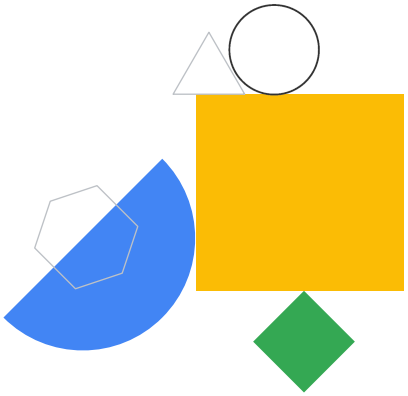
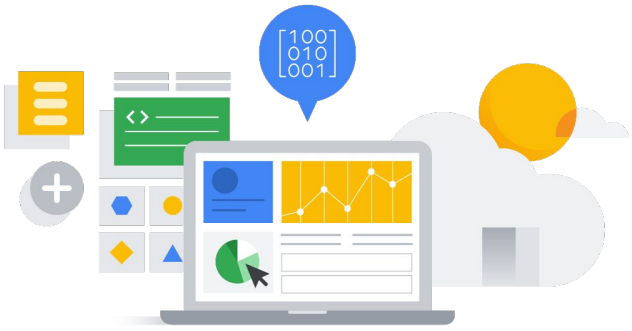
Google Cloud



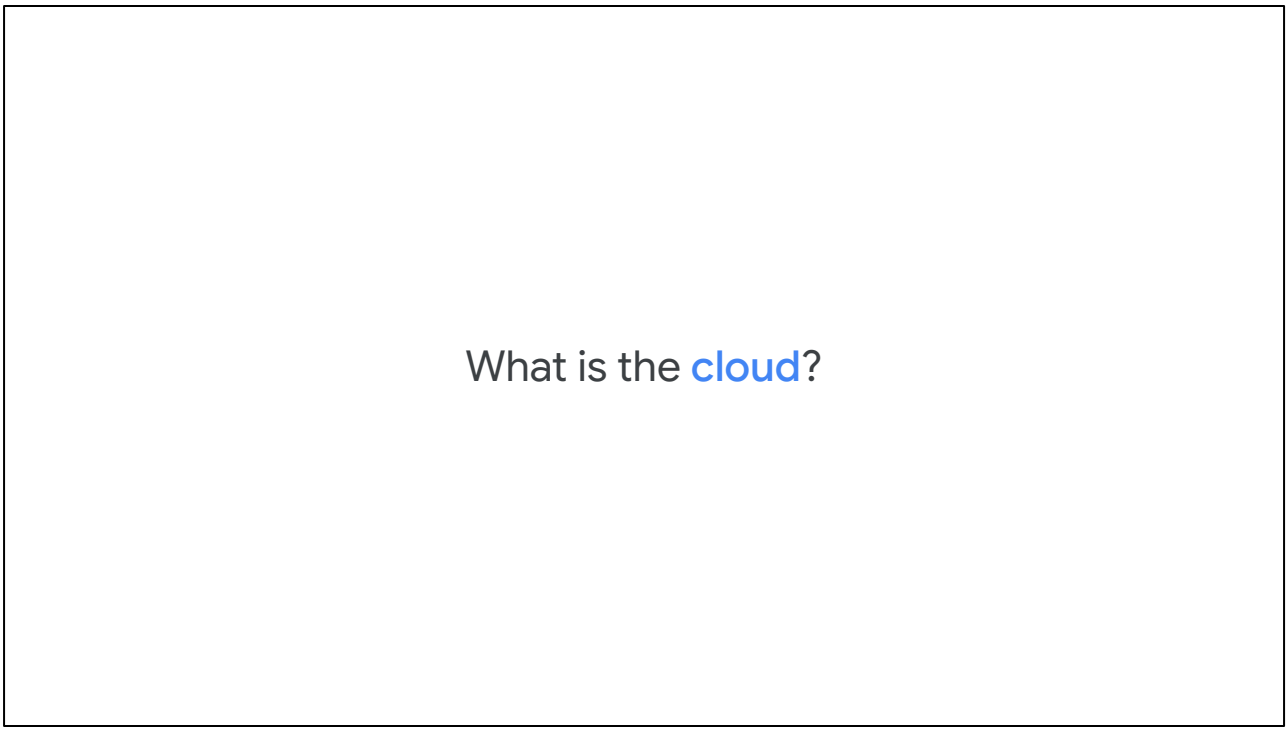
Google Cloud Core Infrastructure Module 1

On-demand course
March 2022



Cloud computing

Let’s start at the beginning with an overview of cloud computing.

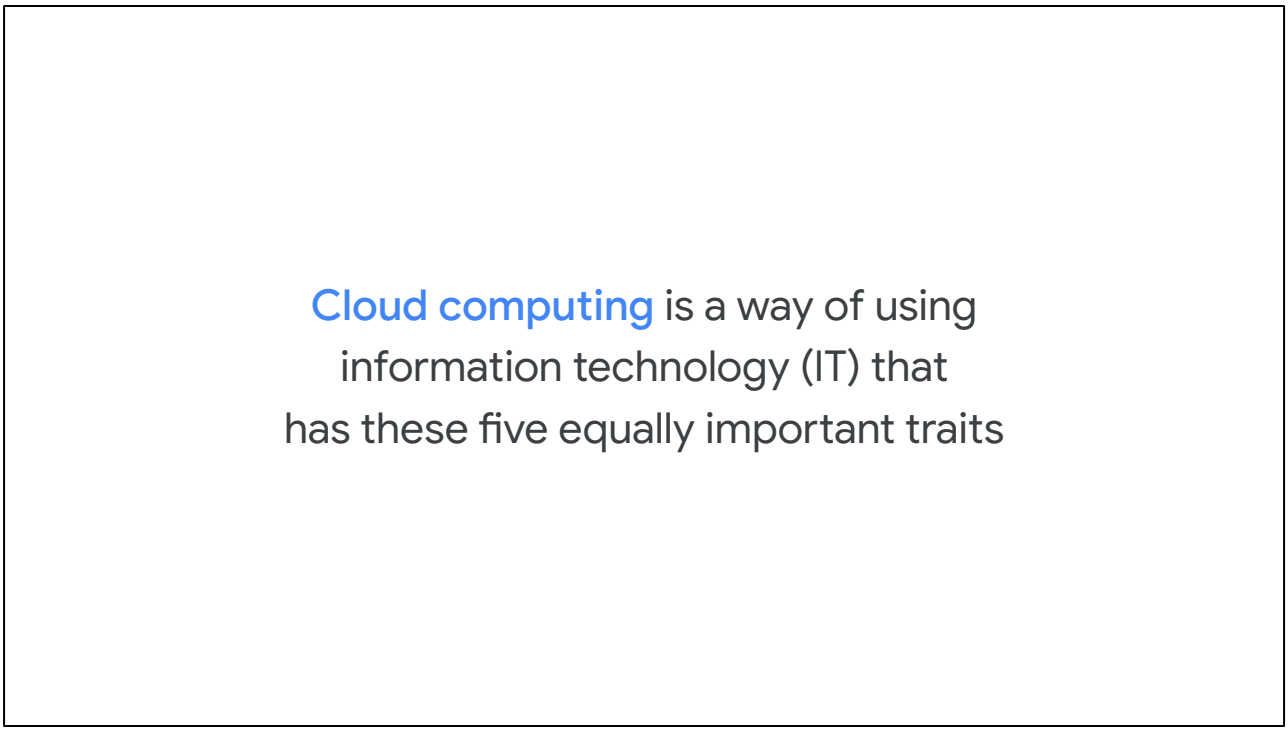


The cloud is a hot topic these days, but what exactly is it?

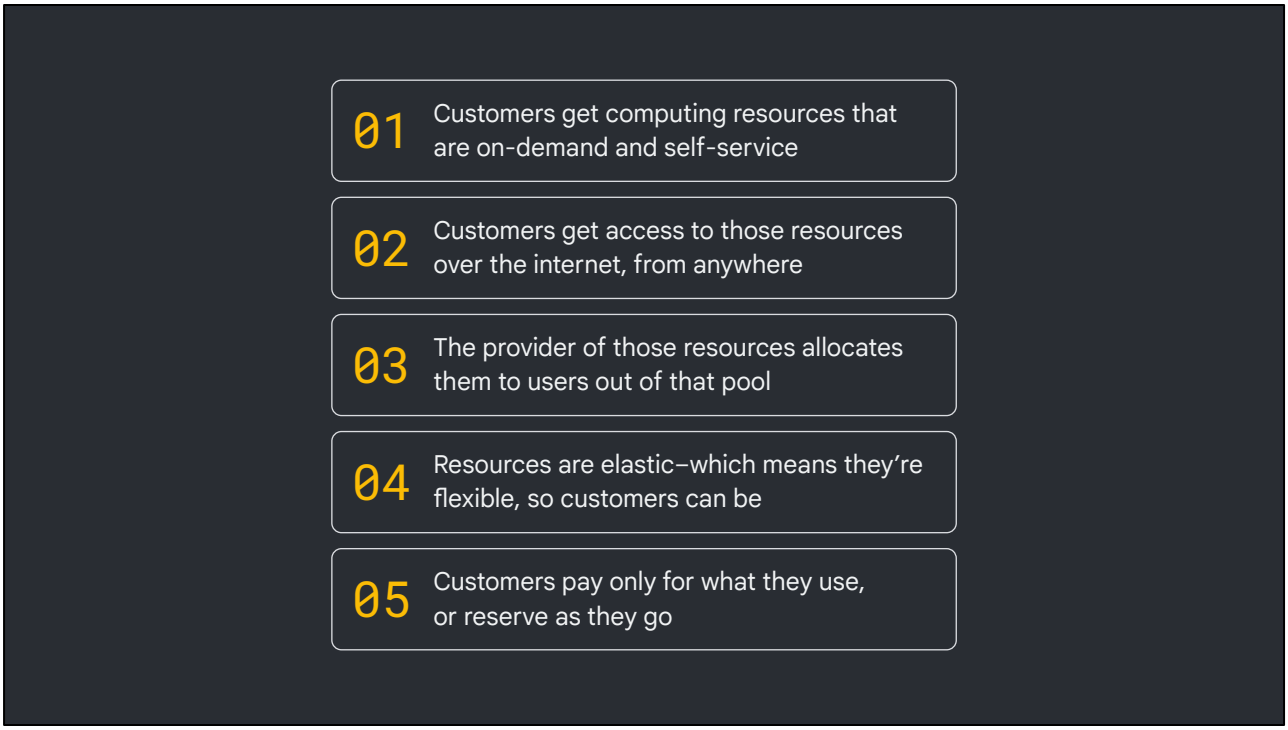
US National Institute of Standards and Technology

Cloud computing

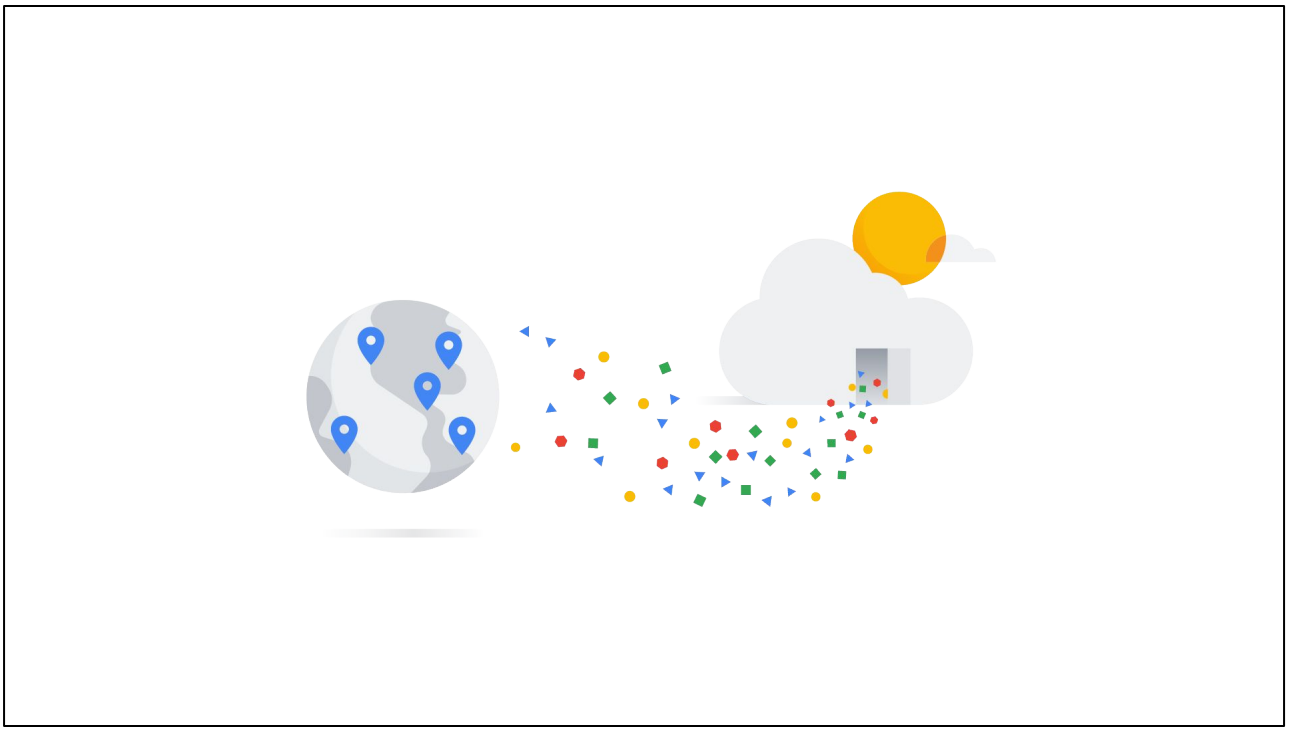
The US National Institute of Standards and Technology created the term **cloud computing**, although there is nothing US-specific about it.



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



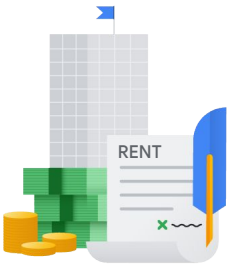
- First, customers get computing resources that are on-demand and self-service. Through a web interface, users get the processing power, storage, and network they need with no need for human intervention.
- Second, customers get access to those resources over the internet, from anywhere they have a connection.
- Third, the cloud provider has a big pool of those resources 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, 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.

Why is the cloud model
so compelling nowadays?

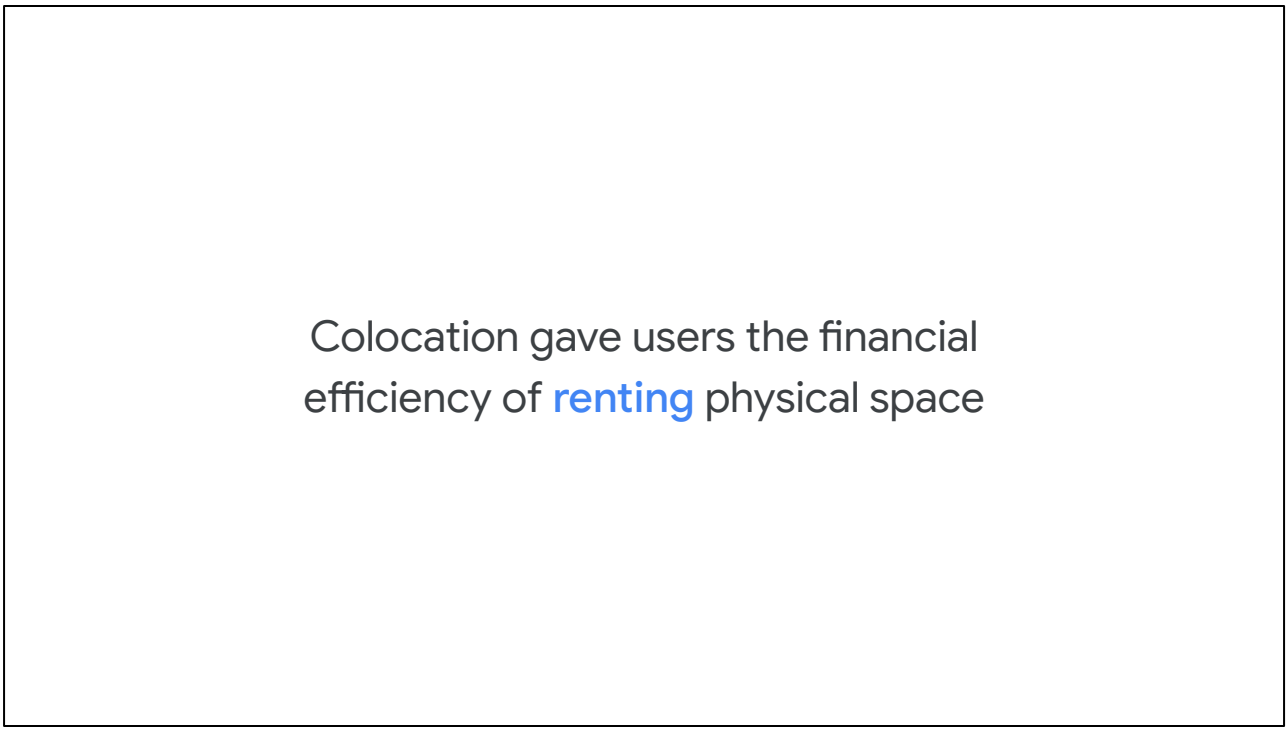
But why is the cloud model so compelling nowadays? To understand why, we need to look at some history.



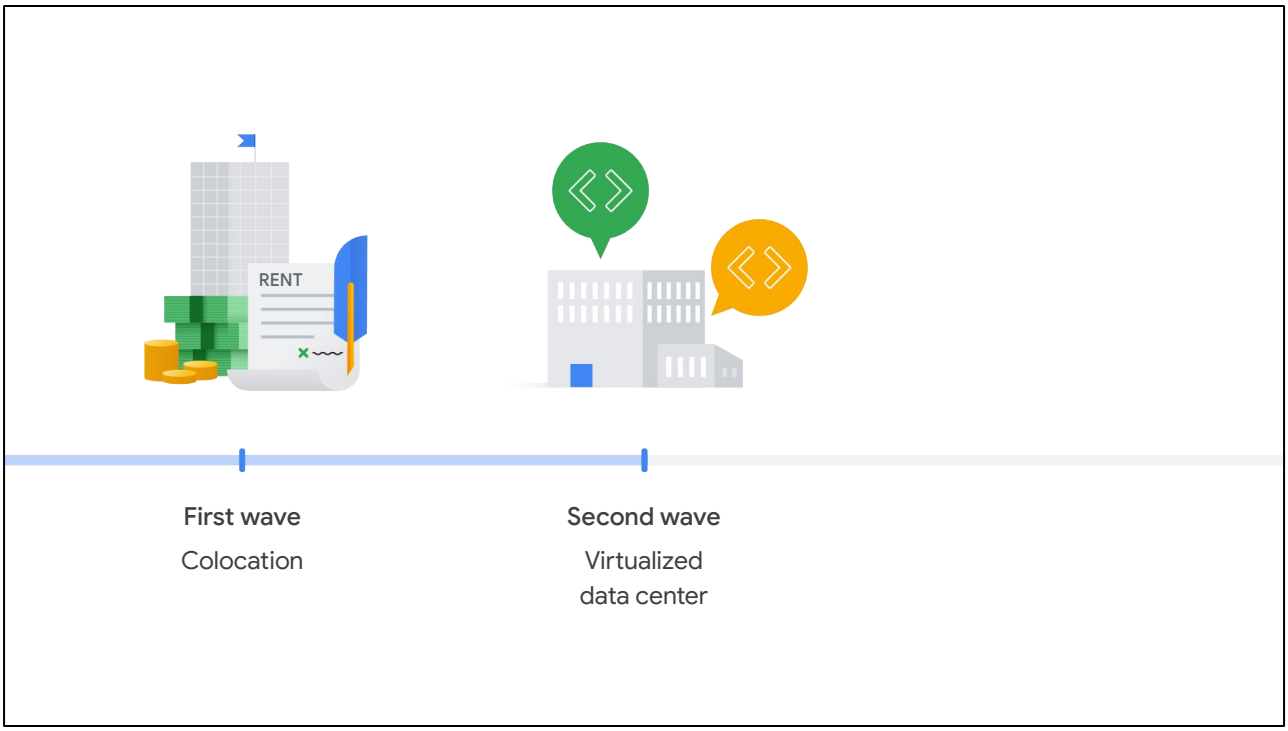
First wave

Colocation

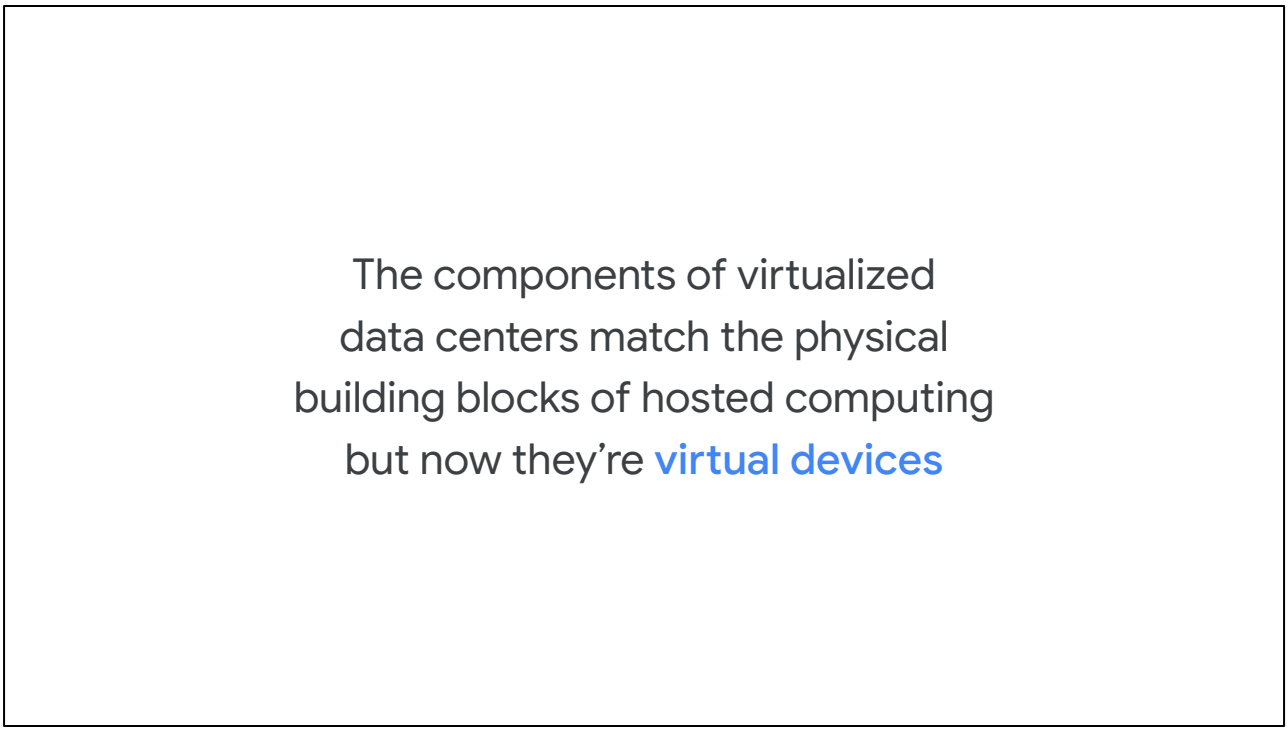
The trend towards cloud computing started with a first wave known as *colocation*.



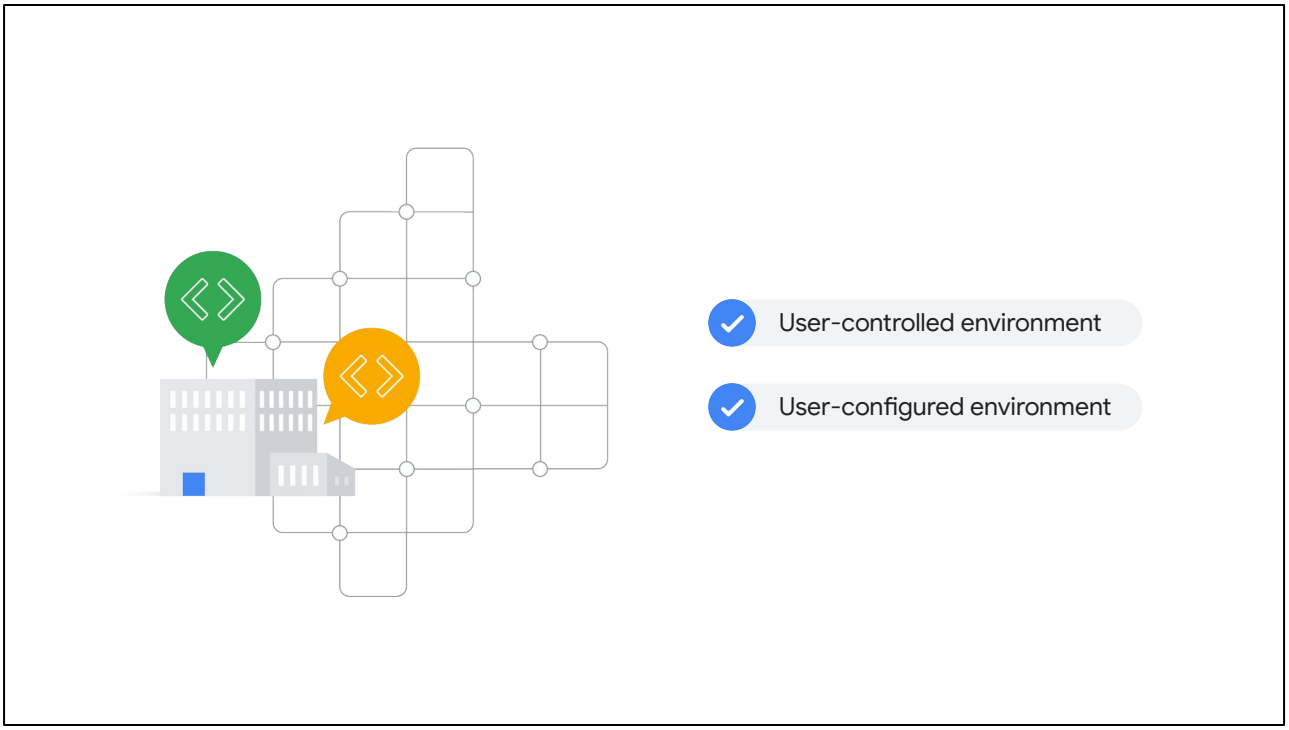
Colocation gave users the financial efficiency of renting physical space, instead of investing in data center real estate.



