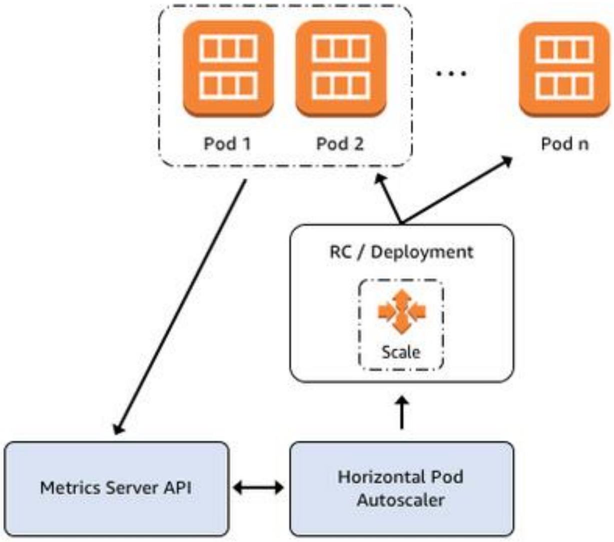
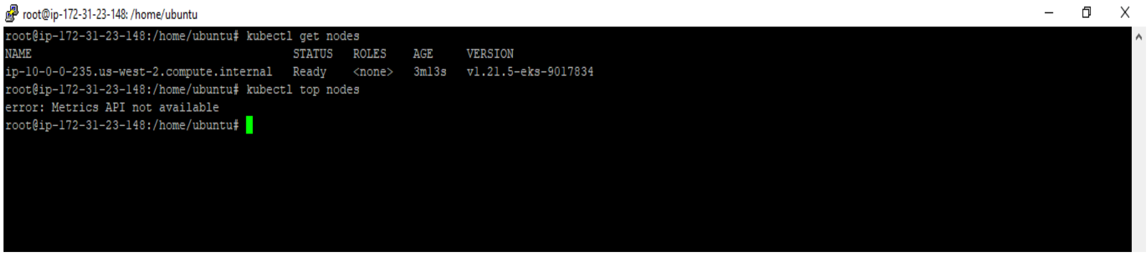
# Horizontal Pod Autoscaling

The Horizontal Pod Autoscaler changes the shape of your Kubernetes workload by automatically increasing or decreasing the number of Pods in response to the workload's CPU or memory consumption, or in response to custom metrics reported from within Kubernetes or external metrics from sources outside of your cluster

Architectural Diagram:



As shown in the image below, until and unless metric server is not deployed in the cluster, it will not show the real time metrics of the system.



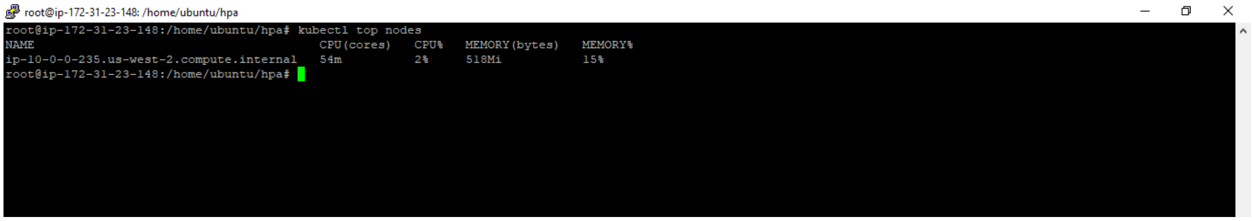
```
root@ip-172-31-23-148:/home/ubuntu# kubectl get nodes
NAME                                        STATUS   ROLES    AGE     VERSION
ip-10-0-0-235.us-west-2.compute.internal   Ready    <none>   3m13s   v1.21.5-eks-9017834
root@ip-172-31-23-148:/home/ubuntu# kubectl top nodes
error: Metrics API not available
root@ip-172-31-23-148:/home/ubuntu#
```

Install Metrics Server commands:

– kubectl apply -f https://github.com/kubernetes-sigs/metrics-server/releases/latest/download/components.yaml

– kubectl get deployment metrics-server -n kube-system

As shown in image below, it is deployed and working fine.



