

---

# Introduction

---

The purpose of this book is to clear up some of the mystery surrounding the topic of cloud computing. In order to understand how computing has evolved, one must understand the evolution of computing from a historical perspective, focusing primarily on those advances that led to the development of cloud computing, such as the transition from mainframes to desktops, laptops, mobile devices, and on to the cloud. We will also need to discuss in some detail the key components that are critical to make the cloud computing paradigm feasible with the technology available today. We will cover some of the standards that are used or are proposed for use in the cloud computing model, since standardization is crucial to achieving widespread acceptance of cloud computing. We will also discuss the means used to manage effectively the infrastructure for cloud computing. Significant legal considerations in properly protecting user data and mitigating corporate liability will also be covered. Finally, we will discuss what some of the more successful cloud vendors have done and how their achievements have helped the cloud model evolve.

Over the last five decades, businesses that use computing resources have learned to contend with a vast array of buzzwords. Much of this *geek-speak* or marketing vapor, over time, has been guilty of making promises that often are never kept. Some promises, to be sure, have been delivered, although others have drifted into oblivion. When it comes to offering technology in a *pay-as-you-use* services model, most information technology (IT) professionals have heard it all—from allocated resource management to grid computing, to on-demand computing and software-as-a-service (SaaS), to utility computing. A new buzzword, *cloud computing*, is presently in vogue in the marketplace, and it is generating all sorts of confusion about what it actually represents.

## What Is the Cloud?

The term *cloud* has been used historically as a metaphor for the Internet. This usage was originally derived from its common depiction in network diagrams as an outline of a cloud, used to represent the transport of data across carrier backbones (which owned the cloud) to an endpoint location on the other side of the cloud. This concept dates back as early as 1961, when Professor John McCarthy suggested that computer time-sharing technology might lead to a future where computing power and even specific applications might be sold through a utility-type business model.<sup>1</sup> This idea became very popular in the late 1960s, but by the mid-1970s the idea faded away when it became clear that the IT-related technologies of the day were unable to sustain such a futuristic computing model. However, since the turn of the millennium, the concept has been revitalized. It was during this time of revitalization that the term *cloud computing* began to emerge in technology circles.

## The Emergence of Cloud Computing

*Utility computing* can be defined as the provision of computational and storage resources as a metered service, similar to those provided by a traditional public utility company. This, of course, is not a new idea. This form of computing is growing in popularity, however, as companies have begun to extend the model to a cloud computing paradigm providing virtual servers that IT departments and users can access on demand. Early enterprise adopters used utility computing mainly for non-mission-critical needs, but that is quickly changing as trust and reliability issues are resolved.

Some people think cloud computing is the next big thing in the world of IT. Others believe it is just another variation of the utility computing model that has been repackaged in this decade as something new and cool. However, it is not just the buzzword “cloud computing” that is causing confusion among the masses. Currently, with so few cloud computing vendors actually practicing this form of technology and also almost every analyst from every research organization in the country defining the term differently, the meaning of the term has become very nebulous. Even among those who think they understand it, definitions vary, and most of those definitions are hazy at best. To clear the haze and make some sense of the new

---

1. <http://computinginthecloud.wordpress.com/2008/09/25/utility-cloud-computingflashback-to-1961-prof-john-mccarthy>, retrieved 5 Jan 2009.

concept, this book will attempt to help you understand just what cloud computing really means, how disruptive to your business it may become in the future, and what its advantages and disadvantages are.

As we said previously, the term *the cloud* is often used as a metaphor for the Internet and has become a familiar cliché. However, when “the cloud” is combined with “computing,” it causes a lot of confusion. Market research analysts and technology vendors alike tend to define cloud computing very narrowly, as a new type of utility computing that basically uses virtual servers that have been made available to third parties via the Internet. Others tend to define the term using a very broad, all-encompassing application of the virtual computing platform. They contend that anything beyond the firewall perimeter is in the cloud. A more tempered view of cloud computing considers it the delivery of computational resources from a location other than the one from which you are computing.

