

Extra Credit Questions

Q 13 c)

Observations:


At roughly 4:15pm I started to run my load tests with my swarm of 100 users at 5/second. At 4:20 pm, I changed this to 300 users at 20/sec. At this point I saw that my instances in the auto-scaling group increased from 2 to 3 at roughly 4:20pm. This was owing to an increase in the desired capacity for instances from 2 to 3 triggered by the scale out alarm. After waiting for another 5 minutes, I noticed the autoscaling group launched another new instance, and the instance was up and running at 4:27pm, bringing the total number of instances to 4. This pattern continued, where every 5 minutes, a new instance was launched, where it took roughly 2 extra minutes for the instance to get up and running from when it was launched. The reason for this behavior is owing to the scaling policy created for scaling out, where when the sum of successful responses exceeds 200 for one consistent minute (and in this case the successful responses are much more), then a new instance is launched. There is a lag of 5 minutes between successive launching of new instances because I set a parameter in the scaling policy which waits 300 seconds, before launching another new instance if the number of successful responses still exceed 200. This pattern continues until a total of 10 instances have been created.

Q 13 d)

Observations:

At 5:09pm I hit the stop button on the locust console. The scaling in process started at 6:07pm, when the first instance terminated. The scaling in process is a much longer process than scaling out in our case, since the scaling policy for scaling in dictates that the CloudWatch alarm on the load balancer is triggered when the target response time is below 10ms for at least one minute. Since the target response time is not purely dictated by the number of requests being received by the load balancer, it is often difficult to get a response time below 10ms, for one consistent minute. Furthermore, the slow process is exacerbated by the 300 seconds wait time before scaling in another instance after terminating an instance. The entire process of scaling in finished at 10:26pm, where all but 2 instances were terminated.

The following are screenshots of the Locust web console when running and after being stopped:

 LOCUST

HOST

https://srirama.ucomps.org

STATUS

RUNNING

300 users

Edit

RPS

152.6

FAILURES

0%

STOP

Reset Stats

Statistics

Charts

Failures

Exceptions

Download Data

Type	Name	# Requests	# Fails	Median (ms)	90%ile (ms)	Average (ms)	Min (ms)	Max (ms)	Average size (bytes)	Current RPS	Current Failures/s
GET	/	213398	0	7	10	8	6	1504	10138	152.6	0
	Aggregated	213398	0	7	10	8	6	1504	10138	152.6	0

Statistics Charts Failures Exceptions Download Data

Type	Name	# Requests	# Fails	Median (ms)	90%ile (ms)	Average (ms)	Min (ms)	Max (ms)	Average size (bytes)	Current RPS	Current Failures/s
GET	/	451637	0	7	9	8	6	3165	10138	150.4	0
	Aggregated	451637	0	7	9	8	6	3165	10138	150.4	0

The following is a screenshot of the timeline as different instances where launched during the scale out process, and finally terminated during the scale in.

Create Auto Scaling group
Actions

Filter: srirama
1 to 2 of 2 Auto Scaling Groups

Name	Launch Configuration	Instances	Desired	Min	Max	Availability Zones	DefaultCooldown	HealthCheckGracePeriod
srirama-web-a...	srirama-web-launch-co...	2	2	2	10	us-east-1e	300	300

Filter: Any Status
Filter scaling history...
1 to 25 of 30 History Items

Status	Description	Start Time	End Time
Successful	Terminating EC2 instance: i-0664c571d688c29e2	2020 June 8 22:26:38 UTC-5	2020 June 8 22:32:10 UTC-5
Successful	Terminating EC2 instance: i-082c1ac0310a56c0c	2020 June 8 22:19:59 UTC-5	2020 June 8 22:25:29 UTC-5
Successful	Terminating EC2 instance: i-0d677fa546c7484af	2020 June 8 22:13:50 UTC-5	2020 June 8 22:19:28 UTC-5
Successful	Terminating EC2 instance: i-0f20406c992f06713	2020 June 8 21:42:40 UTC-5	2020 June 8 21:48:11 UTC-5
Successful	Terminating EC2 instance: i-0e22422e6589913fd	2020 June 8 20:52:36 UTC-5	2020 June 8 20:57:59 UTC-5
Successful	Terminating EC2 instance: i-0204c332898edede2	2020 June 8 20:45:57 UTC-5	2020 June 8 20:51:24 UTC-5
Successful	Terminating EC2 instance: i-017bfa44d980ac16	2020 June 8 20:36:44 UTC-5	2020 June 8 20:42:23 UTC-5
Successful	Terminating EC2 instance: i-0ea156cb5a44a798	2020 June 8 18:07:35 UTC-5	2020 June 8 18:13:02 UTC-5
Successful	Launching a new EC2 instance: i-07d8064ab53b8e55f	2020 June 8 17:04:44 UTC-5	2020 June 8 17:05:17 UTC-5
Successful	Launching a new EC2 instance: i-0e22422e6589913fd	2020 June 8 16:58:05 UTC-5	2020 June 8 16:58:37 UTC-5
Successful	Launching a new EC2 instance: i-0204c332898edede2	2020 June 8 16:51:55 UTC-5	2020 June 8 16:52:28 UTC-5
Successful	Launching a new EC2 instance: i-0f20406c992f06713	2020 June 8 16:45:46 UTC-5	2020 June 8 16:46:19 UTC-5
Successful	Launching a new EC2 instance: i-017bfa44d980ac16	2020 June 8 16:39:07 UTC-5	2020 June 8 16:39:40 UTC-5
Successful	Launching a new EC2 instance: i-082c1ac0310a56c0c	2020 June 8 16:32:58 UTC-5	2020 June 8 16:33:31 UTC-5
Successful	Launching a new EC2 instance: i-0d677fa546c7484af	2020 June 8 16:26:49 UTC-5	2020 June 8 16:27:22 UTC-5
Successful	Launching a new EC2 instance: i-0ea156cb5a44a798	2020 June 8 16:20:10 UTC-5	2020 June 8 16:20:42 UTC-5
Successful	Launching a new EC2 instance: i-0c01393c75dbfe8c8	2020 June 8 15:35:11 UTC-5	2020 June 8 15:35:44 UTC-5
Successful	Launching a new EC2 instance: i-0664c571d688c29e2	2020 June 8 15:35:11 UTC-5	2020 June 8 15:35:44 UTC-5

