Carnegie Mellon University

Introduction to AWS/GCP

16-824 Visual Learning and Recognition (Fall 2022)

Presented by Vanshaj (vanshajc@)

Slides credit to Spring 2022 TA Group, updated by Fall 2022 TAs

Course Logistics

- You will receive $50 of AWS credit per homework
  - We will send this credit out soon via email
- $50 GCP credit sent for HW1
  - We will send $100 more GCP credits later

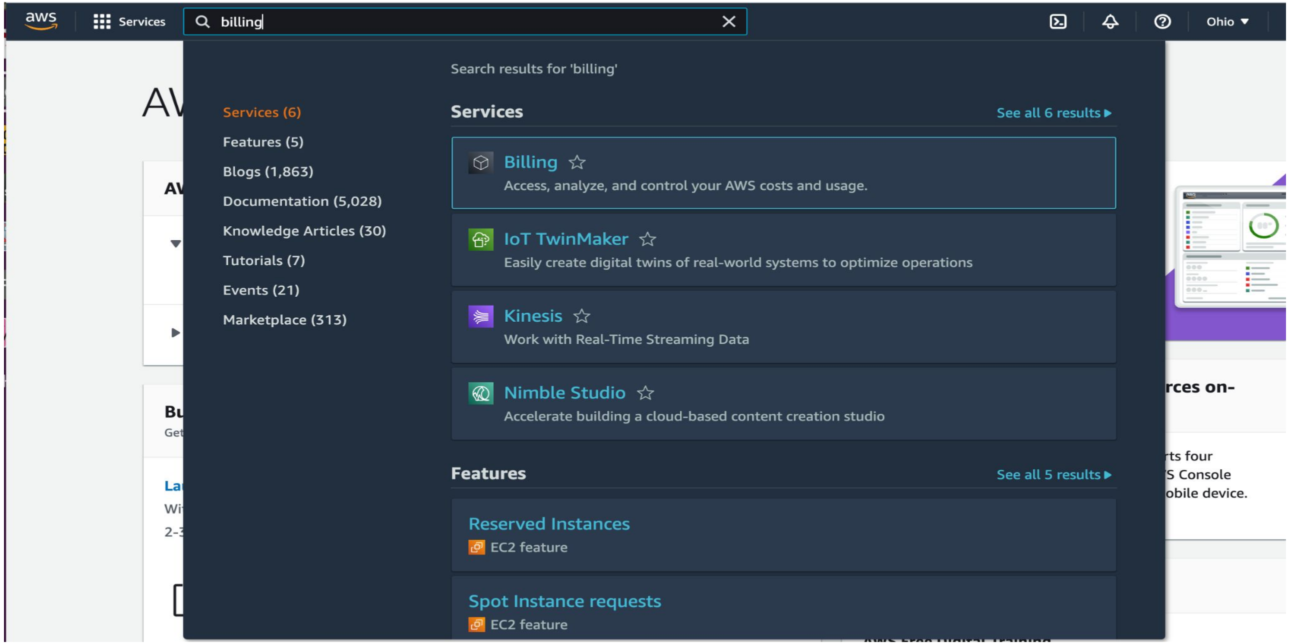
Other Resources

- Personal Laptop / Lab GPU Server
  - We cannot provide installation support
- Google Colab
  - Free GPU access (time limited)
  - Very easy to set up
  - Might need pro ($10/mo) for smoother experience

AWS Account Setup

- **Use andrew email address**
- https://aws.amazon.com/console/
- Redeem credits via billing console

- Apply the credits to your AWS account
    - AWS console - Billing - Credits - Redeem Credit - Enter Promo Code



Search results for 'billing'

- Services (6)
- Features (5)
- Blogs (1,863)
- Documentation (5,028)
- Knowledge Articles (30)
- Tutorials (7)
- Events (21)
- Marketplace (313)

Services See all 6 results ▸

Billing ☆
Access, analyze, and control your AWS costs and usage.

IoT TwinMaker ☆
Easily create digital twins of real-world systems to optimize operations

Kinesis ☆
Work with Real-Time Streaming Data

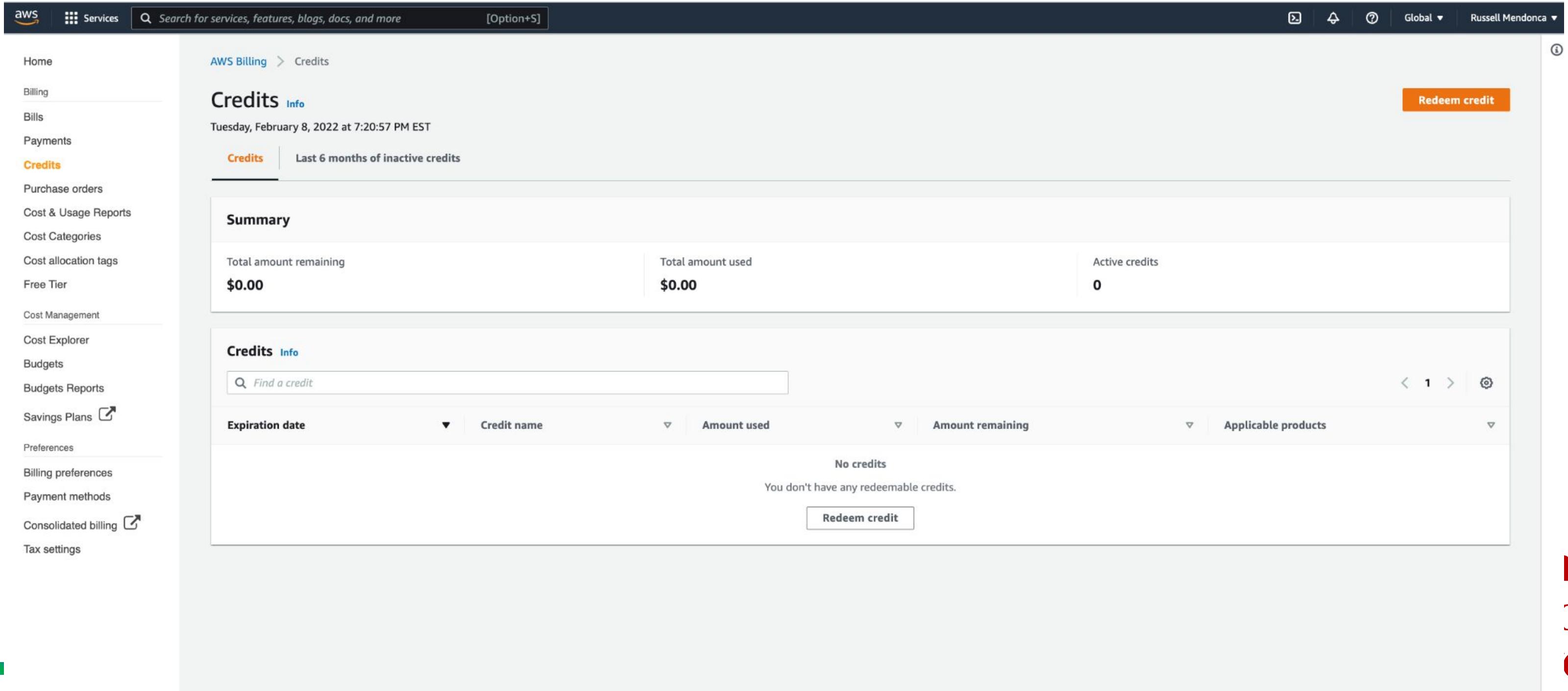
Nimble Studio ☆
Accelerate building a cloud-based content creation studio

Features See all 5 results ▸

Reserved Instances
EC2 feature

Spot Instance requests
EC2 feature

- Apply the credits to your AWS account
    - AWS console - Billing - Credits - Redeem Credit - Enter Promo Code



Credits Info

Tuesday, February 8, 2022 at 7:20:57 PM EST

Summary

Total amount remaining	Total amount used	Active credits
$0.00	$0.00	0

Credits Info

Expiration date	Credit name	Amount used	Amount remaining	Applicable products

No credits

You don't have any redeemable credits.

Amazon Elastic Compute Cloud (EC2)

- Visual learning is computation-intensive, requiring special hardware: GPU
- EC2 provides resizable compute capacity in the cloud
- Easy deployment of hardware, system softwares and data storage according to your demand and budget

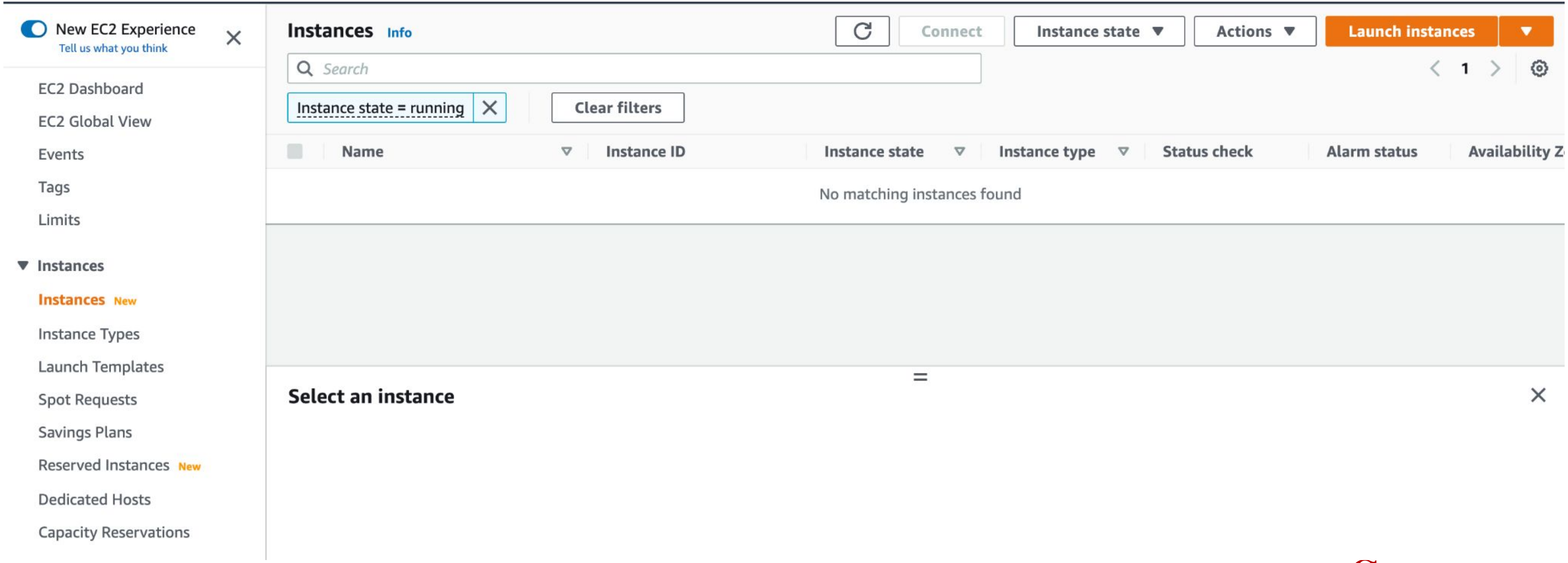
EC2 Instances

- An instance is a virtual machine with specific hardware configuration
- Life cycle of an instance:
  - Launch
  - **Stop - when you are not working on EC2, only charges storage**
  - (Re)start
  - Terminate - be careful, the instance and root storage will be deleted.

