Cloud Computing (CS 435)

Table of Contents:

Lesson No. 01

INTRODUCTION TO COURSE

Module No – 001

- This course is designed to cover different dimensions of Cloud Computing.
- The course aims at providing good knowledge regarding the Cloud Computing.
- The course begins with the background knowledge.
- Cloud Computing has over lapping features with various other forms of computing.
- Basically it is the provision of computing (over virtualized resources) as a (charged for) service over the Internet.
- It is important to build some conceptual foundations.
- Cloud computing has somehow evolved from the fields of Cluster Computing and Grid Computing.
- The course begins with the overview of these field.
- Since the Cloud Computing aims at providing computations over Internet, this course builds and revises the basic and advanced networking concepts such as:
- Network structures, designs, protocols etc.
- Switching, routing and network virtualization.
- Data center networking
- Cloud Computing depends upon a technology called *Virtualization* technology for dynamic creation and provisioning of computing resources.
- Up-to-date knowledge of virtualization is discussed before proceeding.
- Different features of Cloud Computing with reference to course books and National Institute of Standards and Technology (NIST) USA will be introduced/explained as and when an enough background knowledge is gained.
- The course continues with the coverage of Cloud architecture, services, examples and issues related to security and privacy etc. for Cloud Computing.
- Later on, the in-depth coverage of Cloud mechanisms are covered.
- The delivery models of Cloud Computing and Cloud Services are periodically discussed with increasing complexity.
- The consumers of Cloud Computing are under a legal cover of Service Level Agreement (SLA).
- The rights and liabilities of Cloud consumers and Cloud providers are also covered.
- A significant volume of the course covers the security issues related to computers and Cloud computing.
- We shall also discuss the disaster recovery and decision making for Cloud computing.
- Cloud Computing setup and hosting is a complex task.
- Cloud management and Cloud hosting data center architectures are also covered.
- Cloud Computing is also provisioned through mobile phones.
- As special contribution:
    - The course covers the Mobile Cloud Computing field.
    - The emerging field of Software Defined Networking (SDN) is introduced.
- **Recommended books and publications:**

  - "Cloud Computing: Concepts, Technology & Architecture" Book by Ricardo Puttini, Thomas Erl, and Zaigham Mahmood; Prentice Hall/PearsonPTR
  - "Cloud Computing" Book by Kris Jamsa; Jones & Bartlett Publishers
  - Liu, Fang, et al. "NIST cloud computing reference architecture." NIST special publication 500.2011 (2011): 292.
- Please follow these important points:
  - Cover the course according to modules.
  - If a key term is confusing to you, use the recommended course book/s and/or Google.
  - Usually, the detail of the key term is covered in the same module or coming modules, so be patient.
  - Background knowledge is provided to refresh the previous knowledge only. Not for knowledge building.
  - Students are encouraged to study/consult the books related to Networks and Computer Architecture, Computer Security, Network Security etc. as well.
- **Instructor's email:** mtayyabch@yahoo.com

Lesson No. 02

INTRODUCTION TO CLOUD COMPUTING

Module No – 002:

"***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***."

National Institute of Science and Technology (NIST) USA

- Essential Characteristics according to NIST definition:
  - *On-demand self-service*
  - *Broad network access*
  - *Resource pooling*
  - *Rapid elasticity*
  - *Measured service*
- 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 client devices (e.g., mobile phones, tablets, laptops, and workstations).

- Resource pooling: The provider's computing resources are pooled to serve multiple consumers according to consumer demand. The customer generally has no control or knowledge over the exact location of the provided resources (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 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. Typically this is done on a pay-per-use or charge-per-use basis.

Module No – 007:

- Cloud Service Models according to NIST definition:
  - *Software as a Service (SaaS)*
  - *Platform as a Service (PaaS)*
  - *Infrastructure as a Service (IaaS)*
- Software as a Service (SaaS):
  - The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. A cloud infrastructure is the collection of hardware and software that enables the five essential characteristics of cloud computing.
  - The applications are accessible from various client devices a web browser (e.g., web-based email), or a program interface.
  - The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, except for limited user specific application configuration settings.
  -
- Platform as a Service (PaaS):
  - The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider.
- Infrastructure as a Service (IaaS):
  - The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications.
  - The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of selected networking components (e.g., host firewalls).
- Cloud Deployment Models according to NIST definition:
  - *Private cloud*
  - *Community cloud*
  - *Public cloud*
  - *Hybrid cloud*
- Private cloud: The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned,

managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

- Community cloud: The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.
- Public cloud: The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.
- Hybrid cloud: The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public).

Lesson No. 03

HISTORY AND BACKGROUND OF CLOUD COMPUTING

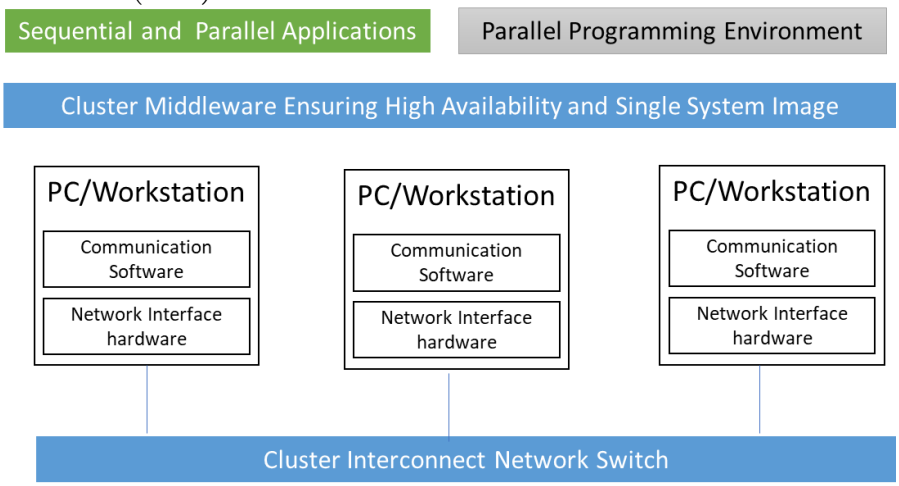
Module No – 003:

- Computer Scientist John McCarthy is attributed with delivering the idea that computations will be provisioned as utilities in future. This idea was presented in 1961.
- In 1960s and 1970s, the *mainframes* (giant powerful computers) were leased out by the manufacturers.
- The idea of *grid computing emerged* in 1990s to use the processing power of networked PCs for scientific calculations during idle times.
- In 1990s, *Salesforce.com* started bringing remotely provisioned software services to the enterprises. *Amazon Web Services (AWS)* were launched in 2002.
- In 2006, the term "*cloud computing*" emerged that enabled organizations to "lease" the computing capacity and processing power from cloud providers.

Module No – 004: Overview of Cluster Computing:

- A computer cluster is a collection of interconnected stand-alone computers which cooperate to work as a single resource pool of computing resources.
- Clusters became popular in 1990s when mainframes and traditional supercomputers were becoming less cost-effective for high performance computing (HPC).
- In 2010, out of top 500 supercomputers. 85% were computer clusters built with homogeneous nodes.
- Cluster computing has laid the foundation of modern day super computers, computational grids and cloud computing.
- Important Benefits of Cluster Computing:
    - Scalability
    - High availability and fault tolerance
    - Use of commodity computers

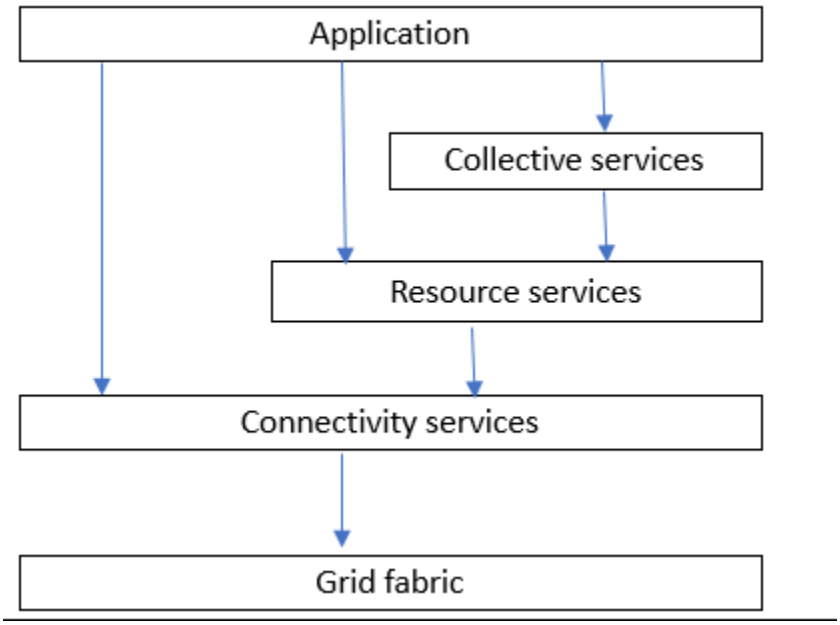
- Cluster Architecture (basic):



Cluster Architecture

Module No – 005: Overview of Grid Computing:

- The grid is an integrated computing infrastructure for bringing together *computers* to create a large collection of compute, storage, and network resources.
- Grid is used to solve large-scale computation problems or to enable fast information retrieval by registered users or user groups.
- *Computers* include PCs, workstations, server clusters, supercomputers, laptops, notebooks, mobile computers, PDAs, etc.
- Building virtual grid through *CPU scavenging*: Creating a grid by using unutilized CPU cycles in a network of computers at night or periods of inactivity.
    - This is done on voluntary basis. The grid hosts donate some RAM, disk space and network bandwidth as well.
    - The most famous example is the SETI@Home which applied over 3 million computers to achieve 23.37 TFlpos as of Sept. 2001.
- Grid Middle ware Layered Architecture (deployed on participant computers):



Grid Computing Overview

- Application: The top layer consisting of user applications to be run on grid.
- Collective Services: Focus on interaction among the resources. implements functions such as resource discovery, scheduling, brokering etc.
- Resource service: Deals with the aggregated computing resources (software and hardware) available for user applications in collective operations.
- Connectivity Layer: Provides the core networking among the computational resources of fabric layer through physical or virtual networking.
- Grid fabric: Consists of all the computational resources such as storage systems, catalogs, network resources, servers and their network connections.

Module No – 006: Difference between Cluster, Grid and Cloud Computing:

- The purpose of *Grid Computing* is to solve large scale computational problems.
- Just like *Clusters*, except that
    - The Grids make use of computational resources are spread across the nation or the globe.
    - These computational resources are owned by different organizations and are shared (as grid resources) by multiple users.
    - Grids heavily depend upon WAN/LAN resources.
    - Virtual Supercomputer term is derived from Grid Computing whereby multiple computers collaborate over network to create an illusion of a single big computer.
- As compared to cloud:
    - The resources do not join or leave the grid dynamically.
    - Majority of the resources are not provisioned from data centers.

- Several organizations may unite to form a grid in the shape of a virtual organization (VO). For example multiple hospitals and research centers may collaborate in a VO to find a cure for cancer.

Module No – 042: Business Drivers for Cloud Computing:

- Various business drivers lure the organizations to start using Cloud.
- These include (but not limited to):
  - IT Capacity Planning
  - Cost Reduction
  - Organizational Agility
  - IT Capacity Planning:
    - It is the estimation and fulfillment of future IT requirements of an organization.
    - The over provisioning of IT happens when acquired equipment is more than the estimated requirements. Resulting in over expenditure.
    - The under provisioning occurs when the equipment turns out to be inadequate to fulfill the IT requirements of the future.
    - IT Capacity planning is a difficult job as it should cover the fluctuating load.
  - Usually the companies adopt any of the following strategies:
    - Lead Strategy: Adding new IT capacity in anticipation of future needs.
    - Lag Strategy: Adding new IT capacity when the IT resources reach the full utilization
    - Match Strategy: Adding IT capacity in small increments.
  - The capacity planning may lead to adopting the option of Cloud Computing and then planning for future needs of Cloud resources rental instead of purchasing the IT equipment.
  - Cost Reduction: The costs include
    - Cost of acquiring the IT infrastructure
    - Operational overheads such as technical personnel salaries, upgrades, utility bills, security, accounts and administrative staff salaries
    - Why not choose the Cloud instead ?
  - Organizational Agility: It is the responsiveness to the change. We consider the *change* in IT for this topic.
- A possible shift, upgrade or acquiring a new software may require to upgrade the hardware.
  - The routine procedures and the business may come to halt or the competitors may out run if the organization fails to invest in IT just because of lack of affordability.
- The Cloud on the other hand, just charges for the usage of IT resources, no need to invest in infrastructure.

Lesson No. 04

BASICS OF COMPUTERS

Module No – 008:

- Mainframe:
  - A mainframe is a large, expensive, powerful server that can handle hundreds or thousands of connected users/servers simultaneously. For example a single mainframe server of IBM's Z series can provide the equivalent computing throughput of at least 500 servers.
  - In 1960s and 1970s, the mainframes were leased out by the manufacturers rather than sold because of enormous cost of ownership.
- Mainframe leasing model:
  - The customers were charged on monthly basis for the use of hardware such as CPU, memory and peripheral devices.
  - The software (compilers, editors etc.) usage was charged for the time of usage.
  - The mainframe leasers used to develop customized software exclusively for a client organization and charged for it.
  - The client was also charged for the maintenance of those customized software.
  - This model still exists in the form of cloud computing.
- Server:
  - A server is a computer which provides services to other computers and/or devices connected to it. Services provided by a server include the controlled access to hardware and software resources and storage.
  - A server can support hundreds and thousands of simultaneous users.
  - Servers are available in a variety of sizes and types.
  - Web server: stores websites and web apps and provides them on your desktops and mobiles through web browsers.
  - Domain Name Server (DNS): Stores domain names and the corresponding IP addresses.
  - Database server: Hosts database and provides access to data and provides data manipulation functionality.
- Desktop:
  - A desktop is a computer which is designed to remain in a stationary position. It is used as a personal computer.
  - Intended to be used by one person at a time.
  - Performs the activities such as
    - Input
    - Processing
    - Output
    - Storage

.

Lesson No. 05

BASICS OF DATA COMMUNICATIONS

Module No – 009:

- Data Communication: Exchange of data over some transmission medium between two devices.
- The following factors are essential for data communication:
  - Data must be delivered to correct destination.
  - There must be timely delivery of the data.
  - There must not be uneven delay among the packet arrival time during audio or video transmission.
- Components:
  - Message: The data to be sent. Can be text, numbers, pictures, audio and video.
  - Sender
  - Receiver
  - Transmission medium: The physical path through which a message travels from sender to receiver.
  - Protocol: The set of agreed-upon communication-rules between sender and receiver devices. Two devices can be connected but not communicating without a protocol.
- Data Representation:
  - Text: Represented by bit pattern called code e.g.; Unicode and American Standard Code for Information Interchange (ASCII).
  - Numbers: Directly converted binary of the number. ASCII is not used to represent numbers.
  - Images: Sent as binary patterns. Image is represented by a matrix of pixels. *Pixel* is a small dot. Each pixel is assigned a bit pattern on the basis of color.
  - Audio: A continuous stream of data. Different from text, numbers and images.
  - Video: Can be a continuous stream or a sequence of image combinations.

Module No – 010: Data Flow:

- Data Flow:
  - Simplex: Unidirectional communication in which either one of the sender or receiver device can transmit. For example: key board, monitor etc.
  - Half Duplex: Both devices can communicate but one at a time. The entire capacity of the transmission medium is available to the transmitting device. For example: walkie-talkies.
  - Full Duplex: Both devices can send and receive at the same time. The transmission medium should provide separate paths (channels) for the transmission of each device. For example telephone conversation is full duplex.

Lesson No. 06

BASICS OF COMPUTER NETWORKING

Module No – 011:

- Computer networking was conceived in 1960s soon after the invention of computers.
- A network is a collection of computers and devices connected together through transmission media.
- Devices:
    - Hosts: Large computers, desktops, laptops, cellular phone or security system.
    - Connecting devices:
        - Router: A device which connects the network with other networks.
        - Switch: A device which connects devices within the network.
        - Modem: A device which changes the form of data (modulates-demodulates).
        -
- Network Criteria:
    - Performance: It is often evaluated by two metrics:
        - Throughput (bulk of data transmitted in unit of time) and delay.
        - Performance: It is often evaluated by two metrics:
            - Increasing the throughput may increase the congestion and hence increase the *network delay*.
            - The *transit time* (message travel time) and response time (time between inquiry and response) indicate the network performance also.
        - Reliability: It is measured in terms of frequency of network failure, time to recover from a failure and robustness from disasters.
        - Security: Protecting data from unauthorized access and damage, and implementation of security policies and procedures for recovery from breaches and data losses.
- Physical Structures:
    - Network Connections: Communication can only take place if the devices are simultaneously connected to the same communication-path or link or connection.
    - Link: A link can be dedicated link (Point to Point) or shared among devices (multipoint).

Module No – 012: Network Topologies:

- Mesh: Every device has a dedicated point to point link to every other device.
    - Advantage: Robustness of network from failure of any link.
    - Disadvantage: The bulk of cabling involved.
- Star: All devices are connected to a central device. Unlike mesh, there is no direct traffic between any two devices but through the central device such as hub.

  - Advantage: Requires only one I/O port in each device as compared to mesh.
  - Disadvantage: If the central device fails, the whole network fails.
  -

- Bus: A multipoint topology in which one long cable is used a network backbone.
  - Advantage: Ease of installation. Requires less cabling than mesh and star.
  - Disadvantage: Difficult to extend, signal drops along the length of cable results in limited number of connections, breaking of backbone cable isolates the network segments and introduces noise.
- Ring: The devices are connected in the form of ring. Each device acts as repeater.
  - Advantages: Easy to expand and alter the network.
  - Disadvantage: Failure of a single device can disable the entire network, transmitting device needs to retain the token signal to perform transmission which slows down the data rate.

Module No – 013: Network Types:

- Local Area Network (LAN): It is a privately owned network and has a scope of an office, building or a campus. A LAN can even extend throughout a company.
  - Each host in a LAN has a unique identifier or address.
  - The communication packets between any two hosts in a LAN contain the source and destination hosts' addresses.
  - Key features:
    - Media type: wired/wireless, twisted pair/cable/fiber, radio, infrared
    - Topologies: Bus, Star, Mesh, Ring, Tree
    - Bit rate: from 1Mbps to 1Gbps
    - Unicast, Broadcast, Multicast
  - Typical LANs:
    - Ethernet (CSMA/CD): Carrier Sense with Multiple Access with Collision Detection (retransmission after collision detection)
    - Local Talk (CSMA/CA): CSMA with Collision Avoidance (reserve the media before transmission)
    - Wireless LAN: IEEE 802.11, Range: < 100 m, Speed: 2Mbps
    - Token Ring: A token travels around the ring, it must be retained by the sender computer to send a single packet, 4,6 or 100 Mbps
    - FDDI: Token ring with fiber optic cable, 100 Mbps
    - ATM: Star based, uses switch, multiple devices can communicate simultaneously, 25, 45, 155, 600+ Mbps
- Wide Area Network (WAN): A network that spans large geographical area such as town, cities, states or even countries. Usually interconnects multiple LANs.
  - Unlike LAN which is owned by the user organization, a WAN is normally created and run by communication companies. It is leased to the user organizations.
  - Types: Point to Point (P2P), Switched

        - P2P WAN: Connecting two devices through wired or wireless media eg; connecting two LANs to form a private internet or internetwork of a company.
        - Switched WAN: A network with more than two ends. It is a combination of several P2P WANs connected by switches.
- Metropolitan Area Network (MAN): It is a computer network covering a large geographical area bigger than LAN and smaller than WAN.
    - Diameter: 5 to 50 km, several buildings or a whole city
    - MAN is not owned by a single organization generally just like WAN. The MAN equipment are usually owned by a service provider.
    - MAN usually provides high speed connectivity to allow sharing regional resources.

Module No – 014: Switching:

- A WAN is a switched network in which a switch connects two links together to forward data from one network to the other.
- Two common types of switched networks are:
    - Circuit-Switched Network: A dedicated physical connection (circuit) is established, maintained and terminated through a carrier network for each communication session.
        - Used extensively by telephone companies. Only useful when all the circuits are being utilized simultaneously; otherwise the network is being underutilized.
    - Packet-Switched Network: It is a WAN switching method in which a single link is shared among multiple network devices.
        - Statistical multiplexing is used to enable devices to share the packet-switching circuits.

Module No – 015: The Internet History and Accessing the Internet:

- Internet: It is a network of interconnected networks which include:
    - *Backbones*: Large networks owned by communication companies such as PTCL, AT&T etc.
    - *Provider Networks*: Use the service of backbone for a fee. Connected to backbone through *peering points*. Sometimes connected to other provider networks as well.
    - *Customer Networks*: Use the services such as internet connectivity provided by provider networks and pay a fee to provider for that.
    - The Backbones and provider networks are also called *Internet Service Providers (ISPs)*.
    - Accessing the Internet:
        - Telephone Networks: Dial-up service, DSL Service
        - Cable Networks
        - Wireless Networks
    - Internet today: World Wide Web, Multimedia, Peer-to-Peer Applications

Module No – 016: TCP/IP Suite:

- TCP/IP Protocol Stack: Transmission Control Protocol (TCP) was proposed in 1973 to ensure a reliable, end-to-end and error free transmission control.
    - It was latter split into TCP and Internet Protocol (IP) layers with IP handling the message routing and TCP performing the error control.
    - Since 1981, TCP/IP is included in the operating systems.
    - Consists of layers of protocols which paved the way for creating today's internet. These layers help in dividing a complex task into several smaller and simpler tasks:
        - Physical Layer: Deals with transmission of bits into signals and transmission of signals over the link.
        - Data-link Layer: Creates the frames of data. Each frame contains the data and is addressed with the MAC address of the receiving device and also contains the MAC address of sending device.
        - Network Layer: Is responsible for host to host communication through their IP addresses and related protocols. No control for error and congestion is performed. Packets are called datagrams.
        - Transport Layer: Responsible for transporting a message from application program running over source host to corresponding application program on destination host. Works on port numbers on corresponding hosts. Main protocols are:
            - Transmission Control Protocol (TCP): Provides flow control, congestion control and error control as it is a connection oriented protocol.
            - User Datagram Protocol (UDP): Is light weight and is not connection oriented.
            - TCP Message = segment
            - UDP Message = datagram
        - Application Layer: Consist of programs running on two hosts and exchanging messages. Applications use these protocols for communication:
            - HTTP
            - FTP
            - SMTP

Module No – 017: IP Addressing:

- The identifier used in Network layer of TCP/IP suit is the address of the internet connection of receiver and sender devices.
- IPv4 is a 32 bit universally unique address while IPv6 is the 128 bit universally unique address.
    - Total IPv4 addresses = 2^{32}
    - Total IPv6 addresses = 2^{128}
- The address is in fact of the connection and may change when the device is moved to another network.

- A device can have two IP addresses if it has two connections with the internet.
- IP address is usually represented by dotted decimal numbers. For example: IP v4 address: 193.63.82.10
- The IP addresses are allocated by the Internet Corporation for Assigned Names and Numbers (ICANN) to ISPs and large organizations.
- Smaller organizations can get IP addresses from ISPs.
- The IP address consists of a prefix part (the Network ID) and postfix part (the Host ID or the Subnet).
- Classification of IPv4 addresses:
    - Class A: 8 bits for Network ID
        - Total networks 2^7
        - Network id starts with '0' binary
        - First byte: 0 to 127
    - Class B: 16 bits for Network ID
        - Total networks 2^{14}
        - Network id starts with '10' binary
        - First byte: 128 to 191
    - Class C: 24 bits for Network ID
        - Total networks 2^{21}
        - Network id starts with '110' binary
        - First byte: 192 to 223
    - Class D: Used for multicasting
        - No prefix or Network ID
        - First byte: 224 to 239
    - Class E: Reserved for future use
        - First byte: 240 to 255
- Address Masking:
    - Classful addressing lead to depletion of IP addresses and/or unused addresses.
    - Solution:
        - Classless addresses with variable sized prefix according to the needs of organizations
        - A notation representing the length of prefix is added at the end of a classless address with a slash '/' to indicate the addresses in a classless address block.

Module No – 018: IP Addressing:

- Dynamic Host Configuration Protocol (DHCP) is used to automatically assign IP addresses from an acquired block of IP addresses.
- Organizations use private IP addressing for the LAN devices and can use Network Address Translation (NAT) mechanism for having a single or a few registered global IP address/es for internet communication.
- NAT enabled router replaces the local address of sending device with the registered global IP address before sending the packets on internet.

- The mapping of incoming internet packets is done through NAT table which contains the source device local address port number of the program along with corresponding IP address of internet device.
- The internal network is supposed to initiate the internet communication in NAT mechanism for mapping to take place.
- Anytime a host or a router needs to find the link-layer address of another host or router in its network, it broadcasts an Address Resolution Protocol (ARP) request packet with the destination IP address and its own IP and link level address.
- The destination device replies to the sender device with its link level address.
- As compared to IPv4, the next generation IP protocol is IPv6. Some important changes are:
    - No more NAT (Network Address Translation)
    - No more private address collisions
- As compared to IPv4, the next generation IP protocol is IPv6. Some important changes are:
    - Built-in authentication and privacy support
    - Easier administration (no more DHCP required)
    - Simplified routing
- There are three categories of IPv6 addresses:
    - Unicast address: For a single connection
    - Multicast address: For a set of interfaces, one message transmitted to all.
    - Anycast address: For a group of interfaces, one message transmitted to a single interface

Module No – 019: Ethernet:

- It is a popular LAN technology for data-link and physical layers.
- Institute for Electrical and Electronic Engineers (IEEE) developed an Ethernet standard known as IEEE Standard 802.3
- TCP/IP does not specify any protocols for data-link and physical layers. It accepts all the protocols working at these layers.
- Ethernet was developed in 1970s and since then it has gone through four generations. This evolution is in fact the reason of vast implementation of Ethernet in the world.
- Data rate 10 Mbps
- Connectionless
- No flow control
- No error control
- No retransmission and acknowledgement
- Hence unreliable like IP and UDP
- Uses link-layer addresses (the 48 bit MAC address)
- CRC is present but corrupted frames are simply discarded by receiver
- Each frame is of 64-1518 bytes of length including 46-1500 bytes of data
- CSMA/CD is used
- Unicast address: Significant bit of first byte is 0
- Multicast address: Significant bit of first byte is 1

- Broadcast address: All 48 bits are 1s.
- Note: All devices on Ethernet receive all the messages but keep only those that are addressed according to above.
- Standard Ethernet types:
    - Bridged
    - Switched
    - Full duplex switches
- Fast Ethernet:
    - Next generation of standard Ethernet
    - Raised speed to 100 Mbps
    - Downward compatible with standard Ethernet (speed is reduced for compatibility)
    - Same 48 bit addressing
    - Frame format is same as of standard Ethernet
    - Uses star topology for connecting three or more devices using switch or hub
- Gigabit Ethernet:
    - 1Gbps speed
    - Compatible with standard and fast Ethernet
    - Star topology using hub or switch
    - Up to 5 kilometers range
- 10 Gigabit Ethernet:
    - 10 Gbps speed
    - Compatible with standard and fast Ethernet
    - Increases the range to tens of kilometers
    - Possibility to interconnect LANs

Module No – 020: Wired LAN vs. Wireless LAN:

- Wired LAN:
    - Medium: Wires
    - Broadcasting and multicasting possible when required
    - Physical connection to network
    - Hosts are connected through link layer switch
    - Connection to other networks through router
- Wireless LAN
    - Medium: Air
    - All devices are broadcasting
    - No physical connection to network
    - No link layer switch exists
    - Connected to other networks through access point (a device that connects a wireless and wired network)
- **IEEE 802.11**
    - It is a wireless LAN standard by IEEE that covers physical and data-link layers
    - Synonyms: WiFi, Wireless LAN
    - Basic architecture consists of an access point (AP) and capable devices connected to Access Point (AP
    - In the absence of AP, the wireless devices connect to form adhoc network

- Multiple overlapping APs are used to cover a larger area
- A device is connected to only one of the nearest APs
- CSMA/CA is used. The sender sends a Request To Send (RTS) packet, the receiver sends Clear To Send (CTS) packet, the sender sends data after receiving CTS, the receiver sends acknowledgement, the other senders can send now.
- If no CTS is received, the sender marks it as a collision
- 802.11 a, b, g, n
- 802.11a: 50 feet, 22 Mbps
- 802.11b: 100 feet, 11 Mbps
- 802.11g: 100 feet, 54 Mbps
- 802.11n: 50 feet, 700 Mbps (to be implemented)

Module No – 021: Bluetooth:

- It is a wireless LAN technology that provides short distance connectivity to devices which have different functionalities for example, mobile phones, headsets, notebooks, desktops, computer peripheral devices, cameras and even the home appliances.
- Multiple devices can be connected through Bluetooth to form a piconet
- Bluetooth supports:
  - Voice and data transmission
  - Adhoc networking for up to 10 meters
- Multiple devices can be connected through Bluetooth to form a piconet
- IEEE standard 802.15 covers the Personal Area Network (PAN) using Bluetooth for an area covering a room
- Versions**:**
  - 1.x: up to 1Mbps, obsolete
  - 2.x: up to 3 Mbps, improved pairing capability between devices from different manufacturers
  - 3.x: up to 24 Mbps using WiFi 802.11
  - 4.x: Up to 24 Mbps, works seamlessly with 4G, works with data collection from sensors an internet of things (IoT)

Module No – 022: WiMAX:

- WiMax stands for Worldwide Interoperability for Microwave Access
- Provides wireless access to Internet for: Homes and offices when the wired access is either not available or is expensive (fixed WiMAX)
- Mobile phones (mobile WiMAX)
- Fixed WiMAX requires the installation of antennas at the premises of the subscriber to receive and send the data from the base station of Internet provider.
- Mobile WiMAX users move from one place to another while connected to the base station of Internet provider .

- WiMAX is the result of IEEE 802.16 project. It is a standard for wireless WAN (or MAN). The subscriber station may be tens of kilometers away from the base station of the provider.
- Remember that 802.11 is the standard for wireless LAN.
- Uses 48-bit MAC address of subscriber station and base station at Data-link layer
- Connection oriented protocol. Each connection has a unique id and hence there is no address field in the frame of WiMAX
- Full duplex communication

Module No – 023: Evolution of Cellular Networks:

- Cellular network or telephony is a radio-based technology
- Radio waves are electromagnetic waves propagated by antennas
- Note: Antenna is a transducer device which converts the altering current into radio waves and vice versa
- 7 billion mobile connections
- 25 billion interconnected devices count predicted in 2020
- Over 100 billion downloads completed in 2013, 270 billion expected in 2017
- The base stations receive from and transmit to cellular phones.
- Cellular Networks have evolved from first generation (1G) to fifth generation (5G). Let us briefly look at these generations:
    - 1G
        - Invented around 1980.
        - First implementation in Tokyo (Japan)
        - Based upon analog technology
        - Expanded to cover all the population of Japan in few years
        - Not secure
        - Anyone with an all-band radio can listen to calls and get the phone number of the subscriber
        - Analog mobiles were larger in size and heavy in weight
    - 2G
        - Invented in 1991, implemented first time in Finland
        - Technologies: Global System for Mobile (GSM) Communication, General Packet Radio Service (GPRS), Code Division Multiple Access (CDMA) [digital signal] and Enhanced Data Rates for GSM Evolution (EDGE)
        - Short Messaging Service (SMS), Multi-Media Messaging Service (MMS
        - Typical data rate: 100 Kbps
        - Email, Web browsing, Camera phones
        - Signal strength decay problem, performance degrades with the rise in number of users in a cell (area maintained by a base station)
    - 3G
        - From 2000 to 2010
        - Technologies: CDMA, WLAN, Bluetooth, Universal Mobile telecommunication Systems (UMTS), High Speed Downlink Packet Access (HSDPA)

- Features: Global Roaming Clarity in voice calls, Fast Communication, Internet, Mobile T.V, Video Conferencing, Video Calls, Multi Media Messaging Service (MMS), 3D gaming and Multiplayer-Gaming, smart phones
- Typical data rate: Up to a few Mbps
- Expensive mobile phones, battery life issue

- 4G
  - Since 2010
  - Technologies: Long Term Evolution (LTE) Standard based on the GSM/EDGE and UMTS/HSPA, Multiple In Multiple Output (MIMO) smart antenna technology, Orthogonal Frequency Digital Multiplexing (OFDM), WiMAX
  - Typical data rate: Up to a few tens of Mbps
  - MAGIC: Mobile multimedia–Anytime anywhere–Global mobile support–Integrated wireless solutions–Customized personal service
  - Maintaining data rate is an issue, not fully implemented in all the world, battery consumption is a bigger problem than 3G
- 5G
  - To be implemented
  - Technologies: New releases of LTE
  - Faster data rate than 4G (> 1Gbps), higher data rate at cell edges
  - Research is still in progress

Module No – 024: Connecting Devices:

- These are the devices used to connect:
  - Hosts to form LANs
  - LANSs to implements WANs and Internet
- The class of each device depends upon the layer/s on which it operates. That is:
  - Hub: Physical layer
  - Link-layer Switch: Physical layer, Data-link layer
  - Router: Network layer, Data-link layer, Physical layer
- Hub
  - It is a multiport repeater device used in star topology.
  - A repeater device regenerates the signal before it become too weak or corrupted.
  - The hub repeats the signal received from any port 'A' to all the other ports except the port 'A' (broadcasting)
  - This is because the hub is a physical layer device. It does not has its own MAC address and can not access the data-link layer address (MAC address) of the data frames.
- Switch
  - It is a multiport bridge device.
  - A bridge joins two logical segments of the same network and intelligently forwards the packets from one segment to other on the basis of destination MAC address and forwarding table.

  - It is a two layer device. It performs functionality on data-link layer as well as it regenerates the signals it receives.
  - A data-link layer switch works by maintaining a switching table and forwarding the packets received from a port 'A' only to the destination port 'B'.
  - Switching table consists of MAC address of the hosts arranged according to the port numbers to which they are attached. It is consulted before forwarding a packet.
  - Switches perform learning to fill the switching table by reading the MAC address of sending device for each port.
  - Broadcast packets are forwarded to all ports
  - In a situation when two LANs are connected through more than one switch then the looping problem can arise.
- Looping
  - The frame/s sent by one host 'X' in LAN1 to a host 'Y' in LAN2 will go through multiple switches and get duplicated when reaching the destination host.
  - Reason: The switches generically forward the frame received for an unregistered MAC to all the ports (except the sender's port).
- Advantages of Switch
  - Collision elimination
  - Connecting heterogeneous devices (in terms of data rate capacity)
- Router
  - It is a three layer device:
  - Physical (regenerating the signals)
  - Data-link layer(checking the MAC addresses of source and destination)
  - It is a three layer device:
    - Network layer (checks the IP addresses of source and destination, connects multiple networks to form bigger networks)
  - Has multiple interfaces. Each interface has a MAC address and IP address.
  - A router:
  - Only processes those packets which are addressed to the interface at which they arrive.
  - Has multiple interfaces. Each interface has a MAC address and IP address.
  - A router changes the source and destination MAC address when it forwards the packets.
- Virtual LAN (VLAN):
  - A logical (not physical) segment of a physical LAN.
  - VLANs are defined by software. Each VLAN is a work group in an organization, has a VLAN ID and receives the broadcast messages addressed to its own ID.
  - A VLAN may span over multiple switches in a LAN.
  - No need to update the physical topology to relocate a person from one VLAN to other, just the software configuration is to be updated.

Module No – 025: Routing:

- In a physical network, multiple LANs and WANs are joined together by the routers.
- Hence there can be more than one route between two hosts.

- Routing is a service of Network layer to find the best route.
- Routing is performed by applying routing protocols and using the decision tables called *routing tables* in each router.
- *Forwarding* is the action performed by a router on the basis of routing protocol and routing table according to the destination address of each packet received at any interface.
- At network layer, each message from higher layer is broken down into *packets*.
- A router performs packet switching.
- Types of routing:
    - *Unicast routing:* A router forwards the packet to only one of the attached networks.
    - *Multicast routing:* A packet is forwarded to multiple attached networks.
- Routing a packet from a source host to destination host can also be defined as routing a packet from a source router (the default router of the source host) to a destination router (the router connected to the destination network) through the intermediate routers using routing algorithms.
- Types of routing:
    - *Connectionless routing:* All packets of the same message are treated independently and may or may not follow the same route.
    - *Connection oriented routing:* All the packets of same message are labeled and routed through a *virtual circuit* or a fixed route.
- An internet can be considered as a graph with each network as an edge and each router as a node.
- In a weighted graph, each edge has a weight or cost.
- Least cost routing can be performed. Example algorithms: Distance-Vector routing, Link-State routing

Lesson No. 07

ADVANCED TOPICS OF COMPUTER NETWORKS

Module No – 026: Broadband Networks & Internet:

- All clouds are inherently dependent upon internetworking or Internet for ubiquitously remote provisioning of IT resources.
- The cloud providers and consumers connect to Internet through ISPs.
- The largest backbone networks of the Internet are strategically interconnected by core routers.
- The core-routers connect the international networks.
- The Internet has become a dynamic and complex aggregate of ISPs.
- There is a hierarchical topology for worldwide interconnectivity composed of tiers.
- There are three tiers of worldwide connectivity:
    - Tier 1 consists of large-scale international connectivity providers.
    - Tier 2 consists of large regional ISPs connected to tier 1.
    - Tier 3 consists of local ISP providers connected to tier 2.
- The cloud providers and users connect directly to tier 3 providers.

Module No – 027: Internet Architecture & Cloud Deployment:

- Internet supports the remote provisioning of IT resources.
- Cloud relies heavily upon Internet.
- The connectivity of end-users of cloud depends upon how the centralized resources of cloud are deployed.
- The cloud resources deployment can be either on-premises or Internet based.
- In cloud deployment using the on-premises, the provider sets up a fully controlled corporate network and a corporate Internet connection for the deployment of IT solutions and applications.
- In the on-premises deployment, the internal users access the cloud through corporate network. The remote users connect through internet by using virtual private network (VPN).
- A VPN creates a secure connection between a remote device and the corporate servers over the internet as if the device is inside the LAN.
- For the internet based deployment, the cloud provider has an Internet connection and al the internal and external users access the cloud resources through cloud provider's internet connection.
- In this deployment, there is an extra charge for internet connectivity.

Module No – 028: Scalable Computing over Internet:

- Scalable computing may refer to the dynamic resizing of the available computing resources (processing, memory, bandwidth, storage etc.) with demand.
- The growth of users and user demands for scalable computing over internet has been accompanied with matching growth in network, computing and resource management technologies.
- The computing platforms have evolved as follows
    - Mainframes (1950-70)
    - Minicomputers (1960-1980)
    - Personal computers (1970-1990)
    - Portable computers (1980-2000)
- Since 1990, the High Performance Computing (HPC) and High Throughput Computing (HTC) have been relying upon clusters, grids and the Internet clouds.
- The speed for HPC systems (supercomputers) has increased from Gflops in early 1990s to now Pflops in 2010.
- The network bandwidth has been doubling each year in the recent past (Gilder's law).
- Processor speed has been doubling every 18 months (Moore's law).
- Means that there has been a steady growth in these technologies.
- Fine grain (instruction level) parallelism and coarse grain (job level) parallelism are available.

- Ubiquitous computing is what refers to computing at any place and time using pervasive devices and wired or wireless communications.
- Utility computing works upon a business model in which the customers pay for computational resources from a provider.
- Cloud computing provides ubiquitous utility computing.

Module No – 029: Technologies for Network based Systems:

- The processor speed and network bandwidth have shown a remarkable growth in last few decades.
- The processor clock rate has risen from 10 MHz in 1970s to over 4GHz in 2010s.
- The network band has increased from 10 Mbps to over 100,000 Mbps
- The excessive heat generation from single processor core with high frequency has limited the maximum speed unless the chip technology matured.
- This has lead to the multi-core architecture of CPUs with dual, quad, six or more cores.
- The graphical processing unit (GPU) development has adopted a many-core architecture with hundreds to thousands of cores.
- Modern architecture of CPUs and GPUs have enhanced the instruction level parallelism (ILP) and the volume of millions of instructions per second (MIPS).
- Sun's Niagara CPU can provide 64 count for ILP.
- Intel's Core i7 990x can provide 159,000 MIPS execution rate
- The CPUs and GPUs are multithreaded, which means that each core can execute multiple processes or threads concurrently.
- A GPU unit has far more (but slower) cores than a multi-core CPU
- The DRAM memory chip capacity has increased from 16 KB in 1976 to 64 GB in 2011.
- The hard disk capacity has increased from 260 MB in 1981 to 3TB a few years ago.
- The flash memory and solid-state drives are rapidly evolving.
- Disk arrays are being utilized to enhance the storage.
- Servers can be connected to network storage such as disk arrays through storage area network (SAN)
- A disk array can be connected to client hosts through network attached storage (NAS)
- The high bandwidth networks in WAN scope can connect the host computers to network storage.
- A single host can be shared among multiple instances of operating systems through v*irtualization technology*. More on this latter.

Module No – 030: Web 2.0

- It is the second generation of world wide web.
- Lets people collaborate and share comments, media and information online.
- The web pages progressed from static to dynamic and interactive.

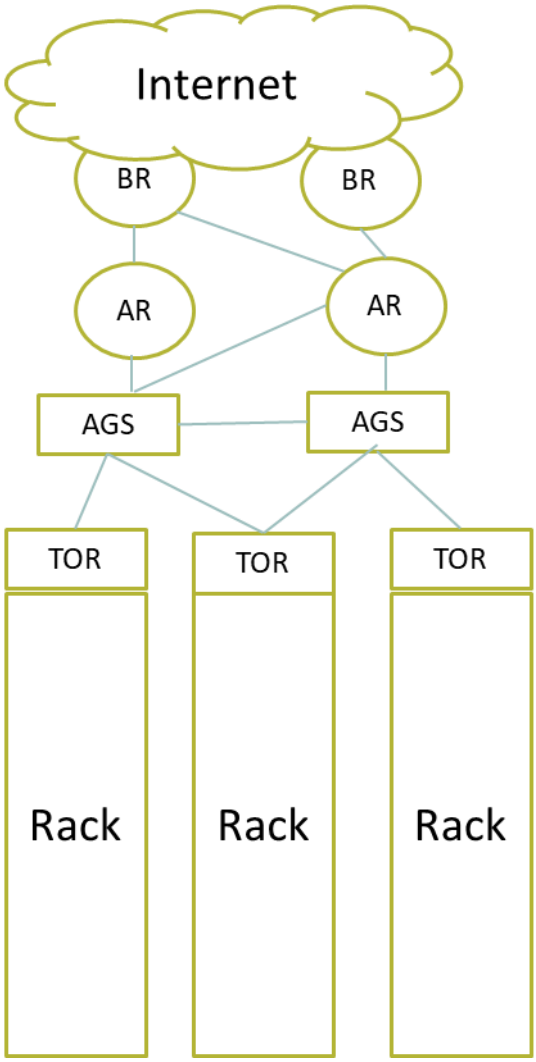
- Through Asynchronous Javascript and XML or Ajax, the web applications can send and receive data from a web server without interfering with the display and behavior of the existing page.
- Social networking and community oriented sites have emerged
- For example myspace.com, facebook.com, twitter.com etc.
- Users can contribute in web based blogs, wikis, online training, online education etc.
- Really Simple Syndication (RSS) feeds continuously keep the subscribers informed about news, follow up updates and products.
- Users can do online chatting and share files through messenger tools such as Yahoo messenger, Skype, WhatsApp etc.

Module No – 038: Virtual Private Network (VPN):

- A VPN extends a private network over public network and enables the users to communicate as if their devices are directly connected to the private network.
- A VPN creates a secured and encrypted network over a less secured network such as the Internet.
- Normally a VPN is provided and managed by a service provider.
- VPN allows the corporate employees to securely access the applications hosted over enterprise LAN.
- VPN is based upon IP tunneling.
- IP tunneling, or port forwarding is the transmission of private network packets over a public network (Internet) as the payload of public network packets such that the routing devices do not come to know about this.
- There are many protocols for VPN establishment and encryption: IP Security (IPSec), Secure Socket Layer(SSL), Point-To-Point Tunneling Protocol (PPTP), Multiprotocol Label Switching (MPLS) etc.
- VPN although provide secured connectivity to extend a private network but the implementation may have performance issues.
- VPN is implementable over Layer 1-3.
- Types of VPN:
    - Remote-access VPN: A VPN client on user's device connected to VPN gateway of the enterprise.
    - Site-to-site VPN: Establishes a VPN between two networks over the Internet by using VPN gateway.
- VPN technology provides access to cloud resources. The VPN gateway exists in the cloud with a secure link provided by the cloud provider.

Module No – 040: Networking Structure of Cloud Hosting Data center:

- At the core of a cloud is a cluster of VMs/ physical servers.
- There can be tens of thousands of physical servers in a data center.
- With each physical server hosting multiple VMs, the cloud hosting networking structure becomes a little complicated.
- The cluster nodes are used for computations. Some nodes are used for workload allocation, some for monitoring and some for load balancing.
- Some node called *gateway nodes* provide the interface of cloud service/s to the outside world (through internet).
- Cloud hosting data center has a layered architecture for the Internet access.
- The servers are physically connected to layer 2 switches. There is a top of rack (TOR) in each rack. One server is connected to only one TOR switch.
- The TORs are connected to aggregate switches (AGS).
- The AGSs provide the cross rack inter VM connectivity.
- There are a few access routers (AR) and border routers (BR) at layer 3.
- The layer 2 AGSs are connected to BR through AR.
- The BRs are connected to Internet.
- Some problems solved by the hypervisor solutions:
    - The VMs hosted on one server may belong to different vLANs.
    - A single vLAN may span over multiple data centers.
    - A company may own multiple data centers and may want to migrate the VMs across data centers.



Lesson No. 08

VIRTUALIZATION

Module No – 031:

- Virtualization is a technology used to enhance the utilization of computing resources.
- A single hardware machine is multiplexed among multiple virtual machines (VMs).
- A software based virtual machine monitor/manager (VMM) or hypervisor is a program that manages the hardware resources for the VMs and also keeps each VM from disrupting other VMs.
- Virtualization implementation levels:
    - Instruction Set Architecture (ISA) level: Executing legacy code over new machines using ISA emulator tool such as an *interpreter* which translate one instruction of source code into corresponding instruction of the target machine.

- Hardware Abstraction level: The hardware components (CPU, RAM, Disk, NIC) of a physical system are virtualized and shared among virtual machines using Virtual Machine Monitor (VMM) tool or *hypervisor* which performs as abstraction layer.
- Operating System Level: The OS running over a server accommodates multiple *containers* or VMs. The host operating system acts as the abstraction layer between hardware and the containers.
- Library support level: The API calls for hardware acceleration such as vCUDA stubs for graphic processing units (GPUs) are available at VM level.
- Application level: An application acts as a VM through wrapping of application in an abstraction layer which isolates it from OS and other applications. Another type is using virtualization layer as programming environment e.g; Java Virtual Machine (JVM).

