



Virtual Machines and Networks in the Cloud

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

Let's begin.

Virtual Machines and Networks in the Cloud

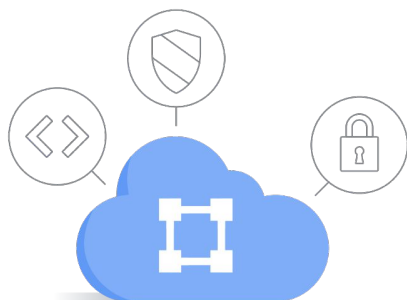
- | | |
|----|-----------------------------------|
| 01 | Virtual Private Cloud networking |
| 02 | Compute Engine |
| 03 | Scaling virtual machines |
| 04 | Important VPC compatibilities |
| 05 | Cloud Load Balancing |
| 06 | Cloud Domain Name Service |
| 07 | Cloud Content Delivery Network |
| 08 | Connecting Networks to Google VPC |



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 is your cloud within the Cloud

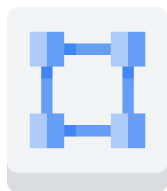


A **virtual private cloud**, or **VPC**, is a secure, individual, private cloud-computing model hosted within a public cloud

Google Cloud

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

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.



Virtual Private Cloud



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

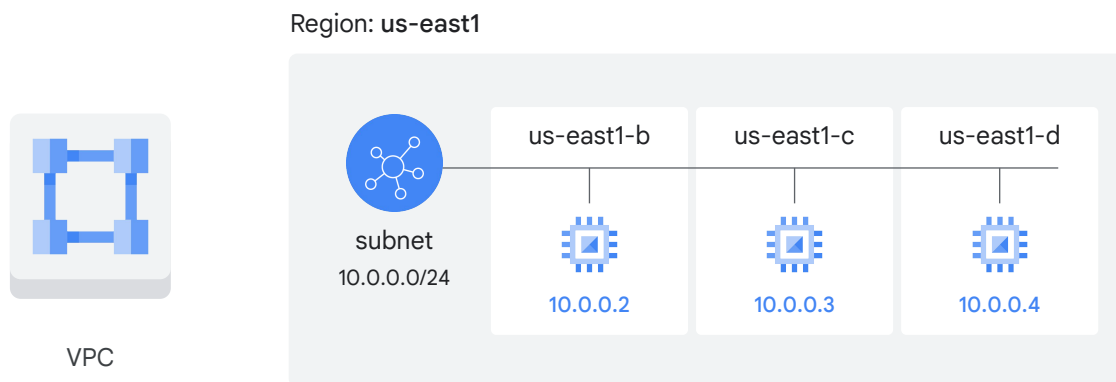


Google VPC networks are global and can have subnets in any Google Cloud region worldwide

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.

VPC subnets connect resources in different zones



The size of a subnet can be increased by expanding the range of IP addresses allocated to it. And doing so won't affect already configured virtual machines.

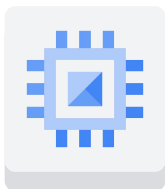
For example, let's take a VPC with **one network** that currently has **one subnet** defined in Google Cloud's *us-east1* region. If the VPC has three Compute Engine VMs attached to it, it means they're neighbors on the same subnet even though they are in different zones! This capability can be used to build solutions that are resilient to disruptions, yet retain a simple network layout.

Virtual Machines and Networks in the Cloud

- | | |
|----|-----------------------------------|
| 01 | Virtual Private Cloud networking |
| 02 | Compute Engine |
| 03 | Scaling virtual machines |
| 04 | Important VPC compatibilities |
| 05 | Cloud Load Balancing |
| 06 | Cloud Domain Name Service |
| 07 | Cloud Content Delivery Network |
| 08 | Connecting Networks to Google VPC |



Earlier in the course, we explored Infrastructure as a Service, or IaaS. Now let's explore Google Cloud's IaaS solution: Google Compute Engine.



