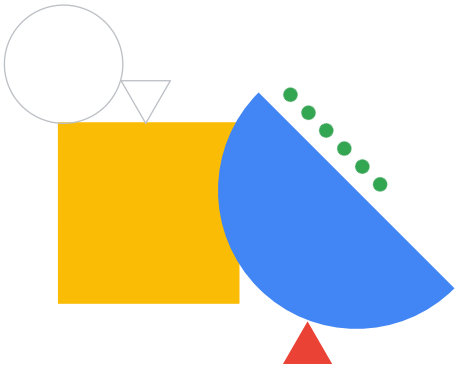
Google Cloud

Google Cloud Fundamentals: Core Infrastructure



Instructor-led training

Introducing Google Cloud

Introducing Google Cloud

Let's start at the beginning with an overview of cloud computing. Some of you might be wondering about the basics: What exactly is Google Cloud? How is it organized? And what makes it unique?

In this module, we'll answer those questions.

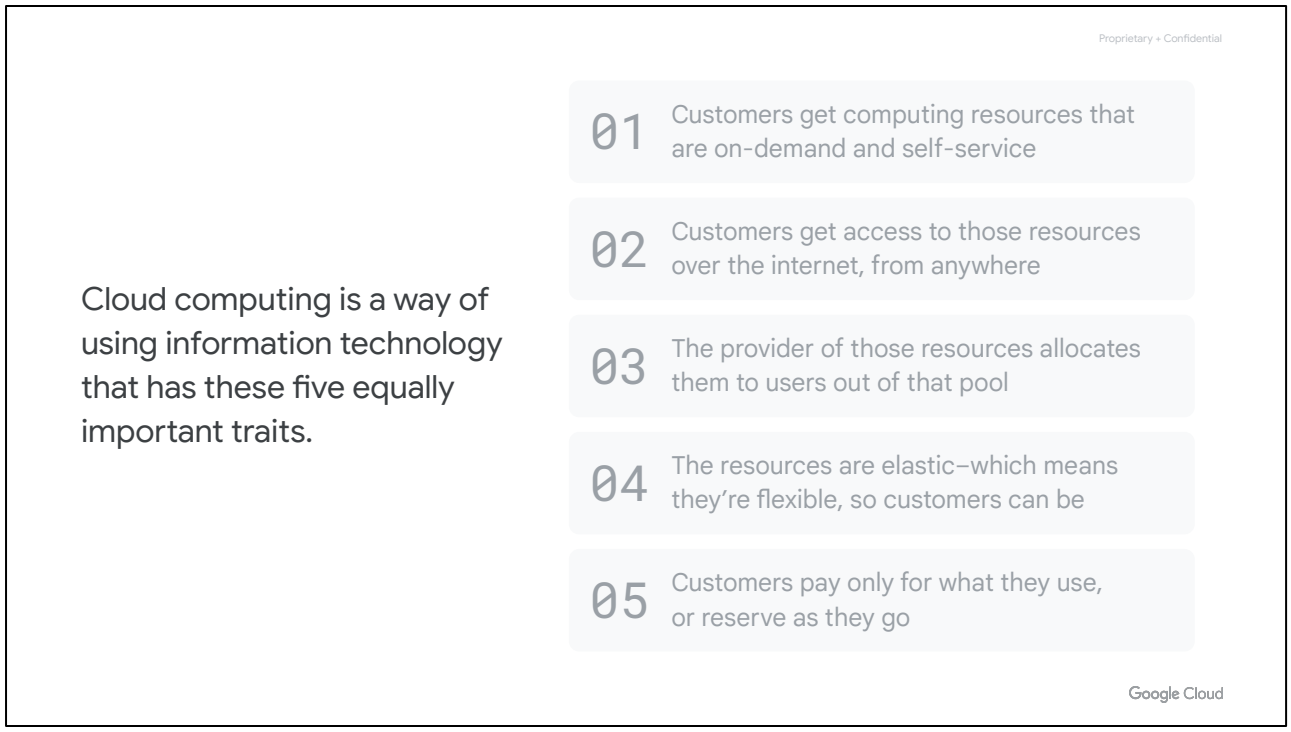
What is Cloud Computing?

US National Institute of Standards and Technology

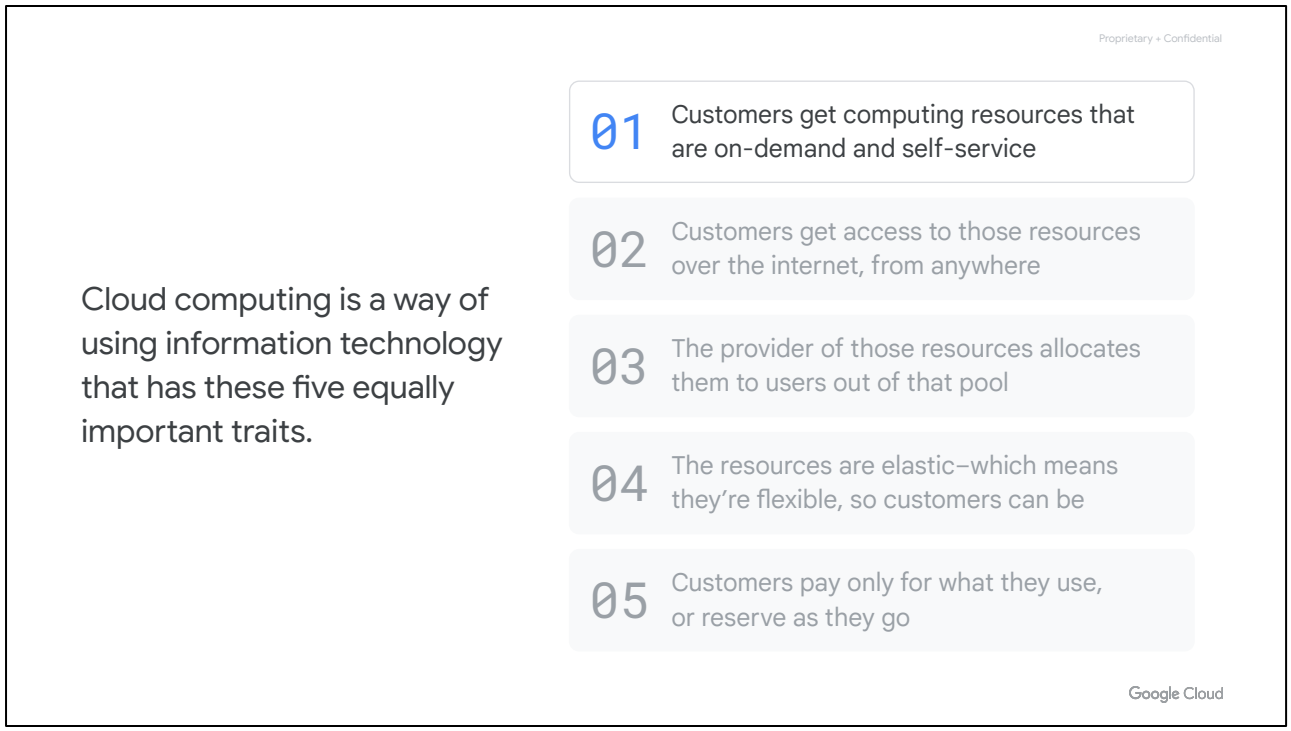
Cloud computing

The cloud is a hot topic these days, but what exactly is it?

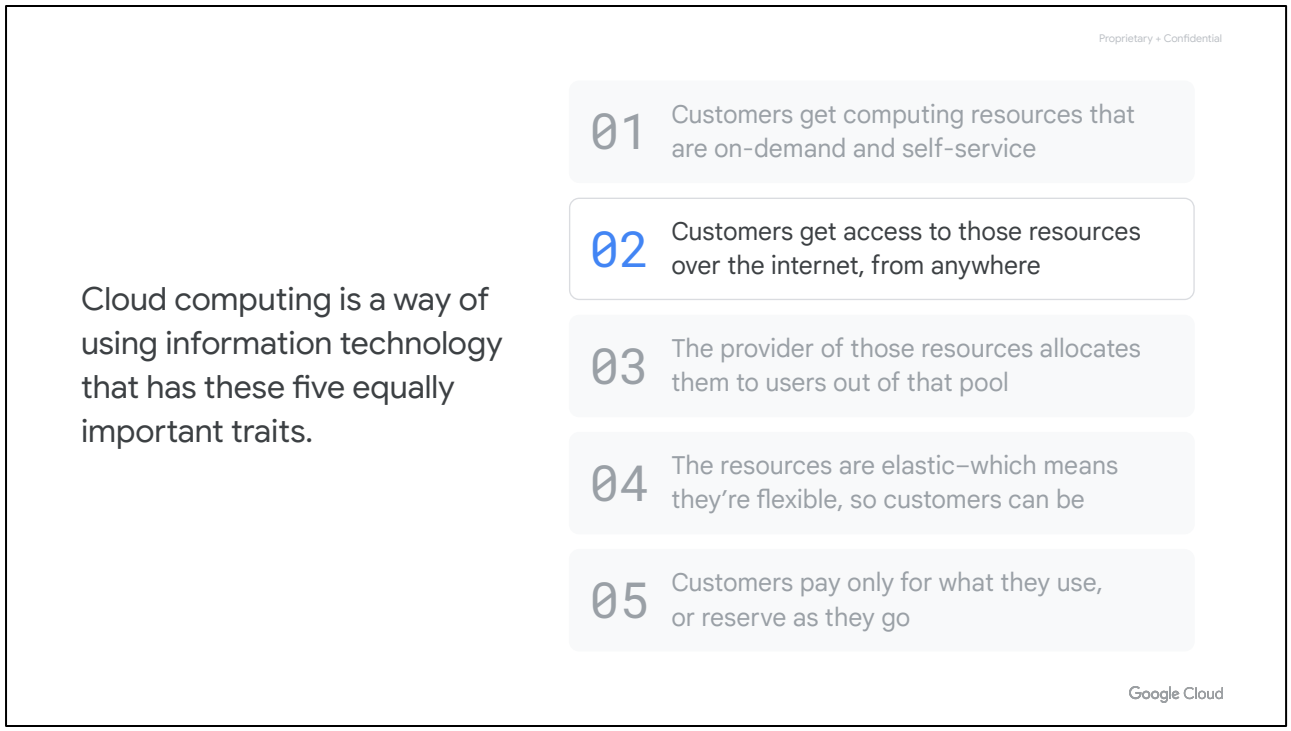
The US National Institute of Standards and Technology created the term cloud computing, although, there is nothing US-specific about it.



