Carnegie Mellon University

Cloud Computing

Code: S24 15619

Name: Mbonabucya James

Andrew_id : Jmbonabu

Spring 2024

MSIT'24

On the 02nd Feb 2024

a)

Looking at the ELB requests, the target RPS has never reached the 35 max. looking at the total requests there exist high spikes in the request.

Target RPS reached 35

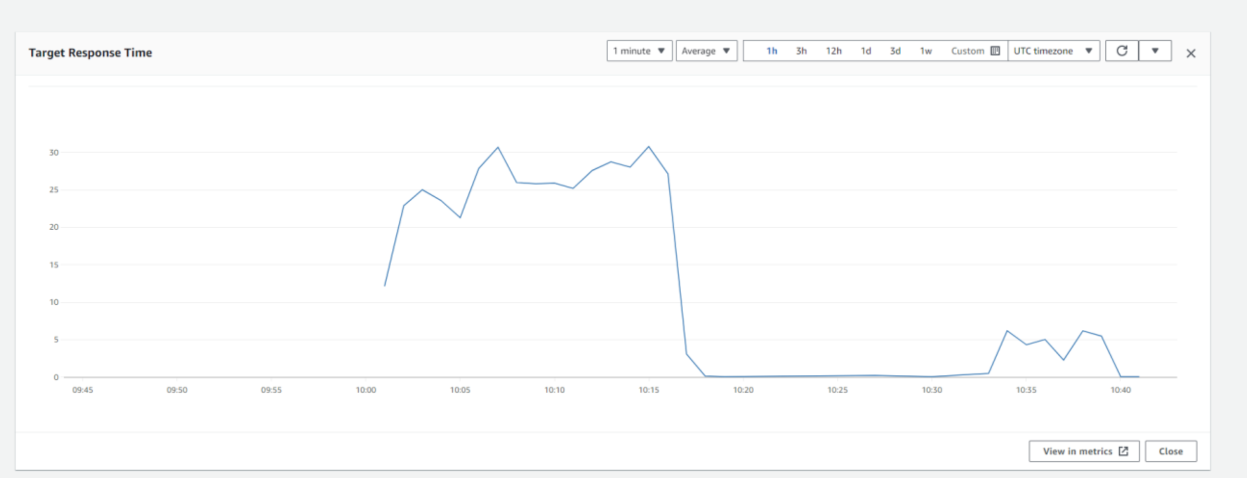


Figure 1 Target RPS

The load balancer   was with a listener configured to respond with a fixed message when receiving HTTP requests on port 80. The graph shows the fluctuations in the total requests patterns that indicates the use of load balancing i.e the load was distributed to the machines available within the auto scaling group

