Cluster Proportional Autoscaler
CPA

OW

Overview

- Unlike traditional autoscalers that rely on metrics like CPU and memory usage, CPA scales your app replicas directly based on the number of nodes or cores in your cluster.
- CPA is a horizontal pod autoscaler that scales replicas based on the number of nodes in a cluster. The proportional autoscaler container watches over the number of schedulable nodes and cores of a cluster and resizes the number of replicas.

Linear Scaling:

- This scaling method will scale the application in direct proportion to the number of nodes or cores available in the cluster. Either the coresPerReplica or nodesPerReplica parameter can be omitted.
- When preventSinglePointFailure is set to true, the controller ensures at least 2 replicas if there are more than one node.
- When includeUnschedulableNodes is set to true, the replicas will be scaled based on the total number of nodes, otherwise they will only scale based on the number of schedulable nodes.

Ladder scaling

- This scaling method uses a step function to determine the ratio of nodes:replicas and/or cores:replicas. The step ladder function uses the data points for core and node scaling from the ConfigMap, and the lookup that yields the higher number of replicas will be used as the target scaling number.
- Either the coresPerReplica or nodesPerReplica parameter can be omitted. Replicas can also be set to 0, unlike in linear mode, which can be used to enable optional features as the cluster grows.

CPA vs HPA

- The key differences between CPA and the Horizontal Pod Autoscaler HPA are:
- CPA scales based on the number of nodes/cores in the cluster, while HPA scales based on CPU/memory utilization.
- CPA does not rely on the Metrics API or Metrics Server, unlike HPA.
- CPA uses a simple control loop to watch cluster size and scale the target workload, while HPA is a Kubernetes API resource.
- CPA is well-suited for cluster services that need to scale with the overall cluster size, while HPA is more general-purpose.