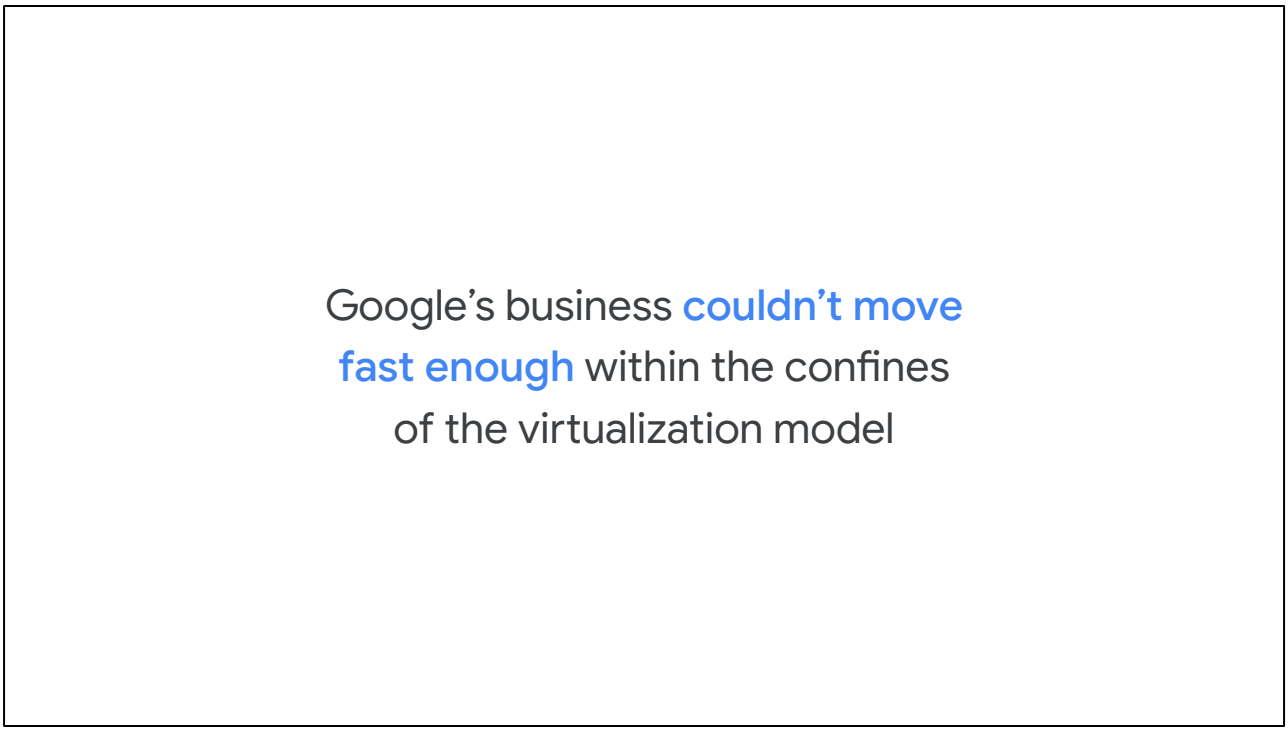
Virtualized data centers of today, which is the **second wave**, share similarities with the private data centers and colocation facilities of decades past.



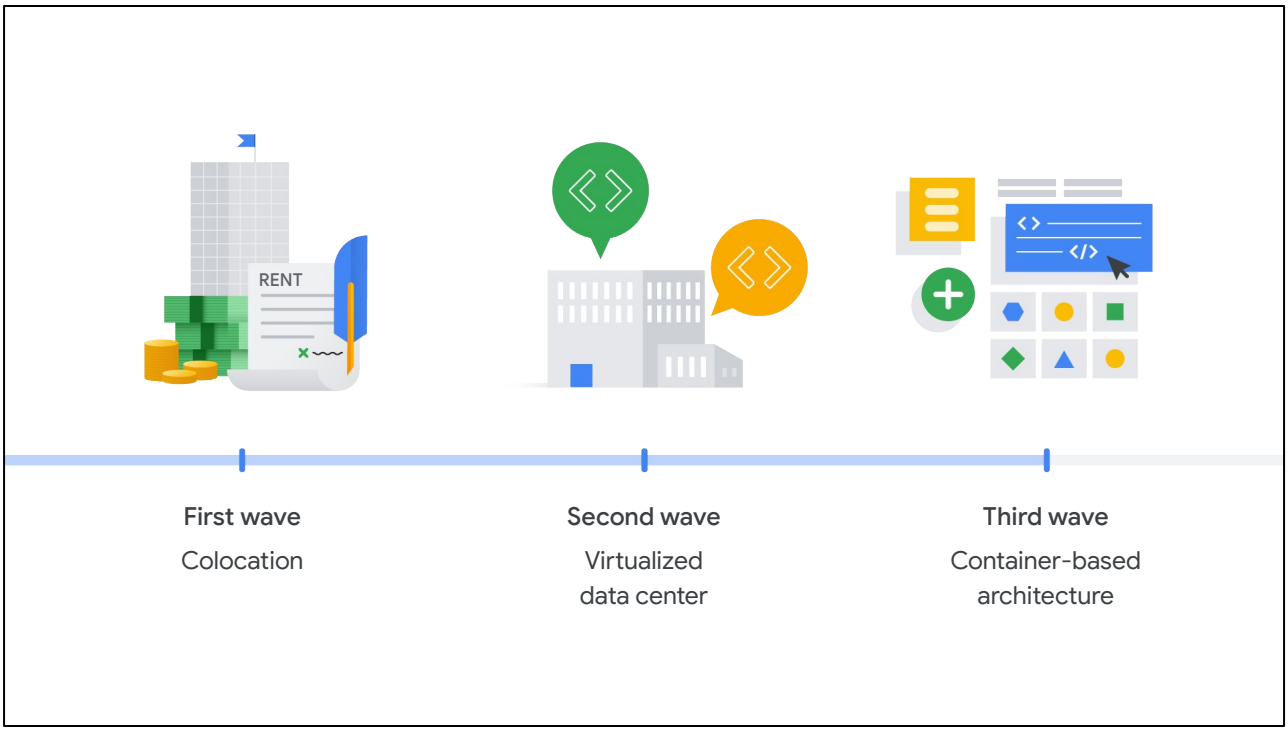
The components of virtualized data centers match the physical building blocks of hosted computing—servers, CPUs, disks, load balancers, and so on—but now they're virtual devices.



With virtualization, enterprises still maintain the infrastructure; but it also remains a user-controlled and user-configured environment.

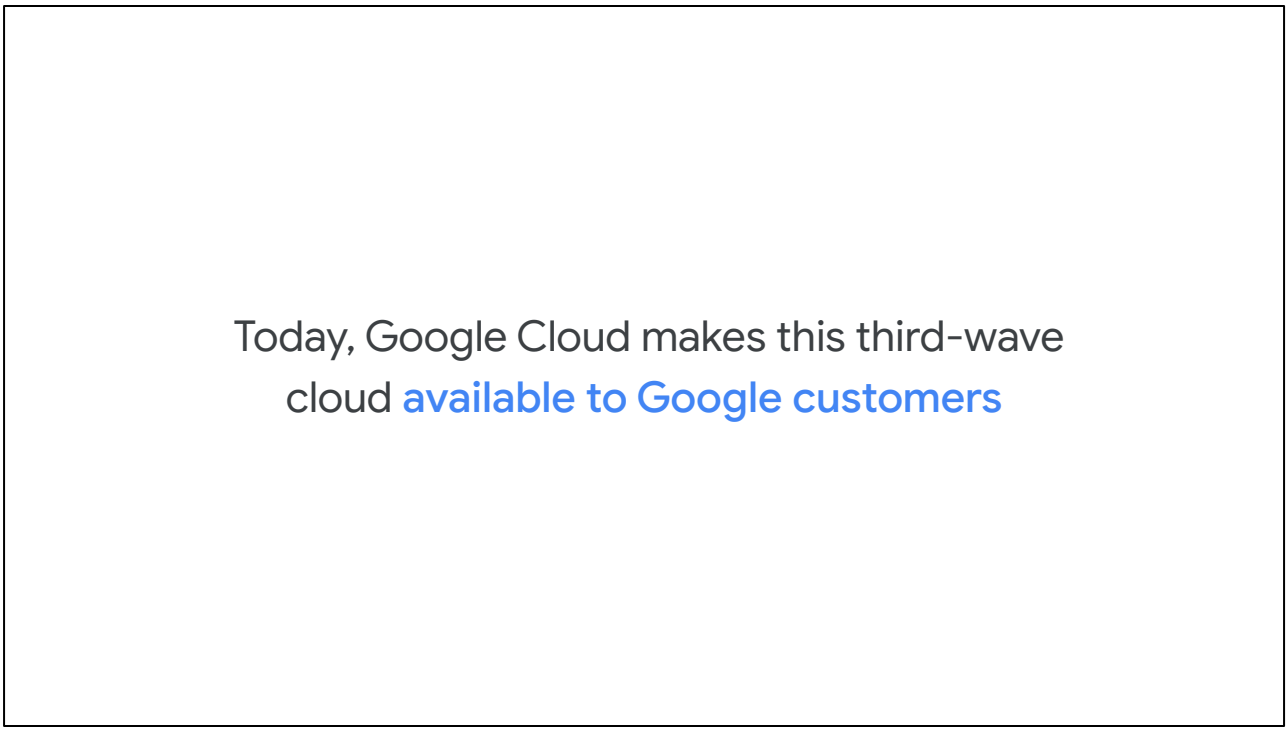


Several years ago, Google realized that its business couldn't move fast enough within the confines of the virtualization model.

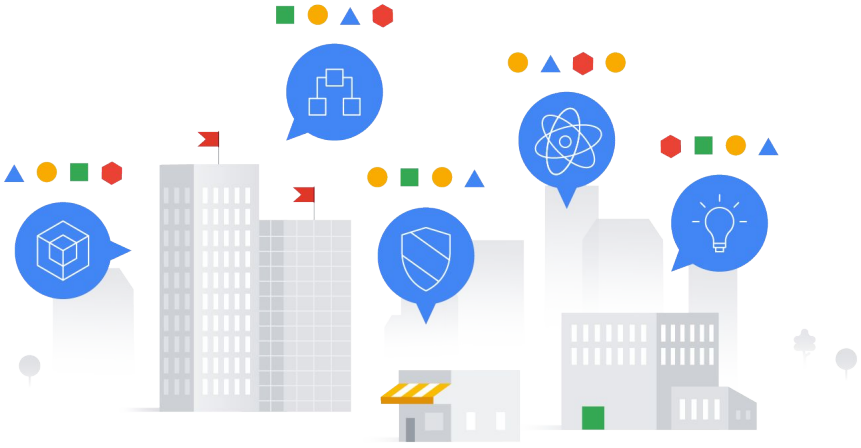


So Google switched to a *container-based architecture*—a fully automated, elastic third-wave cloud that consists of a combination of automated services and scalable data.

Services automatically provision and configure the infrastructure used to run applications.



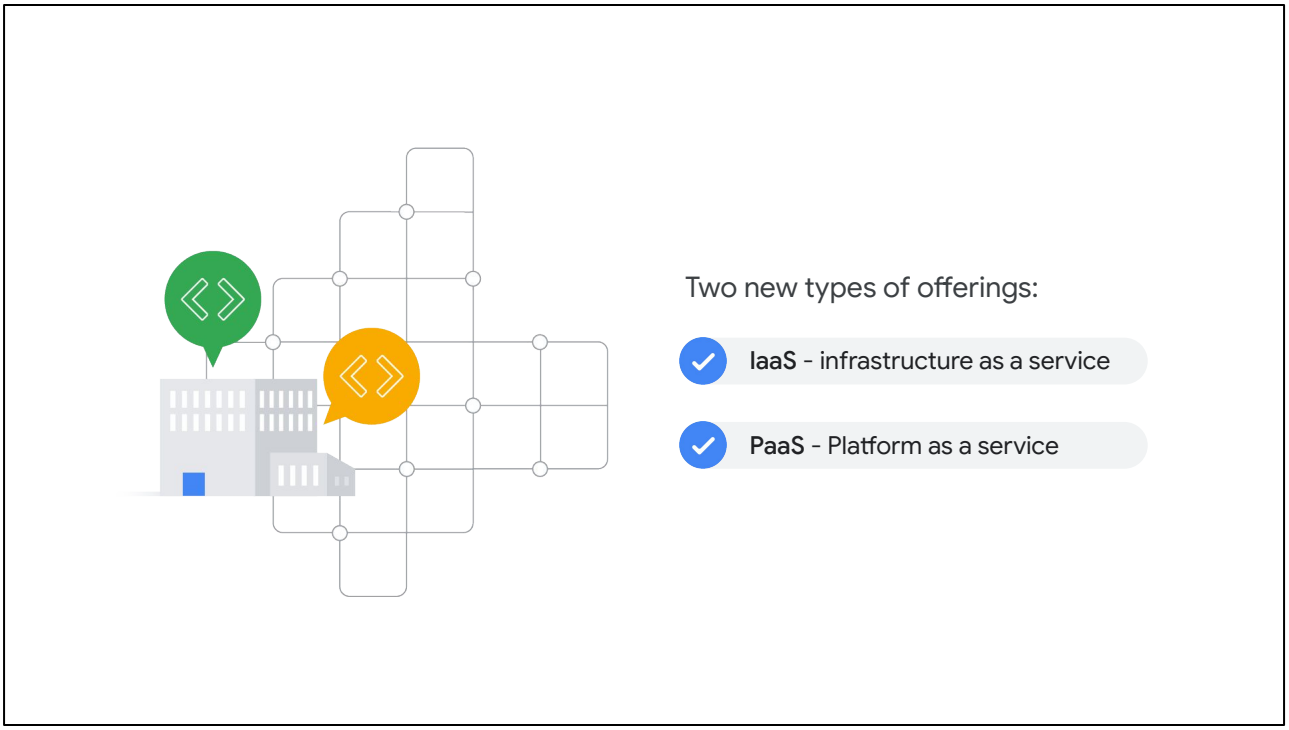
Today, Google Cloud makes this **third-wave** *cloud* available to Google customers.



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.

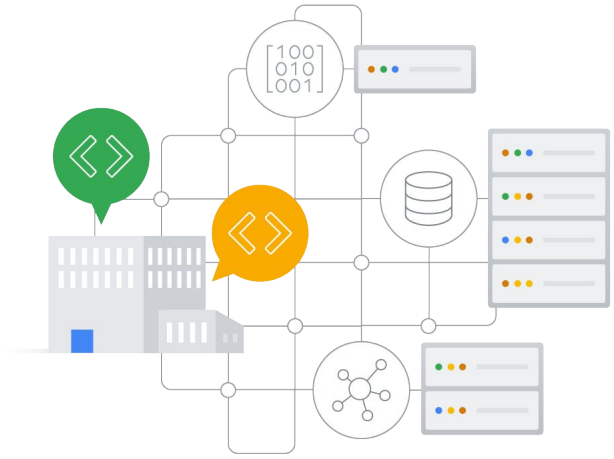


The move to virtualized data centers introduced customers to two new types of offerings:

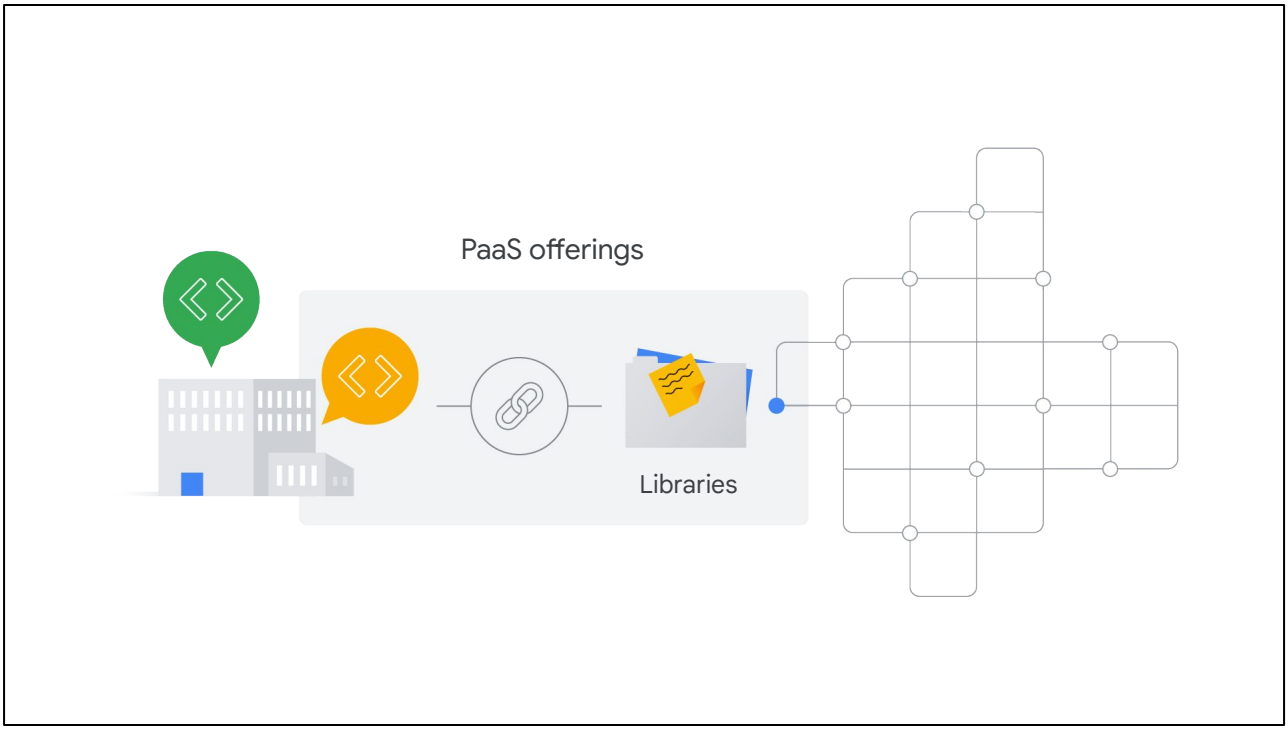
- **infrastructure as a service**, commonly referred to as IaaS, and
- **platform as a service**, or PaaS.

IaaS offerings provide:

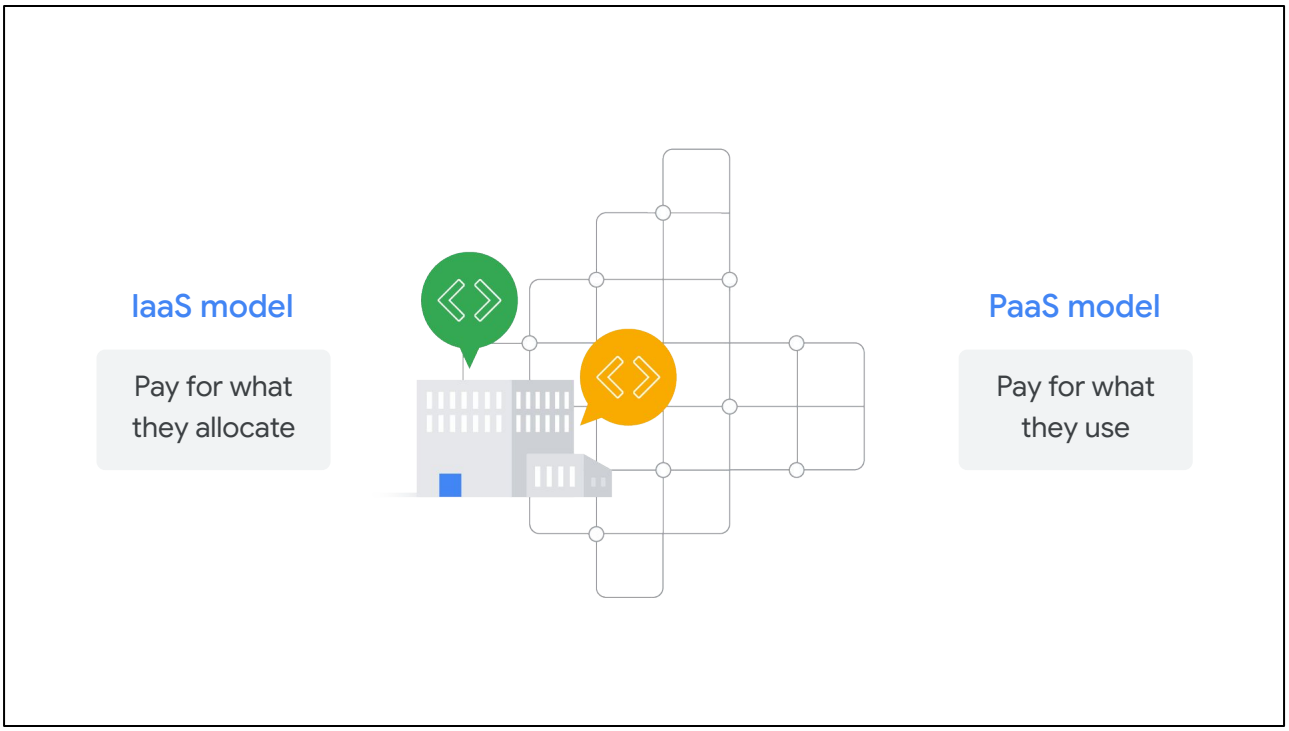
- Raw compute
- Storage
- Network capabilities



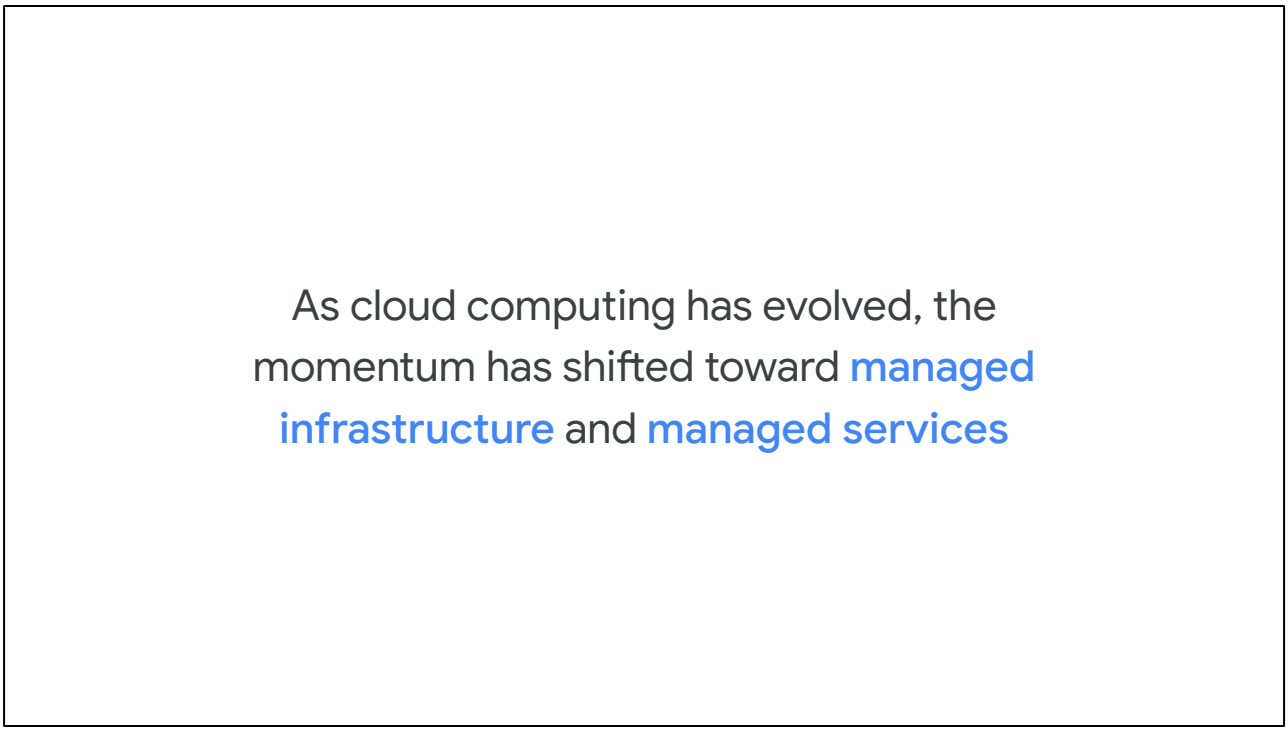
IaaS offerings provide raw compute, storage, and network capabilities, organized virtually into resources that are similar to physical data centers.



