API OBJECTS

API Objects

Autoscaling

In the autoscaling group we find the **Horizontal Pod Autoscalers** (**HPA**). This is a stable resource. HPAs automatically scale Replication Controllers, ReplicaSets, or Deployments based on a target of 50% CPU usage by default. The usage is checked by the kubelet every 30 seconds, and retrieved by the Metrics Server API call every minute. HPA checks with the Metrics Server every 30 seconds. Should a Pod be added or removed, HPA waits 180 seconds before further action.

Other metrics can be used and queried via REST. The autoscaler does not collect the metrics, it only makes a request for the aggregated information and increases or decreases the number of replicas to match the configuration.

The **Cluster Autoscaler** (**CA**) adds or removes nodes to the cluster, based on the inability to deploy a Pod or having nodes with low utilization for at least 10 minutes. This allows dynamic requests of resources from the cloud provider and minimizes expenses for unused nodes. If you are using CA, nodes should be added and removed through **`cluster-autoscaler-`** commands. Scale-up and down of nodes is checked every 10 seconds, but decisions are made on a node every 10 minutes. Should a scale-down fail, the group will be rechecked in 3 minutes, with the failing node being eligible in five minutes. The total time to allocate a new node is largely dependent on the cloud provider.

Another project still under development is the Vertical Pod Autoscaler. This component will adjust the amount of CPU and memory requested by Pods.