Compute Engine



Compute Engine lets users create and run virtual machines on Google infrastructure

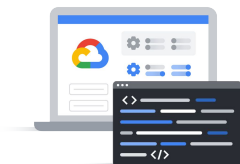


Each VM contains the power and functionality of a full-fledged operating system

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 is designed to be fast and 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 instances are flexible resources



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

Linux Server images

Windows Server images

provided by Google

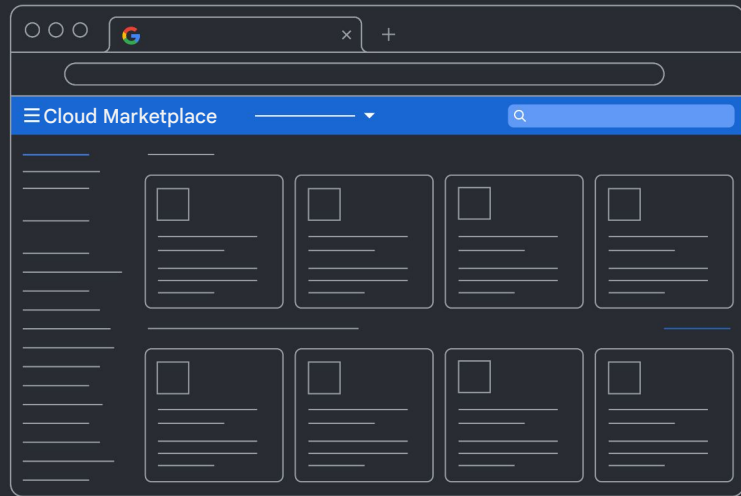
VM can run Linux and Windows Server images or any customized versions of these images



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

You can create a virtual machine instance or create a group of managed instances by using 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.

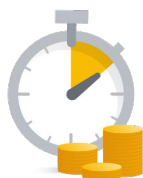
The instance can run Linux and Windows Server images provided by Google, or any customized versions of these images. You can also build and run images of other operating systems and flexibly reconfigure virtual machines.



A quick way to get started with Google Cloud 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, virtual machine instances, storage, or network settings, although many of them can be modified before launch if that's required.

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.

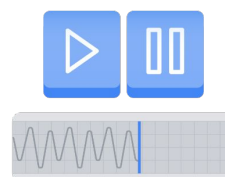
Compute Engine pricing and billing offers discounts



Sustained-use



Committed-use



Preemptible &
Spot VMs

At this point, you might be wondering about pricing and billing related to Compute Engine.

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

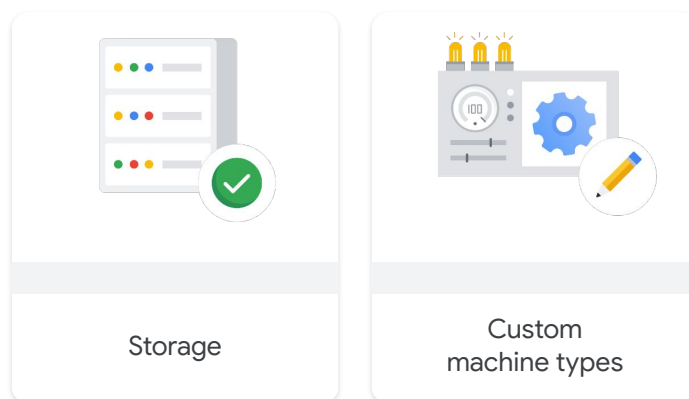
And then there are **Preemptible and Spot VMs**. Let's say you have a workload that doesn't require a human to sit and wait for it to finish—like a batch job analyzing a large dataset, for example. You can save money, in some cases up to 90%, by choosing Preemptible VMs to run the job.

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. While savings can be had with preemptible or spot VMs, you'll need to ensure your job can be stopped and restarted.

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 pricing is “pay for only what you need”



Google Cloud

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.

Compute Engine offers several types of storage options for your instances. Each of the following storage options has unique price and performance characteristics:

- Zonal persistent disk: Efficient, reliable block storage.
- Regional persistent disk: Regional block storage replicated in two zones.
- Local SSD: High performance, transient, local block storage.
- Cloud Storage buckets: Affordable object storage.
- Filestore: High performance file storage for Google Cloud users.

