**Lecture Notes for AWS Solution Architect**

**Module-I**

**On-premises data centers**

"**On-premise**" refers to private data centers that companies house in their own facilities and maintain themselves.

On-premise infrastructure can be used to run private clouds, in which compute resources are virtualized in much the same way as those of public clouds (however, private clouds can also be run on leased third-party hardware).

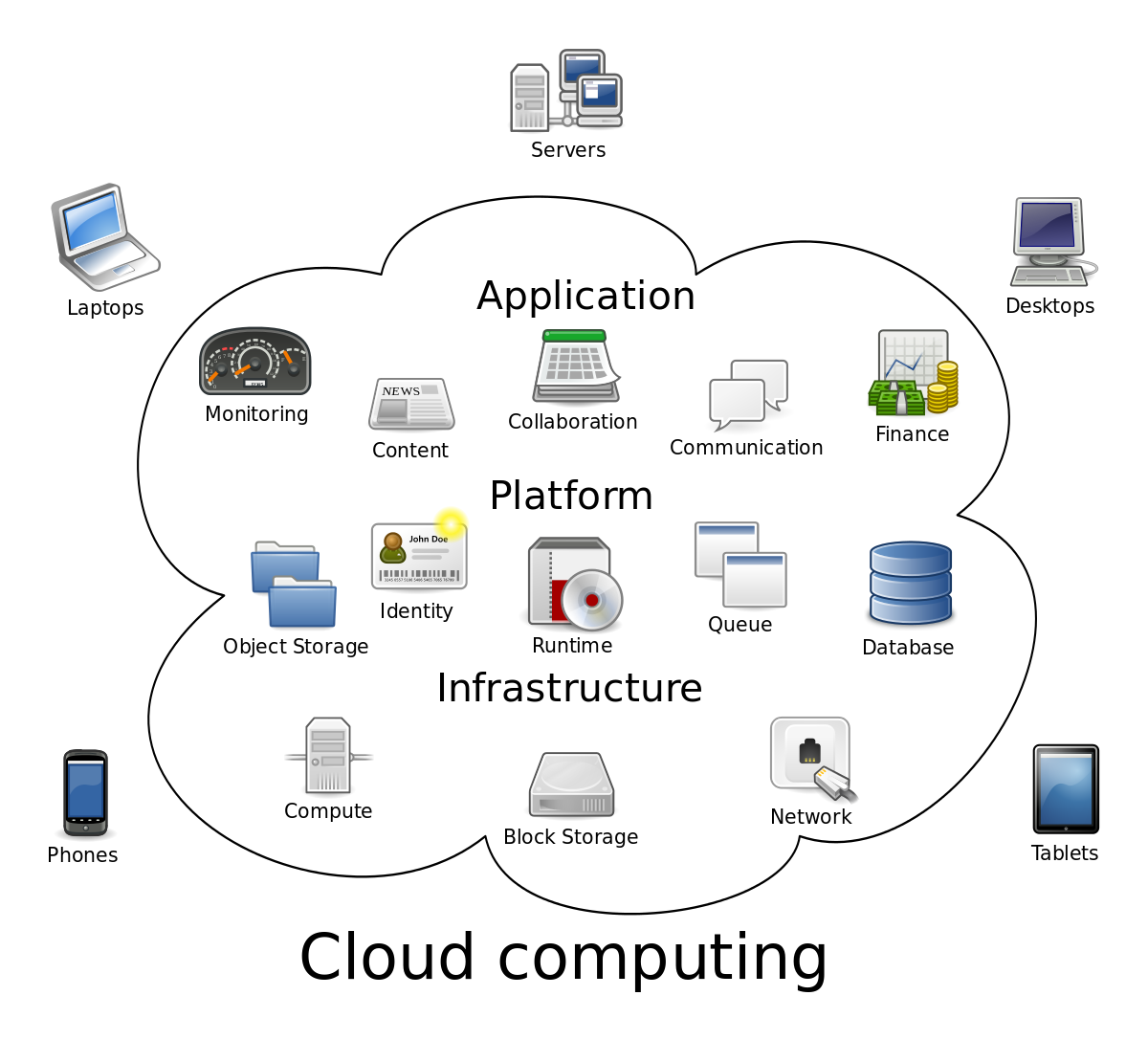
**On-Premise** defined: A solution hosted in-house and usually supported by a third-party.

