Preparing for Gadi

Last updated 11 Jun 2020 (12:00pm AEST)

> ⚠ 26 Mar 2020 **COVID-19 Update**
>
> NCI staff are currently working from home In response to the COVID-19 pandemic. NCI systems remain fully operational and will be supported remotely. NCI user support will continue to operate as normal, with email the preferred mode of support interaction - help@nci.org.au.

NCI will regularly update this page and provide more detailed information as it becomes available. Note that the format of this page (and child pages) may change.

If you have questions or special concerns about how your work may be impacted by the transition from Raijin to Gadi please let us know as soon as possible - contact NCI user support at help@nci.org.au - and we will endeavour to help you as soon as possible.

Gadi Status Summary

UPDATED 04 Feb 2020

- Cascade Lake hugemem nodes are now available on Gadi.
- On Gadi use the PBS directive "-lstorage=<path>" if your job accesses a /g/data directory or the /scratch directory of another project. (Note that POSIX permissions still apply.) Failure to provide these directives will cause a job to fail with a run-time error. See the section below **Filesystems-/g/data** for more information.
- On Gadi, user workflows should reference /g/data directory paths using the form "/g/data/projectcode", i.e. without the alphanumeric filesystem descriptors 1a, 1b, 2, 3, or 4.
- A Raijin run-time compatibility image is provided on Gadi. To use this add the "-limage=raijin" directive in your PBS job script, and modify your cpu request to be Gadi compliant, i.e. a multiple of 48 cpus, if you are using more than one full node. *Use of the Raijin compatibility image will incur a performance penalty. All users are advised to recompile their applications on Gadi as soon as possible.*
- Raijin /short and /home file systems are now offline. These file systems will be decommissioned on Tuesday 28 Jan 2020.

Details on these items can be found in the following sections.

Gadi Timeline

Raijin Sandy Bridge nodes will be decommissioned earlier than originally planned to accommodate electrical power support for Gadi. Please note that this revised schedule is still subject to change.

Date(s)	Events
04 Nov 2019	NCI data centre preparation and Gadi installation phase one - COMPLETED.
15 Nov 2019	Gadi stability and acceptance testing underway. Users preparing for Gadi.
18 Nov 2019	Raijin user home directories will be copied to Gadi home directories ($HOME/raijin_home).
18 Nov 2019	**Transition Phase One** Gadi and Raijin available to users. Gadi pre-production configuration is expected to include one rack of V100 GPU nodes. Gadi allocations will match Raijin Q4 pro-rata allocations. Jobs can be run (independently) on both systems.
18 Nov 2019 - 31 Dec 2019	Raijin /short available read-only on Gadi login and data mover nodes for user file transfers. Progressive deployment of Gadi nodes to full specification, and phased retirement of Raijin Sandy Bridge nodes.
27 Nov 2019	50% of Raijin Sandy Bridge nodes decommissioned to allow power work for Gadi - DONE
17 Dec 2019	Raijin Broadwell nodes offline for power reconfiguration work - DONE
18 Dec 2019	Raijin operational with Broadwell and Skylake nodes only - DONE All Raijin Sandy Bridge nodes decommissioned; "normal" and "express" queues no longer available - DONE
19 Dec 2019	Raijin run-time compatibility environment available on Gadi - DONE
27 Dec 2019 - 06 Jan 2020	**Scheduled Downtime - Gadi** Scheduled downtime for final Gadi configuration and pre-production acceptance testing.
27 Dec 2019 - 31 Dec 2019	Raijin operational with Broadwell and Skylake compute nodes.
09 Jan 2020	**Phase Two** Gadi available to users. Broadwell and Skylake nodes offline for Gadi integration. *UPDATE: The ANU has re-opened its main campus following a multi-day shutdown due to hazardous smoke conditions in the ACT.*

20 Jan 2020 - 22 Jan 2020	ANU campus closure due to severe hailstorm. NCI systems available to users.
20 Jan 2020	Raijin /short and /home file systems decommissioned.
Jan 2020 - TBD	Broadwell and Skylake nodes migrated to Gadi.
02 Apr 2020 - 05 Apr 2020	2020 Q2 Scheduled Maintenance Downtime - Gadi Q2 scheduled quarterly maintenance downtime will be extended to accommodate configuration tuning for Gadi. Details will be provided at a later date.

Please note that this timeline will be updated as often as necessary to reflect progress in data centre preparations, installation activities, and dependencies. NCI must decommission Raijin before Gadi can be configured in its full production capacity.

