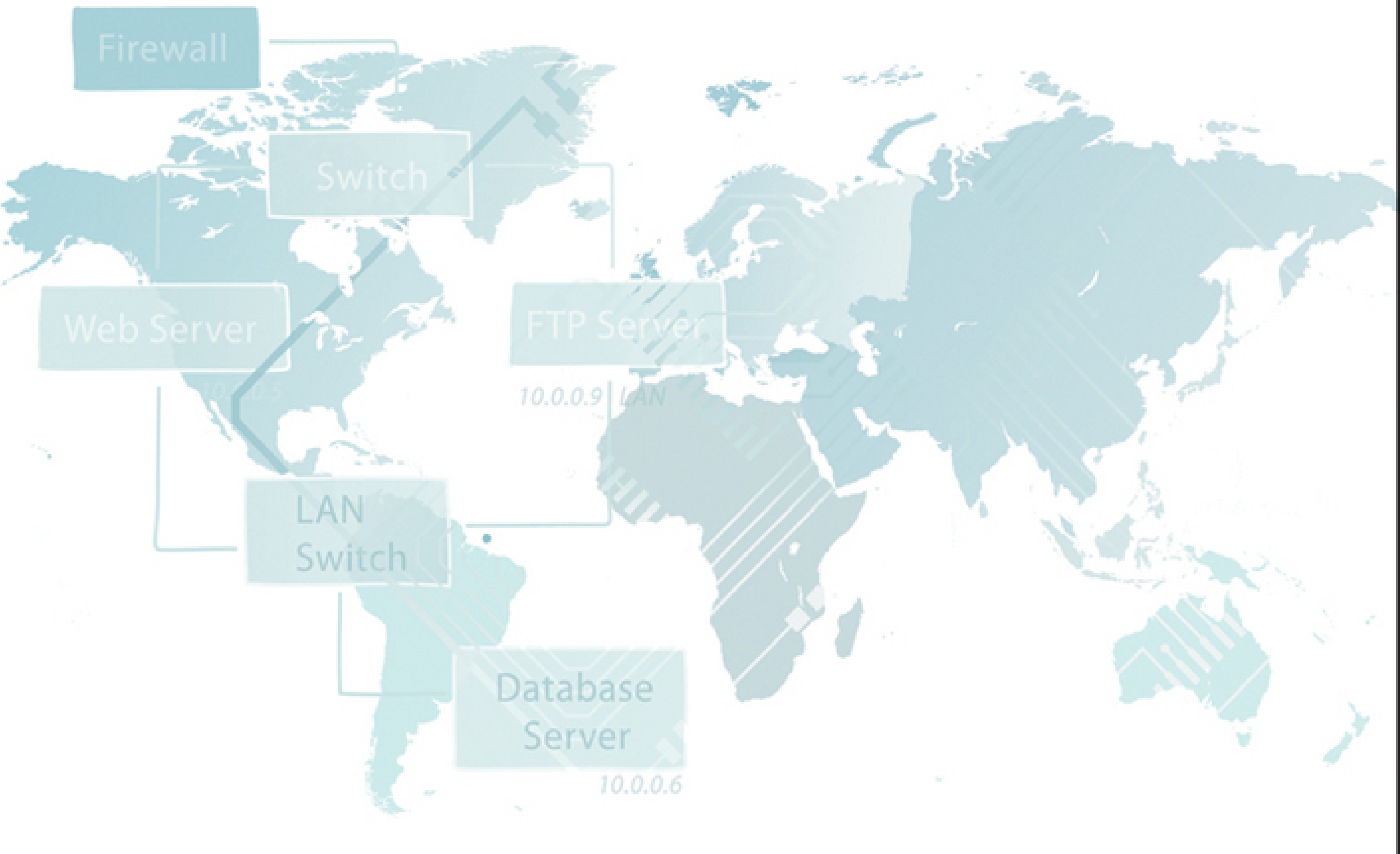
CLOUD SECURITY

A Comprehensive Guide to Secure Cloud Computing



Ronald L. Krutz and Russell Dean Vines

Foreword

Whenever we come upon something new, we try to understand it. A good way of understanding new things is to look for something from our experience that can serve as a metaphor. Sometimes this process works well, sometimes not.

Computer security has long labored under the metaphor of physical security. It stands to reason that we would assume that millennia of experience with keeping physical assets safe would serve us in keeping digital assets safe as well.

