EC2

Compute – Charged per hour while instance is running, if stop – no charge

Virtual h/w dedicated to instance –Instance type

Software loaded – AMI (Amazon machine image)

Instance types – grouped into families (Example m4 family – good choice for many applications.)

Family

c4 Compute optimized—For workloads requiring significant processing

r3 Memory optimized—For memory-intensive workloads

i2 Storage optimized—For workloads requiring high amounts of fast SSD storage

g2 GPU-based

Network performance -  low, moderate, high..

Enhanced networking.

Enhanced networking reduces the impact of virtualization on network performance by enabling a capability called Single Root I/O Virtualization (SRIOV). This results in more Packets Per Second (PPS), lower latency, and less jitter

The Amazon Machine Image (AMI) – OS + Patch + Application/software

X86, Linux and Windows

* Published by AWS— Rhel, ubuntu, Amazon linux distribution, windows
* The AWS Marketplace— published by amazon partners.., additional charge for the software..
* Generated from Existing Instances— customized images + updates + security..
* Uploaded Virtual Servers— using AWS import/Export service – vmdk, ova – customer responsibility to be compliant with licenses.

Accessing Instance

Public DNS Name –

* + generated automatically and cannot be specified by the customer
  + persists only while the instance is running and cannot be transferred to another instance

Public IP

* IP address is assigned from the addresses reserved by AWS and cannot be specified.
* This IP address is unique on the Internet, persists only while the instance is running

Elastic IP

* + unique on the Internet
  + you reserve independently and associate with an Amazon EC2 instance
  + persists until the customer releases it and is not tied to the lifetime or state of an individual instance

Amazon EC2 uses public-key cryptography to encrypt and decrypt login information.

Public-key cryptography uses a public key to encrypt a piece of data and an associated private key to decrypt the data. These two keys together are called a key pair.

AWS stores the public key, and the private key is kept by the customer

the public key is stored in the ∼/.ssh/authorized\_keys

default user – ec2-user

Windows instance, Amazon EC2 generates a random password for the local administrator account and encrypts the password using the public key.

Initial access to the instance is obtained by decrypting the password with the private key, either in the console or through the API. The decrypted password can be used to log in to the instance with the local administrator account via RDP.

Virtual firewall (Security groups)

* allows you to control traffic in and out of your instances
* SG rule attributes - Port/protocol/(source/destination)
* Every instance will have one SG - default deny
* 2 rules - aggregation
* Types

EC2-Classic Security Groups - Control outgoing instance traffic

VPC Security Groups - Control outgoing and incoming instance traffic

* A security group is a stateful firewall; that is, an outgoing message is remembered so that the response is allowed through the security group without an explicit inbound rule being required.
* Security groups are applied at the instance level,

**The Lifecycle of Instances**

**Bootstrapping -** configure instances and install applications programmatically when an instance is launched. The process of providing code to be run on an instance at launch is called *bootstrapping*.

One of the parameters when an instance is launched is a string value called *UserData*. – shell script, batch/powershell script

Patch + AD join + install software + copy script from storage+ agent install + set role

**VM Import/Export**

Company premises VM – can be imported to AWS and then exported back

Export not allowed on VM created using AWS AMI

**Instance Metadata**

*Instance metadata* is data about your instance that you can use to configure or manage the running instance

AWS instance properties + OS level information can be retrieved using API - <http://169.254.169.254/latest/meta-data/>

[ec2-user@ip-172-31-31-227 ~]$ curl http://169.254.169.254/latest/meta-data/

ami-id

ami-launch-index

ami-manifest-path

block-device-mapping/

hostname

instance-action

instance-id

instance-type

local-hostname

local-ipv4

mac

metrics/

network/

placement/

profile

public-hostname

public-ipv4

public-keys/

reservation-id

security-groups

services/[ec2-user@ip-172-31-31-227 ~]$ curl http://169.254.169.254/latest/meta-data/security-groups

default[ec2-user@ip-172-31-31-227 ~]$ curl http://169.254.169.254/latest/meta-data/hostname

ip-172-31-31-227.us-west-2.compute.internal[ec2-user@ip-172-31-31-227 ~]

**Managing Instances**

Tags are key/value pairs you can associate with your instance or other service.

