[H1] How Cloud Computing Works

Dek: Methods and techniques

The foundations of cloud computing are high-powered computers, called *servers*, and the facilities they're gathered inside, called *data centers*. Both servers and data centers predate the cloud by decades. Historically, when organizations increased processing power by concentrating computers in one or more places, they bought and operated that equipment on-premises. Cloud computing's central innovation is placing large numbers of servers in data centers, typically located far from the individuals and organizations that access them.

[H2] How data centers work

Though "data center" can refer to any group of networked computers, in the context of cloud computing the term usually describes the buildings—or compounds of buildings—that house collections of internet-connected servers. Data centers allow cloud providers to connect more servers to one another, and to more users over the internet.

[H3] How servers work

Servers are high-powered computers that use specialized versions of the computing hardware found in consumer desktop or laptop computers. That specialized hardware makes servers more expensive than ordinary computers, but it can also make them:

- **Better at multitasking.** Additional processor components and more short-term memory help servers carry out more tasks at once.

- **Able to handle more data.** Specialized features of their processors can allow bigger data transfers, preventing the bottlenecks that can lead to lag or processing delays.
- **More durable and longer lasting.** Higher-quality components reduce the chance of breakdowns, even when servers are operating at full performance for extended periods.
- **Optimized to work with other servers.** Special networking hardware and software enables servers to more easily coordinate with other servers they're connected to, sharing and shifting processing duties as needed.

