Assignment 3: Auto-Scaling a Local VM to GCP Based on Resource Usage

1. Introduction

This report details the implementation of a system that monitors a local Virtual Machine (VM) for high resource usage and auto-scales to a public cloud (Google Cloud Platform - GCP) when CPU or memory exceeds 75% utilization.

2. System Overview

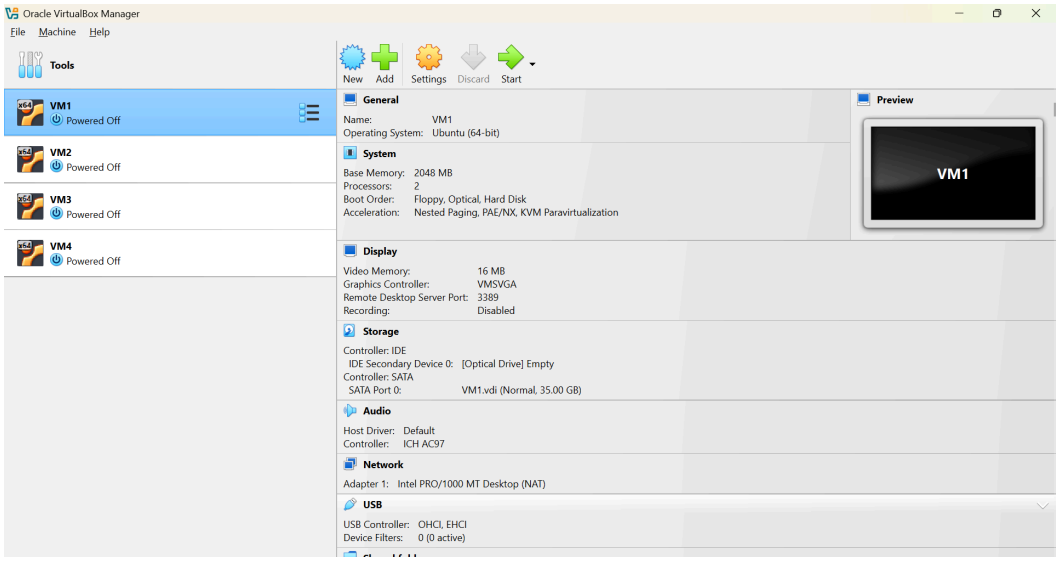
The system consists of:

- A local VM running Ubuntu.
- A resource monitoring script using `psutil`.
- A Google Cloud Function that triggers auto-scaling.
- An instance group on GCP configured for auto-scaling.

3. Step-by-Step Implementation

3.1 Creating a Local VM

1. Install VirtualBox.
2. Set up an Ubuntu 22.04 VM with 2 CPU cores, 2GB RAM.



Install required packages:

```
sudo apt update && sudo apt install -y python3 stress stress-ng htop
python3-pip docker.io
sudo apt update && sudo apt install -y virtualbox-guest-utils
```

3.2 Implementing Resource Monitoring

Install `psutil`:

```
sudo apt update
sudo apt install -y python3-psutil python3-requests
```

Create `monitor_stress.py`:

