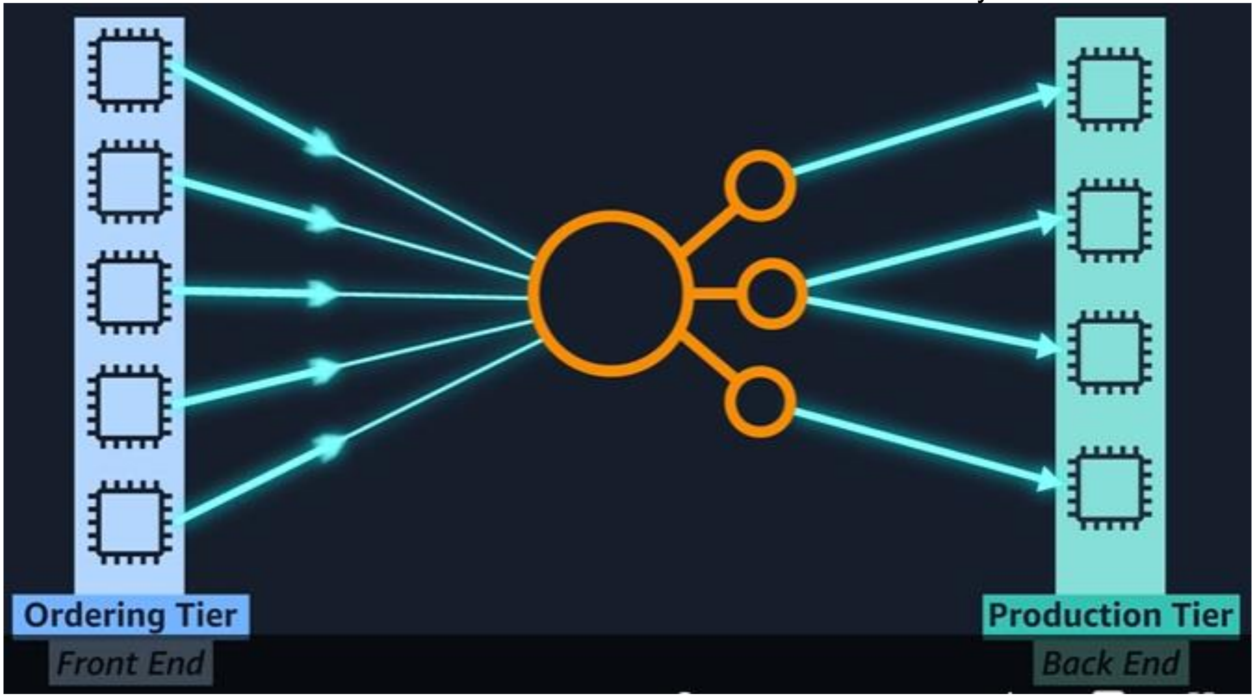
Elastic Compute Cloud (EC2)

>EC2 Overview

This is the most fundamental service in AWS. This is a MUST since it is in and out of the exam. This is a secure, resizable compute capacity in the cloud. Like a VM, only hosted in AWS instead of your own data center. Gives you the capacity you want when you need it. You're in complete control over it. EC2 is essentially a service that provides secure, resizable compute capacity in the cloud.

EC2 allows you to rent and manage virtual servers in the cloud.

- Elastic compute power
    - You can stretch an elastic band far beyond its resting state. But part of what makes it truly elastic is the fact that, when you let go of it, it immediately returns to its original size. The reason the word elastic is used in the names of so many AWS services (Elastic Compute Cloud, Elastic Load Balancing, Elastic Beanstalk, and so on) is because those services are built to be easily and automatically resized.
    - Elastic load balancing
        - Elastic Load Balancing is the AWS service that automatically distributes incoming application traffic across multiple resources, such as Amazon EC2 instances.
        - Think of it as a balancer that "slows" but doesn't really slow down traffic.



The key is choosing the right tool for the right job. As traffic grows, the ELB grows/scales it is designed for the throughput.

