Google Cloud

Partner Certification Academy

Professional Cloud Architect

Google Cloud

The information in this presentation is classified:

Google confidential & proprietary

⚠ This presentation is shared with you under NDA.

- Do **not** record or take screenshots of this presentation.
- Do **not** share or otherwise distribute the information in this presentation with anyone **inside** or **outside** of your organization.

Thank you!

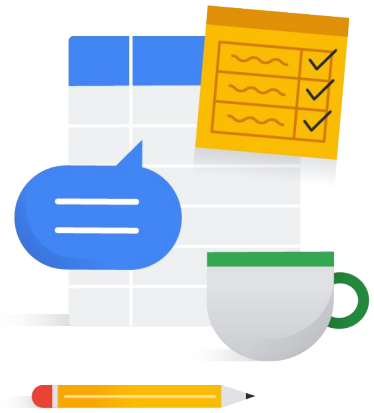
Google Cloud

Session logistics

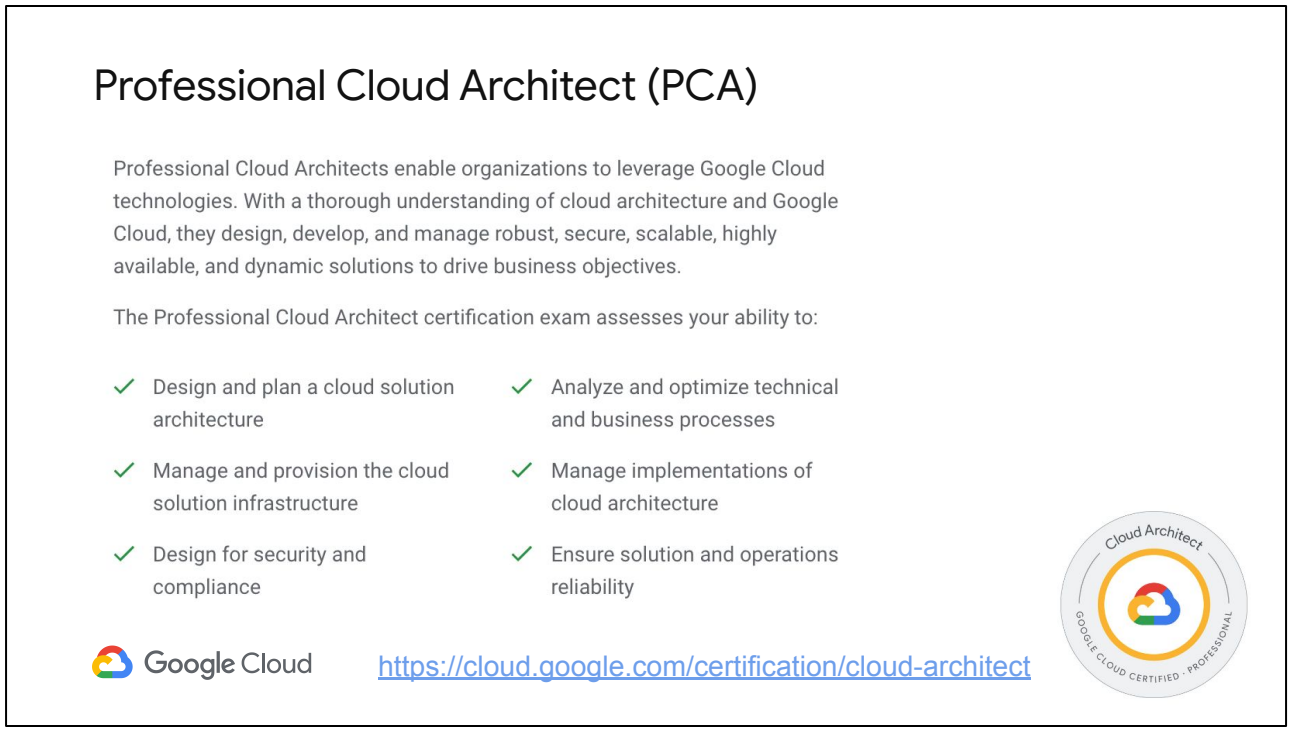
- When you have a question, please:
  - Click the Raise hand button in Google Meet.
  - Or add your question to the Q&A section of Google Meet.
  - Please note that answers may be deferred until the end of the session.
- These slides are available in the Student Lecture section of your Qwiklabs classroom.
- The session is **not recorded.**
- Google Meet does not have persistent chat.
  - If you get disconnected, you will lose the chat history.
  - Please copy any important URLs to a local text file as they appear in the chat.

Program issues or concerns?

- Problems with **accessing** Cloud Skills Boost for Partners
  - partner-training@google.com
- Problems with **a lab** (locked out, etc.)
  - support@qwiklabs.com
- Problems with accessing Partner Advantage
  - https://support.google.com/googlecloud/topic/9198654



Google Cloud



What is a PCA supposed to know/do?

Professional Cloud Architect Certification Page
https://cloud.google.com/certification/cloud-architect

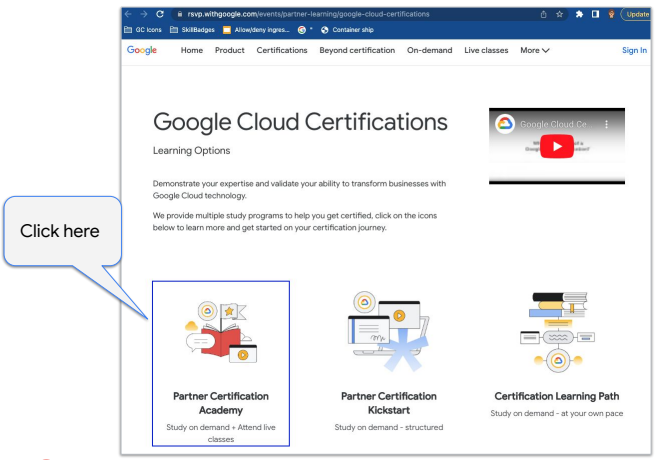
Professional Cloud Architect Exam Guide
https://cloud.google.com/certification/guides/professional-cloud-architect/

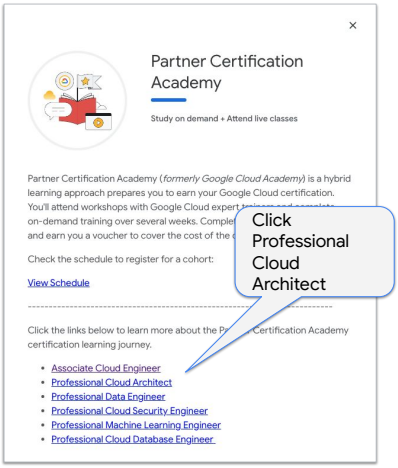
Professional Cloud Architect Sample Questions
https://docs.google.com/forms/d/e/1FAIpQLSdvf8Xq6m0kvyIoysdr8WZYCG32WHENStftiHTSdtW4ad2-0w/viewform

Google Cloud Adoption Framework + Whitepaper
https://cloud.google.com/adoption-framework
https://services.google.com/fh/files/misc/google_cloud_adoption_framework_whitepaper.pdf

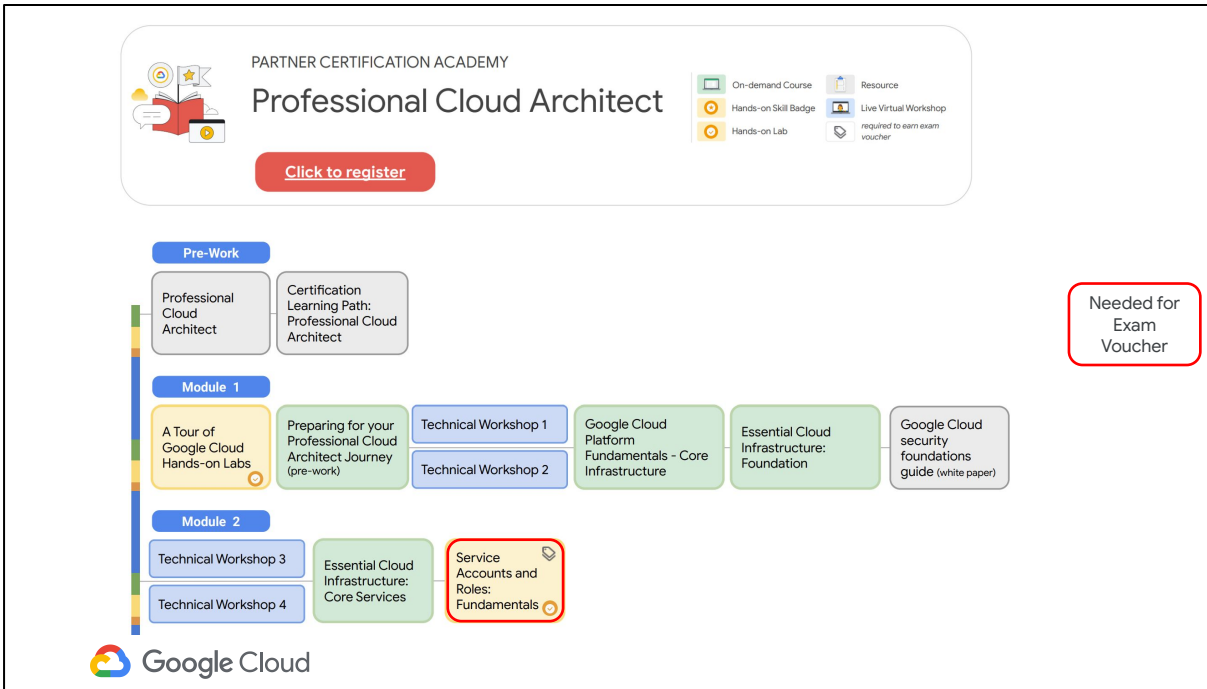
Learning Path - Partner Certification Academy Website

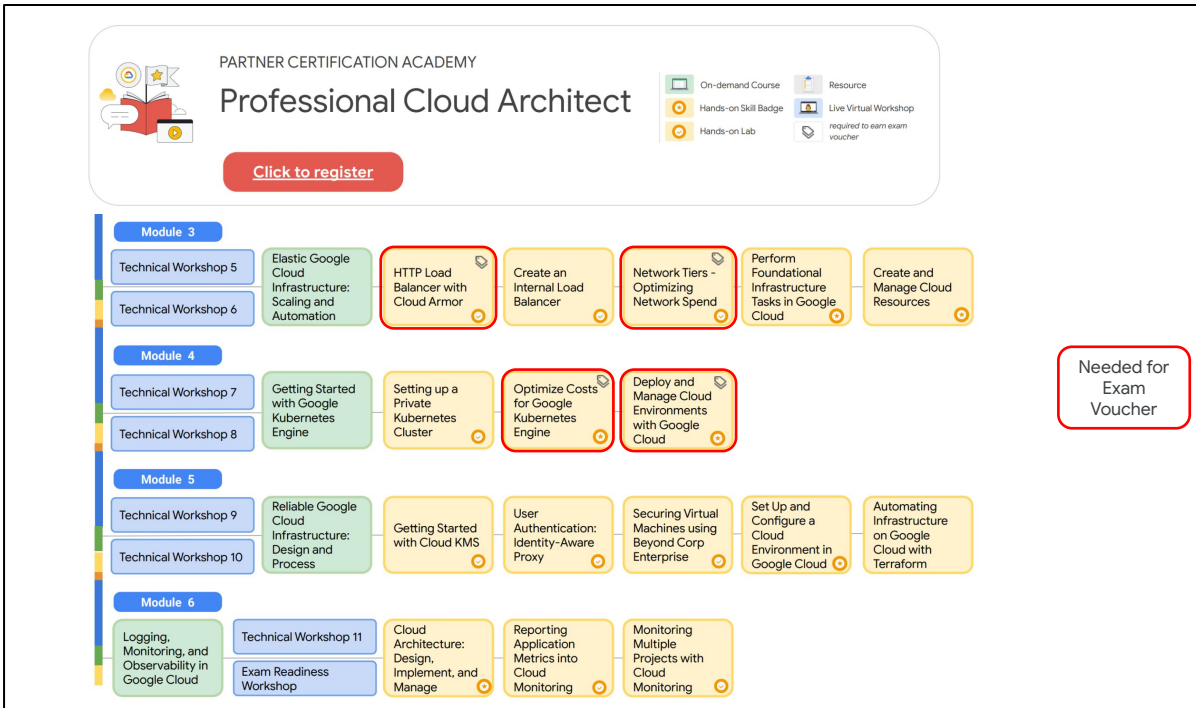
Go to: https://rsvp.withgoogle.com/events/partner-learning/google-cloud-certifications





Google Cloud


PARTNER CERTIFICATION ACADEMY
Professional Cloud Architect
Click to register
On-demand Course
Hands-on Skill Badge
Hands-on Lab
Resource
Live Virtual Workshop
required to earn exam voucher
Pre-Work
Professional Cloud Architect
Certification Learning Path: Professional Cloud Architect
Module 1
A Tour of Google Cloud Hands-on Labs
Preparing for your Professional Cloud Architect Journey (pre-work)
Technical Workshop 1
Technical Workshop 2
Google Cloud Platform Fundamentals - Core Infrastructure
Essential Cloud Infrastructure: Foundation
Google Cloud security foundations guide (white paper)
Module 2
Technical Workshop 3
Technical Workshop 4
Essential Cloud Infrastructure: Core Services
Service Accounts and Roles: Fundamentals
Needed for Exam Voucher
Google Cloud


PARTNER CERTIFICATION ACADEMY
Professional Cloud Architect
On-demand Course
Hands-on Skill Badge
Hands-on Lab
Resource
Live Virtual Workshop
required to earn exam voucher
Click to register
Module 3
Technical Workshop 5
Technical Workshop 6
Elastic Google Cloud Infrastructure: Scaling and Automation
HTTP Load Balancer with Cloud Armor
Create an Internal Load Balancer
Network Tiers - Optimizing Network Spend
Perform Foundational Infrastructure Tasks in Google Cloud
Create and Manage Cloud Resources
Module 4
Technical Workshop 7
Technical Workshop 8
Getting Started with Google Kubernetes Engine
Setting up a Private Kubernetes Cluster
Optimize Costs for Google Kubernetes Engine
Deploy and Manage Cloud Environments with Google Cloud
Module 5
Technical Workshop 9
Technical Workshop 10
Reliable Google Cloud Infrastructure: Design and Process
Getting Started with Cloud KMS
User Authentication: Identity-Aware Proxy
Securing Virtual Machines using Beyond Corp Enterprise
Set Up and Configure a Cloud Environment in Google Cloud
Automating Infrastructure on Google Cloud with Terraform
Module 6
Logging, Monitoring, and Observability in Google Cloud
Technical Workshop 11
Exam Readiness Workshop
Cloud Architecture: Design, Implement, and Manage
Reporting Application Metrics into Cloud Monitoring
Monitoring Multiple Projects with Cloud Monitoring
Needed for Exam Voucher

Key concepts in the on-demand content

Topics in this module

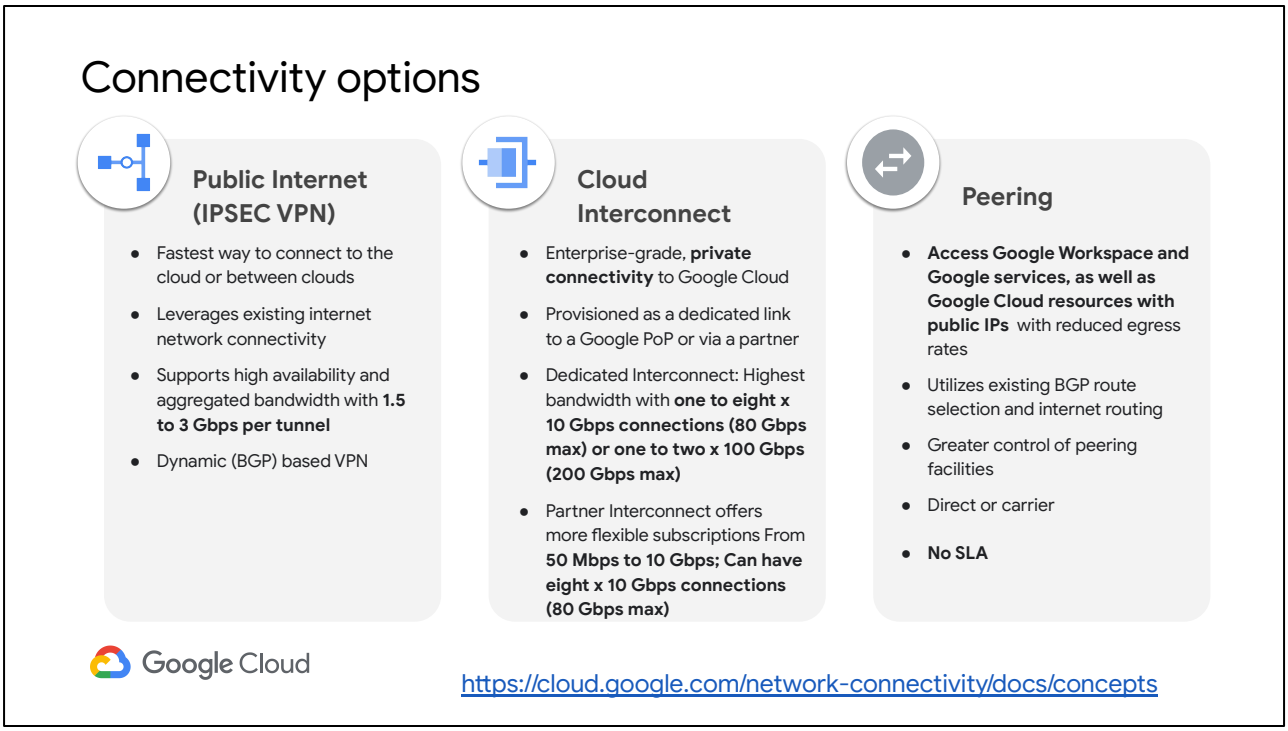
- Hybrid Connectivity
- Instance Groups
- Load Balancing
- Data Analytics pipeline

Google Cloud

Hybrid Connectivity...

