Chapter 3. Understanding Cloud Computing

This is the first of two chapters that provide an overview of introductory cloud computing topics. It begins with a brief history of cloud computing along with short descriptions of its business and technology drivers. This is followed by definitions of basic concepts and terminology, in addition to explanations of the primary benefits and challenges of cloud computing adoption.

3.1. Origins and Influences

A Brief History

The idea of computing in a "cloud" traces back to the origins of utility computing, a concept that computer scientist John McCarthy publicly proposed in 1961:

"If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility. ... The computer utility could become the basis of a new and important industry."

In 1969, Leonard Kleinrock, a chief scientist of the Advanced Research Projects Agency Network or ARPANET project that seeded the Internet, stated:

"As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of 'computer utilities' ...".

The general public has been leveraging forms of Internet-based computer utilities since the mid-1990s through various incarnations of search engines (Yahoo!, Google), e-mail services (Hotmail, Gmail), open publishing platforms (MySpace, Facebook, YouTube), and other types of social media (Twitter, LinkedIn). Though consumer-centric, these services popularized and validated core concepts that form the basis of modern-day cloud computing.

In the late 1990s, Salesforce.com pioneered the notion of bringing remotely provisioned services into the enterprise. In 2002, Amazon.com launched the Amazon Web Services (AWS) platform, a suite of enterprise-oriented services that provide remotely provisioned storage, computing resources, and business functionality.

A slightly different evocation of the term "Network Cloud" or "Cloud" was introduced in the early 1990s throughout the networking industry. It referred to an abstraction layer derived in the delivery methods of data across heterogeneous public and semi-public networks that were primarily packet-switched, although cellular networks used the "Cloud" term as well. The networking method at this point supported the transmission of data from one end-point (local network) to the "Cloud" (wide area network) and then further decomposed to another intended end-point. This is relevant, as the networking industry still references the

use of this term, and is considered an early adopter of the concepts that underlie utility computing.

It wasn't until 2006 that the term "cloud computing" emerged in the commercial arena. It was during this time that Amazon launched its Elastic Compute Cloud (EC2) services that enabled organizations to "lease" computing capacity and processing power to run their enterprise applications. Google Apps also began providing browser-based enterprise applications in the same year, and three years later, the Google App Engine became another historic milestone.

Definitions

A Gartner report listing cloud computing at the top of its strategic technology areas further reaffirmed its prominence as an industry trend by announcing its formal definition as:

"...a style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service to external customers using Internet technologies."

This is a slight revision of Gartner's original definition from 2008, in which "massively scalable" was used instead of "scalable and elastic." This acknowledges the importance of scalability in relation to the ability to scale vertically and not just to enormous proportions.

Forrester Research provided its own definition of cloud computing as:

"...a standardized IT capability (services, software, or infrastructure) delivered via Internet technologies in a pay-per-use, self-service way."

The definition that received industry-wide acceptance was composed by the National Institute of Standards and Technology (NIST). NIST published its original definition back in 2009, followed by a revised version after further review and industry input that was published in September of 2011:

"Cloud computing is 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."

This book provides a more concise definition:

"Cloud computing is a specialized form of distributed computing that introduces utilization models for remotely provisioning scalable and measured resources."

This simplified definition is in line with all of the preceding definition variations that were put forth by other organizations within the cloud computing industry. The characteristics, service models, and deployment models referenced in the NIST definition are further covered in Chapter 4.

Business Drivers

Before delving into the layers of technologies that underlie clouds, the motivations that led to their creation by industry leaders must first be understood. Several of the primary business drivers that fostered modern cloud-based technology are presented in this section.

The origins and inspirations of many of the characteristics, models, and mechanisms covered throughout subsequent chapters can be traced back to the upcoming business drivers. It is important to note that these influences shaped clouds and the overall cloud computing market from both ends. They have motivated organizations to adopt cloud computing in support of their business automation requirements. They have correspondingly motivated other organizations to become providers of cloud environments and cloud technology vendors in order to create and meet the demand to fulfill consumer needs.

Capacity Planning

Capacity planning is the process of determining and fulfilling future demands of an organization's IT resources, products, and services. Within this context, *capacity* represents the maximum amount of work that an IT resource is capable of delivering in a given period of time. A discrepancy between the capacity of an IT resource and its demand can result in a system becoming either inefficient (over-provisioning) or unable to fulfill user needs (under-provisioning). Capacity planning is focused on minimizing this discrepancy to achieve predictable efficiency and performance.

Different capacity planning strategies exist:

• *Lead Strategy* – adding capacity to an IT resource in anticipation of demand

• *Lag Strategy* – adding capacity when the IT resource reaches its full capacity

• *Match Strategy* – adding IT resource capacity in small increments, as demand increases

Planning for capacity can be challenging because it requires estimating usage load fluctuations. There is a constant need to balance peak usage requirements without unnecessary over-expenditure on infrastructure. An example is outfitting IT infrastructure to accommodate maximum usage loads which can impose unreasonable financial investments. In such cases, moderating investments can result in under-provisioning, leading to transaction losses and other usage limitations from lowered usage thresholds.

Cost Reduction

A direct alignment between IT costs and business performance can be difficult to maintain. The growth of IT environments often corresponds to the assessment of their maximum usage requirements. This can make the support of new and expanded business automations an ever-increasing investment. Much of this required investment is funneled into infrastructure expansion because the usage potential of a given automation solution will always be limited by the processing power of its underlying infrastructure.

Two costs need to be accounted for: the cost of acquiring new infrastructure, and the cost of its ongoing ownership. Operational overhead represents a considerable share of IT budgets, often exceeding up-front investment costs.

Common forms of infrastructure-related operating overhead include the following:

• technical personnel required to keep the environment operational

• upgrades and patches that introduce additional testing and deployment cycles

• utility bills and capital expense investments for power and cooling

• security and access control measures that need to be maintained and enforced to protect infrastructure resources

• administrative and accounts staff that may be required to keep track of licenses and support arrangements

The on-going ownership of internal technology infrastructure can encompass burdensome responsibilities that impose compound impacts on corporate budgets. An IT department can consequently become a significant—and at times overwhelming—drain on the business, potentially inhibiting its responsiveness, profitability, and overall evolution.

Organizational Agility

Businesses need the ability to adapt and evolve to successfully face change caused by both internal and external factors. Organizational agility is the measure of an organization's responsiveness to change.

An IT enterprise often needs to respond to business change by scaling its IT resources beyond the scope of what was previously predicted or planned for. For example, infrastructure may be subject to limitations that prevent the organization from responding to usage fluctuations—even when anticipated—if previous capacity planning efforts were restricted by inadequate budgets.

