Introduction to Cloud computing

Origins and influences

- 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 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.

-

Definitions

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

- National Institute of Standards and Technology (NIST)

- 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 ef ort or service provider interaction.

- This cloud model is composed of five essential characteristics, three service models, and four deployment models

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

Business Drivers

- 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

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

- 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

- 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.
- The virtualization process severs this software-hardware dependency, as hardware requirements can be simulated by emulation software running in virtualized environments

- .
- As cloud computing evolved, a generation of modern virtualization technologies emerged to overcome the performance, reliability, and scalability limitations of traditional virtualization platforms

Technology Innovations vs. Enabling Technologies

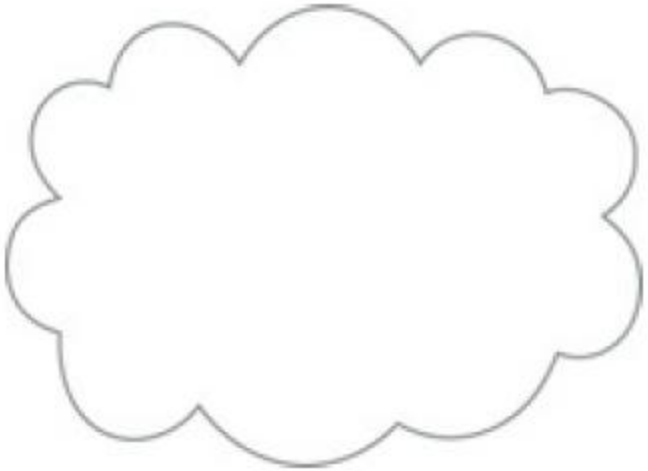
- Broadband Networks and Internet Architecture
- Data Center Technology
- (Modern) Virtualization Technology
- Web Technology
- Multitenant Technology
- Service Technology

Each of these cloud-enabling technologies existed in some form prior to the formal advent of cloud computing

- Basic Concepts and Terminology

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.
- A network of networks providing remote access to a set of decentralized IT resources.



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

Cloud Vs Internet

- A cloud has a finite boundary.
- There are many individual clouds that are accessible via the Internet.
- 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.

- A cloud can be based on the use of any protocols that allow for the remote access to its IT resources

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

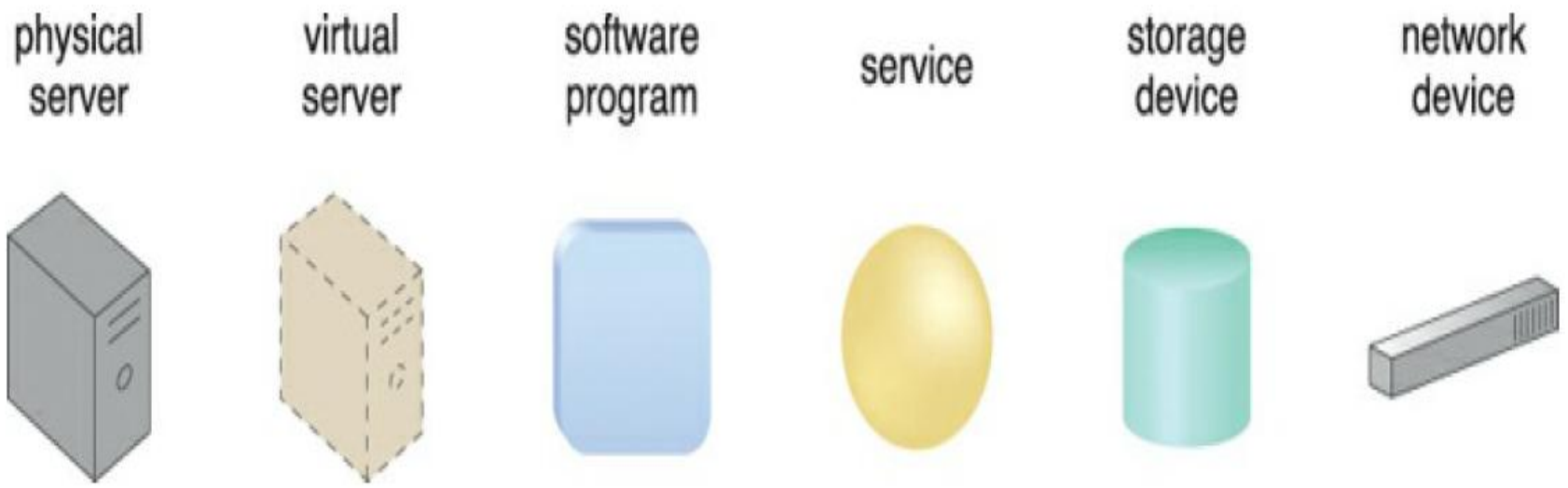
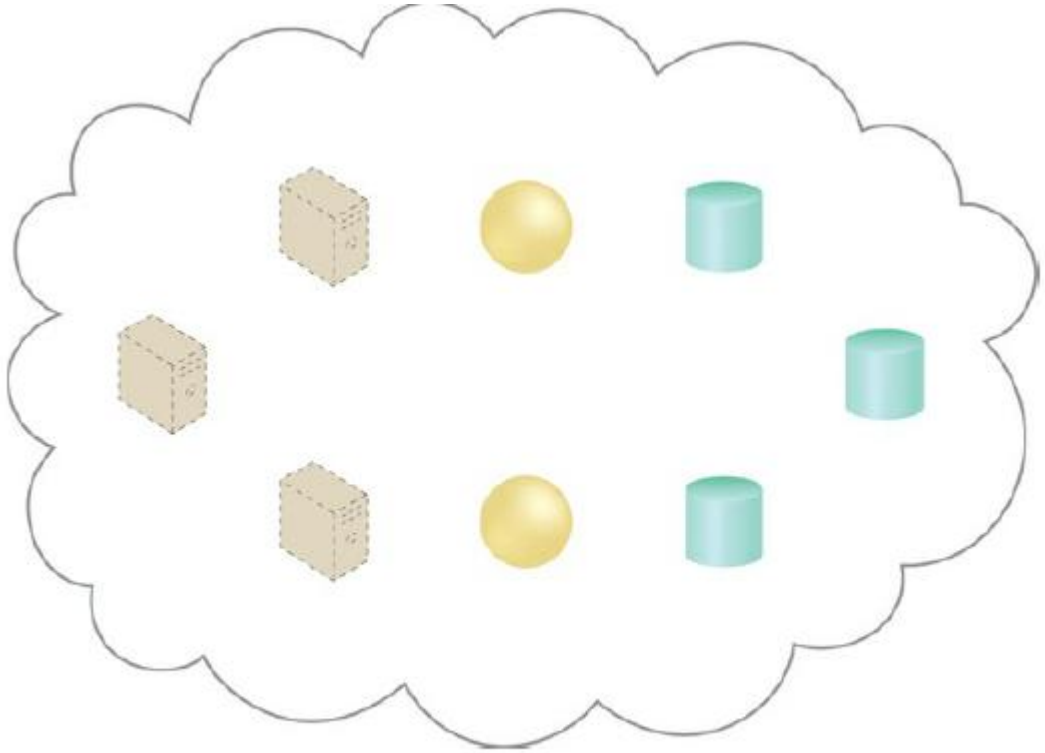


