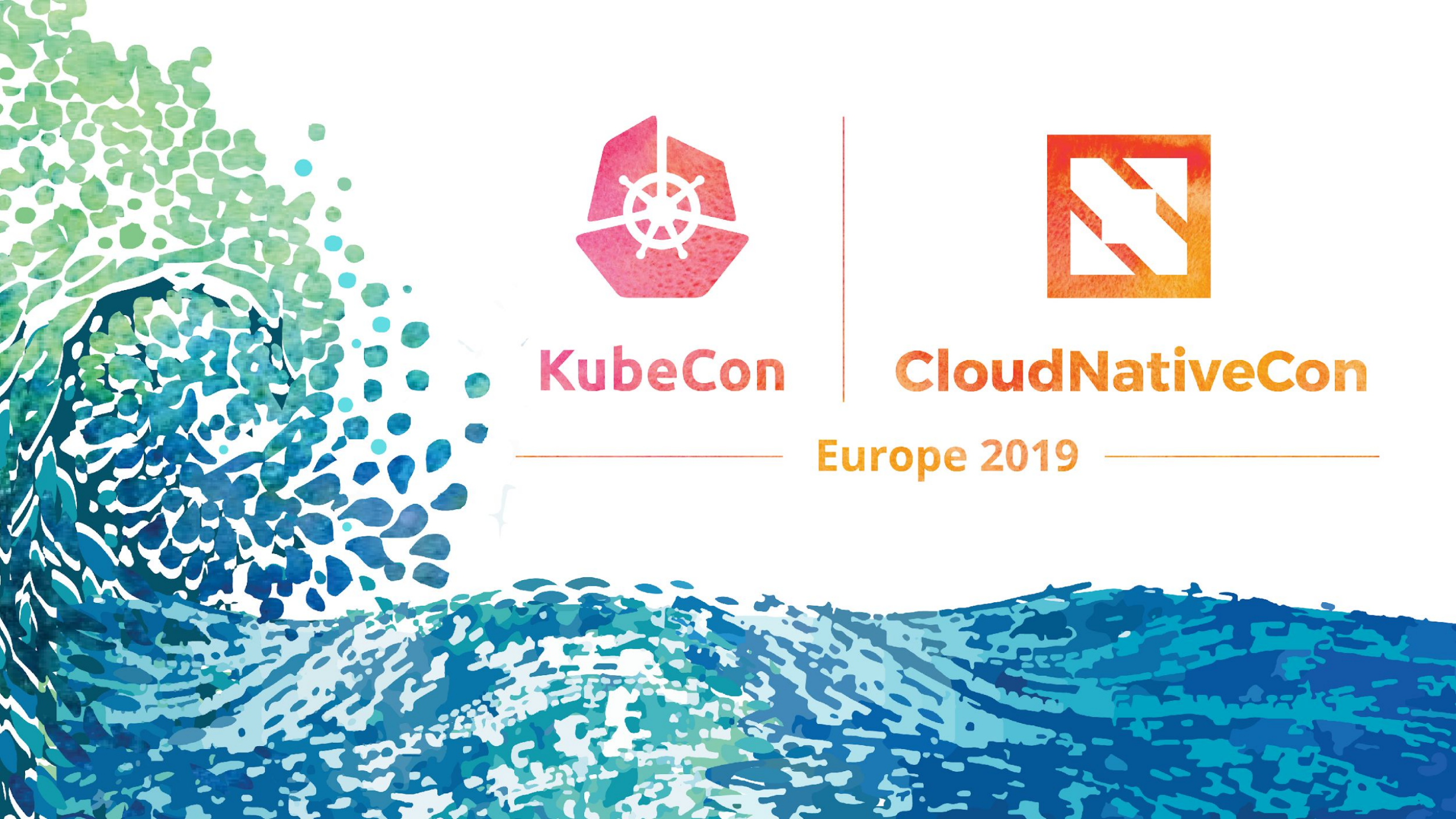

KubeCon
CloudNativeCon
Europe 2019

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://xxxxxxxxxxxxxxxxxxxx
    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

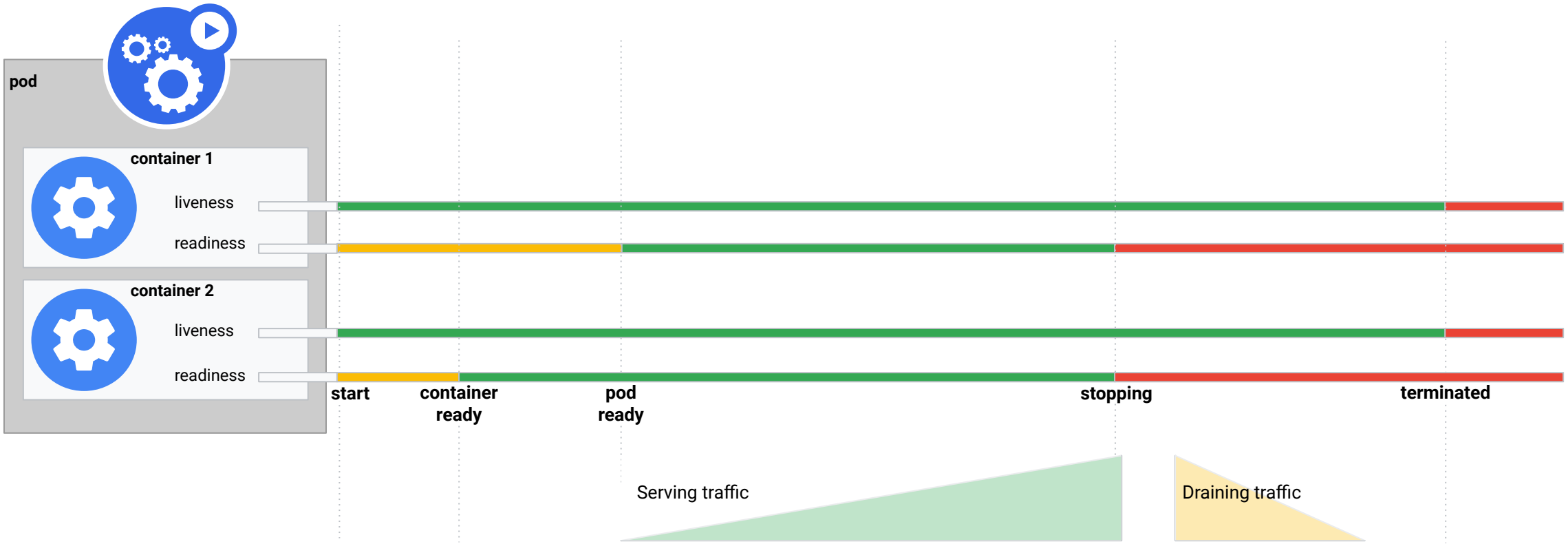

```
kind: Pod
apiVersion: v1
metadata:
  ...
spec:
  ...
status:
  conditions
    - type: PodScheduled
      status: "True"
      lastTransitionTime: "2019
    - type: Initialized
      status: "True"
      lastTransitionTime:
    - type: Ready
      status: "True"
      lastTransitionTime: "2019-05-21T00:01:00Z"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

=

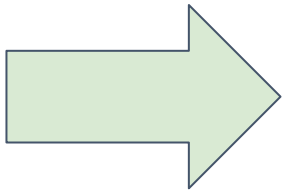
?

Pod Readiness Consumer: Workload


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

Deployment Rolling Update

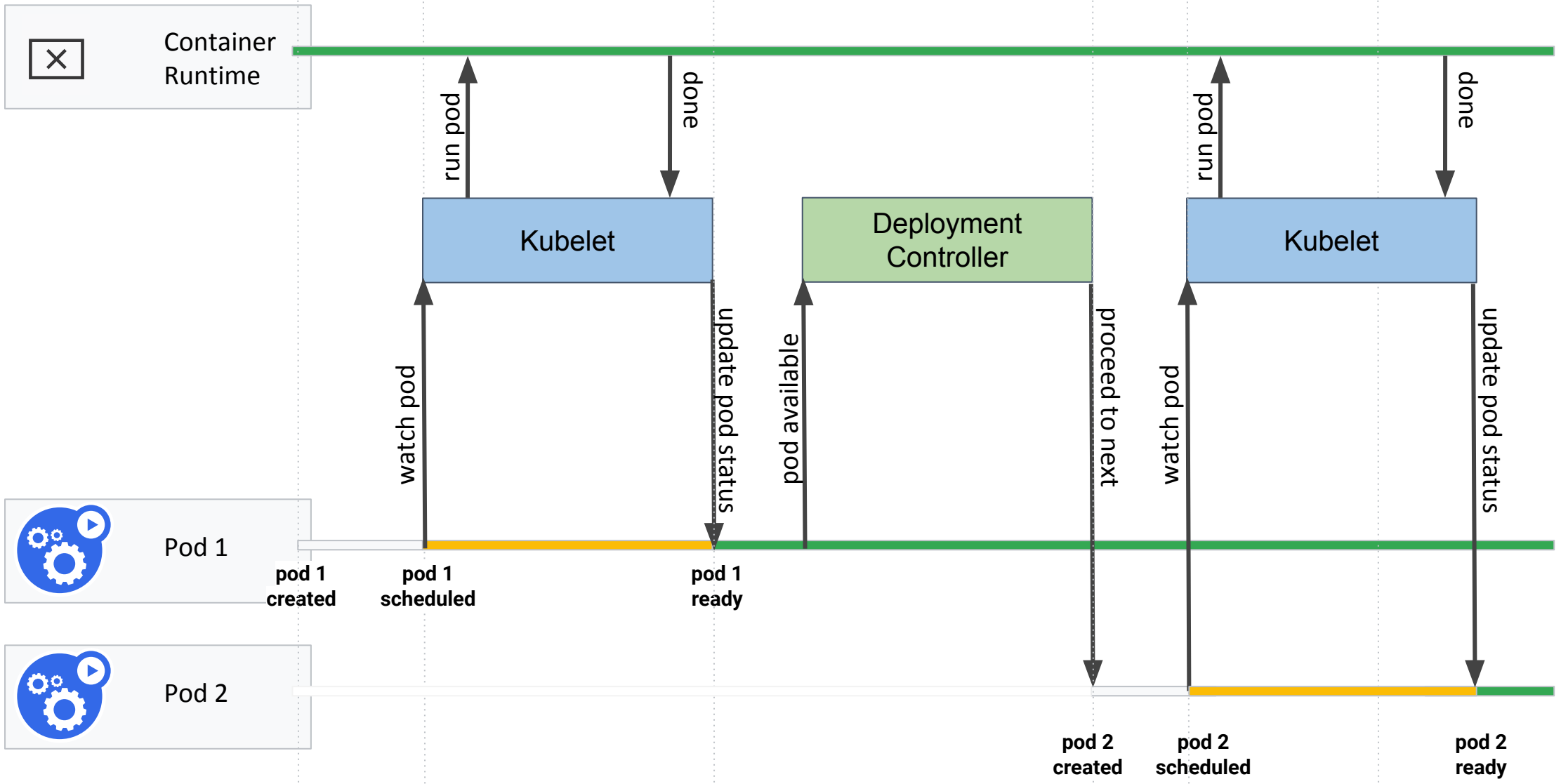