In other cases, changing business needs and priorities may require IT resources to be more available and reliable than before. Even if sufficient infrastructure is in place for an organization to support anticipated usage volumes, the nature of the usage may generate runtime exceptions that bring down hosting servers. Due to a lack of reliability controls within the infrastructure, responsiveness to consumer or customer requirements may be reduced to a point whereby a business' overall continuity is threatened.

On a broader scale, the up-front investments and infrastructure ownership costs that are required to enable new or expanded business automation solutions may themselves be prohibitive enough for a business to settle for IT infrastructure of less-than-ideal quality, thereby decreasing its ability to meet real-world requirements.

Worse yet, the business may decide against proceeding with an automation solution altogether upon review of its infrastructure budget, because it simply cannot afford to. This form of inability to respond can inhibit an organization from keeping up with market demands, competitive pressures, and its own strategic business goals.

Technology Innovations

Established technologies are often used as inspiration and, at times, the actual foundations upon which new technology innovations are derived and built. This section briefly describes the pre-existing technologies considered to be the primary influences on cloud computing.

Clustering

A cluster is a group of independent IT resources that are interconnected and work as a single system. System failure rates are reduced while availability and reliability are increased, since redundancy and failover features are inherent to the cluster.

A general prerequisite of hardware clustering is that its component systems have reasonably identical hardware and operating systems to provide similar performance levels when one failed component is to be replaced by another. Component devices that form a cluster are kept in synchronization through dedicated, high-speed communication links.

The basic concept of built-in redundancy and failover is core to cloud platforms. Clustering technology is explored further in Chapter 8 as part of the *Resource Cluster* mechanism description.

Grid Computing

A computing grid (or "computational grid") provides a platform in which computing resources are organized into one or more logical pools. These pools are collectively coordinated to provide a high performance distributed grid, sometimes referred to as a "super virtual computer." Grid computing differs from clustering in that grid systems are much more loosely coupled and distributed. As a result, grid computing systems can involve computing resources that are heterogeneous and geographically dispersed, which is generally not possible with cluster computing-based systems.

Grid computing has been an on-going research area in computing science since the early 1990s. The technological advancements achieved by grid computing projects have influenced various aspects of cloud computing platforms and mechanisms, specifically in relation to common feature-sets such as networked access, resource pooling, and scalability and resiliency. These types of features can be established by both grid computing and cloud computing, in their own distinctive approaches.

For example, grid computing is based on a middleware layer that is deployed on computing resources. These IT resources participate in a grid pool that implements a series of workload distribution and coordination functions. This middle tier can contain load balancing logic, failover controls, and autonomic configuration management, each having previously inspired similar—and several more sophisticated—cloud computing technologies. It is for this reason that some classify cloud computing as a descendant of earlier grid computing initiatives.

Virtualization

Virtualization represents a technology platform used for the creation of virtual instances of IT resources. A layer of virtualization software allows physical IT resources to provide multiple virtual images of themselves so that their underlying processing capabilities can be shared by multiple users.

Prior to the advent of virtualization technologies, software was limited to residing on and being coupled with static hardware environments. The virtualization process severs this software-hardware dependency, as hardware requirements can be simulated by emulation software running in virtualized environments.

Established virtualization technologies can be traced to several cloud characteristics and cloud computing mechanisms, having inspired many of their core features. As cloud computing evolved, a generation of *modern* virtualization technologies emerged to overcome the performance, reliability, and scalability limitations of traditional virtualization platforms.

As a foundation of contemporary cloud technology, modern virtualization provides a variety of virtualization types and technology layers that are discussed separately in Chapter 5.

Technology Innovations vs. Enabling Technologies

It is essential to highlight several other areas of technology that continue to contribute to modern-day cloud-based platforms. These are distinguished as *cloud-enabling technologies*, the following of which are covered in Chapter 5:

- Broadband Networks and Internet Architecture
- Data Center Technology
- (Modern) Virtualization Technology

• Web Technology
• Multitenant Technology
• Service Technology

Each of these cloud-enabling technologies existed in some form prior to the formal advent of cloud computing. Some were refined further, and on occasion even redefined, as a result of the subsequent evolution of cloud computing.

Summary of Key Points

• The primary business drivers that exposed the need for cloud computing and led to its formation include capacity planning, cost reduction, and organizational agility.

• The primary technology innovations that influenced and inspired key distinguishing features and aspects of cloud computing include clustering, grid computing, and traditional forms of virtualization.

3.2. Basic Concepts and Terminology

This section establishes a set of basic terms that represent the fundamental concepts and aspects pertaining to the notion of a cloud and its most primitive artifacts.

Cloud

A *cloud* refers to a distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources. The term originated as a metaphor for the Internet which is, in essence, a network of networks providing remote access to a set of decentralized IT resources. Prior to cloud computing becoming its own formalized IT industry segment, the symbol of a cloud was commonly used to represent the Internet in a variety of specifications and mainstream documentation of Web-based architectures. This same symbol is now used to specifically represent the boundary of a cloud environment, as shown in Figure 3.1.

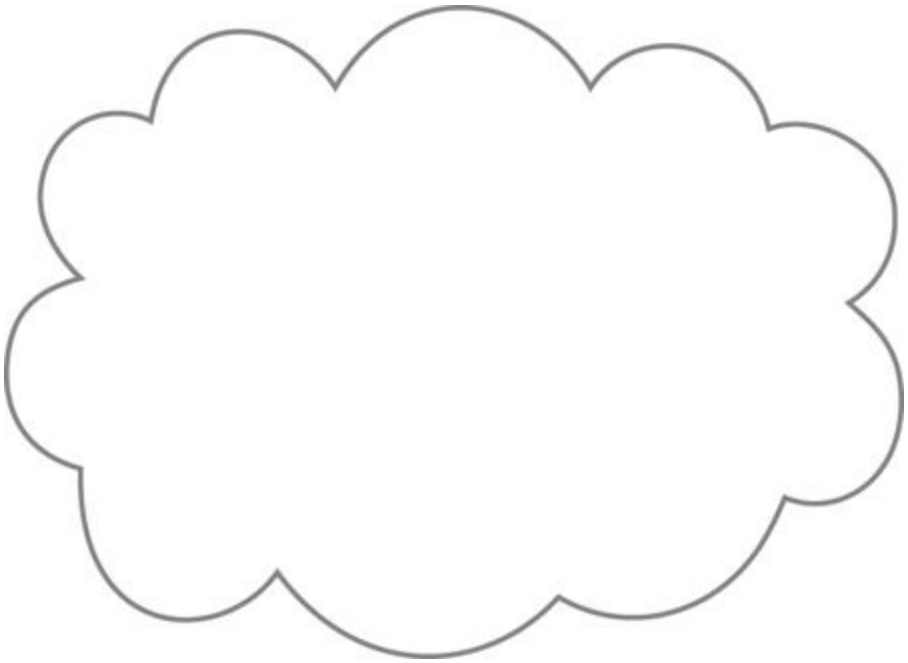

Figure 3.1. The symbol used to denote the boundary of a cloud environment.