Figure 3.2 Examples of common IT resources and their corresponding symbols.



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

On-Premise

- 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.

- 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.

Cloud Consumers and Cloud Providers

- The party that provides cloud-based IT resources is the *cloud provider.*
- *The party that uses cloudbased* IT resources is the *cloud consumer.*
- *These terms represent roles usually assumed by* organizations in relation to clouds and corresponding cloud provisioning contracts

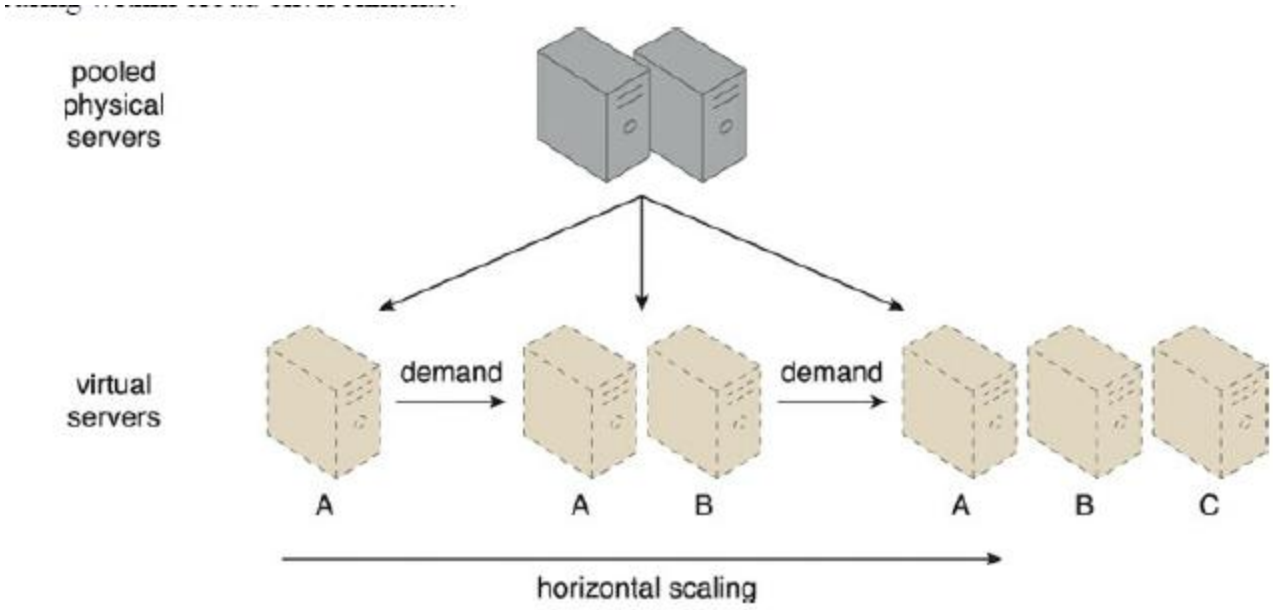
Scaling

- 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*

Horizontal Scaling

- The allocating or releasing of IT resources that are of the same type is referred to as *horizontal scaling.*
- *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