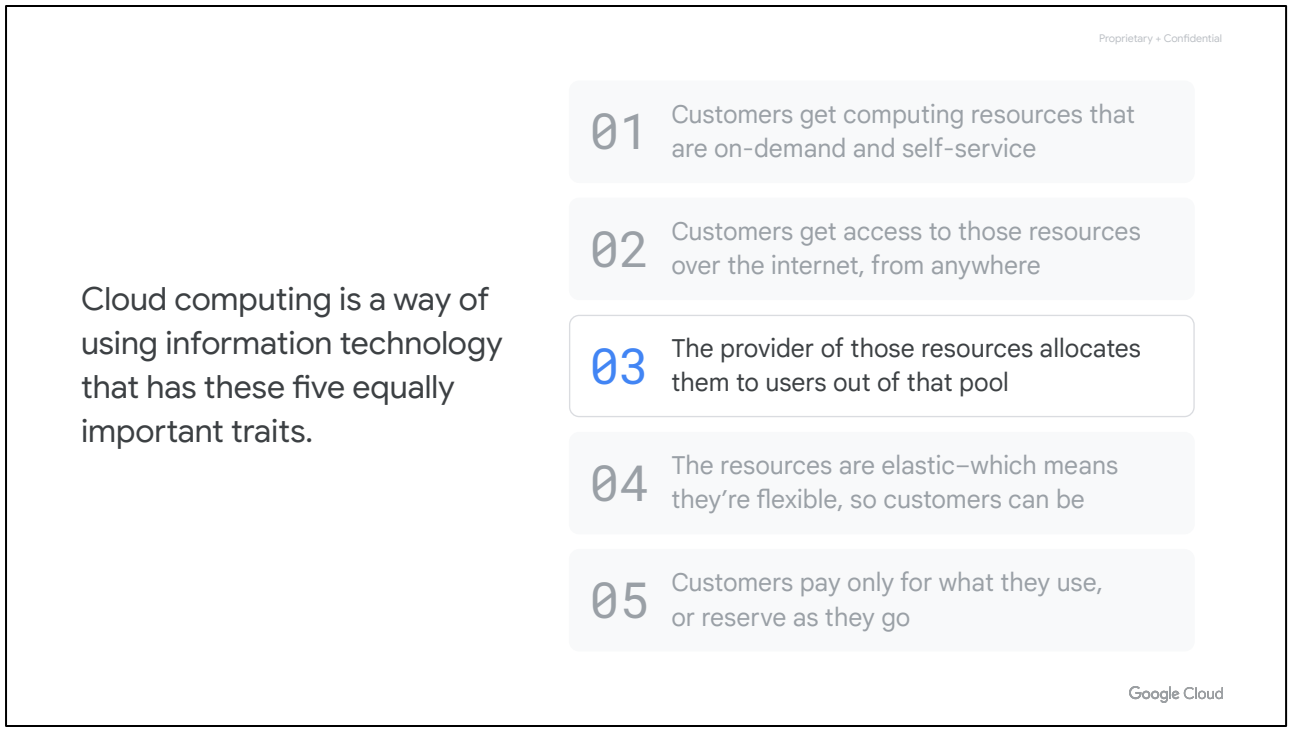
Cloud computing is a way of using information technology (IT) that has these five equally important traits.



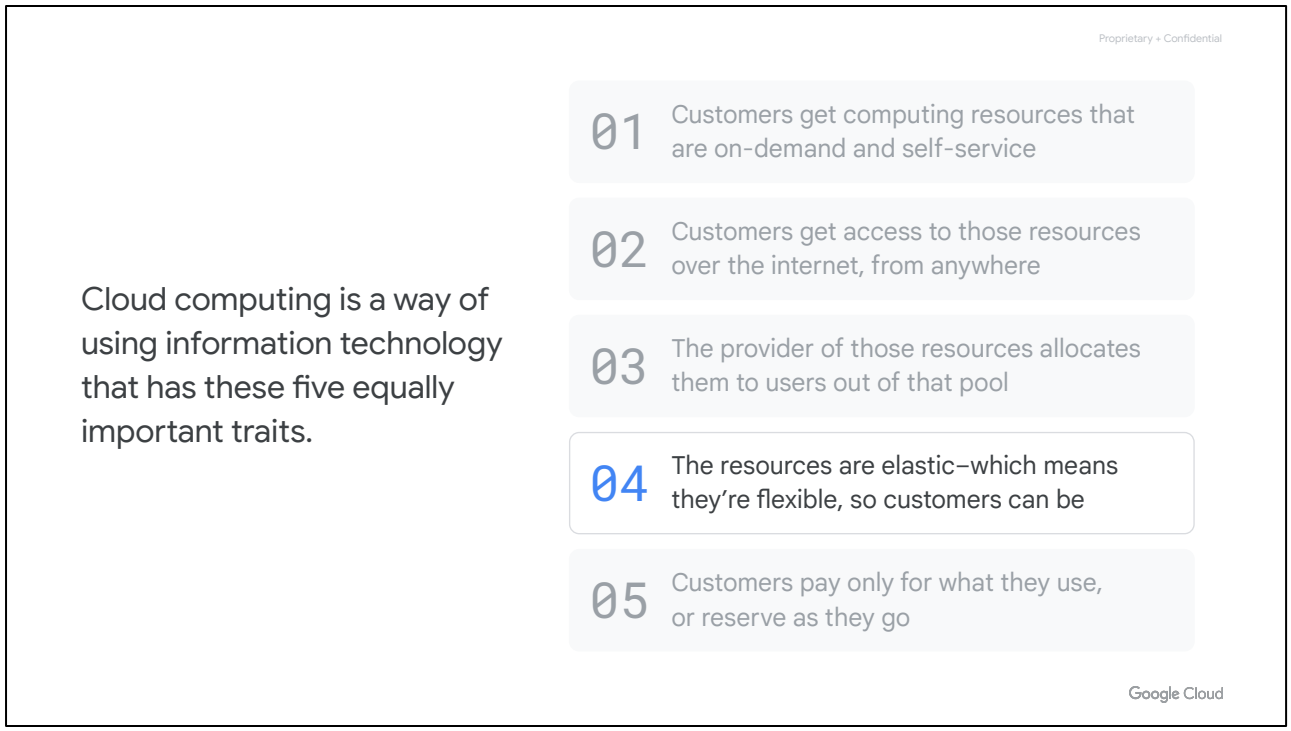
First, customers get computing resources that are on-demand and self-service. Through a web interface, users get the processing power, storage, and network they need with no need for human intervention.



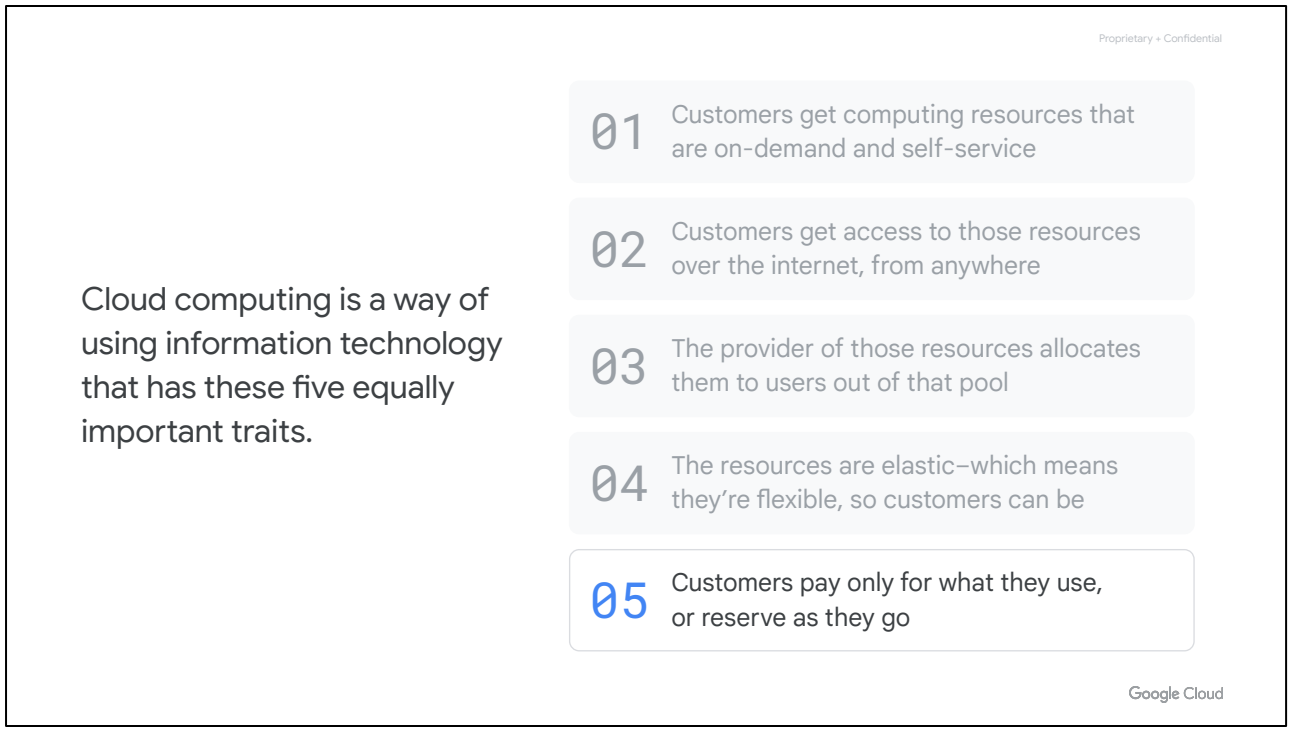
Second, customers get access to those resources over the internet, from anywhere they have a connection.



