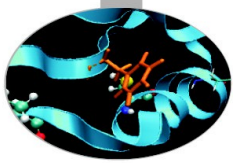
CINECA SCAI
SuperComputing Applications and Innovation



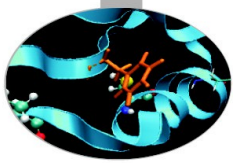
Data Resources and Data Transfer

Author: Alessandro Grottesi
Speaker: Alessandro Marani

13/06/2016

PRACE CINECA

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) or
- permanent (data are never cancelled or cancelled only at the "end");

they can also be:

- user specific (each username has a different data area) or
- project specific (defined for each project - account_no).

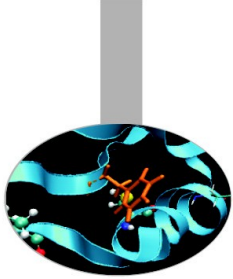
Finally the storage areas can be:

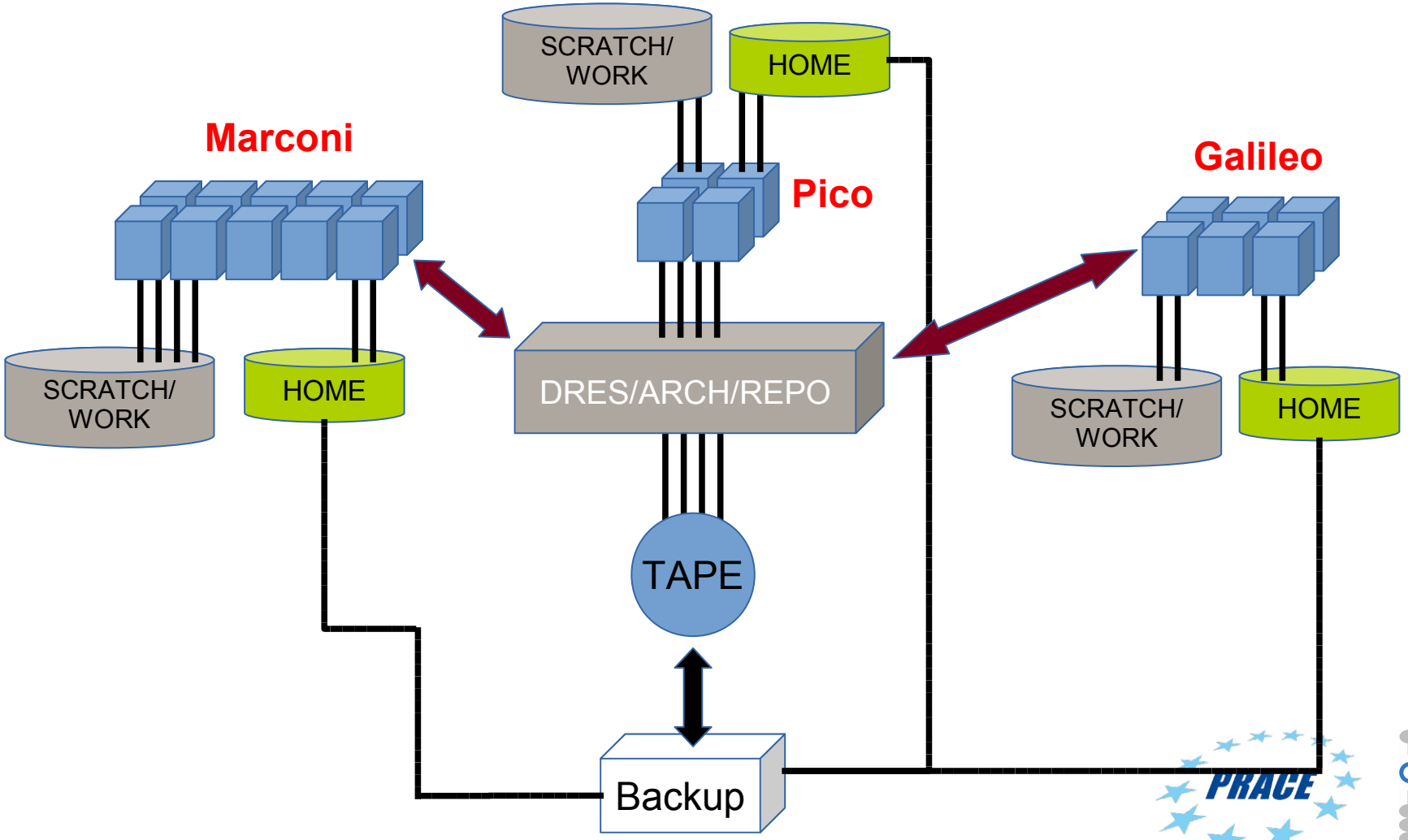
- Local: that means they are specific for each system.
- Shared: The same area can be accessed by all HPC systems