Much of our thinking in computer security has therefore been concerned with putting important things someplace “safe” and then controlling access to it. I distinctly recall a conversation with a security analyst at the beginning of the PC network era. When asked how to ensure the security of data on a PC, he said, “Simple. Put the data on the PC. Put the PC in a safe. Put the safe at the bottom of the ocean.”

We have been challenged over the years with coming up with safe places that allowed access. We have been challenged with even figuring out what “safe” might mean in a world where risks could come from anywhere, including inside our own organizations.

In today’s world, the physical security metaphor continues to deteriorate. We’ve all seen a movie or TV show where some critical piece of data becomes key to the plot. The location of the next terrorist attack is kept on a single USB that is subject to theft, deterioration, or any other number of physical ills designed to increase the drama. That is simply not the nature of data. Data is viral. Where did this data come from? It was never on a hard drive? No one ever emailed anybody about the attack? Can’t somebody plug the damn key in and make a YouTube video about it so that everyone can see it?

As we move to this new era of cloud computing, the last vestiges of our physical world metaphors are swept way. We need to understand data access and validation in a new way — perhaps in the way they should have been understood all along. Data security needs to be understood as something new, requiring new and innovative solutions.

Security professionals are perhaps rightfully overwhelmed by this challenge. Despite increased spending, the average firm finds itself less secure than it was five years ago. Advancements in security tools and techniques have not kept pace with risks and attack vectors. How can the security community respond to these ever-increasing threats when the additional requirements of virtualization and agility drive data assets up into a nebulous “cloud”?

One thing we do know for sure: Security will not drive or control this change. Any business requirement for lower costs and increased agility of cloud computing will eventually rule the day. Security professionals have attempted to slow the growth of several technology initiatives over the years in an attempt to control the risks. E-mail, instant messaging, and web browsing are some that come to mind immediately. We know from past experience, however, that implementing appropriate controls generally works far better than attempting to simply stop these initiatives.

As security professionals, it is incumbent on us to generate innovations in our concepts of data security and integrity. We need tools and processes that recognize the ephemeral nature of data and the reality that physical locational controls simply will not work going forward. With a little hard work, we can achieve security models that minimize risk and enable this new method of computing. We don’t need to give up on security; we simply need to abandon some of our metaphors.

This book serves as a guide for doing just that. As security professionals, we may not want to embrace the cloud, but we're certainly going to have to learn to live with it.

Ken Phelan
CTO Gotham Technology Group

Introduction

Cloud computing provides the capability to use computing and storage resources on a metered basis and reduce the investments in an organization’s computing infrastructure. The spawning and deletion of virtual machines running on physical hardware and being controlled by hypervisors is a cost-efficient and flexible computing paradigm.

In addition, the integration and widespread availability of large amounts of “sanitized’ information such as health care records can be of tremendous benefit to researchers and practitioners.

However, as with any technology, the full potential of the cloud cannot be achieved without understanding its capabilities, vulnerabilities, advantages, and trade-offs. This text provides insight into these areas and describes methods of achieving the maximum benefit from cloud computation with minimal risk.

Overview of the Book and Technology

With all its benefits, cloud computing also brings with it concerns about the security and privacy of information extant on the cloud as a result of its size, structure, and geographical dispersion. Such concerns involve the following issues:

- Leakage and unauthorized access of data among virtual machines running on the same server
- Failure of a cloud provider to properly handle and protect sensitive information
- Release of critical and sensitive data to law enforcement or government agencies without the approval and/or knowledge of the client
- Ability to meet compliance and regulatory requirements
- System crashes and failures that make the cloud service unavailable for extended periods of time
- Hackers breaking into client applications hosted on the cloud and acquiring and distributing sensitive information
- The robustness of the security protections instituted by the cloud provider
- The degree of interoperability available so that a client can easily move applications among different cloud providers and avoid “lock-in”

Cloud users should also be concerned about the continued availability of their data over long periods of time and whether or not a cloud provider might surreptitiously exploit sensitive data for its own gain.

One mitigation method that can be used to protect cloud data is encryption. Encrypting data can protect it from disclosure by the cloud provider or from hackers, but it makes it difficult to search or perform calculations on that data.

This book clarifies all these issues and provides comprehensive guidance on how to navigate the field of cloud computing to achieve the maximum return on cloud investments without compromising information security.

How This Book Is Organized

The text explores the principal characteristics of cloud computing, including scalability, flexibility, virtualization, automation, measured service, and ubiquitous network access, while showing their relationships to secure cloud computing.

