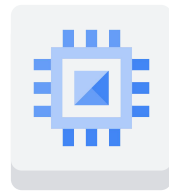
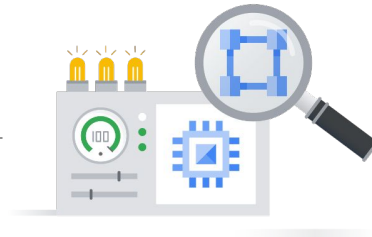


Google Cloud Core Infrastructure Module 3

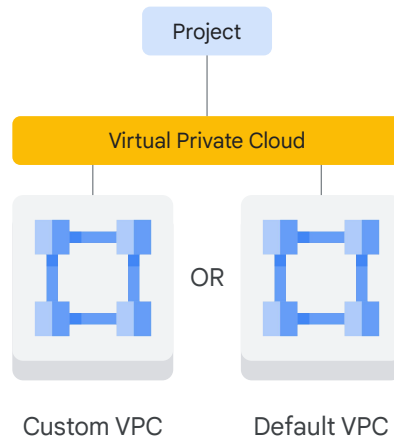
On-demand course
March 2022



Compute Engine



In this section of the course, we're going to explore how Google Compute Engine works with a focus on virtual networking.



Many users start with Google Cloud by defining their own virtual private cloud inside their first Google Cloud project or by starting with the default virtual private cloud.

So, what is a virtual private cloud?



Virtual private cloud (VPC)

A secure, individual, private cloud-computing model hosted within a public cloud

A virtual private cloud, or VPC, is a secure, individual, private cloud-computing model hosted within a public cloud – like Google Cloud!

01

Run code, store data, host websites,
and anything else that can be done in
an ordinary private cloud

02

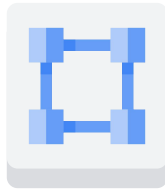
A virtual private cloud is hosted
remotely by a public cloud provider



On a VPC, customers can run code, store data, host websites, and do anything else they could do in an ordinary private cloud, but this private cloud is hosted remotely by a **public** cloud provider.



This means that VPCs combine the scalability and convenience of public cloud computing with the data isolation of private cloud computing.



VPC



VPC networks connect Google Cloud resources to each other and to the internet



Segmenting networks



Using firewall rules to restrict access to instances



Creating static routes to forward traffic to specific destinations



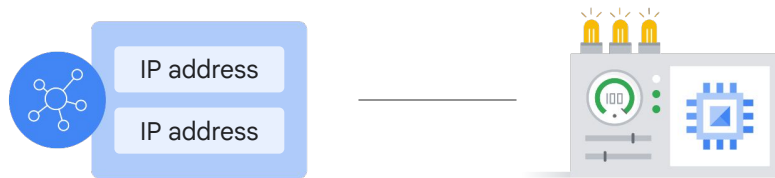
Google VPC networks are global and can have subnets in any Google Cloud region worldwide

VPC networks connect Google Cloud resources to each other and to the internet. This includes:

- segmenting networks,
- using firewall rules to restrict access to instances,
- and creating static routes to forward traffic to specific destinations.

Here's something that tends to surprise a lot of new Google Cloud users: **Google VPC networks are global**. They can also have subnets, which is a segmented piece of the larger network, in any Google Cloud region worldwide.

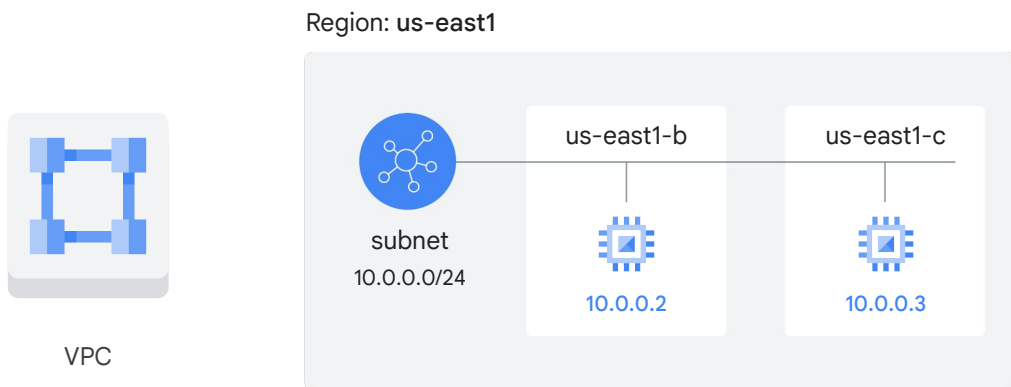
Subnets can span the zones that make up a region. This architecture makes it easy to define network layouts with global scope. Resources can even be in different zones on the same subnet.



The size of a subnet can be increased by expanding the



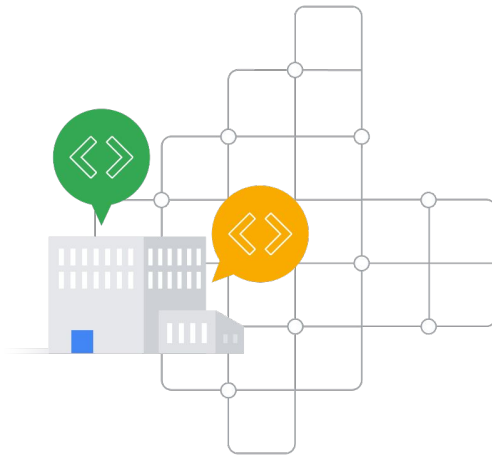
range of IP addresses allocated to it, and doing so won't affect virtual machines that are already configured.



