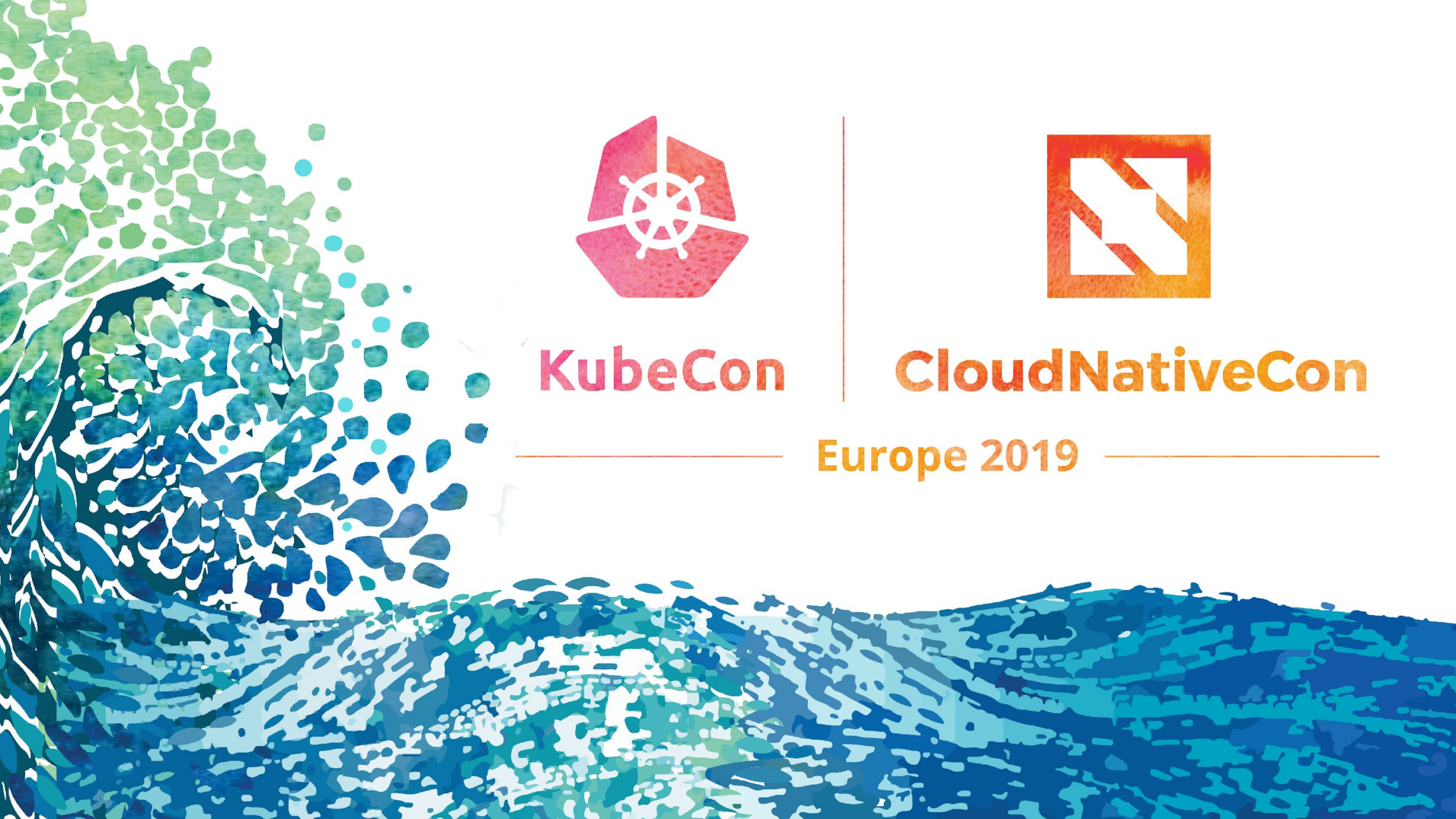


KubeCon | CloudNativeCon

Europe 2019

KubeCon | CloudNativeCon

Europe 2019

Ready? A Deep Dive into Pod Readiness Gates for Service Health Management

Minhan Xia, Software Engineer, Google
Ping Zou, Software Engineer, Intuit

Agenda

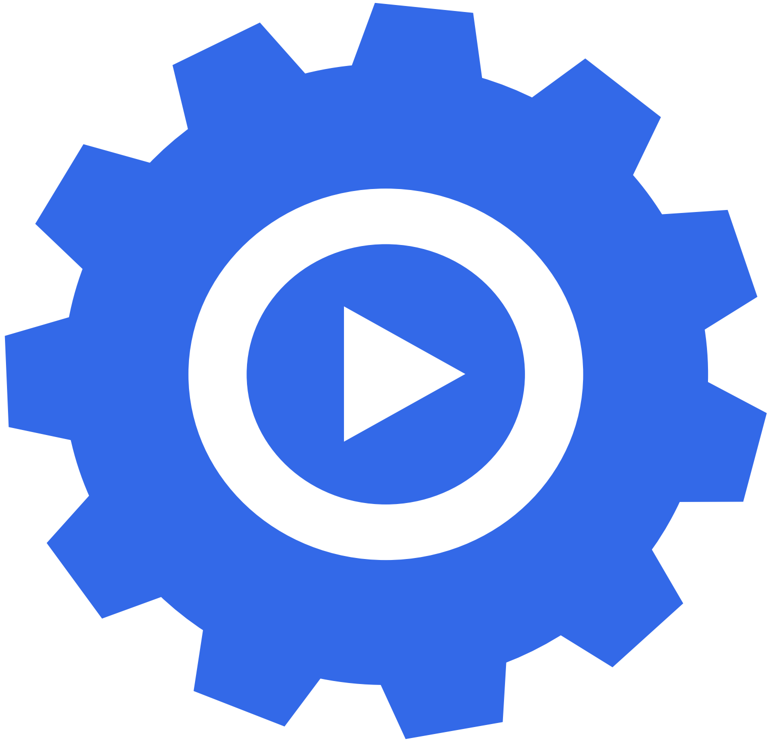
- *Pod Status Recap*
- *Pod ReadinessGate Intro*
- *Kubernetes Engine Use Case*
- *Foremast Use Case*

Pod Status Recap

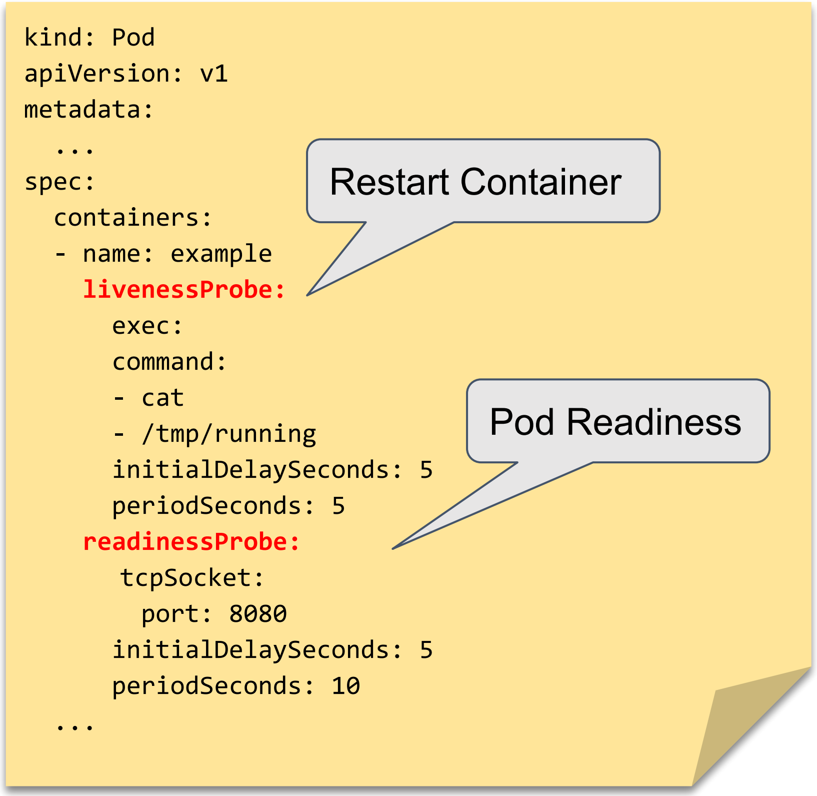
Container Status

```
kind: Pod
apiVersion: v1
metadata:
  ...
spec:
  ...
status:
  ...
  containerStatuses:
  - containerID: docker://xxxxxxxxxxxxxxxxxxxxx
    image: k8s.gcr.io/busybox
    imageID: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
    name: example
    ready: true
    restartCount: 0
    state:
      running:
        startedAt: "2019-05-21T00:00:00Z"

  ...
```



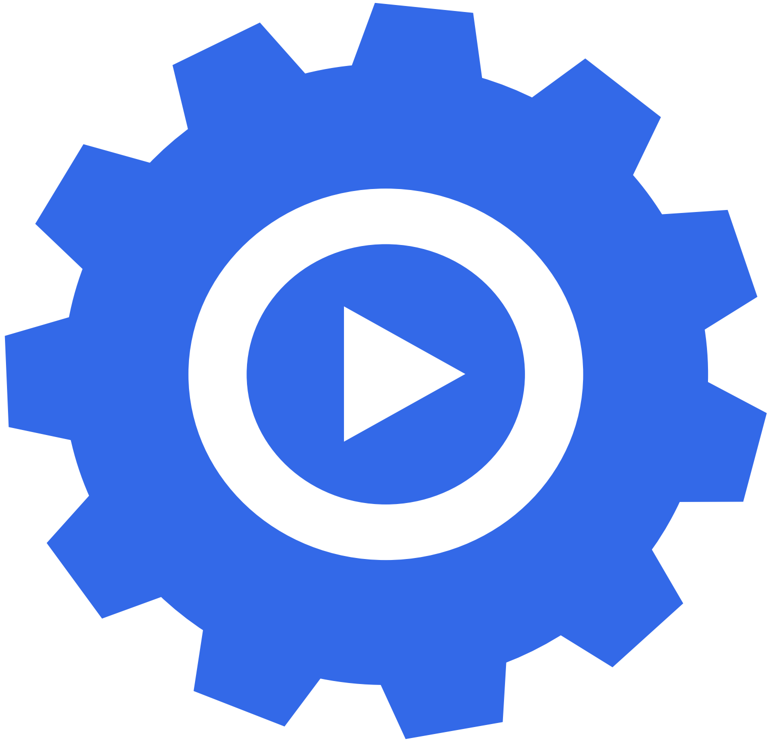
Container Status



```
kind: Pod
apiVersion: v1
metadata:
 ...
spec:
 containers:
 - name: example
  livenessProbe:
   exec:
   command:
   - cat
   - /tmp/running
   initialDelaySeconds: 5
   periodSeconds: 5
  readinessProbe:
   tcpSocket:
    port: 8080
   initialDelaySeconds: 5
   periodSeconds: 10
 ...
```

