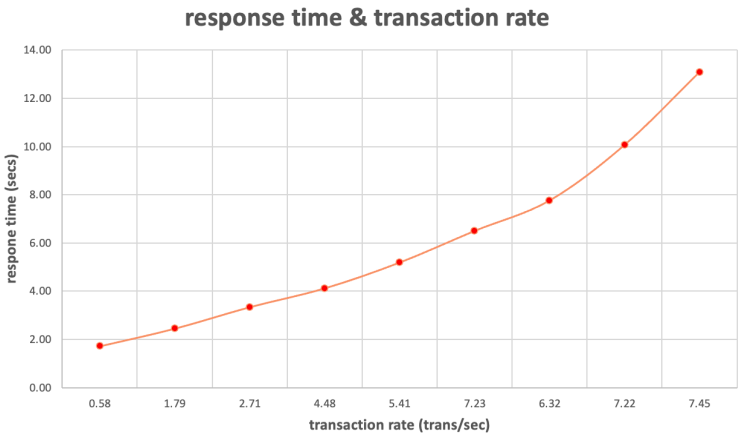
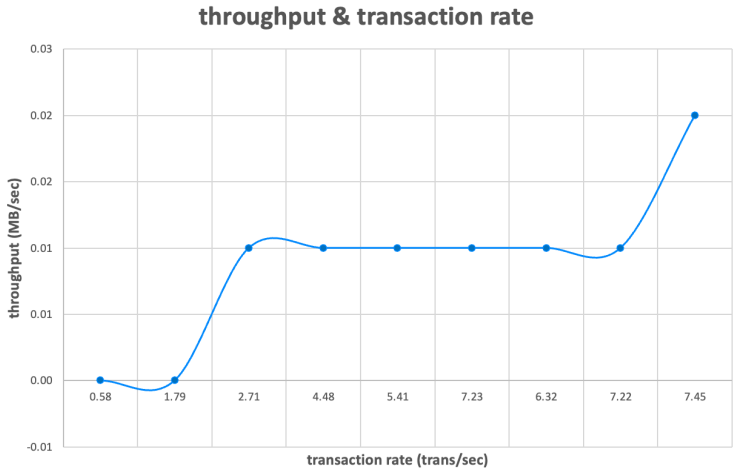
Activity 3 AutoScaling

Activity 3 II PaaS AutoScaling

1. From part D, show the inflection point on both of your curves and explain how you map this to the appropriate system resources (performance metrics) to use to trigger auto-scaling for PaaS.

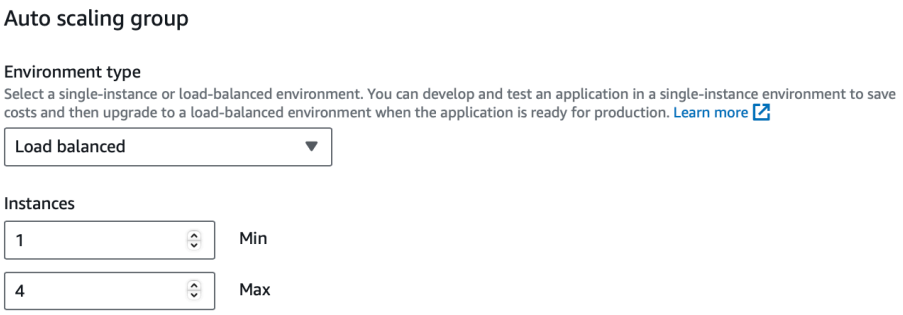
siege -c	response time (secs)	throughput (MB/sec)	transaction rate (trans/sec)
1	1.72	0.00	0.58
5	2.45	0.00	1.79
10	3.34	0.01	2.71
20	4.12	0.01	4.48
30	5.19	0.01	5.41
50	6.50	0.01	7.23
70	7.76	0.01	6.32
100	10.06	0.01	7.22
150	13.08	0.02	7.45
155	13.74	0.02	7.35
160	21.45	0.01	5.10
170	23.78	0.01	4.86
200	30.08	0.01	3.88