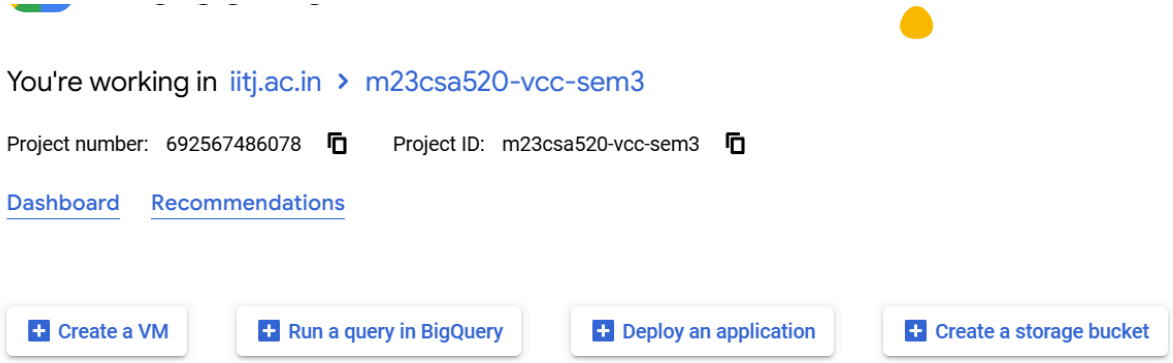
(placed the file in the shared folder)

```
total 52
-rw-rw-r--  1 vboxvm1 vboxvm1  1205 Feb 10 21:30 a
drwxrwxr-x 83 vboxvm1 vboxvm1  4096 Feb 11 11:50 node_modules
-rw-rw-r--  1 vboxvm1 vboxvm1 32979 Feb 11 11:50 package-lock.json
-rw-rw-r--  1 vboxvm1 vboxvm1   318 Feb 11 11:50 package.json
drwxrwxr-x  2 vboxvm1 vboxvm1  4096 Feb 11 12:03 microservices
vboxvm1@vm1:~$ ls /media/
sf_VCC
vboxvm1@vm1:~$ cp /media/sf_VCC/monitor_stress.py ~/
vboxvm1@vm1:~$ ls -lrt
total 56
-rw-rw-r--  1 vboxvm1 vboxvm1  1205 Feb 10 21:30 a
drwxrwxr-x 83 vboxvm1 vboxvm1  4096 Feb 11 11:50 node_modules
-rw-rw-r--  1 vboxvm1 vboxvm1 32979 Feb 11 11:50 package-lock.json
-rw-rw-r--  1 vboxvm1 vboxvm1   318 Feb 11 11:50 package.json
drwxrwxr-x  2 vboxvm1 vboxvm1  4096 Feb 11 12:03 microservices
-rwxrwx---  1 vboxvm1 vboxvm1  1767 Mar 23 13:01 monitor_stress.py
vboxvm1@vm1:~$ chmod 755 monitor*
vboxvm1@vm1:~$ ls -lrt
total 56
-rw-rw-r--  1 vboxvm1 vboxvm1  1205 Feb 10 21:30 a
drwxrwxr-x 83 vboxvm1 vboxvm1  4096 Feb 11 11:50 node_modules
-rw-rw-r--  1 vboxvm1 vboxvm1 32979 Feb 11 11:50 package-lock.json
-rw-rw-r--  1 vboxvm1 vboxvm1   318 Feb 11 11:50 package.json
drwxrwxr-x  2 vboxvm1 vboxvm1  4096 Feb 11 12:03 microservices
-rwxr-xr-x  1 vboxvm1 vboxvm1  1767 Mar 23 13:01 monitor_stress.py
vboxvm1@vm1:~$ _
```

