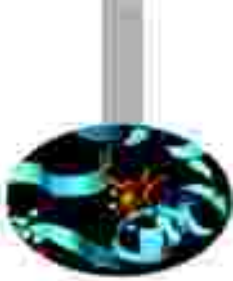
CINECA SCAI
SuperComputing Applications and Innovation

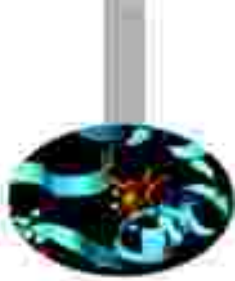


Data Managment @ CINECA

Alessandro Grottesi

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
- prj specific (defined for each project - account_no).

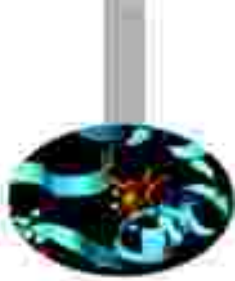
Finally the storage areas can be:

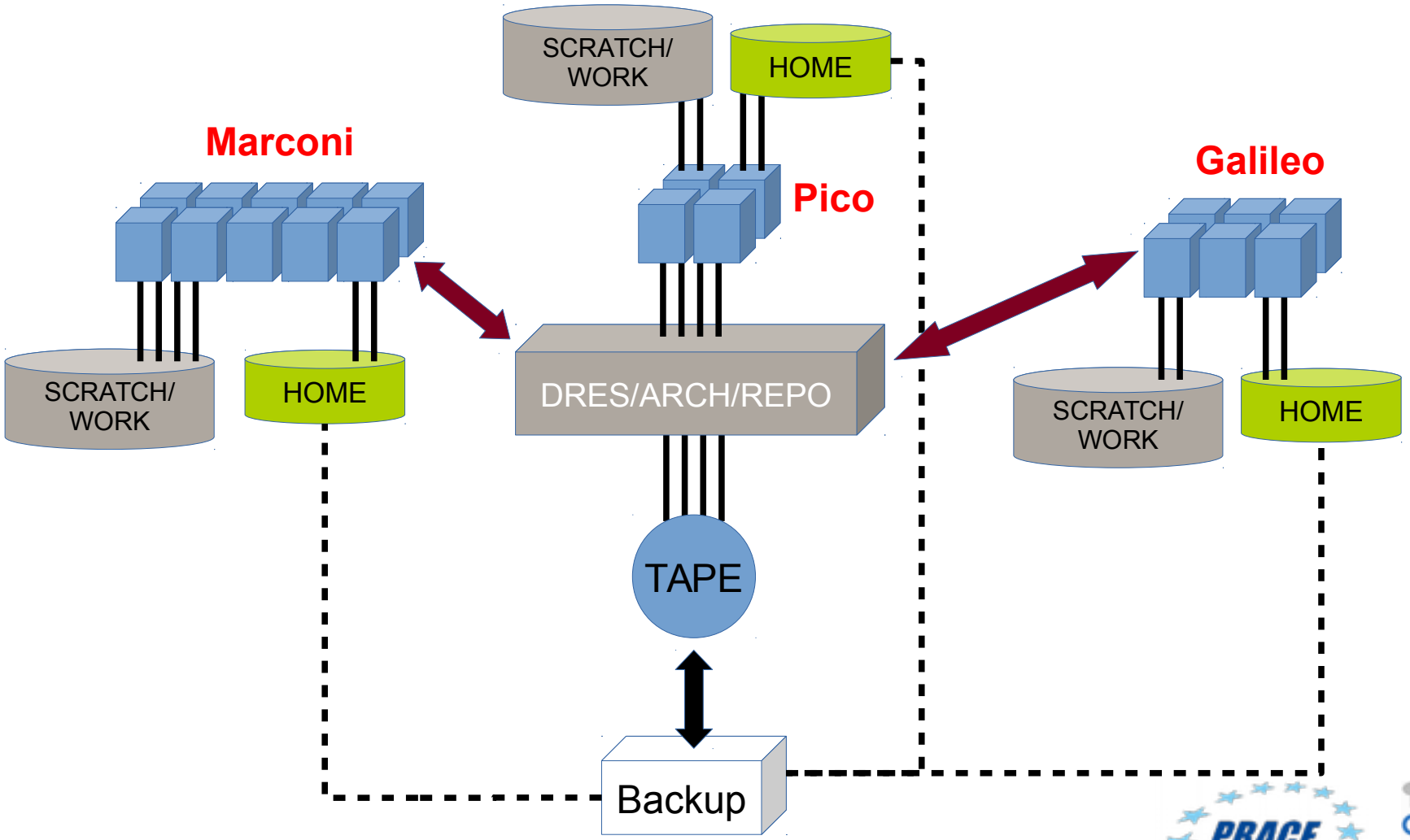
- Local: that means they are specific for each system.
- Shared: The same area can be accessed by all HPC systems

