CS188 Discussion W3

Mingyu Derek Ma

Email: ma@cs.ucla.edu

Reminder

- HW1 released, due at Jan 31 11:59pm
- Project midterm report due Feb 2nd
- For setting up remote machine: start early (there are waiting and manual screening time needed)

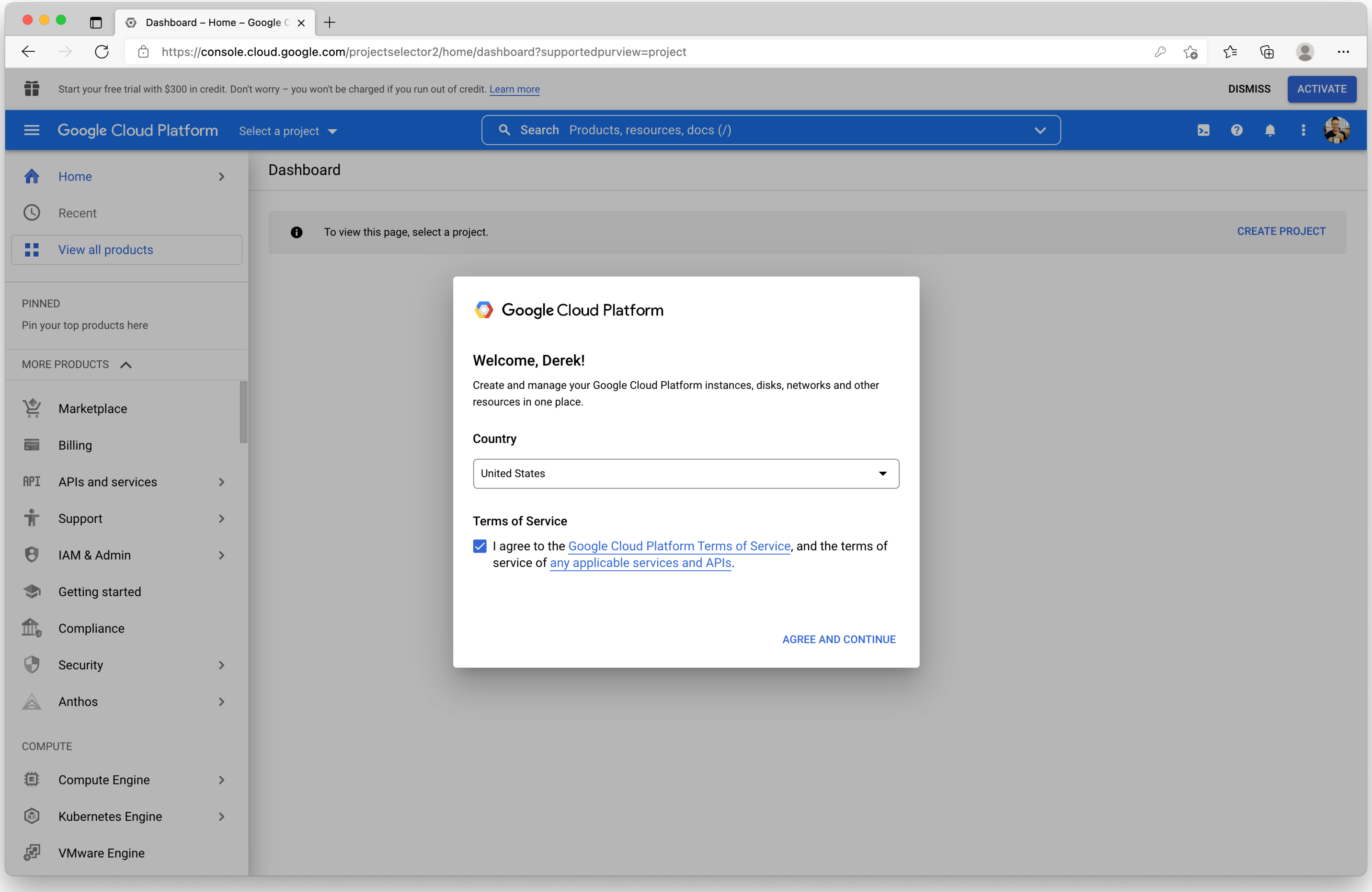
Today: running experiments on cloud

- Set up virtual machines with GPUs on Google Cloud
- Tips for running experiments on a Linux machine

Create Instance with GPU on GCP

1. Create project

- Click "Create Project"
- Or “New Project” after clicking the project name next to the “Google Cloud Platform” title


https://console.cloud.google.com/projectselector2/home/dashboard?supportedpurview=project

Start your free trial with $300 in credit. Don't worry – you won't be charged if you run out of credit. Learn more

To view this page, select a project.

Google Cloud Platform

Welcome, Derek!

Create and manage your Google Cloud Platform instances, disks, networks and other resources in one place.

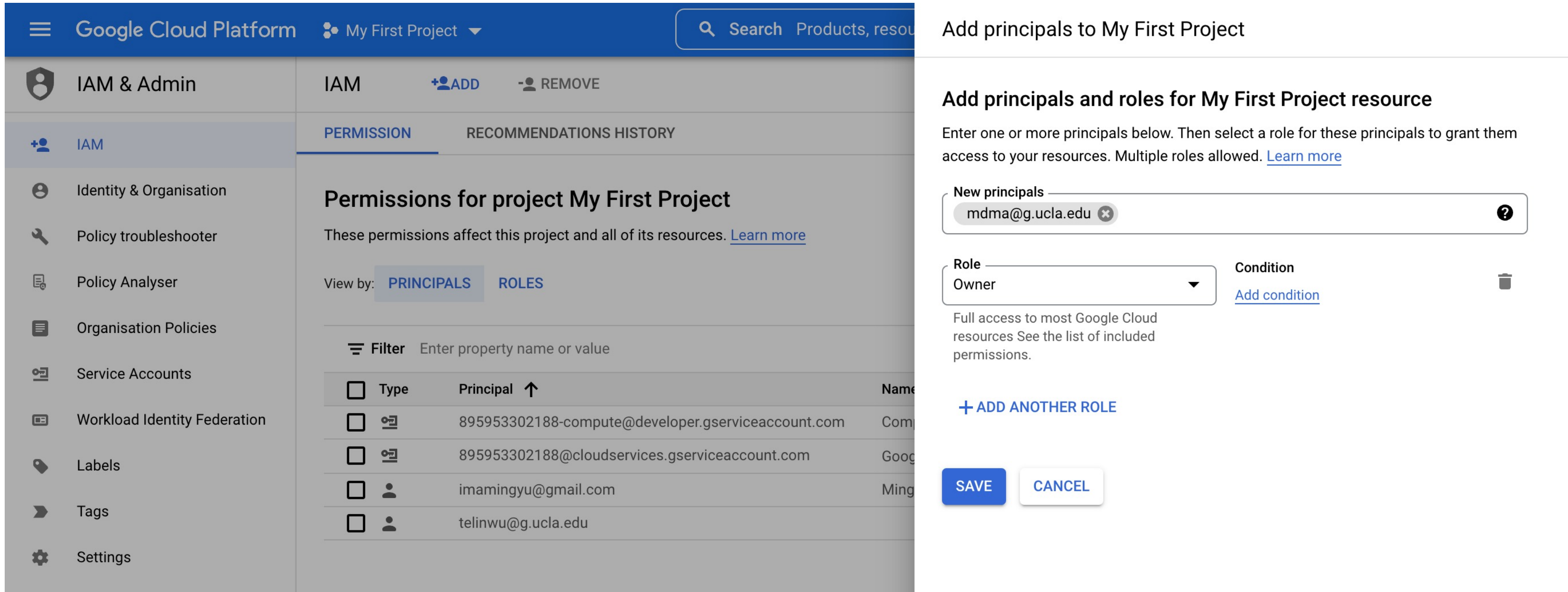
Country: United States

Terms of Service

☑ I agree to the Google Cloud Platform Terms of Service, and the terms of service of any applicable services and APIs.