**Off**-**Premise** defined: A solution hosted by a third-party and usually supported by a different third-party.

**What is cloud computing**?

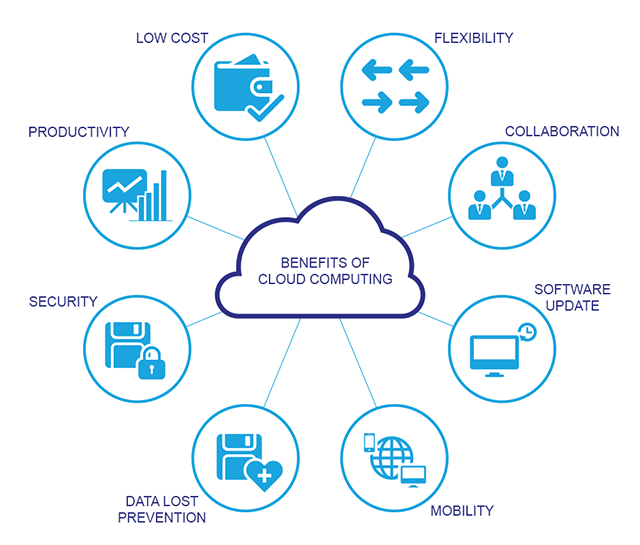
Cloud computing is the on-demand delivery of IT resources over the Internet with pay-as-you-go pricing. Instead of buying, owning, and maintaining physical data centers and servers, you can access technology services, such as computing power, storage, and databases, on an as-needed basis from a cloud service provider.

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.



**Benefits of Cloud Computing**

* Cost Savings
* Security
* Flexibility
* Mobility
* Insight
* Increased Collaboration
* Quality Control
* Disaster Recovery
* Loss Prevention
* Automatic Software Updates
* Competitive Edge
* Sustainability



**Common Characteristics**

* Massive Scale
* Resilient Computing
* Homogeneity
* Geographic Distribution
* Virtualization
* Service Orientation
* Low Cost Software
* Advanced Security

**Essential Characteristics**

* **On-demand self-service**

A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

* **Broad network access**

Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

* **Resource pooling**

The provider’s computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.

**Types of Cloud**

1. Public Cloud
2. Private Cloud
3. Hybrid Cloud
4. Community Cloud

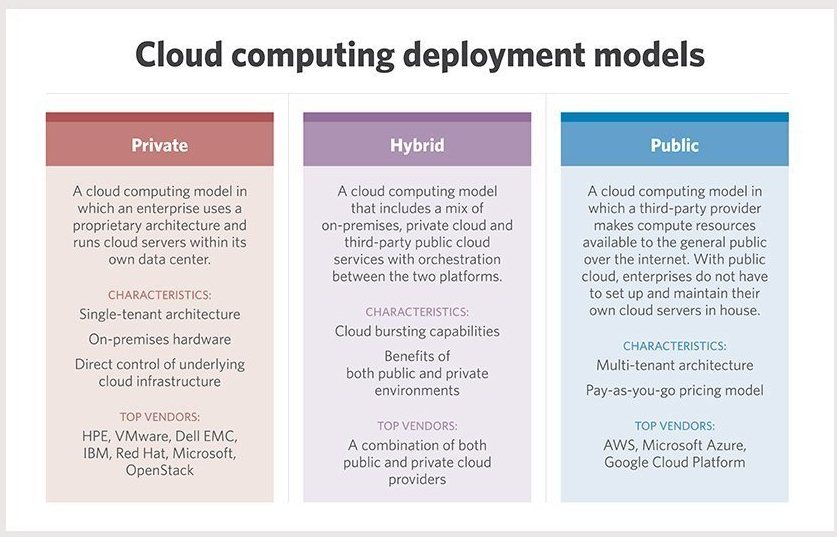
**Public cloud** is defined as computing services offered by third-party providers over the public Internet, making them available to anyone who wants to use or purchase them. They may be free or sold on-demand, allowing customers to pay only per usage for the CPU cycles, storage, or bandwidth they consume.

