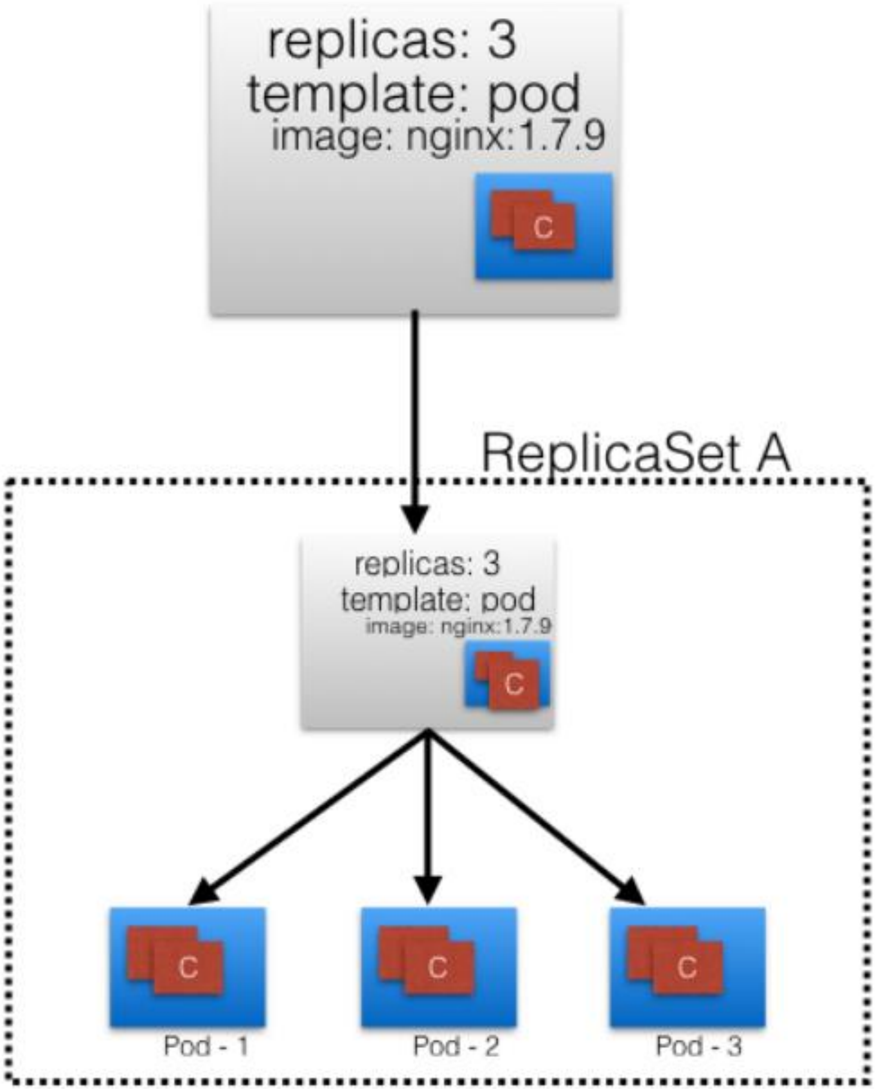
Kubernetes – Deployments

Deployment

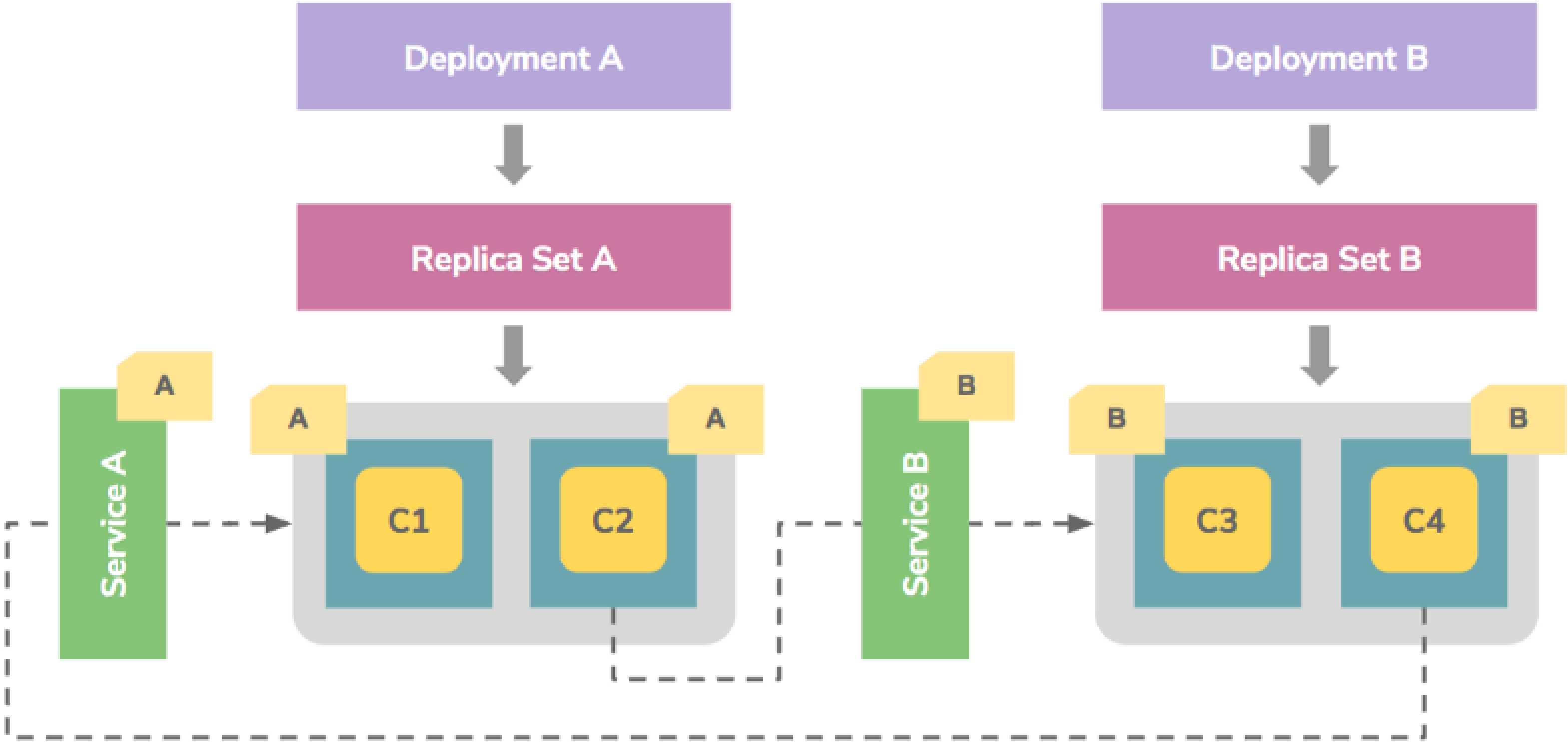
- ReplicaSet used for managing pods with desired state
- For functionalities like deployment, rolling out changes
- Ensures only certain number of pods are down during update e.g. 25 % max unavailable by default
- Deployment works with replicas set to add more capabilities like rollout changes, rollback etc

- maxUnavailable - max number of pods can be un-available during upgrade
- maxSurge - max number of pods can be scheduled above original number of pods
- e.g. maxSurge = 0 then no extra pods will gets launched during upgrade

Deployment


replicas: 3
template: pod
image: nginx:1.7.9
C
ReplicaSet A
replicas: 3
template: pod
image: nginx:1.7.9
C
C
C
C
Pod - 1
Pod - 2
Pod - 3

Deployment


Deployment A
Deployment B
Replica Set A
Replica Set B
A
A
A
Service A
C1
C2
B
B
B
Service B
C3
C4

