Cloud computing

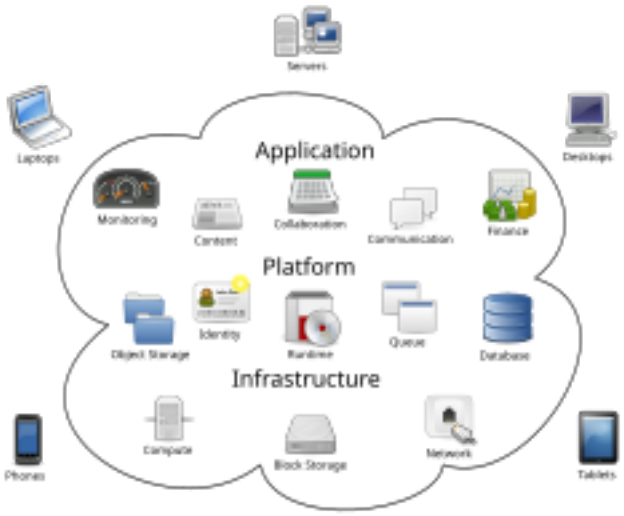

Figure 1: Cloud computing metaphor: the group of networked elements providing services does not need to be addressed or managed individually by users; instead, the entire provider-managed suite of hardware and software can be thought of as an amorphous cloud.

Cloud computing is "a paradigm for enabling network access to a scalable and elastic pool of shareable physical or virtual resources with self-service provisioning and administration on-demand,"according to ISO. It is commonly referred to as "the cloud".

Characteristics

In 2011, the National Institute of Standards and Technology (NIST) identified five "essential characteristics" for cloud systems. Below are the exact definitions according to NIST:

- On-demand self-service: "A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider."
- Broad network access: "Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations)."
- Resource pooling: "The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand."
- Rapid elasticity: "Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear unlimited and can be appropriated in any quantity at any time."
- Measured service: "Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

By 2023, the International Organization for Standardization (ISO) had expanded and refined the list.

History

The history of cloud computing extends to the 1960s, with the initial concepts of time-sharing becoming popularized via remote job entry (RJE). The "data center"model, where users submitted jobs to operators to run on mainframes, was predominantly used during this era. This was a time of exploration and experimentation with ways to make large-scale computing power available to more users through time-sharing, optimizing the infrastructure, platform, and applications, and increasing efficiency for end users.

The "cloud"metaphor for virtualized services dates to 1994, when it was used by General Magic for the universe of "places"that mobile agents in the Telescript environment could "go". The metaphor is credited to David Hoffman, a General Magic communications specialist, based on its long-standing use in networking and telecom. The expression cloud computing became more widely known in 1996 when Compaq Computer Corporation drew up a business plan for future computing and the Internet. The company's ambition was to supercharge sales with "cloud computing-enabled applications". The business plan foresaw that online consumer file storage would likely be commercially successful. As a result, Compaq decided to sell server hardware to internet service providers.

In the 2000s, the application of cloud computing began to take shape with the establishment of Amazon Web Services (AWS) in 2002, which allowed developers to build applications independently. In 2006 Amazon Simple Storage Service, known as Amazon S3, and the Amazon Elastic Compute Cloud (EC2) were released. In 2008 NASA's development of the first open-source software for deploying private and hybrid clouds.

The following decade saw the launch of various cloud services. In 2010, Microsoft launched Microsoft Azure, and Rackspace Hosting and NASA initiated an open-source cloud-software project, OpenStack. IBM introduced the IBM SmartCloud framework in 2011, and Oracle announced the Oracle Cloud in 2012. In December 2019, Amazon launched AWS Outposts, a service that extends AWS infrastructure, services, APIs, and tools to customer data centers, co-location spaces, or on-premises facilities.

Value proposition

Cloud computing can enable shorter time to market by providing pre-configured tools, scalable resources, and managed services, allowing users to focus on their core business value instead of maintaining infrastructure. Cloud platforms can enable organizations and individuals to reduce upfront capital expenditures on physical infrastructure by shifting to an operational expenditure model, where costs scale with usage. Cloud platforms also offer managed services and tools, such as artificial intelligence, data analytics, and machine learning, which might otherwise require significant in-house expertise and infrastructure investment.

