CHAPTER

1 INTRODUCTION

Conceptually, computing can be viewed as another utility, like electricity, water, or gas, accessible to every household in many countries of the world. Computer clouds are the utilities providing computing services. In utility computing the hardware and the software resources are concentrated in large data centers. The users of computing services pay as they consume computing, storage, and communication resources. While utility computing often requires a cloud-like infrastructure, the focus of cloud computing is on the business model for providing computing services.

More than half a century ago, at the centennial anniversary of MIT, John McCarthy, the 1971 Turing Award recipient for his work in Artificial Intelligence, prophetically stated: "If computers of the type I have advocated become the computers of the future, then computing may someday be organized as a public utility, just as the telephone system is a public utility... The computer utility could become the basis of a new and important industry." The prediction of McCarthy is now a technological and social reality.

Cloud computing is a disruptive computing paradigm and, as such, it required major changes in many areas of computer science and computer engineering including data storage, computer architecture, networking, resource management, scheduling, and last but not least, computer security. The chapters of this book cover the most significant challenges posed by the scale of the cloud infrastructure and the very large population of cloud users with diverse applications and requirements.

The Internet made cloud computing possible; we could not even dream of using computing and storage resources from distant data centers without fast communication. The evolution of cloud computing is organically tied to the future of the Internet. The Internet of Things (IoT) has already planted some of its early seeds in computer clouds. For example, Amazon already offers services such as Lambda and Kinesis discussed in Section 2.4.

The number of Internet users has increased tenfold from 1999 to 2013; the first billion was reached in 2005, the second in 2010, and the third in 2014. This number is even larger now, see Figure 1.1. Many Internet users have discovered the appeal of cloud computing either directly or indirectly through a variety of services, without knowing the role the clouds play in their life. In the years to come the vast computational resources provided by the cloud infrastructure will be used for the design and engineering of complex systems, scientific discovery, education, business, analytics, art, and virtually all other aspects of human endeavor. Exabytes of data stored in the clouds are streamed, downloaded, and accessed by millions of cloud users.

This chapter introduces basic cloud computing concepts in Section 1.1. The broader context of network-centric computing and network-centric content is discussed in Section 1.2. Why cloud computing became a reality in the last years after a long struggle to design large-scale distributed systems and computational grids? This question is addressed in Section 1.3, while Section 1.4 coveres the defining attributes of computer clouds and the cloud delivery models. Ethical issues and cloud vulnerability are discussed in Sections 1.5 and 1.6, respectively.

Cloud Computing. DOI: 10.1016/B978-0-12-812810-7.00001-7

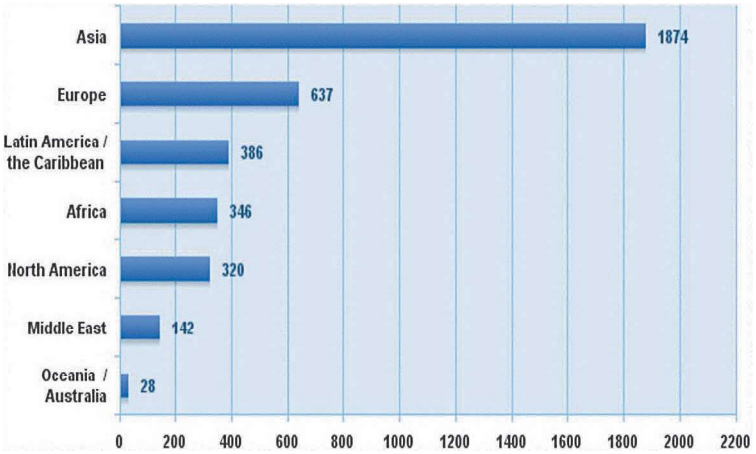

FIGURE 1.1

The number of Internet users in different regions of the world as of March 25, 2017 (in millions), according to http://www.internetworldstats.com/stats.htm.

