

GCP & AWS update GPU limits

CSCI 599
2019-02-12

GCP & AWS by default don't allow users to use GPU machines (GPU limits are 0 initially).

So if you want to use their GPU machine, you have to update the GPU limits.

The latter slides will show how to update the GPU limits.

AWS & GCP are not free!

GPU Instances are expensive!

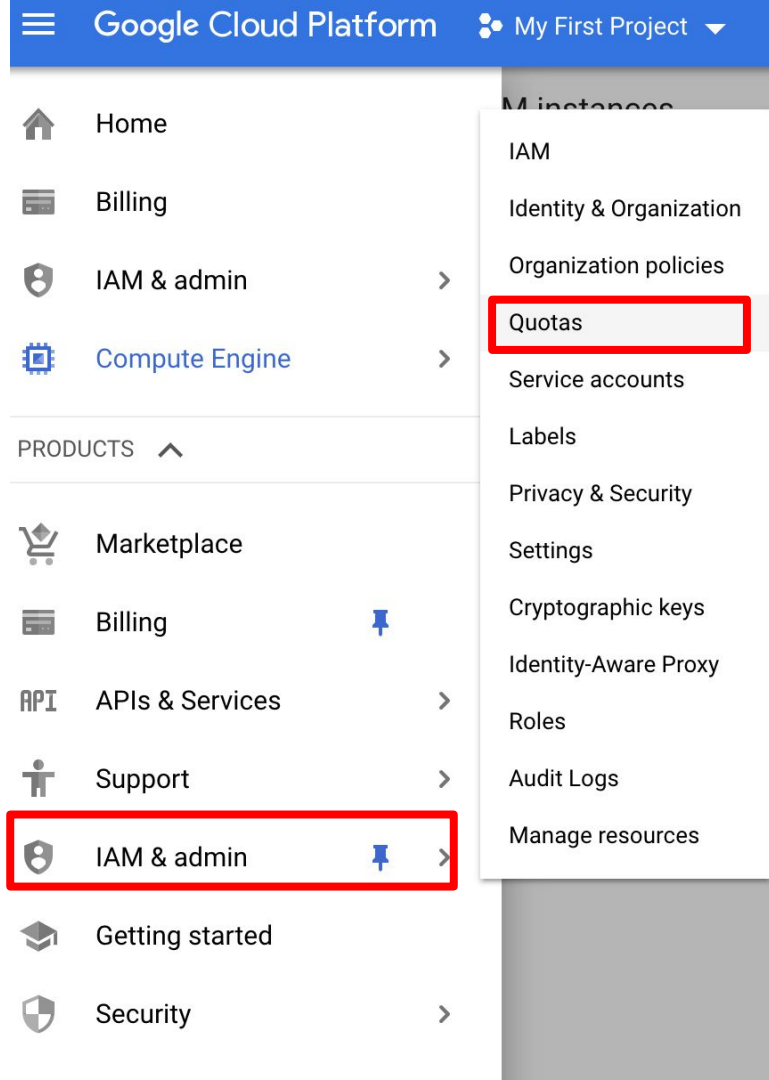
You are responsible for all the billings!

Even if you are not running processes, they will charge you if your machine is running.

Remember to shutdown/terminate your machines when not using them.

GCP: update GPU limits

- first create a project
- click top left menu icon
- hover over “IAM & admin”
- click the “Quotas”



GCP: update GPU limits

note: if your “limit” value is 1 or more, then you are fine. “0” is the default value for new users.

- Inside the Quotas page,
 - set “Metric” filter to be “GPUs (all regions)”
 - set “Location” filter to be Global
- select the “Compute Engine API”
- click “EDIT QUOTAS”

The screenshot shows the Google Cloud Platform interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'My First Project', and a search bar. The left sidebar contains navigation links for IAM & admin, IAM, Identity & Organization, Organization policies, and Quotas. The main content area is titled 'Quotas' and features a red box labeled '3' around the '+ EDIT QUOTAS' button. Below this, there are filters for 'Quota type' (set to 'All quotas'), 'Service' (set to 'All services'), 'Metric' (set to 'GPUs (all regions)' with a red box labeled '1'), and 'Location' (set to 'Global'). A 'Clear' button is also present. The table below shows a single entry for 'Service' with a red box labeled '2' around the 'Compute Engine API GPUs (all regions)' row. The table columns are 'Location' (Global), 'Current Usage' (0), '7 Day Peak Usage' (with a dropdown arrow), and 'Limit' (1). A line from the note above points to the 'Limit' column.