**Private cloud** solutions are dedicated to one organization or business, and often have much more specific security controls than a public cloud. Many medical offices, banking institutions, and other organizations who are required to meet federal and state guidelines for data controls use a private cloud. Using private cloud storage allows them to control highly sensitive data by meeting regulations and industry-based criteria, whether that be medical records, trade secrets, or other classified information.

**Hybrid cloud** solutions are a blend of public and private clouds. This is a more complex cloud solution in that the organization must manage multiple platforms and determine where data is stored. An example of a hybrid cloud solution is an organization that wants to keep confidential information secured on their private cloud, but make more general, customer-facing content on a public cloud.

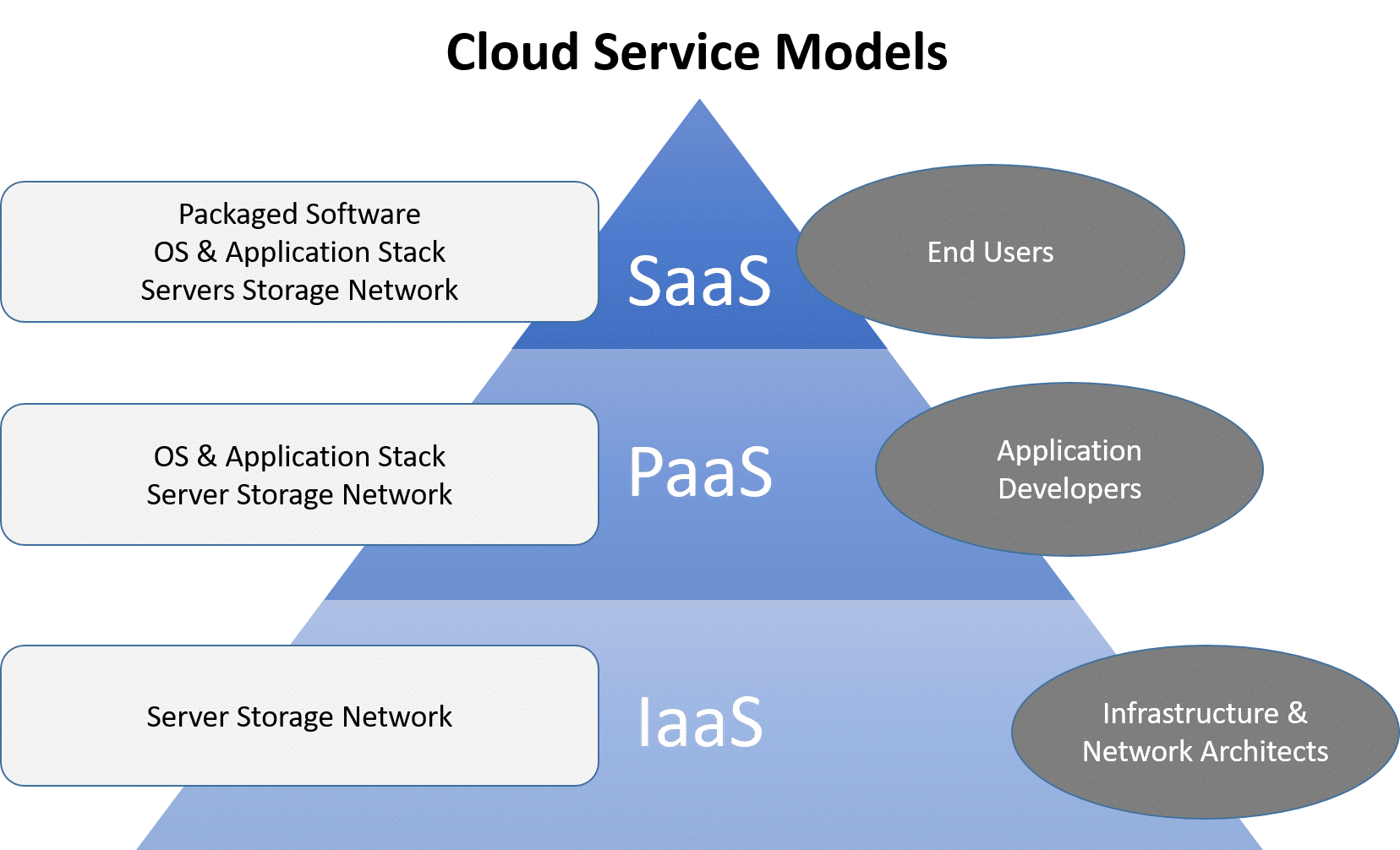
**Community cloud** is a cloud infrastructure that allows systems and services to be accessible by a group of several organizations to share the information. It is owned, managed, and operated by one or more organizations in the community, a third party, or a combination of them.