While cloud computing can offer cost advantages through effective resource optimization, organizations often face challenges such as unused resources, inefficient configurations, and hidden costs without proper oversight and governance. Many cloud platforms provide cost management tools, such as AWS Cost Explorer and Azure Cost Management, and frameworks like FinOps have emerged to standardize financial operations in the cloud. Cloud computing also facilitates collaboration, remote work, and global service delivery by enabling secure access to data and applications from any location with an internet connection.

Cloud providers offer various redundancy options for core services, such as managed storage and managed databases, though redundancy configurations often vary by service tier. Advanced redundancy strategies, such as cross-region replication or failover systems, typically require explicit configuration and may incur additional costs or licensing fees.

Cloud environments operate under a shared responsibility model, where providers are typically responsible for infrastructure security, physical hardware, and software updates, while customers are accountable for data encryption, identity and access management (IAM), and application-level security. These responsibilities vary depending on the cloud service model—Infrastructure as a Service (IaaS), Platform as a Service (PaaS), or Software as a Service (SaaS)—with customers typically having more control and responsibility in IaaS environments and progressively less in PaaS and SaaS models, often trading control for convenience and managed services.

Adoption and suitability

The decision to adopt cloud computing or maintain on-premises infrastructure depends on factors such as scalability, cost structure, latency requirements, regulatory constraints, and infrastructure customization.

Organizations with variable or unpredictable workloads, limited capital for upfront investments, or a focus on rapid scalability benefit from cloud adoption. Startups, SaaS companies, and e-commerce platforms often prefer the pay-as-you-go operational expenditure (OpEx) model of cloud infrastructure. Additionally, companies prioritizing global accessibility, remote workforce enablement, disaster recovery, and leveraging advanced services such as AI/ML and analytics are well-suited for the cloud. In recent years, some cloud providers have started offering specialized services for high-performance computing and low-latency applications, addressing some use cases previously exclusive to on-premises setups.

On the other hand, organizations with strict regulatory requirements, highly predictable workloads, or reliance on deeply integrated legacy systems may find cloud infrastructure less suitable. Businesses in industries like defense, government, or those handling highly sensitive data often favor on-premises setups for greater control and data sovereignty. Additionally, companies with ultra-low latency requirements, such as high-frequency trading (HFT) firms, rely on custom hardware (e.g., FPGAs) and physical proximity to exchanges, which most cloud providers cannot fully replicate despite recent advancements. Similarly, tech giants like Google, Meta, and Amazon build their own data centers due to economies of scale, predictable workloads, and the ability to customize hardware and network infrastructure for optimal efficiency. However, these companies also use cloud services selectively for certain workloads and applications where it aligns with their operational needs.

In practice, many organizations are increasingly adopting hybrid cloud architectures, combining on-premises infrastructure with cloud services. This approach allows businesses to balance scalability, cost-effectiveness, and control, offering the benefits of both deployment models while mitigating their respective limitations.

Challenges and limitations

One of the main challenges of cloud computing, in comparison to more traditional on-premises computing, is data security and privacy. Cloud users entrust their sensitive data to third-party providers, who may not have adequate measures to protect it from unauthorized access, breaches, or leaks. Cloud users also face compliance risks if they have to adhere to certain regulations or standards regarding data protection, such as GDPR or HIPAA.

Another challenge of cloud computing is reduced visibility and control. Cloud users may not have full insight into how their cloud resources are managed, configured, or optimized by their providers. They may also have limited ability to customize or modify their cloud services according to their specific needs or preferences. Complete understanding of all technology may be impossible, especially given the scale, complexity, and deliberate opacity of contemporary systems; however, there is a need for understanding complex technologies and their interconnections to have power and agency within them. The metaphor of the cloud can be seen as problematic as cloud computing retains the aura of something noumenal and numinous; it is something experienced without precisely understanding what it is or how it works.

Additionally, cloud migration is a significant challenge. This process involves transferring data, applications, or workloads from one cloud environment to another, or from on-premises infrastructure to the cloud. Cloud migration can be complicated, time-consuming, and expensive, particularly when there are compatibility issues between different cloud platforms or architectures. If not carefully planned and executed, cloud migration can lead to downtime, reduced performance, or even data loss.

