GCP Virtual Machine Implementation with Auto-Scaling and Security

Assignment Report

Name: Suvodip Som

Roll Number: M23CSA533

Course: Virtualization and Cloud

Computing - CSL7510

Institution: Indian Institute of Technology,

Jodhpur

Table of Contents

- 1. Introduction
- 2. Step-by-Step Implementation Creating a VM Instance
 - ° Configuring Auto-Scaling Policies
 - Implementing Security Measures
 - IAM Role Configuration
 Firewall Rule Setup
- 3. Architecture Design
- 4. Deployment Scripts
- 5. Testing and Verification
- 6. Conclusion

Introduction

This document provides a comprehensive guide to setting up a virtual machine environment in Google Cloud Platform with auto-scaling capabilities and robust security measures.

The implementation follows cloud best practices to ensure scalability, reliability, and security of the deployment.

Step-by-Step Implementation Creating a VM Instance

1. Navigate to Google Cloud Console

° Go to https://console.cloud.google.com/) Select or create a project for your VM deployment

2. Access Compute Engine

- $_{\circ}^{\circ}$ From the navigation menu, select "Compute Engine" > "VM instances"
- Click "Create Instance"

3. Configure the VM

- . Name: web-server-vm-template
- ° Region and Zone: Select based on your latency requirements (e.g., us-central1-a)
- Machine Configuration:
 - Series: E2 (cost-effective)

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

4. Configure Boot Disk

- ° Operating System: Debian 11 (Bullseye)
- 。 Boot disk type: Standard persistent disk (pd-standard)

Size: 20 GB

5. Configure Identity and API Access

- ° Service account: Select a service account with minimal required permissions
 - Access scopes: Set to "Allow default access"

6. Configure Firewall

- ° Allow HTTP traffic: Yes (for web server access)
 - Allow HTTPS traffic: Yes (for secure web access)

7. Advanced Options

_o Management: Add startup script to install and configure web server software:

8. Create the VM

。Click "Create" to provision the VM

9. Create a VM Instance Template

- ° Navigate to "Compute Engine" > "Instance templates"
- 。 Click "Create instance template"
- ° Configure with the same settings as your VM Click "Create"

Configuring Auto-Scaling Policies

1. Create a Managed Instance Group (MIG)

- ° Navigate to "Compute Engine" > "Instance groups"
- 。 Click "Create Instance Group"
- ° Select "New managed instance group (stateless)"
- . Name: web-server-mig
- ° Instance template: Select your created template
- 。 Location: Regional (recommended for high availability)
- Region: Same as your VM template Minimum number of instances: 2

Maximum number of instances: 10

Maximum number of instances: 10

Click "Create"

2. Configure Auto-Scaling Policies

- From your MIG, click "Edit"
- ° Under "Autoscaling", select "On: add and remove instances to the group"

Autoscaling metrics:

- Select "CPU utilization"
- Target CPU utilization: 70%

Cool-down period: 60 seconds

Click "Done" and "Save"

3. Configure Health Check

Navigate to "Compute Engine" > "Health checks"

。 Click "Create health check"

° Name: http-health-check

。 Protocol: HTTP

° Port: 80

。Request path: /

Check interval: 10 seconds

Timeout: 5 secondsHealthy threshold: 2Unhealthy threshold: 3

Click "Create"

Return to your MIG and add this health check

4. Configure Initial Size

° Under your MIG configuration, set:

Initial size: 2 (This ensures high availability)

Implementing Security Measures

IAM Role Configuration

1. Create Custom IAM Roles

Navigate to "IAM & Admin" > "Roles"

- 。Click "Create Role"
- ° Title: VM-Operator
- Description: "Role for operating VMs with limited permissions"
- $^{\circ}$ ID: vm_operator

Add the following permissions:

- compute.instances.get
- compute.instances.list

compute.instances.start
compute.instances.stop
compute.instances.reset

Click "Create"

2. Create a Service Account

- ° Navigate to "IAM & Admin" > "Service Accounts"
- 。 Click "Create Service Account"
- ° Name: vm-operations-sa
- Description: "Service account for VM operations"
- ° Click "Create and Continue"

Assign your custom role: VM-Operator

Click "Done"

3. Assign Service Account to VMs

- ° Navigate to your VM instance template
- 。 Click "Edit"
- ° Under "Service account", select your newly created service account Save the changes

4. Apply Principle of Least Privilege

- ° Review all IAM roles in your project
- 。 Remove unnecessary permissions

Ensure each account has only the permissions required for its function

Firewall Rule Setup

1. Create Firewall Rules

- ° Navigate to "VPC Network" > "Firewall"
 - Click "Create Firewall Rule"

2. Configure Ingress Rules

Name: allow-http-https

Description: "Allow HTTP and HTTPS traffic"

