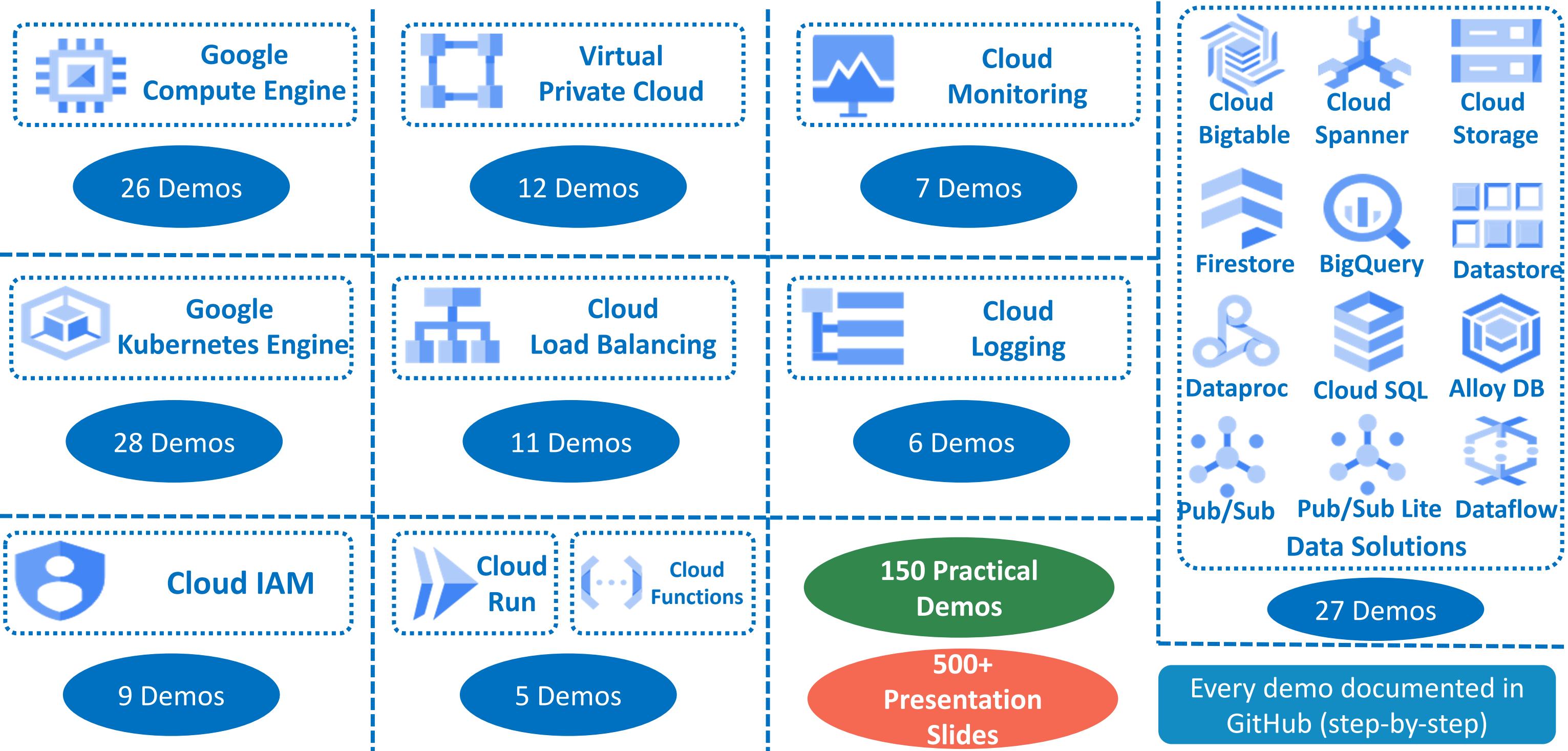


Google Cloud Certification

Associate Cloud Engineer - 130+ Practical Demos

Kalyan Reddy Daida

Google Cloud Associate Cloud Engineer



Google Compute Engine

Compute Engine VMs

VM Instance Basics

1

Startup Scripts

2

Cloud Shell & gcloud

3

Instance Templates

4

Machine Images

5

Spot VMs

6

Attach GPU to VM

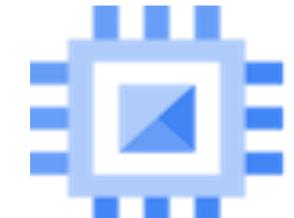
7

Sole-tenant Nodes

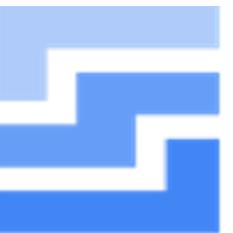
8

Ops Agent

9



**Compute
Engine**



**Persistent
Disks**

Persistent Disks

10

Cloud KMS

11

Attach Non-Boot Disk

12

Resize Disks

13

Regional Persistent Disks

14

Hyper Disks

15

Hyper Disk Storage Pools

16

Disk Images

17

Disk Snapshots

18

Local SSDs

26

**Compute Engine
Practical Demos**

**Every concept has a
practical demo**

Google Compute Engine

Instance Groups

Unmanaged

19

Managed Stateless

20

Managed Stateful

21

SSH Keys

Default SSH Keys

22

Project-level Metadata

23

Instance-level Metadata

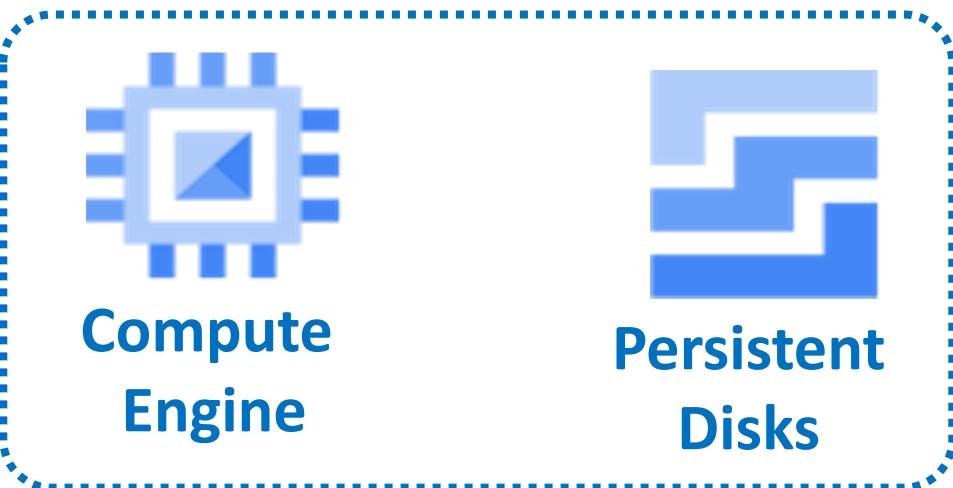
24

SSH OS Login

25

OS Login - Third Party Tools

26



26

Compute Engine
Practical Demos

Every concept with a practical
demo documented on GitHub

- ✓ Compute-Engine
 - ✓ 01-Compute-Engine-VM-Instances
 - > 01-01-VMInstance-Basics
 - > 01-02-VMInstance-with-Startup-Script
 - > 01-03-Cloud-Shell-and-gcloud-cli
 - > 01-04-VMInstance-with-InstanceTemplate
 - > 01-05-VMInstance-with-MachineImages
 - > 01-06-VMInstance-SpotVMs
 - > 01-07-VMInstance-Attach-GPU
 - > 01-08-VMInstance-Sole-tenant-Nodes
 - > 01-09-VMInstances-Ops-Agent
 - ✓ 02-Compute-Engine-Storage
 - > 02-01-Cloud-KMS
 - > 02-02-VMInstance-Attach-NonBootDisk
 - > 02-03-VMInstance-Resize-Disks-Boot-No
 - > 02-04-Regional-Persistent-Disks
 - > 02-05-Hyperdisks
 - > 02-06-Hyperdisks-with-storage-pool
 - > 02-07-VMInstance-with-DiskImage
 - > 02-08-VMInstance-with-DiskSnapshot
 - > 02-09-VMInstance-LocalSSD
 - ✓ 03-Instance-Groups
 - > 03-01-InstanceGroups-Unmanaged
 - > 03-02-InstanceGroups-Stateless
 - > 03-03-InstanceGroups-Stateful
 - ✓ 04-SSH-Keys
 - > 04-01-SSHKeys-default
 - > 04-02-SSHKeys-Project-Level-Metadata
 - > 04-03-SSHKeys-Instance-Level-Metadata
 - > 04-04-SSHKeys-OSLogin
 - > 04-05-SSHKeys-OSLogin-ThirdPartyTools

Google Kubernetes Engine

Kubernetes Workload Resources

Kubernetes Pods

Pods - Imperative

1

Pods - Declarative

2

Kubernetes Deployments

Deployment - Imperative

3

Deployment - Delcarative

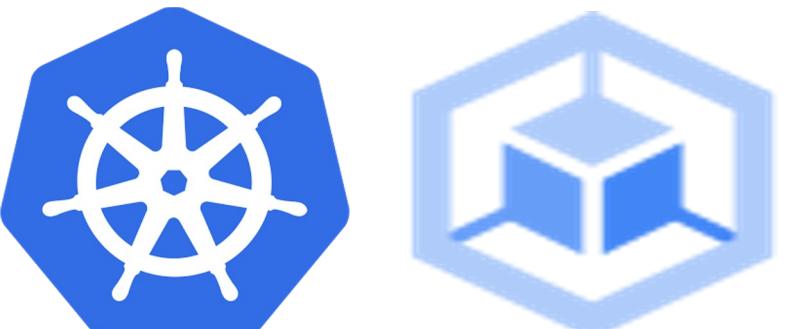
4

Deployment - Update

5

Deployment - Rollback

6



Kubernetes Engine

28

Google Kubernetes
Engine (GKE)
Practical Demos

Every concept has a
practical demo

Kubernetes Workload Resources

Other Workload Resources

7 Kubernetes ReplicaSets

8 Kubernetes StatefulSets

9 Kubernetes DaemonSets

Kubernetes Services

10 Node Port Service

11 Cluster IP Service

12 Headless Service

13 Ingress Service

Google Kubernetes Engine

Kubernetes Workload Resources

Kubernetes Jobs

Job Basics 14

Job Back-off limit 15

Job Completions 16

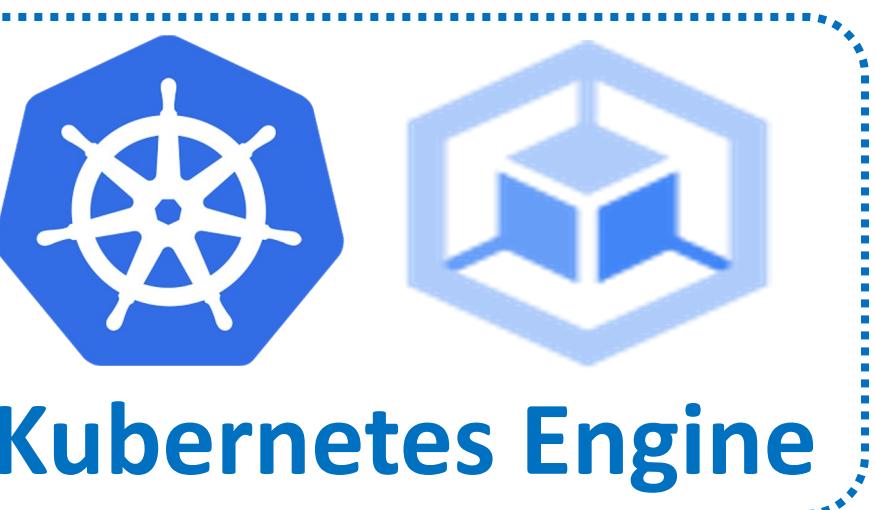
Job Parallelism 17

Job Active Deadline Seconds 18

Cron Jobs 19

Cluster Management

Node Pools & Node Selectors 20



28

Google Kubernetes
Engine (GKE)
Practical Demos

Every concept has a
practical demo

GKE Cluster Types

21 Regional Standard Cluster

22 GKE Auto-pilot Cluster

23 GKE Private Cluster

GKE Autoscaling

24 Cluster Autoscaler

25 Horizontal Pod Autoscaling

26 Vertical Pod Autoscaling

Kubernetes Storage

27 GKE Storage with Compute
Engine Disks

28 Google Artifact Registry

Google Kubernetes Engine

- ✓ google-kubernetes-engine
 - > 01-Create-GKE-Regional-Standard-Cluster
 - ✓ 02-Kubernetes-Pods-and-Services
 - > 01-Pods-and-Services-Imperative
 - > 02-Pods-and-Services-Declarative
 - > 03-Kubernetes-ReplicaSets
 - ✓ 04-Kubernetes-Deployments-and-Services
 - > 04-01-Kubernetes-Deployments-Imperative
 - > 04-02-Kubernetes-Deployments-Update
 - > 04-03-Kubernetes-Deployments-ROLLBACK
 - > 04-04-Kubernetes-Deployments-Declarative
 - > 05-NodePools-and-NodeSelectors
 - > 06-Kubernetes-DaemonSets
 - ✓ 07-Kubernetes-Jobs
 - > 07-01-Job-basics
 - > 07-02-Job-backofflimit
 - > 07-03-Job-Completions
 - > 07-04-Job-Parallelism
 - > 07-05-Job-ActiveDeadlineSeconds
 - > 07-06-Cron-Job
 - > 08-Kubernetes-Services
 - > 09-Kubernetes-Storage
 - > 10-Kubernetes-StatefulSets
 - > 11-GKE-Autoscaling
 - > 12-GKE-Artifact-Registry
 - > 13-GKE-Private-Cluster
 - > 14-GKE-AutoPilot-Cluster



Kubernetes Engine

28

Google Kubernetes
Engine (GKE)
Practical Demos

Every concept with a practical
demo documented on GitHub

08-04-Kubernetes-Ingress-Service

kube-manifests

- ! 01-Nginx-App1-Deployment-and-NodePortService.yaml
- ! 02-Nginx-App2-Deployment-and-NodePortService.yaml
- ! 03-Nginx-App3-Deployment-and-NodePortService.yaml
- ! 04-Ingress-ContextPath-Based-Routing.yaml

Kubernetes YAML files
are well-structured and
well-documented on
GitHub

Google Cloud Identity & Access Management



Cloud IAM

IAM Roles

1

IAM Roles
(gcloud)

2

IAM Policy

3

IAM Conditions

4

IAM Service Accounts

5

IAM Service Accounts
(gcloud)

6

IAM Service Account
Impersonation

7

IAM Service Account Keys
(Long-lived Credentials)

8

IAM Service Account
Short-lived Credentials

9



Cloud IAM

9
Cloud IAM
Practical Demos

- ✓ Cloud-IAM
 - > 01-IAM-Roles
 - > 02-IAM-Roles-gcloud
 - > 03-IAM-Policy
 - > 04-IAM-Conditions
 - > 05-IAM-ServiceAccounts
 - > 06-IAM-ServiceAccounts-gcloud
 - > 07-IAM-ServiceAccount-Impersonation
 - > 08-IAM-Service-Account-Keys
 - > 09-IAM-ServiceAccount-ShortLived-Credentials



Every concept has a practical demo

Every concept with a practical demo documented on GitHub

Google Virtual Private Cloud (VPC)

VPC Types

VPC Auto & Custom

1

VPC Networking

Static
Internal and External IP

2

Cloud NAT &
Cloud Router

3

VPC
Private Google Access

4

Cloud Domains

5

Cloud DNS

6



Virtual Private Cloud

12

VPC Practical Demos

Every concept has a
practical demo

VPC Firewall Rules

7

Ingress Rule Basics

8

Ingress Rule
with Target Tags

9

Ingress Rule
with Service Accounts

10

Ingress Rule
With Destination Filter

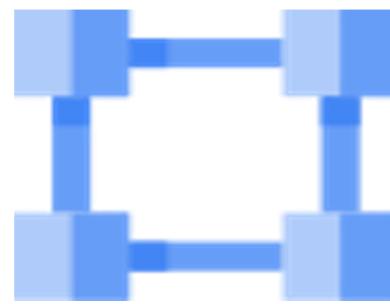
11

Egress
Deny Rule

12

Firewall Policies

Google Virtual Private Cloud (VPC)



Virtual Private Cloud

12
VPC

Practical Demos

Every concept with a practical
demo documented on GitHub

- ✓ Cloud-VPC
 - > 01-VPC-Auto-and-Custom
 - > 02-VPC-Firewall-Rules-Ingress-All
 - > 03-VPC-Firewall-Rules-Ingress-Target-Tags
 - > 04-VPC-Firewall-Rules-Ingress-SA
 - > 05-VPC-Firewall-Rules-Ingress-Destination
 - > 06-VPC-Firewall-Rules-Egress
 - > 07-VPC-Firewall-Policies
 - > 08-VPC-Static-External-Internal-IP
 - > 09-VPC-InternalIP-Only-CloudNAT
 - > 10-VPC-Private-Google-Access
- ✓ Cloud-Domains-and-Cloud-DNS
 - > 01-Cloud-Domains-Basics
 - > 02-Cloud-DNS-Basics

Google Cloud Load Balancing

Cloud Load Balancing

**Regional
Managed Instance Groups**

1

**Global
HTTP LB**

2

**Global HTTPS LB
(Self-signed SSL)**

3

**Global HTTPS LB
(Google Managed SSL)**

4

**Global
TCP Proxy LB**

5

**Global
SSL Proxy LB**

6

**Zonal
Managed Instance Groups**

7

**Regional
HTTP LB**

8

**Regional
HTTP Internal LB**

9



Cloud Load Balancing

11

Cloud
Load Balancing
Practical Demos

Cloud Load Balancing

10

**Regional
TCP Proxy**

11

**Regional
TCP Pass-through**

- ✓ Cloud-LoadBalancer
 - > 01-Regional-Managed-Instance-Groups
 - > 02-Cloud-LoadBalancer-Global-HTTP
 - > 03-Cloud-LoadBalancer-Global-HTTPS-SelfSignedSSL
 - > 04-Cloud-LoadBalancer-Global-HTTPS-GoogleManagedSSL
 - > 05-Cloud-LoadBalancer-Global-TCPProxy
 - > 06-Cloud-LoadBalancer-Global-SSLProxy
 - > 07-Zonal-Managed-Instance-Groups
 - > 08-Cloud-LoadBalancer-Regional-HTTP
 - > 09-Cloud-LoadBalancer-Regional-Internal-HTTP
 - > 10-Cloud-LoadBalancer-Regional-TCP-Proxy
 - > 11-Cloud-LoadBalancer-Regional-TCP-Passthrough

Every concept has a practical demo

Every concept with a practical demo documented on GitHub



Google Cloud Run & Cloud Functions

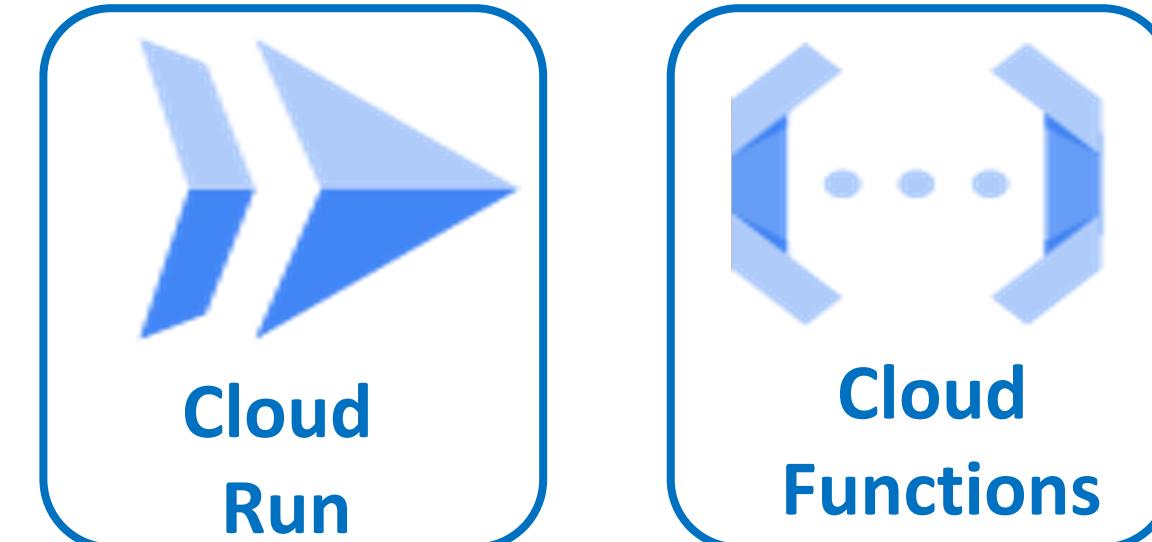
Cloud Run

Cloud Run Services

1

Cloud Run Jobs

2



Cloud Functions

Cloud Functions
HTTP

1

Cloud Functions
Events with Pub/Sub

2

Cloud Functions
Events with Cloud Storage

3

- ✓ Cloud-Run
- > 01-Cloud-Run-Services
- > 02-Cloud-Run-Jobs

5
Cloud Run &
Cloud Functions
Practical Demos

- ✓ Cloud-Functions
 - > 01-Cloud-Functions-HTTP
 - > 02-Cloud-Functions-Events-with-Cloud-PubSub
 - > 03-Cloud-Functions-Events-with-Cloud-Storage

Every concept has a practical demo

Every concept with a practical demo documented on GitHub

Google Cloud Data Solutions

Cloud SQL

Cloud SQL Basics

1

Cloud SQL Backup & Restore

Cloud SQL Import & Export

Cloud Dataflow

Cloud Dataflow Basics

4

Cloud Spanner

Cloud Spanner Basics

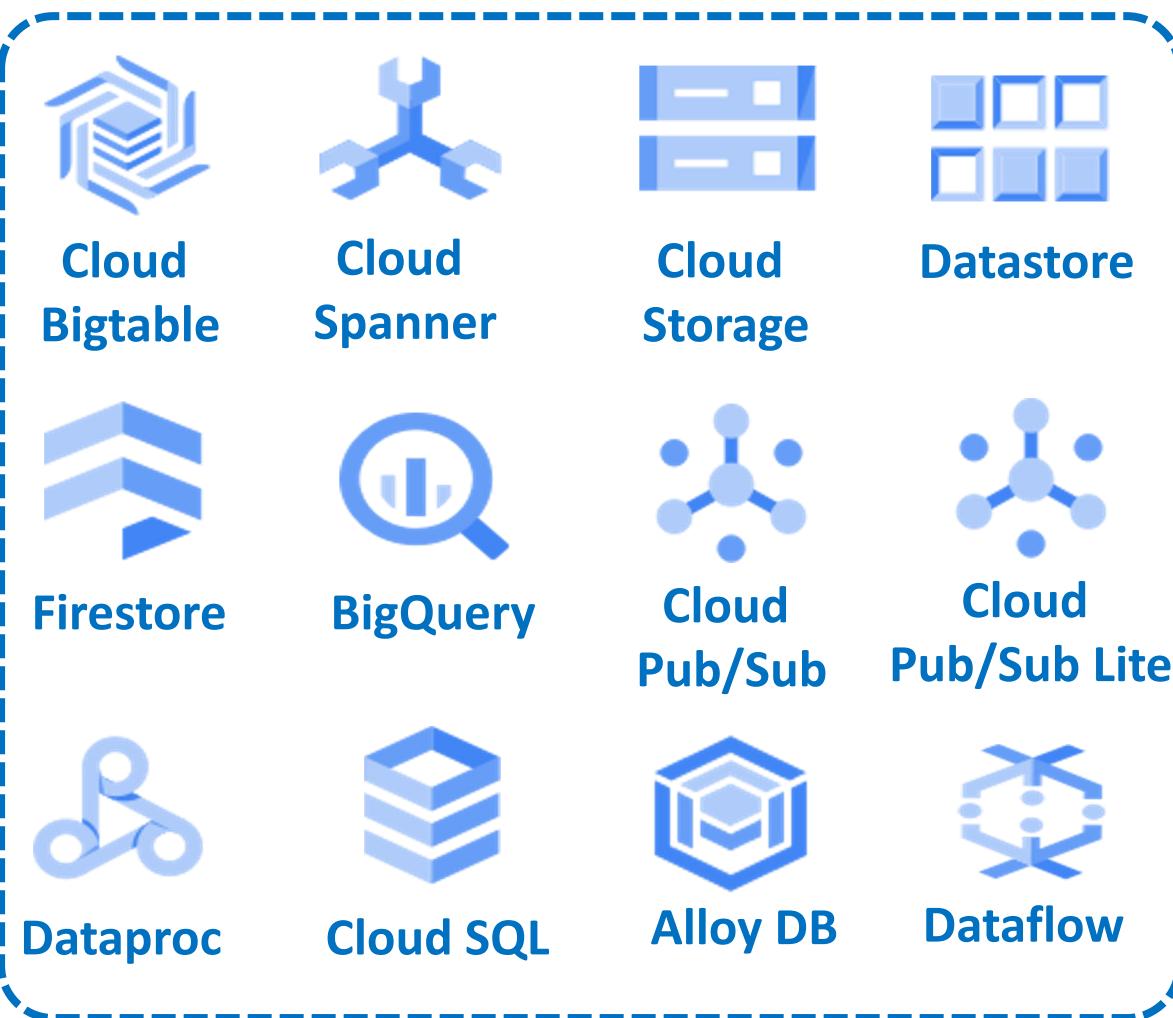
5

Cloud Spanner Import & Export

6

Cloud Spanner Backup & Restore

7



12

GCP Data Services

27

Practical Demos

Alloy DB

8

Alloy DB Basics

Cloud Firestore

9

Cloud Firestore Basics

10

Cloud Firestore gcloud

Cloud Datastore

11

Cloud Datastore Basics

Cloud Bigtable

12

Cloud Bigtable Basics

13

Cloud Bigtable Backup & Restore

Google Cloud Data Solutions

Cloud Dataproc

Cloud Dataproc Cluster and Jobs

14

Cloud Dataproc Serverless Batch

15

Cloud Pub/Sub, Pub/Sub Lite

Cloud Pub/Sub Basics

16

Cloud Pub/Sub using gcloud

17

Cloud Pub/Sub write to Cloud Storage

18

Cloud Pub/Sub Lite Basics

19

Every Data Solution will have 1 or more practical demos



Cloud Bigtable



Cloud Spanner



Cloud Storage



Datastore



Firestore



BigQuery



Cloud Pub/Sub



Cloud Pub/Sub Lite



Dataproc



Cloud SQL



Alloy DB



Dataflow

12

GCP Data Services

27

Practical Demos

Cloud Storage

20

Cloud Storage Basics

21

Cloud Storage Classes

22

Object Lifecycle Management

23

Object Versioning

24

Security ACLs

25

Security UBLA

Cloud BigQuery

26

Query Public Datasets

27

Create Datasets, Tables, Load Data & Query Data

Google Cloud Monitoring

Cloud Monitoring

Uptime Checks

1

Alert Policies

2

Monitoring Groups

3

Synthetic Monitor
Custom Scripts

4

Synthetic Monitor
Mocha Template

5

Synthetic Monitor
Broken-link Checker

6

Custom Dashboards

7



Cloud Monitoring

7

Cloud Monitoring Practical Demos

Every concept has a practical demo

Every concept with a practical demo documented on GitHub

- ✓ Cloud-Monitoring
 - > 01-Monitoring-Uptime-Checks
 - > 02-Monitoring-Alerts
 - > 03-Monitoring-Groups
 - > 04-Synthetic-Monitor-Custom-Script
 - > 05-Synthetic-Monitor-Mocha
 - > 06-Synthetic-Monitor-Brokenlink-checker
 - > 07-Custom-Dashboards



Google Cloud Logging

Cloud Logging

Log Explorer

1

Application Integration
(Nginx Logs & Metrics)

2

Log-based
Metric & Alerts

3

Log-based
Alert Policy

4

Log Storage &
Log Router Sinks

5

Log Analytics &
Big Query Datasets

6



Cloud Logging

6

Cloud Logging Practical Demos

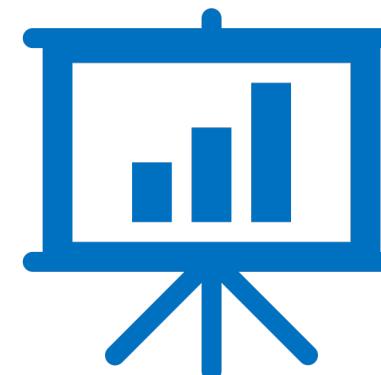
- ✓ Cloud-Logging
 - > 01-Log-Explorer
 - > 02-Application-Integration
 - > 03-Log-based-Metric
 - > 04-Log-based-Alert-Policy
 - > 05-Logs-Storage
 - > 06-Log-Analytics

Every concept has a practical demo

Every concept with a practical demo documented on GitHub

GitHub Repository

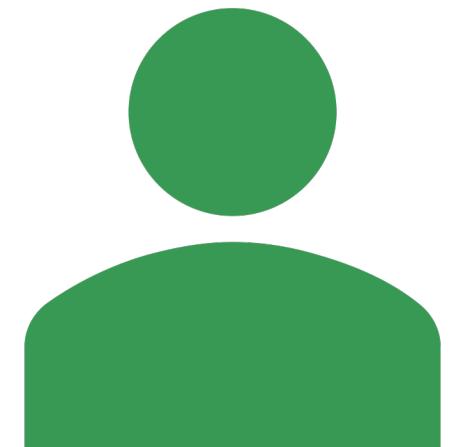
Repository Used For	Repository URL
Course Main Repository with step-by-step documentation	https://github.com/stacksimplify/google-cloud-certifications/
Course Presentation	https://github.com/stacksimplify/google-cloud-certifications/tree/main/course-presentation



500+ presentation slides outlining various concepts we have implemented



Google Cloud Account



Google Cloud Account

 Try Google Cloud for free

Step 1 of 3 Account information



GCP User901

gcpuser901@gmail.com

[SWITCH ACCOUNT](#)

Country

India

What best describes your organisation or needs?

Please select

Personal project

Terms of Service

I have read and agree to the [Google Cloud Platform Terms of Service](#), [Supplemental Free Trial Terms of Service](#) and the Terms of Service of [any applicable services and APIs](#).

Required to continue

[CONTINUE](#)

Access to all Cloud Platform products

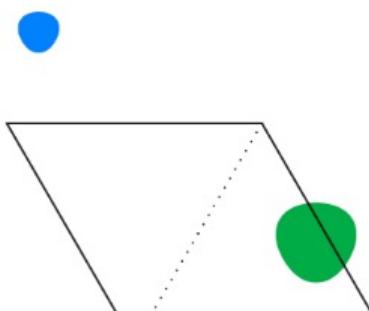
Get everything that you need to build and run your apps, websites and services, including Firebase and the Google Maps API.

\$300 credit for free

Put Google Cloud to work with \$300 in credit to spend over the next 90 days.

No auto-charge after free trial ends

We ask you for your credit card details to make sure that you are not a robot. You won't be charged unless you manually upgrade to a paid account.



Google Cloud Account



Try Google Cloud for free

Step 2 of 3 Identity verification and contact information

We'll send a text message with a six-digit code to verify your identity and confirm where we can reach you about solutions to support your Cloud experience.
Standard rates apply.



Phone number

+91 XXXXXXXXXX

SEND CODE

Google Cloud Account - Individual

 Try Google Cloud for free

Step 3 of 3 Payment information verification

Your payment information helps us to reduce fraud and abuse. You won't be charged unless you turn on automatic billing.

 Account type 

Individual

Only Business accounts can have multiple users. You cannot change the account type after signing up. In some countries, this selection affects your tax options.

[Learn more](#)

Payment method

Card number

#

Card number is required



You'll be charged automatically on the 1st of each month. If your balance reaches your payment threshold before then, you'll be charged immediately. [Learn more](#)



Tax information



The personal information that you provide here will be added to your payments profile. It will be stored securely and treated in accordance with the [Google Privacy Policy](#).

[START MY FREE TRIAL](#)

Google Cloud Account - Business

 Try Google Cloud for free

Step 3 of 3 Payment information verification

Your payment information helps us to reduce fraud and abuse. You won't be charged unless you turn on automatic billing.

 Account type 

Business

Only Business accounts can have multiple users. You cannot change the account type after signing up. In some countries, this selection affects your tax options.

[Learn more](#)

 Business name
stacksimplify

Payment method

Card number

#

Card number is required



Tax information  

Tax status: Business

The personal information that you provide here will be added to your payments profile. It will be stored securely and treated in accordance with the [Google Privacy Policy](#).

START MY FREE TRIAL

Google Cloud Account

Payment method

Card number



Invalid card number

MM YY
01 / 30

Cardholder name

Kalyan Reddy Daida

Address line 1

Address line 1 is required

Address line 2

Town/City

Town/City is required

Postcode



Postcode is required

State

Telangana 

Provide Address Details

Google Cloud Account

The screenshot shows the Google Cloud Platform dashboard with a prominent white survey overlay in the center. The overlay is titled "Welcome, GCP User901!" and informs the user about a free trial with \$300 in credit. It consists of four numbered questions:

- 1 What best describes your organisation or needs?
Please select *
Personal project
- 2 What brought you to Google Cloud?
- 3 What are you interested in doing with Google Cloud?
- 4 What best describes your role?

At the bottom right of the overlay are "CLOSE" and "DONE" buttons. The background of the dashboard shows various pinned services like APIs and services, Compute Engine, and Cloud Storage.

Google Cloud Account



Welcome, GCP User901!

Your free trial includes \$300 in credit to spend over the next 90 days. To help us serve you better, please answer 4 questions.

1 What best describes your organisation or needs?

2 What brought you to Google Cloud?

Please select *

Evaluate technical capabilities of Google Cloud

NEXT

3 What are you interested in doing with Google Cloud?

4 What best describes your role?

CLOSE

DONE



Welcome, GCP User901!

1 What best describes your organisation or needs?

2 What brought you to Google Cloud?

3 What are you interested in doing with Google Cloud?

Websites

Mobile apps

Storage/backup

Data analytics

Artificial intelligence/machine learning

Game development

Containerisation

Data management

Virtual machines (VMs)

Google Maps

Other APIs (e.g. Text-to-Speech, Speech-to-Text, Vision)

Google Photos or Google Workspace

Other

I'm not sure yet

NEXT

4 What best describes your role?

CLOSE

DONE

Google Cloud Account



Welcome, GCP User901!

Your free trial includes \$300 in credit to spend over the next 90 days. To help us serve you better, please answer 4 questions.

- 1 What best describes your organisation or needs?
- 2 What brought you to Google Cloud?
- 3 What are you interested in doing with Google Cloud?
- 4 What best describes your role?

Please select *

Solution/System architect

CLOSE

DONE

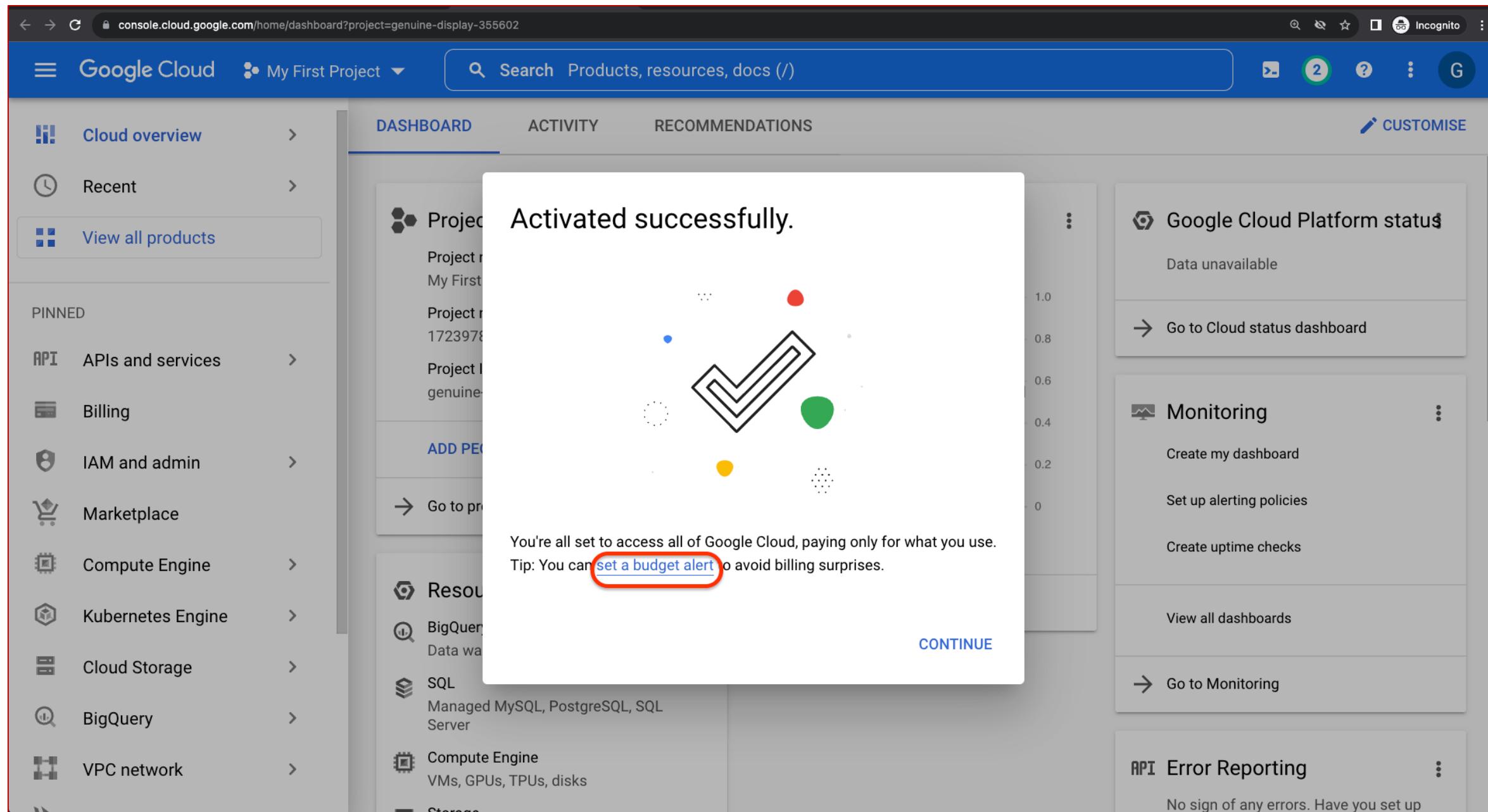
Google Cloud Account - ACTIVATE

The screenshot shows the Google Cloud Platform dashboard for a project named "My First Project". The top navigation bar includes a "Free trial status" message about credit remaining, and a prominent blue "ACTIVATE" button. The dashboard features sections for "Cloud overview", "Recent", "View all products", "PINNED APIs and services", and "Billing". The main content area displays "Project info" (Project name: My First Project, Project number: 172397892561, Project ID: genuine-display-355602), "API APIs" (Requests (requests/sec) chart showing 1.0, 0.8, 0.6, 0.4 levels, with a note: "No data is available for the selected time frame."), and "Google Cloud Platform status" (Data unavailable, with a link to "Go to Cloud status dashboard"). A "Monitoring" section is also visible.

Google Cloud Account - ACTIVATE

The screenshot shows the Google Cloud Platform dashboard for a project named "genuine-display-355602". At the top, there is a banner indicating a free trial status with ₹23,635.88 credit and 91 days remaining. A "DISMISS" button and an "ACTIVATE" button are visible. The dashboard has tabs for DASHBOARD, ACTIVITY, and RECOMMENDATIONS. On the left, a sidebar lists various services: Cloud overview, Recent, View all products (pinned), APIs and services, Billing, IAM and admin, Marketplace, Compute Engine, Kubernetes Engine, Cloud Storage, and BigQuery. In the center, a modal dialog box titled "Activate your full account" lists three benefits: "Keep your cloud running uninterrupted", "Keep any remaining credits to spend during your free trial", and "Pay only for what you use – automatic billing starts once your free trial ends". It includes "CANCEL" and "ACTIVATE" buttons, with the "ACTIVATE" button being highlighted with a red box. To the right of the modal, there are sections for "Google Cloud Platform status" (data unavailable) and "Monitoring" (with options to create a dashboard, set up alerting policies, and create uptime checks).

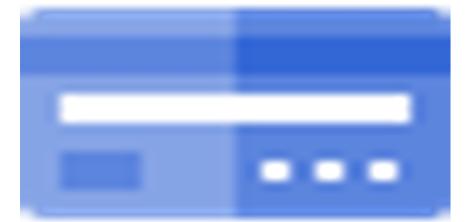
Google Cloud Account – Set a Budget Alert





Google Cloud

Create Budget Alerts



Google Cloud Account - Create Budget Alerts

Google Cloud Search Bill X 2 ? :

Billing ← Create budget LEARN

My Billing Account

- Overview
- Reports
- Cost table
- Cost breakdown
- Commitments
- Commitment analysis
- Budgets & alerts**
- Billing export
- Pricing
- Documents
- Manage resources
- Release notes

1 Scope

Name *

A budget enables you to track your actual spend against your planned spend.

Time range

The month starts on the first of the month and resets at the beginning of each month.

A budget can be scoped to focus on a specific set of resources.

Projects

Services

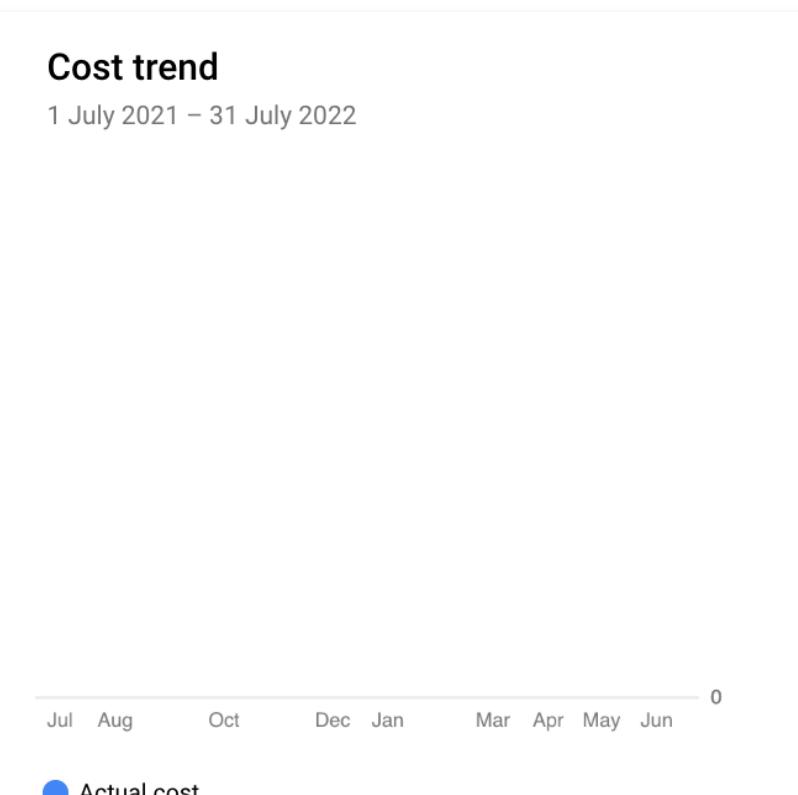
Credits
Selected credits are applied to the total cost. Budget tracks the total cost minus any applicable selected credits

Discounts [?](#)

Promotions and others [?](#)

NEXT

Cost trend
1 July 2021 – 31 July 2022



Jul Aug Oct Dec Jan Mar Apr May Jun 0

Actual cost

→ View report

Google Cloud Account - Create Budget Alerts

The screenshot shows the Google Cloud Billing interface with the following details:

- Left Sidebar:** Shows navigation links for Billing (Overview, Reports, Cost table, Cost breakdown, Commitments, Commitment analysis, Budgets & alerts, Billing export, Pricing, Documents, Manage resources, Release notes).
- Header:** Google Cloud, Search Bar, Notifications (2), Help, and User Profile.
- Current View:** Create budget, Step 2: Amount.
- Form Fields:**
 - Budget type:** Specified amount.
 - Target amount ***: ₹ 2000.
- Actions:** FINISH, CANCEL, NEXT.
- Cost trend:** A chart showing actual cost from July 2021 to June 2022, with a target amount of ₹ 2K marked by a dashed red line.

Google Cloud Account - Create Budget Alerts

≡ Google Cloud

🔍 Search
Bill
X
2
?
⋮
G

Billing
[Create budget](#)
 LEARN

My Billing Account

- Overview
- Reports
- Cost table
- Cost breakdown
- Commitments
- Commitment analysis
- Budgets & alerts**
- Billing export
- Pricing
- Documents
- Manage resources
- Release notes

Actions

Set alert threshold rules

Send email alert notifications after the actual or forecasted spend exceeds a percentage of the budget or a specified amount. [Learn more.](#)

Percentage of bu...	Amount 1 *	Trigger on...
50 %	₹ 1000	Actual
Percentage of bu...	Amount 2 *	Trigger on...
90 %	₹ 1800	Actual
Percentage of bu...	Amount 3 *	Trigger on...
100 %	₹ 2000	Actual

[+ ADD THRESHOLD](#)

Manage notifications

Send email alert notifications to billing admins and users of this billing account.

Email alerts to billing admins and users

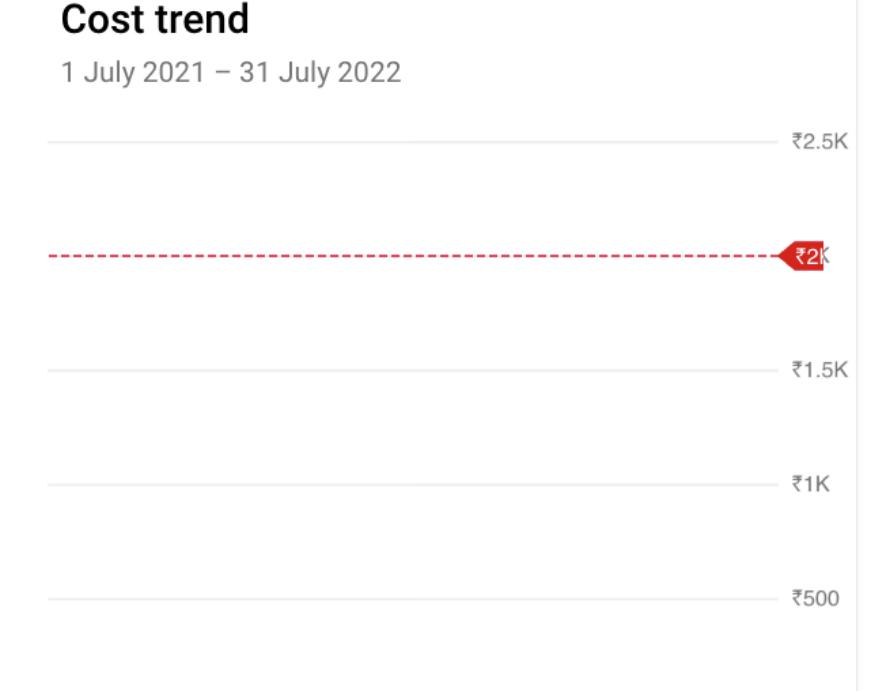
Allow Monitoring email notification channels to receive alerts when this budget reaches thresholds.

Link Monitoring email notification channels to this budget

Select a project and a maximum of 5 Monitoring email notification channels

Cost trend

1 July 2021 – 31 July 2022



Actual cost

View report



Google Cloud Regions & Zones

Google Cloud - Regions

- **What is a Region ?**

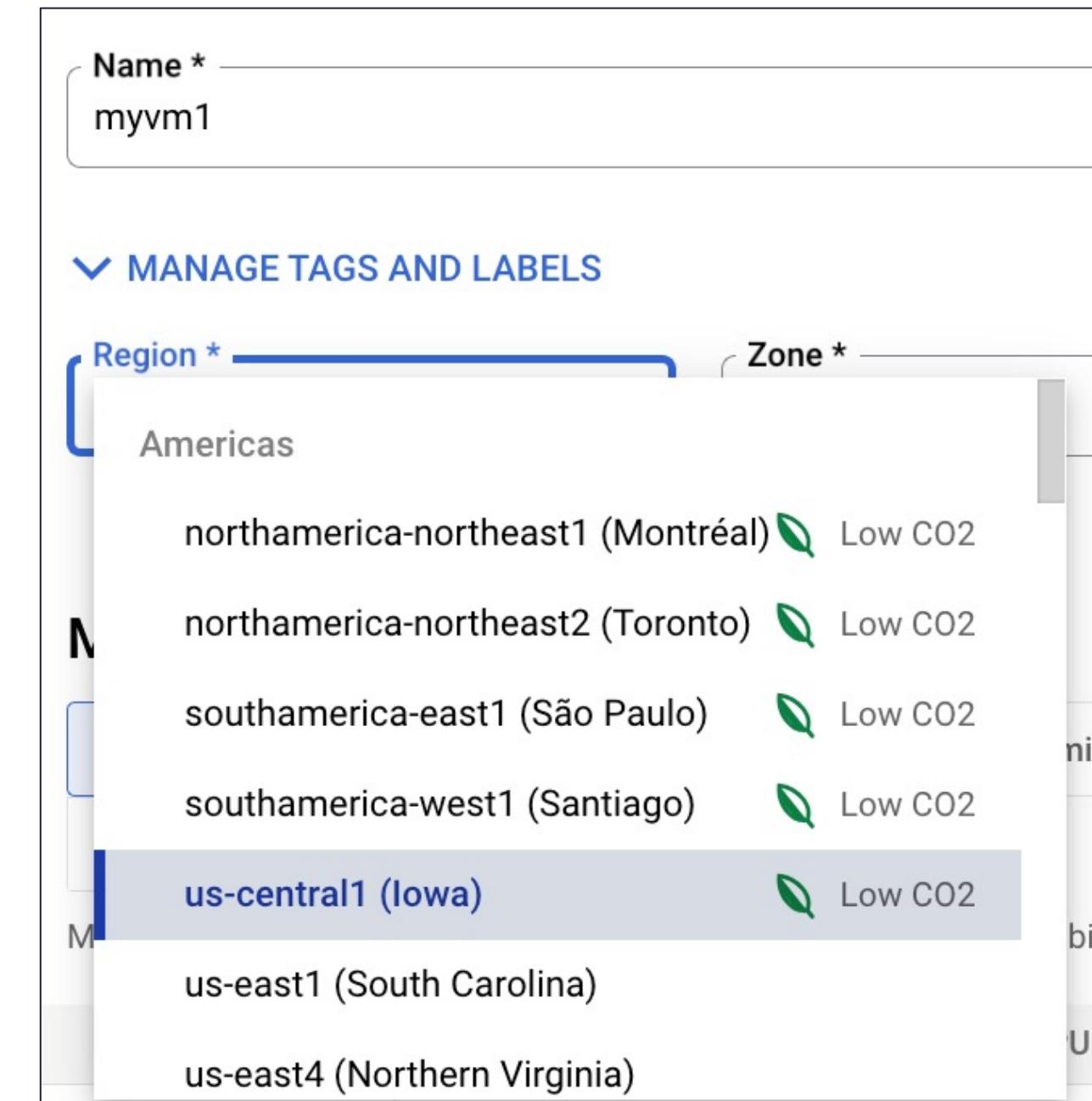
- **Region:** Geographical location
 - Example:
 - us-central1 (Iowa)
 - us-east1 (South Carolina)
 - us-east4 (Northern Virginia)

- **Why do we need Regions ?**

- **High Availability:** To provide High Availability for our applications across Regions
- **Compliance:** To be in compliance with respective region government regulations
- **Low-latency:** Provides Low-latency if your applications hosted nearer to your customers

- **Where do we use Region as a Cloud Admin in Google Cloud ?**

- When creating Google Cloud resources (Example: VM Instance) we need to select the region where that VM instance to be created



Google Cloud - Regions

- How many regions are there in Google Cloud ?
- Google Cloud Platform has 40 regions and growing

40

REGIONS



Reference: <https://cloud.google.com/about/locations#lightbox-regions-map>

Google Cloud - Zones

- **What are Zones ?**

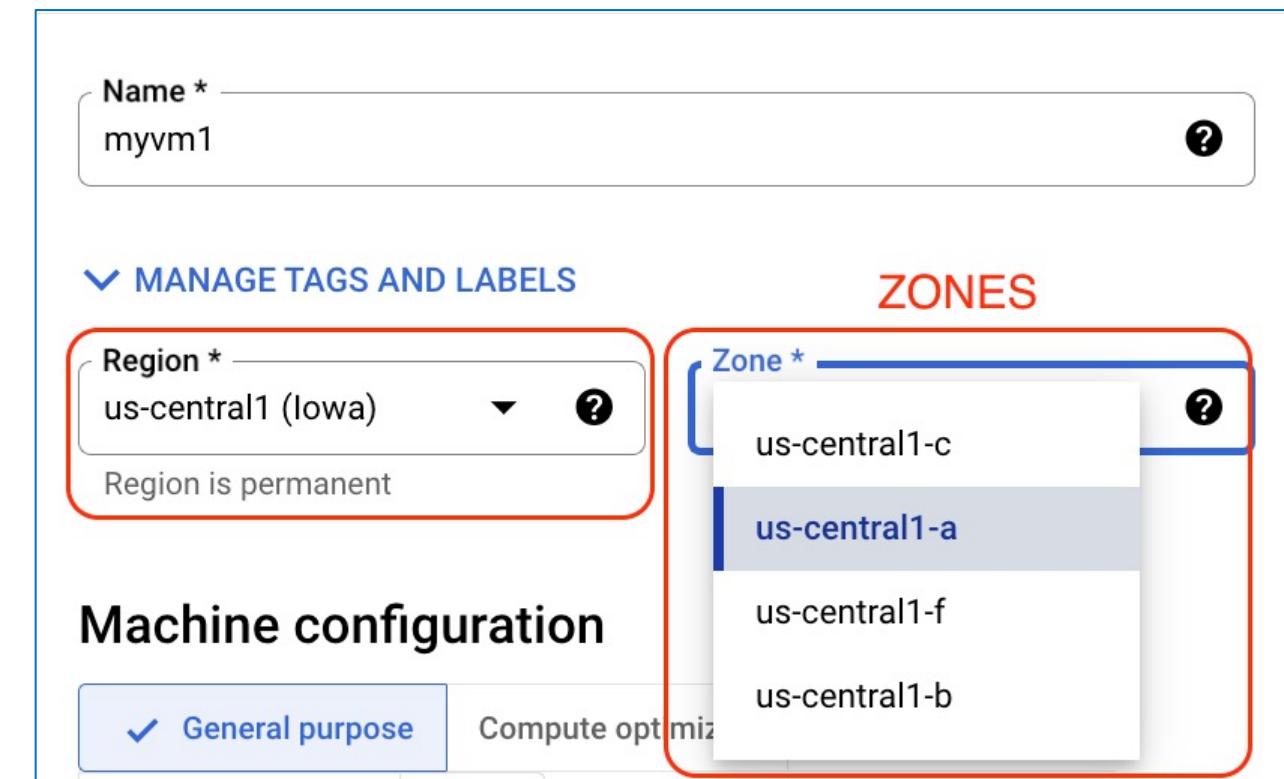
- Zones are **deployment areas** in a region
- In each region we will have 3 or more zones

- **Why do we need multiple Zones in a region ?**

- To achieve **High availability and fault-tolerance** in a region for our applications

- **How are zones connected in a region ?**

- Zones are connected with **low-latency links** (Network performance will be **very high** in between zones)



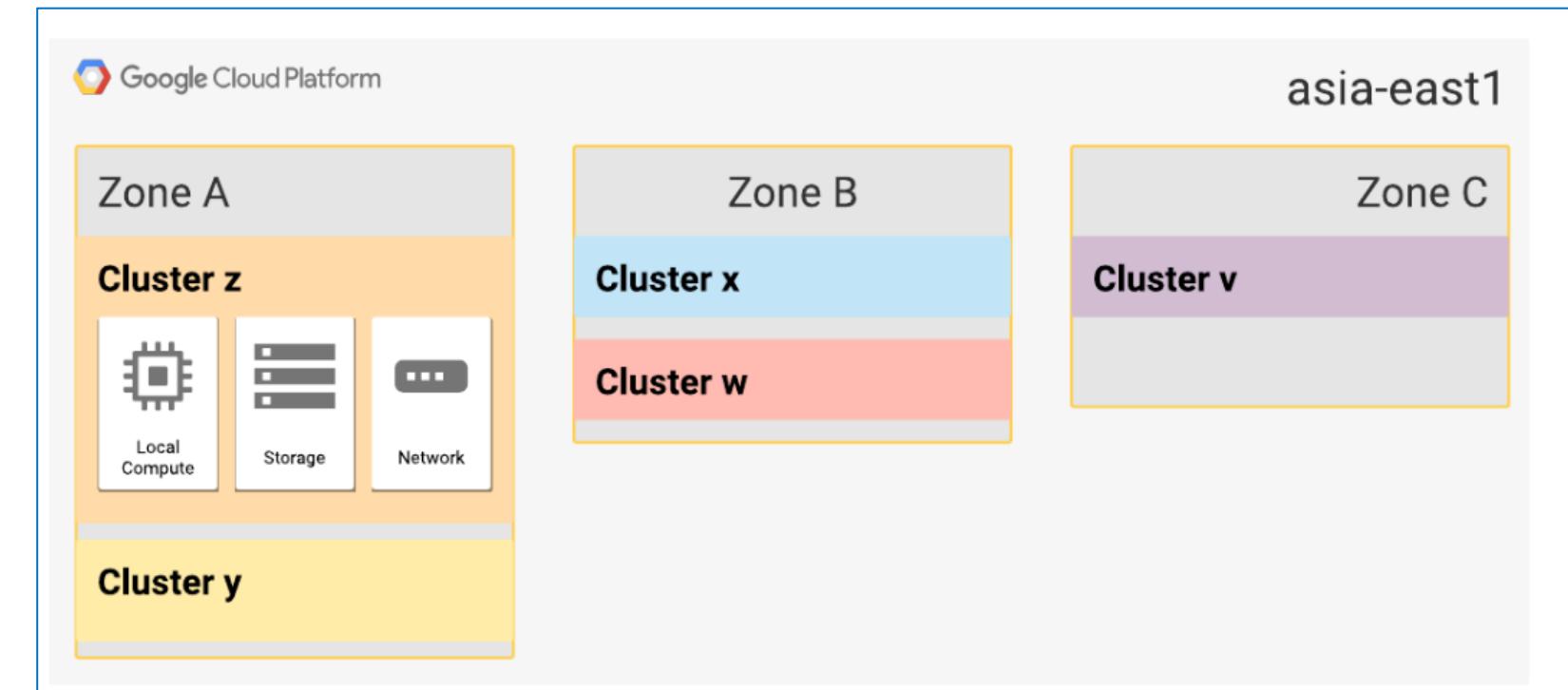
Google Cloud - Clusters in a Zone

- **What are Clusters in a Zone ?**

- In each Zone, we will have **one or more Clusters**
- **Clusters:** A cluster represents a distinct physical infrastructure (compute, network, and storage) that is housed in a data center

- **Understand Zone to Cluster Mapping**

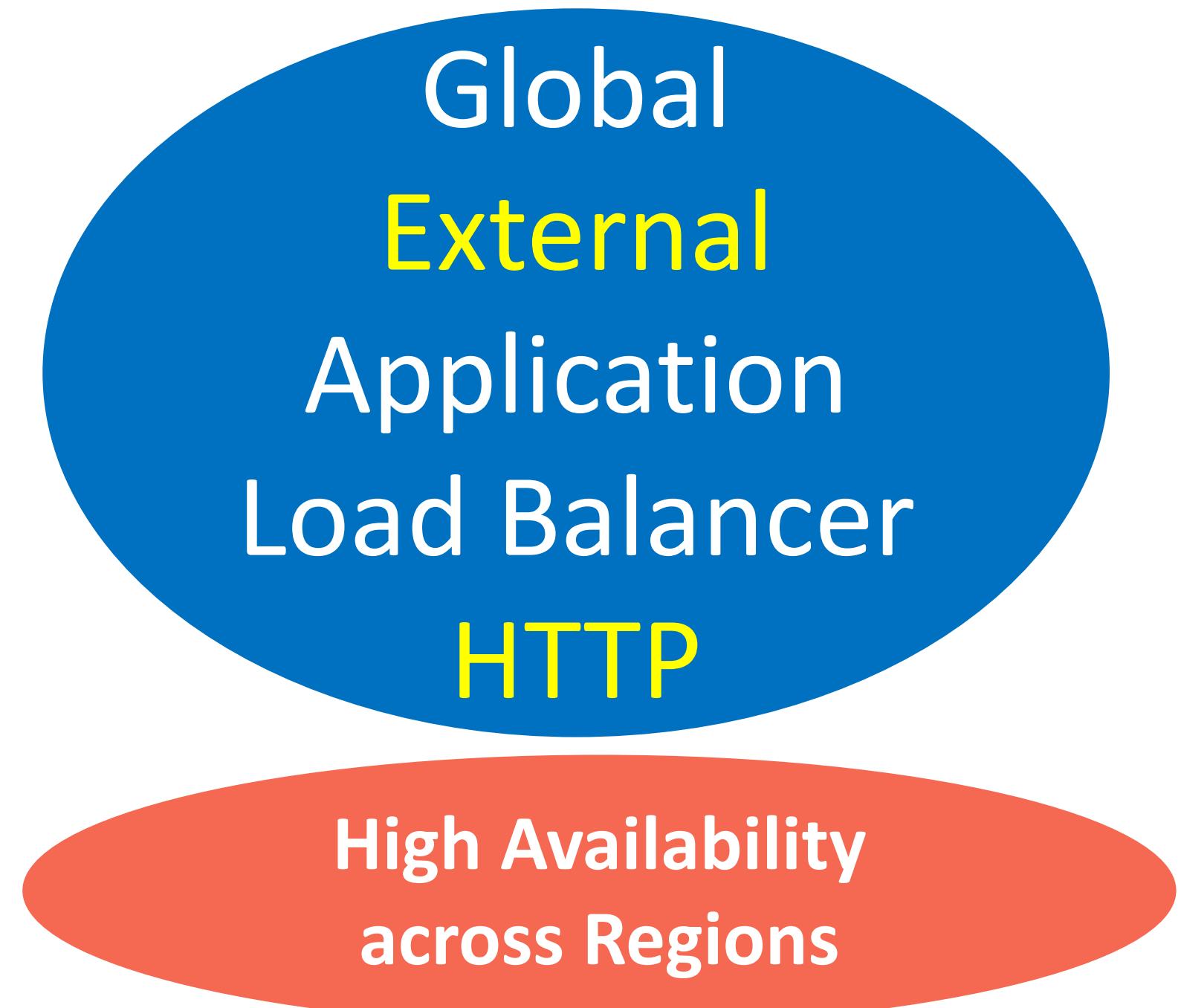
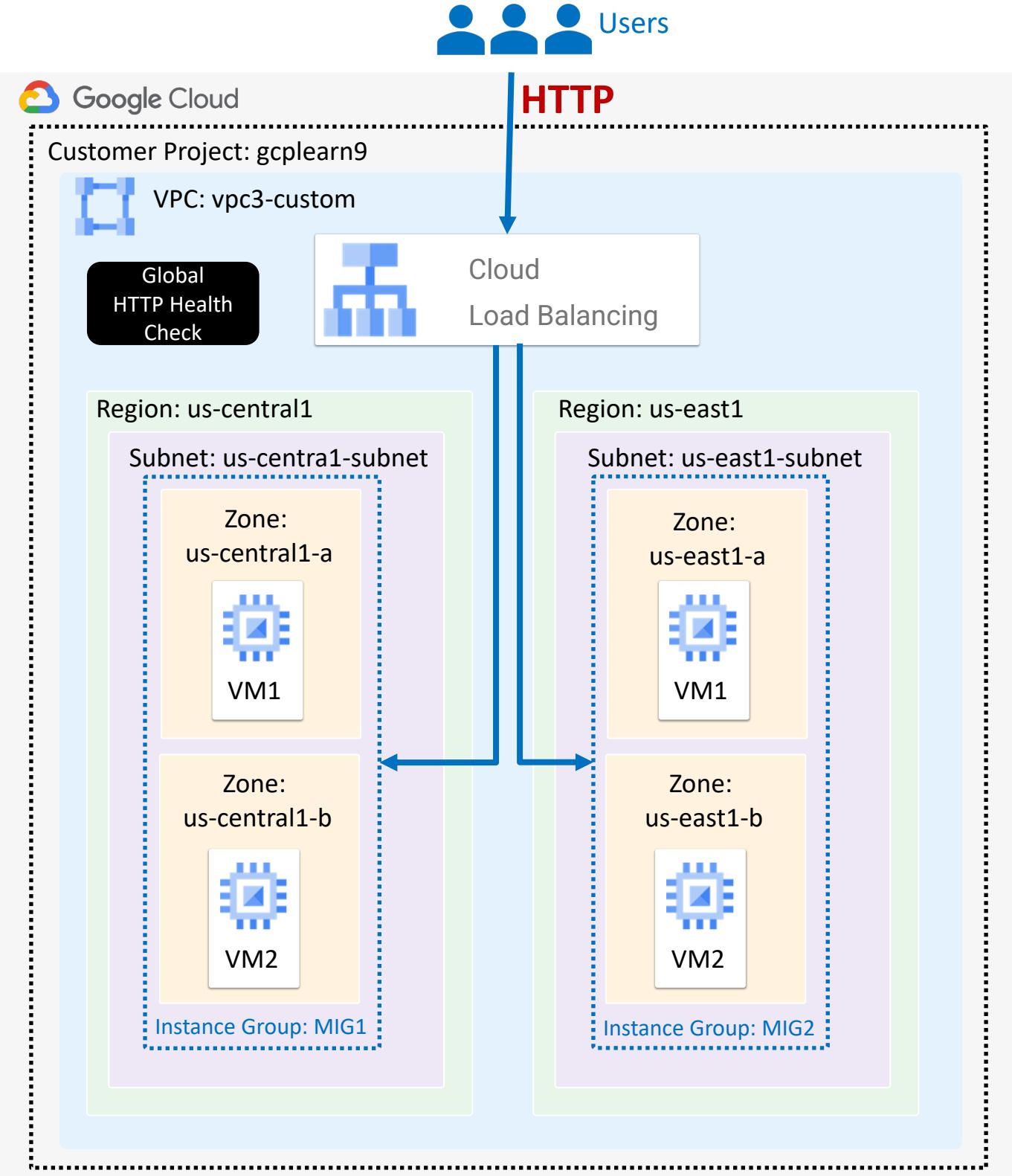
- **Organization A** in **asia-east1-a** zone will have **Cluster Z**
- **Organization B** in **asia-east1-a** zone will have **Cluster Y**
- For most organizations, Compute Engine ensures that all projects in an organization have a **consistent zone to cluster mapping**



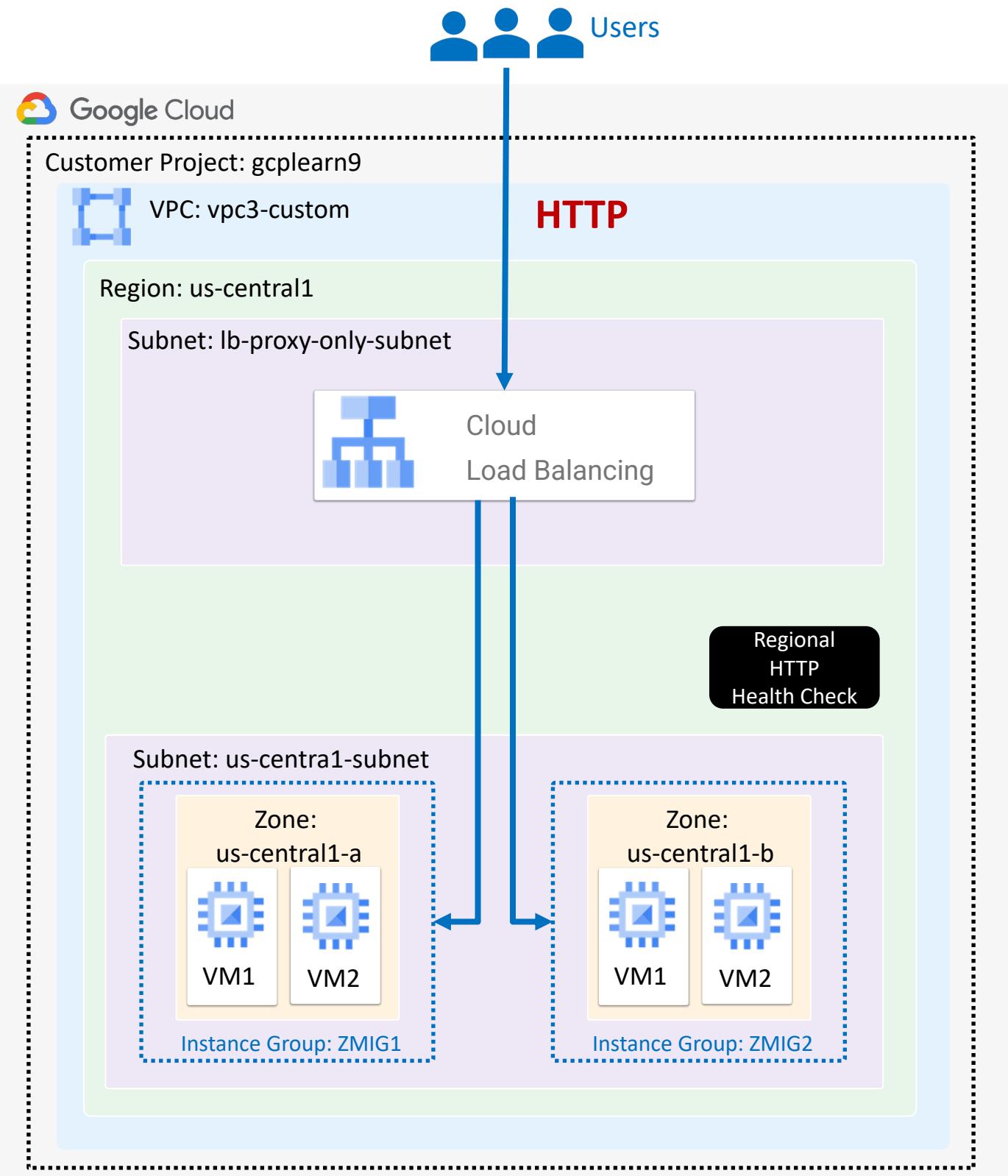
Zone to Cluster Mapping is **not visible** to customers. Google Cloud manages it internally

Reference: <https://cloud.google.com/compute/docs/regions-zones/zone-virtualization>

Cloud Load Balancing



Cloud Load Balancing



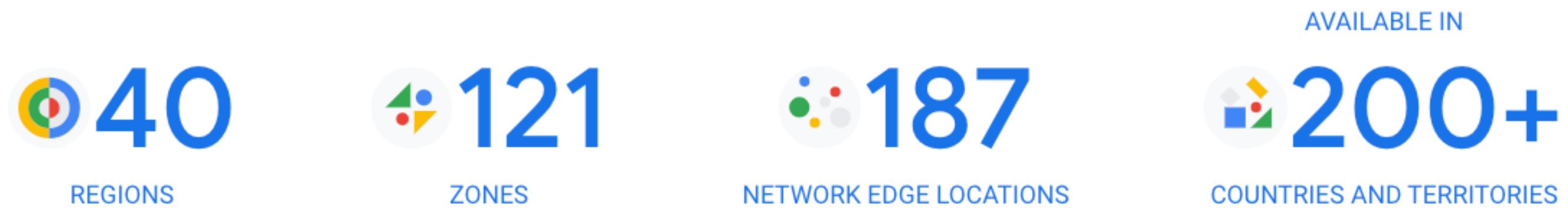
Regional
External
Application
Load Balancer

HTTP

High Availability
in a Regions across
Zones

Google Cloud - Regions & Zones

- Google Cloud keeps adding **more and more regions and zones**.
- It's a **continuous process**
- **As on today** below is the count
- **Global, Regional and Zonal GCP Product Mapping:**
<https://cloud.google.com/about/locations>



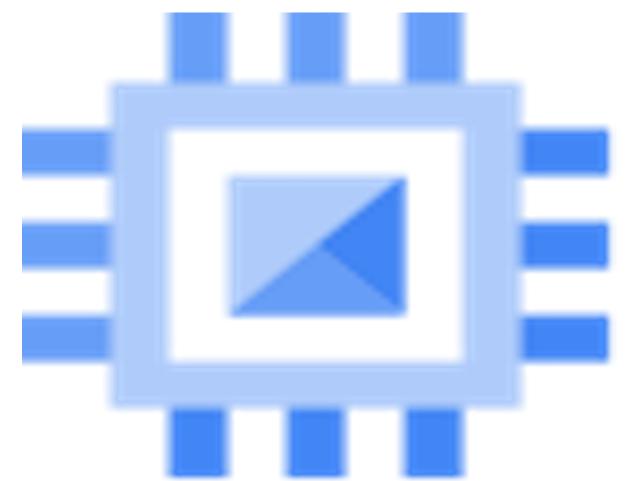
COMING SOON! Google Cloud will continue expanding into the following regions: Mexico, Malaysia, Thailand, New Zealand, Greece, Norway, Austria and Sweden.

Refefence: <https://cloud.google.com/compute/docs/regions-zones/zone-virtualization>

Demo

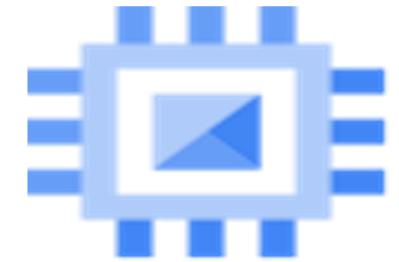


Google Compute Engine Virtual Machine Basics



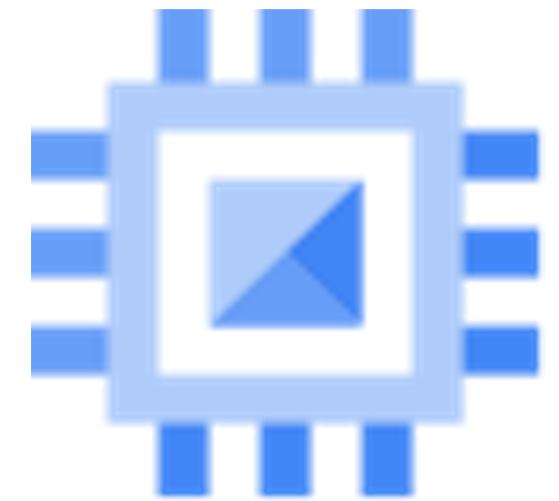
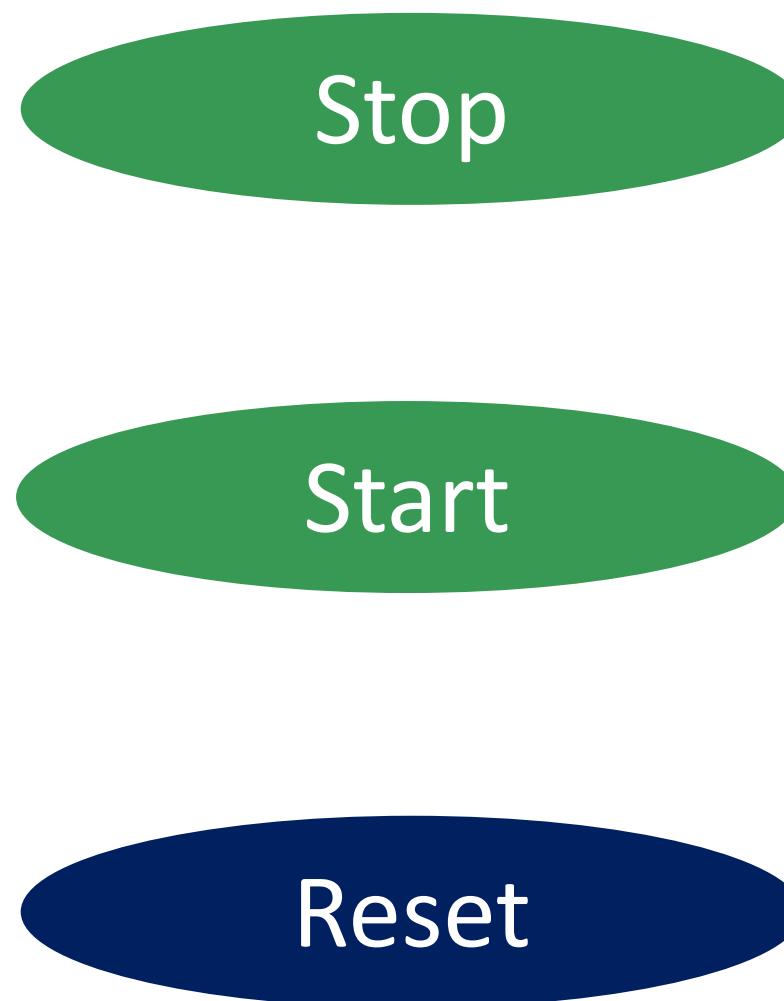
Google Compute Engine - VM Basics

- **Step-01:** Create a **Linux** Virtual Machine
- **Step-02:** Connect to VM using **SSH**
- **Step-03:** Install a simple webserver
- **Step-04:** Access the sample pages via browser
- **Step-05:** Explore VM Actions
 - Stop / Start
 - Suspend / Resume
 - Reset
 - Delete
- **Step-06:** Explore VM
 - Details Tab
 - Observability Tab
 - Screenshot Tab



Google
Compute Engine

Managing VM Instances

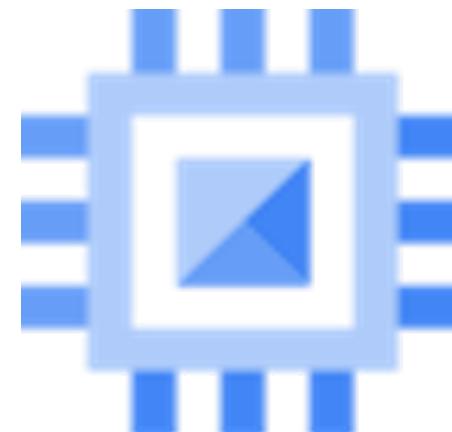


VM Instance

VM Instance Start / Stop



Stop



VM Instance

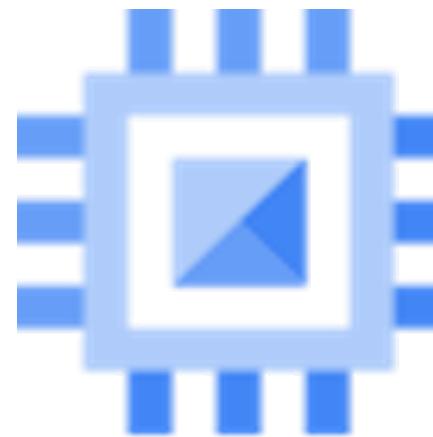


Start

- **What will we lose when we stop the VM ?**
 - External or [Public IP](#) will be released and during the next start [different public IP](#) will be assigned to VM.
- **What will VM retain when we stop the VM ?**
 - Stopped VM retains its [persistent disks](#), [internal IPs](#) and [MAC Address](#).
- **When do you stop the VM instance ?**
 - To [save cost](#) when we don't need that VM in running state
 - To [change](#) VM Instance [Machine Type](#) features
 - To add or remove [disks](#) attached to VM
- **Billing**
 - VMs in the stopped state ([TERMINATED](#)) are [not charged](#)
 - Persistent Disks attached to VM, external IP Address (static) are charged

VM Instance Suspend / Resume

Suspend



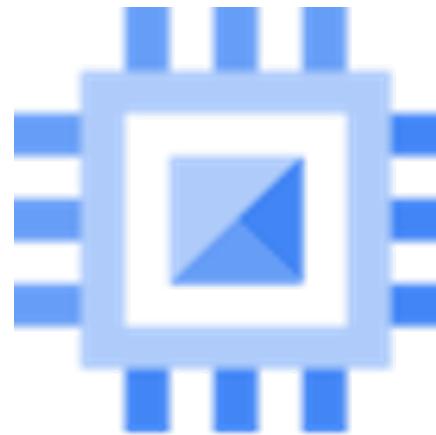
VM Instance

Resume

- Suspended Instances **preserve** the OS Memory, device state and application state
- When you **suspend** your instance, its memory **moves to storage**.
- When you **resume** your instance, its memory **moves from storage to the instance memory**.
- You can only suspend an instance **for up to 60 days** before the VM is **automatically stopped**.
- You cannot suspend instances with more than **120 GB of memory**.
- Most of the OS **support** Suspend and Resume features. Few of the OS, **don't support** these features.

VM Instance Suspend / Resume

Suspend



VM Instance

Resume

- **Usecases**

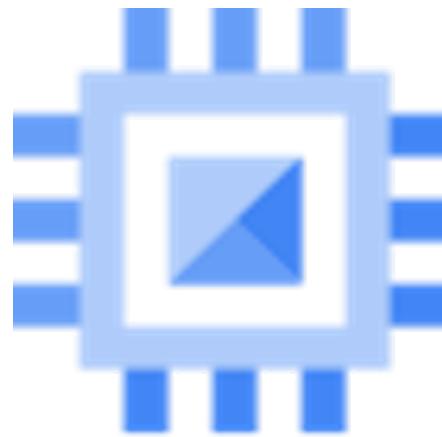
- Dev and Test environments that are not used during non-business hours whose application state to be preserved
- Applications which take long time for initialization when they come online

- **Pricing**

- Google charges for the storage necessary for the instance memory
- Any persistent disk attached the VM Instance
- Any Static IPs attached to the instance

VM Instance Reset / Delete

Reset



VM Instance

Delete

• Reset VM Instance

- Performing a reset on your VM is **similar to doing a hard reset** on your computer.
- Resetting a **VM forcibly wipes the memory contents** of the machine and resets the VM to its initial state.
- The VM **does not perform a clean shutdown** of the guest OS.
- Throughout this process, the VM remains in the **RUNNING** state.

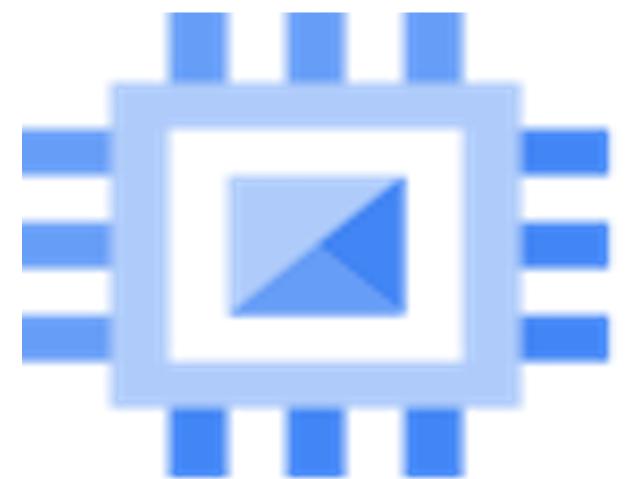
• Delete VM Instance

- Delete an instance to **remove the instance** and the **associated resources** from your project

Demo



Google Compute Engine VM Startup Script



What is a VM Startup Script ?

- A **startup script** is a file that contains commands that run when a virtual machine (VM) instance boots.
- GCE provides support for running startup scripts on **Linux VMs** and **Windows VMs**.
- Startup Scripts can be configured at **VM-level** and **Project-level**
- Startup Scripts configured at VM-level will **override** the Project-level startup scripts
- **Why do we need to configure Project-level startup scripts ?**
 - If we want all the VM's in that respective project need to have **same scripts applied** (install antivirus package) then we can define **project level startup scripts** which applies to all VM's

Startup Script

```
#!/bin/bash
sudo apt install -y telnet
sudo apt install -y nginx
sudo systemctl enable nginx
sudo chmod -R 755 /var/www/html
HOSTNAME=$(hostname)
sudo echo "<!DOCTYPE html> <html> <body style='background-color:rgb(250, 210, 210);'> <h1>Welcome to StackSimplify - WebVM App1</h1> <p><strong>VM Hostname:</strong> $HOSTNAME</p> <p><strong>VM IP Address:</strong> $(hostname -l)</p> <p><strong>Application Version:</strong> V1</p> <p>Google Cloud Platform - Demos</p> </body></html>" | sudo tee /var/www/html/index.html
```

Linux Startup Script

- **Linux Startup Script Types**

- Bash and Non-Bash
- To use a **non-bash** file, designate the interpreter by adding a **#!** to the top of the file
- Example: **#! /usr/bin/python3**

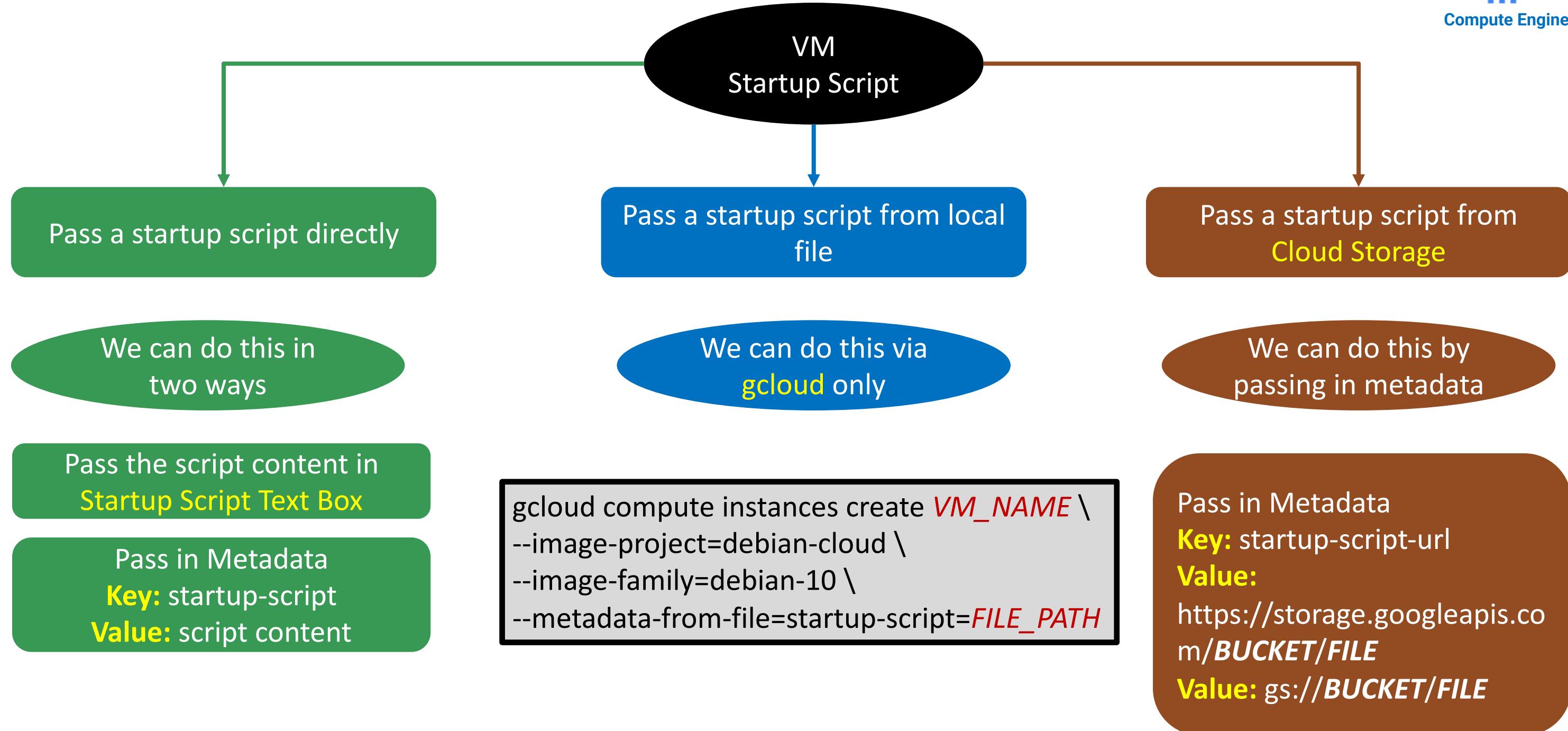
- **What happens when we provide Startup Script to a VM ?**

- Compute Engine **copies** startup script to the VM
- Sets **run** permissions on the startup script
- Runs the startup script as the **root user** when the VM boots

Startup Script

```
#!/bin/bash
sudo apt install -y telnet
sudo apt install -y nginx
sudo systemctl enable nginx
sudo chmod -R 755 /var/www/html
HOSTNAME=$(hostname)
sudo echo "<!DOCTYPE html> <html> <body style='background-color:rgb(250, 210, 210);'> <h1>Welcome to StackSimplify - WebVM App1</h1> <p><strong>VM Hostname:</strong> $HOSTNAME</p> <p><strong>VM IP Address:</strong> $(hostname -l)</p> <p><strong>Application Version:</strong> V1</p> <p>Google Cloud Platform - Demos</p> </body></html>" | sudo tee /var/www/html/index.html
```

Google Compute Engine – VM Startup Script



VM Startup Scripts

- How many startup scripts we can use per VM ?
 - We can use **multiple** startup scripts
- What is the Order of Execution for Startup Scripts ?
 - **Metadata-Key:** [startup-script](#)
 - Startup scripts provided directly or locally will take the **first preference** in order of execution.
 - **Metadata-Key:** [startup-script-url](#)
 - Startup scripts provided in Cloud Storage will take the **second preference** in order of execution
- What is the startup script maximum allowed size ?
 - Scripts up to **256 KB** are allowed

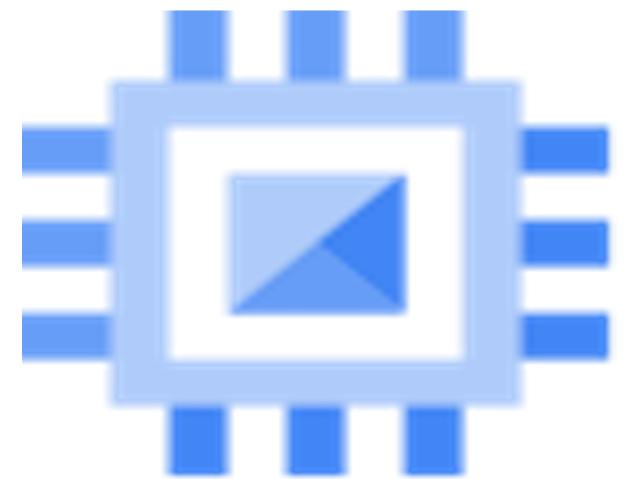
Startup Script

```
#!/bin/bash
sudo apt install -y telnet
sudo apt install -y nginx
sudo systemctl enable nginx
sudo chmod -R 755 /var/www/html
HOSTNAME=$(hostname)
sudo echo "<!DOCTYPE html> <html> <body style='background-
color:rgb(250, 210, 210);'> <h1>Welcome to StackSimplify - WebVM App1
</h1> <p><strong>VM Hostname:</strong> $HOSTNAME</p>
<p><strong>VM IP Address:</strong> $(hostname -l)</p>
<p><strong>Application Version:</strong> V1</p> <p>Google Cloud
Platform - Demos</p> </body></html>" | sudo tee
/var/www/html/index.html
```

Demo

Google Cloud

Cloud Shell



Cloud Shell

- Cloud Shell is an **online development** and **operations** environment
- Accessible anywhere with your **browser**
- Provides **command-line access** to a virtual machine instance in a terminal window
- Favourite command-line tools **pre-installed**, **put up-to-date and ready to use**. Just connect and use
 - Bash, sh, emacs, vim
 - **gcloud**, MySQL, Kubernetes CLI, Docker CLI, minikube and many more
- 5 GB of **persistent disk storage** for \$HOME directory
- **Online Code Editor:** Develop, build, debug, and test your apps anywhere using the Cloud Shell Editor.
- **Source control via Git:** Clone or pull remote repositories or commit your code changes back to your repo via Cloud Shell's git client

Additional Reference: <https://cloud.google.com/shell>

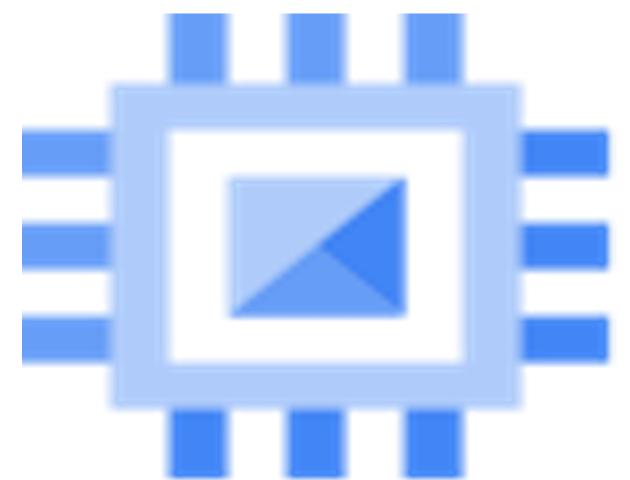
Demo



Google Compute Engine

gcloud Command-line Tool

Cloud SDK



Cloud SDK - gcloud Command-line Tool

- Many **GCP services** can be managed using gcloud cli
 - Google Compute Engine VMs and other resources
 - Cloud SQL Instances, Cloud Functions, Cloud Run
 - GKE Clusters
 - Dataproc Clusters
 - Cloud DNS managed zones and record sets
 - Cloud Deployment Manager deployments
- **Few** of the GCP services have their own CLI Tools
 - Cloud Bigtable - cbt
 - Cloud BigQuery - bq
 - Cloud Storage – gsutil (old), **gcloud is recommended** going forward
 - Kubernetes – kubectl (kubectl helps in managing resources in Kubernetes cluster)

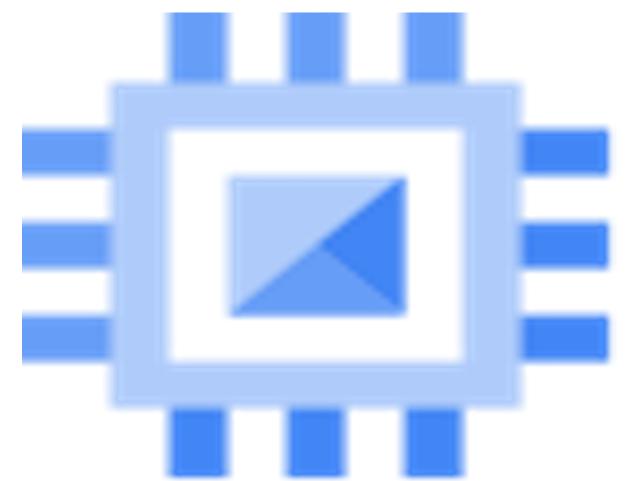
Cloud SDK - gcloud Command-line Tool

- gcloud cli can be used with two options
- **Option-1: Install gcloud in your local desktop**
 - gcloud cli is part of Cloud SDK
 - **Install Cloud SDK:** <https://cloud.google.com/sdk/docs/install>
- **Option-2: Use gcloud from Cloud Shell**
 - gcloud cli is by default installed and ready to use on Cloud Shell
 - We will choose **option-2** for now

Demo

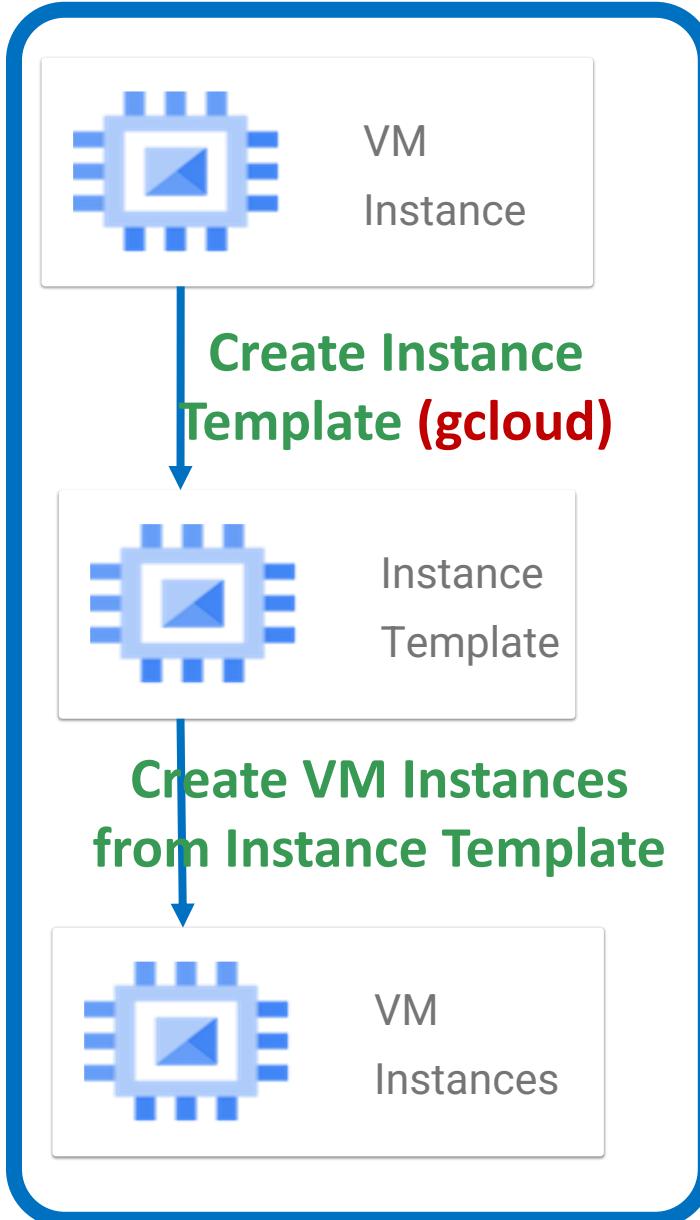


Google Compute Engine Instance Templates



Compute Engine - Instance Template

- **What is an Instance Template ?**
- Instance Templates is used to quickly
 - Create VM Instances
 - Create Managed Instance Groups (MIG – discussed in later demos)
 - Create reservations for VMs
 - Create future reservations for VMs
- Using Instance Templates, we **save** VM configuration information like
 - Machine Configuration, Boot Disk or Container Image, Identity and API Access
 - Firewall, Management, Security, Disks, Networking and Sole-Tenancy
- **Why do we need Instance Templates or When to use them?**
 - To create VMs with **Identical Configurations**
 - In short, define all settings in Instance Templates **once** and **quickly** reference them to **create VM's or groups of VM's (MIG)**.



Compute Engine - Instance Template

- **How to update Instance Templates ?**
 - We **cannot** update an existing Instance Template
 - We can **create the clone** of existing Instance Template using **CREATE SIMILAR** Option and add additional configs to new Instance Template.
- **Can we override configuration settings from Instance Template when creating a VM Instance?**
 - Yes, we can do that.
- **How many ways we can create Instance Templates ?**
 - Create an Instance Template from **scratch**
 - Create an Instance Template from **existing Instance Template** using **CREATE SIMILAR** Option
 - Create Instance Template based on **existing VM Instance**
 - This is possible only using **gcloud command or API**
 - We **cannot** create an Instance Template based on existing VM Instance using **Google Cloud Console**

Compute Engine - Instance Templates

- **What are Deterministic Instance Templates ?**
 - Be explicitly clear about **third-party software versions** used in Startup-Scripts
 - Example: Nginx specific version we want to install using startup-script
 - By creating deterministic instance templates, you minimize **ambiguity and unexpected behaviour** from your instance templates
- **Can we create regional and global instance templates ?**
 - Yes, we can create **regional and global** instance templates

Compute Engine - Instance Templates

	Regional Instance Template	Global Instance Template
Scope	Can be used in specific region only	Can be used in any region
Use case	<ol style="list-style-type: none"> 1. Reduces cross-region dependency 2. Achieves data residency in a region for compliance requirements 3. Example: To meet physical data location requirements confined to specific region 	<ol style="list-style-type: none"> 1. We can reuse the instance templates to create VMs, MIGs and reservations in different regions 2. Define once, use globally in any region

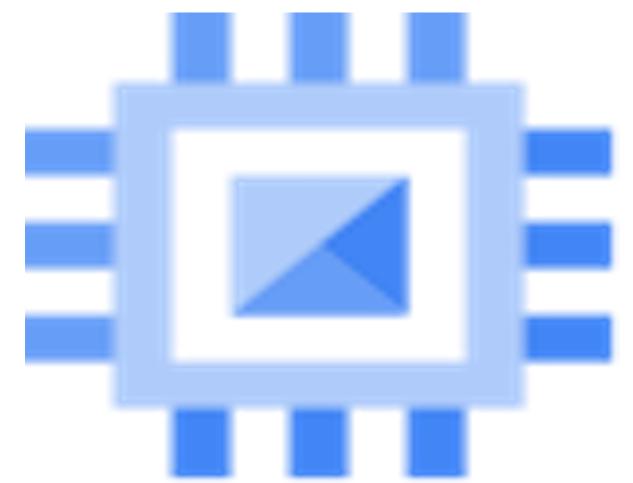
Compute Engine - Instance Templates

- **Can Instance Template be locked to a zone even if the template is a regional or Global template?**
- Yes, if you specify a [zonal resource](#) in your [global](#) or [regional](#) template (example: read-only persistent disk from a zone) where the resource resides, then that [restricts](#) the template to that respective zone only.
- **Can Instance Template be locked to a region even if the template is a Global template?**
- Yes, if you specify a [regional resource](#) in a [global template](#), template is restricted to that region only

Demo



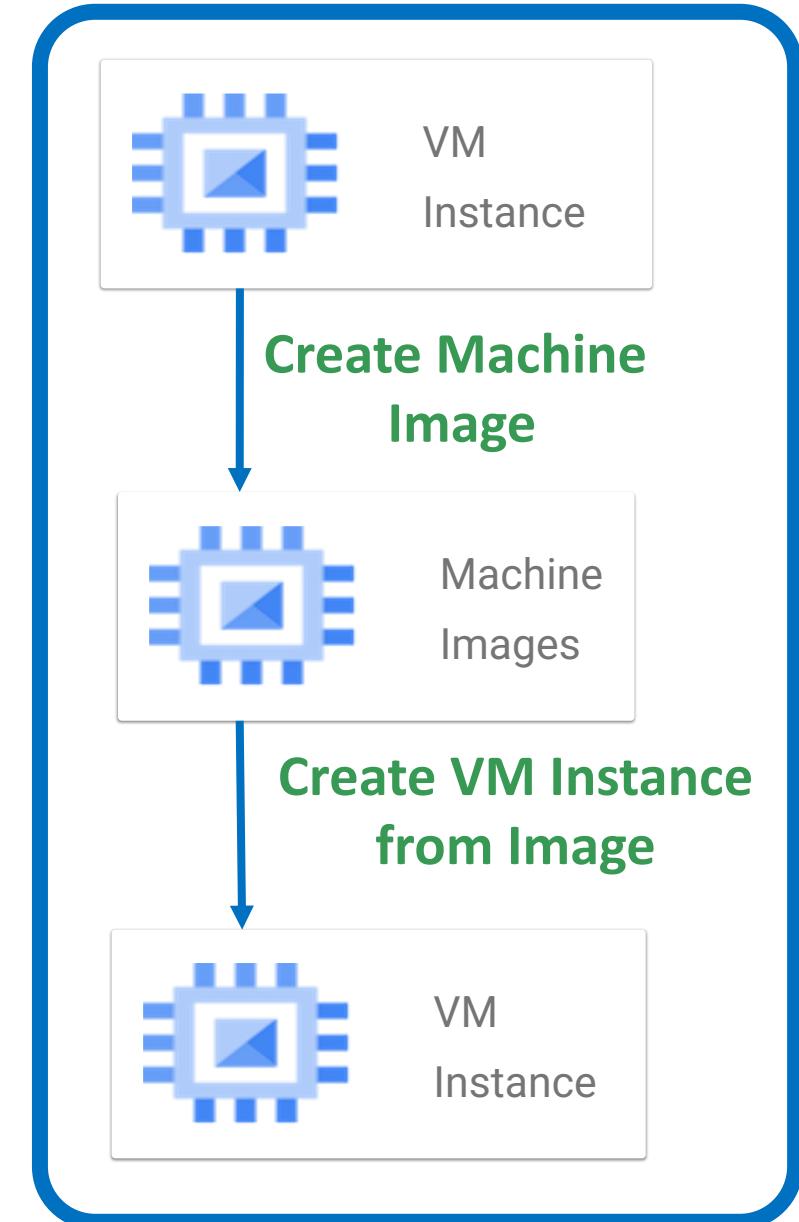
Google Compute Engine Machine Images



Google Compute Engine - Machine Images

- What is a Machine Image ?
- A machine image **contains** a VM's properties, metadata, permissions, and **data** from all its attached disks which helps in creating a new VM Instance.
- You can use a machine image in many **system maintenance scenarios**, such as
 - Instance Creation
 - Backup
 - Recovery
 - Instance Cloning

 Key point here is all **VM configuration** is persisted during Machine Image creation and can be used when creating a VM from MI



Google Compute Engine - Machine Images

- A machine image is **unchangeable** (cannot edit).
- We can **override** almost all the properties of the machine image when **creating a VM instance** from the machine image.
- We can create VM Instances from Machine Images from all 3 options
 - Google Cloud **Console**
 - **gcloud** Command-line Tool
 - Compute Engine **REST API**
- **Where are the Machine Images stored ?**
- We can store machine images in a desired **region or multi-region** (accessible across globe) to meet compliance needs
- They are stored in **GCP managed Cloud Storage buckets** (Buckets not visible to us in our account)
- **Easy Extendibility:** Attach **GPUs and local SSDs** to Spot instances for additional performance and saving

Google Compute Engine - Machine Images

- What all information from the source instance **is collected by machine image?**
 - VM [Instance Configuration](#)
 - VM Description
 - Machine Type
 - Instance Metadata
 - Labels
 - Network Tags
 - Maintenance Policy
 - [Volume Mapping](#) used to create Persistent Disks and local SSDs
 - Data stored on persistent disks at [consistent points in time](#) across disks
- What all information from the source instance **is not collected by machine image?**
 - Data [in memory](#)
 - Data [in local SSD](#). However, a machine image captures the device mapping of local SSDs.
 - Attributes that are specific to the source instance, such as the [name or IP address](#)

Google Compute Engine - Machine Images

- When to use Machine Images ?
- Machine Images are so **powerful** that it is supported in all scenarios related to **backup, cloning, replication and configuration**

Scenarios	Machine image	Persistent disk snapshot	Custom image	Instance template
Single disk backup	Yes	Yes	Yes	No
Multiple disk backup	Yes	No	No	No
Differential backup	Yes	Yes	No	No
Instance cloning	Yes	No	Yes	Yes
Base image for replication	No	No	Yes	No

Upcoming Demos

Reducing Launch Time with Custom Machine Images



VM Instance

- **Step-1:** Launch VM Instance
- **Step-2:** Install OS Patches using startup-script
- **Step-3:** Install Application Software using startup-script
- **Step-4:** Start Application

With **start-up script**, launch time of a VM is very **high**

VM Instance from Custom Machine Image

- **Step-1:** Launch VM Instance from Custom Machine Image
- **Step-2:** Start Application

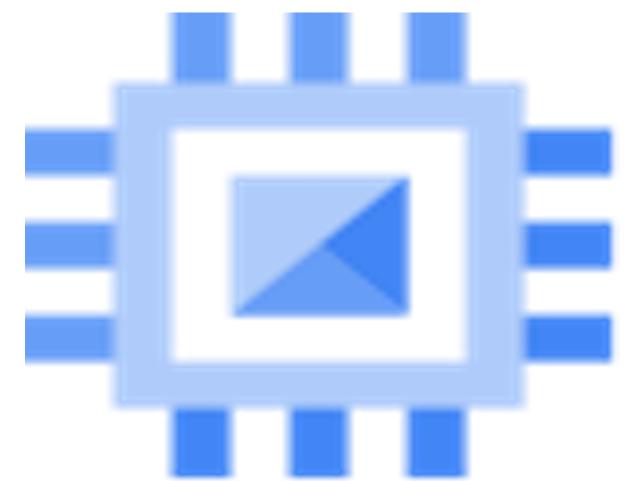
With **custom Machine Image**, launch time of a VM is very very **less**

PREFER using Machine Image over Startup Script to reduce VM launch times

Concept



Google Compute Engine Live Migration & Availability Policy



GCE - Live Migration & Availability Policy

Who will take care of software updates to our VM Instances ?

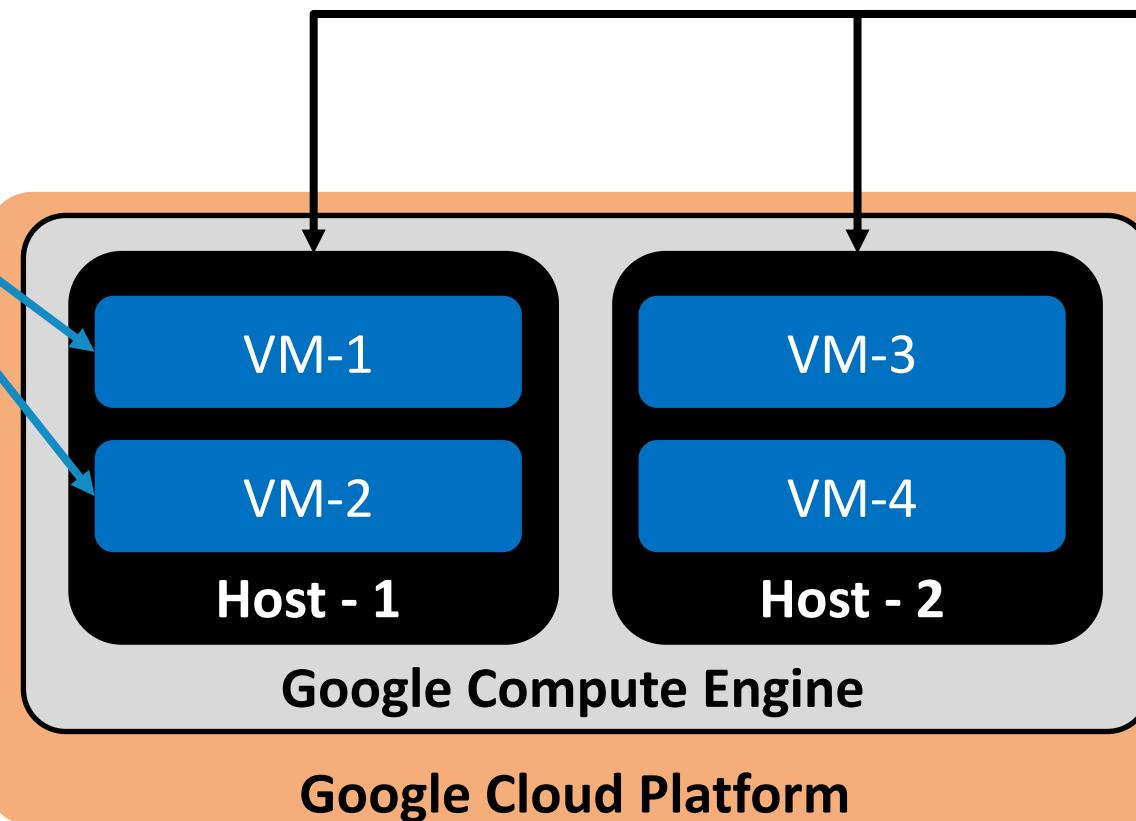
Our admin teams take care of those with **VM Manager** which contains OS Patch Management and OS Configuration Management features

VM Manager

Patch

OS policies

Our Teams



Who will take care of software and hardware updates to HOST MACHINES in Cloud Platform ?

GCP Teams

How to ensure our VM Instances are up and running when the **underlying HOST SYSTEMS** are undergoing HARDWARE or SOFTWARE upgrades ?

Live Migration & Availability Policy settings help us to ensure no impact to our VM Instances when HOST SYSTEMS are undergoing maintenance in GCP

GCE - Live Migration & Availability Policy

- How does LIVE MIGRATION help ?
- LIVE MIGRATION keeps your instances running during
 - Regular Infrastructure maintenance and upgrades
 - Network and power grid maintenance
 - IN SHORT, for all GCP maintenance events
- What happens with LIVE MIGRATION ?
 - Your running VM Instances are migrated from HOST which is prepared for undergoing maintenance to other HOST in same ZONE.
 - NO changes to VM settings or metadata
- Supported for instances with Local SSDs
- Not supported for GPUs, Spot VMs and preemptible VMs

Availability policies

VM provisioning model

Standard

Choose "Spot" to get a discounted, preemptible VM. Otherwise, stick to "Standard". [Learn more](#) 

Set a time limit for the VM 

On VM termination

Choose what happens to your VM when it's preempted or reaches its time limit

On host maintenance

Migrate VM instance (Recommended)

When Compute Engine performs periodic infrastructure maintenance it can migrate your VM instances to other hardware without downtime

Automatic restart

On (recommended)

Compute Engine can automatically restart VM instances if they are terminated

GCE - Live Migration & Availability Policy

Availability policies

VM provisioning model

Standard

Choose "Spot" to get a discounted, preemptible VM. Otherwise, stick to "Standard". [Learn more](#)

Set a time limit for the VM ?

On VM termination

Choose what happens to your VM when it's preempted or reaches its time limit

On host maintenance

Migrate VM instance (Recommended)

Terminate VM instance

Automatic restart

On (recommended)

Compute Engine can automatically restart VM instances if they are terminated

Availability policies

VM provisioning model

Standard

Choose "Spot" to get a discounted, preemptible VM. Otherwise, stick to "Standard". [Learn more](#)

Set a time limit for the VM ?

On VM termination

Choose what happens to your VM when it's preempted or reaches its time limit

On host maintenance

Migrate VM instance (Recommended)

When Compute Engine performs periodic infrastructure maintenance it can migrate your VM instances to other hardware without downtime

Automatic restart

On (recommended)

Off

On host maintenance: What should happen during regular maintenance ?

Migrate VM Instance (RECOMMENDED OPTION)

Terminate VM Instance (Stop VM)

Automatic Restart : What should happen if VMs are terminated for non-user-initiated reasons ?

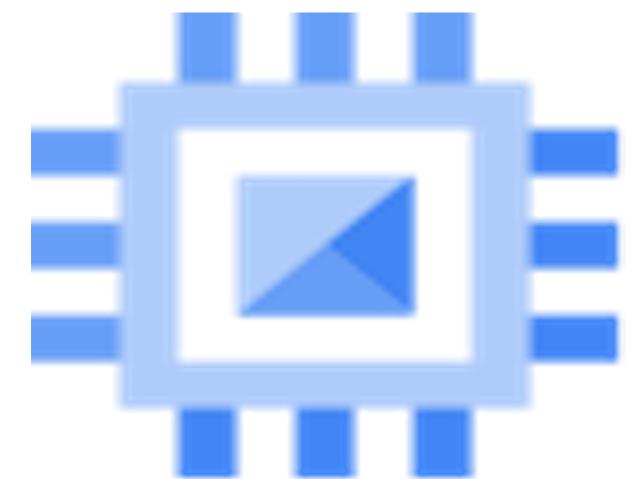
Automatic Restart VMs: **ON** (RECOMMENDED Option)

Automatic Restart VMs: **OFF**

Demo



Google Compute Engine



Temporary VM Instances

Spot VMs and Preemptible VMs

Google Compute Engine - Spot VMs

- What is a Spot VM Instance ?
- Spot VMs are excess Compute Engine capacity
- Spot VMs are available at much lower price **60-91% discount** compared to the price of standard VMs.
- Why are they coming at very low cost ?
- Compute Engine might pre-emptively stop or delete (preempt) these instances if it needs to reclaim those resources for other tasks
 - Stop VM (Default Option): Can be restarted, memory is not preserved
 - Delete VM: Permanently deleted, memory is not preserved
- VM Instances get 30 second warning before it gets preempted by compute engine (Just to save anything on that VM)

Availability policies

VM provisioning model
Spot

Choose "Spot" to get a discounted, preemptible VM. Otherwise, stick to "Standard". [Learn more](#) 

Set a time limit for the VM 

On VM termination

- Stop**
Can be restarted, memory is not preserved
- Delete**
Permanently deleted, memory is not preserved

When Compute Engine performs periodic infrastructure maintenance it can migrate your VM instances to other hardware without downtime

Automatic restart

- Off**

Compute Engine can automatically restart VM instances if they are terminated

Google Compute Engine - Spot VMs

- **What is Fault-tolerance design?**
 - A fault-tolerant design lets a **system keep going**, even if something breaks, instead of crashing completely.
- **At what scenarios we can use Spot VMs ?**
 - We can use Spot VMs for applications that are **fault-tolerant**
 - If your **budgets are cost-sensitive** we can use Spot VMs where we get **60-91% discounts** when compared standard VMs
- **Limitations**
 - Compute Engine might preempt Spot VMs to **reclaim the resources at any time**.
 - Can't **live migrate** to a regular VM instance, it will terminate the VM
 - Can't set to **automatically restart** when there is a maintenance event.
 - NO **SLA**
 - Spot VMs are **finite** Compute Engine resources, so they **might not always be available**
 - The **Google Cloud Free Tier credits** for Compute Engine do not apply to Spot VMs.

Google Compute Engine - Spot VMs

- **Usecases: Regular Workloads**
 - High performance computing
 - Big data and analytics
 - Continuous integration/continuous delivery (CI/CD) Pipelines
 - Batch processing workloads
- **Usecases - Container Workloads**
 - You can use Spot VMs to create [Google Kubernetes Engine Node Pools](#)

VM Provisioning Models

Standard VM

Availability policies

VM provisioning model
Standard

Choose "Spot" to get a discounted, preemptible VM. Otherwise, stick to "Standard". [Learn more](#)

Set a time limit for the VM ?

On VM termination

Choose what happens to your VM when it's preempted or reaches its time limit

On host maintenance

Migrate VM instance (Recommended)

When Compute Engine performs periodic infrastructure maintenance it can migrate your VM instances to other hardware without downtime

Automatic restart

On (recommended)

Compute Engine can automatically restart VM instances if they are terminated

Spot VM

Availability policies

VM provisioning model
Spot

Choose "Spot" to get a discounted, preemptible VM. Otherwise, stick to "Standard". [Learn more](#)

Set a time limit for the VM ?

