UG2.5: Data storage and FileSystems

In this page:

Data Storage architecture

All HPC systems share the same logical disk structure and file systems definition.

The available storage areas can be

- **temporary** (data are cancelled after a given period);
- **permanent** (data are never cancelled or cancelled only a few months after the "end" of the project);

they can also be

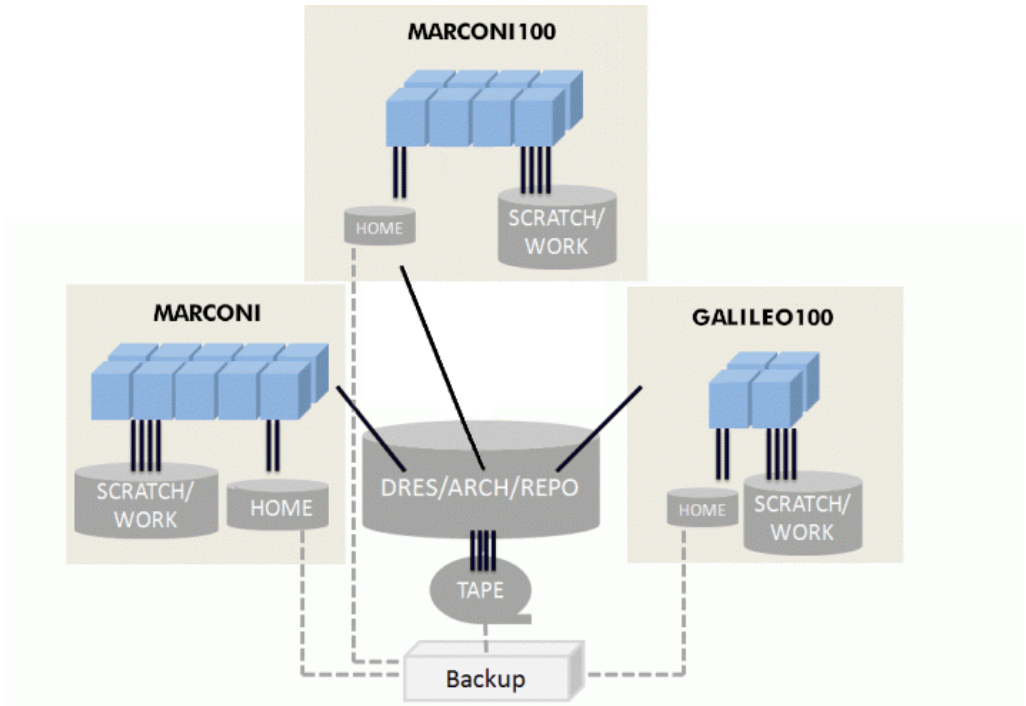
- **user specific** (each username has a different data area);
- **project specific** (accessible by all users within the same project).

Finally, the storage areas can be:

- **Local** (specific for each system);
- **Shared** (the same area can be accessed by all HPC systems)

The available data areas are defined through predefined "environment variables": $HOME, $CINECA_SCRATCH, $WORK.

Important: It is the user's responsibility to backup your important data. We only guarantee a daily backup of data in the $HOME area.



$HOME: permanent/backed up, user specific, local

$HOME is a **local** area where you are placed after the login procedure. It is where system, and user applications store their dot-files and dot-directories (.nwchemrc, .ssh, ...) and where users keep initialization files specific for the systems (.cshrc, .profile, ...). There is a $HOME area for each username on the machine.

This area is conceived to store programs and small personal data. It has a **quota** of 50 GB. Files are **never deleted** from this area. Moreover, they are guaranteed by **daily backups**: if you delete or accidentally overwrite a file, you can ask our Help Desk (superc@cineca.it) to restore it. A maximum of 3 versions of each file is stored as a backup. The last version of the deleted file is kept for two months, then definitely removed from the backup archive. File retention is related to the life of the username; data are preserved until the username remains active.

$WORK: permanent, project specific, local

$WORK is a scratch area for **collaborative work** within a given project. File retention is related to the life of the project. Files in $WORK will be conserved up to 6 months after the project end, and then they will be cancelled. Please note that there is **no back-up** in this area.

This area is conceived for hosting large working data files since it is characterized by the high bandwidth of a parallel file system. It behaves very well when I/O is performed accessing large blocks of data, while it is not well suited for frequent and small I/O operations. This is the main area for maintaining scratch files resulting from batch processing.

