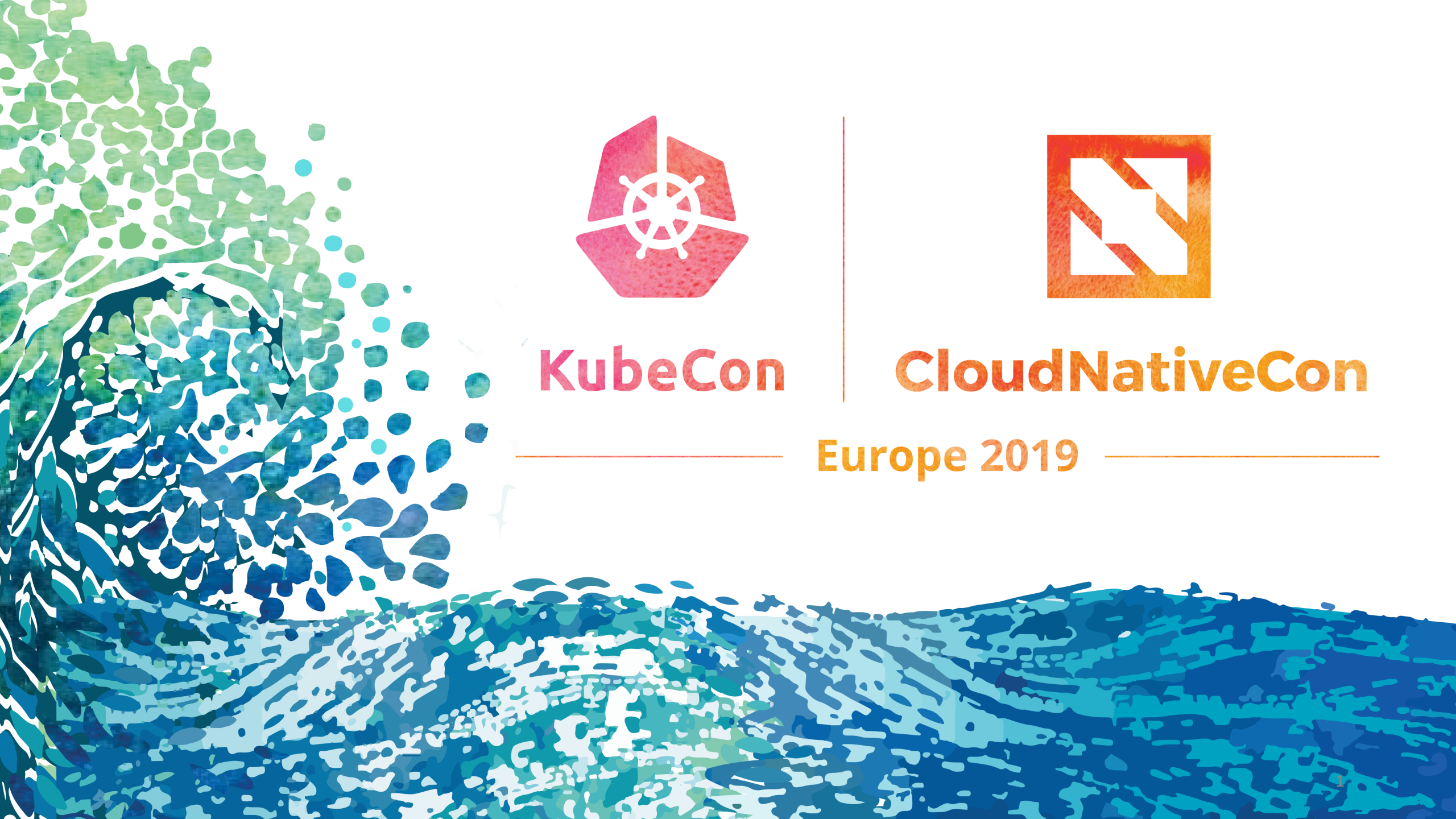
