AWS

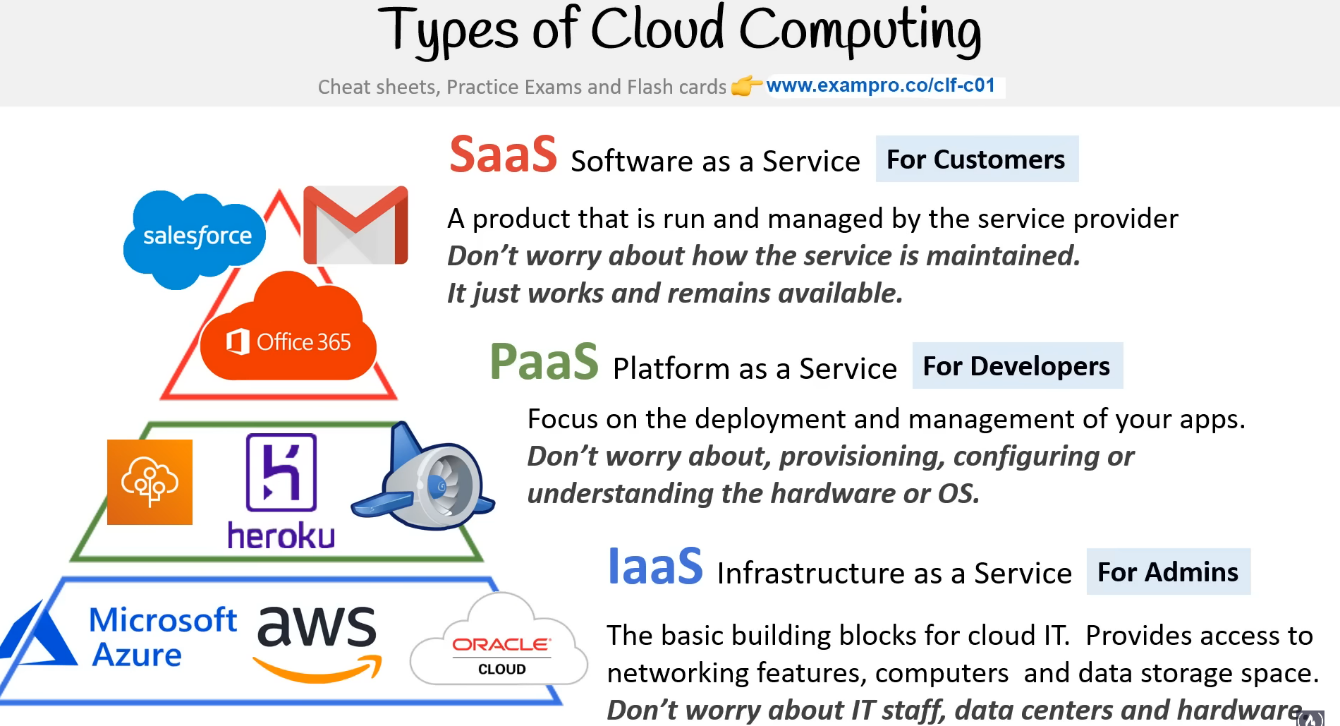
Cloud service provider tahat are infrastructure as a service (IaaS) will always have 4 core service offerings:

Compute - EC2 virtual machines

Storage - EBS (elastic block storage) virtual hard drives

Database - AQL databases

Networking (and content delivery) - VPC private cloud network



Amazon EC2 (Elastic Compute Cloud) is a web service provided by Amazon Web Services (AWS) that allows users to rent virtual servers, called instances, on which they can run their applications. EC2 provides scalable computing resources in the cloud, making it easier for businesses and developers to deploy and manage their applications without having to invest in physical hardware.

Key features of EC2 include:

1. Virtual Server Instances: EC2 allows users to create and manage virtual server instances in the cloud. These instances can be customized based on the specific requirements of the application, such as the desired CPU, memory, storage, and networking capacity.

2. Elasticity and Scalability: EC2 offers the ability to scale instances up or down based on the demand. This elasticity allows users to handle variations in traffic or workload by automatically adding or removing instances to match the application's needs.

3. Multiple Operating Systems: EC2 supports a wide range of operating systems, including various flavors of Linux, Windows Server, and others. This flexibility enables users to run their applications on their preferred operating system.