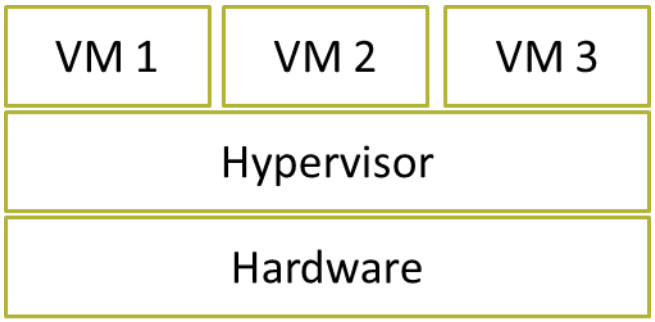
Module No – 032: Virtualization Structures:

- We know that the *virtualization layer* transforms the physical hardware into virtual hardware. There are three classes of VM architectures.
  - Hypervisor Architecture:
    - It is the hardware level virtualization. Also called the bare-metal virtualization
    - The hypervisor sits between the hardware and the VMs and manages the VMs.
    - Example: Xen, VMware
  - Full-virtualization Architecture:
    - The guest operating system (OS) or the VM's OS does not know that it is installed on a VM.
    - The Virtualization layer manages the hardware acceleration. For example VMware
    - The virtualization layer can be installed on hardware or on host's OS.
    - Some of the instructions of a gust VM are directly run on hardware to enhance the performance.
  - Para-virtualization Architecture:
    - The guest OS is modified to comply with virtualization layer. All calls for hardware acceleration are handled by virtualization layer.
    - For example: KVM

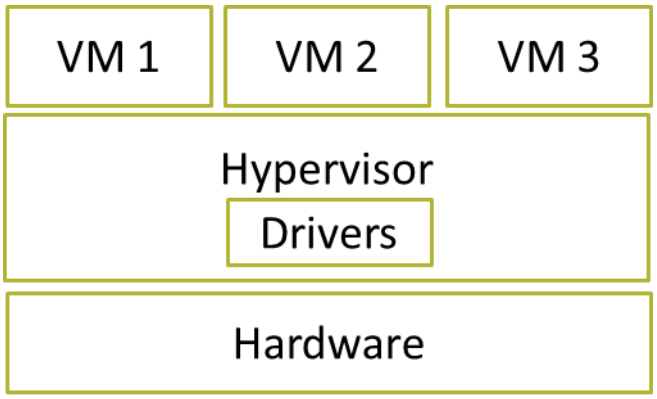
Module No – 033: Virtualization Architectures:

- Hypervisor transforms the physical hardware into virtual hardware.
- Virtualization Architectures:

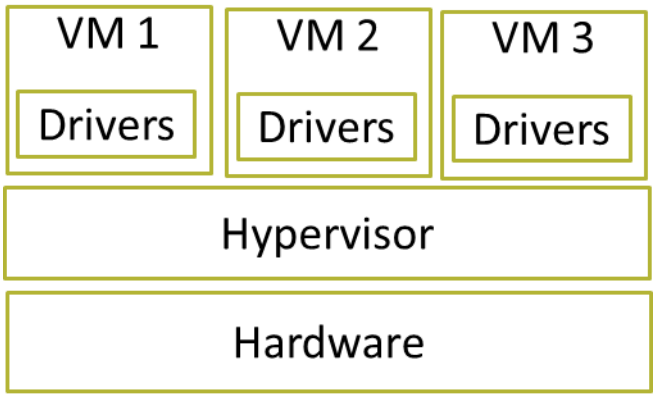
- Generic:



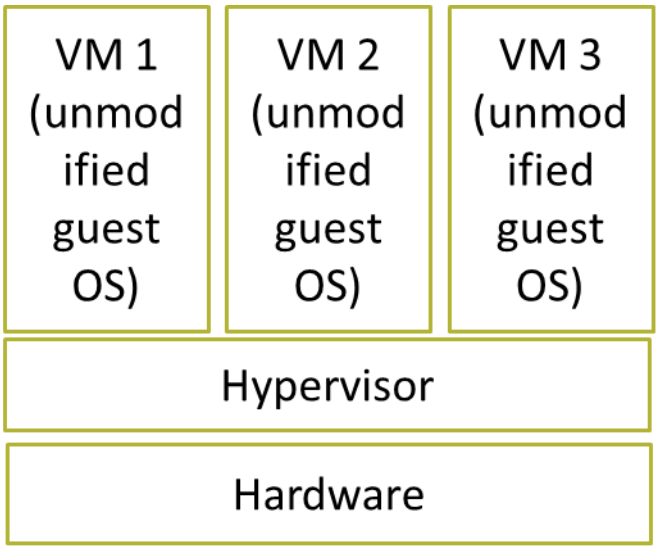
- Monolithic:



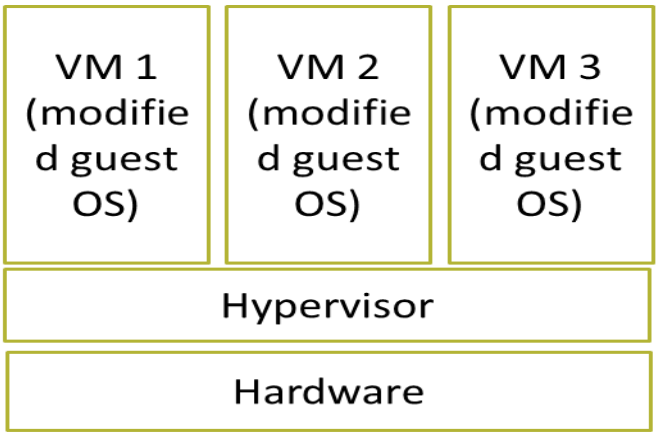
- Microkernel:



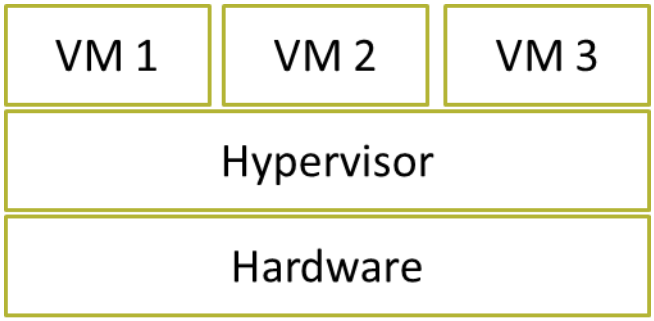
- Full Virtualization:



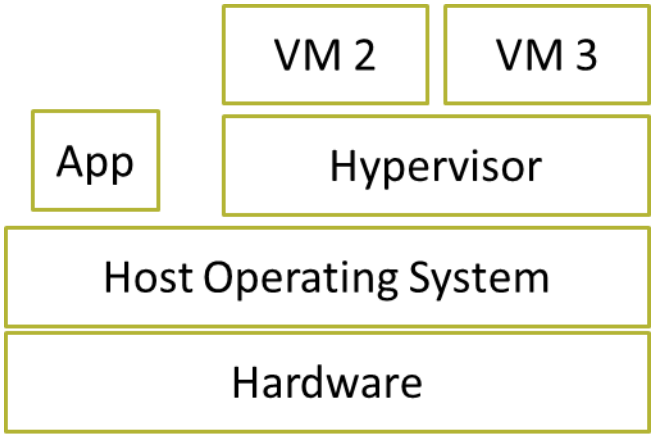
- Para Virtualization:



- Native Virtualization Architecture:



- Hosted Virtualization Architecture:



Module No – 034: Virtualization of CPU, Memory and I/O Devices:

- To support virtualization, processors such as x86 architecture use a special mode and instructions known as *hardware-assisted virtualization.*
- In this way, the hypervisor is able to trap the sensitive instructions of the guest OS and its applications.
- The modern processors allow multiple processes to run simultaneously. Any process can execute a critical instruction to crash the whole system.
- Therefore, the critical instructions are executed in *privileged* or *supervisor* mode of the processor. The OS controls this mode on behalf of the processes being executed.
- The second type of instructions are non-privileged or non-critical instructions which are run in *user-mode* of the processor.
- CPU Virtualization: A CPU is virtualizable if it is able to run the privileged and un-privileged instructions of a VM in user mode and the hypervisor running in supervisor mode.
- Memory Virtualization: Traditionally, the OS performs the mapping of virtual memory to machine memory by using page tables.
- The modern x86 CPUs include the *memory management unit (MMU)* and *translation lookaside buffer (TLB)* to optimize virtual memory performance.
- However, in virtualization environment, the memory virtualization involves the sharing and dynamic allocation of physical memory of the system to the physical memory of the VMs.
- The guest OS performs the virtual to physical memory mapping of the VM, while the hypervisor performs mapping of physical memory to machine memory.
- I/O Virtualization:
  - It is done in either of the three ways:

- Full device emulation: The device is emulated in software located in hypervisor. The hypervisor interacts with the real device. The VM interacts with the virtual device.
- Para-virtualization based I/O: The guest OS interacts with the device through its frontend driver. The frontend driver interacts with a backend driver of the device. The backend driver interacts with the device.
- Direct I/O virtualization: This type of I/O virtualization allows the VMs to directly access the device.

Module No – 035: Virtual Clusters:

- A virtual cluster consists of several VMs hosted on a physical cluster.
- The VMs are interconnected through a virtual network across multiple physical networks.
- The nodes can be physical or virtual machines and can grow or shrink dynamically.
- The failure of a host can make the hosted VMs unavailable but the virtual cluster does not collapse.
- The failure of a VM does not fails the host.
- A physical cluster may host multiple virtual clusters.
- A virtual cluster may span over multiple physical cluster.
- In order to deploy a virtual cluster, several VMs with installed OS and application software are required.
- The deployment time is to be as quick as possible.
- Templates can be used to deploy the VMs from.
- A template is a disk image with preinstalled OS with or without certain applications.
- A suitable template can be copied as disk image of a VM. This saves time of installing and configuring.
- When the VM is ready and up, it is deployed to a suitable host.
- The VM then joins a virtual cluster.
- All of the above can be done manually as well as full or partially automated.
- Reasons of virtualization:
    - Sharing of resources
    - Isolation of users of shared resource
    - Aggregation of smaller resources into a single big virtual resource (e.g., Storage)
    - Dynamic relocation/provisioning of virtual resources is easier than physical resources
    - Easier management of virtual resources/devices/machines.

Module No – 036: Virtual Machine (VM) Migration:

- VMs can be migrated from one host to another for:
    - Server Load balancing
    - Server consolidation

  - Remedy for failover hosts and VMs
  - Remedy for resource shortage for a VM
- A VM can be in any of the following states:
  - Powered-off
  - Suspended
  - Paused
  - Powered-on
- The following options are available for VM migration:
  - Cold migration: The VM has to be powered-off before migration.
  - Warm migration: Suspended VM migration.
  - Live migration: For powered-on VM with zero downtime and full availability.
- A VM is made of two basic components:
  - VM state: The processor and RAM contents
  - Virtual hard disk: Residing on network storage or on host's hard disk
- Live migration of VM means zero downtime of OS, connectivity and applications running on the VM.
- For live migration, the VM state is to be copied from source to destination host. The virtual disk can also be migrated through live storage migration feature of the hypervisor.
- Modern day hypervisors provide unbroken network connectivity of the VM during live migration.
- During the live migration, the state and storage of the VM keeps on working on source host to avoid down-time.
- For live migration of a VM with the virtual hard disk on network accessed shared storage, the virtual hard disk need not to be copied if the destination host can access that network based storage.
- Migrating the virtual hard disk is time consuming as well as network bandwidth consuming due to multi-Gigabyte migration.
- A better solution is to use the network storage.

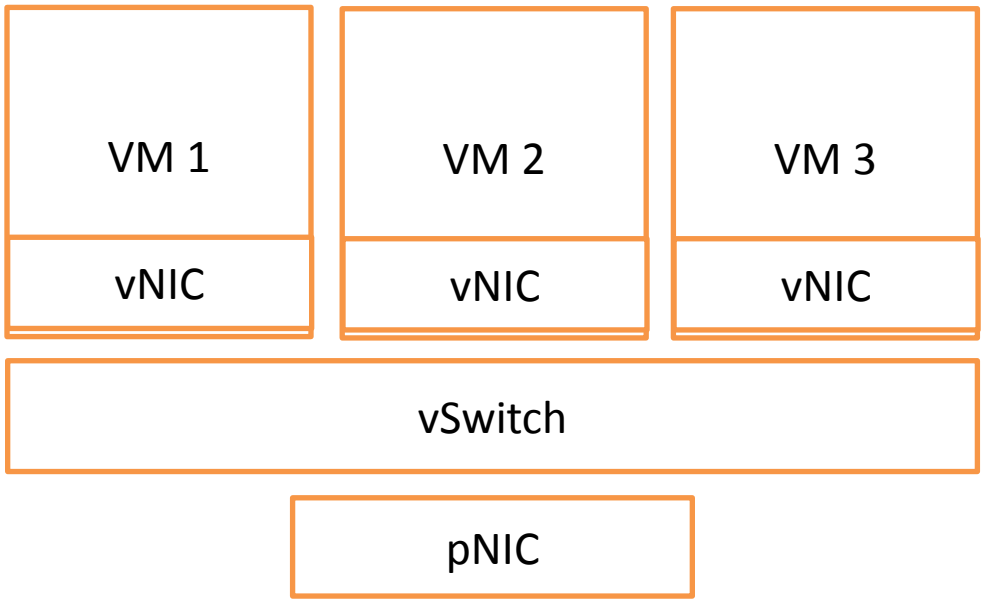
Module No – 037: Virtualization for Data Center Automation:

- A data center is a facility with networked computers and is used by businesses and other organizations to process, store and share large amounts of data.
- Companies like Google, Yahoo, Amazon, Microsoft, IBM, HP, Apple etc. have invested billions of dollars for constructing the data centers.
- Data center automation refers to the dynamic provisioning of hardware and software resources to millions of users simultaneously.
- Data centers can host Clouds.
- Data center automation is triggered by the growth of virtualization products.
- The data center owner has three major considerations:
  - Assuring Performance and QoS
  - Increase resource utilization
  - Saving costs

- Enhanced resource allocation (to jobs and/or VMs) may be performed in data centers to assure performance and QoS.
- The over allocation of computing resources may result in decrease in average utilization of these resources.
- This also leads to increased costs due to power consumption.
- Example: A VM hosted on a server with 1.5 GHz *4 cores and 16 GB of RAM is allocated 1.5GHz * 2 vCPUS, 4 GB vRAM (half of the processing and 1/4th RAM).
- Suppose if there are two such VMs. But the overall average workload of the hosted VMs keeps the physical utilization to less than 50%. This is a resource wastage as 50% of the resources remain idle.
- *Server consolidation* is a technique by which more VMs are aggregated on a single server (by migrating jobs/VMs to it) while assuring performance and QoS.
    - This increases the resource utilization across data center.
    - More servers are available to take more workload. . Otherwise, the idle servers can be shut down to save power.
- Virtualization technology also helps in setting of virtual storage (over VMs) to offer virtual disks to other VMs.
- Virtualization can synchronize with cloud management systems to dynamically provision cloud services and billing systems.
- Hence, virtualization is essential for Cloud computing.

Module No – 039: Network Virtualization:

- Multiple virtual Network Interface Cards (vNIC) are linked to physical NIC or pNIC through a virtual Switch (vSwitch) inside a hypervisor.



Network Virtualization

- A *virtual network* consists of virtual nodes and virtual links.
- Network virtualization establishes the coexistence of multiple virtual networks.
- Network virtualization proposes the decoupling of traditional ISP functionalities such as *infrastructure setup and management* from the *creation and management of virtual networks*.
- It is possible to use physical infrastructures of multiple providers to dynamically compose virtual network/s.
- Network virtualization technologies:
    - Virtual Local Area Network (VLAN): Logically grouping the hosts with common interest into a single broadcasting domain.
    - Virtual Private Networks (VPN): A dedicated communications network of enterprise/s and user/s by using tunneling over public networks (Internet).

Lesson No. 09

ESSENTIAL CHARATERISTICS OF CLOUD COMPUTING

Module No – 041:

- On-demand self-service: The user can automatically be allocated the computing resources without any manual operations (except the initial signing up process). The cloud management software handles the resource management and provisioning.
- Broad Network Access: The cloud resources can be accessed through network through broad range of wired and wireless devices. Various connectivity technologies are available.
- Resource pooling: Resources (Computing, memory, storage, network) are available in volumes and therefore can be pooled. The resources can be physical or virtual. Multiple users can simultaneously share these resources through dynamic allocation and reallocation.
- Rapid elasticity: The cloud resources are virtually unlimited. So much so, the provisioning of these resources can shrink and expand elastically according to demand.
- Measured Service: The resource usage is charged by the provider from users, according to usage.

Module No – 043: Revisiting 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***."

National Institute of Science and Technology (NIST) USA

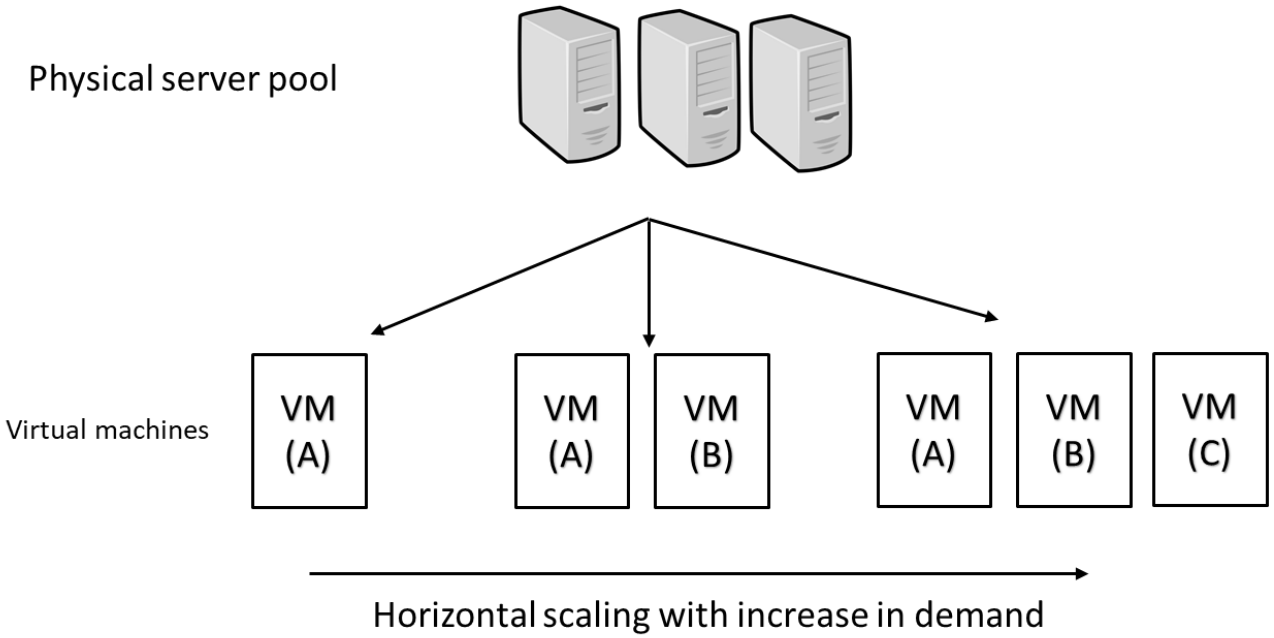
- Cloud computing can fulfill the business drivers such as
    - IT Capacity Planning
    - Cost Reduction
    - Organizational Agility

Module No – 044: Some key terms about Cloud Computing:

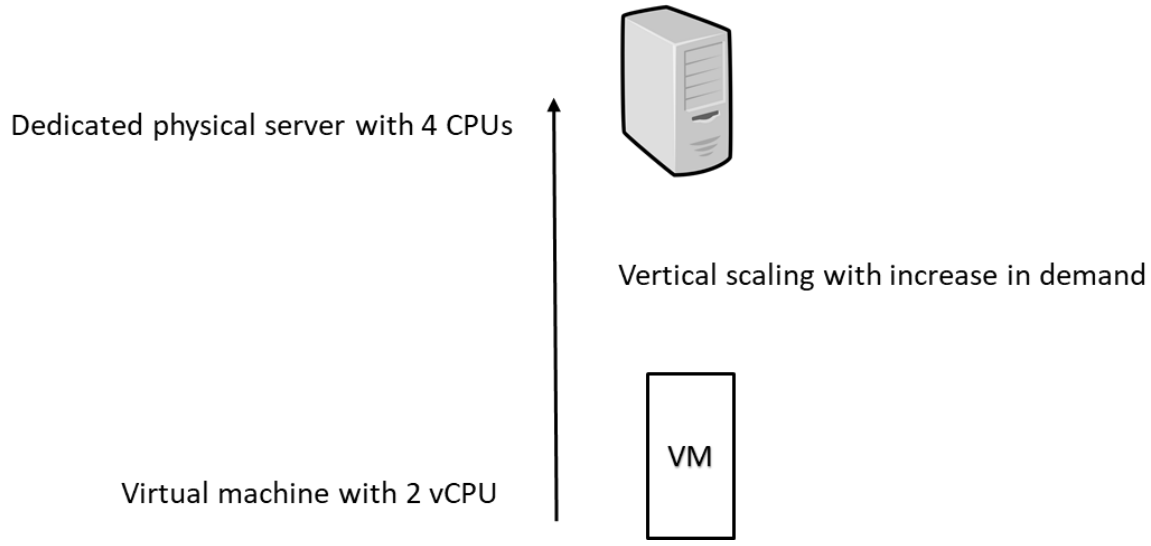
- Some key terms and concepts essential for understanding Cloud Computing course:
  - IT Resources
  - On-premises
  - Cloud Consumers
  - Cloud Providers
- Cloud IT Resources: Can be physical or virtual resources (virtual resources are implemented in software):
  - Physical/Virtual machines/servers
  - Physical/virtual storage
- On-premises: An IT resource which is hosted/located at the enterprise's premises.
  - It is different from a Cloud resource since a Cloud resource is hosted on Cloud.
  - An on-premises IT resource can be connected to a Cloud resource and/or can be moved to a Cloud.
  - However the distinction is difficult for private clouds.
- Cloud Providers: The party providing the cloud-based IT resources.
- Cloud Consumer: The user of cloud-based IT resources is called *cloud consumer*.

Module No – 045: Scaling, Cloud Service Providers & Consumers:

- Scaling: It refers to the ability of an IT resource to handle increased or decreased usage demands.
  - Scaling: It refers to the ability of an IT resource to handle increased or decreased usage demands.
- Following are the types of scaling:
  - Horizontal scaling: It is the scaling out or scaling in of the IT resources of same type. The number of resources increases or decreases according to load.
  - Commodity hardware can do the work, instantly available IT resources, not limited by hardware capacity



- Vertical scaling: When an IT resource is replaced with a resource of higher capacity (scaling up) or when replaced with the resource of lower capacity (scaling down) according to workload.
- Specialized server are required, instantly available IT resources, additional setup is required (downtime required during replacement), limited by maximum hardware capacity, less common in Cloud.



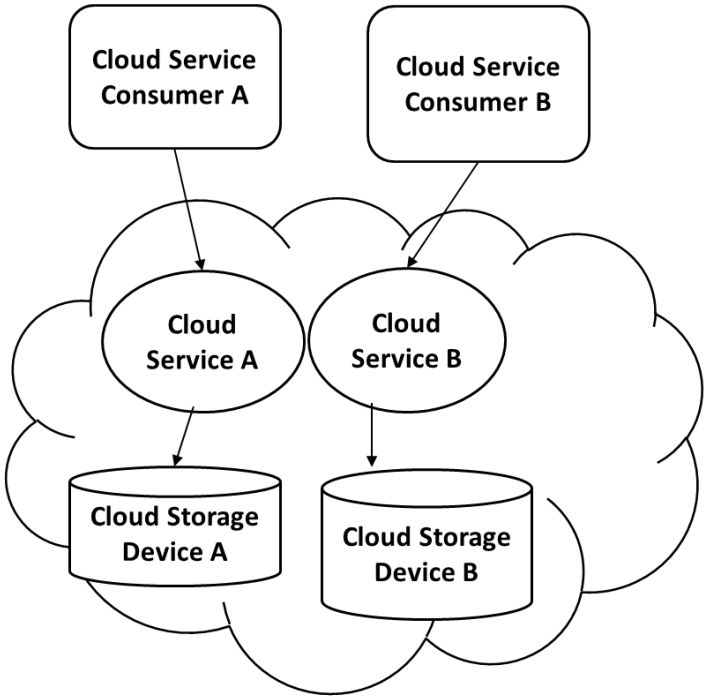
- Cloud Service:
  - Any IT resource (software/VM) that is made remotely available by the cloud provider.

- Remember that not all the IT resources deployed in a cloud environment are remotely accessible. Some resources are used within the Cloud for support and monitoring etc.

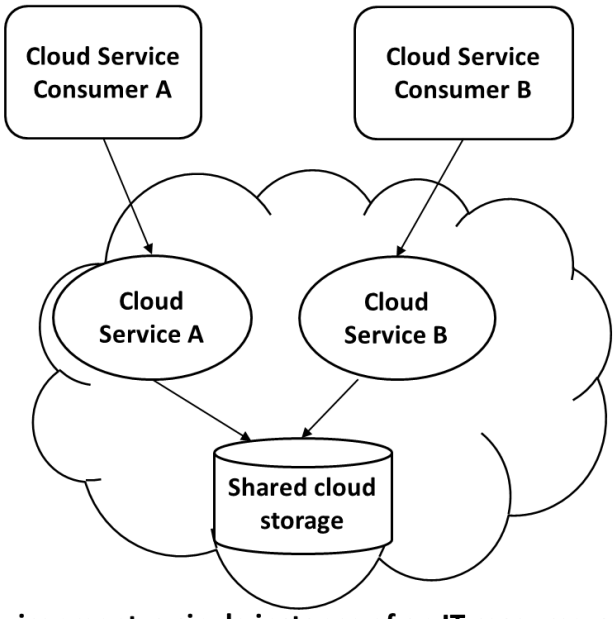
- The human users interact with a leased VM.
- Client programs interact with cloud software service/s through API calls.
- The software program and service accessing a *cloud service* is called a *cloud service consumer*

Module No – 053: Multitenancy:

- A software architecture consisting of software executing over a server and serves different users (tenants) whereby each tenant is isolated from the others.
- Cloud computing serves different cloud consumers by using virtualization software frequently.



In single-tenant environment, there is a separate IT resource for each tenant.

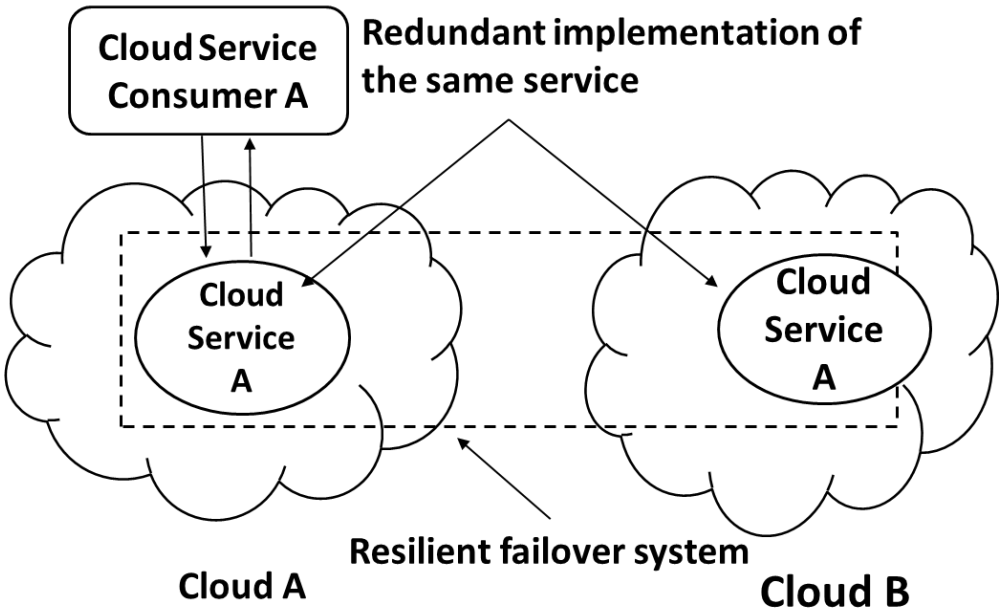


Multi-tenant environment, a single instance of an IT resource such as Cloud storage device serves multiple consumers.

- The cloud provider pools the IT-resources by using multitenancy technology to dynamically assign and reassign these resources according to cloud consumers' demands.
- The physical as well as virtual resources are multi-tenanted (or shared) by using statistical multiplexing.

Module No – 054:

- Resiliency: The ability of a computer system to recover from a failure is called resiliency.
    - The redundant implementation of IT-resources paves the way to a resilient system.
    - The whole system is pre-configured so that as soon as a resource fails, the processing is automatically handed over to the redundant resource.
    - Resiliency is one of the features of cloud computing whereby the redundancy of IT-resources is implemented at different physical locations and/or in different clouds.
    - For example the data can be kept at two different locations and replicated. If the primary hard disk fails, the secondary drive takes the data connectivity.
- A cloud service can be configured at two different VMs (A and B) and each VM is placed on a separate server or a different cloud. VM B is kept as failsafe resource. In case VM A fails, the VM B starts processing the user service user/s requests.



Lesson No. 10

BENEFITS OF CLOUD COMPUTING

Module No – 046:

- The immediate benefit of using Cloud is the reduction in initial cost.
- The initial costs include:
    - Infrastructure costs:
        - IT equipment
        - Software
        - Networking
        - Construction costs
        - Installation costs
        - The infrastructure costs can be regarded as capital investments or ownership costs. The cloud saves the initial upfront ownership costs. The cloud offers affordable and attractive packages for services obtained in large volume. The cloud reduces investment and proportional costs.
    - Proportional cost or operational costs (as discussed before): The cloud rental can replace this cost. The rental costs are highly competitive.
- The cloud provider can increase the profit by increasing the resource utilization, using proven practices and by optimizing the cloud architecture.
- Common measurable benefits for the cloud consumers are:
- Pay-as-you-go rental for short term usage

- The availability of virtually unlimited resources on demand with negligible wait time for provisioning.
- The IT resources can be added or removed in a fine grained level e.g., 1 GB of storage increments
- Applications and resources can be migrated across regions if required.

Module No – 047: Increased Scalability, Availability & Reliability

- Increased scalability: The cloud can dynamically and instantly provide the computing resources.
- This provision can be on demand or as per user configuration.
- Similarly these IT resources can be released automatically or manually with the decrease in processing demand.
- This dynamic scalability avoids the over-provisioning and under-provisioning and the associated disadvantages.
- Availability: The availability of IT resources sometimes can be referred to profit and customer retention.
    - If an IT resource becomes unavailable (such as a database dealing with clients' orders) then this may result in customer dissatisfaction and loss of business.
- Reliability: The reliability of IT resources is very important for continual business data processing and response time.
    - The failure of any IT resource can be cause the collapse the IT system. For example failure of the Ethernet switch may crash a distributed application.
- The modular structure and resource redundancy in cloud increases the availability and reliability. Cloud, on the other hand provides a guaranteed level of availability and reliability through a legal agreement called service level agreement (SLA) between the cloud provider and cloud user.
- The recovery time after failure is the added penalty. It is the time when the system remains unavailable.
- The modular structure and resource redundancy in cloud increases the availability and reliability. It also improves the recovery time.

Lesson No. 11

RISKS AND CHALLENGES OF CLOUD COMPUTING

Module No – 048:

- The term *vulnerability* refers to a state of being attacked.
    - Moving the business data to cloud can introduce vulnerabilities and security risks.
- The term *security framework* refers to the procedures and practices for securing the resource such as data, network and IT infrastructure.

- Unless the cloud provider and cloud user are covered under same security framework, the vulnerabilities are unavoidable.
- The cloud provider and user have to be in a *trust* relationship. The factors affecting the trust may include the following facts:
    - The data is being accessed remotely.
    - There are multiple users sharing the cloud based IT resources such as virtual storage.
    - The cloud provider has a privileged access to the users' data.
    - The security of the data depends upon the security policies of the provider and the consumer.
    - There can be malicious consumers (human and automated) who can benefit from the security vulnerabilities of the cloud environment by stealing and/or damaging the business data.

Module No – 049:

- Reduced operational governance control: The cloud consumer gets a lesser privileged control over the resources leased from the cloud.
- There can be risks arising as to how the cloud provider manages the cloud.
- An unreliable cloud provider may not abide by the guarantees offered in SLA of the cloud services. This will directly affect the quality of cloud consumer solutions (enterprise software) which rely upon these services.
- The cloud consumer should keep track of actual level of service being provided by the cloud provider.
    - The SLA violations can lead to penalties receivable from the cloud provider.
- Limited portability between cloud providers: Due to lack of industry standards for cloud computing, the public clouds environments remain proprietary to their providers.
- It is quite challenging to move a custom-built software from one cloud to another if it has dependencies upon the proprietary environment (such as security framework) of the former cloud.
- Multi-regional compliance and legal issues: Cloud providers tend to set their data centers in regions favoring affordability and/or convenient. This may lead to legal issues for cloud provider as well as cloud consumers.
- Some countries such as some UK laws require the personal data of UK citizens to be hosted inside UK.
- Thus a cloud provider with multi-regional data centers including UK, can not migrate the UK citizen's personal data outside UK.
- The UK citizens are legally bound to keep the personal data on clouds hosted in UK only.
- Some countries such as USA allows government agencies' access to data hosted inside USA.
- Despite that the owners of this data are neither residing inside nor the citizens of USA, but still their data is accessible by the USA government agencies if hosted inside USA.

Lesson No. 12

ROLES AND BOUNDARIES IN CLOUD COMPUTING

Module No – 050:

- Cloud provider: The organization that provides the IT resources.
    - Responsible for providing IT resources as per SLA.
    - Also performs the management and administrative tasks to assure flawless provisioning of cloud services.
    - A cloud provider usually owns the IT resources of the cloud.
    - It is also possible that the cloud provider resells the cloud services leased from another cloud providers.
- Cloud consumer: The organization or individual who has contracted with cloud provider to lease/rent the cloud IT-resources through user interface and/or through software API calls.
    - In the later case, a cloud consumer uses a *cloud service consumer* (a software program) to interact/use a cloud service.
- Cloud Service Owner: Is the one who owns the cloud service. Can be:
    - Cloud consumer: If the deployed service is on leased IT-resources.
    - Cloud provider: If the cloud provider has deployed the service on cloud IT-resources.
    - A cloud service owner may not be the owner of the cloud IT-resource.

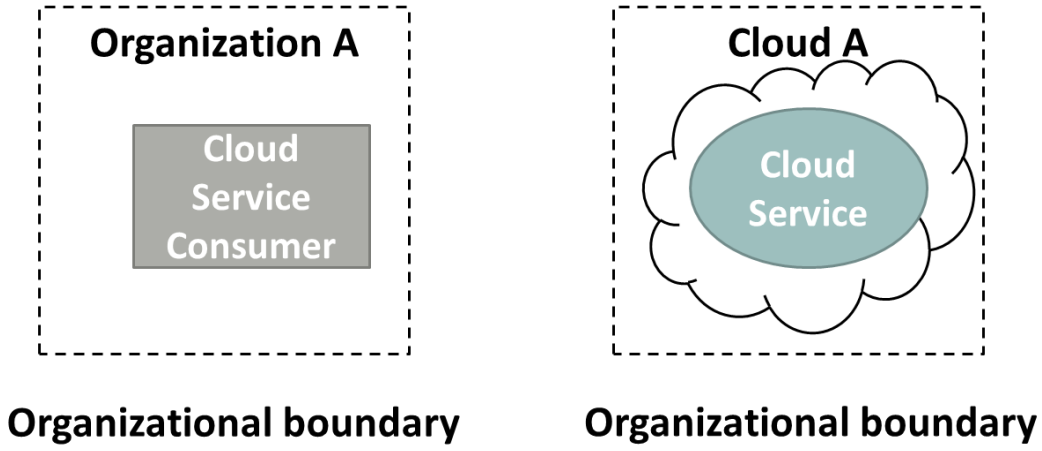
Module No – 051:

- Cloud Resource Administrator: This role is responsible for administering the cloud resources (including cloud services).
    - Cloud resource administrator can be:
        - Cloud consumer (as cloud service owner)
        - Cloud provider (when the service resides inside the cloud)
        - Third party contracted to administer a cloud service
- Additional roles:
    - Cloud Auditor: Provides an unbiased assessment of trust building features of the cloud. These include the security, privacy impact and performance of the cloud. The cloud consumer may rely upon the cloud audit report for choosing a cloud.
    - Cloud Broker: A party that provides mediation services to cloud providers (seller) and cloud consumers (buyer) for the purchasing of cloud services.
    - Cloud Carrier: The party responsible for providing connectivity between cloud provider and cloud consumer. The ISPs can be assumed as cloud carriers.
    - The cloud provider and cloud carrier are in legal agreement (SLA) to assure a certain level of connectivity and network security.

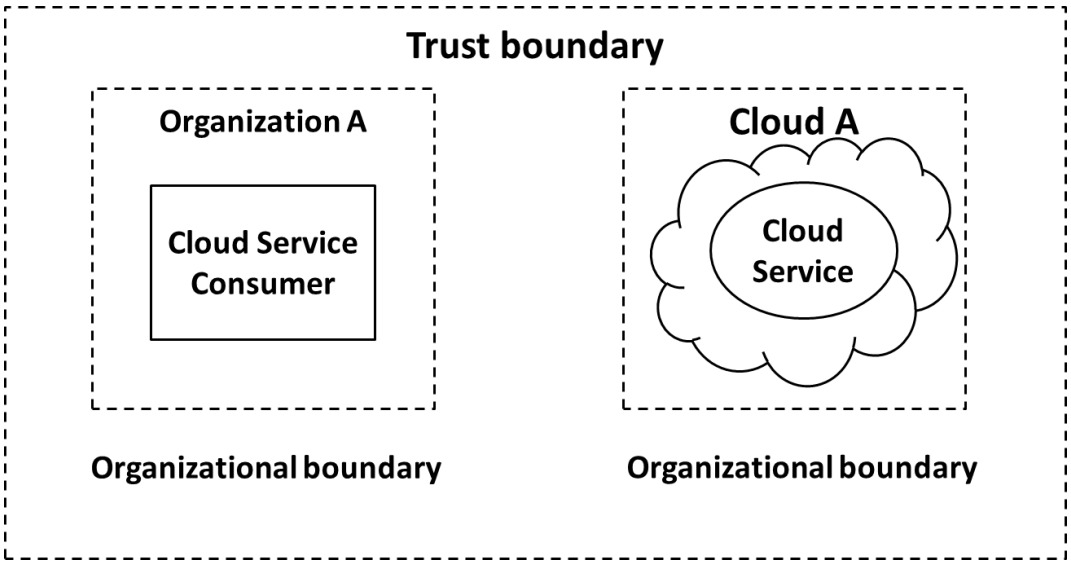
Module No – 052:

- Organizational boundary: This is a boundary of ownership ad governance of IT assets of an organization.

- Similarly, the cloud has its organizational boundary.
- This is a boundary of ownership ad governance of IT assets of an organization.
- Similarly, the cloud has its organizational boundary.



- Trust boundary: When an organization takes the role of cloud consumer, then it has to extend its trust boundary to include the cloud resources. A trust boundary represents a border around trusted IT-resources.



Lesson No. 13

CLOUD SERVICE MODELS

Module No – 055: IaaS, PaaS & SaaS Provisioning:

- IaaS: The IT-resources are typically virtualized and packaged in a preplanned way.
  - The IT-resources are usually freshly instanced e.g., VMs.
  - The cloud consumer has a high level of control and configuration-responsibility.
  - The cloud consumer also has the duty of configuring these resources.
  - Sometimes a cloud provider will contract IaaS offerings from other cloud provider to scale its own cloud environment.
  - The VMs can be obtained specifying the hardware requirements such as processor capacity, memory, storage etc.
- PaaS: Delivers a programming environment containing preconfigured tools to support the development lifecycle of custom applications.
  - PaaS products are available with different development stacks such as Google App Engine provides a Python and Java environment.
  - The PaaS is chosen:
    - To enhance or substitute the on-premises software development environment.
    - To create a cloud service in order to provide a cloud service to other cloud consumers.
  - The PaaS saves the consumer from administrative tasks such as installations and configurations to set up the software development infrastructure.
  - On the other hand the cloud consumer has lower level of control over the underlying infrastructure.
- SaaS: Is the software hosted over cloud infrastructure and offered as a utility services.
  - SaaS is provided as a reusable utility service commercially available to different users.
  - A SaaS can be deployed over IaaS and/or PaaS instance. Whereby the cloud consumer (of IaaS/PaaS) becomes the provider.
  - The service consumer has a very limited control over the underlying SaaS implementation.

Module No – 056:IaaS, PaaS & SaaS Comparison

- Control level:
  - SaaS: Usage and usage related configuration
  - PaaS: Limited administrative
  - IaaS: Full administrative
- Functionality provided to cloud consumer:
  - SaaS: Access to front-end user-interface
  - PaaS: Moderate level of administrative control over programming platform
  - IaaS: Full administrative control over virtual resources of the VMs
- Common activities of cloud consumer:
  - SaaS: Use and configure the service
  - PaaS: Develop, debug and deploy the cloud services and cloud based solutions
  - IaaS: Installation and configuration of software, configure the infrastructure of VM

- Common Cloud Provider's Activities:
  - SaaS: Implementation, management and maintenance of cloud service.
  - PaaS: Providing the pre-configured programming platform, middleware and any other IT resource needed.
  - IaaS: Provisions and manages the VMs and underlying physical infrastructure.
- The three cloud models of cloud delivery can be combined in a way that one delivery model is deployed over another. Such as:
  - PaaS over IaaS
  - SaaS over PaaS
  - SaaS over PaaS over IaaS

Module No – 057: Software as a Service (SaaS) Overview:

- NIST definition of SaaS: *"Software deployed as a hosted service and accessed over the Internet."*
- The SaaS is a software solution having the code and data executing and residing on cloud.
- A user accesses the SaaS through browser.
- Remember: *The cloud service consumer is a temporary runtime role assumed by a software program when it accesses a cloud service.*
- [Thomas Erl [2014], Cloud Computing Concepts, Technology and Architecture, Pearson]
- For the time being we shall assume that the browser acts as *cloud service consumer* when accessing a SaaS.
- SaaS solutions eliminate the need of on-premises (data center based) applications, application administration and data storage.
- The customer is allowed to adopt pay-as-you-go type of rental.
- SaaS offers scalability and device-independent access to the SaaS solution/s.
- SaaS provider assures that the software provided is solidly tested and supported.
- The notable disadvantage of SaaS is that the data resides off-premises.
- The notable disadvantage of SaaS is that the data resides off-premises.
- Therefore the data security is of prime importance because the customers' data may be proprietary and business-sensitive.
- The SaaS provider offers SaaS apps executing over IT-resources. These resources can be from a physical servers or a VM owned/rented by the provider.
- Each instance of a SaaS app (consumed by a user) is allocated separate set of IT-resources.
- Classes of SaaS:
  - Business logic: Connect the suppliers, employees, investors and customers.
    - Example: Invoicing, fund transfer, inventory management, customer relationship management (CRM)
  - Collaboration: Support teams of people work together.
    - Examples: Calendar systems, email, screen sharing, conference management and online gaming.
  - Office productivity: Office environment support.
    - Examples: word processors, spreadsheets, presentation and database software.
  - Software tools: For the support of developing software and solving compatibility problems.

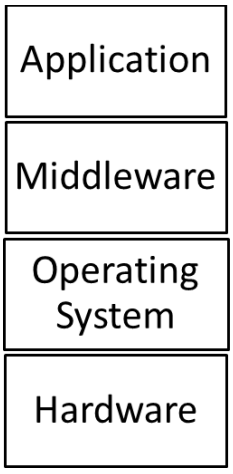
        - Examples: format conversion tools, security scanning, compliance checking and Web development.
- Software that are not suitable for public SaaS offerings (according to NIST):
    - Real-time software: They require precise response time. Due to variable response time and network delays, these software are not suitable to be offered as SaaS. Such as flight control systems and factory robots etc.
    - Bulk-consumer data: When extremely large amount of data is originating physically at the consumer's side such as physical monitoring and patient monitoring data. It is not feasible to transfer this data in real time over WAN to SaaS provider.
    - Critical software: A software is labeled *critical* if its failure or delay in handling can cause loss of life or loss of property. These software are not suitable for SaaS because achieving a continuous acceptable reliability for critical software in public SaaS is quite challenging due to (unreliable) public network based access.
- SaaS billing: Based on
    - Number of users
    - Time in use
    - Per-execution, per-record-processed
    - Network bandwidth consumed
    - Quantity/duration of data stored

Module No – 058:SaaS Examples:

- Salesforce.com SaaS for Customer Relationship Management (CRM)
    - Manage sales contacts and leads.
    - Centralize the contact. information and project details.
    - The sales reports from any place any time.
    - The sales reports from any place any time.
    - Manages and syncs sales contacts and meetings with other tools such as Microsoft Outlook.
- Taleo SaaS for Human Resources Management (HRM):
    - Recruitment tools to manage the applicants' data for hiring purposes.
    - Performance management and tracking tools for employees' evaluation.
    - Performance management and tracking tools for employees' evaluation.
    - Compensation tools for rewarding the employees according to performance.
    - Workforce training and professional development tools
- ADP SaaS for Payroll Processing and HRM:
    - Cloud solution for time management, employees benefits calculation, worker compensation and HR issues.
- Carbonite SaaS for File Backups:
    - Provides backup services for precious business data and personal data. The data is kept securely and redundantly.
- Microsoft Office 365 SaaS for Document Creation, Editing and Sharing:
    - In order to provide the documentation tools at affordable price and to compete with the freeware solutions, Microsoft offers its flagship software suite on monthly rental basis.

Module No – 059:SaaS Software Stack:

- The provider controls most of the software stack.



SaaS Software Stack

- Application: Email
- Middleware: software libraries, run time environments (Java, Python)
- Service provider has admin control over application and total control over the rest of the layers.
- Service consumer has limited admin control over the application and no control over the rest of the stack.
- A consumer can create, send and manage the emails and even the email accounts.
- But the email provider has absolute control over the SaaS software stack in order to perform its duties such as provisioning, management, updates and billing in email app.

Module No – 060: SaaS Benefits:

- Modest software tool footprint: There is no need for complex installation procedures because the SaaS applications and accessible through web browsers. This is one of the reasons of widespread use of SaaS applications.
- Efficient use of software licenses: The license issuance and management procedure is quite efficient. A single client is issued a single license for multiple computers. This is because the software is running directly on provider's infrastructure and thus can be billed and monitored directly.
- Centralized management and data: The consumer's data is stored in cloud. The provider assures the security and availability of data. The data seems centralized for the consumer may in fact be distributed and replicated by the provider. Data backup is provided at possibly additional charges.
- Platform responsibilities managed by providers: Consumer does not has to bother about operating system type, hardware and software configurations, software installation and upgrades.

- Savings in up-front costs: (as discussed before) the up-front costs such as equipment acquisition and hardware provisioning etc. are avoided by SaaS consumer.
- The provider is responsible for operational issues such as backups, system maintenance, security software, upgrades, trouble shooting in software, physical security and hardware management etc.

Module No – 061: SaaS: Issues and Concerns:

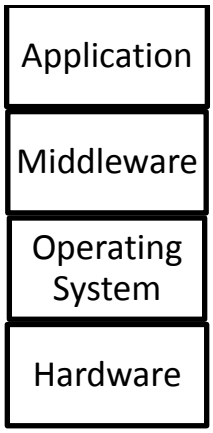
- The NIST has identified few issues and concerns about SaaS.
- Most of these issues are due to network dependency of SaaS.
    - Browser based risks and remedies: Since the SaaS is accessed through browser installed on consumers' device, the inherent vulnerabilities of the web browsers do have impact over SaaS security.
        - Although the browsers apply encryption upon network traffic, yet various network attacks such as brute force and man in the middle attacks are possible upon the SaaS data.
        - The resources leased by a consumer can be hijacked by malicious users due to poor implementation of cryptographic features of browsers.
        - If the consumer's browser is already infected with a security threat (due to a visit to malicious website) then later, the same browser is used for SaaS access, then the SaaS data might get compromised.
        - If a single consumer accesses multiple SaaS services using browser instances, then the data of these SaaS instances may get mixed up.
        - A few suggestions by NIST:
            - Use different browsers to access each different SaaS.
            - Do not use the same web browser for web surfing and SaaS access.
            - Use a VM to access the SaaS.
    - Network dependence: SaaS application depends upon reliable and continuously available network.
        - The reliability of a public network (Internet) can not be guaranteed as compared to dedicated and protected communication links of private SaaS applications.
    - Lack of portability between SaaS clouds:, It may not be trivial to import export data among different SaaS applications deployed over different clouds due to customized development and deployment of SaaS applications and data formats.
    - Isolation vs. Efficiency (Security vs. Cost Tradeoffs): The SaaS provider has to make a trade-off decision as to deploy separate IT-resources (such as VMs) for each client or concurrently server multiple clients through a single deployment of SaaS application.

Module No – 062: NIST Recommendations for SaaS

- Data protection: The consumer should analyze the data protection, configuration, database transaction processing technologies of SaaS provider. Compare them with the confidentiality, integrity, availability and compliance requirement of the consumer.
- Client device/application protection: The consumer's client device (browser running over a computer) should be protected to control the exposure to attacks.
- Encryption: Strong encryption algorithm with key of required strength should be used for each web session as well as for data.
- Secure data deletion: The data deletion through consumer's request should be reliably done.

Module No – 063: Platform As a Service (PaaS) Overview

- According to NIST, PaaS provides a toolkit for conveniently developing, deploying and administering application software which can support a large number of users, process large volumes of data and can be accessed over Internet.
- What does PaaS clouds really provide: a set of software building blocks, a set of development tools (languages and compilers) and supporting environments for run-time of applications developed over PaaS.
- PaaS clouds also provide tools to deploy the developed applications.
- Additionally, the PaaS clouds provide processing, storage and networking resources.
- PaaS consumers:
    - Application developers
    - Application testers
    - Application deployers
    - Application administrators
    - Application end users (SaaS users)
- The consumers are charged according to tools and IT-resources usage.
- PaaS Software stack: The cloud provider fully controls the hardware and OS layers:



PaaS Software stack

- PaaS Provider/ Consumer Scope of Control: The provider has administrative control of middleware.

- The provider has no control over application layer.
- Remember that the application developed by using PaaS is deployed as SaaS and the PaaS consumer has full administrative control over that SaaS.
- The provider however controls the runtime-environment which is necessary for PaaS application.
- PaaS billing: Usually based on:
    - Number of consumers
    - Kind of consumers (e.g., developers vs. application end users)
    - Storage, processing, or network resources consumed by the platform
    - Requests serviced
    - The time the platform is in use.

Module No – 064: PaaS Examples:

- We are going to discuss a few examples of PaaS.
    - Google App Engine (GAE): Allows the users to create and host web based (Java, Python & Go) applications running over the infrastructure and services provided by Google. GAE is a free service until the application grows to a significant size.
    - Force.com as a PaaS: This is a service of Salesforce.com (a SaaS provider). It offers four different programming environments for nonprogrammers, programmers and software vendors.
        - Nonprogrammers can create finance, HR etc. applications and websites without coding by using drag drop of controls.
        - Programmers can develop Java applications and deploy them as SaaS.
        - The software vendors can distribute and update their applications over cloud by using Force.com.
    - LongJump as a PaaS: Supports the entire cycle of software development from requirement gathering to application release and support. It is free of cost.
    - Openshift as a PaaS: It is a PaaS offering from Red Hat which is also the distributor for Red Hat Linux. Openshift PaaS provides the primary development tools for cloud based solutions written in PHP, Python and Ruby.
        - Openshift also provides development tools for Linux-based solutions written in C programming language.
    - Windows Azure and SQL Azure as a PaaS: Provided by Microsoft as a paid service. The users can develop applications in .Net as well as Java, PHP and Ruby.
        - SQL Azure provides database solutions for application developed and running inside Windows Azure.

Module No – 065: Benefits and Disadvantages of PaaS Solutions:

- Benefits:
    - Lower total cost of ownership in terms of hardware and software investment.
    - Lower administrative overhead of system development.

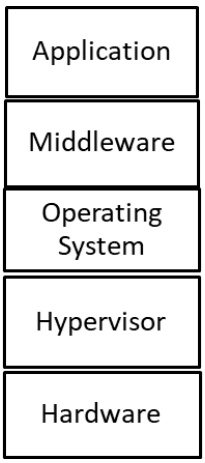
  - No requirement of software upgrades of tools.
  - Faster application development and deployment.
  - Scalable resources available for the applications. The user pays only for the resources used.
- Disadvantages:
  - The inherent problem of data placed offsite raises the security concerns.
  - The integration of PaaS applications with on-site legacy solutions is not trivial.
  - The PaaS provider has to be trusted for data and application security.
  - The issues of SaaS are also the issues of PaaS such as browser based risks, network dependence and isolation vs efficiency.
  - Portability of PaaS applications across different providers may not be possible sue to incompatibility in coding structures (hash, queue, file etc.).

Module No – 066: PaaS Recommendations:

- Generic interfaces: The consumer should make sure that the interfaces for hash tables, queues and files etc. are generic so that there will be less issues of portability (among PaaS providers) and interoperability (of applications) in future.
- Standard language and tools: Choose a PaaS provider which offers standardized language and tools unless it is absolutely unavoidable to use the proprietary languages and tools.
- Data access: The provider with the standardized data access protocol (such as SQL) should be preferred.
- Data protection: The confidentiality, compliance, integrity and availability needs of the organization should be compared with the data protection mechanisms of the provider.
- Application framework: The PaaS providers which offer the features in application development framework for eliminating security vulnerabilities of the application should be chosen.
- Component testing: The software libraries provided by the PaaS provider should be aiming at providing proper functionality and performance.
- Security and secure data deletion: Ensure that the PaaS applications can be configured to run in a secure manner (e.g., using cryptography during communication) and that a reliable mechanism for data deletion is provided by the PaaS provider.