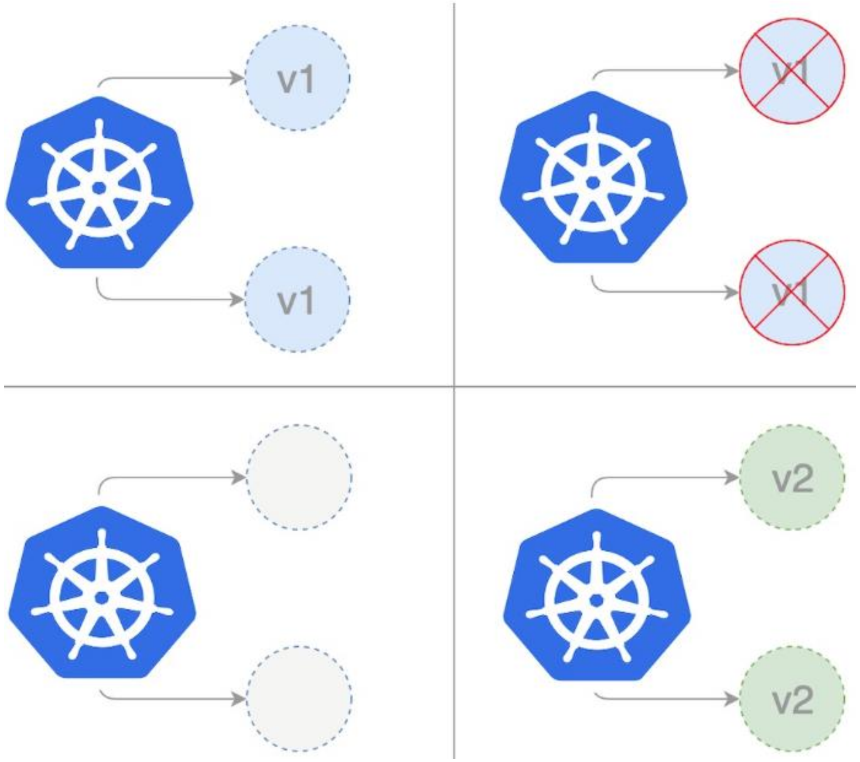
Deployment Strategy

- If v1 is running in prod, how will you update to v2?
- Change or upgrade an application

- Types
  - Recreate
  - RollingUpdate
  - Blue Green
  - Canary

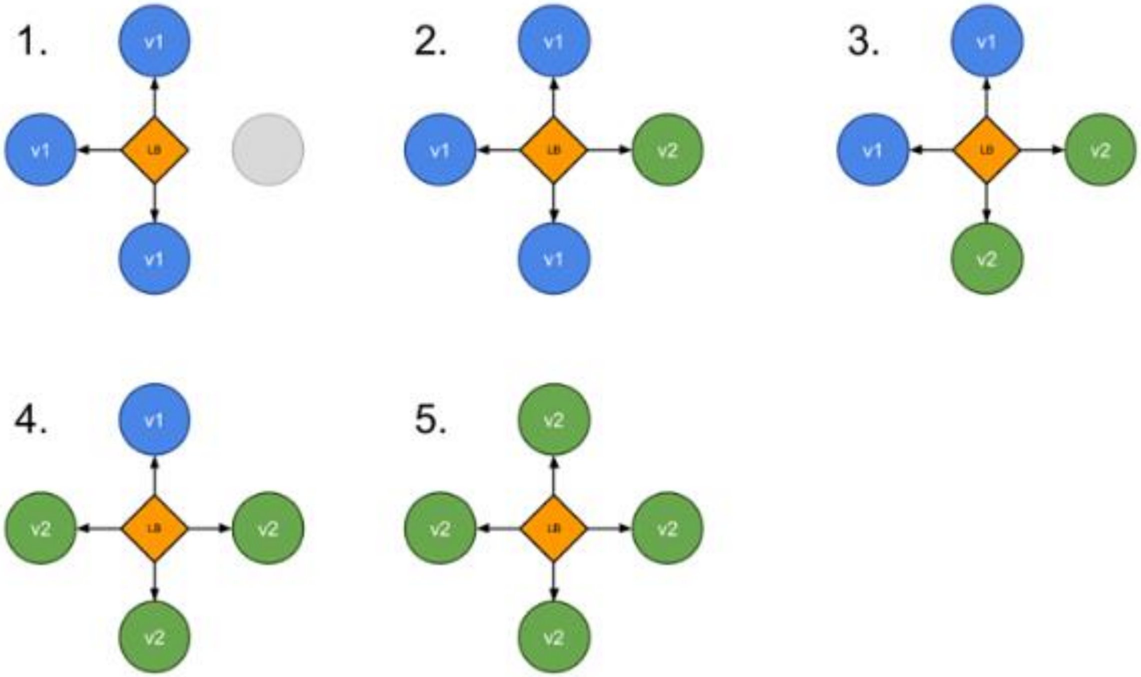
Deployment Strategy- Recreate

- Recreate
  - All pods gets killed all at once and replaced all at once by newer one
  - Application may not available during terminate and create stage
- Dev test env – recommended