```
kind: Deployment
metadata:
  generation: 2
  ...
spec:
  replicas: 10
  strategy:
    rollingUpdate:
      maxSurge: 1
      maxUnavailable: 1
    type: RollingUpdate
  ...
```



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

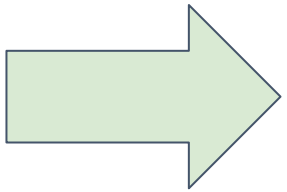
```
kind: ReplicaSet
metadata:
  generation: 2
  ...
spec:
  replicas: 5
  ...
```

Deployment Rolling Update



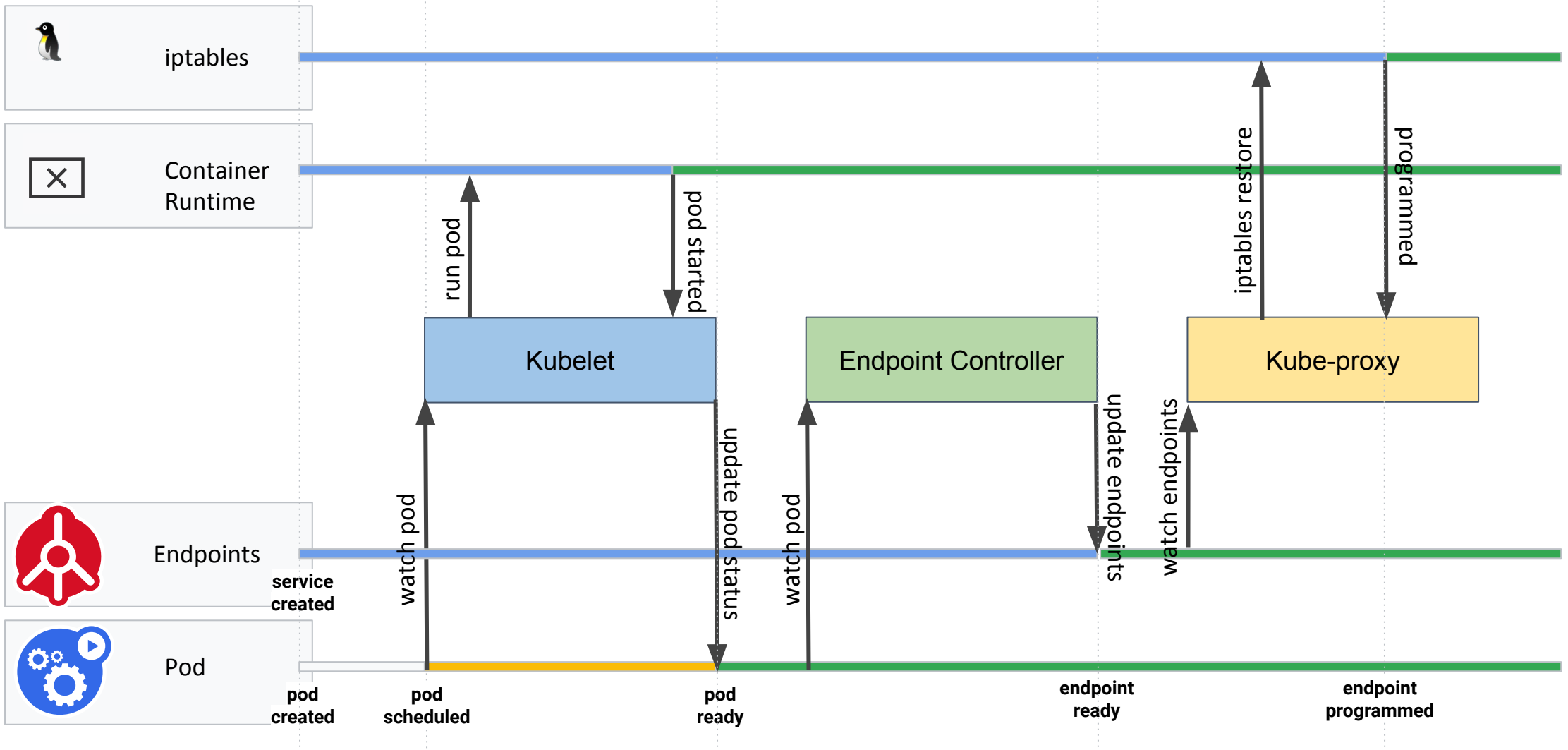
Pod Readiness Consumer: Service


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

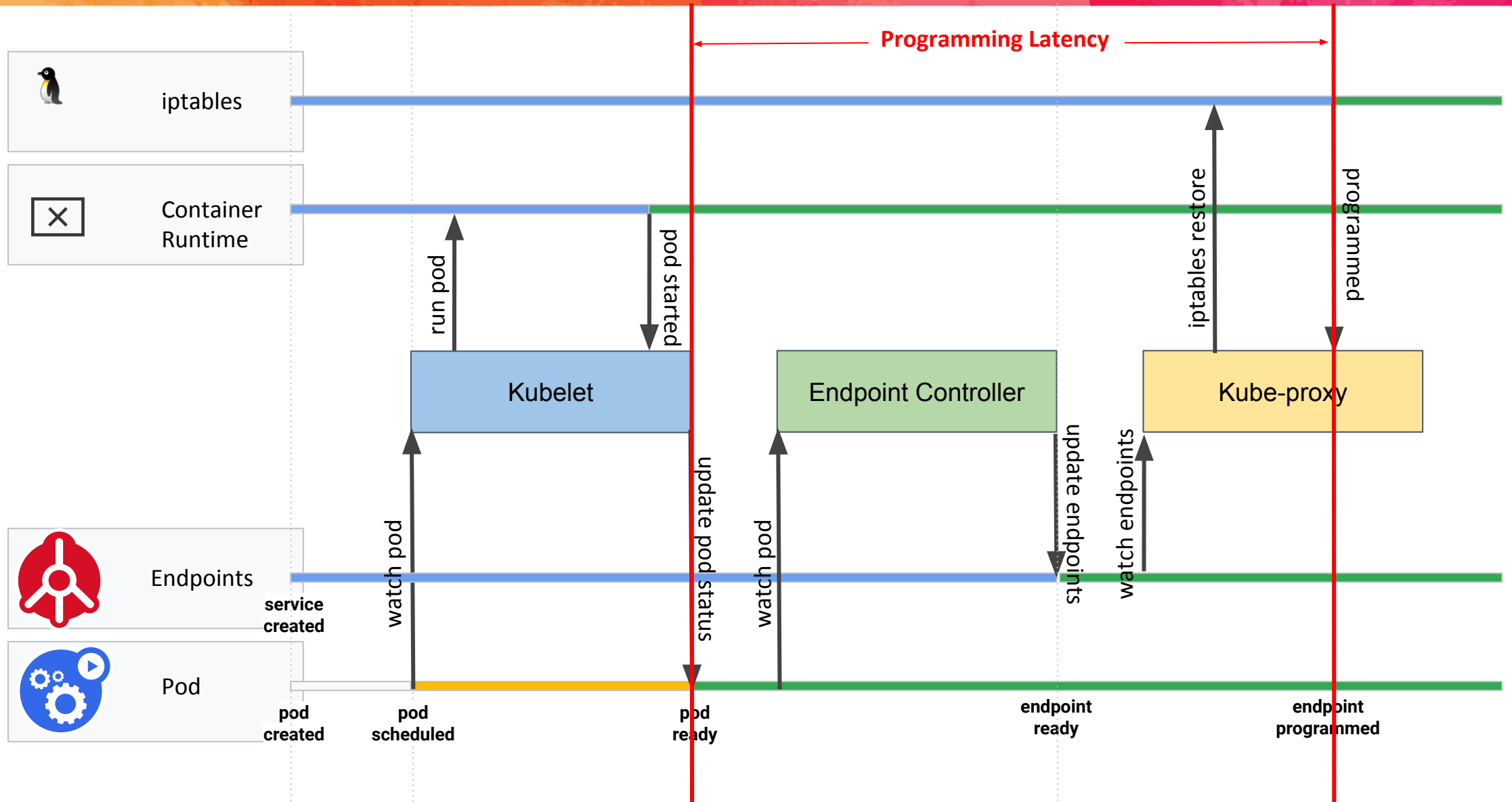