The message of the day on Gadi login nodes will always contain the most up to date information about system status and availability.

User Environment

1. The user's default shell and project will be controlled by the file gadi-login.conf in each user's $HOME/.config directory.
2. Gadi /home quotas are applied on a per-user basis, as on Raijin.
3. Gadi /home quotas will be 10 GB.
4. Gadi login and compute nodes will run the CentOS 8 operating system.
5. Raijin /short and /home directories are available on Gadi via the paths /raijin/short and /raijin/home, respectively, until 20 Jan 2020. These paths can be accessed read-only on login nodes and copyq/datamover nodes only.

User Environment: Compilers and MPI

Package	Versions
Intel Compilers	2019.3.199
Intel MPI	2018.3.222, 2019.3.199
OpenMPI	2.1.6, 3.0.4, 3.1.4, 4.0.1

NCI plans to provide version OpenMPI 4.0.2 at the time of Gadi pre-production, subject to testing and validation.

Processor Comparison: Raijin vs Gadi

Raijin	Gadi
Intel Xeon E5-2670 (Sandy Bridge)	Intel Xeon Platinum 8274 (Cascade Lake)
Two physical processors per node	Two physical processors per node
2.6 GHz clock speed	3.2 GHz clock speed
16 cores per node	48 cores per node
332 GFLOPs per node (theoretical peak)	4915 GFLOPs per node (theoretical peak).

Resources

The computing charge rate on Gadi is 2.0 service units (SU) per cpu-hour. This rate broadly reflects Gadi's performance relative to Raijin.

All NCI allocations for 2020, including NCMAS, will be on Gadi only.

Compute allocations on Gadi are managed by stakeholder scheme managers, as on Raijin. Check this page - https://nci.org.au/scheme-managers - to identify your scheme manager.

Compute allocations on Gadi will apply to projects, as on Raijin.

In 2019 Q4, all active projects will be given Gadi compute quotas which match (pro-rata) their 2019 Q4 Raijin allocations.

During the Gadi pre-production period, compute (job) accounting on Raijin and Gadi will be independent.

Logging in

To login from your local desktop or other NCI computer run ssh:

ssh abc123@gadi.nci.org.au

where abc123 is your own username. Your ssh connection will be to one of ten possible login nodes. As usual, for security reasons we ask that you do not set up passwordless ssh to Gadi. Entering your password every time you login is more secure, or use specialised ssh secure agents.

File Systems - /home

Gadi /home is a new, independent file system.

The quota on Gadi home directories will be 10 GB, as compared to a 2 GB quota on Raijin. Home directories are intended for irreproducible files, e.g. source code and configuration files. Users are expected to utilise /scratch, /g/data and JOBFS file systems for working data.

On 13 Nov 2019 NCI will copy the contents of each user's Raijin home directory to the user's home directory on Gadi. The copy destination will be a subdirectory on Gadi, $HOME/raijin_home. This will be a one-time copy. Users will be responsible for migrating any further home directory files from Raijin to Gadi after 13 Nov 2019.

Users are strongly encouraged to retain only essential files from their Raijin home directories on Gadi.

UPDATE 08 Jan 2020 – Raijin /short and /home directories are available on Gadi via the paths /raijin/short and /raijin/home, respectively, until 20 Jan 2020. These paths can be accessed read-only on login nodes and copyq/datamover nodes only.

File Systems - /scratch

The temporary file system for Gadi users is /scratch. Note that the path '/short', as used on Raijin, will not exist on Gadi.

Raijin /short will be available on Gadi via a temporary, read-only path on login and data mover nodes only until 20 Jan 2020. Users are strongly encouraged to copy only essential files to Gadi /scratch.

The contents of Raijin /short will not be migrated to Gadi /scratch. It is the responsibility of each user or project to transfer any files he/she needs from Raijin /short to Gadi.

Data transfer rates from Raijin /short to Gadi /scratch are expected to be approximately 1 TB per hour. Please plan your transfers accordingly, and do not wait until the last minute.

Gadi /scratch will be subject to an automated file purging policy: files will be removed 90 days after the time of last modification (mtime). In the interest of fairness and transparency, exceptions to this policy are not permitted.

Access time (atime) will not be considered in the /scratch purging policy. Persistent files should be stored in home directories or in project directories on the /g/data file systems.

Safety quotas, to prevent accidental overpopulation of the file system, will be applied to projects on /scratch. NCI is currently developing the safety quota implementation.

NCI is developing tools and notifications to help users track the status of their files in /scratch.

Any attempts to circumvent the 90-day scratch purge policy by using the touch command or other strategies will result in account deactivation.

