UNIT III CLOUD ARCHITECTURE, SERVICES AND STORAGE

Layered Cloud Architecture Design – NIST Cloud Computing Reference Architecture – Public, Private and Hybrid Clouds - laaS – PaaS – SaaS – Architectural Design Challenges – Cloud Storage – Storage-as-a-Service – Advantages of Cloud Storage – Cloud Storage Providers – S3.

3.1 LAYERED ARCHITECTURE:

Generic Cloud Architecture Design:

An Internet cloud is envisioned as a public cluster of servers provisioned on demand to perform collective web services or distributed applications using data-center resources.

- ❖ Cloud Platform Design Goals
- ❖ Enabling Technologies for Clouds
- ❖ A Generic Cloud Architecture

Cloud Platform Design Goals

- Scalability
- Virtualization
- Efficiency
- Reliability
- Security

Cloud management receives the user request and finds the correct resources. Cloud calls the provisioning services which invoke the resources in the cloud. Cloud management software needs to support both physical and virtual machines

Enabling Technologies for Clouds

- Cloud users are able to demand more capacity at peak demand, reduce costs, experiment with new services, and remove unneeded capacity.
- Service providers can increase system utilization via multiplexing, virtualization and dynamic resource provisioning.
- Clouds are enabled by the progress in hardware, software and networking technologies
- Cloud users are able to demand more capacity at peak demand, reduce costs, experiment with new services, and remove unneeded capacity.
- Service providers can increase system utilization via multiplexing, virtualization and dynamic resource provisioning.

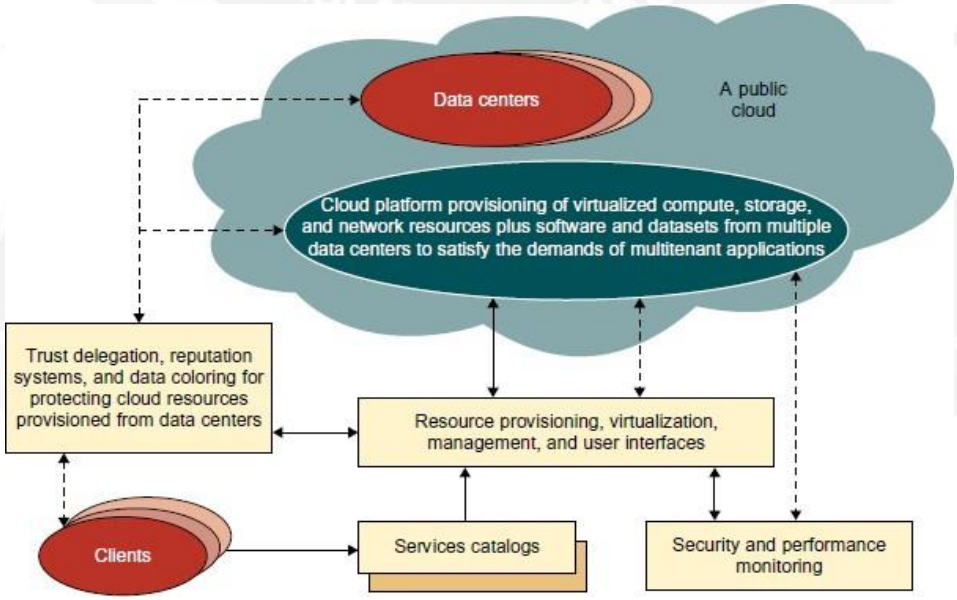
- Clouds are enabled by the progress in hardware, software and networking technologies

Table 4.3 Cloud-Enabling Technologies in Hardware, Software, and Networking

Technology	Requirements and Benefits
Fast platform deployment	Fast, efficient, and flexible deployment of cloud resources to provide dynamic computing environment to users
Virtual clusters on demand	Virtualized cluster of VMs provisioned to satisfy user demand and virtual cluster reconfigured as workload changes
Multitenant techniques	SaaS for distributing software to a large number of users for their simultaneous use and resource sharing if so desired
Massive data processing	Internet search and web services which often require massive data processing, especially to support personalized services
Web-scale communication	Support for e-commerce, distance education, telemedicine, social networking, digital government, and digital entertainment applications
Distributed storage	Large-scale storage of personal records and public archive information which demands distributed storage over the clouds
Licensing and billing services	License management and billing services which greatly benefit all types of cloud services in utility computing

A Generic Cloud Architecture

- The Internet cloud is envisioned as a massive cluster of servers.
- Servers are provisioned on demand to perform collective web services using data-center resources.
- The cloud platform is formed dynamically by provisioning or deprovisioning servers, software, and database resources.
- Servers in the cloud can be physical machines or VMs.
- User interfaces are applied to request services.