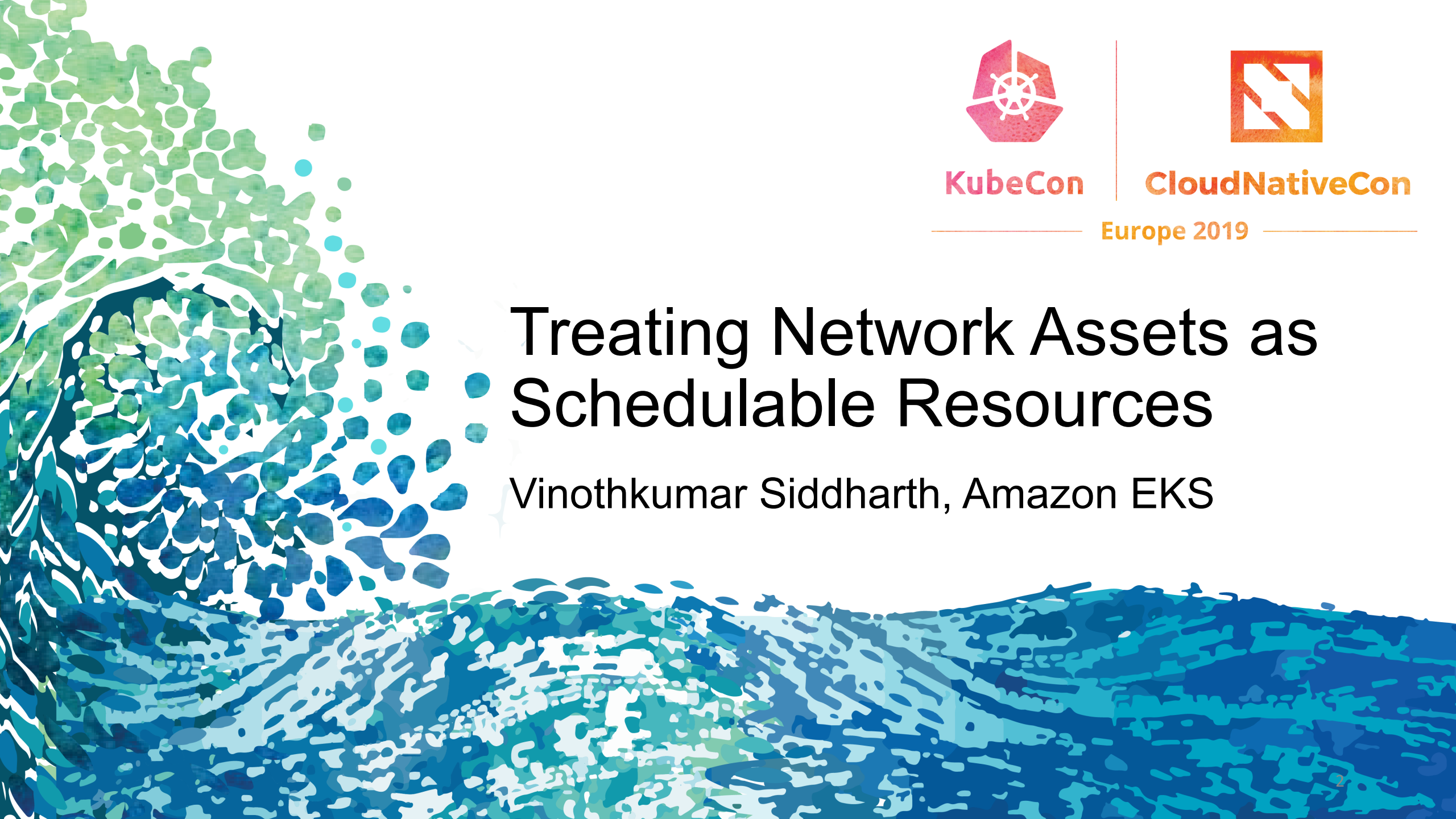