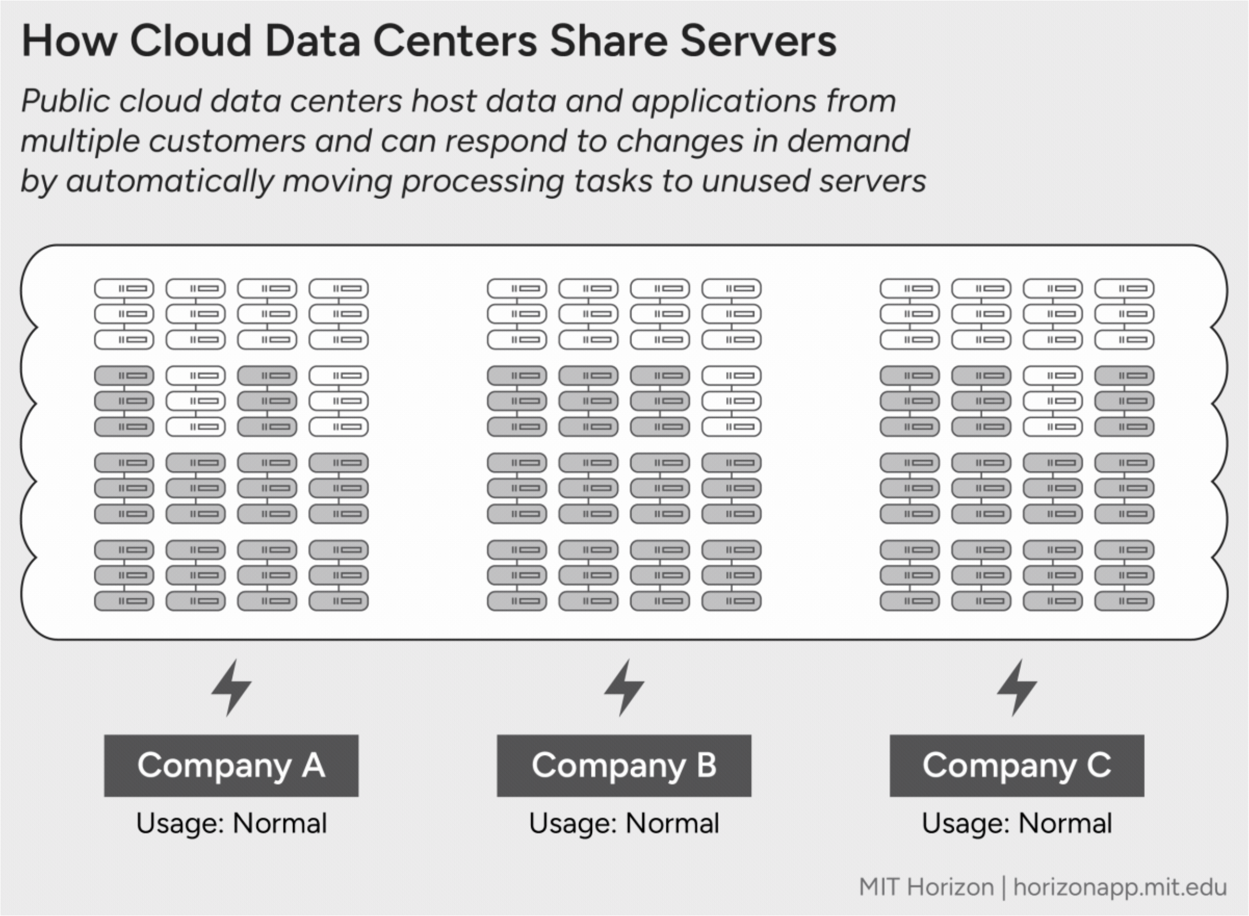
[H3] How groups of servers work

By gathering and connecting servers under one roof, data centers can amplify the efficiencies provided by each server. Data centers vary in size and design, ranging from a dedicated section of a building containing dozens of servers to sprawling facilities filled with hundreds of rows of server racks and cabinets. Either way, consolidating large numbers of reliable and high-performance computers in a single place enables data centers to:

- **Scale processing power and data storage on demand.** Data centers use automation and management software to distribute computing tasks across a pool of servers. This can include shifting workloads to servers that aren't in use, or even to another data center, to accommodate sudden increases in demand such as a spike in users logged into an application. Data centers can also respond to decreases in demand, scaling back computing resources and allowing customers to pay less when an application goes unused.
- **Reduce electricity use and other overhead costs.** All computers consume electricity and produce heat. The high-performance and long-running nature of servers means they

> do both consistently. Collecting servers in one location, and providing them with energy management software, allows the operator of a data center to optimize energy expenditure among its servers—including spreading activity across more servers. A data center can also employ customized approaches to cooling groups of servers. For example, placing rows of server racks and cabinets in floor-to-ceiling enclosures allows HVAC systems to focus on cooling those closed-off aisles, while the spaces between rows—where workers walk through the data center—remain less air-conditioned.

These and other efficiencies lower operating costs and result in equipment that's ready to handle significant computing workloads for long periods. Though some organizations run their own data centers, the larger facilities are operated by cloud providers and organizations effectively rent the use of their servers. (For information about specific cloud providers, see Major Cloud Computing Service Providers.)



How Cloud Data Centers Share Servers

Public cloud data centers host data and applications from multiple customers and can respond to changes in demand by automatically moving processing tasks to unused servers

MIT Horizon | horizonapp.mit.edu

HED: How Cloud Data Centers Share Servers

DEK: Public cloud data centers host data and applications from multiple customers and can respond to changes in demand by automatically moving processing tasks to unused servers

- Company A, B, C
- Usage: Normal or High

[H2] Types of cloud services

By allowing organizations to use computer hardware at a distance, the cloud turns most computing functions into services that a user can access themself without having to own or operate the necessary equipment. For example, an organization might move resources to the cloud to make employee records more accessible to HR staff in multiple locations or to give an existing software application more reach and processing power. This “as a service” approach can be applied to every kind of activity in the cloud. An organization’s chosen approach can help determine not only what resources it needs to move to the cloud but also the technical expertise required to make those resources work remotely.