The available data areas are defined through predefined "environment variables":

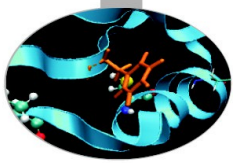
- $HOME
- $CINECA_SCRATCH
- $WORK
- $TAPE
- $DRES

Data storage architecture overview





Accessing environment variables



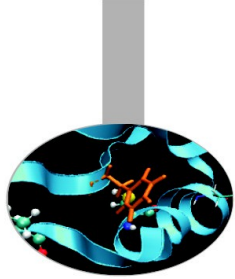
The environment variables $HOME, $WORK and $CINECA_SCRATCH are defined on all HPC Systems, and you can access these areas simply using those names:

```
>cd $HOME
>cd $CINECA_SCRATCH
>cd $WORK
```

Since the storage areas are based on GPFS, you cannot use the usual "quota" project to show quotas and occupancies. A command is available on all HPC systems to check the quota and the occupancy of common data areas visible from the cluster:

```
>cindata -u <username>
>cindata -a
>cindata -h
```

Checking user's storage areas

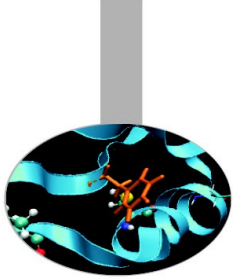


Example of cindata usage:

```
[amarani0@node166 ~]$ cindata -u aparodi0
```

USER	AREADESCR	AREAID	FRESH	USED	QTA	USED%	aUSED	aQTA	aUSED%
aparodi0	/gpfs/scratch/	galileo_scr	2day	0	--	--%	215T	299T	72.2%
aparodi0	/gpfs/work/IscrC_TEWRHYS	galileo_work-IscrC_TEWRHYS	6hou	160M	--	--%	1.0T	1T	100.0%
aparodi0	/galileo/	galileo_hpc	2hou	2.6G	--	--%	6.7T	8.2T	82.4%
aparodi0	/galileo/home	galileo_hpc-home	2hou	2.6G	50G	5.1%	5.5T	--	--%
aparodi0	/gpfs/work/IscrC_SCENE	galileo_work-IscrC_SCENE	6hou	454G	--	--%	454G	1T	44.4%
aparodi0	/gpfs/work/	galileo_work	6hou	454G	--	--%	149T	299T	50.1%
aparodi0	/gss/gss_work/DRES_SCENE	work_ONLINE-DRES_SCENE-ARCH	3hou	875G	--	--%	875G	3T	28.5%
aparodi0	/gss/gss_work/	work_ONLINE	3hou	875G	--	--%	1.2P	1.4P	84.7%
aparodi0	/gss/gss_work/	work_OFFLINE	3hou	2.1T	--	--%	223T	--	--%
aparodi0	/gss/gss_work/DRES_SCENE	work_OFFLINE-DRES_SCENE-ARCH	3hou	2.1T	--	--%	2.1T	3T	71.4%

Data Resources @ CINECA



$DRES (**Data Res**ource): permanent, shared (among platforms and projects)

You need to ask for this kind of resource explicitly, it does not come as part of a project (mailto: superc@cineca.it).

The retention of the files is related to the life of the DRES itself. Files in DRES will be conserved up to 6 months after the DRES completion (which is independent from the project completion), then the DRES folder will be cancelled.