KubeCon
CloudNativeCon
Europe 2019



Treating Network Assets as Schedulable Resources

Vinothkumar Siddharth, Amazon EKS

Who am I?

- Software Engineer
  - Amazon EKS
- Past (Software Engineer)
  - Amazon ECS
  - Citrix Systems R&D, Xen/XAPI Project

Agenda

- Overview
  - CNI
  - Container Scheduling
- Key Challenges
- Solution
- Demo

Overview of CNI

What is CNI?

- Highly overloaded term!

What is CNI?

- What is CNI?
  - Spec (v0.3.1)
    - Add
    - Del
    - Version
  - Spec (v0.4.0)

What is CNI?

- What is CNI?
  - Collection of plugins
    - bridge, loopback, ptp, vlan, …
    - win-bridge, win-overlay
    - dhcp, host-local, static
    - flannel, portmap,

What is CNI?

- What is CNI?
  - Library

CNI Recap

- What is CNI?
  - Spec (v0.3.1)
    - Add
    - Del
    - Version
  - Library
  - Collection of base plugins

- More!

CNI Kubelet Interaction

- How kubelet and CNI interact ?
  - At least one invocation of the CNI plugin binary per Sandbox.

CNI Kubelet Configuration

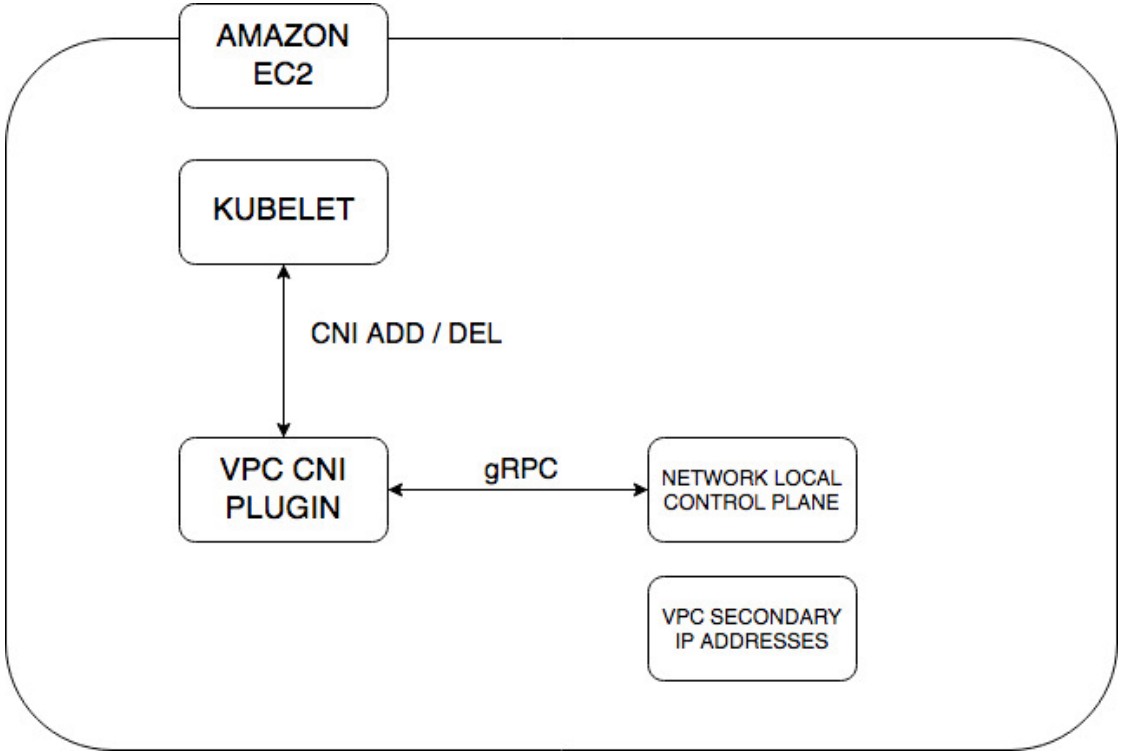
- Kubelet Configuration
  - “--network-plugin”
  - “--cni-bin-dir”
  - “--cni-conf-dir”

