O'REILLY

Google Cloud Professional Cloud Architect Crash Course

Prerequisites

- Familiarity with cloud platforms (AWS, Azure)
- Basic familiarity with the GCP

- This training focuses on breadth - not depth
- Concepts, fundamentals and applications

Introductions

I have experience with the Google Cloud Platform:

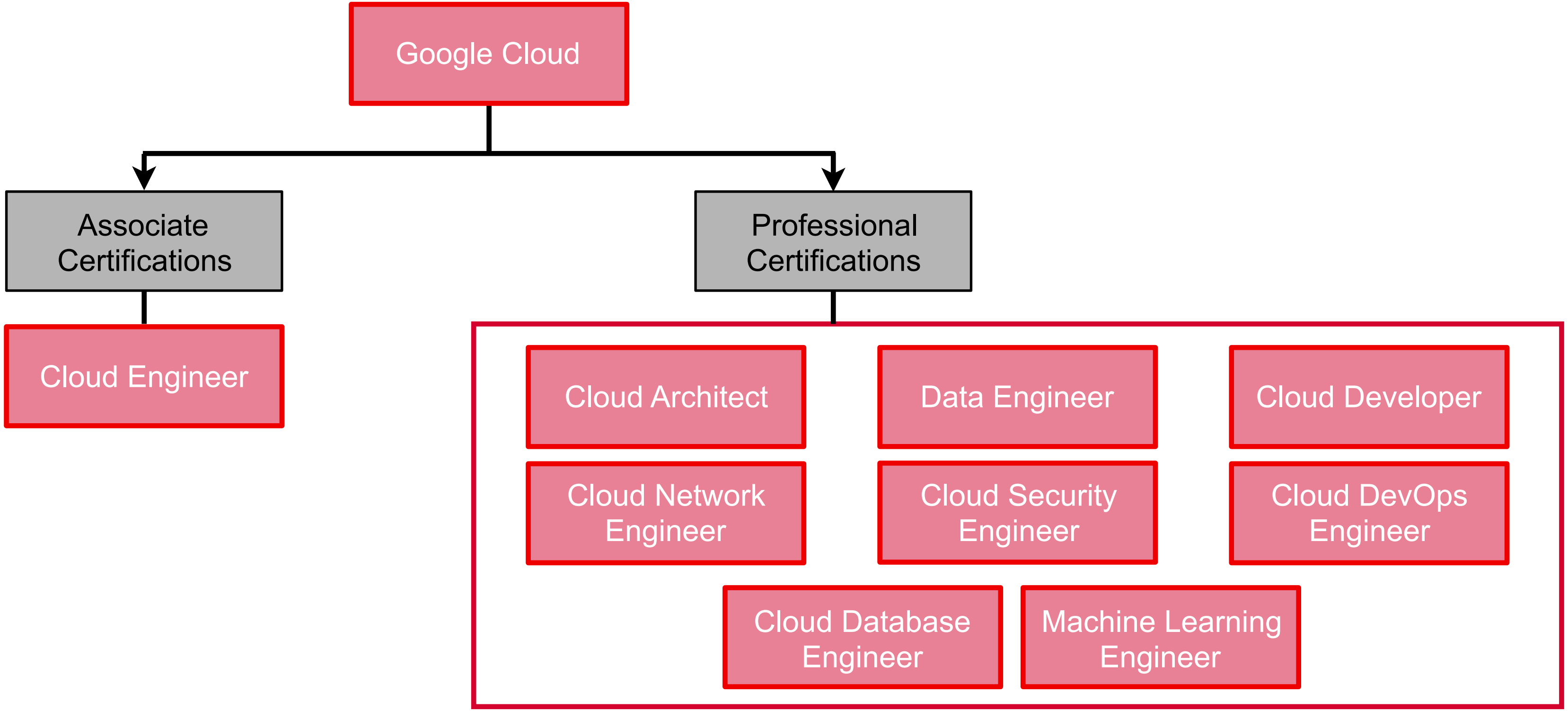
1. No experience at all
2. 0-1 years of experience
3. 2-3 years of experience
4. 3+ years of experience

Introductions

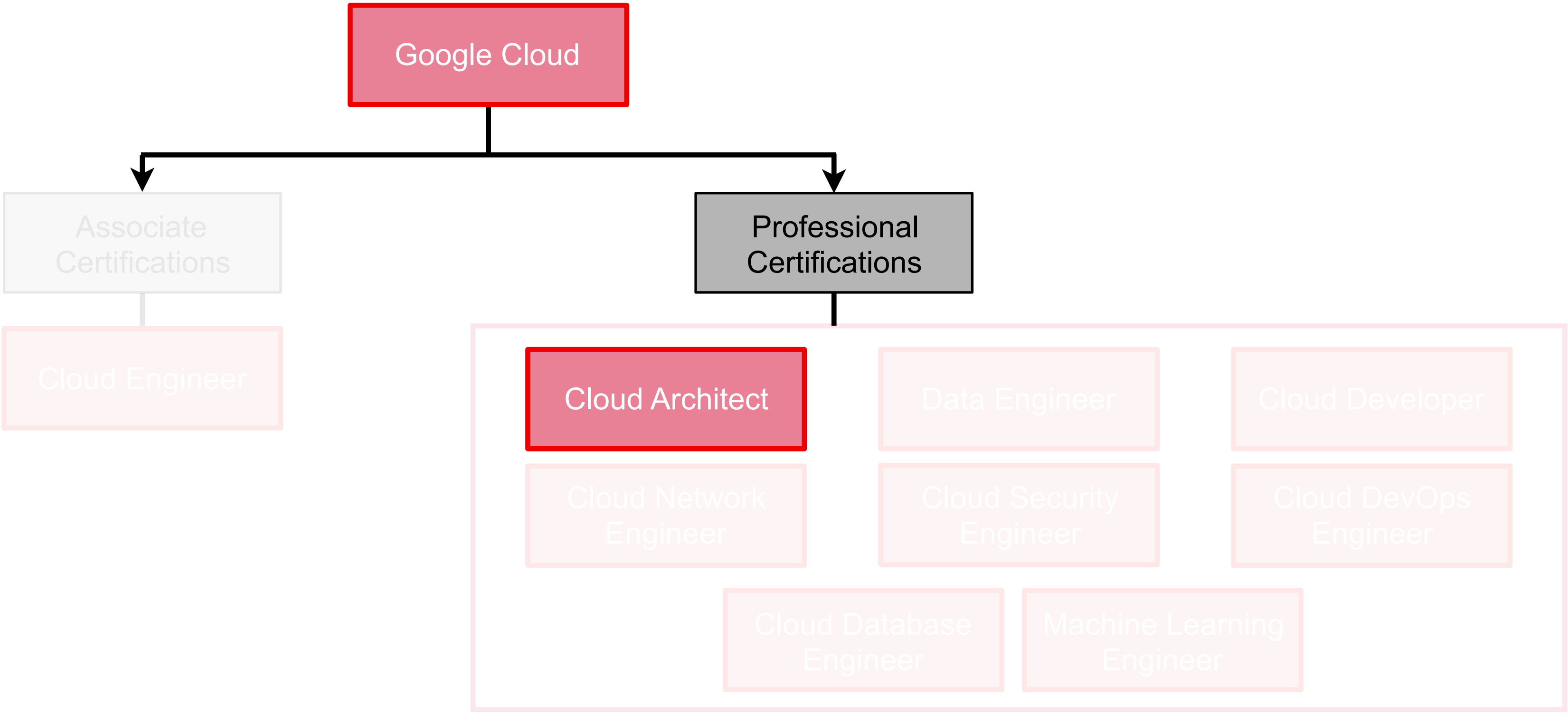
I have worked on other cloud platforms:

1. Mostly AWS
2. Mostly Azure
3. Mostly Oracle
4. Mostly IBM
5. Other cloud platforms

Google Cloud Certifications



Google Cloud Certifications



Professional Cloud Architect

- Test duration: 2 hours
- Registration fee: $200 + taxes
- Languages: English, Japanese
- Exam format: 50-60 multiple choice and multiple select questions
- Case Studies: 2 case studies in each exam - make up 20-30% of the exam
- Recommended: 3+ years industry experience, 1+ year designing and managing solutions on GCP

Professional Cloud Architect

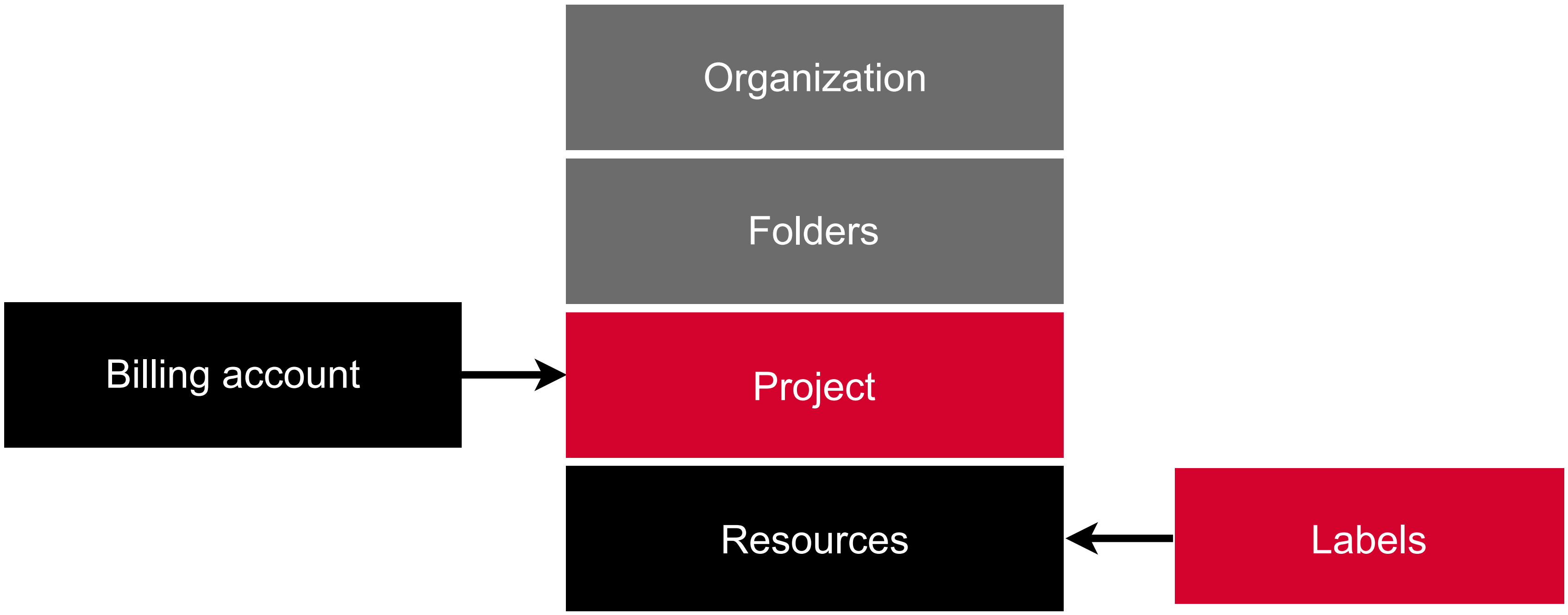
- Vast array of services for a wide variety of use cases
- A good understanding of the specialized strengths of each service
- Main exam link:
  - https://cloud.google.com/certification/cloud-architect

Professional Cloud Architect

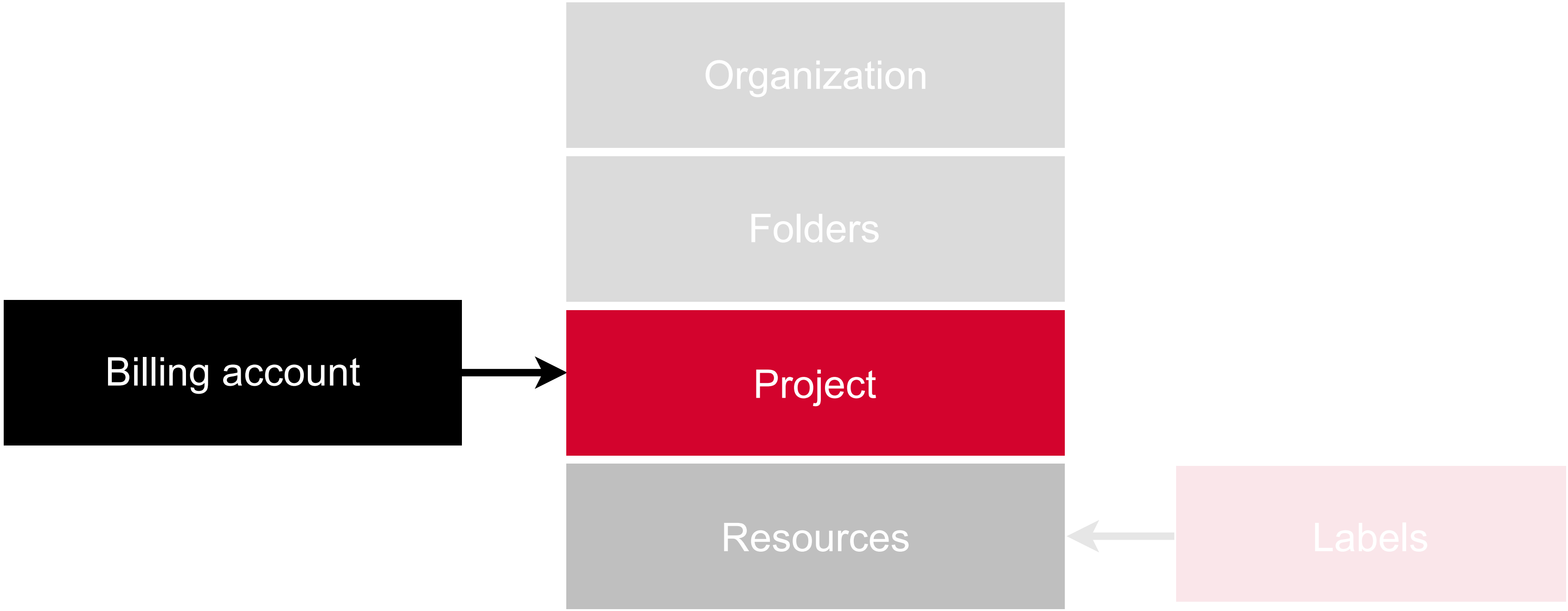
- Cloud Architect Certification training path:
  - https://www.cloudskillsboost.google/paths/12
- Case studies link here:
  - https://cloud.google.com/certification/guides/professional-cloud-architect/
- Extensive labs for hands-on practice:
  - https://codelabs.developers.google.com/?cat=Cloud
- Sample test:
  - https://docs.google.com/forms/d/e/1FAIpQLSf54f7FbtSJcXUY6-DUHfBG31jZ3pujgb8-a5io_9biJsNpqg/viewform?usp=sf_link

Google Cloud Platform Basics

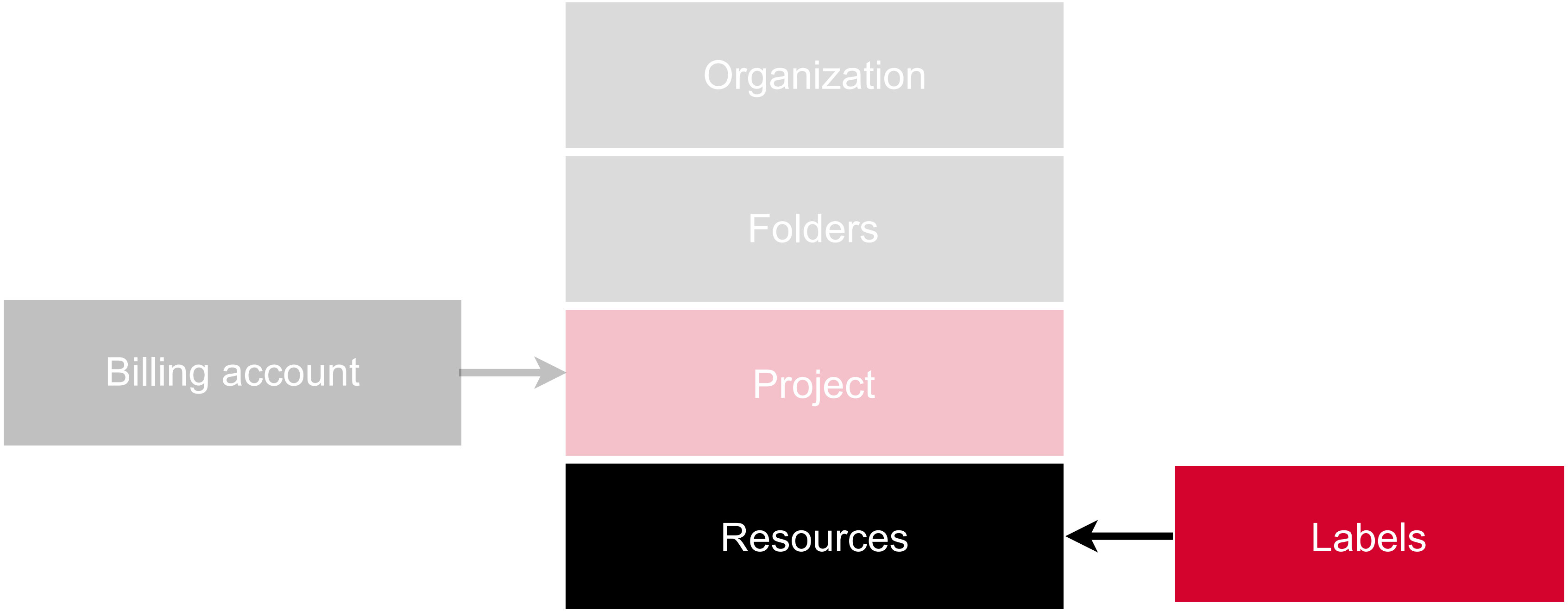
Resource Hierarchy of Components



Billing Accounts Are Associated with Projects



Labels Are Applied to Resources



Labels Help in Allocating Costs

- Categorize resources
  - Different environments
  - Different projects
- Label resources accordingly
  - *env=dev, env=prod*
  - *service=search, service=catalog*
- Can export billing to BigQuery and analyze costs using labels

Using Google Cloud Resources

Cloud Console

Cloud Shell

Command-line Tools

gsutil, bq

APIs and Client Libraries

Choices in Computing

Compute

Where is code executed and how?

Storage

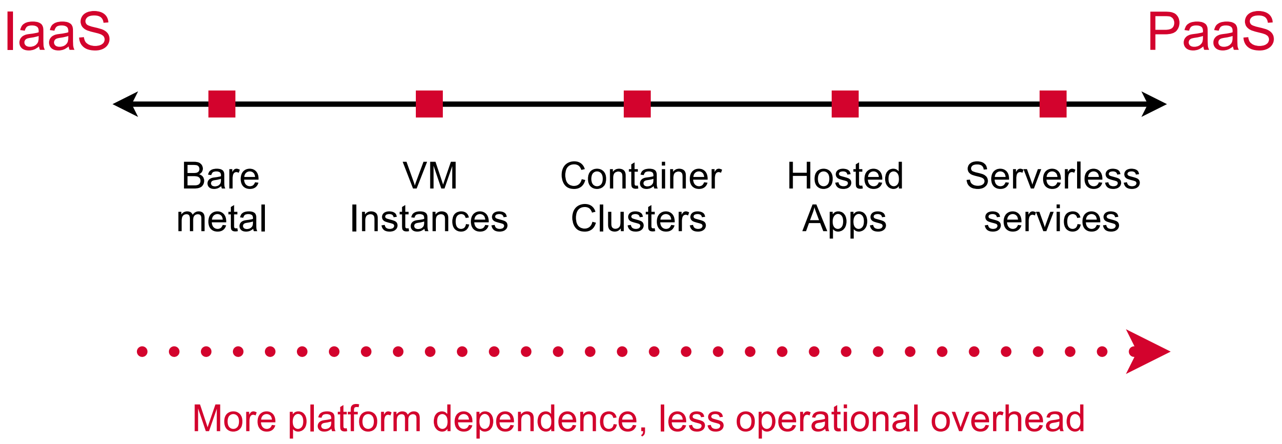
Where is data stored?

Networking, logging, are choices made after this fundamental decision

Compute Choices

Compute Choices

Compute Choices



Compute Choices

IaaS

PaaS

Bare metal

VM Instances

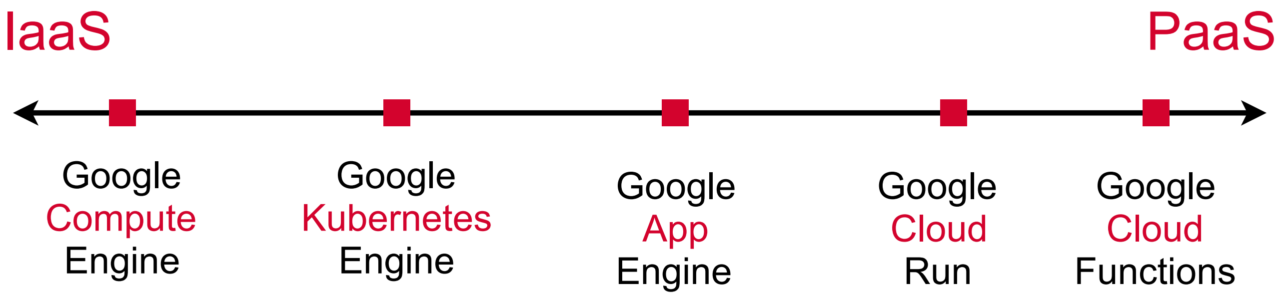
Container Clusters

Hosted Apps

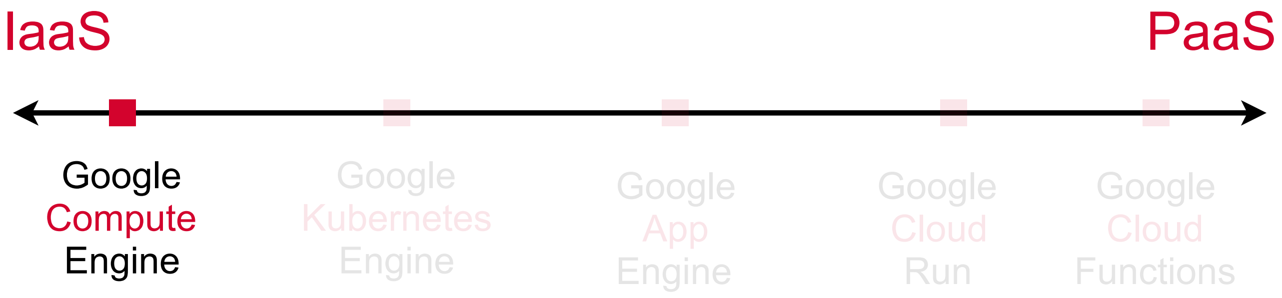
Serverless services

More control, granular access, more administrative overhead

Google Cloud Compute Choices



Google Cloud Compute Choices



Projects

Which of the following best describes a project on the GCP?

1. Logical grouping of resources based on labels
2. Root node in the resource hierarchy
3. Used to group GCP networks
4. Logical grouping for resources, associated with billing

Projects

Which of the following best describes a project on the GCP?

1. Logical grouping of resources based on labels
2. Root node in the resource hierarchy
3. Used to group GCP networks
4. **Logical grouping for resources, associated with billing**

Cloud Shell

Which of the following best describes Cloud Shell?

1. Command-line utility used to work with the GCP services
2. Ephemeral VM which offers a terminal on the browser
3. PaaS offering on the GCP for hosted applications
4. IaaS offering on the GCP

Cloud Shell

Which of the following best describes Cloud Shell?

1. Command-line utility used to work with the GCP services
2. **Ephemeral VM which offers a terminal on the browser**
3. PaaS offering on the GCP for hosted applications
4. IaaS offering on the GCP

Google Compute Engine (GCE)

Zones and Regions

Zone

Availability zone (similar to a datacenter)

Region

Set of zones with high-speed network links

Zones and Regions

Zone

“asia-south1-a”

Region

“asia-south1”

Networks are Global Resources

Network

User-controlled IP addresses, subnets and firewalls

Networks are Global Resources

Network

default

Global, Regional, and Zonal Compute Resources

- Compute engine resources are
  - global
  - regional
  - zonal
- **Scope determines accessibility of a resource to other resources on the Google Cloud**
  - Global resources accessible from any region or zone
  - Regional resources accessible only from the same region
  - Zonal resources accessible from the same zone

Global, Regional and Zonal Compute Resources

- Global:
  - Global static IP addresses
  - Images and snapshots
  - Networks, firewalls, routes
- Regional
  - Subnets
  - Regional static external IP addresses
  - Regional persistent disks
- Zonal
  - Instances
  - Persistent disks

Regions and Zones

- Connecting a persistent disk to a VM requires both to be in the same zone
- Assigning a static IP address to a VM requires both to be in the same region

Configuration Choices

Machine Family

General purpose, compute optimized, memory optimized, accelerator-optimized

Machine Series

Machines have generation numbers where higher generations have newer features

Machine Type

vCPUs count, memory capacity, and storage capacity

Base Image

Public (free or premium), custom, snapshots from boot disks

Machine Type

Predefined

Custom

Machine Type

Predefined

Custom

Fixed set of types with fixed ratios of memory to vCPU count

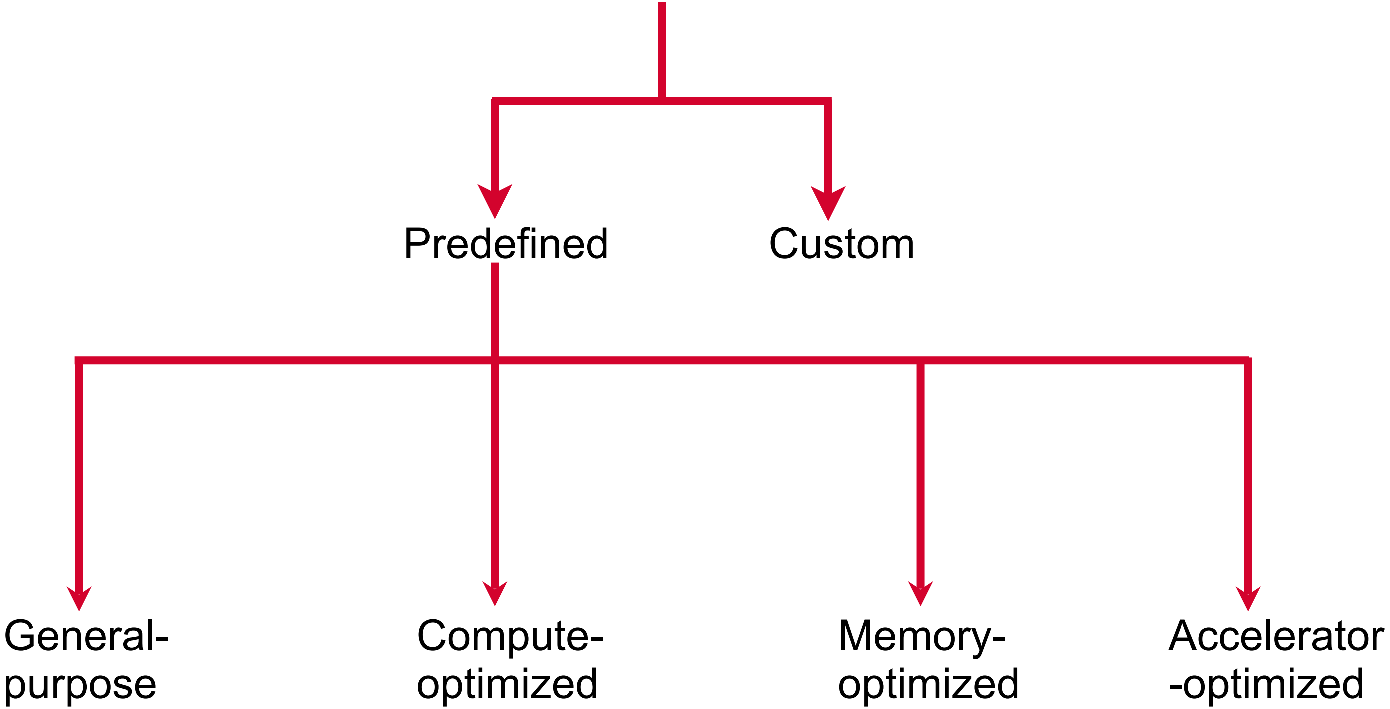
Machine Type

Predefined

Custom

Can independently
specify vCPU count
and amount of memory

Machine Type



Shared-core Machines

- Cost-effective for running non-resource intensive operations
- A single vCPU run for a time period on single hardware
- Offer micro-bursting capabilities for spikes
- Instance will use additional physical CPUs during spikes

Base Images

Public

Custom

Base Images

Public

Custom

Provided and maintained by Google, open-source communities, and third-party vendors

All projects have access to these images and can use them to create instances

Base Images

Public

Custom

Linux, Windows, Container-optimized OS, SQL Server

Base Images

Public

Custom

Available only to your project

First, create a custom image from boot disks and other images; then, use the custom image to create an instance

Spot Instances

An instance that you can create and run at a much lower price than normal instances. However, **GCE might terminate (preempt)** these instances if it requires access to those resources for other tasks.

May not always be available. Not covered by SLAs

- Batch processing and data analysis
- CI/CD pipelines

Preemptible Instances

Similar to Spot VMs (older product and will have fewer features than Spot VMs)

Will definitely be preempted every 24 hours

May not always be available. Not covered by SLAs

- Batch processing and data analysis
- CI/CD pipelines

Sole-tenant Nodes

A sole-tenant node is a physical Compute Engine server that is **dedicated to hosting VM instances** only for your specific **project**

Keeps your instances physically separated from instances in other projects.
Group instances on the same hardware

- Compliance requirements
- Performance-sensitive applications
- Data isolation

Shielded VMs

Shielded VMs provide enhanced security features to protect virtual machines from rootkits, bootkits, and other advanced persistent threats.

Ensures VMs firmware and boot loader not tampered with.