Several types of DRES are available:

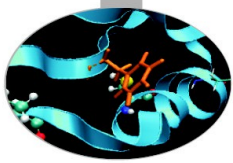
- **FS**: normal filesystem access on hight throughput disks, shared among all HPC platforms (mounted only on the login nodes). This is a repository area for **collaborative work** among different projects and across platforms.
- **ARCH**: magnetic tape archiving with a disk-like interface via LTFS. This is a repository area for **long-term archiviation** of important files.
- **REPO**: smart repository based on iRODS

$DRES is conceived for hosting data files to be used by more than one project, in particular if you need to use them from different platforms.

For example, you would need to post-process data produced on FERMI, taking advantage from the visualization environment of PICO; or you would require a place for your data from experiments to be processed by several related projects.

This filesystem is mounted on the login nodes of FERMI and GALILEO and on all nodes of PICO.

Data Resources @ CINECA



A DRES directory can be created on request of an user. It's only-storage resource, based on GSS technology. It's characterized by:

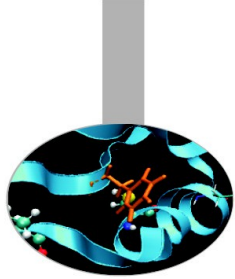
- an Owner (a user who owns that resource and is allowed to manage it),
- some possible Collaborators (users who can access the resource but not manage it)
- a validity time, an extension and a storage type
- some possible computational Projects (all collaborators of the project can access the resource)

DRES files will be moved in the tape storage after certain conditions are met:
- ARCHIVE: the files are older than 3 months and bigger than 50 MB
- FILESYSTEM: the files are older than 3 months and bigger than 100 MB

This policy may be subject to change!!!

BEWARE OF ENDIANNESS!! If you are using a DRES to share with MARCONI files producted on FERMI, keep in mind that the former cluster is little-endian while the latter is big-endian. Proper file conversion may be required to use binary data files on Marconi.

Data Resources @ CINECA



$TAPE: permanent, user specific, shared

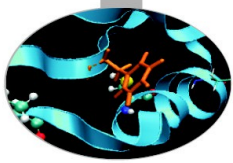
This is a small archive area conceived for saving personal data on magnetic media.

The list of file is maintained on disks, the file content is moved automatically to tape using the LTFS technology. This archive space is not created by default for all users, you have to ask for it, by specifying the maximum space required (mailto: superc@cineca.it).

This filesystem is mounted on the login nodes of FERMI and GALILEO and on all nodes of PICO.

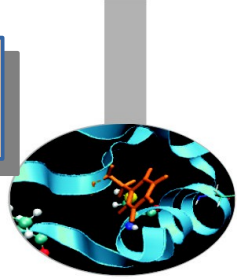
The retention of the files is related to the life of the username, data are preserved until the username remains active.

Examples of data usage



data are critical, not so large, I want to make sure to preserve them safely.	**$HOME** is the right place. The only limitation is the quota limit on this area, usually several GB, you can ask to enlarge up to 50GB.
large data to be shared with all collaborators of my project	**$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.
data to be shared with other users, not necessarily sharing the same common projects	**$CINECA_SCRATCH** is the right place.
data to be maintained even beyond the project. I'll use the data on CINECA hosts	**$DRES** repo or archive or **$TAPE** are the possible solutions.
data to be shared among different platforms	**$DRES** file system

Data Transfer: basic tools



scp is useful to move small amount of data, since it is not optimised. Typically, to copy all files named *.bin in my local pc to a the remote.host in a dir named my_directory, type:

```
>scp -r *.bin myusername@remote.host:/my_directory/
```

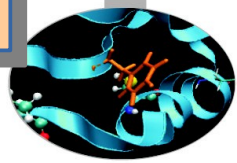
rsync is useful if you need to syncronize the content of a local and a remote dir on a remote host:

```
>rsync -avHS -r *.bin myusername@remote.host:/my_directory/.
```

sftp is a tool to get/put files to/from a remote directory on a remote host:

```
>sftp myusername@remote.host
>...
>mget *.bin
```