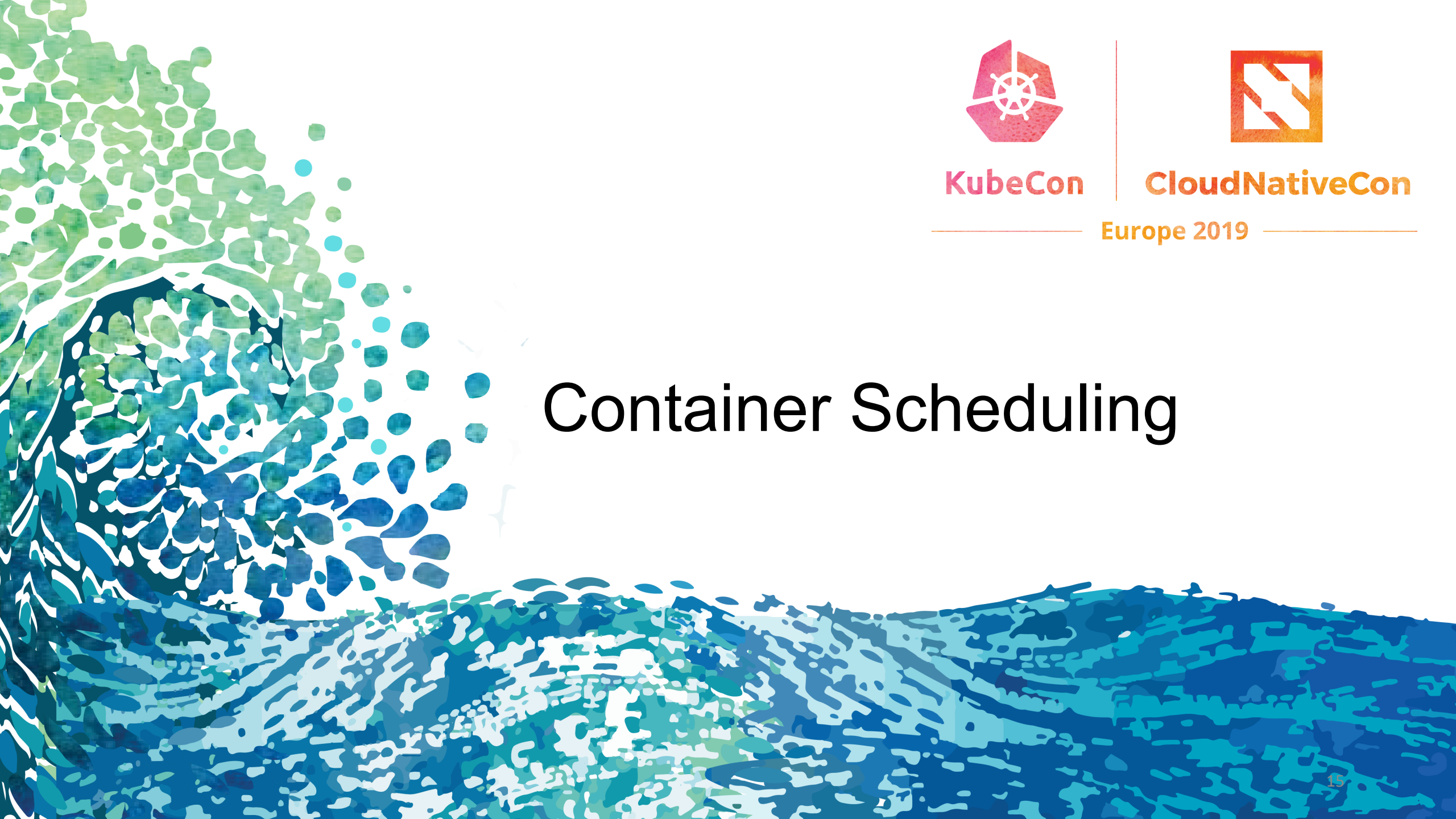
CNI Deployment

- Typical CNI deployments
  - CNI plugin binary
  - CNI config
  - Node level daemon

