1

Introduction

This chapter introduces Cloud computing. The introduction helps the reader to get an overview of Cloud computing and its main challenges. Subsequent chapters of this book assume the reader understands the content of this chapter.

1.1 Overview

Cloud computing originates from industry (commercial requirements and needs). Governments and leading industrial bodies involved academia at early stages of adopting Cloud computing because of its promising future as an Internet-scale critical infrastructure. Involving academia would ensure that Cloud computing is critically analyzed, which helps in understanding its problems and limitations. This would also help in advancing the knowledge of this domain by defining and executing research road maps to establish next-generation trustworthy Cloud infrastructure. Moreover, academia would provide the required education in Cloud computing by developing undergraduate and postgraduate courses in this domain.

Cloud comes with enormous advantages; for example, it reduces the capital costs of newly established businesses, it reduces provisioning time of different types of services, it establishes new business models, it reduces the overhead of infrastructure management, and it extends IT infrastructures to the limits of their hosting Cloud infrastructure. Although Cloud computing is associated with such great features, it also has critical problems preventing its wider adoption by critical business applications, critical infrastructures, or even end-users with sensitive data. Examples of such problems include: security and privacy problems, operational management problems, and legal concerns. The immaturity of Cloud and the generosity of its allocated funds have made Cloud computing, in a relatively short period of time, one of the most in-demand research topics around the world.

Cloud computing is built on complex technologies which are not easy to understand, as an integrated science, for many people working in the industry and academia. A fundamental reason behind this is the lack of resources analyzing current Cloud infrastructure, its properties and limitations [1, 2]. The main objective of this book is to establish the foundations of Cloud computing, which would help researchers and professionals to understand Cloud as an

Cloud Management and Security, First Edition. Imad M. Abbadi.
© 2014 John Wiley & Sons, Ltd. Published 2014 by John Wiley & Sons, Ltd.
Companion Website: www.wiley.com/go/abbadi_cloud

integrated science. Understanding the Cloud structure and properties is key for conducting practical research in this area that could possibly be adopted by industry.

Most current research assumes Cloud computing is a black-box that has physical and virtual resources. The lack of careful understanding of the properties, structure, management, and operation of the black-box results in confusion and misunderstanding. In terms of misunderstanding, this relates to Cloud's limitations and the expectations of what it could practically provide. For example, some people claim that Cloud has immediate and unlimited capabilities, that is immediate and unlimited scalability. This is not practical considering present-day technologies, such as the limitations of hardware resources. There are also many other factors that have not been considered in such strong claims, for example should Cloud provide unlimited resources in case of application software bugs? Should resources be available immediately upon request without users' prior agreement? This book discusses these issues in detail.

This chapter is organized as follows. Section 1.2 discusses the definition of Cloud computing. Section 1.3 clarifies the evolution of Cloud computing. Section 1.4 discusses Cloud services. Section 1.5 discusses Cloud deployment types. Section 1.6 discusses the main challenges of Clouds. Finally, we summarize the chapter in Section 1.7 and provide a list of exercises in Section 1.8.

1.2 Cloud Definition

Cloud computing is a new buzzword in computing terms and it is associated with various definitions. In this book we focus on two definitions: the first is provided by the National Institute of Standards & Technology (NIST) [2] and the second is provided by an EU study of the future directions of Clouds [3]. The main reasons for analyzing these definitions in particular are:

- The good reputation of the organizations behind the definitions. For example, the EU study was edited by representatives of leading universities and industrial bodies such as Oracle, Google, Microsoft, and IBM.
- We found thsse definitions to be unique, such that their combination provides the most important elements of Cloud as covered throughout this book.

> NIST defines Cloud 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 [2].

In contrast:

> An EU study defines Cloud as an elastic execution environment of resources involving multiple stakeholders and providing a metered service and multiple granularities for specified level of quality [3].

Although both definitions come from reputable organizations, they are not consistent. This is not to say that either of them is wrong, but they are incomplete. Both definitions reveal many important keywords reflecting Clouds capabilities; however, a careful analysis of these definitions shows they only have one keyword in common. The first definition uses *'rapidly provisioned and released'* while the second definition uses *'elastic execution.'* These two keywords have the same objective. However, other keywords are not the same, for example 'minimal management effort' as stated by the NIST definition is not stated anywhere in the EU definition. Similarly, the EU definition uses the keyword 'metered service' which is again not stated anywhere in the NIST definition.

