AUTOSCALING IN KUBERNETES

Bhavin Gandhi

@_bhavin192

geeksocket.in

WHAT IS AUTOSCALING?

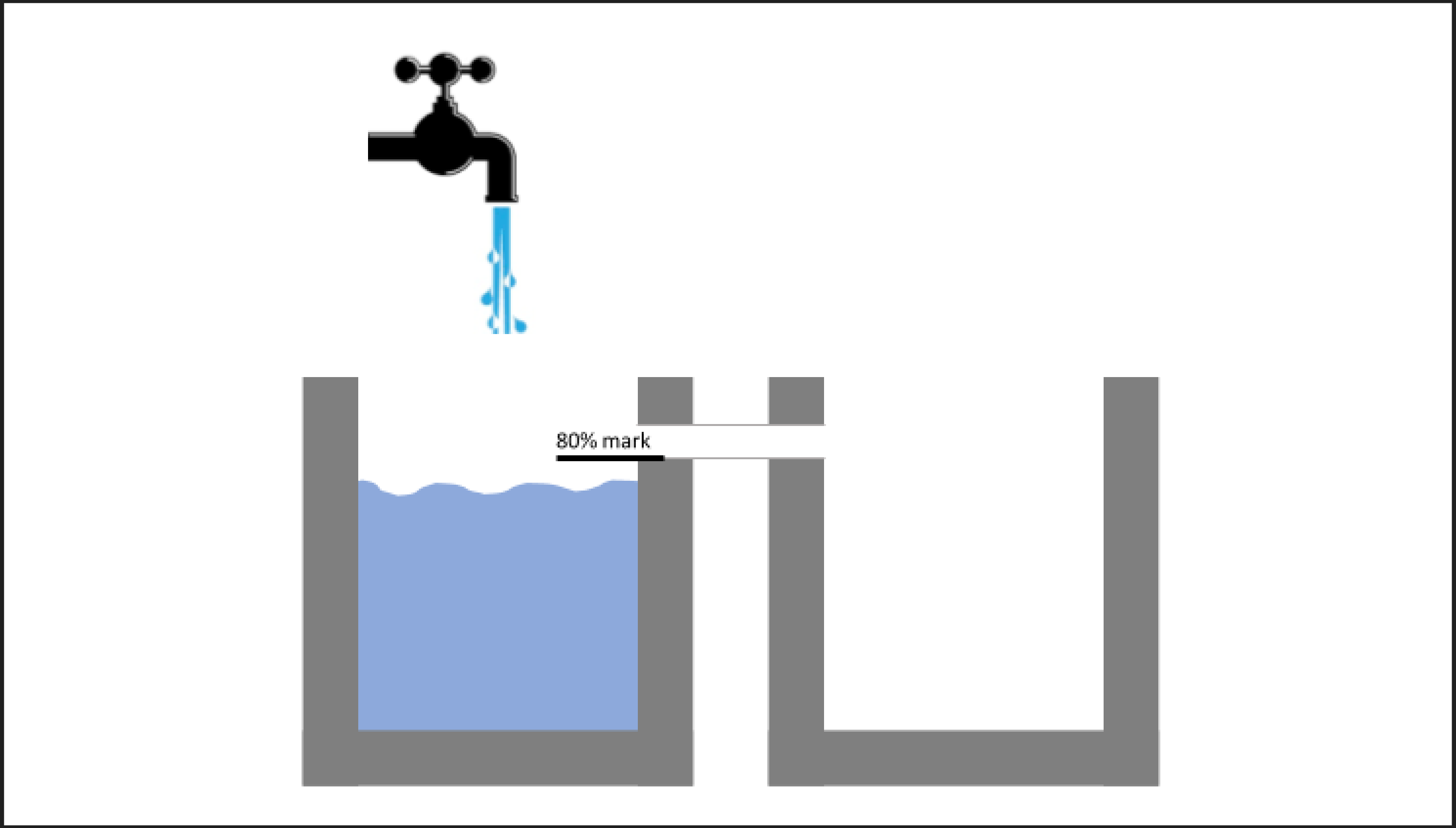

Image Credits: http://blog.infracloud.io/kubernetes-autoscaling-explained

WHY YOU NEED AUTOSCALING?