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

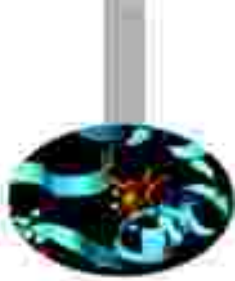
- $HOME
- $CINECA_SCRATCH
- $WORK.

Data storage architecture overview





Data Resources @ CINECA



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

This is a repository area for **collaborative work** among different projects and across platforms. For example, you would need to post-process data produced on Marconi, 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.

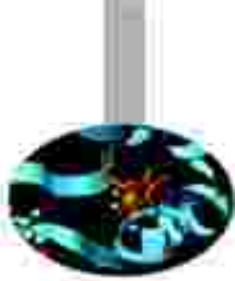
The retention of the files is related to the life of the DRES itself. Files in DRES will be conserved up to 6 months after the DRES completion, then they will be cancelled.

Several types of DRES are available:

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

This filesystem is mounted on the login nodes of MARCONI 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:

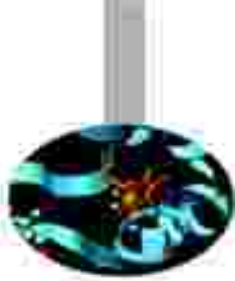
- an Owner (a user who owns that resource and is allowed to manage it),
- some possible Collaborators (users who can access the resource but not manage it)
- a validity time, an extension and a storage type
- some possible computational Projects (all collaborators of the project can access the resource)

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!!!

Data Storage @ CINECA



$TAPE: permanent, user specific, shared

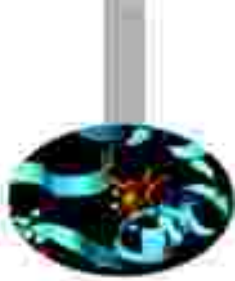
This is an **archive** area conceived for saving personal data on magnetic media.

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 MARCONI and GALILEO and on all nodes of PICO.

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

Data Storage @ CINECA

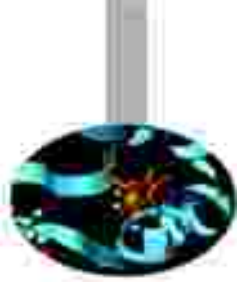


$TAPE: permanent, user specific, shared

Caveat: it is advisable to create single compressed tar files to archive large amount of data so as to optimize the usage of TAPE resource and avoid long retrieve time. To this purpose, users can pack multiple directories into a compressed tar files:

tar -cvf my_data.tar dir1/	creates data.tar archive
tar -rvf my_data.tar dir2/ dir3/ dir4/..	append data to archive
gzip my_data.tar	gzip archive

