Migrating to the upgraded national systems

Sergey Mashchenko
(SHARCNET)

July 30, 2025

Outline

- The context
- How much is changing?
- Changes which might affect you
    - User agreements (CCDB web site)
    - Fatter nodes
    - GPUs: MIGs and MPS
    - Changes to the interconnect
    - Changes to the file systems
    - Changes to the scheduler
    - Changes to the software
    - Interactive access
- Conclusions

The context

National systems upgrades 2025

- Four out of five DRAC national systems are going through major upgrades this spring/summer:
  - Beluga → Rorqual
  - Graham → Nibi
  - Niagara → Trillium
  - Cedar → Fir
- The fifth system (Narval) is the only one which is not been upgraded this time
- The upgrades are to replace our aging hardware with the current technology, achieving a 3-4x increase in the number of CPU cores and amount of memory.
  - The number of GPUs will actually go down, but the compute power will grow by a factor of 3.5x.
  - Storage capacity will not grow much this time.

(cont.)

- Some of the new systems are already fully operational, and the plan is to make all the systems fully operational by the end of July.
  - The new RAC projects should start at that point
- The overview of the migration to the new systems process can be found on this page:

  https://docs.alliancecan.ca/wiki/Infrastructure_renewal

How much is changing?

The good news

- Very little will change in terms of how our users use the systems.
- We did our best to minimize the impact of the upgrades on the existing workflows, which in most cases won’t have to change when migrating to the new systems.
- One area where some workflow adjustments will be required is GPU computing (more about that later).
- You may want to recompile your CPU (definitely GPU) codes for the new systems.
- Most of the things that did change will not affect the existing workflows, but are important to be aware of as they can be very useful for users.
  - Fatter nodes, better performing file systems, better interactive access options etc.

Changes which might affect you

User agreements (CCDB web site)

- To access the new systems, you might have to visit the CCDB web site, and to opt in for specific clusters.
  - Once logged in to CCDB, go to Resources → Access Systems menu item. The direct link: https://ccdb.alliancecan.ca/me/access_systems
  - In the systems table on the left side, the clusters with a green check mark are already opted in; the ones with an exclamation mark require you to go through the opt in step (click on the exclamation mark to start the process).
  - Rorqual requires all the users to go though this step, while other systems (like Nibi) will automatically accept recent Graham users.
  - After opting in, it may take up to an hour before you can login to the cluster and have everything setup (home, SLURM account etc.)

Fatter nodes

- One big (and positive) change is the transition from 32...40 cpu cores per node to 192 cpu cores per node setup, on all four upgraded national systems.
  - Narval – the system which is not upgraded this time – has nodes with 48 and 64 cpu cores.
- This significant change may bring new life to the research codes which rely solely on CPU multi-threading for parallel computing (OpenMP etc.).

(cont.)

- Be aware of the Amdahl’s law though.
  - This is the basic law of parallel computing, postulating that you cannot necessarily make your parallel code run faster by simply adding new CPU (or GPU) cores: at some point, the CPU utilization will drop down (often dramatically), with the code becoming extremely wasteful.
  - One useful workaround is provided by the Gustafson’s law: **increasing the problem size at the same rate as the increase of the CPU cores** will often let you use many more CPU cores efficiently.
  - Double check your workflow: if it currently utilizes the whole node for multi-threading, it needs to be tested with a variable number of CPU cores (up to 192) on the new systems.

(cont.)

- Another important consequence of the fatter nodes: much more memory per node.
  - We maintained our usual 4GB/CPU core ratio in the new systems.
  - As a result, the amount of RAM per node in regular compute nodes increased from 128...192 GB (on the old systems) to 768 GB (on the new systems).

GPUs: MIGs and MPS

- GPUs is likely the only area where you will have to make some adjustments to your existing workflow.
- The reason is that we are replacing our older generation Nvidia GPUs (P100 and V100) with **fewer** H100 (more modern) GPUs.
  - After the upgrades, the combined compute power in GPUs will grow by a factor of 3.5x (from ~6000 RGU to ~21,000 RGU), but the number of GPUs will actually go down – from 3200 to 2100.
  - A side note: Nibi is the only cluster which has 8 H100 GPUs per node (the rest have 4 H100 per node).
- We are already having issues with most jobs under-utilizing the GPUs we have (P100, V100, A100), because the problem size is too small to saturate these GPUs.
- With much more powerful H100 GPUs, this problem would become much worse – if we didn’t do anything.

(cont.)

- To address this issue, many H100 GPUs on our new systems (and also many A100 GPUs on Narval) are segmented into smaller size virtual GPUs, using the MIG technology from Nvidia.
- Segment sizes are
  - 1/8 in terms of computing (with 10 GB RAM)
  - 2/8 in terms of computing (with 20 GB RAM)
  - 3/8 in terms of computing (with 40 GB RAM)
  - Full size H100 (80 GB RAM)

(cont.)

- The exact details of MIGs are cluster specific.
  - For example, on Rorqual one half of all H100 GPUs are split into the MIGs of different sizes, with the names
    - nvidia_h100_80gb_hbm3_1g.10gb
    - nvidia_h100_80gb_hbm3_2g.20gb
    - nvidia_h100_80gb_hbm3_3g.40gb
  - Simply replace the usual GPU name inside your job script with one of the above names (on Rorqual).
- This page will be updated with more details as they become available: https://docs.alliancecan.ca/wiki/Multi-Instance_GPU

(cont.)

