AUTO SCALING

Automatically adjust your service’s desired count up and down in response to CloudWatch alarms

A very interesting feature! **Will automatically increase or decrease the number of running tasks according to demand**. This will only work if EC2 instances in the cluster are able to run multiple containers otherwise the maximum number of tasks will be limited to the maximum number of EC2 instances.

As an Application type Elastic Load Balancer was implemented in this solution, allowing to have multiple tasks per EC2 instance, we can maximize the resources by giving the containers the minimum amount of resources necessary for correct operation. If demand increases, ECS will auto-scale by adding containers to satisfice the requests. If requests decreases, EBS will stop the unnecessary containers.

CloudWatch alarms can be configured with actions to scale up or down our service. For instance, we can define that if memory is > 95%, ECS should remove 1 running task, to make sure server doesn't run out of memory. (More experimentation needed)

AUTO SCALING DISABLED - ROUND ROBIN

If auto scaling is disabled and the load in each container is similar, ELB will send each request to the next container, in a round-robin fashion. (Yes, but why? Research)

IMPORTANT: SEE DOC AWECS INTEGRATION