AWS

Elastic Compute Cloud (EC2)

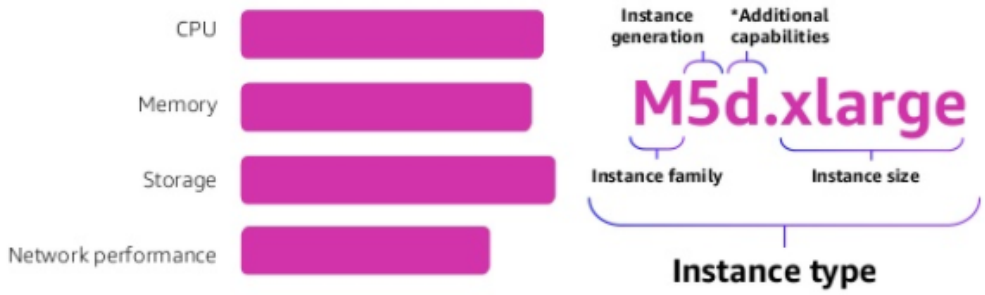
what is it?

EC2 is a service that provides virtual machines in the cloud where you only pay for the capacity you use and choose from 'families' of instance types that are good for different use cases.

what do the letters and numbers mean?

- **Family** – different instance types with resources for different use cases
- **Generation** – AWS phase out older technologies and bring in new ones with more resources using these numbers to show which is which
- **Size** – resources go up in a linear fashion, as well as the price that goes with it

Diagram below from 2018 re:Invent EC2 Fundamentals slides



AWS

Elastic Compute Cloud (EC2)

how do i pick an instance type?

EC2 comes in variety instance types specialised for different roles:

- **General Purpose** - balanced compute, memory, and networking resources
- **Compute Optimised** - ideal for compute-bound applications that benefit from the high performance processor
- **Memory-Optimised** - fast performance for workloads that process large data sets in memory
- **Accelerated Optimised** - hardware accelerators, or co-processors
- **Storage Optimised** - high, sequential read and write access to very large data sets on local storage

General Purpose	Compute Optimised	Memory Optimised	Accelerated Computing	Storage Optimised
A1 ARM based core and custom silicon	C4 Compute - CPU intensive apps and DBs	R4 RAM - Memory intensive apps and DB's	P2 Processing optimised- Machine Learning	H1 High Disk Throughput - Big data clusters
T2 Tiny - Web servers and small DBs		X1 Xtreme RAM - For SAP/Spark	G3 Graphics Intensive - Video and streaming	I3 IOPS - NoSQL DBs
M4 Main - App servers and general purpose		z1d High Compute and High Memory - Gaming	F1 Field Programmable - Hardware acceleration	D2 Dense Storage - Data Warehousing

AWS

Elastic Compute Cloud (EC2)

payment options

On-Demand

- Pay for capacity by per hour or second
- No commitment
- Good for apps being developed or with unpredictable usage spikes

Reserved Instances

- Provides a reservation at 75% off the On-Demand price
- Gives you the ability to launch instances when you need them
- Reduced price as you need to commit to one or three-year terms and decide if you will pay all upfront, partial upfront, or no upfront

Spot Instances

- Bid for spare capacity for up to 90% off the On-Demand price
- Flexible start and end times
- If you're outbid the instance is terminated and you don't pay for the hour
- If you stop the instance you will pay for the hour
- Good for those background jobs which aren't critical

Dedicated Hosts

- Provides capacity on dedicated physical servers
- Good for when can't share capacity due to regulatory reasons or for licensing requirements

Savings Plan

- Provides the benefits of Reserved Instances but with more flexibility
- You will need to commit to a one or three year term but can change instance type within the same family while taking advantage of savings