AWS Cloud EC2 Pricing

AWS EC2 Pricing

With AWS EC2 you pay for the compute time.

You only pay for the compute time that you use.

It offers different pricing options.

On Demand Instances

On Demand Instances are best used for short-term workloads.

It requires no upfront costs or minimum amount on the purchase.

The instances run until you stop them.

You pay for what you use.

AWS EC2 Savings Plan

The savings plan is a commitment for usage over a 1-year or 3-year term.

Committing to a period gives a discounted price.

If you surpass the budget, the cost goes to normal (on-demand) prices.

You will learn more about AWS Cost Explorer later in this tutorial.

AWS Cost Explorer is a tool that helps to plan usage with AWS Cloud.

Reserved Instances

The Reserved Instances are used to reserve instances for an agreed period.

The options are for 1-year or 3-years. The latter one gives the highest discount.

Spot Instances

This pricing model is best for workloads with flexible start and end times, which can take interruptions.

Spot instances can give up to a 90% cost savings.

The reason behind the discount is that AWS can optimize its capacity, giving you better prices.

Dedicated Hosts

Dedicated hosts are physical servers fully dedicated to you.

You can use your existing VM software licenses.

The Dedicated Host is the most expensive model.

AWS Cloud EC2 Scaling

AWS EC2 Scaling

Scaling is about only using the resources that you need.

In addition, have the flexibility to grow freely.

Make sure to have an architecture that can handle changes in demand.

Designing a scalable architecture allow you to only pay for the resources that you need at any given time.

AWS EC2 Auto Scaling

Servers can get more requests than they can handle.

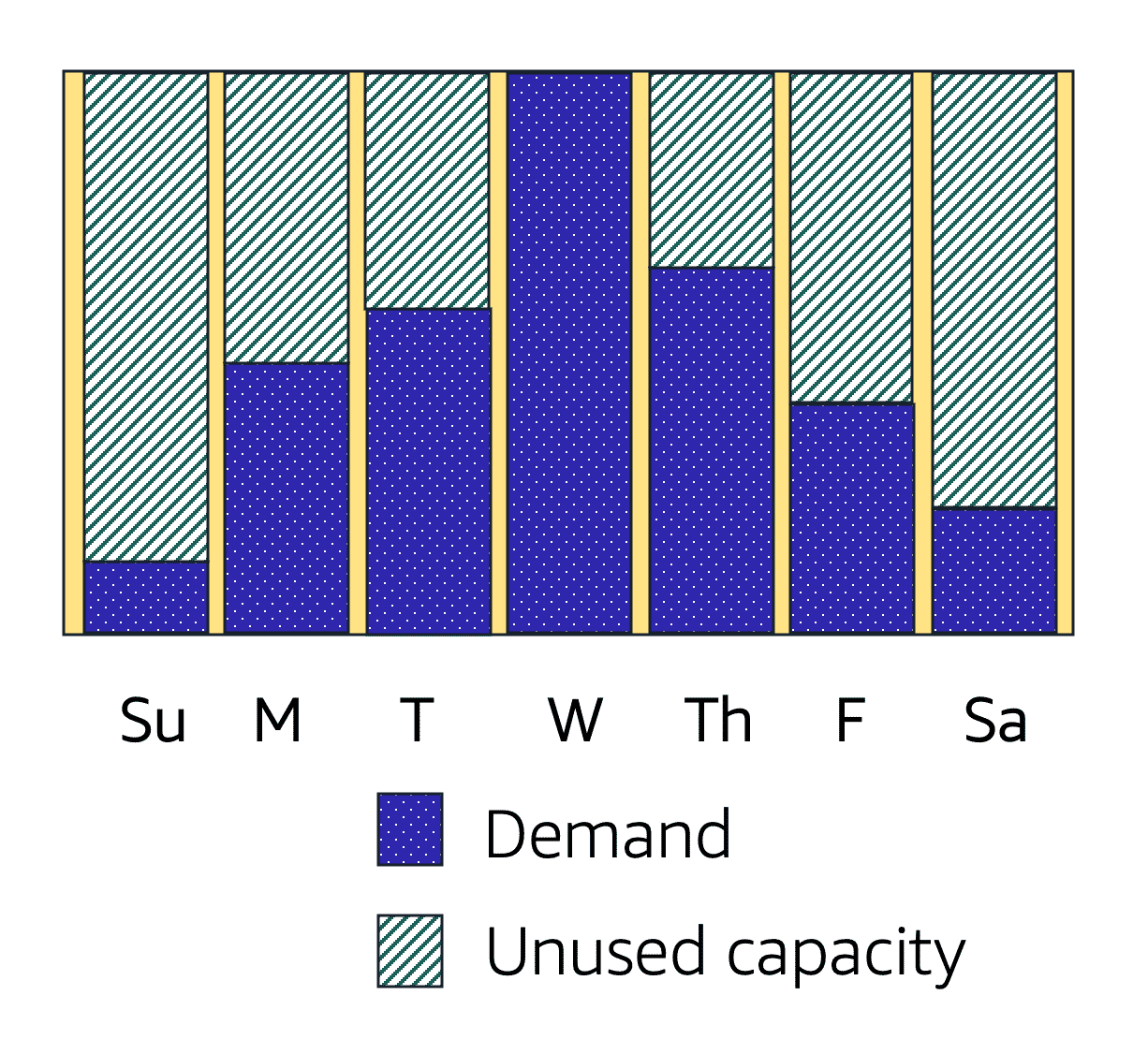
Too many requests can cause timeouts and outages.

AWS EC2 Auto Scaling allows you to add or remove EC2 instances automatically.

It automates the capacity to the demand.

There are two approaches:

* Dynamic scaling: responds to changing demand
* Predictive scaling: schedules the number of instances based on a predicted demand
* Dynamic and Predictive scaling can be combined to scale faster

Image created by Amazon Web Service

The picture illustrates that demand can change during a week based.

AWS Cloud Auto Scaling

AWS EC2 Auto Scaling

EC2 Auto Scaling can be added as a buffer on top of your instances.

It can add new instances to the application when necessary and terminate them when no longer needed.

You can set up a group of instances.

Here you can set a **minimum capacity** of instances that will always be running. The rest will operate when necessary.

You can set the **desired number** of AWS EC2 instances in the scaling group.

However, the desired capacity defaults to your minimum capacity if not specified.

The last configuration is **Maximum capacity**.

Here you set the maximum capacity of instances to be used.

The Auto Scaling groups allow you to have a dynamic environment.

You set the minimum capacity, the desired number, and the maximum capacity.

The group will operate within the config and give you a predictable and cost-effective architecture.