Restart Container

Pod Readiness



Pod Status

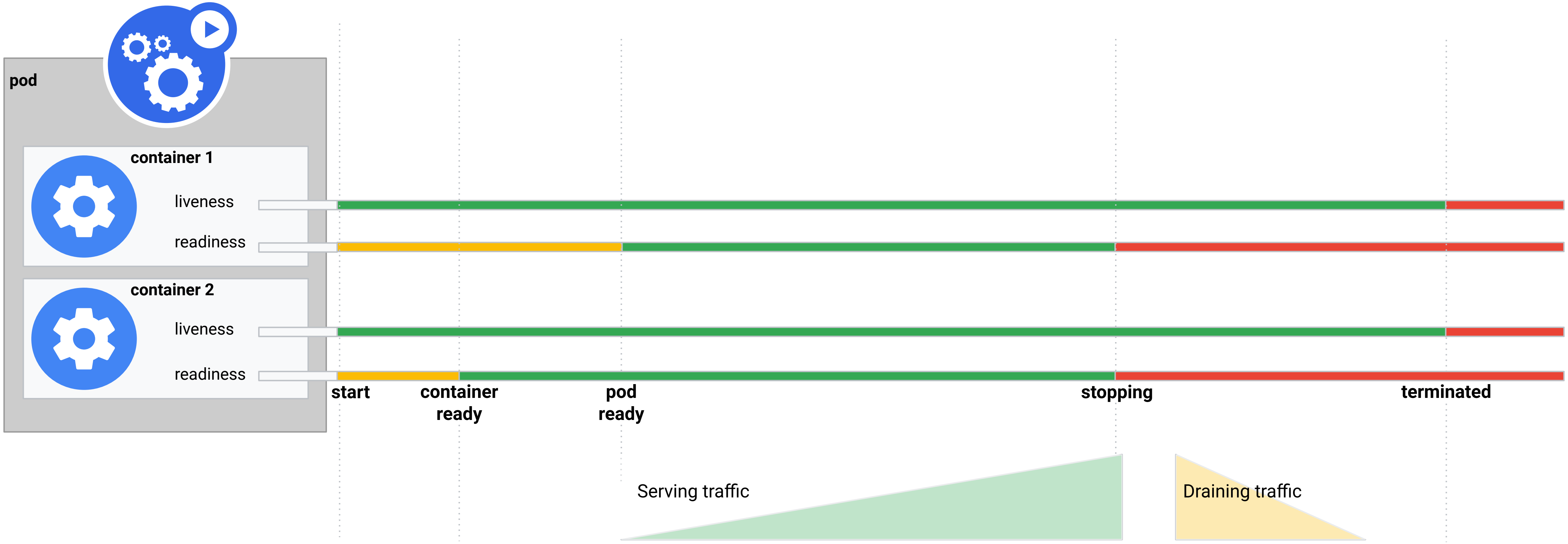
```
kind: Pod
apiVersion: v1
metadata:
  ...
spec:
  ...
status:
  conditions
    - type: PodScheduled
      status: "True"
      lastTransitionTime: "2019
    - type: Initialized
      status: "True"
      lastTransitionTime:
    - type: Ready
      status: "True"
      lastTransitionTime: "2019-05-21T00:01:00Z"
      ...
  phase: Running
  ...
```

Pod has been scheduled to a node

all init containers have started successfully

all containers are ready

Pod LifeCycle



Pod Readiness

All Containers are ready

=

Pod is ready

=

Pod is ready to serve traffic

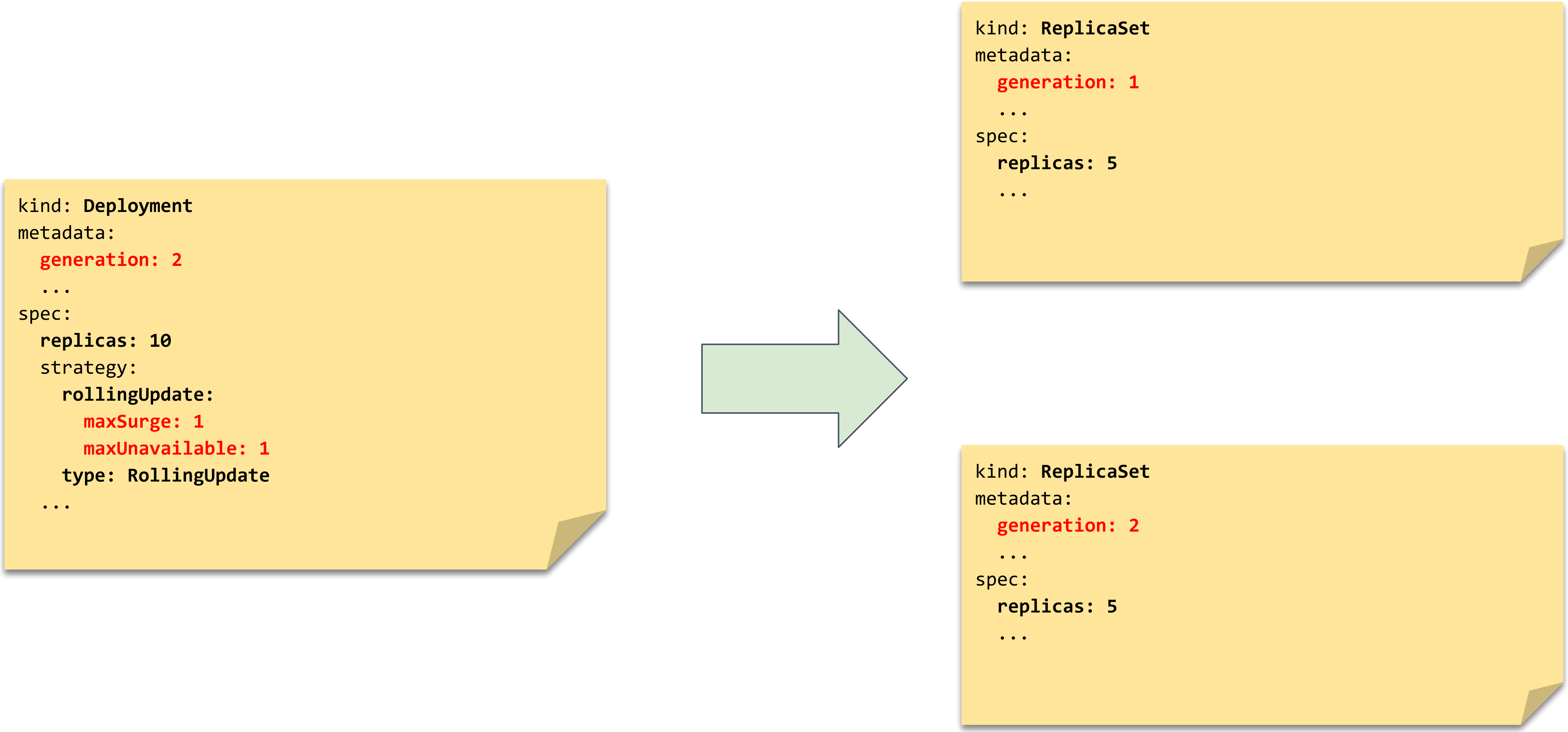
=

?

Pod Readiness Consumer: Workload

```
kind: Deployment
metadata:
  ...
spec:
  replicas: 10
  strategy:
    rollingUpdate:
      maxSurge: 1
      maxUnavailable: 1
    type: RollingUpdate
  ...
```

