CLOUD NATIVE + OPEN SOURCE

Virtual Summit China 2020

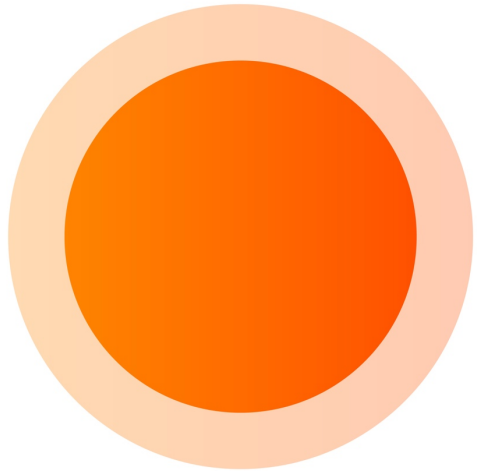
How we Manage our Widely Varied Kubernetes Infrastructures in Alibaba

Ziren Wan - Alibaba Cloud

Jie Chen - Alibaba Cloud

Agenda

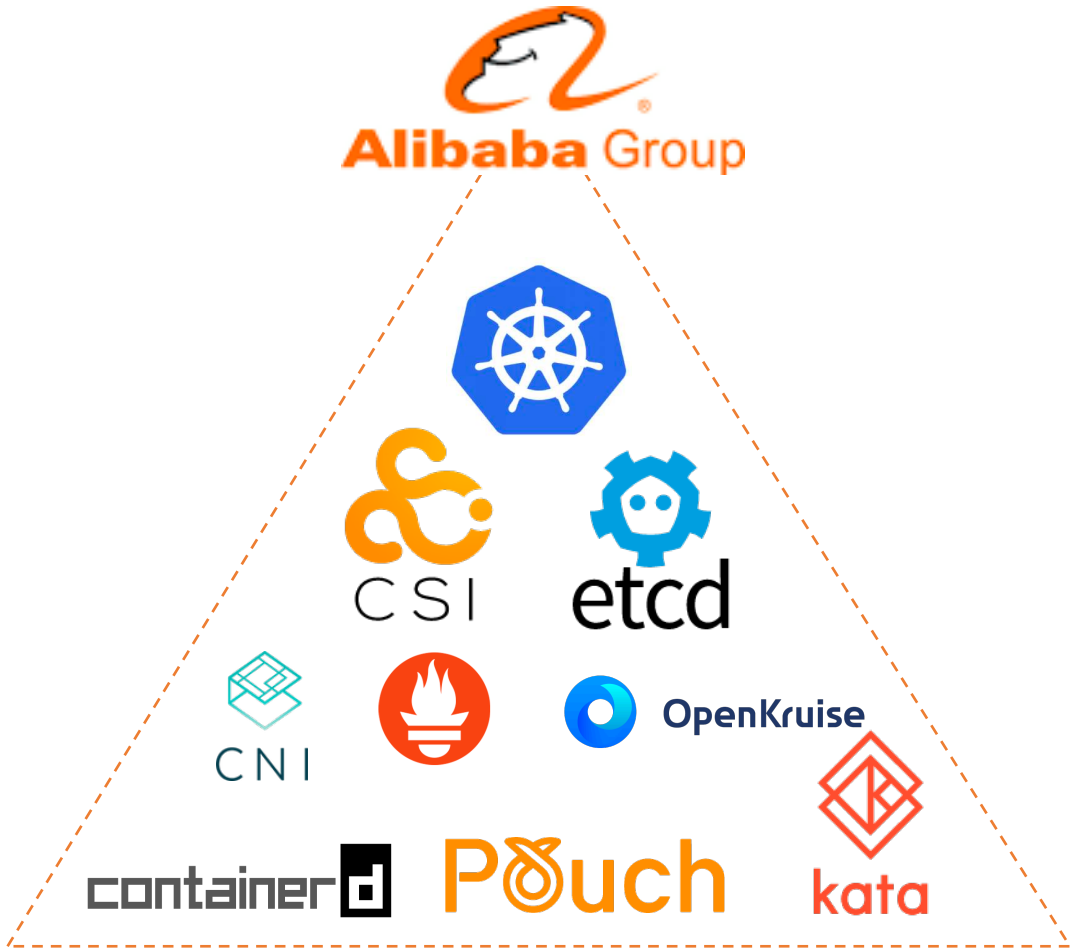
Alibaba Cloud | Worldwide Cloud Services Partner



- Background
- Alibaba Kubernetes Architecture
- Infrastructure Management
- CI/CD Pipelines
- Quick Demo

Background

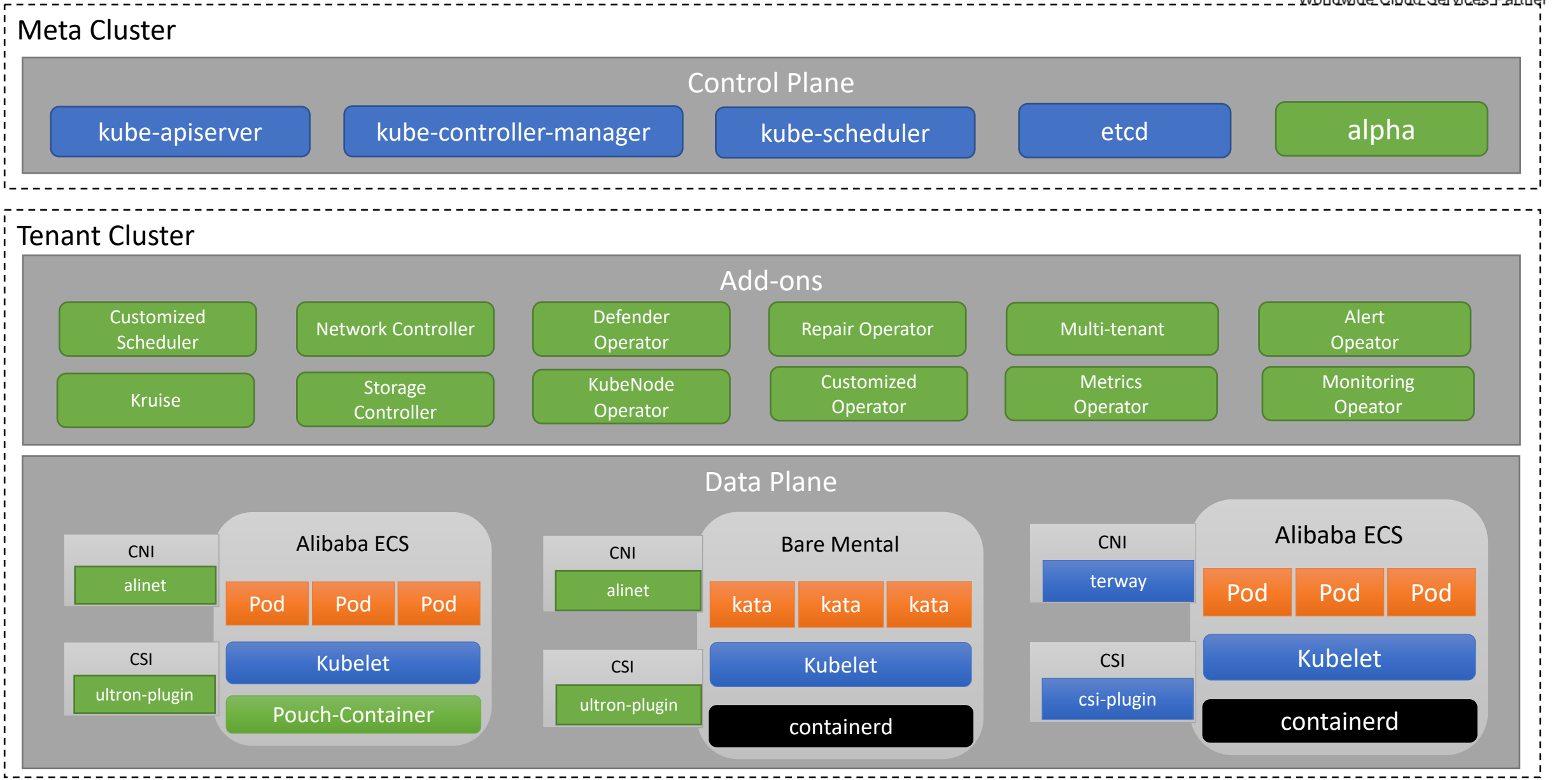
Alibaba Cloud | Worldwide Cloud Services Partner



- Who are we?
- Scale of Alibaba Kubernetes Clusters (handreds of internal clusters, 5k-10k nodes each)
- Variety of Cluster Infrastructures (200+ addons)
- Significance of keeping the stability in large-scale clusters.