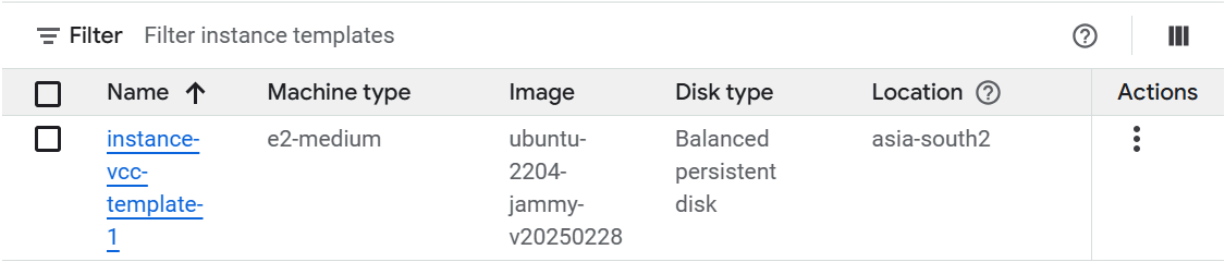
```python
import psutil
import time
import subprocess
import requests
import shutil

# Constants
THRESHOLD = 75  # CPU usage percentage
STRESS_DURATION = 45  #Duration for stress test
STRESS_CORES = 4  # Number of CPU cores to stress
GCP_TRIGGER_URL = "https://scale-up-instance-xk4lfjdp7a-em.a.run.app"

# GCP Instance Group Details
PROJECT = "m23csa520-vcc-sem3"  # Update with your actual project ID
ZONE = "asia-south2-a"  # Update with your actual zone
INSTANCE_GROUP = "vcc-auto-scale-group"  # Update with your actual instance group
NEW_SIZE = 2  # Target size after scaling up

def start_stress():
    """Start a CPU stress test using 'stress-ng' and wait for CPU to rise."""
    print("Low CPU detected! Running stress test...")

    if not shutil.which("stress-ng"):
        print("Error: 'stress-ng' not found. Installing it...")
        subprocess.run(["sudo", "apt", "update"], check=True)
        subprocess.run(["sudo", "apt", "install", "-y", "stress-ng"], check=True)

    # Start stress-ng in background
    process = subprocess.Popen([
        "stress-ng", "--cpu", str(STRESS_CORES),
        "--cpu-method", "matrixprod", "--cpu-load", "80",
        "--timeout", str(STRESS_DURATION)
    ], stdout=subprocess.PIPE, stderr=subprocess.PIPE)

    print("Waiting for CPU load to increase...")

    # **Wait until CPU load crosses THRESHOLD before returning**
    while True:
        time.sleep(5)  # Give some time for stress to take effect
        cpu_usage = psutil.cpu_percent(interval=1)
        print(f"Current CPU usage: {cpu_usage}%")

        if cpu_usage >= THRESHOLD:
            print("CPU load has increased successfully! Triggering cloud deployment...")
            trigger_gcp_scaling()
            break  # Stop waiting and return
```

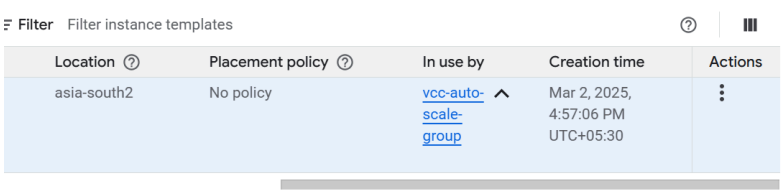
```
def trigger_gcp_scaling():
    """Trigger GCP function to scale up instances."""
    payload = {
        "project": PROJECT,
        "zone": ZONE,
        "instance_group": INSTANCE_GROUP,
        "size": NEW_SIZE
    }

    headers = {"Content-Type": "application/json"}

    try:
        response = requests.post(GCP_TRIGGER_URL, json=payload, headers=headers
        print(f"Cloud response: {response.status_code} - {response.text}")

        if response.status_code != 200:
            print("Check GCP function logs for errors!")

    except requests.RequestException as e:
        print(f"Failed to reach cloud function: {e}")

def check_resources():
    """Monitor CPU usage and trigger stress test if needed."""
    while True:
        cpu_usage = psutil.cpu_percent(interval=5)
        memory_usage = psutil.virtual_memory().percent
        print(f"CPU: {cpu_usage}%, Memory: {memory_usage}%")

        if cpu_usage < THRESHOLD:  # Start stress if CPU is low
            start_stress()

        time.sleep(2)  # Check every 2 seconds

if __name__ == "__main__":
    check_resources()
```