1.1 CLOUD COMPUTING

In 2011, NIST, the US National Institute of Standards and Technology, defined cloud computing as "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."

Cloud computing is characterized by five attributes: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. DBaaS – database as a service is a more recent addition to the three cloud service delivery models: SaaS – software as a service, PaaS – platform as a service, and IaaS – infrastructure as a service. Private cloud, community cloud, public cloud, and hybrid cloud are the four deployment models shown in Figure 1.2, Section 1.4.

Cloud computing era started in 2006 when Amazon offered the Elastic Cloud Computing (EC2) and the Simple Storage Service (S3), the first services provided by Amazon Web Services (AWS). Five years later, in 2012, EC2 was used by businesses in 200 countries. S3 has surpassed two trillion objects and routinely runs more than 1.1 million peak requests per second. The Elastic MapReduce has launched 5.5 million clusters since the start of the service in May 2010 (ZDNet 2013). The range of services offered by Cloud Service Providers (CSPs), and the number of cloud users have increased dramatically during the last few years.

The cloud computing movement is motivated by the idea that data processing and storage can be done more efficiently on large farms of computing and storage systems accessible via the Internet. Computer clouds support a paradigm shift from local to network-centric computing and network-centric content where distant data centers provide the computing and storage resources. In this new paradigm users relinquish control of their data and code to Cloud Service Providers.

Cloud computing offers scalable and elastic computing and storage services. The resources used for these services can be metered and the users can be charged only for the resources they used. Cloud computing is a business reality as a large number of organizations have adopted this paradigm.

Cloud computing is cost-effective because of resource multiplexing. Application data is stored closer to the site where it is used in a manner that is device and location-independent; potentially, this data storage strategy increases reliability, as well as security. The maintenance and the security are ensured by service providers. Organizations using computer clouds are relieved of supporting large IT teams, acquiring and maintaining costly hardware and software, and paying large electricity bills. CSPs can operate more efficiently due to economy of scale.

Data analytics, data mining, computational financing, scientific and engineering applications, gaming and social networking, as well as other computational and data-intensive activities benefit from cloud computing. Storing information on the cloud has significant advantages. Content previously confined to personal devices such as workstations, laptops, tablets, and smart phones need no longer be stored locally. Data stored on computer clouds can be shared among all these devices and it is accessible whenever a device is connected to the Internet. For example, in 2011 Apple announced the *iCloud,* a network-centric alternative for content including music, videos, movies, and personal information. In February 2017 iCloud had 782 million subscribers according to http://appleinsider.com/.

Cloud computing represents a dramatic shift in the design of systems capable of providing vast amounts of computing cycles and storage space. Computer clouds use off-the shelf, low-cost components. During the previous four decades powerful, one-of-a-kind supercomputers, were built at a high cost, with the most advanced components available at the time.

In early 1990s Gordon Bell argued that one-of-a-kind systems are not only expensive to build, but the cost of rewriting applications for them is prohibitive. He anticipated that sooner or later massively parallel computing will evolve into computing for the masses [59].

There are virtually no bounds on composition of digital systems controlled by software, so we are tempted to build increasingly more complex systems including systems of systems [335]. The behavior and the properties of such systems are not always well understood. We should not be surprised that computing clouds will occasionally exhibit an unexpected behavior and large-scale systems will occasionally fail.

The architecture, the coordination mechanisms, the design methodology, and the analysis techniques for large-scale complex systems such as computing clouds will evolve in response to changes in technology, the environment, and the social impact of cloud computing. Some of these changes will reflect changes in communication, in the Internet itself in terms of speed, reliability, security, capacity to accommodate a larger addressing space by migration to IPv6, and so on.

Cloud computing reinforces the idea that computing and communication are deeply intertwined. Advances in one field are critical for the other. Indeed, cloud computing could not emerge as a feasible alternative to the traditional paradigms for high-performance computing before the Internet was able to support high-bandwidth, low-latency, reliable, low-cost communication. At the same time, modern networks could not function without powerful computing systems to manage the network. High performance switches are critical elements of both networks and computer clouds.