It is important to distinguish the term “cloud” and the cloud symbol from the Internet. As a specific environment used to remotely provision IT resources, a cloud has a finite boundary. There are many individual clouds that are accessible via the Internet. Whereas the Internet provides open access to many Web-based IT resources, a cloud is typically privately owned and offers access to IT resources that is metered.

Much of the Internet is dedicated to the access of content-based IT resources published via the World Wide Web. IT resources provided by cloud environments, on the other hand, are dedicated to supplying back-end processing capabilities and user-based access to these capabilities. Another key distinction is that it is not necessary for clouds to be Web-based even if they are commonly based on Internet protocols and technologies. Protocols refer to standards and methods that allow computers to communicate with each other in a pre-defined and structured manner. A cloud can be based on the use of any protocols that allow for the remote access to its IT resources.

Note

Diagrams in this book depict the Internet using the globe symbol.

IT Resource

An *IT resource* is a physical or virtual IT-related artifact that can be either software-based, such as a virtual server or a custom software program, or hardware-based, such as a physical server or a network device (Figure 3.2).

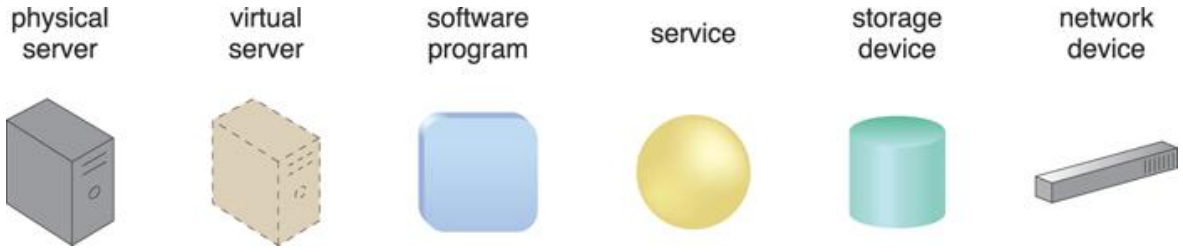

Figure 3.2. Examples of common IT resources and their corresponding symbols.

Figure 3.3 illustrates how the cloud symbol can be used to define a boundary for a cloud-based environment that hosts and provisions a set of IT resources. The displayed IT resources are consequently considered to be cloud-based IT resources.

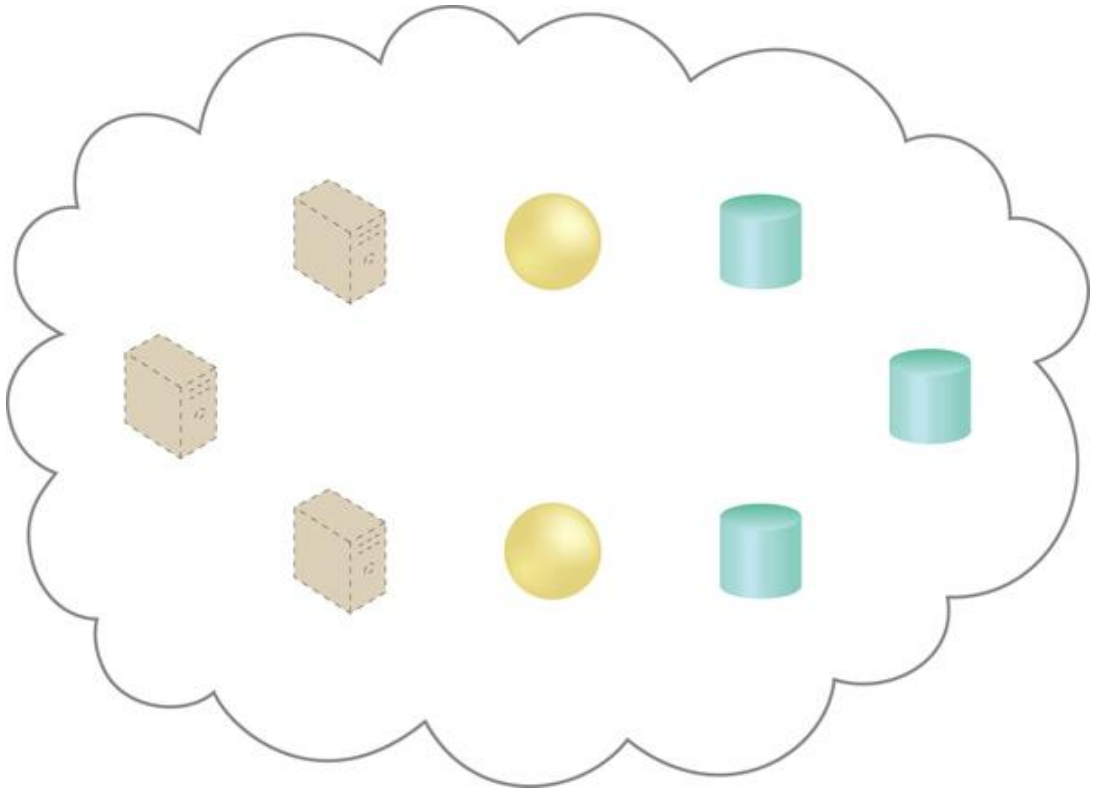

Figure 3.3. A cloud is hosting eight IT resources: three virtual servers, two cloud services, and three storage devices.

Technology architectures and various interaction scenarios involving IT resources are illustrated in diagrams like the one shown in Figure 3.3. It is important to note the following points when studying and working with these diagrams:

• The IT resources shown within the boundary of a given cloud symbol usually do not represent all of the available IT resources hosted by that cloud. Subsets of IT resources are generally highlighted to demonstrate a particular topic.

• Focusing on the relevant aspects of a topic requires many of these diagrams to intentionally provide abstracted views of the underlying technology architectures. This means that only a portion of the actual technical details are shown.

Furthermore, some diagrams will display IT resources outside of the cloud symbol. This convention is used to indicate IT resources that are not cloud-based.

Note

The virtual server IT resource displayed in Figure 3.2 is further discussed in Chapters 5 and 7. Physical servers are sometimes referred to as *physical hosts* (or just *hosts*) in reference to the fact that they are responsible for hosting virtual servers.

On-Premise

As a distinct and remotely accessible environment, a cloud represents an option for the deployment of IT resources. An IT resource that is hosted in a conventional IT enterprise within an organizational boundary (that does not specifically represent a cloud) is considered to be located on the premises of the IT enterprise, or *on-premise* for short. In other words, the term “on-premise” is another way of stating “on the premises of a controlled IT environment that is not cloud-based.” This term is used to qualify an IT resource as an alternative to “cloud-based.” An IT resource that is on-premise cannot be cloud-based, and vice-versa.