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
- 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

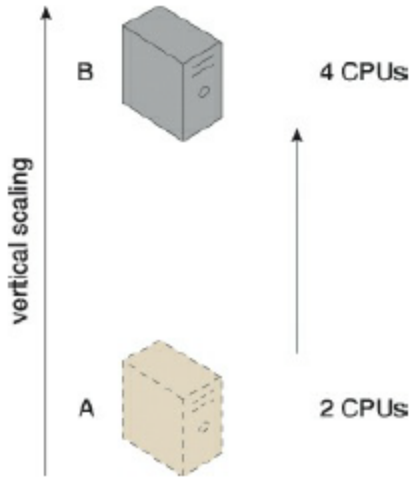


Figure 3.5 An IT resource (a virtual server with two CPUs) is scaled up by replacing it with a more powerful IT resource with increased capacity for data storage (a physical server with four CPUs).

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

Table 3.1 A comparison of horizontal and vertical scaling.

Cloud Service

- 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 resource

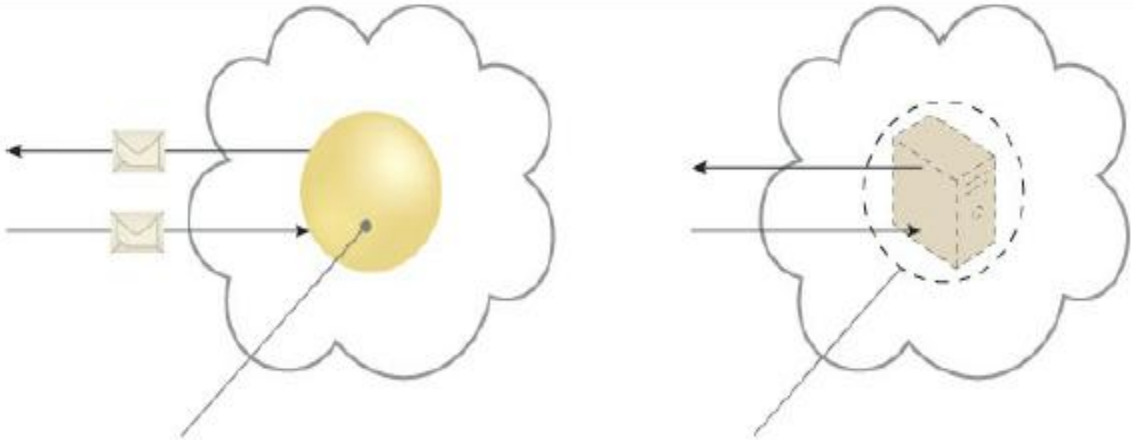


Figure 3.6 A cloud service with a published technical interface is being accessed by a consumer outside of the cloud (left). A cloud service that exists as a virtual server is also being accessed from outside of the cloud's boundary (right). The cloud service on the left is likely being invoked by a 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.

Cloud Service Consumer

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

- 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).

Goals and Benefits

- **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

- 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

- 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

- 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

Common measurable benefits

- 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.

- **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, ondemand or via the cloud consumer’s direct configuration.

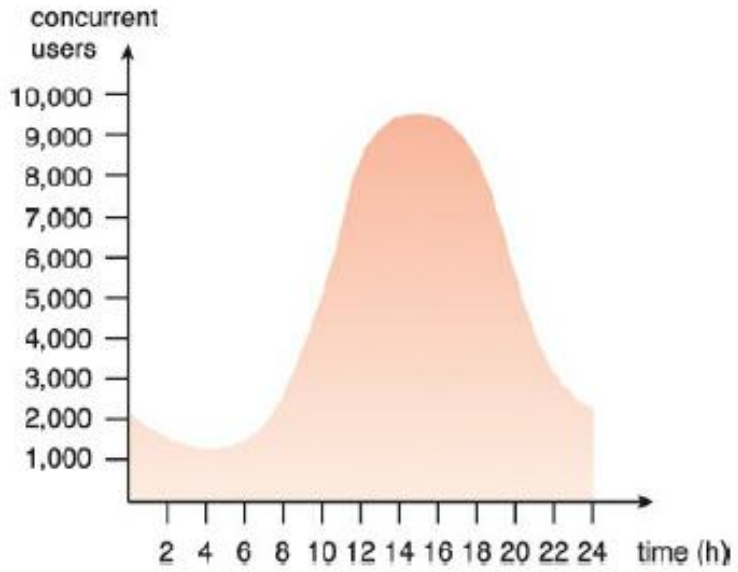


Figure 3.8 An example of an organization's changing demand for an IT resource over the course of a day.

- **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.

Concepts and Models

- **Roles and Boundaries**
- Organizations and humans can assume different types of pre-defined roles depending on how they relate to and/or interact with a cloud and its hosted IT resources.
- Each of the upcoming roles participates in and carries out responsibilities in relation to cloud-based activity.

- **Cloud Provider**
- The organization that provides cloud-based IT resources is the *cloud provider.*

- *When assuming the* role of cloud provider, an organization is responsible for making cloud services available to cloud consumers, as per agreed upon SLA guarantees.