There is **one $WORK area for each active project** on the machine. The default **quota** is 1 TB per project, but extensions can be considered by the Help Desk (mailto: superc@cineca.it) if motivated. The owner of the main directory is the PI (Principal Investigator) of the project. All collaborators are allowed to read/write in there. Collaborators are advised to create a personal directory in $WORK for storing their personal files. By default, the personal directory will be protected (only the owner can read/write), but protection can be easily modified, for example by allowing write permission to project collaborators through chmod command. This second approach does not affect global files security.

The (**chprj - change project**) command makes it easier to manage the different WORK areas for different projects, see here for details.

$FAST: permanent, project specific, local (LEONARDO ONLY)

$FAST is a scratch area for **collaborative work** within a given project. File retention is related to the life of the project. Files in $FAST will be conserved up to 6 months after the project end, and then they will be cancelled. Please note that there is **no back-up** in this area.

This area is conceived for hosting working data files whenever the I/O operations constitute the bottleneck for your applications. It behaves well both when I/O is performed accessing large blocks of data, and for frequent and small I/O operations. Due to the limited size of the area, the main space for maintaining the data resulting from batch processing is the corresponding $WORK area.

There is **one $FAST area for each active project** on the machine. The fixed **quota** is 1 TB per project, and due to the total dimension of the storage, extensions cannot be considered. The owner of the main directory is the PI (Principal Investigator) of the project. All collaborators are allowed to read/write in there. Collaborators are advised to create a personal directory in $FAST for storing their personal files. By default, the personal directory will be protected (only the owner can read/write), but protection can be easily modified, for example by allowing write permission to project collaborators through chmod command. This second approach does not affect global files security.

The (**chprj - change project**) command makes it easier to manage the different FAST areas for different projects, see here for details.

$CINECA_SCRATCH: temporary , user specific, local

This is a **local temporary** storage conceived for temporary files from batch applications. There are important differences with respect to $WORK area. It is **user specific** (not project specific). By default, file access is closed to everyone, in case you need less restrictive protections, you can set them with **chmod** command.

On this area, a **periodic cleaning procedure** could be applied, with a normal retention time of 40 days: files are daily cancelled by an automatic procedure if not accessed for more than 40 days. In each user's home directory ($HOME) a file lists all deleted files for a given day. Please notice that on G100 the variable is $SCRATCH.

```
CLEAN_<yyyymmdd>.log
      <yyyymmdd> = date when files were cancelled
```

There is one $CINECA_SCRATCH area for each username on the machine.

$CINECA_SCRATCH does **not have any disk quota**. Please be aware that on Galileo100 and Marconi100 clusters, to prevent a very dangerous filling condition, a **20TB disk quota** will be **temporarily imposed** to all users when the global quota area reaches **88% of occupancy**; this disk quota will be removed when the global occupancy lowers back to normal.

To verify if and how the cleaning procedure is active on a given cluster, check the Mott-of-the-Day.

$TMPDIR: temporary, user specific, local

Each compute node is equipped with a **local area** whose dimension differs depending on the cluster (please look at the specific page of the cluster for more details).
When a job starts, a **temporary area** is defined on the storage **local to each compute node**. On MARCONI and GALILEO100:

TMPDIR=/scratch_local/slurm_job.$SLURM_JOB_ID

Differently from the other CINECA clusters, on LEONARDO the job's temporary area is managed by the slurm job_container/tmpfs plugin, which provides an equivalent job-specific, private temporary file system space, with private instances of /tmp and /dev/shm in the job's user space:

TMPDIR=/tmp

visible via the command "df -h /tmp". If more jobs share one node, each one will have a **private /tmp** in the job's user space. The tmpfs are removed at the end of the job (and all data will be lost).

Whatever the mechanism, the TMPDIR can be used **exclusively** by the job's owner. During your jobs, you can access the area with the (local) variable $TMPDIR. In your sbatch script, for example, you can move the input data of your simulations to the $TMPDIR before the beginning of your run and also write on $TMPDIR your results. This would further improve the I/O speed of your code.

However, the directory is **removed at the job's end;** hence always remember to save the data stored in such area to a permanent directory in your sbatch script at the end of the run. Please note that the area is located on local disks, so it can be accessed only by the processes running on the specific node. For multinode jobs, if you need all the processes to access some data, please use the shared filesystems $HOME, $WORK, $CINECA_SCRATCH.

Differently from the other CINECA clusters, thanks to the job_container/tmpfs plugin the local storage is considered a "resource" on LEONARDO, and can be explicitly asked **on the diskful nodes only** (DCGP and serial nodes) via the sbatch directive or srun option "-gres=tmpfs:XX" (see the specific Disks and Filesystems section on LEONARDO's User Guide for the allowed maximum values). For the same reason, the requested amount of gres/tmpfs resource **contributes to the consumed budget**, changing the number of accounted equivalent core hours, see the dedicated section on the Accounting on CINECA clusters.