On VM termination

Stop

Can be restarted, memory is not preserved

Delete

Permanently deleted, memory is not preserved

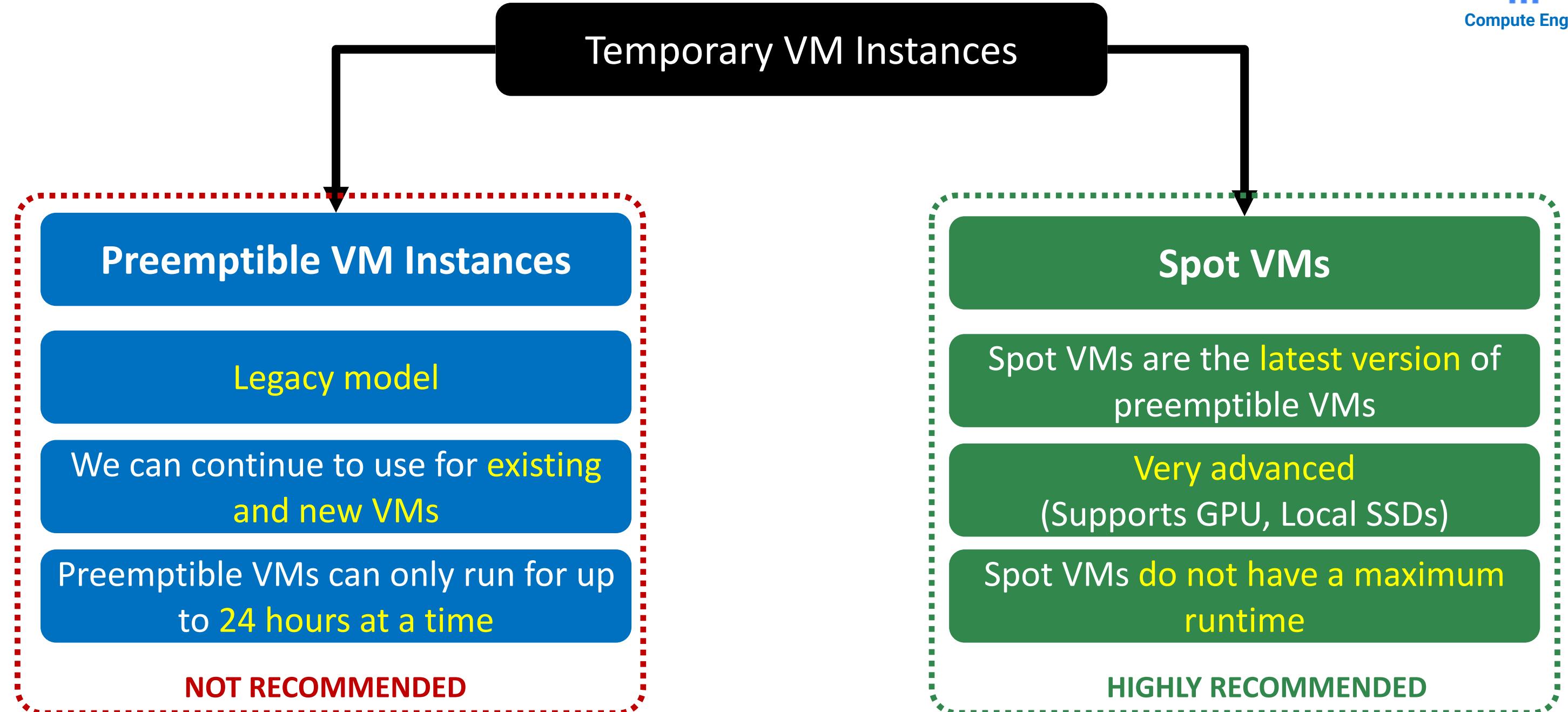
When Compute Engine performs periodic infrastructure maintenance it can migrate your VM instances to other hardware without downtime

Automatic restart

Off

Compute Engine can automatically restart VM instances if they are terminated

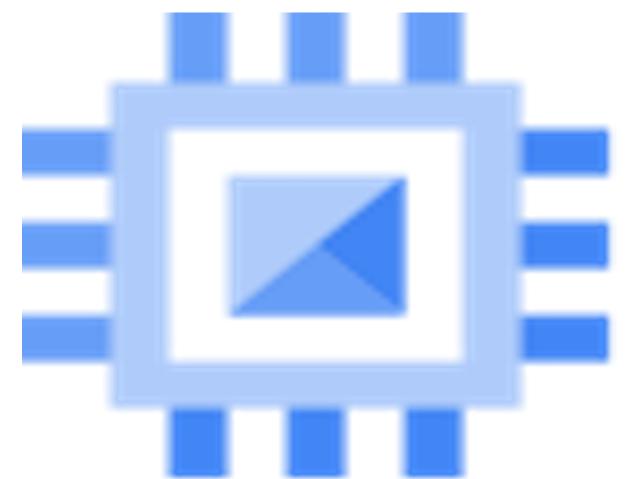
Temporary VM Instances



Concept



Google Compute Engine Machine Families & Types



Choosing the right virtual machine type



Reference: <https://cloud.google.com/compute/>

Google Compute Engine Machine Families

General-purpose workloads			
E2	N2, N2D, N1	C3, C3D	Tau T2D, Tau T2A
Day-to-day computing at a lower cost	Balanced price/performance across a wide range of machine types	Consistently high performance for a variety of workloads	Best per-core performance/cost for scale-out workloads
<ul style="list-style-type: none">• Low-traffic web servers• Back office apps• Containerized microservices• Microservices• Virtual desktops• Development and test environments	<ul style="list-style-type: none">• Low to medium traffic web and app servers• Containerized microservices• Business intelligence apps• Virtual desktops• CRM applications• Data Pipelines	<ul style="list-style-type: none">• High traffic web and app servers• Databases• In-memory caches• Ad servers• Game Servers• Data analytics• Media streaming and transcoding• CPU-based ML training and inference	<ul style="list-style-type: none">• Scale-out workloads• Web serving• Containerized microservices• Media transcoding• Large-scale Java applications

Reference: <https://cloud.google.com/compute/docs/machine-resource>

Optimized workloads			
Storage-optimized	Compute-optimized	Memory-optimized	Accelerator-optimized
Z3 (Preview)	H3, C2, C2D	M3, M2, M1	A3, A2, G2
Highest block storage to compute ratios for storage-intensive workloads	Ultra high performance for compute-intensive workloads	Highest memory to compute ratios for memory-intensive workloads	Optimized for accelerated high performance computing workloads
<ul style="list-style-type: none"> • File servers • Flash-optimized databases • Scale-out analytics • Other databases 	<ul style="list-style-type: none"> • Compute-bound workloads • High-performance web servers • Game Servers • High performance computing (HPC) • Media transcoding • Modeling and simulation workloads • AI/ML 	<ul style="list-style-type: none"> • Medium to extra-large SAP HANA in-memory databases • In-memory data stores, such as Redis • Simulation • High Performance databases such as Microsoft SQL Server, MySQL • Electronic design automation 	<ul style="list-style-type: none"> • Generative AI models such as the following: <ul style="list-style-type: none"> • Large Language Models (LLM) • Diffusion Models • Generative Adversarial Networks (GAN) • CUDA-enabled ML training and inference • High-performance computing (HPC) • Massively parallelized computation • BERT natural language processing • Deep learning recommendation model (DLRM) • Video transcoding • Remote visualization workstation

Google Compute Engine Machine Families

Reference: <https://cloud.google.com/compute/docs/machine-resource>

Google Compute Engine - Machine Types

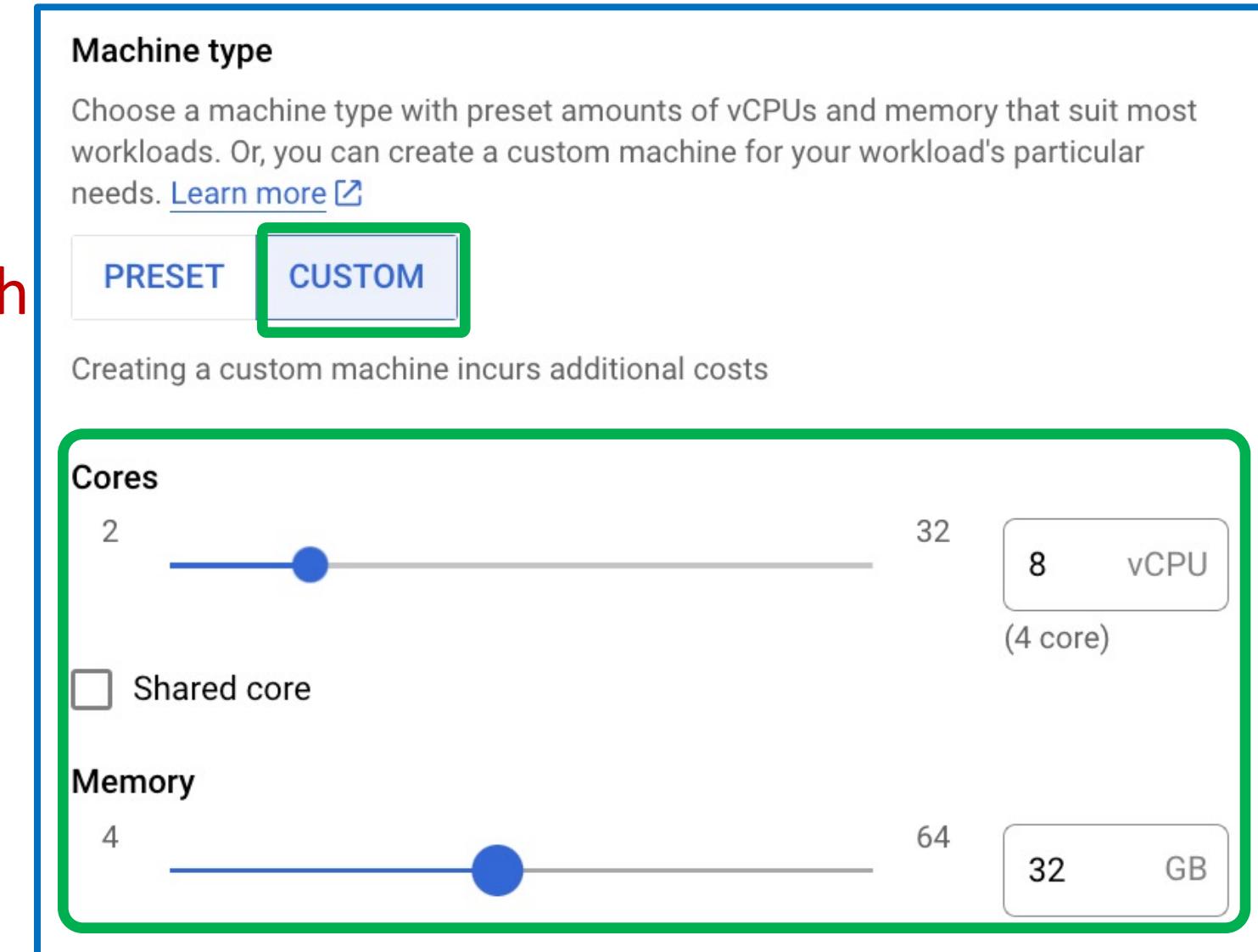
Machine types	vCPUs*	Memory (GB)	Max number of persistent disks (PDs) [†]	Max total PD size (TB)	Local SSD	Maximum egress bandwidth (Gbps) [‡]
e2-standard-2	2	8	128	257	No	4
e2-standard-4	4	16	128	257	No	8
e2-standard-8	8	32	128	257	No	16
e2-standard-16	16	64	128	257	No	16
e2-standard-32	32	128	128	257	No	16

- We can choose machine type based on **CPU**, **Memory** and **Disk** needed for us.
- As number of **vCPU's increases** accordingly memory, disk and networking sizes increase

Google Compute Engine – Customized Machine Type



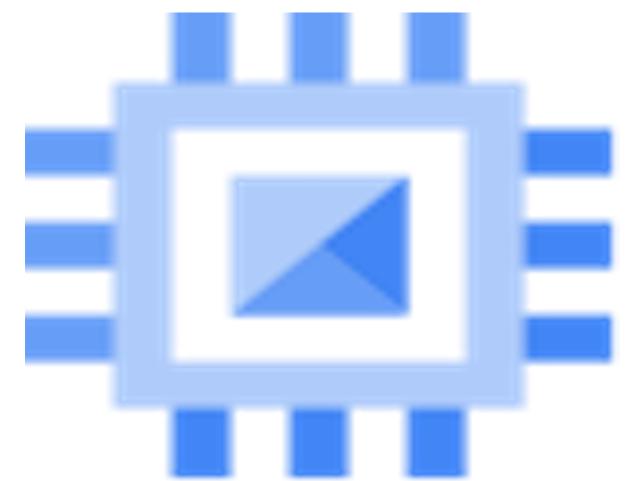
- **Custom Machine Types:** We can configure our **Customized Machine Type**
 - Desired vCPU Cores
 - Desired Memory
 - Desired GPU
- If predefined machine types **doesn't match our workload needs** we can create a VM with **custom machine type**
- This feature is available for **specific Machine families only**
 - **General Purpose:** E2, E2 shared-core, N2, N2D, N1
- Billed per **vCPUs and memory** provisioned



Demo



Google Compute Engine GPUs



Google Compute Engine - GPUs

- What is GPU ?

- Graphics Processing Unit

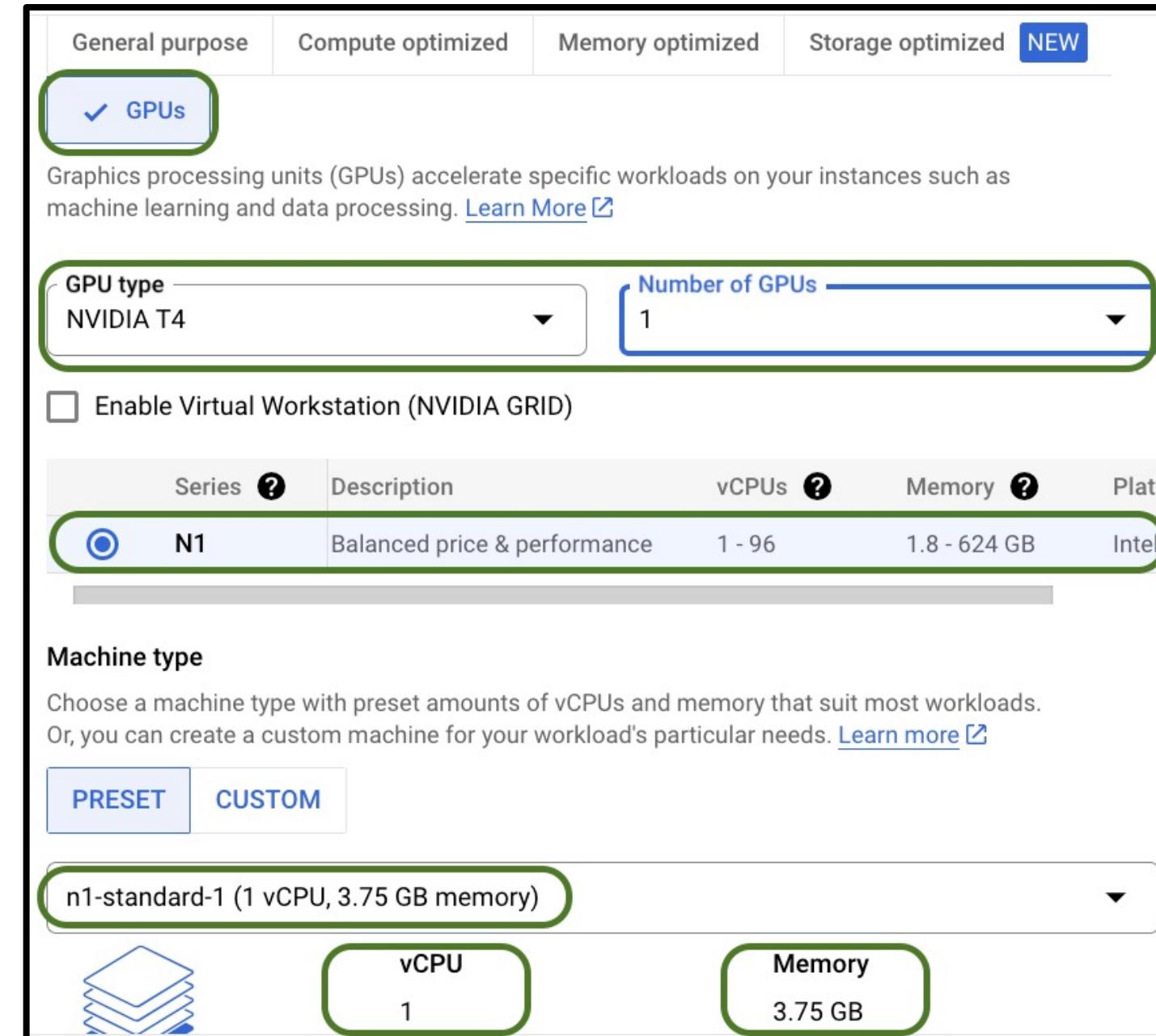
- Where do you use GPUs?

- Graphic Intensive workloads
- Machine Learning
- Scientific Computing (Math Intensive)
- 3D Visualization

- How to use GPUs in GCP GCE ?

- Machine Family: GPU
- Boot Disk: Use Deep Learning Linux OS for supporting GPUs

Machine Family: N1



Google Compute Engine - GPUs

- Which machine families support GPUs?

- N1 (general-purpose)

- Attach the GPU to the [VM](#) during, or after VM creation

- A3, A2, and G2 (accelerator-optimized)

- GPUs are [automatically attached](#) when you create the VM

- You can add GPU to [preemptible](#) and [Spot VM Instances](#)

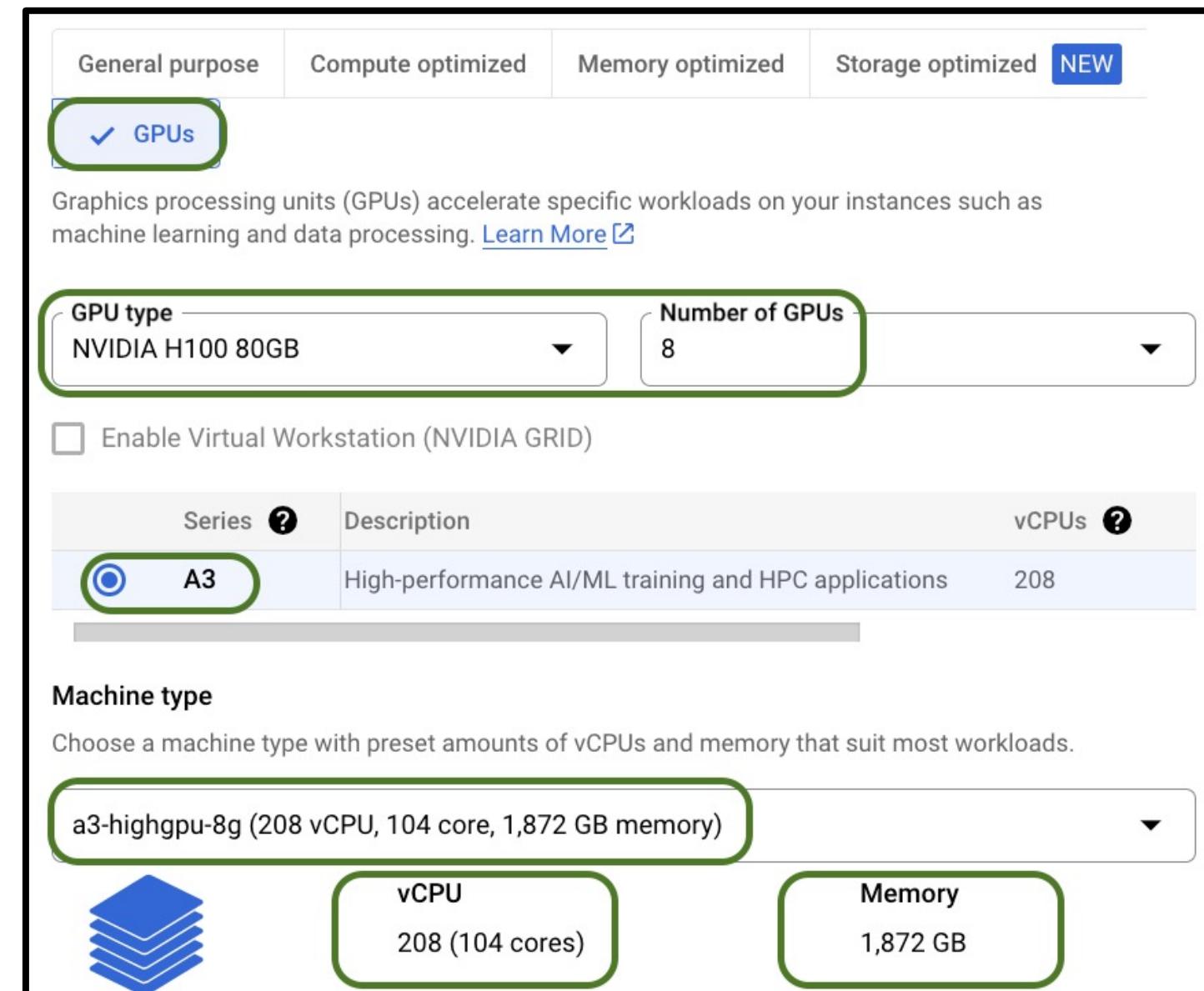
- **GPUs and host maintenance**

- VMs with attached GPUs [cannot live migrate](#) and [must stop](#) for host maintenance events

- **GPUs and block storage**

- You can add [Local SSDs](#) to VMs that have GPUs attached
- Not all GPU types support Local SSDs

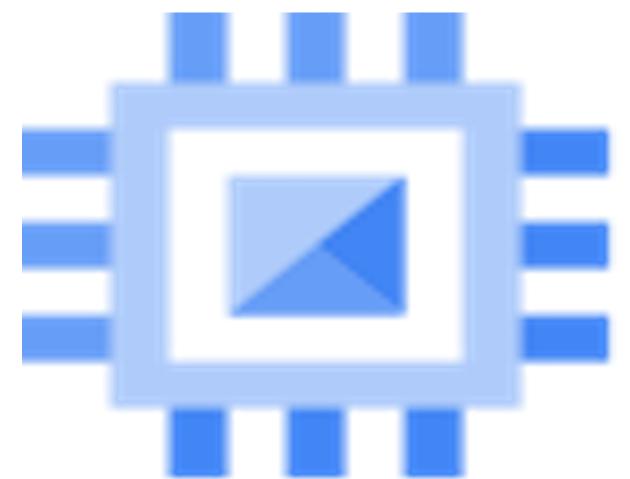
Machine Family: A3



Concept

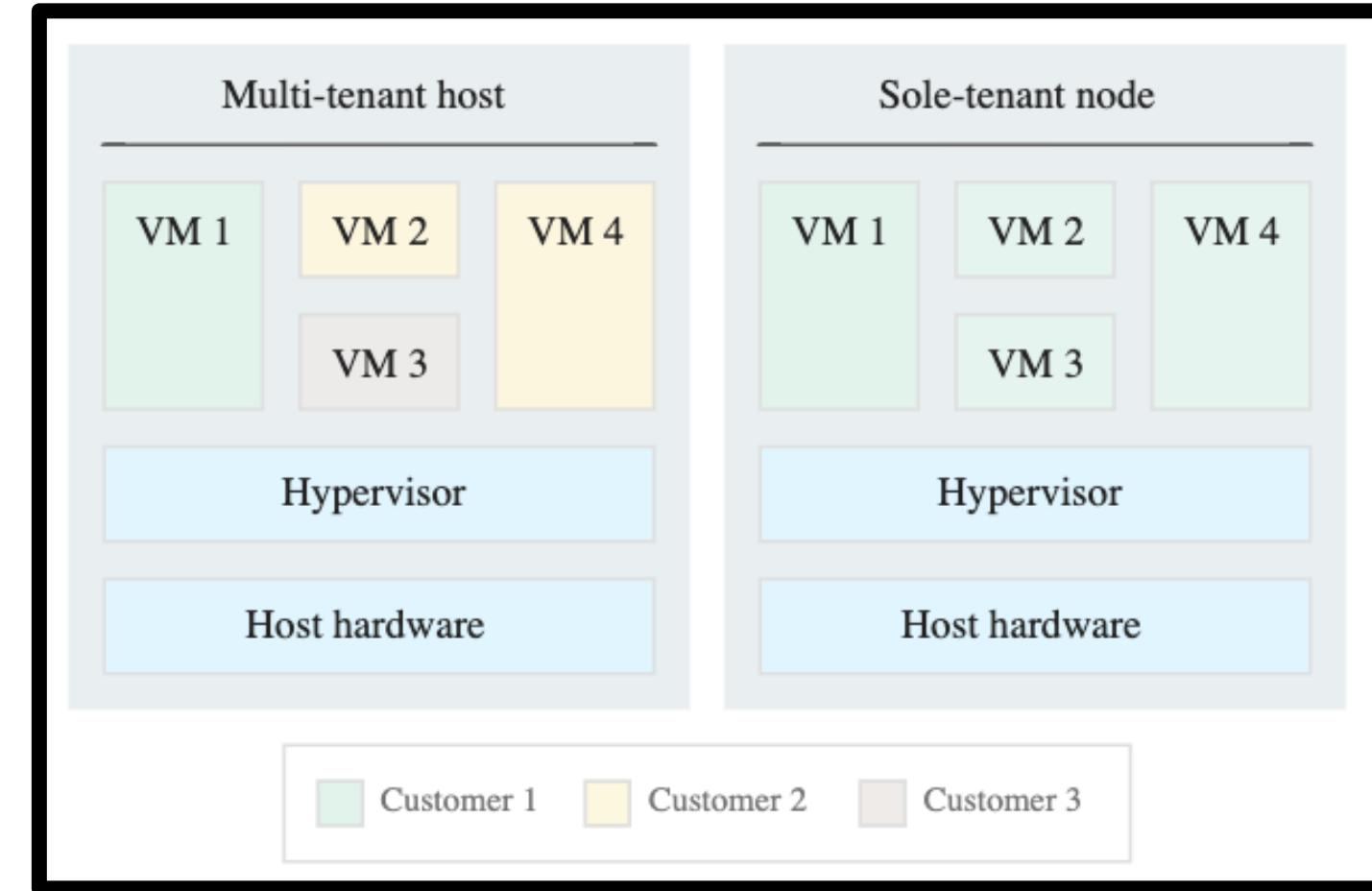


Google Compute Engine Sole-Tenant Nodes



Google Compute Engine - Sole-tenant Nodes

- Dedicated Physical server for hosting only your projects.
- NOT SHARED with any other customers
- Can use all Compute Engine features as other VMs
- Can meet security and compliance requirements with workloads that require physical isolation
- Can meet requirements for BYOL (Bring your own license) scenarios
- Configurable maintenance policy helps us to schedule the Sole-tenant HOST maintenance as per our desired timing.

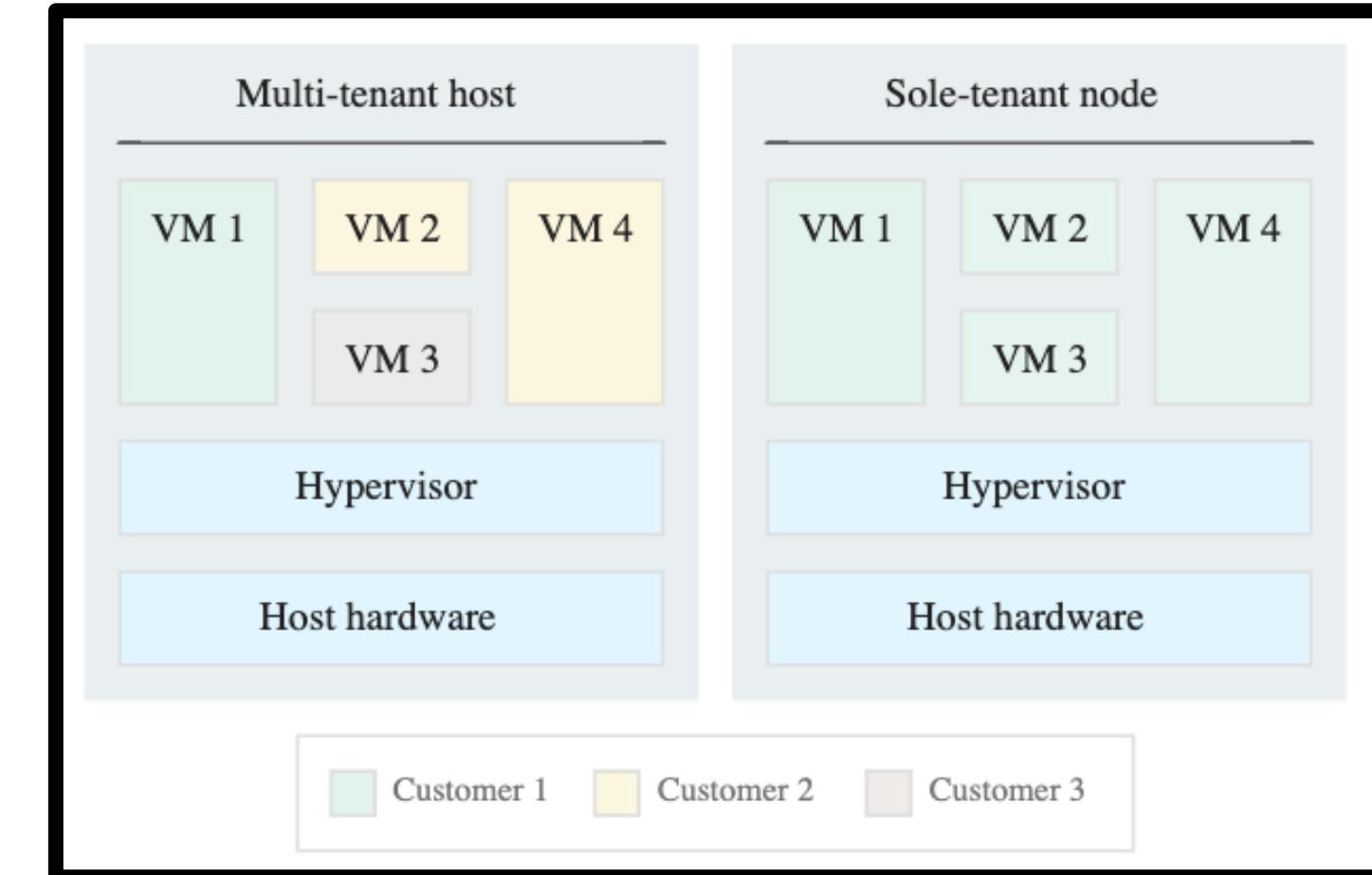


Reference: <https://cloud.google.com>

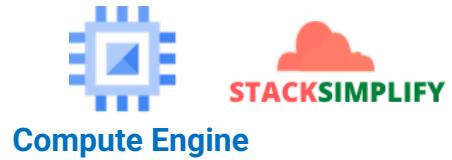
Google Compute Engine - Sole-tenant Nodes

- **Usecases**

- Gaming workloads with **performance requirements**
- Finance or healthcare workloads with **security and compliance requirements**
- Windows workloads with **licensing requirements**
- Machine learning, data processing, or image rendering workloads. For these workloads, consider reserving GPUs.



Sole-tenant Nodes - Node Template



• Node Template

- Defines the **properties** of each sole-tenant node in a node group
- Node Template is a **regional resource**

• Node Type

- m1-node-96-1433
 - vCPU: 96
 - **Memory:** 1433 GB
- Local SSD (Optional)
- GPU Accelerator (Optional)

• Affinity Labels

- Affinity labels **ensure** that VM instances run on the required node groups

Create a node template

Name *
node-template-1

Name is permanent

Node type *
c2-node-60-240

Local SSD
16 disks (6000 GB)

GPU accelerator
NVIDIA Tesla P100 (4 count)

CPU overcommit

CPU overcommit provides dedicated access to a physical server with the ability to control the overcommit levels of each VM instance scheduled on the node. [Learn more](#)



Enable CPU overcommit

Affinity labels

Use node affinity labels to ensure that instances run on the required node groups. [Learn more](#)

Key 1 *
apptype

Value 1 *
bankingapps

Google Compute Engine Sole-tenant Node Types

Node Types

Node type	Processor	vCPU	GB	vCPU:GB	Sockets	Cores:Socket	Total cores
c2-node-60-240	Cascade Lake	60	240	1:4	2	18	36
c3-node-176-352	Sapphire Rapids	176	352	1:2	2	48	96
c3-node-176-704	Sapphire Rapids	176	704	1:4	2	48	96
c3-node-176-1408	Sapphire Rapids	176	1408	1:8	2	48	96
c3d-node-360-708	AMD EPYC Genoa	360	708	1:2	2	96	192
c3d-node-360-1440	AMD EPYC Genoa	360	1440	1:4	2	96	192
c3d-node-360-2880	AMD EPYC Genoa	360	2880	1:8	2	96	192
g2-node-96-384	Cascade Lake	96	384	1:4	2	28	56
g2-node-96-432	Cascade Lake	96	432	1:4.5	2	28	56
h3-node-88-352	Sapphire Rapids	88	352	1:4	2	48	96
m1-node-96-1433	Skylake	96	1433	1:14.9	2	28	56
m1-node-160-3844	Broadwell E7	160	3844	1:24	4	22	88
m2-node-416-11776	Cascade Lake	416	11776	1:28.3	8	28	224
m3-node-128-1952	Ice Lake	128	1952	1:15.25	2	36	72
m3-node-128-3904	Ice Lake	128	3904	1:30.5	2	36	72
n1-node-96-624	Skylake	96	624	1:6.5	2	28	56

Reference: <https://cloud.google.com/compute/docs/nodes/sole-tenant-nodes>

Sole-tenant Nodes - Node Group

- **Node Group**
 - Group sole-tenant nodes to work as a **single unit** in that respective zone
- **Node Group Location**
 - **Region and Zone** where sole-tenant nodes will be created
- **Node Template**
 - Node Type, Affinity labels, Local SSD, GPUs and CPU Environment
- **Autoscaling**
 - **Resize** the node group with sole-tenant nodes based on **high or low loads**
 - Minimum Nodes: 2, Maximum Nodes: 4
- **Maintenance Settings**
 - Define a **desired maintenance window** for sole-tenant nodes
- **Share Settings**
 - **Do not share** this node group with other projects
 - Share this node group with **all projects** within the organization
 - Share this node group with **selected projects** within the organization

← Create a node group

Node group properties
These properties are permanent and cannot be edited after creation.

Name	node-group-1
Region	us-central1
Zone	us-central1-c

Node template properties
This property is permanent and cannot be edited after creation.

Node template	node-template-1
---------------	-----------------

Configure autoscaling
Use autoscaling to allow automatic resizing of this node group for periods of high and low load.

Autoscaling mode	On
Minimum nodes	2
Maximum nodes	4

4 Configure maintenance settings (optional)
Configure the behavior of your node group during host maintenance events. Unless physical server affinity is required, the Default maintenance policy is recommended. [Learn more](#)

Maintenance policy *	<input type="button" value="Default (recommended)"/>
Maintenance window	<input type="text" value="0:00 - 4:00"/> 4-hour window in UTC

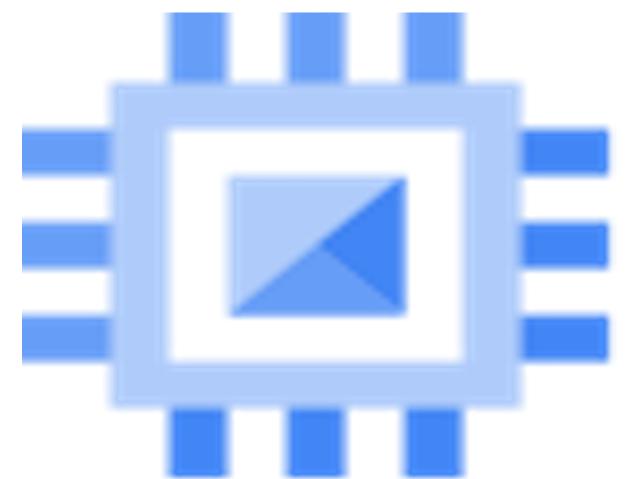
5 Configure share settings (optional)
Enable other projects to configure VMs on this node group. [Learn more](#)

<input checked="" type="radio"/> Do not share this node group with other projects	?
<input type="radio"/> Share this node group with all projects within the organization	?
<input type="radio"/> Share this node group with selected projects within the organization	?

Demo

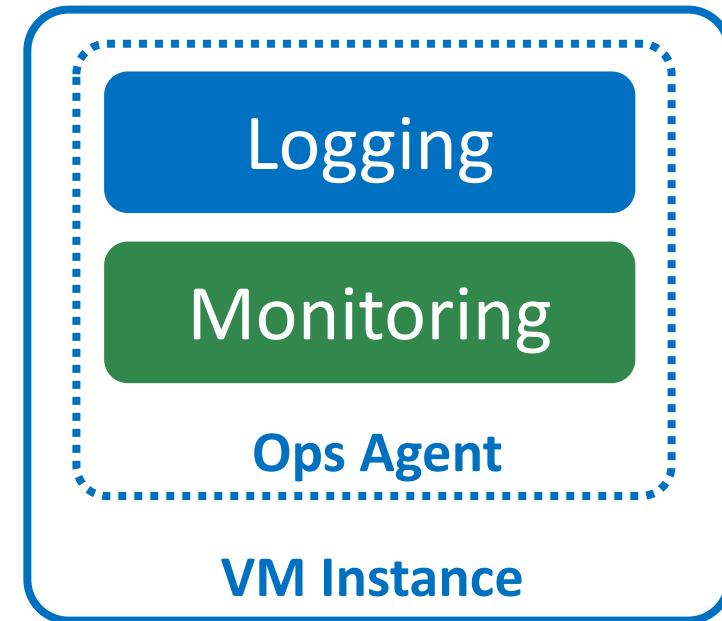


Google Compute Engine Ops Agent Logging and Monitoring



Compute Engine - Ops Agent

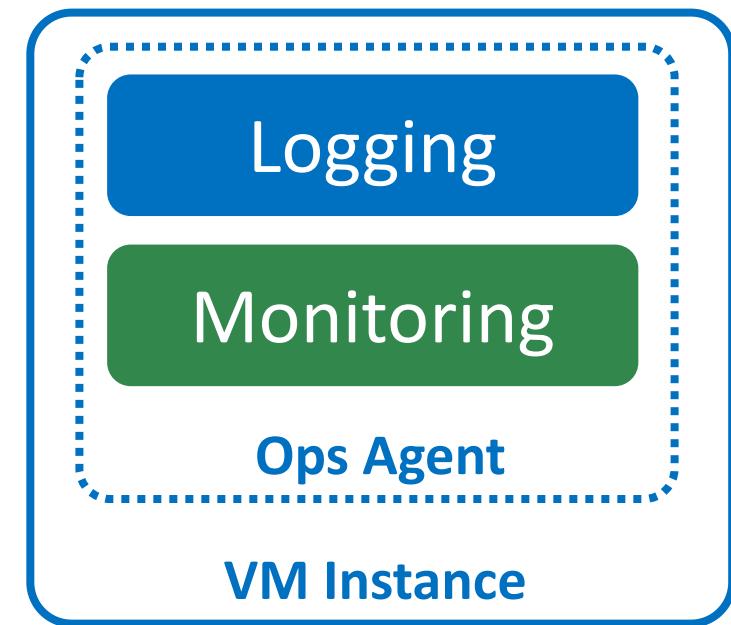
- **Cloud Observability Agents**
 - **Legacy Monitoring Agent & Legacy Logging Agent**
 - No new feature development
 - No support for new operating systems
 - Recommended to use [Ops Agent](#) for all new workloads
 - Transition your existing VMs to Ops Agent
 - **Ops Agent**
 - Latest and greatest
 - Recommended by google for all the new VM Instances
- **Ops Agent:** Single agent for collecting logs and metrics
 - Collects logs and sends them to [Cloud Logging](#)
 - Collects metrics and traces and sends them to [Cloud Monitoring](#)
- Uses [Fluent-bit](#) for logs
- Uses [Open Telemetry Collector](#) for metrics and traces



Compute Engine - Ops Agent

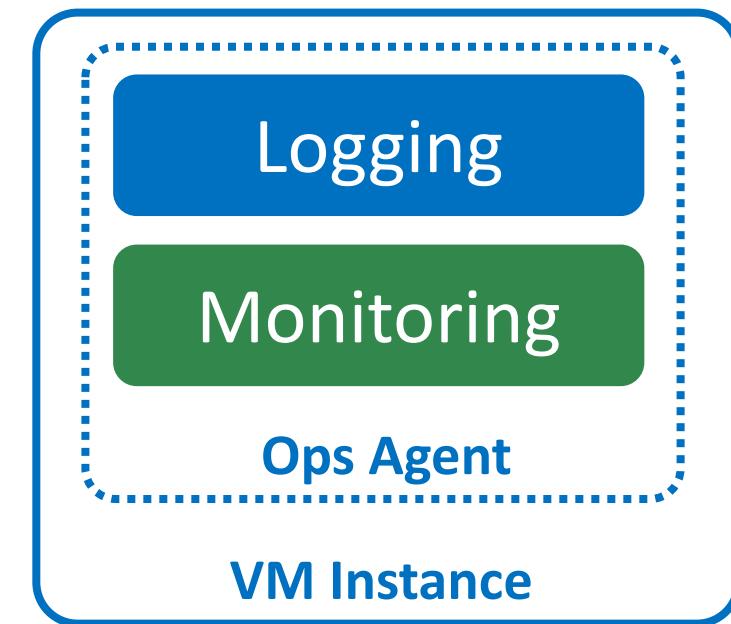
- **Ops Agent Features**

- Provides [YAML-based configurations](#)
- Support for [standard linux and windows distributions](#)
- [Multiple](#) ways to Install / Upgrade the Ops Agent
 - Install [automatically](#) during VM creation
 - Install on fleet of VMs using [gcloud](#)
 - Install on fleet of VMs using [automation tools](#)
 - Ansible
 - Chef
 - Puppet
 - Terraform
 - Agent policy using gcloud CLI
 - Install on Individual VMs using [Cloud Monitoring VM Instances Dashboard](#)
 - Install [manually](#) using commands



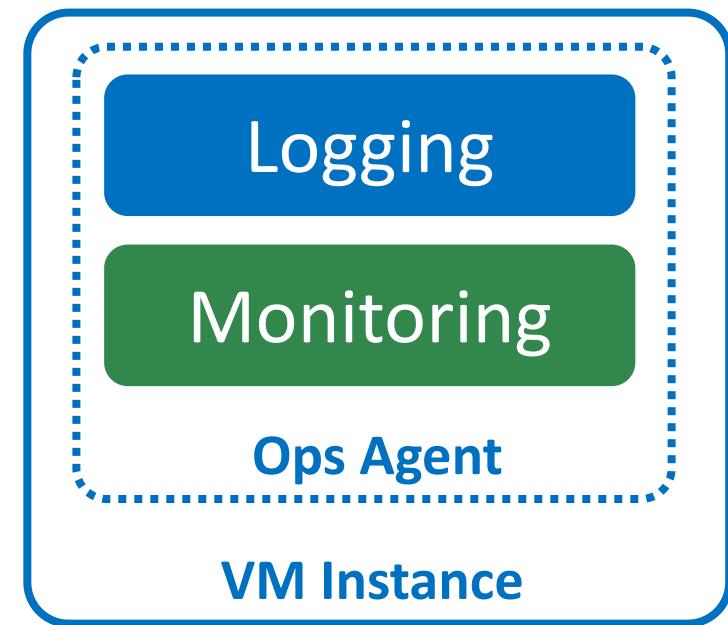
Compute Engine - Ops Agent

- **Ops Agent Logging Features**
- Improved performance when compared to legacy logging agent
- **Collecting logs from various sources**
 - Standard **system** logs (`/var/log/syslog`, `/var/log/messages`)
 - **File** based logs with customizable paths
 - Logs over **TCP** protocol
 - Logs over **Forward** protocol (Used by **Fluent Bit** and **Fluentd**)
- **Flexible processing**
 - Parse **text logs** into structured logs: **JSON-based** and **regular-expression-based** parsing.
 - Exclude **logs** based on labels and regular expressions.
- **Third-party application support**
 - Apache Kafka, Nginx, Apache Flink, Apache Hadoop, Apache Hbase
 - MariaDB, MongoDB, MySQL, Oracle DB, Redis, WildFly, SAP HANA
 - **For complete list:** <https://cloud.google.com/stackdriver/docs/solutions/agents/ops-agent/third-party>



Compute Engine - Ops Agent

- **Ops Agent Monitoring Features**
- System metrics collected with **no configuration**
 - **Linux and Windows:** cpu, disk, interfaces, swap, network, processes, agent self metrics
 - **Linux only:** GPU
 - **Windows only:** iis, mssql, pagefile,
- **Third-party application support**
 - Apache Kafka, Nginx, Apache Flink, Apache Hadoop, Apache Hbase
 - MariaDB, MongoDB, MySQL, Oracle DB, Redis, WildFly, SAP HANA
 - **For complete list:** <https://cloud.google.com/stackdriver/docs/solutions/agents/ops-agent/third-party>
- Collection of **Prometheus metrics** from applications running on Compute Engine
- Collection of **NVIDIA Data Center GPU Manager (DCGM)** metrics

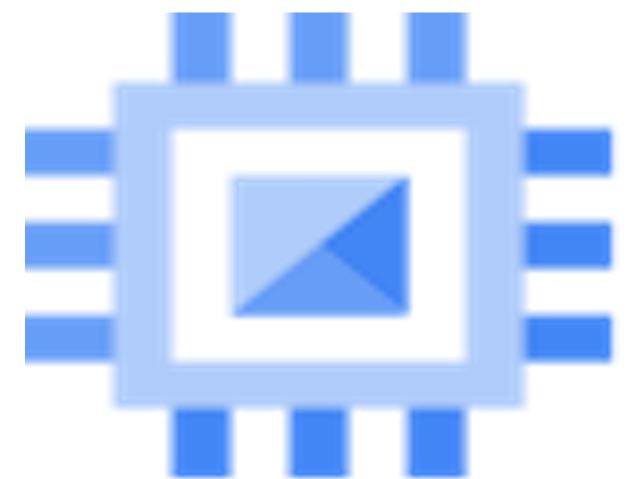


Concept



Google Compute Engine

Sustained & Committed Use
Discounts



Committed Use Discounts

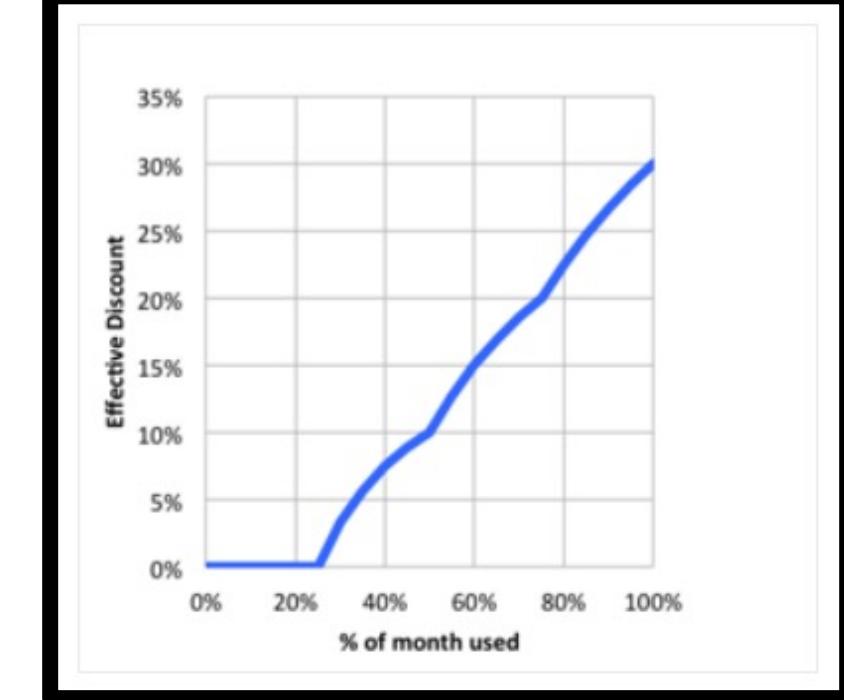
- These are ideal for workloads with **predictable resource** needs.
- Compute Engine resource such as can be purchased at discounted price
 - vCPUs
 - Memory
 - GPUs
 - Local SSDs
- Commit for 1 year or 3 years
- 57% to 70% discount based on Machine types and GPUs
- Applicable for predefined and custom machine types
- Automatically apply to VMs by Google Kubernetes Engine, Dataproc, and Compute Engine.

Committed Use Discounts

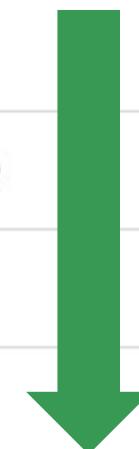
- **RESTRICTION:** Do not apply to VMs created using App Engine flexible environment or Dataflow.
- **RESTRICTION:** Commitments must be purchased on a per-region basis
- **RESTRICTION:** Do not apply for projects that are in the free tier period with free tier credit
- **RESTRICTION:** Projects that do not have any payment history do not qualify for committed use discounts.
- **RESTRICTION:** Do not apply to preemptible, Spot VM instances, N1 shared-core machine types, or extended memory

Sustained Use Discounts

- **Automatic discounts** for running specific Compute Engine resources a significant portion of the billing month
- When VM usage hours **increases**, discount increases.



Resources	Usage level (% of the month)	% at which incremental is charged	Incremental rate (USD/hour) example: c2-standard-4 instance
General-purpose N2 and N2D predefined and custom machine types, and Compute-optimized machine types	0%-25%	100% of base rate	\$0.2088
	25%-50%	86.78% of base rate	\$0.1811
	50%-75%	73.3% of base rate	\$0.1530
	75%-100%	60% of base rate	\$0.1252

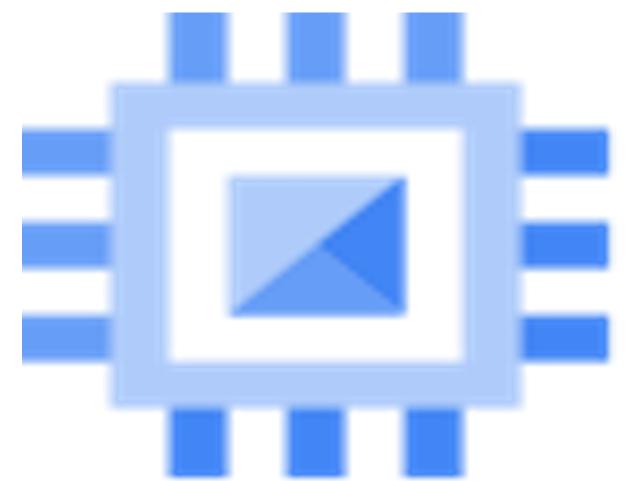


Sustained Use Discounts

- With Sustained use discounts, we get **10% to 50%** discounts based on usage in a month.
- **Automatically** apply to VMs created by both Google Kubernetes Engine and Compute Engine.
- **LIMITATION:** **Do not apply** to VMs created using the App Engine flexible environment and Dataflow.
- **LIMITATION:** Do not apply to E2 and A2 machine types.



Google Cloud Compute Engine Storage



Compute Engine - Storage

Compute Engine
Storage

Disks

Storage Pool

Snapshots

Images

Demo-1

Demo-2

Demo-3

Demo-4

Demo-5

Demo-6

Demo-7

Demo-8

Demo-9

Cloud Key Management Service

Create Non-Boot Persistent Disk and attach to VM, use KMS Encryption keys created in previous for encryption

Resize Disk Size for Boot and Non-Boot Disks

Regional Persistent Disk

Create Hyperdisks

Hyperdisks with Storage pools

Create Disk Image

Create Disk Snapshot

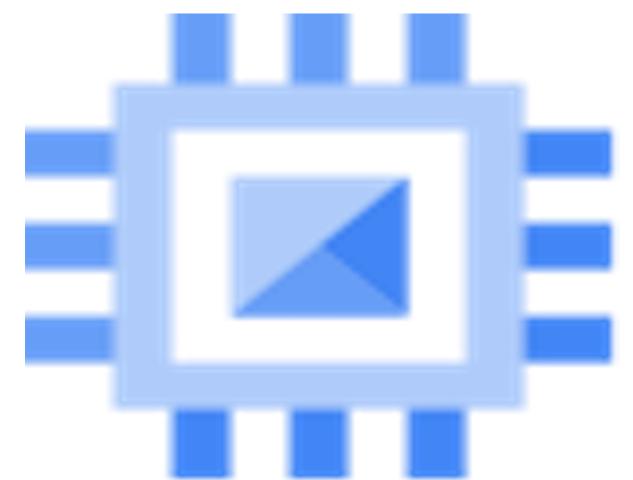
Create Local SSD



Cloud
Key Management Service



Google Cloud Key Management Service



Data States

Data at rest

Example-1: Data stored in Hard Disks

Example-2: Data stored in backups and archives

Data in Transit

Data IN and OUT from cloud over INTERNET

Data in Transit inside cloud (Example: Application talking to DB)

Data in use

Active data currently in processing state

Example: Data in RAM

Data Encryption

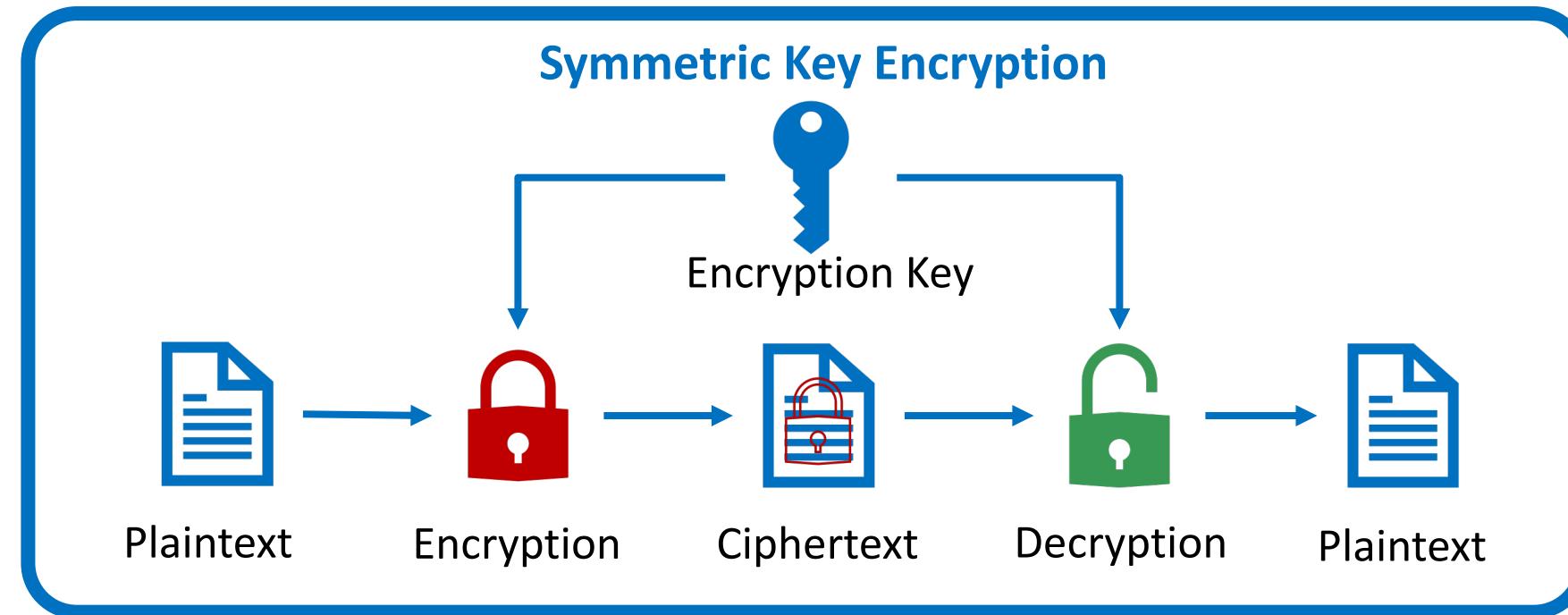
Types of Data Encryption

Symmetric
Encryption

Asymmetric
Key Encryption

It is recommended to encrypt both **Data at Rest** and **Data at Transit**

Symmetric Key Encryption



- Symmetric Key Encryption uses **same key** for both Encryption and Decryption
- **Example Encryption Algorithms**
 - DES – Data Encryption Standard
 - Triple DES
 - AES – Advanced Encryption Standard
 - IDEA - International Data Encryption Algorithm

Symmetric Key Encryption

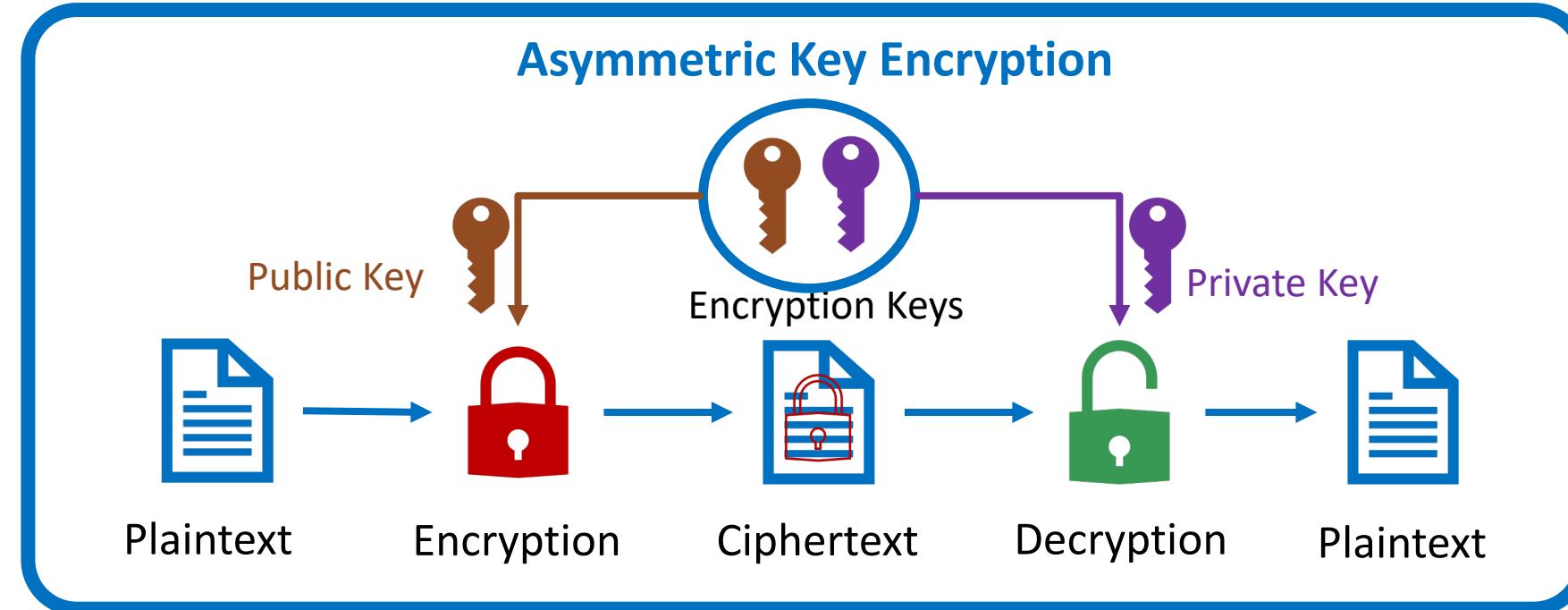
- **Advantages**

- **Security:** Algorithms like AES take billions of years to crack using brute-force attacks.
- **Speed:** Because of its shorter key it is much faster to execute and uses less resources (CPU, Memory) to Encrypt and Decrypt
- **Industry adoption and acceptance:** Algorithms like AES have become the gold standard of data encryption because of their security and speed benefits and hugely in use industry wide.
- RECOMMENDED for Bulk Data Transfers

- **Challenges**

- How to secure encryption key ?
- How to share encryption key ?

Asymmetric Key Encryption



- Asymmetric Key Encryption uses **two keys**: Private and Public Keys
- Encrypts data with **public key** and decrypts with **private key**
- **Example Encryption Algorithms**
 - **RSA**: Digital Signature Standard
 - **DSC**: Digital Signature Standard
 - **DSA**: Digital Signature Algorithm
 - **ECC**: Elliptical Curve Cryptography

Asymmetric Key Encryption

- **Advantages**

- Private key is **not shared**. Overall process is **more secure** when compared to Symmetric key encryption

- **Disadvantages**

- The encryption process is **slow**
- Resource utilization is **very high**
- Not recommended for bulk data transfers

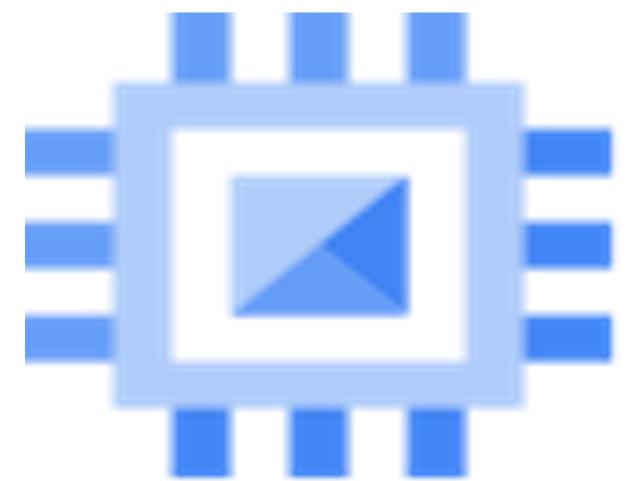
Google Cloud - Key Management Service (KMS)

- Cloud KMS is used to **centrally manage** encryption keys on GCP
- Supports both **Symmetric** and **Asymmetric** key encryptions
- Use KMS generated encryption keys in your **applications** and **GCP Services** (Compute Engine, Cloud SQL)
- KMS provides an **API** to **encrypt, decrypt or sign data** which can be used in our Application Development.
- Key Management **Options** available for use
 - **Google-managed** encryption key (No configuration required)
 - **CMEK:** **Customer-managed** encryption key (Manage via Cloud KMS)
 - **CSEK:** **Customer-supplied** encryption key (Manage outside of Google cloud)

Concept



Google Compute Engine Storage Options



Google Cloud Platform - Storage Options

Block Storage



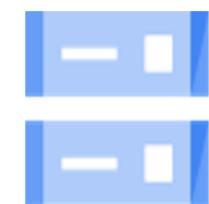
Persistent Disk (Regional & Zonal)
Local SSD, Hyperdisks

File Storage



Filestore

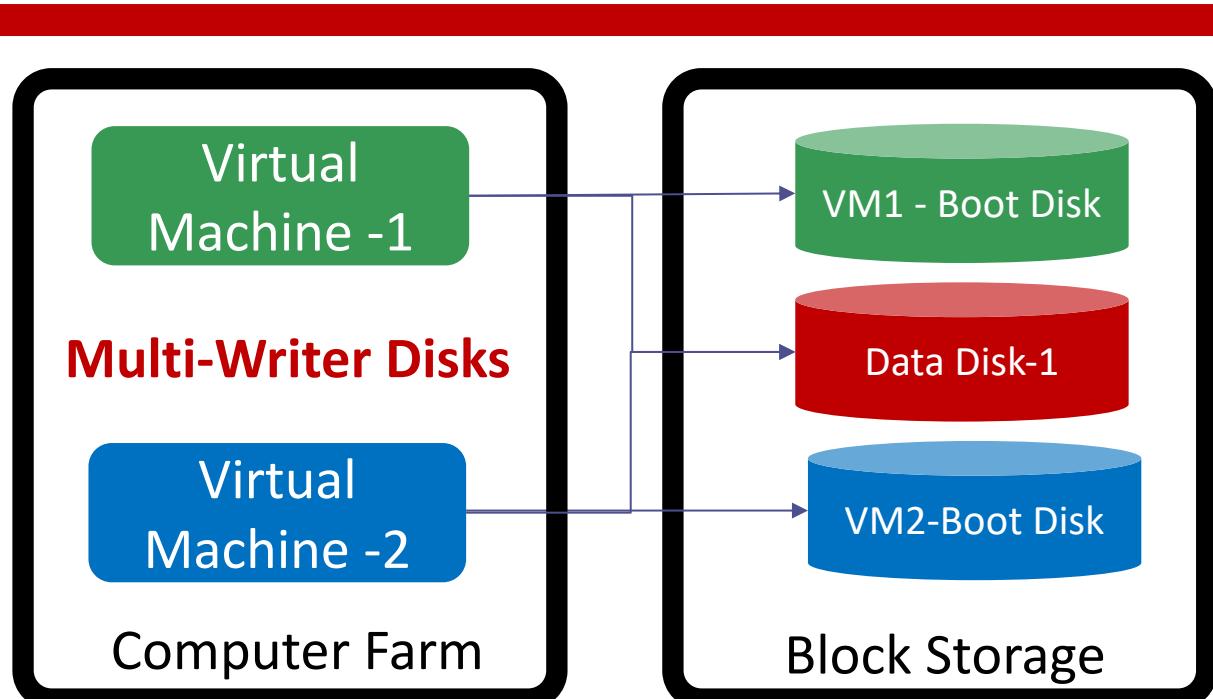
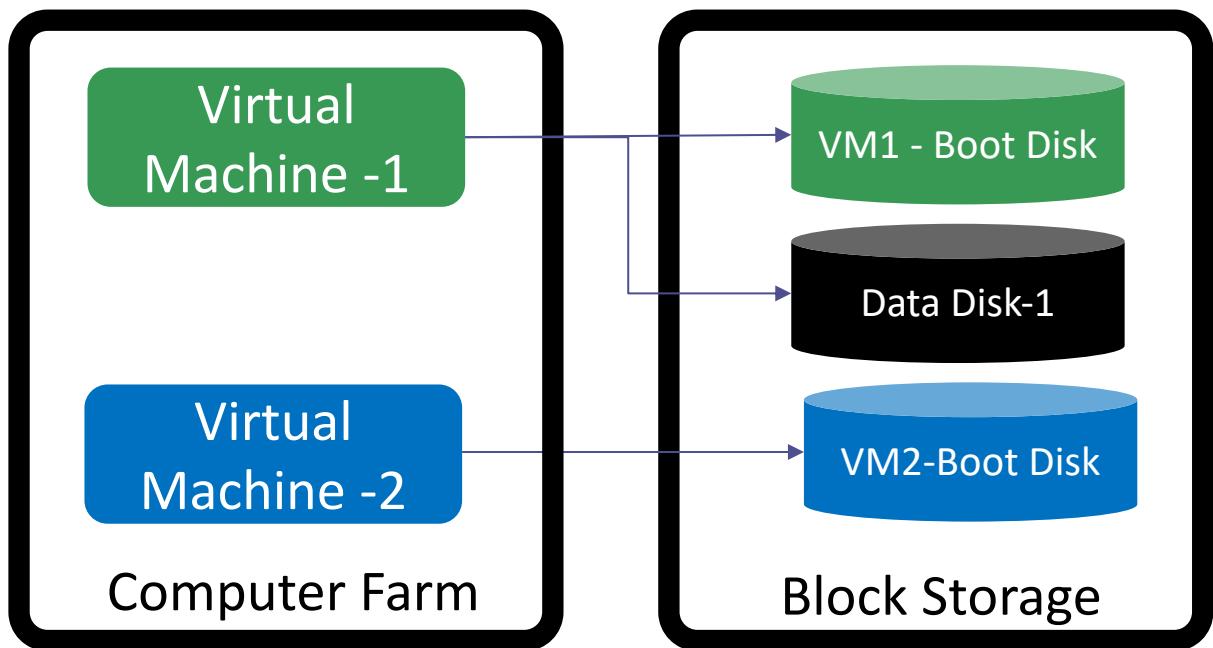
Object Storage



Cloud Storage

Block Storage in General

- **General Use:** Hard Disks attached to VMs
- At any point of time one Block Storage Device can be connected to **one VM in Read-Write Mode**
- We can attach **Read-Only** Block Storage Devices to Multiple VMs
- One VM can be associated with **Multiple** Block Storage Devices
- **Recent & Latest:** Multiple VMs can connect to a Block Storage Device in **Read-Write Mode** and that feature is called **Multi-Writer** Disks
- **Block Storage Types in General**
 - **Direct-attached Storage (DAS):** Regular HDD, SSD
 - **Storage Area Network (SAN):** High-Speed network that provides block level network access to Storage.



Google Compute Engine - Block Storage Options



Block Storage



Zonal Persistent Disk

Data replicated in single zone

Lifecycle not tied to VM Instance

More durable

Regional Persistent Disk

Data replicated in two zones
in same region

Local SSD

Local Block Storage on Host
of the VM Instance

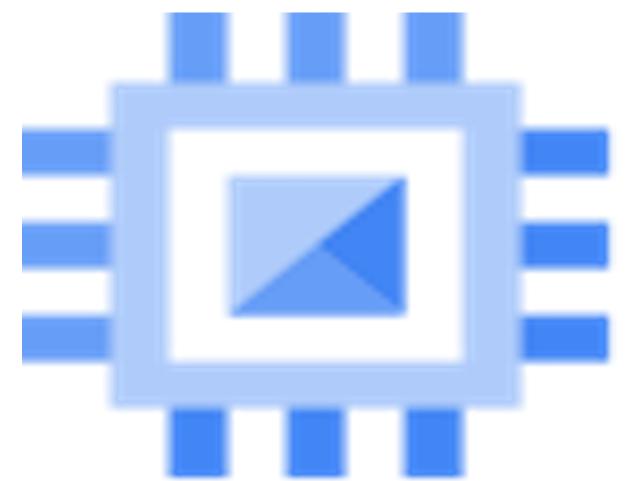
Lifecycle tied to VM Instance

Temporary Data

Demo



Google Compute Engine Persistent Disks (PD)



Compute Engine - Persistent Disks

Zonal Persistent Disk

Data replicated in **single zone**

More **durable**

Regional Persistent Disk

Data replicated in **two zones in same region**

Network block storage attached to your VM Instance

PDs can be used as **boot disk or data disk** for a VM Instance

Lifecycle **not tied** to VM Instance (Attach / Detach from VM Instance to other)

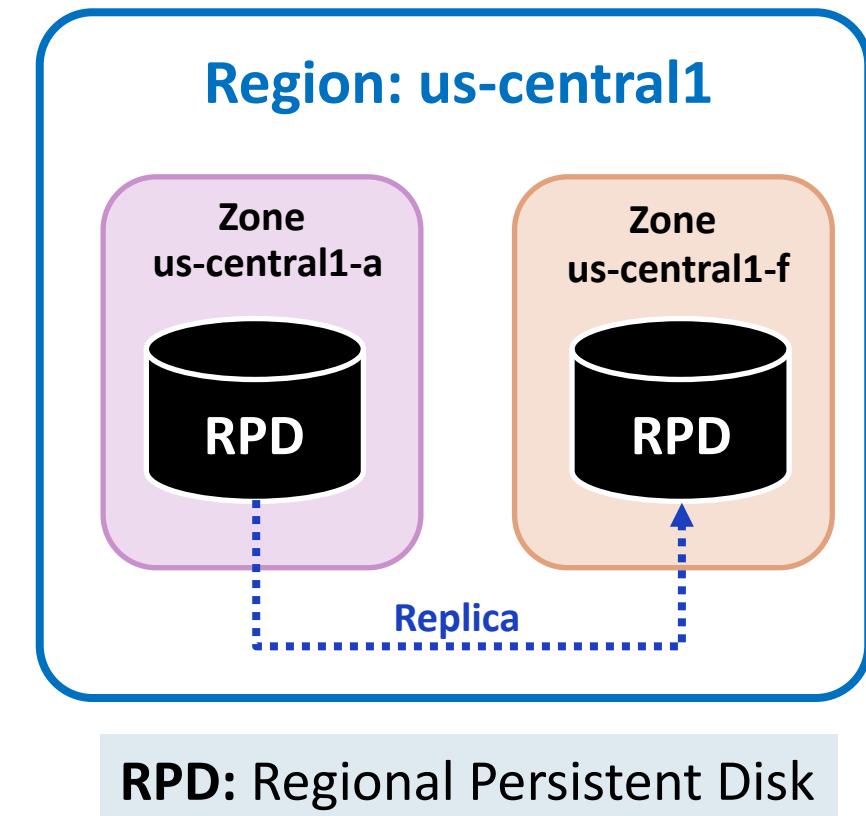
VERY FLEXIBLE: Increase size when needed, Performance scales with Size (For higher performance, increase disk or add more PDs)

Resize Disks: You can only resize a persistent disk to **increase its size**. You cannot reduce the size of a persistent disk.

Resize Disks: You can resize disks at **any time**, whether or not the disk is attached to a running VM.

Compute Engine Storage - Persistent Disks

- How regional persistent disks are different from Zonal persistent disks ?
- You can attach regional PDs only to VMs that use **E2**, **N1**, **N2** and **N2D** machine types
- You **can't use** a regional persistent disk with a memory-optimized, compute-optimized, storage-optimized or accelerator-optimized machine type VM.
- You **cannot use** regional persistent disks as **boot disks**.
- You can create a regional persistent disk from a **snapshot** but **not an image**.
- The **minimum size** of a regional standard persistent disk is **200 GB**.



Compute Engine Storage - Persistent Disks

Persistent Disk Types

Standard Persistent Disk

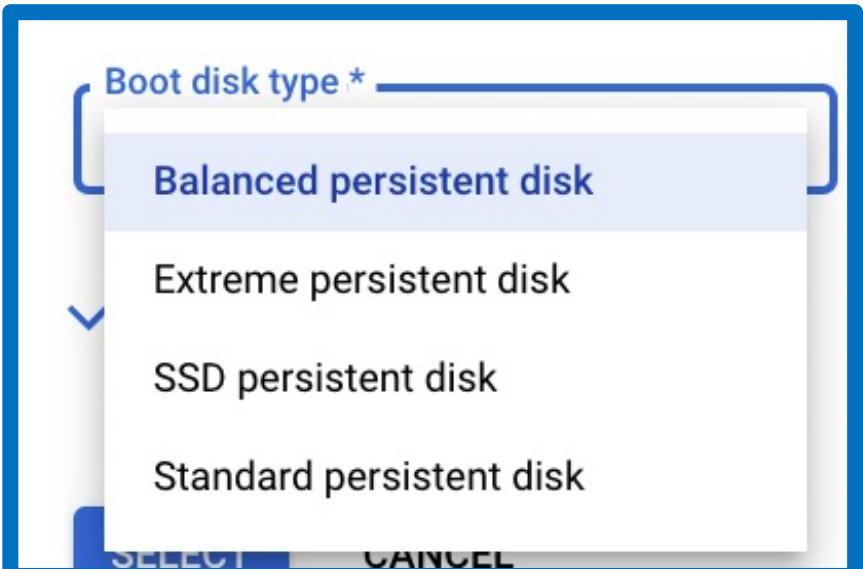
Balanced Persistent Disk

SSD Persistent Disk

Extreme Persistent Disk

Hyperdisk

For VM Instance - Boot Disks



Hyperdisk cannot
be used as VM
Boot Disk

For Data Disks

Disk settings

Disk type *

Hyperdisk Balanced

Hyperdisk Extreme

Hyperdisk Throughput

Balanced persistent disk

Extreme persistent disk

SSD persistent disk

Standard persistent disk

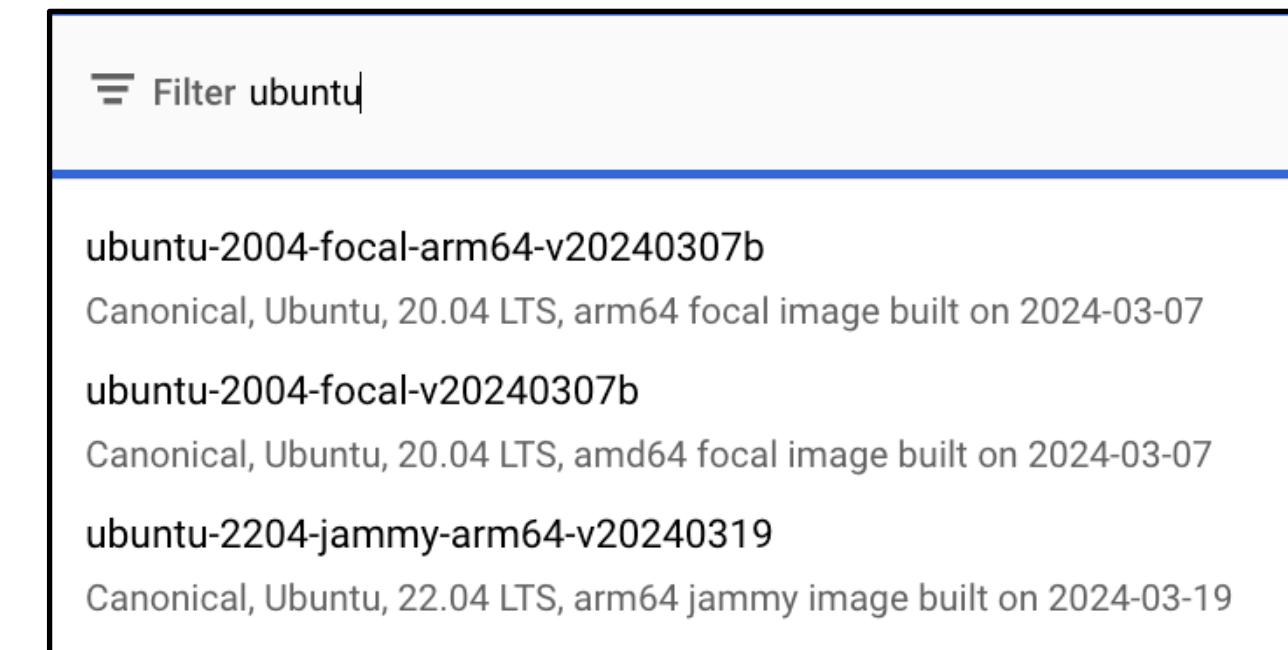
Compute Engine Storage - Persistent Disks

- How many ways we can create the Persistent Disks ?
- Create Persistent Disks with **source** as
 - **Blank Disk:** You can create an empty Disk
 - **Image:** You can create a bootable disk image from another disk image
 - **Snapshot:** You can create a disk from another disk snapshot
 - **Instant Snapshot:** You can create disk from an instant snapshot of another disk
 - **Archive Snapshot:** You can create disk from an archive snapshot of another disk

Disk Source Types



Source: Image



Compute Engine Storage - Persistent Disks

Performance estimate for

Size	Series	CPU Count
500 GB	N2	1

Standard **Balanced** **SSD** **Extreme**

	Optimized for	Read IOPS per instance	Write IOPS per instance	Read throughput per instance	Write throughput per instance
Cost-sensitive, throughput optimized non-boot data drives	General purpose enterprise applications. Best price per GB	375	750	60	60
Performance sensitive, business critical applications. Best price per IOPS.	3,000	3,000	140	140	204
High-end database workload like SAP Hana	15,000	9,000	240	240	240
Up to 15,000					
Up to 9,000					
240					

Size and CPU is directly proportional to performance of disk (If Size or CPU increases, performance of disk increases)

Experience this practically in google cloud console

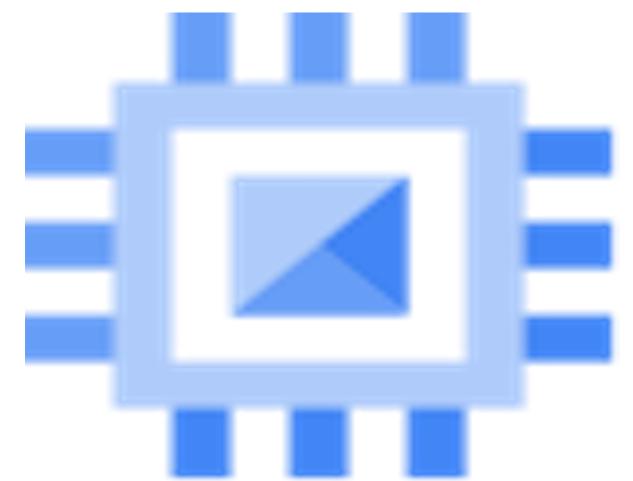
Persistent Disks – Standard vs Balanced vs SSD vs Extreme SSD

Feature	Standard	Balanced	SSD	Extreme SSD
Storage Type	Hard Disk Drive (HDD)	Solid State Drive (SSD)	Solid State Drive (SSD)	Solid State Drive (SSD)
Naming	pd-standard	pd-balanced	pd-ssd	pd-extreme
Cost	Low	Medium	Expensive	Expensive
Durability	99.99%	99.999%	99.999%	99.9999%
Performance - Sequential IOPS (Big Data / Batch Processing)	Good	Good	Very Good	Very Good
Performance – Random IOPS (Transactional Apps)	Bad	Good	Very Good	Very Very Good
Boot / Data Disk	Boot / Data Disk	Boot / Data Disk	Data Disk	Data Disk
Use cases	Cost-sensitive, throughput optimized non-boot data drives	General purpose enterprise applications.	Performance sensitive, business critical applications.	High-end database workload like SAP Hana

Demo



Google Compute Engine Hyperdisk



Compute Engine Storage - Hyperdisk

- **Hyperdisk:** newest generation of network block storage
- **What is the major difference between Persistent Disks and Hyperdisk?**
 - **Persistent Disk:** Performance [scales automatically](#) with [size and cpu](#)
 - **Hyperdisk:** You can provision performance [directly](#)
 - Dedicated [IOPs](#)
 - Dedicated [Throughput](#)
- Most of the features are [same as](#) Persistent Disks
 - Hyperdisk volume can be mounted to VM using [NVMe or SCSI](#) interface
 - [Attach, detach from VM](#) (Lifecycle not tied to VM)
 - Data is [persistent](#) over VM reboots and deletions

← Create a disk

Disk settings

Disk type * Hyperdisk Balanced ?

COMPARE DISK TYPES

Size * 100 GB ?

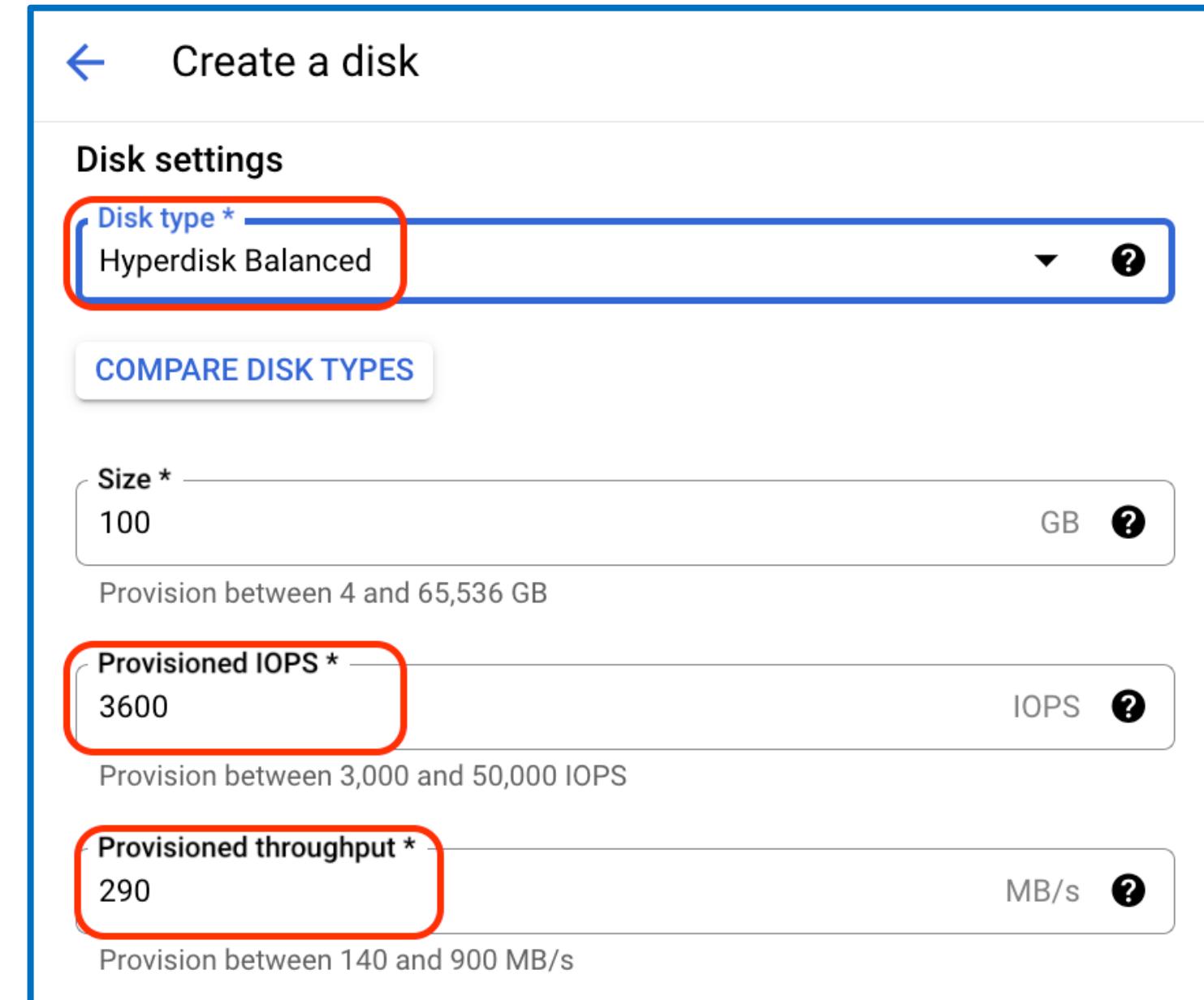
Provision between 4 and 65,536 GB

Provisioned IOPS * 3600 IOPS ?

Provision between 3,000 and 50,000 IOPS

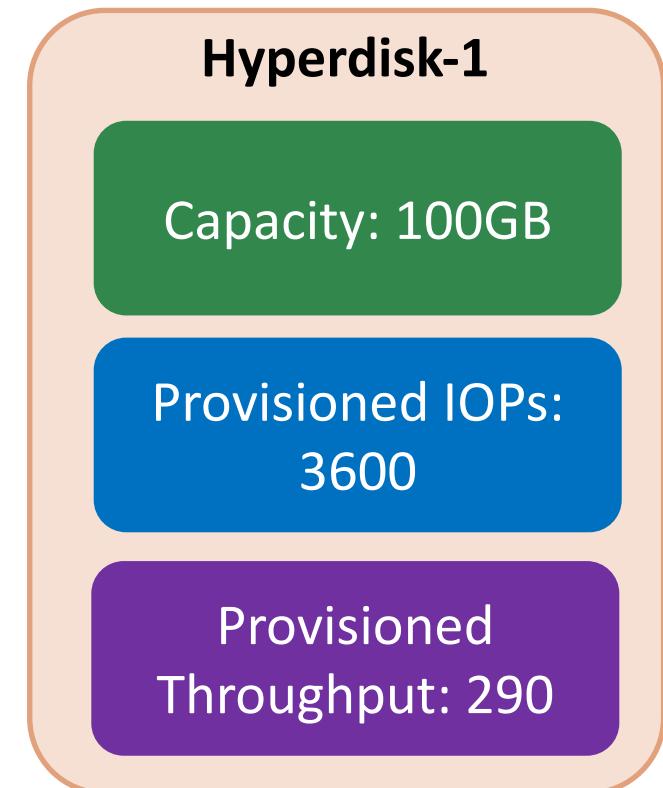
Provisioned throughput * 290 MB/s ?

Provision between 140 and 900 MB/s



Compute Engine Storage - Hyperdisk Types

- **Hyperdisk Balanced**
 - Primarily used for **regular** workloads
 - **Usecases:** Web Applications, medium-tier databases
- **Hyperdisk Extreme**
 - Primarily used for **performance-critical applications** where Extreme Persistent Disk **does not provide** enough performance
 - **Usecases:** High performance databases
- **Hyperdisk Throughput**
 - Lets you flexibly **provision capacity and throughput** as needed for your **scale-out workloads**.
 - **Usecases:** Hadoop, Kafka, data drives for cost-sensitive apps, scale-out analytics



Hyperdisk - Supported Machine Types

Hyperdisk Balanced

Hyperdisk Balanced supports these machine types:

- [H3 machine types](#) with 88 vCPUs
- [C3 machine types](#) with 22, 44, 88, or 176 vCPUs
- [C3D machine types](#) all machine types
- [M1 machine types](#) with 40, 80, 96, or 160 vCPUs
- [M2 machine types](#) with 208 or 416 vCPUs
- [M3 machine types](#) with 32, 64, or 128 vCPUs

Hyperdisk Extreme

Hyperdisk Extreme supports these machine types:

- C3 with 88 or more vCPUs
- C3D with 60 or more vCPUs
- M1 with 80 or more vCPUs
- M2 (all machine types)
- M3 with 64 or more vCPUs
- N2 with 80 or more vCPUs

Hyperdisk Throughput

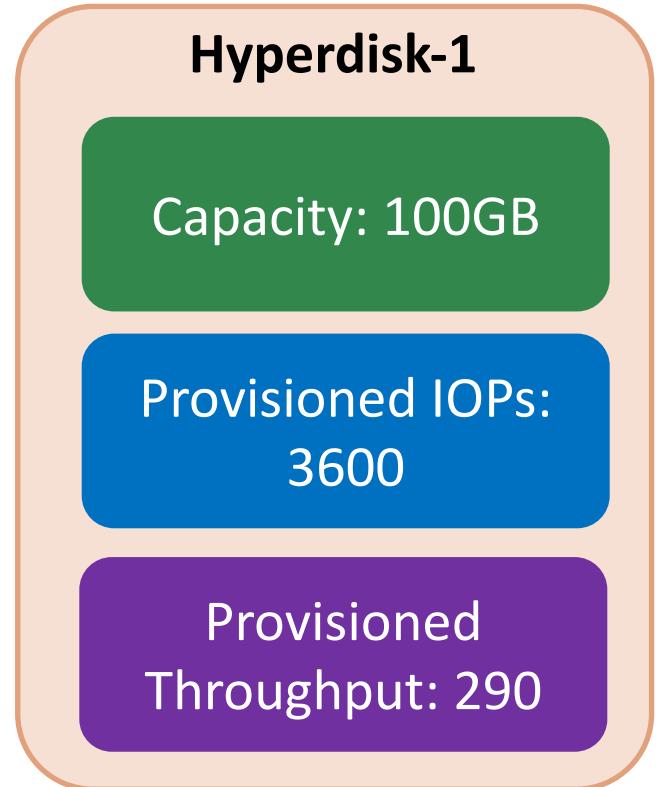
Hyperdisk Throughput supports these machine types:

- A3
- C3
- C3D
- G2
- H3
- M3
- N2
- N2D
- T2D
- Z3 ([Preview](#))

Reference: <https://cloud.google.com/compute/docs/disks/hyperdisks#machine-type-support>

Compute Engine Storage - Hyperdisk Limitations

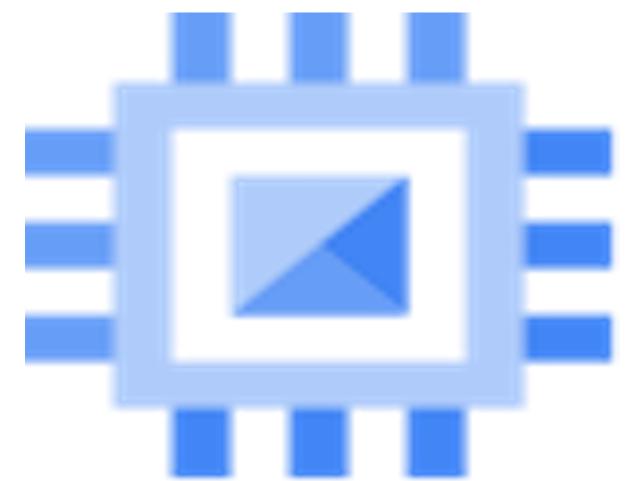
- You can't **create an image or machine image** from a Hyperdisk Extreme or Hyperdisk Throughput volume.
- You can't **clone** a Hyperdisk volume.
- Hyperdisk volumes are **zonal only**. You **can't create regional** Hyperdisk volumes.
- You **can't attach** multiple VMs in **read-only mode** to a Hyperdisk volume.
- Hyperdisk volumes can't be used in **multi-writer mode** or attached to multiple VMs.
- Hyperdisk Extreme and Hyperdisk Throughput volumes can't be used as **boot disks**.



Demo



Google Compute Engine Storage Pools

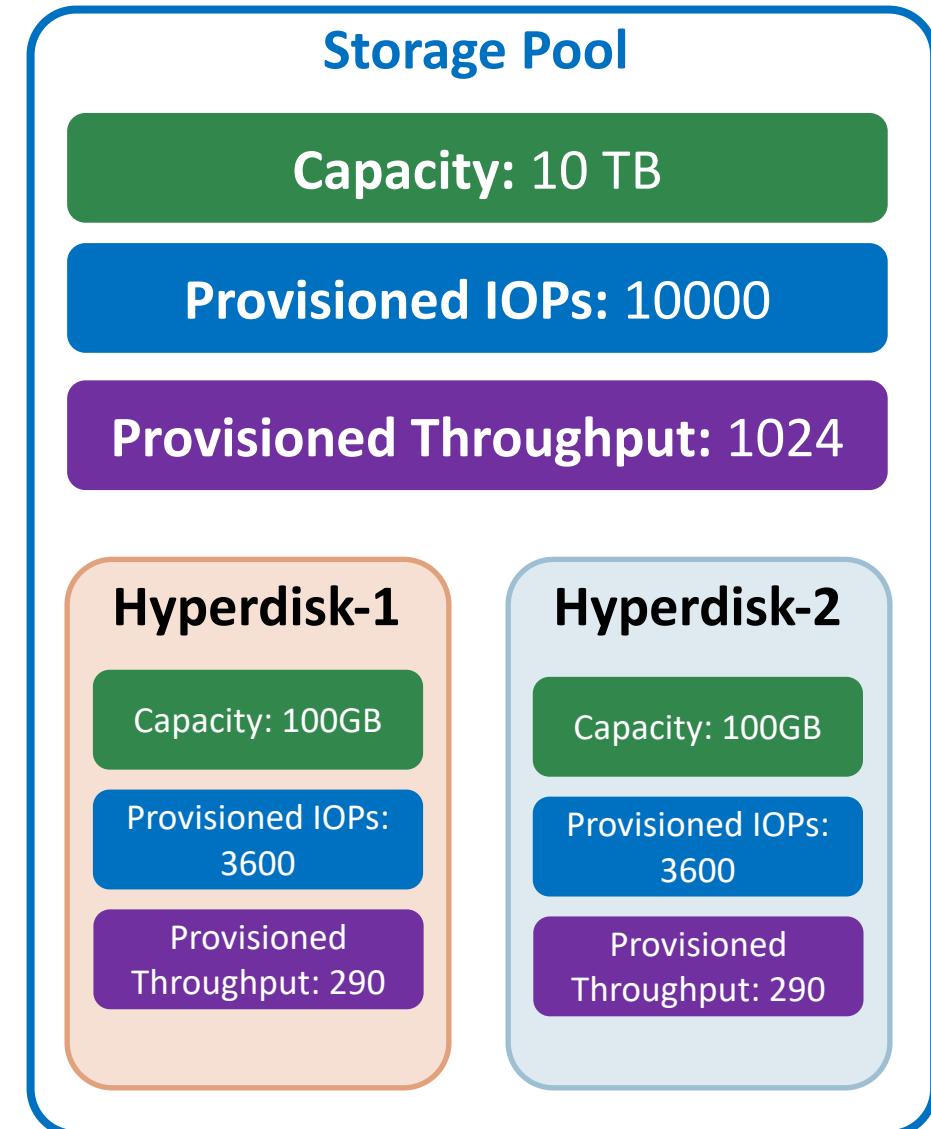


Compute Engine Storage - Storage Pools

- **Storage Pools:** Primarily used for large-scale storage
- Pre-purchased collection of capacity, throughput, and IOPS which you can then provision to your applications as needed

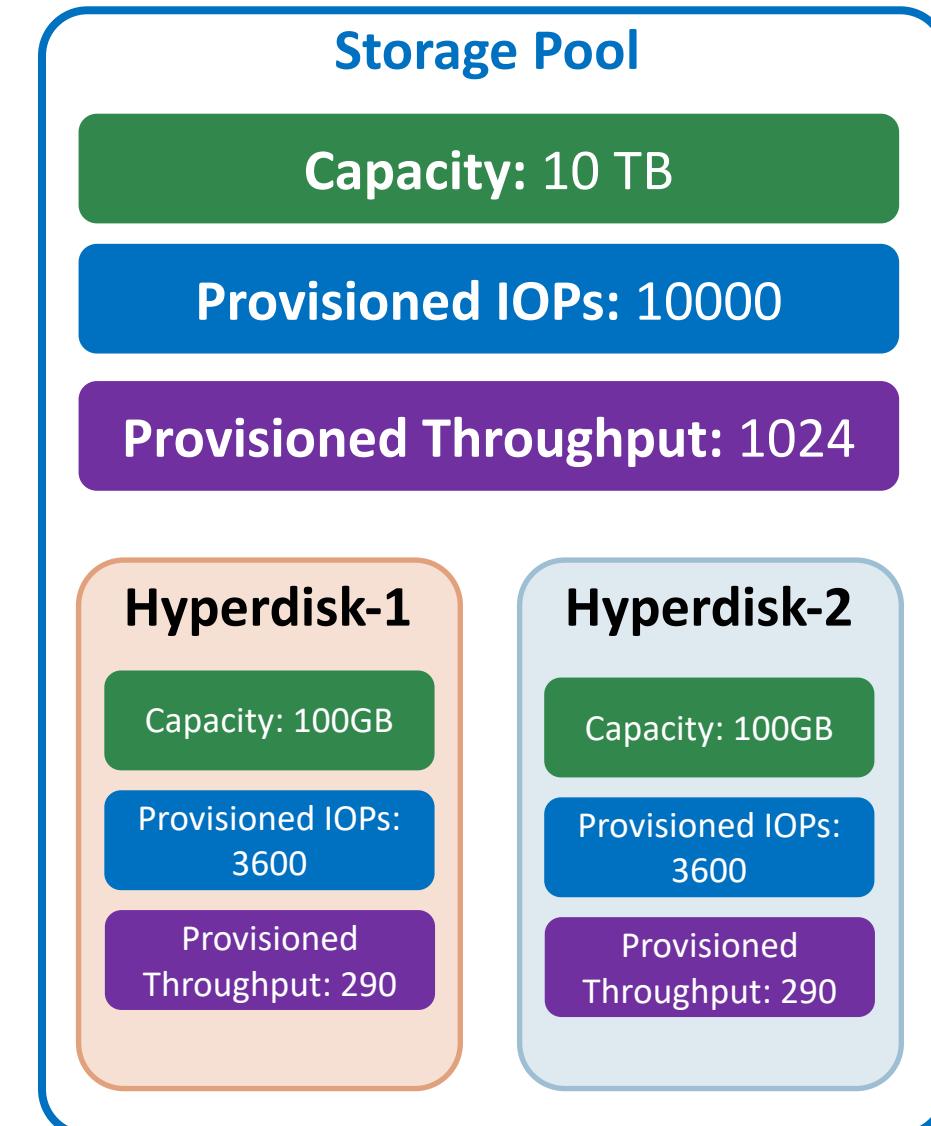
• Storage Pool Types

- **Hyperdisk Balanced:** We specify
 - Capacity
 - Provisioned IOPs
 - Provisioned Throughput
- **Hyperdisk Throughput:** We specify
 - Capacity
 - Provisioned Throughput



Compute Engine Storage - Storage Pools

- What are key features of Storage Pools ?
- **Capacity Thin Provisioning**
 - Blocks are **allocated as needed** instead of allocating all the blocks in advance (during disk creation).
 - In cases like where we **create disks with big sizes (100GB)** but when coming to usage we use only **very less (5GB)**, this feature will be very helpful
- **Data Reduction**
 - Storage pools use a variety of **data reduction technologies** to increase storage efficiency by **compressing data**
- **Auto-grow Capacity**
 - If the storage pool utilization **reaches 80% of the provisioned capacity**, Hyperdisk Storage Pools attempts to **automatically add capacity** to the storage pool to prevent errors related to insufficient capacity.



Compute Engine Storage - Storage Pools

- **When to use Storage Pools ?**

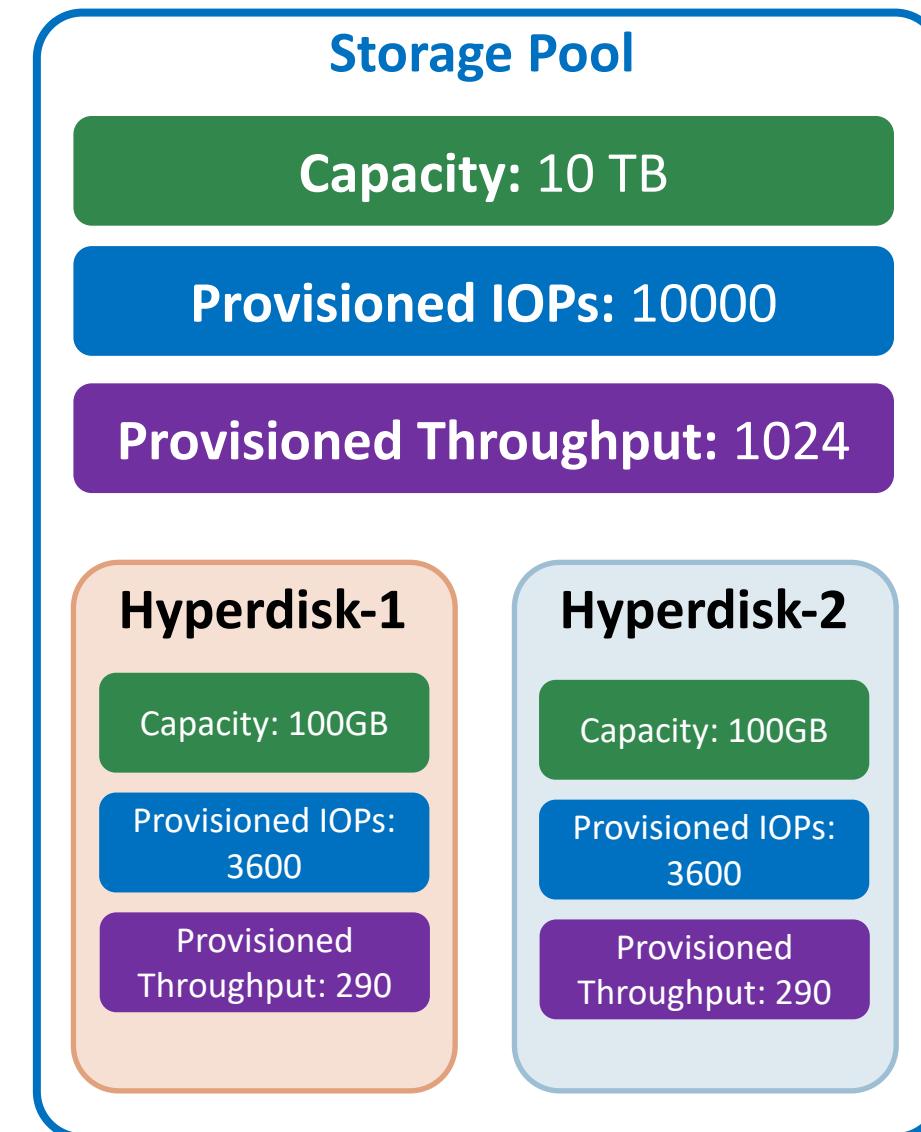
- Difficulty in planning the resource requirements when **migrating workloads from on-premise workloads** that use a **SAN** to Google Cloud

- **Underutilization of Block storage resources**

- Block Storage primarily provisioned for **peak usage but never used**
- Storage pool takes care of **capacity requirements**
- When the storage pool usage reaches 80%, it will attempt to automatically add more capacity to ensure utilization of disk is always less than 80%

- **Complex management**

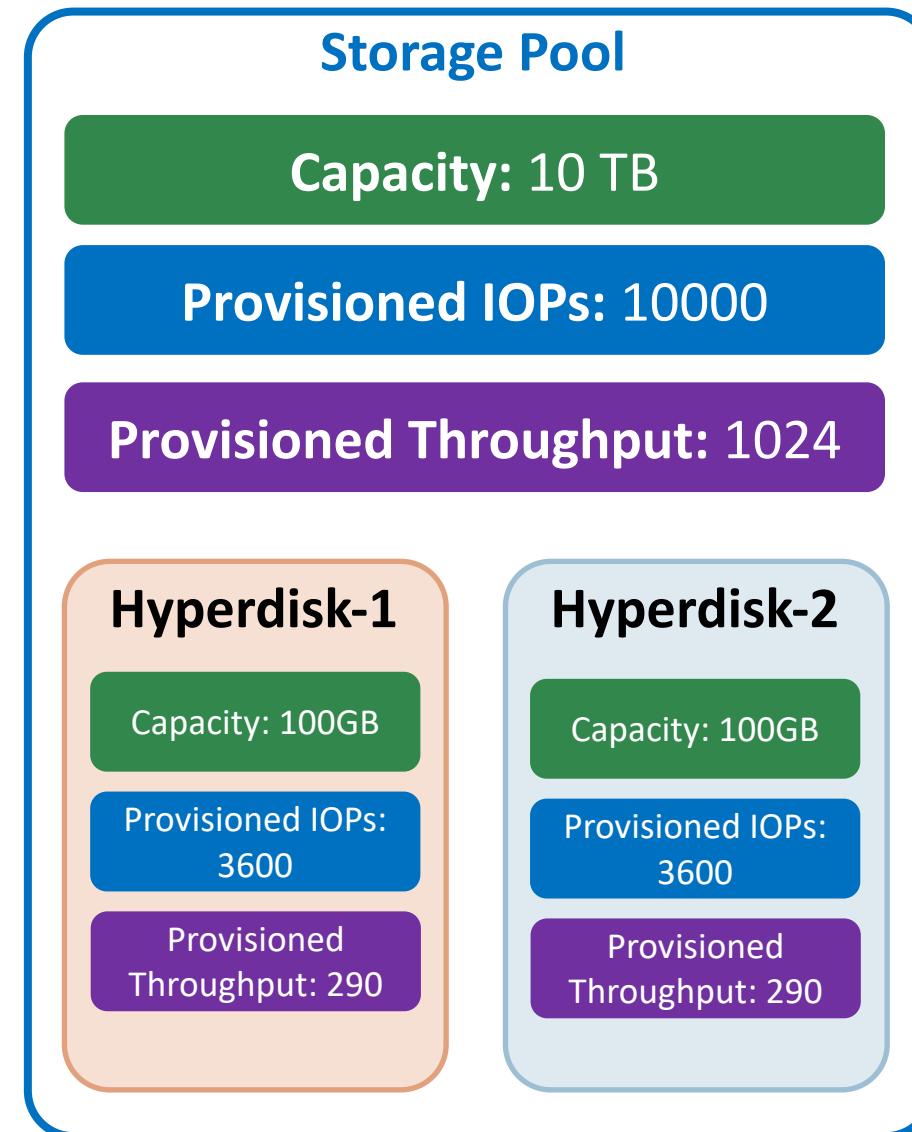
- Managing 100s or 1000s of **disks manually** is time consuming
- Recommended to use Storage pools



Compute Engine - Storage Pool Limitations

- **Storage Pool Limitations**

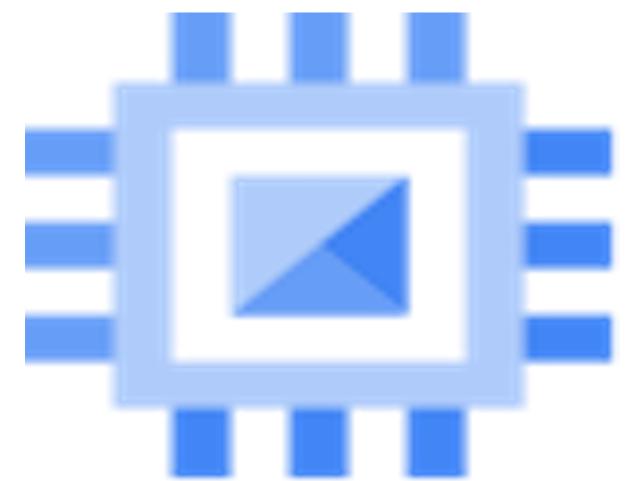
- Storage pools are a **zonal resource**.
- You can use Hyperdisk Storage Pools with only **Compute Engine**. Cloud SQL instances **cannot use** Hyperdisk Storage Pools
- You can create a Hyperdisk Storage Pool with up to **1 PiB** of provisioned capacity.
- You can create up to **1,000 disks** in a storage pool.
- You **can't** create **regional disks** in a storage pool.
- You **can't clone, create instant snapshots** in a storage pool
- Only new disks in the **same project and zone** can be created in a storage pool
- **Complete Limitations Reference:**
https://cloud.google.com/compute/docs/disks/storage-pools#sp_limitations



Demo



Google Compute Engine Disk Images



Compute Engine Storage - Images

- **What is an Image ?**
 - An image is a **replica of a disk** that contains the applications and operating system needed to start a VM
- **How many types of Images available in GCP GCE ?**
 - Public Images
 - Custom Images
- **What are Public Images?**
 - Provided and **maintained by** Google, open-source communities, and third-party vendors
 - By default, **all Google Cloud projects have access** to these images and can use them to **create** VM instances.
 - **Pricing:** Most of them are **free to use (no cost)**, some premium images that do add **additional cost**

Compute Engine Storage - Custom Images

- **What are Custom Images ?**

- You can create custom images from [other boot disks](#)
- You can also create custom images from other public images listed in gcp ([only can be done using gcloud command line and api](#))
- Custom Images are available [only to your](#) cloud project
- **Pricing:**
 - Custom Images imported to GCP adds [no cost](#) for the image.
 - Custom Images do incur an [image storage charge](#) while you keep your custom image in your project.

- **Why do we use custom Images ?**

- To import a boot disk image from our [on-premise](#) environment to GCP
- Create an image from the [boot disks of existing GCE VM Instances](#) ([Pre-configured VM with all application software](#)). Then use that [disk image](#) to create new boot disks for new VM Instances.

Compute Engine Storage - Image Families

- **What are Image Families ?**

- Image families are used to **simplify image versioning**
- Public Images are by default **grouped** into image families
- The image family **always points to the most recent image** in that family, so your instance templates and scripts can use that image **without having to update references** to a specific image version

- **What is the advantages of using Image Families ?**

- Image **Versioning**
- Image **grouping** for the similar type of images (java-apps, webserver-apps)
- **Rollback** to previous version if the latest has any issues

- **When to use Image Families ?**

- If we have **regular updates (more frequent)** to our custom images with newer configurations, it is recommended to use Image Families concept.

Status	Name	Location	Archive size	Disk size	Created by	Family	Architecture
<input checked="" type="checkbox"/>	ubuntu-2004-focal-arm64-v20240307b	asia, eu, us	—	10 GB	Canonical	ubuntu-2004-lts-arm64	Arm64
<input checked="" type="checkbox"/>	ubuntu-2004-focal-v20240307b	asia, eu, us	—	10 GB	Canonical	ubuntu-2004-lts	x86/64

Compute Engine Storage - Image Deprecation States

- **ACTIVE**

- The Image is **active** and can be used as normal
- Image Families point to the **most recent** and active image in a family

- **DEPRECATED**

- The Image is marked as deprecated **but still be used** to create a VM Instance
- Image families **no longer point to this image** even if it is the most recent image in the family.

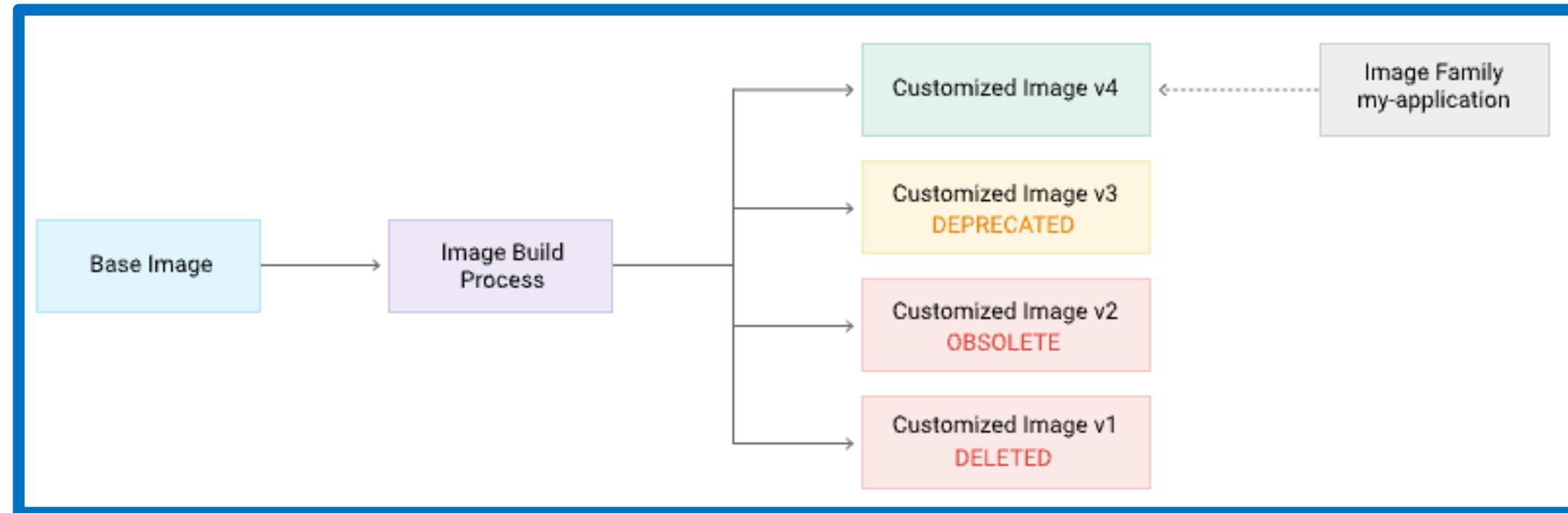
- **OBSOLETE**

- The image is marked obsolete and is **no longer available** for use.
- An **error message** is returned if you try to use this image in a request.

- **DELETED**

- This image is deleted. An **error message** is returned if you try to use a deleted image.

Compute Engine Storage – Image Families

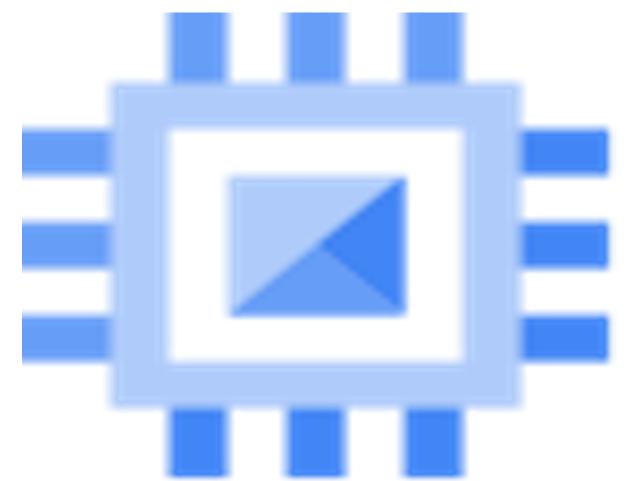


Reference: <https://cloud.google.com/compute/docs/images/image-families-best-practices>

Demo



Google Compute Engine Persistent Disk Snapshots



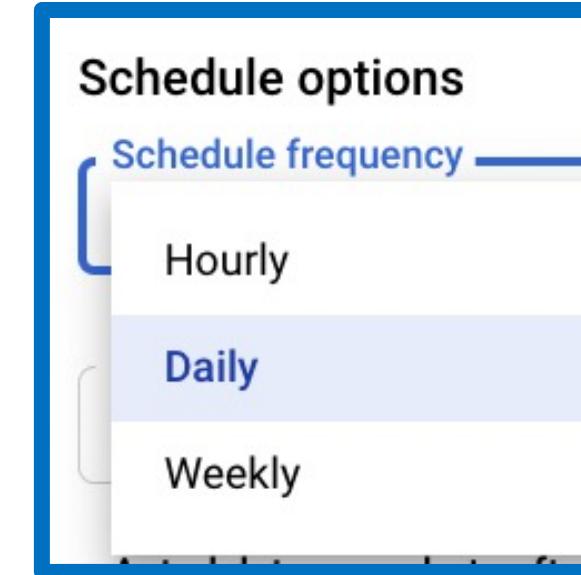
Compute Engine Storage - Persistent Disk Snapshots

- **Why are Persistent Disk Snapshots required ?**
 - Snapshots are required to **periodically backup data** from your zonal or regional persistent disks
- **Can we create Snapshots when the VMs are in running State ?**
 - Yes, we create snapshots from disks even while they are **attached to running VM Instances**
- **Are Snapshots Multi-regional ?**
 - Snapshots can be **Regional and Multi-regional**
 - Snapshots are **global resources** so you can use them to **restore data** to a new disk or instance **within** the same project
 - You can also **share snapshots** across projects

Compute Engine Storage - Persistent Disk Snapshots

- **How frequent we can take Snapshots ?**

- **RECOMMENDED:** Take snapshots **once an hour**
- **SUPER BEST RECOMMENDED:** If there is no need for an hourly snapshots, plan for taking snapshots during **non-business hours** (once or twice per day). Why?
 - Even though disk volume is available during the snapshot creation time, there will be a **slighter performance degradation**.
- **Snapshot Schedules:** We can create a Snapshot Schedule at **hourly, daily and weekly**.



- **Are Snapshots incremental ?**

- **Yes.** We don't lose data by deleting older snapshots
- **RECOMMENDED:** **Delete** older snapshots
- Configure **Deletion Rule** to **delete older snapshots** using **Snapshot schedules**

Deletion rule 

After you delete the disk that uses this schedule:

- Keep snapshots
- Delete snapshots older than 14 days

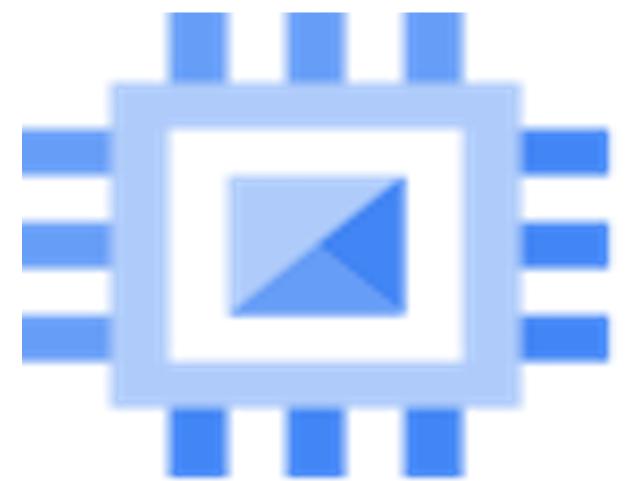
Compute Engine Storage - Persistent Disk Snapshots

- What are advantages of Incremental Snapshots ?
- Snapshots are **incremental by default** which helps us
 - To **avoid billing** you for redundant data
 - To **minimize** use of storage space
 - To **decrease** snapshot creation latency
 - However, to ensure the **reliability of snapshot history**, a snapshot might occasionally capture a full image of the disk.
- If you are **repeatedly creating persistent disk from snapshots**
 - Create an **image from snapshot** and use the image to create VMs.
 - This approach will be **faster** and saves **networking costs**.

Demo



Google Compute Engine Local SSD



Compute Engine Storage - Local SSD

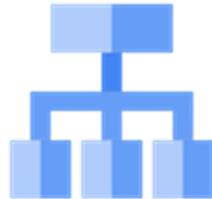
- Local SSDs **physically attached to HOST SERVER** of VM Instance.
- **Data Encryption**
 - Compute Engine **automatically encrypts** your data when it is written to local SSD
 - Local SSDs **cannot use** encryption keys (CMEK or CSEK)
- **Machine Types**
 - Only **few machine types** support Local SSDs
- **Performance**
 - Provides very **high IOPS** and **low latency**
 - Higher throughput (**10x – 100x faster** when compared to PDs)
 - Performance depends on which **interface** you select
 - RECOMMENDED to use **SCSI** and **NVMe** interfaces
 - Use a **NVMe-enabled** or multi-queue **SCSI** VM images for better performance

Compute Engine Storage - Local SSD

- **Data Persistence**
 - Data persists only until instance is **running**
 - When VM Instance is **deleted**, Local SSDs will get **deleted**.
 - Enable **live migration** for data to survive **regular maintenance events**
 - You **CANNOT** detach and attach it to another VM Instance
- **When compared to PDs**
 - Local SSDs will have **lower durability, lower availability and lower flexibility**.
- **Use cases**
 - Caching Services
 - Temporary data that needs **high throughput**

Persistent Disk vs Local SSDs

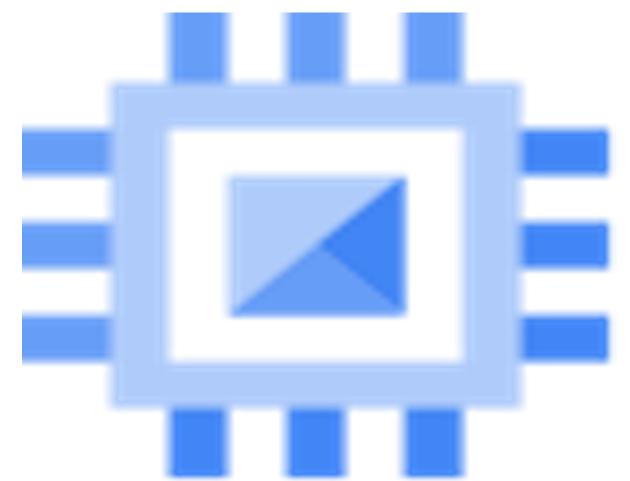
Feature	Persistent Disks (PDs)	Local SSDs
Storage Type	Permanent Storage	Ephemeral Storage (Temporary)
Encryption	GMEK, CMEK, CSEK	Automatically Encrypted
Machine Types	All Machine Types	Only some Machine Types
Performance	Good (adds network latency)	10 – 100x when compared to PDs
Lifecycle	Can attach /detach from one VM to other VM	Tied to only 1 VM (cannot detach and attach)
How will it be attached to VM Instance ?	Attached as Network Drive	Physically attached to HOST Server of VM Instance
Disk Snapshots	Supported	Not Supported



Cloud
Load Balancing



Google Compute Engine Instance Groups



Compute Engine - Instance Groups

- **What is an Instance Group ?**

- **Group of VM Instances** managed as a single entity is called Instance Group

- **How many types of Instance Groups available in GCP ?**

- **Unmanaged** Instance Groups
- Managed Instance Groups(MIG) - **Stateless**
- Managed Instance Groups(MIG) - **Stateful**

- **What is a Zonal MIG ?**

- VM Instances created by MIG will be restricted to specific **SINGLE ZONE**

- **What is a Regional MIG?**

- VM Instances created by MIG will be distributed across **selected Zones**
- Regional MIG gives **HIGH AVAILABILITY (RECOMMENDED)**

Zonal MIG

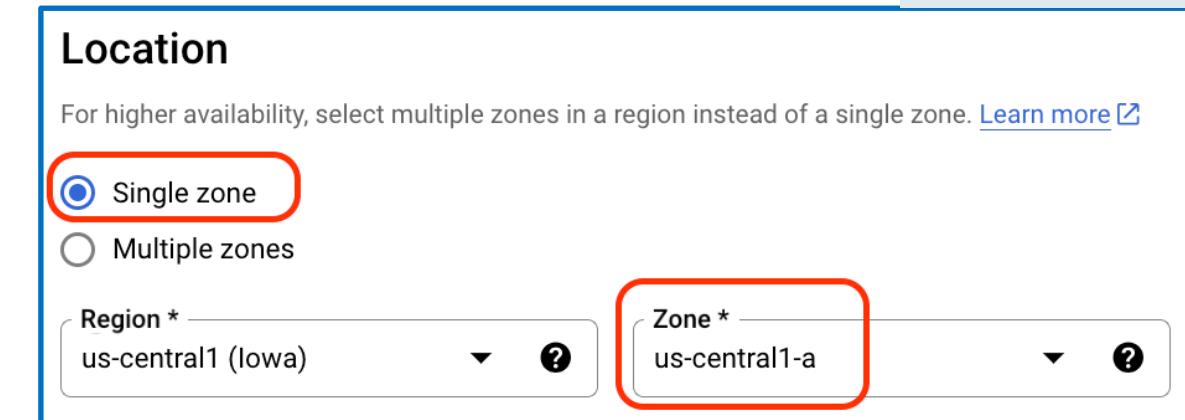
Location

For higher availability, select multiple zones in a region instead of a single zone. [Learn more ↗](#)

Single zone
 Multiple zones

Region * us-central1 (Iowa) ?

Zone * us-central1-a ?



Regional MIG

Location

For higher availability, select multiple zones in a region instead of a single zone. [Learn more ↗](#)

Single zone
 Multiple zones

Region * us-central1 (Iowa) ?

