Unit 2
Cloud Computing Architecture and Management

1-Cloud Architecture

2-Anatomy of the Cloud

3- Network Connectivity in Cloud Computing

4- Applications on the Cloud

5- Managing the Cloud

6- Migrating Application to Cloud

Overview

This unit firstly describes the cloud architecture. Cloud architecture consists of a hierarchical set of components that collectively describe the way the cloud works. The next section explains about the cloud anatomy, followed by network connectivity in the cloud and then the fine details about managing a cloud application. Finally, an overview on migrating applications to the cloud is discussed.

1-Cloud Architecture

Any technological model consists of an architecture based on which the model functions, which is a hierarchical view of describing the technology. 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.

1.1 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. The thin client here refers to a device that is completely dependent on some other system for its complete functionality. In simple terms, they have very low processing capability. Similarly, thick clients are general computers that have adequate processing capability. They have sufficient capability for independent work. Usually, a cloud application can be accessed in the same way as a web application. But internally, the properties of cloud applications are significantly different. Thus, this layer consists of client devices.

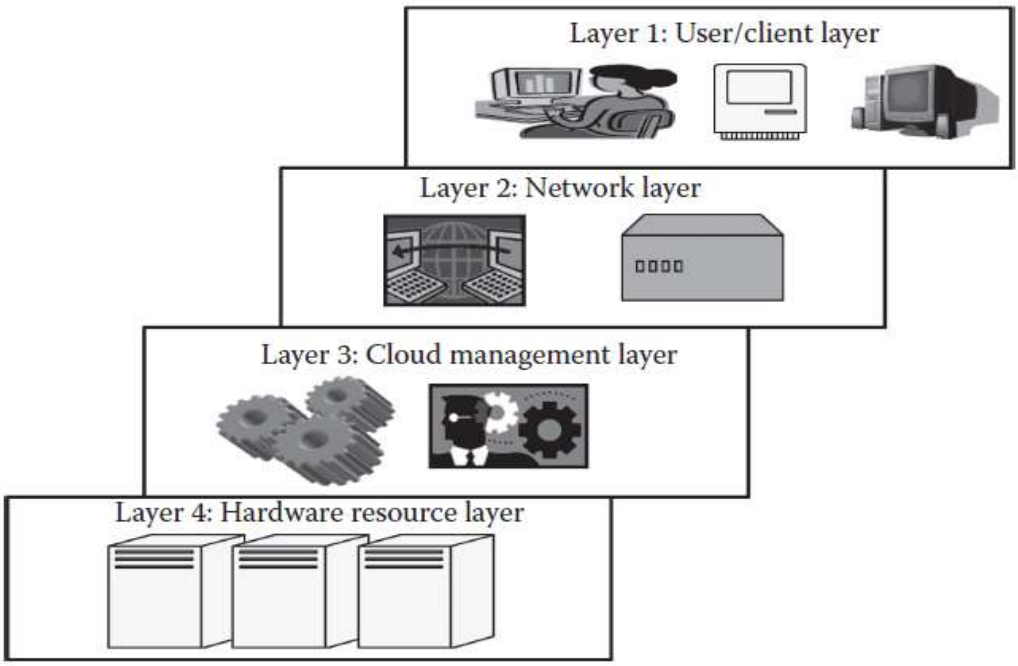

Fig. 1. Cloud architecture.

1.2 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. This is primarily the Internet in the case of a public cloud. The public cloud usually exists in a specific location and the user would not know the location as it is abstract. And, 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). Even in this case, the cloud completely depends on the network that is used. Usually, when accessing the public or private cloud, the users require minimum bandwidth, which is sometimes defined by the cloud providers. This layer does not come under the purview of service-level agreements (SLAs), that is, SLAs do not take into account the Internet connection between the user and cloud for quality of service (QoS).

1.3 Layer 3 (Cloud Management Layer)

This layer consists of softwares 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. This layer comes under the purview of SLAs, that is, the operations taking place in this layer would affect the SLAs that are being decided upon between the users and the service providers. Any delay in processing or any discrepancy in service provisioning may lead to an SLA violation. As per rules, any SLA violation would result in a penalty to be given by the service provider. These SLAs are for both private and public clouds Popular service providers are Amazon Web Services (AWS) and Microsoft Azure for public cloud. Similarly, OpenStack and Eucalyptus allow private cloud creation, deployment, and management.

1.4 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. Similarly, in a private cloud, it can be a data center, which is a huge collection of hardware resources interconnected to each other that is present in a specific location or a high configuration system. 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. Whenever a user accesses the cloud, it should be available to the users as quickly as possible and should be within the time that is defined by the SLAs. As mentioned, if there is any discrepancy in provisioning the resources or application, the service provider has to pay the penalty. Hence, the data center consists of a high-speed network connection and a highly efficient algorithm to transfer the data from the data center to the manager. There can be a number of data centers for a cloud, and similarly, a number of clouds can share a data center.

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.

2-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, whereas architecture wholly defines and describes the technology over which it is working. 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. Thus, anatomy can be considered as a part

of architecture. Figure 2 depicts the most standard anatomy that is the base for the cloud. There are basically five components of the cloud:

FIGURE 3.2

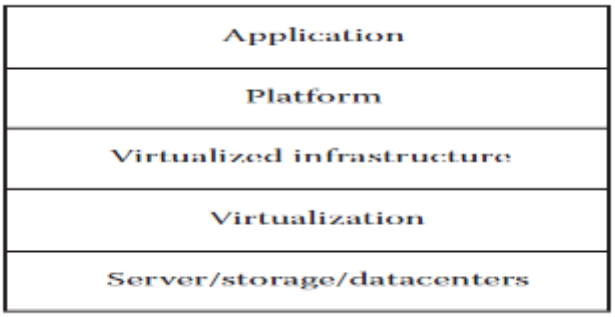

Fig. 2. Cloud structure.

