**UNIT – III**

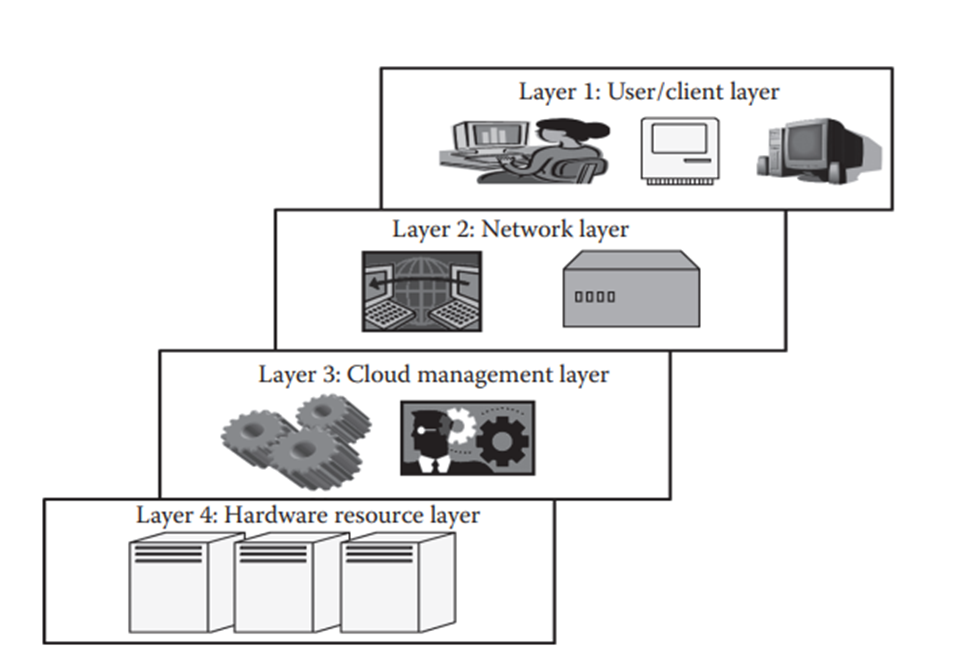
**Cloud Computing Architecture and Management:** Cloud architecture, Layer, Anatomy of the Cloud, Network Connectivity in Cloud Computing, Applications, on the Cloud, Managing the Cloud, Managing the Cloud Infrastructure Managing the Cloud application, Migrating Application to Cloud, Phases of Cloud Migration Approaches for Cloud Migration.

**Cloud architecture**

The cloud also has an architecture that describes its working mechanism. It includes the dependencies on which it works and the components that work over it. The cloud is a recent technology that is completely dependent on the Internet for its functioning.

The cloud architecture can be divided into four layers based on the access of the cloud by the user. They are as follows

Layer 1 (User/Client Layer) This layer is the lowest layer in the cloud architecture. All the users or client belong to this layer. This is the place where the client/user initiates the connection to the cloud. The client can be any device such as a thin client, thick client, or mobile or any handheld device that would support basic functionalities to access a web application.



Layer 2 (Network Layer)

This layer allows the users to connect to the cloud. The whole cloud infrastructure is dependent on this connection where the services are offered to the customers. The public cloud can be accessed all over the world. In the case of a private cloud, the connectivity may be provided by a local area network (LAN).

Layer 3 (Cloud Management Layer)

This layer consists of software’s that are used in managing the cloud. The softwares can be a cloud operating system (OS), a software that acts as an interface between the data center (actual resources) and the user, or a management software that allows managing resources. These softwares usually allow resource management (scheduling, provisioning, etc.), optimization (server consolidation, storage workload consolidation), and internal cloud governance. These SLAs are for both private and public clouds Popular service providers are Amazon Web Services (AWS) and Microsoft Azure for public cloud.