Share project with teammates

- *“IAM & Admin” > “IAM”*
- Add new user to the project, so other teammates can access the instances under this project



Permissions for project My First Project

These permissions affect this project and all of its resources. Learn more

Type	Principal	Name
	895953302188-compute@developer.gserviceaccount.com	Com
	895953302188@cloudservices.gserviceaccount.com	Goog
	imamingyu@gmail.com	Ming
	telinwu@g.ucla.edu	

Add principals to My First Project

Add principals and roles for My First Project resource

Enter one or more principals below. Then select a role for these principals to grant them access to your resources. Multiple roles allowed. Learn more

New principals: mdma@g.ucla.edu

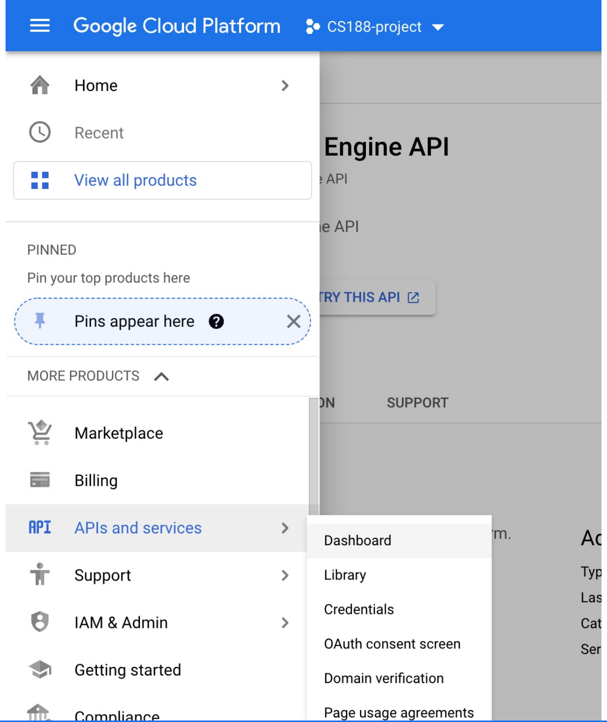
Role: Owner

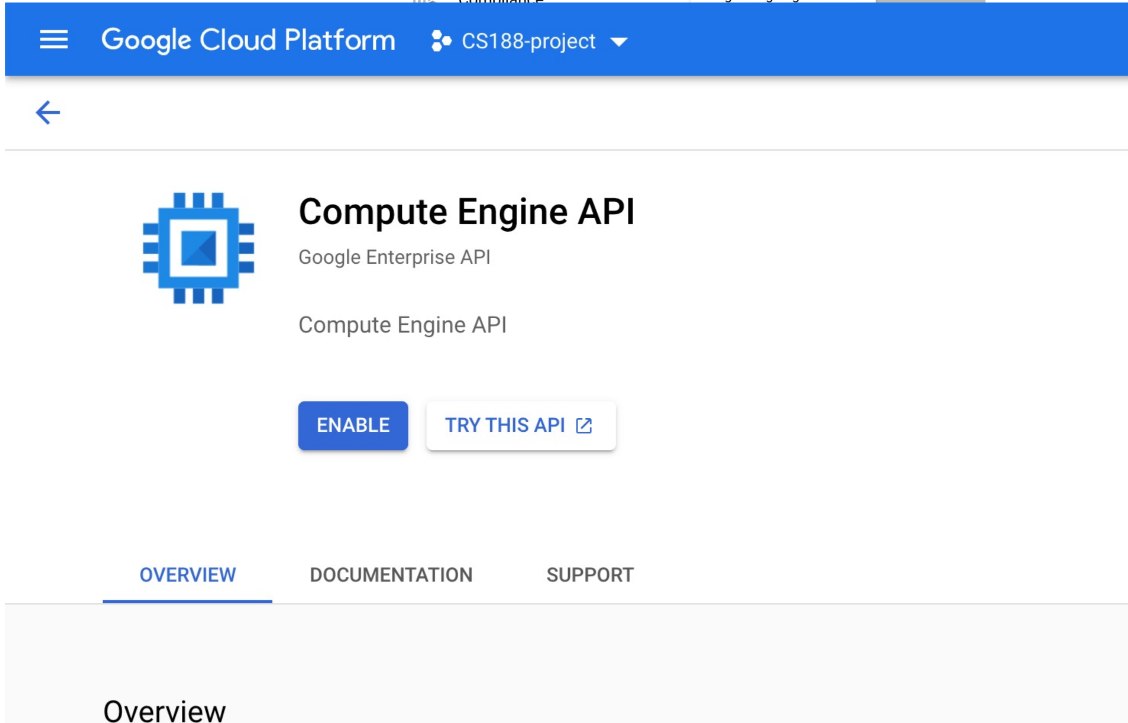
Full access to most Google Cloud resources See the list of included permissions.

Condition: Add condition

2. Enable Compute Engine API

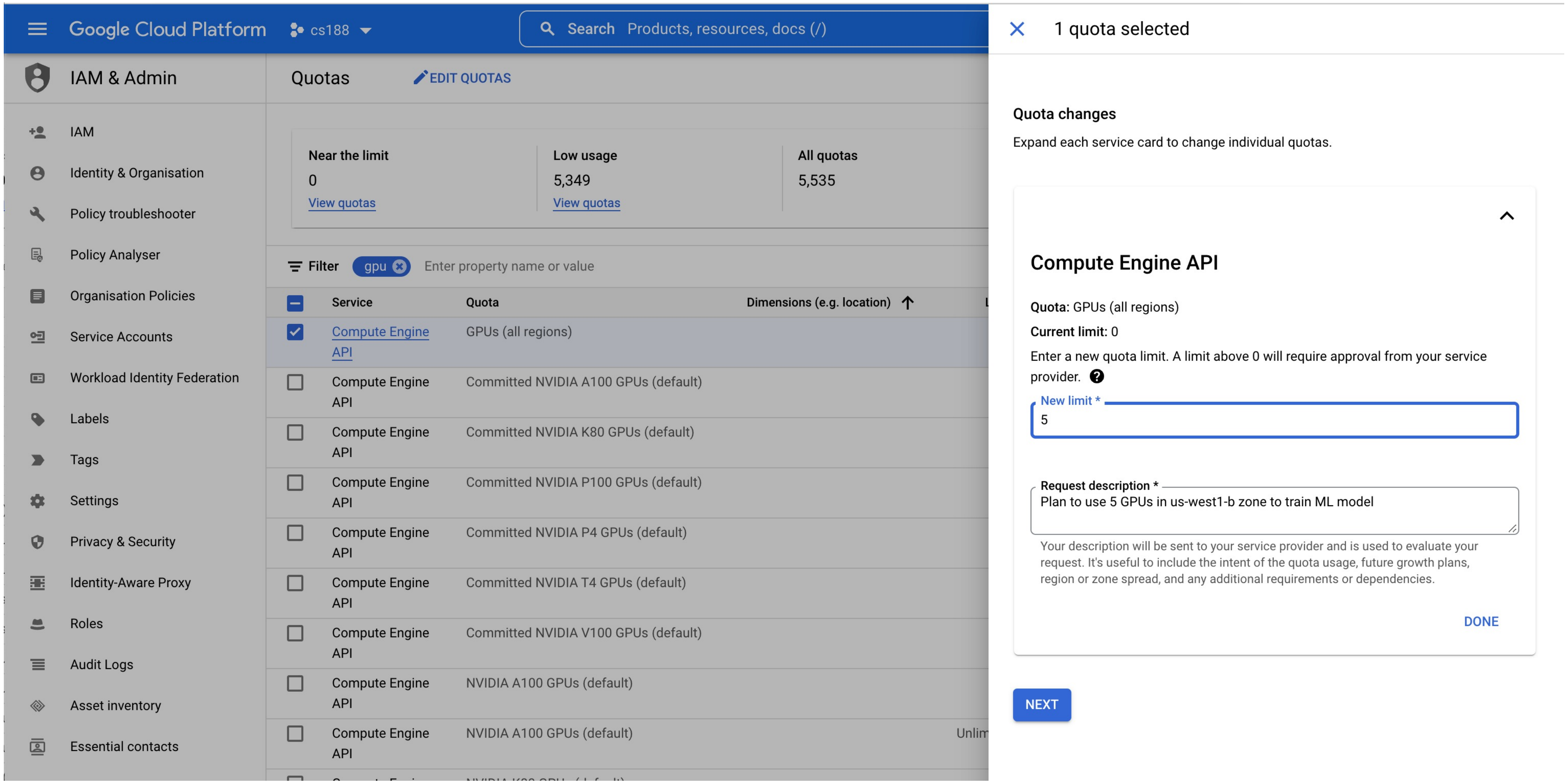
- It will prompt you to enable to the API when you first open the interfaces for Compute Engine
- Otherwise you can enable the API at *“API and services” > “Dashboard” > Search “Compute Engine API” > Enable*