Module No – 067: Infrastructure As a Service (IaaS) Overview:

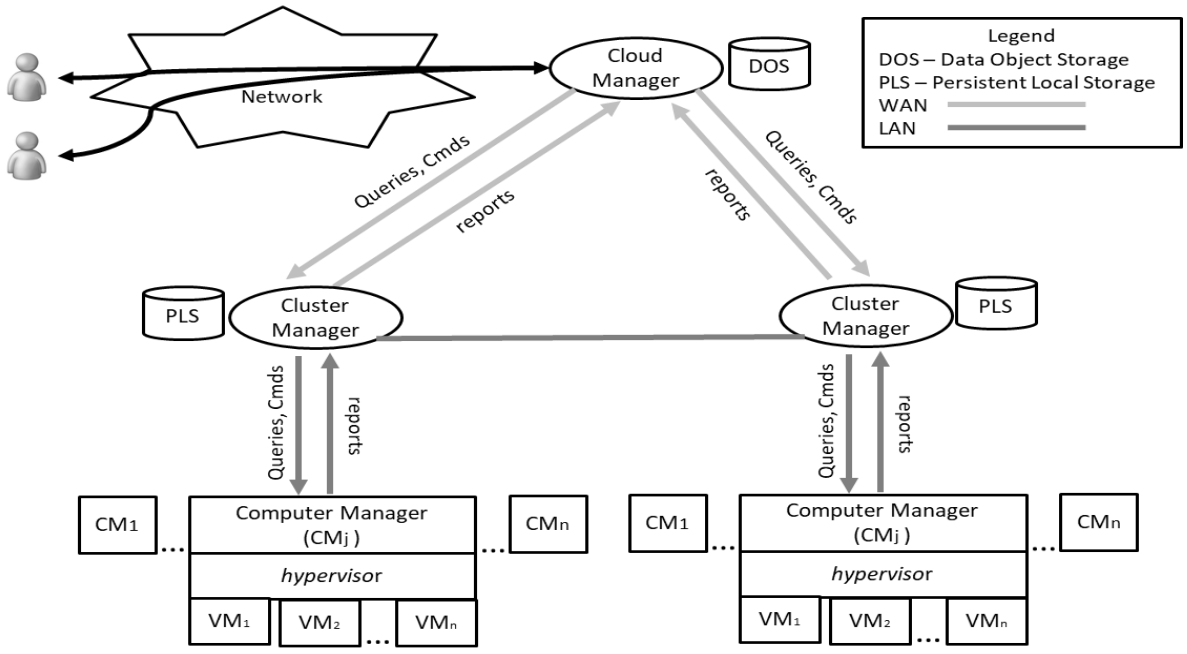
- As an alternative to PaaS, some consumers may prefer to use IaaS in order to have management control over the IT resources.
- The IaaS provider makes available the computing resources in the form of VMs.
- The consumer has the duty of installing OS and software.
- The provider also provides stable network access, network components such as firewalls, and data storage.
- IaaS Provider/Consumer Scope of Control: The provider has no control over top three layers.



IaaS Software Stack

- IaaS Provider/Consumer Scope of Control: The provider has admin control over hypervisor and total control over hardware layer.
- IaaS Provider/Consumer Scope of Control: The consumer has total control over top three layers.
- IaaS Provider/Consumer Scope of Control: The consumer can request the provider to deliver a VM from hypervisor layer.
- The consumer has no control over hardware layer.
- Customer billing:
    - Per CPU hour
    - Data GB stored per hour
    - Network bandwidth consumed, network infrastructure used (e.g., IP addresses) per hour
    - Value-added services used (e.g., monitoring, automatic scaling).

Module No – 068: IaaS Operational Overview:



The operational infrastructure of IaaS

Module No – 069: IaaS Benefits:

- Saving in upfront cost: As in SaaS and PaaS. Although the responsibility of installing OS and software is of the consumer.
- Full administrative control over VM:
  - Start, shut down, pause
  - Installation of OS and applications
  - Accessing VM through network services of VM through a network protocol such as Secure Shell.
  - Flexible and scalable renting: The VMs can be rented in any volume desired by the consumer. The rental for each VM can be on usage (of raw resources such as CPU, memory, bandwidth, storage, firewall, database etc.) basis.
  - Portability and interoperability with legacy applications: Since the consumer has full control over the VM to install OS and other applications, the legacy applications (which are usually installed on consumer owned server/s) can be configured to run with or ported to the VM.

Module No – 070: IaaS Issues and Concerns:

- Network dependence
- Browser based risks (same as discussed for SaaS and PaaS).
- Compatibility with legacy software vulnerabilities: Since the consumer is allowed to install the legacy applications on VMs rented through IaaS, this exposes the VMs to the vulnerabilities in those legacy software.

- Implementation challenges exist for VM isolation: In order to prevent the VMs from eavesdropping other VMs mounted over same server, the isolation features of hypervisor are utilized. But these features may not withstand a sophisticated attacks.
- Dynamic network configuration for VM traffic isolation: A dynamic network path is provided from VM to consumer when a VM is rented. The provider has to isolate VM consumers from accessing the network traffic of other consumers.
- Data erase practices: When a VM is no longer rented by a consumer, the virtual drive of that VM must be erased/overwritten multiple times to eliminate any chance of residual data access by the next consumer of that VM.
- NIST recommendations for IaaS: The provider should implement data and network traffic isolation for the VM consumers. The features of data security as well as secure deletion of residual data of VM consumer.

Lesson No. 14

DATA STORAGE IN CLOUDS

Module No – 073:Network Storage:

- Computers attached to a local area network (LAN) may require additional storage space to support file sharing, file replication and storage for large files.
- Traditionally this additional space is provided through file servers which have larger disk capacity.
- With the evolution of computer networks, the file server was extended through the use of storage area network (SAN).
- The SAN enabled storage devices are attached to the network.
- The software running over SAN devices allows direct access to these devices throughout network.
- Later on, a class of storage devices emerged to be implemented as network attached storage (NAS).
- Advantages of network storage (particularly of SAN) are:
    - Data reliability and reconstruction through replication.
    - Better performance than file server.
    - Compatibility with common file systems and operating systems.
    - Best choice for backups.

Module No – 074:Cloud Based Data Storage:

- Cloud storage is the next step in the evolution of network storage devices.
- Instead of storing the data locally, the data can be stored on cloud and can be accessed through web.
- The user can have virtually unlimited storage space available at affordable rates.
- There are various modes of data access in Cloud:
    - Using web browser interfaces to move the files to and from the cloud storage.
    - Through a mounted disk drive that appears local to the user's computer.

    - Through API calls to access the cloud storage.
- There are a number of cloud storage providers which offer file storage, sharing and synchronization. Such as:
- Carbonite
- pCloud
- Dropbox
- ElephantDrive
- These providers offer a certain volume of free storage as well as paid storage at low prices.

Module No – 075:Cloud Based Data Storage: Advantages & Disadvantages:

- Advantages:
    - Scalability: The user can scale the storage capacity (up or down) according to requirement.
    - Various convenient costing models are available from one time payment to monthly payment to pay as per use.
    - Reliability: The storage providers provide the assurance for data reliability (through replication).
    - The data can be accessed worldwide by using Internet.
    - Various methods of data access are available (as discussed before).
- Disadvantages:
    - Performance: Because of the Internet based access, the cloud storage can never be as fast as SAN or NAS based local storage.
    - Security: Not all the users may be able to trust the cloud provider for the users' data.
    - Data orphans: The user has to trust the data deletion policies of the provider. The files (on cloud storage) deleted by the user may not be immediately (or ever) be deleted from the cloud storage.

Module No – 076: Cloud Based Backup Systems:

- The term *backup* refers to the copying of (data and/or database) files to a secondary site for preservation in case of device or software failures.
- Backup is an important part of disaster recovery plan.
- In case of a disaster, the data can be restored to the state of last backup.
- Cloud based backup system comprises of procedures to send the copy of data over a proprietary or public network to a remote server hosted by the cloud service provider.
- The provider charges the user according to number of accesses or data volume or number of users.
- Cloud based backup or online backup system is implemented through a client software installed on the user's computer. The software collects, compresses and sends the data to cloud backup on timely basis.
- Advantages:
    - The data is backed up in encrypted form.

  - Backup can be performed on the convenience of user (daily, weekly, monthly).
  - The user can easily retrieve the backup files from the cloud.
- Disadvantages / Limitations:
  - Due to security concerns, the critical data backup is preferably stored on local storage.
  - The long term data storage in heavy volume over cloud may have humongous cost.
  - Due to network cost, the incremental backup is preferred.

Module No – 077: Database and Block Storage:

- A Cloud database is a database that resides on Cloud platform.
- The Cloud database can be accessed by:
  - The applications hosted on Cloud
  - The application hosted locally (can access through Internet)
- The cloud database is provisioned in either of the following methods:
  - Installed on a rented VM by the user
  - As part of PaaS
  - Provided as a service by cloud provider or the database companies.
- Advantages of Cloud based Database solutions:
  - Cost effective scalability as per use
  - High availability of database software through redundant hardware (minimizes the downtime in case of failure)
  - High availability of data due to replication of database
  - Reduced administration of database provided as service or as part of PaaS.
- Disadvantages of Cloud based database solutions:
  - The user may not trust the cloud provider regarding sensitive data
  - Due to Internet based access, the Cloud based database is not as fast as a locally installed database.
- There are a number of cloud based database providers such as:
  - Oracle
  - Amazon
  - Microsoft

- Cloud based block storage is a sequence of bits and provided as a block on cloud storage.

- It is suitable in the following situations:

  - When the data may not map properly on a file system or on a database

  - The application developer wants to store data in a customized file system

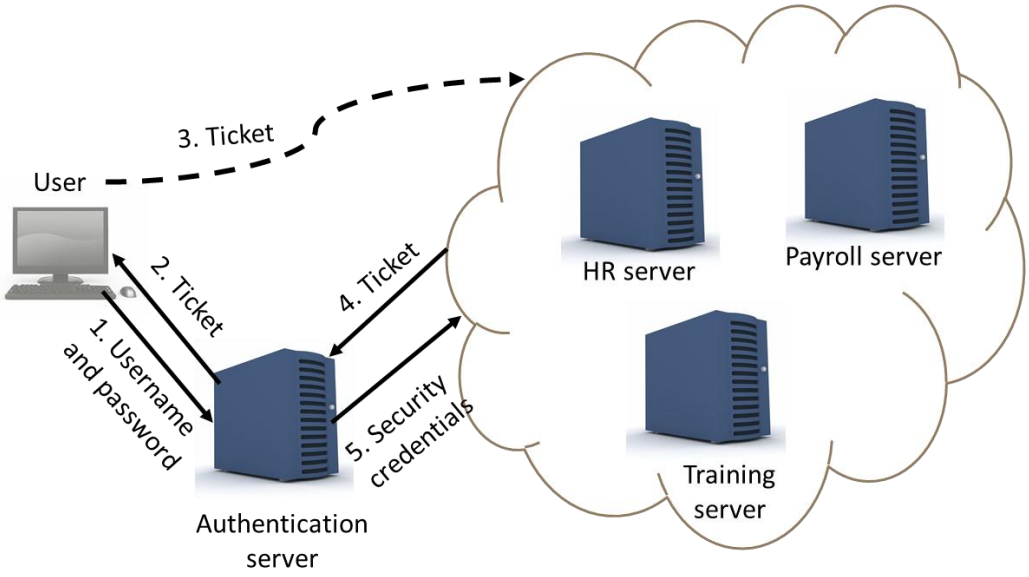
- Amazon Elastic Block Store (EBS) is a highly available, scalable and reliable block storage solution which supports block sizes of up to 1 terabytes.

Lesson No. 15

MISCELLANEOUS SERVICES OF CLOUD COMPUTING

Module No – 071:Identity as a Service (IDaaS):

- Today within most companies, the users may have to log in to several applications servers (on premises and/or cloud) to perform daily tasks. Some of these systems may be cloud based.
- The user has to remember multiple logins and passwords.
- When a user leaves a company, the related logins and passwords must be deleted.
- The identity management is a complex task and therefore provided as a service for cloud consumers.
- For example single sign on (SSO). Single sign on (SSO) software is installed over authentication server.
- Before connecting to application servers, the user connects with the authentication server to obtain a secure ticket.



- The authentication server maintains the user login security credentials required by application servers.
- When the user leaves the company, only the user's login on authentication server is needed to be disabled to block the user's access to all the application servers.
- There are a few examples of IDaaS providers for on-premises and cloud applications such as Ping IDaaS and PasswordBank IDaaS.

Module No – 072:IDaaS: OpenID:

- It is a popular example of Identity as a Service(IDaaS).
- Allows the users to sign-in to multiple websites by using a single account.

- Solves a lot of problems related to multiple log-in accounts per user.
- Why use OpenID:
    - Avoid too many user names and passwords.
    - Overcoming the scarcity of desired user names.
    - Account management is difficult otherwise.
    - Avoid filling long forms for creating logins again and again.
- OpenID is not controlled by any organization and/or person.
- There are a number of companies (providers) which provide OpenID accounts. These include: Google, Microsoft, Yahoo, Amazon, SalesForce etc.
- There are more than 1 billion OpenID accounts which are accepted by over 50,000 websites.
- How does it work:
    - A user creates an OpenID login through a suitable provider.
    - The user visits a website which is compatible with OpenID.
    - The (visited) site prompts the user to sign-in with the OpenID credentials.
    - The user is redirected to the OpenID provider's website.
    - The user opts to share the credentials/token with the (visited) website.
    - The user provides login and password at the OpenID provider's website.
    - If the user is verified, the OpenID provider confirms the (visited) website.
    - The user is redirected to the (visited) website which accepts the user as authenticated user.

Module No – 078: Identity as a Service (IDaaS):

- Collaboration is defined as the process in which two or more people work together to achieve a goal.
- Traditionally, the collaboration has been achieved through face to face meetings in conference rooms.
- Some team members had to travel (from near or far) to attend the meetings.
- Those who could not personally arrive at the meeting had either of the following two choices:
- Phone call to a speaker phone placed at the conference table
- Study the minutes of meeting
- A solution that could reduce the requirements of personal meetings was required to save time and effort and to increase the productivity from the collaborations.
- The web based collaboration began with the web mail.
- Users can compose, send, receive and read the emails by using the web browser and Internet connection.
- A single user can address multiple recipients in a single mail.
- (IM) provide a real time exchange of messages and replies (chat) by using messaging software.
- IM is another form of traditional collaboration. Current tools for IM allow file exchange and audio/video calling.
- Voice over Internet Protocol (VoIP) enables the users to make phone calls over the Internet.
- VoIP tools such as Skype provide a convenient way to perform conference calls by using computers and mobile phones.

Module No – 079: Cloud based Phone & Fax Systems:

- Sending and/or receiving fax traditionally required the fax machine and telephone connection.
- Similarly, phone calling has been dependent upon telephone infrastructure.
- In modern days, many companies have started providing cloud based calling and cloud based fax services.
- These companies have all the calling/fax operations performed over the Cloud and provisioned over the Internet.
- Taking example of Google Voice Phone System: The account holder receives the services of call answering and voice mail.
- The user can even configure the service to forward the incoming phone calls to a cell number.
- Google delivers the voice messages left by the callers as audio messages as well as in the form of text which are receivable anywhere through the Internet.
- Cloud based fax service provided by various companies is provisioned as a separate virtual number to each subscriber. This number corresponds o a virtual fa machine.
- The fax received over the virtual fax machine are delivered through email as PDF attachment.
- Similarly, to send a fax, a simple email (with PDF file) to virtual fax account will send the fax to recipient/s.

Module No – 080: Editing the Shared Files in Cloud:

- As we have seen that data and files can be stored on Cloud storage.
- It is also possible to edit the files (located on Cloud storage) shared among concurrent users.
- Provides another way of collaboration.
- A number of service providers offer the editing of shared files such as text, spreadsheet and presentation files. These include the famous providers:
    - Dropbox
    - Microsoft
    - Google
- Dropbox offers file sharing through public folders among the Dropbox users.
- It is allowed to edit the MSWord, Excel and PowerPoint files in browser and without the MSOffice installed.
- Simultaneous users can edit a shared document.
- Google provided *Google Docs* service offers web-based free access to a word processor, spreadsheet and presentation programs to create, share, edit, print and download the documents stored on Cloud.
- Google Docs can be shared through simple email link.

Module No – 081:Collaboration in the Cloud Through Collaborative Meetings:

- Collaborative meeting can be performed by using the software hosted on Cloud.
- Organizations get a cost effective *virtual meeting* as an alternative to face to face meetings.
- The features of cloud based collaborative meetings are:
    - Streaming video to allow face to face interaction
    - Shared whiteboards to control the presentation
    - Shared applications to demonstrate software in live environment
    - Meeting recordings for playback and sharing
- *GoToMeeting* is one of the leading providers of virtual meetings.
- Can support face to face meetings and web seminars (webinars) with more than 1000 attendees.
- The video recording of virtual meetings and webinars can also be used for virtual training and reference purposes as well.

Module No – 082: Collaboration by Social Media & Video Streaming:

- Social media and streaming video contents provide yet another way for collaboration.
- Cloud hosted social media such as Facebook and SalesForce.com's Chatter tool are available for collaboration among team members.
- The team member can easily exchange updates, comments and reviews regarding different tasks.
- Files can be shared among the team members.
- Photos and videos can be uploaded and shared to demonstrate a situation.
- Live video streaming can also be broadcasted if required.
- YouTube offers a free, reliable and Web accessed cloud storage for video contents worldwide.
- Videos created for collaboration can be shared among team members and publicly as well.
- The collaborative videos may include technical training clips, discussions and/or site coverage etc.
- The viewers can discuss and upload written comments on the video clip.

Lesson No. 16

CLOUD DEPLOYMENT MODELS

Module No – 083: Public Cloud:

- Public cloud is one of the deployment models of Cloud through which the IT resources are publicly available and accessible trough public Internet.
- Characteristics of Public Cloud according to NIST:
    - The consumer is generally not aware of the location of IT resources unless a location restriction is imposed by either of provider or consumer. Still it is difficult for the

consumer to verify the location on map from where the IT resources are being provisioned.
- The consumer workload may be a co-resident of the workload of other consumer (multi-tenancy) which may include the rivals, adversaries and in worst case, the attackers.
- The consumer has limited visibility of the software and procedures of the provider. The consumer has to trust the provider for securing the consumer's data and fully disposing the deleted data.
- The consumer undergoes a limited upfront cost regarding the provisioning of IT resources as compared to in house or locally setting up the IT infrastructure.
- Thanks to the workload management, dynamic collaboration among cloud providers and (generally) large setups, the public clouds can give the illusion of unlimited resources and elasticity to the consumers.
- The provider is in a limited legal Service Level Agreement (SLA) with the consumer. The SLA covers the minimum performance assurance/s by the provider and penalty in case of violation to the assurance/s.

Module No – 084: Private Cloud:

- Characteristics of Private Cloud according to NIST:
  - The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units).
  - It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.
  - The private cloud users depend upon the local area network if the cloud is locally deployed and accessed from a single site.
  - For multi-site access and outsourcing, the dedicated leased secure communication lines should be used.
  - Consumers are needed to be trained for working in Cloud environment.
  - Consumers have no knowledge of the location of their workload. Even in on-site deployment, a consumer can not pinpoint a server for the location of workload.
  - However, in case of outsourced Private Cloud, the consumer organization may have some knowledge of the cluster location and network segment serving the Private Cloud at the provider's end.
  - Consumer workload is vulnerable to cons of multi-tenancy from the insider malicious colleagues.
  - Modest cost for outsourced private Cloud (excludes infrastructure cost): Negotiation with the provider, Upgradation in network equipment, updating of legacy software to work on Cloud, training of staff etc.
  - Significant cost for onsite private Cloud (includes the data center and infrastructure cost): Updating of legacy software to work on Cloud, training of staff etc.
  - Resource limitation in on-site private Cloud but extendible resources available in case of outsourced private Cloud.

Module No – 085: Community Cloud:

- Characteristics of Community Cloud according to NIST:
  - The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations).
  - It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.
  - For the onsite Community Cloud, the resource sharing among the participating organizations has to be decided explicitly or implicitly.
  - At least one member of the community should provide Cloud services.
  - Network dependency: In case of on-site deployment, the network dependency is similar to on-site distributed Private Cloud setup. The performance and security can be enhanced through dedicated secured communication lines.
  - Network dependency: The members can also use encryption over Internet for the network access to the Community Cloud resources.
  - IT skills are required to manage the Community Cloud deployment and operations in both the participants (providing Cloud services) and consumer members of the community.
  - Workload locations are generally hidden from the community members unless a participant member decides to outsource the Cloud services (similar to outsourced Private Cloud). In this case, prior approval and documentation should take place.
  - Multi-tenancy cons are similar to onsite Private Cloud scenario.
  - The upfront cost for consumer-only member is same as of outsourced Private Cloud. While for participant members (onsite deployment), the upfront cost is similar to onsite Private Cloud.
  - The onsite deployment of Community cloud suffers from resource shortage as of onsite Private Cloud because each participant organization has limited resources.
  - Extensive resources are available for outsourced Community Cloud just like outsourced Private Cloud.
  - Due to a number of members, there are a number of security perimeters (hence complex cryptography) and dedicated communication lines in a Community Cloud. This offers a better security from external threats.

Module No – 086: Hybrid Cloud:

- Characteristics of Hybrid Cloud according to NIST:
  - The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public).
  - The hybrid cloud components infrastructures (private, community, or public) remain unique entities.
  - The hybrid cloud components infrastructures (private, community, or public) are bound together by standardized or proprietary technology that enables data and application portability (for load balancing between clouds).

- Hybrid Clouds are often possible when the phenomenon of *Cloud Bursting* is applied whereby a consumer uses a private cloud in routine but may use the services of other types of clouds for load balancing at peak times.
- Hybrid Clouds are also formed when one type of cloud is used to provide backup to another type of cloud.
- An organization may choose to process sensitive data on outsourced private-cloud but choose new software testing on a public cloud.
- It may be cost effective to put the web requests handling for web applications on a PaaS instance while the background processing of those web applications can be done on on-site community cloud.
- Challenges for hybrid clouds exist in security management, identity management and access control of multiple clouds etc.
- More complex scenario arises when the clouds are dynamically joining and exiting the hybrid cloud.
- Network dependence
- IT skills required
- Workload locations are hidden from consumer
- Security risks due to multi-tenancy

Lesson No. 17

SERVICE ORIENTED ARCHITECTURE

Module No – 087: Web Applications & Multitenant Technology:

- Web Applications: These are the applications which use web technologies (URL, HTTP, HTML, XML) and generally use web browser based interface.
- Can be modeled on the basis of three-tier model.
  - Presentation layer
  - Application layer
  - Data layer
- Web Application Architecture 1:

<table>
<tr><th>Layer</th><th colspan="2">Implementation</th></tr>
<tr><td></td><td>Server side</td><td>Client side</td></tr>
<tr><td>Presentation</td><td rowspan="2">Web/
Application Server</td><td>Web client</td></tr>
<tr><td>Application</td><td rowspan="2"></td></tr>
<tr><td>Data</td><td>Data storage server</td></tr>
</table>

- Web Application Architecture 2:

Layer	Implementation	
	Server side	Client side
Presentation	Web server	Web client
Application	Application server	
Data	Data storage server	

- Multi-tenant Technology: The multi-tenant applications allow isolated to simultaneous users (tenants).
  - The data and configuration of each user remains private to other users.
  - The tenants can customize the user interface, business process, data model and access control of the multi-tenant application.
- Common Characteristics of Multi-tenant Applications:
  - Usage isolation
  - Data security
  - Backup and restore is separate for each tenant
  - Application upgrades do not negatively effect the existing users
  - Scalability in terms of number of tenants
  - Metered usage
  - Databases, tables and/or schema isolation for each user

Module No – 088: Service Oriented Architecture

- Web Services are independent units of software (code) which allow network based machine-to-machine interaction.
  - Have no user interface.
  - Process data between the computers through API calls.
  - Examples: SOAP and REST based web services
- Service oriented architecture (SOA) is usually a collection of services (web services)
- These services communicate with each other for the exchange of data and processing.
- Two or more services may be coordinating an activity.

- Examples of web services:
    - Return the weather conditions for a specific zip code
    - Return real-time traffic conditions doe a road or highway
    - Return a stock price for a particular company
- Web services are not web pages.
- To use a web service (which resides on a remote server), a program exchanges messages with the service.
- The user program sends parameters (through API call) such as zip code to the web service and waits for the reply.
- Web services are treated as black box by the programmer.
- Web services are interoperable which means that programs written in dissimilar language/s than the web-based service can call the API functions.
- Web Services: The core technologies are:
    - Web Service Description Language (WSDL): A markup language to define the API of the web service including the functions and the input/output messages associated with each function.
    - Message input/output are in the form of XML and defined by XML schema.
    - The message formatting is according to a common messaging format defined by Simple Object Access Protocol (SOAP) or through Representation State Transfer (REST).
    - Universal Description, Discovery and Integration (UDDI) is a standard which regulates the service registries in which WSDL definitions can be published so that they can be discovered by the users.
- Cloud Service & Web Services:
    - These two are not alike.
    - Can be used independent of each other in a SOA.
    - Cloud services are SaaS, PaaS & IaaS
    - Web services are API Calls.
    - Web services can be the front door for the cloud services running at the backend.
    - Cloud services are often provided over web services.
    - For example Amazon Web Service (AWS) based cloud services (e.g., data processing service deployed by a provider) can be accessed over network through API developed (by the same provider) using Amazon API Gateway.

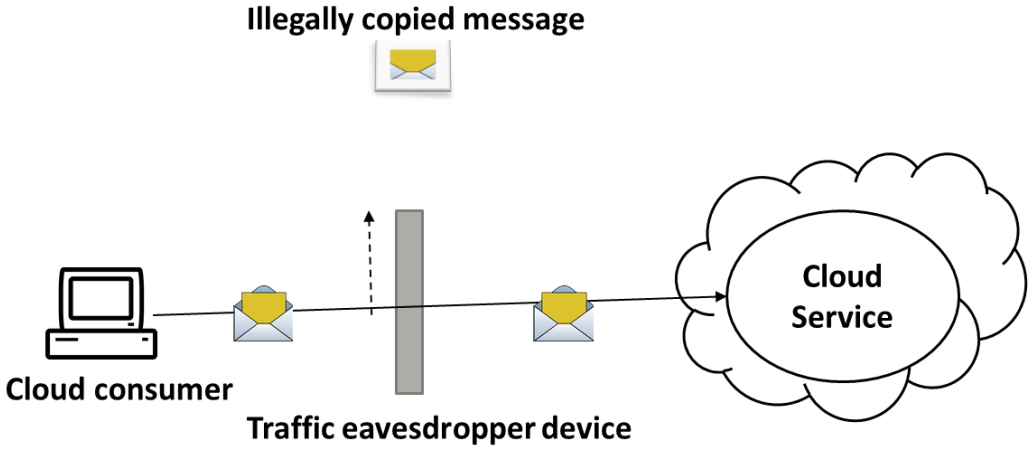
Lesson No. 18

CLOUD SECURITY THREATS

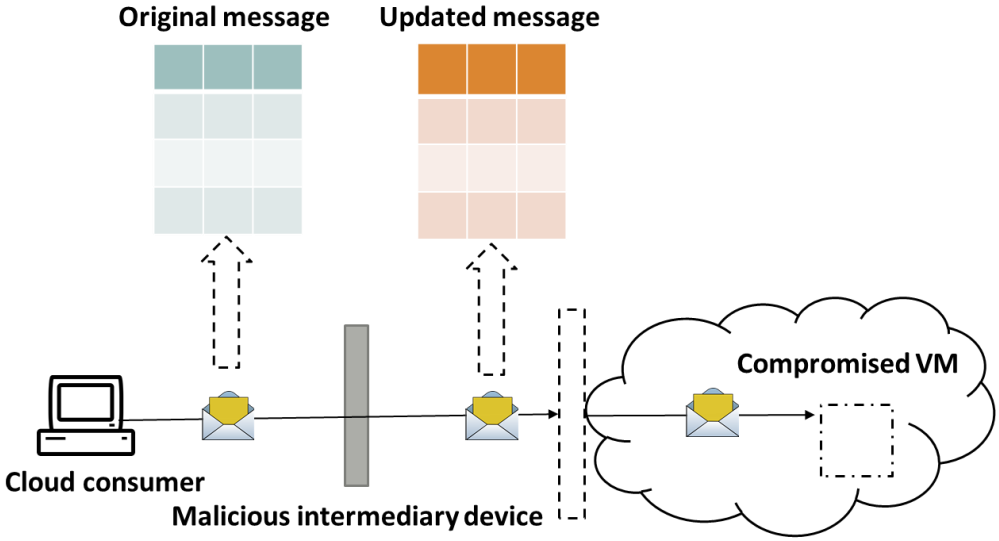
Module No – 089:

- This module is about the prominent security threats to the Cloud computing.
- The following are significant threats to Cloud Security:

- Traffic Eavesdropping: This module is about the prominent security threats to the Cloud computing. Compromises the message contents. Can go undetected for extended periods of time.



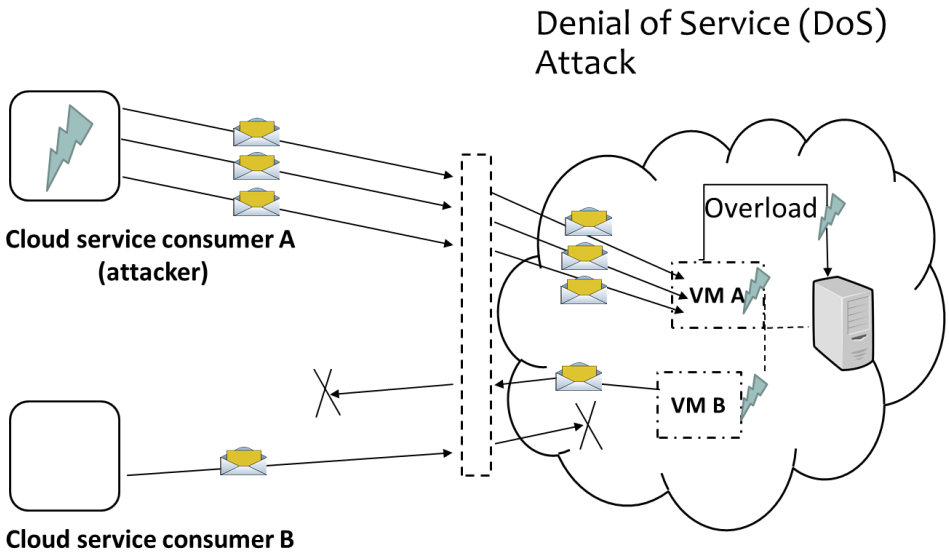
- Malicious Intermediately: The messages are illegally intercepted and then the contents are updated. The updated message is then relayed towards the cloud. The messages are illegally intercepted and then the contents are updated. The updated message is then relayed towards the cloud. The message may be updated with malicious contents which reach the VM hosting the cloud service undetected.



Module No – 090: Cloud Security Threats:

- …continued
  - Denial of Service (DoS): The purpose is to overload the IT resources so the sage where they can not work properly. Can be launched in the following ways: Workload on a cloud service is artificially increased through fake messages or repeated

communication requests. Network is overloaded with traffic to cripple the performance and increasing the response time. Multiple cloud service requests are sent. Each request is designed to consume excessive memory and processing resources.



- Insufficient Authorization based attack: It is a situation when a malicious user gets direct access to IT resources which are supposed to be accessed by trusted users only. Happens when a broad access is provided to the IT resources and/or due to erroneously.
- Weak authentication based attacks: Happen when weak passwords or shared (login) accounts are used to protect the IT resources. The impact of attacks due to insufficient authorization and weak authentication depends upon the range of IT resources and the range of access to those IT resources is compromised.

Module No – 091:Cloud Security Threats:

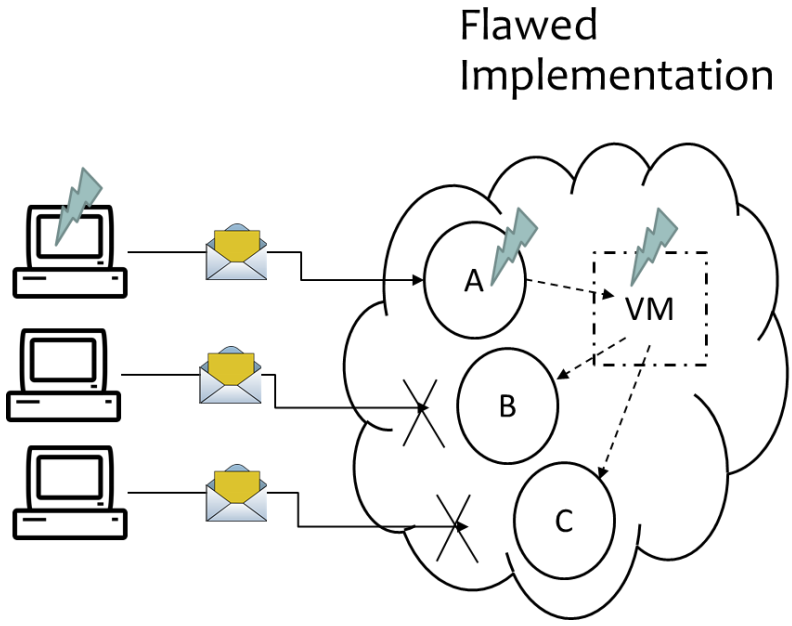
- …continued
  - Virtualization Attack: Based upon the administrative privileges provided to the Cloud consumers and multi-tenancy, it is possible to compromise the underlying physical hardware. It is also possible that the security flaws be arising due to *VM sprawl* (a lack of security patches on OS installed on VM). Another possibility is the installation of VM-aware malware to exploit the security flaws of hypervisor. Following are possible sources in which the physical server may be compromised:
    - By an imposter in disguise of a legitimate consumer. The attacker cracks the (weak) password of a consumer.
    - By a trusted but malicious consumer.
    - In either case, the vulnerabilities in the virtualization platform are exploited over a single VM to take control of the physical server hosting the infected VM. Makes all the VMs hosted on the compromised server as vulnerable.

- A more severe scenario arises when the infected VM is migrated to other server for load balancing. In this case, a number of servers may get compromised.

- Overlapping Trust Boundaries: Moving of consumer data to Cloud means that the provider now shares (with the consumer) the responsibilities of availability, confidentiality and integrity of data. The consumer thus extends the trust boundary to include the cloud provider. This is prone to vulnerabilities. When multiple consumers of a cloud share an IT resource, the trust boundaries overlap. The provider may not be able to provider the security features that can satisfy the security requirement of all the consumers of shared IT resource on a Cloud. More complex scenarios arise when the consumer data is replicated and stored on multiple sites.
- Another complexity arises when the Cloud provider handover the business to a new owner. The data integrity becomes threatened in both cases.

Module No – 092:Cloud Security Threats:

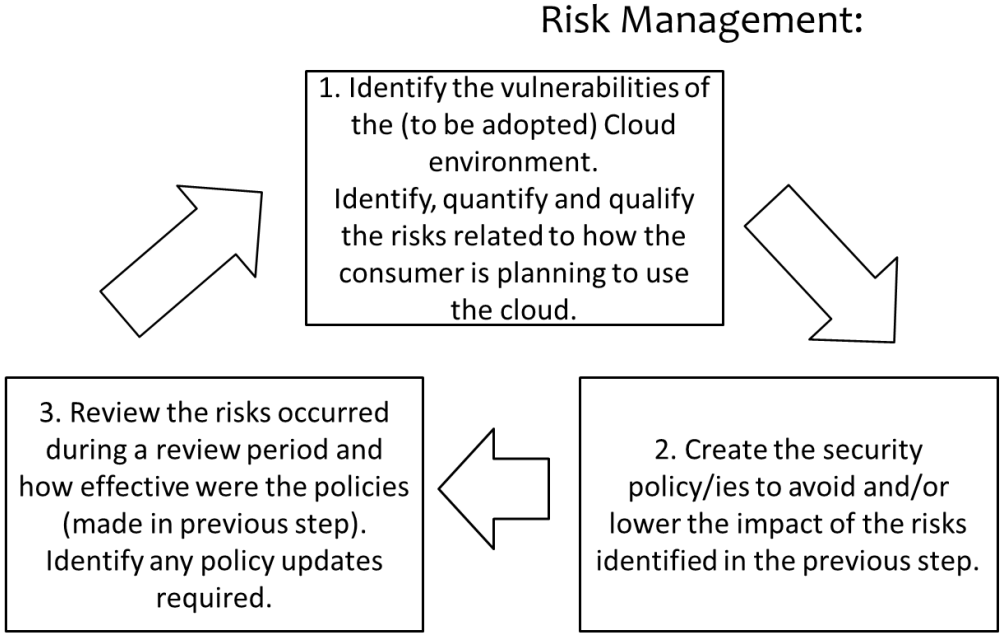
- …continued.
  - Flawed Implementation: The implementation of Cloud services may have some flaws related to configuration resulting into the occurring of unexpected events. Particularly the security and operational weaknesses in Cloud provider's software/hardware can be targeted by the attackers to put the integrity, confidentiality and/or availability of IT resources of the provider at stake. Equally important point is the implementation flaws of Cloud services may result in the crash of VM and thus will effect all the other services on that VM as well. For example service A has some implementational flaws to crash the hosting VM when a certain message is sent. This will also effect the services B and C and can be exploited by an attacker.



- Disparity of Computer Security Policy : A computer security policy defines the set of rules and mechanisms to ensure the security of the computers of the organization. The computer security policies of the consumer and provider may not match. Before opting of outsourcing and/or public cloud, an organization must evaluate the compatibility of provider's security policy with its own. The lack of administrative privileges provided to the consumer makes the implementation of the consumer chosen computer security policy very difficult. Due to the discussed points, the standardization of securing the IT resources leased by a consumer and the consumer data is a challenging task.

Module No – 093: Cloud Security Threats:

- …continued.
  - Contracts: As an additional consideration, the SLA offered by the provider should be carefully examined to clarify the liabilities taken by the provider and the security policy implemented by the provider. This helps in determining the following:
    - If the consumer deploys its own solution over the Cloud resources then it is a situation of consumer's assets deployed over provider's assets. Then how the blame will be determined when a security breach or a runtime failure occurs ?
    - If the consumer can apply its own security policies while the cloud provider keeps the administrative rights to the IT infrastructure. Then how this disparity will be overcome.
  - Risk Management: The cloud consumers should perform a cyclic process of risk management to access the potential threats and challenges related to Cloud adoption. This should be a part of risk management strategy. It is a three stage process.



Lesson No. 19

TRUST ISSUES IN CLOUD

Module No – 094: Brief overview (more in Lesson 39):

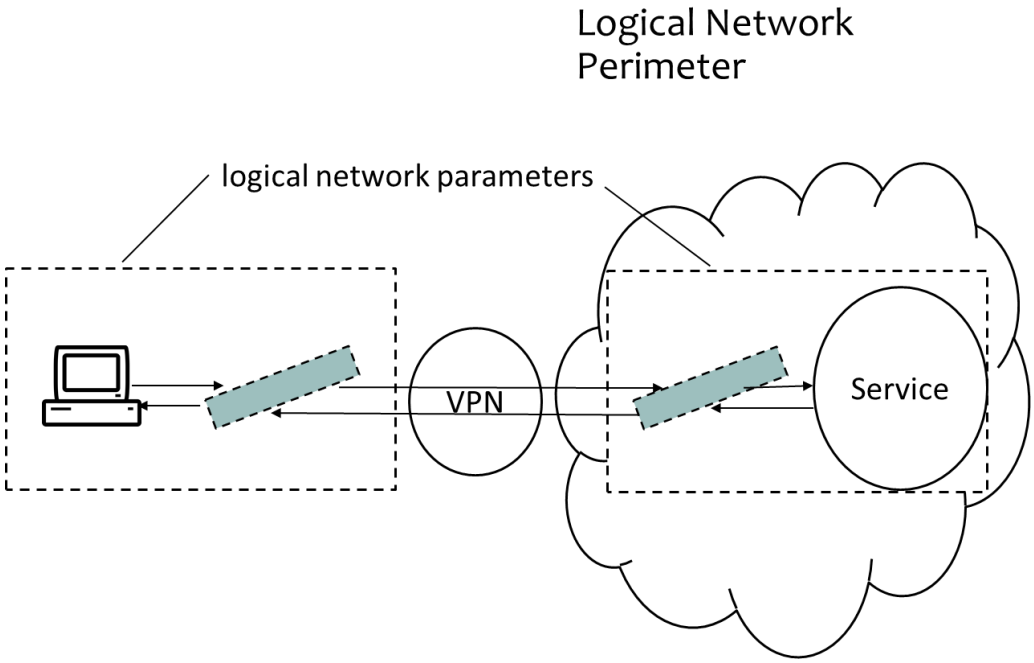
- Link between Privacy, Security and Trust:
  - Privacy: The confidentiality of data related to a person or organization.
  - Security: The preservation of confidentiality, integrity and availability of data.
  - Trust: The state of accepting a vulnerability on the base of positive expectations.
- Privacy issues of Cloud Computing:
  - Lack of user control
  - Lack of training and expertise
  - Possibility of secondary (/unauthorize) use of consumer data
  - Legal compliance
- Security issues of Cloud Computing:
  - Overlapping security boundaries
  - Unauthorized access
  - Lack of interoperability of security policies
  - Uncertainty of data deletion
  - Compromise of management console
  - Backup vulnerabilities
  - Isolation failure in multi-tenant applications
  - Inadequate monitoring and audit
- Trust in Cloud: The consumer's trust in Cloud is affected by the privacy and security vulnerabilities of Cloud as discussed before.
  - Further, due to lack of transparency the blame of responsibility is difficult to be placed if the provider is outsourcing the IT resources from a chain of outsourcing.
  - The pay-as-you-go and *on-demand* provision of cloud resources may be subject to low level of trust.
  - The lack of trust is the key factor for user reluctance to use Cloud services.
  - Consumer feels a *lack of control* in shifting to Cloud.
  - The companies shifting from on-premises setups to public Clouds are more concerned about data security and health than of the servers.
  - Concerns are present regarding foreign governments' access to consumers' data on Cloud.
  - The analysis of tradeoffs of Cloud privacy, security, cost and benefits determines the decision of Cloud usage.
- Conclusion: The consumers' trust can be assured through the safeguarding of personal/confidential/sensitive data. The existence0enhancement of transparency and accountability can increase the trust. Research should be conducted to quantify and model the trust and trust management, so that approaches for strengthening the consumers' trust can be proposed, tested, and/or enhanced.

Lesson No. 20

MECHANISMS RELATED TO CLOUD INFRASTRUCTURE

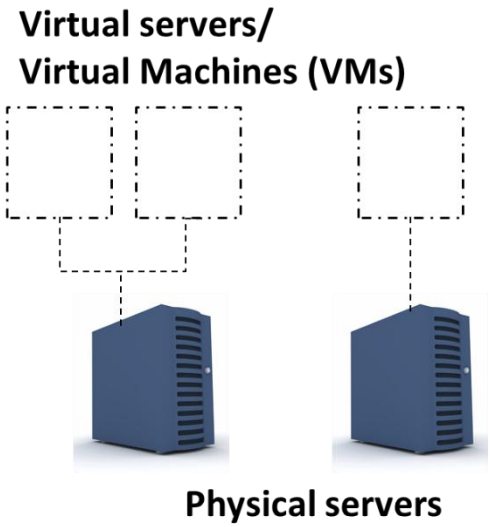
Module No – 095: Logical Network Perimeter:

- It establishes the boundary of virtual network to hold with in and isolate a set of related cloud-IT resources that may be distributed physically. Implemented as virtual environment, it has the following components:
  - *Virtual Firewall* to filter the traffic of isolated network to and from Internet.
  - *Virtual Network* consisting of virtual nodes and virtual links.

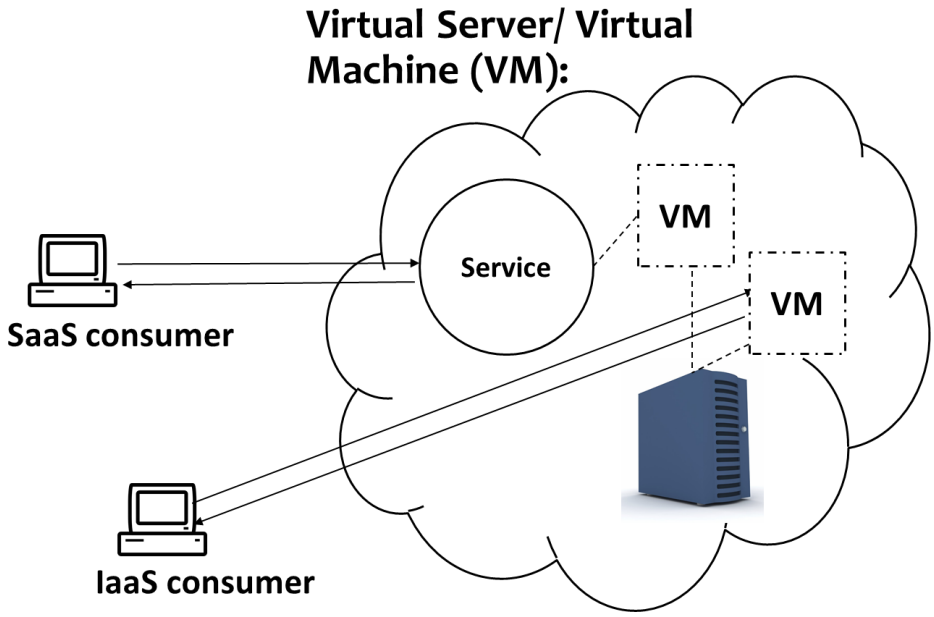


Module No – 096: Virtual Server or Virtual Machine (VM):

- Virtual Server: Virtual servers or Virtual Machines (VMs) emulate the physical servers.
  - Each virtual server can host numerous IT resources, cloud-based solutions and other cloud computing mechanisms. Depending upon the capacity, a physical server may host multiple virtual servers.



Virtual Servers or Virtual Machines(VMs)



Cloud service consumers, Cloud service and VM relationship

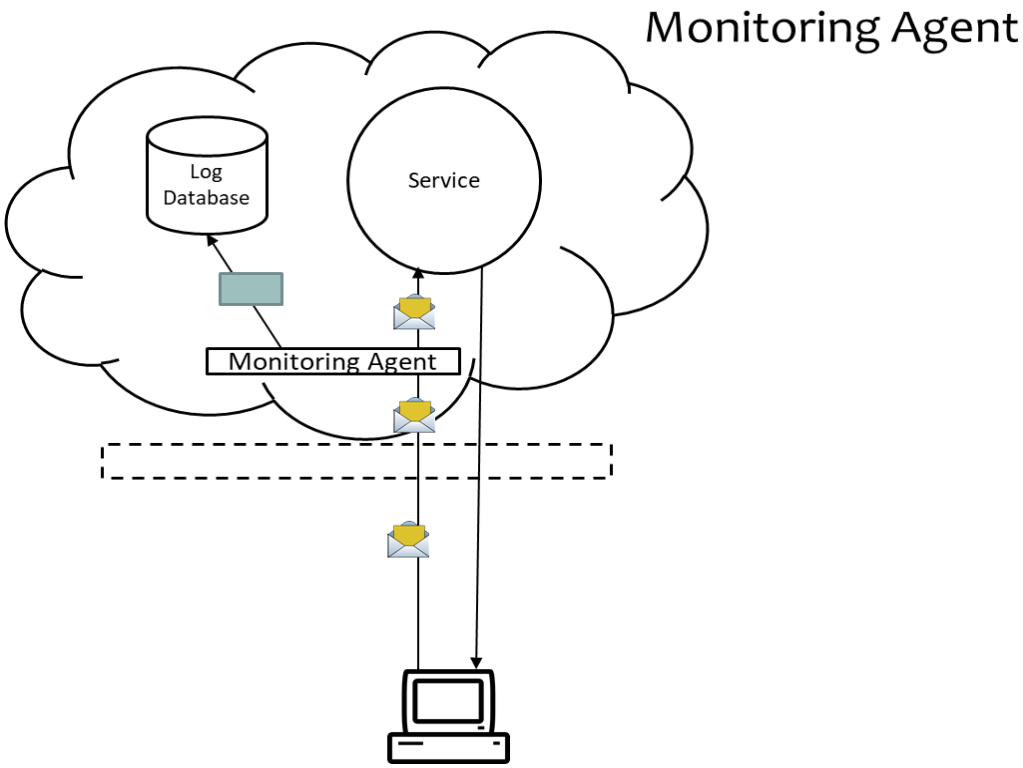
- In order to rapidly provision the VMs with installed and preconfigured software such as OS, programming platforms etc., the virtual servers are *cloned* by templates.
- A *template* is a master copy of virtual server. It contains the configuration, installed software, any configured virtual devices and disk contents.
- A consumer can:
  - Connect to a self-service portal of Cloud provider.
  - Choose a suitable template.
  - Instantiate a virtual server through administrative portal which works with the help of virtual infrastructure manager (VIM) module.
  - Customize the virtual server through usage and administrative portal.

Module No – 097: Cloud Storage Device:

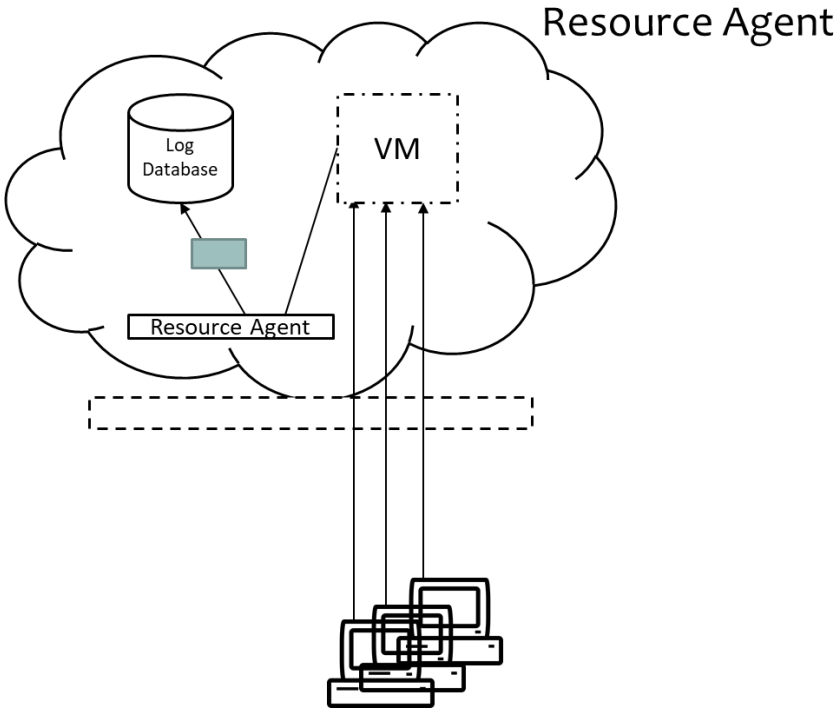
- It represents the storage mechanisms devised specifically for cloud-based provisioning.
    - Instances of these devices can be virtualized.
    - Support dynamic scaling
    - Can be accessed remotely by Cloud storage services.
    - The cloud storage mechanisms support the following (but not limited to) logical units of data storage:
        - Files (data grouped into files that are located in folders)
        - Blocks (the smallest unit of data that is individually accessible)
        - Datasets (such as data arranged in databases)
        - Objects (data and associated meta data)
            - Each of these levels is associated with a certain type of technical interface consisting of a specific type of cloud storage device with a Cloud storage service used to use its API.
        - Network Storage Interface: For file and block storage
        - Object Storage Interface: Based upon technologies that support a range of data and media types. The storage mechanism can be accessed by REST or SOAP based web services.
        - Database Storage Interface: Supports the relational (SQL based) and non-relational databases (NoSQL storage).
        - Database Storage Interface: Data stored in relational database is more structured and normalized than non-relational database. The relational databases have higher processing overhead. While the non-relational have high data-redundancy. Also, transactions and joins are not supported. The relational databases have higher processing overhead than non-relational database. The non-relational databases storage have high data-redundancy and non-structured data. The relational-database functions such as transactions and joins are not supported in non-relational database storage.

