Google Cloud

GCP Fundamentals: Core Infrastructure

Virtual Machines in the Cloud

Google Cloud

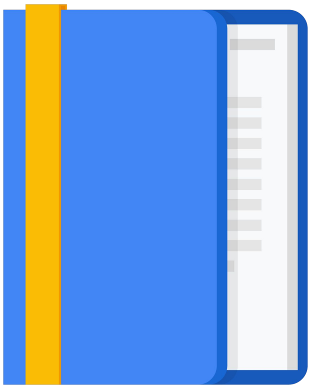
Agenda

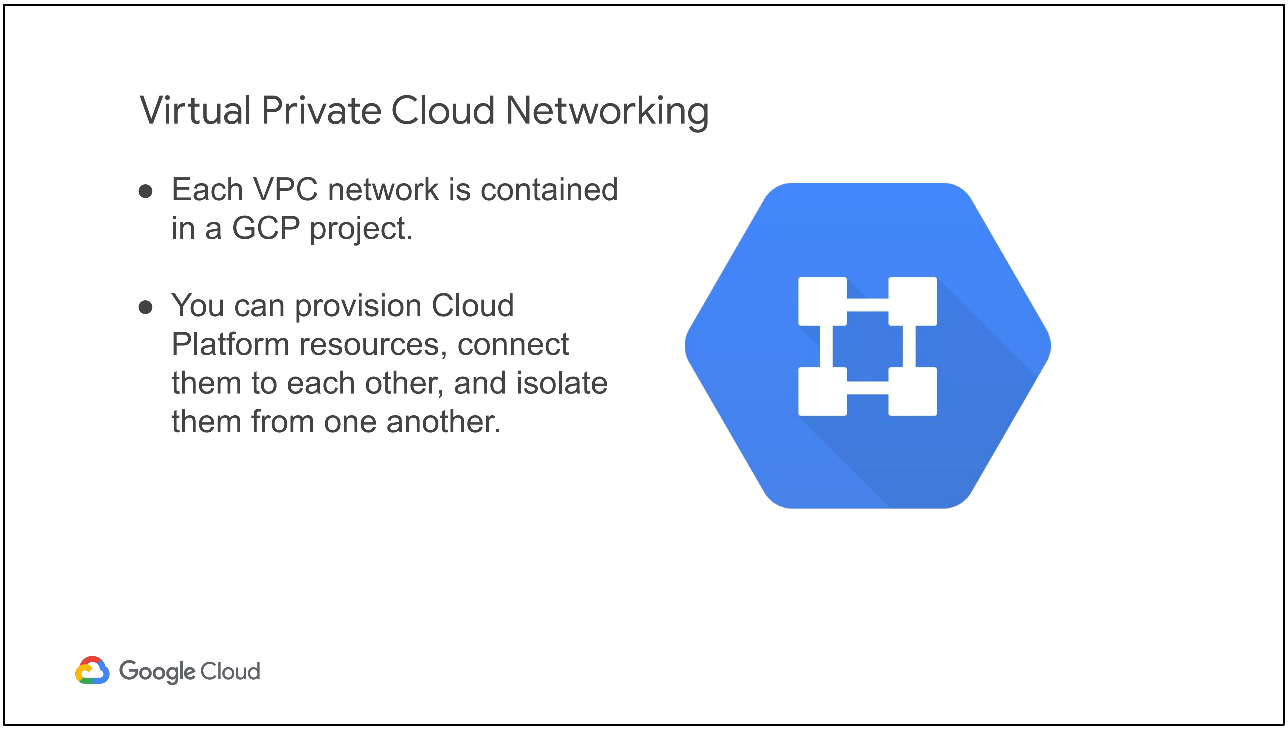
Virtual Private Cloud (VPC) Network

Compute Engine

Important VPC capabilities

Quiz and lab





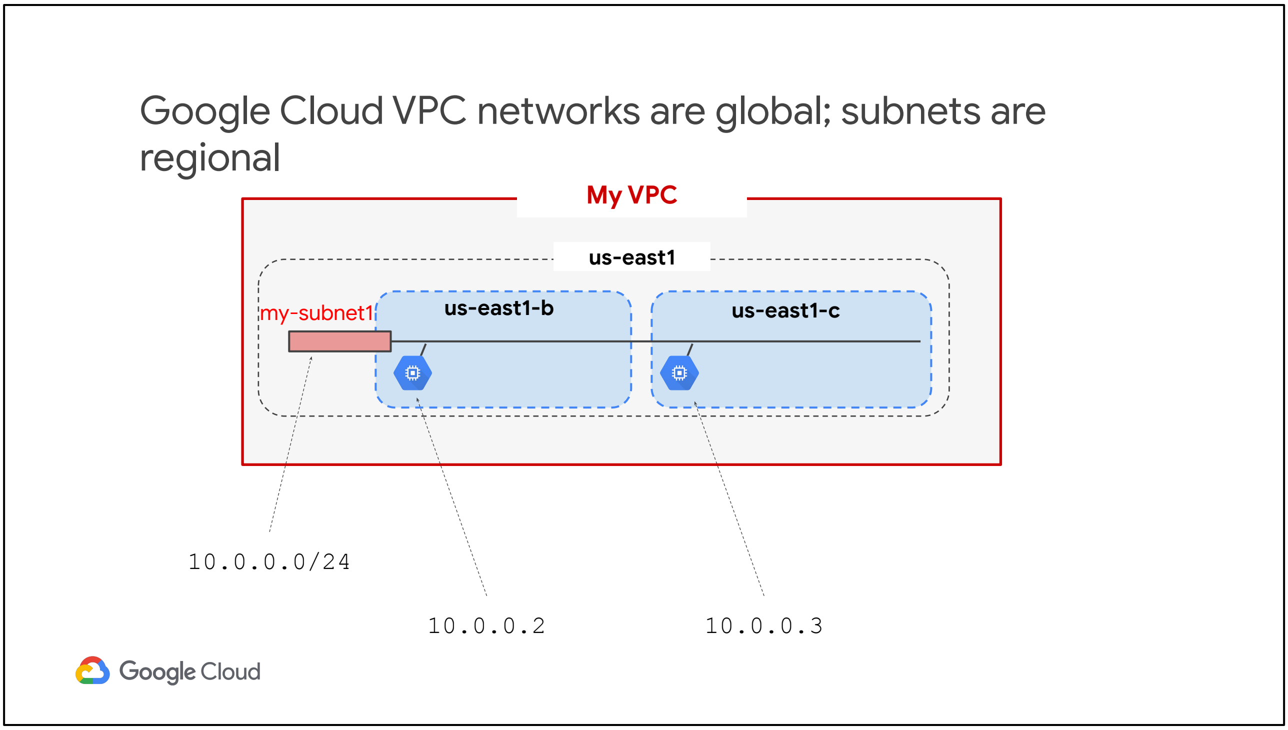
Virtual Private Cloud Networking

- Each VPC network is contained in a GCP project.
- You can provision Cloud Platform resources, connect them to each other, and isolate them from one another.

Google Cloud

Your VPC networks connect your Google Cloud Platform resources to each other and to the internet. You can segment your networks, use firewall rules to restrict access to instances, and create static routes to forward traffic to specific destinations.

Many users get started with GCP is to define their own Virtual Private Cloud inside their first GCP project. Or they can simply choose the default VPC and get started with that.