- Cost saving
- Less downtime

WHAT TO AUTOSCALE

- Application instances (Pods)
- Nodes

WHEN TO AUTOSCALE

- What are metrics

HOW TO AUTOSCALE

AUTOSCALING NODES

- Cluster Autoscaler
- https://git.k8s.io/autoscaler/cluster-autoscaler

AUTOSCALING PODS

- Horizontally
- Vertically

HORIZONTALPODAUTOSCALER

- Controller loop
  - Looks for certain metric values and takes decisions based on those values
- Fetches metrics values from the end point
  - `metrics.k8s.io`
  - `custom.metrics.k8s.io`
  - `external.metrics.k8s.io`

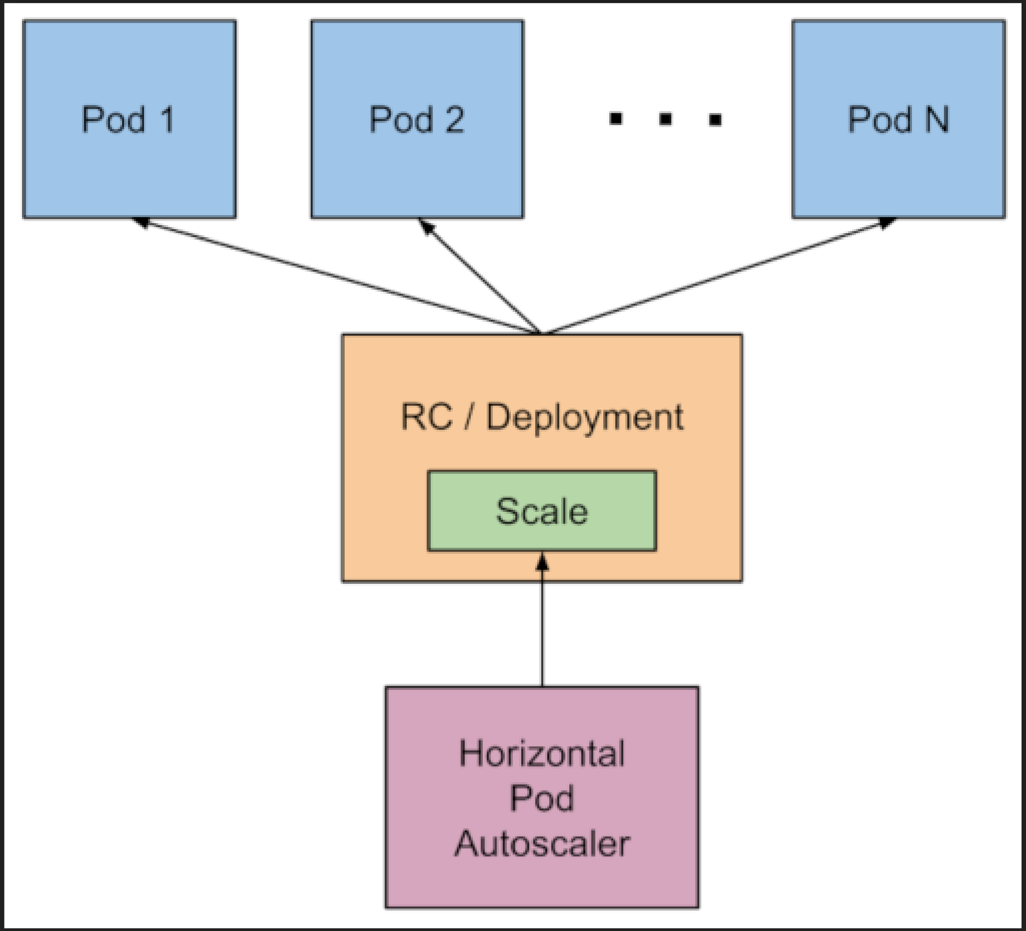

Image Credits: https://git.k8s.io/website (CC BY 4.0)