Choosing a Network Connectivity product

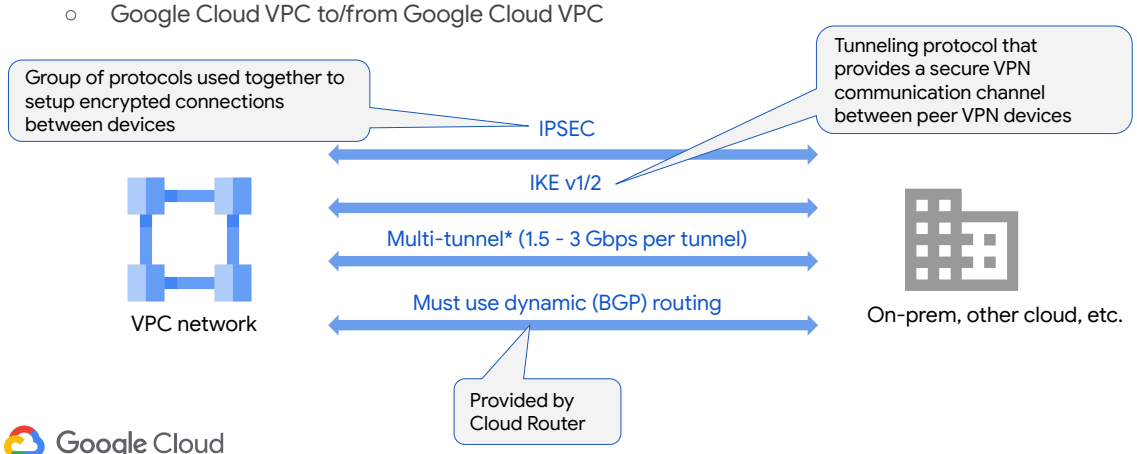
- On-prem network to Google Cloud with with lower throughput, IPsec encryption
  - Cloud VPN
- On-prem network to Google Cloud VPC Networks with high throughput needs
  - Cloud Interconnect (Dedicated or Partner)
- On-prem network to Google Workspace/supported public Google APIs (e.g., gmail)
  - Direct Peering
  - Carrier Peering
- Website which explains all of the above
  - https://cloud.google.com/network-connectivity/docs/how-to/choose-product



Network Connectivity
https://cloud.google.com/network-connectivity/docs/concepts

High Availability VPN overview

- Supports site-to-site VPN for different topologies/configuration scenarios:
  - Google Cloud VPC to/from on-premise
  - Google Cloud VPC to/from Amazon Web Services (AWS) virtual private gateway
  - Google Cloud VPC to/from Google Cloud VPC



On Google Cloud, dynamic routing can be established using Cloud Router. It exchanges network topology information through Border Gateway Protocol (BGP). Cloud Router advertises subnets from its VPC network to another router or gateway via BGP.

HA VPN is a high availability Cloud VPN solution that lets you securely connect your on-premises network to your Virtual Private Cloud (VPC) network through an IPsec VPN connection in a single region. 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 two or four 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 external 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. When you delete the HA VPN gateway, Google Cloud releases the IP addresses for reuse. You can configure an HA VPN gateway with only one active interface and one external IP address; however, this configuration does not provide a 99.99% service availability SLA. VPN tunnels connected to HA VPN gateways must use dynamic (BGP) routing. Depending on the way that you configure route priorities for HA VPN tunnels, you can create an active/active or

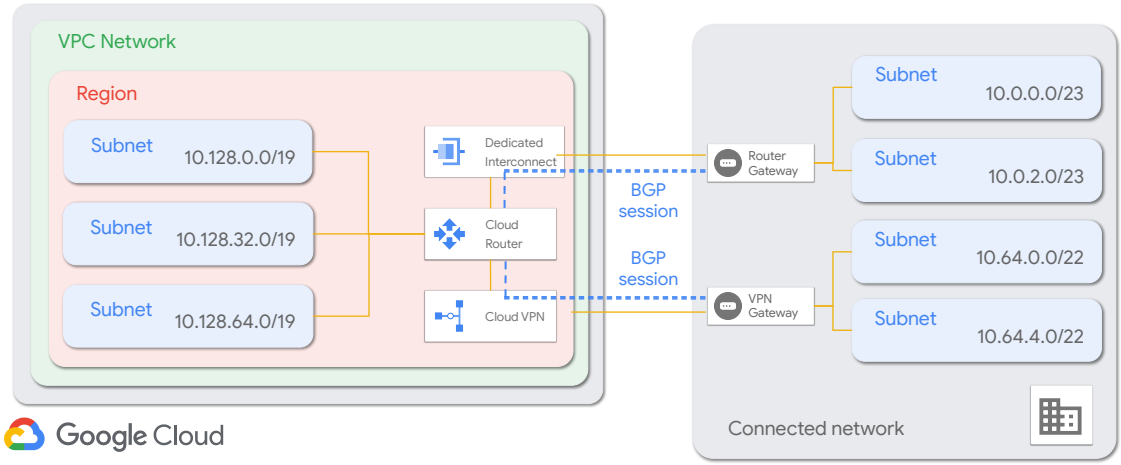
active/passive routing configuration.

HA VPN supports site-to-site VPN in one of the following recommended topologies or configuration scenarios:

- An HA VPN gateway to peer VPN devices
- An HA VPN gateway to an Amazon Web Services (AWS) virtual private gateway
- Two HA VPN gateways connected to each other

Cloud Router

- Enables dynamic discovery of routes between connected networks
- Used with Cloud VPN and Cloud Interconnect
- Free service



Cloud Router overview:
https://cloud.google.com/network-connectivity/docs/router/concepts/overview

Cloud Router pricing
https://cloud.google.com/network-connectivity/docs/router/pricing

Cloud Router is a fully distributed and managed Google Cloud service that uses the Border Gateway Protocol (BGP) to advertise IP address ranges. It programs custom dynamic routes based on the BGP advertisements that it receives from a peer. Instead of a physical device or appliance, each Cloud Router consists of software tasks that act as BGP speakers and responders. A Cloud Router also serves as the control plane for Cloud NAT. Cloud Router provides BGP services for the following Google Cloud products:

- Dedicated Interconnect
- Partner Interconnect
- Cloud VPN, specifically HA VPN
- Router appliance (part of Network Connectivity Center)

Key points:

- An important **building block** of **Cloud VPN** and **Interconnect**
- **Managed service**
- **Regional**
- **Control plane** only. Not part of the data path.
- **Exchanges routes** between your VPC and your on-premises network via **BGP**
- **Dynamic routing** options
    - Regional: Shares routes **only for subnets** in the region **where Cloud Router is provisioned**
    - Global: Shares routes for **all subnets** in the VPC
- **Route advertisement** (on BGP session or Cloud Router level)
    - **Default**: Advertises subnets according to the selected routing option
    - **Custom:** Customise which subnets and IP ranges to advertise, for example:
        - when you want to **skip advertising specific subnets**
        - when you want to **advertise subnets in different BGP sessions** for **separation purposes**
- **Scaling:**
    - Note that **dynamic routes** learned by Cloud Router are l**imited to 100 by default**.
    - The **routes** dynamically announced by **on-premises to Google Cloud** should be **summarized to avoid** hitting this **quota**.

Suggested Lab (if time allows)

https://partner.cloudskillsboost.google/catalog_lab/620

Google Cloud

https://partner.cloudskillsboost.google/catalog_lab/620

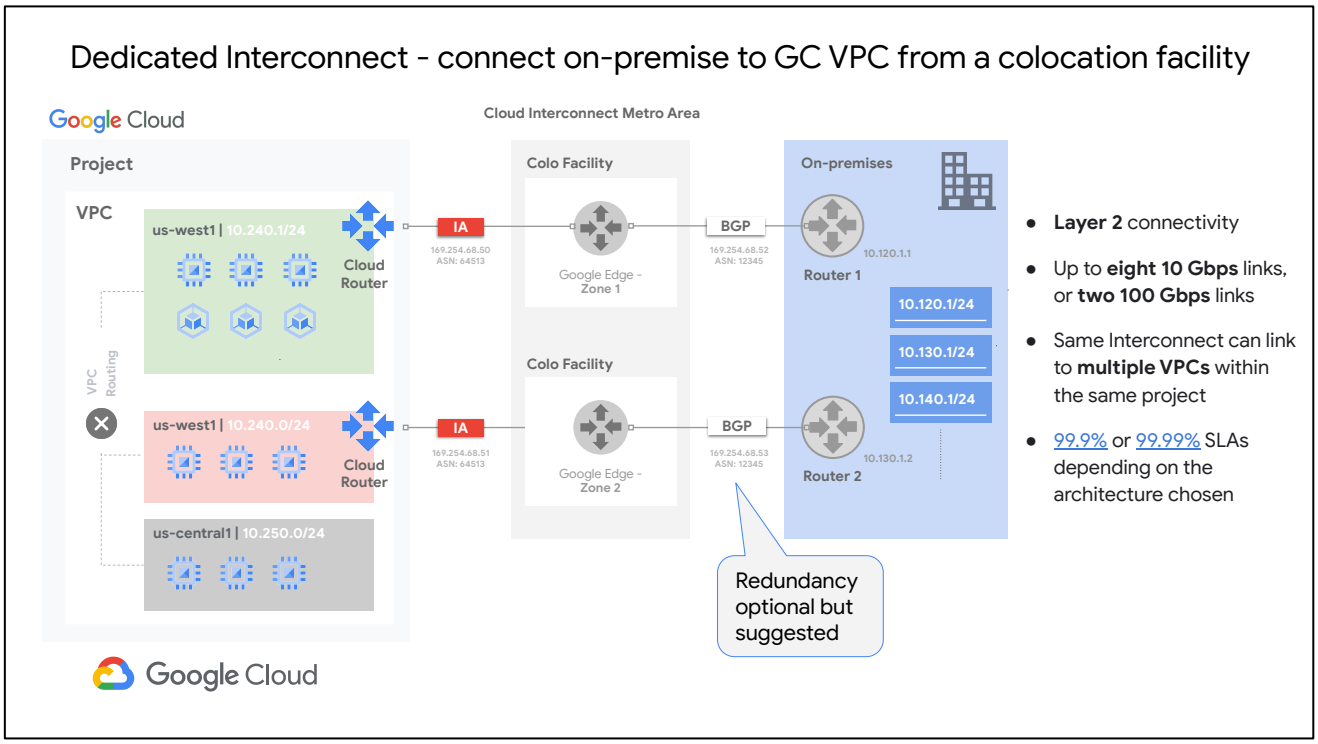
Cloud Interconnect

- Provides a **dedicated high-speed connection** between on-prem and GC VPC
  - Connect to Compute Engine VMs from on-premise using internal IPs
- **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 to a maximum of 8 x 10 Gbps connections (80 Gbps)

Dedicated Interconnect overview
https://cloud.google.com/network-connectivity/docs/interconnect/concepts/dedicated-overview

Partner Interconnect overview
https://cloud.google.com/network-connectivity/docs/interconnect/concepts/partner-overview

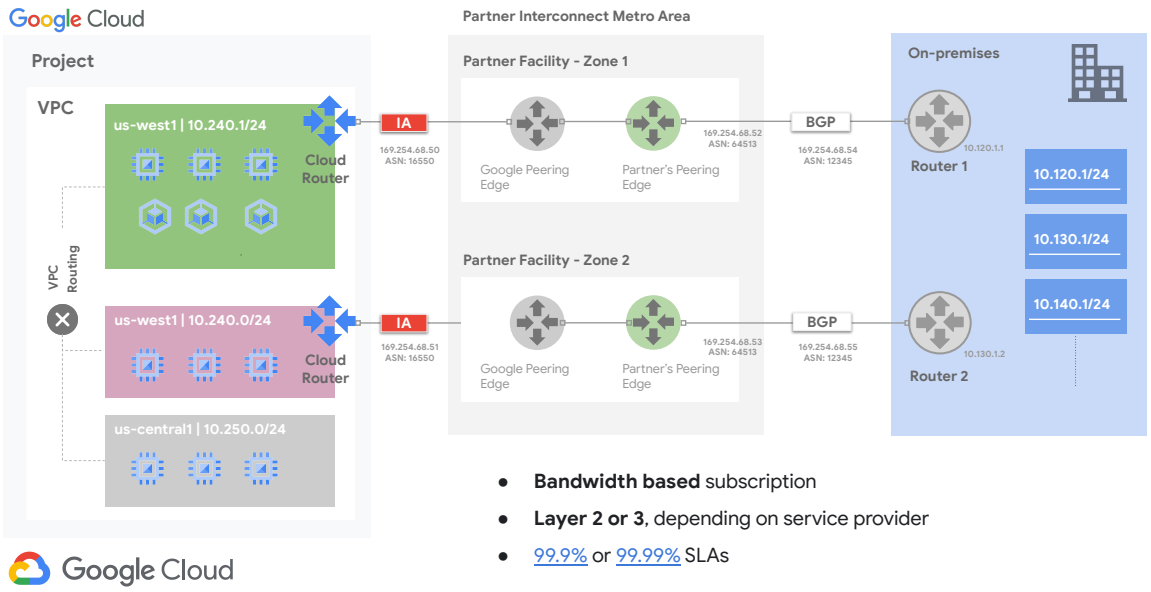
Cloud Interconnect FAQ
https://cloud.google.com/network-connectivity/docs/interconnect/support/faq



- **Key features**
  - Traffic doesn't traverse the public internet, and **takes fewer hops**, which means **higher reliability**.
  - Your **VPC network** internal (**RFC 1918**) IP addresses are **directly accessible from your on-premises** network. VPN or NAT are not needed.
  - **High bandwidth**: **8 x 10Gbps**, or **2 x 100Gbps**
  - **Reduces egress costs** as traffic remains internal
  - Works with **Private Google Access** to allow Google API access through internal links
  - Traffic is **unencrypted**. Use **own-managed IPSEC VPN** if that's a requirement. Cloud VPN over Interconnect is not supported.
- **Components**
  - **VLAN attachments** create a **VLAN** and associates with a **Cloud Router**. Can be **attached to multiple VPCs**. Limits:
    - Max attachments per single interconnect: 16
    - Max bandwidth per attachment: 50 Mbps to 50 Gbps
  - **Interconnect location** is the **colocation facility**. There a a few per **metro area** for **redundancy**. Each serves a list of regions (full list).
  - **Cloud Router** is used to dynamically exchange routes
- **Diagram overview**
  - **On-premises:** Two routers with links to two colo facilities in the same

- metro area for redundancy
- **IA (Interconnect Attachment) / VLAN attachments**: One from each colo facility to a different Cloud Router
- **BGP session** between **Cloud Router and on-premises routers** for routes exchange
- In this case, **Cloud Routers** are on us-west1. To allow **access from on-premises to** resources on **us-central1 subnet**, **global routing** would be needed

Partner Interconnect - connect on-premise to GC VPC using a service provider



- **Bandwidth based** subscription
- **Layer 2 or 3**, depending on service provider
- 99.9% or 99.99% SLAs

Key points:

- Good fit when your data center **can't reach a dedicated interconnect facility**, or if you **don't need an entire 10 Gbps link.**
- **Bandwidth** based subscription is more flexible**: 50 Mbps - 10 Gbps** per link, up to **8 x 10 Gbps links**
- **Connectivity layer** depends on service provider. See full list.
  - **Layer 2**: BGP session required between **on-prem and Google Cloud routers** for **each VLAN attachment**
  - **Layer 3**: **Service provider establishes** a BGP session between **Google Cloud and their edge routers** for **each VLAN attachment**
- **Supported configurations** for 99.9% and 99.99% SLAs.
- Note that when using Partner Interconnect, the ASN configured on the Cloud Routers must be 16550.

Summary: Hybrid connectivity: How will you connect to your VPCs?

VPN over the Internet

- Easy setup, quick to deploy
- HA VPN recommended for redundancy on the Google Cloud side
- Still relies on Internet so less predictable and might compete for on-premises networks Internet bandwidth
- Performance and bandwidth depends on your Internet Gateway but about 1.5 Gbps per VPN tunnel
- Pricing

Partner Interconnect

- Needs to go through a Partner to get it configured
- Redundancy possible with multiple Partner Interconnects connections
- Uses a private link between the Partner and Google so more predictable and reliable
- Flexible bandwidth between 50 Mbps and 50 Gbps
- Pricing

Dedicated Interconnect

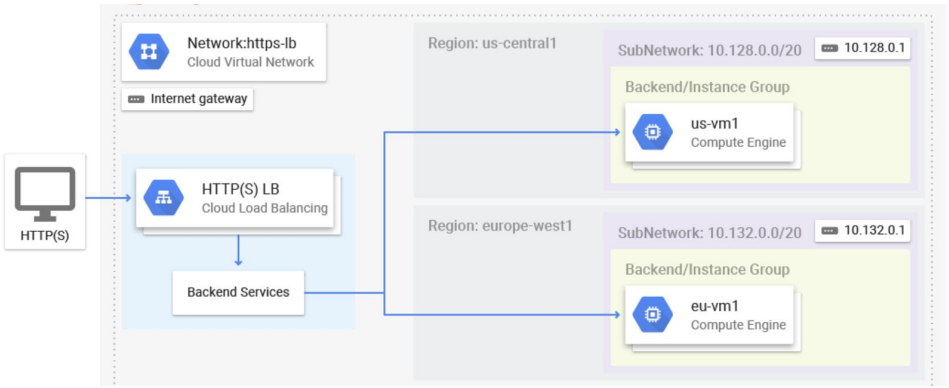
- Needs you to meet Google in a colocation facility
- Redundancy possible with multiple Dedicated Interconnects connections
- Uses a private and dedicated link between you and Google so more predictable and reliable
- Flexible bandwidth between **10 Gbps (minimum)** and 100 Gbps
- Pricing

Best practices

Use HA VPN over the Internet to get started, Partner Interconnect when you need a reliable connection between you and Google Cloud and Dedicated Interconnect when you need bandwidth to Google Cloud above multiple Gbps.

Instance Groups...

