1.3 Evolution of IT leading to cloud computing

Cloud computing didn't sprout fully formed from the technology ether in 2005. Its technological underpinnings developed over the course of the last 40 or so years. The technological process was evolutionary, across several disparate areas. But these advances, aggregated into a bundle, represent a revolutionary change in the way IT will be conducted in the future.

Gillett and Kapor made the first known reference to cloud computing in 1996 in an MIT paper (http://ccs.mit.edu/papers/CCSWP197/CCSWP197.html). Today's common understanding of cloud computing retains the original intent. It was a mere decade later when a real-world instantiation of the cloud came into existence as Amazon repurposed its latent e-commerce resources and went into the business of providing cloud services. From there, it was only a matter of a few months until the term became commonplace in our collective consciousness and, as figure 1.3 shows, in our Google search requests (they're the same thing in today's world, right?).

1.3.1 Origin of the "cloud" metaphor

One common question people ask is, "Where did the term *cloud* come from?" The answer is that for over a decade, whenever people drew pictures of application architectures that involved the internet, they inevitably represented the internet with a cloud, as shown in figure 1.4.

The cloud in the diagram is meant to convey that anonymous people are sitting at browsers accessing the internet, and somehow their browser visits a site and begins to

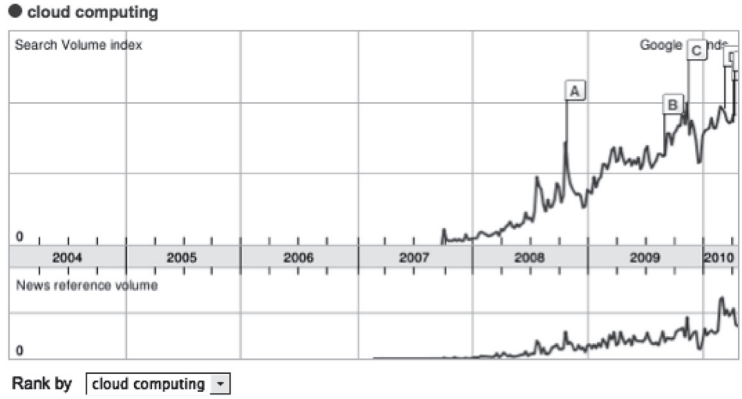

Figure 1.3 Cloud computing as a concept entered our collective consciousness in mid-2007. This figure shows the rapid rise in popularity of the search term *cloud computing* as measured by Google. The labels correspond to major cloud announcements. A: Microsoft announces it will rent cloud computing space; B: *Philadelphia Inquirer* reports, "Microsoft's cloud computing system grow is growing up"; C: *Winnipeg Free Press* reports, "Google looks to be cloud-computing rainmaker." Source: Google Trends (www.google.com/trends), on the term *cloud computing*.

access its infrastructure and applications. From "somewhere out there" you get visitors who can become users who may buy products or services from you. Unlike internal customers to whom you may provide IT applications and services, this constituency exists "somewhere else," outside of your firewall, and hence outside of your domain of control. The image of a cloud is merely a way to represent this vast potential base of anonymous users coming from the internet.

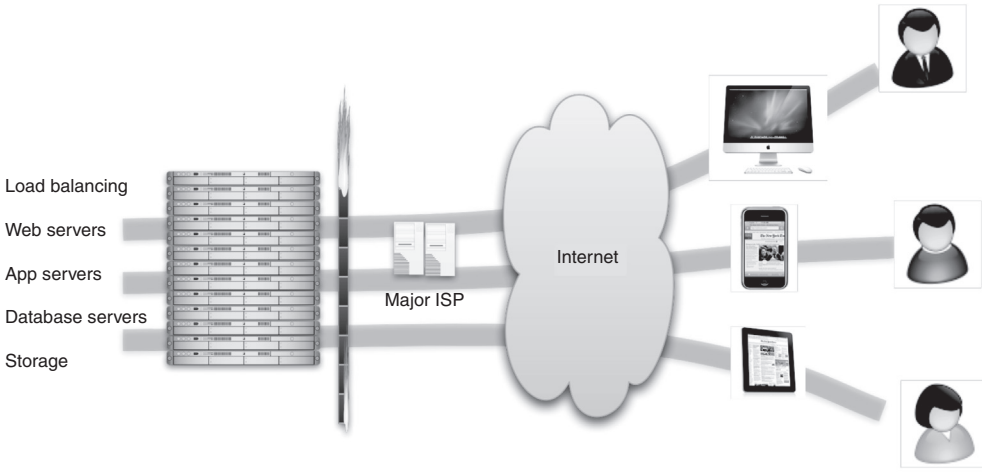

Figure 1.4 A picture of a cloud is a ubiquitous representation of the internet and is used almost universally in discussions or drawings of computer architecture.