The Gadi /scratch purging policy is expected to be activated in 2020 Q2, 01 Apr 2020. Users will have approximately three (3) months to clean up and organise files in /scratch directories before activation of the purging regime.

All compute projects will be provided with a default /g/data directory for storage of persistent data. The default quota for /g/data project directories remains to be finalised. Note that allocations for projects which already have /g/data access will not change in 2020 unless such changes are defined in a contract or agreement.

Plan to modify your workflow(s) to place temporary files on /scratch, and persistent files on /g/data.

File Systems - /g/data

The /g/data file systems will continue to be available on Gadi and Raijin during the Gadi transition phase. Infrastructure work may temporarily impact file system performance during pre-production. Please also note that during transition, while Raijin and Gadi systems are both connected to the /g/data file systems, the file system performance may be impacted, as bandwidth is shared across both systems.

Jobs on Gadi must explicitly declare, via PBS directives, which file systems are to be accessed during the job. As an example, a job which will read or write data in the /scratch/<project> and /g/data/<project> directories must include the directive "-lstorage=scratch/<project>+gdata/<project>". A job that attempts to access a /g/data or scratch directory without this directive will fail during run time. Refer to the Data Collections section (below) to ensure that your access to data collections projects will not be affected by this change.

A user shell on a Gadi login node will not have access to /g/data file system directories of projects for which the user is not a member.

Projects should exercise caution in running workflows on Gadi and Raijin simultaneously during the pre-production period. Jobs which are in flight at the same time on Raijin and Gadi, and which access files on the /g/data file systems, for example, could fail due to file contention.

Project data on the /g/data2 file system was recently migrated to a new file system, /g/data4. A symbolic link /g/data2/g/data4 has been provided for backward compatibility on Raijin. This /g/data2 symbolic link will not be provided on Gadi. All Gadi users are expected to update scripts and workflows to include the new /g/data4 path where needed.

On Gadi, user workflows should reference /g/data directory paths of the form "/g/data/projectcode". A numeric file system descriptor in /g/data paths, for example /g/data**1a**/ab12, should no longer be used.

Jobs

Gadi Cascade Lake normal/express/copyq nodes have 48 CPUs and 192 GB memory.

Users will need to adjust PBS job scripts and workflows from Raijin to suit Gadi: 48 CPUs/node (Cascade Lake, significantly faster than Raijin), 192 GB RAM/node, 400 GB PBS_JOBFS/node, and so on.

Gadi runs PBS Pro version 19.

Job scheduling will be determined at the project level, as on Raijin. It is not possible to schedule jobs on a per-user basis on Gadi.

Gadi will have Normal and Express queues as on Raijin. Gadi's Broadwell and SkyLake queues will conform to Raijin specifications. Queue details are available on the page Gadi Job Queues.

The PBS_JOBFS size on Gadi normal/express/copyq nodes will be limited to 400 GB per node. Jobs that require more than 400 GB/node are expected to use /scratch disk.

Jobs on Gadi must explicitly declare, via PBS directives, which file systems are to be accessed during the job. As an example, a job which will read or write data in the /scratch/<project> and /g/data/<project> directories must include the directive "-lstorage=scratch/<project>+gdata/<project>". A job that attempts to access a /g/data or scratch directory without this directive will fail during run time.

Exercise caution if you use symbolic links in your workflows. The -lstorage directive tells PBSPro which directories to mount for the execution of a job, and therefore must refer to an actual project directory on /scratch or /g/data. The best practice is to always use actual target directories in -lstorage directives. Symbolic links which cross file systems, for example, will fail at run time.

Jobs which use less than a full node (Cascade Lake = 48 cpus) will be charged according to the fractional utilisation of node resources, that is, by number of CPUs or amount of node memory *requested*, whichever is larger. Note that charging on Raijin was based on cpu-hours only, without consideration of the memory requested or used.

Projects which use memory-intensive, low-compute workflows may consume SUs more rapidly than expected on Gadi.

Project job resource exemption (for example, wall time extensions) established on Raijin will not be carried across to Gadi. Most user jobs on Gadi will require less wall time than on Raijin. Job resource exceptions on Gadi will need to be compellingly justified.

Raijin will continue to operate with Broadwell and Skylake compute nodes until 31 Dec 2019. Broadwell and Skylake nodes will be offline for relocation and integration with Gadi during January 2020. NCI will advise users when they are available.

