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).

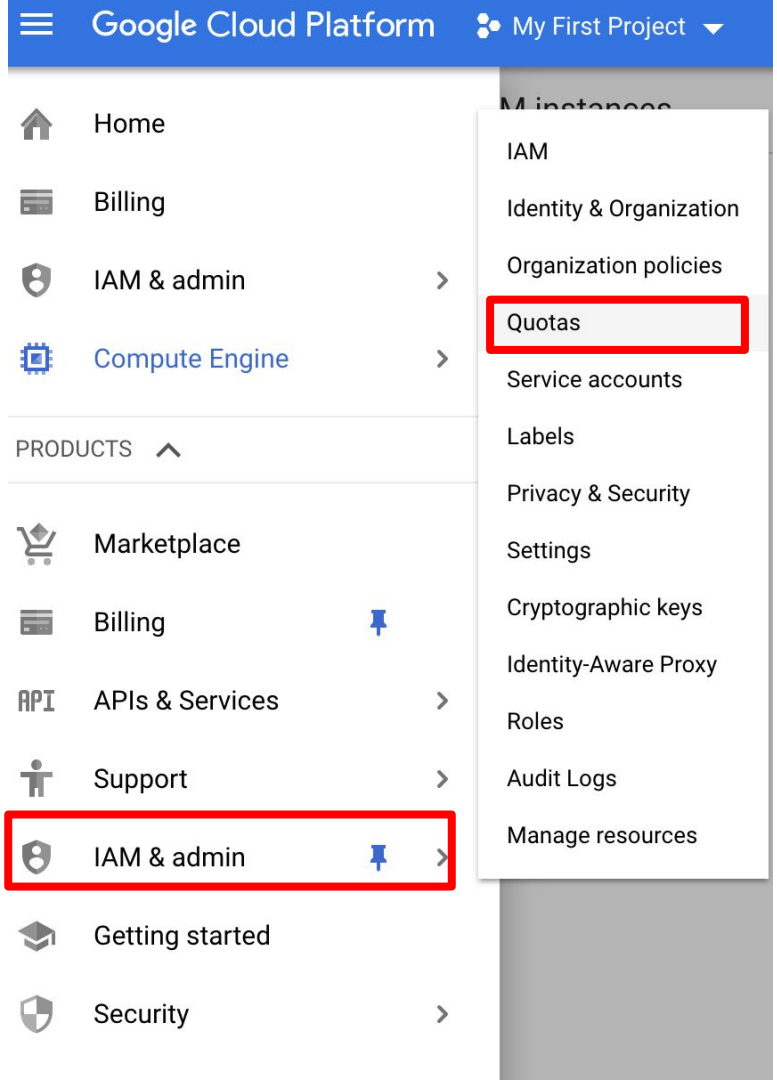
So if you want to use their GPU machine, you have to update the GPU limits.

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

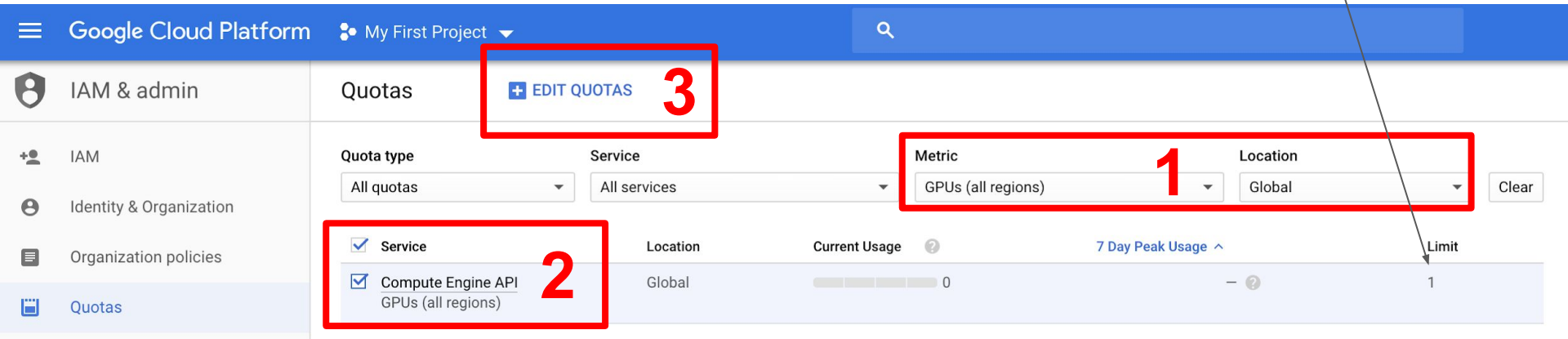
- first create a project
- click top left menu icon
- hover over “**IAM & admin**”
- click the “**Quotas**”