- The cloud provider is further tasked with any required management and administrative duties to ensure the on-going operation of the overall cloud infrastructure.

- Cloud providers normally own the IT resources that are made available for lease by cloud consumers; however, some cloud providers also “resell” IT resources leased from other cloud providers

- **Cloud Consumer**
- A *cloud consumer is an organization (or a human) that has a formal contract or arrangement with a* cloud provider to use IT resources made available by the cloud provider.
- Specifically, the cloud consumer uses a cloud service consumer to access a cloud service

A *cloud consumer* is an organization (or a human) that has a formal contract or arrangement with a cloud provider to use IT resources made available by the cloud provider. Specifically, the cloud consumer uses a cloud service consumer to access a cloud service (Figure 4.1).

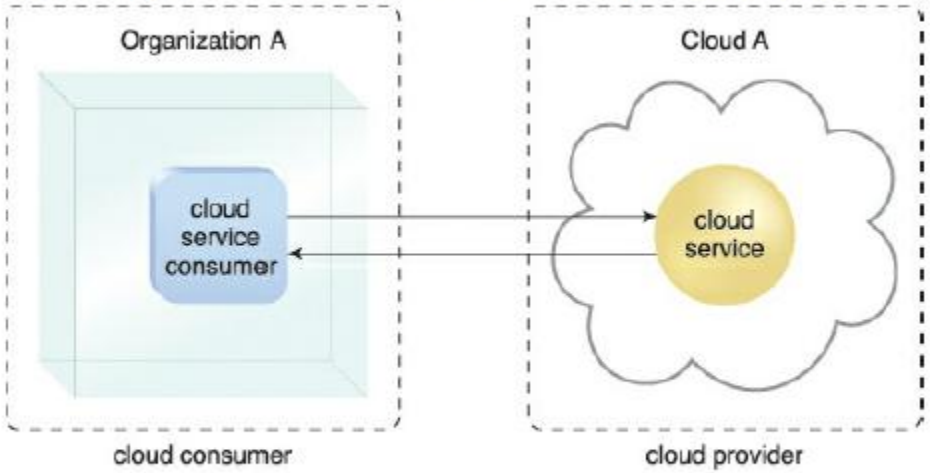
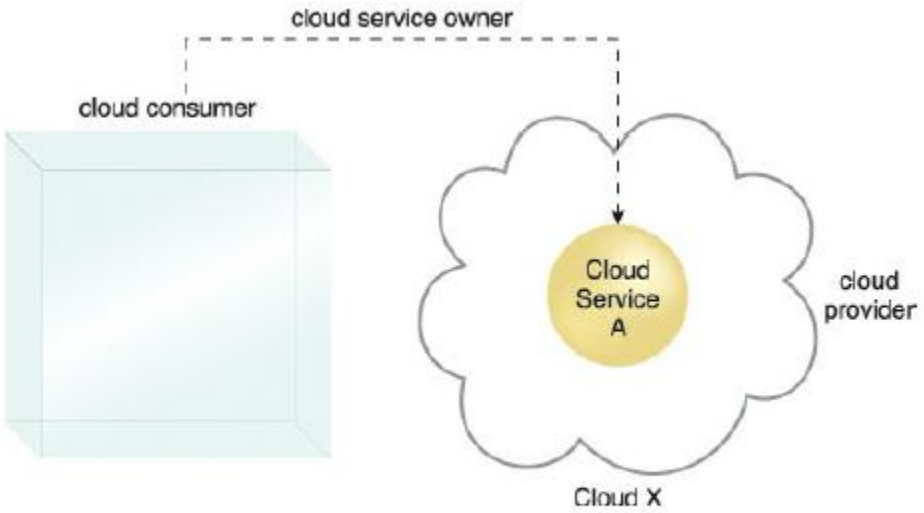
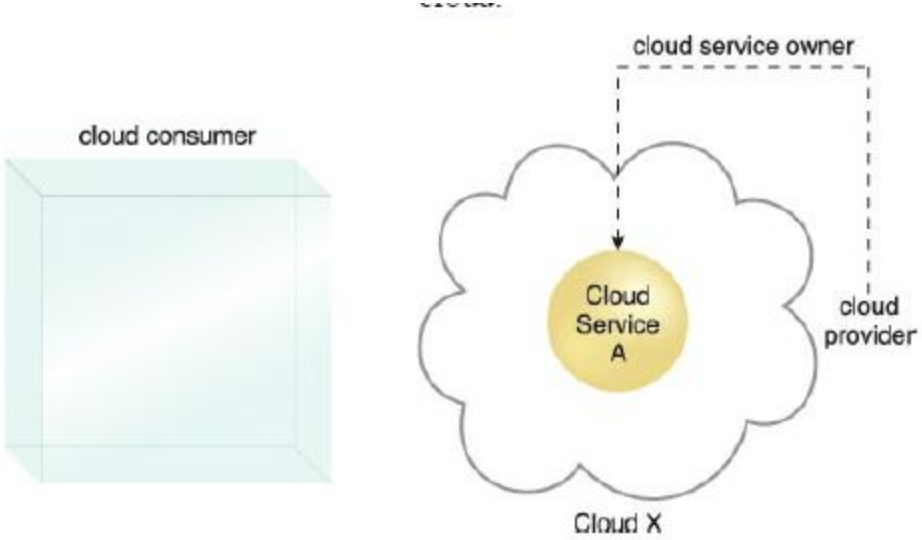


Figure 4.1 A cloud consumer (Organization A) interacts with a cloud service from a cloud provider (that owns Cloud A). Within Organization A, the cloud service consumer is being used to access the cloud service.