Layer 4 (Hardware Resource Layer)

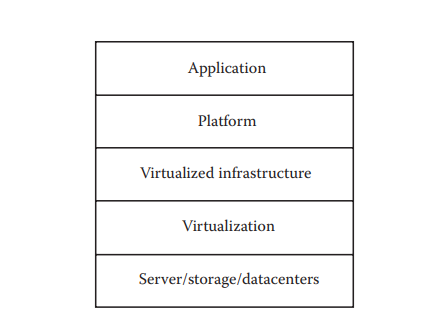
Layer 4 consists of provisions for actual hardware resources. Usually, in the case of a public cloud, a data center is used in the back end.

This layer comes under the purview of SLAs. This is the most important layer that governs the SLAs. This layer affects the SLAs most in the case of data centers.

Thus, this is the architecture of a cloud. The layering is strict, and for any cloud application, this is followed. There can be a little loose isolation between layer 3 and layer 4 depending on the way the cloud is deployed.

**Anatomy of the Cloud**

Cloud anatomy can be simply defined as the structure of the cloud. Cloud anatomy cannot be considered the same as cloud architecture. It may not include any dependency on which or over which the technology works,



Architecture is a hierarchical structural view that defines the technology as well as the technology over which it is dependent or/and the technology that are dependent on it.

1. Application: The upper layer is the application layer. In this layer, any applications are executed.
2. Platform: This component consists of platforms that are responsible for the execution of the application. This platform is between the infrastructure and the application.
3. Infrastructure: The infrastructure consists of resources over which the other components work. This provides computational capability to the user.
4. Virtualization: Virtualization is the process of making logical components of resources over the existing physical resources. The logical components are isolated and independent, which form the infrastructure.
5. Physical hardware: The physical hardware is provided by server and storage units.

**Network Connectivity in Cloud Computing**

Cloud computing is a technique of resource sharing where servers, storage, and other computing infrastructure in multiple locations are connected by networks. In the cloud, when an application is submitted for its execution, needy and suitable resources are allocated from this collection of resources; as these resources are connected via the Internet, the users get their required results.

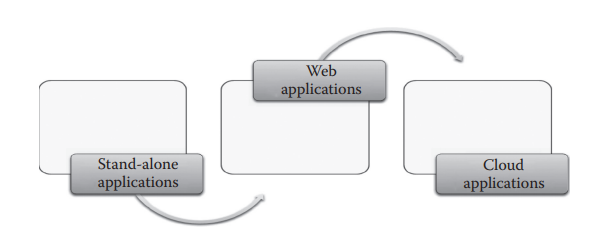
For many cloud computing applications, network performance will be the key issue to cloud computing performance. Since cloud computing has various deployment options, we now consider the important aspects related to the cloud deployment models and their accessibility from the viewpoint of network connectivity.

**Applications on the Cloud**

The power of a computer is realized through the applications. There are several types of applications. The first type of applications that was developed and used was a stand-alone application.

A stand-alone application is developed to be run on a single system that does not use network for its functioning. These stand-alone systems use only the machine in which they are installed. The functioning of these kinds of systems is totally dependent on the resources or features available within the system.

These systems do not need the data or processing power of other systems; they are self-sustaining. But as the time passed, the requirements of the users changed and certain applications were required, which could be accessed by other users away from the systems. This led to the inception of web application.



Though this application is much used, there are shortcomings as discussed in the following:

• The web application is not elastic and cannot handle very heavy loads, that is, it cannot serve highly varying loads.

• The web application is not multitenant.

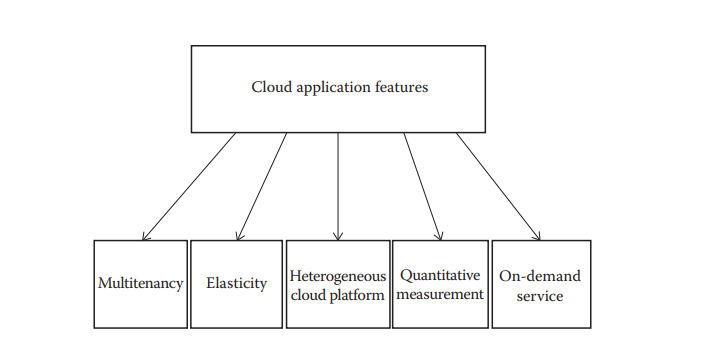
• The web application does not provide a quantitative measurement of the services that are given to the users, though they can monitor the user.

