Google Cloud

Google Cloud and Hybrid Network Architecture

Learning objectives

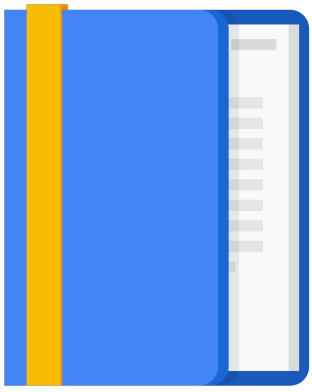
- Design VPC networks to optimize for cost, security, and performance.
- Configure global and regional load balancers to provide access to services.
- Leverage Cloud CDN to provide lower latency and decrease network egress.
- Evaluate network architecture using the Network Intelligence Center.
- Connect networks using peering, VPNs and Cloud Interconnect.

Google Cloud

This module discusses the Google Cloud network architecture. It includes load balancing and the range of load balancing options—global, regional, internal, external—and the traffic type. The connection of on-premises networks with Google Cloud networks is also discussed with the various interconnection options. The different options and how they impact performance, security, and cost and how CDN can be leveraged to manage costs are examined. Managing and diagnosing networks through the Network Intelligence Center is also covered.

Agenda

- Designing Google Cloud Networks
- Design Activity #8
- Connecting Networks
- Design Activity #9



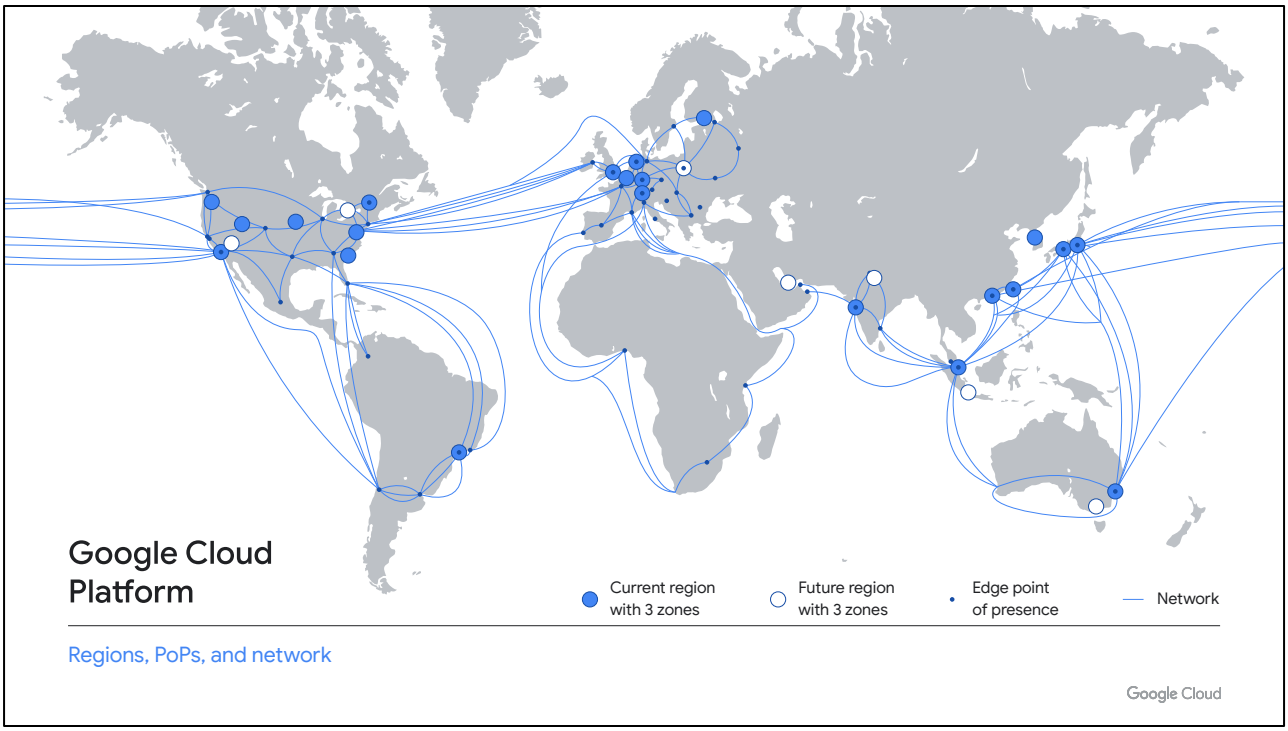
Google Cloud

Google runs a worldwide network that connects regions all over the world

Design your networks based on location, number of users, scalability, fault tolerance, and other service requirements.

Google Cloud

Google Cloud's footprint spans 61 zones and over 130 points of presence across more than 35 countries. High bandwidth connectivity via subsea cables provides unrivalled network performance. Google Cloud customers can use this high bandwidth infrastructure for their own cloud networking needs.



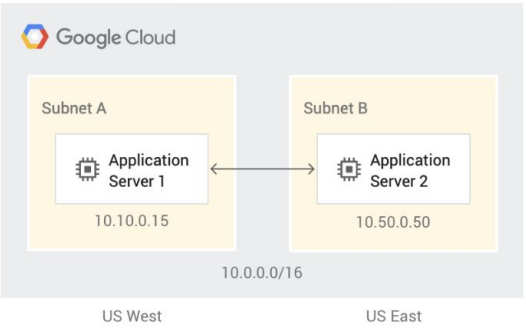
The number of regions and zones is continually increasing. An up-to-date view is here: https://cloud.google.com/about/locations/.

The network and points of presence are here: https://cloud.google.com/about/locations/?tab=network.