Scalable, Fault Tolerant Architecture Illustrated



Google Cloud

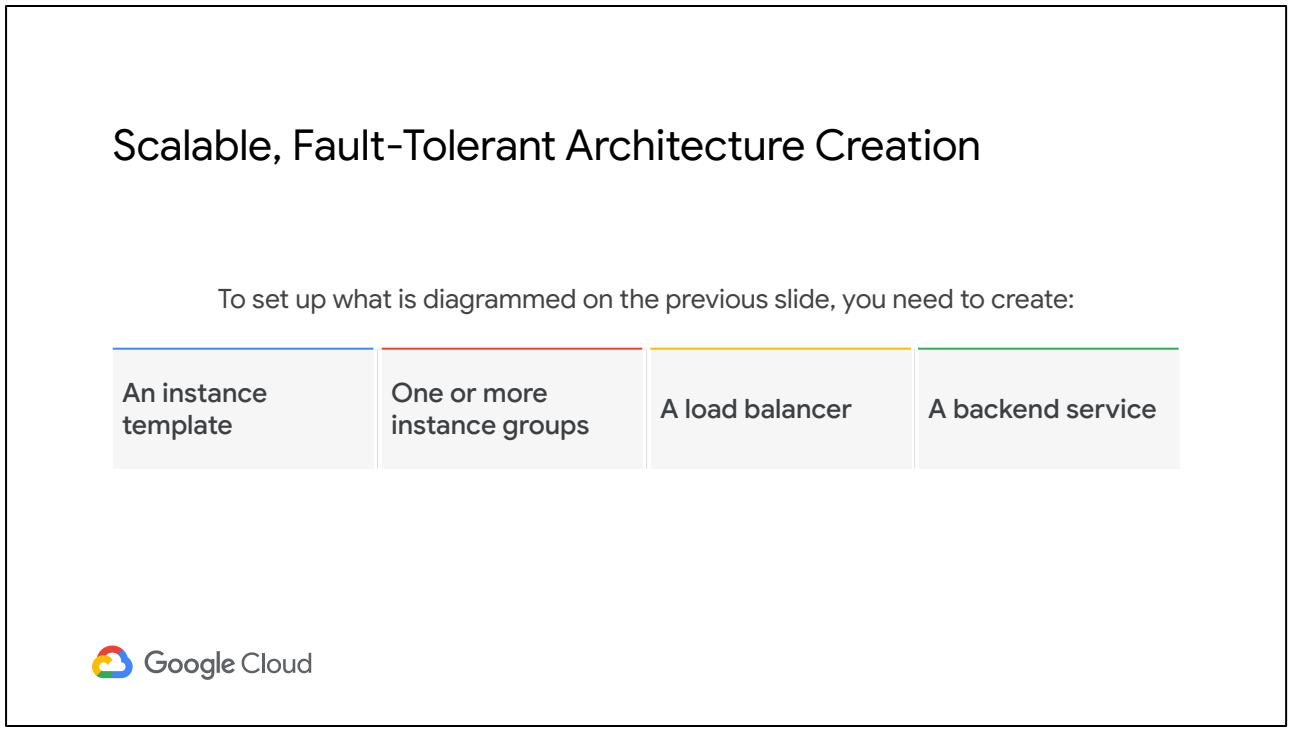
This illustration shows a fault tolerant architecture using load balancing and autoscaling.

In this example, HTTP(S) traffic is being delivered to an Global Load Balancer. The load balancer is a managed service, so performance and high availability is built into the service.

The backend service is responsible for distributing traffic to the appropriate region based on latency.

Each region, in this example, has a managed instance group of instances that will be increased or decreased based on a predefined metric you choose based on the application.

The infrastructure shown can survive a region outage and still function.



In order to build the infrastructure in the previous slide, you would need to:

- Build an instance template
- Create one or more managed instance groups(MIG) at the regions required
- Build a global load balancer with the MIG as the backend service

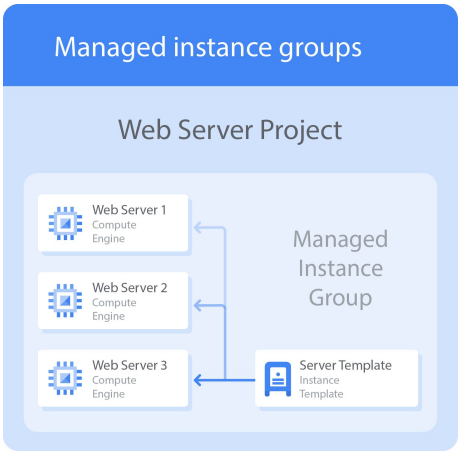
Compute Engine elasticity - Instance Groups

- Two types
  - **Managed instance groups**
    - Group of **identical** machines created from a template
    - Use case: When need VM elasticity based on demand
  - **Unmanaged instance groups**
    - Group of VMs with **different configurations**
    - Use case: Lift and shift of on-premise workloads that need a load balancer to serve traffic
      - No automatic elasticity or automatic healing

https://cloud.google.com/compute/docs/instance-groups

Managed instance groups create VMs based on templates

- Instance templates define the VMs: image, machine type, etc.
  - Test to find the smallest machine type that will run your program
- Instance group manager creates the machines.
- Optimize cost and meet varying user workloads via autoscaling
- Enable auto healing via health checks
- Can be single zone or regional
  - Use multiple zones for high availability



Google Cloud

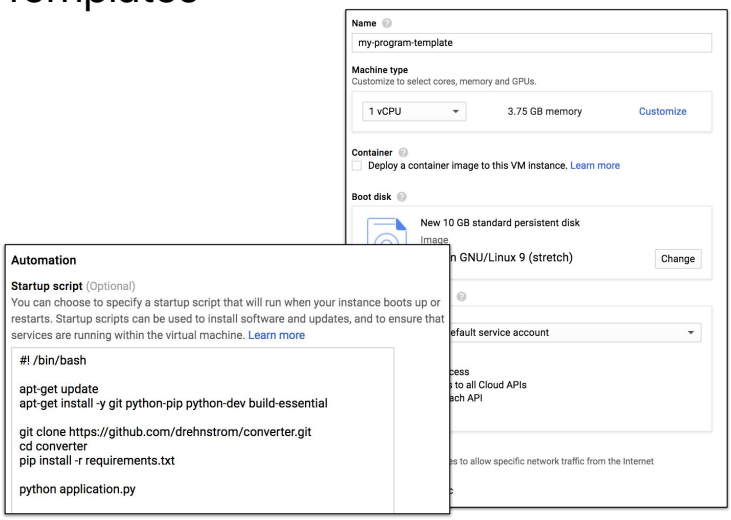
Managed instance groups create VMs based on instance templates. Instance templates are just a resource used to define VMs and managed instance groups. The templates define the boot disk image or container image to be used, the machine type, labels, and other instance properties like a startup script to install software from a Git repository.

The virtual machines in a managed instance group are created by an instance group manager. Using a managed instance group offers many advantages, such as autohealing to re-create instances that don't respond and creating instances in multiple zones for high availability.

Creating an Instance Templates

Contains the information required to build a VM for a deployment

- Like building a VM, but don't create the instance, create the template
- Include a startup script to automate code deployment
- Can also use a custom image



Google Cloud

https://cloud.google.com/compute/docs/instance-templates

Creating an instance template is similar to creating a virtual machine.

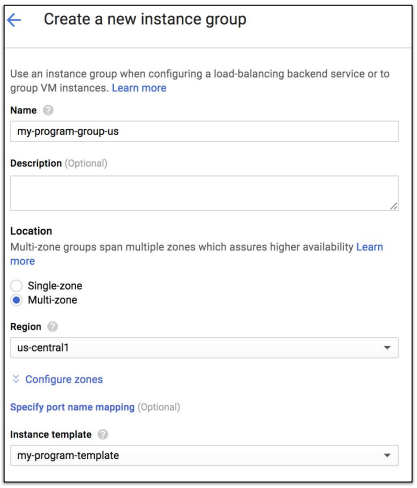
The template will include all information required to build the instance.

The information provided to build the template can include startup scripts to automate code deployment if needed.

Creating an Instance Group

Instance groups create machines based on instance templates

- Specify how many machines to create
- In which region
- Based on what template
- Autoscaler adds and removes machines based on demand
- Health check ensures machines are working



Google Cloud

An instance group enables you to manage a group of virtual machines as a single entity.

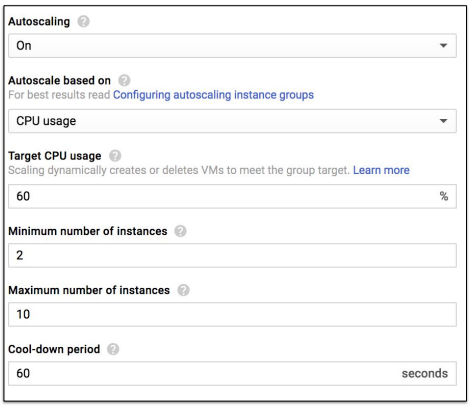
All virtual machines in a managed instance group are built from the same template.

When building the group, you specify:

- Number and location of instances
- Template to build from
- Metric or metrics to use to scale instances
- Health check to verify machines are working properly

Specifying an Autoscaling Configuration

- Defines when to create or destroy machines
- Pick a metric and a threshold
  - CPU utilization
  - Load balancing capacity
  - Monitoring metrics
  - Predictions based on history
- Specify minimum and maximum number of machines



Google Cloud

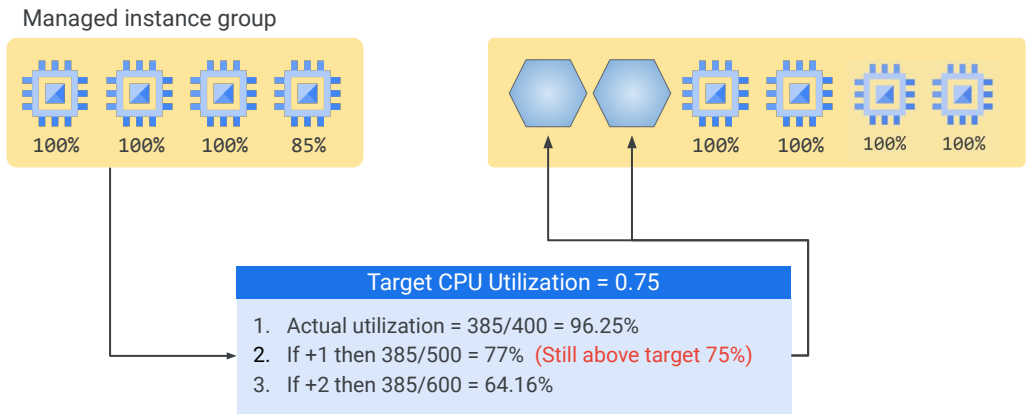
https://cloud.google.com/compute/docs/autoscaler

Part of the configuration of the managed instance group includes how to configure autoscaling. This configuration defines when to create and destroy instances.

Pick the metrics and thresholds to use.

Specify a minimum and maximum number of instances to run.

Example Scale-out policy decision



Google Cloud

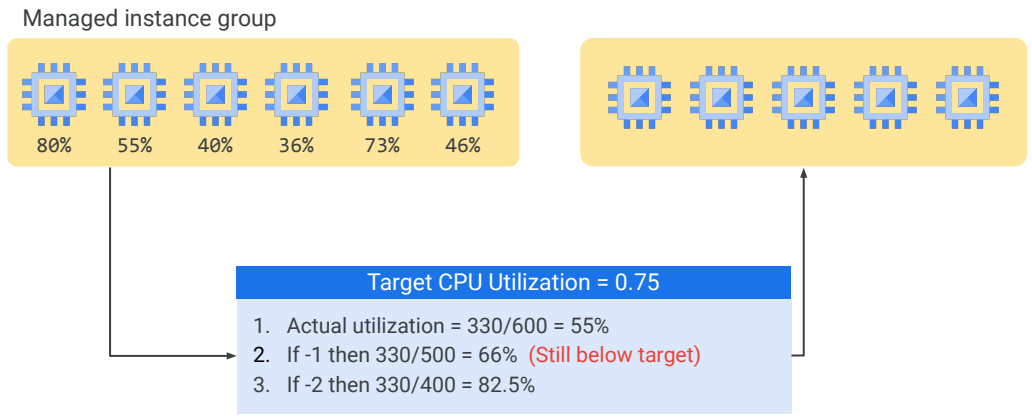
The percentage utilization that an additional VM contributes depends on the size of the group. The 4th VM added to a group offers 25% increase in capacity to the group. The 10th VM added to a group only offers 10% more capacity, even though the VMs are the same size.

In this case shown in the diagram Autoscaler is conservative and rounds up. In other words, it would prefer to start an extra VM that isn't really needed than to possibly run out of capacity.

https://cloud.google.com/compute/docs/autoscaler/understanding-autoscaler-decisions

https://cloud.google.com/compute/docs/autoscaler/multiple-policies

Example Scale-in policy decision



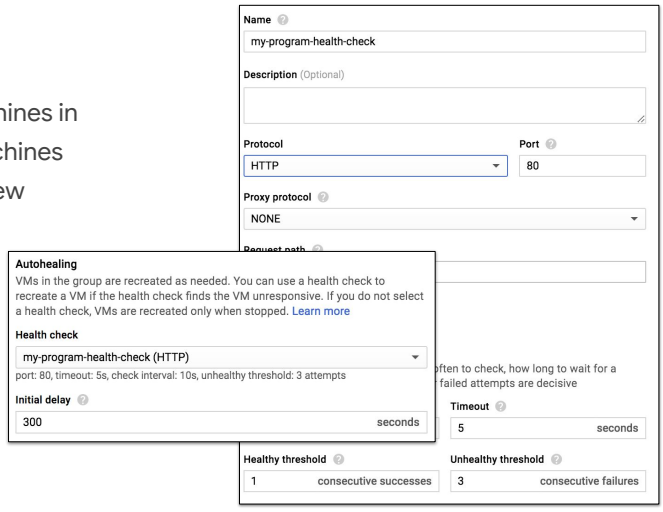
Google Cloud

In this example, removing one VM doesn't get close enough to the target of 75%. Removing a second VM would exceed the target. Autoscaler behaves conservatively. So it will shut down one VM rather than two VMs. It would prefer underutilization over running out of resource when it is needed.

TIP: When would you use Cloud Monitoring metrics for autoscaling? A couple of examples: When the customer has a custom metric (e.g., 'x' number of people playing a game) that they want to use for scaling. Or if the application running on the VMs is reading data from a Pub/Sub queue, maybe scale when the queue reaches a certain size.

Health Checks

- Simply makes requests to the machines in the instance groups, and if the machines don't work, they are shut off and new ones created
- Parameters control:
  - Where to make request
  - Via what port
  - How often



Google Cloud

Set up an application health check and autohealing

Health checks ensure that only fully operable instances are being used from the group. With health checks, instances that fail the health check can be excluded and new instances can take their place.

Parameters required to configure the health check include:

- Where to make the request
- What port to use
- And how often to run the check

Health checks need a firewall rule to allow incoming probes from certain ports. See https://cloud.google.com/compute/docs/instance-groups/autohealing-instances-in-migs

Stateful Managed Instance Groups

- MIGs support both stateful and stateless workloads
  - Stateful workloads preserve individual VM state (for example, a database shard, or app configuration) on the VM's disks
    - Not easy to scale horizontally (add more nodes)
      - Could require data replication, creation or deletion of data shards, or changing the overall application configuration
    - If VM is reported as “unhealthy”
      - Need to retain VM identity (name), IP address, metadata, and data to be used when a new one is created
  - Stateless workloads, like a web frontend, do not retain any state on the individual VMs
    - Easy to scale up/down as needed

Stateful managed instance groups

When to use Stateful Managed Instance Groups

- Databases such as Cassandra, ElasticSearch, MongoDB, and ZooKeeper.
  - You are responsible for replicating data if the database doesn't do it for you
- Data processing applications such as Kafka (Pub/Sub) and Flink (Dataflow)
- Other stateful applications such as TeamCity, Jenkins, Bamboo, or custom stateful workloads
- Legacy monolith applications that store application state on a boot disk or additional persistent disks
- Batch workloads with checkpointing.
  - Can preserve checkpointed results of long-running computation on disk in anticipation of workload or VM failure or instance preemption.
  - Stateful MIGs can recreate a failed machine, while preserving its data disk, so that your computation can continue from the last checkpoint.

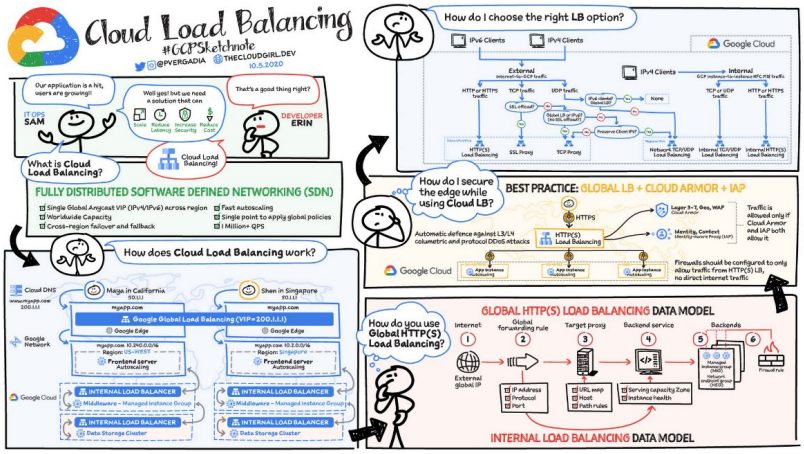
Comparison of Stateless and Stateful MIGs

	Stateless	Stateful
Autoscaling	YES	NO
Disk preservation	NO	YES
Auto-healing	YES	YES
Auto-updating	YES	YES
Load balancing	YES	YES
Multi-zone deployment	YES	YES

Health checks help ensure a Stateful VM can be recreated if a failure occurred

Cloud Load Balancing...

Cloud Load Balancing



Google Cloud

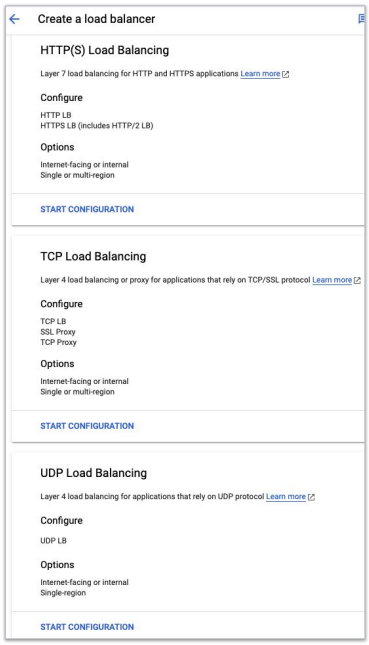
[What is Cloud Load Balancing?]