Module No – 098: Cloud Usage Monitor

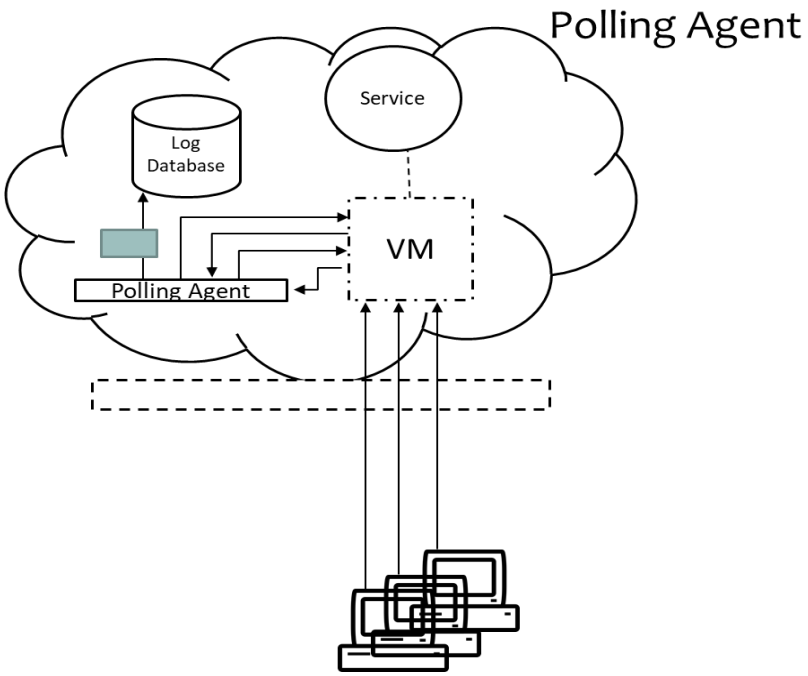
- It is a software used to collect and process the data related to Cloud-based IT resources.
    - The reporting and analysis requirements of the Cloud usage module determines the scope and volume of data collected/extracted.
- There are a few generic types or formats of Cloud usage monitors:
    - Monitoring Agent: It transparently monitors and analyzes the dataflow over communication paths. It measures the network traffic and messages.



- Resource Agent: Collects the resource usage data related to certain events such as initiating, suspending, resuming and vertical scaling. It interacts with the Cloud resource management module.



- Polling Agent: Collects the Cloud service usage data after periodic polling to IT resources. For example the uptime/downtime of a Cloud service. Records the updated status of the resource.



Module No – 099:Resource Replication

- It is a technique by which multiple copies of the IT resources are created to increase the availability and productivity of the IT resources. Virtualization technology is used for Cloud IT resources' replication.
- For example, due to a physical server failure and in order to over come the resultant downtime of a Cloud service deployed over a VM hosted by that physical server, the entire VM along with the software (Cloud service implementation) is replicated to another server.
- Another example is the horizontal scaling of IT resources such as increasing or decreasing of Cloud service instances by replication of VM hosting the service instance, corresponding to workload.
- The resource replication process yields the IT resources which are monitored under the Cloud usage monitor mechanism.
- Resource replication is also essential for pay-as-you-go type of usage & billing.

Module No – 100:Ready-Made Environment

- This mechanism represents the provisioning of preconfigure PaaS instances with ready to use and customizable programming environments. Provide the dependable PaaS instances.
- Time efficient provisioning

- Typically include:
  - Software development tools
  - Databases
  - Middleware
  - Governance tools
- The middleware is provided to support multi-tenant platforms to develop and deploy the complementary web services for SaaS scenarios.
- Overall, the ready-made environment mechanism supports the development and production level deployment of Cloud services.

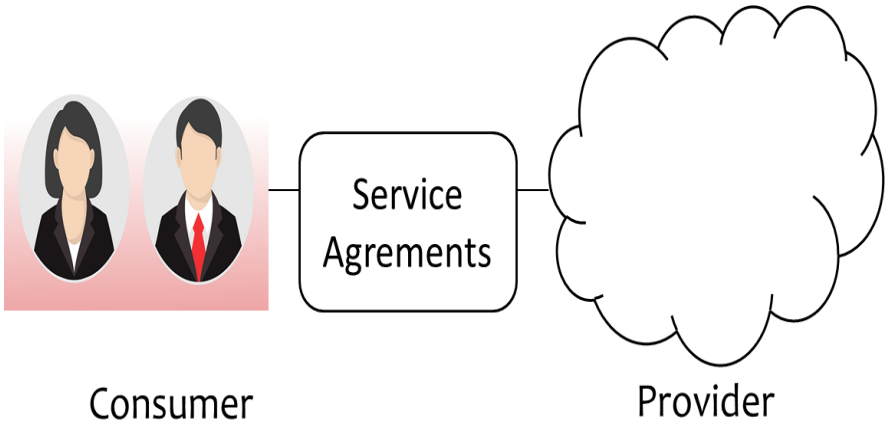
Lesson No. 21

SERVICE AGREMENTS(SAs)

Module No – 101:

- NIST identifies that the consumer and provider are under a legal agreement or terms of service.
- The agreement has two parts:
  - Service Agreement
  - Service Level Agreement (SLA)
- Service agreement contains the legal terms of contract.
- The SLA contains the technical performance promises by the provider and the remedies for performance failures.
- Over all called *Service Agreements* by NIST
- The following promises are made to consumer by the provides:
  - Availability:
    - Usually 99.5% to 100% availability is assured.
    - The assurance is for a time intervals of a billing cycle e.g., 15 minute, 1 hour, 1 Year etc. for which the service status will be "up" for sure.
    - But this has to be clarified that for example time period of assurance is 15 minutes and even if the service is "down" for 14 minutes, then it legally means that the service was not "down" for the whole interval.
    - Typically, several failures in subsystems are required to completely "down" a service for the whole period of billing.
    - The provider may adjust the availability promises on case to case basis.
  - Remedies for Failure to Perform:
    - In case of violation of the promise of *availability (*during a time period) by the provider, the customer will be compensated in terms of service credit for future use of Cloud service.
    - A refund is usually not given.
    - Consumer is responsible to monitor the availability of service and claim for compensation.
    - The following situations result in termination of Cloud IT resources usage for a consumer:

- Voluntarily by consumer
- Terminated by the provider for violating the provider's rule of service and/or for non-payment.
    - The providers usually take no responsibility for preserving the data in later case. While in former case, the preservation is done for few days.



    - Legal Care of Consumer Information:
        - The provider assures for not disclosing/viewing/using/sharing the consumer's data except in case of legal requirement.
        - On the other hand the provider retains the right of monitoring the consumer data as well as may demand a copy of consumer's software for monitoring assistance.
- The following limitations are included in the policies by the provider:
    - Scheduled Outages:
        - Will not be considered as service failure.
        - Will be informed in advance.
        - Will be of a limited time period.
    - Force majeure events:
        - Providers do take the responsibility for the events out of their realistic boundary. Such as:
        - Power failure, natural disaster and unreliable connectivity between consumer and cloud service.
    - Service Agreement Changes:
        - The provider usually retain the right to change the terms of contract, billing amount etc. on limited notice.
        - Consumers should keep a regular check for updated service charges
        - Sometimes the provider inform a specific consumer by email or postage.
        - The changes may take effect immediately or after few weeks.
    - Security:

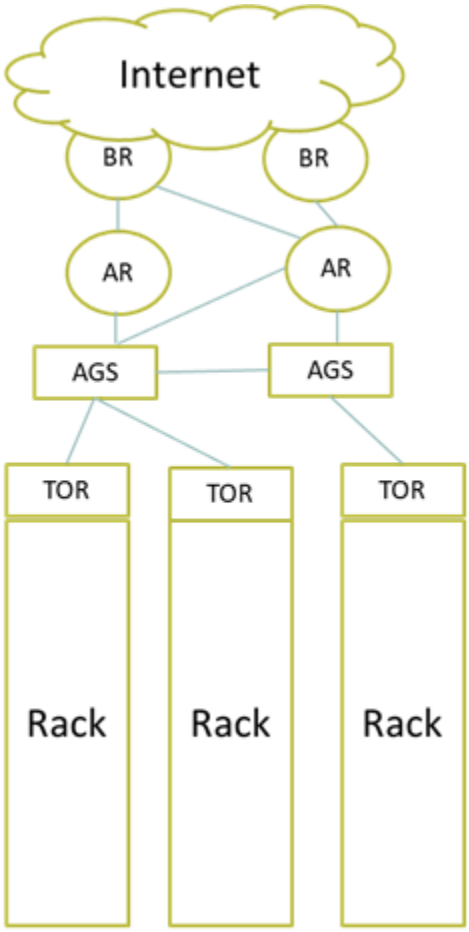
        - The providers do not take liability of data loss, data corruption or unauthorized data usage if they happen due to security breach or due to service interruption caused by a malicious activity.
        - At most, the service credit is compensated in case of data loss.
        - Although the providers promises for best effort security but the responsibility of data security is placed on the consumer.
        - It is difficult for the customer to determine the cause of data loss (malicious activity or some other reason).
    - Service API Changes:
        - The providers generally retain the right to delete or update the service API.
        - Can happen any time and without prior notice.
- Generally the consumer has to agree upon the following **obligations**:
    - Acceptable Use Polices: The consumers are generally required to refrain from:
        - Storing illegal data
        - Conducting security attacks on Cloud infrastructure and/or on any other user.
    - Licensed Software: The provider require the consumer to install and use only the licensed third party software over the Cloud.
    - Timely Payments: The consumer should timely pay the bill from the provider. Otherwise the consumer may get terminated after some time.
- **Recommendations** by NIST:
    - The consumers should carefully study and negotiate the service agreements. Specially take care of the SLA assurances and responsibilities by the provider.
    - Choose the most suitable Cloud provider periodically after review.
    -

Lesson No. 22

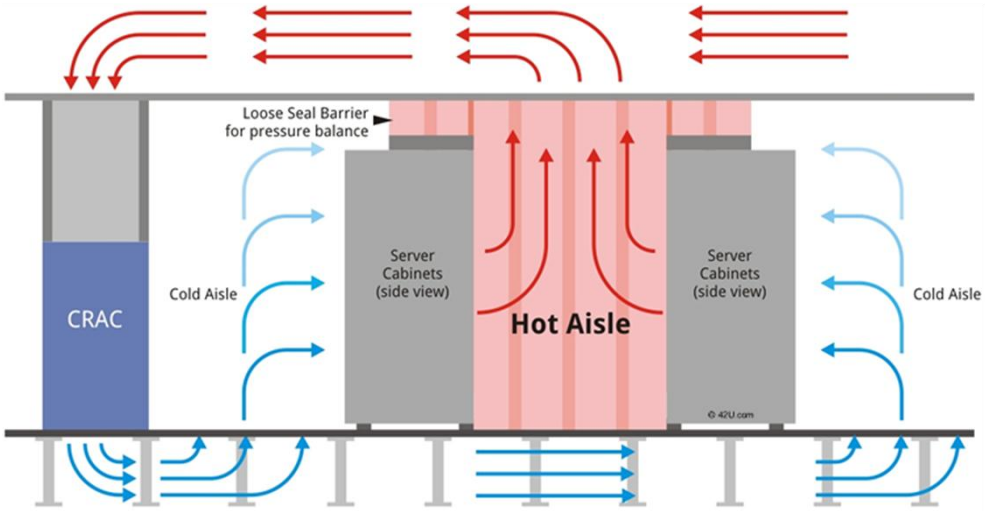
CLOUD HOSTING DATA CENTER DESIGN

Module No – 102:

- Key terms:
    - CRAC: Computer Room Air Conditioning
    - Hot aisle
    - Cold aisle
    - Server cabinets (Racks)
    - Hollow floor
    - Perforated tiles
- Cloud hosting data center has a layered architecture for the Internet access.
- The servers are physically connected to layer 2 switches. There is a top of rack (TOR) in each rack. One server is connected to only one TOR switch.
- The TORs are connected to aggregate switches (AGS).



Cloud hosting data center design



Cloud hosting data center design

- Data centers consume huge amounts of electricity. As much as a small town in USA.
- A large data center can host hundreds of thousands physical servers.
- It is more costly to setup and run a small data center in terms of unit costs (per server, per MB of storage, per GHz, Network bandwidth) and operational costs as compared to larger data centers.
- Google has 900,000 physical servers around the world in its data centers. Together these servers consume 260 million watts of power which accounts to 0.01% of global energy usage.
- Facebook data center servers process 2.4 billion pieces of content and 750TB of data every day.

Module No – 103:Data center Interconnection Networks

- The network connecting the data center servers is called *data center interconnection network*.
- It is a core design of data center.
- The network design must support the following features:
  - Low latency
  - High bandwidth
  - Low cost
  - Message-passing interface (MPI) communication support
  - Fault tolerance
  - Must satisfy both point-to-point and collective communication patterns among all server nodes.

- Application Traffic Support: The data center interconnection network must support the MPI communication and high bandwidth.
    - Example: Distributed file access, Map and Reduce functions etc.
    - Some servers can be configured to be master and others be slaves.
- Network Expandability: The interconnection network must be expandable.
    - Should support load balancing and data movement.
        - No bottlenecks
        - Can be expanded in the unit of *data center container* which contains hundreds of servers and is a building block of large data centers.
        -
- Fault Tolerance and Graceful Degradation: Can be implemented through:
    - Replication in software and hardware resources
    - Redundant links among any two servers
    - No single point of failure or critical links
    - Two layered design should be used (a network layer close to servers and the upper layer or backbone) to support modular (container) based expandable design.

Module No – 104: Modular Data center and Interconnection

- Modular Data Center in Shipping Containers: The modern data centers are a the collection of container based clusters that can be shipped from one location to another through trucks.
- It is an alternative to warehouse based data center.
- Modular Data Center in Shipping Containers:
- For example: The SGI ICE Cube container can house 46,080 processing cores or 30 PB of storage per container.
- Modular Data Center in Shipping Containers:
    - Such a design:
        - Is more energy efficient in terms of cooling cost as compared to warehouse based design.
        - Is more mobile and easily transportable.
        - Is ready to be deployed since it is assembled with servers, networking, power supplies and cooling mechanisms. It is then tested and shipped.
        - Helps in dynamic scalability of data center.
        - Makes the relocation of data center as relatively easier than warehouse based design.
        - Inter-Module Connection Networking requires an extra layer over modular containers to allow dynamic scaling and interconnection.

Module No – 105: Data center Management Issues

- Modern day data centers handle ever larger volumes of data and conduct the processing massive amounts of user requests around the globe.
- In order to maintain user satisfaction and performance, the managing of a data center has become a set of complex tasks. These include (but not limited to):

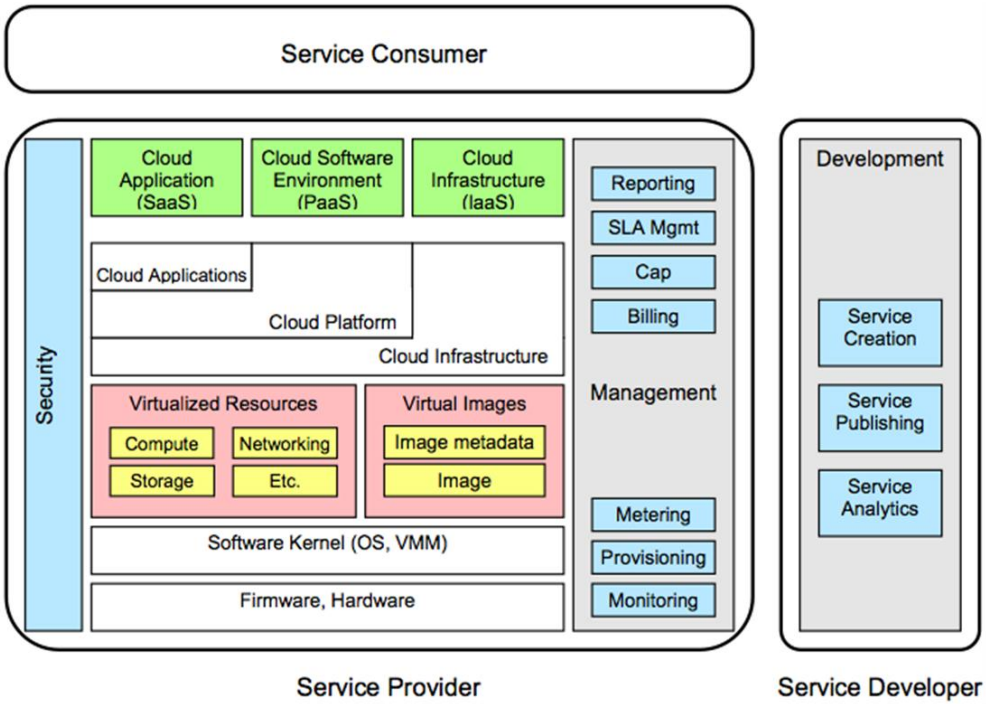
- Making common users happy by providing quality services.
- Ensuring uninterrupted and high availability of (Cloud) services.
- Managing multiple modules concurrently. Such as processing, networking, security and maintenance etc.
- Managing and planning for the scalability of data center.
- Ensuring the reliability of virtual infrastructure through fault tolerant and recovery mechanism to minimize the downtime and data loss.
- Managing and lowering the operational costs and transferring the cost benefit to Cloud providers and consumers.
- Security enforcement and data protection
- Implementation of Green information technology usage to lower the amount of energy consumption.

Lesson No. 23

CLOUD ARCHITECTURE

Module No – 106: Generic Cloud Architecture Considerations:

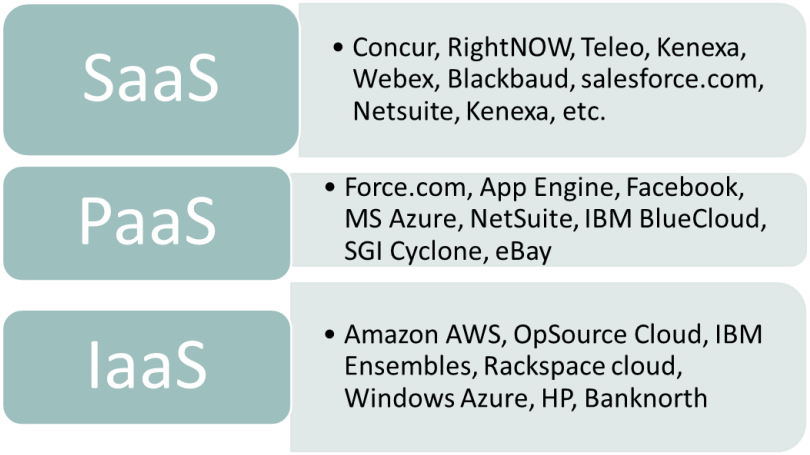
- A generic architecture of a (public) Cloud can be envisioned on the basis of technologies we have studied so far.
- Major goals of a Cloud platform can be:
    - Scalability
    - Virtualization
    - Efficiency
    - Reliability
- A Cloud management software receives the consumers' requests for IT resources and provisions these resource by using various internal services.
- A Cloud architecture has to deal with certain challenges. A few of them are:
    - Establishment of large scale computing (hardware + software) infrastructure.
    - User friendly and efficient management of Cloud infrastructure.
- Ensuring scalability of IT resources.
- Reliable and fault tolerant implementation for processing and data.
- Implementation of disaster recovery mechanisms.
- Cloud architecture should be expandable by adding more hardware.
- Software, hardware and network technologies have emerged as Cloud enabling technologies.
- Enhancement in the following technologies have contributed towards wide spread establishment of Cloud computing:
- Software: Virtualization, multi-tenancy, web applications, SOA, load balancing, monitoring, billing, data storage
- Hardware: CPU, memory, storage, network
- Connectivity: Web2.0



Generic Cloud Architecture

Module No – 107: Layered Cloud Architecture:

- Cloud architecture can be considered as consisting of layers and sub-layers of Services with each layer supporting the upper layer.
- In order of dependency, these layers are grouped at high level as:
  - SaaS
  - PaaS
  - IaaS
- Software Service development and deployment requires a platform service.
- A platform service is deployed over a VM provisioned through IaaS.
- Some services may draw resources from multiple layers/sub-layers.
- The scope of support from vendor side is highest for SaaS and lowest for IaaS.
- Services developed on PaaS require the later to provide support for scalability, security, and must be dependable.



Layered Cloud Architecture

- Unless there is interoperability among the Clouds, a Service deployed on a certain platform instance may not be portable to another platform.

Module No – 108: Virtualization Support and Disaster Recovery:

- The IT resources and data are prone to disasters (natural and/or human made) which damage them partially or fully and thus may crash the whole computing system of an organization.
- Key terms:
  - Failover: It is process through which a system transfers control (usually automatedly)to an alternate deployment upon failure of primary deployment.
  - Failback: The process of restoring of the system from alternative to primary deployment and restoration of original state.
  - The use of virtualization can implement the failover and brings reduction in failback time.
  - As compared to (for example) a data disaster for data stored on magnetic tapes, days are require for restoration/recovery.
  - The redundant deployment of software solutions, data and IT resources is quite easy by using virtualization.
  - One deployment is considered as primary, while other deployment/s are kept as backup.
  - The backup deployment is either updated periodically or the image/snapshot of the primary deployment (e.g., VMs) can be saved.
  - Upon failure, the backup deployment takes over.
  - The primary deployment is then restored from the most recent snapshot.
  - Virtualization has become the core part of disaster recovery plans of major organizations since last decade.
  - Virtualization even allows the testing of disaster recovery plan through emulation and without disturbing the production/primary deployment.
  - Although the failed physical servers have to be re-purchased/repaired, but the virtualization lowers the additional costs and time related to failback.

- The organizations should mark the critical applications and data and use replication of data in virtualized environments to support effective disaster recovery.

Module No – 109: Cloud Architectural Design Challenges:

- **Challenge 1: Service availability and Data Lock-in Problem**:
  - Depending upon a single provider for service deployment results in a single point of failure or lock-in.
- **Challenge 1: Service availability and Data Lock-in Problem**:
  - High availability of a service can be assured by distributed deployment over multiple Clouds.
  - Requires the interoperability/standardization of API calls on different PaaS platforms.
- **Challenge 2: Data Privacy and Security Concerns:**
  - Due to public access of Clouds, multitenancy and sophisticated attacks/malware, the implementation and assurance of privacy and security of consumers' data is a big challenge.
- **Challenge 3: Unpredictable Performance and Bottlenecks:**
  - The unpredictability of processing and data load over Cloud services introduce I/O bottlenecks such as concurrent read/write access requirements to shared storage for large data volumes by multiple VMs.
  - The providers have to carefully analyze the deployment decisions according to surge in computing/data loads and should tune the bottlenecks.

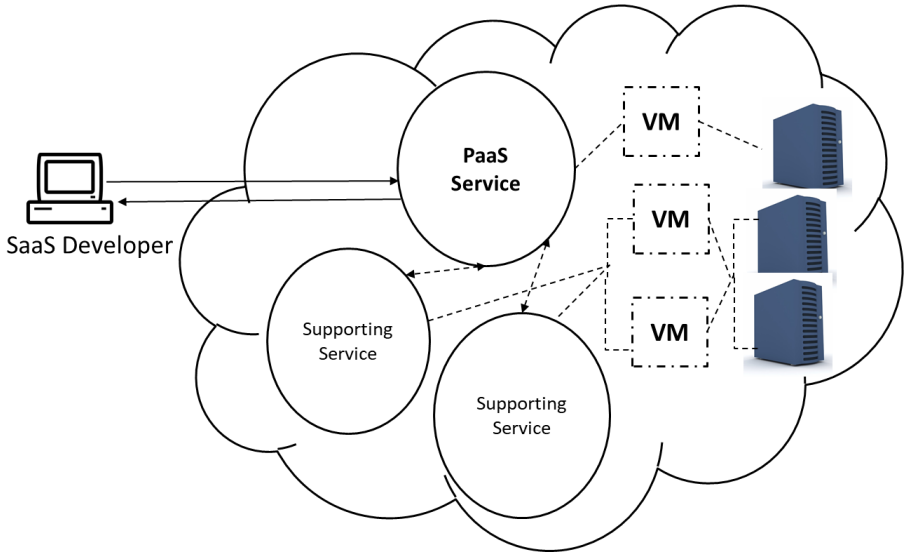
Module No – 110: Cloud Architectural Design Challenges (continued):

- **Challenge 4: Distributed Storage and Widespread Software Bugs:**
  - Ensuring data consistency, durability and high availability is a challenge when the data is distributed.
  - Debugging of data to remove inconsistencies and errors is important but challenging.
- **Challenge 5: Cloud Scalability, Interoperability and Standardization:**
  - Scalability is one of the basic features of Cloud computing and thus requires (for example) dynamic availability of IT resources (hardware) for scaling up.
  - The heterogeneity in hardware and/or hypervisor makes it challenging to dynamically include more hardware/virtualized IT resources.
  - The open virtualization format (OVF) describes and open, secure, efficient, portable and extensible format for packaging and distribution of VMs and the software to be deployed over VMs.
  - OVF allows hypervisor, guest OS and hardware platform independent packaging of VMs and software.
  - Interoperability should be provided for cross hypervisor and cross platform (intel & AMD) live migration of VMs.
- **Challenge 6: Software Licensing and Reputation Sharing:**

- The fact that the license model of commercial software is not suitable for utility computing, the providers have to rely upon open source software and/or bulk usage license.
- If the reputation of a provider is affected (due to consumers' malicious behavior), then there is no service to safe-guard the provider's reputation.

Module No – 111: Public Cloud Platforms Architecture Examples:

- We shall look at a few examples of PaaS platforms on public clouds.



PaaS Platform Architecture

- **Google App Engine (GAE):** It is a popular platform for developing Cloud applications.
  - Based upon technologies:
    - Google File System (GFS): Stores large volumes of data
    - MapReduce: Used in parallel job execution on massive data
    - Chubby (Distributed applications' locking)
    - BigTable (Storage service to access structured data)
  - Consumers are allowed to develop applications in popular languages such as Java, PHP, Go and Python. The following are components of GAE:
    - Datastore
    - Application runtime environment (for web applications)
    - Software Development Kit (SDK) (for local application development)
    - Administration console (management of user application development cycles)
    - Web service infrastructure (interfaces for flexible use of storage and networks resources)
  - Well known applications of GAE are Google Search Engine, Google Docs, Google Earth, and Gmail.

- Consumers can create Cloud applications by using GAE which run on Google data centers.

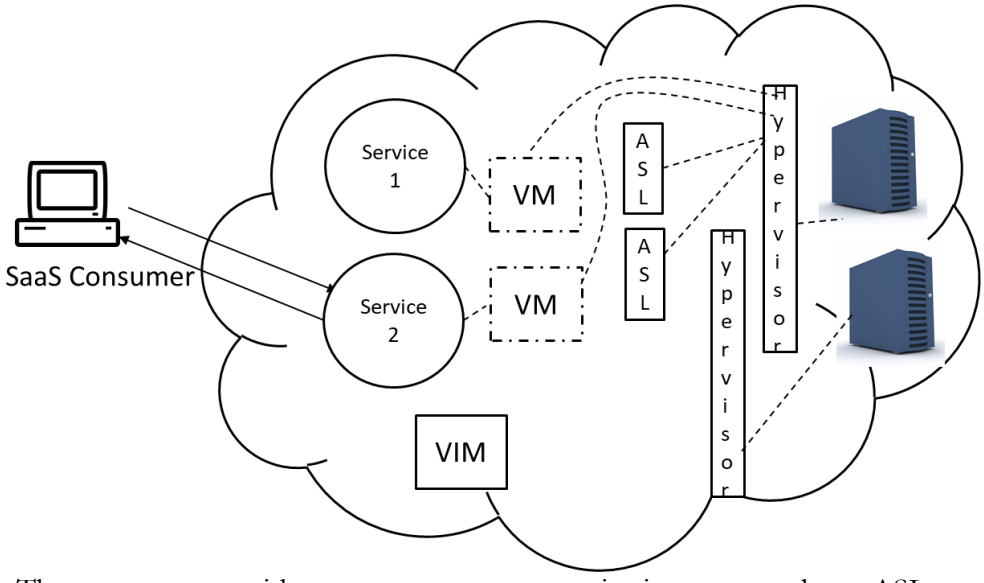
- **Amazon Web Services (AWS):**
  - Amazon provides the SOAP web services and IaaS to the consumers/developers to create and host Cloud services.
  - Amazon Elastic Computing Cloud (EC2) is a web service to provide the VMs for hosting Cloud applications.
  - Simple Storage Service (S3) provides the object-oriented storage service.
  - Elastic Block Service (EBS) provides the block storage interface.
  - Simple Queue Service (SQS) provides inter process message passing.
  - Amazon DevPay service can be used for online billing and account management for the service providers to sell the applications developed and/or hosted on AWS.

Lesson No. 24

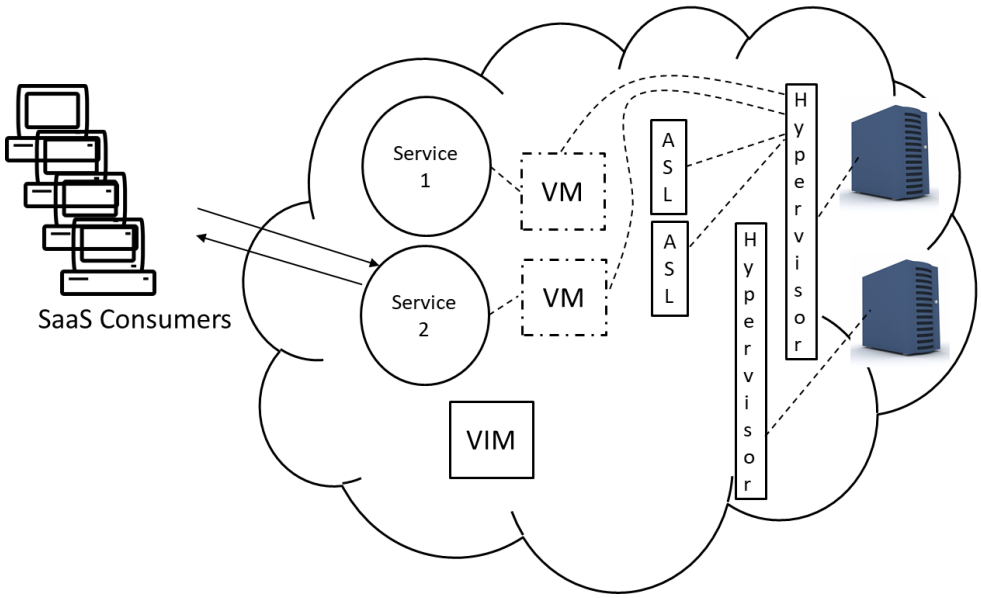
SPECIALIZED CLOUD MECHANISMS

Module No – 112: Automated Scaling Listener (ASL)

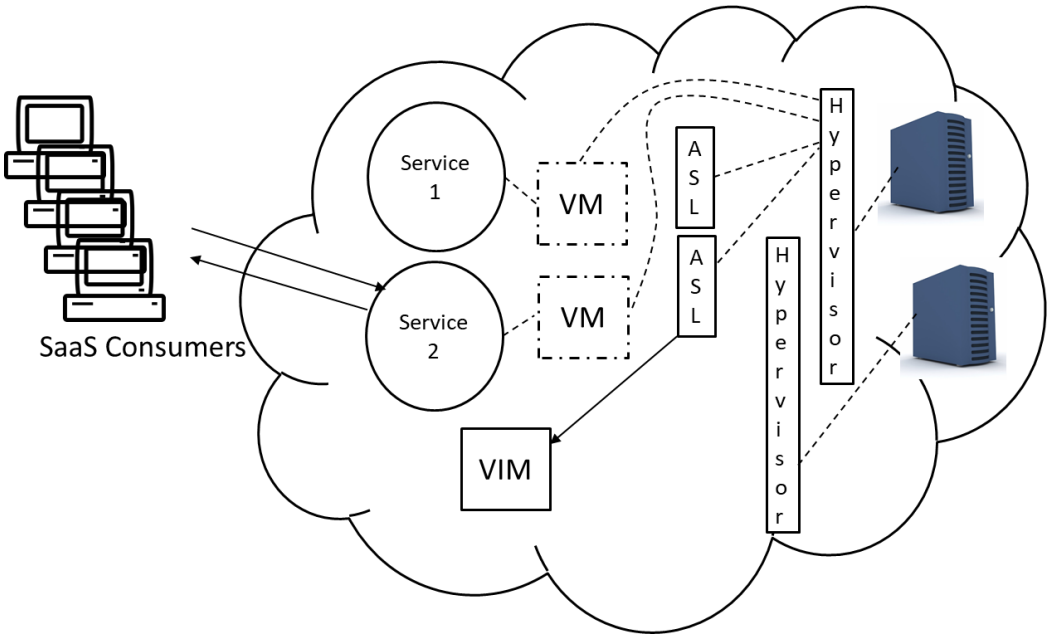
- It is the software module (service agent) which monitors and tracks the communication between Cloud service and the service consumer for dynamic scaling purpose.
  - Can indicate the need for scaling to cloud consumer.
  - Indicates to cloud manager for scaling in/out (if configured to auto scaling by the consumer).



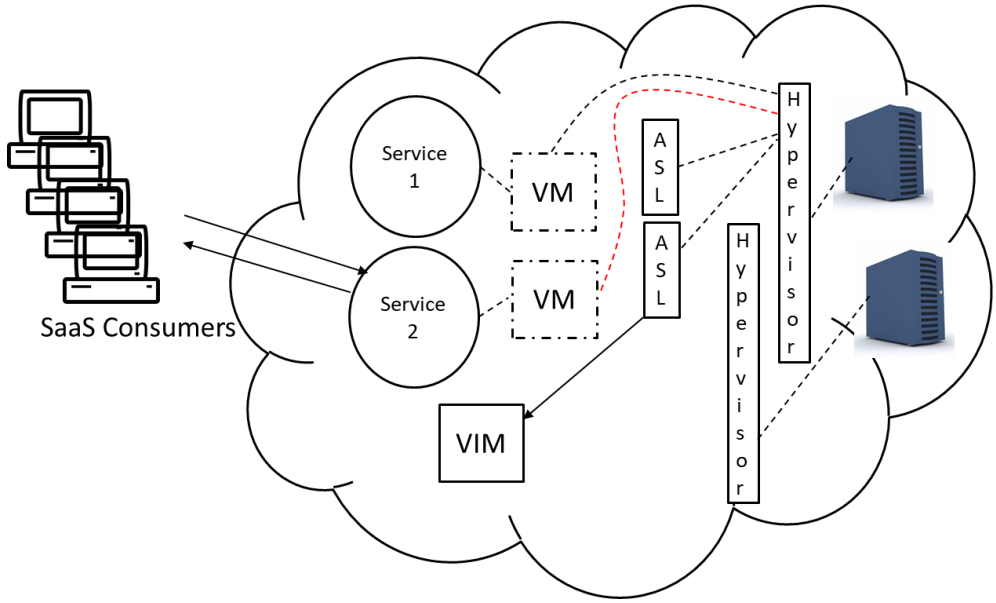
The startup setup with one consumer, two service instances and two ASL modules.



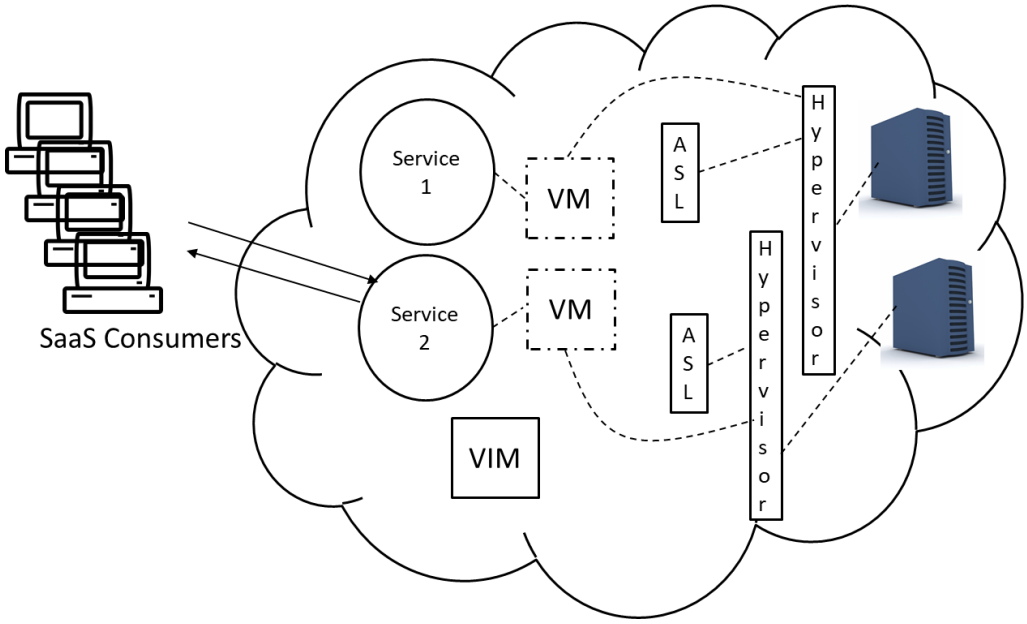
The number of Cloud service 2 consumers are increasing.



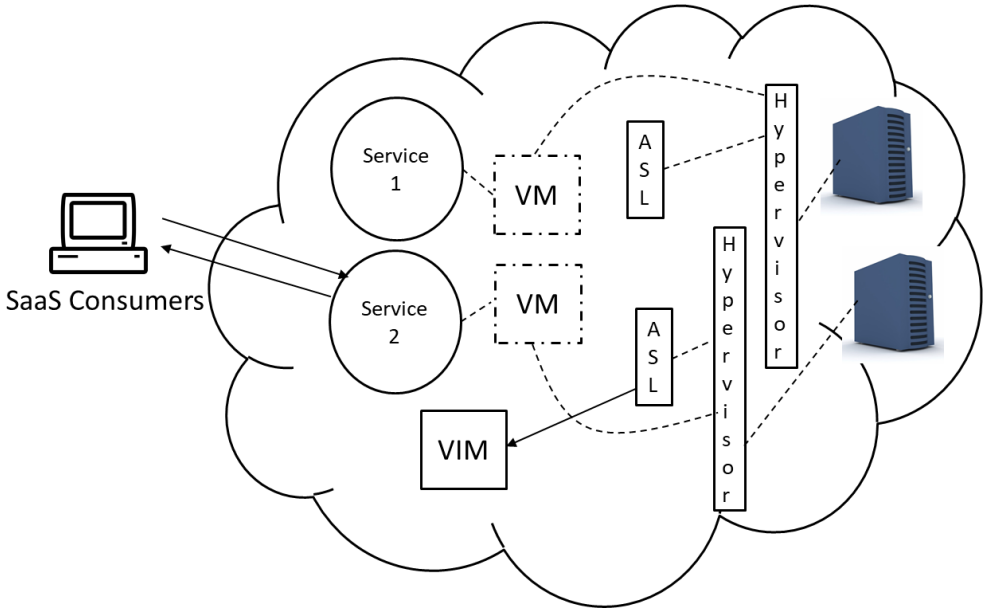
The ASL indicates the Virtual Infrastructure Manager (VIM) for increased load and lack of resources for VM hosting Service 2 on current host/server.



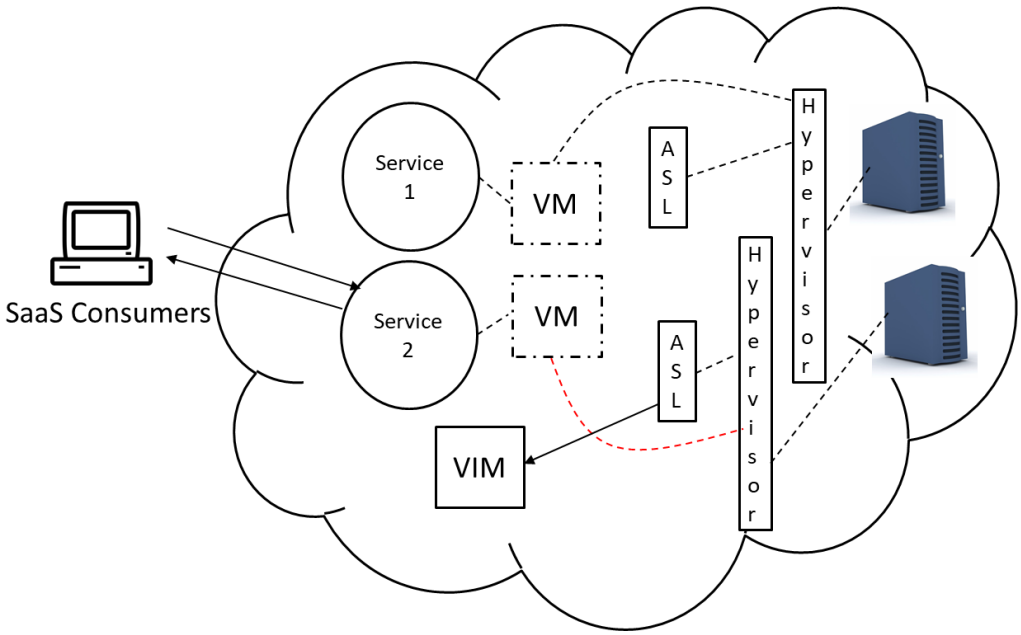
The VIM initiates the migration of VM hosting service 2 to new host for resource availability.



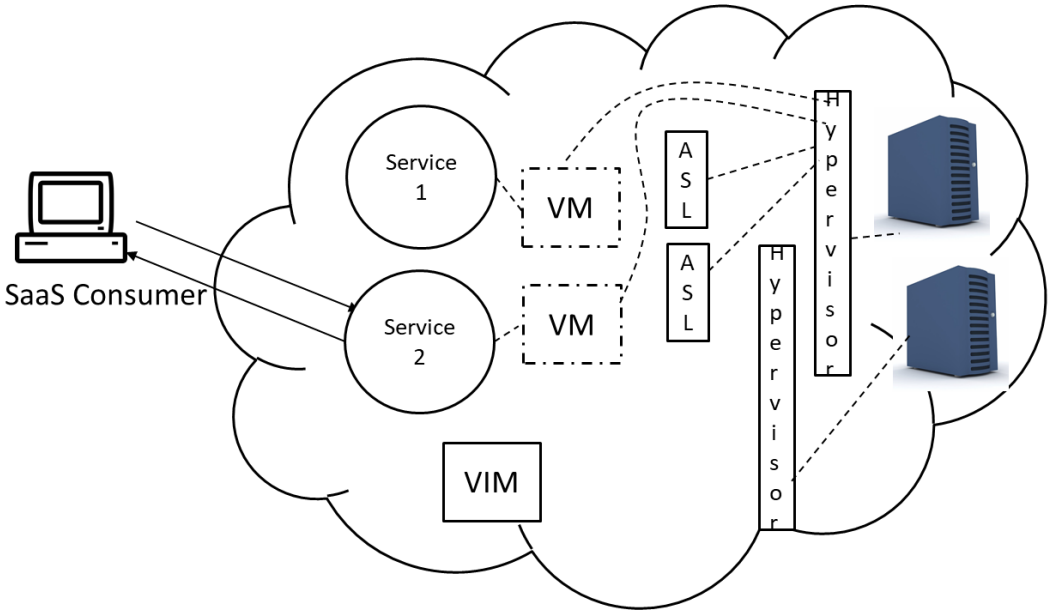
The VM hosting service 2 is migrated to the new host with more resources.



The number of service consumers of Service 2 have decreased. The ASL indicates this to VIM



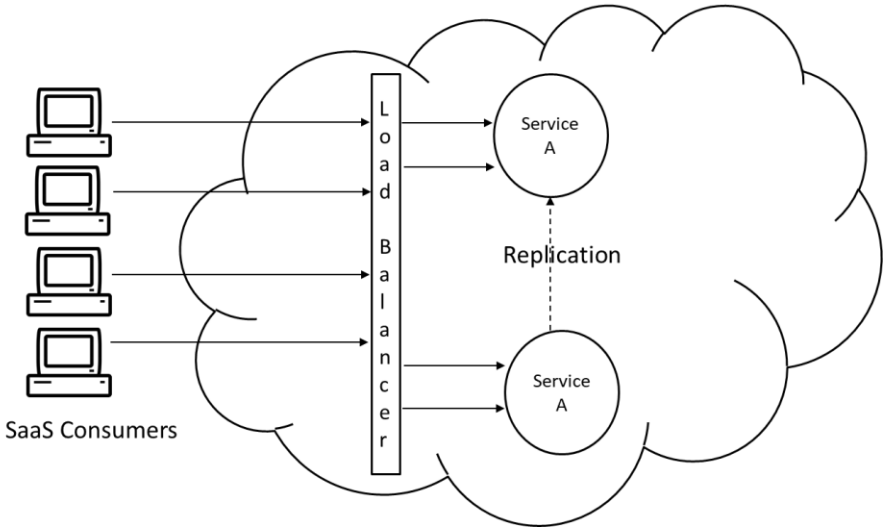
The VIM prepares for migration of service 2 hosting VM for server consolidation.



The VM is migrated to the (previous) host/server.

Module No – 113:

- **Load Balancer:** It is the service agent which distributes workload among multiple processing resources such as multiple service instances**.** Workload is distributed on the basis of:
    - Processing capacity of the IT resource
    - Workload prioritization
    - Content-Aware distribution



Follow the video lecture to understand fully

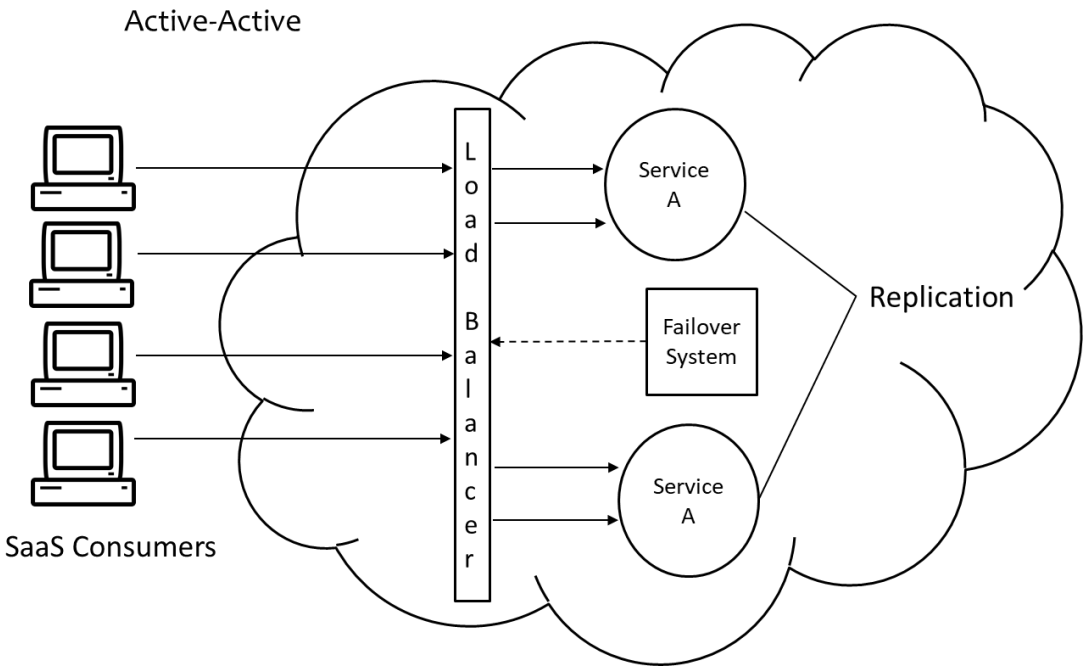
- **SLA Monitor:** Works by pinging (for example) to a service instance to record the "down" status with time.
    - The statistics are used to evaluate the extent of SLA violation.
    - Uses a polling agent (studied before).
- **Pay-per-use Monitor:** It is based upon a monitoring agent (studied before).
    - It collects the resource usage by intercepting the messages sent to a Cloud service by the consumer.
    - Collected data (such as transmitted data volume, bandwidth consumption etc.) is used for billing purpose.
    -

Module No – 114: Failover System

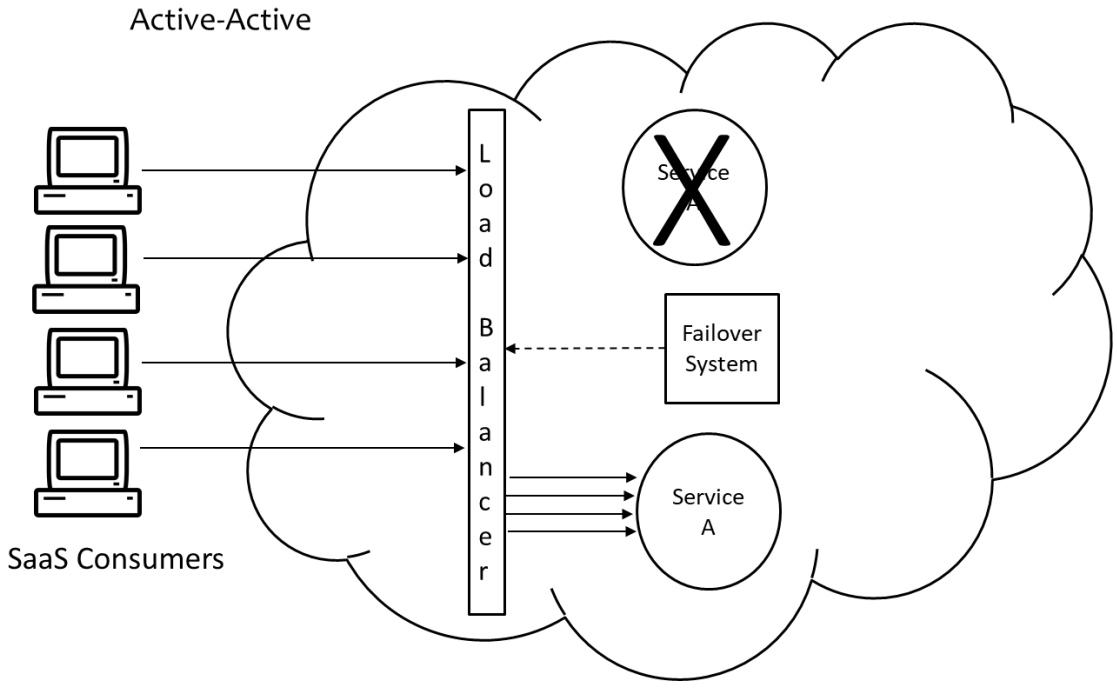
- This mechanism is used to increase the reliability and availability of IT resources by using redundant implementations (for example of Cloud services).
- Used for:
    - Mission critical programs
    - Cloud (supporting) services which can cause a single point of failure.
- The redundant implementations are actively monitored for error detection and unavailability of resources.
- Configurations:
    - Active-Active: The redundant implementation is actively processing the workload. Load balancer implementation is required. The failover system detects the resource failure and directs the load balancer to allocate workload only to active (redundant) implementation. When the failed instance is recovered or replicated, the failover system directs the load balancer to start allocating the workload to all (including replicated) instances.
    - Active-Passive: The redundant instance is passive till the active instance fails. The failover system when detects a failure, it activates a redundant instance and redirects the workload towards the newly activated instance. Upon recovery or replication of failed instance, the failover system puts it to stand-by state while the previously activated instance continues to serve as the active instance.