Google Cloud VPC networks are global; subnets are regional

Google Virtual Private Cloud networks that you define have global scope. They can have subnets in any GCP region worldwide. Subnets can span the zones that make up a region. This architecture makes it easy for you to define your own network layout with global scope. You can also have resources in different zones on the same subnet.

You can dynamically increase the size of a subnet in a custom network by expanding the range of IP addresses allocated to it. Doing that doesn't affect already configured VMs.

In this example, your VPC has one network. So far, it has one subnet defined, in GCP's us-east1 region. Notice that it has two Compute Engine VMs attached to it. They're neighbors on the same subnet even though they are in different zones! You can use this capability to build solutions that are resilient but still have simple network layouts.

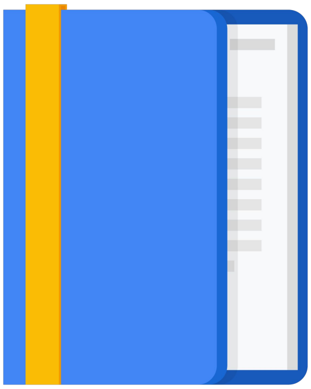
Agenda

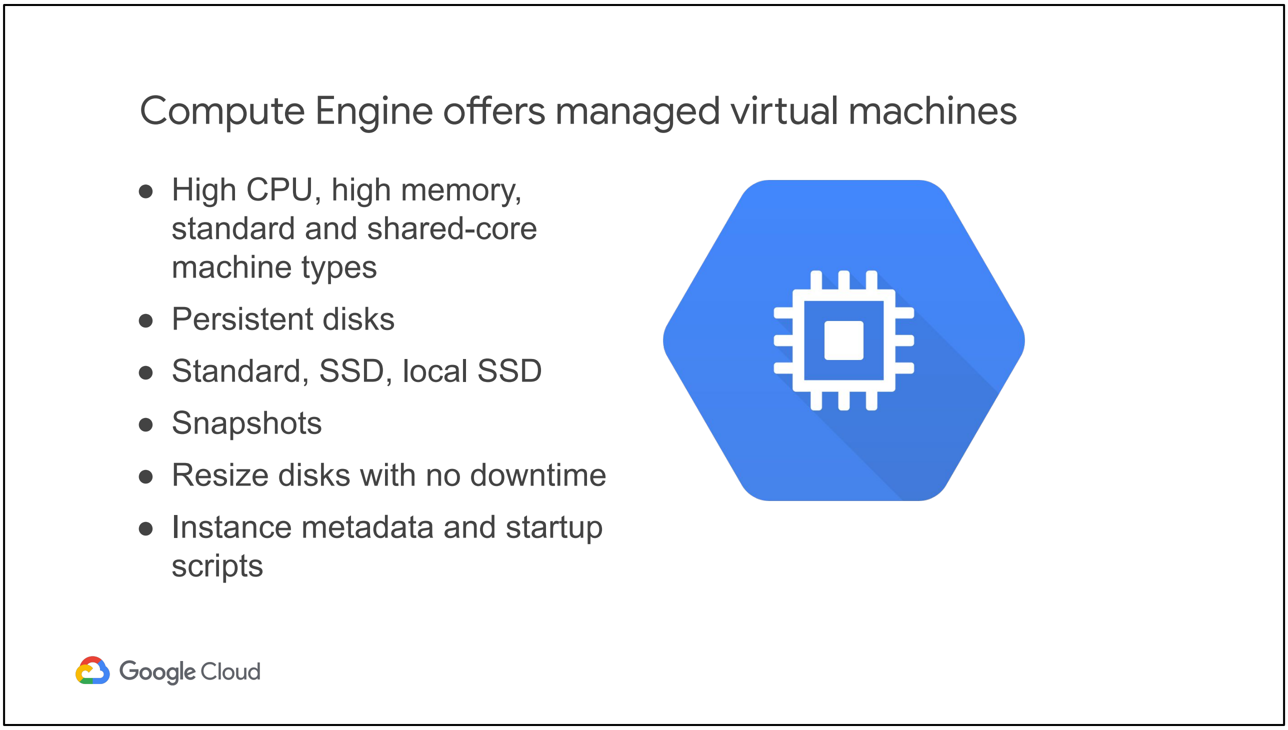
Virtual Private Cloud (VPC) Network

Compute Engine

Important VPC capabilities

Quiz and lab





Compute Engine offers managed virtual machines

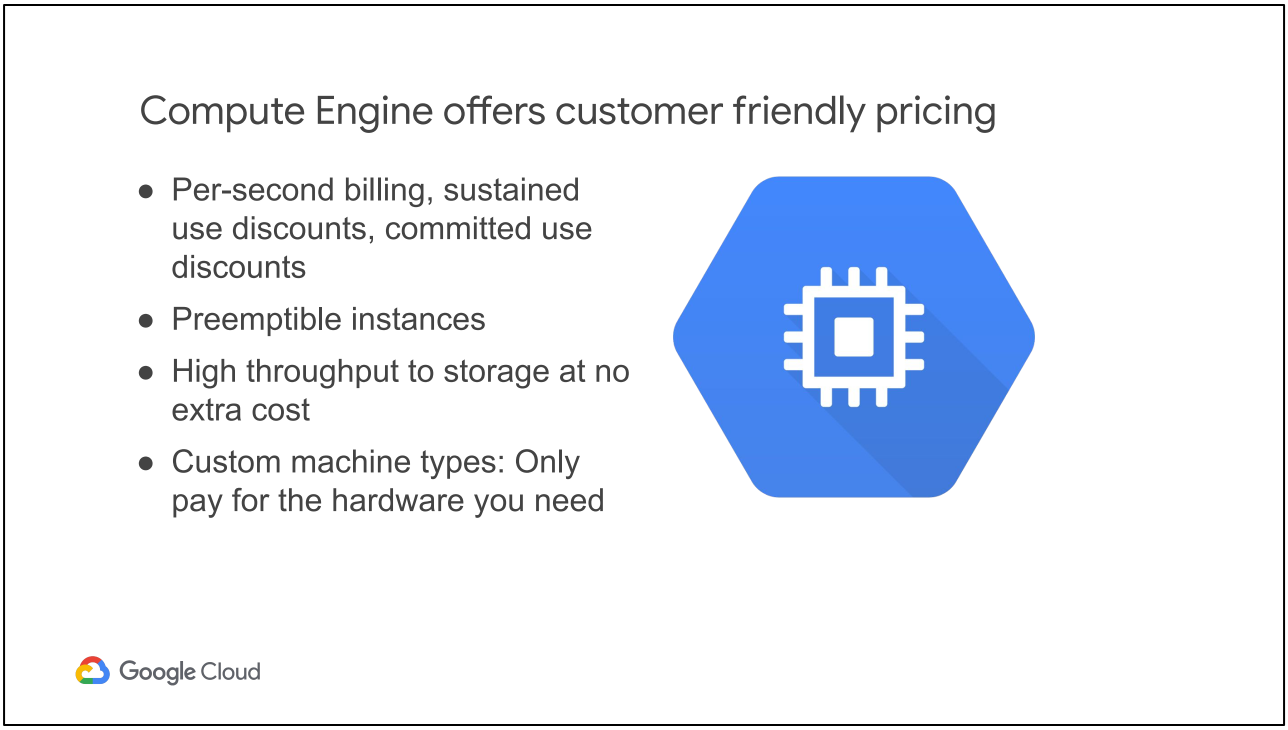
- High CPU, high memory, standard and shared-core machine types
- Persistent disks
- Standard, SSD, local SSD
- Snapshots
- Resize disks with no downtime
- Instance metadata and startup scripts

Google Cloud

Virtual machines have the power and generality of a full-fledged operating system in each. You configure a virtual machine much like you build out a physical server: by specifying its amounts of CPU power and memory, its amounts and types of storage, and its operating system. Compute Engine lets you create and run virtual machines on Google infrastructure. There are no upfront investments, and you can run thousands of virtual CPUs on a system that is designed to be fast and to offer consistent performance.