- Secure boot
- Virtual Trusted Platform Module (vTPM)

Confidential VMs

Confidential VMs are designed to provide advanced security and privacy for your workloads by encrypting data in use. Ensures that data processed within the VM is encrypted

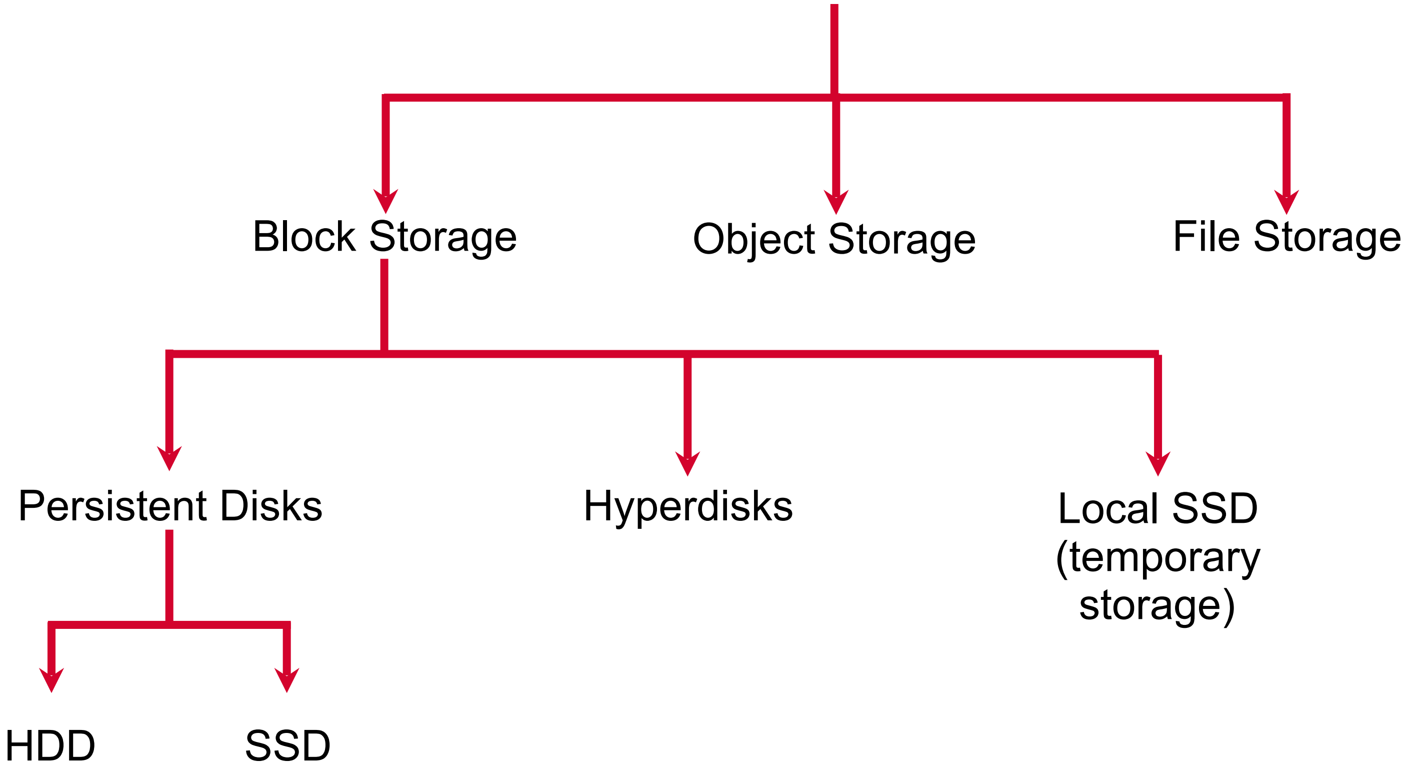
Accessing Storage from VMs

Block Storage

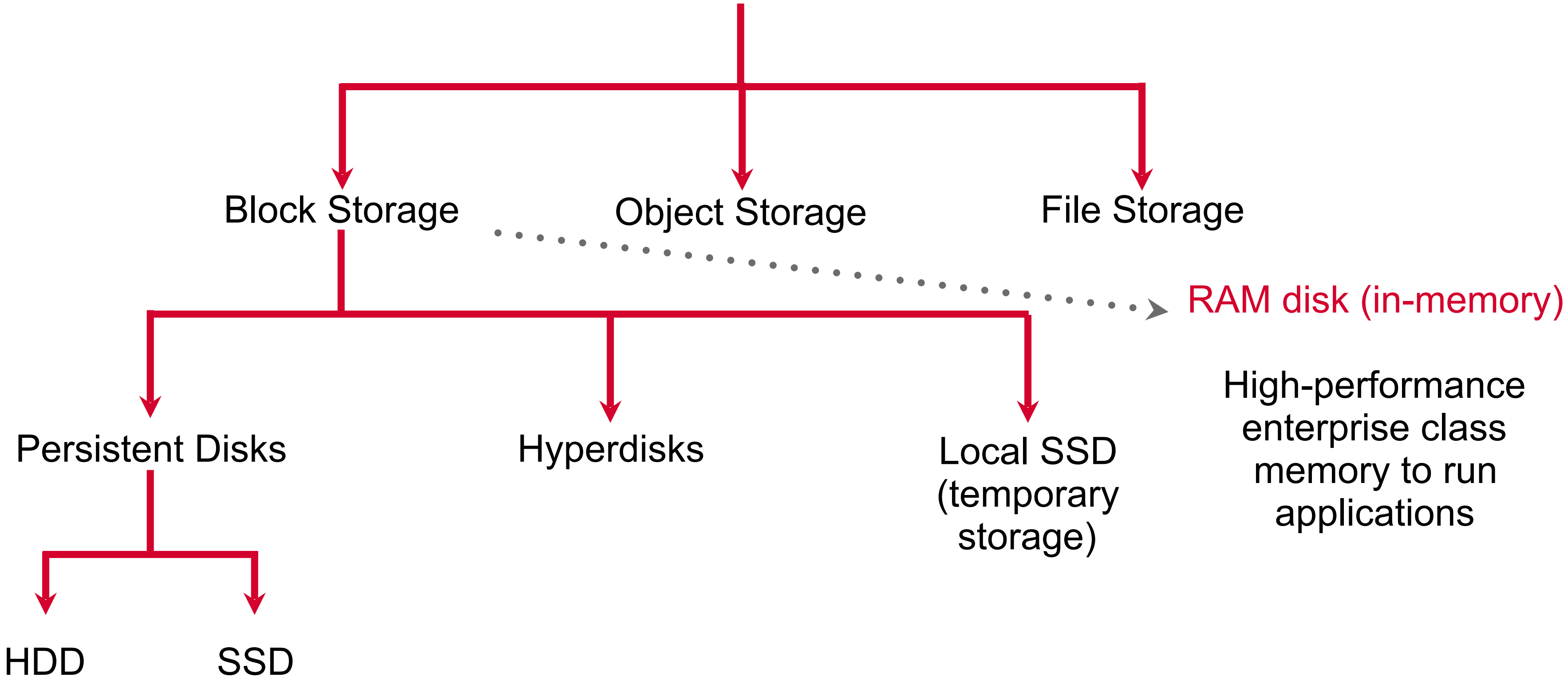
Object Storage

File Storage

Accessing Storage from VMs



Accessing Storage from VMs



Accessing Storage from VMs

Block Storage

Object Storage

GCS Buckets

Pay-what-you-use
blob storage

File Storage

Cloud Filestore

File-system interface
with a managed
shared file storage
service

Persistent Disks vs.Local SSDs

Persistent Disks

- Network-attached storage
- Data redundancy built-in
- Bootable
- Durable
- HDD or SSD
- 64TB max for one volume
- Create snapshots or images
- Relatively slow

Local SSDs

- **Physically attached to instance**
- No data redundancy built-in
- Not bootable
- Not durable
- SSD for better performance
- 9TB max
- Cannot create snapshots or images
- Very fast, especially for random access

O'REILLY

Snapshots and Images

Image

- Binary file used to instantiate VM root disk
- Usually based off OS image
- Also contains boot loader
- Can also contain customizations
- Managed by GCP image service

Snapshot

- Binary file with exact contents of persistent disk
- “Point-in-time” snapshot
- Managed by GCP snapshot service
- Incremental backups possible too
- Used to back up data from persistent disks

Snapshots and Images

Conceptually very similar but many differences in nitty-gritty

Region

Which of the following best describes a region on the GCP?

1. A logical area that may be spread across countries
2. A single datacenter on the GCP
3. A geographical area with multiple datacenters
4. Physically connected hardware devices in a datacenter

Region

Which of the following best describes a region on the GCP?

1. A logical area that may be spread across countries
2. A single datacenter on the GCP
3. **A geographical area with multiple datacenters**
4. Physically connected hardware devices in a datacenter

Local SSD

Which of the following correctly describes a local SSD?

1. Used a as a boot disk and can be snapshotted
2. Offers lower performance as compared with Cloud Storage Buckets
3. Elastic storage which grows as you store more data in it
4. Physically attached to your VM so offers high throughput and low latency

Local SSD

Which of the following correctly describes a local SSD?

1. Used a as a boot disk and can be snapshotted
2. Offers lower performance as compared with Cloud Storage Buckets
3. Elastic storage which grows as you store more data in it
4. **Physically attached to your VM so offers high throughput and low latency**

Google App Engine

Google Cloud Compute Choices

IaaS

PaaS

Google Compute Engine

Google Kubernetes Engine

Google App Engine

Google Cloud Run

Google Cloud Functions

Google App Engine

Web framework and platform for hosting web applications on the Google Cloud

Support for Go, PHP, Java, Python, Node.js, .NET, Ruby and other languages

Google App Engine

Web framework and platform for hosting web applications on the Google Cloud

Support for Go, PHP, Java, Python, Node.js, .NET, Ruby and other languages

Focus on development and code

Infrastructure and scaling taken care of by the platform

App Engine Environments

Standard Environment

Flexible Environment

App Engine Environments

Standard

- App runs in a proprietary sandbox
- Instances start up in seconds
- Code in few languages/versions only
- No other runtimes possible
- Apps cannot access Compute Engine resources
- Can install 3rd party binaries only for selected runtimes

App Engine Environments

Standard

- App runs in a proprietary sandbox
- Instances start up in seconds
- Code in few languages/versions only
- No other runtimes possible
- Apps cannot access Compute Engine resources
- Can install 3rd party binaries only for selected runtimes

Flexible

- Runs in Docker container on GCE VM
- Instance start up in minutes
- Code in far more languages/versions
- Custom runtimes possible
- Apps can access Compute Engine resources, some OS packages
- Can install and access third-party binaries

App Engine Environments

Standard

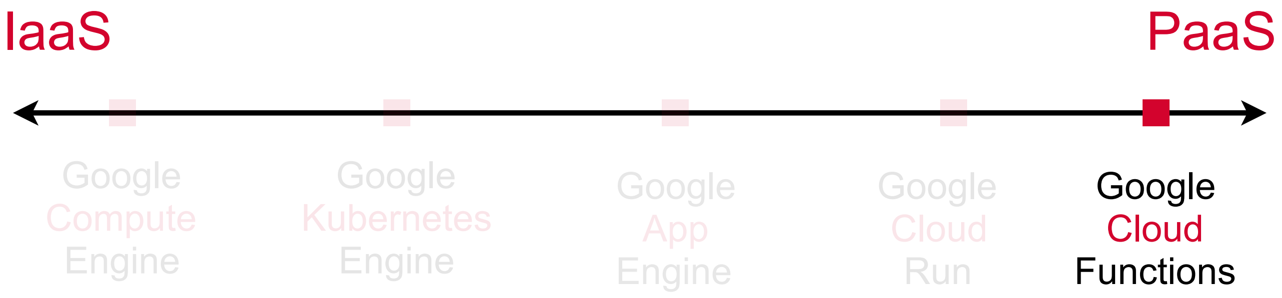
- Apps that experience traffic spikes
- Usually stateless HTTP web apps

Flexible

- Apps that experience consistent traffic
- General purpose apps

Google Cloud Functions

Google Cloud Compute Choices



Cloud Functions

Event-driven serverless compute platform

Event-driven Serverless Compute

Event occurs

Platform triggers execution

Cloud Function code runs

Invokes other Google Cloud services

Types of Events

- HTTP
- Background
  - Cloud Storage
  - Pub/Sub
  - Logging
  - Firestore

Concurrency and Scale

- Spin up function instances based on current load
- Functions receive event parameters from platform
- Functions do not share memory or variables
- An instance processes a single request (generation 1)
- Function concurrency supported (generation 2)
- Functions should be stateless

Google Cloud Run

Google Cloud Compute Choices

IaaS

PaaS

Google Compute Engine

Google Kubernetes Engine

Google App Engine

Google Cloud Run

Google Cloud Functions

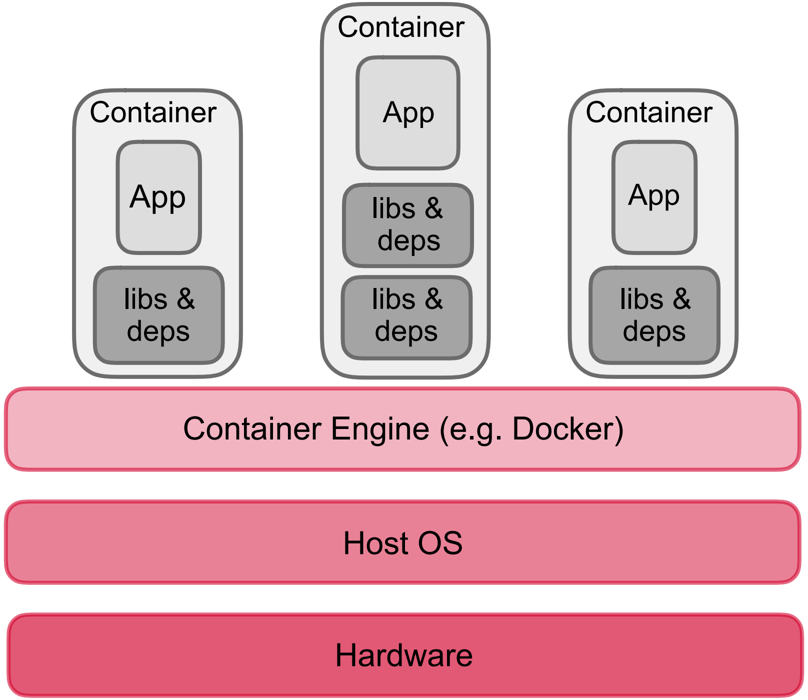
Container

A container image is a lightweight, stand-alone, executable package of a piece of software that includes everything needed to run it; code, runtime, system tools, system libraries, settings

Container

- Contains applications
- And all of the application’s dependencies
- Platform independent
- Runs on layer of abstraction
- Docker Runtime (for Docker containers)

Modern Workloads on Containers



Cloud Run

Serverless, managed platform that lets you run containers directly on top of Google’s scalable architecture

Cloud Run

- Write your code in any programming language
- Create a container image (or use source-based deployment option - Google Cloud will build container image for you)
- Register the container with the artifact registry
- Deploy your container directly using Cloud Run
- No cluster creation no infrastructure management
- Request-based pricing and instance-based pricing

Running Code Using Cloud Run

Cloud Run Services

Cloud Run Jobs

Both use the same environment and have the same integrations with other Google Cloud services

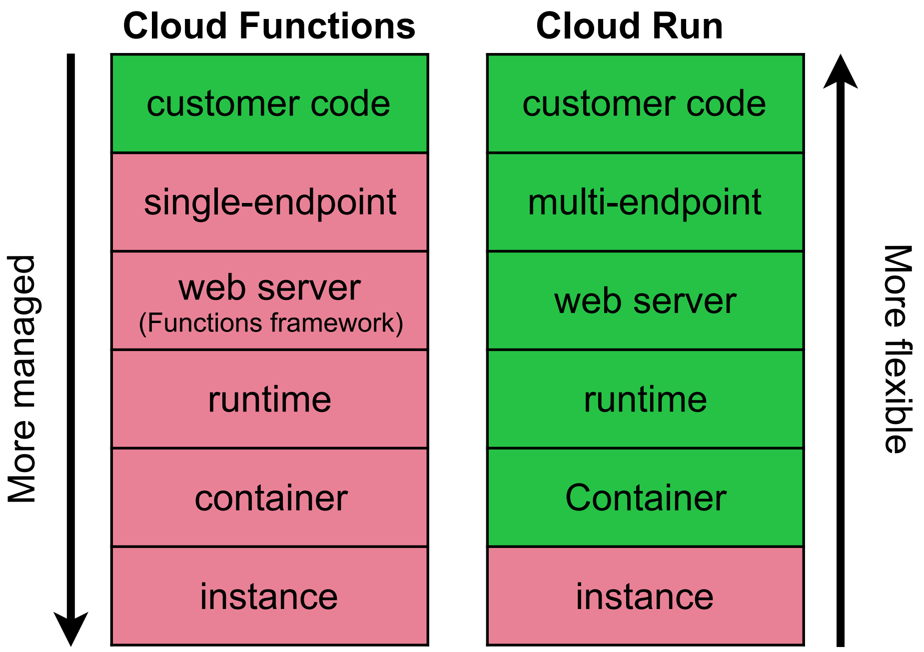
Cloud Run Services

- Used to run code that responds to web requests or events
- Each service located in a Google Cloud region
- Replicated across zones in the region
- Exposes an endpoint
- Automatically scales underlying infrastructure to handle incoming requests
- Version management, rollbacks, traffic management - all handled by the platform

Cloud Run Jobs

- Used to run code that performs work (a job) and quits when the job is done
- Each service located in a Google Cloud region and executes one or more containers to completion
- A job comprises of many tasks executing in parallel - each container runs one task

Cloud Functions vs. Cloud Run



How managed do you want to be?

Cloud Functions vs. Cloud Run

Cloud Functions

- Specific limited runtimes supported
- Can be triggered based on platform events
- No support for running jobs
- 2nd generation functions support concurrency

Cloud Run

- All runtimes that can be run using containers
- Expose endpoints and invoked using HTTP requests
- Support for running jobs
- Great support for concurrent requests

Cloud Functions vs. Cloud Run

Cloud Functions

- Choose Cloud Functions if you primarily want to connect to other cloud services on Google Cloud

Cloud Run

- Choose Cloud Run if you want a simple way to scale and maintain services using containers

AppEngine

Which of the following is true about the standard environment on AppEngine?

1. Can be used with custom runtimes
2. Runs in a proprietary sandbox on the GCP
3. Runs within a Docker container
4. Takes a couple of minutes to startup

AppEngine

Which of the following is true about the standard environment on AppEngine?

1. Can be used with custom runtimes
2. **Runs in a proprietary sandbox on the GCP**
3. Runs within a Docker container
4. Takes a couple of minutes to startup

AppEngine

Which of the following is true about the flexible environment on AppEngine?

1. Cannot install third party libraries
2. Runs in a proprietary sandbox on the GCP
3. Runs within a Docker container
4. Takes only a few seconds to startup

AppEngine

Which of the following is true about the flexible environment on AppEngine?

1. Cannot install third party libraries
2. Runs in a proprietary sandbox on the GCP
3. **Runs within a Docker containe**r
4. Takes only a few seconds to startup

Serverless Applications

When would you choose to use Cloud Functions over Cloud Run?

1. When you need to run a containerized application.
2. When you need to run a function in response to events.
3. When you require fine-grained control over application resources.
4. When you need to deploy a long-running application.

Serverless Applications

When would you choose to use Cloud Functions over Cloud Run?

1. When you need to run a containerized application.
2. When you need to run a function in response to events.
3. When you require fine-grained control over application resources.
4. When you need to deploy a long-running application.

Serverless Applications

Which of the compute options is great for running batch jobs in containers?

1. Cloud Functions
2. AppEngine
3. Cloud Run
4. Apps running on VMs

Serverless Applications

Which of the compute options is great for running batch jobs in containers?

1. Cloud Functions
2. AppEngine
3. **Cloud Run**
4. Apps running on VMs

O'REILLY

Storage on the Google Cloud

Choices in Computing

Compute

Where is code executed and how?

Storage

Where is data stored?

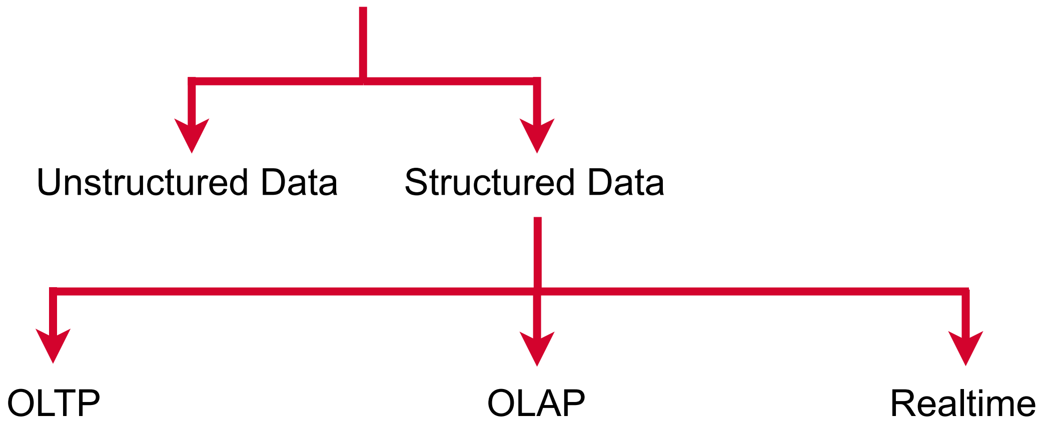
Networking, logging, are choices made after this fundamental decision

Storage Technologies

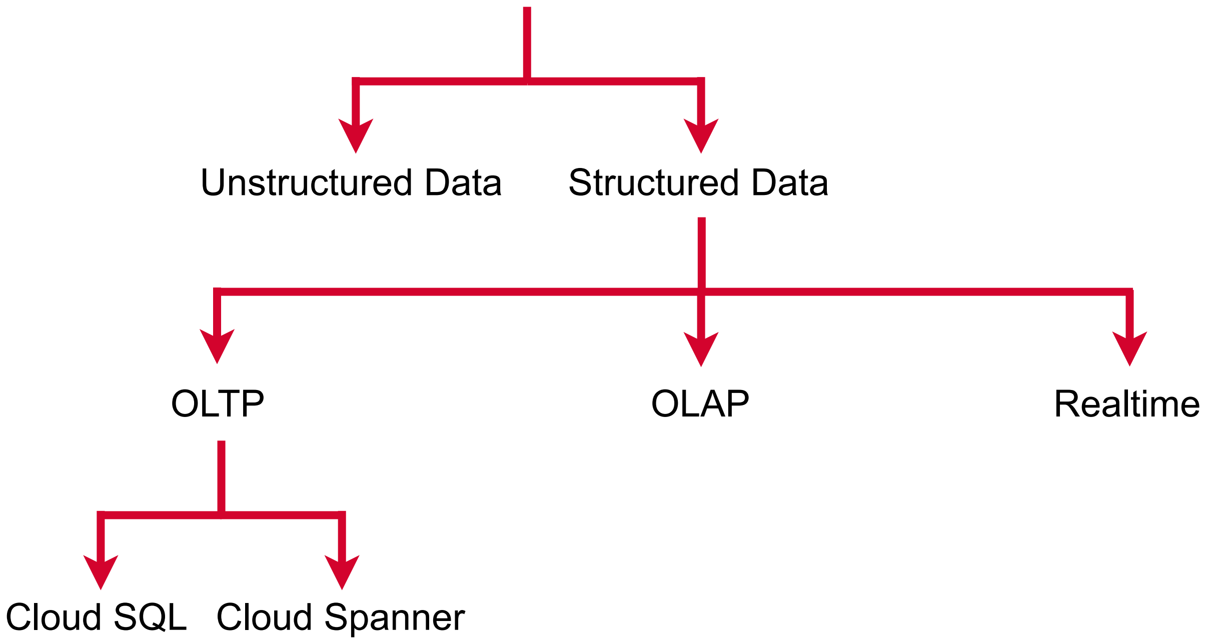
Unstructured Data

Structured Data

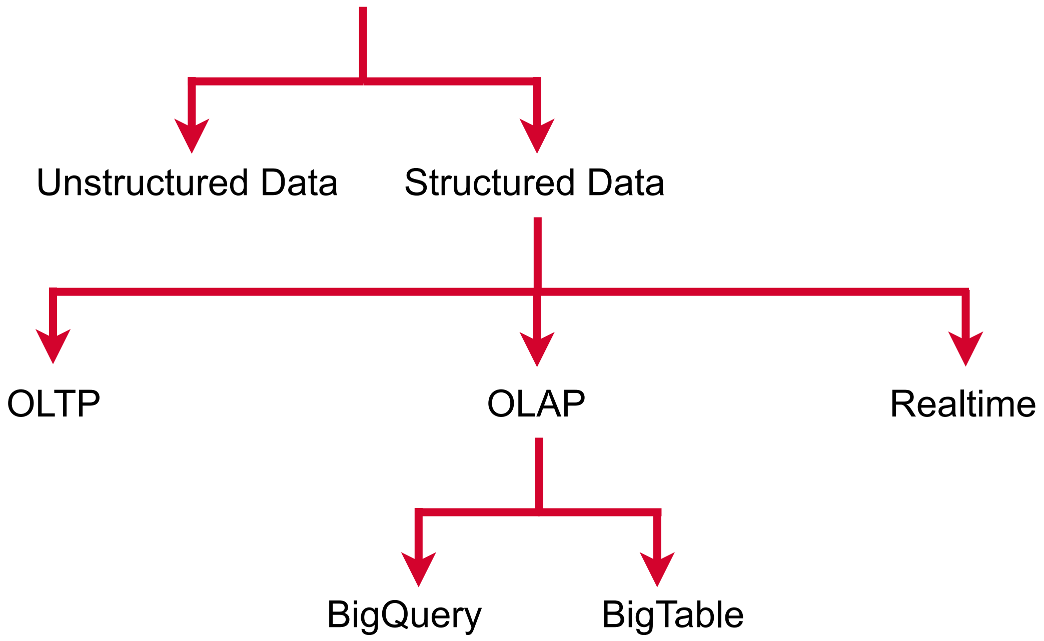
Storage Technologies



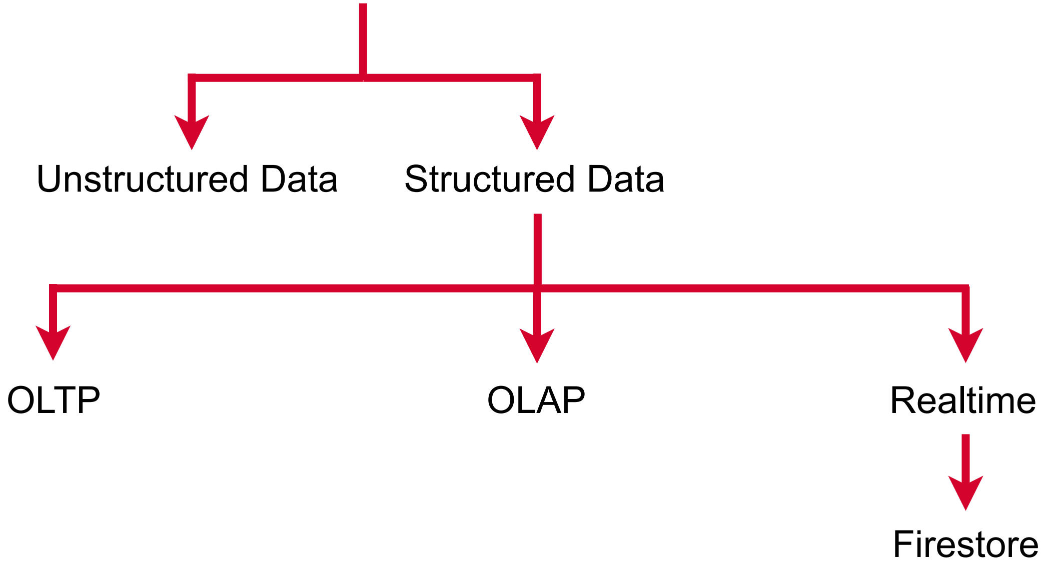
Storage Technologies



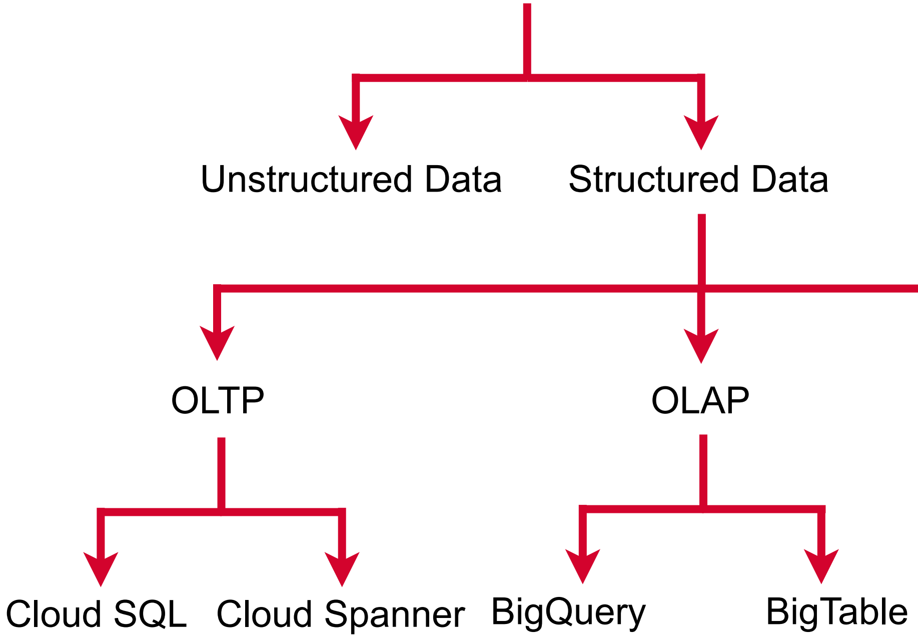
Storage Technologies



Storage Technologies



Storage Technologies



Storage Technologies

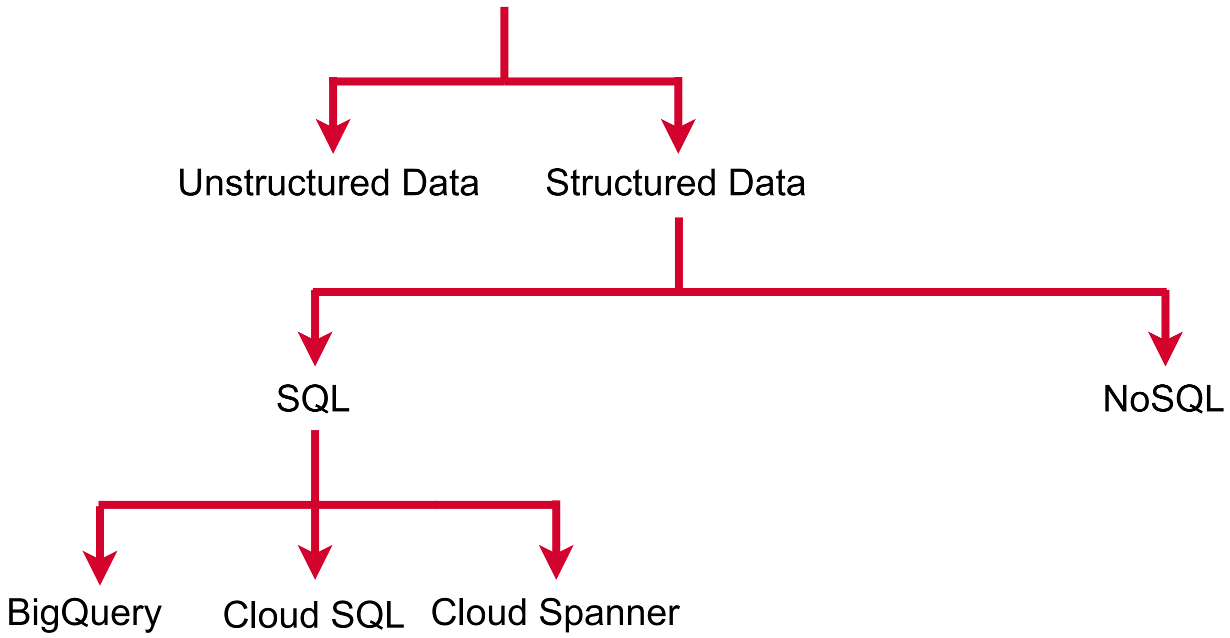
Unstructured Data

Structured Data

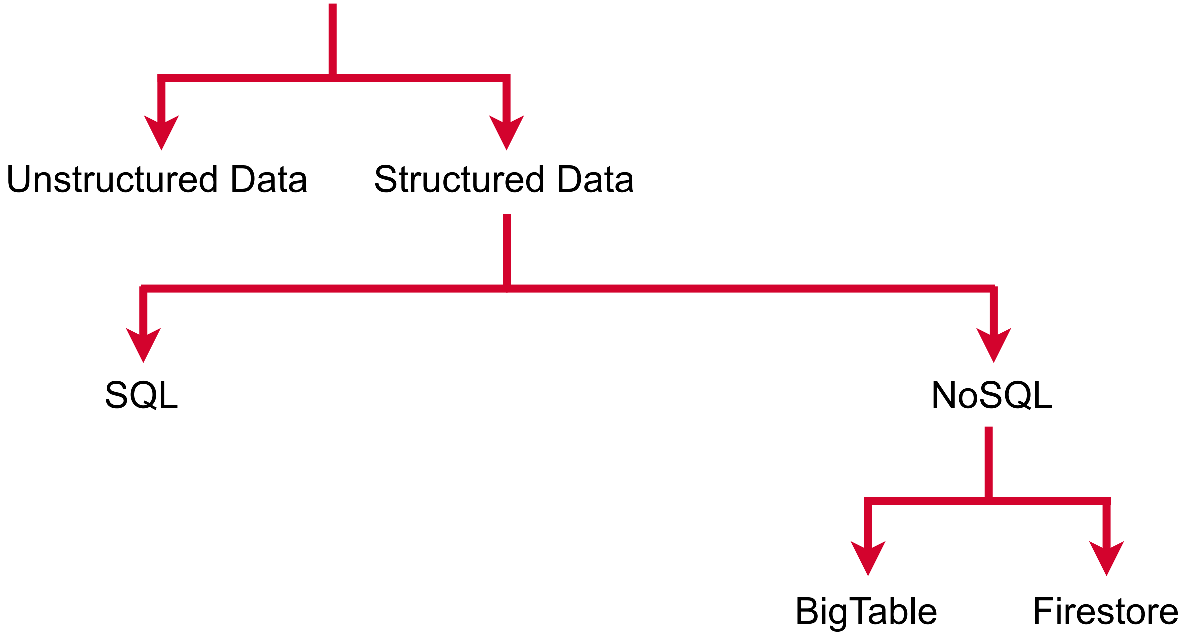
SQL

NoSQL

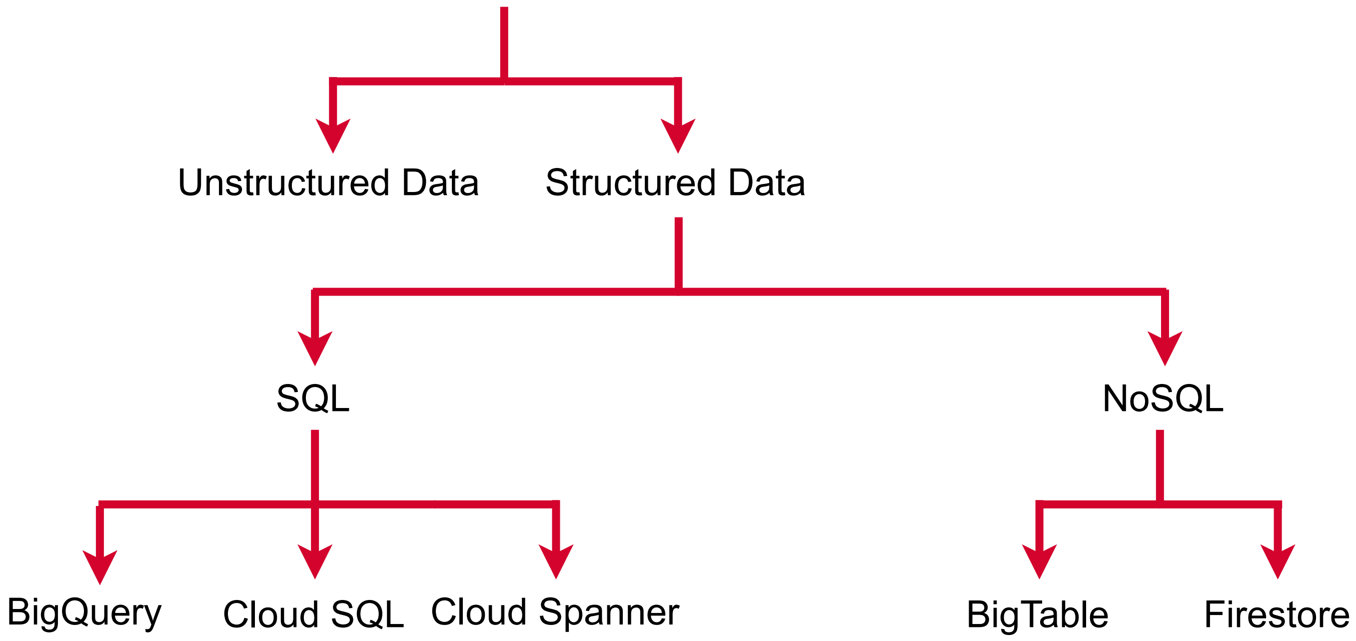
Storage Technologies



Storage Technologies



Storage Technologies



Unstructured Data

Block Storage

File Storage

Object Storage

Unstructured Data

Block Storage

File Storage

Object Storage

Physically addressable storage accessed from compute - data split into uniform blocks

High performance read and write access at the block level

Unstructured Data

Block Storage | **File Storage** | Object Storage

Stores data as a hierarchy of files within directories

Shared concurrent access from multiple machines

Unstructured Data

Block Storage

File Storage

Object Storage

Logically addressable
storage accessed from
compute or by human users

Persistent Disks vs. Buckets

Persistent Disks

- Block storage
- Max 64TB in size
- Pay what you allocate
- Tied to GCE VMs
- Zonal (or regional) access

Buckets

- Object storage
- Infinitely scalable
- Pay what you use
- Independent of GCE VMs
- Global access

Google Cloud Storage

GCS Storage Classes

How often is a data item accessed?

