Cloud computing - Fundamental concepts and models

Date @November 12, 2024

Understanding cloud computing: origins, influences, benefits

History

The first time the term cloud was used was in 1996 during the Compaq meeting. Compaq bought a start-up company which supports the idea that data and application will be in Internet and not in the personal PCs.

Only in 2006 the term “cloud computing” emerged in the commercial arena.

Definition

The National Institute of Standards and Technology (NIST) has given the following definition: cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.

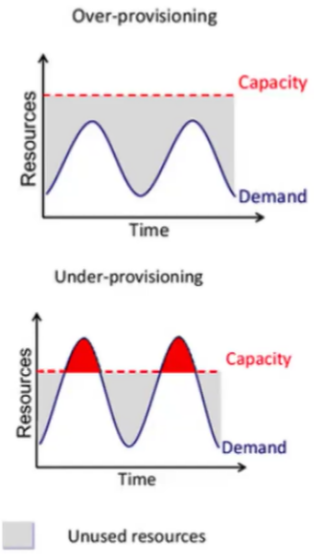
Business drivers

Capacity planning

Capacity corresponds to the maximum number of work that an IT resource is able to do in a given period of time. Capacity planning is the process of planning the IT resources in such a way to use all the capacity of those resources to satisfy some service. It aims to minimize the discrepancy of available resources vs demand.

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):

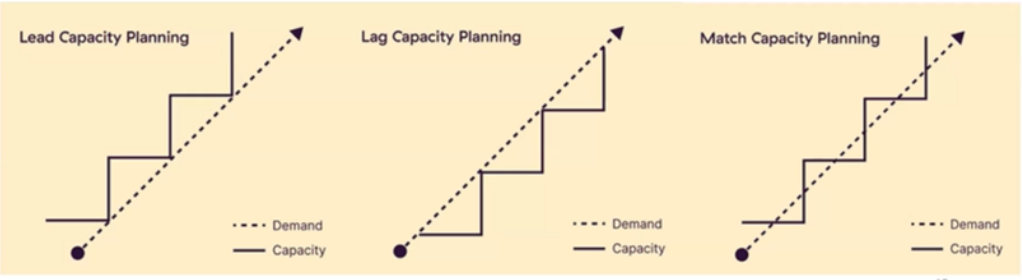
- over-provisioning is costly due to too much infrastructures
- under-provisioning is costly due to potential for business loss from poor quality of service



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.

Different capacity planning strategies exist:

- lead strategy: add capacity to an IT resource in anticipation of demand
- lag strategy: add capacity when the IT resource reaches its full capacity
- match strategy: add IT resource capacity in small increments, as demand increases



Cost reduction

Two costs need to be accounted for:

- the cost of acquiring new infrastructure
- the cost of its ongoing ownership

Much of the required investment is directed into infrastructure expansion because the usage potential of a given automation solution will always be limited by the processing power of its underlying infrastructure.

Operational overhead

Operational overhead represents a considerable share of IT budgets, often exceeding up-front investments costs.

Common forms of infrastructure-related operating overhead include:

- technical personnel to maintain physical IT infrastructure
- system upgrades, patches that add testing to deployment cycles
- utility bills, capital investments for power and cooling
- security and access control measures for server rooms
- admin and accounting staff to track licenses, support agreements, purchases

Organizational agility

Nowadays businesses must be able to react to any unforeseen event, to adapt and to evolve. Organization agility is the measure of an organization's responsiveness to change.

Cloud computing benefits

Reduced investments and proportional costs

The primary economic justification for investing in cloud-based IT resources is the reduction—or complete elimination—of up-front IT investments, particularly hardware and software purchases and their associated ownership costs.

Minimizing up-front financial commitments enables businesses to start small and scale their IT resources as needed. This reduction in initial capital expenses also frees up funds for core business investments.

Increased scalability

Clouds can instantly and dynamically allocate IT resources to cloud consumers, on-demand or via the cloud consumer's direct configuration: on-demand access to pay-as-you-go computing resources on a short-term basis, 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.