```
kind: Endpoints
metadata:
  ...
subsets:
- addresses:
  - ip: ${Pod IP}
    nodeName: ${Node Name}
    targetRef: ${Pod}
  ...
```

Pod Readiness Consumer: Service



Pod Readiness Consumer: Service



Rendezvous


ReplicaSet
Service
Deployment
Pod
Ingress
StatefulSet
NetworkPolicy

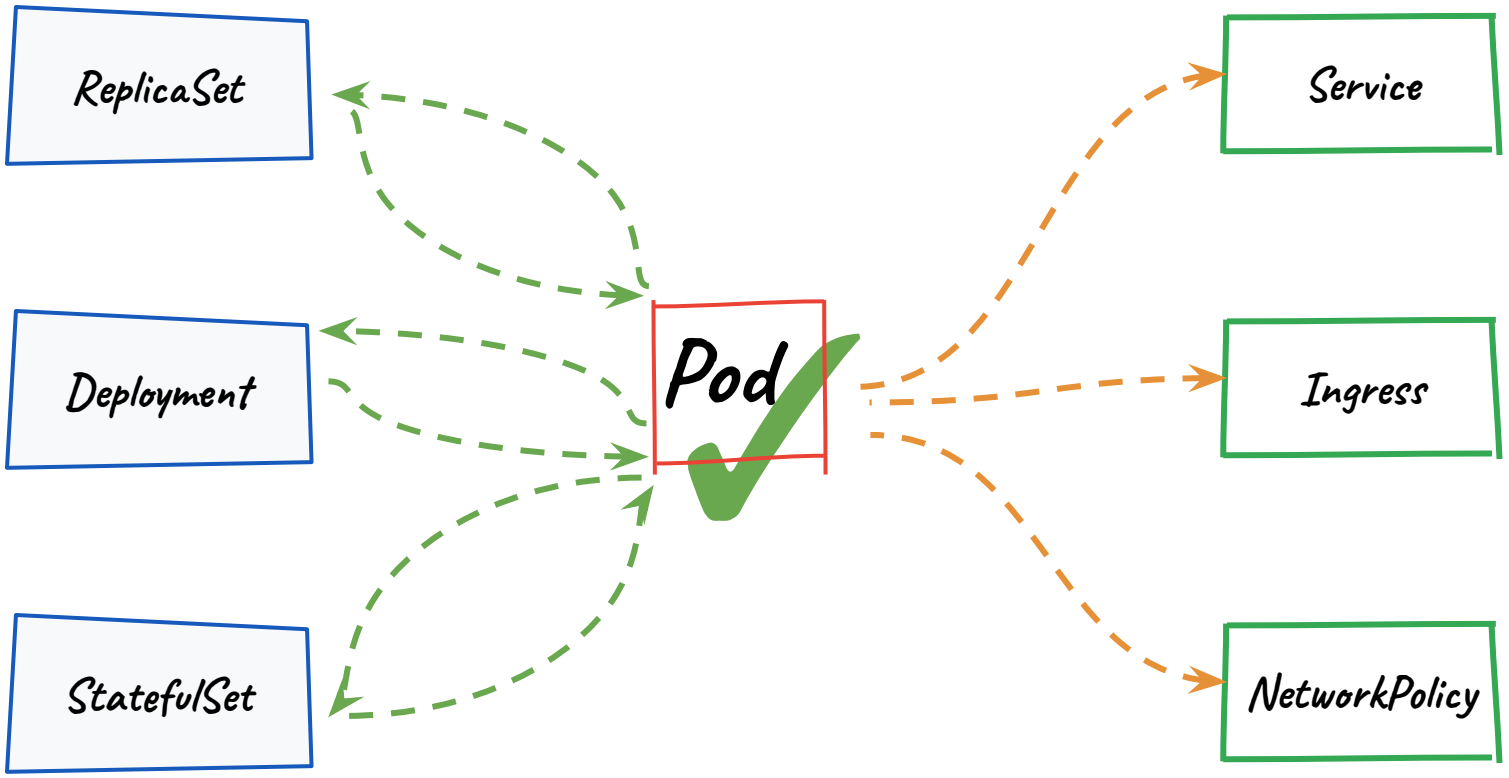
Workload vs. Network Abstractions

KubeCon | CloudNativeCon

Europe 2019

Do they work actually together?

Workloads vs. Network Abstractions



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?

Constraints

Backward Compatibility

Backward Compatibility

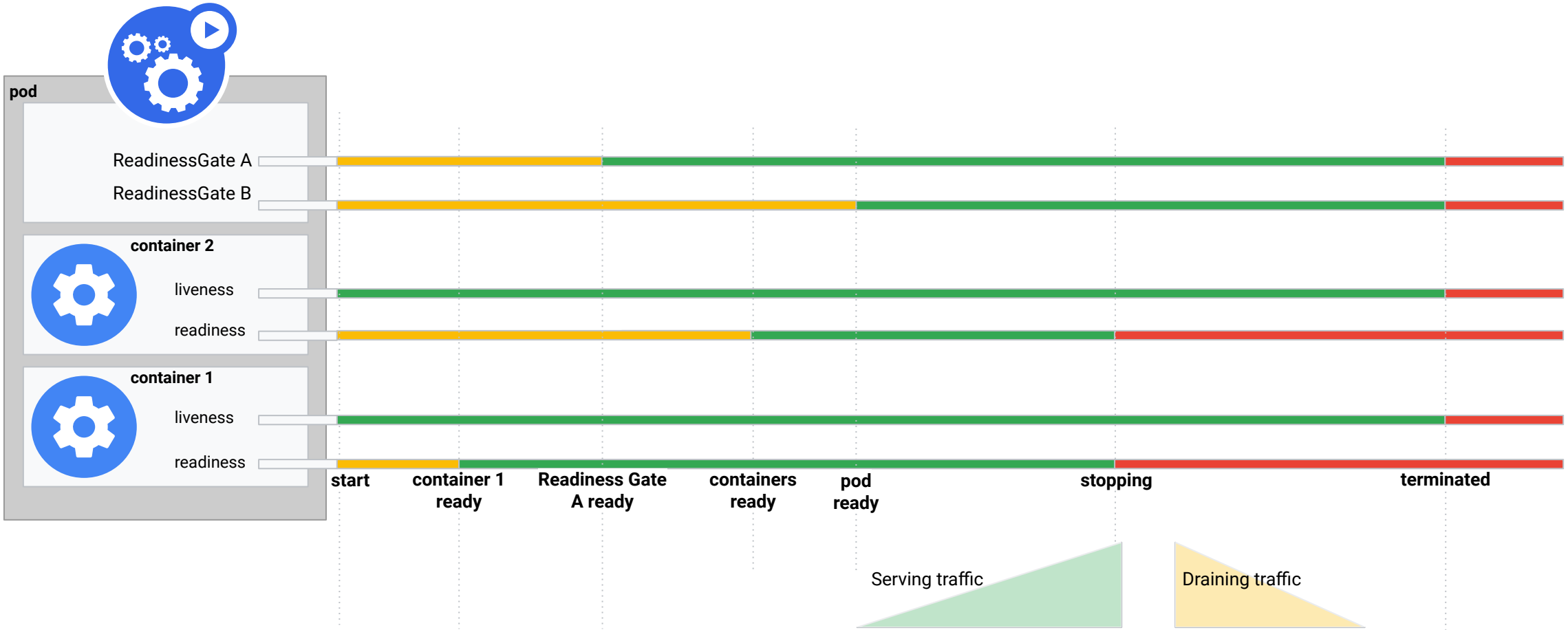
Backward Compatibility

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
...
```

Seamless UX

ReadinessGate Injection?

webhook!

ReadinessGate Condition Update?

PATCH pod status

in custom controller!