Cloud migration challenges

According to the 2024 State of the Cloud Report by Flexera, approximately 50% of respondents identified the following top challenges when migrating workloads to public clouds:

1. “Understanding application dependencies”
2. “Comparing on-premise and cloud costs”

3. “Assessing technical feasibility.”

Implementation challenges

Applications hosted in the cloud are susceptible to the fallacies of distributed computing, a series of misconceptions that can lead to significant issues in software development and deployment.

Cloud cost overruns

In a report by Gartner, a survey of 200 IT leaders revealed that 69% experienced budget overruns in their organizations'cloud expenditures during 2023. Conversely, 31% of IT leaders whose organizations stayed within budget attributed their success to accurate forecasting and budgeting, proactive monitoring of spending, and effective optimization.

The 2024 Flexera State of Cloud Report identifies the top cloud challenges as managing cloud spend, followed by security concerns and lack of expertise. Public cloud expenditures exceeded budgeted amounts by an average of 15%. The report also reveals that cost savings is the top cloud initiative for 60% of respondents. Furthermore, 65% measure cloud progress through cost savings, while 42% prioritize shorter time-to-market, indicating that cloud's promise of accelerated deployment is often overshadowed by cost concerns.

Service Level Agreements

Typically, cloud providers'Service Level Agreements (SLAs) do not encompass all forms of service interruptions. Exclusions typically include planned maintenance, downtime resulting from external factors such as network issues, human errors, like misconfigurations, natural disasters, force majeure events, or security breaches. Typically, customers bear the responsibility of monitoring SLA compliance and must file claims for any unmet SLAs within a designated timeframe. Customers should be aware of how deviations from SLAs are calculated, as these parameters may vary by service. These requirements can place a considerable burden on customers. Additionally, SLA percentages and conditions can differ across various services within the same provider, with some services lacking any SLA altogether. In cases of service interruptions due to hardware failures in the cloud provider, the company typically does not offer monetary compensation. Instead, eligible users may receive credits as outlined in the corresponding SLA.

Leaky abstractions

Cloud computing abstractions aim to simplify resource management, but leaky abstractions can expose underlying complexities. These variations in abstraction quality depend on the cloud vendor, service and architecture. Mitigating leaky abstractions requires users to understand the implementation details and limitations of the cloud services they utilize.

Service lock-in within the same vendor

Service lock-in within the same vendor occurs when a customer becomes dependent on specific services within a cloud vendor, making it challenging to switch to alternative services within the same vendor when their needs change.

Security and privacy

Cloud computing poses privacy concerns because the service provider can access the data that is in the cloud at any time. It could accidentally or deliberately alter or delete information. Many cloud providers can share information with third parties if necessary for purposes of law and order without a warrant. That is permitted in their privacy policies, which users must agree to before they start using cloud services. Solutions to privacy include policy and legislation as well as end-users'choices for how data is stored. Users can encrypt data that is processed or stored within the cloud to prevent unauthorized access. Identity management systems can also provide practical solutions to privacy concerns in cloud computing. These systems distinguish between

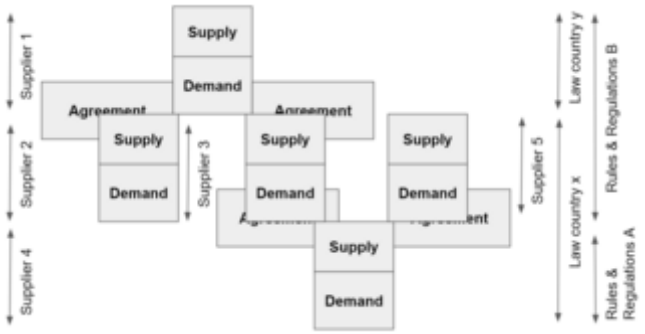

Figure 2: Cloud suppliers security and privacy agreements must be aligned to the demand(s) requirements and regulations.

authorized and unauthorized users and determine the amount of data that is accessible to each entity. The systems work by creating and describing identities, recording activities, and getting rid of unused identities.