The complexity of the cloud computing infrastructure is unquestionable and raises questions such as: How can we manage such systems? Do we have to consider radically new ideas, such as self-management and self-repair for future clouds consisting of millions of servers? Should we migrate from

a strictly deterministic view of such complex systems to a non-deterministic one? Answers to these questions provide a rich set of research topics for the computer science and engineering community.

The cloud movement is not without skeptics and critics. The critics argue that cloud computing is just a marketing ploy, that users may become dependent on proprietary systems, that the failure of a large system such as the cloud could have significant consequences for a very large group of users who depend on the cloud for their computing and storage needs. Security and privacy are major concerns for cloud computing users.

A very important question is if under pressure from the user community the current standardization efforts will succeed. The alternative, the continuing dominance of proprietary cloud computing environments is likely to have a negative impact on the field. The cloud delivery models, SaaS, PaaS, IaaS, together with DBaaS discussed in depth in Chapter 2 will continue to coexist for the foreseeable future.

Services based on SaaS will probably be increasingly more popular because they are more accessible to lay people, while services based on the IaaS will be the domain of computer savvy individuals, large organizations, and the government. If the standardization effort succeeds, then we may see IaaS designed to migrate from one infrastructure to another and overcome the concerns related to vendor lock-in. The popularity of DBaaS services is likely to grow.

1.2 NETWORK-CENTRIC COMPUTING AND NETWORK-CENTRIC CONTENT

Network-centric computing and network-centric content concepts reflect the fact that data processing and data storage takes place on remote computer systems accessed via the ubiquitous Internet, rather than locally. The term *content* refers to any type or volume of media, be it static or dynamic, monolithic or modular, live or stored, produced by aggregation, or mixed.

The two network-centric paradigms share a number of characteristics:

- Most network-centric applications are data intensive. For example, data analytics allow enterprises to optimize their operations; computer simulation is a powerful tool for scientific research in virtually all areas of science from physics, biology, and chemistry, to archeology. Sophisticated tools for computer-aided design such as Catia (Computer Aided Three-dimensional Interactive Application) are widely used in aerospace and automotive industries. The widespread use of sensors generate a large volume of data. Multimedia applications are increasingly more popular; the larger footprint of the media data increases the load placed on storage, networking, and processing systems.
- Virtually all applications are network-intensive. Transferring large volumes of data requires high-bandwidth networks. Parallel computing, computation steering, and data streaming are examples of applications that can only run efficiently on low latency networks. Computation steering in numerical simulation means to interactively guide a computational experiment towards a region of interest.
- Computing and communication resources (CPU cycles, storage, network bandwidth) are shared and resources can be aggregated to support data-intensive applications. Multiplexing leads to a higher resource utilization; indeed, when multiple applications share a system their peak demands for resources are not synchronized and the average system utilization increases.
- Data sharing facilitates collaborative activities. Indeed, many applications in science, engineering, as well as industrial, financial, governmental applications require multiple types of analysis

of shared data sets and multiple decisions carried out by groups scattered around the globe. Open software development sites are another example of such collaborative activities.
- The systems are accessed using *thin clients* running on systems with limited resources. In June 2011 Google released Google Chrome OS designed to run on primitive devices and based on the browser with the same name.
- The infrastructure supports some form of *workflow management*. Indeed, complex computational tasks require coordination of several applications; composition of services is a basic tenet of Web 2.0.