• The web applications are usually in one particular platform.

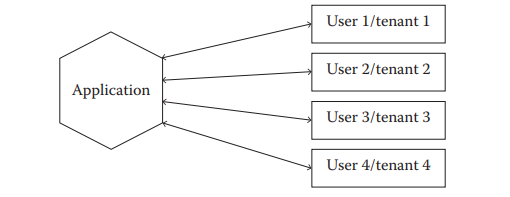
• The web applications are not provided on a pay-as-you-go basis; thus, a particular service is given to the user for permanent or trial use and usually the timings of user access cannot be monitored.

• Due to its non elastic nature, peak load transactions cannot be handled.

The cloud as mentioned can be classified into three broad access or service models, Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). Cloud application in general refers to a SaaS application.



1. Multitenancy: Multitenancy is one of the important properties of cloud that make it different from other types of application in which the software can be shared by different users with full independence. Here, independence refers to logical independence.
2. Elasticity: Elasticity is also a unique property that enables the cloud to serve better. According to Herbst et al. [4], elasticity can be defined as the degree to which a system is able to adapt to workload changes by provisioning and de provisioning resources in an autonomic manner such that at each point in time, the available resources match the current demand as closely as possible. Elasticity allows the cloud providers to efficiently handle the number of users, from one.



3. Heterogeneous cloud platform: The cloud platform supports heterogeneity, wherein any type of application can be deployed in the cloud. Because of this property, the cloud is flexible for the developers, which facilitates deployment. The applications that are usually deployed can be accessed by the users using a web browser.

4. Quantitative measurement: The services provided can be quantitatively measured. The user is usually offered services based on certain charges. Here, the application or resources are given as a utility on a pay-per-use basis. Thus, the use can be monitored and measured. Not only the services are measureable, but also the link usage and several other parameters that support cloud applications can be measured. This property of measuring the usage is usually not available in a web application and is a unique feature for cloud-based applications.

5. On-demand service: The cloud applications offer service to the user, on demand, that is, whenever the user requires it. The cloud service would allow the users to access web applications usually without any restrictions on time, duration, and type of device used.

**Managing the Cloud**

Cloud management is aimed at efficiently managing the cloud so as to maintain the QoS. It is one of the prime jobs to be considered. The whole cloud is dependent on the way it is managed. Cloud management can be divided into two parts: 1. Managing the infrastructure of the cloud 2. Managing the cloud application.

Managing the Cloud Infrastructure The infrastructure of the cloud is considered to be the backbone of the cloud. This component is mainly responsible for the QoS factor. If the infrastructure is not properly managed, then the whole cloud can fail and QoS wou

The core of cloud management is resource management. Resource management involves several internal tasks such as resource scheduling, provisioning, and load balancing. These tasks are mainly managed by the cloud service provider’s core software capabilities such as the cloud OS that is responsible for providing services to the cloud and that internally controls the cloud. A cloud infrastructure is a very complex system that consists of a lot of resources. These resources are usually shared by several users.

Poor resource management may lead to several inefficiencies in terms of performance, functionality, and cost. If a resource is not efficiently managed, the performance of the whole system is affected. Performance is the most important aspect of the cloud, because everything in the cloud is dependent on the SLAs and the SLAs can be satisfied only if performance is good.