CNI Deployment

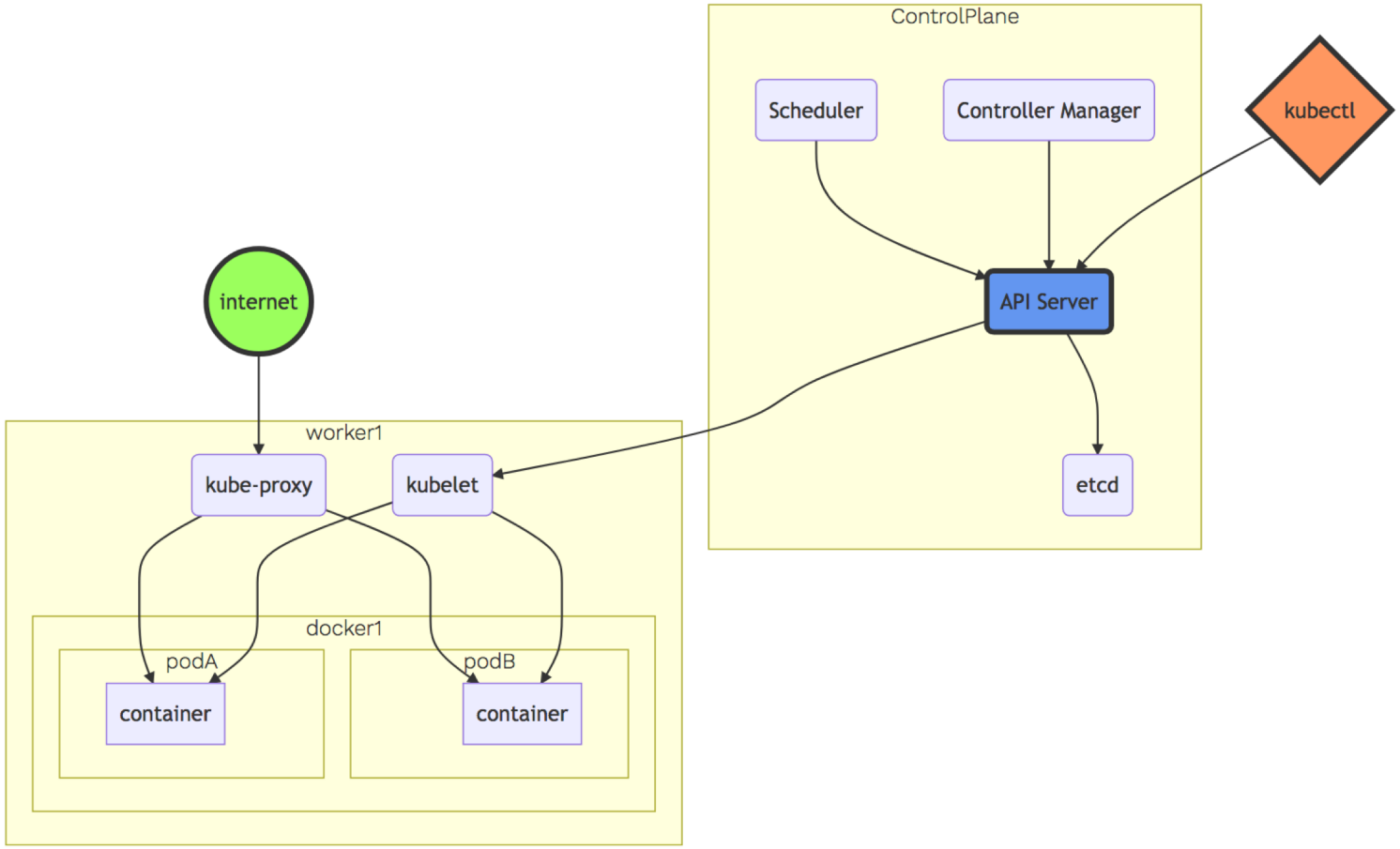
- Typical CNI deployments
  - CNI plugin binary
  - CNI config
  - Node level daemon