GCP: update GPU limits

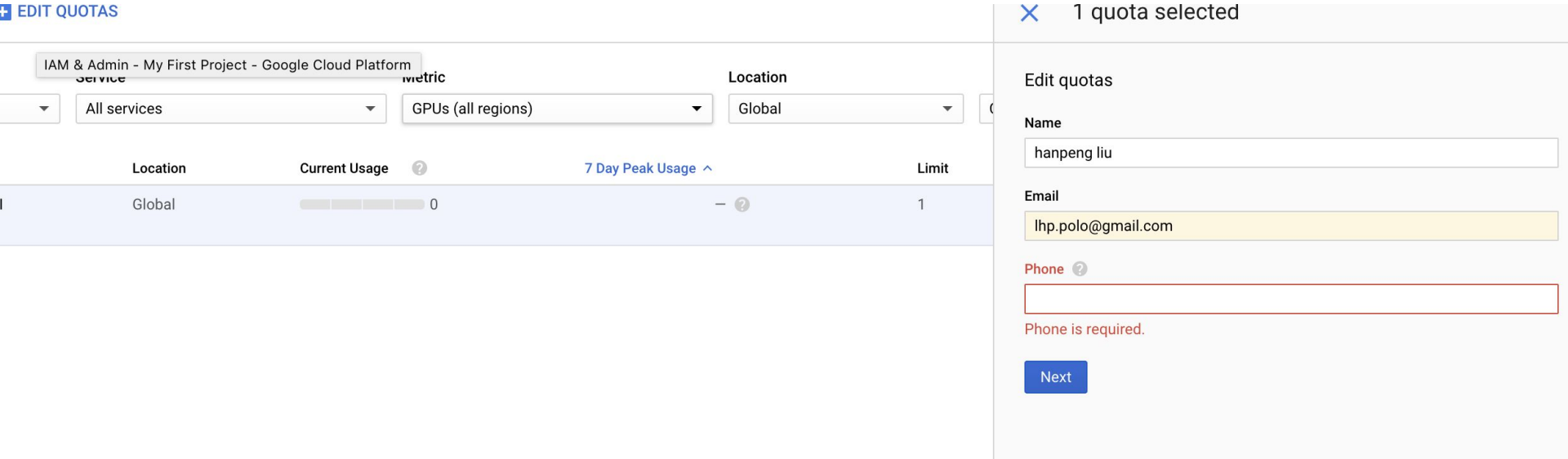
note: if your “limit” value is 1 or more, then you are fine. “0” is the default value for new users.

- 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”