Note the following key points:

• An on-premise IT resource can access and interact with a cloud-based IT resource.

• An on-premise IT resource can be moved to a cloud, thereby changing it to a cloud-based IT resource.

• Redundant deployments of an IT resource can exist in both on-premise and cloud-based environments.

If the distinction between on-premise and cloud-based IT resources is confusing in relation to private clouds (described in the *Cloud Deployment Models* section of Chapter 4), then an alternative qualifier can be used.

Cloud Consumers and Cloud Providers

The party that provides cloud-based IT resources is the *cloud provider*. The party that uses cloud-based IT resources is the *cloud consumer*. These terms represent roles usually assumed by organizations in relation to clouds and corresponding cloud provisioning contracts. These roles are formally defined in Chapter 4, as part of the *Roles and Boundaries* section.

Scaling

Scaling, from an IT resource perspective, represents the ability of the IT resource to handle increased or decreased usage demands.

The following are types of scaling:

• *Horizontal Scaling* – scaling out and scaling in

• *Vertical Scaling* – scaling up and scaling down

The next two sections briefly describe each.

Horizontal Scaling

The allocating or releasing of IT resources that are of the same type is referred to as *horizontal scaling*(Figure 3.4). The horizontal allocation of resources is referred to as *scaling out* and the horizontal releasing of resources is referred to as *scaling in*. Horizontal scaling is a common form of scaling within cloud environments.

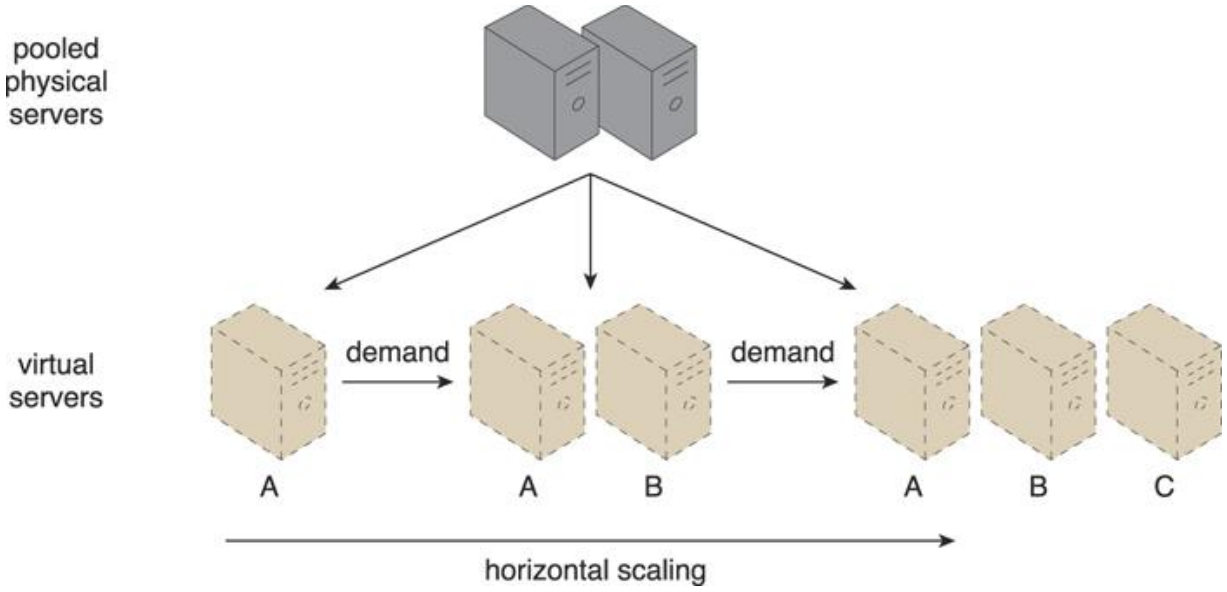

Figure 3.4. An IT resource (Virtual Server A) is scaled out by adding more of the same IT resources (Virtual Servers B and C).

Vertical Scaling

When an existing IT resource is replaced by another with higher or lower capacity, *vertical scaling* is considered to have occurred (Figure 3.5). Specifically, the replacing of an IT resource with another that has a higher capacity is referred to as *scaling up* and the replacing an IT resource with another that has a lower capacity is considered *scaling down*. Vertical scaling is less common in cloud environments due to the downtime required while the replacement is taking place.

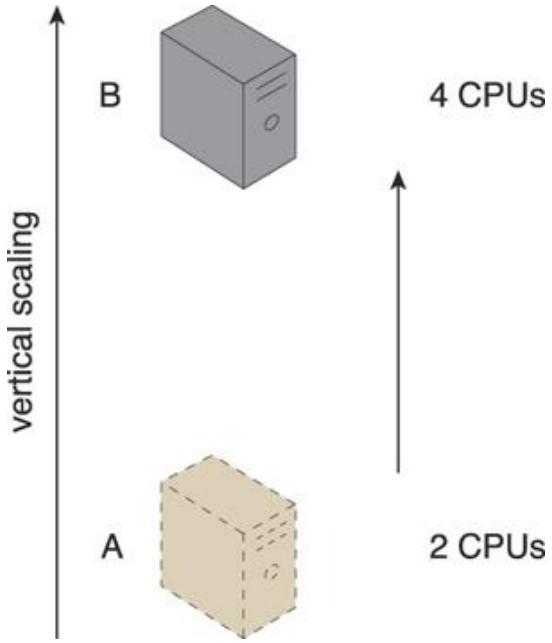

Figure 3.5. An IT resource (a virtual server with two CPUs) is scaled up by replacing it with a more powerful IT resource with increased capacity for data storage (a physical server with four CPUs).

Table 3.1 provides a brief overview of common pros and cons associated with horizontal and vertical scaling.

Table 3.1. A comparison of horizontal and vertical scaling.

Horizontal Scaling	Vertical Scaling
less expensive (through commodity hardware components)	more expensive (specialized servers)
IT resources instantly available	IT resources normally instantly available
resource replication and automated scaling	additional setup is normally needed
additional IT resources needed	no additional IT resources needed
not limited by hardware capacity	limited by maximum hardware capacity

Cloud Service

Although a cloud is a remotely accessible environment, not all IT resources residing within a cloud can be made available for remote

access. For example, a database or a physical server deployed within a cloud may only be accessible by other IT resources that are within the same cloud. A software program with a published API may be deployed specifically to enable access by remote clients.

A *cloud service* is any IT resource that is made remotely accessible via a cloud. Unlike other IT fields that fall under the service technology umbrella—such as service-oriented architecture—the term "service" within the context of cloud computing is especially broad. A cloud service can exist as a simple Web-based software program with a technical interface invoked via the use of a messaging protocol, or as a remote access point for administrative tools or larger environments and other IT resources.