Cloud computing is in fact a combination of both definitions as each definition provides a partial view of the Cloud attributes. Therefore, we could redefine Cloud computing as follows:

> Cloud computing is a model involving multiple stakeholders and 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. The model provides a metered service and multiple granularities for a specified level of quality.

This book focuses primarily on the details behind the elements in the definition which would clarify the Cloud computing black-box.

1.3 Cloud Evolution

Enterprise infrastructures witnessed three major fundamental changes, which were a result of major innovations in computer science. These are as follows:

- *Traditional enterprise infrastructure*. This is the foundation of the virtualization era. Initially, it starts with a few powerful servers (what used to be called mainframes). With advances in technologies and an increased number of required applications, the number of servers increases rapidly. This results in a huge number of resources within an enterprise infrastructure. Despite the complexity of the traditional enterprise infrastructure, the relationship between customers and their resources is simple. Within this, the requirements of customers are carefully analyzed by system analysts. The system analysts forward the analyzed results to enterprise architects. The enterprise architects deliver an architecture which is designed to address the needs of a specific customer application requirement. The resources required by the delivered architecture in most cases run a specific customer applications. This process results in a one-to-one relationship between architecture and customer. Such a relationship causes huge wastage of resources including, for example, computational resources, power consumption, and data-center spaces. In contrast, this relationship results in a relatively more secure and customized design than the other evolution models of enterprise infrastructure.
- *Virtual enterprise infrastructure.* This is the foundation of today's Cloud infrastructure. The problems of the traditional enterprise infrastructure, which affect the green agenda, require novel innovations enabling customers to share resources without losing control or

increasing security risks. This was the start of the virtualization era, which brings tremendous advantages in terms of consolidating resources and results in effective utilization of power, data-center space, etc. A virtual enterprise infrastructure suffers from many problems, such as security, privacy, and performance problems, which restricts many applications from running on virtual machines. As a result, virtual infrastructures for many enterprises support applications that run on virtual resources and those that run directly on physical resources.

The virtualization era changes the mentality of enterprise architects as the relationship between users and their physical resources is no longer one-to-one. This raises a big challenge in terms of how such a consolidated virtualized architecture could satisfy users' dynamic requirements and unique application nature. Enterprise architects address this by studying the environment inherited from the traditional enterprise infrastructure, to find that different architectures have some similarities. The similarities between independent applications enable enterprise architects to split the infrastructure into groups. Each group has architecture-specific static properties. The properties enable the group to address common requirements of a certain category of applications. For example, a group could be allocated to applications that tolerate a single point of failure; another group could be allocated to applications that require full resilience with no single point of failure; a third group could be allocated to applications that are highly computational; a group for archiving systems; and so on.

The second part of the challenging question is how such a grouping, which is associated with almost static properties, could be used to address users' dynamic requirements and their unique application nature. Enterprise architects realize that virtualization can be fine-tuned and architected to support the dynamic application requirements which cannot be provided by the physical group static properties. In other words, a combination of static physical properties and dynamic virtual properties is used to support customer expectations in a virtual enterprise infrastructure.

- *Cloud infrastructure*. This has evolved from the virtual enterprise infrastructure. Chapters 2 and 4 cover the details of Cloud structure and its attributes. Clouds come with many important and promising features, such as direct interaction with customers via supplied APIs, automatically managed resources via self-managed services, and support for a pay-per-use model. In addition, Cloud computing comes with new promising business models that would enable more efficient utilization of resources and quicker time-to-market. Cloud computing inherits the problems of the virtual infrastructure and in addition, it comes with more serious problems including security problems, operational and data management problems. The problems associated with Cloud prevent its wider adoption, especially by critical organizations. This chapter discusses the most important problems in Clouds.

1.4 Cloud Services

Cloud services are also referred to as Cloud types in some references. These are served by Cloud providers to their customers following a pre-agreed service level agreement (SLA). Figure 1.1 illustrates the commonly agreed Cloud services in the context of a Cloud environment. Understanding these services requires understanding the structure of the Cloud, which is discussed in detail in Chapter 2. As illustrated in the figure, the Cloud structure could be viewed