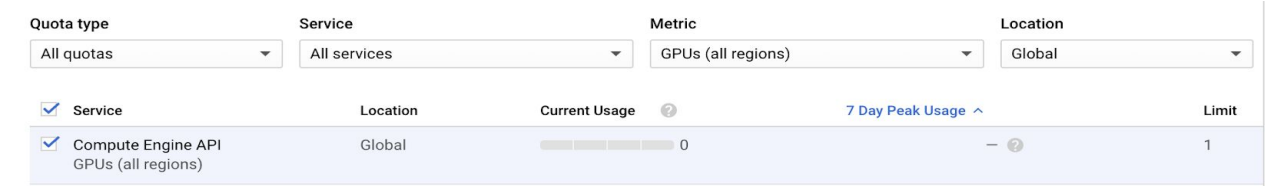
GCP: edit quotas

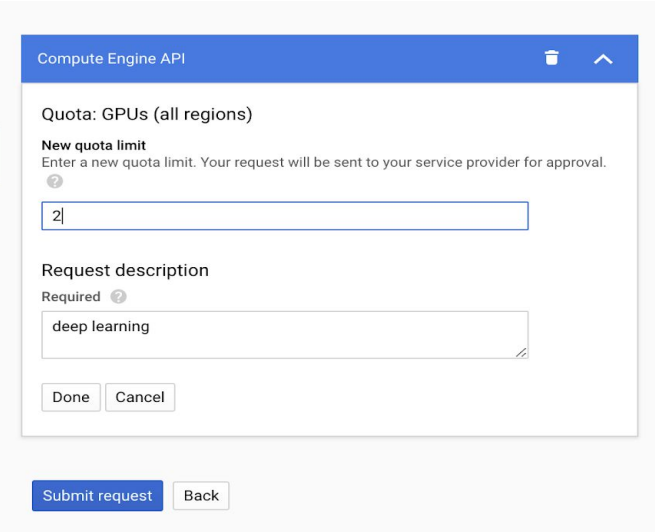
- in the right panel, fill in your name/email/phone, click next



GCP: edit quotas

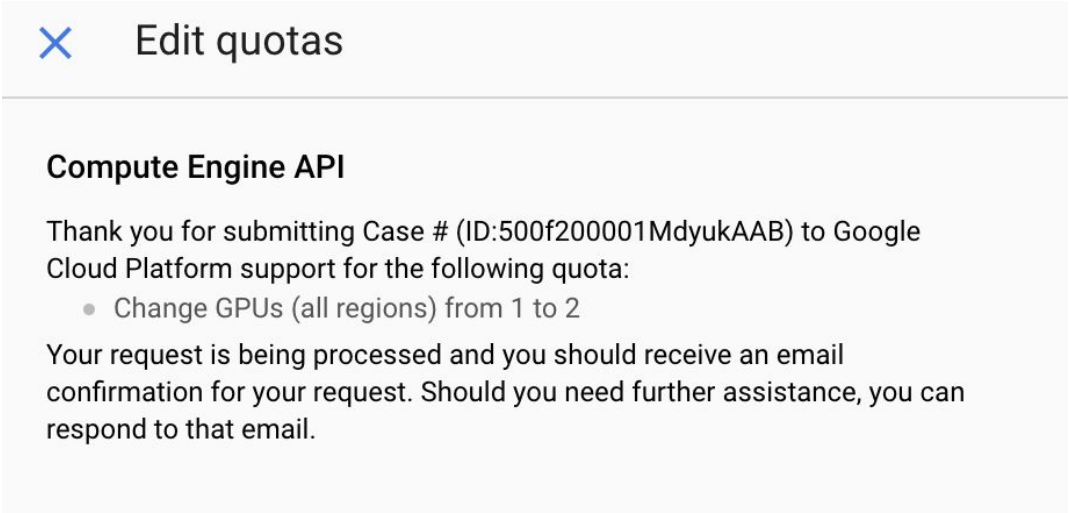
- fill the maximum number of GPUs you want in the “**new quota limit**”, e.g., “2”
- write something to the description, e.g., “deep learning”
- click “**submit request**”