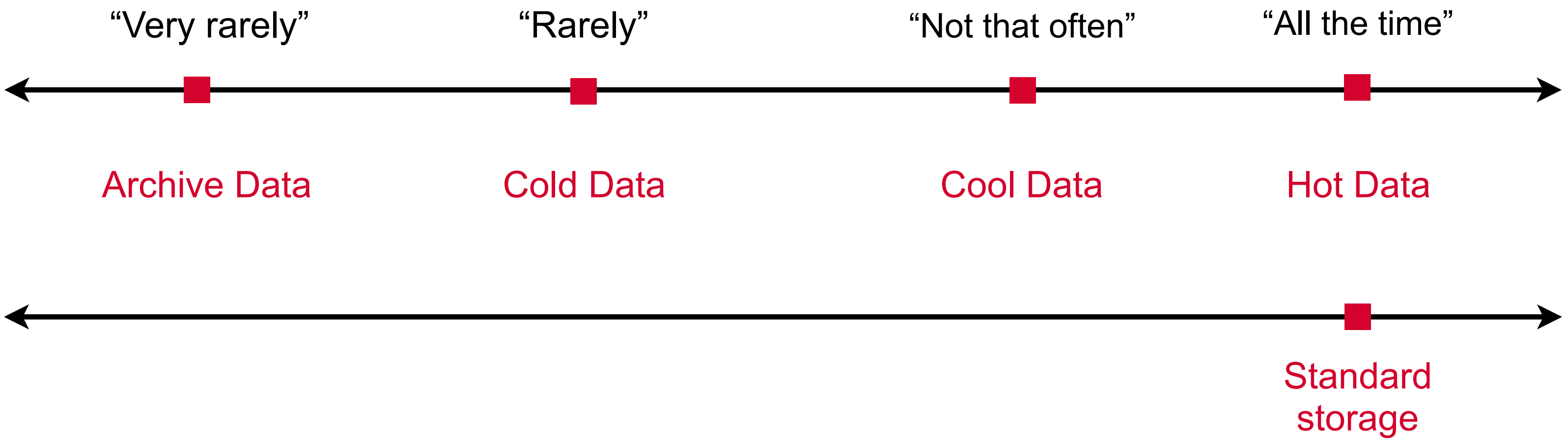
GCS Storage Classes

How often is a data item accessed?

“Very rarely”	“Rarely”	“Not that often”	“All the time”
Archive Data	Cold Data	Cool Data	Hot Data
Less than once a year	Once every 3 months	Once a month	Many times a month

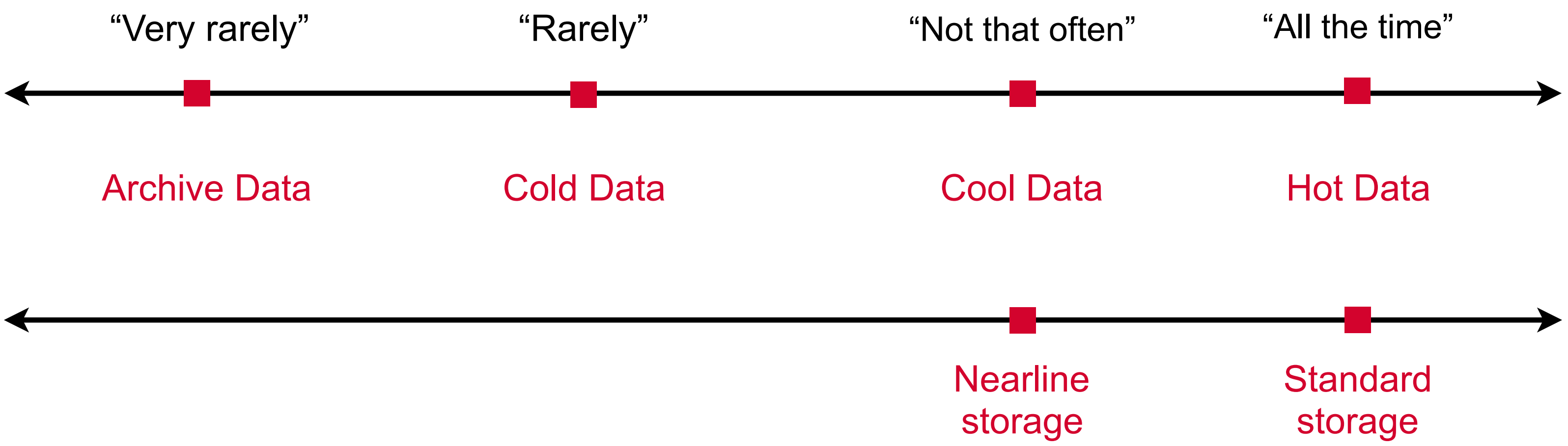
GCS Storage Classes

How often is a data item accessed?



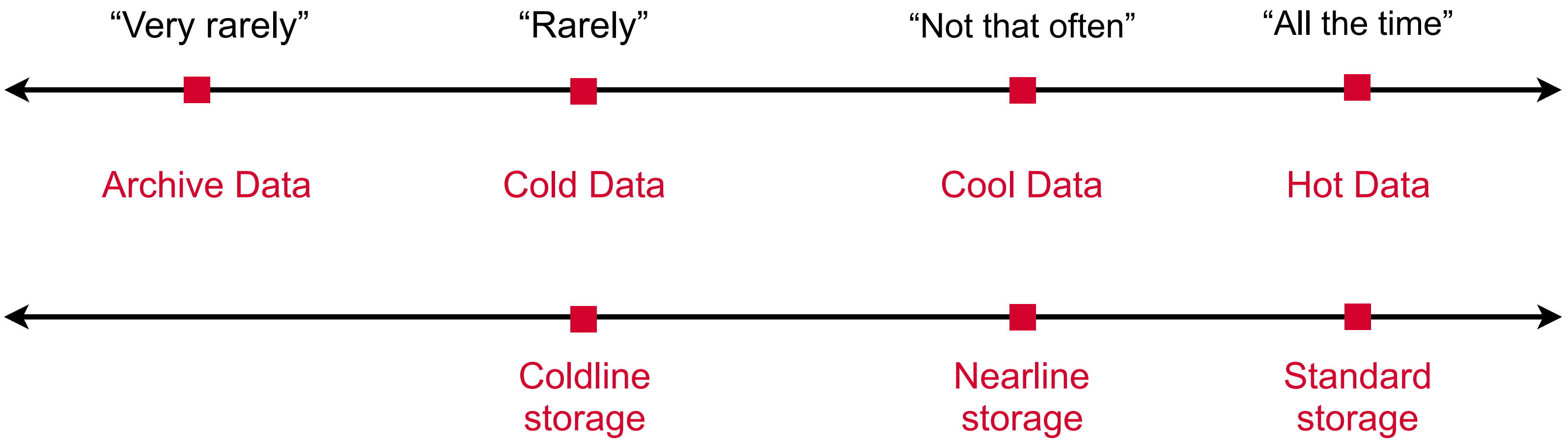
GCS Storage Classes

How often is a data item accessed?



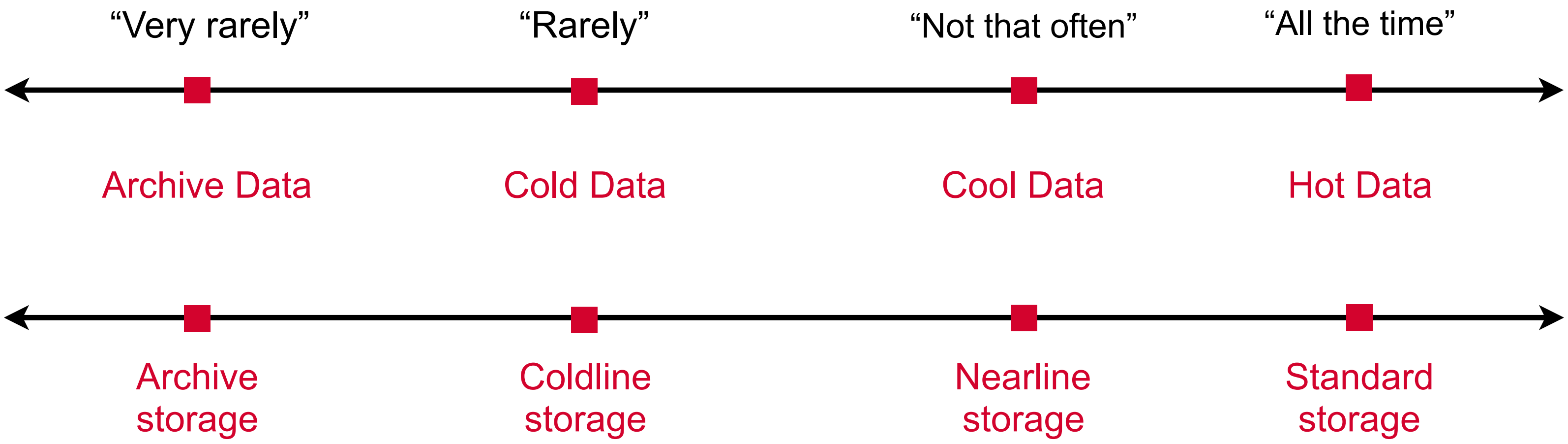
GCS Storage Classes

How often is a data item accessed?

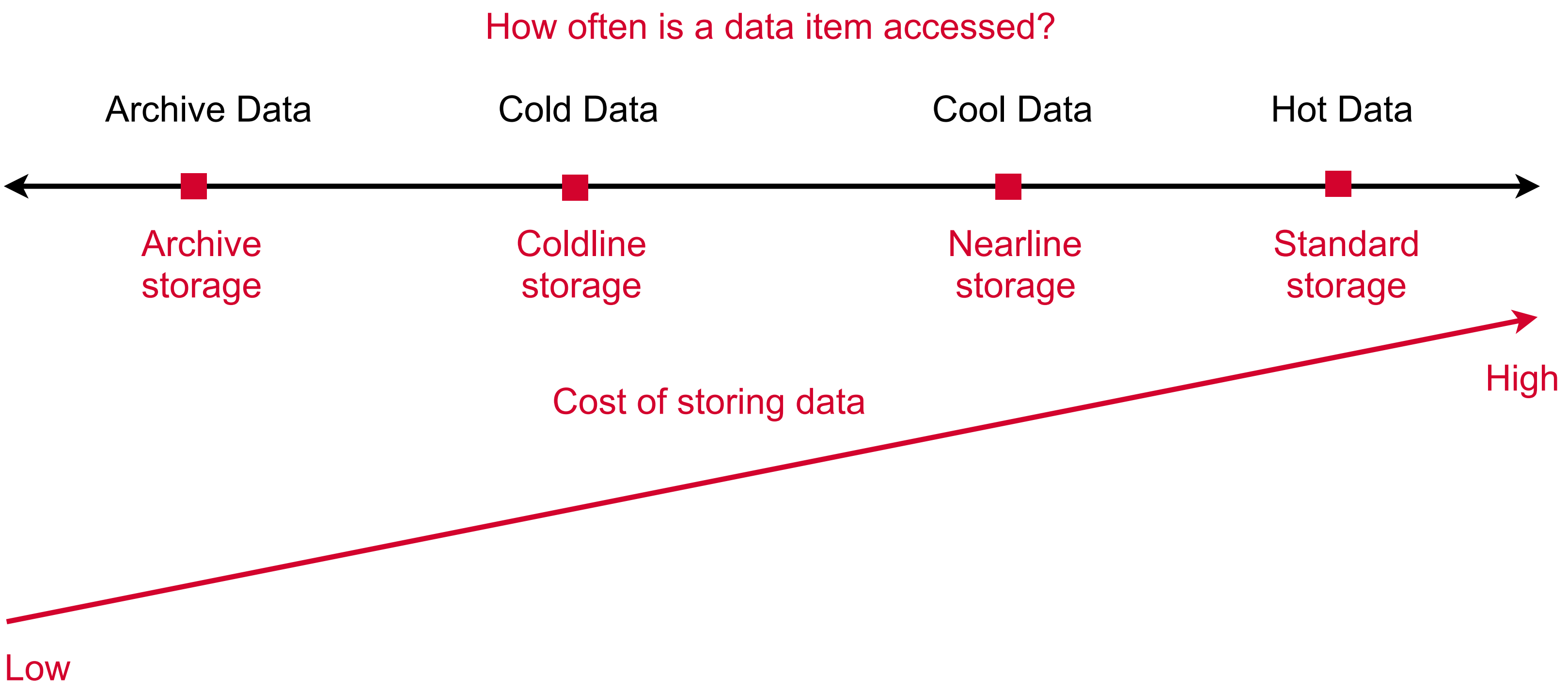


GCS Storage Classes

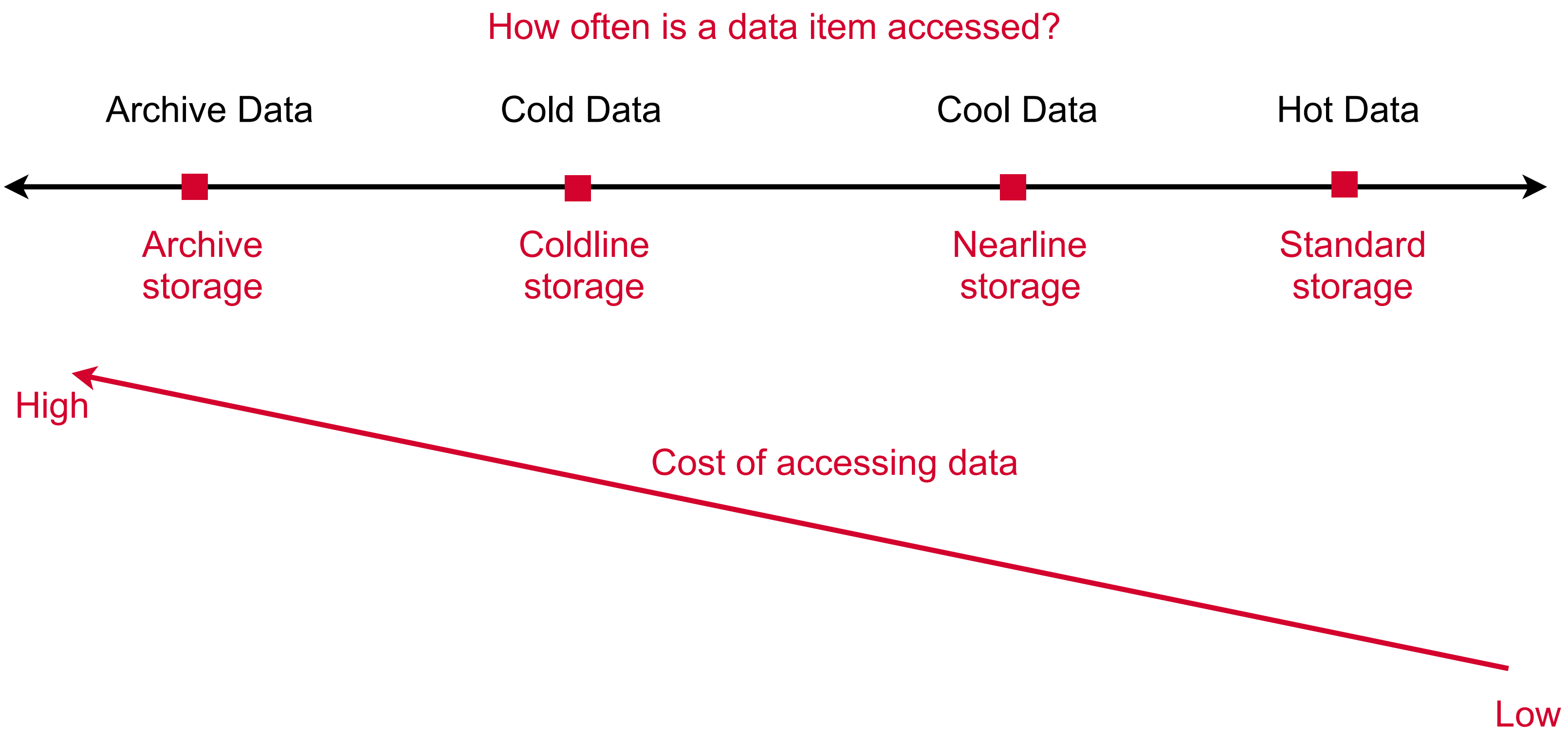
How often is a data item accessed?



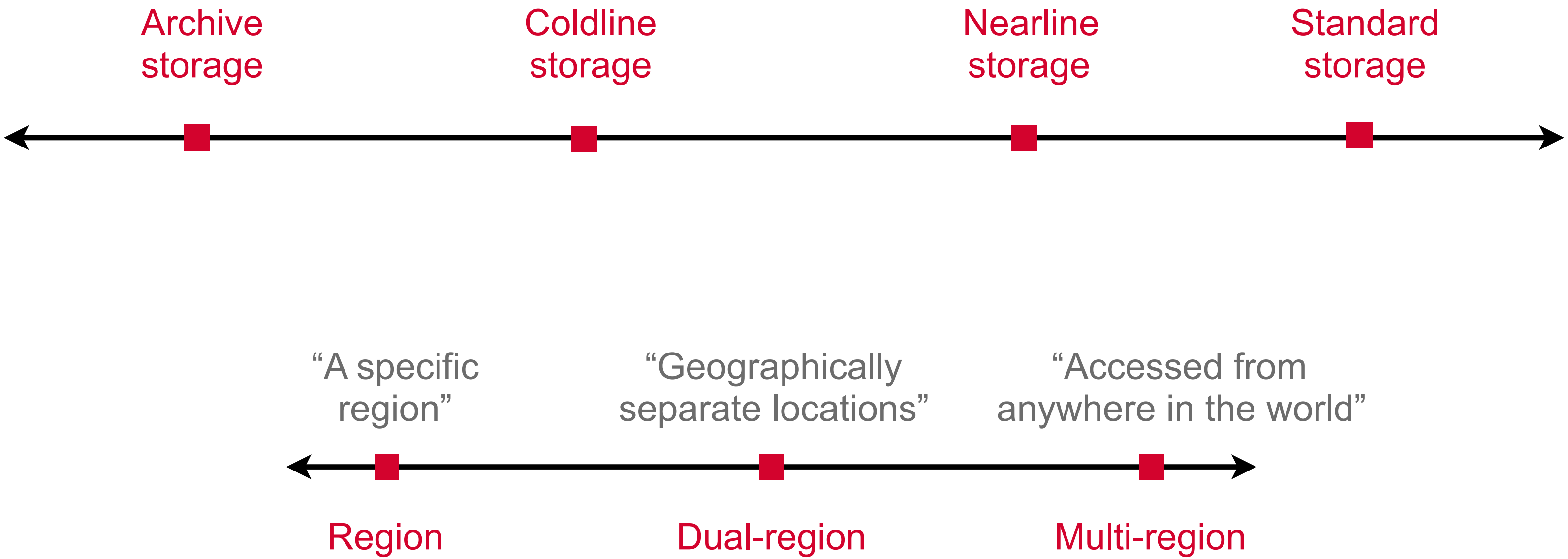
GCS Storage Classes



GCS Storage Classes



All Storage Classes



Autoclass

Moves data that is not accessed to colder storage classes to reduce cost

Moves data that is accessed to standard storage to optimize cost of future access

Coldline and Archive has about the same speed of access as other storage classes (different from AWS Glacier and S3)

Availability

Storage Costs

Retrieval Costs

Durability

Access Frequency

Use Cases

Different storage classes represent different trade-offs

Several parameters along which to compare

Availability

Storage Costs

Retrieval Costs

Durability

Access Frequency

Use Cases

Storage Class	Availability
Standard storage (dual and multi-regional)	99.95%
Standard storage (regional)	99.9%
Nearline (regional)	99.0%
Coldline (regional)	99.0%

Dual-region and multi-region buckets are tied to multi-regional locations: US, EU and Asia

Helps adhere to data storage regulations in the US and EU

Availability

Storage Costs

Retrieval Costs

Durability

Access Frequency

Use Cases

Storage Class	Storage Cost (cents/GB/month)
Standard	2.6
Nearline	1.0
Coldline	0.7
Archive	0.24

Availability

Storage Costs

Retrieval Costs

Durability

Access Frequency

Use Cases

Storage Class	Retrieval Cost (cents/GB)
Standard	**None**
Nearline	1.0
Coldline	2.0
Archive	5.0

- Availability
- Storage Costs
- **Retrieval Costs**
- Durability
- Access Frequency
- Use Cases

Storage Class	Minimum Commitment
Standard	**None**
Nearline	30 days*
Coldline	90 days*
Archive	365 days*

*Early deletion will incur charges

Availability

Storage Costs

Retrieval Costs

Durability

Access Frequency

Use Cases

Storage Class	Durability
Standard	99.999999999%
Nearline	99.999999999%
Coldline	99.999999999%
Archive	99.999999999%

“11 nines”

- Availability
- Storage Costs
- Retrieval Costs
- Durability
- **Access Frequency**
- Use Cases

Storage Class	Access Frequency
Standard	Daily
Nearline	Monthly
Coldline	Quarterly
Archive	Less than once a year

- Availability
- Storage Costs
- Retrieval Costs
- Durability
- Access Frequency
- **Use Cases**

Storage Class	Access Frequency
Standard storage (dual and multi-regional)	Serving websites, interactive workloads, mobile and gaming applications
Standard storage (regional)	Access from Compute Engine VMs or Dataproc cluster
Nearline	Data backup, disaster recovery, archival storage
Coldline/Archive	Legal or regulatory needs; also disaster recovery where recovery time is important

Object Versioning

- Needs to be enabled for bucket
- Once enabled, bucket creates archived versions of each object
- Whenever live object is overwritten or deleted
- Version with unique generation number is created
- Each copy charged separately

Object Lifecycle Management

- Can automatically specify changes to object storage class
  - “Change from regional to nearline after 30 days”
  - “Delete all data created before 1/8/2018”
  - “Delete all but 2 most recent versions”

Encryption

- Encrypted even at rest
- Default: Google generates keys
- Can use CSEK
  - *Customer Supplied Encryption Key*

GCS for Object Storage

File Storage

- Hierarchical structure
- Support for nesting and directories
- File-level locks
- File and directory headers

Object Storage

- Flat, non-nested structure
- Nested structure merely simulated
- No distributed lock - last write wins
- Unstructured series of bytes

Storage Class

Which of the following is true for coldline storage?

1. Low cost of storage, high cost of retrieval
2. Low cost of storage, low cost of retrieval
3. High cost of storage, low cost of retrieval
4. High cost of storage, high cost of retrieval

Storage Class

Which of the following is true for coldline storage?

1. **Low cost of storage, high cost of retrieval**
2. Low cost of storage, low cost of retrieval
3. High cost of storage, low cost of retrieval
4. High cost of storage, high cost of retrieval

Storage Use Cases

Use Case	Appropriate GCP Service	Non-GCP Equivalents
Block storage	Persistent disks or local SSDs	AWS EBS, Azure Disk
Object/blob storage	Cloud Storage (GCS) buckets	AWS S3, Azure Blob Storage
Relational data - small, regional payloads	Cloud SQL	AWS RDS, Azure SQL Database
Relational data - large, global payloads	Cloud Spanner	Aurora DB
HTML/XML documents with NoSQL access	Firestore	AWS DynamoDB, Azure Cosmos DB
Large, naturally ordered data with NoSQL access	BigTable	
Analytics and complex queries with SQL access	BigQuery	AWS Redshift, Azure Synapse Analytics

Cloud SQL

Cloud SQL is the fully-managed MySQL, PostgreSQL and SQL Server database service on the Google Cloud Platform

Transactional support, ACID support

Easiest migration path for on-prem RDBMS

High availability using failover replicas in different zones

Google Cloud Spanner

A global, horizontally scaling, strongly consistent relational database service built on proprietary technology

Scales horizontally by adding nodes

ACID support at scale

Relatively expensive and Google proprietary

Cloud Firestore

Flexible, scalable, NoSQL database for keeping data in sync across client apps.

Mobile and web server development as a part of GCP's Firebase platform

Realtime listeners and offline support

GCP vs. Firebase

GCP

- Makes Google's infrastructure publicly available as services
- Main users are server-side and backend developers
- Services focus on leveraging Google's core infrastructure
- Networking, storage, machine learning, traffic management, scaling

GCP vs. Firebase

GCP

- Makes Google’s infrastructure publicly available as services
- Main users are server-side and backend developers
- Services focus on leveraging Google’s core infrastructure
- Networking, storage, machine learning, traffic management, scaling

Firebase

- Build mobile and web applications quickly
- Mainly used by client-side application developers
- Services to build applications, engage and grow users
- Realtime database, crashlytics, performance management, messaging

BigQuery Features

- Serverless: No cluster, no provisioning
- Structured data with fields
- Can ingest streaming data at scale
- Autoscaling
- Automatic high availability
- Simple SQL queries

Redis

Very popular in-memory key-value NoSQL database

Memcached

General purpose, distributed, memory-caching system

Cloud Memorystore

Google managed service for Redis and Memcached that offers scaling, high availability and a convenient migration path

Google Cloud Bigtable

NoSQL database technology ideal for very large, sparse datasets with sequential ordering in key column; provides very fast writes as well as reads

Choose Bigtable For

- Time series data: Naturally ordered
- Internet of Things data: Constant stream of writes
- Financial data: Often efficiently represented as time series data
- Large datasets > 1 TB with each row < 10 MB

Storage

You have about 5TB of data on your on-premises MySQL database, you want to lift and shift this to the GCP. Which storage technology would you use?

1. Cloud SQL
2. Cloud Spanner
3. BigQuery
4. Cloud Memorystore

Storage

You have about 5TB of data on your on-premises MySQL database, you want to lift and shift this to the GCP. Which storage technology would you use?

1. **Cloud SQL**
2. Cloud Spanner
3. BigQuery
4. Cloud Memorystore

Storage

You have a financial application where transaction support is critical and your clients are distributed globally. Which GCP technology would you use?

1. Cloud SQL
2. Cloud Spanner
3. BigQuery
4. Cloud Memorystore

Storage

You have a financial application where transaction support is critical and your clients are distributed globally. Which GCP technology would you use?

1. Cloud SQL
2. **Cloud Spanner**
3. BigQuery
4. Cloud Memorystore

Storage

You are building a chat application within your product and you want your users to get realtime message updates. Which GCP technology would you choose?

1. Cloud Firestore
2. Cloud Spanner
3. BigQuery
4. Cloud Bigtable

Storage

You are building a chat application within your product and you want your users to get realtime message updates. Which GCP technology would you choose?

1. **Cloud Firestore**
2. Cloud Spanner
3. BigQuery
4. Cloud Bigtable

Storage

You have a realtime stock market application that stores stores stock prices at every tick. You want extremely low latency access to price data at 5 minute intervals. What GCP technology would you choose?

1. Cloud Firestore
2. Cloud Spanner
3. BigQuery
4. Cloud Bigtable

Storage

You have a realtime stock market application that stores stores stock prices at every tick. You want extremely low latency access to price data at 5 minute intervals. What GCP technology would you choose?