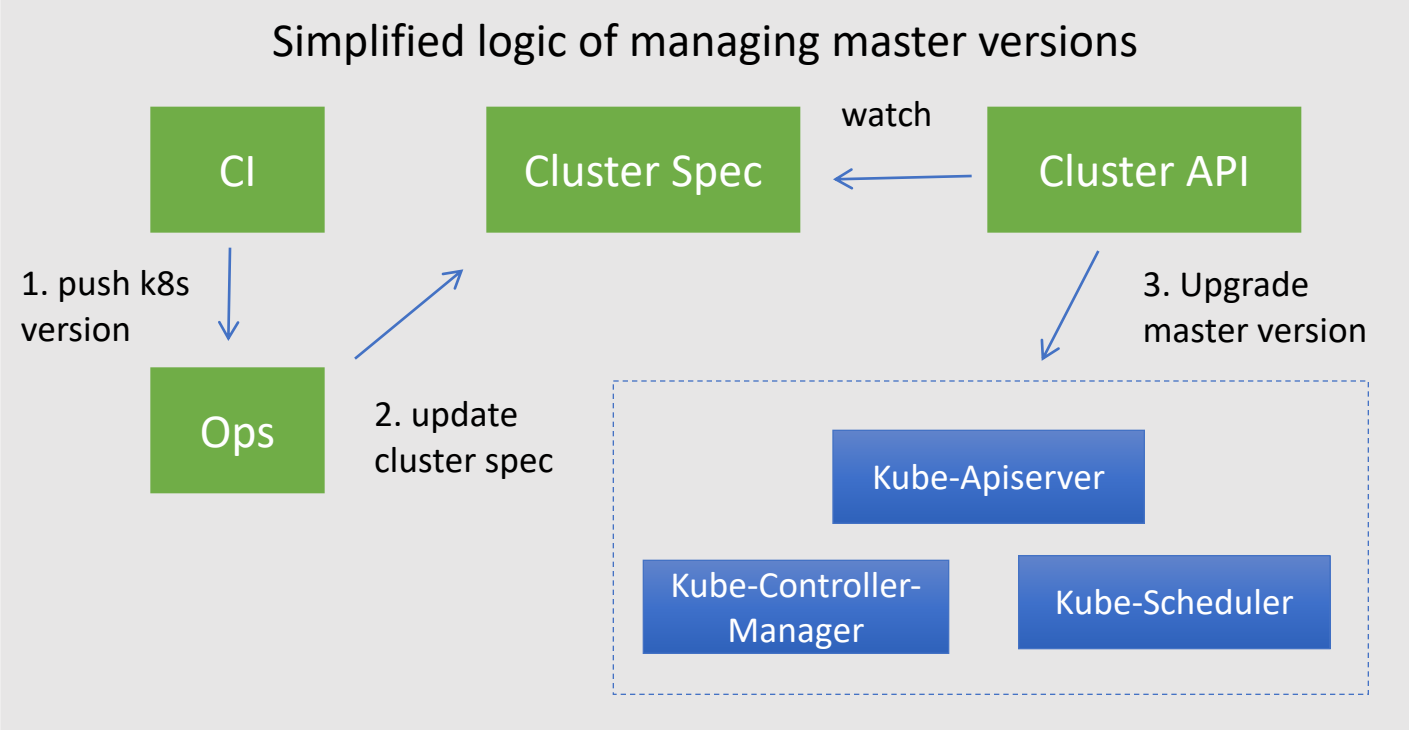
Architecture of Alibaba Kubernetes Infrastructure

Alibaba Cloud | Worldwide Cloud Services Partner



Infrastructure Management - Master

Alibaba Cloud | Worldwide Cloud Services Partner

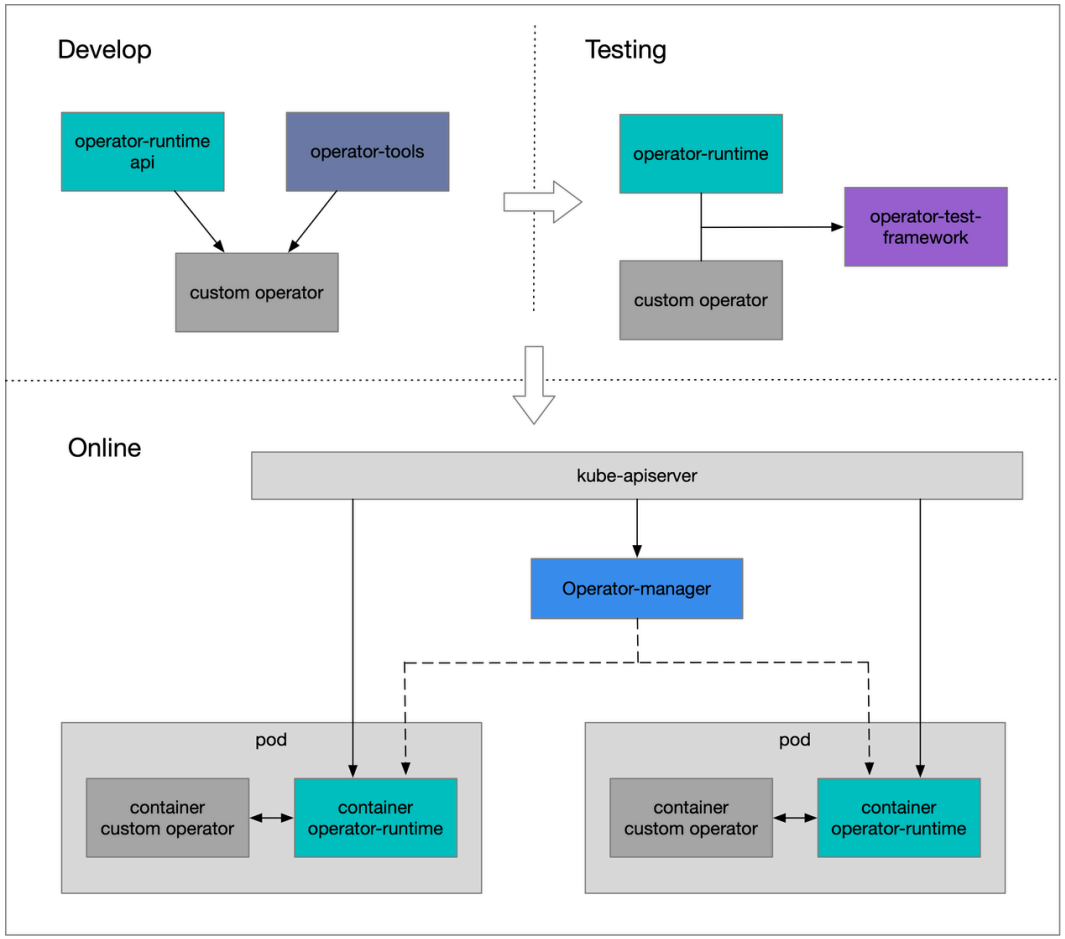



```
apiVersion: alibabacloud.com/v1alpha1
kind: Cluster
metadata:
  labels:
    cluster.id: c3f1b726caecf4d0ca076f73ee781e312
name: kubernetes-cluster
namespace: c3f1b726caecf4d0ca076f73ee781e312
spec:
  kubernetes:
    kcm:
      commit: 0bfce06
      name: kubernetes.kdm.kcm
      replicas: 3
      version: v1.16.3-alibaba.2
    kore:
      name: kubernetes.kdm.korepanel
      replicas: 3
      version: v1.16.3-alibaba.2
    rols:
      name: kubernetes.kdm.roles
      version: v1.16.3-alibaba.2
    scheduler:
      commit: 0bfce06
      name: kubernetes.kdm.scheduler
      replicas: 3
      version: v1.16.3-alibaba.2
```

