Figure 24-3 summarizes how the CA interacts with cloud providers and Kubernetes for scaling out cluster nodes.

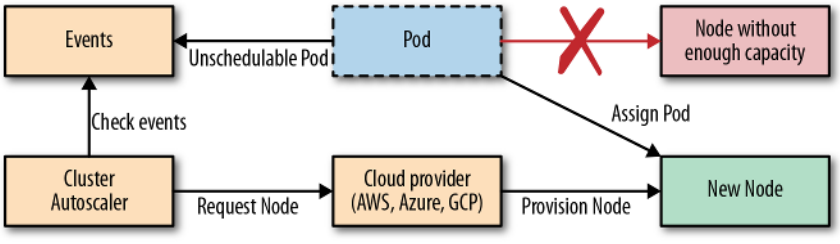

Figure 24-3. Cluster autoscaling mechanism

As you probably figured out by now, scaling Pods and nodes are decoupled but complementary procedures. An HPA or VPA can analyze usage metrics, events, and scale Pods. If the cluster capacity is insufficient, the CA kicks in and increases the capacity. The CA is also helpful when there are irregularities in the cluster load due to batch Jobs, recurring tasks, continuous integration tests, or other peak tasks that require a temporary increase in the capacity. It can increase and reduce capacity and provide significant savings on cloud infrastructure costs.

Scaling Levels

In this chapter, we explored various techniques for scaling deployed workloads to meet their changing resource needs. While a human operator can manually perform most of the activities listed here, that doesn't align with the cloud-native mindset. In order to enable large-scale distributed system management, the automation of repetitive activities is a must. The preferred approach is to automate scaling and enable human operators to focus on tasks that a Kubernetes *operator* cannot automate yet.

Let's review all of the scaling techniques, from the more granular to the more coarse-grained order as shown in Figure 24-4.