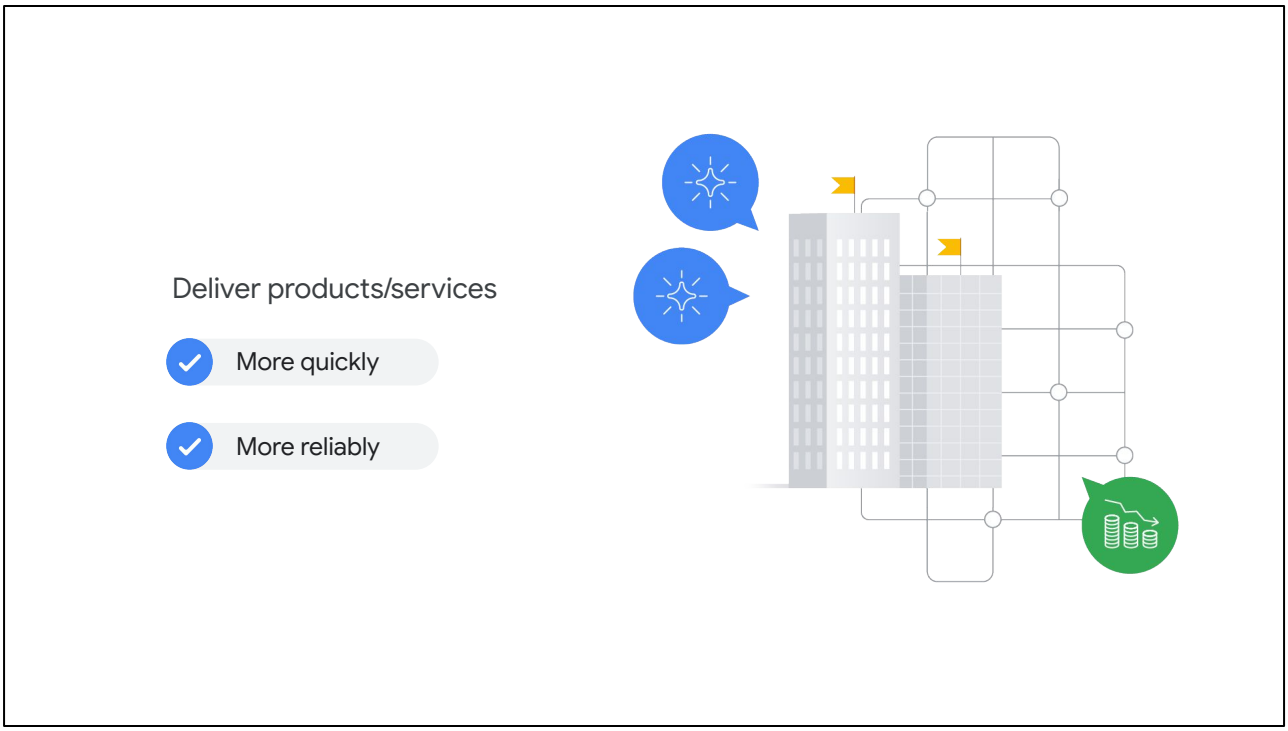
PaaS offerings, in contrast, bind code to libraries that provide access to the infrastructure application needs. This allows more resources to be focused on application logic.



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.

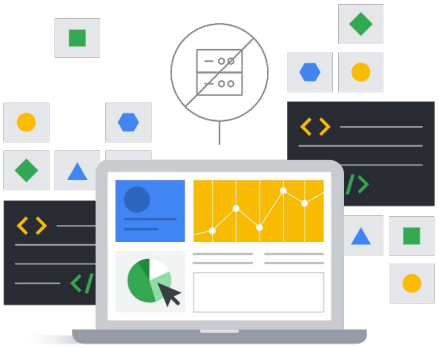


As cloud computing has evolved, the momentum has shifted toward managed infrastructure and managed services.



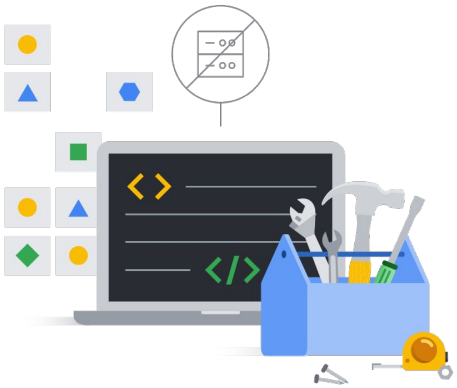
Leveraging managed resources and services allows companies to concentrate more on their business goals and spend less time and money on creating and maintaining their technical infrastructure.

It allows companies to deliver products and services to their customers more quickly and reliably.



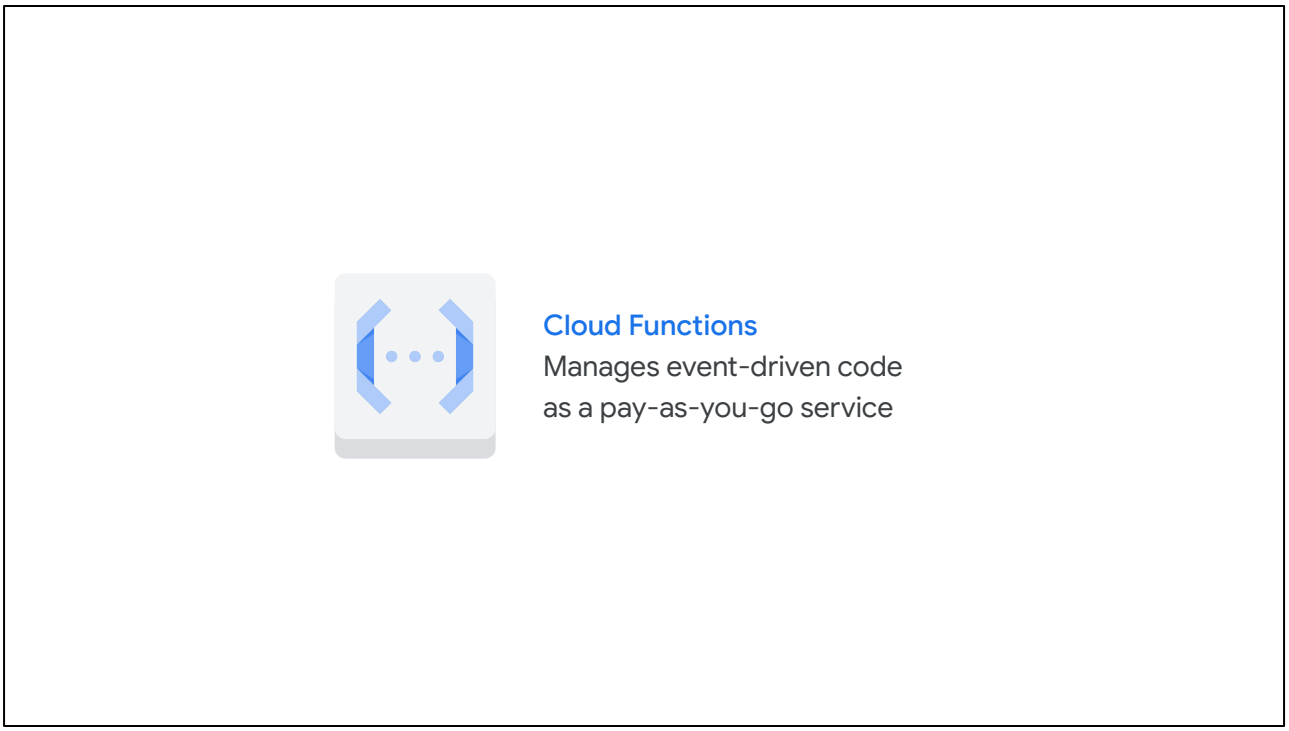
Serverless cloud computing

Serverless is yet another step in the evolution of cloud computing.

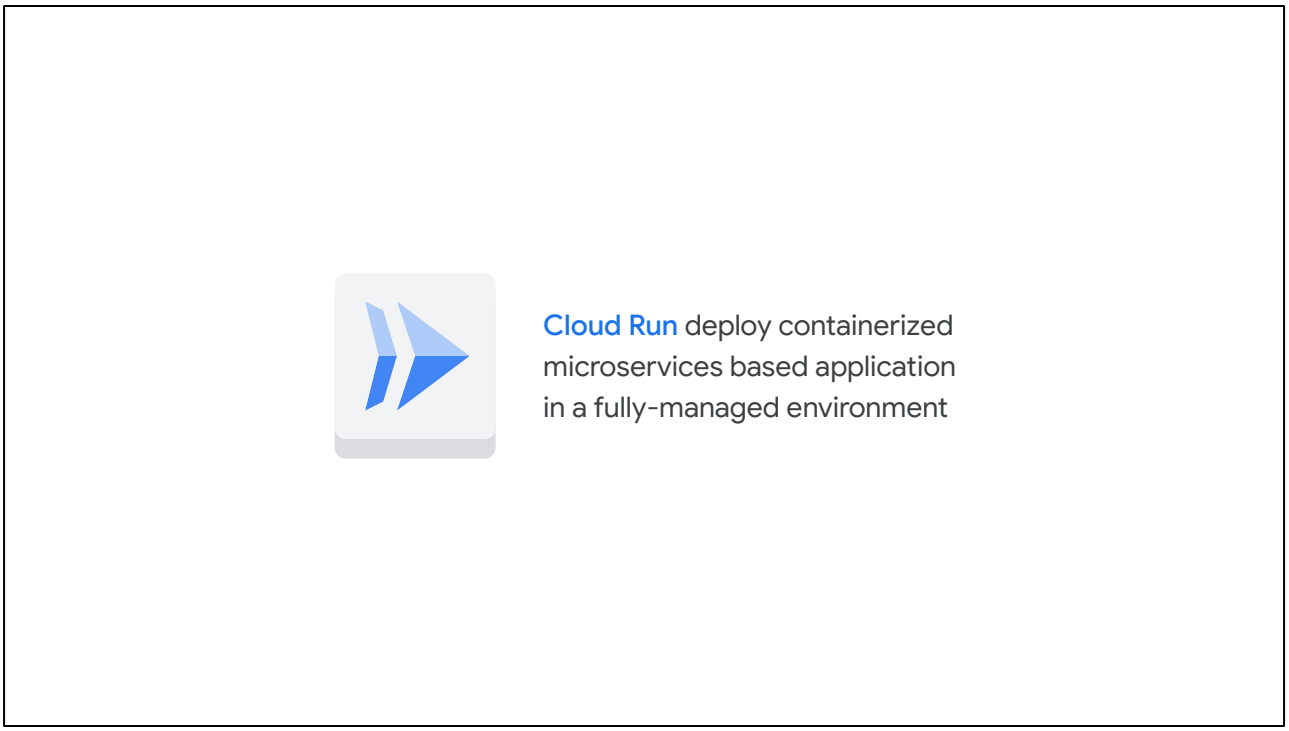


- Allows developers to concentrate on code
- No infrastructure management needed

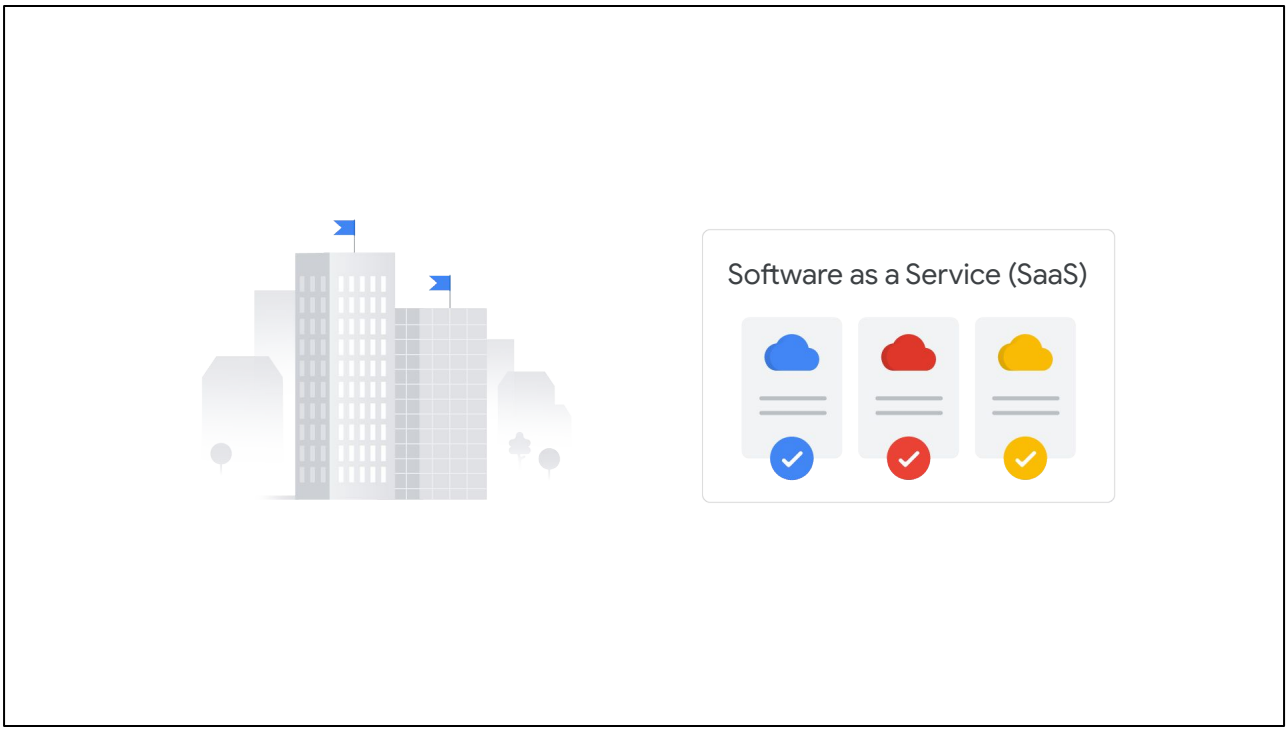
It allows developers to concentrate on their code, rather than on server configuration, by eliminating the need for any infrastructure management. Serverless technologies offered by Google



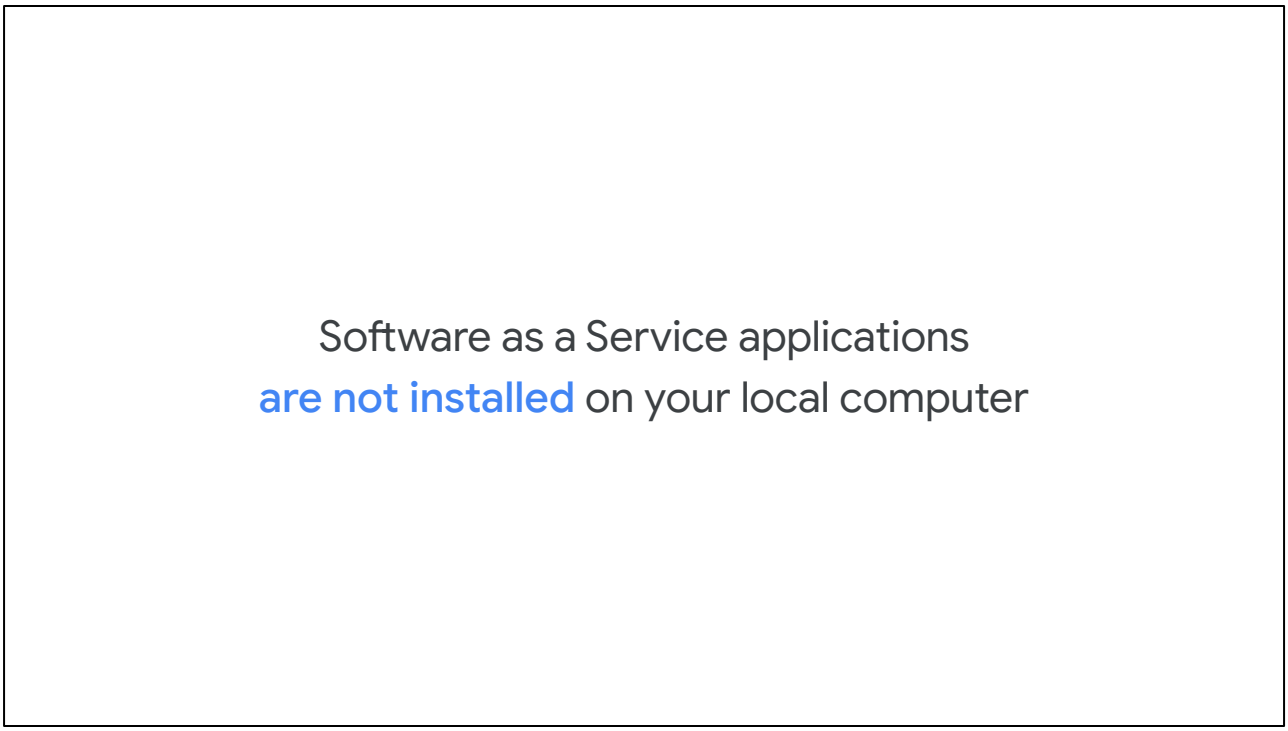
include Cloud Functions, which manages event-driven code as a pay-as-you-go service, and



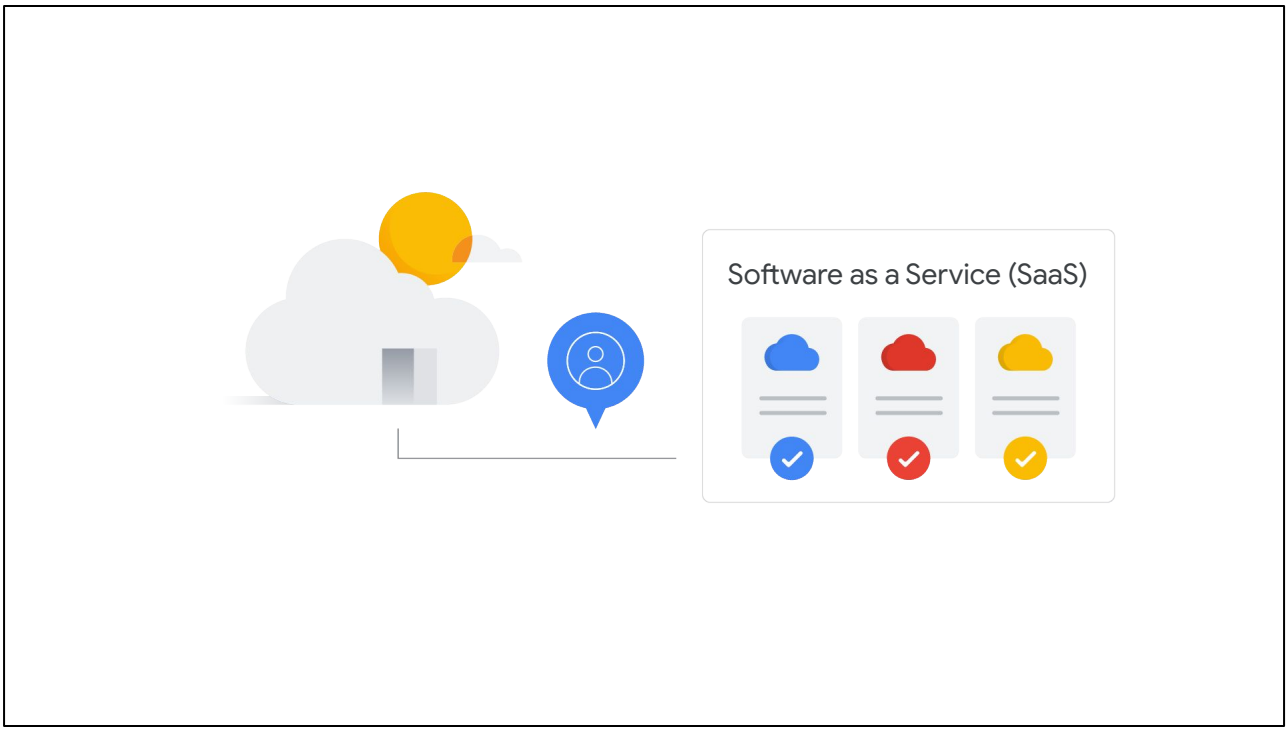
Cloud Run, which allows customers to deploy their containerized microservices based application in a fully-managed environment.



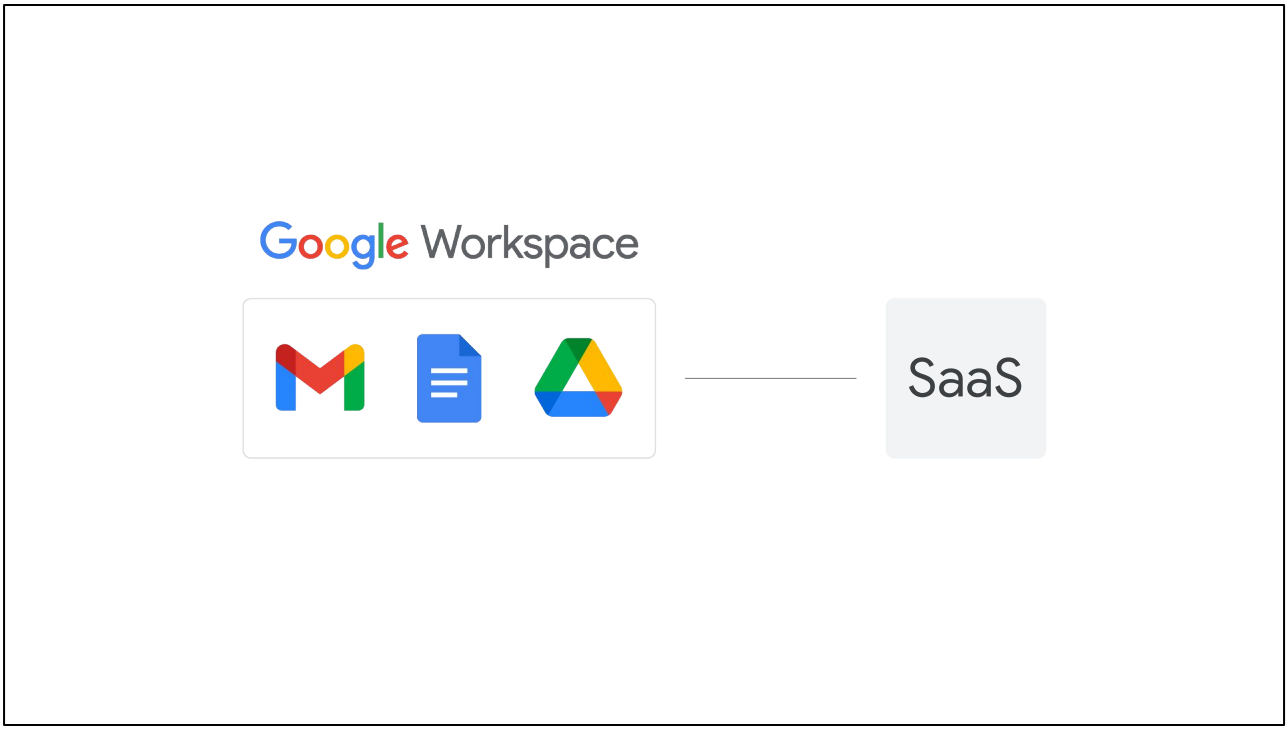
While it's outside the scope of *this* course, you might have heard about software as a service, SaaS, and wondered what it is and how it fits into the Cloud ecosphere.



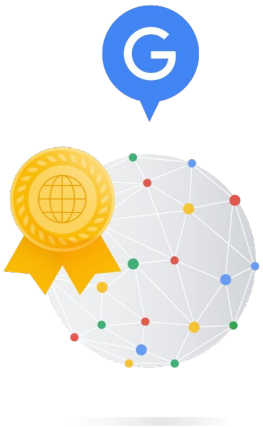
Software as a Service applications are not installed on your local computer.



Instead, they run in the cloud as a service and are consumed directly over the internet by end users.



Popular Google applications such as Gmail, Docs, and Drive, that are a part of Google Workspace, are all examples of SaaS.



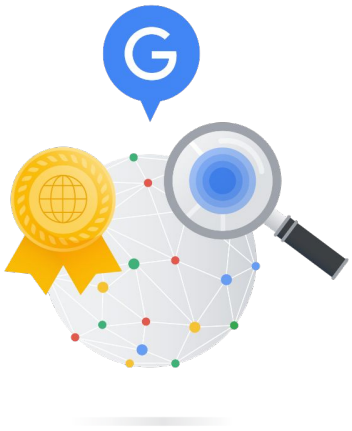
Billions

of dollars invested
over many years

Google Cloud runs on Google's own global network.

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

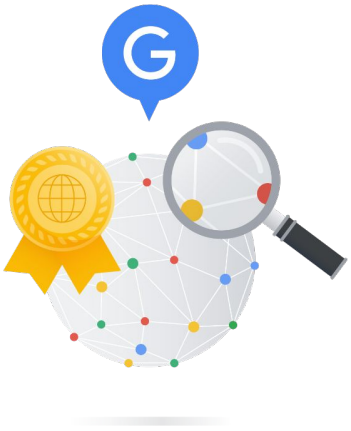
- Highest possible throughput
- Lowest possible latencies
- 100+ content caching nodes worldwide
- High demand content is cached for quicker access



This 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.

