GPU-Enabled Instances

Launching GPU-enabled instances is as simple as launching any other instance in the Rapid Access Cloud, however there are some special processes around the usage of GPU resources.

GPU Availability and Scheduling

Currently, only the Edmonton region (yeg) has two GPU-enabled nodes, each hosting 4 NVIDIA K80 graphics cards equipped with 2 GPU cores and 24GB of onboard RAM. In total, there are 16 GPU cores available. Because we have so few resources at the moment, we have created a *Lease Process* to allow users to launch instances as normal but with a **twenty-four hour lease** on the instance launched. 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.

If you would like to save your work for future use, we ask that users create their own snapshot and do not rely on the shelving process. You can use the snapshot to launch a new instance when the GPU resources are needed again.

There is a possibility that when you try to launch a GPU-enabled instance, you will get an error indicating that there are “not enough resources”, which means all the GPUs are currently in use. At this time the *Lease Process* is on a first-come-first-served basis with no reservations, so if there are no GPU cores available users will need to try again later.

Available GPU Types

At this time, Cybera provides the following GPUs:

GPU Type	Flavour Series
NVIDIA K80	g1
NVIDIA TITAN Xp	g2
NVIDIA TITAN RTX	g3

GPU Utilization Best Practice

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.

Launching a GPU instance

1. Log into the Edmonton region of the Rapid Access Cloud.
2. Select a flavour with the `g1` or `g2` prefix. For example, `g1.24hr-auto-destruct`.
3. (Optional) Select a GPU image if you want to take advantage of pre-installed CUDA and NVIDIA drivers.

Lease Extension

Once a GPU-enabled instance has been launched, users can extend the lease on the instance by twenty-four hours if they require more time. There is no limit imposed by the *Lease Process* on how many times you can extend the lease, however if there are other Rapid Access Cloud users waiting for resources new extensions may not be granted.

1. Log-in to the Rapid Access Cloud dashboard at https://cloud.cybera.ca.
2. In the left-hand panel under “Compute”, click “Instances”.
3. Click the name of a GPU-enabled instance you would like to extend the lease on. Note the “Lease Date”; this will initially be set to the time and date the instance was created plus twenty-four hours.
4. Click the “Submit” button; the Lease Date will be set to twenty-four hours from the time you click Submit.

Lease Expiration and Unshelving

At the end of the lease, the GPU instance will automatically be 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.

If you need to recover a shelved instance, follow this process:

1. Extend the lease on the instance.

2. Execute the following command using the command line tools:

```
nova unshelve <instance uuid>
```

Note that you should create your own snapshot before the instance expires, which will allow you to resume your work whenever without having to unshelve.