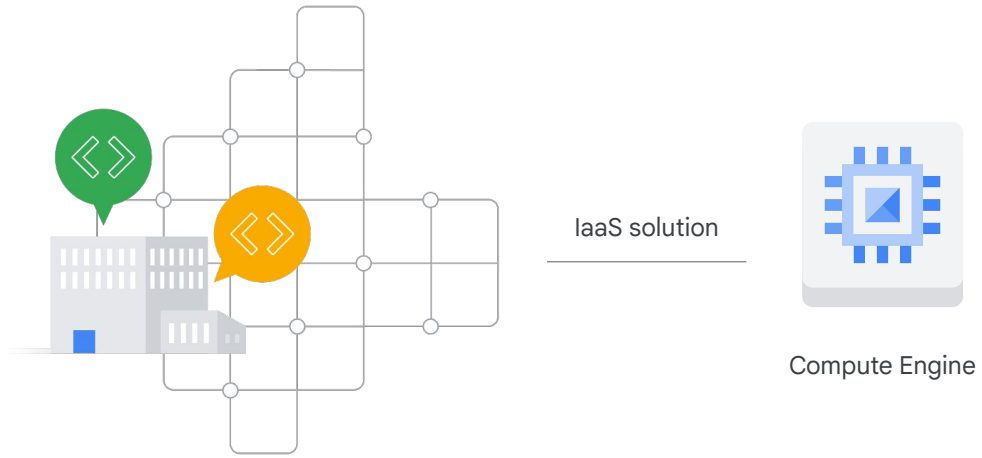
For example, let's take a VPC with **one network** that currently has **one subnet** defined in Google Cloud's *us-east1* region. If the VPC has two Compute Engine VMs attached to it, it means they're neighbors on the same subnet even though they're in different zones!

This capability can be used to build solutions that are **resilient to disruptions** yet retain a **simple network layout**

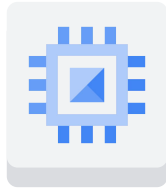
This capability can be used to build solutions that are resilient to disruptions yet retain a simple network layout.



Earlier in the course, we explored infrastructure as a service, or IaaS.



Now let's explore Google Cloud's IaaS solution: Compute Engine.



Compute Engine



Can create and run virtual machines on Google infrastructure



No upfront investments



Thousands of virtual CPUs can run on a system that's designed to be fast and to offer consistent performance



Each VM contains the power and functionality of a full-fledged operating system



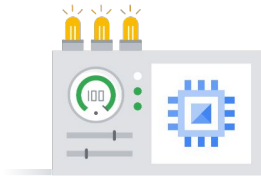
Can be configured much like a physical server

With Compute Engine, users can create and run virtual machines on Google infrastructure.

There are no upfront investments, and thousands of virtual CPUs can run on a system that's designed to be fast and to offer consistent performance.

Each virtual machine contains the power and functionality of a full-fledged operating system.

This means a virtual machine can be configured much like a physical server: by specifying the amount of CPU power and memory needed, the amount and type of storage needed, and the operating system.



Virtual machine

01

Can be created using the Google Cloud console, the Google Cloud CLI, or the Compute Engine API

02

Can run Linux and Windows Server images provided by Google or any customized versions of these images

03

Can build and run images of other operating systems and flexibly reconfigure VMs

A virtual machine instance can be created via the Google Cloud console, which is a web-based tool to manage Google Cloud projects and resources, the Google Cloud CLI, or the Compute Engine API.

The instance can run Linux and Windows Server images provided by Google or any customized versions of these images.

You can also build and run images of other operating systems and flexibly reconfigure virtual machines.

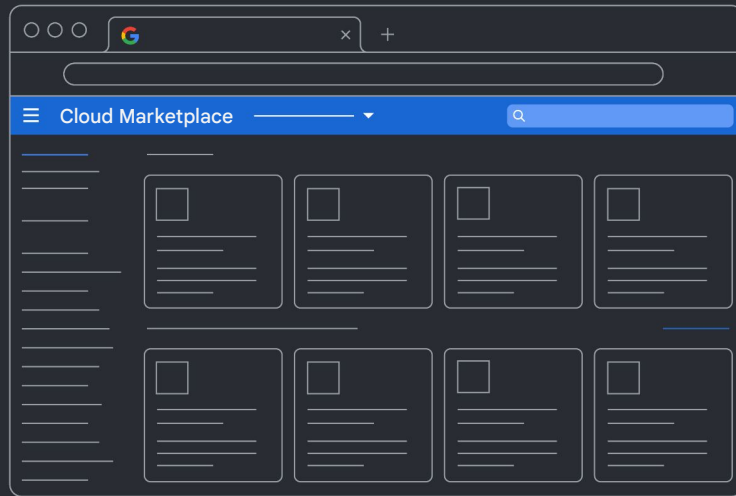


A quick way to get started with Google Cloud



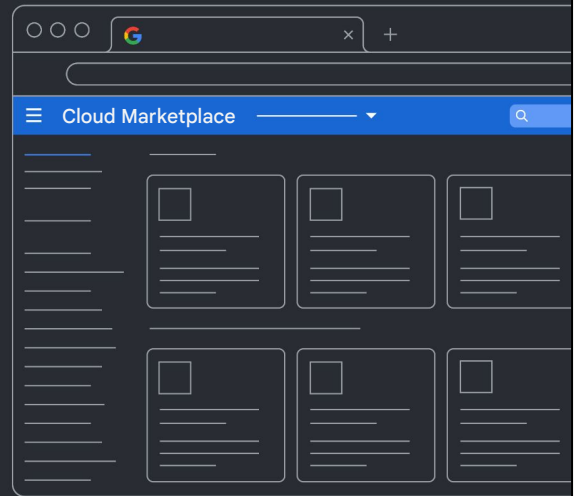
Cloud Marketplace

is through the Cloud Marketplace, which offers



solutions from both Google and third-party vendors. With these solutions, there's

No need to manually configure the software, VM instances, storage, or network settings



no need to manually configure the software, virtual machine instances, storage, or network settings, although many of them can be modified before launch if that's required.