3. Check/change GPU quota after 48 hours

- By default, we can use 0 GPUs
- We need to request an increase in GPU quota
- Resource quotas | Compute Engine Documentation | Google Cloud
- Check quota at “IAM & Admin” > “Quotas”
- Add “gpu” in the filter
- Select quota item, click “Edit Quotas”
- Submit quota change request, need 24-48 hours to get response
  - Submit the quota increase request after 48 hours of creating your project, otherwise it will be declined



Near the limit	Low usage	All quotas
0	5,349	5,535
View quotas	View quotas	

☐	Service	Quota	Dimensions (e.g. location) ↑	L
☑	Compute Engine API	GPUs (all regions)		
☐	Compute Engine API	Committed NVIDIA A100 GPUs (default)		
☐	Compute Engine API	Committed NVIDIA K80 GPUs (default)		
☐	Compute Engine API	Committed NVIDIA P100 GPUs (default)		
☐	Compute Engine API	Committed NVIDIA P4 GPUs (default)		
☐	Compute Engine API	Committed NVIDIA T4 GPUs (default)		
☐	Compute Engine API	Committed NVIDIA V100 GPUs (default)		
☐	Compute Engine API	NVIDIA A100 GPUs (default)		
☐	Compute Engine API	NVIDIA A100 GPUs (default)		Unlim

Quota changes

Expand each service card to change individual quotas.

Compute Engine API

Quota: GPUs (all regions)

Current limit: 0

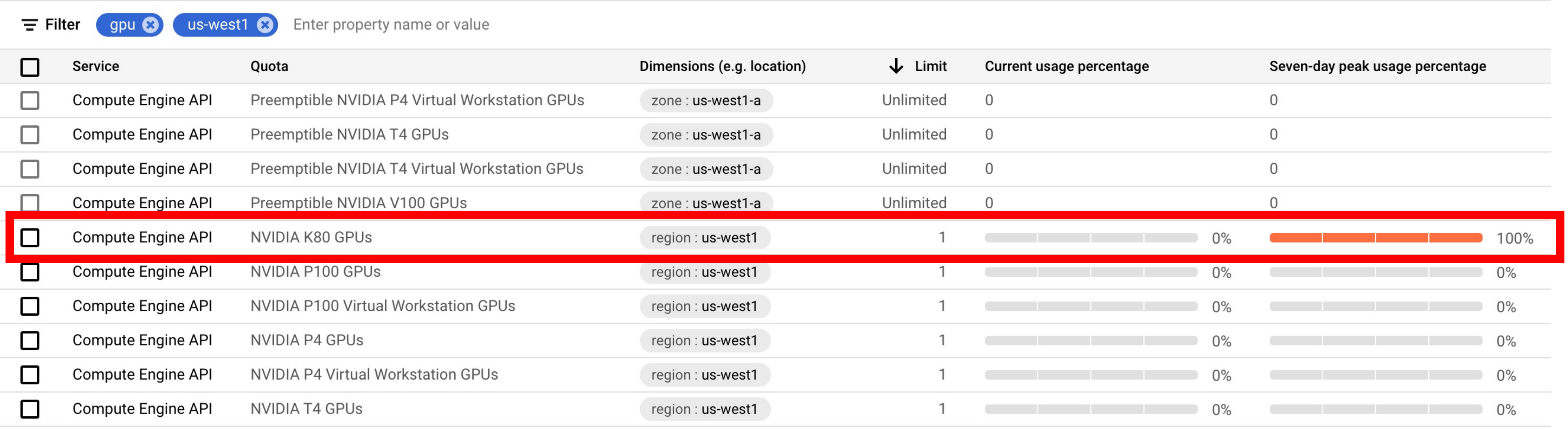
Enter a new quota limit. A limit above 0 will require approval from your service provider. ?

New limit *
5

Request description *
Plan to use 5 GPUs in us-west1-b zone to train ML model

Your description will be sent to your service provider and is used to evaluate your request. It's useful to include the intent of the quota usage, future growth plans, region or zone spread, and any additional requirements or dependencies.

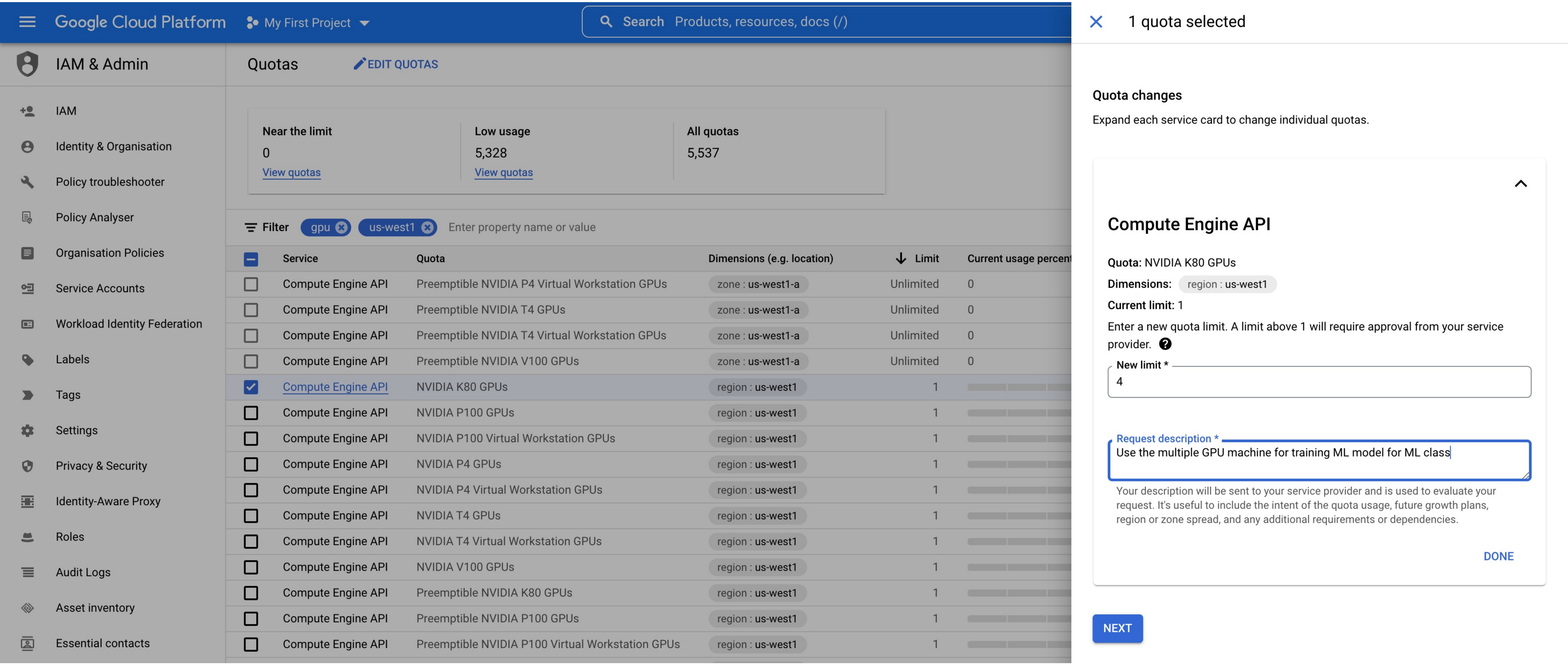
- Increase quota for *GPUs (all regions)*