The edge points of presence are the locations where Google networks are connected with internet service providers to allow users to connect. The Google Cloud network strongly distinguishes Google from other cloud service providers. The points of presence allow Google to provide very low latency network performance.

In Google Cloud, VPC networks are global

- When creating networks, create subnets for the regions you want to operate in.
- Resources across regions can reach each other without any added interconnect.
- If you are a global company, choose regions around the world.
- If your users are close together, choose the region closest to them plus a backup region.
- A project can have multiple networks.

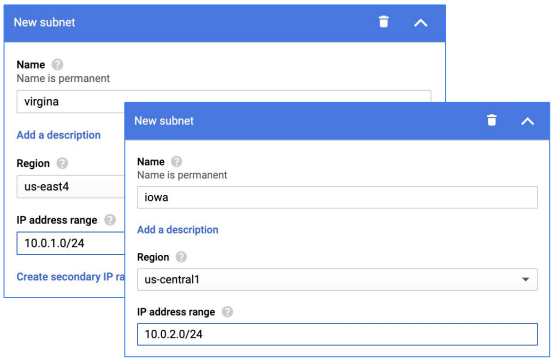


Google Cloud

VPCs are software-designed versions of physical networks; VPC has global scope and so spans regions. When creating a VPC, you can create subnetworks for the regions you want to operate in. Automatic subnet creation mode will create a subnetwork in each region by default. A VPC is associated with a project or an organization. Projects can have multiple VPCs.

When creating custom subnets, specify the region and the internal IP address range

- IP address ranges cannot overlap.
- Machines in the same VPC can communicate via their internal IP address regardless of the subnet region.
- Subnets don't need to be derived from a single CIDR block.
- Subnets are expandable without down time.
- IP Aliasing or Secondary range can be set on the subnet.



Google Cloud

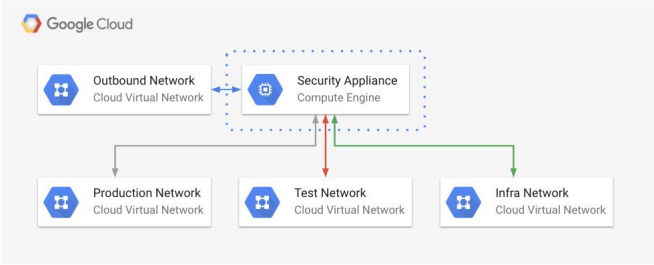
Subnets are regional resources, and each subnet has a range of IP addresses. A network must have at least one subnetwork before it can be used. When a subnet is created, the primary IP address range must be specified. It is possible to select any private CIDR block for the primary IP address range of the subnet. These addresses can be used for VM primary internal IP addresses, VM alias IP addresses, and the IP addresses of internal load balancers. Alias IP addresses are a feature that allows a range of IP addresses to be assigned to a VM's network interfaces. The use case is if multiple services are running on a VM, and each service needs a different IP address. Secondary IP address ranges can be defined, which are separate CIDR blocks and are used for alias IP addresses.

When CIDR blocks are assigned, subnetworks do not need to form a contiguous block, although automode VPC networks create a subnet in each region automatically with contiguous CIDR blocks.

Machines on the same VPC but different subnetworks can communicate using their internal IP addresses.

A single VM can have multiple network interfaces connecting to different networks

- Each network must have a subnet in the region the VM is created in.
- Each interface must be attached to a different VPC.
- Maximum of 8 interfaces per VM.



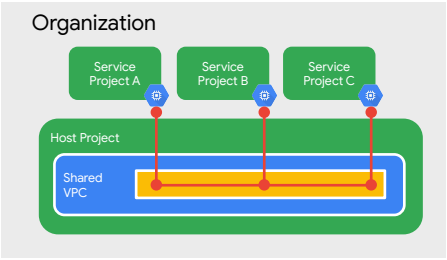
Google Cloud

By default, every instance in a VPC network has a single default network interface. It is possible to add more interfaces, up to a maximum of 8. Each interface must be connected to a different VPC network. Network interfaces can only be added at instance creation and can only be removed by instance deletion.

A Shared VPC is created in one project, but can be shared and used by other projects

Requires an organization

- Create the VPC in the host project.
- Shared VPC admin shares the VPC with other service projects.



Google Cloud

Allows centralized control over network configuration

- Network admins configure subnets, firewall rules, routes, etc.
- Remove network admin rights from developers.
- Developers focus on machine creation and configuration in the shared network.
- Disable the creation of the default network using an organizational policy.

Shared VPC allows an organization to connect resources from multiple projects to a common VPC network so that they can communicate with each other securely and efficiently using internal IPs from that network. When you use Shared VPC, you designate a project as a host project and attach one or more other service projects to it. The VPC networks in the host project are called Shared VPC networks. Eligible resources from service projects can use subnets in the Shared VPC network. Eligible resources include Compute Engine resources, GKE clusters, and App Engine flexible instances.

