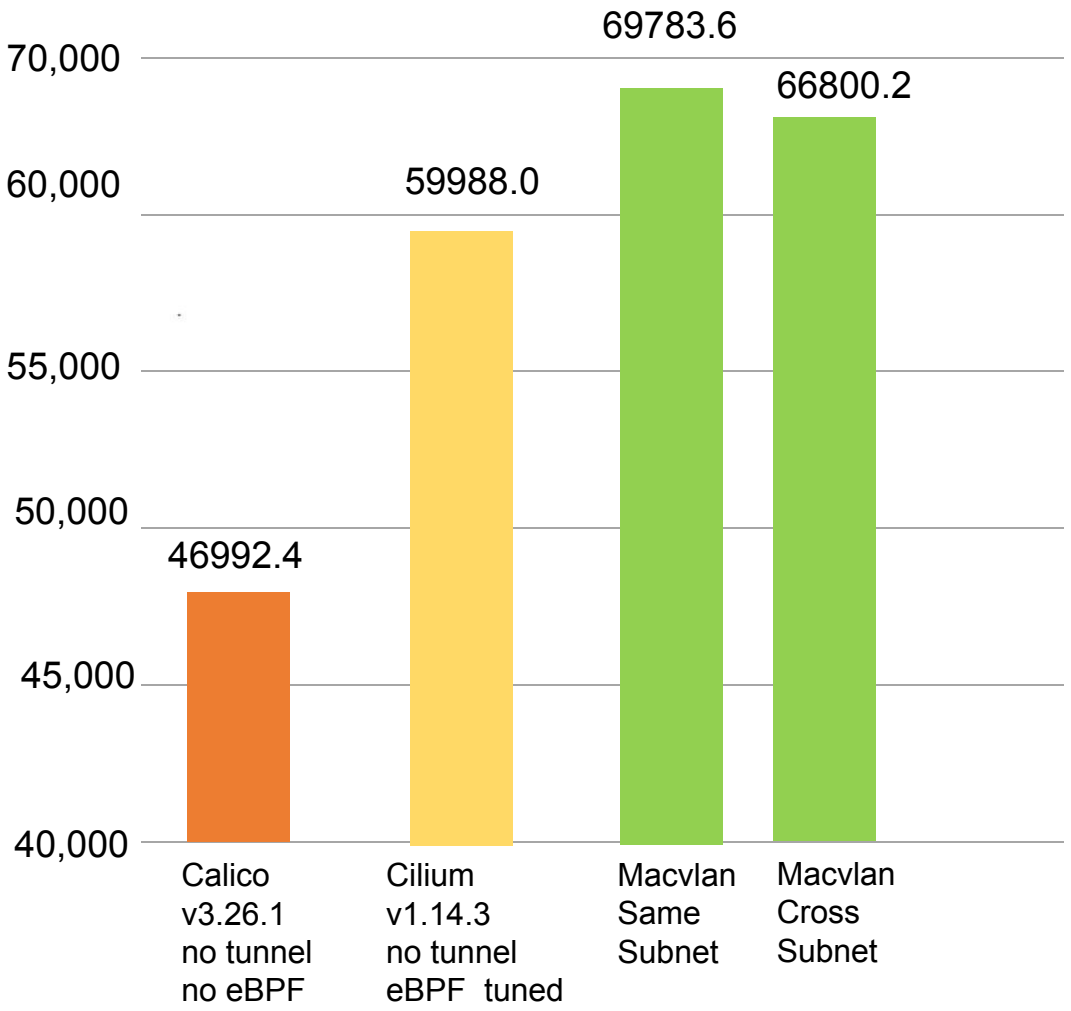




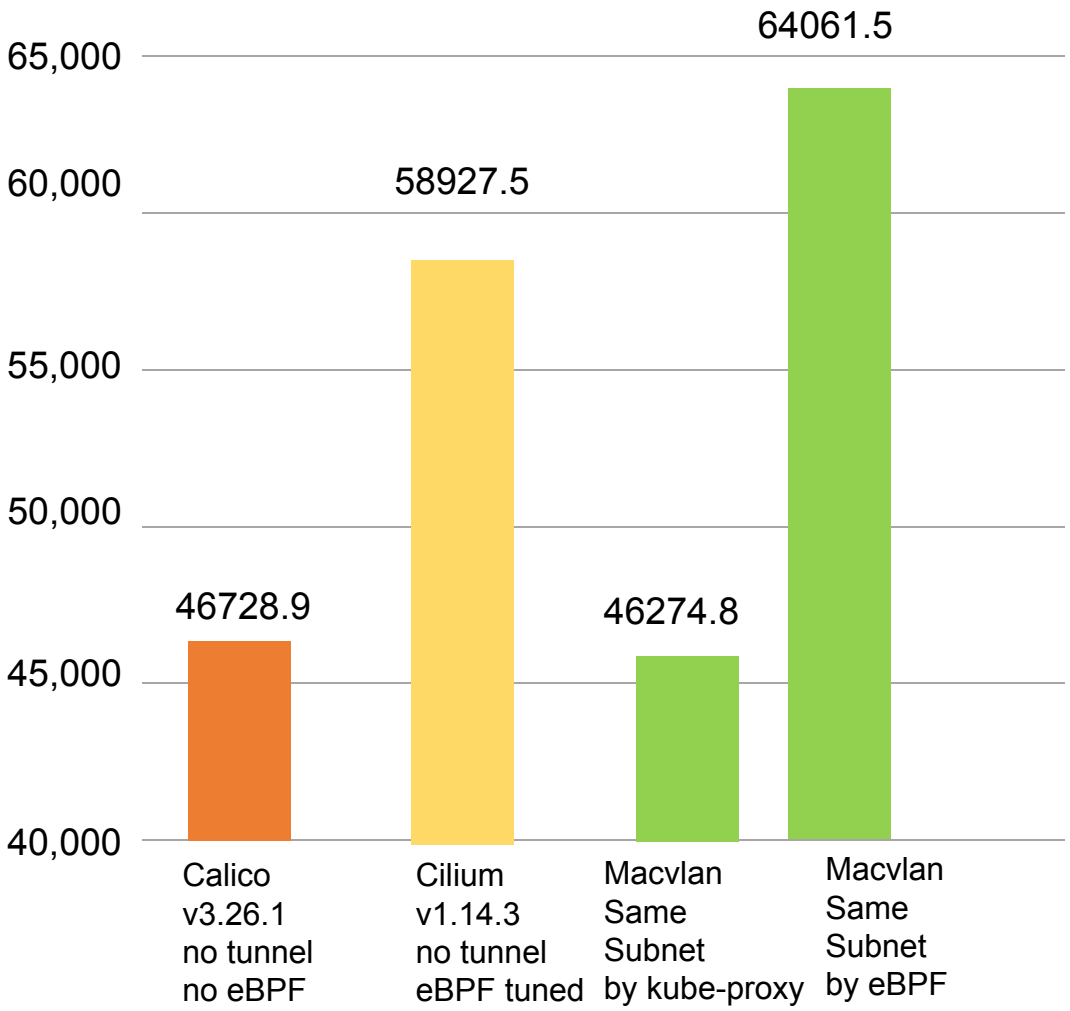
Redis RPS Performance

KubeCon CloudNativeCon North America 2023

Cross-Node Pod to Pod by Pod IP
Get Throughput （rps）
Spiderpool v0.8.0



Cross-Node Pod to Pod by Cluster IP
Get Throughput （rps)
Spiderpool v0.8.0



Contact Us

KubeCon CloudNativeCon North America 2023

Spiderpool : https://github.com/spidernet-io/spiderpool
Slack: https://cloud-native.slack.com/messages/spiderpool

EgressGateway : https://github.com/spidernet-io/egressgateway

Please scan the QR Code above
to leave feedback on this session


KubeCon | CloudNativeCon

North America 2023

Q&A