Killarney User Guide For Vector Researchers

June, 2025

The Killarney Compute Cluster, part of the Pan-Canadian AI Compute Environment (PAICE), is managed by the Vector Institute and hosted by the University of Toronto, Ontario. Killarney is an upgrade from the current on-prem Vaughan cluster both in terms of technology and in its management and governance. Working with the Digital Research Alliance of Canada (DRAC), Killarney follows similar architectural design and controls as the national sites.

Our AI partner Institutions, AMII and MILA also manage compute clusters Vulcan and TamIA, respectively.

This document is intended to assist Vector researchers currently using the Vaughan cluster to migrate their research workloads to Killarney.

Migrating From Vector's Legacy Vaughan Cluster

Current Vector users will see many improvements when moving to Killarney.

The GPUs on Killarmey's compute nodes, Dell R750xa and XE9680 models, are a generational advancement over the deployment at Vaughan, with Killarney's L40S GPUs representing a sizable leap from the A40 tier, and H100 GPUs providing providing dramatically higher performance than the A100 nodes.

Storage performance is improved by making use of a parallel, distributed filesystem provided by the VastData platform with a total of 1.7 petabytes of space. Data is organized differently from Vaughan, with Project directories going from a per-request, custom-access directory to a PI-assigned directory structure. Scratch and Home directories serve similar purposes to that of Vaughan.

Connecting the compute and storage together is a high speed Infiniband HDR based network with the L40S based nodes running at 100 Gbps and the H100 based nodes running at 400 Gbps. This is an improvement over Vaughans' 50 Gbps Ethernet-based network.

Software is provided via the Alliance's extensive library of software that is shared across all advanced research compute clusters. Similar to the Vaughan cluster, users access this software using Modules. Users are encouraged to build shared software environments in their assigned Project space (via python venv, for example) rather than requesting a custom software suite be installed under the global root, if possible.

Requesting Access to Killarney

Principal Investigators and their Research members **must** register their individual accounts in the DRAC's CCDB as described at the Apply for an Account page,

- When registering, please note your **Institution Name** should be your non-Vector affiliation, University of Toronto, for example.

Once your CCDB account has been set up, navigate to https://ccdb.alliancecan.ca/me/access_services and opt-in to the systems you wish to use (*this page is not in the CCDB's menu system at time of writing*).

Using the Cluster

Once your CCDB account has been set up and you have opted-in to use Killarney, you can login the cluster by using SSH to killarney.alliancecan.ca.

Ensure your CCDB account has your SSH public key uploaded and Duo multi-factor authentication configured - further details on these items, as well as general tips for running jobs, how the scheduler works, and so on can be found in the Alliance's Getting Started documentation.

You must have opted in via the CCDB via https://ccdb.alliancecan.ca/me/access_services to be able to log in, and be associated with an active AI project (AIP) in order to submit jobs.

The Killarney environment has two tiers of GPU compute nodes

- Standard compute: 168 nodes, each with 64 cpu cores, 512 GB RAM, and 4x NVIDIA L40S 48GB GPUs, connected with 100 Gbps Infiniband
- Performance compute: 10 nodes, each with 48 cpu cores, 2048 GB RAM, and 8x NVIDIA H100 SXM 80GB GPUs, connected with 400 Gbps Infiniband
- Note there are no dedicated CPU compute nodes and as such no distinction is made between GPU and CPU jobs.

As this is an Alliance-based site, things are a little different compared to Vector's private compute cluster at Vaughan.

- When submitting jobs using sbatch or srun, you must ensure that you specify
    - The slurm Account (-A option)
    - The time required (-t option)

    - A working directory other than your home directory, such as /scratch/<login> or /project/<project> (-D option)
    - GPUs via the --gres option. If no GPU is specified, none will be allocated; if a quantity is specified, but no model, that quantity of L40S GPUs is allocated.

      --gres=gpu:2 -> two L40S GPUs
      --gres=gpu:l40s:2 -> two L40S GPUs
      --gres=gpu:h100:8 -> eight H100 GPUs

- Take care to request an appropriate amount of memory (--mem option); if you only need one GPU, but you request all of the compute node's system RAM, you will block access to the node for other users (since no memory is available), and the remaining GPUs will be unavailable for use. As a guideline, standard compute nodes have 128GB of RAM and 16 cores per GPU and performance compute nodes have 256GB of RAM and 12 cores per GPU.

- Storage is available via the common Alliance paths:
    - /home/$USER - your home directory
    - /scratch/$USER - scratch area for temporary data
    - /project/[project-name] - shared working directory
    - /datasets - read-only, common reference data that is useful for many jobs

    - Quotas, backup, and purge policies are outlined on the Alliance's wiki.

Getting Help

For technical help please see https://docs.alliancecan.ca/ or email support@alliancecan.ca.