In Figure 3.6, the yellow circle symbol is used to represent the cloud service as a simple Web-based software program. A different IT resource symbol may be used in the latter case, depending on the nature of the access that is provided by the cloud service.

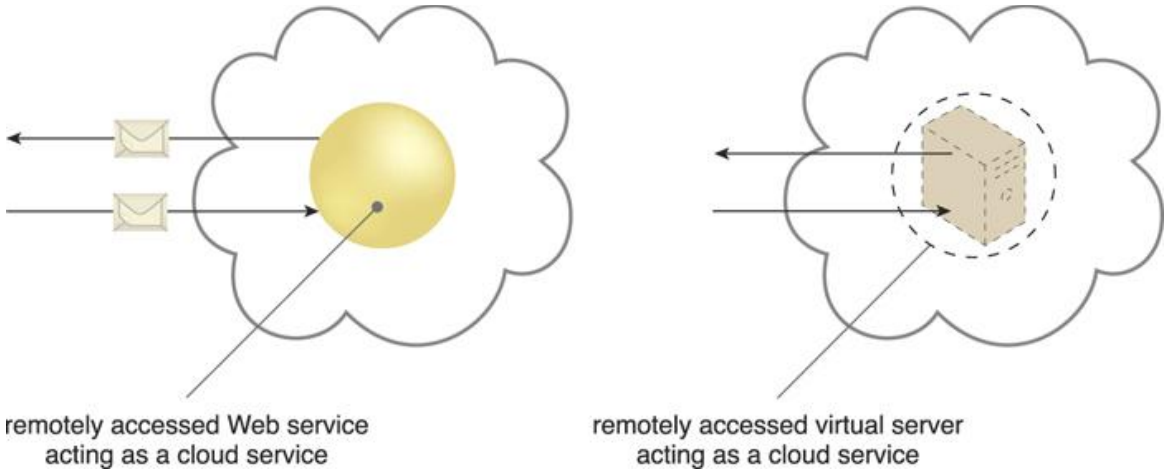

Figure 3.6. A cloud service with a published technical interface is being accessed by a consumer outside of the cloud (left). A cloud service that exists as a virtual server is also being accessed from outside of the cloud's boundary (right). The cloud service on the left is likely being invoked by a consumer program that was designed to access the cloud service's published technical interface. The cloud service on the right may be accessed by a human user that has remotely logged on to the virtual server.

The driving motivation behind cloud computing is to provide IT resources as services that encapsulate other IT resources, while offering functions for clients to use and leverage remotely. A multitude of models for generic types of cloud services have emerged, most of which are labeled with the "as-a-service" suffix.

Note

Cloud service usage conditions are typically expressed in a service-level agreement (SLA) that is the human-readable part of a service contract between a cloud provider and cloud consumer that describes QoS features, behaviors, and limitations of a cloud-based service or other provisions.

An SLA provides details of various measurable characteristics related to IT outcomes, such as uptime, security characteristics, and other specific QoS features, including availability, reliability, and performance. Since the implementation of a service is hidden from the cloud consumer, an SLA becomes a critical specification. SLAs are covered in detail in Chapter 16.

Cloud Service Consumer

The *cloud service consumer* is a temporary runtime role assumed by a software program when it accesses a cloud service.

As shown in Figure 3.7, common types of cloud service consumers can include software programs and services capable of remotely accessing cloud services with published service contracts, as well as workstations, laptops and mobile devices running software capable of remotely accessing other IT resources positioned as cloud services.

Figure 3.7. Examples of cloud service consumers. Depending on the nature of a given diagram, an artifact labeled as a cloud service consumer may be a software program or a hardware device (in which case it is implied that it is running a software program capable of acting as a cloud service consumer).

3.3. Goals and Benefits

The common benefits associated with adopting cloud computing are explained in this section.

Note

The following sections make reference to the terms “public cloud” and “private cloud.” These terms are described in the *Cloud Deployment Models* section in Chapter 4.

Reduced Investments and Proportional Costs

Similar to a product wholesaler that purchases goods in bulk for lower price points, public cloud providers base their business model on the mass-acquisition of IT resources that are then made available to cloud consumers via attractively priced leasing packages. This opens the door for organizations to gain access to powerful infrastructure without having to purchase it themselves.

The most common economic rationale for investing in cloud-based IT resources is in the reduction or outright elimination of up-front IT investments, namely hardware and software purchases and ownership costs. A cloud's Measured Usage characteristic represents a feature-set that allows measured operational expenditures (directly related to business performance) to replace anticipated capital expenditures. This is also referred to as *proportional costs*.

This elimination or minimization of up-front financial commitments allows enterprises to start small and accordingly increase IT resource allocation as required. Moreover, the reduction of up-front capital expenses allows for the capital to be redirected to the core business investment. In its most basic form, opportunities to decrease costs are derived from the deployment and operation of large-scale data centers by major cloud providers. Such data centers are commonly located in destinations where real estate, IT professionals, and network bandwidth can be obtained at lower costs, resulting in both capital and operational savings.

The same rationale applies to operating systems, middleware or platform software, and application software. Pooled IT resources are made available to and shared by multiple cloud consumers, resulting in increased or even maximum possible utilization. Operational costs and inefficiencies can be further reduced by applying proven practices and patterns for optimizing cloud architectures, their management, and their governance.

Common measurable benefits to cloud consumers include:

• On-demand access to pay-as-you-go computing resources on a short-term basis (such as processors by the hour), and the ability to release these computing resources when they are no longer needed.

• The perception of having unlimited computing resources that are available on demand, thereby reducing the need to prepare for provisioning.

• The ability to add or remove IT resources at a fine-grained level, such as modifying available storage disk space by single gigabyte increments.

• Abstraction of the infrastructure so applications are not locked into devices or locations and can be easily moved if needed.

For example, a company with sizable batch-centric tasks can complete them as quickly as their application software can scale. Using 100 servers for one hour costs the same as using one server for 100 hours. This "elasticity" of IT resources, achieved without requiring steep initial investments to create a large-scale computing infrastructure, can be extremely compelling.

Despite the ease with which many identify the financial benefits of cloud computing, the actual economics can be complex to calculate and assess. The decision to proceed with a cloud computing adoption strategy will involve much more than a simple comparison between the cost of leasing and the cost of purchasing. For example, the financial benefits of dynamic scaling and the risk transference of both over-provisioning (under-utilization) and under-provisioning (over-utilization) must also be accounted for. Chapter 15 explores common criteria and formulas for performing detailed financial comparisons and assessments.

Note

Another area of cost savings offered by clouds is the "as-a-service" usage model, whereby technical and operational implementation details of IT resource provisioning are abstracted from cloud consumers and packaged into "ready-to-use" or "off-the-shelf" solutions. These services-based products can simplify and expedite the development, deployment, and administration of IT resources when compared to performing equivalent tasks with on-premise solutions. The resulting savings in time and required IT expertise can be significant and can contribute to the justification of adopting cloud computing.