Filter gpu us-west1 Enter property name or value

□	Service	Quota	Dimensions (e.g. location)	Limit	Current usage percentage	Seven-day peak usage percentage
□	Compute Engine API	Preemptible NVIDIA P4 Virtual Workstation GPUs	zone : us-west1-a	Unlimited	0	0
□	Compute Engine API	Preemptible NVIDIA T4 GPUs	zone : us-west1-a	Unlimited	0	0
□	Compute Engine API	Preemptible NVIDIA T4 Virtual Workstation GPUs	zone : us-west1-a	Unlimited	0	0
□	Compute Engine API	Preemptible NVIDIA V100 GPUs	zone : us-west1-a	Unlimited	0	0
□	Compute Engine API	NVIDIA K80 GPUs	region : us-west1	1	0%	100%
□	Compute Engine API	NVIDIA P100 GPUs	region : us-west1	1	0%	0%
□	Compute Engine API	NVIDIA P100 Virtual Workstation GPUs	region : us-west1	1	0%	0%
□	Compute Engine API	NVIDIA P4 GPUs	region : us-west1	1	0%	0%
□	Compute Engine API	NVIDIA P4 Virtual Workstation GPUs	region : us-west1	1	0%	0%
□	Compute Engine API	NVIDIA T4 GPUs	region : us-west1	1	0%	0%

- Increase quota for specific region and type of GPU you want to use (for example NVIDIA K80 GPUs at us-west1 is limited to 1 in the screenshot)



Near the limit	Low usage	All quotas
0	5,328	5,537

☐	Service	Quota	Dimensions (e.g. location)	Limit	Current usage percent
☐	Compute Engine API	Preemptible NVIDIA P4 Virtual Workstation GPUs	zone : us-west1-a	Unlimited	0
☐	Compute Engine API	Preemptible NVIDIA T4 GPUs	zone : us-west1-a	Unlimited	0
☐	Compute Engine API	Preemptible NVIDIA T4 Virtual Workstation GPUs	zone : us-west1-a	Unlimited	0
☐	Compute Engine API	Preemptible NVIDIA V100 GPUs	zone : us-west1-a	Unlimited	0
☑	Compute Engine API	NVIDIA K80 GPUs	region : us-west1	1	
☐	Compute Engine API	NVIDIA P100 GPUs	region : us-west1	1	
☐	Compute Engine API	NVIDIA P100 Virtual Workstation GPUs	region : us-west1	1	
☐	Compute Engine API	NVIDIA P4 GPUs	region : us-west1	1	
☐	Compute Engine API	NVIDIA P4 Virtual Workstation GPUs	region : us-west1	1	
☐	Compute Engine API	NVIDIA T4 GPUs	region : us-west1	1	
☐	Compute Engine API	NVIDIA T4 Virtual Workstation GPUs	region : us-west1	1	
☐	Compute Engine API	NVIDIA V100 GPUs	region : us-west1	1	
☐	Compute Engine API	Preemptible NVIDIA K80 GPUs	region : us-west1	1	
☐	Compute Engine API	Preemptible NVIDIA P100 GPUs	region : us-west1	1	
☐	Compute Engine API	Preemptible NVIDIA P100 Virtual Workstation GPUs	region : us-west1	1	

1 quota selected

Quota changes

Expand each service card to change individual quotas.

Compute Engine API

Quota: NVIDIA K80 GPUs

Dimensions: region : us-west1

Current limit: 1

Enter a new quota limit. A limit above 1 will require approval from your service provider.

New limit *: 4

Request description *: Use the multiple GPU machine for training ML model for ML class

Your description will be sent to your service provider and is used to evaluate your request. It's useful to include the intent of the quota usage, future growth plans, region or zone spread, and any additional requirements or dependencies.

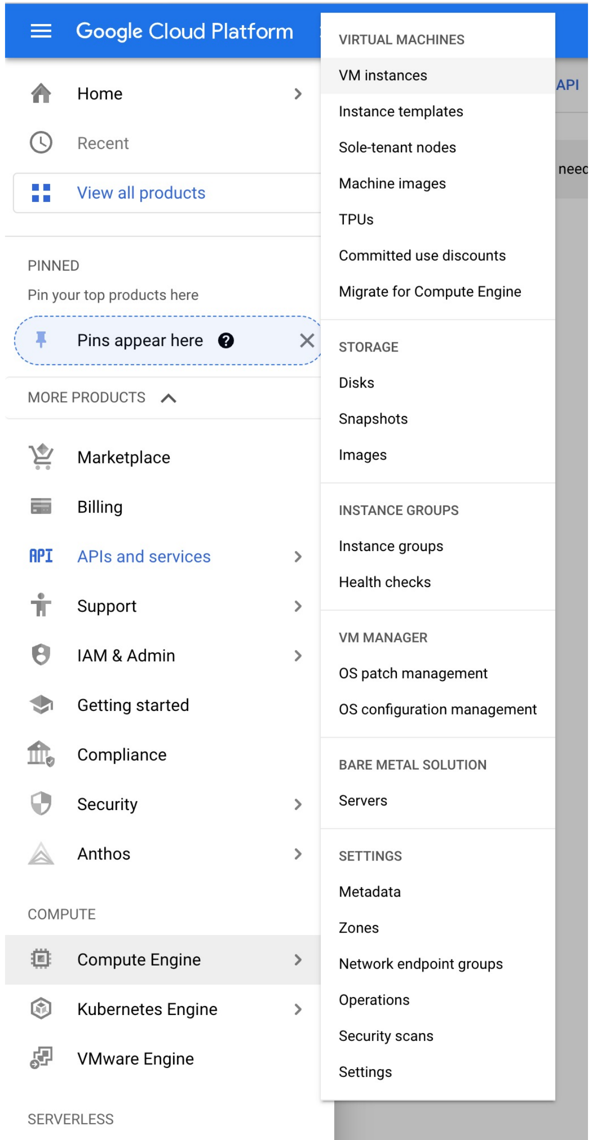
- Increase quota for specific region and type of GPU you want to use (for example NVIDIA K80 GPUs at us-west1 is limited to 1 in the screenshot)

GPU Choices

- GPUs on Compute Engine | Compute Engine Documentation | Google Cloud
- GPUs pricing | Compute Engine: Virtual Machines (VMs) | Google Cloud
- GPU regions and zones availability | Compute Engine Documentation | Google Cloud