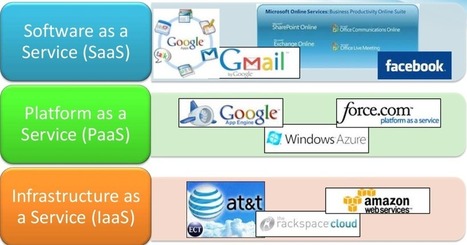
**Comparison between Public, Private and Hybrid Cloud**

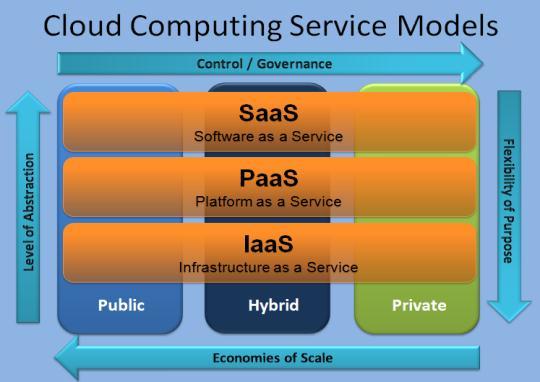


**Cloud Service Models**

* IAAS
* PAAS
* SAAS
* Infrastructure as a service (IaaS) is a cloud computing offering in which a vendor provides users access to computing resources such as servers, storage and networking. Organizations use their own platforms and applications within a service provider's infrastructure. A vendor provides clients pay-as-you-go access to storage, networking, servers and other computing resources in the cloud.
* Platform as a service (PaaS) is a service provider offers access to a cloud-based environment in which users can build and deliver applications. The provider supplies underlying infrastructure.
* Software as a service (SaaS) is a service provider delivers software and applications through the internet. Users subscribe to the software and access it via the web or vendor APIs.







**AWS Global Infrastructure**

The AWS Global Cloud Infrastructure is the most secure, extensive, and reliable cloud platform, offering over 175 fully featured services from data centers globally. Whether you need to deploy your application workloads across the globe in a single click, or you want to build and deploy specific applications closer to your end-users with single-digit millisecond latency, AWS provides you the cloud infrastructure where and when you need it.

With millions of active customers and tens of thousands of partners globally, AWS has the largest and most dynamic ecosystem. Customers across virtually every industry and of every size, including start-ups, enterprises, and public sector organizations, are running every imaginable use case on AWS.

**Regions**

AWS has the concept of a Region, which is a physical location around the world where we cluster data centers. We call each group of logical data centers an Availability Zone.

Each AWS Region consists of multiple, isolated, and physically separate AZ's within a geographic area. Unlike other cloud providers, who often define a region as a single data center, the multiple AZ design of every AWS Region offers advantages for customers.