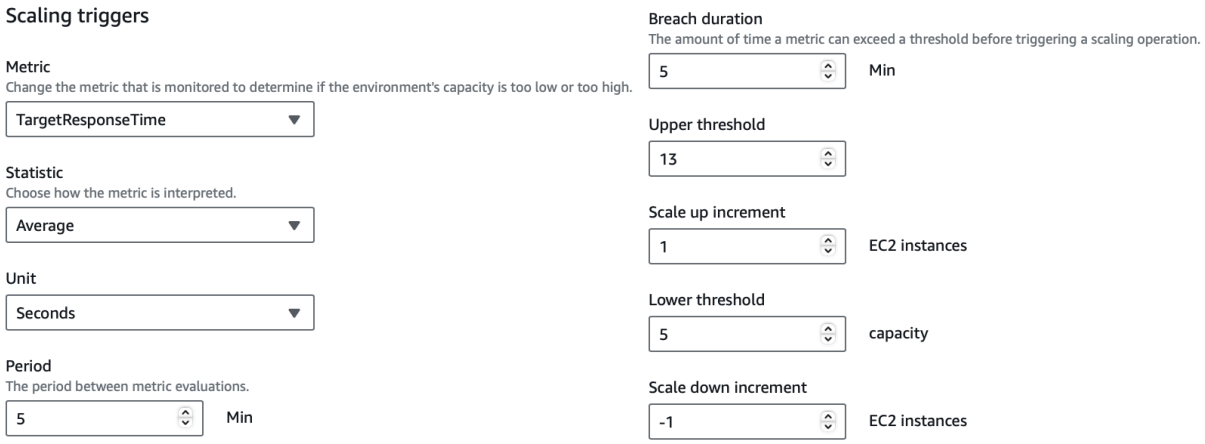




2. Describe the scaling policy you configured to scale up and scale down.

From response time graph, the web server can work normally until response time 13.74 secs. So, I decide to set scaling policy with **target response time** upper threshold at 13 secs to scale up and lower threshold at 5 secs to scale down.





3. How did you run siege (what options) to trigger scaling up and scaling down?

siege option to trigger scaling up.

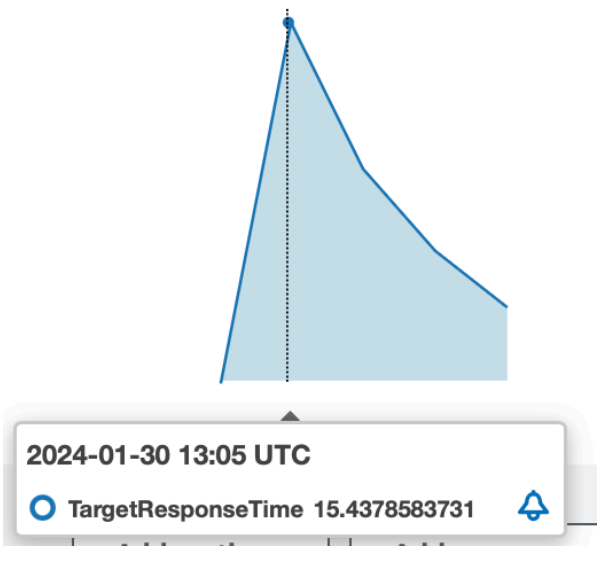
```
siege -c100 -d1 -r10 http://act3-default-env.eba-puzdk9ap.us-west-2.elasticbeanstalk.com
```

siege option to trigger scaling down.