Service	Location	Current Usage	7 Day Peak Usage	Limit
Compute Engine API GPUs (all regions)	Global	0	— ?	1

GCP: edit quotas

- in the right panel, fill in your name/email/phone, click next

+ EDIT QUOTAS

IAM & Admin - My First Project - Google Cloud Platform

Service: All services Location: GPUs (all regions) Global

Location	Current Usage ?	7 Day Peak Usage ^	Limit
Global	0	— ?	1

✕ 1 quota selected

Edit quotas

Name

hanpeng liu

Email

lhp.polo@gmail.com

Phone ?

Phone is required.

Next

GCP: edit quotas

- fill the maximum number of GPUs you want in the “**new quota limit**”, e.g., “2”
- write something to the description, e.g., “deep learning”
- click “**submit request**”

Quota type	Service	Metric	Location
All quotas	All services	GPUs (all regions)	Global

<input checked="" type="checkbox"/> Service	Location	Current Usage	7 Day Peak Usage	Limit
<input checked="" type="checkbox"/> Compute Engine API GPUs (all regions)	Global	<div><div></div></div> 0	—	1

Compute Engine API

Quota: GPUs (all regions)

New quota limit
Enter a new quota limit. Your request will be sent to your service provider for approval.

Request description
Required

Done

Cancel

Submit request

Back

GCP: edit quotas

- you will see your case is submitted and
- usually it takes several hours to get quota approved

Edit quotas

Compute Engine API

Thank you for submitting Case # (ID:500f200001MdyukAAB) to Google Cloud Platform support for the following quota:

- Change GPUs (all regions) from 1 to 2

Your request is being processed and you should receive an email confirmation for your request. Should you need further assistance, you can respond to that email.

AWS & GCP are not free!

GPU Instances are expensive!

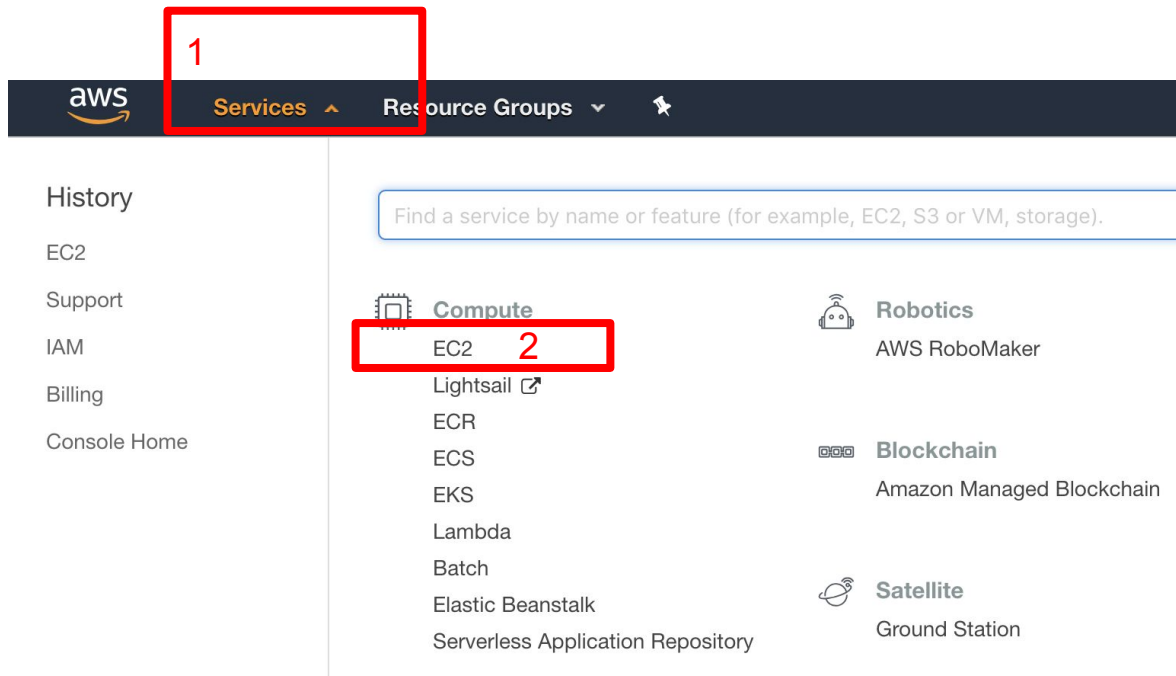
You are responsible for all the billings!

Even if you are not running processes, they will charge you if your machine is running.

Remember to shutdown/terminate your machines when not using them.