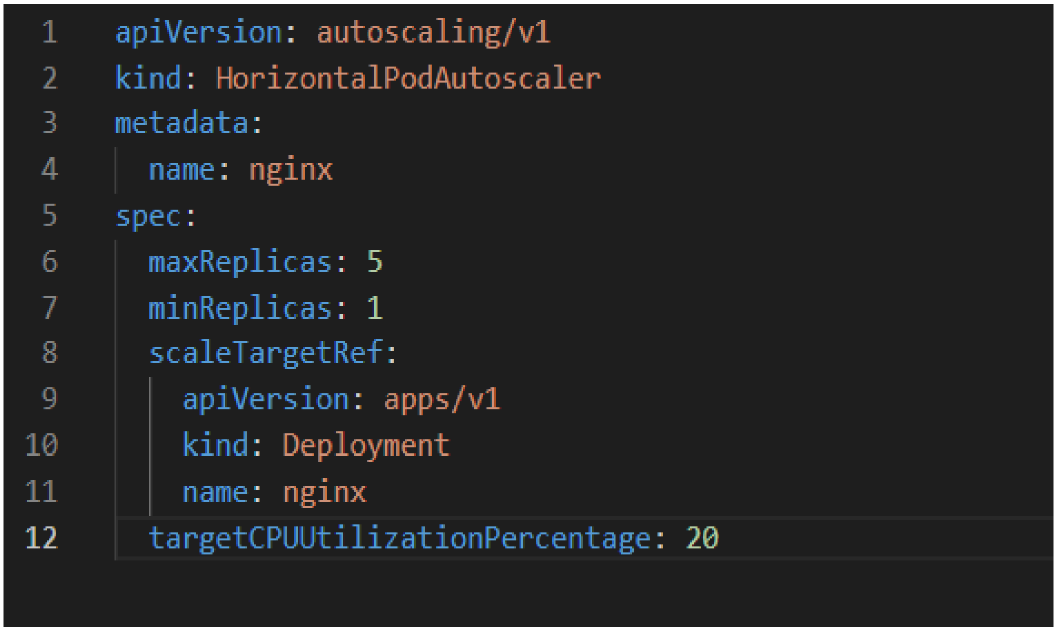
```
root@ip-172-31-23-148:/home/ubuntu/hpa# kubectl top nodes
NAME                                       CPU(cores)   CPU%   MEMORY(bytes)   MEMORY%
ip-10-0-0-235.us-west-2.compute.internal   54m          2%     518Mi           15%
root@ip-172-31-23-148:/home/ubuntu/hpa#
```

Now, using simple yaml files, I deployed Nginx and its service in the cluster. Service is used to expose the nginx pod through NodePort.

Now, to make stress on the Nginx, I have used the Siege tool for stress testing purpose.