- **Cloud Service Owner**
- The person or organization that legally owns a cloud service is called a *cloud service owner. The* cloud service owner can be the cloud consumer, or the cloud provider that owns the cloud within which the cloud service resides



A cloud consumer can be a cloud service owner when it deploys its own service in a cloud.



3 A cloud provider becomes a cloud service owner if it deploys its own cloud service, typically for other cloud consumers to use.

- Several cloud consumer organizations develop and deploy cloud services in clouds owned by other parties for the purpose of making the cloud services available to the general public.
- The reason a cloud service owner is not called a cloud resource owner is because the cloud service owner role only applies to cloud services

- **Cloud Resource Administrator**
- A *cloud resource administrator is the person or organization responsible for administering a cloudbased* IT resource (including cloud services).
- The cloud resource administrator can be (or belong to)
- the cloud consumer or cloud provider of the cloud within which the cloud service resides.
- Alternatively, it can be (or belong to) a third-party organization contracted to administer the cloudbased
- IT resource.

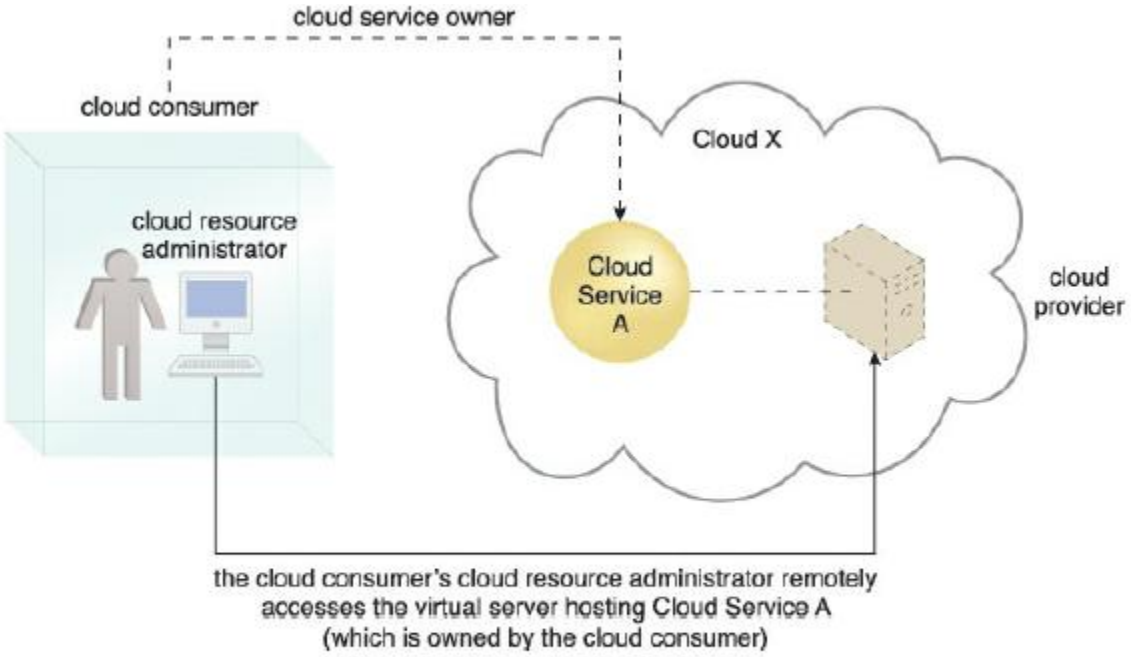


Figure 4.4 A cloud resource administrator can be with a cloud consumer organization and administer remotely accessible IT resources that belong to the cloud consumer.