According to the Cloud Security Alliance, the top three threats in the cloud are Insecure Interfaces and APIs, Data Loss & Leakage, and Hardware Failure—which accounted for 29%, 25% and 10% of all cloud security outages respectively. Together, these form shared technology vulnerabilities. In a cloud provider platform being shared by different users, there may be a possibility that information belonging to different customers resides on the same data server. Additionally, Eugene Schultz, chief technology officer at Emagined Security, said that hackers are spending substantial time and effort looking for ways to penetrate the cloud. "There are some real Achilles'heels in the cloud infrastructure that are making big holes for the bad guys to get into". Because data from hundreds or thousands of companies can be stored on large cloud servers, hackers can theoretically gain control of huge stores of information through a single attack—a process he called "hyperjacking". Some examples of this include the Dropbox security breach, and iCloud 2014 leak. Dropbox had been breached in October 2014, having over seven million of its users passwords stolen by hackers in an effort to get monetary value from it by Bitcoins (BTC). By having these passwords, they are able to read private data as well as have this data be indexed by search engines (making the information public).

There is the problem of legal ownership of the data (If a user stores some data in the cloud, can the cloud provider profit from it?). Many Terms of Service agreements are silent on the question of ownership. Physical control of the computer equipment (private cloud) is more secure than having the equipment off-site and under someone else's control (public cloud). This delivers great incentive to public cloud computing service providers to prioritize building and maintaining strong management of secure services. Some small businesses that do not have expertise in IT security could find that it is more secure for them to use a public cloud. There is the risk that end users do not understand the issues involved when signing on to a cloud service (persons sometimes do not read the many pages of the terms of service agreement, and just click "Accept" without reading). This is important now that cloud computing is common and required for some services to work, for example for an intelligent personal assistant (Apple's Siri or Google Assistant). Fundamentally, private cloud is seen as more secure with higher levels of control for the owner, however public cloud is seen to be more flexible and requires less time and money investment from the user.

The attacks that can be made on cloud computing systems include man-in-the middle attacks, phishing attacks, authentication attacks, and malware attacks. One of the largest threats is considered to be malware attacks, such as Trojan horses. Recent research conducted in 2022 has revealed that the Trojan horse injection method is a serious problem with harmful impacts on cloud computing systems.

Service models

The National Institute of Standards and Technology recognized three cloud service models in 2011: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The International Organization for Standardization (ISO) later identified additional models in 2023, including "Network as a

