

**CLOUD COMPUTING  
(PROFESSIONAL ELECTIVE – IV)**

**B.Tech. IV Year I Sem.**

**Course Code: CS742PE**

**L T P C**  
**3 0 0 3**

**Prerequisites:**

- A course on “Computer Networks”.
- A course on “Operating Systems”.
- A course on “Distributed Systems”.

**Course Objectives:**

- This course provides an insight into cloud computing
- Topics covered include- distributed system models, different cloud service models, service-oriented architectures, cloud programming and software environments, resource management.

**Course Outcomes:**

- Ability to understand various service delivery models of a cloud computing architecture.
- Ability to understand the ways in which the cloud can be programmed and deployed.
- Understanding cloud service providers.

**UNIT - I**

**Computing Paradigms:** High-Performance Computing, Parallel Computing, Distributed Computing, Cluster Computing, Grid Computing, Cloud Computing, Bio computing, Mobile Computing, Quantum Computing, Optical Computing, Nano computing.

**UNIT - II**

**Cloud Computing Fundamentals:** Motivation for Cloud Computing, The Need for Cloud Computing, Defining Cloud Computing, Definition of Cloud computing, Cloud Computing Is a Service, Cloud Computing Is a Platform, Principles of Cloud computing, Five Essential Characteristics, Four Cloud Deployment Models

**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.

**UNIT - IV**

**Cloud Service Models:** Infrastructure as a Service, Characteristics of IaaS. Suitability of IaaS, Pros and Cons of IaaS, Summary of IaaS Providers, Platform as a Service,

Characteristics of PaaS, Suitability of PaaS, Pros and Cons of PaaS, Summary of PaaS Providers, Software as a Service, Characteristics of SaaS, Suitability of SaaS, Pros and Cons of SaaS, Summary of SaaS Providers, Other Cloud Service Models.

## UNIT - V

**Cloud Service Providers:** EMC, EMC IT, Captiva Cloud Toolkit, Google, Cloud Platform, Cloud Storage, Google Cloud Connect, Google Cloud Print, Google App Engine, Amazon Web Services, Amazon Elastic Compute Cloud, Amazon Simple Storage Service, Amazon Simple Queue ,service, Microsoft, Windows Azure, Microsoft Assessment and Planning Toolkit, SharePoint, IBM, Cloud Models, IBM Smart Cloud, SAP Labs, SAP HANA Cloud Platform, Virtualization Services Provided by SAP, Sales force, Sales Cloud, Service Cloud: Knowledge as a Service, Rack space, VMware, Manjra soft, Aneka Platform

### TEXT BOOKS:

1. Essentials of cloud Computing: K. Chandrasekhran, CRC press, 2014

### REFERENCE BOOKS:

1. Cloud Computing: Principles and Paradigms by Rajkumar Buyya, James Broberg and Andrzej M. Goscinski, Wiley, 2011.
2. Distributed and Cloud Computing, Kai Hwang, Geoffery C. Fox, Jack J. Dongarra, Elsevier, 2012.
3. Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance, Tim Mather, Subra Kumaraswamy, Shahed Latif, O'Reilly, SPD, rp2011.

## UNIT-1

<b>1. Computing Paradigms.....</b>
Learning Objectives .....
Preamble.....
1.1   High-Performance Computing .....
1.2   Parallel Computing.....
1.3   Distributed Computing.....
1.4   Cluster Computing .....
1.5   Grid Computing .....
1.6   Cloud Computing .....
1.7   Biocomputing.....
1.8   Mobile Computing.....
1.9   Quantum Computing.....
1.10   Optical Computing .....
1.11   Nanocomputing .....

### **1.1 High-Performance Computing**

In high-performance computing systems, a pool of processors (processor machines or central processing units [CPUs]) connected (networked) with other resources like memory, storage, and input and output devices, and the deployed software is enabled to run in the entire system of connected components.

The processor machines can be of homogeneous or heterogeneous type. The legacy meaning of high-performance computing (HPC) is the supercomputers; however, it is not true in present-day computing scenarios. Therefore, HPC can also be attributed to mean the other computing paradigms that are discussed in the forthcoming sections, as it is a common name for all these computing systems.

Thus, examples of HPC include a small cluster of desktop computers or personal computers (PCs) to the fastest supercomputers. HPC systems are normally found in those applications where it is required to use or solve scientific problems. Most of the time, the challenge in working with these kinds of problems is to perform suitable simulation study, and this can be accomplished by HPC without any difficulty. Scientific examples such as protein folding in molecular biology and studies on developing models and applications based on nuclear fusion are worth noting as potential applications for HPC.

## 1.2 Parallel Computing

Parallel computing is also one of the facets of HPC. Here, a set of processors work cooperatively to solve a computational problem. These processor machines or CPUs are mostly of homogeneous type. Therefore, this definition is *the same* as that of HPC and is broad enough to include supercomputers that have hundreds or thousands of processors interconnected with other resources. One can distinguish between *conventional* (also known as serial or sequential or Von Neumann) computers and parallel computers in the way the applications are executed.

In serial or sequential computers, the following apply:

- It runs on a single computer/processor machine having a single CPU.
- A problem is broken down into a discrete series of instructions.
- Instructions are executed one after another.