AUTOSCALING BASED ON CPU/RAM

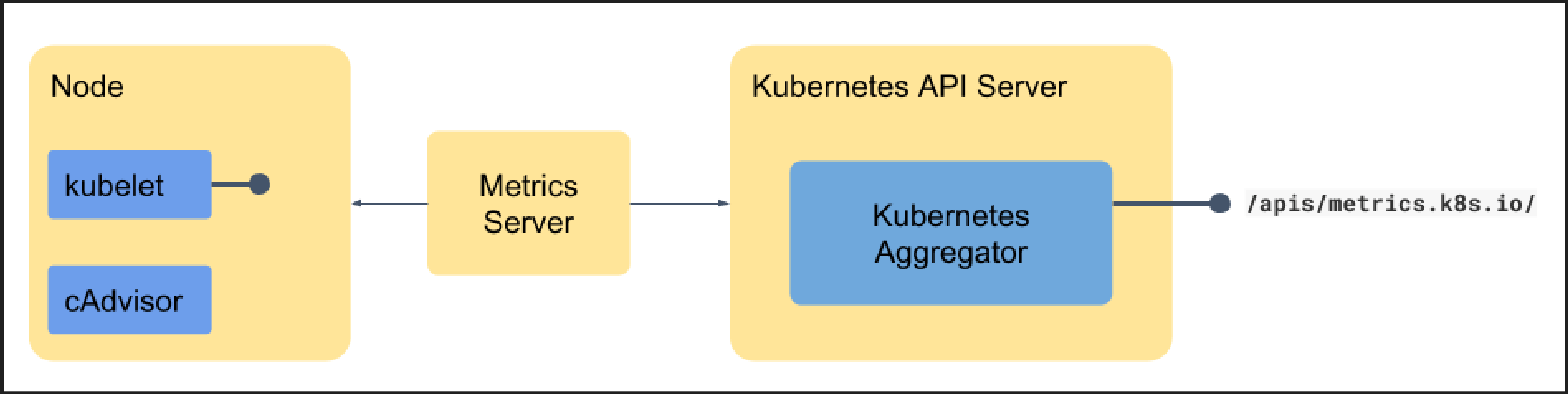

Image Credits: http://blog.infracloud.io/kubernetes-

Speaker notes

- Thist is part of `autoscaling/v1` which was introduced in v1.2 of Kubernetes
- Needs resource limits and requests set in order to calculate the usage etc

DEMO

```
# Create a deployment
$ kubectl run stress-deploy \
  --image=monitoringartist/docker-killer:latest \
  --limits="cpu=200m,memory=512Mi" \
  --requests="cpu=200m,memory=512Mi" \
  --env="TIMEOUT=10000" -- cpubomb

# Autoscale it based on CPU usage
$ kubectl autoscale deployment \
  --max="100" --min="3" \
  --cpu-percent="80" \
  stress-deploy
```

CHECK THE STATUS

```
# Watch HorizontalPodAutoscaler
$ kubectl get hpa -w

# Watch Pods
$ kubectl get pods -w

# Get list of Nodes
$ kubectl get nodes
```

HPA SPECIFICATION

```yaml
apiVersion: autoscaling/v2beta1
kind: HorizontalPodAutoscaler
metadata:
  name: stress-deploy
spec:
  scaleTargetRef:
    apiVersion: apps/v1beta1
    kind: Deployment
    name: stress-deploy
  minReplicas: 3
  maxReplicas: 100
  metrics:
  - type: Resource
    resource:
      name: cpu
      targetAverageUtilization: 80
```

AUTOSCALING BASED ON CUSTOM METRICS

Speaker notes

This is part of `autoscaling/v2beta1` introduced in v1.6 of Kubernetes

USING PROMETHEUS AND PROMETHEUS ADAPTER

- Kubernetes Autoscaling with Custom Metrics
- http://blog.infracloud.io/kubernetes-autoscaling-custom-metrics/

USING DATADOG'S CLUSTER AGENT

- Autoscale your Kubernetes workloads with any Datadog metric
- https://www.datadoghq.com/blog/autoscale-kubernetes-datadog/

WHAT'S NEXT

- http://blog.infracloud.io/kubernetes-autoscaling-explained/
- https://youtu.be/YWLrvj3XOD0
- https://kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale/

THANK YOU