1. Cloud Firestore
2. Cloud Spanner
3. BigQuery
4. **Cloud Bigtable**

O'REILLY

Networking

IP addresses, routes, and firewall rules all exist inside a GCP resource called a VPC Network

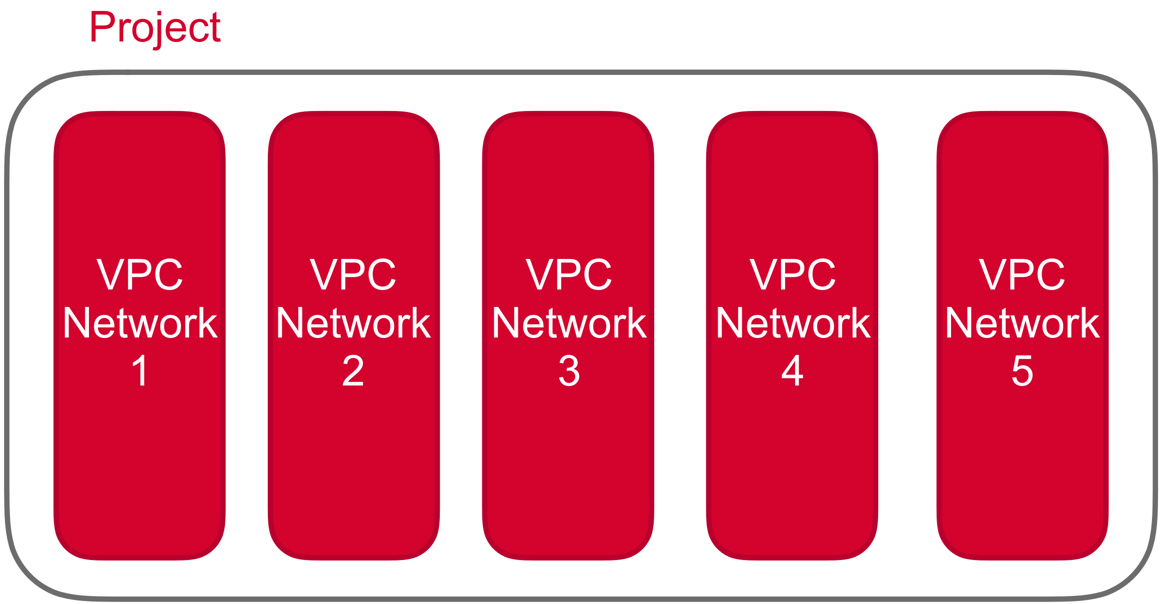
Google Virtual Private Cloud

A VPC network, often just called a network, is a global, private, isolated virtual network partition that provides managed network functionality

Google Virtual Private Cloud

A VPC network, often just called a network, is a global, private, isolated virtual network partition that provides managed network functionality

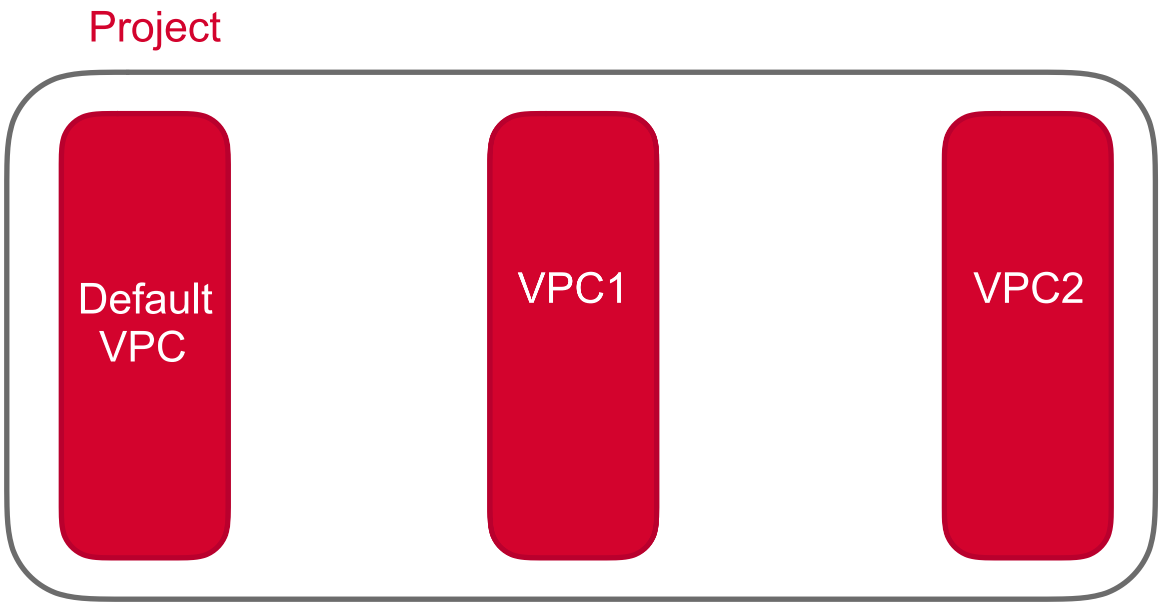
Multiple VPCs in a Project



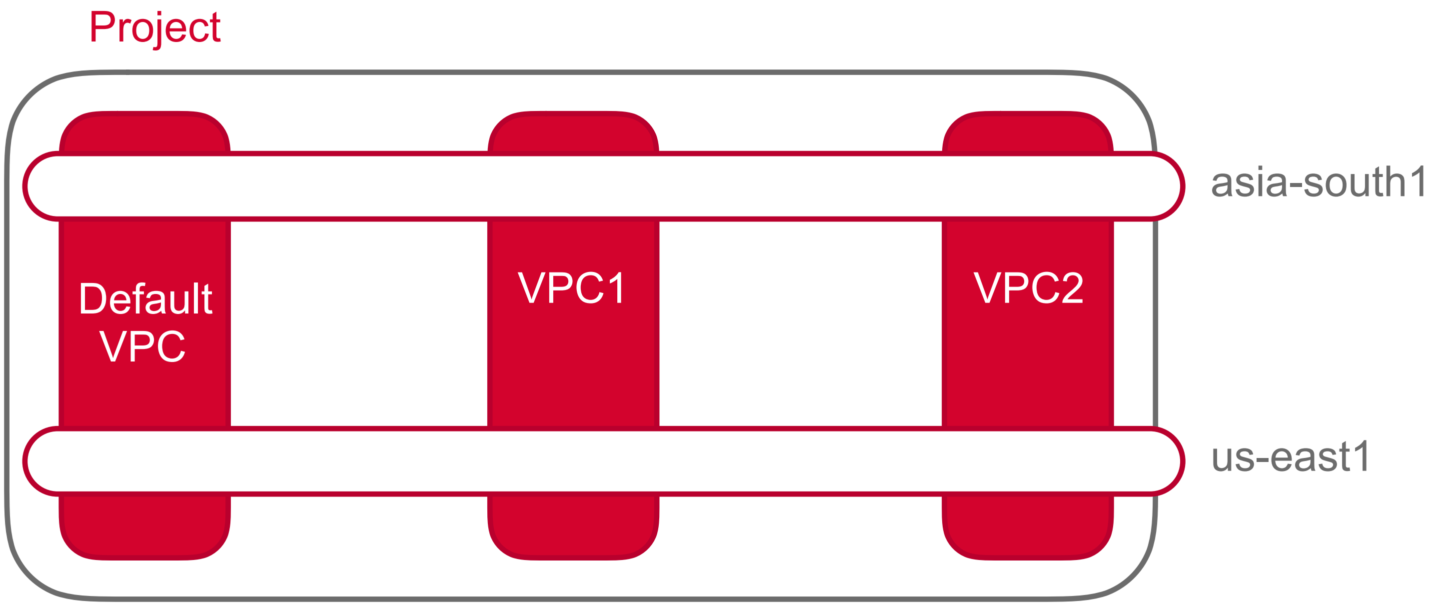
Projects and VPCs

- VPCs are global resources on the GCP
- Each VPC must exist inside a project
- Default VPC pre-created in each project

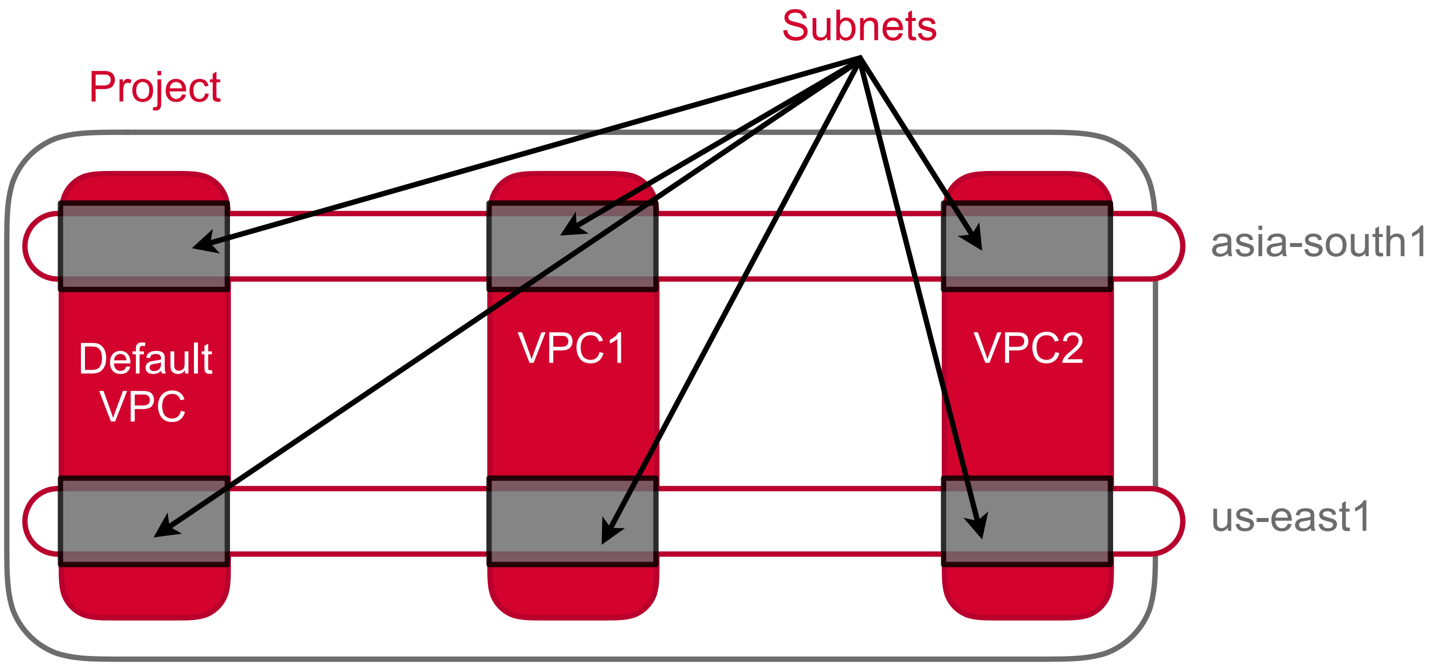
VPCs Are Global



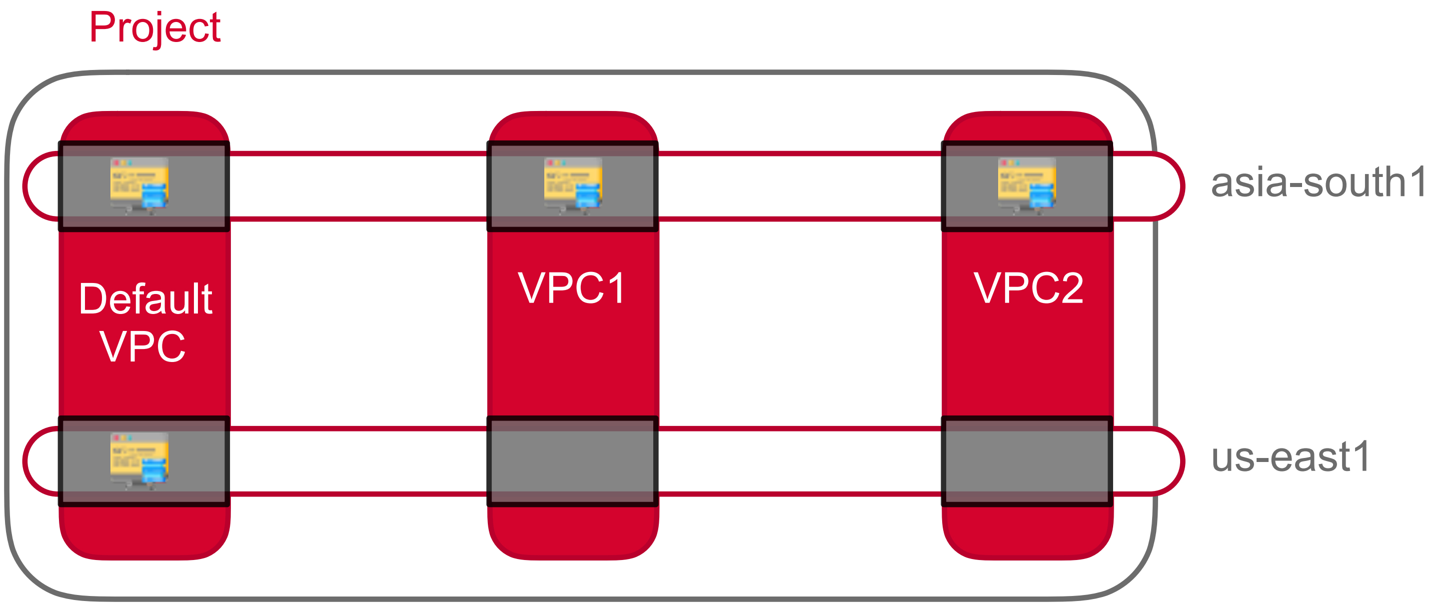
VPCs Are Global



Subnets in Each Region



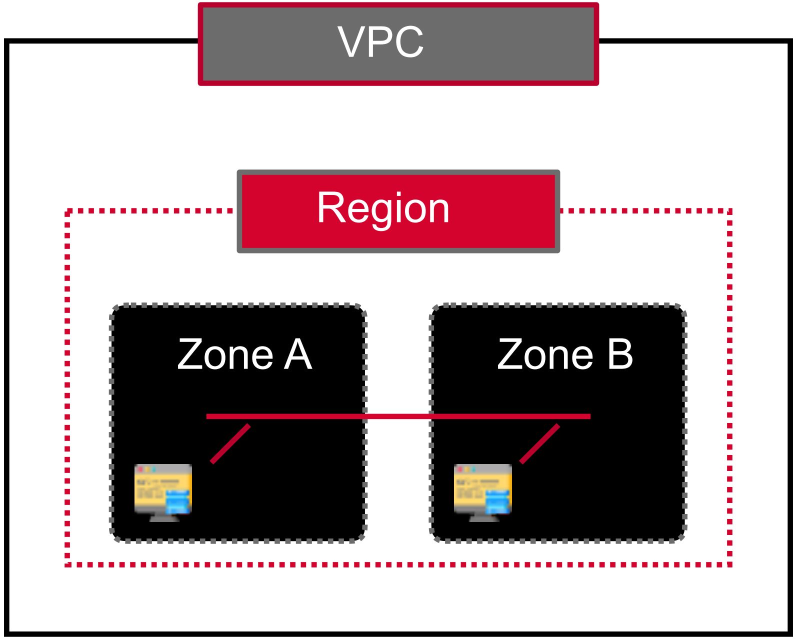
Resources Provisioned on Subnets



Subnets

- **IP range partitions** within global VPCs
- VPCs have no IP ranges
- Subnets are regional - can span zones inside a region
- Network has to have at least one subnet before you can use it

Subnets Span Zones



Subnets and IP Ranges

- Each subnet must have primary address range
- Valid RFC 1918 CIDR block
- Subnet ranges in same network cannot overlap
- Subnet ranges in different networks can overlap

AutoMode and CustomMode VPCs

Auto Mode

Subnets automatically created in each region, default firewall rules

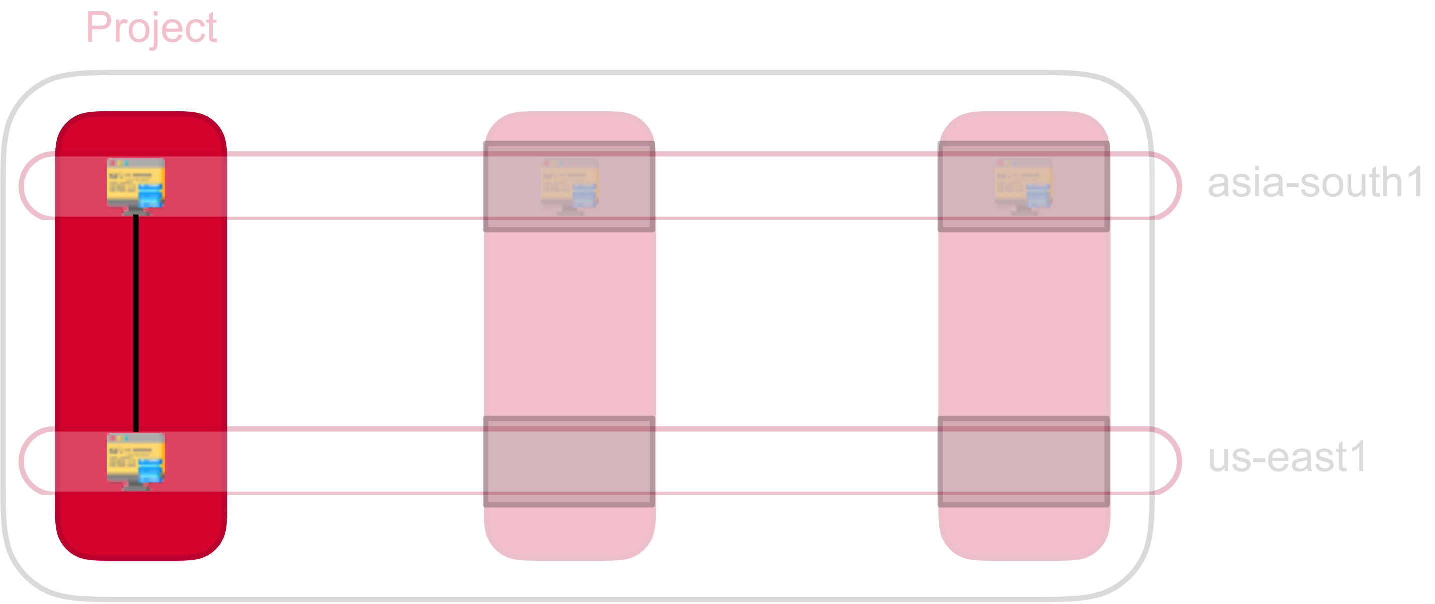
Custom Mode

Manually create subnets in regions, no defaults preconfigured

Auto Mode and Custom Mode VPCs

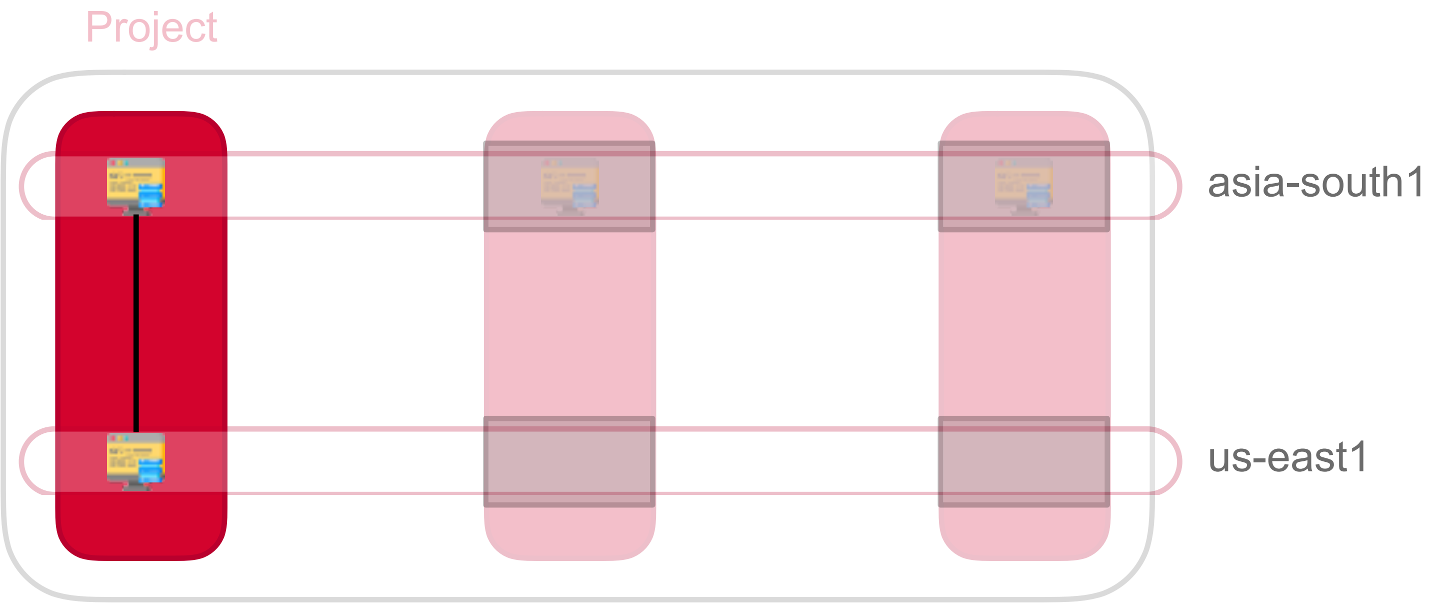
- Auto Mode VPCs have pre-created subnets
  - One in each GCP region
- Custom Mode VPCs start with no subnets
  - Full control over which regions have subnets
  - Can create multiple subnets in a region

Communication on VPCs



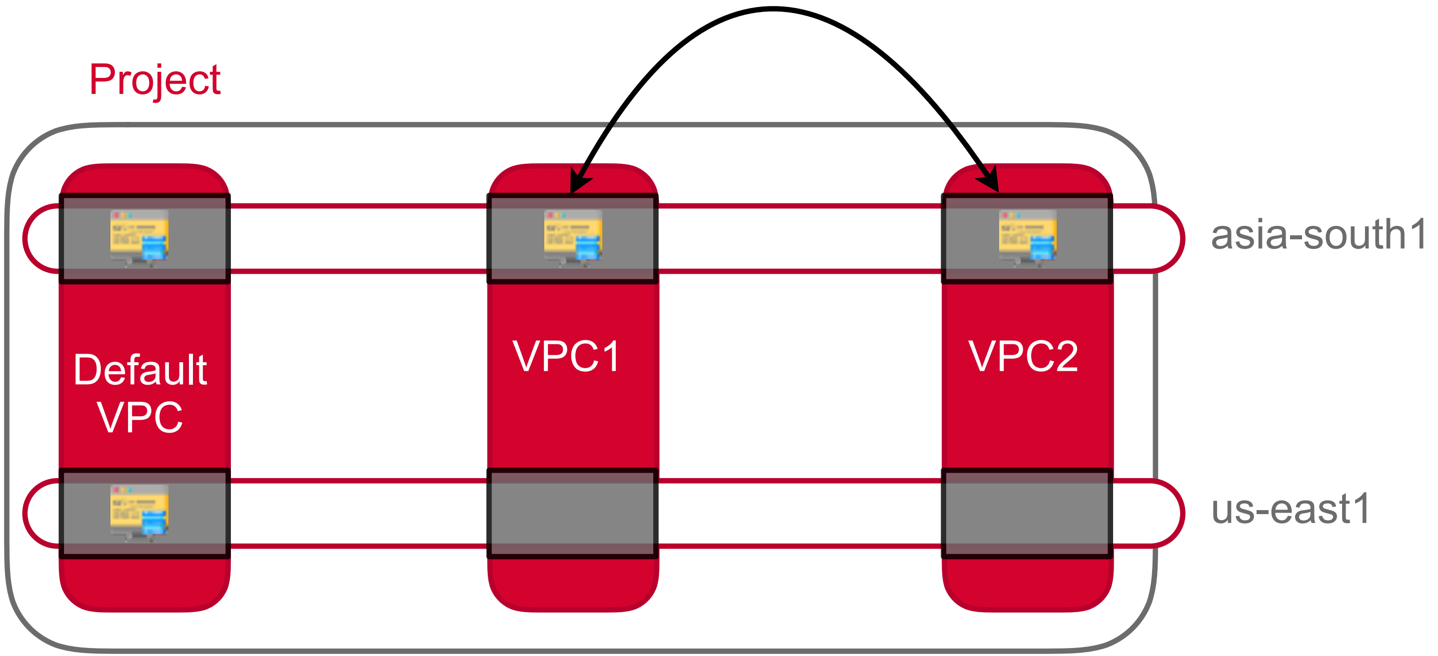
Resources within a VPC communicate using private IP addresses

Communication on VPCs



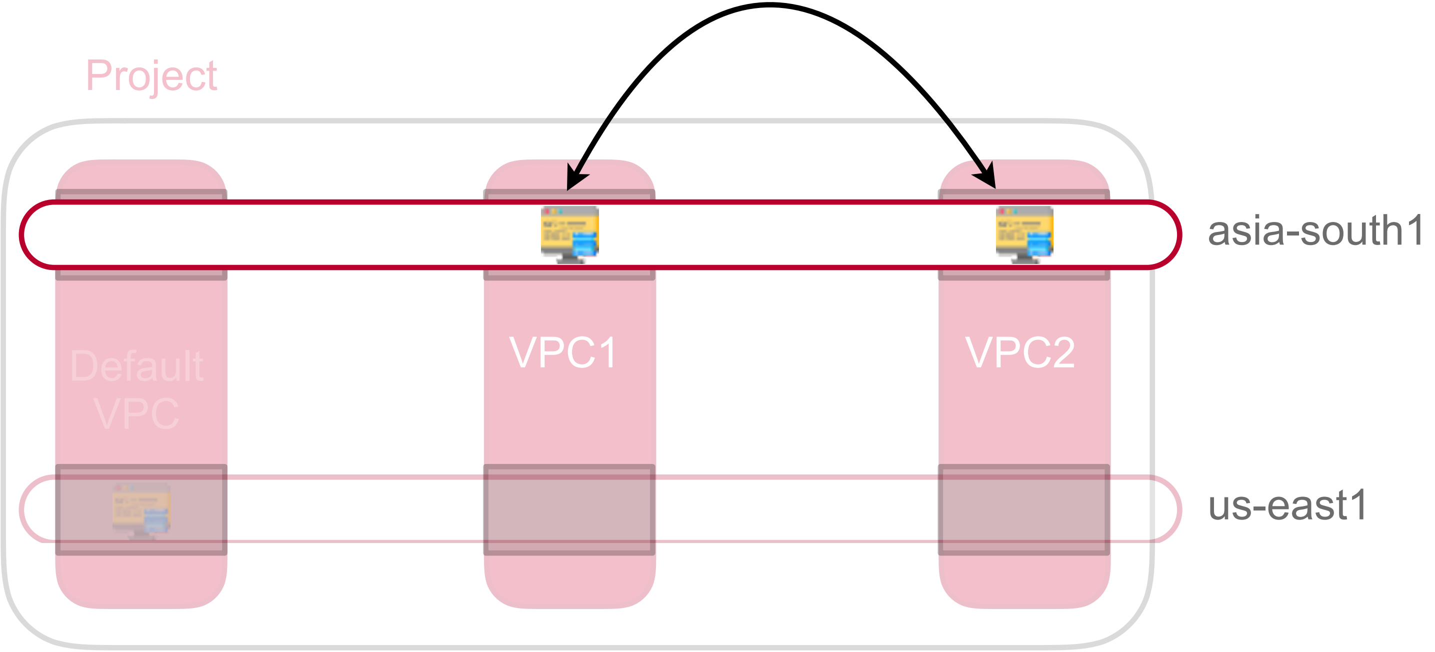
Wherever they are located in the world - irrespective of physical location

Communication on VPCs



Resources on different VPCs communicate over the internet using external IPs

Communication on VPCs



Even though they are in the same region - they may even be in the same zone on the same physical hardware

Default VPC

- Pre-created on every project
- Includes subnet for each GCP region
- New subnets added when new regions are created
- Resources created here by default

Default VPC

- Includes routes for all resources
- All VMs on the default VPC can talk to each other
- Default gateway to internet
- Includes several firewall rules

Firewall Rules

- Every VPC is a distributed firewall
- Firewall rules defined in VPC
- Are applied on per-instance basis
- Can also regulate internal traffic

Firewall Rules

- Every VPC has two permanent rules
    - Implied allow egress
    - Implied deny ingress
- Can be overridden by more specific rules
- In addition, default VPC has several rules

Additional Rules in Default VPC

- default-allow-internal
- default-allow-ssh
- default-allow-rdp
- default-allow-icmp

Networking

Which of the following is true for GCP subnets?

1. They are zonal resources
2. They are global resources
3. Every resource has to be provisioned on a subnet
4. They are physical network partitions

Networking

Which of the following is true for GCP subnets?

1. They are zonal resources
2. They are global resources
3. **Every resource has to be provisioned on a subnet**
4. They are physical network partitions

Networking

How do GCP resources in the same region but on different VPCs communicate with each other?

1. Using private IP addresses
2. Using external IP addresses
3. They cannot communicate with each other
4. Using hostnames

Networking

How do GCP resources in the same region but on different VPCs communicate with each other?

1. Using private IP addresses
2. **Using external IP addresses**
3. They cannot communicate with each other
4. Using hostnames

Networking

Which of the following statements is true for the default VPC?

1. They cannot be manually configured once set up
2. They allow external clients to send traffic to all resources by default
3. They come with no firewall rules configured
4. Subnets in new GCP regions are automatically added

Networking

Which of the following statements is true for the default VPC?

1. They cannot be manually configured once set up
2. They allow external clients to send traffic to all resources by default
3. They come with no firewall rules configured
4. **Subnets in new GCP regions are automatically added**

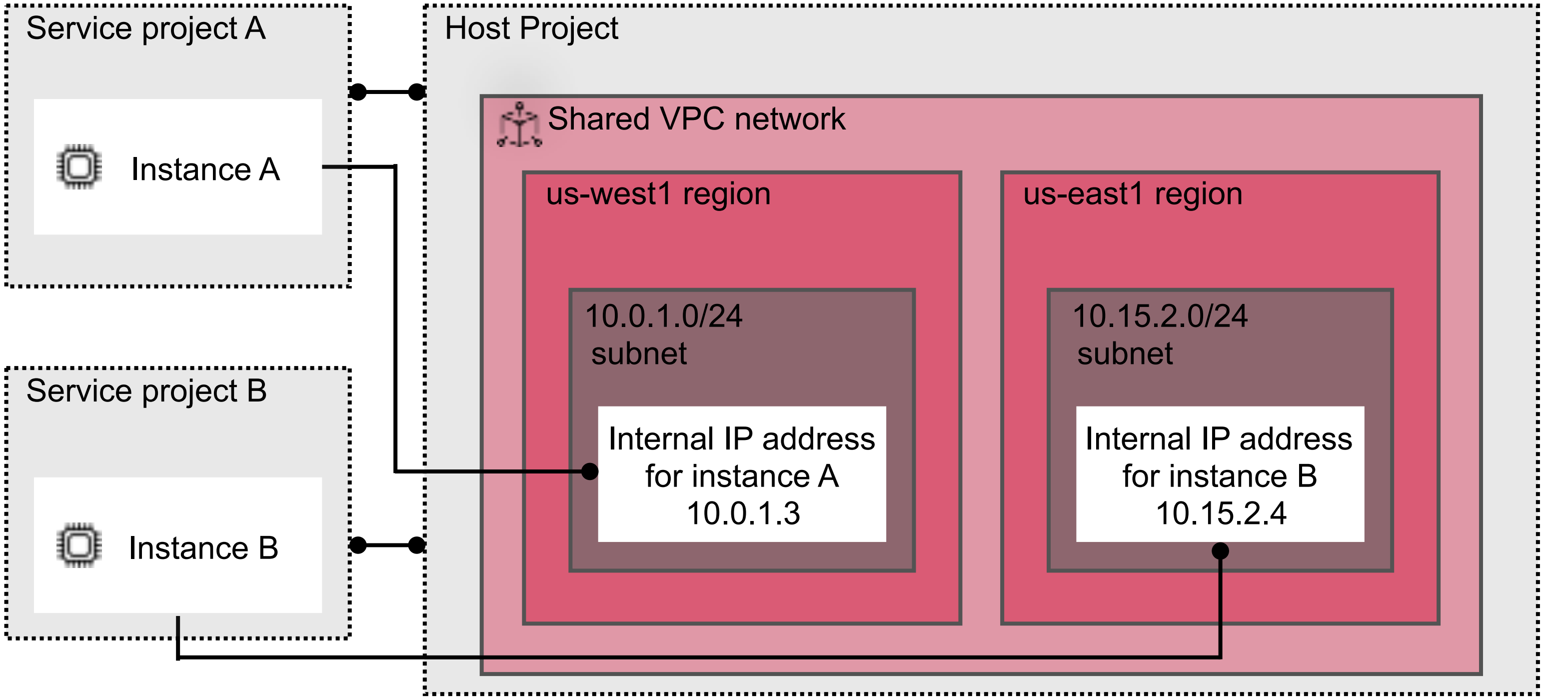
O'REILLY

Connecting Networks

Shared VPC

- Share VPC across projects on GCP
- One VPC shared across projects
- Projects must be in the same organization
- Host project, guest resources
- Shared VPC admin to administer the shared VPC

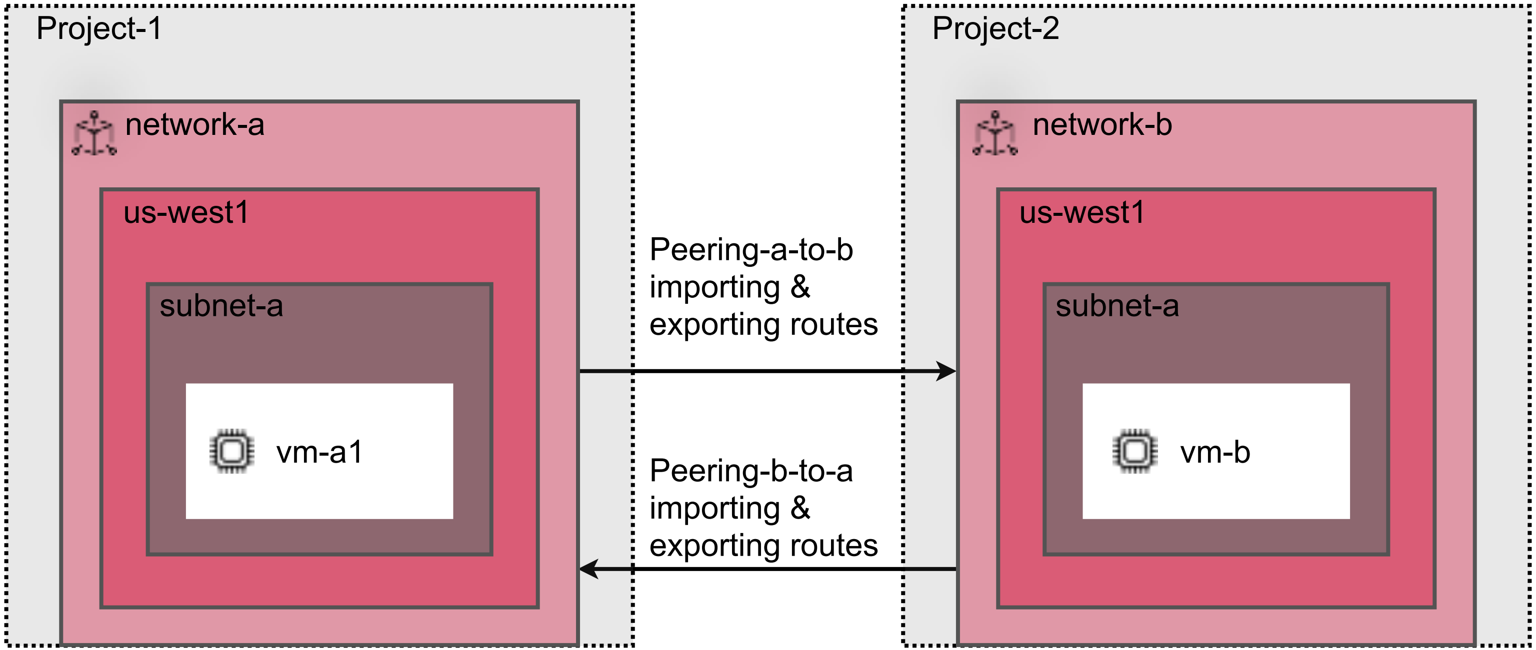
Shared VPC


Service project A
Instance A
Service project B
Instance B
Host Project
Shared VPC network
us-west1 region
us-east1 region
10.0.1.0/24
subnet
10.15.2.0/24
subnet
Internal IP address
for instance A
10.0.1.3
Internal IP address
for instance B
10.15.2.4