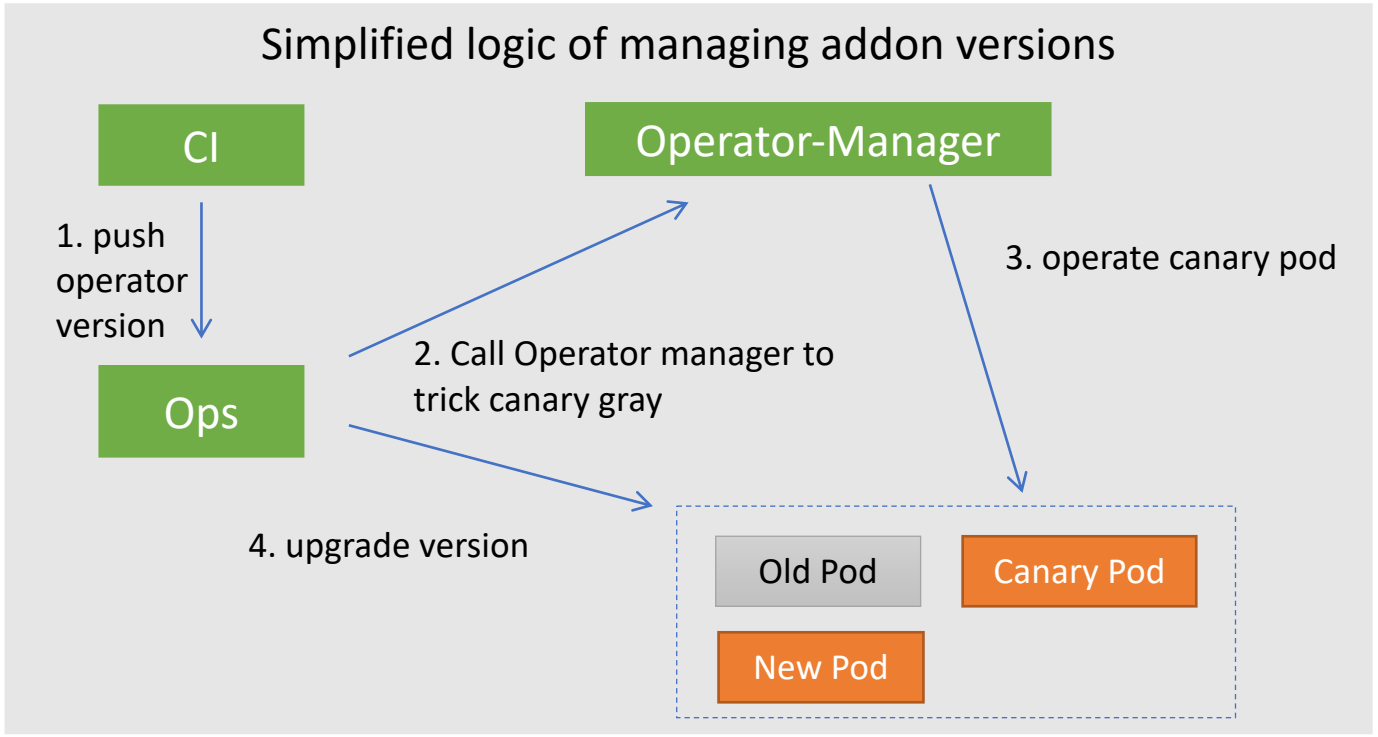
Kubernetes Version

- Use Cluster API manage master version

Infrastructure Management - Addon



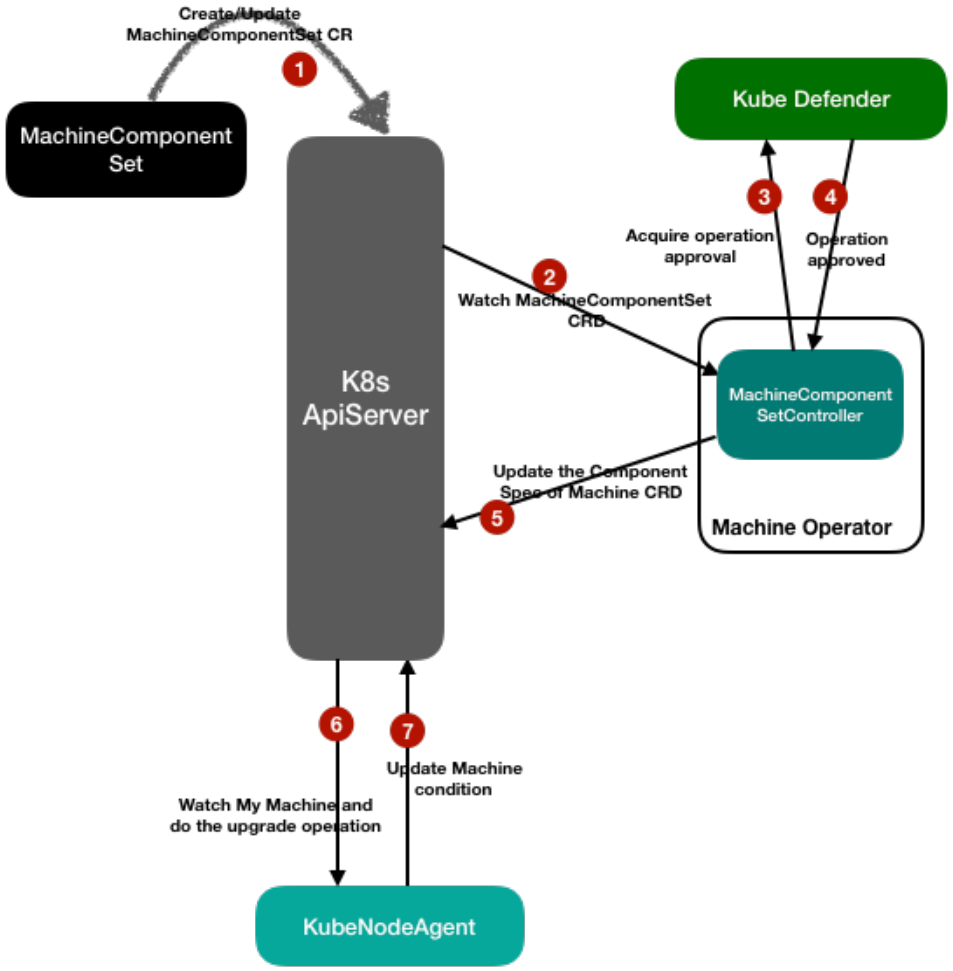
- Operator manager infrastructure

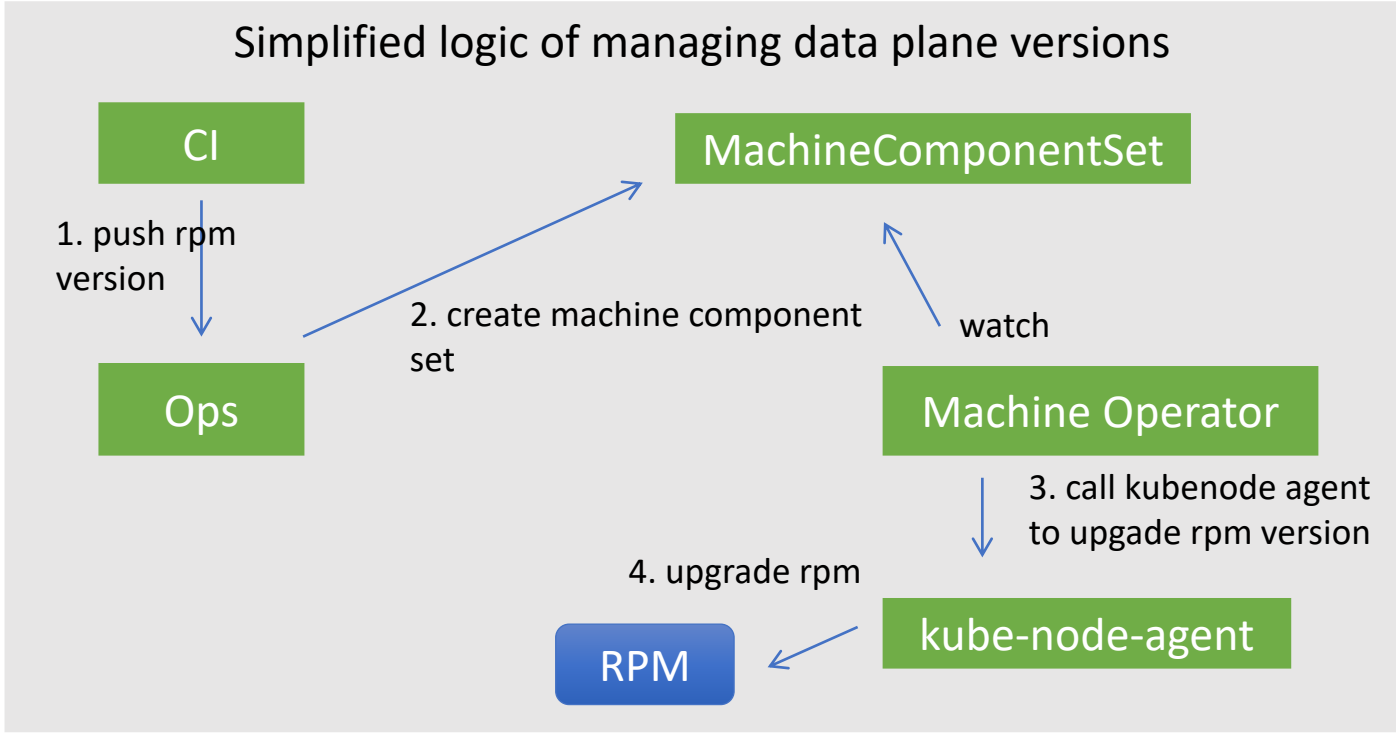


1. first create a canary pod and
2. then update operator rules and watching the canary pod status
3. call UpdateOperatorRule to empty the rules and delete the canary pod
4. upgrade to new version

Infrastructure Management - Dataplane

Alibaba Cloud | Worldwide Cloud Services Partner





- KubeNode: upgrade a dataplane component
- use partition to controller the batch of gray

“Philosophy”

Alibaba Cloud | Worldwide Cloud Services Partner

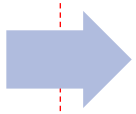
- **Components varied from different clusters**
  - How to manage components
