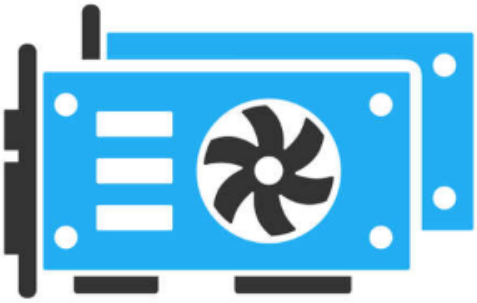
GPU Nodes



Overview

Compute Resources

More detailed information on system resources can be found on our Compute Resources page.

Containers with GPU Support

Singularity containers are available as modules on HPC for GPU-supported workflows. For more information, see our documentation on Containers.

Accessing GPUs

Information on how to request GPUs using SLURM can be found in our SLURM Documentation.

Training

For a list of training resources related to GPU workflows, see our Training documentation.

Cluster Information

Puma

Puma has a different arrangement for GPU nodes than Ocelote and ElGato. Whereas the older clusters have one GPU per node, Puma has four. This has a financial advantage for providing GPU's with lower overall cost, and a technical advantage of allowing jobs that can use multiple GPU's to run faster than spanning multiple nodes. This capability comes from using a newer operating system.
Each node has four Nvidia V100S model GPUs. They are provisioned with 32GB memory compared to 16GB on the P100's.

	V100 PCIe	V100 SXM2	V100S PCIe
GPU Architecture	NVIDIA Volta		
NVIDIA Tensor Cores	640		
NVIDIA CUDA® Cores	5,120		
Double-Precision Performance	7 TFLOPS	7.8 TFLOPS	8.2 TFLOPS
Single-Precision Performance	14 TFLOPS	15.7 TFLOPS	16.4 TFLOPS
Tensor Performance	112 TFLOPS	125 TFLOPS	130 TFLOPS
GPU Memory	32 GB /16 GB HBM2		32 GB HBM2
Memory Bandwidth	900 GB/sec		1134 GB/sec
ECC	Yes		

Ocelote

Ocelote has 45 compute nodes with Nvidia P100 GPUs that are available to researchers on campus. The limitation is a maximum of 10 concurrent jobs. One node with a V100 is also available. Since there is only one, you can feel free to use it for testing and comparisons to the P100, but production work should be run on the P100's. There is also one node with two P100's for testing jobs that use two GPU's. This one should be used to compare with running a job on two nodes.

PERFORMANCE SPECIFICATION FOR NVIDIA TESLA P100 ACCELERATORS

	P100 for PCIe-Based Servers
Double-Precision Performance	4.7 TeraFLOPS
Single-Precision Performance	9.3 TeraFLOPS
Half-Precision Performance	18.7 TeraFLOPS
NVIDIA NVLink™ Interconnect Bandwidth	-
PCIe x16 Interconnect Bandwidth	32 GB/s
CoWoS HBM2 Stacked Memory Capacity	16 GB or 12 GB
CoWoS HBM2 Stacked Memory Bandwidth	732 GB/s or 549 GB/s
Enhanced Programmability with Page Migration Engine	✓
ECC Protection for Reliability	✓

Cuda Modules

Nvidia Nsight Compute (the interactive kernel profiler) is not available. In response to a security alert (CVE-2018-6260) this capability is only available with root authority which users do not have.

The latest Cuda module available on the system is **11.0** and is the only version until newer ones come along. The Cuda driver version can be queried with the `nvidia-smi` command. To see the modules available, in an interactive session simply run:

```
$ module avail cuda

-------------------- /opt/ohpc/pub/moduledeps/gnu8-openmpi3 --------------------
   cp2k-cuda/7.1.0

------------------------- /opt/ohpc/pub/modulefiles --------------------------
   cuda11-dnn/8.0.2    cuda11-sdk/20.7    cuda11/11.0
```

OpenACC

The OpenACC API is a collection of compiler directives and runtime routines that allow you to specify loops and regions of code in standard C, C++, and Fortran that you can offload from a host CPU to the GPU.

We provide two methods of support for OpenACC

1. We support OpenACC in the PGI Compiler. The PGI implementation of OpenACC is considered the best implementation. "module load pgi" on Ocelote. If you are on a GPU node from an interactive session you can run "pgaccelinfo" to test functionality. Remember that the login nodes do not have GPUs or software installed.
   A useful getting-started guide written by Nvidia is available here: https://www.pgroup.com/doc/openacc17_gs.pdf

2. We support OpenACC in the GCC Compiler 6.1 which is automatically loaded as a module when you log into Ocelote. Verify with "module list".
   The GCC 6 release includes a much improved implementation of the OpenACC 2.0a specification.
   A useful quick reference guide is available from: https://gcc.gnu.org/wiki/OpenACC#Quick_Reference_Guide

---

Applications

Many applications have been optimized to run faster on GPU's. These include:

Application	Information	Access
NAMD	Installed as a module	`$ module load namd`
VASP	A restricted license version is installed; only available to licensed users	`$ module load vasp`
GROMACS	Installed as a module	`$ module load gromacs`
LAMMPS	Installed as a module	`$ module load lammps`
ABAQUS	Installed as a module and available as an application through Open OnDemand	`$ module load abaqus`
GAUSSIAN	Installed as a module. See these notes.	`$ module load gaussian/g16`
MATLAB	Installed as a module and available as an application through Open OnDemand. Review the GPU Coder on their website	`$ module load matlab`
ANSYS Fluent	Installed as a module and available as an application through Open OnDemand	`$ module load ansys`
RELION	Available as a Singularity container or as a module.	`$ module load relion`
ML and DL Frameworks	See the section below.	

Python ML/DL including Nvidia RAPIDS

The minimum version of Python that is supported is 3.6:

Framework	Details

numba	RAPIDS: numba is for Cuda programming
cuml	RAPIDS: Cuda Machine Learning has many ML algorithms like K-means, PCA and SVM
cudf	RAPIDS: Cuda Dataframes supports loading and manipulating datasets
tensorflow	TensorFlow is an open source software library for numerical computation using data flow graphs.
torch	PyTorch supports tensor computation and deep neural networks
caffe2	A deep learning framework
tensorrt	Inference server for deep learning
tensorboard	Visualization tool for machine learning