If you are not sure which option to use, the most common solution is to add a persistent disk to your instance.

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

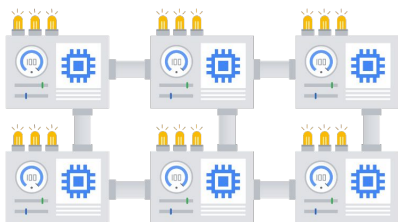
For more information, see Storage options at <https://cloud.google.com/compute/docs/disks>

Virtual Machines and Networks in the Cloud

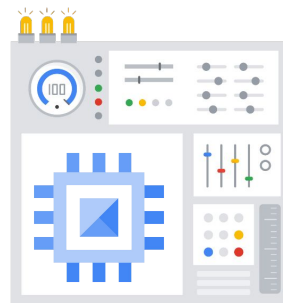
- | | |
|----|-----------------------------------|
| 01 | Virtual Private Cloud networking |
| 02 | Compute Engine |
| 03 | Scaling virtual machines |
| 04 | Important VPC compatibilities |
| 05 | Cloud Load Balancing |
| 06 | Cloud Domain Name Service |
| 07 | Cloud Content Delivery Network |
| 08 | Connecting Networks to Google VPC |



VMs can be auto-scaled to meet demand



Use **Autoscaling** for resilient, scalable applications



Use **big VMs** for memory- and compute-intensive applications

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.

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.

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. 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.

Specs for currently available VM machine types:
cloud.google.com/compute/docs/machine-types

Specs for currently available VM machine types can be found at
<https://cloud.google.com/compute/docs/machine-types>

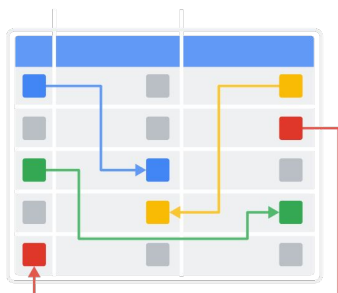
Virtual Machines and Networks in the Cloud

- | | |
|----|-----------------------------------|
| 01 | Virtual Private Cloud networking |
| 02 | Compute Engine |
| 03 | Scaling virtual machines |
| 04 | Important VPC compatibilities |
| 05 | Cloud Load Balancing |
| 06 | Cloud Domain Name Service |
| 07 | Cloud Content Delivery Network |
| 08 | Connecting Networks to Google VPC |



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

VPCs do not require a router to be provisioned



Routing tables



Built-in so you don't have to provision or manage a router



Used to forward traffic from one instance to another

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 are 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.

VPCs also do not require a firewall to be provisioned



Firewall



Restricts access to instances through both incoming and outgoing traffic



Rules can be defined through network tags on Compute Engine instances

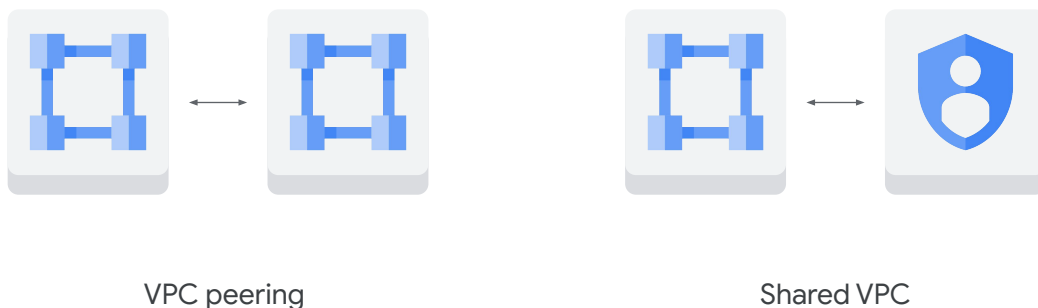
Another thing you don't have to provision or manage for Google Cloud is a **firewall**.

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.

VPC peering and sharing allow projects to communicate



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?

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, then you can configure a **Shared VPC**.

