02 Cloud Computing Intro

Cloud Computing: Core Concepts and AWS Focus

Cloud Computing Introduction

Cloud Computing is the **on-demand delivery** of Information Technology (IT) resources and applications via the **internet** with pay-as-you-go pricing.

- In essence, instead of incurring large capital expenditures (CapEx) to **buy, own, and maintain** physical infrastructure like servers, networking gear, and data centers, organizations **rent and access** these resources as a utility from a third-party cloud service provider (CSP).
- This paradigm shift allows for provisioning and de-provisioning of resources **programmatically and rapidly**.
- Examples of IT resources delivered via the cloud include:
  - **Core Infrastructure**: Virtual Machines (VMs), Databases, Storage, and Networking.
  - **Advanced Services**: Internet of Things (IoT) platforms, Machine Learning (ML), Artificial Intelligence (AI) services, serverless compute, and Big Data analytics tools.

Key Benefits of Cloud Computing

Cloud computing offers compelling advantages that drive digital transformation across industries:

- **Cost Efficiency & Lower Upfront Costs (CapEx to OpEx)**: It shifts spending from large **Capital Expenditures (CapEx)**—buying data centers and equipment—to flexible **Operational Expenditures (OpEx)**—paying only for resources consumed. This eliminates the need for expensive, speculative hardware purchases.
- **Massive Scalability and Elasticity**: Resources can be scaled **up or down automatically and rapidly** based on current demand. This **elasticity** ensures you have capacity when traffic surges and are not paying for idle resources during quiet times.
- **Increased Agility and Speed**: Developers can deploy databases, virtual machines, networks, and complete environments **in minutes** at the click of a button or via an API, dramatically reducing time-to-market for new features and applications.
- **Global Reach**: CSPs have massive, globally distributed data centers. Organizations can deploy their applications in different **geographic regions** to achieve lower latency for global users and meet data residency requirements.
- **Security, Reliability, and High Availability**: CSPs invest heavily in security, often providing better protection than individual companies can afford. They offer services that ensure high **fault tolerance, data backup, and disaster recovery**.
- **Focus on Business Value**: By outsourcing infrastructure management to the CSP, the customer's IT staff is freed up to focus on **core business competencies** and innovation rather than infrastructure patching and maintenance.

The NIST Definition of Cloud Computing

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. This cloud model is composed of five essential characteristics, three service models, and four deployment models.

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. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.

Rapid elasticity. Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

Measured service. Cloud systems automatically control and optimize resource use by leveraging a metering capability[1] at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

CAPEX

Assets purchased upfront to meet business requirements

One-time purchase at the beginning and use till end of life

Need to own and maintain the assets purchased

Examples
- Buying or renting datacentres
- Buying and maintaining hardware e.g. servers, storage arrays etc.
- Buying software license upfront