01

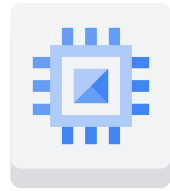
Most software packages are available at no additional charge beyond the normal fees

02

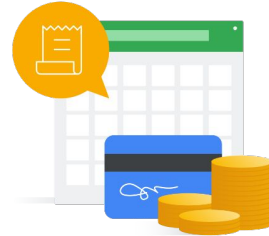
Some Cloud Marketplace images charge usage fees, but they all show estimates of their monthly charges before they're launched

Most software packages in Cloud Marketplace are available at no additional charge beyond the normal usage fees for Google Cloud resources.

Some Cloud Marketplace images charge usage fees, particularly those published by third parties, with commercially licensed software, but they all show estimates of their monthly charges before they're launched.

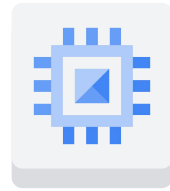
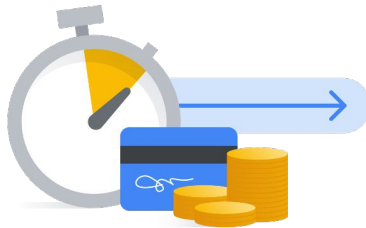


Compute Engine



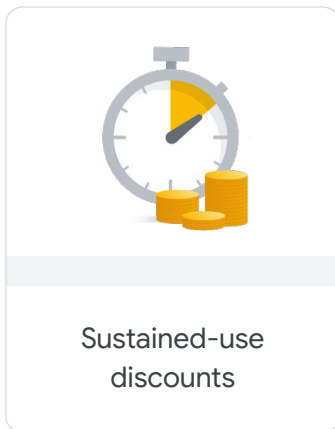
At this point, you might be wondering about Compute Engine's pricing and billing structure.

- ✓ Bills by the second
- ✓ 1 minute minimum

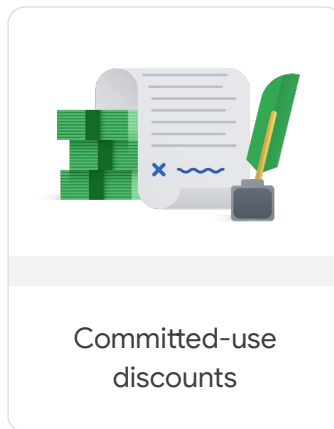


Compute Engine

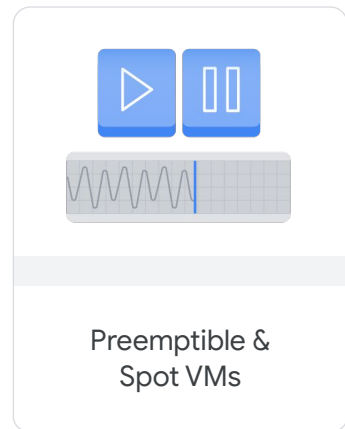
For the use of virtual machines, Compute Engine bills by the second with a one-minute minimum,



Sustained-use
discounts



Committed-use
discounts

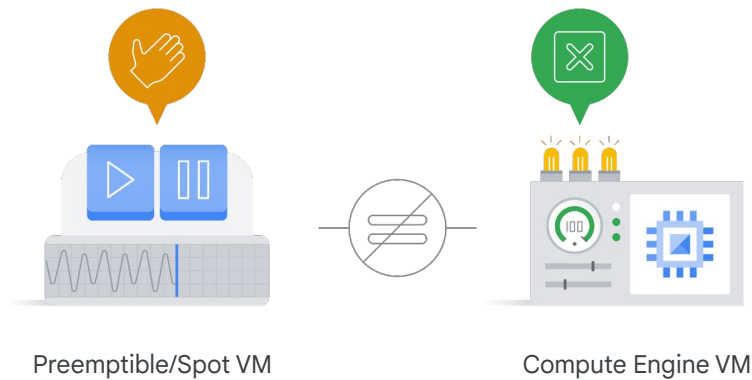


Preemptible &
Spot VMs

and **sustained-use discounts** start to apply automatically to virtual machines the longer they run. So, for each VM that runs for more than 25% of a month, Compute Engine automatically applies a discount for every additional minute.

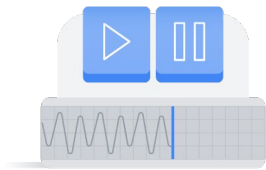
Compute Engine also offers **committed-use discounts**. This means that for stable and predictable workloads, a specific amount of vCPUs and memory can be purchased for up to a 57% discount off of normal prices in return for committing to a usage term of one year or three years.