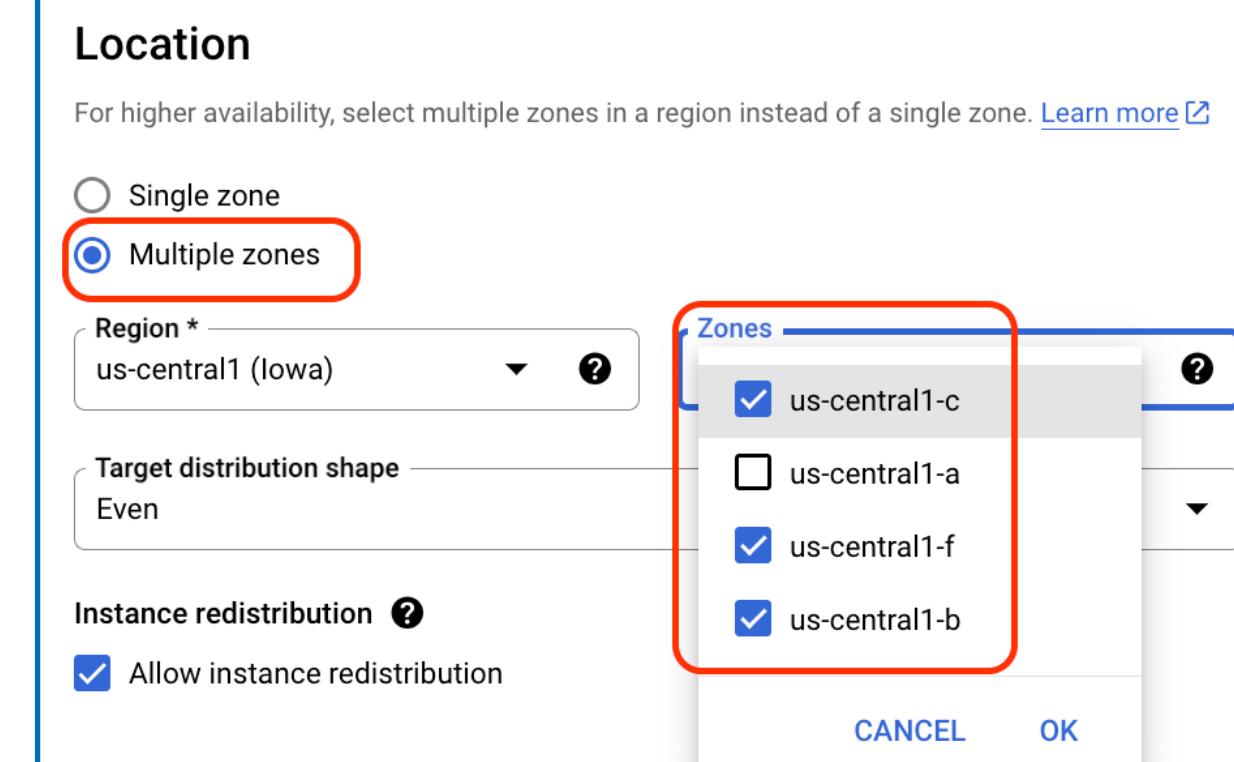
Target distribution shape Even

Instance redistribution ?
 Allow instance redistribution

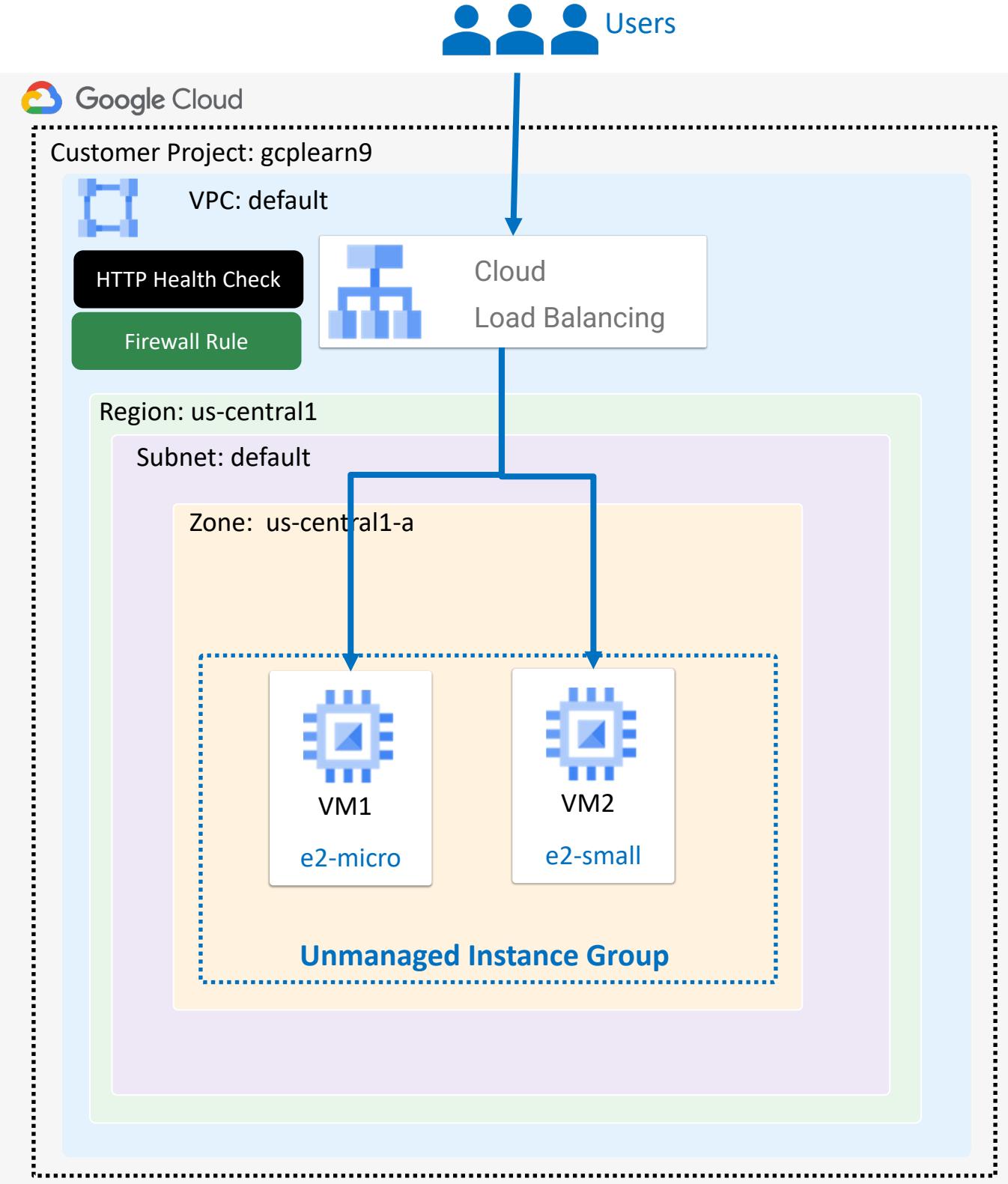
Zones

<input checked="" type="checkbox"/> us-central1-c	?
<input type="checkbox"/> us-central1-a	?
<input checked="" type="checkbox"/> us-central1-f	?
<input checked="" type="checkbox"/> us-central1-b	?

CANCEL OK



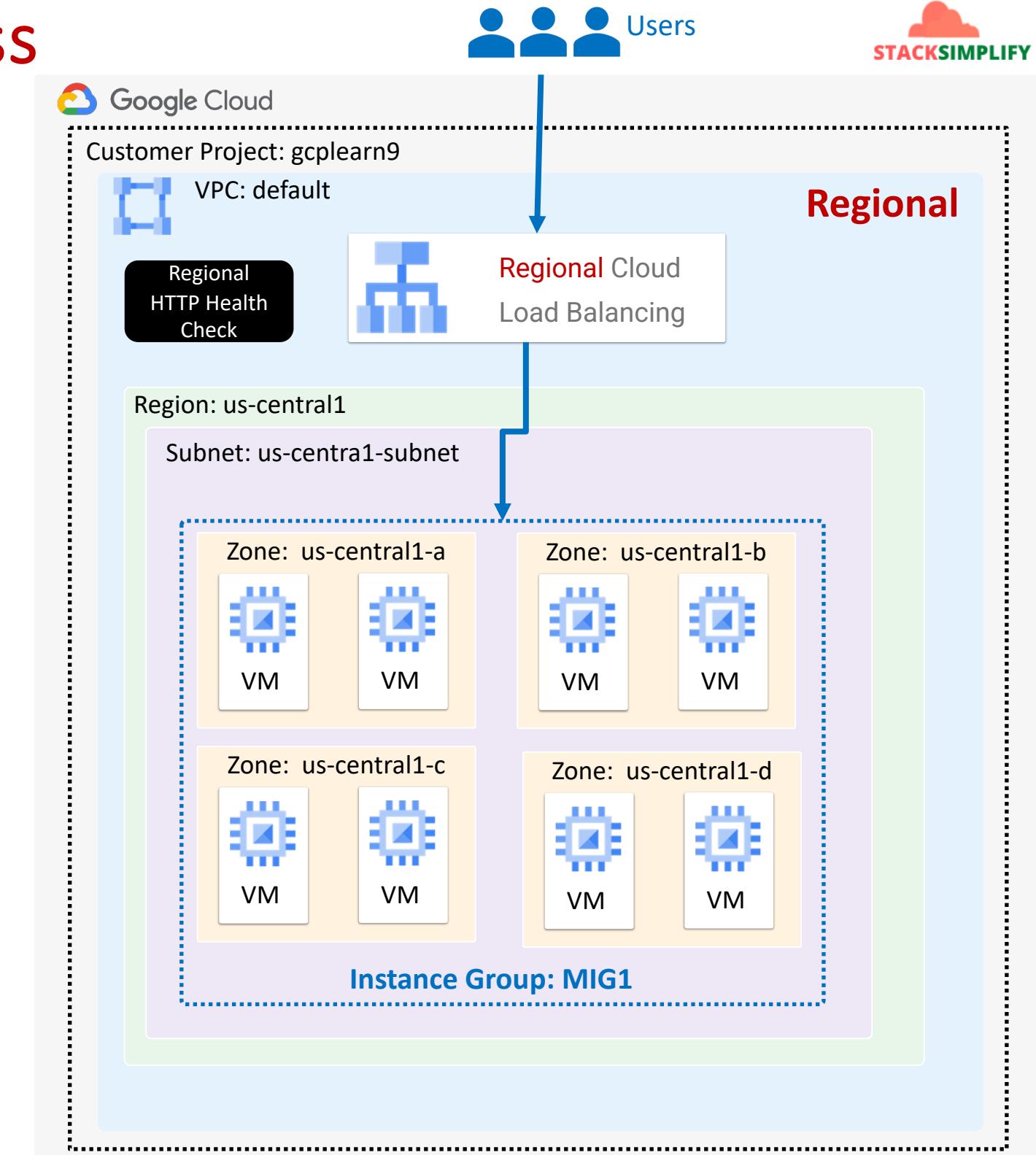
Unmanaged Instance Groups



- Non-Identical VMs can be part of this group (different VM configurations)
Example: e2-small and e2-micor
- Supports Load Balancing
- Autoscaling, auto-healing, auto-updating and multi-zone deployments are **not supported**
- Instance Template **not required**
- **NOT RECOMMENDED** unless we want to maintain **non-identical VMs** in a group or maintain VMs **ourselves**

Managed Instance Groups - Stateless

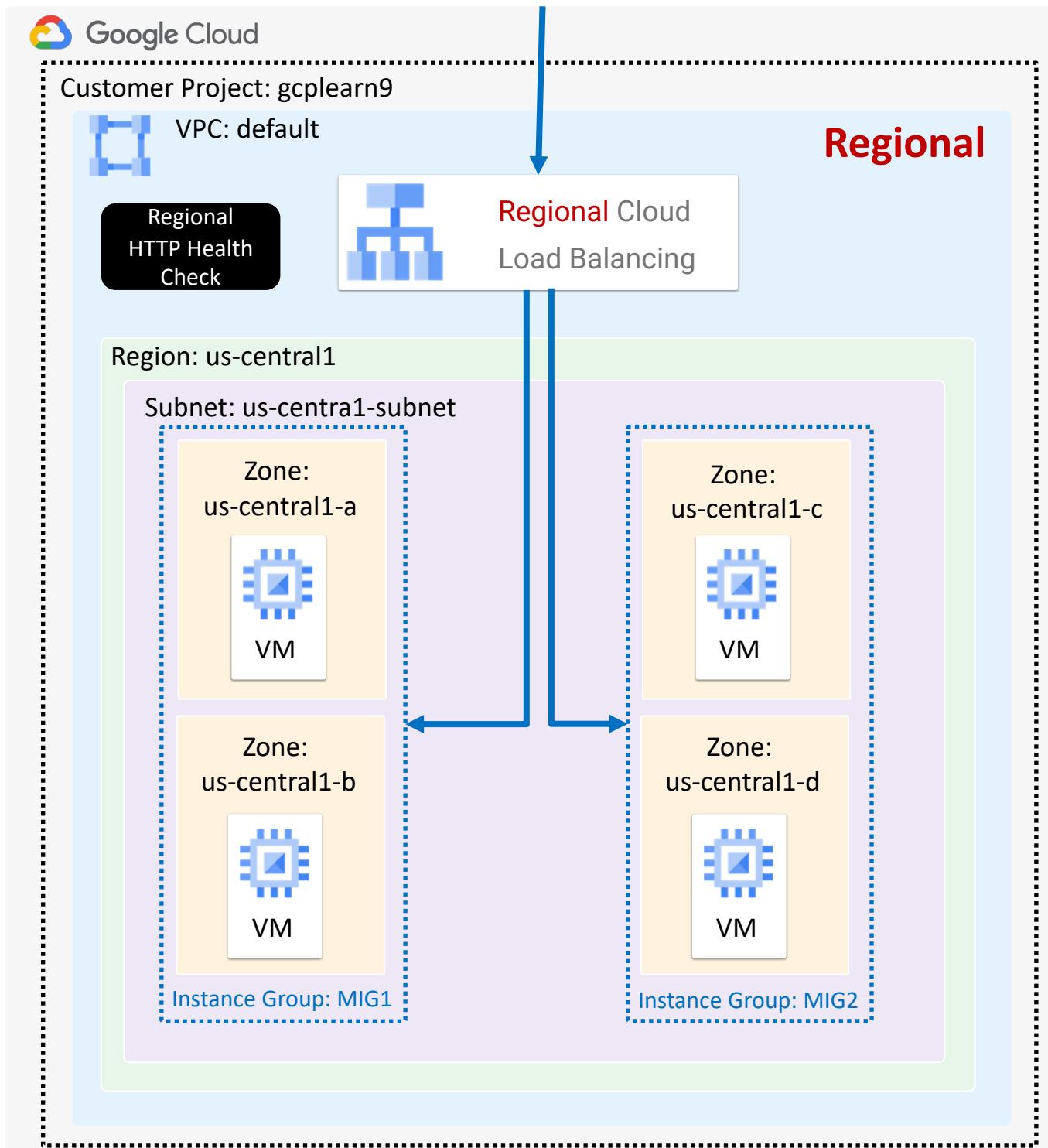
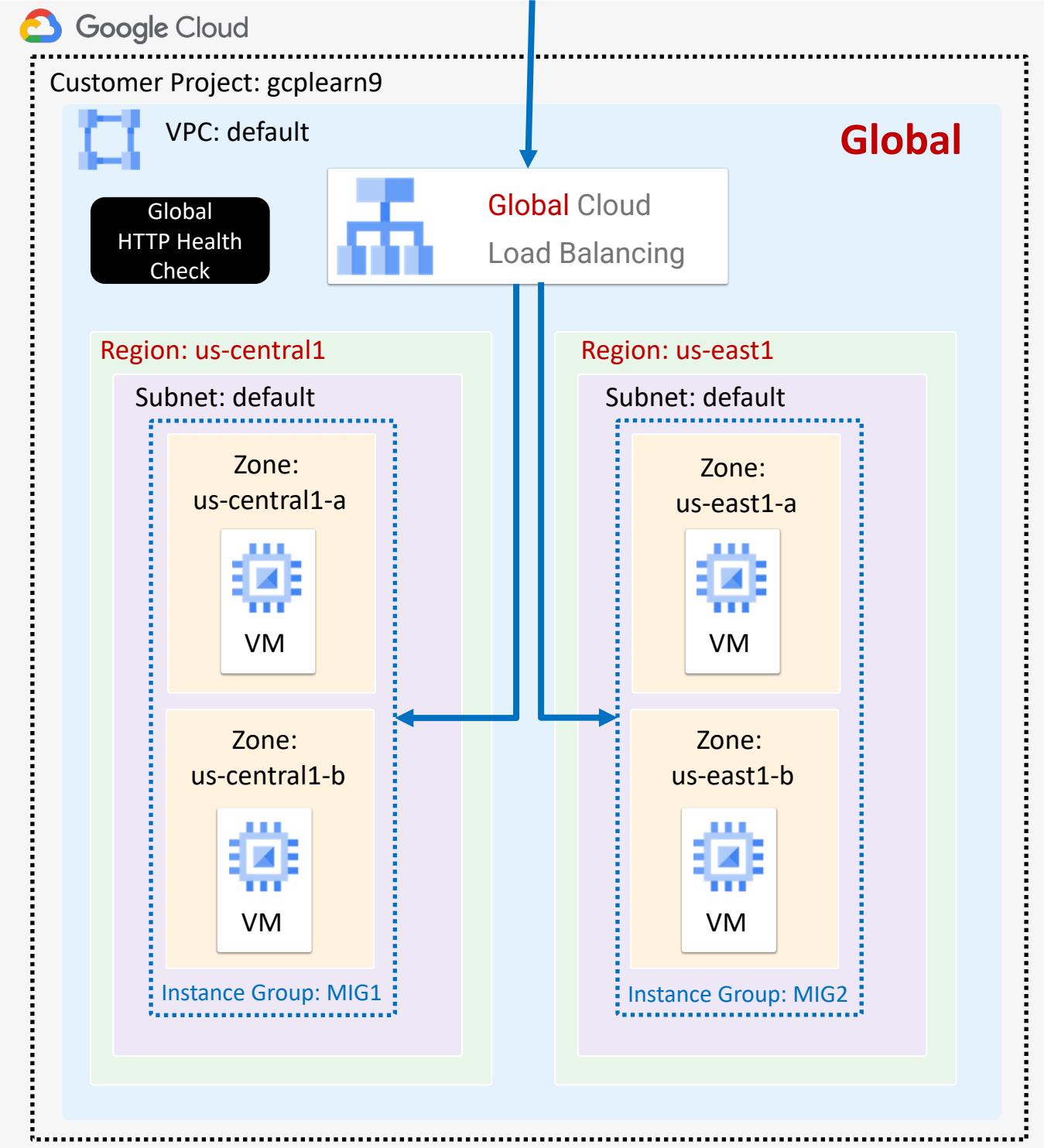
- Identical VMs will be created using Instance Templates
- **Example:** All VM Instances in this group will be of same Instance type (e2-small)
- Instance Template is mandatory
- **Features**
 - Load Balancing
 - Multi-zone deployments
 - Autoscaling
 - Auto-healing (Health Checks)
 - Auto-updating
- **MIG Stateless:** Recommended in 99% of the use-cases





Users

Managed Instance Groups



Compute Engine - Instance Groups Autoscaling

- **What is the core feature of Managed Instance Group(MIG) ?**
 - MIG maintains **configured number of instances** at any point of time
 - If a VM Instance crashes, MIG **automatically** launches a new VM Instance in the Instance Group.
- **What is autoscaling ?**
 - MIGs support autoscaling that **dynamically adds or removes VM instances** from the group in response to increases or decreases **in load**
 - You can configure an **autoscaling policy** to specify how you want to scale the group.
 - In your autoscaling policy, you can set **one or more signals** to scale the group based on **CPU utilization, load balancing capacity, Cloud Monitoring metrics and Schedules**.

Compute Engine - Instance Group Auto-Healing

- **What is auto-healing ?**

- We already know MIG **automatically recreates a new VM Instance** when any of the instance in MIG is not in **RUNNING** state.
- Ideally that is **not sufficient** for maintaining effective high-availability of our application.
- **What happens if application running inside VM Instance freezes, crashes or overloads ?**
 - VM will be in **RUNNING** state but application is having an **issue**
 - At this point, we will have **drop in our High Availability** due to application unavailability and load on other VM Instances increases.
- Application-based **autohealing improves application availability** by relying on a **health checking signal** that detects application-specific issues such as **freezing, crashing, or overloading**.
- If a health check **determines** that an application has **failed** on a VM, the group automatically **recreates** that VM instance

Compute Engine - Instance Group Auto-Updating

- **What is auto-updating ?**

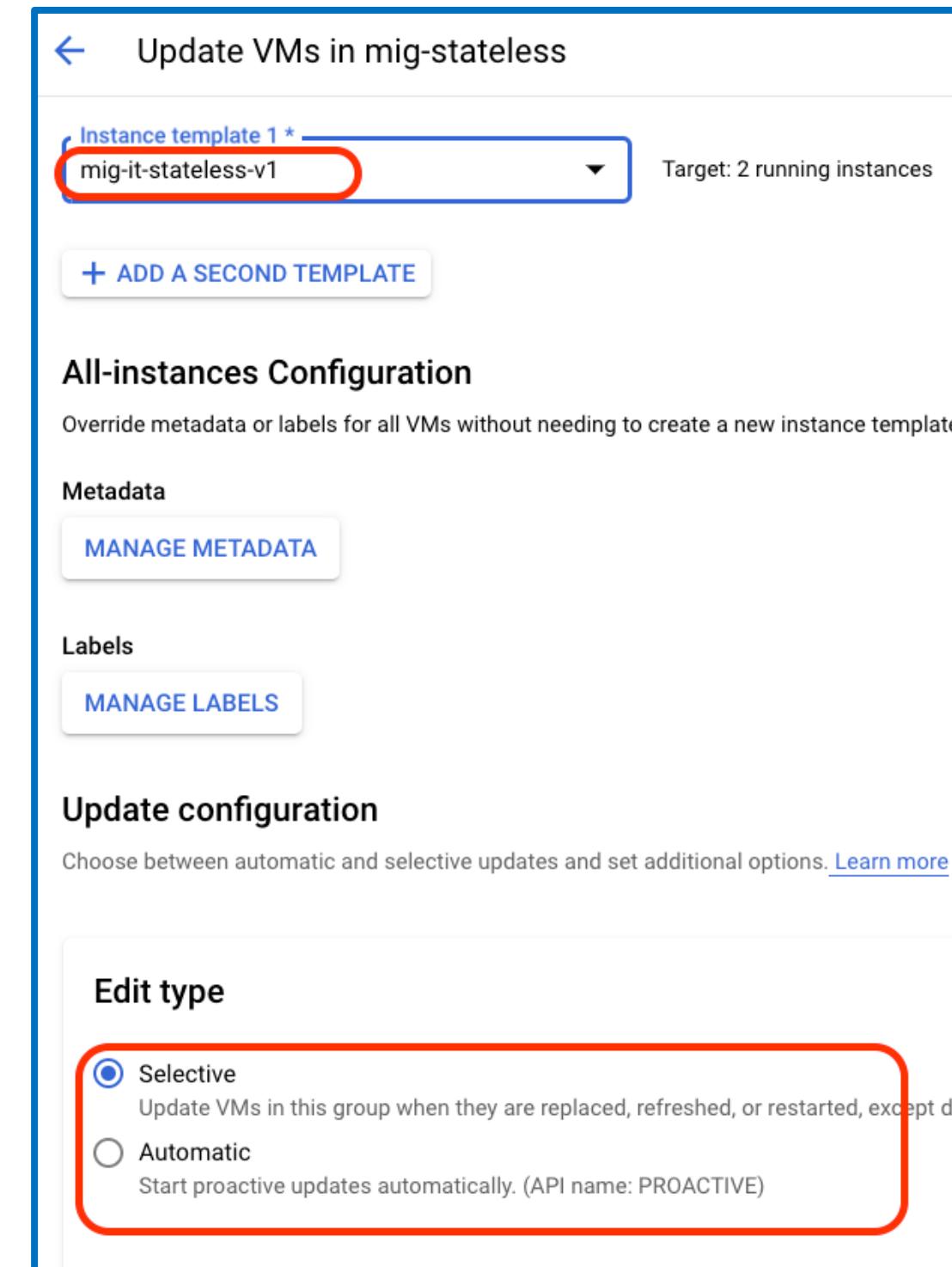
- Helps us to easily and safely deploy **new version of software** to VM Instances in a MIG

- **What are Rolling Updates ?**

- Create **new Instance Template** and update MIG to new instance template
- MIG will **slowly or gradually do the rolling update** of VM Instances to the new Instance Template based on update type selected

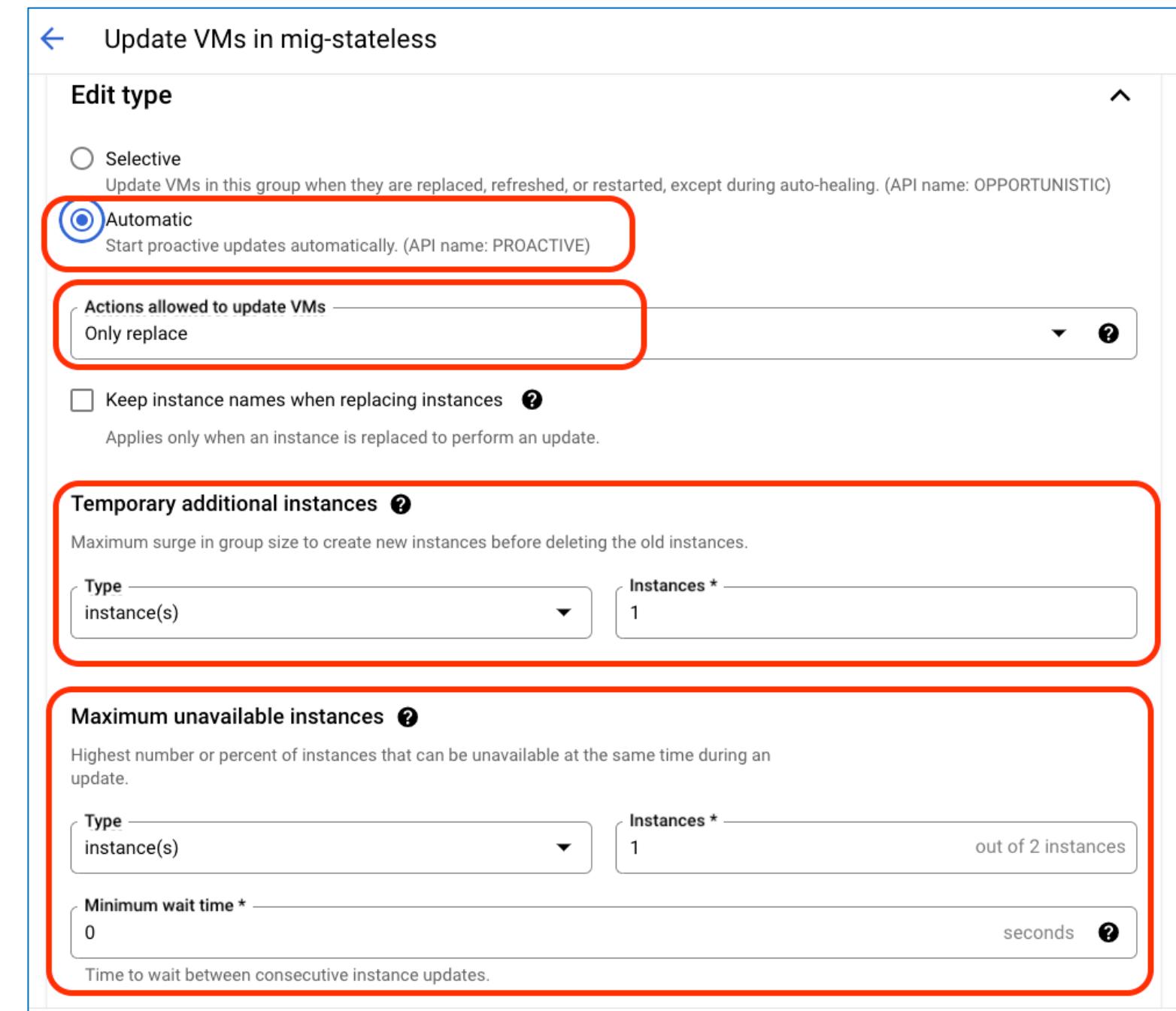
- **When should the VM Instances get updated ?**

- **PROACTIVE (Automatic):** Start updates to **VMs immediately** in the MIG
- **OPPORTUNISTIC (Selective):** Update VMs in MIG when they are **replaced, refreshed or restarted** except during auto-healing use case.



Compute Engine - Instance Groups Auto-Updating

- When Update Type: Automatic
- How should the update happen ?
 - Temporary additional Instances or Maximum Surge: How many temporary instances should be created during deletion of old instances for instance replacement ?
 - Maximum Unavailable: How many instances can be offline at the same time while replacing VM instances ?



Compute Engine - Instance Groups Restart / Repace VMs



- How do the Rolling Restart / Replace VM Instance should happen ?

- **Maximum Surge:** How many **temporary instances** should be creating during the instance replacement due to updates ?
- **Maximum Unavailable:** How many instances can be **offline** at the same time while restarting / replacing VM instances ?

← Restart / replace instances of instance-group-1

Gradual restart or replace of all instances in the group. [Learn more](#)

Current template(s)

demo3-instance-template : 2 instances

Operation

Restart
 Replace
Deletes instances and creates new ones

Maximum surge

Maximum number (or percentage) of temporary instances to add while replacing. [Learn more](#)

instance(s) ▾

Maximum unavailable

Maximum instances (number or percentage) that can be offline at the same time while restarting / replacing. [Learn more](#)

3 instance(s) ▾ out of 2 instances

Minimum wait time

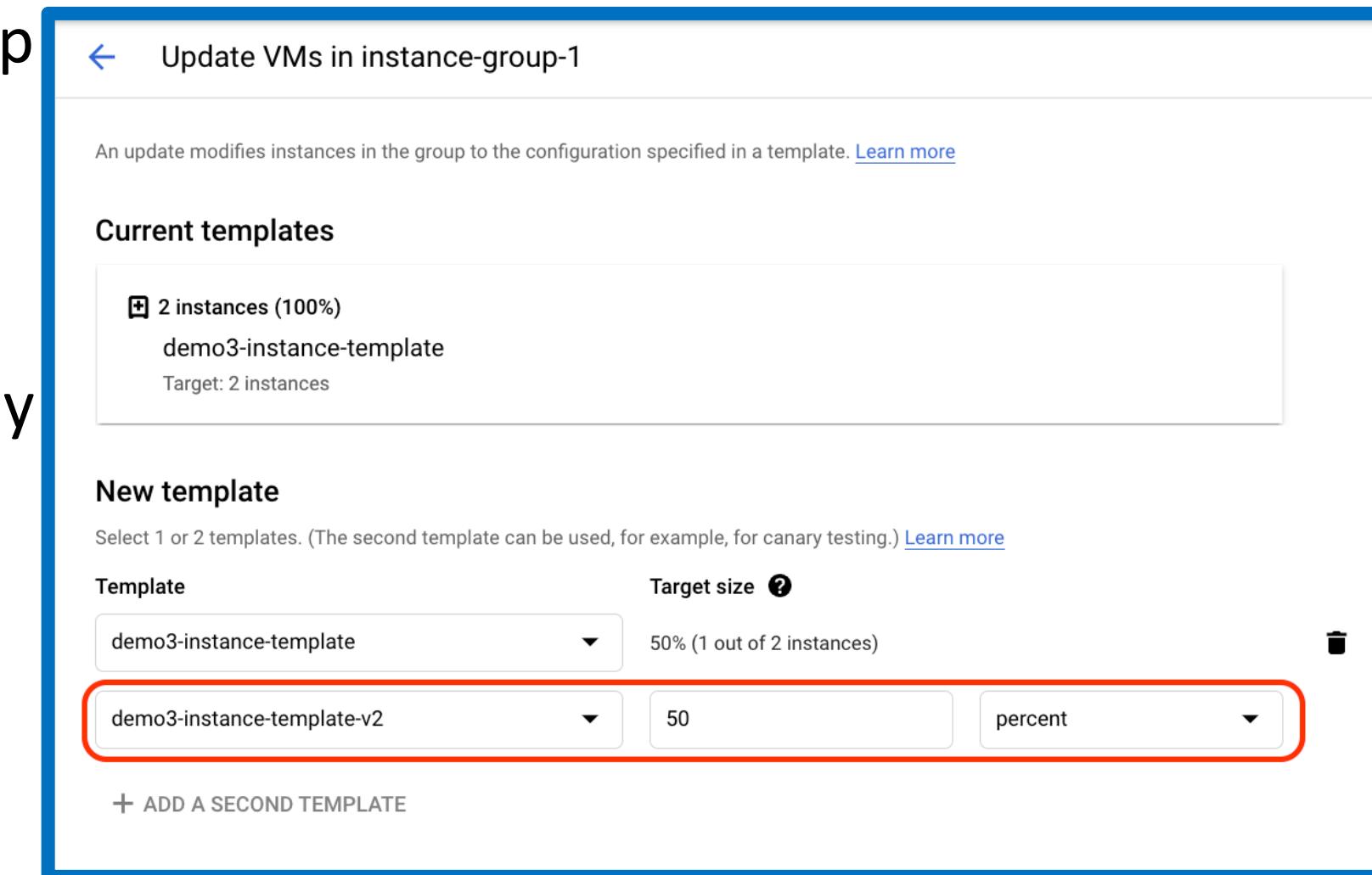
Time to wait between consecutive instance restart / replace operations. [Learn more](#)

0 s

Restart **Cancel**

Compute Engine - Instance Group Canary Updates

- What is Canary Deployment ?
- Test new application version with group of instances before releasing it across all instances.
- In rolling update type, we will completely switch the instance template from old to new and gradually update VMs based on update type
- In canary deployment model, we will add second instance template by specifying Target Size (50% or 1 VM) where new application version should be deployed.



Managed Instance Groups - Stateful

- **MIG Stateful:** can preserve each instance

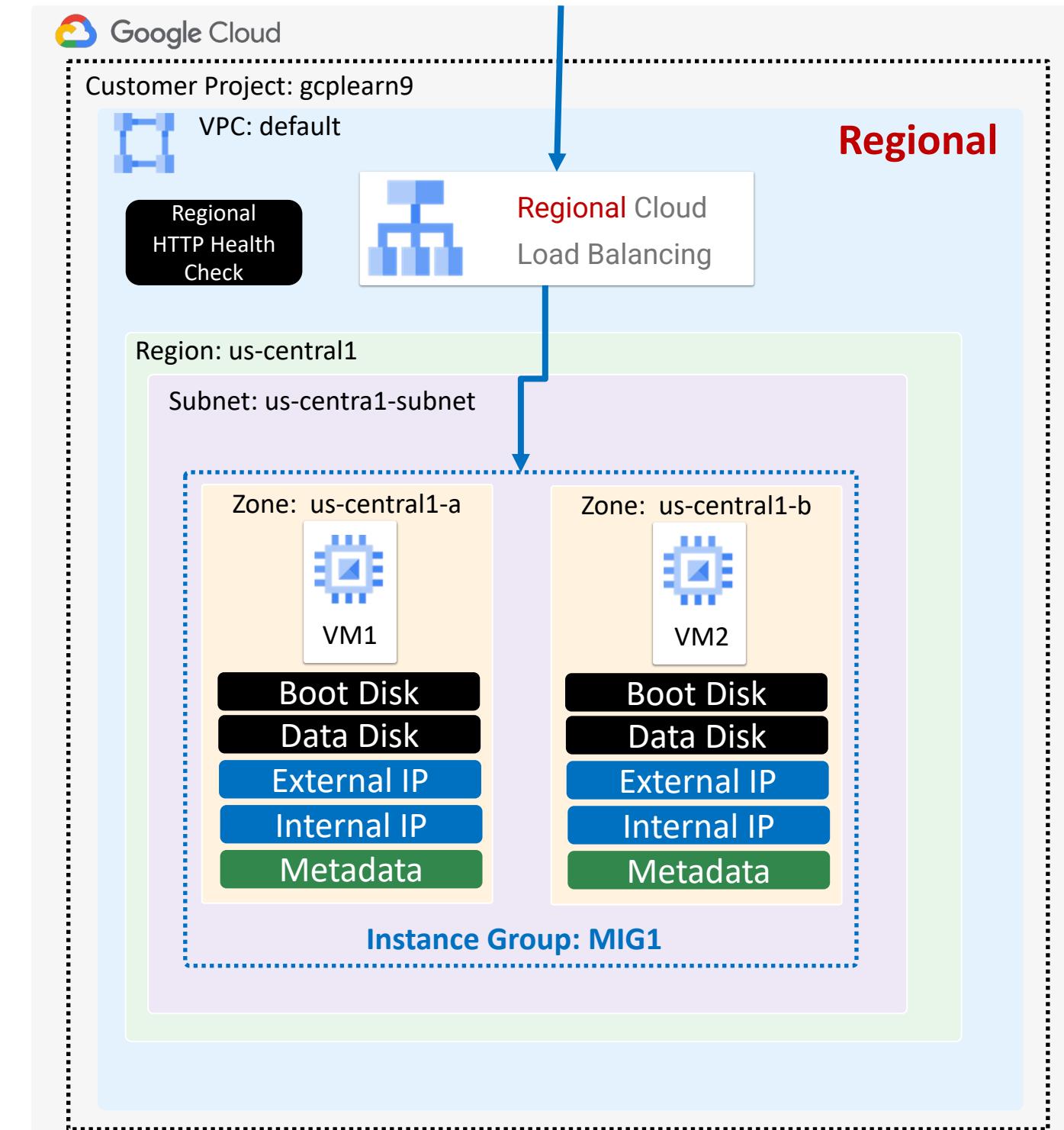
- Unique state
- Persistent Disks
- Instance Metadata
- Customizable Instance names

- **Key Features**

- Autoscaling (NOT SUPPORTED)
- Load Balancing
- Multi-zone deployments
- Auto-healing (Health Checks)
- Auto-updating
- Disk and Metadata Preservation
 - Boot Disks and Data Disks
 - External and Internal IP

- Recommended for applications with **stateful data**

- Databases
- Legacy monolith applications
- Long-running batch computations



Instance Groups

Unmanaged Instance Groups

Non-Identical VMs can be part of this group (different VM configurations) Example: e2-small and e2-medium

Supports Load Balancing

autoscaling, auto-healing, auto-updating and multi-zone deployments **are not supported**

Instance Template **not required**

NOT RECOMMENDED unless we want to maintain non-identical VMs in a group or maintain VMs ourselves

Instance Groups Stateless

These are called **MIGs** (Managed Instance Groups)

Identical VMs will be created using Instance Templates
Example: All VM Instances in this group will be of same Instance type (e2-small)

Supports Load Balancing and multi-zone deployments

Supports autoscaling, auto-healing and auto-updating

Instance Template is **mandatory**

Recommended in **99%** of the use-cases

Autoscaling not supported
Supports auto-healing and auto-updated

Recommended for applications with stateful data or configurations such as databases, legacy monolith applications

GCE Instance Group - Quick Reference

Question

How can applications survive [Zonal failures](#) using Managed Instance Groups ?

How can VM Instances be replaced when they are unhealthy (VM State not in [RUNNING](#) state or [Application](#) in VM is unhealthy)?

How can you preserve the VM State in MIG ?

Answer

Create Regional Managed Instance Group ([Regional MIGs](#) or [Multi-zone MIGs](#))

Configure health checks ([Auto-Healing](#))

Use [Stateful MIG](#) which will preserve VM State (Instance name, persistent disks attached to VM and Metadata)
Recommended for stateful workloads like [databases](#) and [data processing](#) applications.

GCE Instance Group - Quick Reference

Question

How can you group **VMs** of different configurations into an Instance Group ?

In which type of MIGs we use **Instance Template** as a mandatory option?

How do you achieve **high availability** in MIG when there are hardware/software updates from GCP cloud platform perspective ?

What is the **major feature** difference between Stateful-MIG and Stateless-MIG?

Answer

Use **Unmanaged** Instance Groups option

In **both stateful - MIG and stateless - MIG**, Instance templates are mandatory

As usual, use the **instance template** with **availability policy** with options

1. Automatic Restart: Enabled
2. On-Host Maintenance: Migrate

These features ensures live migration and automatic restart of VM Instances

Autoscaling not supported in Stateful – MIGs

Rest all features auto-healing, auto-updating, multi-zone deployments are supported in both type of MIGs

Compute Engine - Instance Groups

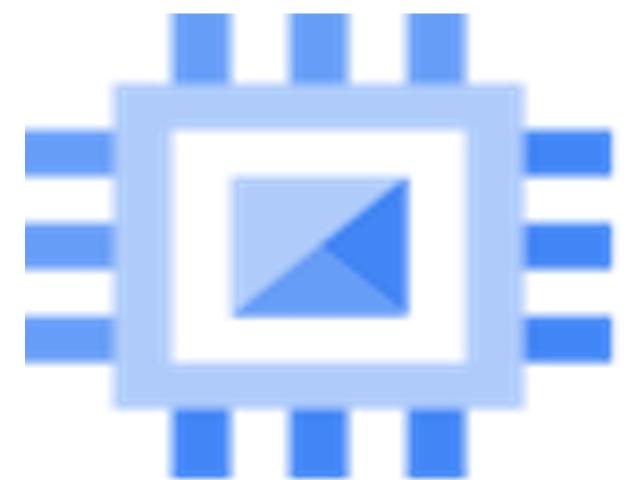
- **Additional Reference**

- <https://cloud.google.com/compute/docs/instance-groups>
- <https://cloud.google.com/compute/docs/instance-groups/creating-groups-of-managed-instances>

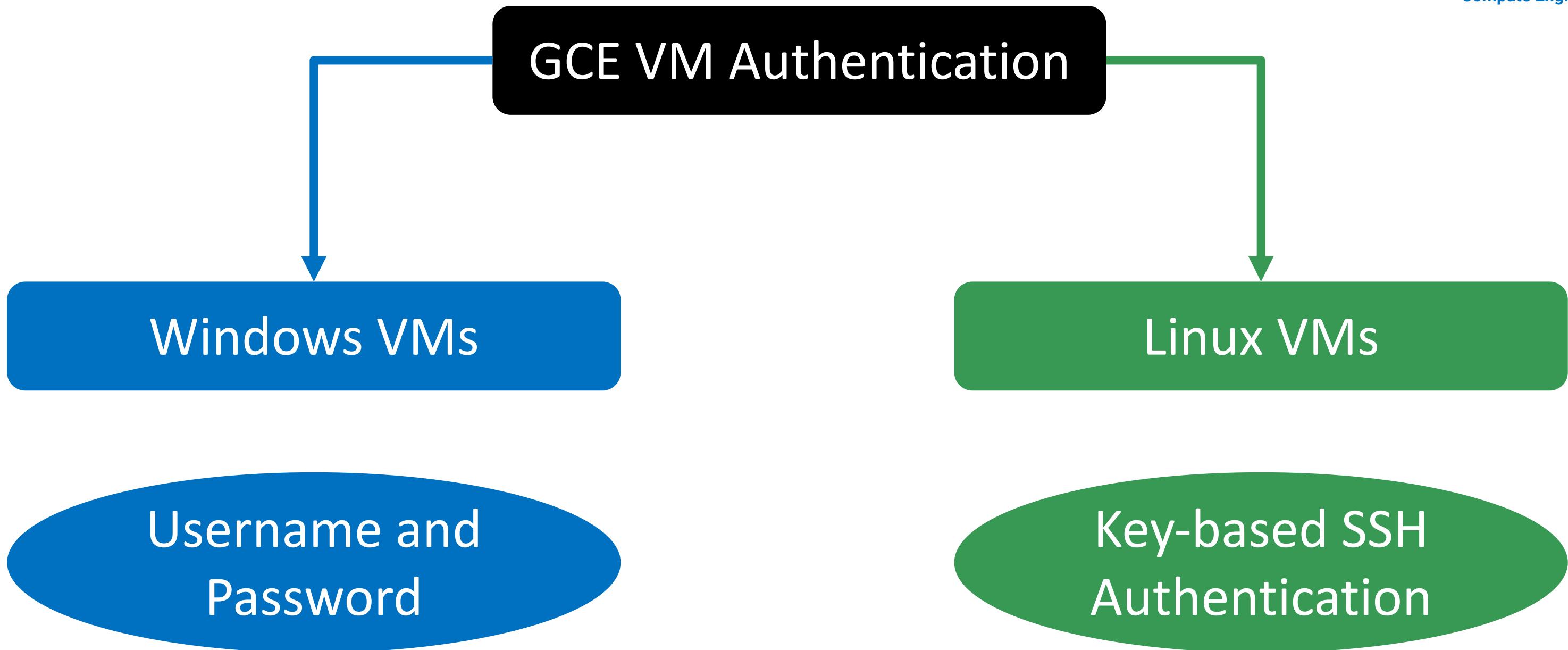
Demo



Google Compute Engine SSH Keys



GCE VM Authentication



GCE Linux VMs - SSH Authentication

Linux VMs Key-based SSH Authentication

Metadata Managed

Manually create and configure SSH Keys or GCE generates Ephemeral Keys

Project-wide public SSH keys

Instance-wide public SSH Keys

OS Login

Manage SSH access without managing individual SSH Keys

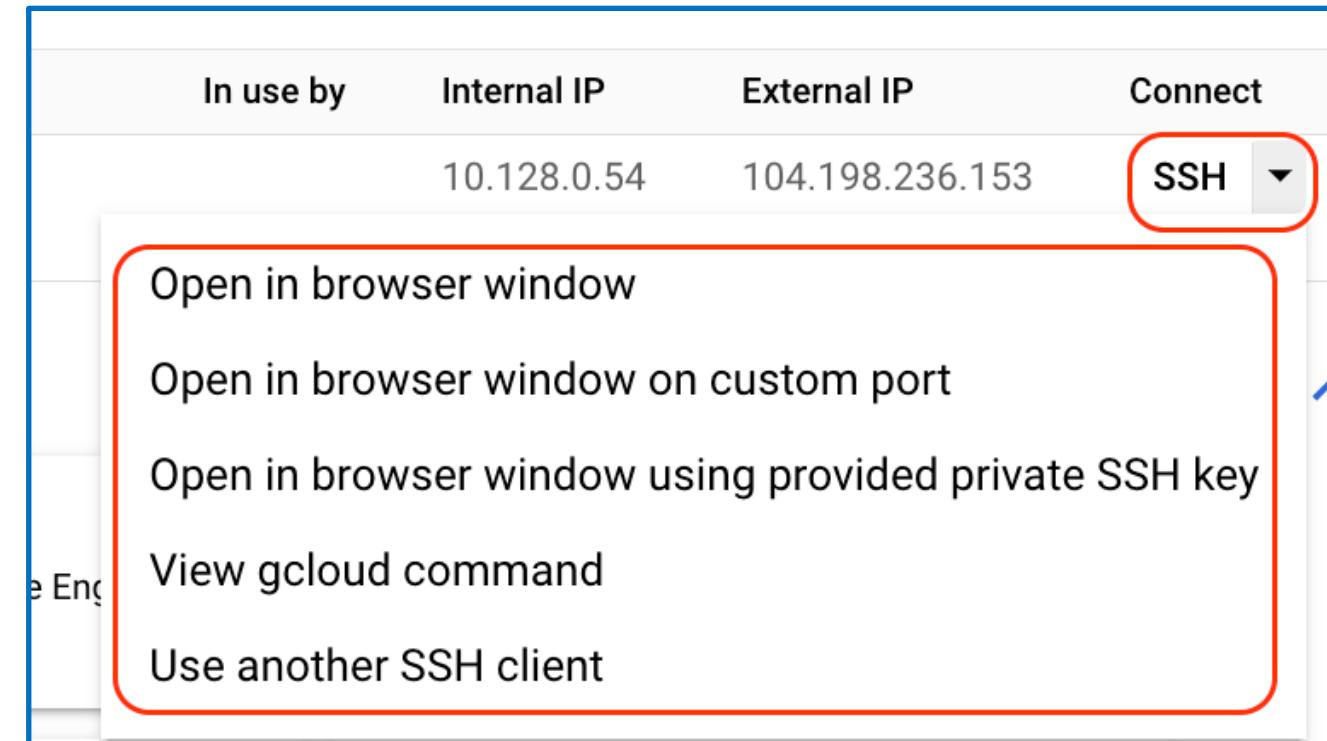
Enable OS Login at project level

Enable OS Login at Instance Level

GCE Linux VMs - SSH Authentication Options

- **Option-1: Google Cloud Console – SSH Button**

- SSH to VM Instance using [Web-based](#) or [browser-based](#)
- Compute Engine generates [Ephemeral SSH keys](#) to SSH to VM
- Your private SSH key is stored in [your browser session](#)
- Google [doesn't have access](#) to your private key



- **Option-2: gcloud cli**

- Compute Engine creates a [username and persistent SSH key pair](#)
- We can [reuse the same SSH key pair](#) for future interactions using gcloud cli

```
# Set GCP Project
gcloud config set project <PROJECT-ID>
gcloud config set project gcplearn9

# SSH to VM using gcloud
gcloud compute ssh --zone <ZONE> <VM-NAME>
gcloud compute ssh --zone "us-central1-a" "vm1"
```

GCE Linux VMs - SSH Authentication Options



- **Option-3: Customized Keys - Metadata managed**

- Generate SSH public and private key using [ssh-keygen](#)
- Upload public SSH key to [project-level](#) metadata or [instance-level](#) ssh keys
- Using the private SSH key on our desktop we can ssh to linux VM using [third-party tools](#) like putty, ssh command etc

Metadata EDIT REFRESH

All instances in this project inherit these SSH keys. [Learn more](#)

METADATA	SSH KEYS
Username ↑	Key
sshcustomuser1	ssh-rsa...

- **Option-4: OS Login managed**

- Set the below Key value pair in Compute Engine Metadata
 - **Key:** enable-oslogin
 - **Value:** TRUE
- We can access the VM instance using
 - [SSH Button](#) using browser
 - Using [gcloud ssh](#) command

Metadata EDIT REFRESH

All instances in this project inherit these key-value pairs.

METADATA	SSH KEYS
Key ↑	Value
enable-oslogin	TRUE

GCE Linux VMs - SSH Authentication Options

- **Option-5: Customized Keys - OS Login managed**
 - Generate SSH public and private key using [ssh-keygen](#)
 - Add public SSH key to the [google cloud account](#)
 - Using the private SSH key on our desktop we can Login (ssh) to linux VM using [third-party tools](#) like putty, ssh command etc

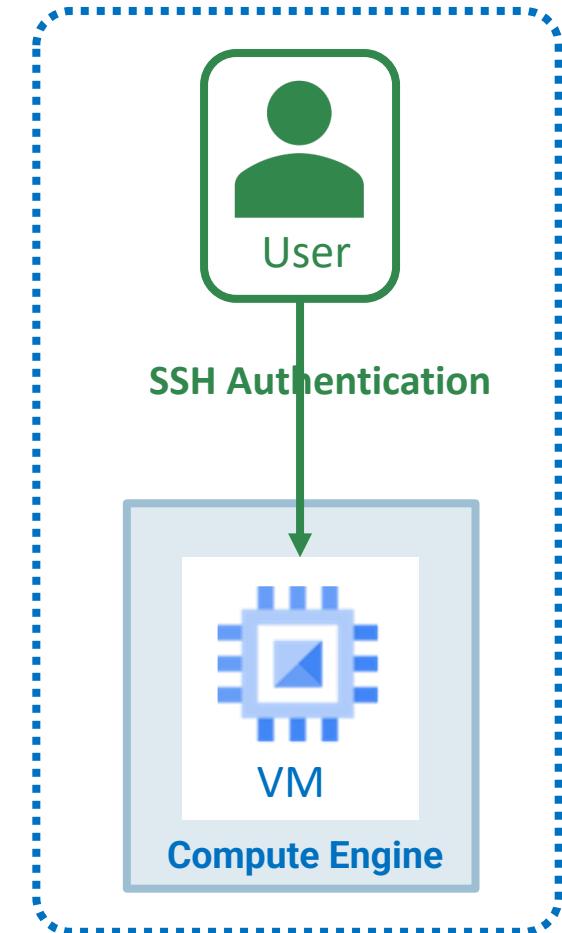
```
# Use the gcloud command-line tool to
gcloud compute os-login ssh-keys add \
    --key-file=KEY_FILE_PATH \
    --ttl=EXPIRE_TIME

gcloud compute os-login ssh-keys add \
    --key-file=ssh-keys-oslogin.pub \
    --ttl=0
```

GCE Linux VMs - SSH Authentication using OS Login

- Why do we need to use OS Login method over Metadata managed for providing access to our Linux VMs ?

- OS Login allows SSH access **without manually managing** SSH Keys
- OS Login is **HIGHLY RECOMMENDED** option for managing access to Linux VM Instances if we need to deal with **huge number of users across multiple instances and google projects.**
- OS Login supports **2-step verification** (Google Authenticator, Text Message, Phone call verification, Phone prompts and Security Key OTP)
- OS Login provides the **ability to import existing Linux user accounts** from on-premise AD or LDAP
- OS Login can be used in combination with **super advanced use cases** like **IAM Organization** (Manage users, groups and centrally control all of your organization's projects and resources)



GCE Linux VMs - SSH Authentication using OS Login

- **Can we separate user and admin access to Linux VMs using OS Login ?**

- **Yes.** We can do that based on roles associated to the user.

- **What roles user need to have for using OS Login SSH Authentication ?**

- **For Normal User:**
roles/compute.osLogin
- **For Admin User:**
roles/compute.osAdminLogin

ID	roles/compute.osLogin
Role launch stage	General Availability
Description	
Access to log in to a Compute Engine instance as a standard (non-administrator) user.	
18 assigned permissions	
compute.disks.listEffectiveTags compute.disks.listTagBindings	

ID	roles/compute.osAdminLogin
Role launch stage	General Availability
Description	
Access to log in to a Compute Engine instance as an administrator user.	
19 assigned permissions	
compute.disks.listEffectiveTags compute.disks.listTagBindings	

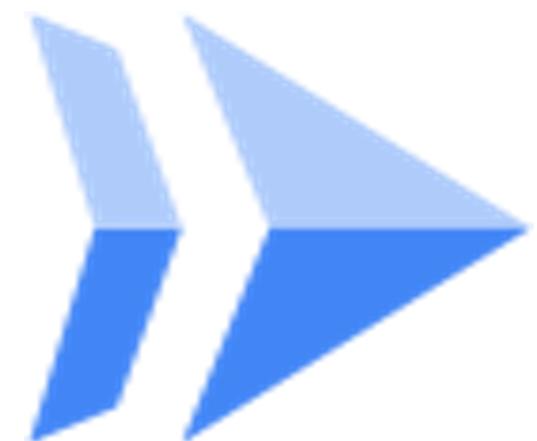


Demo



Google Serverless Cloud Run

Services and Jobs



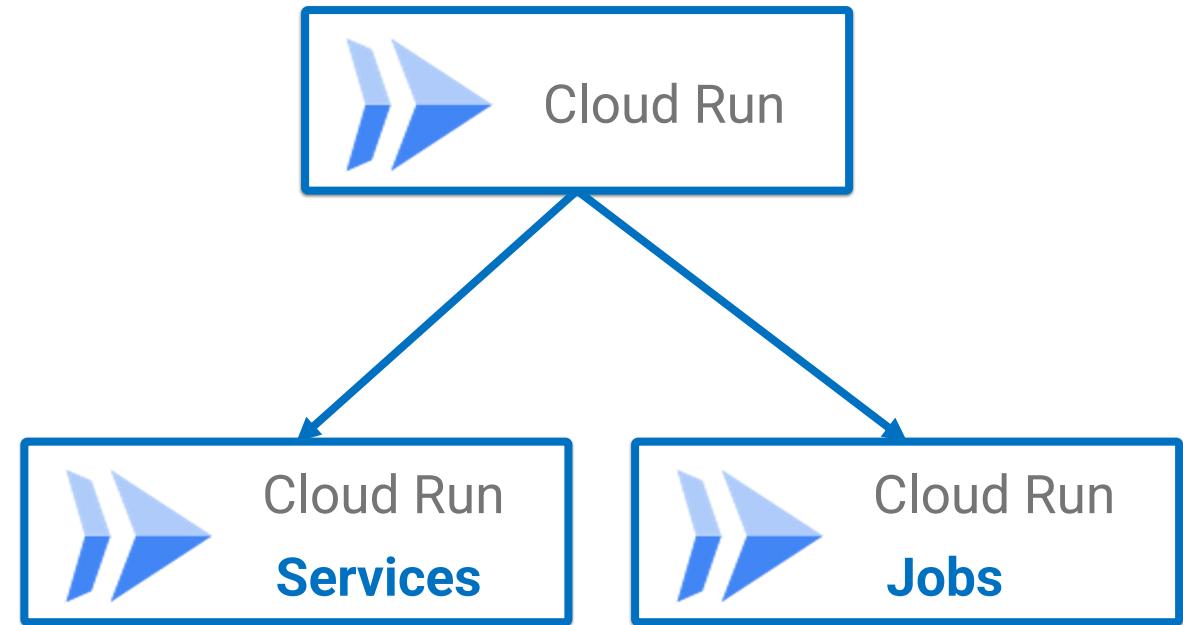
Google Cloud Run

- Cloud Run is a **Serverless** Container Platform
- **Fully managed** Compute Platform
 - Used to run **containers** directly
 - NO **manual** infrastructure management required.
 - NO **visibility** to underlying vm instances
- Any language, any library, any binary supported to run on Cloud Run
- **Fully Integrated** with other Google cloud services to build featured applications (Cloud SQL, Cloud Build, Cloud Logging, Cloud Monitoring, Firebase, Cloud Load Balancing, Cloud Memory Store, Secret Manager, VPC Private Networking, Cloud Tasks)



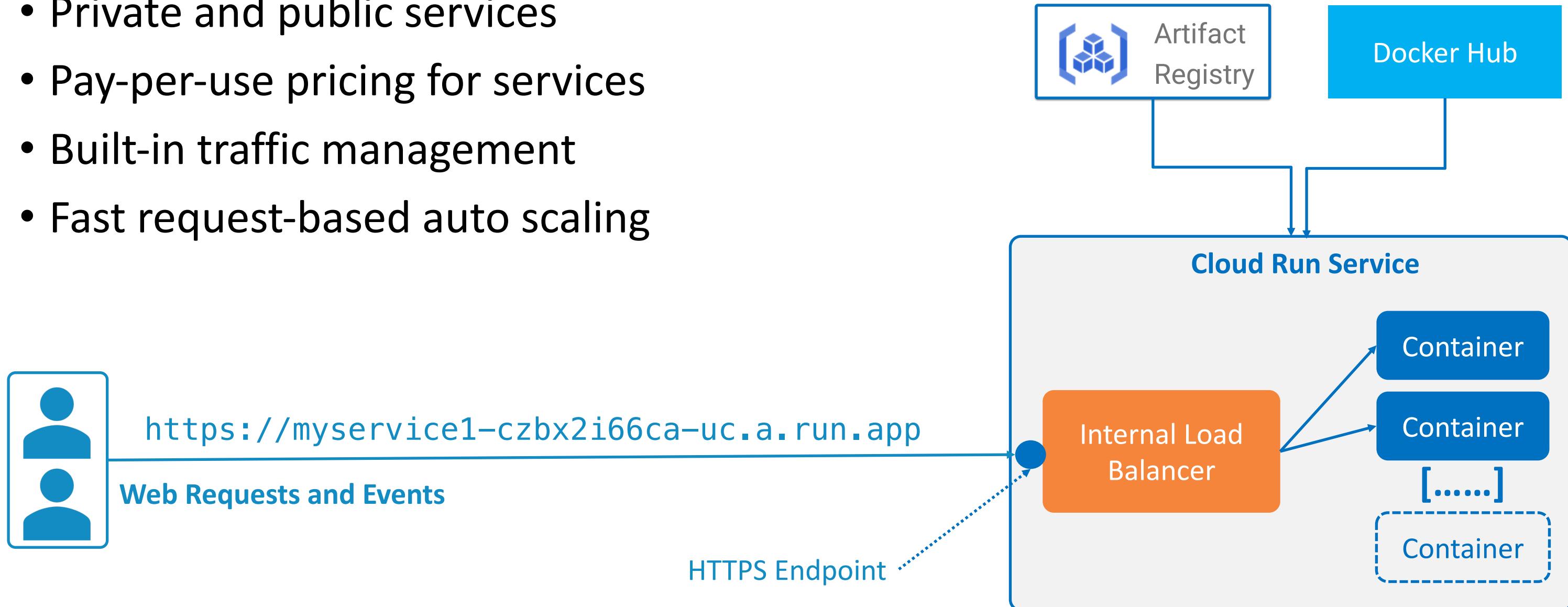
Google Cloud Run

- We can run code on Cloud Run in two ways
- **Cloud Run Services:**
 - Deploy/Run code **that responds** to web requests or events
 - **When to use**
 - Websites and Web Applications
 - APIs and Microservices (supports HTTP and gRPC protocols)
 - Streaming data (Receive events from Eventarc or messages from Pub/Sub Subscriptions)
- **Cloud Run Jobs:**
 - Deploy/Run code that performs work **(a job)** and quits when the work is done
 - In short, our container will execute job and **runs to completion.**
 - **When to use:** Database migrations, scheduled jobs, parallel processing of tasks



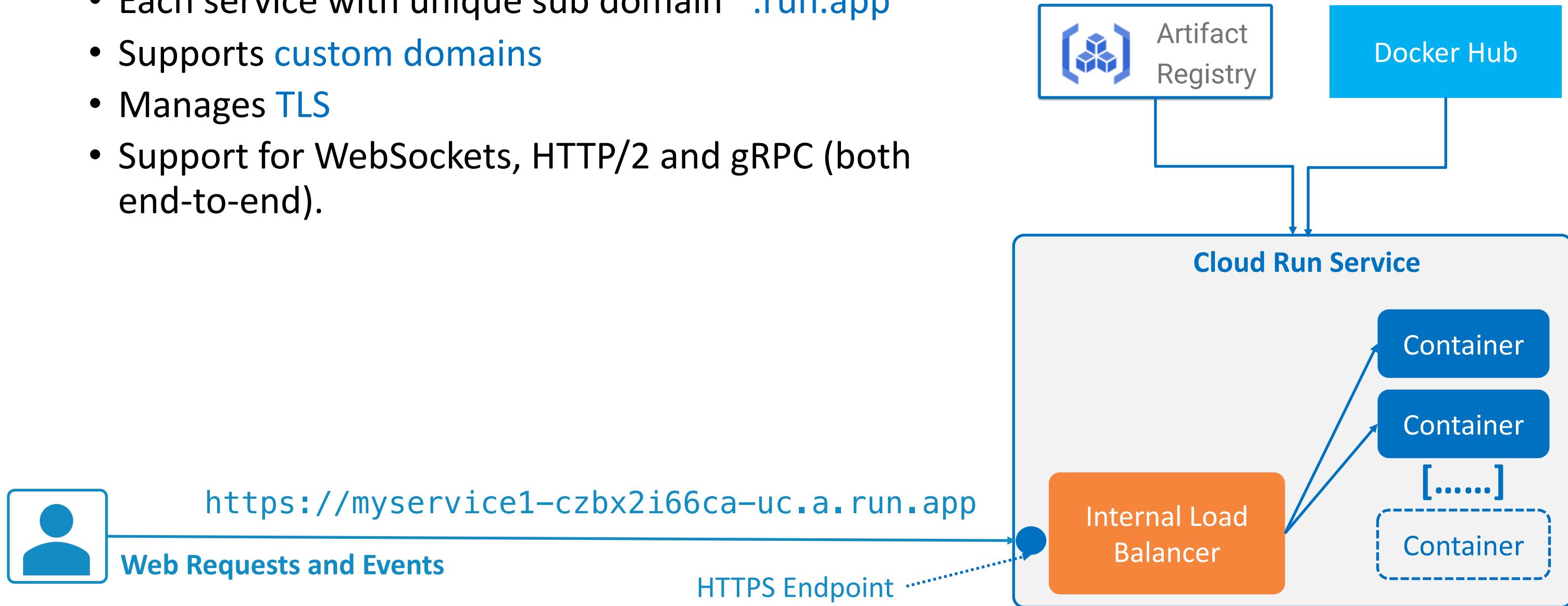
Google Cloud Run Services - Key Features

- Unique HTTPS endpoint for every service
- Private and public services
- Pay-per-use pricing for services
- Built-in traffic management
- Fast request-based auto scaling



Google Cloud Run Services - HTTPS Endpoint

- Unique **HTTPS endpoint** for every service
 - Each service with unique sub domain *.run.app
 - Supports **custom domains**
 - Manages **TLS**
 - Support for WebSockets, HTTP/2 and gRPC (both end-to-end).



Google Cloud Run Services - Pricing Model

- Pay-per-use pricing for services
- Request-based
 - If an instance is not processing requests, the **CPU is not allocated and not charged**. Additionally, we pay a per-request fee.
- Instance-based
 - You're charged for the **entire lifetime** of an instance and the CPU is always allocated. There's no per-request fee.

CPU allocation and pricing

- CPU is only allocated during request processing

You are charged per request and only when the container instance processes a request.

- CPU is always allocated

You are charged for the entire lifecycle of the container instance.

Google Cloud Run Services - Traffic Management

- **Built-in traffic management**

- Route Traffic to
 - Latest revision
 - Roll back to previous revision
 - Split Traffic to multiple revisions at same time (gradual rollout)

Revisions [MANAGE TRAFFIC](#)

[Filter](#) Filter revisions [?](#)

Name	Traffic	Deployed	Revision URLs (tags)	Actions
myservice1-00002-8xf	10%	1 minute ago	myappv2	
myservice1-00001-mjv	90%	2 minutes ago	myappv1	

Google Cloud Run Services - Autoscaling

- **Fast request-based auto scaling**
 - **Minimum Instances:** starts from zero, Set to 1 to reduce cold starts
 - **Maximum Instances:** scale out to 1000 instances and more with a request to increase quota
- **Scale to zero and minimum instances**
 - When minimum instances set to zero and no requests then **active instances will be zero**
 - New instance **created** as soon as the request comes in
 - **Negatively impacts** the response times for the first request

Autoscaling 

Minimum and maximum numbers of instances the created revision scales to.

Minimum number of instances * Maximum number of instances *

Set to 1 to reduce cold starts. [Learn more](#) 

Google Cloud Run Services - Access Modes

- **Ingress Control**

- **Public Service**

- Allow direct access from internet

- **Private Service**

- Allow traffic from VPC
 - Allow traffic from external Application Load Balancers

- **Authentication**

- Un-authenticated Access (Public API or Website)
 - Authenticated Access using Cloud Identity-Aware Proxy (Secure access via web or mobile clients)

Ingress control

Internal

Allow traffic from your project, shared VPC, and VPC service controls perimeter. Traffic from another Cloud Run service must be routed through a VPC. Limitations apply. [Learn more](#)

All

Allow direct access to your service from the internet

Authentication *

Allow unauthenticated invocations

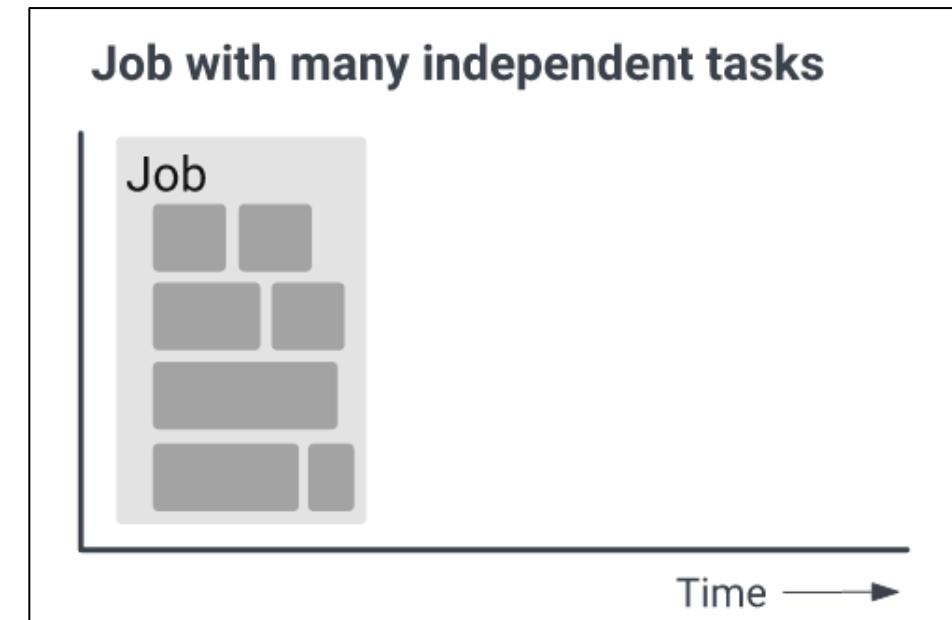
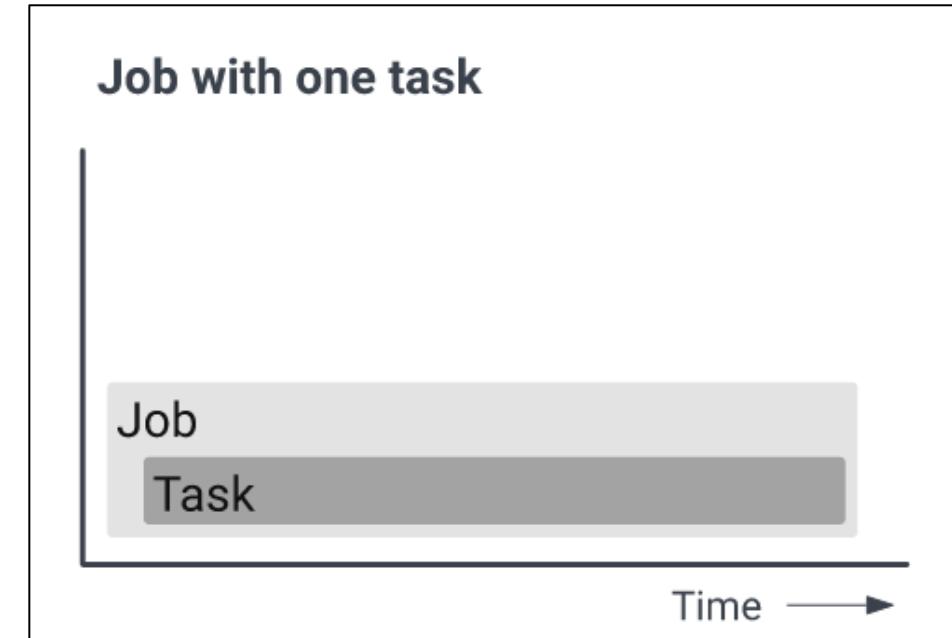
Check this if you are creating a public API or website.

Require authentication

Manage authorized users with Cloud IAM.

Google Cloud Run - Jobs

- **Cloud Run Job:** Execute code to **complete** a task and **terminate** upon completion.
- **Single Job:** Job with **one task**
- **Array Jobs:** Job with **many independent tasks**
 - **Example:** Read 1,000 images from Cloud Storage to resize and crop them
 - Parallel processing will help us to complete the job **faster**.
- **When to use Cloud Run Jobs ?**
 - **Script or tool:** Run database migrations
 - **Array Job:** When parallelized processing needed
 - **Scheduled Job:** run script every day at specified time (10pm)



Demo

Google Serverless

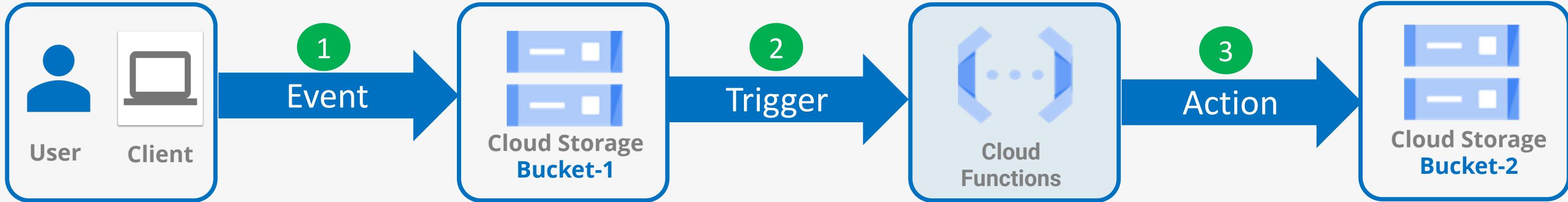
Cloud Functions

Google Cloud Functions

- **How to execute code when an event occurs ?**
 - Example Event-1: When an [HTTPS URL](#) is invoked.
 - Example Event-2: When a [message](#) is published from a Cloud Pub/Sub service.
 - Example Event-3: When a file is [uploaded](#) to Cloud Storage bucket
- **Cloud Functions: FAAS - Function As A Service**
 - Run code when an [event](#) occurs
 - Fully managed [Serverless](#) compute service
 - No need to provision or manage any servers
 - **Autoscaling:** Functions scale automatically in response to the number of incoming events (Scale Horizontally)
 - Multiple [programming languages](#) are supported
 - As on today, Python, Node.js, Go, Java, .NET and Ruby are supported

Cloud Functions - Flow

Google Cloud



Event:
Upload an image
file to Cloud
Storage

Trigger:
Triggers cloud
function in
response to the
event occurred

Action:
Perform action like
“blur background
of an image” and
upload the
modified image to
Cloud Storage

Google Cloud Function – Types

- **HTTPS Functions**

- Invoked using [HTTPS](#) urls
- Common [HTTP methods](#) like GET, PUT, POST, DELETE, and OPTIONS are supported
- We have a dedicated demo to understand in detail

- **Event-Driven Functions**

- Used to [handle events](#) from other Cloud services like Cloud Storage, Cloud Pub/Sub, Cloud Firestore and many more
- 125+ Event sources were integrated with Cloud Functions using [Eventarc](#).
- We have 2 demos planned to understand in detail

Cloud Functions - 1st gen vs 2nd gen

Cloud Function Feature	Cloud Function 1st gen	Cloud Function 2nd gen
Why do we need two versions of Cloud Functions?	First version with limited features	Very advanced which is built on top of Cloud Run and Eventarc
Timeout: If the Cloud Function has not completed by the timeout duration, then the Function will be terminated	Default: 1 minute HTTP Functions: 9 minutes Event-driven Functions: 9 minutes	Default: 1 minute HTTP Functions: 60 minutes Event-driven Functions: 10 minutes Longer processing times
Compute Power (Memory and CPU)	Max Memory: 8GB Max CPU: 4vCPU CPU is auto-allocated. No option to choose	Max Memory: 32GB Max CPU: 8vCPU Have option to select CPU
Concurrency: The maximum number of concurrent requests that can reach each container instance.	No Concurrency option Example: Requests: 10 Container Instances created: 10	Concurrency supported Example: Concurrency: 10 , Requests: 10 Container Instances created: 1 Limitations: not supported for all Runtimes
Traffic Splitting 1. Gradual Rollout (V1: 90%, V2: 10%)	NOT POSSIBLE	As it is built-on Cloud Run, Traffic Splitting is easily applied using gcloud run or Cloud Run web console .

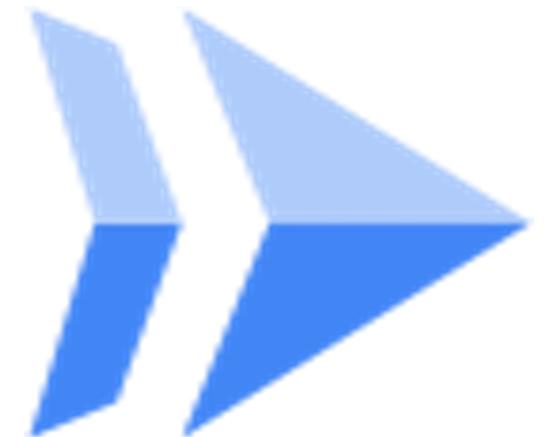
Cloud Functions - 1st gen vs 2nd gen

Cloud Function Feature	Cloud Function 1st gen	Cloud Function 2nd gen
Number of Event Triggers Supported	Limited <ol style="list-style-type: none">1. HTTP2. Cloud Storage3. Cloud Pub/Sub4. Cloud Firestore5. Few Firebase triggers	125+ Triggers supported <ol style="list-style-type: none">1. HTTPS2. Eventarc (Many Google Cloud sources)3. Third-Party Events

Concept



Google Serverless



Cloud Run vs Cloud Run for Anthos

Cloud Run vs Cloud Run for Anthos

Cloud Run Feature	Cloud Run	Cloud Run for Anthos
Deployment Environment	Serverless Platform	Google GKE Enterprise on Anthos Platform
Underlying Infrastructure	Compute Instances	GKE Worker Nodes on 1. On-Premise 2. GCP 3. Multi-Cloud (AWS, Azure Clouds)
Infrastructure Control	No visibility to underlying infra	1. More control over underlying infra 2. Suitable for large organizations that requires greater customization
Autoscaling	Automatic scaling based on demand.	1. Inherits GKE Kubernetes Autoscaling features
Networking	Automatically managed, No control for us	1. Inherits GKE Kubernetes networking capabilities 2. More fine-grained control over networking configurations
Security and Compliance	Provides built-in security and adheres to GCP security standards	1. Additional security controls to support specific compliance requirements. 2. Especially for regulated industries

Cloud Run vs Cloud Run for Anthos

Cloud Run Feature	Cloud Run	Cloud Run for Anthos
Pricing Model	Billed based on number of requests and vCPU usage	<ol style="list-style-type: none"> 1. Part of GKE Enterprise Kubernetes Resource usage 2. Anthos Platform pricing
Use cases	Ideal for small teams (Developers) focusing on building applications without managing infrastructure	<ol style="list-style-type: none"> 1. Suited for organizations with a hybrid or multi-cloud strategy 2. Suited for organizations requiring more control over the Kubernetes underlying infrastructure

Cloud Run for Anthos - Screenshots

Anthos Platform

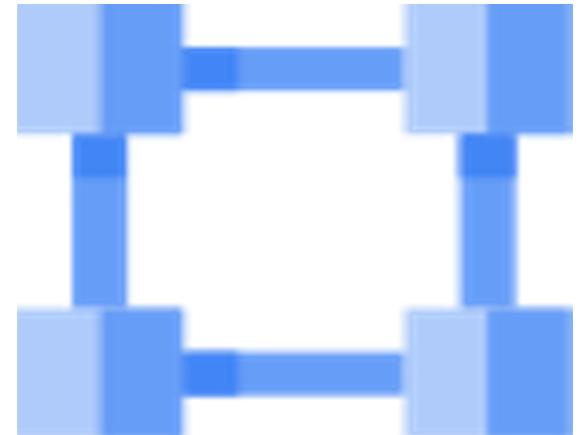
The screenshot shows the Google Cloud Anthos Platform interface. On the left, there's a sidebar with various options like Anthos, Fleet Management, Clusters, Feature Management, Service Mesh, Config, Policy, Security (with a PREVIEW badge), and Cloud Run for Anthos. The Cloud Run for Anthos option is highlighted with a red box. The main content area is titled "Services" and includes "CREATE SERVICE" and "MANAGE CUSTOM DOMAINS" buttons. It displays a message about services being located in specific infrastructure to handle requests. A "Filter" section allows filtering by Name, Req/sec, Region, and GKE Cluster. Below it, a message says "No rows to display".

GKE Enterprise

The screenshot shows the Google Cloud GKE Enterprise interface. On the left, there's a sidebar with "Kubernetes Engine" (highlighted with a red box), "Fleet" (kdaida123 fleet), "Resource Management" (Overview, Clusters, Workloads, Teams), and "Feature management". The main content area is titled "Overview (kdaida123 fleet)". It includes a time range selector (1 hour, 6 hours, 1 day, 1 week, 1 month) and a section for "Clusters in this Fleet" which says "No data" and has a "View all clusters" link.

**Demo**

Google Cloud Networking

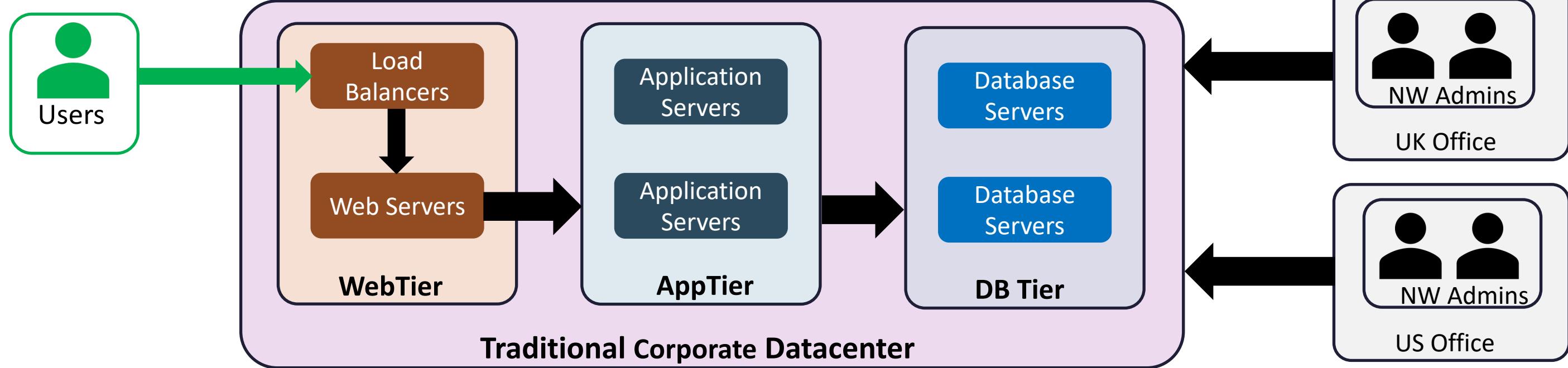


Virtual Private Cloud (VPC)

Auto-mode, Custom-mode

Demo-01

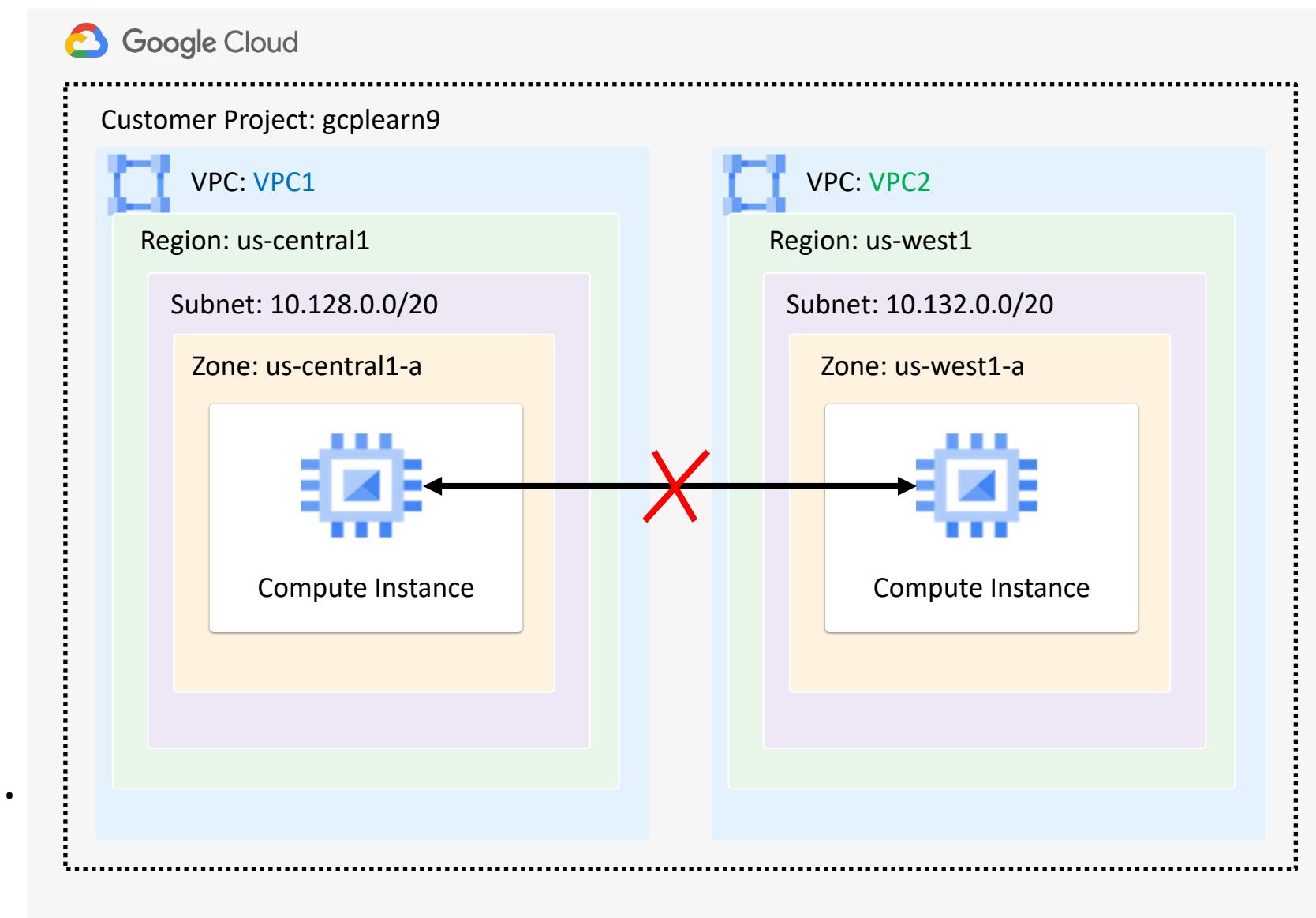
Google Cloud VPC



- **Traditional Corporate Datacenter**
 - Users **can access applications** (www.stack simplify.com) hosted in Datacenter using internet
 - Users **cannot access** any servers directly (Web, App or DB Servers), its a **private network**
 - **Network/System Admins** can access the servers in Datacenter using **Corporate Network** from different locations
- **How to create a Private network in Google Cloud ?**
 - Google Cloud **VPC** (Virtual Private Cloud)

Google Cloud VPC

- **Cloud VPC:** Private Network in Google Cloud
- VPC is a **global resource**, not associated to a region or zone
- VPC resources like **routes and firewalls** also global resources
- **Isolation and Security:**
 - VPC allows you to create a **logically isolated networks**
 - You can create resources (compute instances) **within a network** you created.
 - This isolation provides **security** by **blocking unauthorized access to your VPC from other VPCs** in Google Cloud.



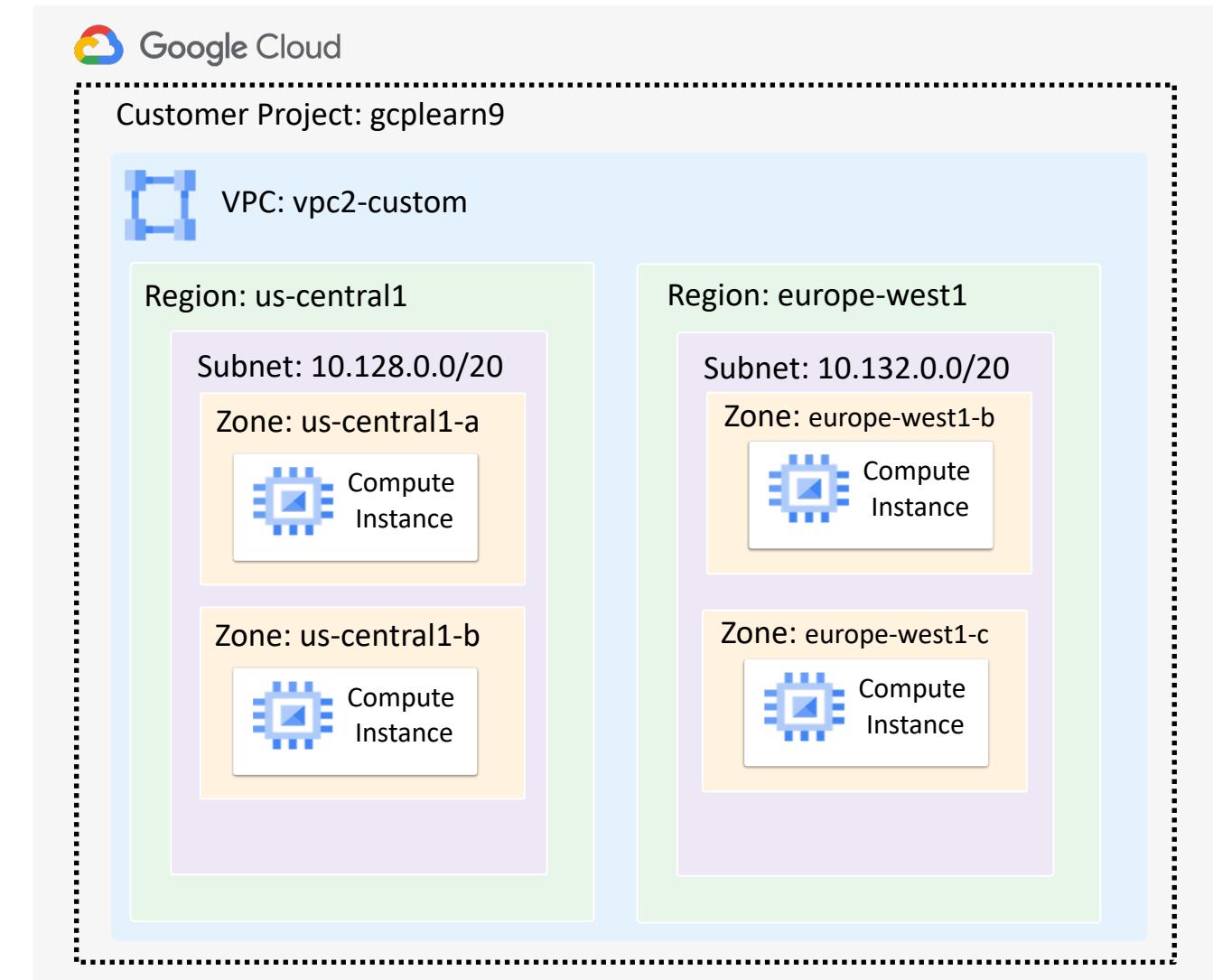
Google Cloud VPC - Subnets

- **Subnets:** Subnets are **regional resources**
- Subnets can **span to multiple zones** in a single region (complete contrary to AWS)
- **IP Address ranges** are associated with subnets (Ex: 10.128.0.0/20)
- VPC network must have **at least one subnet** before you can use it

Subnets [+ ADD SUBNET](#) [FLOW LOGS ▾](#)

Filter Enter property name or value

<input type="checkbox"/> Name ↑	Region	Stack Type	Internal IP ranges
<input type="checkbox"/> default	us-central1	IPv4	10.128.0.0/20
<input type="checkbox"/> default	europe-west1	IPv4	10.132.0.0/20
<input type="checkbox"/> default	us-west1	IPv4	10.138.0.0/20



Google Cloud VPC - Modes

- We can create VPC in **two modes**

- Auto Mode VPC Network
- Custom Mode Network

• Auto mode VPC Network

- Subnets will be created **automatically** (one subnet for each region)
- By default, every new project will have a **default VPC** created and ready-to-use
- This default VPC uses **auto-mode**, so default subnets were **auto-generated** for each region with **predefined IP Address ranges**
- Automatic mode supports **IPv4 (single stack)** only
- As new regions become available, **new subnets in those regions are automatically added to the auto mode VPC network.**

VPC networks			
Filter Enter property name or value			
Name	Subnets	MTU	Mode
default	42	1460	Auto
vpc2-custom	1	1460	Custom

VPC network details			
default			
OVERVIEW SUBNETS STATIC INTERNAL IP ADDRESSES			
Subnets	+ ADD SUBNET	FLOW LOGS	
Filter Enter property name or value			
<input type="checkbox"/>	Name	Stack Type	Internal IP ranges
<input type="checkbox"/>	default	IPv4	10.128.0.0/20
<input type="checkbox"/>	default	IPv4	10.132.0.0/20
<input type="checkbox"/>	default	IPv4	10.138.0.0/20
<input type="checkbox"/>	default	IPv4	10.140.0.0/20
<input type="checkbox"/>	default	IPv4	10.142.0.0/20
<input type="checkbox"/>	default	IPv4	10.146.0.0/20
<input type="checkbox"/>	default	IPv4	10.148.0.0/20

Google Cloud VPC - Modes

- **Custom Mode VPC Network**

- Subnets need to be created **manually**
- We define IP Address ranges (**as desired**)
- Custom mode supports **IPv4 or IPv4 and IPv6 (dual- stack)**
- **Highly recommended** for all type of workloads (Dev to Production)

VPC networks			
Filter		Enter property name or value	
Name ↑	Subnets	MTU	Mode
default	42	1460	Auto
vpc2-custom	1	1460	Custom

Google Cloud VPC - Auto vs Custom modes

Features	VPC Auto Mode	VPC Custom Mode
Subnet creation	One subnet from each region is auto-generated	Need to create subnets manually (more flexible)
IP Ranges	Pre-defined: Chances of overlap with other VPC networks and IP ranges which are external to cloud (On-premise networks)	Complete Control: We decide on the IP Range based on analysis of our current and future IP range data considering all our on-premise Datacenters and VPC networks
Mode Conversion	We can do a one-time conversion from auto-mode to custom-mode	NOT POSSIBLE (custom-mode to auto-mode)
VPC Peering Or Cloud VPN	Every auto-mode vpc network uses same IP Ranges (pre-populated) for all auto-mode VPCs Cannot connect two VPCs with auto-mode to one another using VPC Peering or Cloud VPN	Fully supported both VPC Peering and Cloud VPN to connect VPCs created using VPC custom mode

Google Cloud VPC - Auto vs Custom modes

Features	VPC Auto Mode	VPC Custom Mode
Dual Stack Subnets	Do not support dual-stack subnets (IPv4 and IPv6)	Supports dual-stack subnets (IPv4 and IPv6)
Production	Not recommended for production	Highly recommended for production use

A large, semi-transparent green oval containing the word "Demo" in white, centered above the main title.

Demo

Google Cloud Networking



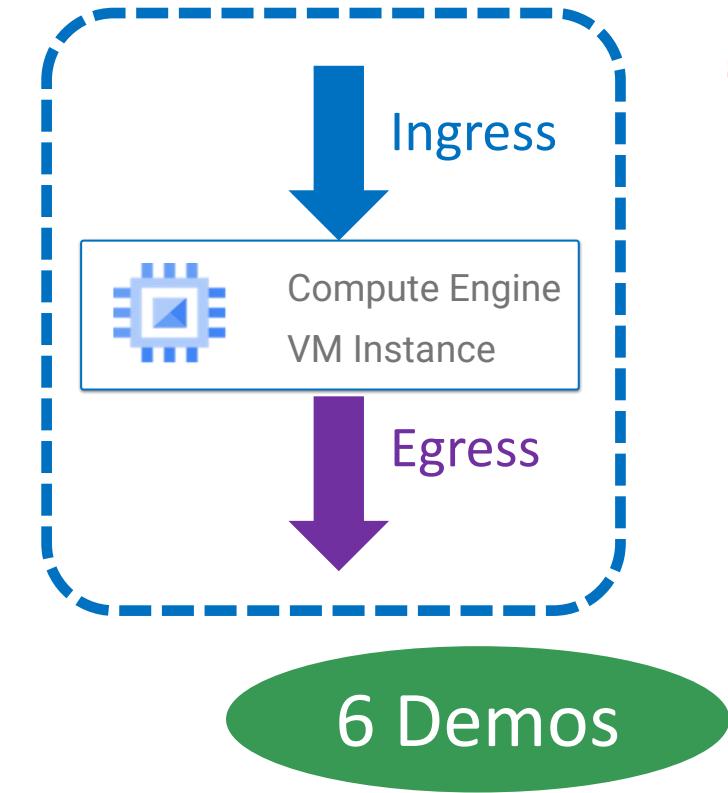
Virtual Private Cloud (VPC) VPC Firewall Rules

A small red oval containing the word "Intro" in white, located in the bottom right corner.

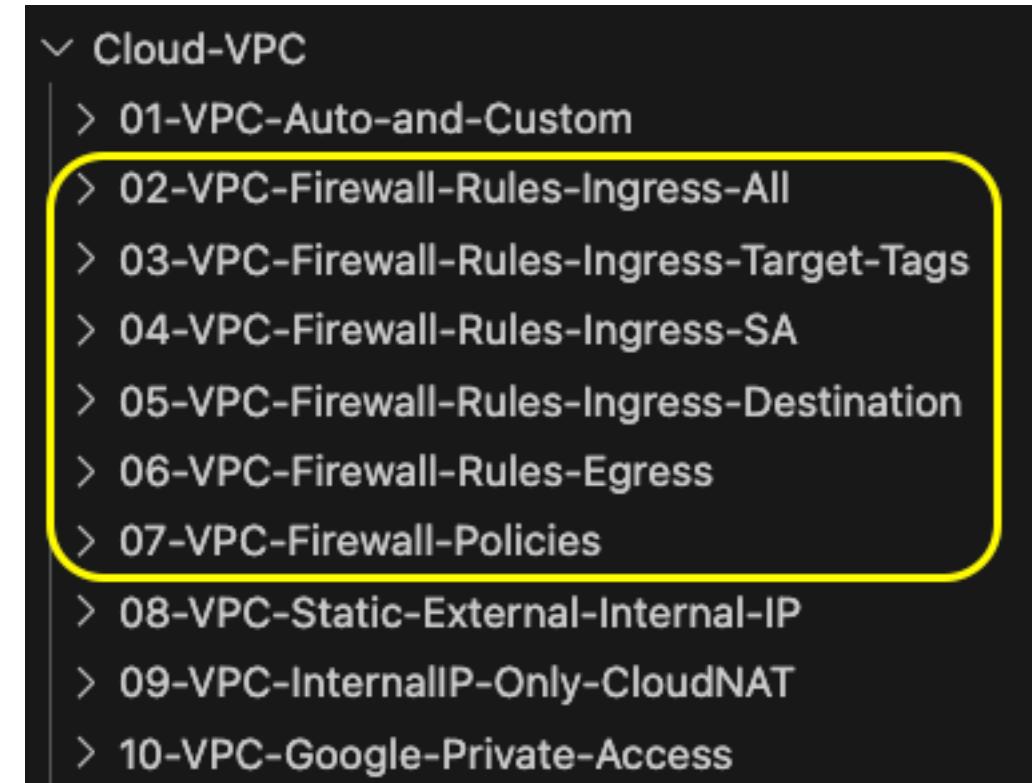
Intro

Google Cloud VPC Firewall Rules

- **VPC Firewall Rules:** let you allow or deny connections **to or from virtual machine (VM)** instances in your VPC network
- **Firewall Rules:** can be applied to **VMs in a single VPC network**
- **Firewall Policies:** can be applied to **VMs in multiple VPC networks**
- **Firewall rule types**
 - Ingress Rules
 - Egress Rules
- Firewall rules can be applied
 - between **other networks to VM instances**
 - between **individual VM instances** in the same VPC network



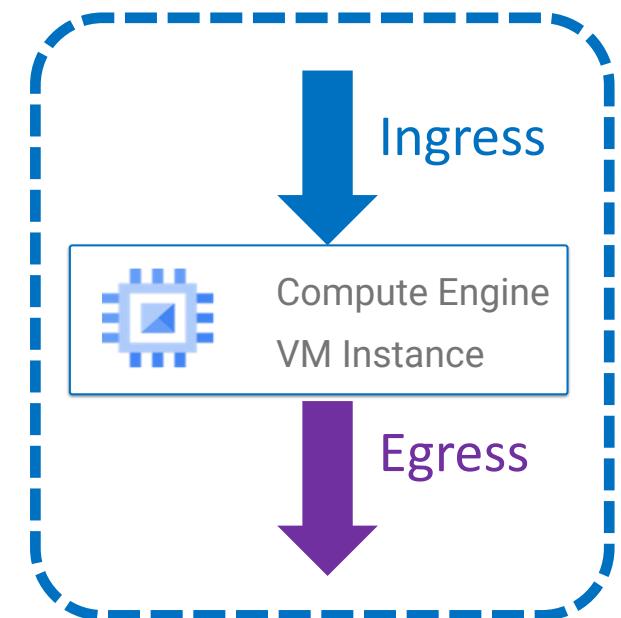
6 Demos



Google Cloud VPC Firewall Rule - Components

- **Firewall Rules Components**

- Direction of Traffic
- Priority
- Target
- Source
- Destination
- Protocol and Ports
- Enforcement Status



Google Cloud VPC Firewall Rule - Components

- **Direction of Traffic**

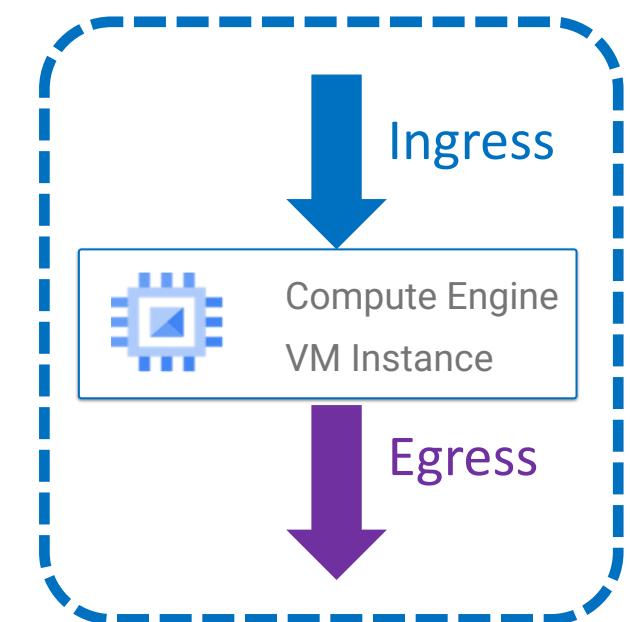
- **Ingress:** Ingress (inbound) describes **packets entering a network interface** of a target (VM Instance)
- **Egress:** Egress (outbound) describes **packets leaving a network interface** of a target (VM Instance)

- **Priority**

- Priority is an integer from 0 to 65535
- Lower integers indicate **higher priorities**
- If priority not specified when creating a rule, it is assigned a priority of **1000**

- **Example:**

- **FW rule-1:** Deny SSH port 22 with priority 200
- **FW rule-2:** Allow SSH port 22 with priority 100
- FW rule-2 has **high priority due to least integer value** which takes **precedence and allows SSH traffic**



Google Cloud VPC Firewall Rule - Components

- **Action on match**

- **Allow:** Allow traffic
- **Deny:** Block traffic

- **Target**

- Targets identify the **network interfaces of instances** to which the firewall rule applies

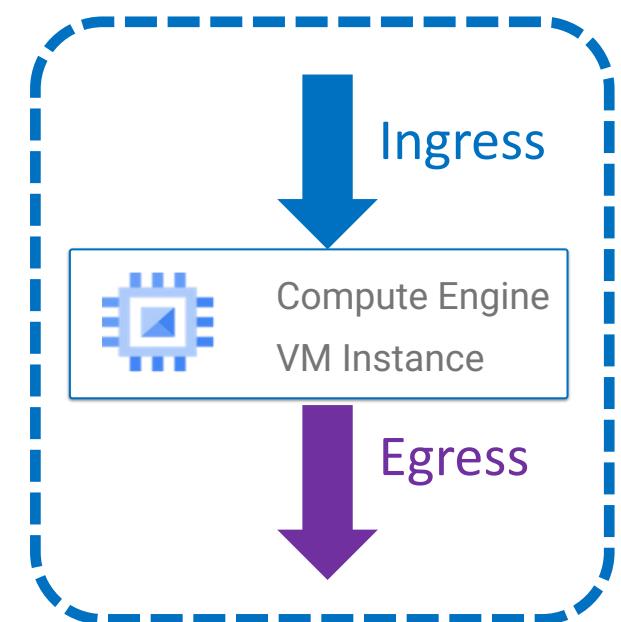
- Compute Engine instances, GKE nodes , App Engine flexible environment instances

- **Target Types**

- **Default target - all instances in the VPC network:** Firewall rule applies to all instances in the VPC network.

- **Instances by target network tags:** firewall rule applies only to instances in the VPC network with a matching network tag.

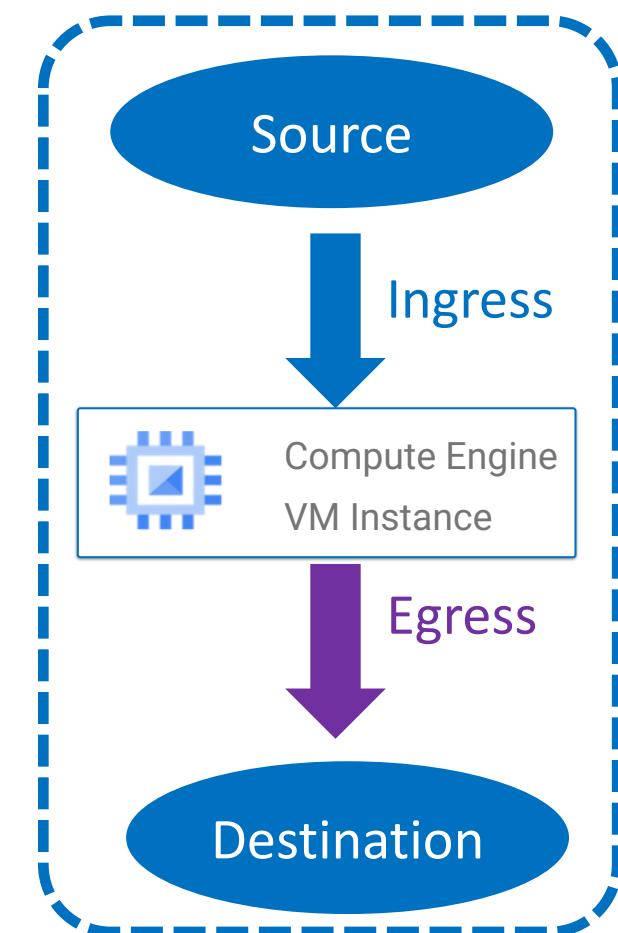
- **Instances by target service accounts:** firewall rule applies only to instances in the VPC network that use a specific service account



3 Practical
Demos for
Target Types

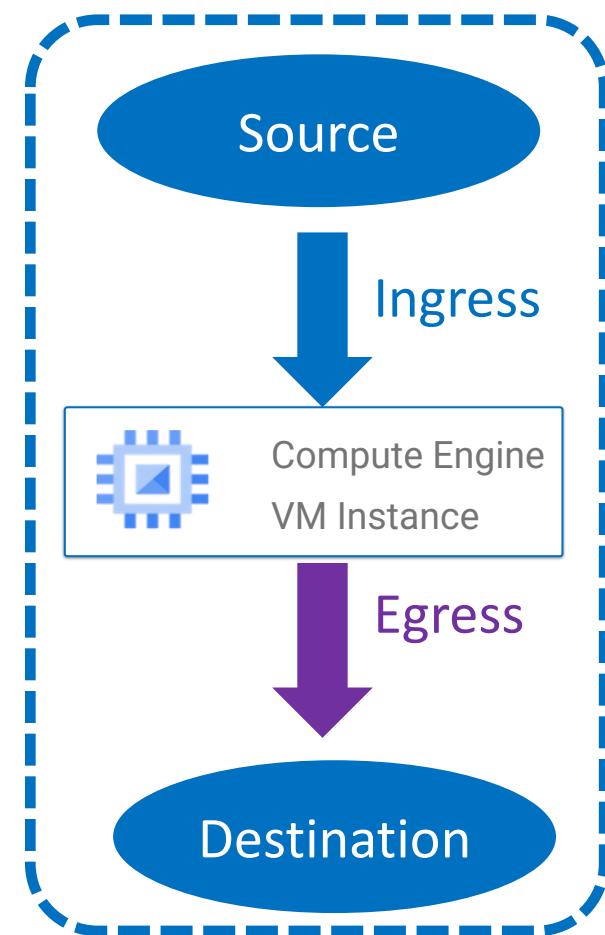
Google Cloud VPC Firewall Rule - Components

- **Source:** Primarily used for [Ingress rules](#)
 - Default source range if unspecified is [0.0.0.0/0](#)
 - We can use the following sources for ingress firewall rules
 - Source IPv4, IPv6 ranges
 - Source network tags
 - Source service accounts
 - Source Combination ([Second source filter](#))
 - source IPv4 + source network tags
 - source IPv6 + source network tags
 - source IPv4 + source service accounts
 - source IPv6 + source service accounts
- **Source for egress rules**
 - We can define [source for egress rules](#) but [99% of the cases we don't need it](#)
 - **Default—implied by target:** By default (if source unspecified), [packet sources are defined implicitly](#) as described in [Targets and IP addresses](#) for egress rules.
 - For egress rules we can define [source IPv4, IPv6 ranges](#)



Google Cloud VPC Firewall Rule - Components

- **Destination:** Primarily used for [egress rules](#)
 - Default destination range if unspecified is 0.0.0.0/0
 - Destination [IPv4](#) and [IPv6](#) ranges
 - **Destinations for ingress rules**
 - **Default—implied by target:** By default (if source unspecified), [packet sources are defined implicitly](#) as described in [Targets and IP addresses](#) for ingress rules.
 - Destination [IPv4](#) and [IPv6](#) ranges



Google Cloud VPC Firewall Rule - Components

- **Protocol and Ports**

- Restrict access using **protocols** or **protocols and destination ports**.

- **Supported Protocols**

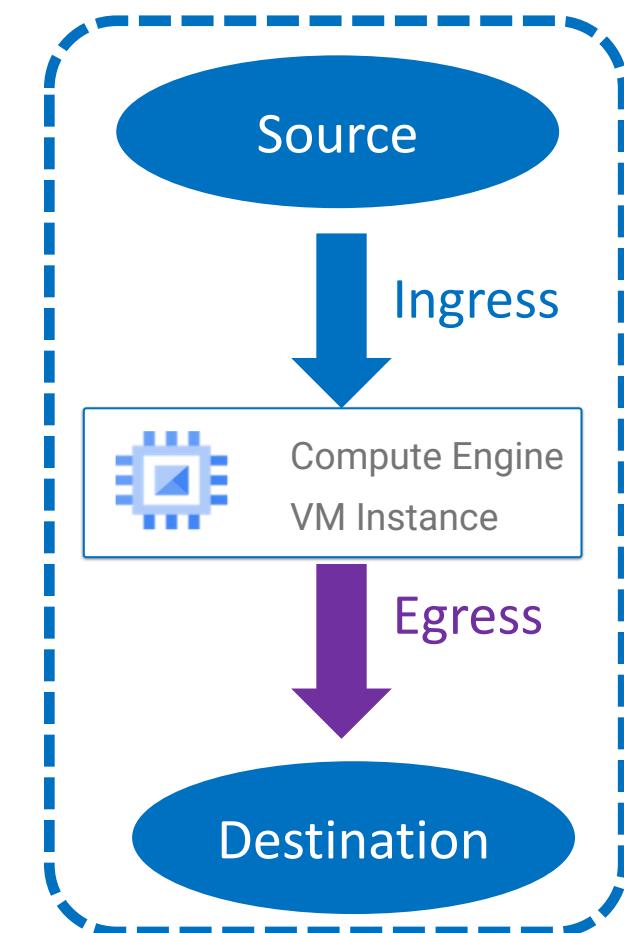
- tcp, udp, icmp, esp, ah, sctp, ipip
- For all other protocols, we can directly use protocol numbers (example: http is 80, https is 443)

- **How do we define them ?**

- **No protocol and port:** All protocols and their applicable ports are allowed
- **Protocol:** tcp
- **Protocol and single port:** tcp:80
- **Protocol and port range:** tcp:80-90
- **Combinations:** icmp,tcp:80, tcp:443, udp:67-69

- **Enforcement Status**

- We can **enable or disable** a firewall rule
- Primarily used for **troubleshooting** and when **performing maintenance** (Example: Allow SSH 22 during maintenance, enable that rule)



Google Cloud VPC Firewall - Implied Rules

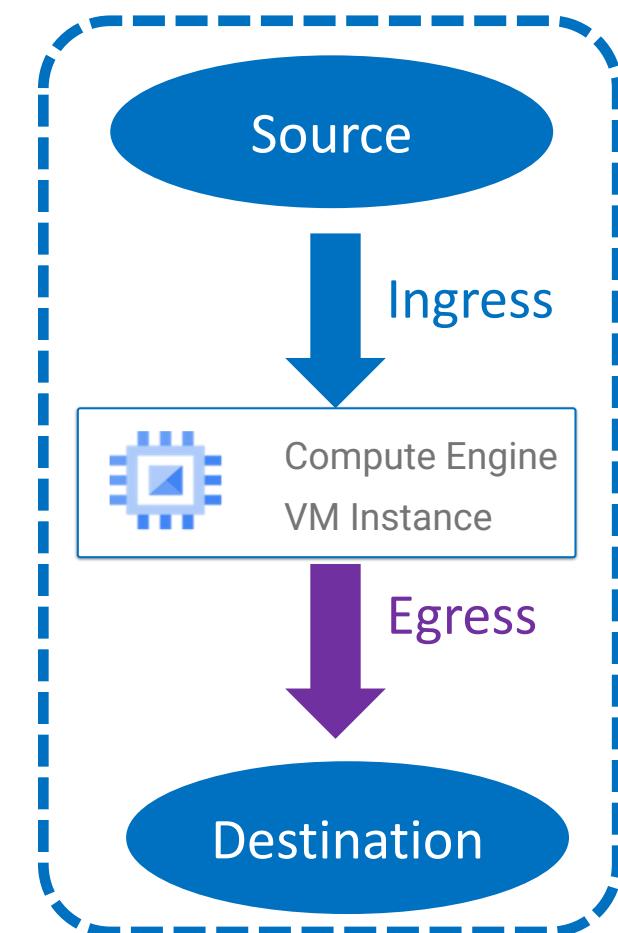
- **Implied Rules**

- **Implied Egress Rule**

- By default, **outbound is open (Allow)**
 - **Action:** Allow
 - **Destination:** 0.0.0.0/0 (IPv4), ::/0 (IPv6)
 - **Priority:** 65536 (least priority)
- We can define a **higher priority rule** to restrict outbound access
- By default, **internet access is allowed** using VM Instance public IP or Cloud NAT (if no other firewall rule explicitly blocks it)

- **Implied Ingress Rule**

- By default, **inbound is blocked (Deny)**
 - **Action:** Deny
 - **Source:** 0.0.0.0/0
 - **Priority:** 65536
- We can define a **higher priority rule** to allow inbound access



Google Cloud VPC Firewall Rules – default VPC

- Default Firewall Rules in VPC Network: default

Rule name	Direction	Priority	Source ranges	Action	Protocols and ports	Description
default-allow-internal	ingress	65534	10.128.0.0/9	allow	tcp:0-65535 udp:0-65535 icmp	Permits incoming connections to VM instances from other instances within the same VPC network.
default-allow-ssh	ingress	65534	0.0.0.0/0	allow	tcp:22	Lets you connect to instances with tools such as <code>ssh</code> , <code>scp</code> , or <code>sftp</code> .
default-allow-rdp	ingress	65534	0.0.0.0/0	allow	tcp:3389	Lets you connect to instances using the Microsoft Remote Desktop Protocol (RDP).
default-allow-icmp	ingress	65534	0.0.0.0/0	allow	icmp	Lets you use tools such as <code>ping</code> .

