GPU Computing

General Questions

What

Cybera's Rapid Access Cloud is offering GPU instances for short-term use to Rapid Access Cloud users. Those doing highly parallelizable computations can benefit from using GPU accelerated computing to greatly speed up their calculations.

> ⚠ These jobs require custom code capable of leveraging GPU resources.

The Rapid Access Cloud provides 3 types of GPUs:

- g1.* instances have Titan K80 GPUs with 2,496 CUDA cores and 12 GB of RAM
- g2.* instances have Titan XP GPUs with 3,840 CUDA cores and 12 GB of RAM
- g3.* instances have Titan RTX GPUs with 4608 CUDA cores and 24 GB of RAM.

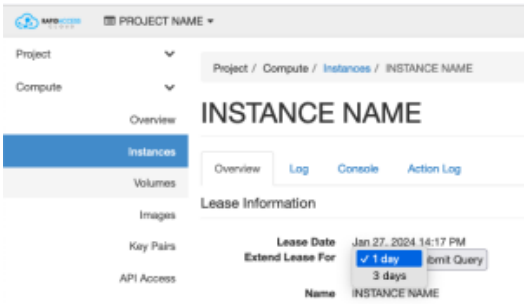
Each default Rapid Access Cloud allocation has enough resources to allow a user to launch 1 VM capable of accessing 1 GPU and are available on a first come, first serve basis.

Who

The GPUs are available to any Albertan to use for pre-production, pre-commercial research, experimentation, or testing purposes. To access the default allocation of 1 GPU instance, simply launch the GPU in the Edmonton region of the Rapid Access Cloud.

How long

The default flavour (`g1.24hr-auto-destruct`) expires after 24 hours. During this initial allocation, users can extend the lease on the virtual machine by an additional 24 or 72 hour period using the Rapid Access Cloud dashboard. This can be found by viewing the details of your instance. Users are free to extend the lease multiple times



For users who would prefer a custom, multi-day allocation, please contact Cybera. Ideal use-cases for multi-day allocations include users who are processing data with the GPUs in such a way that it removes a significant barrier for completing the rest of their work/research.

How

If you already have a Rapid Access Cloud account, simply launch a GPU enabled instance from the dashboard or via the command line interface. If you do not have an account yet, sign up here.

Supported use-cases

Supported use-cases include general purpose GPU applications, such as:

- Deep learning
- Analytics
- Engineering applications

We do not support 3D visualization work at this time.

Contact Info

Email: rac-admin@cybera.ca

Technical/Usage Questions

What is best practice for using the GPUs?

Because there are a limited number of GPUs available, users are expected to only use a GPU enabled instance when it is actually in-use. If there is a period of more than one day during which users know the GPU will not be needed, we ask that users terminate their instance in order to free it up for other users.

In order to resume their work at a later time, users should:

1. Shut-off the instance
2. Take a snapshot of the instance
3. Terminate the instance

When ready to resume their work, users can use the snapshot to relaunch their GPU instance.

What is pre-installed on the GPU images?

We provide several GPU CLI only images with NVIDIA drivers and CUDA pre-installed:

Image Name	CUDA Version
Ubuntu 16.04	10.1
Ubuntu 18.04	10.1
CentOS 7	10.1

How do I free up my GPU instance when it’s not in use?

If there is a period of more than one day during which users know the GPU will not be needed, we ask that users terminate their instance in order to free it up for other users.

In order to resume their work at a later time, users should:

1. Shut-off the instance
2. Take a snapshot of the instance
3. Terminate the instance

When ready to resume their work, users can use the snapshot to relaunch their GPU instance.

What happens at the end of my 24 hour allocation?

At the end of the lease, the instance will be automatically snapshotted and terminated. While the automated shelving process takes a snapshot of your instance, these snapshots should be considered a safety mechanism and will be deleted periodically.

What if I need the GPU for more than 24 hours?

You can extend the lease by 24 or 72 hours if you need the virtual machine for more than one day. Simply go to Dashboard Instances Click on the instance name Select Extend Lease For

How do I access my data after the GPU allocation has ended?

If you snapshot the instance, you can simply re-launch from that snapshot using the `g1.24hr-auto-destruct flavour` and booting from your snapshot. If your data was stored on the ephemeral disk (not recommended), the data will be accessible there. Data stored on a volume can simply be attached to any other instance (including non-GPU instances) and accessed from there.

How do I resume my GPU instance?

If you wish to resume your instance that has been shelved after your allocation has ended you can do so by:

1. Go to your instance's page and extend your lease for 24 hours exactly as you would if you want your GPU for more than 24 hours. Your lease is NOT renewed when you unshelve your instance.
2. Choose Unshelve as an option from the dashboard or issue `openstack server unshelve` if you are using the OpenStack command line clients. If a GPU is available your instance will then launch.

If GPUs are not available, you will receive an error that it could not launch. However your instance's data will not be lost. Please try again later as GPUs do become available.