1

Introduction

> If the seventeenth and early eighteenth centuries are the age of clocks, and the later eighteenth and the nineteenth centuries constitute the age of steam engines, the present time is the age of communication and control.
>
> **Norbert Wiener** (from the 1948 edition of *Cybernetics: or Control and Communication in the Animal and the Machine*).

It is unfortunate that we don't remember the exact date of the extraordinary event that we are about to describe, except that it took place sometime in the Fall of 1994. Then Professor Noah Prywes of the University of Pennsylvania gave a memorable invited talk at Bell Labs, at which two authors[1] of this book were present. The main point of the talk was a proposal that AT&T (of which Bell Labs was a part at the time) should go into the business of providing computing services—in addition to telecommunications services—to other companies by actually running these companies' data centers. "All they need is just to plug in their terminals so that they receive IT services as a utility. They would pay anything to get rid of the headaches and costs of operating their own machines, upgrading software, and what not."

Professor Prywes, whom we will meet more than once in this book, well known in Bell Labs as a software visionary and more than that—the founder and CEO of a successful software company, *Computer Command and Control*—was suggesting something that appeared too extravagant even to the researchers. The core business of AT&T at that time was telecommunications services. The major enterprise customers of AT&T were buying the *customer premises equipment* (such as private branch exchange switches and machines that ran software in support of call centers). In other words, the enterprise was buying things to run on premises rather than outsourcing things to the network provider!

Most attendees saw the merit of the idea, but could not immediately relate it to their day-to-day work, or—more importantly—to the company's stated business plan. Furthermore, at that very moment the Bell Labs computing environment was migrating from the Unix

[1] Igor Faynberg and Hui-Lan Lu, then members of the technical staff at Bell Labs Area 41 (Architecture Area).

Cloud Computing: Business Trends and Technologies, First Edition. Igor Faynberg, Hui-Lan Lu and Dor Skuler.

programming environment hosted on mainframes and Sun workstations to Microsoft Office-powered personal computers. It is not that we, who "grew up" with the Unix operating system, liked the change, but we were told that this was the way the industry was going (and it was!) as far as office information technology was concerned. But if so, then the enterprise would be going in exactly the *opposite* way—by placing computing in the hands of each employee. Professor Prywes did not deny the pace of acceptance of personal computing; his argument was that there was much more to enterprises than what was occurring inside their individual workstations—payroll databases, for example.

There was a lively discussion, which quickly turned to the detail. Professor Prywes cited the achievements in virtualization and massive parallel-processing technologies, which were sufficient to enable his vision. These arguments were compelling, but ultimately the core business of AT&T was networking, and networking was centered on telecommunications services.

Still, telecommunications services were provided by software, and even the telephone switches were but peripheral devices controlled by computers. It was in the 1990s that virtual telecommunications networking services such as *Software Defined Networks*—not to be confused with the namesake development in data networking, which we will cover in Chapter 4—were emerging on the purely software and data communications platform called *Intelligent Network*. It is on the basis of the latter that Professor Prywes thought the computing services could be offered. In summary, the idea was to combine data communications with centralized powerful computing centers, all under the central command and control of a major telecommunications company. All of us in the audience were intrigued.

The idea of computing as a public utility was not new. It had been outlined by Douglas F. Parkhill in his 1966 book [1].

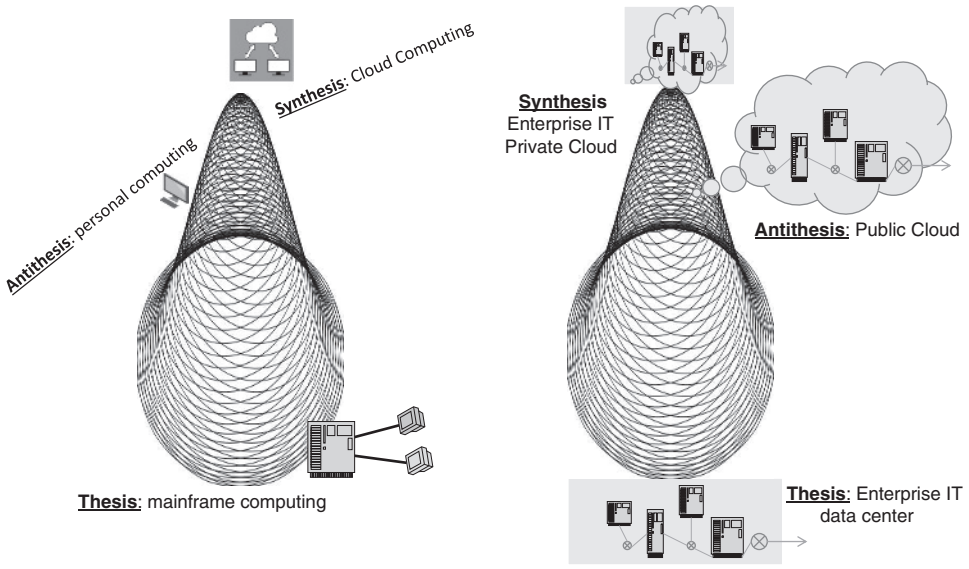
In the end, however, none of us could sell the idea to senior management. The times the telecommunications industry was going through in 1994 could best be characterized as "interesting," and AT&T did not fare particularly well for a number of reasons.[2] Even though Bell Labs was at the forefront of the development of all relevant technologies, recommending those to businesses was a different matter—especially where a proposal for a radical change of business model was made, and especially in turbulent times.