And then there are **Preemptible and Spot VMs**. Let's say you have a workload that doesn't require a human to sit and wait for it to finish—such as a batch job analyzing a large dataset. You can save money, in some cases up to 90%, by choosing Preemptible VMs to run the job.



A Preemptible or Spot VM is different from an ordinary Compute Engine VM in only one respect: Compute Engine has permission to terminate a job if its resources are needed elsewhere.

Although savings are possible with preemptible or spot VMs, you'll need to ensure that your job can be stopped and restarted.



Preemptible/Spot VM

Spot VMs

- More features
- No maximum runtime
- Same pricing

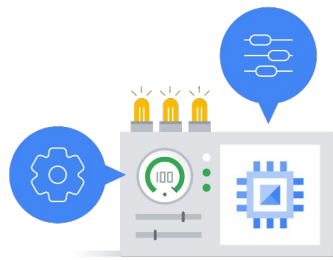
Preemptible VMs

- Less features
- Runtime up to 24h
- Same pricing

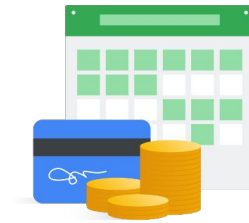
Spot VMs differ from Preemptible VMs by offering more features. For example, preemptible VMs can only run for up to 24 hours at a time, but Spot VMs do not have a maximum runtime. However, the pricing is, currently the same for both.

Compute Engine **doesn't require a particular option or machine type** to get high throughput between processing and persistent disks

In terms of storage, Compute Engine doesn't require a particular option or machine type to get high throughput between processing and persistent disks. That's the default, and it comes to you at no extra cost.



Custom virtual machines

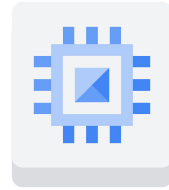


Pay for what you need

And finally, you'll only pay for what you need with custom machine types.

Choose machine properties by:

- ✓ Using a set of predefined machine types
- ✓ Creating your own custom machine types

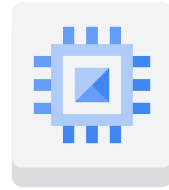


Compute Engine

Compute Engine lets you choose the machine properties of your instances, like the number of virtual CPUs and the amount of memory, by using a set of predefined machine types or by creating your own custom machine types.

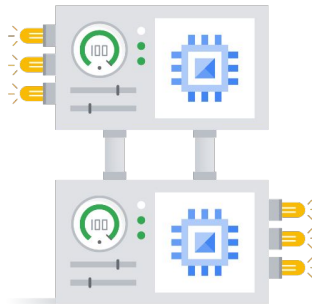
Choose machine properties by:

- ✓ Using a set of predefined machine types
- ✓ Creating your own custom machine types



Compute Engine

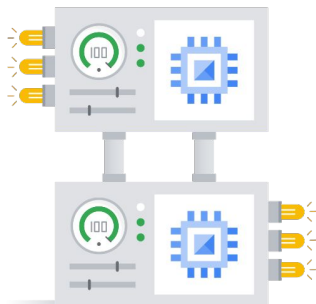
As we've just seen, with Compute Engine, you can choose the most appropriate machine properties for your instances, like the number of virtual CPUs and the amount of memory, by using a set of predefined machine types, or by creating custom machine types.



Autoscaling

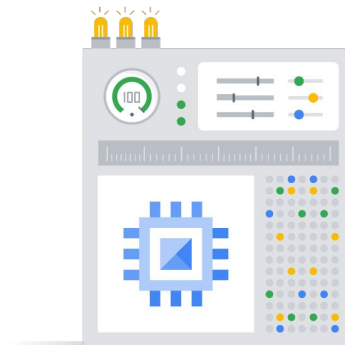
Resilient and scalable apps

To do this, Compute Engine has a feature called Autoscaling, where VMs can be added to or subtracted from an application based on load metrics. The other part of making that work is balancing the incoming traffic among the VMs. Google's Virtual Private Cloud (VPC) supports several different kinds of load balancing, which we'll explore shortly.



Autoscaling

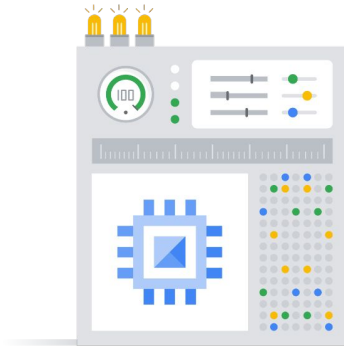
Resilient and scalable apps



Large Virtual Machines

Memory/compute-intensive apps

With Compute Engine, you can in fact configure very large VMs, which are great for workloads such as in-memory databases and CPU-intensive analytics, but most Google Cloud customers start off with scaling *out*, not up.



Maximum number of CPUs per VM

- ✓ Tied to its machine family
- ✓ Constrained by the user's quota

The maximum number of CPUs per VM is tied to its “machine family” and is also constrained by the quota available to the user, which is zone-dependent.

Specifications for currently available
VM machine types can be found at:

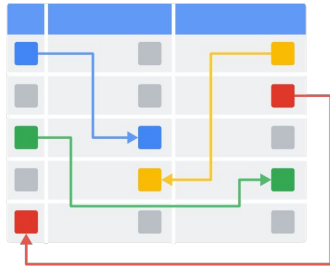
cloud.google.com/compute/docs/machine-types

Specifications for currently available VM machine types can be found at
cloud.google.com/compute/docs/machine-types



Virtual Private Cloud

Now let's explore some of the most important Virtual Private Cloud compatibility features.