Module No – 115: Failover System: Case study

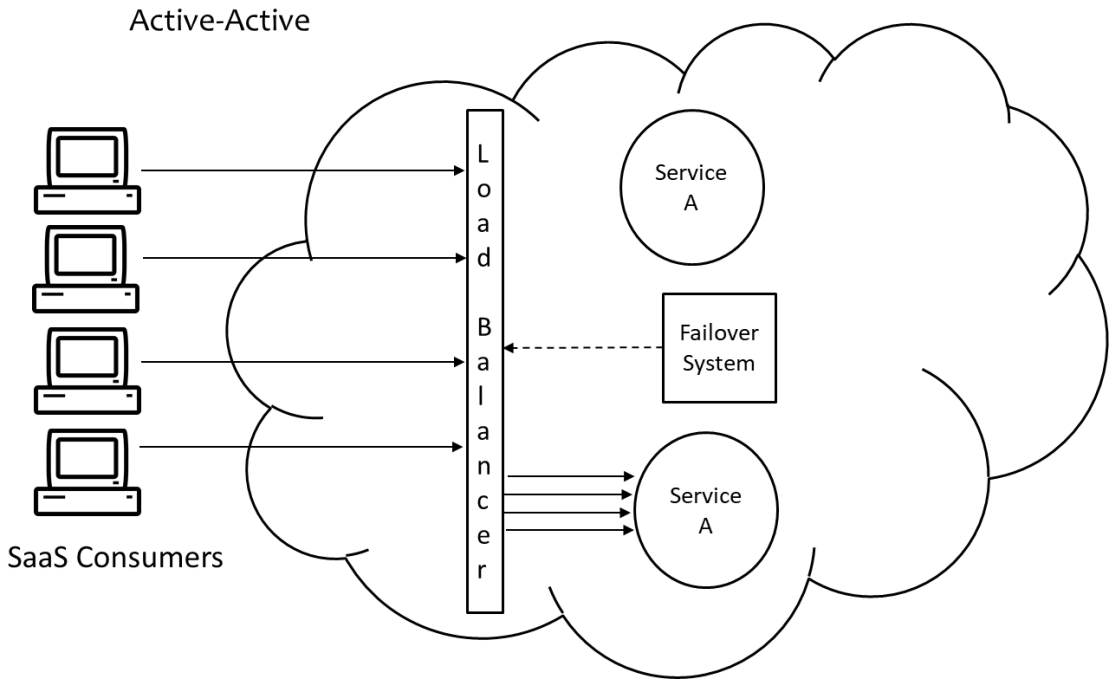
- Let us see the implementations of Failover System:



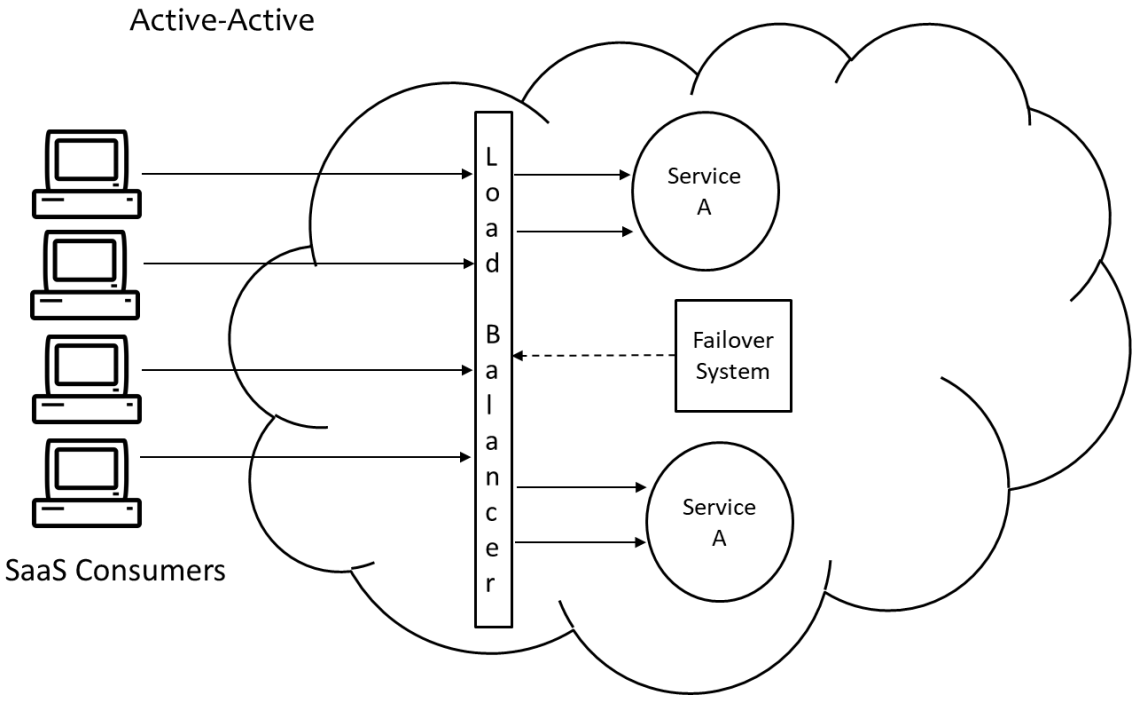
Follow the video lecture to understand fully



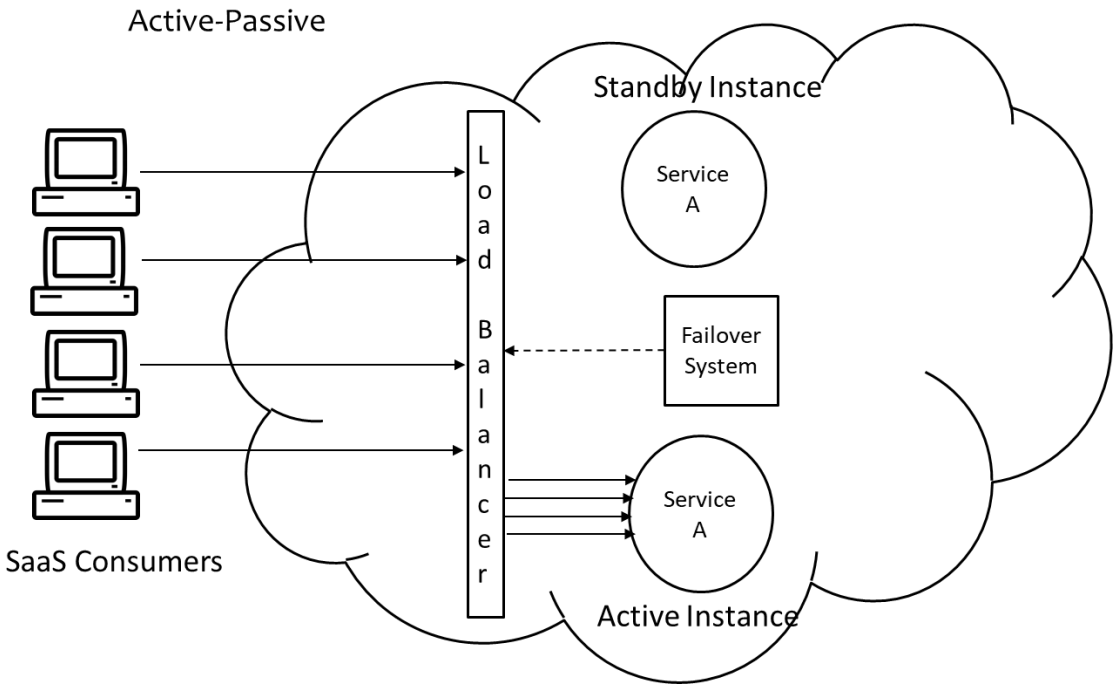
Follow the video lecture to understand fully



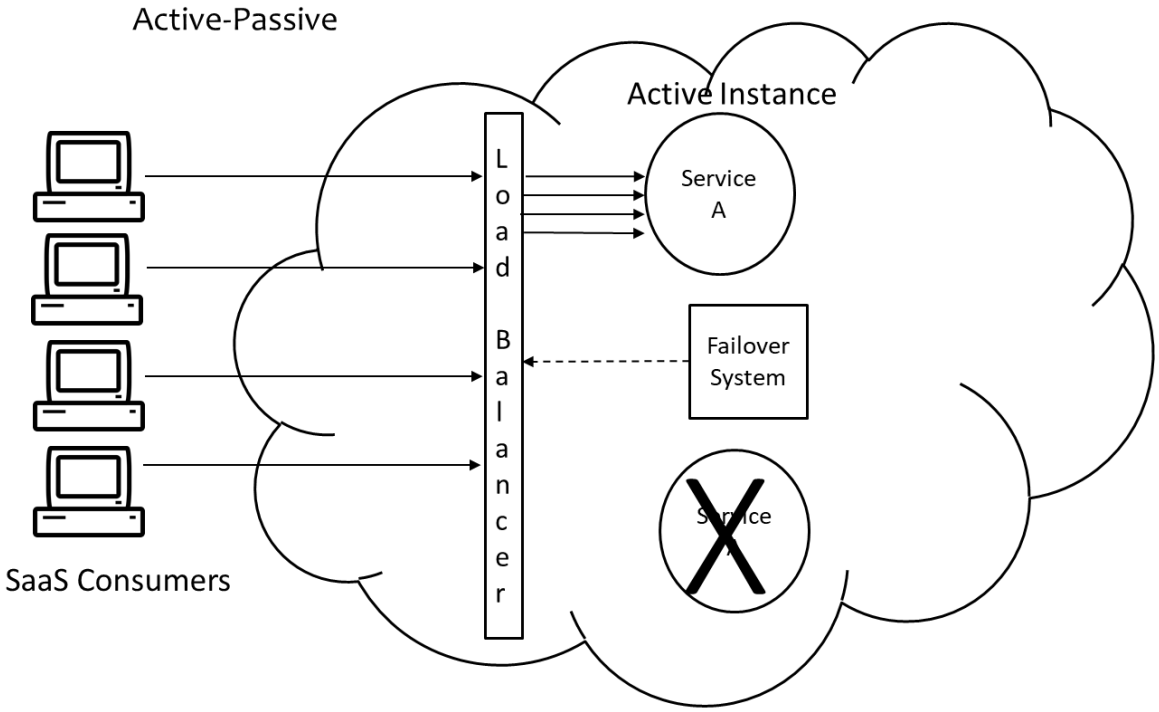
Follow the video lecture to understand fully



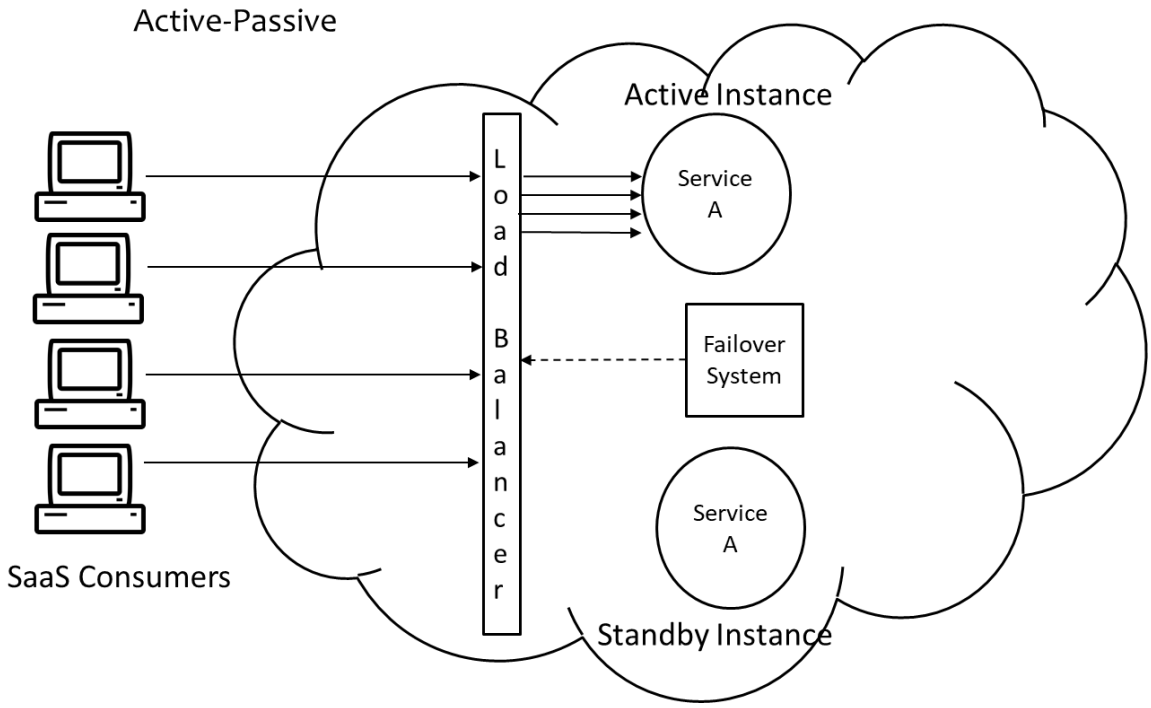
Follow the video lecture to understand fully



Follow the video lecture to understand fully



Follow the video lecture to understand fully



Follow the video lecture to understand fully

Module No – 116: Resource Cluster Mechanism:

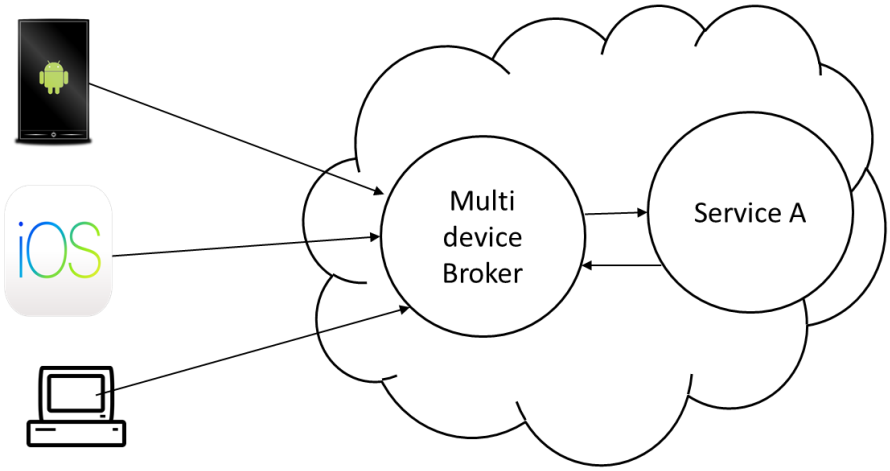
- The Cloud promises virtually unlimited IT resources.
- These IT resources are (although virtualized) but can not be provided through a single physical server.
- It is obvious that the Cloud IT resources are provisioned from multiple physical servers located in a single or multiple data center/s.
- The resource cluster mechanism is used to group multiple IT resources so that they can be used as a single IT resource.
- This increases the computing capacity, load balancing capacity and availability of the clustered IT resources.
- High speed communication links are used to connect the clustered IT resources for:
    - Workload distribution
    - Task scheduling
    - Data sharing
    - System synchronization
- Server clusters may or may not have a shared storage.
- Common types:
    - *Server Cluster*: Consisting of physical or virtual servers. The virtualized clusters support the migration of VMs for scaling and load balancing.

- *Database Cluster*: Is used to keep redundant implementation of databases. It has features to synchronize the data across all the redundant instances.
  - Useful for active-active and active-passive failover systems.
- *Large Dataset Clusters*: This type of cluster is used to partition and distribute large datasets without affecting the data integrity or computing accuracy.
  - Each node processes workloads without any need to depend/communicate with other nodes.

- Additional types:
  - *Load Balanced Cluster*: Implements a load balancer mechanism (discussed before).
  - *HA Cluster*: Implements a failover system (discussed before).

Module No – 117:

- **Multi-Device Broker:** This mechanism is used to transform the messages (received from heterogenous devices of Cloud consumers) into a standard format before conveying them to the Cloud service.
  - The response messages from Cloud service are intercepted and transformed back to the device specific format before conveying to the devices through the multi-device broker mechanism.



Follow the video lecture to understand fully

- **State Management Database:** It is a device used to temporarily store the state data of software programs.
  - State data can be (for example) the configuration and number of VMs being employed to support a user subscription to a PaaS instance.
  - In this way, the programs do not use the RAM for state-caching purposes and thus the amount of memory consumed is lowered.

- The services can then be in a "stateless" condition.
- For example, a PaaS instance (ready-made environment) requires three VMs. If user pauses activity, the state data is saved in state management software and the underlying infrastructure is scaled in to a single VM.
- When the user resumes the activity, the state is restored by scaling out on the basis of data retrieved from state management database.

Lesson No. 25

CLOUD MANAGEMENT

Module No – 118: Remote Administration System

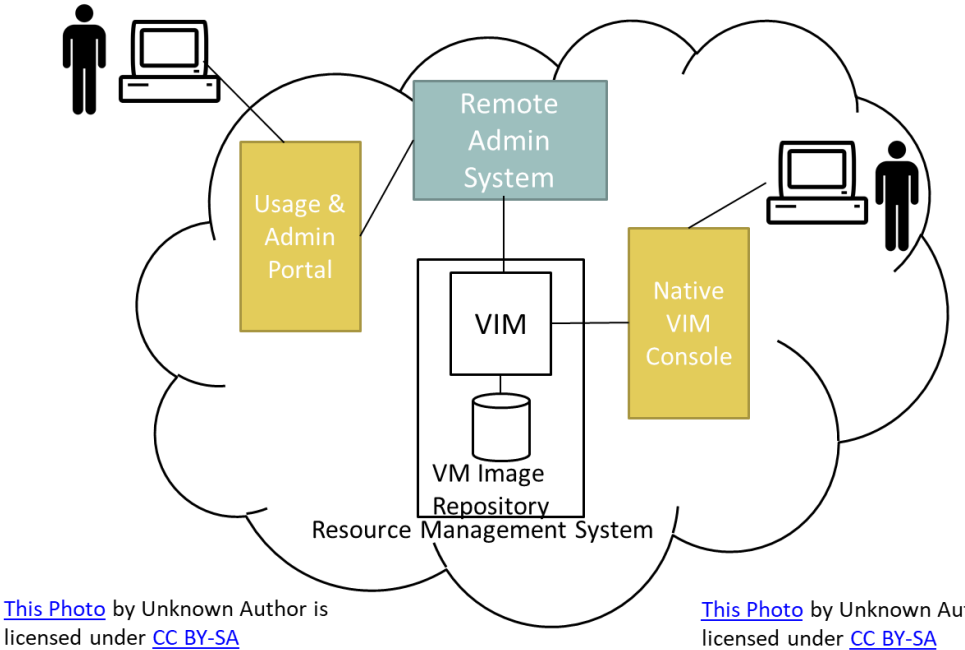
- It is a Cloud mechanism which provides the APIs and tools to the providers to develop and used online portals.
- These portals also provide some administrative controls to the Cloud consumers as well.
- Usage and Administration Portal:
  - Management controlling of Cloud IT resources
  - IT resources usage reports
- Self-Service Portal:
  - The consumer can look at and choose various Cloud services
  - The chosen services/package is submitted to Cloud provider for automated provisioning
- The remote administration console can be used to:
  - Configure and setting cloud services
  - Provision and releasing IT resources for on-demand usage
  - Monitor cloud service status, usage and performance
  - QoS and SLA fulfillment monitoring
  - IT-resource leasing cost and usage fee management
  - Managing user accounts, security credentials, authorization and access control
  - The remote administration console can be used to:
  - Capacity planning
- If allowed, a Cloud consumer can create its own front-end application using API calls of remote administration system.

Module No – 119: Resources Management System

- Utilizes the virtual infrastructure manager (VIM) for creating and managing the virtual IT resources.
- Typical tasks include:
  - Managing the templates used to initialize the VMs
  - Allocating and releasing the virtual IT resources
  - Starting, pausing, resuming and termination of virtual IT resources in response to allocation/release of these resources

- Coordination of IT resources for resource replication, load balancer and failover system
- Implementation of usage and security policies for a Cloud service
- Monitoring the operational conditions of IT resources

- These tasks can be accessed by the cloud resource administrators (personnel) employed by the cloud provider or cloud consumer.
- The provider (and/or the administrator staff of provider) can access the resource management directly through native VIM console.
- The consumer (and/or administrator staff of the consumer) use the remote administration system(created by the provider and) based upon API calls of resource management system.



Resource Management in Cloud

Module No – 120: SLA & Billing Management Systems

- The SLA management system provides features for management and monitoring of SLA.
- Uses a monitoring agent to collect the SLA data on the basis of predefined metrics.
- The SLA monitoring agent periodically pings the service to evaluate the "down" time if occurs.
- The collected data is made available to the usage and administrative portals so that an external and/or internal administrator can access the data for querying and reporting purposes.
- The SLA metrics monitored are in accordance with the SLA agreement.
- The billing management system collects and processes the data related to service usage.
- This data is used to generate consumer invoice and for accounting purposes provider.

- The pay-as-you-go type of billing specifically require the usage data.
- The billing management system can cater for different pricing (pay-per-use, flat rate, per allocation etc.) models as well as custom pricing models.
- Billing arrangement can be pre-usage or post-usage.

Lesson No. 26

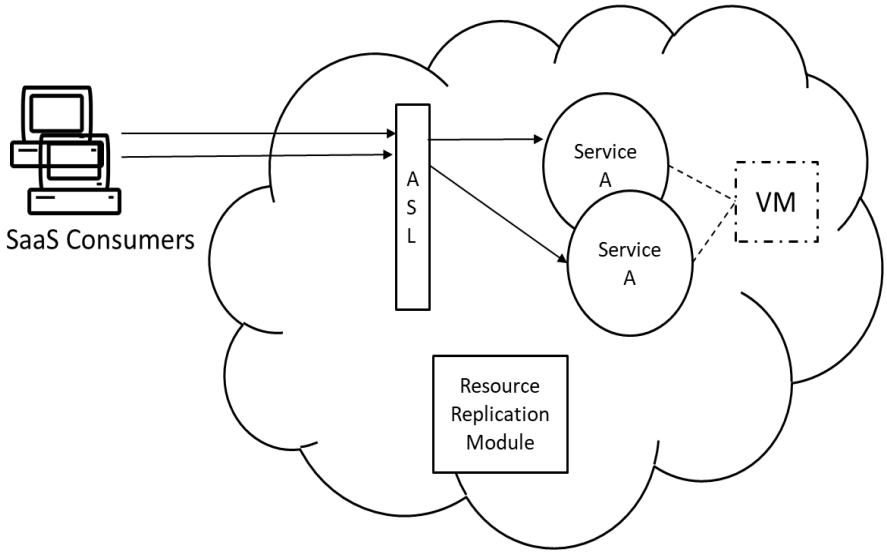
FUNDAMENTAL CLOUD ARCHITECTURES

Module No – 121: Resource Pooling Architecture

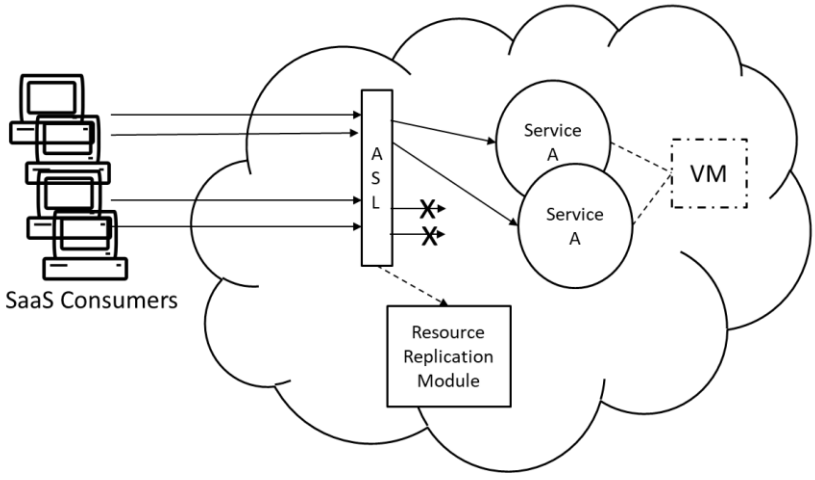
- It is based upon using one or more resource pool in which identical IT resources are grouped and maintained automatically by a system which also ensures that the resource pools remain synchronized.
- A few examples of resources pools are as follows:
    - Physical server pools consisting of (ready to use) networked servers with installed OS and other tools.
    - VM (virtual server) pool/s configured by using one or more templates selected by the consumer during provisioning.
    - Cloud storage pools consisting of file/block based storage structures.
    - Network pools consist of different (preconfigured) network connecting devices that are created for redundant connectivity, load balancing and link aggregation.
    - CPU pools are ready to be allocated to VMs by the multiple of single core.
        - Dedicated pools can be created for each type of IT resources.
        - Individual resource pools can become sub-groups into larger pool.
        - A resource pool can be divided into sibling pools as well as nested pools.
        - Sibling pools are independent and isolated from each other. May have different types of IT resources.
        - Nested pools are drawn from a bigger pool and consist of the same types of IT resources as are present in the parent pool.
- Resource pools created for different consumers are isolated from each other.
- The additional mechanisms associated with resource pooling are:
    - Audit monitor: Tracks the credentials of consumers when they login for IT resource usage.
    - Cloud Usage Monitor
    - Hypervisor
    - Logical Network Perimeter
    - Pay-Per-Use Monitor
    - Remote Administration System
    - Resource Management System
    - Resource Replication

Module No – 122: Dynamic Scalability Architecture:

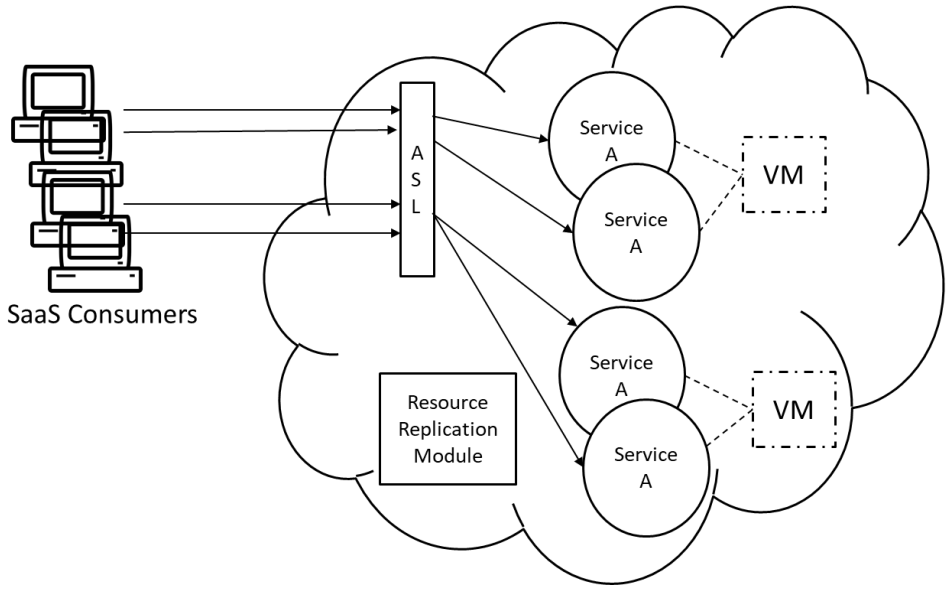
- Dynamic scalability is provided through dynamic allocation of available resources from the resource pool.
- Scaling can be horizontal & vertical and can also be through dynamic relocation. Scaling (considered in this topic) is preconfigured and according to some preset thresholds.
- To implement this architecture, the automated scaling listener (ASL) and Resource Replication Mechanism are utilized.
- Cloud usage monitor and pay-per-use monitor can complement this architecture for monitoring and billing purposes.



Dynamic Scalability Architecture



Dynamic Scalability Architecture



Dynamic Scalability Architecture

Module No – 123:

- **Workload Distribution Architecture:** The workload distribution is required to prevent the following scenarios:
  - Over-utilization of IT resources to prevent the loss in performance.
  - Under-utilization of IT resources to prevent the over expenditure.
  - The workload is distributed on the basis of a load balancing algorithm with the scope/s of VMs, Cloud storage devices and cloud services.
  - Accompanied by the following mechanisms:
    - Audit monitor
    - Cloud usage monitor
    - Logical network perimeter
    - Resource cluster
    - Resource replication
- **Service Load Balancing Architecture:** This architecture specifically is used for workload distribution for Cloud services.
  - Multiple instances of the Cloud services are deployed with load balancing system.
  - The load balancer can be either deployed as an external module.
  - Alternatively the load balancer can be integrated into the Cloud service. In this case, one of the redundant instances of a service becomes the master or workload distributor to dynamically allocate the workload among service instances.
  - The following modules are additionally required:
  - Cloud usage monitor
  - Resource cluster (with active-active failover system configuration)
  - Resource replication

Module No – 124: Elastic Disk Provisioning Architecture

- Cloud costing model for disk storage may charge on the basis of total volume of allocated storage space instead of total space used.
- The elastic disk provisioning architecture implements a dynamic storage provisioning based billing.
- The user is charged only for the consumed storage.
- The technique of thin-provisioning of storage is used.
- Thin-provisioning allocates the storage space dynamically for the VM's storage.
- Requires some extra overhead when more storage space is to be allocated.
- The thin-provisioning software is required to be installed on VMs to coordinate the thin-provisioning process with the hypervisor.
- Requires the implementation of:
    - Cloud usage monitor
    - Resource replication module (for converting thin-provisioning into thick or static disk storage)
    - Pay-per use monitor tracks and reports the granular billing related to disk usage.
- In order to avoid data loss and service unavailability due to disk failure, redundant storage is applied.
- Additionally, in case of network failure, the disruptions in Cloud services can be avoided through redundant storage incident.
- This is part of failover system (active-passive).
- The primary and secondary storage are synchronized so that in case of a disaster, the secondary storage can be activated.
- A storage device gateway (part of failover system) diverts the Cloud consumers' requests to secondary storage device whenever the primary storage device fails.
- The primary and secondary storage locations may be geographically apart (for disaster recovery) with a (possibly leased) network connection among the two sites.

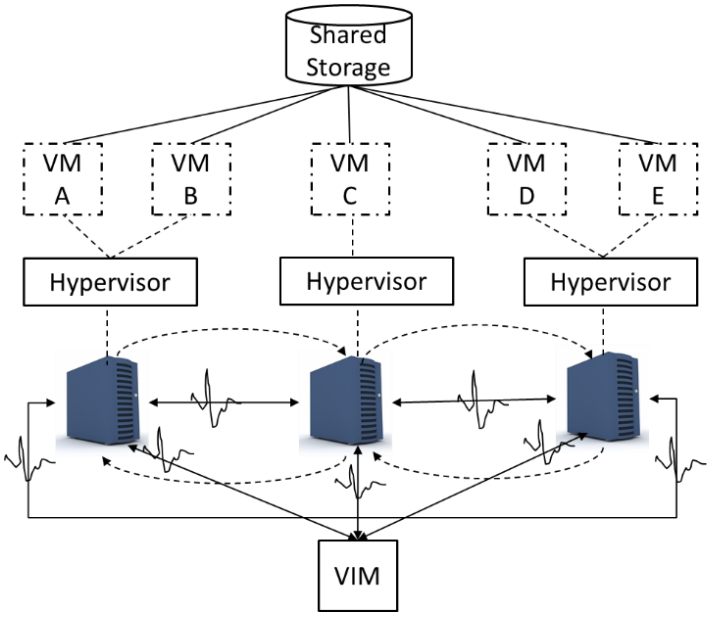
Lesson No. 27

ADVANCED CLOUD ARCHITECTURES

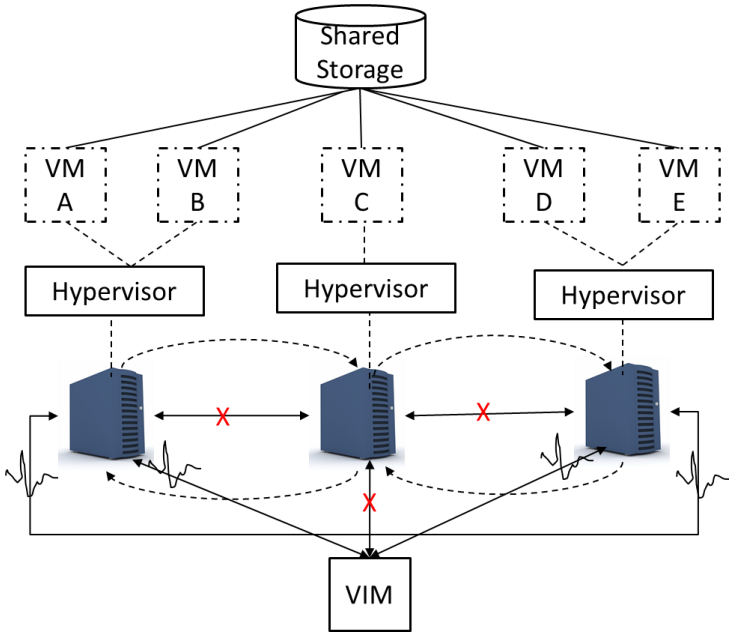
Module No – 125: Hypervisor Clustering Architecture:

- VMs hosted on a physical server are managed by hypervisor.
- The failure of the physical server cripples the hypervisor and therefore the hosted VMs also become unavailable.
- The hypervisor clustering creates a high-availability cluster across multiple physical servers.
- The hypervisor cluster operations such as live VM migration and heartbeat message exchange with hypervisors, are controlled by VIM (virtual infrastructure manager) module.
- The hypervisor cluster uses a shared storage to support a prompt live-migration of VMs.
- The additional modules to be implemented with hypervisor clustering architecture are:
    - Logical network perimeter (to create logical boundary of each hypervisor cluster).

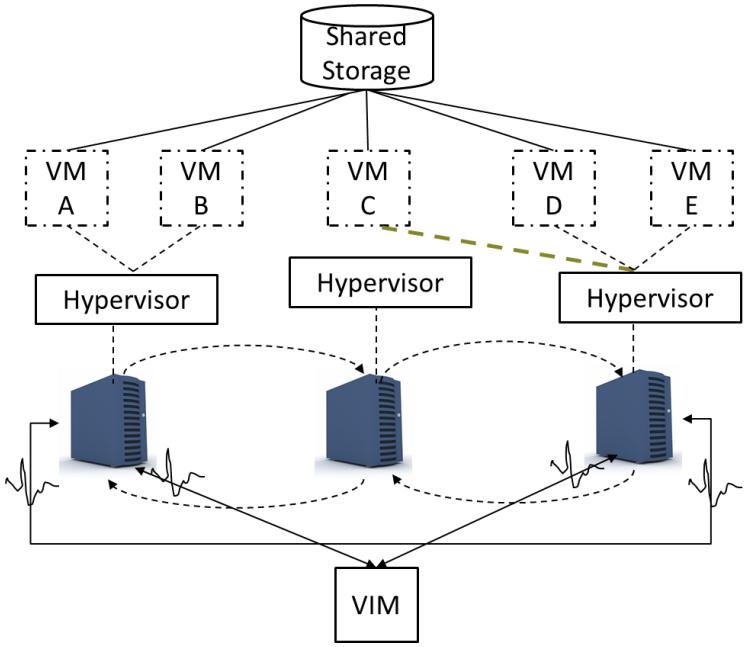
- Resource replication module



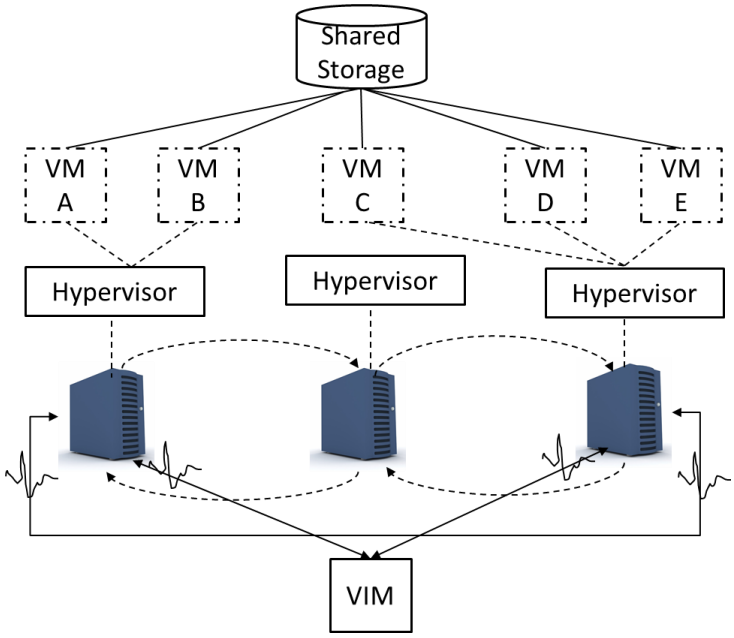
Follow the video lecture to understand fully



Follow the video lecture to understand fully



Follow the video lecture to understand fully

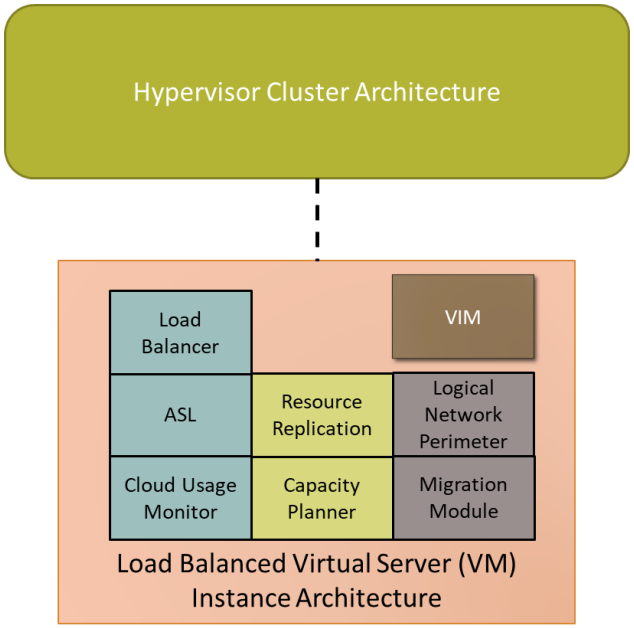


Follow the video lecture to understand fully

Module No – 126: Load Balanced Virtual Server/VM Instances Architecture:

- It balances the physical server utilization through VM migration in the hypervisor cluster architecture.
- Avoids over/under utilization of physical servers.

- Maintains performance of services hosted on VMs.
- Implements a capacity watchdog/monitor system consisting of:
  - Cloud usage monitor
  - Live VM migration module
  - Capacity planner
- The cloud usage monitor tracks the usage of physical server and VMs hosted on that server. In case of fluctuation in usage, it reports to capacity planner module.
- The capacity planner modules dynamically matches the capacities of physical servers and the resource demands of hosted VMs.
- If any VM is facing resource shortage then the capacity planner initiates the VM migration to the suitable server with sufficient capacity.
- The following modules are integrated into this architecture:
- Automated scaling listener (for monitoring workload over VMs) and load balancer
- Logical network perimeter to comply with privacy requirements of SLA
- Resource replication for load balancing



Load Balanced Virtual Server/VM Instances Architecture

Module No – 127: Non-Disruptive Service Relocation Architecture:

- A Cloud service may become disruptive/unavailable/down due to:
  - Over burden of processing load
  - Scheduled updates
- The requests of service consumers are not processed during the time of unavailability.
- By using the specialized Cloud architectures, the non-disruptive service relocation architecture can be implemented.

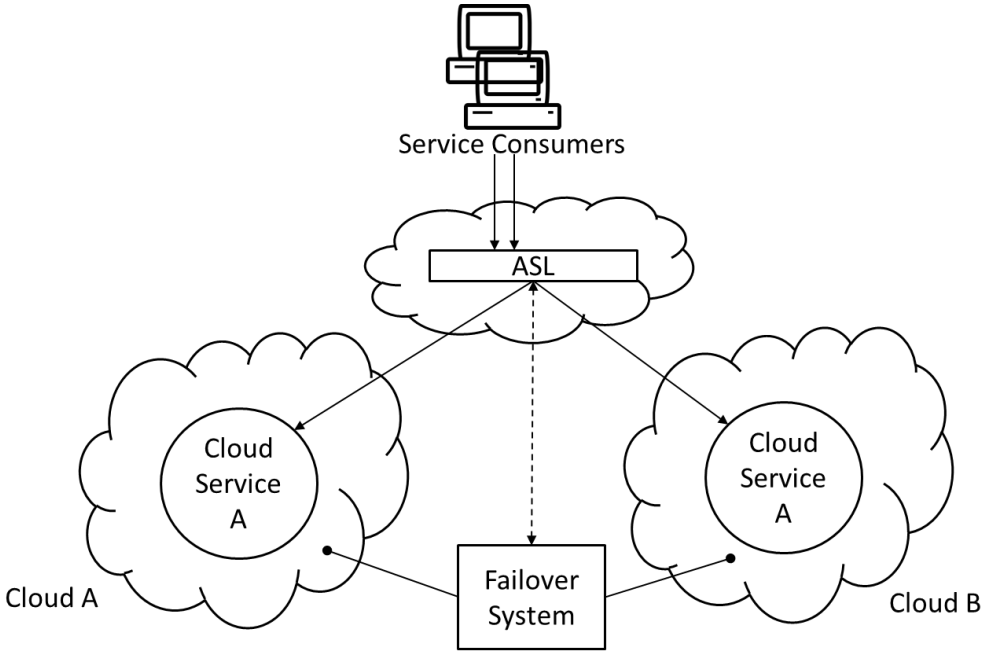
- Either the duplication of cloud service or service migration is used to provide a non-disruptiveness in service.
- By using the specialized Cloud architectures, the non-disruptive service relocation architecture can be implemented.
- Either the duplication of cloud service or service migration is used to provide a non-disruptiveness in service.
- In service-duplication implementation, the service run-time is temporarily replicated to another location and then synchronized with the primary deployment.
- The consumers' requests are diverted to the temporary deployment and then the primary deployment is made unavailable for maintenance.
- In case of migration of service to another location (such as on the indication of automated scaling listener), then it is a permanent relocation. Means, the temporary duplicate copy is not made.
- The service hosting VM's migration procedure depends upon the fact that the VM storage is hosted on shared or local physical server storage is used.
- In later case a replicated copy of to-be-migrated VM is made on the destination server and then powered-on, the consumers' requests are redirected to the duplicated instance through load balancer module. After that, the original VM is deactivated.
- In formal case, the above procedure is not required if the destination server can also access the same shared storage.
- The following are the important modules for the implementation:
    - Automated Scaling Listener
    - Load balancer
    - Hypervisors
    - VMs
    - Cloud storage device
- The additional modules are:
    - Cloud usage monitor
    - Pay per use monitor
    - Resource replication
    - SLA management system
    - SLA monitor

Module No – 128: Zero Downtime Architecture

- The failure of the physical server results in the unavailability ofVMs hosted on that server.
- The services deployed over the unavailable VMs are obviously disrupted.
- The Zero downtime architecture implements a failover system through which the VMs (from the failed physical server) are dynamically shifted to another physical server without any interruption.
- The VMs are required to be stored on a shared storage.
- The additional modules required may include:
    - Cloud usage monitor
    - Logical network perimeter

- Resource cluster group (containing active-active clusters to assure high availability of IT-resources for VM)
- Resource replication

- **Cloud Balancing Architecture:** It is the implementation of failover system across multiple clouds.
  - It improves/increases the following features:
    - Performance and scalability of IT-resources
    - Availability and reliability of IT resources
    - Load balancing
  - Requires an automated scaling listener and failover system.
  - The automated scaling listener redirects the requests of service consumers towards multi-cloud redundant implementations of service instances based on on-going scaling and performance requirements.
  - The failover system (detects any failure/s and) coordinates with automated scaling listener with information regarding the extent of failure so that the automated scaling listener can adjust the relaying of consumers' requests accordingly.



Cloud Balancing Architecture

Module No – 129: Resource Reservation Architecture:

- A situation of *resource constraint* may arise when two or more Cloud-consumers (sharing some IT-resources such as a resource pool) experience a performance loss when the runtime resource demand exceeds the capacity of the provided resources.
- Resource constraint situation may also arise for the IT-resources not configured for sharing such as nested and/or sibling pools when one pool *borrows* the resources from the other

pool. The lending pool may create resource constraints for its consumers later on if the borrowed resources are not returned sooner.

- If each consumer can be assured the availability of a minimum volume of:
    - Single IT resource
    - Portion of an IT resource
    - Multiple It resources
        - Then this implements a resource reservation architecture.
- In case of implementation for resource pools, the reservation system must assure that each pool maintains a certain volume of resource/s in *unborrowable* form.
- The resource management system mechanism (studied earlier) can be utilized for resource reservation.
- The resource/s volume in a pool or the capacity of a single IT resource which exceeds the reservation threshold can be shared among the consumers.
- The resource management system manages the borrowing of IT resources across multiple resource pools.
- The additional modules that can be implemented are:
    - Cloud usage monitor
    - Logical network perimeter (for resource borrowing boundary)
    - Resource replication (just in case new IT resources are to be generated)

Module No – 130: Dynamic Failure Detection and Recovery Architecture

- It may be possible to detect and counter some failures in Cloud environment if there is an automated system with failure diagnosis and solution selection intelligence.
- This architecture establishes a resilient watchdog/module containing the definitions of pre-marked events and the runtime logic to select the best (predefined) routine to coup with those events.
- The resilient module generates alarms/reports the events which are not predefined.
- The resilient watchdog module performs the following five core functions:
    - Monitoring
    - Identifying an event
    - Executing the reactive routine/s
    - Reporting
- This architecture allows the implementation of an automated recovery policy consisting of predefined steps and may involve actions such as:
    - Running a script
    - Sending a message
    - Restarting services
- Can be integrated into a failover system along with SLA management system.

Module No – 131: Bare-Metal Provisioning Architecture:

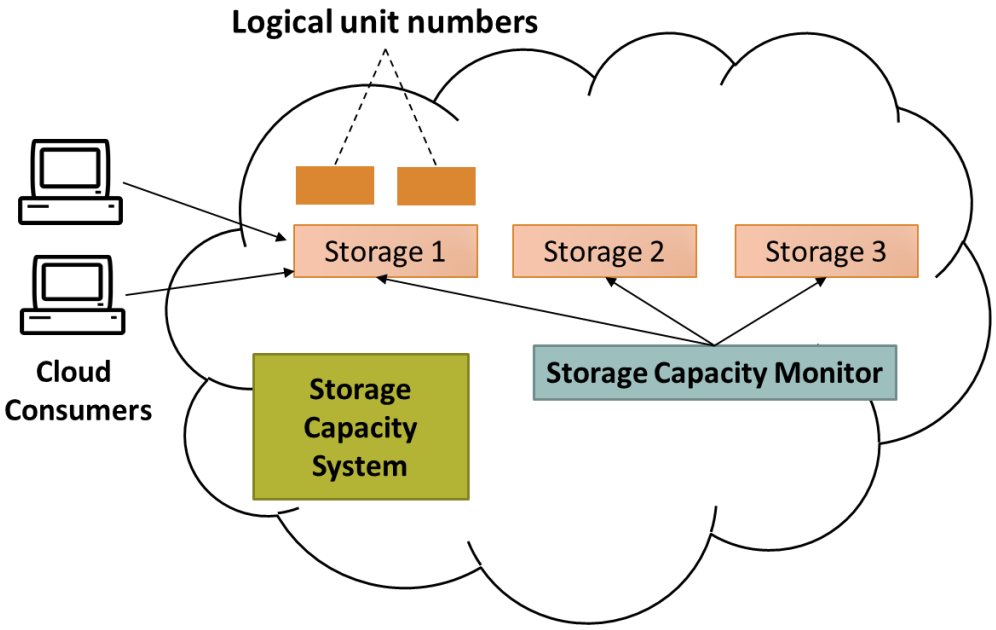
- This architecture implements the provisioning of bare-metal servers on demand.
- These servers do not have an OS or hypervisor installed when the provisioning process is initiated.
- Therefore these servers are required to contain some mechanism to be accessed through a remote management console to install OS/hypervisor.
- This functionality is either built-in into ROM, contained in the chipset or through an expansion slot.
- These can be accessed through remote administration system through web.
- The IP address of the physical server is required for connectivity.
- The IP address can be configured manually or through DHCP service.
- The bare-metal provisioning can be automated to avoid manual deployment errors and time delays in OS/hypervisor deployment through remote management system.
- The automated bare-metal provisioning allows the consumers to obtain multiple servers by using the management software.
- The automated bare-metal provisioning is centrally controlled.
- The controlling software connects with the server management software for OS installation.
- Following are the steps:
    - Consumer connects to central software through self-service portal.
    - The available servers are shown.
    - The consumer choses the server and the OS to be installed.
- The resource management system (studied before) is used to install the required chosen OS.
- The consumer starts using the provisioned server.