Increased Scalability

By providing pools of IT resources, along with tools and technologies designed to leverage them collectively, clouds can instantly and dynamically allocate IT resources to cloud consumers, on-demand or via the cloud consumer's direct configuration. This empowers cloud consumers to scale their cloud-based IT resources to accommodate

processing fluctuations and peaks automatically or manually. Similarly, cloud-based IT resources can be released (automatically or manually) as processing demands decrease.

A simple example of usage demand fluctuations throughout a 24 hour period is provided in Figure 3.8.

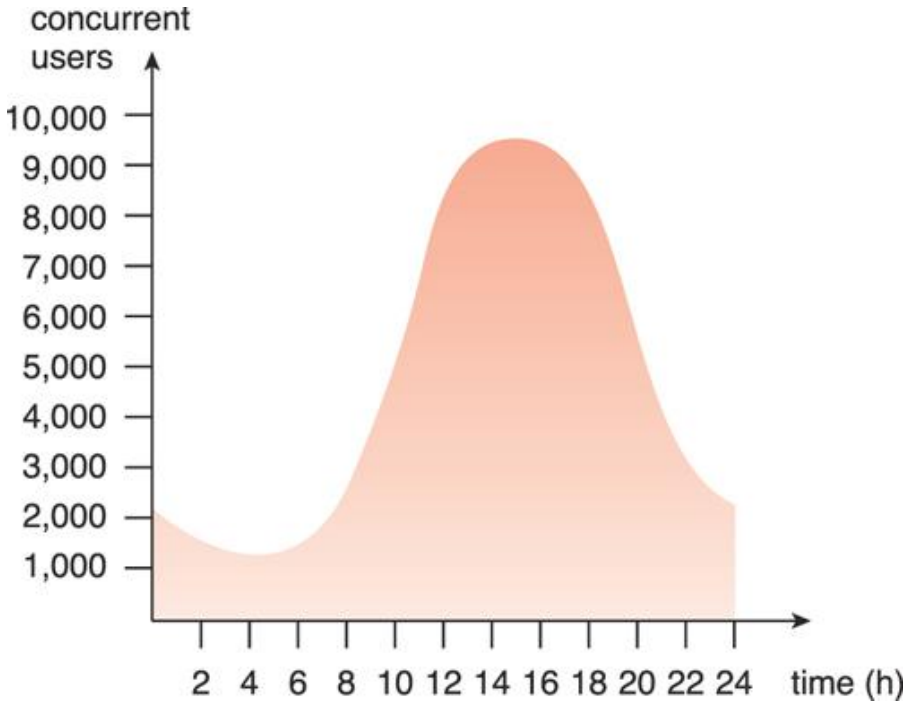

Figure 3.8. An example of an organization's changing demand for an IT resource over the course of a day.

The inherent, built-in feature of clouds to provide flexible levels of scalability to IT resources is directly related to the aforementioned proportional costs benefit. Besides the evident financial gain to the automated reduction of scaling, the ability of IT resources to always meet and fulfill unpredictable usage demands avoids potential loss of business that can occur when usage thresholds are met.

Note

When associating the benefit of Increased Scalability with the capacity planning strategies introduced earlier in the *Business Drivers* section, the Lag and Match Strategies are generally more applicable due to a cloud's ability to scale IT resources on-demand.

Increased Availability and Reliability

The availability and reliability of IT resources are directly associated with tangible business benefits. Outages limit the time an IT resource can be "open for business" for its customers, thereby limiting its usage and revenue generating potential. Runtime failures that are not immediately corrected can have a more significant impact during high-volume usage periods. Not only is the IT resource unable to respond to customer requests, its unexpected failure can decrease overall customer confidence.

A hallmark of the typical cloud environment is its intrinsic ability to provide extensive support for increasing the availability of a cloud-based IT resource to minimize or even eliminate outages, and for increasing its reliability so as to minimize the impact of runtime failure conditions.

Specifically:

• An IT resource with increased availability is accessible for longer periods of time (for example, 22 hours out of a 24 hour day). Cloud providers generally offer "resilient" IT resources for which they are able to guarantee high levels of availability.

• An IT resource with increased reliability is able to better avoid and recover from exception conditions. The modular architecture of cloud environments provides extensive failover support that increases reliability.

It is important that organizations carefully examine the SLAs offered by cloud providers when considering the leasing of cloud-based services and IT resources. Although many cloud environments are capable of offering remarkably high levels of availability and reliability, it comes down to the guarantees made in the SLA that typically represent their actual contractual obligations.

Summary of Key Points

• Cloud environments are comprised of highly extensive infrastructure that offers pools of IT resources that can be leased using a pay-for-use model whereby only the actual usage of the IT resources is billable. When compared to equivalent on-premise environments, clouds provide the potential for reduced initial investments and operational costs proportional to measured usage.

• The inherent ability of a cloud to scale IT resources enables organizations to accommodate unpredictable usage fluctuations without being limited by pre-defined thresholds that may turn away usage requests from customers.

Conversely, the ability of a cloud to decrease required scaling is a feature that relates directly to the proportional costs benefit.

• By leveraging cloud environments to make IT resources highly available and reliable, organizations are able to increase quality-of-service guarantees to customers and further reduce or avoid potential loss of business resulting from unanticipated runtime failures.

3.4. Risks and Challenges

Several of the most critical cloud computing challenges pertaining mostly to cloud consumers that use IT resources located in public clouds are presented and examined.

Increased Security Vulnerabilities

The moving of business data to the cloud means that the responsibility over data security becomes shared with the cloud provider. The remote usage of IT resources requires an expansion of trust boundaries by the cloud consumer to include the external cloud. It can be difficult to establish a security architecture that spans such a trust boundary without introducing vulnerabilities, unless cloud consumers and cloud providers happen to support the same or compatible security frameworks—which is unlikely with public clouds.

Another consequence of overlapping trust boundaries relates to the cloud provider's privileged access to cloud consumer data. The extent to which the data is secure is now limited to the security controls and policies applied by both the cloud consumer and cloud provider. Furthermore, there can be overlapping trust boundaries from different cloud consumers due to the fact that cloud-based IT resources are commonly shared.

The overlapping of trust boundaries and the increased exposure of data can provide malicious cloud consumers (human and automated) with greater opportunities to attack IT resources and steal or damage business data. Figure 3.9 illustrates a scenario whereby two organizations accessing the same cloud service are required to extend their respective trust boundaries to the cloud, resulting in overlapping trust boundaries. It can be challenging for the cloud provider to offer security mechanisms that accommodate the security requirements of both cloud service consumers.