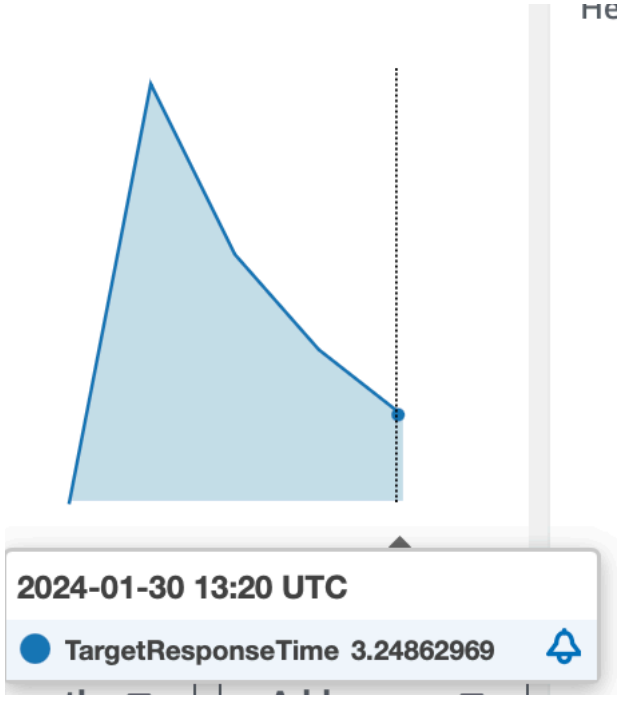
```
siege -c5 -d10 -r10 http://act3-default-env.eba-puzdk9ap.us-west-2.elasticbeanstalk.com
```

4. Was the scaling up and scaling down behavior consistent with your scaling policy?

TargetResponseTime when it is scaling up.



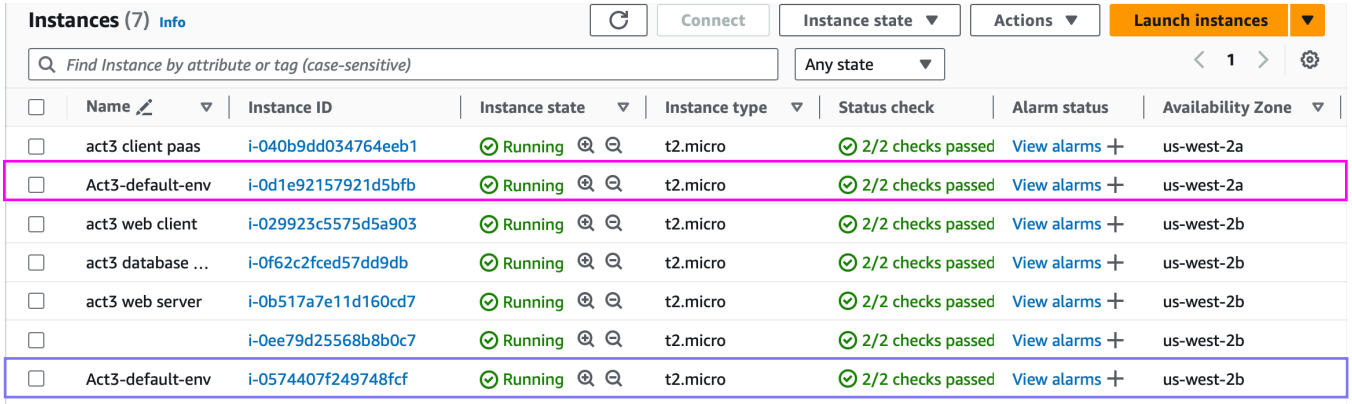
TargetResponseTime when it is scaling down.



When the matric average of TargetResponseTime exceeds the upper threshold of scaling policy, it indicates a need to scale up. Conversely, if it falls below the lower threshold, scaling down is warranted.

5. Include screenshots from Elastic Beanstalk showing the number of instances running and screenshots of resource monitoring data for your environment in your report to confirm that you successfully triggered auto-scaling according to your configured policies.

Instances running while scaling up.