- MIG is a great way to access a fraction of a full size GPU without being affected by other users’ jobs.
- But its disadvantage is that the GPU segmentation is static, and not under user’s control.
- The alternative technology of sharing a GPU - MPS - may work better for many users, as it is allows dynamic sharing of the GPU cores and GPU memory between unrelated processes (but belonging to the same user).
  - Use it for GPU farming, or hybrid OpenMP/CUDA and MPI/CUDA codes.

(cont.)

- Users cannot simply copy their GPU based workflow to the new systems without some modifications and testing.
- They should test the H100 utilization (using cluster portals, or profilers), and if it is not good (**which will be true in most cases**), then either MIG, or MPS, or a combination of the two needs to be tested.
- The good news: neither MIG nor MPS require any changes to the code (no recompiling either). You just need to make a few small corrections to the job script files.
- Consider taking this one-hour long self paced course to master the MIG and MPS technologies: https://training.sharcnet.ca/courses/course/view.php?id=210

Changes to the interconnect

- Things got better overall:
  - Fir offers substantial improvements. Notably, the blocking domain now spans three full racks - approximately 43,000 cores in total (72 nodes × 192 cores × 3 racks). This enables significantly larger, tightly-coupled jobs.
  - Nibi is the only cluster which switched from InfiniBand to Ethernet interconnect (200 Gbit/s for CPU nodes, 400 Gbit/s for GPU nodes). Latency may be larger – test your MPI code for scalability.
  - Trillium has Nvidia’s NDR Infiniband network (400/800 Gbit/s for CPU/GPU nodes).
  - Rorqual: 100 → 200 Gbit/s infiniband.

Changes to the file systems

- Data migration
  - Nibi, Fir: data migration has already happened
  - Trillium: the migration will happen automatically
  - Rorqual: users have to copy their own data
- Nibi and Trillium now use VAST file system
  - NVMe SSD based storage
  - The data is automatically compressed and deduplicated
    - So users do not have to compress their data
  - Snapshots mechanism allows users to access older versions of their files, and to undelete a file (“oops” command on nibi).

- Trillium will now provide default 1TB project space (same as other national systems).
  - Unlike the other systems, this quota cannot be increased by request.
- Unlike Cedar, Fir’s file system doesn’t do automatic compression and deduplication.
- Changes to the project and scratch symbolic links
  - Nibi: same as before, except individual student subdirectories are no longer created automatically inside ~/project/def-xxx directory.
  - Rorqual: the symlinks moved from ~ to ~/links directory.

Scratch

- Scratch (Nibi, and perhaps Trillium): new accounting model
  - If your usage is under 1T, you never have to do anything (no expiration, etc).
  - When you exceed this "soft limit", you trigger an over-quota condition, which includes a 60-day grace period.
  - When the grace period expires, over-quota enforcement starts, disallowing any new allocation. (This implies that your files at that point are still accessible, but adding files or increasing file sizes will fail with an error.)
  - To resolve this, you must bring your usage under the soft limit (which "resets" the clock).
  - There is also a hard quota of 20T.

Changes to the scheduler

- None at the moment. It is still SLURM.
- It is possible some new features will be introduced in the near future.
  - In particular, we are contemplating a possibility of creating a “whole subnode” partition.
  - Similarly to the usual “whole node partition”, we may set aside some nodes which will be scheduled by a fixed (“subnode”) size (say, 32 cpu cores) jobs.
  - This is likely needed as the new nodes are very fat (192 cores) – much larger than on the old systems (32-40 cores).

Changes to the software

- Mostly unchanged.
- StdEnv/2020 is hidden and deprecated, but still loadable if absolutely needed.
- Applications compiled with cuda < 12 do not support the H100
- For Trillium, gentoo/2023 will be loaded by default but not StdEnv. The CCEnv/NiaEnv mechanism on Niagara will not be carried over.

Interactive access

- Using command line: salloc access will be available.
- GUI access:
  - JupyterHub will be available
  - Open OnDemand access is provided for two clusters:
    - Nibi: https://ondemand.sharcnet.ca
    - Trillium: https://ondemand.scinet.utoronto.ca/pun/sys/dashboard
  - OnDemand advantages (attend Tyson’s webinar on August 27 for details!):
    - cut and paste directly works on browsers that support it (and on ones that don't, there is a cut and paste workaround scratch pad), and
    - the desktop automatically resizes (instead of just zooming in and out) to fit you browser window.

Conclusions

- Exiting times: a significant increase in the amount and speed of the resources provided by the national systems!
- A small price to pay: some relatively minor adjustments to do:
  - Copy your data (Rorqual only).
  - Verify your job scripts for using symlinks (project, scratch); correct them as needed.
  - The code you compiled yourself: you may need to recompile it (definitely if it’s a GPU code).
  - CPU parallel codes: redo the scalability test (run with 2, 4, 8, 16… cpu cores; aim at 75% efficiency or better).
  - GPU codes: test it with MPS, and MIGs of different sizes.

References

- https://docs.alliancecan.ca/wiki/Infrastructure_renewal
- Individual new cluster pages
  - https://docs.alliancecan.ca/wiki/Nibi
  - https://docs.alliancecan.ca/wiki/Rorqual
  - https://docs.alliancecan.ca/wiki/Fir
  - https://docs.alliancecan.ca/wiki/Trillium
- https://docs.alliancecan.ca/wiki/Multi-Instance_GPU
- MIG+MPS course: https://training.sharcnet.ca/courses/course/view.php?id=210

Questions?

You can contact me directly (syam@sharcnet.ca)

or send an email to help@sharcnet.ca

The end