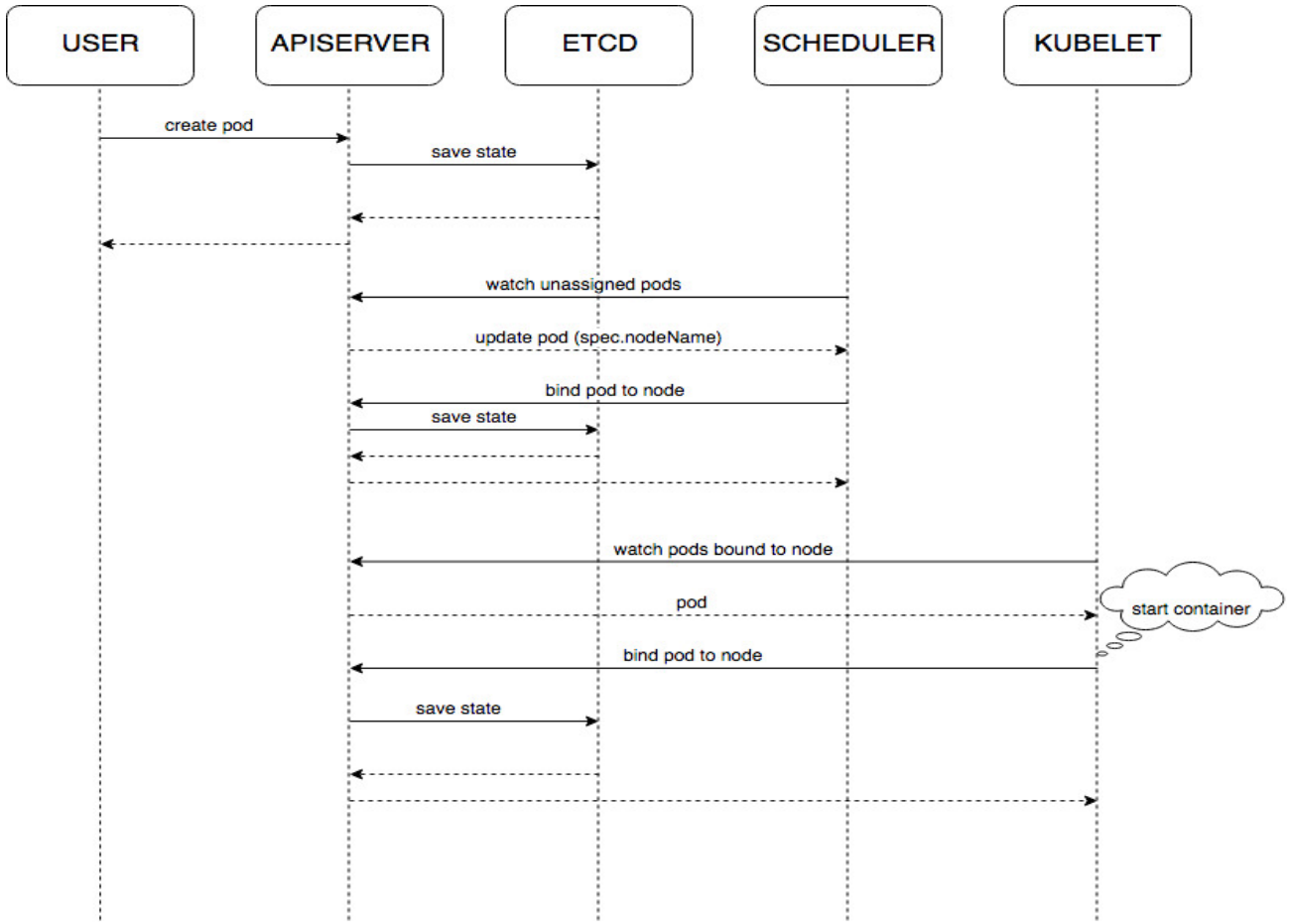


Container Scheduling

Kubernetes Architecture



Kube-Scheduler



Kube-Scheduler

- Find nodes for pods in a cluster

- Complex and feature rich
- Multiple dimensions to optimize

- Default Scheduler:
  - Accounts for CPU and memory in the scheduling process!
  - Supports resource based scheduling

Kube-Scheduler: Resource Based


```
apiVersion: v1
kind: Pod
spec:
  containers:
  - command: …
    resources:
      limits:
        cpu:  “500m”
        memory: “1500Mi”
      requests:
        cpu:  “500m”
        memory: “1500Mi”
```

Custom Schedulers

- Kubernetes provides a framework to incorporate custom schedulers for special needs!
  - Typical deployment managed by the default scheduler.
  - spec.schedulerName: <special-purpose-scheduler>

- Kubernetes enables multiple schedulers within the cluster.