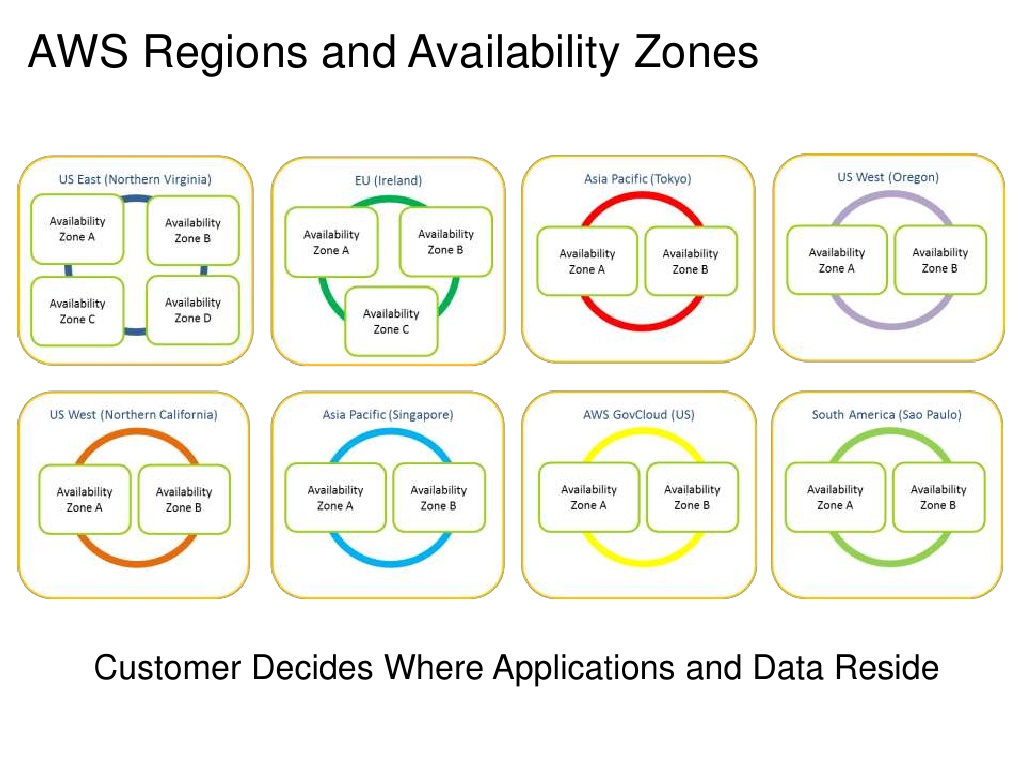
Each AZ has independent power, cooling, and physical security and is connected via redundant, ultra-low-latency networks. AWS customers focused on high availability can design their applications to run in multiple AZ's to achieve even greater fault-tolerance. AWS infrastructure Regions meet the highest levels of security, compliance, and data protection.

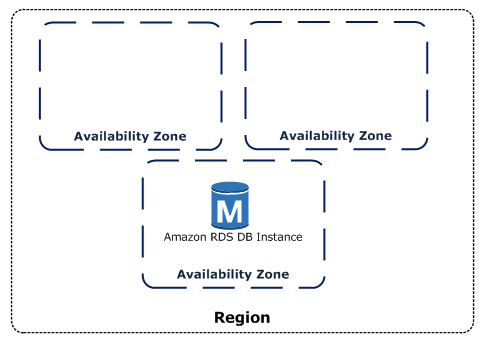
|  |  |  |  |
| --- | --- | --- | --- |
| **Region Name** | **Region** | **Endpoint** | **Protocol** |
| US East (Ohio) | us-east-2 | rds.us-east-2.amazonaws.com | HTTPS |
| US East (N. Virginia) | us-east-1 | rds.us-east-1.amazonaws.com | HTTPS |
| US West (N. California) | us-west-1 | rds.us-west-1.amazonaws.com | HTTPS |
| US West (Oregon) | us-west-2 | rds.us-west-2.amazonaws.com | HTTPS |
| Asia Pacific (Mumbai) | ap-south-1 | rds.ap-south-1.amazonaws.com | HTTPS |
| Asia Pacific (Osaka-Local) | ap-northeast-3 | rds.ap-northeast-3.amazonaws.com | HTTPS |
| Asia Pacific (Seoul) | ap-northeast-2 | rds.ap-northeast-2.amazonaws.com | HTTPS |
| Asia Pacific (Singapore) | ap-southeast-1 | rds.ap-southeast-1.amazonaws.com | HTTPS |
| Asia Pacific (Sydney) | ap-southeast-2 | rds.ap-southeast-2.amazonaws.com | HTTPS |
| Asia Pacific (Tokyo) | ap-northeast-1 | rds.ap-northeast-1.amazonaws.com | HTTPS |
| Canada (Central) | ca-central-1 | rds.ca-central-1.amazonaws.com | HTTPS |
| China (Beijing) | cn-north-1 | rds.cn-north-1.amazonaws.com.cn | HTTPS |