VPC Peering

- Two or more VPCs shared across projects
- Projects need not be in the same organization
- Allows resources on different VPC networks to communicate using internal IP addresses
- Resources on the network use Google infrastructure to communicate
- Reduced latency, higher security and lower cost as compared with using external IPs

VPC Peering


Project-1
network-a
us-west1
subnet-a
vm-a1
Project-2
network-b
us-west1
subnet-a
vm-b
Peering-a-to-b
importing &
exporting routes
Peering-b-to-a
importing &
exporting routes

Shared VPCs vs. Network Peering

Shared VPCs

- Only within **same organization**
- One VPC used across projects
- Host and service projects not peers
- Single level of sharing possible

Network Peering

- Across **organization boundaries**
- Multiple VPCs share resources
- Connected VPCs are peers
- Multiple levels of peering possible

Interconnecting Networks

GCP-to-GCP

VPC Network Peering

Enterprise connectivity

Peering and interconnect options

Interconnecting Networks

GCP-to-GCP

VPC Network Peering

Enterprise connectivity

Peering and interconnect options

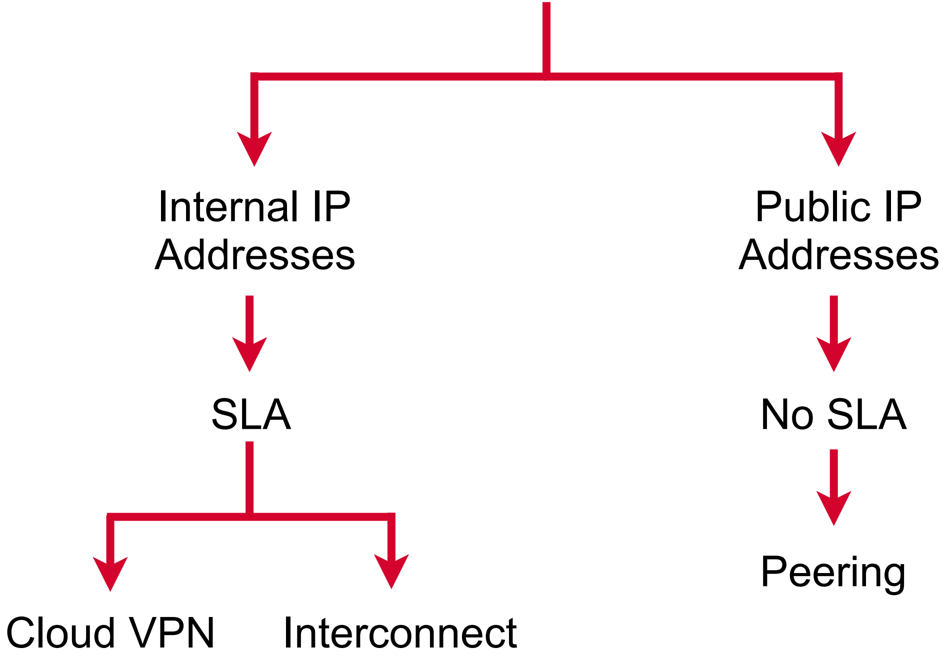
Connect a cloud network with an on-premise network using private or public IP addresses

Enterprise Connectivity

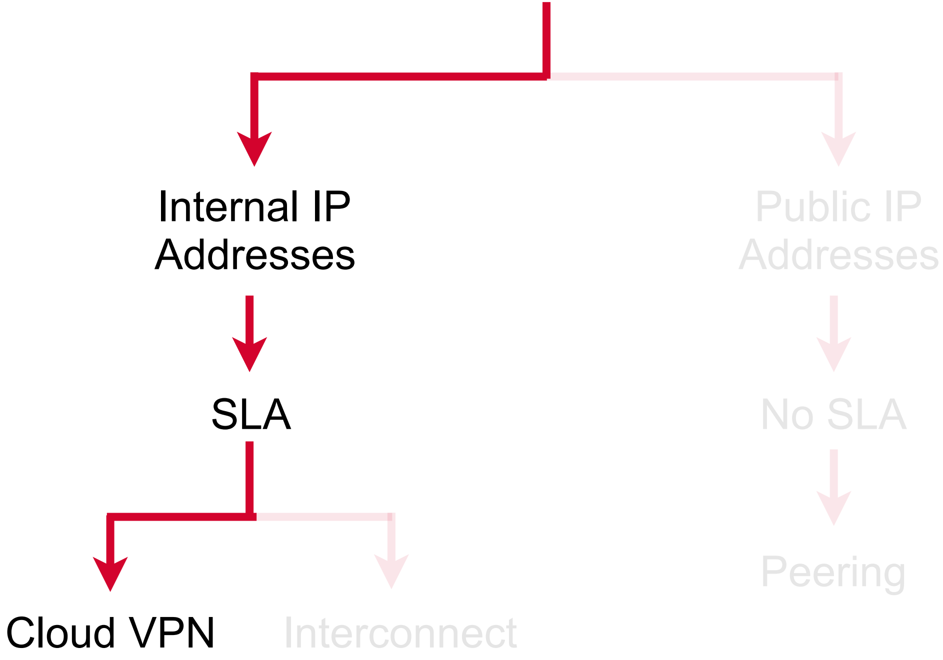
Internal IP Addresses

Public IP Addresses

Enterprise Connectivity



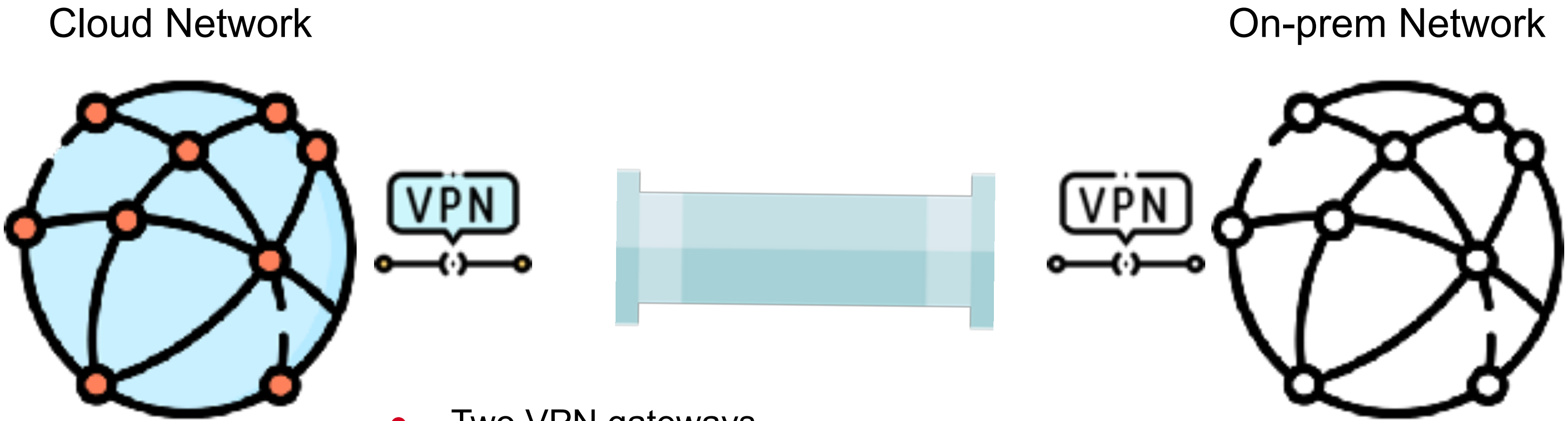
Enterprise Connectivity



Cloud VPN

Configuration Property	Choice
Connection	**Encrypted tunnel to VPC networks through the public internet**
Access Type	**Internal IP** addresses in RFC 1918 address space
Capacity	1.5-3 Gbps for each tunnel
Other Considerations	**Requires a VPN device on your on-premises network**

Cloud VPN



- Two VPN gateways
- One for cloud network, another for on-prem network

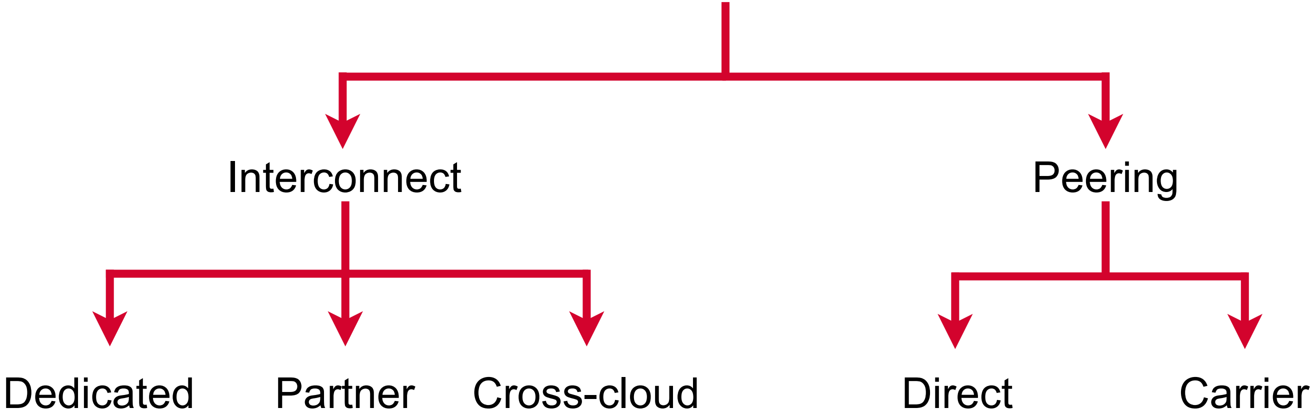
- Traffic encrypted at one gateway
- Decrypted at other gateway
- Keys need to be exchanged

Enterprise Connectivity

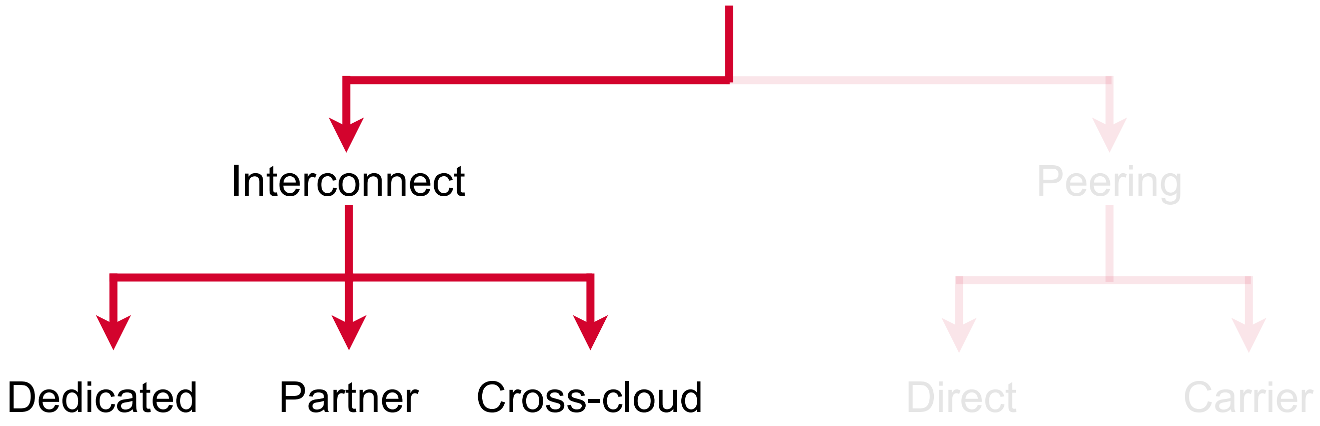
Interconnect

Peering

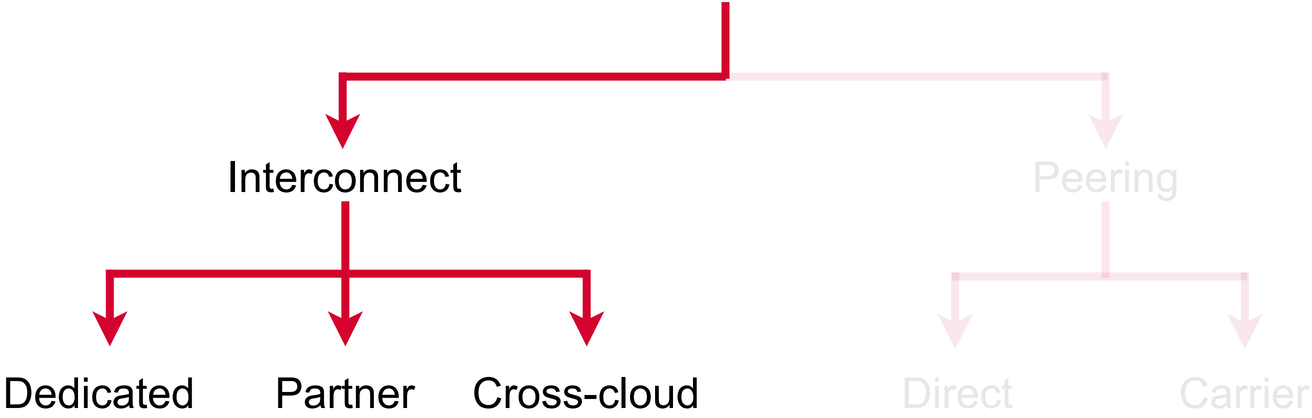
Enterprise Connectivity



Enterprise Connectivity



Enterprise Connectivity



Internal IP addresses in
RFC 1918 address space

With SLA

Traffic between your external network and Google network DOES NOT traverse the public internet

Dedicated Interconnect

Configuration Property	Choice
Connection	Dedicated, direct connection to VPC networks
Access Type	Internal IP addresses in RFC 1918 address space
Capacity	10 Gbps or 100 Gbps connections
Other Considerations	**Must have connection in a Google supported colocation facility that supports the regions you want to connect to**

Partner Interconnect

Configuration Property	Choice
Connection	Dedicated Bandwidth, connection to VPC network through a service provider
Access Type	Internal IP addresses in RFC 1918 address space
Capacity	50Mbps - 50Gbps per connection
Other Considerations	**Service providers might have specific restrictions or requirements**

Cross-cloud Interconnect

- High-bandwidth dedicated connectivity between Google Cloud and another service provider
- Google will provision a dedicated physical connection
- Useful for:
  - Site-to-site data transfer
  - Multi-cloud strategy

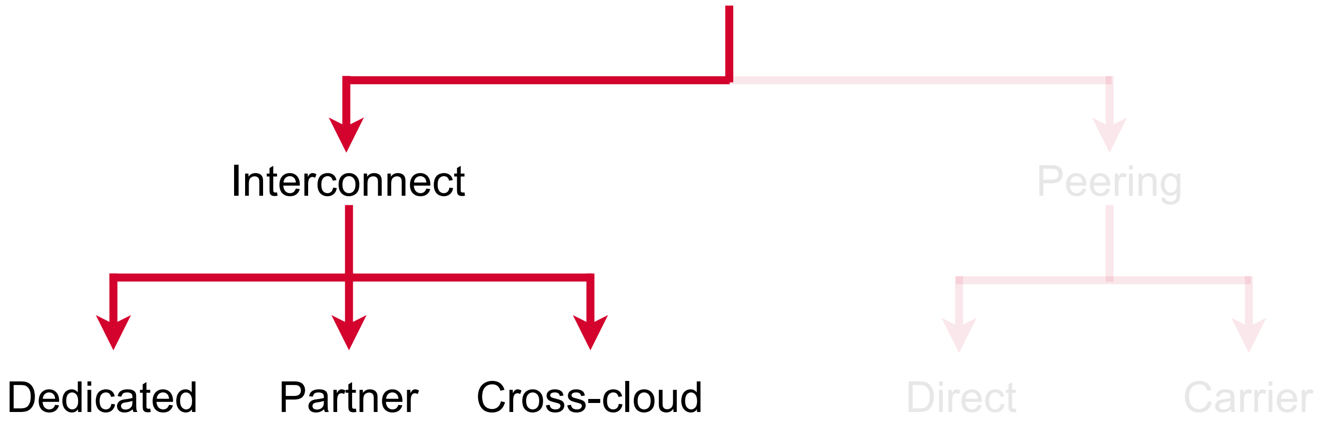
Cross-cloud Interconnect

Configuration Property	Choice
Connection	Dedicated physical connection between Google Cloud and other cloud platform
Access Type	Internal IP addresses in RFC 1918 address space
Capacity	10 Gbps or 100Gbps
Other Considerations	**Supported cloud provides AWS, Azure, Oracle, Alibaba**

Cloud Router

- Cloud Router is a fully distributed and managed Google Cloud service that **dynamically manages routing tables**
- Uses the Border **Gateway Protocol (BGP)** to exchange routes between Google Cloud and on-premise networks
- Allows for **automatic updation** when network changes occur
- Used with Cloud Interconnect and Cloud VPN

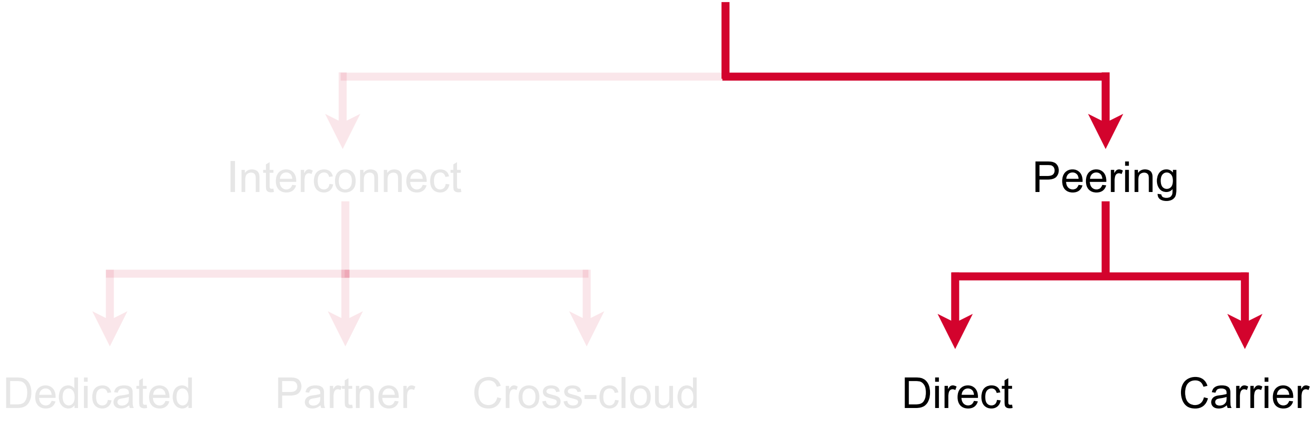
Enterprise Connectivity



Internal IP addresses in
RFC 1918 address space

With SLA

Enterprise Connectivity



Public IP addresses

No SLA

Direct Peering

Configuration Property	Choice
Connection	Provides direct access from your on-premises network to Google Workspace and Google APIs for the full suite of Google Cloud products.
Access Type	Public IP addresses
Other Considerations	**Connects to Google's edge network**

Carrier Peering

Configuration Property	Choice
Connection	Peering through service provider to access Google applications such as Google Workspace and to Google Cloud products that can be exposed through one or more public IP addresses.
Access Type	Public IP addresses
Other Considerations	**Connects to Google's edge network through a service provider. Requirements vary by partner**

Networking

Which of the following is a difference between using Shared VPC and Peering to interconnect networks in different GCP projects?

1. Shared VPC can span projects in multiple organizations but Peering cannot
2. Shared VPC cannot span projects in multiple organizations but Peering can
3. Shared VPC offers lower latency as compared with Peering
4. Shared VPCs allow communication using internal IPs but with Peering you use external IPs

Networking

Which of the following is a difference between using Shared VPC and Peering to interconnect networks in different GCP projects?

1. Shared VPC can span projects in multiple organizations but Peering cannot
2. **Shared VPC cannot span projects in multiple organizations but Peering can**
3. Shared VPC offers lower latency as compared with Peering
4. Shared VPCs allow communication using internal IPs but with Peering you use external IPs

Networking

Among the following interconnect options in the GCP, which one requires your on premise network to physically meet Google's network in a colocation facility?

1. VPN Tunnel
2. Carrier Peering
3. Dedicated Interconnect
4. Partner Interconnect

Networking

Among the following interconnect options in the GCP, which one requires your on premise network to physically meet Google's network in a colocation facility?

1. VPN Tunnel
2. Carrier Peering
3. **Dedicated Interconnect**
4. Partner Interconnect

O'REILLY

VPC Service Controls

VPC Service Controls

Help protect against accidental or targeted data exfiltration risks from Google Cloud services such as Cloud Storage and BigQuery.

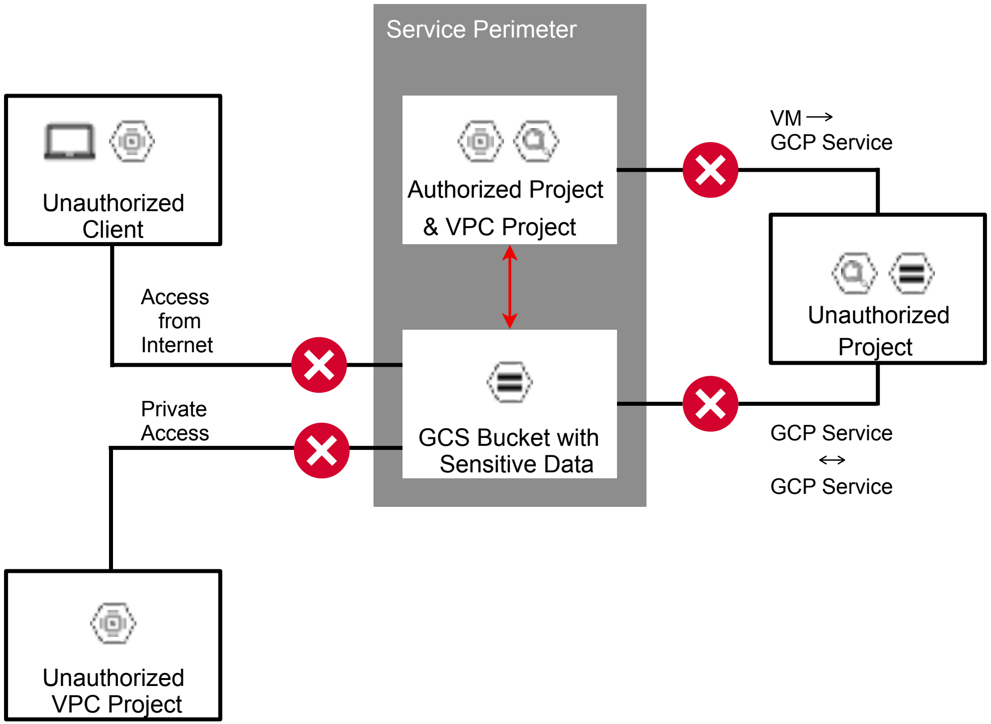
Creates service perimeters that protect the resources or data that you specify

Service Perimeter

A **service perimeter** creates a security boundary around Google Cloud resources.

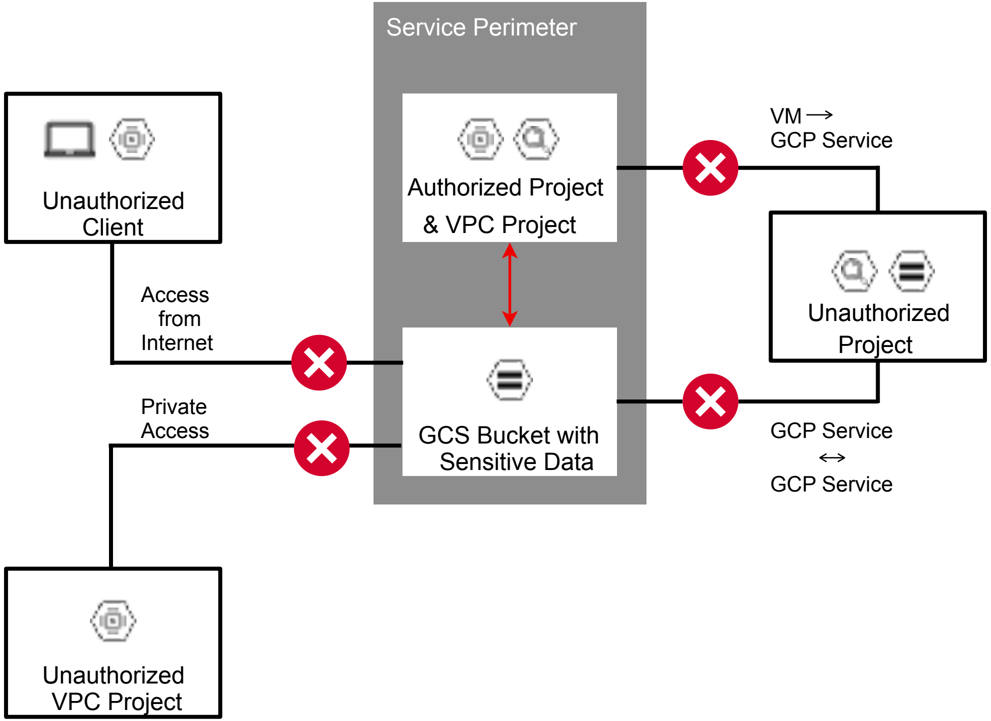
A service perimeter allows free communication within the perimeter but, by default, blocks communication to Google Cloud services across the perimeter.

Define Security Perimeters



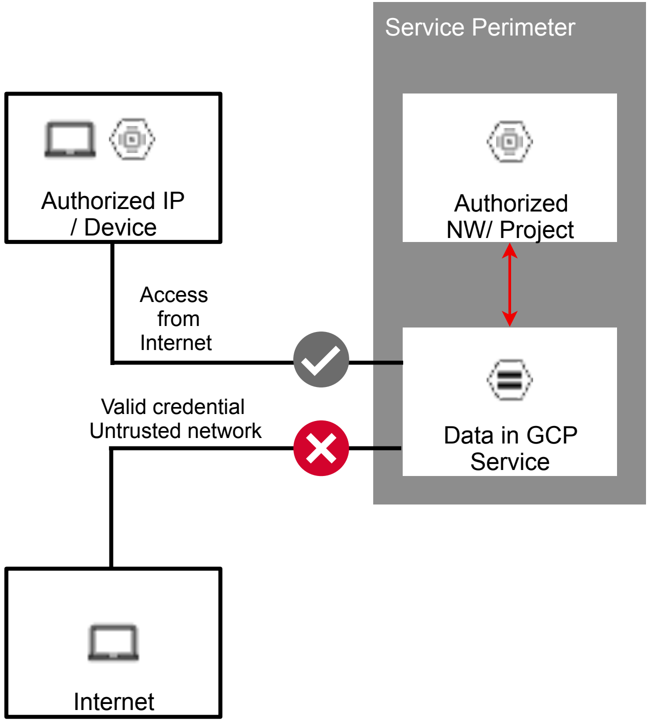
Include resources with sensitive data
Include services with access to those resources

Control Data Movement In and Out of Perimeter



Data cannot be copied to unauthorized resources outside the perimeter
Data exchange across the perimeter controlled by ingress and egress rules

Context Aware Access



Based on identity of the user, device state, network origin, other context signals

O'REILLY

Identity and Access Management

Cloud IAM

Manage identity and access control by defining who (identity) has what access (role) for which resource.

Cloud IAM

Permission to access a resource is not granted directly to the end user. Instead, permissions are grouped into roles, and roles are granted to authenticated principals.

Cloud IAM

Permission to access a resource is not granted directly to the end user. Instead, permissions are grouped into roles, and roles are granted to authenticated principals.

- Principal: GCP identity - user, group, service account
- Role: Collection of permissions
- Policy: Binding members to a role

Role-based Access Control

Identity

Permissions

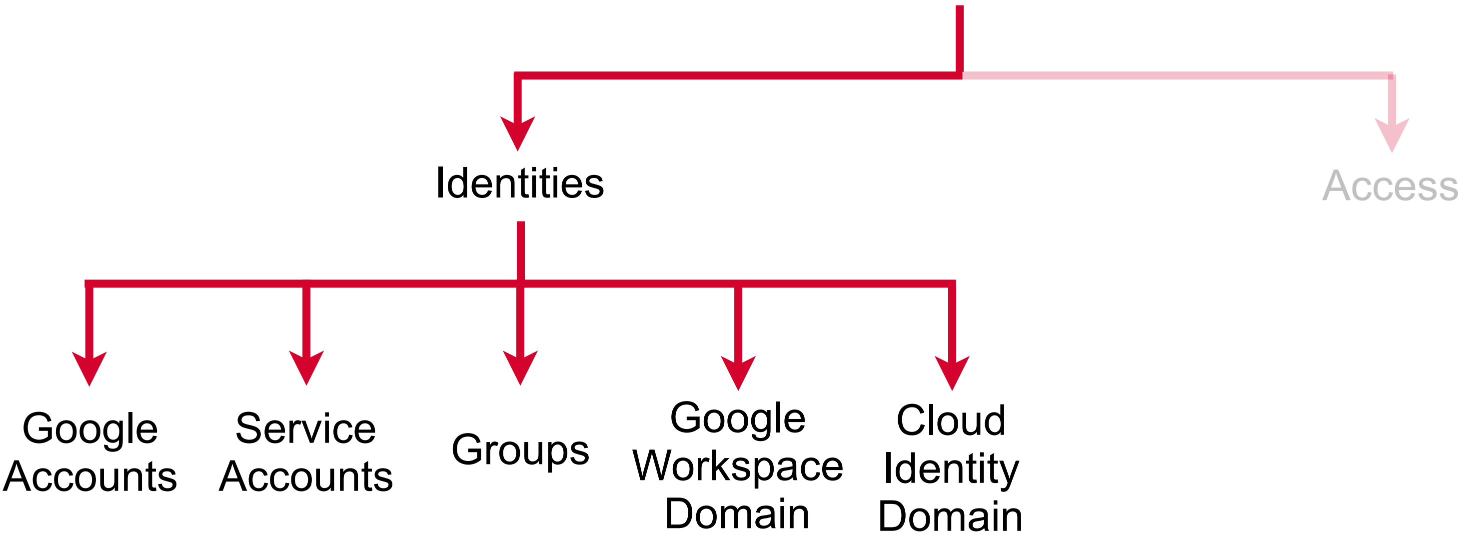
Resource

Identity and Access Management (IAM)

Identities

Access

Identity and Access Management (IAM)



Google Accounts

A Google account represents a developer, an administrator, or any other person who interacts with GCP.

Service Accounts

A service account is an account that belongs to your application instead of to an individual end user.

Google Groups

A Google Group is a named collection of Google accounts and service accounts. Every group has a unique email address that is associated with the group.

Google Workspace Domains

A Google Workspace domain represents a virtual group of all the Google accounts that have been created in an organization’s account.

