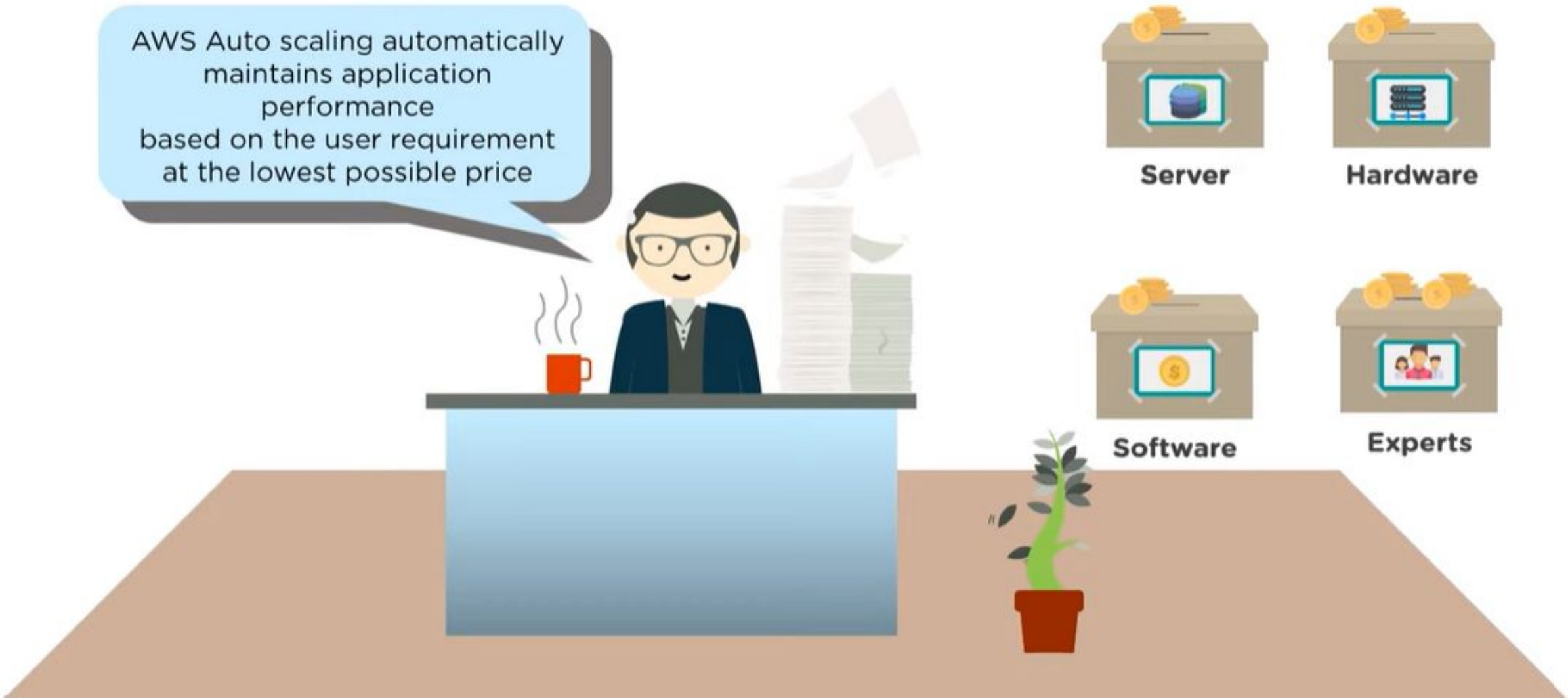
Autoscaling

Cloud scaling is the process of **adding or removing cloud computing resources as you need them**. Cloud infrastructure and platforms are typically priced on a utility model such that you are billed for what you use at a finely grained level. This allows cloud services, systems and applications to be quickly scaled up and down to meet demands.

After AWS Cloud Scaling



What Is AWS Cloud Scaling?

AWS Auto Scaling is a service which helps user to monitor applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost.

Snapshots

- It is used as a backup of a single EBS volume attached to the EC2 instance
- Opt for this when the instance contains multiple static EBS volumes
- Here, pay only for the storage of the modified data
- It is a non-bootable image on EBS volume

AMI

- It is used as a backup of an EC2 instance
- This is widely used to replace a failed EC2 instance
- Here, pay only for the storage that you use
- It is a bootable image on EC2 instance

Snapshots vs AMIs

Snapshots

- It is used as a backup of a single EBS volume attached to the EC2 instance
- Opt for this when the instance contains multiple static EBS volumes
- Here, pay only for the storage of the modified data
- It is a non-bootable image on EBS volume

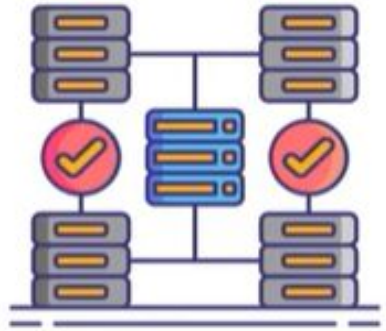
However, creating an AMI image will also create EBS snapshots

How Does AWS Auto Scaling Work?

1. Configure single unified scaling policy per application source

2. Explore application

3. Choose the service you want to scale



4. Select what to optimize

- Cost
- Performance

5. Keep track of scaling

Different Scaling Plans

- Scaling strategy guides the service of AWS Auto Scaling on how to optimize resources in an application
- With scaling strategy, users can create their own strategy based on the required metrics and thresholds

What's an Auto Scaling Group?

- In real-life, the load on your websites and application can change
- In the cloud, you can create and get rid of servers very quickly

- The goal of an Auto Scaling Group (ASG) is to:
  - Scale out (add EC2 instances) to match an increased load
  - Scale in (remove EC2 instances) to match a decreased load
  - Ensure we have a minimum and a maximum number of machines running
  - Automatically Register new instances to a load balancer

Types of Scaling

- Manual Scaling
- Scheduled Scaling
- Dynamic Scaling

Benefits of AWS Auto Scaling

Better fault tolerance

Better cost management

Reliability

Scalability

Flexibility

Better availability