Table of Contents

Foreword

This guide tries to explain what happens when you launch an EC instance in AWS from a high-level perspective. When you launch an EC2 instance using default settings, the VPC, Subnets, and IGW are automatically set up for you thus you don't need to worry about them. Certain additional details are omitted and you are encouraged to deep dive into them on your own.

Concept You need to know

EC2 Instance

Think of it like a server machine, with its own CPU, RAM memory, and hard disk space. Similar to your computer but provided by a Cloud Service Provider (like AWS).

An EC2 instance can have the following actions related to its state:

- **Start** (launch the instance) from a newly created instance or a previously stopped instance or an EBS snapshot.
- **Stop**: Shut down your instance, but keep the EBS volume (hard disk space) associated with it.
- **Terminate (Think first before you click this)**: Shut down your instance and delete the EBS volume (hard disk space) associated with it. which means **any software installed will be gone and you will have to reinstall them** again when you launch a new instance.

One recommended practice is to create an EBS snapshot of your instance whenever you install new software / perform a major configuration change. You can then use the snapshot as a backup, use it to launch a new instance without the need to reinstall the software required, etc. Note to delete older version of EBS snapshot to save cost

An EC2 instance launched with default settings will have a **public IP address** and a **private IP address** automatically assigned to it. Note you need to set up a Security Group (SG) to allow the inbound traffic into your instance.

Public IP Address

Allows you to connect to your EC2 instance over the internet.

By default, AWS will automatically assign a new public IP address when you launch an instance. That means your public IP address will be different every time you start your instance.

One way to fix the public IP address is to associate your instance with the **Elastic IP address.** Then the public IP address will be the same as the Elastic IP address during every start. Note Elastic IP address is not free.

Concept for Your Knowledge

Private IP Address

A private IP Address is one way for your EC2 instance to connect to other resources in your VPC without going through the internet.

Knowledge to set up and use the private IP address is not required to complete the assignments. But if you want to know more:

The private IP address is not visible to the internet. So you can use any address you want without worrying that someone may be able to access it.

In practice, private IPv4 address ranges that you can use are specified in RFC1918

- 10.0.0.0 - 10.255.255.255
- 172.16.0.0 - 172.31.255.255
- 192.168.0.0 - 192.168.255.255

The private IPv4 address is associated with a **subnet**. A subnet will have an associated CIDR block defined. The **CIDR block** defined the range of private IPv4 addresses that can be used within the subnet.

Example CIDR block range for a subnet

- 172.16.0.0/20
- The /20 indicates the non-available bits for the subnet portion of the address.
- An IPv4 address will have 32 bits, where each block number between the dots can have values between 0 – 255 (8 bits).
- So if 20 bits are not available, the available bits are then 12.
- which means there are 2^12 addresses that you can use in your subnet.
- In the example above, these addresses will start from 172.16.0.0 and up to 172.16.15.255.

For AWS, the first four IP addresses and the last IP address in each subnet CIDR block are not available for your use, and they cannot be assigned to a resource.
https://docs.aws.amazon.com/vpc/latest/userguide/subnet-sizing.html

Similar to a Public IP address, the private IP address is automatically assigned when you launch your instance. See the addendum for how to set up a private IP address when launching your EC2 instance.

VPC

A virtual private cloud (VPC) is an abstract concept, it means a secure, isolated private cloud hosted within a public cloud.

VPC is region-specific (region: AWS Data Center Location). All resources you launch within the same region will be assigned to the same VPC by default. You can create separate VPCs for the same region, but the knowledge to do so is beyond the scope of the course and assignments.

The VPC will group the resources into **subnets.** One subnet is usually associated with an Availability Zone (AZ: AWS Data Center).

The VPC will need an Internet GateWay (IGW) to connect to the Internet.

You can view your VPC settings in Your VPCs page (see screenshot below).