At a higher level, other than these three issues, there are few more issues that depend on resource management. These are power consumption and optimization of multiple objectives to further reduce the cost. To accomplish these tasks, there are several approaches followed, namely, consolidation of server and storage workloads. Consolidation would reduce the energy consumption and in some cases would increase the performance of the cloud. According to Margaret Rouse [5], server consolidation by definition is an approach to the efficient usage of computer server resources in order to reduce the total number of servers or server locations that an organization requires

Load fluctuation is the point where the workload of the system changes continuously. This is one of the important criteria and issues that should be considered for cloud applications. Load fluctuation can be divided into two types: predictable and unpredictable. Predictable load fluctuations are easy to handle. The cloud can be preconfigured for handling such kind of fluctuations. Whereas unpredictable load fluctuations are difficult to handle, ironically this is one of the reasons why cloud is preferred by several users.

**Managing the Cloud Application**

Business companies are increasingly looking to move or build their corporate applications on cloud platforms to improve agility or to meet dynamic requirements that exist in the globalization of businesses and responsiveness to market demands. But, this shift or moving the applications to the cloud environment brings new complexities.

So, understanding the availability of an application requires inspecting the infrastructure, the services it consumes, and the upkeep of the application. The composite nature of cloud applications requires visibility into all the services to determine the overall availability and uptime.

Cloud application management is to address these issues and propose solutions to make it possible to have insight into the application that runs in the cloud, as well as implement or enforce enterprise policies like governance and auditing and environment management while the application is deployed in the cloud.

**Migrating Application to Cloud**

Cloud migration encompasses moving one or more enterprise applications and their IT environments from the traditional hosting type to the cloud environment, either public, private, or hybrid. Cloud migration presents an opportunity to significantly reduce costs incurred on applications. This activity comprises, of different phases like evaluation, migration strategy, prototyping, provisioning, and testing.

**Phases of Cloud Migration**

1. Evaluation: Evaluation is carried out for all the components like current infrastructure and application architecture, environment in terms of compute, storage, monitoring, and management, SLAs, operational processes, financial considerations, risk, security, compliance, and licensing needs are identified to build a business case for moving to the cloud.

2. Migration strategy: Based on the evaluation, a migration strategy is drawn—a hot plug strategy is used where the applications and their data and interface dependencies are isolated and these applications can be operationalized all at once. A fusion strategy is used where the applications can be partially migrated; but for a portion of it, there are dependencies based on existing licenses, specialized server requirements like mainframes, or extensive interconnections with other applications.

3. Prototyping: Migration activity is preceded by a prototyping activity to validate and ensure that a small portion of the applications are tested on the cloud environment with test data setup.

4. Provisioning: Pre migration optimizations identified are implemented. Cloud servers are provisioned for all the identified environments, necessary platform software’s and applications are deployed, configurations are tuned to match the new environment sizing, and databases and files are replicated. All internal and external integration points are properly configured. Web services, batch jobs, and operation and management software are set up in the new environments.

5. Testing: Post migration tests are conducted to ensure that migration has been successful. Performance and load testing, failure and recovery testing, and scale-out testing are conducted against the expected traffic load and resource utilization levels.

**Approaches for Cloud Migration**

The following are the four broad approaches for cloud migration that have been adopted effectively by vendors:

1. Migrate existing applications: Rebuild or rearchitect some or all the applications, taking advantage of some of the virtualization technologies around to accelerate the work. But, it requires top engineers to develop new functionality. This can be achieved over the course of several releases with the timing determined by customer demand.

2. Start from scratch: Rather than cannibalize sales, confuse customers with choice, and tie up engineers trying to rebuild existing application, it may be easier to start again. Many of the R&D decisions will be different now, and with some of the more sophisticated development environments, one can achieve more even with a small focused working team.

3. Separate company: One may want to create a whole new company with separate brand, management, R&D, and sales. The investment and internet protocol (IP) may come from the existing company, but many of the conflicts disappear once a new born in the cloud company is established. The separate company may even be a subsidiary of the existing company. What is important is that the new company can act, operate, and behave like a cloud-based start-up.

4. Buy an existing cloud vendor: For a large established vendor, buying a cloud-based competitor achieves two things. Firstly, it removes a competitor, and secondly, it enables the vendor to hit the ground running in the cloud space. The risk of course is that the innovation, drive, and operational approach of the cloud-based company are destroyed as it is merged into the larger acquirer