Deployment Rolling Update

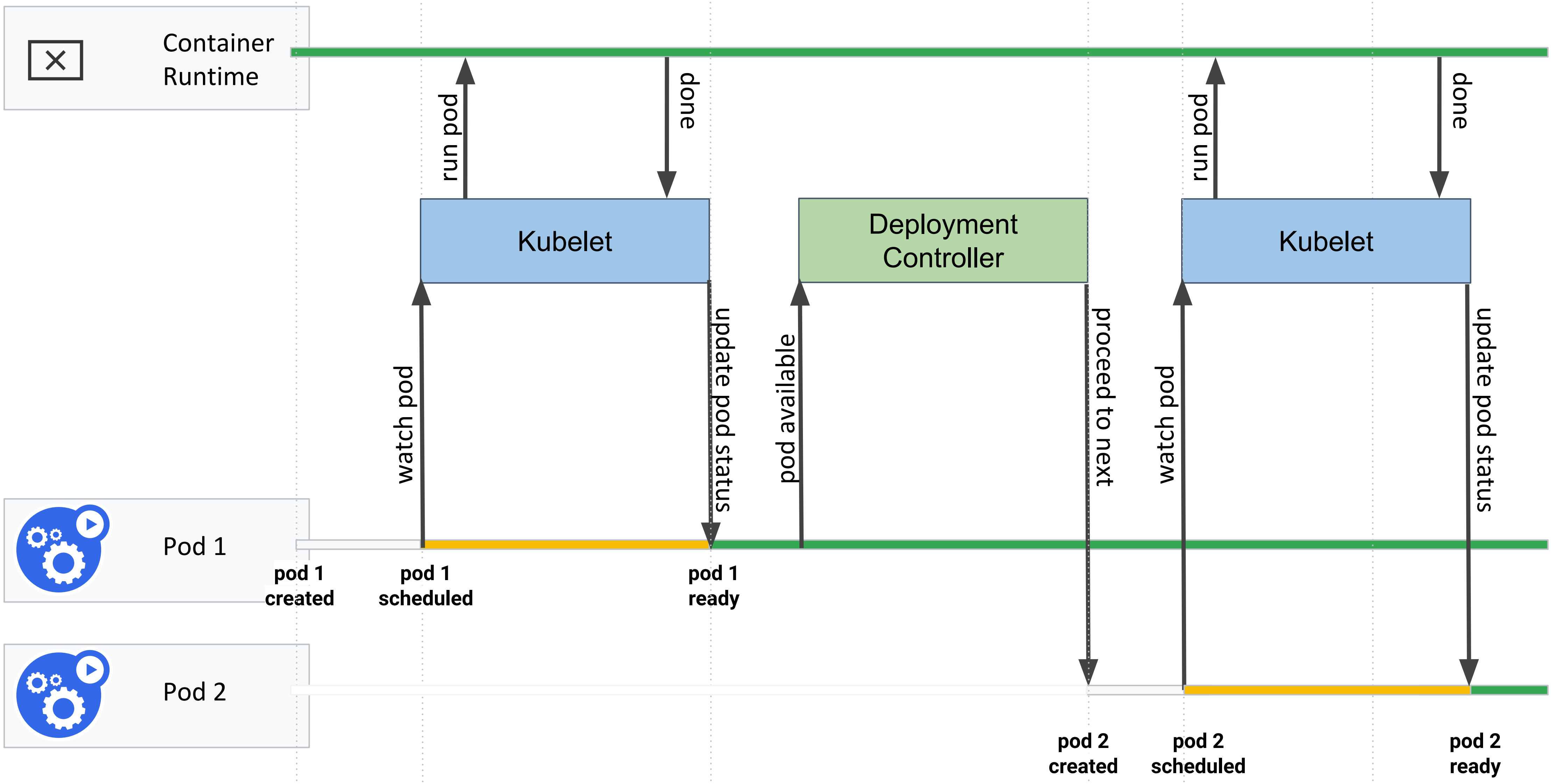


```
kind: Deployment
metadata:
  generation: 2
  ...
spec:
  replicas: 10
  strategy:
    rollingUpdate:
      maxSurge: 1
      maxUnavailable: 1
    type: RollingUpdate
  ...
```

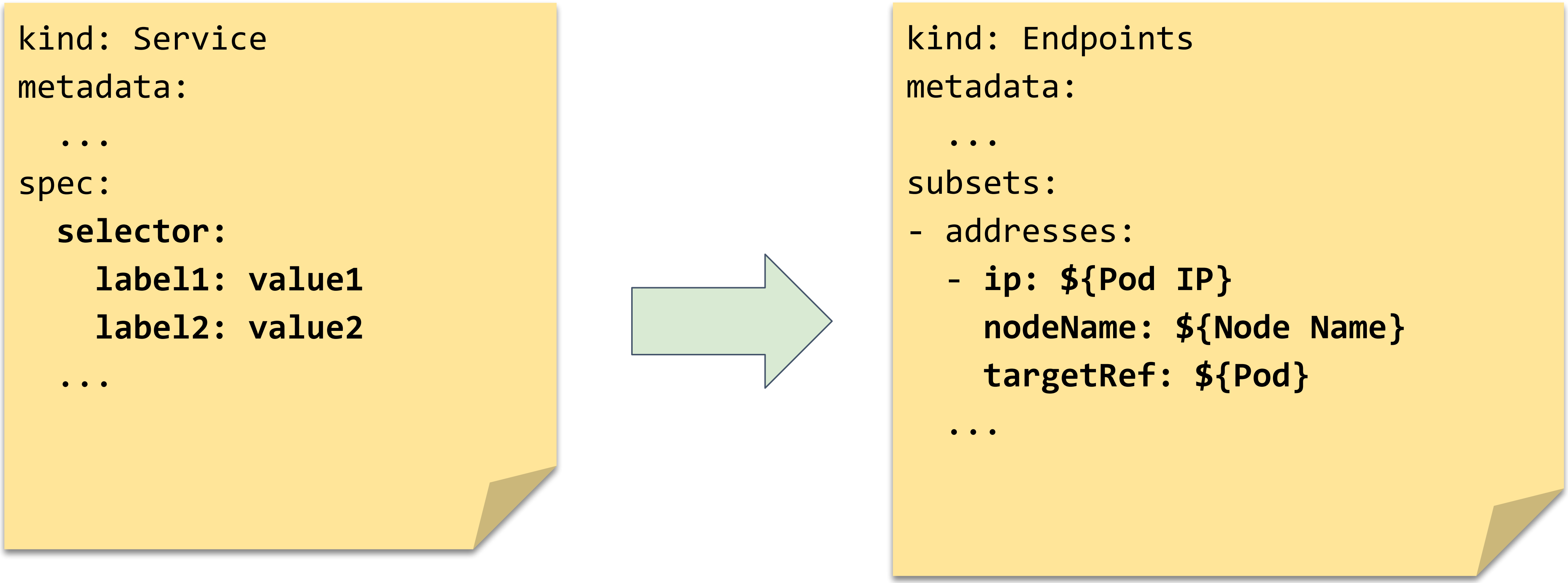
```
kind: ReplicaSet
metadata:
  generation: 1
  ...
spec:
  replicas: 5
  ...
```

```
kind: ReplicaSet
metadata:
  generation: 2
  ...
spec:
  replicas: 5
  ...
```