Video: https://www.youtube.com/watch?v=h8EqM6Xt3MA

Corresponding blog
https://cloud.google.com/blog/topics/developers-practitioners/what-cloud-load-balancing

Load balancing Features

- High performance
- Fully distributed and software defined
- Anycast IP
- Regional/zonal spillover and failover
- Intelligent backend autoscaling and health checks



Google Cloud

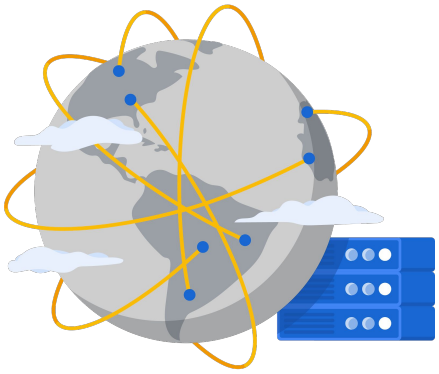
Cloud Load Balancing
https://cloud.google.com/load-balancing

- Load balancers are **software defined and distributed**, which means they are **highly performant** and not bottlenecked by a single appliance
- They are **tightly integrated with Compute Engine and GKE** to allow for **intelligent autoscaling** based on various metrics, and **health checks**, so we can route traffic to the **healthy instances and locations.**
- Load balancers are available based on **geo scope** (regional / global), **network tier** (premium / standard) and **proxy or pass-through**
- **Pass-through** proxies means client IP is preserved. **Proxy** means the client IP is not preserved, and a different connection is established between the LB and the backend instances.
- **High-level review**
  - **Internal**
    - **Regional** and require **premium network tier**
    - **L4 (TCP/UDP)** and **L7 (HTTP/s)**
  - **External**
    - Supports both **network tiers**
      - **Proxy** load balancers (TCP/SSL/HTTPs) are available as **global resources** by **directing traffic** to **healthy**

        - **backends** closest to the end-user. For that, they make use of **Google global network infrastructure**, and accordingly require **Premium Tier.**
        - They are also supported with **standard tier**, in which case they effectively function as **regional load balancers**
- **HTTPs** load balancer easily integrates with **Cloud CDN** and **Cloud Armor (WAF)**
- The **Network Load Balancer** provides a L4 (TCP/UDP) regional load balancer that is pass-through

HTTP(S) load balancing

- Global or regional (internal or external) load balancing
- Anycast IP address
- HTTP on port 80 or 8080
- HTTPs on port 443
- IPv4 or IPv6
- Autoscaling
- URL maps



HTTP(S) Load Balancing

Google Cloud

Google Cloud's HTTP(S) load balancing provides global load balancing for HTTP(S) requests destined for your instances. This means that your applications are available to your customers at a single anycast IP address, which simplifies your DNS setup. HTTP(S) load balancing balances HTTP and HTTPS traffic across multiple backend instances and across multiple regions.

HTTP requests are load balanced on port 80 or 8080, and HTTPS requests are load balanced on port 443.

This load balancer supports both IPv4 and IPv6 clients, is scalable, requires no pre-warming, and enables content-based and cross-region load balancing.

You can configure URL maps that route some URLs to one set of instances and route other URLs to other instances. Requests are generally routed to the instance group that is closest to the user. If the closest instance group does not have sufficient capacity, the request is sent to the next closest instance group that does have capacity.

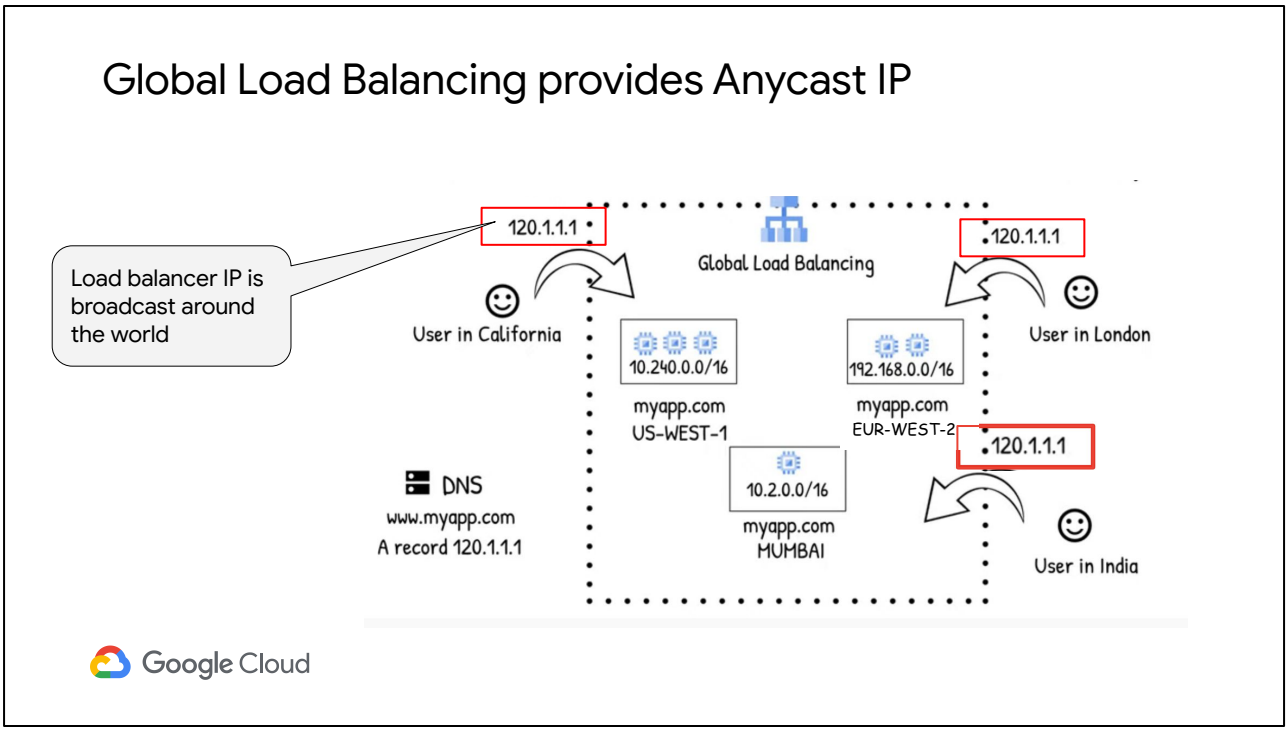
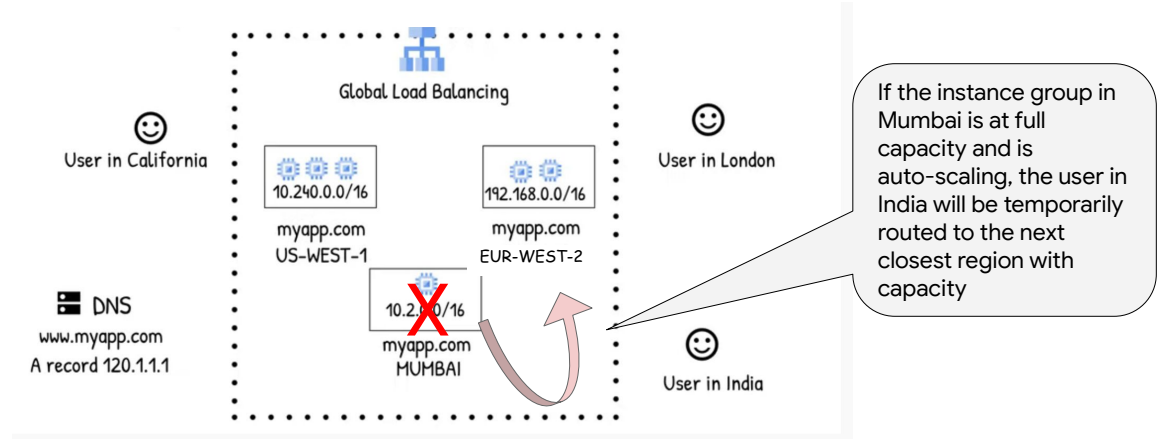


Image from video:
https://www.youtube.com/watch?v=0fQr7TRhnnU&list=PLTWE_Imu2InBzuPmOcgAYP7U80a87cpJd

Corresponding blog:
https://cloud.google.com/blog/topics/developers-practitioners/what-cloud-load-balancing

Regional/Zonal Spillover



Google Cloud

Image from video:
https://www.youtube.com/watch?v=0fQr7TRhnnU&list=PLTWE_Imu2InBzuPmOcgAYP7U80a87cpJd

Corresponding blog:
https://cloud.google.com/blog/topics/developers-practitioners/what-cloud-load-balancing

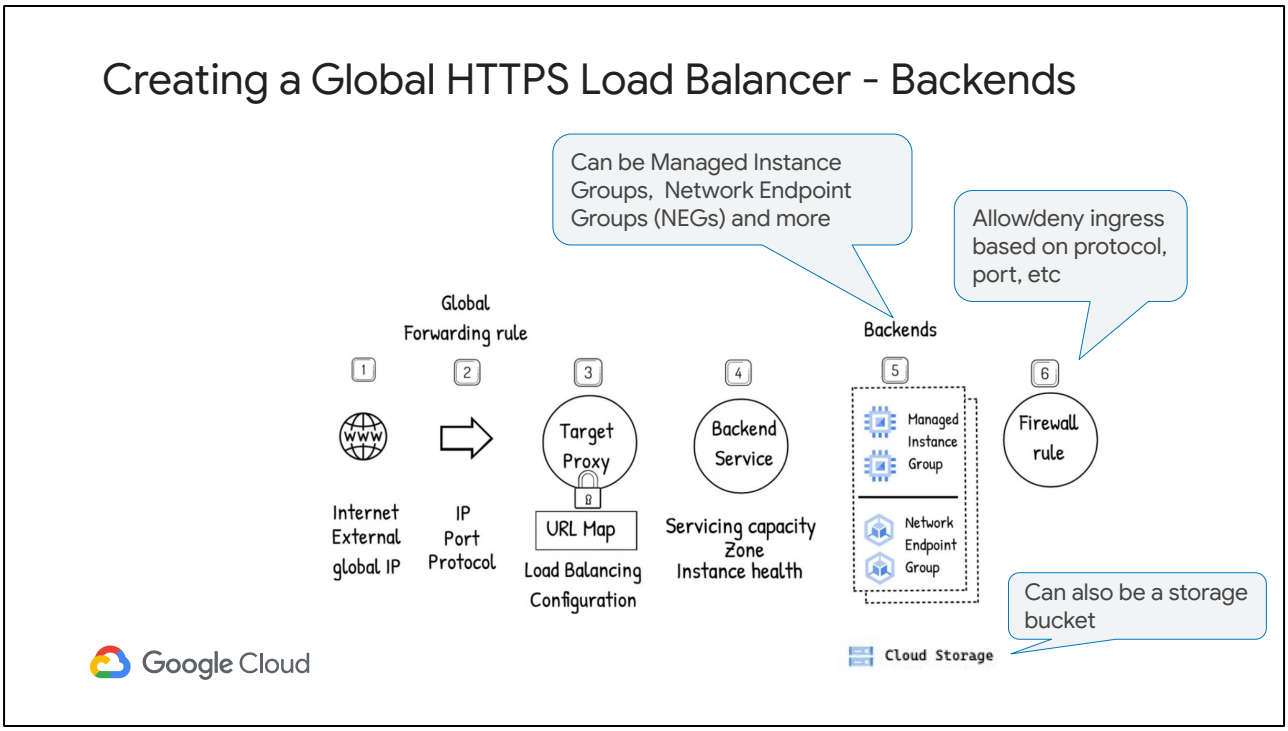


Image from video:
https://www.youtube.com/watch?v=0fQr7TRhnnU&list=PLTWE_Imu2InBzuPmOcgAYP7U80a87cpJd

Corresponding blog:
https://cloud.google.com/blog/topics/developers-practitioners/what-cloud-load-balancing

Backend options

- Backend - one or more endpoints that receive traffic from a load balancer
- Options include
  - Instance groups
  - Cloud Storage buckets
  - Network endpoint groups (NEG)
    - A group of backend endpoints or services
    - Common use case: Services running in containers
  - Zonal NEG
    - A group of VMs in the same network, subnet and zone
  - Serverless NEG
    - Cloud Run, Cloud Function, App Engine, API Gateway
  - Hybrid connectivity NEG
    - Routing traffic to an on-premises location or another cloud

Backend services overview

Backend services overview
https://cloud.google.com/load-balancing/docs/backend-service

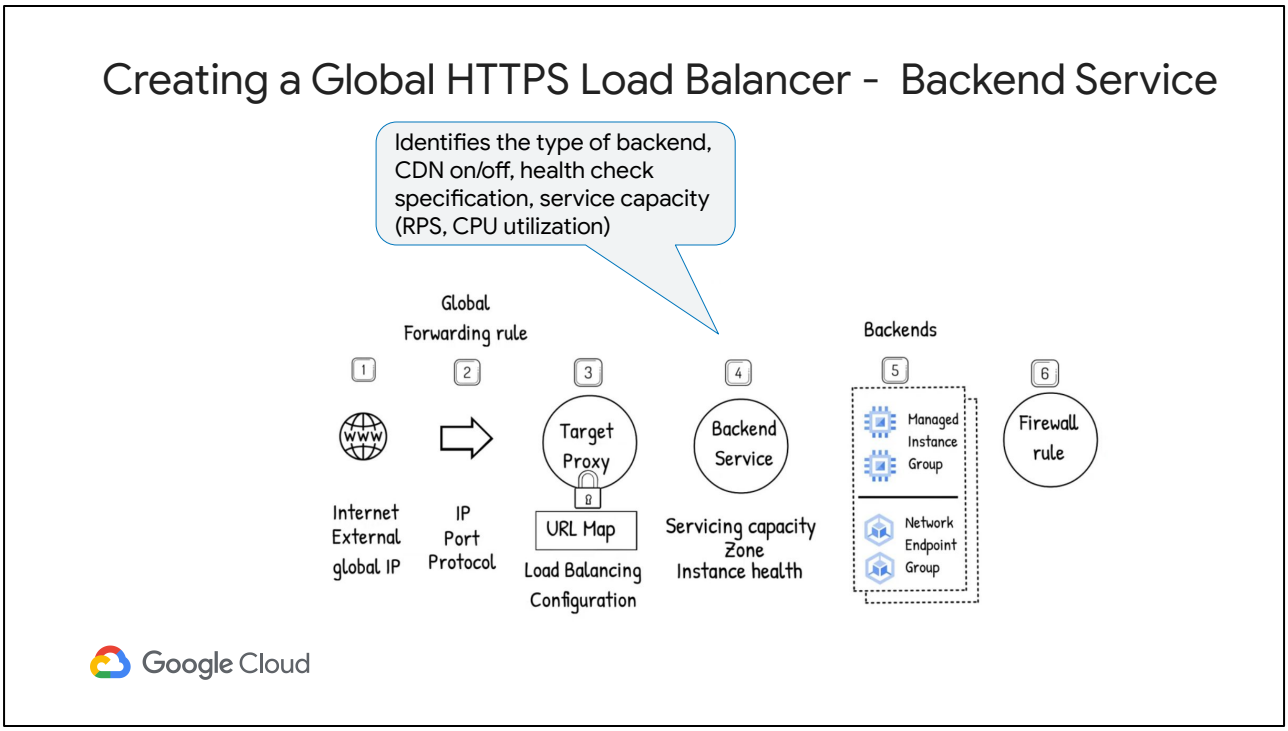
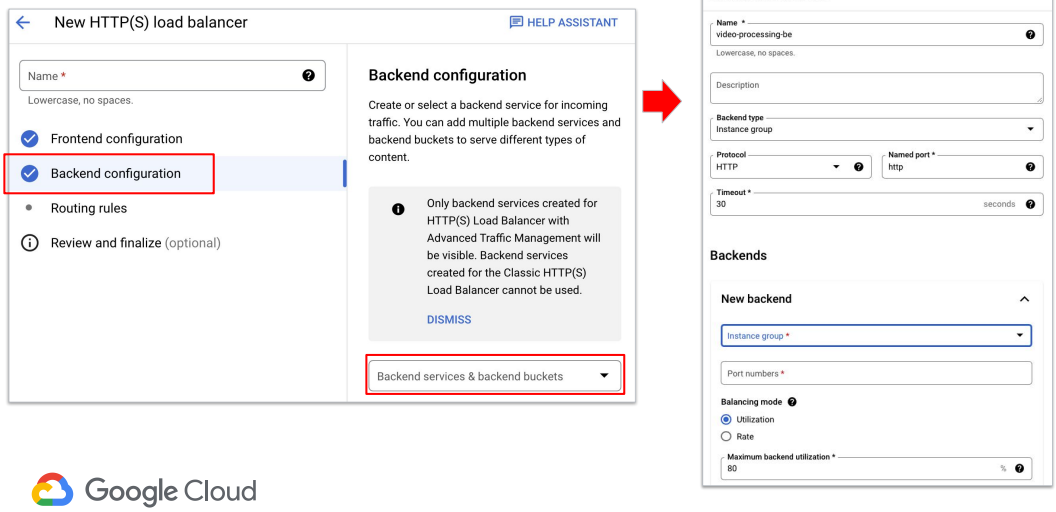


Image from video:
https://www.youtube.com/watch?v=0fQr7TRhnnU&list=PLTWE_Imu2InBzuPmOcgAYP7U80a87cpJd

Corresponding blog:
https://cloud.google.com/blog/topics/developers-practitioners/what-cloud-load-balancing

Setting up an Instance Group backend service



Google Cloud

Instance group backend services
https://cloud.google.com/load-balancing/docs/backend-service#instance_groups