Reference: https://cloud.google.com/firewall/docs/firewalls#default_firewall_rules

Google Cloud VPC Firewall Rules - Best Practices

- **Implement least-privilege**

- Block **all traffic by default** and **allow the specific traffic** you need
- In short, limit the firewall rule to **protocol and port** that you need

- **Allow Rules**

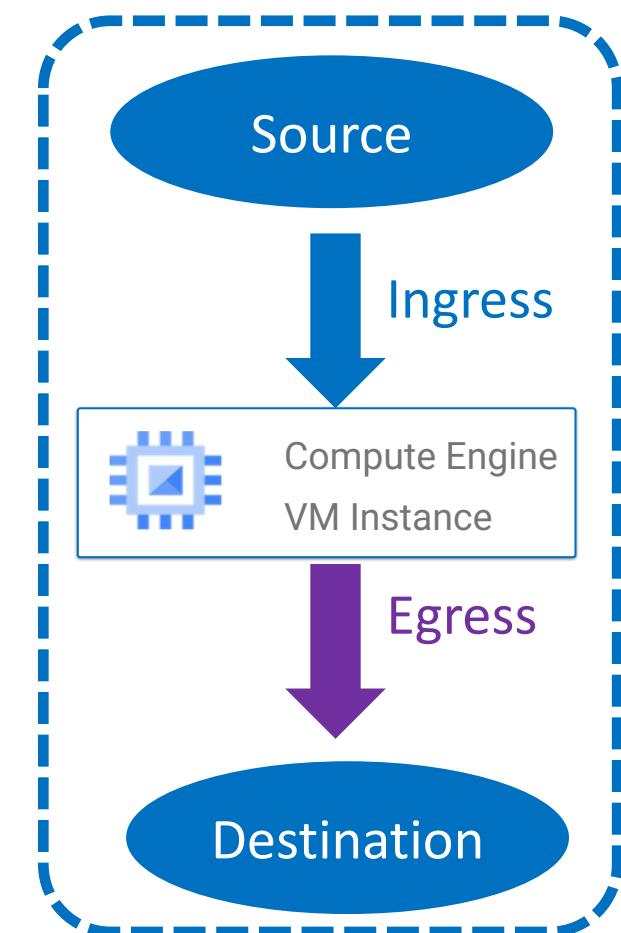
- For allow rules, restrict them to specific VMs by specifying **service account of the VMs**

- **Limit Rules based on IP address**

- Try to minimize the **per IP address** firewall rules
- Try to use **IP Ranges**, so tracking of **rules will be easy** in long run and good **for compliance and auditing** purpose

- **Enable Firewall Rules Logging**

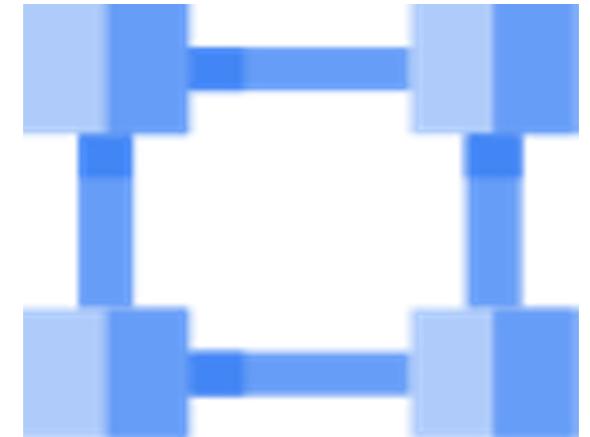
- Enable **logging** to verify that firewall rules are being used in the **intended way**
- This setting will **incur additional costs** and using it **selectively** is recommended



Demo



Google Cloud Networking

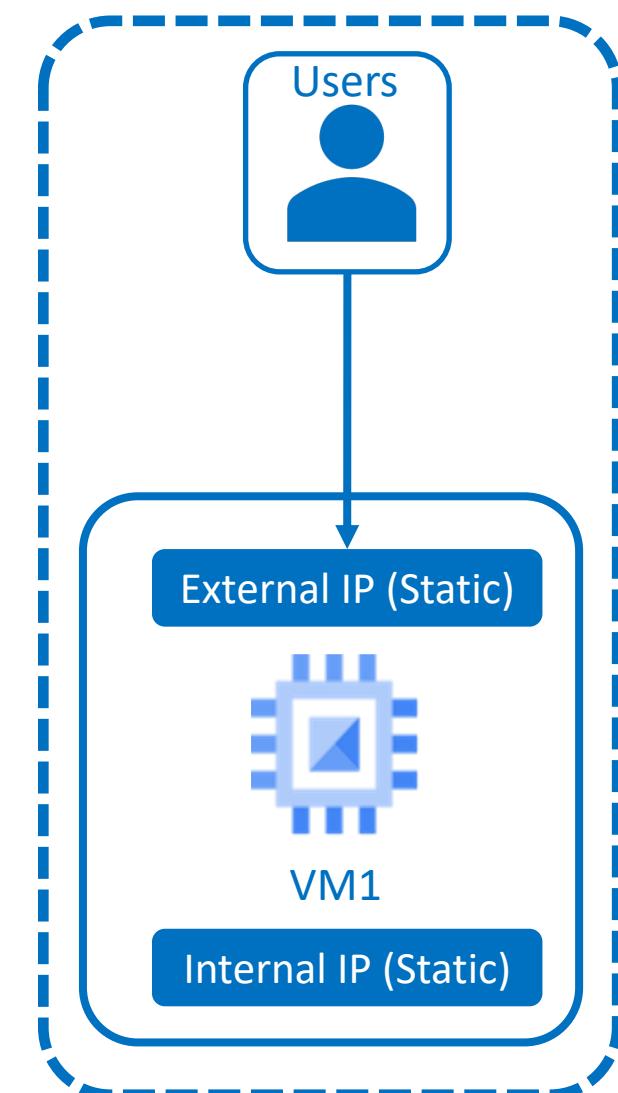


Virtual Private Cloud (VPC)

IP Addresses (External, Internal)

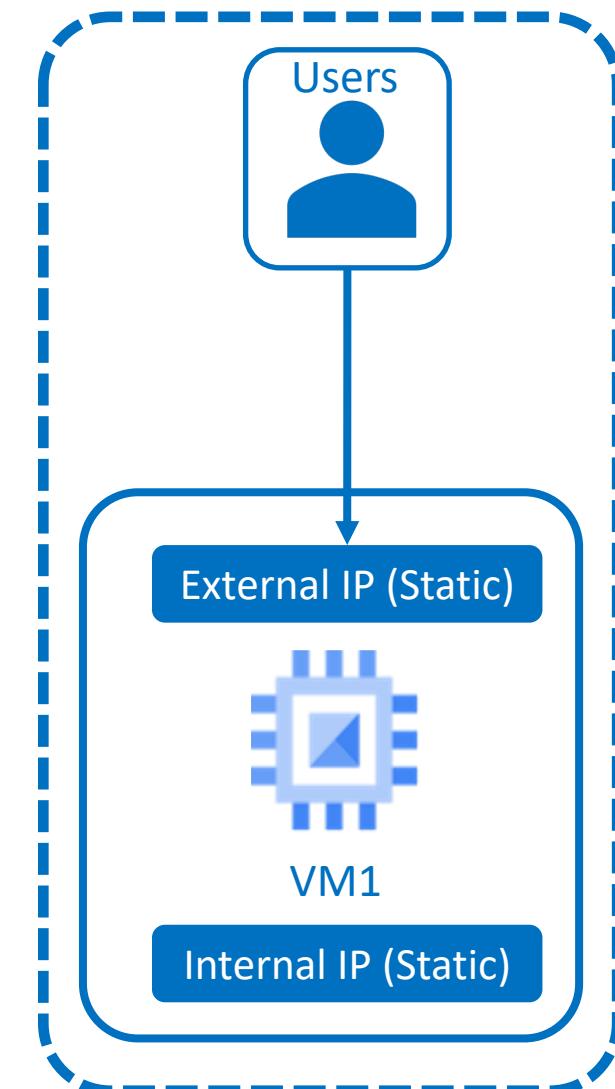
Google Cloud VPC - IP Addresses

- **IP Addresses**
 - External IP (Public IP)
 - Internal IP (Private IP)
- **Ephemeral IP:** This IP **doesn't persist** beyond the life of the resource
 - Google Cloud **automatically assigns** the resource an ephemeral IP address (Ex: VM Instance)
 - Ephemeral IP address is **released if you stop or delete** the resource
- **Static IP:** This IP need to be **reserved and released explicitly**
 - Primarily used for load balancers
 - Ex: DNS register the static IP to hostname (stacksimplify.com)



Google Cloud VPC - IP Addresses

- **External IP (Public IP):** Anyone can access this IP via [internet](#)
 - External [IPv4](#) addresses can be provided by google or you can bring your own IP ([BYOIP](#))
 - External [IPv6](#) addresses are provided by google
 - Primarily used by [VM Instances and Load Balancers](#)
 - Can be created as [Regional or Global](#) resource
 - Regional external IP can use [premium or standard](#) network tiers
 - Used by VM Instances, External passthrough network load balancers
 - Used by Cloud NAT and Cloud VPN
 - Global IP can only use [premium network tier](#) (Googles high quality network backbone)
 - Used by Global External Application load balancers
 - Used by Global External Proxy network load balancers



Google Cloud VPC - Static IP Addresses

- **Internal IP (Private IP): Not accessible via internet**

- Internal IPs are **local** to VPC network
- Internal IPs are accessible
 - within the VPC network
 - from other VPC network using **VPC peering**
 - from on-premise network using **Cloud VPN** or **Cloud Interconnect**

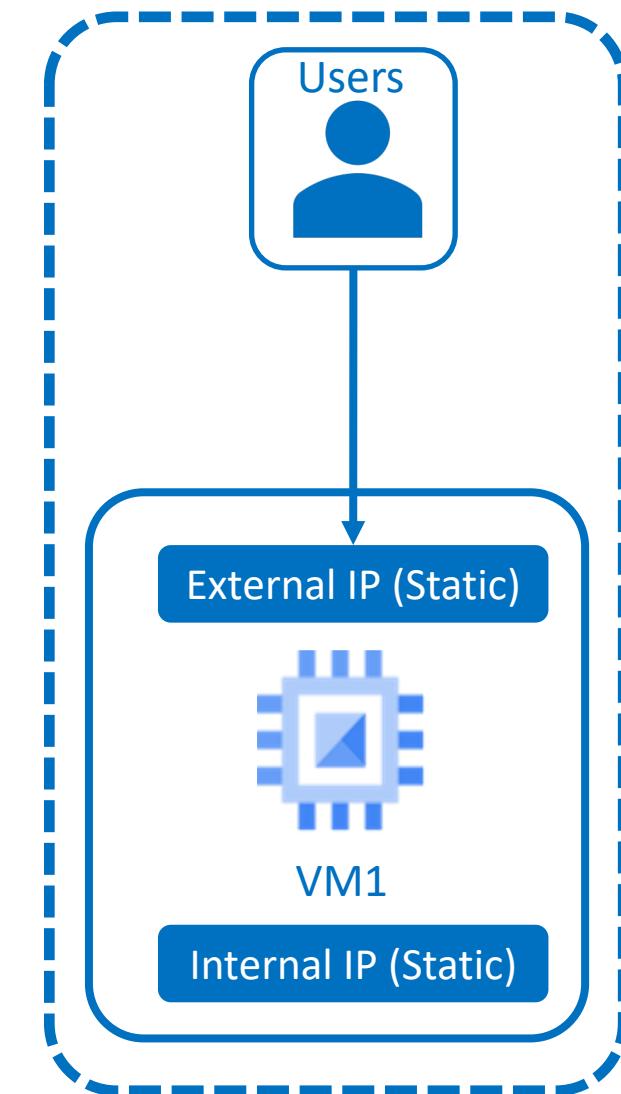
- Internal IPs always use **premium network tier**

- **Regional Internal IP addresses**

- **Static IP:** Can be **reserved and released explicitly** (Regional resource)
- Used by **subnets** (Compute Engine network interfaces, GKE Nodes)
- Used by **internal** application load balancers, passthrough network load balancers, many more

- **Global Internal IP addresses**

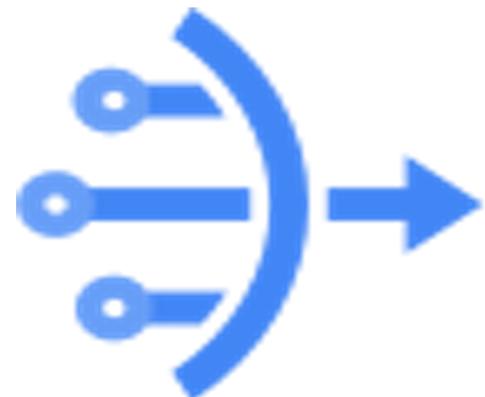
- Private Service Connect endpoints for Google APIs
- **Static IP:** **Cannot reserve global Internal IP**



Demo



Google Cloud Networking Cloud NAT

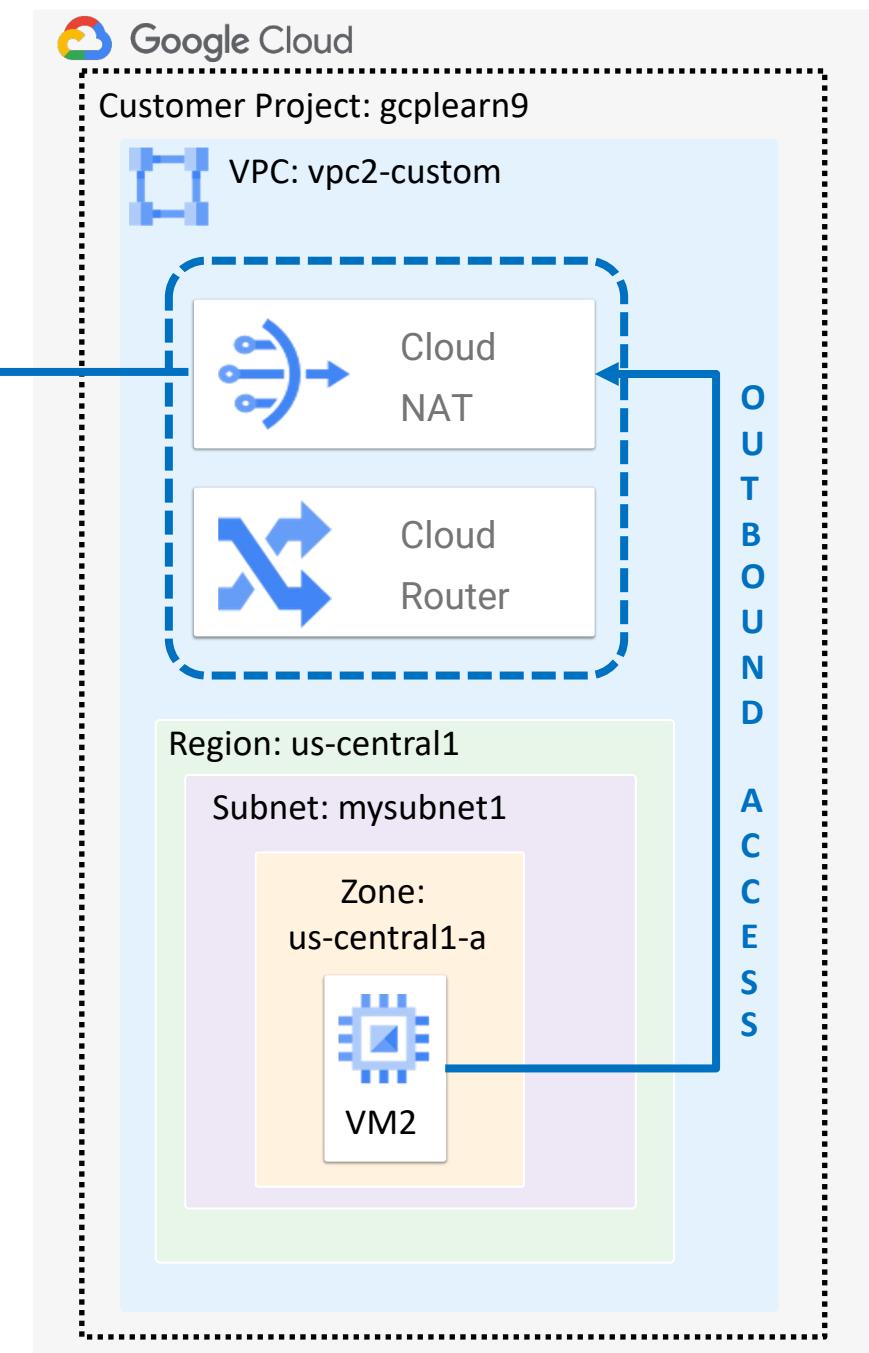
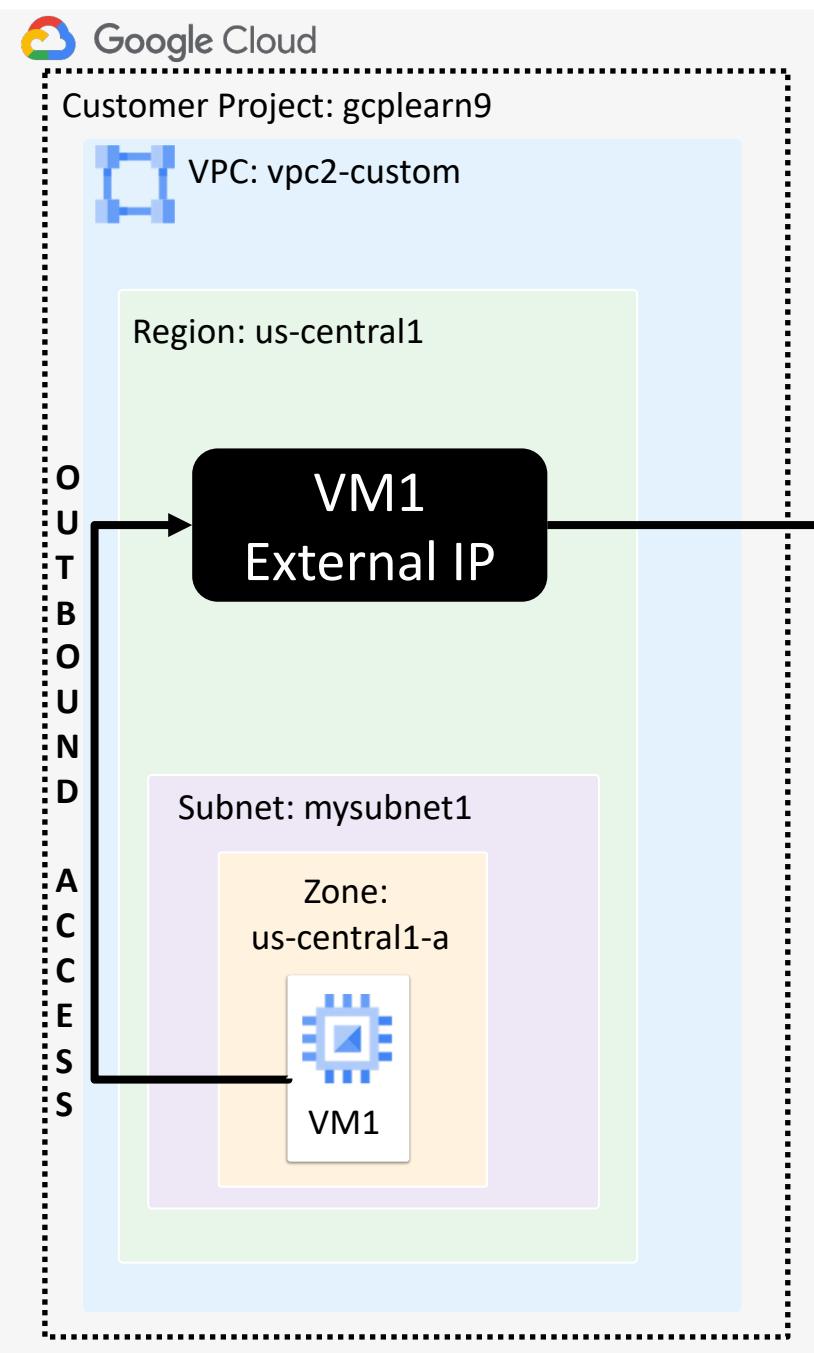


Google Cloud NAT

VM with External IP

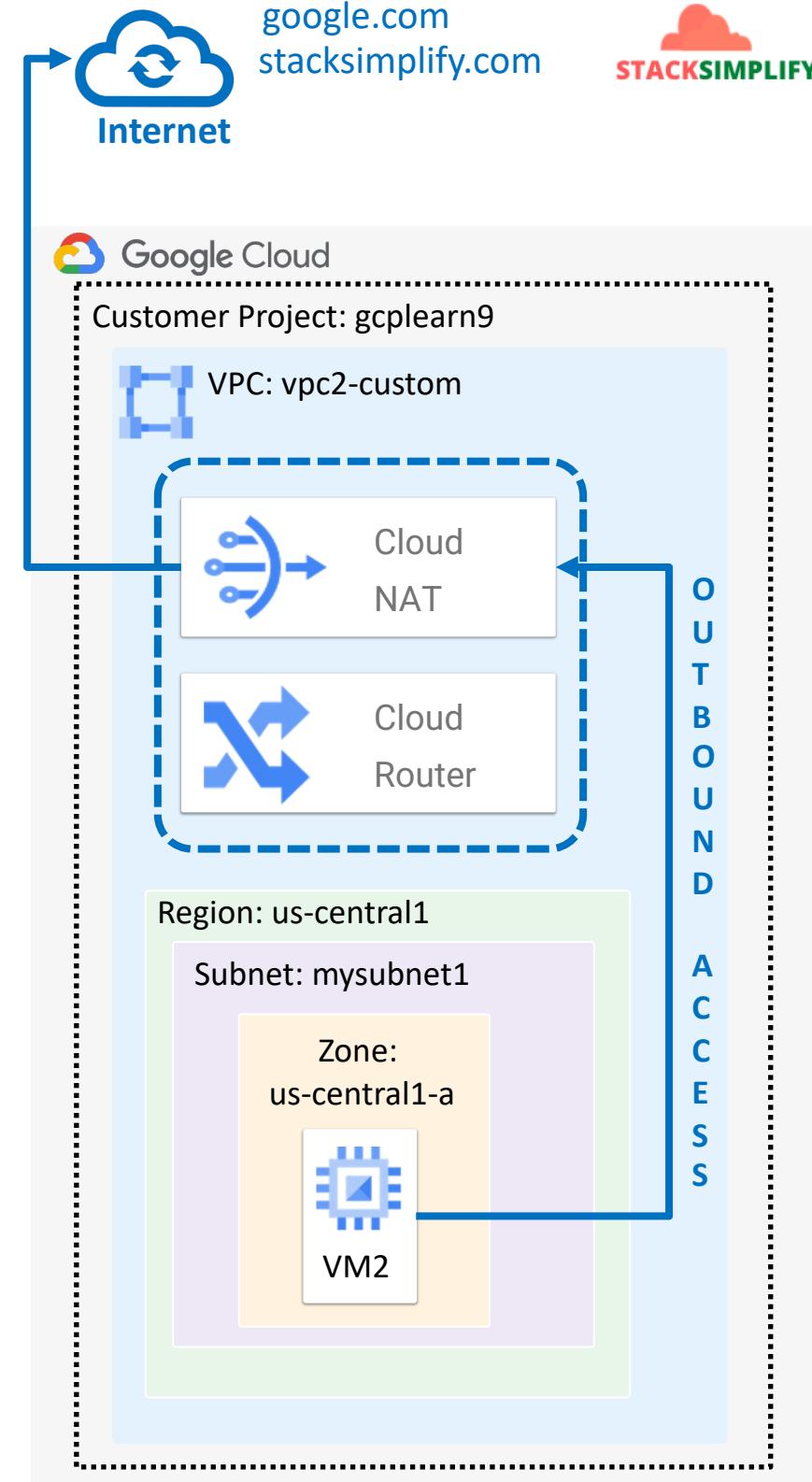
VM without External IP

Why do we need
Cloud NAT ?



Google Cloud NAT

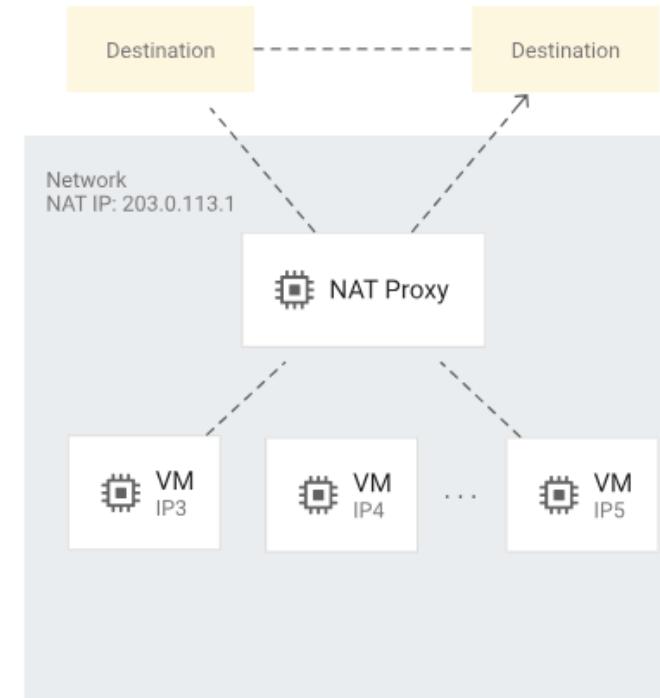
- **Cloud NAT:** primarily used to create **outbound connections** to the internet or to other VPC networks
- Cloud NAT enables **instances in a private subnet** connect to **resources outside your VPC network**
- Cloud NAT provides outgoing connectivity for
 - Compute Engine VM Instances
 - Private Google Kubernetes Engine (GKE) clusters
 - Cloud Run Instances using Serverless VPC Access
 - Cloud Function Instances using Serverless VPC Access
 - App Engine Standard using Serverless VPC Access
- Cloud NAT Types
 - Public NAT
 - Private NAT



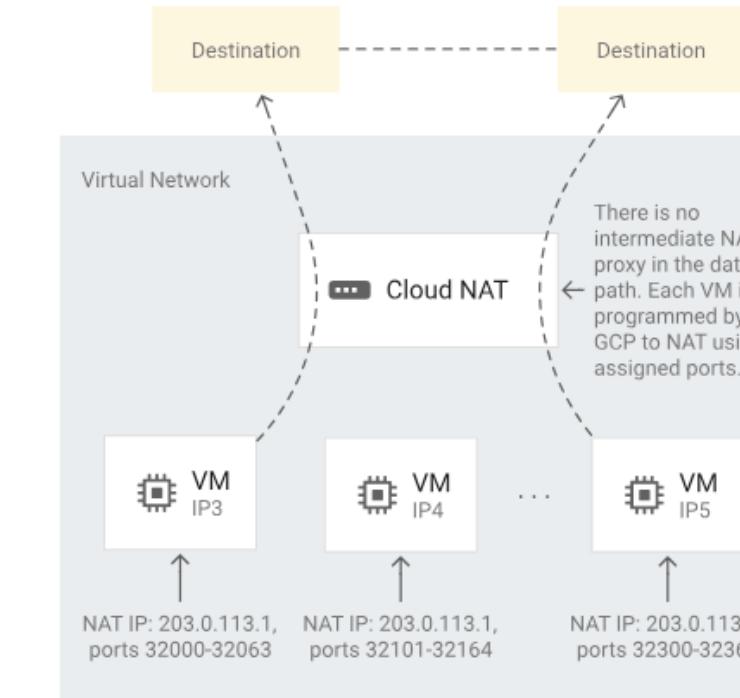
Google Cloud NAT

- **Public NAT**

- Public NAT enables Google Cloud resources **that do not have public IP addresses** communicate with the **internet**
- These VMs use a **set of shared public IP addresses** to connect to the internet
- Public NAT **does not rely on proxy VMs**
- Instead, a Public NAT gateway allocates a **set of external IP addresses and source ports** to each **VM** that uses the gateway to create **outbound connections** to the internet



1. Typical NAT Proxies



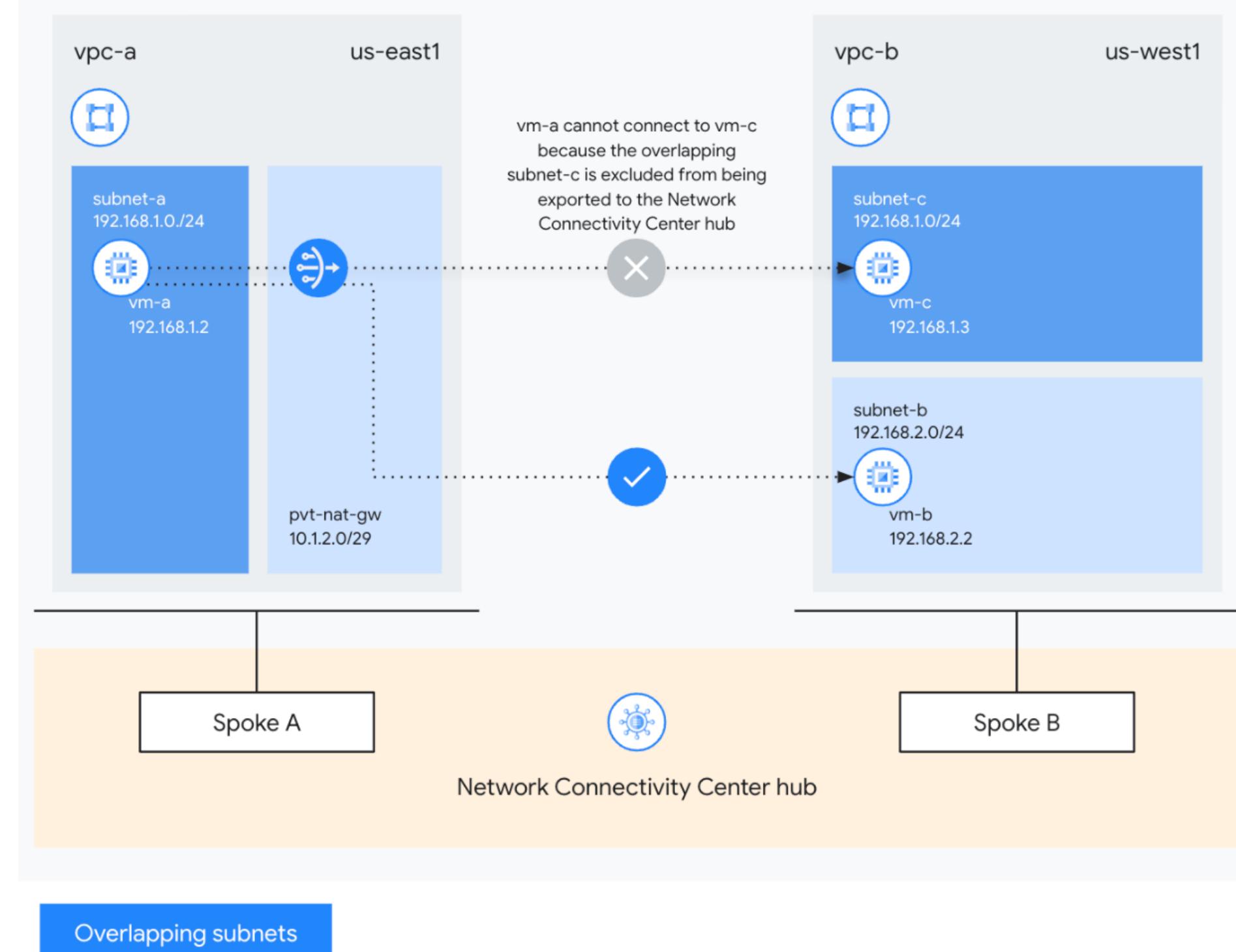
2. Google Cloud NAT

Reference: <https://cloud.google.com/static/nat/images/07.svg>

Google Cloud NAT

- **Private NAT**

- Private NAT enables *private-to-private* translations across Google Cloud networks.
- It helps to perform NAT between multiple VPC networks using Network Connectivity Center



Reference: <https://cloud.google.com/static/nat/images/inter-vpc-nat-flow.png>

Google Cloud NAT

- **Cloud NAT Benefits**

- **Security**

- Reduces the [number of external IPs to VMs](#) which eventually reduces the number egress of firewall rules

- **Availability**

- Cloud NAT is a [distributed, software-defined managed service](#)
 - [No physical VMs or proxy servers](#)

- **Scalability**

- Cloud NAT can be configured to [automatically scale the number of NAT IP addresses](#) that it uses

- **Performance**

- Cloud NAT [does not reduce the network bandwidth](#) per VM
 - Cloud NAT is implemented by Google's [Andromeda software-defined networking](#)

- **Logging**

- For Cloud NAT traffic, you can [trace the connections and bandwidth](#) for compliance, debugging, analytics, and accounting purposes

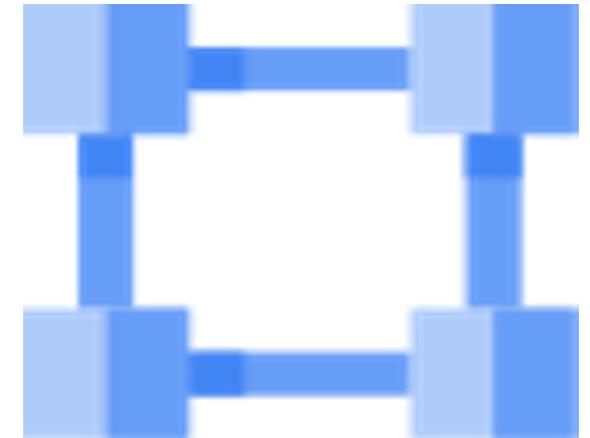
- **Monitoring**

- Cloud NAT exposes key metrics to Cloud Monitoring that [give you insight into your fleet's use of NAT gateways.](#)
 - Metrics [are sent automatically](#) to Cloud Monitoring



Demo

Google Cloud Networking

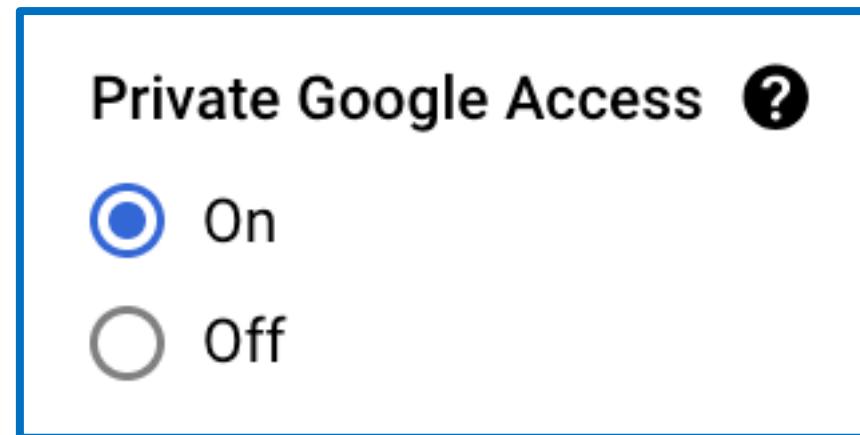


Virtual Private Cloud (VPC)

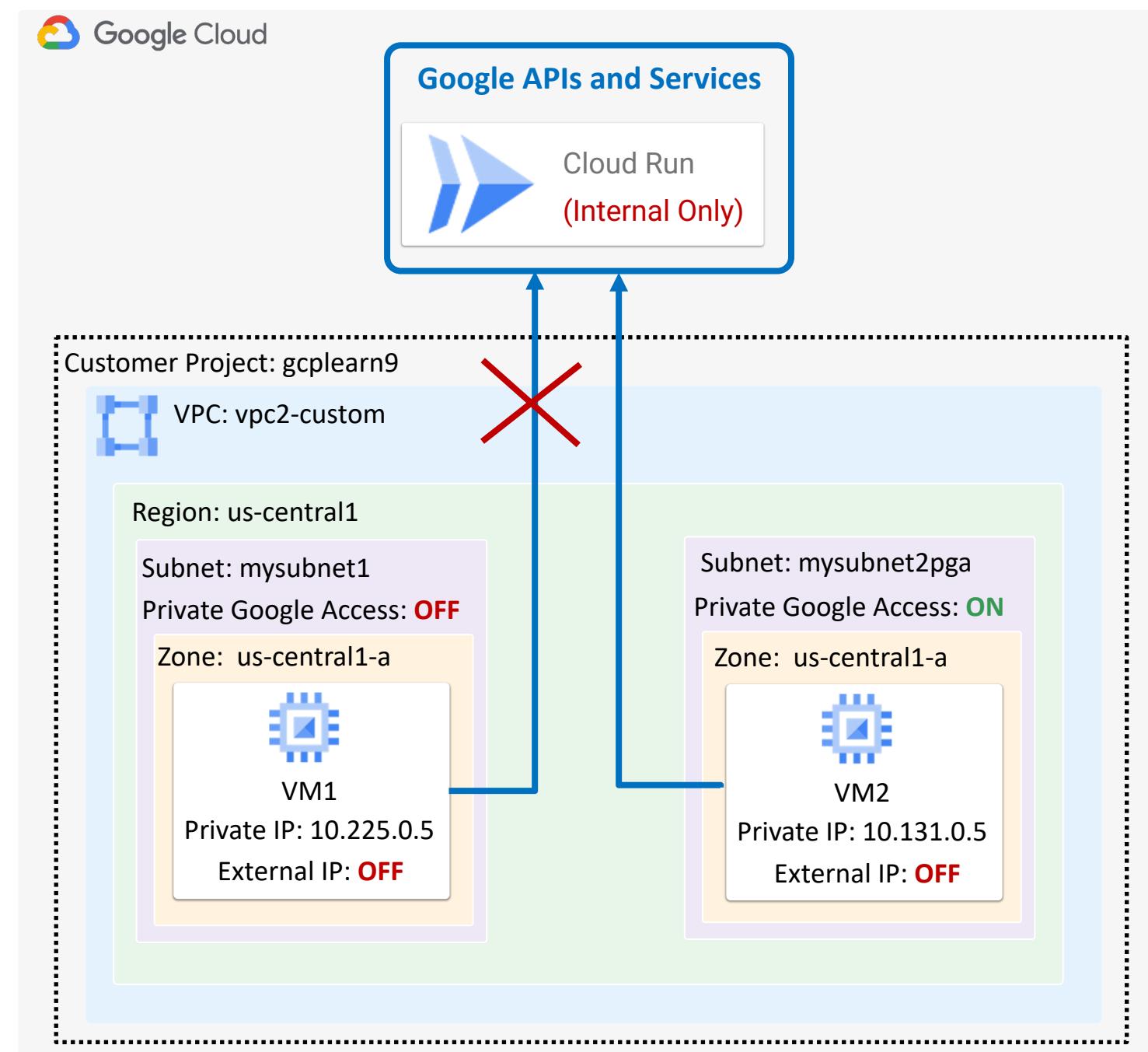
Private Google Access

Google Cloud VPC - Private Google Access

- Private Google Access is a **subnet level setting (ON/OFF)**

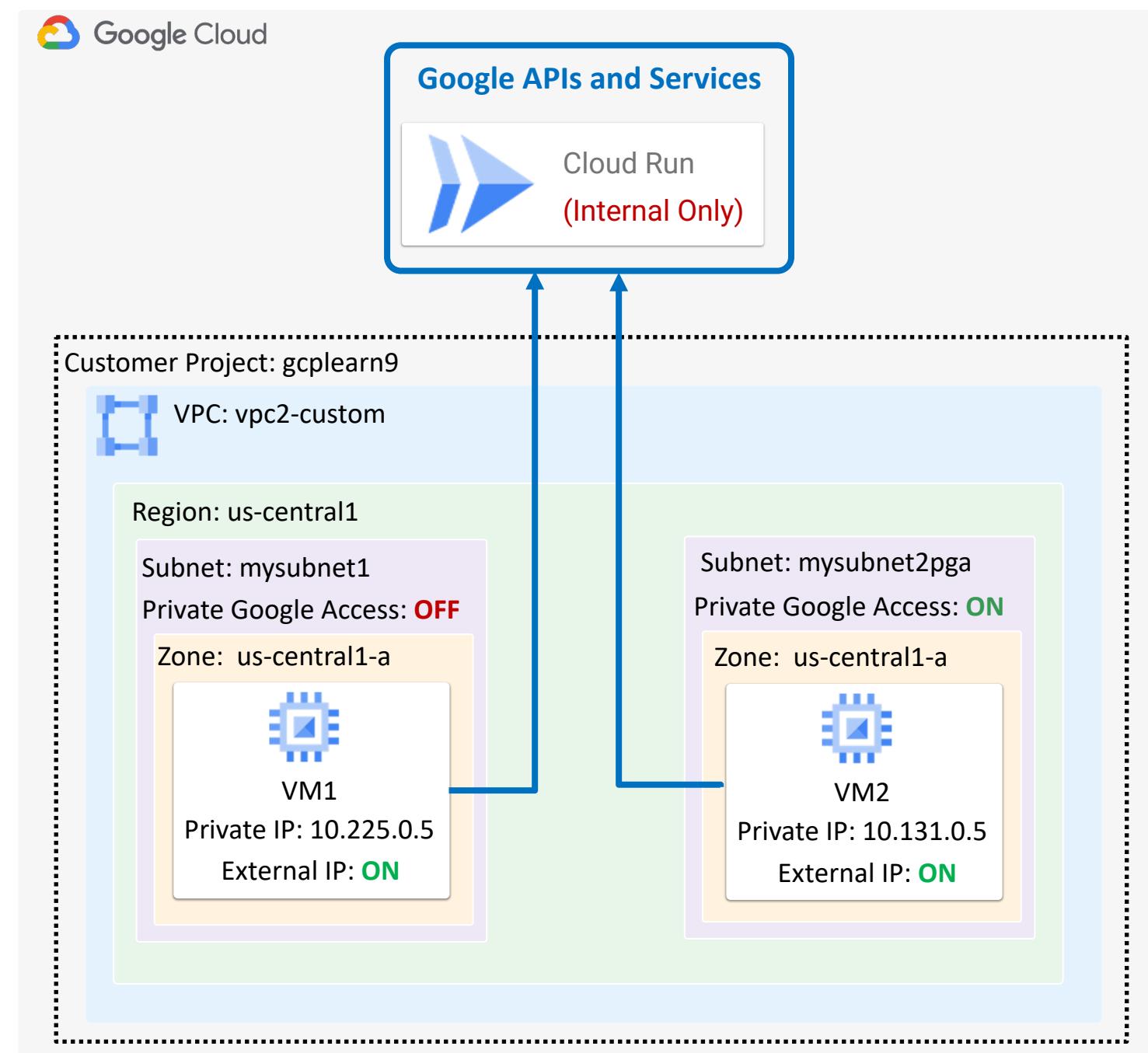


- VM instances that only have **internal IP addresses (no external IP addresses)** can use Private Google Access to **reach** Google APIs and services



Google Cloud VPC - Private Google Access

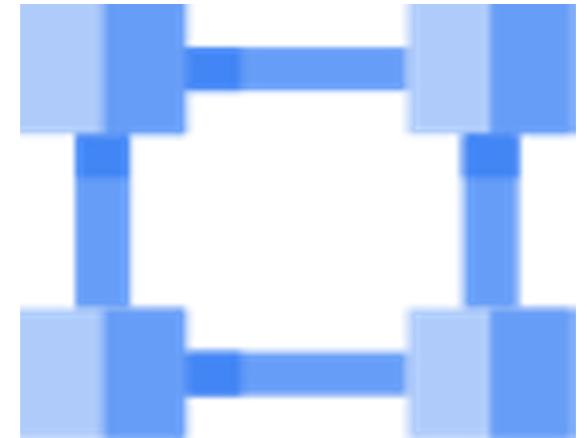
- VMs with **external IPs** can access as usual (no effect)
- **List of Google APIs and Services:**
<https://developers.google.com/apis-explorer/>



Demo



Google Cloud Networking



Virtual Private Cloud (VPC)

VPC Network Peering

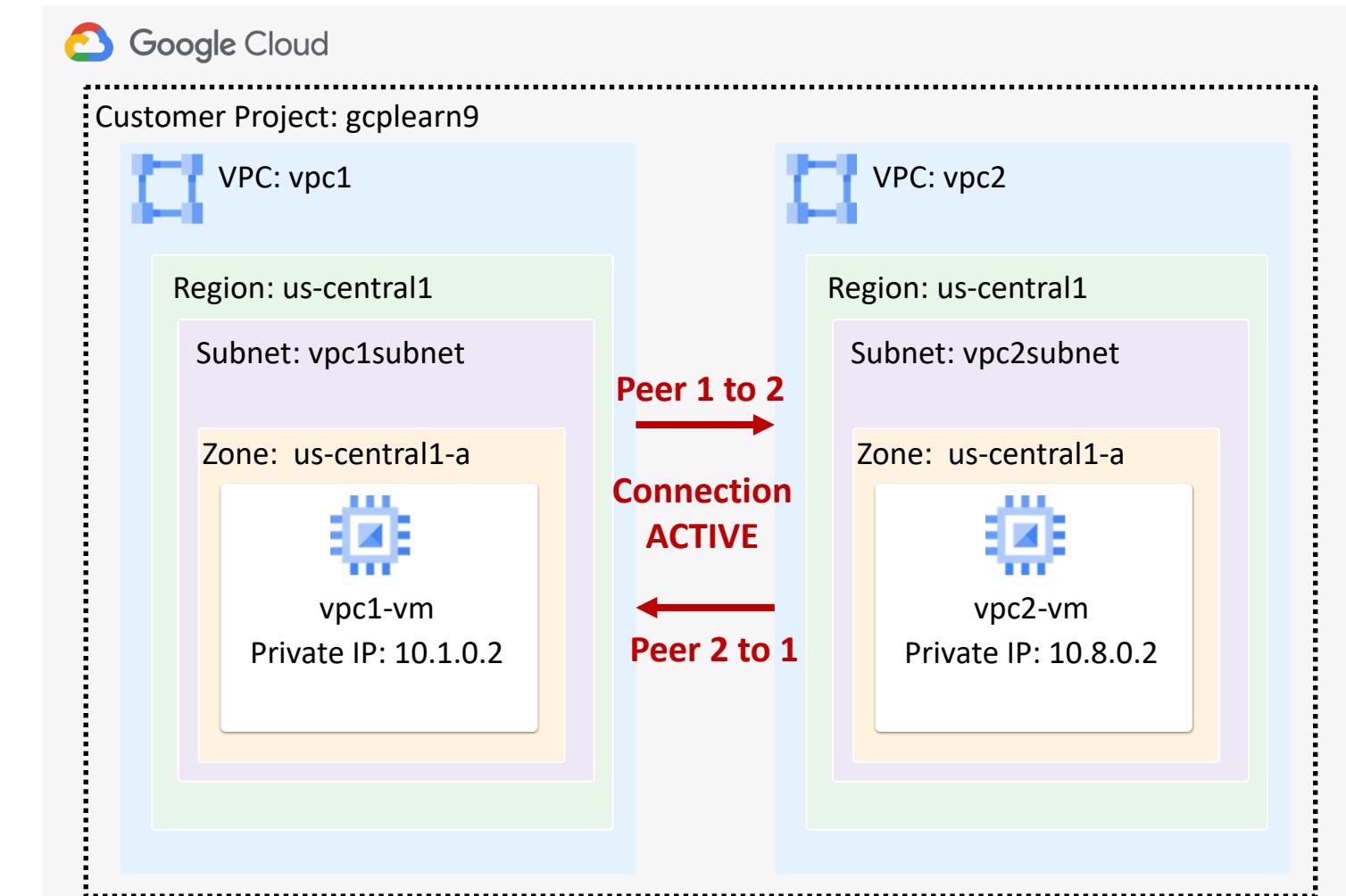
Google Cloud VPC - VPC Network Peering

- **VPC Network Peering:** Connects two VPC networks so that resources in each VPC can communicate to each other.

- Peered VPC network can be in
 - same project
 - different project of same organization
 - different project of different organization

- VPC Network peering works with
 - Compute Engine
 - Google Kubernetes Engine
 - App Engine flexible environments
 - Publish SaaS (Software as a Service) products from Cloud Marketplace

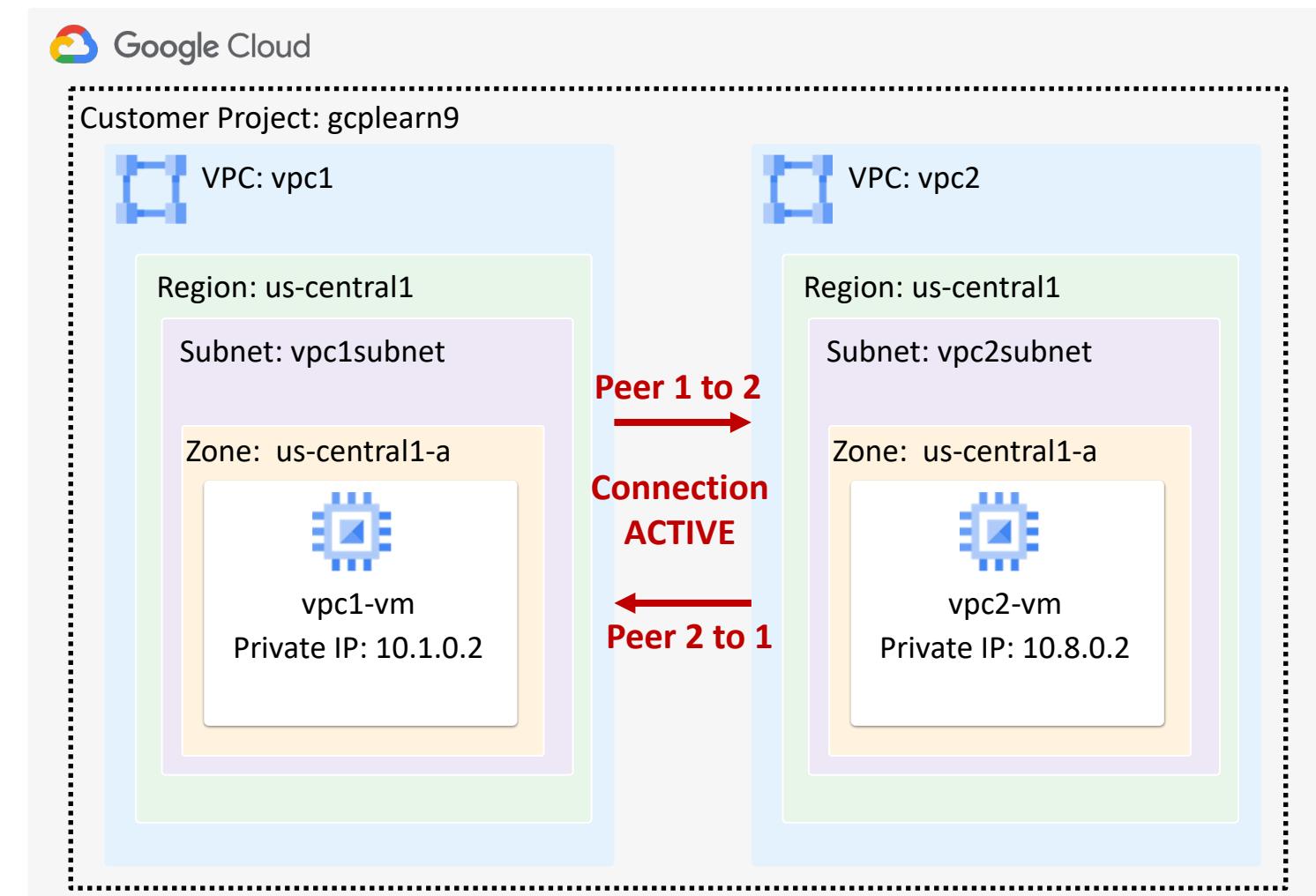
VPC Network Peering



Google Cloud VPC - VPC Network Peering

- Provides low latency, internal IPv4, IPv6 connectivity between VPC networks
- **Important Notes**
- Can't connect two auto-mode VPC networks (because both uses the same CIDR block 10.128.0.0/9)
- For VPC networks participating in peering, it's essential to ensure that their CIDR blocks don't overlap.
- We can connect an auto-mode VPC with a custom mode VPC provided their CIDR blocks don't overlap.

VPC Network Peering



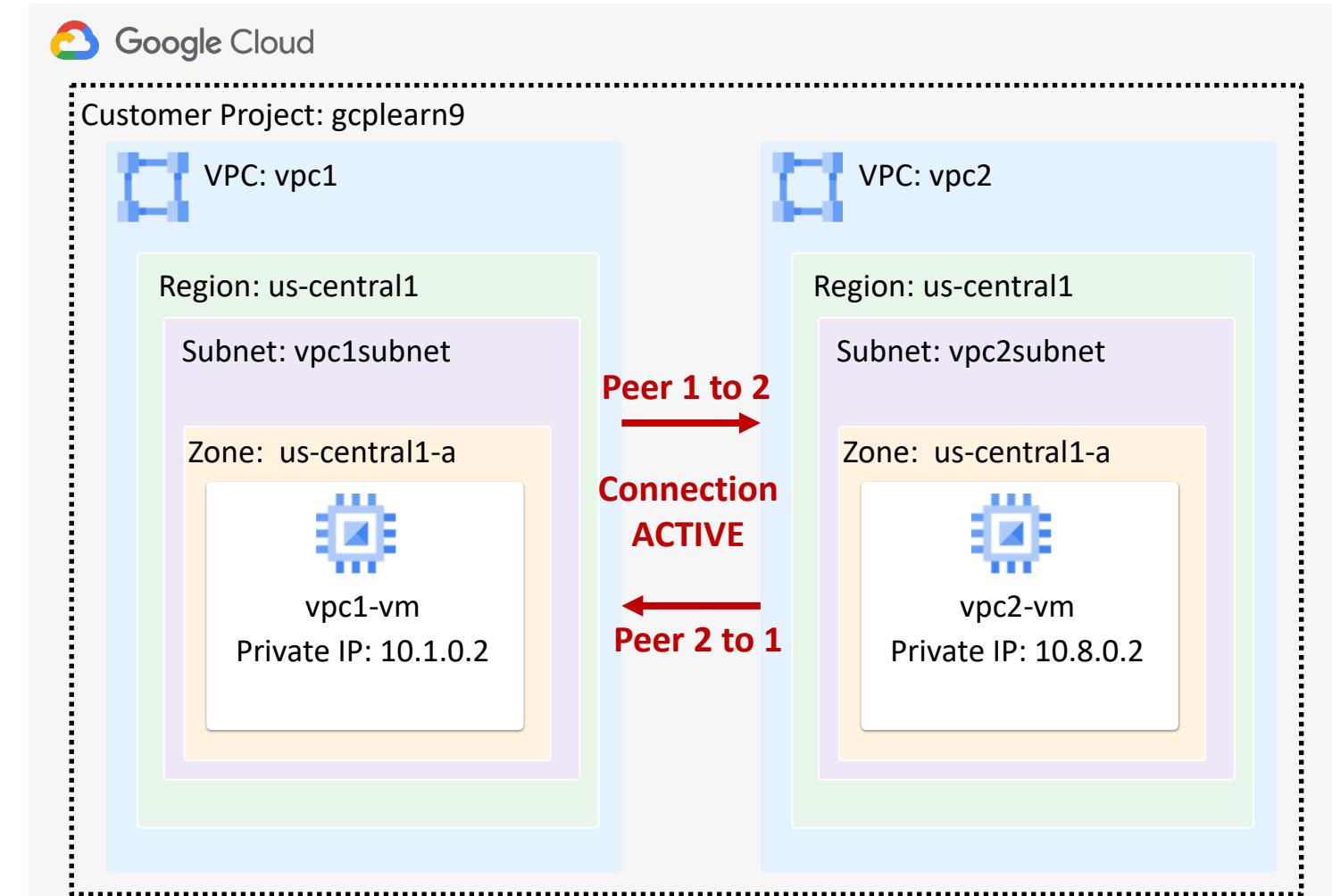
Google Cloud VPC - VPC Network Peering

- **Restrictions:** The following are not going to work across peered VPC networks
 - No subnet IP range **overlap**
 - **Legacy networks** are not supported
 - No Compute Engine DNS
 - Tags and service accounts are not usable
 - GKE is supported only by enabling **IP Aliases or custom routes**
 - Cloud Load Balancing **does not support** having load balancer's frontends and backends in different VPC networks

• Additional Reference:

https://cloud.google.com/vpc/docs/using-vpc-peering#no_dns_across_projects

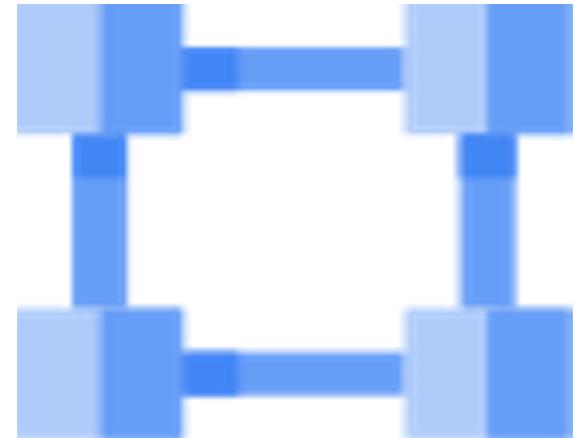
VPC Network Peering



Concept



Google Cloud Networking



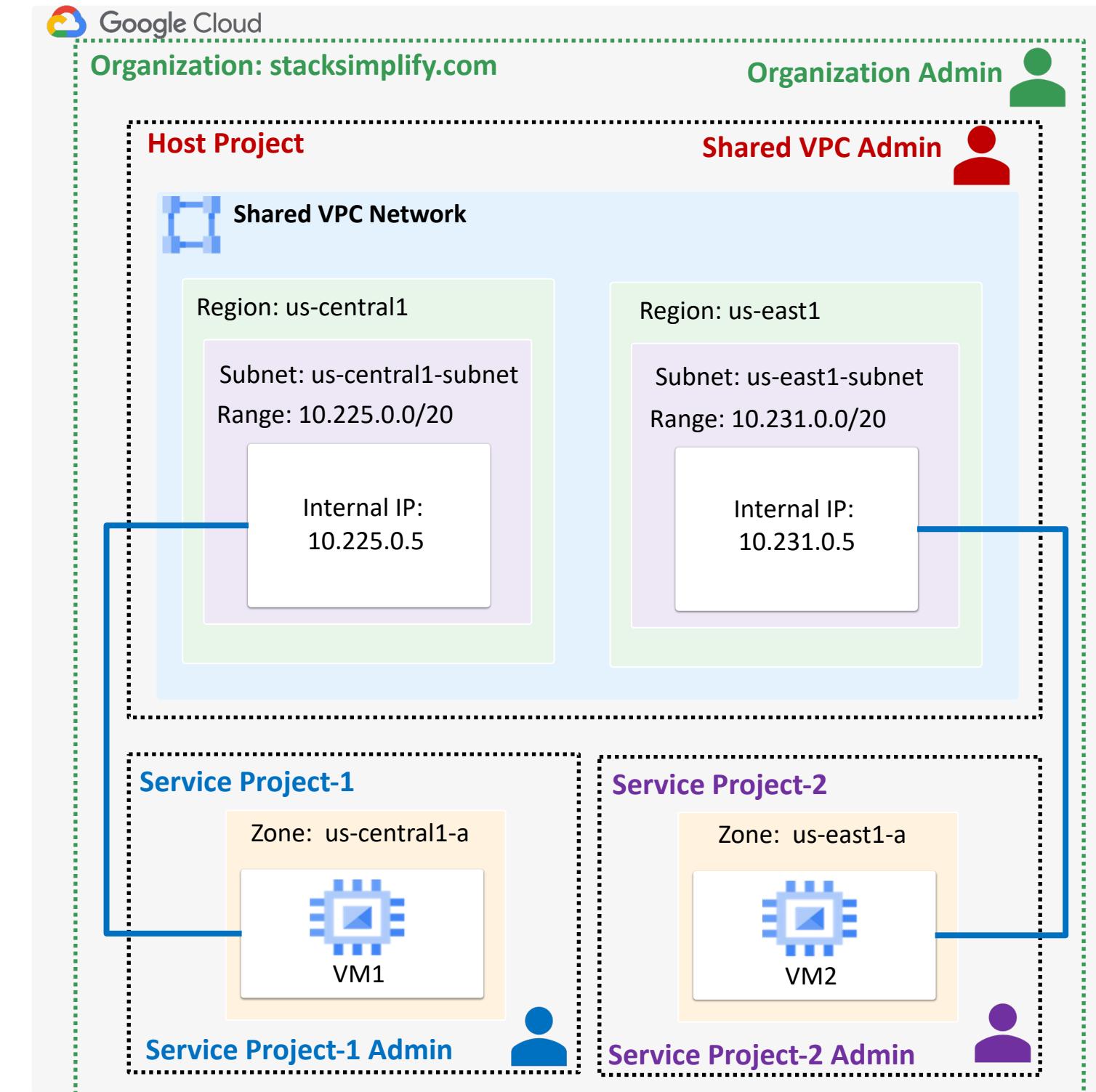
Virtual Private Cloud (VPC)

Shared VPC

Cloud VPC - Shared VPC

- **Shared VPC**

- VPC can be shared **across multiple projects**
- **Important Note:** Shared VPC can be created only with **GCP Organizational accounts**. Not possible via free-tier or personal accounts



Cloud VPC - Shared VPC

- **Standalone Project**

- The project that **does not participate** in shared VPC (regular project and regular vpc)

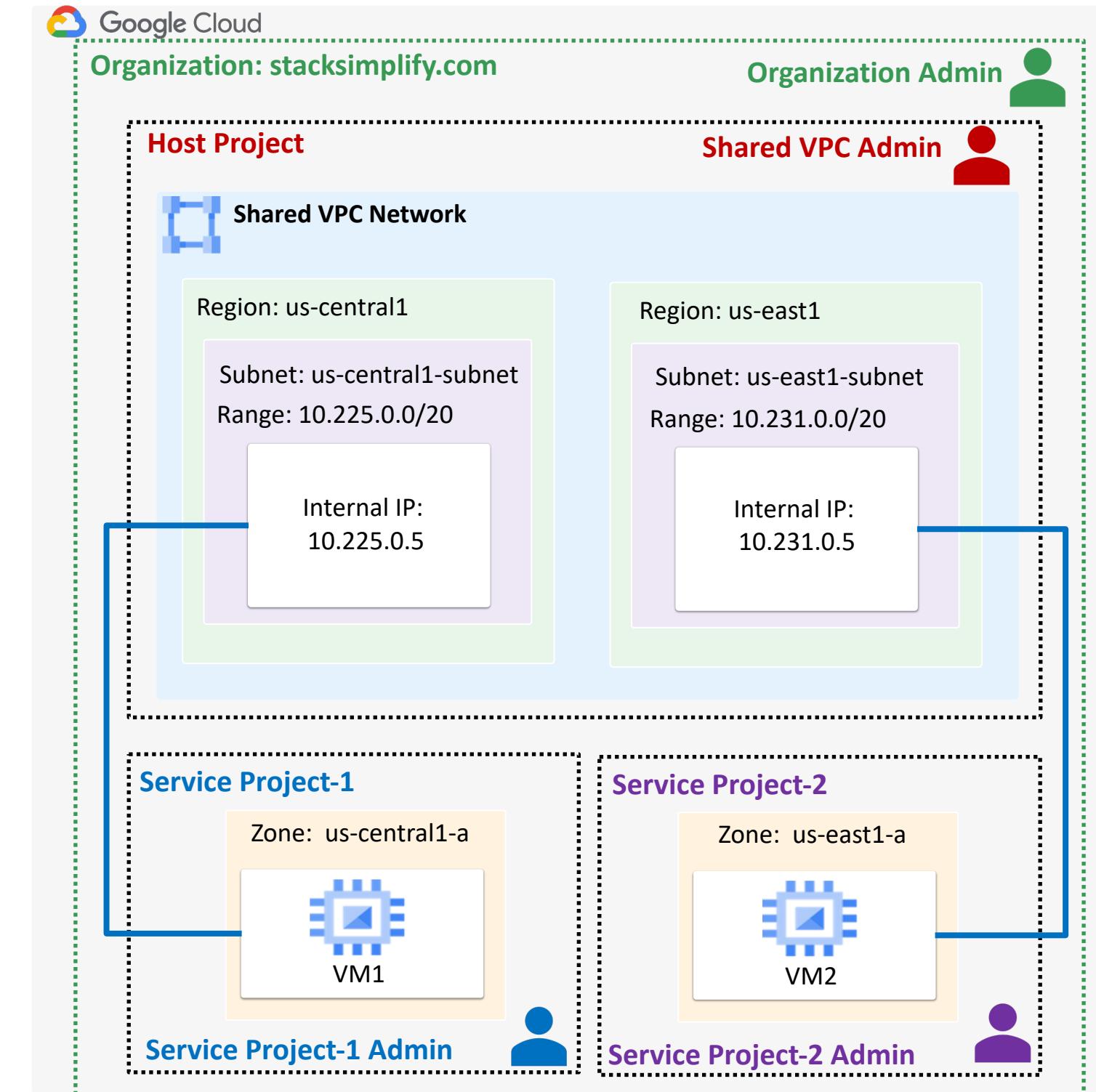
- **Host Project**

- The project where **VPC network is created**

- **Service Project**

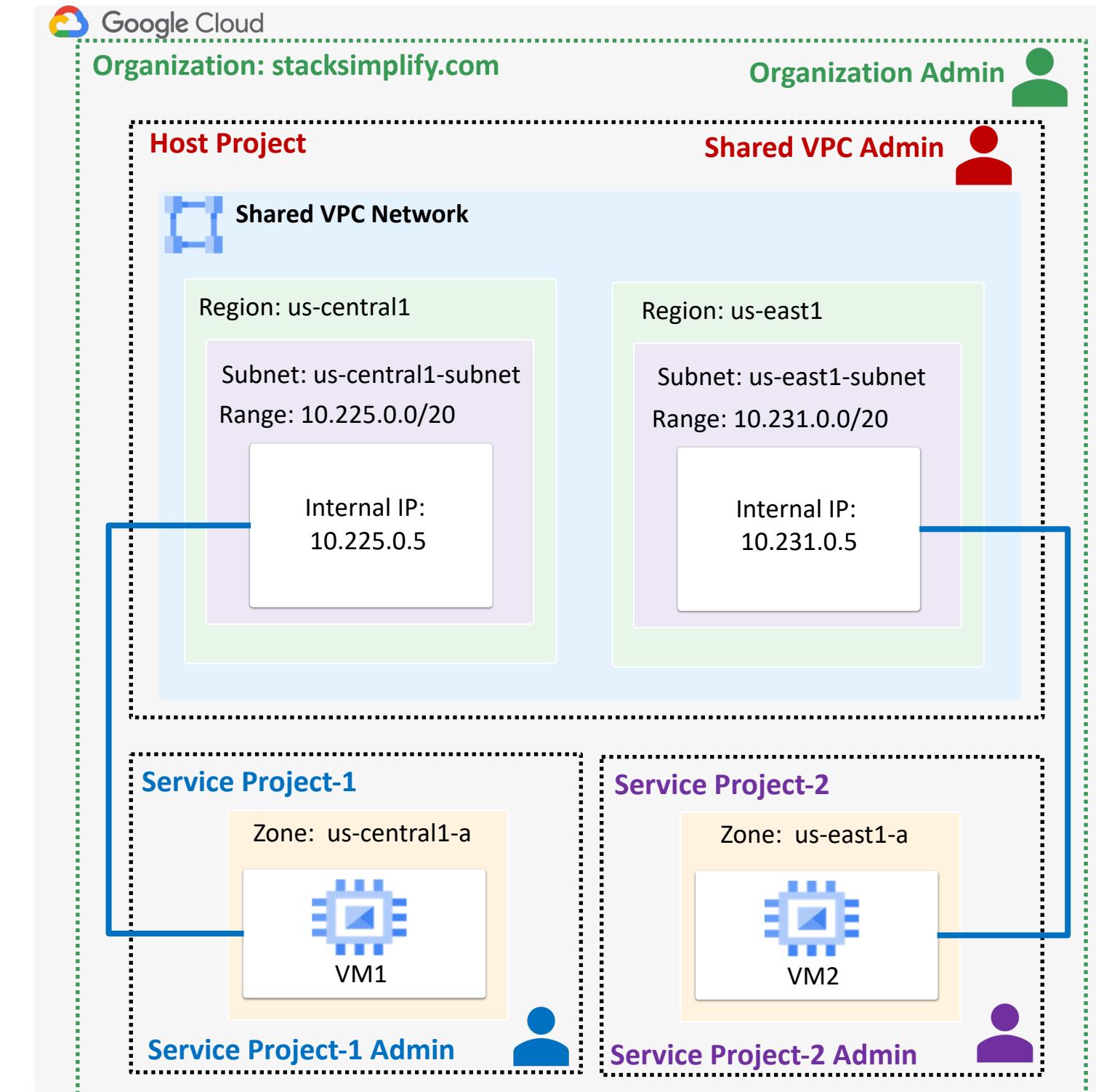
- The projects **which are using the VPC** shared from host project

- **Important Note:** A project cannot be **both** a host and service project



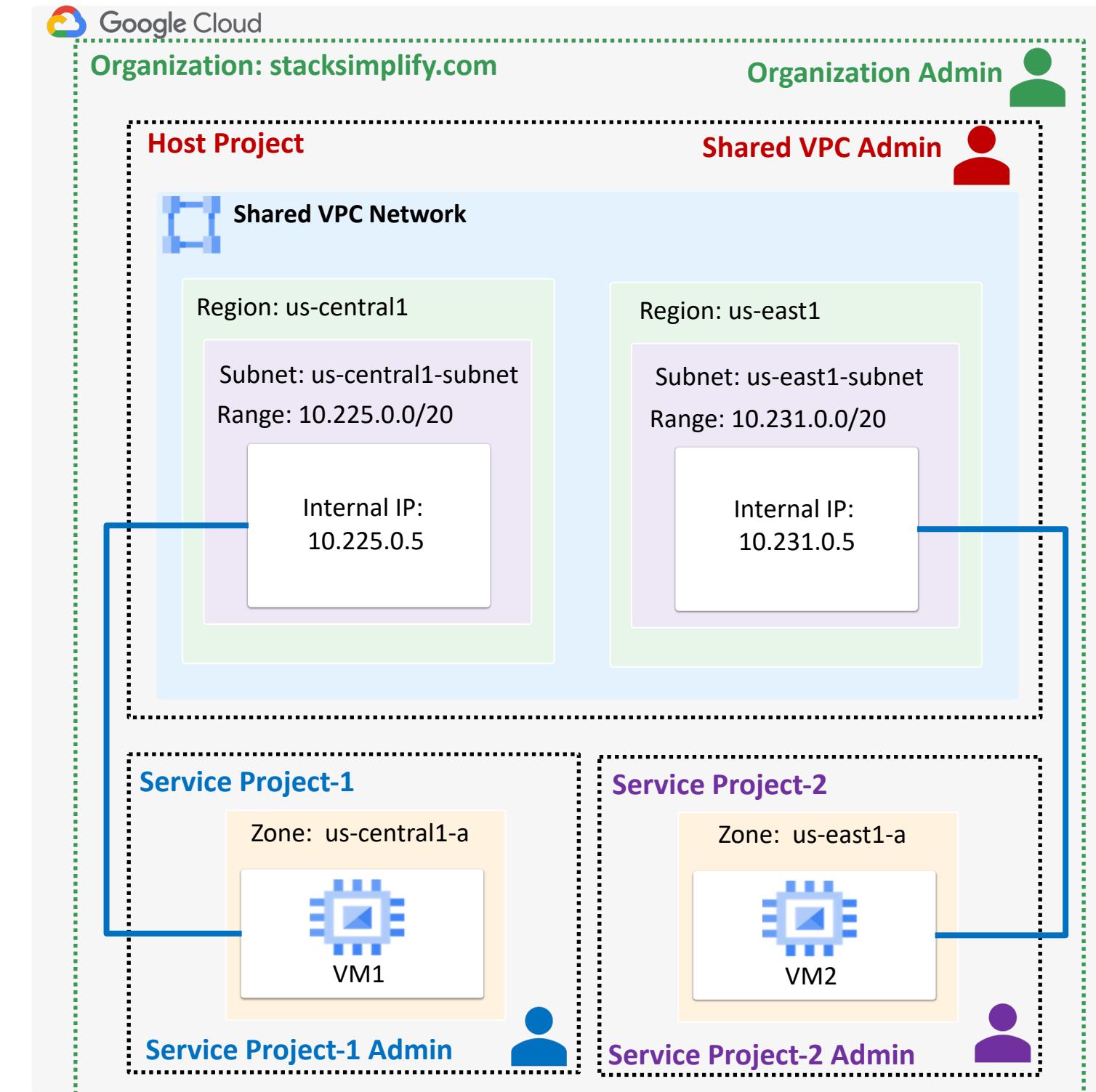
Cloud VPC - Shared VPC

- Why do we need a Shared VPC ?
- Separation of duties
 - Shared VPC Admin takes care of VPC networking
 - Creating Subnets
 - Designing subnet IP ranges
 - All network related
 - Service Project Admin takes care of application resources
 - Creating VM Instances
 - Managing Disks
 - All application related
- Restricting the access
 - Service project admins cannot create or edit or update VPC resources.
 - They cannot make network-impacting changes unknowingly causing major outages.



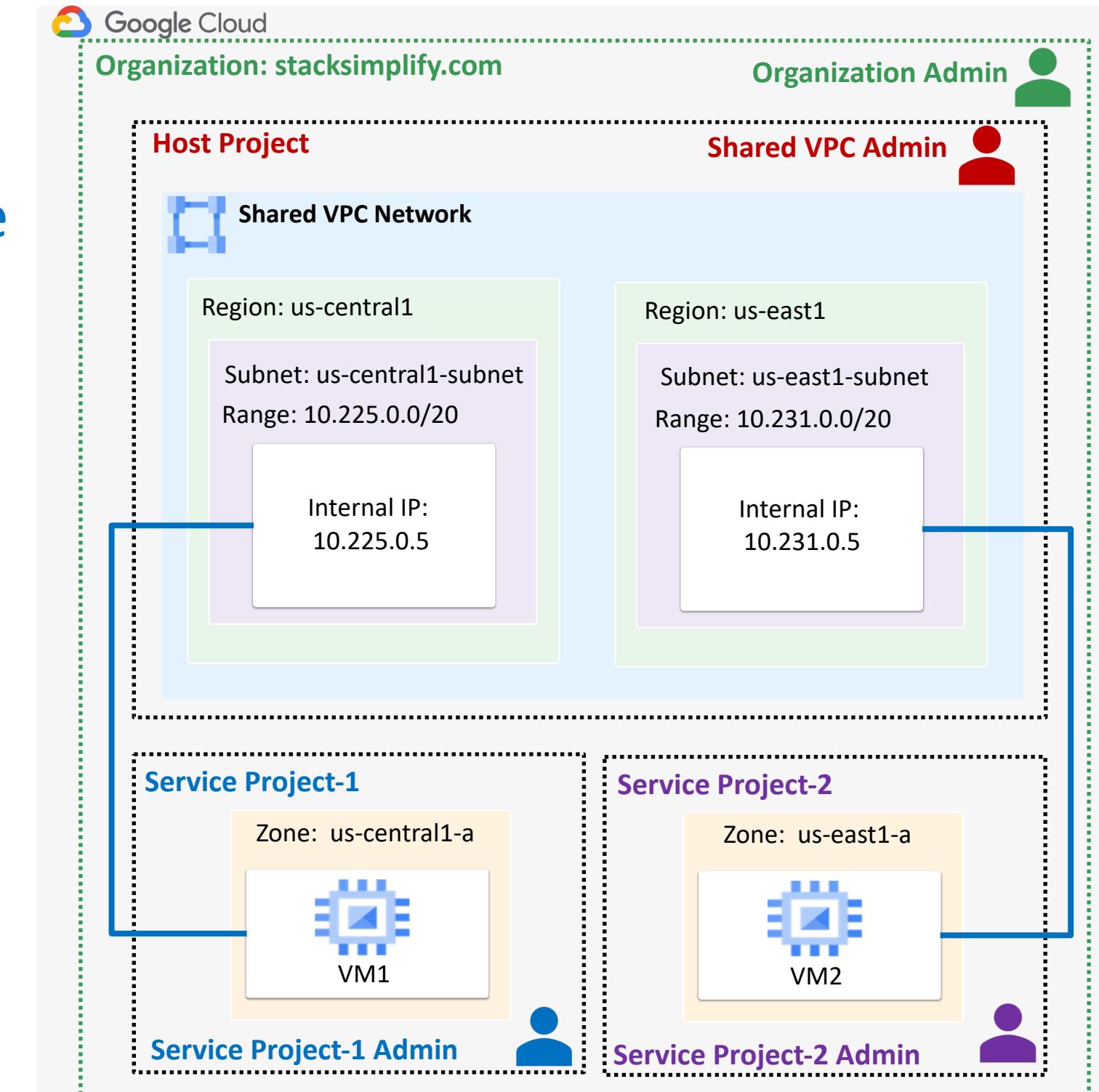
Cloud VPC - Shared VPC

- **Shared VPC Admin:** Can further delegate tasks to
 - **Network Admin**
 - **Role:** compute.networkAdmin
 - Have **full control on all network resources** in host project except firewall rules and SSL certificates
 - **Security Admin**
 - **Role:** compute.securityAdmin
 - Manage **firewall rules and SSL certificates**



Cloud VPC - Shared VPC

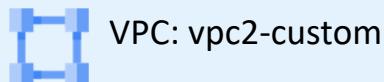
- How does Billing work when using Shared VPC?
- Billing will be attributed to the service project where the resource is located.
- Example-1: If VM instance is located in Service project-1, bill will be attributed to that respective service project only
- Example-2: You can separate your environment bills by having separate service project per environment
- Each service project is billed separately



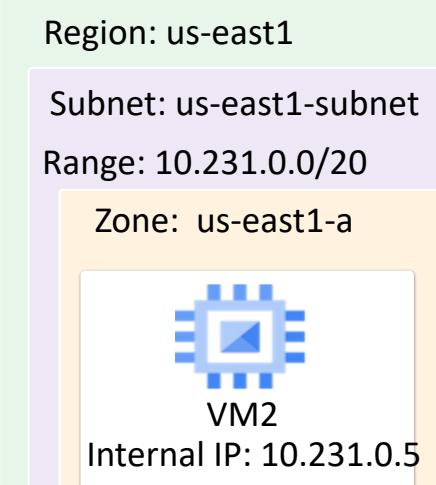
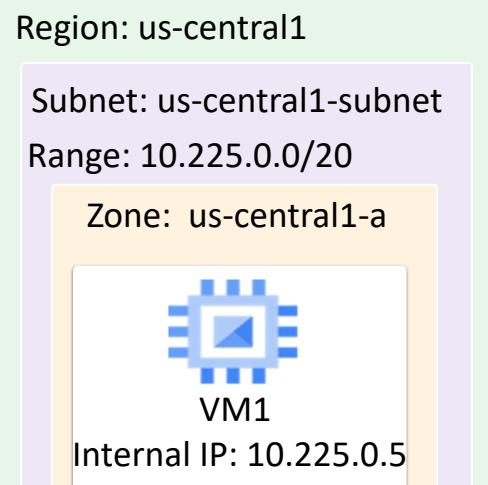
Google Cloud VPC - Shared VPC

Google Cloud

Standalone Project



VPC: vpc2-custom



Organization: stacksimplify.com

Host Project



Shared VPC Network

Region: us-central1

Subnet: us-central1-subnet
Range: 10.225.0.0/20

Internal IP:
10.225.0.5

Organization Admin

Shared VPC Admin

Region: us-east1

Subnet: us-east1-subnet
Range: 10.231.0.0/20

Internal IP:
10.231.0.5

Service Project-1



Service Project-1 Admin

Service Project-2



Service Project-2 Admin



Demo

Google Cloud Networking Cloud Domains



Google Cloud Domains

- **Cloud Domains:** Primarily used for **Domain Registration and management**
- **Key Benefits**
 - Register **new domains**
 - Bills for purchasing/renewing domains will use the same **Cloud Billing account**
 - **Automatic renewal** of registered domains
 - Let's you manage **domain registrations per project**, not per user
 - **Supports DNSSEC** which protects your domains from spoofing and cache poisoning attacks
 - **Tightly integrated** with Google Cloud DNS (Domain Name system) for creating and managing DNS records

Registrations						
Filter		Enter property name or value				
<input type="checkbox"/>	Status	Domain name	DNS	Renewal	Expires/renews on	Privacy protection
<input checked="" type="checkbox"/>	Active	kalyanreddydaida.com	Cloud DNS	Automatic	June 6, 2024	On

Google Cloud Domains

- **Important Note:** From September 7, 2023 onwards Squarespace acquired all domain registrations and related customer accounts from [Google Domains](#)
- Cloud Domains uses [Google Domains](#) as underlying domain registrar
- Now after squarespace acquisition, [Cloud Domains uses Squarespace Domains](#) as underlying registrar
- **How does this impact Cloud Domains in Google Cloud ?**
 - **Complete Faq:** <https://cloud.google.com/domains/docs/faq>
 - In a shorter note, [google supports Cloud Domains even](#) after this acquisition
 - **Following features in Cloud Domain will work as-is**
 - [Searching and registering](#) new domains
 - [Renewing](#) existing domains
 - Update your contact details and DNS settings
 - [Transfer](#) a registered domain to another registrar (GoDaddy, Namecheap, AWS Route53)
 - Transfer domain from [another registrar to Cloud Domain - NOT POSSIBLE](#)

Additional Reference: <https://cloud.google.com/domains/docs/deprecations/feature-deprecations>

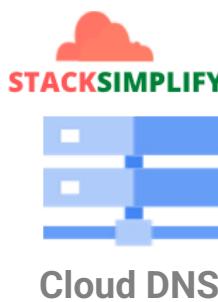
Demo



Google Cloud Networking Cloud DNS



Google Cloud DNS



- **Cloud DNS:** High-performance, resilient, global [Domain Name System \(DNS\)](#) service
- Cloud DNS translates requests for domain names like [stacksimplify.com](#) into IP addresses like 13.224.249.31

```
dkalyanreddy@cloudshell:~ (gcplearn9)$ nslookup stack simplify.com
Server: 169.254.169.254
Address: 169.254.169.254#53

Non-authoritative answer:
Name: stack simplify.com
Address: 13.224.249.31
Name: stack simplify.com
Address: 13.224.249.62
Name: stack simplify.com
Address: 13.224.249.81
Name: stack simplify.com
Address: 13.224.249.73
```

Google Cloud DNS - Terminology

- **DNS Zones:** It is a **container of DNS records** for the same DNS name suffix (Example suffix: stacksimplify.com)

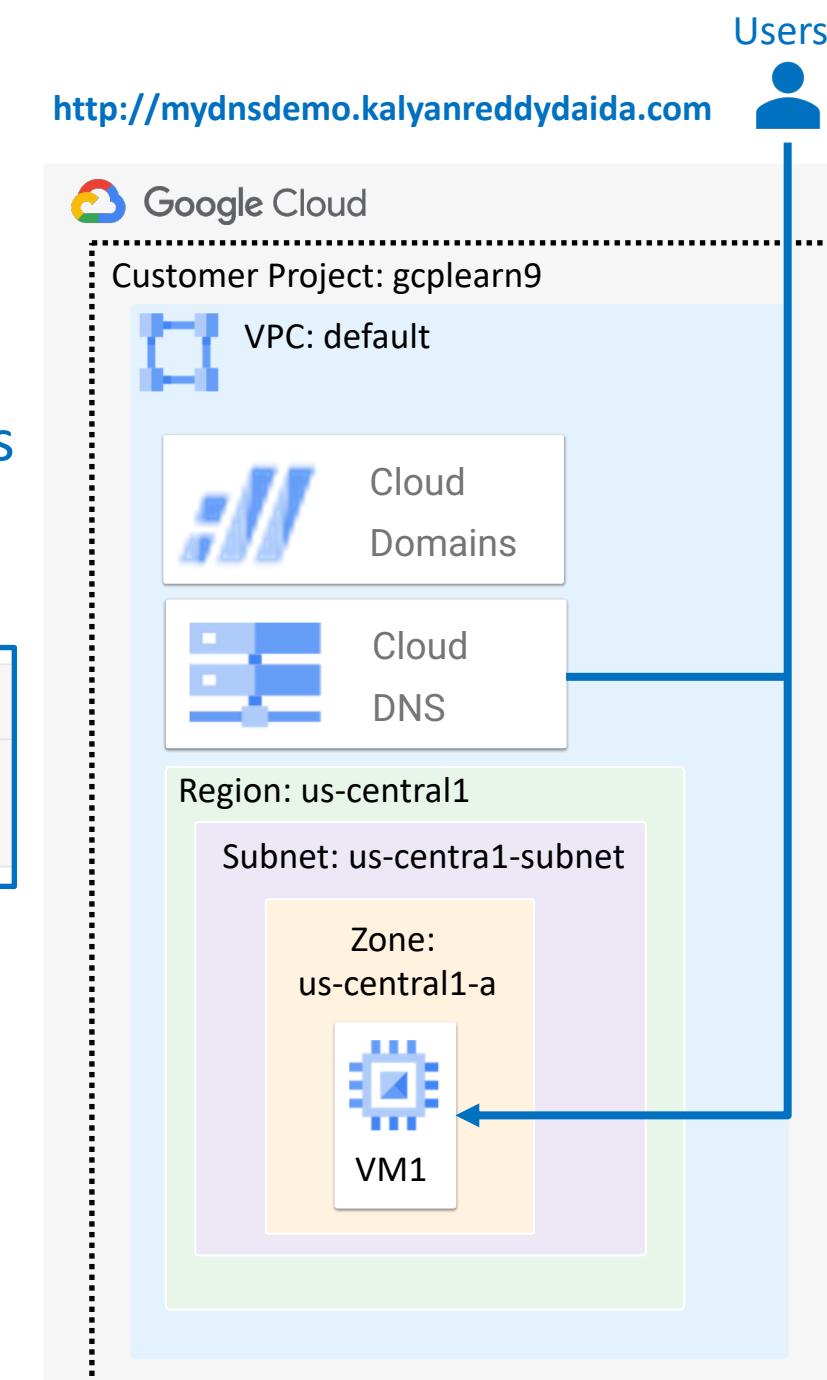
- **Public Zone**

- **Visible to the internet**
- Primarily used for **DNS management** of your **internet facing applications**
- **DNSSEC (DNS Security)** protects your domains from **spoofing and cache poisoning attacks**

Zone name ↑	DNS name	DNSSEC	Description	Zone type
kalyanreddydaida-com	kalyanreddydaida.com.	On	DNS zone for domain: kalyanreddydaida.com	Public

- **Private Zone**

- Contains DNS records that are **only visible internally** within your Google Cloud networks
- Easy to manage **internal DNS solution** for all our internal needs (For Internal applications, VM Instances etc)



Google Cloud DNS - Terminology

- **Record Sets:** actual DNS records
 - DNS Name to IP Address mapping
 - We have different record types, but primarily we use **Address record (A) type** which maps hostnames to IPv4 address

DNS Record Set

Create record set

DNS name: myapp1 .kalyanreddydaida.com.

Resource record type: A

TTL *: 5

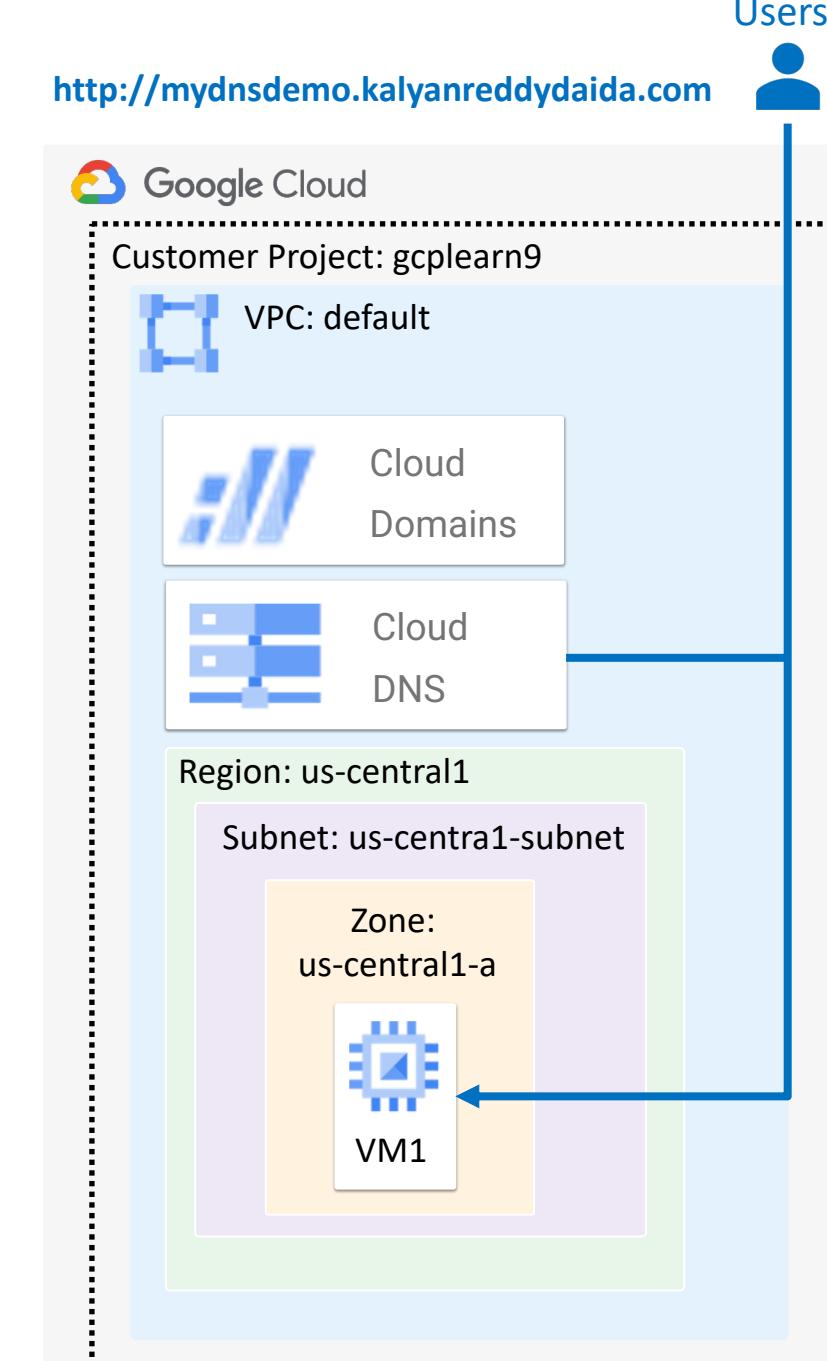
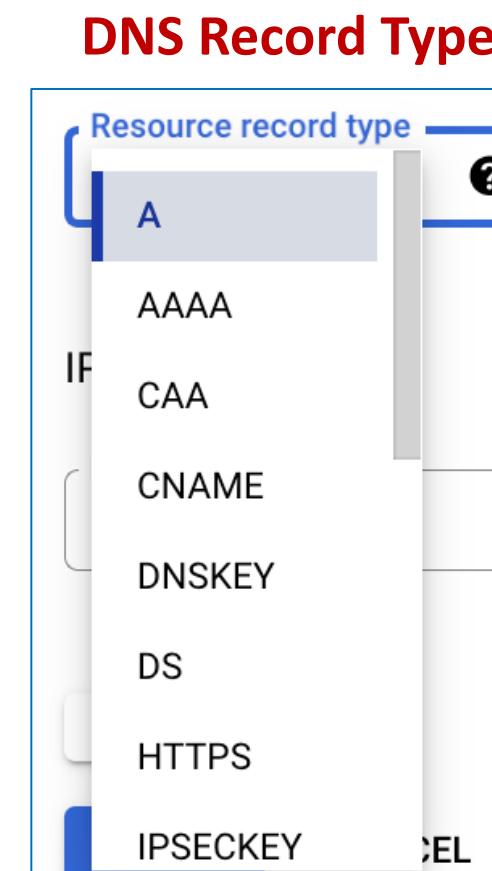
TTL unit: minutes

IPv4 Address 1 *: 108.157.238.34

SELECT IP ADDRESS

+ ADD ITEM

CREATE CANCEL



https://cloud.google.com/dns/docs/records-overview#supported_dns_record_types

Google Cloud DNS - Features

- **Integration with Cloud IAM**

- Provides secure domains management with **full control and visibility** for domain resources

- **Integration with Cloud Logging**

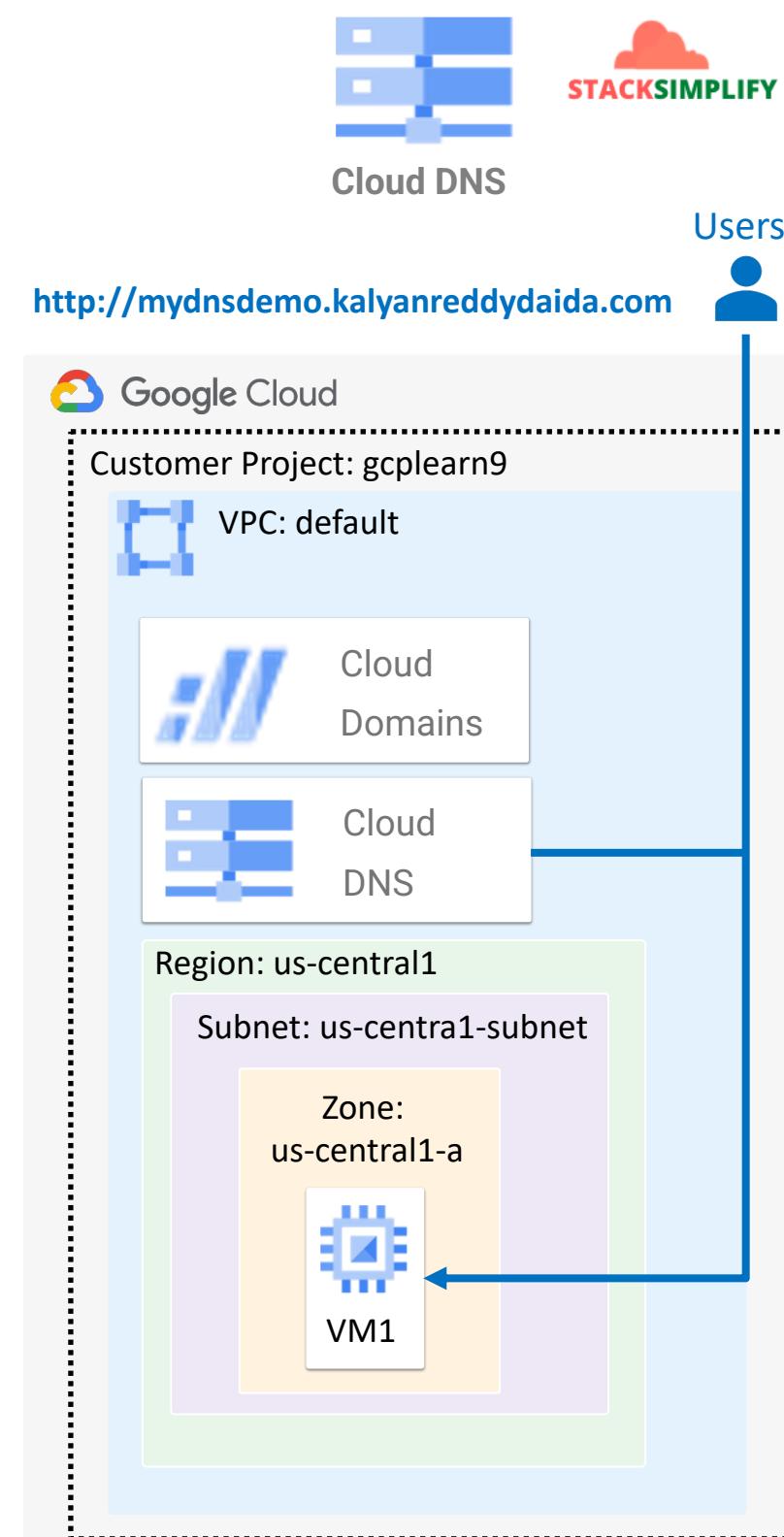
- **Private DNS logs** a record for **every DNS query** in Cloud Logging
- We can **view and export logs** to supported destinations

- **Fast anycast Nameservers**

- **Anycast:** Uses the **nearest nameserver to users** for DNS lookup and resolution
- Provides **very high availability** across globe and **low latency**

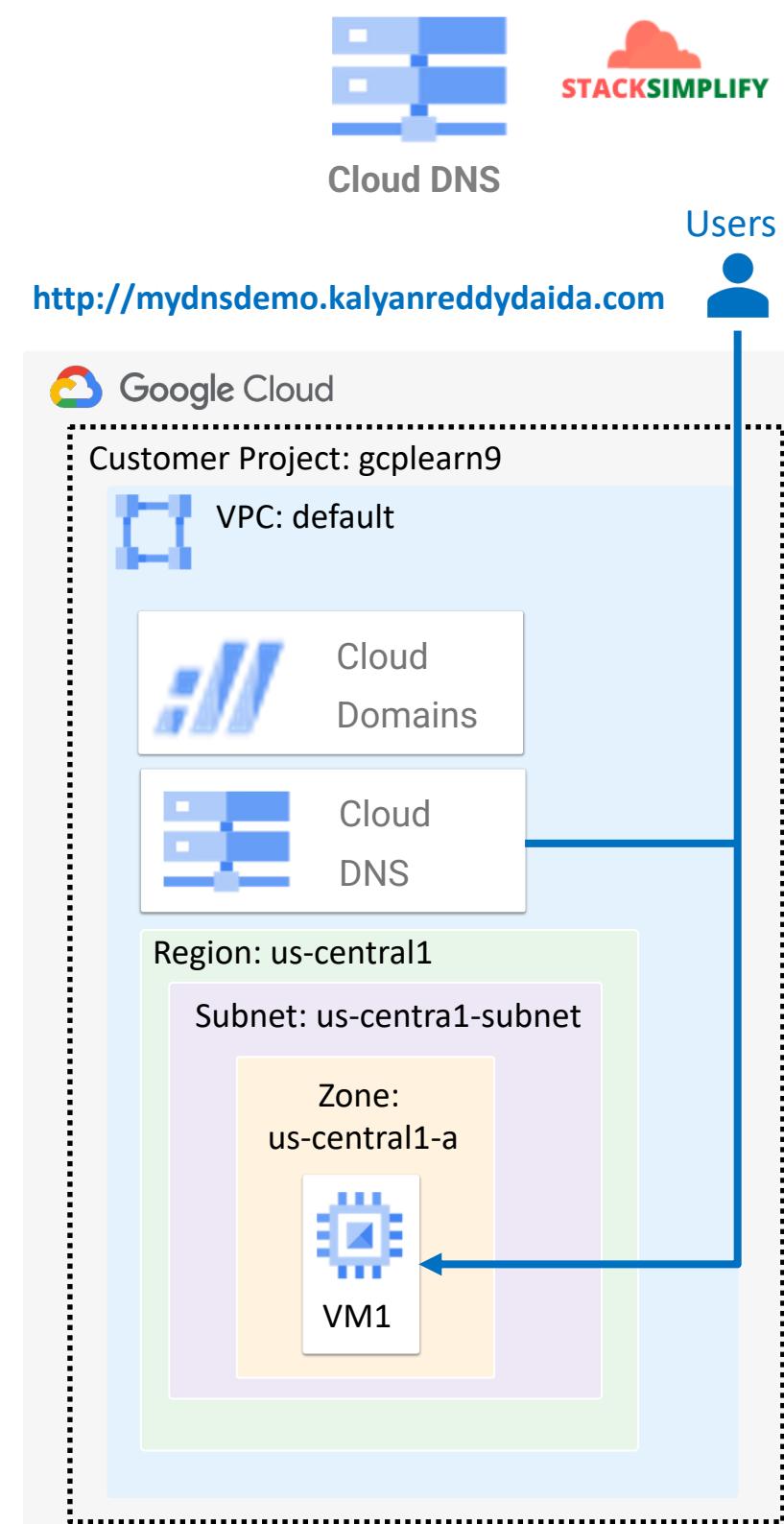
- **DNS registration and management**

- Cloud DNS is **tightly integrated** with Cloud Domains which takes care of DNS registration and management



Google Cloud DNS - Features

- **Container-native Cloud DNS**
 - Natively integrated with Google Kubernetes Engine (GKE)
 - Provides in-cluster Service DNS resolution
 - Provides high-throughput, scalable DNS resolution for every GKE node
- **DNS Peering**
 - Sharing DNS data
 - All or a portion of DNS namespace can be configured to be sent from one network to another
 - Will respect all DNS configurations defined in peered network
- **DNS Forwarding**
 - Primarily used for hybrid-cloud architecture
 - Forward DNS queries from Cloud DNS to on-premise DNS servers where actual DNS resolution takes place



**Demo**

Google Cloud Networking

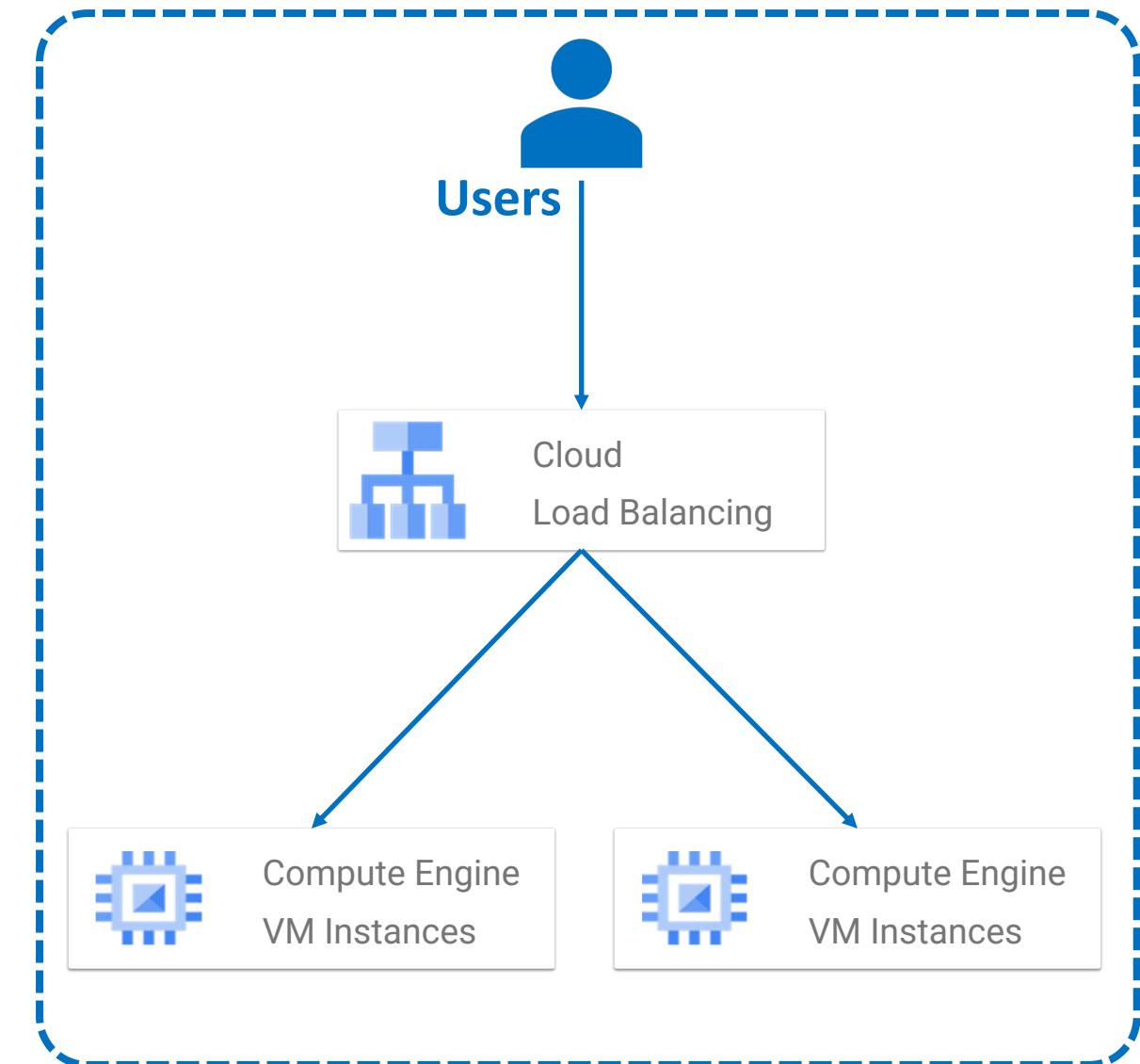
Cloud Load Balancing

Global & Regional

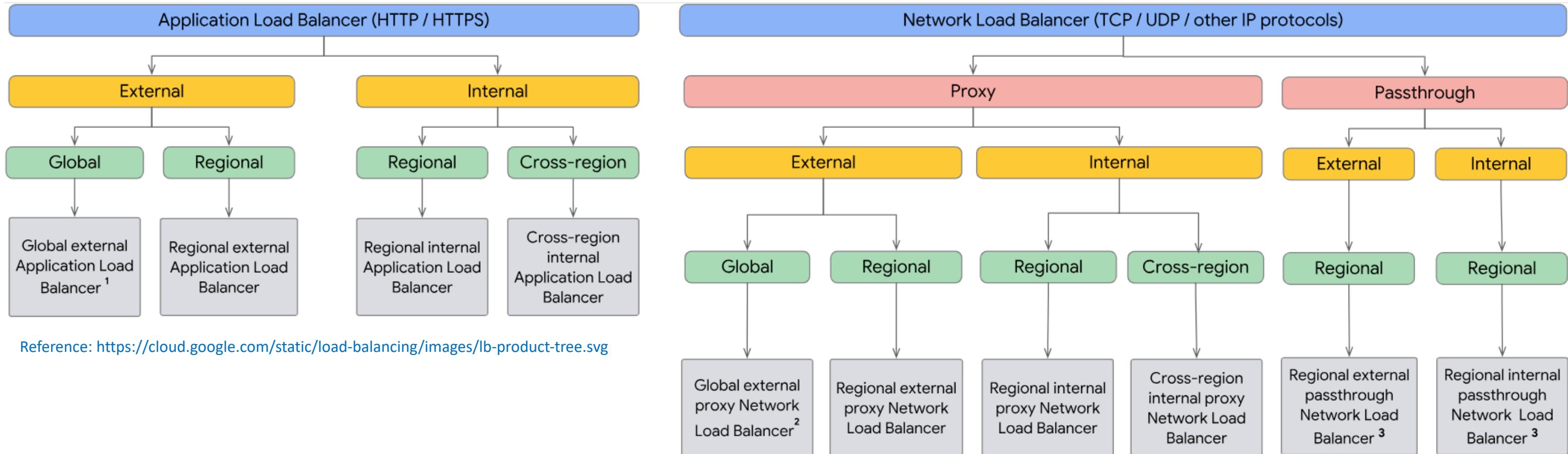
Intro

Google Cloud Load Balancing

- **Cloud Load Balancing:** Distributes user traffic across **multiple instances** of your applications
- Primarily used for providing **HIGH AVAILABILITY** for your applications
- Fully **distributed** managed service
 - **Global:** load balance **across regions**
 - **Regional:** load balancer across zones in a region
- **Seamless autoscaling**
 - **Software-defined** (No VMs or hardware to be provisioned)
 - **Scales** as user traffic grows
 - Zero to full throttle in seconds
- **Global with single anycast IP**
 - **Anycast IP:** Routes **users to nearest available region** (low latency)
 - Provides **cross-region load balancing**, including **automatic multi-region failover**
 - Reacts **instantaneously** to changes in users, traffic, network, backend health, and other related conditions



Google Cloud Load Balancing - Types



- We can create approximately **10 types** of Load Balancers
- On a very high-level, these are **categorized** as
 - Application Load Balancers
 - Network Load Balancers - Proxy
 - Network Load Balancers - Pass-through

Google Cloud Load Balancing - Protocols

OSI Model

- **Layer7 Load Balancer**

- Application load balancer
 - [HTTP or HTTPS](#)

- **Layer4 Load Balancer**

- Network Load Balancers - Proxy

- [TCP \(TCP Proxy\)](#)
- [TCP with optional SSL offload \(SSL Proxy\)](#)

- Network Load Balancers - Pass-through

- [TCP, UDP, ICMP, ICMPv6, SCTP, ESP, AH, and GRE](#)

- **Important Note:** Each OSI layer uses layers below it. (Example: layer7 uses layer 1 to 6 and 7)

Layer	Layer Number	Description	Protocols and Technologies
Application	7	Provides high-level APIs and protocols for application services.	HTTP, SMTP, FTP, DNS, DHCP, SNMP
Presentation	6	Translates data between the application and network layers. Handles encryption, compression, and data formatting.	SSL/TLS, JPEG, GIF, MPEG, ASCII, EBCDIC
Session	5	Establishes, maintains, and terminates connections between applications. Manages sessions and dialogues.	NetBIOS, TLS, SSH, RPC, PPTP, SIP, SOCKS
Transport	4	Manages end-to-end communication between hosts. Ensures reliable data transfer and error recovery.	TCP, UDP, SCTP
Network	3	Routes data packets between devices on different networks. Handles logical addressing and routing.	IP, ICMP, ARP, OSPF, BGP, RIP, IPsec
Data Link	2	Controls the transmission of data over the physical layer. Manages access to the physical medium.	Ethernet, PPP, HDLC, Wi-Fi, VLANs, MAC addresses
Physical	1	Transmits raw data bits over a physical medium. Deals with electrical, mechanical, and procedural aspects.	Ethernet, Wi-Fi, Fiber optics, RS-232, USB, Bluetooth

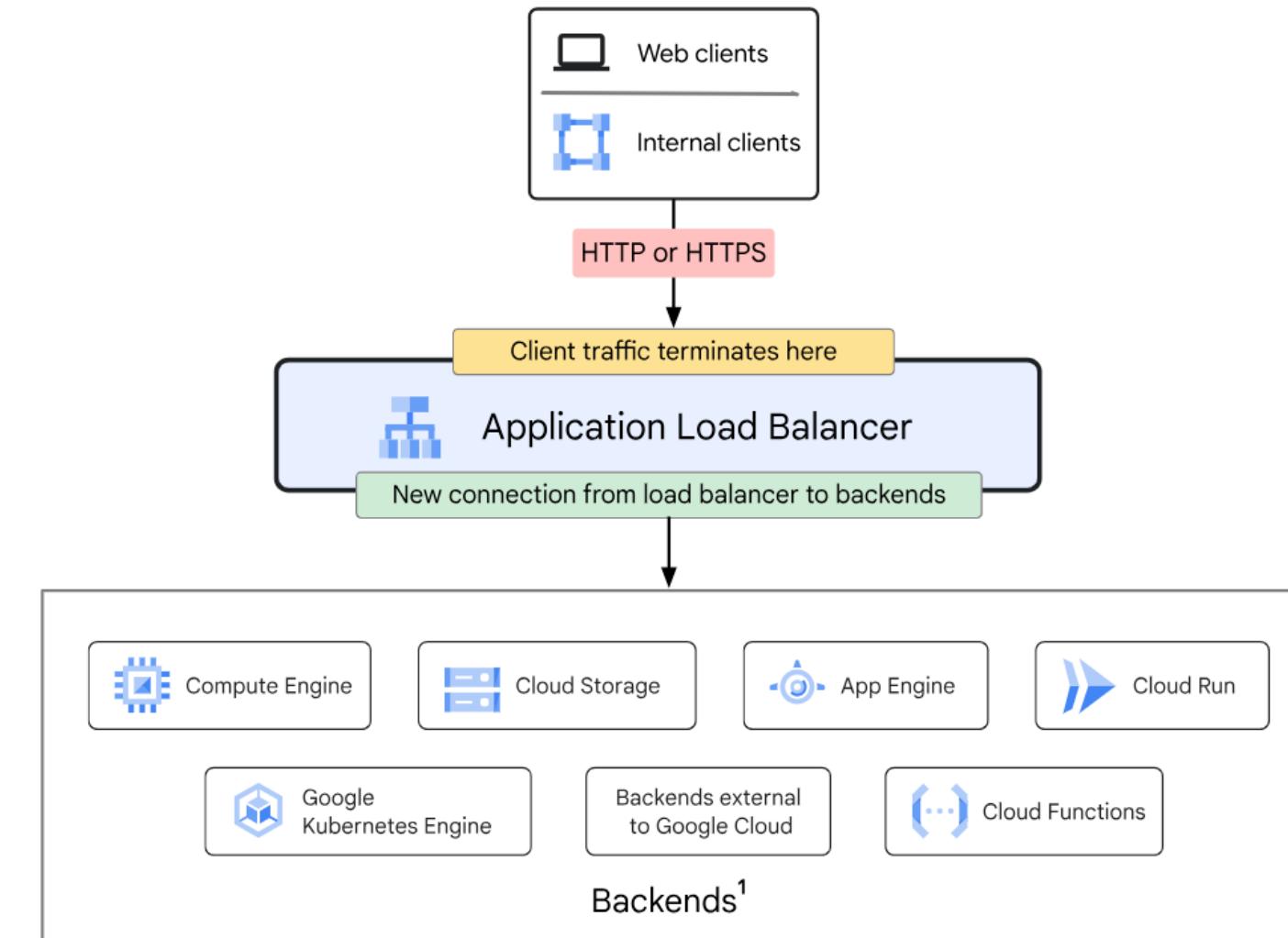
Google Cloud Load Balancing - Terminology

- **Backend**

- Google Cloud services that **receive traffic** from Load Balancer
 - Instance Groups
 - Cloud Storage
 - App Engine
 - Cloud Run
 - GKE
 - Cloud Functions
 - Backends external to Google Cloud

- **Frontend**

- Define **Load Balancer IP address, port, protocol (HTTP, HTTPS, TCP, UDP, SSL etc)**
- We will use this IP address to **access** the load balancer



Reference: <https://cloud.google.com/static/load-balancing/images/application-load-balancer.svg>

Google Cloud Load Balancing - Terminology

- **Routing Rules (Application LBs)**

- **Path based Routing**

- App1: stacksimplify.com/app1
 - App2: stacksimplify.com/app2

- **Host based Routing**

- App1: app1.stacksimplify.com
 - App2: app2.stacksimplify.com

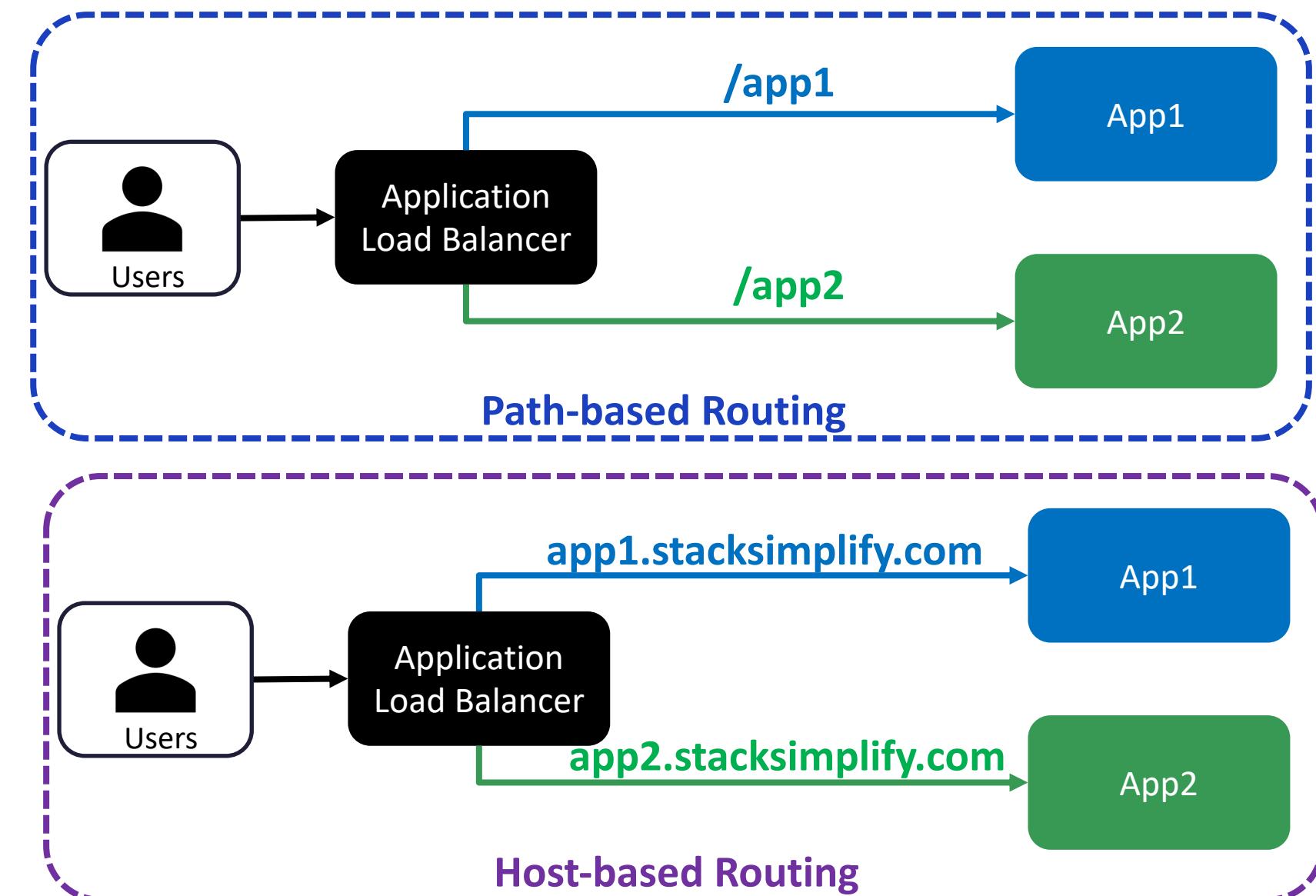
- **Rewrite the Requested URL**

- **Host Rewrite:** app1.stacksimplify.com to app1.terraformguru.com
 - **Path Rewrite:** stacksimplify.com/app1 to stacksimplify.com/app1new

- **Add and remove** request and response headers

- Configure a **URL redirect**

- Many many more things we can do



**Demo**

Google Cloud Networking



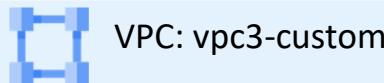
Cloud Load Balancing

Regional Managed Instance Groups

**Demo-1**



Customer Project: gcplearn9

Global
Health Check

Region: us-central1

Subnet: us-central1-subnet

Zone:
us-central1-aZone:
us-central1-b

Instance Group: MIG1

Region: us-east1

Subnet: us-east1-subnet

Zone:
us-east1-aZone:
us-east1-b

Instance Group: MIG2

Cloud Load Balancing

- **Demo-1: Create Regional Managed Instance Groups**
- **Step-1:** Create VPC in *custom mode*
- **Step-2:** Create Subnets
 - us-central1-subnet
 - us-east1-subnet
- **Step-3:** Create Firewall Rules
 - Ingress rule that allows traffic from the Google Cloud health checking systems
- **Step-4:** Create Global Health Check
- **Step-5:** Create Instance Template
- **Step-6:** Create Managed Instance Group in two regions

Demo-1

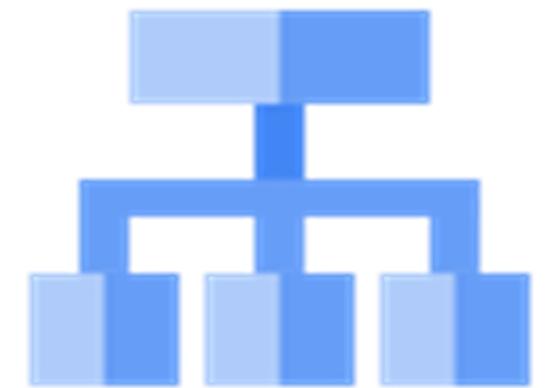
Demo



Google Cloud Networking

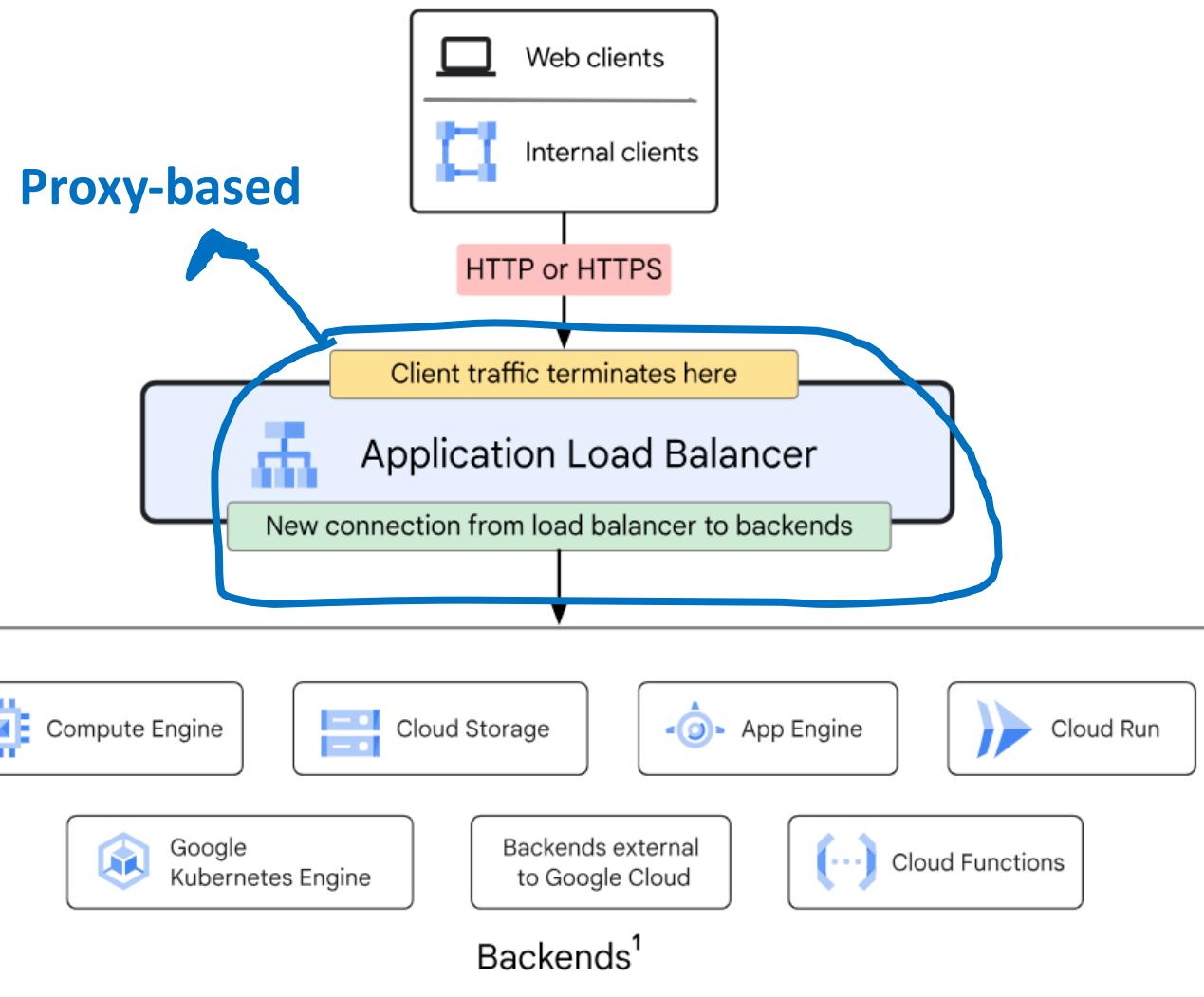
Cloud Load Balancing

Global External Application Load Balancer - HTTP



Demo-2

Cloud Load Balancing - Application Load Balancer (HTTP/S)

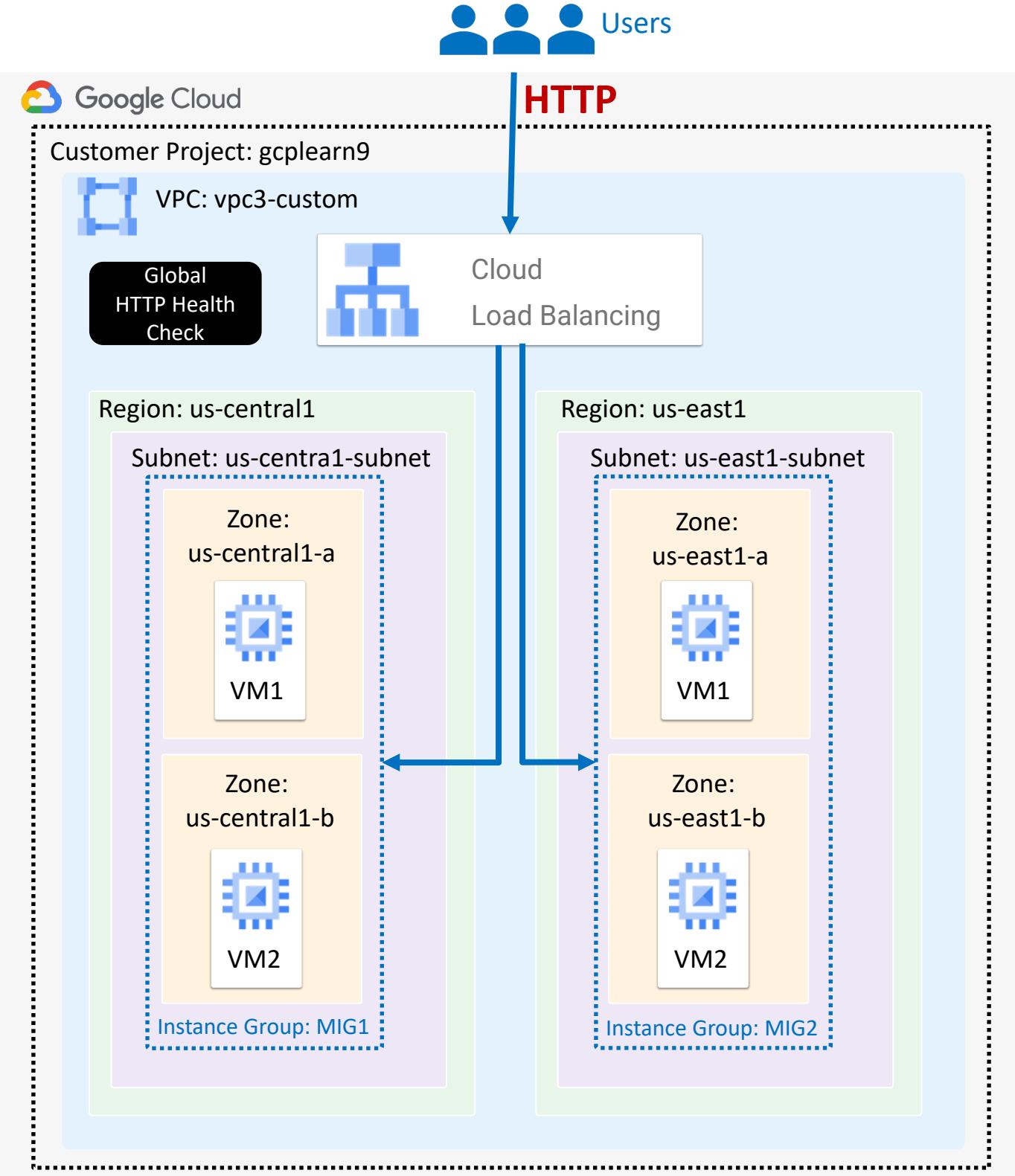


- **Application Load Balancer (HTTP/S)**
- Proxy-based Layer 7 load balancers
- **Proxy-based means**
 - Client traffic terminated on Load Balancer
 - New Connection created from load balancer to backends
- **Provides**
 - content-based routing
 - Application-aware health checks
- **External & Internal**
 - **Global**
 - support backends in multiple regions
 - **Regional**
 - support backends in a single region only
 - **Accessibility**
 - **External:** Accessible via internet
 - **Internal:** Accessible to systems in VPC or systems connected to VPC
- Ideal for web applications, APIs and microservices

Demo-1

Reference: <https://cloud.google.com/static/load-balancing/images/application-load-balancer.svg>