There are sources of concern and benefits of the paradigm shift from local to network-centric data processing and storage:

- The management of large pools of resources poses new challenges as such systems are vulnerable to malicious attacks that can affect a large user population.
- Large-scale systems are affected by phenomena characteristic to complex systems such as phase transitions when a relatively small change of environment could lead to an undesirable system state [328]. Alternative resource management strategies, such as self-organization, and decisions based on approximate knowledge of the system state must be considered.
- Ensuring Quality of Service (QoS) guarantees is extremely challenging in such environments, as total performance isolation is elusive.
- Data sharing poses not only security and privacy challenges but also requires mechanisms for access control for authorized users and for detailed logs of the history of data changes.
- Cost reduction. Concentration of resources creates the opportunity to pay-as-you-go for computing and thus, eliminates the initial investment and reduces significantly the maintenance and operation costs of the local computing infrastructure.
- User convenience and elasticity, the ability to accommodate workloads with very large peak-to-average ratios.

The creation and consumption of audio and visual content is likely to transform the Internet. It is expected that the Internet will support increased quality in terms of resolution, frame rate, color depth, stereoscopic information. It seems reasonable to assume that the Future Internet[1] will be *content-centric*. *Information* is the result of functions applied to content.

The content should be treated as having meaningful semantic connotations rather than a string of bytes; the focus will be on the information that can be extracted by content mining when users request named data and content providers publish data objects. Content-centric routing will allow users to fetch the desired data from the most suitable location in terms of network latency or download time. There are also some challenges, such as providing secure services for content manipulation, ensuring global rights-management, control over unsuitable content, and reputation management.

[1] The term "Future Internet" is a generic concept referring to all research and development activities involved in development of new architectures and protocols for the Internet.

1.3 CLOUD COMPUTING, AN OLD IDEA WHOSE TIME HAS COME

It is hard to point out a single technological or architectural development that triggered the movement towards network-centric computing and network-centric content. This movement is the result of a cumulative effect of developments in microprocessor, storage, and networking technologies coupled with architectural advancements in all these areas and last but not least, with advances in software systems, tools, programming languages and algorithms supporting distributed and parallel computing.

Along the years we have witnessed the breathtaking evolution of solid state technologies which led to the development of multicore processors. Quad-core processors such as AMD Phenom II X4, Intel i3, i5, and i7, and hexacore processors such as AMD Phenom II X6 and Intel Core i7 Extreme Edition 980X are now used to build the servers populating computer clouds. The proximity of multiple cores on the same die allows cache coherency circuitry to operate at a much higher clock rate than it would be possible if signals were to travel off-chip.

Storage technology has also evolved dramatically. For example, solid state disks such as RamSan-440 allow systems to manage very high transaction volumes and larger numbers of concurrent users. RamSan-440 uses DDR2 (double-data-rate) RAM to deliver 600 000 sustained random IOPS (Input/output operations per second) and over 4 GB/second of sustained random read or write bandwidth, with latency of less than 15 microseconds and it is available in 256 GB and 512 GB configurations. The price of memory has dropped significantly; at the time of this writing the price of 1 GB module for a PC is around $5. Optical storage technologies and flash memories are widely used nowadays.

The thinking in software engineering has also evolved and new models have emerged. A software architecture and a software design pattern, the *three-tier model* has emerged. Its components are:

1. *Presentation tier,* the topmost level of the application. Typically, it runs on a desktop, PC, or workstation, uses a standard graphical user interface (GUI), and displays information related to services e.g., browsing merchandize, purchasing, and shopping cart contents. The presentation tier communicates with other tiers.
2. *Application/logic tier* controls the functionality of an application and may consist of one or more separate modules running on a workstation or application server. It may be multi-tiered itself and then the architecture is called an *n-tier architecture*.
3. *Data tier* controls the servers where the information is stored; it runs a relational database management system on a database server or a mainframe and contains the computer data storage logic. The data tier keeps data independent from application servers or processing logic and improves scalability and performance.

Any tier can be replaced independently; for example, a change of operating system in the presentation tier would only affect the user interface code.

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 size centers which 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 size data centers [37].

The networking costs, in dollars per Mbit/sec/month, are $95/13 = 7.1$ larger for medium size data centers. The storage costs, in dollars per GB/month, are $2.2/0.4 = 5.7$ larger for medium size centers.