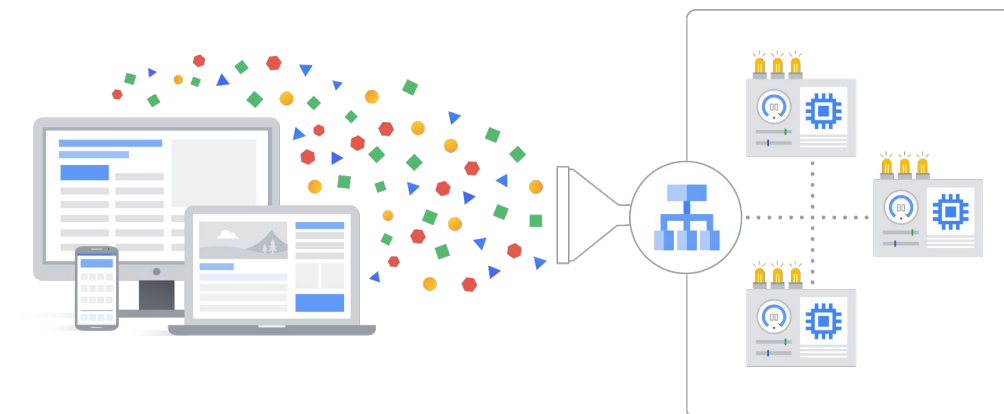
Virtual Machines and Networks in the Cloud

- | | |
|----|-----------------------------------|
| 01 | Virtual Private Cloud networking |
| 02 | Compute Engine |
| 03 | Scaling virtual machines |
| 04 | Important VPC compatibilities |
| 05 | Cloud Load Balancing |
| 06 | Cloud Domain Name Service |
| 07 | Cloud Content Delivery Network |
| 08 | Connecting Networks to Google VPC |



Previously, we explored how virtual machines can autoscale to respond to changing loads. But how do your customers get to your application when it might be provided by four VMs one moment, and by forty VMs at another? That's done through **Cloud Load Balancing**.

Load balancing distributes traffic across instances



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(S), 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

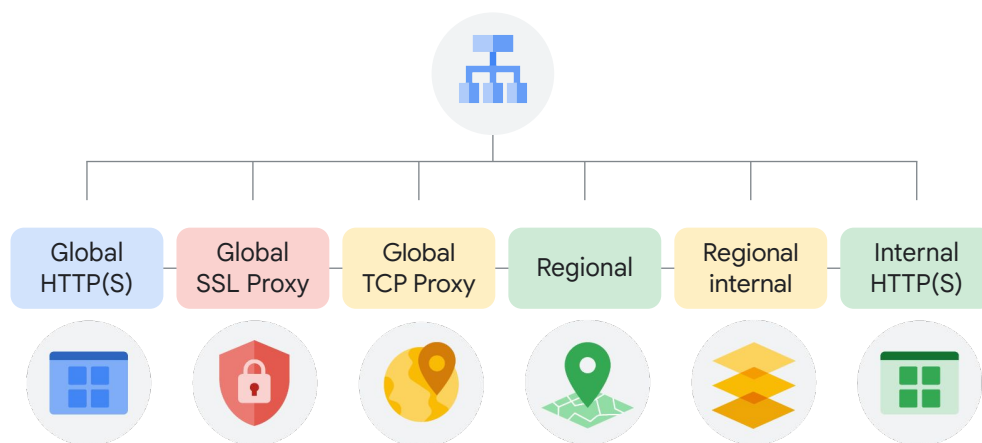


No “pre-warming” is required for anticipated spikes in traffic

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.

There are several load balancing options



Google Cloud

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 does not use Secure Sockets Layer, use the **Global TCP Proxy** load balancer.

Those 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.

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.

