Filesystems User Guide

Summary

Name[1]	Purpose	Availability	Quota[2]	Timelimit[3]	Backup
/home/unigrp/user	Irreproducible data eg. source code	Raijin only	2GB /user	none	Yes
/short/projectid	Large data I/O; data maintained beyond one job	Raijin only	72GB /project	none	**No**
/g/data/projectid[4]	Processing of large data files	global		none	**No**
massdata[5]	Archiving large data files	external – access using the mdss command	20GB	none	Yes: 2 copies in two different locations
$PBS_JOBFS	I/O intensive data over the job's lifetime	local to each individual Raijin node	unlimited[6]	duration of job	**No**

1. Each user belongs to at least two Unix groups:
   - `unigrp`– determined by their host institution, and
   - `projectid(s)` – one for each project they are attached to.
2. Increases to these quotas will be considered on a case-by-case basis.
3. Timelimit defines time after which a file is erased on the file system since its most recent access time, as defined by the file access timestamp.
4. Please make sure you specify **#PBS -lother=gdata1** when submitting jobs accessing files in /g/data1. If /g/data1 filesystem is not available, your job will not start. The following command can be used to monitor the status of /g/data1 on Raijin and can be incorporated inside your jobscript:

   ```
   /opt/rash/bin/modstatus -n gdata1_status
   ```

5. Please make sure you specify **#PBS -lother=mdss** when submitting jobs accessing files in MDSS. If the MDSS filesystem is not available, your job will not start. The following command can be used to monitor the status of MDSS on Raijin and can be incorporated inside your jobscript:

   ```
   /opt/rash/bin/modstatus -n mdss_status
   ```

6. Users request allocation of /jobfs as part of their job submission – the actual disk quota for a particular job is given by the jobfs request. Requests larger than **420GB** for Sandybridge (copyq, normal, express), **700 GB** for the GPU queue, **400GB** for everything else (knl, normalbw, expressbw, normalsp, hugemem, gpupascal) will be automatically redirected to /short (but will still be deleted at the end of the job).

Filesystems

/home

- Intended to be used for source code, executables and irreproducible data (input files etc), NOT large data sets.
- Globally accessible from all nodes within a system.
- **Backed up** on a regular basis.
- Quotas apply – use `lquota` on Raijin to see your disk quota and usage, and see the Disk Quota Policy document for details of the ramifications of exceeding the quotas.
- Requests for an increase in your quota will be considered.

/short

- Intended to be used for job data that must live beyond the lifetime of the job.

- Each project has a directory with pathname `/short/projectid` on each compute system. Users connected to the project have `rwx` permissions in that directory and so may create their own files in those areas.
- Globally accessible from all nodes within a system.
- **NOT backed up** – users should save to MDSS system as necessary.
- Quotas apply on a per project basis – use `lquota` or `nci_account` on each machine to see your disk quota and usage. See the Disk Quota Policy below for details of the ramifications of exceeding the quotas.
- Note that there are also limits on the number of files (inodes) that can be owned by a group (project) on /short. This limit and usage can be seen using `nci_account`. An excessive number of inodes causes a number of filesystem problems, hence the limit.
- While file expiry is not yet in place on Raijin, we may consider implementing this in the future – files that have not been accessed within the expiry time frame may be automatically deleted. Users will be notified before this happens.
- Requests for an increase in either the disk quota or the file time limit will be considered.

Warning:

Lots of small I/O to `/short` (or `/home`) can be very slow and impact other jobs on the system.

- Avoid "dribbly" I/O, eg writing 2 numbers from your inner loop. Writing to /short every second is too often.
- Avoid frequent opening and closing of files (or other file operations).
- Use /jobfs (see below) instead of /short for jobs that do lots of file manipulation.

To achieve good I/O performance, try to read or write *binary* files in large chunks (of around 1MB or greater). To find out more details of how to best tune your I/O contact us.

/g/data

- Global Lustre Filesystem /g/data/ – stores persistent data, mounted on Raijin and Tenjin (cloud) nodes.
- **NOT backed up**
- The global Lustre filesystem is designed to be a high-speed file system that is accessible to the major systems operated by NCI:
  - Raijin via FDR Infiniband
  - OpenStack via 10Gb Ethernet (Lustre export over NFS)
- Please make sure you specify **`#PBS -lother=gdataN`** (where N=1, 2 or 3 depending on which /g/data filesystem is being used) when submitting jobs. If `/g/dataN` filesystem is not available, your job will not start. The following command can be used to monitor the status of `/g/dataN` on Raijin and can be incorporated inside your jobscript:

```
/opt/rash/bin/modstatus -n gdataN_status
```

For more information on how to use the above filesystems (/home, /short, and /g/data) see Lustre Basics and Lustre Best Practices.

massdata / MDSS

- Intended to be used for archiving **large** data files particularly those created or used by batch jobs. (It is a misuse of the system to try to store large numbers of small files – please do NOT do this. See the `netcp -t` command option below.)
- **Backed up** and stored in two locations.
- Each project has a directory on the Mass Data Storage System (MDSS) with pathname `/massdata/projectid` on that system. This path CANNOT be directly accessed from Raijin login.
- Remote access to your massdata directory is by the `mdss` utility or the `netcp` and `netmv` commands (see man `mdss/netcp/netmv` for full details.) The `mdss` commands operate on files in that remote directory.

```
mdss    put      – copy files to the MDSS
        get      – copy files from the MDSS
        mk/rmdir – create/delete directories on the MDSS
        ls       – list directories
```

```
netcp/netmv      netmv and netcp generate a script, then submit a
batch request to PBS to copy files (and directories) from Raijin to
the MDSS. In the case of netmv, remove the files from Raijin if the
copy has succeeded.

-t               create a tarfile to transfer
-z/-Z            gzip/compress the file to be transferred

Please use at least the -t option if you wish to archive a
directory structure of numerous small files.
```

- Users connected to the project have `rwx` permissions in that directory and so may create their own files in those areas.
- NOT to be used as an extension of home directories (files changed/removed on the massdata area are not in general recoverable, as there are no back-ups of previous revisions.)
- Currently batch jobs (other than copyq jobs) cannot use the mdss utilities.

