



LONG-TERM


Guide

Optimizing your cloud usage: A guide to hyperscale reservations





Table of contents

- 03 Introduction
 - 03 Why reserved instances were introduced by cloud providers
 - 04 AWS reserved instances
 - 05 AWS reserved instances vs. AWS savings plans
 - 06 Azure reserved instances
 - 08 Google Cloud zonal reservations, committed use and sustained use discounts
 - 08 The reality of using reserved capacity pricing models
- 



Reserved Instances are

mutually beneficial

for both the consumer and the public cloud provider.

Introduction

In 2009 AWS (Amazon Web Services) [introduced](#) a new EC2 pricing model known as reserved instances. In exchange for an upfront commitment of 1 or 3 years, reserved instances offer significant cost savings (anywhere from ~29% up to 72%) on cloud compute resources compared with on-demand instances. Today other cloud providers such as Microsoft Azure and Google Cloud Platform also offer similar pricing constructs.

Why reserved instances were introduced by cloud providers

Since AWS launched in 2006, developers have enjoyed the ease of spinning up and down a wide variety of EC2 instances, which were reasonably priced, for periodic on-demand usage.

While this new-found flexibility helped drive faster innovation for both enterprises and start-ups (no more waiting on IT to provision a server or being forced to shelve projects due to the prohibitively high cost of buying physical servers), longer-term usage was still expensive at the existing on-demand rates and this limited EC2 adoption to shorter-term projects. Additionally, from Amazon's perspective, having a more predictable amount of demand for EC2 in the form of long-term reservations or commitments, would make it much easier for internal capacity planning and financial forecasting.

To remedy this, Amazon introduced EC2 reserved instances in early 2009. It would be several more years before other cloud providers would offer similar pricing models, but as of 2020, all offer a full range of cloud products and services, including reserved pricing models.

AWS reserved instances

AWS offers a broad range of reserved instance types and flavors. Let's take a look at the main criteria that define EC2 reserved instances.

Commitment length: AWS reserved instances entail either a one or three year commitment.

Standard or convertible class: These [two main classes or categories of reserved instances](#) each have pros and cons. For example, standard reservations can be resold on the [AWS reserved instance marketplace](#) (which provides a flexible off-ramp if your capacity needs change) but cannot be applied to a different instance family type than the one originally specified. On the other hand, convertible reservations cannot be resold, but they can be modified and applied to different family types.

Regional or zonal scope: The scope of both standard and convertible reservations can be further fine-tuned with [regional and zonal reservations](#). The latter provides guaranteed capacity but zero flexibility in terms of modifying the reserved instance's zone, family and size, and the former provides modification flexibility but no guarantees on capacity availability.

Flexibility: Regional reservations are "size flexible" and are automatically applied to different instance sizes within the same instance family. So for example if you purchased one c5.xlarge reserved instance, but instead started running two c5.large instances, the c5.xlarge reservation would be applied to the 2 smaller instances.

Instance type: With the broad range of instance families that AWS offers, [reserved instance pricing](#) is initially defined and applied to a particular instance family. As mentioned previously, you can change the family type for convertible reservations but not for standard reservations.

Payment options: Reserved instances can be paid upfront fully, partially or with no upfront payment. The level of savings will vary with greater levels of commitment providing greater savings.

EC2 reservations used to come with guaranteed capacity. That's no longer the case unless you select zonal reservations.



Convertible reservations

allow for extensible contracting so you can add RIs to cover additional instances without having to start new 1 or 3-year terms.

Platforms supported: The reserved instances are typically associated with Linux and Windows servers.

Services covered: AWS has expanded the usage of reserved instances beyond EC2 and now offers this pricing model for additional services such as RDS, ElastiCache, Redshift and DynamoDB.

AWS reserved instances vs. AWS savings plans

In 2019, AWS introduced Savings Plans (EC2 Savings Plans and Compute Savings Plans). With this new committed capacity pricing plan customers commit to spend any desired amount per hour (e.g. \$35/hour, for either 1 or 3 years). In this example, anything spent up to \$35 will be charged per Savings Plans rates (between 66-72% savings). Any spend that exceeds the committed amount will be charged at on-demand rates.

As always, each option has benefits and limitations. Here are some important ones to consider before choosing the best pricing model for you.

- Standard reserved instances can be bought and sold on the AWS Marketplace allowing for greater flexibility, while Savings Plans cannot, requiring you to keep your committed spend at the level you defined.
- Savings Plans can only be applied to EC2, Lambda, Fargate, and Sagemaker while reserved instances have broader applications for EC2, RDS, Redshift, ElastiCache, etc.
- Amazon offers three Savings Plans types: Compute, EC2 Instance, and Amazon Sagemaker. Compute Savings Plans offer the most flexibility and apply to EC2, Lambda, and Fargate regardless of instance family, size, region, OS, or tenancy. Compared to other savings plans, EC2 Instance Savings Plans offer the largest savings and apply to individual instance families in the same region. And lastly, Sagemaker Savings Plans are a flexible pricing model for Sagemaker usage.



While unused AWS EC2 standard reservations can be sold off in the AWS Marketplace, there is no Azure RI Marketplace.

Azure RIs can be canceled in amounts up to \$50,000 within a 12-month rolling window.

