

Deploying C++ application on GKE

Overview

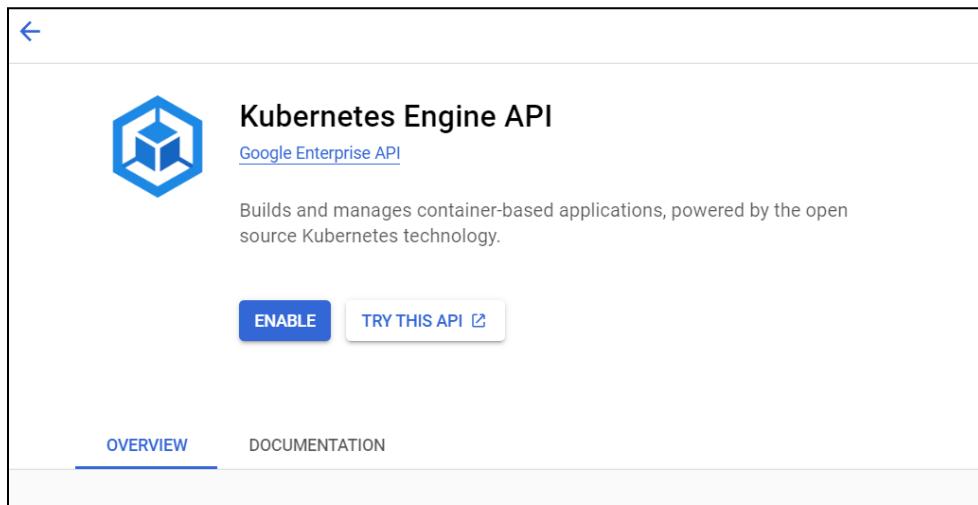
Google Kubernetes Engine (GKE) provides a managed environment for deploying, managing, and scaling your containerized applications using Google infrastructure. The GKE environment consists of multiple machines (specifically, Compute Engine instances) grouped together to form a cluster.

In this hands-on lab, you'll create Google Kubernetes Engine (GKE) cluster using the Google Cloud Console.

Prerequisites

To create a Google Kubernetes Engine (GKE) cluster in GCP following resources such as secondary IP ranges for Nodes and Pods, service account and firewall rule for health checks is required to be provisioned

Enable the Kubernetes Engine API - [Kubernetes Engine API](#)



1. Create a Secondary IP Ranges for Pods and Service

- In the Cloud Console, on the **Navigation menu (≡)**, click **VPC network > VPC networks**.
- Search VPC created previously i.e **labs-vpc**, click on **labs-vpc**.

Name	Region	Stack Type	Internal IP ranges	External IP ranges	Secondary IPv4 ranges	Gateway	Private Google Access	Flow logs
labs-subnet	us-central1	IPv4	10.0.0.0/24	None	None	10.0.0.1	Off	Off

- In the subnet section click on **labs-subnet**, to add Secondary ranges.

IP range	IP version	Access type
10.0.0.0/24	IPv4	Internal

- Click on **EDIT**, to add secondary ranges.
- In **Secondary IPv4 ranges**, click on **ADD IP RANGE**.

Add **subnet range name 1** as pod and **Secondary IPv4 range 1** as 10.91.0.0/22.

Add **subnet range name 2** as svc and **Secondary IPv4 range 2** as 10.90.0.0/22.

The screenshot shows the 'VPC network' configuration page. On the left sidebar, there are several options: VPC networks, IP addresses, Bring your own IP, Firewall, Routes, VPC network peering, Shared VPC, Serverless VPC access, and Packet mirroring. The main panel has sections for 'IP stack type' (set to 'IPv4 (single-stack)'), 'IP ranges' (containing a note about uniqueness and a list box with '10.0.0.0/24'), 'Secondary IPv4 ranges' (with two input fields: 'Subnet range name 1 - pod' set to 'E.g. range-1' and 'Secondary IPv4 range 1 *' set to '10.91.0.0/22'; and 'Subnet range name 2 - svc' set to 'E.g. range-1' and 'Secondary IPv4 range 2 *' set to '10.90.0.0/22')). At the bottom, there is a 'Gateway' field set to '10.0.0.1' and a blue '+ ADD IP RANGE' button.

- Click **SAVE** to add the secondary ranges. The secondary ranges for pods and service are created.

The screenshot shows the 'VPC network' configuration page after saving. The 'VPC Network' is named 'labs-vpc'. The 'Region' is 'us-central1'. The 'IP stack type' is 'IPv4 (single-stack)'. The 'IP ranges' section shows a table with one row: 'IP range' '10.0.0.0/24', 'IP version' 'IPv4', and 'Access type' 'Internal'. The 'Secondary IPv4 ranges' section shows a table with two rows: 'Subnet range name' 'pod' and 'Secondary IPv4 range' '10.91.0.0/22'; and 'Subnet range name' 'svc' and 'Secondary IPv4 range' '10.90.0.0/22'. The 'Gateway' is '10.0.0.1'. The 'Private Google Access' is set to 'Off'.

2. Create Health check Firewall Rule

You need to create ingress firewall rules applicable to all VMs being load balanced to allow traffic from health check prober IP ranges.

- In the Cloud Console, on the **Navigation menu** (≡), click **VPC network > Firewall**.
- To create a new Health check Firewall rule, click **CREATE FIREWALL RULE**.

There are many parameters you can configure when creating a new firewall rule

- Enter firewall **Name** as it is a mandatory field.
- Description can be left blank as it is an optional field.
- Select the VPC **Network** created previously i.e. *labs-vpc* to attach the firewall to that network.

The screenshot shows the 'Create a firewall rule' interface. On the left is a sidebar with icons for VPC networks, IP addresses, Bring your own IP, Firewall (which is selected and highlighted in blue), Routes, VPC network peering, Shared VPC, Serverless VPC access, and Packet mirroring. The main area has a title 'Create a firewall rule' with a back arrow. It contains fields for 'Name' (set to 'labs-health-check-fw'), 'Description' (empty), 'Logs' (with 'Off' selected), 'Network' (set to 'labs-vpc'), 'Priority' (set to '1000'), 'Direction of traffic' (set to 'Ingress'), and a note to 'CHECK PRIORITY OF OTHER FIREWALL RULES'.

- Select **Direction of traffic** as **Ingress** by selecting the checkbox.
- Select Action on match as **Allow** by selecting the checkbox.
- In the Target field, select **All instances in the network**.
- In the Source filter, select IPv4 ranges.
- Enter the Source IPv4 range **35.191.0.0/16 & 130.211.0.0/22**.
- In Protocols and Ports, enable **TCP** protocol with port 80 for HTTP and 8080.

VPC network

Direction of traffic **Ingress**

Action on match **Allow**

Targets: All instances in the network

Source filter: IPv4 ranges

Source IPv4 ranges *: 35.191.0.0/16, 130.211.0.0/22

Second source filter: None

Protocols and ports **TCP**

Ports: 80,8080

CREATE

- Click on **CREATE** to create a Health check firewall.

Second source filter: None

Protocols and ports **TCP**

Ports: 80,8080

UDP

Other

DISABLE RULE

CREATE **CANCEL**

EQUIVALENT COMMAND LINE

- The IAP firewall named `labs-health-check-fw` is created.

Cloud Firewall Rules											
	Name	Type	Targets	Filters	Protocols / ports	Action	Priority	Network	Logs	Hit count	
<input type="checkbox"/>	labs-health-check-fw	Ingress	Apply to all	IP ranges: 35.191.0.0/16, 130.211.0.0/22	tcp:80, 8080	Allow	1000	labs-vpc	Off	Edit	

3. Create Service Account for GKE

A service account is identified by its email address, which is unique to the account. Before creating the service account following the below steps

- Enable the IAM API - [IAM API](#)

Required roles for your IAM account.

To get the permissions that you need to manage service accounts, ask your administrator to grant you the following IAM roles on the project:

To view and create service accounts:

Create Service Accounts (roles/iam.serviceAccountCreator)

- In the Cloud Console, on the **Navigation menu** (≡), click **IAM & Admin > Firewall**.
- To create a new custom service account, click **CREATE SERVICE ACCOUNT**.

There are many parameters you can configure when creating a new firewall rule

- Enter **Service Account Name** as it is a mandatory field.
- Service account ID is auto populated with service account name.
- Description can be kept blank as it is an optional field.

- Click **CREATE AND CONTINUE** to create the custom service account.

Service account details

Service account name: labs-gke-sa

Display name for this service account:

Service account ID *: labs-gke-sa

Email address: labs-gke-sa@[REDACTED].iam.gserviceaccount.com

Service account description:

Describe what this service account will do

CREATE AND CONTINUE

Grant this service account access to project (optional)

Grant users access to this service account (optional)

DONE **CANCEL**

- Other fields are optional and can be skipped. The custom service account name `labs-gke-sa@project-id.iam.gserviceaccount.com` is created
- Ask your administrator to grant **Storage Admin, Kubernetes Engine Cluster Admin, Kubernetes Engine Admin, Compute Admin and Artifact Admin** IAM roles on the created custom service account
- Navigate to IAM, click **Grant ACCESS**.
- Enter the created service account name in **New Principals**.
- In Role, Select **Storage Admin, Kubernetes Engine Cluster Admin and Artifact Admin** from the dropdown.

Principal ⓘ
labs-gke-sa@[REDACTED].iam.gserviceaccount.com

Project [REDACTED]

Assign roles

Roles are composed of sets of permissions and determine what the principal can do with this resource. [Learn more](#)

Role	IAM condition (optional) ⓘ	
Artifact Registry Administrator	+ ADD IAM CONDITION	[REDACTED]
Compute Admin	+ ADD IAM CONDITION	[REDACTED]
Kubernetes Engine Admin	+ ADD IAM CONDITION	[REDACTED]
Kubernetes Engine Cluster A...	+ ADD IAM CONDITION	[REDACTED]
Monitoring Admin	+ ADD IAM CONDITION	[REDACTED]

SAVE **TEST CHANGES** **CANCEL**

4. Create Artifact Repository

- In the Cloud Console, on the **Navigation menu** (≡), click **Artifact Registry > Repositories**.
- To create a new Artifact repository, click **CREATE ARTIFACT REPOSITORY**.

There are many parameters you can configure when creating a new Artifact Repository.

- Enter **Repository Name** as it is a mandatory field.
- Select format as **DOCKER**.
- Select Location type as Region and from the region drop down select **us-central1**.
- Description can be kept blank as it is an optional field.
- For encryption, let it be default google managed encryption keys.
- Click **CREATE**, to create a new artifact repository.