- **Always provide the stable component version**
  - How to make stable releases
- **Continuous and non-disruptive cluster delivery**
  - How to build safe delivery pipelines

Component Management

Image-Oriented

- Only patch container image
- Simple but not fit to all cases

Design for CI



YAML-Oriented

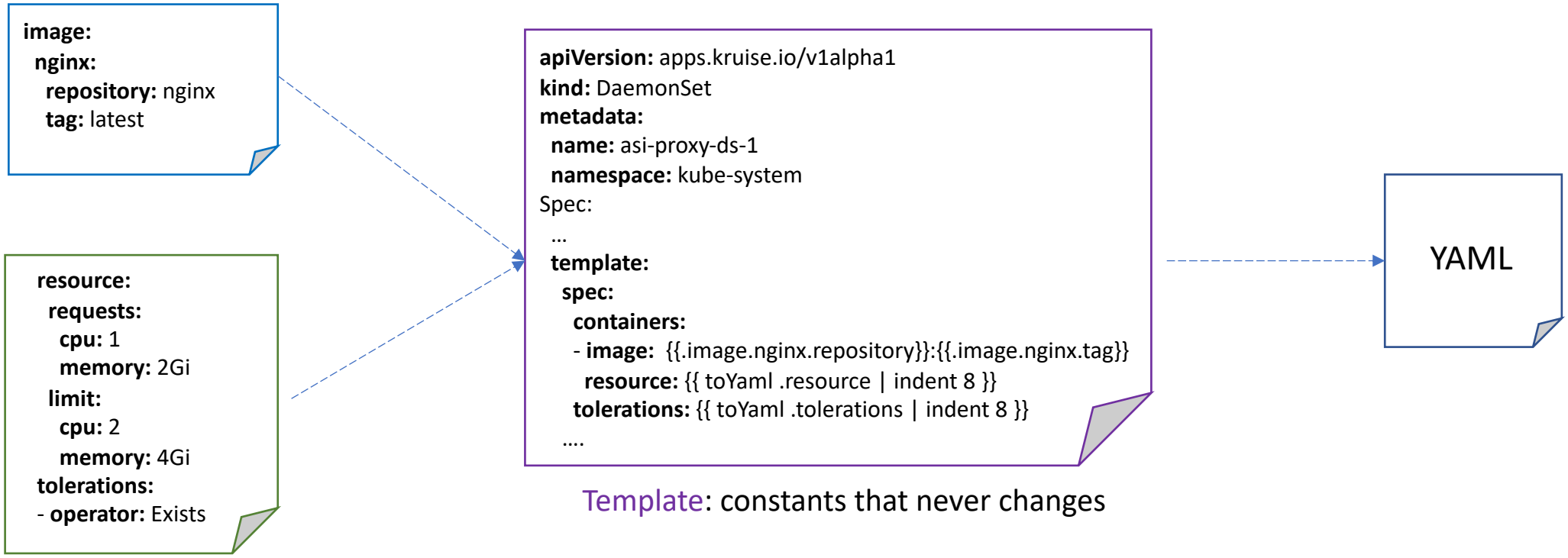
- Helm template
- Separate image and meta-config

Helm + Version Control

Component Management

- Infrastructure Components = YAML = Template + Image + Meta-Config

Image: expected to be the same

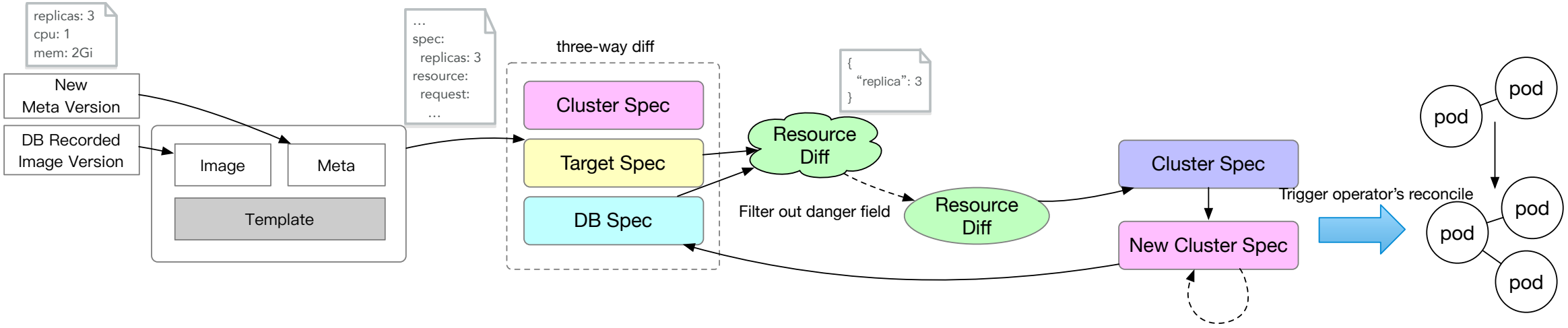


Template: constants that never changes

Meta-Config: Varies from cluster to cluster

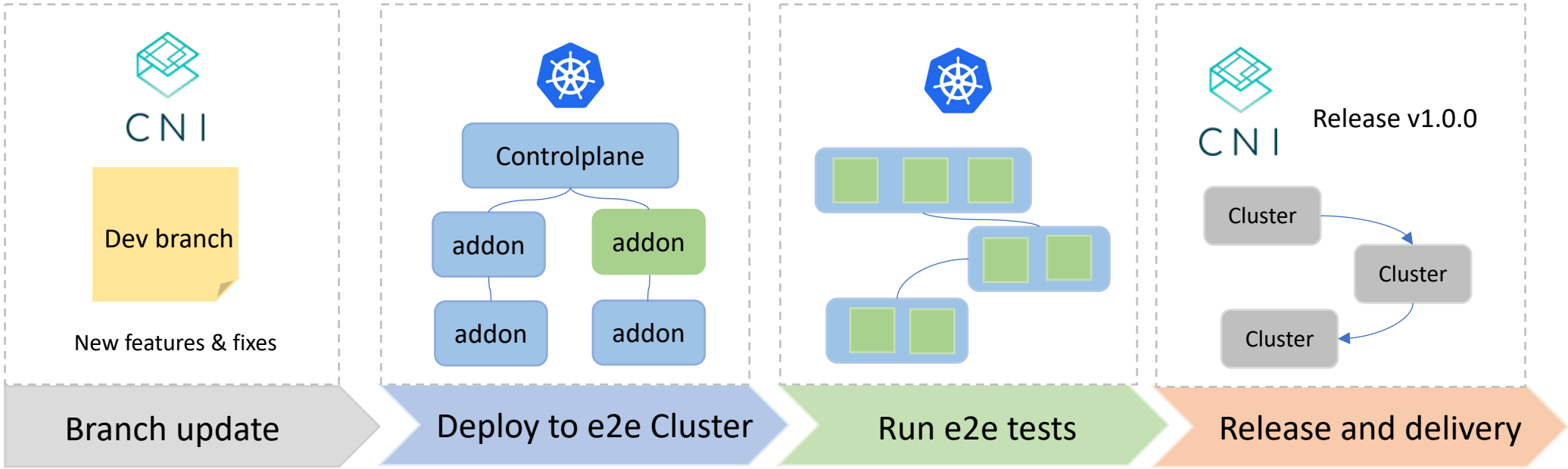
Component Management

- Do things like that **kubectl apply** does
  - Compare with current spec/cluster spec
  - PATCH diff to apiserver



