CS351 - Cloud Computing

Lecture #5

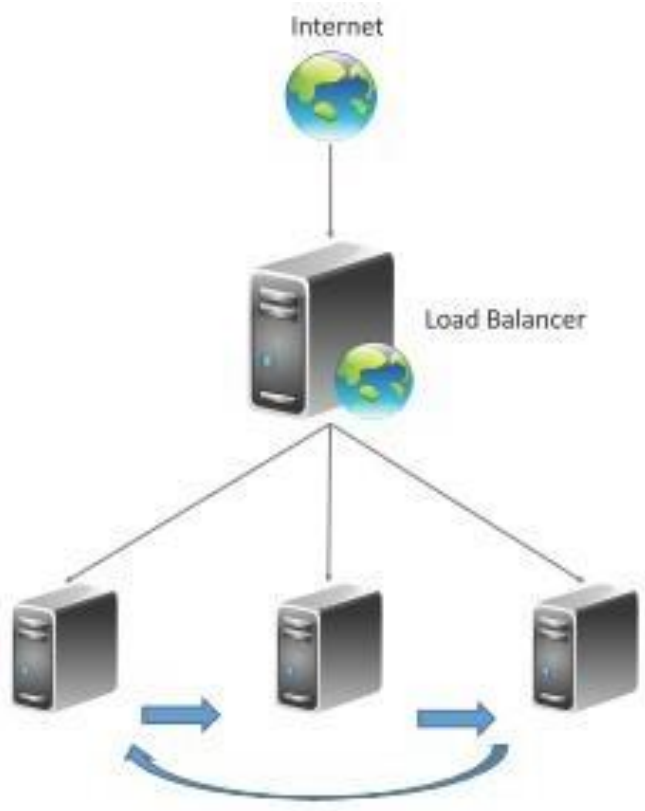
Dr. Ferdous Ahmed Barbhuiya
Indian Institute of Information Technology Guwahati

Enabling Techlogies for Cloud

- Concepts and enabling technologies of cloud computing
  - **Virtualization**
  - Load balancing
  - Scalability & Elasticity
  - Deployment
  - Replication
  - Monitoring
  - Identity and Access Management
  - Service Level Agreements
  - Billing

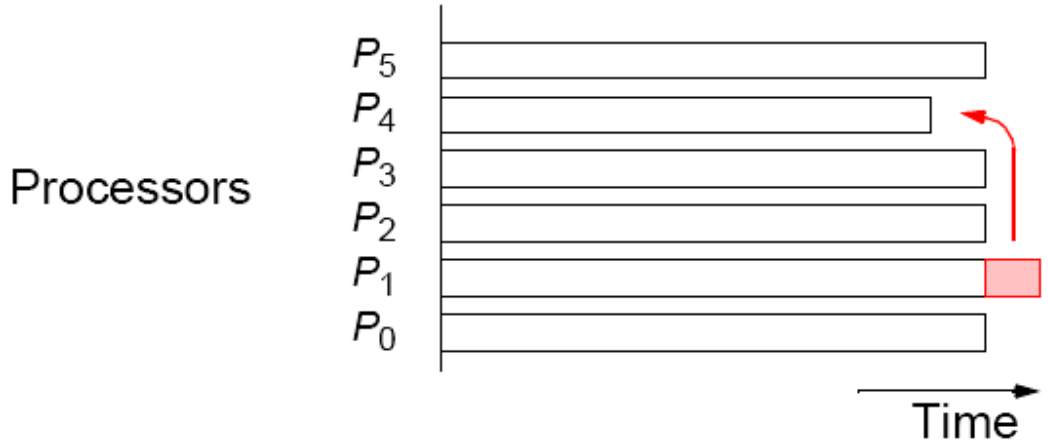
Load Balancing

- Cloud computing resources can be scaled up on demand to meet the performance requirements of applications.
- Load balancing distributes workloads across multiple servers to meet the application workloads.
- The goals of load balancing techniques include:
  - Achieve maximum utilization of resources
  - Minimizing the response times
  - Maximizing throughput