Module No – 132: Rapid Provisioning Architecture:

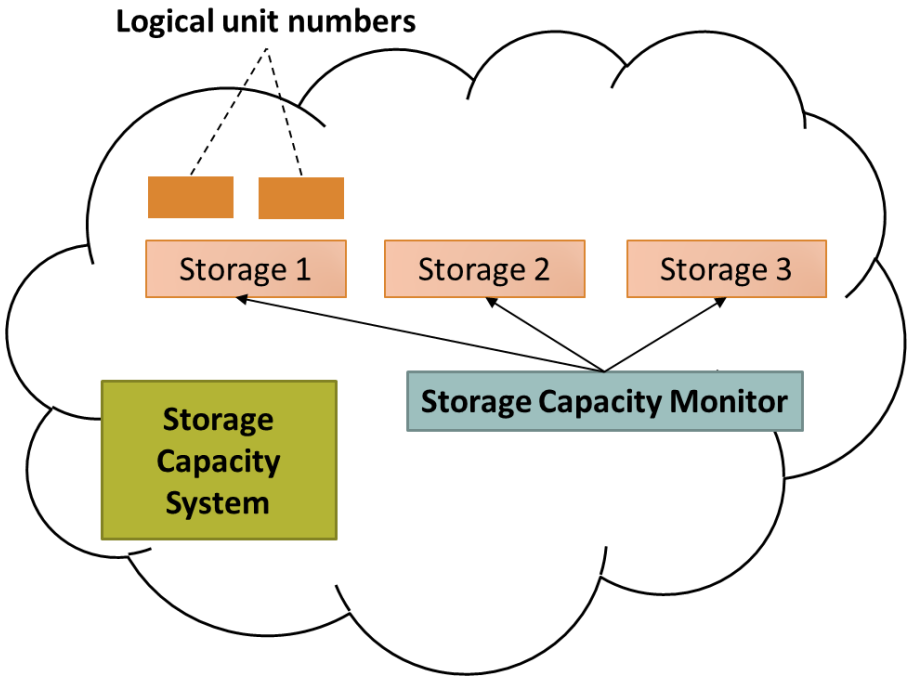
- The provisioning Cloud IT-resources can be automated to save time, reduce human related errors and to increase the throughput.
- For example a consumer can initiate the automated provisioning of 50 VMs simultaneously instead of waiting for one VM at a time.
- The rapid provisioning architecture has a (centralized) control module complemented by:
    - Server templates
    - Server images (for bare-metal provisioning)
    - Applications and PaaS packages (software and applications & environments)
- OS and Application baselines (configuration templates applied after installation of OS and applications)
- Customized scripts and management modules for smooth procedures
- The following steps can be visualized during the automated rapid provisioning:
    - A consumer chooses a VM package through self-service portal and submits the provisioning request.
    - The centralized provisioning module selects an available VM and initiates it through a suitable template.
    - Upon initiation, the baseline/s templates are applied.
    - The VM is ready to use now.

Module No – 133: Storage Workload Management Architecture:

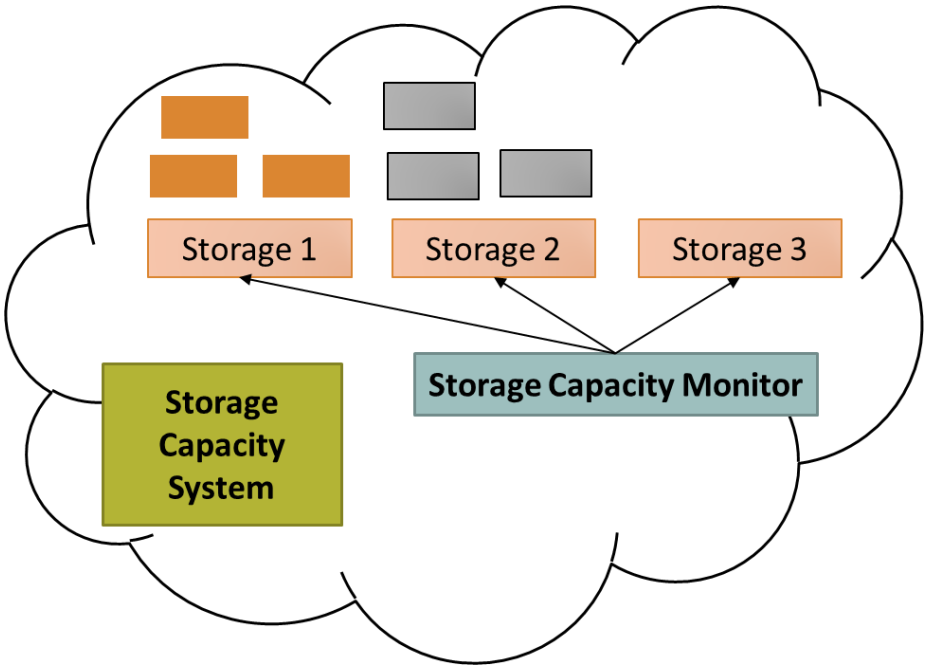
- Logical Unit Number is a logical drive that represents a partition of a physical drive.
- The storage workload management architecture ensures the even distribution of all logical-unit-numbers across the Cloud storage devices.
- The even distribution of logical unit numbers is done through implementation of a storage capacity system and a storage monitoring module.
- The storage capacity monitoring module highlights the overburden storage device.
- The storage capacity system evenly distributes the logical-unit-number drives.
    - Additional modules:
    - Cloud usage monitor
    - Load balancer
    - Logical network perimeter



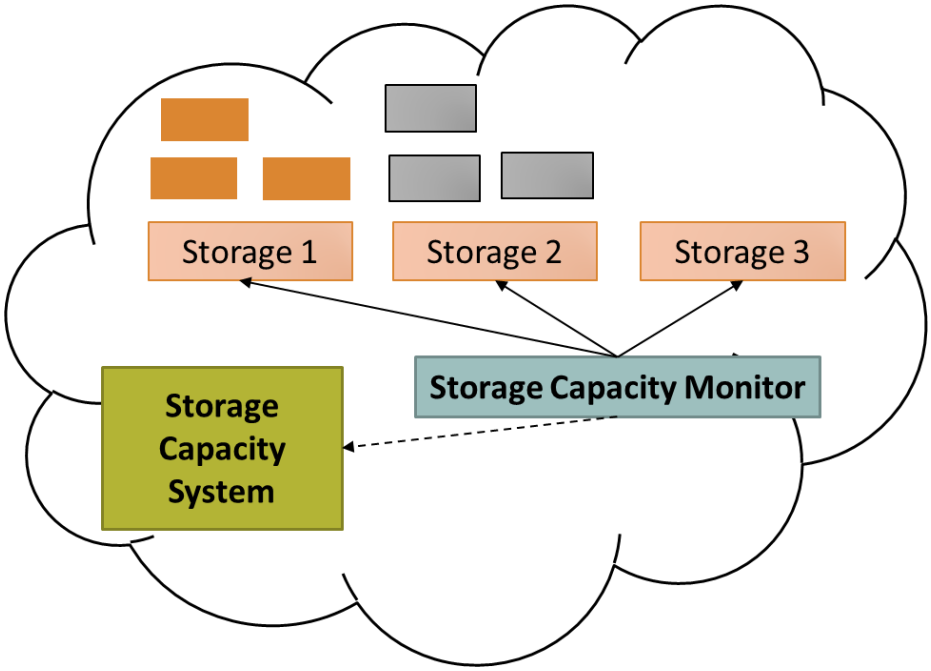
Initial distribution of logical unit numbers across the Cloud storage devices.



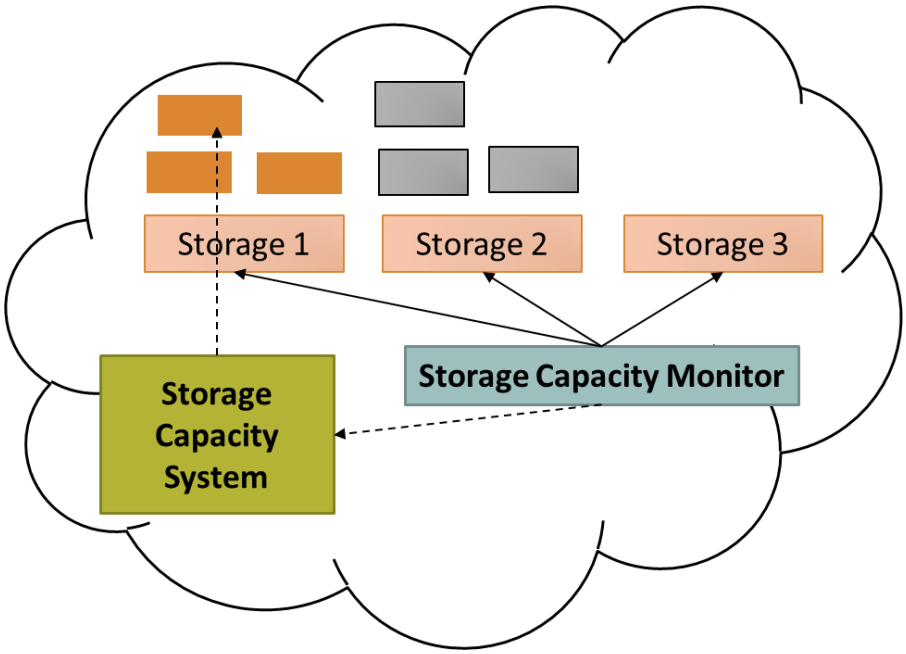
Storage capacity monitoring module checks each storage device.



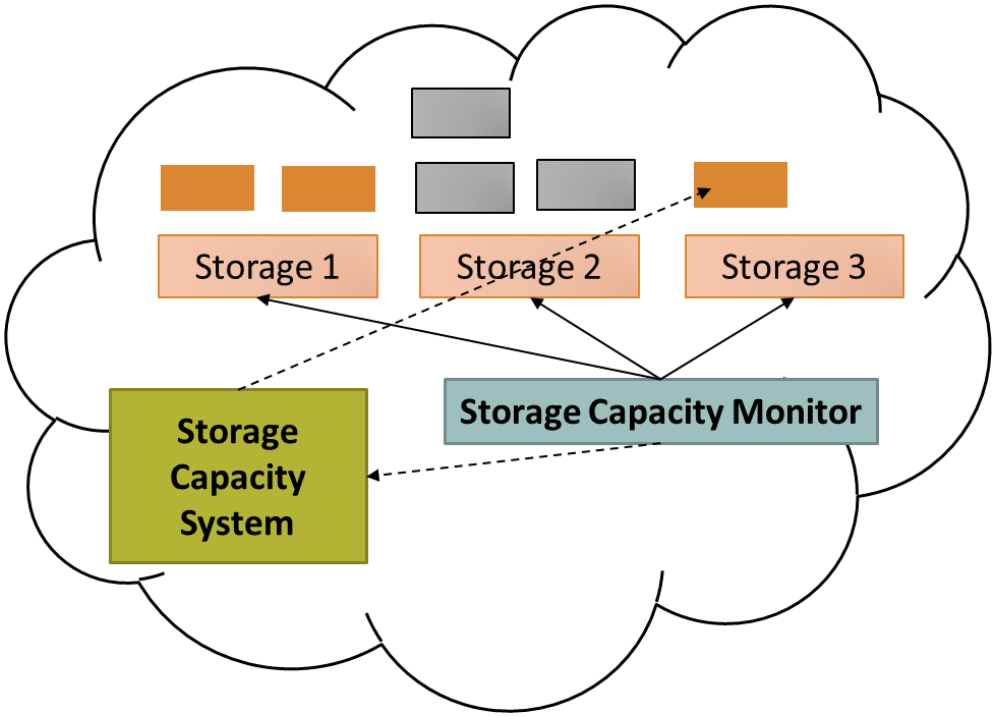
Storage capacity monitoring module highlights the overburden storage device



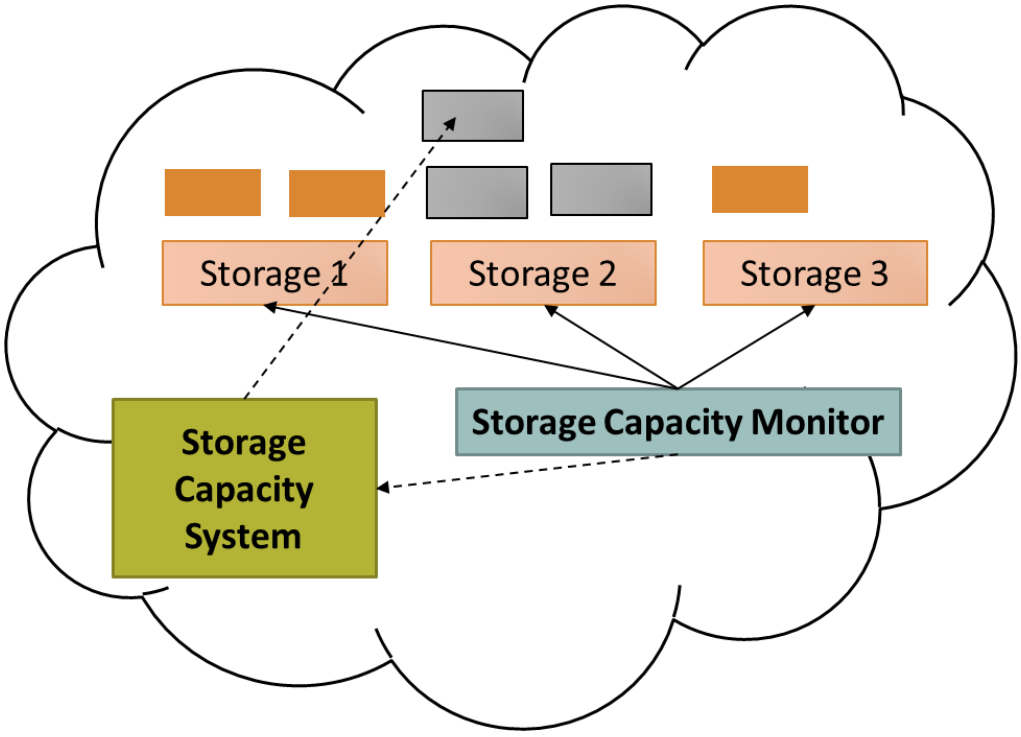
Storage capacity monitoring module indicates the Storage capacity system for migration of logical unit number to another storage device.



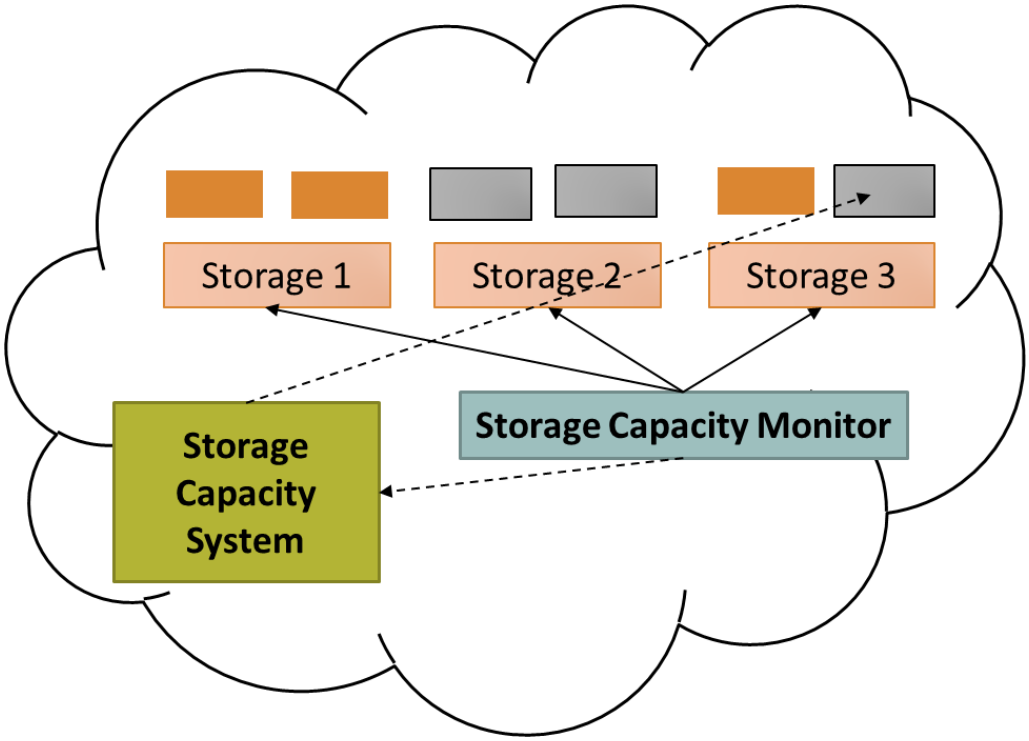
Storage capacity system identifies a logical unit number to migrate.



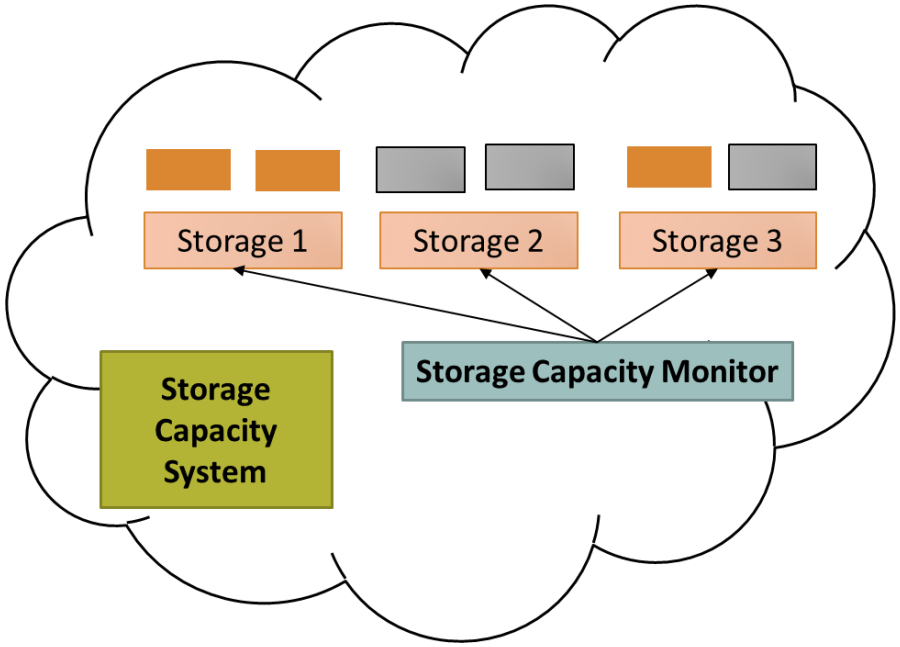
Storage capacity system identifies the destination storage device and shifts the logical unit number to destination device.



Storage capacity monitoring module indicates the Storage capacity system for migration of logical unit number to another storage device.



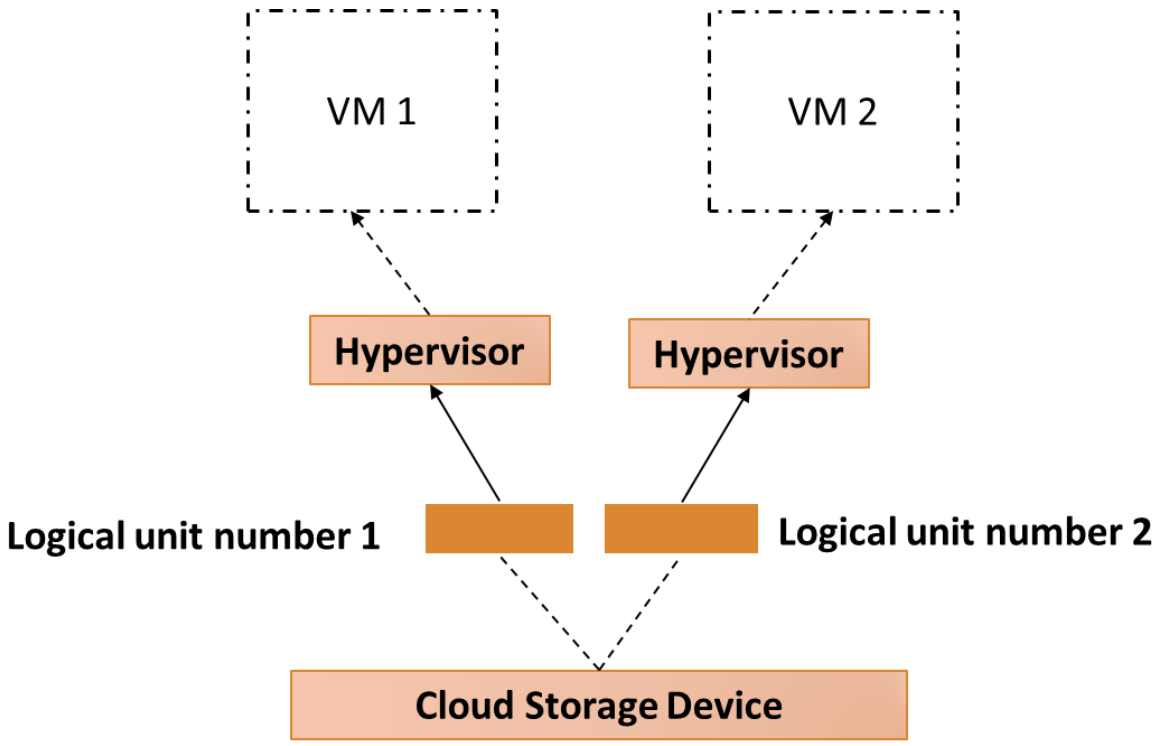
Storage capacity system identifies the destination storage device and shifts the logical unit number to destination device.



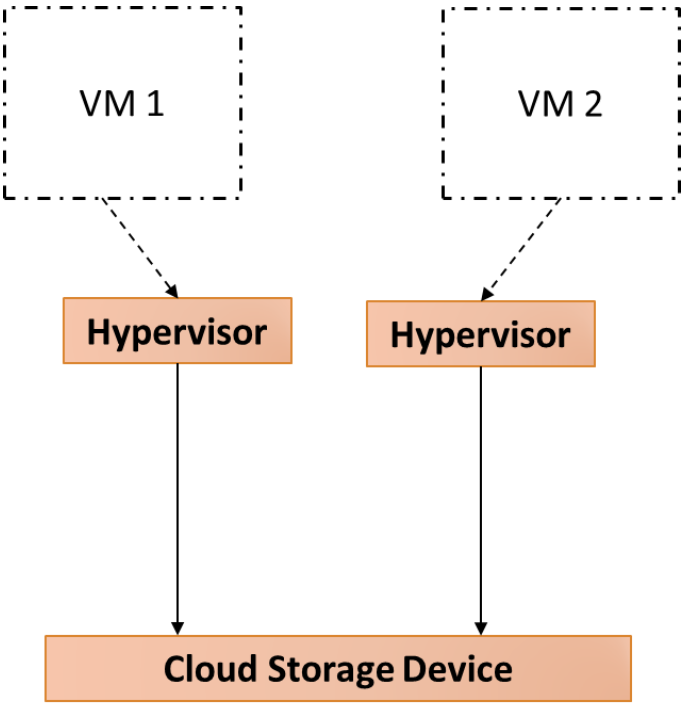
The result is the even distribution of logical unit numbers across all storage devices.

Module No – 134: Direct I/O Architecture:

- The VMs access various physical I/O circuits/cards of the hosting physical server through the hypervisor. This is called I/O virtualization.
- However, times may come when the hypervisor assisted access may become a bottleneck for concurrent I/O requests.
- The direct I/O architecture is the possibility of accessing the physical I/O devices from VMs without intervention of hypervisor.
- The physical server's CPU has to be compatible to direct I/O.
- Additional modules required are:
  - Cloud usage monitor
  - Logical network perimeter (to allow only a limited number of VMs to for direct I/O)
  - Pay-per-use monitor
- It is a type of direct I/O in which the VMs access the logical unit numbers directly.
- The VMs can also be given direct access to block level storage.



A type of direct I/O in which the VMs access the logical unit numbers directly.



VMs can also be given direct access to block level storage.

Module No – 135: Dynamic Data Normalization Architecture:

- The duplication of data over Cloud can cause some problems such as:
- Increase time to store, backup and copy the data
- More space is required
- More cost is to be paid by the consumer
- Data synchronization issues, time consumed in data synchronization and resolving the synchronization related issues.
- The provider has to arrange more storage space and allocate more resources for monitoring and management of replicated data.
- Implements the data de-duplication by preventing the consumers to store replicated data.
- Can be applied to block storage and file based storage.
- Analyzes the received data before sending to storage.
- Each data block is analyzed and a hash code is generated according to the contents.
- The hash code is compared to already stored data blocks
- If a duplicate code is found, the new data block is rejected and a pointer to the already stored block is saved instead.
- The new blocks are saved after the hash code check.
- Can also be applied to backup storage devices.

Module No – 136:Elastic Network Capacity Architecture

- Network bandwidth limit may inhibit the performance and may become a bottleneck.
- It is the software which implements the dynamic scalability of network bandwidth.
- The scalability is provided on per user basis.
- Each user is connected to a separate network port.
- Automated scaling listener, elastic network capacity controller and a resource pool of network ports are used for implementation.
- The automated scaling listener monitors the network traffic and indicates the elastic network capacity controller to enhance the bandwidth and/or number of ports when required.
- When applied to virtual switches, then each virtual switch is configured to induct more physical uplinks.
- Alternatively, the direct I/O can be used to enhance network bandwidth for any VM.

Module No – 137: Cross Storage Device Vertical Tiering Architecture

- The approach is to dynamically shift the logical unit number to another storage device with larger capacity in terms of number of requests processed per second and the amount of data being handled.
    - As compared to traditional approach, it is not constrained by the availability of free space on the physical storage device hosting the logical unit number.
    - Automated scaling listener and storage management modules are required for the implementation.
    - The automated scaling listener monitors the number of requests being sent to the logical unit numbers.
    - When a pre-set threshold of number of requests to a logical unit number is reached, the automated scaling listener signals the storage management module to shift that logical unit number to another device with higher capacity.
    - While moving a logical unit number, the connectivity/availability of data is not interrupted.

Module No – 138: Intra Storage Device Vertical Data tiering Architecture

- Required when there are security and/or legal constraints regarding data migration across different storage devices.
- The data is stored over logical unit numbers.
- This is the implementation of vertical scaling capability over a single cloud storage device.
- The single storage device optimally uses different disks with varying features/capacities.
- Different disks are graded and marked according to capacity.
- Implemented through automated scaling listener and storage management software.
- The automated scaling listener monitors the logical unit numbers.

- A logical unit number is hosted over a disk. The grade of the disk may be chosen randomly or according to a policy.
- Upon rise of performance requirements for a logical unit number, the automated scaling listener signals the storage management program to move the logical unit number to a disk with higher grade.

Module No – 139: Load Balanced Virtual Switches Architecture

- Situations may arise when the consumers' requests may face unlimited delays and packet loss due to network congestion between physical host and network device on uplink.
- The main reason of data packet loss and delays are usually due to single physical uplink.
- This architecture implements a network load balancing mechanism for a physical server (and virtual switch hosted over that server) through the use of multiple physical uplinks.
- This means that more than one NICs are used for each physical server.
- By using *link aggregation* techniques, network traffic can be distributed among multiple physical uplinks.
- The phenomenon of network bottleneck due to single physical uplink can be avoided.
- Can also maintain the availability of the VMs even if any uplink fails.
- The virtual switch has to be configured for compatible and seamless use of multiple physical NICs.

Module No – 140: Multipath Resource Access Architecture

- It is the implementation of multiple access routes/paths to and from a Cloud IT-resource.
- The need for multiple paths arises to provide resiliency when a physical link fails.
- Suppose a Cloud storage device with multiple logical unit numbers deployed over it is connected to (for example) a physical server (hosting multiple VMs with logical unit numbers hosted on that cloud storage).
- If the link to shared storage fails, all the VMs will crash.
- Calling for failover system may take a some time if there are multiple VMs.
- Executing the failover system (when the physical server has not crashed) is expensive.
- An alternative method for resiliency is to provide multiple paths between the physical server and the cloud storage device.

Module No – 141: Persistent Virtual Network Configuration Architecture:

- When a VM is instantiated on a physical host, the network configuration such as the allocated port number (from virtual switch) is set up.
- If the VM is migrated to another physical server, the same port number (of virtual switch of the destination server) may not be available.

- This will jeopardize the network traffic for the migrated VM because the destination environment does not have the network configuration and port number info regarding the migrated VM.
- This architecture ensures that every VM migration uses a persistent information regarding network configuration and port number.
- Implemented through a central virtual switch spanning over multiple physical servers.
- The network configuration and port setting of the VMs (hosted over servers connected through central virtual switch) is centrally stored.
- Each sever is allocated some virtual ports.
- Migration of a VM from one host to another keeps the virtual port number persistent. Thus the connectivity of VM is not lost during and after the migration.

Module No – 142: Redundant Physical Connection for Virtual Servers Architecture

- Redundant hardware devices ca be added to a physical server to add resiliency.
- Working in active-passive manner, the redundant device is kept in a waiting state.
- If the primary device fails, the secondary device takes over.
- This architecture implements redundant NICs to provide high availability and connectivity of the VMs hosted on physical server.
- The redundant NIC is connected to the physical switch through separate links.
- The virtual switch is configured to use all the redundant NIC.
- But only one NIC is kept primary and active.
- The secondary NIC does not forward any packets although it receives packets from VMs until the primary NIC fails.
- The process is transparent to the hosted VMs.

Module No – 143: Storage Maintenance Window Architecture

- The Cloud storage devices needs to undergo for maintenance process in order to maintain their working potential.
- A Cloud storage device hosts multiple logical unit numbers.
- It is not practical to disconnect the storage device/s and then perform maintenance.
- In order to maintain the availability of data, this architecture temporarily copies the data from a to-be-maintained storage device to a secondary device.
- The data is (for example) arranged/stored in the form of logical unit numbers which in-turn are connected to different VMs and/or accessed by different consumers.
- It is therefore important that the logical unit numbers be migrated live.
- The connectivity and availability of data are maintained.
- Once the data is migrated, the primary device is made unavailable. The secondary device serves the data requests even during migration.
- The storage service gateway forwards the consumer requests to secondary storage.

- The data is moved back to the primary storage after the maintenance is over.
- The whole process remains transparent.

Lesson No. 28

CLOUD FEDERATION & BROKERAGE

Module No – 144: Cloud Federation:

- VMs It is the interconnection of Cloud computing infrastructures of two or more Cloud providers for load balancing.
- One of the providers buys the services from the other provider.
- The federation agreement may be timely or permanent.
- It provides revenue for the seller and allows the buyer to extend its resources capacity without setting up or acquiring new hardware resources.
- The VMs and data can be migrated across the providers' Clouds.
- The process of Cloud federation remains transparent to the consumers of the buyer.
- Federation can be performed horizontally or vertically on the basis of extending the SaaS, PaaS and IaaS of the federation buyer.
- The SLA of the consumers (of buyer) is followed over the seller's infrastructure as well.

Module No – 145:Workload Placement in Federated Clouds

- Due to the availability of a finite number of physical resources, a single Cloud can handle a certain number of consumers' requests in a unit time.
- We are supposing that a time deadline exists to process a consumer's request.
- If a Cloud infrastructure cannot meet the requests' deadlines, then it is experiencing resource shortage or *congestion.*
- At this point, the chances of SLA violation start becoming solid.
- The Cloud provider may be heading towards SLA penalties if the situation persists.
- A decision has to be made by the Cloud provider to process the consumer requests that are in excess to the current capacity on the basis of:
- Revenue to be earned from processing the extra requests
- The cost to be paid to other provider/s
- The deadline of the requests vs. latency of remote provider
- A Cloud federation may also be created to fulfill the requests of a remote consumer through the closest provider in that region to reduce network latency.
- Thus federation of Clouds offer a better solution to resource shortage and latency issues in Cloud computing.
- Federation can be horizontal. In this, the Cloud services (IaaS, PaaS and SaaS) are horizontally expanded.

- In vertical federation, a Cloud provider *A* (for example) may host a SaaS/PaaS instant of another provider *B* over its own IaaS to fulfil the requests of provider *A*.
- Federation can also be hybrid.

Module No – 146: Cloud Brokerage:

- It is a third party process of finding the appropriate Cloud provider.
- This process is dependent upon the requirements and/or directions of the consumers.
- Performed by a person or individual called *broker*. Acting as intermediary between the consumer and provider of Cloud services.
- Saves the consumer from spending efforts and time to search for Cloud provider.
- Works closely with the consumer to understand the business process, provisioning needs, budget and other requirements.
- The broker then searches for the best options of Cloud providers for the consumer.
- The consumer is then provided with a list of providers according to requirements and budget of the consumer.
- The broker may also be granted rights to negotiate with the providers on behalf of the consumer. The broker can even settle a contract with the most suitable provider.
- A broker may also provide:
    - API and GUI to the consumer for handling the Cloud resources.
    - Encryption and data transfer & management assistance.
    - Advisory service to the consumer for improving the use and benefits of the Cloud.

Lesson No. 29

CLOUD DELIVERY/SERVICE MODELS' PERSPECTIVES

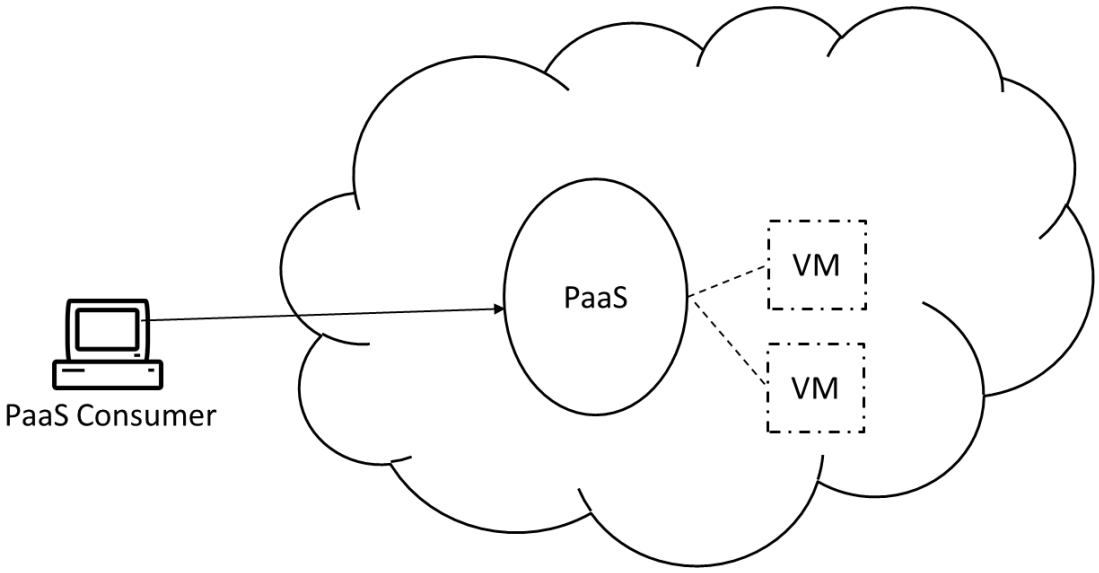
Module No – 147: Cloud Provider's Perspective about IaaS:

- In this and next two modules, we shall discuss the overall perspective of Cloud provider in establishing and managing of Cloud services. Namely:
    - IaaS
    - PaaS
    - SaaS
- The two basic IT resources of IaaS are:
    - VMs
    - Cloud storage
- These are offered along with the:
    - OS
    - (virtual) RAM
    - (virtual) CPU
    - (virtual) Storage
- VMs are usually provisioned through VM images which are predefined configurations.
- Bare-metal provisioning is also provided to the consumers with administrative access.
- Snapshots of VMs can be occasionally taken for failover and replication purposes.

- A cloud may be provisioned through multiple data centers spanning at different geographical locations and connected through highspeed networking.
- VLANs and network access control are used to isolate a networked set of VMs (into a network perimeter) which are provisioned to a single consumer/organization.
- Cloud resource pools and resource management systems can be used to provide scalability.
- Replication is used to ensure high availability and forming a failover system.
- Multipath resource access architecture is used to provide reliability.
- Resource reservation architecture is used for provisioning of dedicated IT resources.
- Different monitors such as pay-per-use monitor and SLA monitors continuously overlook VM lifecycles, data storage and network usage to establish billing system and SLA management.
- Cloud security (encryption, authentication and authorization systems) are to be implemented.

Module No – 148: Cloud Provider's Perspective about PaaS:

- PaaS instances are hired by developers who want to develop Cloud applications.
- Readymade environments are usually created to provide on-demand access to a pre-configured set of software tools and SDK.
- The PaaS environments can simulate a Cloud environment with security procedures to enable the developer to test the application/s being developed.
- The PaaS environments also help in establishing multitenancy and auto scalability features in developed applications.
- Scalability can be provided to an overloaded application on the recommendation and budget of the PaaS consumer.
- Automated scaling listener and load balancers are utilized for workload distribution.
- Non-disruptive service relocation architecture and failover systems are utilized to ensure reliability in PaaS instances.
- PaaS instances may comprise of multiple VMs and can be distributed across different data centers.



PaaS instances may comprise of multiple VMs and can be distributed across different data centers.

- Pay-per-use monitor and SLA monitor can be used to collect data regarding resource usage and failures.
- The security features of IaaS are usually ample for PaaS instances.

Module No – 149: Cloud Provider's Perspective about SaaS:

- SaaS instances are unique from IaaS and PaaS instances due to the existence of concurrent users.
- The SaaS implementations depend upon scalability& workload distribution mechanisms and non-disruptive service relocation architectures for smooth provisioning and overcoming failures.
- Unlike the IaaS and PaaS, every SaaS deployment is unique from other implementations.
- Every SaaS deployment has different programming logic, resource requirements and consumer workloads.
- The diverse SaaS deployments include: Wikipedia, Google talk, email, Android play store, Google search engine etc.
- The implementation mediums include:
  - Mobile apps
  - REST service
  - Web service
- These mediums also provide API calls. The examples include: electronic payments services such as PayPal, mapping and routing services (Google Maps) etc.
- Mobile based SaaS implementations are usually supported by multi-device broker mechanism for heterogeneous device-based access.
- Therefore, SaaS implementation requires the implementation of:

- Service load balancing, Dynamic failure detection and recovery, storage maintenance window, elastic resource/network capacity and Cloud balancing architectures.
- Monitoring is usually performed through pay-per-use monitors to collect consumer usage related data for billing
- Additional security features (as already provided by underlying IaaS environment) may be deployed according to business logic.

Module No – 150: Cloud Consumer's Perspective about IaaS

- A consumer accesses the VM through a remote terminal application. The VM has to have an OS installed.
  - Remote desktop client for Windows
  - SSH client for Mac and Linux based systems
- Cloud storage device can directly be connected to the VM or to a local device on-premises.
- The Cloud storage data can be handled and rendered through Networked file system, storage area network and/or object-based storage accessible through Web-based interface.
- The administrative rights of the IaaS consumer include, controlling of:
  - Scalability
  - Life cycle of VM (powering-On/Off and restarting)
  - Network setting (firewall and network perimeter)
  - Cloud storage attachment
  - Failover setting
  - SLA monitoring
  - Basic software installations (OS and pre installed software)
  - VM initializing image selection
  - Passwords and credentials management for Cloud IT-resources
  - Costs
- IaaS resources are managed through remote administration portals and/or command line interfaces through execution of code scripts.

Module No – 151: Cloud Consumer's Perspective about PaaS

- Typically, a PaaS consumer receives the following:
  - Software libraries
  - Class libraries
  - Frameworks
  - APIs
  - Databases
  - Cloud emulation environment
- The completed applications are deployed to Cloud
- The administrative rights of the PaaS consumer include the control of:
  - Login management of service/s developed/deployed using PaaS instance
  - Choosing the tools in case of ready-made environment
  - Cloud storage device selection

- IT-resource usage cost
- Deployment of automated scaling listener, load balancer and replication etc.
- SLA monitoring

Module No – 152:Cloud Consumer's Perspective about SaaS

- The SaaS deployments are complemented by APIs.
- The API calls enable the wide spread of the SaaS into webpages and applications. For example Google maps
- Many SaaS are free of charge. Although the service provider may gather background data
- The SaaS consumers have least administrative privileges and the least responsibilities regarding deployment and managing the service.
- A few runtime configurations can be controlled by SaaS consumers. These include:
  - Usage cost control
  - SLA monitoring
  - Security related configurations

Lesson No. 30

INTER-CLOUD RESOURCE MANAGEMENT

Module No – 153:

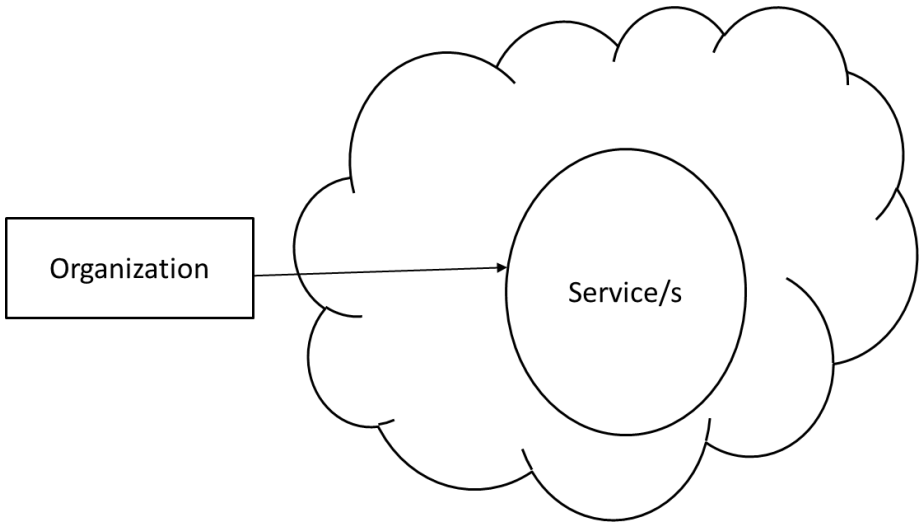
- The term **Inter-Cloud** refers to as *Cloud of Clouds* just as **Internet** is regarded as *network of networks*.
- Cloud computing has proliferated throughout the computing world.
- The providers are of two types:
- With extra (idle) resources
- With resource shortage
- Many providers look for getting reasonable clients to generate more revenue and to make good use of idle resources.
- Cloud federation gives a solution to this problem.
- But a bigger picture lies in Inter-Cloud where the global federation takes place.
- The Inter-Cloud can be established where each member Cloud is connected to other member Clouds just like Internet connects the networks.
- It is the ultimate future of Cloud federation.
- Technological giants such as IBM, HP, CISCO, RedHat etc. are actively working on establishment of cloud-of-clouds.
- We hope that soon the issues of interoperability, inter-cloud communication, security and workload migration will be addressed.

Lesson No. 31

CLOUD COST METRICS AND PRICING MODELS

Module No – 154:

- In next few modules, we shall discuss different cost metrics and pricing models of Cloud.
- **Business Cost Metrics:** The common types of metrices related to cost benefit analysis of Cloud computing.
- **Upfront Costs:** Related to initial investment regarding IT resource acquiring and installations.
  High costs for on-premises installation as compared to leased from Cloud.
- **On-going Costs:** Include the running costs of the IT resources e.g., licensing fee, electricity, insurance and labor.
  The long term ongoing costs of Cloud IT-resource can exceed the on-premises costs.
- **Additional Costs:** These are specialized cost metrics. These may include:
  - **Cost of Capital:** It is the cost of raising a capital amount. It is higher if a high capital is to be arranged in short time. The organization may have to bear some costs in raising a large amount. This is important decision for up-front cost metrics.



Organizational perspective of Cloud services.

  - **Sunk Costs:** These are the costs already spent by the organization over IT-infrastructure. If the Cloud is preferred then these costs are sunk. Hence should be considered along with up-front cost of Cloud. Difficult to justify the leasing of Cloud IT resources in the presence of high sunk costs.

- **Integration Costs:** The time and labor costs required to integrate a Cloud solution which include the testing of Cloud services acquired.
- **Locked-in Costs:** The costs related to being dependent upon a single Cloud provider due to lack of interoperability among different providers. Affects the business benefits of leasing the Cloud based IT-resources.

Module No – 155:

- **Cloud Usage Cost Metrics:** In this module we shall study different metrics related to cost calculation of Cloud IT resource usage.
  - *Network Usage:* Cumulative of, or separate of outbound and inbound network traffic in bytes over the monitored time. Costing may be cumulative or separate for inbound and outbound traffic. Many Cloud providers do not charge for inbound traffic to promote the consumers to shift their data towards Cloud.
    - May also be based upon the *static IP address usage* and network traffic processed by *Virtual Firewall.*
  - *VM Usage:* Related to the number of VMs and the usage of the allocated VMs. Can be static cost, pay-per-use, or according to the features of VM. Applicable to IaaS and PaaS instances.
  - *Cloud Storage Device Usage:* It is charged by the amount of storage used. Usually the *on-demand storage allocation* pattern in used to calculate bill on time basis for example on hourly basis. Another(scarcely used) billing option is to charge on the basis of I/O operations to and from storage.
  - *Cloud Service Usage:* The service usage can be charged on the basis of duration of subscription, number of nominated users and/or number of transaction served by the service.

Module No – 156: Case Study for Total Cost of Ownership (TCO) Analysis

- The TCO includes the costs of acquiring, installing and maintaining the hardware and software to perform the IT tasks of the organization.
- In this module, we shall perform a case study to evaluate the TCO for on-premises and Cloud based solution.
- Suppose a company wants to migrate a legacy application to PaaS. The application requires a database server and 4 VMs hosted on 2 physical servers.
- Next we perform a TCO analysis for 3 years:

Upfront Cost	Cloud Environment (Rupees)	On-Premise Environment (Rupees)
Hardware	0	190,000
Licensing	0	200,000
Labor	40,000	40,000
Total Up-Front Costs	40,000	430,000

Follow the video lecture to understand fully

Monthly On-Going Costs	Cloud Environment (Rupees)	On-Premise Environment (Rupees)
Application Servers	120,000	0
Database Servers	30,000	0
WAN Network	4,000	0
Environment	0	40,000
Software Licensing	0	30,000
Hardware Maintenance	0	15,000
Administration	30,000	80,000
Total On-Going Costs	184,000	165,000

Follow the video lecture to understand fully

	Cloud Environment (Rupees)	On-Premise Environment (Rupees)
Total Upfront Cost	40,000	430,000
3 Months On-Going Costs	(184000 x 3) = 552,000	(165,000 x 3) = 495,000
Gran Total	**592,000**	**925,000**

Follow the video lecture to understand fully

Module No – 157: Cost Management Considerations

- Cost management can take place across the lifecycle phases of Cloud services. These phases may include:
  - Design & Development
  - Deployment
  - Service Contracting
  - Provisioning & Decommissioning
- The cost templates used by the providers depend upon:
  - Market competition
  - Overhead occurred during design, deployment and operations of the service
  - Cost reduction considerations through increased sharing of IT resources
- A pricing model for Cloud services can be composed of:
  - Cost metrics
  - Fixed and variable rates definitions
  - Discount offerings
  - Cost customization possibilities
  - Negotiations by consumers
  - Payment options

Module No – 158:

- **Case study:** We shall now see and example case of different price offering from a Cloud provider.

Instance Name	Configuration	Operating System	Pay-per-use Hourly Rate Rs.
Small VM	1 vCPU, 4GB vRAM, 20GB Storage	Linux	50
		Windows	80
Medium VM	2 vCPUs, 8GB vRAM, 20 GB Storage	Linux	145
		Windows	190
Large VM	8vCPUs, 16GB vRAM, 80GB Storage	Linux	300
		Windows	350
VM with Large VRAM	8vCPUs, 64GB vRAM, 20GB Storage	Linux	800
		Windows	900
VM with Large vCPU	32vCPUs, 16GB vRAM, 20GB Storage	Linux	900
		Windows	1000

Follow the video lecture to understand fully

Instance Name	Configuration	Operating System	I Year Term Pricing Rs.	
			Up-front	Pay-per-use Hourly
Small VM	1 vCPU, 4GB vRAM, 20GB Storage	Linux	500	30
		Windows	700	45
Medium VM	2 vCPUs, 8GB vRAM, 20 GB Storage	Linux	1200	75
		Windows	1500	90
Large VM	8vCPUs, 16GB vRAM, 20GB Storage	Linux	3000	172
		Windows	3500	200
VM with Large VRAM	8vCPUs, 64GB vRAM, 20GB Storage	Linux	700	400
		Windows	800	450
VM with Large VRAM	32vCPUs, 16GB vRAM, 20GB Storage	Linux	750	400
		Windows	850	450

Follow the video lecture to understand fully

Lesson No. 32

CLOUD SERVICE QUALITY METRICS

Module No – 159:

- These metrics are used to define and monitor the SLA.
- The characteristics of Quality of Service (QoS) can be expressed by these metrices.
- These include:
    - *Availability*: up-time, down-time, service duration
    - *Reliability*: minimum time between failures, guaranteed rate of successful response
    - *Performance*: capacity, response time and delivery time guarantees
    - *Scalability*: Capacity fluctuation and responsiveness guarantees
    - *Resiliency*: Mean time to switchover and recovery
    - Service quality metrics are required to be *quantifiable, repeatable, comparable* and *easily obtainable*

Module No – 160: Service Availability Metrics

- *Availability Rate Metric*:
    - The value in % of up-time e.g., 100%.
    - Measured as total up-time/total time.
    - Monitored weekly, monthly and/or yearly.
    - Applied to IaaS, PaaS and SaaS.
    - Expressed as cumulative value.
    - E.g., 99.5% minimum
- *Down-time Duration Metric*:
    - Expresses the maximum and average continuous down-time.
    - Covers the duration of outage.
    - Measured whenever the outage event occurs.
    - Applied to IaaS, PaaS and SaaS.
    - E.g., 1 hr max, 15 min average

Module No – 161: Service Reliability Metrics

- *Reliability* in context to Cloud IT-resources refers to the probability that an IT-resource can be performing its intended function under predefined conditions without experiencing failure.
    - Focuses on the duration in which the service performs as expected.
    - This requires the service to be operational and available during that time.
- *Mean-Time Between Failures Metric*:
    - Expected time between two consecutive failures.