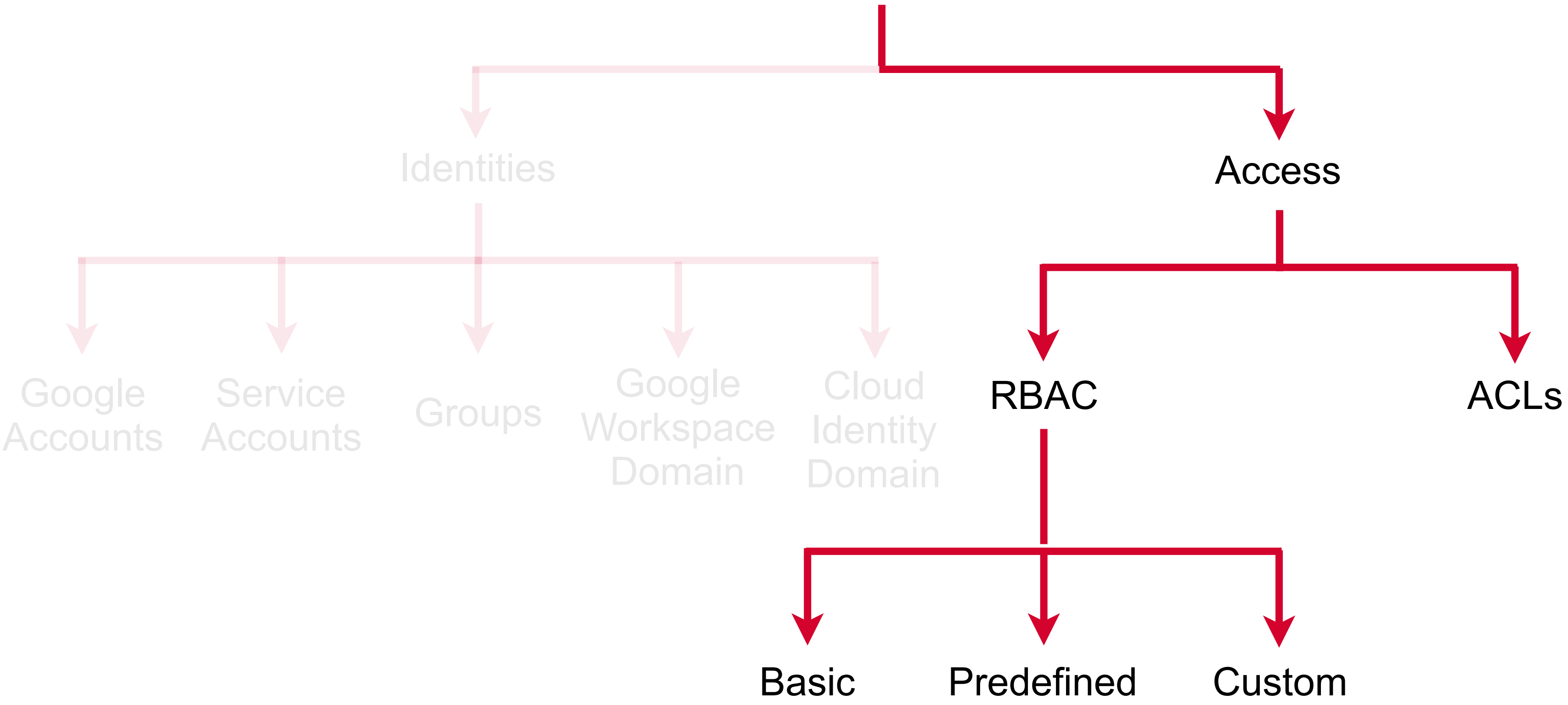
Google Workspace domains represent your organization's Internet domain name.

Cloud Identity Domains

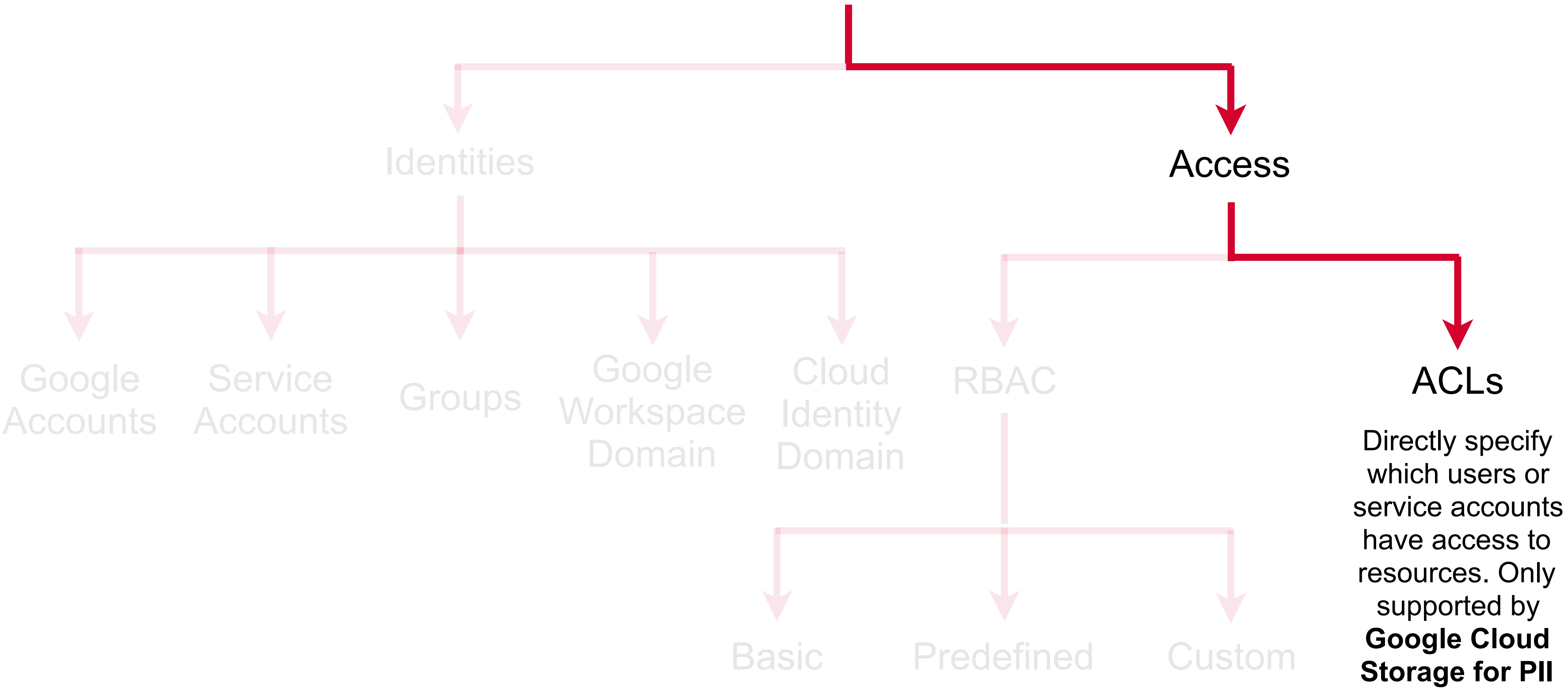
A Cloud Identity domain is like a Google Workspace domain because it represents a virtual group of all Google accounts in an organization.

However, Cloud Identity domain users don't have access to Google Workspace applications and features.

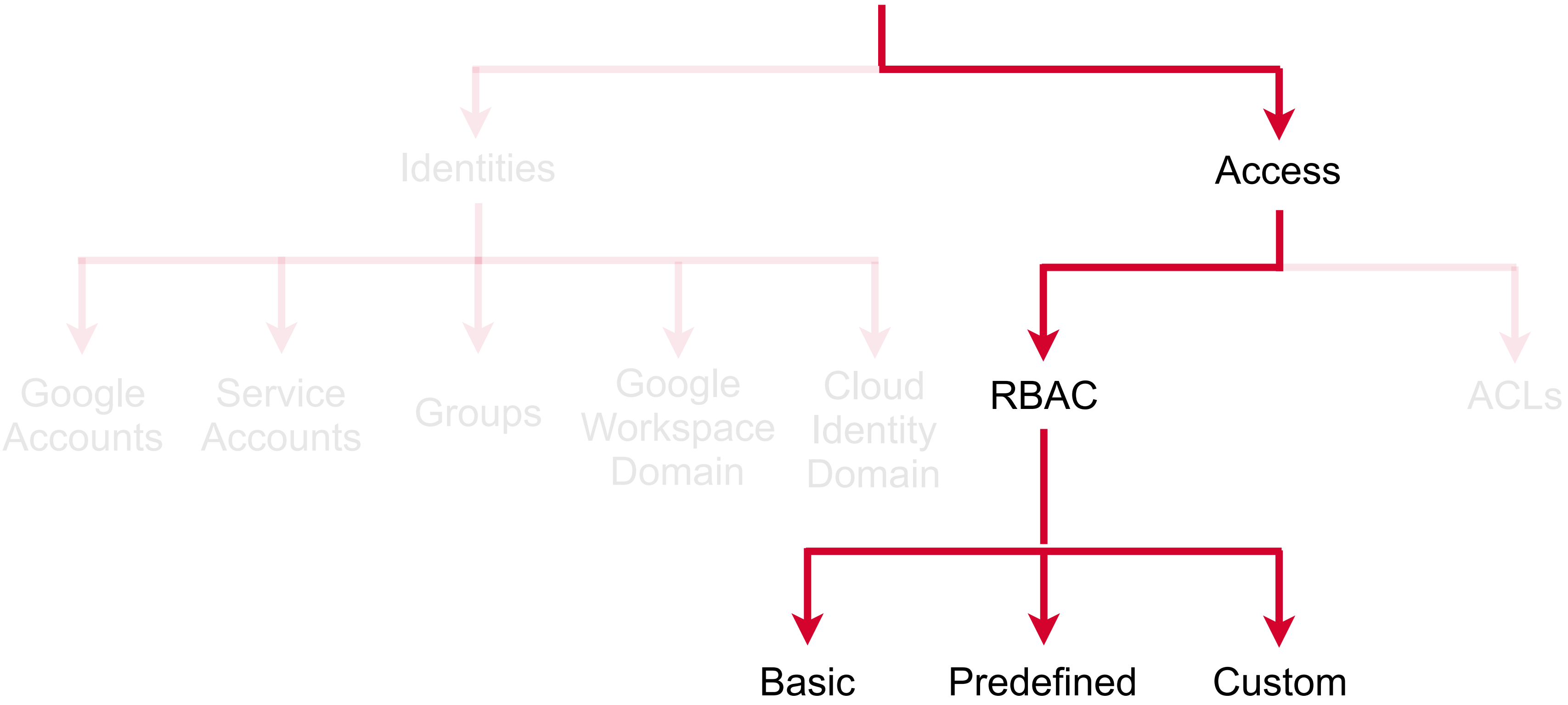
Identity and Access Management (IAM)



ACLs Not Part of the IAM Service on Google



Identity and Access Management (IAM)



Basic Roles

Three concentric roles that existed prior to the introduction of Cloud IAM: Owner, Editor, and Viewer of any resource.

Historically available, not recommended unless there is not alternative.

Predefined Roles

- Project Roles
- App Engine Roles
- BigQuery Roles
- Cloud Bigtable Roles
- Cloud Billing Roles

Predefined Roles

roles/bigquery.dataViewer

bigquery.datasets.get
bigquery.datasets.getIamPolicy
bigquery.models.getData
bigquery.models.getMetadata
bigquery.models.list
bigquery.routines.get
bigquery.routines.list
bigquery.tables.export
bigquery.tables.get
bigquery.tables.getData
bigquery.tables.list
resourcemanager.projects.get
resourcemanager.projects.list

Custom Roles

User-defined roles that bundle one or more supported permissions tailored to meet your specific needs.

Not maintained by Google; when new permissions, features, or services are added to GCP, your custom roles will not be updated automatically.

O'REILLY

Identity Aware Proxy

Identity Aware Proxy (IAP)

A central authorization layer for applications accessed by HTTPS, so you can use an application-level access control model instead of relying on network-level firewalls.

Define access policies centrally and apply them to all of your applications and resources.

Can set up individual or group-based access to applications

IAM and IAP (Identity Aware Proxy)

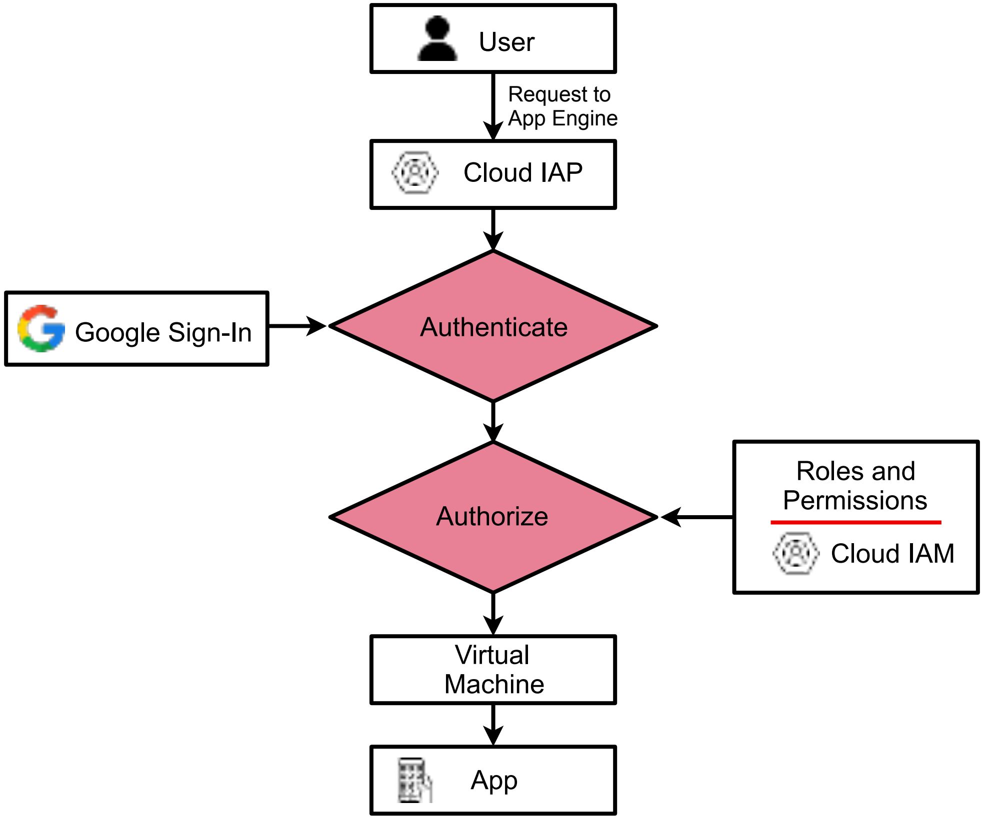
IAM

- Access controls and permissions for Google Cloud resources
- Configured at the resource level (VMs, buckets, datasets)

IAP

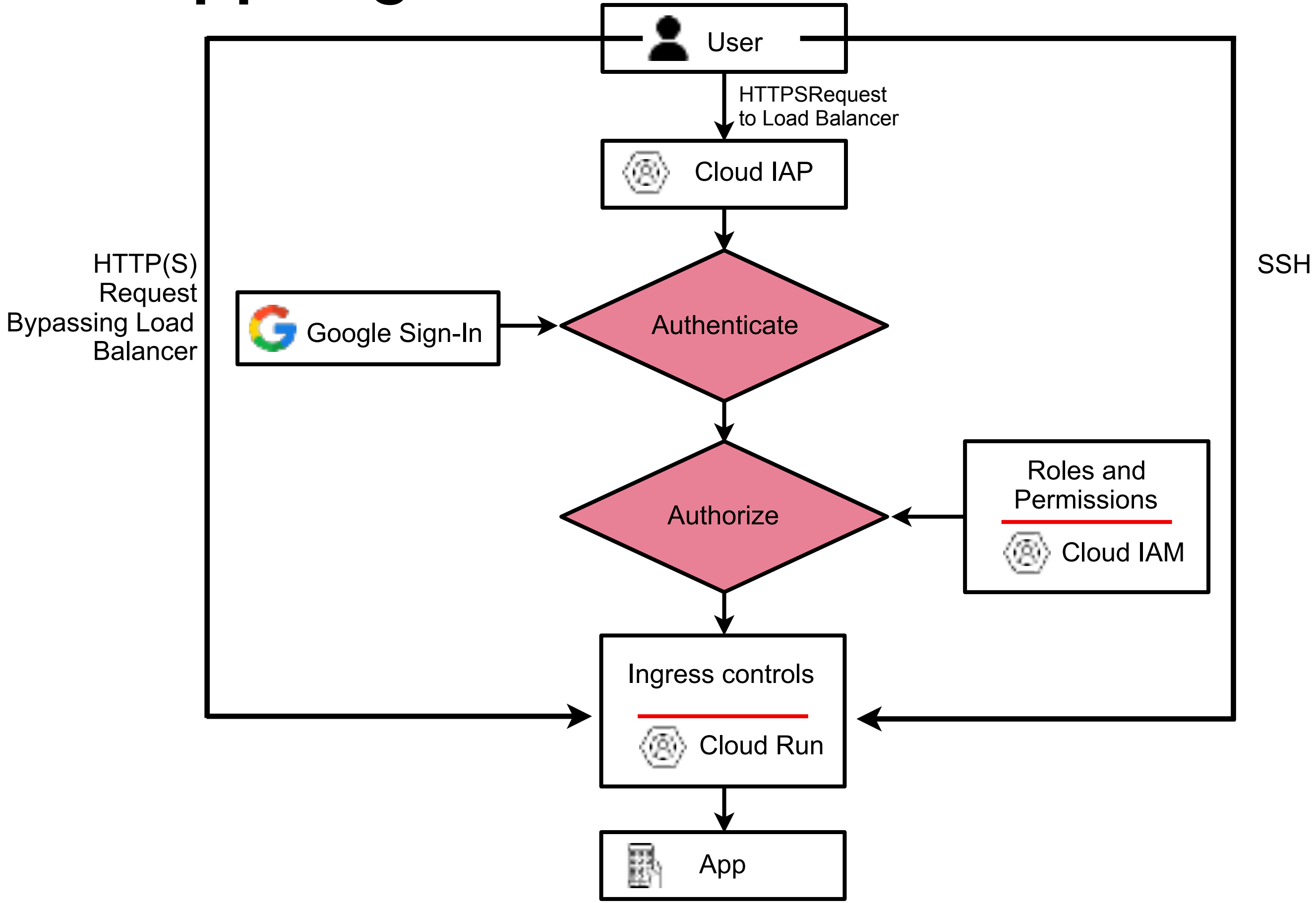
- Security layer that controls access to applications running on Google Cloud
- Configured to protect applications by intercepting requests to them
- Uses identities and roles from IAM to grant access to applications

IAP with App Engine



Can work with Cloud Run, Compute Engine,
GKE, and even On-premise apps

IAP with App Engine



IAP secures authentication and authorization of all requests to App Engine, Cloud Load Balancing (HTTPS), or internal HTTP load balancing.

IAP doesn't protect against activity within a project, such as another VM inside the project.

IAM and IAP

A developer writes an application that invokes various GCP services. Following best practices the application should get its permissions from:

1. The project editor
2. The project owner
3. The developer's identity
4. A service account

IAM and IAP

A developer writes an application that invokes various GCP services. Following best practices the application should get its permissions from:

1. The project editor
2. The project owner
3. The developer's identity
4. **A service account**

IAM and IAP

When new permissions are created, the following entity will not automatically be updated with any additional appropriate permissions:

1. Custom roles
2. Primitive roles
3. Project owner
4. Predefined roles

IAM and IAP

When new permissions are created, the following entity will not automatically be updated with any additional appropriate permissions:

1. **Custom roles**
2. Primitive roles
3. Project owner
4. Predefined roles

IAM and IAP

In order to protect your applications running on a Compute Engine virtual machine what would you use?

1. Access Control Lists (ACLs)
2. Identity and Access Management (IAM)
3. Firewall Rules
4. Identity Aware Proxy (IAP)

IAM and IAP

In order to protect your applications running on a Compute Engine virtual machine what would you use?

1. Access Control Lists (ACLs)
2. Identity and Access Management (IAM)
3. Firewall Rules
4. **Identity Aware Proxy (IAP)**

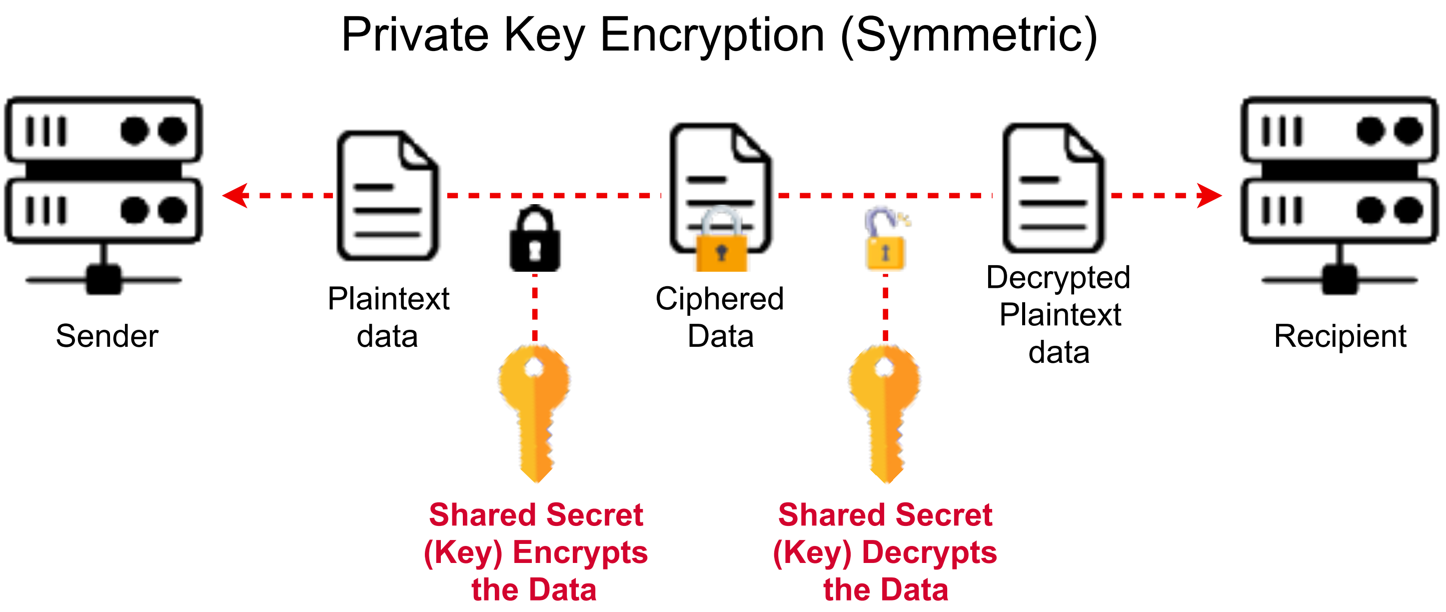
Key Management

Cryptographic Keys

Cryptographic keys serve as the secret codes that enable the encryption and decryption of data.

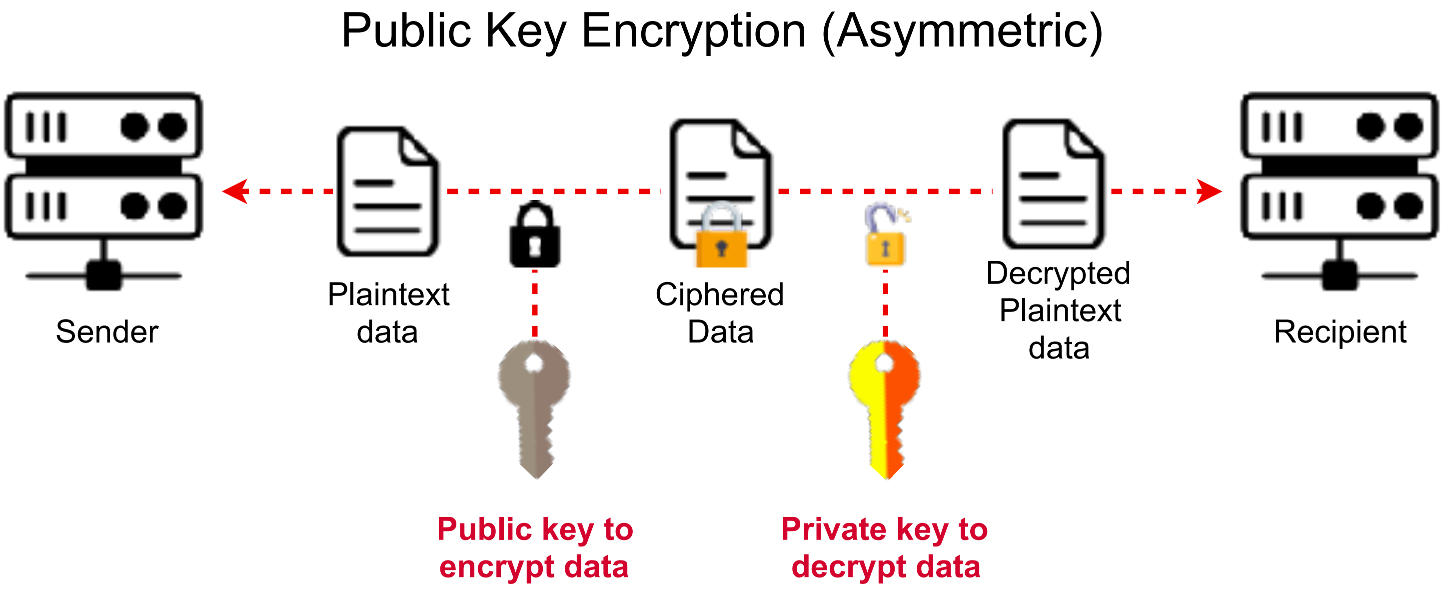
Ensure that only authorized parties with the correct key can access and read the encrypted information

Symmetric Key Encryption



Encrypting the data and decrypting the data make use of the same shared key

Asymmetric Key Encryption



The encryption key is publicly available - the decryption key is private

Default Encryption on the Google Cloud

All Google Cloud services that store data encrypt data by default

No configuration and automatic encryption. Most services automatically rotate keys

Google-owned and Google-managed keys

Customer Managed Encryption Keys (CMEK)

Encryption keys that customers create, own, and manage within cloud services to secure their data

CMEKs give customers greater control over their encryption practices, including key rotation and access policies

Cloud Key Management Service (CMEKs)

Software-based keys

Hardware-based keys

External keys

Cloud Key Management Service (CMEKs)

Cloud KMS (Key Management Service)

Software-based keys

Cloud HSM (Hardware Security Module)

Hardware-based keys

Cloud EKM (External Key Manager)

External keys

Cloud Key Management Service (CMEKs)

Cloud KMS (Key Management Service)

Software-based keys

Cloud HSM (Hardware Security Module)

Hardware-based keys

Cloud EKM (External Key Manager)

External keys

Control keys, key rotation schedule, IAM roles and permissions

Cloud Key Management Service (CMEKs)

Cloud KMS (Key Management Service)

Software-based keys

Control keys, key rotation schedule, IAM roles and permissions

Cloud HSM (Hardware Security Module)

Hardware-based keys

More secure that software keys, stored in a separate physical device

Control keys, key rotation schedule, IAM roles and permissions

Cloud EKM (External Key Manager)

External keys

Cloud Key Management Service (CMEKs)

Cloud KMS (Key Management Service)

Software-based keys

Control keys, key rotation schedule, IAM roles and permissions

Cloud HSM (Hardware Security Module)

Hardware-based keys

More secure that software keys, stored in a separate physical device

Control keys, key rotation schedule, IAM roles and permissions

Cloud EKM (External Key Manager)

External keys

Keys stored outside Google in an external provider, keys never sent to Google

Control keys, key rotation schedule, IAM roles and permissions

Customer Supplied Encryption Keys (CSEK)

Customers provide key materials when needed.

Google keeps keys in-memory, keys not stored permanently on Google's servers

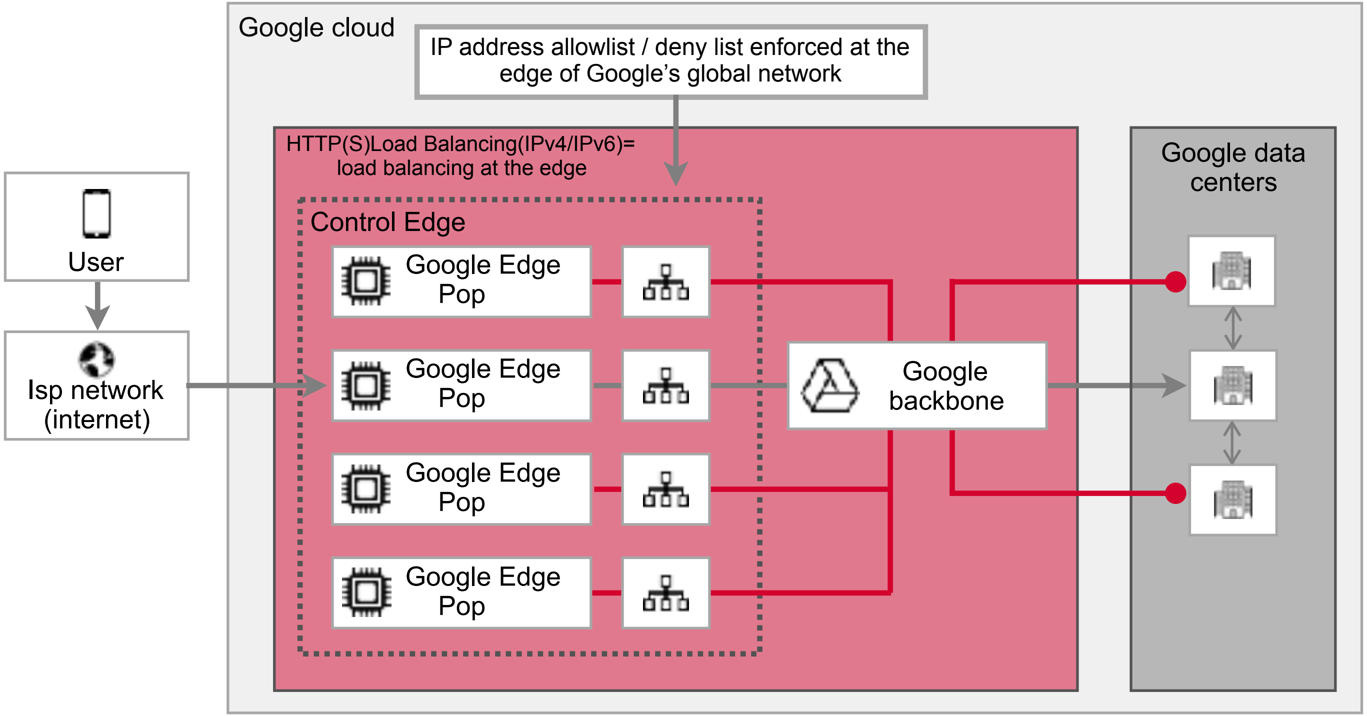
O'REILLY

Cloud Armor and Data Loss Prevention

Cloud Armor

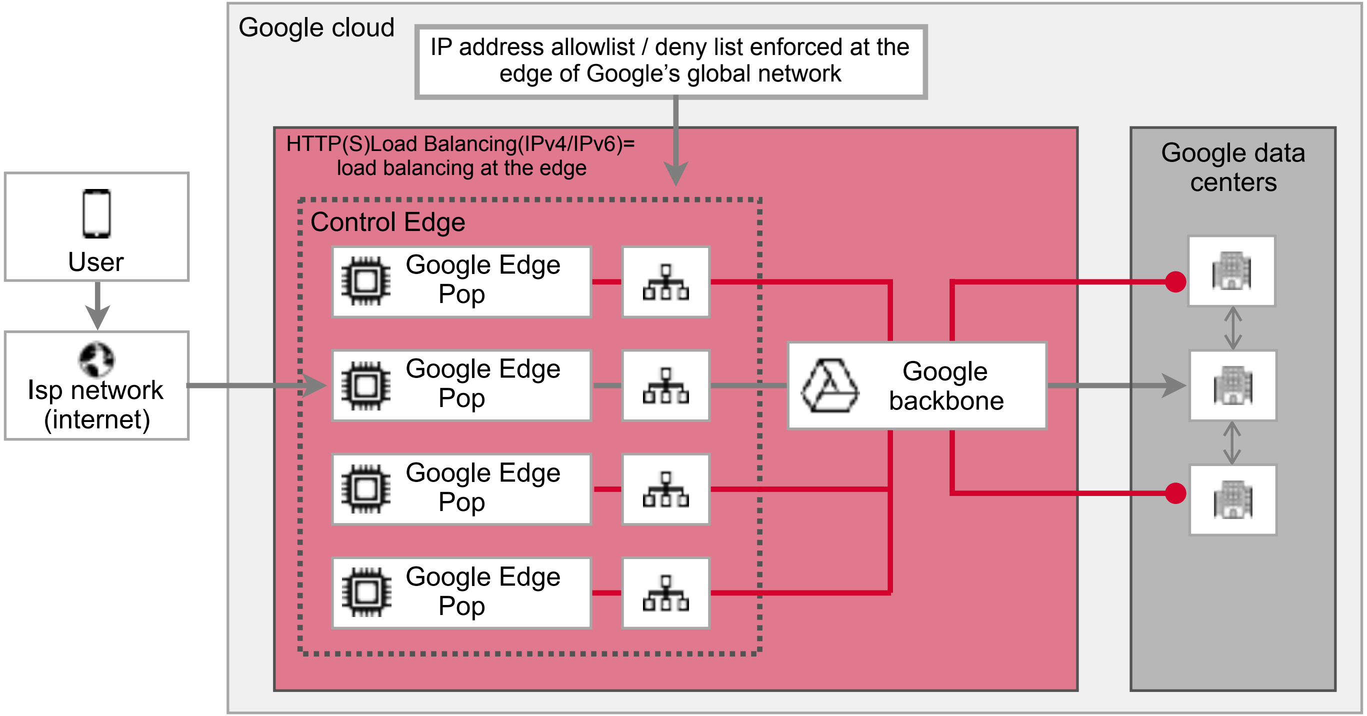
Helps protect your Google Cloud deployments from multiple types of threats, including distributed denial-of-service (DDoS) attacks, cross-site scripting (XSS), and SQL injection (SQLi).

