some fundamental cloud computing terms.

- **Cloud Computing** – Delivery of computing services (e.g., servers, storage, databases, networking, software) over the internet ("the cloud").

- **On-Premises** – Traditional computing where all resources are managed locally within an organization's infrastructure.

- **Cloud Provider** – A company that offers cloud computing services (e.g., AWS, Microsoft Azure, Google Cloud).

- **Public Cloud** – Cloud infrastructure available to multiple customers over the internet (e.g., AWS, Google Cloud, Azure).

- **Private Cloud** – Cloud infrastructure dedicated to a single organization, offering more control and security.

- **Hybrid Cloud** – A combination of public and private clouds, allowing data and applications to be shared between them.

- **Multi-Cloud** – The use of multiple cloud services from different providers.

Cloud Service Models

- **IaaS (Infrastructure as a Service)** – Provides virtualized computing resources like servers and storage (e.g., virtual Machine).
-
- **PaaS (Platform as a Service)** – Provides a platform for developers to build applications without managing infrastructure (e.g., Google App services).
- **SaaS (Software as a Service)** – Software applications hosted and managed by providers, accessible via the internet (e.g., Gmail, Dropbox, Microsoft 365).

Cloud Deployment Models

- **Public Cloud** – Hosted by third-party providers and shared among multiple customers.
- **Private Cloud** – Exclusive cloud infrastructure for a single organization.
- **Hybrid Cloud** – A mix of public and private cloud solutions.

Key Cloud Components

- **Virtual Machine (VM)** – A software-based emulation of a physical computer.
- **Container** – A lightweight, portable unit that includes an application and its dependencies (e.g., Docker, Kubernetes).
- **microservices** – A cloud-native architectural approach where applications are broken into small, independent services.

Storage and Networking

- **Object Storage** – Storage model that manages data as objects (Azure Blob Storage).
- **Block Storage** – Storage model where data is stored in fixed-sized blocks (e.g., Azure Disk Storage).
- **File Storage** – Traditional storage where data is stored in hierarchical file structures (e.g. Azure Files).

Security and Compliance

- **IAM (Identity and Access Management)** – Controls who can access cloud resources (e.g., Azure AD).

- **Encryption** – The process of securing data by converting it into unreadable text.

- **Firewall** – Security system that monitors and controls incoming and outgoing network traffic.

- **DDoS Protection** – Defenses against Distributed Denial of Service attacks, which overwhelm a service with traffic.

Cloud Cost and Billing

- **Pay-as-you-go** – Pricing model where you only pay for the resources you use.
- **Reserved Instances** – Cloud resources purchased for a long-term commitment to save costs.
- **Auto-Scaling** – Automatically adjusts cloud resources based on demand to optimize cost and performance.

Latency in Cloud Computing

Latency refers to the delay or time it takes for data to travel from one point to another in a network. In cloud computing, latency impacts the performance of applications and services, especially those requiring real-time processing.

Types of Latency

1. **Network Latency** – The time it takes for data to travel between a user's device and the cloud server.
2. **Processing Latency** – The delay caused by processing requests on a server.
3. **Storage Latency** – The time taken to retrieve data from cloud storage.
4. **Application Latency** – The delay in response due to application logic and dependencies.

Causes of Latency

- **Physical Distance** – The farther the data has to travel, the higher the latency.
- **Network Congestion** – High traffic can slow down data transfer.
- **Server Load** – Overloaded cloud servers may take longer to process requests.
- **DNS Resolution Time** – The time taken to translate a domain name into an IP address.

- **Security Measures** – Firewalls, encryption, and security checks can introduce delays.

How to Reduce Latency

- **Use CDNs (Content Delivery Networks)** – Store copies of data closer to users.
- **Choose Cloud Regions Wisely** – Deploy resources in cloud regions closer to end users.
- **Implement Edge Computing** – Process data closer to the source instead of centralized cloud data centers.
- **Optimize Network Routing** – Use private or dedicated connections (e.g., AWS Direct Connect, Azure ExpressRoute).
- **Load Balancing** – Distribute traffic across multiple servers to avoid overload

Key Concepts of Availability

1. **High Availability (HA)** – Ensuring minimal downtime by using redundant systems and failover mechanisms.

2. **Fault Tolerance** – The ability of a system to continue functioning even when one or more components fail.

3. **Uptime SLA (Service Level Agreement)** – Cloud providers guarantee a percentage of uptime (e.g., Azure offers 99.9% to 99.99% uptime based on services used).
4. **Failover** – Automatically switching to a backup system in case of failure.

5. **Redundancy** – Duplicating critical components (e.g., multiple data centers, servers, or databases) to prevent single points of failure.

6. **Disaster Recovery (DR)** – Strategies to restore services quickly after a failure or disaster (e.g., backups, geo-replication).

