Automated VM Scaling and Security on GCP

Aryan Kumar M23CSA510

March 1, 2025

GitHub Repository: `https://github.com/Aryank47/gcp-autoscaling`

Abstract

This report outlines step-by-step instructions for creating a VM on Google Cloud Platform (GCP), configuring auto-scaling based on CPU utilization, and implementing security measures such as IAM roles and firewall rules. The architecture ensures scalability, security, and automated testing of the auto-scaling policy.

Contents

1 Introduction

Google Cloud Platform (GCP) is a robust cloud infrastructure for deploying scalable applications. This document demonstrates how to set up a VM instances with auto-scaling and security policies.

2 Step-by-Step Implementation

2.1 Step 1: Create a VM Instance on GCP

2.2 config

```
    ZONE=asia-south1-c
    MACHINE_TYPE=e2-micro
    TEMPLATE_NAME=test-template
    MIG_NAME=test-mig
    TARGET_CPU=0.7
    MAX_INSTANCES=3
    LOAD_DURATION=300
```

2.3 Authenticate and Initialize the CLI

```
gcloud auth login
gcloud config set project [PROJECT_ID]
gcloud config set compute/zone [ZONE]
```

2.4 Create a VM Instance

```
gcloud compute instance-templates create $TEMPLATE_NAME \
        --machine-type=$MACHINE_TYPE \
        --image-family=debian-11 \
        --image-project=debian-cloud \
        --tags=allow-ssh,allow-http \
        --metadata=startup-script='
#!/bin/bash
apt␣update
apt␣install␣-y␣stress-ng␣python3
(
cat␣<<-EOF
import␣os

def␣generate_stress(duration=300):
␣␣␣␣command␣=␣f"stress-ng␣--cpu␣2␣--timeout␣{duration}s"
␣␣␣␣os.system(command)

if␣__name__␣==␣"__main__":
␣␣␣␣generate_stress()
EOF
)␣>␣/usr/local/bin/load_generator.py
chmod␣+x␣/usr/local/bin/load_generator.py
␣␣␣␣␣␣␣␣'
```

2.5 Verify the Instance

```
gcloud compute instances list
```

2.6 Step 2: Configure Auto-Scaling Policy

2.7 Create a Managed Instance Group (MIG)

```
gcloud compute instance-groups managed create $MIG_NAME \
        --template=$TEMPLATE_NAME \
        --base-instance-name=test-vm \
        --zone=$ZONE \
        --size=1
```

2.8 Set Auto-Scaling Policy

```
gcloud compute instance-groups managed set-autoscaling $MIG_NAME \
        --max-num-replicas=$MAX_INSTANCES \
        --min-num-replicas=1 \
        --target-cpu-utilization=$TARGET_CPU \
        --zone=$ZONE
```

2.9 Verify Auto-Scaling Configuration

```
gcloud compute instance-groups managed describe my-mig --zone=[ZONE]
```

2.10 Step 3: Implement Security Measures

2.11 Create Firewall Rules

```
gcloud compute firewall-rules create allow-ssh \
  --allow tcp:22 \
  --target-tags=allow-ssh

gcloud compute firewall-rules create allow-http \
  --allow tcp:80 \
  --target-tags=allow-http

gcloud compute firewall-rules create deny-all \
  --deny all \
  --priority=65534
```

2.12 Configure IAM Roles

```
gcloud iam service-accounts create compute-reader --display-name="Compute␣Reader"
gcloud projects add-iam-policy-binding [PROJECT_ID] \
  --member="serviceAccount:compute-reader@[PROJECT_ID].iam.gserviceaccount.com" \
  --role="roles/compute.instanceViewer"
```