In parallel computing, since there is simultaneous use of multiple processor machines, the following apply:

- It is run using multiple processors (multiple CPUs).
  - A problem is broken down into discrete parts that can be solved concurrently.
  - Each part is further broken down into a series of instructions.
    - Instructions from each part are executed simultaneously on different processors.
    - An overall control/coordination mechanism is employed.
- 

### 1.3 Distributed Computing

Distributed computing is also a computing system that consists of multiple computers or processor machines connected through a network, which can be homogeneous or heterogeneous, but run as a single system. The connectivity can be such that the CPUs in a distributed system can be physically close together and connected by a local network, or they can be geographically distant and connected by a wide area network. The heterogeneity in a distributed system supports any number of possible configurations in the processor machines, such as mainframes, PCs, workstations, and minicomputers. The goal of distributed computing is to make such a network work as a single computer.

Distributed computing systems are advantageous over centralized systems, because there is a support for the following characteristic features:

1. Scalability: It is the ability of the system to be easily expanded by adding more machines as needed, and vice versa, without affecting the existing setup.
2. Redundancy or replication: Here, several machines can provide the same services, so that even if one is unavailable (or failed), work does not stop because other similar computing supports will be available.

## 1.4 Cluster Computing

A cluster computing system consists of a set of the same or similar type of processor machines connected using a dedicated network infrastructure. All processor machines share resources such as a common home directory and have a software such as a message passing interface (MPI) implementation installed to allow programs to be run across all nodes simultaneously. This is also a kind of HPC category. The individual computers in a cluster can be referred to as *nodes*. The reason to realize a cluster as HPC is due to the fact that the individual nodes can work together to solve a problem larger than any computer can easily solve. And, the nodes need to communicate with one another in order to work cooperatively and meaningfully together to solve the problem in hand.

If we have processor machines of heterogeneous types in a cluster, this kind of clusters become a subtype and still mostly are in the experimental or research stage.

---

## 1.5 Grid Computing

The computing resources in most of the organizations are underutilized but are necessary for certain operations. The idea of grid computing is to make use of such nonutilized computing power by the needy organizations, and thereby the return on investment (ROI) on computing investments can be increased.

Thus, grid computing is a network of computing or processor machines managed with a kind of software such as middleware, in order to access and use the resources remotely. The managing activity of grid resources through the middleware is called *grid services*. Grid services provide access control, security, access to data including digital libraries and databases, and access to large-scale interactive and long-term storage facilities.

**TABLE 1.1**

## Electrical Power Grid and Grid Computing

Electrical Power Grid	Grid Computing
<i>Never worry</i> about where the electricity that we are using comes from; that is, whether it is from coal in Australia, from wind power in the United States, or from a nuclear plant in France, one can simply plug the electrical appliance into the wall-mounted socket and it will get the electrical power that we need to operate the appliance.	<i>Never worry</i> about where the computer power that we are using comes from; that is, whether it is from a supercomputer in Germany, a computer farm in India, or a laptop in New Zealand, one can simply plug in the computer and the Internet and it will get the application execution done.
<i>The infrastructure</i> that makes this possible is called <i>the power grid</i> . It links together many different kinds of power plants with our home, through transmission stations, power stations, transformers, power lines, and so forth.	<i>The infrastructure</i> that makes this possible is called <i>the computing grid</i> . It links together computing resources, such as PCs, workstations, servers, and storage elements, and provides the mechanism needed to access them via the Internet.
The power grid is <i>pervasive</i> : electricity is available essentially everywhere, and one can simply access it through a standard wall-mounted socket.	The grid is also <i>pervasive</i> in the sense that the remote computing resources would be accessible from different platforms, including laptops and mobile phones, and one can simply access the grid computing power through the web browser.
The power grid is a <i>utility</i> : we ask for electricity and we get it. We also pay for what we get.	The grid computing is also a <i>utility</i> : we ask for computing power or storage capacity and we get it. We also pay for what we get.

Grid computing is more popular due to the following reasons:

- Its ability to make use of unused computing power, and thus, it is a cost-effective solution (reducing investments, only recurring costs)
- As a way to solve problems in line with any HPC-based application
- Enables heterogeneous resources of computers to work cooperatively and collaboratively to solve a scientific problem

Researchers associate the term *grid* to the way electricity is distributed in municipal areas for the common man. In this context, the difference between electrical power grid and grid computing is worth noting (Table 1.1).

## 1.6 Cloud Computing

The computing trend moved toward cloud from the concept of grid computing, particularly when large computing resources are required to solve a single problem, using the ideas of computing power as a *utility* and other allied concepts. However, the potential difference between grid and cloud is that grid computing supports leveraging several computers in parallel to solve a particular application, while cloud computing supports leveraging multiple resources, including computing resources, to deliver a unified *service* to the end user.

In cloud computing, the IT and business resources, such as servers, storage, network, applications, and processes, can be dynamically provisioned to the user needs and workload. In addition, while a cloud can provision and support a grid, a cloud can also support nongrid environments, such as a three-tier web architecture running on traditional or Web 2.0 applications.

## 1.7 Biocomputing

Biocomputing systems use the concepts of biologically derived or simulated molecules (or models) that perform computational processes in order to solve a problem. The biologically derived models aid in structuring the computer programs that become part of the application.