More details of eligible resources can be found here: https://cloud.google.com/vpc/docs/shared-vpc#resources_that_can_be_attached_to_shared_vpc_networks_from_a_service_project

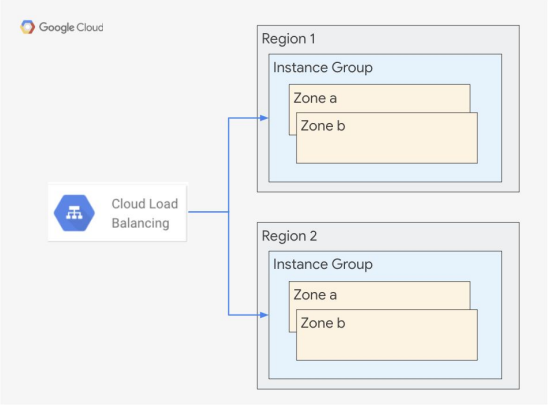
Shared VPC lets organization administrators delegate administrative responsibilities, such as creating and managing instances, to Service Project Admins while maintaining centralized control over network resources like subnets, routes, and firewalls. This model allows organizations to do the following:

1. Implement the security best practice of least privilege for network admin, auditing, and access control. Shared VPC admins delegate admin tasks to admins in the shared network without allowing service project admins to make network-affecting changes. They can only create and manage instances that use the shared VPC.
2. Apply and enforce consistent access control policies at the network level for

1. multiple service projects.

Use a global load balancer to provide access to services deployed in multiple regions

- Global load balancing supported by HTTP load balancer and TCP and SSL proxies.
- HTTP load balancer routes requests to the region closest to the user.
  - Uses a global, anycast IP address.



Google Cloud

To decide which load balancer best suits your implementation of Google Cloud, consider the following aspects of Cloud Load Balancing:

- Global versus regional load balancing
- External versus internal load balancing
- Traffic type

A global load balancer should be used when backends are distributed across multiple regions, and access to the content/application is to be provided with a single anycast IP address.

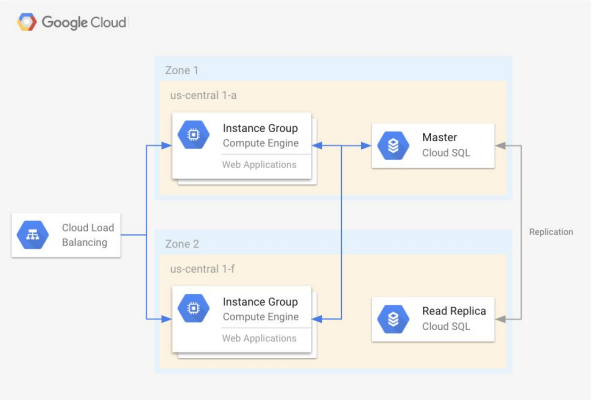
There are three global load balancers described below. These are global for premium tier networks and regional in standard tier.

- **HTTP(S)** balances HTTP(S) load across a set of backend instances. It balances requests among backends in multiple regions by directing requests to the region closest to the user. If that region is at capacity, the load balancer delivers requests to the next closest region that has capacity.
- **SSL proxy** terminates SSL/TLS connections at the load balancing layer and then balances connections across instances. SSL proxy is designed for non-HTTP(S) traffic. As with HTTP(S) load balancing, the requests are sent to the region closest to the user that has capacity.
- **TCP Proxy** terminates TCP sessions at the load balancer and forwards traffic to backend servers. It is intended for non-HTTP traffic. For proxied SSL traffic, you should use SSL proxy. Requests are load balanced to the region closest

- to the user with capacity.

Use a regional load balancer to provide access to services deployed in a single region

- Supported by HTTP, TCP, and UDP load balancers.
- Can have a public or private IP address.
- Can use any TCP or UDP port.



Google Cloud

Regional load balancing is used when backends are in one region and IPv4 termination is required.

HTTP regional load balancing is for internal traffic only (unless using standard tier network when HTTP(S) load balancers are regional by default). Internal traffic does not traverse the public internet.

All backends must be in the same VPC network and the same region as the backend service. The backend service must also be in the same region and VPC network as the forwarding rule.

TCP/UDP load balancing is available for internal and external traffic. Internal traffic is only supported for premium tier networks.

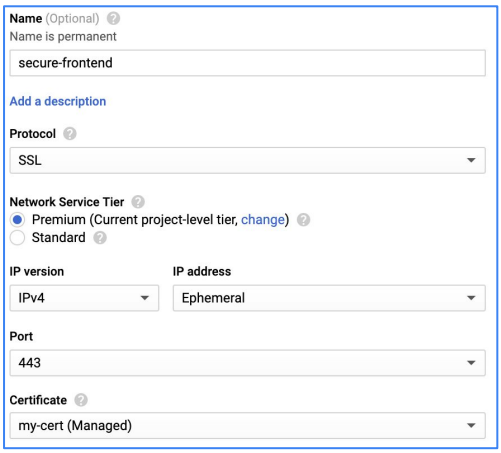
Network TCP/UDP load balancing
Network Load Balancing enables you to load balance traffic on your systems based on incoming IP protocol data, including address, port, and protocol type. It is a regional, non-proxied load balancing system. The use case for Network Load Balancing is for UDP traffic and for TCP and SSL traffic on ports that are not supported by the SSL Proxy and TCP Proxy load balancers. A Network load balancer is a pass-through load balancer that does not proxy connections from clients.

Internal TCP/UDP load balancing
Internal TCP/UDP Load Balancing enables you to load balance TCP/UDP traffic

behind a private IP address that is accessible only to internal virtual machine instances. The use case for internal TCP/UDP load balancing is to configure an internal IP address to act as the frontend to private backend instances. Internal TCP/UDP Load Balancing supports regional managed instance groups, enabling auto scaling across a region and protecting services from zonal failures.

If your load balancers have public IPs, secure them using SSL

- Supported by HTTP and TCP load balancers
- Self-managed and Google-managed SSL certificates



Google Cloud

With public IPs, traffic will be traversing the internet, so it is a best practice to use SSL for both HTTP and TCP load balancers. When configuring an HTTP load balancer, HTTP is the default protocol in Cloud Console for the frontend configuration, and HTTPS requires explicit selection together with the certificate. For a TCP load balancer, SSL is the default in the frontend configuration.