Medium size data 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 used to keep the servers and the networking infrastructure running and heating and cooling the data centers. In 2006 the data centers reportedly consumed 61 billion kWh, 1.5% of all electric energy in the U.S., at a cost of $4.5 billion. We have seen a 4% increase in total data center energy consumption from 2010 to 2014.

In 2014 there were about three million data centers in the U.S. The data centers consumed about 70 billion kWh, representing about 2% of the total energy consumption in the U.S. This is the equivalent of the electric energy consumed by about 6.4 million average homes in the U.S. that year. The energy costs differ from state to state, e.g., one kWh costs 3.6 cents in Idaho, 10 cents in California, and 18 cents in Hawaii. This explains why cloud data centers are placed in regions with lower energy cost.

A natural question to ask is: Why cloud computing could be successful when other paradigms have failed? The reasons why cloud computing could be successful can be grouped in several general categories: technological advances, a realistic system model, user convenience, and financial advantages. A non-exhaustive list of reasons for the success of cloud computing includes:

- 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 to promote the new technologies.
- A cloud consists of a mostly 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 [160,164]; its adoption by industrial organizations, financial institutions, healthcare organizations and so on, has a potentially huge impact on the economy.
- A cloud provides the illusion of infinite computing resources; its elasticity frees the applications 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 industry-wide 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 GM move its IT activities 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 (SLA)s discussed in Section 2.9. A temporary fix but with negative economical implications is *overprovisioning*, i.e., having enough resources to satisfy the largest projected demand.

- Vendor lock-in; once a customer is hooked to one cloud service provider it is hard to move to another. The standardization efforts at NIST attempt to address this problem.
- Data confidentiality and auditability; this is indeed a serious problem analyzed in Chapter 11.
- Data transfer bottlenecks critical for data-intensive applications. Transferring 1 TB of data on a 1 Mbps network takes 8 000 000 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, e.g., a 1 Gbps network would reduce this time to 8 000 seconds, or slightly more than 2 hours.
- Performance unpredictability; this is one of the consequences of resource sharing. Strategies for performance isolation are discussed in Section 10.1.
- Elasticity, the ability to scale up and down quickly. New algorithms for controlling resource allocation and workload placement are necessary. Autonomic computing based on self-organization and self-management seems to be a promising avenue.

There are other perennial problems with no clear solutions at this time, including software licensing and dealing with systems bugs.

1.4 CLOUD DELIVERY MODELS AND DEFINING ATTRIBUTES

Cloud computing delivery models, deployment models, defining attributes, resources, and organization of the infrastructure discussed in chapter are summarized in Figure 1.2. The cloud delivery models, SaaS, PaaS, IaaS, and DBaaS can be deployed as public, private, community, and hybrid clouds.

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

- Cloud computing uses Internet technologies to offer elastic services. The term "elastic computing" refers to the ability of dynamically acquiring computing resources and supporting 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 used.
- The 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.

The term "computer cloud" is overloaded as it covers infrastructures of different sizes, with different management, and a different user population. Several types of clouds 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.