Cloud consumers can scale their cloud-based IT resources automatically or manually to accommodate processing fluctuations and peaks. They have the ability to add or remove IT resources at a fine-grained level—for example, modifying available storage disk space in single gigabyte increments.

Increased availability and reliability

Cloud brings greater flexibility and mobility. In fact, using the cloud, companies can instantly access their accounts through any device anytime and anywhere; moreover,

data can be stored, downloaded, restored or processed easily, saving a lot of time and effort.

The cloud environment has the 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.

Basic concepts and terminology

On-premise infrastructure vs. Cloud-based

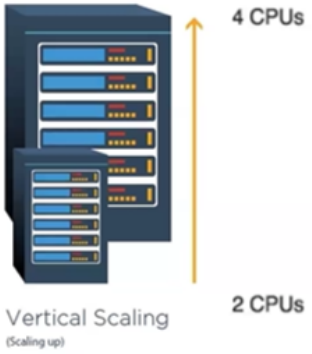
An on-premise infrastructure is the traditional IT infrastructure that is both installed and managed locally by the business itself. A cloud-based infrastructure is an infrastructure that is accessible remotely. An on-premise resource may access and interact with a cloud-based one, or it can be moved 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.

Scaling

Scaling, from an IT resource point of view, means improving the capacity of an IT infrastructure, and there are two main ways:

- vertical scaling
- horizontal scaling

The idea of vertical scaling is that in order to improve the capacity of an IT infrastructure you replace an already existing IT resource with the resource that has higher capacity. This scaling is also called scaling up. When you replace a resource with one with lower capacity we talk about scaling down. Vertical scaling is less common in cloud environments due to the downtime required while the replacement is taking place.



The idea of horizontal scaling, which is also referred to as scaling out, is that existing resources are not replaced but you add new hardware with the same capacity as the

existing ones. When releasing resources it is named scaling in.

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

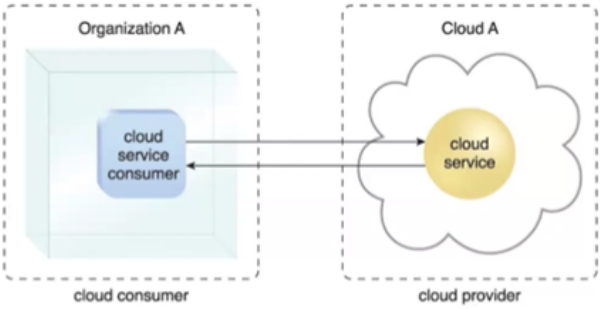
A cloud service is any IT resource that is made remotely accessible via a cloud and it can exist as a simple web-based software program with a technical interface invoked via the use of a messaging protocol, or as a remote access point for administrative tools or larger environments and other IT resources. The cloud service consumer is a temporary runtime role assumed by a software program when it accesses a cloud service.

Cloud provider

A cloud provider is the organization that provides cloud-based IT resources. When assuming the role of cloud provider, an organization is responsible for making cloud services available to cloud consumers. The cloud provider is 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, but some of them also resell IT resources.

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 porivder. Specifically, the cloud consumer uses a cloud service consumer to access a cloud service.



So the cloud consumer is the organization who would like to have some resources, while the cloud service consumer is the software program used by the cloud consumer to interact with the service throughout the computer network.

Service level agreements

SLAs are used in different industries to establish a trust relationship between service providers and consumers. The SLA details the service-level capabilities promised by the providers to be delivered and requirements/expectations stated by consumers. These details are very important as the document establishes a legal binding for both the parties and work as reference in any dispute.

Cloud service owner

The cloud service owner is the entity that owns a cloud service from a legal point of view. Most of the time the owner of the cloud service is the cloud provider itself, but he can also be the cloud consumer.

Cloud resource administrator

A cloud resource administrator is the entity responsible for administering a cloud-based IT resource. He can be the cloud consumer or the cloud provider as well. Or he may be a third-party organization contracted to administer the cloud-based IT resource.