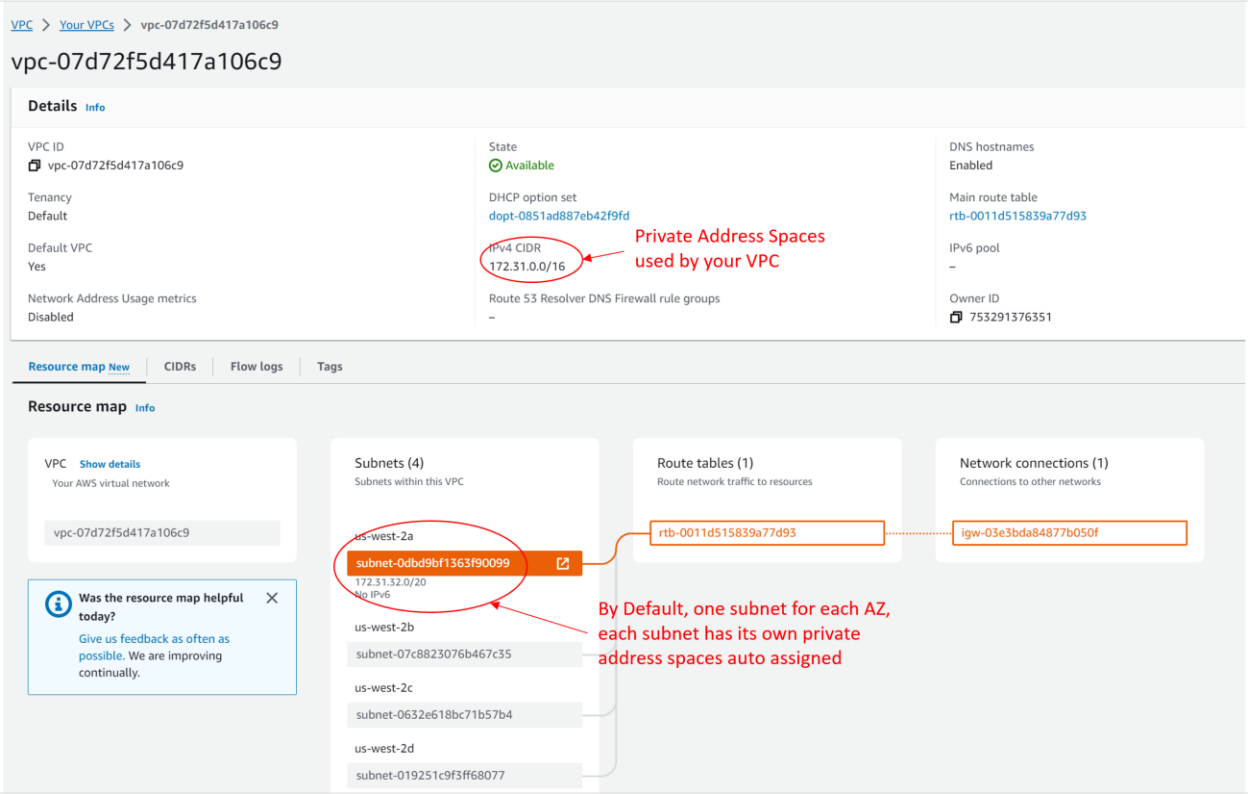


Figure 1: Reference Your VPCs page on AWS