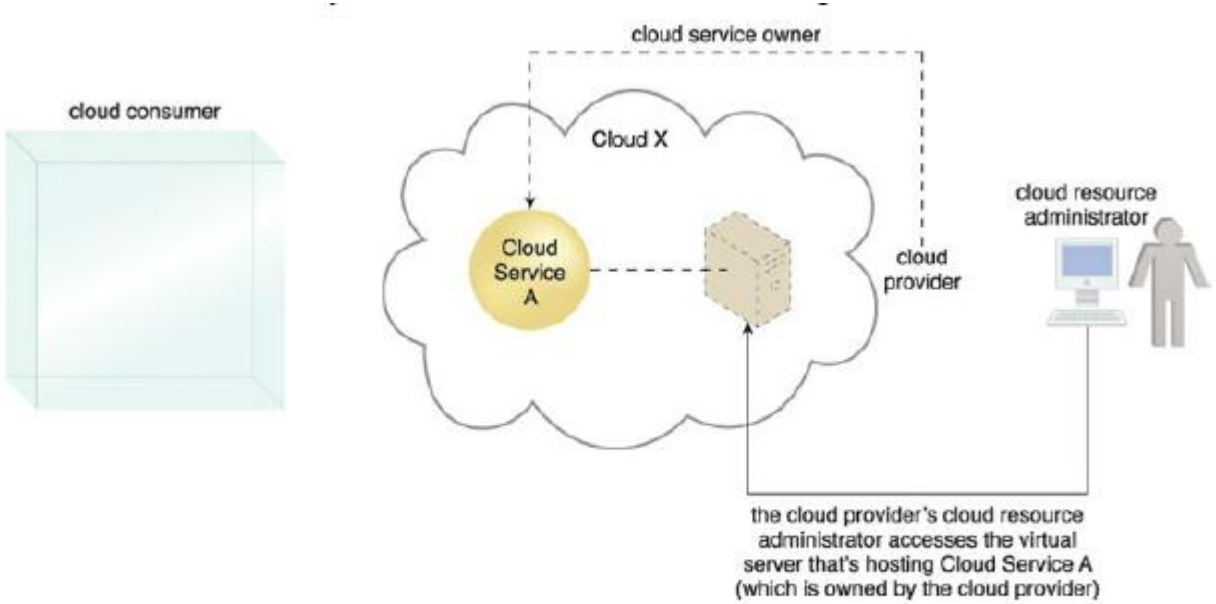


Figure 4.5 A cloud resource administrator can be with a cloud provider organization for which it can administer the cloud provider's internally and externally available IT resources.

- The reason a cloud resource administrator is not referred to as a “cloud service administrator” is
- because this role may be responsible for administering cloud-based IT resources that don’t exist as cloud services.
- For example, if the cloud resource administrator belongs to (or is contracted by) the cloud provider, IT resources not made remotely accessible may be administered by this role (and
- these types of IT resources are not classified as cloud services).

Additional Roles

- *Cloud Auditor – A third-party (often accredited) that conducts independent assessments of* cloud environments assumes the role of the *cloud auditor.*
- *The typical responsibilities* associated with this role include the evaluation of security controls, privacy impacts, and performance.
- The main purpose of the cloud auditor role is to provide an unbiased assessment (and possible endorsement) of a cloud environment to help strengthen the trust relationship between cloud consumers and cloud providers.
- .

- • *Cloud Broker – This role is assumed by a party that assumes the responsibility of managing and* negotiating the usage of cloud services between cloud consumers and cloud providers.
- Mediation services provided by *cloud brokers include service intermediation, aggregation, and* arbitrage.

- *Cloud Carrier – The party responsible for providing the wire-level connectivity between*
- cloud consumers and cloud providers assumes the role of the *cloud carrier. This role is often*
- assumed by network and telecommunication providers

Organizational Boundary

- An *organizational boundary represents the physical perimeter that surrounds a set of IT resources* that are owned and governed by an organization.
- The organizational boundary does not represent the boundary of an actual organization, only an organizational set of IT assets and IT resources.
- Similarly, clouds have an organizational boundary

