**1. What is Amazon EC2 Auto Scaling?**

**Answer:** Amazon EC2 Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application. You create collections of EC2 instances, called Auto Scaling groups. You can specify the minimum number of instances in each Auto Scaling group, and Amazon EC2 Auto Scaling ensures that your group never goes below this size. You can specify the maximum number of instances in each Auto Scaling group, and Amazon EC2 Auto Scaling ensures that your group never goes above this size. If you specify the desired capacity, either when you create the group or at any time thereafter, Amazon EC2 Auto Scaling ensures that your group has this many instances. If you specify scaling policies, then Amazon EC2 Auto Scaling can launch or terminate instances as demand on your application increases or decreases.

For example, the following Auto Scaling group has a minimum size of one instance, a desired capacity of two instances, and a maximum size of four instances. The scaling policies that you define adjust the number of instances, within your minimum and maximum number of instances, based on the criteria that you specify.

**2. Amazon EC2 Auto Scaling benefits**

**Answer:** Adding Amazon EC2 Auto Scaling to your application architecture is one way to maximize the benefits of the AWS Cloud. When you use Amazon EC2 Auto Scaling, your applications gain the following benefits:

Better fault tolerance. Amazon EC2 Auto Scaling can detect when an instance is unhealthy, terminate it, and launch an instance to replace it. You can also configure Amazon EC2 Auto Scaling to use multiple Availability Zones. If one Availability Zone becomes unavailable, Amazon EC2 Auto Scaling can launch instances in another one to compensate.

Better availability. Amazon EC2 Auto Scaling helps ensure that your application always has the right amount of capacity to handle the current traffic demand.

Better cost management. Amazon EC2 Auto Scaling can dynamically increase and decrease capacity as needed. Because you pay for the EC2 instances you use, you save money by launching instances when they are needed and terminating them when they aren't.

**3. How Does AWS Auto Scaling work?**

**Answer:** Configure a single unified scaling policy per application source.

Explore the application and create a system that adds and removes EC2 instances in response to the requirement.

Choose the service that you want to scale up or down.

Select what to optimize. Based on a schedule, scale your application in response to predictable load changes.

Keep tracing the scaling load and maintain a steady count of instances.

**4. What are the Benefits of Auto Scaling?**

**Answer:** **IMPROVE FAULT TOLERANCE**

Amazon EC2 Auto Scaling can detect when an instance is unhealthy, terminate it, and replace it with a new one.

**INCREASE APPLICATION AVAILABILITY**

Amazon EC2 Auto Scaling ensures that your application always has the right amount of compute, and also proactively provisions capacity with Predictive Scaling.

**LOWER COSTS**

Amazon EC2 Auto Scaling adds instances only when needed, and can scale across purchase options to optimize performance and cost.

**5. How does Auto Scaling works?**

**Answer: Fleet Management**

Whether you are running one Amazon EC2 instance or thousands, you can use Amazon EC2 Auto Scaling to detect impaired Amazon EC2 instances and unhealthy applications, and replace the instances without your intervention. This ensures that your application is getting the compute capacity that you expect. Amazon EC2 Auto Scaling will perform three main functions to automate fleet management for EC2 instances:

**Monitor the health of running instances**

Amazon EC2 Auto Scaling ensures that your application is able to receive traffic and that EC2 instances are working properly. Amazon EC2 Auto Scaling periodically performs health checks to identify any instances that are unhealthy.

**Replace impaired instances automatically**

When an impaired instance fails a health check, Amazon EC2 Auto Scaling automatically terminates it and replaces it with a new one. That means that you don’t need to respond manually when an instance needs replacing.

**Balance capacity across Availability Zones**

Amazon EC2 Auto Scaling can automatically balance instances across zones, and always launches new instances so that they are balanced between zones as evenly as possible across your entire fleet.

**6. What are types of Auto Scaling?**

**Answer:** **Scheduled Scaling**

Scaling based on a schedule allows you to scale your application ahead of known load changes. For example, every week the traffic to your web application starts to increase on Wednesday, remains high on Thursday, and starts to decrease on Friday. You can plan your scaling activities based on the known traffic patterns of your web application.

**Dynamic Scaling**

Amazon EC2 Auto Scaling enables you to follow the demand curve for your applications closely, reducing the need to manually provision Amazon EC2 capacity in advance. For example, you can use target tracking scaling policies to select a load metric for your application, such as CPU utilization. Or, you could set a target value using the new “Request Count Per Target” metric from Application Load Balancer, a load balancing option for the Elastic Load Balancing service. Amazon EC2 Auto Scaling will then automatically adjust the number of EC2 instances as needed to maintain your target.

**Predictive Scaling**

Predictive Scaling, now natively supported as an EC2 Auto Scaling policy, uses machine learning to schedule the right number of EC2 instances in anticipation of approaching traffic changes. Predictive Scaling predicts future traffic, including regularly-occurring spikes, and provisions the right number of EC2 instances in advance. Predictive Scaling’s machine learning algorithms detect changes in daily and weekly patterns, automatically adjusting their forecasts. This removes the need for manual adjustment of Auto Scaling parameters as cyclicality changes over time, making Auto Scaling simpler to configure. Auto Scaling enhanced with Predictive Scaling delivers faster, simpler, and more accurate capacity provisioning resulting in lower cost and more responsive applications.