Data Transfer: more on rsync



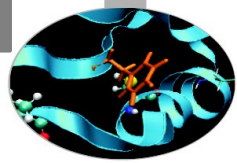
Below is a template of a job set to run in the archive queue, that uses rsync for data transfer:

```
#!/bin/bash
#PBS -l walltime=4:00:00
#PBS -l select=1:mpiprocs=1
## PBS -N myjob
#PBS -o rsync$job_id.out
#PBS -e rsync$job_id.err
#PBS -q archive

. /cineca/prod/environment/module/3.1.6/none/init/bash
cd $PBS_O_WORKDIR

sourc=/gpfs/scratch/.....       ## do not put the / here
dest=/shared/project/data/..... ## put the / here

rsync -avHS -r $sourc $dest  > logrsync.out
```

If your data copy requires more than 4 hours you can run a multisteps job. Each step of this job has up to 4 hours of time limit and will copy the data starting from the file where the previous step was interrupted:

```
qsub -W depend=afterok:JOBID job.sh
```

Data Transfer: GridFTP clients



GridFTP is a very efficient protocol for transferring data, it enhances the standard ftp service making it more reliable and faster. It is being developed by the Globus alliances and is part of an open-source toolkit for HPC applications management.

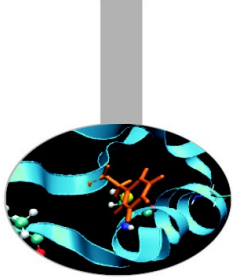
globus-url-copy is a scriptable command line tool that can do multi-protocol data movement supporting GridFTP. It is mainly for Linux/Unix users. It is possible to use globus-url-copy in these cases:

- User Local PC <==> CINECA HPC Cluster
- User Local PC <==> iRODS repository
- CINECA HPC Cluster A <==> CINECA HPC Cluster B
- CINECA HPC Cluster <==> iRODS repository

To use globus-url-copy tool, you must have a **valid x.509 personal certificate.**

Please refer to:
https://wiki.u-gov.it/confluence/display/SCAIUS/globus-url-copy+client

iRODS-based REPO

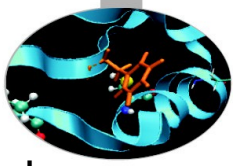


REPO is a Cineca service, implemented through iRODS (Integrated Rule-Oriented Data System), for the management of long lasting data.

This service aims to store and maintain scientific data sets and it is built in a way that allows a user to safely back-up data and at the same time manage them through a variety of clients, such as web browser, graphical desktop and command line interfaces.

It relies on plain filesystems to store data files and on databases to store the metadata. The service's architecture has been carefully designed to scale to millions of files and petabytes of data, joining robustness and versatility.

iRODS-based REPO



The complete set of features available to manage data via iRODS can be summarised as follows:

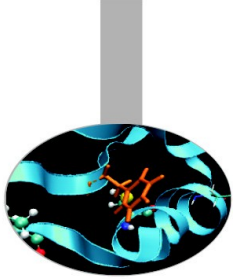
Upload/Download: the system supports high performance transfer protocols like GridFTP,
or iRODS multi-threads transfer mechanism.

- the GridFTP interface for iRODS is at address: data.pico.cineca.it:2811.
- The iRODS commands, official documentation available at https://docs.irods.org/master/icommands/user/

Metadata management: each object can be associated to specific metadata represented as triplets (name,value,unit), or simply tagged and commented. This operation can be performed at any time, not just before the first upload.

Preservation: the long-term accessibility is granted by means of a seamless archiving process,
which is able to move the collections of data from the on-line storage space to a tape based off-line space and back, according to general or per-project policies.

iRODS-based REPO



- **Stage-in/stage-out**: the service is enabled to move data sets, requested as input for computations, towards the HPC machines' local storage space, commonly named “scratch”, and backwards as soon as the results are available.

- **Sharing**: the capability to share single data objects or whole collections is implemented via a unix-like ownership model, which allows to make them accessible to single users or groups. Moreover a ticket based approach is used to provide temporary access tokens with limited rights.

- **Searching**: the data are indexed and the searches can be based on the objects location or on the associated metadata.