  - Measured as normal operation duration/number of failures.
  - Measured as monthly and/or yearly.
  - Applicable to IaaS and PaaS.
  - E.g., 90 days average
- *Service Reliability Rate Metric*: It is the percentage of successful service outcomes.
  - Measures the non-critical errors during the up-time.
  - Measured as total number of successful responses/total number of requests
  - Measured as weekly, monthly and/or yearly.
  - Applicable to SaaS.
  - E.g. minimum 99.5%

Module No – 162: Service Performance Metrics

- *Service performance* refers to the ability of an IT resource to carryout its functions within expected perimeters.
  - Measured with respect to capacity metrics.
- *Network Capacity Metric*: Measured as bandwidth/throughput in bits per second. Applicable to IaaS, PaaS and SaaS. Measured continuously. Expressed as for example MB/sec.
  - *Storage Device Capacity Metric*: Measured as size in GB. Applicable to IaaS, PaaS and SaaS. Continuously measured.
- *VM Capacity Metric*: Measured as features such as number of CPUs, CPU frequency in GHz, RAM size in GB and storage size in GB. Continuously measured. Applied to IaaS and PaaS.
  - *Web Application Capacity Metric:* The number of requests processed in a minute (for example). Applicable to SaaS

Module No – 163: Service Scalability Metrics

- These are related to the IT resource's elastic capacity, the maximum capacity that an IT resource can reach and the adaptability of an IT resource to workload fluctuations.
- For example a VM can be scaled up to 64 cores and 256 GB of RAM or can be scaled out to 8replicated instances.
- *Storage Scalability (Horizontal) Metric*: The permissible capacity change of a storage device in accordance with the increase in workload.
  - Measured in GB.
  - Applicable to IaaS, PaaS and SaaS.
  - E.g., 1000 GB maximum (automatic scaling)
- *Server Scalability (Horizontal) Metric*: The permissible server capacity in response to increased workload.
  - Measure in number of VMs in resource pool.
  - Applicable to IaaS, PaaS
  - E.g., 1 VM minimum up to 10 VMs maximum (automated scaling)
- *Server Scalability (vertical) Metric*: Measured in terms of number of vCPUs, vRAM size in GB.
  - Applicable to IaaS and PaaS.
  - E.g., 256 cores maximum and 256 GB od RAM

Module No – 164: Service Resiliency Metrics

- Refers to the ability of an It resource to recover form operational disturbance.
- When considered with respect to SLA, the resiliency guarantees are based upon redundant implementation, resource replication and disaster recovery systems.
- Can be applied over three phases:
    - *Design phase:* To evaluate the preparation level of systems and service to cope with the known challenge.
    - *Operational phase*: Measure the difference in service levels (in terms of availability, reliability, performance and scalability metrics) before, during and after a downtime event.
    -
- *Recovery phase*: To measure the rate at which an IT resource recovers from downtime. For example the meantime for a system to record a downtime event and switch over to a new VM.
- Two common metrics related to measuring resiliency are as follows:
    - *Mean-Time to Switchover (MTSO) Metric*: The time to switch over to a replicated instance after the failure of a VM.
        - Measured in terms of time (from time of failure to recovery)
        - Measure as per month and/or year
        - Applicable to IaaS and PaaS
        - E.g., 12 minutes average
    - *Mean-Time System Recovery (MTSR) Metric*: The time expected for a resilient system to perform a complete recovery from a VM failure.
        - Measured as total time spent during recovery/total number of failures
        - Measured as monthly and/or yearly
        - Applicable to IaaS, PaaS and SaaS
        - E.g., 100 minutes average

Module No – 165: Service Quality Metrics and SLA Guidelines

- In this module, we shall discuss some of the best practices of Cloud consumers for dealing with SLAs.
- **Mapping of test-cases to the SLAs:** A consumer should highlight some test cases (disasters, performance, workload fluctuations etc.) and evaluate the SLA accordingly. The SLA should be aligned with the consumer's requirements of the outcome of these test-cases.
- **Understanding the scope of SLA:** A clear understanding of the scope of SLA should be made. It is possible that a software solution may be partially covered by an SLA for example the database may be left uncovered.
- **Documenting the guarantees:** It is important to document all the guarantees at proper granularity. Any particular guarantee requirement should also be properly and clearly mentioned in SLA.

- **Defining penalties:** The penalties and reimbursements should be clearly defined and documented in SLA.
- **SLA Monitoring from independent party**: Consider the SLA monitoring from a third party.
- **SLA monitoring data archives**: The consumer may want the provider to delete the monitored data due to privacy requirement. This should be disclosed as an assurance by the provider in SLA.

Lesson No. 33

CLOUD SIMULATOR

Module No – 166:CloudSim: Introduction

- Cloud computing is an ongoing phenomenon which requires periodic updates in terms of architectures (study before), procedures and services.
- Cloud computing is an ongoing phenomenon which requires periodic updates in terms of architectures (study before), procedures and services.
- For a lot of researchers, it is difficult to test their research and theories on a Cloud computing setup hosted on a real data center.
- It is more suitable to use a simulation environment such as CloudSim for testing various new and updated procedures and hypothesizes related to Cloud computing.
- CloudSim is free of cost.
- Can be used by IaaS/PaaS users to tune up performance bottlenecks in the Cloud services before deploying on actual Cloud.
- Can be used by the provider to test various policies related o resource leasing, pricing and workload management of the Cloud.
- The documentation, setup and tutorials are available free of cost from the CloudSim webpage: http://www.cloudbus.org/cloudsim/

Module No – 167: CloudSim: Configuration

- Some configurations are required for the CloudSim. The important requirements are discussed in this module.
- CloudSim requires Sun's Java 8 or newer version. Older versions of Java are not compatible.
- You can download Java for desktops and notebooks from https://java.com/en/download/
- CloudSim requires Sun's Java 8 or newer version. Older versions of Java are not compatible.
- You can download Java for desktops and notebooks from https://java.com/en/download/
- CloudSim setup is just needed to be unpacked before using. If you want to remove CloudSim, remove the folder.
- CloudSim setup comes with various coded examples which can be test run for understanding the CloudSim architecture.
- CloudSim site has video tutorial explaining the step-by-step configuration and execution.

- CloudSim setup is just needed to be unpacked before using. If you want to remove CloudSim, remove the folder.
- CloudSim setup comes with various coded examples which can be test run for understanding the CloudSim architecture.
- CloudSim site has video tutorial explaining the step-by-step configuration and execution.

Module No – 167:CloudSim: Example Code

- CloudSim emulates a virtualized data center environment.
- There are various hosts (physical servers) and *VMs* mounted on the hosts.
- The workload unit is called *Cloudlet*.
- There is a *broker* entity which gets the Cloudlets executed on Data center just as in real life scenarios.
- A service named as *Cloud Information Service (CIS)* contains the registry of the data center.
- The broker object retrieves the information of all the data centers registered with CIS.
- The broker then contacts a suitable data center and executes the Cloudlet/s over the hosted VMs.
- Each host has properties such as processors and RAM. The hosted VMs share the host.
- Each VM can have multiple Cloudlets scheduled over it.
- Different *policy* objects such as VM-allocation, VM-scheduler and Cloudlet-scheduler exist.
- *VM-allocation policy* is used by data center to allocate VMs to hosts.
- *VM-scheduler policy* is used by host for resource allocation.
- *Cloudlet-scheduler policy* is used by the VM.
- These policies are either space shared or time shared.

Lesson No. 34

COMPUTER SECURITY BASICS

Module No – 169: Computer Security Overview:

- As per NIST, the **Computer Security** is defined as:
- *"The protection afforded to an automated information system in order to attain the applicable objectives of preserving the integrity, availability and confidentiality of information system resources (includes hardware, software, firmware, information/data and telecommunications)* "
- **Information System:** It is a software that helps in organize and analyze data.
- **Privacy:** *The right of an individual to control or influence about what information can be collected and stored and by whom and to whom that information may be disclosed.*

"William Stallings [2013] Computer Security Principles and Pactices, Pearson"

- The European Union (EU) definition of *personal data* is:

- *Personal data shall mean any information relating to an identified or identifiable natural person; an identifiable person is one who can be identified, directly or indirectly, in particular to an identification number or to one or more factors specific to his physical, psychological, mental, economic, cultural or social identity.*

  European Commission: Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.

- The key terminologies of privacy are:
  - *Data controller*: An individual or a body which individually or jointly determines the purpose and procedure of processing an item of personal information.
  - *Data processor*: An individual or body which processes the personal information on behalf of the data controller
  - *Data subject:* An identified or identifiable individual to whom personal information relates directly or indirectly.
- Privacy is regarded as human right in Europe while in America it traditionally refers to avoiding harm to people in any context.

Module No – 170: Confidentiality Integrity and Availability (CIA):

- **Confidentiality**: Allowing only authorized access and disclosure to the information
  - For example: The student grade information is the asset of students and confidentiality should be imposed by law in the USA. This information is only available to the students, their parents and employees that require the information for routine job purposes. Lesser level of confidentiality is the enrolment information.
- **Integrity**: Guarding against unauthorized modification and/or destruction of information.
  - For example: The information related to patients' allergy requires high integrity because any inaccuracy in this information may lead to serious harm or even death. While a low integrity is required for an online poll where people may choose inaccurate options or reviews.
- **Availability**: Ensuring timely and reliable access to the information.
  - For example: A service providing authentication verification is needed to be highly available otherwise the users/employees can not access the data and organization will suffer monitory losses due to the halt in data processing operations. An online telephone directory may not be required to e highly available.
- **Authentication:** Proving the identity of a user e.g.; through login and password.
- **Authorization**: Verification of the access rights of an authenticated user e.g.; subscription to basic or premium user access to an online gaming website.

Module No – 171: Computer Security & Trust:

- Trust is a psychological state comprising of intensions to accepts the risks on the basis of positive expectations of the behavior of another person or entity.
- Trust is broader term than security because the trust is also based upon experience and criteria.
- Trust has two types:
    - Hard trust: Requires the usage of security-oriented aspects such as authentication, encryption and security (CIA).
    - Soft trust: Consists of non-security oriented phenomenon such as human psychology, brand loyalty and user friendliness.
- Usually people find it harder to trust online services than offline services.
- Trust on online services can be enhanced/revived by using security features but it is not a guaranteed solution.
- The trust in Cloud computing is of two types: *Persistent trust* (long term) and *Dynamic trust* (short term).
- The trust of Cloud consumer can be enhanced and established through security elements.
- More to come in next modules.

Module No – 172: Cryptography:

- It is a science of providing security for information.
- It is a science of secret communication.
- Has been historically used by the governments and armies for as long ago as 1900 BC.
- It converts the data into a format which is not readable by an un-authorized user.
- *Cryptanalysis*: The science of analyzing and breaking the code of encrypted text.
- *Cryptology*: Involves both cryptography and cryptanalysis.
- The following are the five primary functions:
    - *Privacy*
    - *Authentication*
    - *Integrity*
    - *Non-repudiation*: A mechanism to prove the originality of the sender.
    - Exchange of *crypto keys* which are the strings of bits used to change the format of the data.

Module No – 173: Authentication & Access Control:

- Authentication can be performed through:
- Something the user knows: password, personal identification number (PIN) or answer to a prearranged question.
- Something the user posses: Electronic cards, smart card and physical keys. Also called *token.*

- Something the individual is (static biometric): Fingerprint, retina and face scan.
- Dynamic biometric: voice, handwriting and typing pattern.
- The purpose of access control is to limit the actions or operations that an authenticated user of a computer system can perform.
- This includes the privileges of the user as well as the programs executing on behalf of the user.
- Access control is enforced by a software module which monitors every action performed by the user and the programs executing on behalf of the user.
- The authorization of each user is set by the security administrator according to policy of the organization.
- Access control requires authentication.

Module No – 174: Malware or Malicious Software:

- It is a program which is inserted into a system (usually covertly) to compromise the confidentiality, integrity and/or availability of the victim's data.
- *Adware*: Advertisement integrated into the software. It can result in pop-up ads or redirecting of browser.
- *Attack kit*: Set of tools to generate more malware.
- *Auto-rooter*: Hacking tool to remotely break into a machine.
- *Backdoor (trapdoor):* A mechanism that bypass the normal security check and may allow unauthorized access.
- *Flooders (Denial of Service client):* Generates and propagate a flood of packets over a network to perform denial of service attack.
- *Keyloggers:* Capture the keystrokes on a compromised system.
- *Logic-bomb:* A malicious code which executes on the happening of certain event or time.
- *Rootkit:* Set of tools used by a hacker after compromising a system and gaining root level access.
- *Spammer programs*: Used to send large volumes of junk emails.
- *Spyware:* A software that collects keystrokes, screen data, network traffic or stealing sensitive information from the files and transfer them to remote computers.
- *Trojan horse:* A look-as-legitimate program with hidden malicious contents.
- *Virus*: A malware that replicates itself over to other computers in contact with an infected computer.
- *Worm:* An independent program that replicates itself over networked computers and compromises the security.
- *Zombie/bot*: A program which is activated over remote system to make a team attack over a victim computer.

Module No – 175: Denial of Service (DoS) Attacks:

- It floods the servers, systems and networks with traffic.
- Makes it impossible for legitimate users to work on the affected IT-resources.
- Difficult to recover from DoS attack. Restarting is also not helpful most of the times.
- Executed not for ransom but to cause harm to the victim.
- The United States Computer Emergency Readiness Team (US-CERT) defines the following symptoms:
    - Degradation in network performance
    - Inability to reach a website
    - Higher than usual volumes of spam email.
- Remedies for DoS attacks:
    - Contact ISP to clarify the reason of downgraded network performance.
    - ISP ca help in throttling malicious traffic.
    - Using DoS detection tools.
- Usual victims: Application servers, DNS servers
- Can be in the form of TCP handshake flood, packet flooding with overloaded payload etc.

Module No – 176: Intrusion detection & Firewalls:

- A firewall is a hardware and/or software based module to block unauthorized access (but allowing authorized access) in a networked environment.
    - Stands between a local network and Internet.
    - Filters the harmful traffic.
- Firewall preforms packet filtering on the base of source/destination IP address.
- Firewall checks the packets on the basis of connections (stateful firewall).
- Other types of firewalls also exist.
- Intrusion detection system (IDS) is a software or hardware device installed on a network or a host to detect intrusion attempts, monitors malicious activity or policy violations.
- Intrusion detection system (IDS) is a software or hardware device installed on a network or a host to detect intrusion attempts, monitors malicious activity or policy violations.

Module No – 177: Buffer Overflow Attacks:

- Allows the attacker to control or crash the process or to modify its internal variables.
- Can be launched through DoS attack.
- Can also occur by chance.

- It occurs when a program attempts to write more data to a block of memory or buffer than the allowed volume.
- The overflowed data is written to the adjacent block/s of the memory. Thus overwriting the adjacent blocks.
- If the adjacent memory buffer is overwritten then the attacker may overwrite a chosen address to a function pointer in that buffer. The chosen address is of a memory location with malicious address.
- Now the function pointer is pointing at the malicious code.
- When the (overwritten) function pointer is executed, the malicious code starts to execute and the attacker gets the system control.
- Can occur wherever direct memory access is allowed such as in C and C++.
- C# and Java have reduced coding errors causing buffer overflow.

Module No – 180: Operating System and Virtualization Security:

- The installation of operating system requires some security measures such as:
- **Planning:** The purpose, user, administrator and data to be processed on that system.
- **Installation:** The security measures should start from the base.
- BIOS level access should be secured and with a password.
- The OS should be patched/updated with latest critical security patches before installing any applications.
- Remove unnecessary services, applications and protocols.
- Configure the users, groups and authentication according to security policy.
- Configure the resource control/permissions. Avoid the default permissions. Must go through all the permissions.
- Install additional security tools such as anti-virus, malware removal, intrusion detection system, firewall etc.
- Identify the white listed applications which can execute on the system.
- **Virtualization Security:** The main concern should be:
    - Isolation of all guest OSs.
    - Monitoring all the guest OSs.
    - Maintenance and security of the OS-images and snapshots.
- Can be implemented through:
    - Clean install of hypervisor from secure and known source.
    - Ensure only the administrative access to hypervisor, snapshots and OS images.
    - The guest OS should be preconfigured to not to allow any modifications/access to underlying hypervisor by the users.
    - Proper mapping of virtual devices over physical devices.

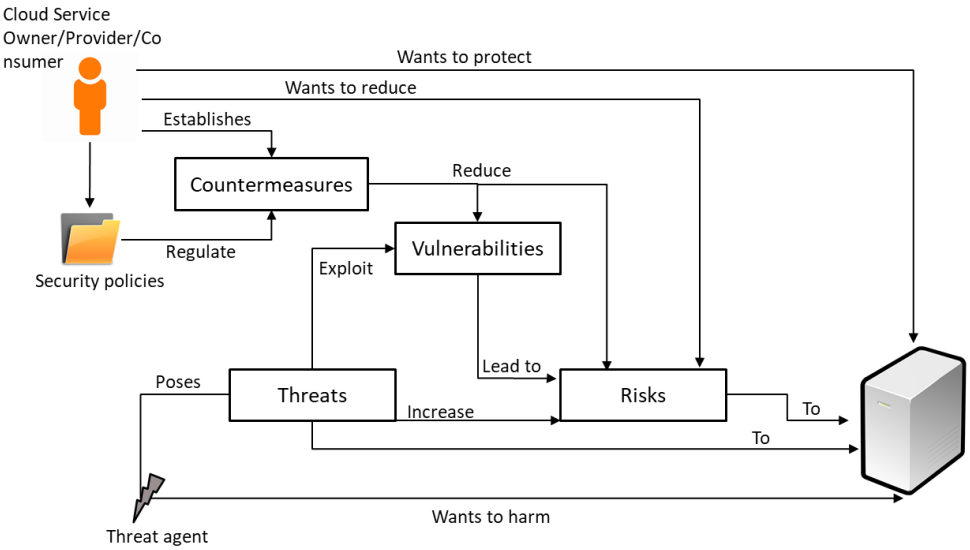
- Network monitoring etc.

Module No – 181: Threat, Vulnerability & Risk:

- **Threat:** It is a potential security breach to affect the privacy and/or cause a harm.
  - Can occur manually and/or automatically.
  - A threat executed results in an *attack*.
  - Threats are designed to exploit the known weaknesses or *Vulnerabilities*.
- **Vulnerability:** It is a (security) weakness which can be exploited.
  - It exists because of:
  - Insufficient protection exists and/or the protection is penetrated through an attack.
  - Configuration deficiencies
  - Security policy weaknesses
  - User error
  - Hardware or firmware weaknesses and software bugs
  - Poor security architecture
- **Risk:** It is a possibility of harm or loss as a result of an activity.
  - Measured according to
    - Threat level
    - Number of possible vulnerabilities
    - Can be expressed as:
    - Probability of occurring of a threat to exploit vulnerabilities
    - The expectation of loss due to compromise of an IT resource

Module No – 182: Threat Agents:

- A threat agent is a factor which is capable of carrying out an attack.
- It can be internal or external and can be human or software.
- **Anonymous Attacker:** A non-privileged service consumer not fully aware of Cloud security measures. Launches network attacks. Can steal user credentials. Can be inhibited by Cloud security measures.
- **Malicious Service Agent:** Can be or acts like a service agent. Has malicious code. Can interpret and forward the network traffic inside Cloud.
- **Trusted Attacker:** Is in the form of legitimate Cloud consumer and launches attacks on other Cloud consumers and the provider to steal information, DoS, hacking of weak authentication processes etc.
- **Malicious Insider:** Typically human threat agent. Can be current or pervious employee. Can cause significant damage with administrative rights.



Follow the video lecture to understand fully

Lesson No. 35

NETWORK SECURITY BASICS

Module No – 178: Internet Security:

- It is a branch of computer security which specifically deals with threats which are Internet based.
- The major threats include the possibilities of unauthorized access to any one or more of the following:
  - Computer system
  - Email account
  - Website
  - Personal details and banking credentials
- Viruses and other malware
- Social engineering
- **Secure Socket Layer (SSL):** It s security protocol for encrypting the communication between a web browser and web server.
  - The website has to enable SSL over its deployment.
  - The browser has to be capable of requesting a secure connection to the websites.
  - Upon request, the website shares its security certificate (issued by a Certificate Authority (CA)) with the browser which the browser confirms for validity.

- Upon confirmation of security certificate, the browser generates the session key for encryption and shares it with website, after this the encrypted communication session starts.
- Websites implementing the SSL use HTTPS (https://...) in the URL instead of HTTP (http://...) and a sign of padlock before the URL.

Module No – 179: Wireless Network Security:

- The wireless network security is applied to wireless networks and is also known as *wireless security*.
- It is used to secure the wireless communication from unauthorized access.
- There are a lot of threats for wireless networks. Such as:
  - The packets can be easily eavesdropped and recorded.
  - The traffic can be modified and retransmitted more easily as compared to wired networks.
  - Prone to DoS attacks at access points (APS).
- Some prominent security protocols for wireless security are:
  - **Wired Equivalent Privacy (WEP):** Designed to provide the same level of security as the wired networks.
    - First standard of 802.11
    - Uses RC4 standard to generate encryption keys of length 40-128 bits.
    - Has a lot of security flaws, difficult to configure and can easily be cracked.
  - **Wi-Fi Protected Access (WPA):** Introduced as an alternative to WEP while a long-term replacement to WEP was being developed.
    - Uses enhanced RC4 through Temporal Key Integrity Protocol (TKIP) which improves wireless security.
    - Backward compatible with WEP.
  - **Wi-Fi Protected Access 2 (WPA2):** Standardized release by IEEE as 802.11i the successor to WPA.
    - Considered as the most secure wireless security standard available
    - Replaces the RC4-TKIP with stronger encryption and authentication methods:
    - Advanced Encryption Standard (AES)
    - Counter Mode with Cipher Block Chaining Message Authentication Code Protocol (CCMP)
    - Allows seamless roaming from one access point to another without reauthentication.

Lesson No. 36

CLOUD SECURITY MECHANISMS

Module No – 183: Encryption:

- The data by default in human readable format called *plaintext*.
- If transmitted over network, the plaintext data is vulnerable to malicious access.

- *Encryption* is a digital coding system to transform the plaintext data into a protected and nonreadable format while preserving the confidentiality and integrity.
- The algorithm used for encryption is called *cypher*.
- The encrypted text is also called *cyphertext*.
- The encryption process uses *encryption key* which is a string of characters. It is secretly created and shared among authorized parties.
- The encryption key is combined with the plaintext to create the encrypted text.
- Encryption helps in countering:
    - Traffic eavesdropping
    - Malicious intermediary
    - Insufficient authorization
    - Overlapping trust boundaries
- This is because the unauthorized user finds it difficult to decrypt the intercepted messages.
- There are two basic types of encryption:
    - **Symmetric Encryption:** It uses single key for encryption and decryption. Also known as *secret key cryptography*. Simpler procedure. Difficult to verify the sender if the key is shared by multiple users.
    - **Asymmetric Encryption:** Uses two different keys (private and public key pair). Also known as *public key cryptography*. A message encrypted with public key can only be decrypted by the respective private key and vice versa.
    - Any party can acquire a public-private key pair. Only the public key is shared publicly.
    - The senders can use the public key of the receiver to encrypt messages. Only the user with corresponding private key can decrypt the message.
    - Successful decryption can ensure confidentiality but does not assure integrity and authenticity of the sender as anyone can encrypt the message using public key.

Module No – 184:

- **Hashing:** It is a process of deriving a hashing code or *message digest* from a message.
    - The message digest is of a fixed length and is shorter than the original message.
    - Uses a hash function to generate the hashing code.
    - A change in message requires the hashing code to be regenerated.
    - The hashing code is attached with the message and sent to the receiver.
    - The receiver applies the same hash function to verify the integrity of the message.
    - If the message was altered during transmission, the receiver side hashing code computation will mismatch the hashing code received with the message. The receiver rejects such messages.
- **Digital Signatures:** It is a mechanism of verifying the authenticity and integrity of a message, software and/or digital comment.
    - It is a digital equivalent of handwritten signature. Used to prevent the tampering and impersonation in digital communication.
    - In many countries, the digital signatures have a legal value.
    - The hashing function is applied to the original message to generate a message digest.

- The message digest text is changed through cryptographic mechanism known only by the sender and receiver.
- The encrypted hash code and hashing algorithm is the digital signature.
- Alteration can be detected at receiver end.
- The administrative tools used by Cloud consumers use the digital signatures with every request to prove the authenticity of each consumer.

Module No – 185: Public Key Infrastructure (PKI):

- It is a mechanism of issuing, supporting and managing the asymmetric encryption keys systematically.
- An encryption key is a string of bits which is paired with the original data to transform it into *encrypted data* or *cyphertext*.
- It is also a system of protocols and data formats etc. to enable the large scale systems to use public key cryptography.
- PKI relies upon digital certificates. Each digital certificate binds the public key to a certificate owner.
- A digital certificate has a validity period and is signed by a certificate authority (CA).
- It is a dependable method to:
  - Implement asymmetric encryption.
  - Managing Cloud consumer and Cloud provider identity information.
  - Defending against malicious intermediary and insufficient authorization threats.

Module No – 186:Identity and Access Management (IAM):

- It is a mechanism comprising of policies and procedures to track and manage the user identities and access privileges for IT resources.
- Consist of four main components:
  - *Authentication*: Usernames+passwords, biometric, remote authentication through registered IP or MAC addresses.
  - *Authorizatio*n: Access control and IT resource availability
  - *User management*: Creating new user-identities, password updates and managing privileges.
  - *Credential management*: It establishes identities and access control rules for defined user accounts.
- As compared to PKI, the IAM uses access control policies and assigns user privileges.
- **Single Sign-On:** Saves the Cloud consumers from signing-in to subsequent services if the consumer is executing an activity which requires several Cloud services.
  - A security broker authorizes the consumer and creates a security context persistent across multiple services.

Module No – 187:

- **Cloud-based Security Groups:** Cloud IT resources are segmented for easy management and provisioning to separate users and groups.
    - The segmentation process creates Cloud-based security groups with separate security policies.
    - These are logical groups which act as network perimeters.
    - Each Cloud-based IT resource is assigned to atleast one logical cloud-based security group.
    - Multiple VMs hosted over same physical server can be allocated to different cloud-based security groups.
    - Safeguard against DoS attacks, insufficient authorization and overlapping trust boundaries threats.
    - Closely related to logical network perimeter mechanism.
- **Hardened Virtual Server Images:** It is a process of removing unnecessary software components from the VM templates.
    - It also includes closing unnecessary ports, removing root access and guest login and disabling unnecessary services.
    - Makes the template more secured than non-hardened server image templates.

Lesson No. 37

PRIVACY ISSUES OF CLOUD COMPUTING

Module No – 188: Lack of user control:

- Data privacy issues such as unauthorized access, secondary usage of data without permission, retention of data and data deletion assurance occur in Cloud Computing.
- With the data of a SaaS user placed in Cloud, there is a lack of user control over that data.
- A few reasons are as follows:
    - *Ownership and control of infrastructure*: The user has neither ownership nor the control of underlying infrastructure of the Cloud. There is a threat of theft, misuse and unauthorized sale of user's data.
    - *Access and transparency*: In many cases, it is not clear that a Cloud service provider can/will access the users' data. It is also not clear that an unauthorized access can be detected by the Cloud user/provider.
    - *Control over data lifecycle*: The Cloud user can not confirm that the data deleted by the user is actually been deleted. There is no assurance for the data deletion of terminated accounts as well. There is no regulation to implement a must-erase liability on Cloud provider.
    - *Changing provider*: It is not clear how to completely retrieve the data from previous provider and how to make sure that the data is completely deleted by the previous provider.
    - *Notification and redress*: It is not clear how to determine the responsibility of (user or provider for) an unauthorized access.

Module No – 189: Lack of Training and Expertise:

- The deployment and running of Cloud service may require the recruitment of highly skilled personals.
- For example the STEM skills (Science, Technology, Engineering and Mathematics) should be present in the recruited people.
- The lack of STEM skilled and/or trained persons can be a Cloud security issue.
- Such people may also lack the understanding of the privacy impact of their decisions.
- Due to the rapid speed and spread of computing devices among the employees, now more employees may introduce a privacy threat on average.
- For example multiple employees may leave their laptops unattended with a further possibility of unencrypted sensitive data.
- The employees can access different public Cloud services through self service portals.
- Care and control must be observed regarding public Cloud access to overcome the privacy issues.

Module No – 190: Unauthorized Secondary Usage:

- There is a high tendency that the data stored or processed over Cloud may be put to unauthorized usage.
- A legal secondary-usage of Cloud consumers' data is to sell the statistics for targeting the advertisements.
- However an illegal secondary-usage example is the selling of sales data to competitors of the consumer.
- Therefore it may be necessary to legally address the usage of consumer's data by the Cloud provider.
- So far there are no measures and means to verify the illegal secondary-usage of consumers' data by the Cloud provider/s.
- In future, a technological solution may be implemented for checking and preventing the unauthorized secondary usage of consumers' data.

Module No – 191: Complexity of Regulatory Compliance:

- The global nature of Cloud computing makes it complex to abide by all the rules and regulations in different regions of the world.
- The legal bindings regarding data location is complex to implement because the data may be replicated on multiple locations at the same time.
- It is also possible that the each replicated copy of the data is managed by different entities for example backup services obtained from two different providers.
- The backup provided by a single provider may be spread across different data centers which may or may not be within the legal location-boundary.

- The rapid provisioning architecture of the Cloud makes it impossible to predict the location of to-be-provisioned Cloud resource such as storage and VMs.
- The cross border movement of data while in transit is very difficult to control. Specially when the data processing is outsourced to another Cloud provider. Then the location assurance of such Cloud provider is a complex task at runtime.

Module No – 192: Addressing Trans-border Data Flow Restrictions:

- The privacy and data protection regulations in many countries restrict the trans-border flow of personal information of the citizens.
- These countries include EU and European Economic Area (EEU) countries, Australia, Canada etc.
- From EU/EEU countries, the personal information can flow to countries which have adequate protection. These include the EU/EEU countries and Canada etc.
- The flow of personal information to other countries is restricted, unless some rules/agreements are followed by those countries.
- For example the information can be transferred from EU to USA if the receiving entity has joined the US Safe Harbor agreement.
- If the receiving country has signed a model contract with the EU country/ies then the personal information can flow towards the receiving country.
- So far the trans-border regulations are not complied with Cloud computing and there is more to be done to implement these data flow restrictions.

Module No – 193: Litigation:

- A Cloud Service Provider (CSP) may be forced to hand over the consumers' data due to a court writ.
- For example, in a case handled by the US court of law, with state vs. the defendant, the US govt. was allowed the access to Hotmail service (of Microsoft) through the court orders.
- . The govt. always wants to check the relevance of evidence with the case. For that, the court can allow access to consumers' data.
- But for private entities, this situation can be avoided through the clauses of legal agreement to bind the CSP for disallowing any access(by a non govt. entity) to the data. OR to govern the response of CSP to any writ from such entities.

Module No – 194: Legal Uncertainty:

- Since the Cloud computing moves ahead of the law, there are legal uncertainties about the privacy rights in the Cloud.
- Also, it is hard to predict the outcome of applying the current legal rules regarding trans-border flow of data to Cloud computing.

- One of the areas of uncertainty is about the procedure of anonymizing or encrypting of personal data requires a legal consent from the owner and the processing related to enhancement of data privacy is exempt from privacy protection requirements?
- Also, it is not clear that the anonymized data (which may or may not contain personal data) is also governed by the trans-border data flow legislations or not.
- In short, the legal uncertainty exists regarding the application of legal frameworks for privacy protection upon Cloud computing.

Module No – 195: Conclusions:

- There is uncertainty in privacy protection globally.
- Cloud globalization has invoked new demands of privacy protection from the existing security-frameworks.
- Policymakers are pushing towards the change in security frameworks and placing more emphasis upon accountability for data privacy related violations.
- USA and EU are currently considering the Privacy Bill of Rights ad the data protection framework respectively for privacy protection.
- Cloud computing offers complex challenges for entities that need to meet global privacy regulations.
- Complying with trans-border data flow restrictions, the difficulty in knowing the geographic location of data processing and storage location are the major challenges for privacy assurance in Cloud computing.
- Data deletion and discarding of virtual storage device/s must be carefully assured.

Lesson No. 38

SECURITY ISSUES OF CLOUD COMPUTING

Module No – 196: Gap in Security:

- Although the security controls for the Cloud are same as of other IT environments, but the lack of user control in Cloud computing introduces security risks.
- These security risks are due to a possible lack of effort for addressing the security issues by the Cloud service provider.
- SLAs do not include any provision of the security procedures made necessary by the consumer or through any standard.
- The gap in security also depends upon the type of service (IaaS, PaaS & SaaS).
- The more privileges given to the consumer (for example in IaaS), the more responsibility of security procedures lies with the consumer.
- The consumer may need to gain the knowledge of the security procedures of provider.
- The provider gives some security recommendations to IaaS and PaaS consumers.

- For SaaS, the consumer needs to implement its own identity management system for access security.
- Generally, it is very difficult to implement protection throughout the Cloud. In few cases the Cloud providers are bound by law for the protection of personal data of the citizens.
- It is difficult to ensure the standardized security when a Cloud provider is outsourcing resources from other providers.
- Currently the providers take no responsibility/liability for deletion, loss or alteration of data.
- The terms of service are usually in favor of the provider.

Module No – 197: Unwanted Access:

- Cloud consumers may experience unwanted access to their data from the governments. There are many laws in the world (for example the US Patriot Act) which allow the government a privileged access to the Cloud consumers' data.
- The other type of unwanted access is from the lack of adequate security when the Cloud provider is in a supply chain link with other providers.
- A malicious employee may have a privileged access to data (because of being the employee).
- Data thieves and even the other consumers of the same service may break into the consumers' data if the data of each consumer is not adequately separated.
- The damage can be far greater than non Cloud environments due to the presence of various roles in Cloud architectures with administrative level access.
- In general the Cloud storage is more prone to risks from malicious behavior than the Cloud processing
- This is because the data may remain in Cloud storage for longer period of time and hence exposed to more risks.

Module No – 198: Vendor Lock-in:

- Cloud computing in today's time lacks interoperability standards.
- There are certain limitations such as
    - Difference between common hypervisors.
    - Gap in standard APIs for management functions.
    - Lack of commonly agreed data formats.
    - Issues with machine-to-machine interoperability of web services.
- The lack of standards makes it difficult to establish security frameworks for heterogeneous environments.
- People mostly depend upon common security best practices.
- Since there is no standardized communication between Cloud providers and no standardized data export format, it is difficult to migrate from one Cloud provider to another or to bring back the data and process it in-house.

Module No – 199: Inadequate Data Deletion:

- So far there is no surety or confirmation functionality for the deleted data being really deleted and non-recoverable by the service provider.
- This is due to lack of consumer control over life cycle of the data (as discussed before).
- This problem is increased with the presence of duplicate copies of the data.
- It might not be possible to delete a virtual disk completely because several consumers might be sharing it or the data of multiple consumers resides over same disk.
- For IaaS and PaaS, the reallocation of VMs to subsequent consumers may introduce the problem of data persistency across multiple reallocations.
- This problem exists until the VM is completely deleted.
- For SaaS, each consumer is one of the users of a multitenant application. The customer's data is available each time the customer logs-in.
- The data is deleted when the SaaS consumer's subscription ends.
- There is correspondingly higher risk to customers' data when the Cloud IT-resources (such as VM and storage) are reused or reallocated to a subsequent consumer.

Module No – 200: Compromise of the Management Interface:

- As discussed previously, the management interfaces are available through remote access via Internet.
- This poses an increased risk compared to traditional hosting providers.
- There can be vulnerabilities associated with browsers and remote access.
- These vulnerabilities can result in the grant of malicious access to a large set of resources.
- This increased risk is persistent even if the access is controlled by a password.

Module No – 201: Backup Vulnerabilities:

- In order to provide high level of reliability and performance, a Cloud provider makes multiple copies of the data and store them at different locations.
- This introduces many vulnerabilities.
- There is a possibility of data loss from Storage as a Service.
- A simple solution is to place data at consumer's premises and use the Cloud to store (possibly encrypted) backup of data.
- A loss of data may occur before taking backup.
- A subset of the data may get separated and unlinked form the rest and thus becomes unrecoverable.
- The failure/loss of data-keys may significantly destroy the data context.
- Sometimes the consumers of traditional (non-Cloud) backup service suffer a complete loss of their data on non-payment of periodic fee.
- In general, the Cloud service show more resiliency than these traditional (non-Cloud) services.

Module No – 202: Isolation Failure:

- The multi-tenant SaaS applications developed by Cloud providers use logical/virtual partitioning of the data of each consumer.
- It is possible that such applications be storing the personal and financial data of the consumers on Cloud.
- This responsibility of securing this data is of the Cloud provider.
- Due to the possibility of the failure of data separation mechanisms, the other tenants can access the sensitive information.
- Virtualization is widely used in Cloud computing. The VMs although are isolated from each other, yet the virtualization based attacks may compromise the hosting server and hence expose all the hosted VMs to the attacker.

Module No – 203: Missing Assurance and Transparency:

- As discussed before, the Cloud provider take lesser liabilities in case of data loss.
- Therefore, the consumers should obtain some assurance from the Cloud provider regarding the safety of their data.
- Consumers may also demand for getting the warning/s regarding any attack/unauthorized access/loss of data.
- A few frameworks exist for security assurance in Cloud. The Cloud providers offer the assurance on the basis of these frameworks.
- However these assurances may not be applied in case of frequent data accesses and/or in case of some instances such as isolation failure (discussed previously).
- Still, there is no compensation offered by the Cloud providers for the incidents of data loss.
- The best assurance for data security in Cloud computing is achievable through keeping the data in private Cloud.
- Although automated data security assurance evaluation frameworks exist but they still need to evolve in order to comply with all the security issues discussed in this course.

Module No – 204: Inadequate Monitoring, Compliance and Audit:

- A Cloud consumer should be able to audit the data processing over Cloud to ensure that the Cloud procedures are in compliance with the security policy of the consumer.
- Similarly the Cloud consumers may want to monitor SLA compliance by the provider but the complexity of Cloud infrastructure makes it very difficult to extract the appropriate information or to perform a correct analysis.
- Cloud providers could implement the internal compliance monitoring controls in addition to external audit process.
- The consumers may even b allowed a 'right to audit' for those particular consumers who have regulatory compliance responsibilities.

- Although the existing procedures for audit can be applied to Cloud computing but the provision of a full audit trail with the public Cloud models is still an unsolved issue.

Module No – 205: Conclusion:

- There are a number of security issues in Cloud computing.
- These issues depend upon the service provision and deployment models.
- Cloud audit is one of the open issues of Cloud security.
- Overall, the adoption of Cloud computing does not necessarily affects the security.
- The security can be outsourced to the experts of security to achieve better security than before.
- The major issue is to find a Cloud provider with suitable controls to assure security, monitoring and audit.

Lesson No. 39

TRUST ISSUES OF CLOUD COMPUTING

Module No – 206: Trust in the Clouds:

- Cloud consumers have to trust the Cloud mechanisms for storing and processing the sensitive data.
- Traditionally, a security perimeter (such as a firewall) is instantiated to setup a trust boundary within which there is a self-control over computing resources and where the sensitive data/information is stored and processed.
- . The network provides trusted links to other trusted end hosts.
- This may work perfectly for the Internet but may not work for public and hybrid Clouds.
- This is because the data may be stored and/or processed beyond the security perimeter such as supply chain issues discussed before.
- The consumers have to extend the trust boundaries to the Cloud provider.
- Therefore, the consumers should only trust the Cloud provider if the information about the reliability of internal mechanisms is provided by trusted entities such as consumer groups, auditors, security experts, reputed companies and established Cloud providers etc.
- The trust relationships can be the decision affecting factors for adopting/accepting a particular security and privacy solution.
- Trust attains a higher level of importance if personal or business critical information is to be stored in Cloud.
- Therefore, the Cloud providers have to have high trust from the consumers.

Module No – 207: Lack of Consumer Trust:

- In the past various surveys in Europe have revealed the lack of consumer trust upon the protection of their data kept online.
- Up to 70% of Europeans were concerned about the non authorized secondary usage of their data.
- The survey about trust on Cloud provider showed the following statistics:
    - Reputation: 29%
    - Recommendation from trusted party: 27%
    - Trial experience: 20%
    - Contractual: 20%
    - Others: 4%
- The consumer trust depends upon the compatibility level of data protection provided by the Cloud provider vs. the consumer's expectations.
- A few such expectations include the regulatory compliance of data handling procedures and control over data lifecycle even in supply chain Cloud provisioning.
- 70% of the business users (in selected regions of the world) are already using private Clouds according to a study.
- However different surveys showed that the enterprises are concerned about:
    - Data security: 70%
    - SLA compliance : 75%
    - Vendor lock-in:79%
    - Interoperability: 63%

Module No – 208: Weak Trust Relationships:

- Although the Cloud provider/s may be using a supply chain mechanism through the IT resources of subcontractors.
- This may jeopardize the security and privacy of the consumers' data (as discussed before) and thus weakens the trust relationships.
- Even if the trust relationships are weak in service delivery chain, but at least some trust exists so that the rapid provisioning of the Cloud services can be performed.
- Significant business risks may arise when critical data is placed on cloud and the consumer has lack of control over the passing of this data to a subcontractor.
- So the trust along the service delivery chain from the consumer to Cloud provider is non-transitive.
- There is a lack of transparency for the consumer in the process of data flow. The consumer may even not know the identity of the subcontractor/s.
- In-fact, the 'On-demand' and 'pay-as-you-go' models may be based upon weak trust relationships.
- This is because new providers have to be added on the go to provide the extra capacity on short notice.

Module No – 209: Lack of Consensus About Trust Management Approaches to Be Used:

- The consensus about the use of trust management approaches for Cloud computing is missing.
- Trust measurement is a major challenge due to the difficulty of contextual representation of trust.
- Some standardized trust models are required to be created for evaluating and assurance of accountability.
- Almost all of the existing models for trust evaluation are not adequate for Cloud computing.
- The existing models of trust evaluation in Cloud computing partially cover the trust categories.
- Trust models are lacking a suitable metrics for accountability.
- There is no consensus on type of evidence required for the verification of the effectiveness of trust mechanisms.

Module No – 210: Conclusions:

- Trust is widely considered as a key concern for consumers, enterprises and regulators.
- Lack of trust is the key factor which inhibits the wide adoption of Cloud services by the end-users.
- People are worried about what will happen to their data when it is placed on Cloud.
- There is a fear of unwanted access, unauthorized access and unauthorized secondary usage of the data.
- There is a lack of user control.
- Enterprises shifting to public Cloud are concerned about the confidentiality and security of their data.
- The regulators are worried about illegal trans-border transfer of data.
- Thus the usage of Cloud is a question of trade-offs between privacy, security, compliance, costs and benefits.
- Trust mechanisms have to be propagated along the chain of service provision.
- Trust measurement models are to be developed to cover all aspects of trust in Cloud computing.

Module No – 211: Trust Management in Cloud Computing:

- In order to monitor and evaluate the trust, the systematic trust-management is required.
- There should be a system to manage the trust.
- The trust management system should be able to measure the "trustfulness" of the Cloud services.
- The following attributes can be considered:
    - Data integrity: Consisting of security, privacy and accuracy.
    - Security of consumers' personal data.
    - Credibility: Measured through QoS.

- Turnaround efficiency: The actual vs. promised turnaround-time. It is the time from placement of consumer's task to the finishing of that task.
- Availability of Cloud service provider's resources and services.
- Reliability or success rate of performing of agreed upon functions within the agreed upon time deadline.
- Adaptability with reference to avoidance of single point of failures through redundant processing and data storage.
- Customer support provided by the Cloud provider.
- The consumer feedback on the service being offered.

- These attributes can be graded and trust computation can be performed. The computed value can be saved for future comparison.

Module No – 218: Approaches to Addressing Privacy, Security and Trust Issues:

- In this module we shall briefly discuss the possible approaches to solve the privacy, security and trust issues in Cloud.
- There are three main dimensions in this regard:
- Innovative regulatory frameworks to facilitate the Cloud operations as well as to solve the possible issues regarding privacy, security and trust.
- Responsible company governance should be exhibited by the provider to show the intension of safeguarding the consumer's data and intension to prove this intension through audit.
- Use of various supporting technologies for privacy enhancement, security mechanisms, encryption, anonymization etc.
- By using a combination of these dimensions, the consumers can be reassured of the security and privacy of their data and the Cloud provider can earn the trust.

Lesson No. 40

OPEN ISSUES IN CLOUD

Module No – 212: Overview:

- Cloud is not a solution for all consumers of IT services.
- Cloud is also not suitable for all applications.
- Cloud computing contains a number of issues which are not necessary unique to Cloud.
- Hardware failures and security compromises are possible in complex computing systems.
- Similarly the software built to fulfill complex requirements of concurrency, dynamic configuration and large scale computations are more prone to bugs and crashes than the commercial scale typical software.
- This should be kept in mind that the Cloud computing(which is based upon complex computing hardware and software) will also exhibit some failures and security compromises.
- But this does not disqualify the Cloud computing from being adopted by the consumers.

- Instead, it means that there are techniques to address, reduce the effects and isolate these failures and compromises.
- We shall discuss the issues related to Cloud computing in the coming modules as highlighted by NIST USA.

Module No – 213: Computing Performance:

- The real time applications require high performance and high degree of predictability.
- Cloud computing shows some performance issues which are similar to those of other forms of distributed computing.
    - **Latency:** As measured through round-trip-time (the time from sending a message to receiving a response) is not predictable for Internet based communications.
    - **Offline Data Synchronization:** For the offline updates in data, the synchronization with all the copies of data on Cloud is a problem. The solution to this problem requires the mechanisms of version control, group collaboration and other synchronization capabilities.
    - **Scalable Programming:** The legacy applications have to be updated to fully benefit from scalable computing capacity feature of Cloud computing.
    - **Data Storage Management:** The consumers require the control over data life cycle and the information regarding any intrusion or unauthorized access to the data.

Module No – 214: Cloud Reliability:

- It is a probability that a system will offer failure-free service for a specified period of time for a specified environment.
- It depends upon the Cloud infrastructure of the provider and the connectivity to the subscribed services.
- Measuring the reliability of a specific Cloud will be difficult due to the complexity of Cloud procedures.
- Several factors affect the Cloud reliability:
    - **Network Dependence:** The unreliability of Internet and the associated attacks affect the Cloud reliability.
    - **Safety-Critical Processing:** The critical applications and hardware such as controls of avionics, nuclear material and medical devices may harm the human life and/or cause the loss of property.
        - These are not suitable to be hosted over Cloud

Module No – 215:

- **Economic Goals:** Although the Cloud provides economic benefits such as saving upfront costs and elimination of maintenance costs and provides consumers with economies of scale.

- However there are a number of economic risks associated with Cloud computing.
- **SLA Evaluation:** The lack of automated mechanisms for SLA compliance by the provider requires the development of a common template that could cover the majority of SLA clauses and could give an overview of SLA complaisance.
- This would be useful in decision making for investing the time and money in manual audit.
- **Portability of Workloads:** The initial barriers to Cloud adoption are the needs of a reliable and secure mechanism for data transfer to Cloud as well as to port the workload to other providers are open issues.
- **Interoperability between Cloud Providers:** The consumers face or are in fear of vendor lock-in due to lack of interoperability among different providers.
- **Disaster Recovery:** The physical and/or electronic disaster recovery requires the implementation of recovery plans for hardware as well as software based disasters so that the provider and consumers can be saved from economic and performance losses.

Module No – 216: Compliance:

- It is with respect to any law and preferred security procedures.
- NIST and other US govt agencies are evolving methods to solve the compliance issues between consumers and providers.
- The consumer is although responsible for compliance but the implementation is actually performed by the provider.
- The consumer has a lack of visibility regarding the actual security procedures being adopted and/or applied by the provider. However the consumer may request for the deployment of monitoring procedures.
- The consumers (having their data processed on provider's premises) need to acquire assurance from the provider regarding the compliance with various laws. For example in US: the health information protection act, payment security standard, information protection accountability act etc.
- The forensics support regarding any incidence should be provided. This will evaluate the type of attack, the extent and damage associated and collection of information for possible legal actions in future.
- The forensic analysis for SaaS is the responsibility of the provider while the forensic analysis of IaaS is the responsibility of the consumer.

Module No – 217: Information Security:

- Related to confidentiality and integrity of data.
- It is also linked with the assurance of availability.
- The following measures can be used by an organization for data security:
    - Application of administrative controls for specifying the authority of specific users to create, update, delete, disclose and transport of the data.

  - Physical controls for the security of data storage devices.
- Technical controls for Identity and Access Management (IAM), data encryption and data audit-handling requirements according to any regulatory need.
- Public and private Clouds have their own typical security exposures.
- The provider however may also provide physical separation of consumers' data in addition to logical separation.
- According to NIST, the provider should provide the monitoring mechanism to satisfy the consumer regarding the security compliance.

Lesson No. 41

DISASTER RECOVERY IN CLOUD COMPUTING

Module No – 219: Understanding the threats:

- **Disk Failure:** disk drives are electro-mechanical devices which wear out and eventually fail.
- Failure can be due to disaster such as fire and floods. Can also be due to theft.
- All mechanical devices have mean time between failure (MTBF).
- The MTBF values given by the manufacturers are usually generic values calculated for a set of devices.
- Therefore, instead of relying upon the MTBF, there must be a disaster recovery plan for the disk failure.
- The following strategies can be utilized:
  - Traditional approach: It is to have backup on separate storage. If the disk fails due to any disaster, the data can be recovered on a new disk from the backup. But if the backup is also destroyed or stolen, then there is a complete loss of data. Also, the recovery process is time consuming.
  - Redundant Array of Independent Disks (RAID): It is a system consisting of multiple disk drives. Multiple copies of data are maintained and stored in a distributed way over the disks. If one disk fails, simply the disk is replaced and the RAID system copies the data over the new disk. But still backup is required because if the entire RAID is destroyed or stolen, there is a complete data loss.
  - Cloud based data storage ad backup: Cloud not only provides the facility of data access over the Internet, but it also provides enhanced data replication. The replication is sometimes performed by default without any extra charges. The Cloud based backup is stored at a remote site so it is an extra advantage as compared to onsite backup placement.
  - Further, the Cloud based backup is readily available and thus reduces the downtime as compared to recovery using traditional tape-based backup.

Module No – 220: Understanding the threats:

- **Power Failure or Disruption:** The Computers can be damaged due to a power surge caused by a storm or some fault in power supply system.
- Power surge may permanently damage the disk storage.
- The user looses all the unsaved data when a power-blackout happens.
- A few disaster recovery plans are as follows:
    - Traditionally, the surge protector devices are used. But these devices are not helpful in saving the (unsaved) data in case of a blackout.
    - The in-house data centers can use huge and expensive uninterruptable power supply (UPS) devices and/or generators.
    - Another solution is to shift the data to another site. But this is expensive and time consuming.
    - The best option is to move the data center to Cloud. The Cloud providers have better(and expensive) power backups and their cost is divided among the consumers. Also, the Cloud mechanism may automatically shift the data to a remote site on another power grid (in case of power failures of longer duration).