4. Security and Networking: EC2 provides several security features, including virtual private clouds (VPCs), security groups, and network access control lists (ACLs) to help users secure their instances. Users can also configure networking settings, such as IP addresses, subnets, and routing tables.

5. Storage Options: EC2 offers various storage options, including Amazon Elastic Block Store (EBS) for persistent block-level storage, Amazon S3 for object storage, and instance store volumes for temporary storage. These options allow users to choose the storage solution that best fits their application's requirements.

6. Integration with Other AWS Services: EC2 seamlessly integrates with other AWS services, such as Amazon S3, Amazon RDS (Relational Database Service), and Amazon VPC, enabling users to build complex and scalable architectures for their applications.

7. Billing and Cost Management: EC2 follows a pay-as-you-go model, allowing users to pay only for the resources they consume. Users can choose from various pricing options, such as on-demand instances, reserved instances, and spot instances, depending on their budget and usage patterns.

Overall, Amazon EC2 provides a flexible and scalable infrastructure for running applications in the cloud. It offers a wide range of features and configurations to accommodate different workloads, making it a popular choice for businesses and developers looking for reliable and cost-effective computing resources.

Amazon S3 (Simple Storage Service) is a highly scalable and durable object storage service provided by Amazon Web Services (AWS). It is designed to store and retrieve any amount of data from anywhere on the web, making it suitable for a wide range of use cases, from simple backup and archiving to large-scale data storage for applications.

Here are the key aspects of Amazon S3:

1. Object Storage: S3 allows you to store and organize data as objects, which consist of the data itself, a unique key, and metadata. Objects can be up to 5 terabytes in size and are stored in containers called buckets. Each object is assigned a unique URL, making it easily accessible via HTTP or HTTPS protocols.

2. Scalability and Durability: S3 is built to provide high scalability and durability. It automatically scales to accommodate any amount of data you store in it and can handle high request rates. S3 stores multiple copies of your data across different availability zones within a region, ensuring durability and high availability.

3. Data Lifecycle Management: S3 offers features for managing the lifecycle of your data. You can define rules to automatically transition objects between storage classes based on their age or other criteria. This allows you to optimize costs by moving less frequently accessed data to lower-cost storage options.

4. Security and Access Control: S3 provides robust security features to protect your data. You can set access control policies at the bucket and object level, defining who can access your data and what they can do with it. S3 also supports encryption at rest and in transit to ensure the confidentiality and integrity of your data.

5. Versioning and Data Replication: S3 allows you to enable versioning for your buckets, which keeps multiple versions of an object as it evolves over time. This helps in data recovery and provides protection against accidental deletions or overwrites. S3 also supports cross-region replication, allowing you to replicate your data to different regions for disaster recovery and low-latency access.

6. Data Transfer and Integration: S3 provides several mechanisms for transferring data into and out of the service. You can use the S3 API, AWS CLI (Command Line Interface), or SDKs (Software Development Kits) to interact with S3 programmatically. Additionally, S3 integrates with other AWS services, such as AWS Lambda, AWS Glue, and Amazon Athena, enabling you to build powerful data processing and analytics workflows.

7. Cost-effective Pricing Model: S3 follows a pay-as-you-go pricing model, where you only pay for the storage you consume, data transfer, and any additional features you use. S3 offers various storage classes with different performance characteristics and pricing options, allowing you to choose the most cost-effective solution for your data storage needs.

Overall, Amazon S3 provides a highly reliable, scalable, and secure object storage solution in the cloud. Its simplicity, durability, and wide range of features have made it a popular choice for storing and managing data for a variety of applications and use cases.

When you programmatically log in to AWS you will use access key and secret key.

Load balancing- involves connecting multiple servers together to form a server pool or a cluster. The load balancer sits in front of these servers and distributes incoming traffic across them based on the load balancing algorithm in use.

The servers in the server pool are typically configured in a similar manner, meaning they have the same application code, content, and configuration. This ensures that each server is capable of handling the same workload and serving the same content.

When a client makes a request to access the application or website, it is received by the load balancer. The load balancer then decides which server in the server pool should handle the request based on the configured load balancing algorithm. It distributes the incoming requests across the servers, ensuring an even distribution of the workload.

