Auto Scaling

- If We need to setup 3 instances for an application then good balancing will be considered if all AZs have almost equal number of instances . For 10 it should be 4 , 3 , 3 etc .
- Auto scaling group tries to balance the instances in all availability zones of the region . Suppose there are 8 , 3 , 2 instances running in 3 AZs . Now How Auto scaling group manages it to balance fully . It firstly tries to increase the running instances in the AZ where number of instances running are less than balance number of instances . For example In the above example firstly 8 , 3, 2 will become 8 , 3, 4 and then 6 , 3, 4 . After that 6 , 4, 4 and then 5 , 4, 4 . Now it would be a perfect balance .
- We define min and max EC2s in Auto scaling group . But during balancing as it increases the instances first in the AZ where there were less number of instances and then removes from the zone where there were more number of instances . This can cause max number of ec2s in an auto scaling group to

be more than max . This is problem . Hence we define policy like max(10% or 1) of max limit can be extra . For example if a ASG has 20 as maximum then it means 2 instances can be extra during rebalancing . If at any point in this group having max 20 has 10 , 3, 3 then 10 , 3, 5 —> 8 , 3, 5 —> 8 , 5, 5, —> 6 , 5, 5 . But if we need to add extra ec2 manually then this addition is not possible . Max rule can be breached only in the case of balancing .

- ASG by default considers health check returned by EC2 only . But ASG must be configured to use health Check done by ELB as well as if ELB considers some instances as faulty then it will stop sending traffic there but AGS will not be aware about it . ASG would think that the traffic is still balanced . Hence ASG must be configured to use health check by ELB as well .
- Grace period of health check is 5 minutes it means once the EC2 instance is launched its status will not be considered before 5 minutes . Only after 5 minutes its status will be considered . Otherwise it can have cascading effect of EC2 starts .
- Auto scaling group will consider an instance

unhealthy after grace period in below situations .

- If status of EC2 is reported other than running . Terminated , standby etc will be considered as faulty .
- If ELB reports that instance is OOS then OSG will consider it as unhealthy .
- In case of rebalancing we increase the instance first but in case of unhealthy instance we decrease the instance first by removing the unhealthy instance .

- EC2 Instances can be added to auto scaling group in below conditions
- EC2 Instance must be in running stage not terminated or stopped .
- AMI used to create EC2 instance still exists otherwise how aws will be able to launch the same instance .
- One instance can be connected with one Auto scaling group .
- Auto scaling group is made on region level and instances from only those sub nets can be added which are attached with the auto scaling group .

- Instances can be removed manually from auto scaling group . Once detached it can be attached with other group also .But this has to be done manually .
- While detaching desired EC2 instances value should also be decreased as once one machine is detached a new one will be added automatically .

- If we delete an ASG then every thing will be become ZERO min , max , desired . Hence it will delete all the EC2s .
- If we don’t want all instances to be killed then first detach the EC2s which you don’t want to be deleted then delete the ASG .
- Once ELB is attached with an already running ASG then all instances need to be registered . But after that newly created instances will be auto registered and deregistered.

- Elastic IP must be release manually for the unhealhty removed instance .
- Auto Scaling policy can be of two types Manual and automatic .
- Manual means min , max and desired is 4 . It means it will always be 4 . If we need to increase or decrease it then it has to be

handled manually .

-
- In Dynamic there are three types
- Target tracking , simple and step scaling policy .
- Target Tracking means (For example) my CPU utilisation must be max 70 % . Now suppose 4 machines has 60 , 60 , 60 , 60 . Then if I remove the 4th one then all the three will be having 80% utilisation . It means I cannot remove 4th . But Now suppose all the three have 50 , 50 , 50 , 50 percent utilisation. Then 4th one can be removed . Now again suppose it increased to 80% . Now If I create 4th one then all 4 will have 60% . Hence one more should be created . Min max can be breached . It has warmup period of default 5 minutes .
- Simple scaling policy means if my CPU utilisation reaches 70 then add 2 more instances . It will not think how much are needed . It has cool down period means when new machines are launched then it will not consider alarms coming in a configurable period (5 min default) . After that if it gets the alarms that the cpu utilisation is not balanced then it will create more instances . But cool

down period will be used . Min max cannot be breached .Even if EC2 instance was ready and handling traffic in 30 seconds it will still wait for 5 minutes and ignore all the alarms . It has cool down period of 5 minutes default . It can be configured that if CU remains >70% for continuous 5 minutes for n number of times then create instances .

- Step scaling is like increase 1 machine if CU is 40 - 50 % , Create 2 when CU is 50-60 % . It don’t have any cool down period . Min max cannot be breached . It has warm up period means it will wait until EC2 instance is not ready . If it is ready in 30 seconds then it will start accepting alarms . It has warmup period not cool down period .
- Predictive scaling uses machine learning techniques to identify when you need more scale . Suppose on Saturday Sunday if you have more traffic then it will scale out during this period.