These are locations where high demand content is cached for quicker access, allowing applications to respond to user requests from the location that will provide the quickest response time.

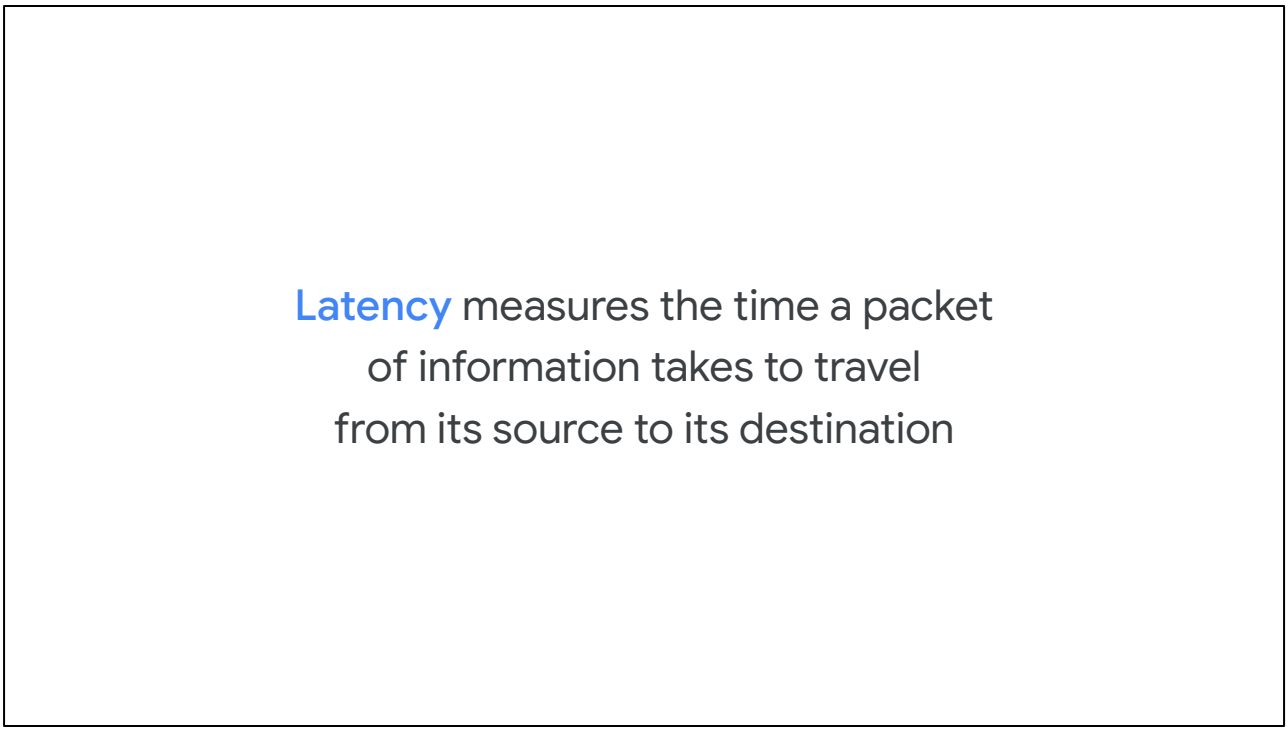
Google Cloud’s infrastructure is based in five major geographic locations: North America, South America, Europe, Asia, and Australia.



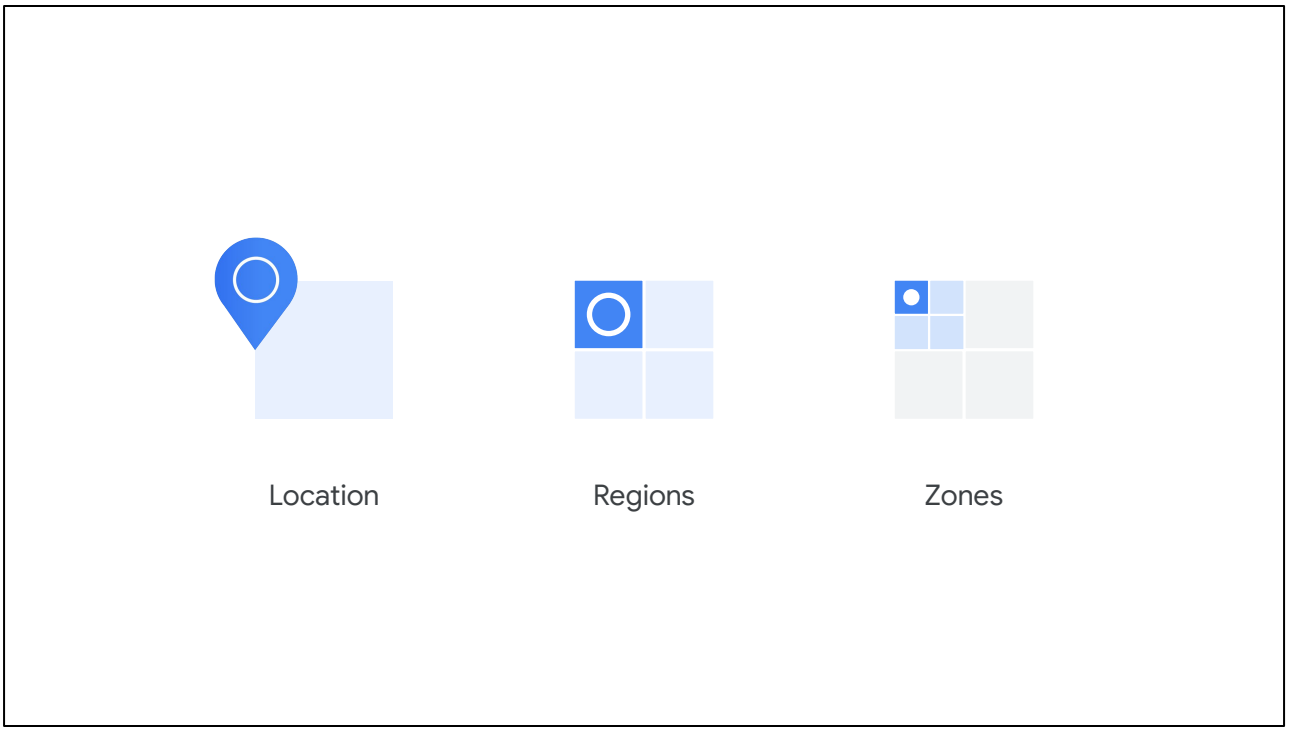
App location affects

- Availability
- Durability
- Latency

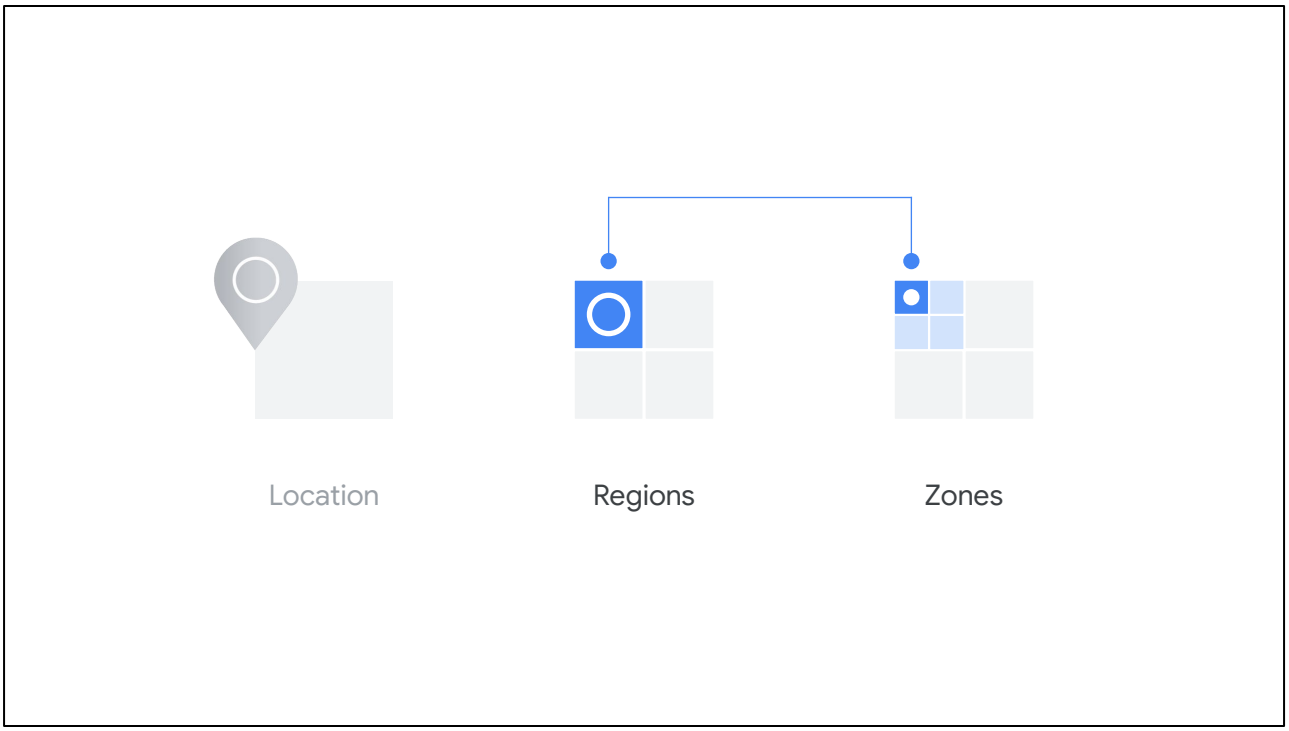
Having multiple service locations is important because choosing where to locate applications affects qualities like availability, durability, and latency, the latter of



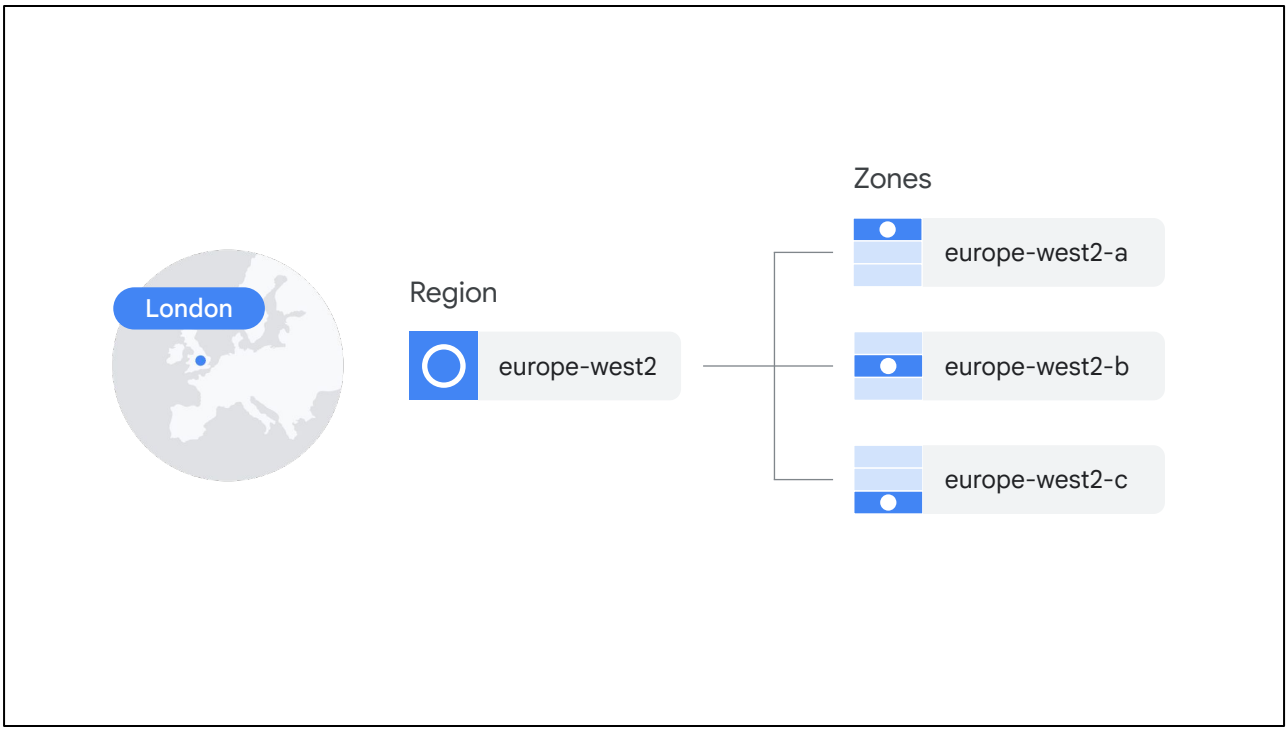
which measures the time a packet of information takes to travel from its source to its destination.



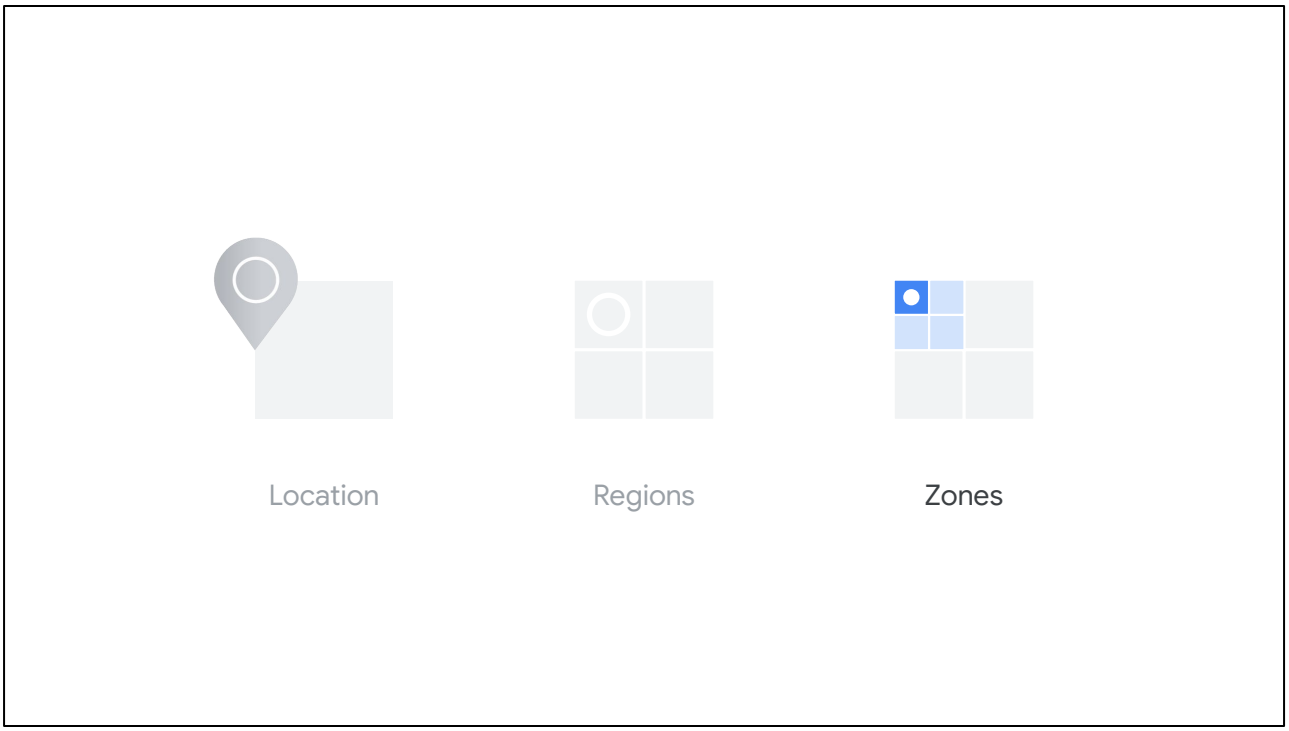
Each of these locations is divided into several different **regions** and **zones**.



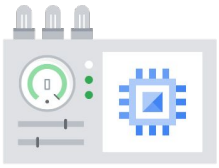
Regions represent independent geographic areas and are composed of zones.



For example, London, or europe-west2, is a region that currently comprises three different zones.

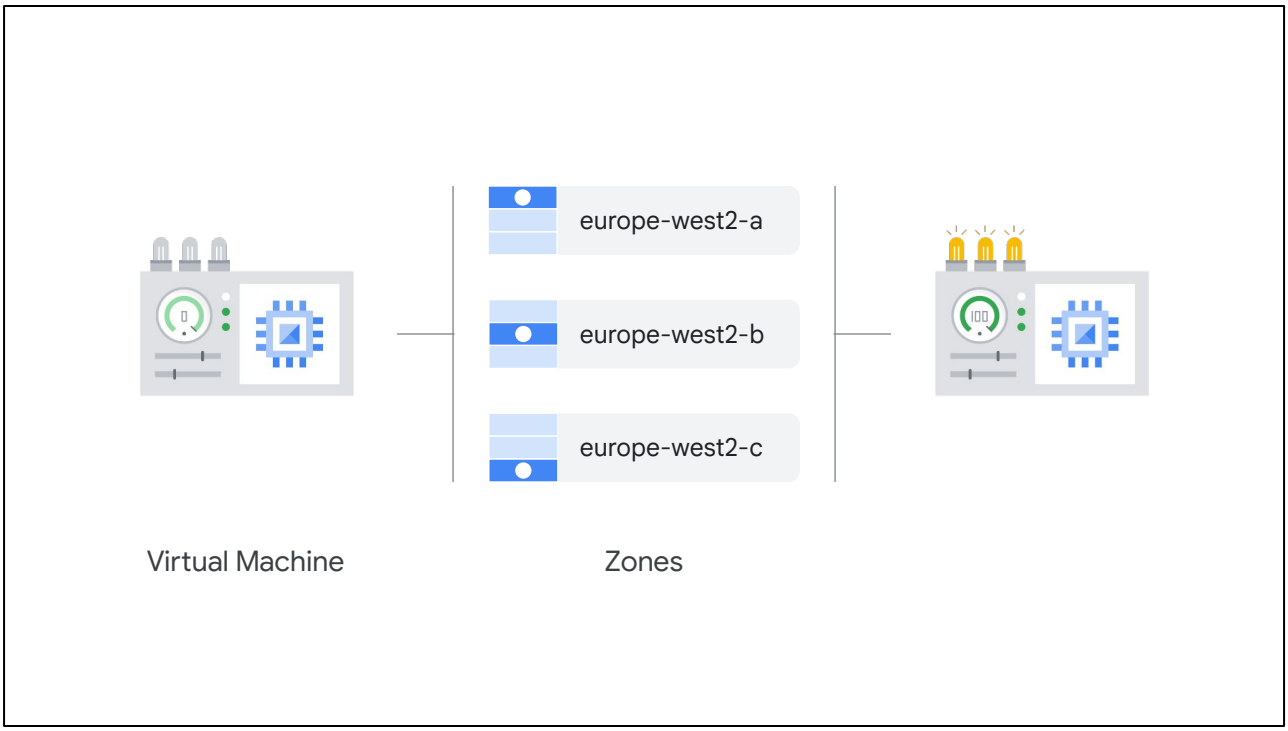


A zone is an area where Google Cloud resources are deployed.

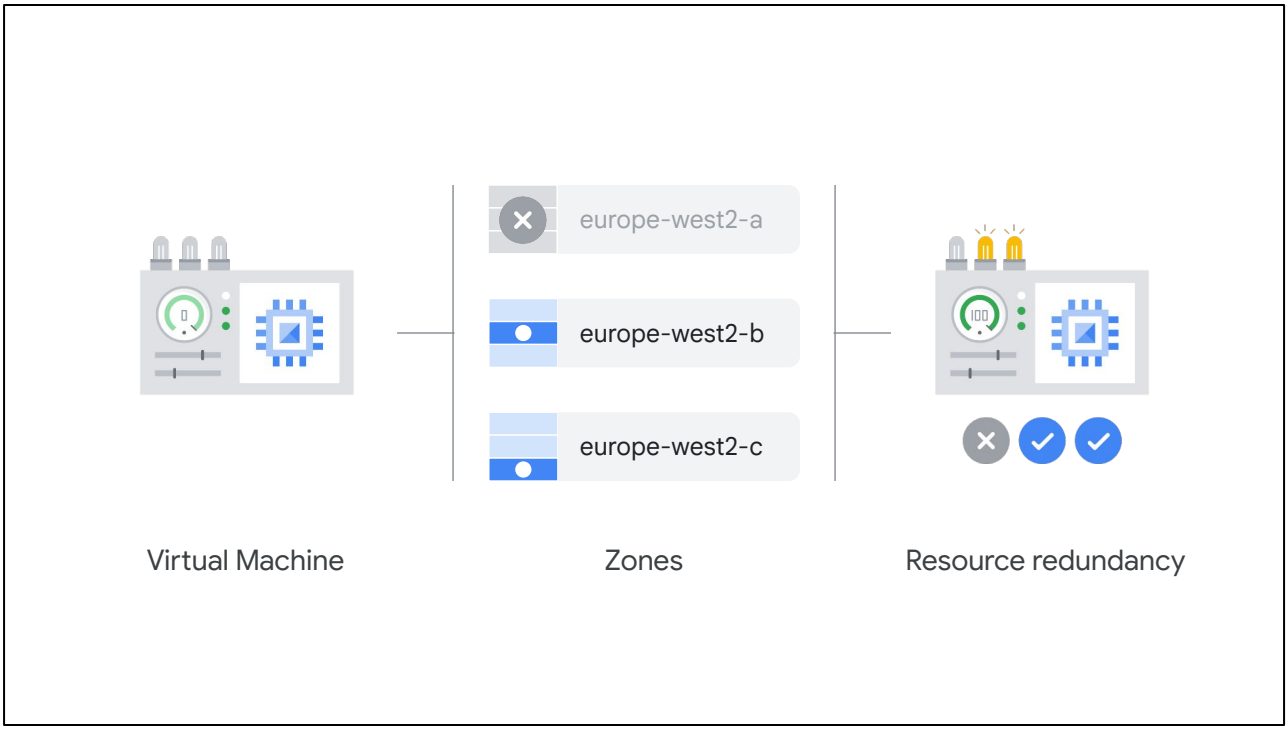


Virtual Machine

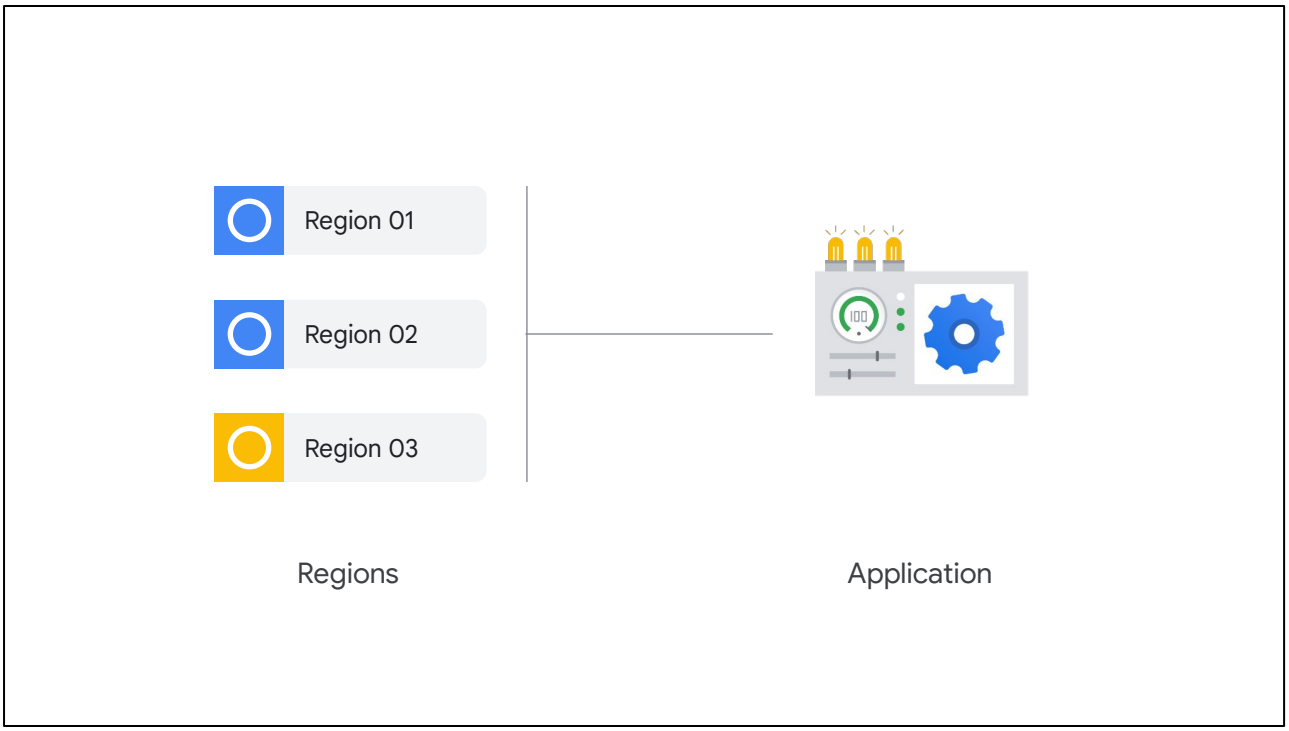
For example, if you launch a virtual machine using Compute Engine



it will run in the zone that you specify

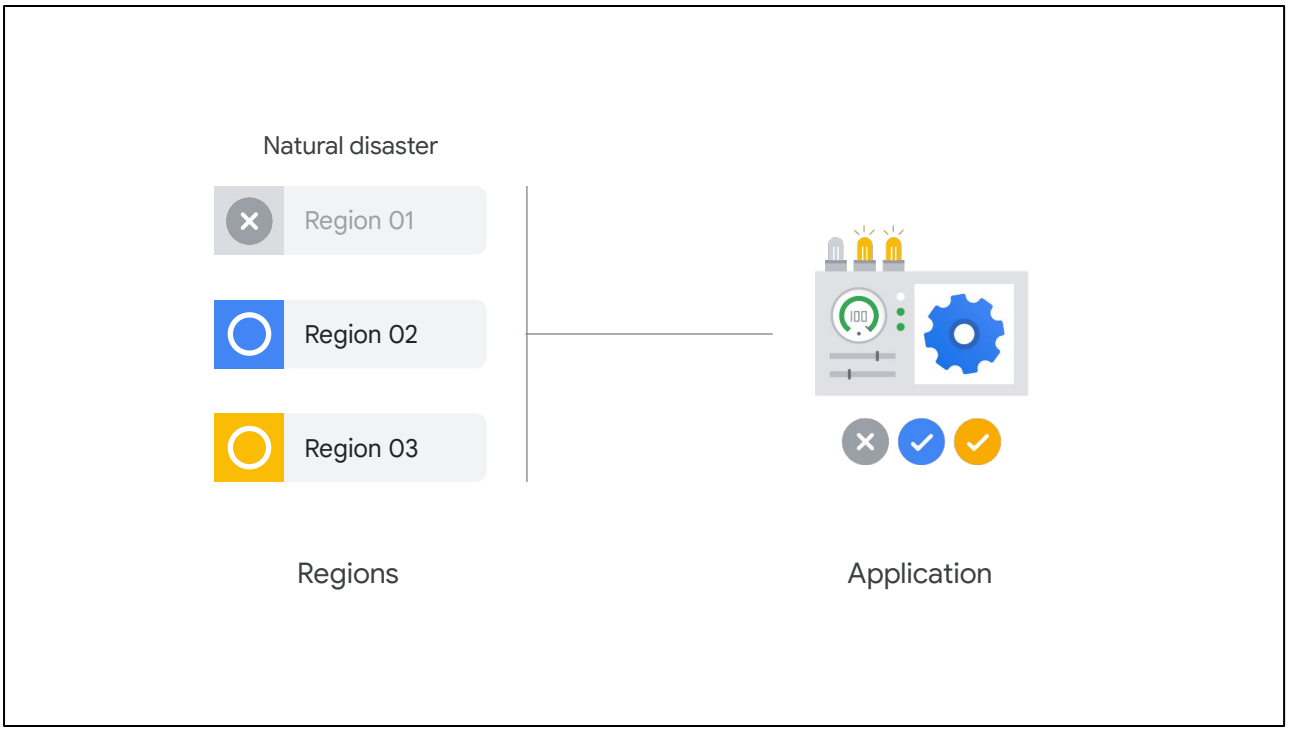


to ensure resource redundancy.