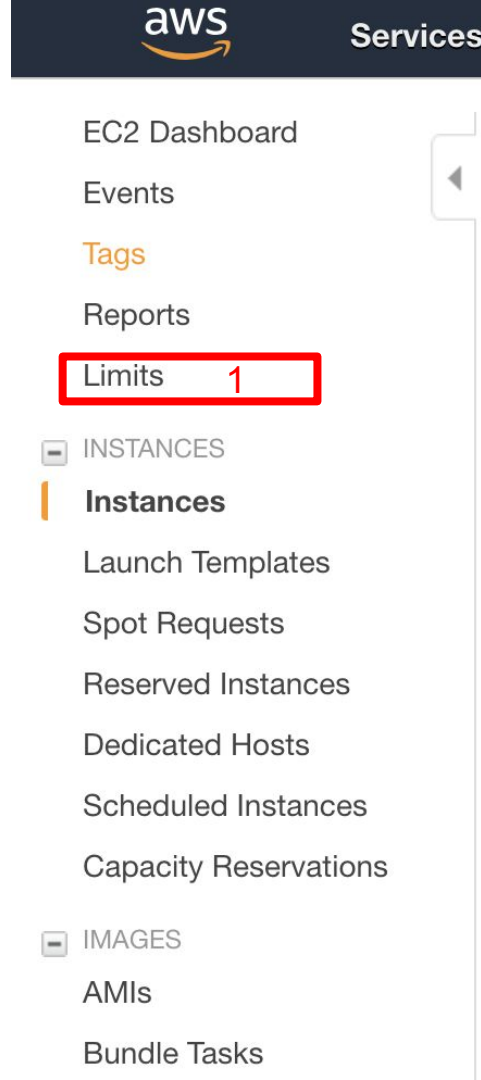
AWS: update GPU limits

- login to AWS console
- click “**Services**”
- then click “**EC2**”



AWS: update GPU limits

- click “**limits**” in the left side bar



AWS: update GPU limits

note: if your “limit” value is 1 or more, then you are fine. “0” is the default value for new users.

- search “p2” in the new page
- click “**Request limit increase**” for “Running On-Demand p2.xlarge instances”

The screenshot shows the AWS Management Console interface. The top navigation bar includes the AWS logo, 'Services', 'Resource Groups', and a search bar containing 'p2'. The left sidebar shows the 'Limits' section under 'INSTANCES'. The main content area displays a table of limits for various instance types. The row for 'Running On-Demand p2.xlarge instances' is highlighted with a red box, and the 'Request limit increase' link for that row is also highlighted with a red box. A red box also highlights the search bar containing 'p2'.

Instance Type	Limit	Action
Running On-Demand m5a.24xlarge instances	0	Request limit increase
Running On-Demand m5a.2xlarge instances	5	Request limit increase
Running On-Demand m5a.4xlarge instances	0	Request limit increase
Running On-Demand m5a.large instances	5	Request limit increase
Running On-Demand m5a.xlarge instances	5	Request limit increase
Running On-Demand m5d.12xlarge instances	0	Request limit increase
Running On-Demand m5d.24xlarge instances	0	Request limit increase
Running On-Demand m5d.2xlarge instances	1	Request limit increase
Running On-Demand m5d.4xlarge instances	0	Request limit increase
Running On-Demand m5d.large instances	5	Request limit increase
Running On-Demand m5d.xlarge instances	2	Request limit increase
Running On-Demand p2.16xlarge instances	0	Request limit increase
Running On-Demand p2.8xlarge instances	0	Request limit increase
Running On-Demand p2.xlarge instances	1	Request limit increase

AWS: update GPU limits

- in the new tab, under “requests”
 - choose the region you want, e.g. Oregon (or whatever you want)
 - choose the machine type: p2.xlarge
 - write down new limit (max GPUs you want), e.g., 1
- under “requests”, fill in the reason, e.g. “deep learning”
- click submit

Requests

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

Request 1

Region
US West (Oregon) ▼

Primary Instance Type
p2.xlarge ▼

Limit
Instance Limit ▼

New limit value
2

Case description

Use case description

deep learning

Maximum 5000 characters (4987 remaining)

► Contact options

Cancel **Submit**

AWS: update GPU limits

- you should see a new page with case ID and other information
- usually it takes several hours to get quota approved

Case ID 5781268701 [Info](#)

Resolve case

Case details

Subject

Limit Increase: EC2 Instances

Case ID

5781268701

Created

2019-02-12T22:47:21.190Z

Case type

Service limits

Status

Unassigned

Severity

General question

Category

Service Limit Increase, EC2 Instances

Additional contacts

-

AWS & GCP are not free!

GPU Instances are expensive!

You are responsible for all the billings!

Even if you are not running processes, they will charge you if your machine is running.

Remember to shutdown/terminate your machines when not using them.

AWS & GCP are not free!

GPU Instances are expensive!

You are responsible for all the billings!

Even if you are not running processes, they will charge you if your machine is running.

Remember to shutdown/terminate your machines when not using them.