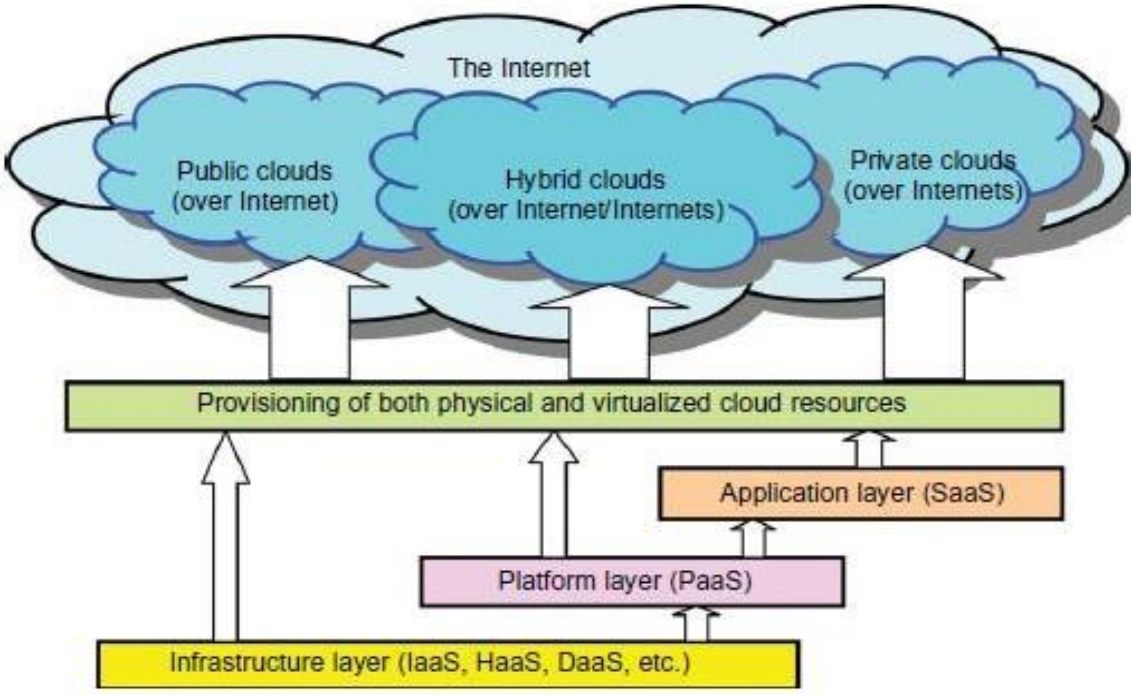


- The cloud computing resources are built into the data centers.
- Data centers are typically owned and operated by a third-party provider.

Consumers do not need to know the underlying technologies

- In a cloud, software becomes a service.
- Cloud demands a high degree of trust of massive amounts of data retrieved from large data centers.
- The software infrastructure of a cloud platform must handle all resource management and maintenance automatically.
- Software must detect the status of each node server joining and leaving.
- Cloud computing providers such as Google and Microsoft, have built a large number of data centers.
- Each data center may have thousands of servers.
- The location of the data center is chosen to reduce power and cooling costs.

Layered Cloud Architectural Development



- The architecture of a cloud is developed at three layers
  - Infrastructure
  - Platform
  - Application

- Implemented with virtualization and standardization of hardware and software resources provisioned in the cloud.

The services to public, private and hybrid clouds are conveyed to users through networking support

Infrastructure Layer

- Foundation for building the platform layer.
- Built with virtualized compute, storage, and network resources.
- Provide the flexibility demanded by users.
- Virtualization realizes automated provisioning of resources and optimizes the infrastructure management process.

Platform Layer

- Foundation for implementing the application layer for SaaS applications.
- Used for general-purpose and repeated usage of the collection of software resources.
- Provides users with an environment to develop their applications, to test operation flows, and to monitor execution results and performance.