Tags can be used to identify attributes of an instance like project, environment (dev, test, and so on), billable department, and so forth.

You can apply up to 10 tags per instance.

**Monitoring Instances**

CloudWatch - provides monitoring and alerting for Amazon EC2 instances, and also other AWS infrastructure

**Modifying an Instance**

**Instance Type – update resources**

**Security Groups  - only instances within Amazon VPC, not outside (EC2-classic) after launch**

**Terminate – stop/remove, to prevent delete – termination protection can be enabled**

Note that this just protects from termination calls from the AWS Management Console, CLI, or API.

It does not prevent termination triggered by an OS shutdown command, termination from an Auto Scaling group (discussed in Chapter 5), or termination of a Spot Instance due to Spot price changes

**Pricing Options**

Per hour , in running state

**On-Demand Instances**

* The price per hour for each instance type published on the AWS website represents the price for *On-Demand Instances*
* Most flexible, no upfront commitment, customer has control over when instance is launched and when to terminate
* Disadvantage - Least cost effective among others
* Pay more, more control

**Reserved Instances**

* Enable customers to reserve capacity for predictable workloads
* Capacity is reserved
* Price 75% less then on-demand
* Cost depends on:
  + Commitment – 1 or 3 years
  + Payment option – All upfront (No monthly), Partial upfront (Rest monthly), no upfront (All monthly payments)

**Spot Instances**

* For workloads that are not time critical and are tolerant of interruption
* Great discount
* customers specify the price they are willing to pay for a certain instance type. When the customer’s bid price is above the current Spot price, the customer will receive the requested instance(s)
* customer will only pay the Spot price for the hours that instance(s) run
* The instances will run until:
  + The customer terminates them.
  + The Spot price goes above the customer’s bid price.
  + There is not enough unused capacity to meet the demand for Spot Instance
  + If Amazon EC2 needs to terminate a Spot Instance, the instance will receive a termination notice providing a two-minute warning prior to Amazon EC2 terminating the instance
  + Usecase - analytics, financial modeling, big data, media encoding, scientific computing, and testing.

Per Second Billing

With per-second billing, you pay for only what you use.

It takes cost of unused minutes and seconds in an hour off of the bill, so you can focus on improving your applications instead of maximizing usage to the hour. Especially, if you manage instances running for irregular periods of time, such as dev/testing, data processing, analytics, batch processing and gaming applications, can benefit.

EC2 usage are billed on one second increments, with a minimum of 60 seconds. Similarly, provisioned storage for EBS volumes will be billed per-second increments, with a 60 second minimum. Per-second billing is available for instances launched in:

* On-Demand, Reserved and Spot forms
* All regions and Availability Zones
* Amazon Linux and Ubuntu

**Architectures with Different Pricing Models**

* workload using a mix of On-Demand and Reserved Instances

**Tenancy Options**

To achieve security and compliance goals

**Shared Tenancy**

* default tenancy model
* single host machine may house instances from different customers
* AWS does not use overprovisioning and fully isolates instances from other instances on the same host, this is a secure tenancy model

**Dedicated Instances**

* Dedicated Instances run on dedicated hardware to a single customer. As a customer runs more Dedicated Instances, more underlying hardware may
* Non-dedicated workloads will go on shared clusters, isolated from dedicated instances be dedicated to their account

**Dedicated Host**

Dedicated Host is a physical server with Amazon EC2 instance capacity fully dedicated to a single customer’s use

Dedicated Hosts can help you address licensing requirements and reduce costs by allowing you to use your existing server-bound software licenses.

The customer has complete control over which specific host runs an instance at launch

This differs from Dedicated Instances in that a Dedicated Instance can launch on any hardware that has been dedicated to the account

**Placement Groups**

A *placement group* is a logical grouping of instances within a single Availability Zone. Placement groups enable applications to participate in a low-latency, 10 Gbps network.

Placement groups are recommended for applications that benefit from low network latency, high network throughput, or both.

Remember that this represents network connectivity between instances.

To fully use this network performance for your placement group, choose an instance type that supports enhanced networking and 10 Gbps network performance.