Deployment Rolling Update



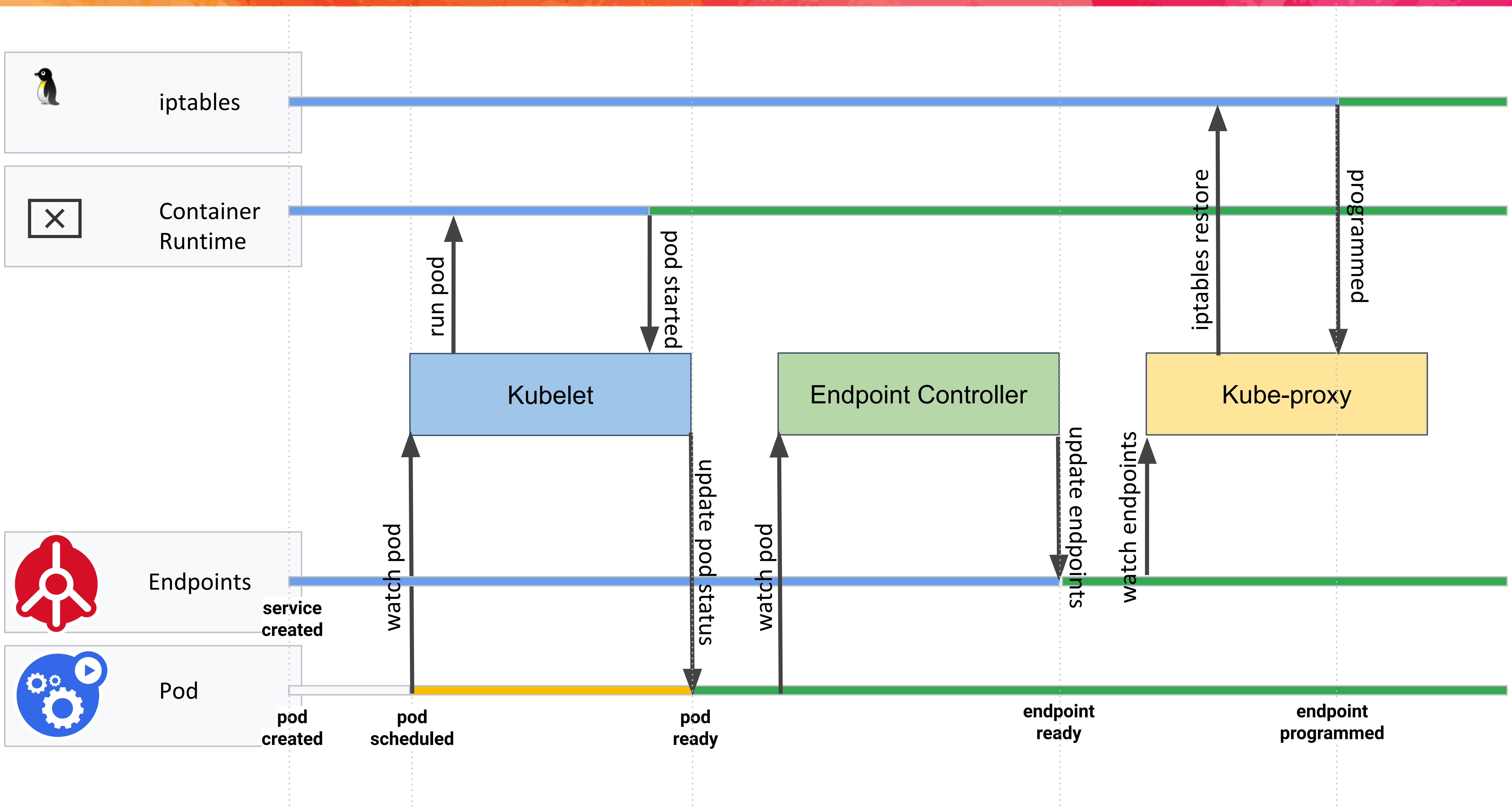
Pod Readiness Consumer: Service



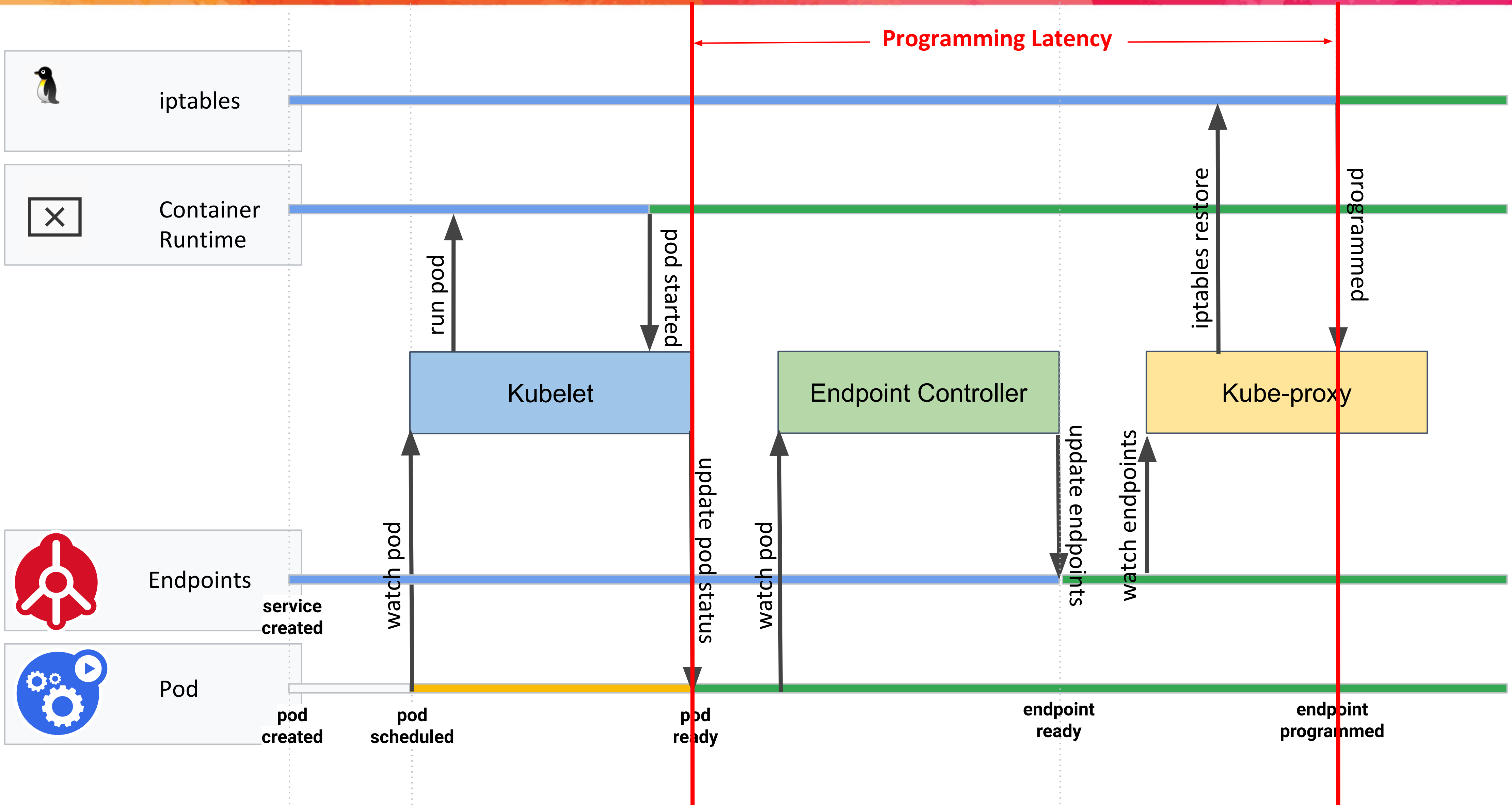
```
kind: Service
metadata:
  ...
spec:
  selector:
    label1: value1
    label2: value2
  ...
```

```
kind: Endpoints
metadata:
  ...
subsets:
- addresses:
  - ip: ${Pod IP}
    nodeName: ${Node Name}
    targetRef: ${Pod}
  ...
```

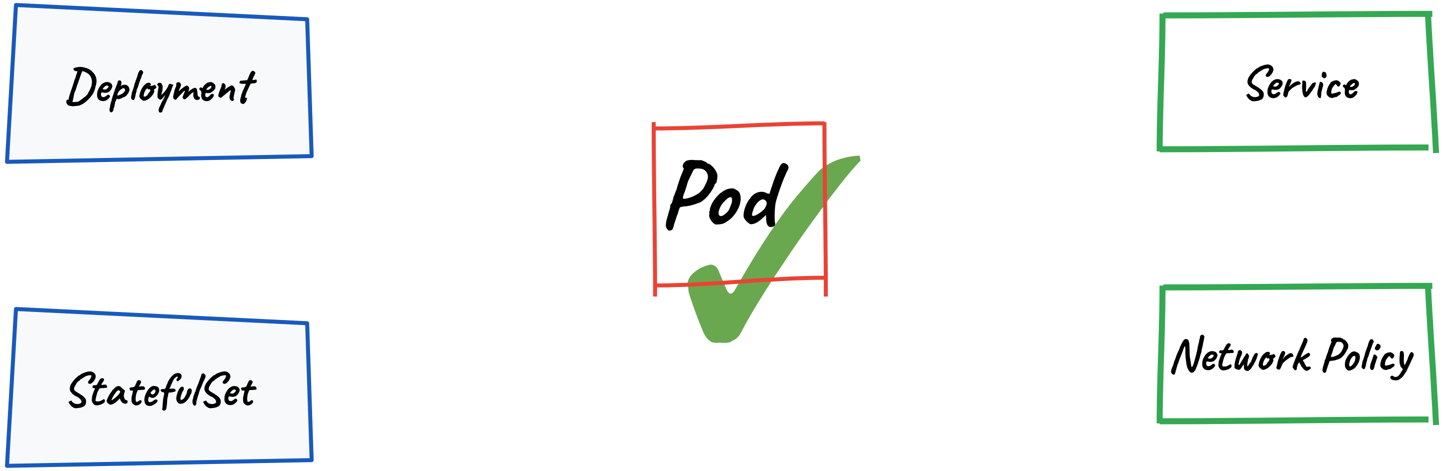
Pod Readiness Consumer: Service



Pod Readiness Consumer: Service



Rendezvous



Deployment

StatefulSet

Pod ✓

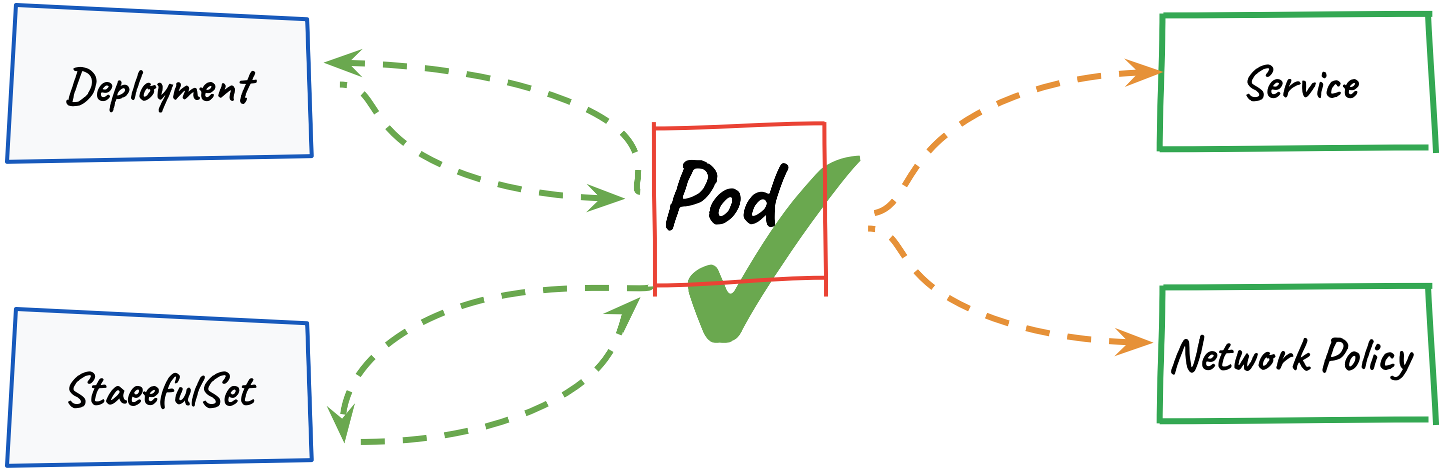
Service

Network Policy

Workload vs. Network Abstractions

Do they work actually together?

Workload vs. Network Abstractions



KubeCon | CloudNativeCon
Europe 2019

Pod ReadinessGate Intro

Pod Ready++?

What if kubelet cannot determine pod readiness?

How to make workloads network aware?

How do service health management solutions better integrate with K8s internal?

Ready++?

Constraints

Backward Compatibility

Backward Compatibility

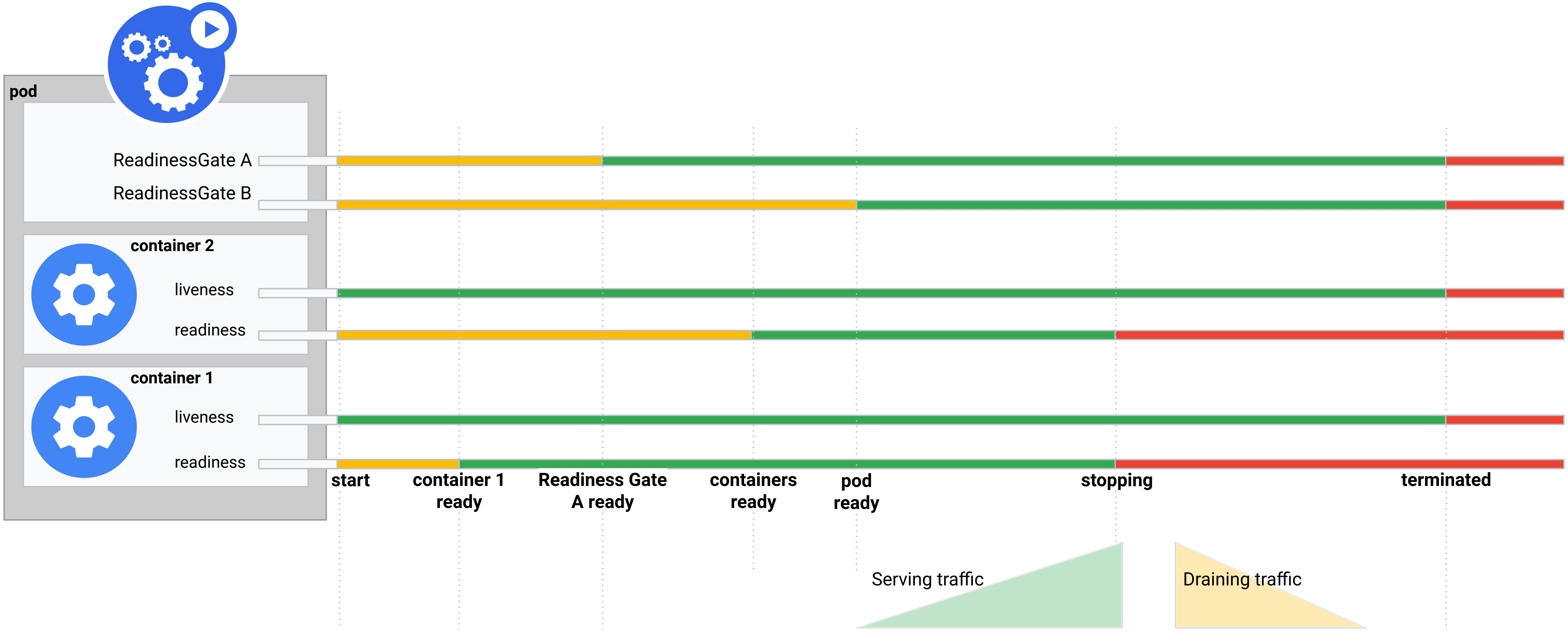
Backward Compatibility

Ready++?