There are two types of certificates supported: self-managed and Google-managed certificates. Details on managing certificates is here: https://cloud.google.com/load-balancing/docs/ssl-certificates

For lower-latency and decreased egress cost leverage Cloud CDN

- Can be enabled when configuring the HTTP global load balancer.
- Caches static content worldwide using Google Cloud edge-caching locations.
- Cache static data from web servers in Compute Engine instances, GKE pods, or Cloud Storage buckets.

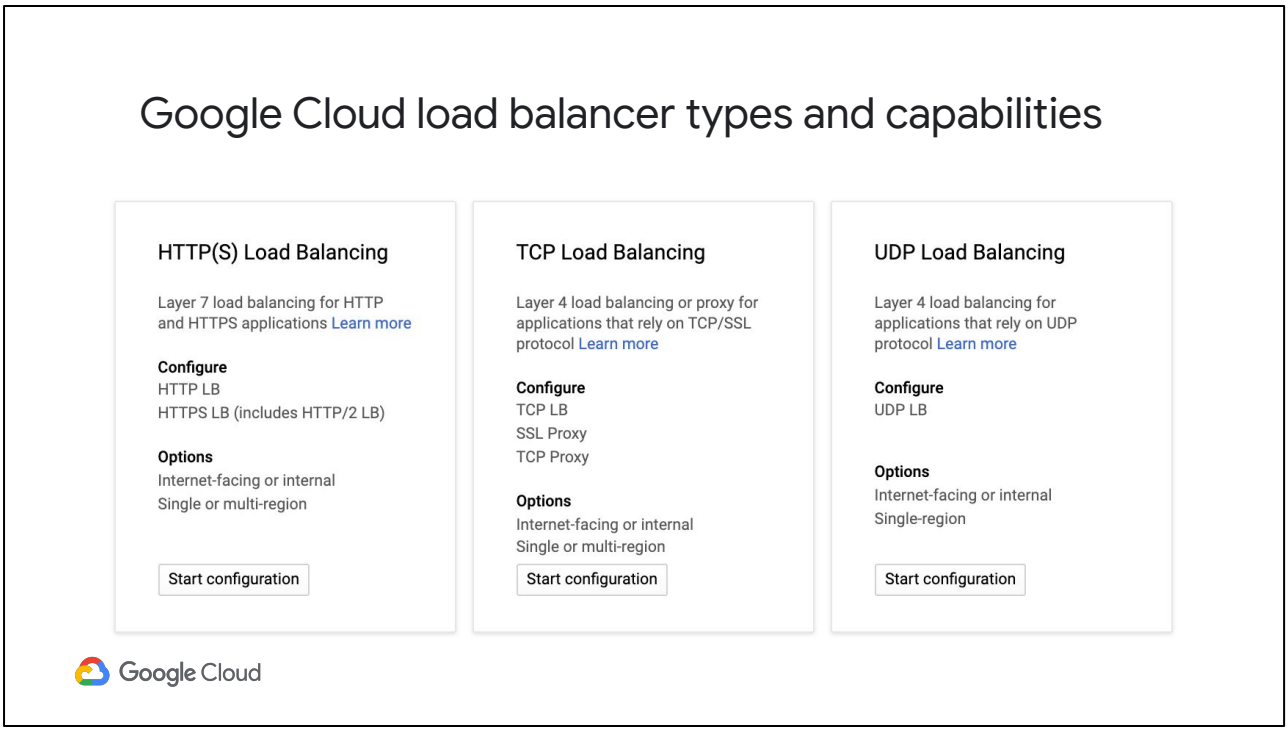
Google Cloud

Cloud CDN uses the globally distributed edge points of presence to speed content delivery for content served. Content sources can be Compute Engine, GKE pods, or Cloud Storage.

Some of the advantages of using Cloud CDN are better user experience through lower network latency and a reduction in serving costs. Cloud CDN is used with global HTTP(S) load balancers.

At a high level, Cloud CDN works as follows:

- When a request for content is made to an HTTP(S) load balancer, the request arrives at a Google Front End (GFE) located at a point of presence as close as possible to the user.
- Assuming that the backend has Cloud CDN configured, then the GFE looks in the Cloud CDN cache for a response to the request. If the GFE finds a cached response, the GFE sends the cached response to the requestor.
- Otherwise, if the GFE can't find a cached response for the request, the GFE makes a request to the appropriate backend (the origin server). If the response to this request is cacheable, the GFE stores the response in the Cloud CDN cache so that the cache can be used for subsequent requests.
- To use Cloud CDN, the load balancer must use premium network tiers.



Load balancers distribute load within a single region or across multiple regions. The different load balancer types are characterized by these features:

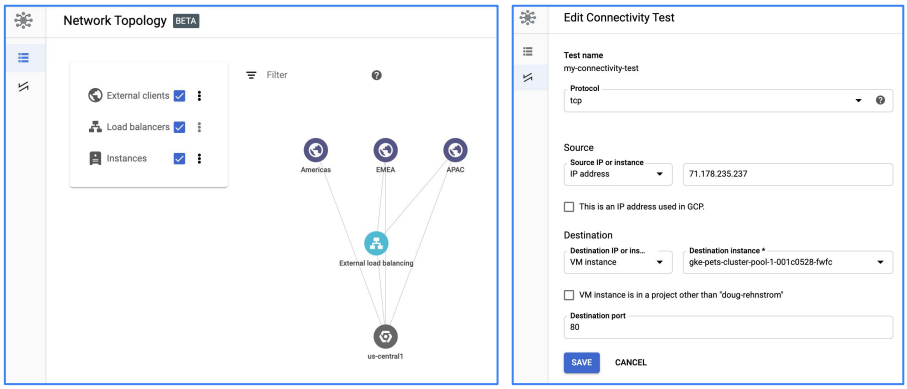
- Global(multi-regional) vs. regional
- External vs. internal
- Traffic type: HTTP(S), TCP, UDP

