Essentials of CLOUD COMPUTING

K. CHANDRASEKARAN

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Essentials of CLOUD COMPUTING

Essentials of CLOUD COMPUTING

K. Chandrasekaran

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2015 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20141014

International Standard Book Number-13: 978-1-4822-0544-2 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Foreword

Cloud computing is sprawling the IT landscape. Driven by several converging and complementary factors, cloud computing is advancing as a viable IT service delivery model at an incredible pace. It has caused a paradigm shift in how we deliver, use, and harness the variety of IT services it offers. It also offers several benefits compared to traditional on-premise computing models, including reduced costs and increased agility and flexibility. Its transformational potential is huge and impressive, and consequently cloud computing is being adopted by individual users, businesses, educational institutions, governments, and community organizations. It helps close the digital (information) divide. It might even help save our planet by providing an overall greener computing environment.

Hence, corporations are eagerly investing in promising cloud computing technologies and services not only in developed economies but also increasingly in emerging economies—including India, China, Taiwan, the Philippines, and South Africa—to address a region's specific needs. Cloud computing is receiving considerable interest among several stakeholders—businesses, the IT industry, application developers, IT administrators and managers, researchers, and students who aspire to be successful IT professionals.

To successfully embrace this new computing paradigm, however, they need to acquire new cloud computing knowledge and skills. In answer to this, universities have begun to offer new courses on cloud computing. Though there are several books on cloud computing—from basic books intended for general readers to advanced compendium for researchers—there are few books that comprehensively cover a range of cloud computing topics and are particularly intended as an entry-level textbook for university students. This book, *Essentials of Cloud Computing,* fills this void and is a timely and valuable addition by Professor K. Chandrasekaran, a well-recognized academic and researcher in cloud computing.

The book, beginning with a brief overview on different computing paradigms and potentials of those paradigms, outlines the fundamentals of cloud computing. Then, it deals with cloud services types, cloud deployment models, technologies supporting and driving the cloud, software process models and programming models for cloud, and development of software application that runs the cloud. It also gives an overview of services available from major cloud providers, highlights currently available open source software and tools for cloud deployment, and discusses security concerns and issues in cloud computing. Finally, it outlines advances in cloud computing such as mobile cloud and green cloud. The book's presentation style supports ease of reading and comprehension. Further, each chapter is supplemented with review questions that help the readers to check their understanding of topics and issues explored in the chapter.

Cloud computing is here to stay, and its adoption will be widespread. It will transform not only the IT industry but also every sector of society. A wide range of people—application developers, enterprise IT architects and administrators, and future IT professionals and managers—will need to learn about cloud computing and how it can be deployed for a variety of applications. This concise and comprehensive book will help readers understand several key aspects of cloud computing—technologies, models, cloud services currently available, applications that are better suited for cloud, and more. It will also help them examine the issues and challenges and develop and deploy applications in clouds.

I believe you will find the book informative, concise, comprehensive, and helpful to gain cloud knowledge.

San Murugesan
Adjunct Professor
University of Western Sydney
Richmond, New South Wales, Australia
Editor in Chief
IEEE IT Professional
Editor
IEEE Computer
Director
BRITE Professional Services
Australia

Preface

Cloud computing is one of the most popular technologies that has become an integral part of the computing world nowadays. The usage and popularity of cloud computing is increasing every day and is expected to increase further. Many frequent Internet users are heavily dependent on cloud-based applications for their day-to-day activities in both professional and personal life. Cloud computing has emerged as a technology to realize the utility model of computing while using Internet for accessing applications.

The past decades have witnessed the success of centralized computing infrastructures in many application domains. Then, the emergence of the Internet brought numerous users of remote applications based on the technologies of distributed computing. Research in distributed computing gave birth to the development of grid computing. Though grid is based on distributed computing, the conceptual basis for grid is somewhat different. Computing with grid enabled researchers to do computationally intensive tasks by using limited infrastructure that was available with them and with the support of high processing power that could be provided by any third party, and thus allowing the researchers to use grid computing, which was one of the first attempts to provide computing resources to users on payment basis. This technology indeed became popular and is being used even now. An associated problem with grid technology was that it could only be used by a certain group of people and it was not open to the public. Cloud computing in simple terms is further extension and variation of grid computing, in which a market-oriented aspect is added. Though there are several other important technical differences, this is one of the major differences between grid and cloud. Thus came cloud computing, which is now being used as a public utility computing software and is accessible by almost every person through the Internet. Apart from this, there are several other properties that make cloud popular and unique. In cloud, the resources are metered, and a user pays according to the usage. Cloud can also support a continuously varying user demands without affecting the performance, and it is always available for use without any restrictions. The users can access cloud from any device, thus reaching a wider range of people.

