The last decades have reinforced the idea that information processing can be done more efficiently

centrally, on large farms of computing and storage systems accessible via the Internet. When computing

resources in distant data centers are used rather than local computing systems, we talk about

network-centric computing and network-centric content. Advancements in networking and other areas

are responsible for the acceptance of the two new computing models and led to the grid computing

movement in the early 1990s and, since 2005, to utility computing and cloud computing.

In utility computing the hardware and software resources are concentrated in large data centers and

users can pay as they consume computing, storage, and communication resources. Utility computing

often requires a cloud-like infrastructure, but its focus is on the business model for providing the

computing services. Cloud computing is a path to utility computing embraced by major IT companies

such as Amazon, Apple, Google, HP, IBM, Microsoft, Oracle, and others.

Cloud computing delivery models, deployment models, defining attributes, resources, and organization

of the infrastructure discussed in this chapter are summarized in Figure 1.1. There are three cloud

delivery models: Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-

Service (IaaS), deployed as public, private, community, and hybrid clouds.

The defining attributes of the new philosophy for delivering computing services are as follows:

• Cloud computing uses Internet technologies to offer elastic services. The term elastic computing

refers to the ability to dynamically acquire computing resources and support a variable workload. A

cloud service provider maintains a massive infrastructure to support elastic services.

• The resources used for these services can be metered and the users can be charged only for the

resources they use.

• Maintenance and security are ensured by service providers.

• Economy of scale allows service providers to operate more efficiently due to specialization and

centralization.

• Cloud computing is cost-effective due to resource multiplexing; lower costs for the service provider

are passed on to the cloud users.

• The application data is stored closer to the site where it is used in a device- and location-independent

manner; potentially, this data storage strategy increases reliability and security and, at the same time,

it lowers communication costs.

Cloud computing is a technical and social reality and an emerging technology. At this time, one

can only speculate how the infrastructure for this new paradigm will evolve and what applications will

migrate to it. The economical, social, ethical, and legal implications of this shift in technology, in which

users rely on services provided by large data centers and store private data and software on systems

they do not control, are likely to be significant.

Scientific and engineering applications, data mining, computational financing, gaming, and social

networking as well as many other computational and data-intensive activities can benefit from cloud

computing. A broad range of data, from the results of high-energy physics experiments to financial or

enterprise management data to personal data such as photos, videos, and movies, can be stored on the

cloud.

In early 2011 Apple announced the iCloud, a network-centric alternative for storing content such

as music, videos, movies, and personal information; this content was previously confined to personal

devices such as workstations, laptops, tablets, or smartphones. The obvious advantage of networkcentric

content is the accessibility of information from any site where users can connect to the Internet.

Clearly, information stored on a cloud can be shared easily, but this approach raises major concerns: Is

the information safe and secure? Is it accessible when we need it? Do we still own it?

In the next few years, the focus of cloud computing is expected to shift from building the infrastructure,

today’s main front of competition among the vendors, to the application domain. This shift in

focus is reflected by Google’s strategy to build a dedicated cloud for government organizations in the

United States. The company states: “We recognize that government agencies have unique regulatory

and compliance requirements for IT systems, and cloud computing is no exception. So we’ve invested

a lot of time in understanding government’s needs and how they relate to cloud computing.”

In a discussion of technology trends, noted computer scientist Jim Gray emphasized that in 2003 the

cost of communication in a wide area network had decreased dramatically and will continue to do so.

Thus, it makes economical sense to store the data near the application [144] – in other words, to store

it in the cloud where the application runs. This insight leads us to believe that several new classes of

cloud computing applications could emerge in the next few years [25].

As always, a good idea has generated a high level of excitement that translated into a flurry of publications

– some of a scholarly depth, others with little merit or even bursting with misinformation. In this

book we attempt to sift through the large volume of information and dissect the main ideas related to

cloud computing. We first discuss applications of cloud computing and then analyze the infrastructure

for the technology.

Several decades of research in parallel and distributed computing have paved the way for cloud

computing. Through the years we have discovered the challenges posed by the implementation, as well

as the algorithmic level, and the ways to address some of them and avoid the others. Thus, it is important

to look back at the lessons we learned from this experience through the years; for this reason we start

our discussion with an overview of parallel computing and distributed systems.