$DRES: permanent, shared (among platforms and projects)

This is a **repository area** for collaborative work among different projects and across platforms. You need to explicitly ask for this kind of resource: it does not come as part of a project (mailto: superc@cineca.it).

File retention is related to the life of DRES itself. Files in DRES will be conserved up to 6 months after DRES completion, then they will be cancelled. Several types of DRES are available, at present:

- **FS**: normal filesystem access on high throughput disks, shared among all HPC platforms (mounted only on login nodes);
- **ARCH**: magnetic tape archiving with a disk-like interface via LTFS;
- **REPO**: smart repository based on iRODS.

This area is conceived for hosting data files to be used by several projects, in particular, if you need to use them among different platforms. For example, you would need to post-process data produced on Marconi, taking advantage of the visualization environment of Galileo; or you would require a place for your data from experiments to be processed by several related projects. **This filesystem is mounted only on login nodes and on the nodes of the "serial" partitions of all HPC clusters in Cineca** (e.g. bdw_all_serial on Marconi, gll_all_serial on Galileo) - please make use of batch jobs on the serial partitions for the rsync transfers of great amounts of data, or to gridftp clients. As a consequence, you have to move data from $DRES to $CINECA_SCRATCH or $WORK area in order for them to be seen by the compute nodes.

WARNING: DRES of type ARCH have a limit in the number of inodes available: 2000 inodes for each TB of quota. This means that no more than 2000 files can be stored in 1 TB of disk space. It is then recommended that you compress your files for storage purposes and that the dimension of each file stored should be in an average quota of 500MB. DRES of types FS and REPO do not have this limitation.

Files stored in DRES of type FS or ARCH will be moved automatically to tape storage, when specific conditions are met:

- for ARCH: files are older than 3 months and bigger than 50 MB
- for FS: files are older than 3 months and bigger than 100 MB

The Data movement is transparent for the user. Only physical support changes, while the logical environment will not be affected (this means that you can reach data stored in tape using the same path you used for storing it in the first place)

Backup policies

Daily backups guarantee the $HOME filesystem. In particular cases, a different agreement is possible: contact the HPC support (superc@cineca.it) for further details.

The backup procedure runs daily, and we preserve a maximum of three different copies of the same file. Older versions are kept for 1 month. The last version of deleted files is kept for 2 months, then definitely removed from the backup archive.

Environment variables

$HOME, $WORK, $CINECA_SCRATCH ($SCRATCH on G100) and $DRES environment variables are defined on all HPC Systems, and you can access these areas simply using those names:

```
cd $HOME
cd $CINECA_SCRATCH
cd $WORK
cd $DRES
```

You are strongly encouraged to use these environment variables instead of full paths to refer to your scripts and codes data.

Summary

$CINECA_SCRATCH	**$WORK**	**$DRES**
Created when username has granted access. Each username has its own area (and only one).	Created when a project is opened. Each project has its own area. All collaborators can write. Each user has as many $WORK areas as active projects.	Created on request. Not connected to a specific project. Data are accessible by all the platforms but visible only to login nodes and nodes of the serial partition. Compute nodes do not see $DRES area.
A clean-up procedure is active. Files older then 40 days are cancelled daily. No backup	Data are preserved up to few months after the end of the project. No backup.	Data are preserved up few months after the expiring date No backup.
No quota.	Default quota of 1 TB. Motivated requests for quota increase will be taken into account.	Quota based on the needs. A limit of 2000 files each TB is present.
By default files are public (read only). The user can change the permission (chmod) and make files private. It is not possible to restrict access to the group (all usernames share the same mail unix group).	By default files are private. The user can change the permission (chmod) and make files visible (R o R/W) to project collaborators.	Same as $WORK area.

Note: the $FAST area is not reported in the Summary since it is defined on Leonardo cluster only.

What to use when...

Data are critical, **not so large**, I want to be sure to preserve them.	**$HOME** is the right place. Pay attention to the 50 GB quota limit for each user on this space.
Large data to be **shared with collaborators** of my project.	**$WORK** is the right place. Here each collaborator can have his own directory. He can open it for reading or even writing and be sure, at the same time, that data are not public. People not included in the project will not be able to access the data. Moreover, data are preserved up to a few months after the project's end.
Data to be **shared with other users**, not necessarily participating in common projects	**$CINECA_SCRATCH** is the right place.