[H3] Infrastructure as a service

The most common purposes of moving data and applications to the cloud include increasing access to them over the internet, lowering costs by using only the physical equipment needed at a given time, and delegating the responsibility for that equipment to another organization. For example, a company that keeps its data in the cloud can make it available to employees working from home while paying for storage space that automatically scales to match the amount of data in use, on servers maintained and secured around the clock by a cloud provider’s data center staff. This is an example of the *infrastructure-as-a-service* (IaaS) service model, where a customer pays to use a provider’s cloud infrastructure.

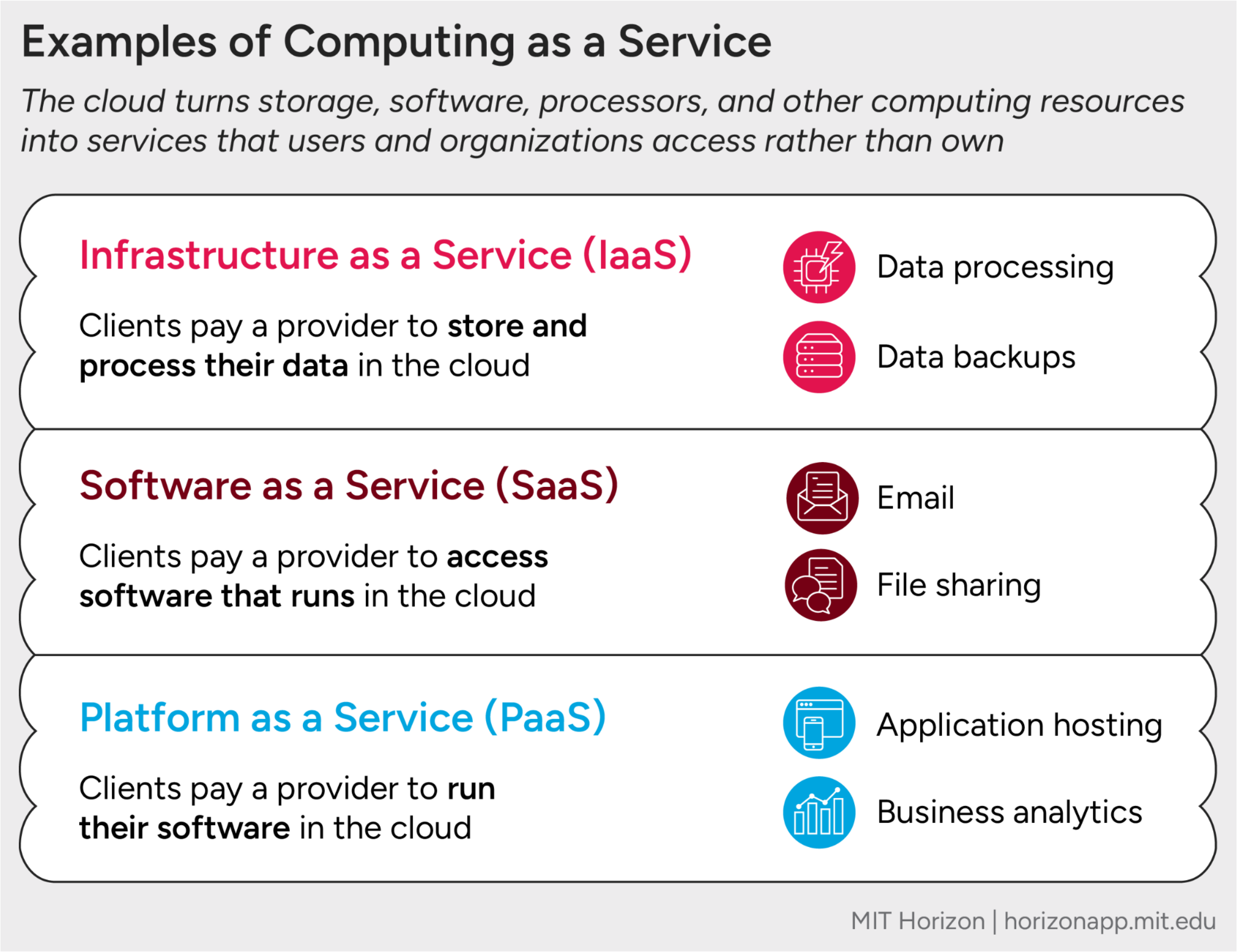
The goal of many IaaS applications isn’t to simply store and access data but to use it in ways that take advantage of its location on remote servers. Distributing data storage and processing throughout the cloud can increase the scale and speed of analysis, as well as allow organizations to incorporate types of data that otherwise wouldn’t be possible to analyze together. This can streamline data analytics and provide more training data to help artificial intelligence algorithms produce insights and predictions. (For more on training algorithms, see How Artificial Intelligence Works.)