Biocomputing provides the theoretical background and practical tools for scientists to explore proteins and DNA. DNA and proteins are nature's

building blocks, but these building blocks are not exactly used as *bricks*; the function of the final molecule rather strongly depends on the *order* of these blocks. Thus, the biocomputing scientist works on inventing the *order* suitable for various applications mimicking biology. Biocomputing shall, therefore, lead to a better understanding of life and the molecular causes of certain diseases.

---

## 1.8 Mobile Computing

In mobile computing, the processing (or computing) elements are small (i.e., handheld devices) and the communication between various resources is taking place using wireless media.

Mobile communication for voice applications (e.g., cellular phone) is widely established throughout the world and witnesses a very rapid growth in all its dimensions including the increase in the number of subscribers of various cellular networks. An extension of this technology is the ability to send and receive data across various cellular networks using small devices such as smartphones. There can be numerous applications based on this technology; for example, video call or conferencing is one of the important applications that people prefer to use in place of existing voice (only) communications on mobile phones.

Mobile computing-based applications are becoming very important and rapidly evolving with various technological advancements as it allows users to transmit data from remote locations to other remote or fixed locations.

## 1.9 Quantum Computing

Manufacturers of computing systems say that there is a limit for cramming more and more transistors into smaller and smaller spaces of integrated circuits (ICs) and thereby doubling the processing power about every 18 months. This problem will have to be overcome by a new *quantum computing*-based solution, wherein the dependence is on quantum information, the rules that govern the subatomic world. Quantum computers are millions of times faster than even our most powerful supercomputers today. Since quantum computing works differently on the most fundamental level than the current technology, and although there are working prototypes, these systems have not so far proved to be alternatives to today's silicon-based machines.

## 1.10 Optical Computing

Optical computing system uses the photons in visible light or infrared beams, rather than electric current, to perform digital computations. An electric current flows at only about 10% of the speed of light. This limits the rate at which data can be exchanged over long distances and is one of the factors that led to the evolution of optical fiber. By applying some of the advantages of visible and/or IR networks at the device and component scale, a computer can be developed that can perform operations 10 or more times faster than a conventional electronic computer.

## 1.11 Nanocomputing

Nanocomputing refers to computing systems that are constructed from nanoscale components. The silicon transistors in traditional computers may be replaced by transistors based on carbon nanotubes.

The successful realization of nanocomputers relates to the scale and integration of these nanotubes or components. The issues of scale relate to the dimensions of the components; they are, at most, a few nanometers in at least two dimensions. The issues of integration of the components are twofold: first, the manufacture of complex arbitrary patterns may be economically infeasible, and second, nanocomputers may include massive quantities of devices. Researchers are working on all these issues to bring nanocomputing a reality.

**UNIT-2**

<b>2. Cloud Computing Fundamentals.....</b>
Learning Objectives .....
Preamble.....
2.1 Motivation for Cloud Computing.....
2.1.1 The Need for Cloud Computing.....
2.2 Defining Cloud Computing.....
2.2.1 NIST Definition of Cloud Computing .....
2.2.2 Cloud Computing Is a Service .....
2.2.3 Cloud Computing Is a Platform.....
2.3 5-4-3 Principles of Cloud computing .....
2.3.1 Five Essential Characteristics.....
2.3.2 Four Cloud Deployment Models .....

## **2.1 Motivation for Cloud Computing**

Let us review the scenario of computing prior to the announcement and availability of cloud computing: The users who are in need of computing are expected to invest money on computing resources such as hardware, software, networking, and storage; this investment naturally costs a bulk currency to the users as they have to buy these computing resources, keep these in their premises, and maintain and make it operational—all these tasks would add cost. And, this is a particularly true and huge expenditure to the enterprises that require enormous computing power and resources, compared with classical academics and individuals.

On the other hand, it is easy and handy to get the required computing power and resources from some provider (or supplier) as and when it is needed and pay only for that usage. This would cost only a reasonable investment or spending, compared to the huge investment when buying the entire computing infrastructure. This phenomenon can be viewed as *capital expenditure* versus *operational expenditure*. As one can easily assess the huge lump sum required for capital expenditure (whole investment and maintenance for computing infrastructure) and compare it with the moderate or smaller lump sum required for the hiring or getting the computing infrastructure only to the tune of required time, and rest of the time free from that. Therefore, cloud

computing is a mechanism of *bringing–hiring or getting the services of the computing power or infrastructure* to an organizational or individual level to the extent required and paying only for the consumed services.

One can compare this situation with the usage of electricity (its services) from its producer-cum-distributor (in India, it is the state-/government-owned electricity boards that give electricity supply to all residences and organizations) to houses or organizations; here, we do not generate electricity (comparable with electricity production-related tasks); rather, we use it only to tune up our requirements in our premises, such as for our lighting and usage of other electrical appliances, and pay as per the electricity meter reading value.

Therefore, cloud computing is needed in getting the services of computing resources. Thus, one can say as a one-line answer to the need for cloud computing that it eliminates a large computing investment without compromising the use of computing at the user level at an operational cost. Cloud computing is very economical and saves a lot of money. A blind benefit of this computing is that even if we lose our laptop or due to some crisis our personal computer—and the desktop system—gets damaged, still our data and files will stay safe and secured as these are not in our local machine (but remotely located at the provider's place—machine).