The screenshot shows the 'Create repository' interface. On the left, there's a sidebar with 'Repositories' and 'Settings'. The main area has fields for 'Name' (set to 'labs-repo'), 'Format' (radio buttons for Docker, Maven, npm, Python, Apt, Yum, Kubeflow Pipelines, and Go, with Docker selected), 'Location type' (radio buttons for Region and Multi-region, with Region selected), and 'Region' (a dropdown menu showing 'us-central1 (Iowa)' which is highlighted with a blue border). There are also sections for 'Release Notes' and 'Description'.

- The repository is created named **labs-repo**.

Repositories	+ CREATE REPOSITORY	DELETE	SETUP INSTRUCTIONS	REFRESH	LEARN	SHOW INFO PANEL
Repositories	ARTIFACT REGISTRY	CONTAINER REGISTRY				
Settings	Filter: labs-repo	Enter property name or value		X	?	☰

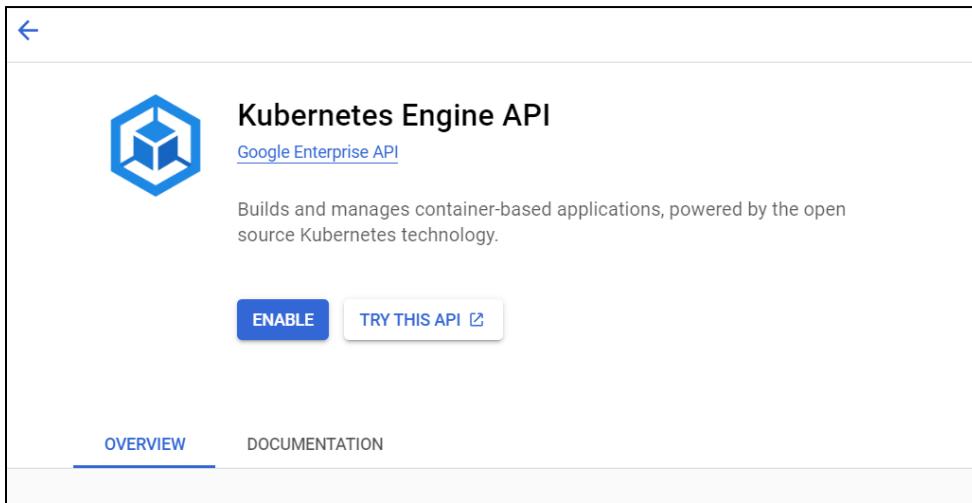
A table listing repositories. One row is shown for 'ARTIFACT REGISTRY' (Container Registry) named 'labs-repo'. The table includes columns for Name, Format, Location, Description, Labels, Version policy, Encryption, Encryption key, Created, and Updated. The 'labs-repo' row shows Docker as the format, us-central1 (Iowa) as the location, and creation times of 4 days ago and 3 days ago respectively.

5. Install Docker

Install Docker based on your Operating System using this [link](#).

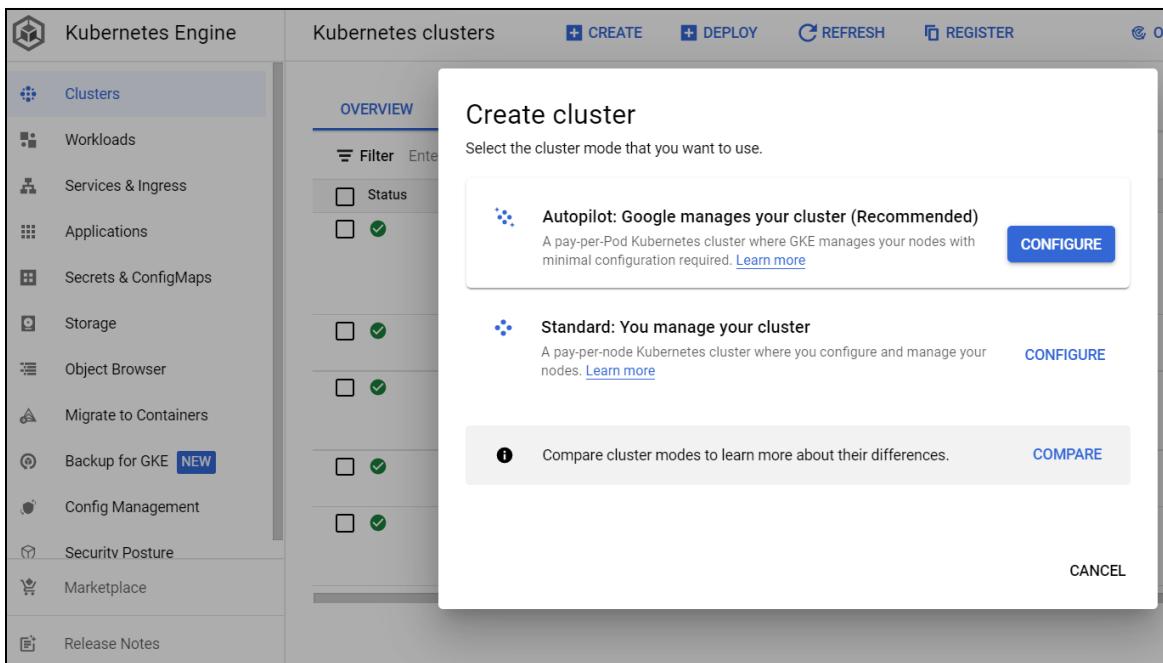
Setup for Google Kubernetes Engine

Enable the Kubernetes Engine API - [Kubernetes Engine API](#)



Create Google Kubernetes Engine Cluster

- In the Cloud Console, on the **Navigation menu** (≡), click **Kubernetes Engine > Cluster**.
- To create a new cluster, click **CREATE**.
- A popup window will appear to select the cluster mode, select **Standard** mode.



- On Cluster basics, Enter the **Name** of cluster. Select location type as **Zonal** and in zone select **us-central1** from dropdown.

Cluster basics

The new cluster will be created with the name, version, and in the location you specify here. After the cluster is created, name and location can't be changed.

To experiment with an affordable cluster, try [My first cluster](#) in the Cluster set-up guides

Name: labs-demo-cluster

Location type: Zonal

Zone: us-central1-c

Specify default node locations

Estimated monthly cost: [PREVIEW](#)
\$158.38
That's about \$0.22 per hour
Pricing is based on the resources you use, management fees, discounts and credits. [Learn more](#)

[SHOW COST BREAKDOWN](#)

CREATE CANCEL Equivalent [REST](#) or [COMMAND LINE](#)

- Rest of the fields can be left with default values.

Cluster basics

Specify default node locations

Increase availability by selecting more than one zone
Current default: us-central1-c

Control plane version

Choose whether you'd like to upgrade the cluster's control plane version manually or let GKE do it automatically. [Learn more](#).

Static version
Manually manage the version upgrades. GKE will only upgrade the control plane and nodes if it's necessary to maintain security and compatibility, as described in the release schedule. [Learn more](#).

Release channel
Let GKE automatically manage the cluster's control plane version. [Learn more](#).

Release channel: Regular channel (default)

Version: 1.23.8-gke.1900 (default)

These versions have passed internal validation and are considered production-quality, but don't have enough historical data to guarantee their stability. Known issues generally have known workarounds. [Release notes](#)

CREATE CANCEL Equivalent [REST](#) or [COMMAND LINE](#)

- Click on default-pool. In Size, Enter **Number of nodes** to **1**. Select the checkbox **Enable cluster autoscaler** for nodes autoscaling. Enter **minimum number of nodes** to **1** and **maximum number of nodes** to **4**.

Create a Kubernetes cluster

[ADD NODE POOL](#) [REMOVE NODE POOL](#) [USE A SETUP GUIDE ▾](#)

Node pool details

Cluster basics

NODE POOLS

- default-pool

CLUSTER

- Automation
- Networking
- Security
- Metadata
- Features

Name default-pool

A node pool is a template for groups of nodes created in this cluster. The new cluster will be created with at least one node pool. More node pools can be added and removed after cluster creation. [Learn more](#)

Node pool names must start with a lowercase letter followed by up to 39 lowercase letters, numbers, or hyphens. They can't end with a hyphen. You cannot change the node pool's name once it's created.

Control plane version - 1.23.8-gke.1900

Compact placement [Beta](#) [?](#)

Size

Number of nodes * 1

Pod address range limits the maximum size of the cluster. [Learn more](#)

Enable cluster autoscaler Cluster autoscaler automatically creates or deletes nodes based on workload needs. [Learn more](#)

Minimum number of nodes * 1

Maximum number of nodes * 4

[CREATE](#) [CANCEL](#) Equivalent [REST](#) or [COMMAND LINE](#)

- In the default-pool sub section, Click on **Nodes**. Select the machine type for your cluster.

These node settings will be used when new nodes are created using this node pool.

NODE POOLS

- default-pool

CLUSTER

- Automation
- Networking
- Security
- Metadata
- Features

Image type Container-Optimized OS with containerd (cos_containerd) (default)

Choose which operating system image you want to run on each node of this cluster. [Learn more](#)

The default Linux node image for newly created clusters and node pools with version 1.23.8-gke.1900 or later is Container-optimized OS with containerd. For Windows node pools using version 1.21 or later, containerd is also the recommended runtime. Since Dockershim is being deprecated by Kubernetes project, [GKE will deprecate Docker node images](#). We recommend that you [migrate to containerd node images](#) as soon as possible. [Learn more](#).

Machine configuration

Choose the machine family, type, and series that will best fit the resource needs of your cluster. You won't be able to change the machine type for this cluster once it's created. [Learn more](#)

Machine family

[GENERAL-PURPOSE](#) [COMPUTE-OPTIMIZED](#) [MEMORY-OPTIMIZED](#) [GPU](#)

Machine types for common workloads, optimized for cost and flexibility

Series E2

CPU platform selection based on availability

Machine type e2-medium (2 vCPU, 4 GB memory)

[CREATE](#) [CANCEL](#) Equivalent [REST](#) or [COMMAND LINE](#)

- Select **Series** as **E2** machine type as **e2-medium**. In Boot disk size enter **30 GB**. Boot disk size can be decided based on application requirements.
- Select the checkbox **Enable nodes on spot VMs**, spot VM's are available for only 24 hours post that they are terminated. By enabling the spot VM option, the total cost of the cluster is reduced from \$158.94 to \$98.61.