Some backend services require health checks

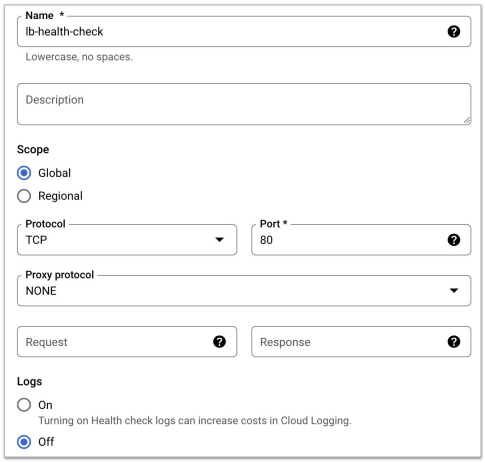
- Instance groups or zonal NEGs backend services must have an associated health check
  - Serverless NEG or an internet NEG backend services must **not** reference a health check
- HTTP(S) Load Balancer uses health checks to determine if a particular backend instance should receive traffic
  - An instance marked as “unhealthy” will not receive traffic
  - Managed Instance Groups also use health checks, but for a different purpose
    - If an instance is reported an unhealthy, it will be removed and replaced with a healthy instance

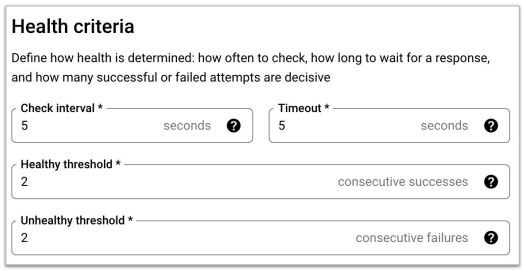
https://cloud.google.com/load-balancing/docs/backend-service#health-checks

Health checks
https://cloud.google.com/load-balancing/docs/backend-service#health-checks

Setting up a Health Check for an Instance Group Backend

- Used by the load balancer to determine whether to send traffic to a particular VM






```
gcloud compute health-checks create http
backend-basic-check --port=80 --no-enable-logging
--check-interval=5 --timeout=5
--unhealthy-threshold=3 --healthy-threshold=2
```

Google Cloud

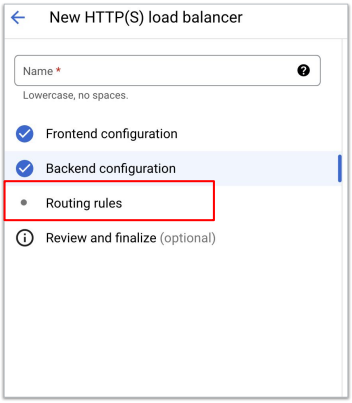
Health Checks need a firewall rule

- Health check probes come from addresses in the ranges 130.211.0.0/22 and 35.191.0.0/16

```
gcloud compute firewall-rules create default-allow-health-check \
 --direction=INGRESS --priority=1000 --network=default \
 --action=ALLOW --rules=tcp:80  \
 --source-ranges=130.211.0.0/22,35.191.0.0/16  \
 --target-tags=allow-health-check
```

Setting up Routing rules

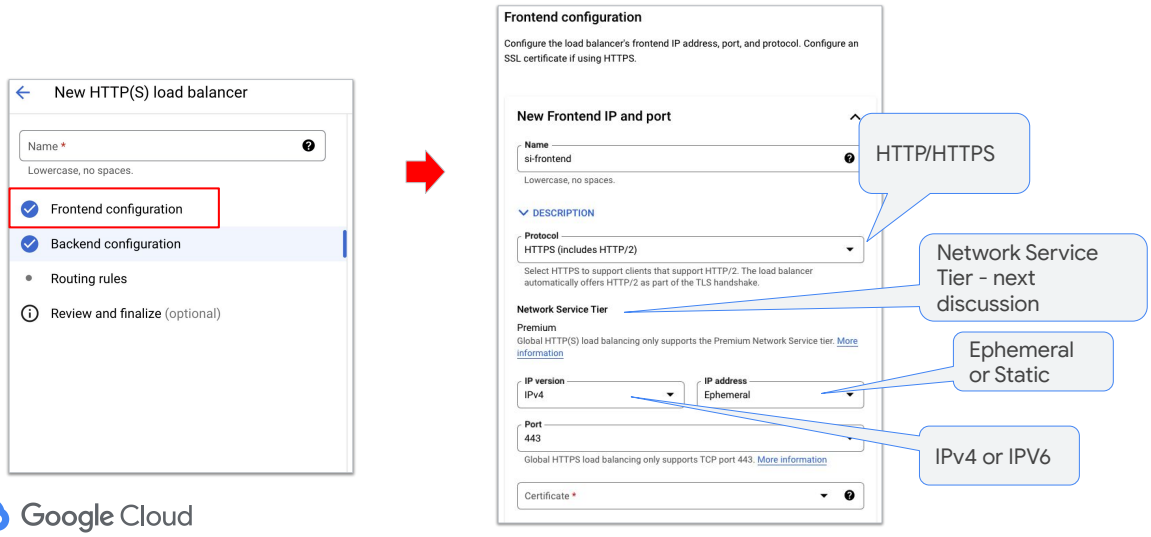
- Paths determines which backend service receives the traffic



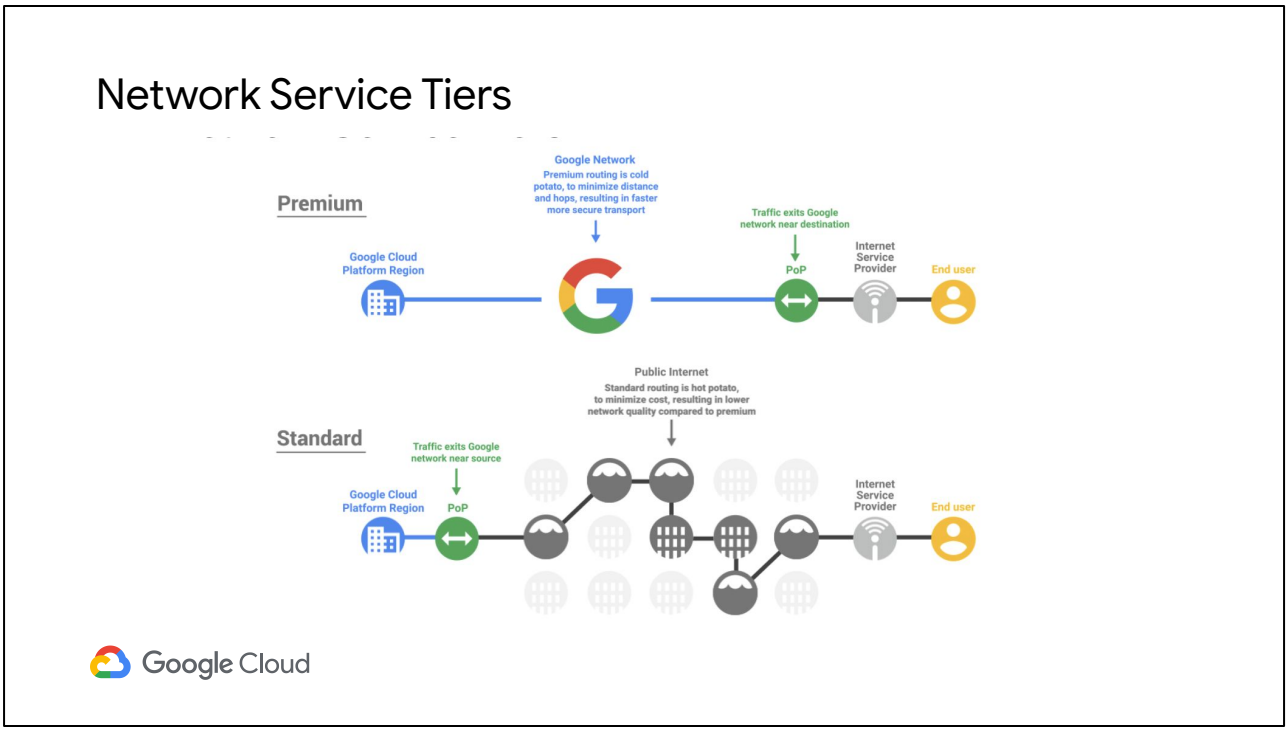
Google Cloud

URL maps overview
https://cloud.google.com/load-balancing/docs/url-map-concepts

Setting up the Frontend configuration



URL maps overview
https://cloud.google.com/load-balancing/docs/url-map-concepts



Premium Tier
https://cloud.google.com/network-tiers/docs/overview#premium_tier

Premium Tier delivers traffic from external systems to Google Cloud resources by using Google's low latency, highly reliable global network. This network consists of an extensive private fiber network with over 100 points of presence (PoPs) around the globe. This network is designed to tolerate multiple failures and disruptions while still delivering traffic.

Premium Tier supports both regional external IP addresses and global external IP addresses for VM instances and load balancers. All global external IP addresses must use Premium Tier. Applications that require high performance and availability, such as those that use HTTP(S), TCP proxy, and SSL proxy load balancers with backends in more than one region, require Premium Tier. Premium Tier is ideal for customers with users in multiple locations worldwide who need the best network performance and reliability.

Standard Tier
https://cloud.google.com/network-tiers/docs/overview#standard_tier

Standard Tier delivers traffic from external systems to Google Cloud resources by routing it over the internet. It leverages the double redundancy of Google's network only up to the point where Google's data center connects to a peering PoP. Packets that leave Google's network are delivered using the public internet and are subject to the reliability of intervening transit providers and ISPs. Standard Tier provides network

quality and reliability comparable to that of other cloud providers.

Standard Tier is priced lower than Premium Tier because traffic from systems on the internet is routed over transit (ISP) networks before being sent to VMs in your VPC network or regional Cloud Storage buckets. Standard Tier outbound traffic normally exits Google's network from the same region used by the sending VM or Cloud Storage bucket, regardless of its destination. In rare cases, such as during a network event, traffic might not be able to travel out the closest exit and might be sent out another exit, perhaps in another region.

Standard Tier offers a lower-cost alternative for the following use cases:

- You have applications that are not latency or performance sensitive.
- You're deploying VM instances or using Cloud Storage that can all be within a single region.

Network Service Tiers

Optimize your network for performance or cost

- **Premium**
  - Delivers traffic on Google's premium backbone
  - Optimized for performance, higher cost
  - For services needing global availability
  - Global HTTP(S) load balancing supports this tier only
  - Backed by SLA
- **Standard**
  - Delivers traffic using regular ISP networks
  - Optimized for cost, lower performance
  - For services hosted within a single region
  - No SLA

Premium
Network Tier

Standard
Network Tier

Google Cloud

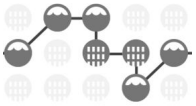
Network Service Tiers

Network Service Tiers
https://cloud.google.com/network-tiers

Network Service Tiers overview
https://cloud.google.com/network-tiers/docs/overview

Google Cloud networking in depth: Understanding Network Service Tiers (May 15, 2019)
https://cloud.google.com/blog/products/networking/google-cloud-networking-in-depth-understanding-network-service-tiers

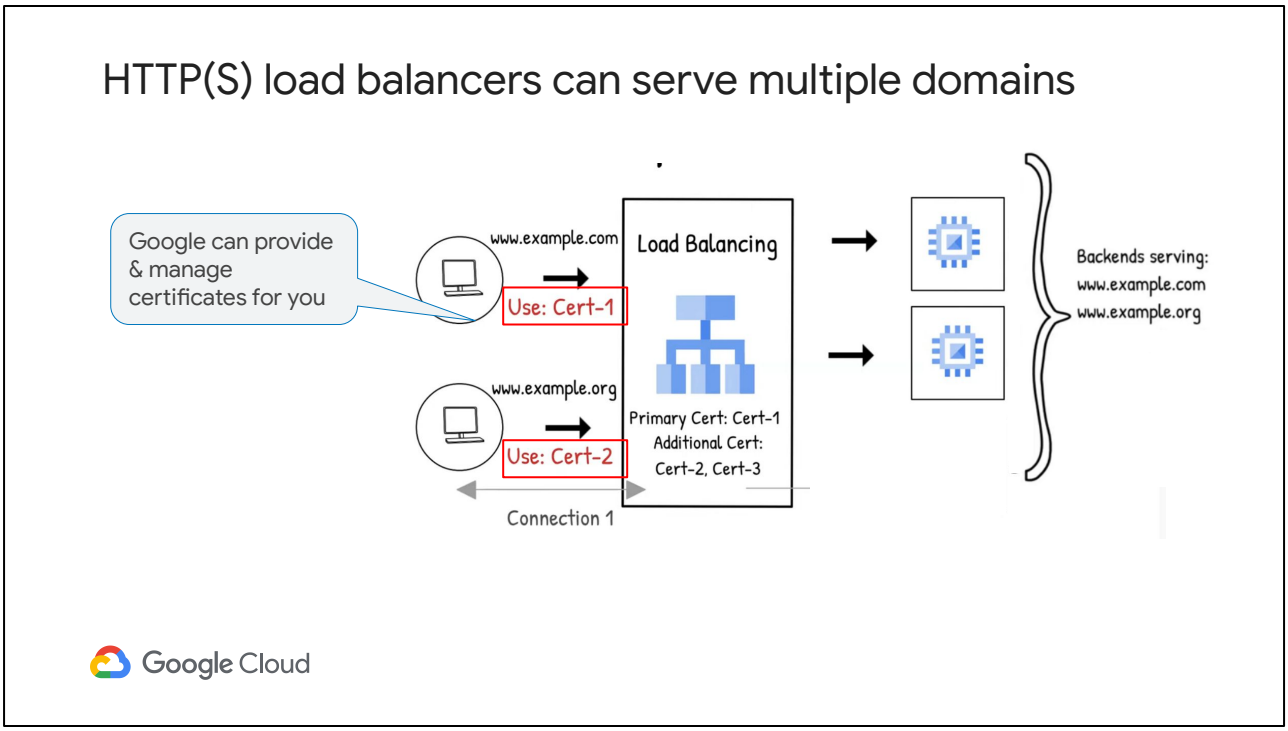
Network Service Tier summary

	Premium	Standard
Choice	 Global LB Performance-optimized	 Regional LB Cost-optimized
Why?	• Performance, Security, Reliability over Google's global private network • Unique to Google Cloud	• Lower Price (~24–33% cheaper) • Lower performance than Premium but comparable to major public clouds
When?	• Recommended Tier • Default for all workloads	• For cost-sensitive workloads • For performance and outbound costs comparable to other public clouds • For non-critical workloads in single region

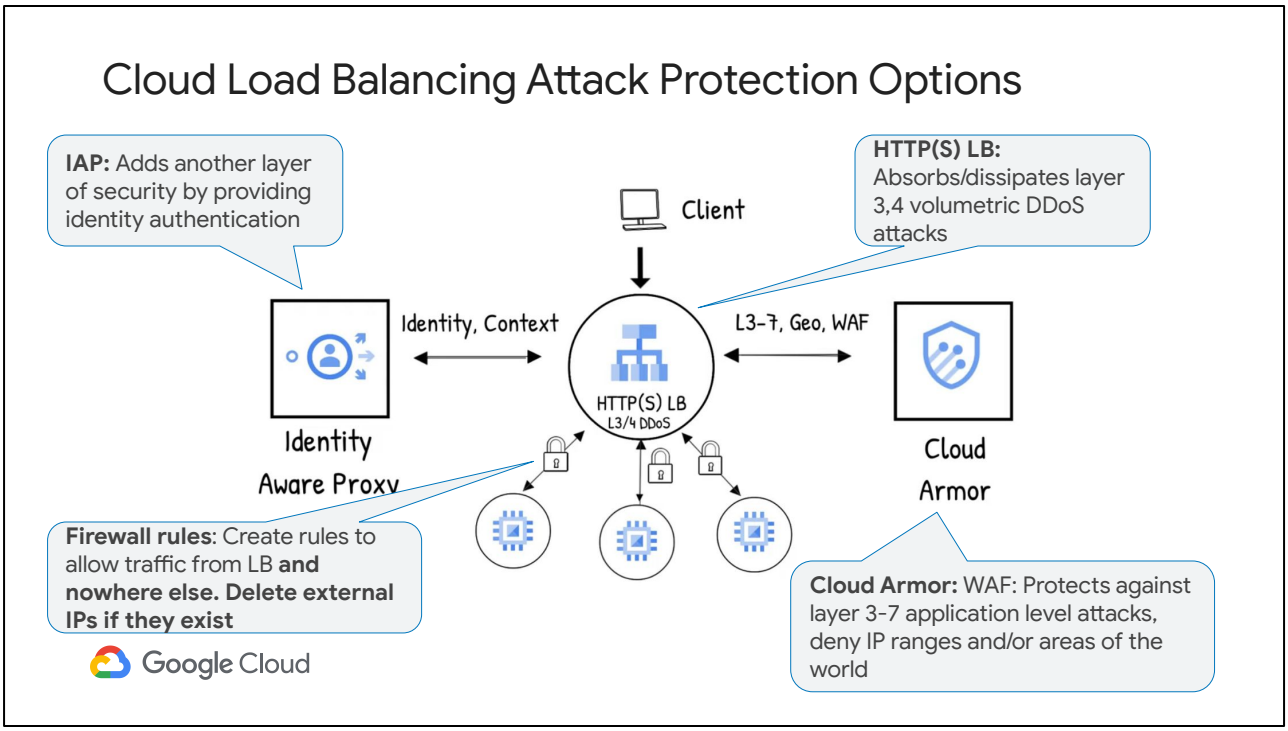
Google Cloud

Ask yourself whether high performance or lower cost is most important for your workload or resource. The Premium Tier is the clear choice for performance. If cost is your main consideration, remember that the Standard Tier has other restrictions in addition to network performance. If you want to deploy your backends or have users in multiple regions, but don't want to use the public internet over Google's network for inter-continental and cross region traffic, you want to choose the Premium Tier. Also, if you want Global Load Balancing or Cloud CDN, you need to use the Premium Tier.

Otherwise, the Standard Tier is a great choice if you don't need any of those services and are okay using the public internet instead of Google's network.