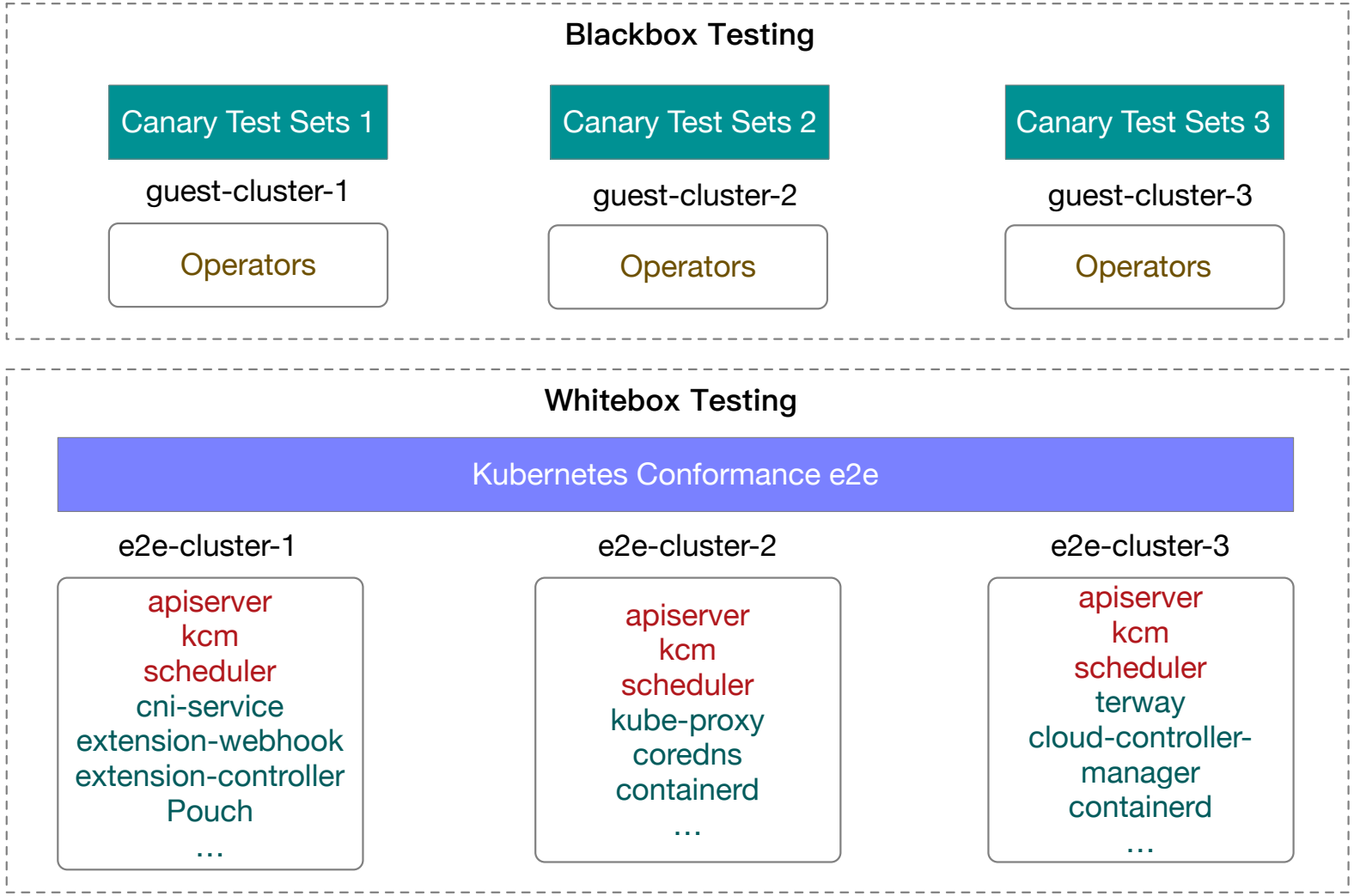
Version Release & Testing

Alibaba Cloud | Worldwide Cloud Services Partner



Version Release & Testing

- e2e testing is not enough
- **Canary tests** runs continuously
  - Create/delete pod/sts/deploy
  - Upgrade sts/deploy
  - Scale up/down sts/deploy
  - Create Job
  - Create CustomResouce
  - …



Intra-cluster upgrade

Alibaba Cloud | Worldwide Cloud Services Partner

- Rolling updates for Kubernetes Workloads
  - Deployment (Kruise)
  - StatefulSet (Kruise)
  - DaemonSet (Kruise)
  - Dataplane components (KubeNode)
- Rollout Policy
- Pause/Resume
- Max unavailable

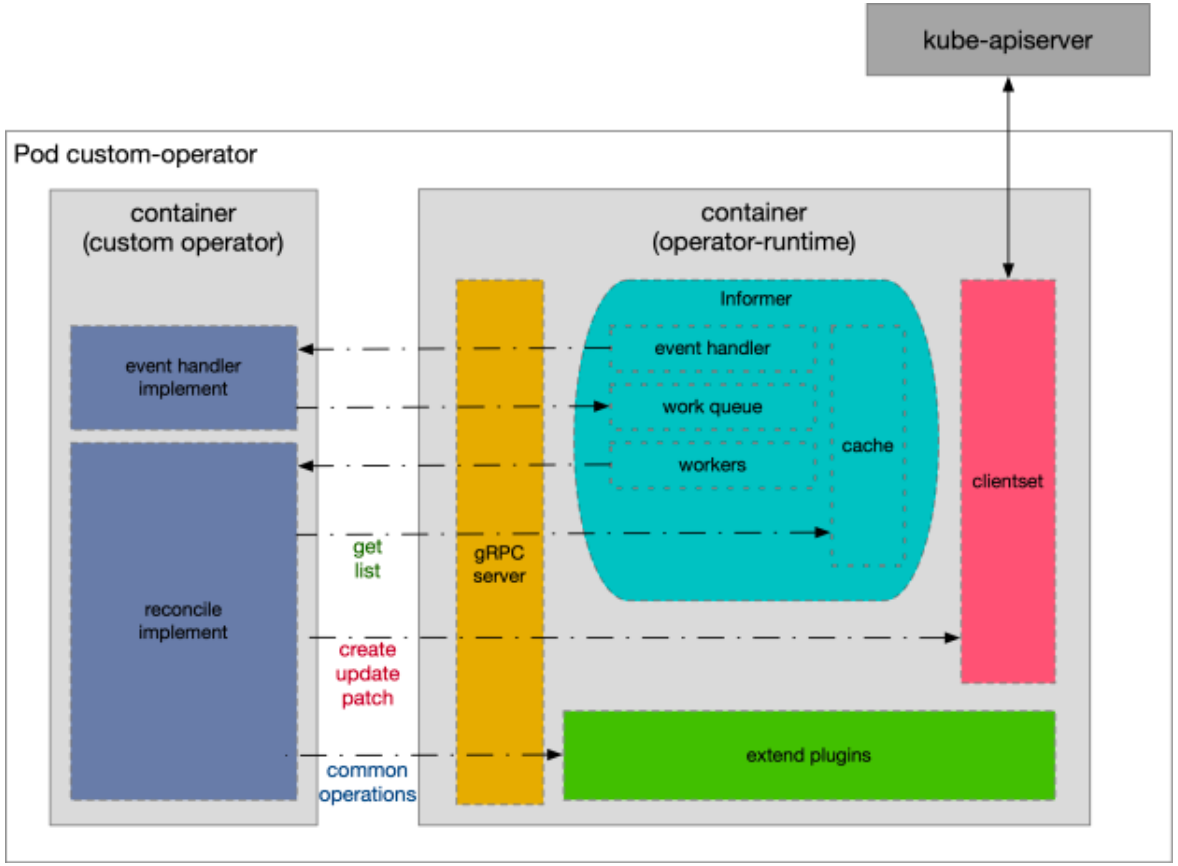
	Deployment	StatefulSet	DaemonSet	Dataplane Components
Rollout Policy	RollingUpdate Canary Deploy	RollingUpdate Canary Deploy	RollingUpdate	RollingUpdate
Pause/Resume	Yes	Yes	Yes	Yes
Max unavailable	Not yet	Yes	Yes	Yes
Partition	No	No	Yes	Yes