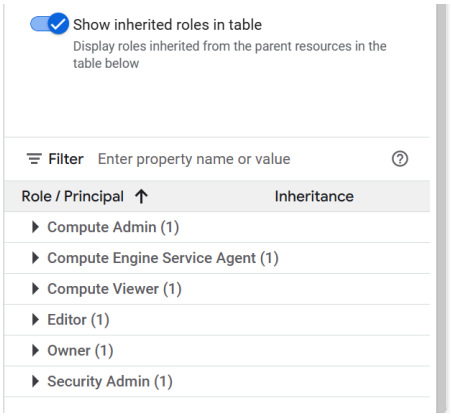
3.3 Configuring Auto-Scaling on GCP



Already have an instance template created for Assignment 2

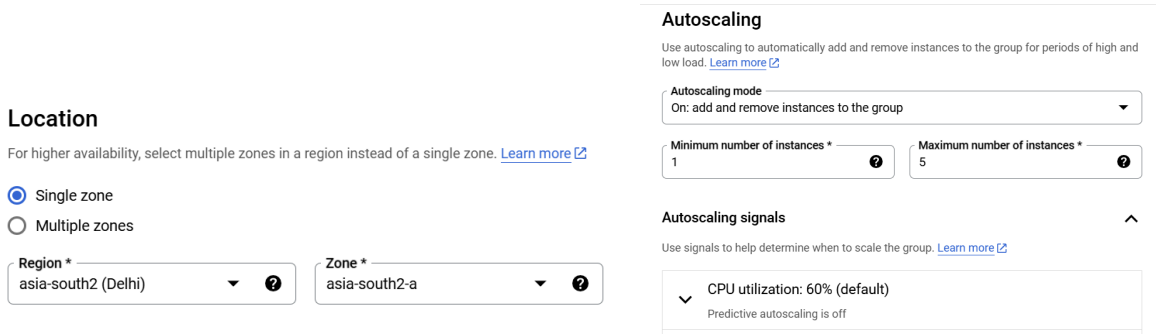




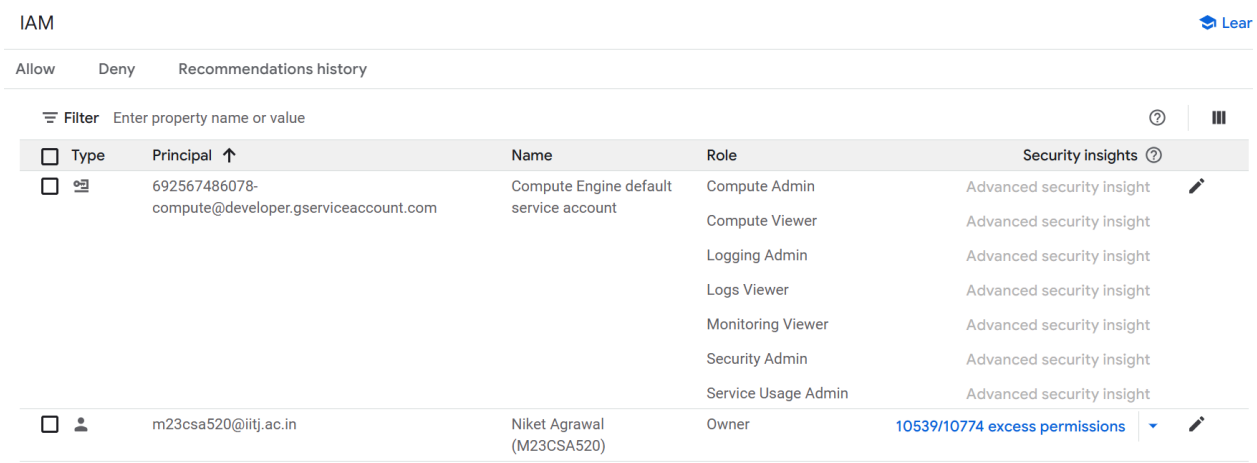


Autoscaling Policy Configuration

- **Minimum instances**: 1
- **Maximum instances**: 5
- **Scaling metric**: CPU utilization
- **Target CPU utilization**: 60%