In addition, one can think to add security while accessing these remote computing resources as depicted in Figure 2.1.

Figure 2.1 shows several cloud computing applications. The *cloud* represents the Internet-based computing resources, and the accessibility is through some

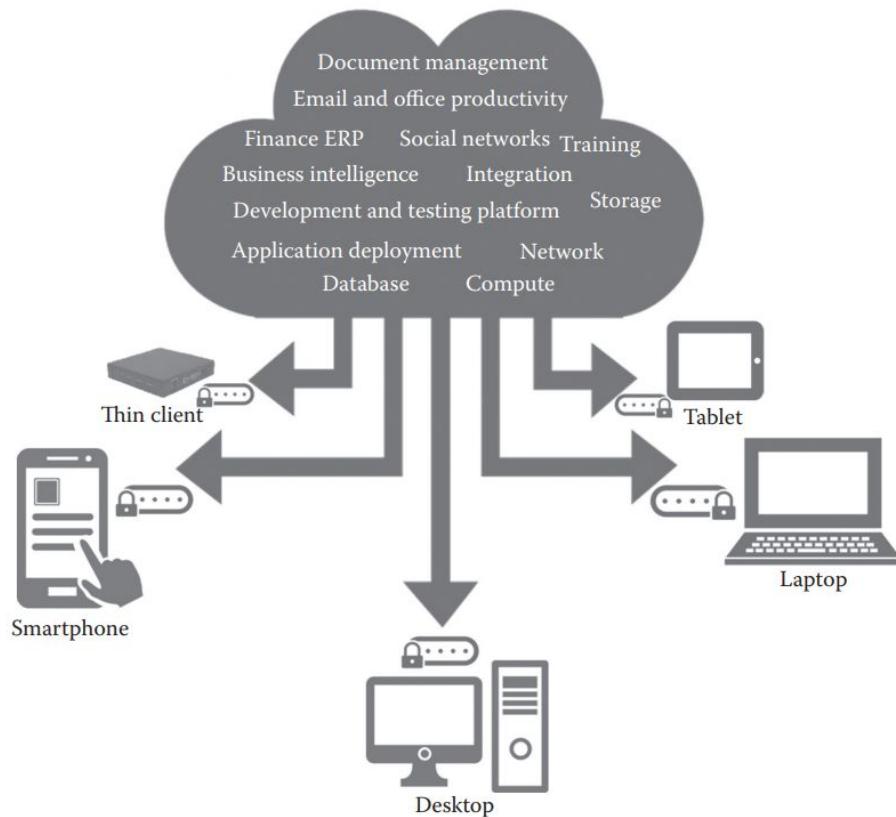


FIGURE 2.1

secure support of connectivity. It is a computing solution growing in popularity, especially among individuals and small- and medium-sized companies (SMEs). In the cloud computing model, an organization's core computer power resides offsite and is essentially subscribed to rather than owned.

Thus, cloud computing comes into focus and much needed only when we think about what computing resources and information technology (IT) solutions are required. This need caters to a way to increase capacity or add capabilities on the fly without investing in new infrastructure, training new personnel, or licensing new software. Cloud computing encompasses the subscription-based or pay-per-use service model of offering computing to end users or customers over the Internet and thereby extending the IT's existing capabilities.

### **2.1.1 The Need for Cloud Computing**

The main reasons for the need and use of cloud computing are convenience and reliability. In the past, if we wanted to bring a file, we would have to save it to a Universal Serial Bus (USB) flash drive, external hard drive, or compact disc (CD) and bring that device to a different place. Instead, saving a file to the cloud

(e.g., use of cloud application Dropbox) ensures that we will be able to access it with any computer that has an Internet connection. The cloud also makes it much easier to share a file with friends, making it possible to collaborate over the web.

While using the cloud, losing our data/file is much less likely. However, just like anything online, there is always a risk that someone may try to gain access to our personal data, and therefore, it is important to choose an access control with a strong password and pay attention to any privacy settings for the cloud service that we are using.

---

## 2.2 Defining Cloud Computing

In the simplest terms, cloud computing means storing and accessing data and programs over the Internet from a remote location or computer instead of our computer's hard drive. This so called *remote location* has several properties such as scalability, elasticity etc., which is significantly different from a simple remote machine. The cloud is just a metaphor for the Internet. When we store data on or run a program from the local computer's hard drive, that is called local storage and computing. For it to be considered *cloud computing*, we need to access our data or programs over the Internet. The end result is the same; however, with an online connection, cloud computing can be done anywhere, anytime, and by any device.

### 2.2.1 NIST Definition of Cloud Computing

The formal definition of cloud computing comes from the National Institute of Standards and Technology (NIST): "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. This cloud model is composed of five essential characteristics, three service models, and four deployment models [1].