[H3] Software as a service

When organizations pay to access another organization's software that runs in the cloud, such as using Zoom for video calls or Slack for messaging, this is referred to as *software as a service* (SaaS). Software that runs entirely in the cloud relies on data centers to handle the bulk of the processing and networking duties, allowing users to access that software with a variety of lower-powered devices. These devices may use a relatively small downloaded application to connect to cloud software. For example, someone can join a 1,000-participant Zoom video conference on their phone because most of the related information is being stored and processed on remote servers.

[H3] Platform as a service

The most complex cloud services often involve *platform-as-a-service* (PaaS) deployments, where an organization operates its own software in the cloud and pays a provider to handle many of the related technical responsibilities, including maintaining the servers that store and run the application. Launching a PaaS application is often as much about developing new software as it is about transferring data to a provider. In some cases, existing code can be modified to work in the cloud, but organizations usually rebuild applications from scratch to run specifically on remote servers.



Examples of Computing as a Service

The cloud turns storage, software, processors, and other computing resources into services that users and organizations access rather than own

Infrastructure as a Service (IaaS)

Clients pay a provider to **store and process their data** in the cloud

Data processing

Data backups

Software as a Service (SaaS)

Clients pay a provider to **access software that runs** in the cloud

Email

File sharing

Platform as a Service (PaaS)

Clients pay a provider to **run their software** in the cloud

Application hosting

Business analytics

MIT Horizon | horizonapp.mit.edu

HED: Examples of Computing as a Service

DEK: The cloud turns storage, software, processors, and other computing resources into services that users and organizations access rather than own

- Infrastructure as a Service (IaaS)
    - Clients pay a provider to store and process their data in the cloud
    - Data processing, data backups
- Platform as a Service (PaaS)
    - Clients pay a provider to run their software in the cloud
    - Application hosting, business analytics
- Software as a Service (SaaS)

- Clients pay a provider to access software that runs in the cloud
- Email, file sharing

[H3] Other service models

There are other types of cloud service models, including some introduced by vendors to help distinguish their products, but all are variations on IaaS, SaaS, and PaaS models. Because cloud service models are defined by how an individual or organization interacts with a given service, a single service can fall under multiple models, depending on who's accessing it. For example, Netflix's video streaming platform is a PaaS for Netflix, which pays a cloud provider to store its content library and host the software that lets millions of subscribers view that content. But Netflix's subscribers use it as an SaaS, paying Netflix to access its streaming service.

[H2] Types of cloud access

Most users connect to the cloud using the same fiber optic cables, satellite arrays, and other infrastructure that enables the public internet. This is also true of many organizations that transfer data to the cloud as part of a larger process called *cloud migration*. Some cloud providers offer migration tools and dedicated higher-speed connections to send information to their servers over the internet, as well as options for physical transportation of large amounts of data by loading it onto devices and physically transporting them to a data center. For example, one major provider, Amazon Web Services (AWS), offers a roughly 50-pound device called Snowball with 80 terabytes of storage (enough to store around 4.5 years of HD video). For even larger data migrations, it has a 45-foot-long shipping container called Snowmobile, which is hauled by a tractor-trailer and filled with 100 petabytes of data storage hardware, or enough to store more than 5,700 years of HD video.