Cloud Load Balancing attributes:

- Distribute load-balanced resources in single or multiple regions
- Meet high-availability requirements
- Put resources behind a single anycast IP address
- Scale resources up or down with intelligent autoscaling
- Use Cloud CDN for optimal content delivery

Content can be served as close as possible to users, on a system that can respond to over 1 million queries per second. Cloud Load Balancing is a fully distributed, software-defined managed service. It is not instance- or device-based, so managing a physical load balancing infrastructure is not necessary.

Network Intelligence Center can be used to visualize network topology and test network connectivity



Google Cloud

Network Intelligence Center, available from the Cloud Console, provides visibility into network topology and a centralized monitoring facility. This helps with troubleshooting and security. The center provides the ability to test connectivity, which supports security and compliance checks.The ability to view traffic flows is very powerful, and metrics support the planning and optimizing of architecture.

There is no need for any configuration: the telemetry data is automatically collected to produce the visualizations. By default, two days' history is preserved, although up to six weeks can be configured to be retained. An important point to note is that the visualization will only display resources that communicated during the selected display period.

Activity 8: Defining network characteristics

Refer to your Design and Process Workbook.

- Specify the network characteristics for your case study VPC.
- Choose the type of load balancer required for each service.

Google Cloud

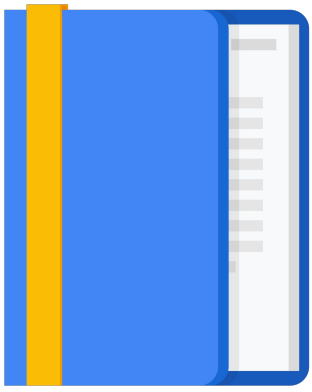
Agenda

Designing Google Cloud Networks

Design Activity #8

Connecting Networks

Design Activity #9



Google Cloud

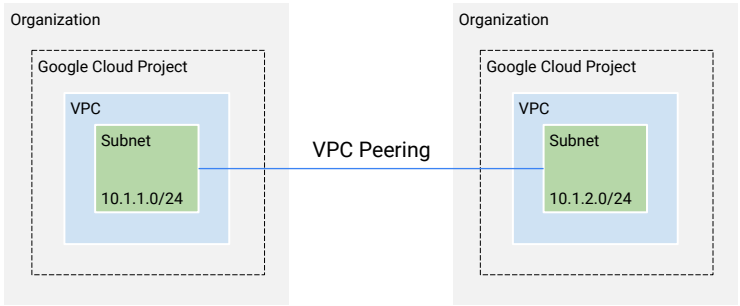
This chapter will introduce the different ways of connecting networks to Google VPC networks.

The options discussed are:

- VPC Peering
- Virtual private networks - Cloud VPNs
- Cloud Interconnect - both Dedicated Interconnect and Partner Interconnect

Use VPC peering to connect networks when they are both in Google Cloud

- Can be the same or different organizations.
- Subnet ranges cannot overlap.
- Network admins for each VPC must approve the peering requests.



Google Cloud

VPC Network Peering allows services to be made available privately across different VPC networks. The networks can be in the same project, different projects, or projects in different organizations.

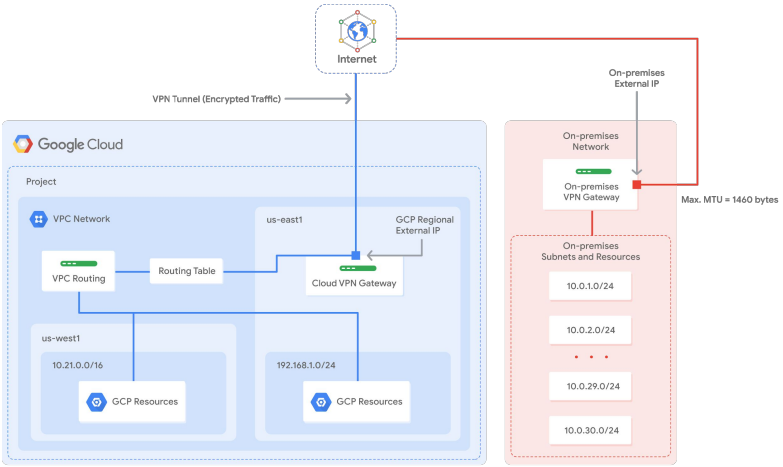
All communication happens by using private RFC 1918 IP addresses. VM instances in each peered network can communicate with one another without using external IP addresses, assuming firewall rules allow this.

Peered networks share subnet routes. Optionally, both networks can be configured to share custom static and dynamic routes too. Network administration for each peered network is unchanged: network admins and security admins for one network do not automatically get those roles for the other network in the peering relationship. If two networks from different projects are peered, project owners, editors, and Compute Instance admins in one project do not automatically receive those roles in the project that contains the other network.

When using VPC peering, care should be taken on VPC limits. The per network VPC limits are no longer relevant, and peer network group limits are applied. As an example, the VPC limit per network for VM instances that can be connected to a VPC is 15,500. But the peer network group limit is also 15,500. So in the slide above, the total number of VM instances across the two VPC networks is 15,500, not 15,500 per network. For more details on VPC peering limits, see: https://cloud.google.com/vpc/docs/quota#vpc-peering.