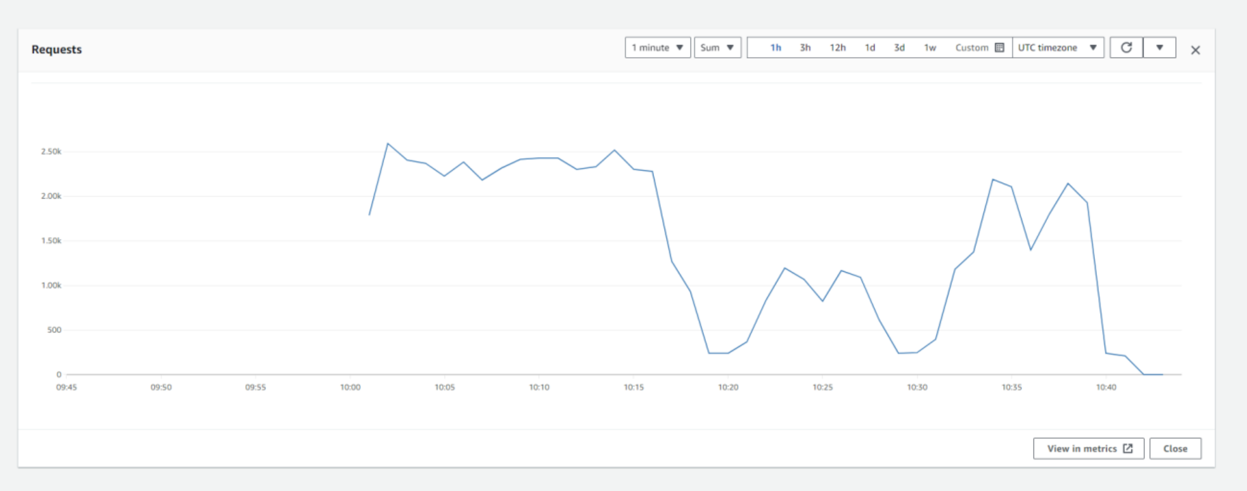


Figure 2 Total requests

Regarding the 504 requests, the patter

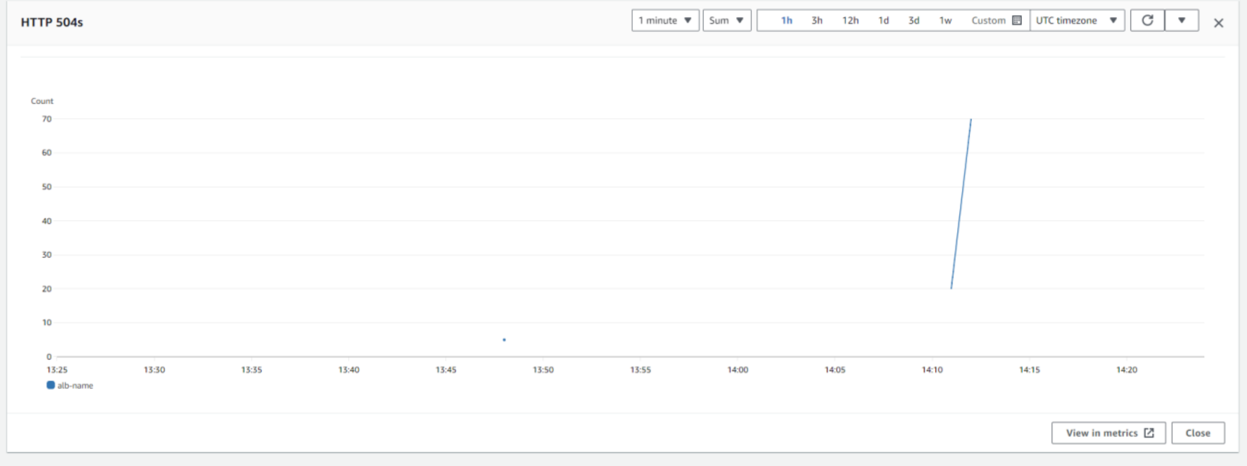