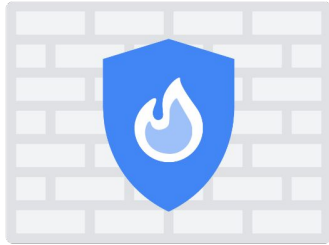
Routing tables

- ✓ Routing tables are built-in
- ✓ No router provisioning or managing
- ✓ Forward traffic from one instance to another
- ✓ No external IP address required

Much like physical networks, VPCs have **routing tables**.

VPC routing tables are built-in so you don't have to provision or manage a router.

They're used to forward traffic from one instance to another within the same network, across subnetworks, or even between Google Cloud zones, without requiring an external IP address.



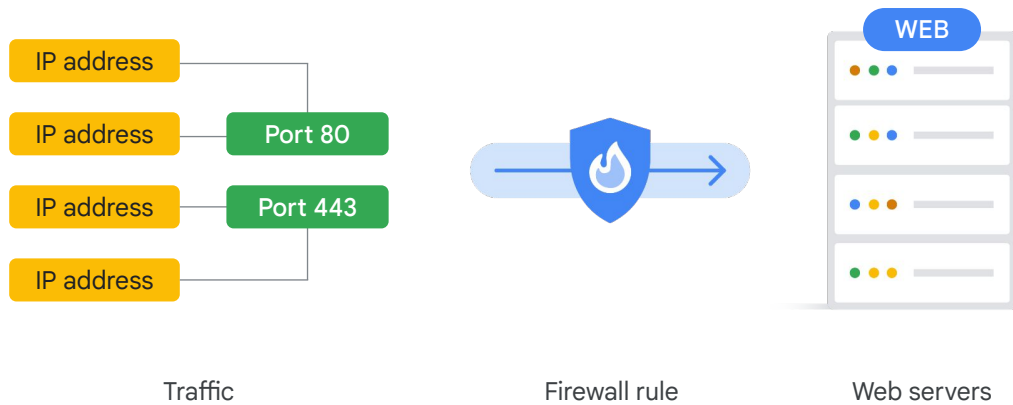
Firewall

- ✓ No router provisioning or managing
- ✓ Restrict access to instances
- ✓ Rules can be defined through network tags

Another thing you don't have to provision or manage for Google Cloud is a **firewall**.

VPCs provide a global distributed firewall, which can be controlled to restrict access to instances through both incoming and outgoing traffic.

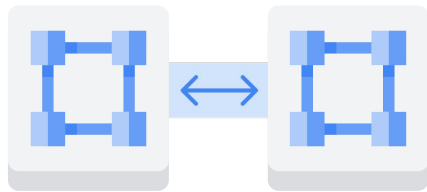
Firewall rules can be defined through network tags on Compute Engine instances, which is really convenient.



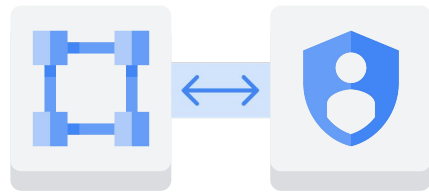
For example, you can tag all your web servers with, say, “WEB,” and write a firewall rule saying that traffic on ports 80 or 443 is allowed into all VMs with the “WEB” tag, no matter what their IP address happens to be.

VPCs belong to Google Cloud projects, but what if your company has **several** Google Cloud projects and the VPCs need to talk to each other?

You'll remember that VPCs belong to Google Cloud projects, but what if your company has several Google Cloud projects, and the VPCs need to talk to each other?



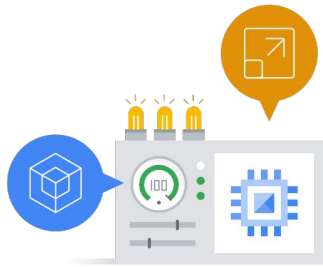
VPC peering



Shared VPC

With **VPC Peering**, a relationship between two VPCs can be established to exchange traffic.

Alternatively, to use the full power of Identity Access Management (IAM) to control who and what in one project can interact with a VPC in another, you can configure a **Shared VPC**.



Virtual machine

Previously, we explored how virtual machines can autoscale to respond to changing loads.

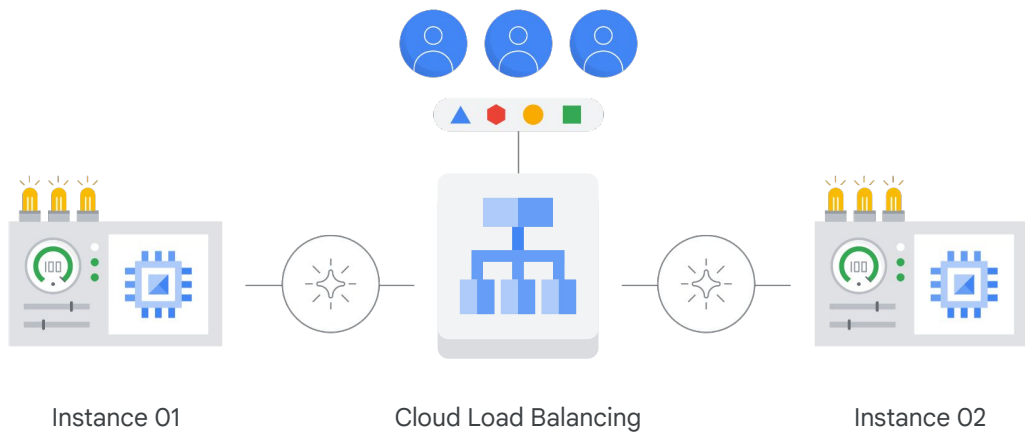
How do your customers get to your
application when it might be provided by
4 VMs one moment, and by 40 VMs at another?