OPEX

Ongoing expenses to run business

Flexible pay as you need

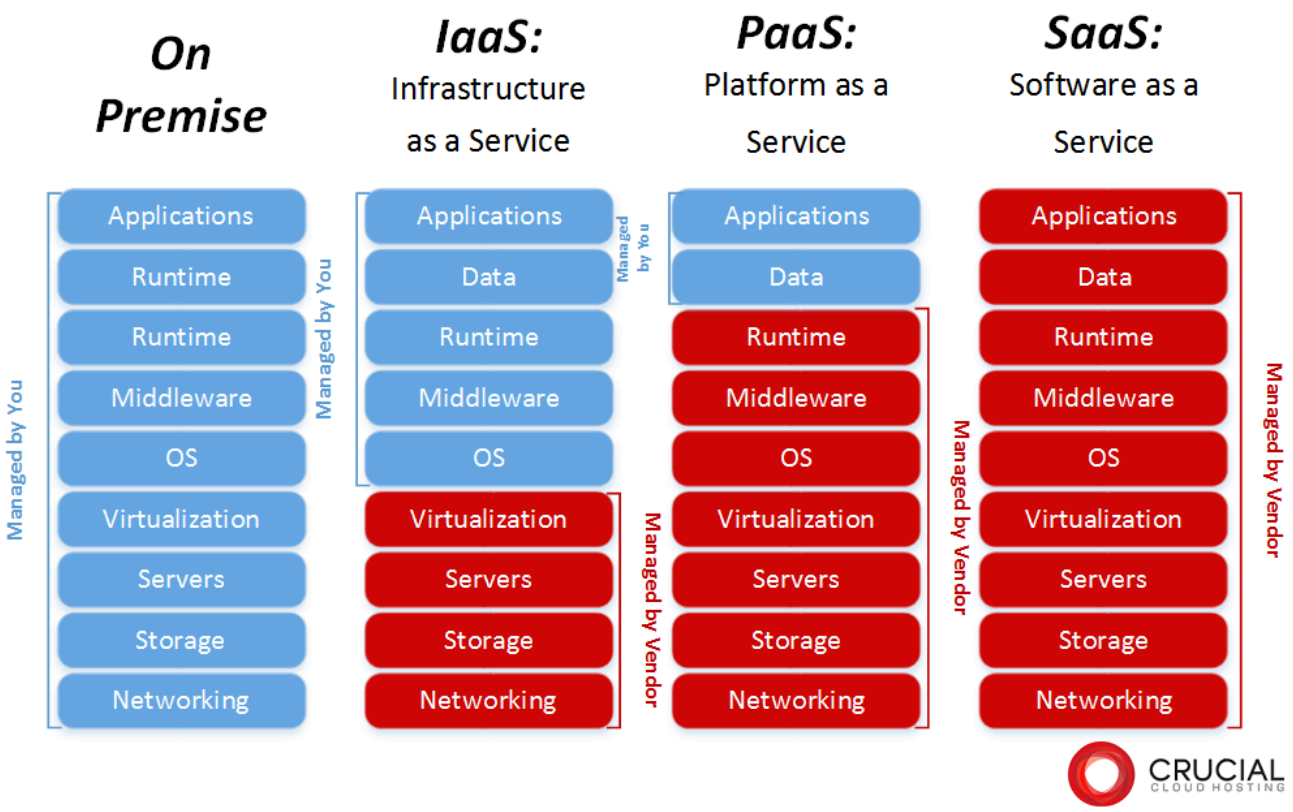
Provider maintains asset while customers focus on their core functionality

Examples
- No need to worry about datacentres
- Paying monthly for Azure services as consumed (e.g. pay monthly for storage and compute based on consumption)
- Subscribing for software and paying monthly based on consumption

The Shared Responsibility Model

In the cloud, security is a **shared responsibility** between the Cloud Service Provider (CSP) and the customer. Understanding this division of labor is crucial for maintaining a secure environment.

Area of Responsibility	Cloud Service Provider (e.g., AWS)	Customer
Security OF the Cloud	Protecting the **global infrastructure** (Regions, Availability Zones, Edge Locations). This includes the physical facilities, hardware, networking, and the underlying virtualization layer.	**Security IN the Cloud**
Examples of CSP Responsibility	Physical security of data centers, foundational networking, hardware patching, operating system management for managed services (e.g., fully managed databases like Amazon RDS).	Customer data, application security, identity and access management (IAM), operating system configuration (for IaaS VMs), network traffic protection (e.g., security groups), and client-side encryption.



Cloud Service Models

Cloud service models define the level of management and control a customer retains over their resources.

1. Infrastructure as a Service (IaaS)

- **Definition**: The most flexible and foundational cloud service, providing the **building blocks for cloud IT**. It provides virtualized computing resources—servers (VMs), storage, and networks—managed over the internet.
- **Control**: Customers are responsible for the **Operating System (OS)**, runtime, application, and data. The CSP manages the underlying infrastructure (servers, storage, virtualization). This offers the **maximum control and flexibility**.

- **Use Cases**: Hosting web applications, traditional "Lift & Shift" migration of on-premises workloads, development/test environments, and big data processing.

AWS IaaS Examples:

1. **Amazon EC2 (Elastic Compute Cloud)**: Provides resizable **virtual servers** (VMs) in the cloud.
2. **Amazon S3 (Simple Storage Service)**: **Massively scalable object storage** for unstructured data.
3. **Amazon EBS (Elastic Block Store)**: Provides **persistent block storage** volumes for use with EC2 instances.
4. **Amazon VPC (Virtual Private Cloud)**: Enables customers to provision a **logically isolated section** of the AWS Cloud to launch resources.
5. **AWS Direct Connect**: A dedicated **private network connection** from a data center to AWS.

2. Platform as a Service (PaaS)

- **Definition**: A middle ground where the CSP manages the hardware, OS, and often the application stack (middleware, runtime, database). The customer focuses solely on **developing and deploying their application code and data**.
- **Control**: The customer has less control over the underlying environment but **increased speed and agility** because they are abstracted away from OS patching and infrastructure maintenance.
- **Analogy**: Think of it as a **ready-to-use factory floor** (platform) where you only need to bring your raw materials (code) and start production.
- **Use Cases**: Rapid development and deployment of applications, web application hosting, and development of frameworks.