> **Note:** always use `-l other=mdss` when using `mdss` commands in `copyq`. This is so that jobs only run when the MDSS system is available.

- Quotas apply – use `nci_account` on the compute machines to see your MDSS quota and usage. See the Disk Quota Policy below for details of the ramifications of exceeding the quotas.
- The **`mdss`** access is intended for relatively modest mass data storage needs. Users with larger capacity storage or more sophisticated access needs should contact us to get an account on the data cluster.

How do I transfer files between massdata and my local machine?

Please note that massdata is not designed for small files. Attempting to store or retrieve files less than a few megabytes will result in extremely poor performance for all users. If you wish to store lots of small files to massdata, please use a utility such as tar to combine them into a single, larger file.

To transfer files between massdata and your local machine, we recommend this two-step workflow.

/jobfs

- Intended for I/O intensive jobs providing scratch space only for the lifetime of the job.
- Only accessible on the execution node (ie. not on a login node).
- Allocated by using the `-ljobfs=??` option to `qsub`, eg. `-ljobfs=5GB` requests 5 GB. Use integers and units of MB or GB (not case-sensitive). The maximum request must be less than or equal to the local disc storage of the node/s
- **NOT backed up**
- /jobfs directories are associated with a currently running job and are automatically deleted at the jobs completion.
- Your batch job can access its jobfs via the environment variable `PBS_JOBFS`. Jobs spanning multiple nodes with local JOBFS space on each node should use the `/opt/pbs/bin/pbsdsh -N` ... command in the batch script to act on all JOBFS directories, e.g.

```
/opt/pbs/bin/pbsdsh -N ls $PBS_JOBFS
```

- For example, if you want local copies of files generated before the current batch run you can do the following to make them available on each nodes' jobfs area.

```
/opt/pbs/bin/pbsdsh -N cp original_file $PBS_JOBFS
```

> Don’t put any quotes around the command issued under pbsdsh.

It is not possible to use the `netmv` command to save data which exists on a `/jobfs` filesystem – files must be copied to `/short` first.

Users who are dealing with large files in **large chunks** (i.e. > 1 MB reads and writes) have a number of options available to them to improve their I/O performance. Contact us for assistance in choosing the best options.

As well as the generally available filesystems listed above, there may be high performance filesystems, utilities or techniques available to improve the I/O performance of your workload. Please contact us if you think this may be relevant to you.

Disk Quota Policy

To avoid the disk usage of one or two users or projects adversely affecting other users, we have implemented a policy of checking disk usage on a regular basis and taking the actions detailed below.

View your disk quotas and current usages on Raijin by running the commands:

`nci_account` **for mdss, /short and /g/data usage**

`lquota` **for /home, /short and /g/data usage**

Administrative disk usage limits

Administrative disk usage limits are imposed on the file systems /home, /short, /g/data and the mdss file storage area.

Users should ensure they have enough space available before starting jobs. Note that adjustments to quotas, on a temporary or a long-term basis, will be made given reasonable justification.

The consequences of a project or user exceeding these administrative disk usage limits are outlined in the following table:

File System	Hard Limit Exceeded
/home	Once quota exceeded, new files will not be able to write to the /home filesystem. No email reminder is currently implemented.
/short /g/data	**Except for the copyq, queued jobs will not start. Currently running jobs will be allowed to continue. The copyq will remain available to allow archiving.** All members of the project will be sent an email to this effect once a day. Queued jobs will be released when the quota system next detects that the disk usage is back under the quota limit. This should occur within 30 minutes of the project getting back under the disk quota limit. If the usage exceeds 90% of the quota, all members of the project will be sent an email to this effect once a month as a reminder. This will not affect the queue until the usage reaches 100% of the limit.
massdata	**Except for the copyq, queued jobs will not start. Currently running jobs will be allowed to continue. The copyq will remain available to allow archiving.** All members of the project will be sent an email to this effect. Due to the relative "expense" of reassessing file space usage under MDSS, reactivation of your queued job will not automatically occur for up to 24 hours after the removal of files. Contact help@nci.org.au to request manual reactivation.

It is important to understand that these limits are **not** the native filesystem quotas available on most Unix filesystems – the behaviour and actions described above are quite different. Exceeding administrative limits should not cause job failures nor stop users from managing files etc. However users should be aware that real filesystem "hard quotas" are also imposed but with a limit much higher than the administrative limits discussed above. If a user or project reaches the relevant filesystem hard quota, jobs *are* likely to fail due to write operations failing.

Note: There are two conventions for defining digital storage capacities, namely, base 2 (1KB = 1024 bytes) and base 10 (1KB = 1000 bytes), and both are in common use. At NCI, the following definitions are used when referring to digital storage capacities (filesystem quota limits, usage, etc.):

1 KiloByte (KB)	2 ^ 10	1024 bytes
1 MegaByte (MB)	2 ^ 20	1048576 bytes
1 GigaByte (GB)	2 ^ 30	1073741824 bytes

1 TeraByte (TB)	2 ^ 40	1099511627776 bytes
1 PetaByte (PB)	2 ^ 50	1125899906842624 bytes

Utilities like `nci_account` and `lquota` use the above formula (base 2), terms and abbreviations. When using other standard OS utilities for measuring storage capacity/usage, verify if the output is based on base 2 or base 10 calculations, irrespective of the terms or abbreviations used.