The book chapters proceed from tracing the evolution of the cloud paradigm to developing architectural characteristics, security fundamentals, cloud computing risks and threats, and useful steps in implementing secure cloud computing.

Chapter 1 defines cloud computing and provides alternative views of its application and significance in the general world of computing. Following this introduction, the chapter presents the essential characteristics of cloud computing and traces the historical architectural, technical, and operational influences that converged to establish what is understand as cloud computing today.

Chapter 2 looks at the primary elements of the cloud computing architecture using various cloud-based computing architecture models. In this chapter we'll examine cloud delivery models (the SaaS, PaaS, and IaaS elements of the SPI framework), cloud deployment models (such as private, community, public, and hybrid clouds), and look at some alternative cloud architecture models, such as the Jericho Cloud Cube.

Chapter 3 explores the fundamental concepts of cloud computing software security, covering cloud security services, cloud security principles, secure software requirements, and testing concepts. It concludes by addressing cloud business continuity planning, disaster recovery, redundancy, and secure remote access.

Chapter 4 examines cloud computing risks and threats in more detail. We'll examine cloud computing risk to privacy assurance and compliance regulations, how cloud computing presents a unique risk to "traditional" concepts of data, identity, and access management (IAM) risks, and how those risks and threats may be unique to cloud service providers (CSPs).

Chapter 5 helps identify management challenges and opportunities. Security management must be able to determine what detective and preventative controls exist to clearly define the security posture of the organization, especially as it relates to the virtualization perimeter. We'll look at security policy and computer intrusion detection and response implementation techniques, and dive deeply into virtualization security management issues.

Chapter 6 addresses the important cloud computing security architectural issues, including trusted cloud computing, secure execution environments, and microarchitectures. It also expands on the critical cloud security principles of identity management and access control and develops the concepts of autonomic systems and autonomic protection mechanisms.

Chapter 7 presents cloud life cycle issues, together with significant standards efforts, incident response approaches, encryption topics, and considerations involving retirement of cloud virtual machines and applications.

Chapter 8 recaps the important cloud computing security concepts, and offers guidance on which services should be moved to the cloud and those that should not. It also reviews questions that a potential user should ask a cloud provider, and lists organizations that provide support and information exchange on cloud applications, standards, and interoperability. Chapter 8 concludes with advice on getting started in cloud computation and a “top ten” list of important related considerations.

Who Should Read This Book

Cloud Security: *A Comprehensive Guide to Secure Cloud Computing* is designed to be a valuable source of information for those who are contemplating using cloud computing as well as professionals with prior cloud computing experience and knowledge. It provides a background of the development of cloud computing and details critical approaches to cloud computing security that affect the types of applications that are best suited to the cloud.

We think that *Cloud Security*: *A Comprehensive Guide to Secure Cloud Computing* would be a useful reference for all of the following:

- Professionals working in the fields of information technology or information system security
- Information security audit professionals
- Information system IT professionals
- Computing or information systems management
- Senior management, seeking to understand the various elements of security as related to cloud computing
- Students attending information system security certification programs or studying computer security

Summary

We hope *Cloud Security*: *A Comprehensive Guide to Secure Cloud Computing* is a useful and readable reference for everyone concerned about the risk of cloud computing and involved with the protection of data.

Issues such as data ownership, privacy protections, data mobility, quality of service and service levels, bandwidth costs, data protection, and support have to be tackled in order to achieve the maximum benefit from cloud computation with minimal risk.

As you try to find your way through a maze of security minefields, this book is mandatory reading if you are involved in any aspect of cloud computing.

CHAPTER 1: Cloud Computing Fundamentals

Out of intense complexities intense simplicities emerge.

—Winston Churchill

Cloud computing evokes different perceptions in different people. To some, it refers to accessing software and storing data in the "cloud" representation of the Internet or a network and using associated services. To others, it is seen as nothing new, but just a modernization of the time-sharing model that was widely employed in the 1960s before the advent of relatively lower-cost computing platforms. These developments eventually evolved to the client/server model and to the personal computer, which placed large amounts of computing power at people's desktops and spelled the demise of time-sharing systems.