Use Cloud VPN to connect a Google Cloud network to a network on-premises or in another cloud

- 99.9% SLA
- For low-volume data connections
- Can configure static or dynamic routes using BGP (Border Gateway Protocol)



Google Cloud

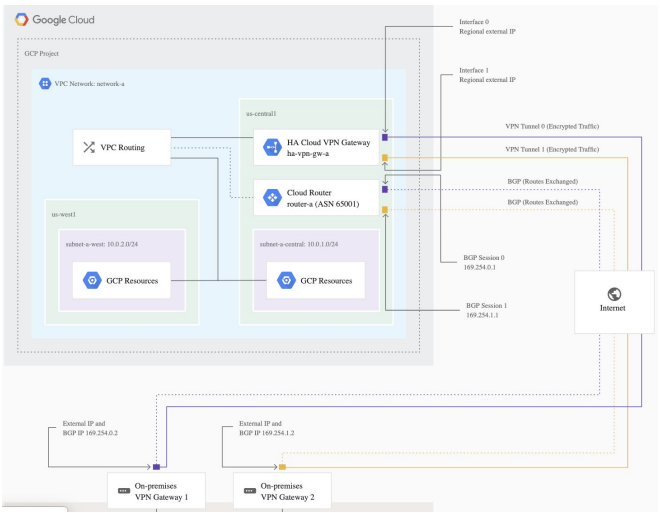
Google provides three connectivity solutions for connecting a peer network to Google Cloud: Cloud VPN, Dedicated Interconnect, and Partner Interconnect. We will discuss all three on subsequent slides but start here with Cloud VPN.

Cloud VPN allows the connection of a peer network to Google's network through an IPsec VPN tunnel. Only IPsec is supported. Traffic is encrypted and travels between the two networks over the public internet. Cloud VPN is useful for low-volume data connections and where you do not require the low latency and high availability of Cloud Interconnect. A Cloud VPN tunnel doesn't require the overhead or costs associated with a direct, private connection. Cloud VPN only requires a VPN device in the peer network location.

Google Cloud offers two types of Cloud VPN: HA VPN and Classic VPN. The SLA and the IP configuration on the slide above refer to the Classic VPN.

High availability VPN ensures 99.99% availability

- VPN gateway has 2 network interfaces.
- Creates two IP addresses.
- Each gateway supports multiple VPN tunnels.



Google Cloud

HA VPN is a high-availability (HA) Cloud VPN solution that lets you securely connect your on-premises network to your Google Cloud Virtual Private Cloud network through an IPsec VPN connection in single region.

HA VPN supports one of the following recommended topologies.

1. HA VPN gateway to peer VPN devices. This requires two VPN tunnels from the perspective of the HA VPN gateway. The vendor of the peer VPN gateway will help to determine which topology is most appropriate from the ones listed below.
   - HA VPN gateway to two separate peer VPN devices where each peer device has its own public IP address.
   - HA VPN gateway to one peer VPN device that has two separate public IP addresses.
   - An HA VPN gateway to one peer VPN device that has one public IP address.

2. HA VPN gateway to an AWS virtual private gateway, which is a peer gateway configuration with four interfaces.

3. HA VPN gateways connected to each other.

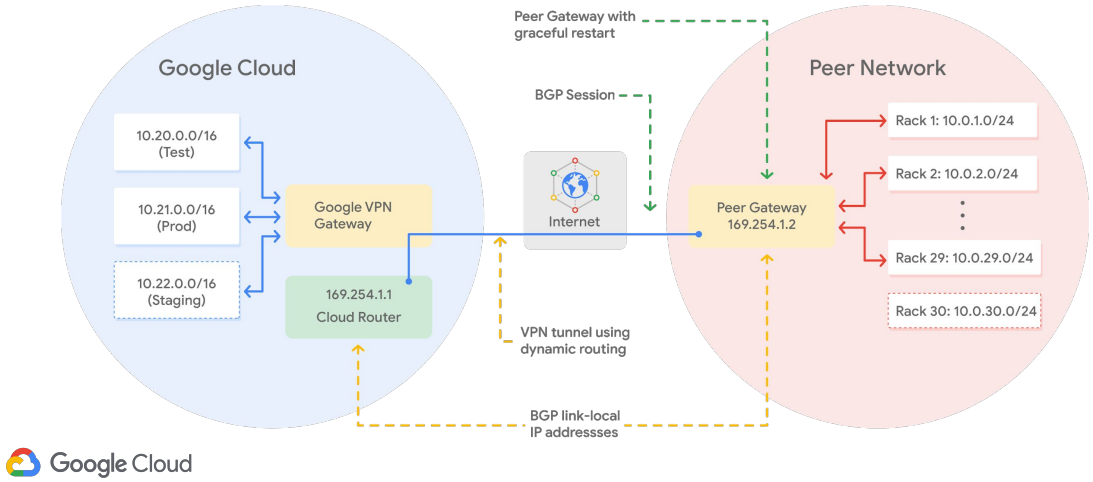
The diagram on the slides shows the two separate peer VPN devices with each peer having its own public IP address.

HA VPN provides an SLA of 99.99% service availability. To guarantee a 99.99%

availability SLA for HA VPN connections, you must properly configure 2 or 4 tunnels from your HA VPN gateway to your peer VPN gateway or to another HA VPN gateway.