**Running Commands on Your Linux Instance at Launch**

When you launch an instance in Amazon EC2, you have the option of passing user data to the instance that can be used to perform common automated configuration tasks and even run scripts after the instance starts. You can pass two types of user data to Amazon EC2: shell scripts and cloud-init directives.

# Launching an EC2 Fleet

An EC2 Fleet contains the configuration information to launch a fleet—or group—of instances. In a single API call, a fleet can launch multiple instance types across multiple Availability Zones, using the On-Demand Instance, Reserved Instance, and Spot Instance purchasing models together.

The EC2 Fleet attempts to launch the number of instances that are required to meet the target capacity specified in your request. The fleet also attempts to maintain its target Spot capacity if your Spot Instances are interrupted due to a change in Spot prices or available capacity

If you have Reserved Instances and you specify On-Demand Instances in your fleet, EC2 Fleet uses your Reserved Instances. For example, if your fleet specifies an On-Demand Instance as c4.large, and you have Reserved Instances for c4.large, you receive the Reserved Instance pricing.

**Spot Fleet** – A set of Spot Instances that is launched based on criteria that you specify. The Spot Fleet selects the Spot Instance pools that meet your needs and launches Spot Instances to meet the target capacity for the fleet. By default, Spot Fleets are set to maintain target capacity by launching replacement instances after Spot Instances in the fleet are terminated. You can submit a Spot Fleet as a one-time request, which does not persist after the instances have been terminated. You can include On-Demand Instance requests in a Spot Fleet request.

**Spot Instance pool** – A set of unused EC2 instances with the same instance type, operating system, Availability Zone, and network platform (EC2-Classic or EC2-VPC).

## Allocation Strategy for Spot Instances

The allocation strategy for the Spot Instances in your Spot Fleet determines how it fulfills your Spot Fleet request from the possible Spot Instance pools represented by its launch specifications. The following are the allocation strategies that you can specify in your Spot Fleet request:

lowestPrice

The Spot Instances come from the pool with the lowest price. This is the default strategy.

diversified

The Spot Instances are distributed across all pools.