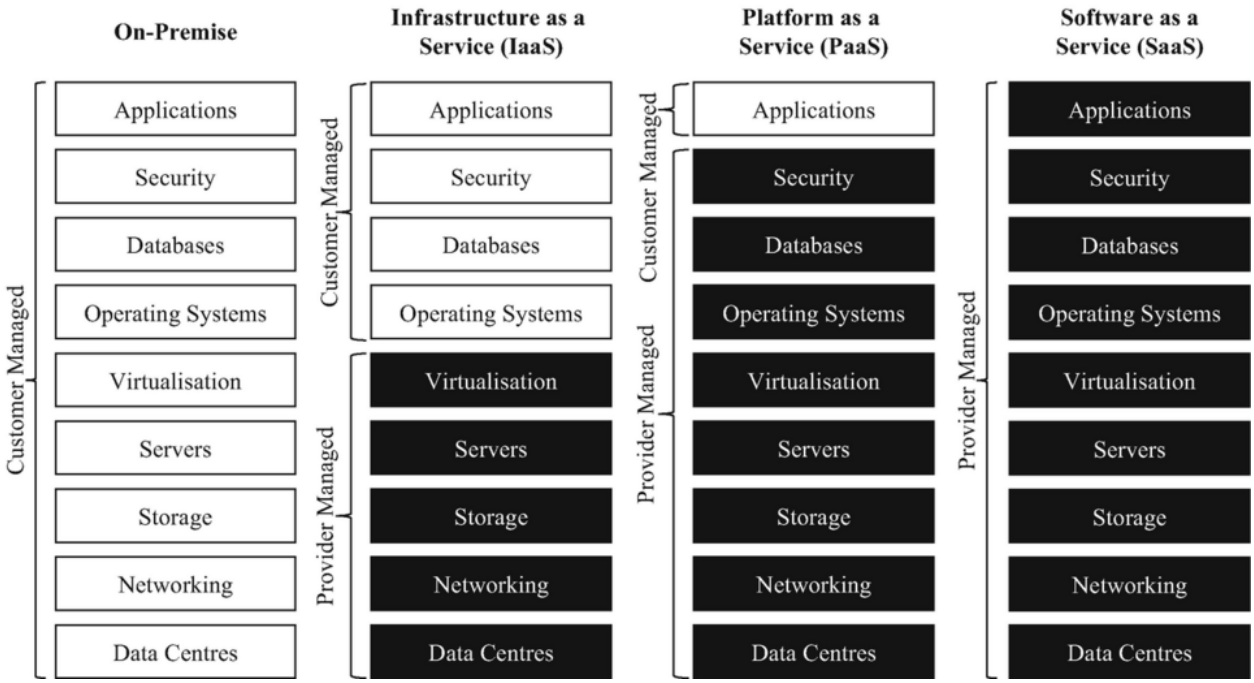

Figure 3: Comparison of on-premise, IaaS, PaaS, and SaaS

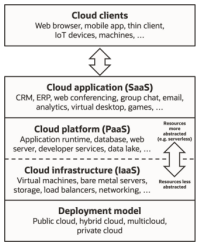

Figure 4: Cloud computing service models arranged as layers in a stack

Service", "Communications as a Service", "Compute as a Service", and "Data Storage as a Service".

Infrastructure as a service (IaaS)

Infrastructure as a service (IaaS) refers to online services that provide high-level APIs used to abstract various low-level details of underlying network infrastructure like physical computing resources, location, data partitioning, scaling, security, backup, etc. A hypervisor runs the virtual machines as guests. Pools of hypervisors within the cloud operational system can support large numbers of virtual machines and the ability to scale services up and down according to customers'varying requirements. Linux containers run in isolated partitions of a single Linux kernel running directly on the physical hardware. Linux cgroups and namespaces are the underlying Linux kernel technologies used to isolate, secure and manage the containers. The use of containers offers higher performance than virtualization because there is no hypervisor overhead. IaaS clouds often offer additional resources such as a virtual-machine disk-image library, raw block storage, file or object storage, firewalls, load balancers, IP addresses, virtual local area networks (VLANs), and software bundles.

The NIST's definition of cloud computing describes IaaS as "where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls)."

IaaS-cloud providers supply these resources on-demand from their large pools of equipment installed in data centers. For wide-area connectivity, customers can use either the Internet or carrier clouds (dedicated virtual private networks). To deploy their applications, cloud users install operating-system images and their application software on the cloud infrastructure. In this model, the cloud user patches and maintains the operating systems and the application software. Cloud providers typically bill IaaS services on a utility computing basis: cost reflects the number of resources allocated and consumed.

Platform as a service (PaaS)

The NIST's definition of cloud computing defines Platform as a Service as:

The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

PaaS vendors offer a development environment to application developers. The provider typically develops toolkit and standards for development and channels for distribution and payment. In the PaaS models, cloud providers deliver a computing platform, typically including an operating system, programming-language execution environment, database, and the web server. Application developers develop and run their software on a cloud platform instead of directly buying and managing the underlying hardware and software layers. With some PaaS, the underlying computer and storage resources scale automatically to match application demand so that the cloud user does not have to allocate resources manually.[need quotation to verify]

Some integration and data management providers also use specialized applications of PaaS as delivery models for data. Examples include iPaaS (Integration Platform as a Service) and dPaaS (Data Platform as a Service). iPaaS enables customers to develop, execute and govern integration flows. Under the iPaaS integration model, customers drive the development and deployment of integrations without installing or managing any hardware or middleware. dPaaS delivers integration—and data-management—products as a fully managed service. Under the dPaaS model, the PaaS provider, not the customer, manages the development and execution of programs by building data applications for the customer. dPaaS users access data through data-visualization tools.