Elasticity in Cloud Computing

Elasticity refers to the ability of a cloud system to **automatically scale resources up or down** based on demand. It ensures optimal performance while minimizing costs by dynamically adjusting computing power.

Scaling in Cloud Computing

Scaling refers to the process of increasing or decreasing cloud resources to meet workload demands. It ensures that applications remain **responsive, efficient, and cost-effective** under varying loads.

Types of Scaling

1. Vertical Scaling (Scaling Up/Down)

- Increases or decreases the **capacity** of an existing resource.
- Example: Upgrading a virtual machine (VM) with more CPU, RAM, or storage.
- **Pros:** Simple and requires no architecture changes.
- **Cons:** Limited by hardware capacity.

💡 **Example in Azure:**
Upgrading an Azure Virtual Machine from **Standard_B2s (2 vCPUs, 4GB RAM)** to **Standard_B4ms (4 vCPUs, 16GB RAM)**.

2. Horizontal Scaling (Scaling Out/In)

- Adds or removes **multiple instances** of resources.
- Example: Increasing the number of VMs or containers to distribute the load.
- **Pros:** Unlimited scalability and better fault tolerance.
- **Cons:** Requires a load balancer to distribute traffic.

Difference Between Scaling and Elasticity in Cloud Computing

Feature	Scaling	Elasticity
Definition	The process of increasing or decreasing cloud resources (manually or	The ability of a system to **automatically** adjust resources dynamically

	automatically) to meet demand.	based on real-time demand.
Scaling Type	Can be **manual or automatic** (Horizontal, Vertical, or Diagonal Scaling).	Always **automatic** and demand-driven.
Time Frame	Can be planned for **long-term growth**.	Works in **real-time** to handle sudden changes in workload.
Resource Adjustment	Resources are increased or decreased as needed, but may not always revert back automatically.	Resources **scale up and down automatically** to match demand, ensuring efficiency.
Flexibility	May require manual intervention or predefined rules.	Fully **automated and adaptive** without user intervention.
Cost Efficiency	Can lead to over-provisioning if not managed well.	Optimized cost management as resources **shrink** when demand decreases.
Example	- Adding more VMs (horizontal scaling) for a growing app. - Upgrading a VM to a higher tier (vertical scaling).	- A serverless function in Azure **scaling up automatically** when requests increase and **scaling down** when traffic decreases.

Analogy for Better Understanding

- **Scaling** = Buying more trucks or upgrading them **before a big shipment** (pre-planned capacity increase).
- **Elasticity** = Having trucks that **automatically appear when needed** and disappear when not (real-time demand adjustment).

Azure Services Supporting Scaling & Elasticity

Service	**Scaling**	**Elasticity**
Azure Virtual Machines (VMs)	Manual or auto-scaling	✕ No auto-reduction
Azure Virtual Machine Scale Sets (VMSS)	Auto-scaling of VMs	✓ Elasticity in VM instances
Azure App Service Auto-Scaling	Auto-scaling of web apps	✓ Elastic scaling based on traffic
Azure Kubernetes Service (AKS)	Scales containers	✓ Elastic container management

Service Level Agreement (SLA) in Cloud Computing

A **Service Level Agreement (SLA)** is a contract between a cloud provider and a customer that defines the level of **service availability, performance, and reliability** guaranteed by the provider. SLAs ensure that businesses receive the expected level of cloud services and compensation if the provider fails to meet the agreement.

Key Components of an SLA

1. **Availability (Uptime Guarantee)**
   - Specifies the percentage of time a service is available.
   - Example: **99.9% uptime guarantee** means a maximum of **8.76 hours of downtime per year**.
2. **Performance Metrics**
   - Defines response times, transaction processing speeds, and network latency limits.
   - Example: API response time must be **less than 100ms**.
3. **Downtime & Maintenance Policies**
   - Specifies planned and unplanned outages.
   - Example: "Scheduled maintenance will occur between 2 AM - 4 AM UTC."
4. **Compensation & Penalties**
   - Refunds or service credits if the provider fails to meet SLA commitments.
   - Example: If availability drops below **99.9%**, users receive a **10% credit**.
5. **Support & Incident Response Times**
   - Defines response and resolution times for different issue severities.

   - Example: **Critical issues resolved within 4 hours**.

6. **Security & Compliance**
   - Includes data protection, encryption, and compliance with regulations (e.g., GDPR, HIPAA).
7. **Termination Clause**
   - Defines conditions for contract termination if the SLA is consistently violated.

Azure SLA Guarantees

Azure Service	**SLA Uptime Guarantee**
Virtual Machines (Single VM)	99.9%
Virtual Machines (with Availability Zones)	99.99%
Azure SQL Database	99.99%
Azure Storage (RA-GRS)	99.99%