In 1961, John McCarthy, a professor at MIT, presented the idea of computing as a utility much like electricity.[1] Another pioneer, who later developed the basis for the ARPANET, the Department of Defense's Advanced Research Projects Agency Network, and precursor to the Internet, was J.C.R. Licklider. In the 1960s, Licklider promulgated ideas at both ARPA and Bolt, Beranek and Newman (BBN), the high-technology research and development company, that envisioned networked computers at a time when punched card, batch computing was dominant. He stated, "If such a network as I envisage nebulously could be brought into operation, we could have at least four large computers, perhaps six or eight small computers, and a great assortment of disc files and magnetic tape units—not to mention remote consoles and teletype stations—all churning away."[2]

The conjunction of the concepts of utility computing and a ubiquitous world-wide network provided the basis for the future evolution of cloud computing.

What Is Cloud Computing?

In an October, 2009 presentation titled “Effectively and Securely Using the Cloud Computing Paradigm,”[3] by Peter Mell and Tim Grance of the National Institute of Standards and Technology (NIST) Information Technology Laboratory, cloud computing is defined as follows:

> *Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable and reliable computing resources (e.g., networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal consumer 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 as follows:

- On-demand self-service
- Ubiquitous network access
- Resource pooling
- Location independence
- Rapid elasticity
- Measured service

The service models are as follows:

- Cloud Software as a Service (SaaS)—Use provider's applications over a network.

- Cloud Platform as a Service (PaaS)—Deploy customer-created applications to a cloud.
- Cloud Infrastructure as a Service (IaaS)—Rent processing, storage, network capacity, and other fundamental computing resources.

The deployment models, which can be either internally or externally implemented, are summarized in the NIST presentation as follows:

- Private cloud—Enterprise owned or leased
- Community cloud—Shared infrastructure for specific community
- Public cloud—Sold to the public, mega-scale infrastructure
- Hybrid cloud—Composition of two or more clouds

These characteristics and models are covered in detail in Chapter 2.

In 2009, the Open Cloud Manifesto was developed by a group of organizations including IBM, Intel, and Google to propose practices for use in the provision of cloud computing services. In the "Open Cloud Manifesto" (www.opencloudmanifesto.org), cloud computing is defined with a set of characteristics and value propositions. The characteristics outlined in the manifesto are as follows:

- The ability to scale and provision computing power dynamically in a cost-efficient way.
- The ability of the consumer (end user, organization, or IT staff) to make the most of that power without having to manage the underlying complexity of the technology.

- The cloud architecture itself can be private (hosted within an organization's firewall) or public (hosted on the Internet).

The value propositions listed in the manifesto are as follows:

- **Scalability on demand**—All organizations have to deal with changes in their environment. The ability of cloud computing solutions to scale up and down is a major benefit. If an organization has periods of time during which their computing resource needs are much higher or lower than normal, cloud technologies (both private and public) can deal with those changes.
- **Streamlining the data center**—An organization of any size will have a substantial investment in its data center. That includes buying and maintaining the hardware and software, providing the facilities in which the hardware is housed, and hiring the personnel who keep the data center running. An organization can streamline its data center by taking advantage of cloud technologies internally or by offloading workload into the public.
- **Improving business processes**—The cloud provides an infrastructure for improving business processes. An organization and its suppliers and partners can share data and applications in the cloud, enabling everyone involved to focus on the business process instead of the infrastructure that hosts it.
- **Minimizing startup costs**—For companies that are just starting out, organizations in emerging markets, or even advanced technology groups in larger organizations, cloud computing greatly reduces startup costs. The new organization starts with an infrastructure already in place, so the time and other resources that would be spent on building a data center are borne by

the cloud provider, whether the cloud is private or public.

From a different perspective, in a ZDNet article titled “The Five Defining Characteristics of Cloud Computing” (http://news.zdnet.com/2100-9595_22-287001.html), Dave Malcolm Surgient proposes the following five defining characteristics of cloud computing:

- **Dynamic computing infrastructure**—A standardized, scalable, dynamic, virtualized, and secure physical infrastructure with levels of redundancy to ensure high levels of availability
- **IT service-centric approach**—As opposed to a server-centric model, the availability of an easily accessible, dedicated instance of an application or service
- **Self-service-based usage model**—The capability to upload, build, deploy, schedule, manage, and report on provided business services on demand
- **Minimally or self-managed platform**—Self-management via software automation employing the following:
    - A provisioning engine for deploying services and tearing them down, recovering resources for high levels of reuse
    - Mechanisms for scheduling and reserving resource capacity
    - Capabilities for configuring, managing, and reporting to ensure that resources can be allocated and reallocated to multiple groups of users
    - Tools for controlling access to resources, and policies for how resources can be used or operations can be performed

- **Consumption-based billing**—Payment for resources as they are used

Important Factors in the Development of Cloud Computing

A number of dynamics such as software interoperability standards, virtualization technologies, high-bandwidth communications, the delivery of enterprise applications, and Web 2.0 contributed to the emergence of cloud computing.