Pod Readiness Gate

```
Kind: Pod
…
spec:
  readinessGates:
  - conditionType: readiness-gate-a
  - conditionType: readiness-gate-b
…
status:
  conditions:
  - lastTransitionTime: 2018-01-01T00:00:00Z
    status: "False"
    type: Ready
  - lastTransitionTime: 2018-01-01T00:00:00Z
    status: "False"
    type: readiness-gate-a
  - lastTransitionTime: 2018-01-01T00:00:00Z
    status: "True"
    type: readiness-gate-b
…
```

Pod LifeCycle with Readiness Gate



Pod Readiness Gate

Pod is Ready

=

All Containers are Ready

AND

All ReadinessGate Conditions are True

Pod Readiness Gate

ContainersReady is True

=

All Containers are Ready

```
Kind: Pod
…
spec:
  readinessGates:
  - conditionType: readiness-gate-a
  - conditionType: readiness-gate-b
…
status:
  conditions:
  - lastProbeTime: null
    lastTransitionTime: 2018-01-01T00:00:00Z
    status: "False"
    type: Ready
  - lastProbeTime: null
    lastTransitionTime: 2018-01-01T00:00:00Z
    status: "True"
    type: ContainersReady
  - lastProbeTime: null
    lastTransitionTime: 2018-01-01T00:00:00Z
    status: "False"
    type: readiness-gate-a
  - lastProbeTime: null
    lastTransitionTime: 2018-01-01T00:00:00Z
    status: "True"
    type: readiness-gate-b
…
```

Kubectl

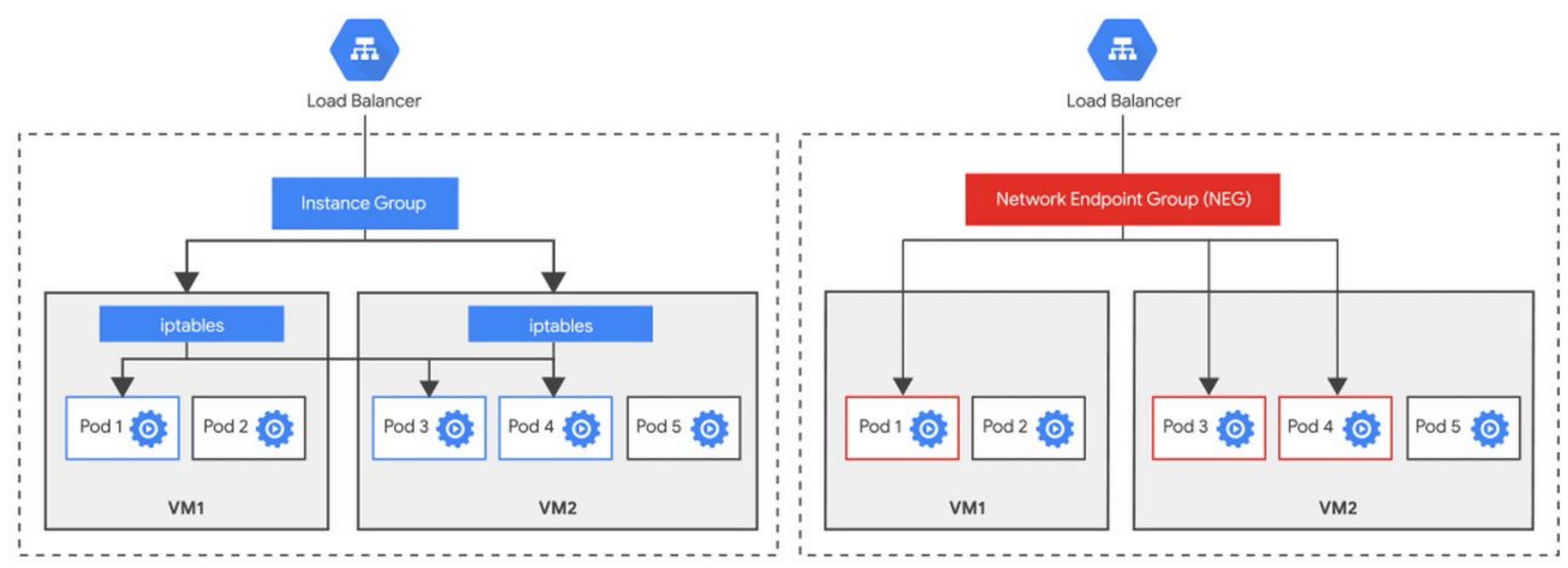
```
$ kubectl get pod -o wide
NAME     READY   STATUS    RESTARTS   AGE    IP           NODE   NOMINATED NODE   READINESS GATES
pod1     1/1     Running   0          11d    10.64.1.96   node   <none>           1/1
pod2     2/2     Running   0          11d    10.64.1.95   node   <none>           2/2
pod3     2/2     Running   0          175m   10.64.2.64   node   <none>           <none>
pod4     3/3     Running   0          175m   10.64.3.85   node   <none>           <none>
```

Containers

Readiness Gates

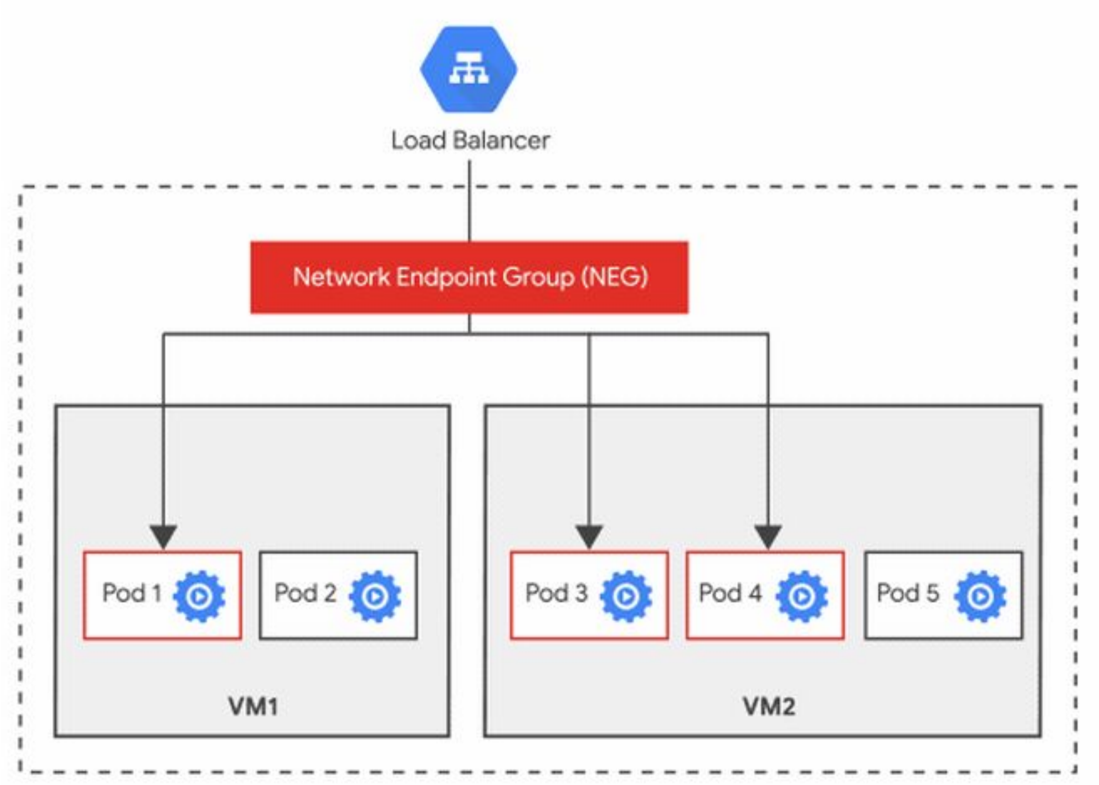
GKE Use Case:
Container Native Load balancing

Container Native Load Balancing



Container Native Load Balancing

- Pods as first class endpoints
- Features like cookie affinity, “Just Work”
- Balances the load without downsides of a second hop