Create a Kubernetes cluster

Machine family

GENERAL-PURPOSE COMPUTE-OPTIMIZED MEMORY-OPTIMIZED GPU

Estimated monthly cost [PREVIEW](#)
\$98.61
That's about \$0.14 per hour
Pricing is based on the resources you use, fees, discounts and credits. [Learn more](#)

NODE POOLS

default-pool

- Nodes
- Networking
- Security
- Metadata

CLUSTER

- Automation
- Networking
- Security
- Metadata
- Features

Machine types for common workloads, optimized for cost and flexibility

Series: E2

Machine type: e2-medium (2 vCPU, 4 GB memory)

vCPU: 1.2 vCPU (1 shared core)

Memory: 4 GB

CPU PLATFORM AND GPU

Boot disk type: Standard persistent disk

Boot disk size (GB): 30

Enable customer-managed encryption for boot disk

Local SSD disks

Enable nodes on spot VMs

CREATE CANCEL Equivalent REST or COMMAND LINE

- In **Networking**, leave with default values.

Node networking

These node networking settings will be used when new nodes are created using this node pool.

The cluster settings specify a maximum of 110 Pods per node, but you can override that setting at the node pool level.

Maximum Pods per node: 110

Mask for Pod address range per node: /24

Network tags

Node Pool Pod Address Range

The cluster settings specify a default cluster level pod address range, but you can override that setting at the node pool level.

Automatically create secondary ranges

Pod secondary CIDR range

CREATE CANCEL Equivalent REST or COMMAND LINE

- In **Security**, In Identity defaults select the service account created previously from the dropdown i.e. **labs-gke-sa**.

Node security

These node security settings will be used when new nodes are created using this node pool.

Identity defaults

Specify the default identity for new auto-provisioned node pools using either a service account or one or more scopes. [Learn more](#)

Service account labs-gke-sa

The service account is used to call Google Cloud APIs. Use the default service account if available.

Access scopes

Access scopes are permanent. Select the type and level of API access to grant the VM. [Learn more](#)

Use IAM roles with service accounts to control VM access. [Learn more](#)

Enable sandbox with gVisor ?

Enable integrity monitoring ?

Enable secure boot ?

CREATE CANCEL Equivalent [REST](#) or [COMMAND LINE](#)

- In **Metadata**, leave with default values.

Node metadata

These node metadata settings will be used when new nodes are created using this node pool.

Kubernetes labels

Use Kubernetes labels to control how workloads are scheduled to your nodes. Labels are applied to all nodes in this node pool and cannot be changed once the cluster is created.

+ ADD KUBERNETES LABEL

Node taints

A node taint lets you mark a node so that the scheduler avoids or prevents using it for certain Pods. Node taints can be used with tolerations to ensure that Pods aren't scheduled onto inappropriate nodes. [Learn more](#)

+ ADD TAINT

GCE instance metadata

Use Kubernetes labels to control how workloads are scheduled to your nodes. Labels are applied to all nodes in this node pool and cannot be changed once the cluster is created. [Learn more](#)

Key 1 *	disable-legacy-endpoints	Value 1 *	true
---------	--------------------------	-----------	------

CREATE CANCEL Equivalent [REST](#) or [COMMAND LINE](#)

- In **Networking**, Select checkbox networks in this project. In **Network**, select the network created previously **labs-vpc** from the dropdown. In the Node subnet,

select the subnet created previously **labs-subnet**. In **Network access**, select the checkbox **Private Cluster**.

Networking

Define how applications in this cluster communicate with each other and with the Kubernetes control plane, and how clients can reach them.

Networks in this project
 Networks shared with me (from host project: searce-playground-host-project)

Network *
 labs-vpc

Node subnet *
 labs-subnet

⚠ Choose a network that has subnetworks in the us-central1 region. To use this network, choose a different region.

Network access

Choose the type of network you want to allow to access your cluster's workloads. [Learn more](#)

Public cluster
 Choose a public cluster to configure access from public networks to the cluster's workloads. Routes aren't created automatically. You cannot change this setting after the cluster is created.

Private cluster
 Choose a private cluster to assign internal IP addresses to Pods and nodes. This isolates the cluster's workloads from public networks. You cannot change this setting after the cluster is created.

CREATE **CANCEL** Equivalent [REST](#) or [COMMAND LINE](#)

- Enable the checkbox, **Access control plane using its external IP address** and **Enable Control plane global access**. In control plane IP range enter **10.1.10.0/28**. Disable checkbox **Automatically create Secondary ranges** & select pod and service. Rest of the fields are auto populated.

Enable Control plane global access

Control plane IP range *
 10.1.10.0/28

Disable Default SNAT

Enable VPC-native traffic routing (uses alias IP)

Automatically create secondary ranges

Pod secondary CIDR range *
 pod (10.91.0.0/22)

Maximum Pods per node
 110

Mask for Pod address range per node: /24

Services secondary CIDR range *
 svc (10.90.0.0/22)

Enable Dataplane V2

⚠ GKE Dataplane V2 has been certified to run up to 500 nodes per cluster, including node autoscaling and surge upgrades. You may request a cluster size of up to 1000 nodes by filing a Support ticket with GCP. For more information, see the note.

CREATE **CANCEL** Equivalent [REST](#) or [COMMAND LINE](#)

- Select the checkbox, enable **control plane authorized network** to connect to your cluster from local machine or vm. In Authorized Networks, Enter **Name** and In Network enter your **public-ip/32**. Click on **DONE**.

The screenshot shows the 'Create a Kubernetes cluster' interface. The left sidebar has sections for Cluster basics, Node pools (with 'default-pool' expanded), and Cluster (with 'Networking' selected). The main area shows configuration options for 'NODE POOLS' and 'CLUSTER'. Under 'NODE POOLS', there are checkboxes for 'Enable NodeLocal DNSCache', 'Enable HTTP load balancing' (checked), 'Enable subsetting for L4 internal load balancers', and 'Enable control plane authorized networks' (checked). Under 'CLUSTER', the 'Networking' section is selected. A modal window titled 'New authorized network' is open, showing fields for 'Name' (set to 'public-ip') and 'Network' (set to '123.201.94.15/32'). Below the modal are 'CANCEL' and 'DONE' buttons.

- In Security, Select the checkbox **Enable Workload Identity**.

The screenshot shows the 'Create a Kubernetes cluster' interface. The left sidebar has sections for Cluster basics, Node pools (with 'default-pool' expanded), and Cluster (with 'Security' selected). The main area shows configuration options for 'NODE POOLS' and 'CLUSTER'. Under 'NODE POOLS', there are checkboxes for 'Enable Binary Authorization', 'Enable Shielded GKE Nodes' (checked), 'Enable Confidential GKE Nodes', 'Encrypt secrets at the application layer', and 'Enable Workload Identity' (checked). A dropdown menu for 'Select workload pool' is open, showing 'searce-playground-v1.svc.id.goog'. Under 'CLUSTER', the 'Security' section is selected. Other checkboxes include 'Enable Google Groups for RBAC'. At the bottom are 'CREATE', 'CANCEL', and 'Equivalent REST or COMMAND LINE' buttons.

- Metadata and Features, leave with default values.
- Cluster creation may take from 5-10 mins. Cluster named labs-demo-cluster is created successfully.

Kubernetes Engine

Clusters

labs-demo-cluster

Cluster creation can take five minutes or more.

33% - Cluster is being configured...

- 1. Configuring
- 2. Deploying
- 3. Health checks

Cluster basics

Name	labs-demo-cluster	🔒
Location type	Zonal	🔒
Control plane zone	us-central1-c	🔒
Default node zones	us-central1-c	✍
Release channel	Regular channel	✍ UPGRADE AVAILABLE

Kubernetes clusters

OVERVIEW **OBSERVABILITY** **COST OPTIMIZATION**

Filter **Name : labs-demo-cluster** **X** Enter property name or value

Status	Name	Location	Number of nodes	Total vCPUs	Total memory	Notifications
<input checked="" type="checkbox"/>	labs-demo-cluster	us-central1-c	1	2	4 GB	⋮

- To connect to cluster, click on cluster name, Click on **CONNECT**

Kubernetes Engine

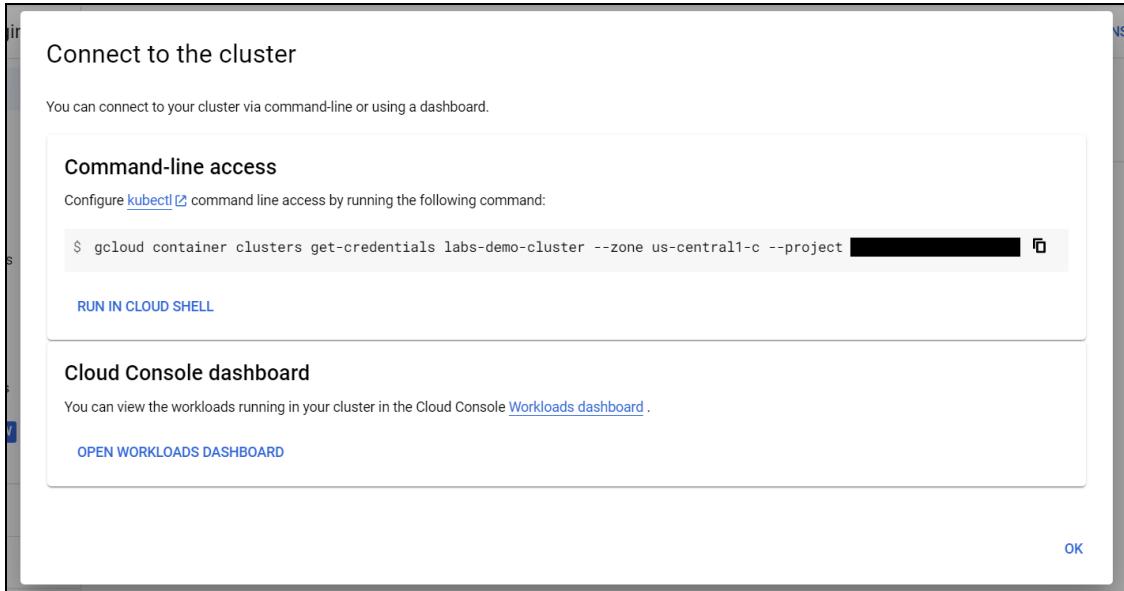
Clusters

labs-demo-cluster

Cluster basics

Name	labs-demo-cluster	🔒
Location type	Zonal	🔒
Control plane zone	us-central1-c	🔒
Default node zones	us-central1-c	✍
Release channel	Regular channel	✍ UPGRADE AVAILABLE
Version	1.23.8-gke.1900	
Total size	1	ⓘ
Endpoint	34.173.24.53	🔒
Show cluster certificate		

- A popup will appear with the gcloud command to connect to your cluster. Copy the command.