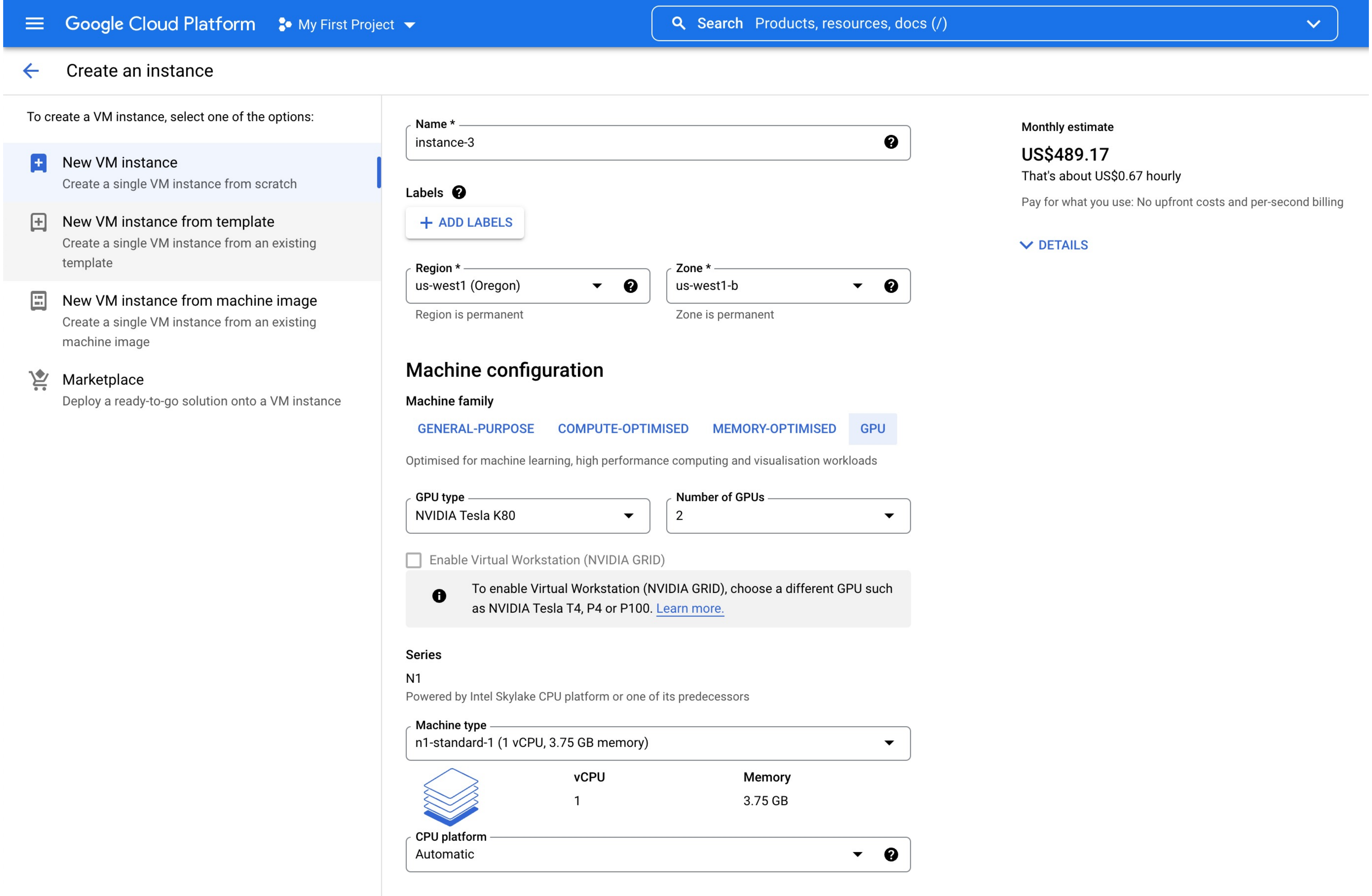
4. Create an instance with attached GPUs

- Enter "Compute Engine" > "VM Instances" > "Create Instance"



4. Create an instance with attached GPUs

- Create an instance
- Choose region and zone that has the GPU you requested
    - Check region supported GPU types in this link
    - For example, we choose “us-west1-b” to use K80 GPU
- Choose “GPU” under “Machine configuration”
- Select GPU type and number



Google Cloud Platform   My First Project ▾   Search Products, resources, docs (/)

← Create an instance

To create a VM instance, select one of the options:

New VM instance
Create a single VM instance from scratch

New VM instance from template
Create a single VM instance from an existing template

New VM instance from machine image
Create a single VM instance from an existing machine image

Marketplace
Deploy a ready-to-go solution onto a VM instance

Name *: instance-3

Labels
+ ADD LABELS

Region *: us-west1 (Oregon)
Region is permanent

Zone *: us-west1-b
Zone is permanent

Machine configuration

Machine family

GENERAL-PURPOSE   COMPUTE-OPTIMISED   MEMORY-OPTIMISED   **GPU**

Optimised for machine learning, high performance computing and visualisation workloads

GPU type: NVIDIA Tesla K80

Number of GPUs: 2

☐ Enable Virtual Workstation (NVIDIA GRID)

To enable Virtual Workstation (NVIDIA GRID), choose a different GPU such as NVIDIA Tesla T4, P4 or P100. Learn more.

Series

N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type: n1-standard-1 (1 vCPU, 3.75 GB memory)

vCPU	Memory
1	3.75 GB

CPU platform: Automatic

Monthly estimate

US$489.17

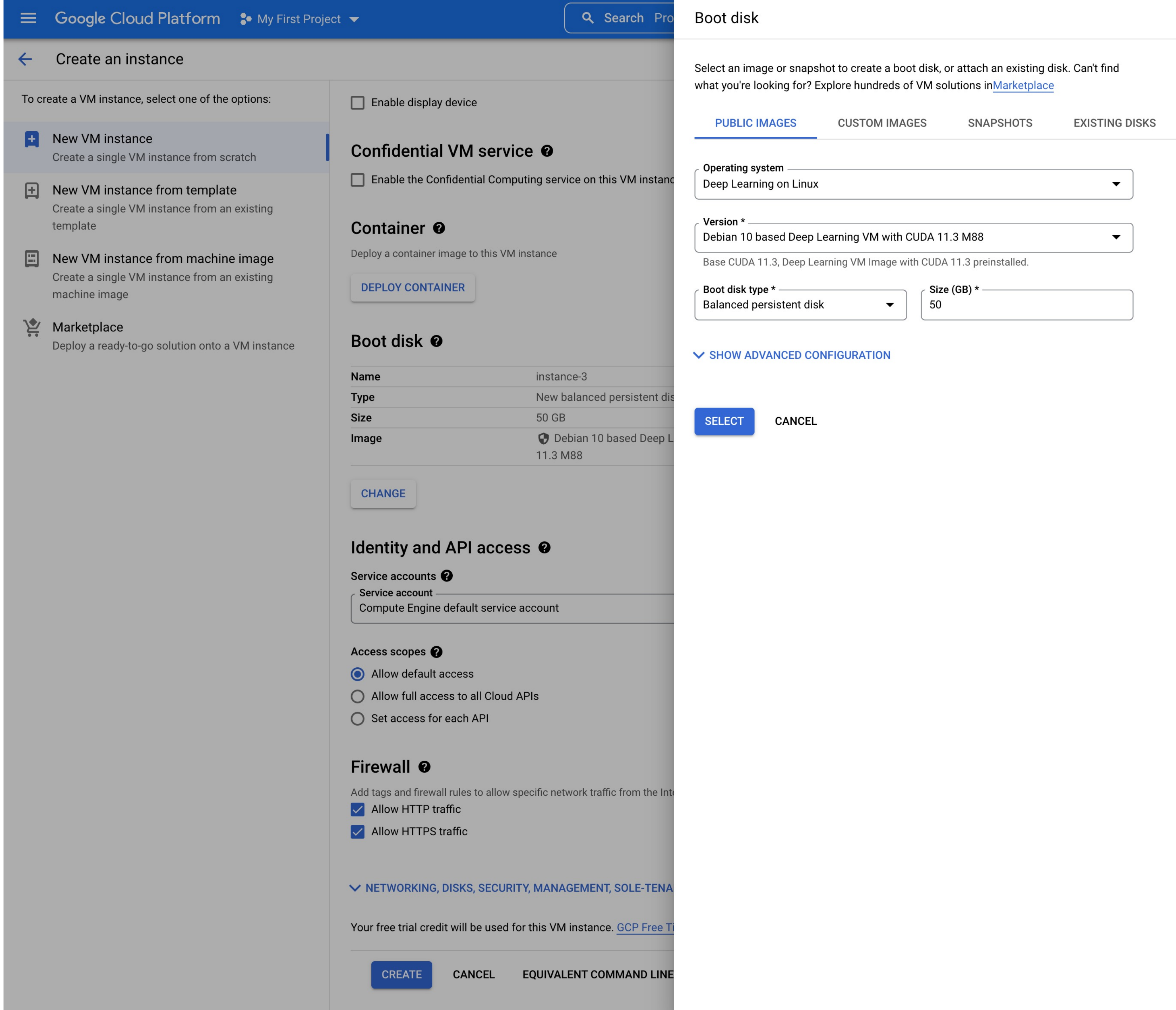
That's about US$0.67 hourly

Pay for what you use: No upfront costs and per-second billing

DETAILS

4. Create an instance with attached GPUs

- Choose Book disk and image
  - Use “Debian 10 based Deep Learning VM with CUDA 11.3” so that CUDA driver is installed already
  - Select size of the boot disk: should be enough for your data + code + saved trained model (the saved model might be large) etc
- Change firewall setting
  - Select allow HTTP and HTTPS traffic, so you can install packages and connect to GitHub server
- Click ″Create”