## The Global Nature of the Cloud

The cloud sees no borders and thus has made the world a much smaller place. The Internet is global in scope but respects only established communication paths. People from everywhere now have access to other people from anywhere else. Globalization of computing assets may be the biggest contribution the cloud has made to date. For this reason, the cloud is the subject of many complex geopolitical issues. Cloud vendors must satisfy myriad regulatory concerns in order to deliver cloud services to a global market. When the Internet was in its infancy, many people believed cyberspace was a distinct environment that needed laws specific to itself. University computing centers and the ARPANET were, for a time, the encapsulated environments where the Internet existed. It took a while to get business to warm up to the idea.

Cloud computing is still in its infancy. There is a hodge-podge of providers, both large and small, delivering a wide variety of cloud-based services. For example, there are full-blown applications, support services, mail-filtering services, storage services, etc. IT practitioners have learned to contend with some of the many cloud-based services out of necessity as business needs dictated. However, cloud computing aggregators and integrators are already emerging, offering packages of products and services as a single entry point into the cloud.

The concept of cloud computing becomes much more understandable when one begins to think about what modern IT environments always

require—the means to increase capacity or add capabilities to their infrastructure dynamically, without investing money in the purchase of new infrastructure, all the while without needing to conduct training for new personnel and without the need for licensing new software. Given a solution to the aforementioned needs, cloud computing models that encompass a subscription-based or pay-per-use paradigm provide a service that can be used over the Internet and extends an IT shop’s existing capabilities. Many users have found that this approach provides a return on investment that IT managers are more than willing to accept.

## Cloud-Based Service Offerings

Cloud computing may be viewed as a resource available as a service for virtual data centers, but cloud computing and virtual data centers are not the same. For example, consider Amazon’s S3 Storage Service. This is a data storage service designed for use across the Internet (i.e., the cloud). It is designed to make web-scale computing easier for developers. According to Amazon:

Amazon S3 provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web. It gives any developer access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. The service aims to maximize benefits of scale and to pass those benefits on to developers.<sup>2</sup>

Amazon.com has played a vital role in the development of cloud computing. In modernizing its data centers after the dot-com bubble burst in 2001, it discovered that the new cloud architecture it had implemented resulted in some very significant internal efficiency improvements. By providing access to its systems for third-party users on a utility computing basis, via Amazon Web Services, introduced in 2002, a revolution of sorts began. Amazon Web Services began implementing its model by renting computing cycles as a service outside a given user’s domain, wherever on the planet that domain might be located. This approach modernized a style of computing whereby IT-related capabilities could be provided “as a service”

---

2. <http://aws.amazon.com/s3>, retrieved 5 Jan 2009.

to users. By allowing their users to access technology-enabled services “in the cloud,” without any need for knowledge of, expertise with, or control over how the technology infrastructure that supports those services worked, Amazon shifted the approach to computing radically. This approach transformed cloud computing into a paradigm whereby data is permanently stored in remote servers accessible via the Internet and cached temporarily on client devices that may include desktops, tablet computers, notebooks, hand-held devices, mobile phones, etc., and is often called *Software as a Service* (SaaS).

SaaS is a type of cloud computing that delivers applications through a browser to thousands of customers using a multiuser architecture. The focus for SaaS is on the end user as opposed to managed services (described below). For the customer, there are no up-front investment costs in servers or software licensing. For the service provider, with just one product to maintain, costs are relatively low compared to the costs incurred with a conventional hosting model. Salesforce.com<sup>3</sup> is by far the best-known example of SaaS computing among enterprise applications. Salesforce.com was founded in 1999 by former Oracle executive Marc Benioff, who pioneered the concept of delivering enterprise applications via a simple web site. Nowadays, SaaS is also commonly used for enterprise resource planning and human resource applications. Another example is Google Apps, which provides online access via a web browser to the most common office and business applications used today, all the while keeping the software and user data stored on Google servers. A decade ago, no one could have predicted the sudden rise of SaaS applications such as these.