Software as a service (SaaS)

The NIST's definition of cloud computing defines Software as a Service as:

The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

In the software as a service (SaaS) model, users gain access to application software and databases. Cloud providers manage the infrastructure and platforms that run the applications. SaaS is sometimes referred to as "on-demand software"and is usually priced on a pay-per-use basis or using a subscription fee. In the SaaS model, cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients. Cloud users do not manage the cloud infrastructure and platform where the application runs. This eliminates the need to install and run the application on the cloud user's own computers, which simplifies maintenance and support. Cloud applications differ from other applications in their scalability—which can be achieved by cloning tasks onto multiple virtual machines at run-time to meet changing work demand. Load balancers distribute the work over the set of virtual machines. This process is transparent to the cloud user, who sees only a single access-point. To accommodate a large number of cloud users, cloud applications can be multitenant, meaning that any machine may serve more than one cloud-user organization.

The pricing model for SaaS applications is typically a monthly or yearly flat fee per user, so prices become scalable and adjustable if users are added or removed at any point. It may also be free. Proponents claim that SaaS gives a business the potential to reduce IT operational costs by outsourcing hardware and software maintenance and support to the cloud provider. This enables the business to reallocate IT operations costs away from hardware/software spending and from personnel expenses, towards meeting other goals. In addition, with applications hosted centrally, updates can be released without the need for users to install new software. One drawback of SaaS comes with storing the users'data on the cloud provider's server. As a result,[citation needed] there could be unauthorized access to the data. Examples of applications offered as SaaS are games and productivity software like Google Docs and Office Online. SaaS applications may be integrated with cloud storage or File hosting services, which is the case with Google Docs being integrated with Google Drive, and Office Online being integrated with OneDrive.

Serverless computing

Serverless computing allows customers to use various cloud capabilities without the need to provision, deploy, or manage hardware or software resources, apart from providing their application code or data. ISO/IEC 22123-2:2023 classifies serverless alongside Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) under the broader category of cloud service categories. Notably, while ISO refers to these classifications as cloud service categories, the National Institute of Standards and Technology (NIST) refers to them as service models.

Deployment models

"A cloud deployment model represents the way in which cloud computing can be organized based on the control and sharing of physical or virtual resources."Cloud deployment models define the fundamental patterns of interaction between cloud customers and cloud providers. They do not detail implementation specifics or the configuration of resources.

Private

Private cloud is cloud infrastructure operated solely for a single organization, whether managed internally or by a third party, and hosted either internally or externally. Undertaking a private cloud project requires significant engagement to virtualize the business environment, and requires the organization to reevaluate decisions about existing resources. It can improve business, but every step in the project raises security issues

