EC2 - Elastic Compute Cloud

EC2 what is it, how it works and everything else on EC2.

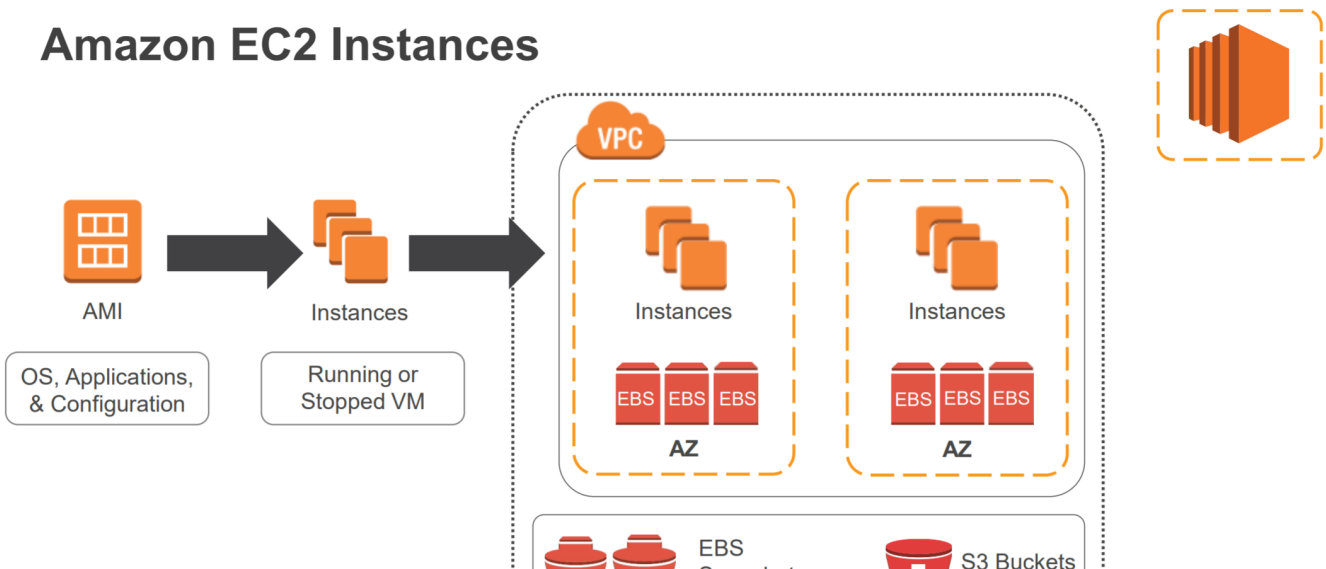
Elastic Compute Cloud (EC2) - Is a web service that provider re-sizable compute capacity in the cloud.

AWS EC-2 reduces the time required to obtain and boot new server instance to minutes allowing you to quickly scale capacity, both up and down, as your computing requirement charge.

In the old world if you needed an app server or db server. You would first need to table to your developer, size what you needed how many cores and then talk to your procurement team and by the time the server is in your data center it could take 2+ months.

Now with EC2, you have this with a click of a button. This also means from a startup perspective you do not have the upfront cost of buying all the hardware you need.

EC2 changes the economics of computing by allowing you to pay only computing by allowing you to pay only for the capacity that you actually use. You also have several tools at the disposal of the developers to build the applications to be resilient and isolate them from failure scenarios.



In the above architecture diagram, you manage Identity and access control via IAM and deploy some EC2 instances in your VPC. These instances are typically your app servers that then would store data in one or more EBS within one Availability Zone.

You would then send periodic EBS snapshots to your S3 buckets that get replicated at a regional level. You could use this architecture for any simple 3 tire architectural that has AZ level redundancy.

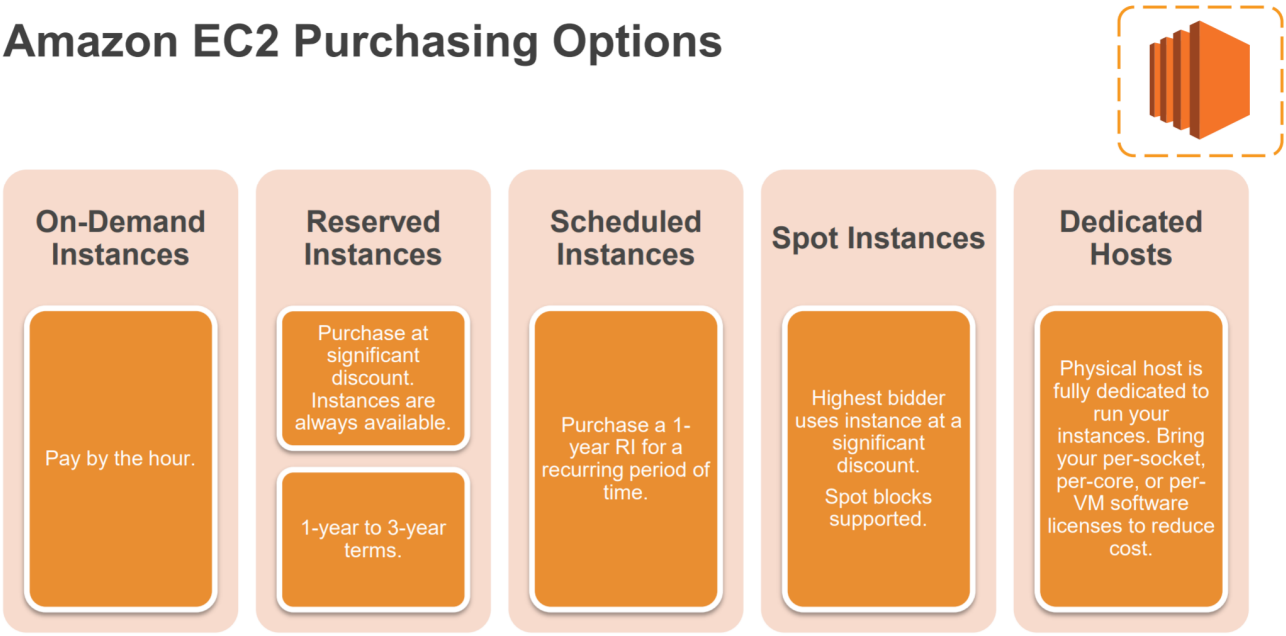
EC2 Purchase Options

On-demand – Fixed rate by the hour or by the second with no commitment

Reserved Instance - You reserve an EC2 instance for your use this gives you a discount. 1 year or 3-year teams

Spot – Enables you to bid whatever price you want to pay for an instance capacity. This is used if your application has flexible start and stop times. e.g. Big Data computation that is not time sensitive

Dedicated Hosts - Can be purchased on-demand, used for regulatory requirements that may not support multi-tenant virtualization



EC2 Types

EC2 Instance Types

Family	Specialty	Use Case
D2	Dense Storage	Fileservers/Data Warehousing/Hadoop
R4	Memory Optimized	Memory Intensive Apps/DBs
M4	General Purpose	Application Servers
C4	Compute Optimized	CPU Intensive Apps/DBs
G2	Graphics Intensive	Video Encoding/ 3D Application Streaming
I2	High Speed Storage	NoSQL DBs, Data Warehousing etc.
F1	Field Programmable Gate Array	Hardware acceleration for your code
T2	Lowest Cost, General Purpose	Web Servers/Small DBs
P2	Graphics/General Purpose GPU	Machine Learning, Bit Coin Mining etc.
X1	Memory Optimized	SAP/HANA/Apache Spark etc.

EC2 Instance Types