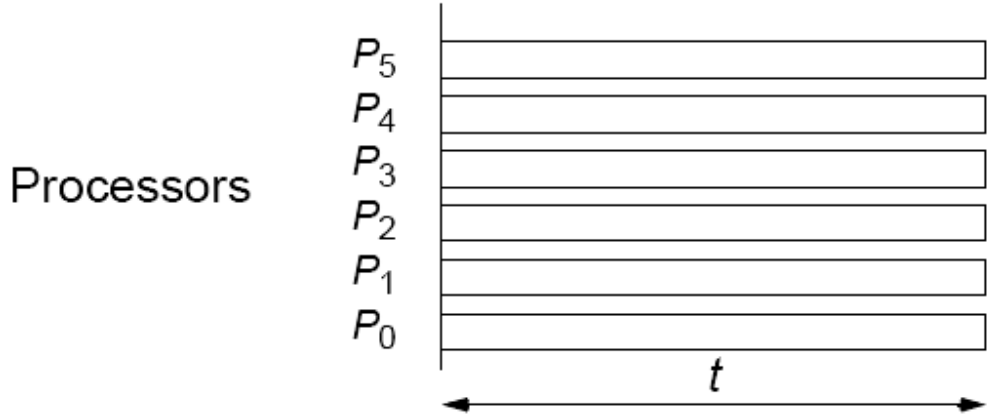


Load Balancing- What it is?

Load balancing – used to distribute computations fairly across processors in order to obtain the highest possible execution speed



(a) Imperfect load balancing leading to increased execution time



(b) Perfect load balancing

Concepts

Web server System

Providing web services

Trend:

1. Increasing number of clients
2. Growing complexity of web applications

Scalable Web server systems

The ability to support large numbers of accesses and resources while still providing adequate performance