Kubectl

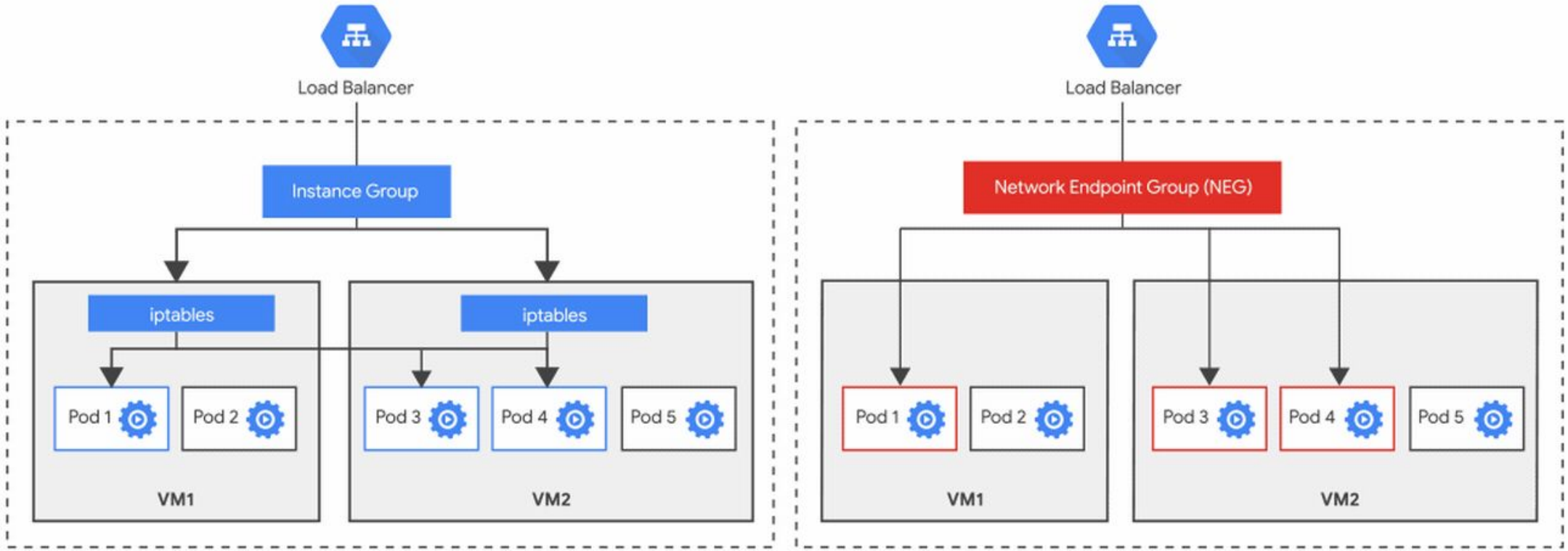

```
$ kubectl get pod -o wide
NAME      READY   STATUS    RESTARTS   AGE    IP             NODE    NOMINATED NODE   READINESS GATES
pod1      1/1     Running   0          11d    10.64.1.96     node    <none>           1/1
pod2      2/2     Running   0          11d    10.64.1.95     node    <none>   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

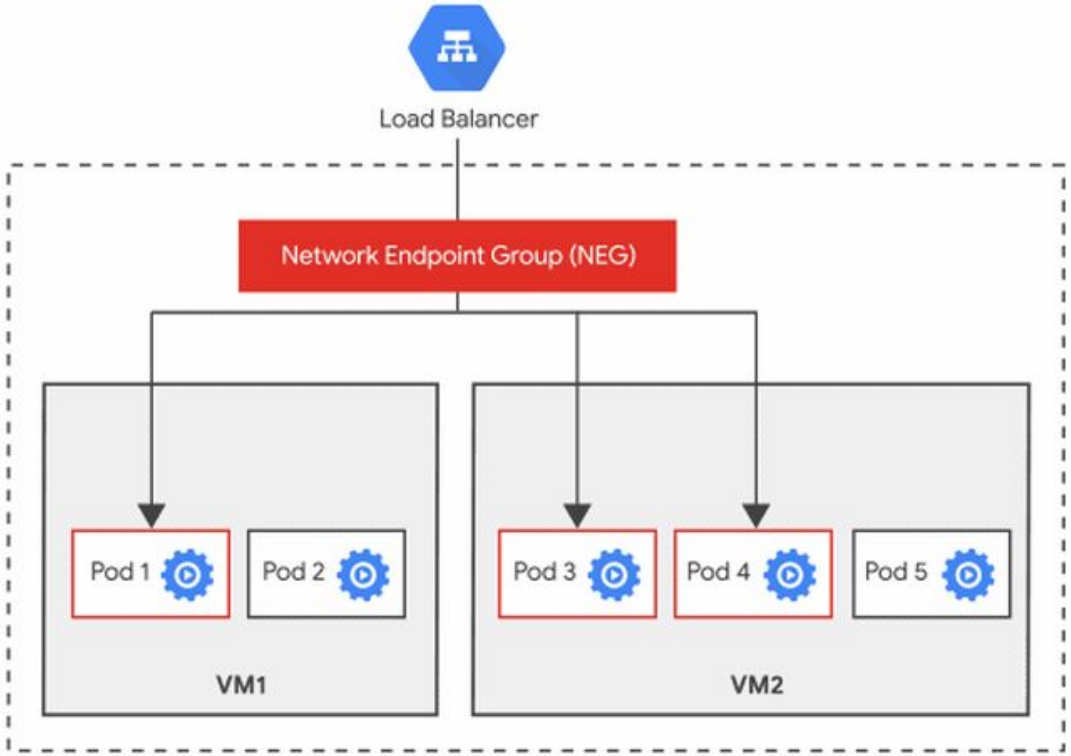
GKE Use Case:
Container Native Load balancing

Container Native Load Balancing



Container Native Load Balancing

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

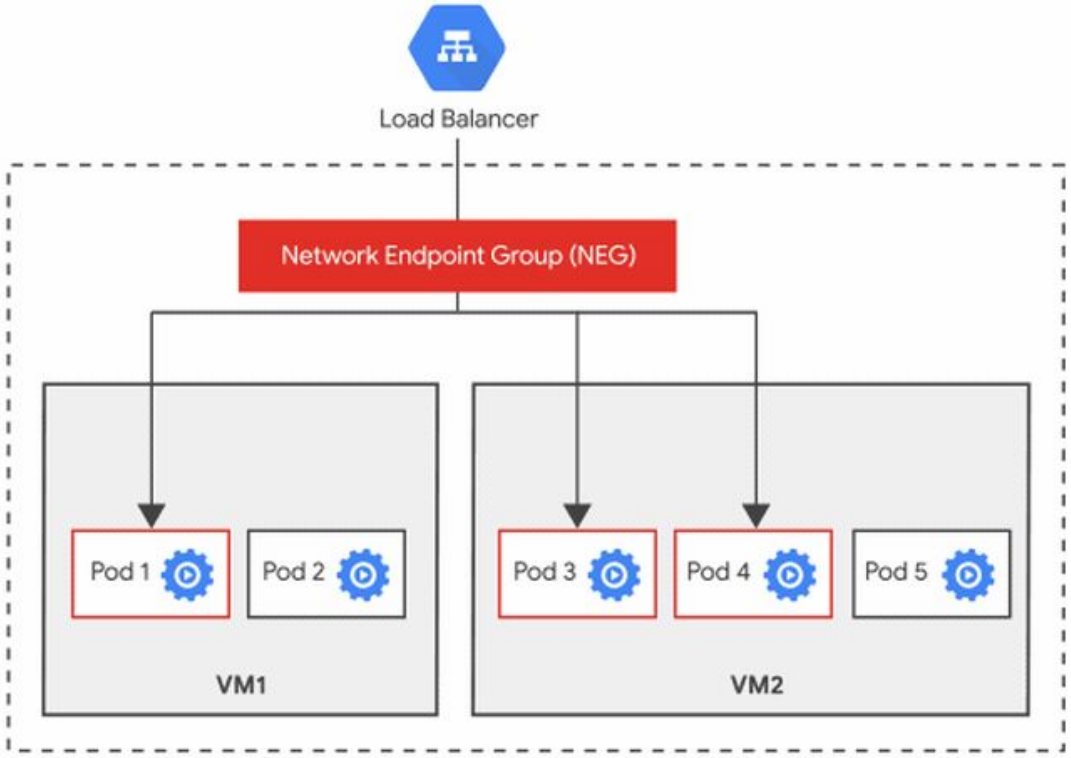


Container Native Load Balancing

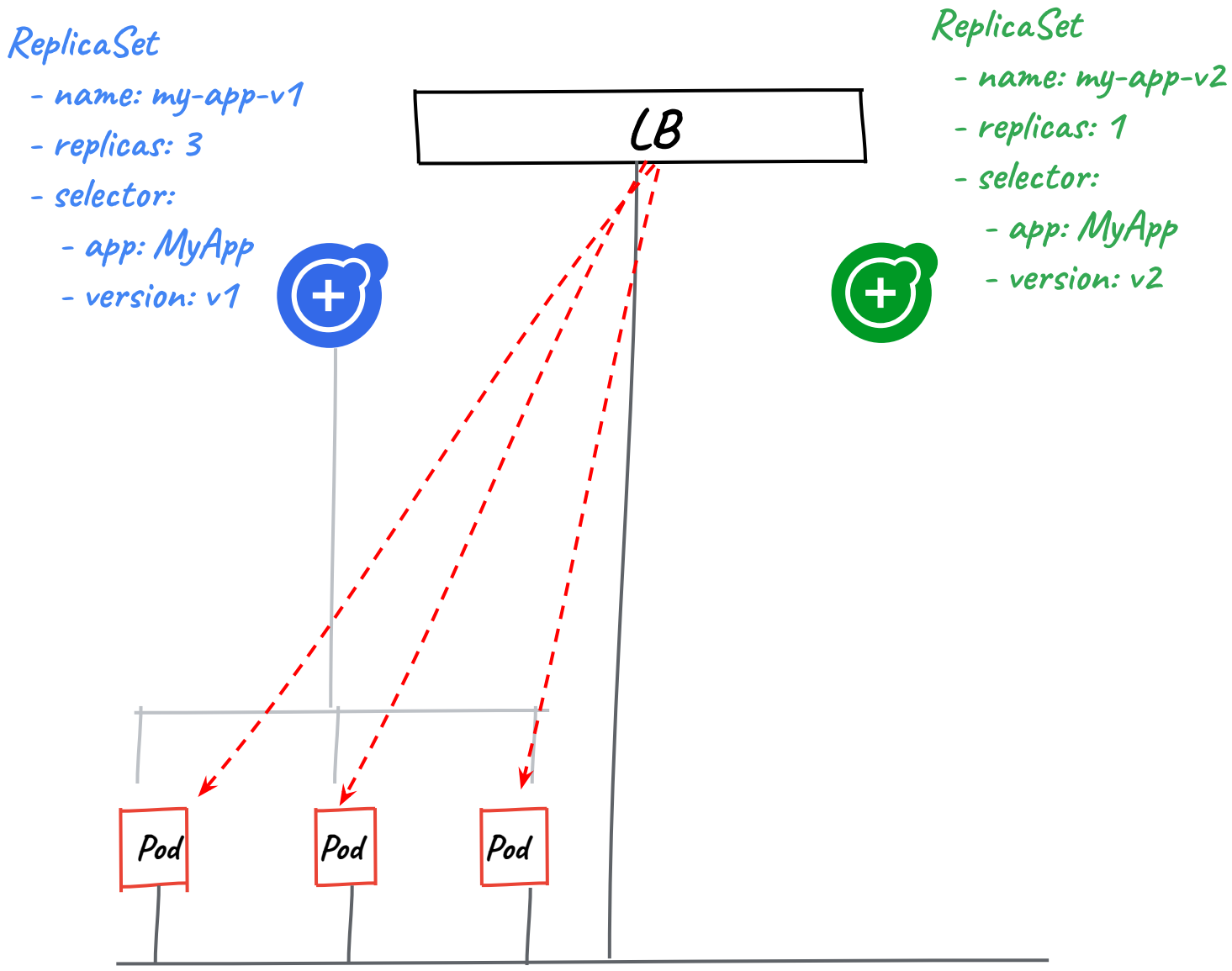
Rolling Update Challenge:

Programming external LBs is slower than iptables

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

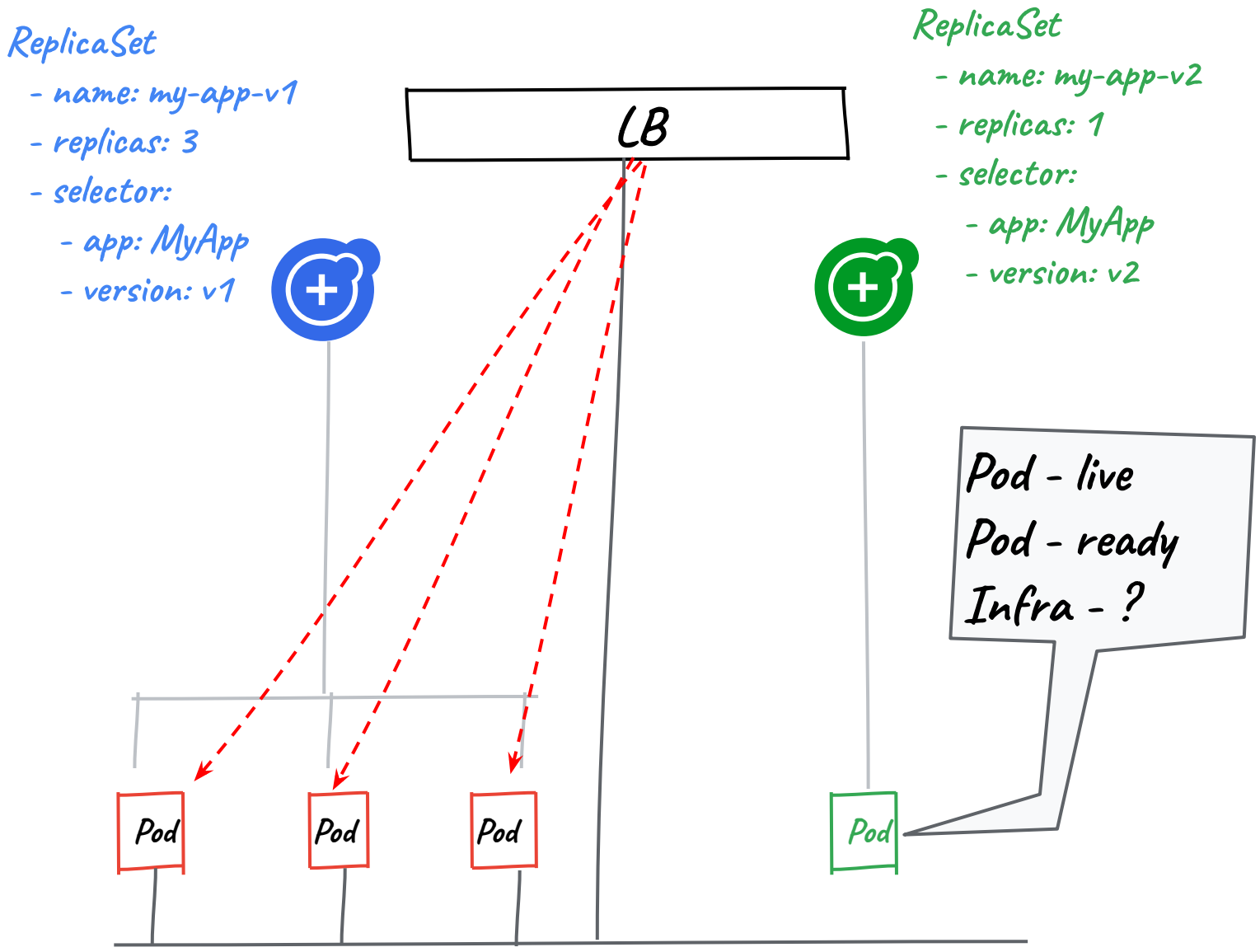


Rolling Update



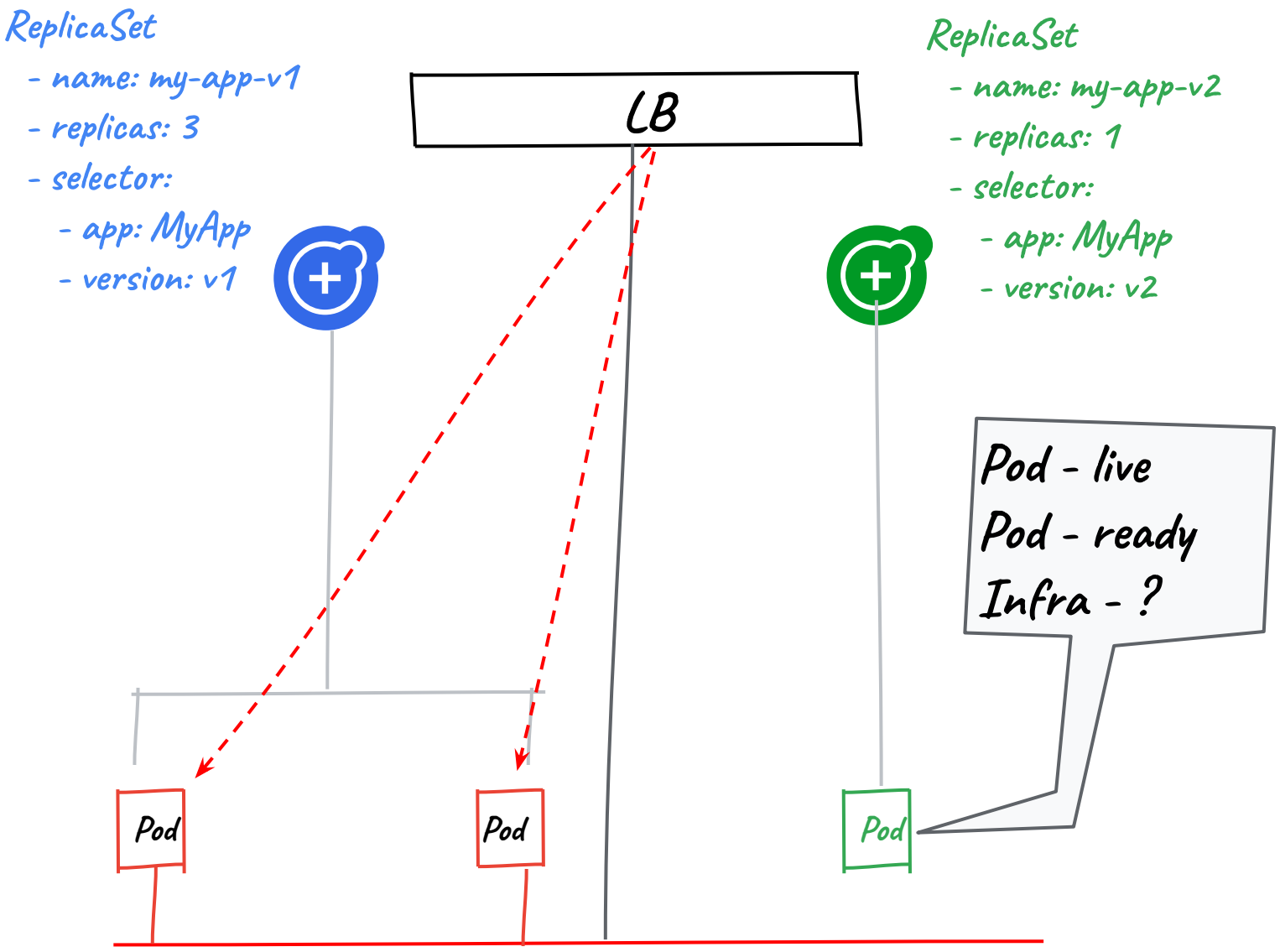
Rolling Update

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



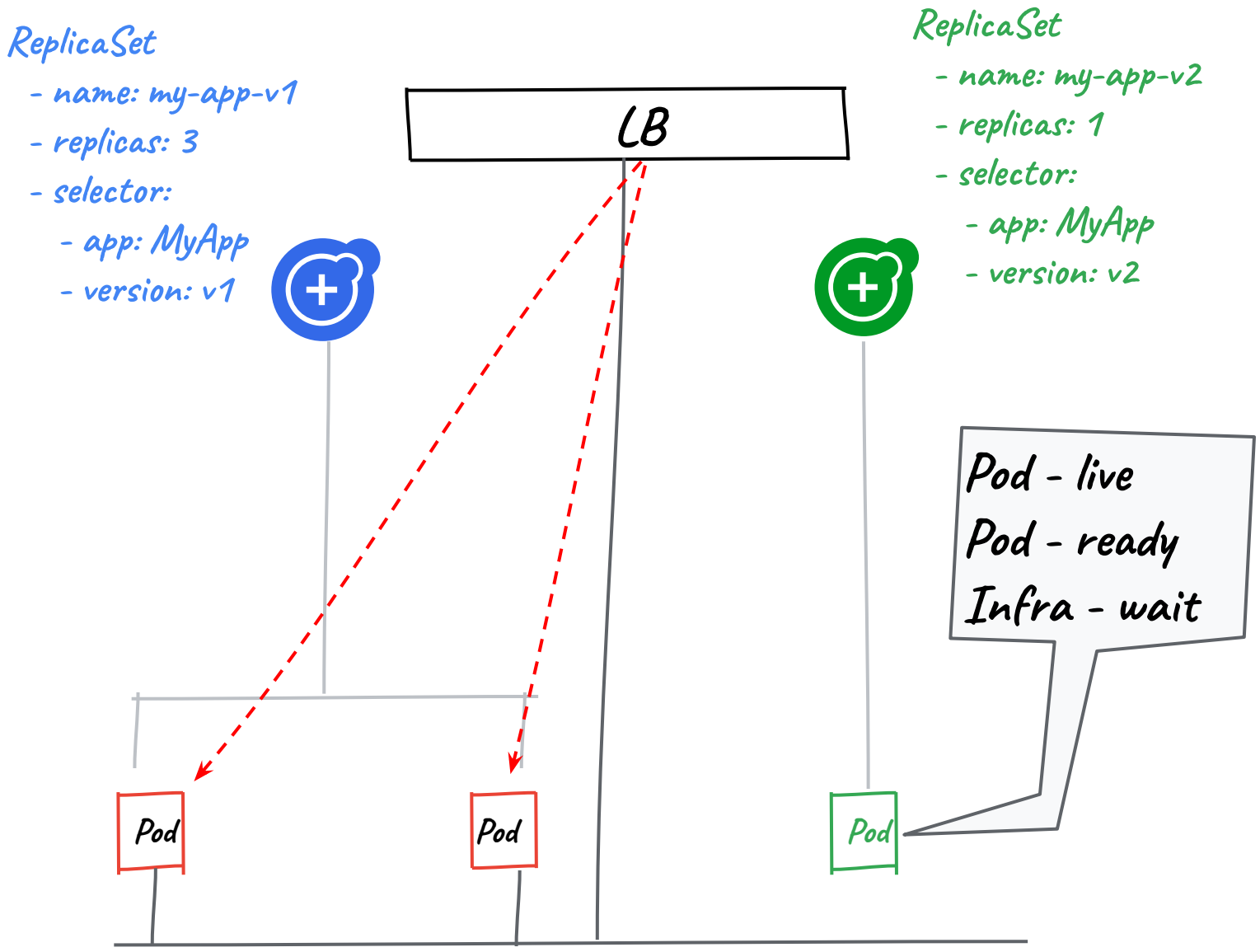
Wait for Infrastructure?

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