Network: default

° Priority: 1000

Direction of traffic: Ingress

° Action on match: Allow

Targets: Specified target tags

° Target tags: web-server

Source filter: IP ranges

° Source IP ranges: 0.0.0.0/0 (consider restricting to specific IP ranges for production)

Protocols and ports:

Select "Specified protocols and ports"

Check "TCP" and specify ports: 80,443

Click "Create"

3. Configure SSH Access Rule

Name: allow-ssh-restricted

Description: "Allow SSH from company IP range only"

Network: default

° Priority: 900 (higher priority than default SSH rule)

Direction of traffic: Ingress

° Action on match: Allow

Targets: Specified target tags

° Target tags: web-server

。Source filter: IP ranges

° Source IP ranges: YOUR_COMPANY_IP_RANGE (e.g.,

192.168.1.0/24)

0

Protocols and ports:

Select "Specified protocols and ports"

Check "TCP" and specify port: 22

Click "Create"

4. Deny All Other Traffic (Optional but Recommended)

° Name: deny-all-other

Description: "Deny all other incoming traffic"

° Network: default

Priority: 2000 (lower priority)

° Direction of traffic: Ingress

Action on match: Deny

° Targets: All instances in the network

Source filter: IP ranges

Source IP ranges: 0.0.0.0/0

Protocols and ports: All

Click "Create"

5. Apply Firewall Tags to VMs

- ° Navigate to your VM instance template
- 。 Click "Edit"
- ° Under "Network tags", add web-server

Save the changes

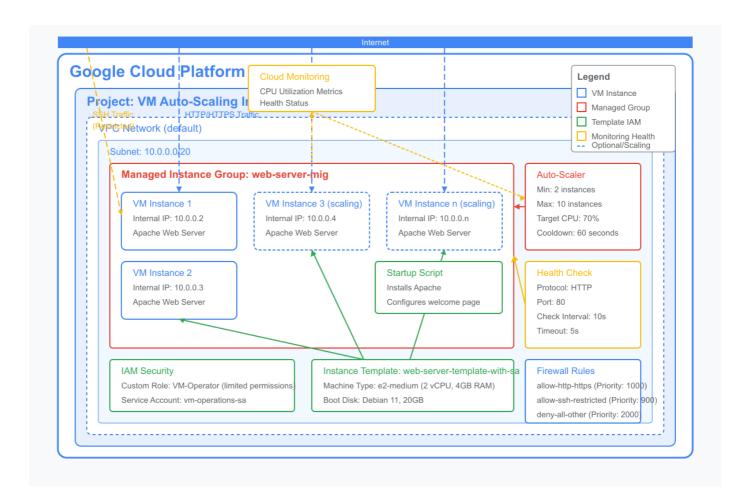
Architecture Design
The architecture consists of:

- Instance Template with Debian 11 and Apache web server
- . Managed Instance Group (MIG) with autoscaling based on CPU utilization
- HTTP/HTTPS Load Balancer (optional, for distributed traffic)
- . Custom IAM roles and service accounts with least privilege

Firewall rules allowing only necessary traffic

Health checks for instance monitoring

The detailed architecture diagram is shown below:



Testing and Verification

To test the implementation, follow these steps:

1. Verify VM Creation

- ° Check the Compute Engine dashboard to confirm VMs are running
- Verify they are part of the MIG and have the correct template applied

2. Test Auto-Scaling

Generate CPU load to trigger auto-scaling:

```
# SSH into one of the
instances gcloud compute ssh
<instance-name> --
zone=<zone>

# Generate load
sudo apt-get
update sudo apt-
get install -y
stress stress --
cpu 2 --timeout
300
```

_o Monitor the MIG dashboard to observe new instances being created

3. Verify Security Configuration

。Check that IAM roles are correctly applied

```
gcloud iam service-accounts get-iam-policy vm-operations-sa@sa@project-id>.iam.gserviceaccount.com
```

- ° Test firewall rules by attempting to access:
 - HTTP/HTTPS: Should succeed from any IP
 - SSH: Should only succeed from company IP range

Other ports: Should be blocked

4. Load Testing

- ° Use tools like Apache Benchmark or JMeter to generate HTTP load
 - Observe auto-scaling behavior under different load patterns

Conclusion

This implementation provides a scalable, secure virtual machine environment in Google Cloud Platform. The auto-scaling policies ensure resource efficiency by automatically

adjusting the number of instances based on workload, while the security measures protect the environment from unauthorized access.

