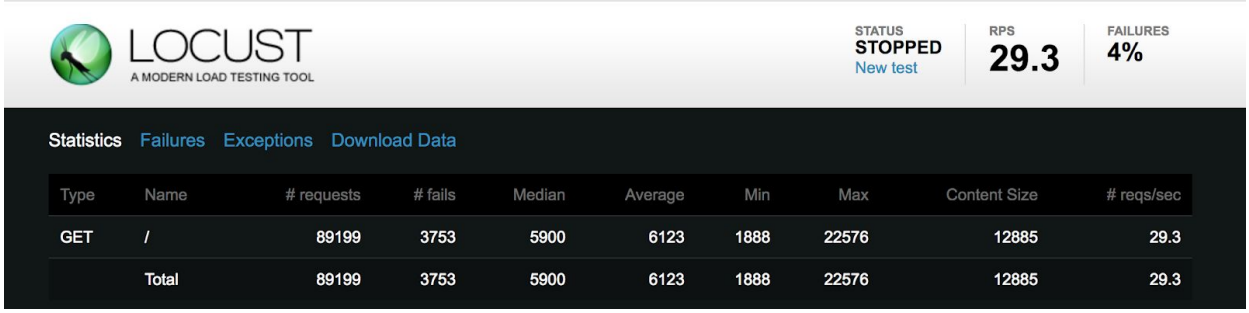
Load test on web servers :

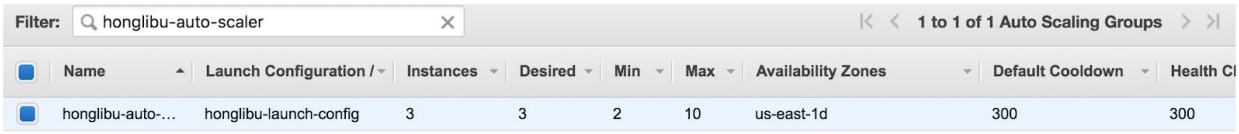
The instances incremented by 1 every 5 minutes. It's because the load balancer kept receiving high-load requests per minute. So the state of the alarm "honglibu-web-scale-out" kept being "ALARM". And,I had set scale-out policy as "300 seconds to warm up". So the auto scaler would wait for 5 minutes before adding a new instance. When the instance number reached the maximum number of instances we set, the increment stopped.

Here is my locust screenshot when instance number reached the maximum number:

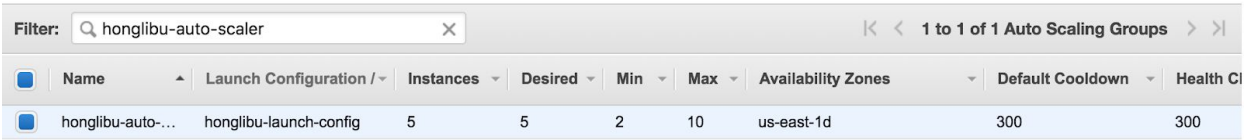


Type	Name	# requests	# fails	Median	Average	Min	Max	Content Size	# reqs/sec
GET	/	89199	3753	5900	6123	1888	22576	12885	29.3
	Total	89199	3753	5900	6123	1888	22576	12885	29.3

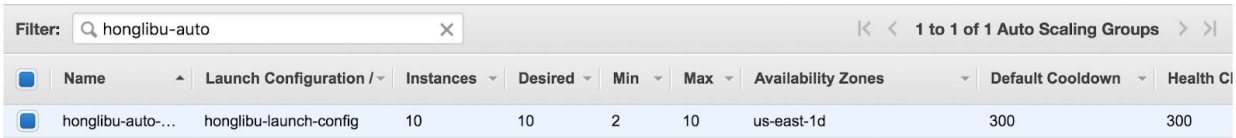
Below several screenshots show the stage when scaling out. At the beginning, two instances is running and ultimately, the instances reach 10.



Name	Launch Configuration /	Instances	Desired	Min	Max	Availability Zones	Default Cooldown	Health Cl
honglibu-auto-...	honglibu-launch-config	3	3	2	10	us-east-1d	300	300



Name	Launch Configuration /	Instances	Desired	Min	Max	Availability Zones	Default Cooldown	Health Cl
honglibu-auto-...	honglibu-launch-config	5	5	2	10	us-east-1d	300	300



Name	Launch Configuration /	Instances	Desired	Min	Max	Availability Zones	Default Cooldown	Health Cl
honglibu-auto-...	honglibu-launch-config	10	10	2	10	us-east-1d	300	300

After I stop Locust test, web farms start to scale in according to my 'honglibu-web-scale-in' policy. In this policy i defined the threshold as 'TargetResponseTime < 0.05 for 1 datapoint within 1 minute'. So if I stop the load test or reset the load test to be in low request workload(e.g. 1 user, at 1/sec), web response time will take less than 50ms, then the ELB will coordinate Auto Scaling group to terminated instances by 1 every 5 minutes. Of course, if consider server latency, the scale in alarm might be bounce between 'ok' and 'alarm'.

Load test on anntools servers:

Here is my screenshots of anntools auto scaling group when i was running autotest script in my local machine. As the anntools alarm, when i continuously send job requests, the sum of job in 15 minutes will reach 30(Threshold). Consequently, the anntools farm will scale out by increasing 1 instance per 5 minutes. When I stopped, the message received will be decreasing, and then the anntools farm will scale in. Finally, there will still have two instances running.

Filter: honglibu-auto

1 to 1 of 1 Auto Scaling Groups

Name	Launch Configuration /	Instances	Desired	Min	Max	Availability Zones	Default Cooldown	Health Cl
honglibu-auto-…	honglibu-launch-config	9	8	2	10	us-east-1d	300	300

Filter: honglibu-annotate

1 to 1 of 1 Auto Scaling Groups

Name	Launch Configuration /	Instances	Desired	Min	Max	Availability Zones	Default Cooldown	Health Cl
honglibu-annot…	honglibu-annotate-laun…	4	4	2	10	us-east-1d	300	300