Google Cloud Platform   My First Project   Search Pro

Create an instance

To create a VM instance, select one of the options:

New VM instance
Create a single VM instance from scratch

New VM instance from template
Create a single VM instance from an existing template

New VM instance from machine image
Create a single VM instance from an existing machine image

Marketplace
Deploy a ready-to-go solution onto a VM instance

□ Enable display device

Confidential VM service

□ Enable the Confidential Computing service on this VM instanc

Container

Deploy a container image to this VM instance

DEPLOY CONTAINER

Boot disk

Name	instance-3
Type	New balanced persistent dis
Size	50 GB
Image	Debian 10 based Deep L 11.3 M88

CHANGE

Identity and API access

Service accounts

Service account
Compute Engine default service account

Access scopes

● Allow default access
○ Allow full access to all Cloud APIs
○ Set access for each API

Firewall

Add tags and firewall rules to allow specific network traffic from the Int

☑ Allow HTTP traffic
☑ Allow HTTPS traffic

NETWORKING, DISKS, SECURITY, MANAGEMENT, SOLE-TENA

Your free trial credit will be used for this VM instance. GCP Free Ti

CREATE   CANCEL   EQUIVALENT COMMAND LINE

Boot disk

Select an image or snapshot to create a boot disk, or attach an existing disk. Can't find what you're looking for? Explore hundreds of VM solutions inMarketplace

PUBLIC IMAGES   CUSTOM IMAGES   SNAPSHOTS   EXISTING DISKS

Operating system
Deep Learning on Linux

Version *
Debian 10 based Deep Learning VM with CUDA 11.3 M88
Base CUDA 11.3, Deep Learning VM Image with CUDA 11.3 preinstalled.

Boot disk type *
Balanced persistent disk

Size (GB) *
50

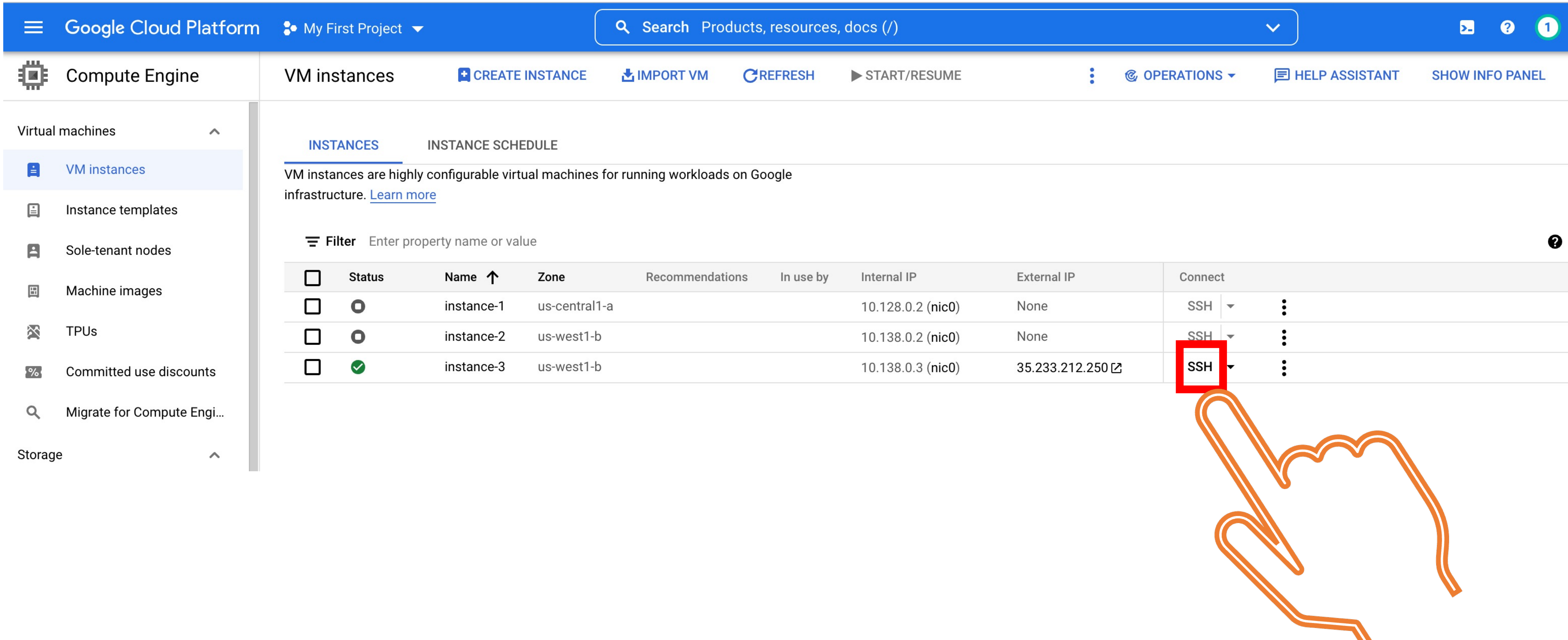
SHOW ADVANCED CONFIGURATION

SELECT   CANCEL

5. Install GPU driver

- If you choose the image with CUDA, your GPU driver will be installed automatically when you first login your machine
    - SSH into your machine in the Google Cloud portal (you have to login using your admin account to install the driver)
    - Input “y” when it prompts “Would you like to install the NVIDIA driver? ”