**Availability Zones**

An Availability Zone (AZ) is one or more discrete data centers with redundant power, networking, and connectivity in an AWS Region. AZ’s give customers the ability to operate production applications and databases that are more highly available, fault tolerant, and scalable than would be possible from a single data center.

All AZ’s in an AWS Region are interconnected with high-bandwidth, low-latency networking, over fully redundant, dedicated metro fiber providing high-throughput, low-latency networking between AZ’s. All traffic between AZ’s is encrypted. The network performance is sufficient to accomplish synchronous replication between AZ’s. AZ’s make partitioning applications for high availability easy.





**Amazon EC2**

Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster. You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage. Amazon EC2 enables you to scale up or down to handle changes in requirements or spikes in popularity, reducing your need to forecast traffic.

**Features of Amazon EC2**

Amazon EC2 provides the following features:

* Virtual computing environments, known as instances
* Preconfigured templates for your instances, known as Amazon Machine Images (AMIs), that package the bits you need for your server (including the operating system and additional software)
* Various configurations of CPU, memory, storage, and networking capacity for your instances, known as instance types
* Secure login information for your instances using key pairs (AWS stores the public key, and you store the private key in a secure place)
* Storage volumes for temporary data that's deleted when you stop or terminate your instance, known as instance store volumes
* Persistent storage volumes for your data using Amazon Elastic Block Store (Amazon EBS), known as *Amazon EBS volumes*
* Multiple physical locations for your resources, such as instances and Amazon EBS volumes, known as *Regions* and *Availability Zones*
* A firewall that enables you to specify the protocols, ports, and source IP ranges that can reach your instances using *security groups*
* Static IPv4 addresses for dynamic cloud computing, known as *Elastic IP addresses*
* Metadata, known as *tags*, that you can create and assign to your Amazon EC2 resources
* Virtual networks you can create that are logically isolated from the rest of the AWS cloud, and that you can optionally connect to your own network, known as *virtual private clouds* (VPCs)

**Amazon Virtual Private Cloud**

**Amazon Virtual Private Cloud** (Amazon VPC) lets you provision a logically isolated section of the AWS Cloud where you can launch AWS resources in a virtual network that you define.

* You have complete control over your virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways.
* You can use both IPv4 and IPv6 in your VPC for secure and easy access to resources and applications.
* You can easily customize the network configuration of your Amazon VPC.
* For example, you can create a public-facing subnet for your web servers that have access to the internet. You can also place your backend systems, such as databases or application servers, in a private-facing subnet with no internet access.
* You can use multiple layers of security, including security groups and network access control lists, to help control access to Amazon EC2 instances in each subnet.
* A subnet is a range of IP addresses in your VPC. You can launch AWS resources into a specified subnet. Use a public subnet for resources that must be connected to the internet, and a private subnet for resources that won't be connected to the internet .
* To protect the AWS resources in each subnet, you can use multiple layers of security, including security groups and network access control lists (ACL).

**Route table concepts**

The following are the key concepts for route tables.

Main route table—The route table that automatically comes with your VPC. It controls the routing for all subnets that are not explicitly associated with any other route table.

Custom route table—A route table that you create for your VPC.

Edge association - A route table that you use to route inbound VPC traffic to an appliance. You associate a route table with the internet gateway or virtual private gateway, and specify the network interface of your appliance as the target for VPC traffic.

Route table association—The association between a route table and a subnet, internet gateway, or virtual private gateway.