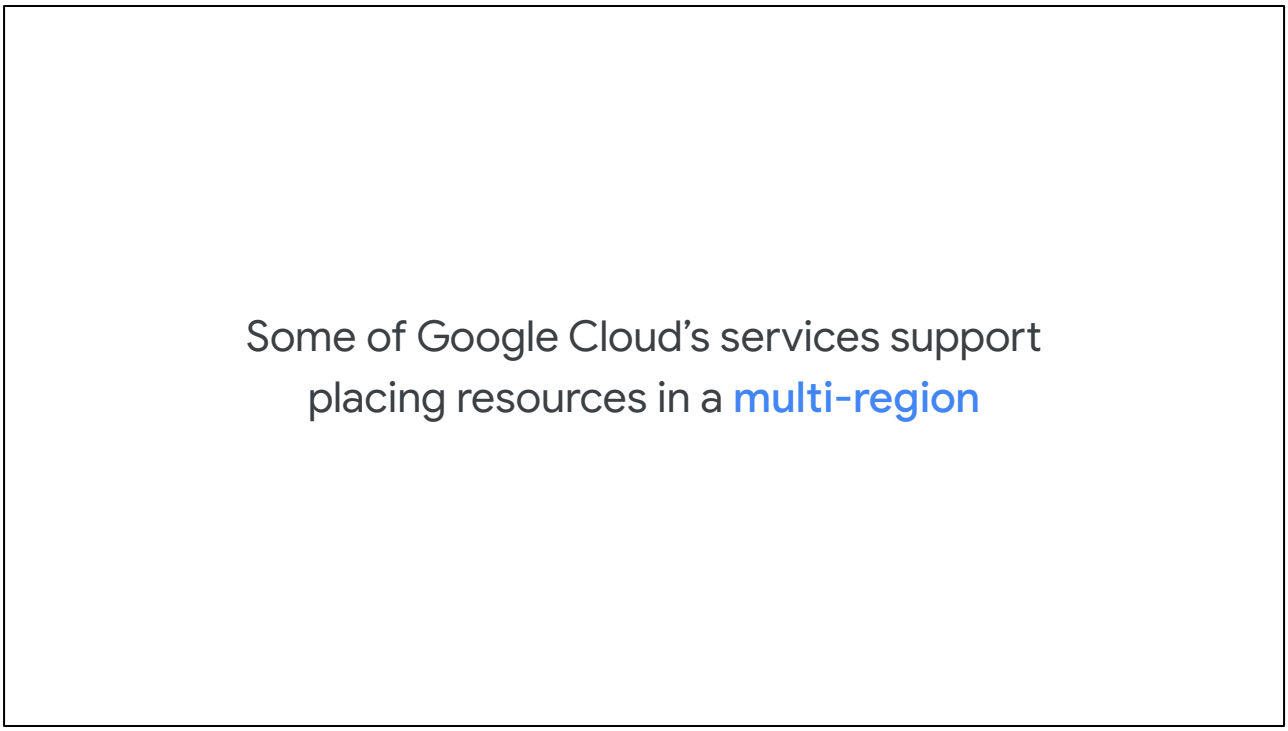


You can run resources in different regions.

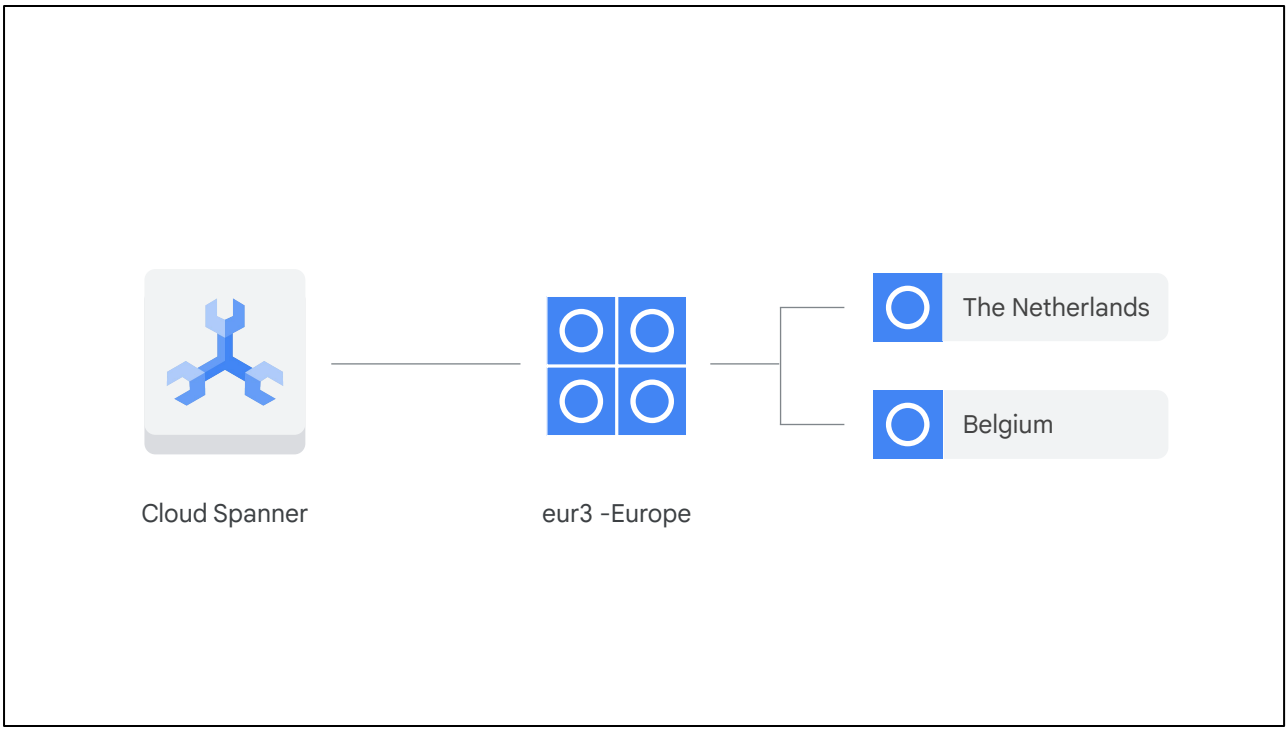
This is useful for bringing applications closer to users around the world, and also for protection in case there are issues with an entire region,



say, due to a natural disaster.

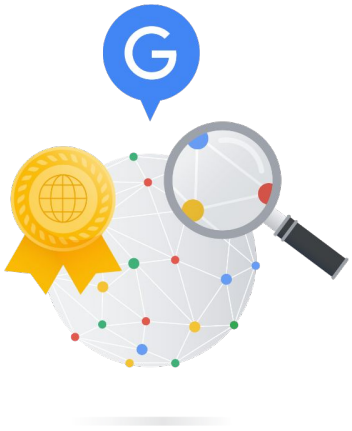


Some of Google Cloud's services support placing resources in what we call a **multi-region**.



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.

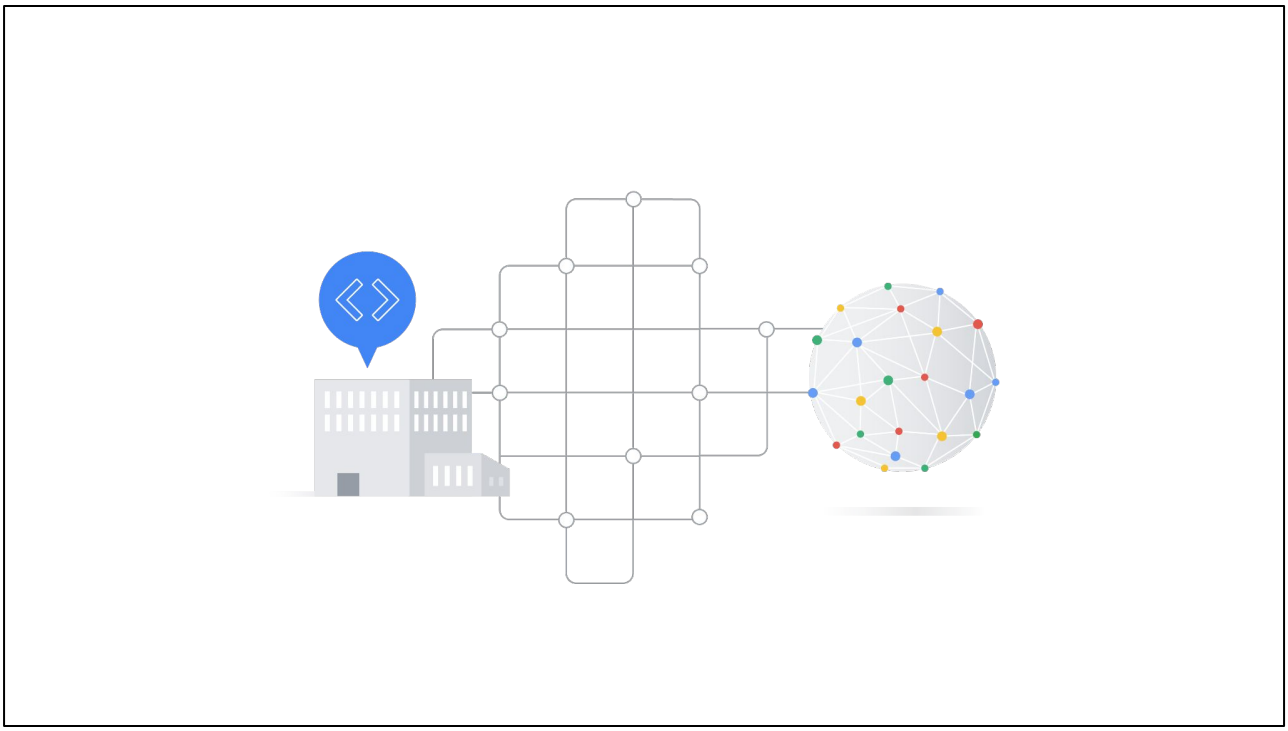


103 Zones

34 Regions

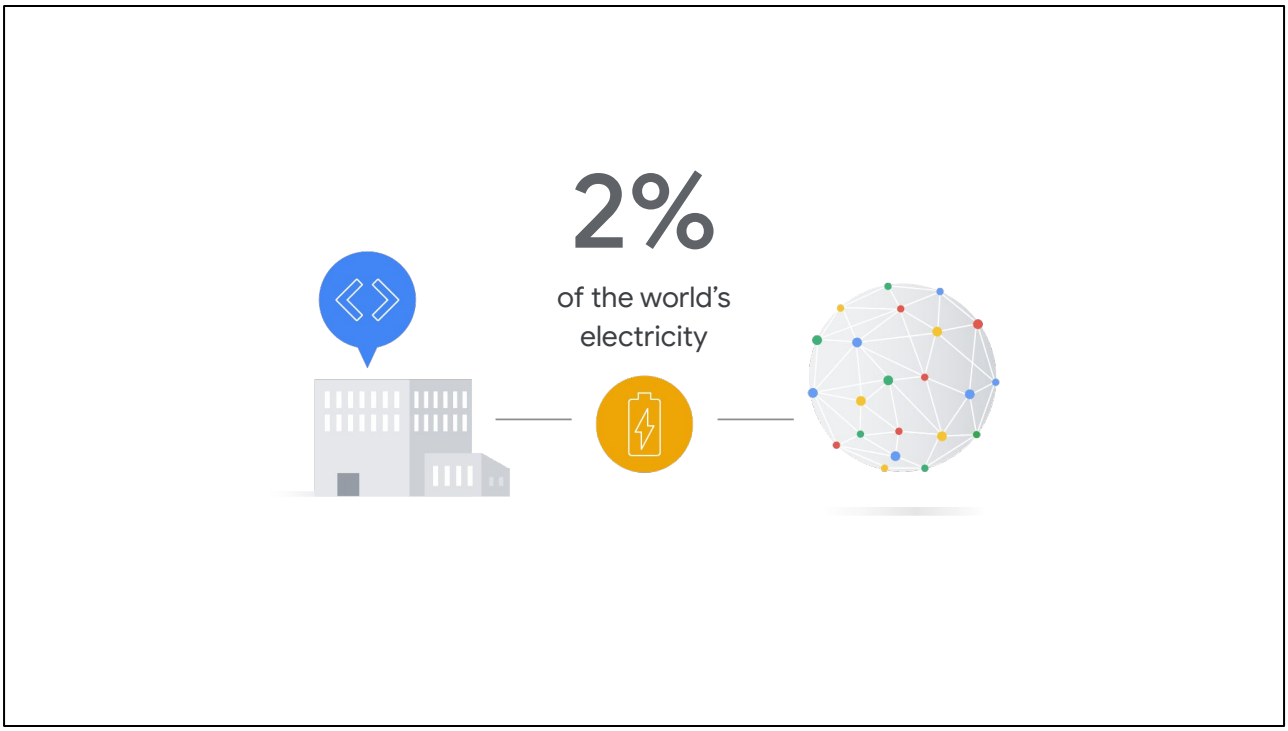
Google Cloud currently supports 103 zones in 34 regions, although this number is increasing all the time.

You can find the most up-to-date numbers at cloud.google.com/about/locations.

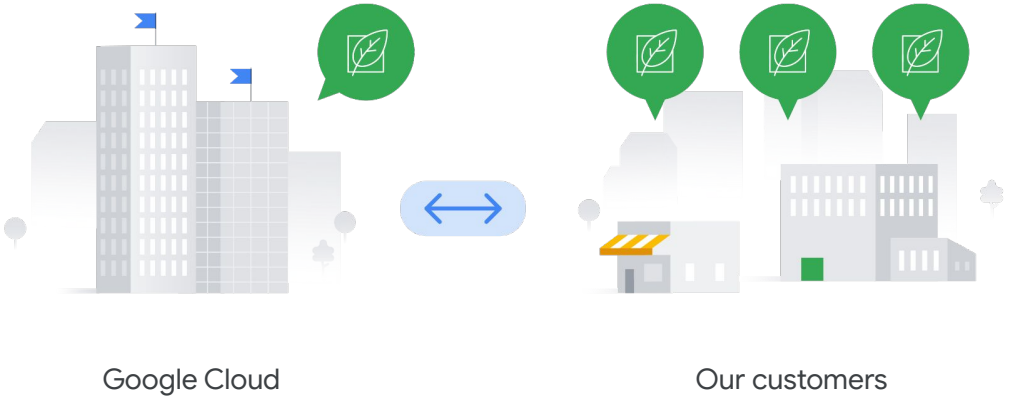


The virtual world, which includes Google Cloud’s network, is built on physical infrastructure, and all those racks of humming servers

use huge amounts of energy.



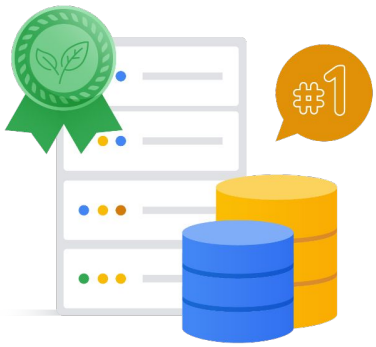
Altogether, existing data centers use roughly 2% of the world's electricity. With this in mind, Google works to make their data centers run as efficiently as possible.



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 on Google Cloud can be a part of meeting those goals.

Therefore, it's useful to note that

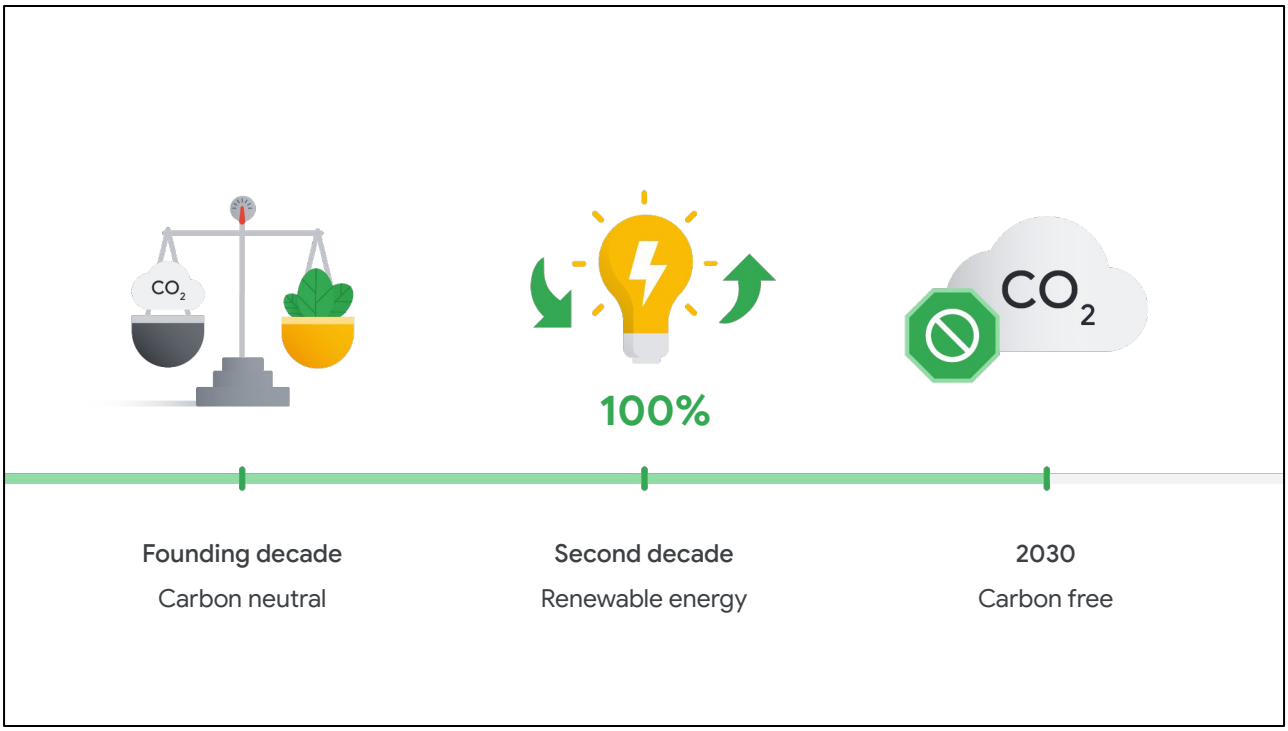
Google's data centers were the first to achieve **ISO 14001 certification**



Google's data centers were the first to achieve ISO 14001 certification, which is a standard that maps out a framework for an organization to enhance its environmental performance through improving resource efficiency and reducing waste.

As an example of how this is being done, here's Google's data center in Hamina, Finland.

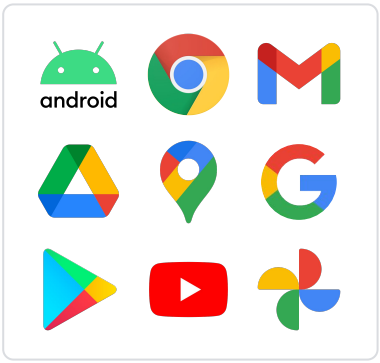
This facility is one of the most advanced and efficient data centers in the Google fleet. Its cooling system, which uses sea water from the Bay of Finland, reduces energy use and is the first of its kind anywhere in the world.



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.

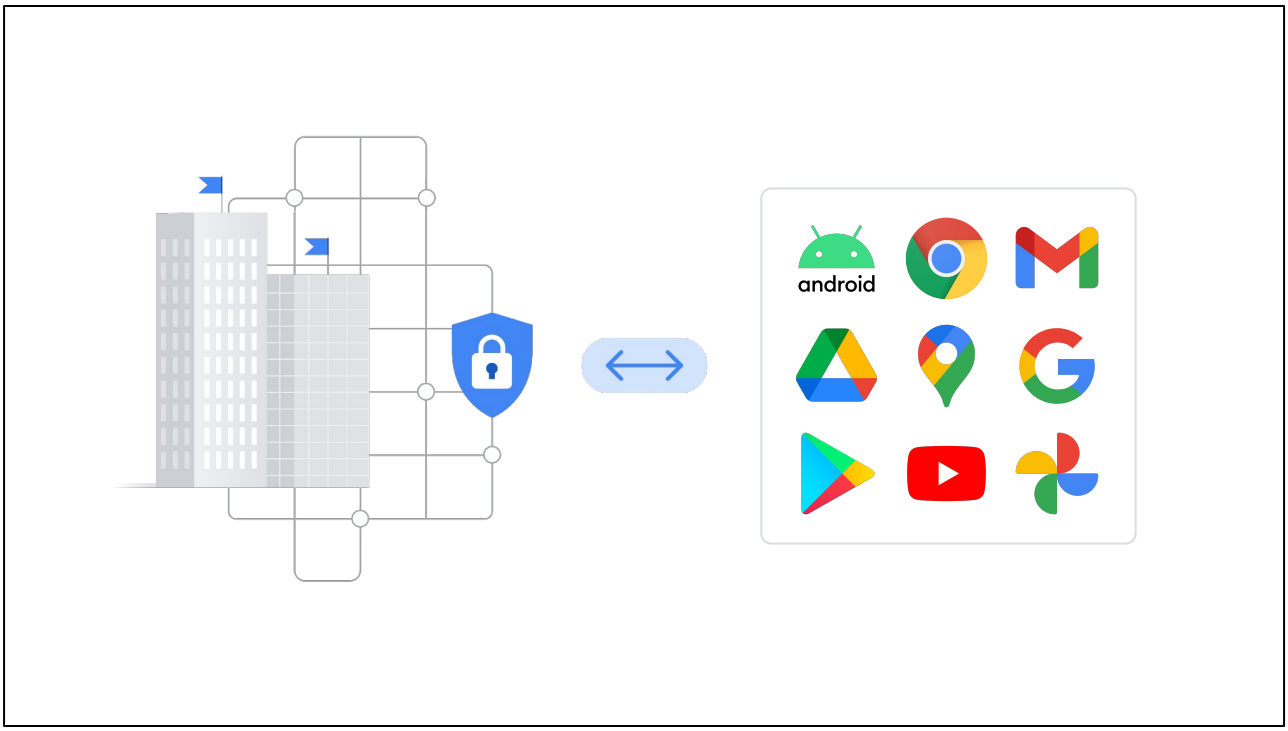
By 2030, we aim to be the first major company to operate completely carbon free.