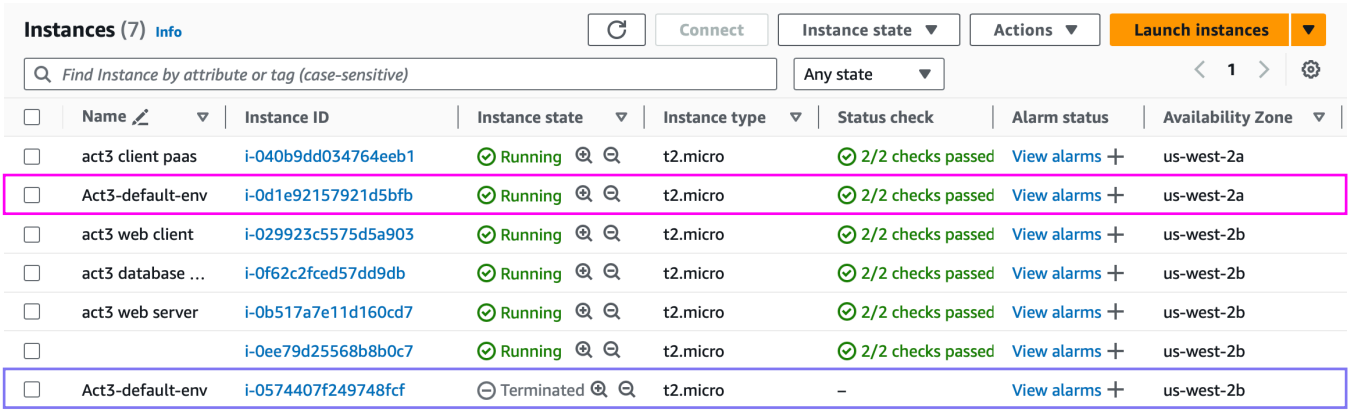
Instances (7) Info

Connect | Instance state | Actions | Launch instances

Find Instance by attribute or tag (case-sensitive) | Any state | 1

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone
act3 client paas	i-040b9dd034764eeb1	Running	t2.micro	2/2 checks passed	View alarms	us-west-2a
Act3-default-env	i-0d1e92157921d5bfb	Running	t2.micro	2/2 checks passed	View alarms	us-west-2a
act3 web client	i-029923c5575d5a903	Running	t2.micro	2/2 checks passed	View alarms	us-west-2b
act3 database ...	i-0f62c2fced57dd9db	Running	t2.micro	2/2 checks passed	View alarms	us-west-2b
act3 web server	i-0b517a7e11d160cd7	Running	t2.micro	2/2 checks passed	View alarms	us-west-2b
	i-0ee79d25568b8b0c7	Running	t2.micro	2/2 checks passed	View alarms	us-west-2b
Act3-default-env	i-0574407f249748fcf	Running	t2.micro	2/2 checks passed	View alarms	us-west-2b

Instances running while scaling down.



Instances (7) Info

Connect | Instance state | Actions | Launch instances

Find Instance by attribute or tag (case-sensitive) | Any state | 1

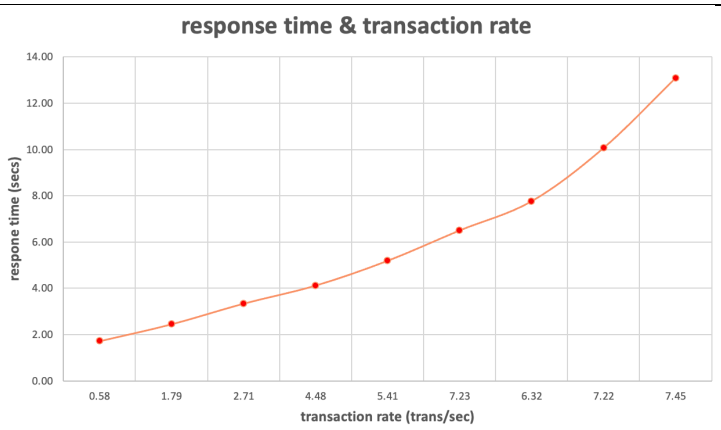
Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone
act3 client paas	i-040b9dd034764eeb1	Running	t2.micro	2/2 checks passed	View alarms	us-west-2a
Act3-default-env	i-0d1e92157921d5bfb	Running	t2.micro	2/2 checks passed	View alarms	us-west-2a
act3 web client	i-029923c5575d5a903	Running	t2.micro	2/2 checks passed	View alarms	us-west-2b
act3 database ...	i-0f62c2fced57dd9db	Running	t2.micro	2/2 checks passed	View alarms	us-west-2b
act3 web server	i-0b517a7e11d160cd7	Running	t2.micro	2/2 checks passed	View alarms	us-west-2b
	i-0ee79d25568b8b0c7	Running	t2.micro	2/2 checks passed	View alarms	us-west-2b
Act3-default-env	i-0574407f249748fcf	Terminated	t2.micro	–	View alarms	us-west-2b