Third, the provider of those resources has a big pool of them, and allocates them to users out of that pool. That allows the provider to buy in bulk and pass the savings on to the customers. Customers don't have to know or care about the exact physical location of those resources.



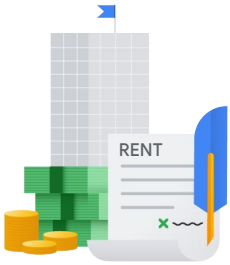
Fourth, the resources are elastic–which means they're flexible, so customers can be. If they need more resources they can get more, and quickly. If they need less, they can scale back.



And **finally**, the customers pay only for what they use, or reserve as they go. If they stop using resources, they stop paying.

That's it. That's the definition of cloud.

The history of cloud computing



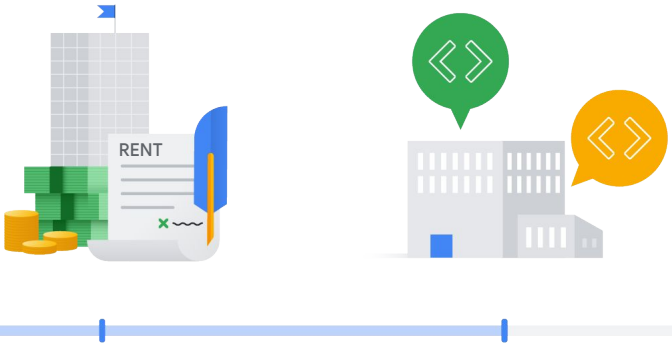
First wave

Colocation

But why is the cloud model so compelling nowadays? To understand why, we need to look at some history.

The trend towards cloud computing started with a **first wave** with something called *colocation*. Colocation gave users the financial efficiency of renting physical space, instead of investing in data center real estate.

The history of cloud computing



First wave

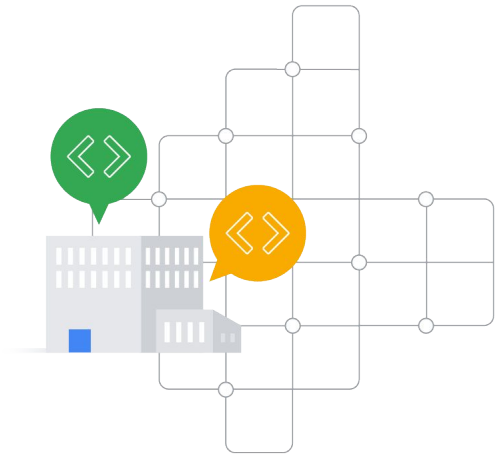
Colocation

Second wave

Virtualized data center

Virtualized data centers of today, which is the **second wave**, share similarities with the private data centers and colocation facilities of decades past. The components of virtualized data centers match the physical building blocks of hosted computing—servers, CPUs, disks, load balancers, and so on—but now they're virtual devices.

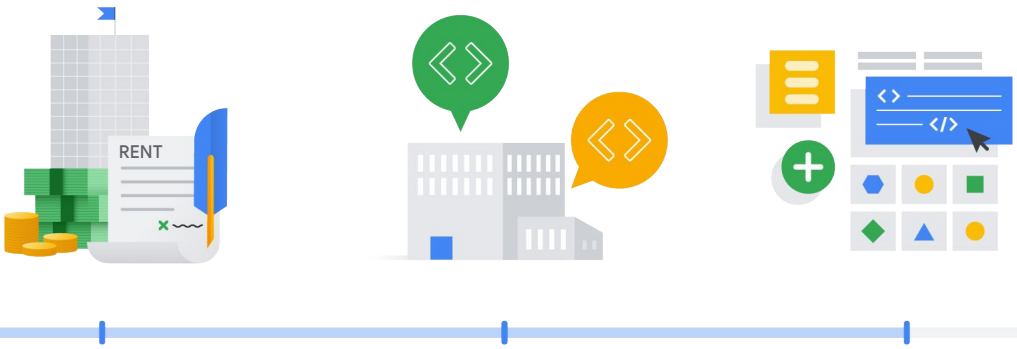
Enterprises still maintain the infrastructure



- User-controlled environment
- User-configured environment

With virtualization, enterprises still maintained the infrastructure; it's still a user-controlled and user-configured environment.

The history of cloud computing



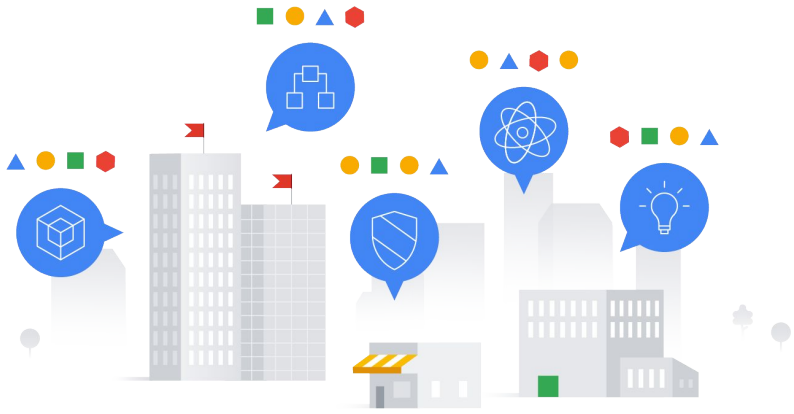
First wave
Colocation

Second wave
Virtualized data center

Third wave
Container-based architecture

Several years ago, Google realized that its business couldn't move fast enough within the confines of the virtualization model. So Google switched to a container-based architecture—a fully automated, elastic **third-wave** cloud that consists of a combination of automated services and scalable data. Services automatically provision and configure the infrastructure used to run applications.

Third-wave cloud is available to Google customers

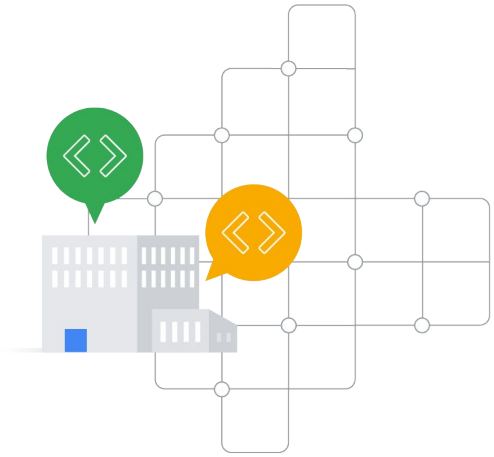


Today, Google Cloud makes this third-wave cloud available to Google customers.

Google believes that, in the future, every company—regardless of size or industry—will differentiate itself from its competitors through technology. Increasingly, that technology will be in the form of software. Great software is based on high-quality data. This means that every company is, or will eventually become, a data company.

Introducing Google Cloud

Cloud service offerings

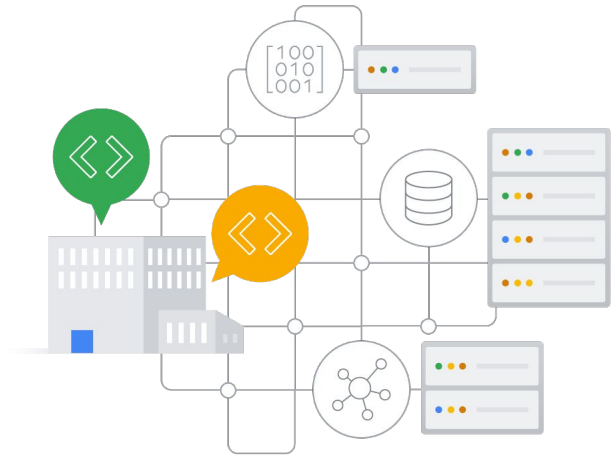


- **IaaS** - Infrastructure as a service
- **PaaS** - Platform as a service

This move to virtualized data centers introduced customers to two new types of offerings: infrastructure as a service, commonly referred to as **IaaS**, and platform as a service, or **PaaS**.

