EECE 7398: Machine Learning with Small Data

Utilizing the Discovery Cluster for Advanced Research

Prof. Sarah Ostadabbas

Electrical and Computer Engineering @Northeastern University

September 4, 2024

Introduction to the Discovery Cluster

- What is the Discovery Cluster?
- Why is it critical for machine learning research?
- What are the capabilities of Northeastern's Research Computing?

What is the Discovery Cluster?

- The Discovery Cluster is Northeastern's primary **high-performance computing (HPC)** resource.
- It enables complex computational tasks including:
  - Large-scale data analysis
  - Machine learning and artificial intelligence
  - Simulations and modeling in scientific research
- **Over 900 nodes** with CPU and GPU access, optimized for research workflows.

Northeastern University
Research Computing

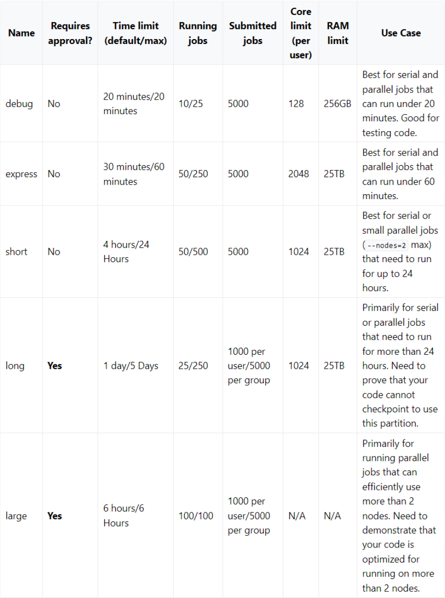
Why Use the Discovery Cluster?

- The Discovery Cluster accelerates machine learning workloads by providing:
    - **GPU acceleration** for deep learning
    - High-performance multi-core CPUs for parallel tasks
    - Access to **large memory nodes** for data-intensive applications
- It is designed to scale up experiments from small datasets to large data with faster results.

Key Features of the Discovery Cluster

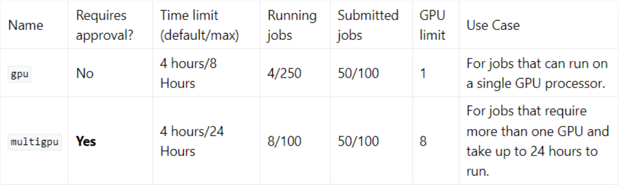
- High-Performance Computing Resources:
  - CPU nodes with multiple cores
  - GPU nodes for accelerated machine learning and AI
- Flexible Environment:
  - Access via Open OnDemand (OOD) or SSH
  - Batch job scheduling using SLURM
  - Customizable environments with Conda, Python, and other modules
- Collaboration and Support:
  - Access to RC's documentation, forums, and dedicated helpdesk
  - Collaborative projects across departments

CPU Partition on the Discovery Cluster



Name	Requires approval?	Time limit (default/max)	Running jobs	Submitted jobs	Core limit (per user)	RAM limit	Use Case
debug	No	20 minutes/20 minutes	10/25	5000	128	256GB	Best for serial and parallel jobs that can run under 20 minutes. Good for testing code.
express	No	30 minutes/60 minutes	50/250	5000	2048	25TB	Best for serial and parallel jobs that can run under 60 minutes.
short	No	4 hours/24 Hours	50/500	5000	1024	25TB	Best for serial or small parallel jobs (`--nodes=2` max) that need to run for up to 24 hours.
long	**Yes**	1 day/5 Days	25/250	1000 per user/5000 per group	1024	25TB	Primarily for serial or parallel jobs that need to run for more than 24 hours. Need to prove that your code cannot checkpoint to use this partition.
large	**Yes**	6 hours/6 Hours	100/100	1000 per user/5000 per group	N/A	N/A	Primarily for running parallel jobs that can efficiently use more than 2 nodes. Need to demonstrate that your code is optimized for running on more than 2 nodes.

GPU Partition on the Discovery Cluster

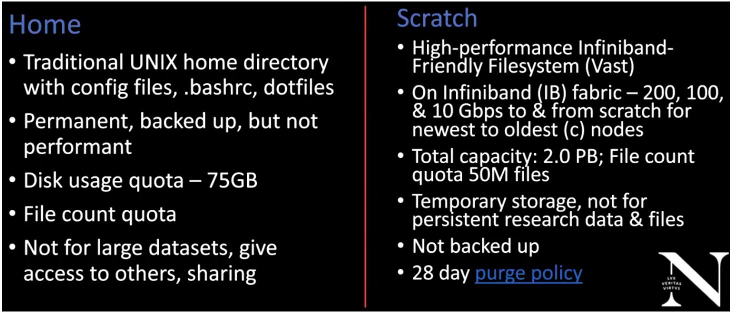


Name	Requires approval?	Time limit (default/max)	Running jobs	Submitted jobs	GPU limit	Use Case
`gpu`	No	4 hours/8 Hours	4/250	50/100	1	For jobs that can run on a single GPU processor.
`multigpu`	**Yes**	4 hours/24 Hours	8/100	50/100	8	For jobs that require more than one GPU and take up to 24 hours to run.