But how do your customers get to your application when it might be provided by four
VMs one moment, and by 40 VMs at another?



Cloud Load Balancing

That's done through **Cloud Load Balancing**.



The job of a load balancer is to distribute user traffic across multiple instances of an application. By spreading the load, load balancing reduces the risk that applications experience performance issues.



Cloud Load Balancing



Fully distributed, software-defined, managed service



You can put Cloud Load Balancing in front of all of your traffic:



HTTP(S)



TCP traffic



SSL traffic



UDP traffic

Cloud Load Balancing is a fully distributed, software-defined, managed service for all your traffic. And because the load balancers don't run in VMs that you have to manage, you don't have to worry about scaling or managing them.

You can put Cloud Load Balancing in front of all of your traffic: HTTP or HTTPS, other TCP and SSL traffic, and UDP traffic too.



Cloud Load Balancing



Provides single as well as cross-region load balancing, including automatic multi-region failover

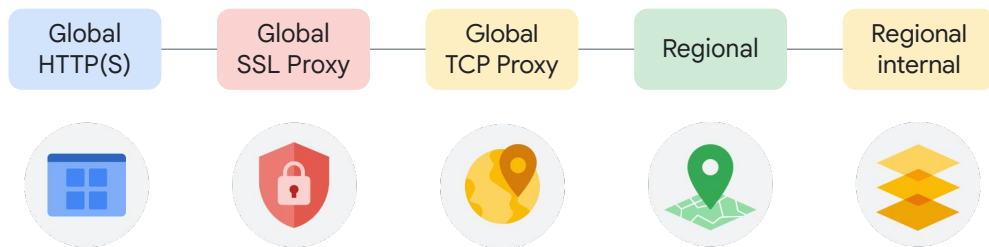


No “pre-warming” is required for anticipated spikes in traffic

Cloud Load Balancing provides cross-region load balancing, including automatic multi-region failover, which gently moves traffic in fractions if backends become unhealthy. Cloud Load Balancing reacts quickly to changes in users, traffic, network, backend health, and other related conditions.

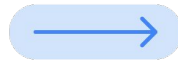
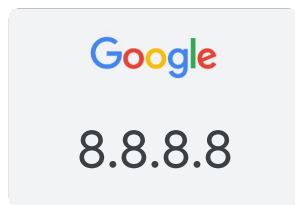
And what if you anticipate a huge spike in demand?

Say, your online game is already a hit; do you need to file a support ticket to warn Google of the incoming load? No. No so-called “pre-warming” is required.



VPC offers a suite of load-balancing options:

- If you need cross-regional load balancing for a web application, use **Global HTTP(S)** load balancing.
- For Secure Sockets Layer traffic that is not HTTP, use the **Global SSL Proxy** load balancer.
- If it's other TCP traffic that doesn't use SSL, use the **Global TCP Proxy** load balancer.
- Those last two proxy services only work for specific port numbers, and they only work for TCP. If you want to load balance UDP traffic, or traffic on any port number, you can still load balance across a Google Cloud region with the **Regional** load balancer.
- Finally, what all those services have in common is that they're intended for traffic coming into the Google network from the internet. But what if you want to load balance traffic inside your project, say, between the presentation layer and the business layer of your application? For that, use the **Regional internal** load balancer. It accepts traffic on a Google Cloud internal IP address and load balances it across Compute Engine VMs.



Domain Name Service

One of the most famous free Google services is 8.8.8.8, which provides a public **Domain Name Service** to the world.



Domain Name Service

Hostname 01

Hostname 02



IP address 01

IP address 02

DNS is what translates internet hostnames to addresses, and as you might imagine,

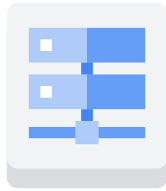
Google has a highly developed DNS infrastructure that makes **8.8.8.8 available** so that **everyone** can take advantage of it

Google has a highly developed DNS infrastructure. It makes 8.8.8.8 available so that everyone can take advantage of it.



Cloud DNS

But what about the internet hostnames and addresses of applications built in Google Cloud? Google Cloud offers **Cloud DNS** to help the world find them.



Cloud DNS



Managed DNS service that runs on the same infrastructure as Google.



Low latency, high availability, and cost-effective.



The DNS information you publish is served from redundant locations around the world.



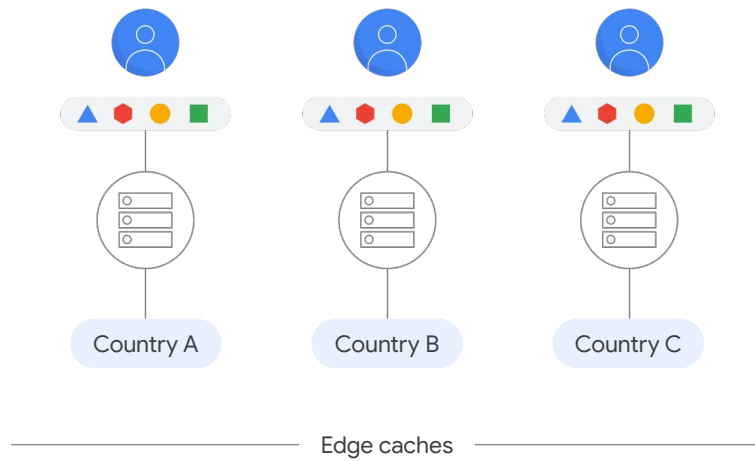
Cloud DNS is programmable. You can publish and manage millions of DNS zones and records using the Google Cloud console, the command-line interface, or the API.