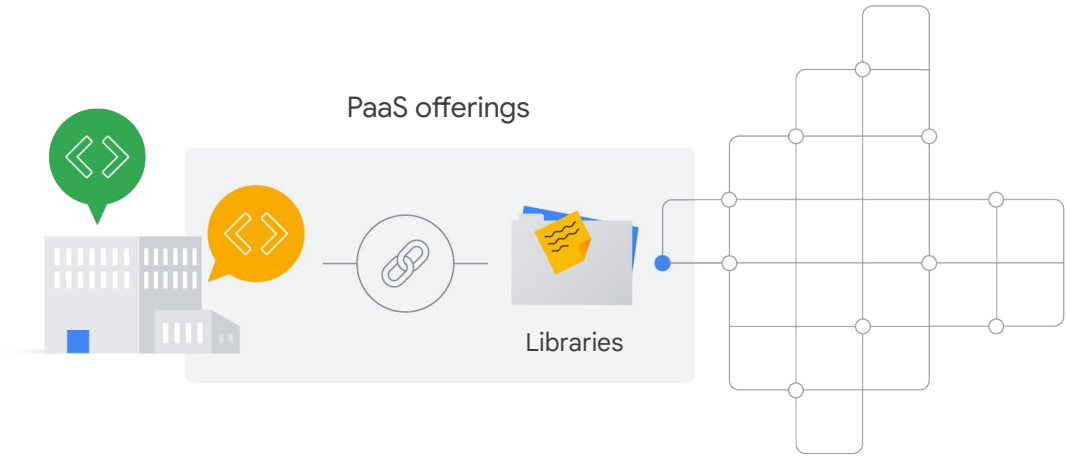
Infrastructure as a Service (IaaS)

- Raw compute
- Storage
- Network capabilities



IaaS offerings provide raw compute, storage, and network capabilities, organized virtually into resources that are similar to physical data centers.

Platform as a Service (PaaS)

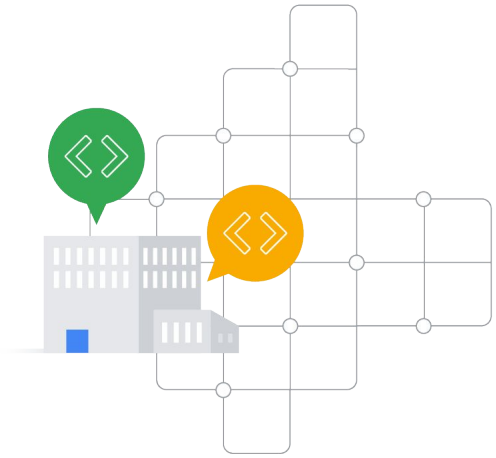


PaaS offerings, on the other hand, bind code to libraries that provide access to the infrastructure application needs. This allows more resources to be focused on application logic.

Payment models

IaaS

Pay for what they allocate



PaaS

Pay for what they use

In the IaaS model, customers pay for the resources they allocate ahead of time; in the PaaS model, customers pay for the resources they actually use.

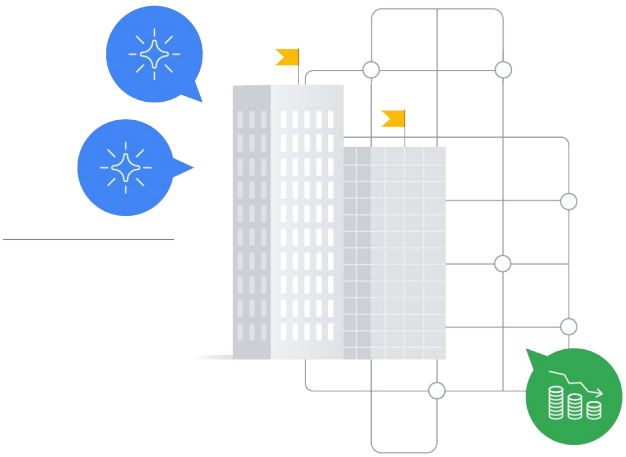
The evolution of cloud computing

Managed resources

Managed services

Deliver products/services

- More quickly
- More reliably

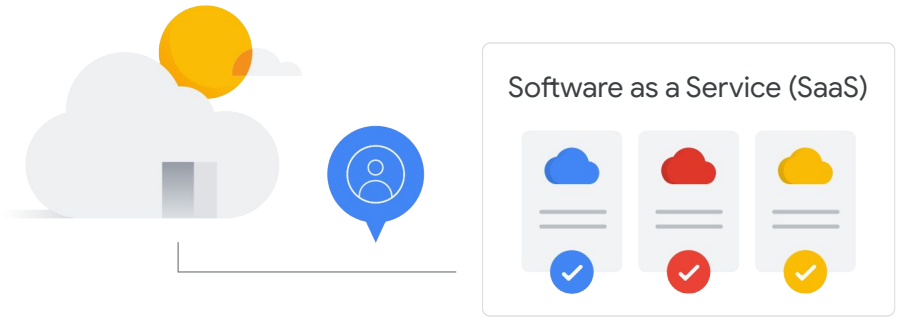


As cloud computing has evolved, the momentum has shifted toward managed infrastructure and managed services.

Leveraging managed resources and services allows companies to concentrate more on their business goals and spend less time and money on creating and maintaining their technical infrastructure. It allows companies to deliver products and services to their customers more quickly and reliably.

Although not shown on this slide, **serverless** is yet another step in the evolution of cloud computing. Serverless computing allows developers to concentrate on their code, rather than on server configuration, by eliminating the need for any infrastructure management. Serverless technologies offered by Google include Cloud Functions which manages event-driven code as a pay-as-you-go service, and Cloud Run, which allows customers to deploy their containerized microservices based application in a fully-managed environment.

What about SaaS?



You might have heard about Software as a Service (SaaS) and wondered what it is and how it fits into the Cloud ecosphere.

Software as a Service applications are not installed on your local computer - they run in the cloud as a service and are consumed directly over the internet by end users.

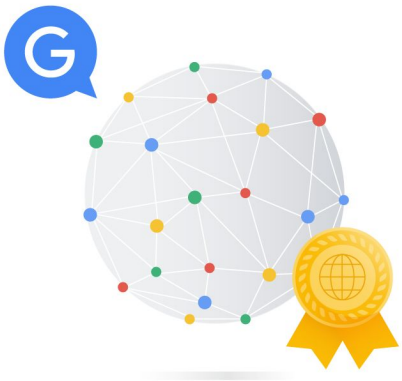
Google's popular applications like Gmail, Docs, and Drive, collectively known as Google Workspace, are all classified as SaaS.

You can learn more about these products or our other course offerings, but they are outside the scope of this particular course.

Introducing Google Cloud

Now that we've learned about or had a refresher on the basics of cloud computing, let's explore Google's own network that Google Cloud runs on.

Largest network of its kind



Google's investment in its network runs into the billions of dollars

Google's network is the largest network of its kind, and Google has invested billions of dollars over the years to build it.

Designed for high throughput

Google's global network is designed to give customers the highest possible throughput and lowest possible latencies for their applications by leveraging more than 100 content caching nodes worldwide–which are locations where high demand content is cached for quicker access–to respond to user requests from the location that will provide the quickest response time.

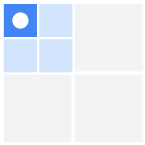
Infrastructure locations

Google Cloud's infrastructure is based in five major geographic locations: North America, South America, Europe, Asia, and Australia. Having multiple service locations is important because choosing where to locate applications affects qualities like availability, durability, and latency, the latter of which measures the time a packet of information takes to travel from its source to its destination.

Geographic locations contain regions and zones

Location

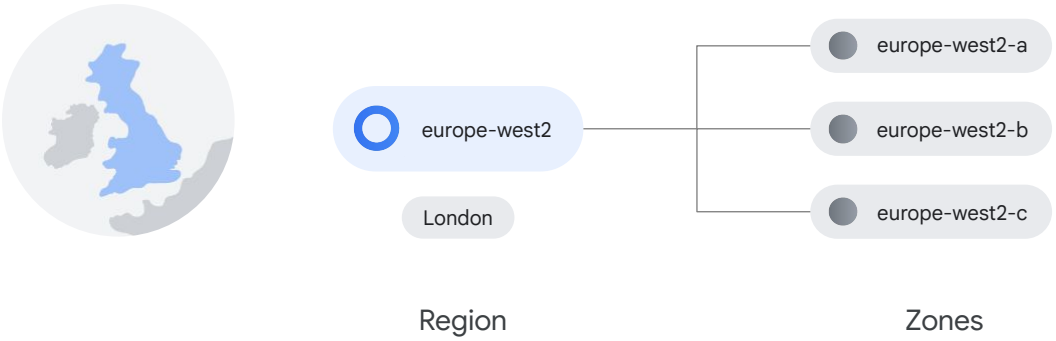
Regions



Zones

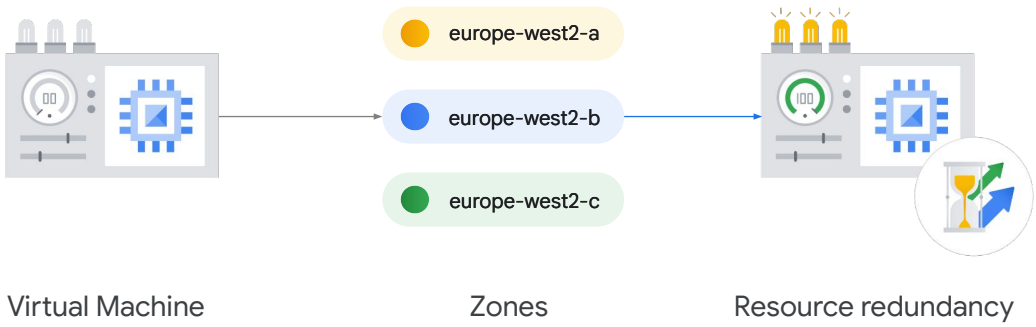
Each of these locations are divided into a number of different **regions and zones**.