Cloud Load Balancing



Global
External
Application
Load Balancer
HTTP

Demo-2

Demo



Google Cloud Networking

Cloud Load Balancing

**Global External Application Load Balancer - HTTPS
(Self-Signed SSL)**



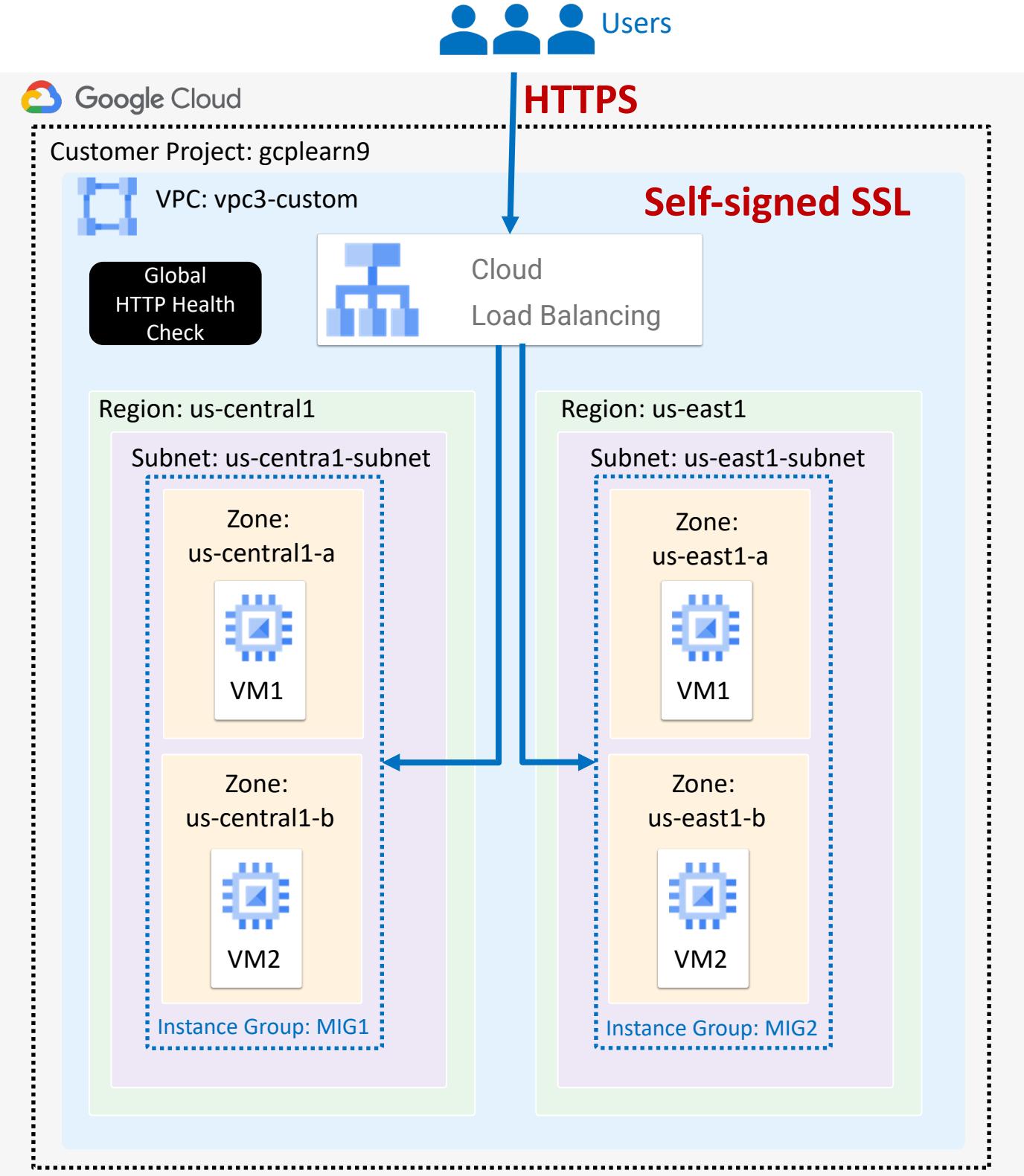
Demo-3

Cloud Load Balancing

Global
External
Application
Load Balancer
HTTPS

Self-signed SSL Certificates

Demo-3



Demo



Google Cloud Networking

Cloud Load Balancing

**Global External Application Load Balancer - HTTPS
(Google Managed SSL Certificates)**



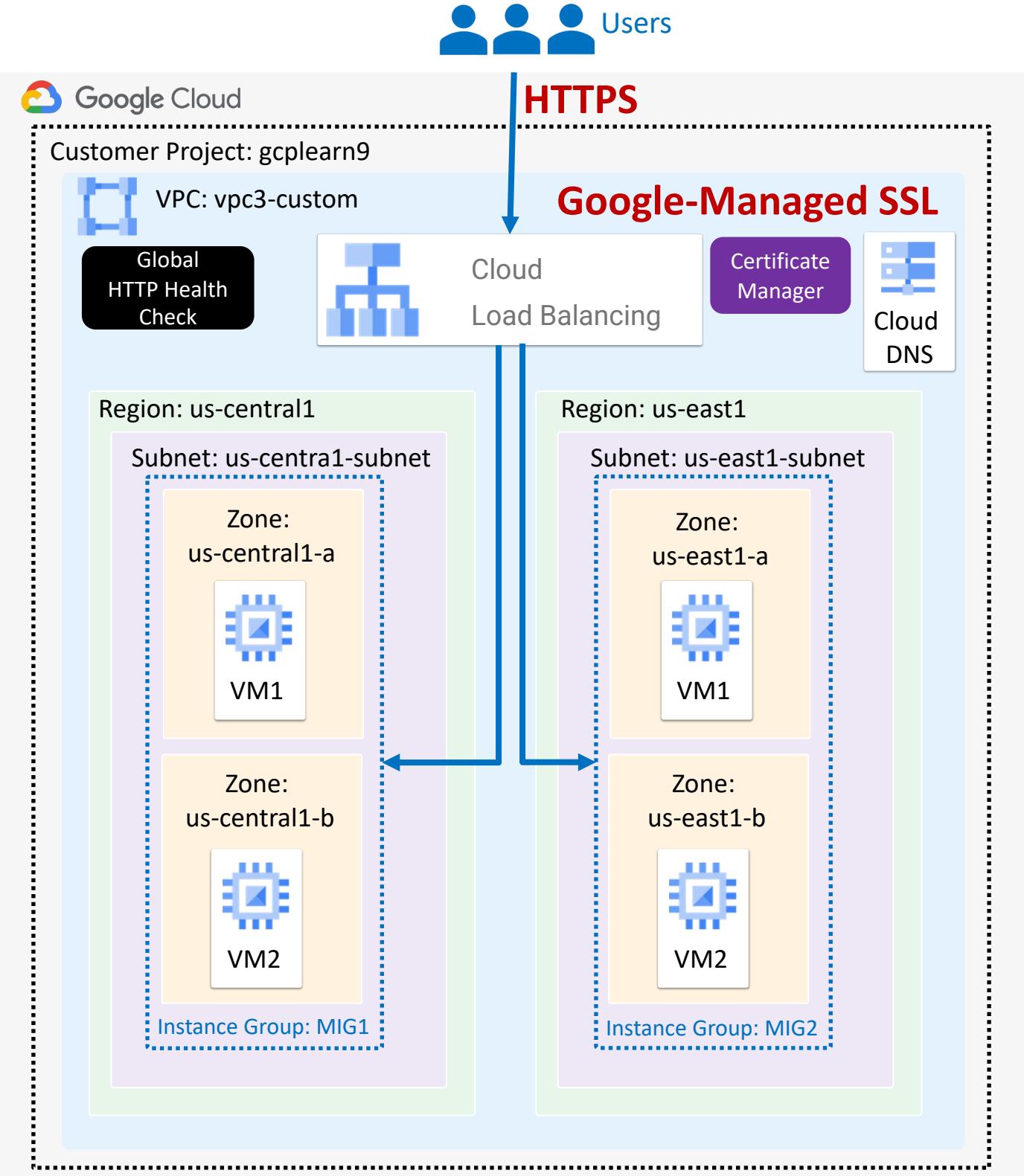
Demo-4

Cloud Load Balancing

Global
External
Application
Load Balancer
HTTPS

Google-managed SSL Certificates

Demo-4



**Demo**

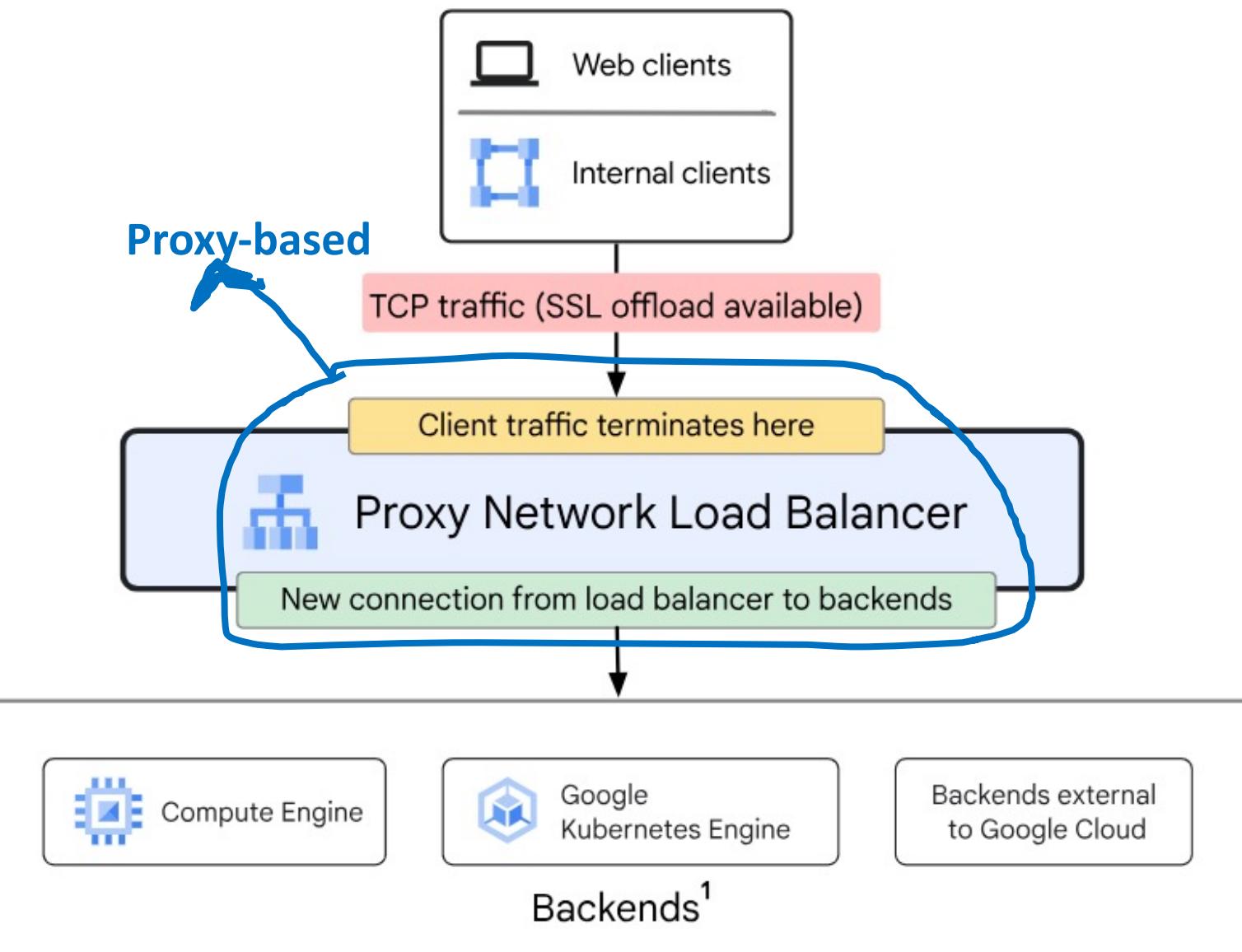
Google Cloud Networking

Cloud Load Balancing

Global External Network Load Balancer - TCP Proxy

**Demo-5**

Cloud Load Balancing - Network Load Balancer (TCP/TLS)

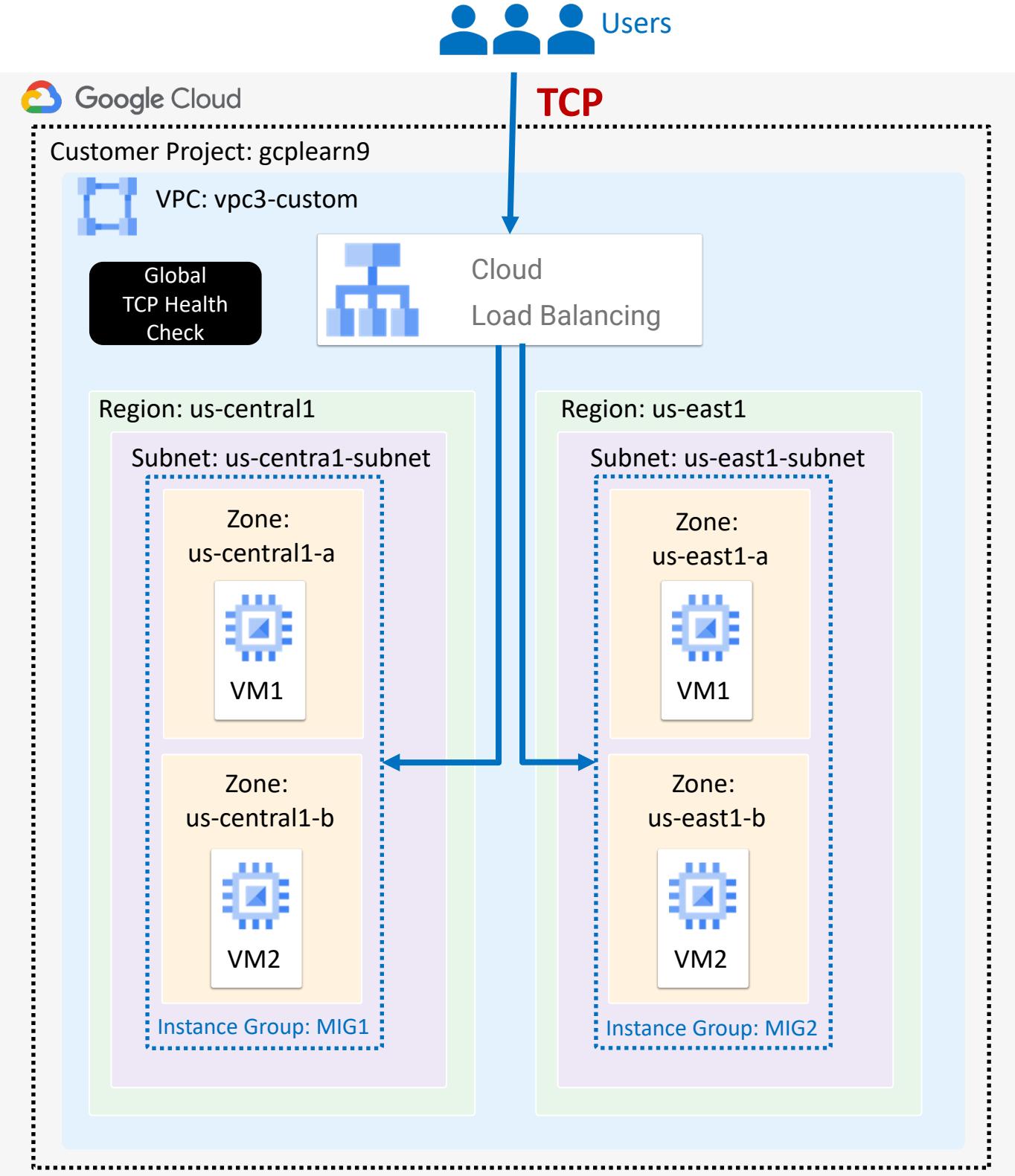


- **Network Load Balancer (TCP/SSL)**
- **Proxy-based Layer 4** load balancers
- **Protocols Supported**
 - TCP (TCP Proxy)
 - SSL (SSL Proxy)

Demo-5

Reference: <https://cloud.google.com/static/load-balancing/images/application-load-balancer.svg>

Cloud Load Balancing



Global
External
Network
Load Balancer
TCP Proxy

Demo-5

Demo



Google Cloud Networking

Cloud Load Balancing

Global External Network Load Balancer - SSL Proxy



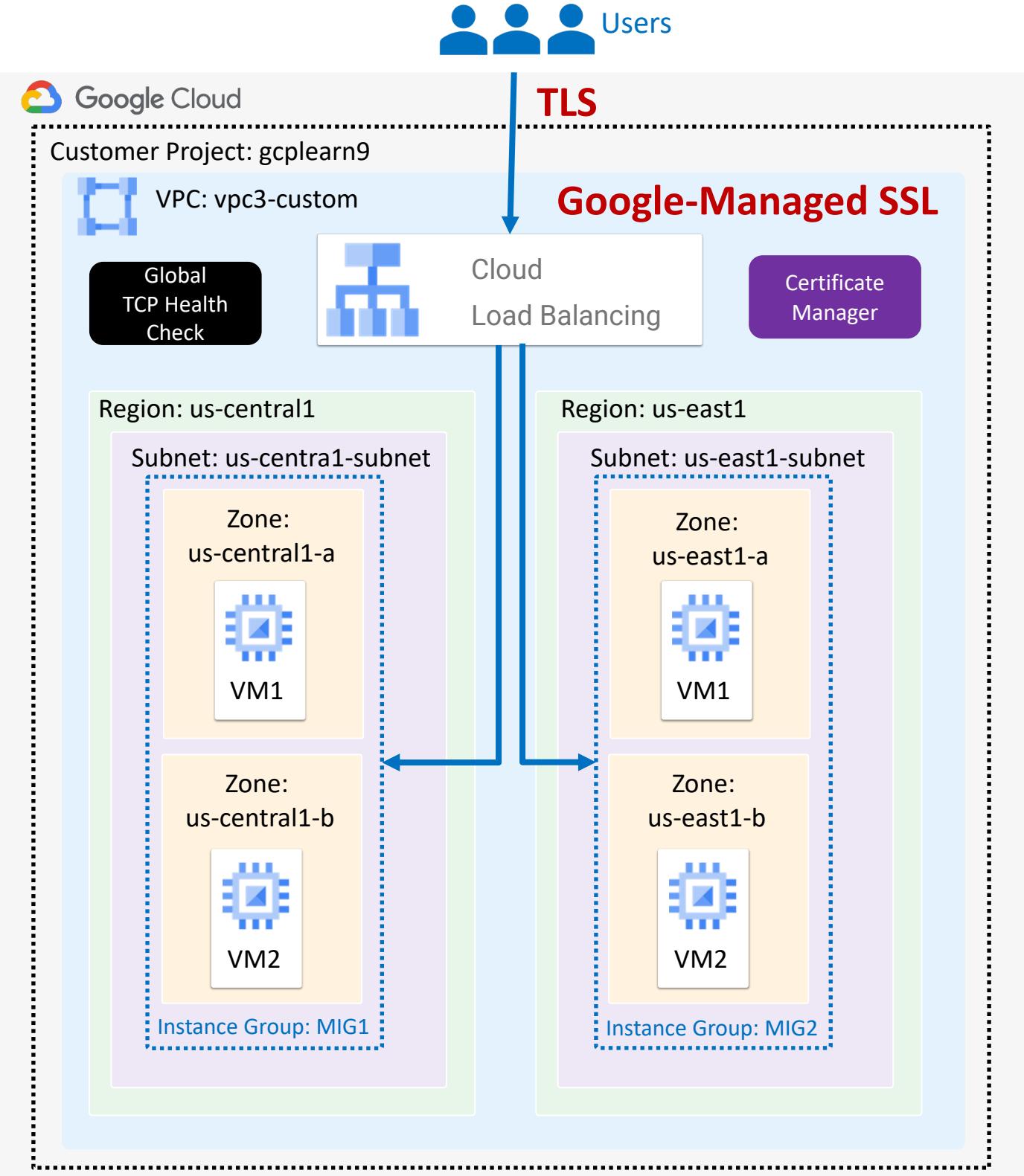
Demo-6

Cloud Load Balancing

Global
External
Network
Load Balancer
SSL Proxy

Google-managed SSL Certificates

Demo-6



**Demo**

Google Cloud Networking

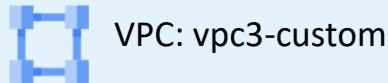
Cloud Load Balancing Zonal Managed Instance Groups

Demo-7

Cloud Load Balancing



Customer Project: gcplearn9

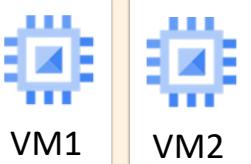


VPC: vpc3-custom

Region: us-central1

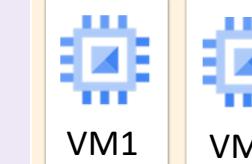
Subnet: us-central1-subnet

Zone:
us-central1-a



Instance Group: ZMIG1

Zone:
us-central1-b



Instance Group: ZMIG2

Regional
Health Check

Zonal
Managed
Instance
Groups

Demo-7

**Demo**

Google Cloud Networking

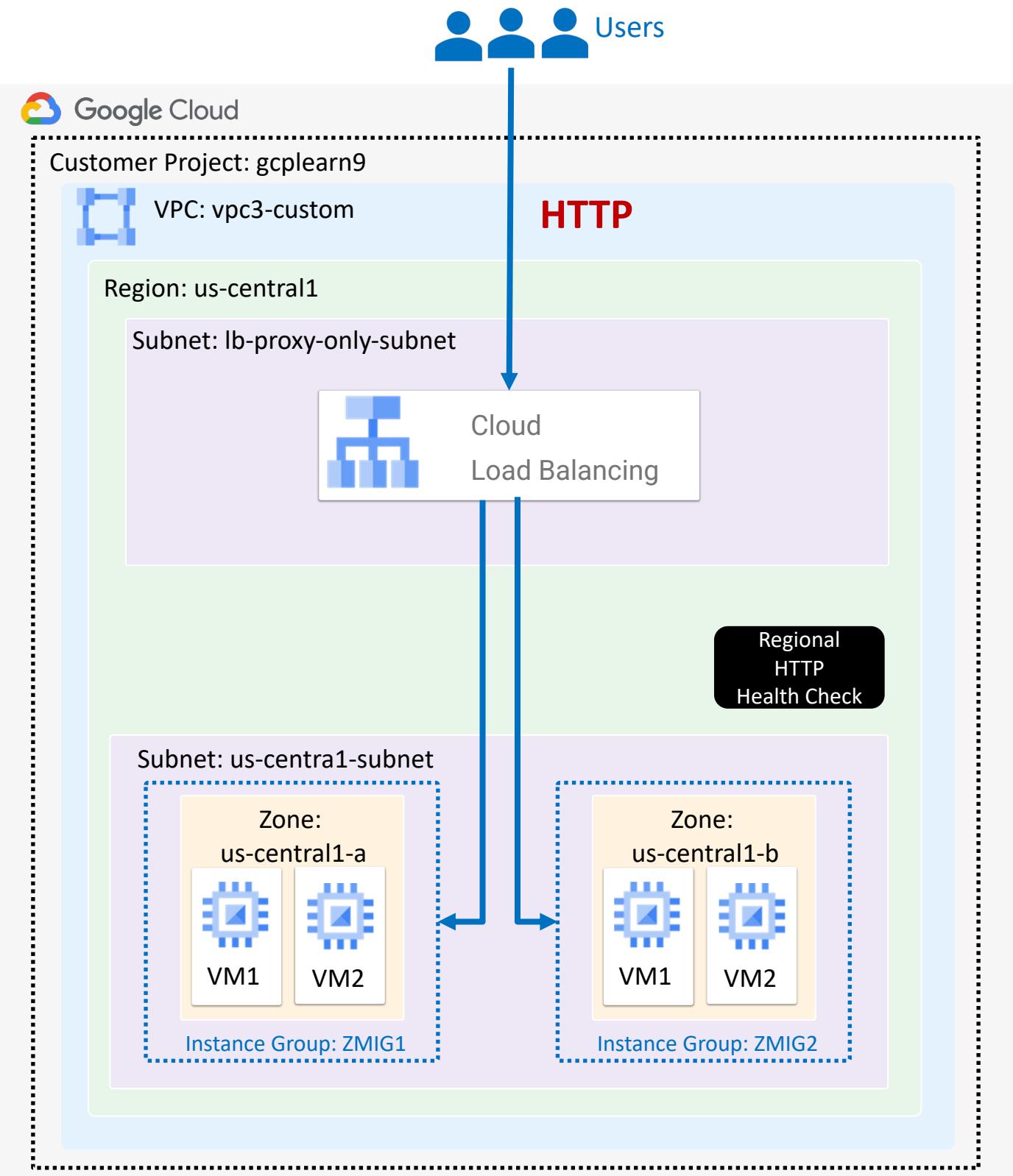


Cloud Load Balancing

Regional External Application Load Balancer HTTP

Demo-8

Cloud Load Balancing



Regional
External
Application
Load Balancer
HTTP

Demo-8

Demo



Google Cloud Networking

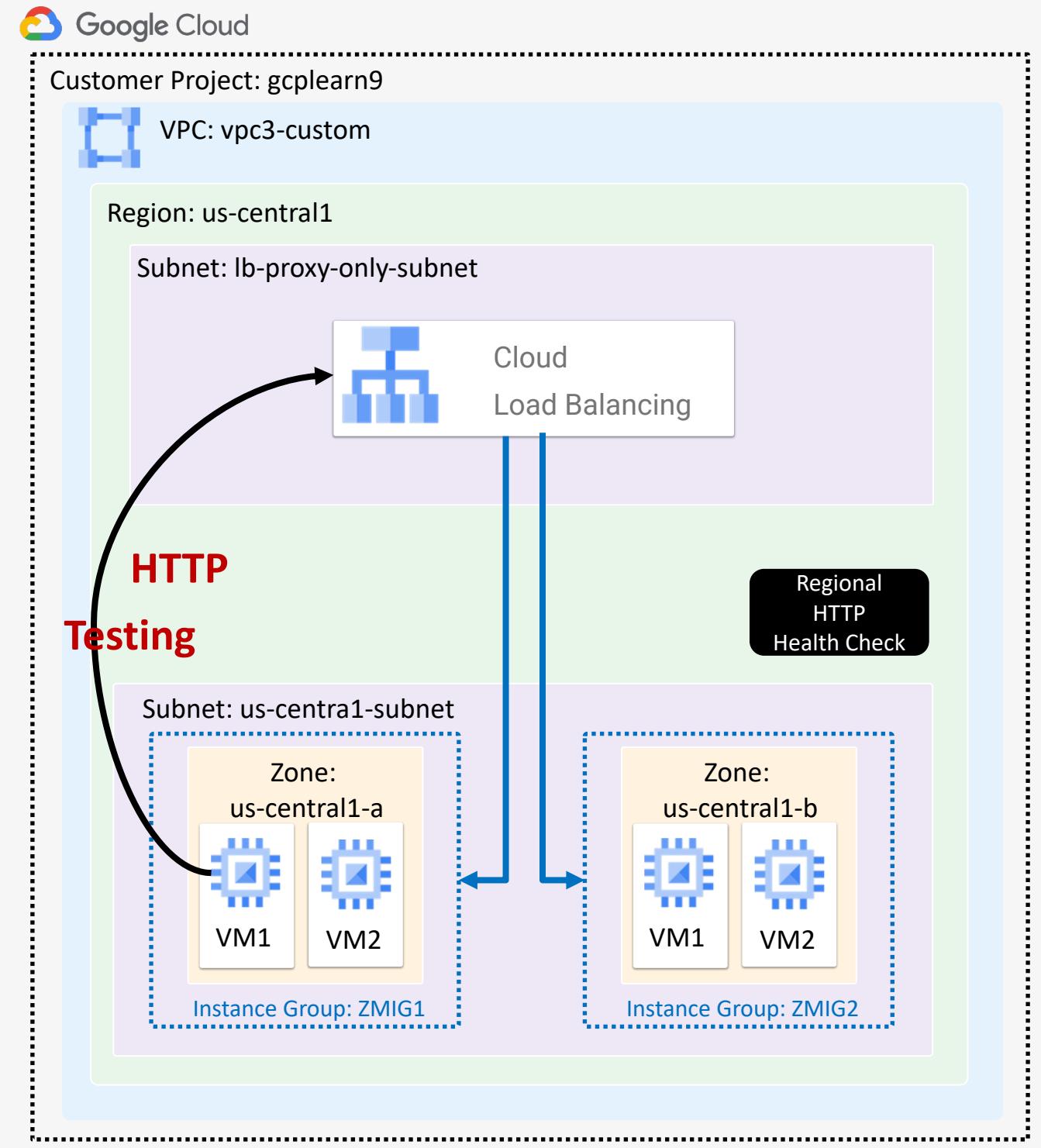


Cloud Load Balancing

Regional Internal Application Load
Balancer HTTP

Demo-9

Cloud Load Balancing



Regional
Internal
Application
Load Balancer
HTTP

Demo-9



Demo

Google Cloud Networking

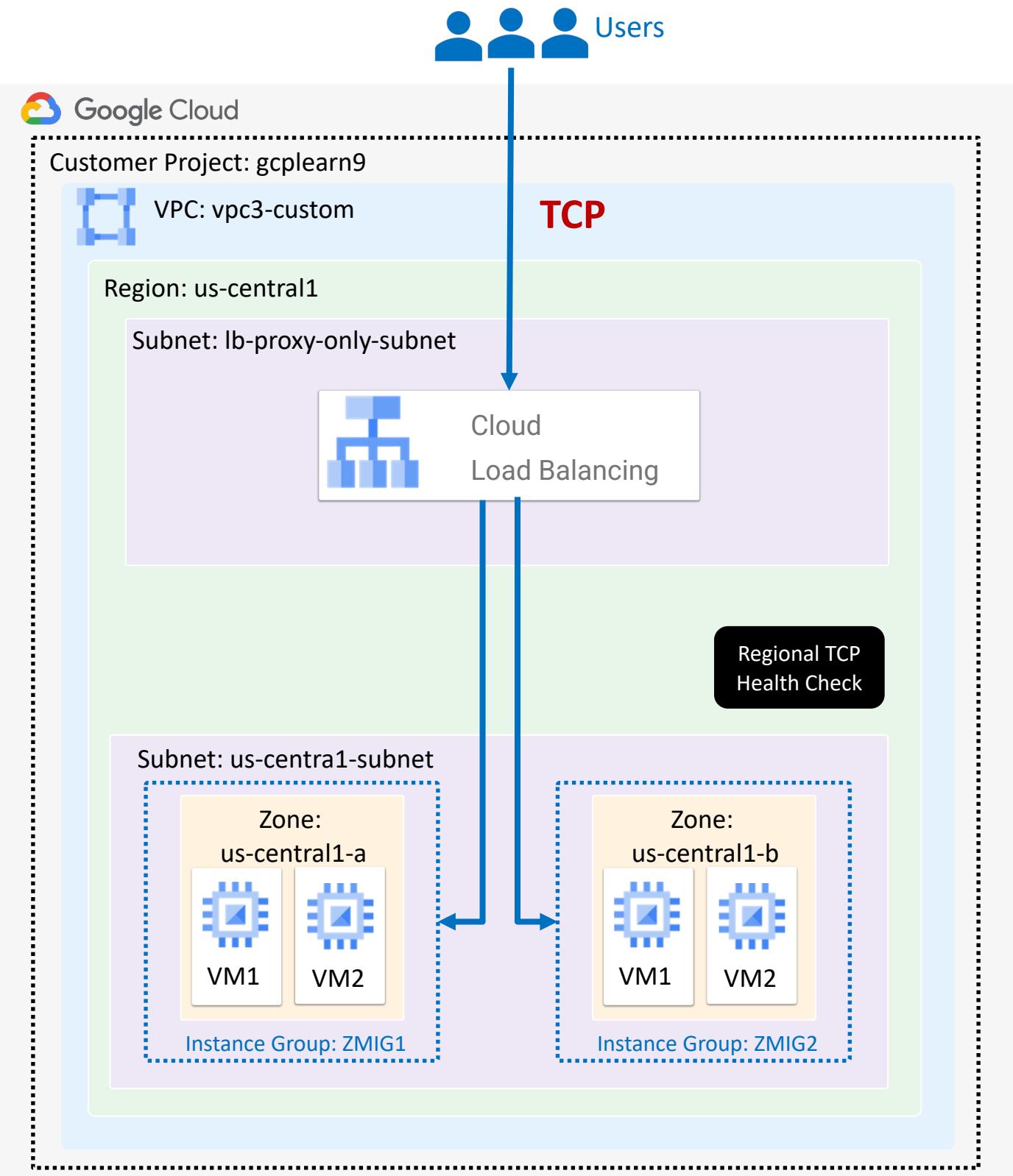


Cloud Load Balancing

Regional External Network Load Balancer - TCP Proxy

Demo-10

Cloud Load Balancing



Regional
External
Network
Load Balancer
TCP Proxy

Demo-10

**Demo**

Google Cloud Networking

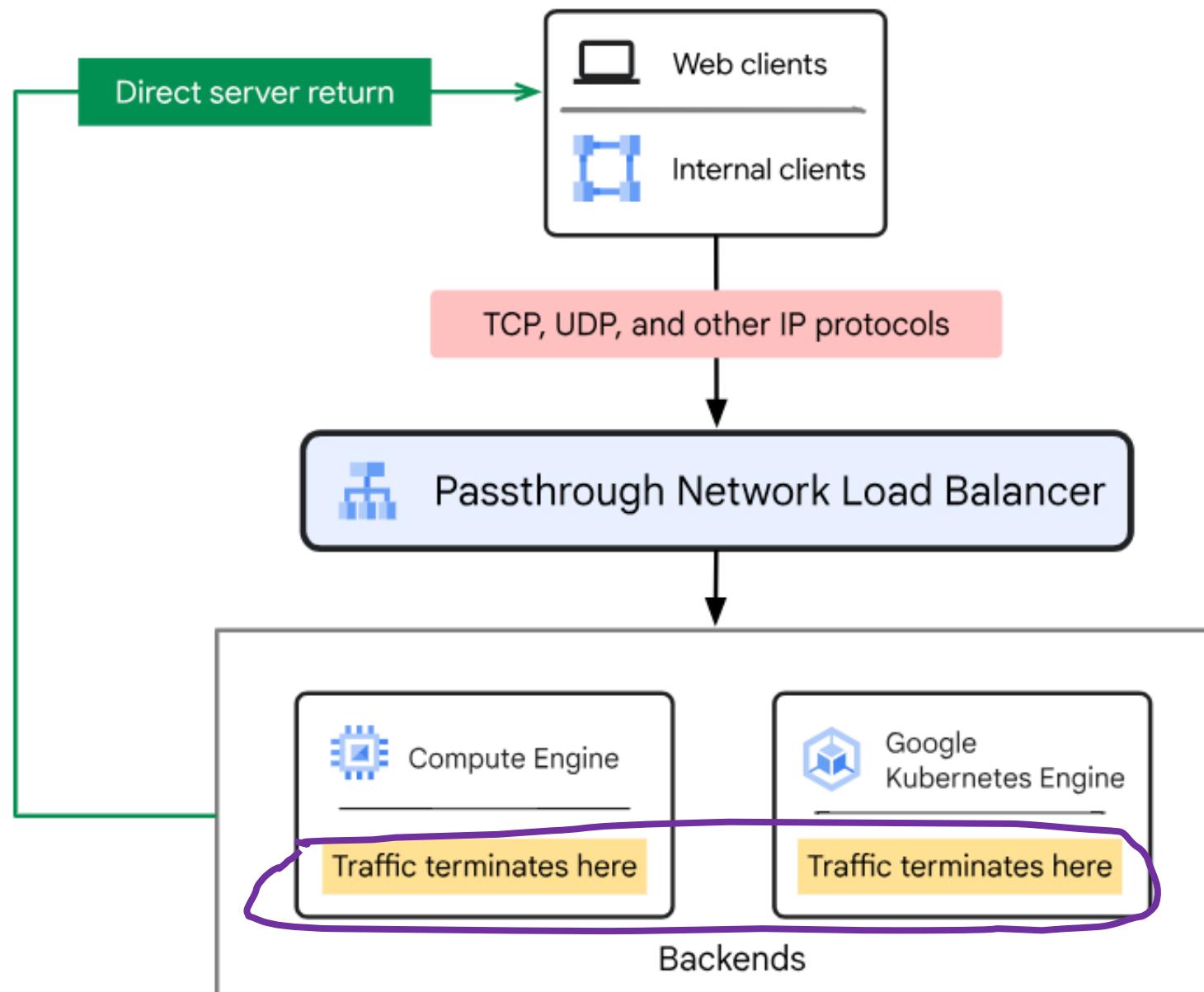


Cloud Load Balancing

Regional External Network Load Balancer - TCP Pass-through

Demo-11

Cloud Load Balancing - Network Load Balancer (TCP Pass-through)

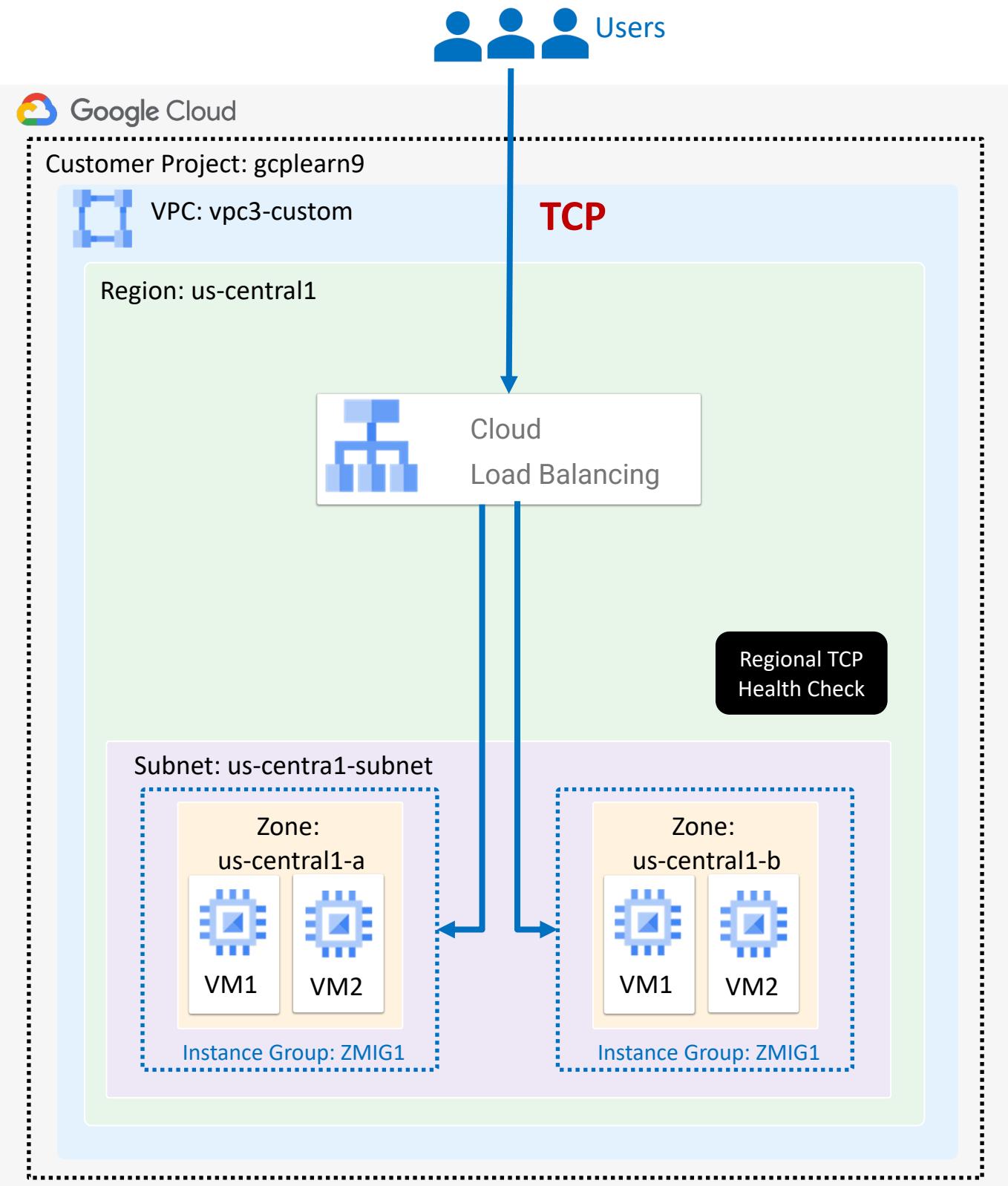


- **Network Load Balancer (TCP Pass-through)**
- **Pass-through** Layer 4 load balancers
- **Connection termination happens at backends**
- **External & Internal**
 - **Regional Only (No Global)**
 - support backends in a **single region only**
 - **Accessibility**
 - **External:** Accessible via **internet**
 - **Internal:** Accessible to **systems in VPC or systems connected to VPC**

Pass-through – Connection termination happens at backend

Demo-11

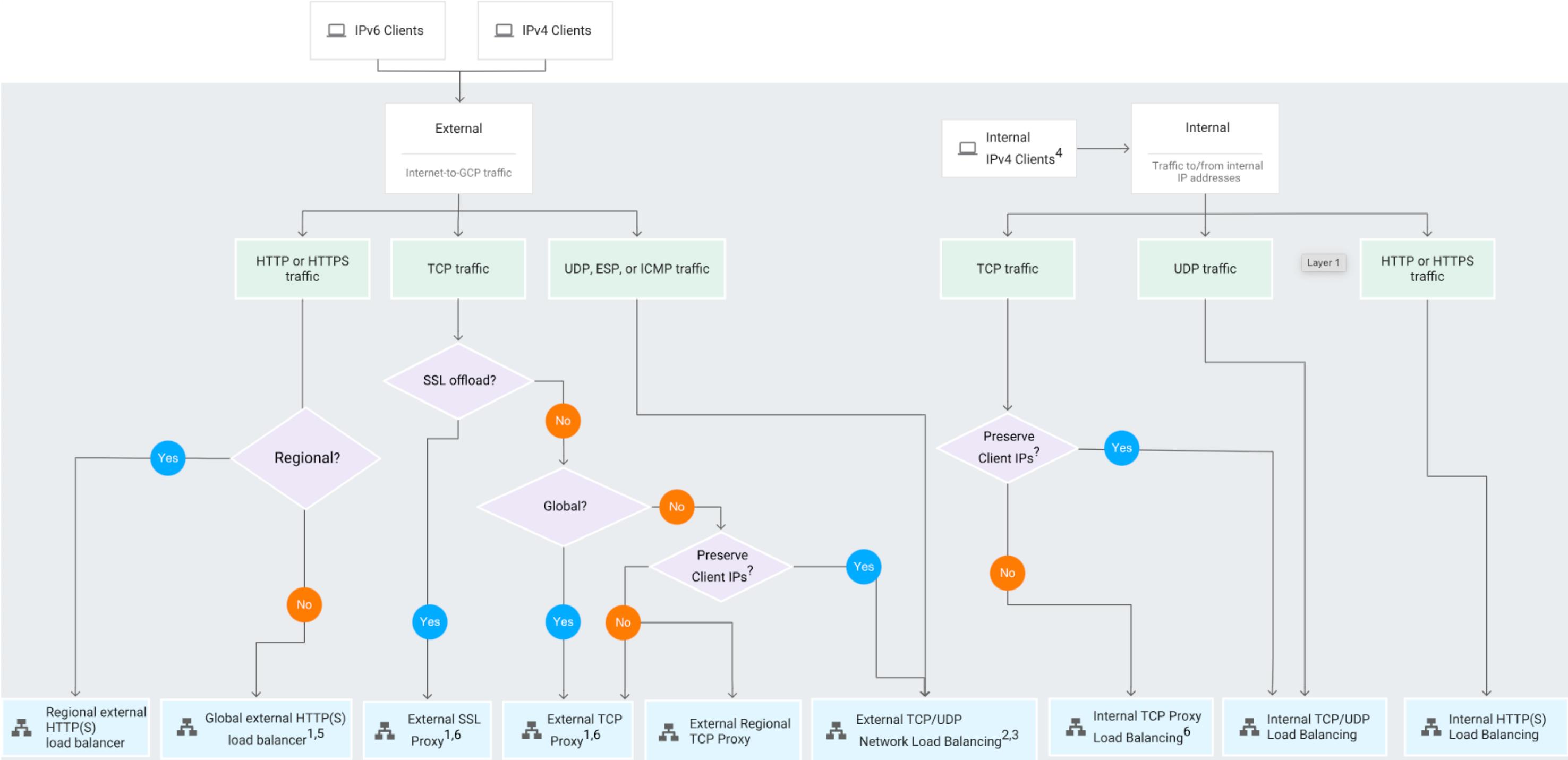
Cloud Load Balancing



Regional
External
Network
Load Balancer
TCP Pass-through

Demo-11

Google Cloud Load Balancing



Reference: <https://cloud.google.com/load-balancing/images/choose-lb.svg>

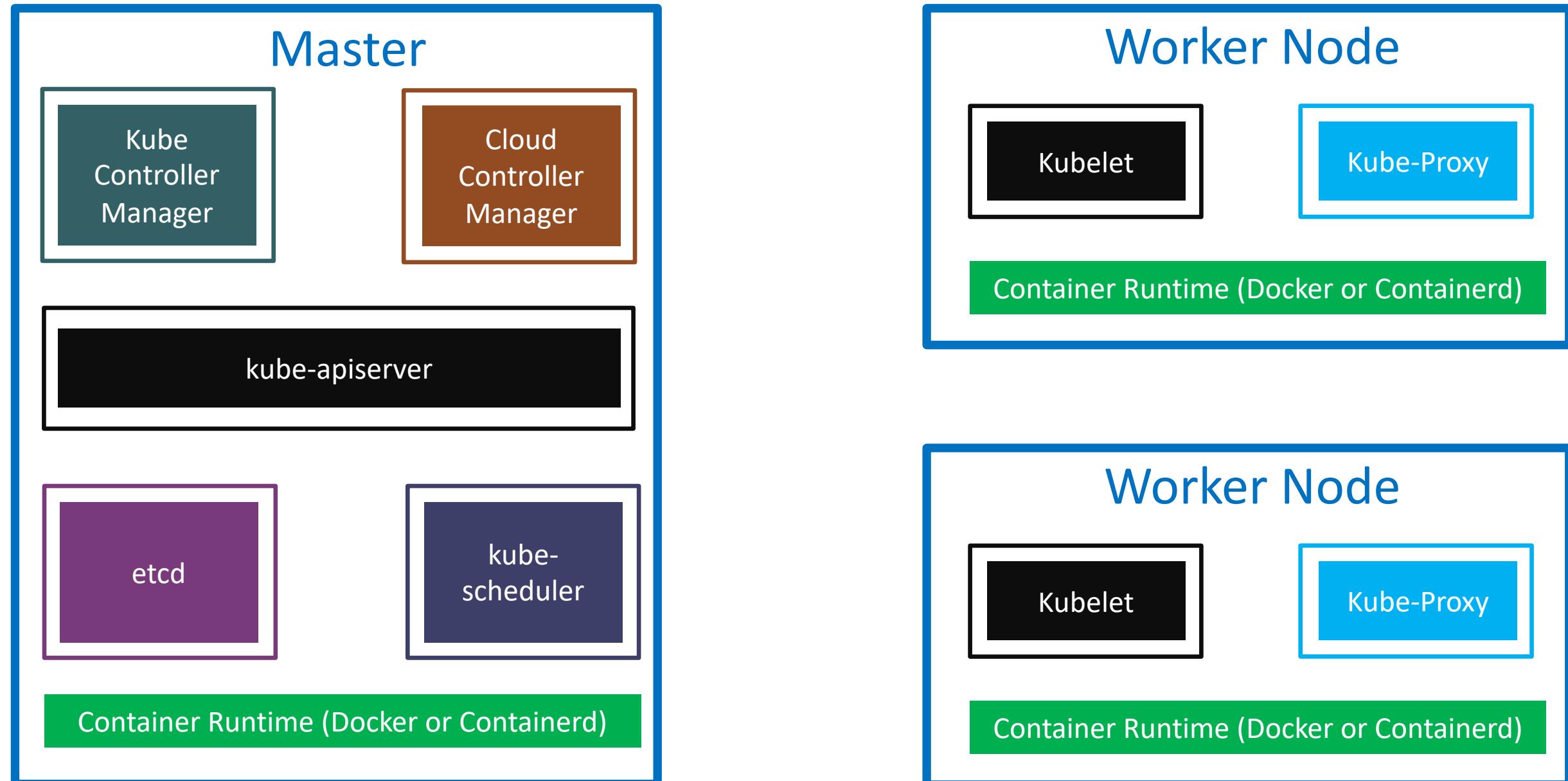


Google Kubernetes Engine

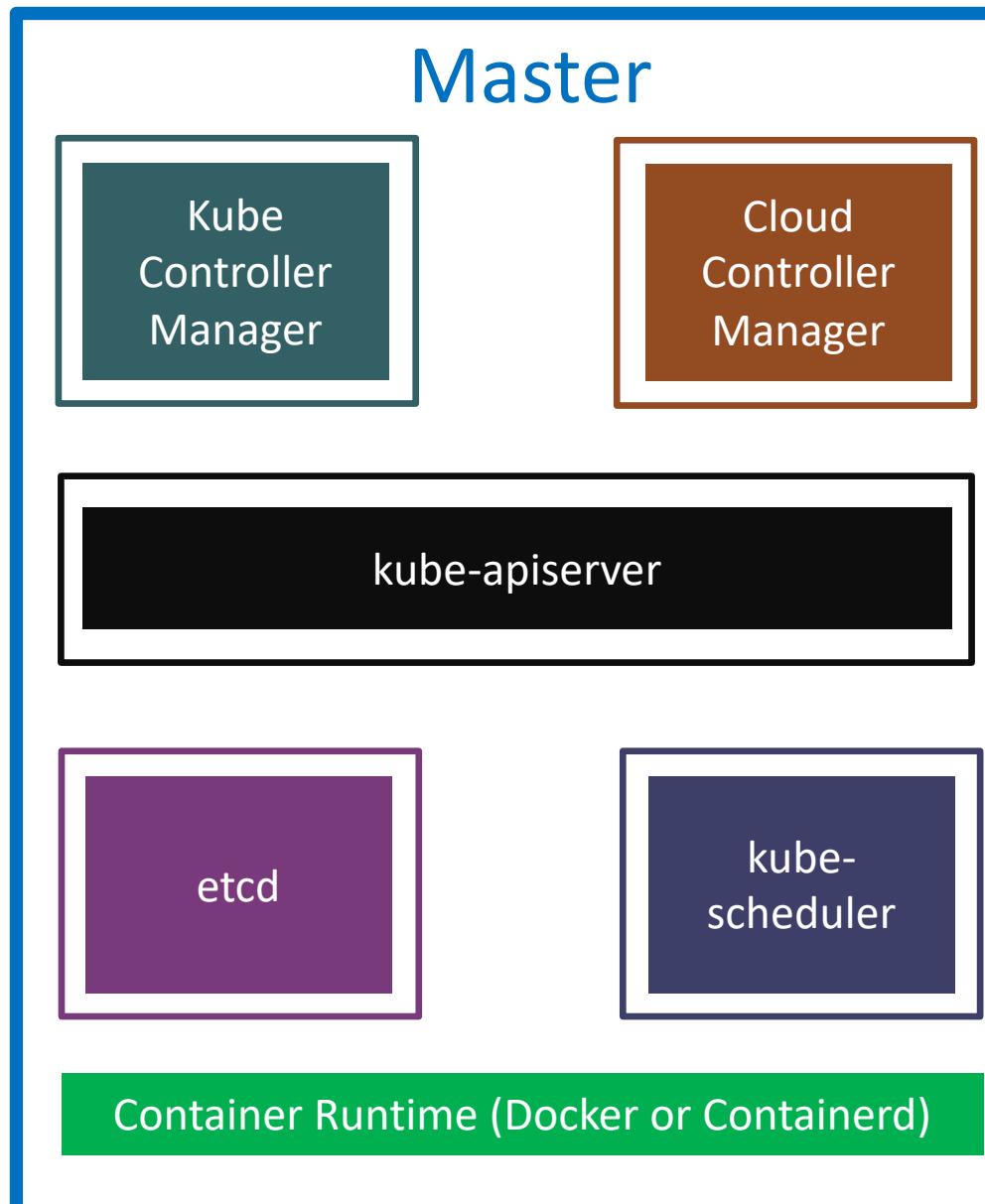
Kubernetes Architecture Quick Overview



Kubernetes - Architecture



Kubernetes Architecture - Master



- **kube-apiserver**

- It acts as **front end** for the Kubernetes control plane. It **exposes** the Kubernetes API
- Command line tools (like kubectl), Users and even Master components (scheduler, controller manager, etcd) and Worker node components like (Kubelet) **everything talk** with API Server.

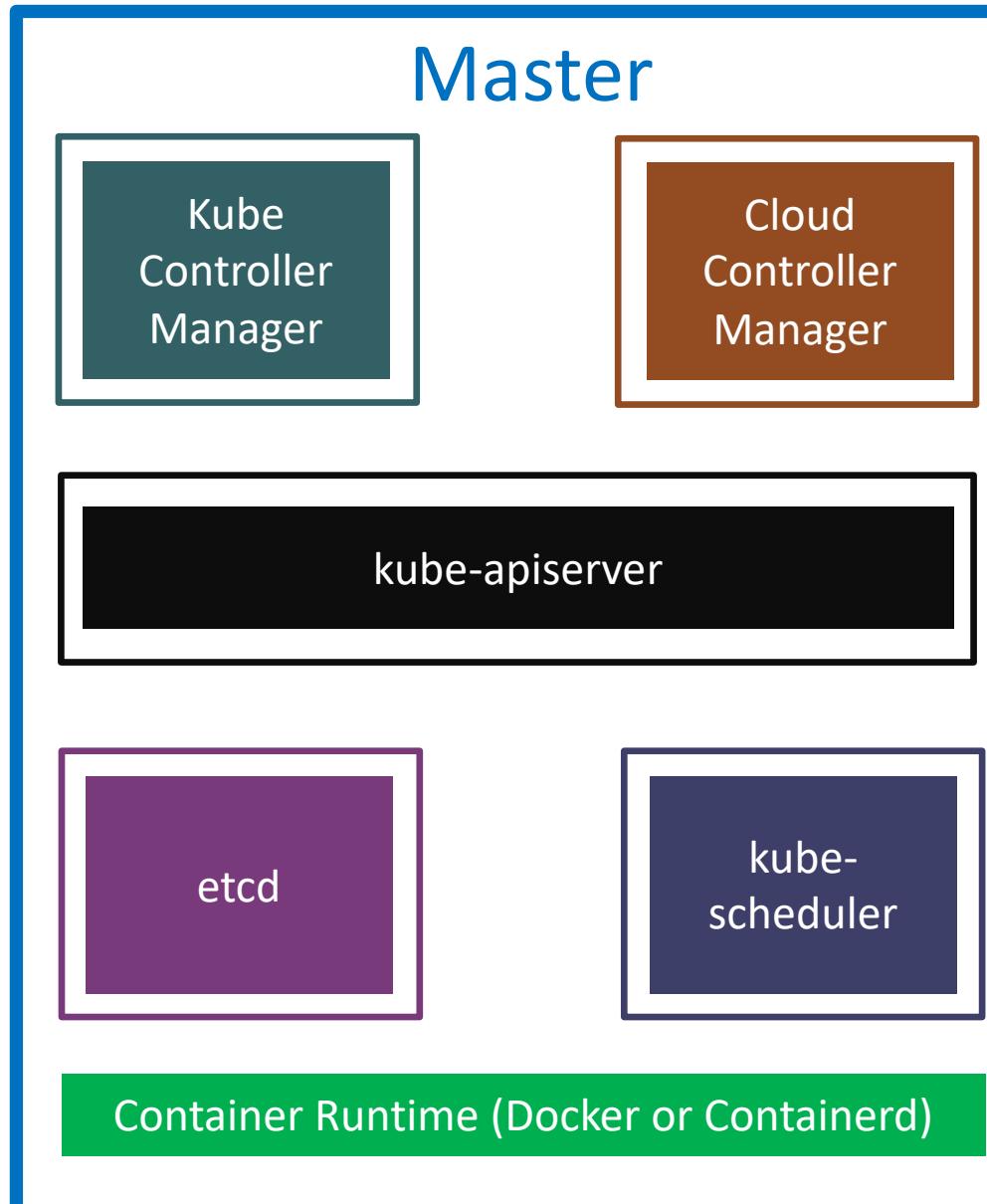
- **etcd**

- Consistent and highly-available **key value store** used as Kubernetes' **backing store** for all cluster data.
- It **stores** all the masters and worker node information.

- **kube-scheduler**

- Scheduler is responsible for **distributing** containers across multiple nodes.
- It **watches** for newly created Pods with no assigned node, and selects a node for them to run on.

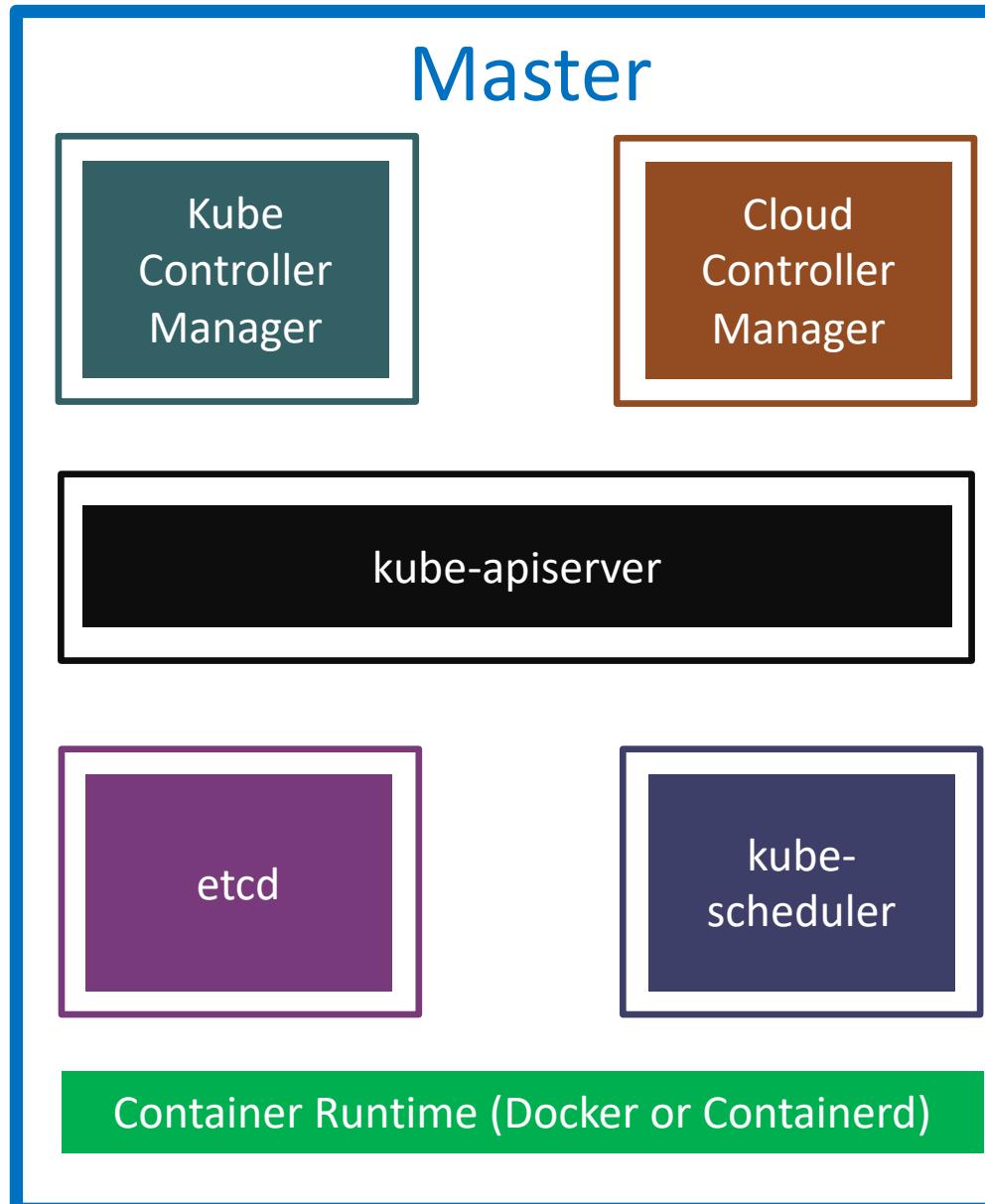
Kubernetes Architecture - Master



- **kube-controller-manager**

- Controllers are responsible for noticing and responding when nodes, containers or endpoints **go down**. They make decisions to bring up new containers in such cases.
- **Node Controller**: Responsible for noticing and responding when **nodes go down**.
- **Replication Controller**: Responsible for maintaining the **correct number of pods** for every replication controller object in the system.
- **Endpoints Controller**: **Populates** the Endpoints object (that is, joins Services & Pods)
- **Service Account & Token Controller**: Creates default accounts and API Access for **new namespaces**.

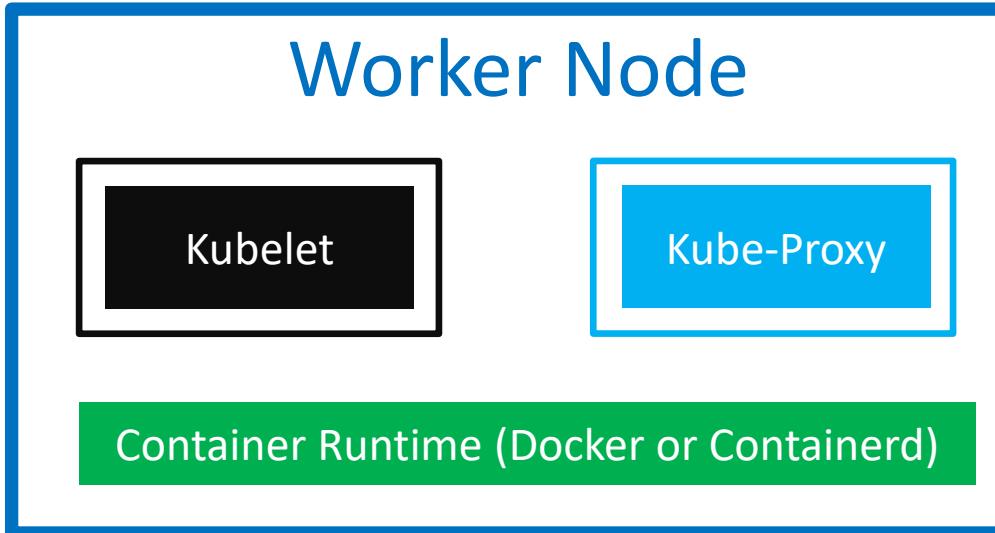
Kubernetes Architecture - Master



- **cloud-controller-manager**

- A Kubernetes control plane component that embeds **cloud-specific control logic**.
- It only runs controllers that are **specific** to your cloud provider.
- **On-Premise** Kubernetes clusters will not have this component.
- **Node controller:** For **checking** the cloud provider to determine if a node has been deleted in the cloud after it stops responding
- **Route controller:** For setting up **routes** in the underlying cloud infrastructure
- **Service controller:** For creating, updating and deleting cloud provider **load balancer**
- **Many more** controllers might be present and will differ from cloud to cloud based on that respective cloud Kubernetes Platform design and Integrations to their Cloud products

Kubernetes Architecture – Worker Nodes



- **Container Runtime**

- Container Runtime is the **underlying software** where we run all these Kubernetes components.
- In GKE, default is **containerd** but we also have options like **Ubuntu with Containerd**, **Ubuntu with Docker** and **Windows**

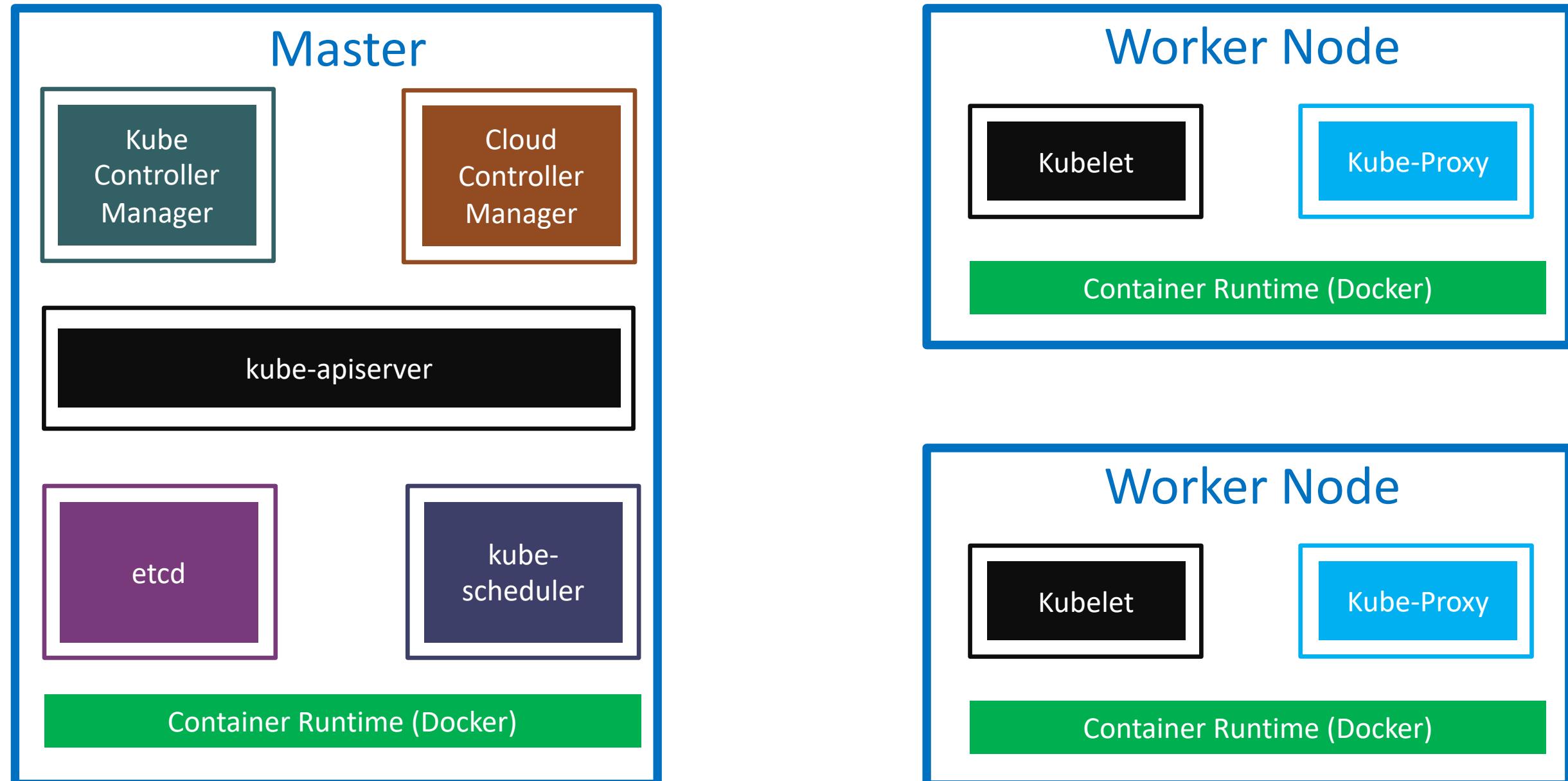
- **Kubelet**

- Kubelet is the **agent** that runs on every node in the cluster
- This agent is **responsible** for making sure that containers are running in a Pod on a node.

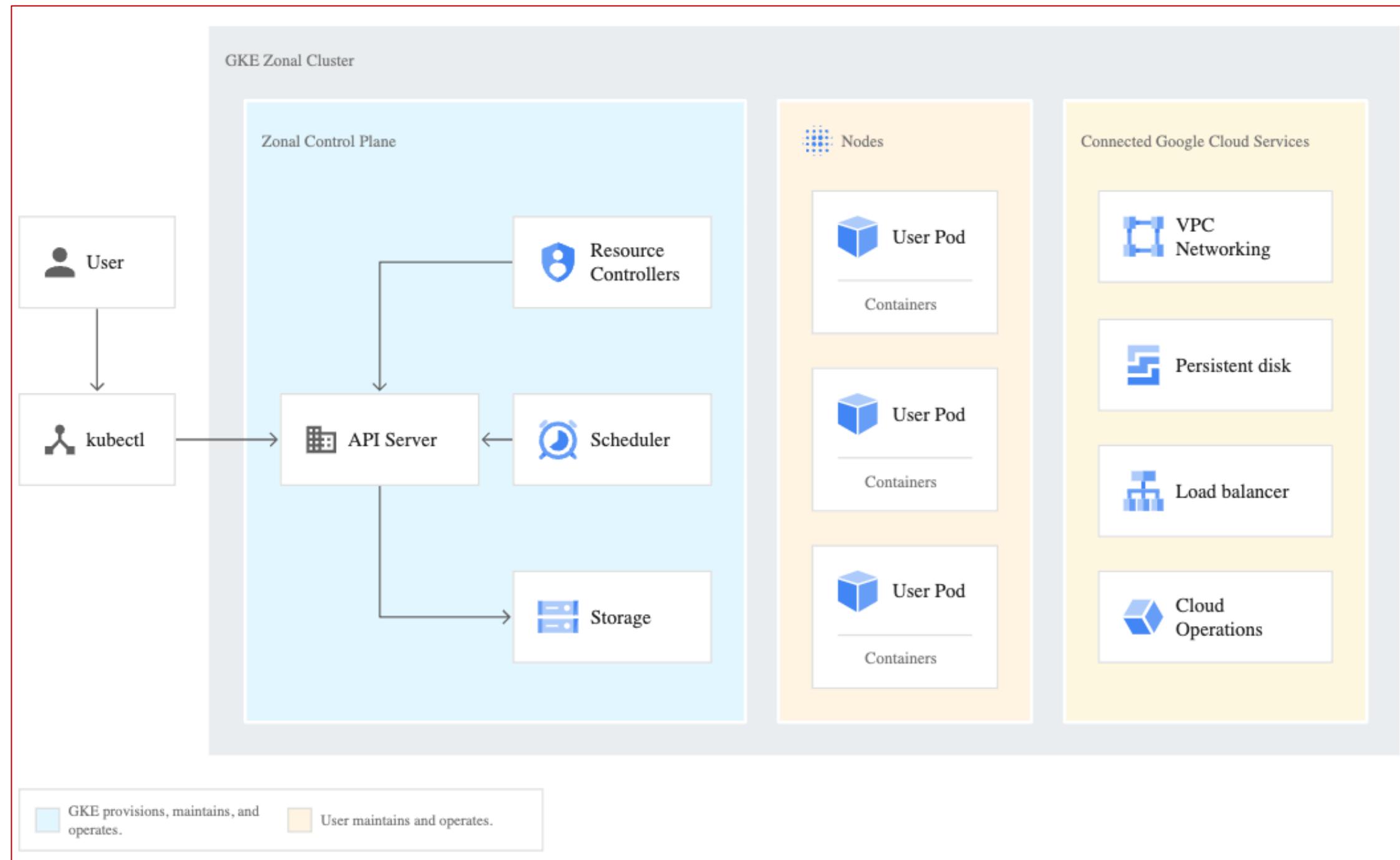
- **Kube-Proxy**

- It is a **network proxy** that runs on each node in your cluster.
- It maintains **network rules** on nodes
- In short, these network rules **allow** network communication to your Pods from network sessions inside or outside of your cluster.

Kubernetes - Architecture



GKE Standard Cluster Architecture



Reference: <https://cloud.google.com/kubernetes-engine/docs/concepts/cluster-architecture>



Google Kubernetes Engine GKE Standard Regional Cluster



GKE Cluster Modes & Types

GKE Cluster Modes

GKE Standard

GKE Autopilot



GKE Cluster Types

GKE Zonal Cluster

GKE Public Cluster

GKE Alpha Cluster

GKE Regional Cluster

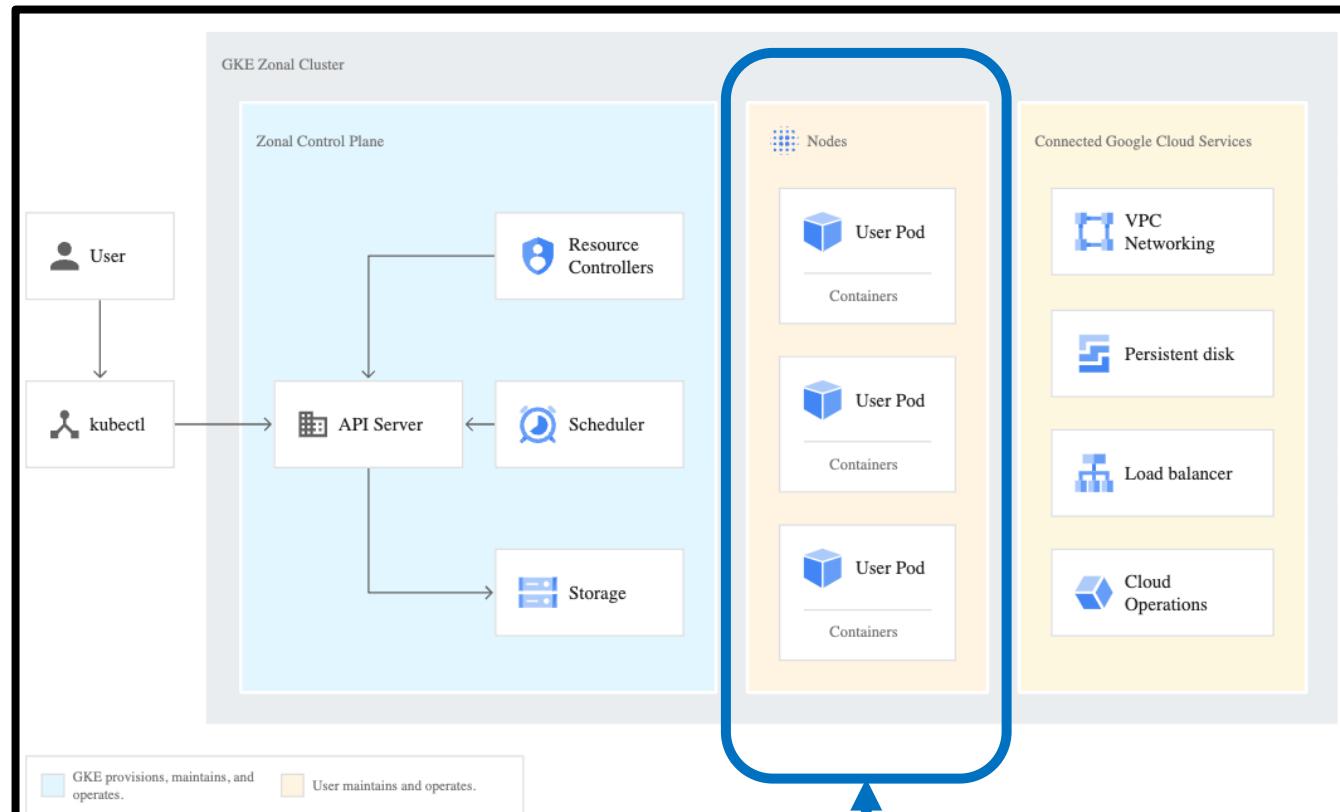
GKE Private Cluster

GKE Cluster using
Windows Node
Pools

GKE Standard vs Autopilot Cluster Architecture

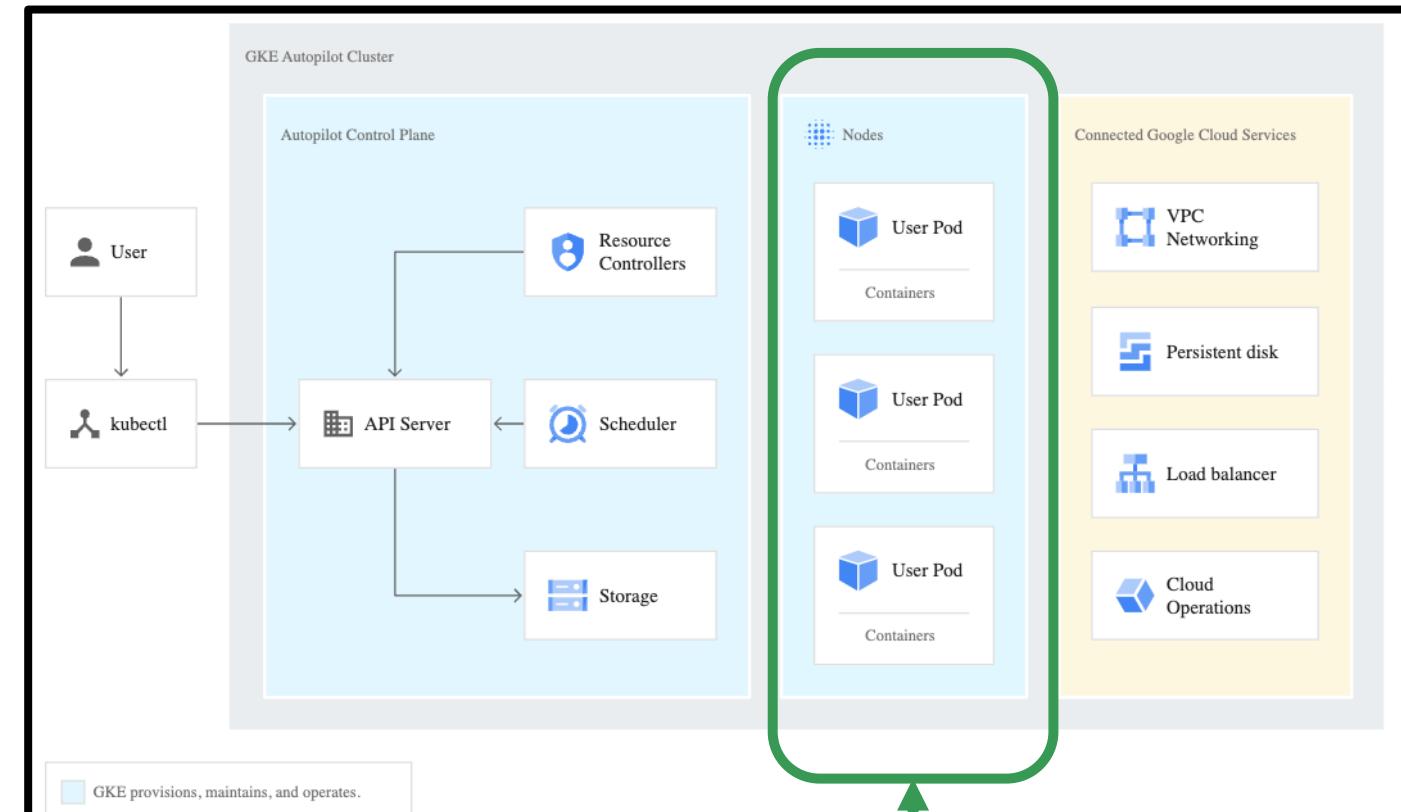


GKE Standard Cluster



User maintains and operates

GKE Autopilot Cluster



GKE Provisions, maintains and operates

Reference: <https://cloud.google.com/kubernetes-engine/docs/concepts/cluster-architecture>

Reference: <https://cloud.google.com/kubernetes-engine/docs/concepts/autopilot-architecture>

Google Cloud: GKE Standard vs Autopilot



Features	GKE Standard Cluster	GKE AutoPilot Cluster
GKE Control plane	Managed by GKE	Managed by GKE
Node and Node Pools	Created, configured and managed by you	Managed by GKE
Provisioning Resources	You manually provision additional resources and set overall cluster size	GKE dynamically provisions resources based on your Pod specification
Billing / Pricing	Pay per node (CPU, memory, boot disk)	<ol style="list-style-type: none">1. Pay per Pod2. Requests (CPU, memory, and ephemeral storage)
Location Availability	Regional or Zonal	Regional
For Complete Comparison	https://cloud.google.com/kubernetes-engine/docs/resources/autopilot-standard-feature-comparison	

Kubernetes

Fundamentals



pod



ReplicaSet



Deployment



Service



Kubernetes Fundamentals

A POD is a **single instance** of an Application.



A POD is the **smallest object**, that you can create in Kubernetes.

A ReplicaSet will maintain a **stable set** of replica Pods running at any given time.



ReplicaSet

In short, it is often used to **guarantee the availability** of a specified number of identical Pods

A Deployment runs **multiple replicas** of your application and **automatically replaces** any instances that **fail** or become **unresponsive**.



Deployment

Deployments are well-suited for **stateless** applications.

A service is an abstraction for pods, providing a stable, so-called **virtual IP (VIP)** address.



Service

In simple terms, service sits **Infront** of a POD and **acts** as a load balancer.

Kubernetes - Imperative & Declarative

Kubernetes Fundamentals

Imperative

Declarative

kubectl

Pod

ReplicaSet

Deployment

Service

YAML & kubectl

Pod

ReplicaSet

Deployment

Service



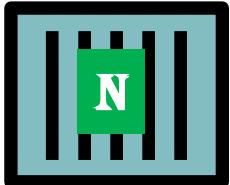
Imperative

Kubernetes Pod

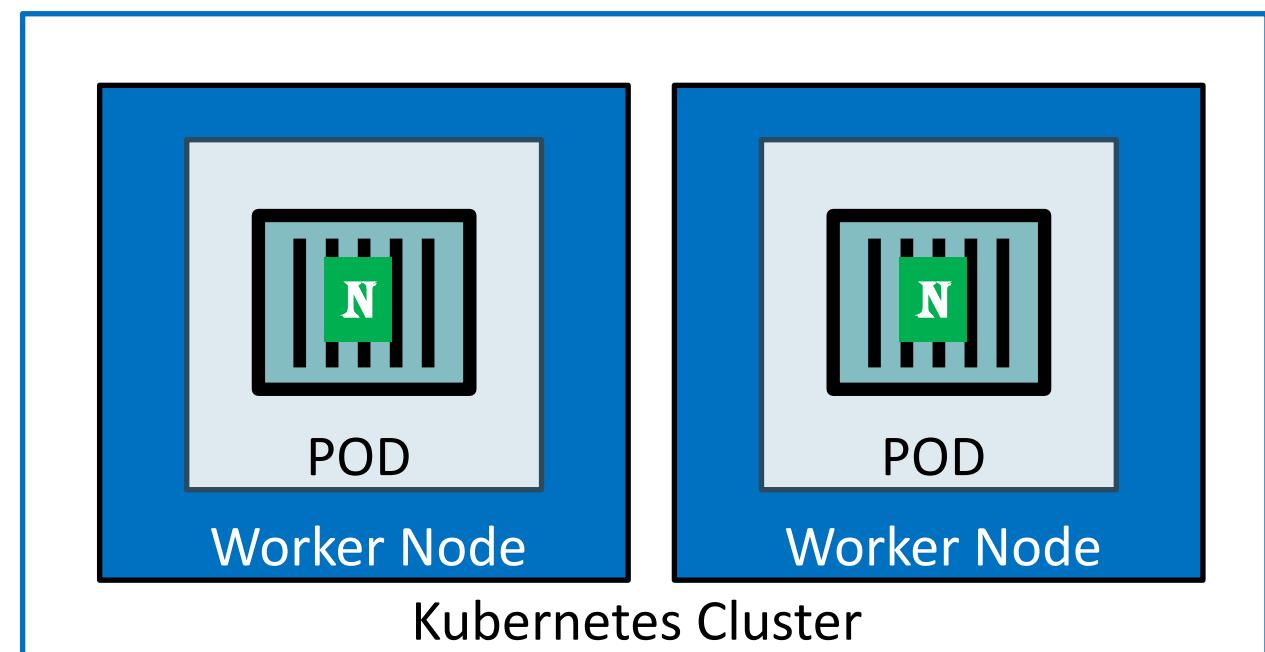


Kubernetes - POD

- With Kubernetes our core goal will be to deploy our applications in the form of **containers** on **worker nodes** in a k8s cluster.
- Kubernetes **does not** deploy containers directly on the worker nodes.
- Container is **encapsulated** in to a Kubernetes Object named **POD**.
- A POD is a **single instance** of an application.
- A POD is the **smallest object** that we can create in Kubernetes.

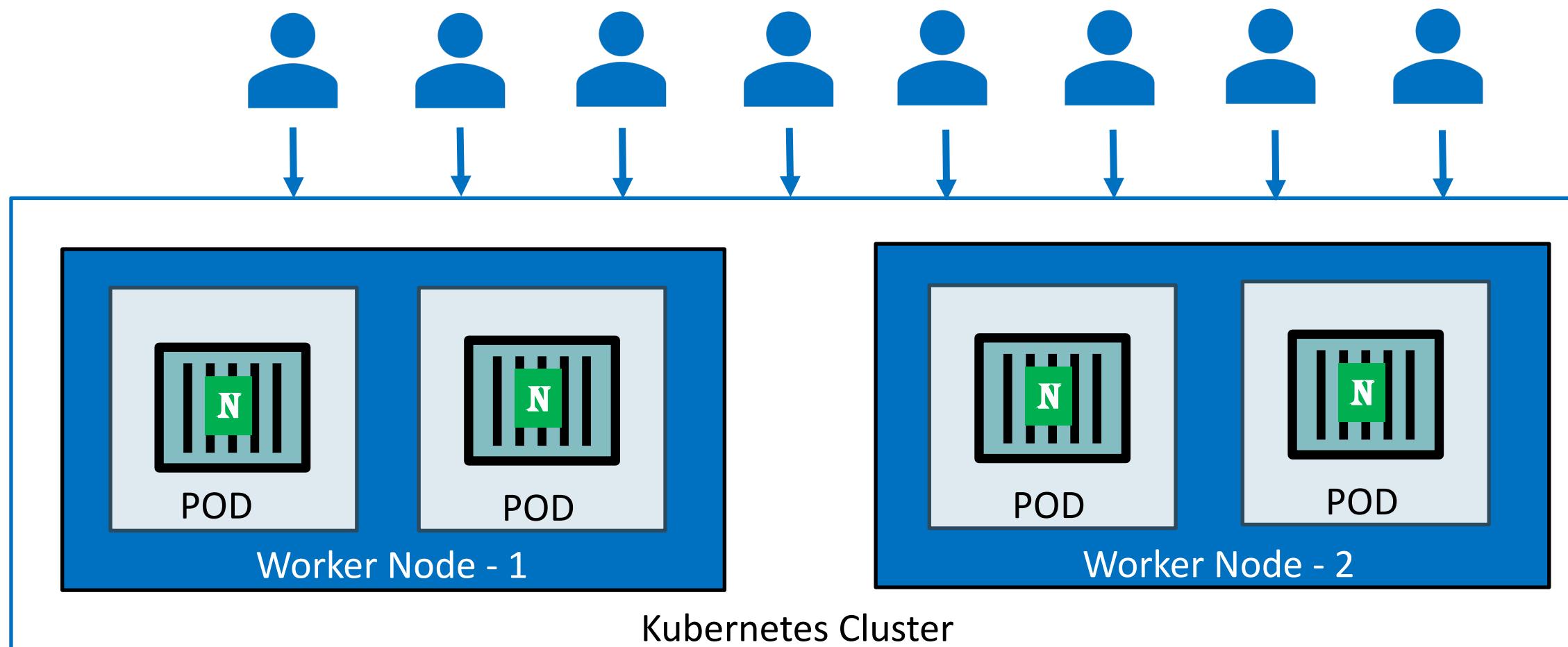


Nginx Container Image



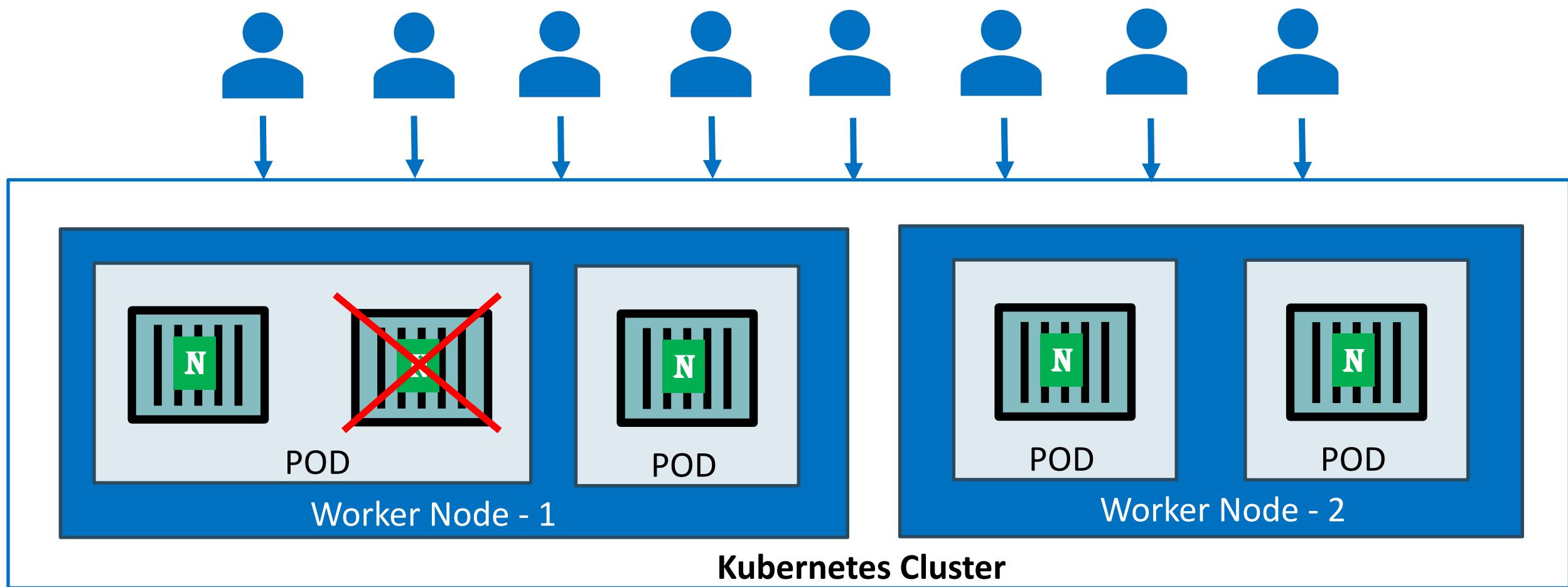
Kubernetes - POD

- PODs generally have **one to one** relationship with containers.
- To scale up we **create** new POD and to scale down we **delete** the POD.



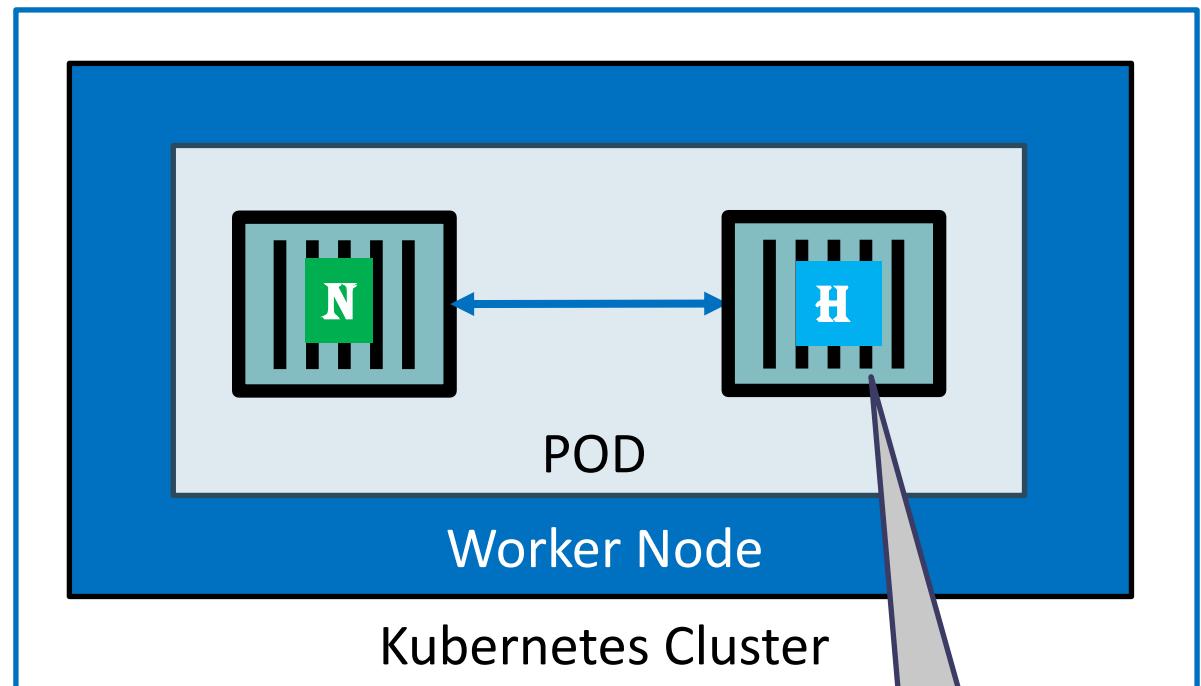
Kubernetes – PODs

- We cannot have multiple containers of **same kind** in a single POD.
- Example: Two NGINX containers in single POD serving same purpose is **not recommended**.

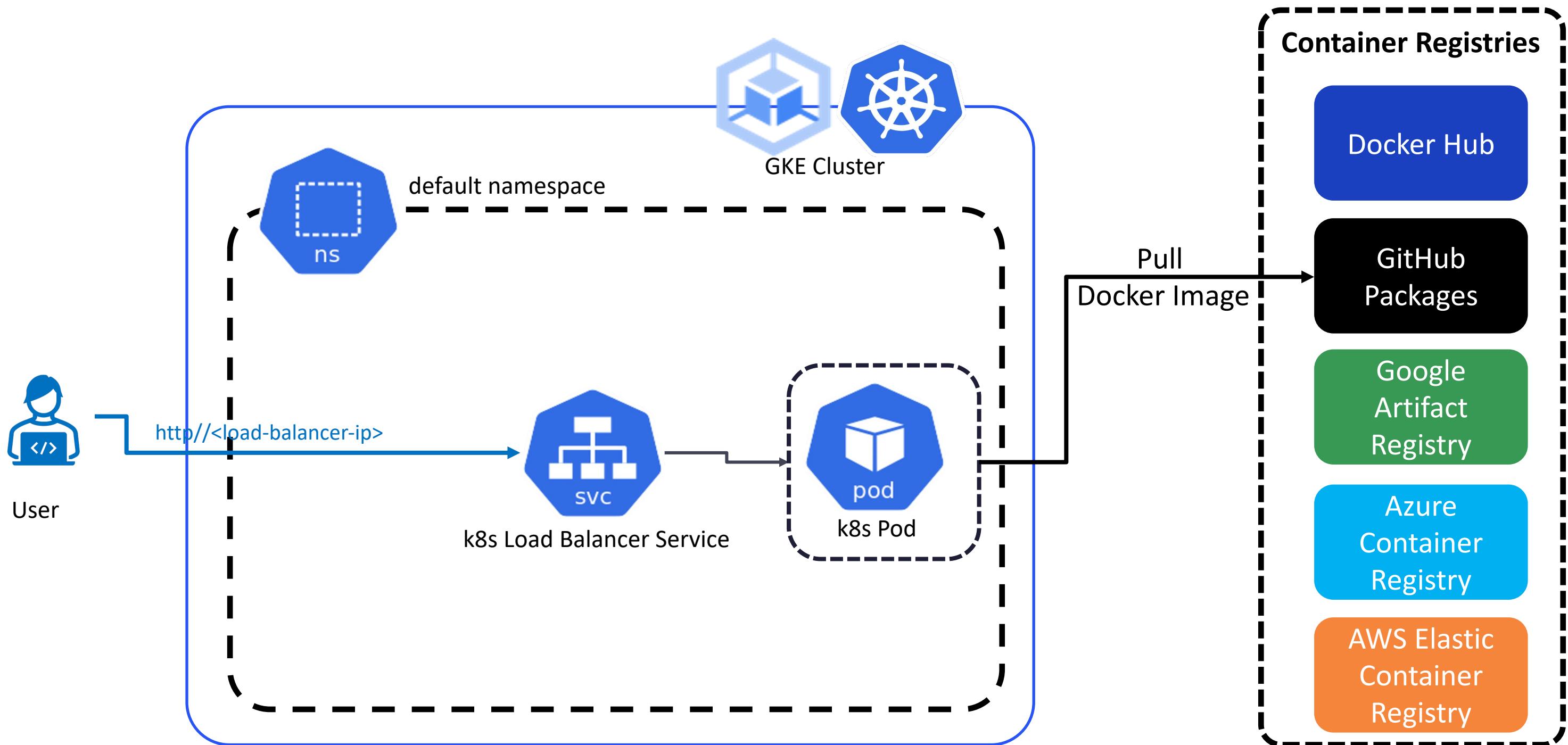


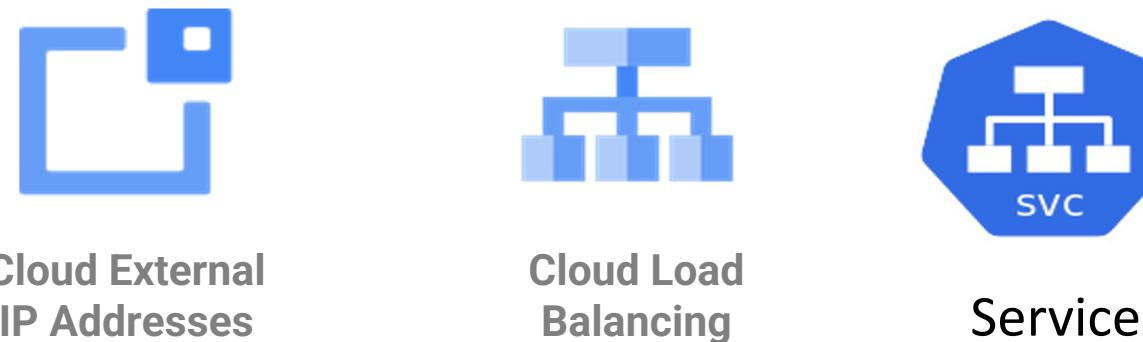
Kubernetes - Multi-Container Pods

- We can have multiple containers in a single POD, provided **they are not of same kind**.
- **Helper Containers (Side-car)**
 - **Data Pullers:** Pull data required by Main Container
 - **Data pushers:** Push data by collecting from main container (logs)
 - **Proxies:** Writes static data to html files using Helper container and Reads using Main Container.
- **Communication**
 - The two containers can easily communicate with each other easily as they share same **network space**.
 - They can also easily share **same storage space**.
- Multi-Container Pods is a **rare use-case** and we will try to focus on core fundamentals.



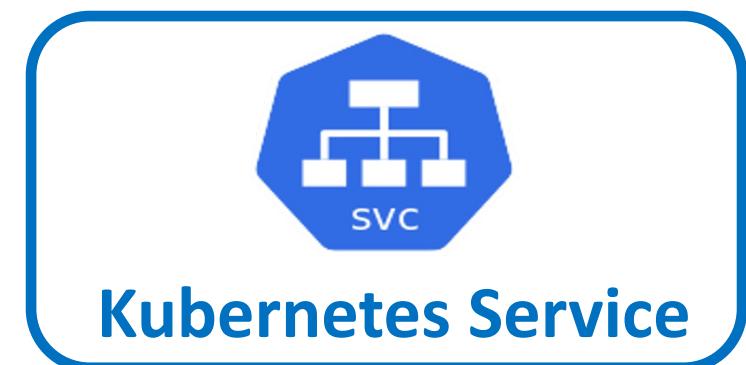
Kubernetes - Pods and Services





Kubernetes Services

Load Balancer



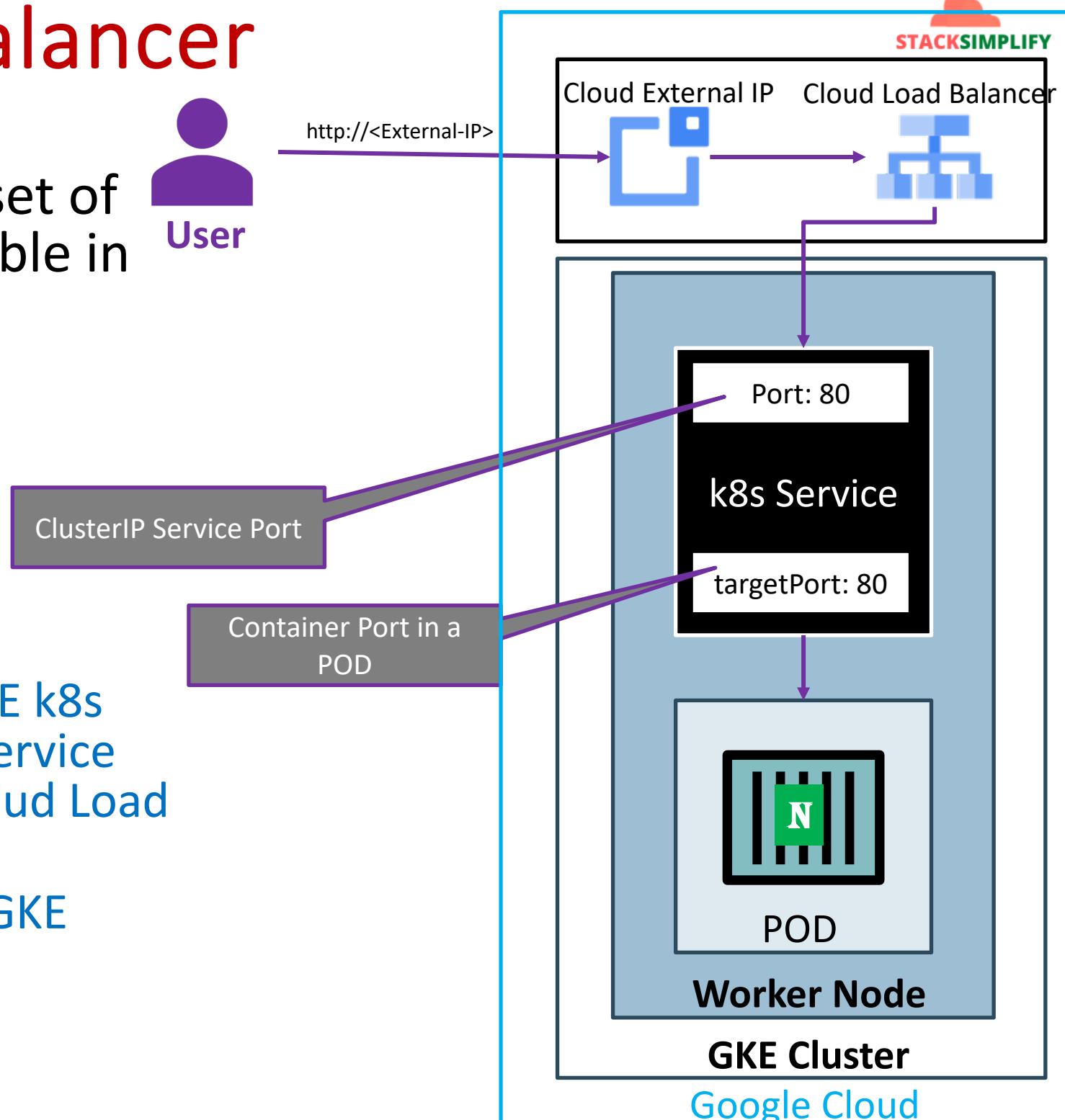
Kubernetes Service - LoadBalancer

- We can expose an application running on a set of PODs using different types of Services available in k8s.

- ClusterIP Service(Internal to k8s cluster)
- NodePort Service (Internet + internal)
- LoadBalancer Service (Internet + internal)
- Ingress Service (Internet + internal)

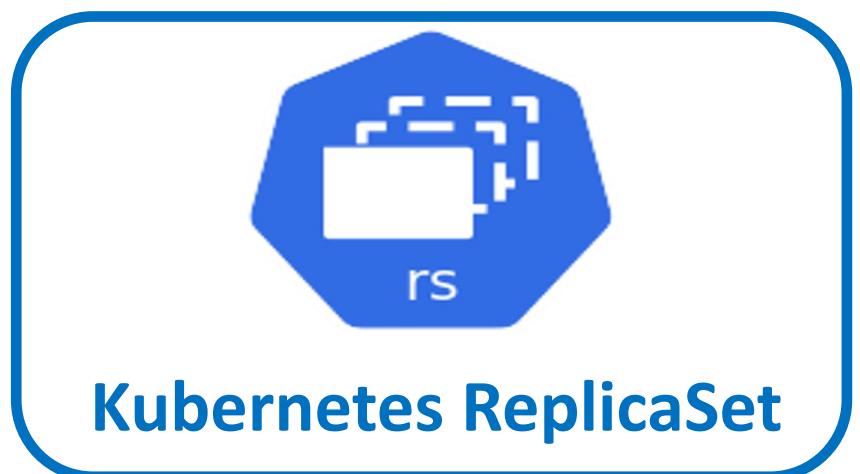
• LoadBalancer Service

- To access our application outside of Google GKE k8s cluster, we can use Kubernetes LoadBalancer service which will be eventually mapped to Google Cloud Load Balancer.
- When we deploy k8s load balancer service in GKE Cluster, the following will be created
 - Google Cloud Load Balancer
 - Google Cloud External IP



Declarative

Kubernetes ReplicaSet



Kubernetes - ReplicaSet

High
Availability

Scaling



ReplicaSet

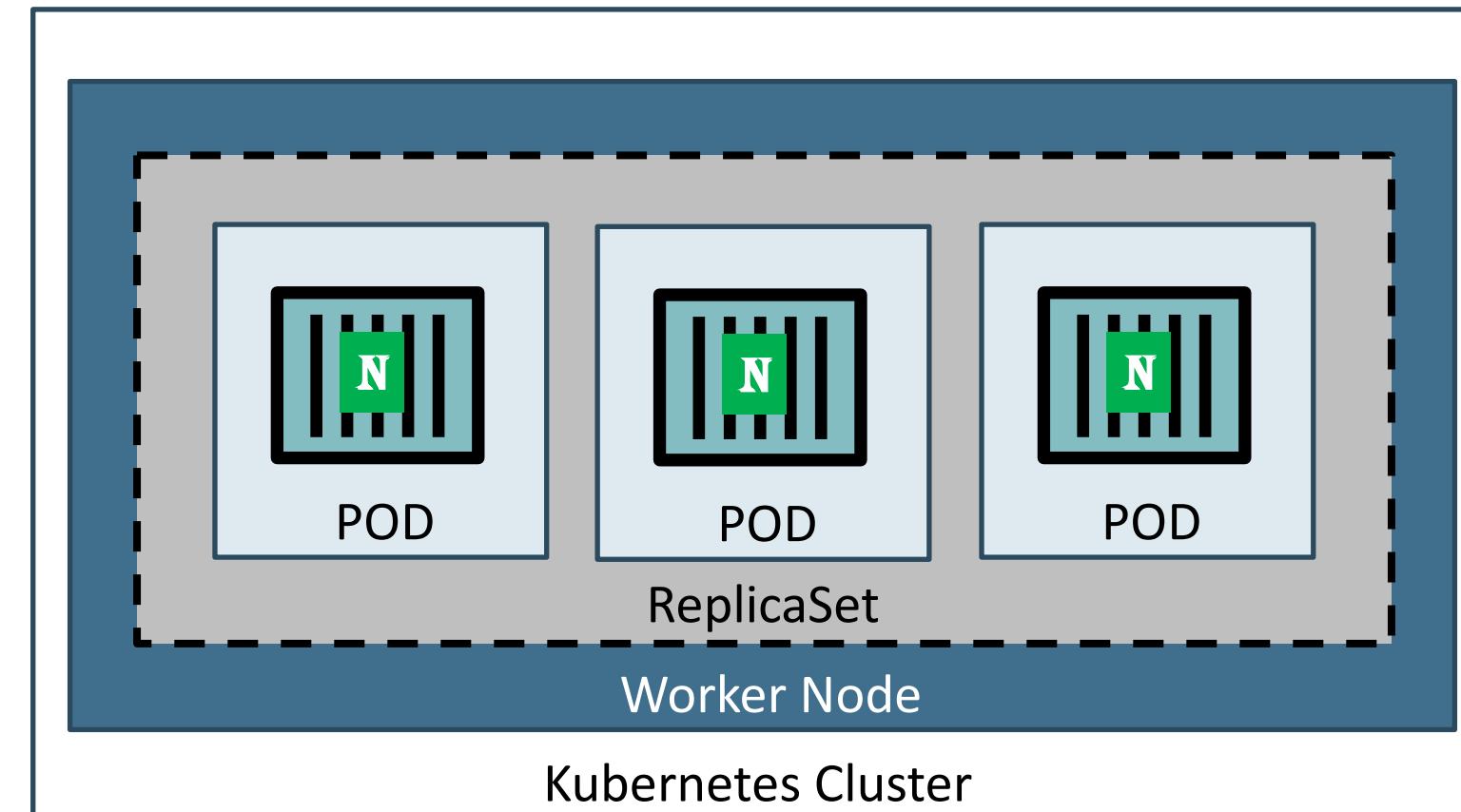
Load Balancing

Labels &
Selectors

Kubernetes - ReplicaSet

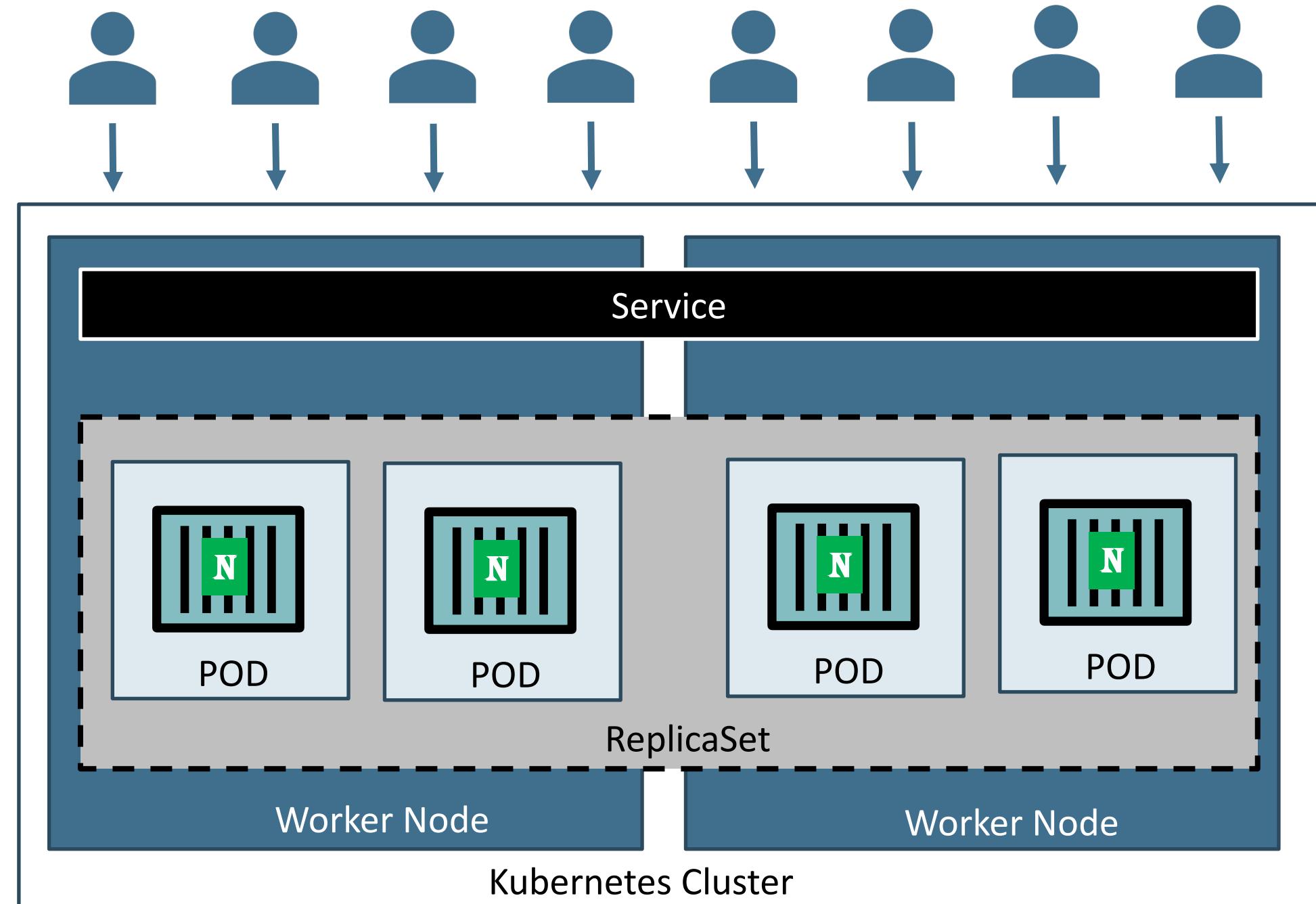
- A ReplicaSet's purpose is to maintain a **stable set of replica Pods** running at any given time.
- If our application crashes (any pod dies), replicaset will **recreate** the pod immediately to ensure the configured number of pods running at any given time.

Reliability
Or
High Availability



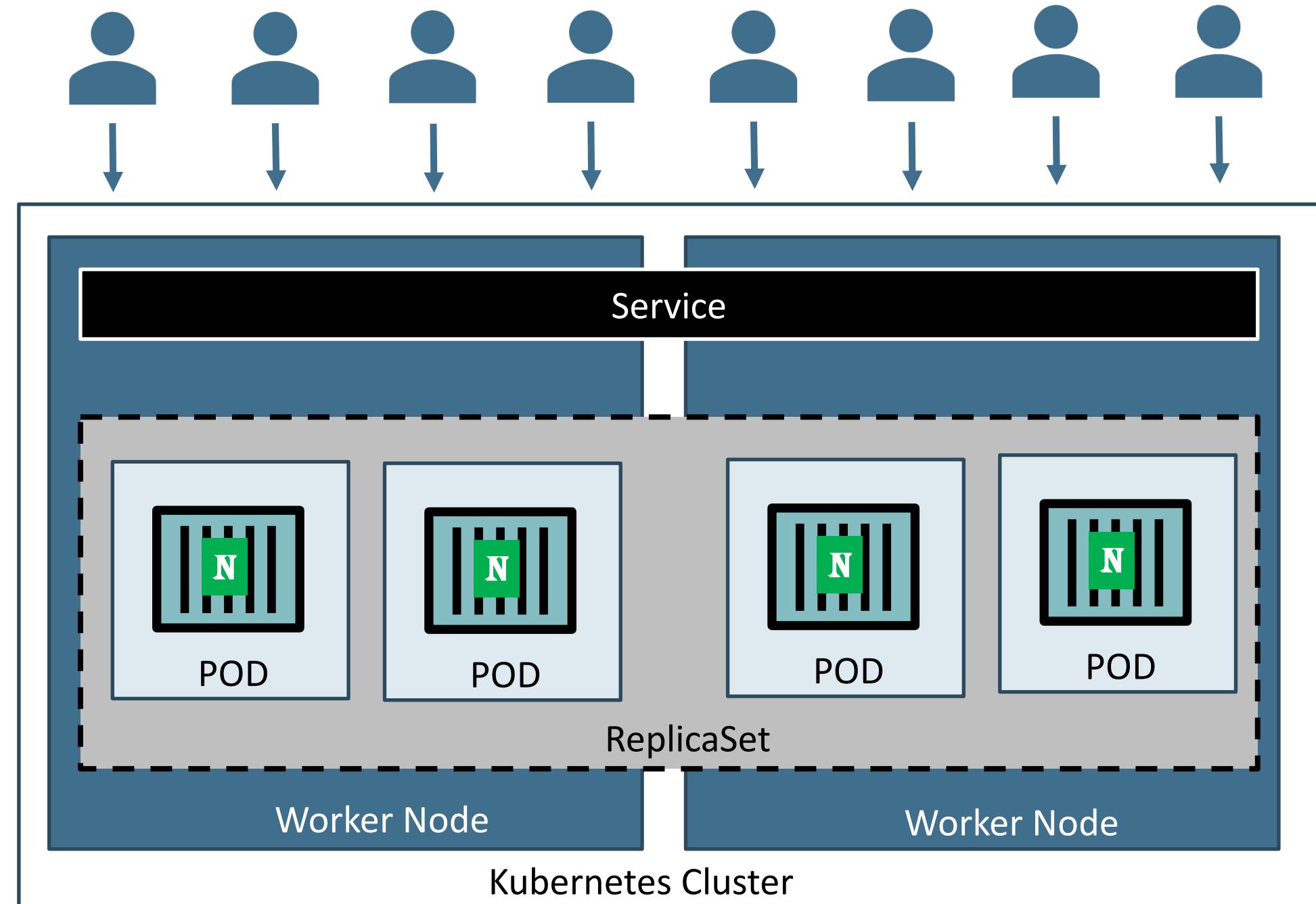
Kubernetes - ReplicaSet

- Load Balancing
- To avoid overloading of traffic to single pod we can use **load balancing**.
- Kubernetes provides pod load balancing **out of the box** using **Services** for the pods which are part of a ReplicaSet
- **Labels & Selectors** are the **key items** which **ties** all 3 together (Pod, ReplicaSet & Service), we will know in detail when we are writing YAML manifests for these objects

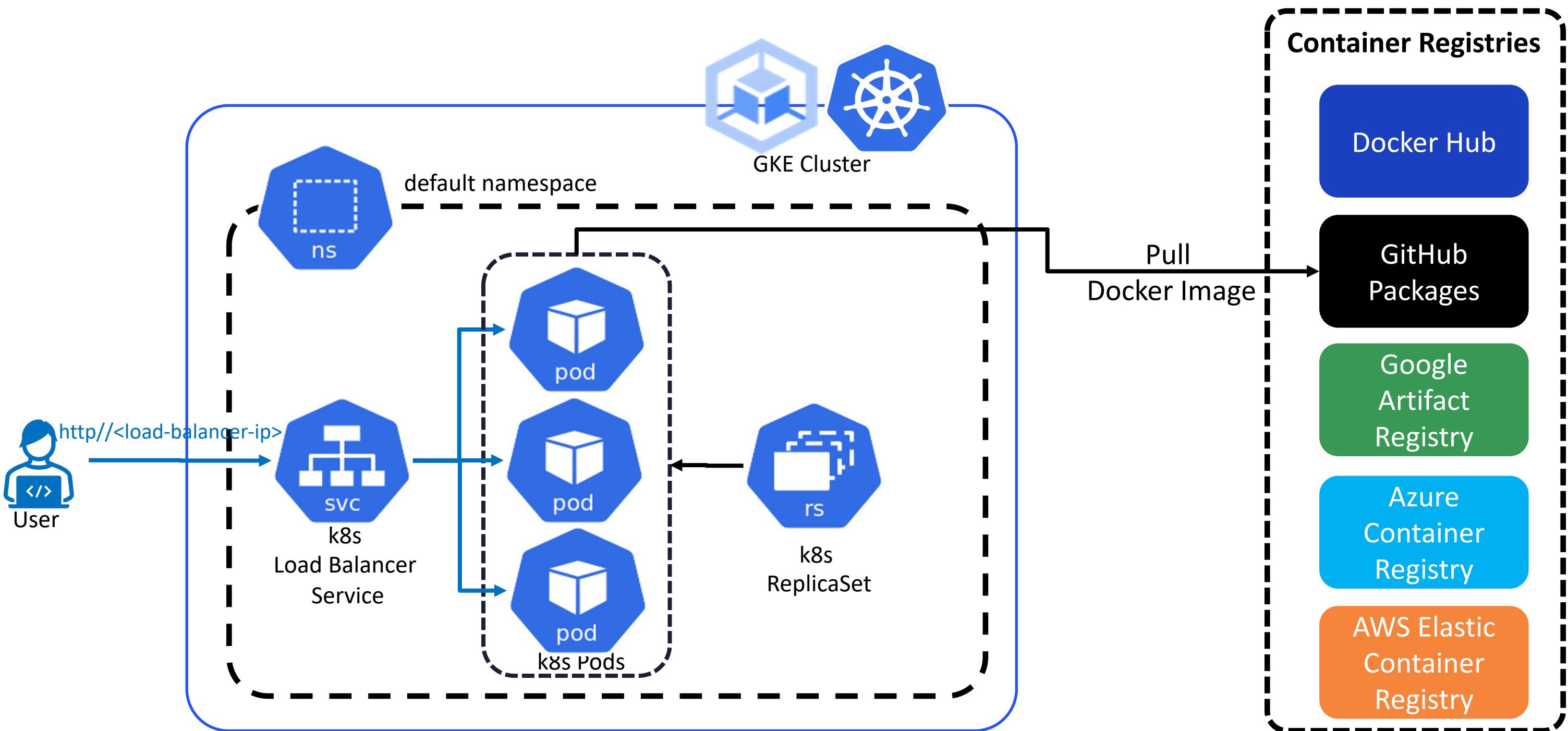


Kubernetes - ReplicaSet

- Scaling
- When load become too much for the number of existing pods, Kubernetes enables us to easily **scale** up our application, adding additional pods as needed.
- This is going to be **seamless and super quick**.