2.13 Verify IAM Policy

```
gcloud projects get-iam-policy [PROJECT_ID] --flatten="bindings[].members"
```

2.14 Step 4: Test Auto-Scaling

2.15 Generate CPU Load

```
gcloud compute ssh my-vm --zone=[ZONE] --command="stress-ng␣--cpu␣2␣--timeout␣300s"
```

2.16 Monitor Scaling Activity

```
gcloud compute instance-groups managed list-instances my-mig --zone=[ZONE]
```

3 Architecture Diagram

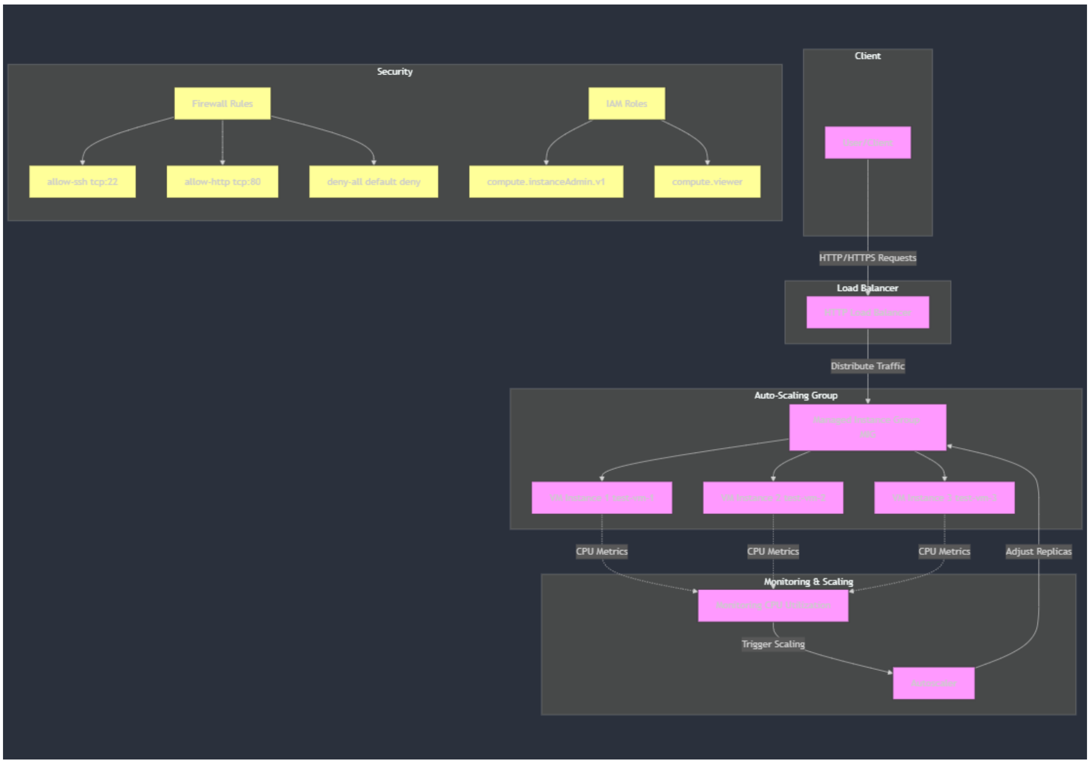
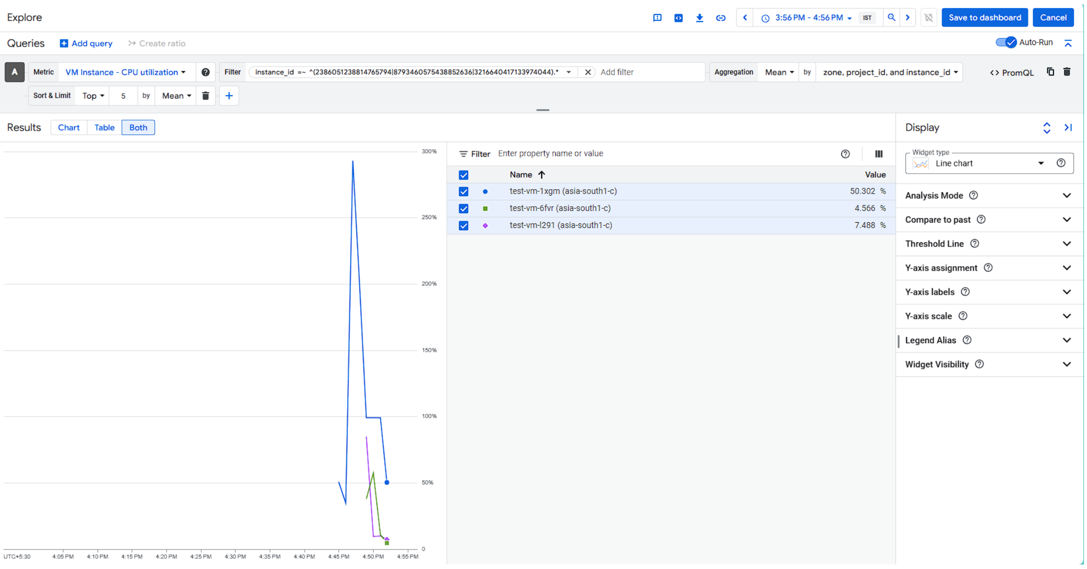