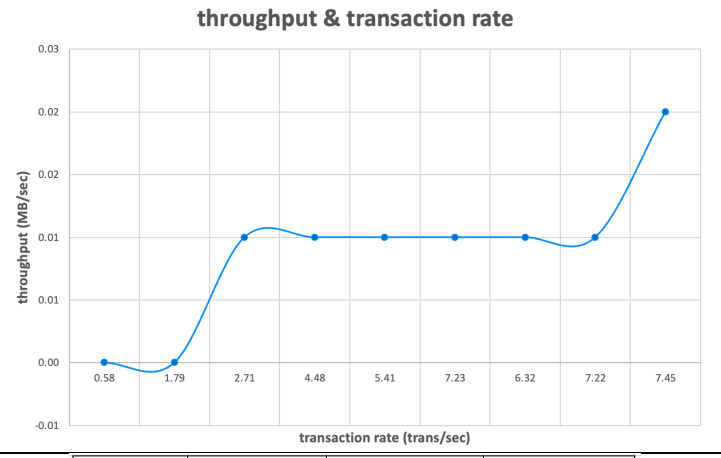
Activity 3 III IaaS AutoScaling

1. Consider observed baseline performance in terms of response time and throughput as a function of offered load (# of client requests/second) when you deploy the same web app to to PaaS vs. IaaS. Double-check to see if they are deployed on the same actual instance sizes and check the AZ.

1.1. Discuss the differences in baseline performance in terms of response time and throughput as a function of offered load (# of client requests/second), and conclude which provides better performance?