The platform should be able to assure users that they have scalability, dependability, and security protection

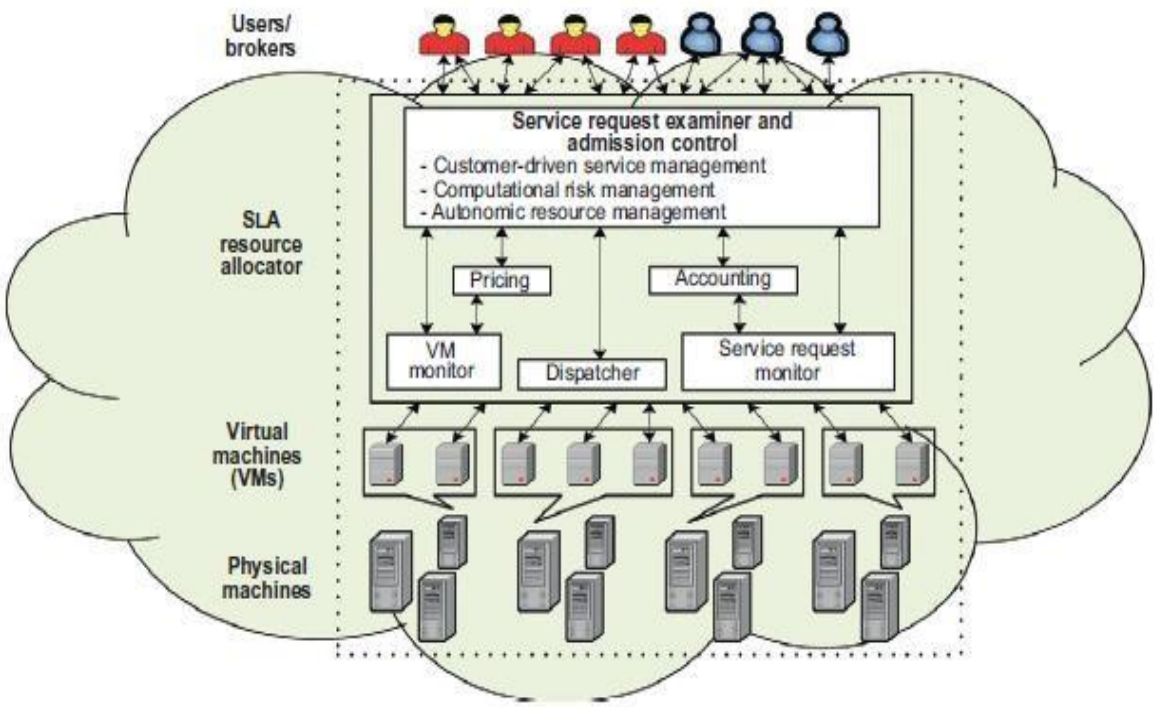
Application Layer

- Collection of all needed software modules for SaaS applications.
- Service applications in this layer include daily office management work, such as information retrieval, document processing, and authentication services.
- The application layer is also heavily used by enterprises in business marketing and sales, consumer relationship management (CRM) and financial transactions.
- Not all cloud services are restricted to a single layer.
- Many applications may apply resources at mixed layers.
- Three layers are built from the bottom up with a dependence relationship.

Market-Oriented Cloud Architecture

- High-level architecture for supporting market-oriented resource allocation in a cloud computing environment.
- Users or brokers acting on user's behalf submit service requests to the data center.
- When a service request is first submitted, the service request examiner interprets the submitted request for QoS requirements.

Accept or Reject the request.



- **VM Monitor**: Latest status information regarding resource availability.
- **Service Request Monitor**: Latest status information workload processing
- **Pricing mechanism:**Decides how service requests are charged.
- **Accounting mechanism:**Maintains the actual usage of resources by requests to compute the final cost.
- VM Monitor mechanism keeps track of the availability of VMs and their resource entitlements.
- Dispatcher starts the execution of accepted service requests on allocated VMs.

  Service Request Monitor mechanism keeps track of the execution progress of service requests.

  Multiple VMs can be started and stopped on demand

Quality of Service Factors

QoS parameters

- Time
- Cost
- Reliability
- Trust/security

  QoS requirements cannot be static and may change over time.