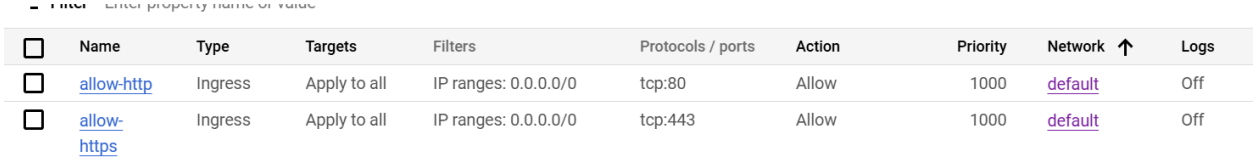


IAM ADMIN Rules



Configure Firewall Rules

1. Navigate to **VPC Network → Firewall**.
2. Click **Create Firewall Rule**.
3. Set up the rule:
    - Name: `allow-http`
    - Network: Default
    - Direction: Ingress
    - Action: Allow
    - Targets: All instances
    - Protocols & Ports: Select `TCP` and enter `80`



Create a Google Cloud Function to trigger instance creation:

main.py

```python
import functions_framework
import google.auth
from googleapiclient.discovery import build

@functions_framework.http
def scale_up(request):
    """Cloud Function to scale up the instance group."""
    credentials, project = google.auth.default()
    service = build('compute', 'v1', credentials=credentials)

    request_json = request.get_json()
    instance_group = request_json.get('instance_group')
    zone = request_json.get('zone')

    if not instance_group or not zone:
        return "Missing instance_group or zone in request", 400

    try:
        request = service.instanceGroupManagers().resize(
            project=project,
            zone=zone,
            instanceGroupManager=instance_group,
            size=2  # Increase instances
        )
        request.execute()
        return "Scaling request sent.", 200
    except Exception as e:
        return f"Error: {str(e)}", 500
```

```
Version: 1
m23csa520@vcc-auto-scale-group-p2wb:~$ gcloud functions deploy scale_up_instance \
  --runtime python310 \
  --trigger-http \
  --allow-unauthenticated \
  --entry-point scale_up \
  --region asia-south2 \
  --source=$(pwd)
```

```
Preparing function...done.
⠹ Updating function (may take a while)...
  ⠹ [Build] Build in progress... Logs are available at [https://console.cloud.google.com/cloud-build/builds;reg
  ion=asia-south2/0d3b6c1a-02c4-41e8-8435-62f1a129bf19?project=692567486078]
  . [Service]
  . [ArtifactRegistry]
  . [Healthcheck]
  . [Triggercheck]
```

```
m23csa520@vcc-auto-scale-group-p2wb:~$ gcloud functions logs read scale_up_instance --region=asia-south2 --limit=50
LEVEL    NAME              EXECUTION_ID  TIME_UTC                 LOG
I        scale-up-instance               2025-03-23 14:14:09.614
I        scale-up-instance               2025-03-23 14:08:32.903
I        scale-up-instance               2025-03-23 14:05:37.982
E        scale-up-instance               2025-03-23 14:03:41.903
WARNING  scale-up-instance               2025-03-23 14:03:14.375
WARNING  scale-up-instance               2025-03-23 14:02:13.242
I        scale-up-instance               2025-03-23 14:00:55.246  Default STARTUP TCP probe succeeded after 1 attempt for container "worker" on port 8080.
m23csa520@vcc-auto-scale-group-p2wb:~$
```

Get the URL for the function deployed

```
url: https://asia-south2-m23csa520-vcc-sem3.cloudfunctions.net/scale_up_instance
m23csa520@vcc-auto-scale-group-p2wb:~$ gcloud functions describe scale_up_instance --region=asia-south2 --format="value(serviceConfig.uri)"
https://scale-up-instance-xk4lfjdp7a-em.a.run.app
m23csa520@vcc-auto-scale-group-p2wb:~$
```