It means that the computing resource or infrastructure—be it server hardware, storage, network, or application software—all available from the cloud vendor or provider's site/premises, can be accessible over the Internet from any remote location and by any local computing device. In addition, the usage or accessibility is to cost only to the level of usage to the customers based on their needs and demands, also known as the *pay-as-you-go* or *pay-as-per-use* model. If the need is more, more quantum computing resources are made available (provisioning with elasticity) by the provider. Minimal management effort implies that at the customer's side, the maintenance of computing systems is very minimal as they will have to look at these tasks only for their local computing devices used for accessing cloud-based resources, not for those computing resources managed at the provider's side. Details of five essential characteristics, three service models,

and four deployment models are provided in the 5-4-3 principles in Section 2.3. Many vendors, pundits, and experts refer to NIST, and both the International Standards Organization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE) back the NIST definition.

Now, let us try to define and understand cloud computing from two other perspectives—as a service and a platform—in the following sections.

### 2.2.2 Cloud Computing Is a Service

The simplest thing that any computer does is allow us to store and retrieve information. We can store our family photographs, our favorite songs, or even save movies on it, which is also the most basic service offered by cloud computing. Let us look at the example of a popular application called *Flickr* to illustrate the meaning of this section.

While Flickr started with an emphasis on sharing photos and images, it has emerged as a great place to store those images. In many ways, it is superior to storing the images on your computer:

1. First, Flickr allows us to easily access our images no matter where we are or what type of device we are using. While we might upload the photos of our vacation from our home computer, later, we can easily access them from our laptop at the office.
2. Second, Flickr lets us share the images. There is no need to burn them to a CD or save them on a flash drive. We can just send someone our Flickr address to share these photos or images.

3. Third, Flickr provides data security. By uploading the images to Flickr, we are providing ourselves with data security by creating a backup on the web. And, while it is always best to keep a local copy—either on a computer, a CD, or a flash drive—the truth is that we are far more likely to lose the images that we store locally than Flickr is of losing our images.

### 2.2.3 Cloud Computing Is a Platform

The World Wide Web (WWW) can be considered as the operating system for all our Internet-based applications. However, one has to understand that we will always need a local operating system in our computer to access web-based applications.

The basic meaning of the term *platform* is that it is the support on which applications run or give results to the users. For example, Microsoft Windows is a platform. But, a platform does not have to be an operating system. Java is a platform even though it is not an operating system.

Through cloud computing, the web is becoming a platform. With trends (applications) such as Office 2.0, more and more applications that were originally available on desktop computers are now being converted into

---

web-cloud applications. Word processors like Buzzword and office suites like Google Docs are now available in the cloud as their desktop counterparts. All these kinds of trends in providing applications via the cloud are turning cloud computing into a platform or to act as a platform.

---

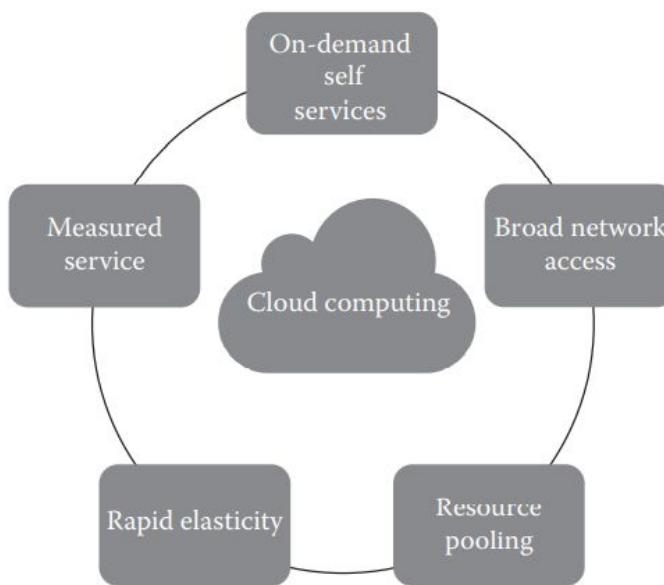
### 2.3 5-4-3 Principles of Cloud computing

The 5-4-3 principles put forth by NIST describe (a) the five essential characteristic features that promote cloud computing, (b) the four deployment models that are used to narrate the cloud computing opportunities for customers while looking at architectural models, and (c) the three important and basic service offering models of cloud computing.

#### 2.3.1 Five Essential Characteristics

Cloud computing has five essential characteristics, which are shown in Figure 2.2. Readers can note the word *essential*, which means that if any of these characteristics is missing, then it is not cloud computing:

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 each service's provider.
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 personal digital assistants [PDAs]).

**FIGURE 2.2**

The essential characteristics of cloud computing.

3. *Elastic resource pooling*: The provider's computing resources are pooled to serve multiple consumers using a multitenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify the location at a higher level of abstraction (e.g., country, state, or data center). Examples of resources include storage, processing, memory, and network bandwidth.
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.
5. *Measured service*: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported providing transparency for both the provider and consumer of the utilized service.