A load balancer acts as a single point of contact for all incoming web traffic to your Auto Scaling group. This means that as you add or remove Amazon EC2 instances in response to the amount of incoming traffic, these requests route to the load balancer first. Then, the requests spread across multiple resources that will handle them. For example, if you have multiple

Amazon EC2 instances, Elastic Load Balancing distributes the workload across the multiple instances so that no single instance has to carry the bulk of it.

- A load balancer distributes incoming application traffic across multiple EC2 instances in multiple Availability Zones. This increases the fault tolerance of your applications. Elastic Load Balancing detects unhealthy instances and routes traffic only to healthy instances.

Servers are physical compute hardware running in a data center. Which EC2 instances are the virtual servers running on these physical servers. Instances are not considered serverless. Because they actually exist on servers somewhere in the data center.

With over 500 instances and choice of the latest processor, storage, networking, operating system, and purchase model to help you best match the needs of your workload. We are the first major cloud provider that supports Intel, AMD, and Arm processors, the only cloud with on-demand EC2 Mac instances, and the only cloud with 400 Gbps Ethernet networking.

EC2 Real World
Deploy a database to EX2 gives you full control over the database. Whereas deploying a web application allows multiple AZs to make the web application highly available.

You can access this through AWS management console, SSH, EC2 Instance connect, AWS Systems Manager

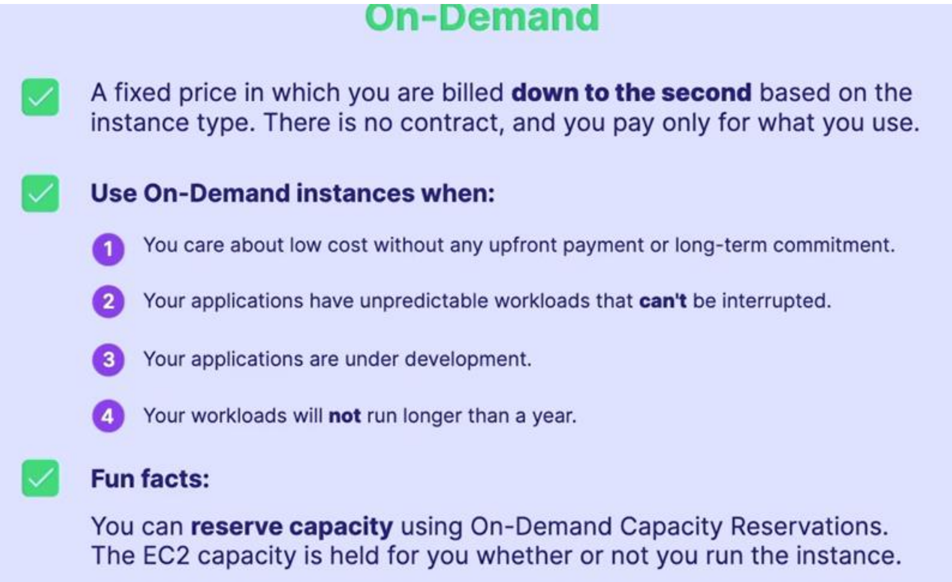
- AWS management console
  - You're able to configure and manage your instances via web browser.
- SSH
  - SSH allows a secure connection
- EC2 Instance connect
  - EIC allows you to use IAM policies to control SSH access to your instances, removing the need to manage SSH keys
- AWS Systems Manager
  - Allows you to manage your EC2 instances via a web browser or AWS CLI

The most common way to connect to Linux EC2 instances is via SSH. which the first thing you'll do is generate a key pair which consists of a private key and a public key. Proves your identity when connecting to an EC2

> EC2 Pricing options

Several pricing options to choose from your EC2 instances

- On-Demand



On-Demand

✓ A fixed price in which you are billed **down to the second** based on the instance type. There is no contract, and you pay only for what you use.