1 Billion+

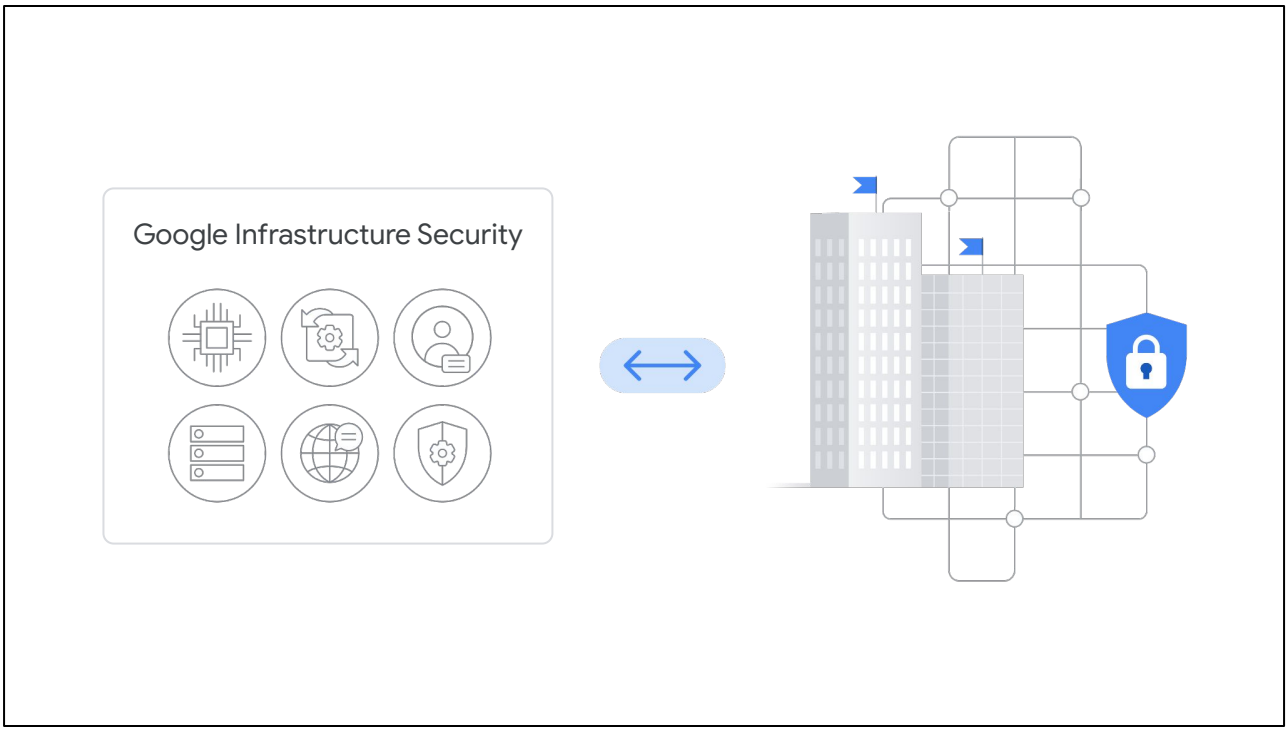
Users

Because Google has nine services with more than a billion users, you can be assured that security is always on the minds of Google's employees.

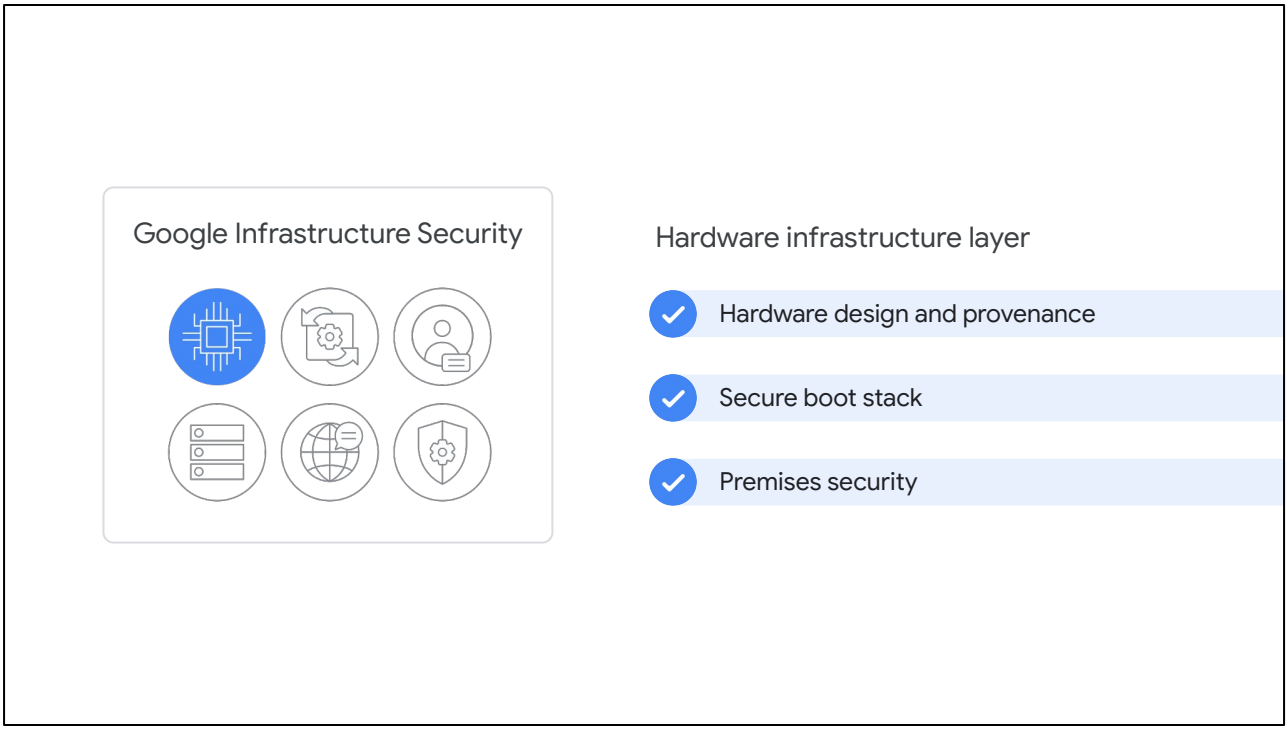


Design for security is prevalent throughout the infrastructure that Google Cloud and Google services run on.

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



The security infrastructure can be explained in progressive layers, starting from the physical security of our data centers, continuing on to how the hardware and software that underlie the infrastructure are secured, and finally, describing the technical constraints and processes in place to support operational security.

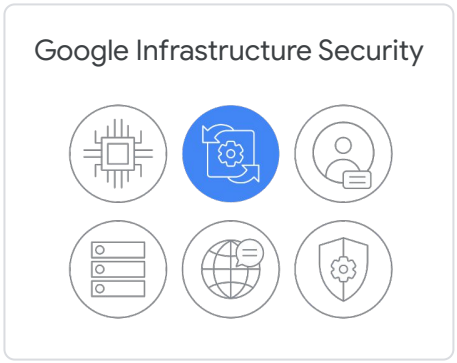


We begin with the **Hardware infrastructure** layer which comprises three key security features:

- The first is **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.

- The next feature is a **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.

- This layer's final feature is **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 number 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.



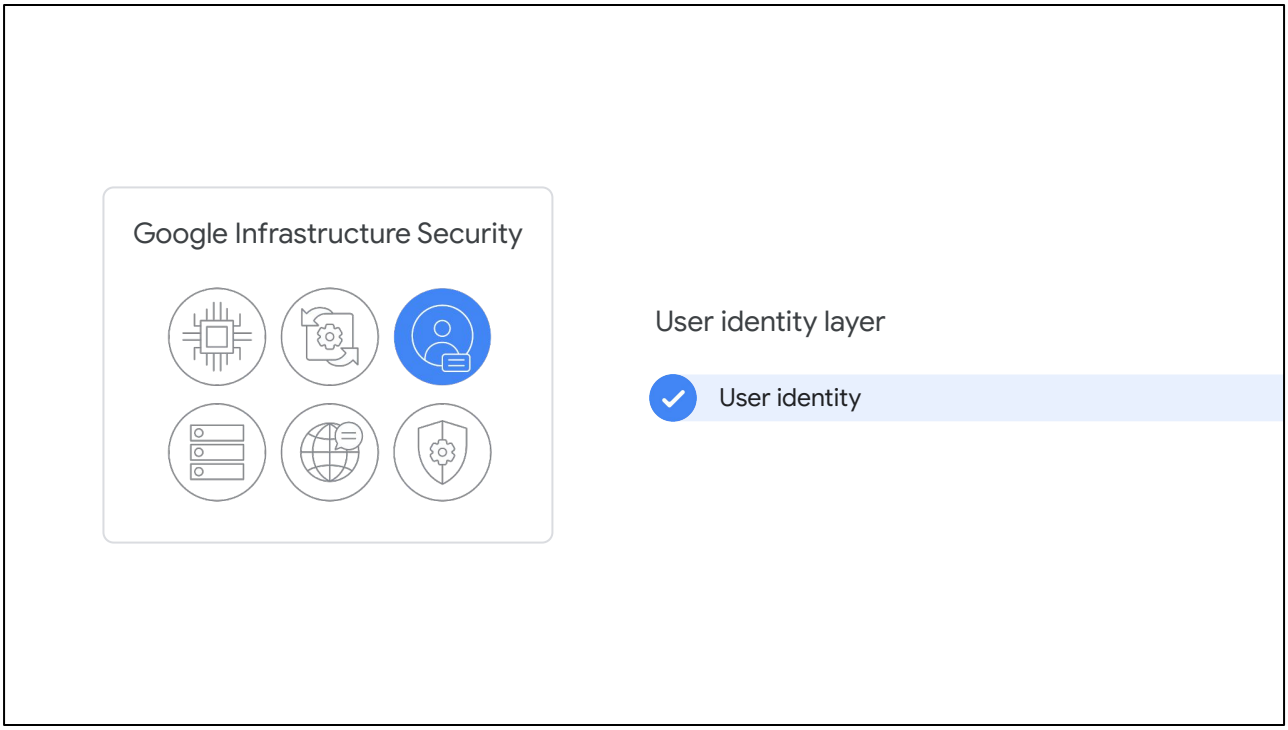
Service deployment layer

Encryption of inter-service communication

Next is the **Service deployment** layer, where the key feature is 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 that goes between data centers.

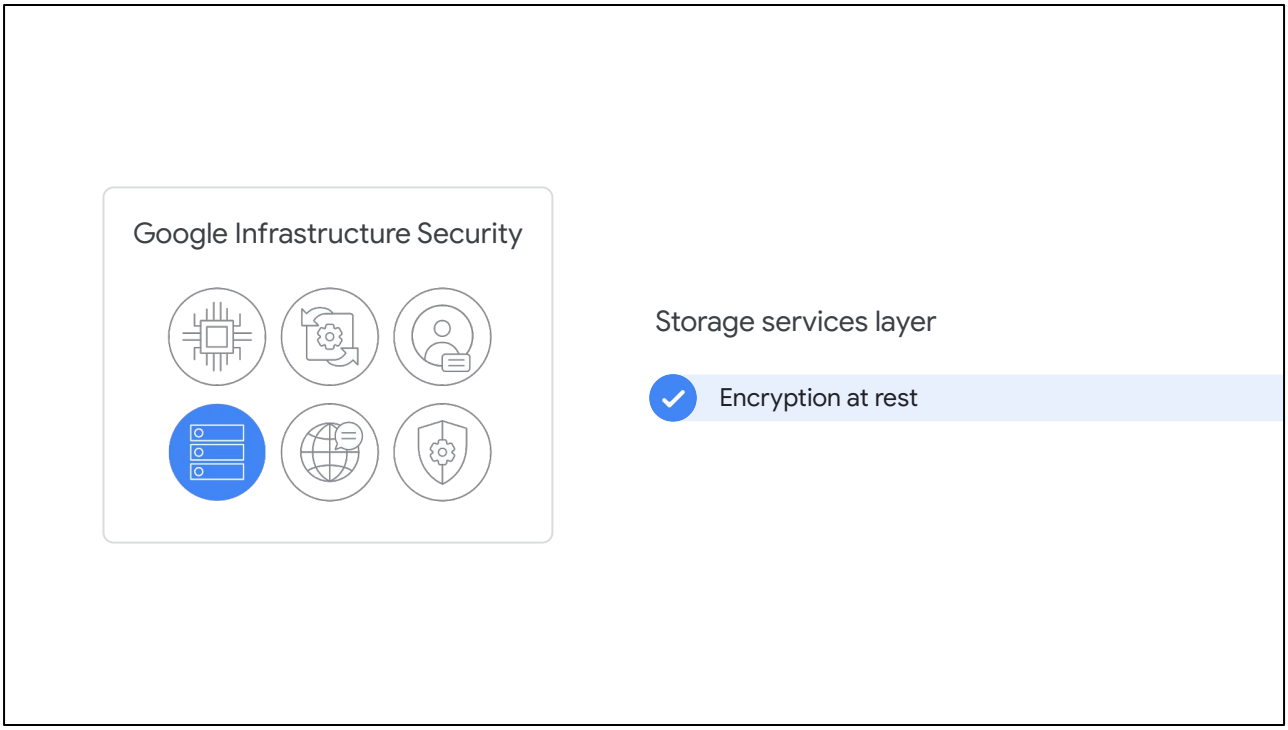
Google has started to deploy hardware cryptographic accelerators that will allow it to extend this default encryption to all infrastructure RPC traffic inside Google data centers.



Then we have the **User identity** layer.

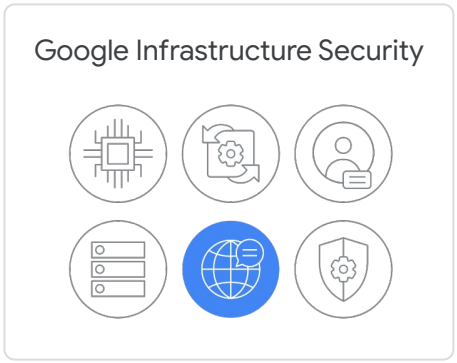
Google's central identity service, which usually manifests to end users as the Google login page, goes beyond asking for a simple username and password. The service also intelligently challenges users for additional information based on risk factors such as whether they have logged in from the same device or a similar location in the past.

Users can also employ secondary factors when signing in, including devices based on the Universal 2nd Factor (U2F) open standard.



On the **Storage services** layer we find the **encryption at rest** security feature.

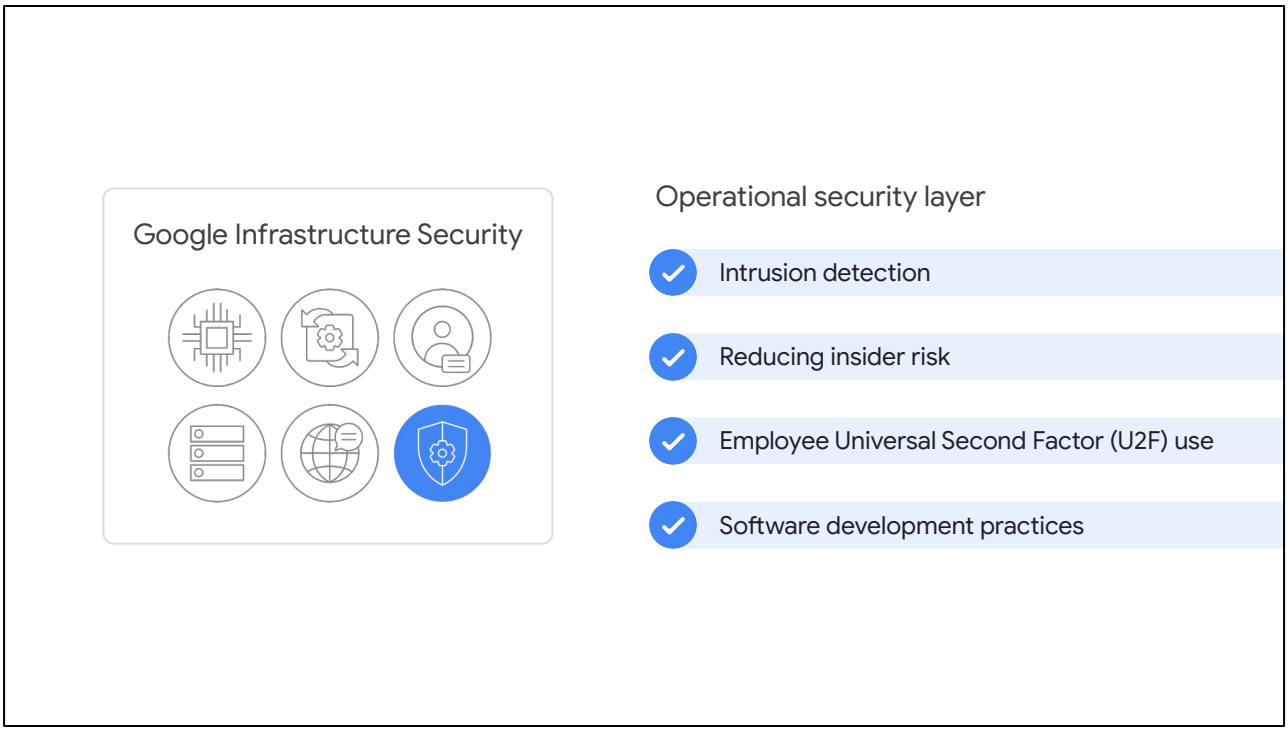
Most applications at Google access physical storage (in other words, “file storage”) indirectly via storage services, and encryption using centrally managed keys is applied at the layer of these storage services. Google also enables hardware encryption support in hard drives and SSDs.



The next layer up is the **Internet communication** layer, and this comprises two key security features.

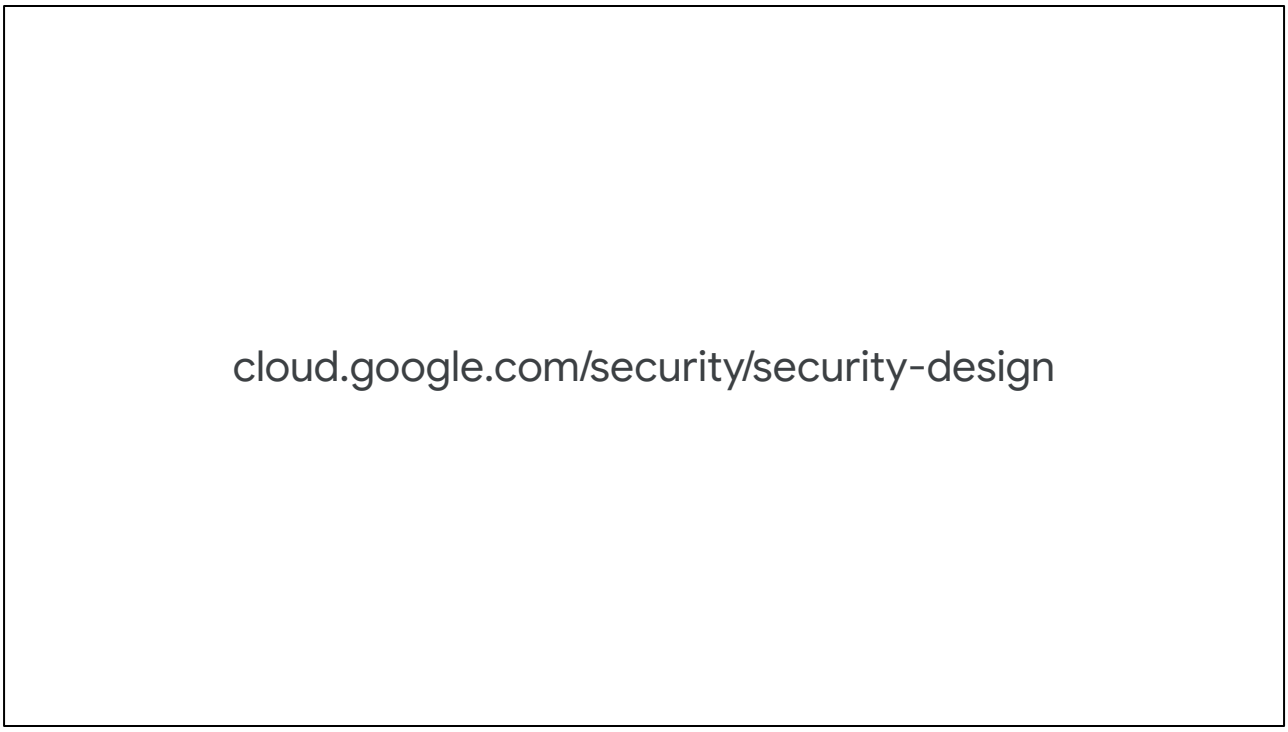
- Google services that are being made available on the internet, register themselves with an infrastructure service called the **Google Front End**, which ensures that all TLS connections are ended using a public-private key pair and an X.509 certificate from a Certified Authority (CA), as well as following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

- Also provided is **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.

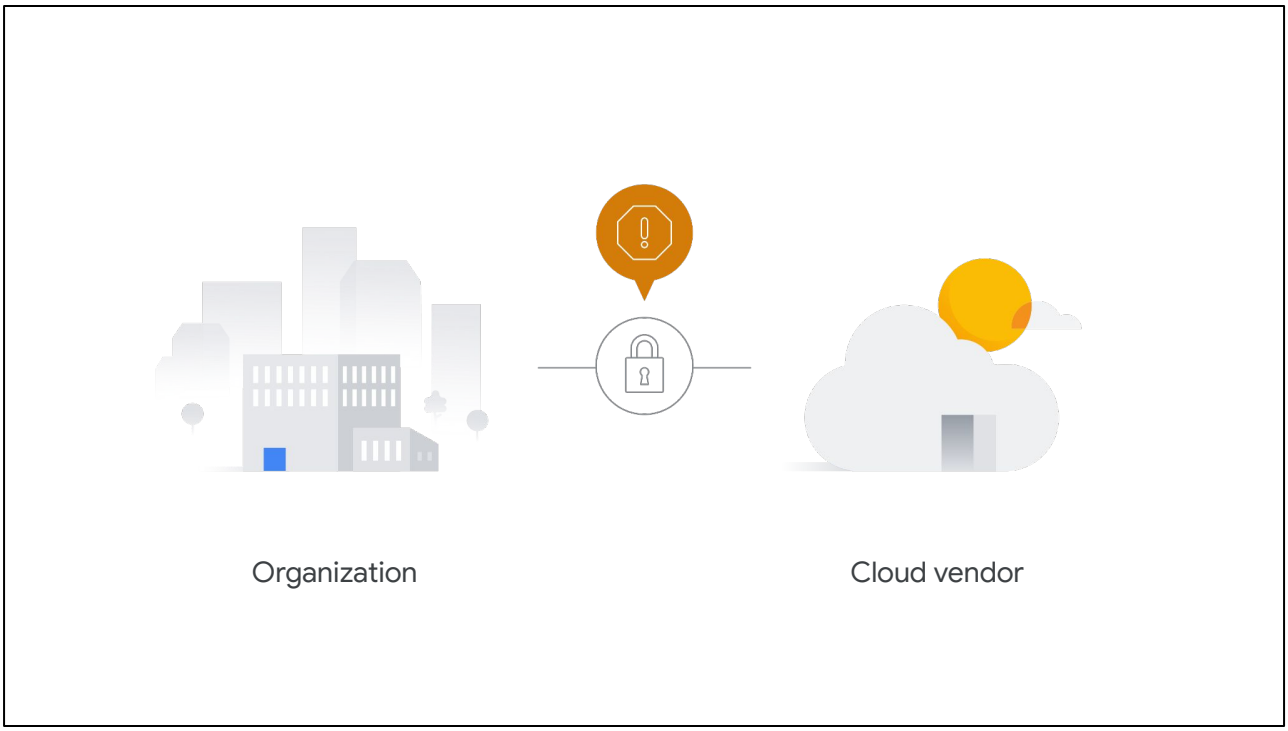


The final layer is Google's **Operational security** layer which provides four key features.