You can flexibly reconfigure Compute Engine virtual machines. And a VM running on Google's cloud has unmatched worldwide network connectivity.

You can create a virtual machine instance by using the Google Cloud Platform Console or the gcloud command-line tool. A Compute Engine 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.



Compute Engine offers customer friendly pricing

- Per-second billing, sustained use discounts, committed use discounts
- Preemptible instances
- High throughput to storage at no extra cost
- Custom machine types: Only pay for the hardware you need

Google Cloud

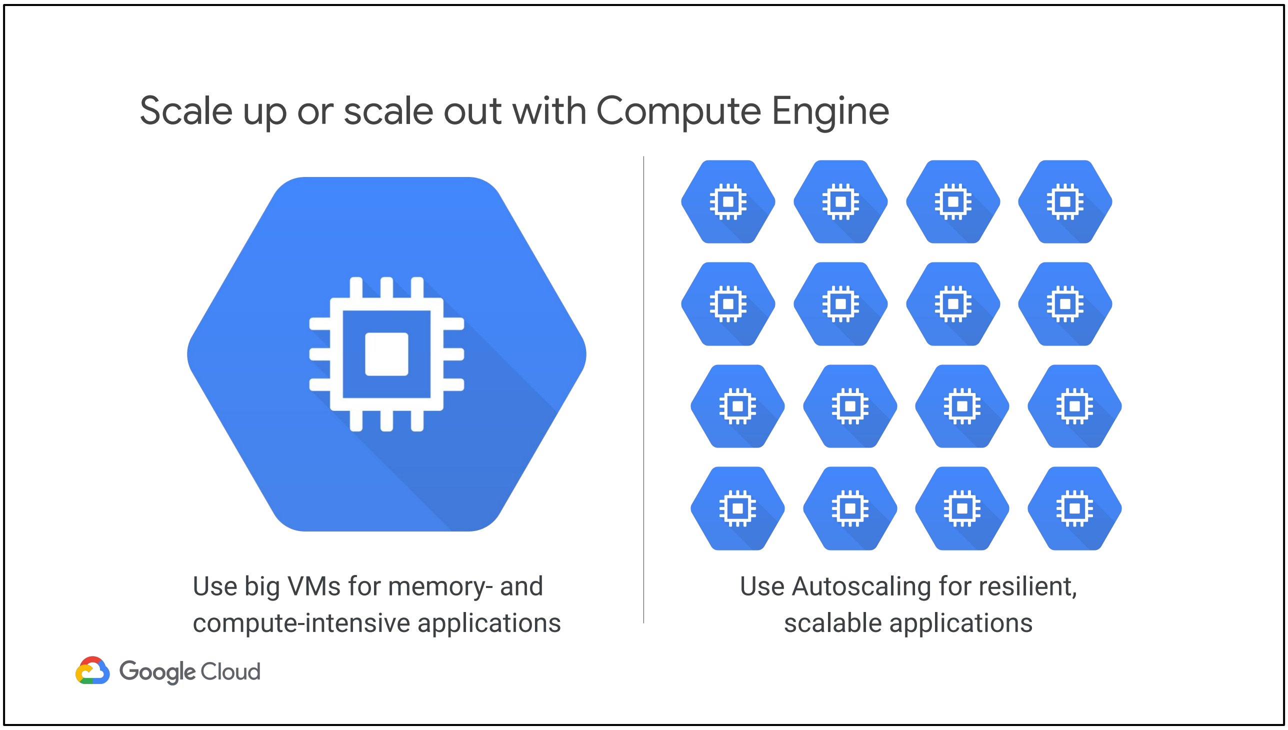
Compute Engine bills by the second for use of virtual machines, with a one-minute minimum. And discounts apply automatically to virtual machines that run for substantial fractions of a month. For each VM that you run for more than 25% of a month, Compute Engine automatically gives you a discount for every incremental minute. You can get up to a 30% net discount for VMs that run the entire month.

Compute Engine offers the ability to purchase committed use contracts in return for deeply discounted prices for VM usage. These discounts are known as committed use discounts. If your workload is stable and predictable, you can purchase a specific amount of vCPUs and memory for up to a 57% discount off of normal prices in return for committing to a usage term of 1 year or 3 years.

Suppose you have a workload that no human being is sitting around waiting to finish. Say, a batch job analyzing a large dataset. You can save money by choosing Preemptible VMs to run the job. A Preemptible VM is different from an ordinary Compute Engine VM in only one respect: you've given Compute Engine permission to terminate it if its resources are needed elsewhere. You can save a lot of money with preemptible VMs, although be sure to make your job able to be stopped and restarted.

You don't have to select a particular option or machine type to get high throughput between your processing and your persistent disks. That’s the default.

You can choose the machine properties of your instances, such as 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.



Scale up or scale out with Compute Engine

Use big VMs for memory- and compute-intensive applications

Use Autoscaling for resilient, scalable applications

Google Cloud

You can make very large VMs in Compute Engine. At the time this deck was produced, the maximum number of virtual CPUs in a VM was zone-dependent and at 96, and the maximum memory size was at 624 GB (6.5 GB per CPU).

You can also use a mega-memory machine type that scales to 1.4 TB memory.

Check the GCP website to see where these maximums are today.

These huge VMs are great for workloads like in-memory databases and CPU-intensive analytics. But most GCP customers start off with scaling out, not up. Compute Engine has a feature called Autoscaling that lets you add and take away VMs from your application based on load metrics. The other part of making that work is balancing the incoming traffic among the VMs. And Google VPC supports several different kinds of load balancing! We'll consider those in the next section.