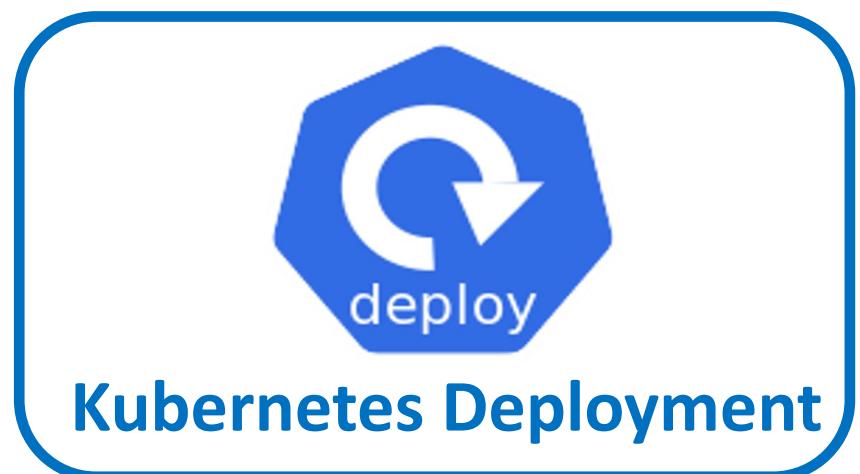


Kubernetes – ReplicaSets and Services

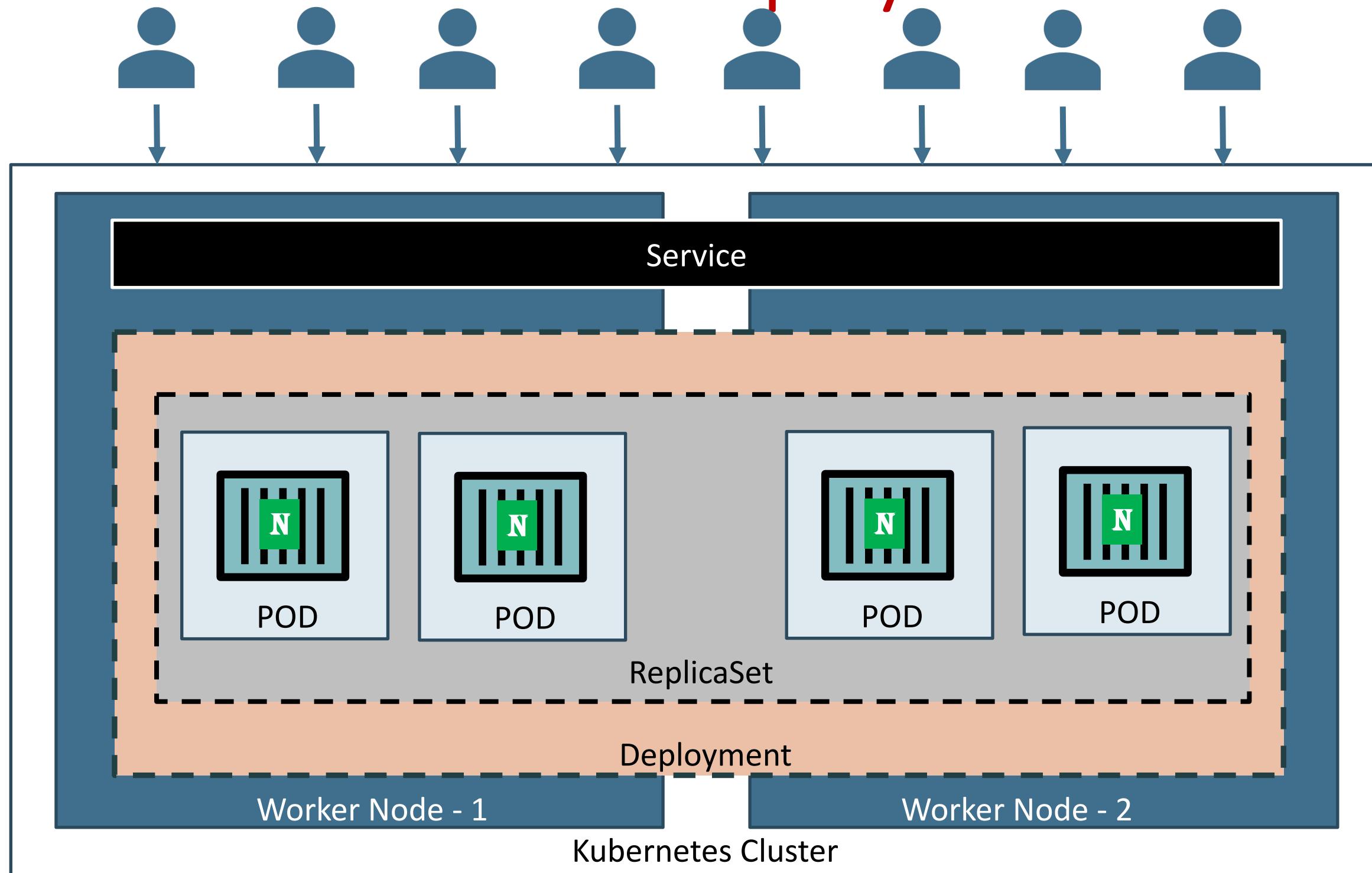


Imperative

Kubernetes Deployment



Kubernetes - Deployment



Kubernetes - Deployment Usecases

Deployment + Service
(Imperative)

1



Kubernetes
Deployment

Update Deployment
(Set Image, Edit)

2

Roll back to Deployment
revisions
(specific or previous version)

3

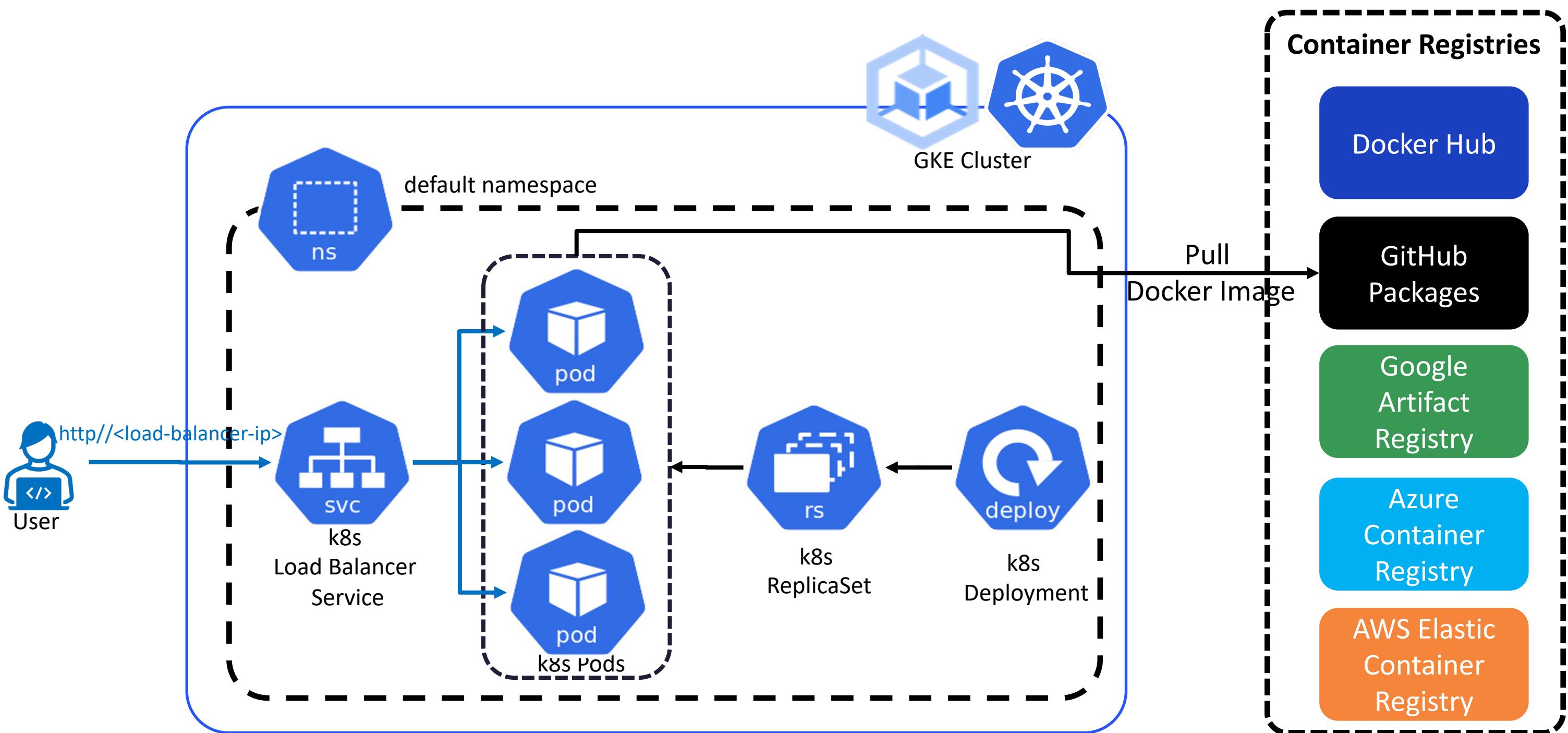
Deployments
(Scale Up / Scale Down)

4

5

Deployment + Service
(Declarative)

Kubernetes - Deployments





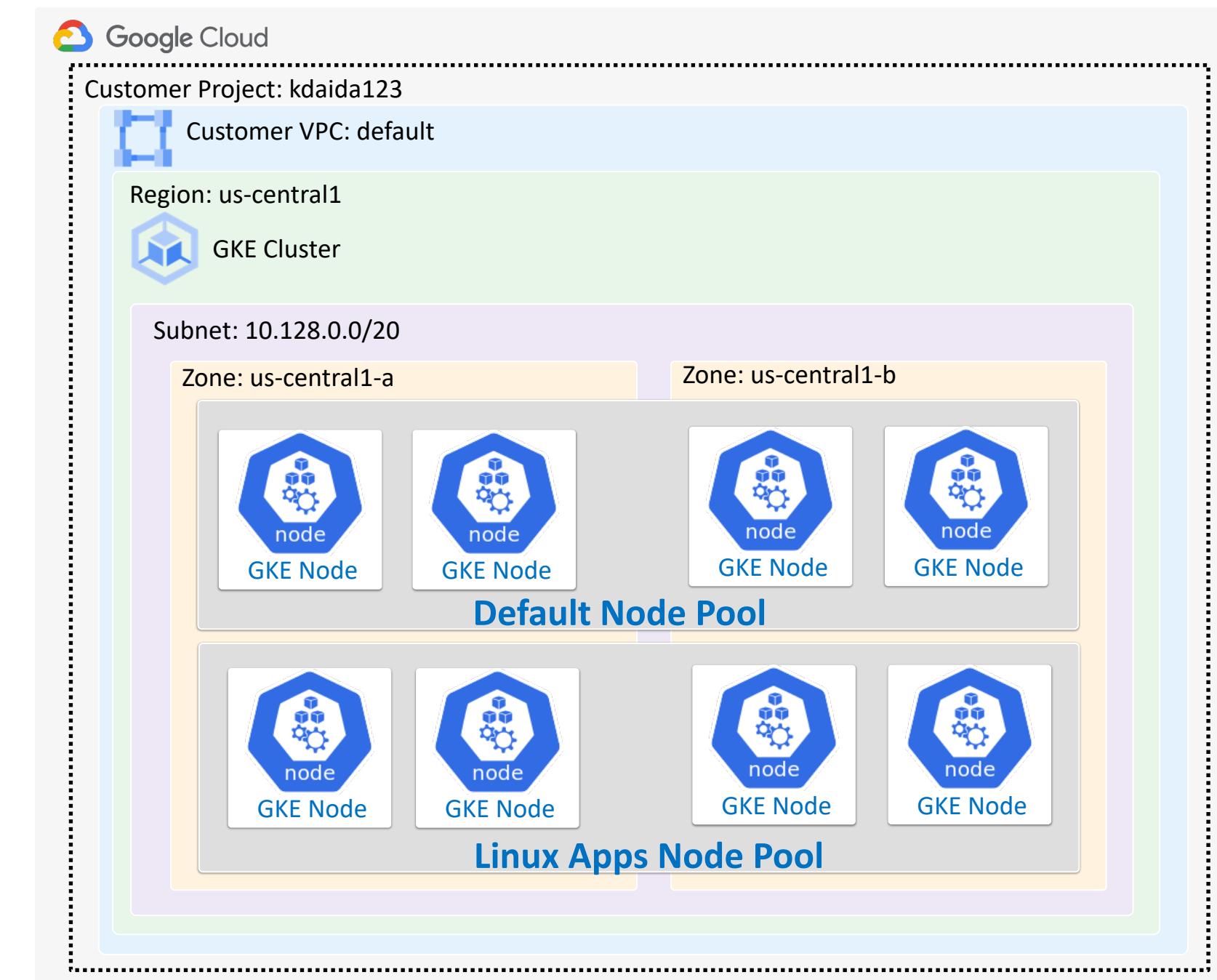
Google Kubernetes Engine

Node Pools and Node Selectors



GKE Node Pools

- **Node Pools:** Group of nodes within a cluster that all have the same configuration
- Each node in the pool has a Kubernetes node label, cloud.google.com/gke-nodepool, which has the node pool's name as its value
 - [cloud.google.com/gke-nodepool: default-pool](#)
 - [cloud.google.com/gke-nodepool: linuxapps-pool](#)
- When you create a cluster, a [default node pool](#) is created
- We can create [more](#) nodepools based on need



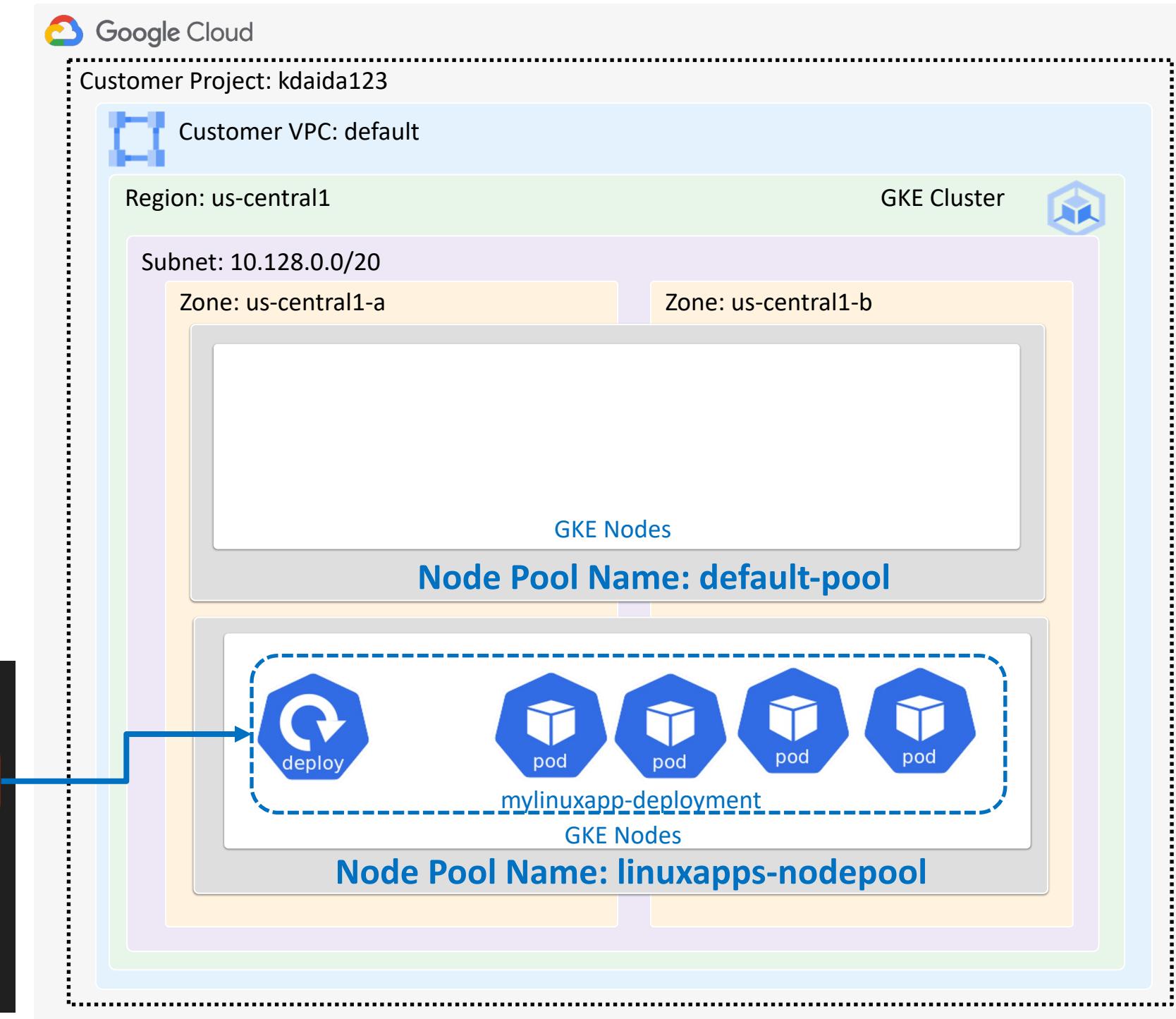
GKE Node Pools

- Node Pools can be created with different configs based on our application needs (categorize)
 - Node pool with [local SSDs](#)
 - Node pool with [minimum CPU Platform](#)
 - Node pool with [spot VMs](#)
 - Node pool with [specific node image](#)
 - Node pool with [specific machine types](#)
- Node pools can be [created, updated and deleted](#) without affecting the whole cluster
- We [cannot make changes to a single node](#) in a node pool, any change will be applied to all nodes in a node pool
- We can [resize node pool](#) by adding or removing nodes
- We can [enable Cluster Autoscaler](#) in node pool to automatically increase or decrease nodes in a node pool based on usage

Kubernetes Node Selectors

- **Node Selectors:** Node selectors are a simple way to control where Pods get scheduled in a Kubernetes cluster, ensuring that they land on nodes with specific characteristics or attributes
- It is a field of PodSpec that specifies a map of key-value pairs.
- You can explicitly deploy a Pod to a specific node pool by setting a nodeSelector in the Pod manifest.

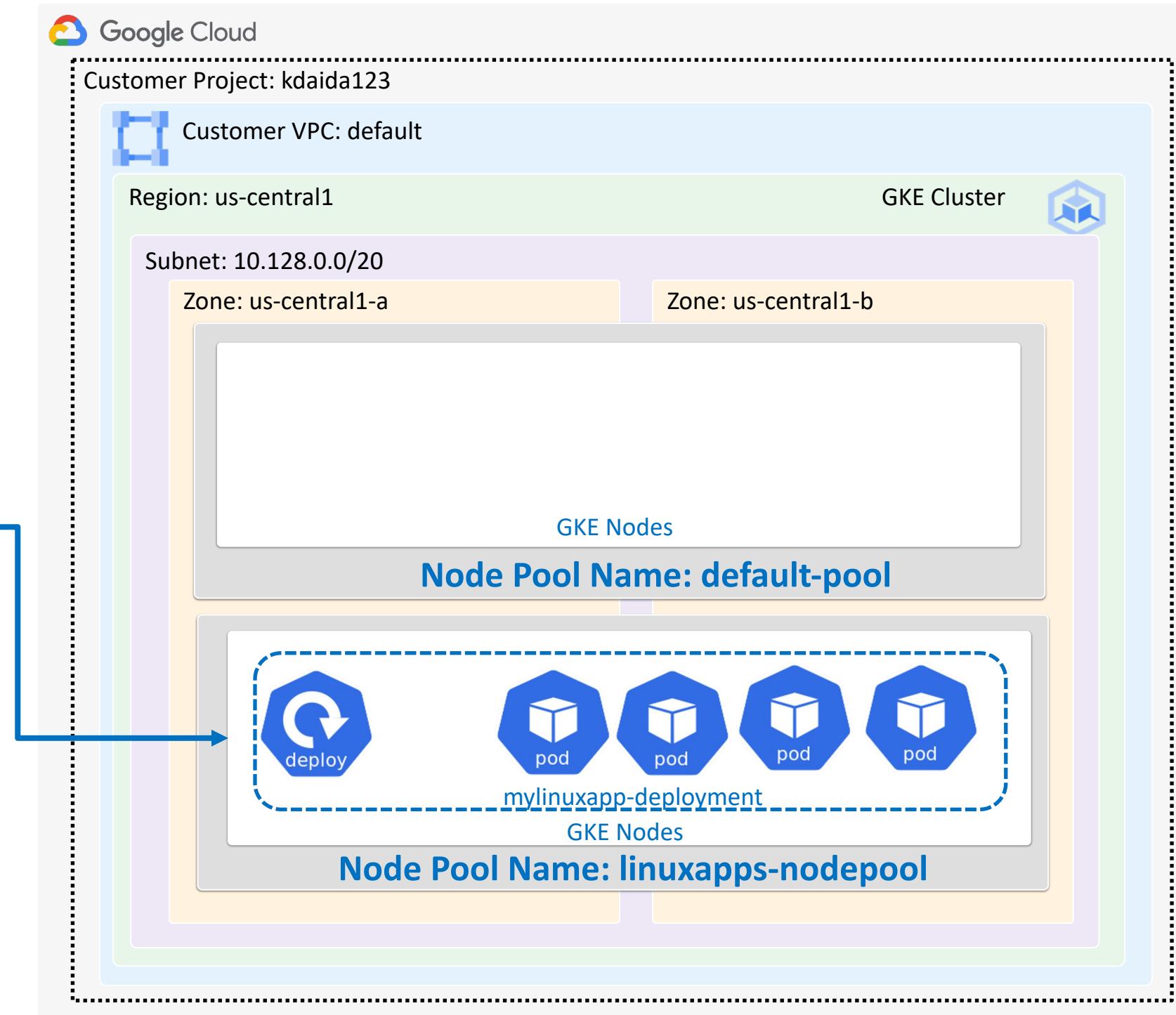
```
spec:  
  # To schedule pods based on NodeSelectors  
  nodeSelector:  
    cloud.google.com/gke-nodepool: linuxapps-nodepool  
  containers:  
    - name: mylinuxapp-container  
      image: ghcr.io/stacksimplify/kubenginx:1.0.0  
      ports:  
        - containerPort: 80
```



GKE Node Pools and Node Selectors

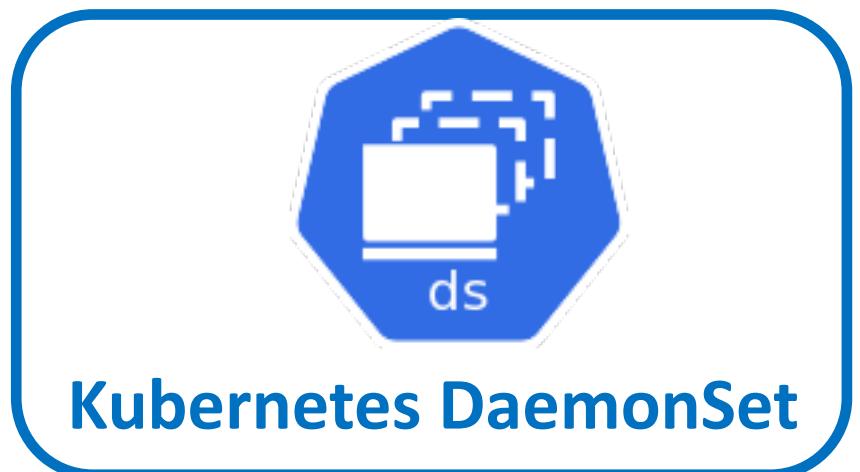
```

apiVersion: apps/v1
kind: Deployment
metadata:
  name: mylinuxapp-deployment
spec:
  replicas: 3
  selector:
    matchLabels:
      app: mylinuxapp
  template:
    metadata:
      name: mylinuxapp-pod
      labels:
        app: mylinuxapp
    spec: Will schedule pods on linuxapps-nodepool
      # To schedule pods on based on NodeSelectors
      nodeSelector:
        cloud.google.com/gke-nodepool: linuxapps-nodepool
      containers:
        - name: mylinuxapp-container
          image: ghcr.io/stacksimplify/kubenginx:1.0.0
          ports:
            - containerPort: 80
  
```



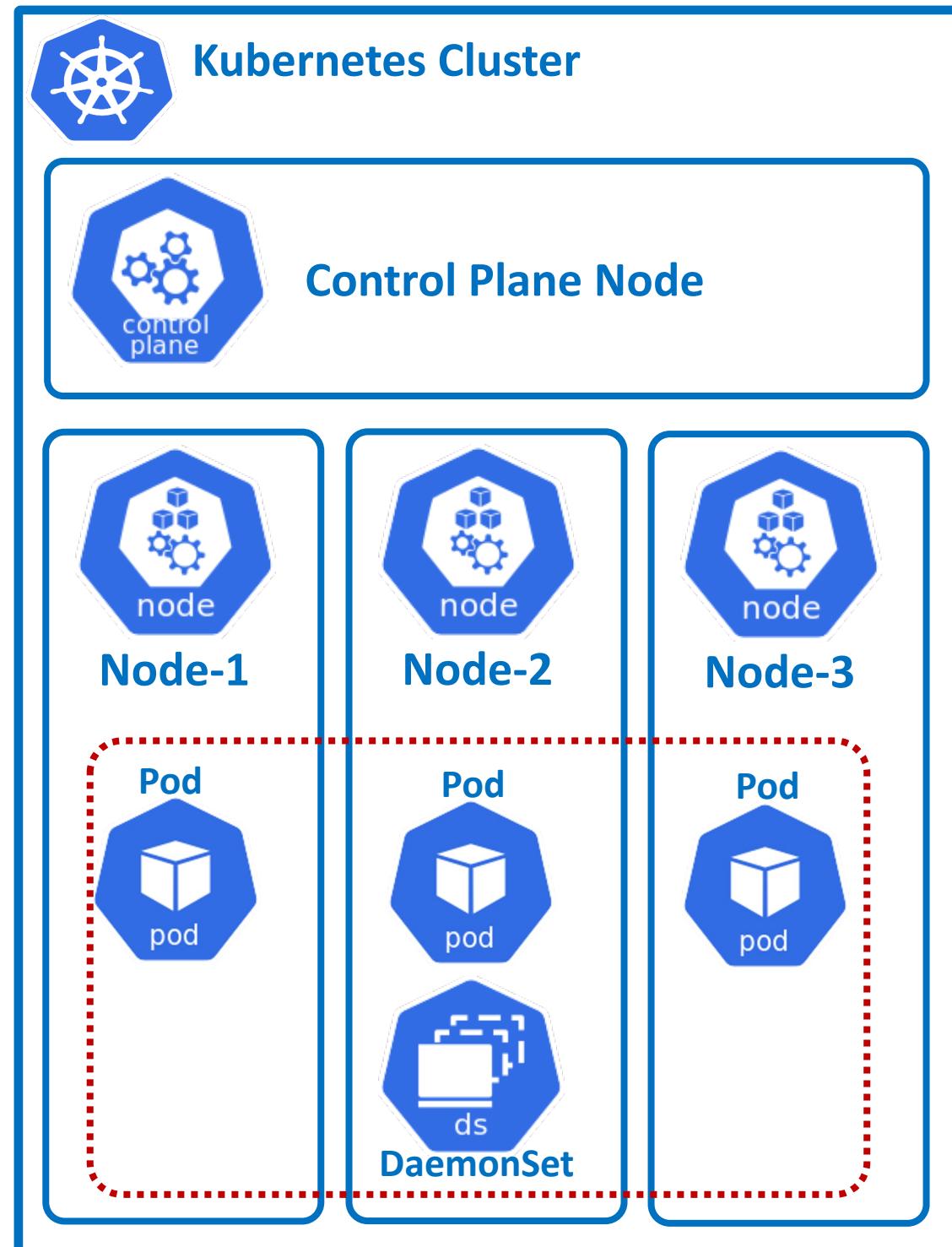


Kubernetes DaemonSet



Kubernetes DaemonSet

- **DaemonSet**: A DaemonSet ensures that **all** (or **some**) **Nodes** run a copy of a Pod.
- As nodes are **added to the cluster**, Pods are added to them.
- As nodes are **removed from the cluster**, those Pods are garbage collected.
- Deleting a DaemonSet will **clean up** the Pods it created
- **Usecases**
 - running a **cluster storage daemon** on every node
 - running a **logs collection daemon** on every node
 - running a **node monitoring daemon** on every node





Kubernetes Job



Kubernetes Jobs

- **Job:** A Job **creates pods for executing tasks** (Example: Log analysis)
- The Job **monitors the completion status of pods** as they finish their tasks.
- Pods (tasks) run to their **completion**
- When a specified number of successful completions is **reached**, the Job will be marked as **complete**.
- **Deleting a Job** will clean up the Pods it created.
- **Suspending a Job** will delete its active Pods until the Job is **resumed again**.

Kubernetes Job				
NAME	COMPLETIONS	DURATION	AGE	
job1	1/1	34s	51s	

Kubernetes Pod created by Job					
NAME	READY	STATUS	RESTARTS	AGE	
job1-dddvs	0/1	Completed	0	55s	

Kubernetes Jobs

- **Usecases**

- **Batch Processing:** ETL Jobs, Log analysis and report generation
- **Parallel Processing:** parallelizing data analysis or image processing

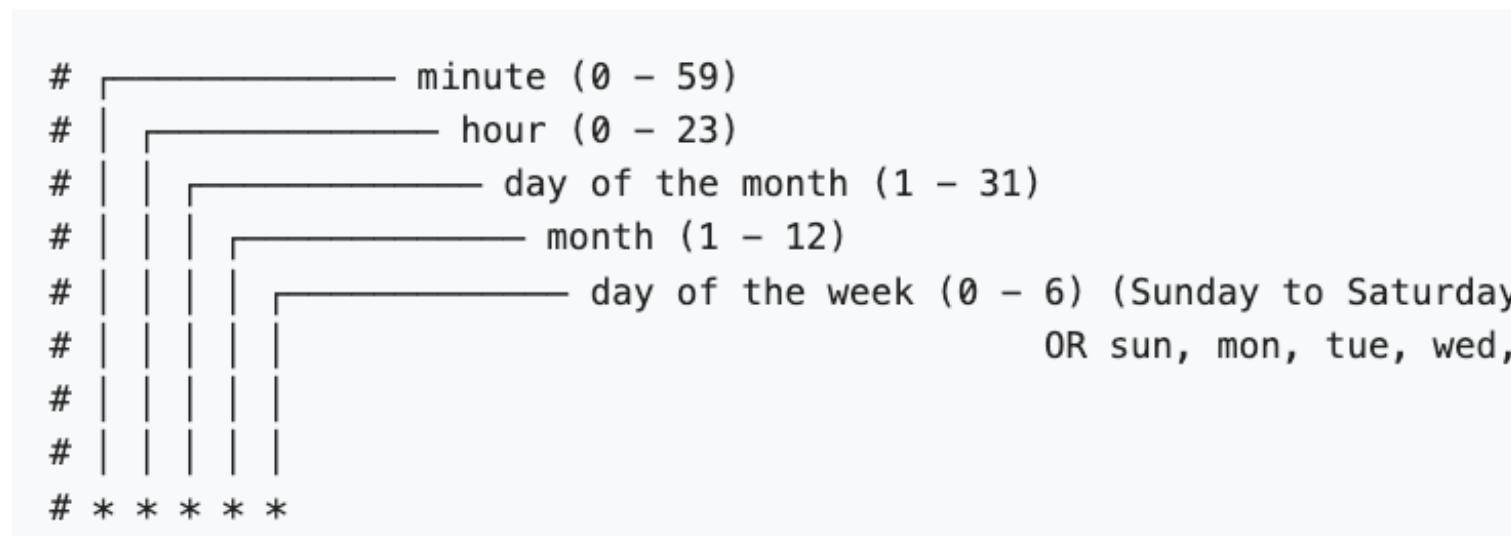
4 Important Settings for a k8s Job we will master with 4 demos

```
apiVersion: batch/v1
kind: Job
metadata:
  # Unique key of the Job ins
  name: job1
spec:
  template:
    metadata:
      name: job1
    spec:
      containers:
        - name: job1
          image: alpine
          command: ['sh', '-c',
# Do not restart contain
restartPolicy: Never
# backoffLimit: Number of r
backoffLimit: 4 # Default v
# completions: Specify how
completions: 4
# parallelism: Specifies th
parallelism: 2
# activeDeadlineSeconds: Sp
activeDeadlineSeconds: 5 #
```

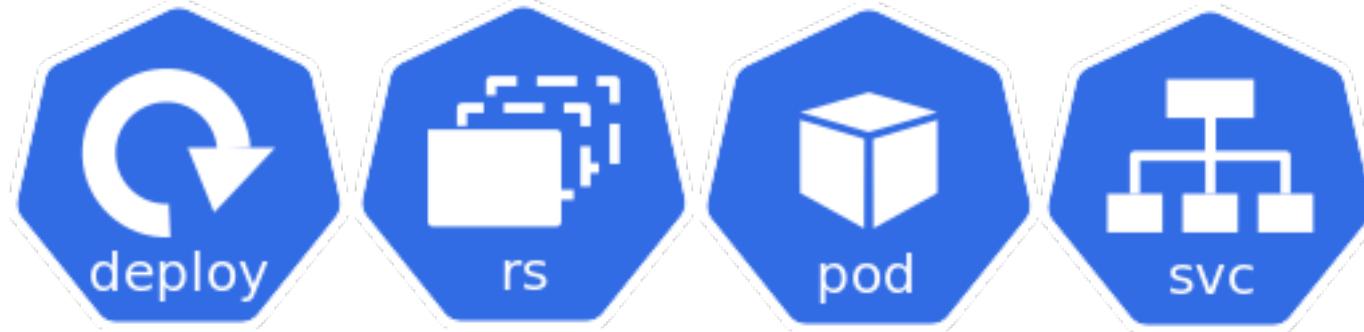
Kubernetes Cron Job



- **Cron Job:** Used for performing regular scheduled actions such as
 - Backups
 - Log rotation
 - Data cleanup
 - One CronJob object is like one line of a crontab file on a Unix system
 - **spec.schedule**



```
# cronjob.yaml
apiVersion: batch/v1
kind: CronJob
metadata:
  name: cron-job-demo
spec:
  schedule: "*/1 * * * *"
  concurrencyPolicy: Allow
  startingDeadlineSeconds: 100
  suspend: false
  successfulJobsHistoryLimit: 3
  failedJobsHistoryLimit: 1
  jobTemplate:
    spec:
      template:
        spec:
          containers:
            - name: hello
              image: busybox
              args:
                - /bin/sh
                - -c
                - date; echo "Hello, World!"
        restartPolicy: OnFailure
```



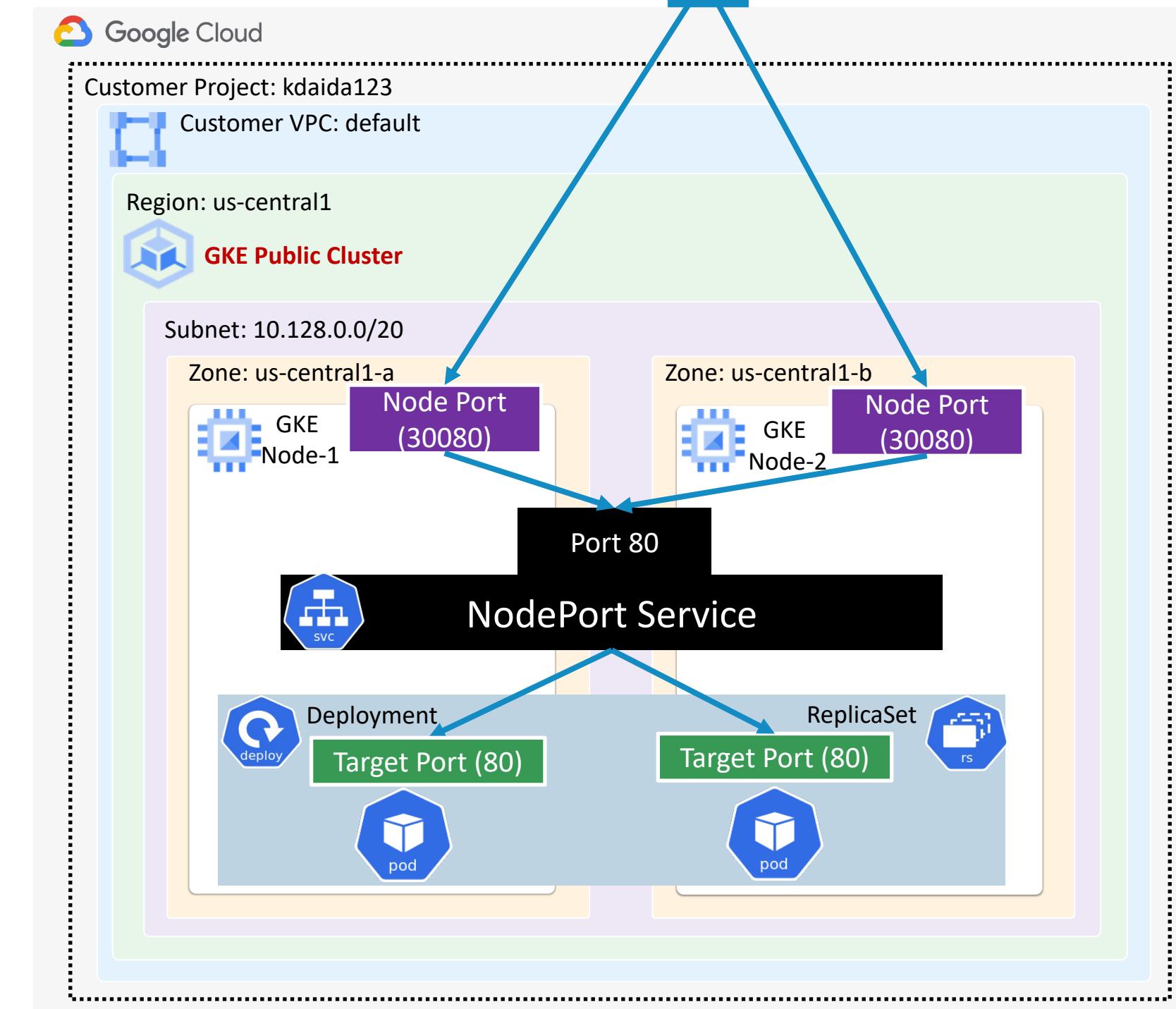
Google Kubernetes Engine

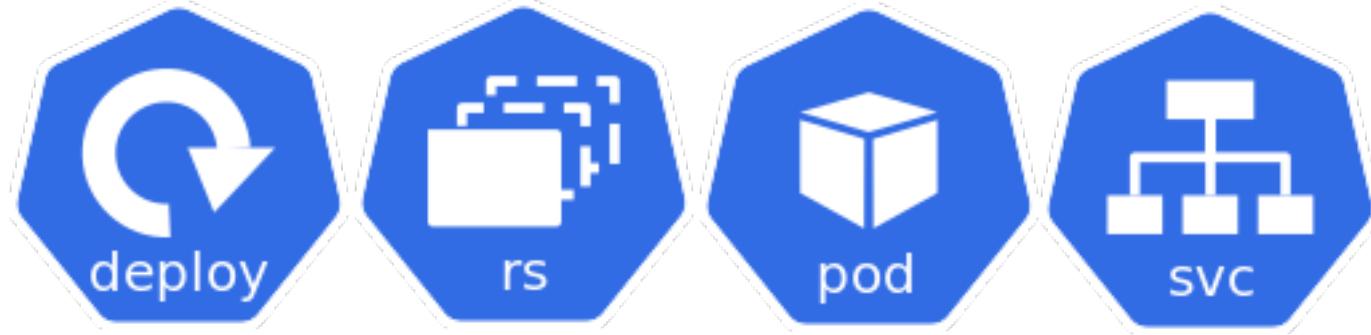


Kubernetes NodePort Service

Kubernetes NodePort Service

- NodePort service allows external clients to access pods via network ports opened on the Kubernetes nodes
- Node Port service port range 30000 to 32768 on Kubernetes Nodes (Range is customizable)
- In real-world, NodePort Services are not used in production grade implementations
- NodePort Services are generally used to test our application Internet provided our Kubernetes Nodes are at internet edge[Public Cluster] (Saves Load balancer costs in testing phases)
- http://NODE-IP:NODEPORT is generally not a recommended practice to use in production for our application.





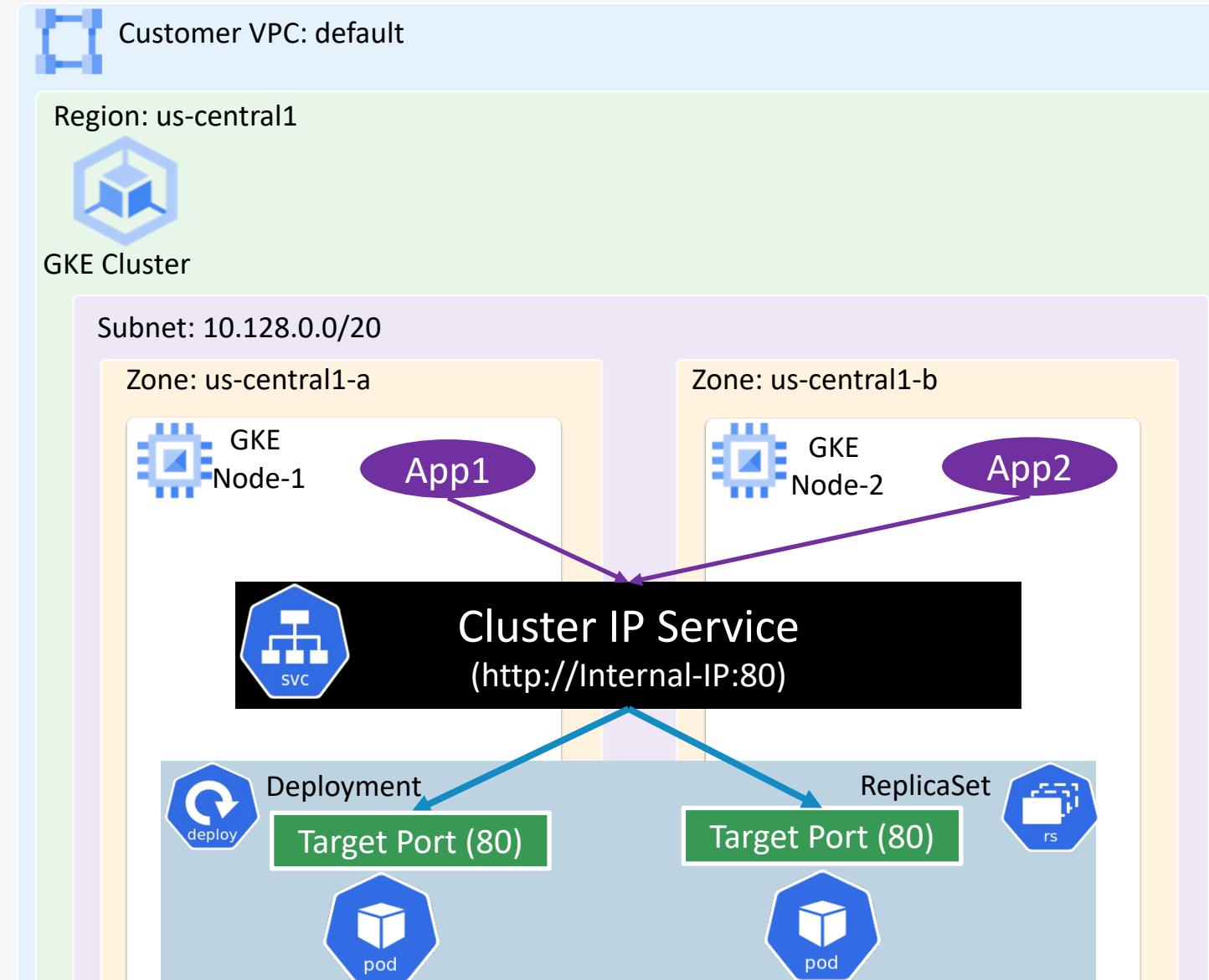
Google Kubernetes Engine



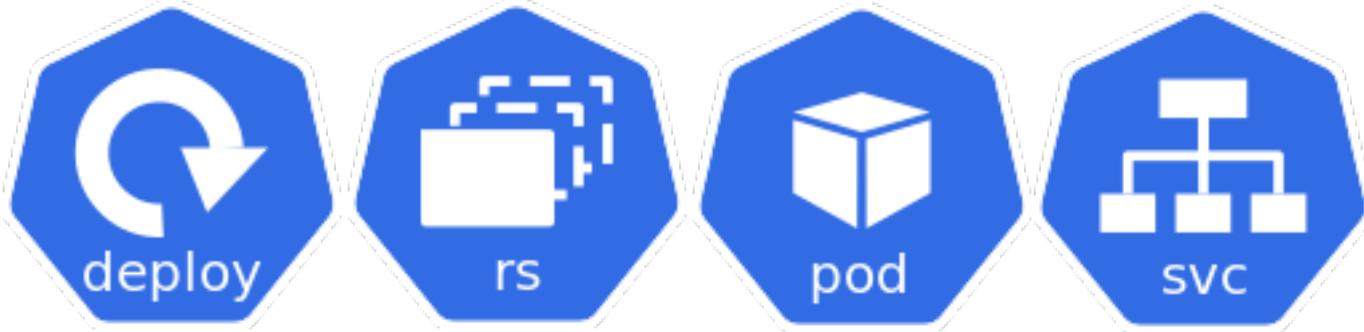
Kubernetes ClusterIP Service

Kubernetes ClusterIP Service

Customer Project: kdaida123



- **Cluster IP Service :** Internal clients send requests to a **stable internal IP address or DNS name**
- Simple we can call it as **Layer4 internal load balancer**
- **Why do we need it ?**
- Pods IPs are **not static**, when pod restarts those Ips will change.
- To load balance traffic to pods, we need a **static DNS name**, that we get via Cluster IP service for internal clients
- **Cluster IP Service DNS Name:**
 - <servicename>.<namespace>.svc.cluster.local
 - **my-clusterip-service.default.svc.cluster.local**



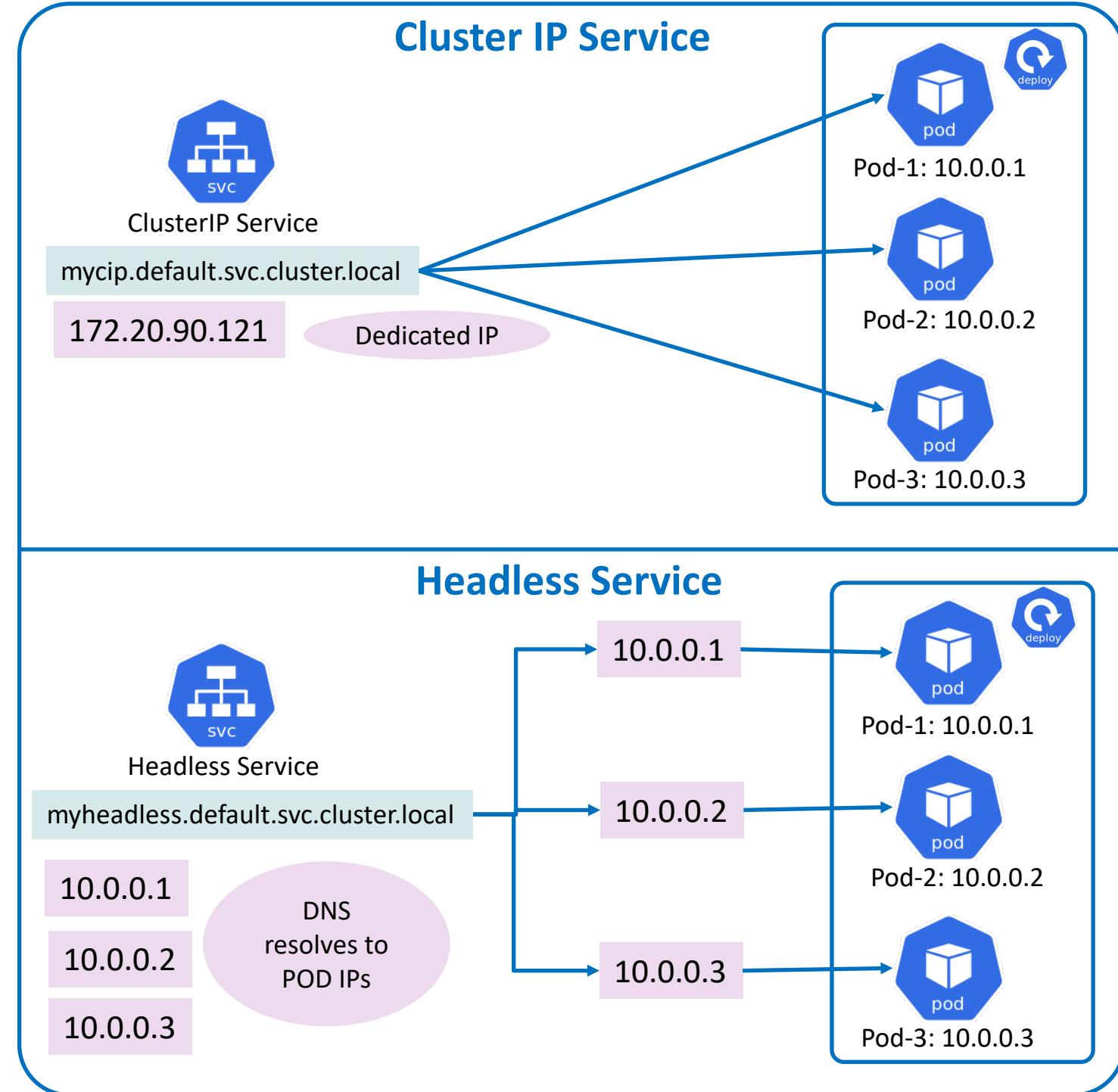
Google Kubernetes Engine



Kubernetes Headless Service

Kubernetes Headless Service

- For Headless Service, ClusterIP is not allocated
- Headless Service DNS directly resolves to POD IP addresses
- Headless Service directly sends traffic to Pod with Pod IP
- **Usecases**
 - Statefulsets
 - Database Clusters (MySQL, Cassandra)
 - Messaging Systems (Kafka, RabbitMQ)



Kubernetes Headless Service

```
dkalyanreddy@cloudshell:~/08-03-Kubernetes-Headless-Service (gcplearn9)$ kubectl get pods -o wide
NAME                               READY   STATUS    RESTARTS   AGE     IP           NODE
curl-pod                           1/1    Running   2 (5m52s ago)   25m    10.36.2.11   gke-standard-pu
blic-clus-default-pool-d36a776f-75tc  <none>
myapp1-deployment-b99ccfb9d-285dm   1/1    Running   0          14m    10.36.2.12   gke-standard-pu
blic-clus-default-pool-d36a776f-75tc  <none>
myapp1-deployment-b99ccfb9d-9vgkx   1/1    Running   0          14m    10.36.1.10   gke-standard-pu
blic-clus-default-pool-de8808a7-6lcx  <none>
myapp1-deployment-b99ccfb9d-f4dbq   1/1    Running   0          14m    10.36.2.13   gke-standard-pu
blic-clus-default-pool-d36a776f-75tc  <none>
myapp1-deployment-b99ccfb9d-twr99   1/1    Running   0          14m    10.36.0.15   gke-standard-pu
blic-clus-default-pool-a509c7cc-zcsj  <none>
dkalyanreddy@cloudshell:~/08-03-Kubernetes-Headless-Service (gcplearn9)$
```

```
apiVersion: v1
kind: Service
metadata:
  name: myapp1-headless-service
spec:
  #type: ClusterIP # ClusterIP, # NodePort
  clusterIP: None
  selector:
    app: myapp1
  ports:
  - name: http
    port: 8080 # Service Port
    targetPort: 8080 # Container Port
```

How do you
create
Headless
Service ?

```
~ $ nslookup myapp1-headless-service.default.svc.cluster.local
Server: 10.81.128.10
Address: 10.81.128.10:53

Name: myapp1-headless-service.default.svc.cluster.local
Address: 10.36.2.12
Name: myapp1-headless-service.default.svc.cluster.local
Address: 10.36.2.13
Name: myapp1-headless-service.default.svc.cluster.local
Address: 10.36.0.15
Name: myapp1-headless-service.default.svc.cluster.local
Address: 10.36.1.10
```



Google Kubernetes Engine

Ingress Service



OSI Layers

L4 Layer

Routes Traffic based on network information port (80, 8080) and protocol (TCP, UDP)

L7 Layer

L7 layer has awareness of Application Information like HTTP URL Path, Host Header, Custom Headers.

Additional Reference: <https://cloud.google.com/kubernetes-engine/docs/concepts/service-networking>

GKE Services supported based on OSI Layers

L4 Layer

Protocols

TCP

UDP

GKE Services

Cluster IP Service

NodePort Service

Internal Load Balancer Service

External Load Balancer Service

L7 Layer

Protocols

HTTP

HTTPS

HTTP2

GKE Services

Internal Ingress Service

External Ingress Service

Multi Cluster Ingress Service



Ingress Controller

In GKE, Google hosts
the Ingress Controller.

We just need to ensure that
below setting is enabled in
NETWORKING section in our
GKE Cluster

HTTP Load Balancing

Enabled

Ingress Manifest - Terminology

Ingress Annotations
(Load Balancer Settings)

Ingress Spec
(Define Ingress
Routing Rules, Default
Backend)

Ingress Features: <https://cloud.google.com/kubernetes-engine/docs/how-to/ingress-features>

```

apiVersion: networking.k8s.io/v1
kind: Ingress
metadata:
  name: ingress-ssl
  annotations:
    # External Load Balancer
    kubernetes.io/ingress.class: "gce"
    # Static IP for Ingress Service
    kubernetes.io/ingress.global-static-ip-name: "gke-ingress-extip1"
    # Google Managed SSL Certificates
    networking.gke.io/managed-certificates: managed-cert-for-ingress
spec:
  defaultBackend:
    service:
      name: app3-nginx-nodeport-service
      port:
        number: 80
  rules:
    - http:
        paths:
          - path: /app1
            pathType: Prefix
            backend:
              service:
                name: app1-nginx-nodeport-service
                port:

```

Annotations

Routing Rules

Ingress Context Path Based Routing

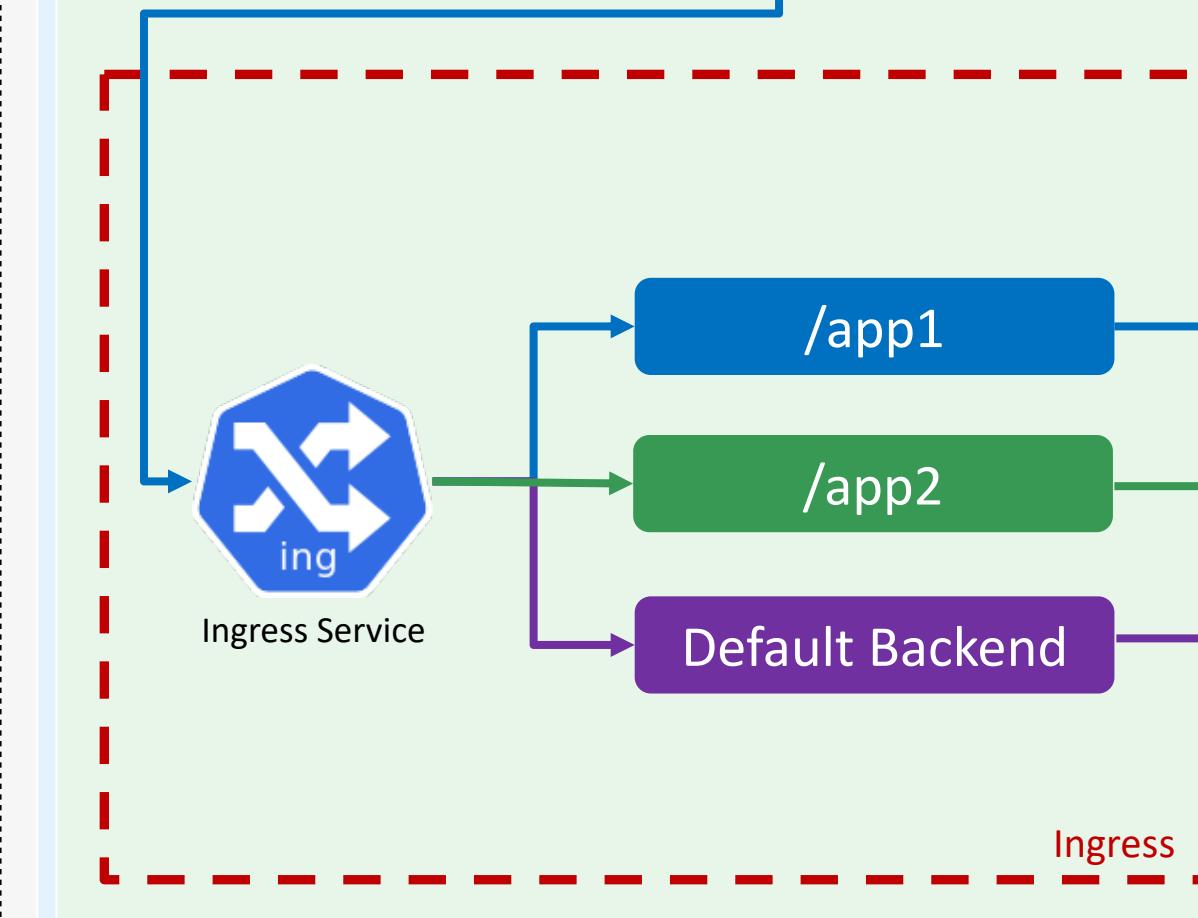
Customer Project: kdaida123

Customer VPC: default

Region: us-central1



GKE Cluster

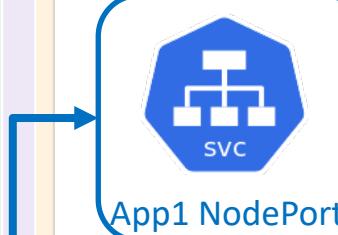
<http://<LB-IP>/app1/index.html><http://<LB-IP>/app2/index.html><http://<LB-IP>>

Subnet: 10.128.0.0/20

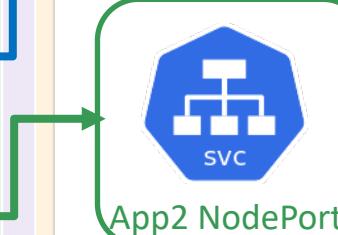
Zone: us-central1-a



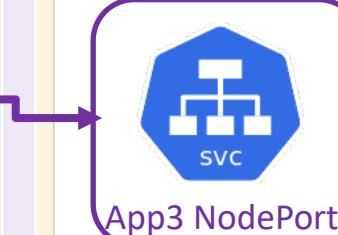
GKE Node-1



App1 NodePort



App2 NodePort



App3 NodePort



pod



rs



deploy



pod



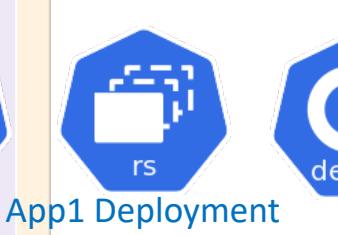
rs



deploy



pod



rs



deploy



pod



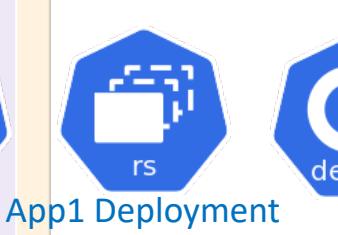
rs



deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



deploy



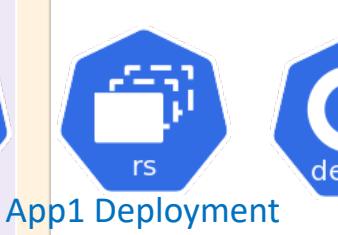
pod



rs



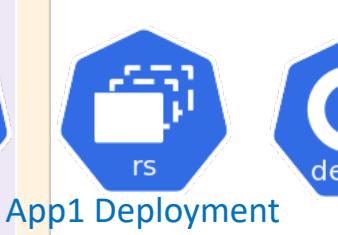
deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



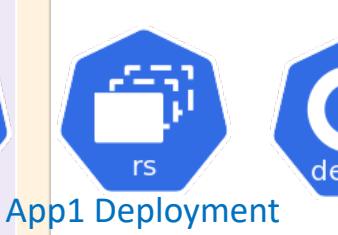
deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



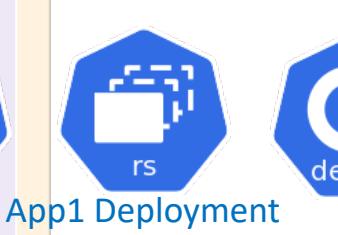
deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



deploy



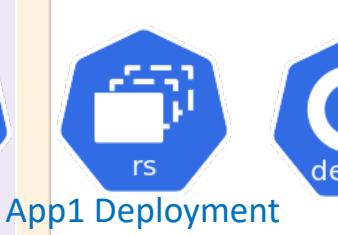
pod



rs



deploy



pod



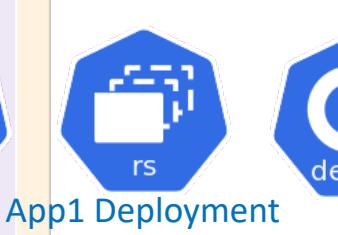
rs



deploy



pod



rs



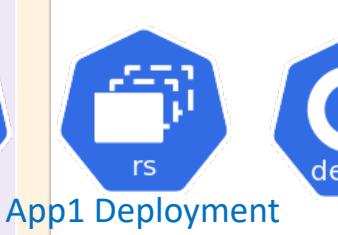
deploy



pod



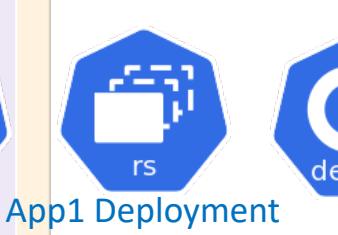
rs



deploy



pod



rs



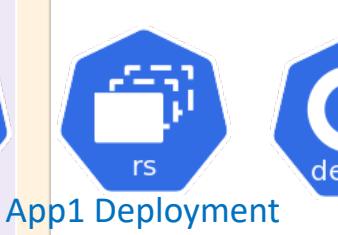
deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



deploy



pod



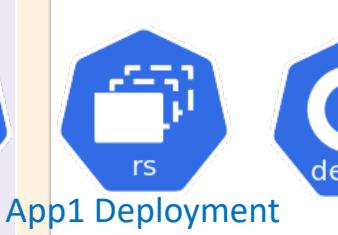
rs



deploy



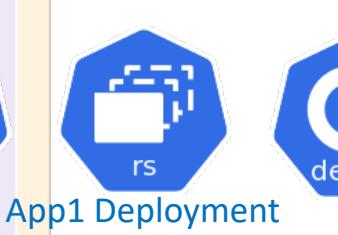
pod



rs



deploy



pod



rs



deploy



pod



rs



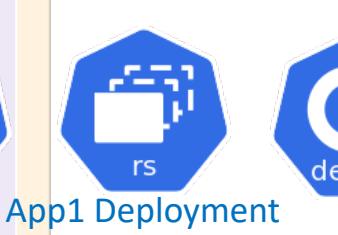
deploy



pod



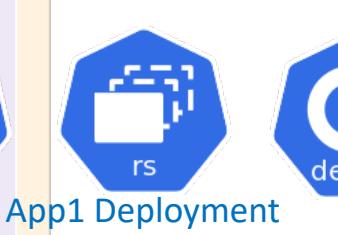
rs



deploy



pod



rs



deploy



pod



rs



deploy



pod



rs



deploy



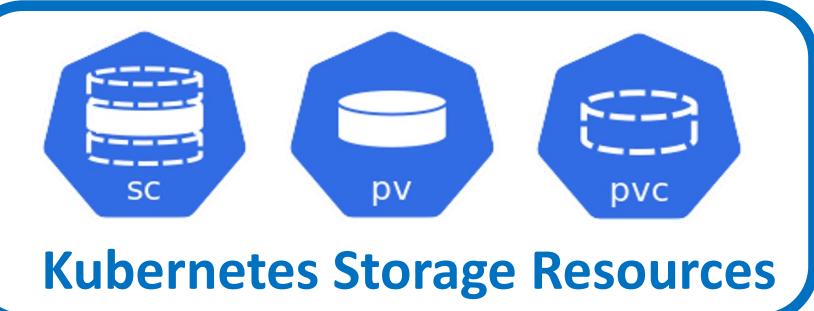
pod</div



Google Kubernetes Engine

Kubernetes Storage

 Persistent Disk



Kubernetes Storage Resources

The diagram shows three hexagonal icons representing Kubernetes storage resources: a solid blue cylinder labeled 'SC' (Storage Class), a white cylinder labeled 'pv' (Persistent Volume), and a dashed blue cylinder labeled 'pvc' (Persistent Volume Claim).

 Helm



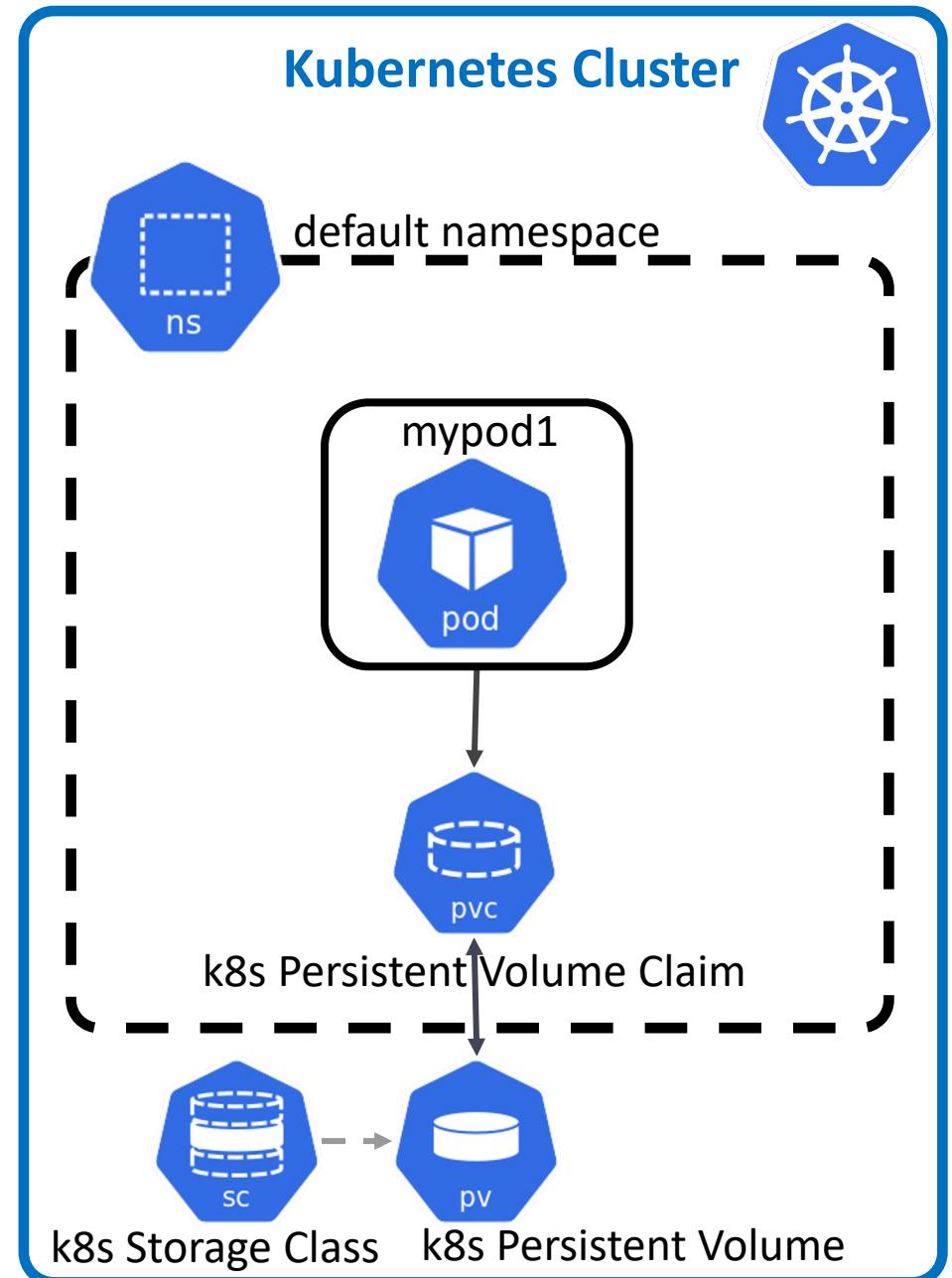
Kubernetes Storage Class

- **Storage Class:** Provides a way for administrators to define the **classes of storage they offer**
 - **standard-rwo:** Provides balanced disks
 - **premium-rwo:** Provides SSD disks

```
Kalyans-Mac-mini:google-kubernetes-engine kalyanreddy$ kubectl describe sc premium-rwo
Name:           premium-rwo
IsDefaultClass: No
Annotations:    components.gke.io/component-name=pdcsi,components.gke.io/component-version=0.13.2,components.gke.io/layer=addon
Provisioner:    pd.csi.storage.gke.io
Parameters:     type=pd-ssd
AllowVolumeExpansion: True
MountOptions:   <none>
ReclaimPolicy: Delete
VolumeBindingMode: WaitForFirstConsumer
Events:         <none>

Kalyans-Mac-mini:google-kubernetes-engine kalyanreddy$ kubectl describe sc standard-rwo
Name:           standard-rwo
IsDefaultClass: Yes
Annotations:   components.gke.io/layer=addon,storageclass.kubernetes.io/is-default-class=true
Provisioner:    pd.csi.storage.gke.io
Parameters:     type=pd-balanced
AllowVolumeExpansion: True
MountOptions:   <none>
ReclaimPolicy: Delete
VolumeBindingMode: WaitForFirstConsumer
Events:         <none>

Kalyans-Mac-mini:google-kubernetes-engine kalyanreddy$
```



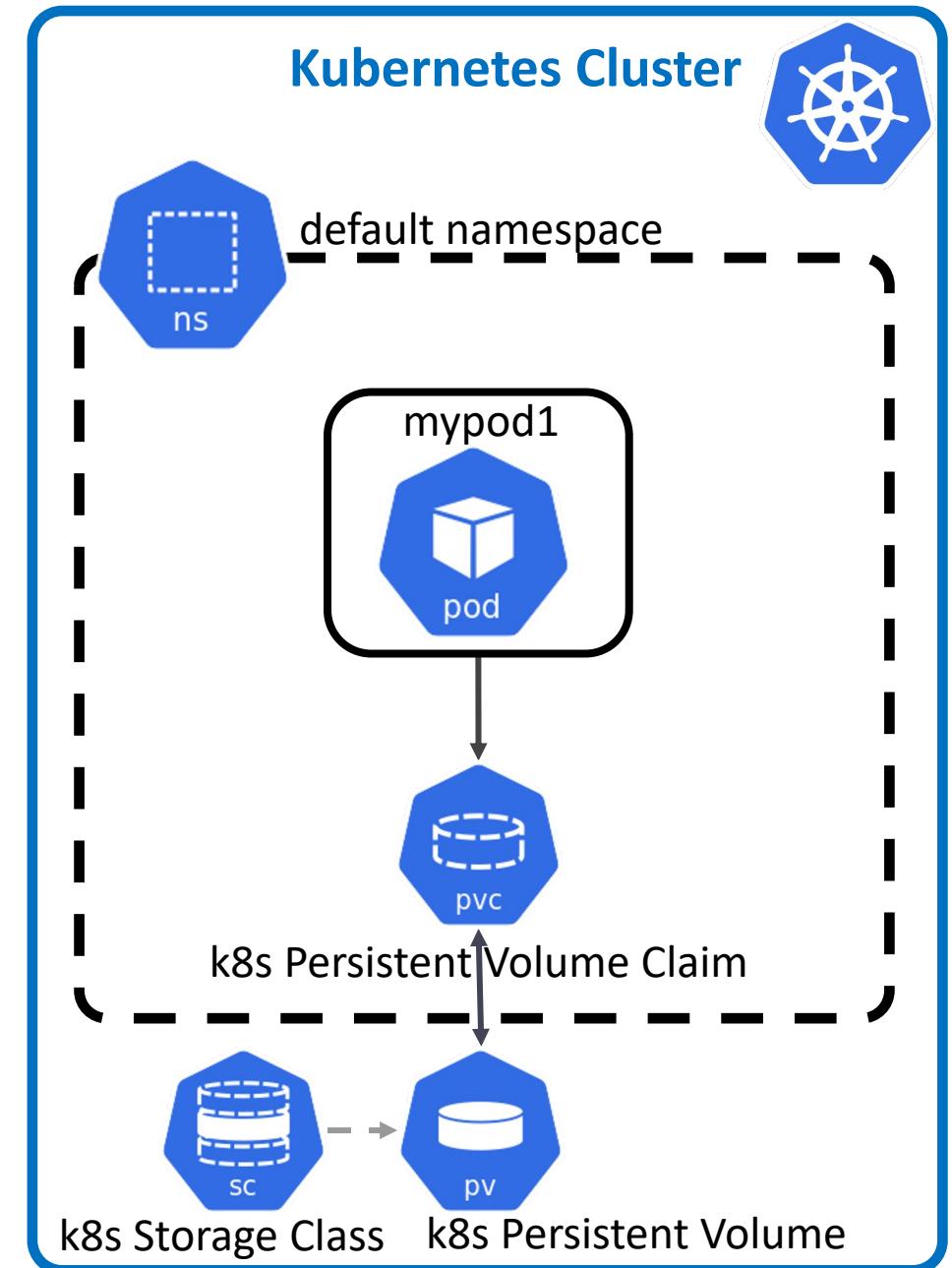
Kubernetes Persistent Volume Claim

- **Persistent Volume Claim (PVC):** PVC objects request a **specific size, access mode, and StorageClass** for the PersistentVolume
- If a PersistentVolume that satisfies the request **exists or can be provisioned**, the PVC is bound to that PersistentVolume.
- **Access Modes**
 - **ReadWriteOnce:** The volume can be mounted as **read-write by a single node**.
 - **ReadOnlyMany:** The volume can be mounted **read-only by many nodes**.
 - **ReadWriteMany:** The volume can be mounted as **read-write by many nodes**.
 - PersistentVolume resources that are backed by **Compute Engine persistent disks don't support** this access mode.

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: mypvc1
spec:
  accessModes:
    - ReadWriteOnce
  storageClassName: standard-rwo
  resources:
    requests:
      storage: 1Gi
```

Kubernetes Persistent Volume

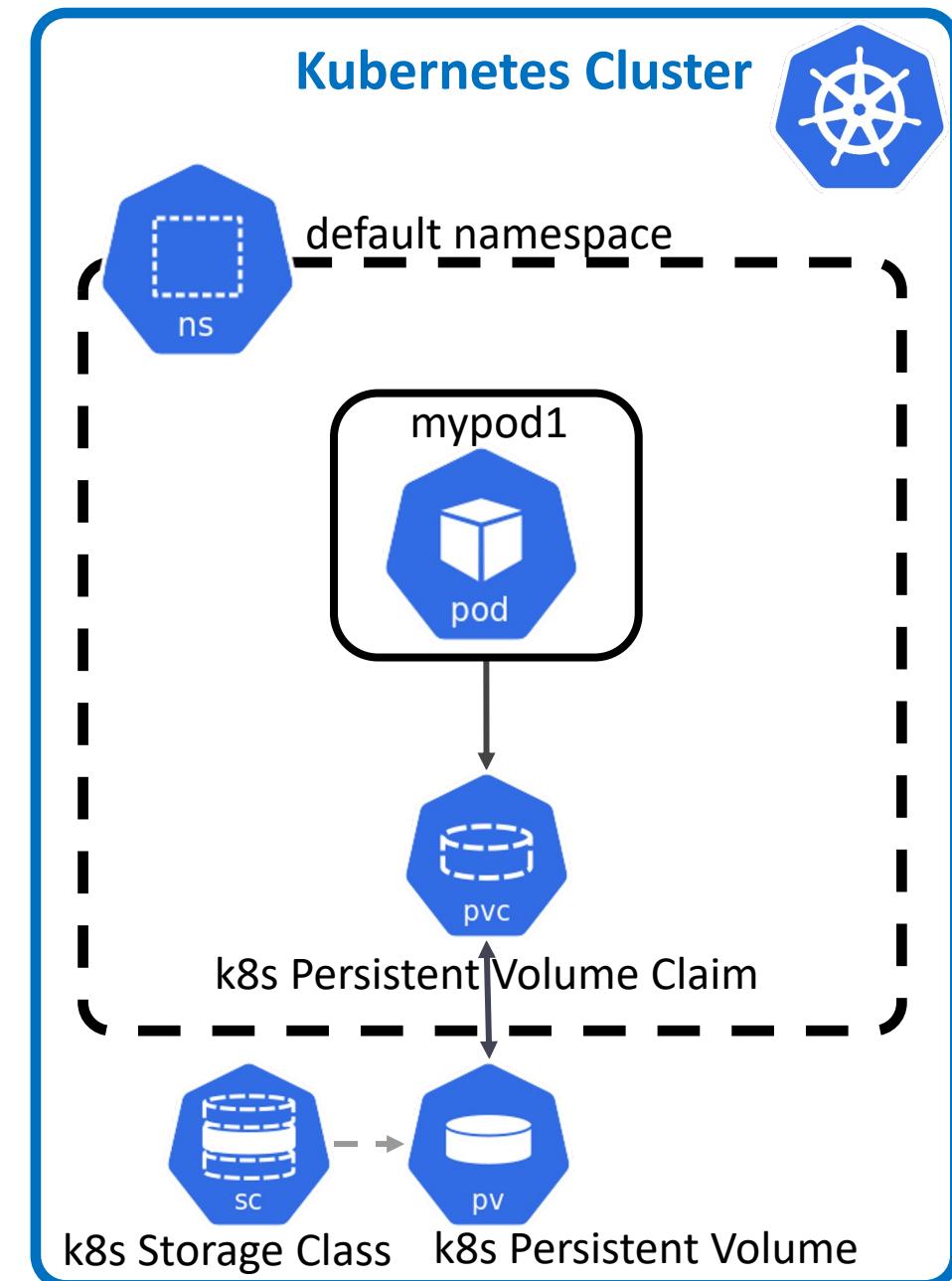
- **Persistent Volume:** In GKE, a `PersistentVolume` is typically backed by a `Google persistent disk`
- A `PersistentVolume` can be `dynamically provisioned`, we do not have to `manually` create and delete the backing storage
- `PersistentVolume` resources are cluster resources that exist `independently` of Pods.
 - Doesn't impact due to `cluster changes`
 - Doesn't impact due to `pods deleted or recreated`
- Persistent Volume will be `provisioned based on configurations` defined in Storage Class and Persistent Volume Claim
 - **Storage Class:** Storage type (`pd-balanced`, `pd-ssd`)
 - **PVC:**
 - **Size:** 1GB, 2GB
 - **Access Modes:** `ReadWriteOnce`, `ReadOnlyMany`, `ReadWriteMany`



Kubernetes Persistent Volume

- **Resource Types [IMPORTANT]**

- Storage Class and Persistent Volume or **cluster level resources**
- Persistent Volume Claim (PVC) is a **namespace level** resource



GKE Cluster - Default Kubernetes Storage Classes



Persistent Disks
Default Storage Class

standard
kubernetes.io/gce-pd

Uses In-Tree Provisioner - Old

NOT RECOMMENDED

standard-rwo
(Balanced Disk)

Uses **CSI Provisioner** (Container Storage Interface – GKE PD CSI) – LATEST & GREATEST

RECOMMENDED

premium-rwo
(SSD Disk)

NAME	PROVISIONER	RECLAIMPOLICY	VOLUMEBINDINGMODE	ALLOWVOLUMEEXPANSION	AGE
premium-rwo	pd.csi.storage.gke.io	Delete	WaitForFirstConsumer	true	27h
standard	kubernetes.io/gce-pd	Delete	Immediate	true	27h
standard-rwo (default)	pd.csi.storage.gke.io	Delete	WaitForFirstConsumer	true	27h

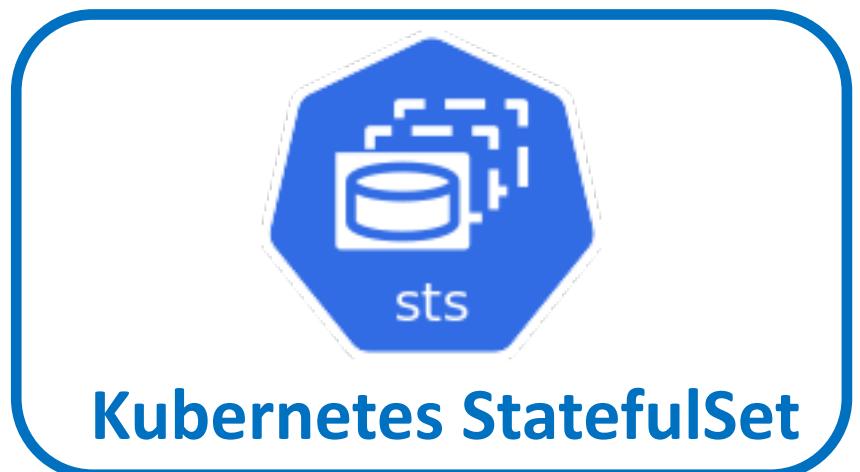
Cluster Feature
should be enabled

Compute Engine persistent disk CSI Driver

Enabled

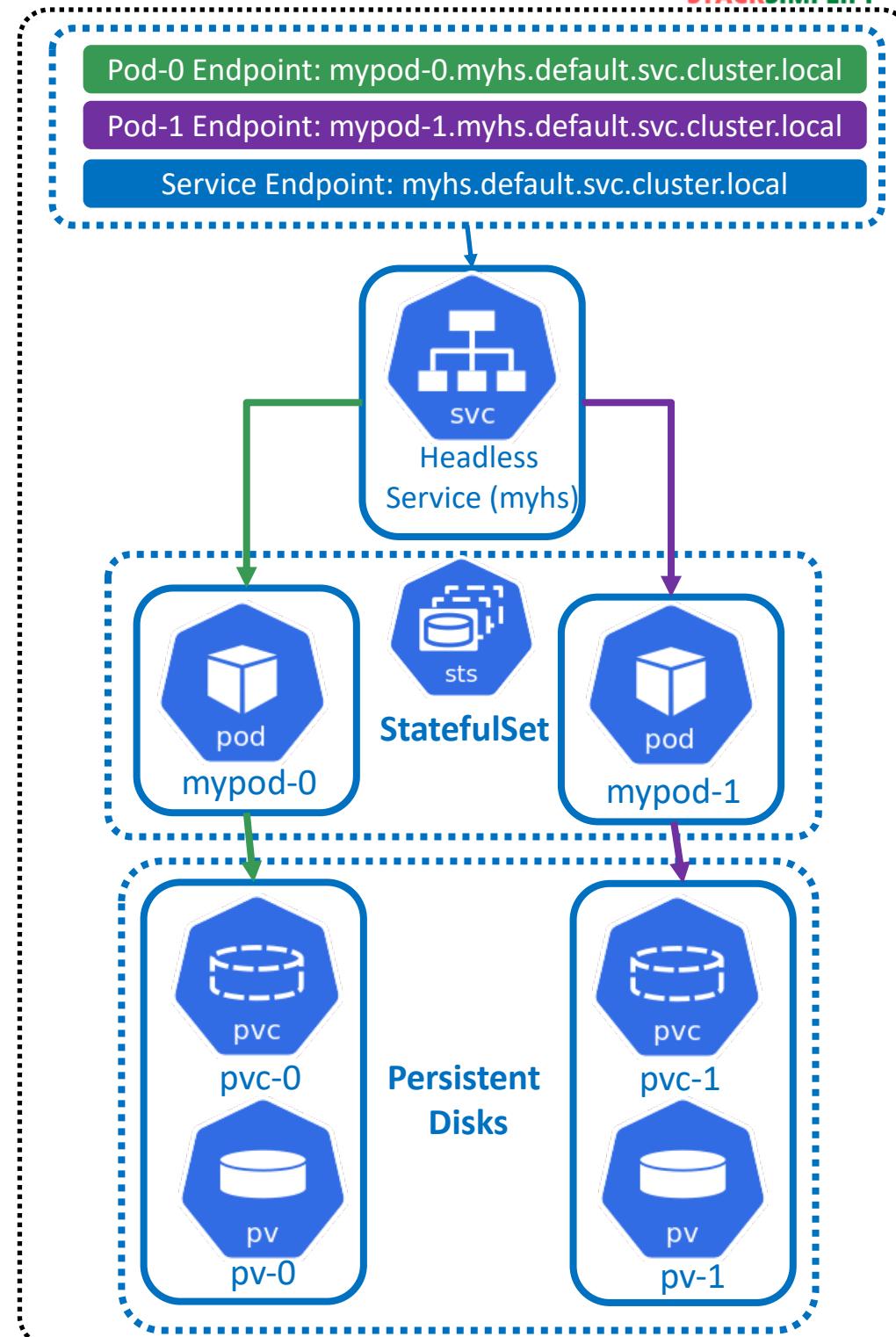


Kubernetes StatefulSets



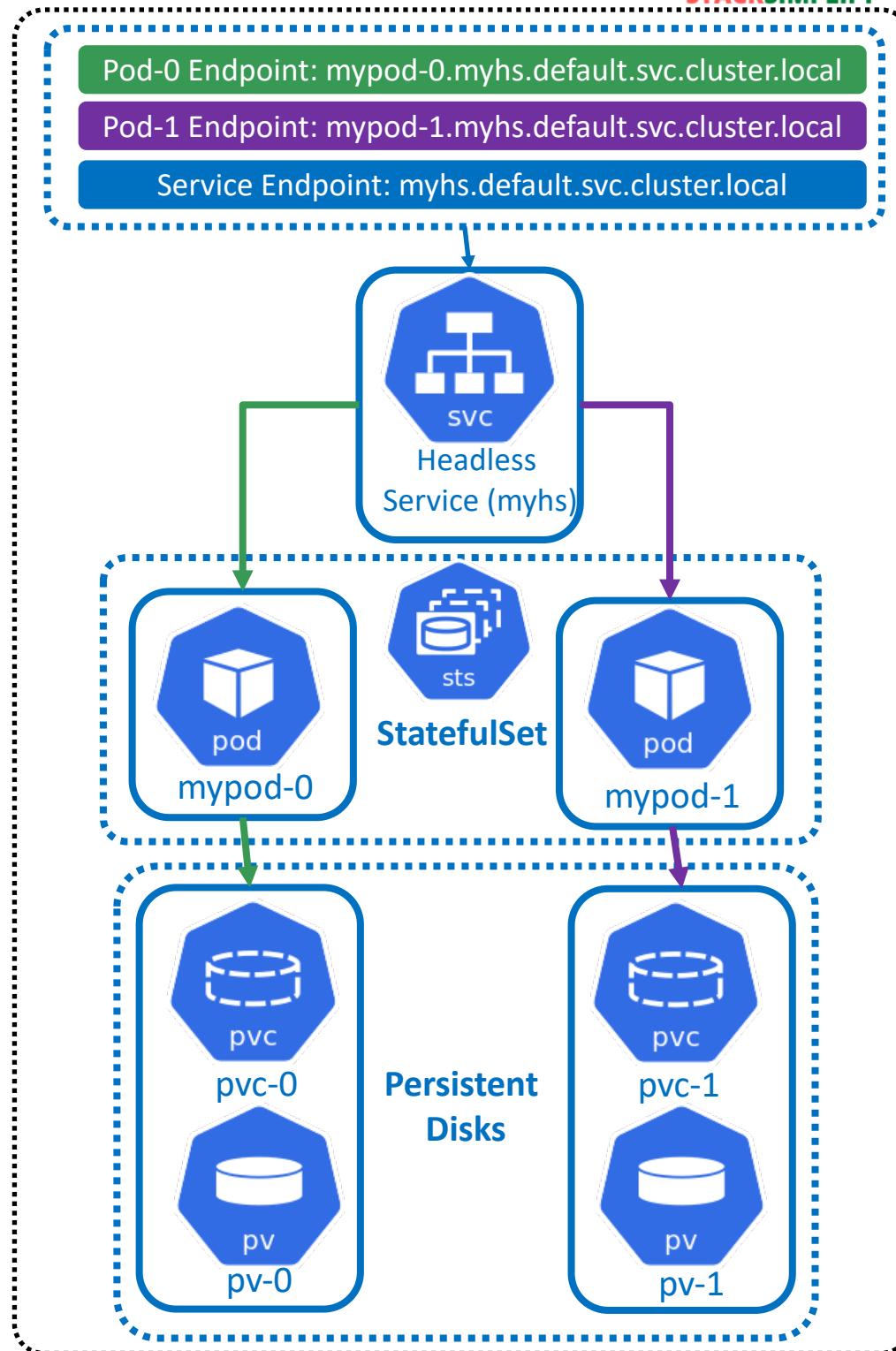
Kubernetes StatefulSet

- **StatefulSets:** Represent a set of Pods with **unique, persistent identities, and stable hostnames** that Kubernetes maintains regardless of where they are scheduled
- The **state information** and **other resilient data** for any given **StatefulSet Pod** is maintained in **persistent volumes** associated with each Pod in the StatefulSet
- StatefulSet Pods can be restarted at any time.
- When restarted or recreated, always it creates the **pod with same name** and attaches the same **persistent disk**
- **For Stateful Applications:** StatefulSets
- **For Stateless Applications:** Deployments



Kubernetes StatefulSet

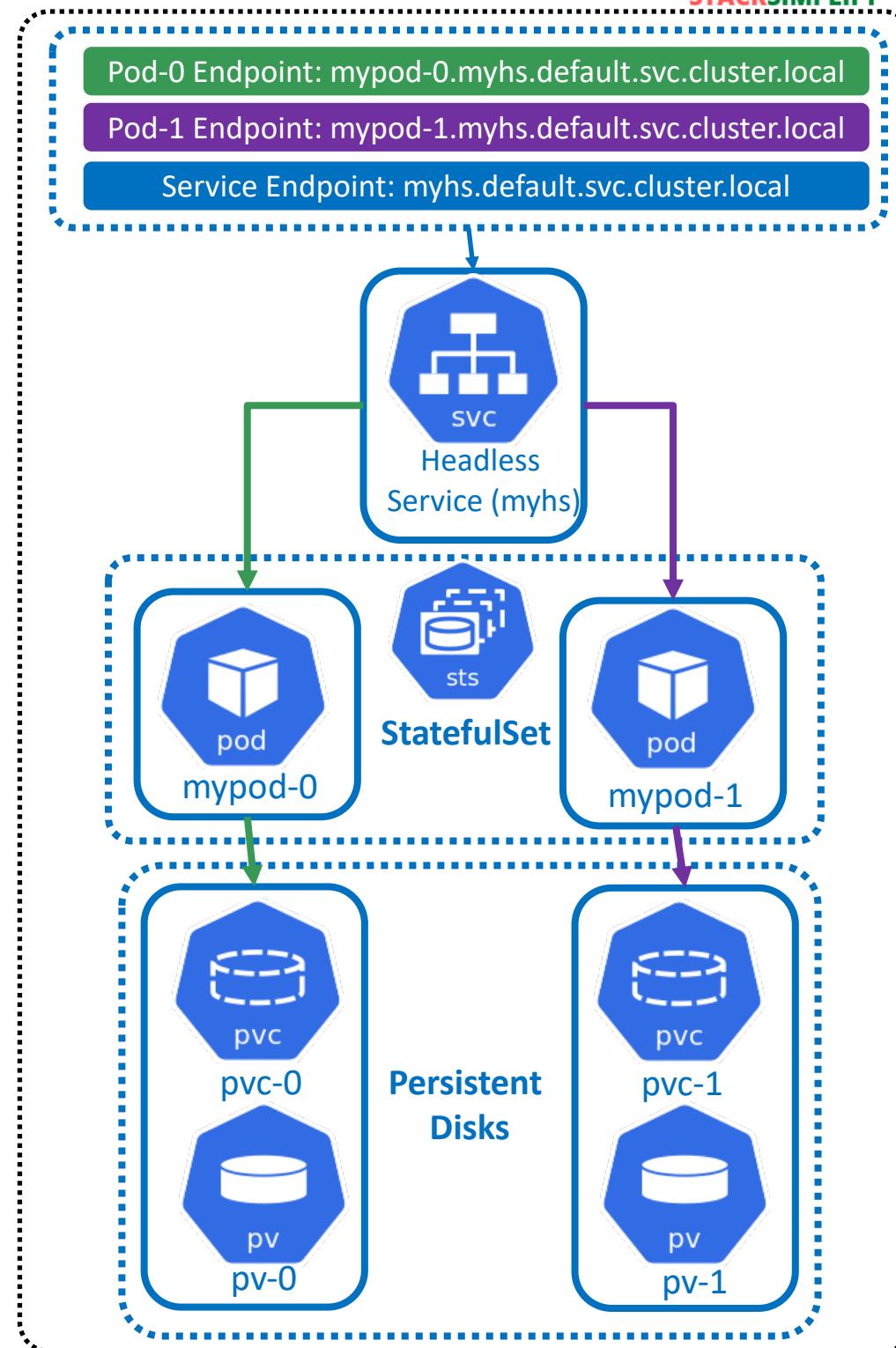
- StatefulSets require a Headless service
 - Each pod will have a **dedicated headless service endpoint** with pod id
 - **Pod-0 Endpoint:** `mypod-0.myhs.default.svc.cluster.local`
 - **Pod-1 Endpoint:** `mypod-1.myhs.default.svc.cluster.local`
 - Headless service endpoints load balance traffic to all pods
 - **Service Endpoint:** `myhs.default.svc.cluster.local`
- StatefulSets are used for Applications that require
 - Stable, unique network identifiers.
 - Stable, persistent storage.
 - Ordered, graceful deployment and scaling.
 - Ordered, automated rolling updates.



Kubernetes StatefulSet

- **Usecases**

- MySQL Database
 - **Pod-0:** Master server (accepts reads and writes)
 - **Pod-1:** Replica Server (accepts reads only)
 - **Pod-2:** Replica Server (accepts reads only)
- PostgreSQL
- Elasticsearch
- Kafka
- Redis
- Cassandra
- Zookeeper



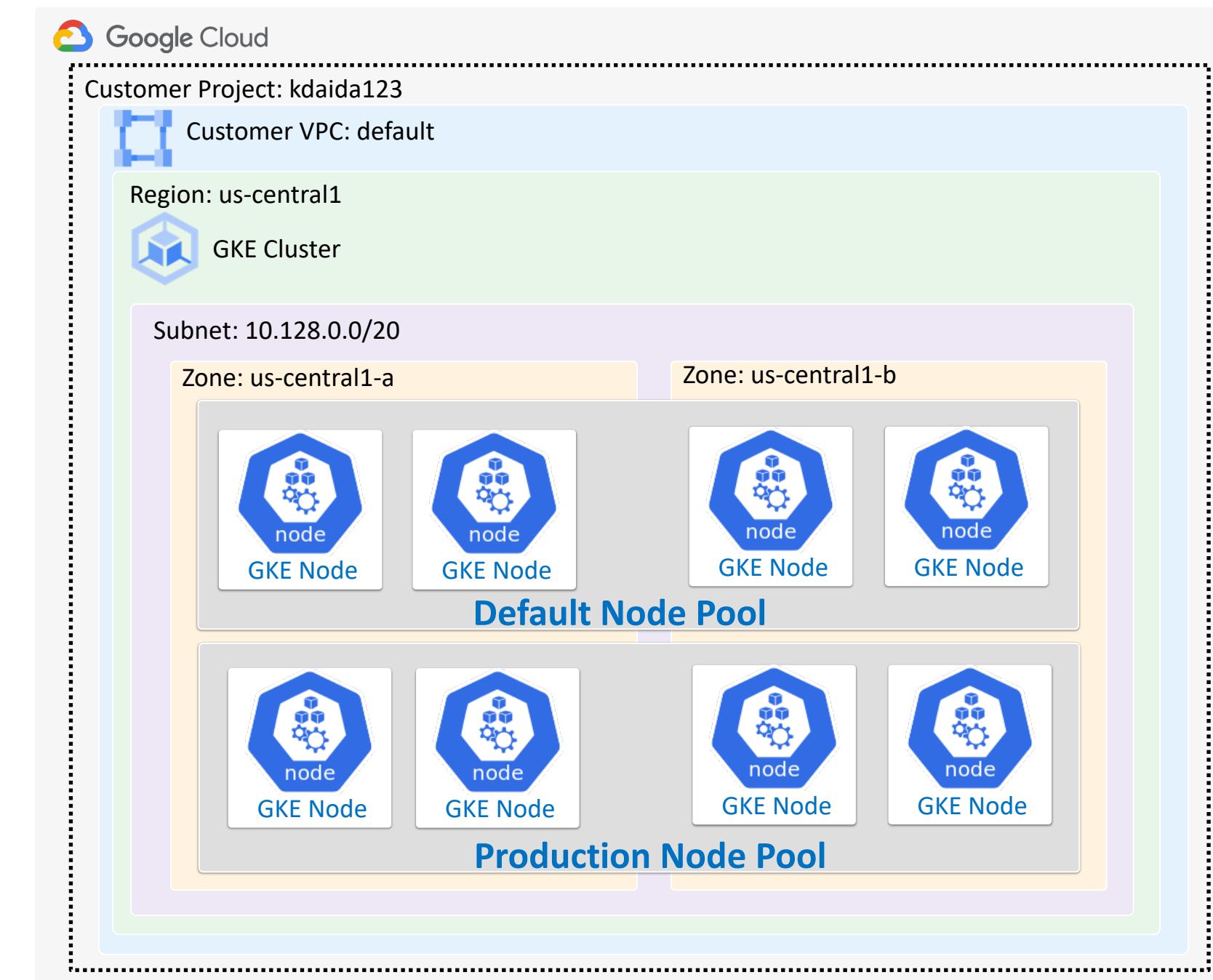


Google Kubernetes Engine Cluster Autoscaler



Kubernetes Cluster Autoscaler

- **Cluster Autoscaler:** **Automatically resizes** the number of nodes in a node pool, based on the **demands of your workloads**.
 - Minimum Size: 3
 - Maximum Size: 12
- **Scale-Out:** **Gradually increases** nodes to maximum size based on need (When pods are unschedulable)
- **Scale-In:** **Decreases nodes to minimum size** when nodes are idle (no workloads scheduled on them)



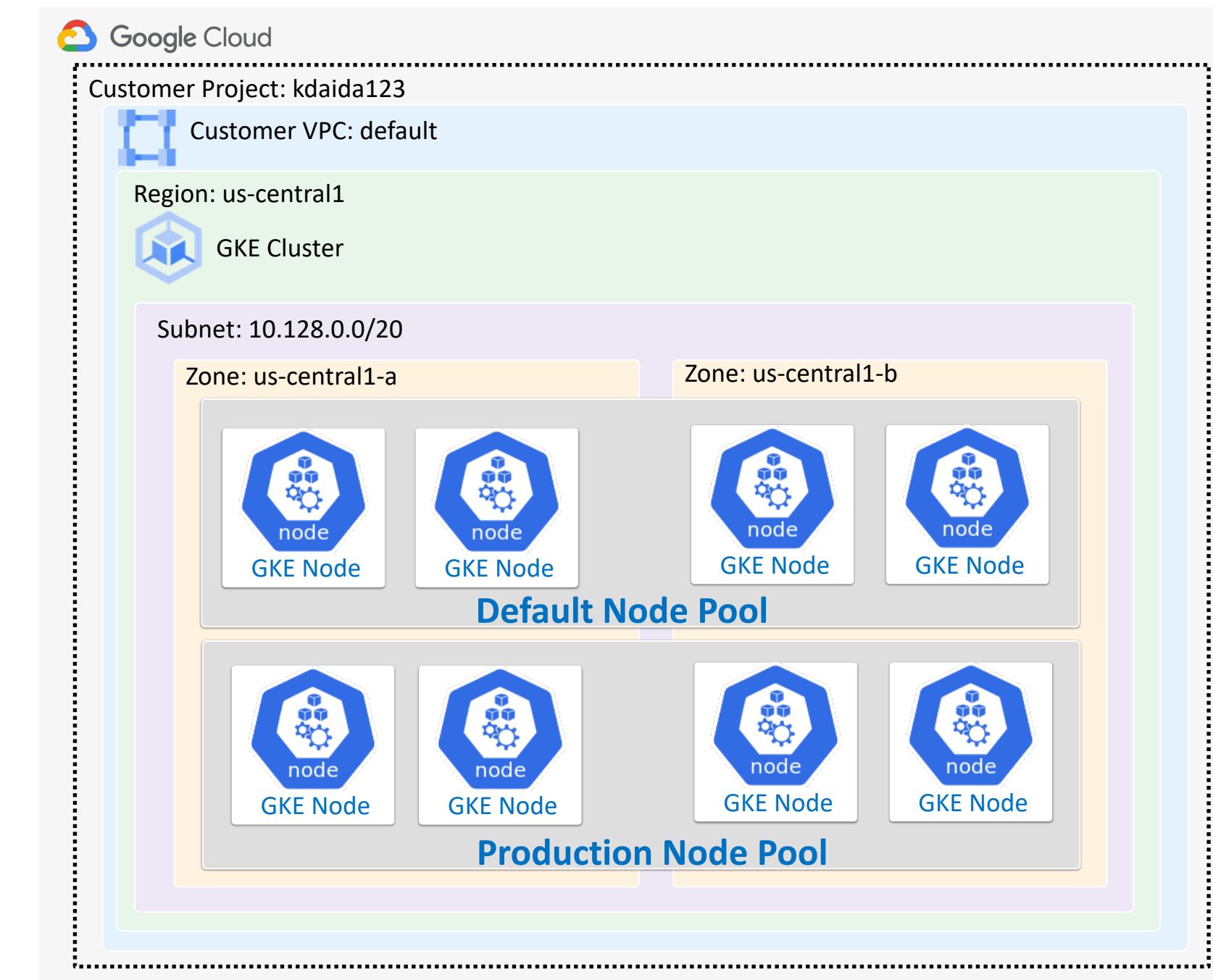
Kubernetes Cluster Autoscaler

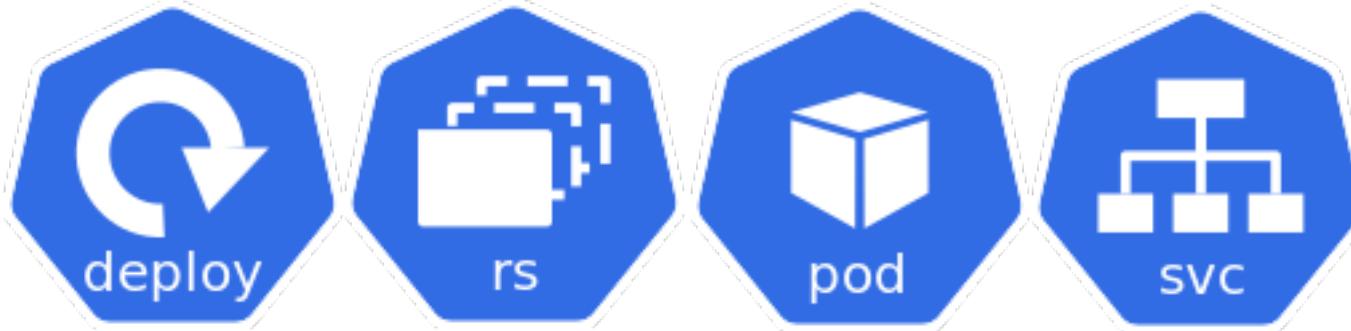
- Benefits

- Increases the **availability** of your workloads
- **Control** the costs
- **Effective** use of system resources (CPU, memory)

- Standard vs Autopilot Clusters

- Standard Cluster uses **Cluster Autoscaler**
- Autopilot cluster uses **Node auto-provisioning**
 - GKE automatically takes care of nodes and manage node pools.





Google Kubernetes Engine Horizontal Pod Autoscaling



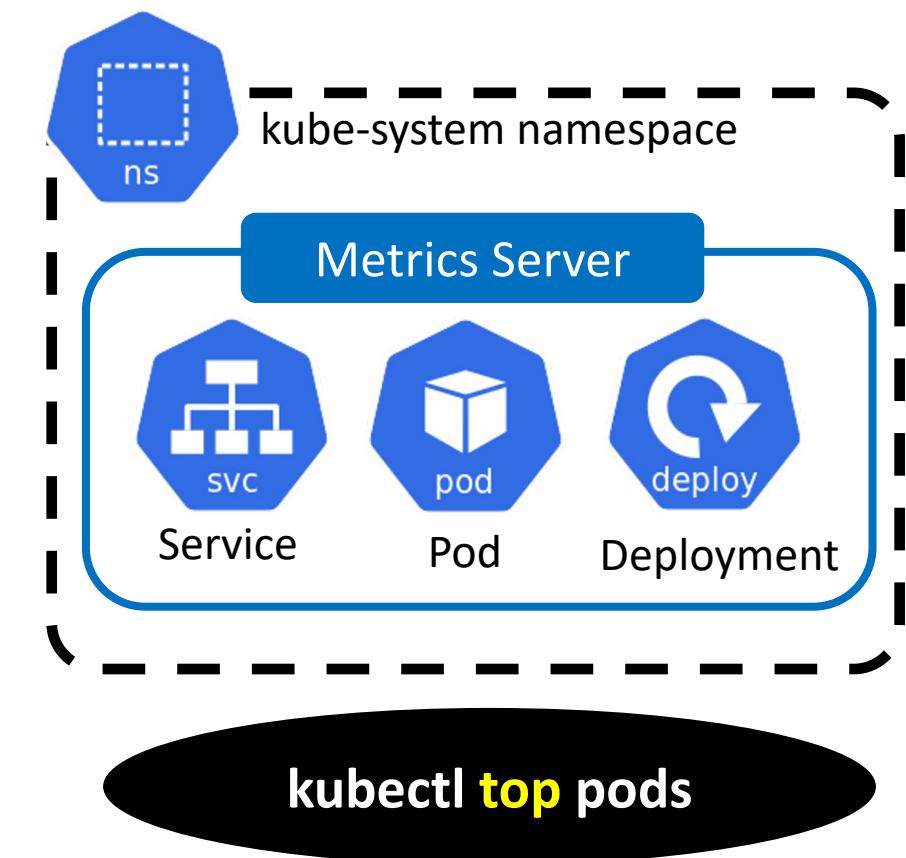
Kubernetes Horizontal Pod Autoscaling

- Automatically increase or decrease number of pods in response to
 - Workloads CPU and Memory Utilization
 - Custom metrics reported from within Kubernetes Cluster
 - External metrics (Load Balancers, Messaging Services ...)
 - Custom metrics using Managed Service for Prometheus
- HPA automatically scales the pods in workload types
 - Kubernetes ReplicaSet
 - Kubernetes Replication Controller
 - Kubernetes Deployment
 - Kubernetes StatefulSet
- This can help our applications
 - Scale out to meet increased demand or
 - Scale in when resources are not needed, thus freeing up your worker nodes for other applications

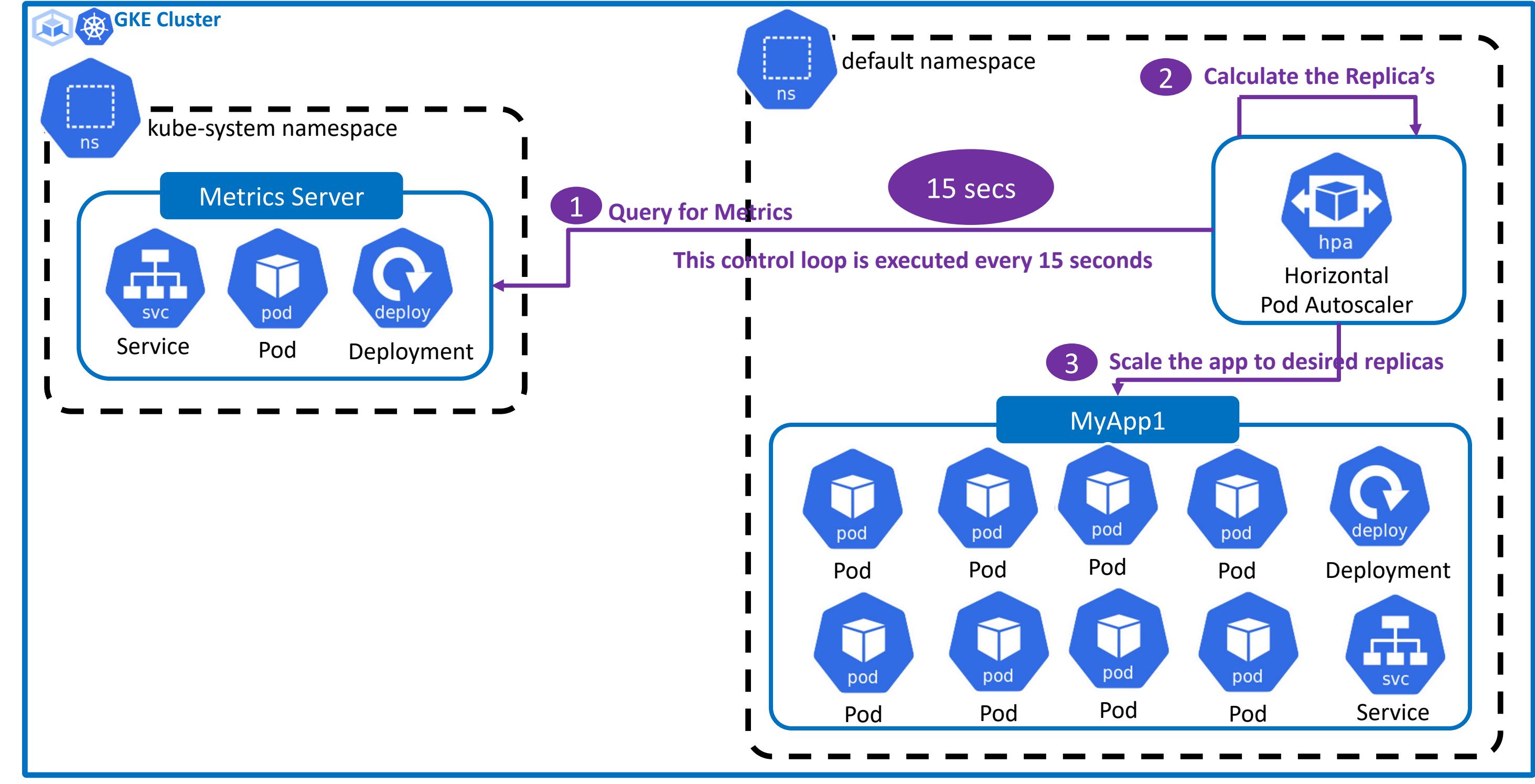
```
HPA Imperative Command: kubectl autoscale deployment my-app --max 6 --min 4 --cpu-percent 50
```

Kubernetes Metrics Server

- Metrics Server collects resource metrics from [Kubelets](#) and [exposes them in Kubernetes apiserver](#) through [Metrics API](#)
- Metrics API can also be accessed by [kubectl top](#), making it easier to debug autoscaling pipelines.
- Fast Autoscaling solution, [collecting metrics every 15 seconds](#)
- Metrics Server is not meant for [non-autoscaling purposes](#). For example, don't use it to forward metrics to monitoring solutions.
- Resource efficiency, [using 1 mili core of CPU and 2 MB of memory](#) for each node in a cluster
- Metrics Server used for [CPU/Memory](#) based [horizontal autoscaling](#)
- Metrics Server used for [automatically adjusting/suggesting](#) resources needed by containers ([Vertical Autoscaling](#))



Kubernetes Horizontal Pod Autoscaling



Kubernetes Horizontal Pod Autoscaling

Autoscaling v1

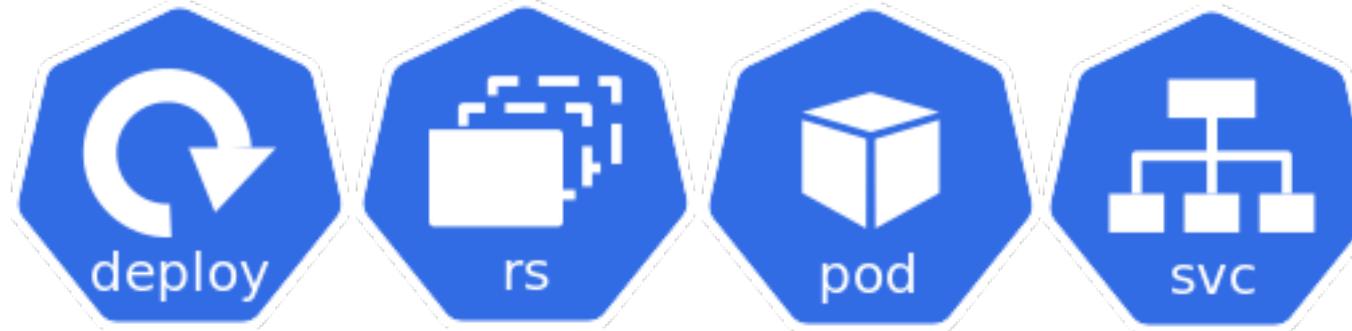
```
apiVersion: autoscaling/v1
kind: HorizontalPodAutoscaler
metadata:
  name: hpa-myapp1
spec:
  scaleTargetRef:
    apiVersion: apps/v1
    kind: Deployment
    name: myapp1-deployment
  minReplicas: 1
  maxReplicas: 10
  targetCPUUtilizationPercentage: 50
```

Autoscaling v2