Additional roles

In cloud there are other three roles, essential to ensure that everything works as expected:

- cloud broker
- cloud auditor

- cloud carrier

Cloud broker

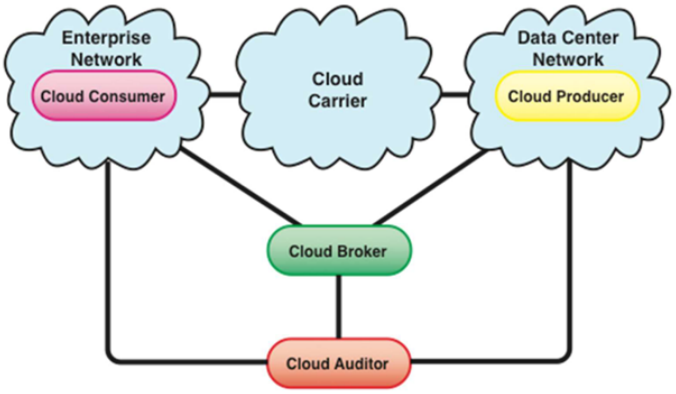
The cloud broker role is assumed by a party that assumes the responsibility of managing and negotiating the usage of cloud services between cloud consumers and cloud providers. He can provide some mediation services such as aggregation and arbitrage.

Cloud auditor

The cloud auditor is a third-party who conducts independent assessments of cloud environments. His responsibilities are the evaluation of security controls, privacy impacts and performance. The main purpose of the cloud auditor role is to provide an unbiased assessment of a cloud environment to help strengthen the trust relationship between cloud consumers and cloud providers.

Cloud carrier

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



Cloud characteristics

There are some characteristics which are common to the most of cloud environments.

On-demand usage

Once configured, usage of the self-provisioned IT resources can be automated, requiring no further human involvement by the cloud consumer or cloud provider. This results in an on-demand usage environment. Also known as “on-demand self service usage,” this characteristic enables the service-based and usage-driven features found in mainstream clouds.

Ubiquitous access

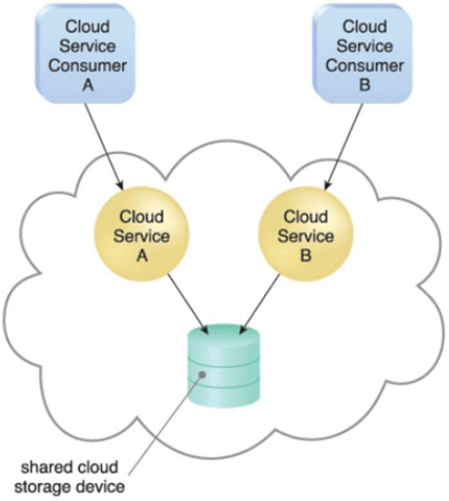
Ubiquitous access represents the ability for a cloud service to be widely accessible. Establishing ubiquitous access for a cloud service can require support for a range of devices, transport protocols, interfaces and security technologies. To enable this level of access generally requires the cloud service architecture to be tailored to the particular needs of different cloud service consumers.

Multitenancy (or resource pooling)

A multitenancy software is a software program that enables an instance of the program to
serve different consumers (tenants) whereby each is isolated from the other.

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.

Through the use of multitenancy technology, IT resources can be dynamically assigned and reassigned, according to cloud service consumer demands.



In a multitenant environment, a single instance of an IT resource, such as a cloud storage device

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. It 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