Managed service providers (MSPs) offer one of the oldest forms of cloud computing. Basically, a managed service is an application that is accessible to an organization’s IT infrastructure rather than to end users. Services include virus scanning for email, antispam services such as Postini,<sup>4</sup> desktop management services such as those offered by CenterBeam<sup>5</sup> or Everdream,<sup>6</sup> and

---

3. <http://www.salesforce.com>, retrieved 5 Jan 2009.

4. In September 2007, Google acquired Postini, recognized as a global leader in on-demand communications security and compliance solutions. This is further evidence of the aggregation of cloud service providers.

5. CenterBeam delivers services over the Internet using a SaaS model.

6. In November 2007, Dell signed an agreement to acquire Everdream, a leading provider of SaaS solutions for remote service management. The planned acquisition was a key component in Dell’s strategy of enabling customers to simplify IT. Everdream’s capabilities complement those provided by the recently acquired SilverBack Technologies, further enabling end-to-end remote management of customers’ IT environments.

application performance monitoring. Managed security services that are delivered by third-party providers also fall into this category.

Platform-as-a-Service (PaaS) is yet another variation of SaaS. Sometimes referred to simply as web services in the cloud, PaaS is closely related to SaaS but delivers a platform from which to work rather than an application to work with. These service providers offer application programming interfaces (APIs) that enable developers to exploit functionality over the Internet, rather than delivering full-blown applications. This variation of cloud computing delivers development environments to programmers, analysts, and software engineers as a service. A general model is implemented under which developers build applications designed to run on the provider's infrastructure and which are delivered to users in via an Internet browser. The main drawback to this approach is that these services are limited by the vendor's design and capabilities. This means a compromise between freedom to develop code that does something other than what the provider can provide and application predictability, performance, and integration.

An example of this model is the Google App Engine. According to Google, "Google App Engine makes it easy to build an application that runs reliably, even under heavy load and with large amounts of data."<sup>7</sup> The Google App Engine environment includes the following features

- Dynamic web serving, with full support for common web technologies
- Persistent storage with queries, sorting, and transactions
- Automatic scaling and load balancing
- APIs for authenticating users and sending email using Google Accounts
- A fully featured local development environment that simulates Google App Engine on your computer

Currently, Google App Engine applications are implemented using the Python programming language. The runtime environment includes the full Python language and most of the Python standard library. For extremely lightweight development, cloud-based mashup platforms (Ajax modules that are assembled in code) abound, such as Yahoo Pipes or Dapper.net.

---

7. <http://code.google.com/appengine/docs/whatisgoogleappengine.html>, retrieved 5 Jan 2009.

## Grid Computing or Cloud Computing?

Grid computing is often confused with cloud computing. Grid computing is a form of distributed computing that implements a *virtual supercomputer* made up of a cluster of networked or Internetworked computers acting in unison to perform very large tasks. Many cloud computing deployments today are powered by grid computing implementations and are billed like utilities, but cloud computing can and should be seen as an evolved next step away from the grid utility model. There is an ever-growing list of providers that have successfully used cloud architectures with little or no centralized infrastructure or billing systems, such as the peer-to-peer network BitTorrent and the volunteer computing initiative SETI@home.<sup>8</sup>

Service commerce platforms are yet another variation of SaaS and MSPs. This type of cloud computing service provides a centralized service hub that users interact with. Currently, the most often used application of this platform is found in financial trading environments or systems that allow users to order things such as travel or personal services from a common platform (e.g., Expedia.com or Hotels.com), which then coordinates pricing and service delivery within the specifications set by the user.

## Is the Cloud Model Reliable?

The majority of today's cloud computing infrastructure consists of time-tested and highly reliable services built on servers with varying levels of virtualized technologies, which are delivered via large data centers operating under service-level agreements that require 99.99% or better uptime. Commercial offerings have evolved to meet the quality-of-service requirements of customers and typically offer such service-level agreements to their customers. From users' perspective, the cloud appears as a single point of access for all their computing needs. These cloud-based services are accessible anywhere in the world, as long as an Internet connection is available. Open standards and open-source software have also been significant factors in the growth of cloud computing, topics we will discuss in more depth later.

