The efforts to support large-scale distributed computing have encountered major difficulties over the

years. The users of these systems discovered how difficult it was to locate the systems able to run an

application. They soon realized that it is equally difficult to scale up and down to accommodate a dynamic

load, to recover after a system failure, and to efficiently support checkpoint or restarting procedures.

At the same time, the providers of computing cycles and storage realized the difficulties in managing

a large number of systems and providing guarantees for the quality of service. Any economic advantage

offered by resource concentration was offset by the cost of management and the relatively low utilization

of costly resources.

Cloud computing is very attractive to users for several economic reasons: It requires a very low

infrastructure investment because there is no need to assemble and maintain a large-scale system and

it has low utility-based computing costs because customers are only billed for the infrastructure used.

At the same time, users benefit from the potential to reduce the execution time of compute-intensive

and data-intensive applications through parallelization. If an application can partition the workload in

n segments and spawn n instances of itself, the execution time could be reduced by a factor close to n.

Moreover, because application developers enjoy the advantages of a just-in-time infrastructure, they

are free to design an application without being concerned with the system where the application will

run. Often, an application becomes the victim of its own success, attracting a user population larger than

the system can support. Cloud elasticity allows such an application to absorb the additional workload

without any effort from application developers.

Cloud computing is also beneficial for the providers of computing cycles because it typically leads

to more efficient resource utilization. The future success of cloud computing rests on the ability of

companies promoting utility computing to convince an increasingly large segment of the user population

of the advantages of network-centric computing and content. This translates into the ability to provide

satisfactory solutions to critical aspects of security, scalability, reliability, quality of service, and the

requirements enforced by SLAs.

The appeal of cloud computing is its focus on enterprise applications. This clearly differentiates it

from the grid computing effort, which was largely focused on scientific and engineering applications.

Of course, the other major advantage of the cloud computing approach over grid computing is the

concentration of resources in large data centers in a single administrative domain.

It is expected that utility computing providers such as Amazon, Apple, Google, HP, IBM,Microsoft,

Oracle, and others will in the future develop application suites to attract customers. Microsoft seems

well positioned in the effort to attract customers to its Azure platform since it already has many applications

for enterprise customers, while Red Hat and Amazon may choose to stay with their “infrastructure

only” approach [27].

The main attraction of cloud computing is the ability to use as many servers as necessary to optimally

respond to the cost and the timing constraints of an application. This is possible only if the workload

can be partitioned in segments of arbitrary size and can be processed in parallel by the servers available

in the cloud. In Section 4.6 we discuss arbitrarily divisible workloads that can be partitioned into an

arbitrarily large number of segments; the arbitrarily divisible load-sharing model is common to many

applications, and these are precisely the applications suitable for cloud computing.

Web services, database services, and transaction-based services are ideal applications for cloud

computing. The resource requirements of transaction-oriented services are dependent on the current

load, which itself is very dynamic; the cost/performance profile of such applications benefits from an

elastic environment in which resources are available when needed and users pay only for the resources

they consume.

Not all applications are suitable for cloud computing; applications for which the workload cannot be

arbitrarily partitioned or that require intensive communication among concurrent instances are unlikely

to perform well on a cloud. An application with a complex workflow and multiple dependencies, as

is often the case in high-performance computing, could require longer execution times and higher

costs on a cloud. The benchmarks for high-performance computing discussed in Section 4.9 show that

communication and memory-intensive applications may not exhibit the performance levels shown when

running on supercomputers with low latency and high-bandwidth interconnects.