Locally Distributed Web System

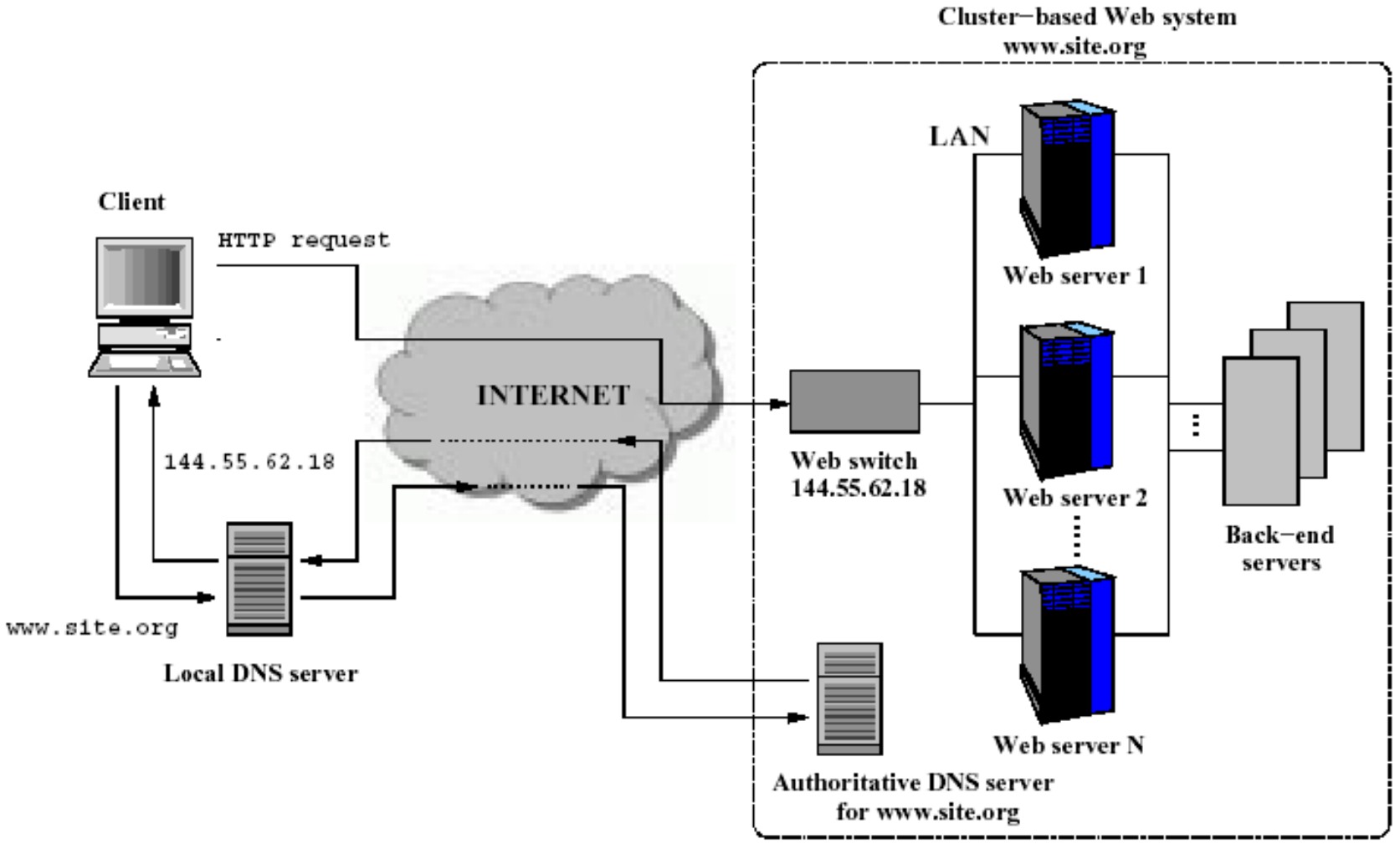
Cluster Based Web System

the server nodes mask their IP addresses to clients, using a Virtual IP address corresponding to one device (web switch) in front of the set of the servers – Web switch receives all packets and then sends them to server nodes

Distributed Web System

the IP addresses of the web server nodes are visible to clients. No web switch, just a layer 3 router may be employed to route the requests

Architecture



Load Balancing Algorithms

- Round Robin load balancing
- Weighted Round Robin load balancing
- Low Latency load balancing
- Least Connections load balancing
- Priority load balancing
- Overflow load balancing

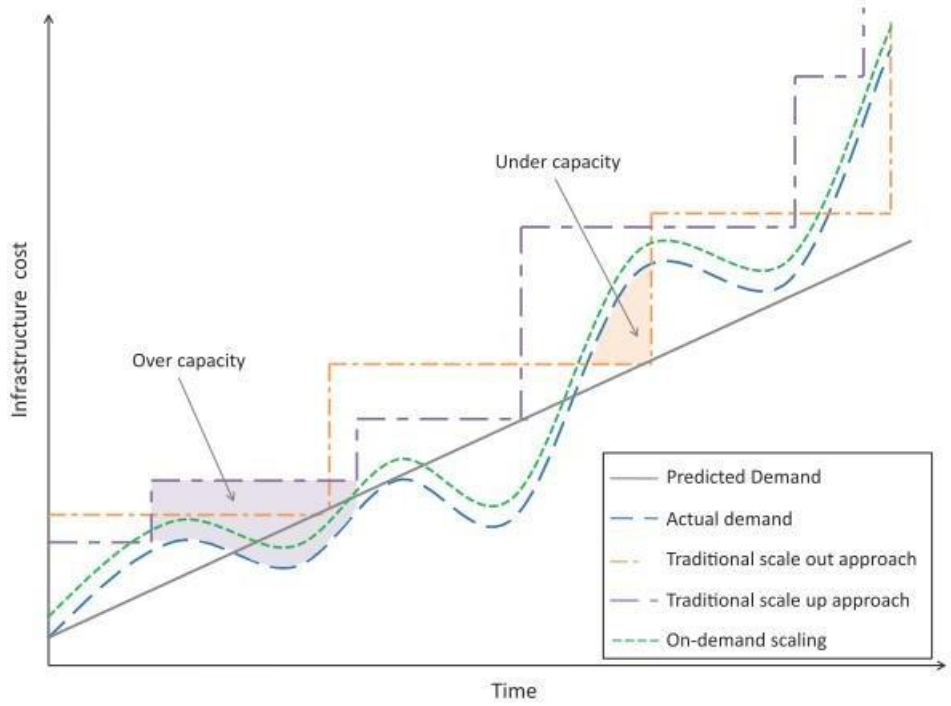
Load Balancing - Persistence Approaches

- Since load balancing can route successive requests from a user session to different servers, maintaining the state or the information of the session is important.

- Persistence Approaches
  - Sticky sessions
  - Session Database
  - Browser cookies
  - URL re-writing

Scalability & Elasticity

- Multi-tier applications such as e-Commerce, social networking, business-to-business, etc. can experience rapid changes in their traffic.
- Capacity planning involves determining the right sizing of each tier of the deployment of an application in terms of the number of resources and the capacity of each resource.
- Capacity planning may be for computing, storage, memory or network resources.