- SSH into your VM created previously (in Lab1), since cluster and VM are in the same network, VM can connect with an internal ip address. Use the above copied command with **--internal-ip** at the end. Install kubectl on the VM. Follow the documentation - [kubectl](#). Post installation of kubectl run command **kubectl get nodes** to verify nodes details.

```
root@labs-demo-vm:~# gcloud container clusters get-credentials labs-demo-cluster --zone us-central1-c --project
<project-id> --internal-ip
Fetching cluster endpoint and auth data.
CRITICAL: ACTION REQUIRED: gke-gcloud-auth-plugin, which is needed for continued use of kubectl, was not found or is not executable. Install gke-gcloud-auth-plugin for use with kubectl by following https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
kubeconfig entry generated for labs-demo-cluster.
root@labs-demo-vm:~# kubectl get nodes
W1114 09:41:40.595756 2558 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+, use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME STATUS ROLES AGE VERSION
gke-labs-demo-cluster-default-pool-5814bf37-9tfw Ready <none> 20m v1.23.8-gke.1900
root@labs-demo-vm:~#
```

Build Docker Image and Push it to Artifact Repository

- So far, we are able to connect to Cluster. Now it's time to start the deployment of the application and validate node & pod autoscaling.
- Login to Github and clone the following repository using command:
`git clone https://github.com/doketanmay98/labs-sample-app.git`
- Folder named labs-sample-app will be created in the present working directory(PWD).
- The repository consists of C++ application codes and Dockerfile to containerize the application.

```
root@labs-demo-vm:~# ll | grep labs
drwxr-xr-x 6 root root 4096 Nov 10 21:10 labs-sample-app/
root@labs-demo-vm:~#
```

- Change directory to Web-application using cd
labs-sample-app/Web-application. This folder contains C++ web applications and Dockerfile.

```
root@labs-demo-vm:~/labs-sample-app/Web-application# ls
Dockerfile WebApp
root@labs-demo-vm:~/labs-sample-app/Web-application#
```

- Create a docker image locally using the following command

```
cd labs-sample-app/Web-application/
```

```
docker build -t web-app .
```

```
docker images
```

```
root@labs-demo-vm:~/labs-sample-app/Web-application# docker build -t web-app .
Sending build context to Docker daemon 18.94kB
Step 1/7 : FROM amytabb/docker_ubuntu16_essentials
latest: Pulling from amytabb/docker_ubuntu16_essentials
8ee29e426c26: Already exists
6e83b260b73b: Already exists
e26b65fd1143: Already exists
40dca07f8222: Already exists
b420ae9e10b3: Already exists
dfaef13193fe3: Already exists
Digest: sha256:aalaeee49d49b6d641a81afeedb252bef55fbbce6a6badc50e61cf649dc9e8e6
Status: Downloaded newer image for amytabb/docker_ubuntu16_essentials:latest
--> 13f6d277d91b
Step 2/7 : COPY WebApp /WebApp
--> 59a40c51e465
Step 3/7 : WORKDIR /WebApp/
--> Running in e743d9d5cf88
Removing intermediate container e743d9d5cf88
--> b734564ff59e
Step 4/7 : RUN g++ -o WebApp WebApp.c
--> Running in 96e053187698
Removing intermediate container 96e053187698
--> e85a58f61128
Step 5/7 : CMD ["./WebApp"]
--> Running in 65430e7cdd47
Removing intermediate container 65430e7cdd47
--> 10f0b6358638
Step 6/7 : FROM nginx
latest: Pulling from library/nginx
e9995326b091: Already exists
71689475aec2: Already exists
f88a23025338: Already exists
0df440342e26: Already exists
eef26ceb3309: Already exists
8e3ed6a9e43a: Already exists
Digest: sha256:943c25b4b66b332184d5ba6bb18234273551593016c0e0ae906bab111548239f
Status: Downloaded newer image for nginx:latest
--> 76c69feac34e
Step 7/7 : COPY WebApp/index.html /usr/share/nginx/html/index.html
--> d788554c740f
Successfully built d788554c740f
Successfully tagged web-app:latest
root@labs-demo-vm:~/labs-sample-app/Web-application#
```

```
root@labs-demo-vm:~/labs-sample-app/Web-application# docker images
REPOSITORY           TAG      IMAGE ID      CREATED             SIZE
web-app              latest   d788554c740f  About a minute ago  142MB
<none>               <none>   10f0b6358638  About a minute ago  807MB
nginx                latest   76c69feac34e  2 weeks ago       142MB
amytabb/docker_ubuntu16_essentials  latest   13f6d277d91b  4 years ago        807MB
root@labs-demo-vm:~/labs-sample-app/Web-application#
```

- Run locally to validate the image working using the following command.

```
docker run -itd -p 8080:80 web-app:latest
```

```
curl localhost:8080
```

```
root@labs-demo-vm:~/labs-sample-app/Web-application# docker run -itd -p 8080:80 web-app:latest
57f21d35f11169b4f983d6ab6ae71fe729a661ddcccf2e1ef0cb6b2390bb81b9
root@labs-demo-vm:~/labs-sample-app/Web-application# docker ps -a
CONTAINER ID        IMAGE               COMMAND                  CREATED             STATUS              PORTS
 NAMES
57f21d35f111       web-app:latest     "/docker-entrypoint..."  4 seconds ago      Up 3 seconds      0.0.0.0:8080->80/tcp, :
::8080->80/tcp    nervous_booth
2d0e44020ad4       web-app:latest     "/docker-entrypoint..."  13 seconds ago     Created          happy_williamson
root@labs-demo-vm:~/labs-sample-app/Web-application# curl localhost:8080
Files can be tricky, but it is fun enough!root@labs-demo-vm:~/labs-sample-app/Web-application#
```

- Authorize with artifact credentials. Navigate to your Artifact Repository. Click on **SETUP INSTRUCTIONS**.

```
gcloud auth configure-docker \
us-central1-docker.pkg.dev
```

Follow the steps below to configure your client to push and pull packages using this repository. You can also view more detailed instructions [here](#). For more information about working with artifacts in this repository, see the [documentation](#).

Initialize gcloud

The [Google Cloud SDK](#) is used to generate an access token when authenticating with Artifact Registry. Make sure that it is installed and initialized with [Application Default Credentials](#) before proceeding.

Configure Docker

Run the following command to configure `gcloud` as the credential helper for the Artifact Registry domain associated with this repository's location:

```
$ gcloud auth configure-docker \
us-central1-docker.pkg.dev
```

```
root@labs-demo-vm:~/labs-sample-app/Web-application# gcloud auth configure-docker \
>     us-central1-docker.pkg.dev
WARNING: Your config file at [/root/.docker/config.json] contains these credential helper entries:

{
  "credHelpers": {
    "us-central1-docker.pkg.dev": "gcloud"
  }
}
Adding credentials for: us-central1-docker.pkg.dev
gcloud credential helpers already registered correctly.
root@labs-demo-vm:~/labs-sample-app/Web-application#
```

- Tag the locally built image with artifact details using the following command.
docker tag web-app us-central1-docker.pkg.dev/<project-id>/labs-repo/web-app

```
root@labs-demo-vm:~/labs-sample-app/Web-application# docker images
REPOSITORY          TAG      IMAGE ID   CREATED        SIZE
web-app              latest   d788554c740f  49 minutes ago  142MB
<none>              <none>   10f0b6358638  49 minutes ago  807MB
nginx               latest   76c69feac34e  2 weeks ago   142MB
amytabb/docker_ubuntu16_essentials  latest   13f6d277d91b  4 years ago   807MB
root@labs-demo-vm:~/labs-sample-app/Web-application# docker tag web-app us-central1-docker.pkg.dev/searce-playground-v1/labs-repo/web-app
root@labs-demo-vm:~/labs-sample-app/Web-application# docker images
REPOSITORY          TAG      IMAGE ID   CREATED        SIZE
web-app              latest   d788554c740f  52 minutes ago  142MB
us-central1-docker.pkg.dev/searce-playground-v1/labs-repo/web-app  latest   d788554c740f  52 minutes ago  142MB
<none>              <none>   10f0b6358638  52 minutes ago  807MB
nginx               latest   76c69feac34e  2 weeks ago   142MB
amytabb/docker_ubuntu16_essentials  latest   13f6d277d91b  4 years ago   807MB
root@labs-demo-vm:~/labs-sample-app/Web-application#
```

- Push the image to the Artifact repository.

docker push us-central1-docker.pkg.dev/<project-id>/labs-repo/web-app:latest

```
root@labs-demo-vm:~/labs-sample-app/Web-application# docker push us-central1-docker.pkg.dev/searce-playground-v1/labs-repo/web-app
Using default tag: latest
The push refers to repository [us-central1-docker.pkg.dev/searce-playground-v1/labs-repo/web-app]
55fdad469d63: Pushed
a2e59a79fae0: Layer already exists
4091cd312f19: Layer already exists
9e7119c28877: Layer already exists
2280b348f4d6: Layer already exists
e74d0d8d2def: Layer already exists
a12586ed027f: Layer already exists
latest: digest: sha256:0d3b243aa1864c97f40f15b705194d42a597d3e22d54315c3750d34ccece3caa size: 1777
root@labs-demo-vm:~/labs-sample-app/Web-application#
```

- Verify the pushed image in the artifact repository.

Images for labs-repo		
◀ DELETE SETUP INSTRUCTIONS		
 Repositories		
 Settings		
 Filter web-app Enter property name or value		
<input type="checkbox"/> Name ↑	 Created	 Updated
<input type="checkbox"/> web-app	 2 minutes ago	 2 minutes ago

Deploying C++ Web Application on GKE Cluster

- Change directory to gke using cd **/labs-sample-app/gke**. This folder contains Kubernetes deployment, service and Horizontal Pod Autoscaling manifest files.
cd /labs-sample-app/gke.
- Edit the deployment.yaml file. Insert your artifact repository url in the **image** parameter to pull the image.
vi deployment.yaml

```

apiVersion: apps/v1
kind: Deployment
metadata:
  name: myapp
spec:
  selector:
    matchLabels:
      app: myapp
  template:
    metadata:
      labels:
        app: myapp
    spec:
      containers:
        - name: myapp
          image: us-central1-docker.pkg.dev/<project-id>/labs-repo/web-app:latest
          resources:
            requests:
              memory: "128Mi"
              cpu: "100m"
            limits:
              memory: "128Mi"
              cpu: "100m"
          ports:
            - containerPort: 80

```

Press **escape** and type **:wq** to save the changes.

