Once the technological elements were in place, it was only a matter of time until the economical

advantages of cloud computing became apparent. Due to the economy of scale, large data centers –

centers with more than 50,000 systems – are more economical to operate than medium-sized centers

that have around 1,000 systems. Large data centers equipped with commodity computers experience

a five to seven times decrease of resource consumption, including energy, compared to medium-sized

centers [25]. The networking costs, in dollars per Mbit/s/month, are 95/13 = 7.1 times larger, and the

storage costs, in dollars per Gbyte/month, are 2.2/0.4 = 5.7 times larger for medium-sized centers.

Medium-sized centers have a larger administrative overhead – one system administrator for 140 systems

versus one for 1,000 systems for large centers.

Data centers are very large consumers of electric energy to keep servers and the networking infrastructure

running and for cooling. For example, there are 6,000 data centers in the United States and in

2006 they reportedly consumed 61 billion KWh, 1.5% of all electric energy in the U.S., at a cost of

$4.5 billion. The power demanded by data centers was predicted to double from 2006 to 2011. Peak

instantaneous demand was predicted to increase from 7 GW in 2006 to 12 GW in 2011, requiring the

construction of 10 new power plants. In the United States the energy costs differ from state to state;

for example 1 KWh costs 3.6 cents in Idaho, 10 cents in California, and 18 cents in Hawaii. Thus, data

centers should be placed at sites with low energy cost.

The term computer cloud is overloaded, since it covers infrastructures of different sizes,with different

management and different user populations. Several types of cloud are envisioned:

• Private cloud. The infrastructure is operated solely for an organization. It may be managed by the

organization or a third party and may exist on or off the premises of the organization.

• Community cloud. The 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 premises

or off premises.

• Public cloud. The infrastructure is made available to the general public or a large industry group

and is owned by an organization selling cloud services.

• Hybrid cloud. The infrastructure is a composition of two or more clouds (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).

A private cloud could provide the computing resources needed for a large organization, such as a

research institution, a university, or a corporation. The argument that a private cloud does not support

utility computing is based on the observation that an organization has to invest in the infrastructure and

a user of a private cloud pays as it consumes resources [25]. Nevertheless, a private cloud could use the same hardware infrastructure as a public one; its security requirements will be different from those for

a public cloud and the software running on the cloud is likely to be restricted to a specific domain.

A natural question to ask is: Why could cloud computing be successful when other paradigms

have failed? The reasons that cloud computing could be successful can be grouped into several general

categories: technological advances, a realistic system model, user convenience, and financial advantages.

A nonexhaustive list of reasons for the success of cloud computing includes these points:

• Cloud computing is in a better position to exploit recent advances in software, networking, storage,

and processor technologies. Cloud computing is promoted by large IT companies where these new

technological developments take place, and these companies have a vested interest in promoting the

new technologies.

• A cloud consists of a homogeneous set of hardware and software resources in a single administrative

domain. In this setup, security, resource management, fault tolerance, and quality of service are less

challenging than in a heterogeneous environment with resources in multiple administrative domains.

• Cloud computing is focused on enterprise computing; its adoption by industrial organizations, financial

institutions, healthcare organizations, and so on has a potentially huge impact on the economy.

• Acloud provides the illusion of infinite computing resources; its elasticity frees application designers

from the confinement of a single system.

• A cloud eliminates the need for up-front financial commitment, and it is based on a pay-as-you-go

approach. This has the potential to attract new applications and new users for existing applications,

fomenting a new era of industrywide technological advancements.

In spite of the technological breakthroughs that have made cloud computing feasible, there are still

major obstacles for this new technology; these obstacles provide opportunity for research.We list a few

of the most obvious obstacles:

• Availability of service. What happens when the service provider cannot deliver? Can a large company

such as General Motors move its IT to the cloud and have assurances that its activity will not be

negatively affected by cloud overload? A partial answer to this question is provided by service-level

agreements (SLAs).6 A temporary fix with negative economical implications is overprovisioning,

that is, having enough resources to satisfy the largest projected demand.

• Vendor lock-in. Once a customer is hooked to one provider, it is hard to move to another. The

standardization efforts at National Institute of Standards and Technology (NIST) attempt to address

this problem.

• Data confidentiality and auditability. This is indeed a serious problem; we analyze it in Chapter 9.

• Data transfer bottlenecks. Many applications are data-intensive. A very important strategy is to store

the data as close as possible to the site where it is needed. Transferring 1 TB of data on a 1Mbps

network takes 8 million seconds, or about 10 days; it is faster and cheaper to use courier service

and send data recoded on some media than to send it over the network. Very high-speed networks

will alleviate this problem in the future; for example, a 1 Gbps network would reduce this time to

8,000 s, or slightly more than 2 h.

• Performance unpredictability. This is one of the consequences of resource sharing. Strategies for

performance isolation are discussed in Section 5.5.

• Elasticity, the ability to scale up and down quickly. Newalgorithms for controlling resource allocation

and workload placement are necessary. Autonomic computing based on self-organization and selfmanagement

seems to be a promising avenue.

There are other perennial problems with no clear solutions at this time, including software licensing

and system bugs.