Scaling Approaches

- Vertical Scaling/Scaling up:
  - Involves upgrading the hardware resources (adding additional computing, memory, storage or network resources).

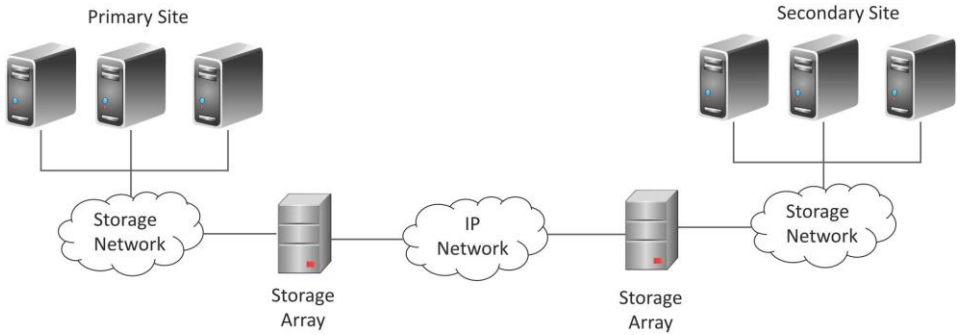
- Horizontal Scaling/Scaling out
  - Involves addition of more resources of the same type.

Deployment

- Cloud application deployment design is an iterative process that involves:
  - Deployment Design
    - The variables in this step include the number of servers in each tier, computing, memory and storage capacities of severs, server interconnection, load balancing and replication strategies.
  - Performance Evaluation
    - To verify whether the application meets the performance requirements with the deployment.
    - Involves monitoring the workload on the application and measuring various workload parameters such as response time and throughput.
    - Utilization of servers (CPU, memory, disk, I/O, etc.) in each tier is also monitored.
  - Deployment Refinement
    - Various alternatives can exist in this step such as vertical scaling (or scaling up), horizontal scaling (or scaling out), alternative server interconnections, alternative load balancing and replication strategies, for instance.

Replication

- Replication is used to create and maintain multiple copies of the data in the cloud.
- Cloud enables rapid implementation of replication solutions for disaster recovery for organizations.
- With cloud-based data replication organizations can plan for disaster recovery without making any capital expenditures on purchasing, configuring and managing secondary site locations.
- Types:
  - Array-based Replication
  - Network-based Replication
  - Host-based Replication



Array-based Replication

Monitoring

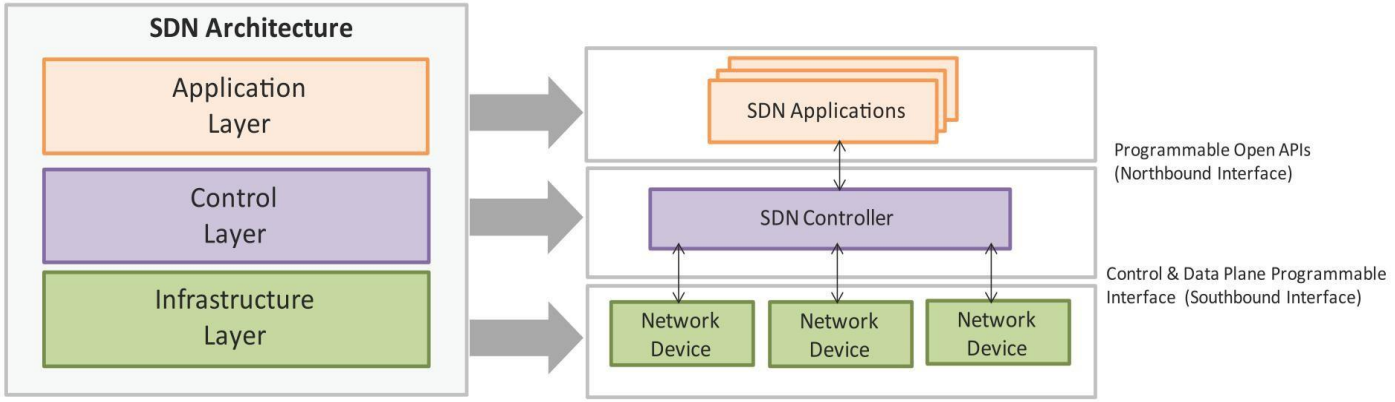
- Monitoring services allow cloud users to collect and analyze the data on various monitoring metrics.
- A monitoring service collects data on various system and application metrics from the cloud computing instances.
- Monitoring of cloud resources is important because it allows the users to keep track of the health of applications and services deployed in the cloud.