There are several applications of cloud computing already being witnessed and experienced. As cloud is elastic, it can be used in places where varying load is one of the main characteristics. It can also be used where on-demand access is required. Similarly, because of its property of multitenancy, it can be used in places where several applications are to be operated. Cloud computing can also be used for data-intensive applications for data analytics and several data-related tasks.

As this is considered a promising technology, several companies such as Google, Microsoft, Amazon, HP, and IBM have invested their time and other resources for further development of cloud computing–related technologies. In return, the companies make profit as cloud applications become more popular and easier to use.

The main objective of this book is to present the readers with the introductory details, technologies, applications development, security, and some advanced topics in cloud computing. It is expected that the book will serve as a reference for a larger audience base, including students of undergraduate and postgraduate programs, practitioners, developers, and new researchers.

This book will be a timely contribution to cloud computing, a field that is gaining momentum in all dimensions such as academia, research, and business. As cloud computing is recognized as one of the top five emerging technologies that will have a major impact on the quality of science and society over the next 20 years, its knowledge will help position our readers at the forefront of the field.

This book discusses in detail the essentials of cloud computing in a way suitable for beginners as well as practitioners who are in need to know or learn more about cloud computing. It can also be used as a handbook for cloud. It contains 14 chapters that follow a standard format for easy and useful reading: Learning Objectives, Preamble, Introduction, and details related to the chapter title with several subsections, supported by a suitable number of diagrams, tables, figures, etc., followed by Summary, Review Points, Review Questions, and References for further reading.

To start with, Chapter 1 aims to give a brief description about available paradigms of computing. This provides the required basic knowledge about computing paradigms to start with cloud technology. It includes several computing paradigms such as high-performance computing, cluster computing, grid computing, and distributed computing.

Chapter 2 gives a basic introduction and discusses the fundamental concepts of cloud. The topics include cloud computing definition, the need for cloud, cloud principles, cloud applications, and several other topics.

Chapter 3 gives an introduction to cloud computing technologies. This includes the basic concepts in cloud such as cloud architecture, cloud anatomy, network connectivity in cloud, cloud management, applications in cloud, and migrating applications to cloud.

Chapter 4 discusses in detail the deployment models such as private, public, community, and hybrid. Their applications, use, and design are also discussed, thereby giving a clear picture and facilitating a proper choice of deployment models.

Chapter 5 discusses in detail the cloud service models such as Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) with several other service models that have emerged recently.

This chapter gives an idea on the properties, architecture, and application of these cloud service models.

Chapter 6 discusses the technological drivers for cloud. The topics covered in these chapters are service-oriented architecture and cloud, virtualization, multicore technology, software models for cloud, pervasive computing, and several other related concepts. This chapter gives an elaborate view on how these technological drivers are related to cloud and promote it further in the context of application development and research.

Chapter 7 gives a detailed description about virtualization. Virtualization is considered to be the basis of cloud computing. Here, opportunities and approaches related to virtualization are discussed. Hypervisors are discussed in detail. This chapter also gives a description on how virtualization is used in cloud computing.

Chapter 8 discusses the programming models that are available for cloud. Here, both existing programming models useful to migrate to cloud and new programming models specific to cloud are discussed in detail.

Chapter 9 describes cloud from a software development perspective, the different perspectives of SaaS development and its challenges, and cloud-aware software development in PaaS.

Chapter 10 deals with the networking aspects in the cloud computing environment. This chapter also presents an overview and issues related to the data center environment. Transport layer issues and Transmission Control Protocol enhancements in data center networks are also discussed.

Chapter 11 gives a brief description of major service providers known in the cloud arena and discusses in detail about the services they offer.

Chapter 12 is especially for open-source users. This chapter gives a list and description of several open-source support and tools available for cloud computing. These are divided according to the service models, that is, SaaS, PaaS, and IaaS. There is also a separate note on open-source tools for research, which describes the tools that can be worked upon in from a research-oriented perspective. It also has an additional note on distributed computing tools that are used for managing distributed systems.

Chapter 13 discusses the security issues in cloud, an important issue in cloud. It discusses about security aspects in general, platform-related security, audit, and compliance in cloud.

The final chapter, Chapter 14, discusses advanced concepts in cloud, such as intercloud, cloud management, mobile cloud, media cloud, cloud governance, green cloud, cloud analytics, and several other allied topics.

The contents of this book reflect the author's lectures on this topic. The author wishes to acknowledge the following for their valuable time and contributions in developing, improving, and formatting the chapters: Mohit P. Tahiliani, Marimuthu C., Raghavan S., Manoj V. Thomas, Rohit P. Tahiliani, Alaka A., Usha D., Anithakumari S., and Christina Terese Joseph.