- With convertible reserved instances, you can increase commitment (i.e. add additional reservations to cover more EC2 Instances) during the contracted term without the need to increase the term period. This is especially helpful when needs change and committing to a new 1 to 3-year term doesn't make sense. With Savings Plans, any addition to the original contract is done with a new contract that starts from day 0.
- EC2 Instance Savings Plans will apply usage across any given instance family, regardless of OS or tenancy. Standard reserved instances can also apply to usage across any given instance type but require the instances to be Linux and default tenancy.
- Convertible reserved instances are scoped to a specific instance type, OS tenancy and region while Compute Savings Plans will apply across all of your usage types in multiple regions.

Azure reserved instances

[Microsoft Azure reserved instances](#) or RIs are similar to AWS's offering with commitment lengths of 1 or 3 years and potential savings of up to 72%. However, there is sufficient nuance to warrant reviewing the main characteristics of Azure's RIs.

Azure account and subscription: Before diving in, it's essential to understand these terms. An Azure account is similar to AWS's master payer account and Azure subscriptions are similar to AWS's sub-accounts. The Azure account is used for overall billing purposes, but each subscription generates its own set of billing data that is helpful for cost allocation purposes. Additionally, each subscription can be used for access control and isolating environments.

Scope: Reserved instances can be either "single" and applied to VMs (or other services) in a specific subscription, they can be "shared" and applied to reservations in any other subscriptions that you are the owner of and that share the same billing context as the reservation, or they can be applied to a resource group, which is a container that holds related resources for an Azure solution.

Flexibility: When buying an Azure reserved instance you can choose “instance size flexibility” which ensures that the reservation will be applied to any size VM in the same instance size flexibility group.

For example, if you bought a reservation for a Standard_DS4_v2 with a footprint of 8 but you actually ran two Standard_DS2_v2 sized VMs with a footprint of 2 each and a Standard_DS3_v2 sized VM with a footprint of 4, your initial reserved instance purchase would fully cover what you actually consumed.

Prioritized capacity: For reserved instances that have a “single” scope you can select “Capacity priority” which as the name implies, prioritizes data center capacity for your deployments. This option is similar to AWS’s zonal reserved instances.

Services covered: Azure provides a very [detailed list](#) of where reservations can be applied. Aside from Azure VMs, some of the services covered include Blob storage capacity, the compute component of Azure Database for MariaDB, MySQL, and PostgreSQL, as well as quite a few other Azure services.

Exchanging or cancelling reservations: If you no longer need the reserved instance you purchased you may exchange it for a different instance family, region, etc. You can also cancel up to \$50,000 worth of reservations, currently without any penalty. [Read more about Azure’s exchange and refund terms.](#)

Azure hybrid benefit: For customers with Windows Server and SQL Server on-premises licenses, [running them on Azure with RIs](#) can deliver up to 80% cost savings.

Google Cloud zonal reservations, committed use and sustained use discounts



Unique pricing model

Google offers a unique pricing model called sustained use where you receive discounts without prior commitment, provided you use the VMs for at least 25% of the month.

Google has taken a slightly different approach to long-term cloud compute consumption. While they offer the ability to [reserve guaranteed capacity](#) in an availability zone of your choice, you still could end up paying on-demand rates unless you specifically opt for their [committed use discounts](#). In this pricing models you commit to a specific amount of usage (e.g. vCPUs, memory, GPUs, local SSDs, etc.) for 1 or 3 years and in return receive a discount of anywhere from 57% to 70%.

Google also offers a unique pricing model called [sustained use](#) where you can receive automatic discounts of up to 30% without any need for prior commitment. However this requires that you run your workloads anywhere from 25%-100% of the month. Less than 25% overall usage will not trigger any discount.

The reality of using reserved capacity pricing models

At the end of the day, even with all the flexibility, the cloud providers have built into their reserved capacity pricing, you still are to a large degree financially locked-in for the term of your commitment.

Selling off your reservations on the AWS reserved instance marketplace requires you to find a buyer. AWS Savings Plans cannot be sold on the AWS marketplace at all. For Azure reservations, you can only get refunded on up to \$50K worth of RIs and Azure may start imposing early termination fines in the near term.

As such, if your project suddenly ends or resource requirements change, in all likelihood you will need to pay for the compute capacity you signed up for or figure out how to best recycle the committed capacity.



>> See how Eco helps FinOps

professionals deliver more savings with less commitment.

Even if the changes in your environment are not so drastic, managing reserved capacity across multiple accounts or subscriptions and ensuring that your investment is being properly utilized is a full-time job.

All this necessitates very careful planning and [ongoing assessment of actual cloud compute usage](#) across your environments.

To address the growing need for reserved cloud capacity management and optimization, Eco by Spot has become the go-to solution for FinOps professionals.

Currently available for AWS and Azure, Eco intelligently manages the entire Reserved Instance lifecycle for a hands-free, risk-free, and high-ROI engagement. With a comprehensive analysis of historic and projected consumption, Eco generates custom strategies for every deployment and fully implements them by buying a blend of shorter-term reservations alongside standard and convertible reservations as well as Savings Plans, with 1 or 3-year terms as appropriate. This saves time and effort and improves the overall utilization of your reserved capacity. In the event of unused reservations, Eco actively identifies and offloads them in the AWS Marketplace, recouping the original investment, and eliminating the issue of financial loss and lock-in. In short, more savings, less commitment.

Going beyond cloud analytics and recommendations, Spot actively optimizes AWS, Azure and Google Cloud deployments with SLA-backed availability, fully automated infrastructure management and up to 90% cost reduction. With Spot, cloud consumers can effortlessly and affordably scale any workload, from stateful, single instances, to cloud-native clusters made up of thousands of nodes.

For more information: www.spot.io >>