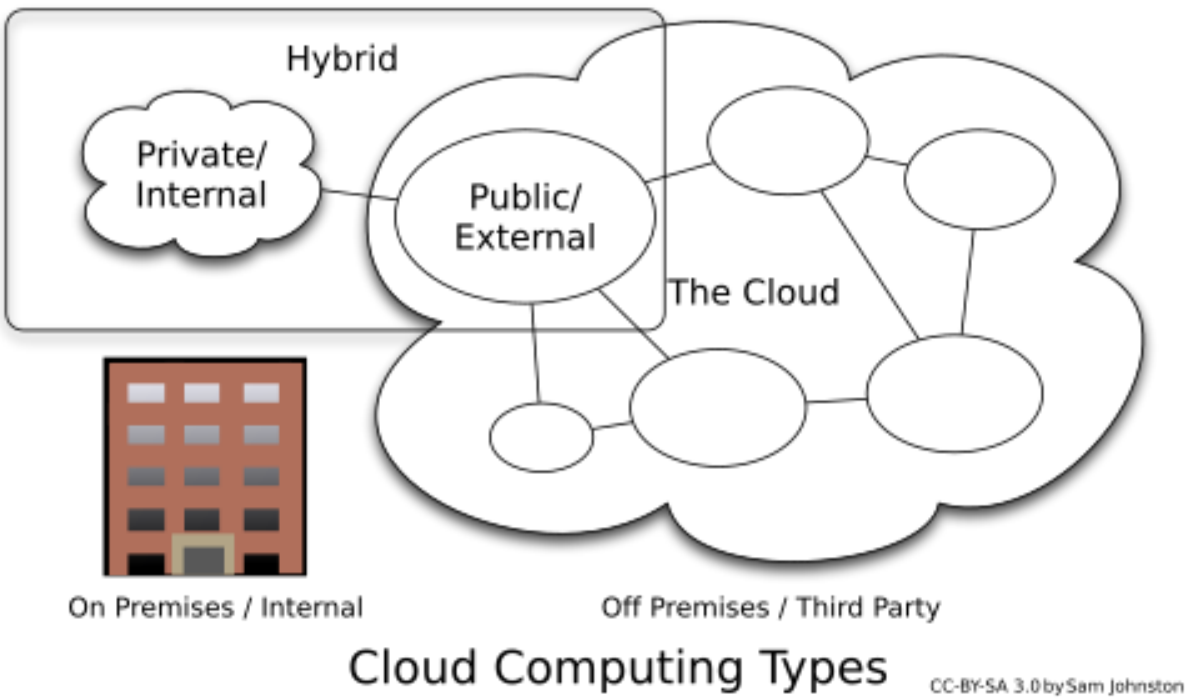

Figure 5: Cloud computing types

that must be addressed to prevent serious vulnerabilities. Self-run data centers are generally capital intensive. They have a significant physical footprint, requiring allocations of space, hardware, and environmental controls. These assets have to be refreshed periodically, resulting in additional capital expenditures. They have attracted criticism because users "still have to buy, build, and manage them"and thus do not benefit from less hands-on management, essentially "[lacking] the economic model that makes cloud computing such an intriguing concept".

Public

Cloud services are considered "public"when they are delivered over the public Internet, and they may be offered as a paid subscription, or free of charge. Architecturally, there are few differences between public- and private-cloud services, but security concerns increase substantially when services (applications, storage, and other resources) are shared by multiple customers. Most public-cloud providers offer direct-connection services that allow customers to securely link their legacy data centers to their cloud-resident applications.

Several factors like the functionality of the solutions, cost, integrational and organizational aspects as well as safety & security are influencing the decision of enterprises and organizations to choose a public cloud or on-premises solution.

Hybrid

Hybrid cloud is a composition of a public cloud and a private environment, such as a private cloud or on-premises resources, that remain distinct entities but are bound together, offering the benefits of multiple deployment models. Hybrid cloud can also mean the ability to connect collocation, managed or dedicated services with cloud resources. Gartner defines a hybrid cloud service as a cloud computing service that is composed of some combination of private, public and community cloud services, from different service providers. A hybrid cloud service crosses isolation and provider boundaries so that it cannot be simply put in one category of private, public, or community cloud service. It allows one to extend either the capacity

or the capability of a cloud service, by aggregation, integration or customization with another cloud service.

Varied use cases for hybrid cloud composition exist. For example, an organization may store sensitive client data in house on a private cloud application, but interconnect that application to a business intelligence application provided on a public cloud as a software service. This example of hybrid cloud extends the capabilities of the enterprise to deliver a specific business service through the addition of externally available public cloud services. Hybrid cloud adoption depends on a number of factors such as data security and compliance requirements, level of control needed over data, and the applications an organization uses.

Another example of hybrid cloud is one where IT organizations use public cloud computing resources to meet temporary capacity needs that can not be met by the private cloud. This capability enables hybrid clouds to employ cloud bursting for scaling across clouds. Cloud bursting is an application deployment model in which an application runs in a private cloud or data center and "bursts"to a public cloud when the demand for computing capacity increases. A primary advantage of cloud bursting and a hybrid cloud model is that an organization pays for extra compute resources only when they are needed. Cloud bursting enables data centers to create an in-house IT infrastructure that supports average workloads, and use cloud resources from public or private clouds, during spikes in processing demands.

Community

Community cloud shares infrastructure between several organizations from a specific community with common concerns (security, compliance, jurisdiction, etc.), whether it is managed internally or by a third-party, and hosted internally or externally, the costs are distributed among fewer users compared to a public cloud (but more than a private cloud). As a result, only a portion of the potential cost savings of cloud computing is achieved.