GCP: edit quotas

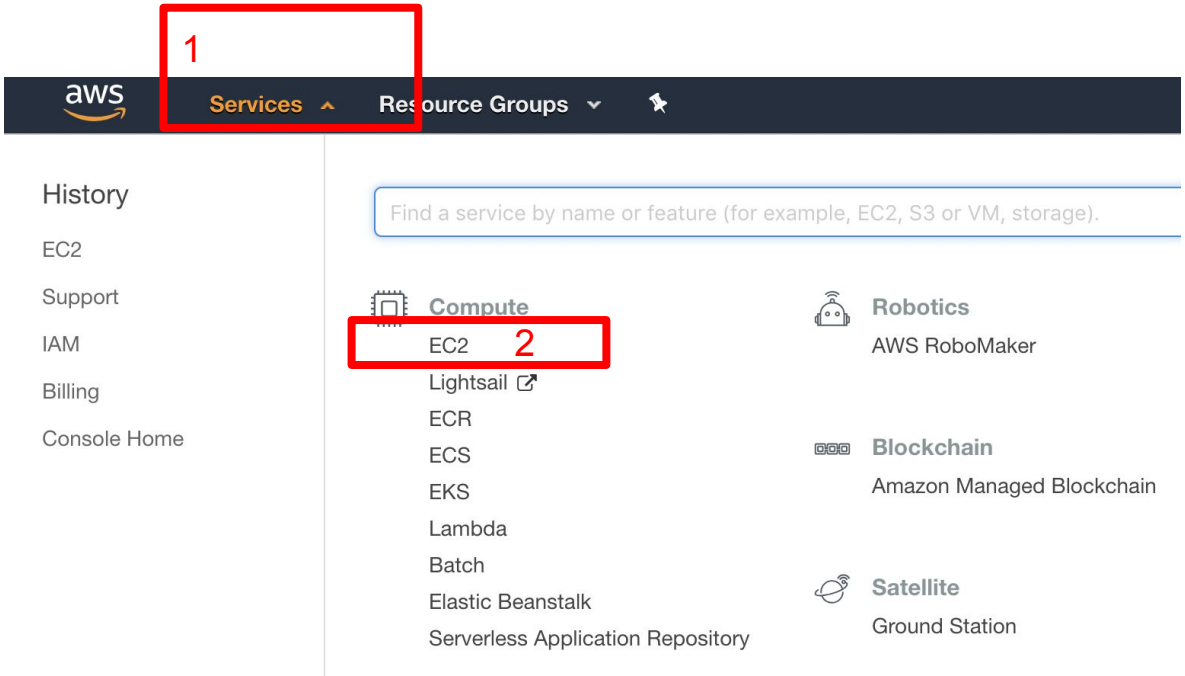
- you will see your case is submitted and
- usually it takes several hours to get quota approved



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.

