Node Operator: Kubernetes
Node Management Made Simple

陈俊, Ant Financial
Agenda

• Background and Motivation
• Introduction of Operators
• Node-Operator
• Advanced Topic: Kube-on-Kube-Operator
• Achievement
• Q&A
Background: DC/OS

From Sigma 2.0(Swarm) to Sigma 3.1(Kubernetes)
Background: Operation Requirements

• Apply to large-scale Cluster
• Setup & Teardown Cluster fast and convenient
• Add & delete Node at any time
• Upgrade Master & Node Components reliably
• Canary Rollout
• Master & Node Component Versions Management
Motivation: Work Order Deployment

Worker Order

- Upgrade Nodes Versions
- Upgrade Node 10.10.10.1
  - Upgrade docker
  - Upgrade kubelet
- Upgrade Node 10.10.10.2
  - Upgrade docker
  - Upgrade kubelet

...
Motivation: Work Order Deployment

Disadvantages

- Inconsistency
- Non-failure-aware
- Complicated architecture

Work order deployment system can not meet the requirements of resource management.
• **Observe**: watch desired resource and actual resource
• **Analyze**: difference from desired and actual config
• **Action**: manage resource to desired config
Operator: Advantages

- Declarative system
  - Manage resource to final state continually
- kube-apiserver oriented programming
  - CustomResourceDefinition (CRD)
- Built on Kubernetes APIs
  - Kubernetes repo support
- Agile, flexible and convenient
Node-Operator: Overview

- **User**: SREs who can scale & offline Nodes through posting Machine CRs.
- **Node-Operator**: difference Machine and Node state, manage Node softwares and configure files.
- **Machine**: the instance of Machine CRD with node basic information, which represent a node desired in the Kubernetes.
- **NPD(Node Problem Detector)**: post Node state to kube-apiserver.
Node-Operator: Scale Nodes

apiVersion: ops.sigma.alipay.com/v1alpha1
kind: MachinePackageVersion
metadata:
  name: kubectl-1.0.1
spec:
  packageName: kubectl
  config:
    rpm: https://kubectl-rpm-url

apiVersion: ops.sigma.alipay.com/v1alpha1
kind: Machine
metadata:
  name: my-node-1
spec:
  ip: 10.10.10.1
  hostname: my-node-1
  machineDriverOptions:
    driver: ssh
    secretName: my-ssh-key
  pkgs:
    - pkgName: kubectl
      pkgVersion: kubectl-1.0.1

status:
  phase: Running
  conditions:
  - type: KubeletOK
    status: "True"
    lastTransitionTime: 2018-10-12T09:05:06Z
    lastUpdateTime: 2018-10-17T09:45:56Z
    message: Package applied successfully
    reason: PackageApplied
  pkgs:
    - pkgName: kubectl
      pkgVersion: kubectl-1.0.1
Node-Operator: Upgrade Nodes

```yaml
apiVersion: ops.sigma.alipay.com/v1alpha1
kind: MachinePackageVersion
metadata:
  name: kubelet-1.0.2
spec:
  machineDriverOptions:
    driver: ssh
  packageName: kubelet
  pkgs:
    pkgName: kubelet
    pkgVersion: kubelet-1.0.2
status:
  conditions:
    - type: KubeletOK
  reason: PackageApplied
  pkgs:
    pkgName: kubelet
    pkgVersion: kubelet-1.0.2
```

Node 10.10.10.1

NPD Pod
docker, CNI, kubelet, Configure Files
Node-Operator: Grayscale Rollout

apiVersion: ops.sigma.alipay.com/v1alpha1
kind: MachinePackageBetaPublish
metadata:
  name: test-beta
spec:
  selector:
    matchLabels:
      is-beta: "true"
  pkgs:
    - packageName: kubelet
      versionName: kubelet-1.0.2
      strategy:
        type: OneByOne
      status:
        phase: Succeeded
        succeeded:
          - my-node-1

Node 10.10.10.1

NPD Pod
docker, CNI, kubelet, Configure Files

apiVersion: ops.sigma.alipay.com/v1alpha1
kind: Machine
metadata:
  name: my-node-1
spec:
  ip: 10.10.1.1
  hostname: my-node-1
  machineDriverOptions:
    driver: ssh
    secretName: my-ssh-key
  pkgs:
    - pkgName: kubelet
      pkgVersion: kubelet-1.0.1
      betaPublishName: test-beta
      status:
        phase: Running
        conditions:
          - type: KubeletOK
            status: "True"
            lastTransitionTime: 2018-10-12T09:05:06Z
            lastUpdateTime: 2018-10-17T09:42:56Z
            message: Package applied successfully
            reason: PackageApplied
          - pkgName: kubelet
            pkgVersion: kubelet-1.0.2
• **Biz-Cluster**: used to deploy our application.

• **Meta-Cluster**: used to set up Biz-Cluster master components. We add Biz-Cluster master nodes to Meta-Cluster.

• **User**: SREs who can setup & upgrade Biz-Cluster by posting Cluster CRs.

• **Kube-on-Kube-Operator**: difference Biz-Cluster CRs and Biz-Cluster master components state, and manage Biz-Cluster master components through Kubernetes resource, such as Deployment, Pod, etc.
Work Together

Biz-Cluster-1 Worker
kubelet, docker, CNI

Biz-Cluster-2 Worker
kubelet, docker, CNI

Meta-Cluster

Biz-Cluster-1 Master
(node-operator)
etcd, kube-apiserver, etc.
kubelet, docker, CNI

Meta-Cluster Master
(kube-on-kube-operator)
etcd, kube-apiserver, etc.
kubelet, docker, CNI

Biz-Cluster-2 Master
(node-operator)
etcd, kube-apiserver, etc.
kubelet, docker, CNI
Achievement

• Anyone can operate and maintenance Kubernetes Cluster
• Set up & tear down Kubernetes Cluster in two Minutes
• Automated rollouts and rollbacks
• Cluster & Node self-healing
THANKS
/感谢聆听

-------- Q&A Section --------

陈俊
WeChat: answer1991chen
Background: Cluster Scale

• Production environment:
  • Dozens of Cluster
  • 5k+ Nodes / Cluster
  • 10k+ Nodes / largest Cluster

• Testing environment
  • Hundreds of Cluster for CI/CD
  • 500+ Nodes / Cluster