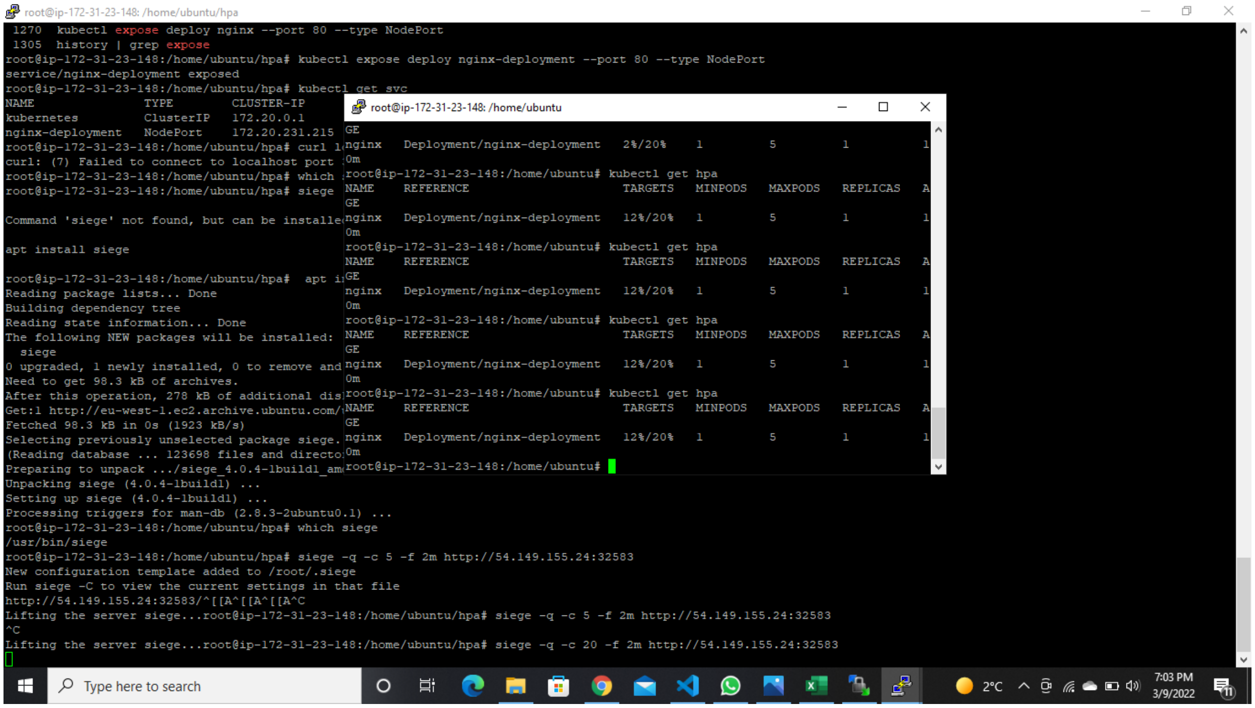
Command to install siege:                         apt-get install siege
Command to create stress over Nginx:   siege -q -c 20 -f 2m http://url

For enabling Horizontal Pod Autoscaling, I used the below shown file named hpa.yaml.
Command used: kubectl -f apply hpa.yaml



```yaml
apiVersion: autoscaling/v1
kind: HorizontalPodAutoscaler
metadata:
  name: nginx
spec:
  maxReplicas: 5
  minReplicas: 1
  scaleTargetRef:
    apiVersion: apps/v1
    kind: Deployment
    name: nginx
  targetCPUUtilizationPercentage: 20
```

As shown in the image below, commands are run and stress on the Nginx is increased.