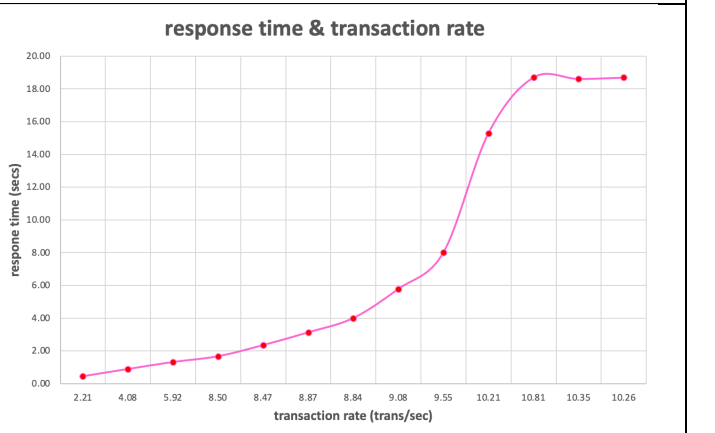
Paas

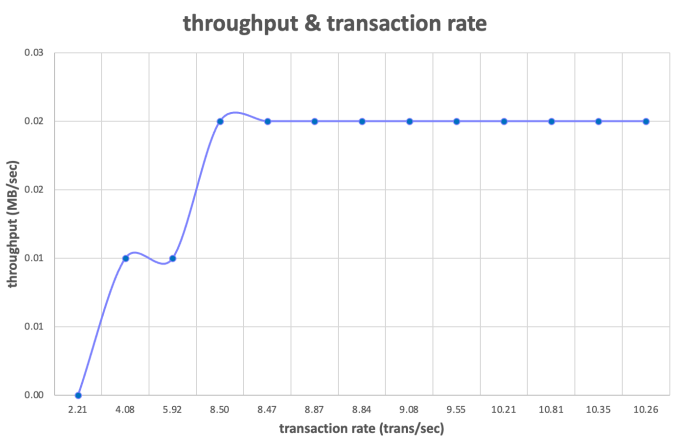




siege -c	response time (secs)	throughput (MB/sec)	transaction rate (trans/sec)
1	1.72	0.00	0.58
5	2.45	0.00	1.79
10	3.34	0.01	2.71
20	4.12	0.01	4.48
30	5.19	0.01	5.41
50	6.50	0.01	7.23
70	7.76	0.01	6.32
100	10.06	0.01	7.22
150	13.08	0.02	7.45
155	13.74	0.02	7.35
160	21.45	0.01	5.10
170	23.78	0.01	4.86
200	30.08	0.01	3.88

Iaas





siege -c	response time (secs)	throughput (MB/sec)	transaction rate (trans/sec)
1	0.45	0.00	2.21
5	0.89	0.01	4.08
10	1.32	0.01	5.92
20	1.67	0.02	8.50
30	2.35	0.02	8.47
40	3.13	0.02	8.87
50	4.00	0.02	8.84
70	5.78	0.02	9.08
100	8.00	0.02	9.55
200	15.29	0.02	10.21
300	18.70	0.02	10.81
500	18.61	0.02	10.35
520	18.69	0.02	10.26
521	36.63	0.02	6.40
522	36.69	0.02	6.36
525	36.41	0.02	6.43
530	35.57	0.02	6.54
550	36.70	0.02	6.38
600	36.67	0.02	6.38
700	36.08	0.02	6.51

The following information demonstrates that IaaS can support more clients, offer faster response times, and handle greater throughput compared to PaaS. Therefore, IaaS exhibits better performance, whereas PaaS involves fewer deployment steps than IaaS.

1.2. Are there any reasons that you think would sensibly explain the performance differences?

Considerations from user experimentation indicate that users desire faster access, which is why IaaS offers better performance.

2. For IaaS, what scaling policy did you use to scale up and scale down?

I set target tracking policy with **Network in** at **3MB** to be scaling policy.

3. How did you run siege (what options) to trigger scaling up and scaling down?

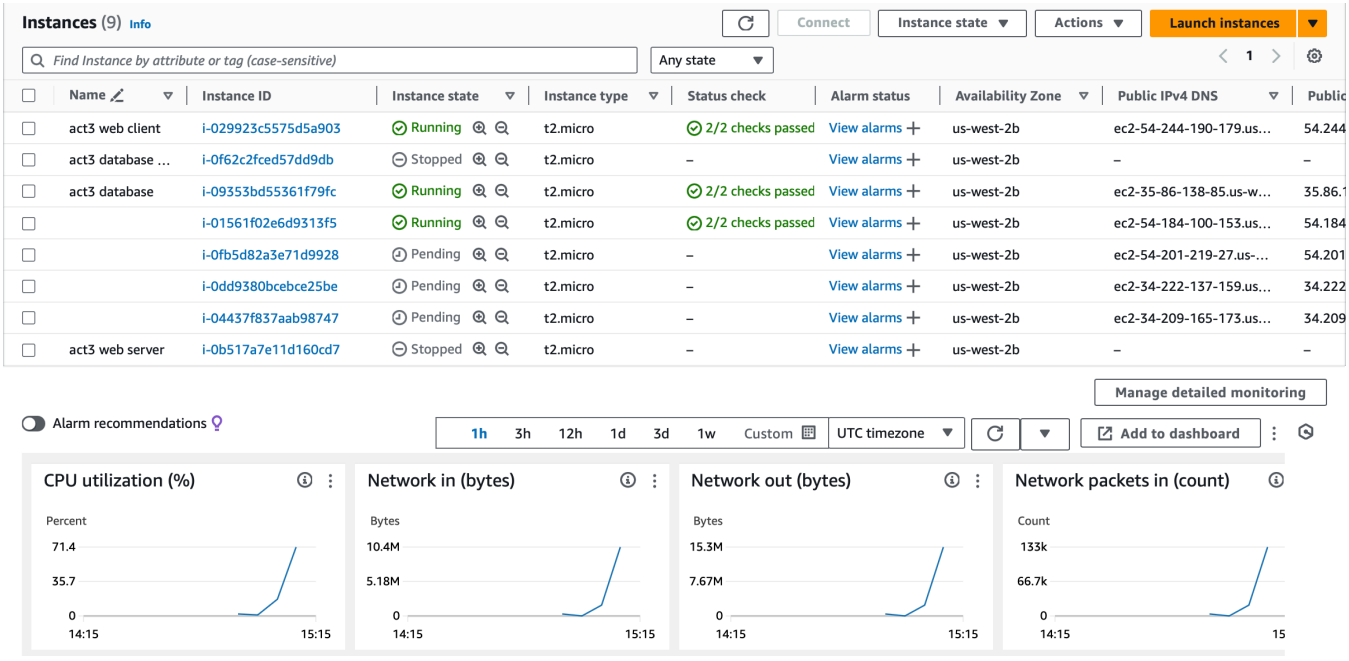
siege option to trigger scaling up.

```
siege -c200 -d1 -r10 http://phpiaasgroup-1-288227399.us-west-2.elb.amazonaws.com/index.php
```