It's a managed DNS service that runs on the same infrastructure as Google.

It has low latency and high availability, and it's a cost-effective way to make your applications and services available to your users.

The DNS information you publish is served from redundant locations around the world.

Cloud DNS is also programmable. You can publish and manage millions of DNS zones and records using the Google Cloud console, the command-line interface, or the API.



Google also has a global system of edge caches. Edge caching refers to the use of caching servers to store content closer to end users.



Cloud CDN

You can use this system to accelerate content delivery in your application by using **Cloud CDN** - Content Delivery Network.



Cloud CDN



Lower network latency



Origins of content will experience reduced load



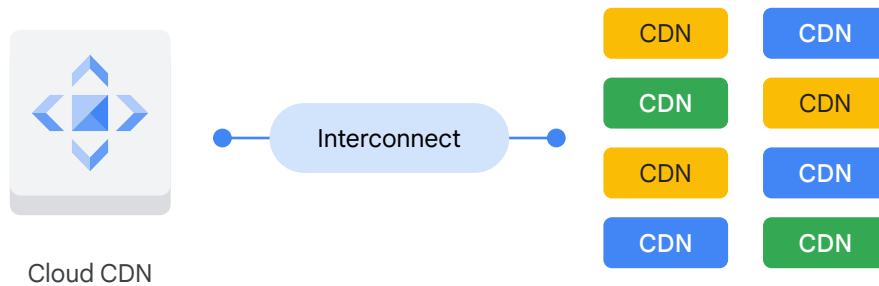
Save money



Enabled with a single checkbox

This means your customers will experience lower network latency, the origins of your content will experience reduced load, and you can even save money.

After HTTP(S) Load Balancing is set up, Cloud CDN can be enabled with a single checkbox.

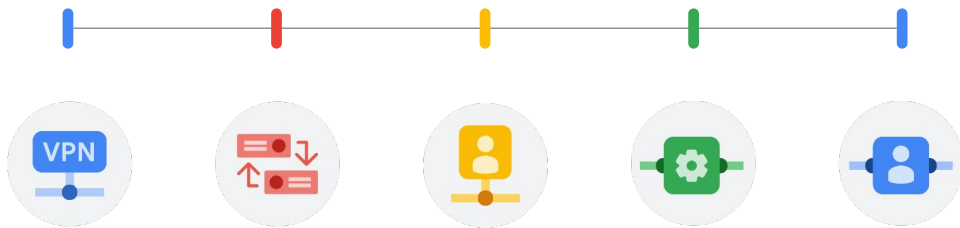


There are many other CDNs available out there, of course. If you are already using one, chances are, it's a part of Google Cloud's CDN Interconnect partner program, and you can continue to use it.

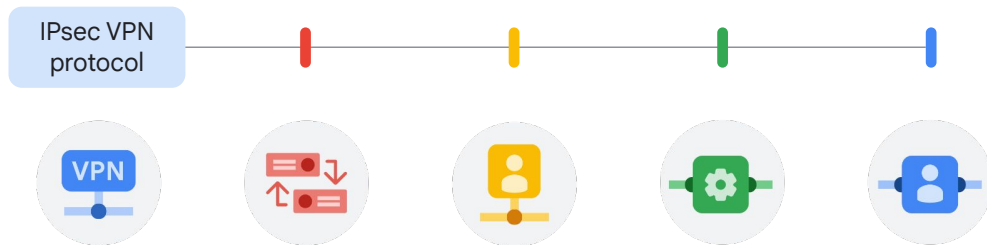


Virtual Private Cloud

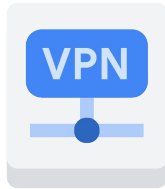
Many Google Cloud customers want to connect their Google Virtual Private Clouds to other networks in their system, such as on-premises networks or networks in other clouds.



There are several effective ways to accomplish this.



One option is to start with a Virtual Private Network connection over the internet and use the **IPsec VPN protocol** to create a “tunnel” connection.



IPsec VPN protocol



Uses **Cloud Router** to make the connection dynamic



Lets other networks and Google VPC exchange route information over the VPN using the **Border Gateway Protocol**

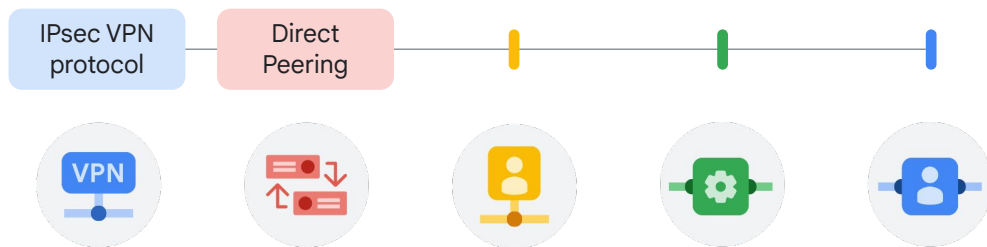


Not always the best option because of security concerns or bandwidth reliability

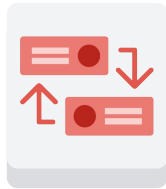
To make the connection dynamic, a Google Cloud feature called Cloud Router can be used.