```
apiVersion: autoscaling/v2
kind: HorizontalPodAutoscaler
metadata:
  name: cpu
spec:
  scaleTargetRef:
    apiVersion: apps/v1
    kind: Deployment
    name: myapp1-deployment
  minReplicas: 1
  maxReplicas: 10
  metrics:
  - type: Resource
    resource:
      name: cpu
      target:
        type: Utilization
        averageUtilization: 30
```



Horizontal
Pod Autoscaler

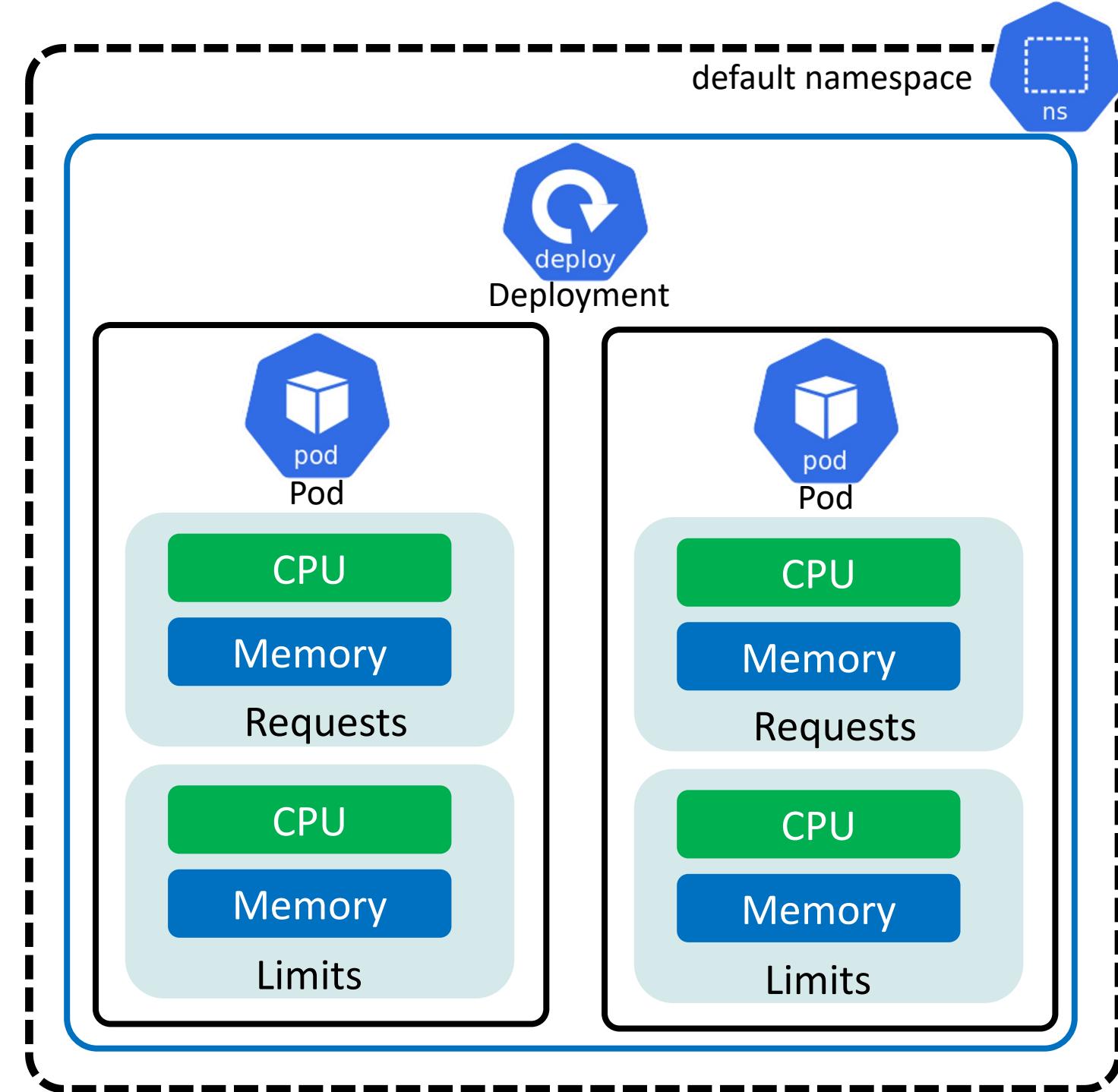


Google Kubernetes Engine Vertical Pod Autoscaling VPA



Kubernetes Vertical Pod Autoscaling

- **Vertical Pod autoscaling:** Lets you **analyze and set** CPU and memory resources required by Pods
- **Manual**
 - VPA **recommends** cpu and memory requests and limits
 - We can review the recommendation and update the values **manually**
- **Automatic**
 - VPA **recommends and automatically updates** the cpu and memory requests and limits
 - VPA **evicts and recreates** the pod during the resource request updates



Kubernetes Vertical Pod Autoscaling

- **Standard vs Autopilot clusters**

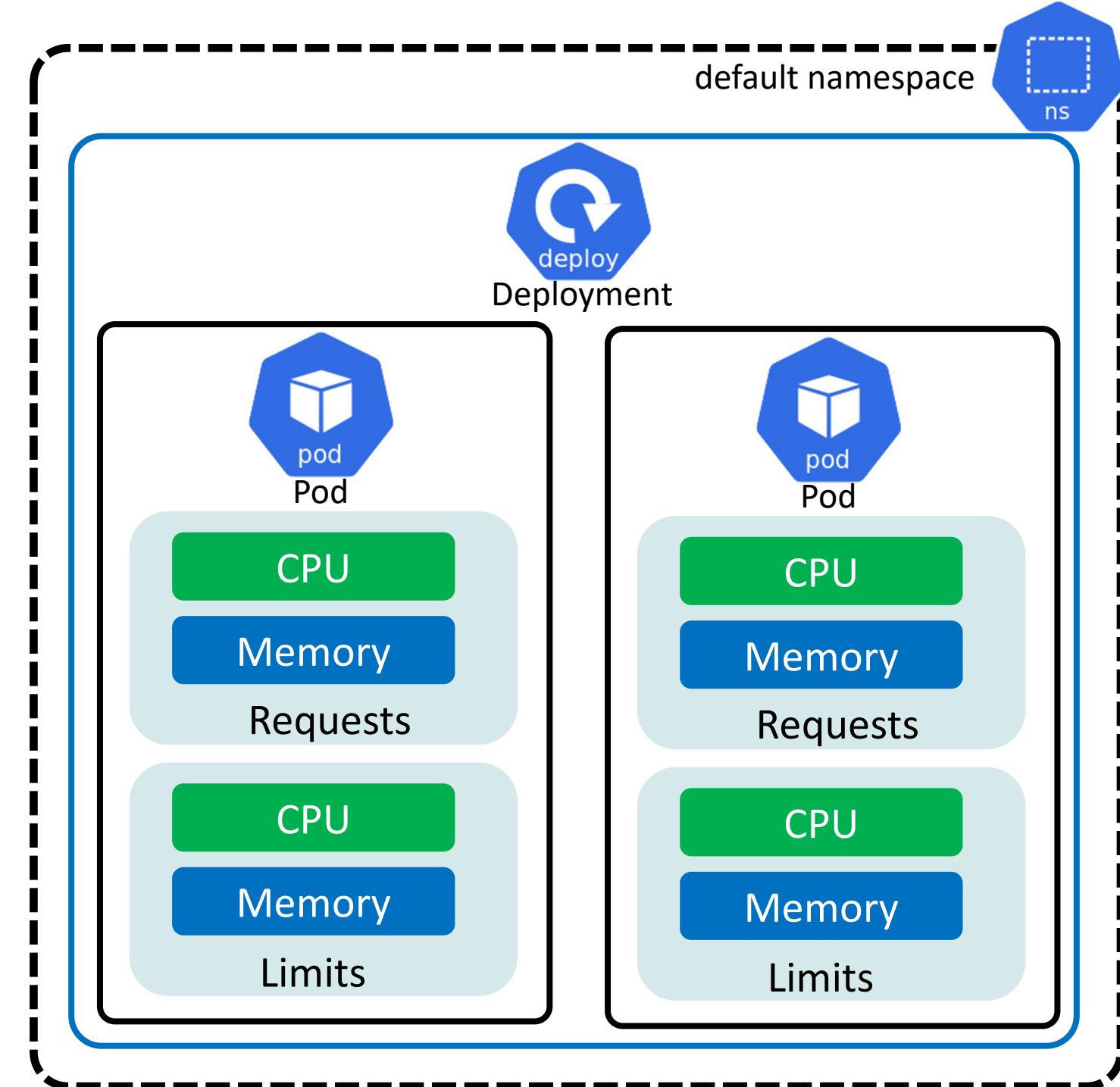
- VPA need to be enabled at [workload level](#) for standard clusters
- VPA is [by default enabled](#) in Autopilot clusters

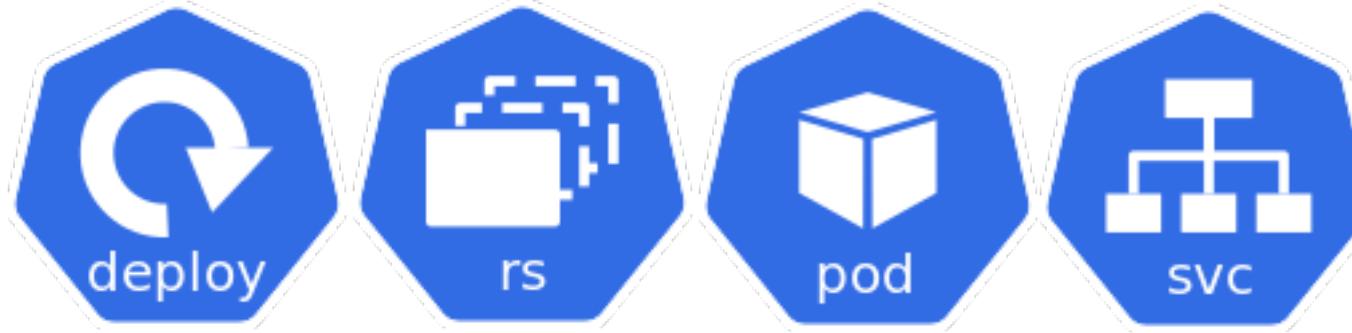
- **Benefits**

- Cluster nodes are used [efficiently](#) because Pods use exactly what they need
- You don't have to run time-consuming [benchmarking tasks to determine](#) the correct values for CPU and memory requests
- Pods are scheduled onto nodes that have the [appropriate resources available](#)
- Runs Vertical Pod Autoscaler Pods as [control plane processes](#) instead of Deployment on worker nodes ([GKE special feature](#))

- **Important Note:**

- [Enable Cluster Autoscaler](#) before enabling VPA
- VPA [can talk to Cluster Autoscaler](#) to increase the number of nodes if needed for the VPA enabled workloads





Google Kubernetes Engine Artifact Registry



What is Google Artifact Registry ?

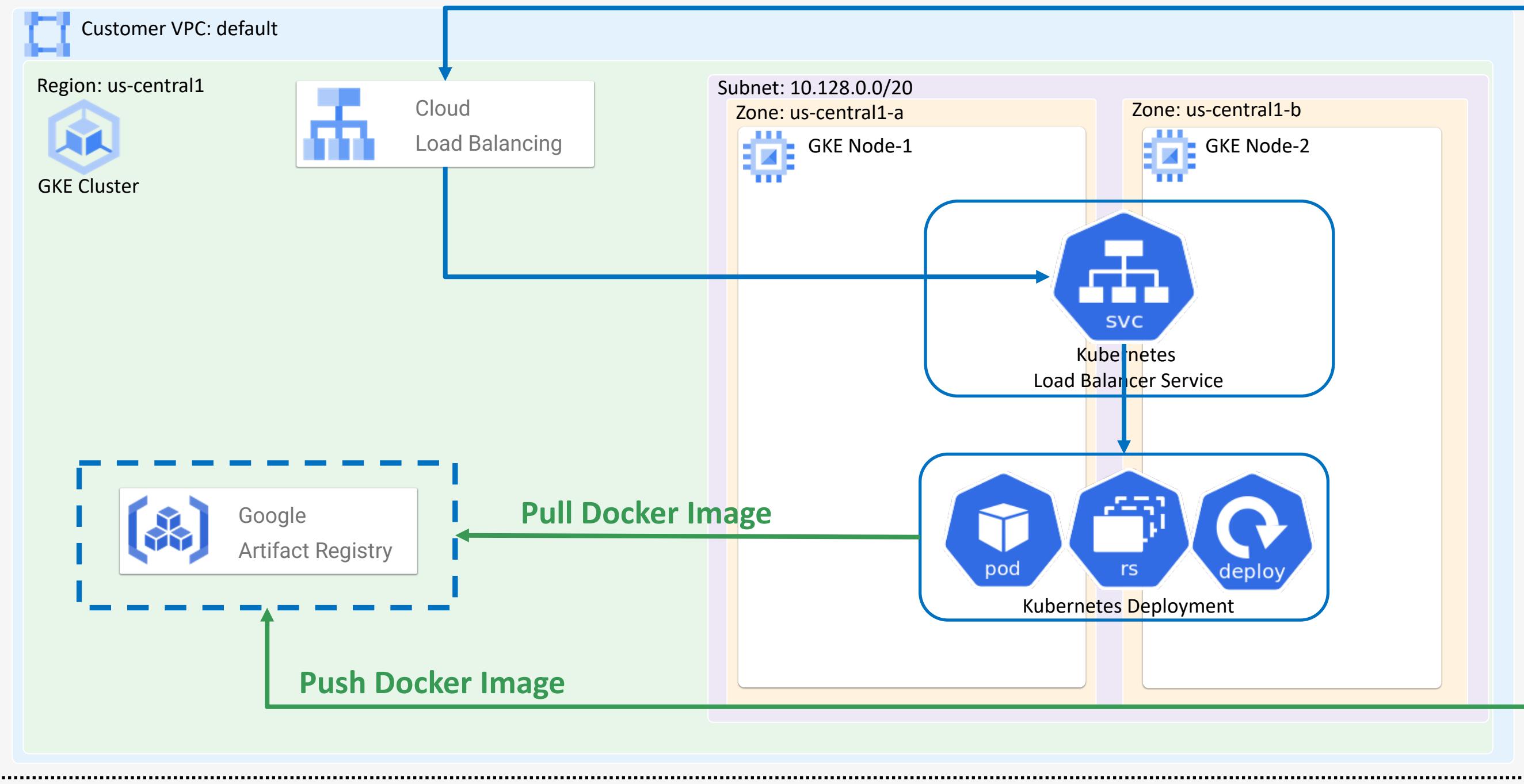
- Artifact Registry provides a **single location** for storing and managing your **packages** and **Docker** container images
- We can call it as **Universal Build Artifact Management Service**
 - Manage Docker Container Images
 - Manage Packages (Maven and npm)
- Google **Container Registry** is **very old service** and **not recommended**
- Google **Artifact** Registry expands on the capabilities of Google Container Registry and **recommended** for use going forward.
- Supports **Regional and Multi-Regional** Repositories
- **Additional Reference:** <https://cloud.google.com/artifact-registry>

What are we going to learn ?

- Step-01: Build a Docker Image on local desktop or Cloud Shell
- Step-02: Run Docker Image and Test it on local desktop
- Step-03: Create Google Artifact Repository
- Step-04: Configure Google Artifact Repository Authentication on local desktop or Cloud Shell
- Step-05: Tag and Push Docker Image to Google Artifact Repository
- Step-06: Verify Docker Image on Google Artifact Repository
- Step-07: Update Docker Image from Google Artifact Repository in Kubernetes Deployment
- Step-08: Deploy Kubernetes Manifests and Verify if Docker image successfully pulled from Google Artifact Registry



Users

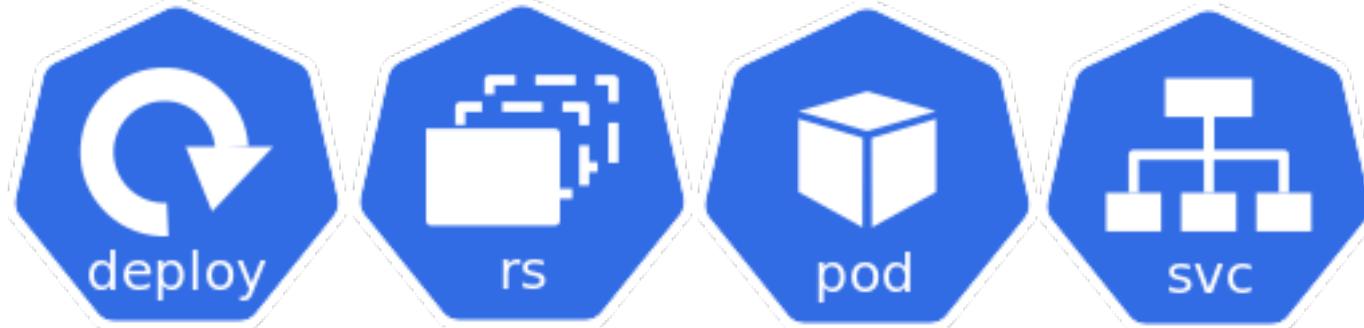
<http://<Load-Balancer-IP-Address>>

NETWORK DESIGN

Google Artifact Registry



Admin

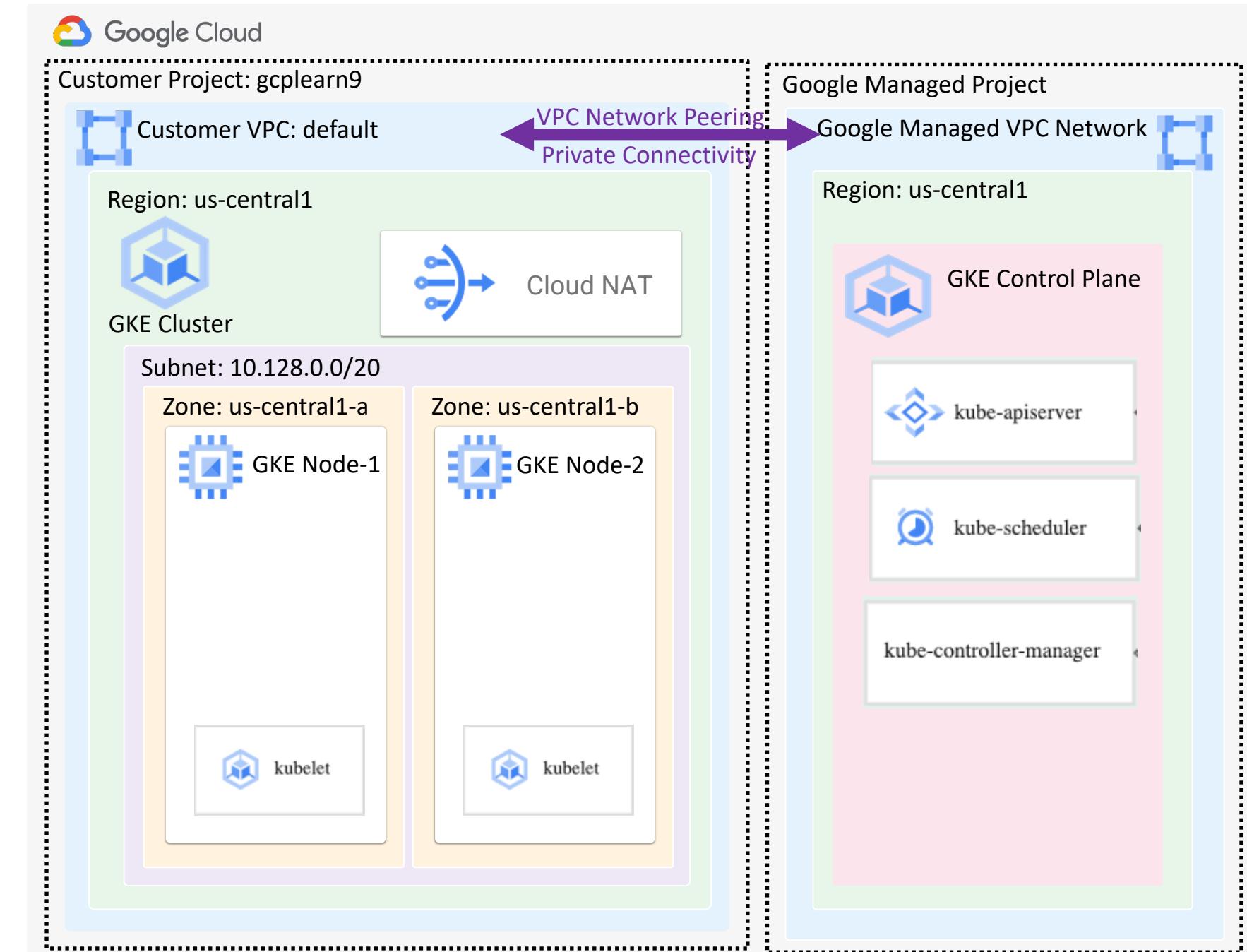


Google Kubernetes Engine Private Cluster (Type: Standard)



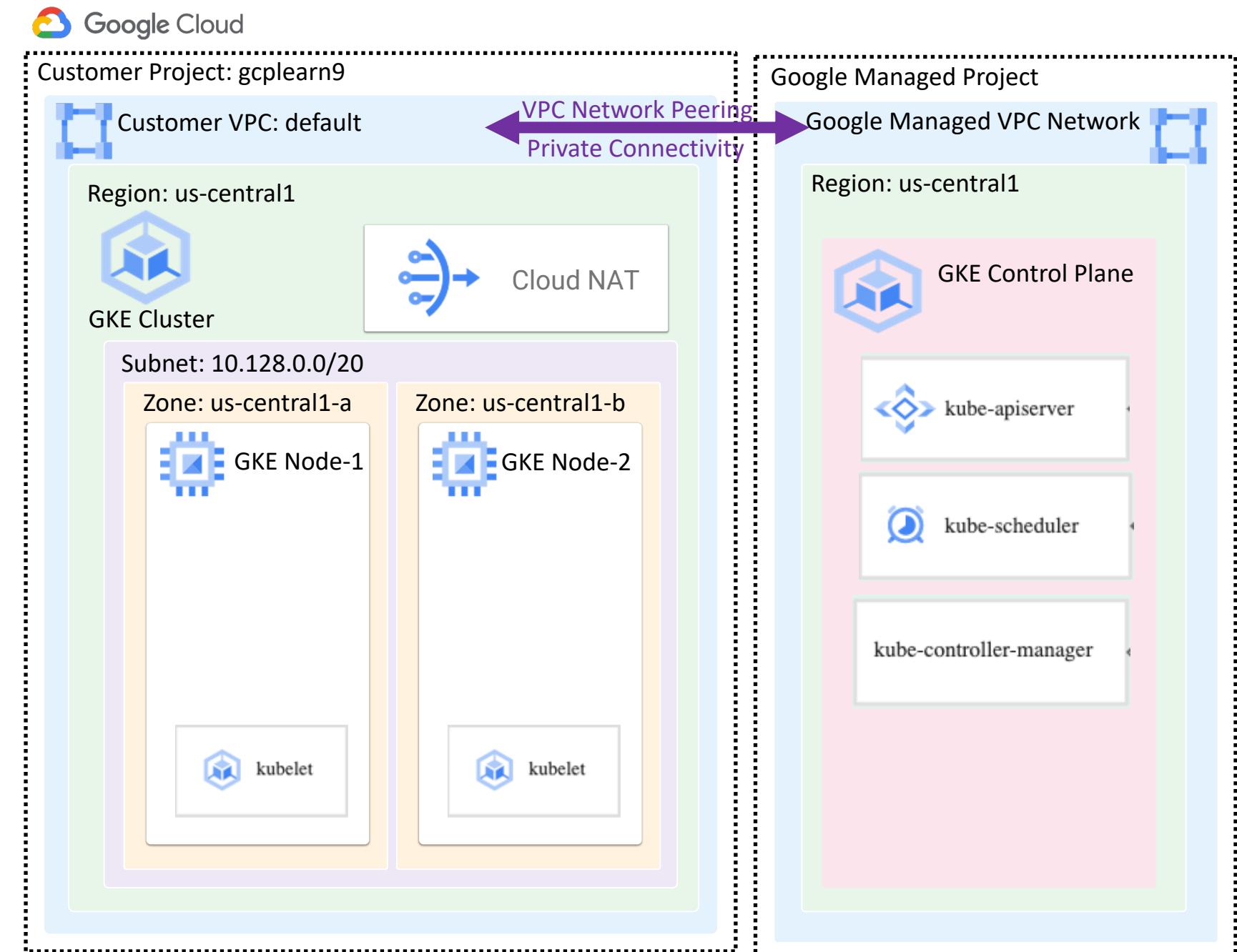
GKE Private Cluster - Network Design

- **Private Clusters:** Control Plane VPC network is connected to our VPC network using **VPC Network peering**
- Our VPC network contains **GKE nodes**
- Google managed VPC network contains cluster **Control plane**
- Traffic between nodes and control plane is routed using **internal IP addresses only**



GKE Private Clusters

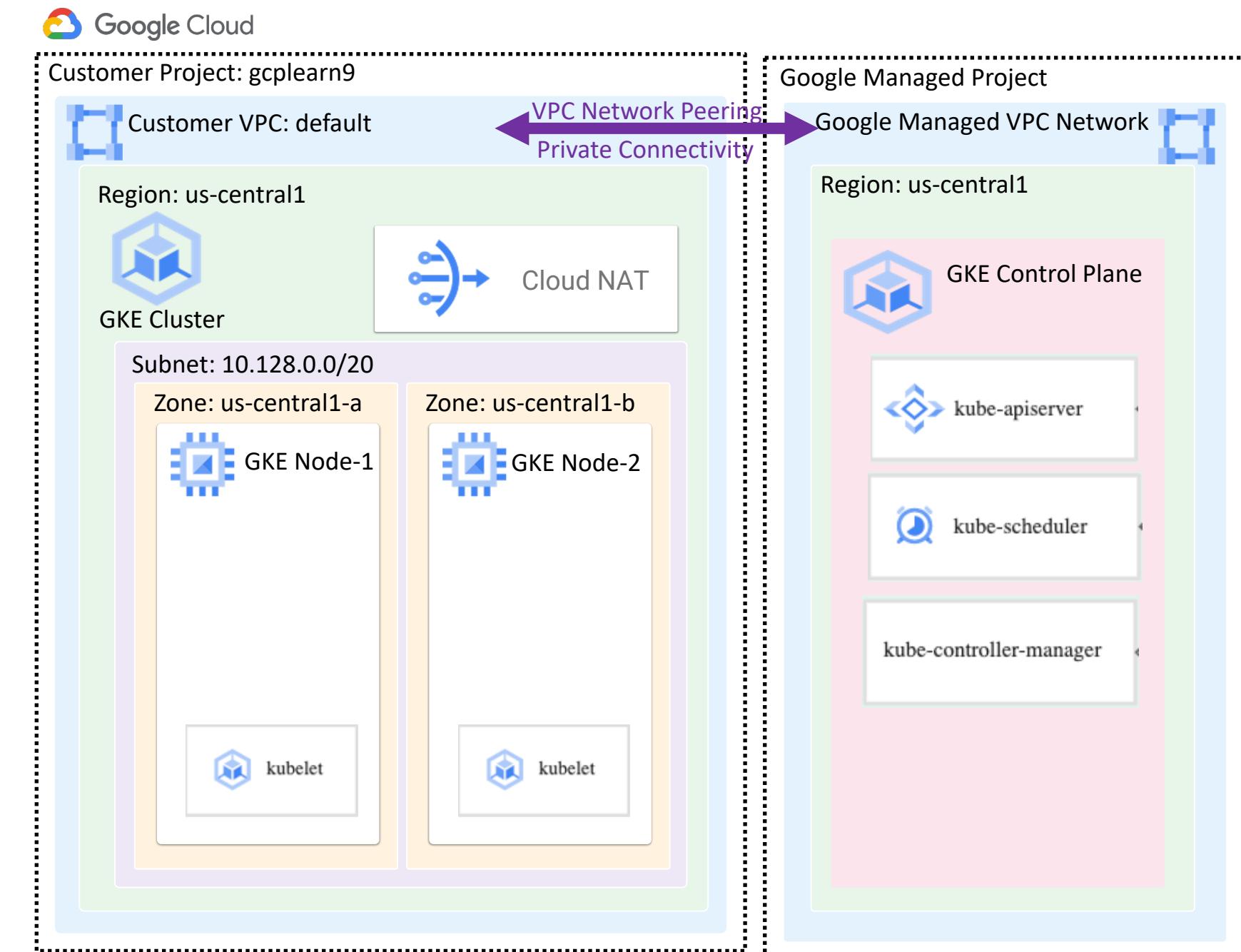
- You can create and configure private clusters in **Standard or Autopilot**.
- Private clusters will have nodes that **do not have** external IP addresses
- If you want to provide outbound internet access for certain private nodes, you can use **Cloud NAT**.



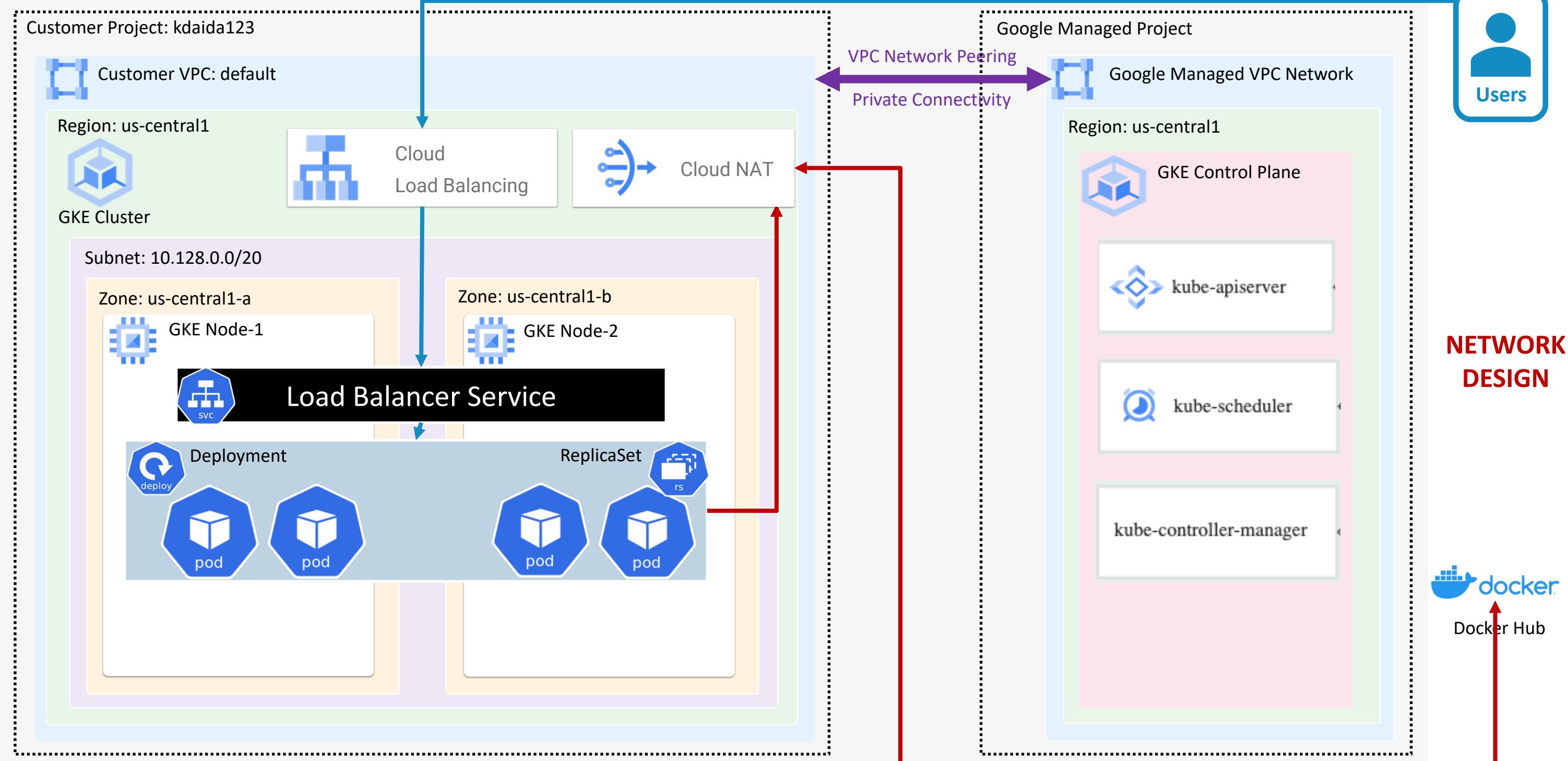
<https://cloud.google.com/kubernetes-engine/docs/concepts/alias-ips#benefits>

GKE Private Clusters

- In private clusters, **private google access enabled by default on VPC subnet** to access other Google cloud APIs and services using private network
 - **Example-1:** To access container images from [Artifact Registry](#)
 - **Example-2:** To send logs to [Cloud Logging](#)



GKE Private Cluster - Pull Docker Image from Docker Hub



GKE Private Cluster - Access using kubectl

- **Least secure Option**

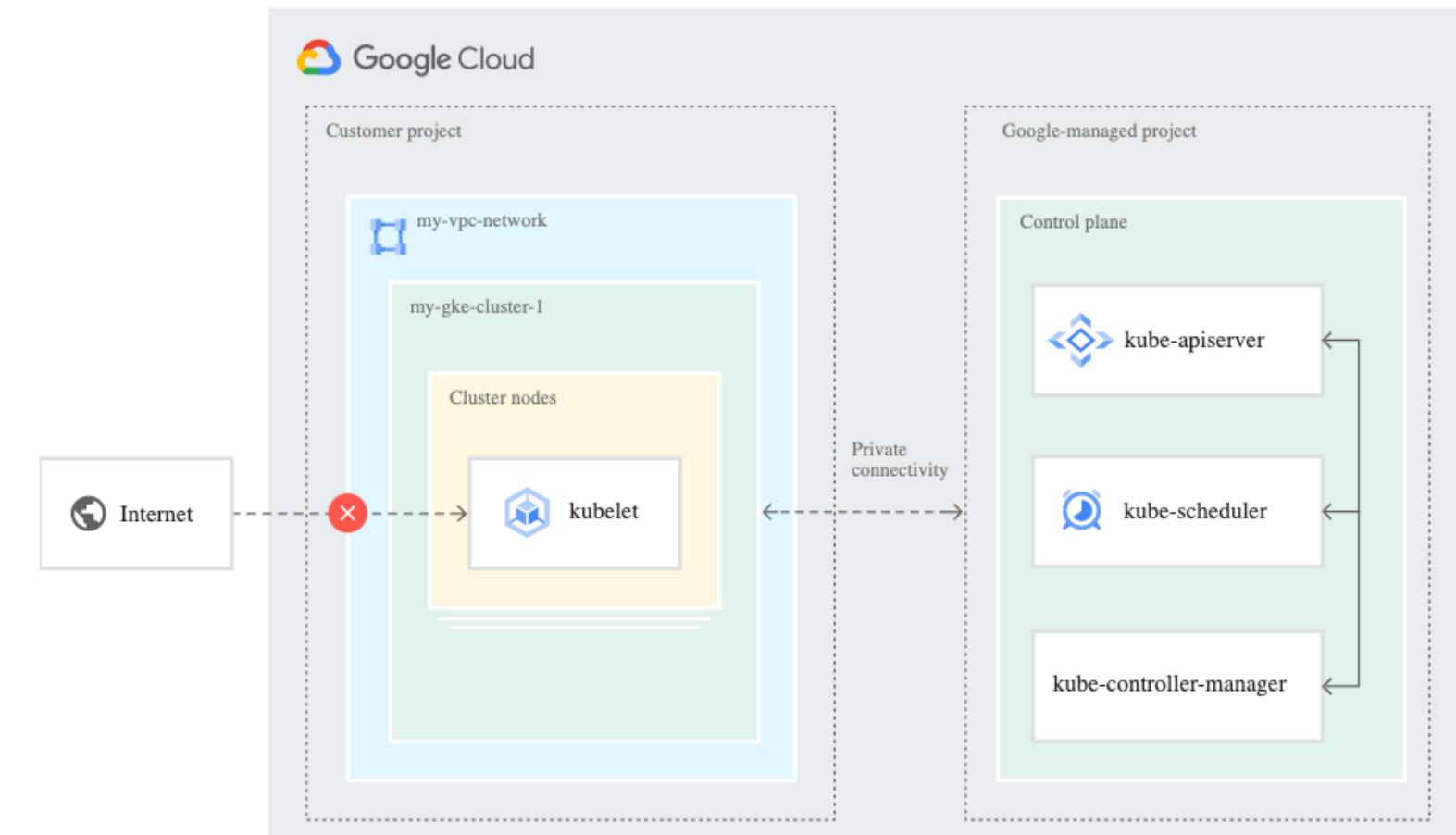
- Public endpoint access: **Enabled**
- Authorized Networks: **Disabled**
- Accessible via **Internet**

- **Medium secure option**

- Public endpoint access: **Enabled**
- Authorized Networks: **Enabled**
- Accessible via **authorized internet IP ranges** (Example: CloudShell, local desktop)

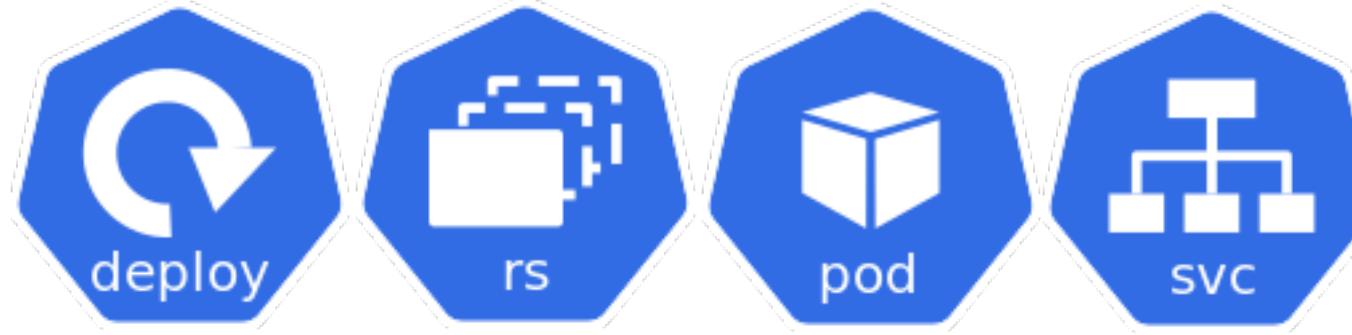
- **High secure option**

- Public endpoint access: **Disabled**
- Accessible via
 - VM in google cloud VPC network
 - **On-premise network** provided Cloud VPN or Cloud Interconnect is configured





Google Kubernetes Engine Autopilot Cluster



Autopilot mode

Optimized Kubernetes cluster with a hands-off experience

[CONFIGURE](#)
[TRY THE DEMO](#)

Standard mode

Kubernetes cluster with node configuration flexibility

[CONFIGURE](#)
[TRY THE DEMO](#)

Scaling

Automatic based on workload

You configure scaling

Nodes

Google manages and configures your nodes

You manage and configure your nodes



Configuration

Streamlined configuration ready to use

You can configure all options

Workloads supported

Most workloads except [these limitations](#)

All Kubernetes workloads

Billing method

[Pay per pod](#)

[Pay per node \(VM\)](#)

SLA

[Kubernetes API and node availability](#)

[Kubernetes API availability](#)

Complete Comparison: <https://cloud.google.com/kubernetes-engine/docs/resources/autopilot-standard-feature-comparison>

GKE Autopilot Cluster

GKE Autopilot Cluster

Pricing: Pay per pod, Pay for the CPU, memory and storage used by our workloads

Pricing: Not charged for system workloads, unused capacity on nodes, Operating System Costs

Security: Clusters have a super hardened configs with many security settings enabled by default.

Security: GKE automatically applies the security patches.

Node Management: Google manages Worker Nodes (Creation, Upgrades, Repairs)

Networking: very advanced. All pod traffic passes through VPC Firewalls.



Scaling: When high load, GKE provisions new nodes for those pods (Node Autoprovisioning)

Resource Management: If we don't specify resource values in workloads, Autopilot sets pre-configured default values

Release Management: All Autopilot clusters are enrolled in GKE release channel which ensures control plane and nodes run on latest versions in that channel

Managed Flexibility: Autopilot supports specific hardware requirements (Compute Classes: General Purpose, Balanced, Scale-Out)

Reduced Operational Complexity: Autopilot reduces platform administration overhead by removing need to continuously monitor nodes, scaling and scheduling operations

Autopilot Cluster

OVERVIEW OBSERVABILITY COST OPTIMIZATION

Filter Enter property name or value ? III

Status	Name ↑	Location	Mode	Number of nodes	Total vCPUs	Total memory
✓	autopilot-cluster-private-1	us-central1	Autopilot		0	0 GB

Observation: Once the Autopilot cluster is created, if we don't deploy any workloads, after some time Number of nodes, Total vCPUs and Total Memory all will come to zero.

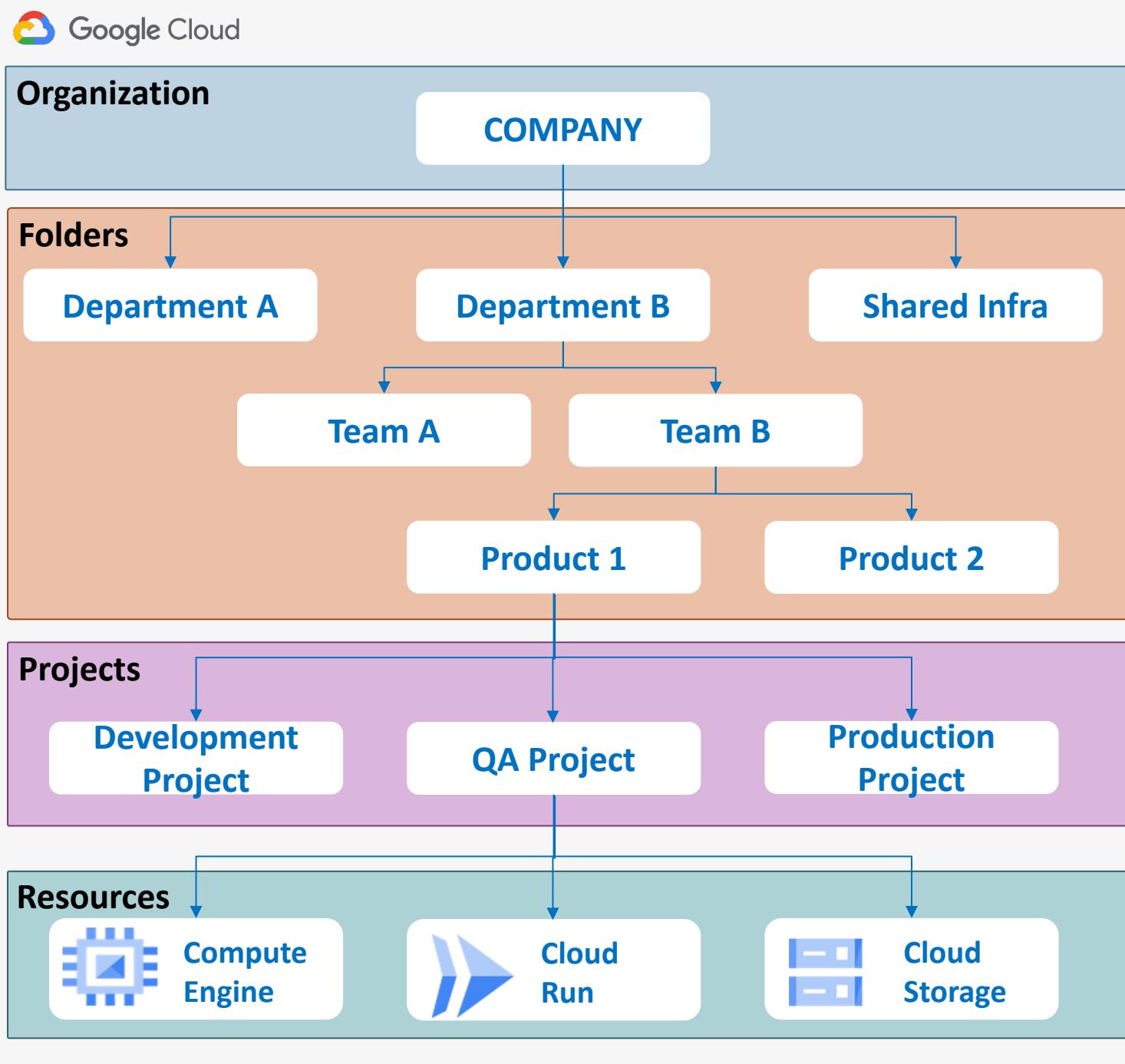
Demo



Google Cloud Resource Manager Resource Hierarchy

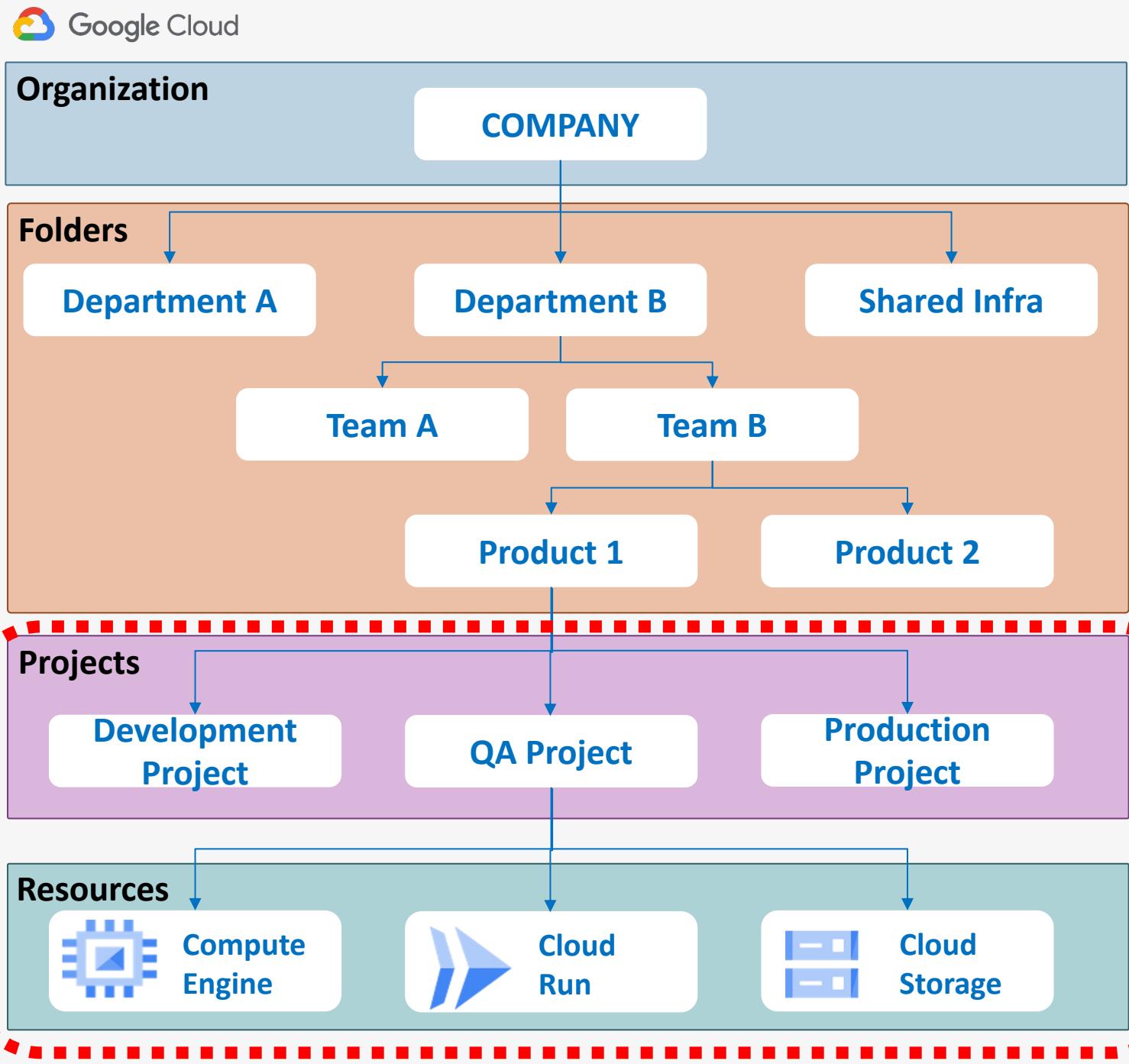


Cloud Resource Manager



- Google cloud resources are organized **hierarchically**
- **TOP LEVEL:** All the resources will have **one parent** each except the highest resource (**Organization**)
- **LOWEST LEVEL:** At the **lowest level** we will have the **Cloud Resources** (**Compute Engine, Cloud Run, Cloud Storage**)
- Google workspace or Cloud Identity account can create ORGANIZATION as the **root node**.
- **Folder Resources**
 - Optional grouping mechanism between organization resources and project resources.
 - An **organization resource** is required as a **prerequisite** to use folders.

Cloud Resource Manager

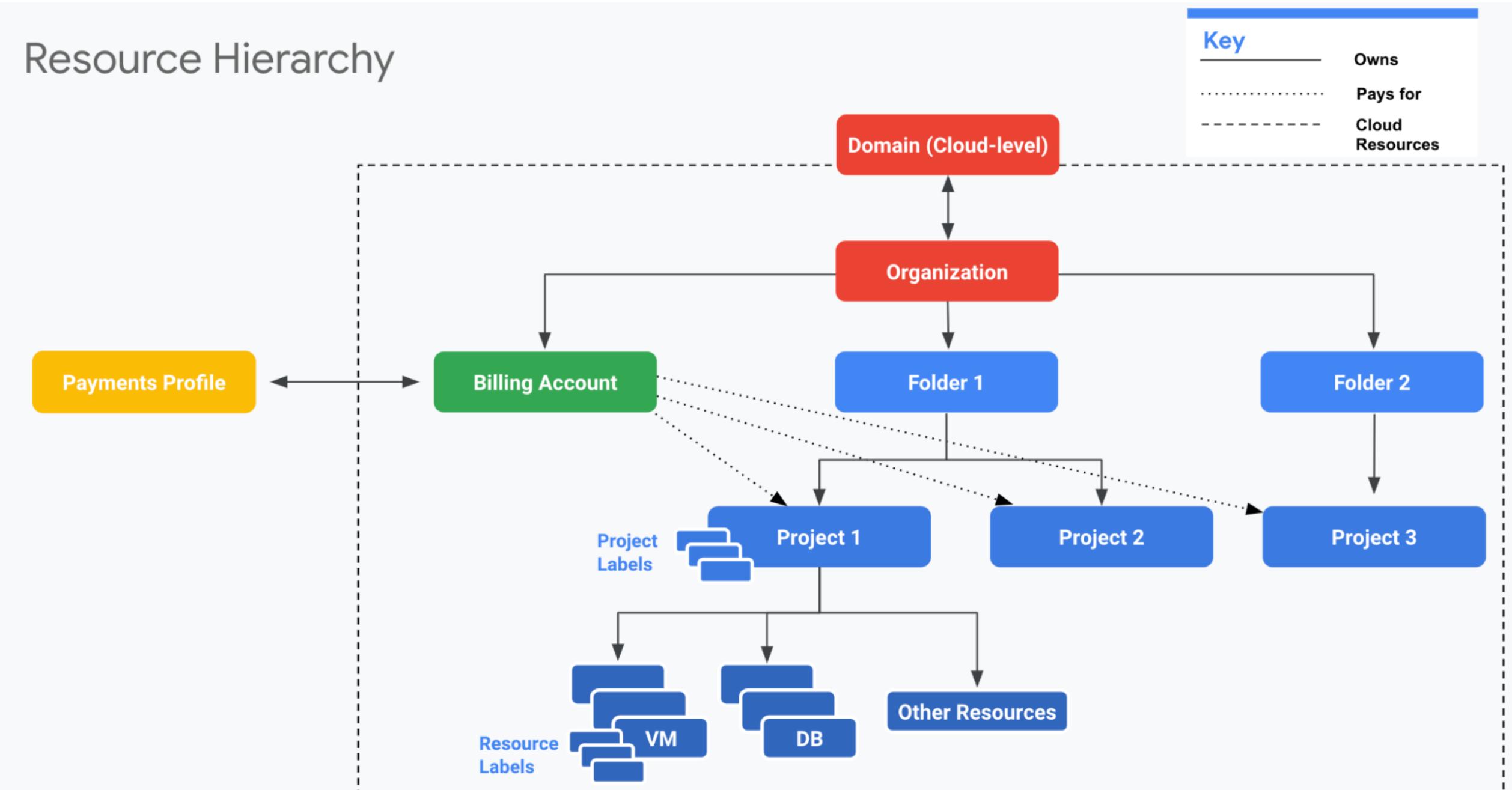


- For Free trial, free tier users will have **NO ORGANIZATION** as root node with projects under it
- We can create **new projects** and under the projects we can create **Cloud Resources**
 - Cloud Storage buckets
 - Compute Engine VM Instances
 - Cloud Run services and many more

Name	ID
▼ No organization	0
☆ gcplearn9	gcplearn9
☆ kdaida123	kdaida123

Cloud Resource Manager

Resource Hierarchy



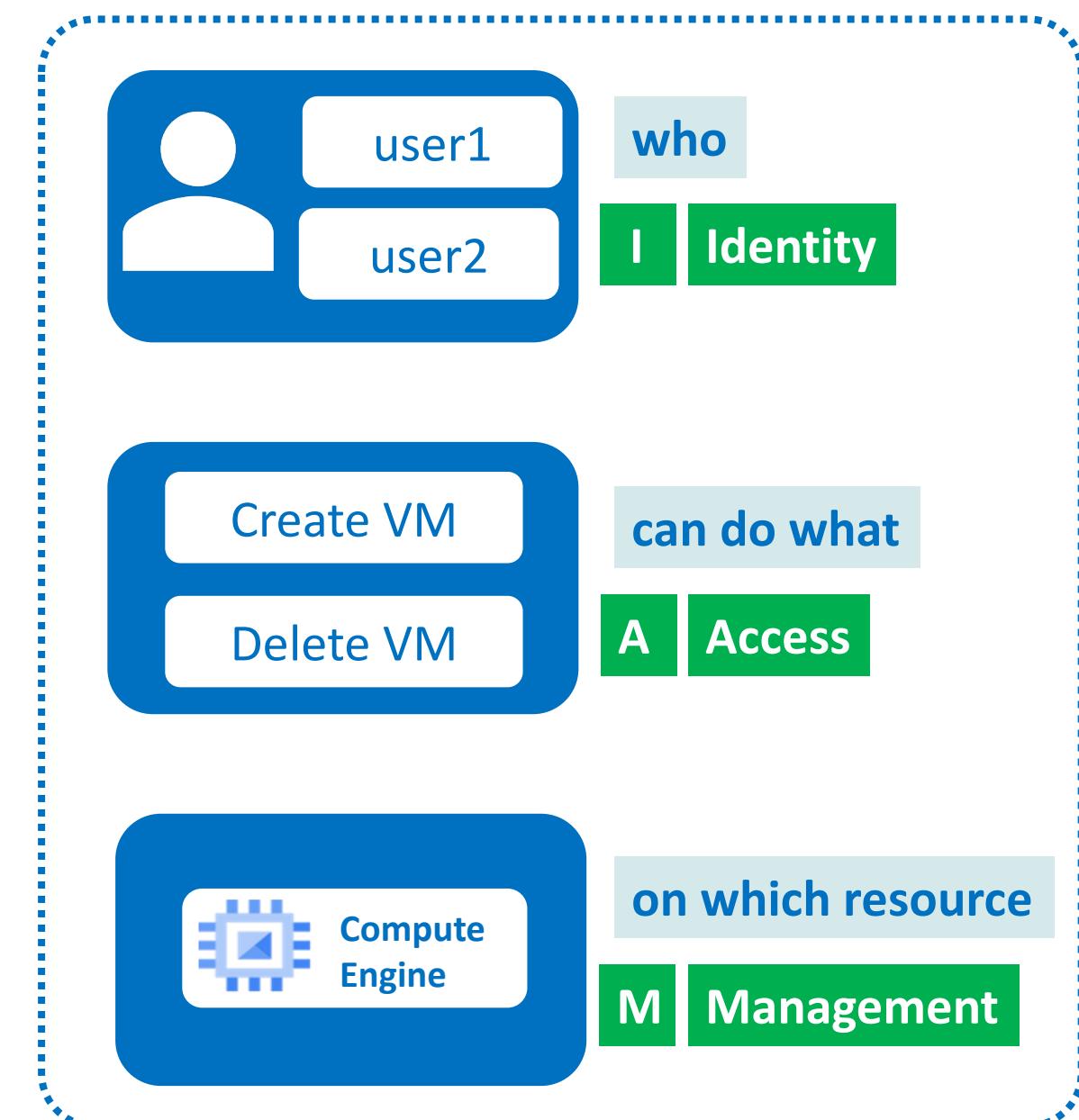
Reference: <https://cloud.google.com>

Demo

Google Cloud IAM IAM Roles

Cloud IAM

- **What is IAM - Identity and Access Management?**
- **Identity / Principal / Member:** In gcp, who can be considered as **identity or principals**
 - Google Accounts
 - Service Accounts
 - Google Groups
 - Google Workspace Accounts
 - Cloud Identity Domain
- **Access:** Roles with Permissions
 - **Role:** Compute Admin, Storage Admin
 - **Permission:** compute.instance.create
 - Example: Compute Admin (**Role**) can create, update, delete VM Instances (**Permissions**)
- **Management:** Manage access to resources
 - GCP Resources
 - Compute Engine, Google Kubernetes Engine
 - Cloud Run, Cloud SQL, GCP services



Cloud IAM Roles

- **What are IAM Permissions?**
- **IAM Permission:** Permissions are **operations that are allowed on a resource**
 - **Example:**
 - compute.instances.create
 - compute.instances.list
 - compute.instances.start
 - compute.instances.stop
- **What are IAM Roles?**
- **IAM Roles:** A role is a **collection of Permissions**
 - **Example:**
 - Compute Admin, Storage Admin
 - When we associate / grant a role to a **Principal (User)** we are granting **all permissions that role contains** to that user

Add permissions

Filter permissions by role

Compute Admin Role

Filter Enter property name or value

Permission ↑	Status
compute.acceleratorTypes.get	Supported
compute.acceleratorTypes.list	Supported
compute.addresses.create	Supported
compute.addresses.createInternal	Testing ⓘ
compute.addresses.delete	Supported
compute.addresses.deleteInternal	Testing ⓘ
compute.addresses.get	Supported
compute.addresses.list	Supported
compute.addresses.setLabels	Testing ⓘ
compute.addresses.use	Supported

Observe the no. of permissions for Compute Admin Role

Permissions

1 – 10 of 839 < >

Cloud IAM Roles

- **IAM Role Types**
 - Basic Roles (Primitive Roles)
 - Predefined Roles
 - Custom Roles
- **Basic Roles (Primitive Roles)**
 - **OWNER:** Full access
 - Example: when we created a google cloud account the [email id](#) which we have used will have this **OWNER role** assigned
 - **EDITOR:** [Edit + View](#) access across google cloud services
 - **VIEWER:** [View only or Read-Only](#) access across google cloud services
- **Important Note:** Assigning these basic roles to multiple users is **not recommended**. In short, **NOT RECOMMENDED FOR PRODUCTION USE**

Type	Principal ↑	Name	Role	Security insights
	dkalyanreddy@gmail.com	Kalyan Reddy Daida	Owner	8244/9340 excess permissions

Cloud IAM Roles

- **Predefined Roles:** Pre-created by google and ready to use
- Provides **Fine-grained** access control
- **Example:**
 - Compute Admin
 - Compute Viewer
 - Compute Network Admin
 - Compute Network Viewer
- Each role serves **different objective**
 - **Compute Admin:** Full access to Compute Engine
 - **Compute Viewer:** Read-only access to Compute Engine

Title	Used in ↑	Status
Compute Admin	Compute Engine	Enabled
Compute Future Reservation Admin	Compute Engine	Enabled
Compute Future Reservation User	Compute Engine	Enabled
Compute Future Reservation Viewer	Compute Engine	Enabled
Compute Image User	Compute Engine	Enabled
Compute Instance Admin (beta)	Compute Engine	Enabled
Compute Instance Admin (v1)	Compute Engine	Enabled
Compute Load Balancer Admin	Compute Engine	Enabled
Compute Load Balancer Services User	Compute Engine	Enabled
Compute Network Admin	Compute Engine	Enabled
Compute Network User	Compute Engine	Enabled
Compute Network Viewer	Compute Engine	Enabled

Cloud IAM Roles

- **Custom Roles:** We can create a [new role](#) by assigning [desired permissions](#) to it
- **When do we create a Custom Role ?**
- When there is [no predefined role](#) satisfying our requirement, we can create a custom role
- **Example:**
 - We can create a custom role which will have permissions [to stop and start a VM Instance](#)

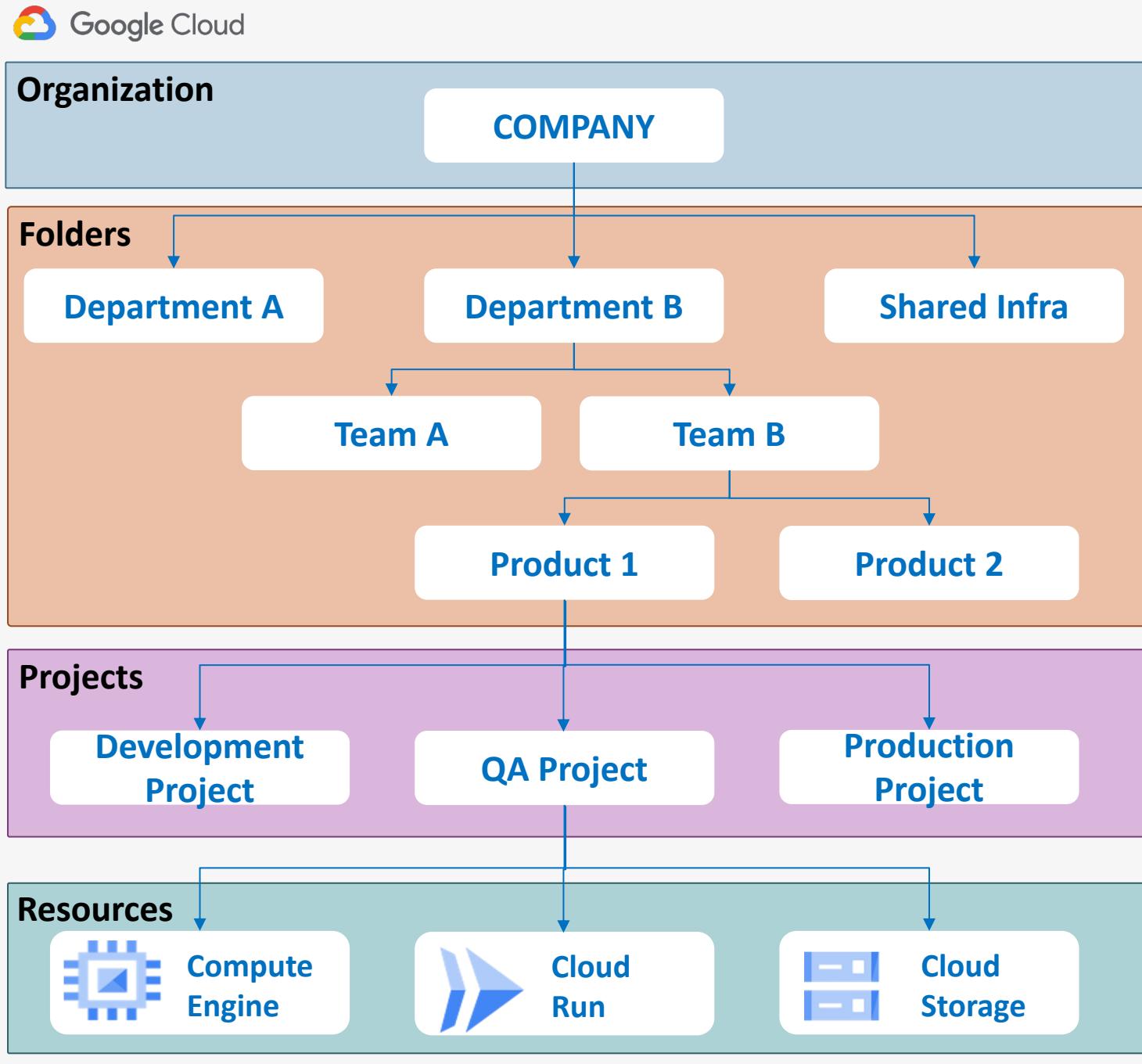
Type	Title	Used in
	Custom Compute Instance Delete Role 101	Custom
	Custom Compute Instance Reset Role 101	Custom
	custom-start-stop	Custom

ID	projects/gcplearn9/roles/CustomRole487
Role launch stage	Alpha
Description	
Created on: 2024-04-15	
2 assigned permissions <div style="border: 1px solid red; padding: 5px; margin-top: 10px;"> compute.instances.start compute.instances.stop </div>	

Demo

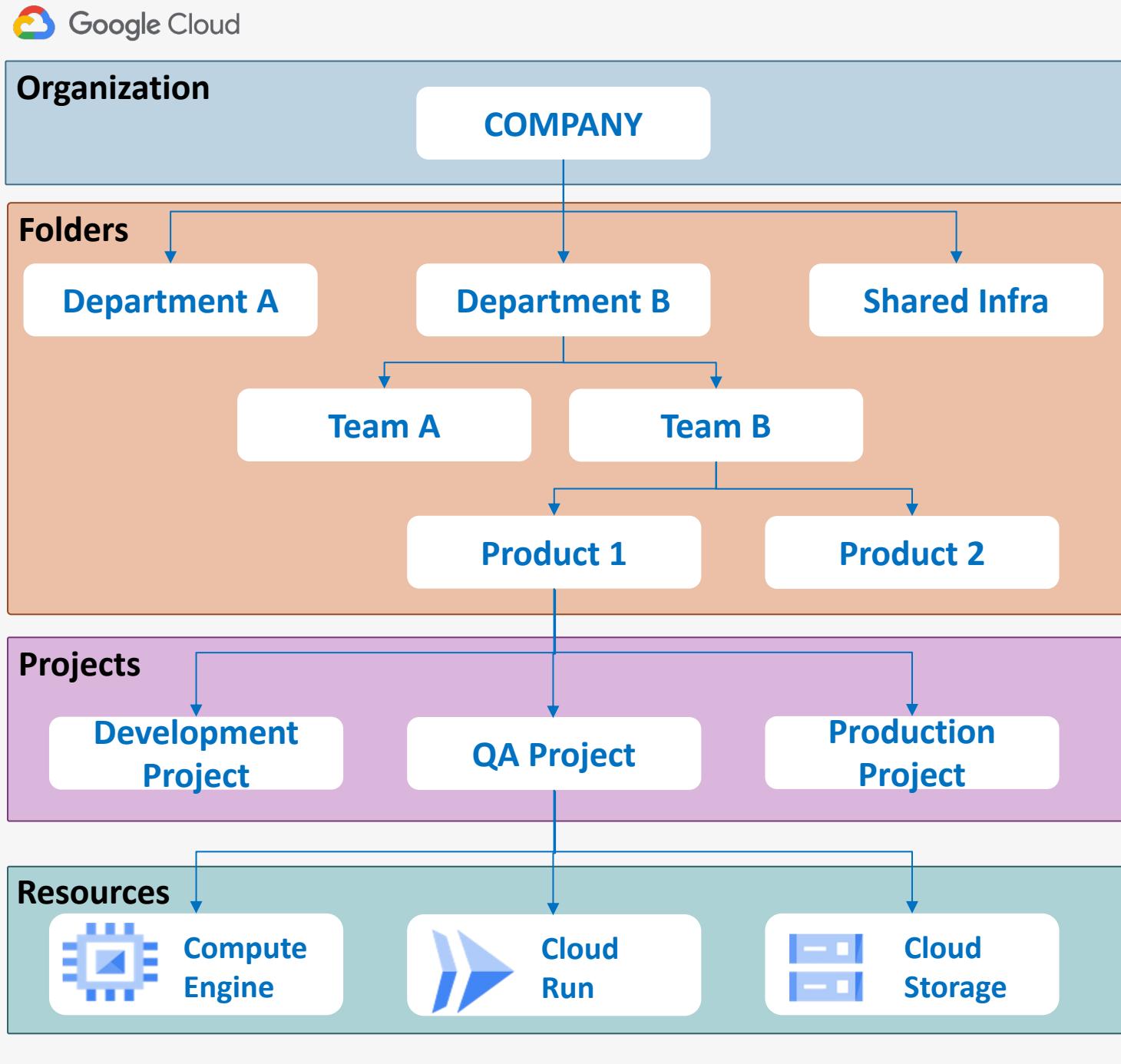
Google Cloud IAM IAM Policy

Cloud Resource Manager + IAM Policy



- We can **set IAM Policy** at
 - Organization level
 - Folder level
 - Project level
 - Resource level (in some cases)
- IAM Policy **set at organization level** is **inherited** by all its child folder, projects and resources
- IAM Policy **set at project level** is **inherited** by all the child resources (Cloud resources like compute engine, cloud run etc)

Cloud Resource Manager + IAM Policy



- Effective policy on a resource is the union of policy set at that resource and policy inherited from its ancestors

Example:

- Development Project IAM Policy = IAM Policy (COMPANY + Department B + Team B + Product 1 + Development Project)



Role Binding

Role

roles/storage.admin

Principals / Members

Google Account
gcpuser08@gmail.comService Account
mysvc103@gcplearn9.iam.gserviceaccount.comGoogle Groups
mygroup1@stacksimplify.com

Role Binding

Role

roles/compute.admin

Principals / Members

Cloud Identity Domain
Stacksimplify.comGoogle Workspace Account
Stacksimplify.com

Cloud IAM Policy

- **IAM Role Binding**

- Bind one or more **principals** to an individual IAM Role
- Principals or Members + IAM Role

- **IAM Policy (Default: Allow Policy)**

- **Collection of role bindings** that bind one or more principals to an individual role
- IAM Policy can have **one or more role bindings**
- An allow policy is **attached to a resource**
 - Example: Organization, Folder, Project or Cloud Resource (Storage Bucket, VM Instance)
- An allow policy will **enforce access control** whenever that resource is accessed.
- **Policy Inheritance:** Policy applied at organization or folder, or project level **will be inherited** to cloud resource level (Storage Bucket or VM Instance)



Role Binding

Role

roles/storage.admin

Principals / Members

Google Account
gcpuser08@gmail.comService Account
mysvc103@gcplearn9.iam.gserviceaccount.comGoogle Groups
mygroup1@stacksimplify.com

Role Binding

Role

roles/compute.admin

Principals / Members

Cloud Identity Domain
Stacksimplify.comGoogle Workspace Account
Stacksimplify.com

Cloud IAM Policy

- **IAM Role Binding**

- **add-iam-policy-binding:** Add IAM policy binding for a resource
- **get-iam-policy:** Get IAM policy for a resource
- **remove-iam-policy-binding:** Remove IAM policy binding for a resource
- **set-iam-policy:** Set IAM policy for a resource

- **Resource: Project: gcplearn9**

- **ADD:** gcloud projects add-iam-policy-binding gcplearn9 --member user:gcpuser08@gmail.com --role=roles/storage.admin
- **GET:** gcloud projects get-iam-policy gcplearn9
- **REMOVE:** gcloud projects remove-iam-policy-binding gcplearn9 --member user:gcpuser08@gmail.com --role=roles/storage.admin



Role Binding

Role

roles/storage.admin

Principals / Members

Google Account

gcpuser08@gmail.com

Service Account

mysvc103@gcplearn9.iam.gserviceaccount.com

Google Groups

mygroup1@stacksimplify.com

Role Binding

Role

roles/compute.admin

Principals / Members

Cloud Identity Domain

Stacksimplify.com

Google Workspace Account

Stacksimplify.com

Cloud IAM Policy

IAM Policy - JSON

```
{  
  "bindings": [  
    {  
      "role": "roles/storage.admin",  
      "members": [  
        "user:gcpuser08@gmail.com",  
        "serviceAccount:mysvc103@gcplearn9.iam.gserviceaccount.com"  
      ]  
    },  
    {  
      "role": "roles/compute.admin",  
      "members": [  
        "group:mygroup1@stacksimplify.com",  
        "domain:stacksimplify.com"  
      ]  
    }  
  ]  
}
```

Role Binding - 1

Role Binding - 2



Role Binding

Role

roles/storage.admin

Principals / Members

Google Account

gcpuser08@gmail.com

Service Account

mysvc103@gcplearn9.iam.gserviceaccount.com

Google Groups

mygroup1@stacksimplify.com

Role Binding

Role

roles/compute.admin

Principals / Members

Cloud Identity Domain

Stacksimplify.com

Google Workspace Account

Stacksimplify.com

Cloud IAM Policy

IAM Policy - YAML**bindings:****- role: roles/storage.admin****members:****- user:gcpuser08@gmail.com****- serviceAccount:mysvc103@gcplearn9.iam.gserviceaccount.com**

Role Binding - 1

- role: roles/compute.admin**members:****- group:mygroup1@stacksimplify.com****- domain:stacksimplify.com**

Role Binding - 2

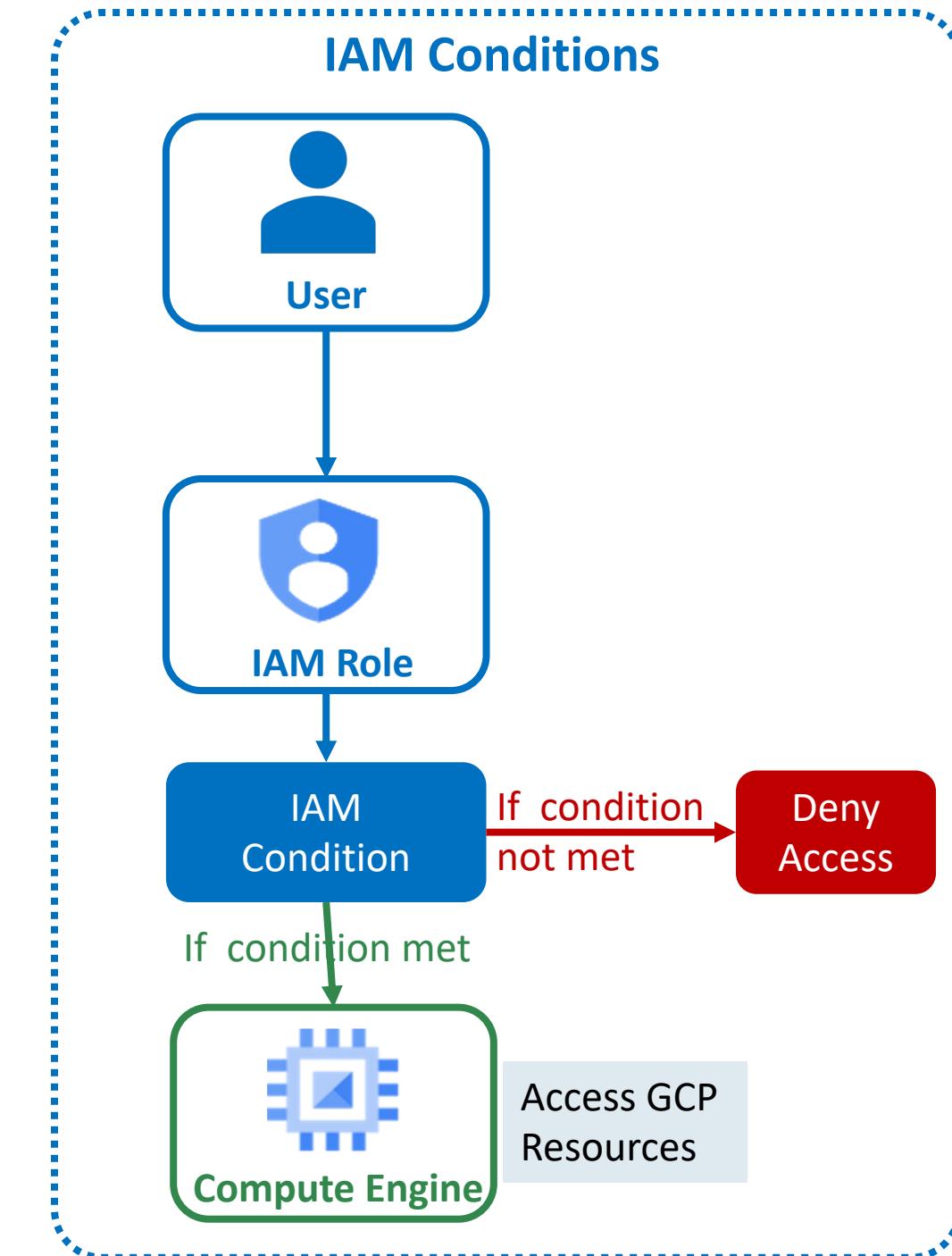
Demo

Google Cloud IAM

IAM Conditions

Cloud IAM Conditions

- **IAM Conditions:** used for **enforcing conditional, attribute-based access control** for Google Cloud resources
- Grant access to **principals** only if **specific condition** is met
 - Grant temporary access to users for a **specified amount of time** to troubleshoot a production issue
 - Grant access on a **specific day or hour**
- **IMPORTANT NOTE:** You **cannot use conditions** for basic roles (Owner, Editor and Viewer)





Role Binding

Role

roles/compute.viewer

Principals / Members

Google Account
gcpuser08@gmail.com

Condition

IAM Conditions

Role Binding

Role

roles/compute.viewer

Principals / Members

Cloud Identity Domain
Stacksimplify.comGoogle Groups
mygroup1@stacksimplify.com

Cloud IAM Policy with IAM Condition

IAM Policy - YAML

```
bindings:  
- role: roles/compute.viewer  
  members:  
  - user:gcpuser08@gmail.com  
    condition:  
      expression: request.time.getDayOfWeek("Asia/Calcutta") == 0  
      title: access-on-a-day  
- role: roles/compute.admin  
  members:  
  - group:mygroup1@stacksimplify.com  
  - domain:stacksimplify.com
```

IAM Condition

Demo



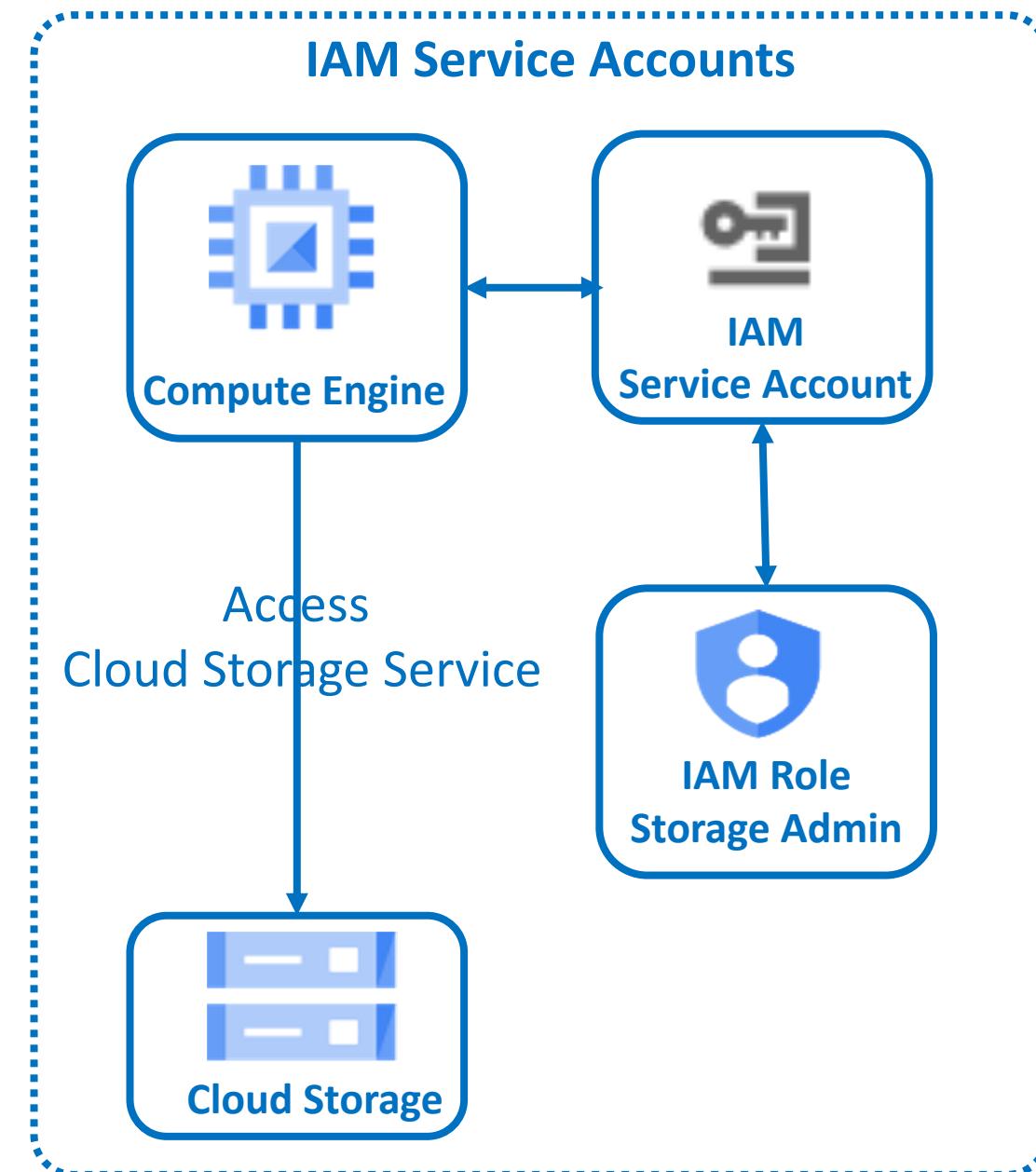
Google Cloud IAM

IAM Service Accounts



Cloud IAM Service Accounts

- How do you access GCP resources using Users ?
 - We will associate required **IAM Role** to User
 - User: gcpuser08@gmail.com
 - Role: Compute Viewer, Storage Admin
- How do GCP Services access other GCP Services?
 - Compute Engine want to access Cloud Storage Buckets
 - **Solution:** Use IAM Service Accounts
- Using Service Accounts
 - GCP Services can access other GCP Services
 - On-prem Services (Application workloads) can access GCP Services



Cloud IAM Service Account - Types

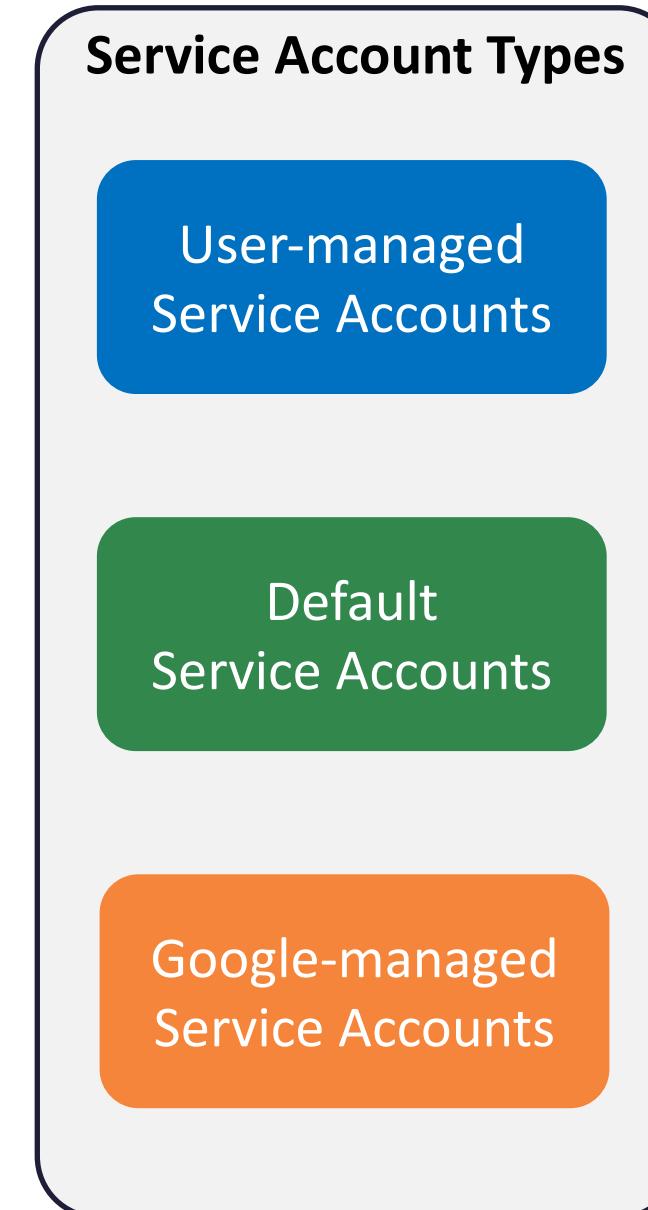
- **Types of Service Accounts**

- **User-managed service accounts (RECOMMENDED)**

- You create and manage these service accounts
- **Example:**
 - Email: mysvc103@gcplearn9.iam.gserviceaccount.com

- **Default service account (NOT RECOMMENDED)**

- Created automatically when you enable certain Google Cloud services
- **Example:**
 - **Compute Engine:**
 - project-number-compute@developer.gserviceaccount.com
 - 899156651629-compute@developer.gserviceaccount.com
 - These are google-created but user-managed service accounts
 - Automatically grants Editor Role (Basic role) which contains huge permissions (NOT RECOMMENDED TO USE IT, least-privilege approach is recommended)



Cloud IAM Service Account - Types

- **Types of Service Accounts**

- **Google-managed service accounts**

- Google-created and Google-managed service accounts
- Google Cloud services **need access to your resources** so that they can **act on your behalf**
- We **don't have access to edit or modify** these service accounts
- We can **see them** in our project allow-policy
 - gcloud projects get-iam-policy PROJECT-ID
 - gcloud projects get-iam-policy gcplearn9
- In short, we **don't need to worry** about these service accounts

Service Account Types

User-managed
Service Accounts

Default
Service Accounts

Google-managed
Service Accounts

Demo



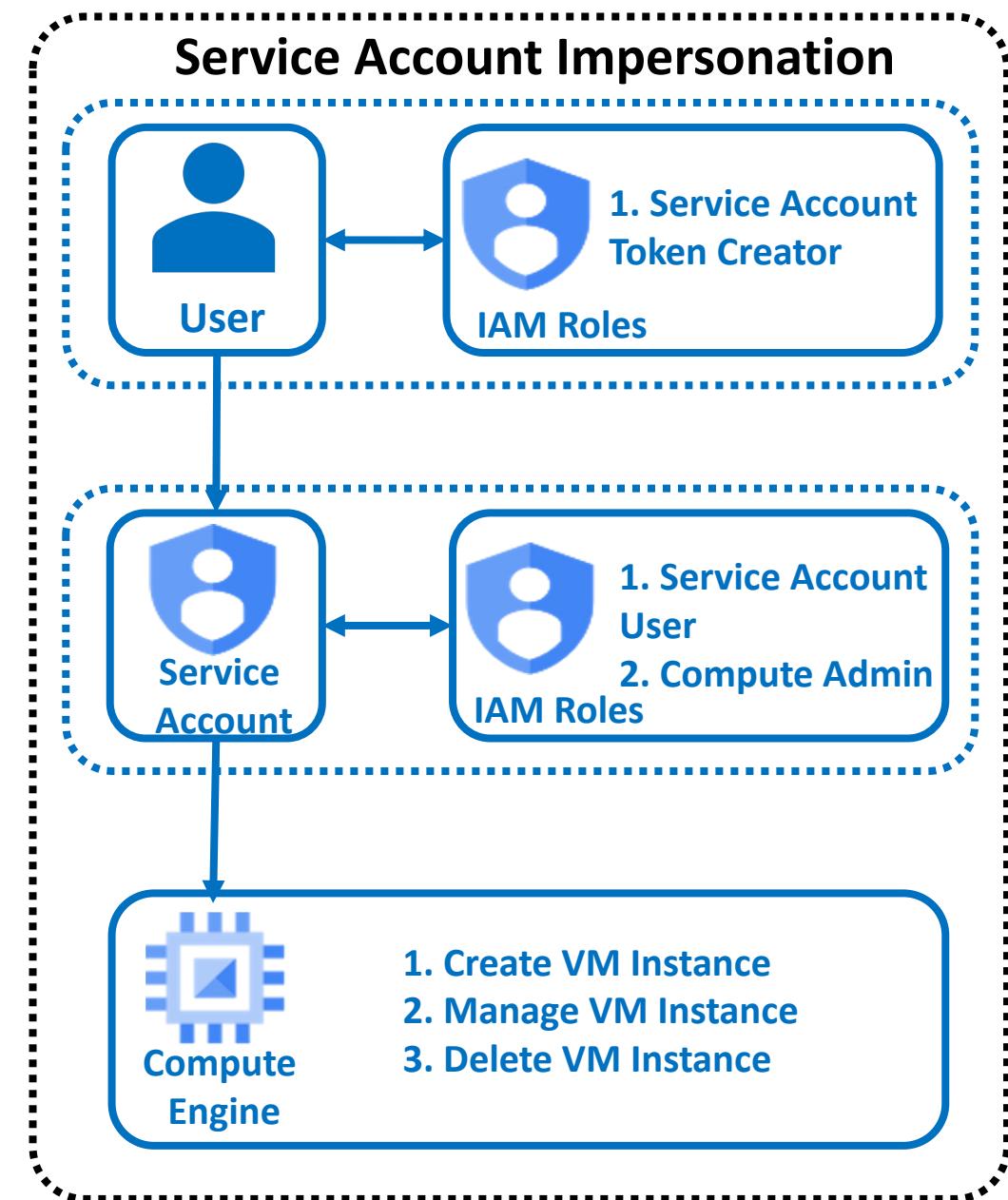
Google Cloud IAM



IAM Service Account Impersonation

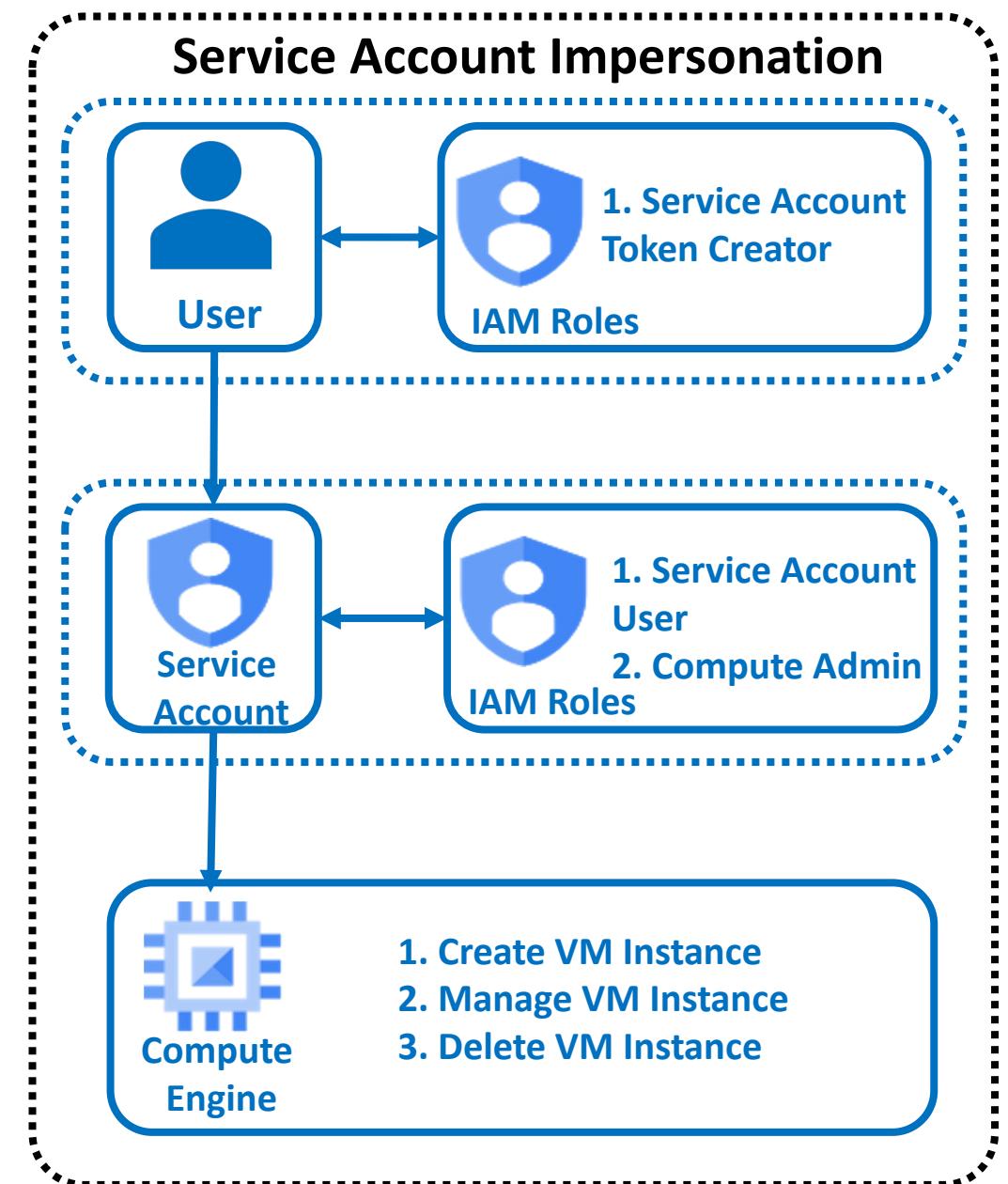
Cloud IAM Service Account - Impersonation

- Scenario-1: How to Create a VM Instance using a normal google cloud user who doesn't have Compute Admin Permissions?
- Step-1: For a normal, User associate Service Account Token Creator IAM role
- Step-2: Create Service Account
- Step-3: Associate the roles Service Account User, Compute Admin to newly created Service account
- Step-4: Create VM Instance using gcloud with flag *--impersonate-service-account*



Cloud IAM Service Account - Impersonation

- IAM Service Account Impersonation:**
 Impersonating a service account lets an **authenticated principal access** whatever the **service account can access**.
- User or a service account can impersonate other service account**
- Usecases**
 - Temporarily grant a user **elevated access**
 - To test specific set of permissions is **sufficient** for a task
 - Authenticate applications that **run outside of GCP**
 - Locally develop applications **that can only run** as a service account



Demo



Google Cloud IAM



IAM Service Account Long-lived Credentials

Cloud IAM Service Account - Credentials

- How do you associate Service Accounts to applications hosted in on-premise or from your local development environment to access to GCP Services?
- We cannot directly associate it using cloud console.
- We can do that by generating credentials using Service Account API and by using ADC (Application Default Credential)
- Service Account API can create the following types of Credentials
 - Long-lived Credentials
 - Service Account Keys
 - Short-lived Credentials
 - OAuth 2.0 Access Tokens
 - OpenID Connect (OIDC) ID Token
 - Self-signed JSON Web Tokens (JWTs)
 - Self-signed binary blobs

Cloud IAM Service Account - Keys (Long-lived)

- **Service Account Keys (Long-lived)**
- Each service account is associated with a [public/private RSA key pair](#).
- **User-managed Key pairs**
 - You can create key pair using [google cloud console](#) or [gcloud](#) for a Service Account
 - `gcloud iam service-accounts keys create`
 - Private key will be [downloaded when the key pair is created](#), and you can use that to authenticate to google cloud
 - Private key is called [SERVICE ACCOUNT KEY](#)
 - By default, Service account key [never expires](#)
 - You can set an [expiry time](#) for all newly created keys in your [project](#), [folder](#), or [organization level](#) using [IAM Policy](#) that enforces a constraint `constraints/iam.serviceAccountKeyExpiryHours`

Cloud IAM Service Account - Keys (Long-lived)

- **When to set expiry time for service account Keys ?**
 - In scenarios where we need to provide **temporary access**, we can set expiry time
 - Access to a **developer** to a specific service, during the development period, later access should be revoked
 - Access to **third party tools**
- **When to not set expiry time for service account keys ?**
 - **Production workloads** should not have an expiry time, this will cause **major outage** if keys are expired
 - Non-production workloads that **need permanent access** (CI CD pipelines)

Demo



Google Cloud IAM



IAM Service Account Short-lived Credentials

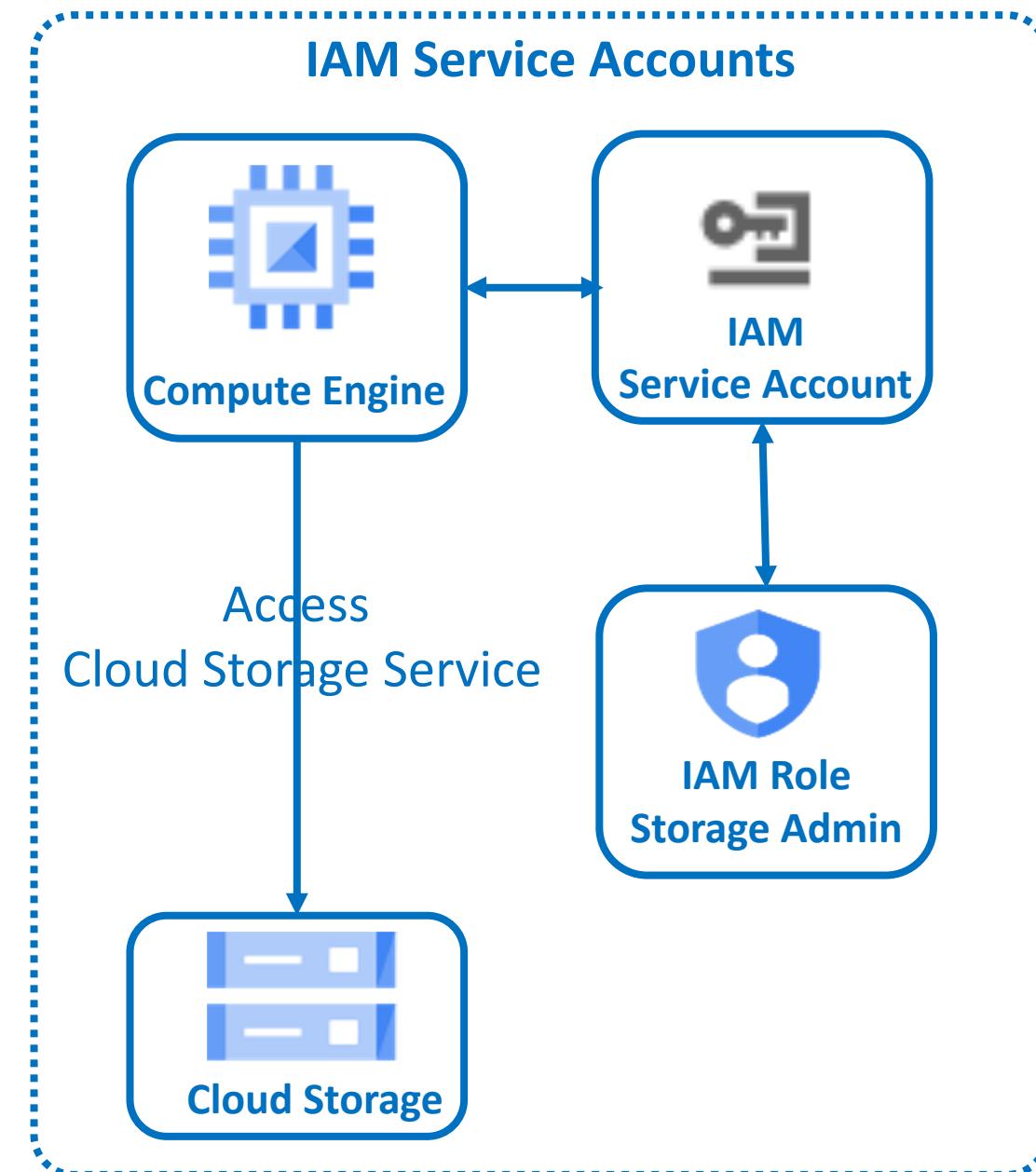
Cloud IAM Service Account - Keys (Short-lived)

- Short-lived service account credentials
- The most secure way to authenticate as a service account is to obtain short-lived credentials for the service account in the form of an OAuth 2.0 access token.
- By default, these tokens expire after 1 hour.
- These can be generated automatically using Cloud client libraries
- Google-managed Key pairs
 - Used by App Engine, Compute Engine to generate short-lived credentials for service accounts, sign blobs and JWTs
 - Automatically rotated
 - Private key is not accessible directly, always held in escrow
- Example Scenario-1:
 - The below command using gcloud will generate a short-lived OAuth 2.0 access token valid for 1 hour
 - Create SERVICE_ACCOUNT with necessary roles (Example: Storage Admin)
 - `gcloud auth print-access-token --impersonate-service-account=<SERVICE_ACCOUNT>`
 - Use ACCESS_TOKEN to create Google Cloud Resources

Cloud IAM Service Account - Keys (Short-lived)

- **Example Scenario-2:**

- You have an [application \(workload\)](#) deployed on Compute Engine VM Instance which needs access to Cloud Storage bucket
- You can [attach a service account](#) to VM Instance which has permissions to Cloud Storage bucket
- The workload on VM Instance can use [Cloud Client libraries](#) to access Cloud Storage Bucket
- **What does Cloud Client Libraries use to obtain short-lived Credentials ?**
- Cloud Client Libraries uses [ADC \(Application Default Credentials\)](#) to obtain short-lived credentials for a service account



Cloud IAM Service Account - ADC

- **What is Application Default Credentials (ADC)?**

- ADC is a strategy used by the [Google authentication libraries](#) to automatically find [credentials](#) based on the application environment.
- The authentication libraries [make those credentials available](#) to Cloud Client Libraries and Google API Client Libraries.
- When you use ADC, your code can run in either a development or production environment [without changing how your application authenticates](#) to Google Cloud services and APIs.

- **ADC Search Order (Priority high to low)**

- [GOOGLE_APPLICATION_CREDENTIALS](#) environment variable
- User credentials set up by using the Google Cloud CLI
 - **Linux OS:** `$HOME/.config/gcloud/application_default_credentials.json`
 - **Windows OS:** `%APPDATA%\gcloud\application_default_credentials.json`
- The [attached service account](#), returned by the metadata server

- **For complete reference:**

- <https://cloud.google.com/docs/authentication/provide-credentials-adc>
- <https://cloud.google.com/docs/authentication/application-default-credentials>

Cloud IAM Service Accounts - ADC

How to provide credentials to ADC ?

Local development environments

Or

On-premise environments

Solution

1. We can use [User accounts or Service Accounts with Service Account Keys](#)

User accounts: *gcloud auth application-default login*

Service Accounts: *gcloud auth application-default login --impersonate-service-account SERVICE_ACCT_EMAIL*

2. [Download](#) the Service account keys (json file)
3. Set the environment variables [**GOOGLE_APPLICATION_CREDENTIALS**](#) to the path of service account key JSON file
4. Configure cloud client libraries in your application to use ADC
5. When you set the **GOOGLE_APPLICATION_CREDENTIALS** environment variable, [ADC checks this location first](#), then checks other locations only if necessary.

Google Cloud cloud-based development environment

1. Cloud Shell or Cloud Code uses the [credentials you provided when you signed in](#), and manages any authorizations required
2. You cannot use the gcloud CLI to provide credentials to ADC in these environments
3. NO SPECIFIC MANUAL ADC setup is needed

Cloud IAM Service Accounts - ADC

How to provide credentials to ADC ?

Google Cloud services that support attaching a service account

Solution

1. Compute Engine, App Engine, and Cloud Functions, support [attaching a user-managed service account](#)
2. The [code running on these resources](#) can use that service account as its identity.
3. NO SPECIFIC MANUAL ADC setup is needed

GKE or GKE Enterprise

1. Uses workload identity
2. NO SPECIFIC MANUAL ADC setup is needed

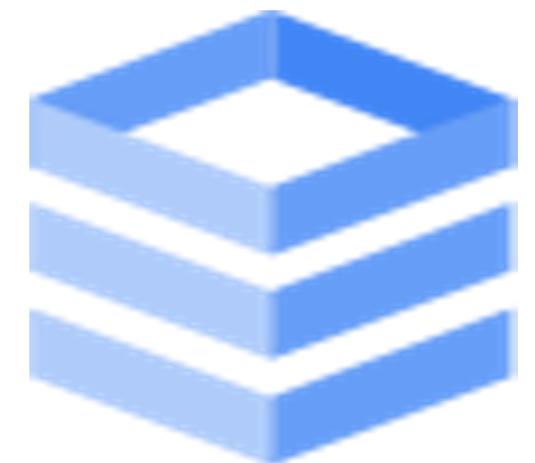
Demo



Google Cloud Data Solutions

Cloud SQL

MySQL, SQL Server and PostgreSQL



Google Cloud SQL

- **Cloud SQL:** Full managed relational database service
- **Supports**

- MySQL
- SQL Server
- PostgreSQL


 Reference: <https://cloud.google.com/sql>

Google Cloud SQL

- Provides built-in **High Availability** with
 - Automatic [failover](#) across zones
 - [99.99%](#) SLA availability for business-critical transactional workloads
 - [99.95%](#) SLA availability to lower costs
- **Reliability and Scalability**
 - We can create [multiple read replicas](#) across zones and regions
 - We can create [automatic backups](#) in a single region or in multiple regions
 - Supports [manual](#) backups (create backup anytime)
 - Supports [point in time recovery](#) (MySQL, PostgreSQL)
 - Instance deletion [protection](#): Avoid accidental deletion
 - [Vertical scaling](#) (Add processor cores, RAM and Storage)
- **Storage**
 - **HDD**: low performance, low cost
 - **SSD**: low latency, high cost, excellent performance, very high QPS (Queries per second)
 - Also supports [automatic storage increase](#)

Google Cloud SQL

- **Data Security**
 - Automatic data [encryption at rest](#)
 - For database tables
 - Temporary files
 - Backups
- **Network Security**
 - Access database with [Public IP or Private IP](#)
 - Supports [private connectivity](#) with VPC (Ex: for GCE, GKE workloads)
 - Contains network firewall allowing to [control public network access](#)
- **Maintenance**
 - [Near-zero downtime](#) of less than [10 seconds](#) for planned maintenance
 - Flexible maintenance window
 - Advance email notification
 - Reschedule by 28 days
 - Pick specific times for maintenance
 - Deny maintenance for 90 days

Google Cloud SQL

- **Cloud DMS - Data Migration Service**

- Migrate from on-premises, Google Cloud, or other clouds to Cloud SQL
- Replicate data continuously for minimal downtime migrations
- Serverless and easy to set up

Demo



Google Cloud Data Solutions

Dataflow Streaming Analytics Service



Google Cloud: Dataflow

- **Dataflow:** Fully managed **data processing service, serverless, fast and cost effective**
- We can use Dataflow to
 - **create jobs** that read from one or more sources,
 - transform the data, and
 - **write the data** to a destination.
- **Automated provisioning and management of processing resources**
 - **No manual** provisioning or management
 - Processing resources gets **provisioned and run during the tenure of dataflow job** and as soon as the job is completed (Example: Batch Job), all resources will be **de-provisioned automatically**.
- **Autoscaling**
 - **Horizontal and vertical autoscaling** of worker resources to maximize resource utilization

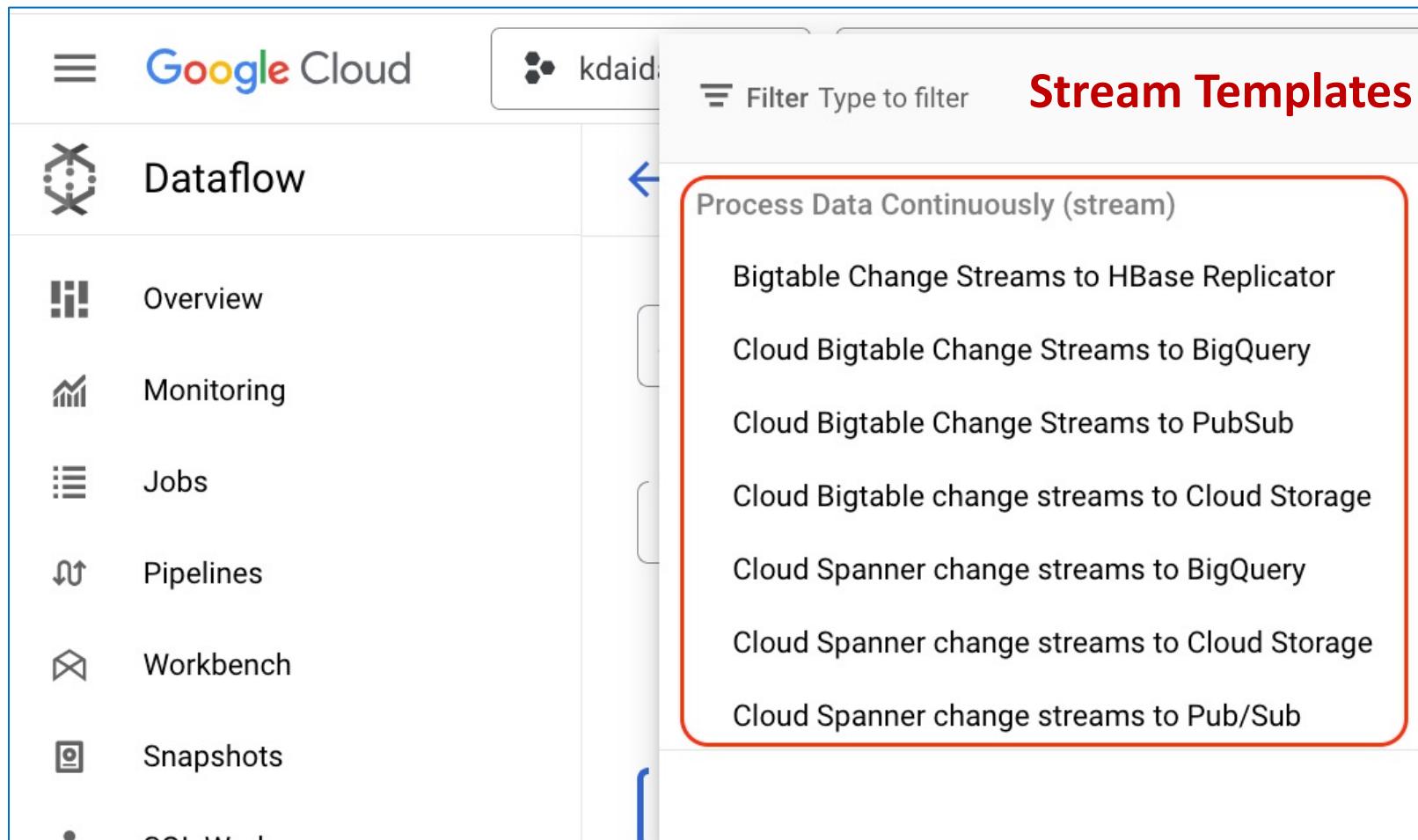
Google Cloud: Dataflow

- **Dataflow Jobs**
 - **Stream:** Process data **continuously**
 - **Batch:** Process data in **bulk**
- **Ready-to-use-real-time AI**
 - Can leverage **out-of-the-box ML features**, including NVIDIA GPU.
 - We can build intelligent solutions for **predictive analytics, anomaly detection, and real-time personalization**.
 - Train, deploy, and manage **complete ML pipelines**, supporting both batch and streaming workflows.

Google Cloud: Dataflow

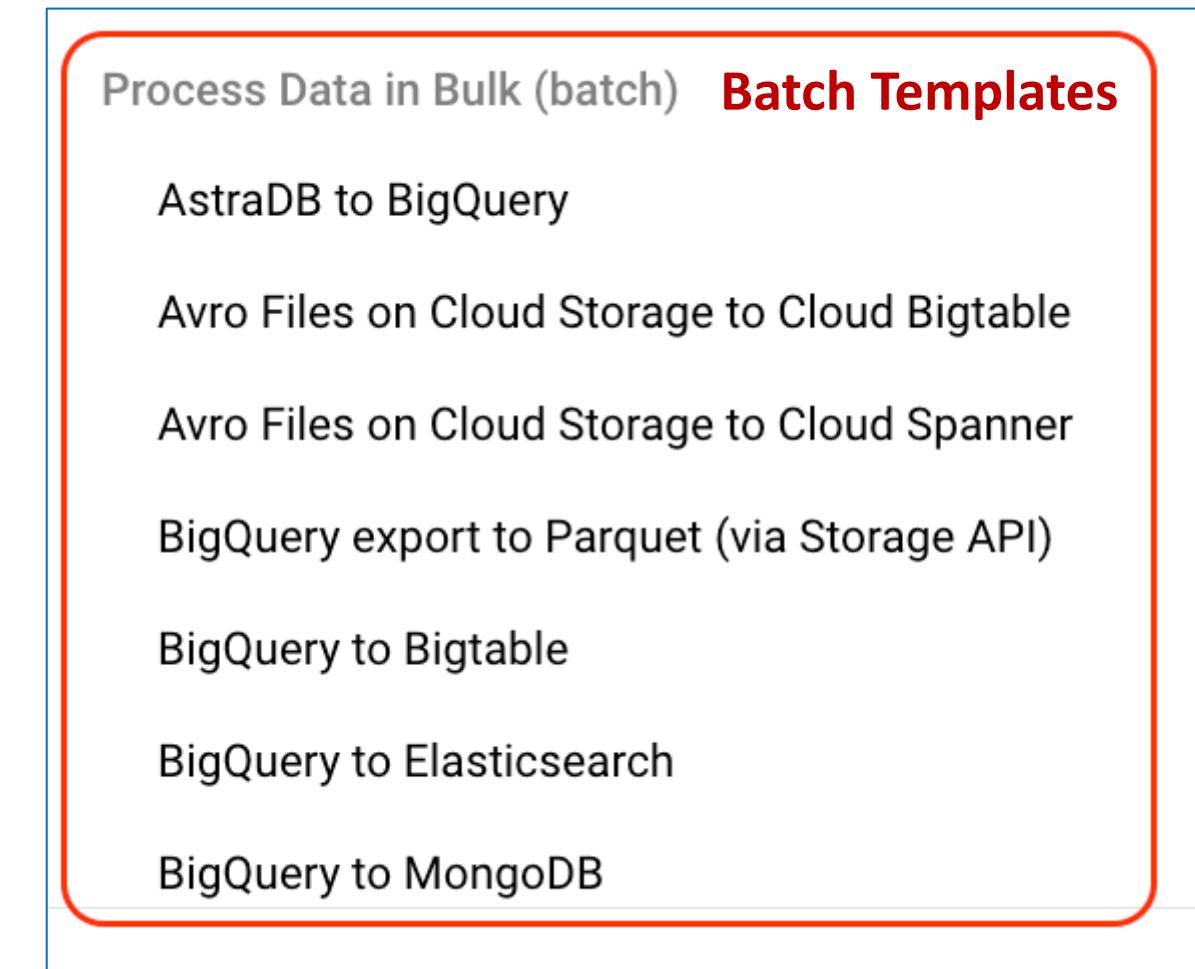
- **Dataflow Templates:**

- Predefined Dataflow templates (ready to use) makes our job easier
- We can also create **custom templates** and share those templates with team.



The screenshot shows the Google Cloud Dataflow Stream Templates interface. On the left, there's a sidebar with icons for Overview, Monitoring, Jobs, Pipelines, Workbench, and Snapshots. The main area has a title "Stream Templates" and a "Filter Type to filter" dropdown. A red box highlights a section titled "Process Data Continuously (stream)" containing the following items:

- Bigtable Change Streams to HBase Replicator
- Cloud Bigtable Change Streams to BigQuery
- Cloud Bigtable Change Streams to PubSub
- Cloud Bigtable change streams to Cloud Storage
- Cloud Spanner change streams to BigQuery
- Cloud Spanner change streams to Cloud Storage
- Cloud Spanner change streams to Pub/Sub



The screenshot shows the Google Cloud Dataflow Batch Templates interface. A red box highlights a section titled "Process Data in Bulk (batch) Batch Templates" containing the following items:

- AstraDB to BigQuery
- Avro Files on Cloud Storage to Cloud Bigtable
- Avro Files on Cloud Storage to Cloud Spanner
- BigQuery export to Parquet (via Storage API)
- BigQuery to Bigtable
- BigQuery to Elasticsearch
- BigQuery to MongoDB

Demo



Google Cloud Data Solutions

Cloud Spanner

GoogleSQL and PostgreSQL



Google Cloud Spanner

- **Cloud Spanner:** Full managed relational database service
- **Reliability and Scalability**
 - Write and read scalability with no limits
 - Scales Horizontally
 - Each compute capacity (node) can process both reads and writes.
 - Apps backed by spanner can read and write up-to-date strongly consistent data globally
 - When running a multi-region instance, Spanner database is protected against a regional failure and offers industry-leading 99.999% availability.
- **SQL Dialects:** When creating a Cloud Spanner Instance, we need to choose the SQL dialect
 - GoogleSQL
 - PostgreSQL

Google Cloud Spanner

- **Automated maintenance**

- Synchronous replication is built-in and automated
- Maintenance (Upgrades, Patches) is built-in with **Zero Downtime**
- 100% online schema changes allowed with **Zero Downtime**

- **Backup and Restore**

- Supports **manual** backup and restore
- Also supports **PITR** (Point in Time Recovery to a **microsecond** granularity)

- **Security**

- Customer-managed encryption keys (CMEK)
- Data-layer encryption
- **Fine-grained** Access control using Cloud IAM
 - Authorize access to Spanner data at the **table and column level**
 - Comprehensive audit logging

Google Cloud Spanner

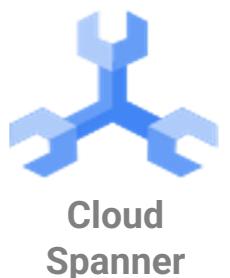
- **Cloud Spanner Data Boost**

- To run operational tasks **without affecting the existing transactional workload**
 - Analytical Queries
 - Batch processing jobs
 - Data export operations
- It is always **hot and ready** to process queries

Google Cloud Spanner

Database Attribute	Relational Database	Non-Relational Database	Cloud Spanner
Schema	Static	Dynamic	Dynamic
SQL	Yes	No	Yes
Transactions	ACID (atomicity, consistency, isolation, durability)	Eventual	Strong-ACID with TrueTime Ordering
Scalability	Vertical (use a bigger machine)	Horizontal (add more machines)	Horizontal (Automatic)
Availability	Failover (Downtime)	High	High 99.999% SLA
Replication	Configurable	Configurable	Automatic

Cloud Spanner is a combination of features from standard Relational and Non-Relational Databases.



Demo

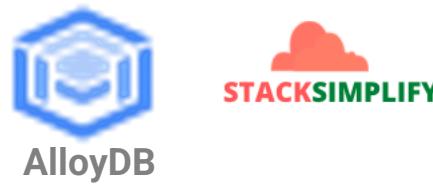


Google Cloud Data Solutions



AlloyDB for PostgreSQL Relational Database

Google AlloyDB for PostgreSQL



Features	Cloud SQL for PostgreSQL	AlloyDB for PostgreSQL
Use cases	Used for smaller and medium workloads like 1. web applications, 2. content management system (CMS) 3. General purpose database needs	Used for high demanding workloads like 1. Large datasets 2. Data warehousing 3. HTAP – Hybrid Transactional and Analytical Processing
Scalability	Supports Vertical Scaling only	Supports Vertical and Horizontal Scaling
Performance	Performance will be low when running complex queries and large datasets	Delivers high performance and scales up and down to handle varying workloads
Key Features	Built-In High Availability and Disaster Recovery	Very advanced features like 1. Automatic Data Tiering 2. Analytics Acceleration 3. Built-in machine learning for optimizing performance and managing complex workloads

Google AlloyDB for PostgreSQL

Features	Cloud SQL for PostgreSQL	AlloyDB for PostgreSQL
Cost	Less expensive when compared to AlloyDB	Very expensive due to its advanced features
Ease of Use	Simple to manage	<ol style="list-style-type: none">1. Requires more configuration and management due to its advanced features2. Steeper learning curve is needed

Demo



Google Cloud Data Solutions

Cloud Firestore No-SQL Databases



Cloud Firestore

- **Cloud Firestore:** Fully managed, scalable, and serverless document database
- **Two modes**
 - **Native Mode(Firebase):** Recommended for all servers, mobile apps, and web apps
 - **Datastore Mode:** Use Datastore mode if your app requires the [Datastore API](#)
- **Serverless**
 - Scales up or down to meet [any demand](#)
 - [No maintenance windows or downtime](#)
- **Replication**
 - Automatic [regional replication](#) with [99.99%](#) SLA availability
 - Automatic [multi-regional replication](#) with [99.999%](#) SLA availability

Firestore VS Datastore

Native mode		
Datastore mode		
Fully managed, scalable, and serverless document database with offline support and Real-time synchronization.	Learn more ↗	Fully managed, scalable NoSQL database.
Learn more ↗		Learn more ↗
SELECT		SELECT
API	Firestore	Datastore
Real-time updates	✓	✗
Mobile/web client libraries with offline data persistence	✓	✗
Query consistency	Strong	Strong
Data model	Documents / collections	Entities / kinds
Web console	Firestore page in Google Cloud and Firebase	Datastore page in Google Cloud

Cloud Firestore

- **AI Integrations**

- Seamless integrations with [AI services](#) with few clicks
- With simple steps enable [AI usecases](#) like
 - Automated language translations
 - Image classification and many more

- **Strong Consistency**

- Allows us to run sophisticated [ACID transactions](#) against document data.
- This gives us [more flexibility in the way we structure](#) our data.

- **Rich Development library support**

- Client Side: Web, iOS, Android, Flutter, C++, and Unity
- Server Side: Node.js, Java, Go, Ruby, and PHP

Cloud Firestore

- What is offline data persistence mode (built-in) and Live synchronization?
- Firestore supports offline data persistence
 - Cloud Firestore Offline Mode caches data for offline access, ensuring continued app functionality without internet.
 - Cached data supports write, read, listen, and query operations even when offline.
 - Upon reconnection, Cloud Firestore syncs local changes to the cloud backend seamlessly.
 - No code changes are required to enable offline persistence in Cloud Firestore.
 - The Firestore client library automatically handles offline and online data access, syncing changes when online.

Cloud Firestore

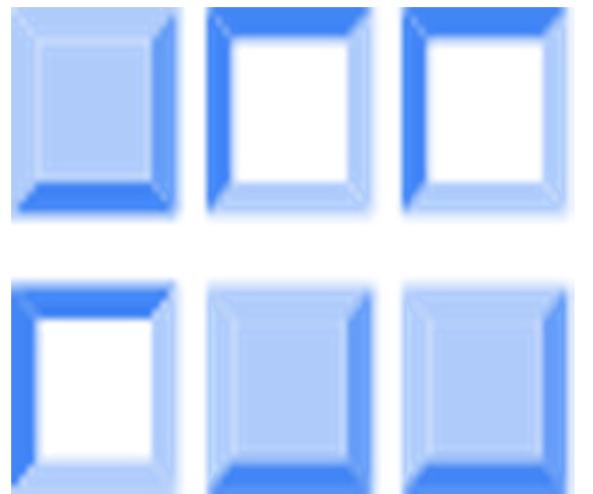
- **What type of applications can use Firestore?**
- **Live synchronization and offline data persistence mode (built-in)** makes it easy to build different types of applications
 - Multi-user, collaborative applications on mobile web
 - IoT devices
 - Live asset tracking or Activity tracking Apps
 - Real-time analytics
 - Social user profiles
 - Gaming leaderboards
 - and many more.....

Demo



Google Cloud Data Solutions

Cloud Datastore No-SQL Database



Cloud Datastore

- **Cloud Datastore:** Fully managed, highly scalable, and serverless NoSQL Database
- **Firestore** is the next generation of Datastore
 - For all **new applications** start using Firestore directly
 - Existing Cloud Datastore databases **automatic upgradation to Firestore** started from 2021 onwards.
- **Schema less database**
 - Easy to manage **underlying data structure** as application evolves
- **Scales automatically** to handle applications load
- **Strong Consistency**
 - **ACID Transactions**
 - To ensure the integrity of data we can execute **multiple datastore operations in a single transaction with ACID characteristics**, so all the grouped operations either **succeed** or **all fail**.
 - SQL-like Queries

Highly recommended to use Firestore

Cloud Datastore

- **Replication**
 - Automatically handles sharding and replication
 - Automatic regional replication with 99.99% SLA availability
 - Automatic multi-regional replication with 99.999% SLA availability
- **What type of applications can use Datastore?**
 - **Web and Mobile Applications:** To store user profiles, preferences and other application data
 - **User Authentication and Authorization:** To store user authentication and authorization information, such as user credentials, access control lists (ACLs), and session data for secure and scalable user management systems
 - **E-commerce Platforms:** To store product catalogs, inventory information, and user shopping carts, ensuring fast and reliable access to data for online transactions.
 - IOT Platforms
 - CRM Applications
 - Gaming Leader Boards
 - And many more

Highly
recommended
to use Firestore

Demo



Google Cloud Data Solutions

Cloud Bigtable No-SQL Database



Google Cloud Bigtable

- **Cloud Bigtable:** Fully managed, key-value, wide-column NoSQL database
- Optimized for high reads/writes per second
- Primarily used for applications that need low latency and high throughput
- **What type of applications can use Cloud Bigtable?**
 - Ideal for apps that need fast access to structured, semi-structured, or unstructured data
 - IOT Apps: It is needed for real-time processing and analysis of streaming data from sensors and devices.
 - Financial Trading Apps: Its ability to process and retrieve large datasets in real-time is crucial for making quick and informed trading decisions.
 - HPC (High Performance Computing) Apps: Batch Analytics, Training ML Models

Google Cloud Bigtable

- **Highly scalable database**
 - Supports **Horizontal scaling**
 - Can process **reads and writes equally** on each node
 - **Automatic scaling** based on **CPU and Storage** utilization
- **Multi-region deployments with automatic replication**
 - From a single zone up to **8 regions** at once
 - Database is protected against **region failures**
 - Provides **99.999%** availability
- **Easy migrations from other NoSQL Databases**
 - **Live migrations:** enables faster and simple onboarding by ensuring accurate data migration with reduced effort
 - **HBase Bigtable replication library:** allows for **no-downtime** live migrations
 - **Dataflow templates:** simplify migrations from Cassandra to Bigtable.

Google Cloud Bigtable

- **Maintenance**
 - Replication is automatic
 - Maintenance is automatic with **Zero downtime**
- **Rich Application and Tool support**
 - We can build **data-driven applications faster** with seamless integration with
 - Apache Spark, Hadoop, GKE, Dataflow, Dataproc, and BigQuery

Google Cloud BigTable

- **BigTable in Real-World**
 - **AdTech and retail**
 - Google [ad personalization](#)
 - Home Depot delivers [personalized experiences](#)
 - OpenX serves over [150 billion ad requests per day](#)
 - **Media**
 - Youtube
 - Twitter's [ad engagement analytics](#)
 - Spotify serves [music recommendations](#)
 - **Time series and IOT**
 - Cognite to manage [industrial time series data](#)
 - Ecobee improved performance by 10x migrating [smart home data](#)
 - **Machine Learning**
 - Tamr and Discord uses Bigtable to deliver [ML-driven experiences](#)
 - Credit Karma uses Bigtable to make [60 billion predictions](#) per day

Demo



Google Cloud Data Solutions



BigQuery Enterprise Data warehouse

Google Cloud: BigQuery

- **BigQuery:** BigQuery is a serverless and cost-effective enterprise data warehouse
- It can access and analyze data across clouds
- **Autoscaling:** It can scale as data grows
- Easily apply built-in machine learning (ML) to all data types using simple SQL
- **Duet AI in BigQuery**
 - It provides contextual code assistance for writing SQL and Python
 - It auto-suggests functions, code blocks, and fixes
- **Query any type of data:** structured, semi-structured and unstructured
- **Data Security**
 - Fine-grained governance control down to column level and row level
 - Data is encrypted at rest and in transit by default

Google Cloud: BigQuery

- **Real-time analytics with streaming data pipelines**
 - Natively integrated with streaming products like [Dataflow](#)
 - [Ingest streaming data and make it immediately available](#) for query
- **Built-in Business Intelligence (BI)**
 - Analyze large datasets interactively with [BigQuery BI Engine](#), an in-memory analysis service that offers [sub-second query response time and high concurrency](#).
- **BigQuery Omni**
 - **Cross Cloud Data Analytics:** We can run [BigQuery analytics](#) on data stored in [Amazon S3](#) or [Azure Blob Storage](#)
 - We can prevent [unnecessary duplication of data](#) for cost reduction and sustainability reasons

Dataflow Job Templates

Process Data Continuously (stream)

Bigtable Change Streams to HBase Replicator

Cloud Bigtable Change Streams to BigQuery

Cloud Bigtable Change Streams to PubSub

Cloud Bigtable change streams to Cloud Storage

Cloud Spanner change streams to BigQuery

Cloud Spanner change streams to Cloud Storage

Cloud Spanner change streams to Pub/Sub

Google Cloud: BigQuery Studio

- **BigQuery Studio** : Single UI for all data teams
- Simplifies [analytics workflows](#) like
 - Data ingestion
 - Preparation to [data exploration](#)
 - Visualization to [ML model creation and use](#)
- Collaborative workspace that helps [accelerate data](#) to AI workflows
- We can use [SQL](#), [Python](#), [Spark](#) or [natural language](#) directly within BigQuery and leverage those code assets [easily across Vertex AI and other products](#) for specialized workflows
- Best practices such as [CI/CD](#), [version history](#) and [source control](#) to analytics assets
- **Unified security and governance**
 - Admins can [uniformly enforce security policies](#) for data assets
 - [Unified credential management](#) across BigQuery and Vertex AI

Demo



Google Cloud Data Solutions



Dataproc

Open Source Data Analytics

Google Cloud: Dataproc

- **Dataproc:** Used to run **open source data analytics** at scale, with enterprise grade security
- It is a **fully managed and highly scalable** service for running
 - Apache Hadoop,
 - Apache Spark,
 - Apache Flink,
 - Presto, and 30+ open source tools and frameworks
- We can create **Dataproc Hadoop clusters** on
 - Compute Engine
 - Google Kubernetes Engine
- In addition, we can use **Serverless** to run **Spark batch workloads** without provisioning and managing a cluster.

Google Cloud: Dataproc

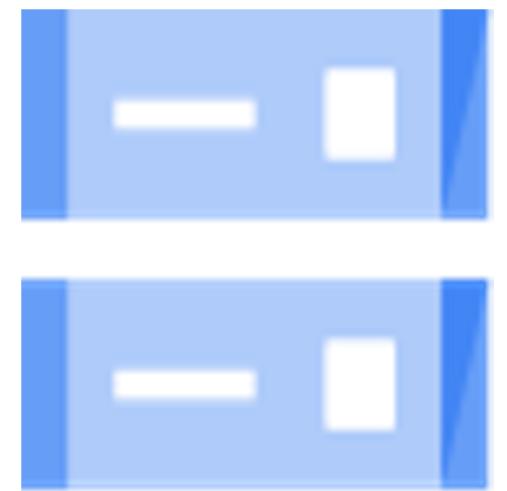
- **Autoscaling Serverless Spark:**
 - can autoscale **without any** manual infrastructure provisioning or tuning
- **Autoscaling Dataproc Clusters:**
 - **Predefined autoscaling policies** available that can apply to cluster
 - Never scale down (Scales up to meet demands)
 - Spark with dynamic allocation (default option)
 - Spark without dynamic allocation (scales rapidly)
 - Map Reduce (Scales gradually)
 - Tez (Scales rapidly)
 - We can automatically **add or remove cluster nodes** based on Autoscaling policy associated to cluster

Demo



Google Cloud Data Solutions

Cloud Storage Object Storage



Google Cloud Storage

- **Cloud Storage:** Fully managed **Object Storage service** to store **unstructured data**
- It can store **any type** of data (text, binary anything)
- It can store **any amount** of data (petabytes)
- We can retrieve data as **many times** as needed
- Low latency (time to first byte typically **tens of milliseconds**)
- Worldwide **accessibility** and worldwide **storage locations**
- **Durability**
 - Very very high durability (99.99999999% annual durability).
- **Redundancy**
 - Redundant **across regions** if the data is stored in a **multi-region or dual-region**

Google Cloud Storage - Buckets

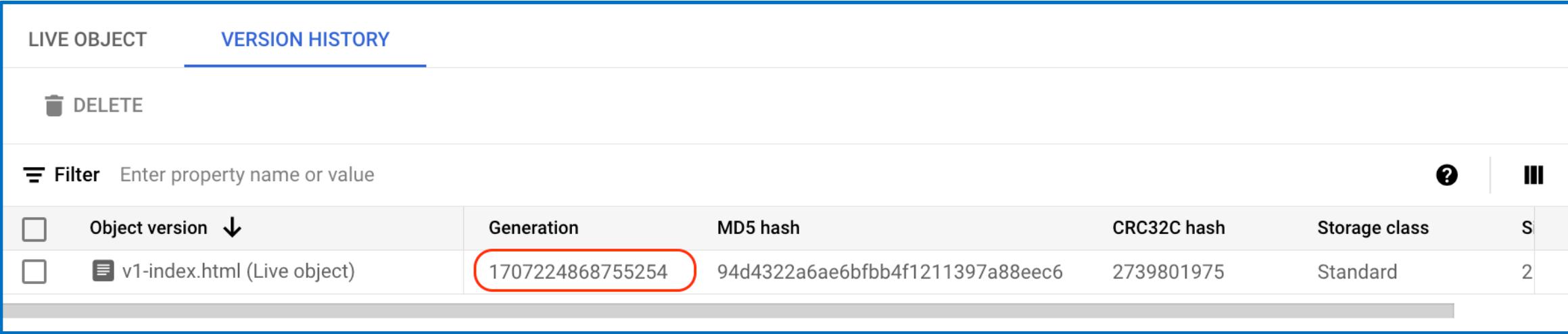
- **Cloud Storage Bucket**

- Bucket name should be **globally unique** on google cloud
- Buckets are associated to **projects**
- **No limit** to the number of buckets we can have in a project
- We cannot change the **name and location** of bucket once created
- We can use **Cloud IAM** to control access to buckets
- Once the bucket is deleted, **name can be re-used** by anyone for new bucket
- **Bucket Names**
 - Must start and end with number or letter, lowercase letters, numeric, dashes, underscores and dots are allowed.
 - 3-63 characters
 - **Spaces are not allowed**
 - Bucket names cannot begin with the “**goog**” prefix or “**google**” or close misspellings, such as “**g00gle**”

Google Cloud Storage - Objects

- **Object in Cloud Storage Bucket**

- **No limit** on the number of objects that we can create in a bucket
- Minimum Object Size: **No minimum size**, Maximum Object Size: **5 TiB**
- Object size is **limited to 5TiB** but Cloud Storage Bucket size is **unlimited**.
- Objects have two components
 - **Object Data**: actual data
 - **Object Metadata**: name-value pairs that describe **various object qualities**.
 - Two **important values** from Object Metadata which can be used to **uniquely identify Object**
 - **Object Name**: name of the object
 - **Generation Number**: **Auto-generated** by cloud storage



LIVE OBJECT		VERSION HISTORY			
		Generation	MD5 hash	CRC32C hash	Storage class
<input type="checkbox"/>	Object version ↓	1707224868755254	94d4322a6ae6bfbb4f1211397a88eec6	2739801975	Standard
<input type="checkbox"/>	v1-index.html (Live object)				2

Google Cloud Storage - Accessibility

- **Accessibility (To manage buckets and objects in a bucket)**

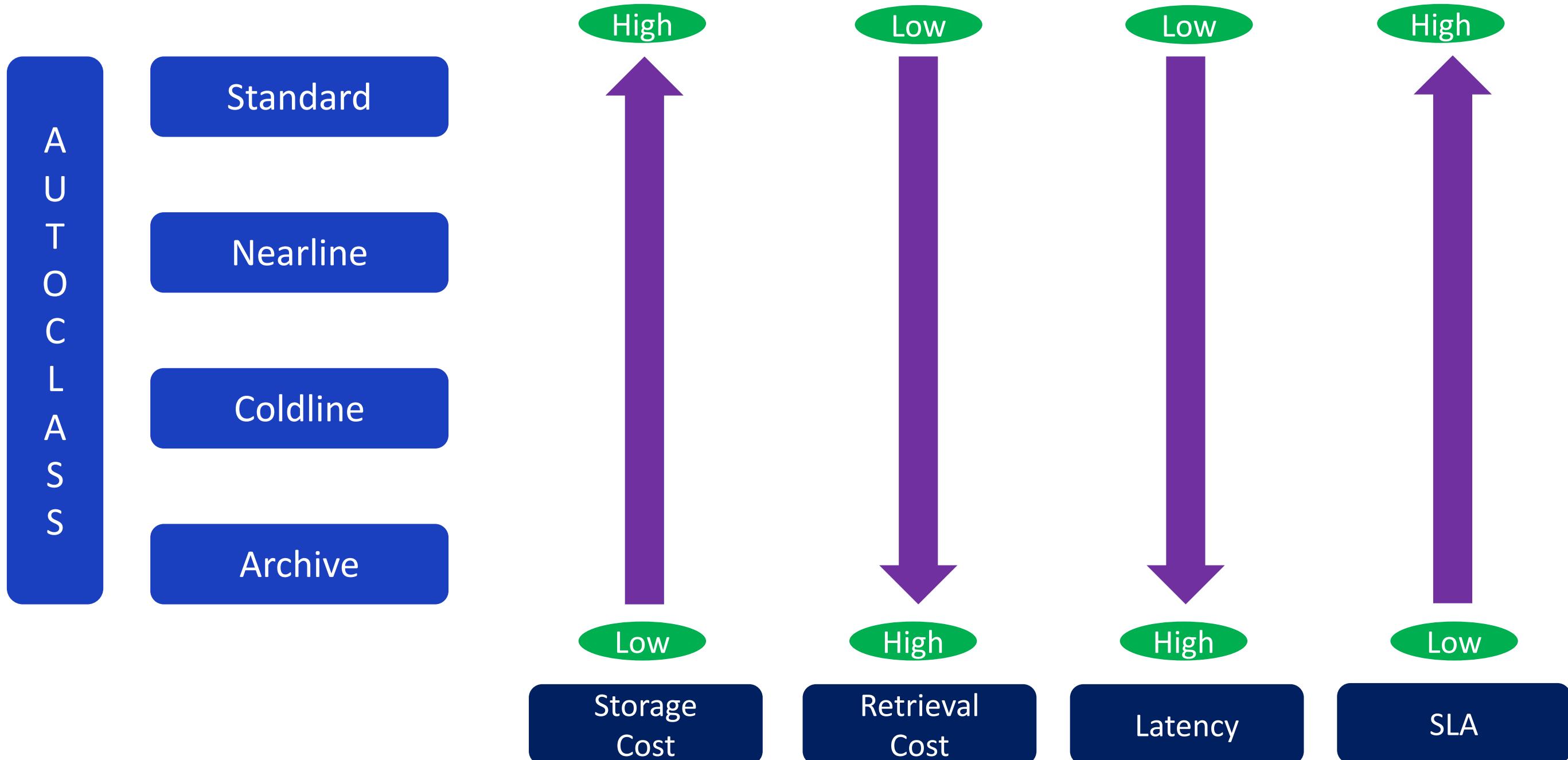
- Google Cloud Web console
- REST API
- Client libraries (Java, Node.js, PHP, Python, Ruby, C++ and C#)
- **CLI: gsutil (old cli – Not recommended)**
- **CLI: gcloud storage (latest and recommended)**
 - File transfers are **faster** when compared to gsutil
 - Uses **faster hashing tools** for data integrity checks
 - It utilizes a **new parallelization strategy**, which allows more work to be done in parallel with far less overhead
 - Automatically **detects optimal settings** and **speeds up transfers** without requiring any flags from the users
 - All operations happen in **parallel**
 - **How to transition existing gsutil scripts ?**
 - Existing gsutil scripts **can be executed as gcloud storage** to avail **performance benefits**
 - **SHIM:** Set `use_gcloud_storage=True` in the `.boto` config file under the [GSUtil] section