### 2.3.2 Four Cloud Deployment Models

Deployment models describe the ways with which the cloud services can be deployed or made available to its customers, depending on the organizational structure and the provisioning location. One can understand it in this manner too: cloud (Internet)-based computing resources—that is, the locations where data and services are acquired and provisioned to its customers—can take various forms. Four deployment models are usually distinguished, namely, public, private, community, and hybrid cloud service usage:

1. *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.
2. *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.
3. *Community cloud*: The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on premise or off premise.
4. *Hybrid cloud*: The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

**UNIT 3 Cloud Computing Architecture and Management**

<b>3. Cloud Computing Architecture and Management.....</b>
Learning Objectives.....
Preamble.....
3.1 Introduction .....
3.2 Cloud Architecture .....
3.2.1 Layer 1 (User/Client Layer) .....
3.2.2 Layer 2 (Network Layer) .....
3.2.3 Layer 3 (Cloud Management Layer) .....
3.2.4 Layer 4 (Hardware Resource Layer).....
3.3 Anatomy of the Cloud.....
3.4 Network Connectivity in Cloud Computing .....
3.4.1 Public Cloud Access Networking.....
3.4.2 Private Cloud Access Networking .....
3.4.3 Intracloud Networking for Public Cloud Services.....
3.4.4 Private Intracloud Networking.....
3.4.5 New Facets in Private Networks .....
3.4.6 Path for Internet Traffic .....
3.5 Applications on the Cloud .....
3.6 Managing the Cloud.....
3.6.1 Managing the Cloud Infrastructure .....
3.6.2 Managing the Cloud Application.....
3.7 Migrating Application to Cloud .....
3.7.1 Phases of Cloud Migration .....
3.7.2 Approaches for Cloud Migration .....

### 3.1 Introduction

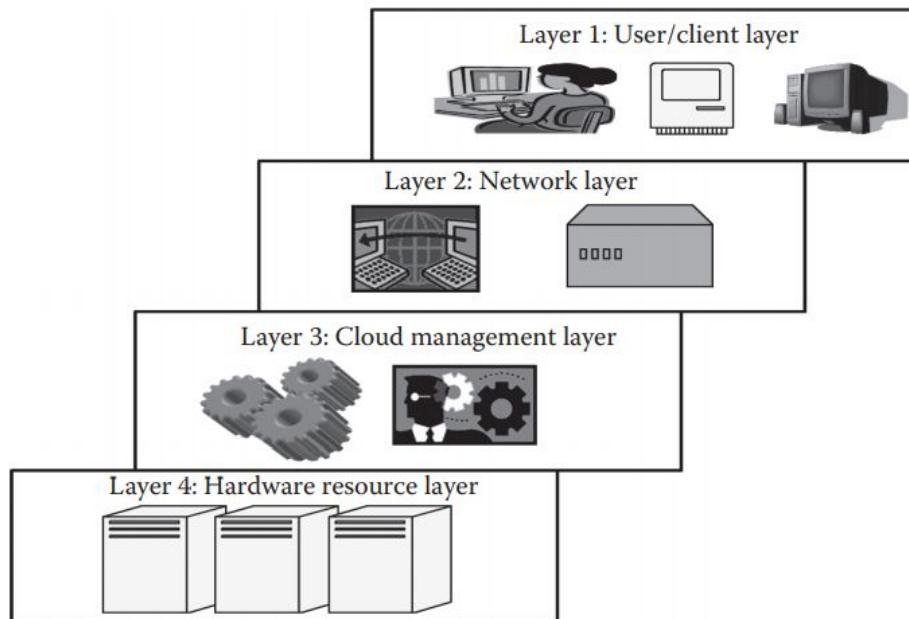
Cloud computing is similar to other technologies in a way that it also has several basic concepts that one should learn before knowing its core concepts. There are several processes and components of cloud computing that need to be discussed. One of the topics of such prime importance is architecture. Architecture is the hierarchical view of describing a technology. This usually includes the components over which the existing technology is built and the components that are dependent on the technology. Another topic that is related to architecture is anatomy. Anatomy describes the core structure of the cloud. Once the structure of the cloud is clear, the network connections in the cloud and the details about the cloud application need to be known. This is important as the cloud is a completely Internet-dependent technology. Similarly, cloud management discusses the important management issues and ways in which the current cloud scenario is managed. It describes the way an application and infrastructure in the cloud are managed. Management is important because of the quality of service (QoS) factors that are involved in the cloud. These QoS factors form the basis for cloud computing. All the services are given based on these QoS factors. Similarly, application migration to the cloud also plays a very important role. Not all applications can be directly deployed to the cloud. An application needs to be properly migrated to the cloud to be considered a proper cloud application that will have all the properties of the cloud.

## 3.2 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. Figure 3.1 depicts the architecture. The cloud architecture can be divided into four layers based on the access of the cloud by the user. They are as follows.

### 3.2.1 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



**FIGURE 3.1**  
Cloud architecture.

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.

### 3.2.2 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).

### 3.2.3 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.

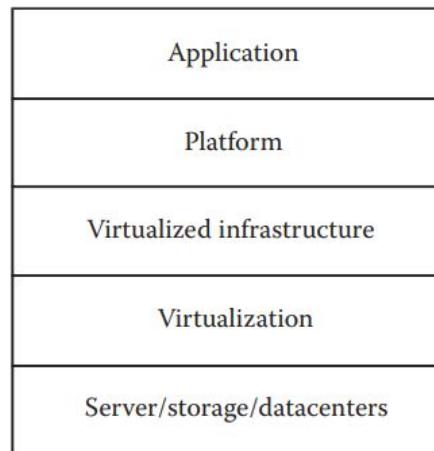
### 3.2.4 Layer 4 (Hardware Resource Layer)

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.