However an organization moves its resources to a data center, the purpose of most migrations is to make those resources accessible over the internet. But accessing a cloud service shouldn't be confused with directly accessing a particular data center. Instead, accessing that information means interacting with any number of layered and largely unseen intermediaries.

The most common of these intermediaries are *APIs*, or *application programming interfaces*—sets of software tools that allow different applications to work with one another. APIs are used in a variety of other computing contexts, but in cloud computing they enable interfaces for people to use the data and services housed in data centers. So a company checking its cloud-stored backup data might see files organized into folders, even if its information is scattered and moving throughout the provider's data center.

Organizations also use APIs to update or add features to their cloud services. Because they're often designed to work with a given cloud provider's servers, these APIs can speed the process of making modular changes to one part of a cloud application without having to rewrite larger amounts of code or ask users to download anything as an update. The next time a customer logs into a cloud videoconferencing application, they'll simply be accessing code that's been updated in the cloud.

This ability to quickly and regularly update services, and to access those updates immediately, is central to the cloud's widespread popularity and impact. The APIs, security measures, and even specific devices that allow for cloud access depend on the type of data center a cloud service uses. And while data centers employ a variety of designs and technologies, they fall into three general categories of cloud access.

[H3] Public clouds

The largest data centers are run by *public* cloud providers such as AWS, Google Cloud, and Microsoft Azure, and they are called "public" because nearly any customer can pay to use their servers. Public cloud data centers host data and applications from multiple customers at once and can be accessed using standard internet connections—though individual applications typically require passwords or similar authentications. SaaS and PaaS deployments almost always run in public cloud environments, because they offer users such easy access. Public cloud data centers are also typically cheaper to work with than other cloud environments, since public providers can distribute and optimize equipment use across large facilities.

[H3] Private clouds

Data centers can also be owned by the organizations that use them, or by a cloud provider that makes them available to only a single customer. These are called *private* cloud environments, and while they're typically more expensive than public cloud solutions, they can offer greater security by imposing more restrictions on access. Some private clouds require an encrypted internet connection (such as a VPN) or a trusted device (such as a work computer), allowing a properly equipped user to access services from almost anywhere. Others blur the line between cloud access and more traditional computer networks, requiring the use of the wireless routers located in the company's offices.