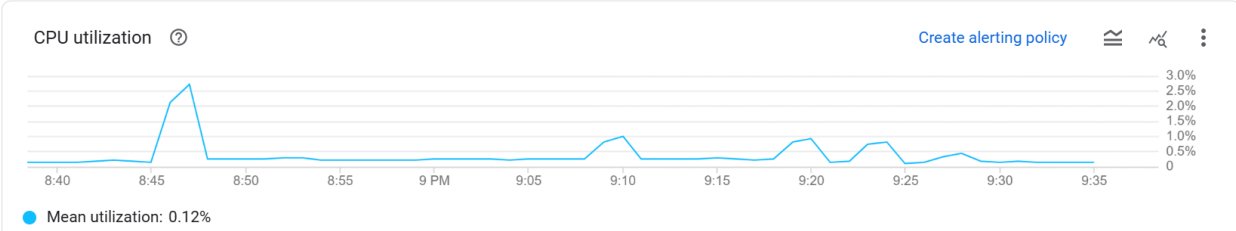
3.4 Run the stress test & monitor on Local VM and check instances on GCP.

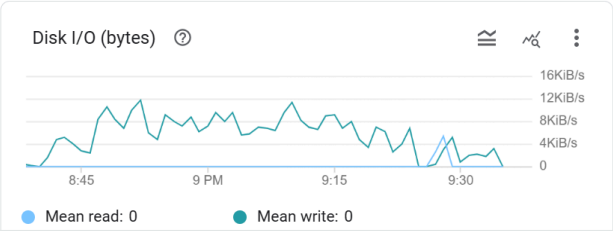
```
vboxvm1@vm1:~$ cp /media/sf_VCC/monitor_stress.py ~/
vboxvm1@vm1:~$
vboxvm1@vm1:~$ python3 monitor_stress.py
CPU: 0.7%, Memory: 22.9%
Low CPU detected! Running stress test...
Waiting for CPU load to increase...
Current CPU usage: 100.0%
CPU load has increased successfully! Triggering cloud deployment...
Cloud response: 200 - Scaling request sent.
CPU: 96.2%, Memory: 23.8%
CPU: 53.0%, Memory: 23.8%
Low CPU detected! Running stress test...
Waiting for CPU load to increase...
Current CPU usage: 100.0%
CPU load has increased successfully! Triggering cloud deployment...
Cloud response: 200 - Scaling request sent.
CPU: 100.0%, Memory: 25.0%
CPU: 85.7%, Memory: 23.8%
^[      CPU: 96.2%, Memory: 23.8%
```

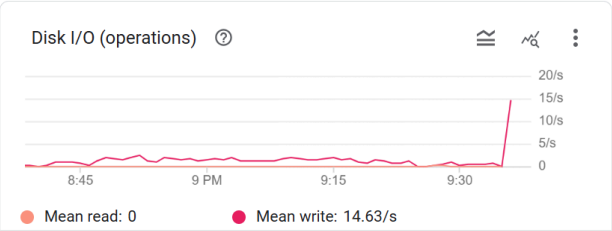
```
m23csa520@vcc-auto-scale-group-p2wb:~$ gcloud functions logs read scale_up_instance --region=asia-south2 --limit=50
LEVEL    NAME              EXECUTION_ID  TIME_UTC                 LOG
I        scale-up-instance               2025-03-23 16:05:34.117
I        scale-up-instance               2025-03-23 16:05:08.897
WARNING  scale-up-instance               2025-03-23 15:56:51.660
WARNING  scale-up-instance               2025-03-23 15:48:33.684
I        scale-up-instance               2025-03-23 15:48:19.118  Default STARTUP TCP probe succeeded after 1 attempt for container "worker" on port 8080.
WARNING  scale-up-instance               2025-03-23 15:48:18.307
I        scale-up-instance               2025-03-23 14:14:09.614
I        scale-up-instance               2025-03-23 14:08:32.903
I        scale-up-instance               2025-03-23 14:05:37.982
E        scale-up-instance               2025-03-23 14:03:41.903
WARNING  scale-up-instance               2025-03-23 14:03:14.375
WARNING  scale-up-instance               2025-03-23 14:02:13.242
I        scale-up-instance               2025-03-23 14:00:55.246  Default STARTUP TCP probe succeeded after 1 attempt for container "worker" on port 8080.
m23csa520@vcc-auto-scale-group-p2wb:~$ date
Sun Mar 23 16:05:49 UTC 2025
m23csa520@vcc-auto-scale-group-p2wb:~$
```