OpenKruise https://github.com/openkruise/kruise

Rollout for operators

Alibaba Cloud | Worldwide Cloud Services Partner

- Enhance the ability of Operator (StatefulSet / Deployment)
  - Implement operator as the way kubebuilder does
  - Sidecar container which contains clientset, informer and plugins
  - Serving operator with gRPC requests

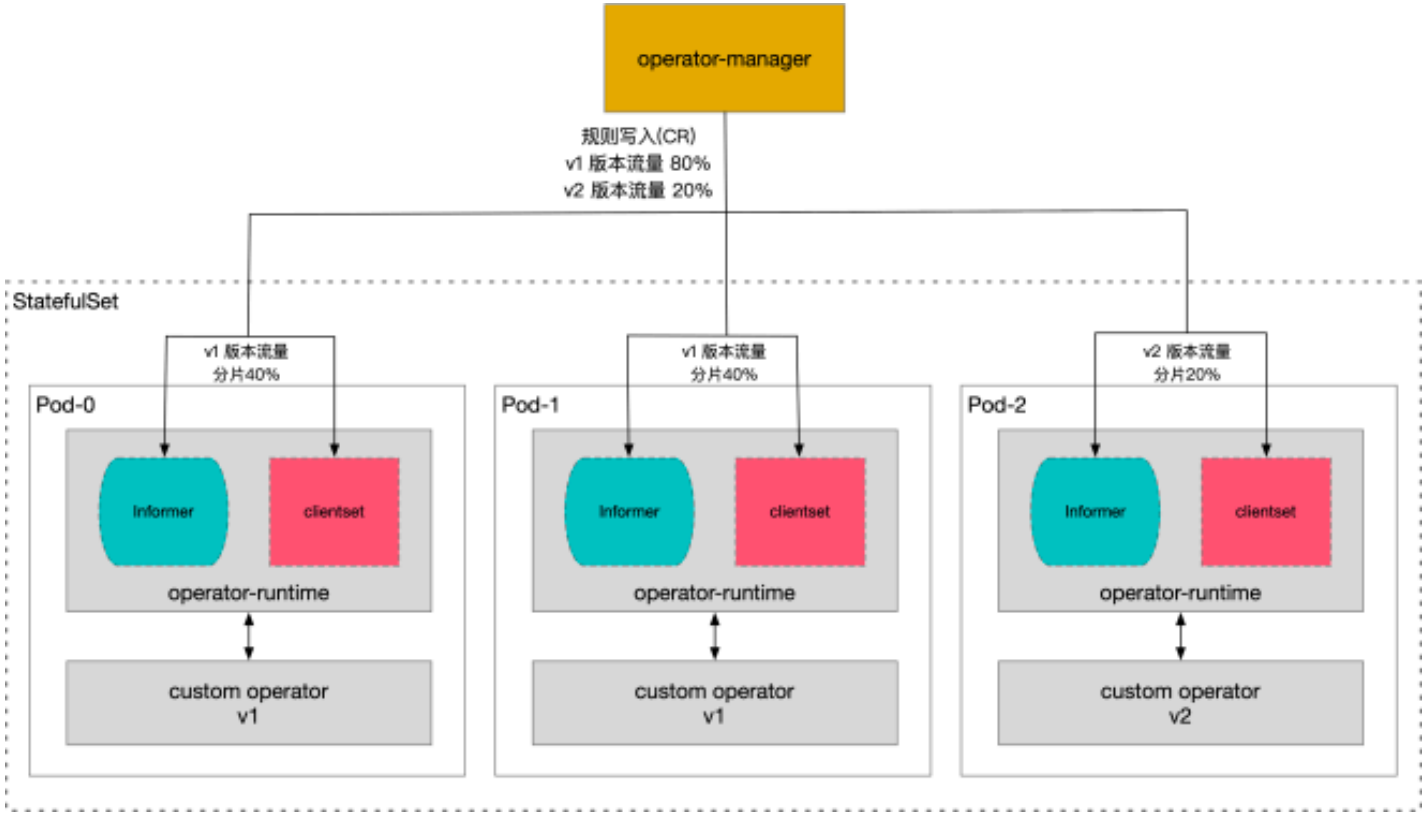


OpenKruise https://github.com/openkruise/kruise

Rollout for operators

Alibaba Cloud | Worldwide Cloud Services Partner

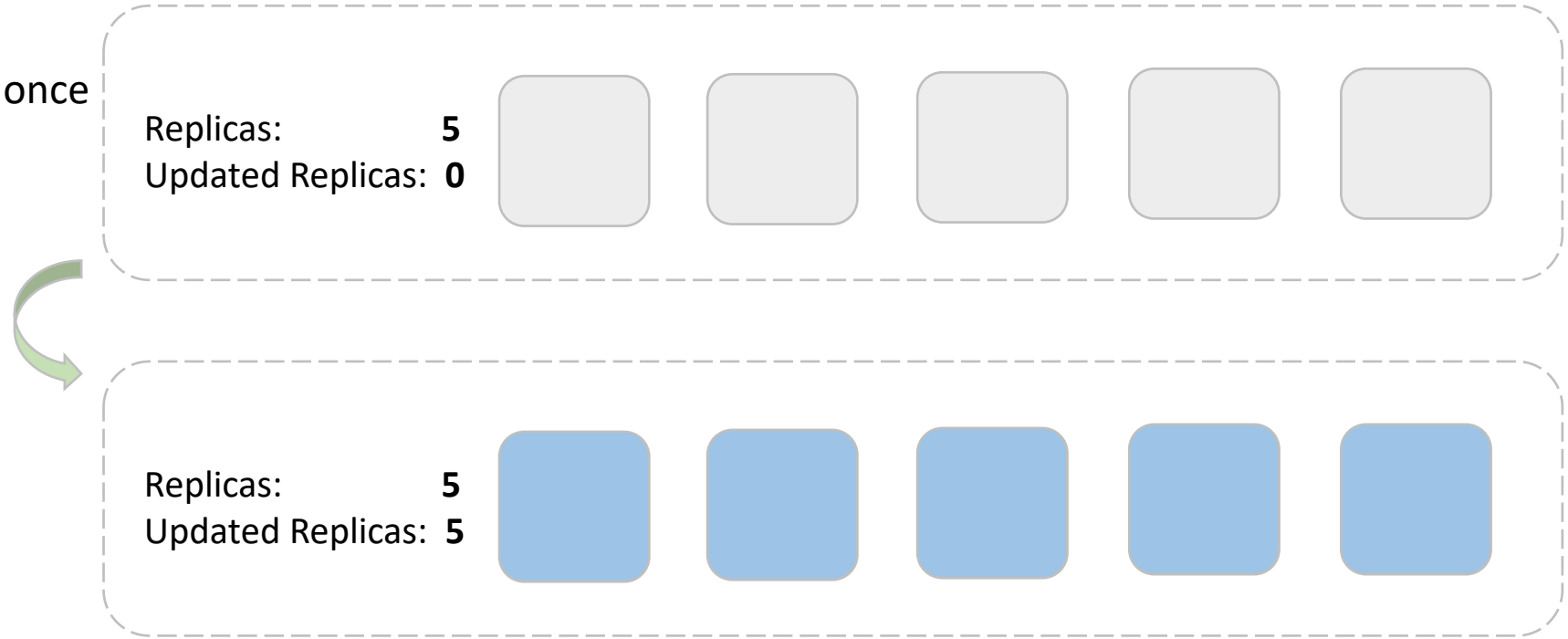
- Canary deploy for Operators
  - Flow control on a monilithic manager
  - Flow slice controlled by rule (Custom Resource)
  - Rolling update