How Does Cloud Armor Work?



Protection against volumetric DDoS attacks. Protection for applications and services running behind a load balancer

How Does Cloud Armor Work?



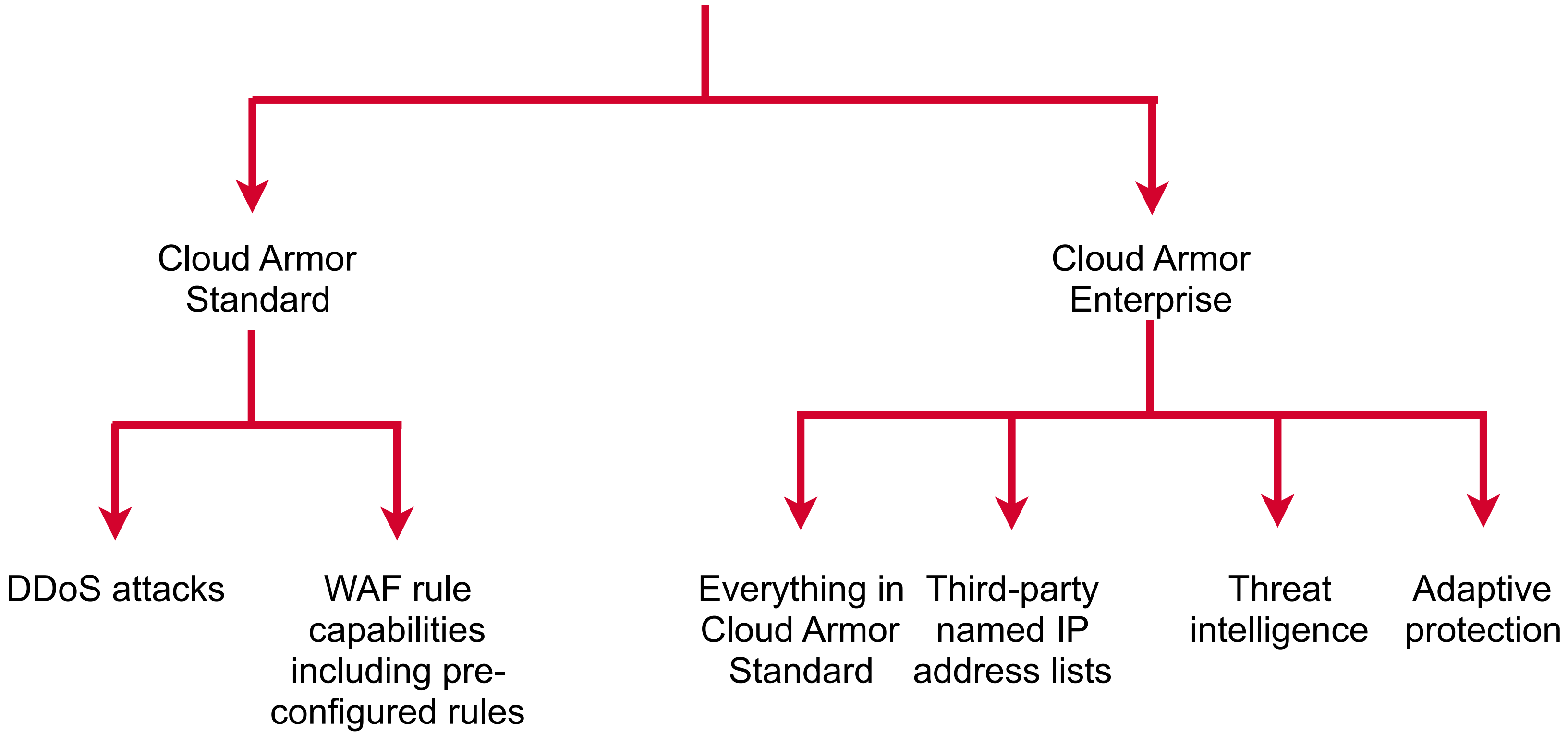
Security policies enforce custom Layer 7 filtering policies including preconfigured web application firewall (WAF) rules to mitigate OWASP top 10 web application vulnerability risks

Cloud Armor Products

Cloud Armor Standard

Cloud Armor Enterprise

Cloud Armor Products



Cloud Armor Enterprise

Third-party named IP address lists

Can use third party lists of malicious IP address - don't have to set up and configure the IP addresses yourself

Threat intelligence

Use Google's continuously updated data about known threats e.g. malicious activity and malware distribution points with preconfigured rules

Adaptive protection

Builds machine learning models to detect and alert anomalous activity, generate a signature for the attack and generate a custom WAF rule to block the signature

Sensitive Data Protection

A fully managed service designed to help you discover, classify, and protect your valuable data assets.

The Cloud Data Loss Prevention APIs are now part of this family of managed services. Provides API access to all the services for sensitive data protection.

Sensitive Data Protection

Sensitive data discovery

Storage inspection

Hybrid inspection

Content inspection

Content de-identification

Sensitive Data Protection

Sensitive data discovery

Storage inspection

Hybrid inspection

Content inspection

Content de-identification

Scan for sensitive data stored in your databases and data warehouses. Use scan configurations to specify what data you are looking for. Constructs data profiles that help you discover sensitive data

Sensitive Data Protection

Sensitive data discovery

Storage inspection

Hybrid inspection

Content inspection

Content de-identification

Scan for and find data **stored in Google Cloud Storage** in unstructured formats e.g chat logs.

Sensitive Data Protection

Sensitive data discovery

Storage inspection

Hybrid inspection

Content inspection

Content de-identification

Scan for and find data stored **outside of Google Cloud** in unstructured formats e.g chat logs.

Sensitive Data Protection

Sensitive data discovery

Storage inspection

Hybrid inspection

Content inspection

Content de-identification

Perform data inspection in near real time - used to integrate into custom workloads, applications, or pipelines

Sensitive Data Protection

Sensitive data discovery

Storage inspection

Hybrid inspection

Content inspection

Content de-identification

Masking, tokenizing, or de-identifying sensitive data in near real time - used to integrate into custom workloads, applications, or pipelines

BeyondCorp

Zero-trust Model

The zero-trust model is a security framework based on the principle that no entity, whether inside or outside the network, should be trusted by default.

Zero-trust Model

The zero-trust model is a security framework based on the principle that no entity, whether inside or outside the network, should be trusted by default.

Every access request verified before being granted access to resources

Traditional models are **perimeter-based** models assume that everything within an organisation's network can be trusted

Pillars of Zero-trust Security

Identity

Device

Network

Application

Data

Pillars of Zero-trust Security

Identity

Device

Network

Application

Data

Strong identity verification mechanisms, such as multi-factor authentication (MFA), to ensure that users are who they claim to be

Pillars of Zero-trust Security

Identity

Device

Network

Application

Data

Ensuring that devices accessing the network are secure and meet predefined security standards. This includes managing device health and compliance

Pillars of Zero-trust Security

Identity

Device

Network

Application

Data

Micro-segmentation and least privilege access to reduce the risk of lateral movement within the network. Network traffic is monitored and analyzed continuously.

Pillars of Zero-trust Security

Identity

Device

Network

Application

Data

Ensuring that applications are secure and can only be accessed by authenticated and authorized users and devices.

Pillars of Zero-trust Security

Identity

Device

Network

Application

Data

Protecting sensitive data through encryption and strict access controls.
Ensuring data is accessible only to authorized users and devices.

BeyondCorp

- A security architecture that focuses on zero trust principles.
- Assumes no implicit trust and verifies each access request individually.
- Employs identity and context-aware access control

Chrome Enterprise Protection

- A suite of security features designed to protect enterprise users and data.
- Offers threat and data protection, rich access controls, and security insights.

In essence, Chrome Enterprise Protection strengthens the BeyondCorp framework by providing specific tools and features to protect users, devices, and data

Security

If you want to programmatically ensure that you de-identify the data that you use to train your AI and ML models what service would you use?

1. Storage inspection
2. Data loss prevention APIs
3. Key management service
4. BeyondCorp

Security

If you want to programmatically ensure that you de-identify the data that you use to train your AI and ML models what service would you use?

1. Storage inspection
2. **Data loss prevention APIs**
3. Key management service
4. BeyondCorp

Security

What is Google's implementation of the zero-trust architecture called?

1. Kubernetes
2. Identity Aware Proxy
3. Cloud HSM
4. BeyondCorp

Security

What is Google’s implementation of the zero-trust architecture called?

1. Kubernetes
2. Identity Aware Proxy
3. Cloud HSM
4. **BeyondCorp**

Containers and Kubernetes

Google Cloud Compute Choices

IaaS

PaaS

Google Compute Engine

Google Kubernetes Engine

Google App Engine

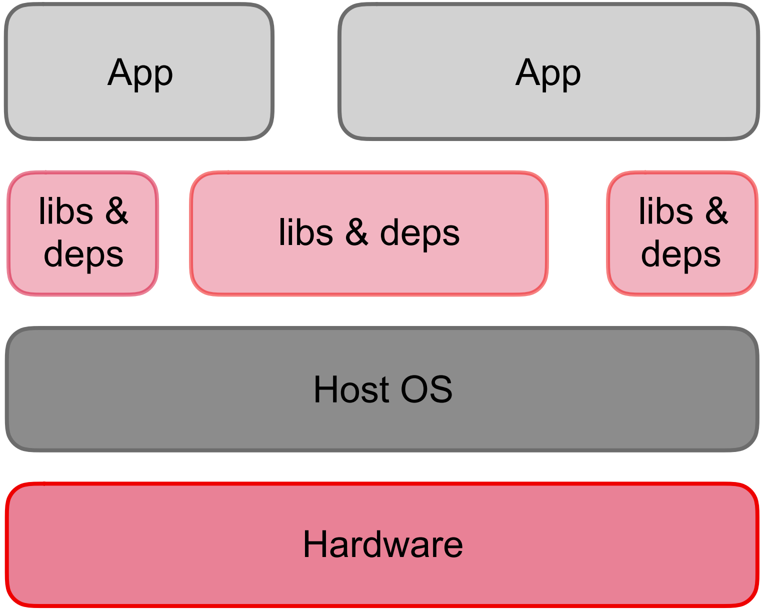
Google Cloud Run

Google Cloud Functions

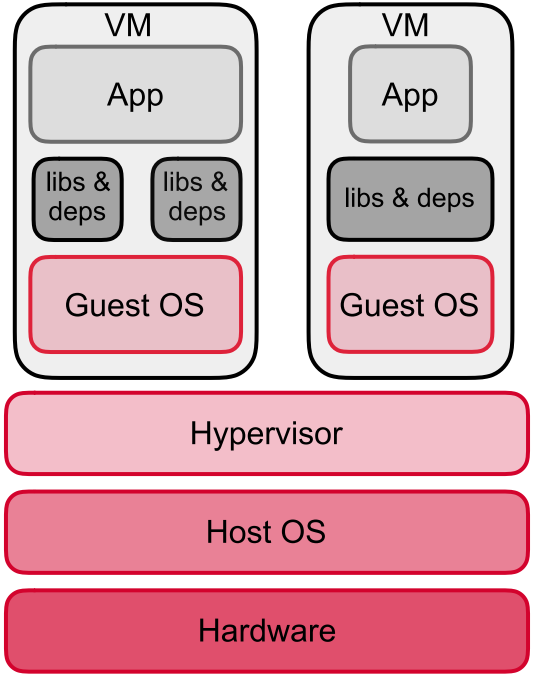
IaaS

PaaS

Traditional Compute on Bare Metal



Modern Workloads on VMs



Drawbacks of VMs

- Contain guest OS
  - Introduces platform dependency
  - Bloats image size to GB (apps far smaller)
- Heavyweight
  - Slow to boot up
  - Slow to scale
- Not trivial to migrate
  - VM migration tools needed

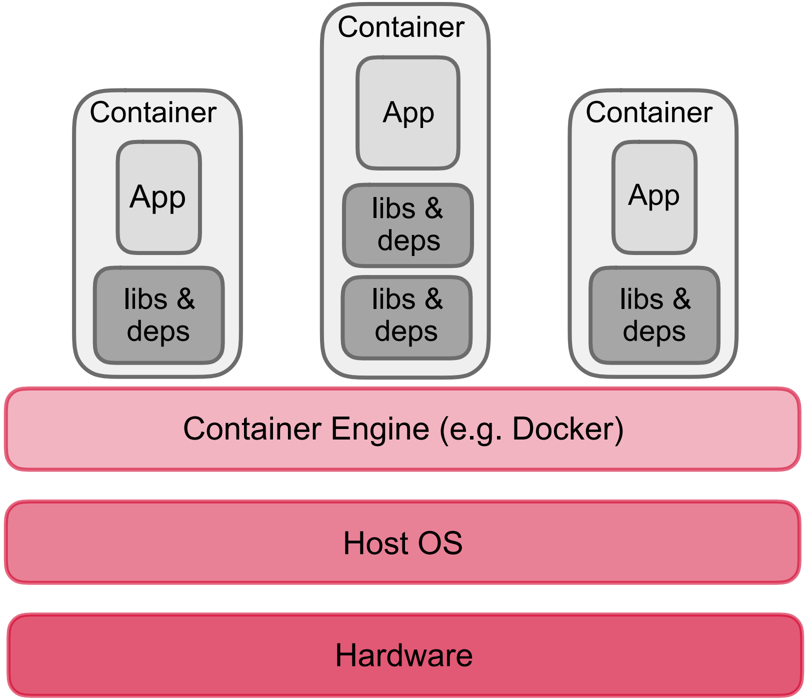
Container

A container image is a lightweight, stand-alone, executable package of a piece of software that includes everything needed to run it; code, runtime, system tools, system libraries, settings

Container

- Contains applications
- And all of the application’s dependencies
- Platform independent
- Runs on layer of abstraction
- Docker Runtime (for Docker containers)

Modern Workloads on Containers



Attractions of Containers

- No guest OS
  - Platform independent
  - Considerably smaller than VM images
- Lightweight
  - Small and fast
  - Quick to start
  - Speeds up autoscaling
- Hybrid, multi-cloud
  - Hybrid: Work on-premise and on cloud
  - Multi-cloud: Not tied to any specific cloud platform

Standalone Container Limitations

- No autohealing
  - Crashed containers won’t restart automatically
  - Need higher level orchestration
- No scaling or autoscaling
  - Overloaded containers don’t spawn more automatically
  - Need higher level orchestration
- No load balancing
  - Containers can’t share load automatically
  - Need higher level orchestration
- No isolation
  - Crashing containers can take each other down
  - Need sandbox to separate them

Kubernetes

Orchestration technology for containers - convert isolated containers running on different hardware into a cluster

Kubernetes is fast emerging as middle-ground between IaaS and PaaS in a hybrid, multi-cloud world

IaaS vs. PaaS

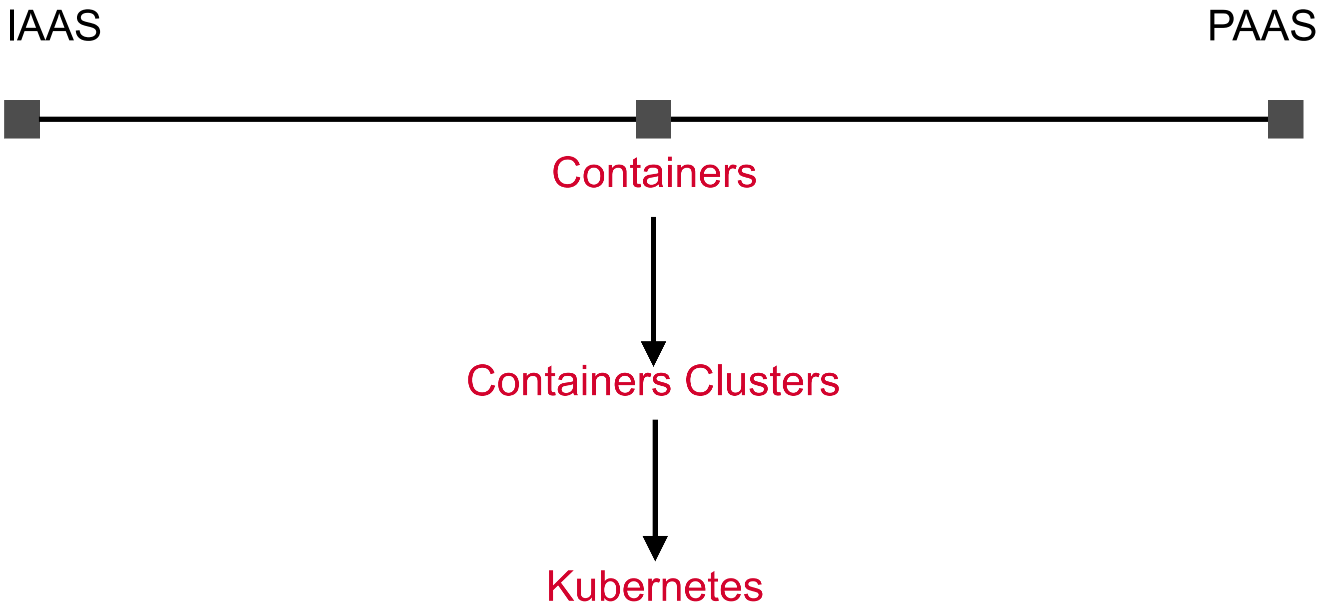
Infrastructure-as-a-Service

- Heavy operational burden
- Migration is hard

Platform-as-a-Service

- Provider lock-in
- Migration is very hard

Compute Choices



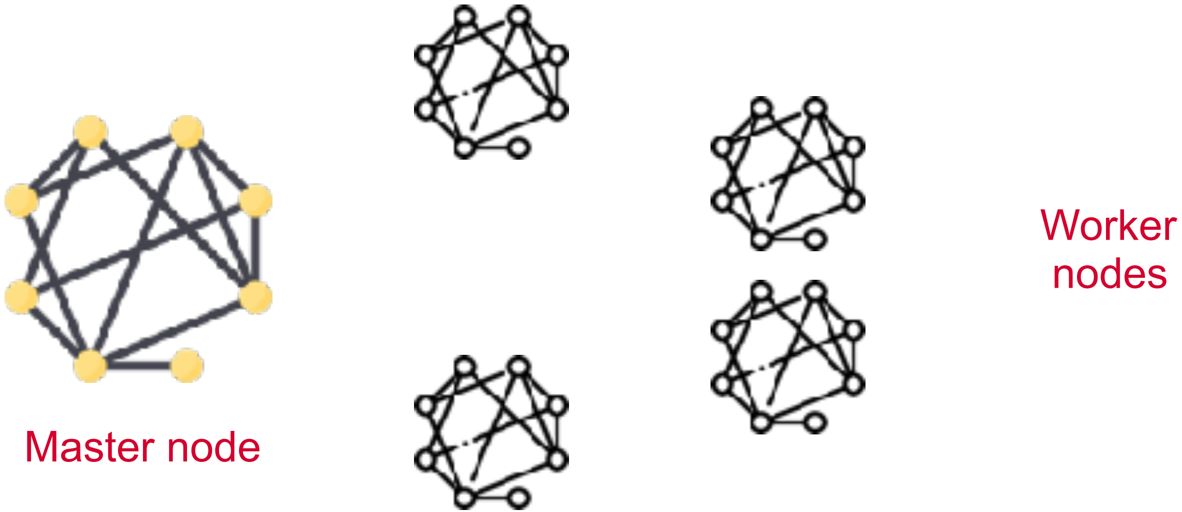
Kubernetes as Orchestrator

- Fault-tolerance
- Autohealing
- Isolation
- Scaling
- Autoscaling
- Load balancing

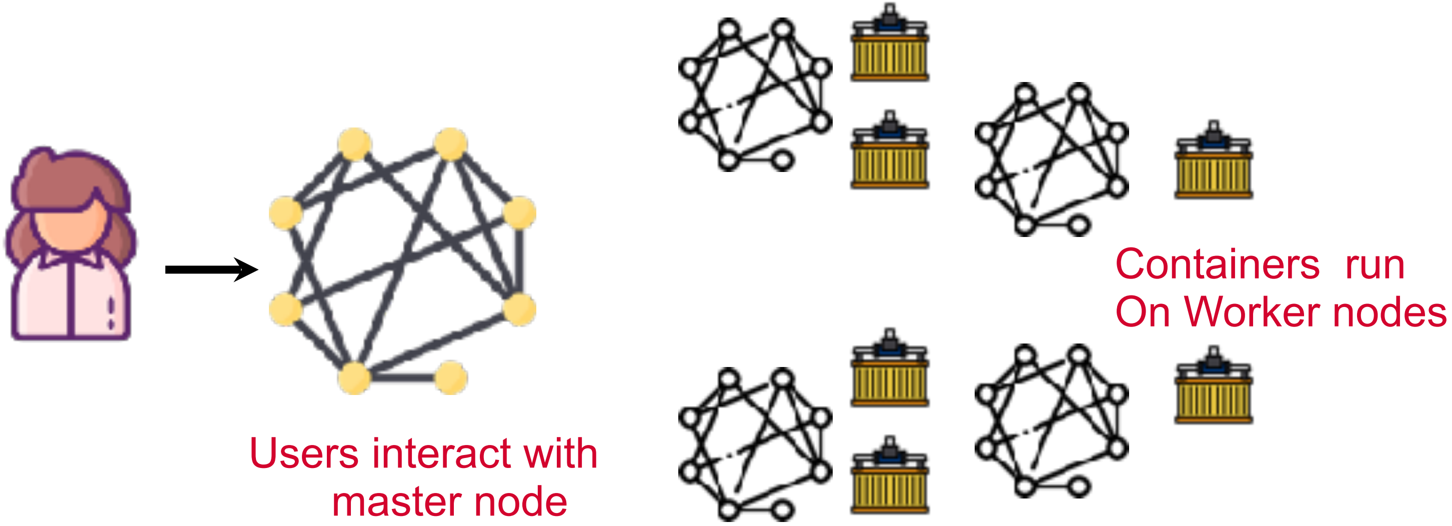
Google Kubernetes Engine (GKE)

- Service for working with Kubernetes clusters on GCP
- Runs Kubernetes on GCE VM instances
- Many more abstractions and a lot more support than using plain Kubernetes on-premises

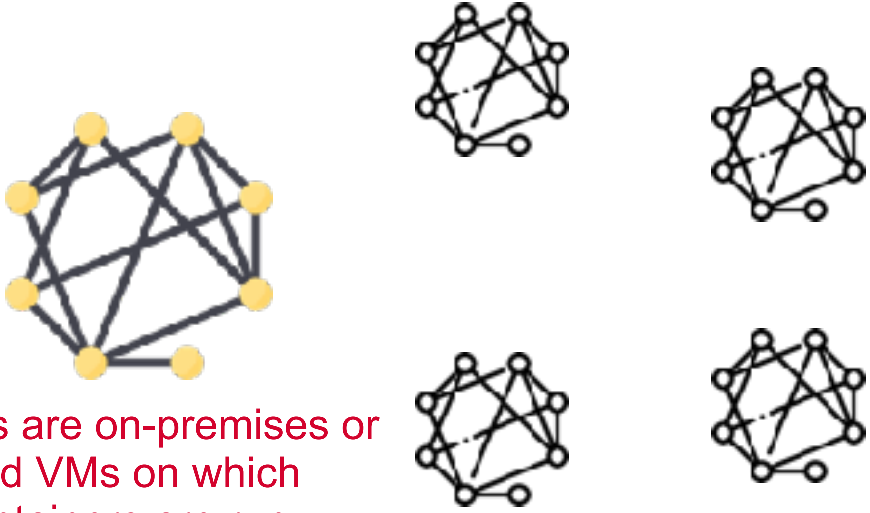
Kubernetes Clusters



Kubernetes Clusters

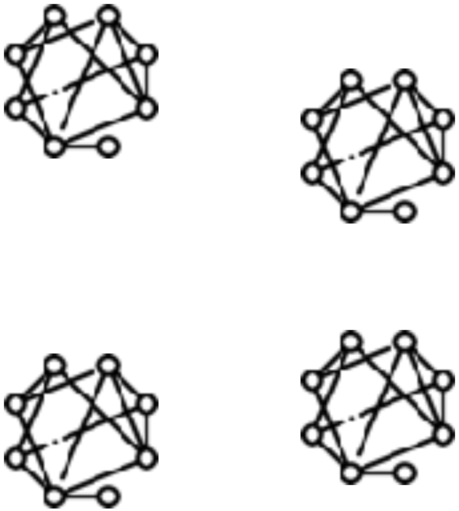


Nodes



Nodes are on-premises or cloud VMs on which containers are run

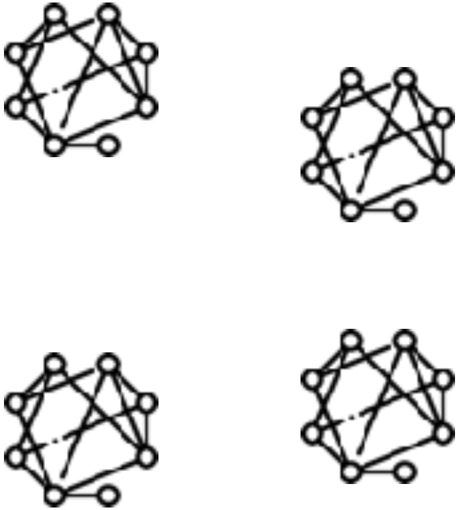
Nodes



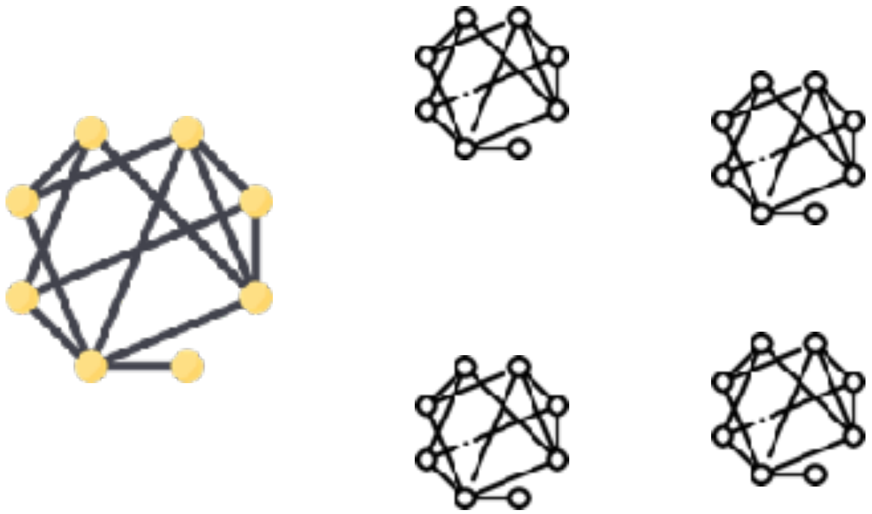
Run the services needed to host Docker containers - communicate with the master

Node Images

Special operating system images are available on the Google Cloud to run on Kubernetes nodes

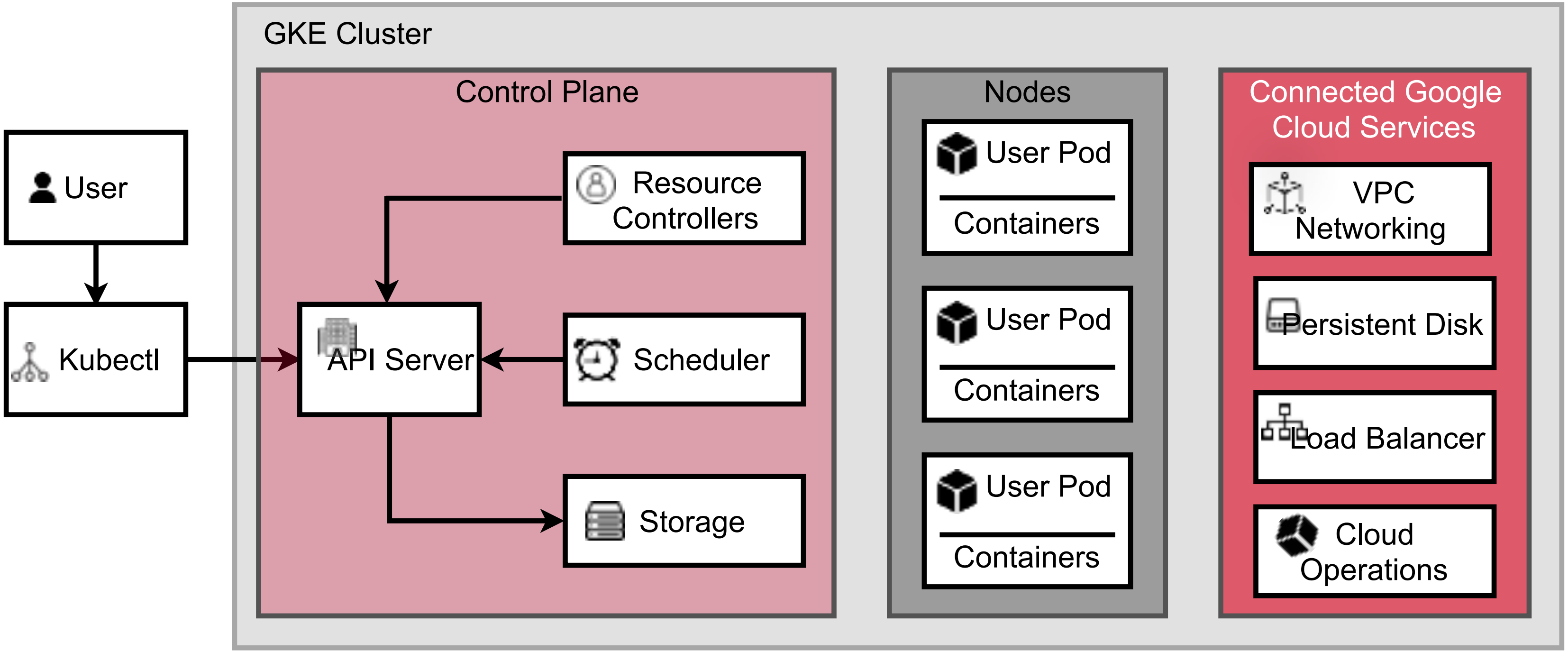


Nodes Pools



Nodes in your cluster that have the same configuration settings

GKE Cluster


User
Kubectl
GKE Cluster
Control Plane
Resource Controllers
API Server
Scheduler
Storage
Nodes
User Pod
Containers
User Pod
Containers
User Pod
Containers
Connected Google Cloud Services
VPC Networking
Persistent Disk
Load Balancer
Cloud Operations

Benefits of GKE

- Use GCP's load balancing for VMs
- Automatic scaling of nodes in cluster
- Automatic upgrades for software on nodes
- Node auto-repair for node health and availability
- Logging and monitoring using GCP's cloud monitoring

GKE Mode of Operation

Autopilot Mode	Standard Mode

Autopilot Mode

- More managed GKE experience
- GKE manages the underlying infrastructure
- Node configuration, autoscaling, auto-upgrades, baseline security and networking configurations
- Implements best practices for security, scalability, and cost optimization by default

Autopilot Mode

- **Cost effective**: Only pay for compute resources that your workloads use while running
- **Automation**: Google creates and manages nodes, scales nodes and workloads based on traffic
- **Security**: Enables many security settings and automatically applies security patches

Standard Mode

- Complete control over all GKE configuration settings
- Manage configurations for node pools, security, scheduling, scaling, resource management, version management and software upgrades

Use Standard Mode If:

- You want granular control over your configuration settings
- You want to install or modify software running on the nodes themselves i.e. change node OS
- Use certain features that are only available in the Standard Mode (GKE Sandbox, Cloud TPU)
- Test alpha features in open source Kubernetes


O'REILLY
Anthos (GKE Enterprise)

Hybrid and Multicloud Environments

- **Workloads on-premises**
  - Data sovereignty and compliance
  - Low latency and performance needs
  - Already existing investment in infra
- **Workloads on another cloud**
  - Mitigating vendor lock-in
  - Building resilience

Multi-cluster management

- Organizations might deploy multiple clusters to meet technical and business needs
  - Separate production and non-production environments
  - Adhere to regulatory requirements
  - Organize services by tiers, locations, or teams
- Multiple clusters introduce challenges in configuration, security, and management

Anthos

Anthos is a modern application management platform that enables organizations to run applications across on-premises, multi-cloud, and hybrid cloud environments.

Anthos

Anthos is a modern application management platform that enables organizations to run applications across on-premises, multi-cloud, and hybrid cloud environments.