Web 2.0 is a term that refers to Web design resulting in an interactive transport mechanism, rather than conventional static screens. Web 2.0 is viewed as a platform for running software applications instead of running them on desktop PCs. Tim O'Reilly of O'Reilly Media is generally acknowledged as coining the term “Web 2.0.” Some of the characteristics commonly associated with Web 2.0 are as follows:

- Use of asynchronous JavaScript and XML (Ajax)
- Combination of services from a number of sources to create a new service (mashup)
- Free Web services
- Use of Really Simple Syndication (RSS)
- Social networking
- Interactive dictionaries and encyclopedias
- Blogging
- Collaborative applications
- Sophisticated gaming
- Wikipedia and other wikis
- Optimized search engines

In 1999, Salesforce.com was formed to deliver enterprise applications over the Internet. This capability was followed in 2002 by the provision of Amazon Web Services, and in 2006 by Amazon's Elastic Compute Cloud (EC2) commercial Web service for running customers' applications. In 2009, Google and Microsoft began offering enterprise application services.

Cloud computing developed from technologies and business approaches that emerged over a number of years. The major building blocks range from Internet technology to cloud service providers, as illustrated in Figure 1.1.

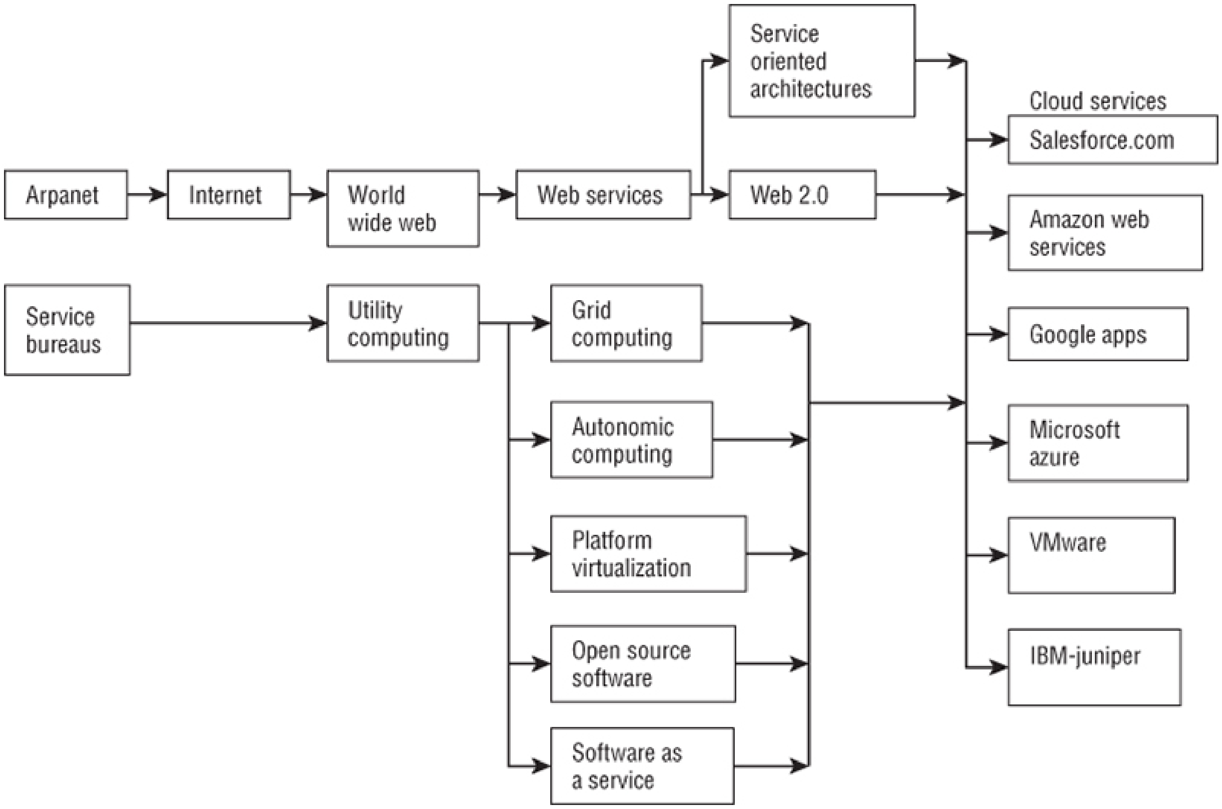


Figure 1.1 Origins of cloud computing

The important elements in the origination of cloud computing will be explored in detail in this book, but a few of the major items are summarized in Table 1.1 for background.