Those users must log in from a PC to access the internet. Technically, each one needs an Internet Service Provider (ISP) that may be a telecom company, their employer, or a dedicated internet access company (such as AOL). Each ISP needs a bank of machines that people can access and that in turn has access to the internet.

Simply put, the earliest concept of the cloud consisted of large aggregations of computers with access to the internet, accessed by people through their browsers. The concept has remained surprisingly true to that early vision but has evolved and matured in important ways. We'll explore those ways in detail in this book.

1.3.2 Major computing paradigm shifts: mainframes to client-server to web

In the 1960s, we saw the development of the first commercial mainframes. In the beginning, these were single-user systems, but they evolved in the 1970s to systems that were time-shared. In this model, the large computing resource was *virtualized*, and a virtual machine was allocated to individual users who were sharing the system (but to each, it seemed that they had an entire dedicated machine).

Virtual instances were accessed in a thin-client model by green-screen terminals. This mode of access can be seen as a direct analog of the concept of virtualized instances in the cloud, although then a single machine was divided among users. In the cloud, it's potentially many thousands of machines. The scarcity of the computing resource in the past drove the virtualization of that resource so that it could be shared, whereas now, the desire to fully utilize physical compute resources is driving cloud virtualization.

As we evolved and entered the client-server era, the primacy of the mainframe as the computing center of the universe dissolved. As computing power increased, work gradually shifted away from centralized computing resources toward increasingly powerful distributed systems. In the era of the PC-based desktop applications, this shift was nearly complete: computing resources for many everyday computing tasks moved to the desktop and became thick client applications (such as Microsoft Office). The mainframe retained its primacy only for corporate or department-wide applications, relegating it to this role alone.

The standardization of networking technology simplified the ability to connect systems as TCP/IP became the protocol of the burgeoning internet in the 1980s. The ascendancy of the web and HTTP in the late 1990s swung the pendulum back to a world where the thin-client model reigned supreme. The world was now positioned to move into the era of *cloud computing*. The biggest stages of the evolution of IT are diagrammed vertically in a timeline in figure 1.5.

The computing evolution we are still in the midst of has had many stages. Platform shifts like mainframe to client-server and then client-server to web were one dimension of the evolution. One that may be less apparent but that is having as profound an impact is the evolution of the data center and how physical computing resources are housed, powered, maintained, and upgraded.

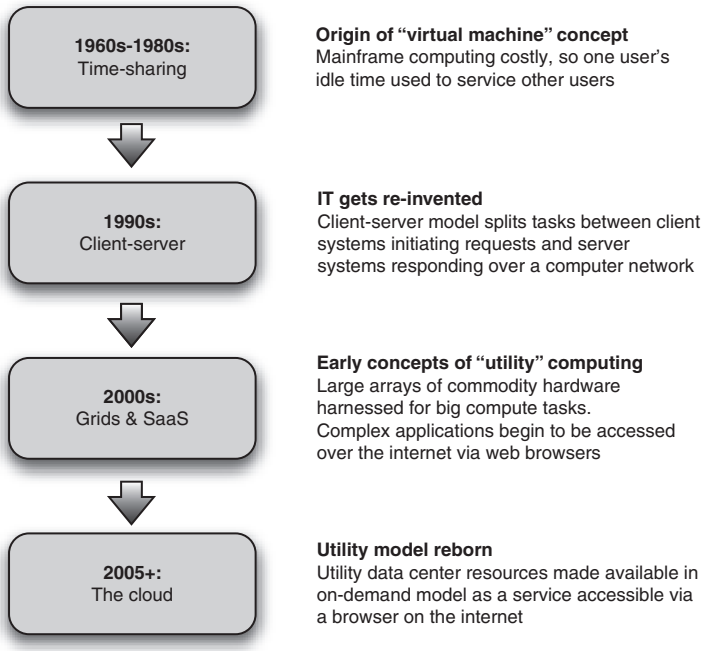

Figure 1.5 Cloud computing is best understood as an evolutionary change. The key elements and concepts of cloud computing emerged gradually over several decades through the various predominant computing paradigms.