Allows you to manage multiple Kubernetes clusters for enterprise workloads at scale

GKE Enterprise

Advanced version of Google Kubernetes Engine designed to meet the needs of large organizations with complex, large-scale Kubernetes deployments

Makes it easier to implement hybrid and multicloud strategies

Anthos (GKE Enterprise) Fleets

- A way to logically group and normalize Kubernetes resources
    - Manage groups of clusters rather than individual clusters
- Resources in a fleet generally related to one another
    - Resources with large cross-service communication benefit from being part of the same fleet
-

Benefits of Fleets

- Unified management of clusters
- Consistent operations across clusters
- Enhanced visibility over the entire system

Kubernetes

Which of the following statements regarding standalone containers are true?

1. They can automatically heal themselves
2. They can spawn new container to handle additional load
3. Higher level abstractions are needed for container clusters
4. Containers only contain your application code

Kubernetes

Which of the following statements regarding standalone containers are true?

1. They can automatically heal themselves
2. They can spawn new container to handle additional load
3. **Higher level abstractions are needed for container clusters**
4. Containers only contain your application code

Kubernetes

How are multiple clusters managed using GKE Enterprise?

1. Use the node pool abstraction
2. Using the fleet abstraction
3. Using the pod abstraction
4. Using the Deployment abstraction

Kubernetes

How are multiple clusters managed using GKE Enterprise?

1. Use the node pool abstraction
2. **Using the fleet abstraction**
3. Using the pod abstraction
4. Using the Deployment abstraction

Instance Groups

Instance Groups

A collection of virtual machines that you can manage as a single entity

Instance Groups

Managed Instance Groups

Unmanaged Instance Groups

Instance Groups

Managed Instance Groups

Unmanaged Instance Groups

Managed Instance Group

Group of identical GCE VM instances, created from the same instance template that are managed by the platform

Managed Instance Group

Group of identical GCE VM instances, created from the same instance template that are managed by the platform

Instances have the exact same configuration

Managed Instance Group

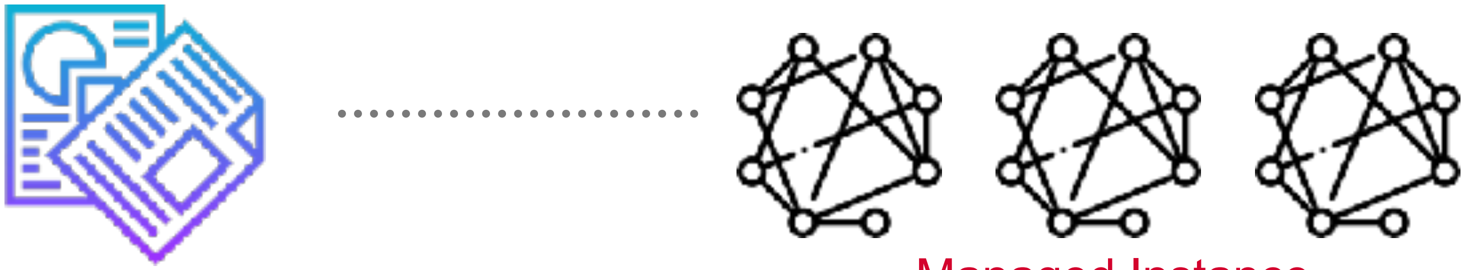
Group of identical GCE VM instances, created from the same instance template that are managed by the platform

The configuration is specified in an instance template

Instance Template

A specification of machine type, boot disk (or container image), zone, labels and other instance properties that can be used to instantiate either individual VM instances or a Managed Instance Group

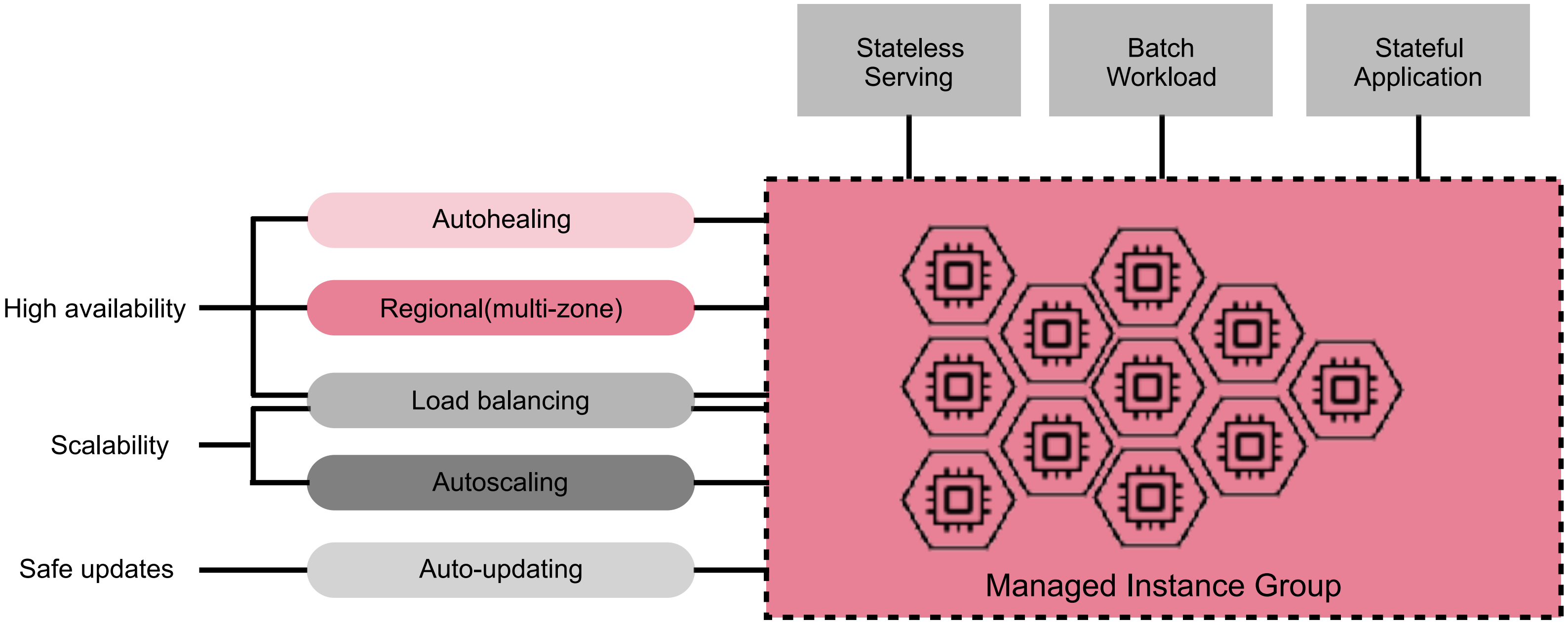
Instance Template to Create Instances



Instance Template

Managed Instance Group

Managed Instance Groups



Instance Groups

Managed Instance Groups

Unmanaged Instance Groups

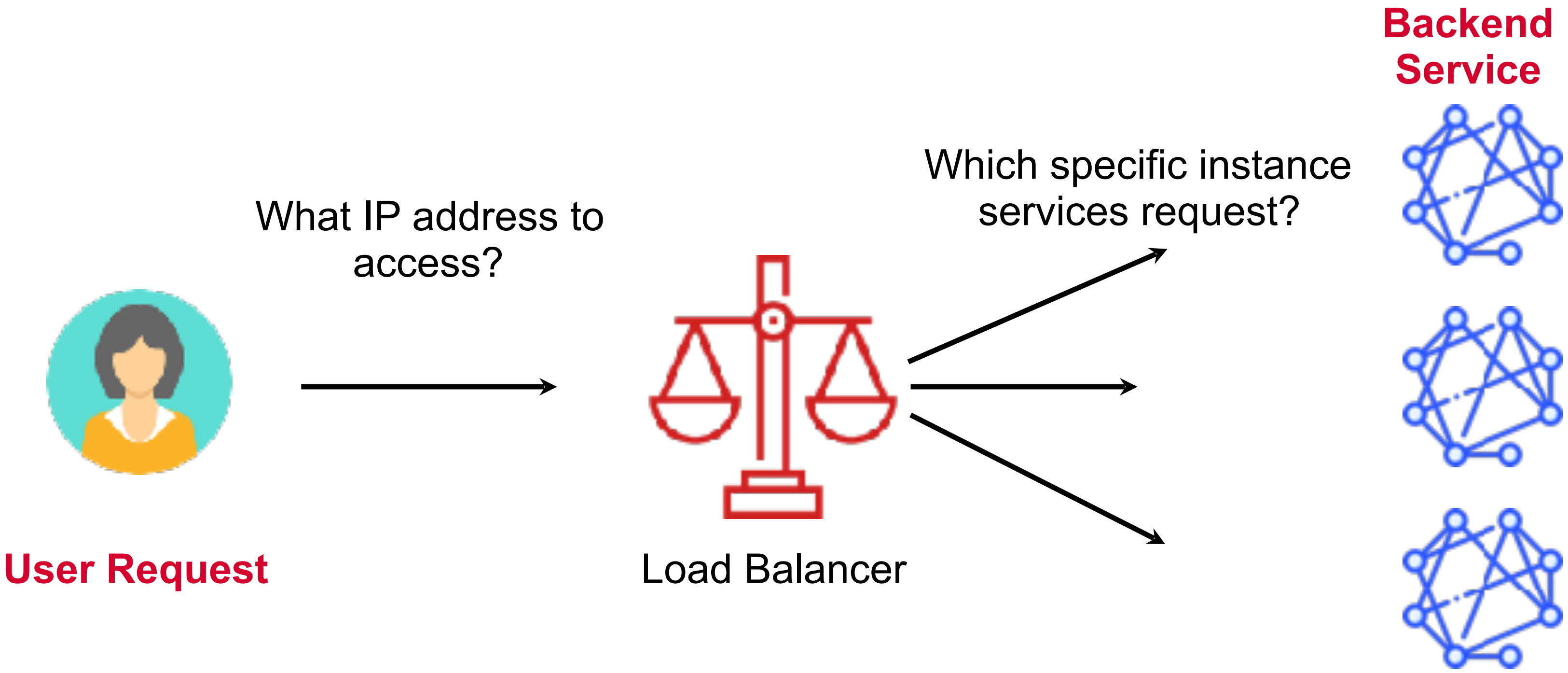
Unmanaged instance groups can contain heterogeneous instances that you can arbitrarily add and remove from the group.

Do not offer autoscaling, auto healing - can be used with a load balancer

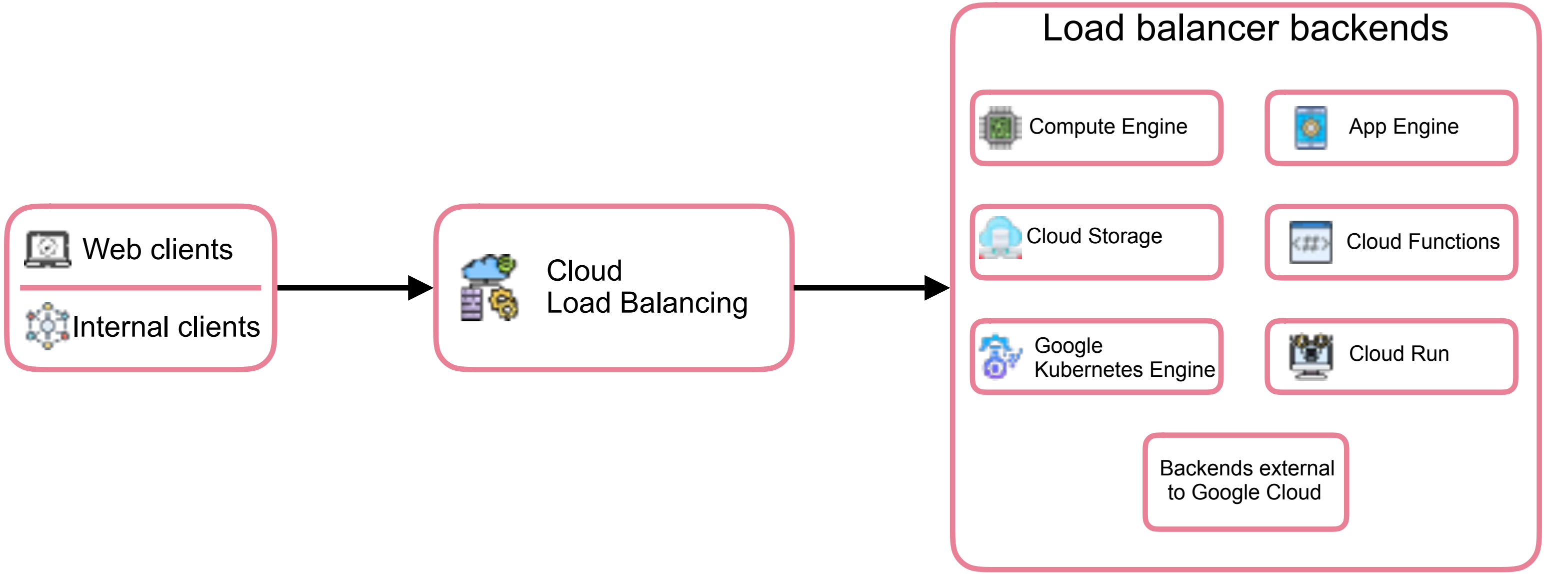
O'REILLY

Load Balancing

Load Balancers



Load Balancers Used with Multiple Backends



Load Balancers

- Complex service
- Many moving parts
- Basic idea
  - Stable front-end IP
  - Forwarding rules to funnel traffic
  - Connect to backend service
  - Distribute load intelligently
  - Health checks to avoid unhealthy instances

Load balancers **distribute** traffic to resources close to users and meet **high-availability** requirements

Load Balancers on the GCP

- Fully managed, software-defined, redundant and highly available
- Supports > 1 million queries per second with high performance and low latency
- Autoscaling to meet increased traffic

Load Balancing Categorization

Global

Use when your users and instances are globally distributed, load balanced resources lie across multiple regions

Regional

Use when instances and users are concentrated in one region

Load Balancing Categorization

External

Distributes traffic from the internet to the Google Cloud

Internal

Distributes traffic only within the Google Cloud, all clients are inside of the Google Cloud

7 Layer OSI Network Stack

Routing decisions based on attributes of the request i.e. HTTP headers and the URL

User
Application Layer
Presentation Layer
Session Layer
Transport Layer
Network Layer
Data Link Layer
Physical Layer

Direct traffic based on data from network and transport layer protocols such as TCP, UDP, ESP, GRE, ICMP, and ICMPv6

Two Types of Load Balancers

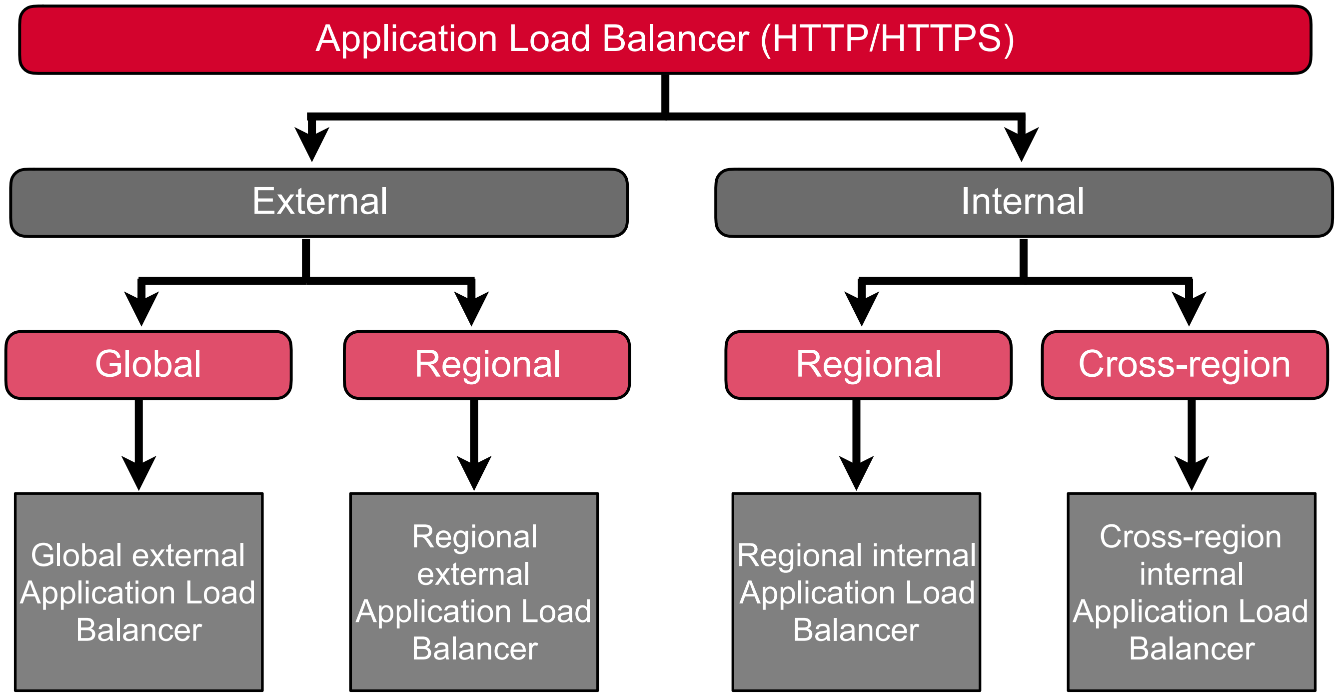
Application Load Balancers

Network Load Balancers

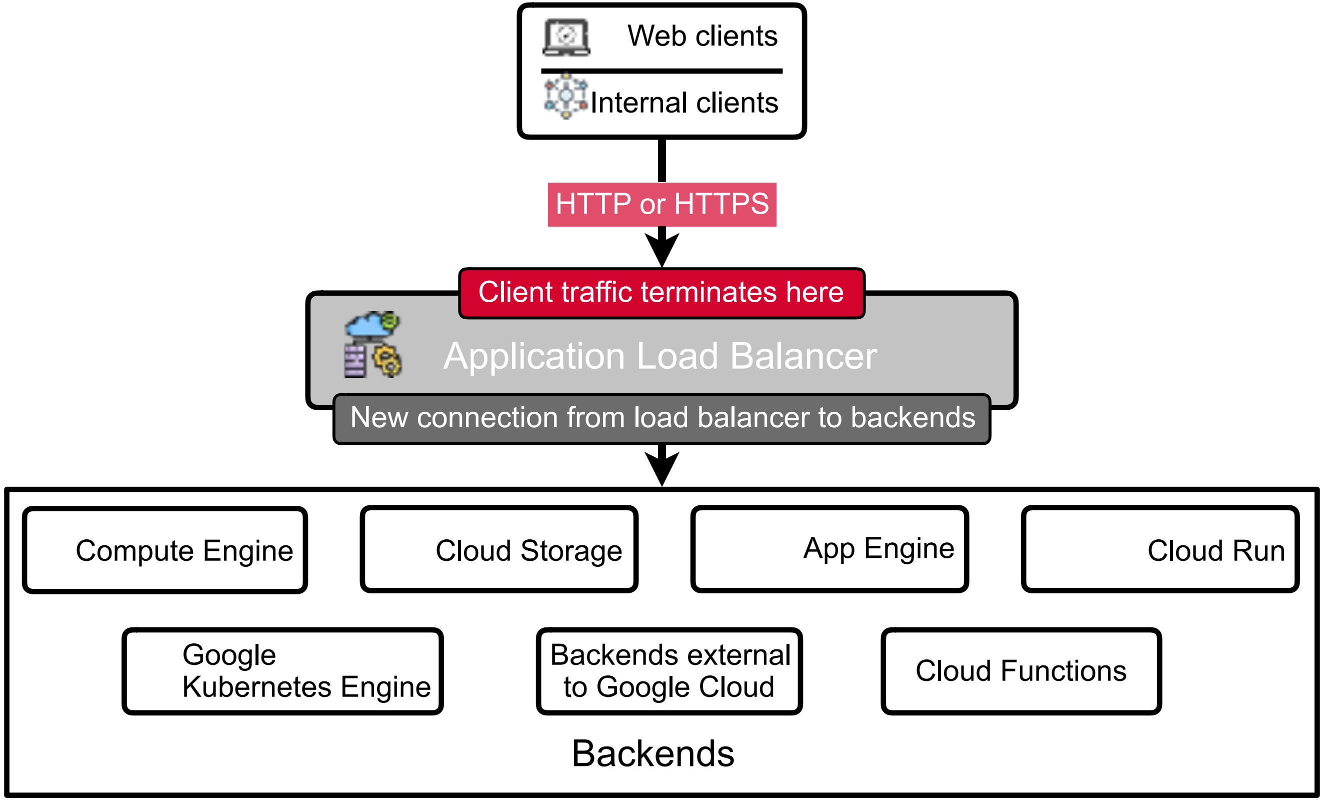
Application Load Balancers

- **Proxy-based** layer 7 load balancers
- Allow you to scale your services behind a single IP
- Distributes HTTP and HTTPS traffic to Google backends and external backends
  - Compute Engine, GKE, Cloud Run

Application Load Balancers



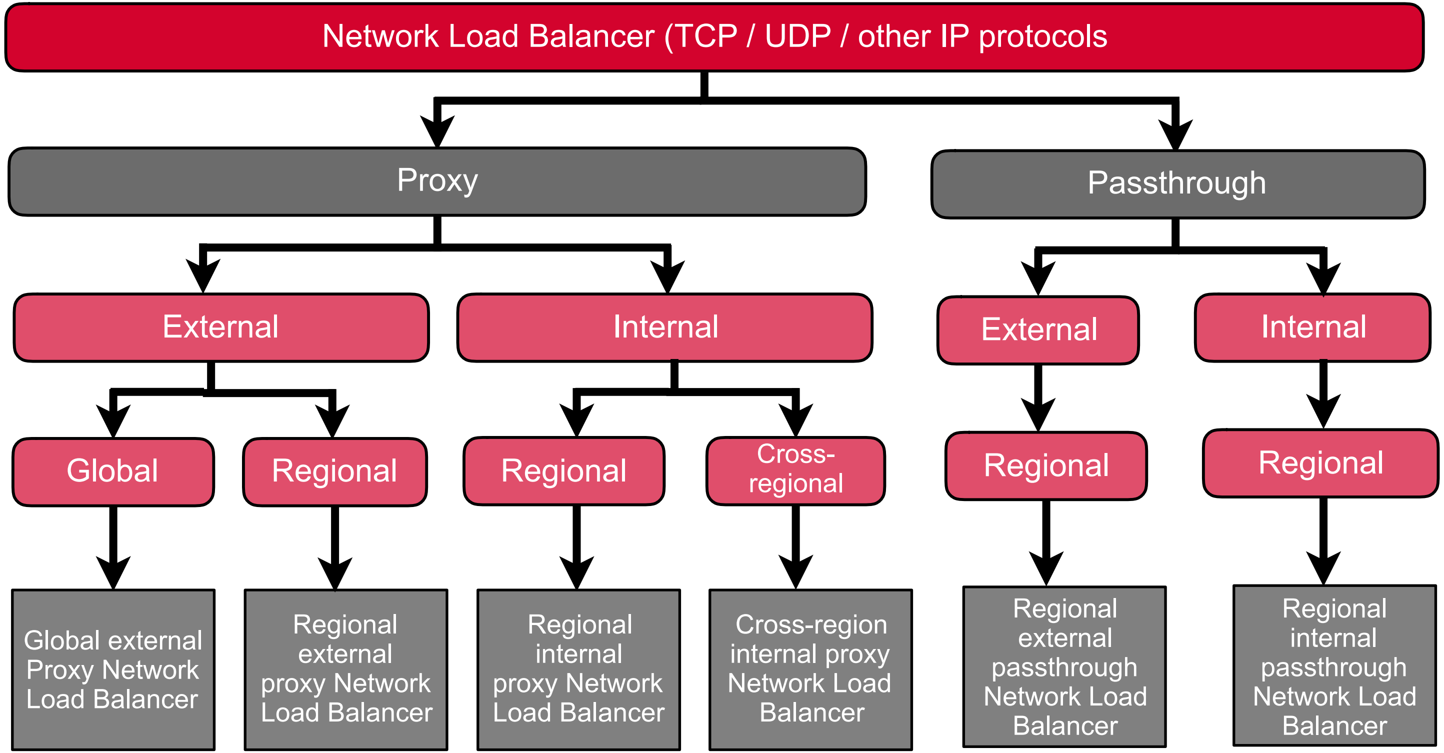
Application Load Balancers



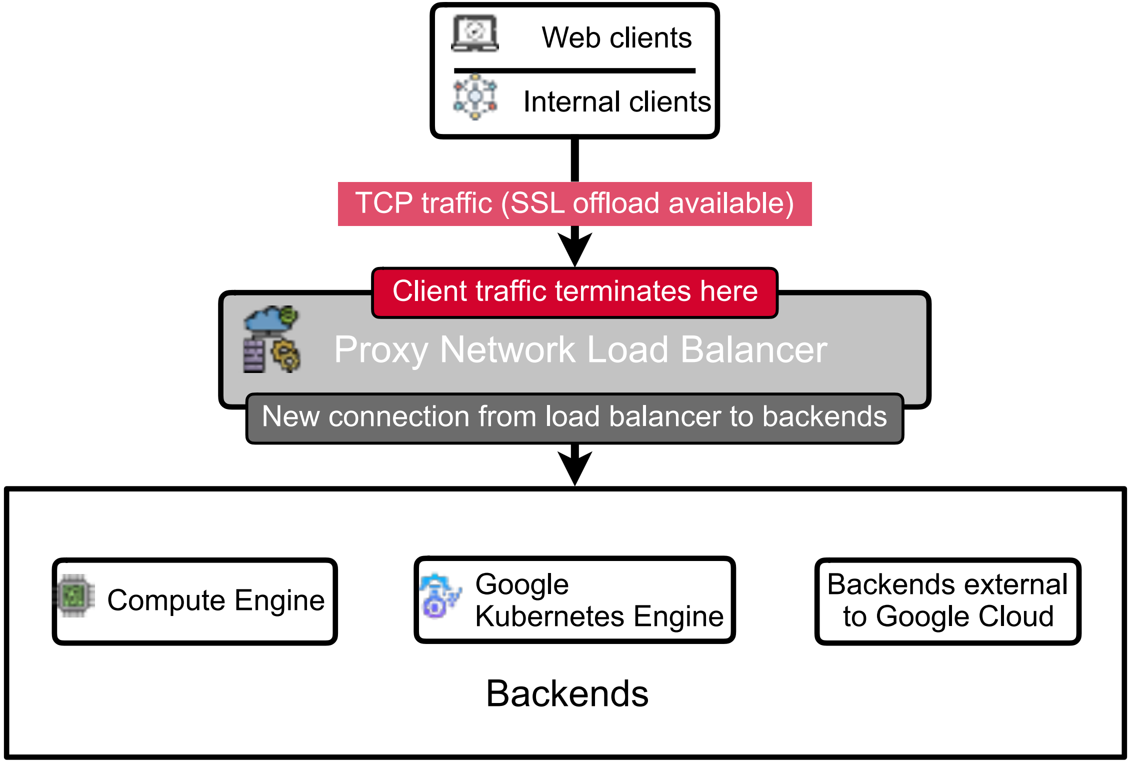
Network Load Balancers

- Layer 4 load balancers
- Handle TCP, UDP, or other IP protocol traffic
- Can be of two types
  - Proxy
  - Passthrough

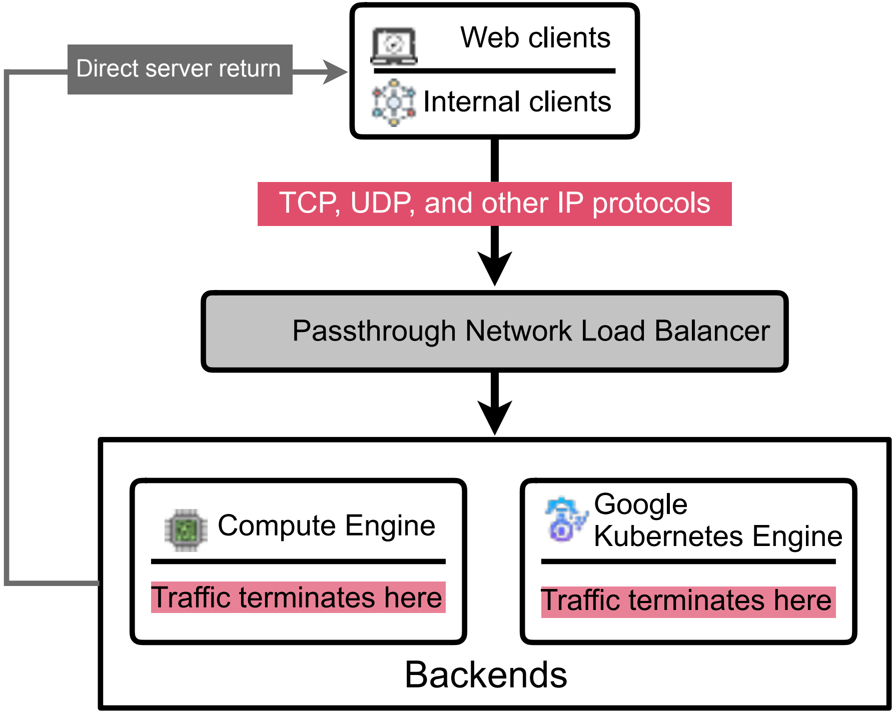
Network Load Balancers



Network Proxy Load Balancers



Network Passthrough Load Balancers



O'REILLY

Cloud CDN

Cloud CDN

Cloud CDN (Content Delivery Network) uses Google's global edge network (Points of Presence or PoPs) to serve content closer to users, which accelerates your websites and applications.

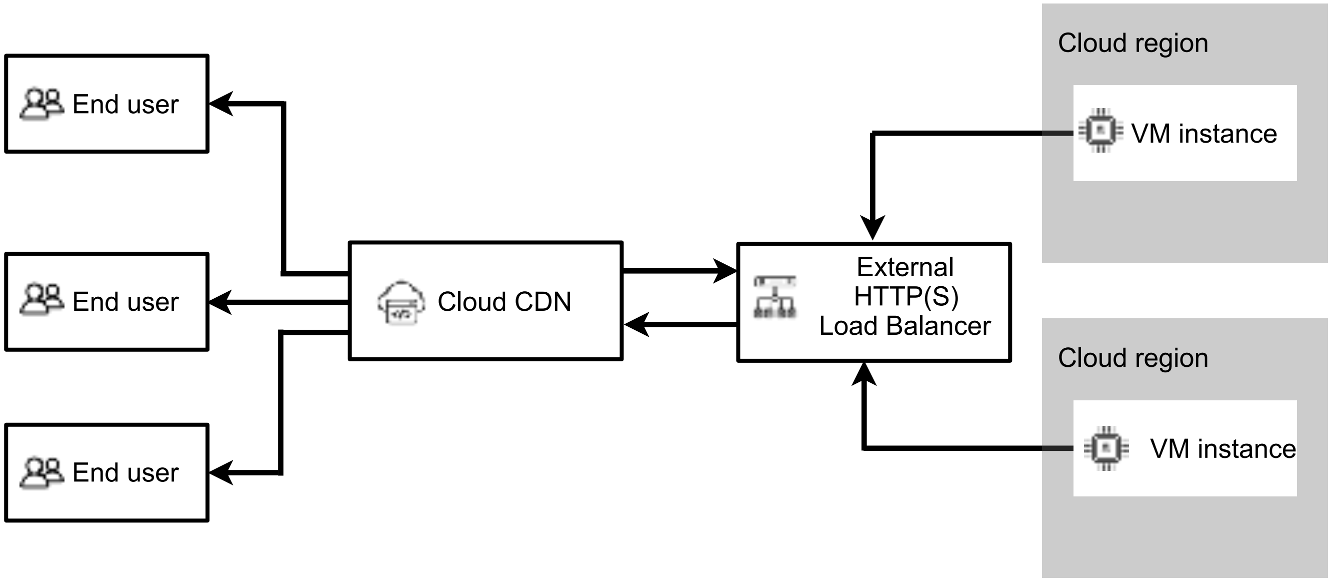
*Google Edge Network consists of numerous edge locations that are spread across various cities and countries globally. These edge locations are situated closer to users than Google’s central data centers, reducing latency by ensuring that users' data and requests travel shorter distances.

Cloud CDN

Usually sits in front of a load balancer and caches content from various types of backends.

Backends referred to as origin servers.

Cloud CDN



The Cloud CDN cache stores and manages content so that future requests for that content can be served faster. The cached content is a copy of cacheable content that is stored on origin servers.

Apigee

Apigee

Apigee is an API management platform developed by Google that enables organizations to design, secure, deploy, monitor, and scale APIs.

Apigee

- Manage API lifecycle
- Traffic management, authentication, analytics, monitoring
- Bot and misconfigured API detection
- Tools to package and manage APIs
- Governance policies

Apigee