Figure 3 HTPP 504 requests

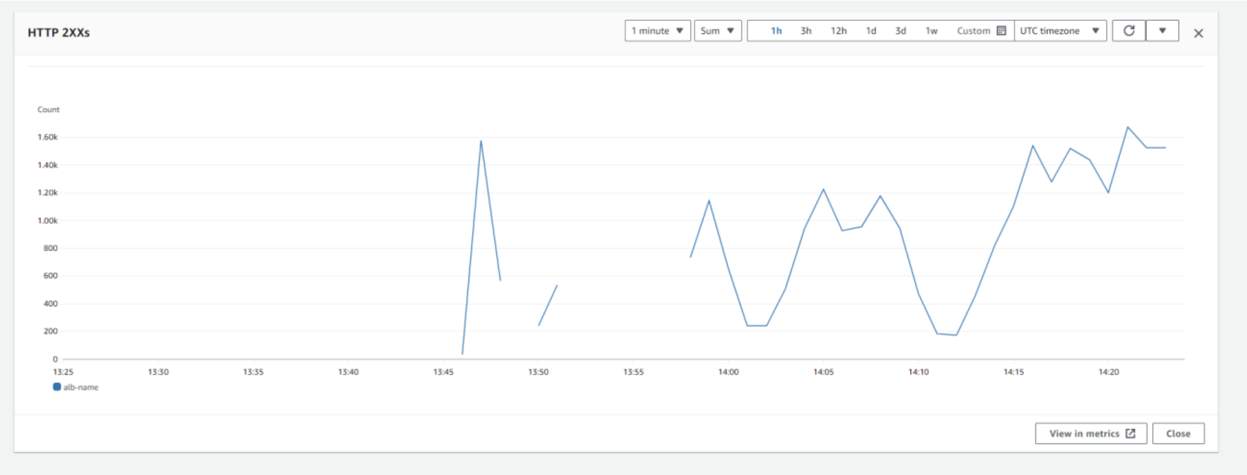
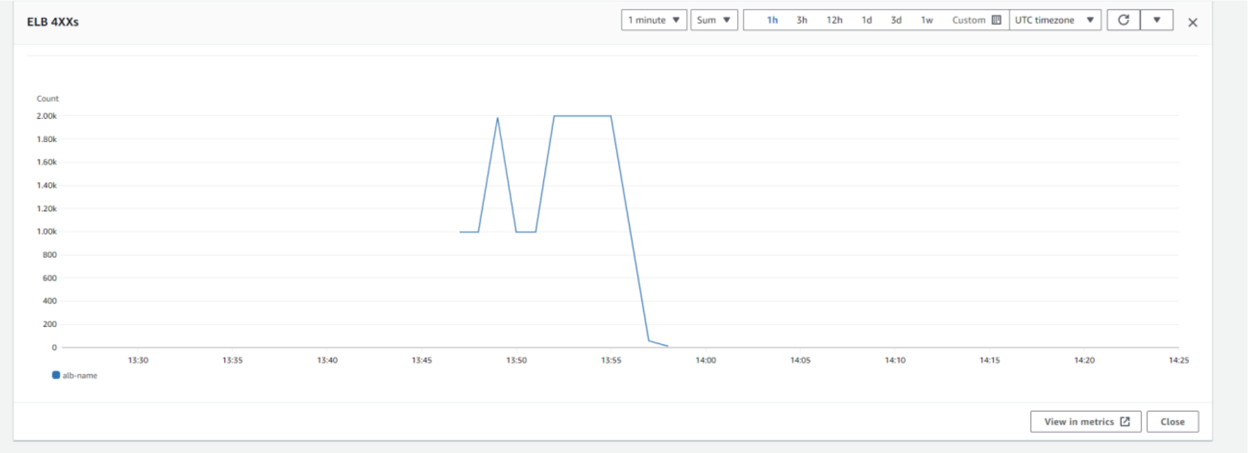
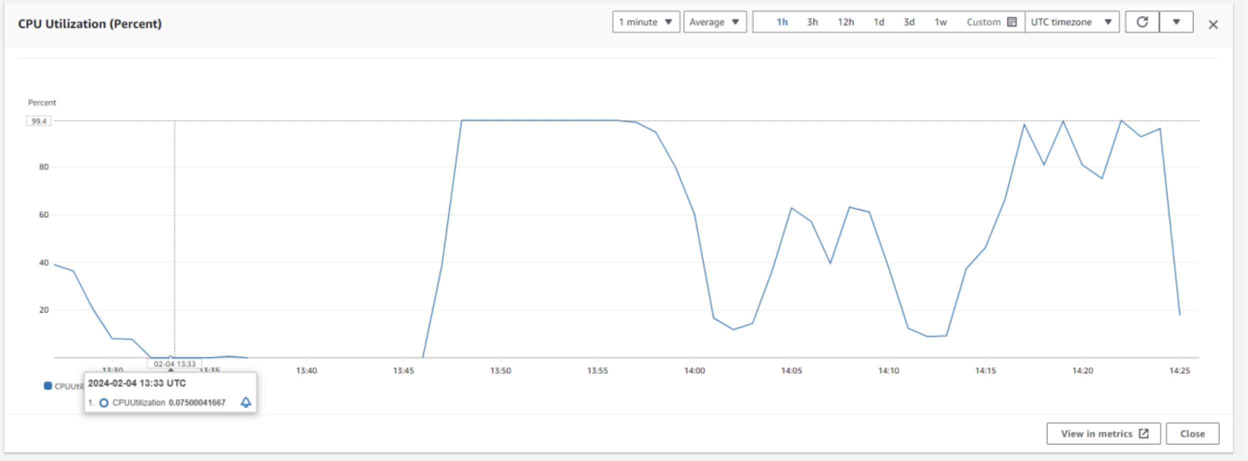