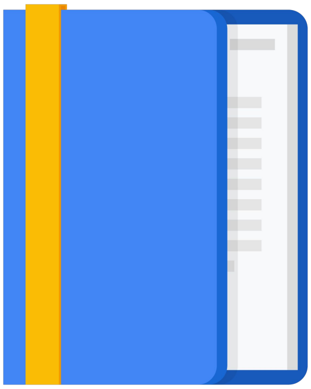
Agenda

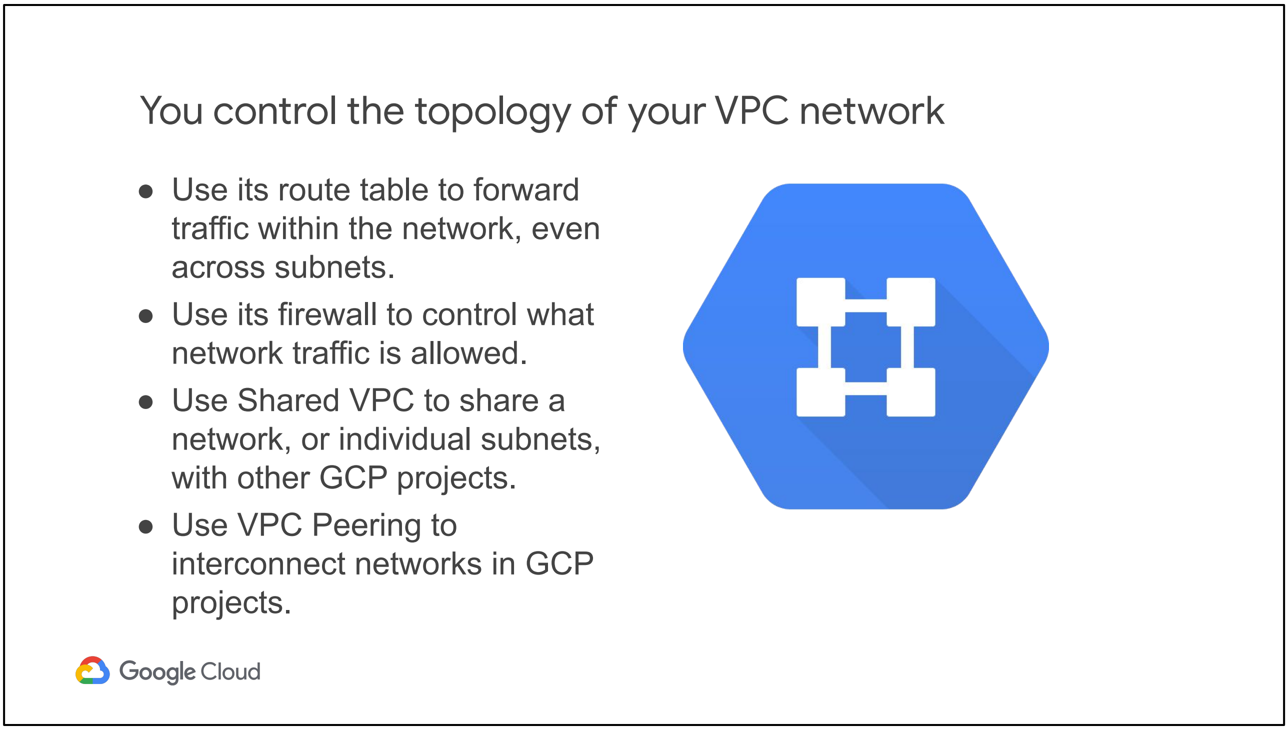
Virtual Private Cloud (VPC) Network

Compute Engine

Important VPC capabilities

Quiz and lab





You control the topology of your VPC network

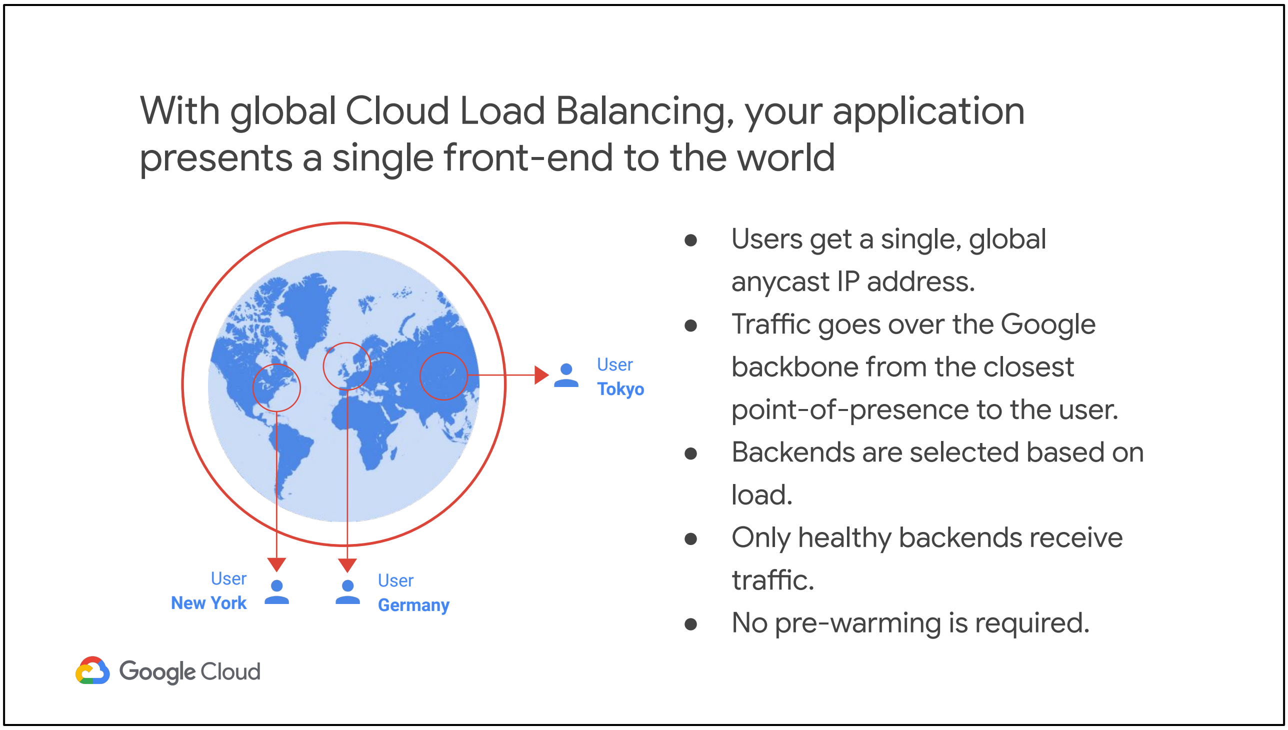
- Use its route table to forward traffic within the network, even across subnets.
- Use its firewall to control what network traffic is allowed.
- Use Shared VPC to share a network, or individual subnets, with other GCP projects.
- Use VPC Peering to interconnect networks in GCP projects.

Google Cloud

Much like physical networks, VPCs have routing tables. These are used to forward traffic from one instance to another instance within the same network, even across subnetworks and even between GCP zones, without requiring an external IP address. VPCs' routing tables are built in; you don't have to provision or manage a router.

Another thing you don't have to provision or manage for GCP: a firewall. VPCs give you a global distributed firewall you can control to restrict access to instances, both incoming and outgoing traffic. You can define firewall rules in terms of metadata 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.

Recall that VPCs belong to GCP projects. But what if your company has several GCP projects, and the VPCs need to talk to each other? If you simply want to establish a peering relationship between two VPCs, so that they can exchange traffic, configure VPC Peering does. On the other hand, if you want to use the full power of IAM to control who and what in one project can interact with a VPC in another, configure Shared VPC.



With global Cloud Load Balancing, your application presents a single front-end to the world

- Users get a single, global anycast IP address.
- Traffic goes over the Google backbone from the closest point-of-presence to the user.
- Backends are selected based on load.
- Only healthy backends receive traffic.
- No pre-warming is required.