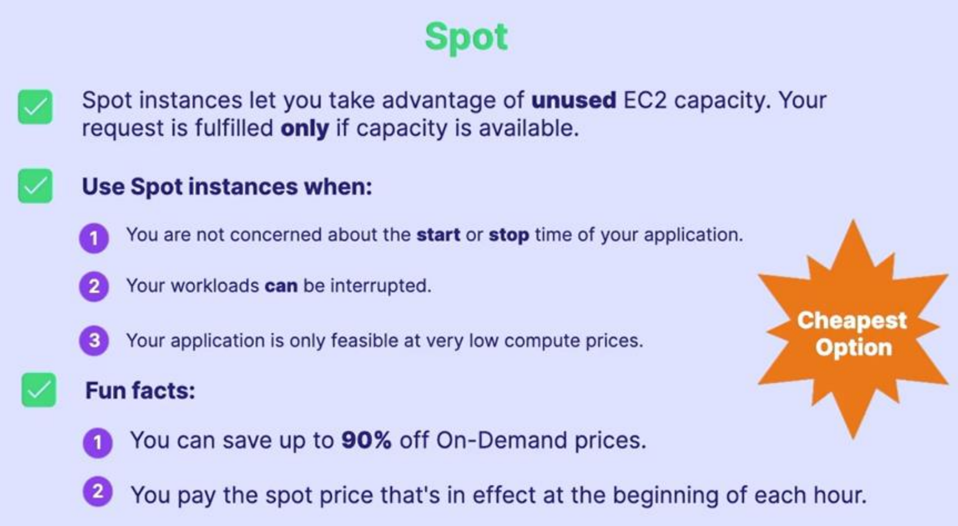
✓ **Use On-Demand instances when:**

1. You care about low cost without any upfront payment or long-term commitment.
2. Your applications have unpredictable workloads that **can't** be interrupted.
3. Your applications are under development.
4. Your workloads will **not** run longer than a year.

✓ **Fun facts:**

You can **reserve capacity** using On-Demand Capacity Reservations. The EC2 capacity is held for you whether or not you run the instance.

- Spot



Spot

✓ Spot instances let you take advantage of **unused** EC2 capacity. Your request is fulfilled **only** if capacity is available.

✓ **Use Spot instances when:**

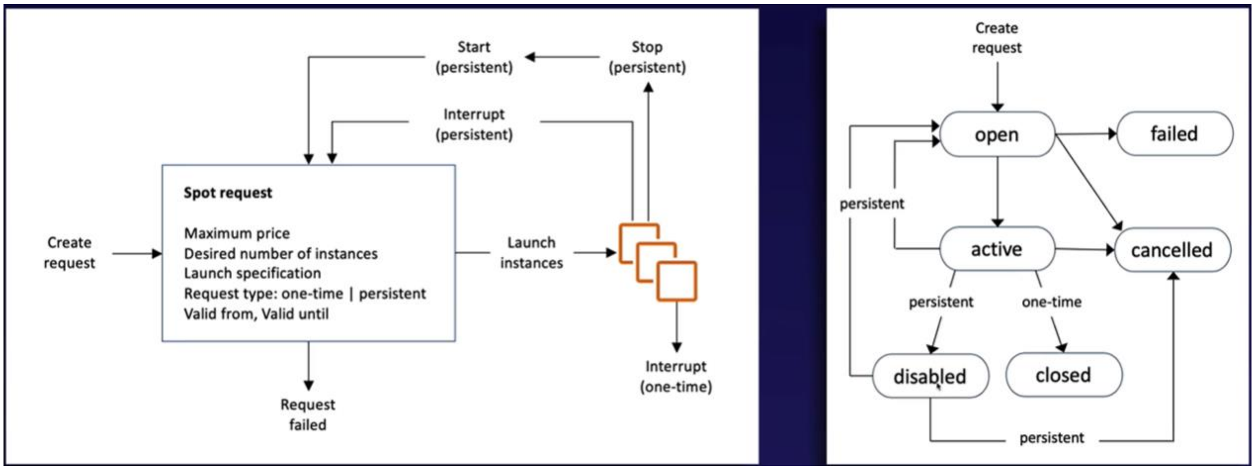
1. You are not concerned about the **start** or **stop** time of your application.
2. Your workloads **can** be interrupted.
3. Your application is only feasible at very low compute prices.

✓ **Fun facts:**

1. You can save up to **90%** off On-Demand prices.
2. You pay the spot price that's in effect at the beginning of each hour.