Finally the Google Cloud **Internal HTTP(S)** Load Balancing is a proxy-based, regional Layer 7 load balancer that also enables you to run and scale your services behind an internal load balancing IP address. Internal HTTP(S) Load Balancing is a managed

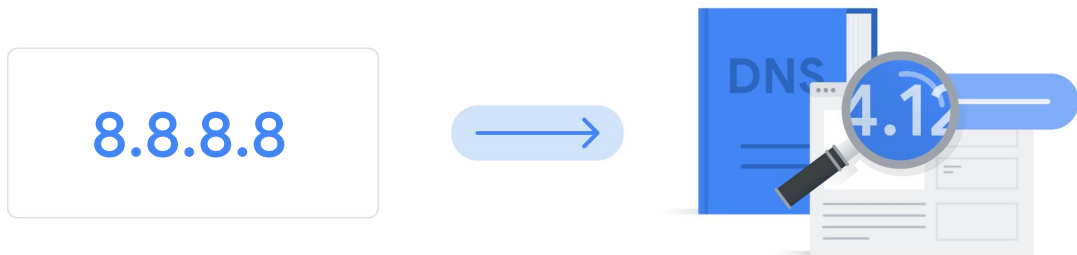
service based on the open source Envoy proxy. This enables rich traffic control capabilities based on HTTP(S) parameters. After the load balancer has been configured, it automatically allocates Envoy proxies to meet your traffic needs.

Virtual Machines and Networks in the Cloud

- | | |
|----|---|
| 01 | Virtual Private Cloud networking |
| 02 | Compute Engine |
| 03 | Scaling virtual machines |
| 04 | Important VPC compatibilities |
| 05 | Cloud Load Balancing |
| 06 | Cloud Domain Name Service |
| 07 | Cloud Content Delivery Network |
| 08 | Connecting Networks to Google VPC |

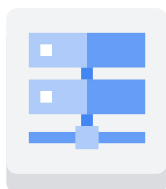


Google provides public DNS services



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. DNS is what translates Internet hostnames to addresses, and as you might imagine, Google has a highly developed DNS infrastructure. It makes 8.8.8.8 available so that everybody can take advantage of it.



Cloud DNS

- ✓ Managed DNS service that runs on the same infrastructure as Google
- ✓ Low latency, high availability, and cost-effective way to make applications and services available to users
- ✓ 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.

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. 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.

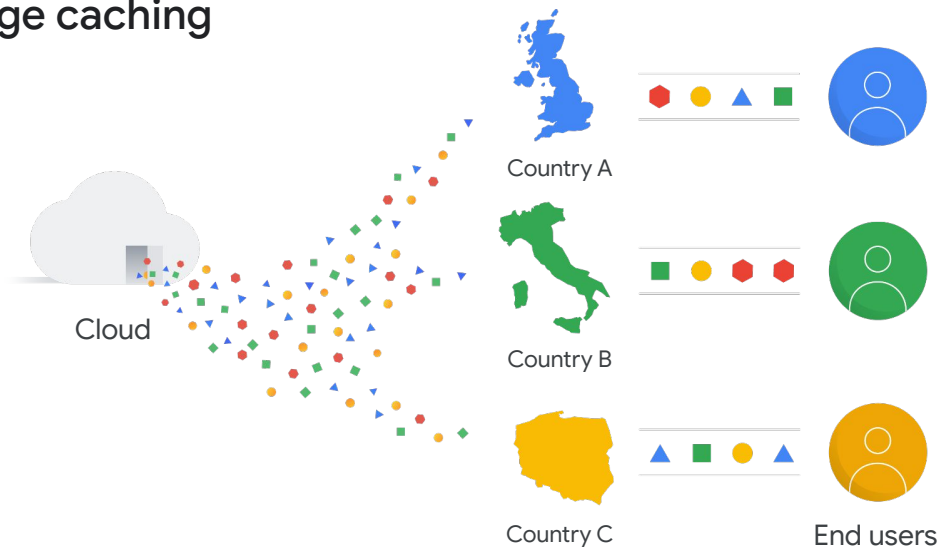
Cloud DNS is also programmable. You can publish and manage millions of DNS zones and records using the Google Cloud console, the command-line interface, or the API.