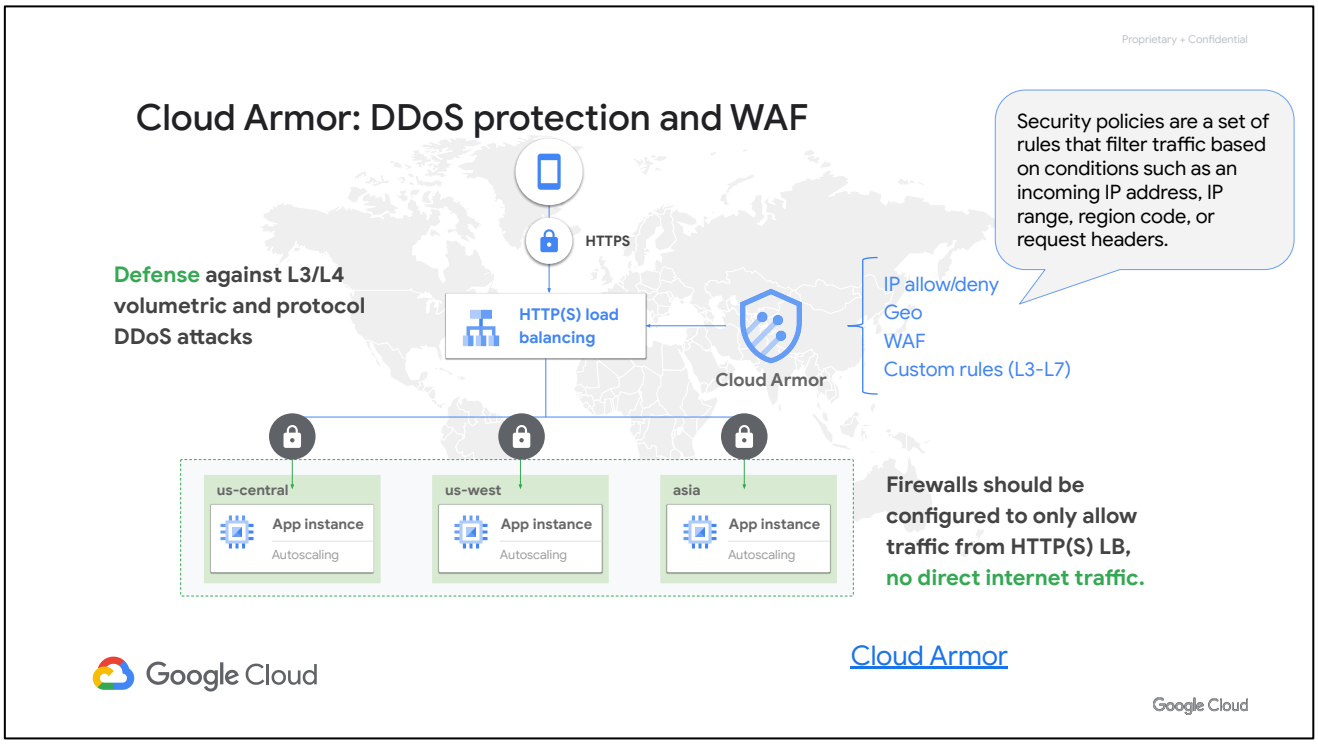


Cloud Load Balancing supports multiple SSL certificates, as well, if you wanted to serve multiple domains using the same load-balancing IP address and port.



WAF: Web application firewall

What is a volumetric attack? A volumetric attack sends a high amount of traffic, or request packets, to a targeted network in an effort to overwhelm its bandwidth capabilities. These attacks work to flood the target in the hopes of slowing or stopping their services.



Google Cloud Armor
https://cloud.google.com/armor

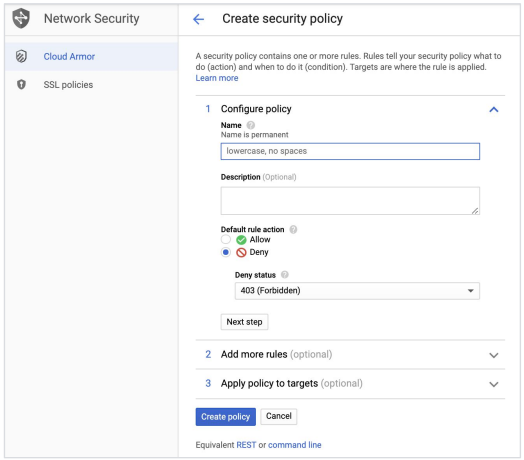
Key points:

- When using **External HTTP(s) LB,** you **benefit** from an unrivaled level of **L3/L4 DDoS protection** from a wide variety of attack types.
- **Cloud Armor**
  - **Managed WAF solution**
  - Runs **distributed over Google's edge**
  - Integrates with **External HTTP(s) Load Balancer**
  - **Provides**
    - **Layer 7 and Application layer** protection
    - **Access controls** to backend services based on **IP and Geo**
    - **Security policies** based on **L3 to L7 request and client attributes**
    - **Request throttling**
    - **Pre-configured** rules (XSS and SQLi based on OWASP Modsecurity)
    - **Real time telemetry** in the form of **Cloud Operations logs** containing Cloud Armor's decisions on a per-request basis, as well as a **monitoring dashboard** that gives granular views of

        - allow, denied, or previewed traffic.
    - **Security policies and IP deny list** are **not supported with Cloud CDN**

Use Google Cloud Armor to create network security policies

- Can allow or deny access to your Google Cloud resources using IP addresses or ranges.
- Create allow lists to allow known addresses.
- Create deny lists to block known attackers.

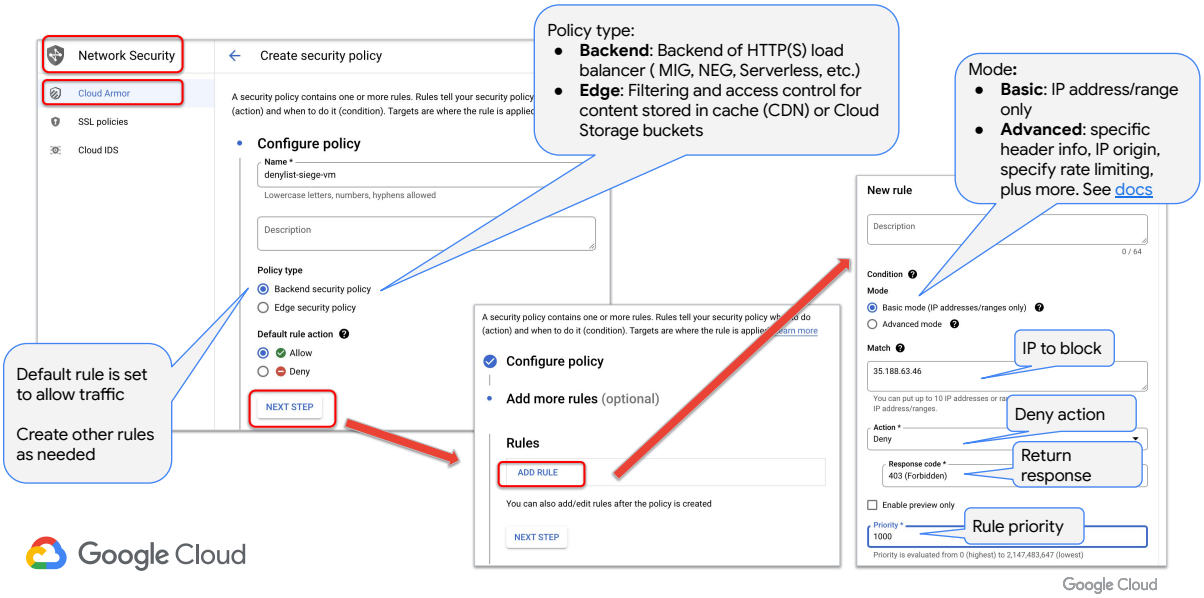


Google Cloud

For additional features over built-in DDoS, such as IPv4 and IPv6 allow or deny, and defense against application-aware attacks such as cross-site scripting and SQL injection, Google offers Google Cloud Armor, which works in conjunction with global HTTP/HTTPS load balancing and enables you to deploy and customize defenses for your internet-facing applications. It's based on the same technologies and global infrastructure that we use to protect Google services like Search, Gmail, and YouTube.

Google Cloud Armor security policies enable the access or denial of HTTP(S) requests to load balancers at the Google Cloud edge as close as possible to the source of incoming traffic. This prevents unwelcome traffic from consuming resources or entering the VPC networks.

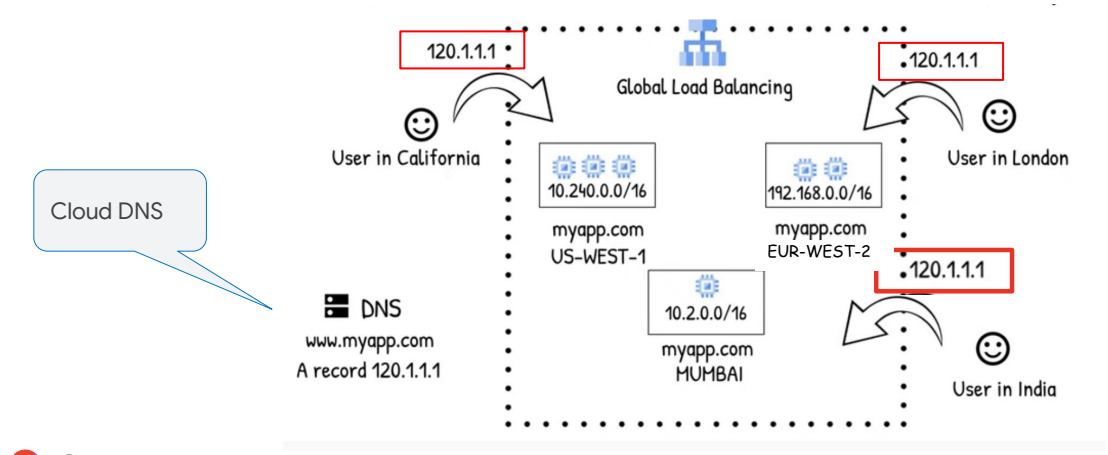
Cloud Armor example - Deny access to specific IP



Google Cloud

Google Cloud

Cloud DNS can be used to map Load Balancer IP to domain name

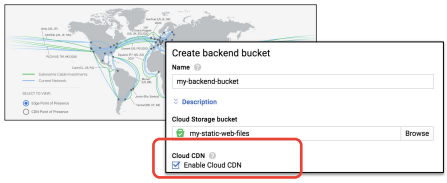


Google Cloud

https://cloud.google.com/dns

Google Cloud provides two Content Delivery Network (CDN) options

- Cloud CDN caches regularly accessed static content
  - Optimized for serving static web assets such as CSS, JavaScript, or non-dynamic HTML files
- Media CDN is a media delivery platform
  - Optimized for high-throughput egress workloads such as streaming video and large file downloads
- Both use Google's globally distributed Edge locations to **cache content** close to your users
- **170+ locations**, with **single IP** across multiple regions



Cloud CDN

Google Cloud

Choose a CDN product

The load balancer can also be configured to leverage Google's content delivery network, called Cloud CDN.

By enabling Cloud CDN, you can cache content all over the world, closer to the user.

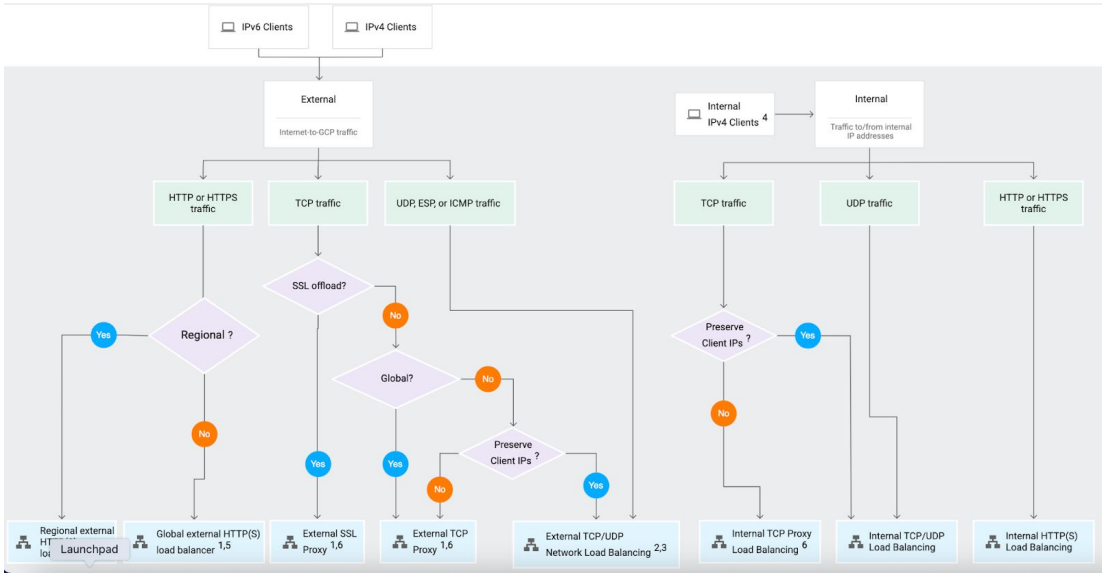
Introducing Media CDN—the modern extensible platform for delivering immersive experiences
https://cloud.google.com/blog/products/networking/introducing-media-cdn/

Choose a CDN product
https://cloud.google.com/media-cdn/docs/choose-cdn-product

How do you decide on the type of load balancer to use?

- Based on use case
- The right option depends on
  - Whether the traffic is Internal or external
  - The type of traffic, e.g. HTTP(S), TCP, UDP, etc
- Use global load balancing when
  - Backends are distributed across multiple regions
    - Provides a single anycast IP address, and supports IPv6 addresses
- Use regional load balancing when
  - Backends are in single region
  - Only require IPv4 termination
- Refer to Choosing a Load Balancer

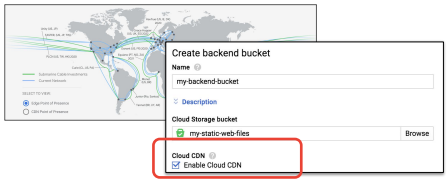
Load Balancer Decision Tree



Choosing a load balancer:
https://cloud.google.com/load-balancing/docs/choosing-load-balancer

Google Cloud provides two Content Delivery Network (CDN) options

- Cloud CDN caches regularly accessed static content
  - Optimized for serving static web assets such as CSS, JavaScript, or non-dynamic HTML files
- Media CDN is a media delivery platform
  - Optimized for high-throughput egress workloads such as streaming video and large file downloads
- Both use Google's globally distributed Edge locations to **cache content** close to your users
- **170+ locations**, with **single IP** across multiple regions



Cloud CDN

Google Cloud

Choose a CDN product

The load balancer can also be configured to leverage Google's content delivery network, called Cloud CDN.

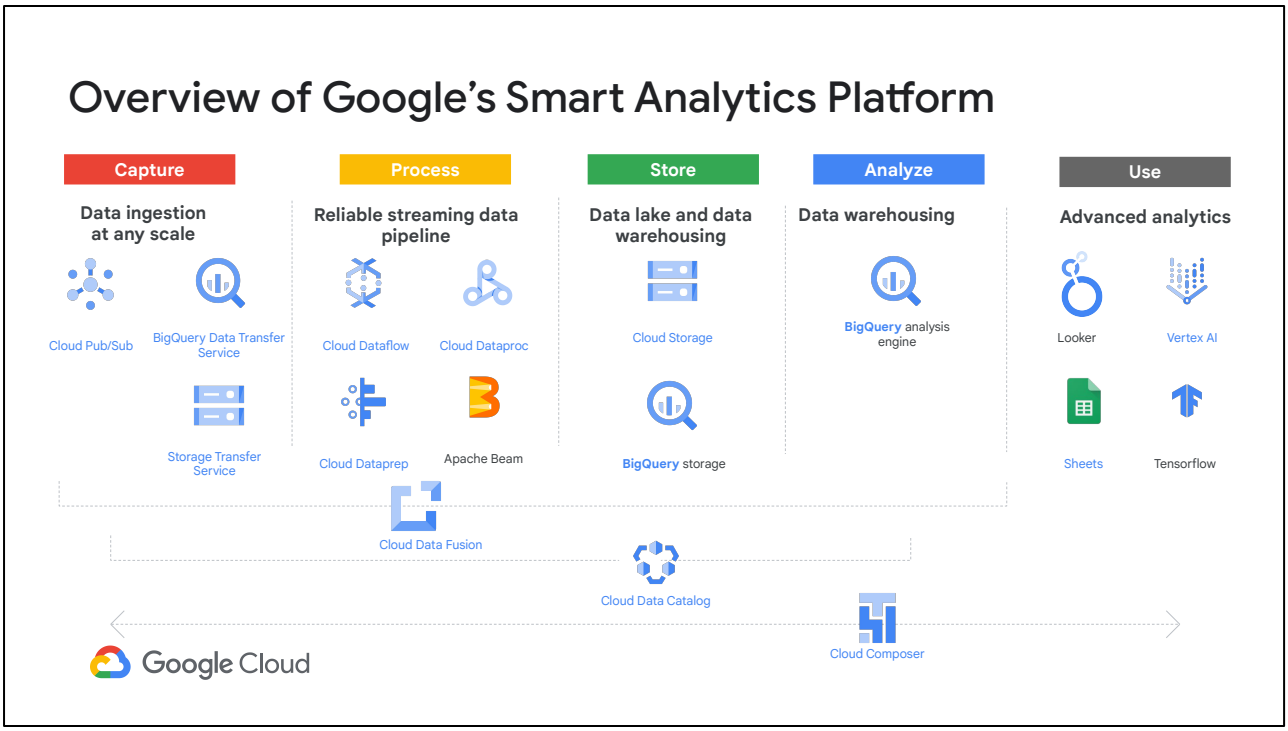
By enabling Cloud CDN, you can cache content all over the world, closer to the user.

Introducing Media CDN—the modern extensible platform for delivering immersive experiences
https://cloud.google.com/blog/products/networking/introducing-media-cdn/

Choose a CDN product
https://cloud.google.com/media-cdn/docs/choose-cdn-product

Data Analytics Pipeline...

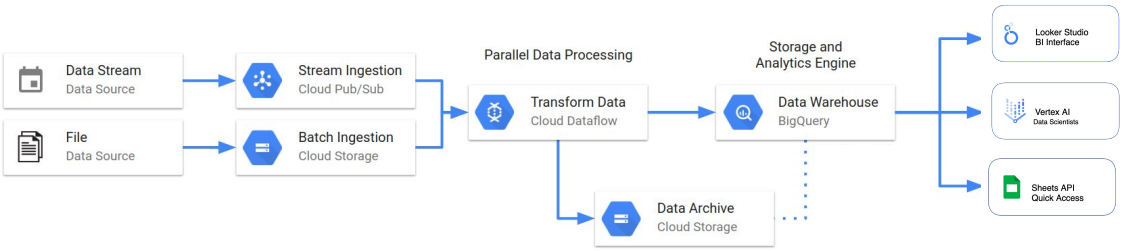
Google Cloud



BigQuery is part of Google Cloud's comprehensive data analytics platform that covers the analytics value chain from Ingest >> process >> store >> advanced analytics and collaboration. BigQuery is deeply integrated with the GCP's analytical and data processing offering, allowing customers to build an enterprise ready cloud native data warehouse.