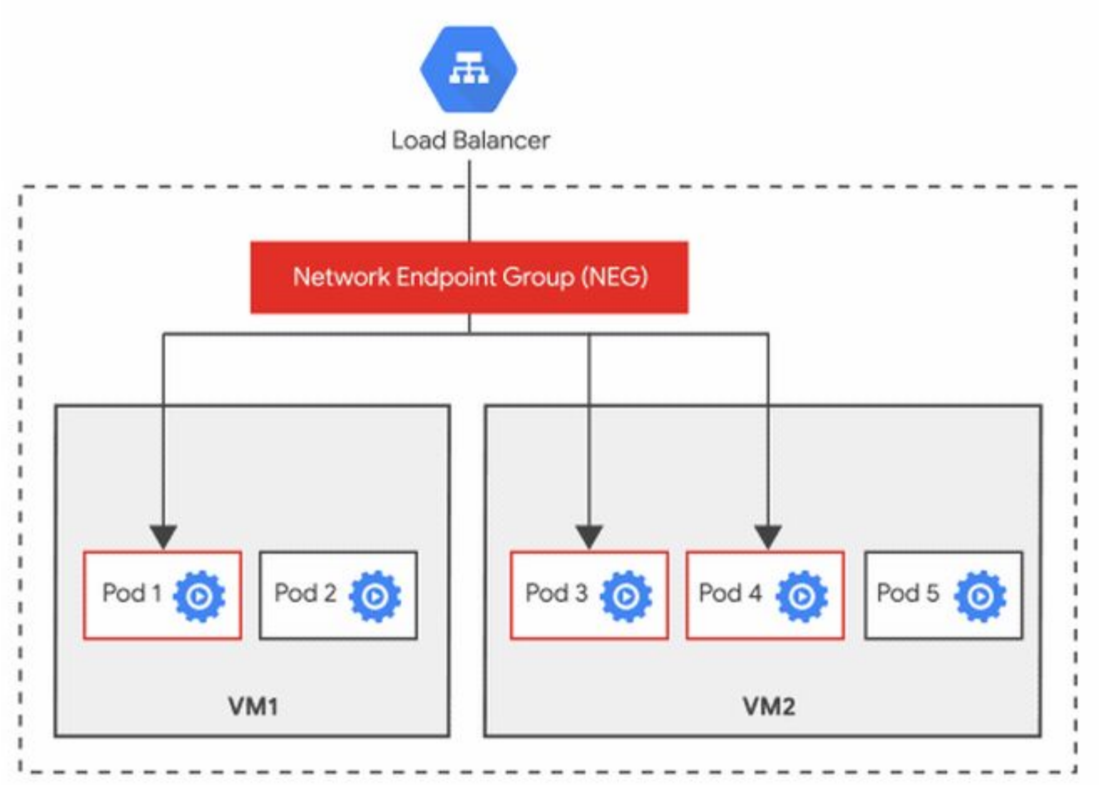


Container Native Load Balancing

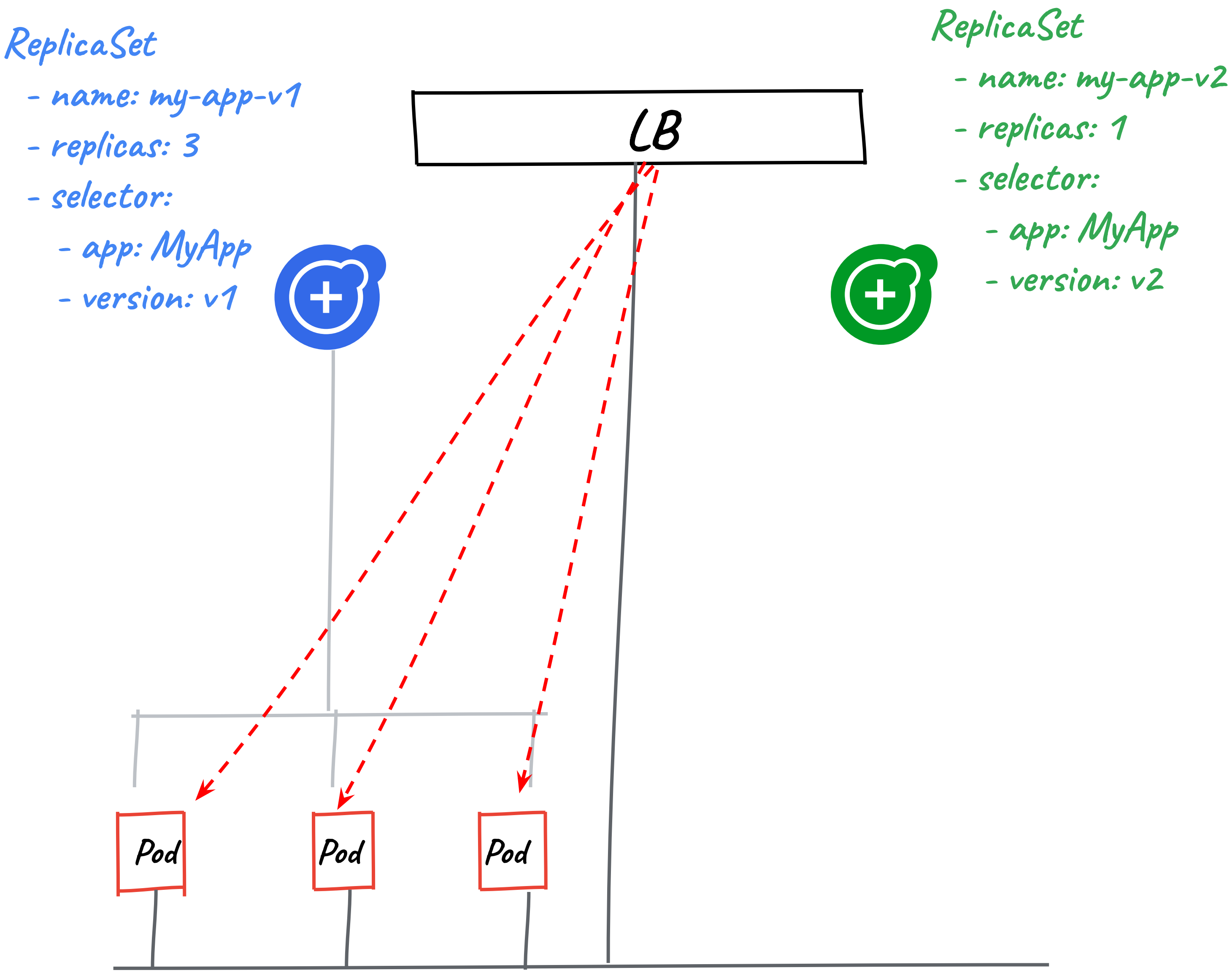
Rolling Update Challenge:

Programming external LBs is slower than iptables

Possible to cause an outage by rolling update going faster than LB

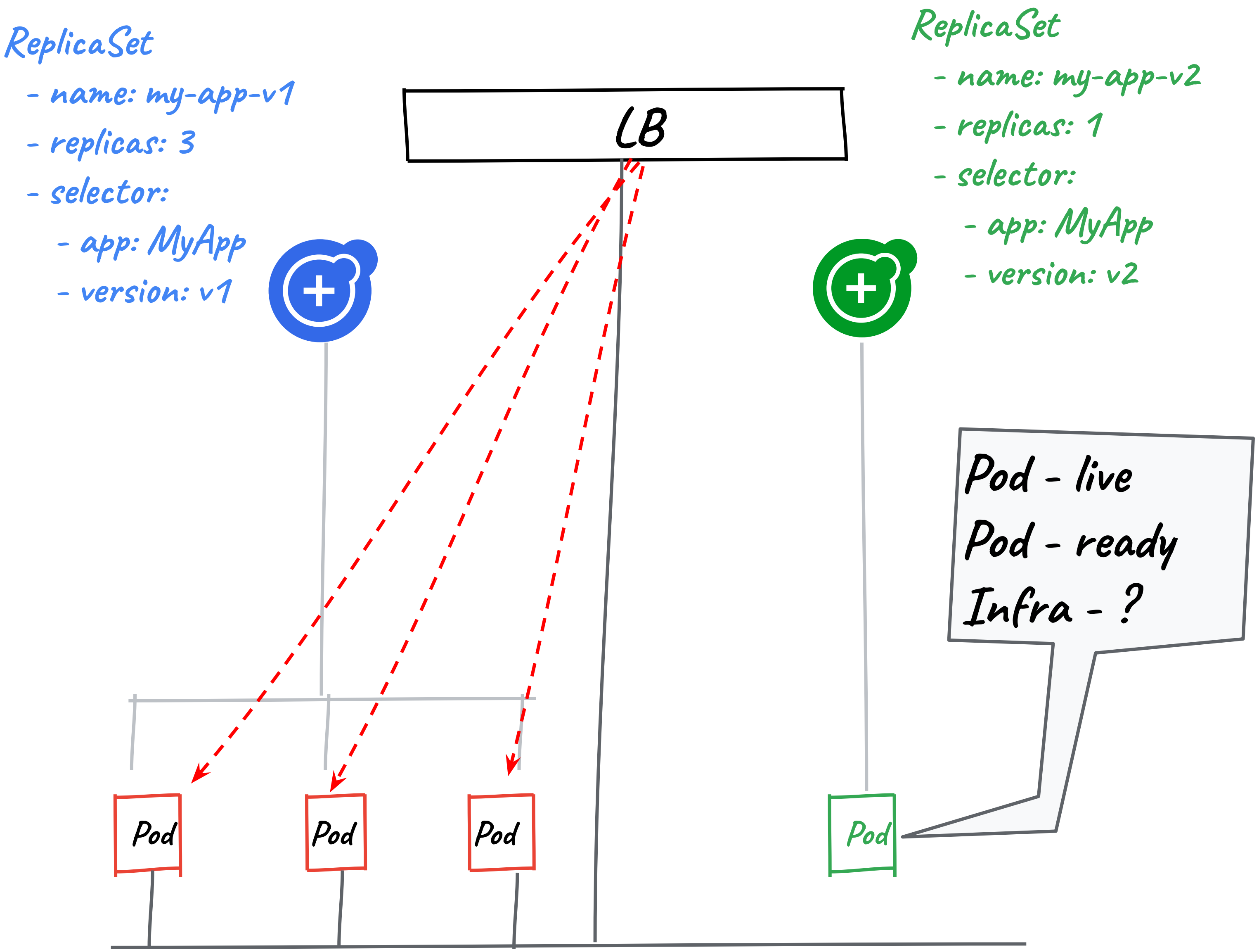


Rolling Update



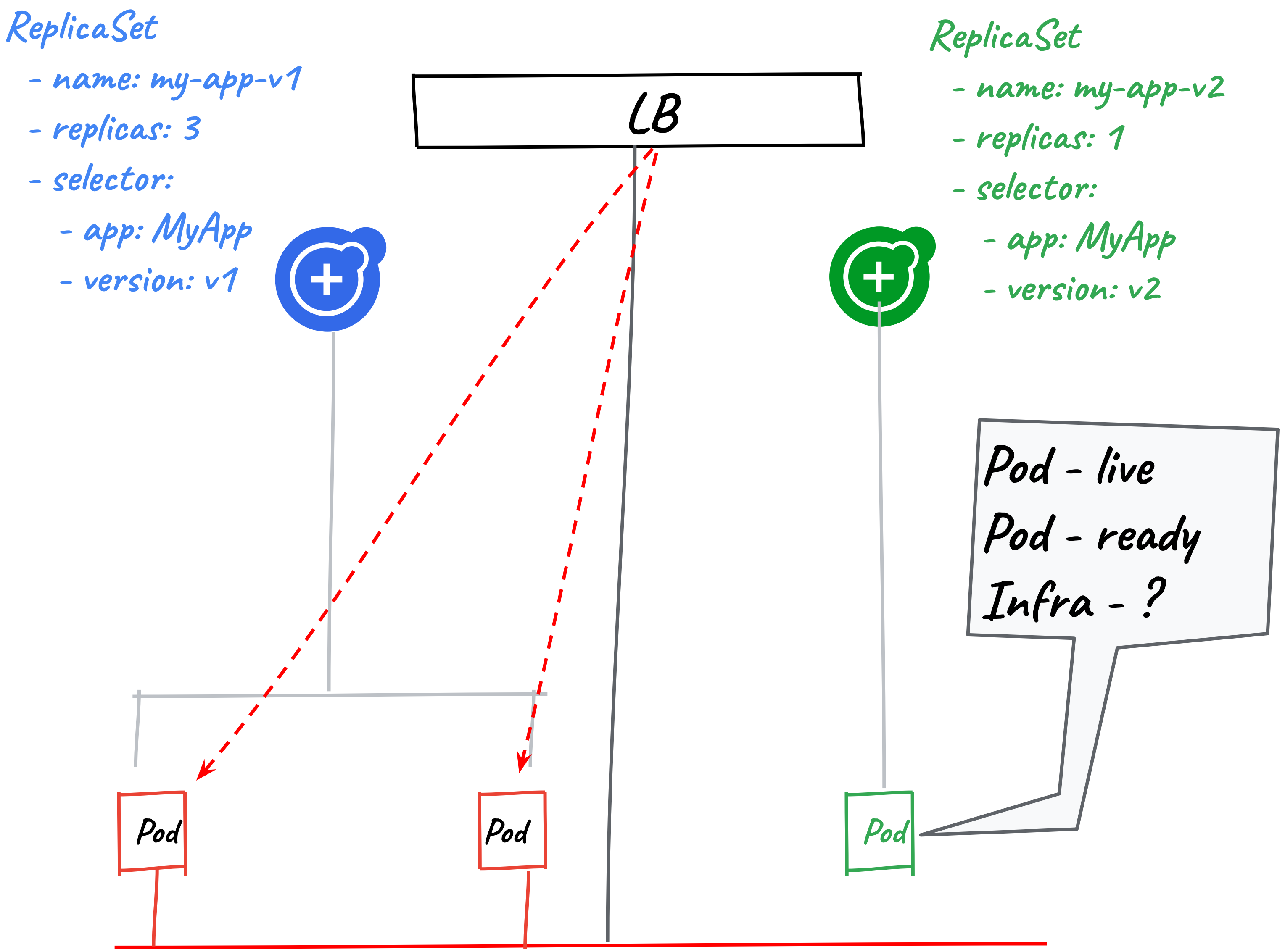
Rolling Update

- Pod Liveness : state of application in pod -alive or not
- Pod Readiness : ready to receive traffic



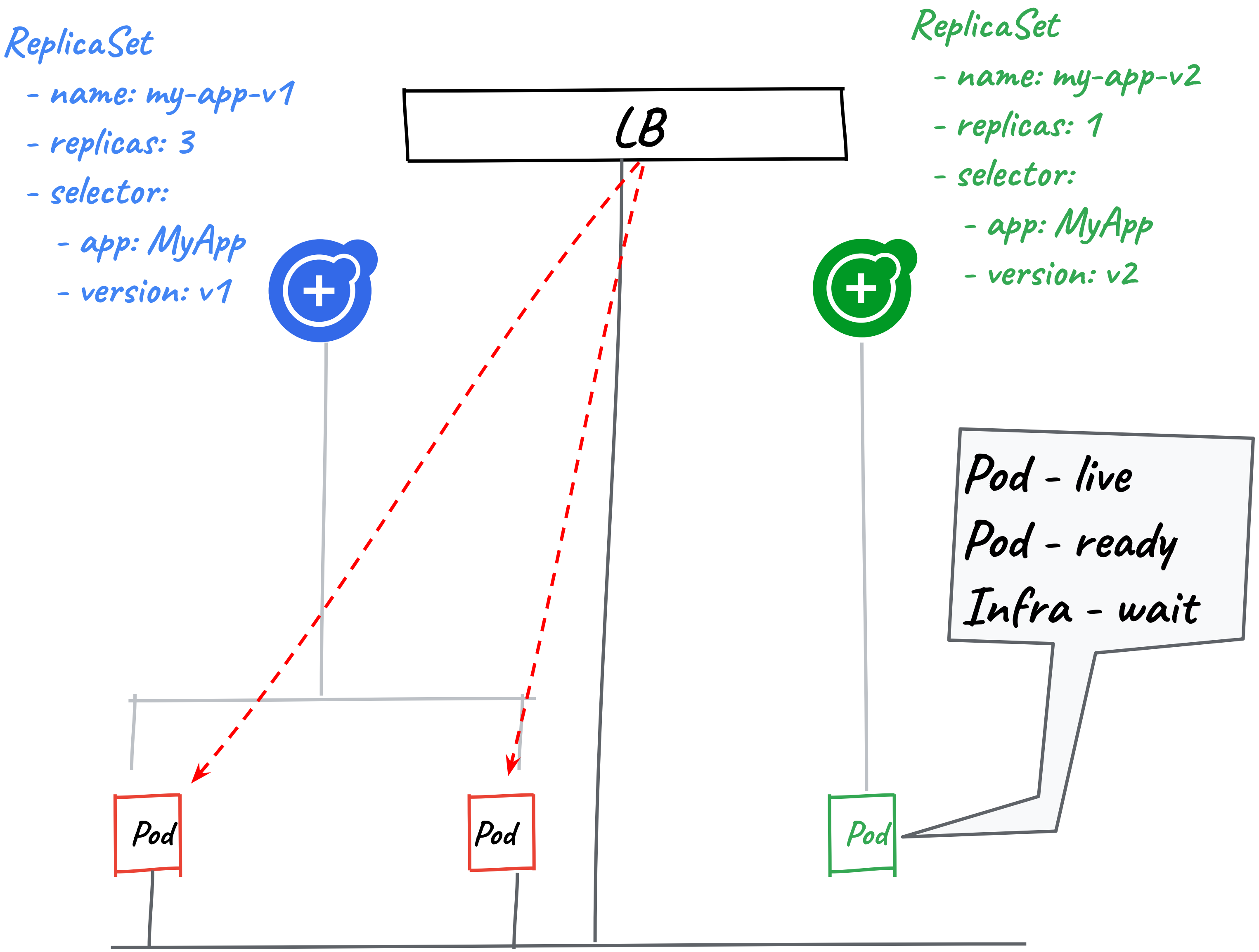
Wait for Infrastructure?

- LB not programmed but Pod reports ready
- Pod from previous replicaset removed.
- Capacity reduced !.