EC2 Instances Types

- Various instance types available - different in RAMs, CPUs, GPUs
- t2.micro: baseline CPU performance
  - 750 hours of free usage per month
  - (Optional) write code; test part of code
- p2.xlarge: what we use for assignments and projects
  - 4-core CPU, Nvidia Tesla K80 (11,441MB memory)
  - **costs $0.9 per Hour**
- g4dn instances: alternative for initial testing with GPU

EC2 Dashboard



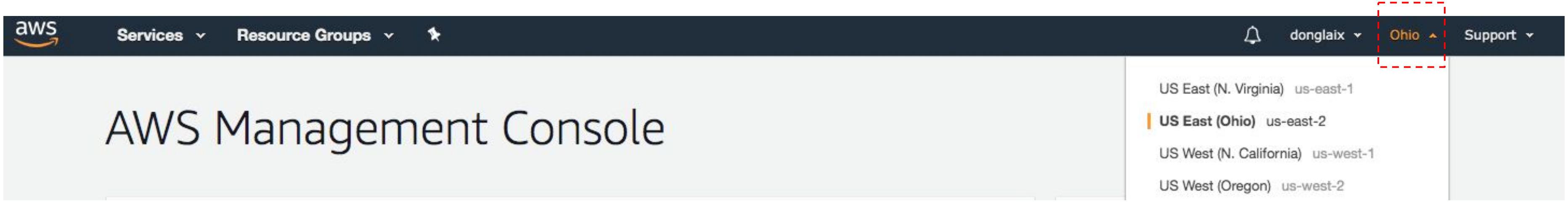
Instance state = running ✕

□	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Z

No matching instances found

Select an instance

EC2 Regions



AWS Management Console

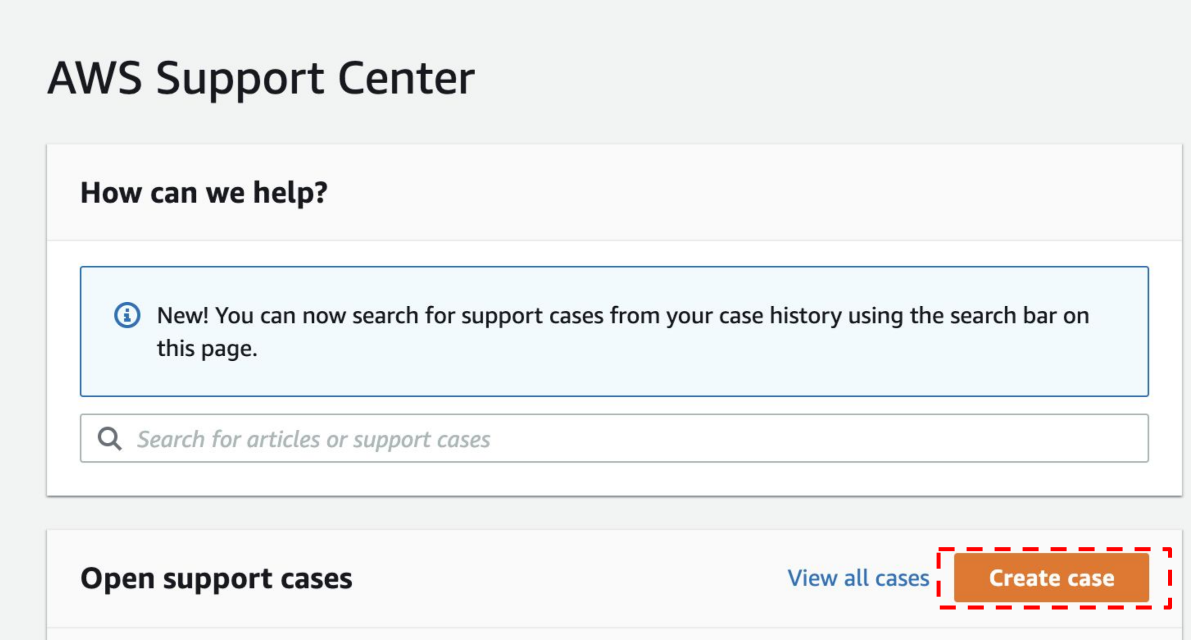
US East (N. Virginia) us-east-1

US East (Ohio) us-east-2

US West (N. California) us-west-1

US West (Oregon) us-west-2

EC2 Limits



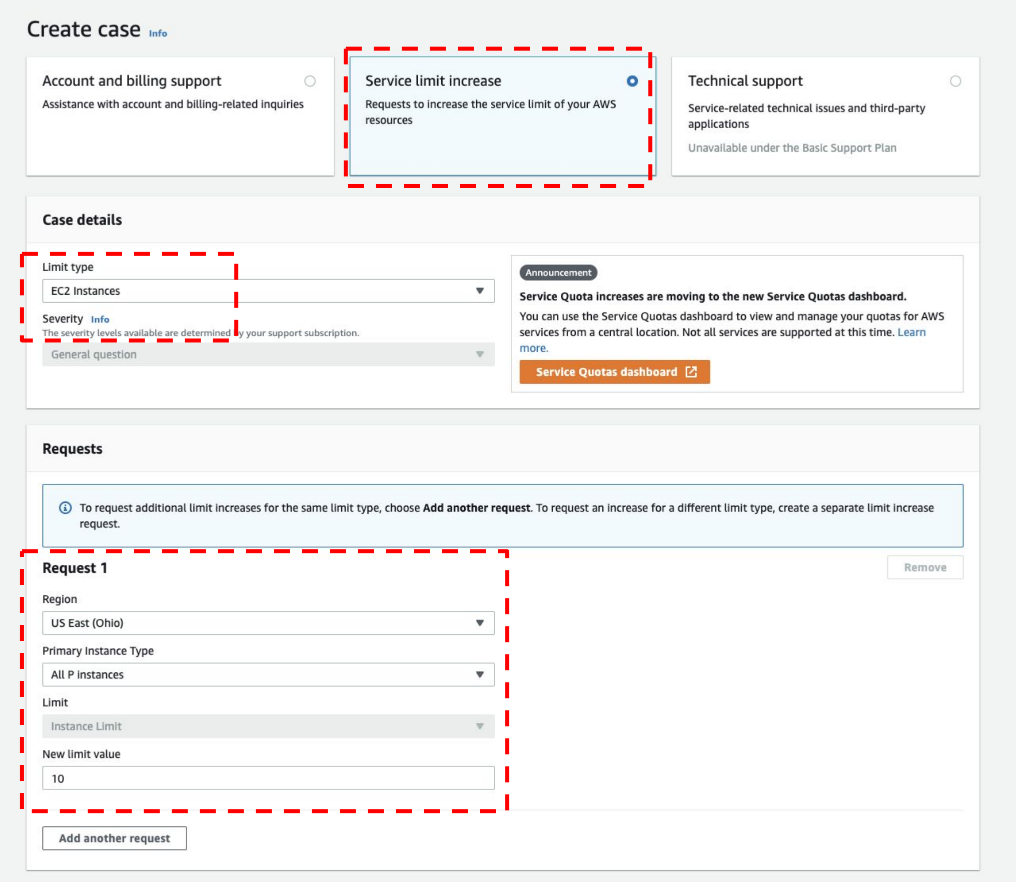
AWS Support Center

How can we help?

New! You can now search for support cases from your case history using the search bar on this page.

- Increase limit for P instances and G instances
- You won’t be able to launch any instances with GPU until you request this service limit increase
- **Do this ASAP - it takes a while for AWS to approve**

EC2 Limits



Create case Info

Account and billing support ○
Assistance with account and billing-related inquiries

Service limit increase ●
Requests to increase the service limit of your AWS resources

Technical support ○
Service-related technical issues and third-party applications
Unavailable under the Basic Support Plan

Case details

Limit type: EC2 Instances

Severity Info
The severity levels available are determined by your support subscription.
General question

Announcement

Service Quota increases are moving to the new Service Quotas dashboard.

You can use the Service Quotas dashboard to view and manage your quotas for AWS services from a central location. Not all services are supported at this time. Learn more.

Service Quotas dashboard

Requests

To request additional limit increases for the same limit type, choose **Add another request**. To request an increase for a different limit type, create a separate limit increase request.

Request 1 Remove

Region: US East (Ohio)

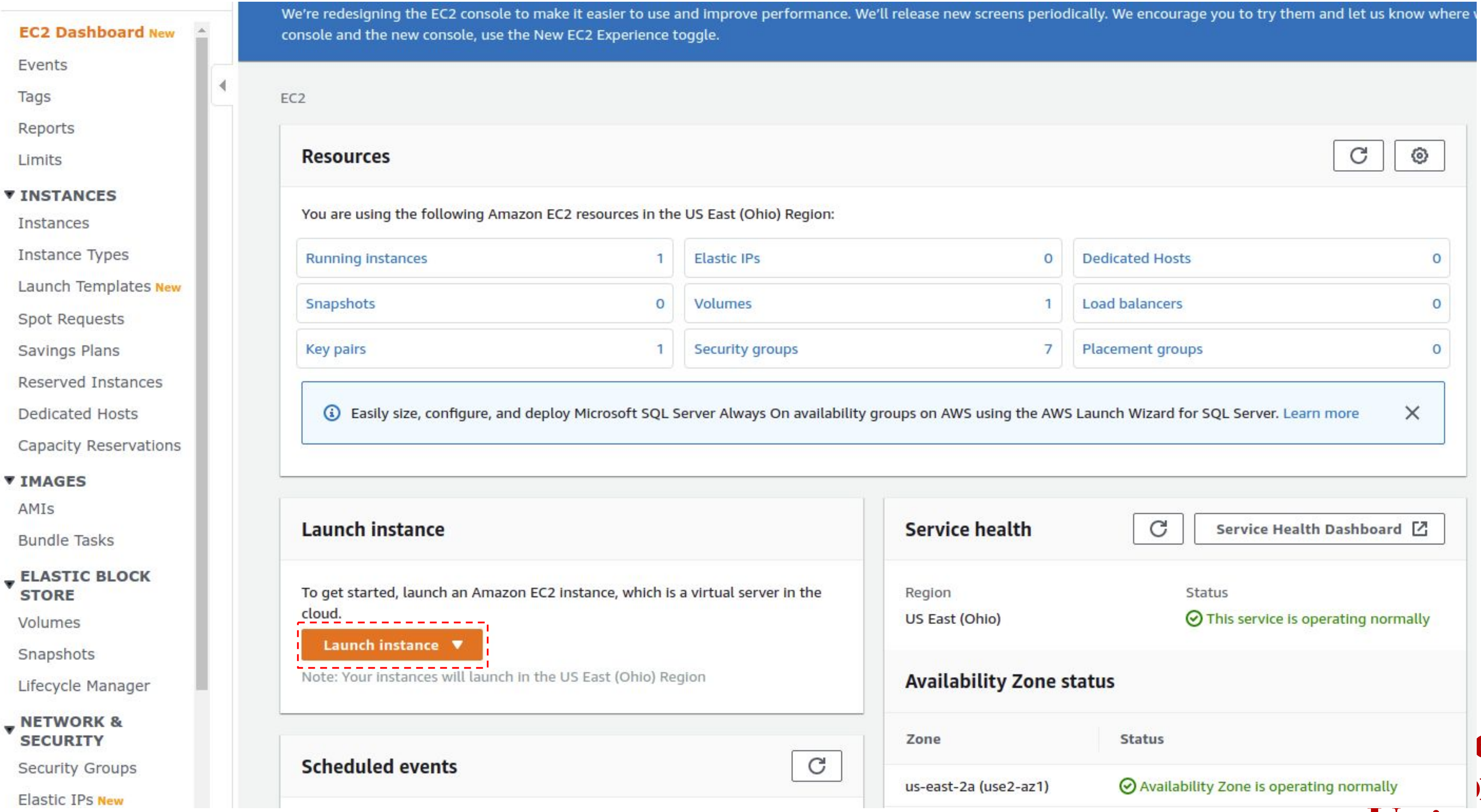
Primary Instance Type: All P instances

Limit: Instance Limit

New limit value: 10

Add another request

- Increase limit for P instances and G instances
- You won’t be able to launch any instances with GPU until you request this service limit increase
- **Do this ASAP - it takes a while for AWS to approve**



We're redesigning the EC2 console to make it easier to use and improve performance. We'll release new screens periodically. We encourage you to try them and let us know where
console and the new console, use the New EC2 Experience toggle.

Resources

You are using the following Amazon EC2 resources in the US East (Ohio) Region:

Running instances	1	Elastic IPs	0	Dedicated Hosts	0
Snapshots	0	Volumes	1	Load balancers	0
Key pairs	1	Security groups	7	Placement groups	0

ⓘ Easily size, configure, and deploy Microsoft SQL Server Always On availability groups on AWS using the AWS Launch Wizard for SQL Server. Learn more ✕

Launch instance

To get started, launch an Amazon EC2 instance, which is a virtual server in the cloud.