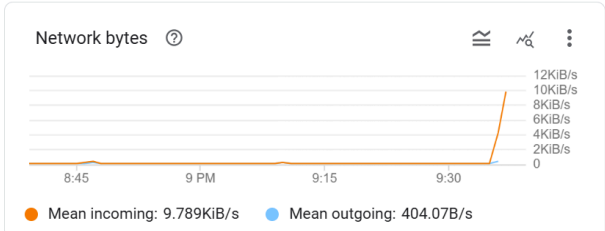
```
m23csa520@vcc-auto-scale-group-p2wb:~$ gcloud compute instance-groups managed list-instances vcc-auto-scale-group     --zone=asia-south2-a     --project=m23csa52
0-vcc-sem3
NAME                        ZONE           STATUS      HEALTH_STATE  ACTION    INSTANCE_TEMPLATE        VERSION_NAME  LAST_ERROR
vcc-auto-scale-group-1g7z   asia-south2-a  STOPPING                  DELETING
vcc-auto-scale-group-df2r   asia-south2-a  STOPPING                  DELETING
vcc-auto-scale-group-fqdw   asia-south2-a  TERMINATED                NONE      instance-vcc-template-1
vcc-auto-scale-group-mnpw   asia-south2-a  STOPPING                  DELETING
vcc-auto-scale-group-p0s2   asia-south2-a                            CREATING  instance-vcc-template-1
vcc-auto-scale-group-p2wb   asia-south2-a  RUNNING                   NONE      instance-vcc-template-1
m23csa520@vcc-auto-scale-group-p2wb:~$
```

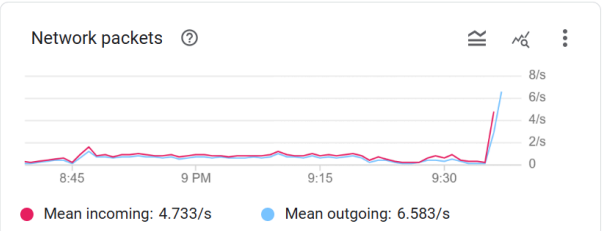
vcc-auto-scale-group-1g7z	Mar 23, 2025, 9:36:25 PM UTC+05:30	-	10.190.0.26 (nic0)	34.131.91.188	SSH
vcc-auto-scale-group-df2r	Mar 23, 2025, 9:35:40 PM UTC+05:30	-	10.190.0.25 (nic0)	34.131.178.220	SSH
vcc-auto-scale-group-fqdw	Mar 2, 2025, 9:38:43 PM UTC+05:30	instance-vcc-template-1 (Regional)	10.190.0.14 (nic0)	None	SSH
vcc-auto-scale-group-mnpw	Mar 23, 2025, 9:35:16 PM UTC+05:30	-	10.190.0.24 (nic0)	34.131.43.78	SSH
vcc-auto-scale-	Mar 23, 2025, 9:36:49 PM	instance-vcc-template-1	10.190.0.27 (nic0)	34.131.173.69	SSH





