☰ Kubernetes Part 3 (Scaling) 2 of 9

Horizontal pod accelerator

Ref - https://kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale/

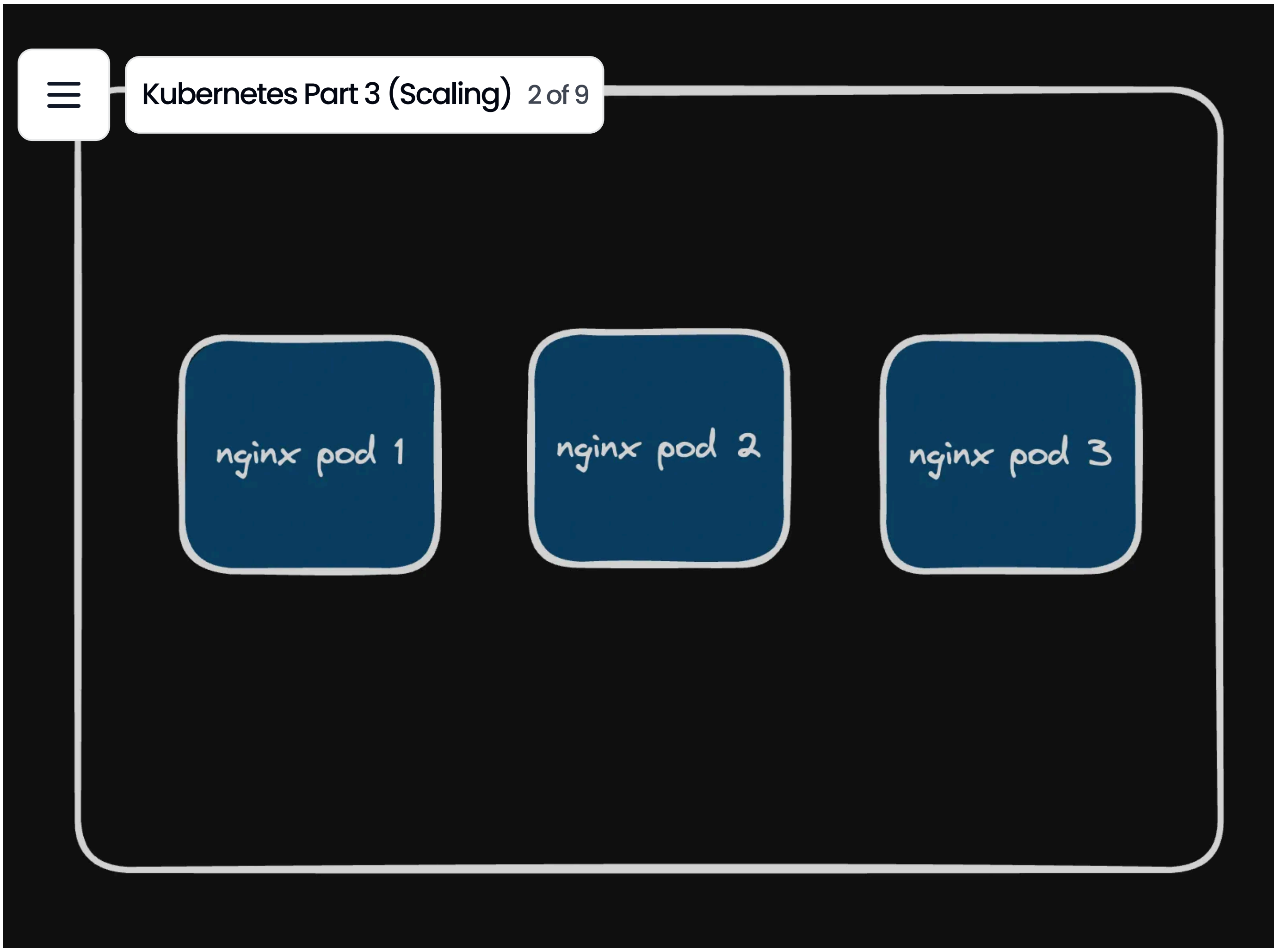
A Horizontal Pod Autoscaler (HPA) is a Kubernetes feature that automatically adjusts the number of pod replicas in a deployment, replica set, or stateful set based on observed metrics like CPU utilisation or custom metrics.

This helps ensure that the application can handle varying loads by scaling out (adding more pod replicas) when demand increases and scaling in (reducing the number of pod replicas) when demand decreases.

Horizontal scaling

As the name suggests, if you add more pods to your cluster, it means scaling `horizontally` . Horizontally refers to the fact that you havent increased the `resources` on the machine.

