Amazon Elastic Cloud Compute (EC2): EC2 Price Optimizations

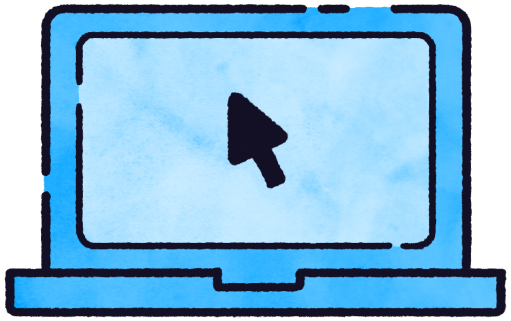
Andru Estes
Principal Author
andru-estes

Reserved Instances and Capacity Reservations

Amazon EC2 Review

On-demand, scalable computing service offered by AWS. Get what you need, when you need it, and only pay for what you use.

On-Demand Pricing



On-Demand

Pay for compute capacity by the hour or second (minimum of 60 seconds) without any long-term commitment

Baseline Cost

On-Demand pricing is the baseline pricing for all compute, meaning it will be the most expensive option (*per hour*)

AWS has introduced multiple options to obtain savings on your EC2 pricing.

One option is called Reserved Instances.

Reserved Instances Concepts

Allows up to 72% savings on EC2 on-demand prices

1-year or 3-year time commitment

No Upfront, Partial Upfront, All Upfront payment options

Choose if you want a Zonal (AZ) or Regional Reserved Instance

Convertible option allows you to change the instance type, family, OS, and scope

Perfect for long-running, critical applications

Less of a discount if you use this option!

You can buy and sell Reserved Instances within the AWS Marketplace!

However, Reserved Instances are slowly being phased out for Savings Plans, which are discussed later in this module.

Image Source: https://unsplash.com/

Capacity Reservations

Allow you to reserve compute capacity for your EC2 instances in a specific AZ for any duration of time

Charged for On-demand pricing even if you are not running an instance!

Perfect for:

- Business-critical workloads
- Apps requiring specific long or short-term capacity

Savings Plans

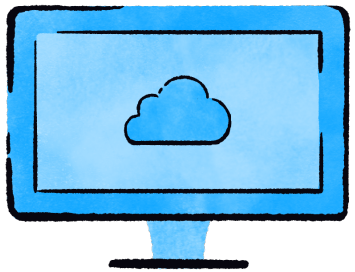
Savings Plans

“Savings Plans offer a flexible pricing model that provides savings on AWS usage.”

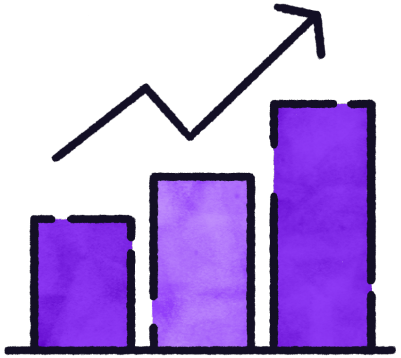
Citation: https://docs.aws.amazon.com/savingsplans/latest/userguide/what-is-savings-plans.html

The successor to using Reserved Instances for specific categories of compute!

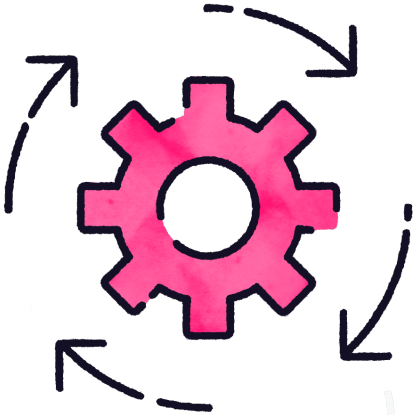
Savings Plan Types



Compute Savings Plans

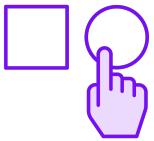


EC2 Instance Savings Plans



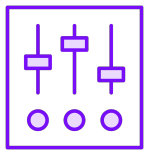
SageMaker Savings Plans

Compute Savings Plans

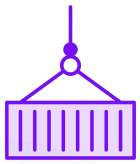


Offer the most flexibility for choosing different instances

Prices can be up to 66% off compared to On-Demand pricing



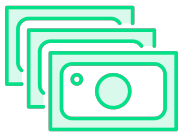
Automatically apply to EC2 instance usage regardless of family, size, Region, OS, and tenancy



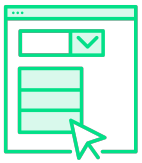
Also includes the following services: AWS Lambda and Fargate

EC2 Instance Savings Plans

Committed to a specific instance family in a chosen AWS Region



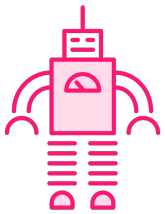
Provide the most savings, offering up to 72% off of On-Demand pricing



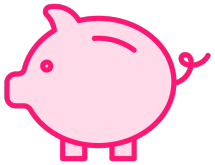
You can only change size, OS, and tenancy

These provide the most savings, but you give up a lot of flexibility!