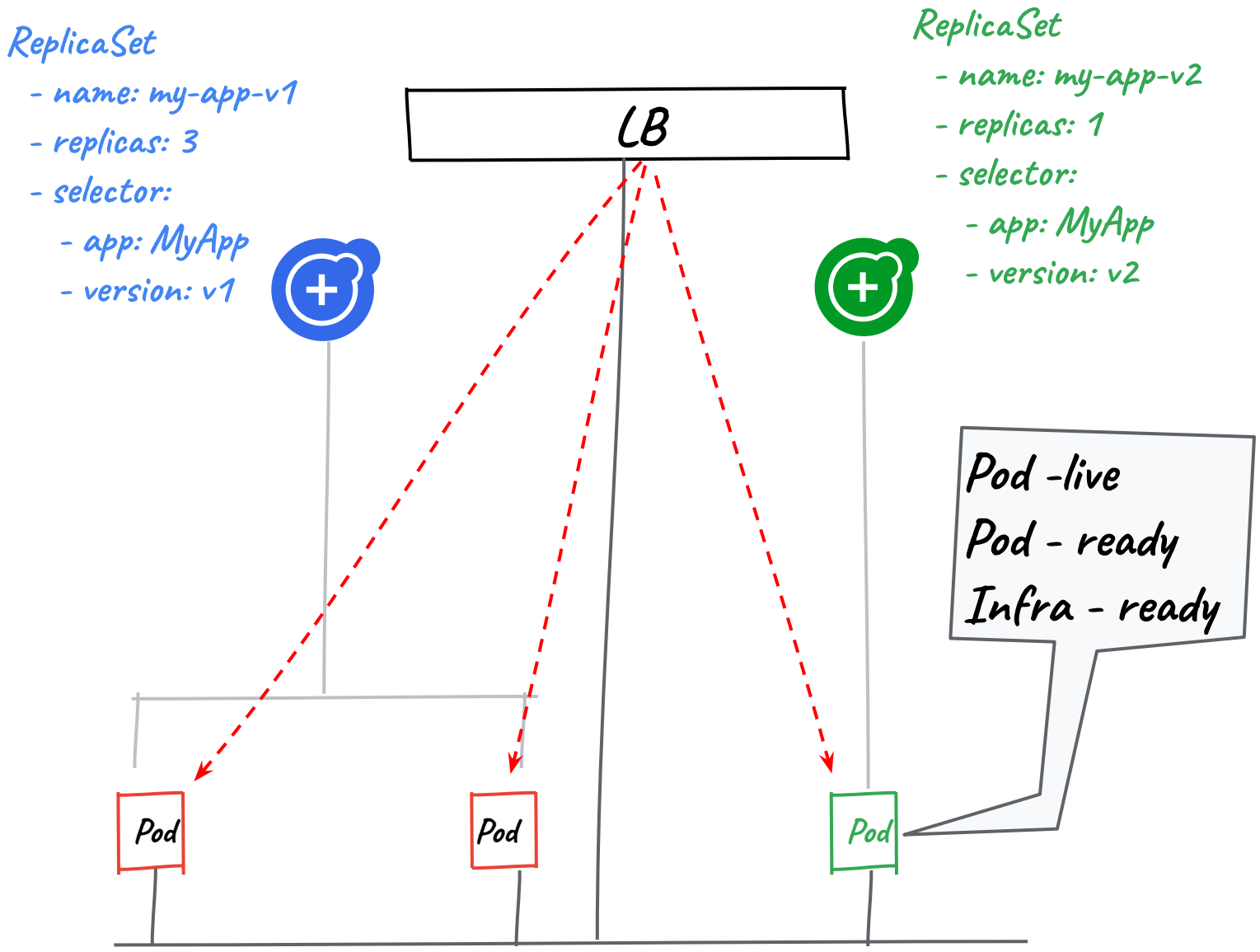
Pod Ready ++

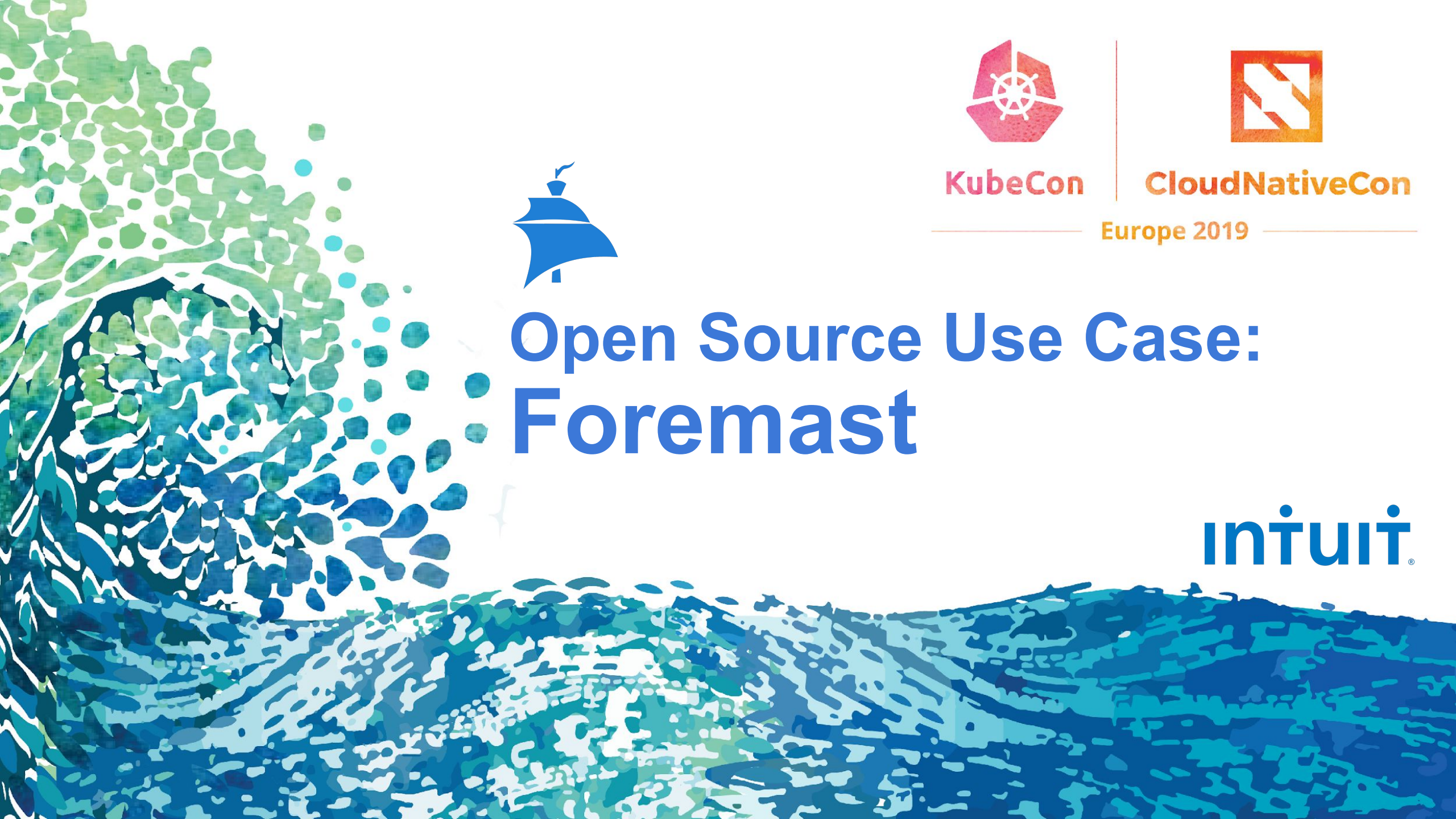
- New state in Pod life cycle to wait - Pod Ready ++



Pod Ready ++

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



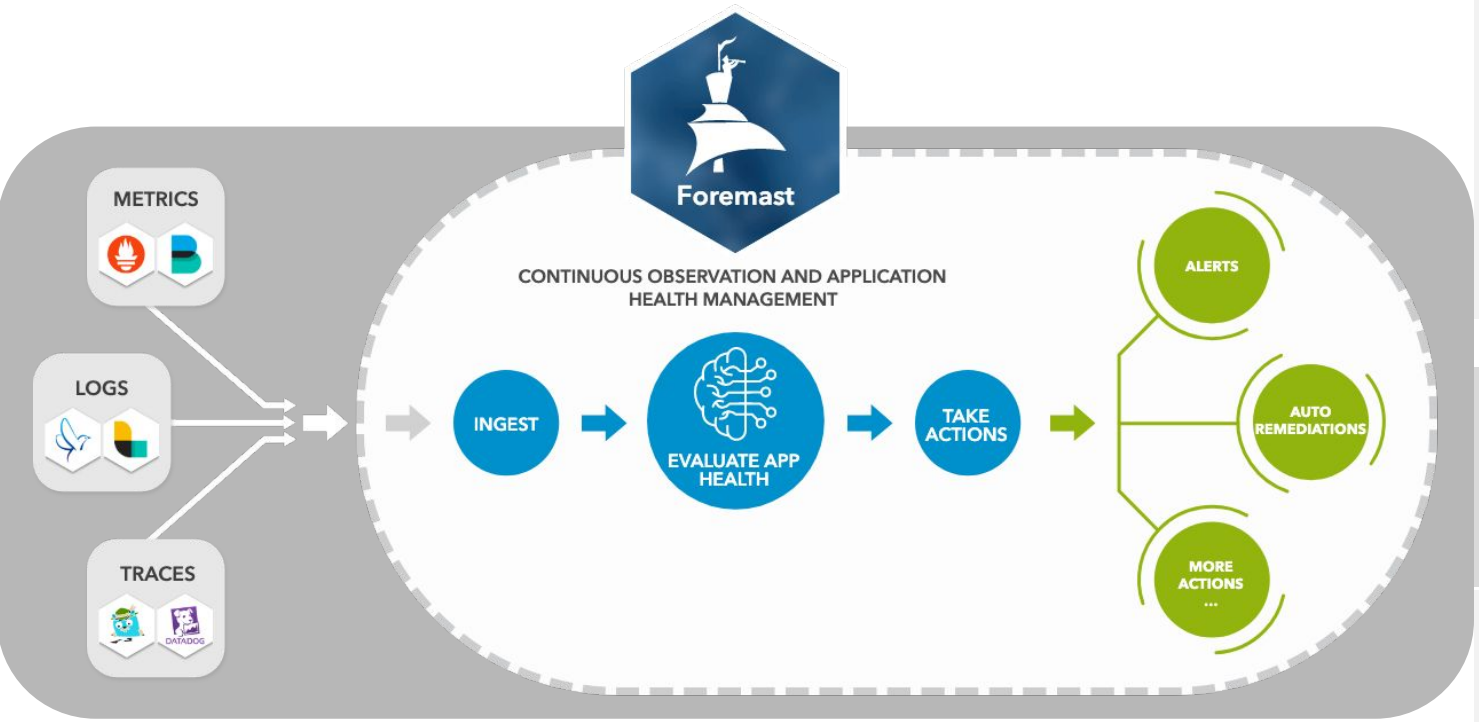


KubeCon | CloudNativeCon Europe 2019

Open Source Use Case: Foremast

INTUIT

Foremast



What is Foremast?

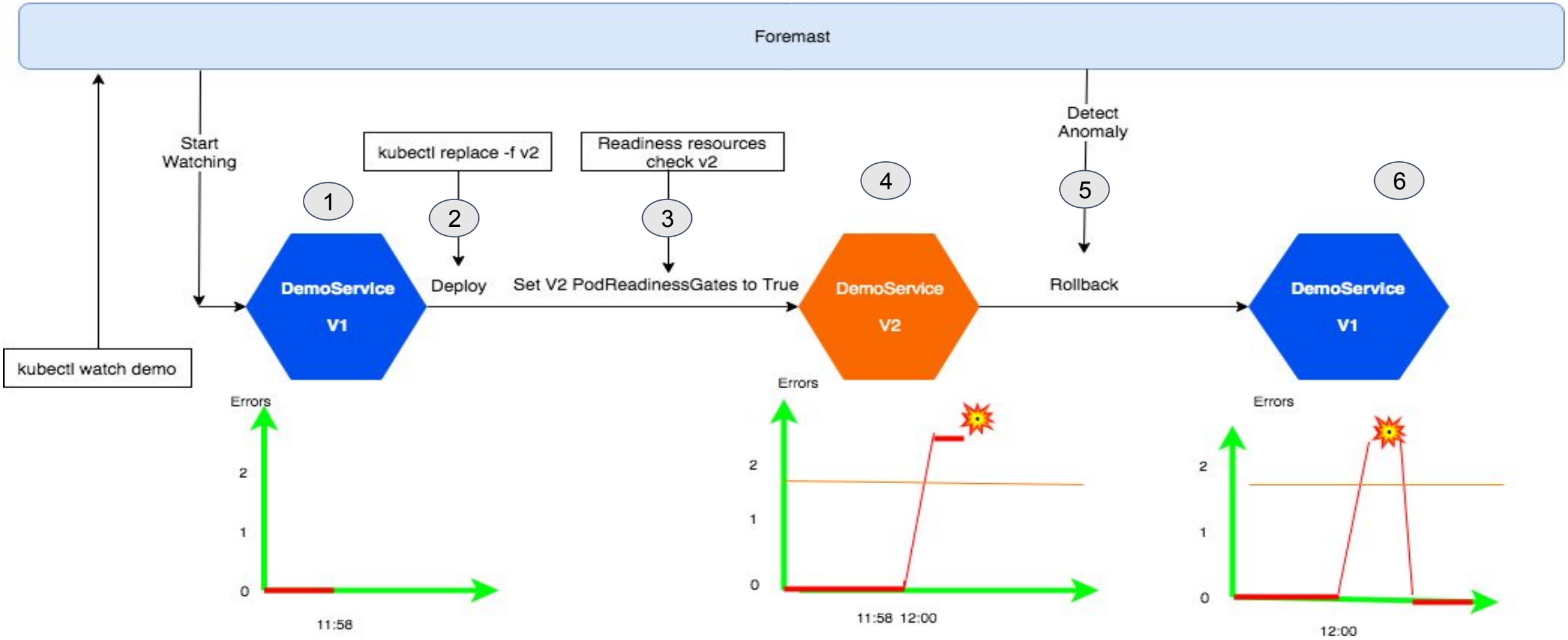
- Cloud Native, **Application Health Manager** for K8s
- AI based **Anomaly Detection** and Remediation
- **Observability** for Metrics, Logs, Traces
- Open sourced by **Intuit**

PodReadinessGates with Foremast

- Continuous Health Checks
- Avoid false alerts

intuit

Foremast Demo



Join the Foremast Team

KubeCon | CloudNativeCon Europe 2019

https://github.com/intuit/foremast

Dawei Ding

@dwding18

Ping Zou

@pzou1974

Sheldon Shao

@shaoxt

Sen Lin

@formuzi

Ed Lee

@edlee2121

Mukulika Kapas

@mukulikak

Kian Jones

@kianjones4

David Masselink

@davemasselink

Srivathsan Canchi

@srivathsanvc

Debashis Saha

@debashissaha

intuit

Q & A

Foremast Demo (1)

Foremast-examples V2 Readiness Gates Status : (Not Set)

