**02. Install VPA Components & Sample Demo Application**

--- Reference - <https://github.com/stacksimplify/aws-eks-kubernetes-masterclass/tree/master/16-EKS-VPA-Vertical-Pod-Autoscaler>

**Introduction**

* The Kubernetes Vertical Pod Autoscaler automatically adjusts the CPU and memory reservations for your pods to help "right size" your applications.
* This adjustment can improve cluster resource utilization and free up CPU and memory for other pods.

**Pre-requisite - Metric server**

* Metric server should be installed & ready, we have done that as part of HPA

--- **kubectl get pods -n kube-system**



--- you can see that the metric server is up and running.

**Deploy the Vertical Pod Autoscaler (VPA)**

**# Clone Repo**

--- **git clone https://github.com/kubernetes/autoscaler.git**

**# Navigate to VPA**

--- **cd autoscaler/vertical-pod-autoscaler/**

**# Uninstall VPA (if we are using old one)**

--- **./hack/vpa-down.sh**

**# Install new version of VPA**

--- **./hack/vpa-up.sh**

**# Verify VPA Pods**

--- **kubectl get pods -n kube-system**



--- **note** - 3 VPA pods are running as part of the vertical auto scalar installation.

**Review & Deploy our Application Manifests (Deployment & Service)**

--- Make a note of resources we have defined spec.containers.resources.requests.

--- limits we define in VPA definition

        resources:

          requests:

            cpu: "5m"

            memory: "5Mi"

--- **01-VPA-DemoApplication.yml**

apiVersion: apps/v1

kind: Deployment

metadata:

  name: vpa-demo-deployment

  labels:

    app: vpa-nginx

spec:

  replicas: 4

  selector:

    matchLabels:

      app: vpa-nginx

  template:

    metadata:

      labels:

        app: vpa-nginx

    spec:

      containers:

      - name: vpa-nginx

        image: stacksimplify/kubenginx:1.0.0

        ports:

        - containerPort: 80

        resources:

          requests:

            cpu: "5m"

            memory: "5Mi"

---

apiVersion: v1

kind: Service

metadata:

  name: vpa-demo-service-nginx

  labels:

    app: vpa-nginx

spec:

  type: NodePort

  selector:

    app: vpa-nginx

  ports:

  - port: 80

    targetPort: 80

    nodePort: 31232

--- **02-VPA-Manifest.yml**

apiVersion: "autoscaling.k8s.io/v1beta2"

kind: VerticalPodAutoscaler

metadata:

  name: kubengix-vpa

spec:

  targetRef:

    apiVersion: "apps/v1"

    kind: Deployment

    name: vpa-demo-deployment

  resourcePolicy:

    containerPolicies:

      - containerName: '\*'

        minAllowed:

          cpu: 5m

          memory: 5Mi

        maxAllowed:

          cpu: 1

          memory: 500Mi

        controlledResources: ["cpu", "memory"]

**Deploy**

**# Deploy Application**

--- **kubectl apply -f kube-manifests/01-VPA-DemoApplication.yml**

**# List Pods, Deploy & Service**

--- **kubectl get pod,svc,deploy**



**# Describe Pod**

--- **kubectl describe pod <pod-name>**

**# Access Application (If our NodeGroup is in Public Subnet, if not ignore)**

--- **kubectl get nodes -o wide**

--- **http://<Worker-Node-Public-IP>:31232**