1. Cloud Pub/sub - Scaled messaging platform
2. BigQuery Data Transfer Service - Ads data for marketing cloud
3. Beam - Stream and batch processing with single programing model with Dataflow
4. Dataproc - Managed Hadoop and Spark platform
5. Dataprep - Analyst can now do data prep using visual tool
6. Data Fusion - Fully managed, code-free data integration service to manage ETL/ELT pipelines and also track lineage of that data.
7. BigQuery cloud-native, highly scalable data warehouse
8. Cloud Storage as your data lake for structured and unstructured data
9. Vertex AI & Tensorflow for machine learning on top of data on BigQuery and Cloud Storage
10. Looker Studio and Sheet for your analysis

Typical Data Processing Pipeline

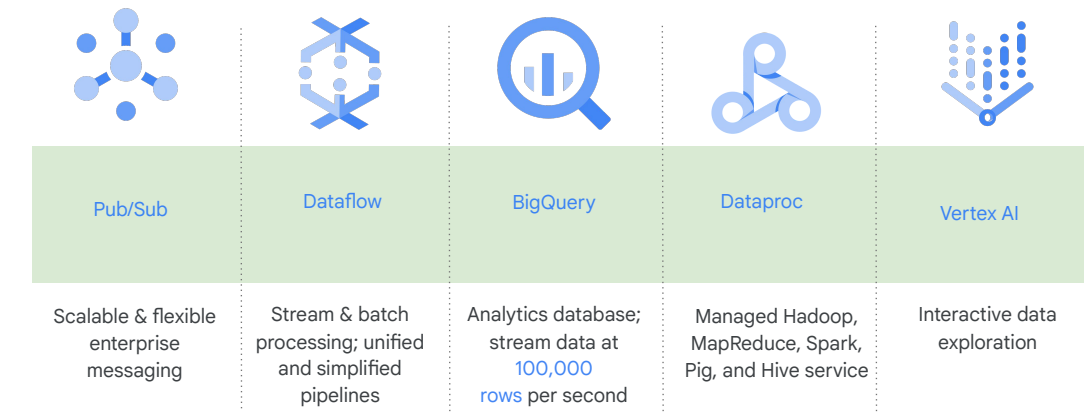


Google Cloud

How to Build a data pipeline with Google Cloud

How to Build a data pipeline with Google Cloud
https://www.youtube.com/watch?v=yVUXvabnMRU

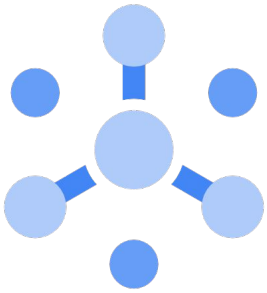
Google Cloud big data services are fully managed and scalable



Google Cloud Big Data solutions are designed to help you transform your business and user experiences with meaningful data insights. It is an integrated, serverless platform. “Serverless” means you don’t have to provision compute instances to run your jobs. The services are fully managed, and you pay only for the resources you consume. The platform is “integrated” so Google Cloud data services work together to help you create custom solutions.

Pub/Sub is scalable, reliable messaging

- Fully managed, massively scalable messaging service
  - It allows messages to be sent between independent applications
  - Can scale to millions of messages per second
- Messages are sent and received via HTTP(S)
- Supports multiple senders and receivers simultaneously
- Global service
  - Messages are copied to multiple zones for greater fault tolerance
  - Dedicated resources in every region for fast delivery worldwide
- Pub/Sub messages are encrypted at rest and in transit



Google Cloud

What is Cloud Pub/Sub?

What is Cloud Pub/Sub?
https://www.youtube.com/watch?v=JrKEErlWvzA&list=PLTWE_Imu2InBzuPmOcgAYP7U80a87cpJd

Cloud Pub/Sub is a fully managed, massively scalable messaging service that can be configured to send messages between independent applications, and can scale to millions of messages per second.

Pub/Sub messages can be sent and received via HTTP and HTTPS.

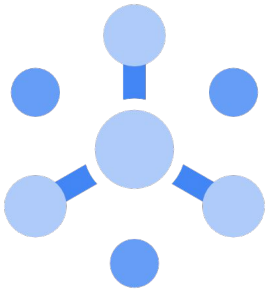
It also supports multiple senders and receivers simultaneously.

Pub/Sub is a global service. Fault tolerance is achieved by copying the message to multiple zones and using dedicated resources in every region for fast worldwide delivery.

All Pub/Sub messages are encrypted at rest and in transit.

Why use Pub/Sub?

- Building block for data ingestion in Dataflow, Internet of Things (IoT), Marketing Analytics, etc.
- Provides push notifications for cloud-based applications.
- Connects applications across Google Cloud (push/pull between components (e.g. GCE and App Engine)



Google Cloud

Pub/Sub is an important building block for applications where data arrives at high and unpredictable rates, like Internet of Things systems. If you're analyzing streaming data, Dataflow is a natural pairing with Pub/Sub.

Pub/Sub also works well with applications built on Google Cloud's compute platforms. You can configure your subscribers to receive messages on a "push" or a "pull" basis. In other words, subscribers can get notified when new messages arrive for them, or they can check for new messages at intervals.

Pub/Sub - customer use case

- Sky is one of Europe's leading media and communications companies, providing Sky TV, streaming, mobile TV, broadband, talk, and line rental services to millions of customers in seven countries
- **Pub/Sub** is used to stream diagnostic data from millions of Sky Q TV boxes
- Data is then parsed through **Cloud Dataflow** to **Cloud Storage** and **BigQuery**, monitored on its way by Stackdriver (**Operations**), which triggers email and Slack alerts should issues occur.

"This fully functioning Google Cloud solution will act as a blueprint for future Sky projects. We can capture all diagnostic data in Google Cloud and use it to inform our future strategy. Sky management sees this as the beginning of a new era in data management, analytics, and data science."

—Oliver Tweedie, Director of Data Engineering, Sky

Sky: Scaling for success with Sky Q diagnostics

Dataproc is managed Hadoop

- Fast, easy, managed way to run Hadoop and Spark/Hive/Pig on Google Cloud
- Create clusters in 90 seconds or less on average
- Scale clusters up and down even when jobs are running
- Dataproc storage
  - Automatically installs the HDFS-compatible Cloud Storage connector
    - Run Apache Hadoop or Apache Spark jobs directly on data in Cloud Storage
  - Alternatively can use boot disks to store data
    - Deleted when the Dataproc cluster is deleted

Google Cloud

Overview
https://cloud.google.com/dataproc

Dataproc Cloud Storage connector
https://cloud.google.com/dataproc/docs/concepts/connectors/cloud-storage

Customer use case: Best practices for migrating Hadoop to Dataproc by LiveRamp
https://cloud.google.com/blog/products/data-analytics/best-practices-for-migrating-hadoop-to-gcp-dataproc

Apache Hadoop is an open-source framework for big data. It is based on the MapReduce programming model, which Google invented and published. The MapReduce model, at its simplest, means that one function -- traditionally called the “map” function -- runs in parallel across a massive dataset to produce intermediate results; and another function -- traditionally called the “reduce” function -- builds a final result set based on all those intermediate results. The term “Hadoop” is often used informally to encompass Apache Hadoop itself and related projects, such as Apache Spark, Apache Pig, and Apache Hive.

Dataproc is a fast, easy, managed way to run Hadoop, Spark, Hive, and Pig on Google Cloud. All you have to do is to request a Hadoop cluster. It will be built for you

in 90 seconds or less, on top of Compute Engine virtual machines whose number and type you can control. If you need more or less processing power while your cluster's running, you can scale it up or down. You can use the default configuration for the Hadoop software in your cluster, or you can customize it. And you can monitor your cluster using Operations.

Hadoop history

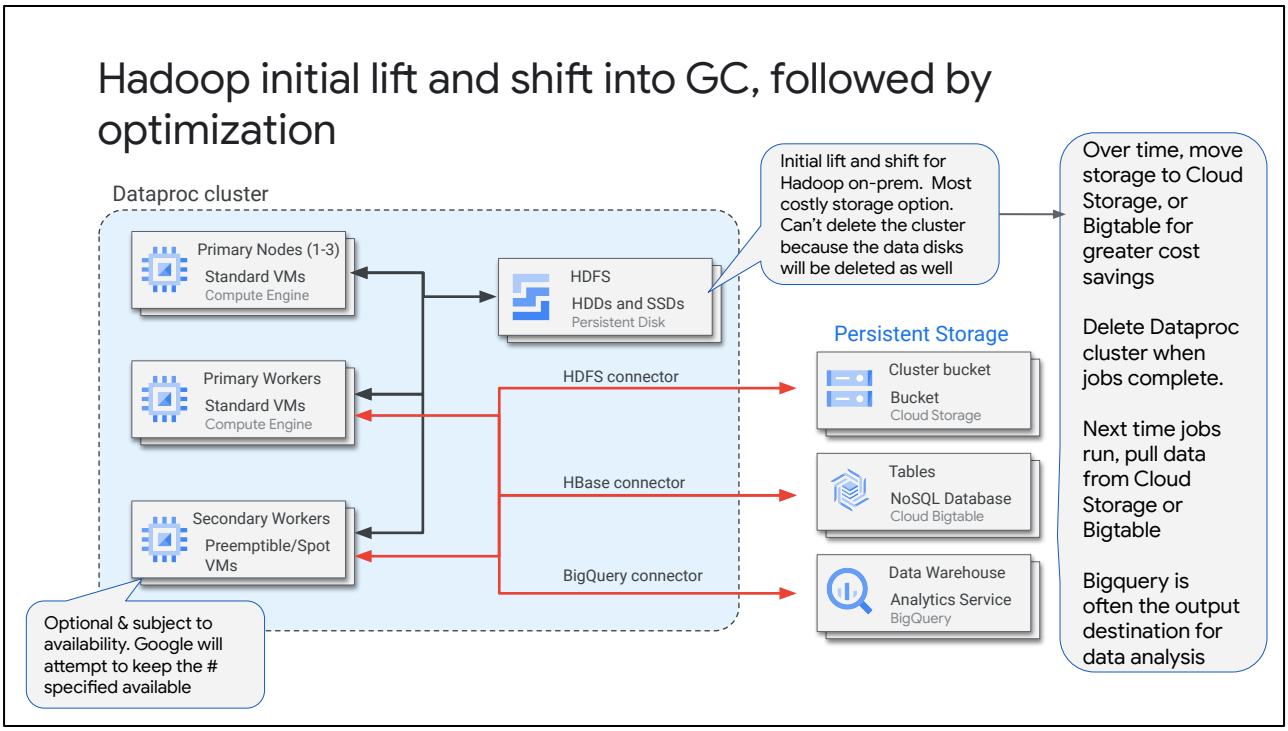
- Google and Yahoo were looking for ways to analyze mountains of user internet search results
  - Google published a white paper in 2004
  - Yahoo implemented the concepts and open sourced it in 2008
- Two main components when running on-premise
  - Multiple nodes (VMs) to process data
    - May consist of 1,000s of nodes
    - Shares computational workloads and works on data in parallel
  - Hadoop Distributed File System (HDFS) to store the data
    - Persistent disk storage

Google's whitepaper:
https://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

Yahoo released Hadoop as an open source project to Apache Software Foundation in 2008

Example use case - clickstream data

- Websites track every click made by every user on every page visited
  - Results in millions of rows of data
- Analysts would like to know why someone added items to a shopping cart, proceeded to checkout, and then abandoned the cart
  - Don't need *all* the data - just data for users who abandoned their cart
    - Phase 1 (aka "Map")
      - Process the data and get the users, the contents of their carts and the last page visited
    - Phase 2 (aka "Reduce")
      - Aggregate the total the number and value of carts abandoned per month
      - Plus total the most common final pages that someone viewed before ending the user session



The lines in black are usually the initial implementation. Customers gain greater cost savings when they transition to the red flow. This requires making some modifications to the jobs to use the connectors.

HDFS: Hadoop data file system. Cloud storage was originally named DFS (distributed file system)
HBase: open-source, NoSQL, distributed big data store. Runs on top of HDFS. The GC equivalent is Bigtable

DFS - distributed file system = Cloud Storage

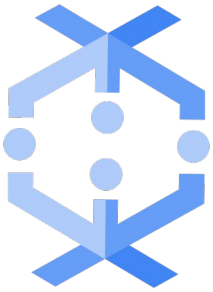
HA Dataproc has 3 masters (one is a witness); With no HA , have 1 master, no failover. All are in the same zone

Primary workers can use autoscaling. Secondary workers are MIGs but do not scale. However if you tell Google you want 4 workers, it will try to keep 4 workers at al times. These can be spot VMs

HBASE - database that lives in HDFS. Equivalent to Cloud Bigtable

Dataflow offers managed data pipelines

- Processes data using Compute Engine instances.
  - Clusters are sized for you
  - Automated scaling, no instance provisioning required
- Write code once and get batch and streaming
  - Transform-based programming model



Google Cloud

Dataflow, the backbone of data analytics

Dataflow, the backbone of data analytics
https://cloud.google.com/blog/topics/developers-practitioners/dataflow-backbone-data-analytics

Dataflow Under the Hood: Comparing Dataflow with other tools (Aug 24, 2020)
https://cloud.google.com/blog/products/data-analytics/dataflow-vs-other-stream-batch-processing-engines

Dataproc is great when you have a dataset of known size, or when you want to manage your cluster size yourself. But what if your data shows up in realtime? Or it's of unpredictable size or rate? That's where Dataflow is a particularly good choice. It's both a unified programming model and a managed service, and it lets you develop and execute a big range of data processing patterns: extract-transform-and-load, batch computation, and continuous computation. You use Dataflow to build data pipelines, and the same pipelines work for both batch and streaming data.

Dataflow is a unified programming model and a managed service for developing and executing a wide range of data processing patterns including ETL, batch computation, and continuous computation. Dataflow frees you from operational tasks like resource management and performance optimization.

Dataflow features:

Resource Management: Dataflow fully automates management of required processing resources. No more spinning up instances by hand.

On Demand: All resources are provided on demand, enabling you to scale to meet your business needs. No need to buy reserved compute instances.

Intelligent Work Scheduling: Automated and optimized work partitioning which can dynamically rebalance lagging work. No more chasing down “hot keys” or pre-processing your input data.

Auto Scaling: Horizontal auto scaling of worker resources to meet optimum throughput requirements results in better overall price-to-performance.

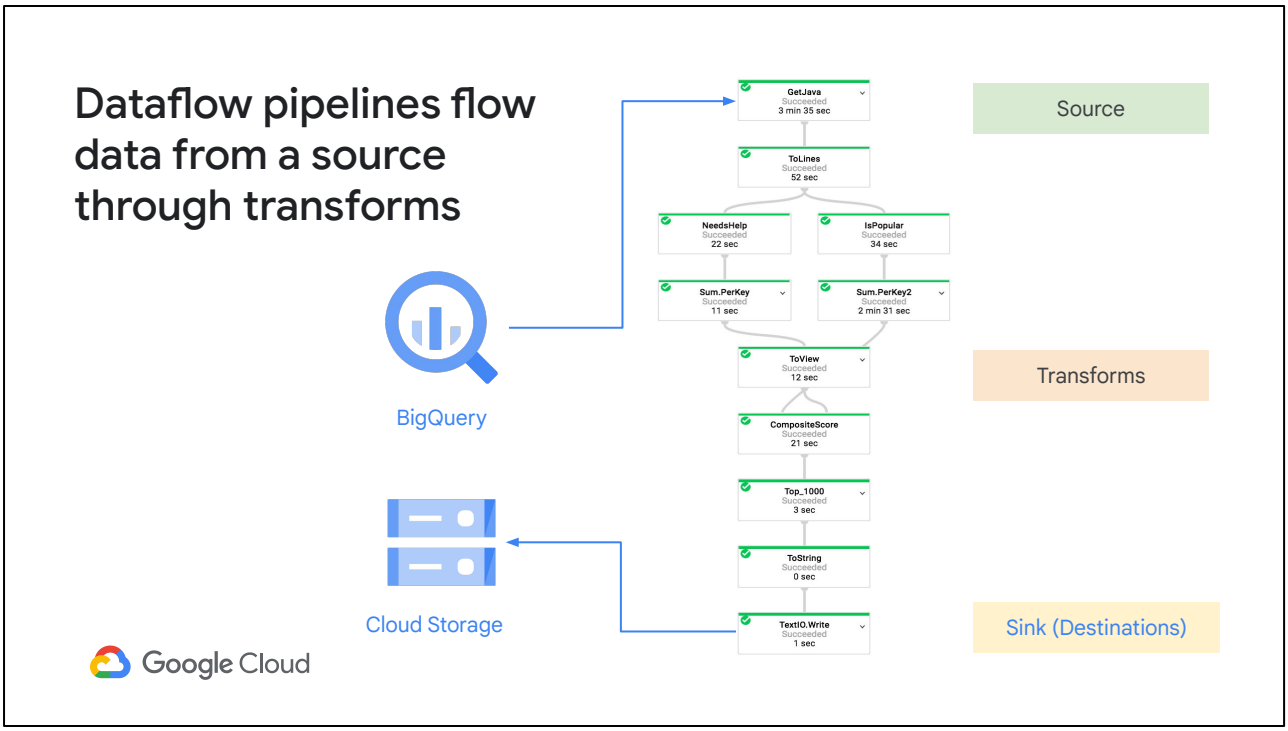
Unified Programming Model: The Dataflow API enables you to express MapReduce like operations, powerful data windowing, and fine grained correctness control regardless of data source.

Open Source: Developers wishing to extend the Dataflow programming model can fork and or submit pull requests on the Java-based Dataflow SDK. Dataflow pipelines can also run on alternate runtimes like Spark and Flink.