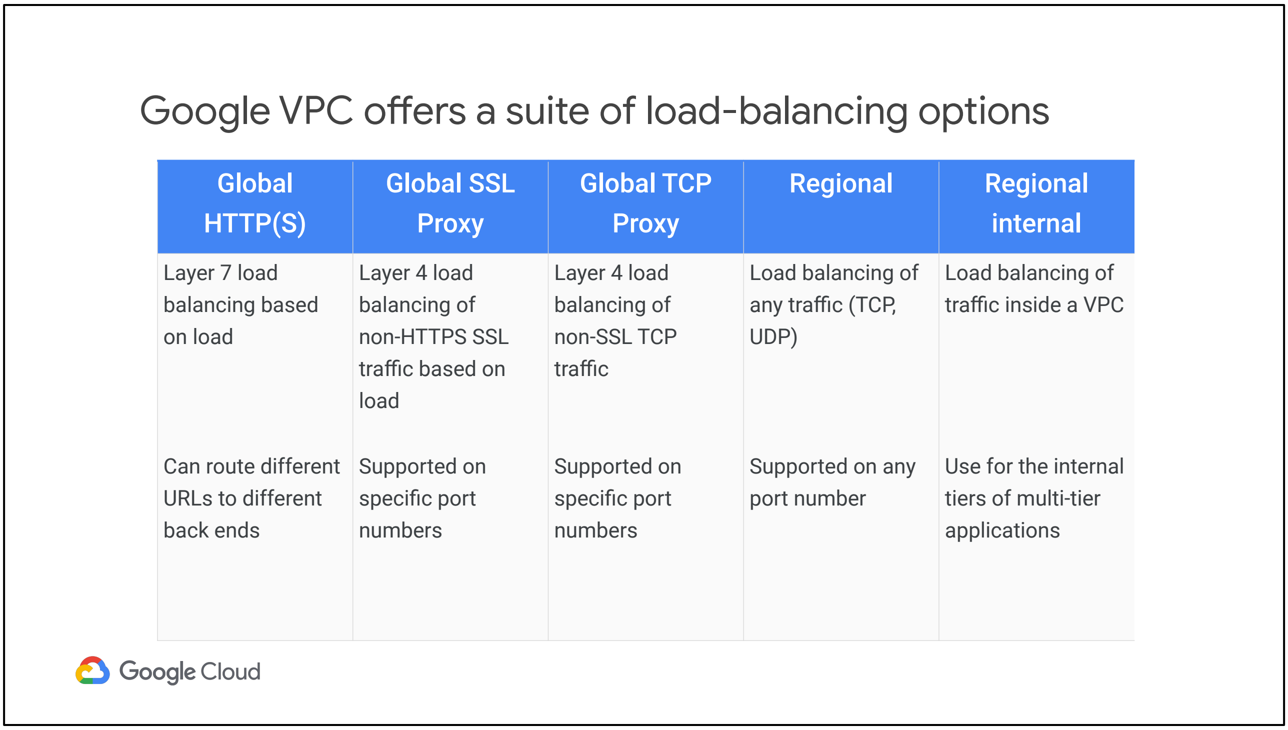
Google Cloud

A few slides back, we talked about how virtual machines can autoscale to respond to changing load. But how do your customers get to your application when it might be provided by four VMs one moment and forty VMs at another? Cloud Load Balancing is the answer.

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 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.

With Cloud Load Balancing, a single anycast IP front-ends all your backend instances in regions around the world. It 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.



Google VPC offers a suite of load-balancing options

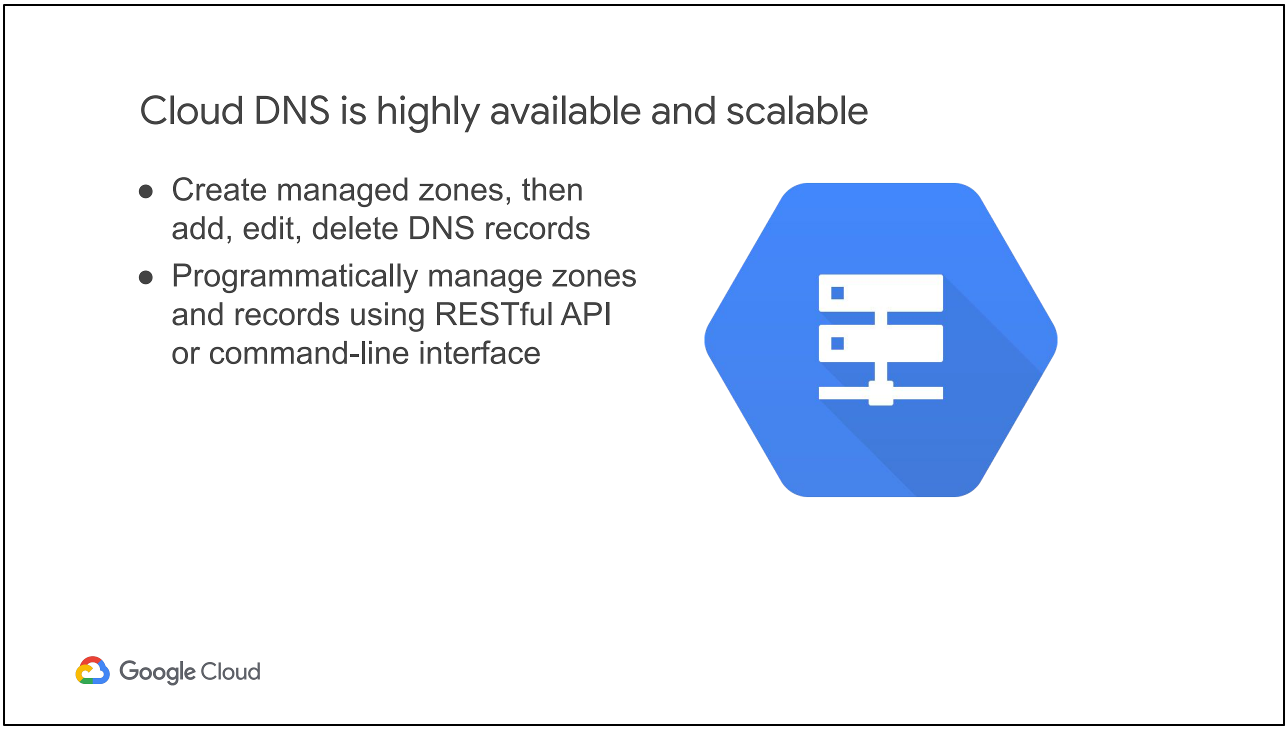
Global HTTP(S)	Global SSL Proxy	Global TCP Proxy	Regional	Regional internal
Layer 7 load balancing based on load	Layer 4 load balancing of non-HTTPS SSL traffic based on load	Layer 4 load balancing of non-SSL TCP traffic	Load balancing of any traffic (TCP, UDP)	Load balancing of traffic inside a VPC
Can route different URLs to different back ends	Supported on specific port numbers	Supported on specific port numbers	Supported on any port number	Use for the internal tiers of multi-tier applications

Google Cloud

If you need cross-regional load balancing for a Web application, use 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 GCP 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 Internal load balancer. It accepts traffic on a GCP internal IP address and load balances it across Compute Engine VMs.