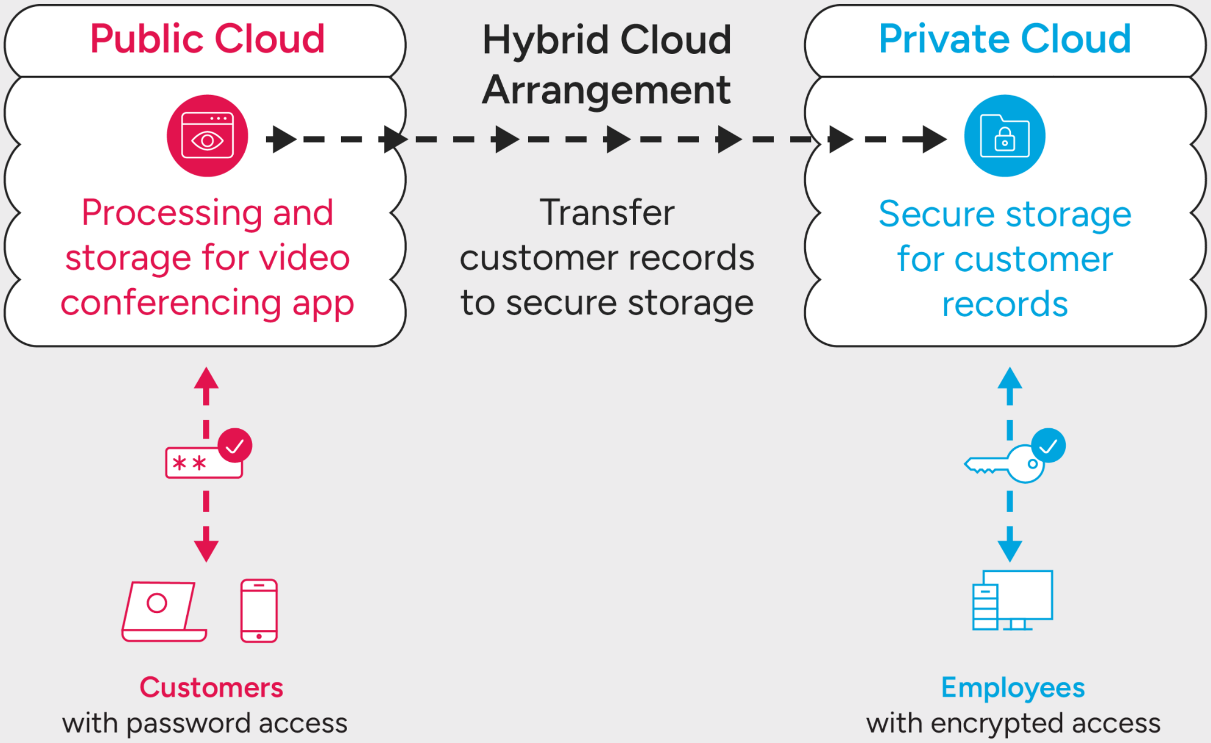
Private cloud data centers are often used for IaaS applications to store information that's only accessible to certain employees, such as financial records or proprietary research. But even when an organization has set up this kind of private access, it doesn't receive a map of servers within that data center and a choice of which one or ones to connect to at a given time. Private clouds still require APIs and other intermediaries to access stored resources.

[H2] Hybrid clouds

When organizations split data and applications across public and private clouds, this is referred to as a *hybrid* cloud environment. This is the most common type of cloud setup for companies. For example, an organization might use a public cloud provider to host its videoconferencing software (a PaaS application) and a private cloud provider for its customer service records (an IaaS application). Hybrid cloud arrangements can also include organizations storing some information—typically sensitive data or information that must be accessed instantaneously—in on-premises data centers while other information is stored remotely in public and private clouds. This combination of different types of data centers can lead to complex access requirements, particularly when it's necessary to set up a direct connection between resources on public and private clouds. In most cases, though, hybrid clouds simply reflect the growing diversity of cloud solutions, with each service requiring the passwords, authorized devices, and other measures associated with accessing public and private clouds.

Public, Private, and Hybrid Cloud Environments

A video conferencing company can run its software in a public cloud, store customer records in a restricted-access private cloud, and create a hybrid cloud that transfers data from public to private server


Public Cloud
Processing and storage for video conferencing app
Hybrid Cloud Arrangement
Transfer customer records to secure storage
Private Cloud
Secure storage for customer records
Customers
with password access
Employees
with encrypted access

MIT Horizon | horizonapp.mit.edu

HED: Public, Private, and Hybrid Cloud Environments

DEK: A video conferencing company can run its software in a public cloud, store customer records in a restricted-access private cloud, and create a hybrid cloud that transfers data from public to private server

- Public Cloud: Processing and storage for video conferencing app. Customers with password access
- Hybrid Cloud Arrangement: Transfer customer records to secure storage
- Private Cloud: Secure storage for customer records. Employees with encrypted access

When setting up a hybrid cloud environment, organizations can choose a single public cloud provider or spread their resources across several different public providers in an arrangement called *multicloud*. Increasingly, organizations are choosing the multicloud approach, which can maximize the advantages of each public cloud provider, securing lower cost options and greater flexibility. And in dividing their resources, organizations can also minimize the impact of a breach or an outage.