Regions contain multiple zones



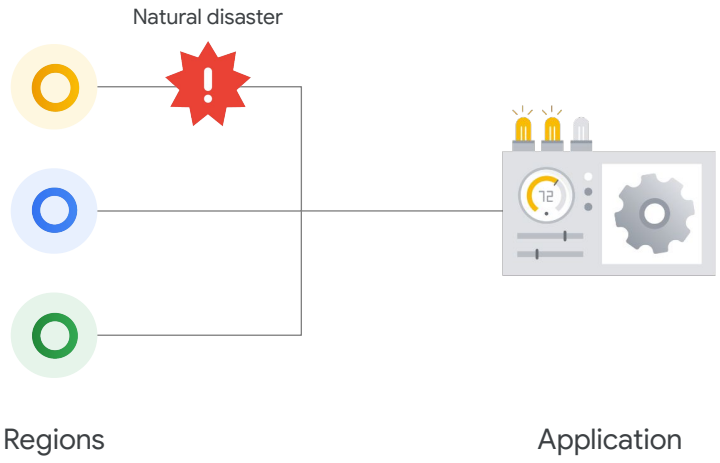
Regions represent independent geographic areas, and are composed of zones. For example, London, or europe-west2, is a region that currently comprises three different zones.

Zones are where Cloud resources are deployed



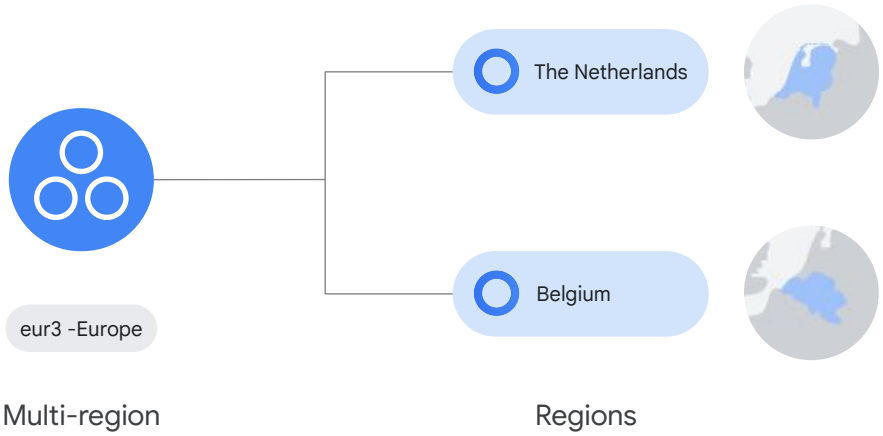
A **zone** is an area where Google Cloud resources get deployed. For example, let's say you launch a virtual machine using Compute Engine–more about Compute Engine in a bit–it will run in the zone that you specify to ensure resource redundancy.

Resources can run in different regions



You can run resources in different regions. This is useful for bringing applications closer to users around the world, and also for protection in case there are issues with an entire region, say, due to a natural disaster.

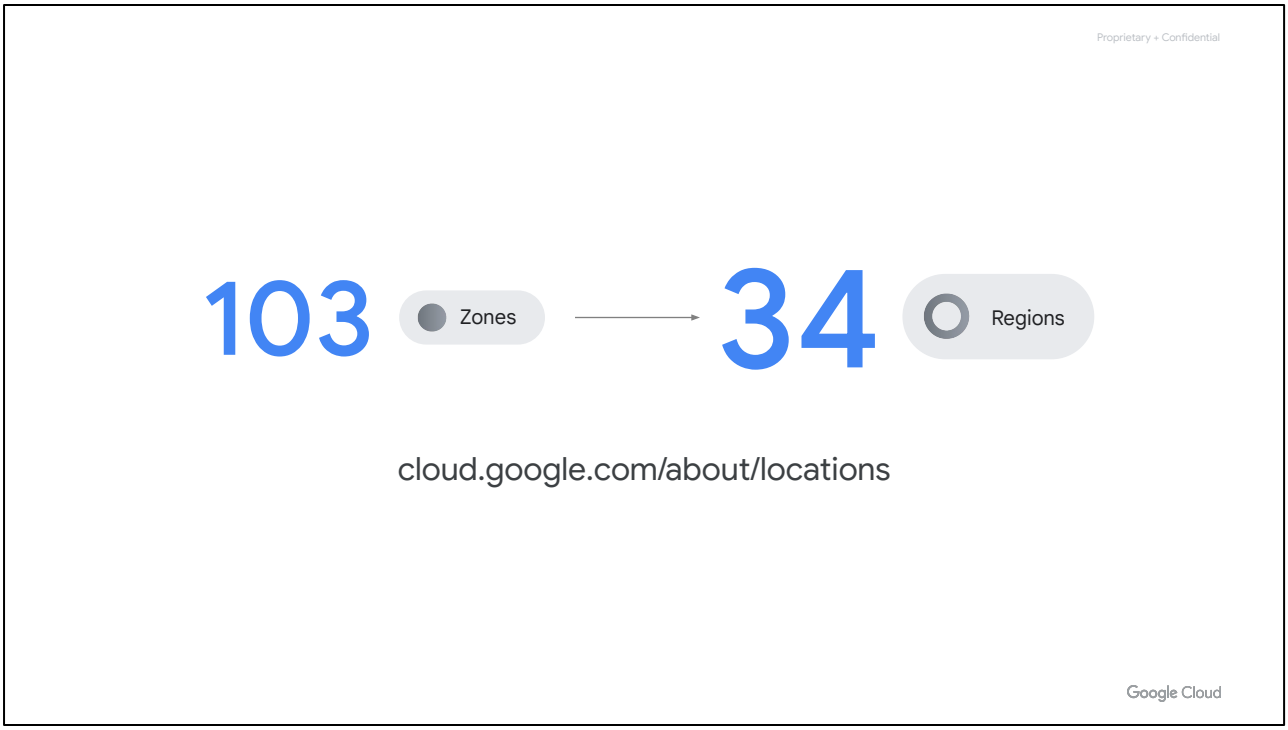
Some services can run in multiple geographic locations



Some of Google Cloud's services support placing resources in what we call a multi-region.

For example, Cloud Spanner multi-region configurations allow you to replicate the database's data not just in multiple zones, but in multiple zones across multiple regions, as defined by the instance configuration.

These additional replicas enable you to read data with low latency from multiple locations close to or within the regions in the configuration, like The Netherlands and Belgium.



Google Cloud currently supports 103 zones in 34 regions, though this is increasing all the time. The most up to date info can be found at cloud.google.com/about/locations.

Introducing Google Cloud

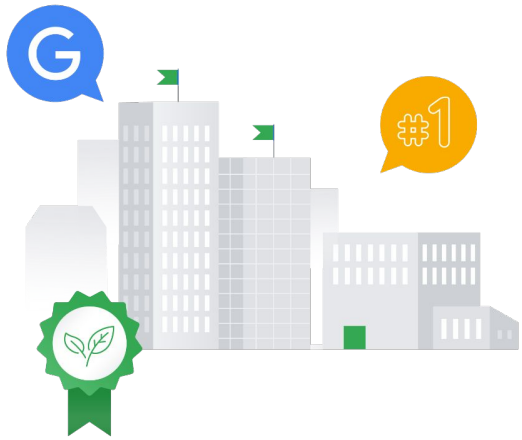
Data center energy consumption

2%

of the world's electricity

The virtual world, which includes Google Cloud's network, is built on physical infrastructure, and all those racks of humming servers use huge amounts of energy. Together, all existing data centers use roughly 2% of the world's electricity. So, Google works to make data centers run as efficiently as possible.

Google aims to improve efficiency and reduce waste



Google's data centers were the first to achieve ISO 14001 certification

Just like our customers, Google is trying to do the right things for the planet. We understand that Google Cloud customers have environmental goals of their own, and running their workloads in Google Cloud can be a part of meeting them.

Therefore, it's important to note that Google's data centers were the first to achieve ISO 14001 certification, which is a standard that maps out a framework for improving resource efficiency and reducing waste.

The data center cooling system in Finland is the first of its kind anywhere in the world.

Google's data center, Hamina, Finland

This is Google's data center in Hamina, Finland. The facility is one of the most advanced and efficient data centers in the Google fleet. Its cooling system, which uses sea water from the Bay of Finland, reduces energy use and is the first of its kind anywhere in the world.

In our founding decade, Google became the first major company to be carbon neutral. In our second decade, we were the first company to achieve 100% renewable energy. By 2030, we aim to be the first major company to operate carbon free.