- Deploy the application using the following command

```
#deploy the application
kubectl apply -f deployment.yaml
```

```
#check the status of pods
kubectl get pods
```

```

root@labs-demo-vm:~/labs-sample-app/gke# kubectl apply -f deployment.yaml
W1114 11:13:15.846520    4287 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
deployment.apps/myapp created
root@labs-demo-vm:~/labs-sample-app/gke# kubectl get pods
W1114 11:13:27.270393    4308 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME           READY   STATUS      RESTARTS   AGE
myapp-6df67d5d8b-2wqlb  0/1   ContainerCreating   0          7s
root@labs-demo-vm:~/labs-sample-app/gke# 
```

Wait till Pod status becomes Running.

```

root@labs-demo-vm:~/labs-sample-app/gke# kubectl get pods
W1114 11:14:31.273515    4317 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME           READY   STATUS      RESTARTS   AGE
myapp-6df67d5d8b-2wqlb  1/1   Running     0          71s
root@labs-demo-vm:~/labs-sample-app/gke# 
```

- Expose the service as a load balancer service to view the application.

```
#deploy load balancer service
kubectl apply -f service.yaml
```

```
#check the status of service
kubectl get svc
```

```

root@labs-demo-vm:~/labs-sample-app/gke# kubectl apply -f service.yaml
W1114 11:16:57.659652    4336 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
service/myapp created
root@labs-demo-vm:~/labs-sample-app/gke# kubectl get svc
W1114 11:17:10.501661    4344 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME        TYPE      CLUSTER-IP   EXTERNAL-IP     PORT(S)   AGE
kubernetes  ClusterIP  10.91.0.1   <none>        443/TCP   118m
myapp       LoadBalancer 10.91.3.17  <pending>     80:30170/TCP 12s
root@labs-demo-vm:~/labs-sample-app/gke#

```

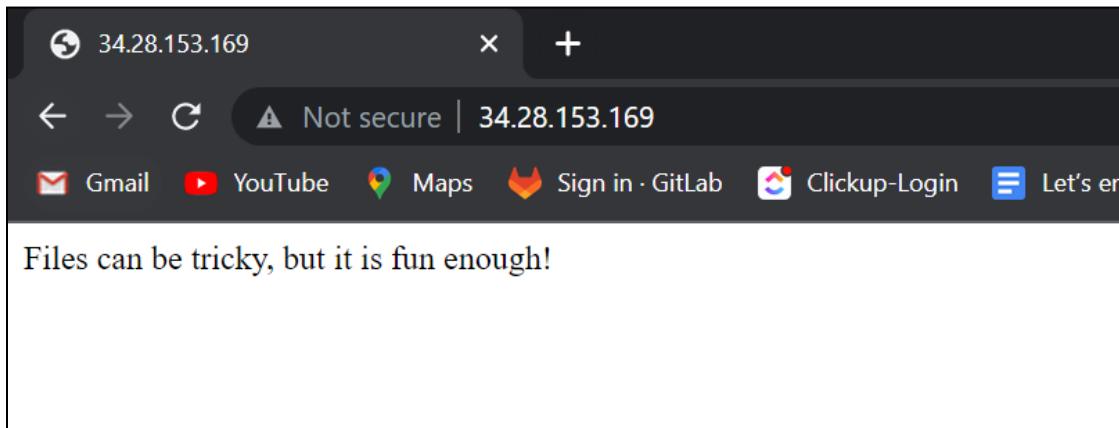
Service will create an external load balancer, wait till it assigns a public ip.

```

root@labs-demo-vm:~/labs-sample-app/gke# kubectl get svc
W1114 11:17:47.163593    4352 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME        TYPE      CLUSTER-IP   EXTERNAL-IP     PORT(S)   AGE
kubernetes  ClusterIP  10.91.0.1   <none>        443/TCP   118m
myapp       LoadBalancer 10.91.3.17  34.28.153.169  80:30170/TCP 49s
root@labs-demo-vm:~/labs-sample-app/gke#

```

Once the external-ip is assigned to service, use the external ip and open browser to view the application. Application is successfully deployed on GKE.



AutoScaling in Google Kubernetes Engine

Overview

When demand is high, the cluster autoscaler adds nodes to the node pool. When demand is low, the cluster autoscaler scales back down to a minimum size that you designate. This can increase the availability of your workloads when you need it, while controlling costs.

Cluster autoscaler increases or decreases the size of the node pool automatically by adding or removing virtual machine (VM) instances in the underlying Compute Engine Managed Instance Group (MIG) for the node pool.

When you first deploy your workload to a Kubernetes cluster, you may not be sure about its resource requirements and how those requirements might change depending on usage patterns, external dependencies, or other factors. Horizontal Pod autoscaling

helps to ensure that your workload functions consistently in different situations, and allows you to control costs by only paying for extra capacity when you need it.

Setup for Autoscaling

- The previously created cluster already has Cluster autoscaler enabled to automatically scale the number of nodes.
- For Pods Autoscaling, Horizontal Pod Autoscaler Kubernetes object manifest file needs to be deployed.
- Change directory to gke using cd **/labs-sample-app/gke**. This folder contains Kubernetes deployment, service and Horizontal Pod Autoscaling manifest files.
cd /labs-sample-app/gke.
- Deploy the HPA using the following command

```
#deploy the HPA object
kubectl apply -f hpa.yaml
```

```
#check the status of hpa
kubectl get hpa
```

The following HPA deployment scales pods from 1 to 6 if CPU utilization goes above 5%. If load on application is greater than the threshold value, it triggers pods to autoscale.

```
root@labs-demo-vm:~/labs-sample-app/gke# cat hpa.yaml
apiVersion: autoscaling/v1
kind: HorizontalPodAutoscaler
metadata:
  name: myapp
spec:
  scaleTargetRef:
    apiVersion: apps/v1
    kind: Deployment
    name: myapp
  minReplicas: 1
  maxReplicas: 6
  targetCPUUtilizationPercentage: 5
root@labs-demo-vm:~/labs-sample-app/gke# kubectl apply -f hpa.yaml
W1114 11:35:12.088688 4435 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
horizontalpodautoscaler.autoscaling/myapp created
root@labs-demo-vm:~/labs-sample-app/gke# kubectl get hpa
W1114 11:35:16.131737 4440 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME      REFERENCE   TARGETS   MINPODS   MAXPODS   REPLICAS   AGE
myapp    Deployment/myapp <unknown>/5%  1         6          0           4s
root@labs-demo-vm:~/labs-sample-app/gke#
```

- To validate the load testing install the Apache Benchmark tool. Follow the documentation - [Apache Benchmark](#).

```
root@labs-demo-vm:~/labs-sample-app/gke# ab -V
This is ApacheBench, Version 2.3 <$Revision: 1807734 $>
Copyright 1996 Adam Twiss, Zeus Technology Ltd, http://www.zeustech.net/
Licensed to The Apache Software Foundation, http://www.apache.org/
root@labs-demo-vm:~/labs-sample-app/gke#
```

- Open 2 more tabs of VM, one to view pod autoscaling and other to view node autoscaling.

Before autoscaling: Watch the nodes & Nodes

Terminal 2: kubectl get nodes -w

Terminal 3: kubectl get pods -w

For autoscaling : Get the external IP of your service. Replace it with your service IP address

Terminal 1 (In black): Enter **ab -n 10000 -c 100 'http://<svc-external-ip>/'**

-n = number of request

-c = number of concurrent request

Terminal 2 (In blue): View Node autoscaling

The number of nodes autoscaled from 1 to 3. Maximum number of nodes will be autoscaled is 4 depending on the traffic/load.

Terminal 3 (In green): View Pod Autoscaling

The number of pods autoscaled from 1 to 6. Many pods will get created but only 6 will be in active/running state as per the HPA definitions.

The screenshot shows three terminal windows side-by-side:

- Terminal 1 (Black):** Shows the command `ab -n 10000 -c 100 'http://<svc-external-ip>/'` being run.
- Terminal 2 (Blue):** Shows the command `kubectl get nodes -w` output. It lists four nodes: `gke-labs-demo-cluster-default-pool-5814bf37-9tfw` (Ready, v1.23.8-gke.1900), `gke-labs-demo-cluster-default-pool-5814bf37-9tfw` (Ready, v1.23.8-gke.1900), `gke-labs-demo-cluster-default-pool-5814bf37-9tfw` (Ready, v1.23.8-gke.1900), and `gke-labs-demo-cluster-default-pool-5814bf37-9tfw` (Ready, v1.23.8-gke.1900).
- Terminal 3 (Green):** Shows the command `kubectl get pods -w` output. It lists 12 pods in the `myapp` namespace, all in Pending state. The names include `myapp-6df67d5d8b-2wqlb`, `myapp-6df67d5d8b-z51kx`, `myapp-6df67d5d8b-z51kx`, `myapp-6df67d5d8b-1hpdp8`, `myapp-6df67d5d8b-hd6rp`, `myapp-6df67d5d8b-1hpdp8`, `myapp-6df67d5d8b-1hpdp8`, `myapp-6df67d5d8b-z51kx`, `myapp-6df67d5d8b-hd6rp`, `myapp-6df67d5d8b-z51kx`, `myapp-6df67d5d8b-55v6c`, `myapp-6df67d5d8b-55v6c`, and `myapp-6df67d5d8b-dfsqf`.

Post Autoscaling:

Nodes and Pods count will decrease to its initial value pre autoscaling.

AutoScaling in Google Kubernetes Engine based on Pub/Sub Triggers

Overview

There are certain use cases where scaling horizontally based on cpu usage does not really work well. Let's say you have a consumer worker pool running in Kubernetes. The consumers are pulling messages from a PubSub topic. When the queue is filling up we want more workers to process the messages quickly. On the other hand, when the queue is empty, we don't want to pay for a big worker pool that sits idle. With PubSub Stackdriver metrics adapter running on GKE we can easily autoscale our worker pool for minimum latency and maximum cost-effectivity.