Virtual Machines and Networks in the Cloud

- | | |
|----|--|
| 01 | Virtual Private Cloud networking |
| 02 | Compute Engine |
| 03 | Scaling virtual machines |
| 04 | Important VPC compatibilities |
| 05 | Cloud Load Balancing |
| 06 | Cloud Domain Name Service |
| 07 | Cloud Content Delivery Network |
| 08 | Connecting Networks to Google VPC |



Edge caching



Google 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

- ✓ Lower network latency
- ✓ Origins of content will experience reduced load
- ✓ Save money
- ✓ Enabled with a single checkbox

You can use this system to accelerate content delivery in your application using **Cloud CDN** - Content Delivery Network. This means your customers will experience lower network latency, the origins of your content will experience reduced load, and you can even save money. Once HTTP(S) Load Balancing is set up, Cloud CDN can be enabled with a single checkbox.

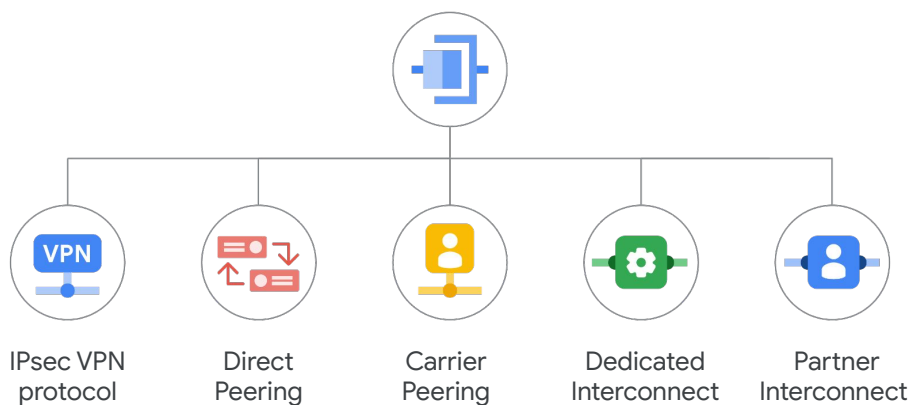
There are many other CDNs available out there, of course. If you are already using one, chances are, it's a part of Google Cloud's CDN Interconnect partner program, and you can continue to use it.

Virtual Machines and Networks in the Cloud

- | | |
|----|---|
| 01 | Virtual Private Cloud networking |
| 02 | Compute Engine |
| 03 | Scaling virtual machines |
| 04 | Important VPC compatibilities |
| 05 | Cloud Load Balancing |
| 06 | Cloud Domain Name Service |
| 07 | Cloud Content Delivery Network |
| 08 | Connecting Networks to Google VPC |



Google VPNs can connect to other networks



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 a number of effective ways to accomplish this.

IPsec VPN creates dynamic connections

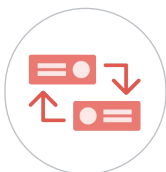


IPsec VPN protocol

- ✓ Creates a VPN “tunnel” connection
- ✓ 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

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. 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.

Direct peering routes traffic through a Google PoP



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



Connects to more than 100 Google points of presence around the world

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.

A second option is to consider “peering” with Google using **Direct Peering**. Peering means putting a router in the same public datacenter as a Google point of presence and using it to exchange traffic between networks. Google has more than 100 points of presence (PoPs) around the world.

Carrier peering allows access to Google Workspace



Carrier Peering



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



Not covered by a Google Service Level Agreement

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

Dedicated Interconnect is a direct connection to Google



Dedicated
Interconnect



Good solution for getting the highest uptimes for interconnection



Allows for one or more direct, private connections to Google



Can be covered by up to a 99.99% SLA if connections have topologies that meet Google's specifications



Connections can be backed up by a VPN for even greater reliability

If getting the highest uptimes for interconnection is important, then using **Dedicated Interconnect** would be a good solution. 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 up to a 99.99% SLA. Also, these connections can be backed up by a VPN for even greater reliability.