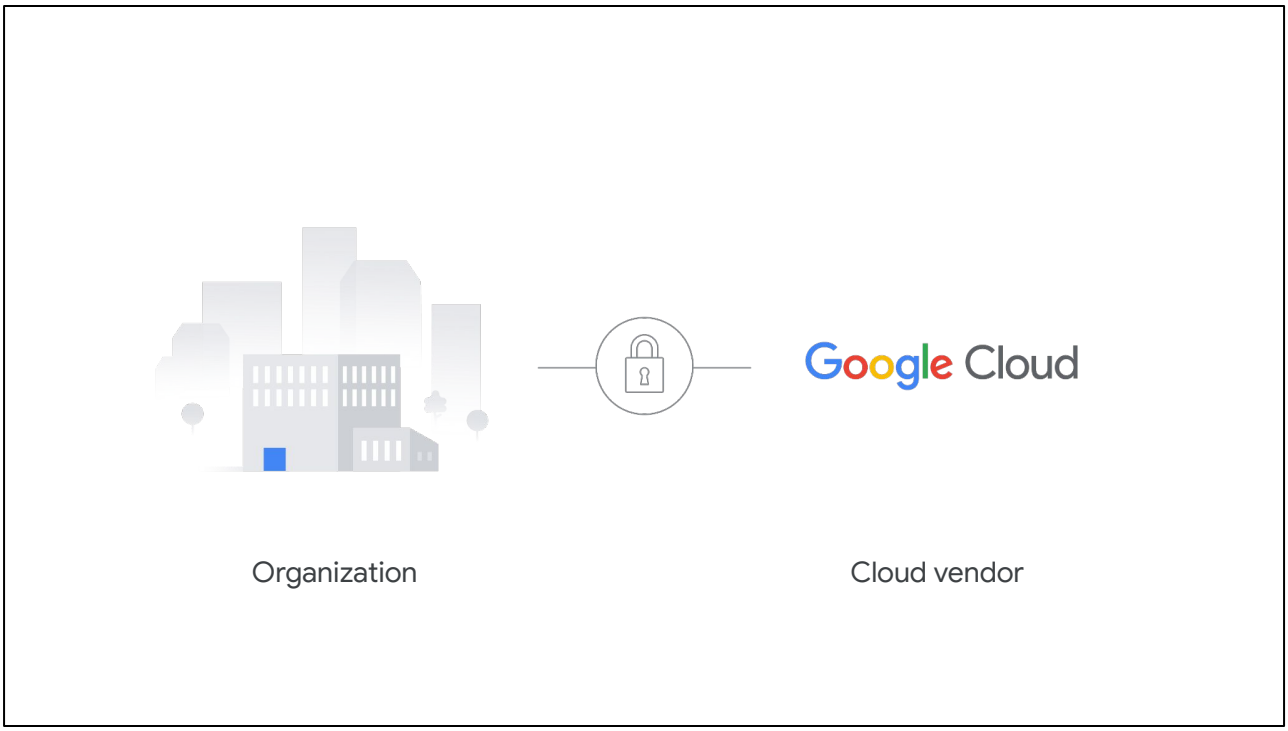
- First is **intrusion detection**. Rules and machine intelligence give Google's operational security teams warnings of possible incidents. Google conducts Red Team exercises to measure and improve the effectiveness of its detection and response mechanisms.
- Next is **reducing insider risk**. Google aggressively limits and actively monitors the activities of employees who have been granted administrative access to the infrastructure.
- Then there's **employee U2F use**. To guard against phishing attacks against Google employees, employee accounts require use of U2F-compatible Security Keys.
- Finally, there are stringent **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. Additionally, Google runs a Vulnerability Rewards Program where we pay anyone who is able to discover and inform us of bugs in our infrastructure or applications.



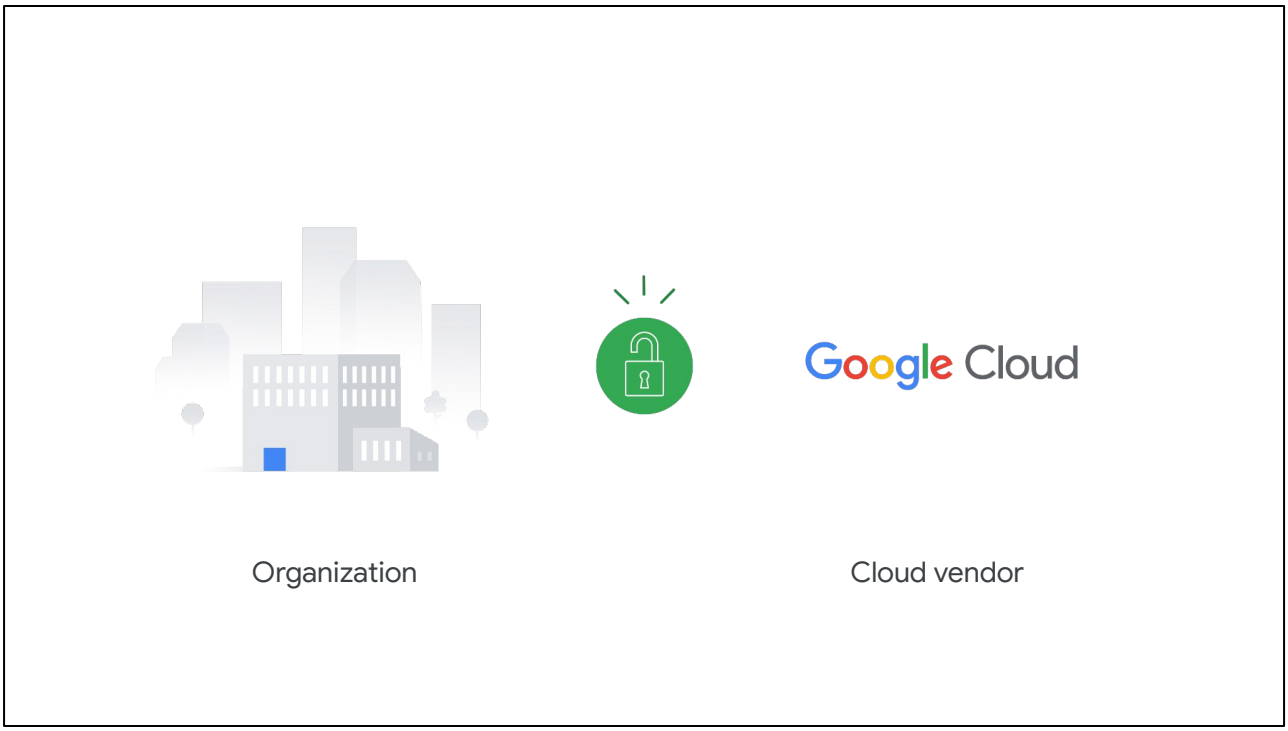
You can learn more about Google's technical-infrastructure security at cloud.google.com/security/security-design.



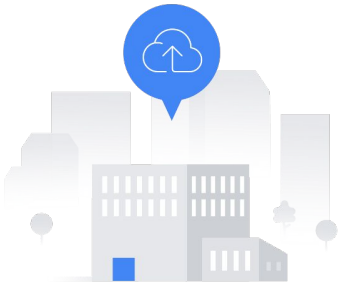
Some organizations are afraid to bring their workloads to the cloud because they're afraid they'll get locked into a particular vendor.



However, if, for whatever reason, a customer decides that Google



is no longer the best provider for their needs, we provide them with the ability

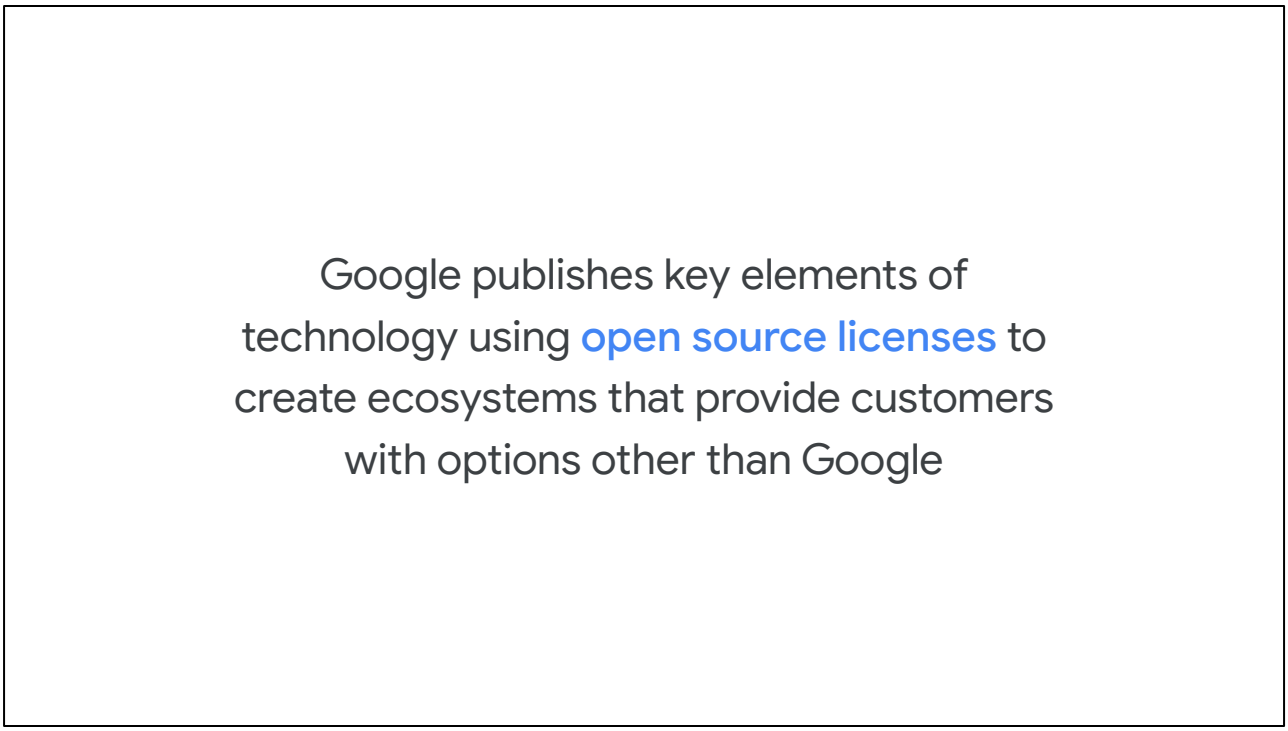


Google Cloud

Organization

Cloud vendor

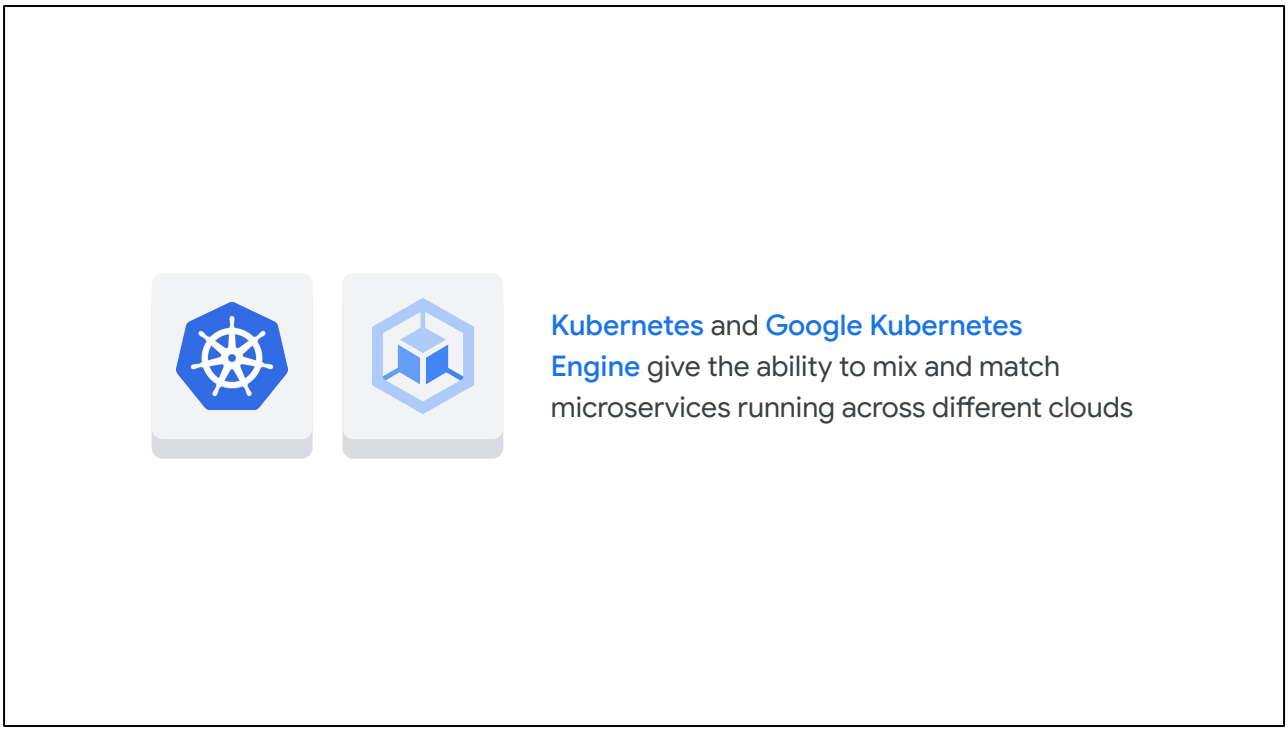
to run their applications elsewhere.



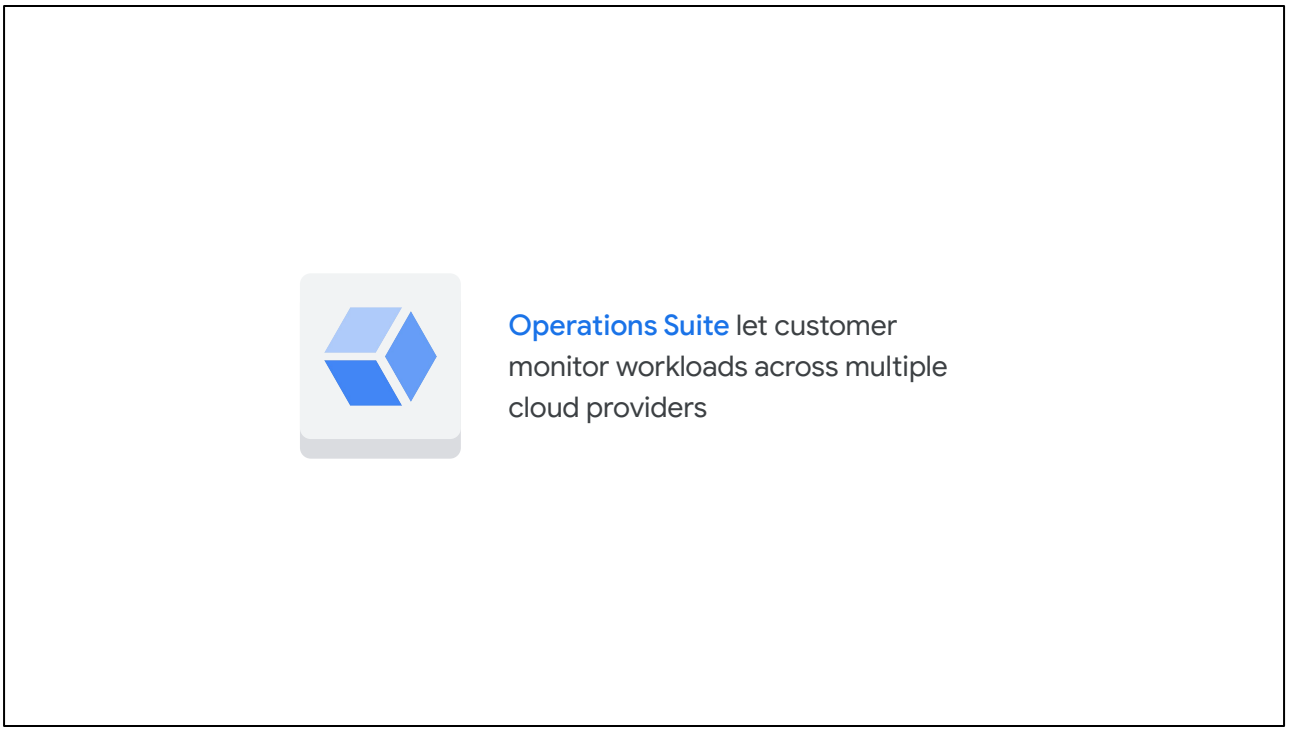
Google publishes key elements of technology using open source licenses to create ecosystems that provide customers with options other than Google.

TensorFlow is an open source library for machine learning at the heart of a strong open source ecosystem

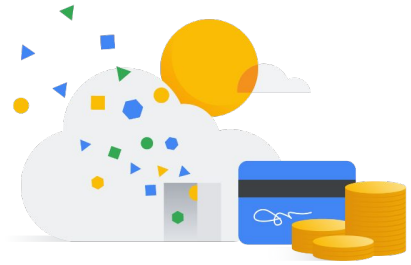
For example, TensorFlow, an open source software library for machine learning developed inside Google, is at the heart of a strong open source ecosystem.



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,

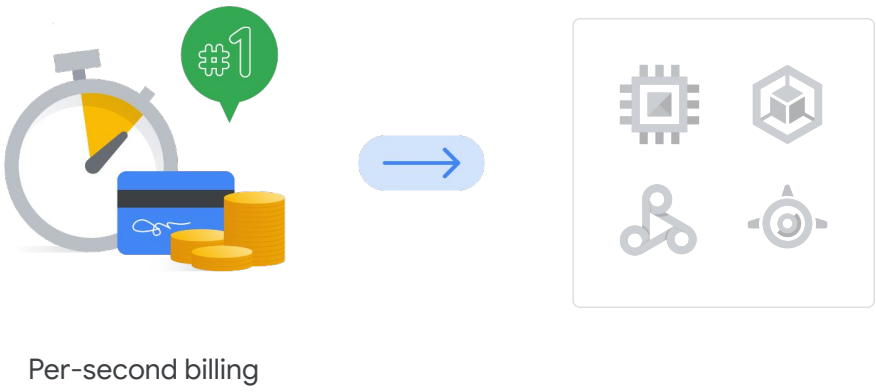


while Google Cloud’s operations suite lets customers monitor workloads across multiple cloud providers.

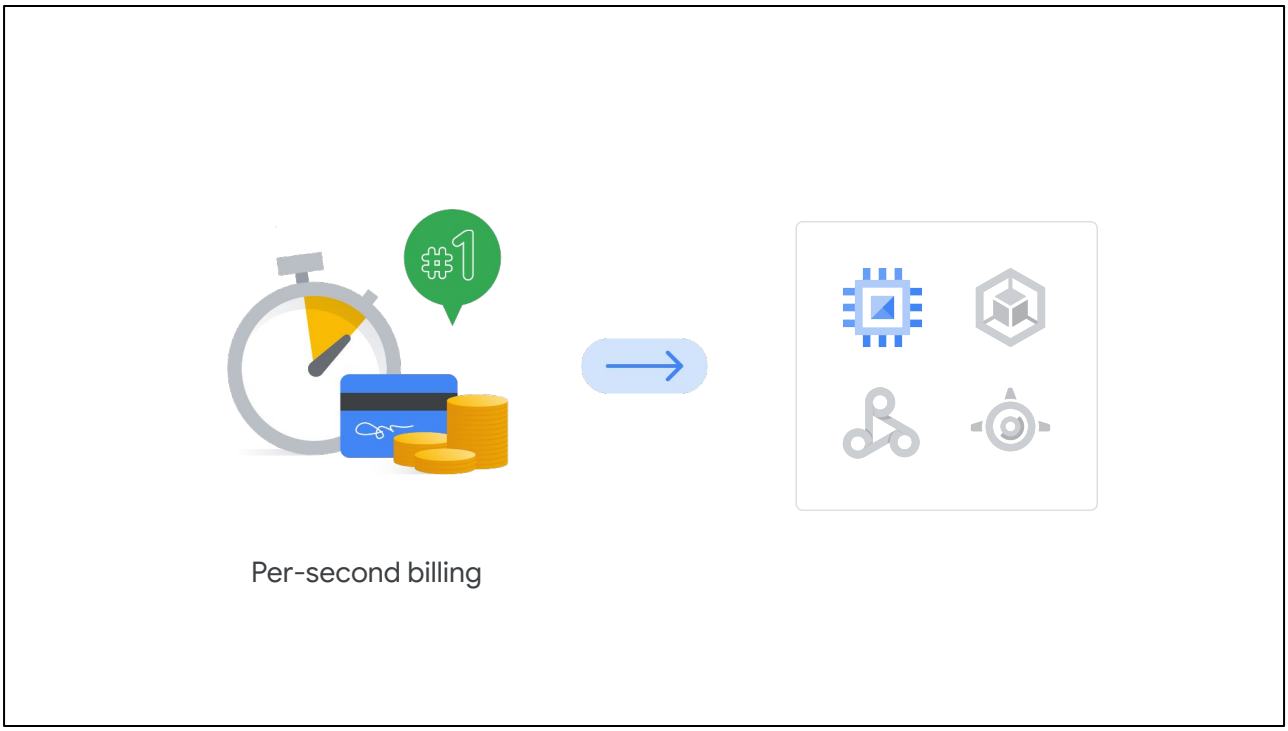


Google Cloud

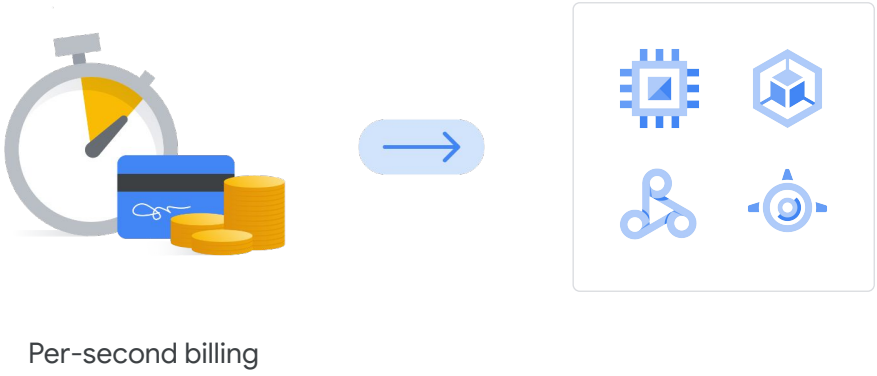
To round off this section of the course, let's take a brief look at Google Cloud's pricing structure.



Google was the first major cloud provider to deliver per-second billing for its infrastructure-as-a-service compute offering,

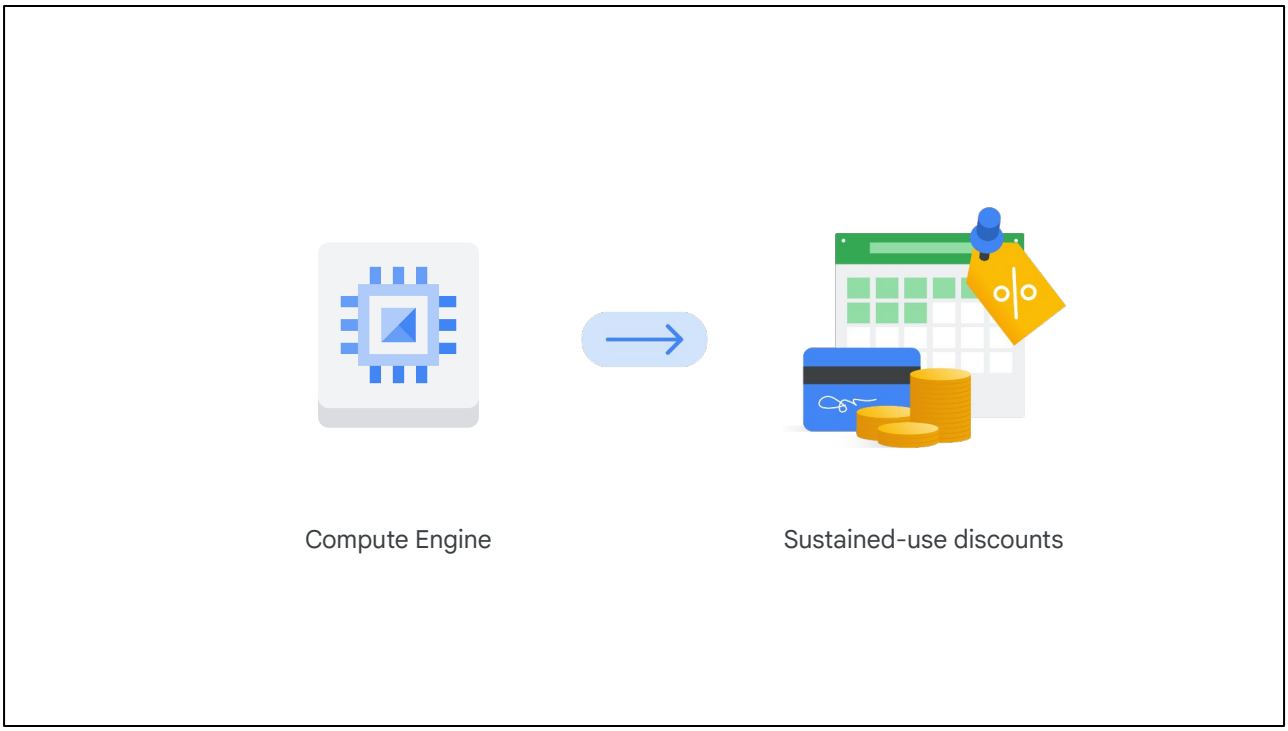


Compute Engine. In addition, per-second billing is now *also* offered for users of

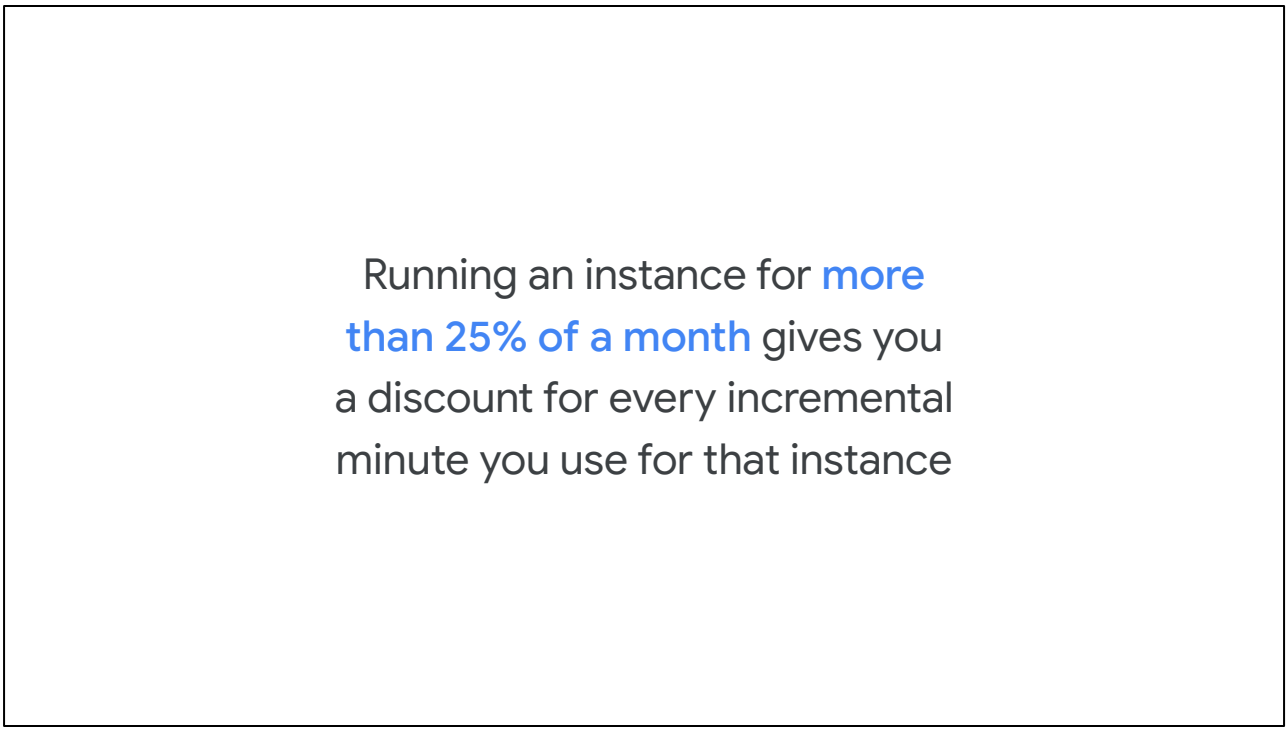


Google Kubernetes Engine (our container infrastructure as a service), Dataproc (which is the equivalent of the big data system Hadoop, but operating as a service), and App Engine flexible environment VMs (a platform as a service).