Figure 1: **GCP Auto-Scaling Architecture**

4 CPU Utilization Metrics

4.1 Initial Load Generation



Metric: VM Instance - CPU utilization

Filter: instance_id =~ ^(2386051238814765794|8793460575438852636|3216640417133974044).*

Aggregation: Mean by zone, project_id, and instance_id

Sort & Limit: Top 5 by Mean

3:56 PM - 4:56 PM IST

Auto-Run

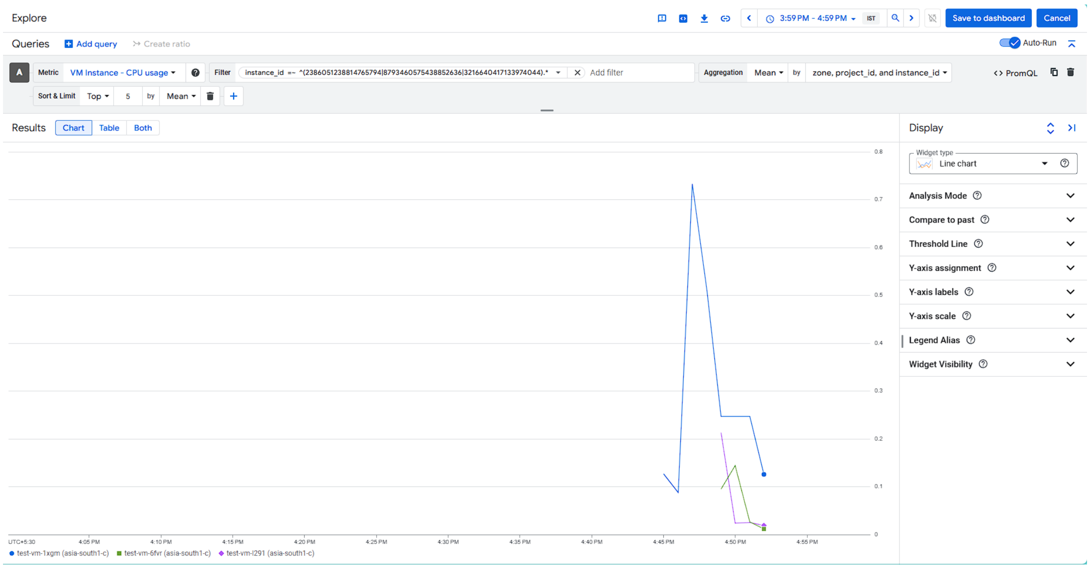
Widget type: Line chart

Name	Value
● test-vm-1xgm (asia-south1-c)	50.302 %
■ test-vm-6fvr (asia-south1-c)	4.566 %
◆ test-vm-l291 (asia-south1-c)	7.488 %

Figure 2: Initial CPU utilization spike during load generation

Description: This graph shows the CPU utilization of the VM instances during the initial load generation phase. The blue line represents the primary instance experiencing a significant spike in CPU usage, triggering the auto-scaling policy.