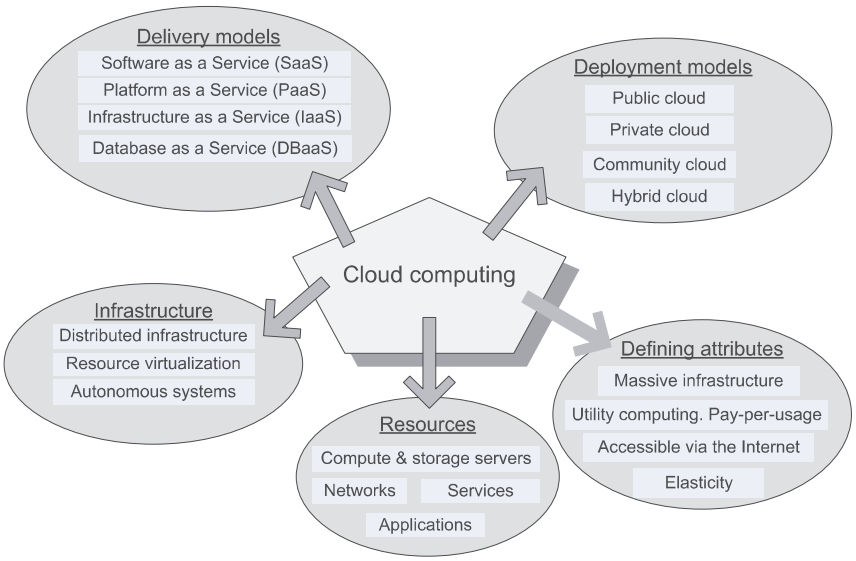

FIGURE 1.2

Cloud computing: delivery models, deployment models, defining attributes, resources, and organization of the infrastructure.

- 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 can provide the computing resources needed by a large organization, e.g., 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 the user of a private cloud does pays as it consumes resources [37]. 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.

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, when the 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.

The obvious advantage of network centric content is the accessibility of information from any site where one could connect to the Internet. Clearly, information stored on a cloud can be shared easily, but this approach raises also major concerns: Is the information safe and secure? Is it accessible when we need it? Do we still own it?

In the next 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 that: "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 the technology trends, Jim Gray emphasized that the cost of communication in a wide area network has decreased dramatically and will continue to do so. Thus, it makes economical sense to store the data near the application [202], 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 [37].

As always, a good idea has generated a high level of excitement 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 cloud computing.

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 along the years from this experience. This is the reason we discuss concurrency in Chapter 3 and parallel and distributed systems in Chapter 4.

1.5 ETHICAL ISSUES IN CLOUD COMPUTING

Cloud computing is based on a paradigm shift with profound implications on computing ethics. The main elements of this shift are:

1. The control is relinquished to third party services.
2. The data is stored on multiple sites administered by several organizations.
3. Multiple services interoperate across the network.

Unauthorized access, data corruption, infrastructure failure, and service unavailability are some of the risks related to relinquishing the control to third party services; moreover, whenever a problem occurs it is difficult to identify the source and the entity causing it. Systems can span the boundaries of multiple organizations and cross the security borders, a process called *de-perimeterisation*. As a result

of de-perimeterisation "not only the border of the organizations IT infrastructure blurs, also the border of the accountability becomes less clear" [485].

The complex structure of cloud services make it difficult to determine who is responsible for each action. Many entities contribute to an action with undesirable consequences and no one can be held responsible, the so-called "problem of many hands."

Ubiquitous and unlimited data sharing and storage among organizations test the self-determination of information, the right and/or the ability of individuals to exercise personal control over the collection, the use, and the disclosure of their personal data by others. This tests the confidence and trust in today's evolving information society. Identity fraud and theft are made possible by the unauthorized access to personal data in circulation and by new forms of dissemination through social networks. All these factors could also pose a danger to cloud computing.

Cloud service providers have already collected petabytes of sensitive personal information stored in data centers around the world. The acceptance of cloud computing will be determined by the effort dedicated by the CSPs and the countries where the data centers are located to ensure privacy. Privacy is affected by cultural differences; while some cultures favor privacy, other cultures emphasize community and this leads to an ambivalent attitude towards privacy in the Internet which is a global system.

The question of what can be done proactively about ethics of cloud computing does not have easy answers as many undesirable phenomena in cloud computing will only appear in time. However, the need for rules and regulations for the governance of cloud computing are obvious. Governance means the manner something is governed or regulated, the method of management, the system of regulations. Explicit attention to ethics must be paid by governmental organizations providing research funding; private companies are less constraint by ethics oversight and governance arrangements are more conducive to profit generation.