### 3.3 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,



**FIGURE 3.2**  
Cloud structure.

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. The basic structure of the cloud is described in Figure 3.2, which can be elaborated, and minute structural details can be given. Figure 3.2 depicts the most standard anatomy that is the base for the cloud. It depends on the person to choose the depth of description of the cloud. A different view of anatomy is given by Refs. [1,2].

There are basically five components of the cloud:

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.

### **3.4 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.

#### **3.4.1 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. One of the possible approaches toward the support of security is to promote connectivity through encrypted tunnels, so that the information may be sent via secure pipes on the Internet. This procedure will be an overhead in the connectivity, and using it will certainly increase delay and may impact performance.

If we want to reduce the delay without compromising security, then we have to select a suitable routing method such as the one reducing the delay by minimizing transit *hops* in the end-to-end connectivity between the cloud provider and cloud consumer. Since the end-to-end connectivity support is via the Internet, which is a complex federation of interconnected providers (known as Internet service providers [ISPs]), one has to look at the options of selecting the path.

### **3.4.2 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.

### **3.4.3 Intracloud Networking for Public Cloud Services**

Another network connectivity consideration in cloud computing is intra-cloud networking for public cloud services. Here, the resources of the

cloud provider and thus the cloud service to the customer are based on the resources that are geographically apart from each other but still connected via the Internet. Public cloud computing networks are internal to the service provider and thus not visible to the user/customer; however, the security aspects of connectivity and the access mechanisms of the resources are important. Another issue to look for is the QoS in the connected resources worldwide. Most of the performance issues and violations from these are addressed in the SLAs commercially.

#### 3.4.4 Private Intracloud Networking

The most complicated issue for networking and connectivity in cloud computing is private intracloud networking. What makes this particular issue so complex is that it depends on how much intracloud connectivity is associated with the applications being executed in this environment. Private intracloud networking is usually supported over connectivity between the major data center sites owned by the company. At a minimum, all cloud computing implementations will rely on intracloud networking to link users with the resource to which their application was assigned. Once the resource linkage is made, the extent to which intracloud networking is used depends on whether the application is componentized based on *service-oriented architecture (SOA)* or not, among multiple systems. If the principle of SOA is followed, then traffic may move between components of the application, as well as between the application and the user. The performance of those connections will then impact cloud computing performance overall. Here too, the impact of cloud computing performance is the differences that exist between

the current application and the network relationships with the application.

There are reasons to consider the networks and connectivity in cloud computing with newer approaches as globalization and changing network requirements, especially those related to increased Internet usage, are demanding more flexibility in the network architectures of today's enterprises. How are these related to us? The answers are discussed later.

### 3.4.5 New Facets in Private Networks

Conventional private networks have been architected for on-premise applications and maximum Internet security. Typically, applications such as e-mail, file sharing, and *enterprise resource planning* (ERP) systems are delivered to on-premise-based servers at each corporate data center. Increasingly today, software vendors are offering Software as a Service (SaaS) as an alternative for their software support to the corporate offices, which brings more challenges in the access and usage mechanisms of software from data center servers and in the connectivity of network architectures. The traditional network architecture for these global enterprises was not designed to optimize performance for cloud applications, now that many applications including

mission-critical applications are transitioning (moving) from on-premise based to cloud based, wherein the network availability becomes as mission critical as electricity: the business cannot function if it cannot access applications such as ERP and e-mail.

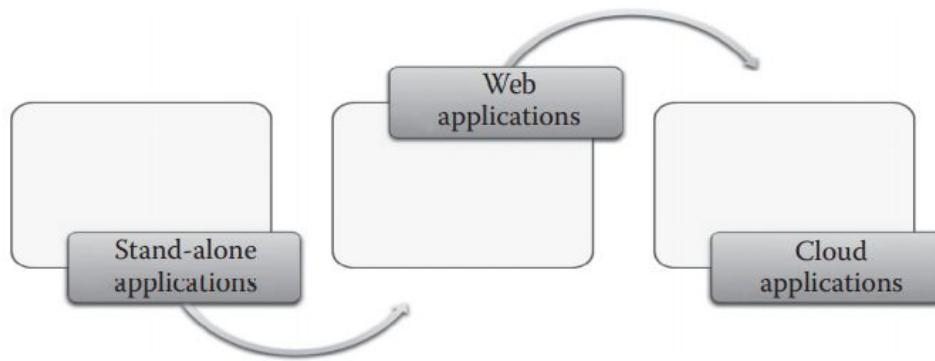
#### **3.4.6 Path for Internet Traffic**

The traditional Internet traffic through a limited set of Internet gateways poses performance and availability issues for end users who are using cloud-based applications. It can be improved if a more widely distributed Internet gateway infrastructure and connectivity are being supported for accessing applications, as they will provide lower-latency access to their cloud applications. As the volume of traffic to cloud applications grows, the percentage of the legacy network's capacity in terms of traffic to regional gateways increases. Applications such as video conferencing would hog more bandwidth while mission-critical applications such as ERP will consume less bandwidth, and hence, one has to plan a correct connectivity and path between providers and consumers.

### 3.5 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

**FIGURE 3.3**

Computer application evolution.

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.