Note: Your instances will launch in the US East (Ohio) Region

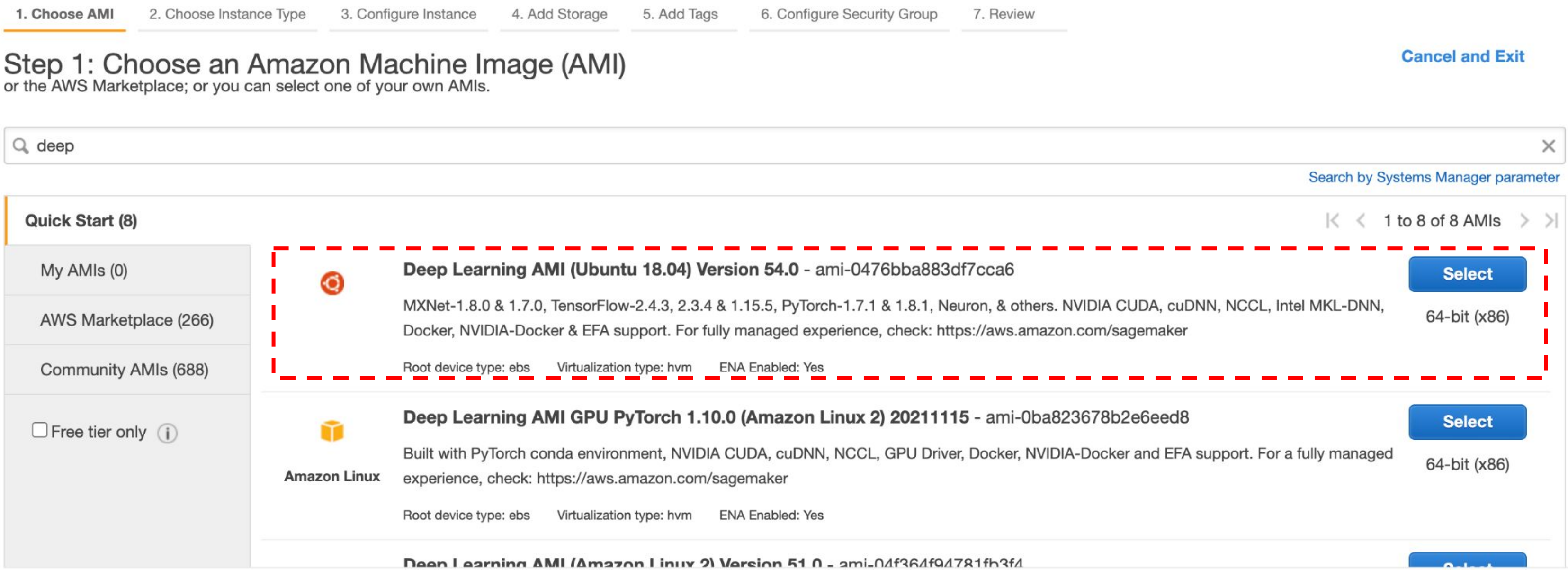
Scheduled events

Service health

Region	Status
US East (Ohio)	✓ This service is operating normally

Availability Zone status

Zone	Status
us-east-2a (use2-az1)	✓ Availability Zone is operating normally



Step 1: Choose an Amazon Machine Image (AMI)

or the AWS Marketplace; or you can select one of your own AMIs.

deep

Deep Learning AMI (Ubuntu 18.04) Version 54.0 - ami-0476bba883df7cca6

MXNet-1.8.0 & 1.7.0, TensorFlow-2.4.3, 2.3.4 & 1.15.5, PyTorch-1.7.1 & 1.8.1, Neuron, & others. NVIDIA CUDA, cuDNN, NCCL, Intel MKL-DNN, Docker, NVIDIA-Docker & EFA support. For fully managed experience, check: https://aws.amazon.com/sagemaker

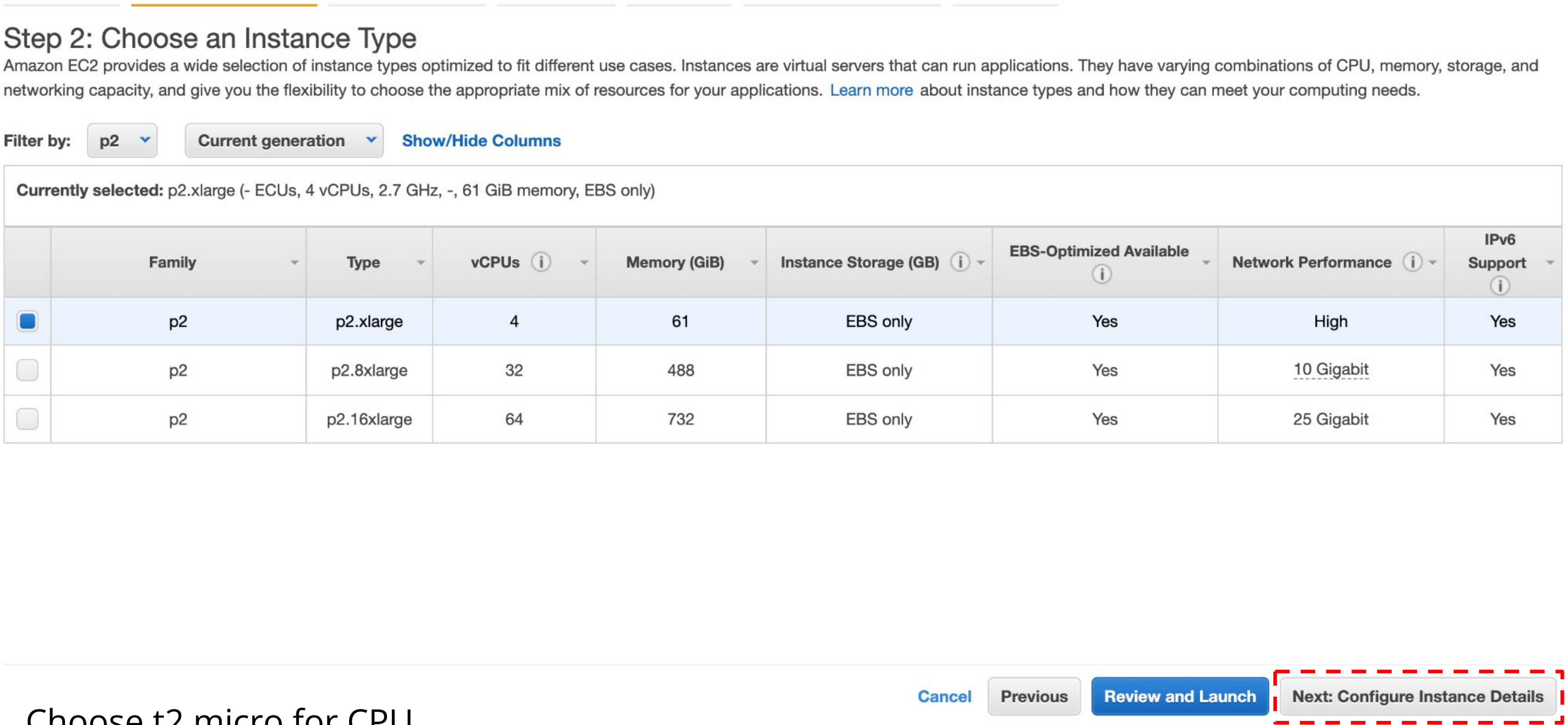
Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

Deep Learning AMI GPU PyTorch 1.10.0 (Amazon Linux 2) 20211115 - ami-0ba823678b2e6eed8

Built with PyTorch conda environment, NVIDIA CUDA, cuDNN, NCCL, GPU Driver, Docker, NVIDIA-Docker and EFA support. For a fully managed experience, check: https://aws.amazon.com/sagemaker

Root device type: ebs Virtualization type: hvm ENA Enabled: Yes

Deep Learning AMI (Amazon Linux 2) Version 51.0 - ami-04f364f94781fb3f4



Step 2: Choose an Instance Type

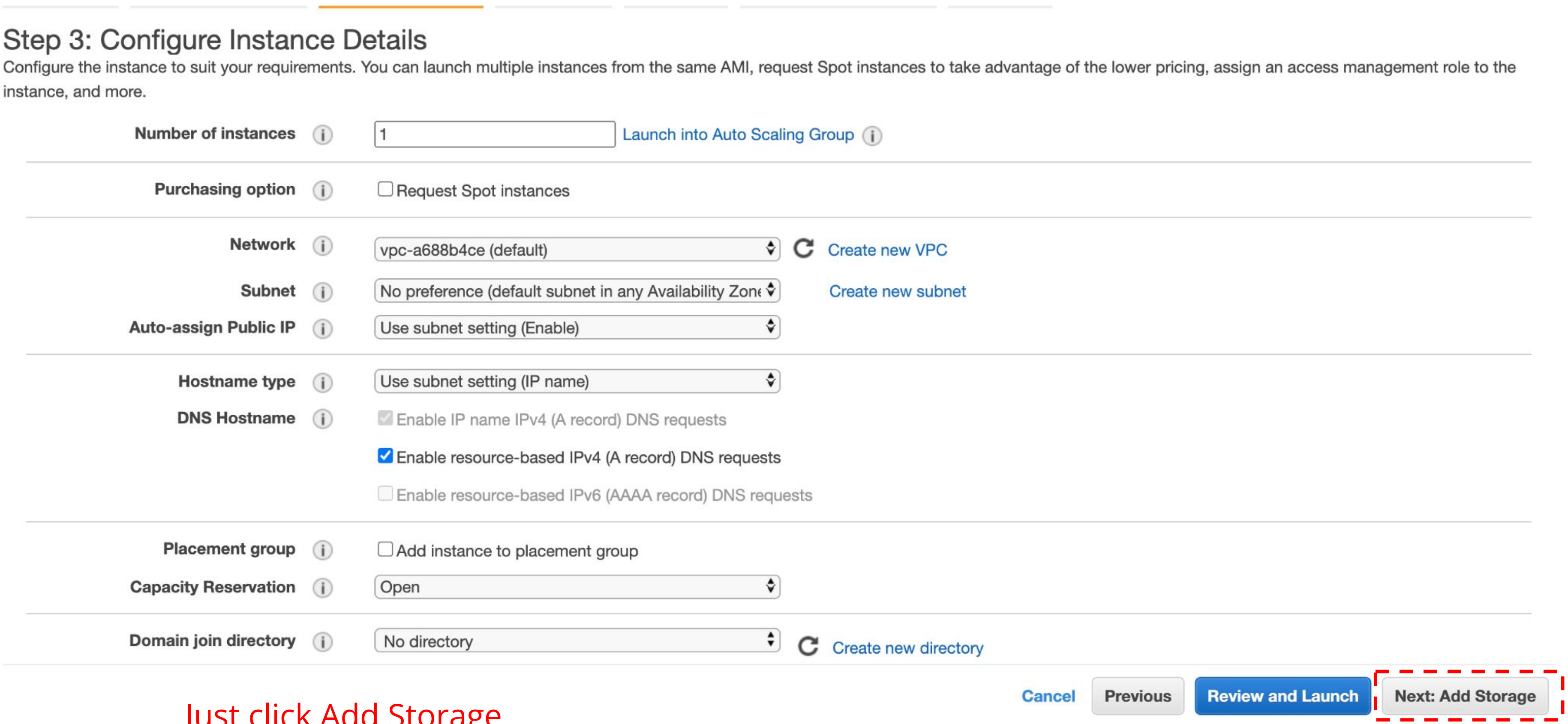
Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. Learn more about instance types and how they can meet your computing needs.

Filter by: p2 | Current generation | Show/Hide Columns

Currently selected: p2.xlarge (- ECUs, 4 vCPUs, 2.7 GHz, -, 61 GiB memory, EBS only)

	Family	Type	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance	IPv6 Support
☑	p2	p2.xlarge	4	61	EBS only	Yes	High	Yes
□	p2	p2.8xlarge	32	488	EBS only	Yes	10 Gigabit	Yes
□	p2	p2.16xlarge	64	732	EBS only	Yes	25 Gigabit	Yes

Choose t2.micro for CPU,
or if P instance type limits aren’t set



Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign an access management role to the instance, and more.

Number of instances: 1

Purchasing option: ☐ Request Spot instances

Network: vpc-a688b4ce (default)

Subnet: No preference (default subnet in any Availability Zone

Auto-assign Public IP: Use subnet setting (Enable)

Hostname type: Use subnet setting (IP name)

DNS Hostname: ☑ Enable IP name IPv4 (A record) DNS requests

☑ Enable resource-based IPv4 (A record) DNS requests

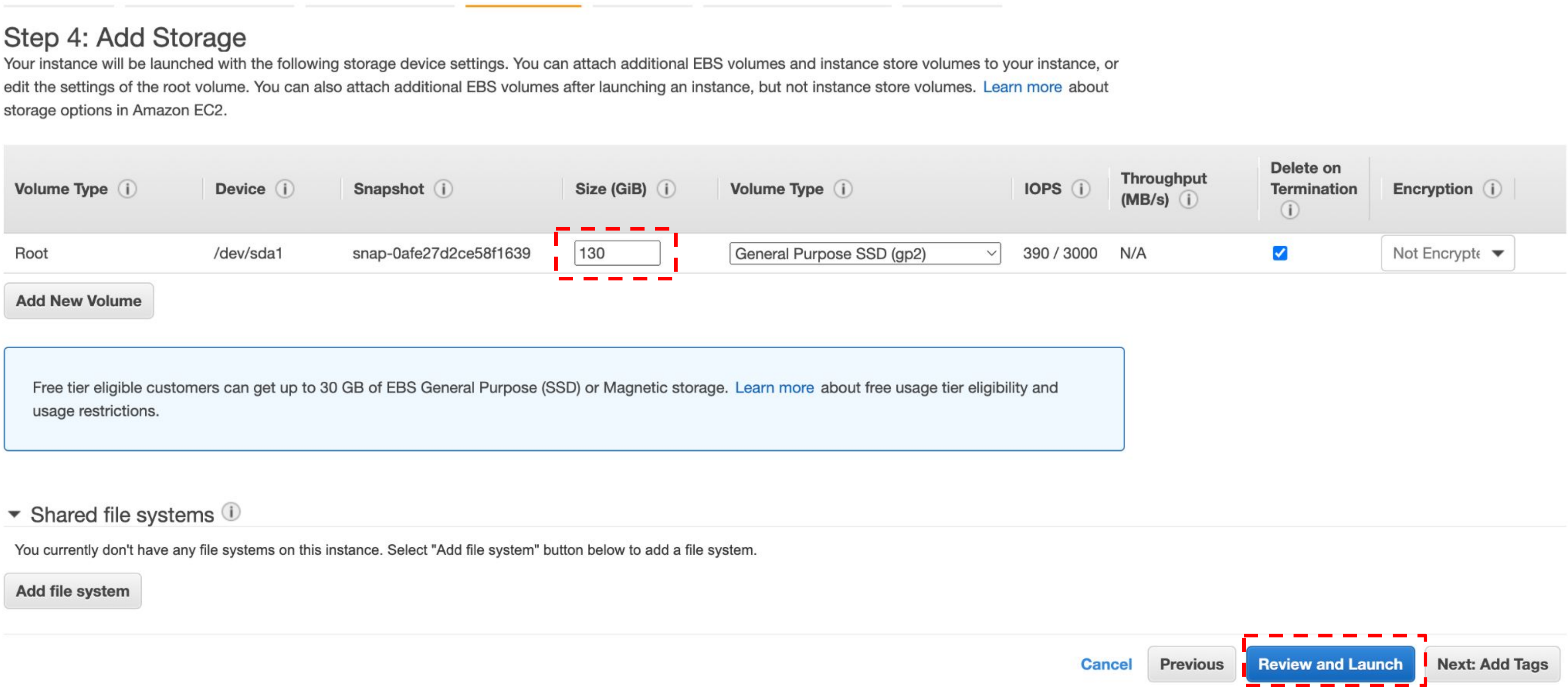
☐ Enable resource-based IPv6 (AAAA record) DNS requests