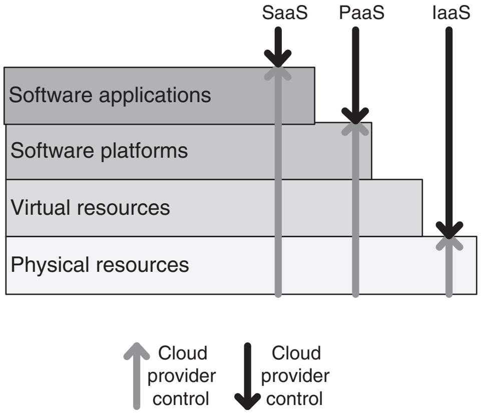

Figure 1.1 Cloud services

based on the hosting relationship as the following hierarchical layers: physical layer, virtual layer, software platform and software application layer. The physical layer is composed of all physical components and their management software components, including the operating system and the hypervisor. The virtual layer is composed of virtual machines, virtual storage, and a virtual network managed by the physical layer. The software application and software platforms are self-explanatory, and could be served either by the virtual layer or directly by the physical layer.

The management of Cloud services is a shared responsibility between the Cloud provider and their customers. The level of responsibility is Cloud service specific, as explained next. Cloud computing has the following main services.

- *Infrastructure as a service (IaaS).* IaaS provides virtual compute and store resources as a service to customers. Cloud providers in IaaS manage the physical resources and their hypervisors. Cloud customers run their software stack and manage the content of their allocated virtual resources, including guest operating system. Customers in this type should, in principle, have overall control of their data. At the time of writing, Cloud providers have ultimate control of customer data.
- *Platform as a service (PaaS).* PaaS provides the environment and software platforms that Cloud customers can use to develop and host their own software applications. Unlike IaaS, PaaS customers do not manage the software platforms provided by the Cloud, but only need to manage their own software stack. Cloud providers of PaaS expose their own APIs, which are used directly by customer applications. The exposed APIs, at the time of writing, do not follow any standard. As a result, Cloud customers of PaaS cannot move their applications transparently across competing Cloud providers.
- *Software as a service (SaaS).* SaaS provides ready-to-consume software applications which address the needs of specific business functions and processes. Cloud providers manage the software applications and the hosting environment completely. Cloud customers might need to manage their specific configurations within the supported software application.

We conclude from the above that Cloud computing provides full outsourcing support for SaaS, partial outsourcing support for PaaS, and minimal outsourcing support for IaaS. That is, IaaS in theory provides customers with the greatest control over their resources, while SaaS provides Cloud providers with the greatest control over their customers' data and Cloud customers with the least control over their resources.

The above services are the main services a Cloud provider supports. Some references discuss other services, such as backup as a service, log as a service, etc. These services would be categorized under the above main services. For example, backup as a service could be viewed as SaaS. A Cloud customer does not necessarily need to stick to one service. It is, rather, likely for a Cloud customer to have a combination of different services. The selection of the service should be based on different complex factors, such as: the nature of the hosted application that will be using the service, the customer level of competence in using IT, the desired level of control, security and privacy requirements, cost factors, and legal requirements. We discuss these in detail throughout the book.

1.5 Cloud Deployment Types

Clouds have the following main deployment types (also referred to as deployment models in some references):

- *Public Cloud.* The infrastructure of a public Cloud is owned by the Cloud provider, but leased to Cloud customers. The Cloud provider typically manages its physical infrastructure, but it could outsource specific functions to a third party as in the case of outsourcing hardware maintenance. Example of this type includes Amazon and RackSpace.
- *Private Cloud.* A private Cloud deployment type is owned and used by a specific enterprise. That is, the enterprise employees are the only customer of the private Cloud. The private Cloud could either directly manage its own infrastructure or it could outsource the management to a third party. Example of a private Cloud deployment type includes most banks and telecoms infrastructure.
- *Community Cloud.* Organizations sharing common business functions and/or objectives could collaborate and establish their own specific community Cloud infrastructure. Example of this include Associated Newspapers which is a group of newspapers and publishing media that establish a community Cloud infrastructure to serve their common needs.
- *Hybrid Cloud.* This deployment type is a mixture of private, community, and/or public Cloud. This is important to support higher resilience, availability, and reliability.

Public Cloud has many more customers than private and community Clouds. As a result, public Cloud hosts more services and has intensive interactions with customers. Managing the huge customer base of public Cloud necessitates the public Cloud only hosting services that could be fully managed automatically with minimal human intervention. Automation hides the complexity of the infrastructure and increases its resilience. At the current time fully automated management services are not yet available for most types of applications and virtual resources. Such a lack of automated management services forces public Cloud providers to mainly support basic services which can be automated. These basic services currently cover the needs of casual users, small businesses, and uncritical applications.