Data to be maintained even beyond the end of the project. I'll use the data on different CINECA hosts	**$DRES** is a better solution.
Data to be shared among different **platforms**	**$DRES** allows this.

Monitoring the occupancy

A command is available on all HPC systems to check the status of the occupancy of all the areas accessible by the user and eventually the presence of a quota limit.

Option "-h" shows the help for this command.

```
cindata
```

If you are a DRES user, the output of cindata command will contain similar lines:

```
USER           AREADESCR                        AREAID                  FRESH    USED      QTA    USED%    aUSED
    aQTA    aUSED%

myuser00  /gpfs/work/<AccountName>       galileo_work-Acc-name          9hou     114G      --      --
%      14T      30T    48.8%

myuser00  /gpfs/scratch/                 galileo_scr                    9hou     149G      --      --
%     341T     420T    81.2%

myuser00  /galileo/home                  galileo_hpc-home               9hou     5.7
G      50G     11.4%    16T       --      --%

myuser00  /gss/gss_work/DRES_myAcc       work_OFFLINE-DRES_myAcc-FS     9hou     2.9G      --      --
%      11T      15T    73.3%

myuser00  /gss/gss_work/DRES_myAcc       work_ONLINE-DRES_myAcc-FS      9hou     1.2T      --      --%     2.8
T       4T    70.0%
```

This may be tricky to interpret. OFFLINE area is the DRES data that has been stored on tape, after 3 months of storage (see DRES description above). ONLINE area is the DRES data still in FS or ARCH area. The total quota of storage capability assigned to your DRES is indicated in the "aQTA" parameter of the line "OFFLINE". When DRES is empty, the corresponding value of the line "ONLINE" will be the same as "OFFLINE", but when files begin to be moved it will decrease, while the "aUSED" parameter of OFFLINE will increase accordingly. It means that you have less space that you can use to store your data since some of them already used space has been moved to tape. Similarly, deleting offline data will decrease the aUSED parameter in OFFLINE, and increase the aQTA parameter in ONLINE of the same amount. The formula to remember is:

TOTAL DRES STORAGE = aQTA-OFF = aQTA-ON + aUSED-OFF

An additional tool for monitoring the disk occupancy is also available on GALILEO100 and LEONARDO, since the command is working only on Lustre filesystems

```
cinQuota
```

The output of the command will contain the following information:

```
--------------------------------------------------------------------------------------------------------------------------
    Filesystem                                 used          quota       grace       files
--------------------------------------------------------------------------------------------------------------------------
 /g100/home/userexternal/myuser00           22.66G       50G       -          194295
 /g100_scratch/userexternal/myuser00        1.955T         0k       -           41139
```

```
/g100_work/<AccountName>                366.3G          1T      -         548665

-------------------------------------------------------------------------------------------------------------------------------------------------
```

The tool is available in the module *cintools,* which is automatically loaded in your environment. However, the module can be unloaded as all the other modules (Modules).

File permissions

$WORK and $DRES are environmental variables automatically set in the user environment.

$WORK variable points to a directory (fileset) specific for one of the user projects:

```
/gpfs/work/<account_name>
```

$DRES variable points to space where all of the dres are defined:

```
/gss/gss_work/
```

In order to use a specific dres type the following path:

```
$DRES/<dres_name>
```

The owner of the root directory is the "Principal Investigator" (PI) of the project or the "owner" of the DRES, the group corresponds to the name of the project or the name of the DRES. Default permissions are:

```
own: rwx
group: rwx
other: -
```

In this way, all project collaborators, sharing the same project group, can read/write into the project/dres fileset, whereas other users can not.

Users are advised to create a personal subdirectory under $WORK and $DRES. By default, files into the subdirectory are private, but the owner can easily share the files with the other collaborators by opening the subdirectory:

```
chmod 777 mydir
chmod 755 mydir
```

Since the $WORK/$DRES fileset is closed to non collaborators, the data sharing is active only among the project collaborators.

Pointing $WORK to a different project: the chprj command

The user can modify the project pointed to by the variable $WORK using the "change project" command.

To list all your accounts (both active or completed) and the default project:

```
chprj -l
```

To set $WORK to point to a different <account_no> project:

```
chprj -d <account_no>
```

More details are in the help page of the command:

```
chprj -h
chprj --help
```

On LEONARDO only: The command applies to the $FAST variable as well.

Endianness

Endianness is the attribute of a system that indicates whether integers are represented from left to right or right to left. At present, all clusters in Cineca are "little-endian".

Managing your data

A comprehensive discussion on how to manage your data can be found in a specific document.