4.2 Auto-Scaling Response



Metric: VM Instance - CPU usage

Filter: instance_id =~ ^(2386051238814765794|8793460575438852636|3216640417133974044).*

Aggregation: Mean by zone, project_id, and instance_id

Sort & Limit: Top 5 by Mean

3:59 PM - 4:59 PM IST

Auto-Run

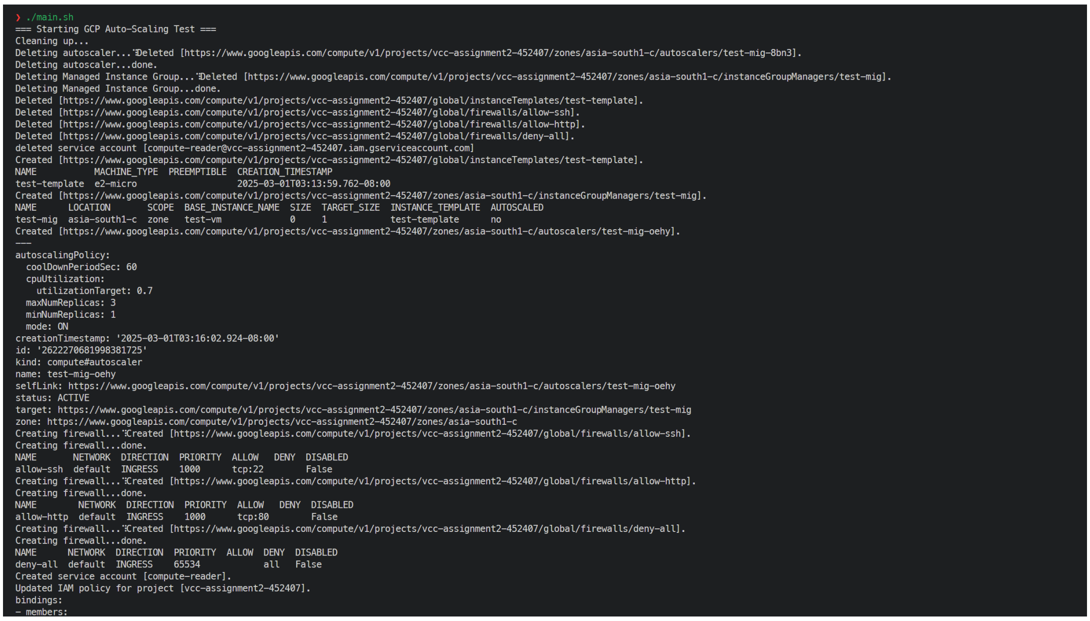
Widget type: Line chart

Figure 3: Auto-scaling response and load distribution

Description: This graph illustrates the auto-scaling response where additional instances are created to handle the load. The different colored lines represent multiple VM instances sharing the load, demonstrating the effectiveness of the auto-scaling configuration.