```
Name:                 demo-6cb994687d-wzklx
Namespace:            foremast-examples
Priority:             0
PriorityClassName:    <none>
Node:                 minikube/10.0.2.15
Start Time:           Fri, 17 May 2019 13:02:16 -0700
Labels:               app=demo
                      pod-template-hash=6cb994687d
                      version=v2
Annotations:          prometheus.io/path: /actuator/prometheus
                      prometheus.io/port: 8080
                      prometheus.io/scheme: http
                      prometheus.io/scrape: true
Status:               Running
IP:                   172.17.0.15
Controlled By:        ReplicaSet/demo-6cb994687d
Containers:
  app:
    Container ID:   docker://e50a6b3f22494c91cd5e760146820a016043c4f8a4bd394070156bb44b86830d
    Image:          docker.io/foremast/k8s-metrics-demo:0.0.122
    Image ID:       docker-pullable://foremast/k8s-metrics-demo@sha256:af6af61ff8103ea9ad3e7ea986dff05c8e4f475d972c
a5e54740e95e91f8364a
    Port:           8080/TCP
    Host Port:      0/TCP
    State:          Running
      Started:      Fri, 17 May 2019 13:02:18 -0700
    Ready:          True
    Restart Count:  0
    Environment:
      APP_NAME:    demo
      JAVA_OPTS:   -DerrorType=5xx -Dfilename=/app/resources/data2.txt -DVALIDATION_TIME=26000 -DMEMORY_SIZE=9900000
0
    Mounts:
      /var/run/secrets/kubernetes.io/serviceaccount from default-token-lflgm (ro)
Readiness Gates:
  Type                  Status
  POD_GATE_CONDITION    <none>
Conditions:
  Type              Status
  Initialized       True
  Ready             False
  ContainersReady   True
  PodScheduled      True
Volumes:
```

Foremast Demo (2)

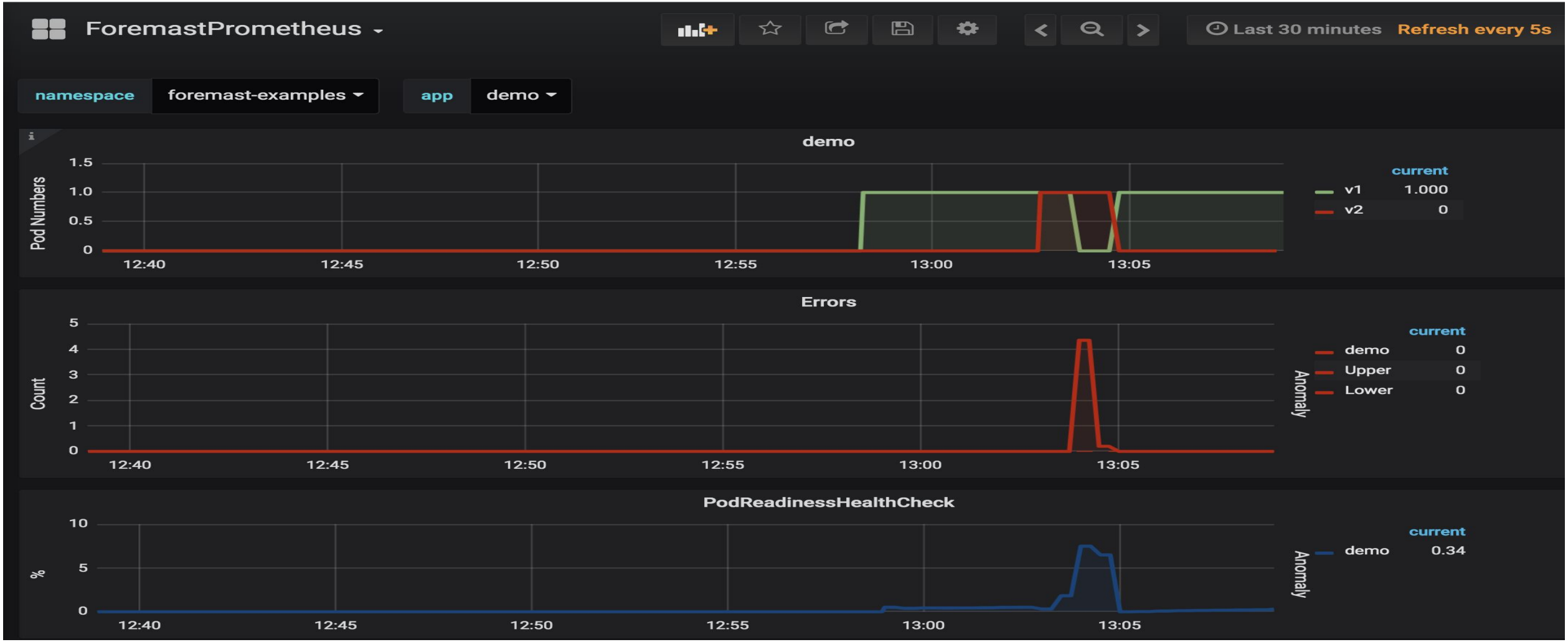
Foremast-examples V2 Readiness Gates Status : Ready after pod health check