Placement group: ☐ Add instance to placement group

Capacity Reservation: Open

Domain join directory: No directory

Just click Add Storage
Do NOT change any settings



Step 4: Add Storage

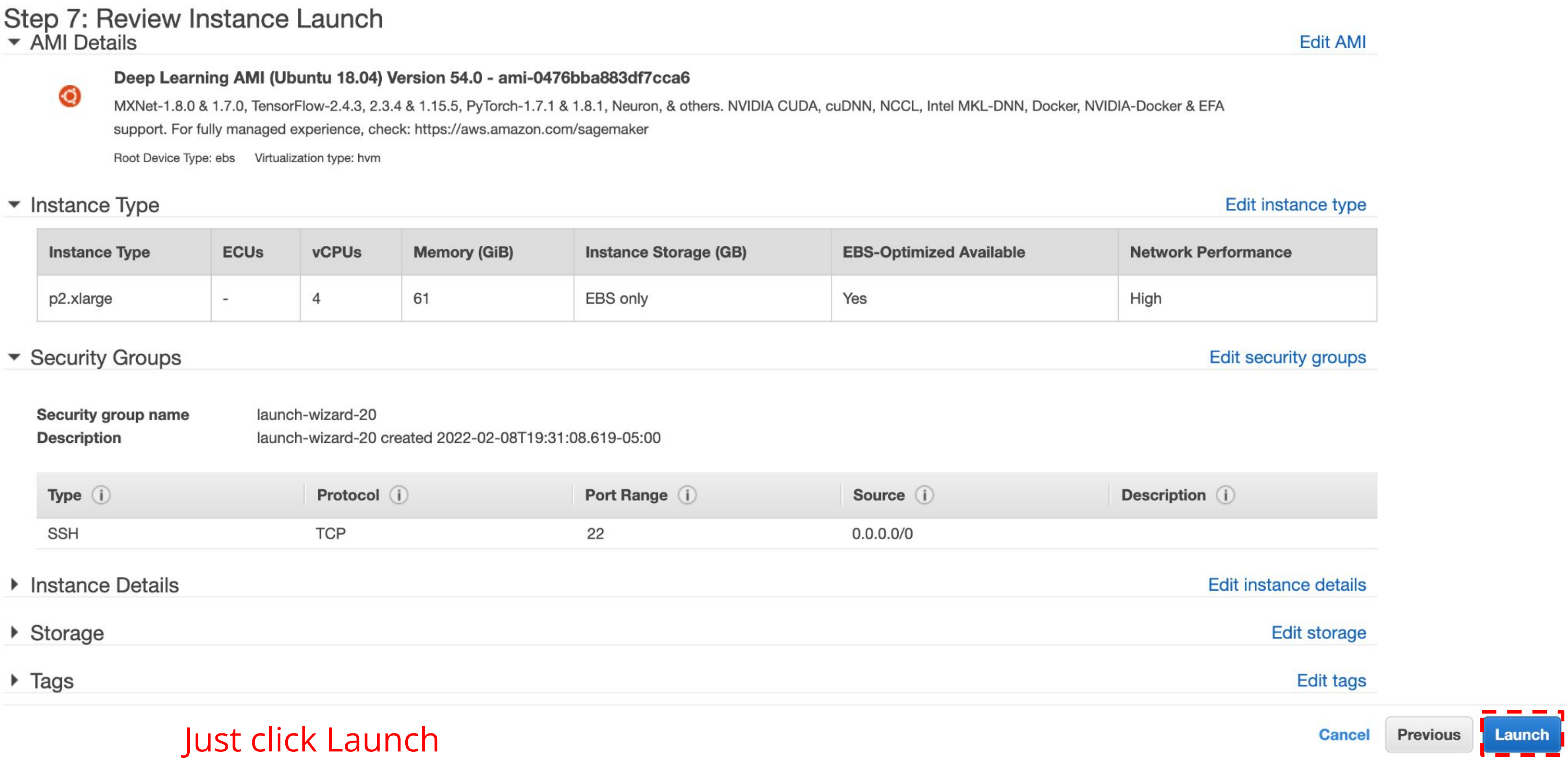
Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. Learn more about storage options in Amazon EC2.

Volume Type	Device	Snapshot	Size (GiB)	Volume Type	IOPS	Throughput (MB/s)	Delete on Termination	Encryption
Root	/dev/sda1	snap-0afe27d2ce58f1639	130	General Purpose SSD (gp2)	390 / 3000	N/A	☑	Not Encrypte

Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage. Learn more about free usage tier eligibility and usage restrictions.

Shared file systems

You currently don't have any file systems on this instance. Select "Add file system" button below to add a file system.



Step 7: Review Instance Launch

▾ AMI Details — Edit AMI

Deep Learning AMI (Ubuntu 18.04) Version 54.0 - ami-0476bba883df7cca6

MXNet-1.8.0 & 1.7.0, TensorFlow-2.4.3, 2.3.4 & 1.15.5, PyTorch-1.7.1 & 1.8.1, Neuron, & others. NVIDIA CUDA, cuDNN, NCCL, Intel MKL-DNN, Docker, NVIDIA-Docker & EFA support. For fully managed experience, check: https://aws.amazon.com/sagemaker

Root Device Type: ebs   Virtualization type: hvm

▾ Instance Type — Edit instance type

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
p2.xlarge	-	4	61	EBS only	Yes	High

▾ Security Groups — Edit security groups

Security group name launch-wizard-20
Description launch-wizard-20 created 2022-02-08T19:31:08.619-05:00

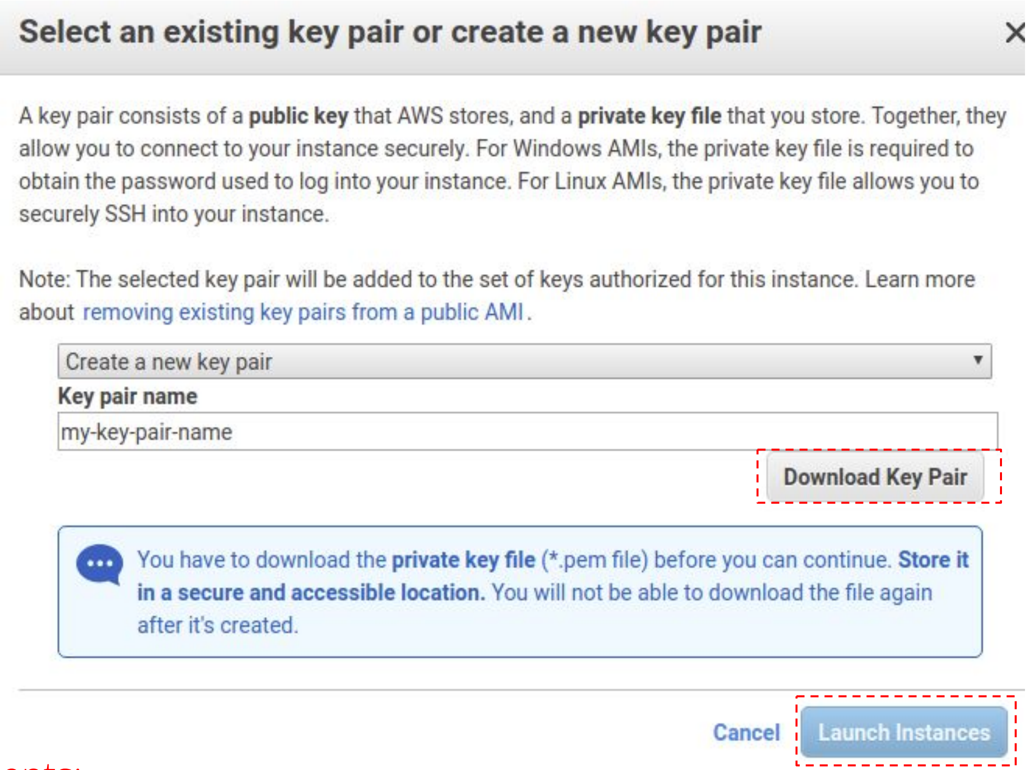
Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ	Description ⓘ
SSH	TCP	22	0.0.0.0/0	

▸ Instance Details — Edit instance details

▸ Storage — Edit storage

▸ Tags — Edit tags

Just click Launch
Do NOT change any settings



Select an existing key pair or create a new key pair

A key pair consists of a **public key** that AWS stores, and a **private key file** that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about removing existing key pairs from a public AMI .

Create a new key pair

Key pair name

my-key-pair-name

You have to download the **private key file** (*.pem file) before you can continue. **Store it in a secure and accessible location.** You will not be able to download the file again after it's created.

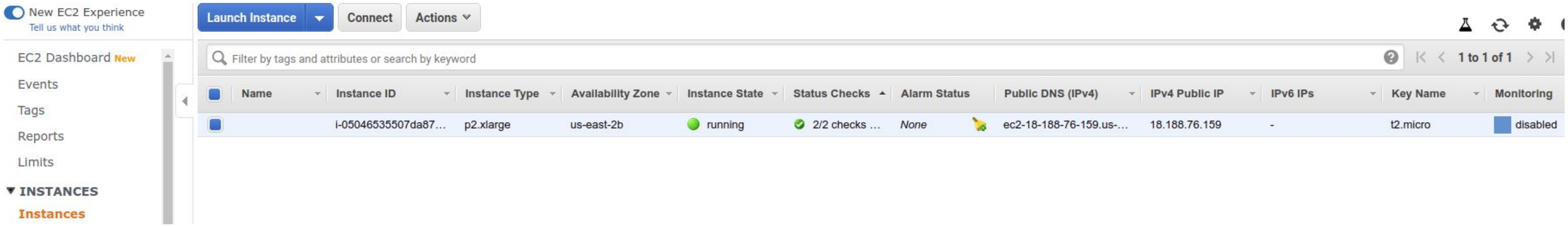
1. Name your key pair
2. Download key pair
3. Launch

Comments:

1. Make sure that you store key pair file safely and privately. You won’t be able to log into an instance if its key pair is lost.
2. You may choose to reuse an existing key pair when you launch a second instance. Make sure you can still find the key pair.

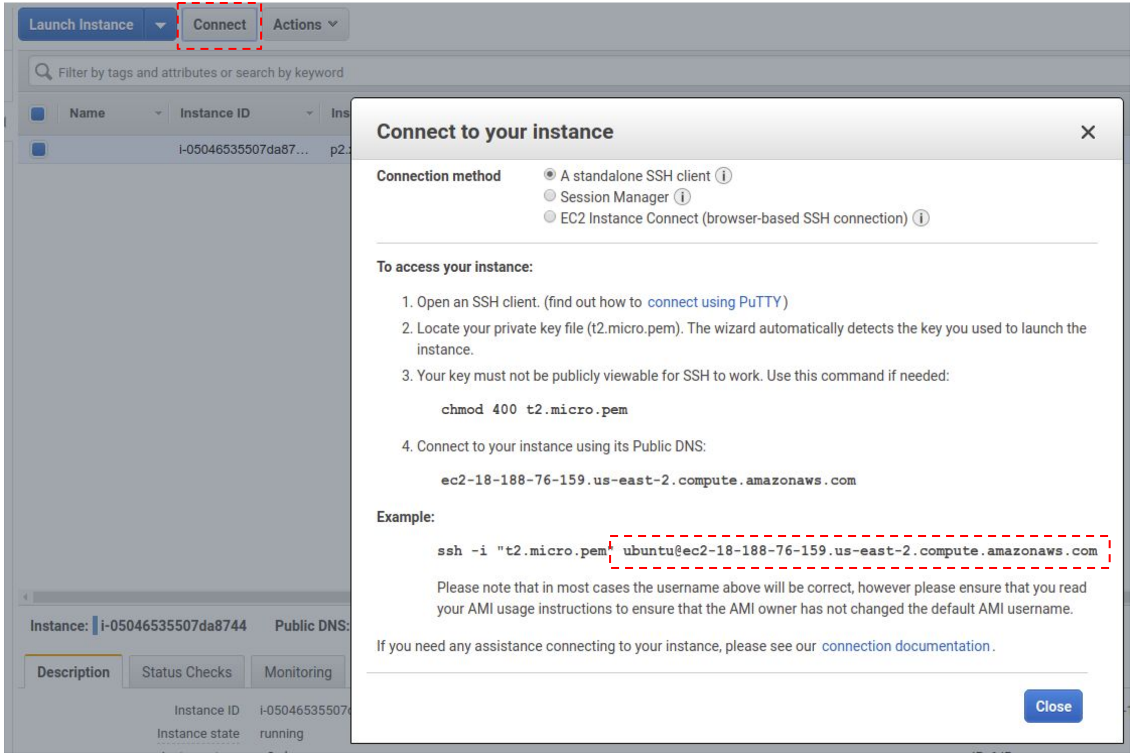
Carnegie Mellon University

Check your instance status



■	Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)	IPv4 Public IP	IPv6 IPs	Key Name	Monitoring
■		i-05046535507da87…	p2.xlarge	us-east-2b	● running	✓ 2/2 checks …	*None*	ec2-18-188-76-159.us-…	18.188.76.159	-	t2.micro	disabled

Instructions for login



Connect to your instance

Connection method
- ● A standalone SSH client
- ○ Session Manager
- ○ EC2 Instance Connect (browser-based SSH connection)

To access your instance:

1. Open an SSH client. (find out how to connect using PuTTY)
2. Locate your private key file (t2.micro.pem). The wizard automatically detects the key you used to launch the instance.
3. Your key must not be publicly viewable for SSH to work. Use this command if needed:

```
chmod 400 t2.micro.pem
```

4. Connect to your instance using its Public DNS:

```
ec2-18-188-76-159.us-east-2.compute.amazonaws.com
```

Example:

```
ssh -i "t2.micro.pem" ubuntu@ec2-18-188-76-159.us-east-2.compute.amazonaws.com
```

Please note that in most cases the username above will be correct, however please ensure that you read your AMI usage instructions to ensure that the AMI owner has not changed the default AMI username.