Introducing Google Cloud

Google Cloud is designed with security in mind

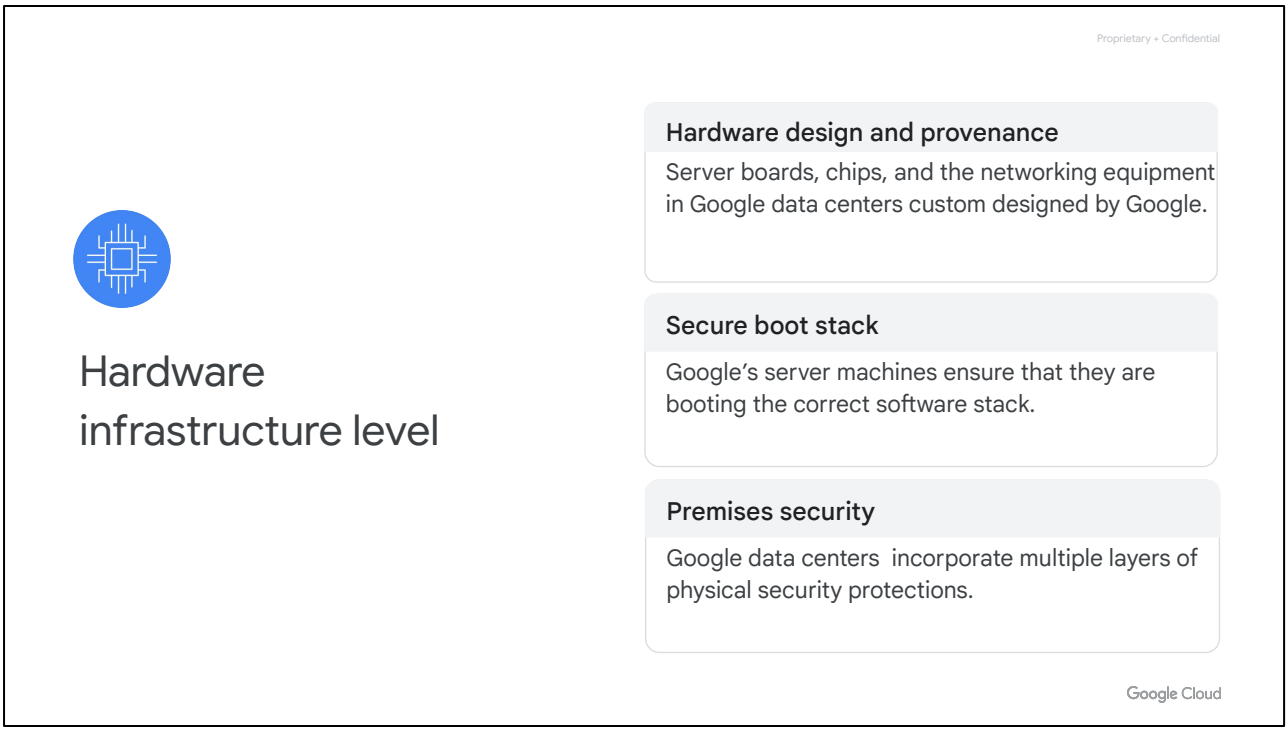
Because Google has nine* services with more than a billion users, you can bet security is always on the minds of Google's employees. Design for security is prevalent, throughout the infrastructure, that Google Cloud and Google services run-on.

Let's talk about a few ways Google works to keep customers' data safe.

*[Android, Chrome, Gmail, Drive, Maps, Photos, Search, Play Store, YouTube]

Hardware infrastructure level

At the **Hardware infrastructure** level:



Hardware design and provenance: Both the server boards and the networking equipment in Google data centers are custom designed by Google. Google also designs custom chips, including a hardware security chip that's currently being deployed on both servers and peripherals.

Secure boot stack: Google server machines use a variety of technologies to ensure that they are booting the correct software stack, such as cryptographic signatures over the BIOS, bootloader, kernel, and base operating system image.

Premises security: Google designs and builds its own data centers, which incorporate multiple layers of physical security protections. Access to these data centers is limited to only a very small fraction of Google employees. Google additionally hosts some servers in third-party data centers, where we ensure that there are Google-controlled physical security measures on top of the security layers provided by the data center operator.

Service
Deployment level

At the **Service deployment** level:

Service deployment level

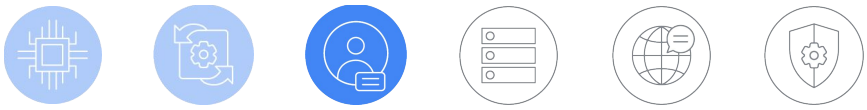
Encryption of inter-service communication

Services communicate with each other using remote procedure calls ("RPCs").

Google's infrastructure encrypts all infrastructure RPC traffic between data centers.

Encryption of inter-service communication: Google's infrastructure provides cryptographic privacy and integrity for remote procedure call ("RPC") data on the network. Google's services communicate with each other using RPC calls. The infrastructure automatically encrypts all infrastructure RPC traffic which goes between data centers. Google has started to deploy hardware cryptographic accelerators that will allow it to extend this default encryption to all infrastructure RPC traffic inside Google data centers.

User identity level



At the **User identity** level:

User identity level

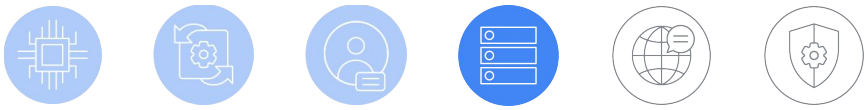
User identity

Intelligently challenges users for additional information based on certain risk factors.

Users can use secondary factors when signing in, including Factor (U2F) open standard.

User identity: Google's central identity service, which usually manifests to end users as the Google login page, goes beyond asking for a simple username and password. The service also intelligently challenges users for additional information based on risk factors such as whether they have logged in from the same device or a similar location in the past. Users also have the option of employing secondary factors when signing in, including devices based on the Universal 2nd Factor (U2F) open standard.

Storage
services level



At the **Storage services** level:

Storage services level

Encryption at rest

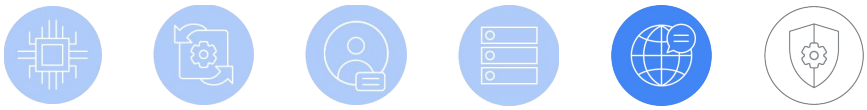
Most Google applications access file storage indirectly via storage services and encryption.

Using centrally managed keys, encryption is applied at the layer of these storage services.

Google also enables hardware encryption support in hard drives and SSDs.

Encryption at rest: Most applications at Google access physical storage (in other words, “file storage”) indirectly via storage services, and encryption (using centrally managed keys) is applied at the layer of these storage services. Google also enables hardware encryption support in hard drives and SSDs.

Internet
communication level



At the **Internet communication** level:

Internet communication level

Google Front End ("GFE")

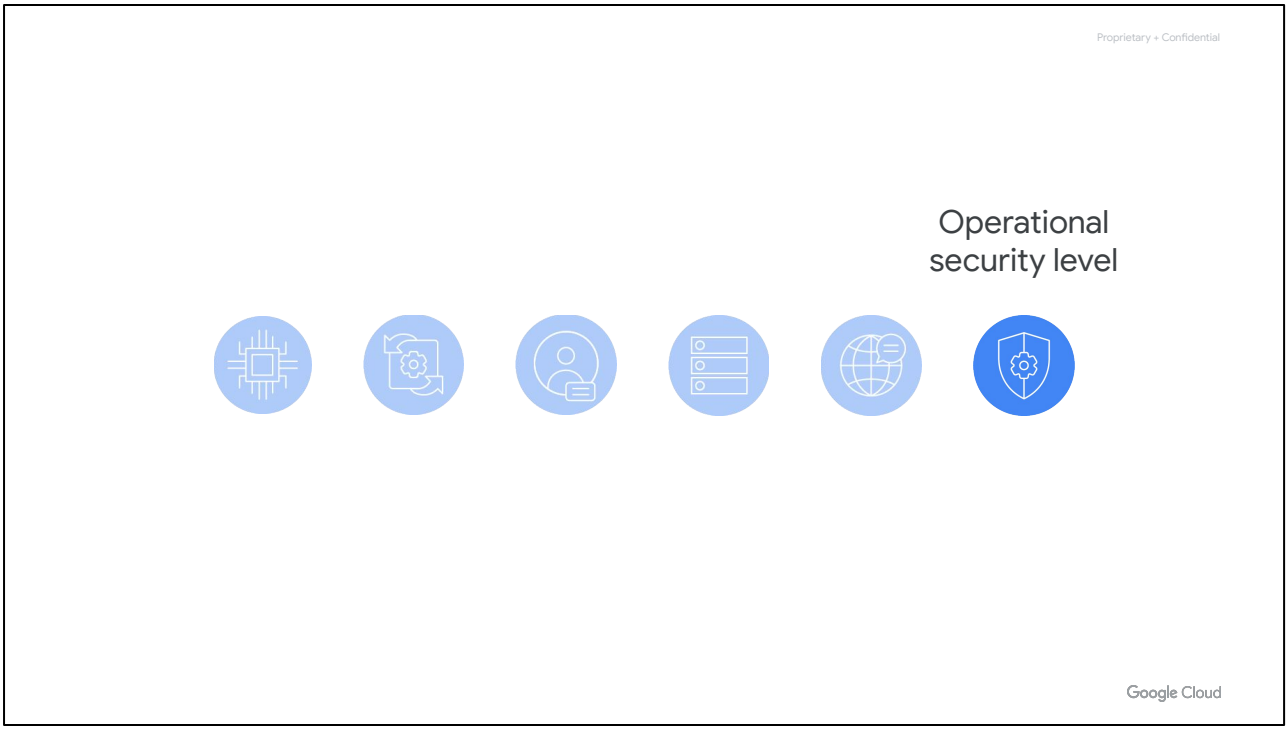
Google Front End ensures that all registered services use TLS connections incorporating the correct certificates and following best practices.