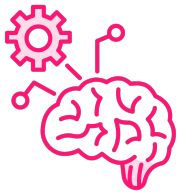
SageMaker Savings Plans



Automatically apply to your SageMaker instance usage regardless of family, size, Region, OS, and tenancy



Provide up to 64% off of On-Demand pricing for SageMaker instances



Only relevant to when you need to save on costs specific to the Amazon SageMaker service!

Image Source: https://unsplash.com/

Savings Plans Concepts

Gain savings by committing to usage of compute over a span of time

Offered in 1-year and 3-year commitments

Similar to Reserved Instances for payment options:

- No Upfront (*least savings*)
- Partial Upfront
- All Upfront (*most savings*)

Cost saving scenarios involving long-term usage for static compute will likely involve one of these plans!

Dedicated Hosts and Instances

Two Dedicated Options

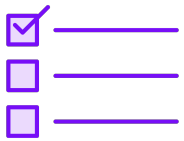
Dedicated Hosts

Physical server that is fully dedicated for your use

Dedicated Instances

EC2 instances that run on hardware dedicated to a single AWS account

Dedicated Hosts

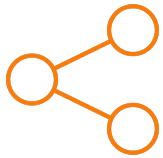


Allow for better, stronger compliance requirements needs

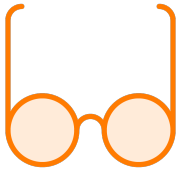
Help address binding software license issues (BYOL) related to servers

The most expensive option for running EC2 compute (Per-host)

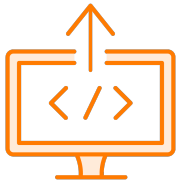
Dedicated Instances



Instances do not run on hardware shared with other AWS accounts



No visibility/control over instance placement, and no host affinity



Partially supports Bring Your Own License (BYOL) requirements

Choosing Dedicated Hosts versus Dedicated Instances comes down to the amount of control you need!

Purchasing Spot Instances

Spot Instance

An instance that uses spare EC2 capacity that is available for much less than the On-Demand price (up to 90% off)

Spot Instance Concepts

Hourly price for a Spot Instance is called a Spot price

Spot price of each instance type in each AZ is set by Amazon EC2

Spot price is adjusted gradually based on long-term supply and demand

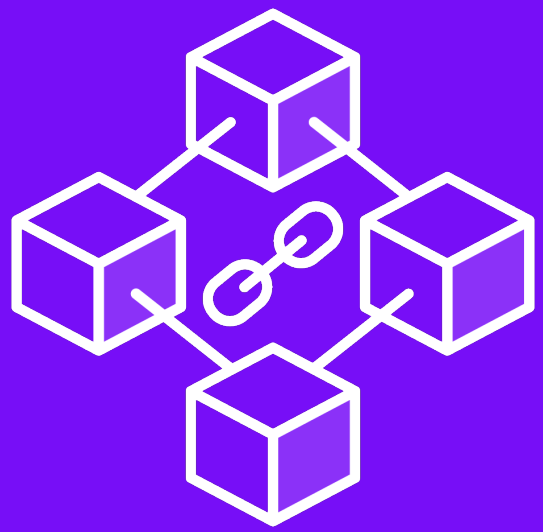
Only run whenever spot capacity is actually available

Perfect, cost-effective choice for flexible applications

Spot Instances are subject to interruption (shutting down) without a guaranteed notice...

AWS attempts to give you a 2-minute warning before a spot instance will be reclaimed, but it is not guaranteed.

This means you should NEVER run important applications or workloads on Spot Instances!



Spot Blocks

A spot instance that you can specify to run for 1-6 hours and better avoid interruptions. More expensive than their Spot Instance counterparts.

EC2 Spot Fleets

Used to launch a fleet of tens, hundreds, or thousands of Amazon EC2 instances in a single operation

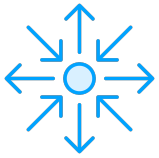
Launch mix of:

- Multiple instance types
- Distributing across multiple AZs
- Using combination of On-Demand and Spot Prices
- Automatic replacement of Spot instances

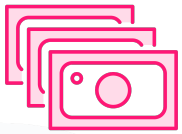
EC2 Spot Fleet Allocation Strategies

Price capacity optimized: Chooses the pools for instances first based on capacity, and then the lowest price available. Recommended approach.

Capacity optimized: Select the pools that have the best amount of capacity for the Spot instances.

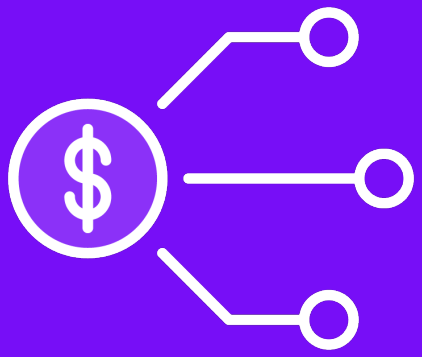


Diversified: Spot instances get distributed across all Spot capacity pools. Better for availability requirements.