5 Terminal Log Screenshots

5.1 Auto-Scaling Test Execution

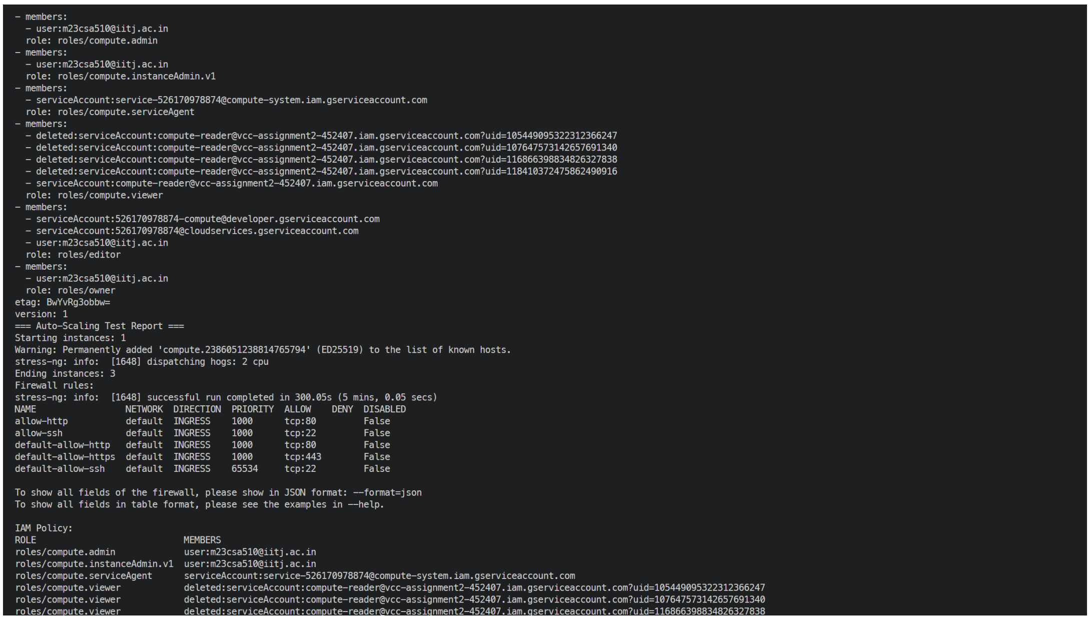


```
❯ ./main.sh
=== Starting GCP Auto-Scaling Test ===
Cleaning up...
Deleting autoscaler...⠹Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/zones/asia-south1-c/autoscalers/test-mig-8bn3].
Deleting autoscaler...done.
Deleting Managed Instance Group...⠹Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/zones/asia-south1-c/instanceGroupManagers/test-mig].
Deleting Managed Instance Group...done.
Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/instanceTemplates/test-template].
Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/firewalls/allow-ssh].
Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/firewalls/allow-http].
Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/firewalls/deny-all].
deleted service account [compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com]
Created [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/instanceTemplates/test-template].
NAME           MACHINE_TYPE  PREEMPTIBLE  CREATION_TIMESTAMP
test-template  e2-micro                   2025-03-01T03:13:59.762-08:00
Created [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/zones/asia-south1-c/instanceGroupManagers/test-mig].
NAME      LOCATION       SCOPE  BASE_INSTANCE_NAME  SIZE  TARGET_SIZE  INSTANCE_TEMPLATE  AUTOSCALED
test-mig  asia-south1-c  zone   test-vm             0     1            test-template      no
Created [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/zones/asia-south1-c/autoscalers/test-mig-oehy].
---
autoscalingPolicy:
  coolDownPeriodSec: 60
  cpuUtilization:
    utilizationTarget: 0.7
  maxNumReplicas: 3
  minNumReplicas: 1
  mode: ON
creationTimestamp: '2025-03-01T03:16:02.924-08:00'
id: '2622270681998381725'
kind: compute#autoscaler
name: test-mig-oehy
selfLink: https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/zones/asia-south1-c/autoscalers/test-mig-oehy
status: ACTIVE
target: https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/zones/asia-south1-c/instanceGroupManagers/test-mig
zone: https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/zones/asia-south1-c
Creating firewall...⠹Created [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/firewalls/allow-ssh].
Creating firewall...done.
NAME       NETWORK  DIRECTION  PRIORITY  ALLOW   DENY  DISABLED
allow-ssh  default  INGRESS    1000      tcp:22        False
Creating firewall...⠹Created [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/firewalls/allow-http].
Creating firewall...done.
NAME        NETWORK  DIRECTION  PRIORITY  ALLOW   DENY  DISABLED
allow-http  default  INGRESS    1000      tcp:80        False
Creating firewall...⠹Created [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/firewalls/deny-all].
Creating firewall...done.
NAME      NETWORK  DIRECTION  PRIORITY  ALLOW  DENY  DISABLED
deny-all  default  INGRESS    65534            all   False
Created service account [compute-reader].
Updated IAM policy for project [vcc-assignment2-452407].
bindings:
- members:
```

Figure 4: Auto-scaling test execution log showing instance scaling from 1 to 3

Description: This screenshot shows the terminal output during the auto-scaling test. The script starts with 1 instance, generates CPU load using `stress-ng`, and scales up to 3 instances as the load increases. The firewall rules and IAM policies are verified at the end.