---

8. SETI@home is a scientific experiment that uses Internet-connected computers in the Search for Extraterrestrial Intelligence (SETI). For more information, see <http://www.seti.org>.

## Benefits of Using a Cloud Model

Because customers generally do not own the infrastructure used in cloud computing environments, they can forgo capital expenditure and consume resources as a service by just paying for what they use. Many cloud computing offerings have adopted the utility computing and billing model described above, while others bill on a subscription basis. By sharing computing power among multiple users, utilization rates are generally greatly improved, because cloud computing servers are not sitting dormant for lack of use. This factor alone can reduce infrastructure costs significantly and accelerate the speed of applications development.

A beneficial side effect of using this model is that computer capacity increases dramatically, since customers do not have to engineer their applications for peak times, when processing loads are greatest. Adoption of the cloud computing model has also been enabled because of the greater availability of increased high-speed bandwidth. With greater enablement, though, there are other issues one must consider, especially legal ones.

## What About Legal Issues When Using Cloud Models?

Recently there have been some efforts to create and unify the legal environment specific to the cloud. For example, the United States–European Union Safe Harbor Act provides a seven-point framework of requirements for U.S. companies that may use data from other parts of the world, namely, the European Union. This framework sets forth how companies can participate and certify their compliance and is defined in detail on the U.S. Department of Commerce and Federal Trade Commission web sites. In summary, the agreement allows most U.S. corporations to certify that they have joined a self-regulatory organization that adheres to the following seven Safe Harbor Principles or has implemented its own privacy policies that conform with these principles:

1. Notify individuals about the purposes for which information is collected and used.
2. Give individuals the choice of whether their information can be disclosed to a third party.
3. Ensure that if it transfers personal information to a third party, that third party also provides the same level of privacy protection.



4. Allow individuals access to their personal information.
5. Take reasonable security precautions to protect collected data from loss, misuse, or disclosure.
6. Take reasonable steps to ensure the integrity of the data collected.;
7. Have in place an adequate enforcement mechanism.

Major service providers such as Amazon Web Services cater to a global marketplace, typically the United States, Japan, and the European Union, by deploying local infrastructure at those locales and allowing customers to select availability zones. However, there are still concerns about security and privacy at both the individual and governmental levels. Of major concern is the USA PATRIOT Act and the Electronic Communications Privacy Act's Stored Communications Act. The USA PATRIOT Act, more commonly known as the Patriot Act, is a controversial Act of Congress that U.S. President George W. Bush signed into law on October 26, 2001. The contrived acronym stands for "Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001" (Public Law P.L. 107-56). The Act expanded the definition of terrorism to include domestic terrorism, thus enlarging the number of activities to which the USA PATRIOT Act's law enforcement powers could be applied. It increased law enforcement agencies' ability to surveil telephone, email communications, medical, financial, and other records and increased the range of discretion for law enforcement and immigration authorities when detaining and deporting immigrants suspected of terrorism-related acts. It lessened the restrictions on foreign intelligence gathering within the United States. Furthermore, it expanded the Secretary of the Treasury's authority to regulate financial transactions involving foreign individuals and businesses.

The Electronic Communications Privacy Act's Stored Communications Act is defined in the U.S. Code, Title 18, Part I, Chapter 121, § 2701, Unlawful Access to Stored Communications. Offenses committed under this act include intentional access without authorization to a facility through which an electronic communication service is provided or intentionally exceeding an authorization to access that facility in order to obtain, alter, or prevent authorized access to a wire or electronic communication while it is in electronic storage in such a system. Persons convicted under

this Act can be punished if the offense is committed for purposes of commercial advantage, malicious destruction or damage, or private commercial gain, or in furtherance of any criminal or tortious act in violation of the Constitution or laws of the United States or any state by a fine or imprisonment or both for not more than five years in the case of a first offense. For a second or subsequent offense, the penalties stiffen to fine or imprisonment for not more than 10 years, or both.