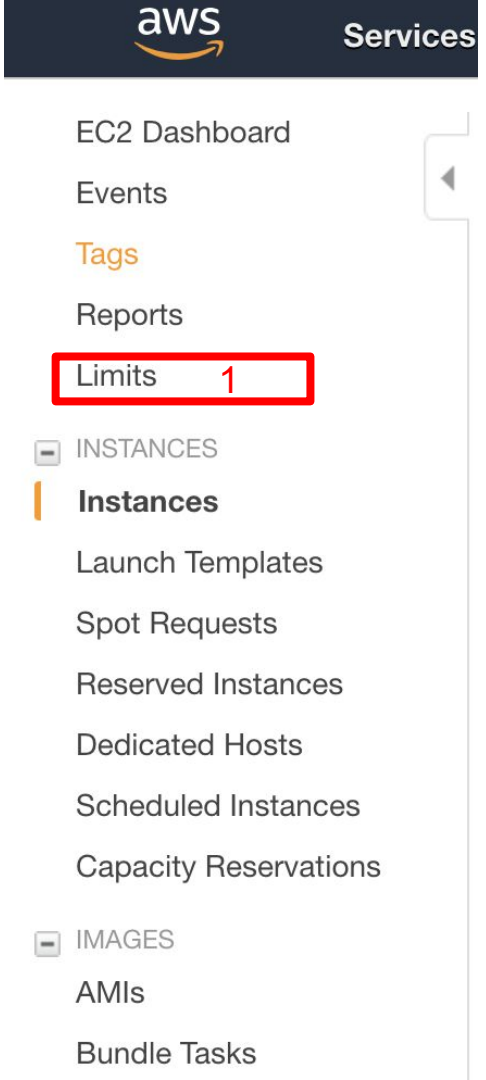
AWS: update GPU limits

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



AWS: update GPU limits

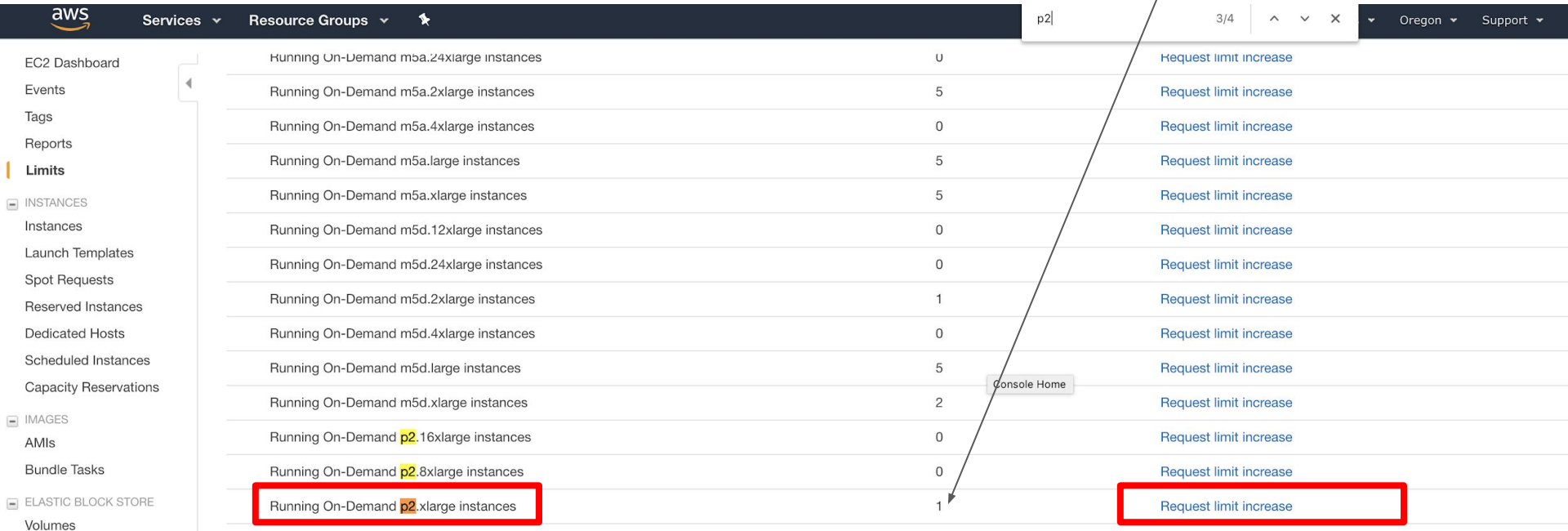
- click “**limits**” in the left side bar



AWS: update GPU limits

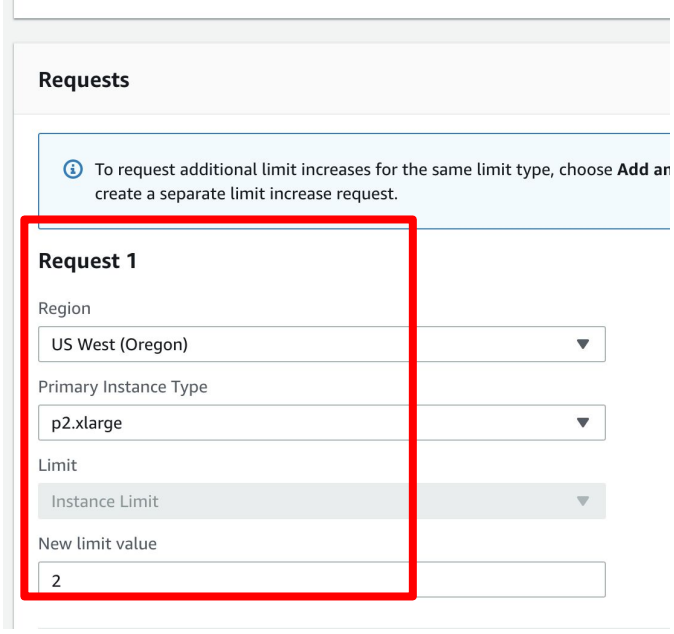
note: if your “limit” value is 1 or more, then you are fine. “0” is the default value for new users.