Community and private Cloud deployment types, however, establish strong relations with their customers. That is, customers typically have a relationship of mutual benefit or shared goals with the Cloud provider; customers may also be contractually bound to good behavior. These characteristics give rise to a substantial degree of trust in the Cloud; its architecture is also important, but perhaps less so. By contrast, users of public Clouds are much more reliant on infrastructure properties in order to establish trust.

The hybrid Cloud model is different from the above as it is a mixture of different Cloud deployment types. Carefully managing it could result in higher reliance, reliability or even a reduction in costs. For example, a hybrid Cloud could be composed of a public and private Cloud such that the private Cloud hosts the critical and dependent application and the public Cloud acts as a web front-end or stores protected backup. The hybrid Cloud could also result in higher risk if badly configured and managed. For example, if a hybrid Cloud is composed of a highly secure private Cloud and a public Cloud, an attacker could attack the weakest link (i.e., the public Cloud) and from there get into the private Cloud. Therefore, careful risk analysis and management would need to be conducted not only when outsourcing services into public and community Cloud types, but importantly when moving into a hybrid Cloud type.

1.6 Main Challenges of Clouds

The EU study of Cloud [3] states the following:

> Cloud technologies and models have not yet reached their full potential and many of the capabilities associated with Clouds are not yet developed and researched to a degree that allows their exploitation to the full degree, respectively meeting all requirements under all potential circumstances of usage.

This strong statement implicitly indicates that Cloud is still at an early stage of development and there are lots of challenges that still need to be addressed in this domain. In this section we highlight the most commonly discussed challenges in Clouds. This book discuss the challenges and how they could be managed using today's technologies. The main challenges in Clouds are as follows:

- *Operational management*. The scale, heterogeneity, and number of services and users of Cloud computing are by far more complex than traditional enterprise infrastructure. This requires automating the process of managing the Cloud environment as the management of Cloud computing is beyond the capabilities of typical human administrations and current system management tools. Providing fully automated management services is one of the key challenges in Cloud, which is discussed in great detail throughout this book. The following are example of cases which currently require excessive human intervention:
  - *Automated and effective elasticity property.* This means that resources which are used by a service should reflect the real needs of the service. For example, running applications should immediately utilize allocated virtual computation, storage, and memory resources without the need to do further updates and/or restarts. This is not provided effectively at the current time, which results in an increase in operational management costs and,

in addition, affects the green agenda. Such a case would require an optimized scheduler which considers the green agenda, SLA and QoS. For example, it is more efficient to not power up resources and delay execution if (i) utilized resources will be available shortly and (ii) SLA/QoS are maintained.

  - *Self-detection of failure and automated recovery.* Failure management within enterprise infrastructures is provided manually with support from the limited available management tools. Such a semi-automatic process reduces the resilience and availability of the infrastructural resources.

- *Data management*. The amount of stored data in the Cloud is huge and increasing massively. Controlling the distribution of data is a big challenge that requires full consideration of legislation, security, privacy, and performance factors. This problem is considered in the first part of this book. The following are examples of data management problems:
  - The huge volume of Cloud data affects data availability and transmission, as the greater the size of data the more complex it is to control its movement across the distributed elements of Clouds.
  - The lack of automated data management mechanisms has a direct effect on the provided QoS.
  - Data management is a major concern when scaling and shrinking resources, which is a result of Cloud elasticity. Cloud elasticity requires ensuring consistency and security of data when replicated and shrunk.
  - Classical DBMS may break in Cloud considering the latency of accessing disks and the cache coherency across a very large number of nodes.
- *Privacy, security, and trust*. Establishing trust in Clouds is the ultimate objective of most research in this domain. Other discussed challenges will eventually help in establishing trust in the Cloud. Privacy, security, and trust is a top challenge of Cloud that directly prevents its wider adoption, especially by critical infrastructure. Clouds suffer from major security concerns, for example: physical resources shared by many (possibly competing) customers – what is known as the *multi-tenant architecture*; vulnerability to the insiders threat of traditional enterprises; complex and heterogeneous architecture increasing security vulnerabilities. In addition, the Cloud elasticity results in security vulnerabilities when replicating, distributing, and shrinking data. This process must validate the non-existence of security holes in remote servers. Equally importantly, in current Cloud users do not have control over their resources, for example users cannot be assured about the way Cloud manages resources, about the integrity of their bills, and about Cloud's compliance with the greed SLA.
- *Forensic and provenance in Clouds*. This is one of the main issues in Clouds, and it helps in addressing many other challenges. A key fundamental requirement for establishing trust in Cloud is having a trustworthy provenance mechanism. Provenance helps in supporting proactive service management, assuring the integrity of bills, providing incident management, and lessening the impacts of insider threats, which increase Cloud trustworthiness. We devote Chapter 10 to this important topic.
- *Federation and interoperability*. The future vision of Cloud computing is to be the Internet-scale critical infrastructure. This strong vision requires trustworthy and resilient Cloud infrastructure that can survive even with failures of multiple Cloud providers. Addressing such a requirement requires establishing a Cloud-of-Clouds (what is also referred to as federated Clouds). The future vision of Cloud computing also enables customers to switch