Subnet route table—A route table that's associated with a subnet.

Gateway route table—A route table that's associated with an internet gateway or virtual private gateway.

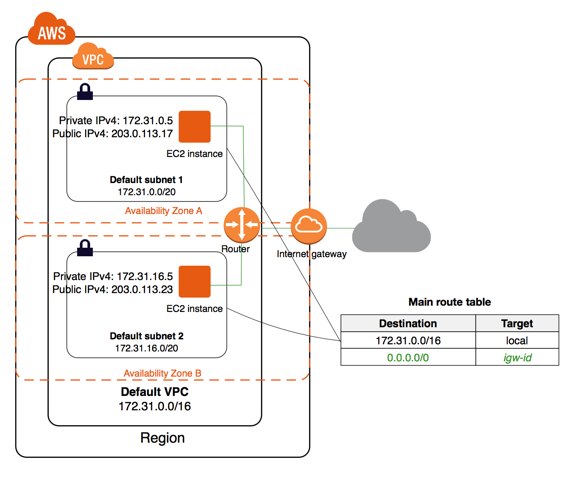
Local gateway route table—A route table that's associated with an Outposts local gateway. For information about local gateways, see Local Gateways in the AWS Outposts User Guide.

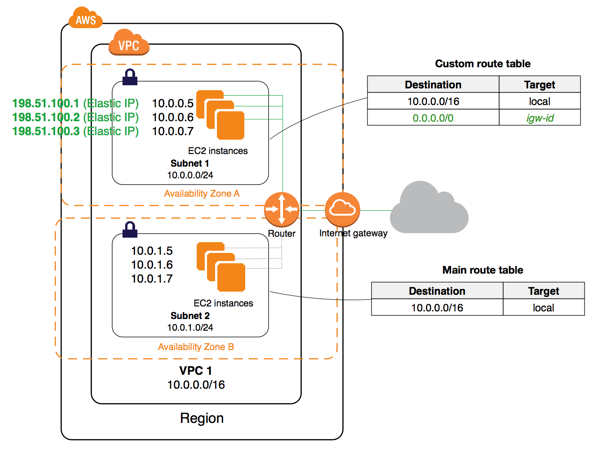
Destination—The range of IP addresses where you want traffic to go (destination CIDR). For example, an external corporate network with a 172.16.0.0/12 CIDR.

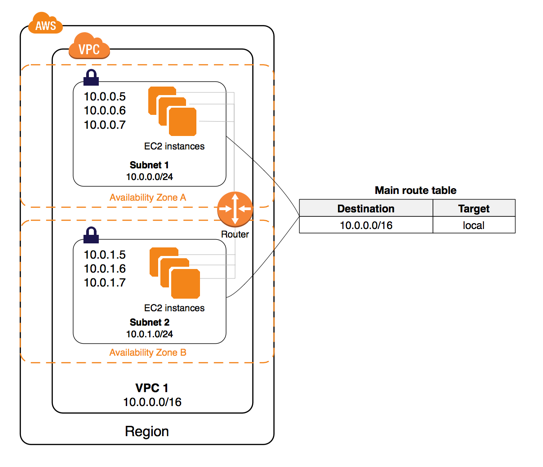
Propagation—Route propagation allows a virtual private gateway to automatically propagate routes to the route tables. This means that you don't need to manually enter VPN routes to your route tables. For more information about VPN routing options, see Site-to-Site VPN routing options in the Site-to-Site VPN User Guide.

Target—The gateway, network interface, or connection through which to send the destination traffic; for example, an internet gateway.

Local route—A default route for communication within the VPC.







**Internet Gateway**

An internet gateway is a horizontally scaled, redundant, and highly available VPC component that allows communication between your VPC and the internet. An internet gateway serves two purposes: to provide a target in your VPC route tables for internet-routable traffic, and to perform network address translation (NAT) for instances that have been assigned public IPv4 addresses.

An internet gateway supports IPv4 and IPv6 traffic. It does not cause availability risks or bandwidth constraints on your network traffic. There's no additional charge for having an internet gateway in your account.