To use spot instances, you must first decide on your max spot price. Which are hourly spot price (varies on capacity and region). Spot instances are useful in such as big data, containerized workloads, CI/CD, high-performance computing (HPC) and image and media rendering. Spot instances are not good for persistent workloads, critical jobs, and databases. Basically don’t work on nothing critical when working with spot instances.

- Spot Block
  - To stop your spot instances from being terminated even if the spot price goes over your max Spot price. You can set spot blocks for between 1 to 6 hours currently.

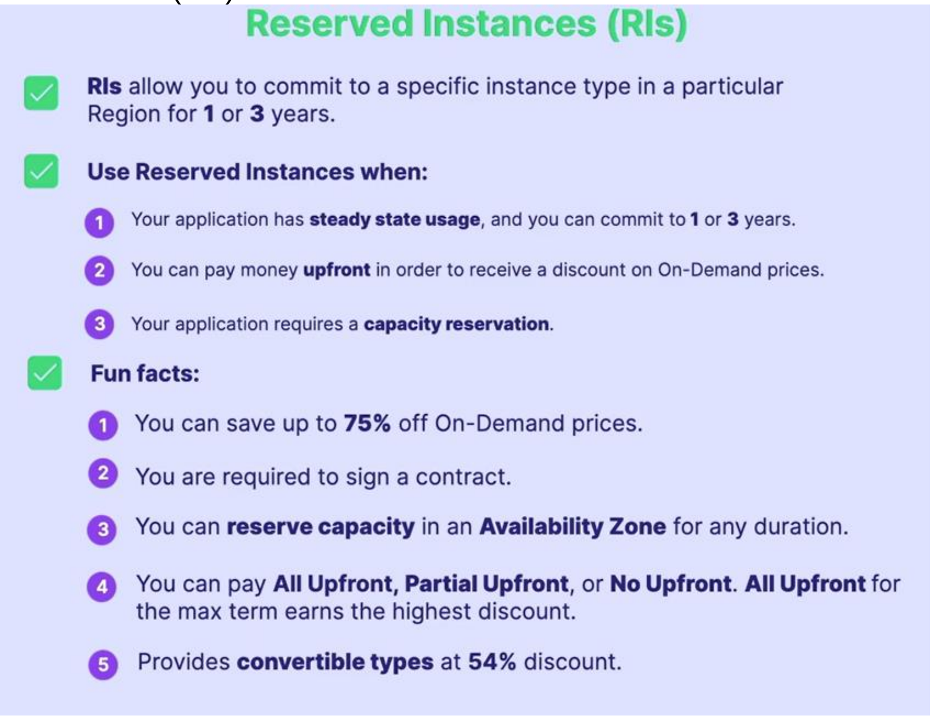


- ❖ Spot Fleet
    - ○ This attempts to launch the number of Spot instances and on-demand instances to meet the target capacity you specified in the spot fleet request.
        - ▪ Max price you specified in the request exceeds the current spot price.
    - ○ Spot fleets will try and match the target capacity with your price restraints

Strategies

- **Capacity optimized:** spot instances come from the pool with capacity for the number of instances launching
- **Lowest price:** The spot instances come from the pool with the lowest price.
- **Diversified:** The spot instances are distributed across all pools
- **InstancePoolsToUseCount:** Spot instances are distributed across the number of spot instance pools you specify
-

- Reserved Instances (RIs)



Reserved Instances (RIs)

✓ **RIs** allow you to commit to a specific instance type in a particular Region for **1** or **3** years.

