How do I configure GPU compute jobs for the cluster?

Tell Me
If you would like to utilize the NVIDIA GPUs on the cluster for your compute job, below are some tips to help your job do so.

Request a GPU (or GPUs) in your Submit Script
Make sure you ask the scheduler for a GPU in your job request (submit script). You append the GPU request on the #PBS directive in which you ask for CPUs, in your submit script, for example:

```bash
#PBS -l nodes=1:ppn=1:gpus=1
```

Load a CUDA Environment Module
Unless your code has built-in GPU support (for example, Matlab), you may want to load one of the available CUDA Toolkit modules; currently we offer 3: cuda/7.5, cuda/8.0, or cuda/9.0. You can load one of the 3 available by adding a "module load..." line to your submit script. You can also issue a "module list" command to display what modules are currently loaded. The CUDA binaries (like nvcc) and libraries should now be available to your compute job:

```bash
module avail cuda
----------------- /apps/usr/pbs/modules/apps -----------------
cuda/10.0 cuda/10.2 cuda/9.2 (default)
module load cuda/10.2
```

NVIDIA CUDA Deep Neural Network (cuDNN)
(Optional) If your code depends on The NVIDIA CUDA Deep Neural Network (cuDNN) GPU-accelerated library, you must load an available cuDNN module to set up your $LD_LIBRARY_PATH. There are several cudnn modules to choose from, depending on what cudnn version *and* what CUDA Toolkit version you require. Please use the command "module avail cudnn" to see what’s available.

```bash
module avail cudnn
------------------- /apps/usr/pbs/modules/lib -------------------
cudnn/7.2.1-cuda9.2 cudnn/7.4.2-cuda9.2 (default) cudnn/7.6.5-cuda10.2
```

```bash
module load cudnn/7.6.5-cuda10.2
```

Request GPU by Model
(Optional) If you would like to target a specific model of GPU, you can add a "feature" tag to your request. For example the following directive requests one node with one traditional computing core and one GTX-1080ti GPU. There is also a "k80" tag for requesting one of the existing Telsa K80 GPUs. The following directive requests one node with one traditional computing core and one K80 GPU:

```bash
### If you prefer an NVIDIA Tesla GTX-1080ti, specify the "gtx1080ti" feature tag:
#PBS -l nodes=1:ppn=1:gpus=1:gtx1080ti
```
### If you prefer an NVIDIA Tesla K80, specify the "k80" feature tag:
```
#PBS -l nodes=1:ppn=1:gpus=1:k80
```

Request GPU by Precision

(Optional) Some codes require higher-accuracy mathematical calculations. In these applications, data is represented by values that are twice as large, using 64 binary bits instead of 32 bits. The larger, more accurate values are called “double-precision” (dp), while the smaller, less accurate values are called “single-precision” (sp). All of the NVIDIA GPUs we have on the various compute nodes support both sp and dp calculations, however, the performance is significantly higher in some cards than others. If you know your code requires dp calculations (GROMACS, for example), you can request dp GPU(s) from the scheduler by adding the “dp” property to your #PBS request. Likewise, if you know your code does not require dp (Machine Learning, Tensorflow, etc), then you can add the “sp” property to your #PBS request. Here are some examples:

- To request a double-precision GPU, specify the "dp" feature tag:
  ```
  #PBS -l nodes=1:ppn=1:gpus=1:dp
  ```

- To request a single-precision GPU, specify the "sp" feature tag:
  ```
  #PBS -l nodes=1:ppn=1:gpus=1:sp
  ```

Related FAQs

- How do I configure GPU compute jobs for the cluster?
- Where can I find Copperhead user notes?
- Can I run parallel jobs on the cluster?
- Can I log in to Research Computing clusters from home?
- HPC Cluster Questions