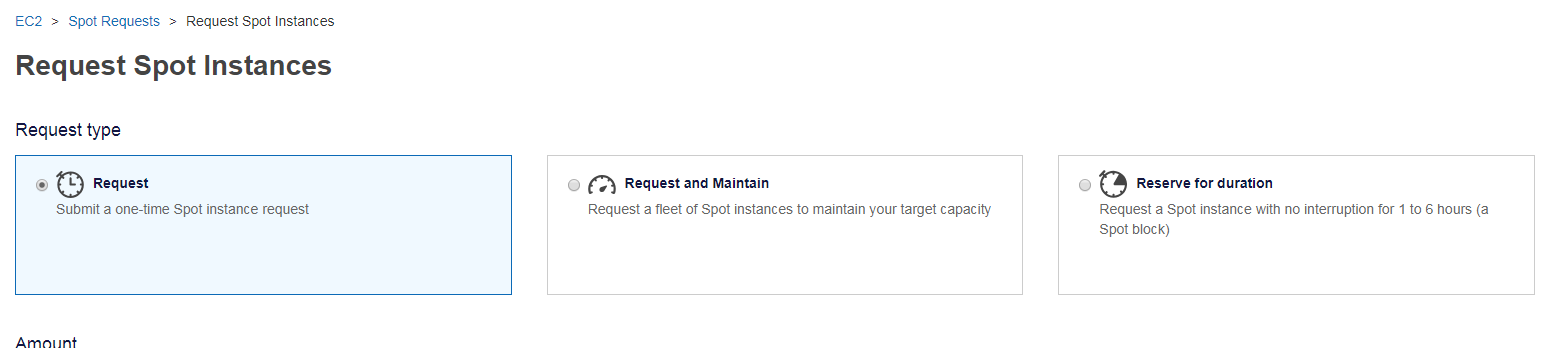
InstancePoolsToUseCount

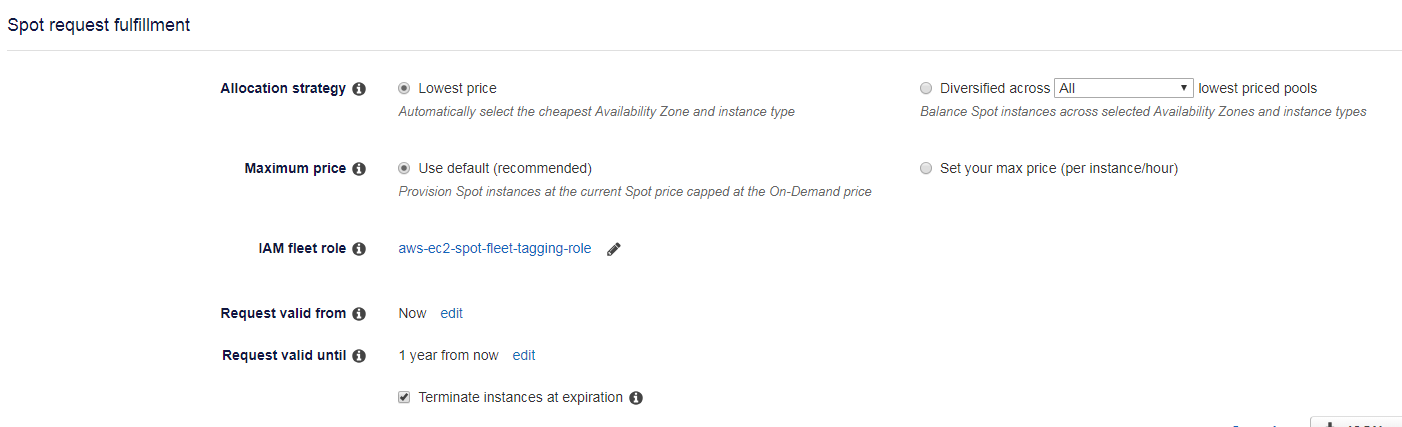
The Spot Instances are distributed across the number of Spot pools that you specify. This parameter is valid only when used in combination with lowestPrice.

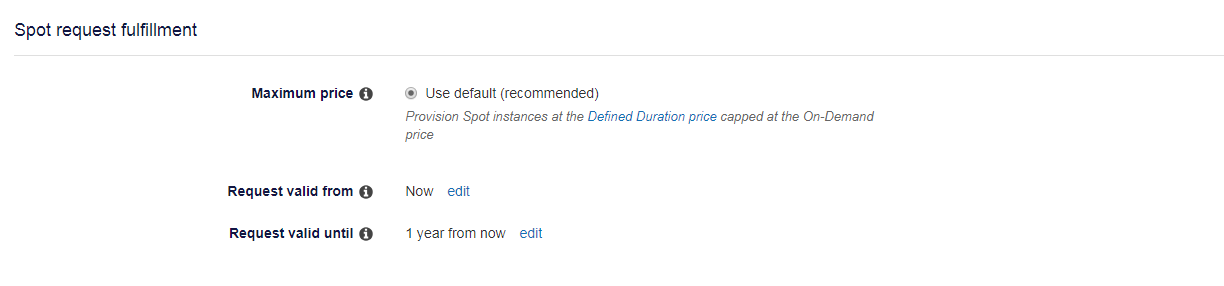
## Specifying a Duration for Your Spot Instances

Amazon EC2 does not terminate Spot Instances with a specified duration (also known as Spot blocks) when the Spot price changes. This makes them ideal for jobs that take a finite time to complete, such as batch processing, encoding and rendering, modeling and analysis, and continuous integration.

You can specify a duration of 1, 2, 3, 4, 5, or 6 hours. The price that you pay depends on the specified duration.







# inux AMI Virtualization Types

Linux Amazon Machine Images use one of two types of virtualization: paravirtual (PV) or hardware virtual machine (HVM). The main differences between PV and HVM AMIs are the way in which they boot and whether they can take advantage of special hardware extensions (CPU, network, and storage) for better performance.