Table 1.1 Important Elements in the Origination of Cloud Computing

Item	**Description**
Utility Computing	The packaging and delivery of computing resources to a customer who pays for these resources as a metered service when needed. The objective is to use services effectively while reducing associated costs. The term “utility” is used to compare this type of computing resource utilization and payment to those of utilities such as providers of electricity or natural gas.
Grid Computing	The application of the processing power of multiple networked computing resources to solve a specific problem. It is a form of parallel processing conducted on a network of computers. In grid computing, servers, storage, and networks are combined to form powerful computing resource nodes that can be dynamically provisioned as needed.
Autonomic Computing	The functioning of a computer system without external control. The term is based on the autonomic nervous system of the human body, which controls breathing, heart functioning, and so on without conscious input from the individual. The objective of autonomic computing is to have the computer perform critical and complex functions without any major intervention by a user.

Item	**Description**
Platform Virtualization	The logical partitioning of physical computing resources into multiple execution environments, including servers, applications, and operating systems. Virtualization is based on the concept of a virtual machine running on a physical computing platform. Virtualization is controlled by a Virtual Machine Monitor (VMM), known as a *hypervisor*. Xen, an open-source hypervisor, is widely used for cloud computing.
Software as a Service (SaaS)	A software distribution and deployment model in which applications are provided to customers as a service. The applications can run on the users' computing systems or the provider's Web servers. SaaS provides for efficient patch management and promotes collaboration.
Service Oriented Architectures (SOA)	A set of services that communicate with each other, whose interfaces are known and described, whose functions are loosely coupled (the type of interface is not tied to the implementation), and whose use can be incorporated by multiple organizations. The SOA service interfaces are specified in XML and the services are expressed in WSDL.
	Applications can access services in a UDDI (Universal Description, Definition, and Integration) registration directory.
Cloud Services Examples	Salesforce.com provides enterprise cloud computing services in 1999.

Item	Description
	Cloud computing services provided by Amazon Web Services in 2002.
	Elastic Compute Cloud (EC2) commercial services offered by Amazon to small companies and individuals whereby computing resources can be rented.
	Google offers Google Apps, which include Web applications such as Gmail, Docs, and Calendar.
	Microsoft Azure Services Cloud Platform supports applications to be hosted and run at Microsoft data centers.
	VMware is a company that provides virtualization software for a variety of platforms.
	IBM and Juniper Networks formed a collaborative partnership in the delivery of cloud computing services.

What Cloud Computing Isn't

Even though cloud computing can incorporate some of the computing paradigms listed in Table 1.1, it is not synonymous with them. For example, cloud computing is not the same as utility computing. Cloud computing does not always employ the metered service pricing of utility computing, and cloud computing can use distributed, virtualized platforms instead of a centralized computing resource.

Is cloud computing the equivalent of grid computing? Grid computing does employ distributed virtual machines, but unlike cloud computing, these machines are usually focused on a single, very large task.

Sometimes client/server computing is viewed as cloud computing, with the cloud appearing in the server role. However, in the traditional client-server model, the server is a specific machine at a specific location. Computations running in the cloud can be based on computers anywhere, split among computers, and can use virtualized platforms, all unknown to the user. All the user knows is that he or she is accessing resources and using processing and storage somewhere to get results.

Cloud computing is not Software as a Service, which is software that an organization can purchase and manage; it is run on the user's hardware or someone else's machines.

Nor is cloud computing virtualization, although it can be used as an element to implement cloud computing. Operating system virtualization can be employed on an organization's local computers or in a data center, which is not cloud computing. However, virtualization can be employed in computing resources out in the cloud.

Cloud computing is not the same as service-oriented architecture (SOA), which supports the exchange of data among different applications engaged in business processes.

In short, although the preceding terms are not synonymous with cloud computing, depending on the implementation they can be a constituent of the cloud.