Cloud DNS is highly available and scalable

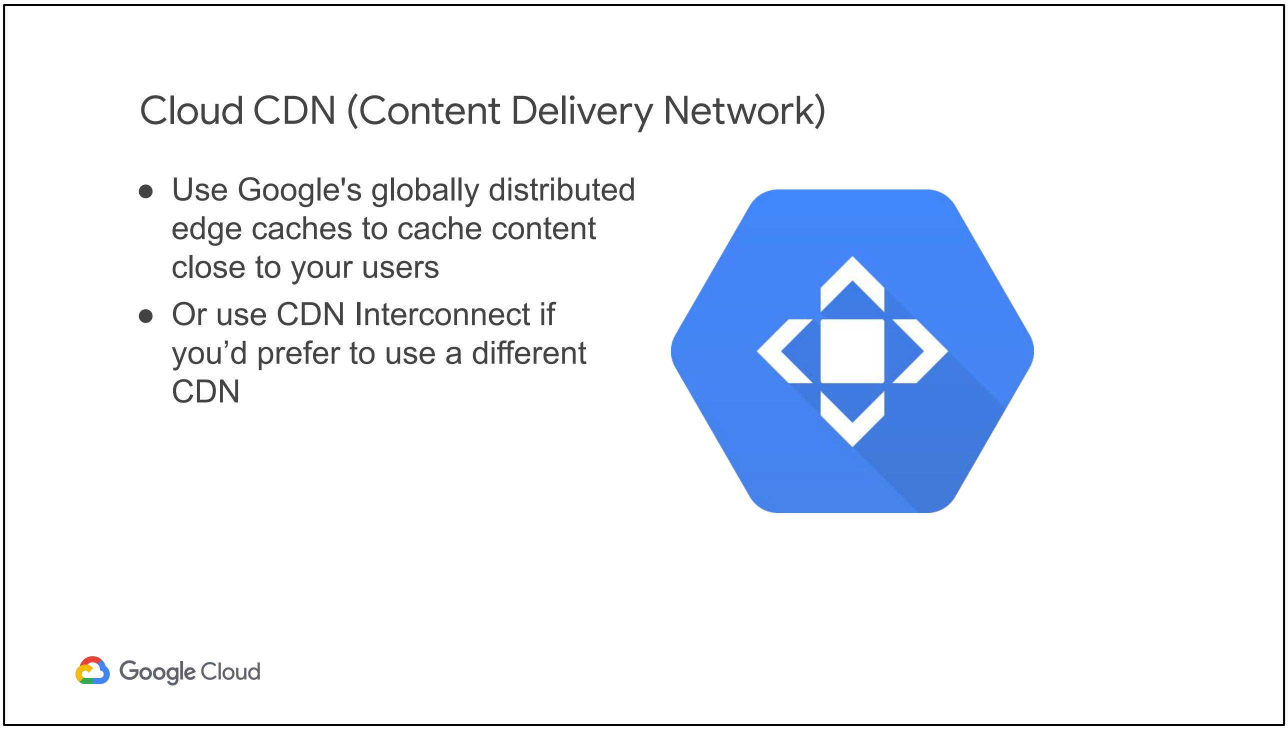
- Create managed zones, then add, edit, delete DNS records
- Programmatically manage zones and records using RESTful API or command-line interface

Google Cloud

One of the most famous Google services that people don’t pay for 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 would imagine, Google has a highly developed DNS infrastructure. It makes 8.8.8.8 available so that everybody can take advantage of it.

But what about the Internet hostnames and addresses of applications you build in GCP?
GCP offers Cloud DNS to help the world find them. It’s a managed DNS service running 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 GCP Console, the command-line interface, or the API.



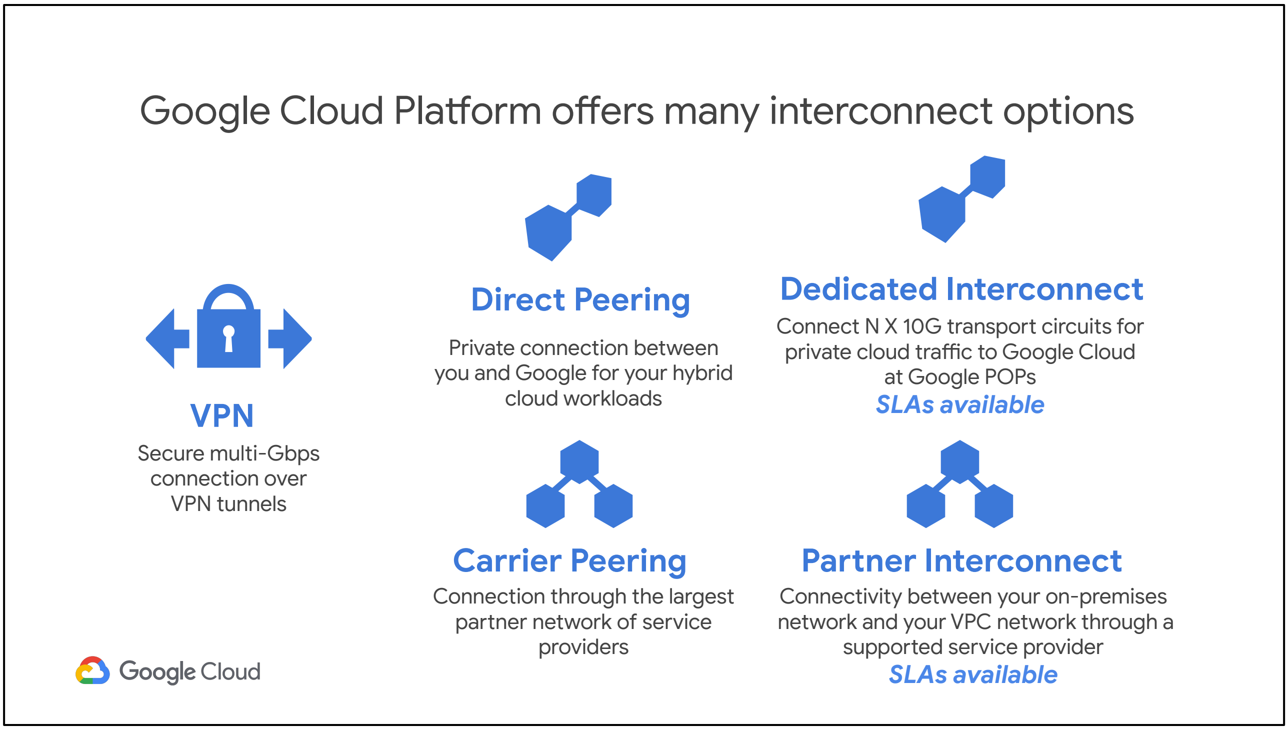
Cloud CDN (Content Delivery Network)

- Use Google's globally distributed edge caches to cache content close to your users
- Or use CDN Interconnect if you’d prefer to use a different CDN

Google Cloud

Google has a global system of edge caches. You can use this system to accelerate content delivery in your application using Google Cloud CDN. Your customers will experience lower network latency, the origins of your content will experience reduced load, and you can save money too. Once you've set up HTTP(S) Load Balancing, simply enable Cloud CDN with a single checkbox.

There are lots of other CDNs out there, of course. If you are already using one, chances are, it is a part of GCP’s CDN Interconnect partner program, and you can continue to use it.



Google Cloud Platform offers many interconnect options

VPN

Secure multi-Gbps connection over VPN tunnels

Direct Peering

Private connection between you and Google for your hybrid cloud workloads

Dedicated Interconnect

Connect N X 10G transport circuits for private cloud traffic to Google Cloud at Google POPs

SLAs available

Carrier Peering

Connection through the largest partner network of service providers

Partner Interconnect

Connectivity between your on-premises network and your VPC network through a supported service provider

SLAs available

Google Cloud

Lots of GCP customers want to interconnect their other networks to their Google VPCs. such as on-premises networks or their networks in other clouds. There are many good choices.

Many customers start with a Virtual Private Network connection over the Internet, using the IPsec protocol. To make that dynamic, they use a GCP feature called Cloud Router. Cloud Router lets your other networks and your Google VPC exchange route information over the VPN using the Border Gateway Protocol. For instance, if you add a new subnet to your Google VPC, your on-premises network will automatically get routes to it.

But some customers don't want to use the Internet, either because of security concerns or because they need more reliable bandwidth. They can consider peering with Google using Direct Peering. Peering means putting a router in the same public datacenter as a Google point of presence and exchanging traffic. Google has more than 100 points of presence around the world. Customers who aren't already in a point of presence can contract with a partner in the Carrier Peering program to get connected.