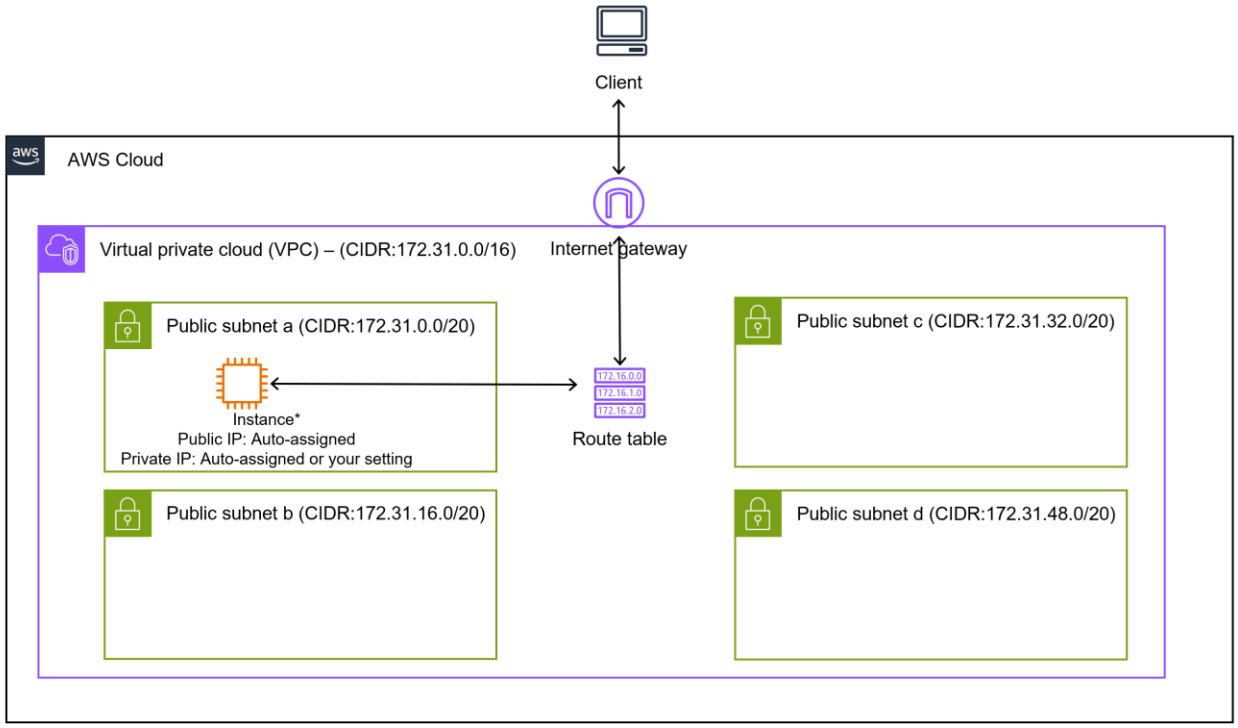
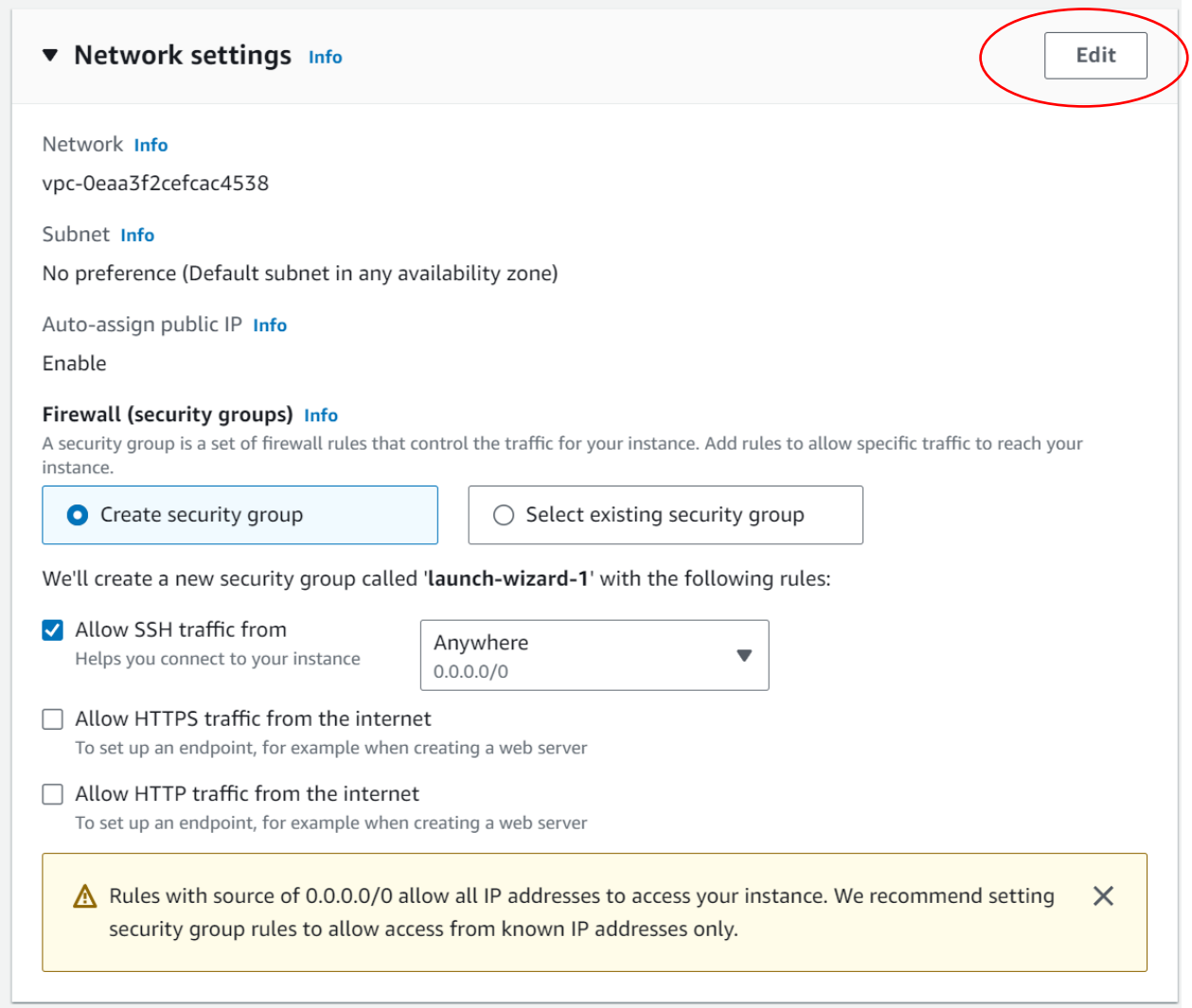