Alternative Views

A number of prominent people view cloud computing as pure hype and really nothing new. In an online video blog (http://www.techcentral.ie/article.aspx?id=13775), Oracle CEO Larry Ellison bluntly states, “What the hell is cloud computing? ... When I read these articles on cloud computing, it is pure idiocy.... Some say it is a using a computer that is out there.... The people that are writing this are insane.... When is this idiocy going to stop?”

Noted information security expert Bruce Schneier, in his June 4, 2009 online newsletter *Schneier on Security* (www.schneier.com/blog/archives/2009/06/cloud_computing.html), says “This year's overhyped IT concept is cloud computing.... But, hype aside, cloud computing is nothing new. It's the modern version of the timesharing model from the 1960s, which was eventually killed by the rise of the personal computer. It's what Hotmail and Gmail have been doing all these years, and it's social networking sites, remote backup companies, and remote email filtering companies such as MessageLabs. Any IT outsourcing—network infrastructure, security monitoring, remote hosting—is a form of cloud computing.”

In a February 10, 2009 *Information Week* article titled “HP on the Cloud: The World Is Cleaving in Two” (http://www.informationweek.com/news/services/business/showArticle.jhtml?articleID=213402906), Russ Daniels of Hewlett Packard states, “Virtually every enterprise will operate in hybrid mode,” with some of its operations on the premises and some in the cloud, he predicted. Contrary to some theories put forth, he says that cloud computing is not a replacement for the data center. “The idea that we're going to one day throw a switch and move everything out to one of a small number of external data centers, located next to a low-cost power source, is nonsensical. It's not going to happen. Cloud computing is not the end of IT.”

Another interesting view of cloud computing can be found at the hardware level. In an online article from EDN (Electronics Design, Strategy, News, at www.edn.com/blog/1690000169/post/1490048349.html), one mode of cloud computing is discussed as clusters of chips. The article reviews presentations from *Hot Chips 21*, *The Symposium on High-Performance Chips*, August 23–25, 2009 (www.hotchips.org/hc21/main_page.htm).

One of the conclusions that can be drawn from the symposium is that silicon designers have their own view of cloud computing that is related to chip architecture. Even though talking about cloud computing from the silicon chip level seems incongruous, it is valuable to understand their perspective.

According to the EDN article, silicon designers view cloud computing as a hierarchy of three elements, as follows:

1. Computing kernels—Processor cores or groups of cores enclosed within a secure perimeter and united by a single coherent address space. This definition is general enough that it could encompass a processor in a PC or a large multiprocessor system.
2. Clusters—Groups of kernels that are connected by a private local area network and whose respective tasks communicate among each other over low-bandwidth links.
3. Systems—Clusters connected through public networks and employing communications that cross security perimeter boundaries. These transactions are necessarily slower than intercluster communications.

Using these definitions, a conventional cloud would be viewed as large server farms that incorporate clusters and use kernels as server boards. An alternative approach broached at the symposium proposed the use of Sony PlayStation 3 (PS3) platforms containing the Cell Broadband processor as low-cost clusters and connecting these clusters through a public network to establish a robust cloud. The processors in this cluster would be powerful, with parallel floating-point hardware and high-speed internal communications. Using the PS3 or future equivalents, this type of cloud could be implemented at relatively low cost, be made widely available, and be amenable to open-source collaborations.

Essential Characteristics

The NIST definition of cloud computing[4] states that the cloud model comprises five essential characteristics. These characteristics are explored in the following sections.

On-Demand Self-Service

On-demand self-service enables users to use cloud computing resources as needed without human interaction between the user and the cloud service provider. With on-demand self-service, a consumer can schedule the use of cloud services such as computation and storage as needed, in addition to managing and deploying these services. In order to be effective and acceptable to the consumer, the self-service interface must be user-friendly and provide effective means to manage the service offerings. This ease of use and elimination of human interaction provides efficiencies and cost savings to both the user and the cloud service provider.

BroadNetwork Access

For cloud computing to be an effective alternative to in-house data centers, high-bandwidth communication links must be available to connect to the cloud services. One of the principal economic justifications for cloud computing is that the lowered cost of high-bandwidth network communication to the cloud provides access to a larger pool of IT resources that sustain a high level of utilization.