In about a year, AT&T announced its trivestiture. The two authors had moved, along with a large part of Bell Labs, into the equipment manufacturing company which became Lucent Technologies and, 10 years later, merged with Alcatel to form Alcatel-Lucent.

At about the same time, Amazon launched a service called *Elastic Compute Cloud* (*EC2*), which delivered pretty much what Professor Prywes had described to us. Here an enterprise user—located anywhere in the world—could create, for a charge, *virtual* machines in the "Cloud" (or, to be more precise, in one of the Amazon data centers) and deploy any software on these machines. But not only that, the machines were *elastic*: as the user's demand for computing power grew, so did the machine power—magically increasing to meet the demand—along with the appropriate cost; when the demand dropped so did the computing power delivered, and also the cost. Hence, the enterprise did not need to invest in purchasing and maintaining computers, it paid only for the computing power it received and could get as much of it as necessary!

As a philosophical aside: one way to look at the computing development is through the prism of dialectics. As depicted in Figure 1.1(a), with mainframe-based computing as the

[2] For one thing, the regional Bell operating companies and other local exchange carriers started to compete with AT&T Communications in the services market, and so they loathed buying equipment from AT&T Network Systems—a manufacturing arm of AT&T.



(a) From mainframe to Cloud (b) From the IT data center to Private Cloud

Figure 1.1 Dialectics in the development of Cloud Computing: (a) from mainframe to Cloud; (b) from IT data center to Private Cloud.

thesis, the industry had moved to personal-workstation-based computing—the *antithesis*. But the spiral development—fostered by advances in data networking, distributed processing, and software automation—brought forth the Cloud as the *synthesis*, where the convenience of seemingly central on-demand computing is combined with the autonomy of a user's computing environment. Another spiral (described in detail in Chapter 2) is depicted in Figure 1.1(b), which demonstrates how the *Public Cloud* has become the *antithesis* to the *thesis* of traditional IT data centers, inviting the outsourcing of the development (via "*Shadow IT*" and *Virtual Private Cloud*). The synthesis is *Private Cloud*, in which the Cloud has moved computing back to the enterprise but in a very novel form.

At this point we are ready to introduce formal definitions, which have been agreed on universally and thus form a standard in themselves. The definitions have been developed at the National Institute of Standards and Technology (NIST) and published in [2]. To begin with, Cloud Computing is defined 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." This Cloud model is composed of five essential characteristics, three service models, and four deployment models.

The five essential characteristics are presented in Figure 1.2.

The three service models, now well known, are Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS). NIST defines them thus:

1. *Software-as-a-Service* (*SaaS*). The capability provided to the consumer is to use the provider's applications running on a Cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser

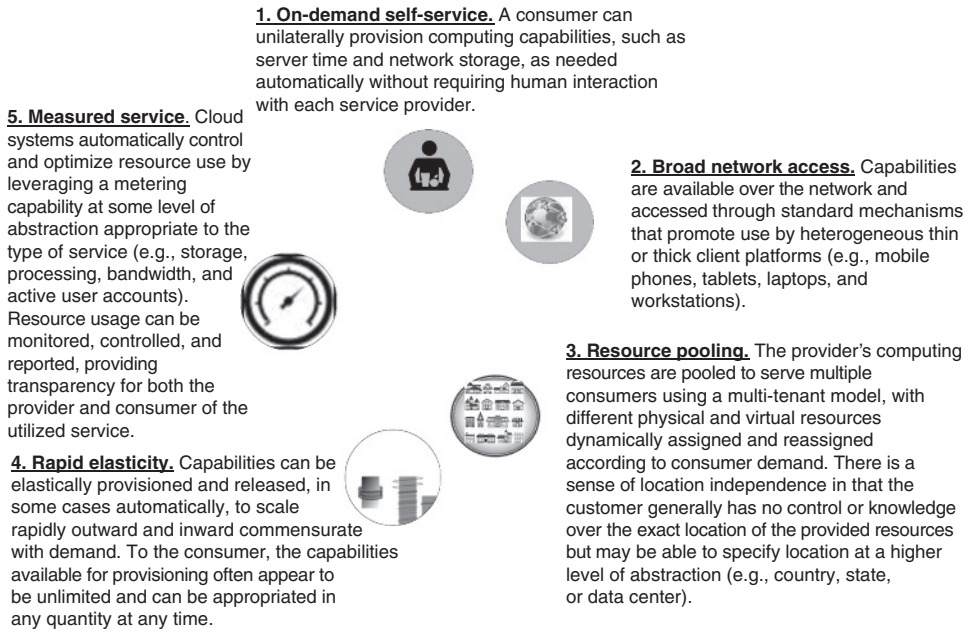


Figure 1.2 Essential characteristics of Cloud Computing. *Source:* NIST SP 800-145, p. 2.