Q14)

Scaling Out Observations:

I started my ann_load python program at roughly 8:12pm. There was a new instance launched in the auto scaling group every five minutes. This occurred because my ann_load python program was sending a continuous stream of job requests to my annotator SNS topic, which breached the CloudWatch alarm threshold of number of messages being greater than 50 for 600 seconds. A new instance was launched after every 300 seconds, since the wait time for launching a new instance was specified to be 5 minutes. Hence the number of instances continuously increased roughly every 5 minutes, until 10 instances where reached.

Scaling In Observations:

I stopped my ann_load python program at 8:58pm. Again, like the case of the web server, the scaling in process took much longer than scaling out, since the number of requests received to be processed was very large. Hence the number of messages available for the queue was much greater than 5 (roughly 60,000), and hence this exceeded the threshold for messages being received as less than 5 for at least 600 seconds, and hence the scale in alarm could not have been triggered until much later (~20days if 2 messages were being read and deleted every second). In order for the scale in to effectively start, I had to purge my queue, so there were no more requests being received by the annotator. After roughly 10 minutes of purging (10:34pm), an instance terminated every five minutes as instructed by the wait time of 300 seconds.

The following is a screenshot of the timeline as different instances were launched during the scale out process, and finally terminated during the scale in:

Create Auto Scaling group

Actions

Filter:

1 to 2 of 2 Auto Scaling Groups

Name	Launch Configuration	Instances	Desired	Min	Max	Availability Zones	Default Cooldown	Health Check Grace
srirama-ann-a...	srirama-ann-launch-co...	2	2	2	10	us-east-1e	300	300

Filter: Any Status

1 to 25 of 34 History Items

Status	Description	Start Time	End Time
Successful	Terminating EC2 instance: i-053557c0354aaa0b	2020 June 8 23:20:31 UTC-5	2020 June 8 23:26:53 UTC-5
Successful	Terminating EC2 instance: i-06452e282d257926f	2020 June 8 23:14:52 UTC-5	2020 June 8 23:15:15 UTC-5
Successful	Terminating EC2 instance: i-0977e7282b5a5853e	2020 June 8 23:08:44 UTC-5	2020 June 8 23:09:06 UTC-5
Successful	Terminating EC2 instance: i-09b47826f9a6838ac	2020 June 8 23:02:35 UTC-5	2020 June 8 23:02:58 UTC-5
Successful	Terminating EC2 instance: i-0f72d734045ea62	2020 June 8 22:55:55 UTC-5	2020 June 8 22:56:36 UTC-5
Successful	Terminating EC2 instance: i-053d14ee6a03ae99	2020 June 8 22:49:46 UTC-5	2020 June 8 22:50:09 UTC-5
Successful	Terminating EC2 instance: i-0316aa16c0d1ac2bb	2020 June 8 22:43:37 UTC-5	2020 June 8 22:43:59 UTC-5
Successful	Terminating EC2 instance: i-0344d08c79bd01f70	2020 June 8 21:11:32 UTC-5	2020 June 8 21:11:55 UTC-5
Successful	Launching a new EC2 instance: i-0f72d734045ea62	2020 June 8 20:58:46 UTC-5	2020 June 8 20:59:18 UTC-5
Successful	Launching a new EC2 instance: i-093d14ee6a03ae99	2020 June 8 20:52:06 UTC-5	2020 June 8 20:52:36 UTC-5
Successful	Launching a new EC2 instance: i-0316aa16c0d1ac2bb	2020 June 8 20:45:57 UTC-5	2020 June 8 20:46:29 UTC-5
Successful	Launching a new EC2 instance: i-05199ddee4b145e71	2020 June 8 20:39:48 UTC-5	2020 June 8 20:40:20 UTC-5
Successful	Launching a new EC2 instance: i-06452e282d257926f	2020 June 8 20:33:08 UTC-5	2020 June 8 20:33:40 UTC-5
Successful	Launching a new EC2 instance: i-0977e7282b5a5853e	2020 June 8 20:26:58 UTC-5	2020 June 8 20:27:31 UTC-5
Successful	Launching a new EC2 instance: i-09b47826f9a6838ac	2020 June 8 20:20:50 UTC-5	2020 June 8 20:21:22 UTC-5
Successful	Launching a new EC2 instance: i-0344d08c79bd01f70	2020 June 8 20:14:40 UTC-5	2020 June 8 20:15:12 UTC-5
Successful	Launching a new EC2 instance: i-053557c0354aaa0b	2020 June 8 15:34:54 UTC-5	2020 June 8 15:35:26 UTC-5
Successful	Launching a new EC2 instance: i-0e5637c05874d2fc1	2020 June 8 15:34:54 UTC-5	2020 June 8 15:35:27 UTC-5