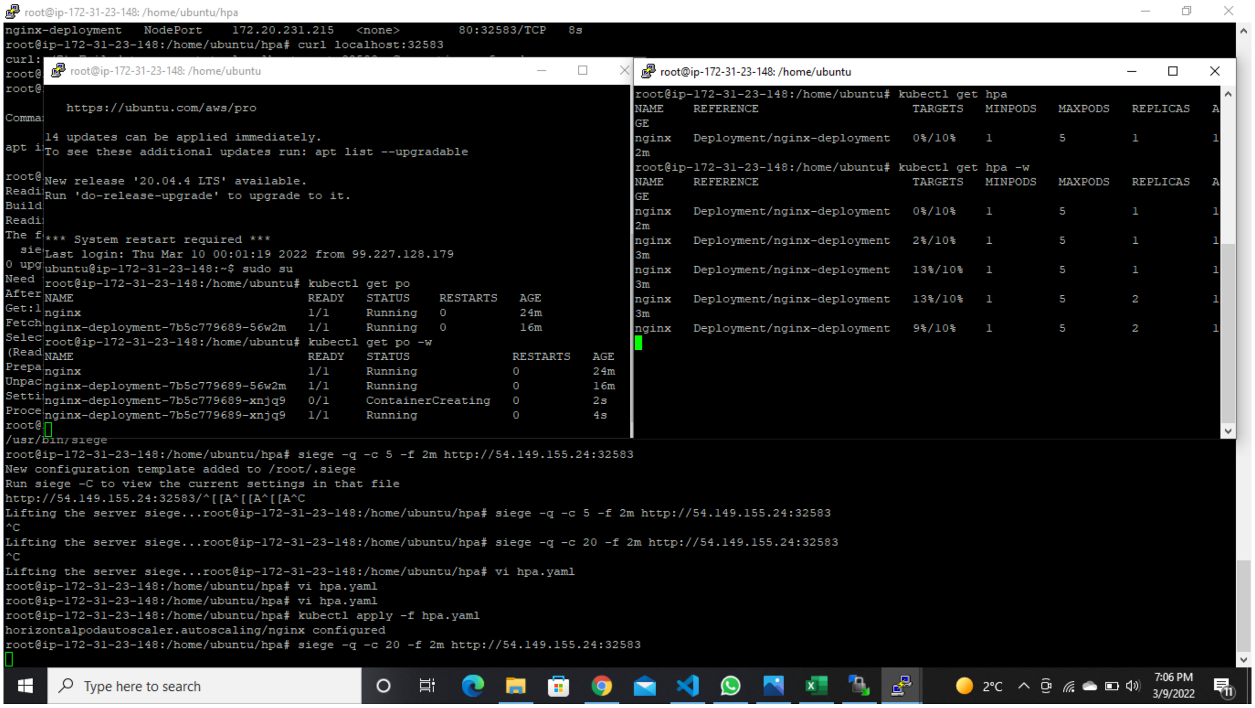


```
 1270  kubectl expose deploy nginx --port 80 --type NodePort
 1305  history | grep expose
root@ip-172-31-23-148:/home/ubuntu/hpa# kubectl expose deploy nginx-deployment --port 80 --type NodePort
service/nginx-deployment exposed
root@ip-172-31-23-148:/home/ubuntu/hpa# kubectl get svc
NAME                 TYPE        CLUSTER-IP
kubernetes           ClusterIP   172.20.0.1
nginx-deployment     NodePort    172.20.231.215
root@ip-172-31-23-148:/home/ubuntu/hpa# curl l
curl: (7) Failed to connect to localhost port
root@ip-172-31-23-148:/home/ubuntu/hpa# which
root@ip-172-31-23-148:/home/ubuntu/hpa# siege

Command 'siege' not found, but can be installe

apt install siege

root@ip-172-31-23-148:/home/ubuntu/hpa#  apt i
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following NEW packages will be installed:
  siege
0 upgraded, 1 newly installed, 0 to remove and
Need to get 98.3 kB of archives.
After this operation, 278 kB of additional dis
Get:1 http://eu-west-1.ec2.archive.ubuntu.com/
Fetched 98.3 kB in 0s (1923 kB/s)
Selecting previously unselected package siege.
(Reading database ... 123698 files and directo
Preparing to unpack .../siege_4.0.4-1build1_am
Unpacking siege (4.0.4-1build1) ...
Setting up siege (4.0.4-1build1) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
root@ip-172-31-23-148:/home/ubuntu/hpa# which siege
/usr/bin/siege
root@ip-172-31-23-148:/home/ubuntu/hpa# siege -q -c 5 -f 2m http://54.149.155.24:32583
New configuration template added to /root/.siege
Run siege -C to view the current settings in that file
http://54.149.155.24:32583/^[[A^[[A^[[A^C
Lifting the server siege...root@ip-172-31-23-148:/home/ubuntu/hpa# siege -q -c 5 -f 2m http://54.149.155.24:32583
^C
Lifting the server siege...root@ip-172-31-23-148:/home/ubuntu/hpa# siege -q -c 20 -f 2m http://54.149.155.24:32583
```

root@ip-172-31-23-148: /home/ubuntu

```
GE
nginx   Deployment/nginx-deployment   2%/20%    1         5         1            1
0m
root@ip-172-31-23-148:/home/ubuntu# kubectl get hpa
NAME    REFERENCE                     TARGETS   MINPODS   MAXPODS   REPLICAS   A
GE
nginx   Deployment/nginx-deployment   12%/20%   1         5         1            1
0m
root@ip-172-31-23-148:/home/ubuntu# kubectl get hpa
NAME    REFERENCE                     TARGETS   MINPODS   MAXPODS   REPLICAS   A
GE
nginx   Deployment/nginx-deployment   12%/20%   1         5         1            1
0m
root@ip-172-31-23-148:/home/ubuntu# kubectl get hpa
NAME    REFERENCE                     TARGETS   MINPODS   MAXPODS   REPLICAS   A
GE
nginx   Deployment/nginx-deployment   12%/20%   1         5         1            1
0m
root@ip-172-31-23-148:/home/ubuntu# kubectl get hpa
NAME    REFERENCE                     TARGETS   MINPODS   MAXPODS   REPLICAS   A
GE
nginx   Deployment/nginx-deployment   12%/20%   1         5         1            1
0m
root@ip-172-31-23-148:/home/ubuntu#
```