## What Are the Key Characteristics of Cloud Computing?

There are several key characteristics of a cloud computing environment. Service offerings are most often made available to specific consumers and small businesses that see the benefit of use because their capital expenditure is minimized. This serves to lower barriers to entry in the marketplace, since the infrastructure used to provide these offerings is owned by the cloud service provider and need not be purchased by the customer. Because users are not tied to a specific device (they need only the ability to access the Internet) and because the Internet allows for location independence, use of the cloud enables cloud computing service providers' customers to access cloud-enabled systems regardless of where they may be located or what device they choose to use.

Multitenancy<sup>9</sup> enables sharing of resources and costs among a large pool of users. Chief benefits to a multitenancy approach include:

- Centralization of infrastructure and lower costs
- Increased peak-load capacity
- Efficiency improvements for systems that are often underutilized
- Dynamic allocation of CPU, storage, and network bandwidth
- Consistent performance that is monitored by the provider of the service

Reliability is often enhanced in cloud computing environments because service providers utilize multiple redundant sites. This is attractive to enter-

---

9. <http://en.wikipedia.org/wiki/Multitenancy>, retrieved 5 Jan 2009. Multitenancy refers to a principle in software architecture where a single instance of the software runs on a SaaS vendor's servers, serving multiple client organizations (tenants).

prises for business continuity and disaster recovery reasons. The drawback, however, is that IT managers can do very little when an outage occurs.

Another benefit that makes cloud services more reliable is that scalability can vary dynamically based on changing user demands. Because the service provider manages the necessary infrastructure, security often is vastly improved. As a result of data centralization, there is an increased focus on protecting customer resources maintained by the service provider. To assure customers that their data is safe, cloud providers are quick to invest in dedicated security staff. This is largely seen as beneficial but has also raised concerns about a user's loss of control over sensitive data. Access to data is usually logged, but accessing the audit logs can be difficult or even impossible for the customer.

Data centers, computers, and the entire associated infrastructure needed to support cloud computing are major consumers of energy. Sustainability of the cloud computing model is achieved by leveraging improvements in resource utilization and implementation of more energy-efficient systems. In 2007, Google, IBM, and a number of universities began working on a large-scale cloud computing research project. By the summer of 2008, quite a few cloud computing events had been scheduled. The first annual conference on cloud computing was scheduled to be hosted online April 20–24, 2009. According to the official web site:

This conference is the world's premier cloud computing event, covering research, development and innovations in the world of cloud computing. The program reflects the highest level of accomplishments in the cloud computing community, while the invited presentations feature an exceptional lineup of speakers. The panels, workshops, and tutorials are selected to cover a range of the hottest topics in cloud computing.<sup>10</sup>

It may seem that all the world is raving about the potential of the cloud computing model, but most business leaders are likely asking: "What is the market opportunity for this technology and what is the future potential for long-term utilization of it?" Meaningful research and data are difficult to find at this point, but the potential uses for cloud computing models are wide. Ultimately, cloud computing is likely to bring supercomputing capa-

---

10. <http://cloudslam09.com>, retrieved 5 Jan 09.

bilities to the masses. Yahoo, Google, Microsoft, IBM, and others are engaged in the creation of online services to give their users even better access to data to aid in daily life issues such as health care, finance, insurance, etc.

## Challenges for the Cloud

The biggest challenges these companies face are secure data storage, high-speed access to the Internet, and standardization. Storing large amounts of data that is oriented around user privacy, identity, and application-specific preferences in centralized locations raises many concerns about data protection. These concerns, in turn, give rise to questions regarding the legal framework that should be implemented for a cloud-oriented environment. Another challenge to the cloud computing model is the fact that broadband penetration in the United States remains far behind that of many other countries in Europe and Asia. Cloud computing is untenable without high-speed connections (both wired and wireless). Unless broadband speeds are available, cloud computing services cannot be made widely accessible. Finally, technical standards used for implementation of the various computer systems and applications necessary to make cloud computing work have still not been completely defined, publicly reviewed, and ratified by an oversight body. Even the consortiums that are forming need to get past that hurdle at some point, and until that happens, progress on new products will likely move at a snail's pace.