↗ ↑



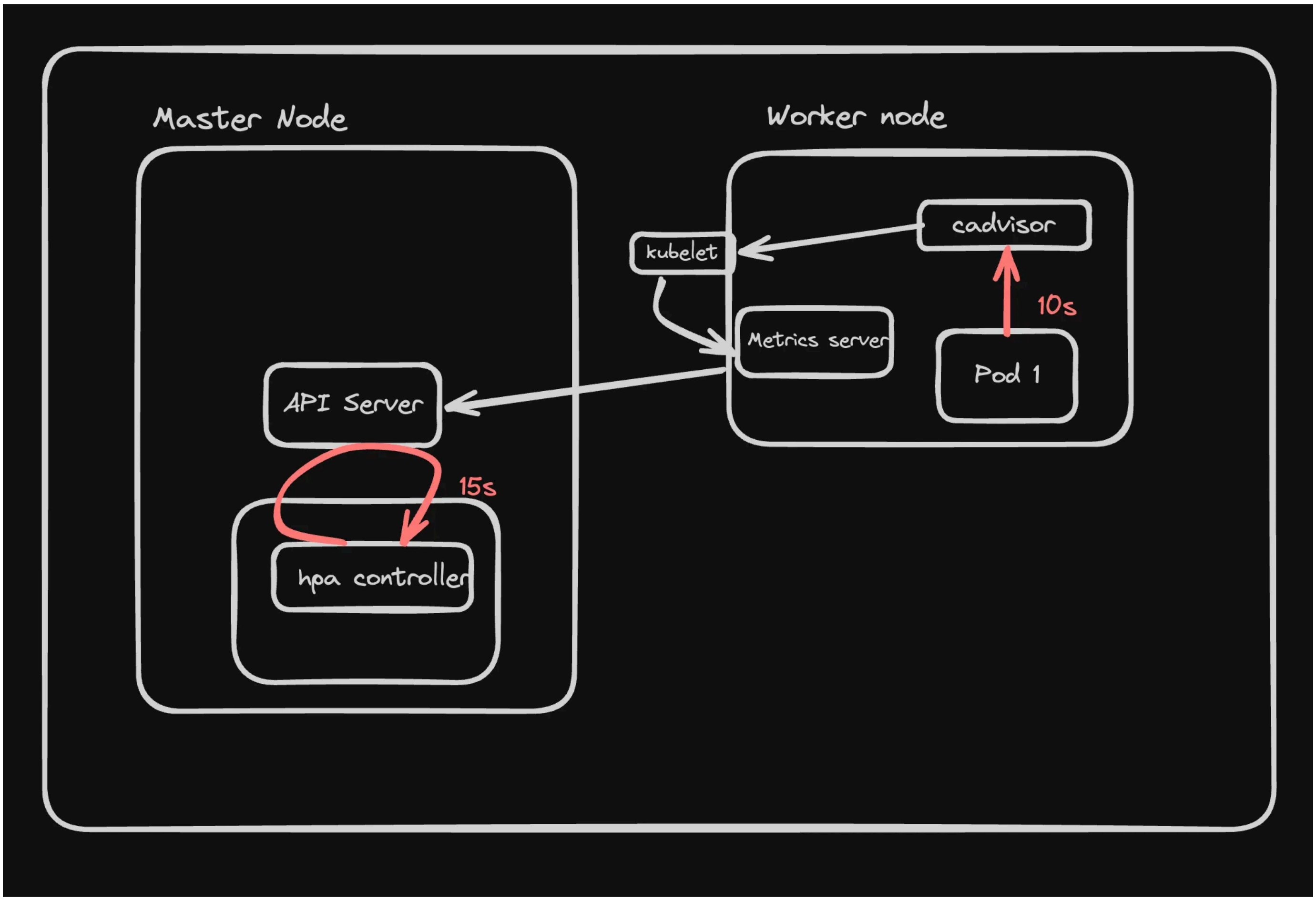
Architecture

Kubernetes implements horizontal pod autoscaling as a control loop that runs intermittently (it is not a continuous process) (once every 15s)

- cadvisor - https://github.com/google/cadvisor
- Metrics server - The Metrics Server is a lightweight, in-memory store for metrics. It collects resource usage metrics (such as CPU and memory) from the kubelets and exposes them via the Kubernetes API (Ref - https://github.com/kubernetes-sigs/metrics-server/issues/237)

```
kubectl apply -f https://github.com/kubernetes-sigs/metrics-server/releases/
or
Apply from here - https://github.com/100xdevs-cohort-2/week-28-manifests
```

```
kubectl top pod -n kube-system
```



Sample request that goes from hpa controller to the API server

```
GET https://338eb37e-2824-4089-8eee-5a05f84fb85e.vultr-k8s.com:6443/ap
```