If you need any assistance connecting to your instance, please see our connection documentation.

Close

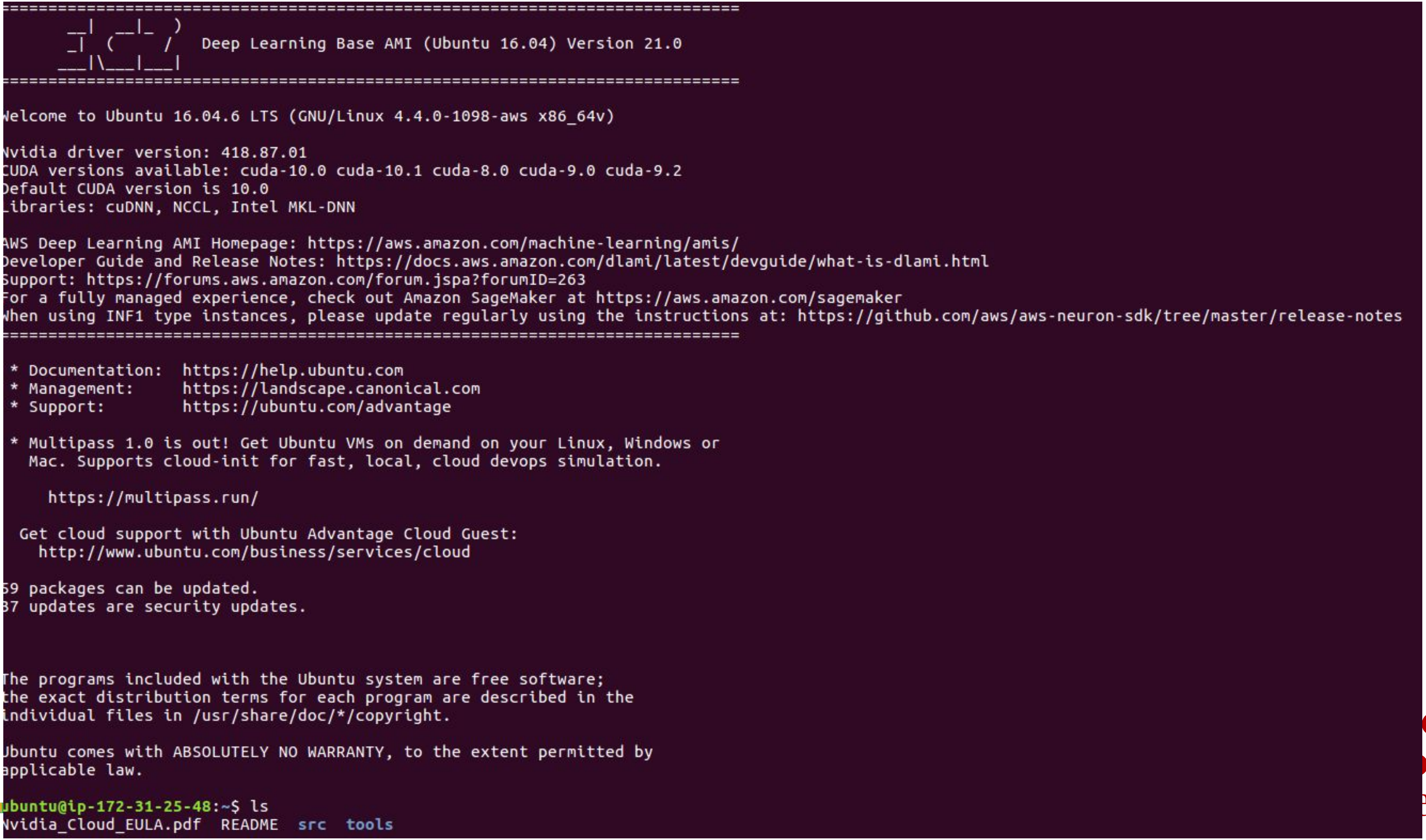
Login via SSH

Open a terminal on Mac OS /Linux, or similar alternatives on Windows
Change directory (cd) to where you store the key pair;
Change the key pair to be only readable by yourself

```
chmod 0400 <path_to_key_file>
```

After that, log in via ssh. Check the username and public ip from the previous slide

```
ssh -i <path_to_key_file> <username>@<public_ip_from_EC2>
```



```
=============================================================================
       __|  __|_  )
       _|  (     /   Deep Learning Base AMI (Ubuntu 16.04) Version 21.0
      ___|\___|___|
=============================================================================

Welcome to Ubuntu 16.04.6 LTS (GNU/Linux 4.4.0-1098-aws x86_64v)

Nvidia driver version: 418.87.01
CUDA versions available: cuda-10.0 cuda-10.1 cuda-8.0 cuda-9.0 cuda-9.2
Default CUDA version is 10.0
Libraries: cuDNN, NCCL, Intel MKL-DNN

AWS Deep Learning AMI Homepage: https://aws.amazon.com/machine-learning/amis/
Developer Guide and Release Notes: https://docs.aws.amazon.com/dlami/latest/devguide/what-is-dlami.html
Support: https://forums.aws.amazon.com/forum.jspa?forumID=263
For a fully managed experience, check out Amazon SageMaker at https://aws.amazon.com/sagemaker
When using INF1 type instances, please update regularly using the instructions at: https://github.com/aws/aws-neuron-sdk/tree/master/release-notes
=============================================================================

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Multipass 1.0 is out! Get Ubuntu VMs on demand on your Linux, Windows or
   Mac. Supports cloud-init for fast, local, cloud devops simulation.

     https://multipass.run/

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

59 packages can be updated.
37 updates are security updates.



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

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

ubuntu@ip-172-31-25-48:~$ ls
Nvidia_Cloud_EULA.pdf  README  src  tools
```

Configure Anaconda Environment

```
$ wget https://repo.continuum.io/archive/Anaconda3-2019.10-Linux-x86_64.sh
$ bash Anaconda3-2019.10-Linux-x86_64.sh
```

Agree to terms and conditions; yes to everything (including conda init); use default installation path.

```
$ source ~/.bashrc
$ conda create -n pytorch_p36 python=3.6
```

Create an environment called pytorch_p36, with the Python 3.6 interpreter.

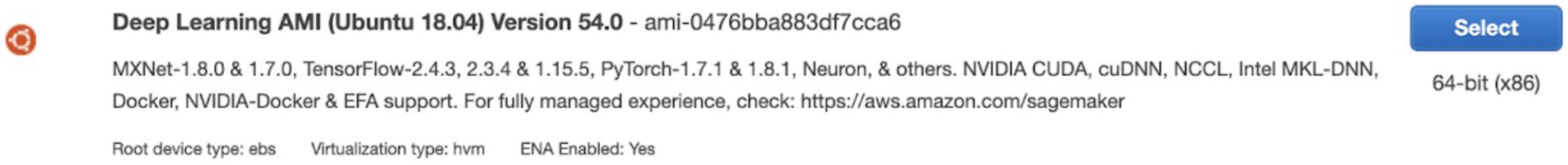
```
$ conda activate pytorch_p36
$ conda install pytorch=1.3.0 torchvision cudatoolkit=10.1 -c pytorch
```

Now you should be able to use PyTorch on your instance.

Carnegie Mellon University

Amazon Machine Images (AMI)

- A template of pre-installed system softwares and libraries
  - For deep learning: Linux OS, Nvidia GPU driver, CUDA, cuDNN, etc.
- When an instance is launched, a root storage is created with a copy of AMI.
- **Deep Learning AMI - we prefer base AMI w/ Ubuntu 18**



Deep Learning AMI (Ubuntu 18.04) Version 54.0 - ami-0476bba883df7cca6

MXNet-1.8.0 & 1.7.0, TensorFlow-2.4.3, 2.3.4 & 1.15.5, PyTorch-1.7.1 & 1.8.1, Neuron, & others. NVIDIA CUDA, cuDNN, NCCL, Intel MKL-DNN, Docker, NVIDIA-Docker & EFA support. For fully managed experience, check: https://aws.amazon.com/sagemaker

Root device type: ebs   Virtualization type: hvm   ENA Enabled: Yes

64-bit (x86)

Storage - Amazon Elastic Block Store (EBS)

- General Purpose SSD (gp2)
  - **$0.1 per GB-month**
  - 30 GB-month per month of free usage for the first 12 months
- Incurs charges as long as the EBS is **active**
  - ”Active” means that it is allocated to your account, until it is deleted.
  - Even if it stores no data
  - Even if it is not attached to your instance

Root Storage vs. Additional Storage

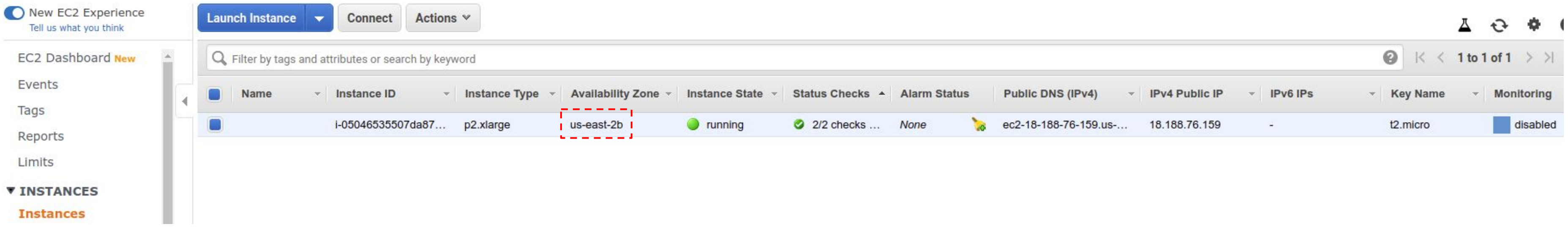
- **Root storage**
- Automatically created when an instance is launched
- Stores OS, libraries
- Automatically deleted when an instance is terminated
- Instance cannot boot when root storage is detached
- 50GB for the deep learning AMI

- **Additional storage**
- Need to create manually
- **Stores your code, model checkpoints & dataset**
- Need to delete manually
- Can be detached from or attached to different instances
- 20 GB for the 1st assignment
- Must be in the same availability zone as the instance