By distributing the traffic among multiple servers, load balancing helps to increase the overall capacity and scalability of the system. It allows the application or website to handle a larger number of concurrent users and ensures that no single server gets overwhelmed with traffic.

Additionally, load balancers provide fault tolerance and high availability. If one server in the server pool becomes unavailable or experiences issues, the load balancer can automatically redirect traffic to the remaining healthy servers. This ensures that the application or website remains accessible even in the event of a server failure.

Overall, load balancing helps to optimize resource utilization, improve performance, and enhance the reliability of applications or websites by distributing traffic across multiple servers.

It has a “Health check” method for checking the status of the servers. If they are running or not.

In AWS, there are two main types of load balancers: Application Load Balancer (ALB) and Network Load Balancer (NLB). Here are the differences between the two:

1. Layer of Operation:

- ALB operates at Layer 7 (the application layer) of the OSI model. It can intelligently distribute traffic based on application-specific content, such as HTTP headers, URL paths, or request methods. ALB is ideal for web applications that require advanced routing capabilities and support for HTTP and HTTPS protocols.

- NLB operates at Layer 4 (the transport layer) of the OSI model. It balances traffic based on IP addresses and ports. NLB is primarily designed for applications that require high throughput, low latency, and direct TCP or UDP traffic forwarding, without the need for advanced HTTP-level routing.

2. Routing and Load Balancing:

- ALB supports advanced routing features, including content-based routing, host-based routing, path-based routing, and redirecting HTTP requests to HTTPS. It can route requests to different backend services based on specific conditions. ALB also supports HTTP/2 and WebSocket protocols.

- NLB provides simple load balancing and operates at the TCP or UDP level. It performs efficient load balancing without modifying the packet headers, making it suitable for protocols that require direct access to client IP addresses or for scenarios where source IP preservation is essential, such as in a VPN setup.

3. Target Types:

- ALB supports a wide range of target types, including EC2 instances, IP addresses, and AWS Lambda functions. It can route traffic to multiple targets within a target group based on the defined routing rules.

- NLB primarily supports EC2 instances as targets, providing high-performance load balancing for TCP and UDP traffic. It is often used for scenarios that require ultra-low latency and high throughput, such as gaming, media streaming, and IoT applications.

4. SSL/TLS Termination:

- ALB provides built-in SSL/TLS termination, allowing it to handle HTTPS traffic directly. It can offload the SSL/TLS encryption and decryption process from the backend servers, improving performance and reducing the processing load on the application servers.

- NLB does not provide SSL/TLS termination. It forwards encrypted traffic (HTTPS) directly to the backend servers, requiring the servers to handle SSL/TLS decryption themselves.

5. Availability Zones:

- Both ALB and NLB support distributing traffic across multiple Availability Zones for high availability and fault tolerance. However, NLB can handle millions of requests per second while maintaining low latencies, making it suitable for extreme-scale applications.

Choosing between ALB and NLB depends on the specific requirements of your application. ALB is often the preferred choice for most web applications due to its advanced routing capabilities and support for HTTP/HTTPS protocols. NLB is typically used in scenarios that require high throughput and direct access to client IP addresses, such as high-performance network-intensive applications.

AMI – amazon machine image – it’s bootable copy of an instance.

Volume- virtual hard drive.

Volumes exist on EBS

Snapshots exist on S3

Snapshots are point in time copies of volumes.

You can create AMIs from Volumes, Instances and Snapshots

Vertical scaling- is powering a computer or a server with resources, like by CPU ram or disk...

Automating scaling- min and max of servers by definition. And if there is less or more it controls the amount. And manages the workflow by demand. It can grow and shrink when needed.

The main differences between S3 buckets and EBS volumes are:

* Purpose: S3 buckets are used for object storage, while EBS volumes are used for block storage.
* Access Protocol: S3 buckets are accessed via HTTP/HTTPS protocols, while EBS volumes are accessed by EC2 instances as block devices.
* Storage Type: S3 buckets store objects of various sizes and types, while EBS volumes provide raw, block-level storage.
* Use Cases: S3 buckets are commonly used for backup, archiving, content distribution, and static website hosting. EBS volumes are used as the main storage for EC2 instances, databases, and file systems.

