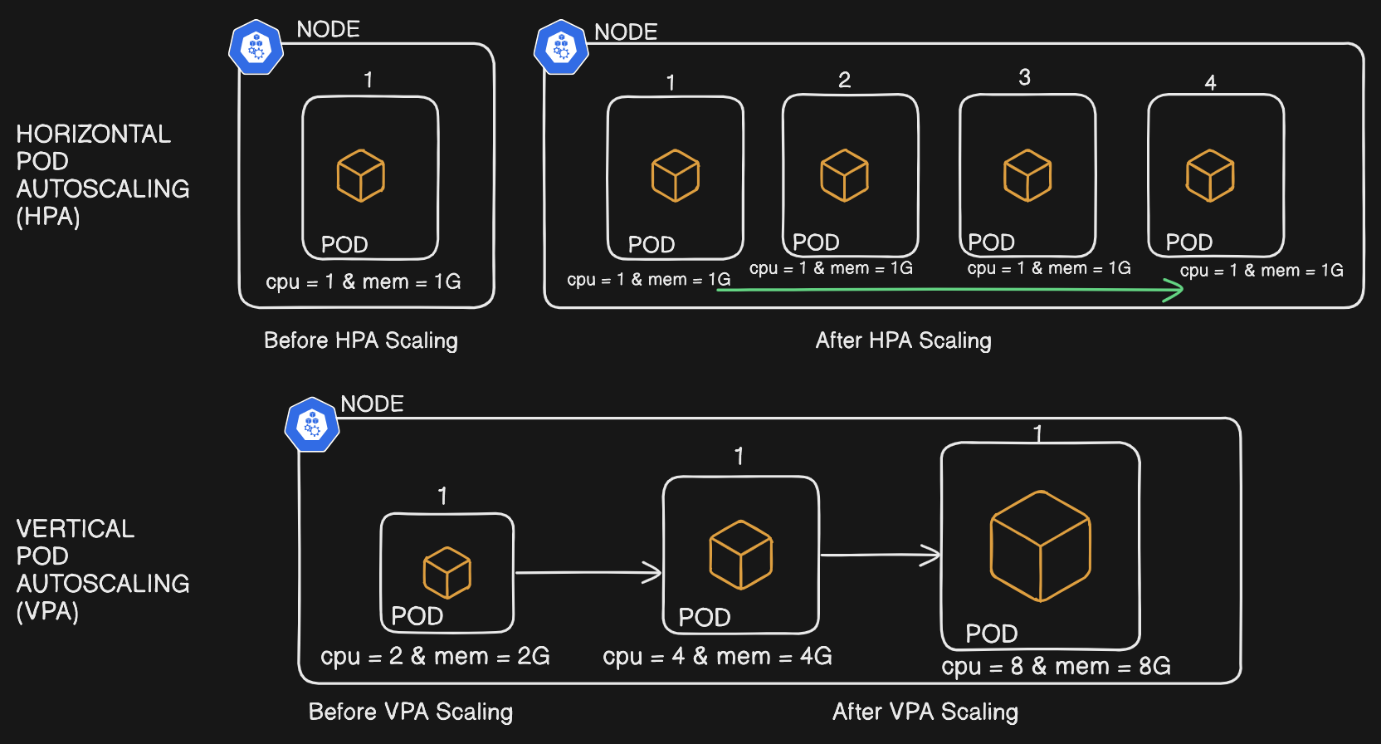
**HPA (Horizontal Pod Autoscaler)** and **VPA (Vertical Pod Autoscaler)** are mechanisms that manage the scaling of resources for applications running within the cluster.



**1. HPA - Horizontal Pod Autoscaler**

* **Purpose**: HPA automatically adjusts the number of pod replicas in a deployment, replicaset, or statefulset based on observed CPU, memory usage, or other custom metrics.
* **Usage**: Best suited for stateless applications where multiple instances of the application can run simultaneously without data synchronization issues.
* **When to Use**:
  + When you expect varying loads and want to handle traffic surges.
  + For applications where scaling horizontally (adding more pods) is easier than increasing resources of a single pod.
* **Advantages**:
  + Increased resilience and load balancing since multiple pods can share the workload.
  + Ideal for applications with distributed or stateless architectures.

**2. VPA - Vertical Pod Autoscaler**

* **Purpose**: VPA automatically adjusts the resource requests and limits (CPU, memory) of existing pods based on observed usage patterns, thereby “scaling up” the resource capabilities of each pod individually.
* **Usage**: Ideal for stateful applications where a single instance with more resources is more efficient than multiple replicas.
* **When to Use**:
  + For applications where horizontal scaling isn't feasible, such as those requiring shared storage or maintaining a unique state.
  + For workloads that require more memory or CPU over time but don't need additional replicas.
* **Advantages**:
  + Can help reduce resource underutilization by right-sizing pods.
  + Prevents resource overcommitment by adjusting limits based on actual usage patterns.