The key benefits of this implementation include:

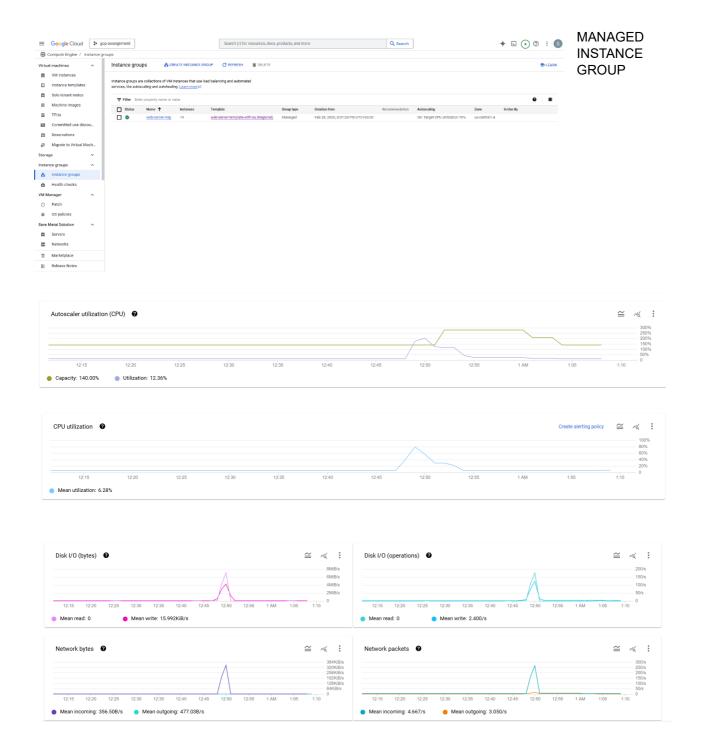
- 1. Scalability: Automatic scaling from 2 to 10 instances based on CPU utilization.
- 2. **Cost Efficiency**: Resources are allocated only when needed and scaled down when demand decreases.

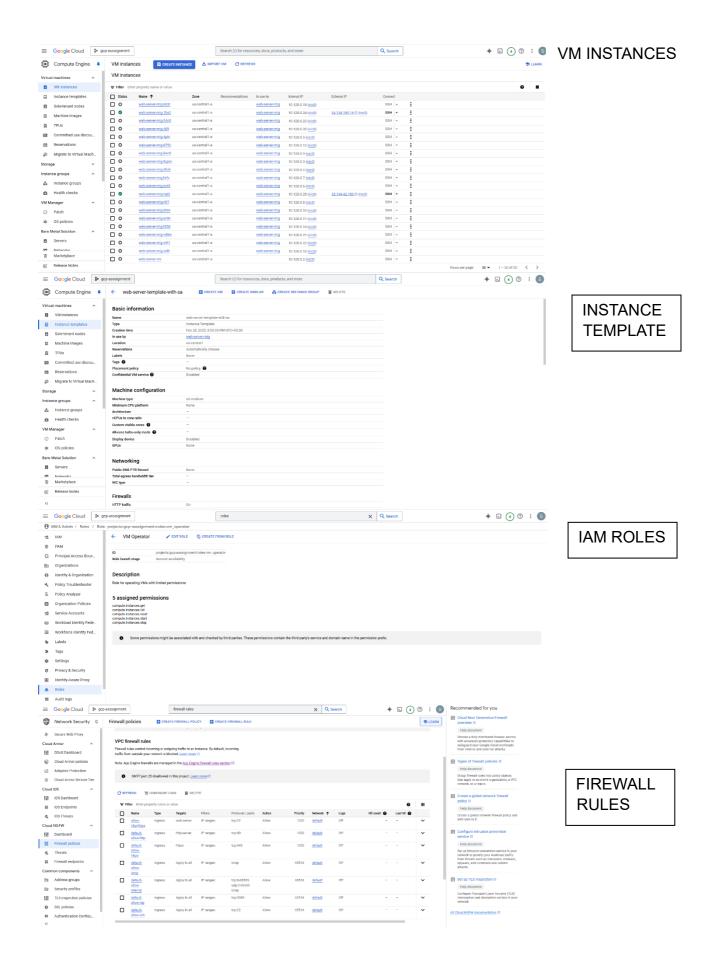
- 3. High Availability: Multiple instances across different zones ensure service continuity.
- 4. **Security**: Robust security measures including custom IAM roles and firewall rules.
- 5. Monitoring: Health checks to ensure all instances are functioning properly.

For production environments, consider:

- Implementing additional security measures like VPC Service Controls
- Setting up monitoring and alerting using Cloud Monitoring
- Implementing disaster recovery procedures
- Regularly reviewing and updating security policies

SOME SCREENSHOTS







Repository Information

The complete source code for this implementation is available in the GitHub repository:

https://github.com/suvosom1997/VCC_GCP_assignment

(https://github.com/suvosom1997/VCC GCP assignm ent)

Video Demonstration

A video demonstration of this implementation is available at:

https://youtu.be/gQxDO1bvn9U (https://youtu.be/gQxDO1bvn9U)

Plagiarism Statement

I, Suvodip Som (M23CSA533), declare that this assignment is my own work and has not been copied from any other source without proper acknowledgment. I understand that

plagiarism is a serious academic offense that can result in voiding my submission and other disciplinary actions from the institute.

Date: March 01, 2025 Signature: Suvodip Som