AWS PaaS Examples:

1. **AWS Elastic Beanstalk**: An **easy-to-use service** for deploying and scaling web applications and services.
2. **AWS Lambda**: A **serverless compute service** that lets you run code without provisioning or managing servers. (Often considered Function-as-a-Service, a subset of PaaS).
3. **Amazon RDS (Relational Database Service)**: A **managed relational database service** that handles setup, patching, and backups for popular database engines (e.g., MySQL, PostgreSQL, Oracle).
4. **AWS Fargate**: A **serverless compute engine** for containers (Amazon ECS and EKS), eliminating the need to manage the underlying server infrastructure.
5. **Amazon SageMaker**: A fully managed service that allows data scientists and developers to **build, train, and deploy Machine Learning models** quickly.

3. Software as a Service (SaaS)

- **Definition**: The most complete cloud offering, delivering a **fully functional application** over the internet. The entire application stack—from the underlying infrastructure to the application itself—is managed by the CSP.
- **Control**: The customer has the **least flexibility** in terms of underlying infrastructure but is only responsible for user access management (IAM) and their data. It is the easiest to deploy and use.
- **Analogy**: Using a **ready-made, off-the-shelf product** like an email service.

- **Use Cases**: Email (Gmail, Outlook), Customer Relationship Management (CRM) tools (Salesforce), and Enterprise Resource Planning (ERP) tools.

AWS SaaS Examples:

1. **Amazon WorkSpaces**: A **managed, secure Desktop-as-a-Service (DaaS)** solution.
2. **Amazon Connect**: A **cloud-based contact center service**.
3. **Amazon QuickSight**: A scalable, serverless **Business Intelligence (BI) service** used to analyze data and create interactive dashboards.
4. **Amazon Chime**: A **communications service** for online meetings, video conferencing, and chat.
5. **AWS Marketplace**: An **online store** where you can find, buy, and immediately deploy software and services from third-party vendors.

Cloud Deployment Models

Cloud deployment models define the location and management of the cloud infrastructure.

1. Public Cloud

- **Definition**: Cloud services are owned and operated by a **third-party vendor** (like AWS) and delivered over the public internet. Resources are **shared** among multiple customers (tenants).
- **Characteristics**:
  - **Shared Infrastructure (Multi-tenant)**: Resources are shared among many organizations, optimizing utilization and cost.
  - **High Scalability & Agility**: Pay-as-you-go model allows for massive, rapid scaling.
  - **Cost-Effective**: Low upfront costs.
- **Example**: **Amazon Web Services (AWS)**, which offers its vast array of services to the general public and businesses worldwide.

2. Private Cloud

- **Definition**: Cloud computing infrastructure operated **solely for a single organization**. It can be physically located in the organization's on-premises data center or hosted by a third-party service provider in a dedicated environment.
- **Characteristics**:
  - **Dedicated & Isolated**: Resources are not shared with other organizations, providing enhanced privacy and security.
  - **Full Control**: Organizations have complete control over the infrastructure, which is crucial for meeting strict regulatory and compliance requirements.
  - **High Cost**: Higher initial investment and ongoing maintenance compared to Public Cloud.
- **Example**: A financial institution building a cloud environment within its own data center to handle sensitive customer transactions.

3. Hybrid Cloud

- **Definition**: A combination of a **Public Cloud** and a **Private Cloud** (either on-premises or a dedicated environment) that are connected by technology that allows **data and applications to**

be shared and moved between them.

- **Characteristics**:
  - **Flexibility and Workload Placement**: Organizations can place sensitive or mission-critical workloads in the Private Cloud while leveraging the Public Cloud for non-sensitive, elastic workloads.
  - **Cloud Bursting**: Using the Public Cloud to handle spikes in traffic that the Private Cloud cannot manage.
  - **Compliance and Optimization**: Balances the security/compliance of a private environment with the cost-effectiveness/scale of a public one.
- **Example**: Using an on-premises private cloud for core customer data and an AWS Public Cloud to host the customer-facing e-commerce front end that handles seasonal traffic surges.

4. Multi-Cloud

- **Definition**: The utilization of **multiple public cloud providers** (e.g., using both AWS and Azure simultaneously). Note that a Hybrid Cloud involves one Public Cloud and one Private Cloud, while Multi-Cloud involves two or more Public Clouds.
- **Characteristics**:
  - **Vendor Lock-in Avoidance**: Reduces reliance on a single CSP.
  - **Best-of-Breed Services**: Allows the organization to pick the best service from different providers (e.g., using a specific ML service from one CSP and a database service from another).
  - **Increased Resiliency**: Provides disaster recovery options across different cloud platforms.