OpenKruise https://github.com/openkruise/kruise

Rollout for DaemonSet

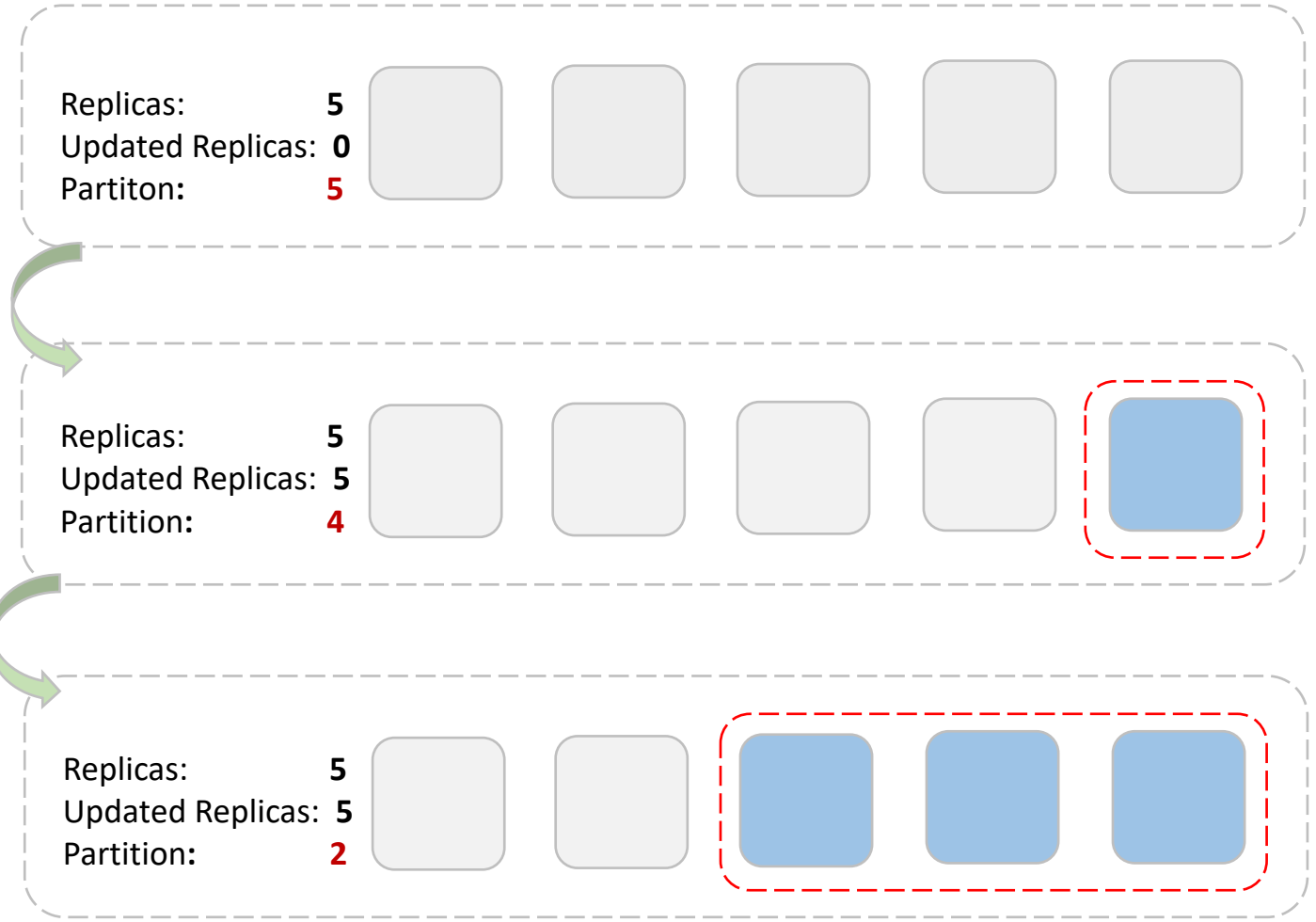
- **Original DaemonSet**
  - Lack of the ability of rolling update
  - always updates all pods once image changes
  - **OnDelete** ?



Rollout for DaemonSet

Alibaba Cloud | Worldwide Cloud Services Partner

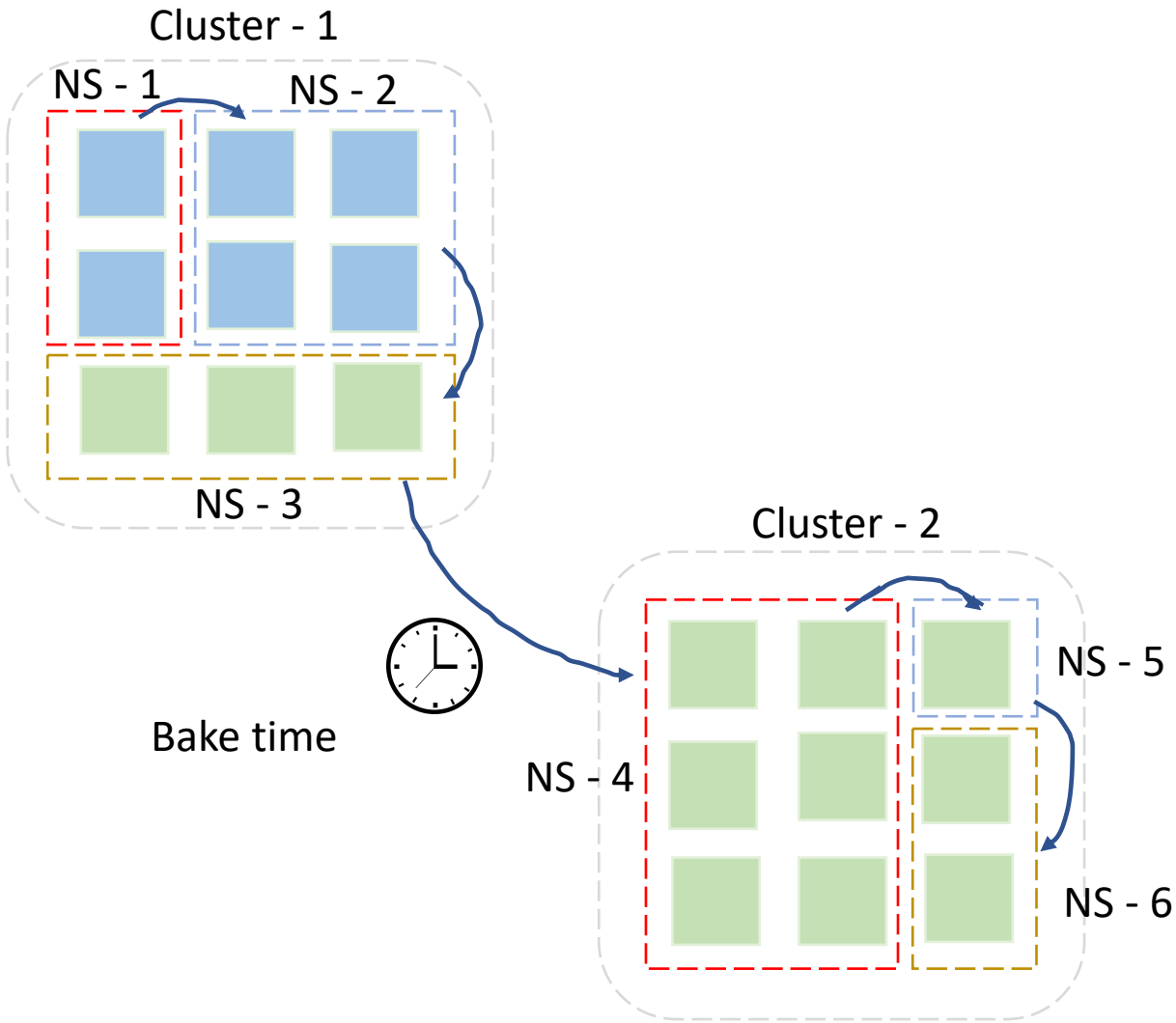
- **Kruise**: Enhance the ability of DaemonSets
  - **Partition**: the number of pods remained to be old version
  - **MaxUnavailable:** the maximum number of pods can be unavailable during rolling update