## Cross-Region AMI Copy

Copying an AMI across geographically diverse regions provides the following benefits:

* Consistent global deployment: Copying an AMI from one region to another enables you to launch consistent instances in different regions based on the same AMI.
* Scalability: You can more easily design and build global applications that meet the needs of your users, regardless of their location.
* Performance: You can increase performance by distributing your application, as well as locating critical components of your application in closer proximity to your users. You can also take advantage of region-specific features, such as instance types or other AWS services.
* High availability: You can design and deploy applications across AWS regions, to increase availability.

When you first copy an instance store-backed AMI to a region, we create an Amazon S3 bucket for the AMIs copied to that region.

## Cross-Account AMI Copy

You can share an AMI with another AWS account. Sharing an AMI does not affect the ownership of the AMI.

The following table shows encryption support for various scenarios. Note that while it is possible to copy an unencrypted snapshot to yield an encrypted snapshot, you cannot copy an encrypted snapshot to yield an unencrypted one.

|  |  |  |
| --- | --- | --- |
| **Scenario** | **Description** | **Supported** |
| 1 | Unencrypted-to-unencrypted | Yes |
| 2 | Encrypted-to-encrypted | Yes |
| 3 | Unencrypted-to-encrypted | Yes |
| 4 | Encrypted-to-unencrypted | No |

Q. What price will I pay for a Spot instance?

You pay the Spot price that’s in effect at the beginning of each instance-hour for your running instance. If Spot price changes after you launch the instance, the new price is charged against the instance usage for the subsequent hour.

Q. Which operating systems are available as Spot instances?

Linux/UNIX and Windows Server are available. Windows Server with SQL Server is not currently available.

. Q. Can I use a Spot instance with a paid AMI for third-party software (such as IBM’s software packages)?

Not at this time.

Q. When would my Spot instance get interrupted?

Over the last 3 months, 92% of Spot instance interruptions were from a customer manually terminating the instance because the application had completed its work.

In the circumstance EC2 needs to reclaim your Spot instance it can be for two possible reasons, with the primary one being Amazon EC2 capacity requirements (e.g. On Demand or Reserved Instance usage). Secondarily, if you have chosen to set a “**maximum Spot price”** and the Spot price rises above this, your instance will be reclaimed with a two-minute notification. **This parameter determines the maximum price you would be willing to pay for a Spot instance hour, and by default, is set at the On-Demand price.** As before, you continue to pay the Spot market price, not your maximum price, at the time your instance was running, charged in per-second increments.

Q. What is the difference between Stop and Hibernate interruption behaviors?

In the case of Hibernate, your instance gets hibernated and the RAM data persisted. In the case of Stop, your instance gets shutdown and RAM is cleared.

Q. What if my EBS root volume is not large enough to store memory state (RAM) for Hibernate?

You should have sufficient space available on your EBS root volume to write data from memory. If the EBS root volume does not enough space, hibernation will fail and the instance will get shutdown instead. Ensure that your EBS volume is large enough to persist memory data before choosing the hibernate option.

Q. How will I be charged if my Spot instance is interrupted?

If your Spot instance is terminated or stopped by Amazon EC2 in the first instance hour, you will not be charged for that usage. However, if you terminate the instance yourself, you will be charged to the **nearest second**. If the Spot instance is terminated or stopped by Amazon EC2 in any subsequent hour, you will be charged for your usage to the nearest second. **If you are running on Windows and you terminate the instance yourself, you will be charged for an entire hour.**

Q. How will I be charged if my Spot instance is interrupted?

If your Spot instance is terminated or stopped by Amazon EC2 in the first instance hour, you will not be charged for that usage. However, if you terminate the instance yourself, you will be charged to the nearest second. If the Spot instance is terminated or stopped by Amazon EC2 in any subsequent hour, you will be charged for your usage to the nearest second. If you are running on Windows and you terminate the instance yourself, you will be charged for an entire hour.

Q: Are Spot blocks (Fixed Duration Spot instances) ever interrupted?

Spot blocks are designed not to be interrupted and will run continuously for the duration you select, independent of Spot market price. In rare situations, Spot blocks may be interrupted due to AWS capacity needs. In these cases, we will provide a two-minute warning before we terminate your instance ([termination notice](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-interruptions.html#spot-instance-termination-notices)), and you will not be charged for the affected instance(s).