**Which One is Better: HPA or VPA?**

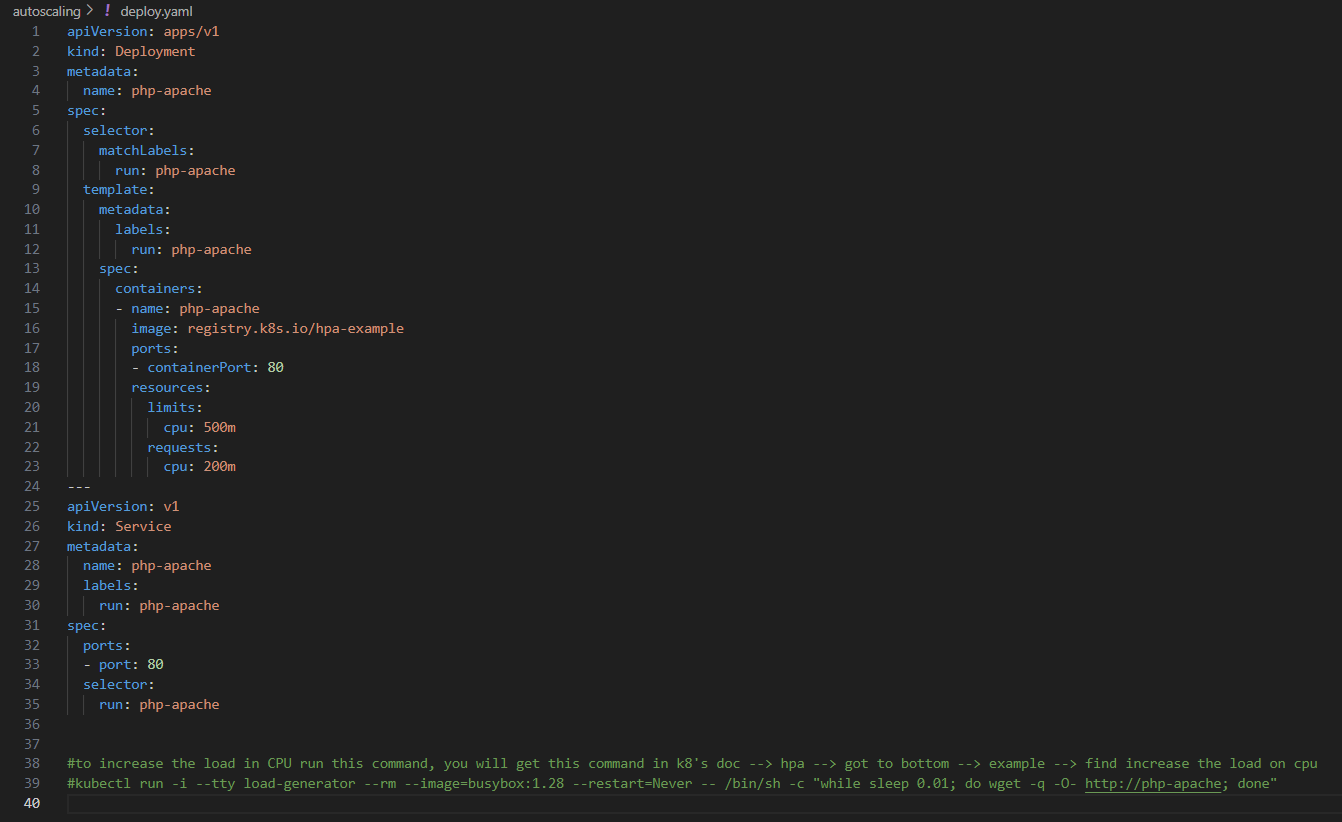
* **Use Cases**: HPA is generally better for **stateless** and **scalable applications**, while VPA is beneficial for stateful applications or those that benefit from increasing the power of individual instances rather than adding new ones.
* **Combined Approach**: Often, HPA and VPA can be used together. HPA handles the number of replicas, while VPA ensures each pod has the correct resources. However, their combined use should be carefully tested, as scaling horizontally while also changing individual pod sizes can introduce complexity.