Challenges

Challenges

- How to account for networking assets in the scheduling process?

Challenges

- How to eliminate the node level agent running on worker nodes?

Challenges

- How to minimize the permissions required to manage the lifecycle of networking assets?

KubeCon | CloudNativeCon Europe 2019

Solution

Solution

- Use extended resources
- Custom resource controllers
- Admission webhooks
- CNI plugin

VPC Resource Controller

- Manages VPC resources for the cluster
- Pluggable resource model
- Provider interface
- Platform agnostic

VPC Resource Controller: Provider


```
type Provider interface {
   GetResourceName() string
   GetDesiredWarmPoolSize() (int, int)
   InitResourcePool(node Node) (*Pool, error)
   CreateResource(node Node, quantity int) (resourceIDs []string, err error)
   DeleteResource(node Node, resourceID string) error
}
```

VPC Resource Controller: Provider

- Provider Interface Implementations
  - Elastic Network Interface Provider
  - IP Address Provider

VPC Resource Controller: Responsibilities

- Watch for node objects
  - Advertise extended resources

VPC Resource Controller: Responsibilities

- Watch for node objects
  - Advertise extended resources

```
apiVersion: v1
kind: Node
spec:
  providerID: aws:///us-west-2a/i-094fe8fb054fd0b07
status:
  allocatable:
    cpu: "4"
    memory: 8023644Ki
    pods: "110"
    vpc.amazonaws.com/ENI: "1"
    vpc.amazonaws.com/PrivateIPv4Address: "14"
  capacity:
    cpu: "4"
    memory: 8023644Ki
    pods: "110"
    vpc.amazonaws.com/ENI: "1"
    vpc.amazonaws.com/PrivateIPv4Address: "14"
```

VPC Resource Controller: Responsibilities

- Watch for pod objects
  - Annotate pod spec with metadata

VPC Resource Controller: Responsibilities

- Watch for pod objects
  - Annotate pod spec with metadata

```
apiVersion: v1
kind: Pod
metadata:
  annotations:
    vpc.amazonaws.com/PrivateIPv4Address: 192.168.113.175
  name: windows-servercore-webserver-5659f96674-cts74
  namespace: default
spec:
  containers:
  - command: …
    name: windows-servercore-webserver
    resources:
      limits:
        vpc.amazonaws.com/PrivateIPv4Address: "1"
      requests:
        vpc.amazonaws.com/PrivateIPv4Address: "1"
```

VPC Admission Webhook

- Inject extended resource requirements for relevant pods

VPC Admission Webhook

- Inject extended resource requirements for relevant pods

```
apiVersion: v1
kind: Pod
metadata:
  annotations:
    vpc.amazonaws.com/PrivateIPv4Address: 192.168.113.175
  name: windows-servercore-webserver-5659f96674-cts74
  namespace: default
spec:
  containers:
  - command: …
    name: windows-servercore-webserver
    resources:
      limits:
        vpc.amazonaws.com/PrivateIPv4Address: "1"
      requests:
        vpc.amazonaws.com/PrivateIPv4Address: "1"
```

CNI Plugin: VPC Shared ENI

- Simple executable (binary)
- Read pod metadata
  - Annotations

- Provides
  - Connectivity
  - Reachability

KubeCon | CloudNativeCon

Europe 2019

Demo

Benefits

- Incorporate network resources in the scheduling process
- Eliminate long running node local agents
- Obtain cluster level network resource accounting
- Easy to add support for new VPC resource abstractions
- Reduced set of permissions on worker nodes

KubeCon | CloudNativeCon Europe 2019

Thank You!