When you create an HA VPN gateway, Google Cloud automatically chooses two public IP addresses, one for each of its fixed number of two interfaces. Each IP address is automatically chosen from a unique address pool to support high availability. Each of the HA VPN gateway interfaces supports multiple tunnels. You can also create multiple HA VPN gateways. In addition, you don't need to create any forwarding rules for HA VPN gateways.

Cloud Router enables dynamic discovery of routes between connected networks



Cloud Router enables dynamic route updates between a Cloud VPN and a non-Google network. Cloud Router eliminates the need to configure static routes and automatically discovers network topology changes. It peers with the peer network VPN gateway or router and exchanges topology information using BGP. Any topology changes are automatically propagated in both directions.

Static routes could be configured on Cloud Router, but these must be manually maintained because topology changes and traffic cannot be rerouted when link failures occur.

Use Cloud Interconnect when a dedicated high-speed connection is required between networks

- Dedicated Interconnect provides a direct connection to a colocation facility.
  - From 10 to 200 Gbps
- Partner Interconnect provides a connection through a service provider.
  - Can purchase less bandwidth from 50 Mbps
- Allows access to VPC resources using internal IP address space.
- Private Google Access allows on-premises hosts to access Google services using private IPs.

Google Cloud

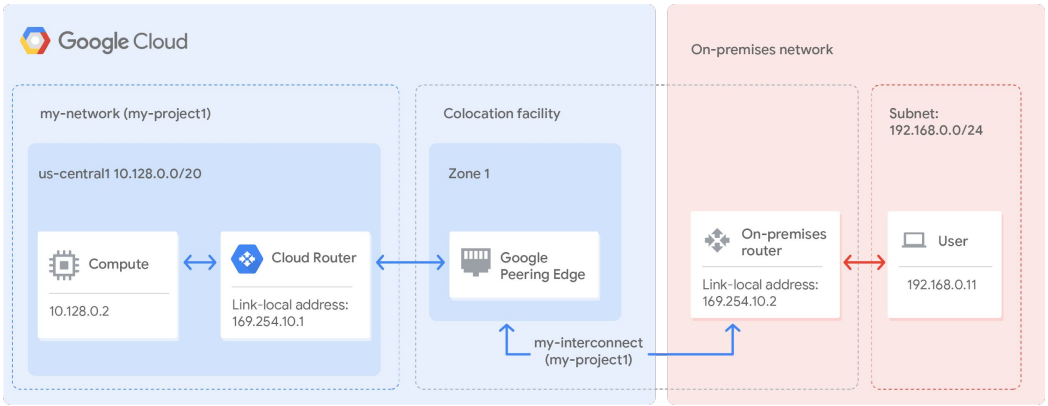
Cloud Interconnect provides low latency, highly available connections that enable you to reliably transfer data between on-premises and Virtual Private Cloud networks. Cloud Interconnect connections provide RFC 1918 communication, meaning internal (private) IP addresses are directly accessible from both networks.

Two options are available for extending on-premises networks:

1. Dedicated Interconnect, which provides a direct physical connection between an on-premises network and Google's network.
2. Partner Interconnect, which provides connectivity between on-premises and Google Cloud VPC networks through a supported service provider.

Traffic using Cloud Interconnect does not travel the public internet, which means fewer network hops and with fewer points of failure. The bandwidth can be scaled to meet your requirements incrementally. The scaling units depend on the type of interconnect and are discussed on the next slides.

Dedicated Interconnect provides direct physical connections

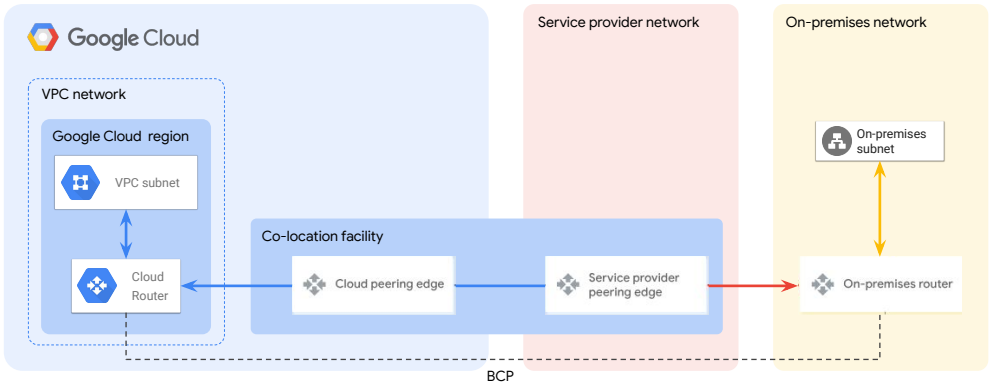


Google Cloud

For high bandwidth needs, Dedicated Interconnect is often a cost-effective solution. It requires the provisioning of a cross connect, dedicated internet connection between the Google network and your own router in a colocation facility. The diagram on the slide above shows a basic setup. A cross connect is provisioned between the Google network and the on-premises router in a common colocation facility. To exchange routes, a BGP session is configured over the interconnect between the Cloud Router and on-premises router. Then, traffic from the on-premises network can reach the VPC network, and vice versa.

A single interconnect can be a single 10-Gb link, a single 100-Gb link, or a link bundle (up to 8 x 10Gbps or 2 x 100Gbps), connected to a single Cloud Router.

Partner Interconnect provides connectivity through a supported service provider



Google Cloud

The use case is for those that have high bandwidth needs but cannot physically meet Google's network in a colocation facility. In this scenario, it is possible to use Partner Interconnect to connect via one of a variety of service providers that connect directly to Google. The connection capacities for each interconnect attachment (VLAN) supported are from 50 Mbps to 10 Gbps with a maximum capacity of 8 x 10 Gbps VLANs.