Finally, the author would like to thank Aastha Sharma, commissioning editor at CRC Press, for her constant communication and follow-up and support throughout the process of getting the book into print.

Readers are requested to visit the website http://www.cloudrose.org/ for further updates and e-mail interactions with the author.

K. Chandrasekaran

1

Computing Paradigms

Learning Objectives

The objectives of this chapter are to

- Give a brief description of major of computing
- Examine at the potential of these paradigms

Preamble

The term paradigm conveys that there is a set of practices to be followed to accomplish a task. In the domain of computing, there are many different standard practices being followed based on inventions and technological advancements. In this chapter, we look into the various computing paradigms: namely high performance computing, cluster computing, grid computing, cloud computing, bio-computing, mobile computing, quantum computing, optical computing, nanocomputing, and network computing. As computing systems become faster and more capable, it is required to note the features of modern computing in order to relate ourselves to the title of this book on cloud computing, and therefore it becomes essential to know little on various computing paradigms.

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
Never worry 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.	*Never worry* 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.
The infrastructure that makes this possible is called *the power grid*. It links together many different kinds of power plants with our home, through transmission stations, power stations, transformers, power lines, and so forth.	*The infrastructure* that makes this possible is called *the computing grid*. 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 *pervasive*: electricity is available essentially everywhere, and one can simply access it through a standard wall-mounted socket.	The grid is also *pervasive* 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 *utility*: we ask for electricity and we get it. We also pay for what we get.	The grid computing is also a *utility*: 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.

We will be looking at the details of cloud computing in different chapters of this book.

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.

1.12 Network Computing

Network computing is a way of designing systems to take advantage of the latest technology and maximize its positive impact on business solutions and their ability to serve their customers using a strong underlying network of computing resources. In any network computing solution, the client component of a networked architecture or application will be with the customer or client or end user, and in modern days, they provide an essential set of functionality necessary to support the appropriate client functions at minimum cost and maximum simplicity. Unlike conventional PCs, they do not need to be individually configured and maintained according to their intended use. The other end of the client component in the network architecture will be a typical *server* environment to *push* the services of the application to the client end.

Almost all the computing paradigms that were discussed earlier are of this nature. Even in the future, if any one invents a totally new computing paradigm, it would be based on a networked architecture, without which it is impossible to realize the benefits for any end user.

1.13 Summary

We are into a post-PC era, in which a greater number and a variety of computers and computing paradigms with different sizes and functions might be used everywhere and with every human being; so, the purpose of this chapter is to illustrate briefly the ideas of all these computing domains, as most of these are ubiquitous and pervasive in its access and working environment.

Key Points

- *Mobile computing*: Mobile computing consists of small processing elements (i.e., handheld devices) and the communication between various resources is by using wireless media (see Section 1.8).
- *Nanocomputing*: Makes use of nanoscale components (see Section 1.11).

Review Questions

1. Why is it necessary to understand the various computing paradigms?
2. Compare grid computing with electric power grid
3. Will mobile computing play a dominant role in the future? Discuss
4. How are distributed computing and network computing different or similar?
5. How may nanocomputing shape future devices?

Further Reading

Ditto, W. L., A. Miliotis, K. Murali, and S. Sinha. The chaos computing paradigm. *Reviews of Nonlinear Dynamics and Complexity* 3: 1–35, 2010.

2

Cloud Computing Fundamentals

Learning Objectives

The objectives of this chapter are to

- Understand the basic ideas and motivation for cloud computing
- To define cloud computing
- Understand the 5-4-3 principles of cloud computing and cloud ecosystem
- Understand the working of a cloud application
- Have a brief understanding on the benefits and drawbacks in cloud computing

Preamble

Modern computing with our laptop or desktop or even with tablets/smartphones using the Internet to access the data and details that we want, which are located/stored at remote places/computers, through the faces of applications like Facebook, e-mail, and YouTube, brings the actual power of information that we need instantaneously within no time. Even if millions of users get connected in this manner, from anywhere in the world, these applications do serve what these users–customers want. This phenomenon of supply of information or any other data and details to all the needy customers, as and when it is asked, is the conceptual understanding and working of what is known as cloud computing. This chapter is devoted to give basic understanding on cloud computing.

