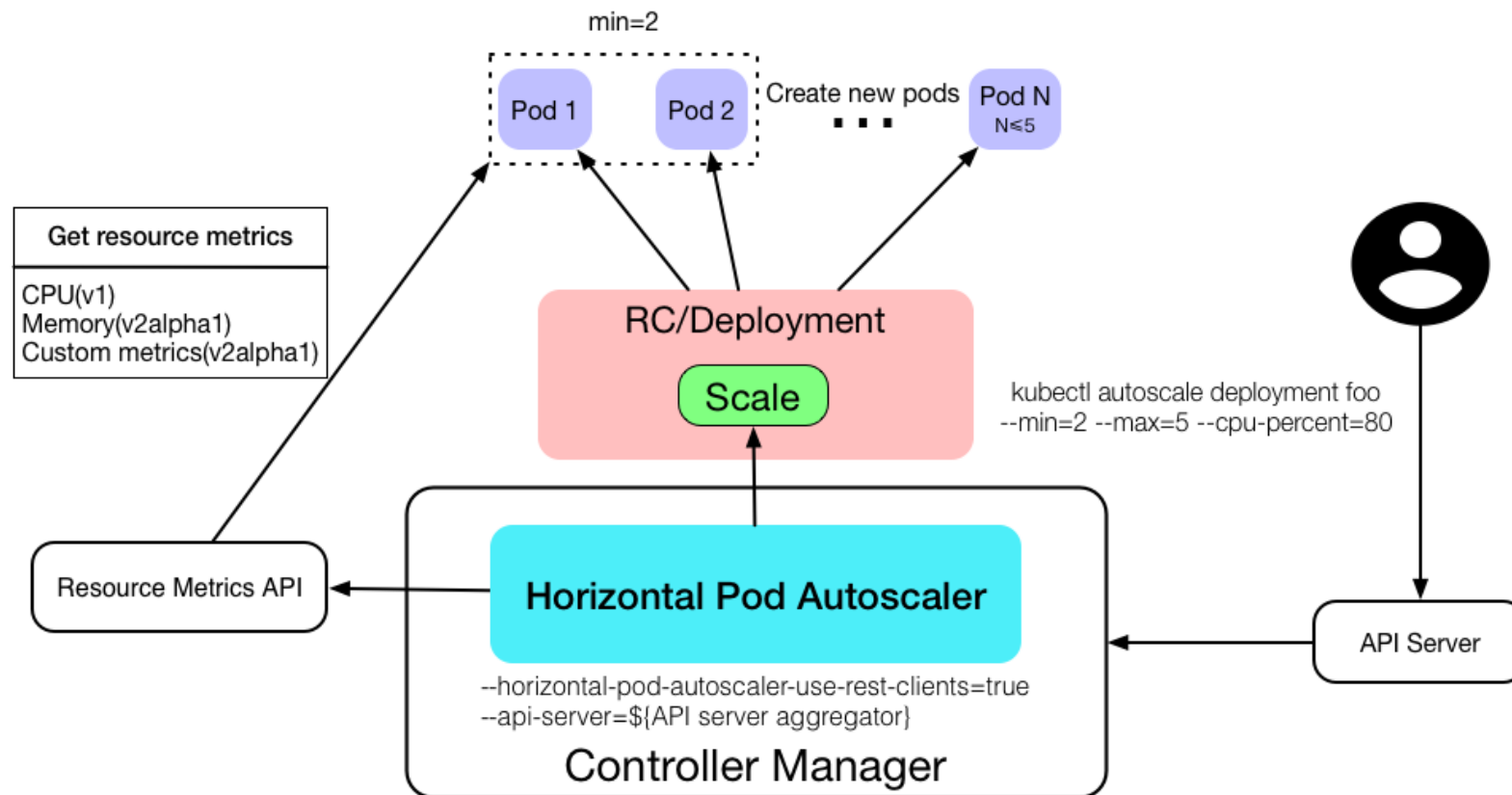
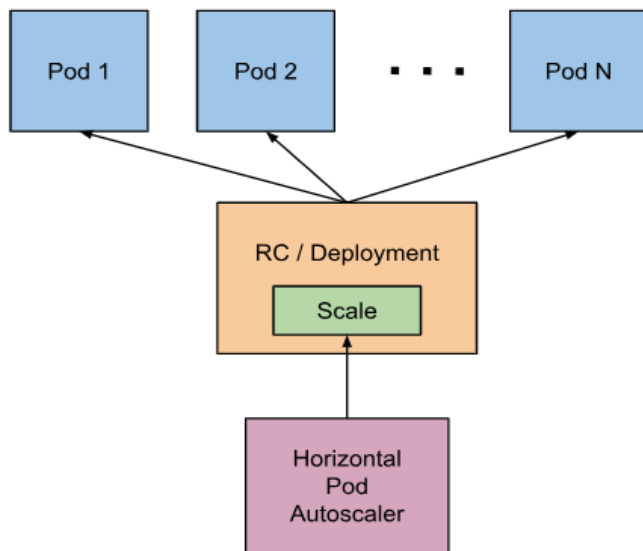


Why use a Horizontal Pod Autoscaler?

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 (or, with [custom metrics](#) 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.

Horizontal Pod Autoscaling



Demo