Activity 9: Diagramming your network

Refer to your Design and Process Workbook.

- Draw a diagram that depicts your network requirements.

Google Cloud

Quiz

You are deploying a large-scale web application with users all over the world and a lot of static content. Which load balancer configuration would likely be the best?

A. TCP load balancer with SSL configured

B. HTTP load balancer with SSL configured

C. HTTP load balancer with SSL configured and the CDN enabled

D. UDP load balancer with SSL configured and the CDN enabled

Google Cloud

You are deploying a large-scale web application with users all over the world and a lot of static content. Which load balancer configuration would likely be the best?
A. TCP load balancer with SSL configured
B. HTTP load balancer with SSL configured
C. HTTP load balancer with SSL configured and the CDN enabled
D. UDP load balancer with SSL configured and the CDN enabled

Quiz

You are deploying a large-scale web application with users all over the world and a lot of static content. Which load balancer configuration would likely be the best?

A. TCP load balancer with SSL configured

B. HTTP load balancer with SSL configured

C. HTTP load balancer with SSL configured and the CDN enabled

D. UDP load balancer with SSL configured and the CDN enabled

A. This answer is not correct. TCP load balancers are not intended for HTTP(S) traffic. In addition, the static content suggests use of CDN, which is not supported with TCP load balancers.

B. This answer is not correct. An HTTP load balancer with SSL configured is a good fit but not the best because CDN is not enabled, which would help with the large amount of static content.

C. This answer is correct. The traffic is HTTP(S), the load balancer should be external and global, and CDN enabled will help performance and cost.

D. This answer is not correct. The traffic type is not UDP. UDP load balancers are also not global.

Quiz

You are a large bank deploying an online banking service to Google Cloud. The service needs high-volume access to mainframe data on-premises. Which connectivity option would likely be best?

A. VPN

B. HTTPS

C. Cloud Interconnect

D. Peering

Google Cloud

You are a large bank deploying an online banking service to Google Cloud. The service needs high-volume access to mainframe data on-premises. Which connectivity option would likely be best?

A. VPN
B. HTTPS
C. Cloud Interconnect
D. Peering

Quiz

You are a large bank deploying an online banking service to Google Cloud. The service needs high-volume access to mainframe data on-premises. Which connectivity option would likely be best?

A. VPN

B. HTTPS

C. Cloud Interconnect

D. Peering

Google Cloud

A. This answer is not correct. VPN is an option for low data volumes.

B. This answer is not correct. HTTPS will not be able to provide bandwidth; there may be internet costs, and the traffic will be moved over the public internet.

C. This answer is correct. Cloud Interconnect provides high bandwidth and low latency. It does need encryption at the application level.

D. This answer is not correct. Peering is for connectivity to services such as G Suite.

Quiz

You have a contract with a service provider to manage your Google VPC networks. You want to connect a network they own to your VPC. Both networks are in Google Cloud. Which connection option should you choose?

A. VPN

B. VPN with high availability and Cloud Router

C. Cloud Interconnect

D. VPC peering

Google Cloud

You have a contract with a service provider to manage your Google VPC networks. You want to connect a network they own to your VPC. Both networks are in Google Cloud. Which connection option should you choose?
A. VPN
B. VPN with high availability and Cloud Router
C. Cloud Interconnect
D. VPC peering

Quiz

You have a contract with a service provider to manage your Google VPC networks. You want to connect a network they own to your VPC. Both networks are in Google Cloud. Which connection option should you choose?

A. VPN

B. VPN with high availability and Cloud Router

C. Cloud Interconnect

D. VPC peering

Google Cloud

A, B, and C are not correct. These options are all for connecting external networks to a VPC.

D. This answer is correct. VPC peering allows connectivity across two VPC networks regardless of whether they belong to the same project or same organization.

Quiz

You want a secure, private connection between your network and a Google Cloud network. There is not a lot of volume, but the connection needs to be extremely reliable. Which configuration below would you choose?

A. VPN

B. VPN with high availability and Cloud Router

C. Cloud Interconnect

D. VPC peering

Google Cloud

You want a secure, private connection between your network and a Google Cloud network. There is not a lot of volume, but the connection needs to be extremely reliable. Which configuration below would you choose?

A. VPN
B. VPN with high availability and Cloud Router
C. Cloud Interconnect
D. VPC peering

Quiz

You want a secure, private connection between your network and a Google Cloud network. There is not a lot of volume, but the connection needs to be extremely reliable. Which configuration below would you choose?

A. VPN

B. VPN with high availability and Cloud Router

C. Cloud Interconnect

D. VPC peering

A. This answer is not correct. VPN is the correct connectivity choice but does not offer HA.

B. This is the correct choice. This offers a secure extremely reliable connection and is more cost-effective than Cloud Interconnect.

C. This answer is not correct. Cloud Interconnect is for high data volumes.

D. This answer is not correct. VPC peering is for interconnection two VPC networks.

Review

Google Cloud and Hybrid Network Architecture

Google Cloud

In this module, you learned about Google Cloud networking and how to design networks that meet your application's security, performance, reliability, and scalability requirements.

We also covered the different options to connect networks using peering, VPN and Cloud Interconnect.

More resources

Cloud networking products
https://cloud.google.com/products/networking/

Google Cloud Hybrid Connectivity
https://cloud.google.com/hybrid-connectivity/

Google Cloud

The links provide access to some useful resources on Google Cloud networking.


Google Cloud