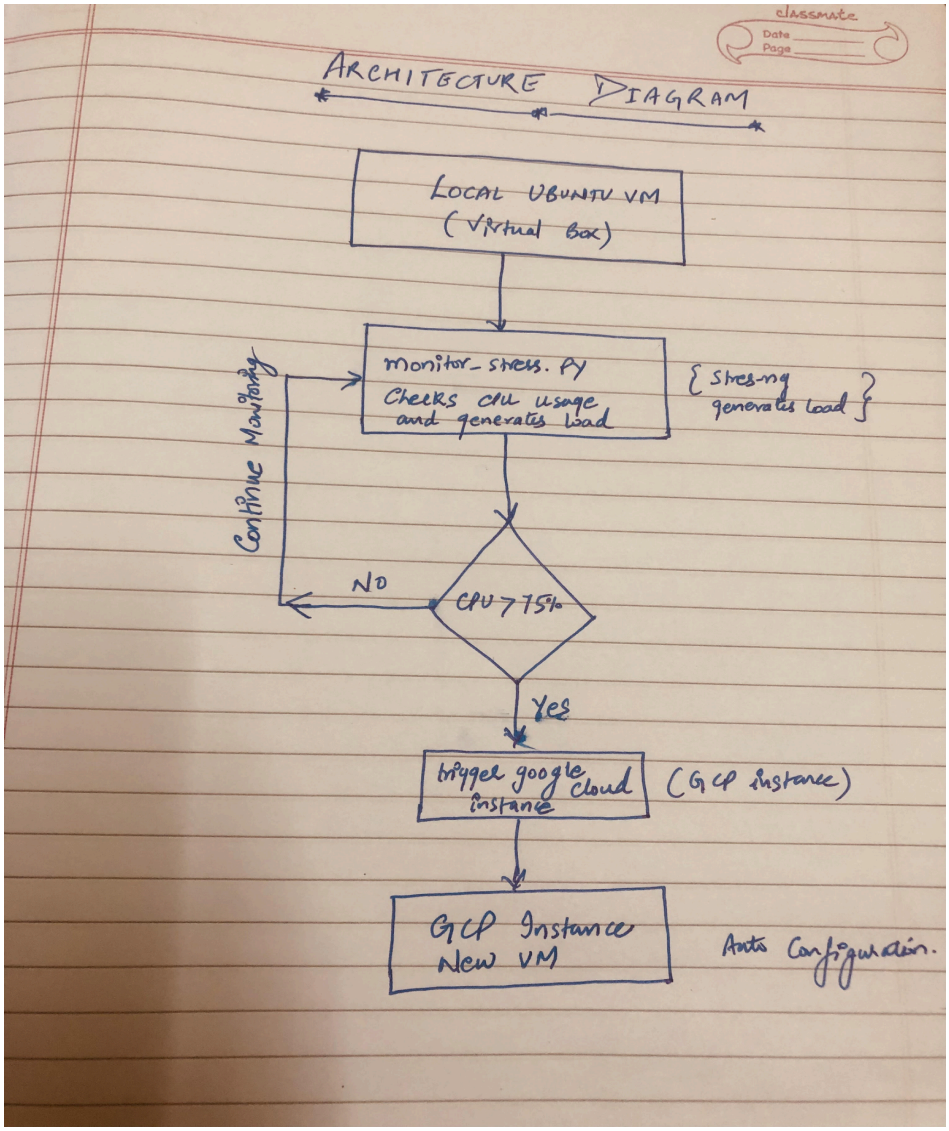






2025-03-23 21:35:10.656 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…
2025-03-23 21:35:11.211 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…
2025-03-23 21:35:35.382 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…
2025-03-23 21:35:35.864 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…
2025-03-23 21:36:19.995 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…
2025-03-23 21:36:20.473 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…
2025-03-23 21:36:44.208 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…
2025-03-23 21:36:44.674 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…
2025-03-23 21:37:09.580 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…
2025-03-23 21:37:10.041 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…
2025-03-23 21:37:34.369 IST	Compute Engine	resize	asia-south2-a:vcc-auto-scale-group	692567486078-compute@developer.gservi…	{@type: type.googleap…

4. Architecture Diagram



5. Source Code Repository

GitHub repository link containing:

- `monitor_stress.py`
- Cloud Function code (gcp_function)
- GCP deployment scripts (main.py, requirements.txt , & miscellaneous commands)

  https://github.com/M23CSA520/VCC_PA3/tree/main

6. Video Demo Link :

https://drive.google.com/file/d/1eX5iwjKKJk07JOMFvA4FtYw4yWpgf4r-/view?usp=sharing