One downside of peering, though, is that it isn't covered by a Google Service Level Agreement. Customers who want the highest uptimes for their interconnection with

Google should use Dedicated Interconnect, in which customers get 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. These connections can be backed up by a VPN for even greater reliability.

Partner Interconnect provides connectivity between your on-premises network and your VPC network through a supported service provider. A Partner Interconnect connection is useful if your data center is in a physical location that can't reach a Dedicated Interconnect colocation facility or if your data needs don't warrant an entire 10 Gbps connection. Depending on your availability needs, you can configure Partner Interconnect 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.

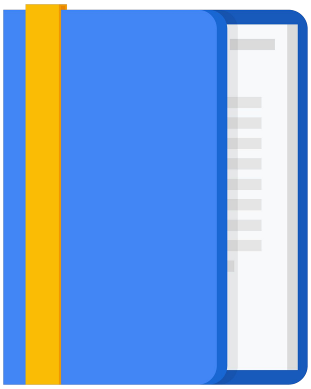
Agenda

Virtual Private Cloud (VPC) Network

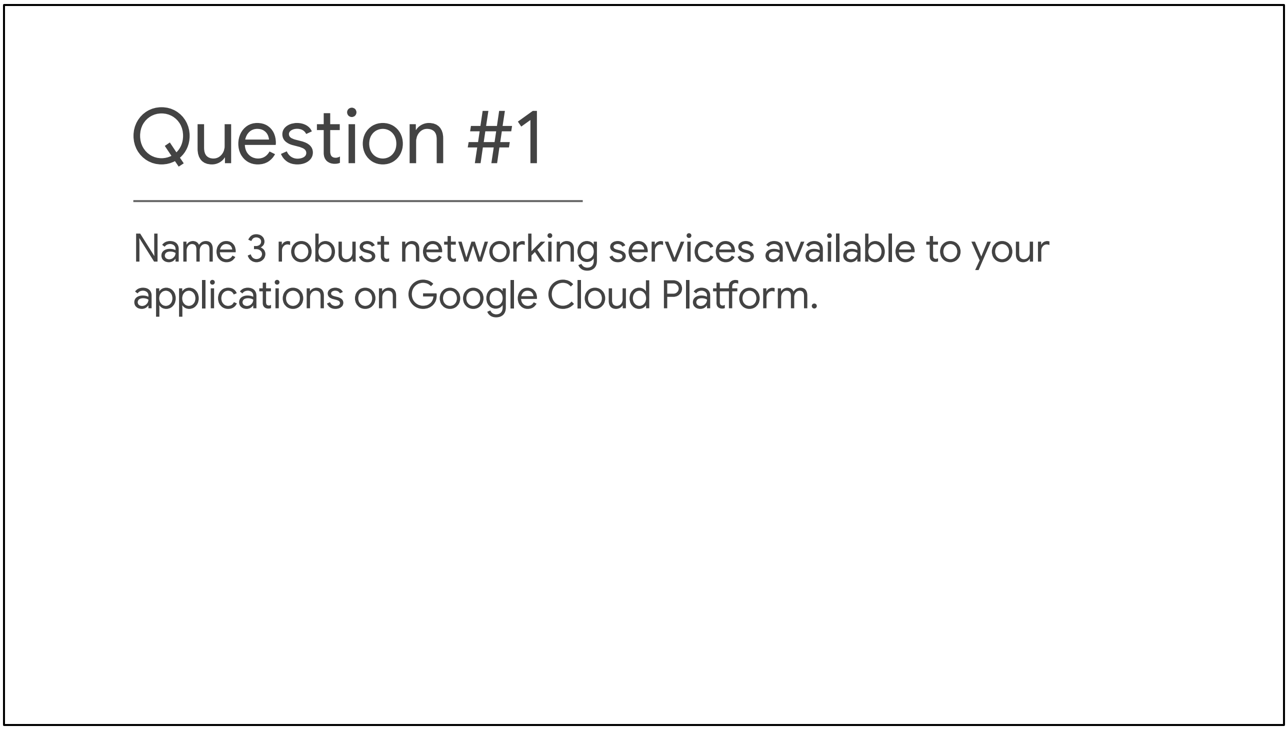
Compute Engine

Important VPC capabilities

Quiz and lab

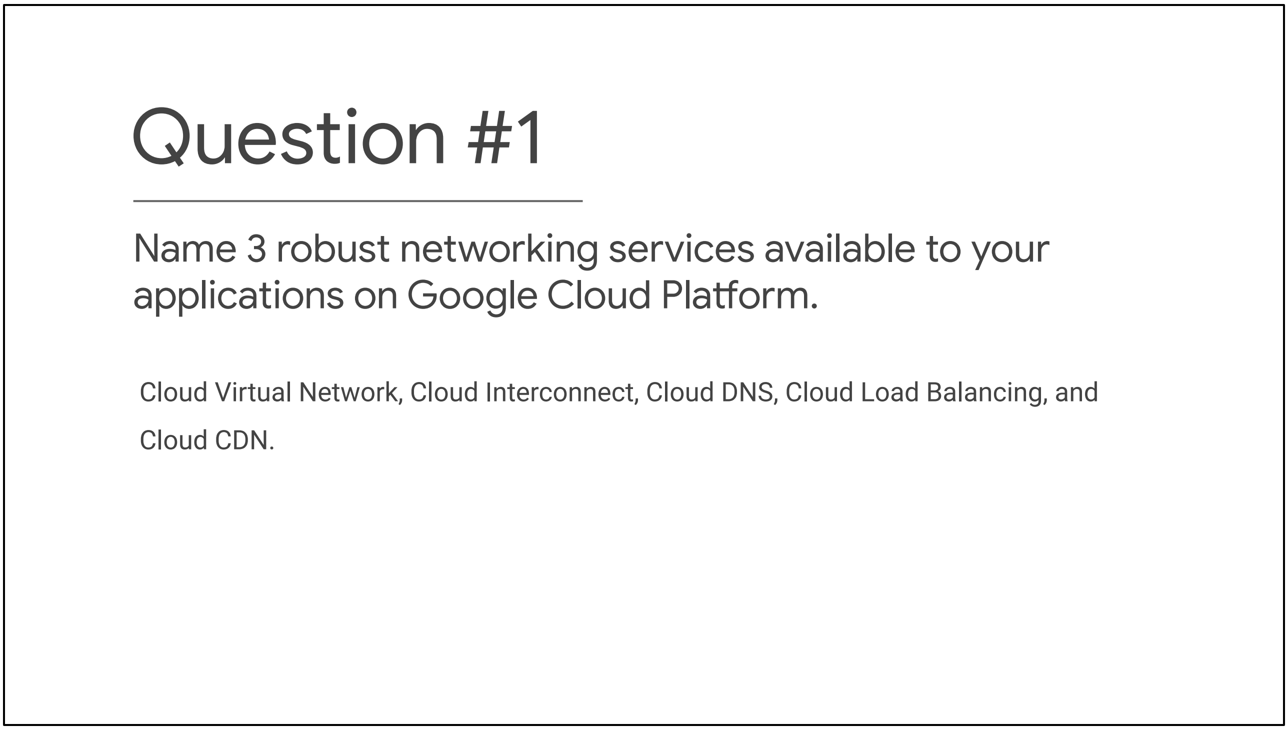


Google Cloud



Question #1

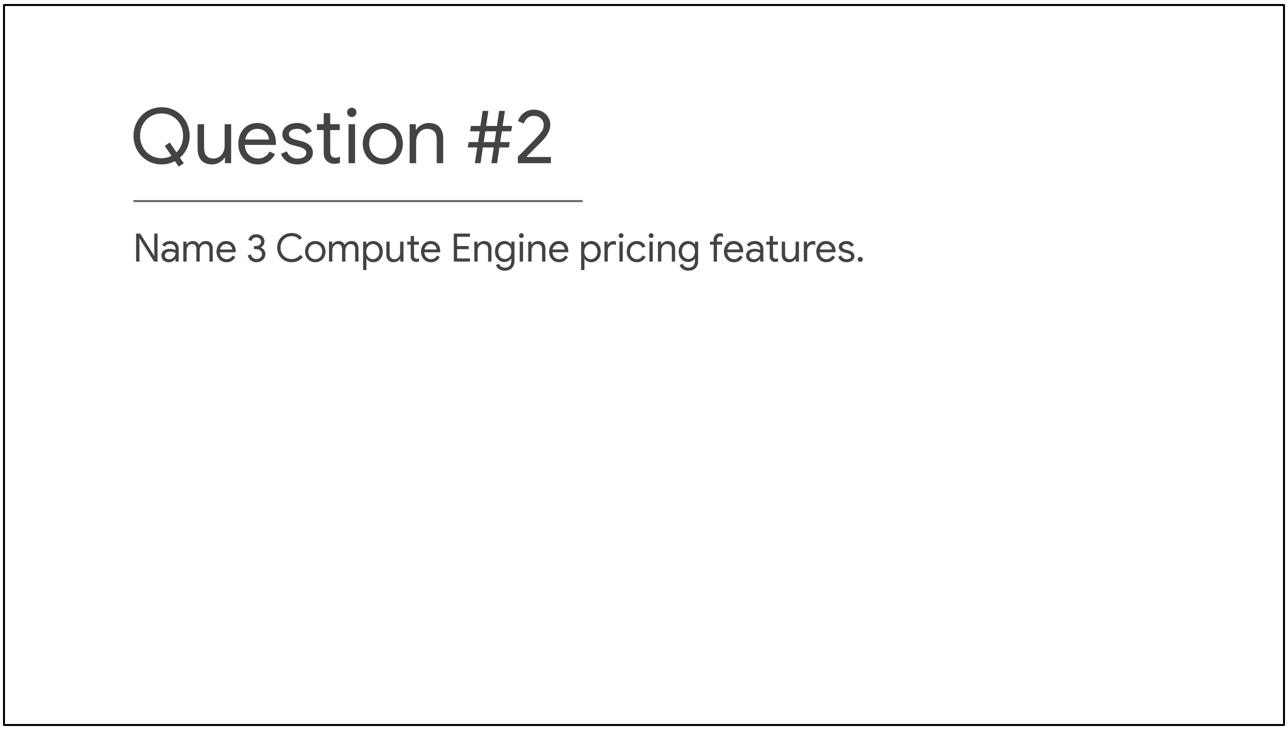
Name 3 robust networking services available to your applications on Google Cloud Platform.



Question #1