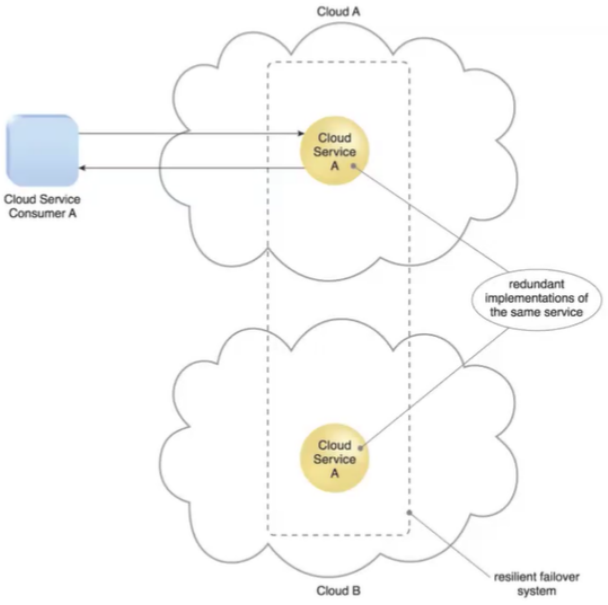
The measured usage characteristic represents the ability of a cloud platform to keep track of the usage of its IT resources. Based on what is measured, the cloud provider can charge a cloud consumer only for the IT resources actually used and/or for the timeframe during which access to the IT resources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

Measured usage also includes the general monitoring of IT resources and related usage reporting. Therefore, measured usage is also relevant to clouds that do not charge for usage.

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.

Cloud consumers can increase both the reliability and availability of their applications by leveraging the resiliency of cloud-based IT resources.



Cloud deployment models

A deployment model specifies how a cloud-based system works in terms of ownership, size and access. You may have four cloud deployment models:

- public
- private

- community
- hybrid

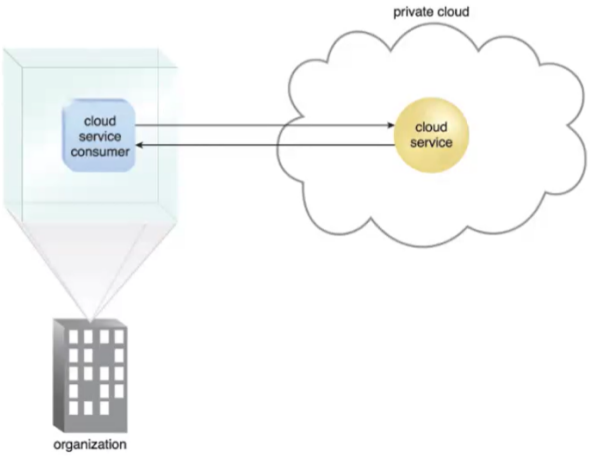
Public clouds

A public cloud is a publicly accessible cloud environment owned by a third-party cloud provider. The IT resources on public clouds are generally offered to cloud consumers at a cost or are commercialized via other avenues, such as advertisement. The cloud provider is responsible for the creation and on-going maintenance of the public cloud and its IT resources. A public cloud may be owned, managed and operated by a business, academic or government organization, or some combination of them.

Private clouds

With a private cloud, the same organization is technically both the cloud consumer and cloud provider. In order to differentiate these roles:

- a separate organizational department typically assumes the responsibility for provisioning the cloud and assumes the cloud provider role
- departments requiring access to the private cloud assume the cloud consumer role



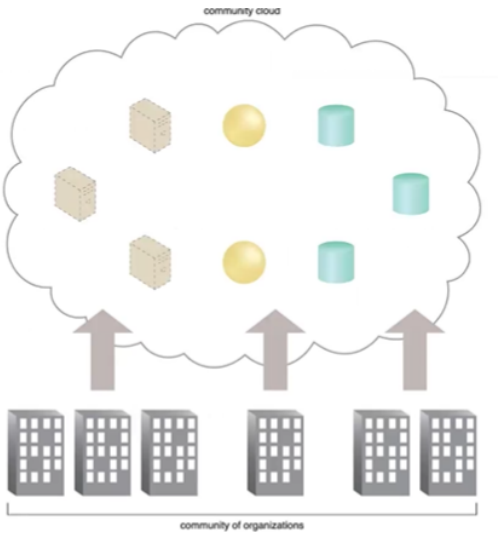
A private cloud is owned by a single organization and private clouds enable an organization to use computing technology as a means of centralizing access to IT resources by different parts, locations or departments of the organization.