```
[GSUtil]
use_gcloud_storage=True
```

Cloud Storage - gsutil and gcloud storage CLI

- **CLI: gcloud storage (LATEST AND RECOMMENDED)**
 - **CREATE:** gcloud storage buckets `create gs://BUCKET_NAME`
 - **LIST:** gcloud storage buckets `list`
 - **COPY(upload):** gcloud storage `cp local/*.html gs://BUCKET_NAME/myapp1`
 - **MOVE:** gcloud storage `mv gs://BUCKET_NAME/myapp1 gs://BUCKET_NAME/myapp2`
 - **DELETE:** gloud storage buckets `delete gs://BUCKET_NAME`
 - **Reference:** <https://cloud.google.com/sdk/gcloud/reference/storage>
- **CLI: gsutil (OLD AND NOT RECOMMENDED)**
 - **CREATE:** gsutil `mb gs://BUCKET_NAME`
 - **LIST BUCKETS:** gsutil `ls`
 - **LIST OBJECTS IN BUCKET:** gsutil `ls gs://BUCKET_NAME`
 - **Reference:** <https://cloud.google.com/storage/docs/gsutil/commands/help>

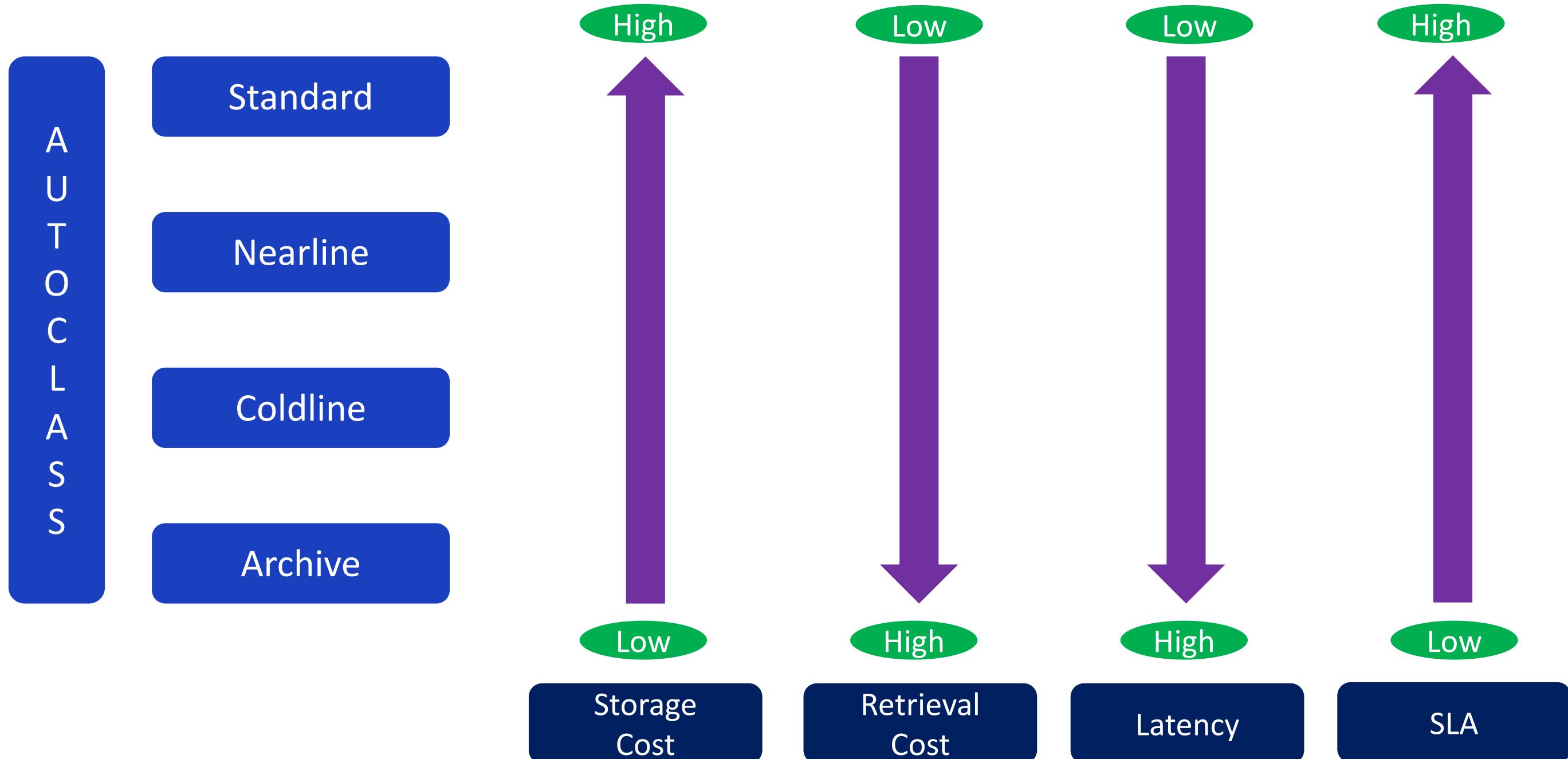
Cloud Storage - Storage Classes



Google Cloud Storage - Storage Classes

Storage Class	Description	Minimum Storage Duration	Retrieval Fees	Typical Monthly Availability
Standard	1. Data that is frequently accessed (hot data) 2. Can be used for websites, streaming videos, and mobile apps	None	None	<ul style="list-style-type: none"> >99.99% in multi-regions and dual-regions 99.99% in regions
Nearline	Low cost, infrequently accessed data (accessed once in a month)	30 days	Yes	<ul style="list-style-type: none"> 99.95% in multi-regions and dual-regions 99.9% in regions
Coldline	A very low cost , infrequently (accessed data once in a quarter)	90 days	Yes	<ul style="list-style-type: none"> 99.95% in multi-regions and dual-regions 99.9% in regions
Archive	1. The lowest cost (accessed once in a year) 2. Data is available in milliseconds when needed (Complete contrary to other cloud providers which takes hours to days)	365 days	Yes	<ul style="list-style-type: none"> 99.95% in multi-regions and dual-regions 99.9% in regions

Cloud Storage - Storage Classes



Cloud Storage - Storage Classes

- **Autoclass:** Automatically transitions each object to Standard or Nearline class based on object-level activity
- **How does Autoclass work ?**
- All data stored in **Standard Class** for **first 30 days**
- Data **that hasn't been accessed for 30 days** will transition to **Nearline storage class**
- **Colder** data that gets **accessed** will transition back to **Standard class**
- Objects **smaller than 128KB** will be **excluded** from Autoclass management and will always remain in Standard class
- Primarily used to optimize for **cost and latency**
- Recommended when **usage frequency is unpredictable**
- Can be changed to a **default class** at any time

Autoclass ?

Automatically transitions each object to Standard or Nearline class based on object-level latency. Recommended if usage frequency may be unpredictable. Can be changed to [Coldline](#).

Set a default class

Applies to all objects in your bucket unless you manually modify the class per object. If usage is highly predictable. Can't be changed to Autoclass once the bucket is created.

Standard ?

Best for short-term storage and frequently accessed data

Nearline

Best for backups and data accessed less than once a month

Coldline

Best for disaster recovery and data accessed less than once a quarter

Archive

Best for long-term digital preservation of data accessed less than once a year

Default storage class	Managed with Autoclass ?
Included classes	Standard, Nearline ?

Cloud Storage - Storage Classes

- We can also transition to **Coldline** and **Archive classes**
- Data in Coldline and Archive classes will have a **lower availability SLA**
- Enabling this option may lead to **increased latency of data reads**

Choose a storage class for your data

A storage class sets costs for storage, retrieval, and operations, with minimal differences in uptime. Choose if you want objects to be managed automatically or specify a default storage class based on how long you plan to store your data and your workload or use case. [Learn more](#)

Autoclass [?](#)

Automatically transitions each object to Standard or Nearline class based on object-level activity, to optimize for cost and latency. Recommended if usage frequency may be unpredictable. Can be changed to a default class at any time. [Pricing details](#)

Opt-in to allow object transitions to Coldline and Archive classes

Default storage class

Managed with Autoclass 

Included classes

Standard, Nearline, Coldline, Archive 

Cloud Storage - Object Lifecycle Management

- **Object Lifecycle Management (OLM)**: Automates **data management tasks** in Google Cloud Storage
- **Example-1**: Automatic transition of **infrequently accessed data (not accessed for 30 days)** to a **lower-cost storage class**, such as Nearline or Coldline, based on predefined rule (OLM Rule)
- **Example-2**: Automatic **deletion of objects** after a **specified period (365 days)** or based on custom conditions defined by lifecycle management rules.
- Optimizes **storage costs** and **simplifies** data management.
- **Benefits**
 - Cost saving
 - Compliance and Governance

Cloud Storage - Object Lifecycle Management Rule

- **OLM Actions**

- **Action-1: Switch Storage Classes**

- Standard to Nearline, Coldline or Archive
- Nearline to Coldline or Archive
- Coldline to Archive

- **Action-2: Delete Objects**

- Delete objects

- **Very Important Note:**

After we add or edit an OLM rule, it may take **up to 24 hours** to take effect

After you add or edit a rule, it may take up to 24 hours to take effect.

- **Select an action**

- Set storage class to Nearline

Best for backups and data accessed less than once a month



Coldline and Archive objects will not be changed to Nearline.

- Set storage class to Coldline

Best for disaster recovery and data accessed less than once a quarter

- Set storage class to Archive

Best for long-term digital preservation of data accessed less than once a year

- Delete object

- Delete multi-part upload

Sets a time limit and removes unfinished or idle multi-part uploads

Cloud Storage - Object Lifecycle Management Rule

- **OLM Conditions**

- **Set Rule Scopes**

- Use **prefix** and **suffix** objects to filter objects by their paths

- **Set Conditions**

- We can set **single** condition or **multiple** conditions
- For the OLM Action to happen, **all the selected conditions should be satisfied**

Select object conditions

This rule will apply the action to current and future objects that meet all the selected conditions below. [Learn more](#)

Set Rule Scopes

Use prefix and suffix rule scopes to filter objects. Up to 50 prefix and 50 suffix matches per bucket, a maximum of 100 total.

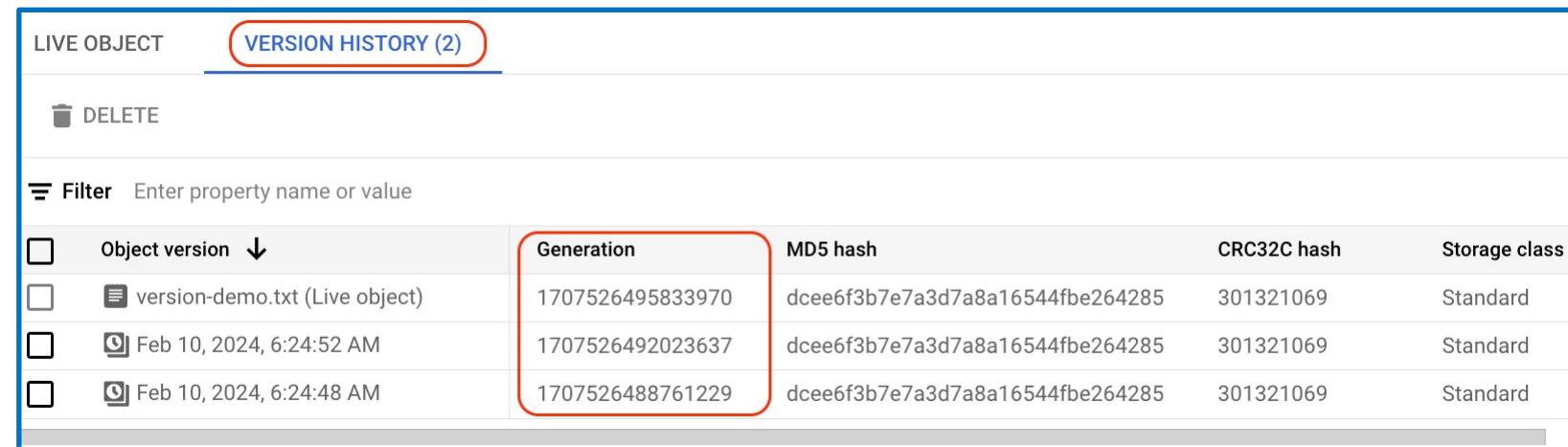
- Object name matches prefix
- Object name matches suffix

Set Conditions

- Age ?
- Created before ?
- Storage class matches
- Number of newer versions ?
- Days since becoming noncurrent ?
- Became noncurrent before ?
- Live state
- Days since custom time ?
- Custom time before ?

Cloud Storage - Object Versioning

- How to restore an object which is accidentally deleted or replaced (overwritten) ?
- **Object Versioning:** Helps to retrieve the **deleted or replaced objects**
- **Live Object:**
 - Live object is the **latest version** of object
- **Noncurrent Object:**
 - When we try to **delete or replace**, live object **will become noncurrent object version**
 - Noncurrent versions **retain** the name of the object but are **uniquely identified** by their **generation number** (Object Name + Generation Number).



The screenshot shows the 'VERSION HISTORY (2)' tab selected in the AWS S3 console. It displays three generations of an object named 'version-demo.txt'. The columns shown are Generation, MD5 hash, CRC32C hash, and Storage class.

	Generation	MD5 hash	CRC32C hash	Storage class
<input type="checkbox"/>	1707526495833970	dcee6f3b7e7a3d7a8a16544fbe264285	301321069	Standard
<input type="checkbox"/>	1707526492023637	dcee6f3b7e7a3d7a8a16544fbe264285	301321069	Standard
<input type="checkbox"/>	1707526488761229	dcee6f3b7e7a3d7a8a16544fbe264285	301321069	Standard

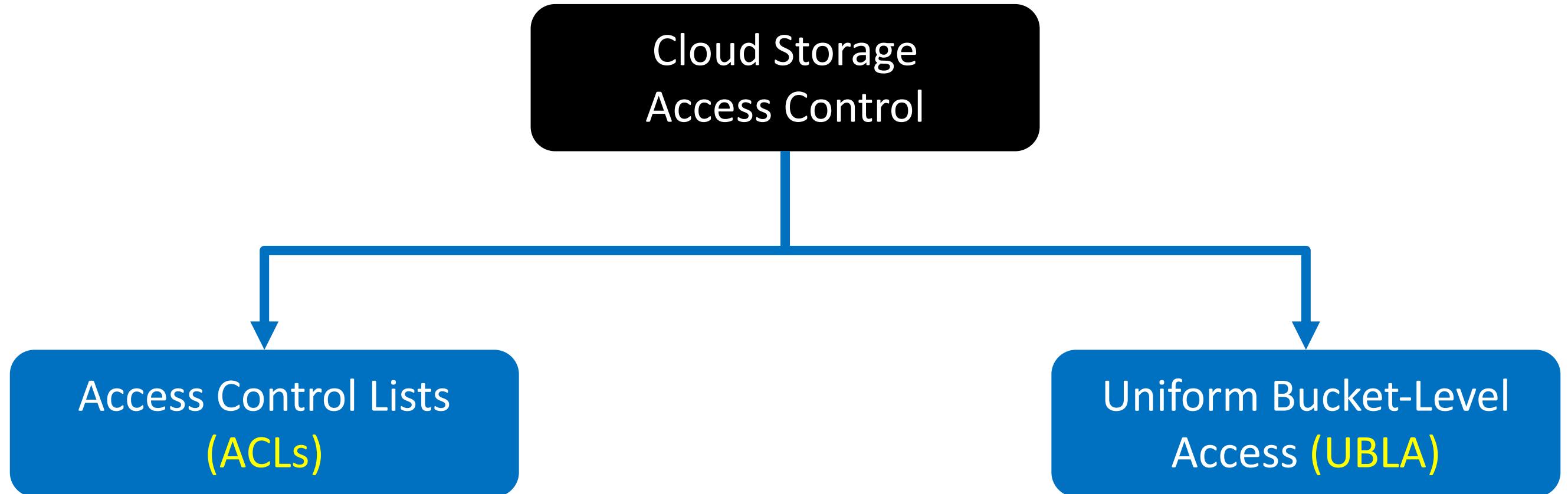
Cloud Storage - Object Versioning

- What happens when we delete a noncurrent object version ?
 - It will get deleted permanently
- We can enable or disable OBJECT VERSIONING
 - This setting is at bucket level, applies to all objects in a bucket
 - This setting can be disabled or enabled anytime
- We can enable default OLM rules to manage noncurrent object versions

Rules	ADD A RULE	DELETE ALL	
Action	Object condition	Works with	
Delete object	Object is noncurrent 2+ newer versions	Object versioning	
Delete object	7+ days since object became noncurrent	Object versioning	



Cloud Storage - Access Control



Google Cloud Storage - Access Control

Feature	Fine-grained Access Control (ACLs)	Uniform Bucket-Level Access (UBLA)
Generation	Legacy Mechanism	Newer approach, simplifies access management (HIGHLY RECOMMENDED)
How does it work?	Allows granular control over permissions on individual objects, including read, write, and delete operations.	Access to all objects within a bucket is controlled uniformly at the bucket level .
IAM Tools and Features	<p>Needs co-ordination between two different access control systems</p> <ol style="list-style-type: none"> 1. Cloud IAM 2. ACLs 	<ol style="list-style-type: none"> 1. UBLA uses Cloud IAM alone to manage permissions 2. In addition, IAM also allows us to use features that are not available when working with ACLs, such as <ol style="list-style-type: none"> 1. Managed folders 2. IAM Conditions 3. Domain restricted sharing 4. Workforce identity federation
Important Note	Not Applicable	Once we enable uniform bucket-level access, we have 90 days to switch back to fine-grained access before uniform bucket-level access becomes permanent .

Google Cloud Storage - Access Control

Feature	Fine-grained Access Control (ACLs)	Uniform Bucket-Level Access (UBLA)
Pros	<ol style="list-style-type: none"> Granular control: ACLs enable fine-grained permissions on individual objects, allowing specific access to different users or entities. Flexibility: Users can control access to individual objects independently within a bucket. 	<ol style="list-style-type: none"> Simplified management: UBLA eliminates the need to manage ACLs at the object level, reducing complexity and overhead. Consistent access control: Access permissions are applied uniformly to all objects within a bucket, ensuring consistency and easier management.
Cons	<ol style="list-style-type: none"> Complexity: Managing ACLs for large numbers of objects or buckets can become complex and challenging to maintain Inconsistency: ACLs operate at the object level, which can lead to inconsistencies and chances of exposure to security threats unknowingly Due to two Access Control Systems Coordination: <ol style="list-style-type: none"> There is an increased chance of unintentional data exposure Auditing who has access to resources is more complicated. Particularly if we have objects that contain sensitive data, such as personally identifiable information 	<ol style="list-style-type: none"> Reduced granularity: Limits the granularity of access control to individual objects when compared to ACLs. Limited flexibility: UBLA is not suitable for scenarios requiring different access permissions for individual objects within a bucket.

Cloud Storage - Concepts Quick Review - 1

Question / Scenario

You run a video streaming service and need to store large video files that are **accessed frequently**. Which storage class should you use?

Your company maintains compliance records that must be retained for seven years but are **rarely accessed once in a year after** the initial upload. Which storage class is appropriate?

You have a backup system for your database that needs to store daily backups, but they are **accessed once or twice in a month** during data loss events. Which storage class should you select?

You're a media company storing archived articles and videos, with occasional requests **once in a quarter** for access from historical researchers. What storage class would you choose?

Answer

Standard Storage Class

Archive Storage Class

Nearline Storage Class

Coldline Storage Class

Cloud Storage - Concepts Quick Review - 2

Question / Scenario

You accidentally [overwrite a critical document](#) in your cloud storage and lost it. How do you [overcome](#) such type of issues in future ?

You're a data analyst working with a large dataset stored in a cloud storage bucket, and you want to move noncurrent objects by [archiving versions older than six months](#) to a lower-cost storage class and [deleting them after one year](#). How do you achieve this ?

Your e-commerce business is expanding rapidly, and you need a reliable and scalable solution to store and manage your increasing volume of [unstructured data](#), including product images, customer data, and transaction records securely. In this scenario, which google cloud service would you consider to meet your storage needs effectively?

Answer

By enabling [Object Versioning](#) in Cloud Storage bucket

By implementing [Object Lifecycle Management Rules](#)

1. [Set Storage Class](#) to [Archive](#) for files older than 6 months
2. [Delete Objects](#) older than one year

Google Cloud Storage

Cloud Storage - Concepts Quick Review - 3

Question / Scenario

In your organization's cloud storage bucket, you have **sensitive financial documents that require exclusive access restricted to senior C-suite staff**, while all other folders and data should remain accessible to all other employees. Which feature would you utilize within Google Cloud Storage to implement **granular bucket access control** to meet these requirements effectively?

Answer

Fine-grained Access Control (ACLs)

Your company operates in a highly regulated industry where data access must be strictly controlled. You need to ensure that all objects within your cloud storage bucket are **accessible only to authorized personnel**. Which feature in Google Cloud Storage would you use to enforce uniform access control at the bucket level, ensuring that only authorized users can access any object within the bucket?

Uniform Bucket-Level Access

Cloud Storage - Concepts Quick Review - 4

Question / Scenario

Your company operates a media streaming service, where you need to store a vast collection of video files ranging from high-demand recent releases to less frequently accessed older content. However, manually managing storage classes for each file is cumbersome and time-consuming. What feature of Google Cloud Storage would you utilize to automatically optimize storage costs based on access patterns and usage, ensuring efficient storage management for your media library?

Your company is migrating a large dataset consisting of thousands of files to Google Cloud Storage. However, you're experiencing significant delays in the transfer process using the [gsutil command-line tool](#), impacting project timelines. How can you expedite the data migration process and improve performance?

Answer

Storage Class: Autoclass

1. We can use [CLI “gcloud storage”](#) which uses new parallelization strategy
2. It automatically [detects optimal settings](#) and [speeds up transfers](#) without requiring any flags from the users

Cloud Storage - Concepts Quick Review - 5

Question / Scenario

Your company has recently deployed a cloud storage bucket to store critical project data for a new initiative. However, due to changes in project requirements and organizational restructuring, you need to **rename the storage bucket and relocate it to a different geographical region**. How can you update the cloud storage bucket's name and location to align with the updated project specifications?

Answer

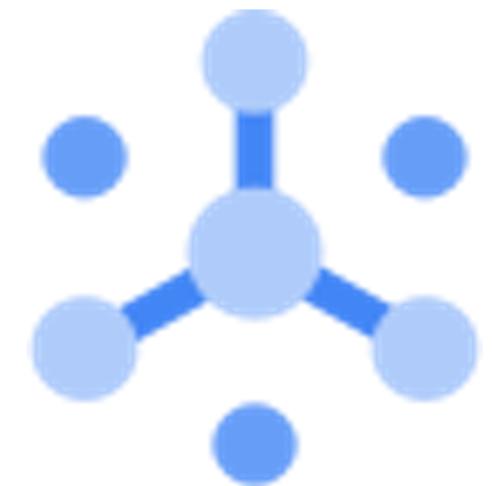
1. We **cannot** change the Cloud Storage **bucket name and location** once it is created
2. To address the updated project specifications, you would need to **create a new storage bucket** with the desired name and location and **migrate the data** from the existing bucket to the new one.

Demo



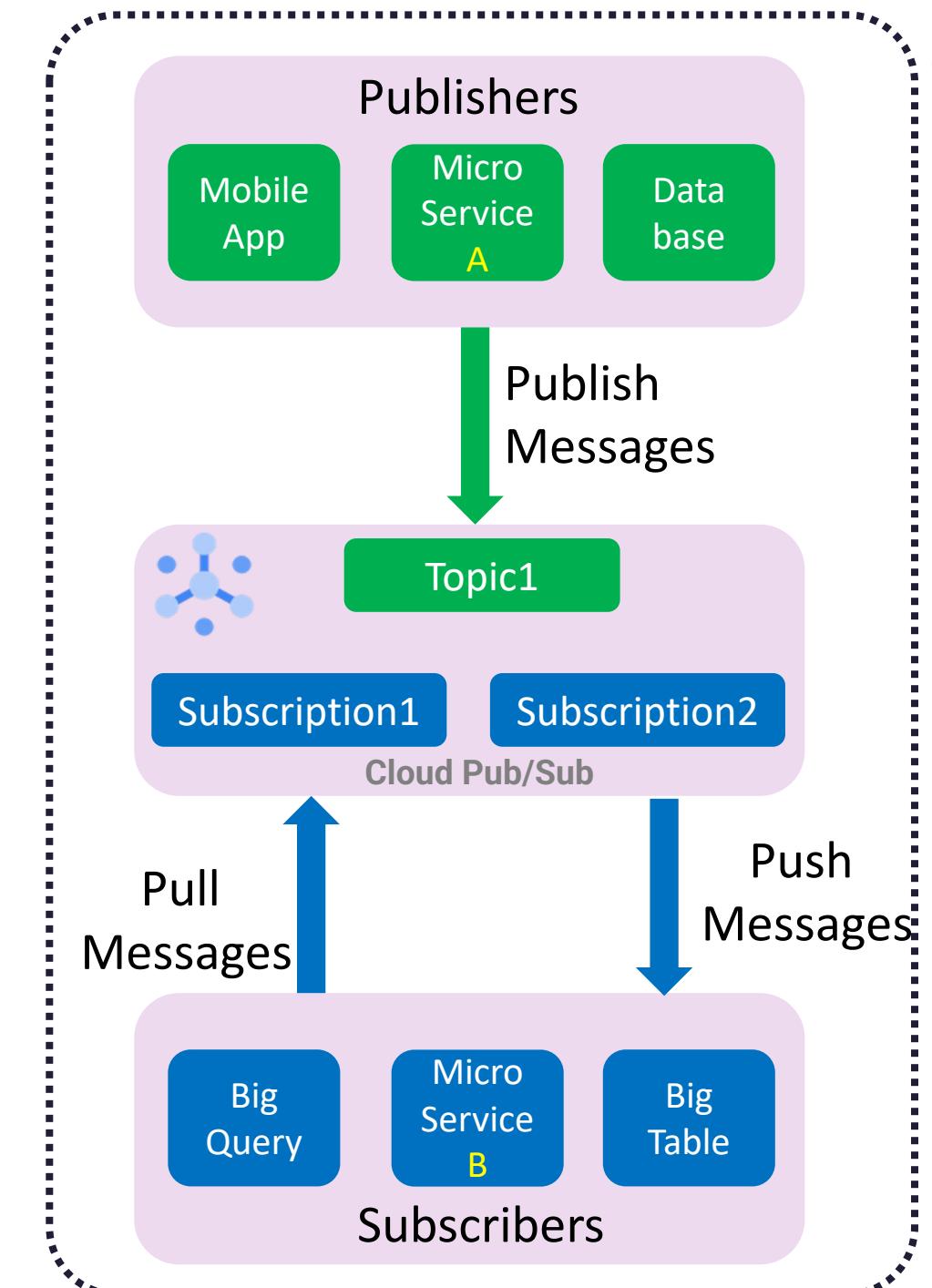
Google Cloud Data Solutions

Cloud Pub/Sub Global Real-time messaging



Google Cloud Pub/Sub

- **Cloud Pub/Sub:** Pub/Sub is **fully-managed asynchronous messaging service** designed to be highly reliable and scalable
- Google products, such as **Ads, Search, and Gmail**, send **500 million messages per second** using this infrastructure, totaling over **1TB/s of data**.
- Primarily used for **real-time data streaming** and **event-driven systems**
- **Stream Analytics (real-time data streaming)**
 - Very powerful feature
 - Ingest **analytic events of our applications** and stream them to BigQuery, with Dataflow



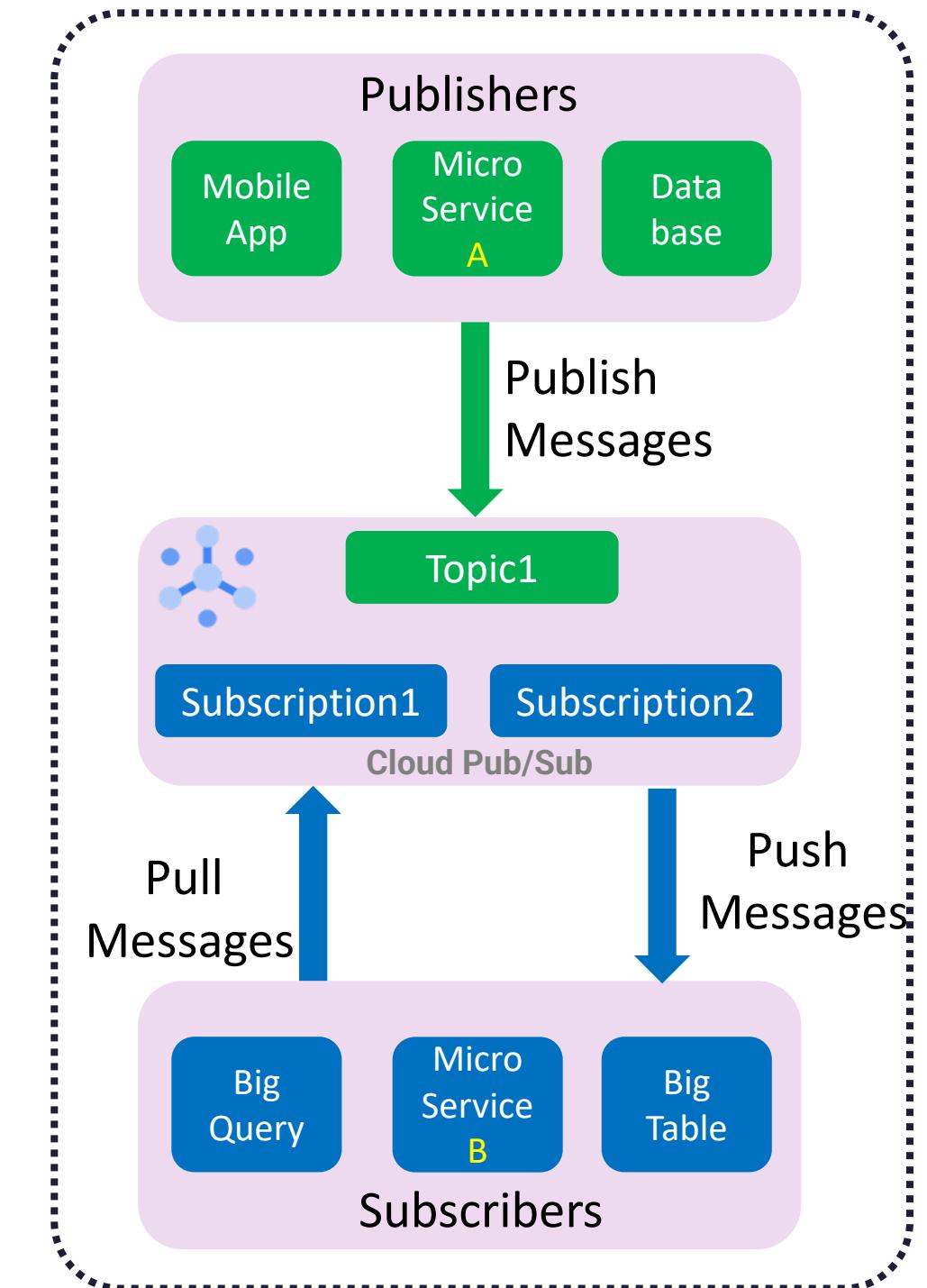
Google Cloud Pub/Sub

- **Publishers**

- Services that produce messages
- Publishers **send events to Pub/Sub Topics**, without worrying about when or how these events will be handled.

- **Subscribers**

- Services that process those messages
- Subscribers subscribe to a **Pub/Sub subscription**
- Subscriptions have the following delivery types
 - **Pull**
 - Subscriber **need to pull** the messages
 - **Push, Write to BigQuery, Write to Cloud Storage**
 - As soon as message arrived at Pub/Sub topic, **subscription will push them** to registered subscribers



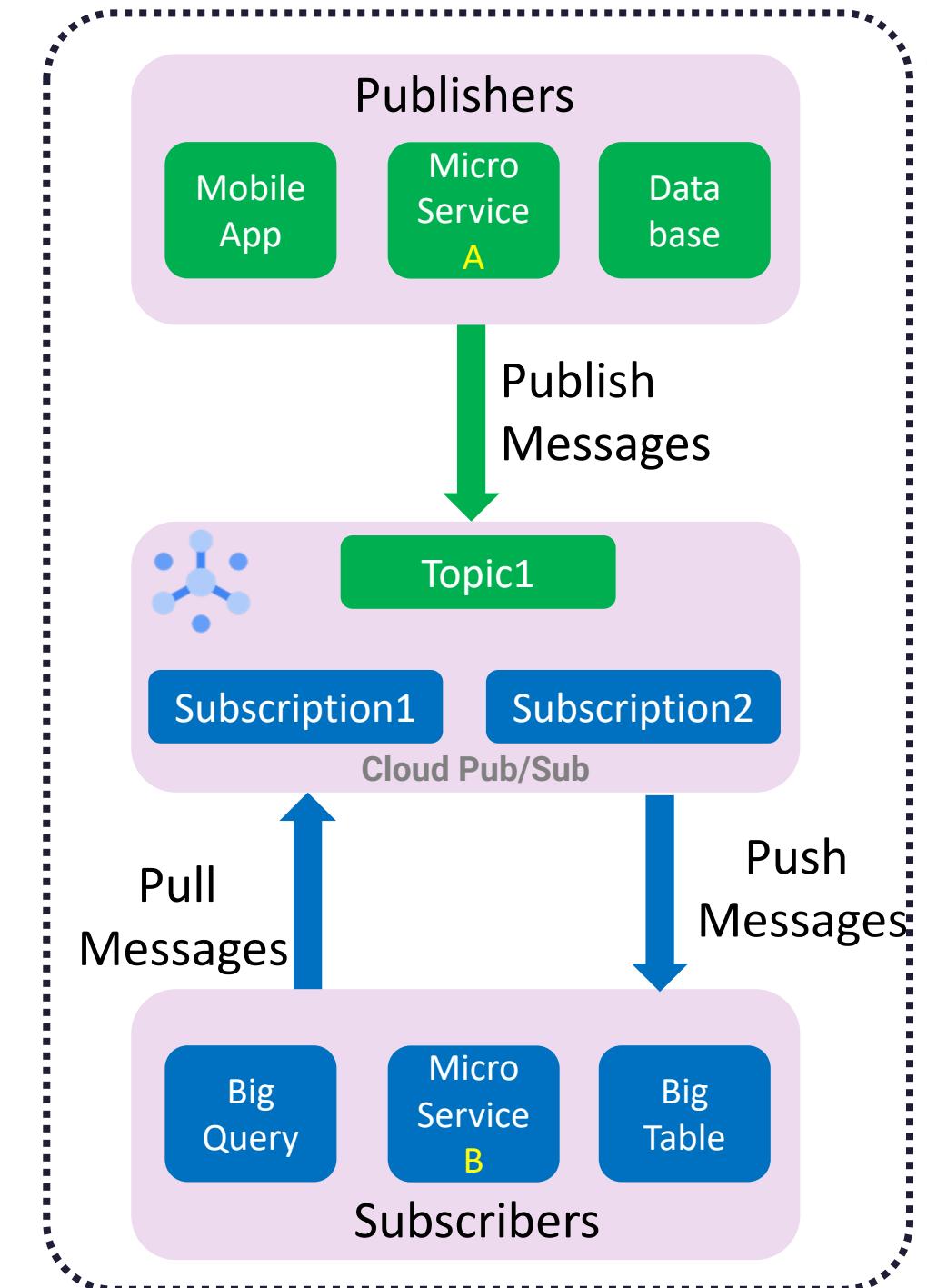
Google Cloud Pub/Sub

- **Global Service**

- Pub/Sub is a **global** service
- Topics and subscriptions are **not region-specific**
- Messages flow within the Pub/Sub service **between regions** when needed
- When using the **global endpoint** (pubsub.googleapis.com), publishers and subscribers connect to the **nearest network region** where Pub/Sub runs.
- When using the **locational endpoints** (`us-central1-pubsub.googleapis.com`), publishers and subscribers connect to Pub/Sub in the **specified region**

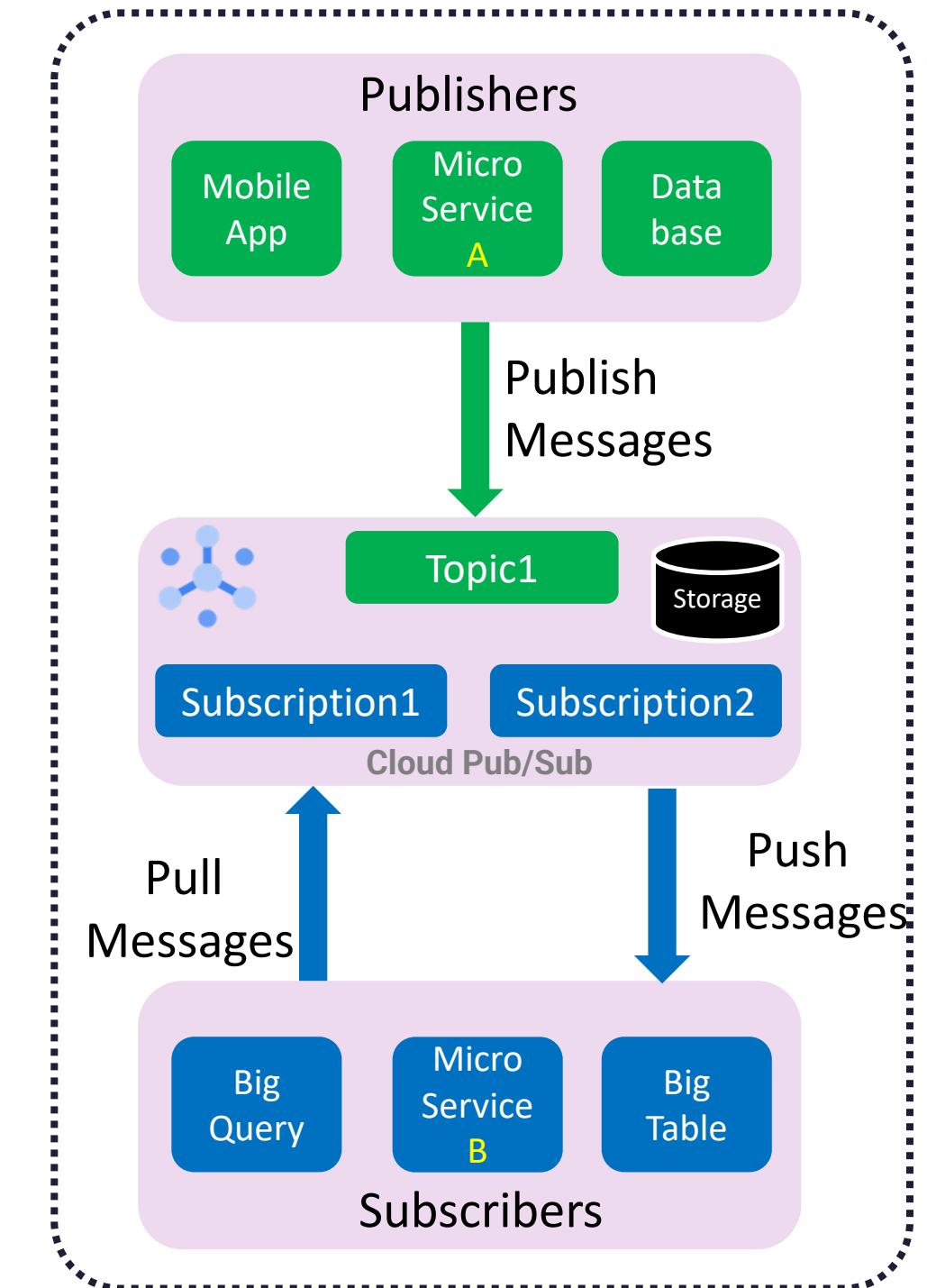
- **Autoscaling**

- Designed to **scale horizontally**
- **No provisioning**, not visible, everything happens in the background
- Auto-everything



Google Cloud Pub/Sub - Life of a Message

1. A publisher **sends a message** to Pub/Sub.
2. The message is **written to** Pub/Sub storage.
3. Pub/Sub **sends an acknowledgement** to the publisher that it has **received the message** and **guarantees** its delivery to all attached subscriptions.
4. At the **same time** as writing the message to storage, Pub/Sub **delivers it to subscribers**.
5. Subscribers send an **acknowledgement** to Pub/Sub that they have **processed** the message.
6. Once at least one subscriber for each subscription has **acknowledged** the message, Pub/Sub **deletes the message** from storage.



Google Cloud Pub/Sub

- **Compliance and Security**

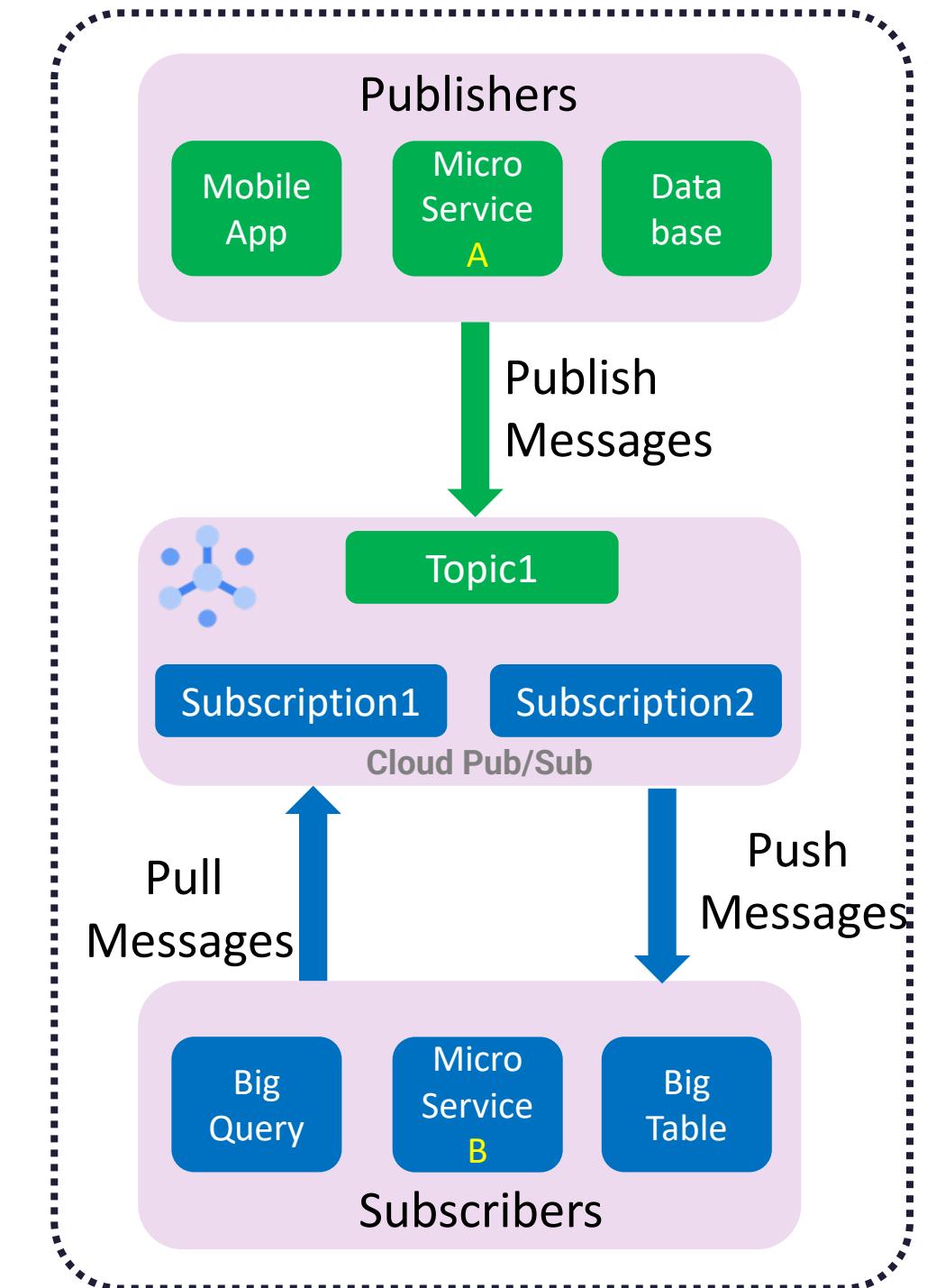
- HIPAA-compliant service
- End to end encryption
- Fine-grained access control

- **Google Cloud-native Integrations**

- Cloud Functions for serverless event-driven computing
- Dataflow (super powerful service in entire google cloud) for Stream Analytics
- Cloud Logging

- **Message Filtering**

- Subscribers will only receive messages that match the filter
- Helps reducing delivery volume to subscribers



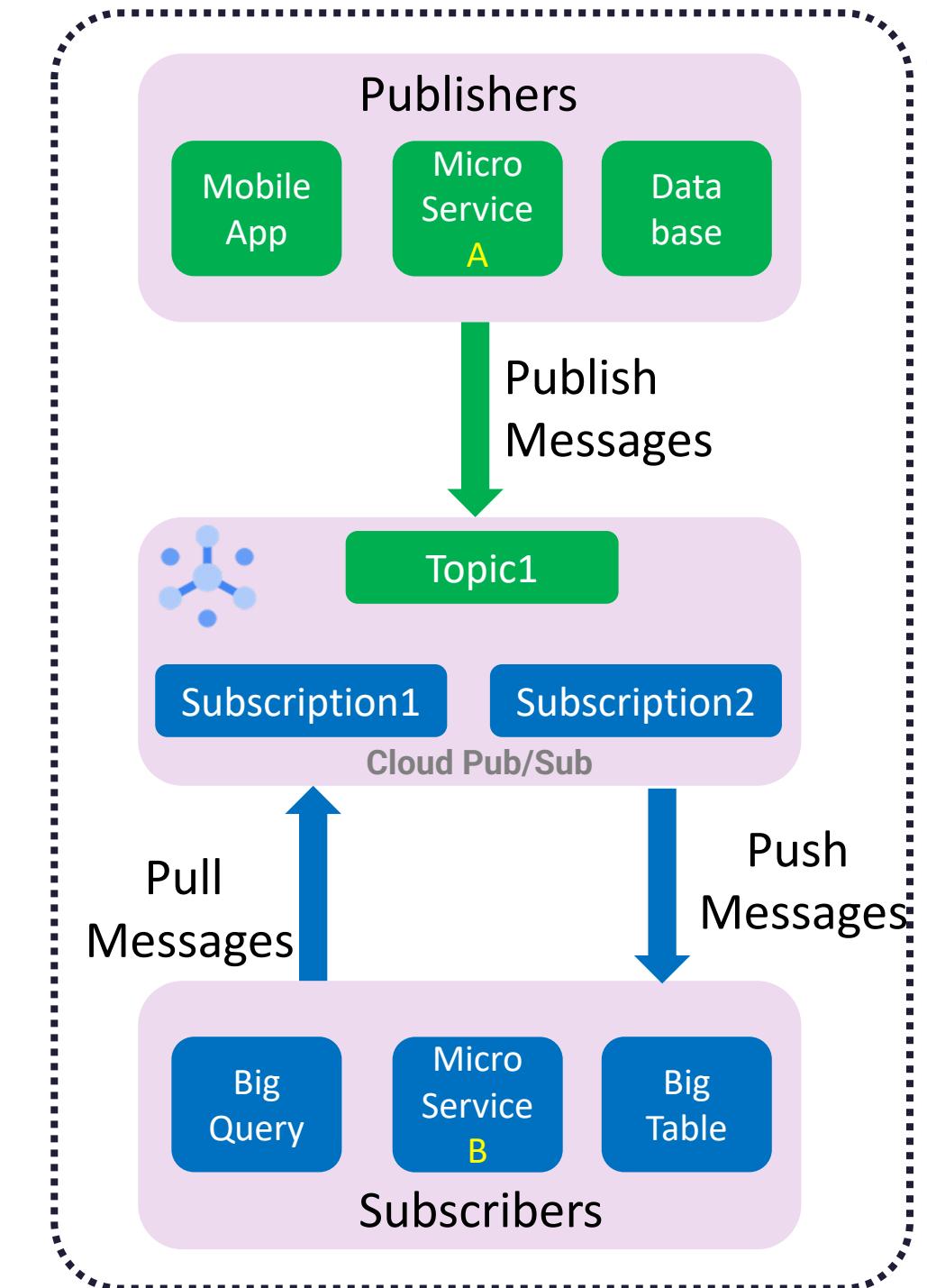
Google Cloud Pub/Sub

- **Dead Letter Topics**

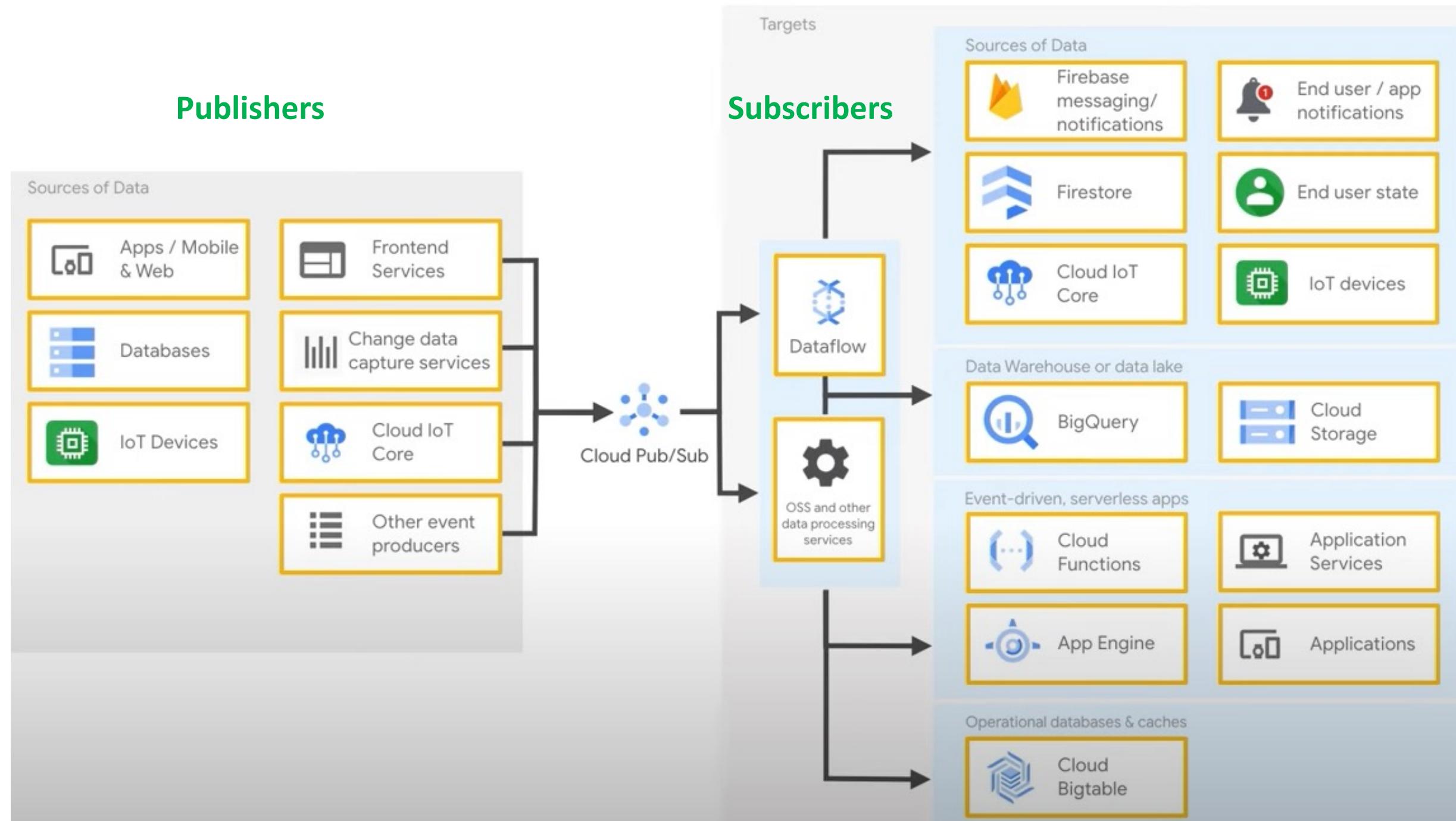
- To enable, **need to create** a Dead letter topic
- Messages **unable to be processed** are published to Dead letter topic for later review and troubleshooting
- This ensures that other messages **aren't held up** while issues are addressed.

- **Exactly once delivery**

- Messages sent to the subscription are **guaranteed not to be resent** before the message's **acknowledgement deadline expires**
- Acknowledged messages **will not be resent** to the subscription



Google Cloud Pub/Sub


 Reference: <https://cloud.google.com>

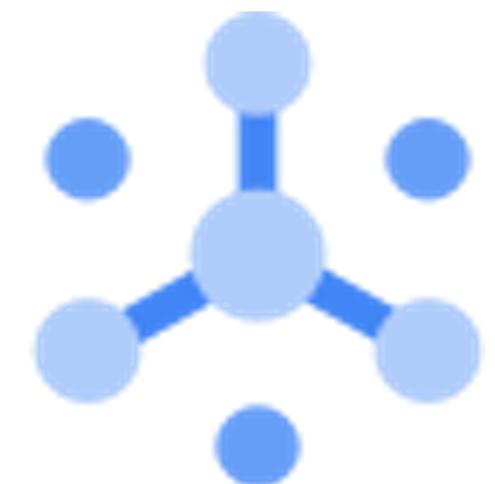
Demo



Google Cloud Data Solutions

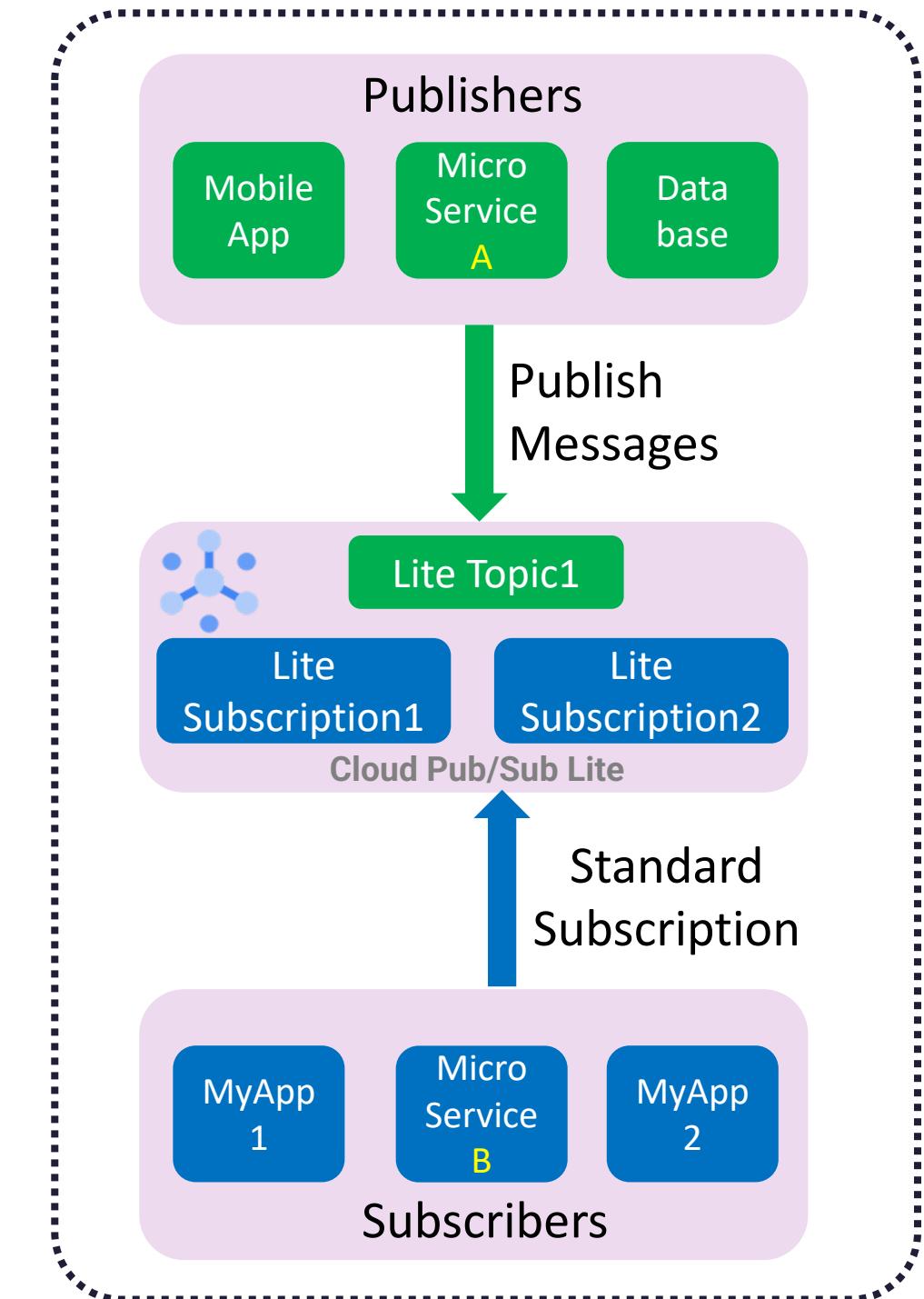
Cloud Pub/Sub Lite

Partition based Messaging Service



Google Cloud Pub/Sub Lite

- **Cloud Pub/Sub Lite:** Messaging service built for **lower cost and lower reliability**
- It is a **high-volume partition-based** messaging service
- **No global availability**
- **Manual Provisioning**
 - Manually reserve and manage resource capacity
 - Storage
 - Throughput
- We can choose **zonal or regional Lite topics**
- Regional Lite topics offer the **same availability SLA** as Pub/Sub topics



Google Cloud Pub/Sub vs Pub/Sub Lite

Feature	Pub/Sub	Pub/Sub Lite
Capacity	Automatically provisioned	Provision manually (Storage and Throughput capacity)
Pricing	Pay for what you use	Pay for the capacity that you provision (Low cost when compared to Pub/Sub)
Reliability	Very high	Low Reliability
Message Routing / Regional Availability	Global	Regional or Zonal
Message Replication	Supports synchronous replication of all data to at least two zones and best-effort replication to a third additional zone	<ol style="list-style-type: none"> Zonal Lite topics are stored in only one zone (No Replication) Regional Lite topics replicate data to a second zone asynchronously
Client Library	Java, Python, Go, Node.js, C++, C#, PHP,	Java, Python, Go
Languages Support	Ruby, SAP, ABAP	
Customer managed encryption keys	Yes	No

Google Cloud Pub/Sub vs Pub/Sub Lite

Feature	Pub/Sub	Pub/Sub Lite
Dead letter topics	Yes	No
Cross-project subscriptions	Yes	No
Exactly-once delivery	Yes	No
Message filtering	Yes	No
Message schema validation	Yes	No
REST endpoints	Yes	No
Storage	Unlimited	Unlimited
Service endpoints	Global and Regional	Regional and Zonal
Integrations	Dataflow, Cloud Functions, Cloud Logging, Kafka Connect, Apache Flink	Dataflow, Apache Spark, Apache Flink, Apache Kafka, Kafka Connect
Subscriptions	Push, Pull, BigQuery and Cloud Storage	Standard Subscription Export Subscription (export to Pub/Sub)

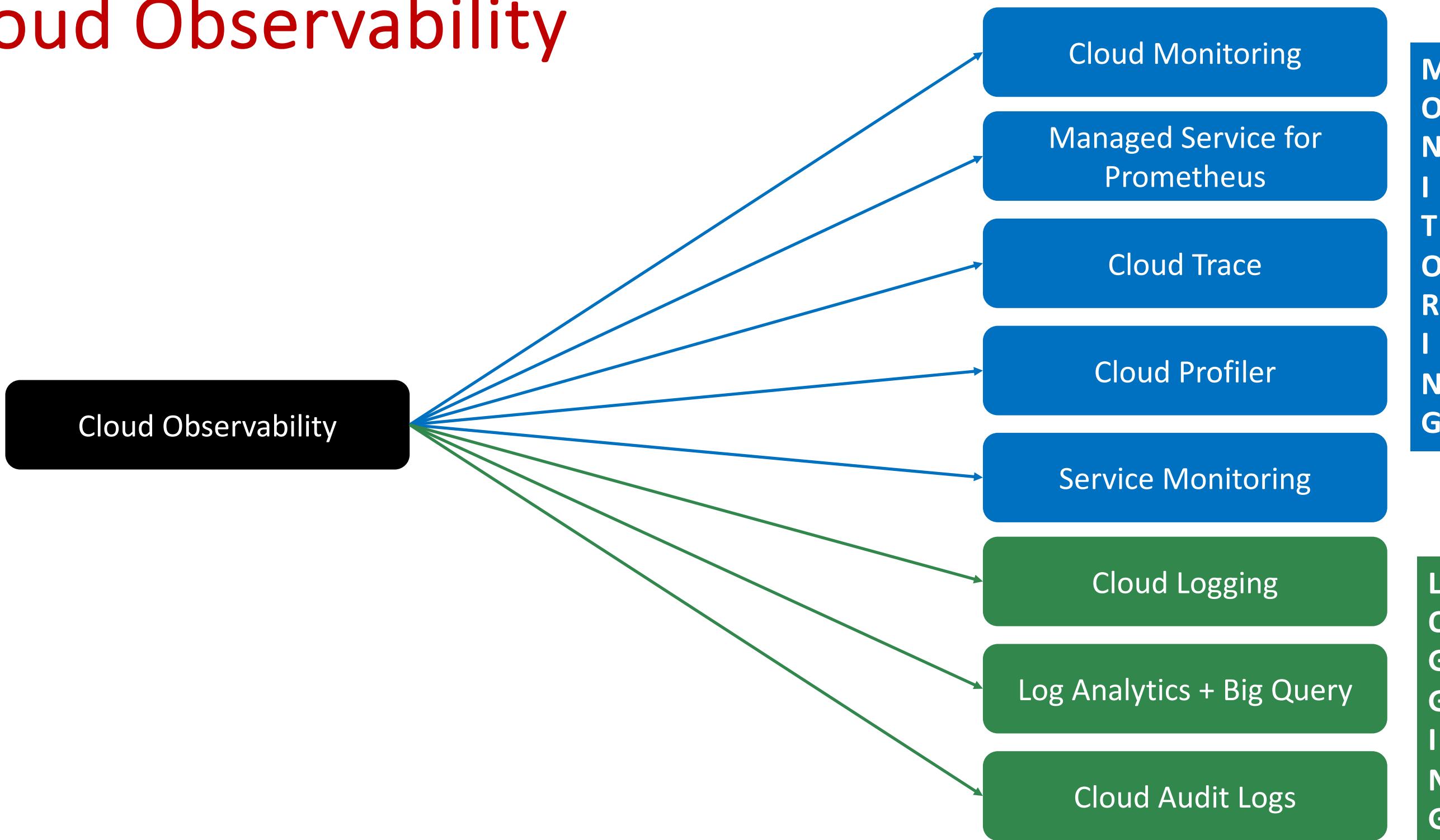
Demo



Google Cloud Observability Monitoring & Logging



Cloud Observability



Cloud Observability

- **Cloud Monitoring**
 - Monitor and alert on all google cloud services you are using and also your workloads when they are having an issue.
 - Supports various alerting mechanisms (Google Chat, Pager Duty, Slack, Webhooks, email, SMS and Pub/Sub)
- **Managed Service for Prometheus**
 - Globally monitor and alert on your workloads, using Prometheus, without having to manually manage and operate Prometheus at scale.
- **Cloud Trace**
 - Latency management solution for our workloads, primarily used for microservices to identify service to service latency
 - Development work involved to enable trace libraries (SDK) and use them in our Application code
- **Cloud Profiler**
 - Performance and cost management solution
 - Provides continuous profiling of resource consumption (Collects CPU and memory usage) in your production applications, helping you identify and eliminate potential performance issues.

Cloud Observability

- **Service Monitoring**
 - If we want to manage our services like [how google manages its own services](#), then we can use this feature
 - Primarily used for [Microservices](#) to define [SLO \(Service level Objectives\)](#)
- **Cloud Logging**
 - [End to end log management](#) for All logs (User logs, platform logs, audit logs, application logs)
- **Log Analytics**
 - Run [queries](#) that [analyze](#) your log data, and then you can [view or chart the query results](#).
 - Helps to query our logs using [SQL \(Relational queries\)](#)
 - Can [create logging datasets](#) in serverless BigQuery platform which can perform aggregation at petabyte scale
- **Cloud Audit Logs**
 - All user activity on Google Cloud Platform is logged to Cloud Audit logs.
 - [Who](#) did what
 - [When](#) they did
 - [Where](#) they did

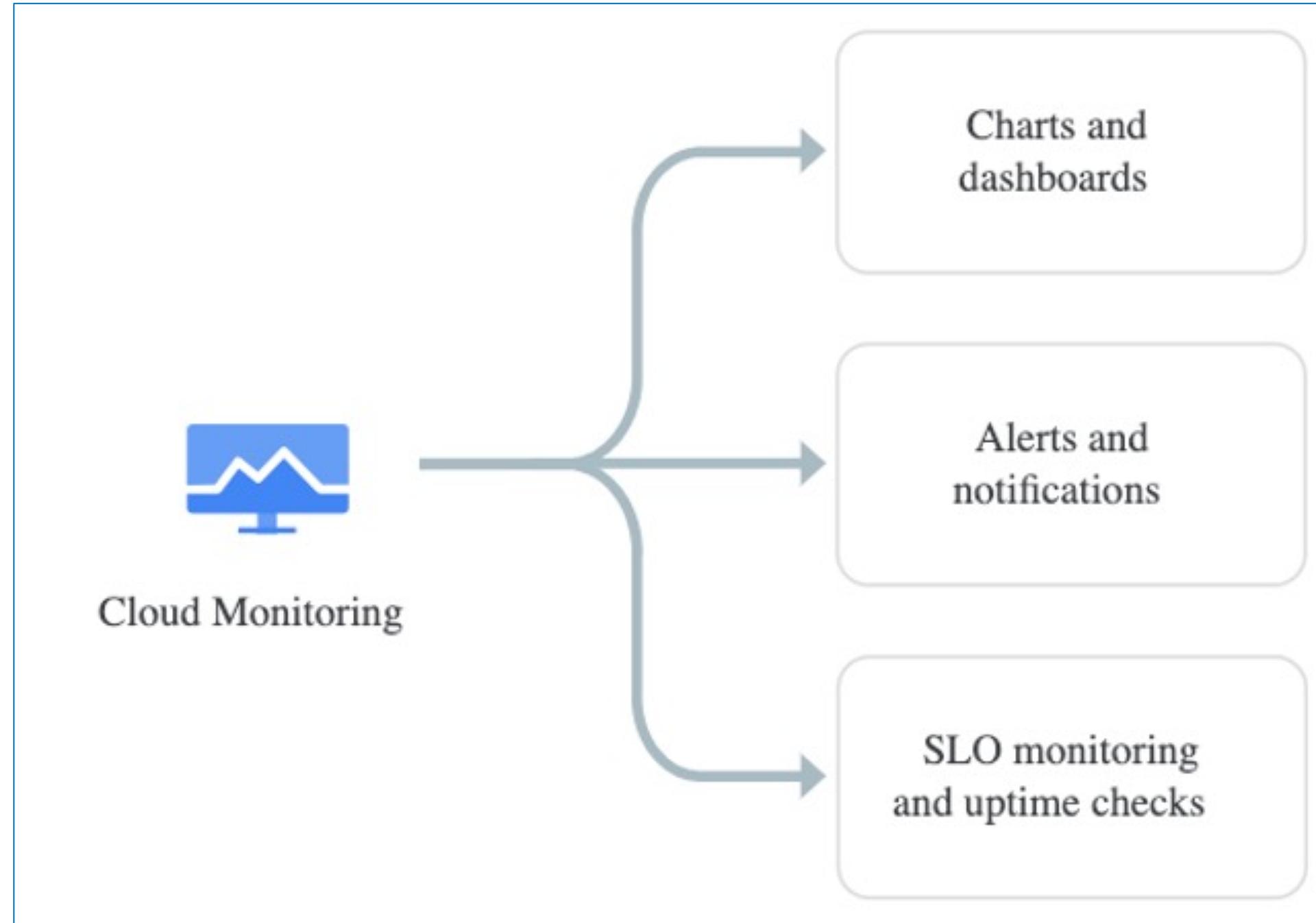
Demo



Google Cloud Observability Cloud Monitoring



Cloud Monitoring



Reference: <https://cloud.google.com/products/operations>

Cloud Monitoring - Usecase

- **Why do we need Monitoring ?**
- **Usecase:** A **simple webserver** deployed to a Compute Engine VM Instance
- **What will be our monitoring scope on a very high level in this usecase**
 - **Application Monitoring**
 - Application Uptime Checks
 - Current Connections
 - Accepted Connections
 - Requests Rate
 - **VM Instance Monitoring**
 - High CPU Utilization
 - High Memory Utilization
 - High Disk Utilization
 - Host Error log detected
 - **Notifications**
 - Send Notifications (email) **when there is an issue** (Application down, high cpu or high memory usage)

Cloud Monitoring - Alerts and Notifications

- **Alerts (Alert Policy)**
 - Create an alert policy for a scenario:
 - When my VM CPU reaches 80%
 - When my application uptime check fails
- **Notifications (Notification Channel)**
 - Notify the respective team when VM CPU reaches 80%
 - **Notification Channels:** Google Chat, Pager Duty, Slack, Webhooks, email, SMS and Pub/Sub
- **What happens with Alert Policy and Notification Channels ?**
 - **Step-1:** We create an Alert policy (when my App uptime check fails) and associate it to a notification channel (email)
 - **Step-2:** An incident will be created and an email (Notification Channel) will be sent with all the details of the incident
 - **Step-3:** We need to review and address the incident (Take necessary action)

Demo

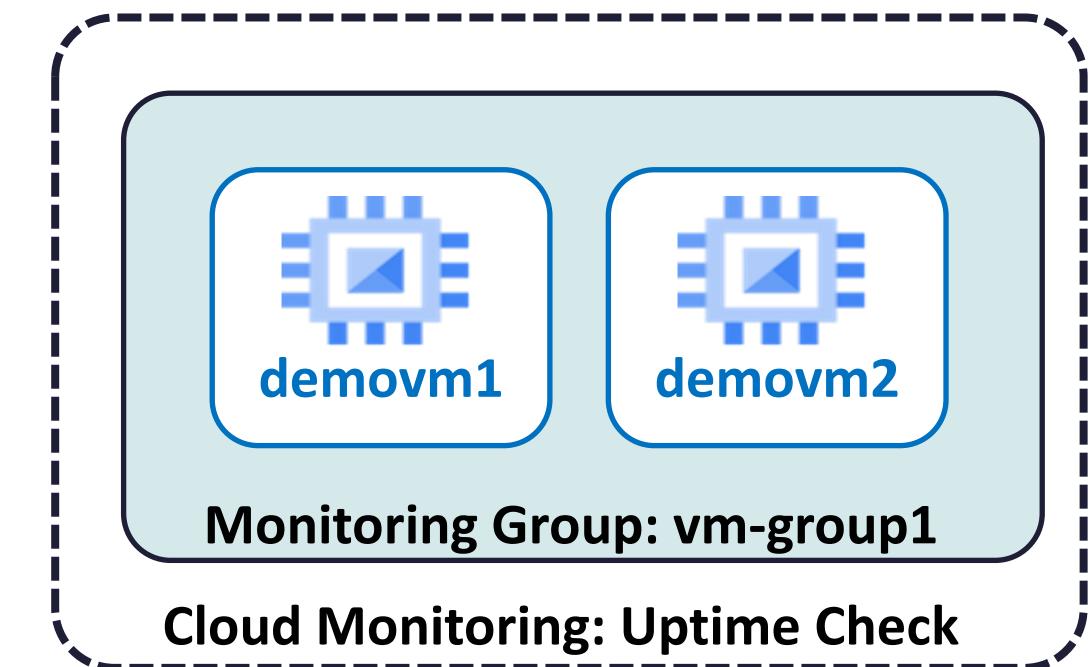


Google Cloud Observability Cloud Monitoring Groups



Cloud Monitoring - Resource Groups

- **Resource Groups:** Dynamic collection of resources monitored as a single resource
 - Compute Engine instances whose names start with the string “demovm”
 - Resources with the tag test-cluster.
- **Example:** We can configure a single uptime check monitor for all the group of VMs whose name starts with “demovm”



Demo



Google Cloud Observability Proactive Monitoring



Cloud Monitoring - Proactive Monitoring

- **Proactive Monitoring and Validation**

- We can create **synthetic monitors** to **test the**
 - Availability, consistency, and performance of your services, applications, web pages, and APIs

- **Uptime Checks**

- Let Google Cloud **periodically query an application** that responds to **HTTP, HTTPS, or TCP requests**
- Can test **public or private endpoints**
- They can **validate the response data** (Example: Some text on web page)

- **Custom and Mocha based Synthetic Monitors**

- We can deploy a **suite of tests** to test our applications
- Can be written in **desired programming language** supported by **Cloud Functions**
- **Cloud Functions trigger periodically** (every 1 or 5 or 10 minute) to test our application
- If you have access to **Gemini Code Assist** in this project, then you can provide a prompt to **generate your test code**

- **Broken-link checkers**

- Let Google Cloud **periodically test a URI** and **test a configurable number of links** found at that URI.

Cloud Monitoring - Proactive Monitoring

- We can create Proactive monitoring (uptime checks and synthetic monitors) using **various ways**

	Google Cloud console	Cloud Monitoring API	Terraform	Client libraries
Uptime checks	Y	Y	Y	Y
Synthetic monitors	Y	Y	Y	
Broken-link checkers	Y	Y	Y	

Reference: <https://cloud.google.com/monitoring/uptime-checks/introduction>

Demo



Google Cloud Observability Monitoring Data Visualization



Cloud Monitoring - Data Visualization

- **Data Visualization:** To visualize the data, we can use Dashboards
 - **Default Dashboards**
 - [Automatically created](#) when we create resources in our GCP project
 - **Custom Dashboards**
 - We can also create custom dashboards with [desired data, view and display format](#) (line graph, bar graph)
 - Your custom dashboards can [display charts, tables, logs and error groups, alerting policies and incidents](#)
 - You can also [share custom dashboards with people or groups](#) in your organization
 - **Chart Service (Metrics Explorer)**
 - Let's you quickly [visualize and explore time-series data](#)
 - The chart settings let you compare [current data to previous data](#), display multiple metrics.
 - You can also [save charts to a custom dashboard](#).

Cloud Monitoring - Data Visualization

	Name
<input type="checkbox"/>	 VM Instances
<input type="checkbox"/>	 Autoscaler Monitoring
<input type="checkbox"/>	 Disks
<input type="checkbox"/>	 Firewalls
<input type="checkbox"/>	 from-metrics-explorer
<input type="checkbox"/>	 GCE VM Instance Monitoring
<input type="checkbox"/>	 GCE VM Lifecycle Events Monitoring
<input type="checkbox"/>	 Infrastructure Summary
<input type="checkbox"/>	 Logs Dashboard
<input type="checkbox"/>	 temp101

Default Dashboards automatically created when we create resources in our GCP project

Example: We have created a VM Instance and then automatically 6 dashboards created due to that

Concept



Google Cloud Observability Monitoring Data Collection and Storage



Cloud Monitoring - Data collection and storage

- **Data Collection and Storage:** Cloud Monitoring collects and stores the following types of metric data
 - System metrics generated by Google Cloud services
 - System and application metrics that the Ops Agent collects about system resources and applications running on Compute Engine instances.
 - You can configure the Ops Agent to collect metrics from third-party plugins such as Apache or Nginx web servers, or MongoDB or PostgreSQL databases.
 - User-defined metrics that are created by using the Cloud Monitoring API or by using a library such as OpenTelemetry.
 - [External metrics][metrics-external] that are defined by some open source libraries or third-party providers.
 - Prometheus metrics that are collected by Google Cloud Managed Service for Prometheus
 - Log-based metrics that record numeric information about the logs written to Cloud Logging

Demo

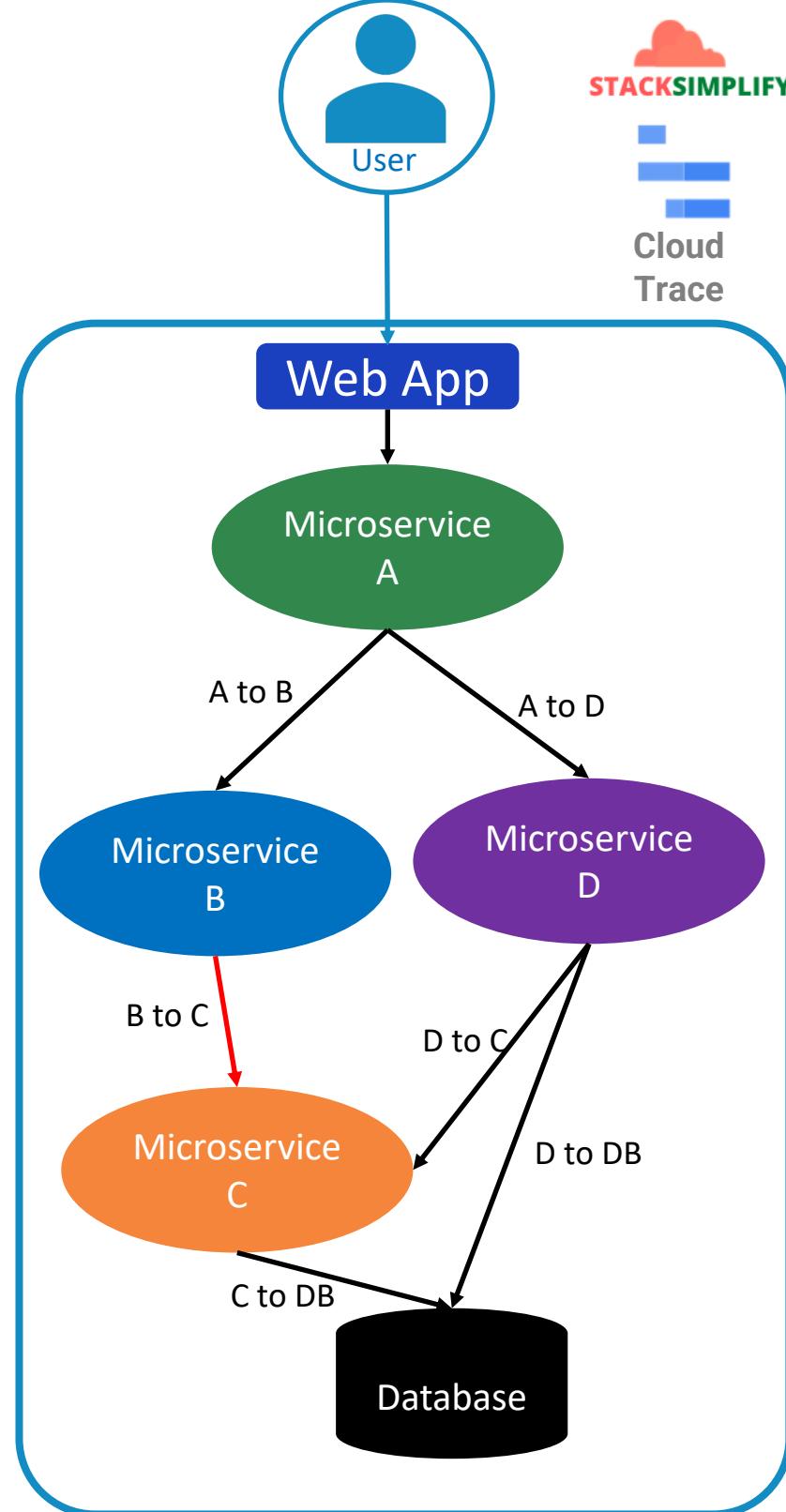


Google Cloud Observability Cloud Trace



Cloud Trace

- **Cloud Trace:** Distributed tracing system that collects latency data from your applications and displays it in the Google Cloud Console
- You can track how requests propagate through your application and receive detailed near real-time performance insights.
- **Find performance bottlenecks:** At what part of our overall application architecture we are seeing performance issues (Example: Microservice B to C)
- **Fast, automatic issue detection**
 - Trace continuously gathers and analyzes trace data and generates analysis reports
 - Cloud Trace will automatically alert you if it detects a significant shift in your app's latency profile.
- **Broad platform support**
 - Language-specific SDKs can analyze projects running on VMs (even those not managed by Google Cloud)
 - **Trace SDK** is currently available for Java, Node.js, Ruby, and Go
 - The **Cloud Trace API** lets you send latency data to, and retrieve latency data from, Cloud Trace (Reference: <https://cloud.google.com/trace/docs/reference>)



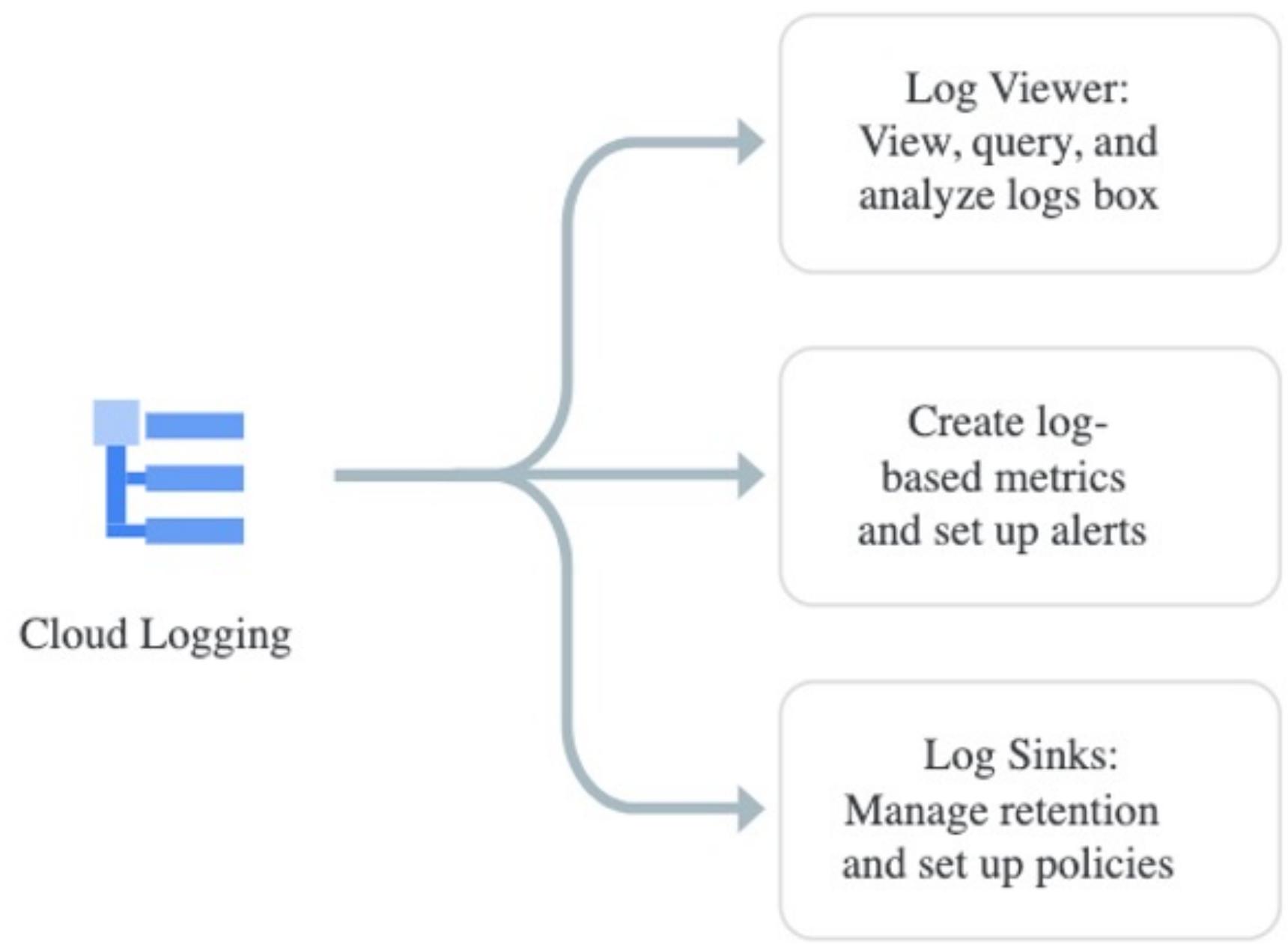
Demo



Google Cloud Observability Cloud Logging



Cloud Logging



Key Concepts

Log Explorer
(Log Viewer)

Log Analytics

Log-based Metrics

Log-based Alert Policies

Log Router (Log Sinks)

Log Storage

Reference: <https://cloud.google.com/products/operations>

Cloud Logging

- **Cloud Logging:** Real-time log-management system with storage, search, analysis, and monitoring support
- **Collect Logs:**
 - Cloud Logging automatically collects logs from Google Cloud resources
 - You can also collect logs from your applications, on-premise resources, and resources from other cloud providers
- **Monitor Logs:**
 - You can also configure alerting policies so that Cloud Monitoring notifies you if certain kinds of events are reported in your logs.
 - You can also create log-based metrics and add them to custom dashboards
- **Compliance:**
 - For regulatory or security reasons, you can determine where your log data is stored using Log Router (Log sinks).

Cloud Logging - Log Explorer

- **Troubleshoot and Analyze Logs:** You can view and analyze logs using [Logs Explorer](#) and [Log Analytics](#) in google cloud console

- They use [different query languages](#)
 - [Log Explorer](#): Logging Query Language
 - [Log Analytics](#): SQL
- They have [different capabilities](#).

- **Log Explorer**

- You use Log Explorer to [troubleshoot and analyze the performance](#) of your services and applications
- This interface is designed to let you view [individual log entries](#) and find related log entries
- Uses [Logging Query Language](#) (Simple and Straight forward) (Reference: <https://cloud.google.com/logging/docs/view/logging-query-language>)

Log Explorer - Sample Queries

```
# Query-1: Verify Nginx Logs
resource.type="gce_instance"
(log_id("nginx_access") OR log_id("nginx_error"))

# Query-2: Verify Nginx Logs with Instance ID
resource.type="gce_instance"
(log_id("nginx_access") OR log_id("nginx_error"))
resource.labels.instance_id="1493921793379482560"
```

Log Analytics - Sample Query

```
# Log Analytics Query
SELECT
  timestamp, severity, resource.type, log_name, text_payload, proto_payload, json_payload
FROM
  `gcplearn9.global.mylogbucket101._AllLogs`
LIMIT 10000
```



Demo



Cloud Observability

Cloud Logging

Log Collection Options & Integrations



Cloud Logging - Log Collection Options

- **Log Collection Options (Applications and Third-party Software)**

- **Option-1: Client Library**

- Instrument (Develop) your application with Google Cloud Logging client library so that logs from your applications can be sent to Cloud Logging in your GCP Project
<https://cloud.google.com/logging/docs/reference/libraries>

- **Option-2: Ops Agent**

- Client library is not required for all applications. For few supported third-party software, we can use Ops Agent
- Install and configure OpsAgent to send logs to Cloud Logging in your GCP Project (Example: Nginx)

Ops Agent - config.yaml

```

metrics:
receivers:
nginx:
  type: nginx
  stub_status_url: http://127.0.0.1:80/nginx_status
service:
pipelines:
nginx:
  receivers:
    - nginx
logging:
receivers:
nginx_access:
  type: nginx_access
nginx_error:
  type: nginx_error
service:
pipelines:
nginx:
  receivers:
    - nginx_access
    - nginx_error
  
```

Cloud Logging - Log Collection Options

- **Ops Agent Integrations**

- Integrations brings together **metrics, logs, dashboards, and alerts** to give you quick access to rich data for your application stack
- **Integration Examples:** Nginx, Microsoft IIS, Apache Webserver, MongoDB, Oracle DB and many more

Integrations

Quick filters	Filter	?
All 55	Filter integrations	
Deployment Platform <ul style="list-style-type: none"> Kubernetes Engine 33 Compute Engine 32 		
Aerospike The Aerospike integration collects key namespace and system metrics, such as disk and memory usage, scans, and connections. The integration collects these metrics using the official client API provided by Aerospike.		
Apache ActiveMQ The Apache ActiveMQ integration collects storage usage and message metrics. Storage metrics include memory and disk usage. Message metrics include number of waiting messages, average wait time, and expired messages.		
Apache Airflow Apache Airflow is an open-source platform for developing, scheduling, and monitoring batch-oriented workflows. Airflow is deployable in many ways, varying from a single		

VIEW DETAILS

VIEW DETAILS

VIEW DETAILS

VIEW DETAILS

Demo



Cloud Observability

Cloud Logging

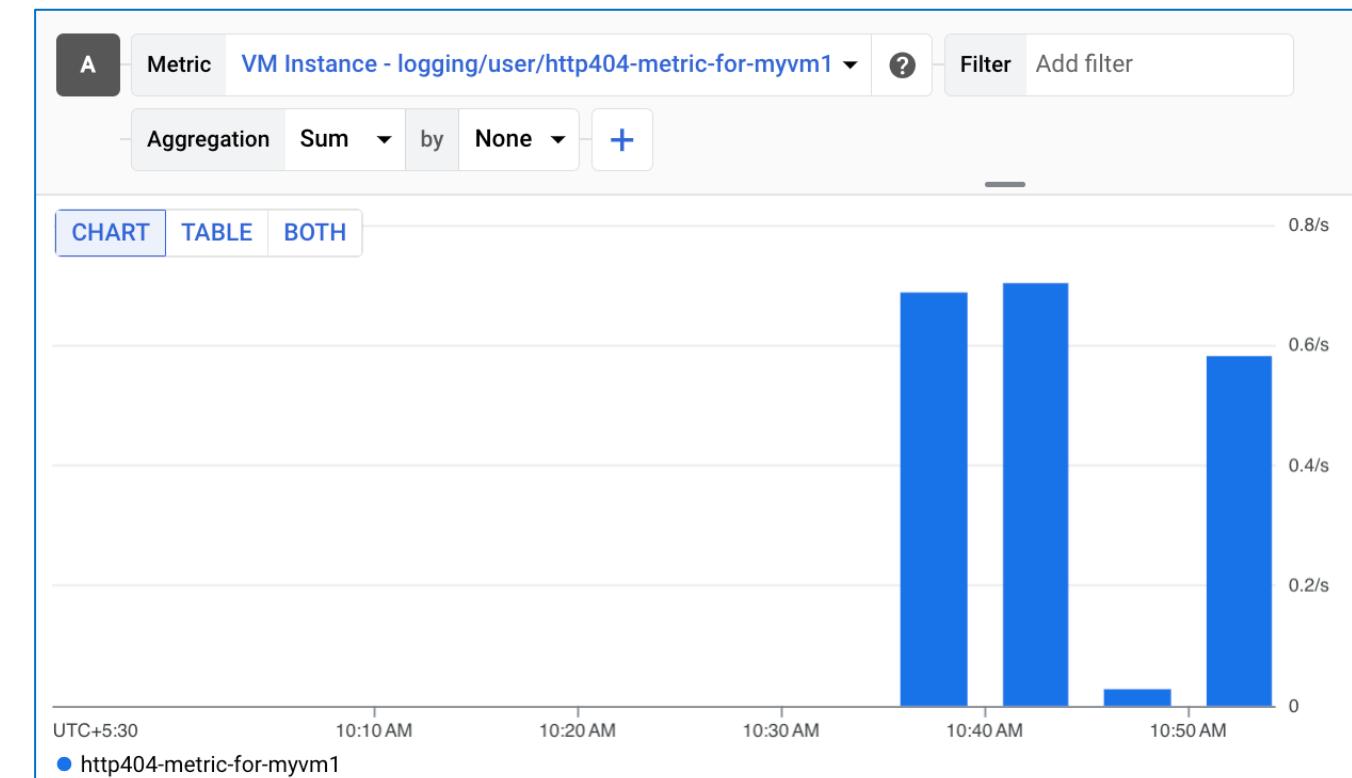
Monitor your Logs - Log based Metrics



Cloud Logging - Monitor your logs

- **Log-based Metric:** If you want to monitor **trends** or the **occurrence of events over time**, create log-based metric
- **Count:** It can **count the number of log entries** that match some pattern
- **Distribution:** It can **extract and organize** information like **response times** into histograms
- **Usecases**
 - Count the occurrences of a message, like a **warning or error**, in your logs and **receive a notification** when the number of occurrences crosses a threshold.
 - Create **charts to display** the numeric data extracted from your logs and add them on a Custom Dashboard.

```
# Search 404 logs
resource.type="gce_instance"
log_id("nginx_access")
http_request.status="404"
```





Demo

Google Cloud Observability

Cloud Logging

Monitor your Logs - Log based Alert Policy



Cloud Logging - Monitor your logs

- **Log-based Alert Policy:** Notify when certain kind of events occur in your log
- **Example-1:**
 - In nginx access log, if we want to get notified when unusual HTTP Methods like PUT, DELETE, CONNECT are being used (someone trying to hack) then we can configure Log based alert policy
- **Example-2:** You want to be notified when an event appears in an audit log; for example, a user accesses the security key of a service account.
- **Example-3:** Your application writes deployment messages to logs, and you want to be notified when a deployment change is logged.

```
# Search for HTTP Methods PUT, CONNECT, DELETE in nginx access logs
resource.type="gce_instance"
log_id("nginx_access")
httpRequest.requestMethod="PUT" OR httpRequest.requestMethod="DELETE" OR httpRequest.
requestMethod="CONNECT"
labels."compute.googleapis.com/resource_name"="myvm1"
```

Demo



Google Cloud Observability

Cloud Logging

Log Storage, Log Router Sinks, Log Categories



Cloud Logging - Log Storage

- **Log Storage:** By default, your Google Cloud project automatically stores all logs it receives in a **Cloud Logging log bucket**
- **Cloud Logging Bucket**
 - This is different from Cloud Storage bucket
 - By default, every GCP project creates two Cloud Logging buckets (**_Default**, **_Required**)
 - All the logs other than Audit logs will be stored in **_Default Log bucket**
 - You can create **custom Log Buckets**

Log buckets					
		Name ↑	Description	Previous month storage	Current month storage
<input type="checkbox"/>		_Default	Default bucket	35.68 MiB	12.78 MiB
<input type="checkbox"/>		_Required	Audit bucket	<i>Not billed</i>	<i>Not billed</i>
<input type="checkbox"/>	...	mylogbucket1	mylogbucket1	10.12 KiB	2.72 MiB
<input type="checkbox"/>	...	mylogbucket2	mylogbucket2	0 B	55.11 KiB

Cloud Logging - Log Storage

- **Log Router Sinks:** Using Log router sinks you can [configure to route logs](#) to desired [Cloud Logging bucket](#) or [other destinations](#)
- **Cloud Logging Destinations**
 - Cloud Logging Bucket
 - BigQuery Dataset
 - Cloud Storage Bucket
 - Pub/Sub topic
 - Splunk
 - Other Google Cloud Project

Log Router Sinks					
<input type="checkbox"/> Enabled		Type	Name ↑	Description	Destination
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Logging bucket	_Default		logging.googleapis.com/projects/gcplearn9/locations/global/buckets/_Default
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Logging bucket	_Required		logging.googleapis.com/projects/gcplearn9/locations/global/buckets/_Required

Cloud Logging - Log Categories

- **Log Categories:** Help describe the [logging information](#) available to you
- **Platform Logs:** Logs written by your [Google cloud services](#)
- **Component Logs:** Logs generated by [Google-provided software components](#) that run on your systems (On-premise VMs – GKE components)
- **Security Logs**
 - **Cloud Audit Logs:** Administrative activities and accesses within your Google Cloud resources
 - **Access Transparency Logs:** Logs of [actions taken by Google staff](#) when accessing [your Google Cloud content](#). Access Transparency logs can help you [track compliance with your legal and regulatory requirements](#) for your organization
- **User-written logs**
 - Ops-Agent
 - Cloud Logging API
 - Cloud Logging client libraries
- **Multi-cloud logs and Hybrid:** Logs from other cloud providers like [Microsoft Azure](#) and logs from [on-premises](#) infrastructure.

Demo



Google Cloud Observability

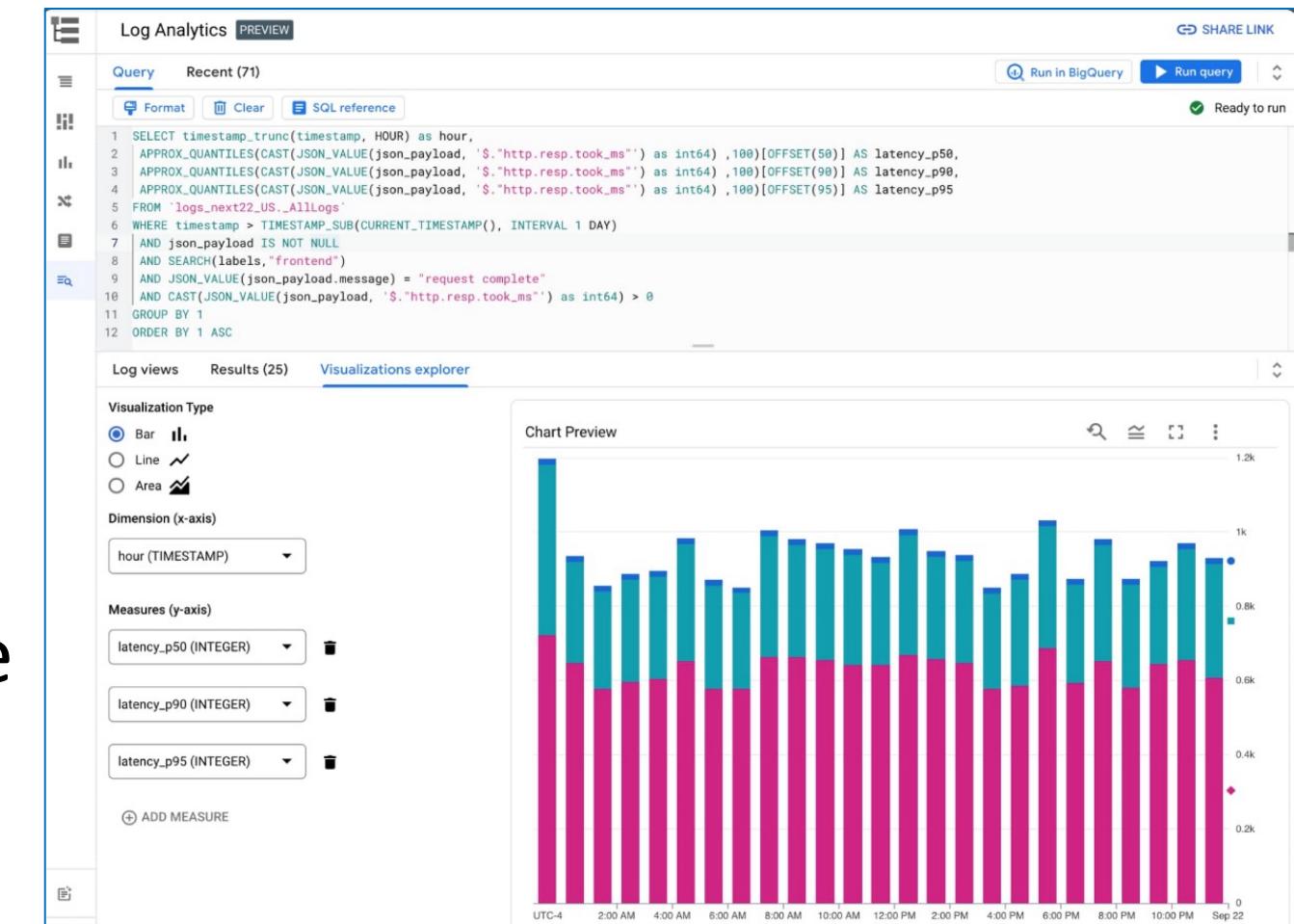
Cloud Logging

Log Analytics



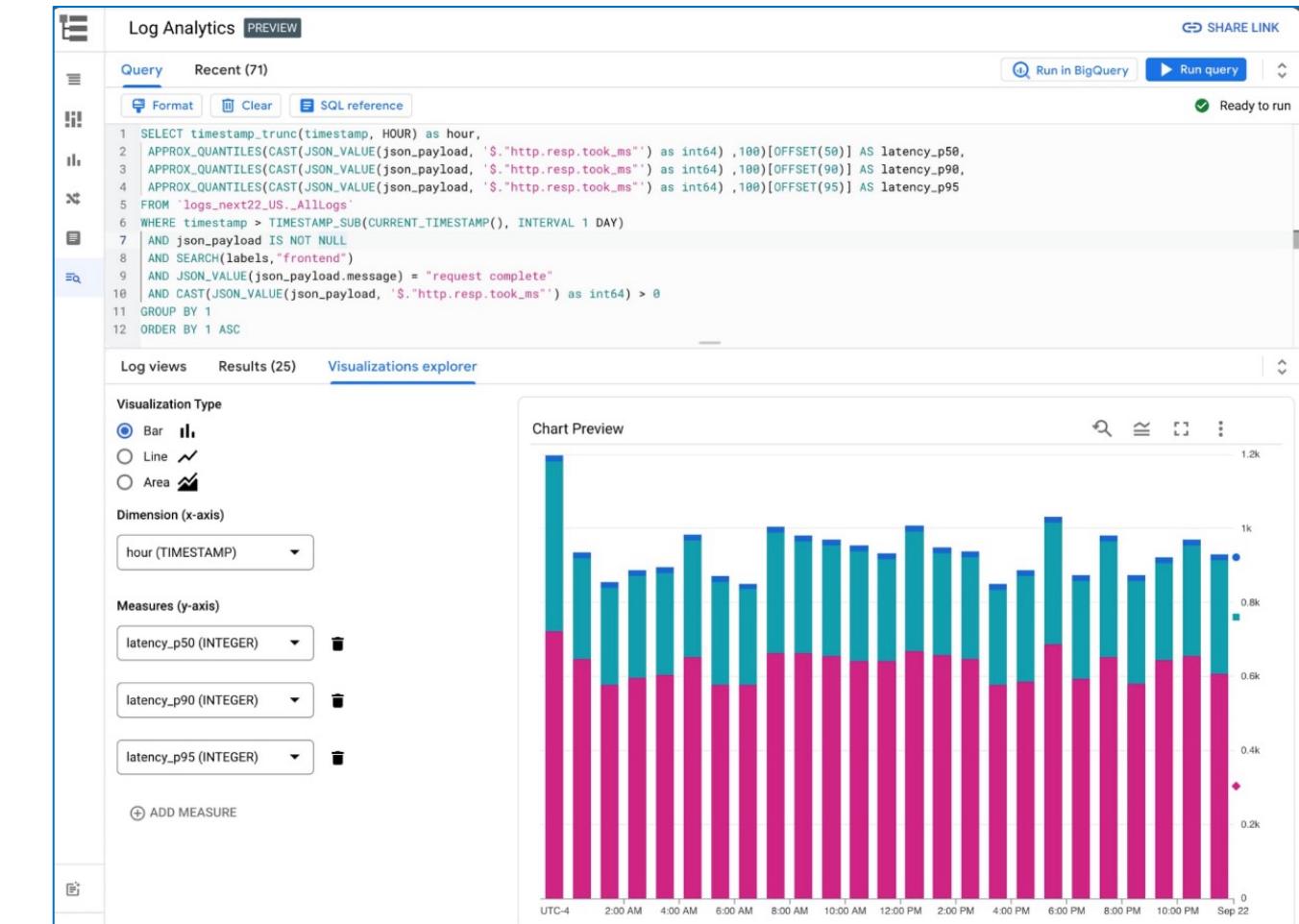
Cloud Logging - Log Analytics

- **Log Analytics:** When you're interested in **performing aggregate operations on your logs**, then you can use Log Analytics
 - **Example:** **counting** the number of log entries that contain a **specific pattern**
- Log Analytics leverages the **power of BigQuery** to perform Analytics
- To use Log Analytics, we need to **upgrade** the Log Bucket.
- **Query Language:** You use **SQL** to query the log data
- **Scalable platform:** Can scale using the **serverless BigQuery** platform and perform aggregation at petabyte scale



Cloud Logging - Log Analytics Benefits

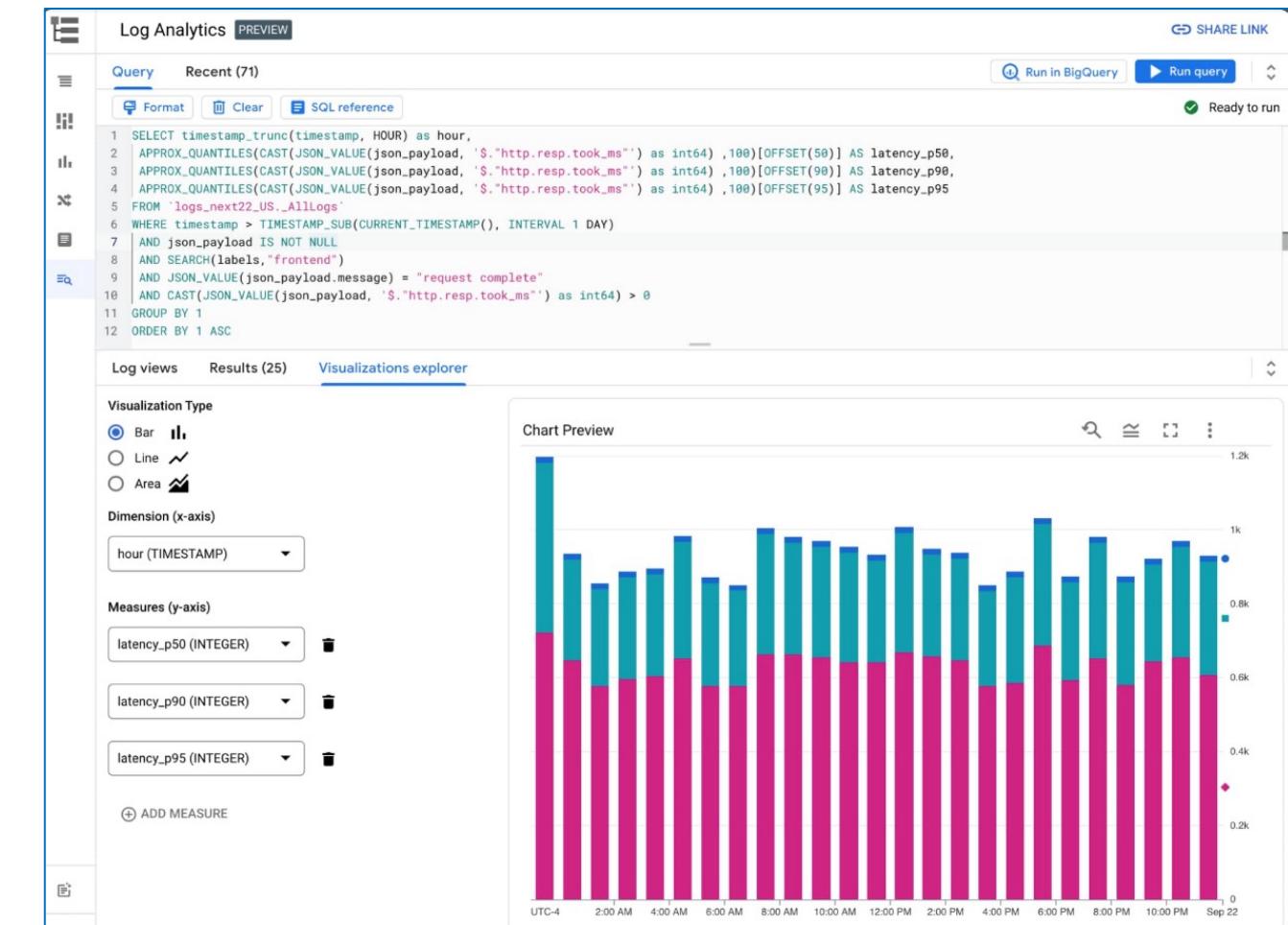
- **Centralized Logging:**
 - Cloud Logging collects and stores the log data in a dedicated Log Bucket.
 - Log Analytics uses the same bucket for Data analysis.
 - You don't need to make duplicate copies of the data.
- **Reduced cost and complexity**
 - Log Analytics allows reuse of data across the organization (same log bucket), effectively saving cost and reducing complexities
- **Big Query Linked Datasets**
 - Log Analytics also let you use BigQuery to query your data
 - You can create a linked dataset in BigQuery to query data from BigQuery



Cloud Logging - Log Analytics Restrictions

- **Restrictions**

- Not **all regions** are supported for Log Analytics
- To upgrade an existing bucket to use Log Analytics
 - **Condition-1:** The log bucket is **unlocked** unless it is the **_Required** bucket.
 - **Condition-2:** There aren't **pending updates** to the bucket.
- Log buckets that are upgraded to use Log Analytics, **we cannot rollback** (**cannot downgrade or remove Log Analytics feature**)
- **VERY VERY IMPORTANT:** Only log entries **written after the upgrade has completed** are available for analytics.



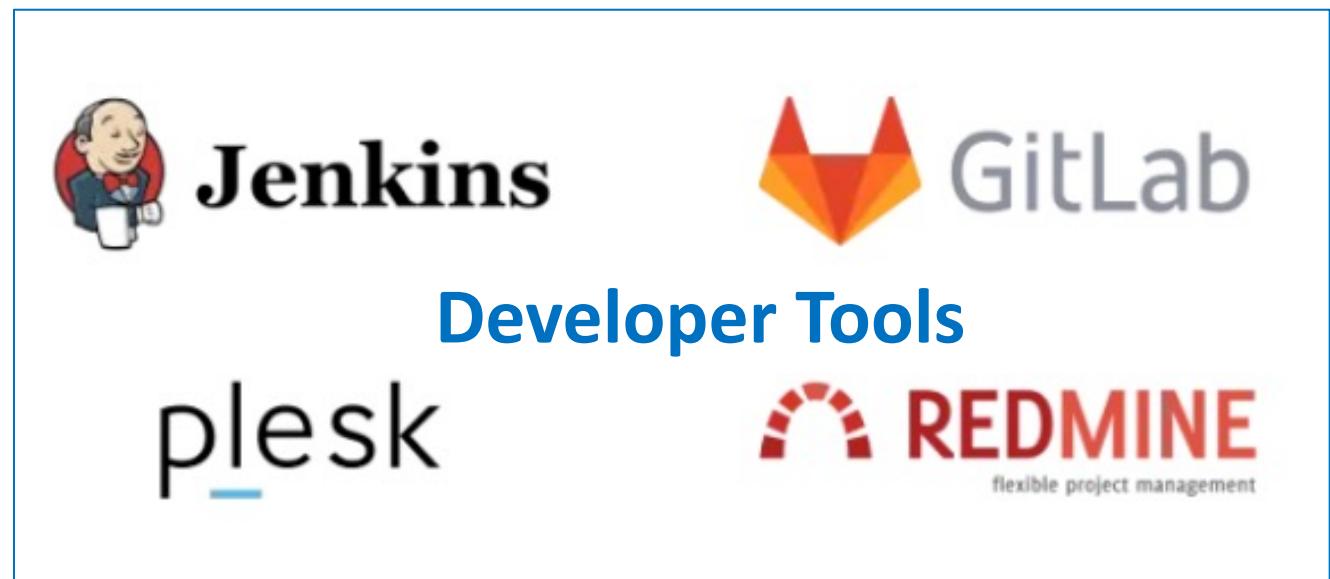
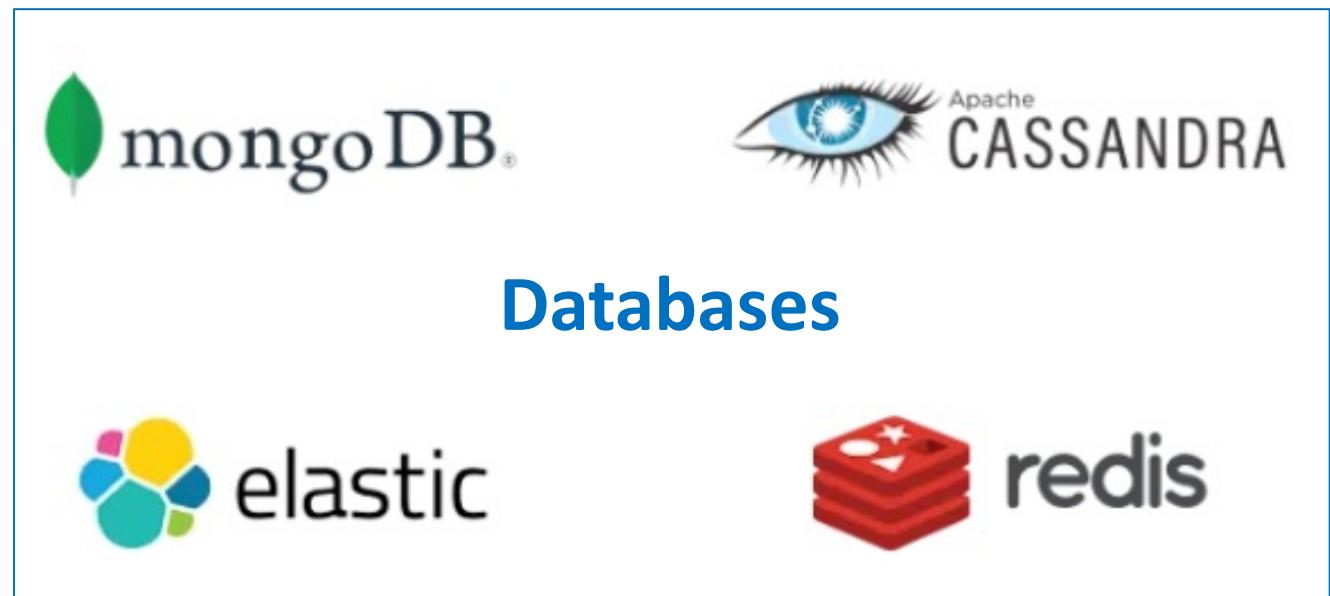
Demo

Google Cloud Cloud Marketplace



Cloud Marketplace

- Cloud Marketplace: Ready to deploy solutions
- Deploy as a software package
- We don't need deeper knowledge of Google Cloud Services (VMs, Storage, VPC networks)
- Simple and easy to follow
- It's a Modern Procurement Platform with flexible billing options
- Thoroughly pre-tested on google cloud



Cloud Marketplace - Billing

• Billing: Free Products

- Many software packages in Marketplace are **free to use**
- You only pay the **standard usage fees for the Google Cloud resources** that you run the software on (Example: VM Instances)

• Billing: Commercial Software

- If you buy a commercial software, you pay for **software + GCP resources where it is hosted**

• Commercial Software Example:

- Ubuntu Pro
- Windows Servers
- Splunk

• You can **negotiate with partners** to reduce cost

• **Just One Bill:** Amount spent on Cloud Marketplace will be **added to Google Cloud Invoice** (No multiple billing systems)

Google Cloud Marketplace Product Types

Price	^
Free Trial	(109)
Free	(1,044)
Paid	(1,883)
BYOL	(264)

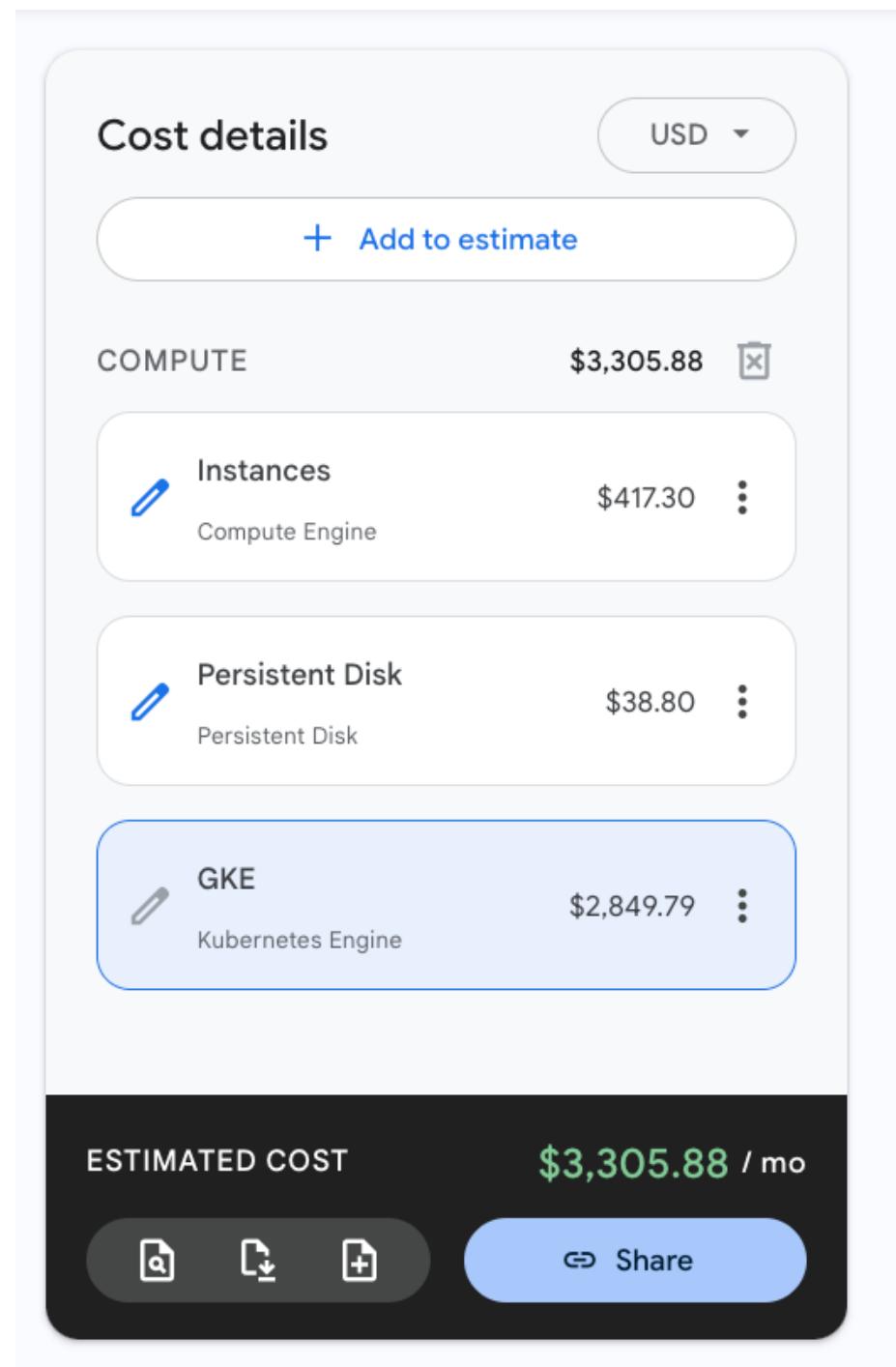
Demo



Google Cloud Pricing Calculator

Google Cloud - Pricing Calculator

- **Pricing Calculator:** We can generate an estimated cost for Google Cloud Resources
- We can
 - get estimates in our currency (USD, INR)
 - can download the estimates
 - create duplicates for our estimates
 - can share the estimates link
 - can do multiple GCP services estimates in a single screen
- **Link to Pricing Calculator:**
<https://cloud.google.com/products/calculator>



ALL LIVE SLIDES ARE BEFORE THIS SLIDE