Accountability is a necessary ingredient of cloud computing; adequate information about how data is handled within the cloud and about allocation of responsibility are key elements for enforcing ethics rules in cloud computing. Recorded evidence allows us to assign responsibility; but there can be tension between privacy and accountability and it is important to establish what is being recorded, and who has access to the records.

Unwanted dependency on a cloud service provider, the so-called *vendor lock-in*, is a serious concern and the current standardization efforts at NIST attempt to address this problem. Another concern for the users is a future with only a handful of companies which dominate the market and dictate prices and policies.

1.6 CLOUD VULNERABILITIES

Clouds are affected by malicious attacks and failures of the infrastructure, e.g., power failures. Such events can affect the Internet domain name servers and prevent access to a cloud or can directly affect the clouds. For example, an attack at Akamai on June 15, 2004 caused a domain name outage and a major blackout that affected Google, Yahoo, and many other sites. In May 2009, Google was the target of a serious denial of service (DNS) attack which took down services like Google News and Gmail for several days.

Lightning caused a prolonged down time at Amazon on June 29–30, 2012; the AWS cloud in the East region of the US which consists of ten data centers across four availability zones, was initially

troubled by utility power fluctuations, probably caused by an electrical storm. Availability zones are locations within data center regions where public cloud services originate and operate. A June 29, 2012 storm on the East Coast took down some of Virginia based Amazon facilities and affected companies using systems exclusively in this region. Instagram, a photo sharing service, was one of the victim of this outage according to http://mashable.com/2012/06/30/aws-instagram/.

The recovery from the failure took a very long time and exposed a range of problems. For example, one of the ten centers failed to switch to backup generators before exhausting the power that could be supplied by UPS units. AWS uses "control planes" to allow users to switch to resources in a different region and this software component also failed. The booting process was faulty and extended the time to restart EC2 and EBS services.

Another critical problem was a bug in the Elastic Load Balancer (ELB), used to route traffic to servers with available capacity. A similar bug affected the recovery process of the Relational Database Service (RDS). This event brought to light "hidden" problems that occur only under special circumstances.

The stability risks due to interacting services are discussed in [177]. A cloud application provider, a cloud storage provider, and a networks provider could implement different policies and the unpredictable interactions between load-balancing and other reactive mechanisms could lead to dynamic instabilities. The unintended coupling of independent controllers which manage the load, the power consumption, and the elements of the infrastructure could lead to undesirable feedback and instability similar with the one experienced by the policy-based routing in the Internet BGP (Border Gateway Protocol).

For example, the load balancer of an application provider could interact with the power optimizer of the infrastructure provider. Some of these couplings may only manifest under extreme condition and be very hard to detect under normal operating condition, but could have disastrous consequences when the system attempts to recover from a hard failure, as in the case of the AWS 2012 failure.

Clustering resources in data centers located in different geographical areas lowers the probability of catastrophic failures. This geographic dispersion of resources could have additional positive side effects such as reduction of communication traffic, lowering energy costs by dispatching the computations to sites where the electric energy is cheaper, and improving performance by an intelligent and efficient load balancing strategy.

Sometimes, a user has the option to decide where to run an application; we shall see in Section 2.3 that an AWS user has the option to choose the regions where the instances of his/her applications will run, as well as the regions of the storage sites. System's objective, maximize throughput, resource utilization, and financial benefits have to be carefully balanced with the user needs, low cost and response time and maximum availability.

The price to pay for any system optimization is an increased system complexity. For example, the latency of communication over a Wide Area Network (WAN) is considerably larger than the one over a Local Area Network (LAN) and requires the development of new algorithms for global decision making.

Chapter 2 takes a closer look at the cloud ecosystem as of late 2016. The next two chapters, Chapters 3 and 4 discuss concurrency concepts and parallel and distributed computing principles concepts relevant to cloud computing. Both subjects are very broad and we only cover aspects particularly relevant to cloud computing. An in depth analysis of networking access to computer clouds and cloud data storage are the subjects of the Chapters 5 and 6, respectively.