Python/Pytorch Development on AWS

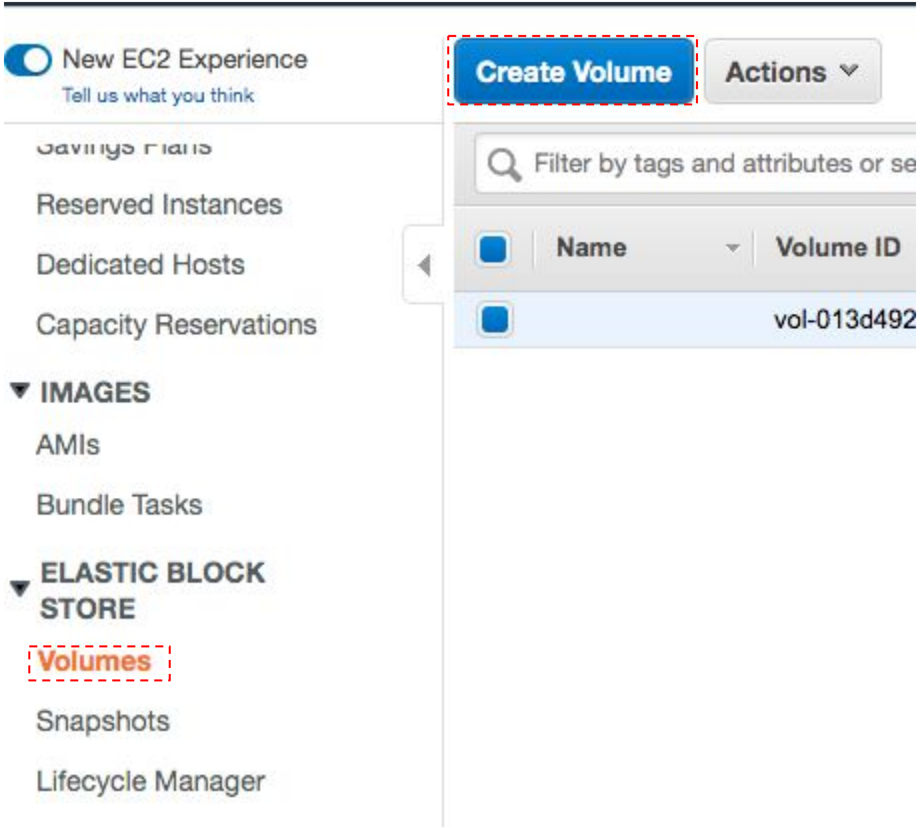
- Deep Learning Base AMI provides almost everything except Python
- Anaconda enables easy Python library management
- Create an environment with Python 3.6
- Alternative: virtualenv & pip
    - we don’t provide instruction
- Use IDE like VSCode to write code w/ a nice UI (some conda integration)

Create additional storage (Optional)

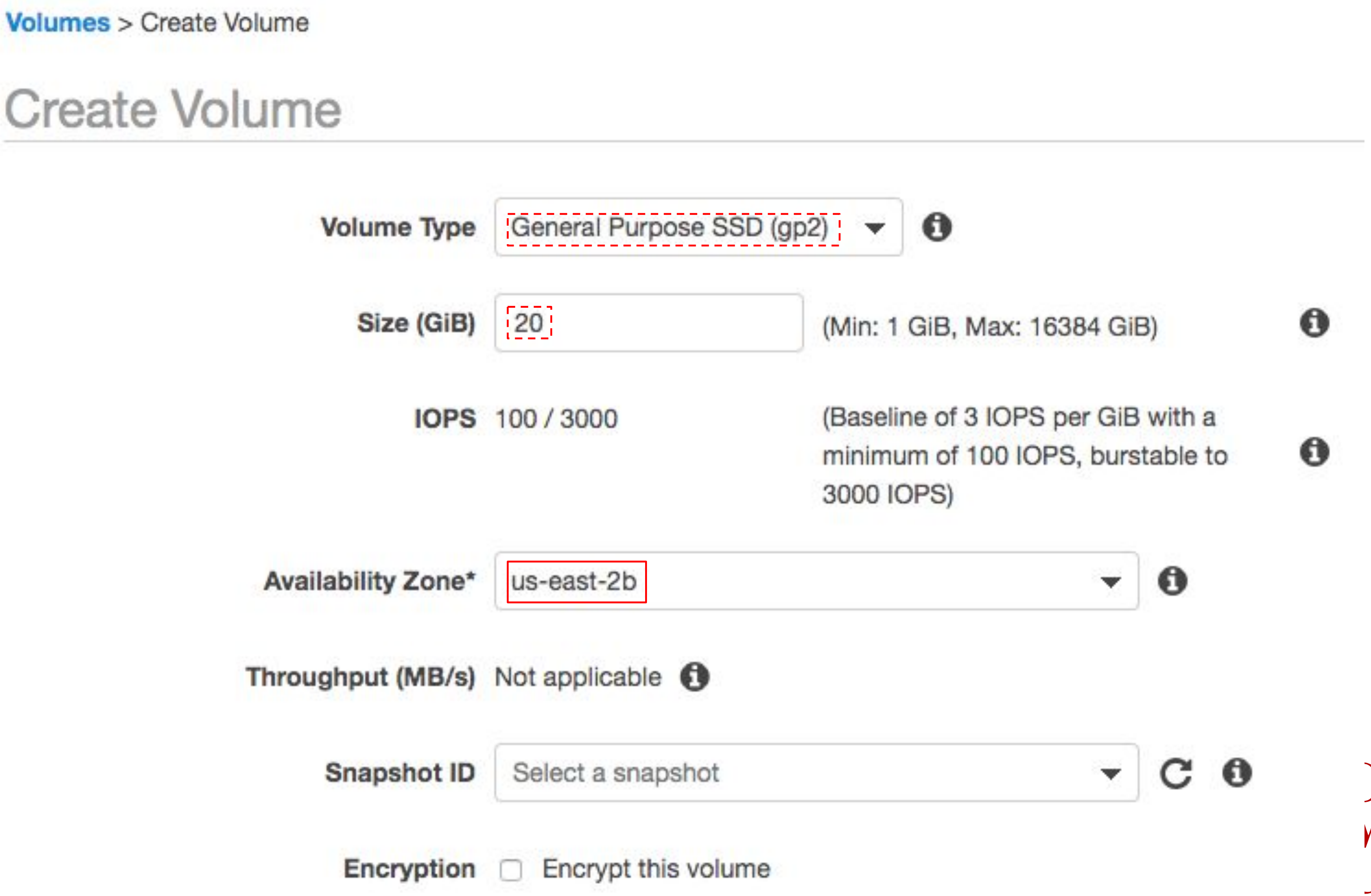


■	Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)	IPv4 Public IP	IPv6 IPs	Key Name	Monitoring
■		i-05046535507da87…	p2.xlarge	us-east-2b	● running	✓ 2/2 checks …	None	ec2-18-188-76-159.us-…	18.188.76.159	-	t2.micro	disabled

Important: check the availability zone of your instance; your additional storage must be in the same zone.



New EC2 Experience
Tell us what you think
Savings Plans
Reserved Instances
Dedicated Hosts
Capacity Reservations
▼ IMAGES
AMIs
Bundle Tasks
▼ ELASTIC BLOCK STORE
Volumes
Snapshots
Lifecycle Manager
Create Volume
Actions
Filter by tags and attributes or se
Name
Volume ID
vol-013d492



Volumes > Create Volume

Create Volume

Volume Type: General Purpose SSD (gp2)

Size (GiB): 20 (Min: 1 GiB, Max: 16384 GiB)

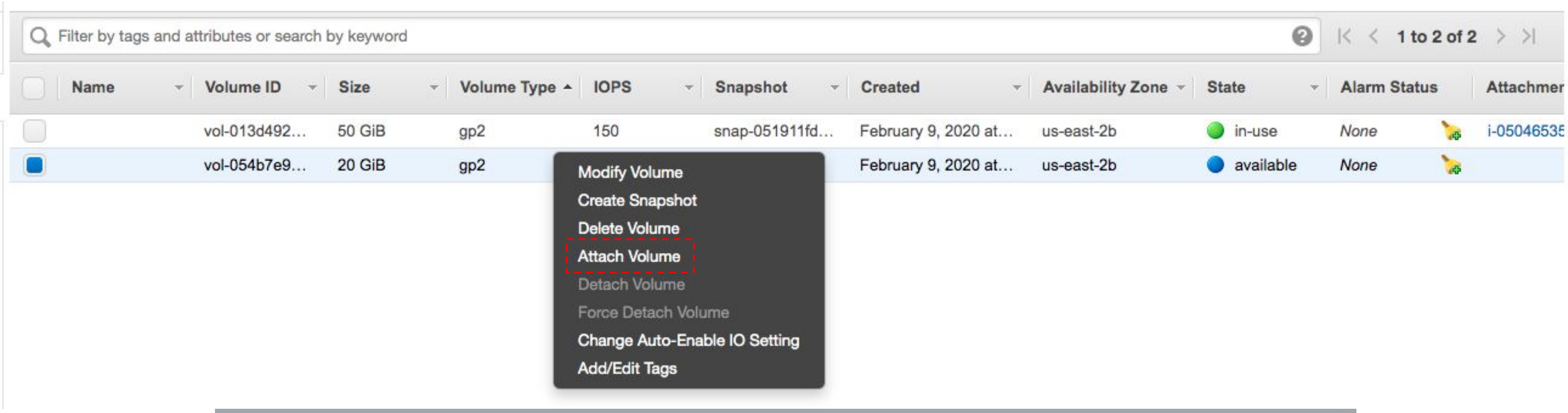
IOPS: 100 / 3000 (Baseline of 3 IOPS per GiB with a minimum of 100 IOPS, burstable to 3000 IOPS)

Availability Zone*: us-east-2b

Throughput (MB/s): Not applicable

Snapshot ID: Select a snapshot

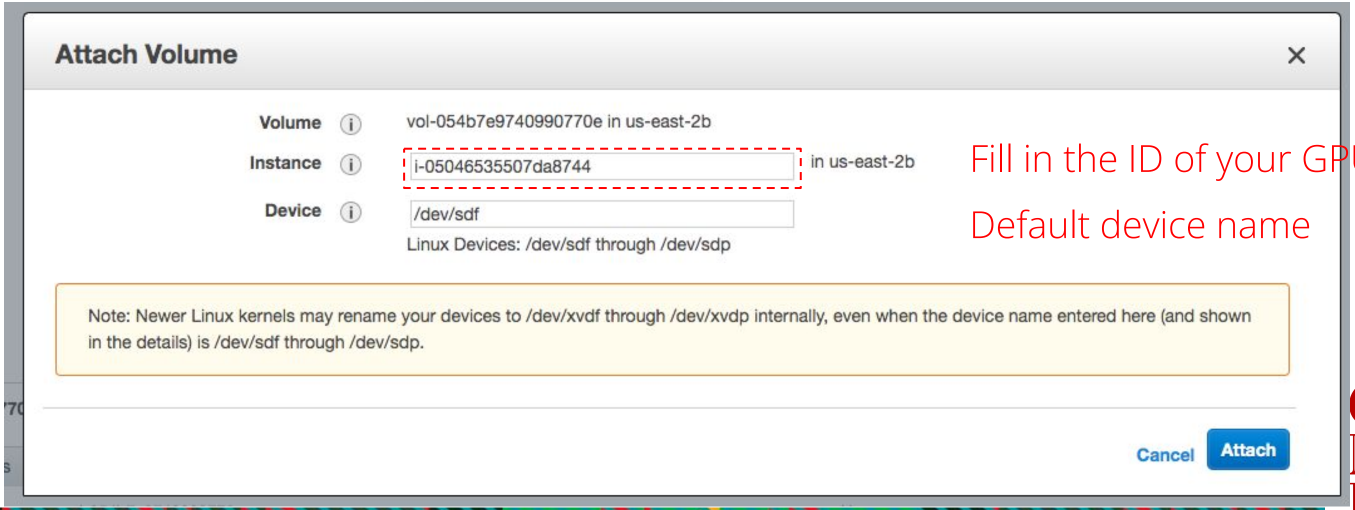
Encryption: □ Encrypt this volume



Filter by tags and attributes or search by keyword

1 to 2 of 2

□	Name	Volume ID	Size	Volume Type	IOPS	Snapshot	Created	Availability Zone	State	Alarm Status	Attachmer
□		vol-013d492…	50 GiB	gp2	150	snap-051911fd…	February 9, 2020 at…	us-east-2b	● in-use	None	i-05046535
■		vol-054b7e9…	20 GiB	gp2			February 9, 2020 at…	us-east-2b	● available	None	



Attach Volume

Volume: vol-054b7e9740990770e in us-east-2b

Instance: i-05046535507da8744 in us-east-2b

Device: /dev/sdf

Linux Devices: /dev/sdf through /dev/sdp

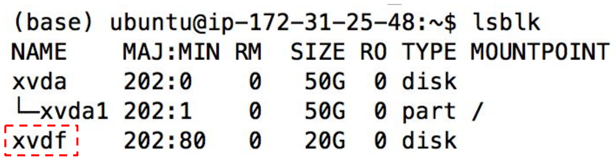
Note: Newer Linux kernels may rename your devices to /dev/xvdf through /dev/xvdp internally, even when the device name entered here (and shown in the details) is /dev/sdf through /dev/sdp.