1.3.3 Housing of physical computing resources: data center evolution

Over the past four decades, there have been tremendous changes in hardware capabilities, specifically in computing power and storage. The ability to quickly process prodigious amounts of data on inexpensive and mass-produced commodity servers means that a few inexpensive racks of servers can handle problems that were tackled on NSA-sized budgets as recently as the early 1990s.

One measure of the progress in computational power is the cost in Floating Point Operations Per Second, or FLOPS. FLOPS are simple mathematical operations (such as addition, multiplication, and division) that can be performed in a single operation by a computer. Comparing the number of operations that two computers can perform in one second allows for a rough measure of their computational strength. In 1976, the state-of-the-art Cray-1 was capable of delivering roughly 150 million FLOPS (megaFLOPS) at the price point of $5 million, or over $33,000/MegaFLOPS. A typical quad-core-processor-based PC today can be purchased for under $1,000 and can perform 50 GigaFLOPS (billion FLOPS), which comes out to about $0.02/MegaFLOPS.

Similarly, the cost of storage has decreased dramatically over the last few decades as the capacity to store data has kept pace with the ability to produce terabytes of digital content in the form of high-definition HD video and high-resolution imagery. In the

early 1980s, disk space costs exceeded $200/MB; today, this cost has come down to under $0.01/MB.

Network technologies have advanced as well, with modern bandwidth rates in the 100–1000 Gbps range commonplace in data centers today. As for WAN, the turn of the millennium saw a massive build-out of dark fiber, bringing high-speed broadband to most urban areas. More rural areas have satellite coverage, and on-the-go, high-speed wireless networks mean almost ubiquitous broadband connectivity to the grid.

To support the cloud, a huge data-center build-out is now underway. Google, Microsoft, Yahoo!, Expedia, Amazon, and others are deploying massive data centers. These are the engine rooms that power the cloud, and they now account for more than 1.2 percent of the U.S.'s total electricity usage (including cooling and auxiliaries),[2] which doubled over the period from 2000 to 2005. We'll present the economies of scale and much more detail about how these mega data centers are shaping up in chapter 2.

1.3.4 Software componentization and remote access: SOA, virtualization, and SaaS

On the software side of the cloud evolution are three important threads of development: virtualization, SOA, and SaaS. Two of these are technological, and the third relates to the business model.

The first important thread is virtualization. As discussed previously, virtualization isn't a new concept, and it existed in mainframe environments. The new innovation that took place in the late 1990s was the extension of this idea to commodity hardware. Virtualization as pioneered by VMware and others took advantage of the capacity of modern multicore CPUs and made it possible to partition and time-slice the operation of commodity servers. Large server farms based on these commodity servers were partitioned for use across large populations of users.

SOA is the second software concept necessary for cloud computing. We see SOA as the logical extension of browser-based standardization applied to machine-to-machine communication. Things that humans did through browsers that interacted with a web server are now done machine-to-machine using the same web-based standard protocols and are called *SOA*. SOA makes practical the componentization and composition of services into applications, and hence it can serve as the architectural model for building composite applications running on multiple virtualized instances.

The final software evolution we consider most pertinent to the cloud is SaaS. Instead of being a technological innovation, this is a business model innovation. Historically, enterprise software was sold predominantly in a perpetual license model. In this model, a customer purchased the right to use a certain software application in perpetuity for a fixed, and in many cases high, price. In subsequent years, they paid for support and maintenance at typically around 18 percent of the original price. This entitled the

[2] Jonathan G. Koomey, Ph.D. (www.koomey.com), Lawrence Berkeley National Laboratory & Stanford University.

customer to upgrades of the software and help when they ran into difficulty. In the SaaS model, you don't purchase the software—you rent it. Typically, the fee scales with the amount of use, so the value derived from the software is proportional to the amount spent on it. The customer buys access to the software for a specified term, which may be days, weeks, months, or years, and can elect to stop paying when they no longer need the SaaS offering. Cloud computing service providers have adopted this *pay-as-you-go* or *on-demand* model.

This brings up an important point we need to consider next. SaaS is one flavor or layer in a stack of cloud types. A common mistake people make in these early days of the cloud is to make an apples-to-oranges comparison of one type of cloud to another. To avoid that, the next section will classify the different layers in the cloud stack and how they compare and contrast.