Denial of Service ("DoS") protection

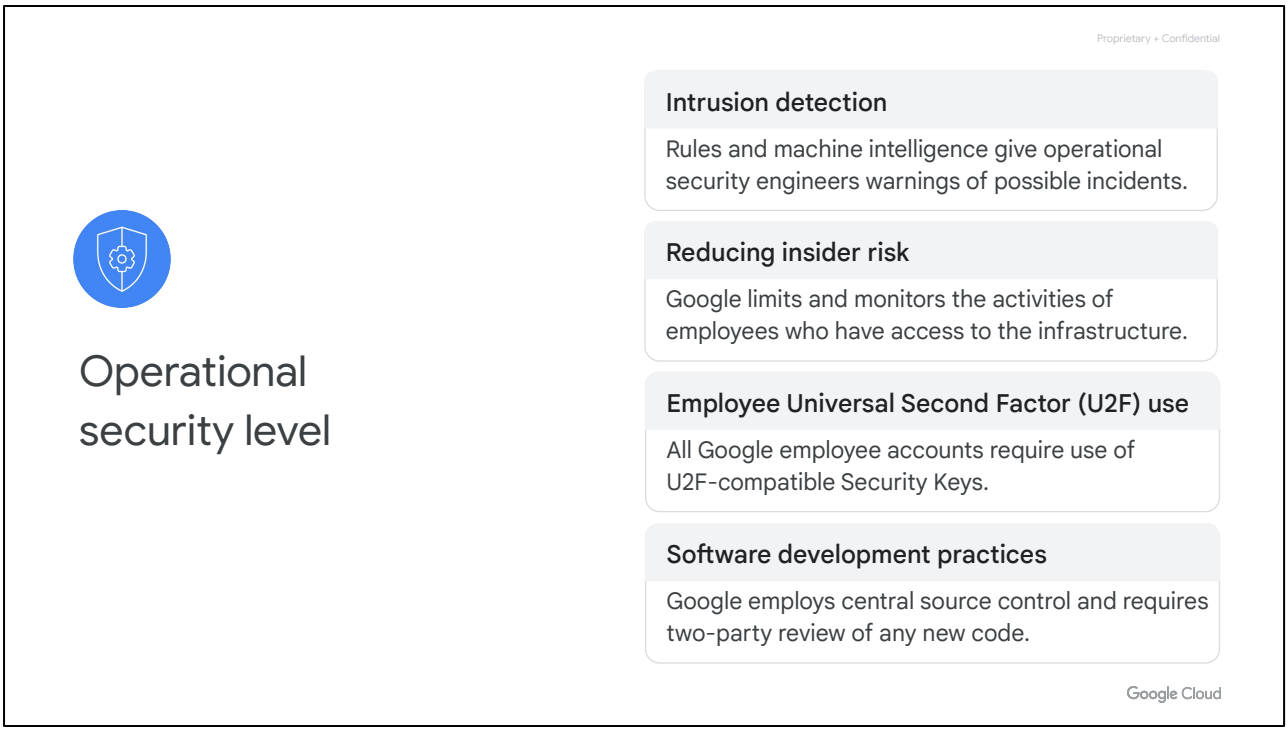
Google has multi-tier, multi-layer DoS protections that reduce the risk of any DoS impact on a service running behind a GFE

Google Front End ("GFE"): Google services that want to make themselves available on the Internet register themselves with an infrastructure service called the Google Front End, which ensures that all TLS connections are ended using a public-private key pair and an X.509 certificate from a Certified Authority (CA) as well as following best practices such as supporting perfect forward secrecy. The GFE additionally applies protections against Denial of Service attacks.

Denial of Service ("DoS") protection: The sheer scale of its infrastructure enables Google to simply absorb many DoS attacks. Google also has multi-tier, multi-layer DoS protections that further reduce the risk of any DoS impact on a service running behind a GFE.



Finally, at Google's **Operational security** level:



Intrusion detection: Rules and machine intelligence give Google's operational security teams warnings of possible incidents. Google conducts Red Team exercises to measure and improve the effectiveness of its detection and response mechanisms.

Reducing insider risk: Google aggressively limits and actively monitors the activities of employees who have been granted administrative access to the infrastructure.

Employee U2F use: To guard against phishing attacks against Google employees, employee accounts require use of U2F-compatible Security Keys.

Software development practices: Google employs central source control and requires two-party review of new code. Google also provides its developers libraries that prevent them from introducing certain classes of security bugs. Google also runs a Vulnerability Rewards Program where we pay anyone who is able to discover and inform us of bugs in our infrastructure or applications.

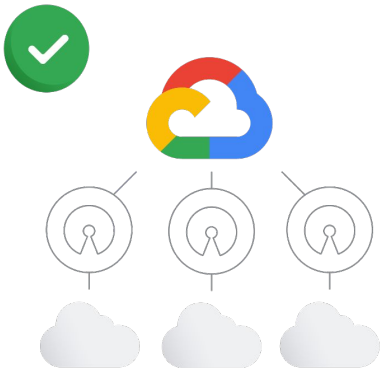
cloud.google.com/security/security-design

You can learn more about Google's technical-infrastructure security at cloud.google.com/security/security-design.

Introducing Google Cloud

Google Cloud is open-source friendly

Data locked into
a particular vendor



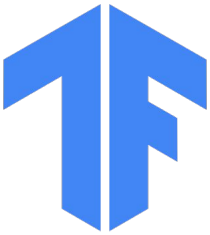
Google give customers the ability to run their
applications elsewhere

Some people are afraid to bring their workloads to the cloud because they're afraid they'll get locked into a particular vendor.

But if, for whatever reason, a customer decides that Google is no longer the best provider for their needs, we give customers the ability to run their applications elsewhere.

Google publishes key elements of technology using open source licenses to create ecosystems that provide customers with options other than Google.

Open-source TensorFlow works with Google Cloud



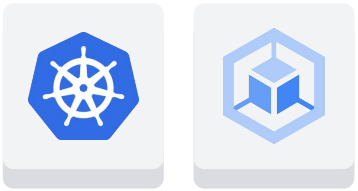
TensorFlow is an open source library for machine learning

TensorFlow is at the heart of a strong open source ecosystem

For example, TensorFlow, an open source software library for machine learning developed inside Google, is at the heart of a strong open source ecosystem.

Interoperability at multiple layers of the stack

Kubernetes GKE



Mix and **match microservices**
running across different clouds

Google Cloud's
operations suite

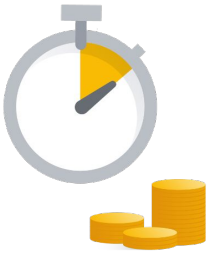
Monitor workloads across
multiple cloud providers

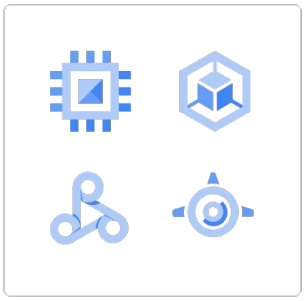
Google provides interoperability at multiple layers of the stack. Kubernetes and Google Kubernetes Engine give customers the ability to mix and match microservices running across different clouds. Google Cloud's operations suite lets customers monitor workloads across multiple cloud providers.

Introducing Google Cloud

Now a word on Google Cloud's pricing structure.

Google Compute products are billed per-second

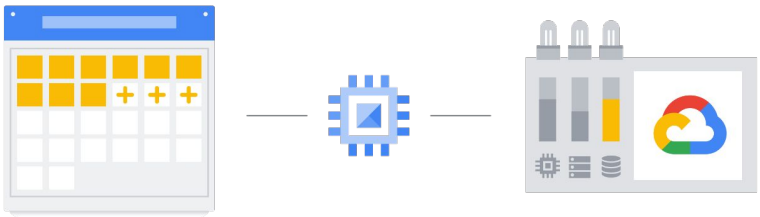




Google was the first major cloud provider to deliver per-second billing for its Infrastructure-as-a-Service compute offering, Compute Engine.

We'll explore these products and services later in this course, but per-second billing is offered for users of Compute Engine, Google Kubernetes Engine (container infrastructure as a service), Dataproc (which is the equivalent of the big data system Hadoop, but operating as a service), and App Engine flexible environment VMs (a platform as a service).