Fill in the ID of your GPU instance

Default device name

Initialize & mount additional storage

- `lsblk`
  - show the attached devices (find the name of the attached EBS volume)
- `sudo file -s /dev/<device_name>`
  - show whether the device has a filesystem
- `sudo mkfs -t ext4 /dev/<device_name>`
  - `e.g. sudo mkfs -t ext4 /dev/xvdf`
  - create the ext4 file system on the device (do this only if the previous command returned “data”)
- `sudo mkdir <mount_point>`
  - `e.g. sudo mkdir /mnt/data`
  - make a folder for mounting the volume
- `sudo mount <device_name> <mount_point>`
  - `e.g. sudo mount /dev/xvdf /mnt/data`
  - mount the volume
- `sudo chmod 777 -R <mount_point>`
  - add permissions to access the folder

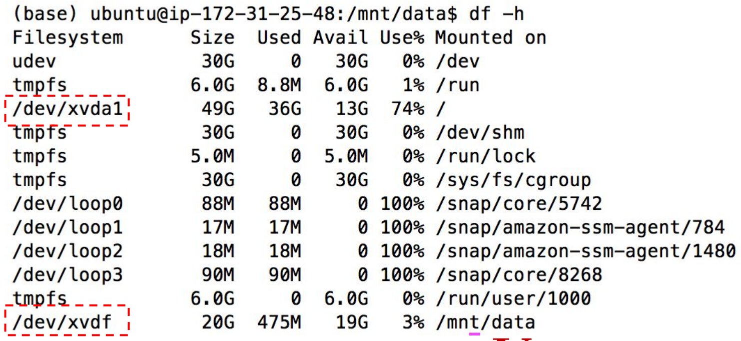


```
(base) ubuntu@ip-172-31-25-48:~$ lsblk
NAME    MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
xvda    202:0    0   50G  0 disk
└─xvda1 202:1    0   50G  0 part /
xvdf    202:80   0   20G  0 disk
```

/dev/sdf in previous slide renamed to /dev/xvdf

Test additional storage

- `cd <mount point>`
  - go to the mount point of the additional storage, e.g., /mnt/data
- `wget http://host.robots.ox.ac.uk/pascal/VOC/voc2007/VOCtest_06-Nov-2007.tar`
  - download the dataset required for the 1st assignment
  - also store your code and model checkpoints on additional EBS storage
- `df -h`
  - show the usage on the disks
  - monitor root storage and additional storage

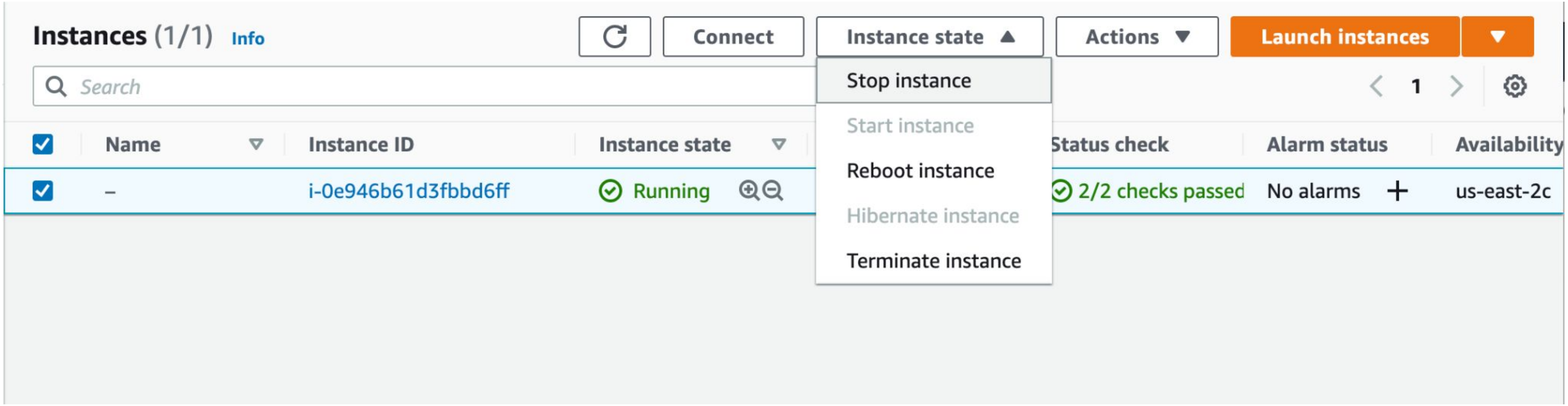


root storage

additional storage

```
(base) ubuntu@ip-172-31-25-48:/mnt/data$ df -h
Filesystem      Size  Used Avail Use% Mounted on
udev             30G     0   30G   0% /dev
tmpfs           6.0G  8.8M  6.0G   1% /run
/dev/xvda1       49G   36G   13G  74% /
tmpfs            30G     0   30G   0% /dev/shm
tmpfs           5.0M     0  5.0M   0% /run/lock
tmpfs            30G     0   30G   0% /sys/fs/cgroup
/dev/loop0       88M   88M     0 100% /snap/core/5742
/dev/loop1       17M   17M     0 100% /snap/amazon-ssm-agent/784
/dev/loop2       18M   18M     0 100% /snap/amazon-ssm-agent/1480
/dev/loop3       90M   90M     0 100% /snap/core/8268
tmpfs           6.0G     0  6.0G   0% /run/user/1000
/dev/xvdf        20G  475M   19G   3% /mnt/data
```

Final Word: Stop your instance! For now..



☑	Name	Instance ID	Instance state	Status check	Alarm status	Availability
☑	–	i-0e946b61d3fbbd6ff	✓ Running	✓ 2/2 checks passed	No alarms	us-east-2c

Google Cloud Platform

Same Idea

- structure
- billing strategy
- increase quotas
- create instance (pay attention to many details).

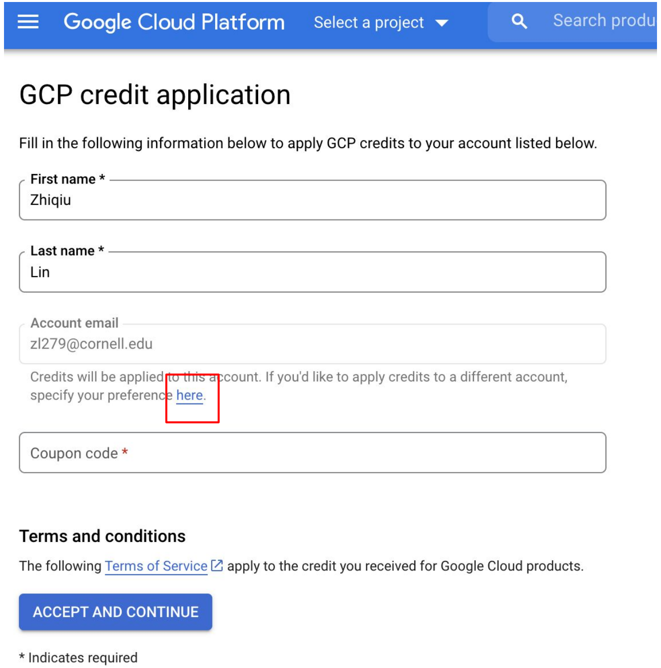
GCP Account Setup

- **Use personal email address**
  - Do not use andrew account (Will likely face organization error)
- https://cloud.google.com/
- Redeem credits at: https://console.cloud.google.com/education

Google Cloud Platform

- Redeem coupon to your personal account
  -

https://console.cloud.google.com/education



GCP credit application

Fill in the following information below to apply GCP credits to your account listed below.

First name *
Zhiqiu

Last name *
Lin

Account email
zl279@cornell.edu

Credits will be applied to this account. If you'd like to apply credits to a different account, specify your preference here.

Coupon code *

Terms and conditions

The following Terms of Service apply to the credit you received for Google Cloud products.

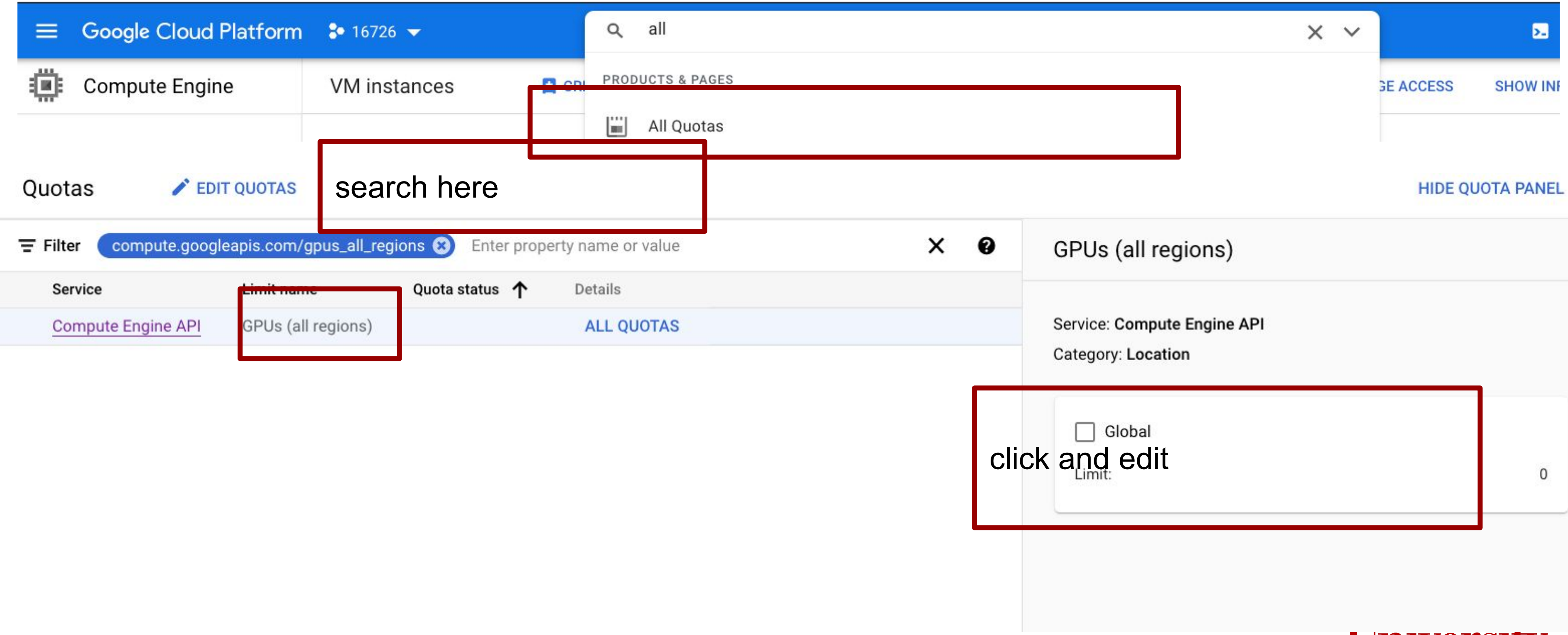
ACCEPT AND CONTINUE

* Indicates required

Google Cloud Platform (Basic steps)

- Upgrade to paid account (not free account)
- Enable “Compute Engine API”
- Create a project
- **Increase your quota for “gpus_all_region”**
- Create “VM Instance”
- Select the zone you are residing in

Increase Quotas (otherwise you cannot use all types of GPUs)



all

search here

Service	Limit name	Quota status	Details
Compute Engine API	GPUs (all regions)		ALL QUOTAS

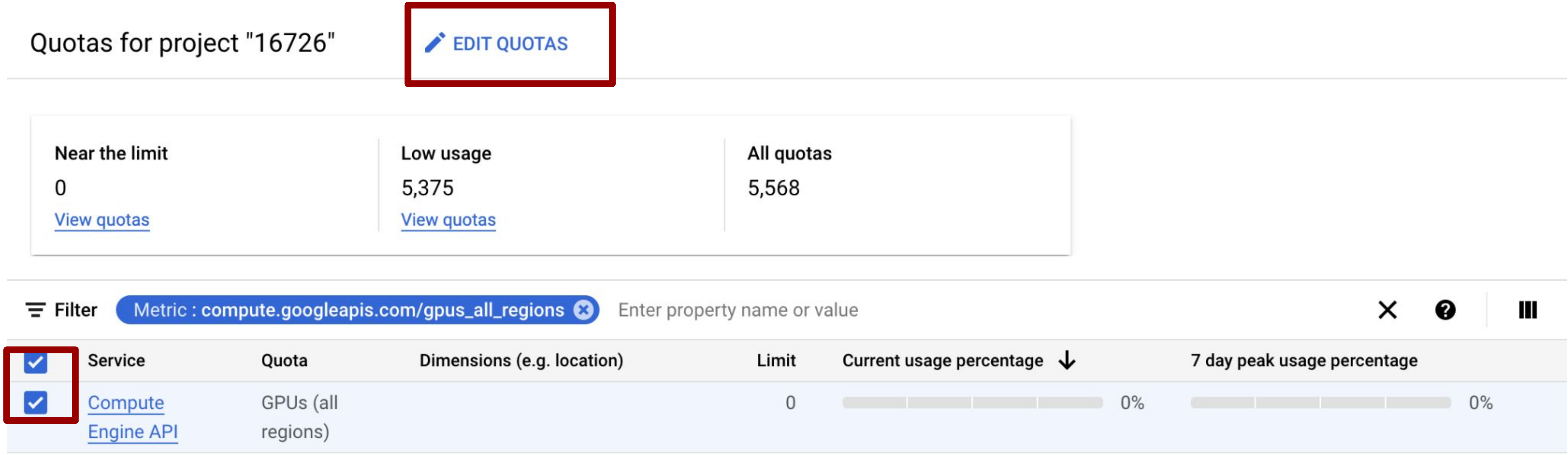
Service: **Compute Engine API**

Category: **Location**

click and edit

Limit: 0

Increase Quotas (otherwise you cannot use all types of GPUs)