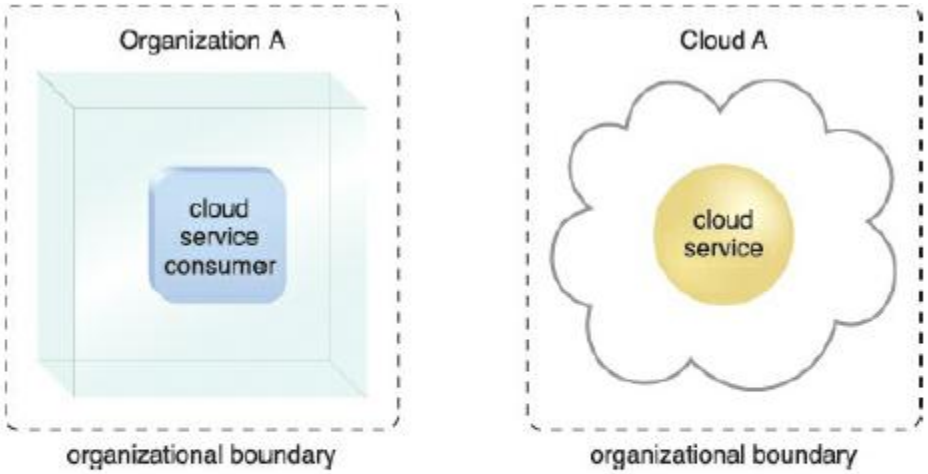


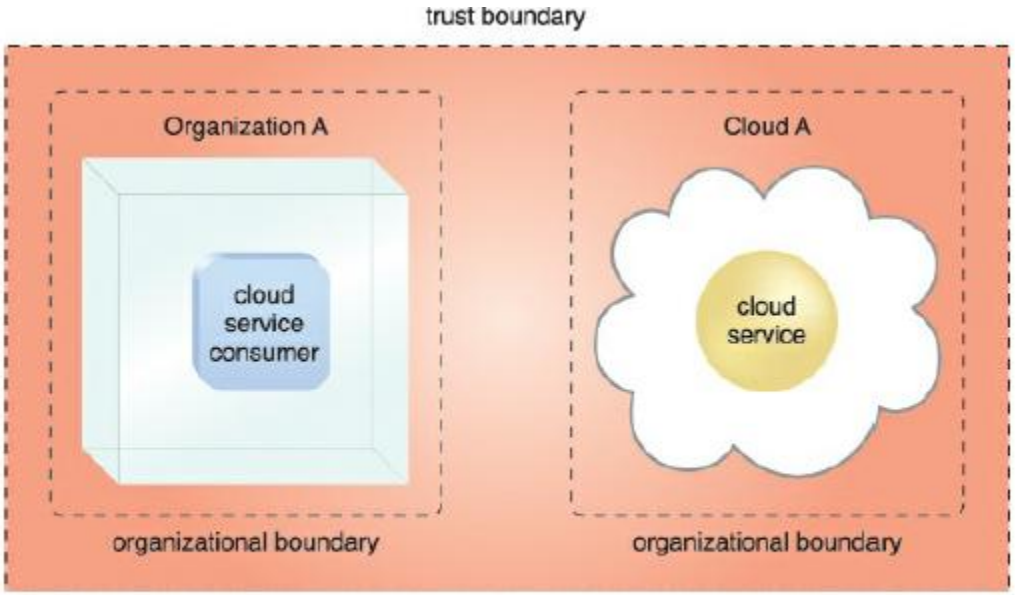
Figure 4.6 Organizational boundaries of a cloud consumer (left), and a cloud provider (right), represented by a broken line notation.

Trust Boundary

- When an organization assumes the role of cloud consumer to access cloud-based IT resources, it needs to extend its trust beyond the physical boundary of the organization to include parts of the cloud environment.

- A *trust boundary is a logical perimeter that typically spans beyond physical boundaries to represent* the extent to which IT resources are trusted .

- When analyzing cloud environments, the trust boundary is most frequently associated with the trust issued by the organization acting as the cloud consumer.



gure 4.7 An extended trust boundary encompasses the organizational boundaries of the cloud provider and the cloud consumer.

Cloud Characteristics

- on-demand usage
- ubiquitous access
- multitenancy (and resource pooling)
- elasticity
- measured usage
- resiliency

on-demand usage

- A cloud consumer can unilaterally access cloud-based IT resources giving the cloud consumer the freedom to self-provision these IT resources.
- Once configured, usage of the self-provisioned IT resources can be automated, requiring no further human involvement by the cloud consumer or cloud provider

- *Ubiquitous access represents the ability for a cloud service to be widely accessible.*

- *Establishing* ubiquitous access for a cloud service can require support for a range of devices, transport protocols, interfaces, and security technologies.

- To enable this level of access generally requires that the cloud service architecture be tailored to the particular needs of different cloud service consumers

Multitenancy

- The characteristic of a software program that enables an instance of the program to serve different consumers (tenants) whereby each is isolated from the other, is referred to as *multitenancy.*
- *A cloud* provider pools its IT resources to serve multiple cloud service consumers by using multitenancy models that frequently rely on the use of virtualization technologies.

- Resource pooling allows cloud providers to pool large-scale IT resources to serve multiple cloud consumers.

- Different physical and virtual IT resources are dynamically assigned and reassigned according to cloud consumer demand, typically followed by execution through statistical multiplexing.

- Resource pooling is commonly achieved through multitenancy technology, and therefore encompassed by this multitenancy characteristic

Figures 4.8 and 4.9 illustrate the difference between single-tenant and multitenant environments.

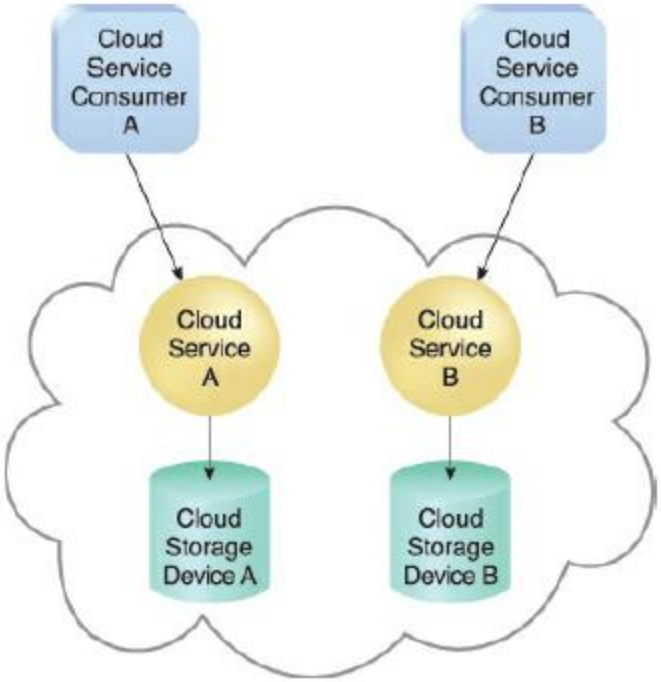


Figure 4.8 In a single-tenant environment, each cloud consumer has a separate IT resource instance.