Multi cloud

According to ISO/IEC 22123-1: "multi-cloud is a cloud deployment model in which a customer uses public cloud services provided by two or more cloud service providers". Poly cloud refers to the use of multiple public clouds for the purpose of leveraging specific services that each provider offers. It differs from Multi cloud in that it is not designed to increase flexibility or mitigate against failures but is rather used to allow an organization to achieve more than could be done with a single provider.

Market

According to International Data Corporation (IDC), global spending on cloud computing services has reached $706 billion and is expected to reach $1.3 trillion by 2025. Gartner estimated that global public cloud services end-user spending would reach $600 billion by 2023. According to a McKinsey & Company report, cloud cost-optimization levers and value-oriented business use cases foresee more than $1 trillion in run-rate EBITDA across Fortune 500 companies as up for grabs in 2030. In 2022, more than $1.3 trillion in enterprise IT spending was at stake from the shift to the cloud, growing to almost $1.8 trillion in 2025, according to Gartner.

The European Commission's 2012 Communication identified several issues which were impeding the development of the cloud computing market:: Section 3

- fragmentation of the digital single market across the EU
- concerns about contracts including reservations about data access and ownership, data portability, and change control
- variations in standards applicable to cloud computing

The Communication set out a series of "digital agenda actions"which the Commission proposed to undertake in order to support the development of a fair and effective market for cloud computing services.: Pages 6–14

Cloud Computing Vendors

As of 2025, the three largest cloud computing providers by market share, commonly referred to as hyperscalers, are Amazon Web Services (AWS), Microsoft Azure, and Google Cloud. These companies dominate the global cloud market due to their extensive infrastructure, broad service offerings, and scalability.

In recent years, organizations have increasingly adopted alternative cloud providers, which offer specialized services that distinguish them from hyperscalers. These providers may offer advantages such as lower costs, improved cost transparency and predictability, enhanced data sovereignty (particularly within regions such as the European Union to comply with regulations like the General Data Protection Regulation (GDPR)), stronger alignment with local regulatory requirements, or industry-specific services.

Alternative cloud providers are often part of multi-cloud strategies, where organizations use multiple cloud services—both from hyperscalers and specialized providers—to optimize performance, compliance, and cost efficiency. However, they do not necessarily serve as direct replacements for hyperscalers, as their offerings are typically more specialized.

Similar concepts

The goal of cloud computing is to allow users to take benefit from all of these technologies, without the need for deep knowledge about or expertise with each one of them. The cloud aims to cut costs and helps the users focus on their core business instead of being impeded by IT obstacles. The main enabling technology for cloud computing is virtualization. Virtualization software separates a physical computing device into one or more "virtual"devices, each of which can be easily used and managed to perform computing tasks. With operating system–level virtualization essentially creating a scalable system of multiple independent computing devices, idle computing resources can be allocated and used more efficiently. Virtualization provides the agility required to speed up IT operations and reduces cost by increasing infrastructure utilization. Autonomic computing automates the process through which the user can provision resources on-demand. By minimizing user involvement, automation speeds up the process, reduces labor costs and reduces the possibility of human errors.

Cloud computing uses concepts from utility computing to provide metrics for the services used. Cloud computing attempts to address QoS (quality of service) and reliability problems of other grid computing models.

Cloud computing shares characteristics with:

- Client–server model –Client–server computing refers broadly to any distributed application that distinguishes between service providers (servers) and service requestors (clients).
- Computer bureau –A service bureau providing computer services, particularly from the 1960s to 1980s.
- Grid computing –A form of distributed and parallel computing, whereby a 'super and virtual computer' is composed of a cluster of networked, loosely coupled computers acting in concert to perform very large tasks.
- Fog computing –Distributed computing paradigm that provides data, compute, storage and application services closer to the client or near-user edge devices, such as network routers. Furthermore, fog computing handles data at the network level, on smart devices and on the end-user client-side (e.g. mobile devices), instead of sending data to a remote location for processing.
- Utility computing –The "packaging of computing resources, such as computation and storage, as a metered service similar to a traditional public utility, such as electricity."
- Peer-to-peer –A distributed architecture without the need for central coordination. Participants are both suppliers and consumers of resources (in contrast to the traditional client-server model).
- Cloud sandbox –A live, isolated computer environment in which a program, code or file can run without affecting the application in which it runs.