It is important to use the terms “on-premise” and “cloud-based” correctly within the context of a private cloud. Even though the private cloud may physically reside on the organization’s premises, IT resources it hosts are still considered “cloud-based” as long as they are made remotely accessible to cloud consumers. IT resources hosted outside of the private cloud by the departments acting as cloud consumers are therefore considered “on-premise” in relation to the private cloud-based IT resources.

Community clouds

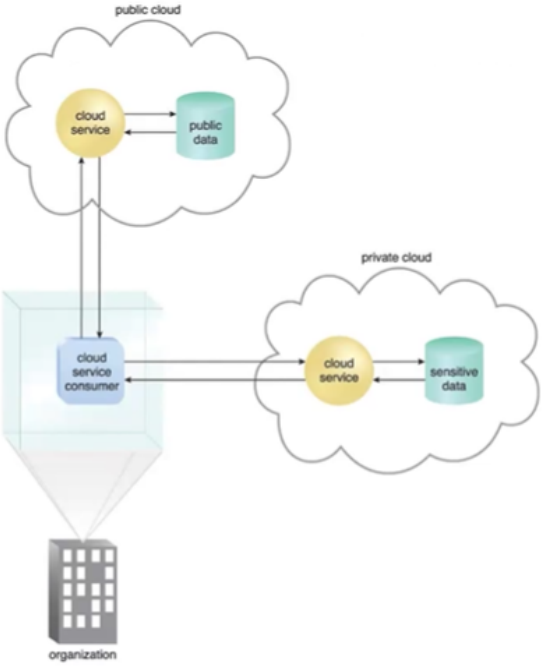
A community cloud shares characteristics of private and public clouds. Like a public cloud, the cloud resources are shared among a number of independent organizations, and like a private cloud, a community cloud has restricted access. The community cloud may be jointly owned by the community members or by a third-party cloud provider that provisions a public cloud with limited access. The member cloud consumers of the community typically share the responsibility for defining and evolving the community cloud.

Membership in the community does not necessarily guarantee access to or control of all the cloud’s IT resources. Parties outside the community are generally not granted access unless allowed by the community.



Hybrid clouds

A hybrid cloud is a cloud environment comprised of two or more different cloud deployment models. For example, a cloud consumer may choose to deploy cloud service processing sensitive data to a private cloud and other, less sensitive cloud services to a public cloud. Hybrid deployment architectures can be complex and challenging to create and maintain due to the potential disparity in cloud environments and the fact that management responsibilites are typically split between the private cloud provider organization and the public cloud provider.



	Private	Community	Public	Hybrid
Scalability	Limited	Limited	Very high	Very high
Security	Most secure option	Very secure	Moderately secure	Very secure
Performance	Very good	Very good	Low to medium	Good
Reliability	Very high	Very high	Medium	Medium to high
Cost	High	Medium	Low	Medium

Cloud delivery models

A cloud delivery model represents a specific, pre-packages combination of IT resources offered by a cloud provider. Three common cloud delivery models have become widely established and formalized:

- infrastructure-as-a-service (IaaS)
- platform-as-a-service (PaaS)
- software-as-a-service (SaaS)

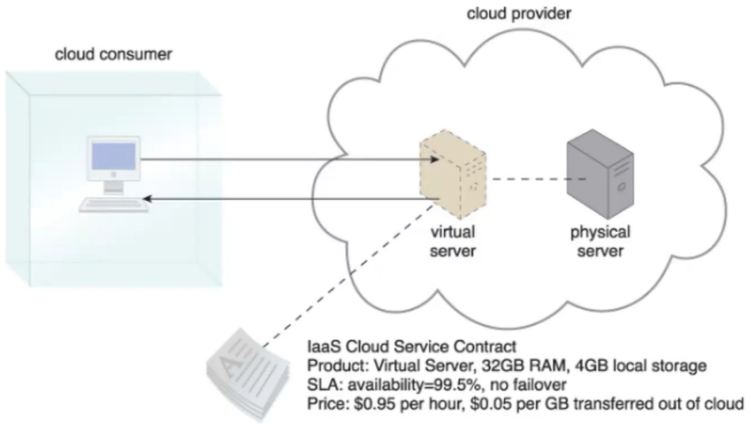
Infrastructure-as-a-service

The IaaS delivery model represents a self-contained IT environment comprised of infrastructure-centric IT resources that can be accessed and managed via cloud service-based interfaces and tools. This environment can include hardware, network, connectivity, operating systems and other raw IT resources.