**Summary**

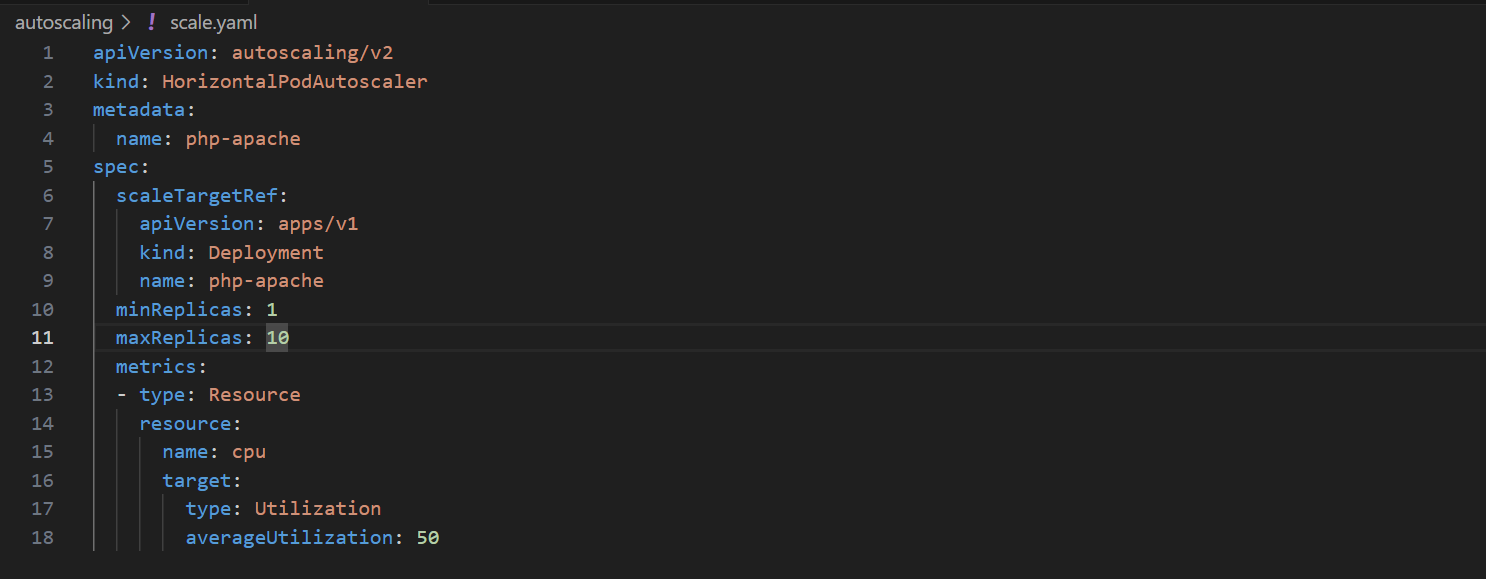
* Use **HPA** for workloads with fluctuating traffic and load that can be distributed across multiple instances.
* Use **VPA** for applications that benefit from increased resources per pod rather than additional replicas.