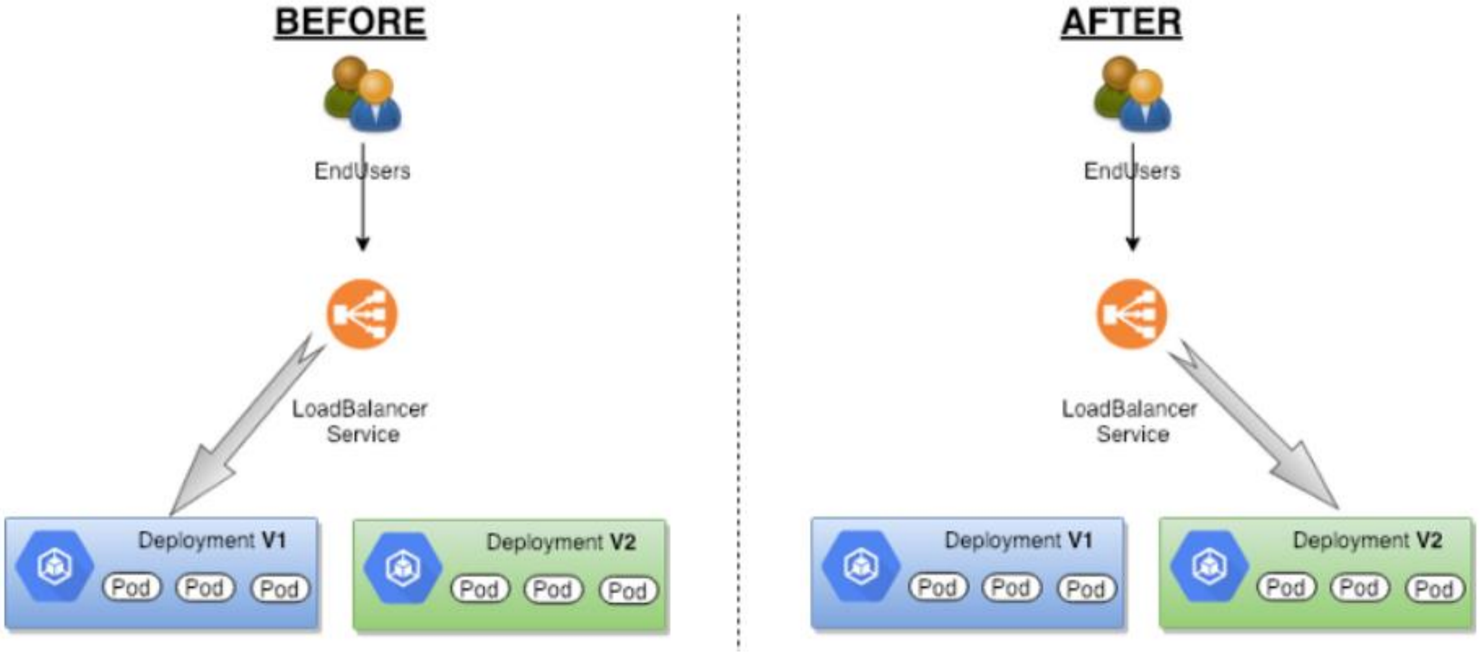
Deployment Strategy- RollingUpdate

- Upgrade happens one batch of instance at a time and not all
- Zero downtime



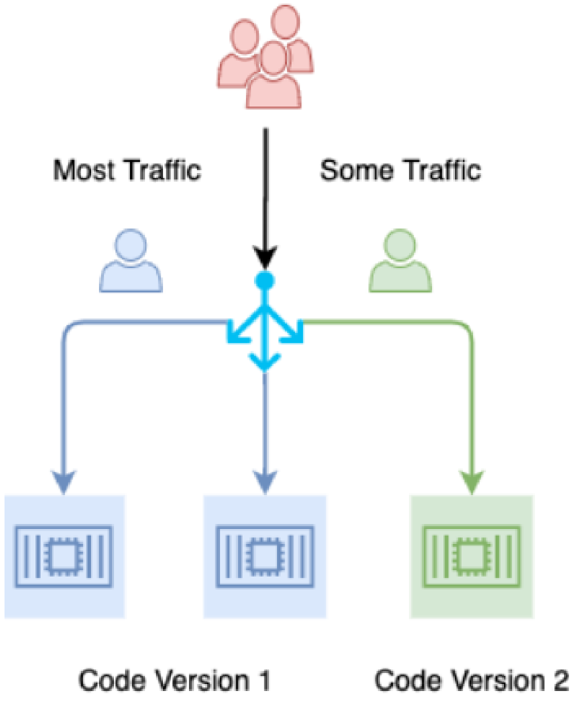
Deployment Strategy- Blue Green

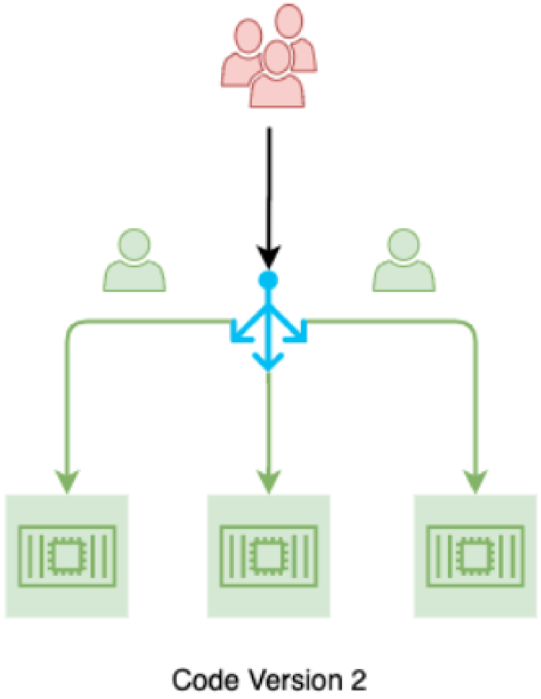
- Technique which reduces downtime and risk running two identical prod environment called blue and green
- Pros - Rollback immediately if required
- Cons - Need double amount of resources to run the parallel environment