5.2 Resource Cleanup

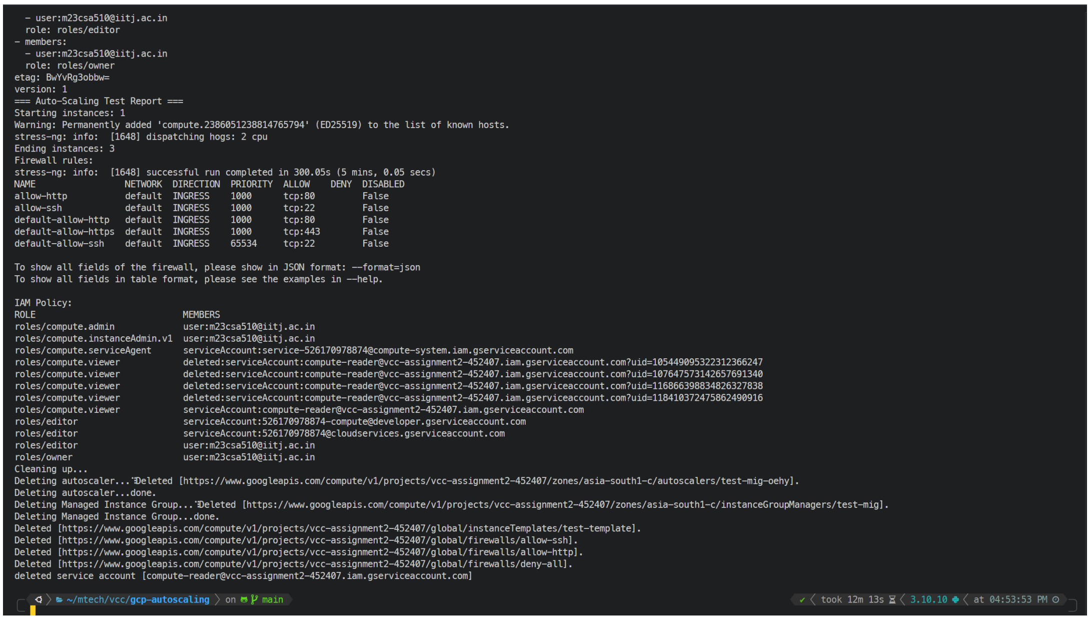


```
- members:
  - user:m23csa510@iitj.ac.in
  role: roles/compute.admin
- members:
  - user:m23csa510@iitj.ac.in
  role: roles/compute.instanceAdmin.v1
- members:
  - serviceAccount:service-526170978874@compute-system.iam.gserviceaccount.com
  role: roles/compute.serviceAgent
- members:
  - deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=105449095322312366247
  - deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=107647573142657691340
  - deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=116866398834826327838
  - deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=118410372475862490916
  - serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com
  role: roles/compute.viewer
- members:
  - serviceAccount:526170978874-compute@developer.gserviceaccount.com
  - serviceAccount:526170978874@cloudservices.gserviceaccount.com
  - user:m23csa510@iitj.ac.in
  role: roles/editor
- members:
  - user:m23csa510@iitj.ac.in
  role: roles/owner
etag: BwYvRg3obbw=
version: 1
=== Auto-Scaling Test Report ===
Starting instances: 1
Warning: Permanently added 'compute.2386051238814765794' (ED25519) to the list of known hosts.
stress-ng: info:  [1648] dispatching hogs: 2 cpu
Ending instances: 3
Firewall rules:
stress-ng: info:  [1648] successful run completed in 300.05s (5 mins, 0.05 secs)
NAME                  NETWORK  DIRECTION  PRIORITY  ALLOW    DENY  DISABLED
allow-http            default  INGRESS    1000      tcp:80         False
allow-ssh             default  INGRESS    1000      tcp:22         False
default-allow-http    default  INGRESS    1000      tcp:80         False
default-allow-https   default  INGRESS    1000      tcp:443        False
default-allow-ssh     default  INGRESS    65534     tcp:22         False

To show all fields of the firewall, please show in JSON format: --format=json
To show all fields in table format, please see the examples in --help.

IAM Policy:
ROLE                             MEMBERS
roles/compute.admin              user:m23csa510@iitj.ac.in
roles/compute.instanceAdmin.v1   user:m23csa510@iitj.ac.in
roles/compute.serviceAgent       serviceAccount:service-526170978874@compute-system.iam.gserviceaccount.com
roles/compute.viewer             deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=105449095322312366247
roles/compute.viewer             deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=107647573142657691340
roles/compute.viewer             deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=116866398834826327838
```