This lab shows autoscaling based on number of undelivered messages in a Cloud Pub/Sub subscription,

Setup for Pub/Sub Metrics based Autoscaling

- In the Cloud Console, on the **Navigation menu** (≡), click **Pub/Sub > Topics**.
- To create a new topic, click **CREATE TOPIC**.
- A dialogue box will appear. Enter **Topic ID** as it is a mandatory field. Leave other fields with default values. Click on **CREATE TOPIC** to create a new topic.

The screenshot shows the 'Create a topic' dialog box. At the top, it says 'Create a topic'. Below that, a descriptive text states: 'A topic forwards messages from publishers to subscribers.' The 'Topic ID *' field contains 'echo'. The 'Topic name' field shows 'projects [REDACTED]/topics/echo'. There are three checkboxes: 'Add a default subscription' (checked), 'Use a schema' (unchecked), and 'Set message retention duration (not free)' (unchecked). Under 'Encryption', there are two options: 'Google-managed encryption key' (selected) and 'Customer-managed encryption key (CMEK)' (unchecked). The 'Google-managed encryption key' option includes the note 'No configuration required'. At the bottom, there are 'CANCEL' and 'CREATE TOPIC' buttons.

Topic named echo is created.

Pub/Sub		Topics	+ CREATE TOPIC	DELETE	SHOW INFO PANEL			
Pub/Sub		LIST METRICS						
Topics		Filter echo Filter topics		X ?				
		Topic ID ↑	Encryption key	Topic name	Retention			
		echo	Google-managed	projects/[REDACTED]/topics/echo	-	⋮		

- Create a custom subscription. Click on topic, In Subscription it will contain the default subscription which was created with topic creation.

Pub/Sub		← echo	EDIT	+ TRIGGER CLOUD FUNCTION	IMPORT	DELETE
Pub/Sub		Topic name projects/searce-playground-v1/topics/echo				
Topics						
Subscriptions						
Schemas						
Pub/Sub Lite		SUBSCRIPTIONS SNAPSHOT MESSAGES METRICS DETAILS				
Lite Reservations		Only subscriptions attached to this topic are displayed. A subscription captures the stream of messages published to a given topic. You can also stream m creating a subscription from a Cloud Dataflow job. Learn more				
Lite Topics		CREATE SUBSCRIPTION EXPORT				
Lite Subscriptions		Filter subscriptions				
		Subscription ID ↑	Subscription name	Project	⋮	
		echo-sub	projects/searce-playground-v1/subscriptions/echo-sub	searce-playground-v1	⋮	

- Click on **CREATE SUBSCRIPTION**, to create a custom subscription.
- Enter Subscription ID as it is a mandatory field. Select Delivery type as **Pull**.

Pub/Sub		← Add subscription to topic
Pub/Sub		A subscription directs messages on a topic to subscribers. Messages can be pushed to subscribers immediately, or subscribers can pull messages as needed.
Topics		Subscription ID * <input type="text" value="echo-read"/>
Subscriptions		Subscription name: projects/[REDACTED]/subscriptions/echo-read
Schemas		Topic name projects/searce-playground-v1/topics/echo
Pub/Sub Lite		Delivery type <input checked="" type="radio"/> Pull <input type="radio"/> Push <input type="radio"/> Write to BigQuery
Lite Reservations		Message retention duration Duration is from 10 minutes to 7 days
Lite Topics		Days <input type="text" value="7"/> Hours <input type="text" value="0"/> Minutes <input type="text" value="0"/>
Lite Subscriptions		<input type="checkbox"/> Retain acknowledged messages When enabled, acknowledged messages are retained for the message retention duration specified above. This increases message storage fees. Learn more
Release Notes		
<		

- Keep the rest of the fields with default values and click on **CREATE**.

Dead lettering

Enable dead lettering

Subscriptions may configure a maximum number of delivery attempts. When a message cannot be delivered, it is republished to the specified dead letter topic.

Retry policy

Retry policy will be triggered on NACKs or acknowledgement deadline exceeded events for a given message. [Learn more](#)

Retry immediately

Retry after exponential backoff delay

CREATE

- Subscription named **echo-read** is created.

The screenshot shows the Google Cloud Pub/Sub interface. On the left, there's a sidebar with 'Pub/Sub' and 'Pub/Sub Lite' sections. Under 'Pub/Sub', there are 'Topics', 'Subscriptions', 'Schemas', and 'Schemas'. Under 'Pub/Sub Lite', there are 'Lite Reservations', 'Lite Topics', and 'Lite Subscriptions'. The main area shows a topic named 'echo' with a 'Topic name' field containing 'projects/[REDACTED]topics/echo'. Below the topic name, there are tabs for 'SUBSCRIPTIONS', 'SNAPSHOTS', 'MESSAGES', 'METRICS', and 'DETAILS'. A note says 'Export options have moved to the Create subscription dropdown menu under the Subscriptions tab below.' There's a 'CREATE SUBSCRIPTION' button and an 'EXPORT' button. A 'Filter' section allows filtering subscriptions by 'Subscription ID' and 'Subscription name'. Two subscriptions are listed: 'echo-read' and 'echo-sub'. Both are associated with 'projects/[REDACTED]' and have a 'Project' column with a redacted value. Each subscription has a three-dot menu icon on the right.

Subscription ID	Subscription name	Project
echo-read	projects/[REDACTED]subscriptions/echo-read	[REDACTED]
echo-sub	projects/[REDACTED]/subscriptions/echo-sub	[REDACTED]

- Deploy the deployment.yaml and service.yaml from the cloned repository. Follow **Deploying C++ Web Application on GKE** described above.

```

root@labs-demo-vm:~/labs-sample-app/gke# kubectl apply -f deployment.yaml
W1116 10:03:00.790901 25901 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
deployment.apps/myapp created
root@labs-demo-vm:~/labs-sample-app/gke# kubectl apply -f service.yaml
W1116 10:03:11.658616 25908 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
service/myapp created
root@labs-demo-vm:~/labs-sample-app/gke# kubectl get pods
W1116 10:03:18.182939 25913 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME READY STATUS RESTARTS AGE
myapp-6df67d5d8b-pjqkc 1/1 Running 0 16s
root@labs-demo-vm:~/labs-sample-app/gke# kubectl get svc
W1116 10:03:21.590736 25917 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME TYPE CLUSTER-IP EXTERNAL-IP PORT(S) AGE
kubernetes ClusterIP 10.90.0.1 <none> 443/TCP 9m28s
myapp LoadBalancer 10.90.3.49 <pending> 80:32564/TCP 9s
root@labs-demo-vm:~/labs-sample-app/gke# █

```

- Delete the previously created HPA object by using the following command.

kubectl delete -f hpa.yaml

Or

kubectl get hpa
kubectl delete <hpa-name>

- Create a custom service account which will be required for stackdriver authorization. Follow the below commands

Note: Add your Project ID in variable \$GCP_PROJECT_ID.

\$GCP_PROJECT_ID = <project-id>

```

#create service account
gcloud iam service-accounts create custom-metrics-sd-adapter --project
"$GCP_PROJECT_ID"

#add roles to created service account
gcloud projects add-iam-policy-binding "$GCP_PROJECT_ID" \
--member
"serviceAccount:custom-metrics-sd-adapter@$GCP_PROJECT_ID.iam.gservice
account.com" \
--role "roles/monitoring.editor"

```

```

root@labs-demo-vm:~/labs-sample-app/gke# gcloud iam service-accounts create custom-metrics-sd-adapter --project
"searce-playground-v1"
Created service account [custom-metrics-sd-adapter].
root@labs-demo-vm:~/labs-sample-app/gke# gcloud projects add-iam-policy-binding "searce-playground-v1" \
>   --member "serviceAccount:custom-metrics-sd-adapter@searce-playground-v1.iam.gserviceaccount.com" \
>   --role "roles/monitoring.editor"
[1] EXPRESSION=request.time < timestamp("2022-03-16T19:34:18.912Z"), TITLE=expire_10_hours
[2] EXPRESSION=request.time < timestamp("2022-03-21T20:44:35.924Z"), TITLE=expire_access_12_hours
[3] EXPRESSION=request.time < timestamp("2022-04-01T20:52:43.232Z"), TITLE=expire_access_6_hours
[4] EXPRESSION=request.time < timestamp("2022-05-09T20:33:53.852Z"), TITLE=expire_access_6_hours
[5] EXPRESSION=request.time < timestamp("2022-05-26T19:39:42.851Z"), TITLE=expire_in_8_hours
[6] EXPRESSION=request.time < timestamp("2022-05-26T19:40:05.950Z"), TITLE=expire_in_8_hours
[7] EXPRESSION=request.time < timestamp("2022-10-17T15:30:00.000Z"), TITLE=expire_after_8_hours
[8] EXPRESSION=request.time < timestamp("2022-10-17T18:28:28.149Z"), TITLE=expire_access_8_hours
[9] EXPRESSION=request.time < timestamp("2022-11-11T16:30:00.000Z"), TITLE=valid till 11th nov
[10] EXPRESSION=request.time.getDayOfWeek() <= 6, TITLE=A week access
[11] None
[12] Specify a new condition
The policy contains bindings with conditions, so specifying a condition is required when adding a binding.
Please specify a condition.: 10

```

- To deploy a stackdriver adapter, various IAM roles and bindings are created. To deploy stackdriver IAM roles follow the following steps.

cd /labs-sample-app/gke/stackdriver

#deploy stackdriver iam-roles-bindings
kubectl apply -f iam-adapter.yaml

Few resources are created and deployed by using the above command.

```

root@labs-demo-vm:~/labs-sample-app/gke/stackdriver# kubectl apply -f iam-adapter.yaml
W1116 10:18:48.018201    26140 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v
1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
namespace/custom-metrics created
serviceaccount/custom-metrics-stackdriver-adapter created
clusterrolebinding.rbac.authorization.k8s.io/custom-metrics:system:auth-delegator created
rolebinding.rbac.authorization.k8s.io/custom-metrics-auth-reader created
clusterrolebinding.rbac.authorization.k8s.io/custom-metrics-resource-reader created
root@labs-demo-vm:~/labs-sample-app/gke/stackdriver#

```

- Impersonating & Binding the Google Managed Service account with Kubernetes service account using the following commands. Enter your project ID.