(e.g., web-based e-mail), or a program interface. The consumer does not manage or control the underlying Cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

2. *Platform-as-a-Service* (*PaaS*). The capability provided to the consumer is to deploy onto the Cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying Cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.
3. *Infrastructure-as-a-Service* (*IaaS*). The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying Cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls).

Over time, other service models have appeared—more often than not in the marketing literature—but the authors of the well-known "Berkeley view of Cloud Computing" [3] chose to "eschew terminology such as 'X as a service (XaaS),'" citing the difficulty of agreeing "even among ourselves what the precise differences among them might be," that is, among the services for some values of X . . .

Finally, the four Cloud deployment models are defined by NIST as follows:

1. *Private Cloud.* The Cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.
2. *Community Cloud*. The Cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.
3. *Public Cloud*. The Cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the Cloud provider.
4. *Hybrid Cloud*. The Cloud infrastructure is a composition of two or more distinct Cloud infrastructures (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).

Cloud Computing is not a single technology. It is better described as a business development, whose realization has been enabled by several disciplines: computer architecture, operating systems, data communications, and network and operations management. As we will see, the latter discipline has been around for as long as networking, but the introduction of Cloud Computing has naturally fueled its growth in a new direction, once again validating the quote from Norbert Wiener's book that we chose as the epigraph to this book.

As Chapter 2 demonstrates, Cloud Computing has had a revolutionary effect on the information technology industry, reverberating through the telecommunications industry, which followed suit. Telecommunications providers demanded that vendors provide software only, rather than "the boxes." There have been several relevant standardization efforts in the industry, and—perhaps more important—there have been open-source software packages for building Cloud environments.

Naturally, standardization was preceded by a significant effort in research and development. In 2011, an author[3] of this book established the *CloudBand* product unit within Alcatel-Lucent, where, with the help of Bell Labs research, the telecommunications Cloud platform has been developed. It was in the context of *CloudBand* that we three authors met and the idea of this book was born.

We planned the book first of all as a textbook on Cloud Computing. Our experience in developing and teaching a graduate course on the subject at the Stevens Institute of Technology taught us that even the brightest and best-prepared students were missing sufficient knowledge in Central Processing Unit (CPU) virtualization (a subject that is rarely taught in the context of computer architecture or operating systems), as well as a number of specific points in data communications. Network and operations management has rarely been part of the modern computer science curriculum.

[3] Dor Skuler, at the time Alcatel-Lucent Vice President and General Manager of the *CloudBand* product unit.

In fact, the same knowledge gap seems to be ubiquitous in the industry, where engineers are forced to specialize, and we hope that this book will help fill the gap by providing an overarching multi-disciplinary foundation.

The rest of the book is structured as follows:

- Chapter 2 is mainly about "what" rather than "how." It provides definitions, describes business considerations—with a special case study of *Network Function Virtualization*—and otherwise provides a bird's eye view of Cloud Computing. The "how" is the subject of the chapters that follow.
- Chapter 3 explains the tenets of CPU virtualization.
- Chapter 4 is dedicated to networking—the nervous system of the Cloud.
- Chapter 5 describes network appliances, the building blocks of Cloud data centers as well as private networks.
- Chapter 6 describes the overall structure of the modern data center, along with its components.
- Chapter 7 reviews operations and management in the Cloud and elucidates the concepts of orchestration and identity and access management, with the case study of *OpenStack*—a popular open-source Cloud project.
- The Appendix delves into the detail of selected topics discussed earlier.

The references (which also form a bibliography on the respective subjects) are placed separately in individual chapters.

Having presented an outline of the book, we should note that there are three essential subjects that do not have a dedicated chapter. Instead, they are addressed in each chapter inasmuch as they concern that chapter's subject matter.

One such subject is security. Needless to say, this is the single most important matter that could make or break Cloud Computing. There are many aspects to security, and so we felt that we should address the aspects relevant to each chapter within the chapter itself.

Another subject that has no "central" coverage is standardization. Again, we introduce the relevant standards and open-source projects while discussing specific technical subjects. The third subject is history. It is well known in engineering that many existing technical solutions are not around because they are optimal, but because of their historical development. In teaching a discipline it is important to point these out, and we have tried our best to do so, again in the context of each technology that we address.

References

[1] Parkhill, D.F. (1966) *Challenge of the Computer Utility*. Addison-Wesley, Reading, MA.

[2] Mell, P. and Grance, T. (2011). Special Publication 800-145: The NIST Definition of Cloud Computing. Recommendations of the National Institute of Standards and Technology. US Department of Commerce, Gaithersburg, MD, September, 2011.

[3] Armbrust, M., Fox, A., Griffith, R., *et al.* (2009) Above the Clouds: A Berkeley view of Cloud Computing. Electrical Engineering and Computer Sciences Technical Report No. UCB/EECS-2009-2A, University of California at Berkeley, Berkeley, CA, February, 2009.