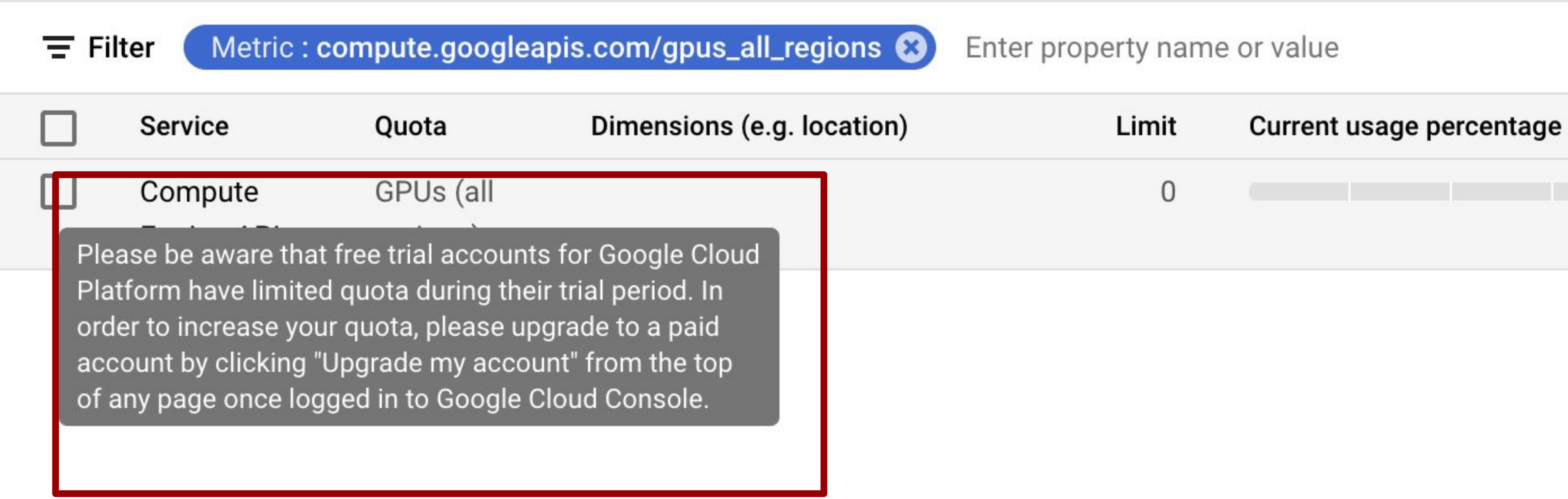
Quotas for project "16726"

Near the limit	Low usage	All quotas
0	5,375	5,568
View quotas	View quotas	

Filter Metric : compute.googleapis.com/gpus_all_regions Enter property name or value

☑	Service	Quota	Dimensions (e.g. location)	Limit	Current usage percentage ↓	7 day peak usage percentage
☑	Compute Engine API	GPUs (all regions)		0	0%	0%

You must have a paid account before increasing Quotas (otherwise you cannot use all types of GPUs)

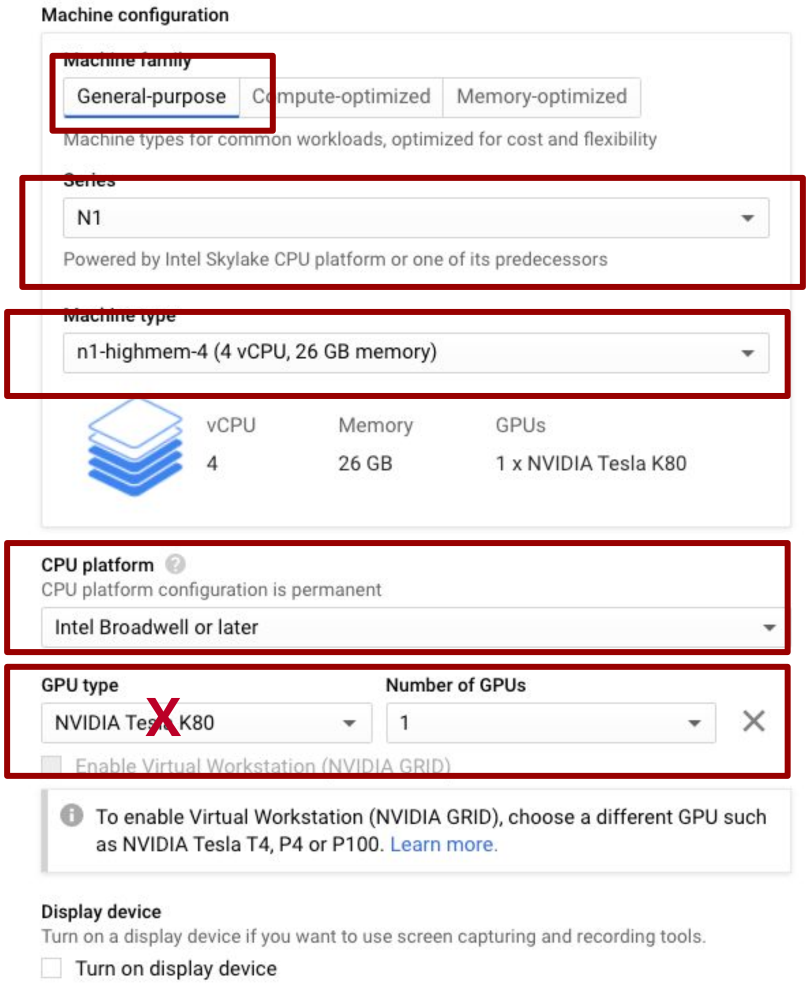


Filter Metric : compute.googleapis.com/gpus_all_regions

Enter property name or value

☐	Service	Quota	Dimensions (e.g. location)	Limit	Current usage percentage
☐	Compute	GPUs (all		0	

Please be aware that free trial accounts for Google Cloud Platform have limited quota during their trial period. In order to increase your quota, please upgrade to a paid account by clicking "Upgrade my account" from the top of any page once logged in to Google Cloud Console.



Machine configuration

Machine family

General-purpose | Compute-optimized | Memory-optimized

Machine types for common workloads, optimized for cost and flexibility

Series

N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type

n1-highmem-4 (4 vCPU, 26 GB memory)

vCPU	Memory	GPUs
4	26 GB	1 x NVIDIA Tesla K80

CPU platform

CPU platform configuration is permanent

Intel Broadwell or later

GPU type	Number of GPUs
NVIDIA Tesla K80 X	1

□ 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.

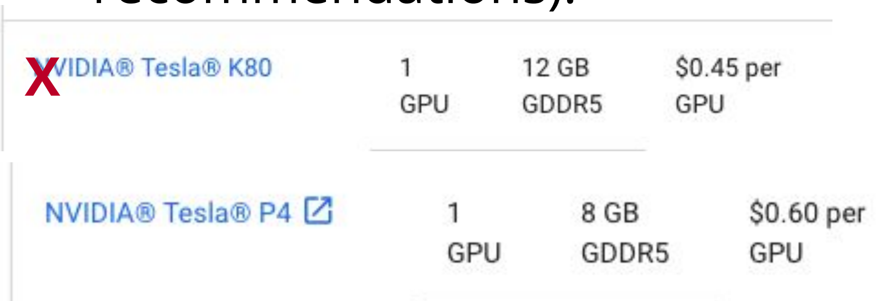
Display device

Turn on a display device if you want to use screen capturing and recording tools.

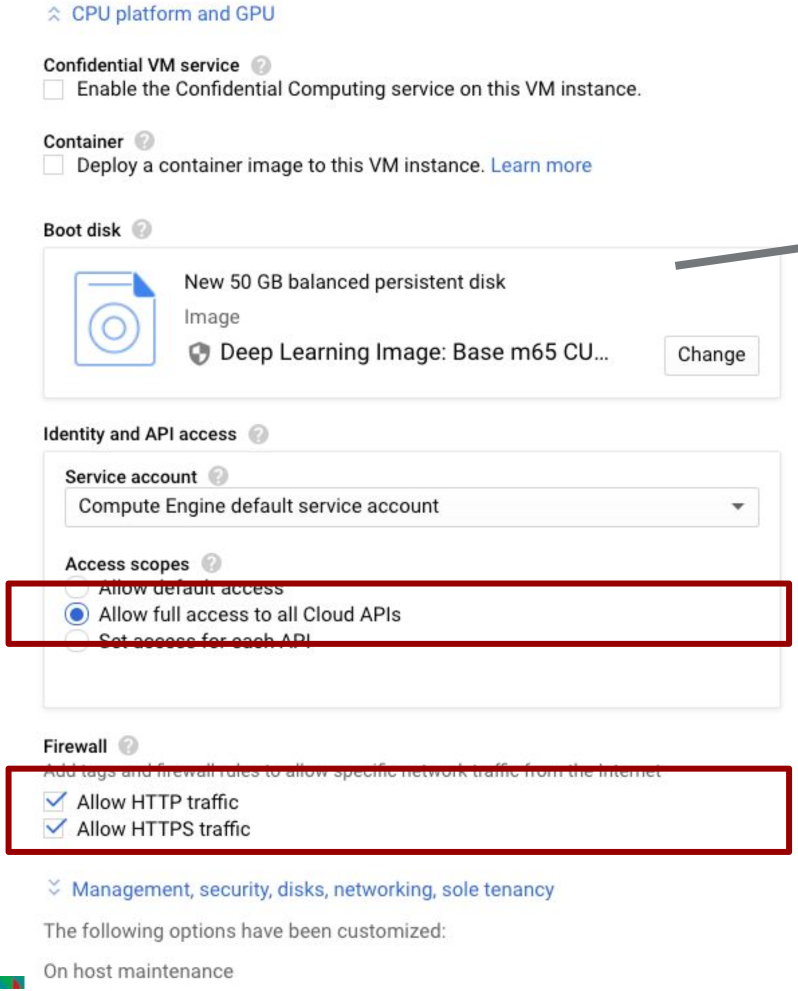
□ Turn on display device

Now you can proceed to:

- Compute Engine -> VM instances -> Create Instance
- Only Tesla P4 is available (as of 2022.02)
- You can use your preferred GPUs (we won't be making recommendations).



X NVIDIA® Tesla® K80	1 GPU	12 GB GDDR5	$0.45 per GPU
NVIDIA® Tesla® P4	1 GPU	8 GB GDDR5	$0.60 per GPU



CPU platform and GPU

Confidential VM service

□ Enable the Confidential Computing service on this VM instance.

Container

□ Deploy a container image to this VM instance. Learn more

Boot disk

New 50 GB balanced persistent disk

Image

Deep Learning Image: Base m65 CU...

Change

Identity and API access

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 Internet

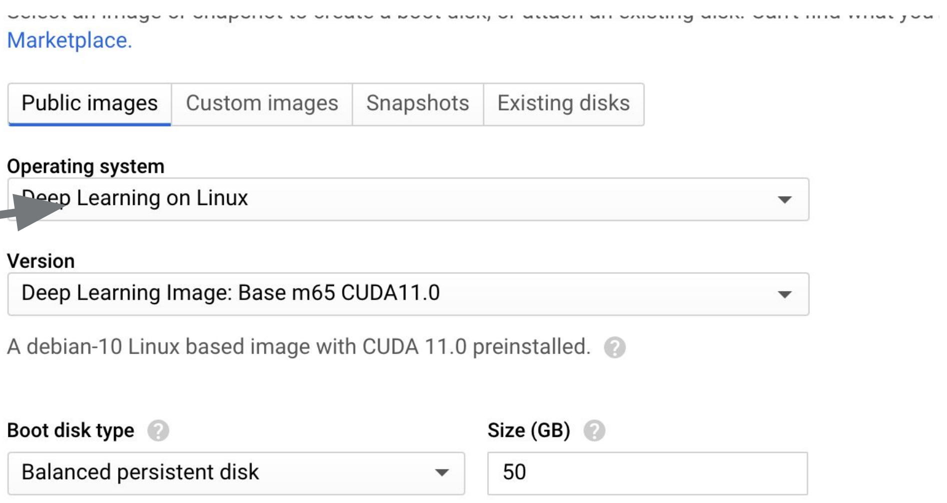
☑ Allow HTTP traffic

☑ Allow HTTPS traffic

Management, security, disks, networking, sole tenancy

The following options have been customized:

On host maintenance



Select an image or snapshot to create a boot disk, or attach an existing disk. Can't find what you

Marketplace.

Operating system

Deep Learning on Linux

Version

Deep Learning Image: Base m65 CUDA11.0

A debian-10 Linux based image with CUDA 11.0 preinstalled.

Boot disk type

Balanced persistent disk

Size (GB)

50

- Set SSH Keys: https://cloud.google.com/compute/docs/instances/adding-removing-ssh-keys#createsshkeys
- add disk: youtube
- Pricing: https://cloud.google.com/compute/gpus-pricing

Final words

- **Stop an instance if you are not working on it!**
- $150 AWS and GCP for all 3 assignments and course project
- Use AWS/GCP wisely & responsibly
- Enjoy the deep learning alchemy!