```
MTVL160740c95:foremast pzou$ kubectl describe pod          demo-6cb994687d-wzklx            -n foremast-examples
Name:                  demo-6cb994687d-wzklx
Namespace:             foremast-examples
Priority:              0
PriorityClassName:     <none>
Node:                  minikube/10.0.2.15
Start Time:            Fri, 17 May 2019 13:02:16 -0700
Labels:                app=demo
                       pod-template-hash=6cb994687d
                       version=v2
Annotations:           prometheus.io/path: /actuator/prometheus
                       prometheus.io/port: 8080
                       prometheus.io/scheme: http
                       prometheus.io/scrape: true
Status:                Running
IP:                    172.17.0.15
Controlled By:         ReplicaSet/demo-6cb994687d
Containers:
  app:
    Container ID:    docker://e50a6b3f22494c91cd5e76014820a016043c4f8a4bd394070156bb44b86830d
    Image:           docker.io/foremast/k8s-metrics-demo:0.0.122
    Image ID:        docker-pullable://foremast/k8s-metrics-demo@sha256:af6af61ff8103ea9ad3e7ea986dff05c8e4f475d972c
a5e54740e95e91f8364a
    Port:            8080/TCP
    Host Port:       0/TCP
    State:           Running
      Started:       Fri, 17 May 2019 13:02:18 -0700
    Ready:           True
    Restart Count:   0
    Environment:
      APP_NAME:    demo
      JAVA_OPTS:   -DerrorType=5xx -Dfilename=/app/resources/data2.txt -DVALIDATION_TIME=26000 -DMEMORY_SIZE=9900000
0
    Mounts:
      /var/run/secrets/kubernetes.io/serviceaccount from default-token-lflgm (ro)
Readiness Gates:
  Type                    Status
  POD_GATE_CONDITION      True
Conditions:
  Type                    Status
  POD_GATE_CONDITION      True
  Initialized             True
  Ready                   True
  ContainersReady         True
  PodScheduled            True
```

Foremast Demo (3)

KubeCon CloudNativeCon Europe 2019

Foremast leverage PodReadinessGate to avoid false alarm



Foremast Demo (4)

Foremast detected unhealth

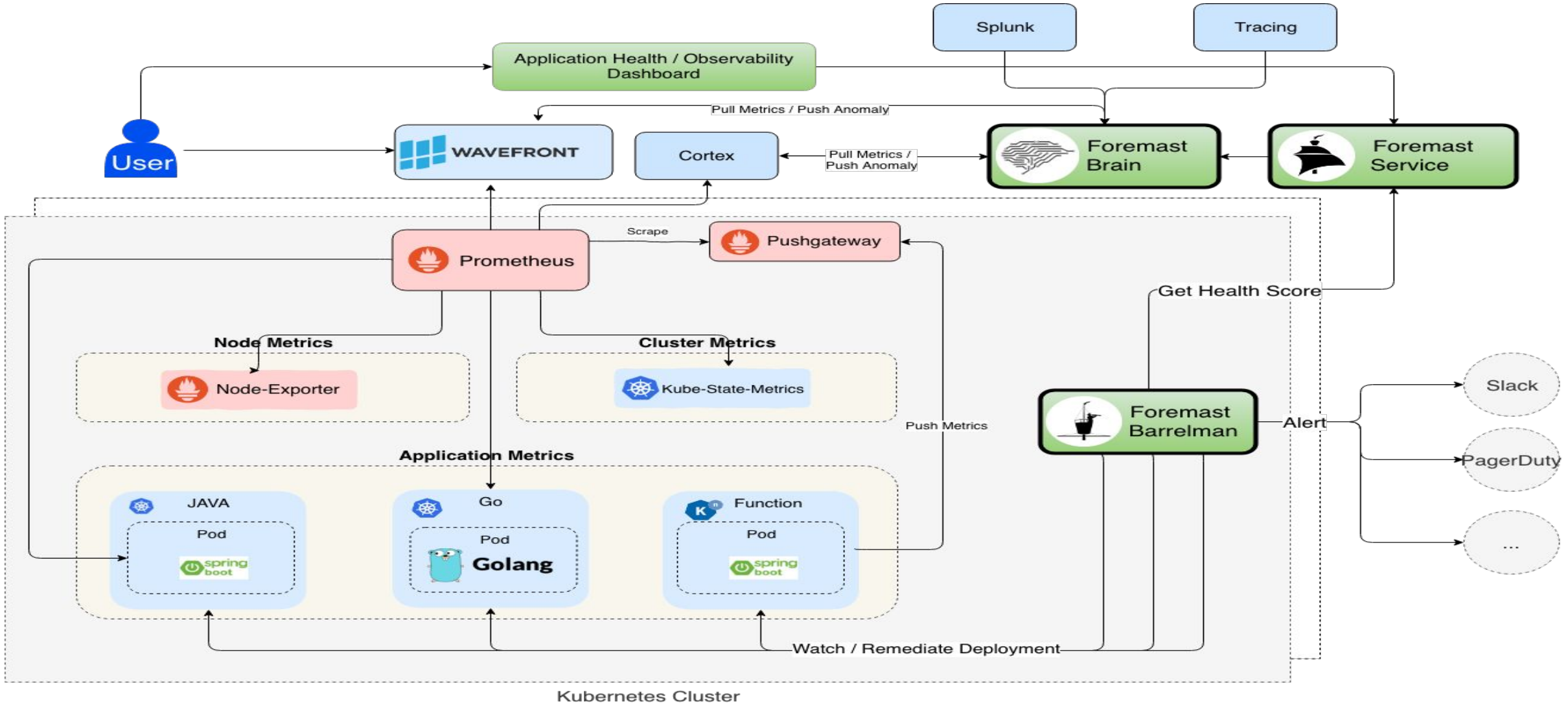
```
MTVL160740c95:foremast pzou$ kubectl get deploymentmonitor demo -o yaml -n foremast-examples
apiVersion: deployment.foremast.ai/v1alpha1
kind: DeploymentMonitor
metadata:
  annotations:
    deployment.kubernetes.io/name: demo
  creationTimestamp: "2019-05-17T19:57:42Z"
  generation: 8
  name: demo
  namespace: foremast-examples
  resourceVersion: "214598"
  selfLink: /apis/deployment.foremast.ai/v1alpha1/namespaces/foremast-examples/deploymentmonitors/demo
  uid: 07dea401-78de-11e9-b059-0800272fac8e
spec:
  analyst:
    endpoint: http://foremast-service.foremast.svc.cluster.local:8099/v1/healthcheck/
    version: 0.0.3
  metrics:
    dataSourceType: prometheus
    endpoint: http://prometheus-k8s.monitoring.svc.cluster.local:9090/api/v1/
    monitoring:
    - metricAlias: error5xx
      metricName: http_server_requests_error_5xx
      metricType: counter
  remediation:
    option: AutoRollback
  rollbackRevision: 3
  selector:
    matchLabels:
      app: demo
  startTime: "2019-05-17T22:35:50Z"
  waitUntil: "2019-05-17T23:05:50Z"
status:
  anomaly: {}
  expired: false
  jobId: 628b73c13d12cfc16930a73a50a22b8c3f7207b14b9428a0a018778b4326f325
  phase: Unhealthy
  remediationTaken: true
  timestamp: "2019-05-17T22:36:55Z"
```

Foremast Demo (5.1)

Foremast-examples V1 (Rollback) --- After detected V2 error anomaly