Examples of Monitoring Metrics

Type	Metrics
CPU	CPU-Usage, CPU-Idle
Disk	Disk-Usage, Bytes/sec (read/write), Operations/sec
Memory	Memory-Used, Memory-Free, Page-Cache
Interface	Packets/sec (incoming/outgoing), Octets/sec(incoming/outgoing)

Software Defined Networking

- Software-Defined Networking (SDN) is a networking architecture that separates the control plane from the data plane and centralizes the network controller.
- Conventional network architecture
  - The control plane and data plane are coupled. Control plane is the part of the network that carries the signaling and routing message traffic while the data plane is the part of the network that carries the payload data traffic.
- SDN Architecture
  - The control and data planes are decoupled and the network controller is centralized.



SDN - Key Elements

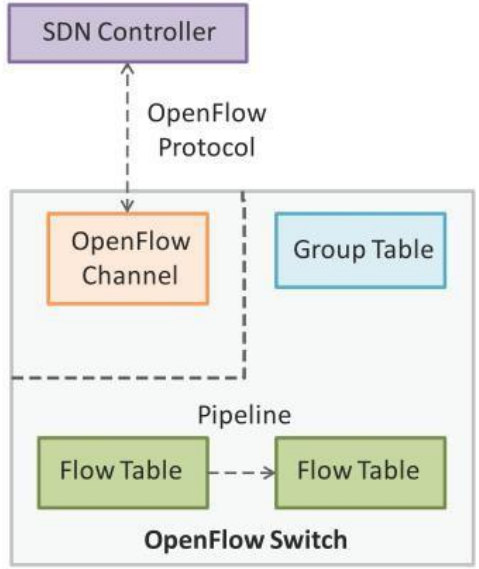
- Centralized Network Controller
  - With decoupled the control and data planes and centralized network controller, the network administrators can rapidly configure the network.

- Programmable Open APIs
  - SDN architecture supports programmable open APIs for interface between the SDN application and control layers (Northbound interface). These open APIs that allow implementing various network services such as routing, quality of service (QoS), access control, etc.

- Standard Communication Interface (OpenFlow)
  - SDN architecture uses a standard communication interface between the control and infrastructure layers (Southbound interface). OpenFlow, which is defined by the Open Networking Foundation (ONF) is the broadly accepted SDN protocol for the Southbound interface.

OpenFlow

- OpenFlow is the broadly accepted SDN protocol for the Southbound interface.
- With OpenFlow, the forwarding plane of the network devices can be directly accessed and manipulated.
- OpenFlow uses the concept of flows to identify network traffic based on pre-defined match rules.
- Flows can be programmed statically or dynamically by the SDN control software.
- OpenFlow protocol is implemented on both sides of the interface between the controller and the network devices.



OpenFlow switch comprising of one or more flow tables and a group table, which perform packet lookups and forwarding, and OpenFlow channel to an external controller.

Network Function Virtualization

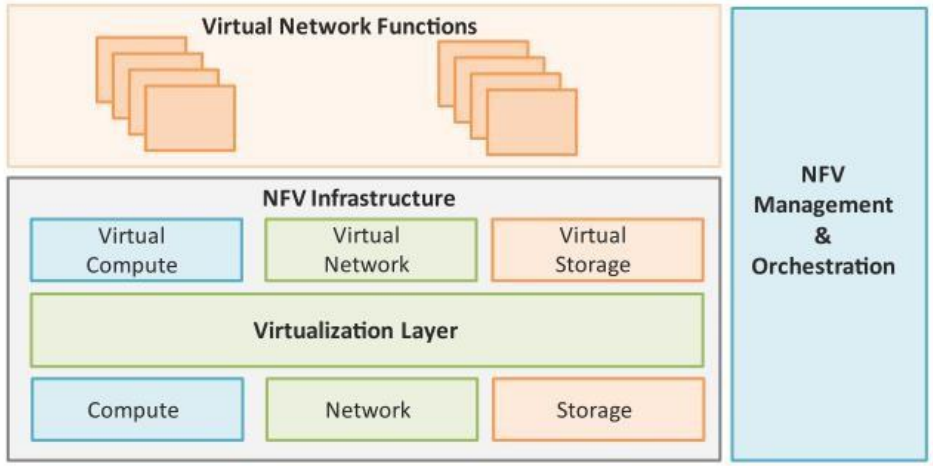
- Network Function Virtualization (NFV) is a technology that leverages virtualization to consolidate the heterogeneous network devices onto industry standard high volume servers, switches and storage.