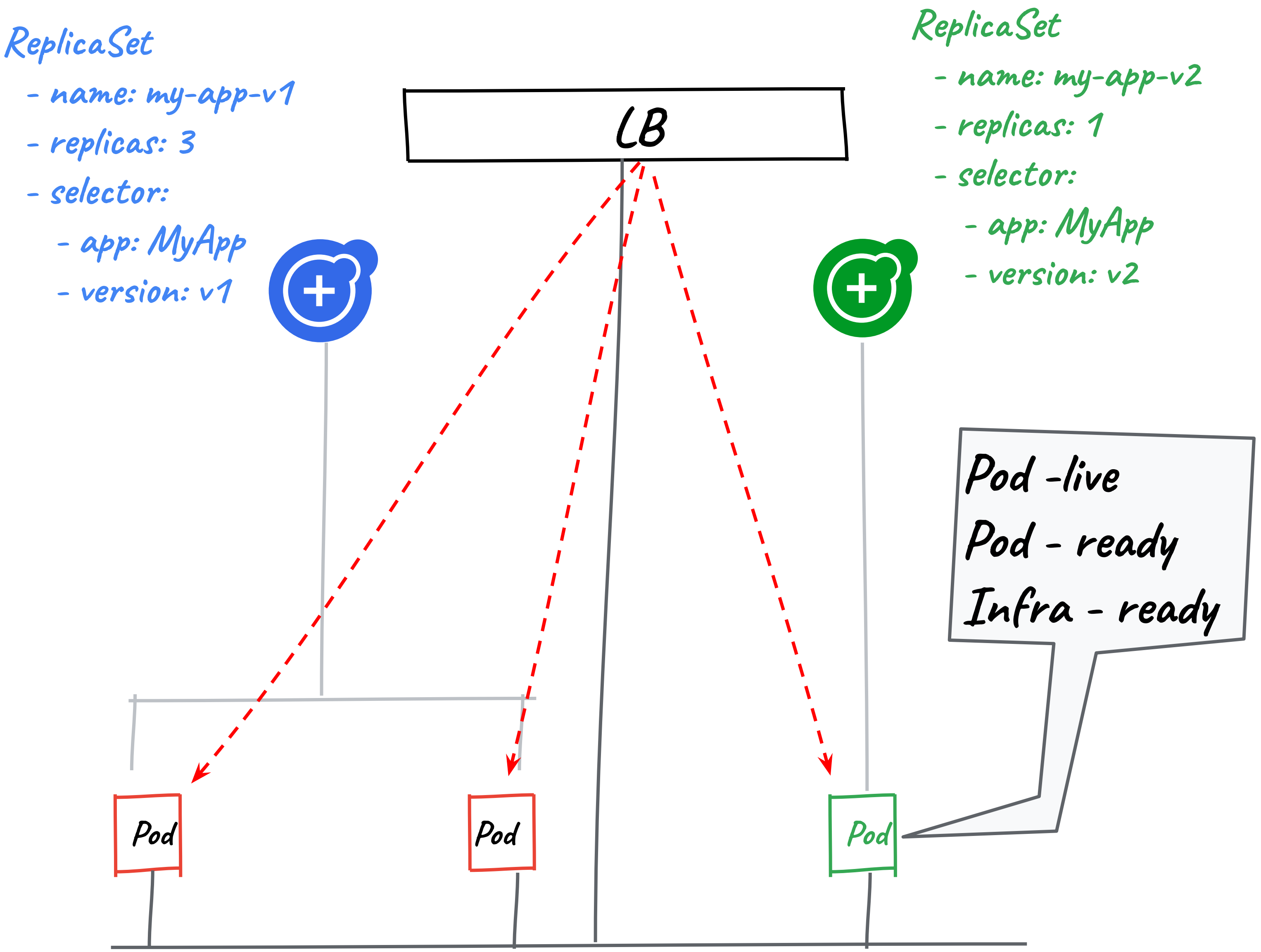
Pod Ready ++

- New state in Pod life cycle to wait - Pod Ready ++



Pod Ready ++

- New state in Pod life cycle to wait - Pod Ready ++

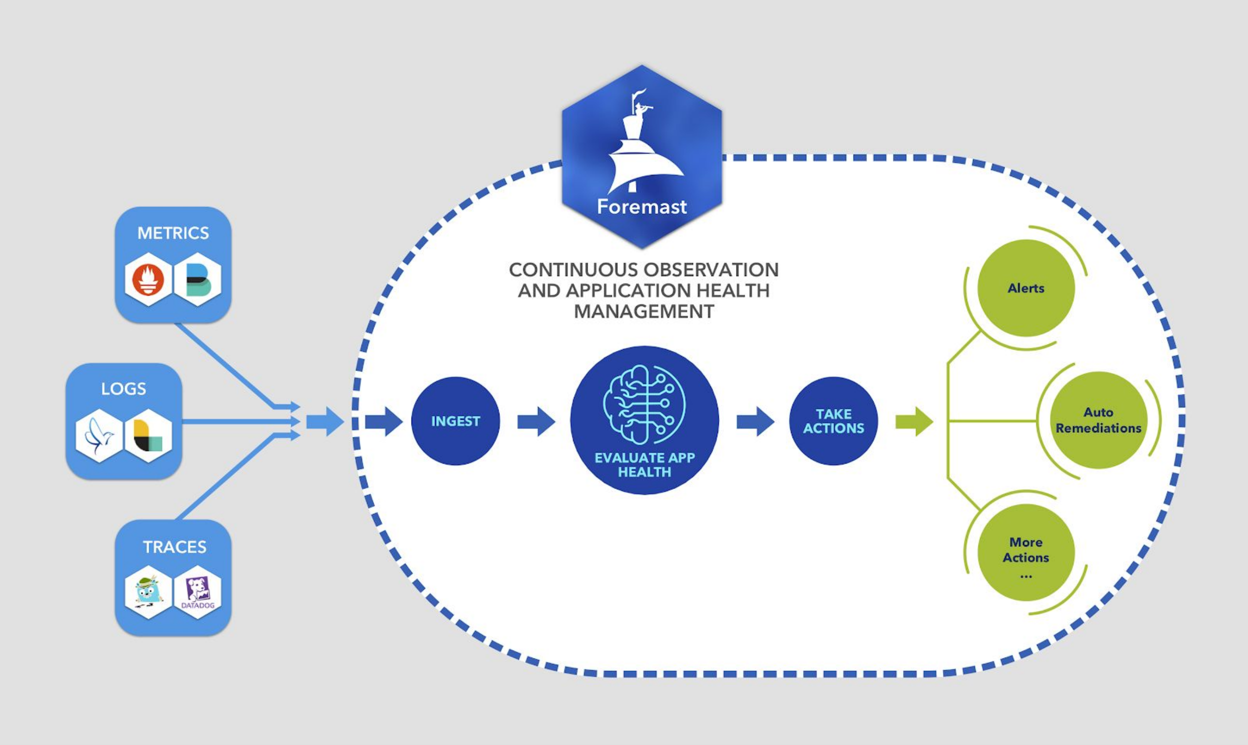


Intuit Use Case: Foremast

Foremast and Pod Readiness Gates

What is Foremast ?

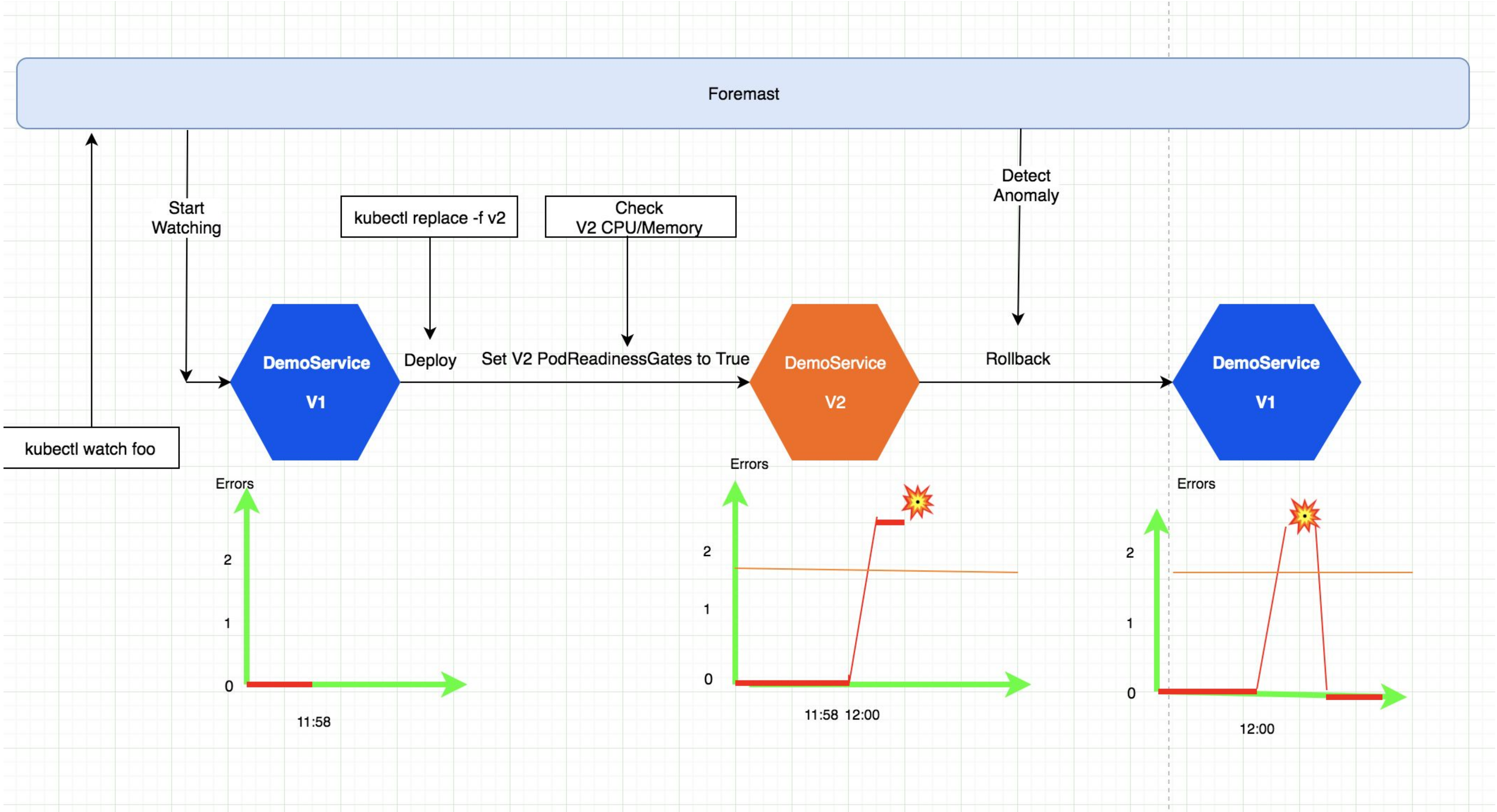
- ➢ Intuit sponsored Open Source Cloud Native health manager platform running on K8s
- ➢ Leverage Metrics, Logs and Traces observability signals
- ➢ Monitor continuously any new deployment rollout strategy like Canary or Blue/Green
- ➢ Use machine learning on the application health signals, detect anomalies and perform remediation



Foremast Pod Readiness Gates feature user cases

- ○ Make sure pod is started and in steady, healthy condition, then set Pod Readiness to true to start to serve traffic
- ○ Reset Pod Readiness Condition to not ready if Pod health check failed.

Foremast Leverage PodReadinessGates Demo



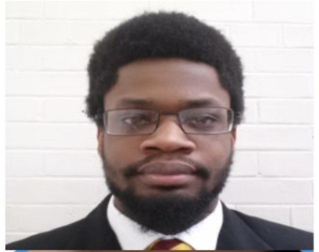
Foremast Team

Foremast Contributors

<- Dawei Ding
Ed Lee ->

Ping Zou ->
<- Sheldon Shao

<- Sen Lin
Dave Masselink



Kian Jones ->
<-Mukulika Kapas

<-Debashis Saha
SrivathsanCanchi>

Reference:
http://foremast.io

GitRepo:
http://github.com/intuit/foremast
http://github.vom/intuit/foremast-brain

Q & A

Appendix -- Foremast

Appendix -- Foremast

Backup Slides

Agenda

1. PodReadinessGate API Intro
   a. Pod Ready?
   b. Container Ready
   c. Pod Life Cycle
   d. Readiness Gate
   e. Custom conditions
2. GCP use case
   a. Rolling Update
   b. disconnect between K8s network primitives and workloads
   c.
3. Foremast Use case
   a. Foremast detected deployment change != pod/container(application) ready and able to serve traffic
   b. Foremast detected deployment change and make sure containers ready then trigger monitoring as service request to monitor if there is any anomaly for new version,