As it reached the threshold, we set for the horizontal pod autoscaling. It will start creating new pods of the same kind. Hence, our target is accomplished by having more pods in real time to manage the stress on the application.



```
nginx-deployment     NodePort    172.20.231.215   <none>        80:32583/TCP   8s
root@ip-172-31-23-148:/home/ubuntu/hpa# curl localhost:32583
curl:
root@
root@

Comma

apt i

root@
Readi
Build
Readi
The f
  sie
0 upg
Need
After
Get:1
Fetch
Selec
(Read
Prepa
Unpac
Setti
Proce
root@
/usr/bin/siege
root@ip-172-31-23-148:/home/ubuntu/hpa# siege -q -c 5 -f 2m http://54.149.155.24:32583
New configuration template added to /root/.siege
Run siege -C to view the current settings in that file
http://54.149.155.24:32583/^[[A^[[A^[[A^C
Lifting the server siege...root@ip-172-31-23-148:/home/ubuntu/hpa# siege -q -c 5 -f 2m http://54.149.155.24:32583
^C
Lifting the server siege...root@ip-172-31-23-148:/home/ubuntu/hpa# siege -q -c 20 -f 2m http://54.149.155.24:32583
^C
Lifting the server siege...root@ip-172-31-23-148:/home/ubuntu/hpa# vi hpa.yaml
root@ip-172-31-23-148:/home/ubuntu/hpa# vi hpa.yaml
root@ip-172-31-23-148:/home/ubuntu/hpa# vi hpa.yaml
root@ip-172-31-23-148:/home/ubuntu/hpa# kubectl apply -f hpa.yaml
horizontalpodautoscaler.autoscaling/nginx configured
root@ip-172-31-23-148:/home/ubuntu/hpa# siege -q -c 20 -f 2m http://54.149.155.24:32583
```

root@ip-172-31-23-148: /home/ubuntu

```

  https://ubuntu.com/aws/pro

14 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

New release '20.04.4 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


*** System restart required ***
Last login: Thu Mar 10 00:01:19 2022 from 99.227.128.179
ubuntu@ip-172-31-23-148:~$ sudo su
root@ip-172-31-23-148:/home/ubuntu# kubectl get po
NAME                                READY   STATUS    RESTARTS   AGE
nginx                               1/1     Running   0          24m
nginx-deployment-7b5c779689-56w2m   1/1     Running   0          16m
root@ip-172-31-23-148:/home/ubuntu# kubectl get po -w
NAME                                READY   STATUS              RESTARTS   AGE
nginx                               1/1     Running             0          24m
nginx-deployment-7b5c779689-56w2m   1/1     Running             0          16m
nginx-deployment-7b5c779689-xnjq9   0/1     ContainerCreating   0          2s
nginx-deployment-7b5c779689-xnjq9   1/1     Running             0          4s
```

root@ip-172-31-23-148: /home/ubuntu

```
root@ip-172-31-23-148:/home/ubuntu# kubectl get hpa
NAME    REFERENCE                     TARGETS   MINPODS   MAXPODS   REPLICAS   A
GE
nginx   Deployment/nginx-deployment   0%/10%    1         5         1          1
2m
root@ip-172-31-23-148:/home/ubuntu# kubectl get hpa -w
NAME    REFERENCE                     TARGETS   MINPODS   MAXPODS   REPLICAS   A
GE
nginx   Deployment/nginx-deployment   0%/10%    1         5         1          1
2m
nginx   Deployment/nginx-deployment   2%/10%    1         5         1          1
3m
nginx   Deployment/nginx-deployment   13%/10%   1         5         1          1
3m
nginx   Deployment/nginx-deployment   13%/10%   1         5         2          1
3m
nginx   Deployment/nginx-deployment   9%/10%    1         5         2          1
```