Partner Interconnect links your on-prem with your VPC



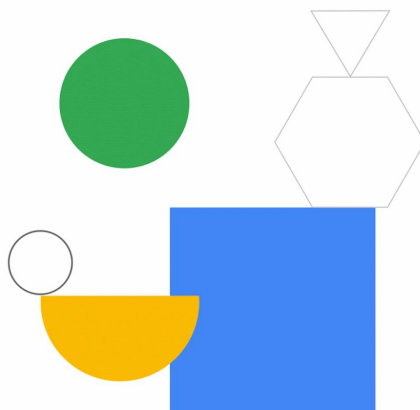
Partner
Interconnect

- ✓ Provides connectivity between an on-premises network and a VPC network through a supported service provider
- ✓ 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 if connections have topologies that meet Google's specifications

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. 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 doesn't warrant an entire 10 Gbps 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 up to a 99.99% SLA, but note that Google is not responsible for any aspects of Partner Interconnect provided by the third party service provider nor any issues outside of Google's network.

Module Quiz



Quiz | Question 1

Question

In Google Cloud VPCs, what scope do subnets have?

- A. Zonal
- B. Regional
- C. Global
- D. Multi-regional

Quiz | Question 1

Answer

In Google Cloud VPCs, what scope do subnets have?

- A. Zonal
- B. Regional
- C. Global
- D. Multi-regional



In Google Cloud VPCs, what scope do subnets have?

A: Zonal

Feedback: Sorry, that's not correct. VPC subnets can span the zones that make up a region. This is beneficial because your solutions can incorporate fault tolerance without complicating your network topology.

B: Regional

Feedback: That's correct. VPC subnets can span the zones that make up a region. This is beneficial because your solutions can incorporate fault tolerance without complicating your network topology.

C: Global

Feedback: Sorry, that's not correct. VPC subnets can span the zones that make up a region. This is beneficial because your solutions can incorporate fault tolerance without complicating your network topology.

D: Multi-regional

Feedback: Sorry, that's not correct. VPC subnets can span the zones that make up a region. This is beneficial because your solutions can incorporate fault tolerance without complicating your network topology.

Quiz | Question 2

Question

What is the main reason customers choose Preemptible or Spot VMs?

- A. To improve performance
- B. To reduce cost
- C. To use custom machine types
- D. To reduce cost on premium operating systems

Quiz | Question 2

Answer

What is the main reason customers choose Preemptible or Spot VMs?

- A. To improve performance
- B. To reduce cost
- C. To use custom machine types
- D. To reduce cost on premium operating systems



What is the main reason customers choose Preemptible VMs?

A: To improve performance.

Feedback: Sorry, that's not correct. Preemptible and spot VMs have the same performance as ordinary VMs of the same machine type.

B: To reduce cost.

Feedback: That's correct! The per-hour price of preemptible and spot VMs incorporates a substantial discount.

C: To use custom machine types

Feedback: Sorry, that's not correct. Both Preemptible/spot VMs and regular VMs can have custom configurations.

D: To reduce cost on premium operating systems.

Feedback: Sorry, that's not correct. Preemptible and spot instances don't reduce the cost of premium operating systems and don't change the way that you're billed for the use of those operating systems. If Compute Engine stops a preemptible or spot instance that runs a premium operating system, you are billed for that operating system as if you stopped the instance yourself.

Quiz | Question 3

Question

For which of these interconnect options is a Service Level Agreement available?

- A. Dedicated Interconnect
- B. Standard Network Tier
- C. Direct Peering
- D. Carrier Peering

Quiz | Question 3

Answer

For which of these interconnect options is a Service Level Agreement available?

- A. Dedicated Interconnect
- B. Standard Network Tier
- C. Direct Peering
- D. Carrier Peering



For which of these interconnect options is a Service Level Agreement available?

A: Dedicated Interconnect

Feedback: Correct!

B: Standard Network Tier

Feedback: Review the lecture "Connecting networks to Google VPC."

C: Direct Peering

Feedback: Review the lecture "Connecting networks to Google VPC."

D: Carrier Peering

Feedback: Review the lecture "Connecting networks to Google VPC."

Lab Intro

Getting Started with VPC Networking and Google Compute Engine

In this lab, you create an auto mode VPC network with firewall rules and two VM instances. Then, you explore the connectivity for the VM instances.