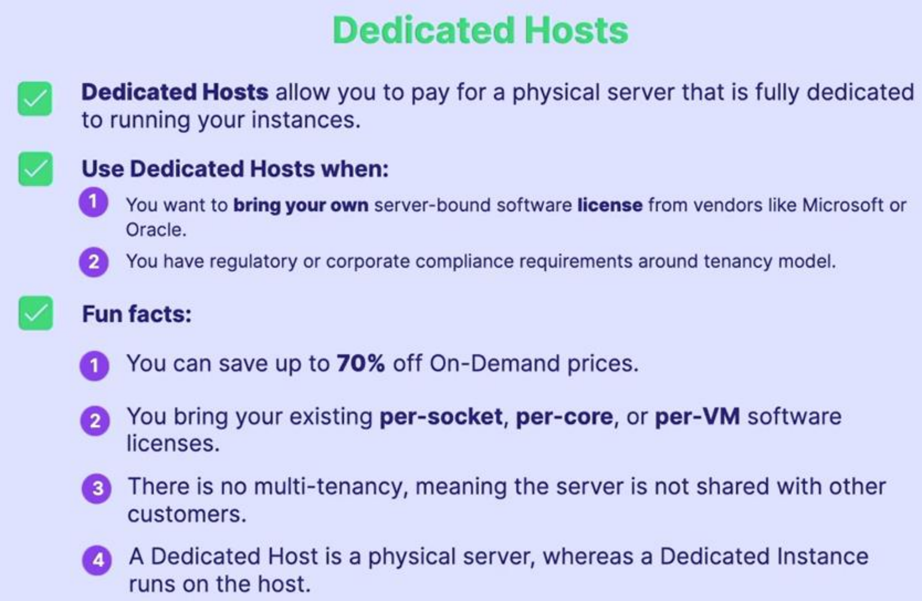
✓ **Use Reserved Instances when:**

1. Your application has **steady state usage**, and you can commit to **1** or **3** years.
2. You can pay money **upfront** in order to receive a discount on On-Demand prices.
3. Your application requires a **capacity reservation.**

✓ **Fun facts:**

1. You can save up to **75%** off On-Demand prices.
2. You are required to sign a contract.
3. You can **reserve capacity** in an **Availability Zone** for any duration.
4. You can pay **All Upfront, Partial Upfront**, or **No Upfront**. **All Upfront** for the max term earns the highest discount.
5. Provides **convertible types** at **54%** discount.

○

- Dedicated Hosts
    - 

Dedicated Hosts

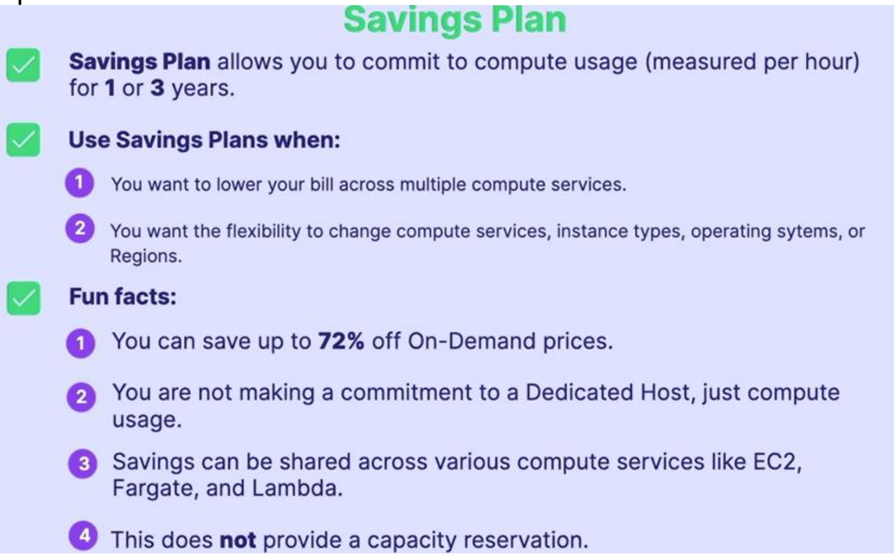
☑ **Dedicated Hosts** allow you to pay for a physical server that is fully dedicated to running your instances.

☑ **Use Dedicated Hosts when:**

1. You want to **bring your own** server-bound software **license** from vendors like Microsoft or Oracle.
2. You have regulatory or corporate compliance requirements around tenancy model.

☑ **Fun facts:**

1. You can save up to **70%** off On-Demand prices.
2. You bring your existing **per-socket**, **per-core**, or **per-VM** software licenses.
3. There is no multi-tenancy, meaning the server is not shared with other customers.
4. A Dedicated Host is a physical server, whereas a Dedicated Instance runs on the host.