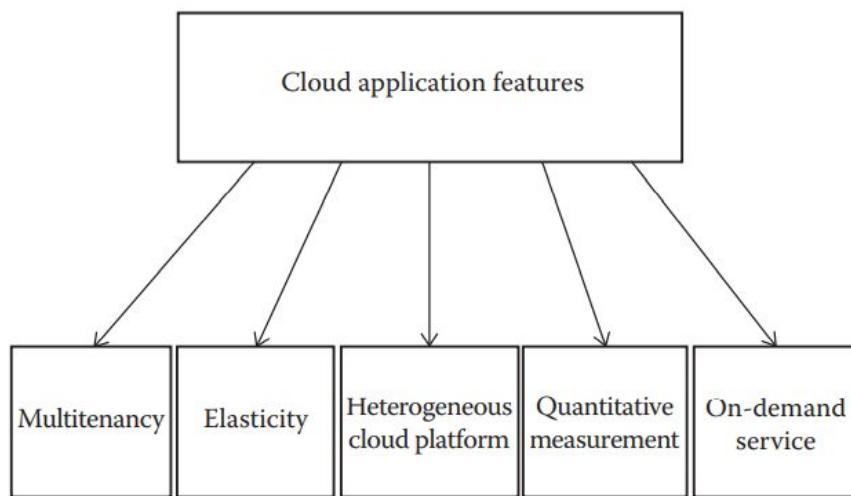
- The web applications are not provided on a pay-as-you-go basis; thus, a particular service is given to the user for permanent or trial use and usually the timings of user access cannot be monitored.
- Due to its nonelastic nature, peak load transactions cannot be handled.

Primarily to solve the previously mentioned problem, the cloud applications were developed. Figure 3.3 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 3.4 depicts the features of a cloud application):

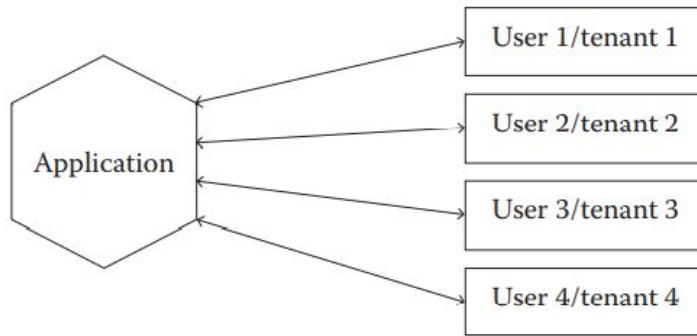
1. *Multitenancy*: Multitenancy is one of the important properties of cloud that make it different from other types of application in which the software can be shared by different users with full independence. Here, independence refers to logical independence.



**FIGURE 3.4**  
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 3.5 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



**FIGURE 3.5**  
Multitenancy.

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.

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.

### 3.6 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

#### 3.6.1 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 would

be adversely affected. 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. Similarly, the basic functionality of the cloud should always be provided and considered at any cost. Even if there is a small discrepancy in providing the functionality, the whole purpose of maintaining the cloud is futile. A partially functional cloud would not satisfy the SLAs.

Lastly, the reason for which the cloud was developed was cost. The cost is a very important criterion as far as the business prospects of the cloud are concerned. On the part of the service providers, if they incur less cost for managing the cloud, then they would try to reduce the cost so as to get a strong user base. Hence, a lot of users would use the services, improving their profit margin. Similarly, if the cost of resource management is high, then definitely the cost of accessing the resources would be high and there

is never a lossy business from any organization and so the service provider would not bear the cost and hence the users have to pay more. Similarly, this would prove costly for service providers as they have a high chance of losing a wide user base, leading to only a marginal growth in the industry. And, competing with its industry rivals would become a big issue. Hence, efficient management with less cost is required.

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.

The previously discussed prospects are mostly suitable for IaaS. Similarly, there are different management methods that are followed for different types of service delivery models. Each of the type has its own way of management. All the management methodologies are based on

load fluctuation. 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.

This is as far as cloud management is concerned. Cloud governance is another topic that is closely related to cloud management. Cloud governance is different from cloud management. Governance in general is a term in the corporate world that generally involves the process of creating value to an organization by creating strategic objectives that will lead to the growth of the company and would maintain a certain level of control over the company. Similar to that, here cloud organization is involved.

There are several aspects of cloud governance out of which SLAs are one of the important aspects. SLAs are the set of rules that are defined between the user and cloud service provider that decide upon the QoS factor. If SLAs are not followed, then the defaulter has to pay the penalty. The whole cloud is governed by keeping these SLAs in mind. Cloud governance is discussed in detail in further chapters.

### 3.6.2 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. But, this shift or moving the applications to the cloud environment brings new complexities. Applications become more composite and complex, which requires leveraging not only capabilities like storage and database offered by the cloud providers but also third-party SaaS capabilities like e-mail and messaging. 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. These cloud-based monitoring and management services can collect a multitude of events, analyze them, and identify critical information that requires additional remedial actions like adjusting capacity or provisioning new services. Additionally,

application management has to be supported with tools and processes required for managing other environments that might coexist, enabling efficient operations.

---

### 3.7 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.

#### 3.7.1 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 hotplug 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:* Premigration optimizations identified are implemented. Cloud servers are provisioned for all the identified environments, necessary platform softwares 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:* Postmigration 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.

### 3.7.2 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.