Lowest Price: Chooses the pools for instances based on the lowest price available. Comes with the highest risk of interruption!

Reducing Spend Using AWS Compute Optimizer

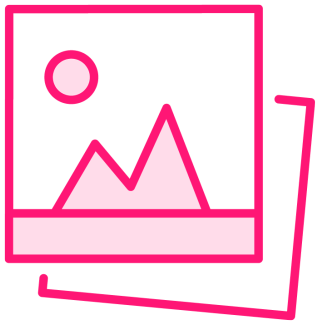


AWS Compute Optimizer

Service that uses machine learning to recommend optimal AWS Compute resources for your workloads to help reduce costs and improve performance

What Does AWS Compute Optimizer Do?

Analyzes configurations and utilization metrics of your AWS resources



Reports current usage optimizations and potential recommendations

Provides graphical history data and projected utilization metrics

Use graphs, metric data, and recommendations for moving or resizing resources

What AWS Resources Work with It?

- EC2 Instances
- EBS Volumes
- Elastic Container Store (ECS) services
- AWS Lambda Functions
- Amazon Relational Database Service (RDS)

You must opt-in and enable this feature to use it!

Module Summary and Exam Tips

EC2 Instance Pricing Options

On-demand: Pay by the hour or the second, depending on the type of instance you run. Great for flexibility. Get what you need, when you need it. They will always be available.

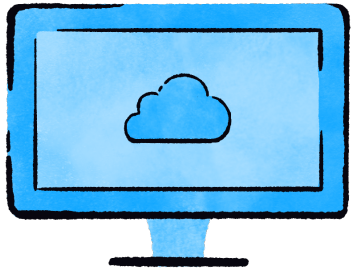
Spot: Purchase unused capacity at a discount of up to 90%. Prices fluctuate with supply and demand. Great for applications with flexible start and end times. Do not run critical workloads on these.

Reserved: Reserved capacity for 1 or 3 years. Up to 72% discount on the hourly charge. Great if you have known, fixed requirements, but they are being phased out.

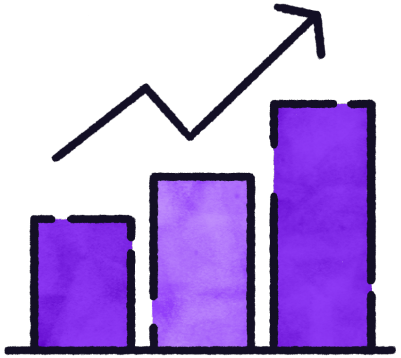
Dedicated: Instances that do not run on hardware shared with other AWS accounts. Partially supports BYOL requirements. The most expensive instance option!

Spot Blocks allow you to run EC2 for 1-6 hours and better avoid interruptions, but they are more expensive than Spot Instances.

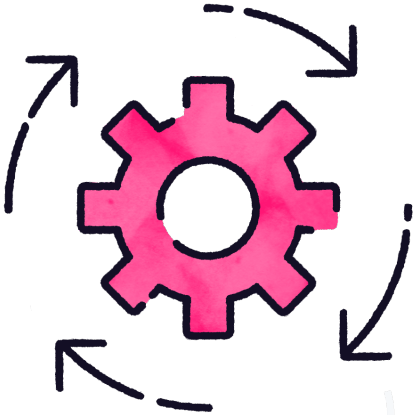
Savings Plans Types



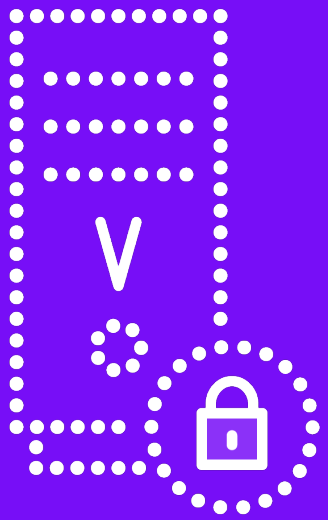
Compute Savings Plans
Most flexible offering for EC2, Fargate, and Lambda. 66% discounts.



EC2 Instance Savings Plans
Only EC2 coverage. Offers 72% discounts.



SageMaker Savings Plans
SageMaker-specific. 64% discounts



Dedicated Hosts

Physical server with EC2 instance capacity fully dedicated to your use. Perfect for scenarios about using existing per-socket, per-core, or per-VM software licenses.

AWS Compute Optimizer

Recommends optimal AWS compute and EBS resources for your workloads

Perfect for scenarios about analyzing the best compute resources for your workloads to optimize costs

Comparing EC2 Pricing Examples

M5a.large instance type running Linux in us-east-1

Purchasing Option	Price (*USD per hour*)
On-Demand	$0.0860 hourly
Reserved Instance - 1 year All Upfront	$0.0506 hourly
Reserved Instance - 3 year All Upfront	$0.0323 hourly
EC2 Instance Savings Plan - 1 Year	$0.051 hourly
EC2 Instance Savings Plan - 3 Year	$0.032 hourly
Spot Instance	$0.0372 hourly