- Savings plans
    - 

Savings Plan

☑ **Savings Plan** allows you to commit to compute usage (measured per hour) for **1** or **3** years.

☑ **Use Savings Plans when:**

1. You want to lower your bill across multiple compute services.
2. You want the flexibility to change compute services, instance types, operating sytems, or Regions.

☑ **Fun facts:**

1. You can save up to **72%** off On-Demand prices.
2. You are not making a commitment to a Dedicated Host, just compute usage.
3. Savings can be shared across various compute services like EC2, Fargate, and Lambda.
4. This does **not** provide a capacity reservation.

>Using Roles

Like users and groups, IAM roles define the limits for what can be done within your AWS account. The important difference is that, unlike users and groups, roles are, for the most part, used by applications and services rather than people. Must specify exactly what permissions you want to give the role or, in other words, what you want the beneficiary processes to be able to do. From that point any authenticated mobile app users will have access to those S3 resources.

- IAM Policies

  - Conditions in IAM policies can look at resource tags to determine whether to allow a particular action. For example, you can specify a condition that permits an EC2 instance to access a production database only if the instance has the Environment tag with the value Production.
  - AWS does not allow users to add an IAM role to an IAM group at this time

In the real world, you can attach a role to an instance that provides privileges to applications running on the instance. Roles help you avoid sharing long-term credentials like access keys and protect your instances.

- IAM Credential Reports
  - This report lists all users in your account and the status of their various credentials.
  - This lists all users in your account and the status of their various credentials, including passwords, access keys, and MFA devices. You can get a credential report from the AWS Management Console, the AWS SDKs and Command Line Tools, or the IAM API.
  - Accessed from the IAM Dashboard, a credential report displays a simple interface with no more (or less) than one lonely button: Download Report. We'll let you handle the practical details from there.Accessed from the IAM Dashboard, a credential report displays a simple interface with no more (or less) than one lonely button: Download Report. We'll let you handle the practical details from there.
  - Suggest downloading the comma-separated values (CSV) files the service generates.

>Security Groups and bootstrap scripts

Down below are ways computers communicate

- Linux | SSH | Port 22
- Windows | RDP | Port 3389
- HTTP | Web Browsing | Port 80
- HTTPS | Web Browsing (SSL\Secure) | Port 443

~ Of course, there is more, but this is all we need to know for taking the exam

- Security groups
  - Security groups are basically virtual firewalls for your EC2 instances. Which by default everything is blocked. For example, after you associate a security group with an EC2 instance, it controls the inbound and outbound traffic for the instance.

        - Here everything will be blocked, you won't be able to SSH, or use it as a web server
            - In order to have communication to your EC2 instance such as SSH/RDP/HTTP you will need to open up the correct ports.

0.0.0.0/0

^ open that IP range. You wouldn't want to do this for SSH or RDP because this opens for an attack within your EC2 instances

- Bootstrap Scripts
    - A script that runs when the instance first runs. Which this performs at root levels
- #!/bin/bash
- yum update -y
- yum install httpd -y
- service httpd start
- cd /var/www/html
- echo "<html><body><h1>Hello my name is Cameron</h1></body></html>" > index.html

In the how to document for bootstrap, remember if we deleted port 80 of course we would no longer go into the webpage.

Tip: a bootstrap script is a script that runs when the instance first runs. It passes user data to the EC2 instance and can be used to install applications (like web servers and databases) as well as do updates and more.

https://docs.aws.amazon.com/vpc/latest/userguide/VPC_SecurityGroups.html

>EC2 Metadata and User Data

Metadata is simply data about your EC2 instance such as…

- Private IP address
- Public IP address
- Hostnames
- Security Groups

To retrieve your metadata you need to type in a command using the curl command. To see how to do this, go to "How To's" folder on the EC2 Metadata and User data

The "http://169.254.169.254/latest/meta-data/" can be used to view instance metadata.