- Relationship to SDN
  - NFV is complementary to SDN as NFV can provide the infrastructure on which SDN can run.
  - NFV and SDN are mutually beneficial to each other but not dependent.
  - Network functions can be virtualized without SDN, similarly, SDN can run without NFV.

- NFV comprises of network functions implemented in software that run on virtualized resources in the cloud.
- NFV enables a separation the network functions which are implemented in software from the underlying hardware.

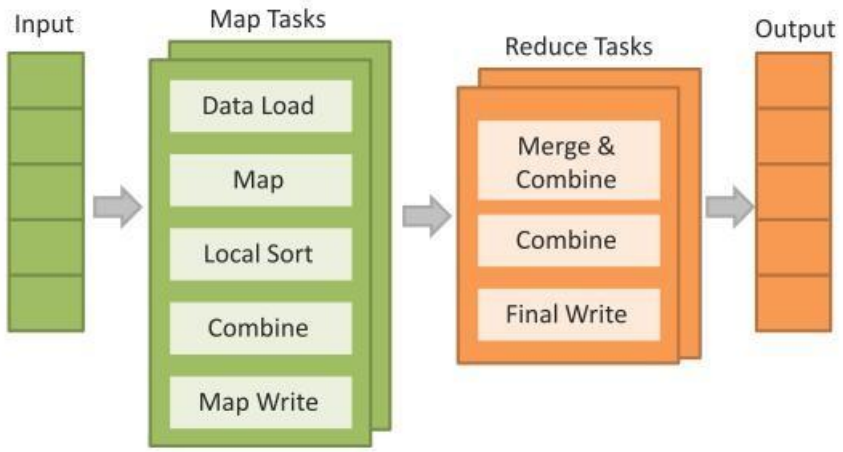
NFV Architecture

- Key elements of the NFV architecture are
  - Virtualized Network Function (VNF): VNF is a software implementation of a network function which is capable of running over the NFV Infrastructure (NFVI).
  - NFV Infrastructure (NFVI): NFVI includes compute, network and storage resources that are virtualized.
  - NFV Management and Orchestration: NFV Management and Orchestration focuses on all virtualization-specific management tasks and covers the orchestration and lifecycle management of physical and/or software resources that support the infrastructure virtualization, and the lifecycle management of VNFs.



MapReduce

- MapReduce is a parallel data processing model for processing and analysis of massive scale data.
- MapReduce phases:
  - **Map Phase**: In the Map phase, data is read from a distributed file system, partitioned among a set of computing nodes in the cluster, and sent to the nodes as a set of key-value pairs.
  - The Map tasks process the input records independently of each other and produce intermediate results as key-value pairs.
  - The intermediate results are stored on the local disk of the node running the Map task.
  - **Reduce Phase**: When all the Map tasks are completed, the Reduce phase begins in which the intermediate data with the same key is aggregated.



Identity and Access Management

- Identity and Access Management (IDAM) for cloud describes the authentication and authorization of users to provide secure access to cloud resources.
- Organizations with multiple users can use IDAM services provided by the cloud service provider for management of user identifiers and user permissions.
- IDAM services allow organizations to centrally manage users, access permissions, security credentials and access keys.
- Organizations can enable role-based access control to cloud resources and applications using the IDAM services.
- IDAM services allow creation of user groups where all the users in a group have the same access permissions.
- Identity and Access Management is enabled by a number of technologies such as OpenAuth, Role-based Access Control (RBAC), Digital Identities, Security Tokens, Identity Providers, etc.

Billing

Cloud service providers offer a number of billing models described as follows:

- Elastic Pricing
  - In elastic pricing or pay-as-you-use pricing model, the customers are charged based on the usage of cloud resources.
- Fixed Pricing
  - In fixed pricing models, customers are charged a fixed amount per month for the cloud resources.
- Spot Pricing
  - Spot pricing models offer variable pricing for cloud resources which is driven by market demand.