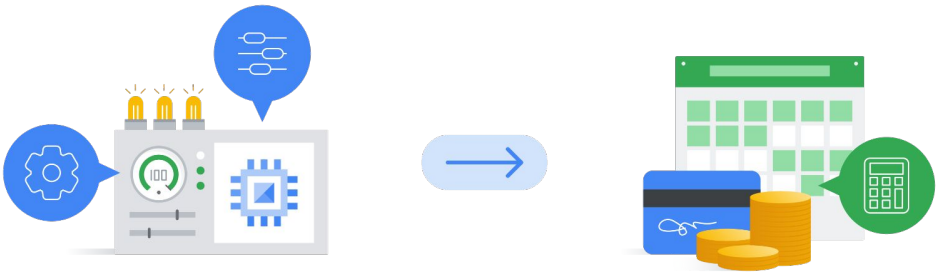
We'll explore these products and services later in the course.



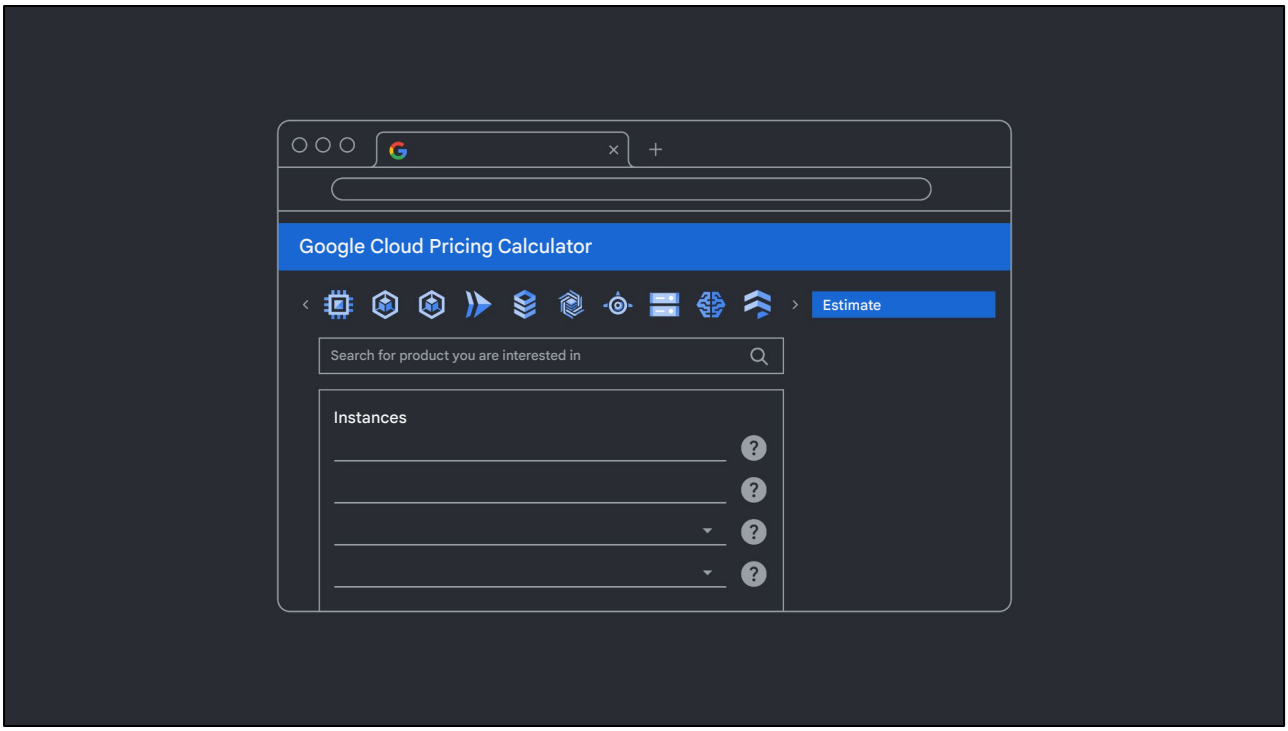
Compute Engine offers automatically applied sustained-use discounts, which are automatic discounts that you get for running a virtual machine instance for a significant portion of the billing month.



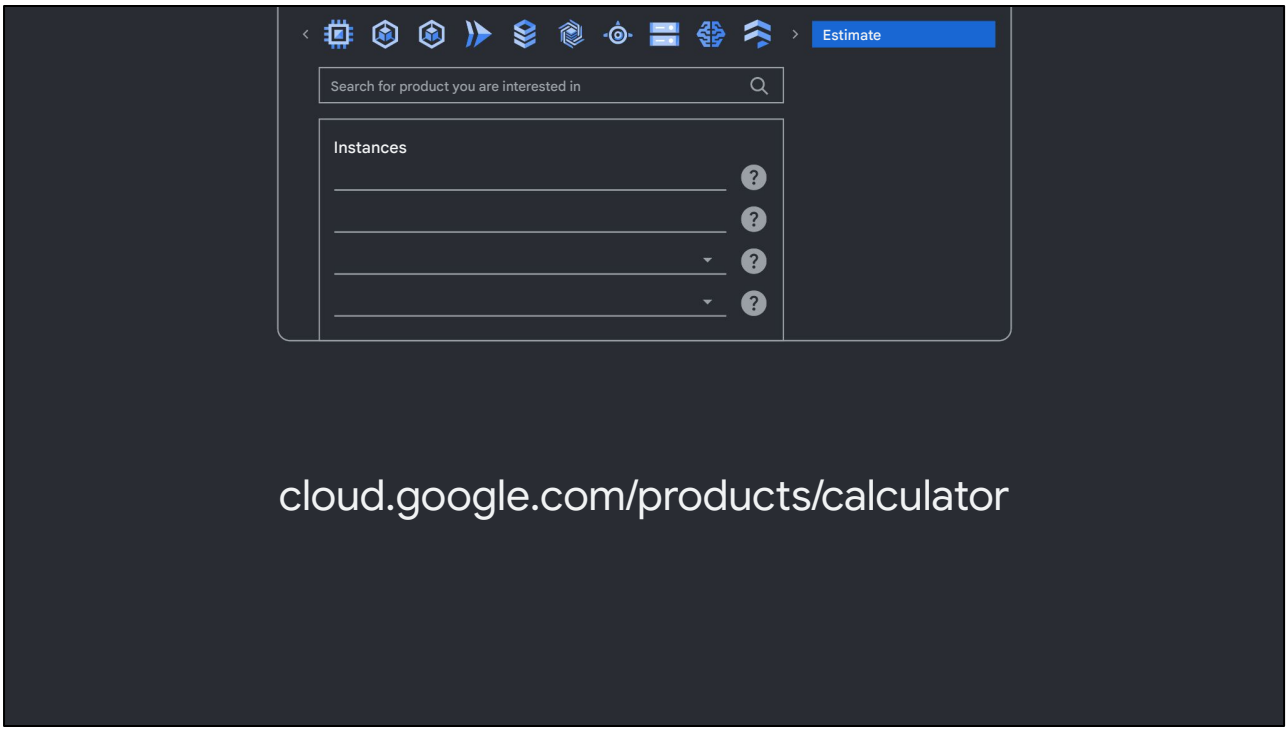
Specifically, when you run an instance for more than 25% of a month, Compute Engine automatically gives you a discount for every incremental minute you use for that instance.



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.

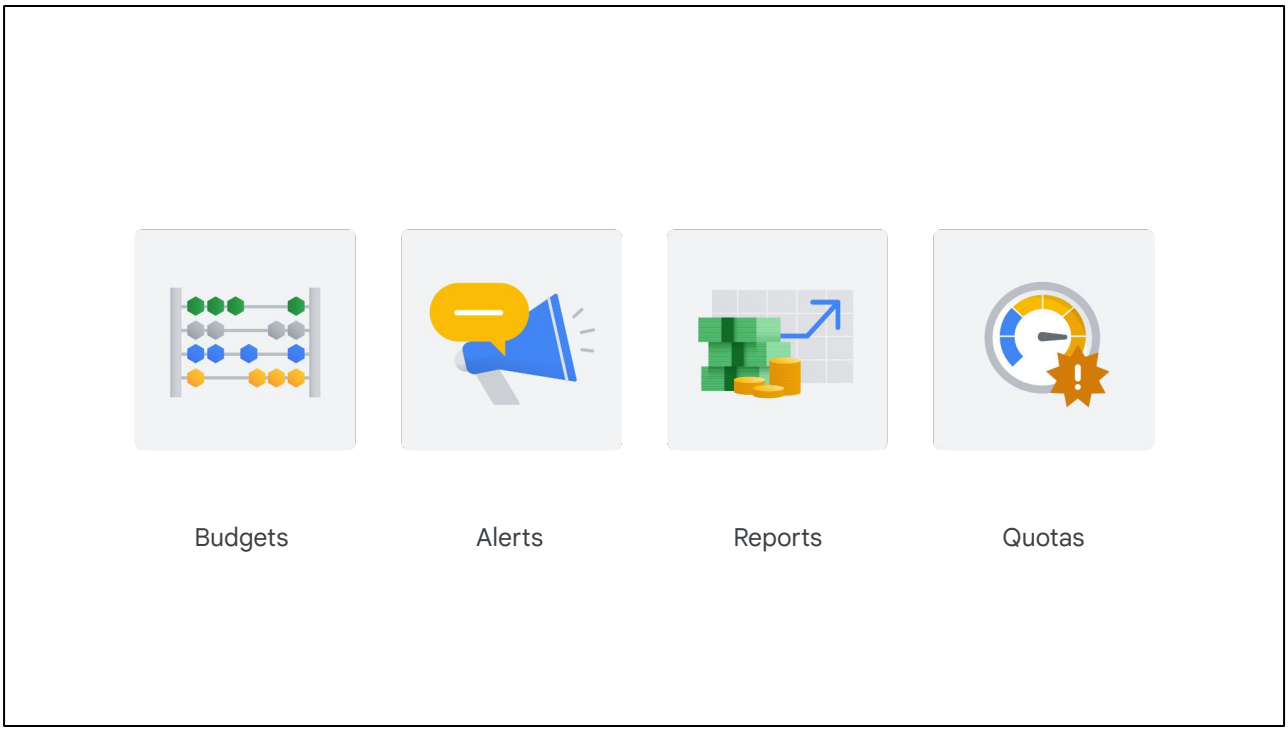


Visit cloud.google.com/products/calculator to try it out.

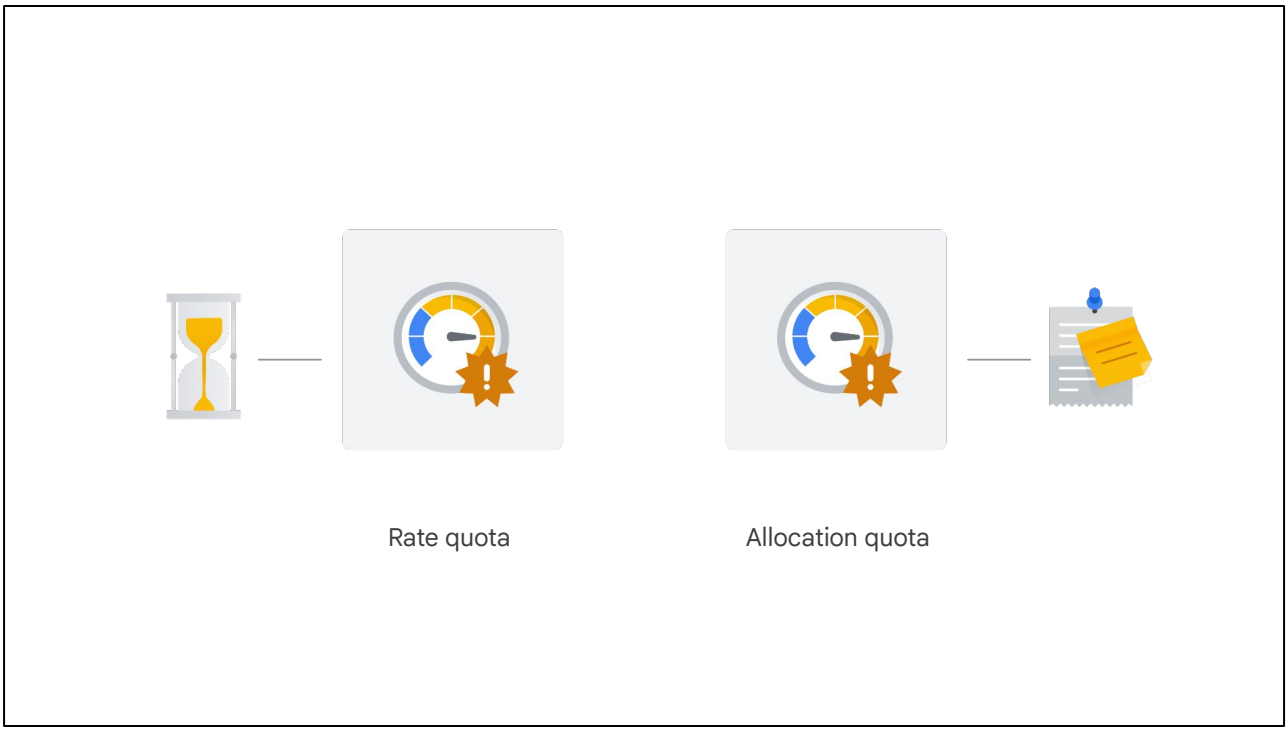
How can I make sure I don’t accidentally run up a big Google Cloud bill?

Now, you’re probably thinking, “How can I make sure I don’t accidentally run up a big Google Cloud bill?”

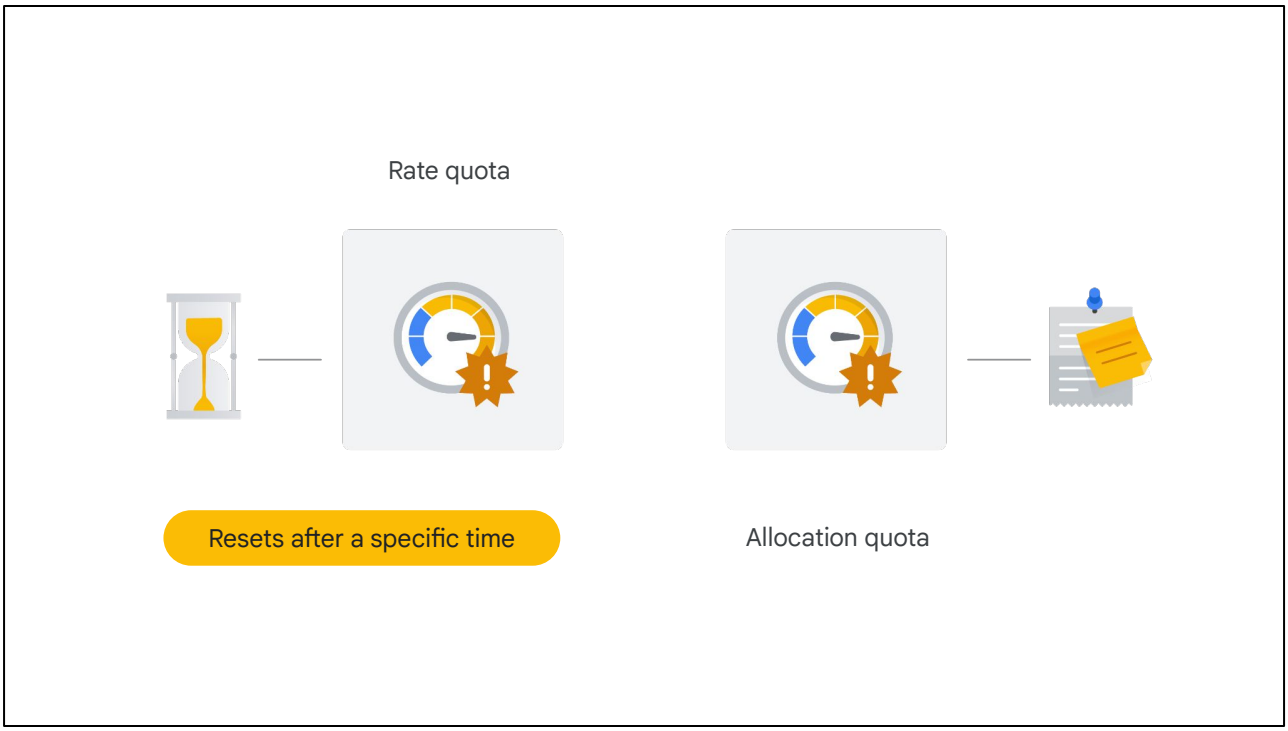
We provide a few tools to help.



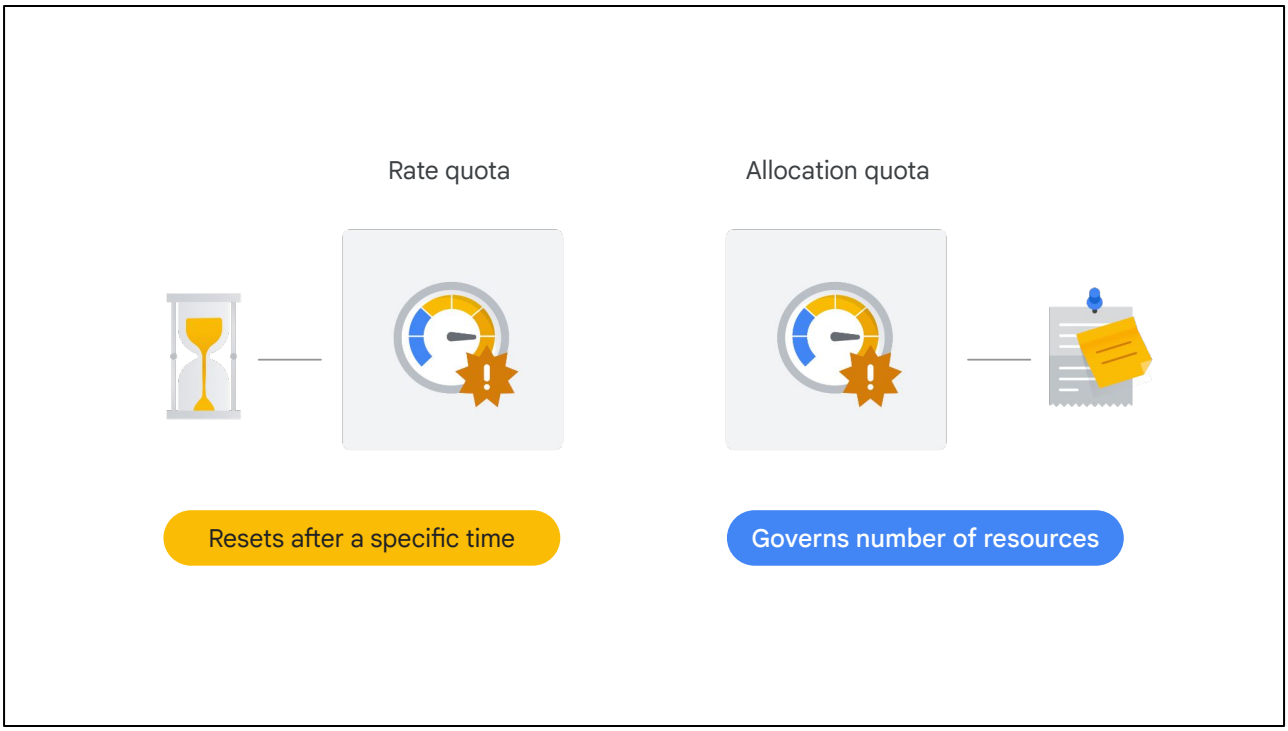
- You can define **budgets** at the billing account level or at the project level. A budget can be a fixed limit, or it can be tied to another metric; for example, a percentage of the previous month's spend.
- To be notified when costs approach your budget limit, you can create an **alert**. For example, 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.
- **Reports** is a visual tool in the Google Cloud console that allows you to monitor expenditure based on a project or services.
- 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.



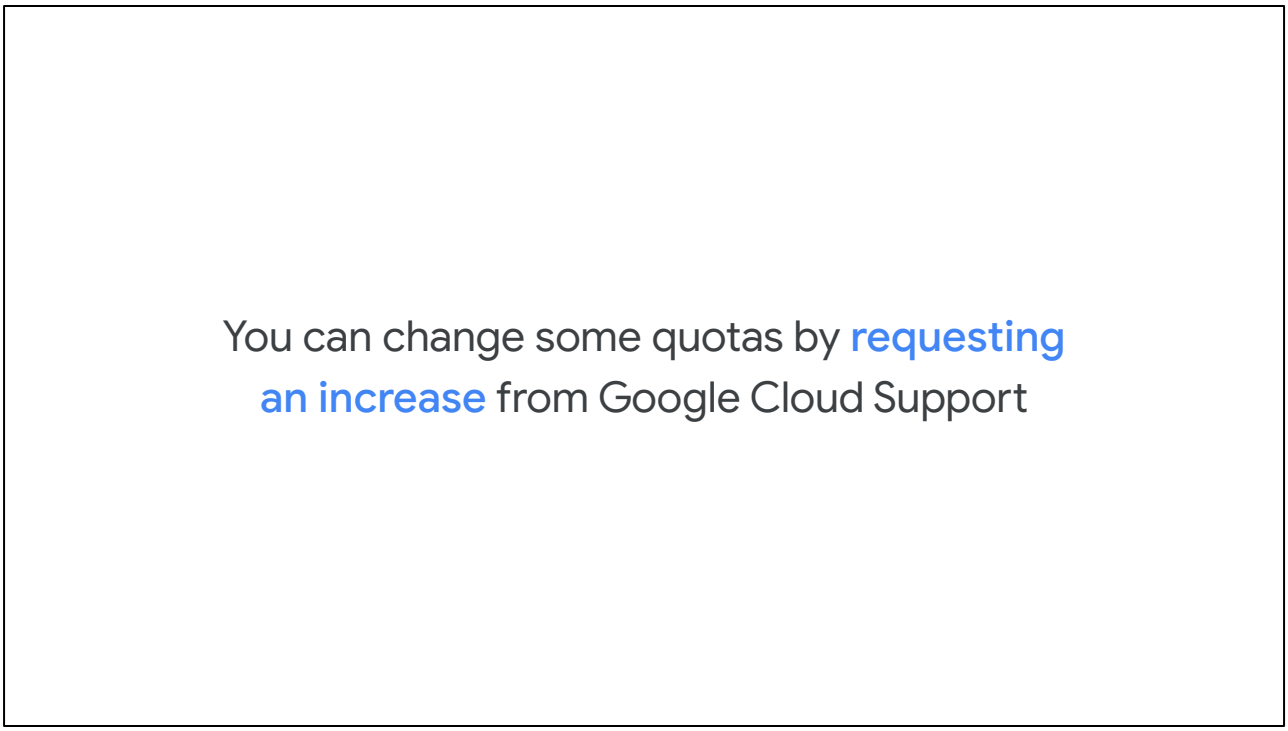
There are two types of quotas: **rate quotas** and **allocation quotas**. Both are applied at the project level.



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.



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 five 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.