Aside from the challenges discussed in the previous paragraph, the reliability of cloud computing has recently been a controversial topic in technology circles. Because of the public availability of a cloud environment, problems that occur in the cloud tend to receive lots of public exposure. Unlike problems that occur in enterprise environments, which often can be contained without publicity, even when only a few cloud computing users have problems, it makes headlines.

In October 2008, Google published an article online that discussed the lessons learned from hosting over a million business customers in the cloud computing model.<sup>11</sup> Google's personnel measure availability as the average uptime per user based on server-side error rates. They believe this reliability metric allows a true side-by-side comparison with other solutions. Their

---

11. Matthew Glotzbach, Product Management Director, Google Enterprise, "What We Learned from 1 Million Businesses in the Cloud," <http://googleblog.blogspot.com/2008/10/what-we-learned-from-1-million.html>, 30 Oct 2008.

measurements are made for every server request for every user, every moment of every day, and even a single millisecond delay is logged. Google analyzed data collected over the previous year and discovered that their Gmail application was available to everyone more than 99.9% of the time.

One might ask how a 99.9% reliability metric compares to conventional approaches used for business email. According to the research firm Radicati Group,<sup>12</sup> companies with on-premises email solutions averaged from 30 to 60 minutes of unscheduled downtime and an additional 36 to 90 minutes of planned downtime per month, compared to 10 to 15 minutes of downtime with Gmail. Based on analysis of these findings, Google claims that for unplanned outages, Gmail is twice as reliable as a Novell GroupWise solution and four times more reliable than a Microsoft Exchange-based solution, both of which require companies to maintain an internal infrastructure themselves. It stands to reason that higher reliability will translate to higher employee productivity. Google discovered that Gmail is more than four times as reliable as the Novell GroupWise solution and 10 times more reliable than an Exchange-based solution when you factor in planned outages inherent in on-premises messaging platforms.

Based on these findings, Google was confident enough to announce publicly in October 2008 that the 99.9% service-level agreement offered to their Premier Edition customers using Gmail would be extended to Google Calendar, Google Docs, Google Sites, and Google Talk. Since more than a million businesses use Google Apps to run their businesses, Google has made a series of commitments to improve communications with customers during any outages and to make all issues visible and transparent through open user groups. Since Google itself runs on its Google Apps platform, the commitment they have made has teeth, and I am a strong advocate of “eating your own dog food.” Google leads the industry in evolving the cloud computing model to become a part of what is being called Web 3.0—the next generation of Internet.<sup>13</sup>

In the following chapters, we will discuss the evolution of computing from a historical perspective, focusing primarily on those advances that led to the development of cloud computing. We will discuss in detail some of the more critical components that are necessary to make the cloud com-

---

12. The Radicati Group, 2008, “Corporate IT Survey—Messaging & Collaboration, 2008–2009,” <http://www.marketwatch.com/news/story/The-Radicati-Group-Releases-New-story.aspx?guid=%7B80D6388A-731C-457F-9156-F783B3E3C720%7D>, retrieved 12 Feb 2009.

13. [http://en.wikipedia.org/wiki/Web\\_3.0](http://en.wikipedia.org/wiki/Web_3.0), retrieved 5 Jan 2009.

puting paradigm feasible. Standardization is a crucial factor in gaining widespread adoption of the cloud computing model, and there are many different standards that need to be finalized before cloud computing becomes a mainstream method of computing for the masses. This book will look at those various standards based on the use and implementation issues surrounding cloud computing. Management of the infrastructure that is maintained by cloud computing service providers will also be discussed. As with any IT, there are legal considerations that must be addressed to properly protect user data and mitigate corporate liability, and we will cover some of the more significant legal issues and even some of the philosophical issues that will most likely not be resolved without adoption of a legal framework. Finally, this book will take a hard look at some of the cloud computing vendors that have had significant success and examine what they have done and how their achievements have helped to shape cloud computing.