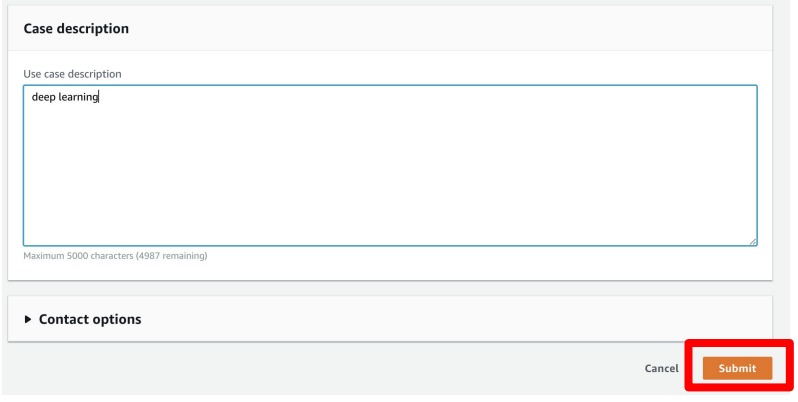
- search “p2” in the new page
- click “**Request limit increase**” for “Running On-Demand p2.xlarge instances”



AWS: update GPU limits

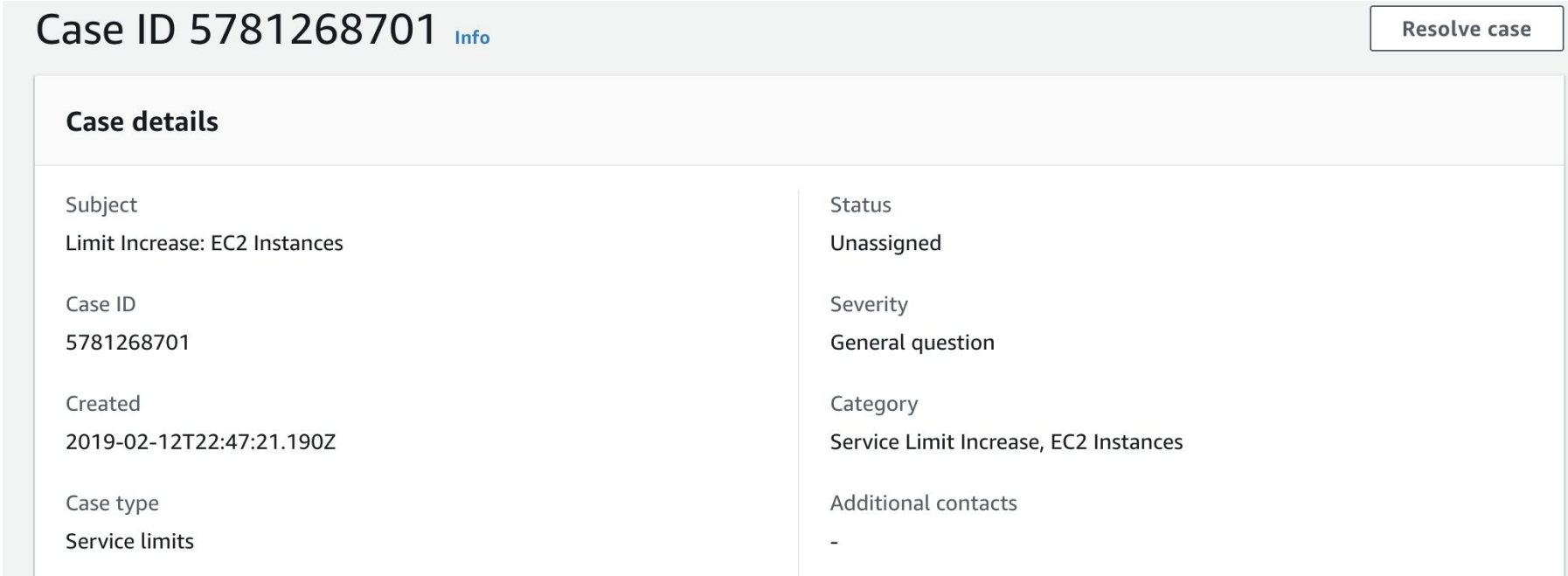
- in the new tab, under “requests”
  - choose the region you want, e.g. Oregon (or whatever you want)
  - choose the machine type: p2.xlarge
  - write down new limit (max GPUs you want), e.g., 1
- under “requests”, fill in the reason, e.g. “deep learning”
- click submit





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



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.