Monitoring: Integrated into the Cloud Console, Dataflow provides statistics such as pipeline throughput and lag, as well as consolidated worker log inspection—all in near-real time.

Integrated: Integrates with Cloud Storage, Pub/Sub, Datastore, Cloud Bigtable, and BigQuery for seamless data processing. And can be extended to interact with others sources and sinks like Apache Kafka and HDFS.

Reliable & Consistent Processing: Dataflow provides built-in support for fault-tolerant execution that is consistent and correct regardless of data size, cluster size, processing pattern or pipeline complexity.

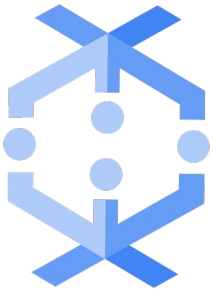


This example Dataflow pipeline reads data from a BigQuery table (the “source”), processes it in various ways (the “transforms”), and writes its output to Cloud Storage (the “sink”). Some of those transforms you see here are map operations, and some are reduce operations. You can build really expressive pipelines.

Each step in the pipeline is elastically scaled. There is no need to launch and manage a cluster. Instead, the service provides all resources on demand. It has automated and optimized work partitioning built in, which can dynamically rebalance lagging work. That reduces the need to worry about “hot keys” -- that is, situations where disproportionately large chunks of your input get mapped to the same custer.

Why use Dataflow?

- *ETL* (extract/transform/load) pipelines to move, filter, enrich, shape data
- *Data analysis*: batch computation or continuous computation using streaming
- *Orchestration*: create pipelines that coordinate services, including external services
- Integrates with Google Cloud services like Cloud Storage, Pub/Sub, BigQuery, and Cloud Bigtable
  - Open source Java, Python SDKs



Google Cloud

People use Dataflow in a variety of use cases. For one, it serves well as a general-purpose ETL tool.

And its use case as a data analysis engine comes in handy in things like these: fraud detection in financial services; IoT analytics in manufacturing, healthcare, and logistics; and clickstream, Point-of-Sale, and segmentation analysis in retail.

And, because those pipelines we saw can orchestrate multiple services, even external services, it can be used in real time applications such as personalizing gaming user experiences.

Dataprep

- Allows data analysts, business analysts, data engineers, and data scientists to **visually** explore, clean, and prepare big data
- Connects to BigQuery, Cloud Storage, Google Sheets, and hundreds of other cloud applications and traditional databases
- Build on top of Dataflow and BigQuery
    - Data transformation and cleaning rules can easily translate into Dataflow jobs or BigQuery SQL statements

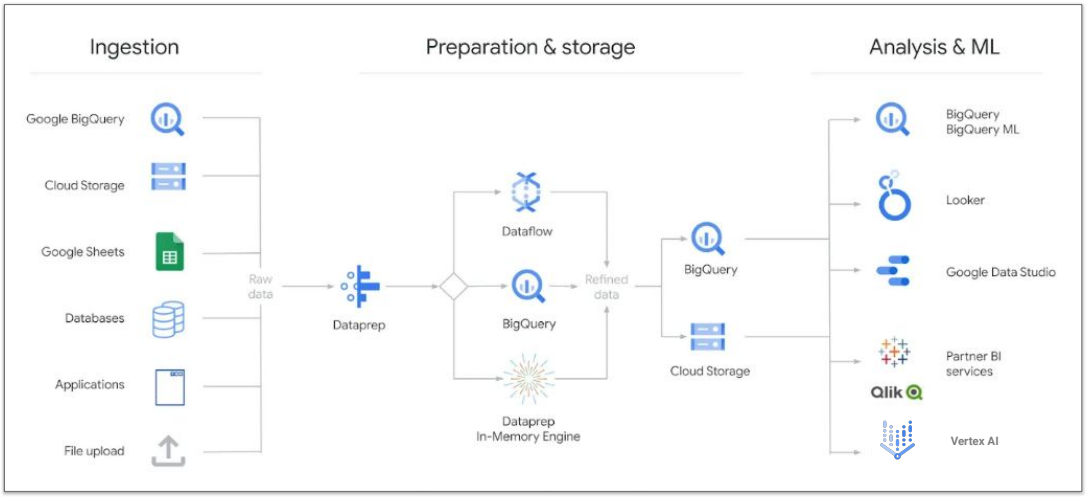
Google Cloud

Dataprep by Trifacta
https://cloud.google.com/dataprep

Dataprep cheat sheet
https://cloud.google.com/blog/topics/developers-practitioners/google-cloud-dataprep-trifacta-cheat-sheet

Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis, reporting, and machine learning. Because Dataprep is serverless and works at any scale, there is no infrastructure to deploy or manage. Your next ideal data transformation is suggested and predicted with each UI input, so you don't have to write code.

Dataprep ELT pipeline architecture



Google Cloud

https://cloud.google.com/dataprep

Once you've defined your sequence of transformations, Dataprep uses Dataflow or BigQuery under the hood, enabling you to process structured or unstructured datasets of any size with the ease of clicks, not code.

BigQuery is a fully managed data warehouse

- Provides near real-time interactive analysis of massive datasets (hundreds of TBs).
- Query using SQL syntax (ANSI SQL 2011)
- No cluster maintenance is required
- Compute and storage are separated with a terabit network in between.
- You only pay for storage and processing used.
- Automatic discount for long-term data storage.

Google Cloud

https://cloud.google.com/bigquery

Bigquery
https://cloud.google.com/blog/topics/developers-practitioners/query-big-bigquery-cheat-sheet

If, instead of a dynamic pipeline, you want to do ad-hoc SQL queries on a massive dataset, that is what BigQuery is for. BigQuery is Google's fully managed, petabyte scale, low cost analytics data warehouse.

BigQuery is NoOps: there is no infrastructure to manage and you don't need a database administrator, so you can focus on analyzing data to find meaningful insights, use familiar SQL, and take advantage of our pay-as-you-go model. BigQuery is a powerful big data analytics platform used by all types of organizations, from startups to Fortune 500 companies.

BigQuery's features:

Flexible Data Ingestion: Load your data from Cloud Storage or Datastore, or stream it into BigQuery at 100,000 rows per second to enable real-time analysis of your data.

Global Availability: You have the option to store your BigQuery data in European locations while continuing to benefit from a fully managed service, now with the option of geographic data control, without low-level cluster maintenance.

Security and Permissions: You have full control over who has access to the data

stored in BigQuery. If you share datasets, doing so will not impact your cost or performance; those you share with pay for their own queries.

Cost Controls: BigQuery provides cost control mechanisms that enable you to cap your daily costs at an amount that you choose. For more information, see Cost Controls.

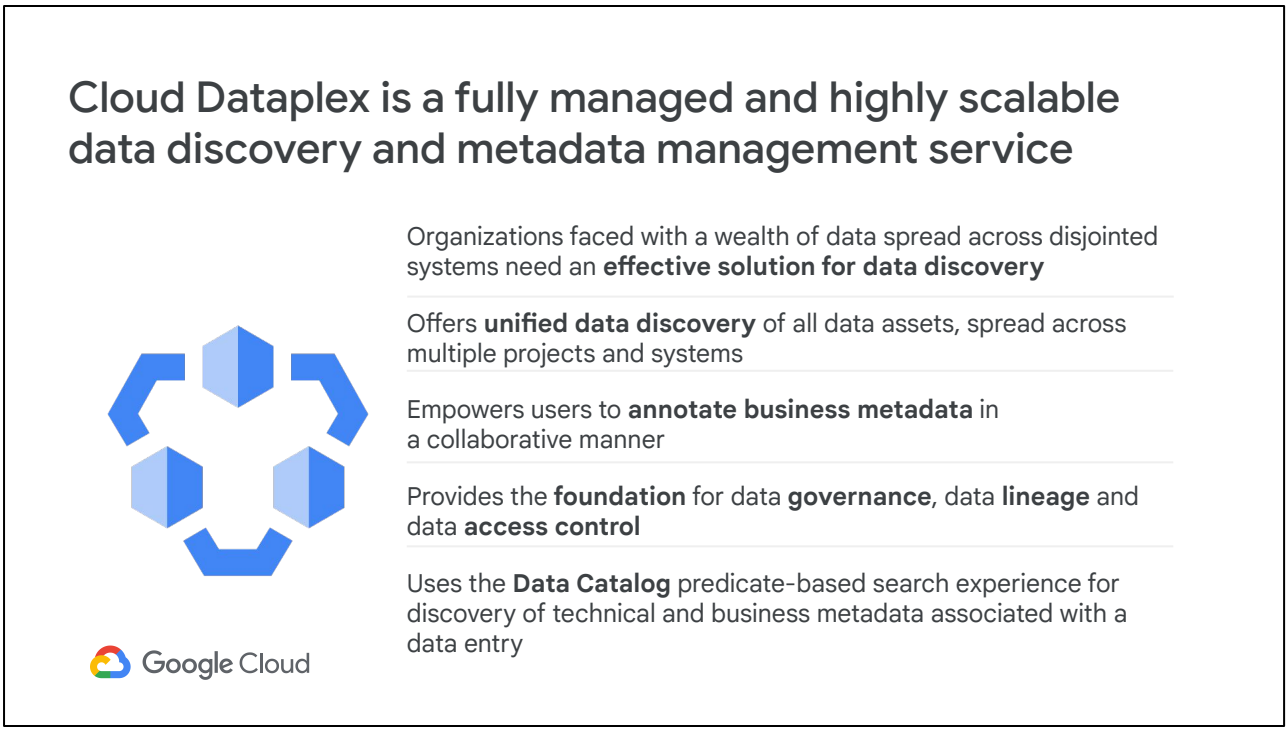
Highly Available: Transparent data replication in multiple geographies means that your data is available and durable even in the case of extreme failure modes.

Super Fast Performance: Run super-fast SQL queries against multiple terabytes of data in seconds, using the processing power of Google's infrastructure.

Fully Integrated In addition to SQL queries, you can easily read and write data in BigQuery via Dataflow, Spark, and Hadoop.

Connect with Google Products: You can automatically export your data from Google Analytics Premium into BigQuery and analyze datasets stored in Google Cloud Storage, Google Drive, and Google Sheets.

BigQuery can make Create, Replace, Update, and Delete changes to databases, subject to some limitations and with certain known issues.



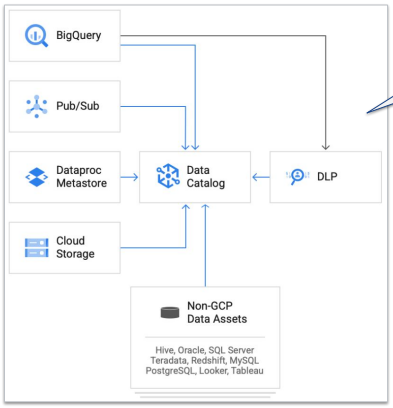
Dataplex: https://cloud.google.com/dataplex
Data Catalog: https://cloud.google.com/data-catalog/docs/concepts/overview
Blog: https://cloud.google.com/blog/products/data-analytics/manage-and-govern-data-with-the-unified-dataplex-and-data-catalog

Use Dataplex to organize, manage, and govern data

From the blog: Prior to this unification, data owners, stewards and governors had to use two different interfaces - Dataplex to organize, manage, and govern their data, and Data Catalog to discover, understand, and enrich their data. Now with this unification, we are creating a single coherent user experience where customers can now automatically discover and catalog all the data they own, understand data lineage, check for data quality, augment with business knowledge, organize data into domains, and then use that combined metadata to power data management.

Cloud Data Catalog

- A fully managed, scalable metadata management service within Dataplex



Can scan Google Cloud services and external data

Example of the kind of metadata that might be kept for BigQuery table

Field display name	Field ID	Required field	Type
Source of data asset	source	Yes	String
Rows in the asset	num_rows	No	Double
Has PII	has_pii	No	Boolean
PII type	pii_type	No	Enumerated Add values `EMAIL_ADDRESS`, `US_SOCIAL_SECURITY_NUMBER`, and `NONE`.
Context	context	No	Richtext

Google Cloud

Protect sensitive data with Data Loss Prevention

- Find Personally Identifiable Information (PII), such as name, email, address, and credit card number
- Allow custom data types to be defined
- Works with data in Cloud Storage, BigQuery, Firestore, and other locations, including data outside Google Cloud
- Works with structured and semistructured data (.pdf, .docx, .csv, .pptx, etc.) as well as unstructured data such as images (.bmp, .jpeg, .png, etc.)
- Optionally redact (natively with DLP; sensitive information (replacement character, encrypted version of the data, hash, etc.)

https://cloud.google.com/dlp

Cloud DLP in action

Fully managed service to discover, classify, and protect sensitive data

[Improving security, compliance, and governance]

Before

ID	Job Title	Phone	Comments
359740	Senior Engineer	307-964-0673	Please email them at jane@imadethisup.com
981587	VP, Engineer	713-910-6787	none
394091	Lawyer	692-398-4146	Updated phone to: 692-398-4146
986941	Senior Ops Manager	294-967-5508	none
490456	Junior Ops Manager	791-954-3281	Tried to verify account with their SSN 222-44-5555

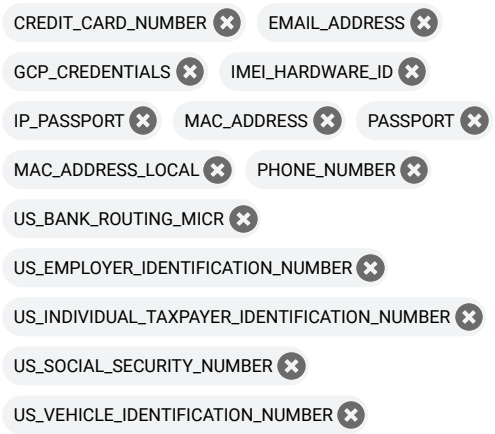
After

ID (FPE)	Job Title	Phone	Comments
438422	Engineer	307-###-####	Please email them at [Found Email]
530375	Engineer	713-###-####	none
496534	Lawyer	692-###-####	Updated phone to: 692-###-####
242348	Ops	294-###-####	none
593887	Ops	791-###-####	Tried to verify account with their SSN [Found SSN]

Google Cloud

Data Loss Prevention API protects sensitive data

- Scans data in Cloud Storage, BigQuery, or Datastore
- Can also scan images.
- Detects many different types of sensitive data, including:
  - Emails
  - Credit cards
  - Tax IDs
- You can add your own information types.
- Can delete, mask, tokenize, or just identify the location of the sensitive data.



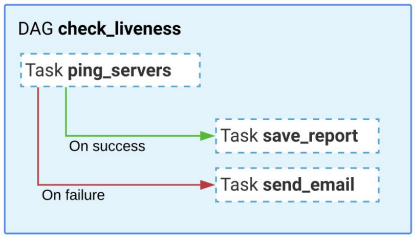
Google Cloud

Cloud DLP helps users better understand and manage sensitive data. It provides fast, scalable classification and redaction for sensitive data elements like credit card numbers, names, social security numbers, US and selected international identifier numbers, phone numbers, and Google Cloud credentials. Cloud DLP classifies this data using more than 90 predefined detectors to identify patterns, formats, and checksums, and even understands contextual clues. You can optionally redact data as well, using techniques like masking, secure hashing, tokenization, bucketing, and format-preserving encryption. Custom detectors can be added.

A big benefit of DLP is the ability to discover, classify and report on data from BigQuery, Cloud Storage, Datastore, and a streaming content API that enables support for additional data sources and applications.

Cloud Composer

- Fully managed workflow orchestration service built on Apache Airflow
  - Benefit from the best of Airflow with no installation or management overhead
- Cross platform orchestration tool that supports AWS, Azure and GCP (and more) with management, scheduling and processing abilities
  - Freedom from vendor lock-in and portability across platforms
- Workflows are implemented using Directed Acyclic Graphs (DAGS) created in Python
  - A collection of tasks that execute at a specific time in a specific order



Google Cloud

Better service orchestration with Workflows

A step up from dataflow in that it can orchestrate tasks across clouds

https://cloud.google.com/composer
https://cloud.google.com/composer/docs/concepts/overview

Better service orchestration with Workflows
https://cloud.google.com/blog/topics/developers-practitioners/better-service-orchestration-workflows/

Cloud Composer
Built on Apache Airflow

Robust Community of contributors and users

Well-established interfaces and library of connectors; operated using Python

Google contributing back to OSS community

Google Cloud

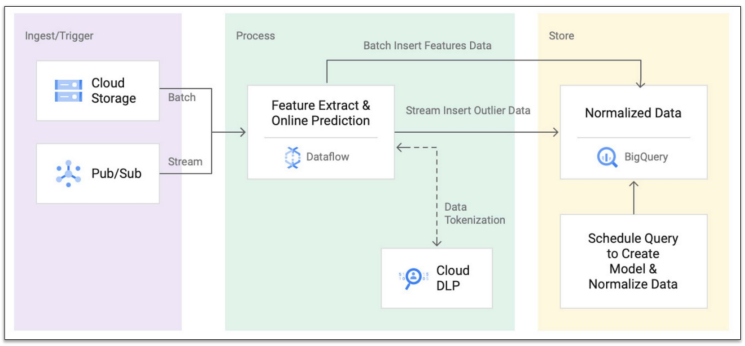
Summary: Big data ingestion and storage

Resource	
Real time ingestion	Pub/Sub
Data Lake / Raw Data	Cloud Storage
Managed ETL	Dataflow, DataPrep, BigQuery
Data Catalog / Data fabric and governance	Data Catalog / Dataplex
Data Warehouse	BigQuery

Resource	
Managed Spark / Hadoop	Dataproc
Serverless Analytics	BigQuery
Data Visualization	Data Studio, Looker
Sensitive data protection	Data Loss Prevention
Messaging	Pub/Sub
Notifications	Pub/Sub

Data analytics design patterns

- Review some of these
  - Be familiar with the products used
  - https://cloud.google.com/architecture/reference-patterns/overview



Google Cloud

Image from:
https://cloud.google.com/blog/products/data-analytics/anomaly-detection-using-streaming-analytics-and-ai

Data analytics design patterns
https://cloud.google.com/architecture/reference-patterns/overview

What Data Pipeline Architecture should I use?
https://cloud.google.com/blog/products/data-analytics/solve-your-data-analytics-etl-or-machine-learning-challenges


Google Cloud