#iam binding
**gcloud iam service-accounts add-iam-policy-binding **
**--role roles/iam.workloadIdentityUser **
--member
"serviceAccount:\$GCP_PROJECT_ID.svc.id.goog[custom-metrics/custom-metri
**cs-stackdriver-adapter]" **
"custom-metrics-sd-adapter@\$GCP_PROJECT_ID.iam.gserviceaccount.com"
#annotating k8s service account
**kubectl annotate serviceaccount custom-metrics-stackdriver-adapter **
"iam.gke.io/gcp-service-account=custom-metrics-sd-adapter@\$GCP_PROJECT
**_ID.iam.gserviceaccount.com" **
--namespace custom-metrics

```

root@labs-demo-vm:~/labs-sample-app/gke/stackdriver# gcloud iam service-accounts add-iam-policy-binding \
>   --role roles/iam.workloadIdentityUser \
>   --member "serviceAccount:searce-playground-v1.svc.id.goog[custom-metrics/custom-metrics-stackdriver-adapter]" \
" \
>   "custom-metrics-sd-adapter@searce-playground-v1.iam.gserviceaccount.com"
Updated IAM policy for serviceAccount [custom-metrics-sd-adapter@searce-playground-v1.iam.gserviceaccount.com].
bindings:
- members:
  - serviceAccount:searce-playground-v1.svc.id.goog[custom-metrics/custom-metrics-stackdriver-adapter]
    role: roles/iam.workloadIdentityUser
etag: BwXtk95fyZA=
version: 1
root@labs-demo-vm:~/labs-sample-app/gke/stackdriver# kubectl annotate serviceaccount custom-metrics-stackdriver-
adapter \
>   "iam.gke.io/gcp-service-account=custom-metrics-sd-adapter@searce-playground-v1.iam.gserviceaccount.com" \
>   --namespace custom-metrics
W1116 10:23:56.865843    26207 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v
1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
serviceaccount/custom-metrics-stackdriver-adapter annotated
root@labs-demo-vm:~/labs-sample-app/gke/stackdriver# 
```

- Deploying the stackdriver adapter deployment using the following command.

```
cd /labs-sample-app/gke/stackdriver
```

```
#deploy stackdriver deployment
kubectl apply -f deploy-adapter.yaml
```

```
#get all objects deployed
kubectl get all -n custom-metrics
```

```

root@labs-demo-vm:~/labs-sample-app/gke/stackdriver# kubectl apply -f deploy-adapter.yaml
W1116 10:25:26.168779    26213 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v
1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
deployment.apps/custom-metrics-stackdriver-adapter created
service/custom-metrics-stackdriver-adapter created
apiservice.apiregistration.k8s.io/v1beta1.custom.metrics.k8s.io created
apiservice.apiregistration.k8s.io/v1beta2.custom.metrics.k8s.io created
apiservice.apiregistration.k8s.io/v1beta1.external.metrics.k8s.io created
Warning: resource clusterroles/external-metrics-reader is missing the kubectl.kubernetes.io/last-applied-configuration
annotation which is required by kubectl apply. kubectl apply should only be used on resources created declaratively by either kubectl create --save-config or kubectl apply. The missing annotation will be patched automatically.
clusterrole.rbac.authorization.k8s.io/external-metrics-reader configured
clusterrolebinding.rbac.authorization.k8s.io/external-metrics-reader created
root@labs-demo-vm:~/labs-sample-app/gke/stackdriver# kubectl get all -n custom-metrics
W1116 10:25:39.962379    26218 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v
1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME                           READY   STATUS    RESTARTS   AGE
pod/custom-metrics-stackdriver-adapter-5885cc597f-jsftr   0/1     Pending   0          14s
NAME                         TYPE        CLUSTER-IP      EXTERNAL-IP   PORT(S)      AGE
service/custom-metrics-stackdriver-adapter   ClusterIP   10.90.3.138   <none>       443/TCP     14s
NAME                           READY   UP-TO-DATE   AVAILABLE   AGE
deployment.apps/custom-metrics-stackdriver-adapter   0/1     1           0          14s
NAME                           DESIRED  CURRENT    READY   AGE
replicaset.apps/custom-metrics-stackdriver-adapter-5885cc597f   1         1         0          14s
root@labs-demo-vm:~/labs-sample-app/gke/stackdriver# 
```

- Stackdriver adapter is successfully deployed. It's time to deploy Pub/sub **number of undelivered/unacknowledged** Metric based Horizontal Pod Autoscaler using the following command.

```
cd /labs-sample-app/gke
```

```
kubectl apply -f pubsub-hpa.yaml
```

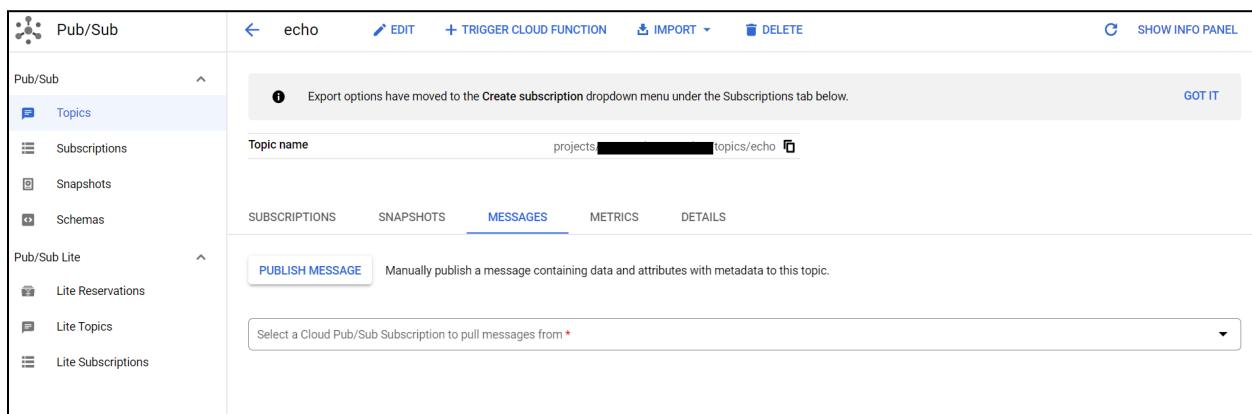
```
kubectl get hpa
```

We'll autoscale between 1–5 replicas, based on external metric **pubsub.googleapis.com|subscription|numUndeliveredMessages** from our echo-read subscription.

Target value is 2 undelivered messages.

```
root@labs-demo-vm:~/labs-sample-app/gke# kubectl apply -f pubsub-hpa.yaml
W1116 11:21:16.111308 26740 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
Warning: autoscaling/v2beta1 HorizontalPodAutoscaler is deprecated in v1.22+, unavailable in v1.25+; use autoscaling/v2 HorizontalPodAutoscaler
horizontalpodautoscaler.autoscaling/myapp created
root@labs-demo-vm:~/labs-sample-app/gke# kubectl get hpa
W1116 11:21:33.437378 26746 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME      REFERENCE      TARGETS      MINPODS      MAXPODS      REPLICAS      AGE
myapp    Deployment/myapp  <unknown>/2 (avg)   1            6            0           16s
root@labs-demo-vm:~/labs-sample-app/gke#
```

- Navigate to pub/sub topic created previously named **echo**. Open the topic, go to the **MESSAGES** section. Here we will publish some messages. Click on **PUBLISH MESSAGE**.



- A dialogue box will appear to publish the message. Lets publish 10 messages.
Enter **Number of messages** to **10**.
In the **Message body** enter your message like **hello**.
Finally, click on **PUBLISH** to publish the messages.

Publish message

Topic name
projects/[REDACTED]/topics/echo

PUBLISH MESSAGE

Number of messages *
10

Enter an amount between 1-100.

Message interval (seconds) *
1

How long to wait before publishing the next message

Message body

The message you want to publish to this topic. Either message or attribute will be required to publish.

Message *
Hello

PUBLISH **CANCEL**

- It will start publishing the messages.

The publish job has completed.

Pending	0
Success	10
Error	0

DISMISS

Export options have moved to the Create subscription dropdown menu under the Subscriptions tab below.

GOT IT

Topic name
projects/[REDACTED]/topics/echo

SUBSCRIPTIONS SNAPSHOTTS **MESSAGES** METRICS DETAILS

PUBLISH MESSAGE Manually publish a message containing data and attributes with metadata to this topic.

- As there are 10 undelivered/unacknowledged messages in queue, as per Horizontal Pod Autoscaler it will trigger to scale more nodes. Open 2 terminals to view Node & Pod autoscaling.

Terminal 1(Black): Node autoscaling

The number of nodes autoscaled from 1 to 3.

Terminal 2(Green): Pod Autoscaling

The number of pods autoscaled from 1 to 5. Many pods will get created but only 5 will be in active/running state as per the HPA definitions.

```

SSH-in-browser
root@labs-demo-vm:~/labs-sample-app/gke# kubectl get nodes -w
W1116 11:24:10.682267 27034 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME STATUS ROLES AGE VERSION
gke-labs-demo-cluster-default-pool-b7belb34-qx4r Ready <none> 8m26s v1.23.12-g
ke.100
gke-labs-demo-cluster-default-pool-b7belb34-sfdv Ready <none> 3m52s v1.23.12-g
ke.100
gke-labs-demo-cluster-default-pool-b7belb34-sfdv Ready <none> 5m1s v1.23.12-g
ke.100

SSH-in-brov
root@labs-demo-vm:~# kubectl get pods -w
W1116 11:24:23.155658 27062 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME READY STATUS RESTARTS AGE
myapp-6df67d5d8b-rxx6s 1/1 Running 0 4m4s
myapp-6df67d5d8b-tczzm 0/1 Pending 0 0s
myapp-6df67d5d8b-tczzm 0/1 Pending 0 0s
myapp-6df67d5d8b-zng2d 0/1 Pending 0 0s
myapp-6df67d5d8b-jlpdg 0/1 Pending 0 0s
myapp-6df67d5d8b-zng2d 0/1 Pending 0 0s
myapp-6df67d5d8b-zng2d 0/1 Pending 0 0s
myapp-6df67d5d8b-jlpdg 0/1 Pending 0 0s
myapp-6df67d5d8b-tczzm 0/1 ContainerCreating 0
myapp-6df67d5d8b-zng2d 0/1 ContainerCreating 0
myapp-6df67d5d8b-tczzm 1/1 Running 0 7s
myapp-6df67d5d8b-zng2d 1/1 Running 0 0s
myapp-6df67d5d8b-d89b6 0/1 Pending 0 0s
myapp-6df67d5d8b-d89b6 0/1 Pending 0 0s