Compute Engine offers billing flexibility

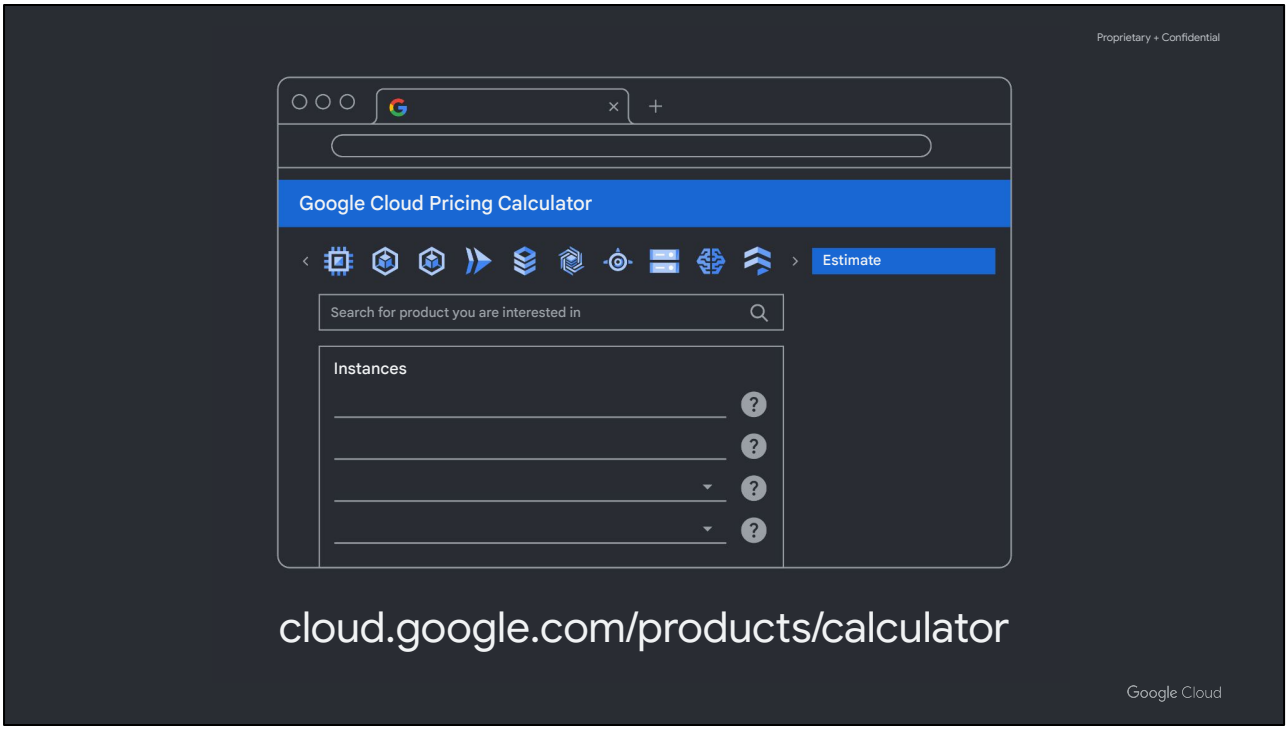


Sustained-use discounts

Custom virtual machine types

Compute Engine offers automatically applied sustained-use discounts, which are automatic discounts that you get for running a virtual-machine instance for a significant portion of the billing month. Specifically, when you run an instance for more than 25% of a month, Compute Engine automatically gives you a discount for every incremental minute you use for that instance.

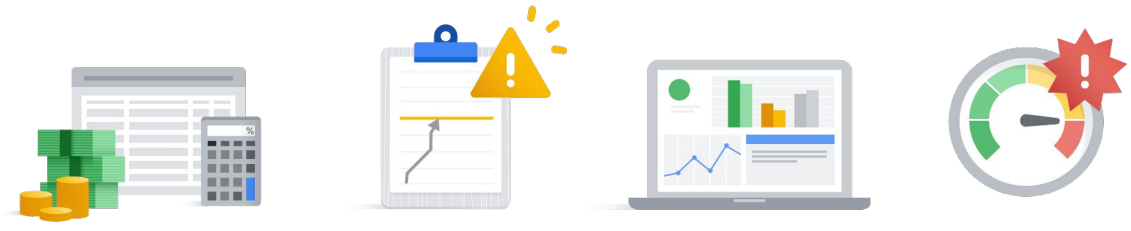
Custom virtual machine types allow Compute Engine virtual machines to be fine-tuned with optimal amounts of vCPU and memory for their applications, so that you can tailor your pricing for your workloads.



Our online pricing calculator can help estimate your costs.

Visit cloud.google.com/products/calculator to try it out.

Billing tools help to budget and monitor usage



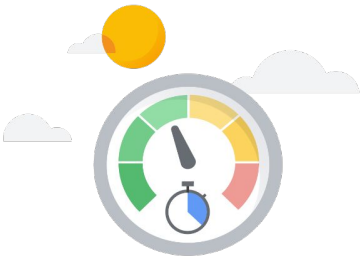
Budgets Alerts Reports Quotas

You're probably thinking, "How can I make sure I don't accidentally run up a big Google Cloud bill?"

We provide a few tools to help.

1. You can define **budgets** at the billing account level or at the project level. A budget can be a fixed limit, or it can be tied to another metric - for example, a percentage of the previous month's spend.
2. To be notified when costs approach your budget limit, you can create an **alert**. For example, with a budget limit of $20,000 and an alert set at 90%, you'll receive a notification alert when your expenses reach $18,000. Alerts are generally set at 50%, 90% and 100%, but can also be customized.
3. **Reports** is a visual tool in the Google Cloud console that allows you to monitor expenditure based on a project or services.
4. Finally, Google Cloud also implements **quotas**, which are designed to prevent the over-consumption of resources because of an error or a malicious attack, protecting both account owners and the Google Cloud community as a whole.

Quotas are allocated at project-level



Rate quota

Resets after
a specific time

Allocation quota

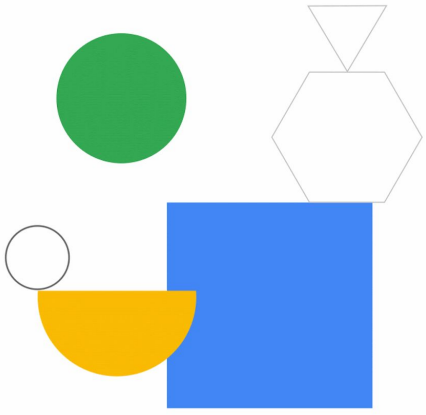
Governs the number
of resources in a project

There are two types of quotas: **rate quotas** and **allocation quotas**. Both are applied at the project level.

1. Rate quotas reset after a specific time. For example, by default, the GKE service implements a quota of 1,000 calls to its API from each Google Cloud project every 100 seconds. After that 100 seconds, the limit is reset.
2. Allocation quotas govern the number of resources you can have in your projects. For example, by default, each Google Cloud project has a quota allowing it no more than 5 Virtual Private Cloud networks.

Although projects all start with the same quotas, you can change some of them by requesting an increase from Google Cloud Support.

Module Quiz



Quiz | Question 1

Question

Why might a Google Cloud customer use resources in several zones within a region?

A. For improved fault tolerance

B. For better performance

C. For expanding services to customers in new areas

D. For getting discounts on other zones

Quiz | Question 1

Answer

Why might a Google Cloud customer use resources in several zones within a region?

A. For improved fault tolerance 

B. For better performance

C. For expanding services to customers in new areas

D. For getting discounts on other zones

Why might a Google Cloud customer use resources in several zones within a region?

***A: For improved fault tolerance**
Feedback: Correct. As part of building a fault-tolerant application, you can spread your resources across multiple zones in a region.

B: For better performance
Feedback: That is not correct. The primary benefit of spreading your resources across multiple zones in a region is protection against failure.

C: For expanding services to customers in new areas
Feedback: That is not correct. The primary benefit of spreading your resources across multiple zones in a region is protection against failure.

D: For getting discounts on other zones
Feedback: That is not correct. The primary benefit of spreading your resources across multiple zones in a region is protection against failure.

Quiz | Question 2

Question

What type of cloud computing service lets you bind your application code to libraries that give access to the infrastructure your application needs?

A. Hybrid Cloud

B. Infrastructure as a Service

C. Software as a Service

D. Platform as a Service

E. Virtualized data centers

Quiz | Question 2

Answer

What type of cloud computing service lets you bind your application code to libraries that give access to the infrastructure your application needs?

A. Hybrid Cloud

B. Infrastructure as a Service

C. Software as a Service

D. Platform as a Service 

E. Virtualized data centers

What type of cloud computing service lets you bind your application code to libraries that give access to the infrastructure your application needs?

A: Hybrid cloud
Feedback: Review the lecture "GCP computing architectures."

B: Infrastructure as a Service
Feedback: Review the lecture "GCP computing architectures."

C: Software as a Service
Feedback: Review the lecture "GCP computing architectures."

D: Platform as a Service
Feedback: Correct!

E: Virtualized data centers
Feedback: Review the lecture "GCP computing architectures."

Quiz | Question 3

Question

Why might a Google Cloud customer use resources in several regions around the world?

A. To improve security

B. To earn discounts

C. To offer localized application versions in different regions

D. To bring their applications closer to users around the world, and for improved fault tolerance

Quiz | Question 3

Answer

Why might a Google Cloud customer use resources in several regions around the world?

A. To improve security

B. To earn discounts

C. To offer localized application versions in different regions

D. To bring their applications closer to users around the world, and for improved fault tolerance

Why might a Google Cloud customer use resources in several regions around the world?

A: To improve security
Feedback: Running an application in several regions does not necessarily improve security. Instead, using resources in multiple regions may give better performance for global users, and it also can protect against failures of entire regions.

B: To earn discounts
Feedback: Running an application in several regions does not necessarily earn discounts. Instead, using resources in multiple regions may give better performance for global users, and it also can protect against failures of entire regions.

C: To offer localized application versions in different regions.
Feedback: Google Cloud allows you to develop different localized versions of your application even if your application is based in a single region. Running an application in several regions has no impact on the localized versions of your application.

***D: To bring their applications closer to users around the world, and for improved fault tolerance**
Feedback: That is correct.