In contrast to traditional hosting or outsourcing environments, with IaaS, IT resources are typically virtualized and packaged into bundles that simplify up-front runtime scaling and customization of the infrastructure.

The main advantage of IaaS is that the consumer has an extremely high level of freedom in the configuration of the resources. After getting access to the virtualized hardware he can control them as he wants. IaaS provides cloud consumers with a high level of control and responsibility over its configuration and utilization.

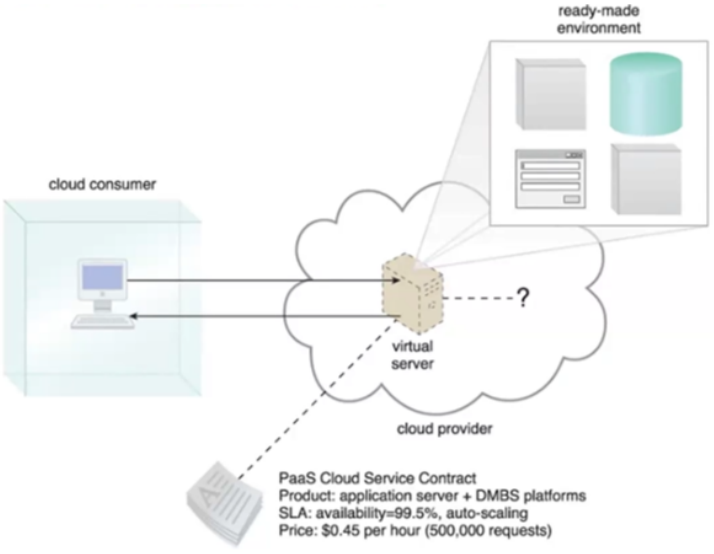
The IT resources provided by IaaS are generally not pre-configured, placing the administrative responsibility directly upon the cloud consumer. This means that the cloud consumer has to patch the OS with the latest security vulnerabilities, he has to enforce security and he has to configure network and intermediate middleboxes.



Platform-as-a-service

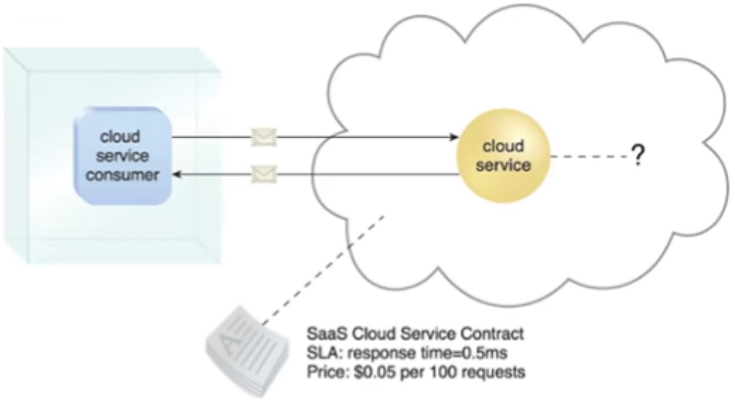
The PaaS delivery model represents a pre-defined "ready-to-use" environment typically comprised of already deployed and configured IT resources. By working within a ready-made platform, the cloud consumer avoids the administrative burden of setting up and maintaining the basic infrastructure IT resources typically provided through the IaaS model. However, the cloud consumer is granted a lower level of control over the underlying IT resources that host and provision the platform.

The platform handles deploying services across one or multiple servers, abstracting details like operating systems, network connectivity, and automatic scaling. In this model, users develop and install their apps on the provided infrastructure. While this enables faster software development cycles, it may come with limitations on supported programming languages. Users also lack control over hardware, security models, and operating systems.