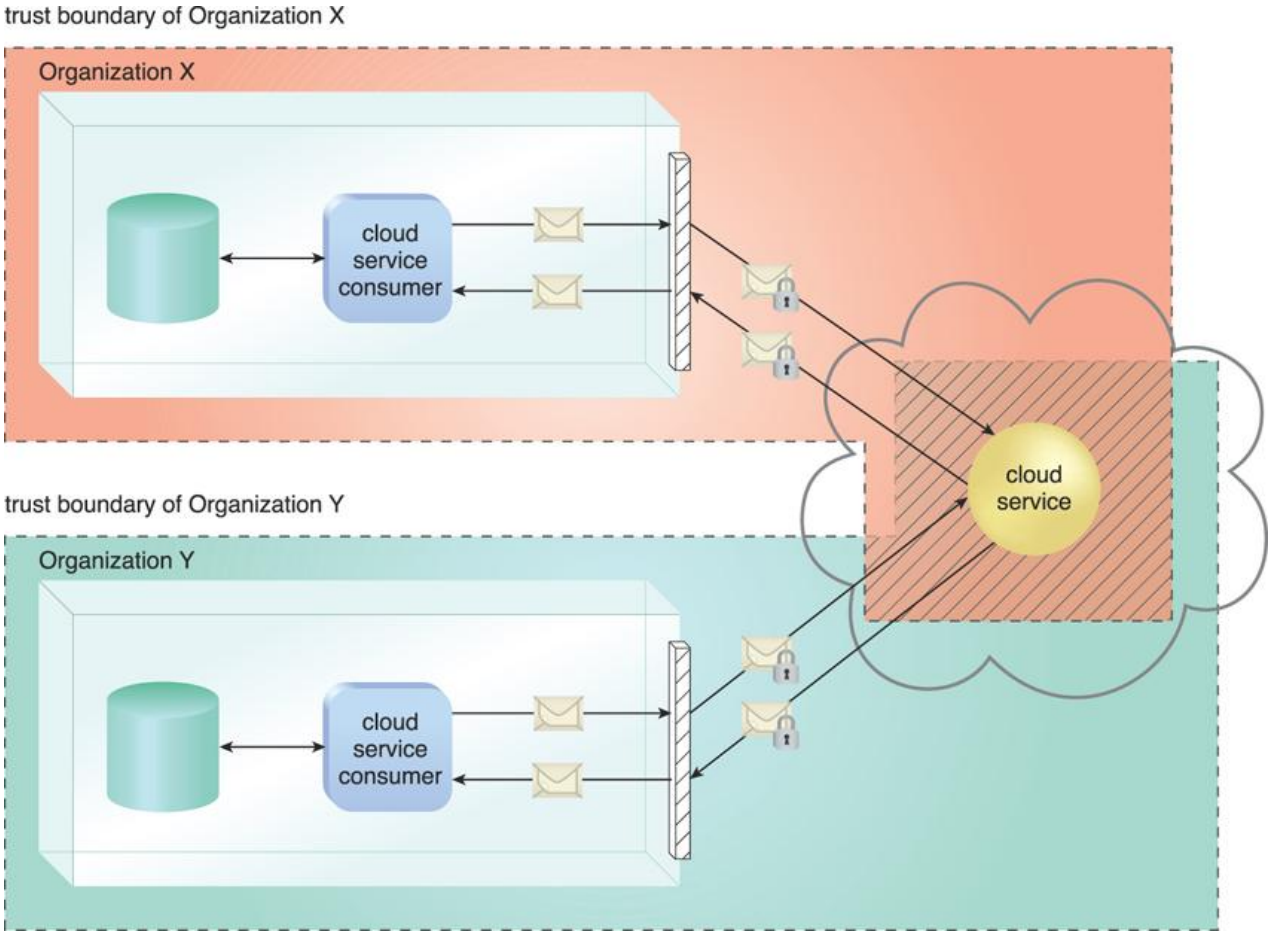

Figure 3.9. The shaded area with diagonal lines indicates the overlap of two organizations' trust boundaries.

Overlapping trust boundaries is a security threat that is discussed in more detail in Chapter 6.

Reduced Operational Governance Control

Cloud consumers are usually allotted a level of governance control that is lower than that over on-premise IT resources. This can introduce risks associated with how the cloud provider operates its cloud, as well as the external connections that are required for communication between the cloud and the cloud consumer.

Consider the following examples:

• An unreliable cloud provider may not maintain the guarantees it makes in the SLAs that were published for its cloud services. This can jeopardize the quality of the cloud consumer solutions that rely on these cloud services.

• Longer geographic distances between the cloud consumer and cloud provider can require additional network hops that introduce fluctuating latency and potential bandwidth constraints.

The latter scenario is illustrated in Figure 3.10.

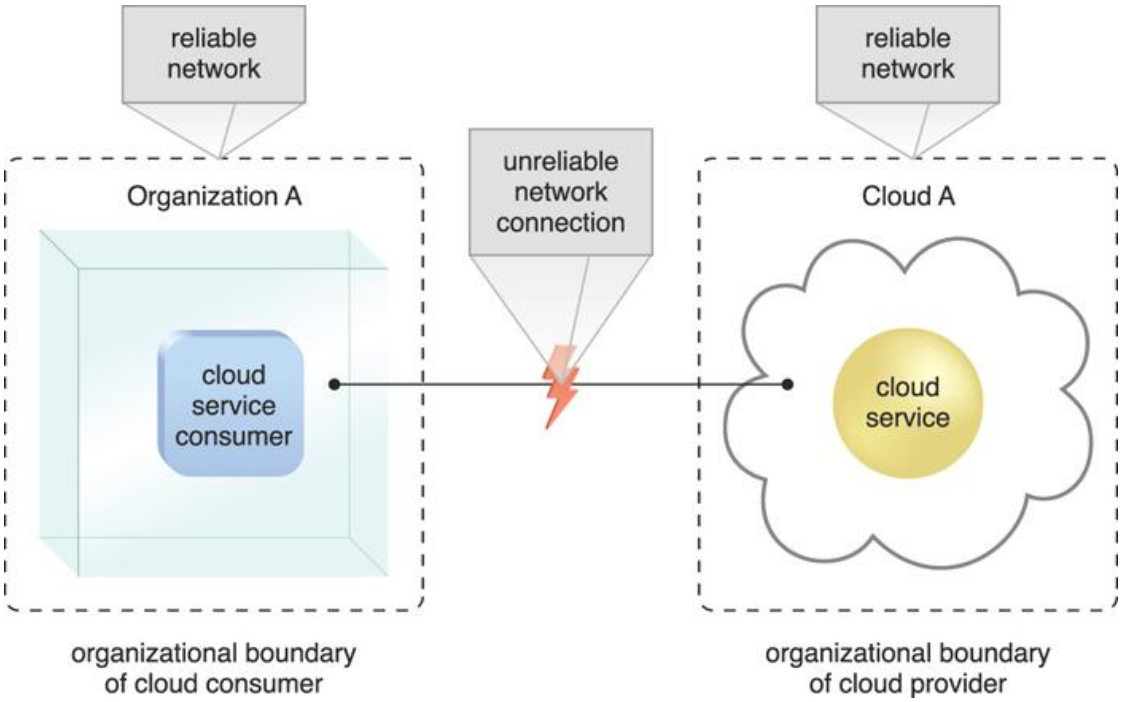

Figure 3.10. An unreliable network connection compromises the quality of communication between cloud consumer and cloud provider environments.