**Enabling internet access**

* To enable access to or from the internet for instances in a subnet in a VPC, you must do the following.
* Create an internet gateway and attach it to your VPC.
* Add a route to your subnet's route table that directs internet-bound traffic to the internet gateway.
* Ensure that instances in your subnet have a globally unique IP address (public IPv4 address, Elastic IP address, or IPv6 address).
* Ensure that your network access control lists and security group rules allow the relevant traffic to flow to and from your instance.

**Subnet**

Subnet is a key component in VPC. A VPC can contain all public subnets (or) public/private subnet combination. Private Subnet is a subnet which doesn’t have a route to the internet gateway. A subnet can be configured as a VPN-only subnet by routing traffic via virtual private gateway.

**Default subnets**

By default, a default subnet is a public subnet, because the main route table sends the subnet's traffic that is destined for the internet to the internet gateway. You can make a default subnet into a private subnet by removing the route from the destination 0.0.0.0/0 to the internet gateway. However, if you do this, no EC2 instance running in that subnet can access the internet.

**Security group**

A security group acts as a virtual firewall for your EC2 instances to control incoming and outgoing traffic. Inbound rules control the incoming traffic to your instance, and outbound rules control the outgoing traffic from your instance. When you launch an instance, you can specify one or more security groups.

**NACL**

A network access control list (ACL) is an optional layer of security for your VPC that acts as a firewall for controlling traffic in and out of one or more subnets. You might set up network ACLs with rules similar to your security groups in order to add an additional layer of security to your VPC.

The following are the basic things that you need to know about network ACLs:

* Your VPC automatically comes with a modifiable default network ACL. By default, it allows all inbound and outbound IPv4 traffic and, if applicable, IPv6 traffic.
* You can create a custom network ACL and associate it with a subnet. By default, each custom network ACL denies all inbound and outbound traffic until you add rules.
* Each subnet in your VPC must be associated with a network ACL. If you don't explicitly associate a subnet with a network ACL, the subnet is automatically associated with the default network ACL.
* You can associate a network ACL with multiple subnets. However, a subnet can be associated with only one network ACL at a time. When you associate a network ACL with a subnet, the previous association is removed.
* A network ACL contains a numbered list of rules. We evaluate the rules in order, starting with the lowest numbered rule, to determine whether traffic is allowed in or out of any subnet associated with the network ACL. The highest number that you can use for a rule is 32766. We recommend that you start by creating rules in increments (for example, increments of 10 or 100) so that you can insert new rules where you need to later on.
* A network ACL has separate inbound and outbound rules, and each rule can either allow or deny traffic.
* Network ACLs are stateless, which means that responses to allowed inbound traffic are subject to the rules for outbound traffic (and vice versa).

**Network Address Translation**

Network Address Translation (NAT) is a process in which one or more local IP address is translated into one or more Global IP address and vice versa in order to provide Internet access to the local hosts. Also, it does the translation of port numbers i.e. masks the port number of the host with another port number, in the packet that will be routed to the destination. It then makes the corresponding entries of IP address and port number in the NAT table. NAT generally operates on router or firewall.

* **NAT** translates the IP addresses of **computers** in a local **network** to a single IP address. This address is often used by the router that connects the **computers** to the Internet. The router can be connected to a DSL modem, cable modem, T1 line, or even a dial-up modem.
* Network address translation is a method of remapping an IP address space into another by modifying network address information in the IP header of packets while they are in transit across a traffic routing device.
* For example a computer on an internal address of 192.168. 1.10 wanted to communicate with a web server somewhere on the internet, NAT would translate the address 192.168. 1.10 to the company's public address, lets call this 1.1. 1.1 for example. so that the internal address is identified as the public address when communicating with the outside world.
* When a packet traverse outside the local (inside) network, then NAT converts that local (private) IP address to a global (public) IP address. When a packet enters the local network, the global (public) IP address is converted to a local (private) IP address.
* If NAT run out of addresses, i.e., no address is left in the pool configured then the packets will be dropped and an Internet Control Message Protocol (ICMP) host unreachable packet to the destination is sent.