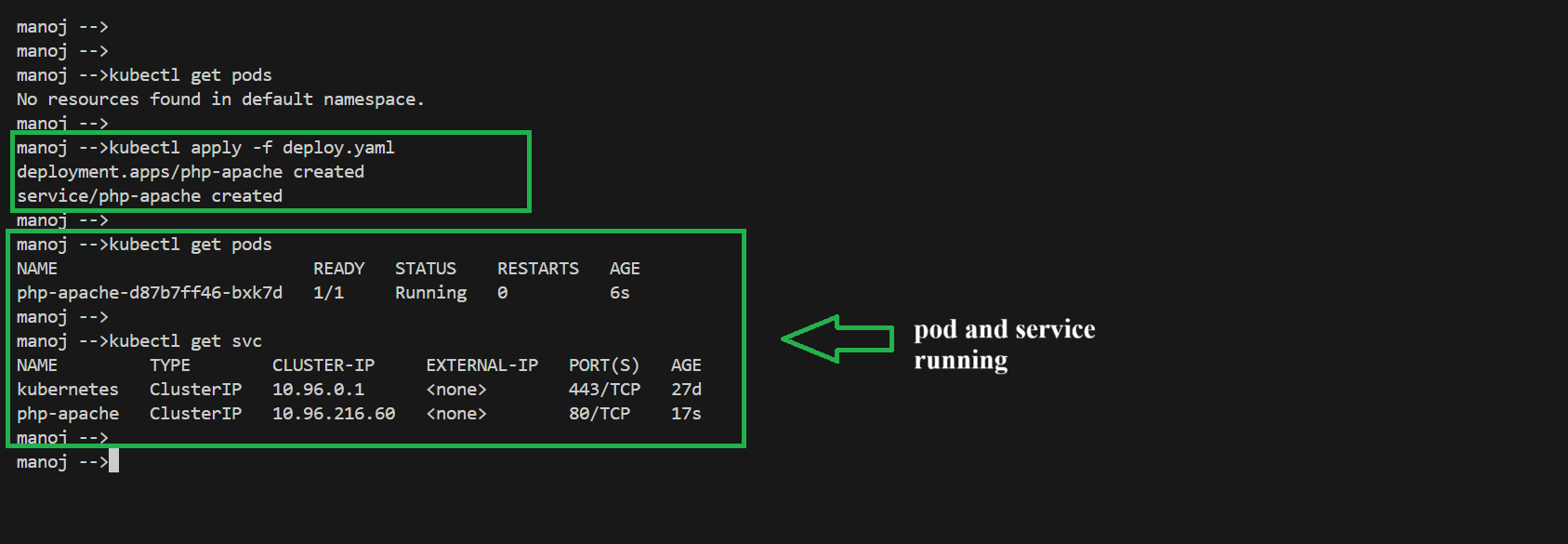
**Hands-on (HPA)**

Deployment and Service manifest

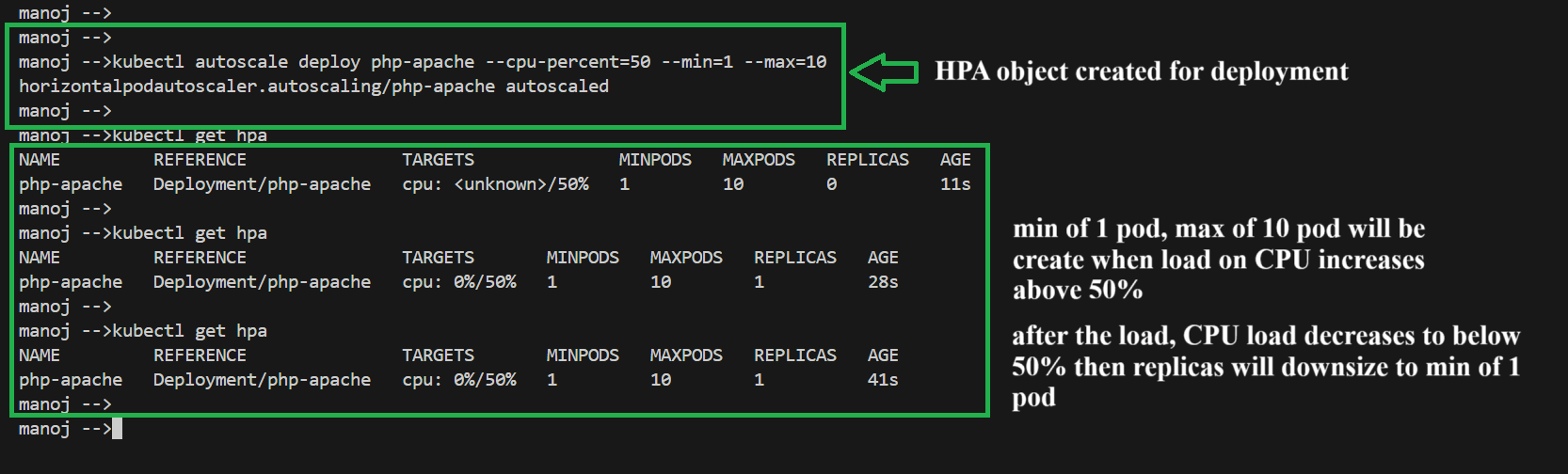
****

Horizontal pod autoscaling manifest, where ever the CPU utilization cross 50% new pod will be created.