transparently between Cloud providers. Such visions (i.e., federated Clouds and flexibility in changing Cloud providers) are not available at the current time. One of the main reasons for this is the lack of standardization in this domain. The first part of this book presents the taxonomy of federated Clouds and briefly discusses this challenge.

- *Performance management*. This is a key subject for the success of Clouds, especially when considering the complexity, enormous customer base, and criticality of the Cloud. For example, high performance is a key for: managing the operation of the Cloud (e.g., scalability and resource scheduling), copying large amounts of data within the Cloud infrastructure and across federated Clouds, copying large amounts of data between Cloud customers and the Cloud infrastructure, and copying large amounts of data across distant locations within the Cloud infrastructure and across federated Clouds.
- *Legislation and policies*. Different countries have their own legislation in terms of where data could be hosted and which data is allowed. Cloud computing has many limitations for complying with different legislations. For example, current Cloud does not have the capabilitly to allow users to enforce the location of where their data could be stored and/or processed. In addition, current Cloud computing does not provide users with the capability to enforce their requirements (e.g., data privacy and security) and neither does it provide the assurance of their enforcement. This book does not cover the details of legal issues in Clouds; however, the frameworks which are discussed in the second part of the book look at how it addresses some of the legal requirements.
- *Economical aspects*. It is not always the case that switching to Cloud would provide the most economical approach. This is especially the case for well-established businesses that already have an enterprise infrastructure. Organizations would need to carefully balance and understand the risk and economical values when switching to Clouds. This book discusses the factors that would need to be considered when switching to Cloud, what services to outsource into Cloud, and the Cloud type that best suits an organization's needs.

  Cloud computing helps in supporting green IT. For example, it offers possibilities to reduce carbon emission through more efficient resource usage; however, this needs to be counterweighed with the indirect carbon footprint arising from more experimental and thus more overall usage of resources, and the pressure on Cloud providers to update their infrastructure more regularly and faster than the average user.

1.7 Summary

Cloud computing is a recent term in IT, which started in 2006 with Amazon EC2. Cloud computing has emerged from commercial requirements and thus it draws an enormous amount of attention from both industry and academia, because of its promising future. Cloud comes with great advantages to help with economic growth, such as supporting the green agenda, reducing operational man-power, and providing effective utilization of resources. The lack of widely accepted academic studies that formally analyze the current Cloud infrastructure results in confusion over realizing its potential features, misunderstanding of some Cloud properties, and underestimating the challenges involved in achieving some of the potential features of Cloud. Discussing these was one of the main objectives of this chapter. The chapter also discussed Cloud services, deployment types, and main challenges. Subsequent chapters of the book build on the concepts presented in this chapter.

1.8 Exercises

Q1. What are the main features of Cloud which differentiate it from traditional data centers and enterprise infrastructures?

Q2. Cloud provides different services (i.e., IaaS, PaaS, and SaaS). Discuss the main differences between Cloud services.

Q3. Discuss the different Cloud deployment types.

Q4. What are the advantages and disadvantages of Clouds?

Q5. Organizations should understand the risks involved when outsourcing their data and services to public Clouds, and they should consider the available security and privacy options provided by Clouds. Can you identify some of the risks and how they could be managed?

Q6. The NIST definition of Cloud computing includes the statement ‘minimal management effort or service provider interaction.’ Discuss the importance of this statement in the Cloud definition.

References

[1] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy H. Katz, Andrew Konwinski *et al*. Above the Clouds: A Berkeley View of Cloud Computing. Technical Report No. UCB/EECS-2009-28, University of California, Berkeley, CA, February 2009.

[2] Peter Mell and Tim Grance. The NIST Definition of Cloud Computing, 2009.

[3] Keith Jeffery and Burkhard Neidecker-Lutz. The Future of Cloud Computing – Opportunities for European Cloud Computing and Beyond, 2010.