How to Access the Discovery Cluster (Part 1)

- There are several methods to access the Discovery Cluster:
- **Open OnDemand (OOD) Web Portal:**
  - A web-based interface for managing files, submitting jobs, and launching interactive apps (e.g., Jupyter, RStudio).
  - Access it through your browser: [Open OnDemand](https://ood.discovery.northeastern.edu)
  - You can upload and download small data files directly using the OOD file transfer feature.
- **SSH (Secure Shell) Access:**
  - Direct terminal-based access for running commands and managing jobs.
  - To connect, use: ssh your-username@discovery.rc.northeastern.edu
  - This method requires an SSH key or password-based authentication.

Storage Configuration



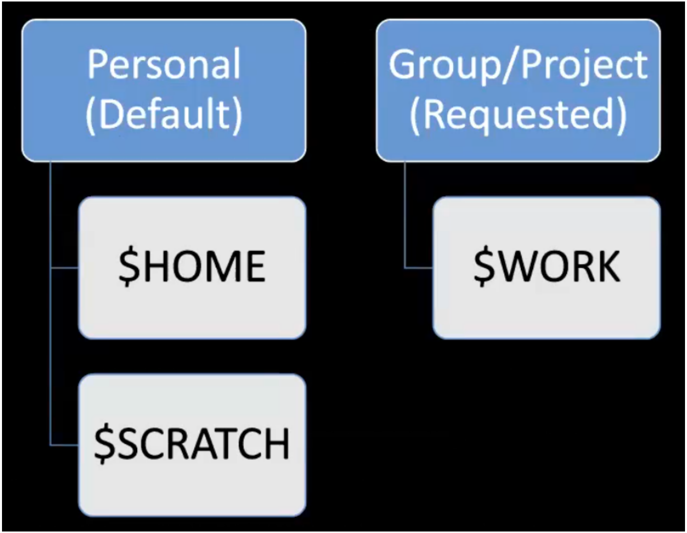
Home

- Traditional UNIX home directory with config files, .bashrc, dotfiles
- Permanent, backed up, but not performant
- Disk usage quota – 75GB
- File count quota
- Not for large datasets, give access to others, sharing

Scratch

- High-performance Infiniband-Friendly Filesystem (Vast)
- On Infiniband (IB) fabric – 200, 100, & 10 Gbps to & from scratch for newest to oldest (c) nodes
- Total capacity: 2.0 PB; File count quota 50M files
- Temporary storage, not for persistent research data & files
- Not backed up
- 28 day purge policy

Storage Hierarchy Configuration



Personal (Default)

$HOME

$SCRATCH

Group/Project (Requested)

$WORK

How to Access the Discovery Cluster (Part 2)

- **VSCode Remote SSH:**
  - Use Visual Studio Code with the **Remote - SSH** extension to work directly on the Discovery Cluster.
  - Provides an integrated development environment for coding, file management, and terminal access.
- **Jupyter Notebooks:**
  - Launch Jupyter Notebooks via the Open OnDemand portal for interactive development.
  - Jupyter can be configured to leverage GPU resources for machine learning tasks.

Setting Up Your Research Environment

- Use Conda to create an isolated Python environment.
- Install key libraries:
    - PyTorch for machine learning
    - NumPy, SciPy, and other dependencies
- Manage software modules, including CUDA for GPU access.

PyTorch and GPU Utilization

- PyTorch is a powerful deep learning framework.
- Ensure you install PyTorch with GPU support for optimal performance.
- Verify that GPUs are available on the Discovery Cluster:
    - Monitor GPU resources with nvidia-smi

Running Experiments Efficiently on the Cluster

- Use SLURM for job scheduling:
  - Submit jobs for batch processing
  - Request appropriate resources (CPUs, GPUs, memory)
- Use Jupyter Notebooks via Open OnDemand for interactive development.
- Monitor and manage GPU resources during training.

Monitoring and Experiment Tracking

- Weights and Biases (or similar tools) for experiment tracking:
    - Track metrics, model versions, and data.
    - Analyze training progress in real-time.
- Use NVIDIA tools to monitor GPU usage:
    - Track memory, temperature, and performance.
    - Run diagnostic tools during training.

Best Practices and Support

- ”Login Node Use Warning” – don’t perform compute intensive jobs on login node when possible
- Keep/home under quota by cleaning it regularly.
- Clear unused packages and caches from /.conda: conda clean -all conda env remove –name ¡name¿
- File Management perform post-processing or tarball files to keep storage under check.

Q&A

- Open discussion on accessing and using the Discovery Cluster.
- Troubleshooting common issues with cluster access or job scheduling.
- Ask about specific research needs or project requirements.

Conclusion and Next Steps

- Recap of key points: Accessing and setting up the Discovery Cluster.
- Apply today's knowledge to your own research projects.
- Research Computing Office Hours
- Join Wednesday Office Hours : 3 - 4 p.m. ET
- Join Thursday Office Hours : 11 a.m. - 12 p.m. ET