Software-as-a-service

SaaS (Software-as-a-Service) offers software programs as shared cloud services, available as products or generic utilities. The SaaS delivery model is typically used to make reusable cloud services widely available—often commercially—to a range of cloud consumers. A thriving marketplace exists for SaaS products, which can be leased and used for various purposes under different terms.



Cloud Delivery Model	Common Cloud Consumer Activities	Common Cloud Provider Activities
SaaS	uses and configures cloud service	implements, manages, and maintains cloud service monitors usage by cloud consumers
PaaS	develops, tests, deploys, and manages cloud services and cloud-based solutions	pre-configures platform and provisions underlying infrastructure, middleware, and other needed IT resources, as necessary monitors usage by cloud consumers
IaaS	sets up and configures bare infrastructure, and installs, manages, and monitors any needed software	provisions and manages the physical processing, storage, networking, and hosting required monitors usage by cloud consumers

Other cloud services

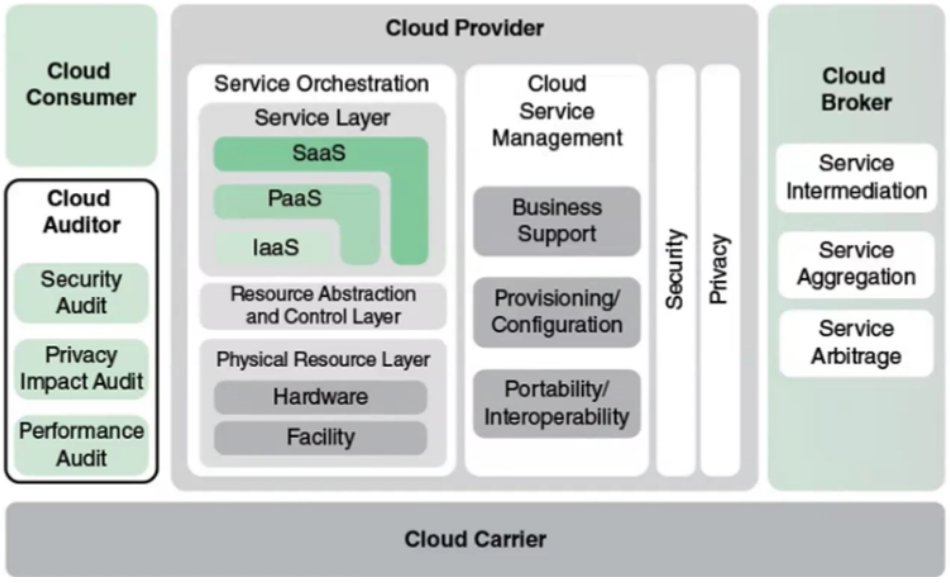
Many specialized variations of the three base cloud delivery models have emerged, each comprised of a distinct combination of IT resources. Some examples include:

- storage-as-a-service
- database-as-a-service
- security-as-a-service
- communication-as-a-service
- integration-as-a-service
- testing-as-a-service
- process-as-a-service

XaaS

XaaS is the latest development in the provisioning of cloud services. The acronym has three generally accepted interpretations, all of which mean essentially the same thing:

- Anything as a Service: Where "anything" refers to any service other than the three traditional cloud services (SaaS, PaaS, IaaS).
- Everything as a Service: This version is somewhat misleading, as no vendor offers every possible cloud service. Instead, it suggests that the cloud service provider is offering a wide range of services. Some providers package together SaaS, PaaS, and IaaS so that customers can do one-stop shopping for the basic cloud services that enterprises increasingly rely on.
- X as a Service: Where "X" can represent any possible cloud service option.



Cloud-enabling technology

Modern-day clouds are built upon a foundation of key technology components that collectively enable the essential features and characteristics of contemporary cloud computing. These fundamental technologies include:

- Broadband networks and internet architecture
- Data center technology
- Virtualization technology
- Web technology
- Multitenant technology
- Containerization