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.

3-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.

3.1 Public Cloud Access Networking

In this option, the connectivity is often through the Internet, though some cloud providers may be able to support virtual private networks (VPNs) for customers. Accessing public cloud services will always create issues related to security, which in turn is related to performance.

3.2 Private Cloud Access Networking

In the private cloud deployment model, since the cloud is part of an organizational network, the technology and approaches are local to the in-house network structure. This may include an Internet VPN or VPN service from a network operator. If the application access was properly done with an organizational network—connectivity in a precloud configuration—transition to private cloud computing will not affect the access performance.

4-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.

The web applications were different from the stand-alone applications in many aspects. The main difference was the client server architecture that was followed by the web application. Unlike stand-alone applications, these systems were totally dependent on the network for its working. Here, there are basically two components, called as the client and the server. The server is a high-end machine that consists of the web application installed. This web application is accessed from other client systems. The client can reside anywhere in the network. It can access the web application through the Internet. This type of application was very useful, and this is extensively used from its inception and

now has become an important part of day-to-day life. 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 nonelastic nature, peak load transactions cannot be handled.

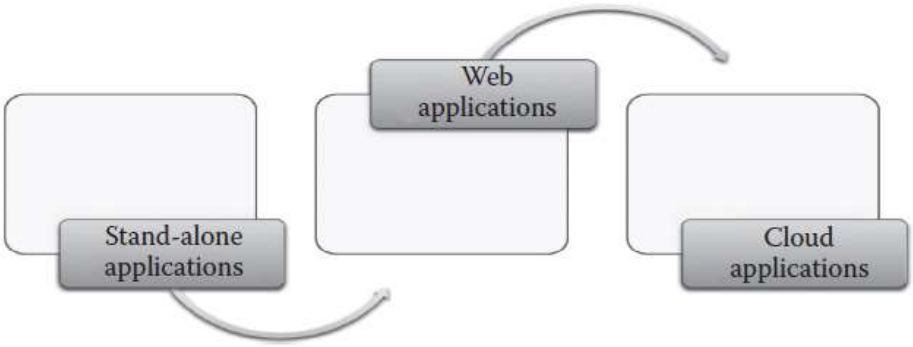

Fig.3. Computer application evolution.

Primarily to solve the previously mentioned problem, the cloud applications were developed. Figure 3 depicts the improvements in the applications.

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. A cloud application is different from other applications; they have unique features. A cloud application usually can be accessed as a web application but its properties differ. According to NIST [3], the features that make cloud applications unique are described in the following (Figure 4 depicts the features of a cloud 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.

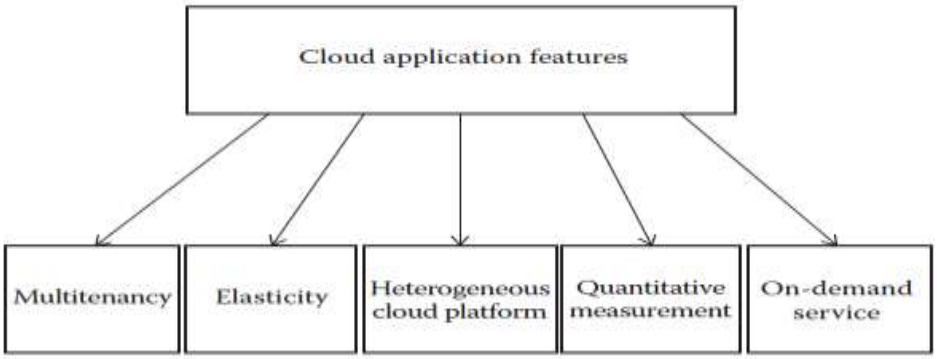

Fig. 4. Features of cloud.

Each user will have a separate application instance and the changes in one application would not affect the other. Physically, the software is shared and is not independent. The degree of physical isolation is very less. The logical independence is what is guaranteed. There are no restrictions in the number of applications being shared. The difficulty in providing logical isolation depends on the physical isolation to a certain extent. If an application is physically too close, then it becomes difficult to provide multitenancy. Web application and cloud application are similar as the users use the same way to access both. Figure 5 depicts a multitenant application where several users share the same application.

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 deprovisioning 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 to several hundreds of users at a time. In addition to this, it supports the rapid fluctuation of loads, that is, the increase or decrease in the number of users and their usage can rapidly change.

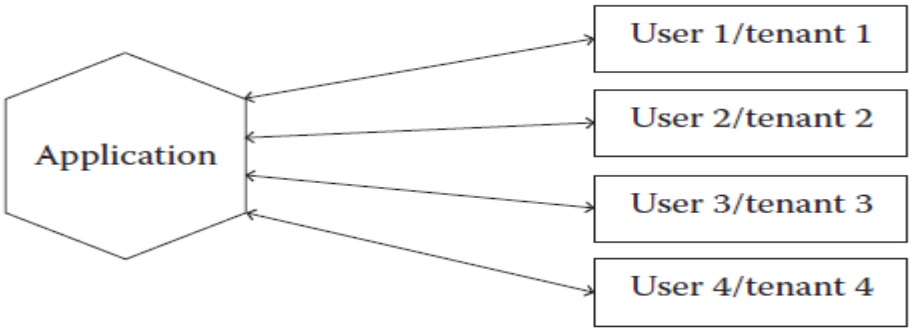

Fig. 5. Multitenancy.

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.

The previously mentioned properties are some of the features that make cloud a unique application platform. These properties mentioned are specific to the cloud hence making it as one of the few technologies that allows application developers to suffice the user's needs seamlessly without any disruption.

5-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: 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 would be adversely affected.

2. 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.

6-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.