```

- Nodes and Pods autoscaled Successfully based on number of undelivered/unacknowledged messages in Pub/Sub.

```

root@labs-demo-vm:~/labs-sample-app/gke# kubectl get nodes
W1116 11:26:34.549975 27087 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME STATUS ROLES AGE VERSION
gke-labs-demo-cluster-default-pool-b7belb34-7w4m Ready <none> 51s v1.23.12-g
ke.100
gke-labs-demo-cluster-default-pool-b7belb34-qx4r Ready <none> 10m v1.23.12-g
ke.100
gke-labs-demo-cluster-default-pool-b7belb34-sfdv Ready <none> 6m16s v1.23.12-g
ke.100
root@labs-demo-vm:~/labs-sample-app/gke# 

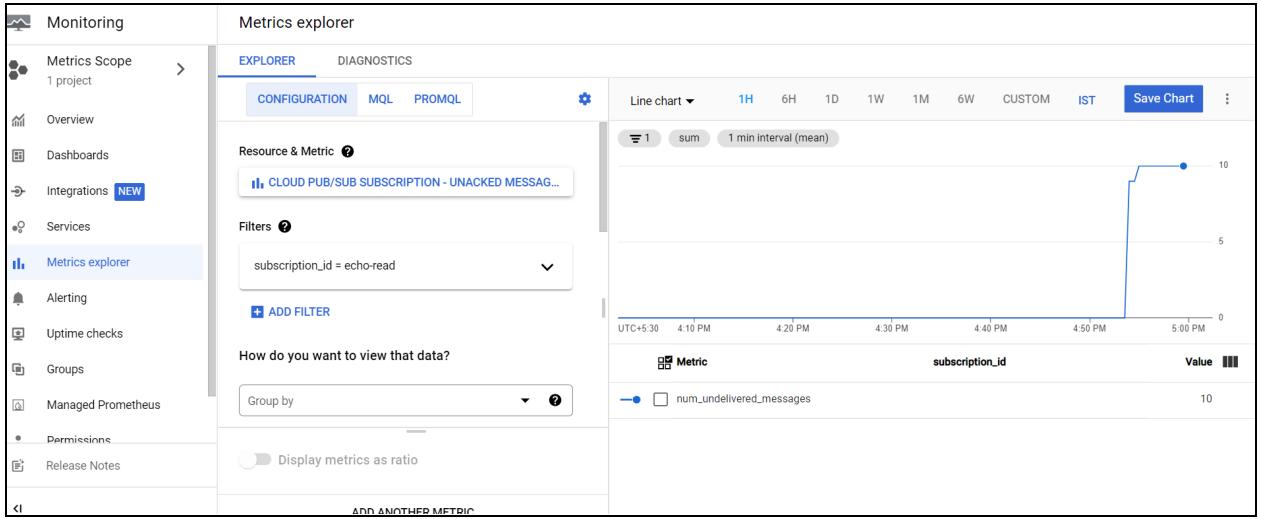
root@labs-demo-vm:~# kubectl get pods
W1116 11:26:30.216736 27083 gcp.go:119] WARNING: the gcp auth plugin is deprecated in v1.22+, unavailable in v1.26+; use gcloud instead.
To learn more, consult https://cloud.google.com/blog/products/containers-kubernetes/kubectl-auth-changes-in-gke
NAME READY STATUS RESTARTS AGE
myapp-6df67d5d8b-d89b6 1/1 Running 0 84s
myapp-6df67d5d8b-jlpdg 1/1 Running 0 99s
myapp-6df67d5d8b-rxx6s 1/1 Running 0 6m11s
myapp-6df67d5d8b-tczzm 1/1 Running 0 99s
myapp-6df67d5d8b-zng2d 1/1 Running 0 99s
root@labs-demo-vm:~# 

```

- Open a new tab. In the Cloud Console, on the **Navigation menu (≡)**, click **Monitoring > Metric Explorer**.
- Click on Resource & Metric, search Pub/Sub. In Pub/Sub search for **Unacked messages** metrics and click on **Apply**.

The screenshot shows the Google Cloud Metrics Explorer interface. On the left, the navigation menu is open, showing options like Monitoring, Metrics Scope, Overview, Dashboards, Integrations, Services, Metrics explorer (which is selected), Alerting, Uptime checks, Groups, Managed Prometheus, Permissions, and Release Notes. The Metrics explorer section has a sidebar with 'ACTIVE METRICS' and a list of metrics related to 'Unacked messages'. The main area displays the 'Metrics explorer' configuration for a 'Cloud Pub/Sub Subscription'. It shows the 'Name' as 'Unacked messages', 'Description' as 'Number of unacknowledged messages (a.k.a. backlog messages) in a subscription.', 'Metric' as 'pubsub.googleapis.com/subscription/num_unacked_messages', 'Resource types' as 'pubsub_subscription', 'Unit' as '1', 'Kind' as 'GAUGE', and 'Value type' as 'INT64'. Below this, there's a list of 'Cloud Pub/Sub Subscription' and 'Cloud Pub/Sub Topic' metrics, each with 29 and 17 metrics respectively. A toggle switch at the bottom allows selecting active resources and metrics. At the bottom right, there are 'Cancel' and 'Apply' buttons.

- It will show no of undelivered/unacknowledged messages in Pub/Sub. As we have published 10 messages the graph shows the same 10 messages.



- Let's acknowledge the messages, navigate to Pub/Sub topic. Go in **MESSAGES**, select the subscription **echo-read**.

The publish job has completed.

Pending	0
Success	10
Error	0

Export options have moved to the Create subscription dropdown menu under the Subscriptions tab below.

Topic name: projects/[REDACTED]/topics/echo

Filter [Type to filter]

projects/[REDACTED]/subscriptions/echo-sub
projects/[REDACTED]/subscriptions/echo-read

REFRESH CREATE A SUBSCRIPTION CANCEL OK

SUBSCRIPTIONS SNAPSHOTS MESSAGES METRICS DETAILS

PUBLISH MESSAGE Manually publish a message containing data and attributes with metadata to this topic.

Select a Cloud Pub/Sub Subscription to pull messages from * projects/seaice-playground-v1/subscriptions/echo-read

Click Pull to view messages and temporarily delay message delivery to other subscribers. Select Enable ACK messages and then click ACK next to the message to permanently prevent message delivery to other subscribers.

PULL Enable ack messages

Filter Filter messages

Publish time	Attribute keys	Message body	Ordering key	Ack
No message found yet				↑

- As it is a Pull based subscription, Click on **PULL** & select the checkbox **Enable ack messages**. It will list all the published messages with the **ACK** field at the end.

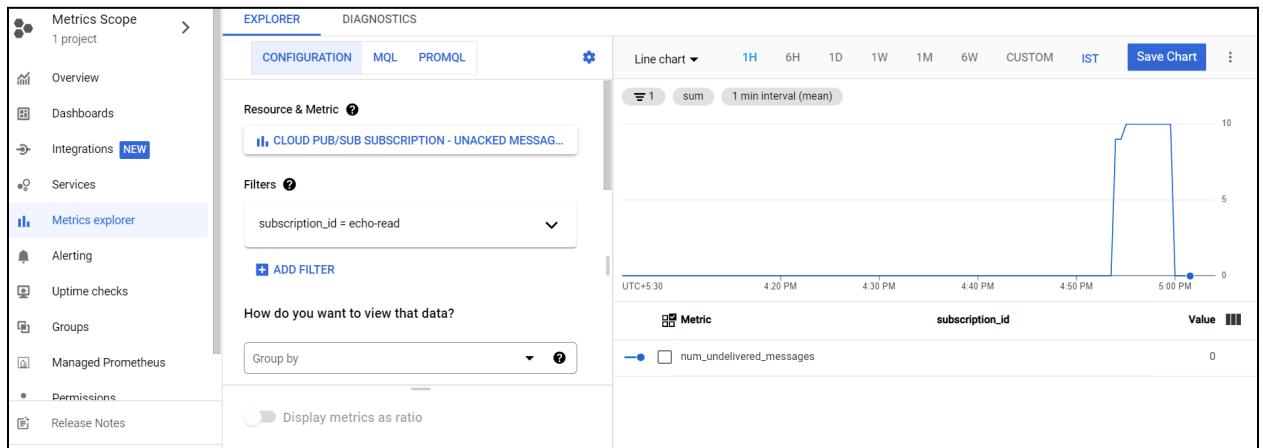
Note : ACK will show as a deadline exceed in some second so complete the ACK step ASAP

	Publish time	Attribute keys	Message body	Ordering key	Ack
Nov 16, 2022, 4:53:16 PM	—	Hello	—	ACK	
Nov 16, 2022, 4:53:16 PM	—	Hello	—	ACK	
Nov 16, 2022, 4:53:17 PM	—	Hello	—	ACK	
Nov 16, 2022, 4:53:18 PM	—	Hello	—	ACK	
Nov 16, 2022, 4:53:20 PM	—	Hello	—	ACK	
Nov 16, 2022, 4:53:20 PM	—	Hello	—	ACK	
Nov 16, 2022, 4:53:21 PM	—	Hello	—	ACK	
Nov 16, 2022, 4:53:22 PM	—	Hello	—	ACK	
Nov 16, 2022, 4:53:23 PM	—	Hello	—	ACK	
Nov 16, 2022, 4:53:24 PM	—	Hello	—	ACK	

- Click on **ACK** of each message which is at the end of row or last column. As soon as you click on **ack** messages will disappear from the list.

	Publish time	Attribute keys	Message body	Ordering key	Ack
Nov 16, 2022, 4:53:23 PM	—	Hello	—	ACK	

- Visit the monitoring tab which shows the details above unacknowledged messages. Refresh the page to view the dip in unacknowledged messages from 10 to 0.



- As there are no messages left unacknowledged HPA will be triggered to scale down the pods as well as nodes.

Conclusion

In this lab, we deployed C++ web application on the GKE cluster. Autoscaled the Nodes & pods of application by using Horizontal Pod Autoscaler based on CPU utilization. Also, Autoscaled the Nodes & pods of application by using Horizontal Pod Autoscaler based on Pub/Sub trigger using external Pub/Sub metric.

References:

- Google Kubernetes Engine - [GKE](#)
- Pub/Sub in GCP - [Pub/Sub](#)
- Overview about Kubernetes - [K8s](#)