OpenKruise https://github.com/openkruise/kruise

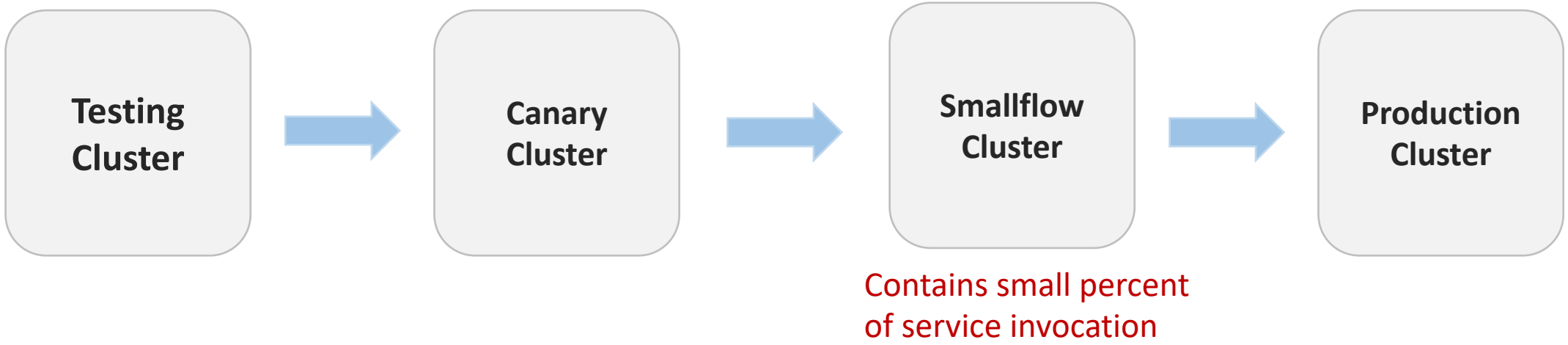
Rollout for Dataplane

- Kubelet / Pouch / containerd …
- Similar to Kruise Daemonset on patition control
- **NodeSet**: a group of nodes which has the same characters, minimum rollout unit
  - Rolling update in each NodeSet
  - Upgrade NodeSet sequentially



Inter-cluster upgrades

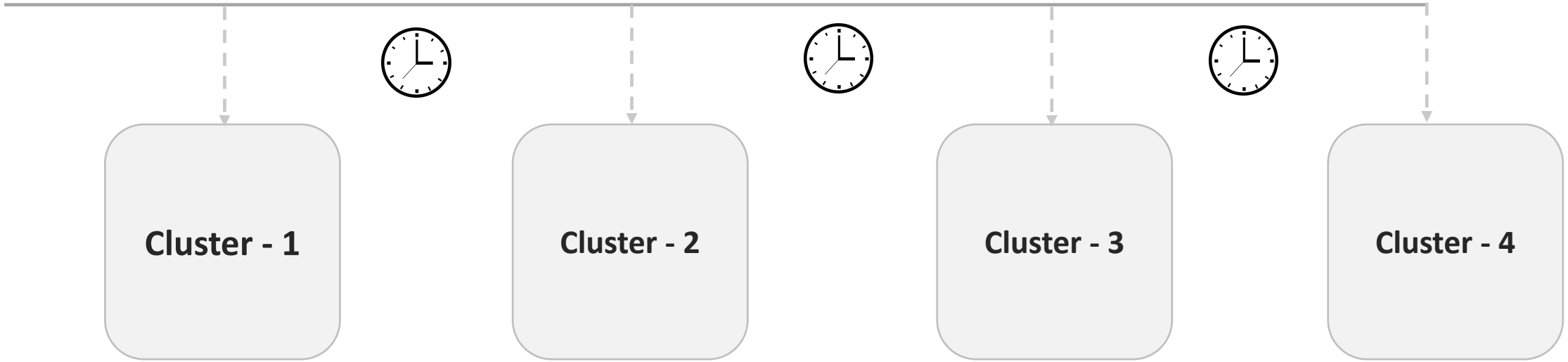
- Inter-cluster rollout pipelines
  - Orchestrate clusters with scale / importance of upper biz apps
  - Build a gray release pipeline
    - Tekton-liked implementation



Inter-cluster upgrades

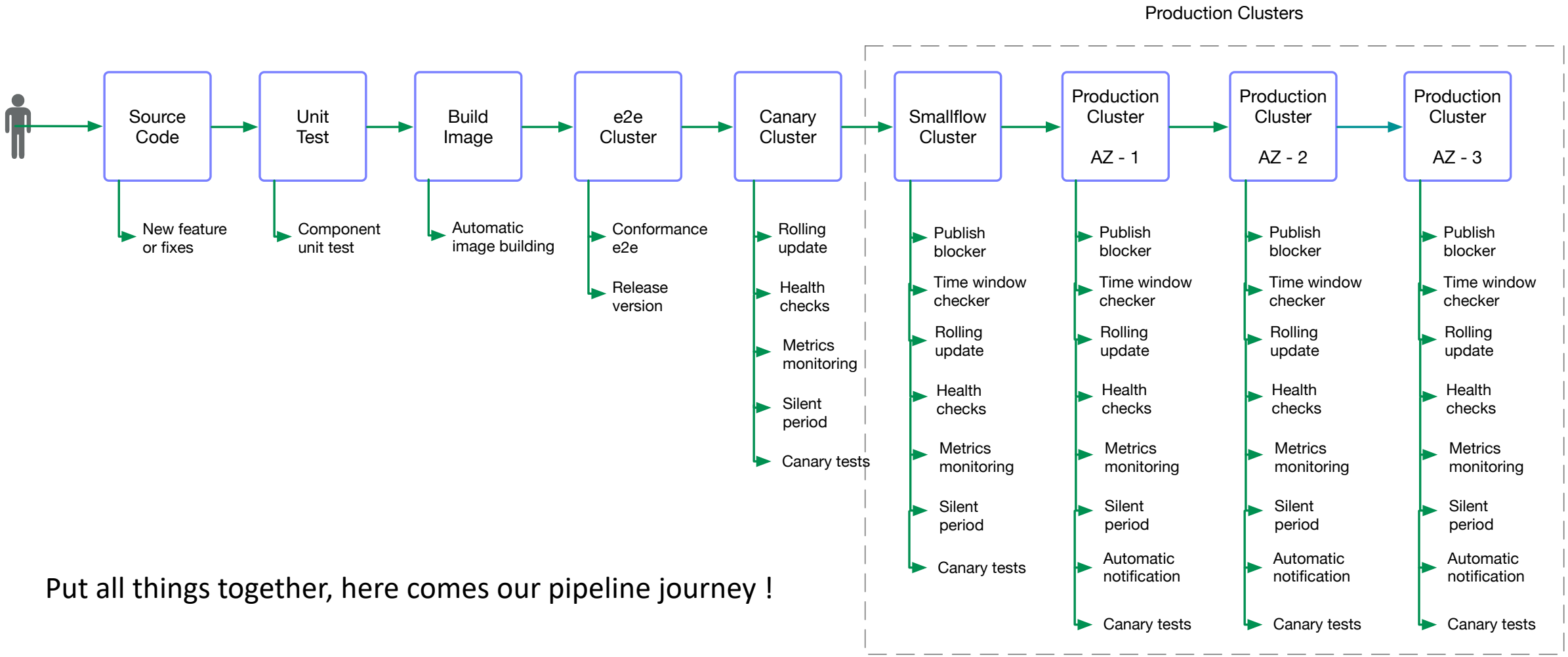
Alibaba Cloud | Worldwide Cloud Services Partner

- Inter-cluster rollout pipelines
  - Silent period between each clusters
  - Pre-checking and post-checking
    - time window checker / rule-based blocker
    - Metric monitoring / health checks



Pipelines – inter-cluster

Alibaba Cloud | Worldwide Cloud Services Partner



Put all things together, here comes our pipeline journey !

CLOUD NATIVE + OPEN SOURCE

Virtual Summit China 2020