4. Was the scaling up and scaling down behavior consistent with your scaling policy?

Yes. Because it creates new instances when Network in exceeds 3 MB.

5. Include screenshots from EC2 monitoring data in the report to make sure I can see that you did trigger auto-scaling according to your configured policies.



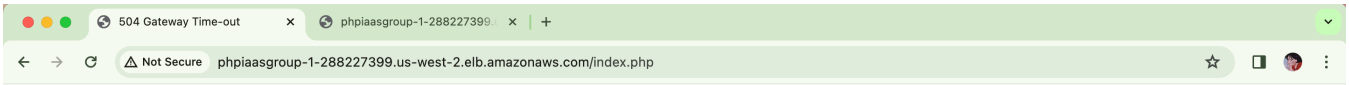
6. Consider what Amazon did when you terminated your EC2 instance(s) in part D (fault tolerance).

6.1. Is Amazon able to launch new instances to replace the one(s) you killed?

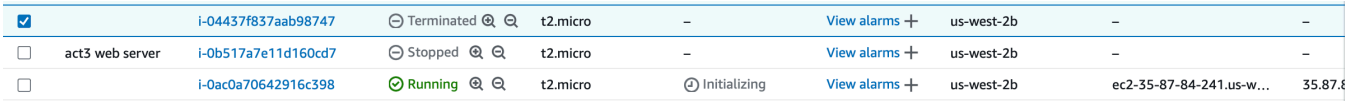
After killing the instance, the browser will display a Gateway Time-out error, and then it will launch a new instance. In siege will appear red path a moment then appears blue path.

6.2. When the new instances are launched, are your web clients (browser or siege) able to access the web application like normal? Provide evidence of what your web clients see.

Yes. After the web server error and cannot access them, it will provide new instance and be able to access a browser like normal.



504 Gateway Time-out



7. Discuss the differences between IaaS auto-scaling vs. PaaS auto-scaling. What are the strengths of each approach?

IaaS auto-scaling provides the automatic adjustment of the number of resources based on predefined policies and conditions. This approach gives the users a greater control over the environment with more flexibility. But in PaaS, it automatically adjusts computing resources, such as the number of application instances or containers, based on the changing demands of an application, which make auto-scaling in PaaS platforms is often easier to configure and deploy.

8. Discuss the differences between IaaS vs. PaaS. What are the strengths of each approach?

IaaS provides virtualized computing resources over the internet, allowing users to run and manage applications on virtualized infrastructure. But PaaS provides a platform that includes not only computing resources but also development tools, databases, middleware, and other components, which is easier for users to focus on writing code without worrying about the underlying infrastructure.