Figure 5: Resource cleanup log showing deletion of GCP resources

Description: This screenshot shows the terminal output during resource cleanup. All created resources including the autoscaler, managed instance group, instance template, firewall rules, and service account are deleted to avoid billing charges.

5.3 IAM Policy Verification



```
  - user:m23csa510@iitj.ac.in
  role: roles/editor
- members:
  - user:m23csa510@iitj.ac.in
  role: roles/owner
etag: BwYvRg3obbw=
version: 1
=== Auto-Scaling Test Report ===
Starting instances: 1
Warning: Permanently added 'compute.2386051238814765794' (ED25519) to the list of known hosts.
stress-ng: info:  [1648] dispatching hogs: 2 cpu
Ending instances: 3
Firewall rules:
stress-ng: info:  [1648] successful run completed in 300.05s (5 mins, 0.05 secs)
NAME                 NETWORK  DIRECTION  PRIORITY  ALLOW    DENY  DISABLED
allow-http           default  INGRESS    1000      tcp:80         False
allow-ssh            default  INGRESS    1000      tcp:22         False
default-allow-http   default  INGRESS    1000      tcp:80         False
default-allow-https  default  INGRESS    1000      tcp:443        False
default-allow-ssh    default  INGRESS    65534     tcp:22         False

To show all fields of the firewall, please show in JSON format: --format=json
To show all fields in table format, please see the examples in --help.

IAM Policy:
ROLE                            MEMBERS
roles/compute.admin             user:m23csa510@iitj.ac.in
roles/compute.instanceAdmin.v1  user:m23csa510@iitj.ac.in
roles/compute.serviceAgent      serviceAccount:service-526170978874@compute-system.iam.gserviceaccount.com
roles/compute.viewer            deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=105449095322312366247
roles/compute.viewer            deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=107647573142657691340
roles/compute.viewer            deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=116866398834826327838
roles/compute.viewer            deleted:serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com?uid=118410372475862490916
roles/compute.viewer            serviceAccount:compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com
roles/editor                    serviceAccount:526170978874-compute@developer.gserviceaccount.com
roles/editor                    serviceAccount:526170978874@cloudservices.gserviceaccount.com
roles/editor                    user:m23csa510@iitj.ac.in
roles/owner                     user:m23csa510@iitj.ac.in
Cleaning up...
Deleting autoscaler...Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/zones/asia-south1-c/autoscalers/test-mig-oehy].
Deleting autoscaler...done.
Deleting Managed Instance Group...Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/zones/asia-south1-c/instanceGroupManagers/test-mig].
Deleting Managed Instance Group...done.
Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/instanceTemplates/test-template].
Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/firewalls/allow-ssh].
Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/firewalls/allow-http].
Deleted [https://www.googleapis.com/compute/v1/projects/vcc-assignment2-452407/global/firewalls/deny-all].
deleted service account [compute-reader@vcc-assignment2-452407.iam.gserviceaccount.com]

 ~/mtech/vcc/gcp-autoscaling  on  main                                          ✓  took 12m 13s   3.10.10   at 04:53:53 PM
```

Figure 6: IAM policy verification log showing assigned roles and members

Description: This screenshot shows the output of the IAM policy verification command. It lists all the roles assigned to different members, including the compute admin, instance admin, service agent, and viewer roles for the service account.

6 Best Practices

- Use the default VPC network for simplicity.
- Keep zones consistent across resources.
- Delete resources after testing to avoid costs.

7 Conclusion

This implementation provides a scalable and secure VM setup on GCP. The auto-scaling policy ensures resource optimization, while IAM and firewall rules enhance security. Always validate your setup by triggering load and monitoring the environment.