```
MTVL160740c95:foremast pzou$ kubectl describe pod demo-b5c6988df-f2vd6  -n foremast-examples
Name:                   demo-b5c6988df-f2vd6
Namespace:              foremast-examples
Priority:               0
PriorityClassName:      <none>
Node:                   minikube/10.0.2.15
Start Time:             Fri, 17 May 2019 13:04:15 -0700
Labels:                 app=demo
                        pod-template-hash=b5c6988df
                        version=v1
Annotations:            prometheus.io/path: /actuator/prometheus
                        prometheus.io/port: 8080
                        prometheus.io/scheme: http
                        prometheus.io/scrape: true
Status:                 Running
IP:                     172.17.0.4
Controlled By:          ReplicaSet/demo-b5c6988df
Containers:
  app:
    Container ID:   docker://f27678807becf6bc6cd9e379c517352f6f0fbb9d54a5eba299179790dad549cb
    Image:          docker.io/foremast/k8s-metrics-demo:0.0.122
    Image ID:       docker-pullable://foremast/k8s-metrics-demo@sha256:af6af61ff8103ea9ad3e7ea986dff05c8e4f475d972c
a5e54740e95e91f8364a
    Port:           8080/TCP
    Host Port:      0/TCP
    State:          Running
      Started:      Fri, 17 May 2019 13:04:17 -0700
    Ready:          True
    Restart Count:  0
    Environment:
      APP_NAME:  demo
    Mounts:
      /var/run/secrets/kubernetes.io/serviceaccount from default-token-lflgm (ro)
Conditions:
  Type              Status
  Initialized       True
  Ready             True
  ContainersReady   True
  PodScheduled      True
```

Foremast Architecture



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,