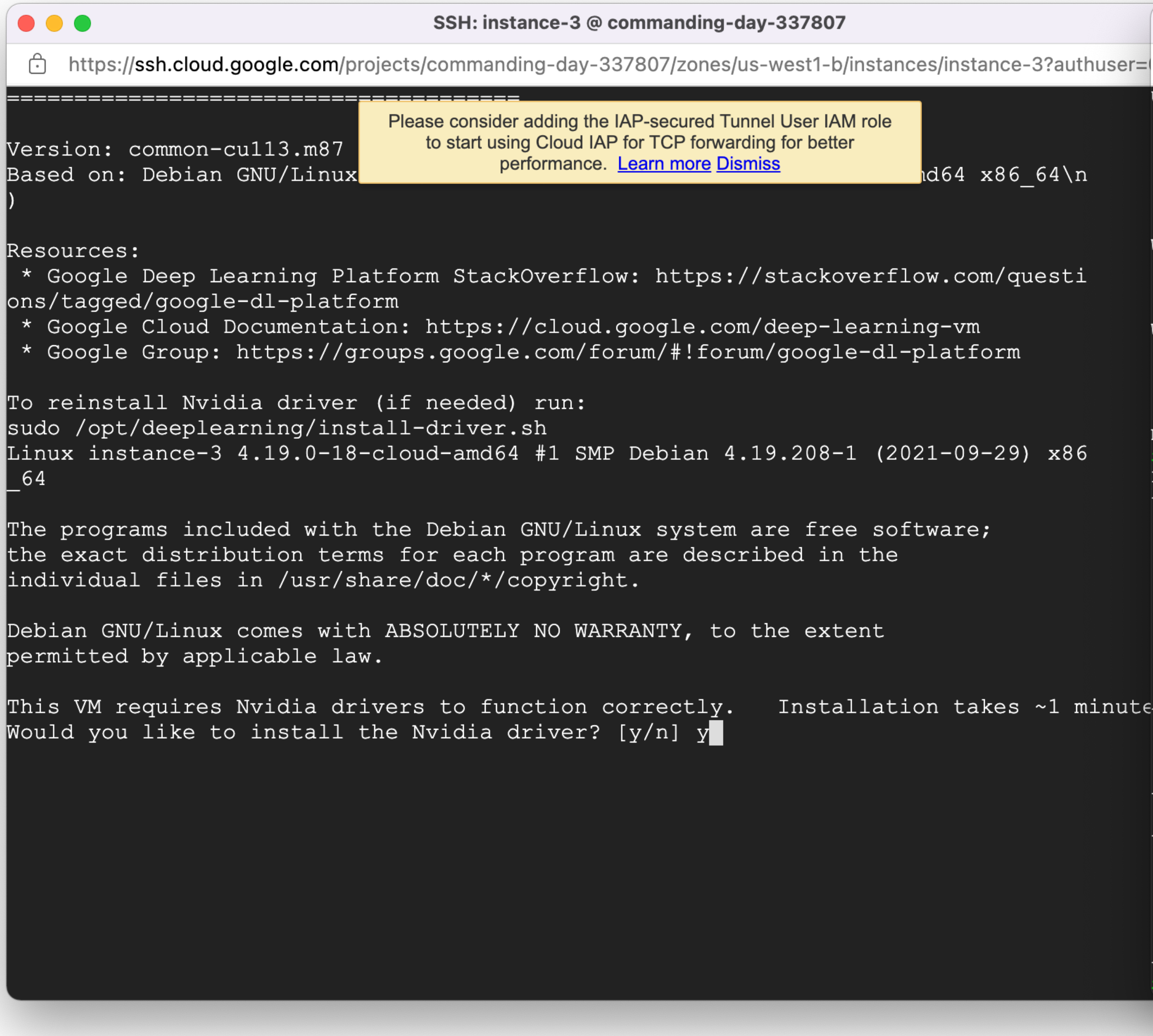
5. Install GPU driver



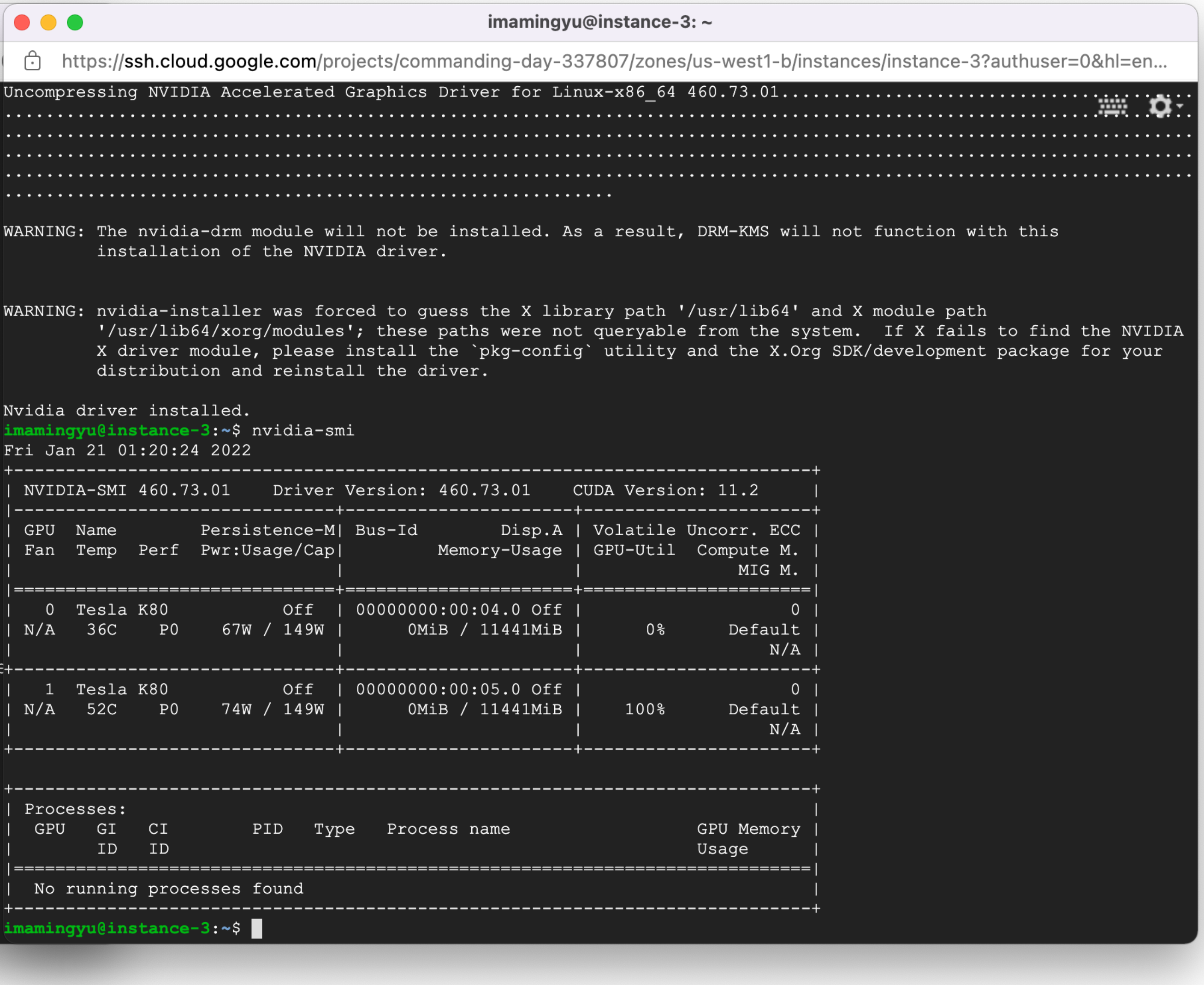
VM instances are highly configurable virtual machines for running workloads on Google infrastructure. Learn more

	Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect
□	○	instance-1	us-central1-a			10.128.0.2 (nic0)	None	SSH
□	○	instance-2	us-west1-b			10.138.0.2 (nic0)	None	SSH
□	✓	instance-3	us-west1-b			10.138.0.3 (nic0)	35.233.212.250	SSH

5. Install GPU driver


SSH: instance-3 @ commanding-day-337807
https://ssh.cloud.google.com/projects/commanding-day-337807/zones/us-west1-b/instances/instance-3?authuser=
Please consider adding the IAP-secured Tunnel User IAM role to start using Cloud IAP for TCP forwarding for better performance. Learn more Dismiss

```
======================================

Version: common-cu113.m87
Based on: Debian GNU/Linux                                    nd64 x86_64\n
)

Resources:
 * Google Deep Learning Platform StackOverflow: https://stackoverflow.com/questi
ons/tagged/google-dl-platform
 * Google Cloud Documentation: https://cloud.google.com/deep-learning-vm
 * Google Group: https://groups.google.com/forum/#!forum/google-dl-platform

To reinstall Nvidia driver (if needed) run:
sudo /opt/deeplearning/install-driver.sh
Linux instance-3 4.19.0-18-cloud-amd64 #1 SMP Debian 4.19.208-1 (2021-09-29) x86
_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.

This VM requires Nvidia drivers to function correctly.   Installation takes ~1 minute
Would you like to install the Nvidia driver? [y/n] y
```


