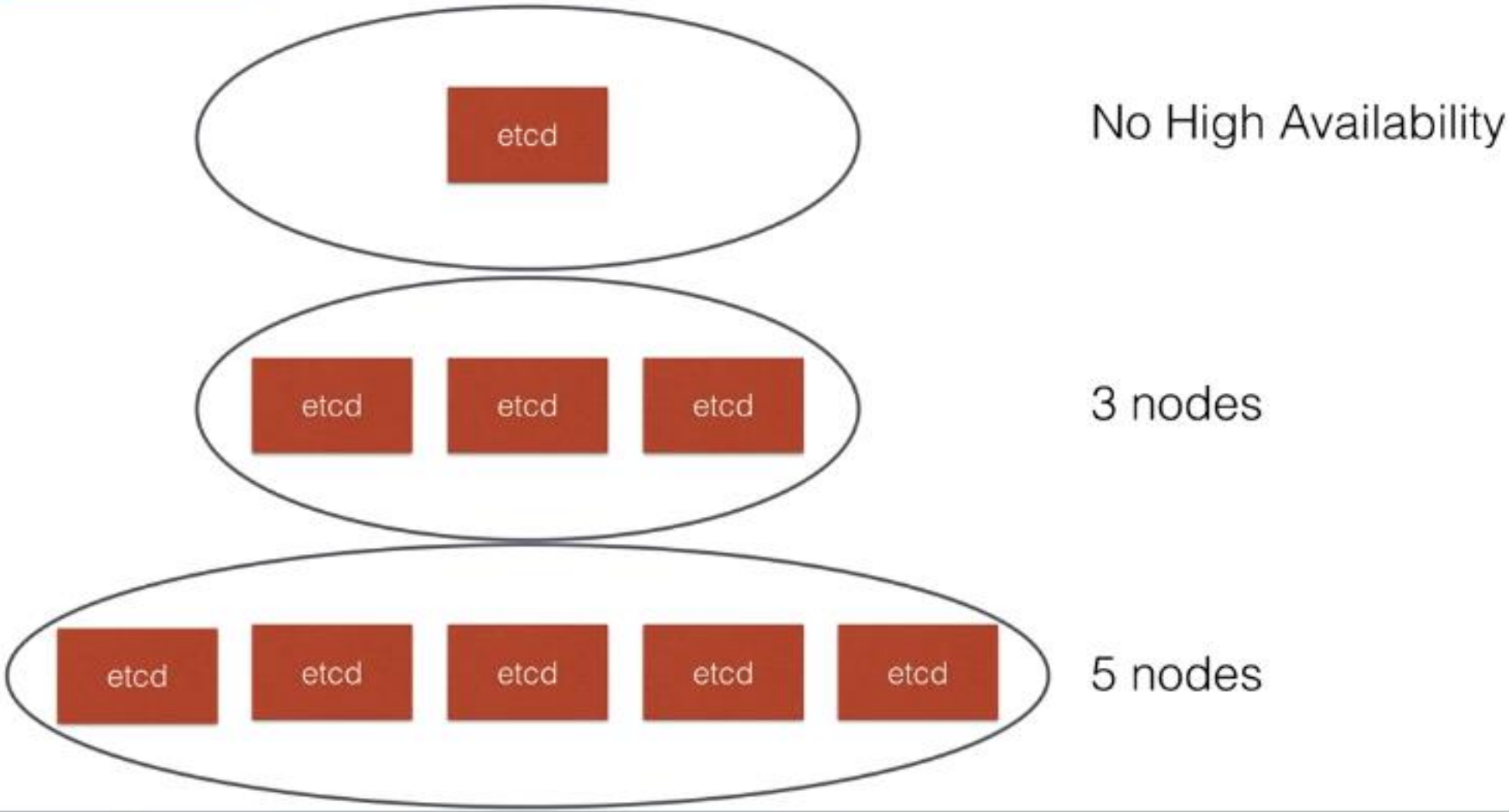
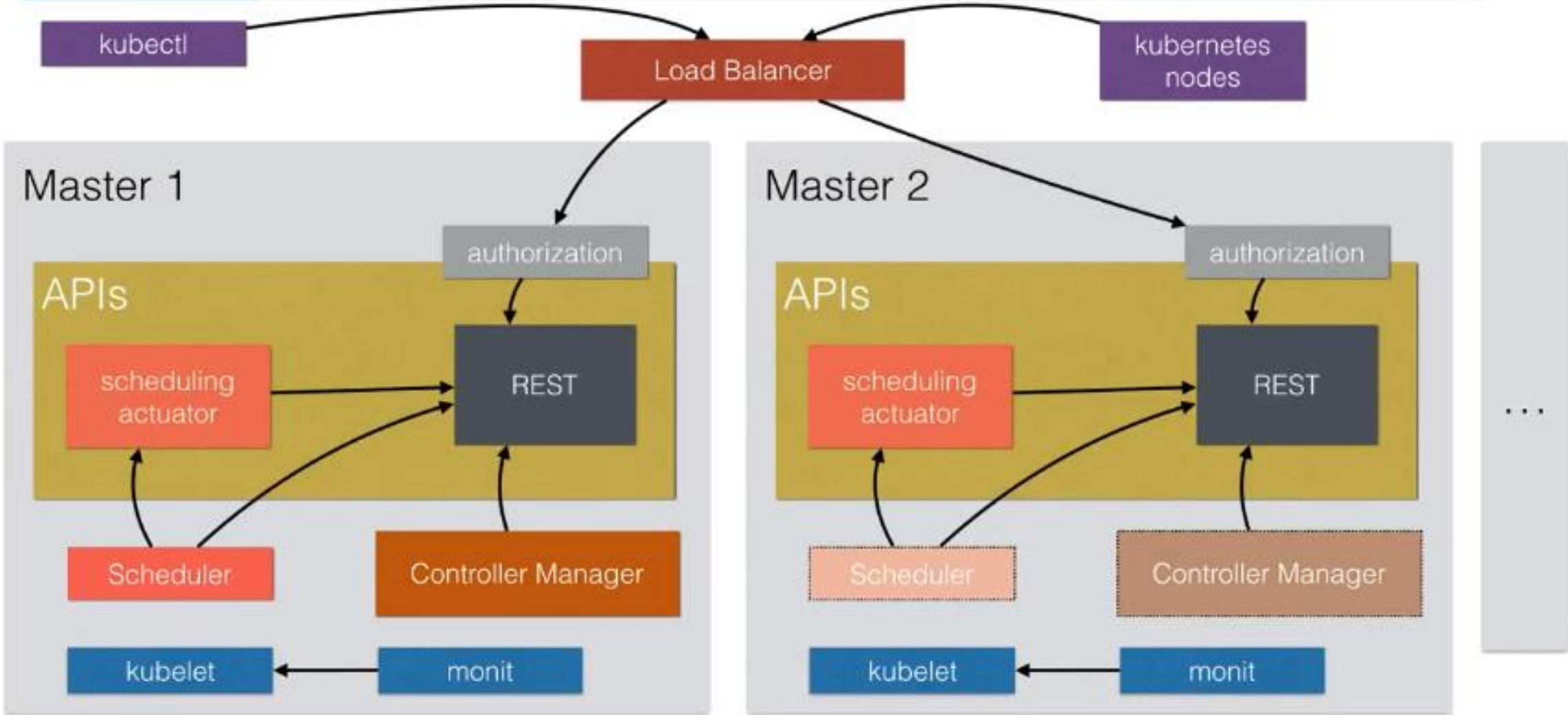
High Availability

- If you're going to run your cluster in production, you're going to want to have all your master services in a **high availability (HA)** setup
- The setup looks like this:
  - **Clustering etcd**: at least run 3 etcd nodes
  - **Replicated API servers** with a LoadBalancer
  - Running multiple instances of the **scheduler** and the **controllers**
    - Only one of them will be the leader, the other ones are on stand-by

Architecture overview - HA



Architecture overview - HA



High Availability

- A cluster like minikube doesn't need HA - it's only a one node cluster
- If you're going to use a production cluster on AWS, **kops** can do the heavy lifting for you
- If you're running on an other cloud platform, have a look at the **kube deployment tools** for that platform
    - **kubeadm** is a tool that is in alpha that can set up a cluster for you
- If you're on a platform without any tooling, have a look at http://kubernetes.io/docs/admin/high-availability/ to implement it yourself