Many organizations use a three-tier architecture to connect a variety of computing platforms such as laptops, printers, mobile phones, and PDAs to the wide area network (WAN). This three-tier architecture comprises the following elements:

- Access switches that connect desktop devices to aggregation switches
- Aggregation switches that control flows
- Core routers and switches that provide connection to the WAN and traffic management

This three-tier approach results in latency times of 50 microseconds or more, which causes problematic delays when using cloud computing. For good performance, the switching environment should have a latency time of 10 microseconds or less. A two-tier approach that eliminates the aggregation layer can meet this requirement, using 10G (10 Gigabits/sec) Ethernet switches and the forthcoming 100G Ethernet switches.

Location-Independent Resource Pooling

The cloud must have a large and flexible resource pool to meet the consumer's needs, provide economies of scale, and meet service level requirements. Applications require resources for their execution, and these resources must be allocated efficiently for optimum performance. The resources can be physically located at many geographic locations and assigned as virtual components of the computation as needed. As stated by NIST,[5] “There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter).”

Rapid Elasticity

Rapid elasticity refers to the ability of the cloud to expand or reduce allocated resources quickly and efficiently to meet the requirements of the self-service characteristic of cloud computing. This allocation might be done automatically and appear to the user as a large pool of dynamic resources that can be paid for as needed and when needed.

One of the considerations in enabling rapid elasticity is the development and implementation of loosely coupled services that scale independently of other services and are not dependent on the elasticity of these other services.

Measured Service

Because of the service-oriented characteristics of cloud computing, the amount of cloud resources used by a consumer can be dynamically and automatically allocated and monitored. The customer can then be billed based on the measured usage of only the cloud resources that were allotted for the particular session.

The NIST view of measured service is "Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported providing transparency for both the provider and consumer of the utilized service."[6]

Architectural Influences

The realization of cloud computing was affected by a number of architectural developments over the past decades. These influences range from advances in high-performance computing to scaling and parallelism advances. Some of the principal architectural developments that support cloud computing are summarized in the following sections.

High-Performance Computing

Because of the Internet and high-performance computers, an evolution is occurring in computing. This evolution is the movement from tasks that are computationally intensive to those problems that are data intensive. This evolution characterizes some types of cloud computing applications, which are practical to run because of high-performance computers. These computers play a key role in cloud computing, and some of the major milestones in their development are presented in this section.

The computers known as *supercomputers* evolved during the 1960s. In 1961, IBM developed the IBM 7030 “Stretch,” which was the first transistor-based supercomputer. It was built for the Los Alamos National Laboratory and was specified at 1.2 MFLOPS (million floating-point operations per second.)

High-performance computing and supercomputing cannot be discussed without acknowledging Seymour Cray, who is credited with developing the first “real” supercomputers. While at Control Data Corporation (CDC), Cray developed the 3 MFLOP CDC 6600 in 1964 and the 36 MFLOP CDC 7600 in 1969. These were based on the relatively new silicon transistor technology. Cray left CDC in 1972 to form his own supercomputing company, Cray Research.

CDC continued on the supercomputer path and delivered the 100 MFLOP CDC STAR-100 in 1974. The STAR-100 was a vector processor, meaning it could operate on multiple arrays of data simultaneously.

Supercomputing technology developments accelerated during the next three decades with a variety of products. Detailing every one is beyond the scope of this text, but some of the key machines are summarized in Table 1.2. In the table, Gigaflops (GFLOPS) represent one billion (10^9) floating point operations per second, Teraflops (TFLOPS) refer to one trillion (10^{12}) floating point operations per second, and Petaflops (PFLOPS) represent are quadrillion (10^{15}) floating point operations per second.

An interesting milestone along the path of supercomputer development was the idea of connecting low-cost, commercially available personal computers in a network cluster to form a high-performance computing system. This idea was formulated in 1993 as the Beowulf computing cluster concept, developed by Thomas Sterling and Donald Becker of NASA. Beowulf uses open-source operating systems such as Solaris or Linux. One of the main characteristics of Beowulf is that all the connected machines appear as a powerful, single resource to the user.

The first prototype in the Beowulf project used 16 Intel DX4 processors connected by 10Mbit/second Ethernet. The DX4 processor is an Intel chip with triple clocking. Because the DX4 processor speed was too great for a single Ethernet bus, a “channel-bonded” Ethernet was developed by spreading the communications across two or more Ethernet buses. This approach is no longer necessary with the advent of Gigabit Ethernet. This initial cluster demonstrated the ability of COTS (commercial off the shelf) products to implement high-performance computing systems.

In general, a Beowulf architecture has the following characteristics:

- It is designed for parallel computing.