Name 3 robust networking services available to your applications on Google Cloud Platform.

Cloud Virtual Network, Cloud Interconnect, Cloud DNS, Cloud Load Balancing, and Cloud CDN.



Question #2

Name 3 Compute Engine pricing features.

Question #2

Name 3 Compute Engine pricing features.

Per-second billing, custom machine types, preemptible instances.

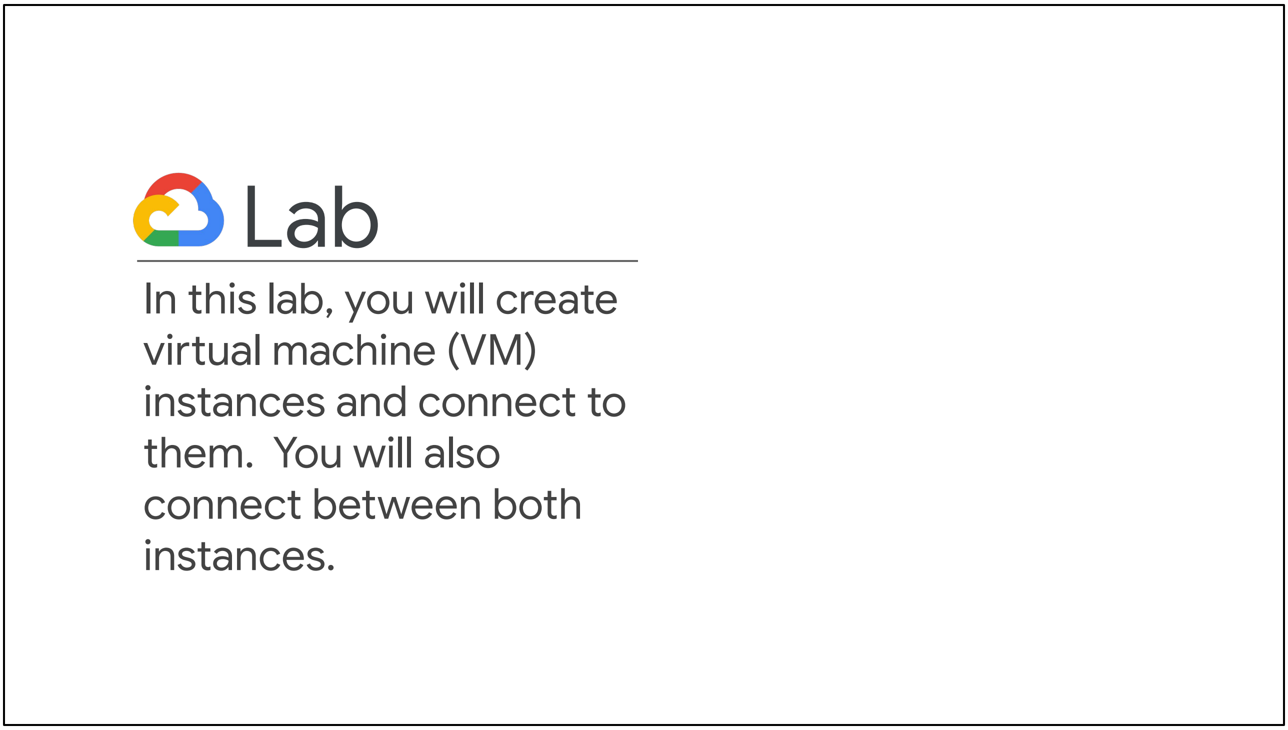
Question #3

True or False: Google Cloud Load Balancing lets you balance HTTP traffic across multiple Compute Engine regions.

Question #3

True or False: Google Cloud Load Balancing lets you balance HTTP traffic across multiple Compute Engine regions.

True



Lab

In this lab, you will create virtual machine (VM) instances and connect to them. You will also connect between both instances.

Lab Objectives

- Create a Compute Engine virtual machine using the Google Cloud Platform Console
- Create a Compute Engine virtual machine using the gcloud command-line interface
- Connect between the two instances

More resources

Google Compute Engine https://cloud.google.com/compute/docs/

Google Cloud Platform VPC https://cloud.google.com/compute/docs/vpc/

Google Cloud Stackdriver https://cloud.google.com/stackdriver/docs/

`gcloud` tool guide https://cloud.google.com/source-repositories/docs/