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

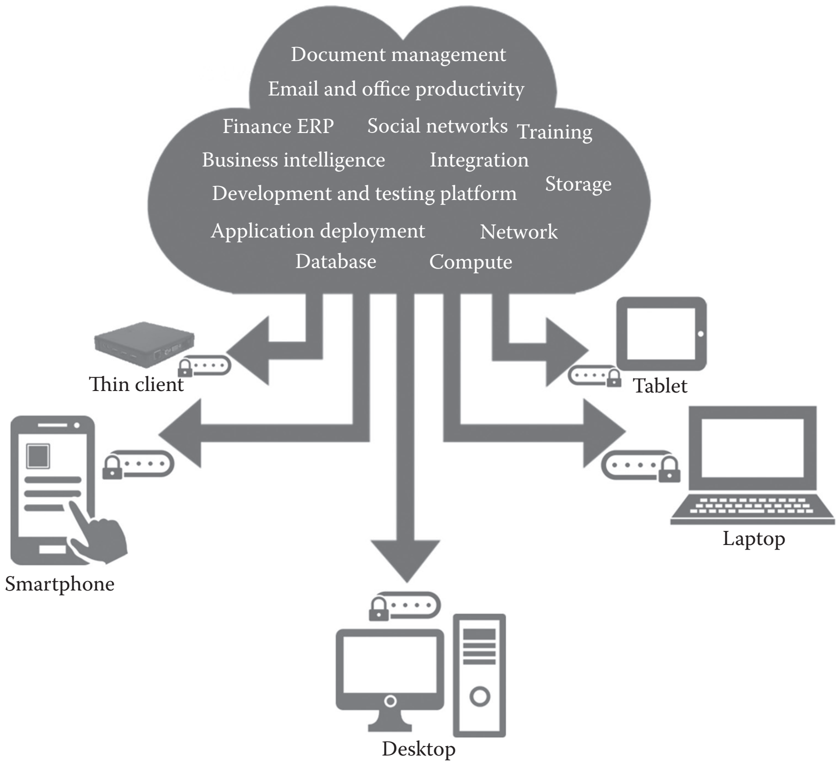

FIGURE 2.1
Cloud computing.

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]).

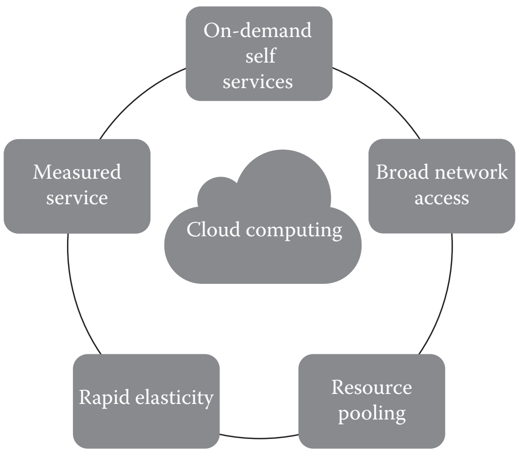

FIGURE 2.2
The essential characteristics of cloud computing.

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).

2.3.3 Three Service Offering Models

The three kinds of services with which the cloud-based computing resources are available to end customers are as follows: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). It is also known as the service–platform–infrastructure (SPI) model of the cloud and is shown in Figure 2.3. SaaS is a software distribution model in which applications (software, which is one of the most important computing resources) are hosted by a vendor or service provider and made available to customers over a network, typically the Internet. PaaS is a paradigm for delivering operating systems and associated services (e.g., computer aided software engineering [CASE] tools, integrated development environments [IDEs] for developing software solutions) over the Internet without downloads or installation. IaaS involves outsourcing the equipment used to support operations, including storage, hardware, servers, and networking components.

1. *Cloud SaaS*: The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure, including network, servers, operating systems, storage, and even individual application capabilities, with the possible exception of limited user-specific application configuration settings. The applications are accessible from various client devices through either a thin client

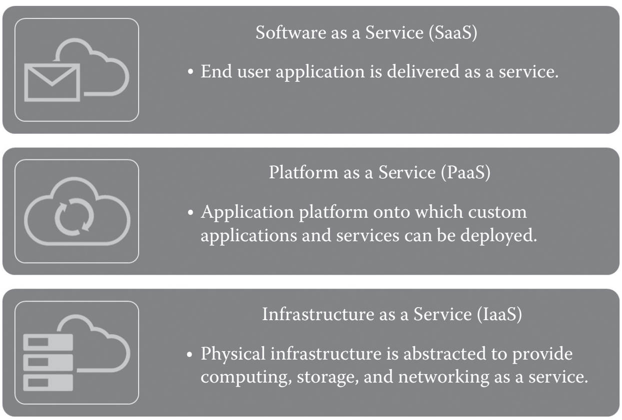

FIGURE 2.3
SPI—service offering model of the cloud.