Resource requests and limits

Ref - https://kubernetes.io/docs/concepts/configuration/manage-resources-containers/

When you specify a <u>Pod</u>, you can optionally specify how much of each resource a <u>container</u> needs. The most common resources to specify are CPU and memory (RAM).

There are two types of resource types

Resource requests

The kubelet reserves at least the *request* amount of that system resource specifically for that container to use.

Resource limits

When you specify a resource *limit* for a container, the <u>kubelet</u> enforces those limits so that the running container is not allowed to use more of that resource than the limit you set.

Difference b/w limits and requests

If the node where a Pod is running has enough of a resource available, it's possible (and allowed) for a container to use more resource than its request for that resource specifies. However, a container is not allowed to use more than its resource limit.

Experiments

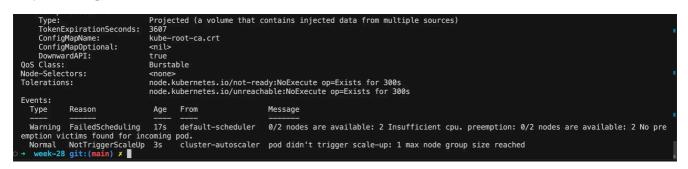
▼ 30% CPU usage on a single threaded Node.js	app
--	-----

A though this is a Node.js app where

it can go up to 100%

```
apiVersion: apps/vl
kind: Deployment
metadata:
name: cpu-deployment
spec:
replicas: 2
selector:
  matchLabels:
   app: cpu-app
template:
  metadata:
  labels:
    app: cpu-app
  spec:
   containers:
   - name: cpu-app
    image: 100xdevs/week-28:latest
    ports:
    - containerPort: 3000
    resources:
     requests:
      cpu: "100m"
     limits:
      cpu: "300m"
```

Try hitting the server



▼ Request 2 vCPU in 10 replicas

Try requesting more resources than available in the cluster.



```
replicas: 10
selector:
 matchLabels:
  app: cpu-app
template:
 metadata:
  labels:
   app: cpu-app
 spec:
  containers:
  - name: cpu-app
   image: 100xdevs/week-28:latest
   ports:
   - containerPort: 3000
   resources:
    requests:
     cpu: "1000m"
    limits:
     cpu: "1000m"
```