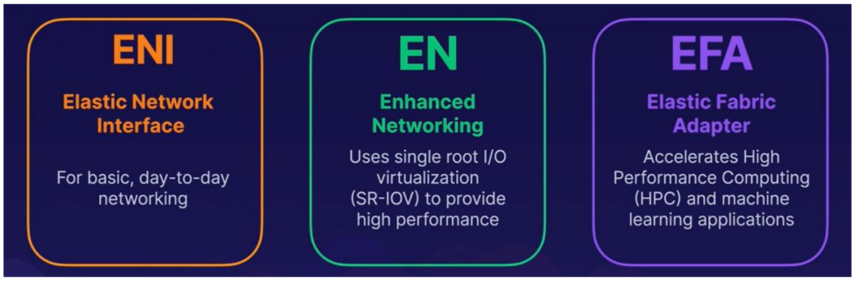
User data is commonly used for bootstrapping an EC2 instance as it comes online.

Remember… if you reboot your EC2 instance, the user data will **not** automatically be updated when the reboot is initiated. User data runs one time and one time only

https://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/instancedata-add-user-data.html

>Networking with EC2

There are 3 different types of virtual networking cards to your EC2 instances



ENI

Elastic Network Interface

For basic, day-to-day networking

EN

Enhanced Networking

Uses single root I/O virtualization (SR-IOV) to provide high performance

EFA

Elastic Fabric Adapter

Accelerates High Performance Computing (HPC) and machine learning applications

1. ENI
   a. Private and public IPv4 Address
   b. Many IPv6 Addresses
   c. MAC Address
   d. 1 or more security Groups

Common ENI Cases…

- Create a management network
- Use network and security appliances in your VPC
- Create dual-homed instances with workloads/roles on distinct subnets
- Create a low-budget, high availability solution

ENI is a good way to create a low budget and high availability solutions. When creating an EC2 it puts ENI to it by default.

2. Enhanced Networking

Enhanced networking uses single root I/O virtualization (SR-IOV) to provide high-performance networking capabilities on supported instance types.

All current generation instance types support enhanced networking, except for T2 instances.

You can enable enhanced networking using one of the following mechanisms:

Elastic Network Adapter (ENA)
The Elastic Network Adapter (ENA) supports network speeds of up to 100 Gbps for supported instance types.
The current generation instances use ENA for enhanced networking, except for C4, D2, and M4 instances smaller than m4.16xlarge.

Intel 82599 Virtual Function (VF) interface
The Intel 82599 Virtual Function interface supports network speeds of up to 10 Gbps for supported instance types.
The following instance types use the Intel 82599 VF interface for enhanced networking: C3, C4, D2, I2, M4 (excluding m4.16xlarge), and R3.

   a. 10 Gbps – 100 Gbps
   b. Provides higher bandwidth, higher packet per second performance

3. EFA (Elastic Fabric Adapter)

   a. A network device you can attach to your Amazon EC2 instance to accelerate high performance computing (HPC) and machine learning applications

Elastic Fabric Adapter (EFA) is a network interface for Amazon EC2 instances that enables customers to run applications requiring high levels of inter-node communications at scale on AWS. Its custom-built operating system (OS) bypass hardware interface enhances the performance of inter-instance communications, which is critical to scaling these applications. With EFA, High Performance Computing (HPC) applications using the Message Passing Interface (MPI) and Machine Learning (ML) applications using NVIDIA Collective Communications Library (NCCL) can scale to thousands of CPUs or GPUs.

>Optimizing with EC2 Placement Groups

Within these placement groups there are only 3 groups

1. Cluster
   a. A cluster placement group can't span multiple Availability Zones.
   b. You can launch multiple instance types into a cluster placement group. However, this reduces the likelihood that the required capacity will be available for your launch to succeed. We recommend using the same instance type for all instances in a cluster placement group.
   c. You would want to use a cluster placement group when you want to reduce network latency in your application.
2. Spread
   a. supports a maximum of seven running instances per Availability Zone.
   b. If you need more than seven instances in an Availability Zone, we recommend that you use multiple spread placement groups.
   c. You can't use Capacity Reservations to reserve capacity in a spread placement group.
3. Partition
   a. A partition placement group supports a maximum of seven partitions per Availability Zone.
   b. When instances are launched into a partition placement group, Amazon EC2 tries to evenly distribute the instances across all partitions. Amazon EC2 doesn't guarantee an even distribution of instances across all partitions.

https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/placement-groups.html#placement-groups-spread

>Timing workloads with Spot Instances and Spot Fleets.