Module No – 221: Understanding the threats:

- **Computer Viruses:** While surfing the web, the users may potentially be downloading and installing software and/or share the drive such as junk drives over their computing devices.
- These devices are at the risk of attacks through computer virus and spyware.
- Traditionally, the following techniques have been used for safeguarding against the virus attacks:
    - Making sure each computer has anti-virus installed and set to auto-update to get the most recent virus and spyware signatures.
    - Restrict the user privilege to install software.
    - Using a firewall over router or on the computer or around the LAN.
- Cloud computing presents difficulties for non-Cloud based viruses to penetrate. This is because of the complexities of virtualization technologies. Also the Cloud providers ensure reasonable security measures for the consumer's data and software.

Module No – 222: Understanding the threats:

- **Fire, Flood & Disgruntled Employees:** The fire as well as the fire extinguishing practices can destroy the computing resources, data and backup.
- Similarly the heavy and/or unexpected rainfall may cause an entire block or whole city including the computing equipment to be affected by a flood.
- . Similarly an angry employee can cause harm by launching a computer virus, deleting files and leaking the passwords.

- Traditionally the office equipment is ensured to lower the monitory damage. Backup is used for data protection. Data centers use special mechanisms for fire-extinguishing without water sprinkles.
- By residing the data center over Cloud, the consumer is freed from making efforst and expenditures for fire prevention systems as well as for data recovery. The cloud provider manages all these procedures and includes the cost as minimal part of the rental.
- Unlike fire, the floods can not be avoided or put-off.
- The only possibility to avoid the damage due to floods is to avoid setting up the data center in a flood zone.
- Similarly, choose a Cloud provider which is outside any flood zone.
- Companies apply access control and backup to limit the access to data as well as the damage to data due to unsatisfied employees.
- In Cloud, the Identity as a Service (IDaaS) based single sign-on excludes the access privileges of terminated employees as quickly as possible to prevent any damages.

Module No – 223: Understanding the threats:

- **Lost Equipment & Desktop Failure:** The loss of equipment such as a laptop may immediately leads to the loss of data and a possible loss of identity.
- If the data stored on the lost device is confidential then this may lead to even more damage.
- Traditionally the risk of damage due to lost or stolen devices is reduced by keeping backup and to safeguard the sensitive data, login and strong password for the devices are used.
- But even the strong passwords are not difficult to break for the experienced hackers. Yet most of the criminals are still prevented to access the data.
- For the Cloud computing, the data can be synchronized over multiple devices using the Cloud service. Therefore the user can get the data from online interface or from other synced devices.
- In case of desktop failure, the user (such as an employee of a company) becomes offline until the worn out desktop is replaced.
- If there was no backup, the data stored on the failed desktop may become unrecoverable.
- Traditionally, data backup is kept for the desktops in an enterprise. The backup is stored on a separate computer. In case of desktop failure, the maintenance staff tries to provide alternative desktop and restore the data s soon as possible.
- Whereas in Cloud, the employees work on the instances of IaaS or Desktop as a Service by using the local desktops.
- In case of desktop failure, the employee can just walk to another computer and log in to the Cloud service to resume the work.

Module No – 224: Understanding the threats:

- **Server failure & Network Failure:** Just like the desktops, the severs can also fail.

- The replacement of blade server is relatively simple process and mostly the blade servers are preferred by the users.
- Ofcourse there has to be a replacement server in stock to replace with the failed server.
- Traditionally the enterprises keep redundant servers to quickly replace a failed server.
- In case of Cloud computing, the providers of IaaS and PaaS manage to provide 99.9% up-time through server redundancy and failover systems. Therefore the Cloud consumers do not have to worry about server failure.
- The network failure can occur due to a faulty device and will cause downtime.
- Traditionally, the users keep 3G and 4G wireless hotspot devices as a backup. While the enterprises obtain redundant Internet connections from different providers.
- Since the Cloud consumers access the Cloud IT resources through the Internet, the consumers have to have redundant connections and/or backup devices for connectivity.
- Same is true for the Cloud service provider. The 99.9% up-time is assured due to backup/redundant Network connections.

Module No – 225: Understanding the threats:

- **Database System Failure & phone system failure:** Most of the companies rely upon database systems to store a wide range of data.
- There are many applications dependent upon database in corporate environment such as customers record keeping, sale-purchase and HR systems etc.
- The failure of data base will obviously makes the dependent application unavailable.
- . Traditionally, the companies either use a backup or replication of database instances. The former case results in downtime of database system while the latter results in minimum downtime or no downtime but is more complicate to implement.
- The Cloud based storage and database systems use replication to minimize the downtime with the help of failover systems.
- Many companies maintain phone systems for conference calling, voice mail and call forwarding.
- Although the employees can switch to using mobile phones in case the phone system fails. But the customers are left unaware of the phone number to connect to the company till the phone system recovers.
- Traditionally, the solutions are applied to reduce the impact of phone failure.
- Cloud based phone systems on the other hand provide reliable and failure safe telephone service. Internally, the redundancy is used in the implementation.

Module No – 226: Measuring Business impact, disaster recovery plan template:

- The process of reducing risks will often have some cost. For example the resource redundancy and backups etc.
- This indicates that investment on risk-reduction mechanisms will be limited.

- The IT staff should therefore evaluate and classify each risk according to its impact upon the routine operations of the company.
- A tabular representation of the risks, the probability of occurrence and the business continuity impact can be shown.
- The next step is to formally document the disaster recovery plan (DRP).
- A template of DRP can contain the plan overview, goals and objectives, types of events covered, risk analysis and the mitigation techniques for each type of risk identified in earlier step.

Risk	Occurrence Probability	Business Continuity Impact
User disk failure	Medium	Low
Server disk failure	Low	High
Network failure	Low	High
Database failure	Medium	High
Server power failure	High	High
Fire	Low	High
Flood	Low	High

Follow the video lecture to understand fully

  - **Standard Programming Languages:** Whenever possible, the consumers should prefer those Clouds which work in standardized programming languages and tools.
  - Right from the start, there should be a clearly documented plan of returning of data/resources to the consumer at the time of termination of service usage.
- **Continuity of Operations:** Consumer should assure that the Cloud provider act upon the requested parameters of disaster recovery plan of the consumer for the business-critical software and data hosted on Cloud.
  - In case of service interruption, the consumer should demand compensation in addition to reversal of service charges from provider.
  - Otherwise the consumer should host such critical applications/data/processing locally.
- **Compliance:** The consumer should determine that the provider has implemented necessary procedures and controls to comply with various legal, ISO and audit standards required as a provider and/or for the consumer to fulfill.
- **Administrator Staff:** The consumer should make sure that the internal procedures and policies of the provider are sufficient to protect against malicious insiders.
- **Licensing:** The consumer should make sure that proper licensing is obtained and provided by the provider for the proprietary software.

Module No – 228: Data Governance:

- **Data Access Standards:** Before developing the Cloud based applications, the consumers should make sure that the application interfaces provided in Cloud are generic and/or data adaptors could be developed for portability and interoperability of the Cloud applications can happen when required.
- **Data Separation:** The consumer should make sure that proactive measures are implemented at the provider's end for separation of sensitive and non-sensitive data.
- **Data Integrity:** Consumers should use checksum and replication technique to ensure the integrity of the data to detect any violations of data integrity.
- **Data Regulations:** The consumer is responsible to ensure that the provider is complying with all the regulations regarding data which are applicable to consumer regarding data storage and processing.
- **Data Disposition:** The consumer should make sure that the provider offer such mechanisms which delete the data of consumer whenever the consumer requests for it. Also make sure that the evidence or proof of data deletion is generated.
- **Data Recovery:** The consumer should examine the data backup, archiving and recovery procedures of the provider and make sure they are satisfactory.

Module No – 229: Security & Reliability:

- **Consumer-side Vulnerabilities:** Consumers should ensure the implementation of proper security and hardening of consumer platforms to avoid browser or other client devices based attacks.
- **Encryption:** Consumer should require that a strong encryption is applied for web sessions, data transfer and data storage.
- **Consumer-side Vulnerabilities:** Consumers should ensure the implementation of proper security and hardening of consumer platforms to avoid browser or other client devices based attacks.
- **Encryption:** Consumer should require that a strong encryption is applied for web sessions, data transfer and data storage.
- **Physical:** Consumers should consider the appropriateness of the physical security implementations and procedures of the provider. There should be a recovery plan for physical attack on provider's site. The providers having multiple installations in different geographical regions should be preferred.
- **Authentication:** Consumers should consider the use of advanced procedures for authentications provided by some providers to avoid account hijacking or identity thefts.
- **Identity & Access Management:** Consumers should have the visibility in the capabilities of provider for:
    - Authentication and access management procedures supported by the provider.
    - The tools available for consumers to extract the authentication information.
    - The tools available to consumer for granting authorization to consumer users and other applications without the involvement of provider.

- **Performance Requirements:** Consumers should benchmark the performance scores of an application before deploying it to Cloud in terms of responsiveness and performance regarding the input/output of bulk data.
- These scores should be used to establish key performance requirements after deploying the application over the Cloud.

Module No – 230: VMs, Software & Applications:

- **VM vulnerabilities:** When the provider is offering Cloud IT resources in the form of VMs, the consumer should make sure that the provider has implemented sufficient mechanisms to avoid attacks from other VMs, physical host and network.
- Also make sure the existence of IDS/IPS systems and network segmentation techniques such as VLANs.
- **VM Migration:** The consumers should plan for VM migration across different providers just in case.
- **Time-critical Software:** Since the public Clouds have unreliable response time therefore the consumers should avoid using the Cloud for the deployment of time-critical software.
- **Safety-critical Software:** Due to the unconfirmed reliability of Cloud subsystems, the use of Cloud for deployment of safety-critical software is discouraged.
- **Application development Tools:** When using the application development tools provided by the service provider, preference should be given to the tools which support the application development lifecycle with security features integrated.
- **Application Runtime Support:** Before deploying an application over Clouds, the consumer should make sure that the libraries calls used in application work correctly and all those libraries are dependable in terms of performance and functionality.
- **Application Configuration:** The consumer should make sure that the applications being deployed over the Cloud can be configured to run in a secured environment such as in a VLAN segment.
- Also make sure that various security frameworks can be integrated with the applications according to requirements of security policies of the consumer.
- **Standard Programming Languages:** Whenever possible, the consumers should prefer those Clouds which work in standardized programming languages and tools.

Lesson No. 42

MIGRATING TO THE CLOUD

Module No – 231: Define System Goals and Requirements:

- The migration to Cloud should be well planned. The first step should be to define the system goals and requirements. The following considerations are important:
    - Data security and privacy requirements

  - Site capacity plan: The Cloud IT resources needed initially for application to operate.
  - Scalability requirements at runtime
  - System uptime requirements
  - Business continuity and disaster requirements
  - Budget requirements
  - Operating system and programming language requirements
  - Type of Cloud: public, private or hybrid
  - Single tenant or multitenant solution requirements
  - Data backup requirements
  - Client device support requirements such as for desktop, tab or smartphone
  - Training requirements
  - Programming API requirements
  - Data export requirements
  - Reporting requirements
- [Jamsa, K. (2012). Cloud computing. Jones & Bartlett Publishers]

Module No – 232: Protect existing data and know your application characteristics:

- It is highly recommended that before migrating to Cloud, the consumer should backup the data. This will help in restoring the data to a certain time.
- The consumer should discuss with provider and agree upon a periodic backup plan.
- The data life cycle and disposal terms and conditions should be finalized at the start.
- If the consumer is required to fulfill any regulatory requirements regarding data privacy, storage and access then this should be discussed with the provider and be included in the legal document of the Cloud agreement.
- The consumer should know the IT resource requirements of the application being deployed over the Cloud.
- The following important features should be known:
  - High and low demand periods in terms of time
  - Average simultaneous users
  - Disk storage requirements
  - Database and replication requirements
  - RAM usage
  - Bandwidth consumption by the application
  - Any requirement related to data caching

Module No – 233: • Establish a realistic deployment schedule, Review budget and Identify IT governance issues:

- Many companies use a planned schedule for Cloud migration to provide enough time for training and testing the application after deployment.
- Some companies use a beta-release to allow employees to interact with the Cloud based version to provide feedback and to perform testing.

- Many companies use key budget factors such as running cost of in-house datacenter, payrolls of the IT staff, software licensing costs and hardware maintenance costs.
- This helps in calculation of total cost of ownership (TCO) of Cloud based solution in comparison.
- Many Cloud providers offer solutions at lower price than in-house deployments.
- Regarding the IT governance requirements, the following are important point:
    - Identify how to align the Cloud solution with company's business strategy.
    - Identify the controls needed within and outside the Cloud based solution so that the application can work correctly.
    - Describe the access control policies for various users
    - Describe how the Cloud provider logs the errors and system events and how to access the log and performance monitoring tools made available to the consumer.

Module No – 234: Designing Cloud based Solution:

- **Identify functional and non-functional requirements:** Before beginning the design phase of a Cloud application, the system requirements must be obtained and finalized.
- Personal meeting may be very helpful in this regard.
- Identification of errors and omission at early stage will save considerable cost and time later.
- The system requirements are of two types:
    - Functional
    - Non-functional
- **Functional requirements:** Define the specific tasks the system will perform. These are provided by the system analyst to the designer.
- **Non-functional requirements**: These are usually related to quality metrics such as performance, reliability and maintainability.

Module No – 235: Designing Cloud based Solution:

- **Identify Cloud Solution Design Metrics:**
    - *Accessibility*: The Cloud solution should consider either to maximize the user access or to restrict the access to authorized users only.
    - *Audit*: The Cloud solution should contain the logging and exception handling at critical processing points so that the log data can be used for audit later.
    - *Availability*: Identify the uptime requirements of the Cloud application being designed. Use the redundant deployment accordingly.
    - *Backup*: The cost of backup and the time to restore (from backup) should be considered while designing the Cloud based solution. If the data is handled by the provider, the backup policies of the provider should be reviewed and renegotiated if found not appropriate.

Module No – 236: Cloud Solution Design Metrics:

- **Existing & Future capacity:** If the application is being migrated to Cloud, then the current requirement of IT resources should be evaluated and used for initial deployment as well as for horizontal or vertical scaling configuration.
- **Configuration management:** Since the Cloud based solutions are accessed through any OS, browser and device, therefore the interfaces of the application should be able to render the contents with respect to OS, browser and user device.
- **Deployment:** The deployment issues such as related to OS, browser and devices should be addressed for the initial deployment as well as for future updates.
- **Environment (Green computing):** Design consideration for the Cloud based solution should contain considerations for power efficient design in order to reduce the environmental effect of carbon footprint of the Cloud based solution.
- **Disaster recover:** The Cloud solution design should have consideration of disaster recovery mechanisms. The potential risks for business continuity should be identified and cost effective mitigation techniques should be configured for these risks.
- **Interoperability:** The design consideration should contain possibility of interoperability between different Cloud solutions in terms of exchange of data.
- **Maintainability:** The Cloud solution should be designed to increase the reusability of code through loose coupling of the modules. This will lower down the maintenance cost.

Module No – 237: Cloud Solution Design Metrics:

- **Performance:** In order to enhance the performance, various considerations should be used such as reduce graphics on key pages, reduce network operations and use of data and application caching. Potential bottlenecks for performance should be identified and addressed.
- **Price:** The design of Cloud solution should be considered with respect to price and budget in short and long term. An inexpensive solution at the time of deployment may prove to be expensive in long run. The scalability, disaster recovery and performance etc. features should be implemented according to budget.
- **Privacy:** Privacy assurance should be implemented. Specifically for healthcare, educational and monitory data related solutions should have privacy assurance for internal as well as external unauthorize access. Same care is to be done for replicated databases and backups.
- **Portability:** The Cloud solution should be designed to be portable to other platforms when required. For that, instead of using the provider specific APIs, the generic or Opensource APIs and programming environments should be proffered for developing the Cloud solution.
- **Recovery:** In addition to disaster recovery, the solution should contain features to recover from other events not covered in disaster recovery such as user error, programing bugs, power outage etc.

Module No – 238: Cloud Solution Design Metrics:

- **Reliability:** Design should include the consideration for hardware failure events. The redundant configuration might be applied according to mean time between failure (MTBF) for each hardware device or establish a reasonable downtime.
- **Response time:** The response time should be as less as possible. Specifically for the online form submissions and reports.
- **Robustness:** It refers to the continuous working capability of the solution despite the errors or system failure. This can be complemented with Cloud resource usage monitoring for timely alarm for critical events.
- **Security:** The developer should consider the Cloud based security and privacy issues while designing.
- **Testability:** Test cases should be developed to test the fulfilment of functional and non-functional requirements of the solution.
- **Usability:** The design can be improved for usability by implementing a prototype and getting users' reviews to enhance the ease of usability of the Cloud solution.

Lesson No. 43

CLOUD APPLICATION SCALABILITY AND RESOURCE SCHEDULING

Module No – 239: Cloud Application Scalability:

- **Review Load Balancing Process:** Cloud based solutions should be able to scale up or down according to demand.
    - Remember, the scaling out and scaling up mean acquiring new resources and upgrading the resources respectively. Scaling in and scaling down are exactly the reverse of these.
    - There should be a load balancer module specially in case of horizontal scaling.
    - Load balancing or load allocating (in this regard) is performed by distribution of workload (which can be in the form of clients' requests) to Cloud IT resources acquired by the Cloud solution.
    - The allocation paten can be through round robin, random or a more complex algorithm containing multiple parameters. (More on this in later module).
- **Application Design:** Cloud based solutions should neither be having no-scaling nor the unlimited scaling.
    - There should be a balanced design of Cloud application regarding scaling with reasonable expectations.
    - Both horizontal and vertical scaling options should be explored either individually or in combination.

Module No – 240: Cloud Application Scalability:

- **Minimize objects on key pages:** Identify the key pages such as home page, forms and frequently visited pages of Cloud based solution.
    - Reduce the number f objects such as graphics, animation, audio etc. from these pages so that they can load quickly.
- **Selecting measurement points:** Remember a rule that a 20% of code usually performs the 80% of processing.
    - Identify such code and apply scaling to it.
    - Otherwise applying scaling may not have the desired performance improvements.
- **Analyze database operations:** The read/write operations should be analyzed for improving performance.
    - The read operations are non-conflicting and hence can be performed on replicated databases (horizontal scaling).
    - But write operations on one replica database requires the synchronization of all database instances and hence the horizontal scaling becomes time consuming.
    - The statistics of database operations should be used for decision about horizontal scaling.
- **Evaluate system's data logging requirements:** The monitoring system regarding the performance and event logging may be consuming disk space and CPU.
    - Evaluate the necessity of logging operations before applying or periodically afterwards and tune them to reduce disk storage and CPU wastage

Module No – 241: Cloud Application Scalability:

- **Capacity planning vs Scalability:** Capacity planning is planning for the resources needed at a specific time by the application.
    - Scalability means acquiring additional resources to process the increasing workload.
    - Both capacity planning and scalability should be performed in harmony.
- **Diminishing return:** The scaling should not be performed beyond a point where there is no corresponding improvement in performance.
- **Performance tuning:** In addition to scaling, the application performance should be tuned by reducing graphics, page load time and response time.
    - Additionally the use of caching should be applied. It is the use of faster hard disks, using RAM contents for content rendering and optimizing the code using 20/80 rule.

Module No – 242: Cloud Resource Scheduling Overview:

- Effective resource scheduling reduces:
    - Execution cost
    - Execution time
    - Energy consumption

- Fulfills QoS requirements:
    - Reliability
    - Security
    - Availability
    - Scalability
- Remember that provider wants to earn more profit and maximize the resource usage.
- Cloud consumer wants to minimize the cost and time of execution of workload.
- The Cloud resource scheduling can be performed on various grounds as discussed in coming modules.
    - [Singh, S., & Chana, I. (2016). A survey on resource scheduling in cloud computing: Issues and challenges. Journal of grid computing, 14(2), 217-264.]

Module No – 243: Cloud Resource Scheduling Overview:

- **Cost-Based Resource Scheduling:** This type of resource scheduling is performed on the basis of cost and budget constraints.
    - The users' requests are processed in first come first served basis along with QoS and time constraints considerations.
    - The cost constraint may take the priority over other constraints and thus may introduce starvation for some tasks.

Module No – 244: Cloud Resource Scheduling Overview:

- **Time-Based Resource Scheduling:** This type of resource scheduling prioritizes the processing deadline of the users' requests.
    - The resources are allocated to those jobs with deadline approaching faster than other requests.
    - Evaluated through the statistics such as number of deadlines missed and the overall cost etc.

Module No – 245: Cloud Resource Scheduling Overview:

- **Cost & Time-Based Resource Scheduling:** Time based scheduling may miss some tasks' deadlines or may prove to be expensive if over provisioning of IT resources is used to meet deadlines.
    - The cost based scheduling may miss some deadlines and/or cause starvation to some costs.
    - Better to use a hybrid approach for resource scheduling to gain cost as well as to minimize task deadline violations.

Module No – 246: Cloud Resource Scheduling Overview:

- **Bargain-Based Resource Scheduling:** This type of scheduling considers that a resource market exists with providers making offers to the users.
    - The users can negotiate for the processing cost.
    - The bargain-based scheduling can achieve low cost and meet deadline if negotiation is successful.
    - It is an evolving technique so far

Module No – 247: Cloud Resource Scheduling Overview:

- **Profit-Based Resource Scheduling:** This type of scheduling aims at increasing the profit of Cloud provider.
    - This can be done either by reducing the cost or increasing the number of simultaneous users.
    - The SLA violation is to be considered while making the profit based scheduling decisions.
    - The penalties of SLA violations may nullify the profit gained.

Module No – 248: Cloud Resource Scheduling Overview:

- **SLA & QoS Based Resource Scheduling:** In this scheduling, the SLA violations are avoided and QoS is maintained.
    - The more load put on IT resources, the more tasks may be completed in a unit time.
    - Yet it may cause SLA violation when IT resources are overloaded.
    - Hence the QoS consideration is applied to ensure SLA is not violated.
    - Suitable for homogeneous tasks for which the estimation can be performed for expected workload and expected time of completion.

Module No – 249: Cloud Resource Scheduling Overview:

- **Energy-Based Resource Scheduling:** The objective is to save energy at data center level to decrease the running cost and to contribute towards environment.
    - Energy consumption estimation is required for each scheduling decision. There can be a number of possible task distribution across servers and VMs.
    - Only that distribution is preferred which shows the least energy consumption for a batch of tasks at hand.

Module No – 250: Cloud Resource Scheduling Overview:

- **Optimization-Based Resource Scheduling:** Optimization is the process of perfecting an algorithm to make it more beneficial. Optimized solution to a problem is the best possible option (under some constraints) available for the problem at hand.

- Popular considerations used by researchers are revenue maximization, lowering communication overhead, output efficiency, energy efficiency, reducing completion time of tasks etc.
- Popular techniques used by researchers include Bayesian assumptions, Stochastic Integer Programming etc.
- The only drawback is the computational time of optimal solution. It may result in deadlines missing for some time critical jobs.

Module No – 251: Cloud Resource Scheduling Overview:

- **Priority-Based Resource Scheduling:** In this type of scheduling, the task starvation can be avoided. Specially for the situation of resource contention.
  - If there is a task classification e.g., on the basis of type, user, resource requirement etc., so that the priority of one task can cause the resource scheduler to preempt the other low priority tasks.
  - May cause the low priority tasks to suffer starvation.
  - *Aging* factor can be applied to increase the priority of low-priority tasks to avoid or lower the starvation.

Module No – 252: Cloud Resource Scheduling Overview:

- **VM-Based Resource Scheduling:** Since the VMs can host Cloud based applications and the VMs can be migrated, the resource scheduling can be performed on VM level.
  - The overall demand of all applications hosted on a VM is considered for scheduling. If a VM is facing resource starvation, it can be migrated to another server with available IT resources.
  - The disadvantage is, there is no guarantee that the destination host also runs out of IT resources due to already deployed VMs

Module No – 253: Cloud Resource Scheduling Overview:

- **VM-Based Resource Scheduling:** Resource scheduling may lead to formation of hybrid Cloud in-order to provision additional IT resources.
  - There is going to be a budget issue regarding outsourcing.
  - So in order to finish the task/s within deadline and budget, the selection of tasks to reschedule and the Cloud IT resources to be availed are very important.

Lesson No. 44

MOBILE CLOUD COMPUTING

Module No – 254:

- **Introduction:** Mobile devices are frequently being used throughout the world.
  - Over the time, the users have started to rely more and more upon mobile devices due to no constraints of time and location.
  - The applications installed over mobiles are of various types and of various computational requirements.
- **Overview:** The mobile devices are inherently constrained by resources shortage such as processing, memory, storage, bandwidth and battery etc.
  - There might be a number of situations when mobile devices become incapable of processing or running the applications due to resource shortage.
  - On the other hand, Cloud computing offers unlimited IT resources over Internet on-the-go.
- **Definition:** Mobile cloud computing at its simplest, refers to an infrastructure where both the data storage and data processing happen outside of the mobile device. Mobile cloud applications move the computing power and data storage away from mobile phones and into the cloud, bringing applications and Mobile Computing to not just smartphone users but a much broader range of mobile subscribers'.
  - [Dinh, H. T., Lee, C., Niyato, D., & Wang, P. (2013). A survey of mobile cloud computing: architecture, applications, and approaches. Wireless communications and mobile computing, 13(18), 1587-1611.]

Module No – 255: Need for Mobile Cloud Computing:

- There are various scenarios which indicate the need of a Mobile Cloud computing environment.
- This module presents a few examples in this regard.
- Optical character recognition (OCR) is used to identify and translate the text from one language to another. An OCR application could be installed over a mobile device for tourists.
- But due to resource shortage over the mobile devices, a better solution is to develop a Mobile Cloud application.
- Data sharing such as images form a site of disaster can be performed over Mobile Cloud application to help in developing an overall view of the site.
- The readings from sensors of multiple mobile devices spread across a vast region can not be otherwise collected and processed except through a Mobile Cloud application.

Module No – 256: Applications of Mobile Cloud:

- We shall discuss a few example applications of Mobile Cloud computing.

- **Mobile Commerce:** The applications of mobile commerce face the complexities such as bandwidth limitation, device configuration and security. In order to address these issues, the mobile commerce applications are integrated into Cloud computing.
- **Mobile Learning:** The mobile learning apps face the limitations in terms of high cost of devices & data plan and network bandwidth. Along with the limitation of storage space over mobile devices, these limitations can be overcome through shifting these applications over Cloud. This results in rendering of larger sized tutorials, faster processing and battery efficiency.
- **Mobile Healthcare:** The mobile healthcare applications based upon Mobile Cloud computing offer the following benefits in addition to assuring the security and privacy:
    - Remote monitoring of pulse rate, blood pressure etc. for patients over Internet.
    - Timely and effective cautioning and guidance to ambulances in case of medical emergencies.
- **Mobile Gaming:** Rendering of contents over mobile devices while executing the game engine over Cloud. Only the screens of the mobile devices are used, the rest is being done on Cloud.

Module No – 257: Mobile Cloud Computing Architecture:

- This module presents a generic architecture of Mobile Cloud computing.
- Mobile devices are connected to mobile networks through base stations such as access point, base transceiver stations or through satellite.
- The base stations establish the link between mobile devices and the mobile networks.
- The base stations convey the ID and location information along with the request to the central processors that are connected to the servers providing mobile network services.
- The mobile networks can provide the authentication, authorization and accounting on the basis of subscribers' data stored in databases.
- The subscribers' requests are then delivered to a Cloud through Internet.
- The Cloud controllers receive the subscribers' requests and provide services accordingly.
- The subscribers' requests are then delivered to a Cloud through Internet.
- The Cloud controllers receive the subscribers' requests and provide services accordingly.

Module No – 258: Mobile Cloud Models:

- There are various models of setting up of Mobile Clouds:
- A mobile device accessing an application/service hosted on Cloud servers such as email through 3G connection.
- Some mobile devices can provide resources to other mobile devices in a Cloud setup using mobile peer-to-peer network.
- The mobile devices can be connected to a *cloudlet* which is a set of multi-core computers connected to remotely placed Cloud servers. These cloudlets are usually in closed vicinity of the mobile devices save the network latency.

- [Fernando, N., Loke, S. W., & Rahayu, W. (2013). Mobile cloud computing: A survey. Future generation computer systems, 29(1), 84-106.]

Module No – 259: Advantages of Mobile Cloud Computing:

- **Extending battery lifetime:** The offloading of computational workload to Cloud saves battery life and reduces the response time.
- **Improving data storage capacity and processing power:** The storage limitation is overcome through Cloud storage. For example images uploaded to Cloud right after capturing. The running cost (in terms of energy and time) of compute-intensive application is also reduced when the mobile Cloud application is used.
- **Improving reliability:** The backup and disaster recovery features of Cloud computing provide reliability surety of data and applications for mobile Cloud computing as well.
- **Additionally:** The benefits of dynamic provisioning and scalability are also there.

Module No – 260: Cost Benefit Analysis of Mobile Cloud Computing:

- Cost benefit analysis proves to be useful for deciding about offloading the workload to Cloud.
- This analysis may consider the total investment (initial and running costs) and compare with the benefits of Mobile Cloud computing.
- Considering the goals of performance, energy conservation and quality to decide which server should receive the offload from mobile devices. Thus the cost benefit analysis in this case is from the point of view of Cloud infrastructure. Prediction can be used to estimate the performance, energy consumption and quality.
- The data related to devices' energy consumption, network throughput and application characteristics can be used to decide for offloading a task (of the profiled application) to Cloud or execute it locally in order to (for example) conserve battery.
- Security and privacy requirements may also be the base of task migration to Cloud.

Module No – 261: Mobile Cloud Computing : Security:

- There are a number of data security and privacy issues in Mobile Cloud computing. These are in addition to the security and privacy issues of Cloud computing.
- The following are the key areas for mobile Cloud security:
- Mobile devices themselves: Attacks, Virus and other malwares
- Mobile network: Related to wireless security
- Vulnerabilities in mobile Cloud applications: regarding the security and privacy bugs.

Module No – 262: Mobile Cloud Computing: Issues: Communications:

- There are some communication issues regarding Mobile Cloud computing. The researchers have also proposed different solutions in this regard.
- **Low Bandwidth:** It is one of the biggest issues for mobile Cloud computing because the radio resource for wireless networks is much scarce as compared with the traditional networks.
- **Availability:** The availability of the service becomes an important issue when the mobile deice has lost contact with mobile Cloud application due to network failure, congestion and loss of signals.
- **Heterogeneity:** The mobile devices accessing a Mobile Cloud Computing application are of numerous types and use various wireless technologies such as 2G, 3G etc and WLAN. An important issue is how to maintain the wireless connectivity along with satisfying the requirements of Mobile Cloud computing such as high-availability, scalability and energy efficiency etc.

Module No – 263: Mobile Cloud Computing: Issues: Computing:

- The mobile device has to make the decision for offloading the computational workload to the Cloud.
- If the offloading is not performed efficiently then the desired performance may not be achieved. Also the battery may get depleted faster than executing the workload locally.
- There are two main types of computational offloading:
    - Static: In which the offloading decisions (consisting of workload partitioning) are made at the execution start of a task or a batch of tasks.
    - Dynamic: The offloading decisions depend upon the run-time conditions of dynamic parameters such as network bandwidth, congestion and battery life etc.
- The static offloading decisions may not turn out to be fruitful if the dynamic parameters change unexpectedly.
- Better not to offload if the time/battery consumption (cost) for offloading is higher than the cost of locally processing the task.

Module No – 264: Mobile Cloud Computing: Issues: End User Related:

- **Incentives:** In the model of Mobile Cloud computing where a mobile device shares resources with other devices, there should be an incentive to share those resources.
- The incentive can be monetary or non-monetary. The non-monetary incentive can be of a common interest such as visitors of a museum trying to translate the same text on an artifact.
- **Presentation issues:** The mobile devices have a high level of heterogeneity. Developing separate user interface for each type of device is unrealistic.
- Since the display area on a mobile device is limited, designing a user friendly GUI requires extra effort.

- **Service Availability and Performance assurance issues:** Mobile devices undergo loss of connectivity due to signal loss, network error, battery depletion etc. and thus service availability and performance assurance turn to become challenges.

Module No – 265: Mobile Cloud Computing: Issues: Data Access:

- Data access over mobile Cloud applications may be challenging in case of low bandwidth, signal loss and/or battery life.
- Accessing the files through mobile devices may turn out to be expensive in terms of data transmission cost, network delays and energy consumption.
- Data access approaches are needed to be developed/polished to maintain a performance level and to save energy.
- Some approaches have optimized the data access patterns.
- Another approach is to use mobile cloudlets which are intermediate devices acting as file cache.
- Interoperability of data is also a challenge to provision data across heterogeneous devices and platforms. A generic representation of data should be preferred.

Module No – 266: Mobile Cloud Computing: Issues: Miscellaneous:

- **Performance:** The performance can be increased by optimally balancing the workload offloading.
    - If the IT resources are being acquired from the surrounding devices, then the performance depends upon the extent of resources available as compared to task at hand.

- **Resource Management:** A mobile Cloud application can acquire all the IT resources from Cloud. Another method is to use the cloudlets which are individual computers or even clusters in the vicinity of the mobile device running the mobile Cloud application.

    - In worst case, the mobile device resources are utilized. All these situations require separate resource management techniques.

Module No – 267: Mobile Cloud Computing Issues: Miscellaneous:

- **Processing Power:** The processing power of a single mobile device is not at all comparable to Cloud.
- The issue is how to efficiently utilize the huge processing power of Cloud to execute the tasks of mobile Cloud applications.
- **Battery Consumption:** Computational offloading becomes more energy efficient if the code size is large and vice versa.

- For example, offloading 500KB of code will take 5% of battery as compared to 10% battery usage if this code is locally processed. Thus 50% of battery is saved when offloading the code.
- But 250KB code is only 30% battery efficient if uploaded.

Module No – 268: Mobile Cloud Computing: Issues and Challenges:

- Support of mobility while assuring connectivity to the Cloud. The network connectivity becomes very important in this case. Cloudlets can support the connectivity but these only exist at certain locations such as cafés and malls.
- The adhoc creation of mobile Cloud over a set of a set of mobiles around a location depends upon the availability of capable devices and cost benefit analysis.
- Assurance of security is an on-going challenge to ensure privacy and security and to establish trust between the mobile device users and the service provider/resource provider.
- The conduct and manage the incentives among the resource lenders (in case of a mobile adhoc Cloud) requires the establishment of trust, payment method and methOds to prevent free riders.

Module No – 269: Mobile Cloud Computing vs Cloud Computing:

- Typically, both the Cloud computing and Mobile Cloud computing are dependent upon remote usage of IT resources offered by Cloud.
- Cloud computing traditionally works to provide various Cloud services such as IaaS, PaaS and SaaS etc. to the consumers.
- Mobile Cloud computing is however more towards providing Cloud based application over mobile devices and to deal with the connectivity, security and performance issues.
- Cloud computing deals with user requirements from a single user to an enterprise level.
- Mobile Cloud applications are more accessed by individual users for personal computing purposes.
- There are multiple models of Cloud computing such as Private, Public, Community and Hybrid.
- Mobile Cloud can be setup over Cloud, Cloudlets and on adhoc basis by using the capable and resource rich mobile devices sharing a common location on map.

Lesson No. 45

SPECIAL TOPICS IN CLOUD COMPUTING AND CONCLUSION OF COURSE

Module No – 270: Big Data Processing in Clouds: Overview of Big Data:

- The term "Big Data" refers to such enormous volume of data that can not be processed through traditional database technologies.

- The following are the three generic characteristics of big data:
    - Data volume is huge
    - Data may not be categorized into regular relational databases
    - The data is generated, captured and processed at high speed.
- Another definition of Big Data can be:
  Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. [Anuar, N.B., Gani, A., Hashem, I.A., Khan, S.U., Mokhtar, S., & Yaqoob, I. (2015). The rise of "big data" on cloud computing: Review and open research issues. Inf. Syst., 47, 98-115]

Module No – 271: Big Data Processing in Clouds: Overview of Big Data:

- The characteristics of big data can be represented by 4vs:
    - **Volume:** The volume of data continues to rise.
    - **Variety:** It refers to the heterogeneous types of data collected through sensors, smartphones, social networks etc. and can be of video, audio, image, text formats. The data can be structured, semi structured or unstructured.
    - **Velocity:** It refers to the speed of transfer of data. The contents of the data also change constantly.
    - **Value:** The analysis of Big Data can generate valuable information.

Module No – 272: Relationship between Cloud computing and Big Data:

- Cloud computing infrastructure can fulfill the data storage and processing requirements to store and analyze the Big Data.
- The data can be stored in large fault tolerant databases. Processing can be performed trough parallel and distributed algorithms.
- Cloud storage can be used to host Big Data while the processing can be done locally on commodity computers.
- Big data Cloud applications can be built to host and process the big data on Cloud.
- There are three popular models for big data:
    - Distributed Map Reduce model popularized by Hadoop
    - NoSQL model used for non-relational, non-tabular storage
    - SQL RDBMS model for relational tabular storage of structured data.
- Traditional tools for big data processing for example can be deployed over Cloud.
- Top-rated Hadoop options include Apache Hadoop, SAP's HANA/Hadoop combination, Hortonworks, Hadapt and VMware's Cloud Foundry, as well as services provided by IBM, Microsoft and Oracle.
- For NoSQL, consider Cassandra, Hbase or MongoDB. IBM also offers NoSQL for the cloud, and there are plenty of other NoSQL providers.
- [http://searchcloudapplications.techtarget.com/tip/How-to-choose-the-best-cloud-big-data-platform]

Module No – 273: Big Data on Cloud Case Studies:

- In this module we shall cover a few examples of usage of Cloud computing for Big Data hosting and processing as case studies.
- **SwiftKey:** It is a smart prediction technology for mobile device virtual keyboards.
    - Terabytes of data is collected and analyzed for active users around the globe for prediction and correction of text through an artificial engine.
    - Uses Amazon Simple Storage Service and Amazon Elastic Cloud to host Hadoop.
- **Halo Game:** More than 50 million copies have been sold worldwide.
    - Collects the game usage data for the players globally for player-ranking in online gaming tournaments.
    - Windows Azure HDInsight Service (based on Hadoop) is used for this purpose.
- **Nokia:** The well known mobile manufacturer collects terabytes of data for analysis,
    - Uses Teradata Enterprise Data Warehouse, Oracle and MySQL data marts, visualization technologies, and Hadoop.

Module No – 274: Big Data Storage and Data processing in Clouds:

- Popular platforms for Big Data processing are:
    - **Hadoop:** It is an open-source project of Apache Software Foundation. It is Java based. Used for batch jobs processing. Has two primary components:
        - Hadoop Distributed File System (HDFS)
        - MapReduce programming framework
    - **Spark:** It is a data processing framework compatible with Hadoop data sources. Suitable for machine learning tasks. Faster than MapReduce. However it is unable to execute concurrent Reduce methods unlike MapReduce of Hadoop.
- Hadoop MapReduce is however currently a more popular computational model of Cloud providers for Big Data processing.

Module No – 275: Challenges & Issues of Big Data processing on Cloud:

- In this module we shall briefly discuss a few challenges and issues related to Big Data processing on Cloud.
- Scalability assurance for storage of rising volume of Big Data.
- Availability assurance of any data out of Big Data stored on Cloud storage is a challenge.
- Data quality refers to the possibility of data-source verification. It is a challenging task for Big Data (for example) collected from mobile phones.
- Simultaneously handling heterogeneous data is challenging.
- Privacy issues arise when the processing of Big Data (through data mining techniques) may lead to sensitive and personal information. Another issue is the lack of established laws and regulations in this regard.

Module No – 276: Multimedia Cloud Computing: Overview & Introduction:

- Internet multimedia is emerging as a service with the development of Web 2.0.
- Multimedia computing has emerged as a prominent technology to provide rich media contents.
- Millions of subscribers and users worldwide obtain multimedia over heterogeneous devices.
- Thus multimedia rendering can rely upon Cloud computing for providing the processing and storage resources.
- Utilizing multimedia-Cloud can have advantages as well as challenges.
- The advantages include the provision of scalable IT resources and the fault tolerant Cloud infrastructure.
- The challenges include the rendering of a variety of multimedia types, formats and services. Maintaining QoS across heterogeneous devices, users and multimedia contents and services is yet another challenge.
- [Zhu, W., Luo, C., Wang, J., & Li, S. (2011). Multimedia cloud computing. IEEE Signal Processing Magazine, 28(3), 59-69.]

Module No – 277: Multimedia Cloud Computing:

- **Architecture & Processing:** In this module we shall consider the example of multimedia edge Cloud (MEC) consisting of cloudlets.
- The multimedia Cloud providers can use the IT resources of cloudlets which are physically placed over the edge (means very close to the multimedia service consumers) to reduce network latencies.
- There can be multiple MECs which are geographically distributed.
- The MECs are connected to central servers through content delivery network (CDN).
- The MECs provide multimedia services and maintain the QoS.

Module No – 278: Cloud-aware Multimedia Applications & Rendering:

- The Cloud offers various applications/services for multimedia. Such as:
- Storage: There are various providers of multimedia storage services for example IOmega (www.iomega.com) , AllMyData (www.amdwebhost.com) etc.
- **Sharing:** The multimedia contents once uploaded, can be shared to other users. The cloud-client network link is usually more robust, capable and reliable than client-client link.
- **Multimedia authoring:** Refers to the services and tools for editing, merging and enhancing the multimedia contents over Cloud.
- **Rendering:** Refers to the creation of images and multimedia to be displayed over the user's device. The multimedia Cloud can perform rendering for less capable devices in order to maintain QoS and minimize the delay.

Module No – 279: Introduction to SDN:

- Network operators often have to configure the devices (switches & routers) separately and by using vendor specific commands.
- Thus, implementing high level network policies is hard and complex in traditional IP networks.
- The dynamic response and reconfiguration is almost non-existent in current IP networks. Enforcing the network policies dynamically is therefore challenging.
- Further, the control plane (the decision making and forwarding rules) and the data plane (which performs traffic forwarding according to the decisions made by control plane) are bundled inside the networking device.
- All this reduces the flexibility, innovation and evolution of the networking infrastructure.
- Software Defined Networking (SDN) is the new paradigm of networking that separates the control plane from data plane.
- It reduces the limitations of traditional networks.
- The switches become the forwarding-only devices, while the control plane is handled by a software controller.
- The controller and switch have a software interface between them.
- Controller directly exercises direct control over the data plane devices through a well defined application program interface (API) such as OpenFlow.
    - [Jain, R., & Paul, S. (2013). Network virtualization and software defined networking for cloud computing: a survey. IEEE Communications Magazine, 51(11), 24-31.]
    - [Kreutz, D., Ramos, F. M., Verissimo, P. E., Rothenberg, C. E., Azodolmolky, S., & Uhlig, S. (2015). Software-defined networking: A comprehensive survey. Proceedings of the IEEE, 103(1), 14-76.]
    - [Nunes, B. A. A., Mendonca, M., Nguyen, X. N., Obraczka, K., & Turletti, T. (2014). A survey of software-defined networking: Past, present, and future of programmable networks. IEEE Communications Surveys & Tutorials, 16(3), 1617-1634.]

Module No – 280: History of SDN:

- SDN has its roots in history as long ago as 80s and 90s with the development of Network Control Point (NCP) technology.
- NCP was introduced by AT&T as probably the first established technique to separate the data plane and control plane.
- *Active Networks* was another attempt to introduce computational and packet modification capabilities to the network nodes.
- *Network virtualization* is a recent development which allows hypervisor like environment to network infrastructure.
- OpenFlow based network operating systems such as ONOS have emerged to make network administration easier and to develop/deploy new protocols and management applications.

Module No – 281: Network Virtualization:

- Each computer system needs at least one L2 NIC (Ethernet card) for communication.
- A physical system must have at least one physical NIC (pNIC).
- Each VM has at least one virtual NIC (vNIC).
- All the vNICs on a physical host (server) are interconnected through a virtual switch (vSwitch).
- The vSwitch is connected to the pNIC.
- Multiple pNICs are connected to a physical switch (pSwitch)
- There are a number of standards available for NIC virtualization.
- A physical ethernet switch can be virtualized by implementing IEEE Bridge Port Extension standard 802.1BR
- The VLANS can span over multiple data centers and there are several approaches to manage the VLANS.
- A VM can be migrated across different data centers by following multiple techniques proposed by researchers.
- The modern processors allow the implementation of software based network devices such as L2 switch, L3 router etc.

Module No – 282: Architecture of SDN:

- SDN has four characteristics:
    - **Separation of control and data planes:** In SDN, the switches are forwarding devices. The network traffic is forwarded by following the forwarding tables. The forwarding tables are created/updated through a controller software.
    - **Centralization of control plane:** SDN relies upon centralized control instead of distributed control.
    - The controller software actually controls a subset of the whole network which is small enough to be controlled by a single controller software.
    - **Programmable control plane:** The controller software is controllable through API calls.
    - This helps in rapid implementation of network policies because the control plane is centralized and not distributed.
    - **Standardized APIs:** There is a southbound API for communication with forwarding devices and a northbound API for communication with network applications.
    - The main southbound API is OPenFlow standardized by the Open Networking Foundation. The northbound APIs have not been standardized yet.

Module No – 283: SDN In Cloud:

- The implementation of SDN in Cloud is an ongoing research work.

- The rise in demand for network virtualization has attracted the virtualization software vendors to integrate SDN features in their products.
- Centralized controllers such as *Beacon* can handle more than 12 million flows per second and can fulfill the requirements of enterprise level networks and data centers for hosting Cloud.
- The SDN can be helpful in monitoring, filtering and managing the network traffic over virtual as well as physical networks inside a Cloud hosting data center.

Module No – 284: Future of SDN:

- In near future, the SDN has to overcome a few challenges.
- Controller and switch design
- Scalability and performance in SDNs
- Controller-service interfacing (specially the northbound interface)
- Virtualization and cloud service applications (the SDN based Cloud infrastructure)
- Information centric networking: The future architecture of Internet aiming at increasing the efficiency of content delivery and availability.
- Enabling heterogeneous networking with SDN: The wired, wireless and adhoc networks.

Module No – 285: Fog Computing:

- It is an emerging paradigm of Cloud computing.
- *Fog Computing* or *Fog* extends the Cloud computing and services to the edge of the network.
- Provides data, computing, storage ad application services to end-users that can be hosted at the network edge or end devices such as set-top-boxes or access points.
- Fog will support Internet of Everything (IoE) applications such as industrial automation, transportation, network of sensors and actuators etc..
- These applications demand real-time/predictable latency and mobility.
- Fog can therefore be considered a candidate technology for beyond 5G networks.
- Fog will result in the defusing of Cloud among the client devices.
- Fog Computing is a scenario where huge number of heterogeneous, ubiquitous and decentralized devices communicate and potentially cooperate among them and the network to perform storage and processing tasks without the intervention of third parties.
- Network virtualization and SDN are going to be the essential parts of Fog computing.
    - [Kitanov, S., Monteiro, E., & Janevski, T. (2016, April). 5G and the Fog—Survey of related technologies and research directions. In Electrotechnical Conference (MELECON), 2016 18th Mediterranean (pp. 1-6). IEEE]

Module No – 286 Cloud Gaming:

- Online gaming consists of four basic building blocks:
- Input module: To receive control information from the player

- Game logic module: To process the game events
- Networking module: To process how the avatars interact with each other
- Rendering module: Formulates the game video and presents it to the players.
- Traditionally, the Internet gaming servers are used to exchange the game information while the input, logic and rendering modules of the game are installed on users' devices.
- Computer games can be hosted as service on Cloud.
- One of the deployment models is to deploy only the I/O module on users' devices and the rest of the modules on Cloud.
- The benefits include the easy deployment, reduced cost and time for the players and formulation of various flexible business models such as pre-pay, post-pay and pay-per-play etc.
    - Cai, W., Chen, M., & Leung, V. C. (2014). Toward gaming as a service. IEEE Internet Computing, 18(3), 12-18.]

Module No – 287: Conclusion & End of Course:

- Cloud computing provides the computing/IT resources to the users over the Internet in a pay-as-you-go type of business model.
- This course has covered almost all the aspects of Cloud computing and the advanced topics related to Cloud.
- We are hopeful that you will find this course interesting, informative and comprehensive.
- We hope that the students of Cloud Computing subject will surely know the importance and ubiquity of Cloud computing.
- We hope that this subject will become the foundation of advanced courses and an initial source of knowledge for all in this regard.

Module No – 288: Short Revision:

- In this course we have had an in-depth study of Cloud computing.
- We came across various technologies forming the building blocks of the Cloud.
- We had detailed study of Cloud deployment and study models.
- We also look at various Cloud mechanisms.
- We also studied some advanced topics related to Cloud computing.
- In particular, the main topics of the course are as follows:
    - Introduction
    - History of Cloud Computing
    - Background (Basics Knowledge)
    - Background (Advanced Knowledge)
    - Cloud Computing in Detail
    - Benefits and Challenges of Cloud Computing
    - Roles and Boundaries of Cloud Computing
    - Cloud Service Models

- Data Storage in Cloud
- Miscellaneous Services of Cloud Computing
- Cloud Deployment Models
- Cloud Security
- Trust Issues in Cloud
- Cloud Infrastructure Basics
- Service Level Agreement
- Cloud Hosting Data Center Design
- Cloud Architectures
- Specialized Cloud Mechanisms
- Cloud Management
- Specialized Cloud Architecture
- Cloud Federation
- Cloud Brokerage
- Cloud Costing Models
- Cloud Metrics
- Privacy Issues for Cloud Computing
- Security Issues for Cloud Computing
- Open Issues in Cloud
- Disaster Recovery in Cloud Computing
- General Recommendations of Cloud Computing
- Migrating to the Clouds
- Designing Cloud based Solutions
- Cloud Applications Scalability
- Cloud Resource Scheduling
- Mobile Cloud Computing
- Special Topics in Cloud Computing
- Advanced Topics
- Cloud Resource Scheduling
- Mobile Cloud Computing
- Special Topics in Cloud Computing
- Advanced Topics

- **Reference Books:**
  - Thomas Erl [2014], Cloud Computing Concepts, Technology and Architecture, Pearson
  - Jamsa, K. (2012). Cloud computing. Jones & Bartlett Publishers
  - Pearson, S., & Yee, G. (Eds.). [2012]. Privacy and security for cloud computing. Springer Science & Business Media.
  - Cloud Computing Synopsis and Recommendations, NIST USA, http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-146.pdf
- **Instructor's email:** mtayyabch@yahoo.com