Legal contracts, when combined with SLAs, technology inspections, and monitoring, can mitigate governance risks and issues. A cloud governance system is established through SLAs, given the "as-a-service" nature of cloud computing. A cloud consumer must keep track of the actual service level being offered and the other warranties that are made by the cloud provider.

Note that different cloud delivery models offer varying degrees of operational control granted to cloud consumers, as further explained in Chapter 4.

Limited Portability Between Cloud Providers

Due to a lack of established industry standards within the cloud computing industry, public clouds are commonly proprietary to various extents. For cloud consumers that have custom-built solutions with

dependencies on these proprietary environments, it can be challenging to move from one cloud provider to another.

Portability is a measure used to determine the impact of moving cloud consumer IT resources and data between clouds (Figure 3.11).

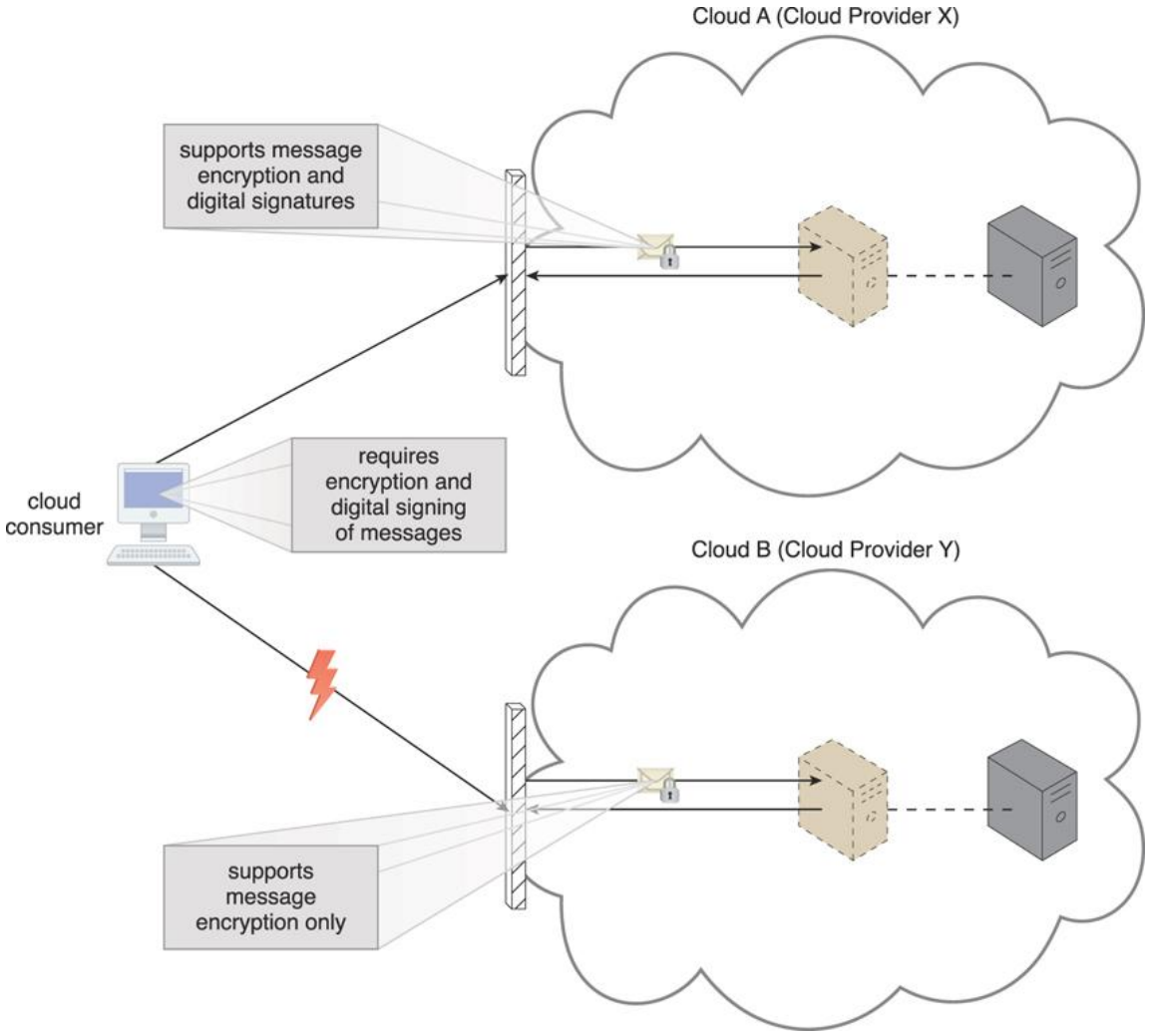

Figure 3.11. A cloud consumer's application has a decreased level of portability when assessing a potential migration from Cloud A to Cloud B, because the cloud provider of Cloud B does not support the same security technologies as Cloud A.

Multi-Regional Compliance and Legal Issues

Third-party cloud providers will frequently establish data centers in affordable or convenient geographical locations. Cloud consumers will often not be aware of the physical location of their IT resources and data when hosted by public clouds. For some organizations, this can pose serious legal concerns pertaining to industry or government regulations

that specify data privacy and storage policies. For example, some UK laws require personal data belonging to UK citizens to be kept within the United Kingdom.

Another potential legal issue pertains to the accessibility and disclosure of data. Countries have laws that require some types of data to be disclosed to certain government agencies or to the subject of the data. For example, a European cloud consumer's data that is located in the U.S. can be more easily accessed by government agencies (due to the U.S. Patriot Act) when compared to data located in many European Union countries.

Most regulatory frameworks recognize that cloud consumer organizations are ultimately responsible for the security, integrity, and storage of their own data, even when it is held by an external cloud provider.

Summary of Key Points

- Cloud environments can introduce distinct security challenges, some of which pertain to overlapping trust boundaries imposed by a cloud provider sharing IT resources with multiple cloud consumers.
- A cloud consumer's operational governance can be limited within cloud environments due to the control exercised by a cloud provider over its platforms.
- The portability of cloud-based IT resources can be inhibited by dependencies upon proprietary characteristics imposed by a cloud.
- The geographical location of data and IT resources can be out of a cloud consumer's control when hosted by a third-party cloud provider. This can introduce various legal and regulatory compliance concerns.