3.1.1 CLOUD REFERENCE ARCHITECTURE

Definitions

- A model of computation and data storage based on “pay as you go” access to “unlimited” remote data center capabilities.
- A cloud infrastructure provides a framework to manage scalable, reliable, on-demand access to applications.
- Cloud services provide the “invisible” backend to many of our mobile applications.

High level of elasticity in consumption.

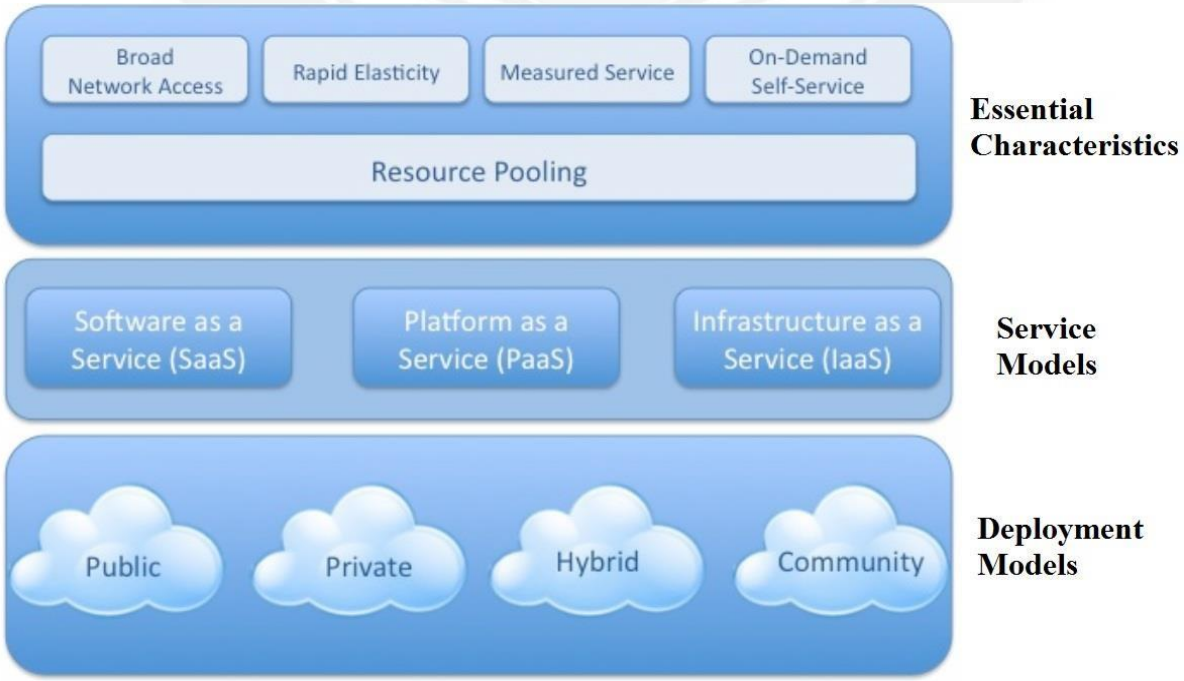
NIST Cloud Definition:

The National Institute of Standards and Technology (NIST) defines cloud computing as a

> **"pay-per-use model for enabling available, convenient and 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.**"

Architecture

- Architecture consists of 3 tiers
  - Cloud Deployment Model
  - Cloud Service Model
  - Essential Characteristics of Cloud Computing .



Essential Characteristics 1

- 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 a service provider.

Essential Characteristics 2

- Broad network access.
  - Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs) as well as other traditional or cloudbased software services.

Essential Characteristics 3

- Resource pooling.
  - The provider's computing resources are pooled to serve multiple consumers using a **multi-tenant model**, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.

Essential Characteristics 4

- **Rapid elasticity.**
  - Capabilities can be rapidly and elastically provisioned - in some cases automatically - to quickly scale out; and rapidly released to quickly scale in.
  - To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

Essential Characteristics 5

- **Measured service.**
  - Cloud systems automatically control and optimize resource usage byleveraging a metering capability at some level of abstraction appropriate to thetype of service.

Resource usage can be monitored, controlled, and reported - providing transparency for both the provider and consumer of the service.