What are EC2 spot instances?

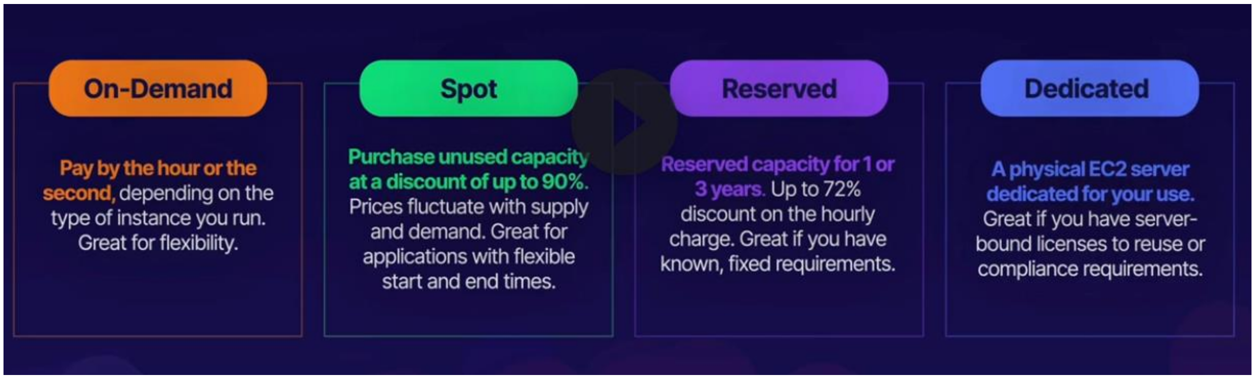
They let you take advantage of unused EC2 capacity in the AWS cloud. They are available at up to 90% discount compared to on-demand prices.

- We would need spot instances for stateless, fault-tolerant, or flexible applications.
  - Such as big data, containerized workloads, CI/CD, high-performance computing (HPC)

-

EC2 Exam Tips

1. EC2 is like a VM hosted in AWS instead of your own Data Center
2. Pricing Options



On-Demand

Pay by the hour or the second, depending on the type of instance you run. Great for flexibility.

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.

Reserved

Reserved capacity for 1 or 3 years. Up to 72% discount on the hourly charge. Great if you have known, fixed requirements.

Dedicated

A physical EC2 server dedicated for your use. Great if you have server-bound licenses to reuse or compliance requirements.

3. AWS Command Line
    a. Allow only minimum account of access required to do their job
    b. Use groups, create IAM groups and assign your users to groups
4. Secret access key
    a. You will only see this once, so if you lose it you can delete the access key ID and secret access key and regenerate them
    b. DO NOT SHARE KEY PAIRS
    c. Command line supports Linux, Windows, and MACos
5. Know the placement groups
    a. Cluster
    b. Partition
    c. Spread
6. Security Groups
    a. Changes to security groups take effect immediately
    b. You can have any number of EC2 instances within a security group
    c. All inbound traffic is blocked by default whereas outbound is allowed
    d. You can have multiple secure
7. User Data vs Metadata
    a. User data is simply bootstrap scripts
    b. Metadata is data about your EC2 instances
    c. You can use bootstrap scripts (user data) to access metadata
8. Networking with EC2
    a. ENI – basic networking perhaps a separate management network from your production network or a separate logging network. Need to do this at a low cost

    b. EFA – For when you need accelerate high performance computing (HPC) and machine learning applications
    c. Enhanced Networking – For when you need speeds between 10 Gbps and 100 Gbps. Anywhere you need reliable, high throughput

9. Solving Licensing
    a. Any questions that talks about special licensing requirements
        i. Think of dedicate host – physical server with EC2 instance capacity fully dedicated to your use. Allows you to use your existing licenses, servers, cores, ect
10. Spot Instances
    a. Saves up 90% of the cost of On-Demand instances
    b. Useful for any type of computing where you don't need persistent storage.
    c. By using spot block, this helps instances from terminating
    d. Spot Fleet is a collection of spot instances and on-demand instances

https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/placement-groups.html