Why use a Horizontal Pod Autoscaler?

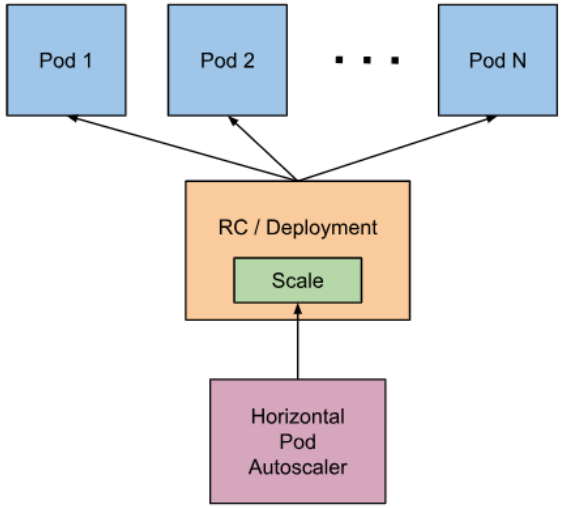
As the name suggests, this component would scale your application automatically. In the cloud, this can really help you reduce the compute and memory resources you will be billed for. Since the Autoscaler is sensitive to the resource utilization, when it sees that a lot of pods are just sitting idle it scales the application down and when the demand on those pods increases it scales the application up by creating new pods and the load gets distributed to those.

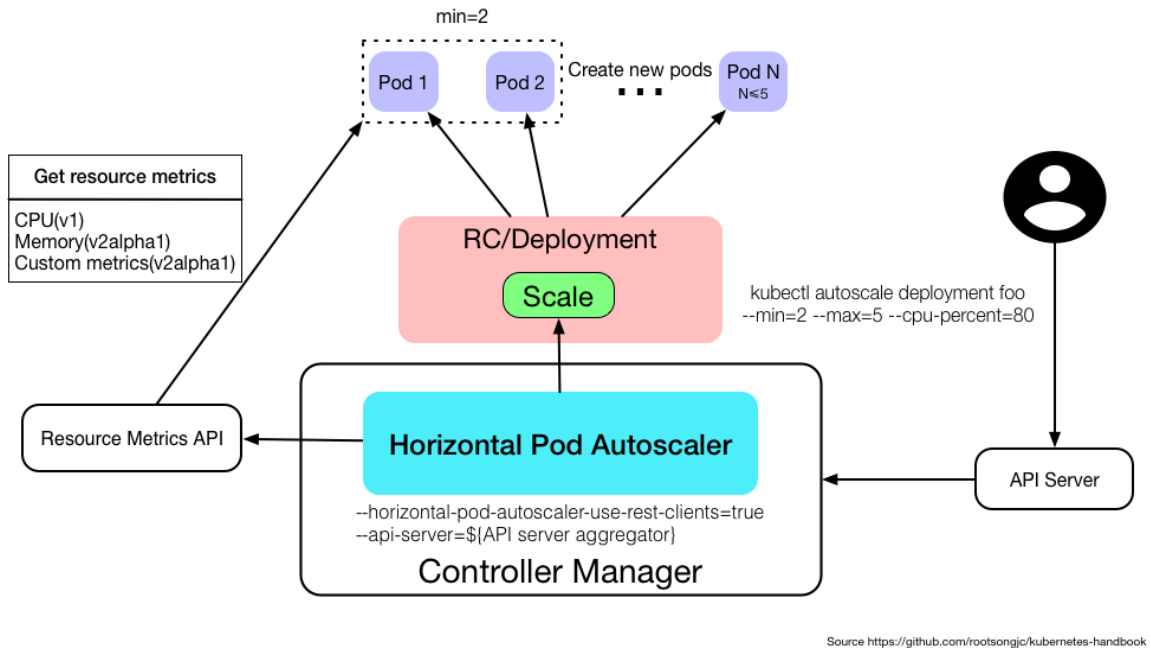
The Horizontal Pod Autoscaler automatically scales the number of pods in a replication controller, deployment or replica set based on observed CPU utilization (support, on some other application-provided metrics). Note that Horizontal Pod Autoscaling does not apply to objects that can’t be scaled, for example, DaemonSets.

Autoscaling in Kubernetes

- Autoscaling is an approach to automatically scale workloads up or down based on resource usage.
- Autoscaling in Kubernetes has two dimensions:
- the **Cluster Autoscaler** that deals with node scaling operations and
- the **Horizontal Pod Autoscaler (HPA)** that automatically scales the number of pods in a deployment or replica set.
- The Cluster Autoscaling together with the Horizontal Pod Autoscaler (HPA) can be used to dynamically adjust the computing power as well as the level of parallelism that your system needs to meet SLAs.
- While the Cluster Autoscaler is highly dependent on the underlying capabilities of the cloud provider that's hosting your cluster, the HPA can operate independently of your IaaS/PaaS provider.

Horizontal Pod Autoscaling





Source https://github.com/rootsongjc/kubernetes-handbook

hpa

Demo


THANK
YOU!