Figure 2: Simplify AWS Resources Architecture Diagram

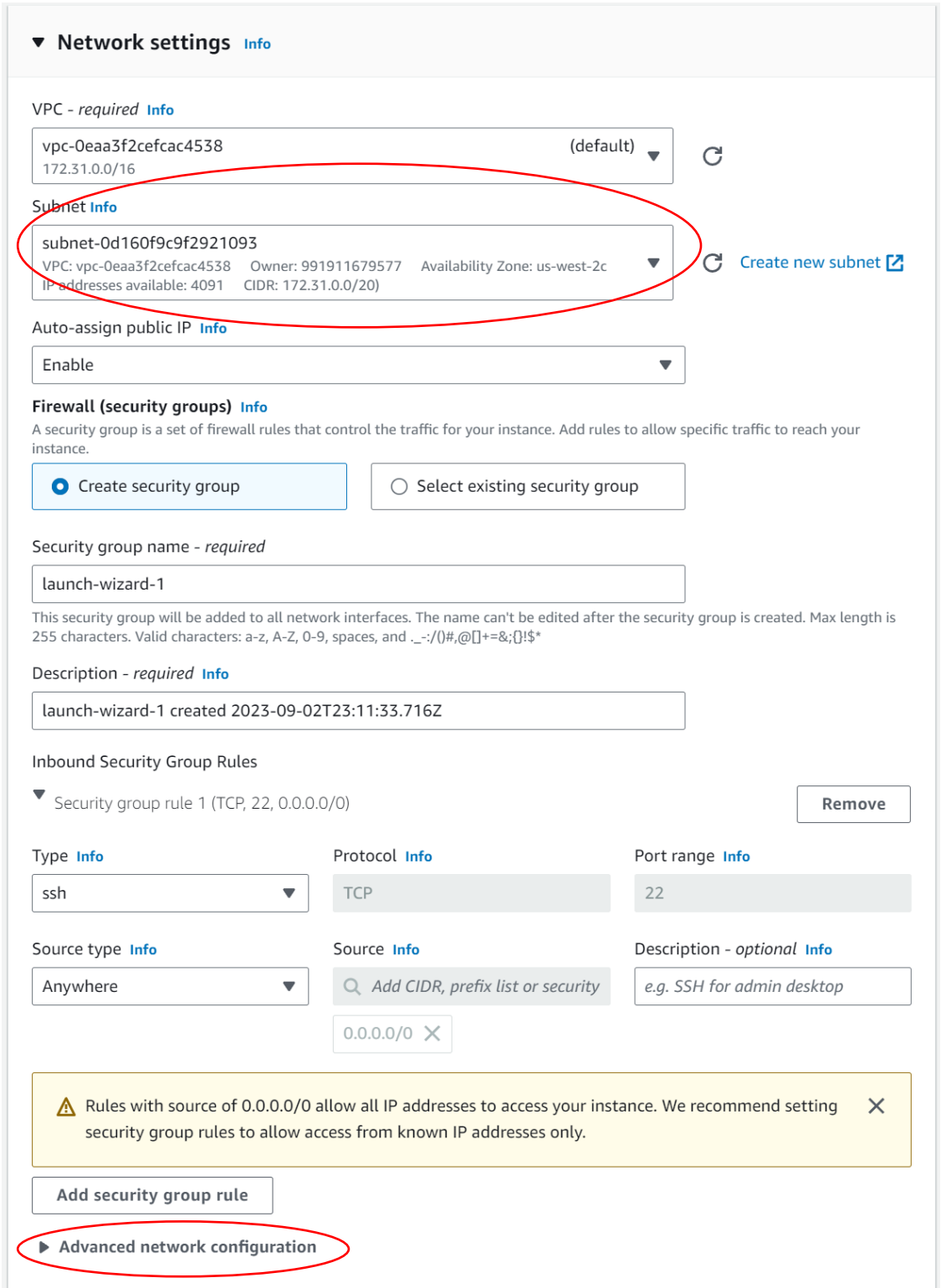
Addendum

How to set the private IP address when launching EC2

1. When launching a new EC2 instance from the template, click Edit Network Settings



2. Select a subnet from the default VPC, then click the text “Advanced network connection” at the bottom.



3. Set your Primary IP for the Private IP address, note the address must be valid according to the subnet chosen.