Unit-2: Virtualization and Resource Management

Elasticity and Scalability in Cloud Computing

Introduction to Cloud Computing

- Cloud computing provides on-demand computing resources, allowing businesses and individuals to use computing power, storage, and services without maintaining physical infrastructure.

- Two key concepts that enhance the efficiency of cloud computing are Elasticity and Scalability.

Scalability in Cloud Computing

Definition

Scalability is the ability of a cloud system to handle increasing workloads by adding (or removing) resources efficiently.

Types of Scalability

Vertical Scalability (Scaling Up/Down)

- Increasing or decreasing the capacity of a single machine (e.g., upgrading CPU, RAM, or storage).

Horizontal Scalability (Scaling Out/In)

- Adding or removing multiple machines to distribute the workload across multiple servers.

Diagonal Scalability

- A combination of both vertical and horizontal scaling for optimized performance.

Key Characteristics

- Ensures performance stability even with increased workloads.

- Involves proactive planning to accommodate future growth.

- Essential for high availability applications, such as banking and social media platforms.

Examples of Scalability

- Netflix scales its infrastructure horizontally to support millions of concurrent users.
- Facebook and Google Cloud Bigtable use distributed databases to handle massive amounts of data.
- Amazon Web Services (AWS) Elastic Load Balancer (ELB) distributes traffic efficiently among multiple servers.

Elasticity in Cloud Computing

Definition

Elasticity refers to the ability of a cloud system to dynamically allocate or deallocate resources based on demand, ensuring optimal performance and cost efficiency.

Key Characteristics

- Cloud platforms can automatically adjust resources in real-time.
- Users pay only for the resources they use, reducing costs.
- Resources are added or removed based on workload fluctuations.
- Best suited for applications with fluctuating resource demands (e.g., e-commerce sites during sales events).

Examples of Elasticity

AWS Auto Scaling

- Automatically adjusts EC2 instances based on demand.

Microsoft Azure Virtual Machine Scale Sets

- Dynamically adjusts the number of VM instances.

Google Cloud Compute Engine Autoscaler

- Allocates or deallocates resources in response to traffic.

Comparison: Elasticity vs. Scalability

Feature	Elasticity	Scalability
Definition	Adjusts resources dynamically based on demand	Expands infrastructure to accommodate growth
Response Time	Real-time, automatic adjustment	Requires pre-planning and resource allocation
Cost Efficiency	High, as resources are allocated only when needed	Moderate, as it requires additional infrastructure
Workload Suitability	Ideal for unpredictable workloads	Best for long-term growth and stable workloads
Example	Auto-scaling e-commerce platform	Expanding data centers for a growing business

Case Studies on Virtualization Technologies

1. VMware Virtualization in Healthcare

2. Microsoft Hyper-V in the Financial Sector

3. Google Cloud Virtualization for E-Commerce

https://www.youtube.com/watch?v=fHnkuAdii_0