imamingyu@instance-3: ~
https://ssh.cloud.google.com/projects/commanding-day-337807/zones/us-west1-b/instances/instance-3?authuser=0&hl=en...

```
Uncompressing NVIDIA Accelerated Graphics Driver for Linux-x86_64 460.73.01.................................
.........................................................................................................
.........................................................................................................
.........................................................................................................
.........................................................................................................
..........................................................

WARNING: The nvidia-drm module will not be installed. As a result, DRM-KMS will not function with this
         installation of the NVIDIA driver.


WARNING: nvidia-installer was forced to guess the X library path '/usr/lib64' and X module path
         '/usr/lib64/xorg/modules'; these paths were not queryable from the system.  If X fails to find the NVIDIA
         X driver module, please install the `pkg-config` utility and the X.Org SDK/development package for your
         distribution and reinstall the driver.

Nvidia driver installed.
imamingyu@instance-3:~$ nvidia-smi
Fri Jan 21 01:20:24 2022
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 460.73.01    Driver Version: 460.73.01    CUDA Version: 11.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla K80           Off  | 00000000:00:04.0 Off |                    0 |
| N/A   36C    P0    67W / 149W |      0MiB / 11441MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   1  Tesla K80           Off  | 00000000:00:05.0 Off |                    0 |
| N/A   52C    P0    74W / 149W |      0MiB / 11441MiB |    100%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
imamingyu@instance-3:~$
```

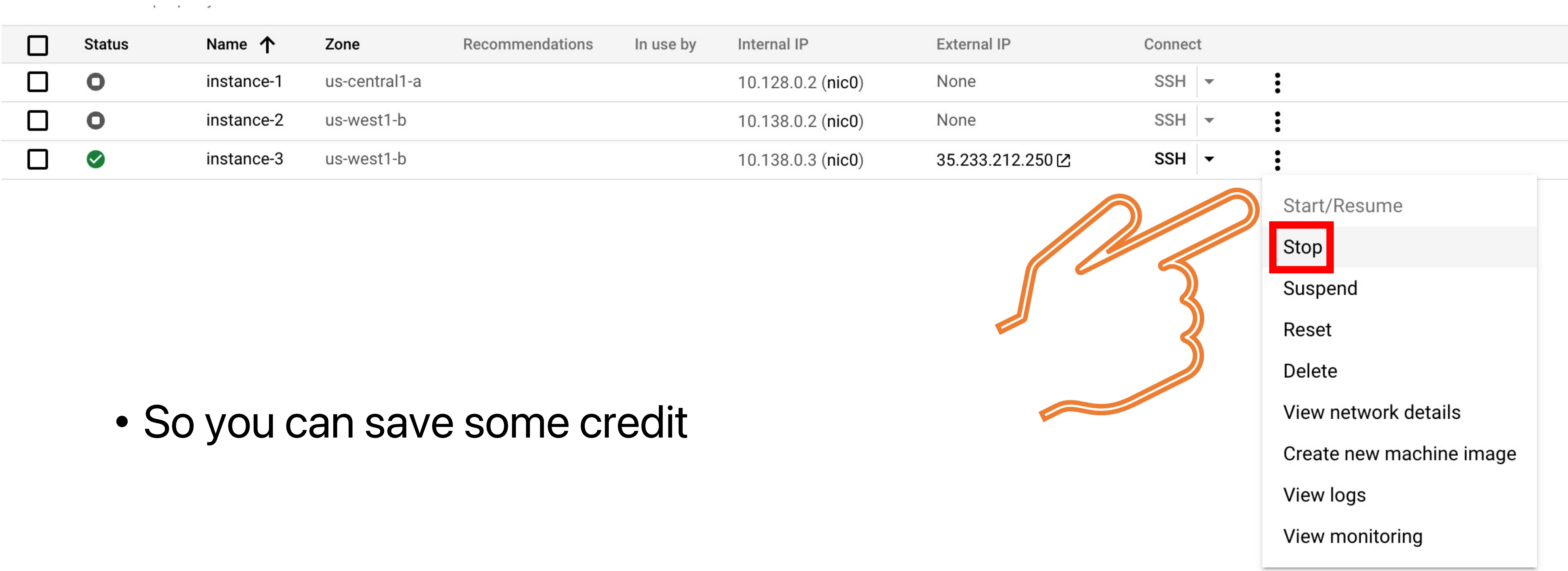
5. Install GPU driver

- Verify the GPU driver is installed
  - Type “nvidia-smi” command, you should see this if the driver is installed successfully

5. Install GPU driver

- Otherwise you could following steps in the following link
  - Installing GPU drivers | Compute Engine Documentation | Google Cloud
- We need to install
  - NVIDIA driver
  - CUDA toolkit
  - CUDA runtime

Turn off your machine when it’s not using



□	Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect	
□	■	instance-1	us-central1-a			10.128.0.2 (nic0)	None	SSH ▾	⋮
□	■	instance-2	us-west1-b			10.138.0.2 (nic0)	None	SSH ▾	⋮
□	✓	instance-3	us-west1-b			10.138.0.3 (nic0)	35.233.212.250	SSH ▾	⋮

- So you can save some credit

Run Sample Codebase

New environment file for the class project!

- Follow updated environment set up instruction and download the new “requirements.txt” file as shown in this commit

Install environment and run code

- Using git clone to download codebase
- Following project README to install environment and train your model

Install miniconda

```
>>> wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
>>> sh Miniconda3-latest-Linux-x86_64.sh
```

Create conda environment

```
>>> conda create -n cs188 python==3.8
>>> conda activate cs188
```

Install dependencies needed

```
>>> conda install pip
>>> pip3 --no-cache-dir install torch==1.10.1+cu113 torchvision==0.11.2+cu113
torchaudio==0.10.1+cu113 -f https://download.pytorch.org/whl/cu113/torch_stable.html
>>> pip install -r requirements.txt
```

Run a training script

```
>>> sh scripts/train_com2sense.sh
```

Tips for Experiments on Remote Machine

Connect to your instance

- If you would like to connect to your machine using terminal directly, instead of using the browser-based ssh window
- Create key (Detailed tutorial: [How to Use SSH Public Key Authentication – ServerPilot](#))
  - Using command `ssh-keygen`
  - You will keep the private key (for example `id_rsa`) in your local computer
- Add key
  - Add public key (like `id_rsa.pub`) to your Google Cloud instance setting
  - Click into your instance, click “Edit” in the top navigation bar, find “SSH key”, click “Add Item”, enter your public SSH key content there
- Connect your remote instance from your local terminal

```
ssh –i key_path username@external_ip_address
```

- [Connecting to Linux VMs using advanced methods | Compute Engine Documentation | Google Cloud](#)

Access file and coding remotely

- You will need to edit code and run the updated codebase with new implementation
- Choice 1: VS Code
  - Developing on Remote Machines using SSH and Visual Studio Code
- Choice 2: PyCharm
  - Getting started with remote development | PyCharm (jetbrains.com)
- Choice 3: transfer files by scp/sftp
  - Using scp/sftp to transfer file/code from your local machine to the remote machine

Monitor and specify GPU usage

- Check whether your job is running on GPU, memory usage, job ID etc
  - `nvidia-smi`
- Specify which GPU(s) to use
  - `export CUDA_VISIBLE_DEVICES="0”`
  - `export CUDA_VISIBLE_DEVICES="0,1,2”`
  - `export CUDA_VISIBLE_DEVICES=""`

Run experiments in background

- Use tmux to run your job in background, so your job can continue running if your ssh session broke
- `tmux new –s exp1`
  - Create a new tmux session
- `control + b`, then press `d`
  - Exit the session
- `tmux a –t exp1`
  - Enter the session exp1 again
- `tmux ls`
  - See all active sessions

Use Jupyter Notebook on Google Cloud

- Running Jupyter Notebook on Google Cloud Platform in 15 min | by Amulya Aankul | Towards Data Science

Google Colab

- Another choice for using GPU
- It has a free version, but you cannot use your Google Cloud credit for Colab
- We will introduce how to use colab in our demo next week