Cloud Router lets other networks and Google VPC, exchange route information over the VPN using the Border Gateway Protocol. Using this method, if you add a new subnet to your Google VPC, your on-premises network will automatically get routes to it.

But using the internet to connect networks isn't always the best option for everyone, either because of security concerns or because of bandwidth reliability.



So, a second option is to consider “peering” with Google using **Direct Peering**.



Direct Peering



Puts a router in the same public datacenter as a Google point of presence (PoP)



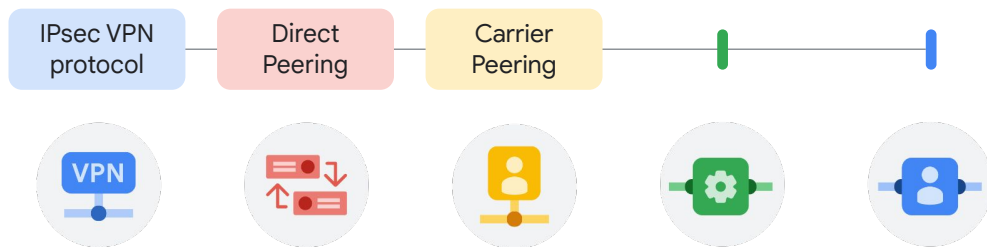
Uses a router to exchange traffic between networks



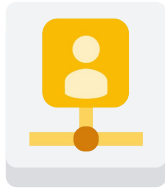
More than 100 Google points of presence around the world

Peering means putting a router in the same public data center as a Google point of presence and using it to exchange traffic between networks.

Google has more than 100 points of presence around the world.



Customers who aren't already in a point of presence can work with a partner in the **Carrier Peering** program to get connected.



Carrier Peering



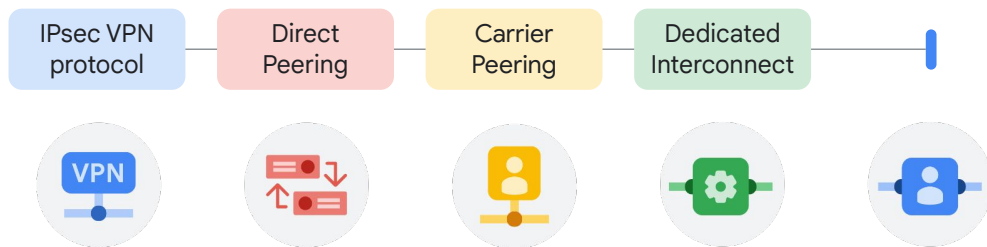
Gives direct access from an on-premises network through a service provider's network



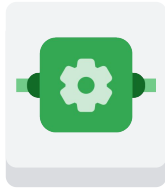
Not covered by a Google Service Level Agreement

Carrier peering gives you direct access from your on-premises network through a service provider's network to Google Workspace and to Google Cloud products that can be exposed through one or more public IP addresses.

One downside of peering, though, is that it isn't covered by a Google Service Level Agreement.



If getting the highest uptimes for interconnection is important, using **Dedicated Interconnect** would be a good solution.



Dedicated Interconnect



Allows for one or more direct, private connections to Google



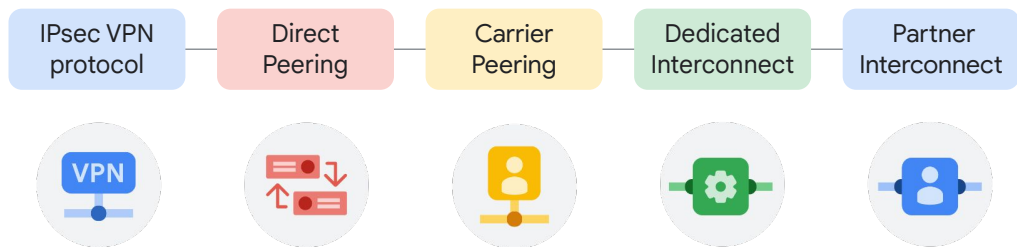
Can be covered by up to a 99.99% SLA



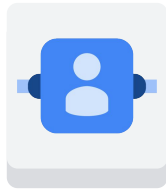
Connections can be backed up by a VPN

This option allows for one or more direct, private connections to Google.

If these connections have topologies that meet Google's specifications, they can be covered by an SLA of up to 99.99%. Also, these connections can be backed up by a VPN for even greater reliability.



And the final option we'll explore is **Partner Interconnect**, which provides connectivity between an on-premises network and a VPC network through a supported service provider.



Partner Interconnect

- ✓ Useful if a data center is in a physical location that can't reach a Dedicated Interconnect colocation facility
- ✓ Useful if the data needs don't warrant an entire 10 GigaBytes per second connection
- ✓ Can be configured to support mission-critical services or applications that can tolerate some downtime
- ✓ Can be covered by up to a 99.99% SLA

A Partner Interconnect connection is useful if a data center is in a physical location that can't reach a Dedicated Interconnect colocation facility, or if the data needs don't warrant an entire 10 GigaBytes per second connection.

Depending on availability needs, Partner Interconnect can be configured to support mission-critical services or applications that can tolerate some downtime.

As with Dedicated Interconnect, if these connections have topologies that meet Google's specifications, they can be covered by an SLA of up to 99.99%, but note that Google isn't responsible for any aspects of Partner Interconnect provided by the third-party service provider, nor any issues outside of Google's network.