Job Charging - Examples

Gadi Cascade Lake node = 48 CPUs, 192 GB memory

1 cpu-hour = 2 service units (SU)

Queue	CPUs	Memory (GB)	Walltime	Charge	Comments
Normal	4	16 GB	5 hours	4 x 5 x 2 = 40 SU	Satisfies 1 CPU <= 4 GB memory.
Normal	8	16 GB	5 hours	8 x 5 x 2 = 80 SU	CPU request dominates.
Normal	8	128 GB	5 hours	32 x 5 x 2 = 320 SU	Memory request dominates. 32 cpus is proportion of node resources.
Normal	8	192 GB	5 hours	48 x 5 x 2 = 480 SU	Memory request dominates. 192GB = 100% of node memory.
Express	8	16 GB	5 hours	8 x 5 x 2 x 3 = 240 SU	CPU request dominates (as above). Express multiplier is x3.

Software

NCI strongly recommends that all users recompile their applications to obtain optimum performance and compatibility with the Gadi run-time environment.

Binary executables from Raijin are expected to be compatible with Gadi *if required dependencies and run-time libraries are available*. Applications which rely on old dependencies are particularly at risk. *Users are strongly encouraged to recompile on Gadi if possible.*

Work is currently in progress porting third-party application software to Gadi. More information is available on the following page: Gadi Software Catalogue.

NCI can assist with local builds of third-party software for individual research groups on Gadi, as on Raijin. Please note that during the transition to Gadi staff time may be limited and software assistance may be deferred until Gadi is fully operational.

The environment modules command will be available on Gadi, and will work in the same manner as on Raijin.

All Python users are encouraged to move to Python 3 as soon as possible. Python 2.7.16 will be provided on Gadi, however this will be the final version of Python 2 installed on the system. Development of Python 2 officially ceased on 01 Jan 2020.

Containers will be available on Gadi, however, NCI staff will need to build the container image to ensure it satisfies security and compatibility criteria. Singularity is the preferred container type at this time. Users who require containers on Gadi should contact NCI user support at help@nci.org.au.

Work is in progress on a container environment to support Raijin backward compatibility on Gadi. This is intended to be a stop-gap solution for projects which require more time to adapt to Gadi. This "Raijin in a container" is expected to be available to users in Q4 and 2020 Q1 for a limited time only - details to be confirmed. Users are again strongly encouraged to rebuild all applications on Gadi for long-term stability and performance.

Raijin Run-time Compatibility Image

NEW 19 Dec 2019

Gadi provides a containerised environment which duplicates the run-time environment available on Raijin. This capability is provided to maintain operational continuity for projects which may require more time to port workflows and tools to the native environment on Gadi.

To use the Raijin compatibility image:

1. Add the following flag to your PBSPro job script: *-limage=raijin*.
2. Modify your job script to request a multiple of 48 CPUs if more than 48 are required.
3. Submit your job on Gadi.

The Raijin compatibility image has several limitations:

- Use of older versions of OpenMPI in the compatibility image will revert to TCP data communications, which will degrade application performance.
- Users who rely on files in Raijin:/projects will need to ensure those files are available on Gadi. The contents of Raijin:/projects is too large to copy for use with the compatibility image.

All projects are encouraged to migrate their applications and workflows to Gadi as soon as possible.

Virtual Desktop Infrastructure (VDI)

NCI's VDI service will continue to be available to users as Gadi enters service in 2019 Q4 and 2020 Q1. Overall VDI functionality is expected to remain unchanged.

The current VDI application software stack will continue to be available. Email help@nci.org.au if you have any questions or requests for new software packages on the VDI.

As is the case now on Raijin, user home directories on VDI will continue to be separate from home directories on Gadi.

VDI-to-Gadi job submission functionality is now available. Please note that this will be implemented as the default option overnight Monday 9 December. For more information about how to use this feature during the transition period see the VDI User Guide https://opus.nci.org.au/display/Help/VDI+User+Guide#VDIUserGuide-4.2.PBS.

Data Collections

Gadi users who require access to NCI data collections should ensure they are members of the required data collection projects.

PBS jobs which read from files in data collections will need to use the requisite job directives to flag collection access, for example, "-lstorage=gdata/<project>". A data collection path on the file system will not be available to a job unless the appropriate PBS directive is provided.

More comprehensive updates on VDI will be provided to users in January-February 2020. If you have specific questions or concerns about VDI please contact NCI User Support - help@nci.org.au.

Training

Transition to Gadi - as presented at the ALCS 2019 Training Day.

Questions?

If you have further questions or concerns about the transition from Raijin to Gadi please contact NCI user support at help@nci.org.au.