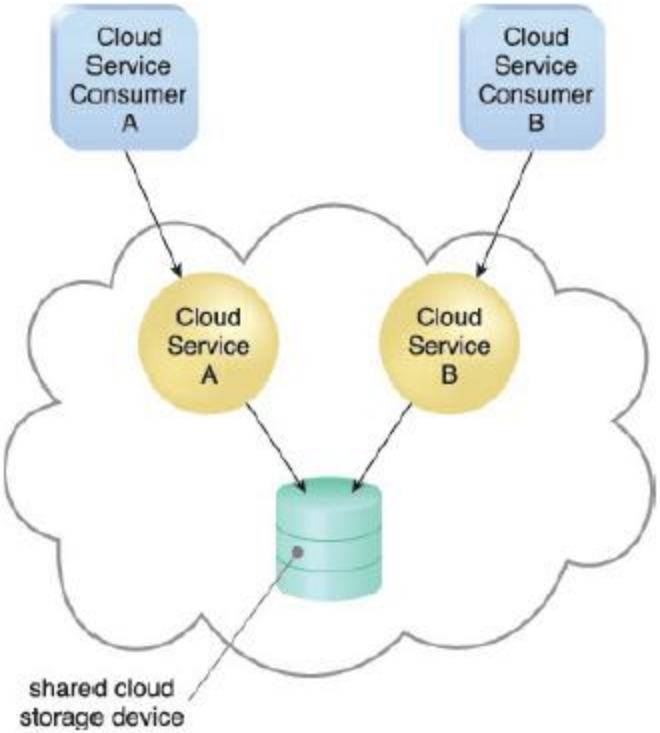


Figure 4.9 In a multitenant environment, a single instance of an IT resource, such as a cloud storage device, serves multiple consumers.

Elasticity

- *Elasticity is the automated ability of a cloud to transparently scale IT resources, as required in* response to runtime conditions or as pre-determined by the cloud consumer or cloud provider.

- Elasticity is often considered a core justification for the adoption of cloud computing, primarily due to the fact that it is closely associated with the Reduced Investment and Proportional Costs benefit.

- Cloud providers with vast IT resources can offer the greatest range of elasticity.

Measured Usage

- The *measured usage characteristic represents the ability of a cloud platform to keep track of the* usage of its IT resources, primarily by cloud consumers.

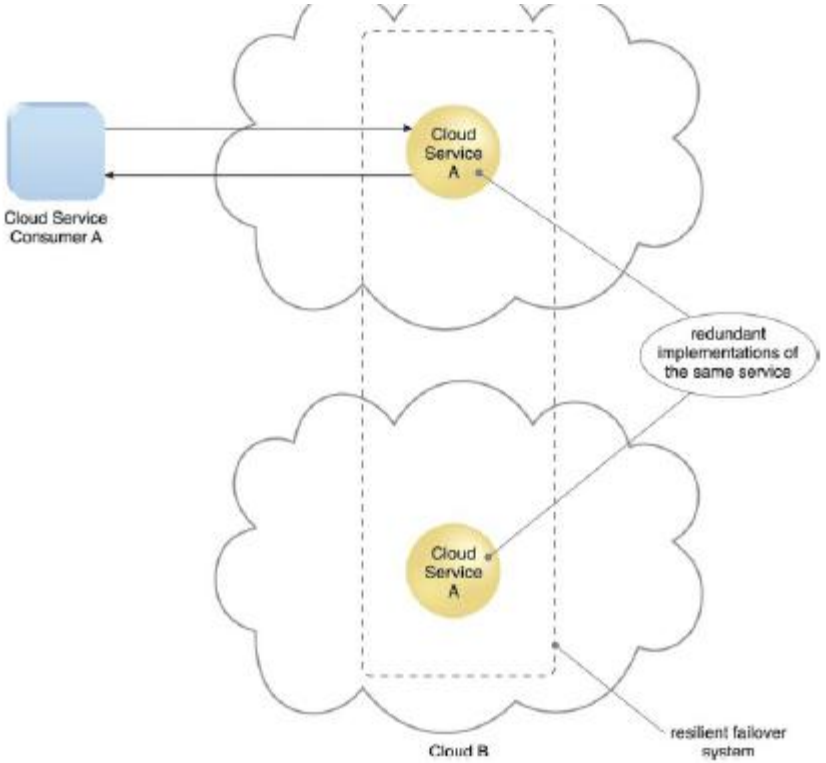
- Based on what is measured, the cloud provider can charge a cloud consumer only for the IT resources actually used and/or for the timeframe during which access to the IT resources was granted. In this context, measured usage is closely related to the on-demand characteristic

Resiliency

- Resilient computing is a form of failover that distributes redundant implementations of IT resources across physical locations.

- IT resources can be pre-configured so that if one becomes deficient, processing is automatically handed over to another redundant implementation.

- Within cloud computing, the characteristic of *resiliency can refer to redundant IT resources within the same cloud* (but in different physical locations) or across multiple clouds



igure 4.10 A resilient system in which Cloud B hosts a redundant implementation of Cloud Service A to provide failover in case Cloud Service A on Cloud A becomes unavailable.