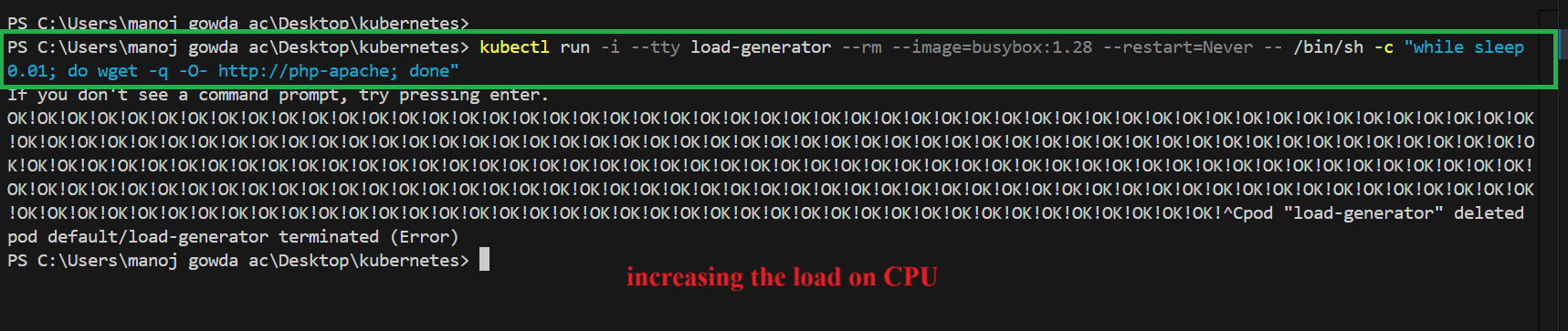
****

****

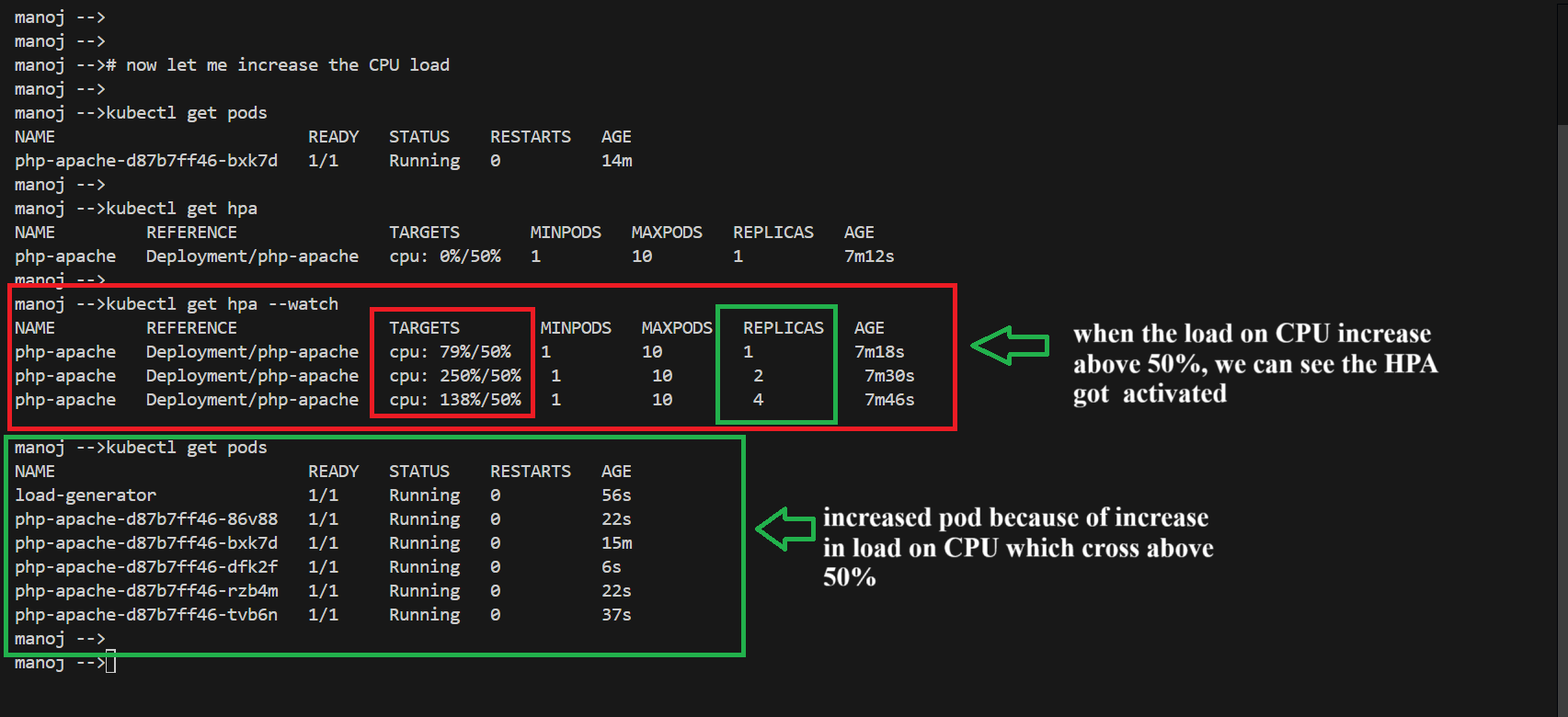
HPA Object created



Applying stress to increase the load on CPU

****

We can see when CPU utilization crossed 50% new pod got created as we specified in HPA manifest.

****