- Zero downtime



Deployment Strategy- Canary

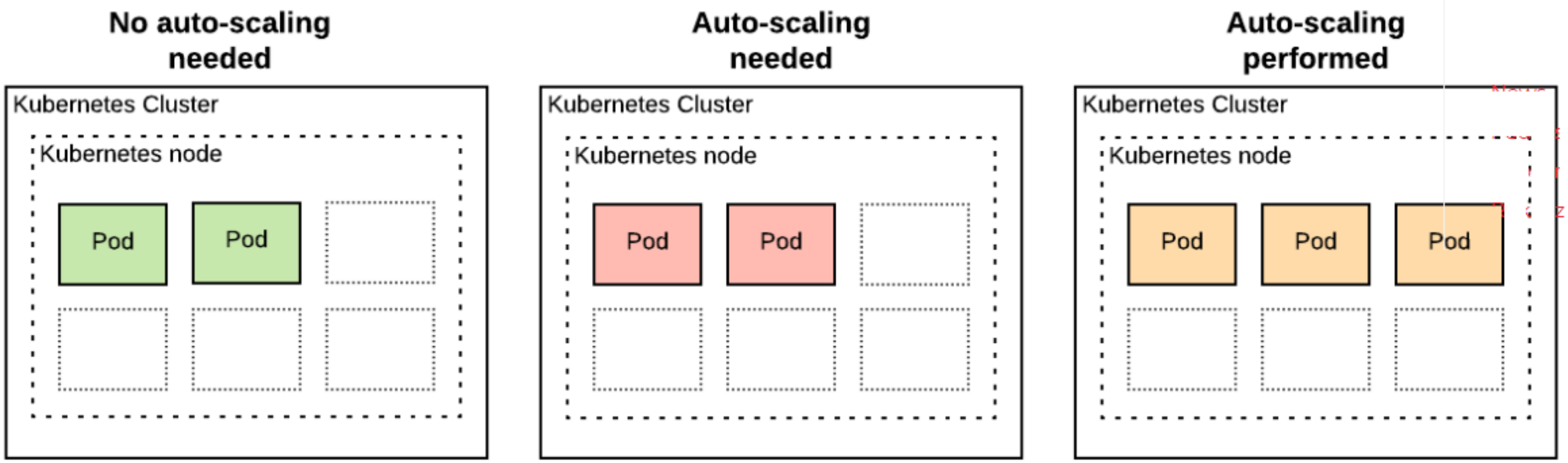
- During upgrade, some loads are getting shifted to newer version of app to see whether app is working successful
- No straight forward implementation supported in k8s
- External tools like estio
- Route53 - weighted policy





Scaling Deployment

- To automatic scale up and scale down the deployment



No auto-scaling needed

Auto-scaling needed

Auto-scaling performed

High CPU usage
Medium CPU usage
Low CPU usage