Figure 4 ELB HTTP 2xx

There is fluctuation in the requests that indicates the balancing the load among the instances



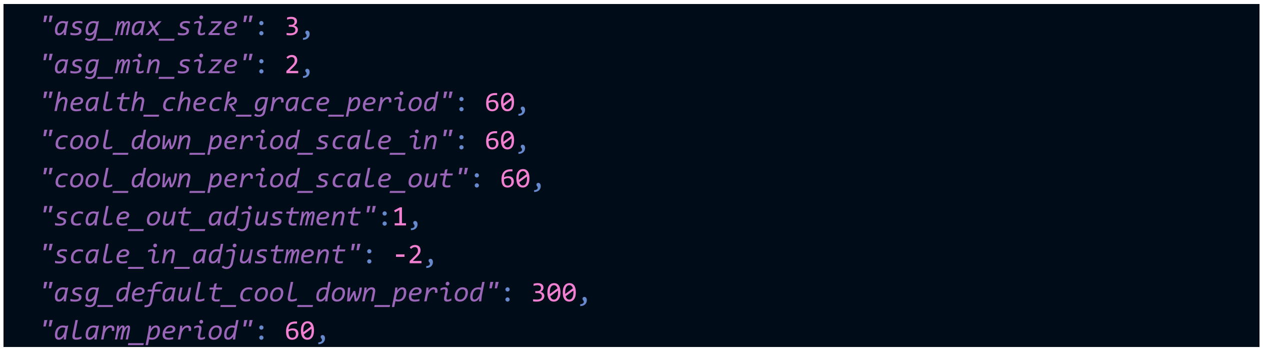
Auto scaling group CPU utilization shows that there were high CPU utilization at the beginning of the



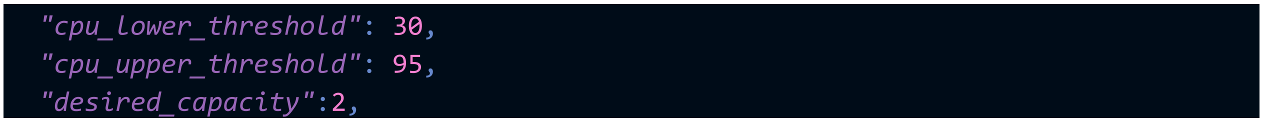
b)

after many successive tests , I came to realize that the maximum instances needed to handle the load were 2 . by reaching to the 2 instances the additional instance was becoming idle.

Here are the configurations



```
  "asg_max_size": 3,
  "asg_min_size": 2,
  "health_check_grace_period": 60,
  "cool_down_period_scale_in": 60,
  "cool_down_period_scale_out": 60,
  "scale_out_adjustment":1,
  "scale_in_adjustment": -2,
  "asg_default_cool_down_period": 300,
  "alarm_period": 60,
```



```
  "cpu_lower_threshold": 30,
  "cpu_upper_threshold": 95,
  "desired_capacity":2,
```

The CPU threshold to was reaching to 99 percent with single instance and I had to scale out while the lower threshold was 3o percent

Health check, grace period and alarm period were put to one minute because looking at the load the requests were minimum and I had to be more aggressive

As per above screenshots, the CPU utilization for the auto scaling group shows the high spike at the beginning of the test since I started with single instance, after scaling out with additional instance , the load was stable hence reduced CPU utilization.