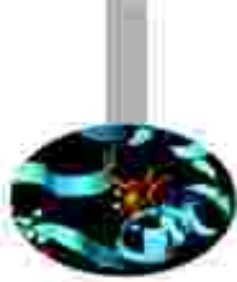
cindata



$ cindata

USER	AREADESCR	AREAID	FRESH	USED	QTA	USED%	aUSED	aQTA	aUSED%
sbuenomi	/gss/gss_work/cin_staff	work_OFFLINE-cin_staff-__UNCAT__	5hou	0	--	--%	0	--	--%
sbuenomi	/marconi_work/cin_tmpaccM1	marconi_work-cin_tmpaccM1	3wee	0	--	--%	0	1T	0.0%
sbuenomi	/gss/gss_work/	work_OFFLINE	5hou	0	--	--%	223T	--	--%
sbuenomi	/gss/gss_work/DRES_cin_fs	work_OFFLINE-DRES_cin_fs-FS	5hou	0	--	--%	0	4.9T	0.0%
sbuenomi	/gss/gss_work/cin_staff	work_ONLINE-cin_staff-__UNCAT__	5hou	1.5M	--	--%	11T	30T	37.3%
sbuenomi	/marconi_work/cin_staff	marconi_work-cin_staff	3wee	30M	--	--%	4.7T	15T	31.5%
sbuenomi	/marconi_work/	marconi_work	3wee	30M	--	--%	38T	7.0P	0.5%
sbuenomi	/gss/gss_work/DRES_cin_fs	work_ONLINE-DRES_cin_fs-FS	5hou	4.3G	--	--%	1.9T	4.9T	38.9%
sbuenomi	/gss/gss_work/	work_ONLINE	5hou	4.3G	--	--%	1.3P	1.4P	92.6%
sbuenomi	**/marconi/home**	marconi_hpc-home	**2day**	**7.6G**	**50G**	**15.1%**	1.3T	--	--%
sbuenomi	/marconi/	marconi_hpc	2day	7.6G	--	--%	1.6T	200T	0.8%
sbuenomi	**/marconi_scratch/**	marconi_scr	**2hou**	**27G**	**--**	**--%**	360T	2.5P	14.3%

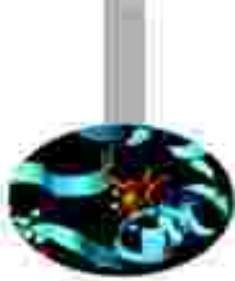
Data Storage @ CINECA



$TAPE: permanent, user specific, shared

USER	AREADESCR	AREAID	FRESH	U
sbuenomi	/marconi_work/cin_tmpaccM1	marconi_work-cin_tmpaccM1	4wee	
sbuenomi	/gss/gss_work/DRES_cin_fs	work_OFFLINE-DRES_cin_fs-FS	28hou	
sbuenomi	/gss/gss_work/cin_staff	work_OFFLINE-cin_staff-__UNCAT__	28hou	
sbuenomi	/gss/gss_work/	work_OFFLINE	28hou	
sbuenomi	/gss/gss_work/cin_staff	work_ONLINE-cin_staff-__UNCAT__	28hou	1
sbuenomi	/marconi_work/	marconi_work	4wee	
sbuenomi	/marconi_work/cin_staff	marconi_work-cin_staff	4wee	
sbuenomi	/gss/gss_work/DRES_cin_fs	work_ONLINE-DRES_cin_fs-FS	28hou	4
sbuenomi	/gss/gss_work/	work_ONLINE	28hou	4
sbuenomi	/marconi/home	marconi_hpc-home	5day	7
sbuenomi	/marconi/	marconi_hpc	5day	7
sbuenomi	/marconi_scratch/	marconi_scr	10hou	

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/
```

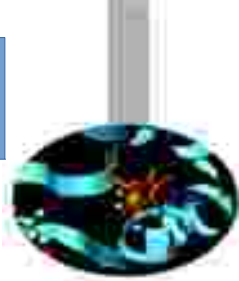
rsync is useful if you need to syncronize the content of a local and a remote dir on a remote host:

```
>rsync --r vzu -e ssh *.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: basic tools



Some Caveat/Hints about data transfer

- On our HPC platforms all interactive programs are killed after 10 minutes of cpu time, so big data movements have to be done using batch script
- To reduce the size of the file to transfer use compressing program (e.g. gzip)
- The latency of transfering a file is quite high, so to exploit bandwith it is better to copy 1 big file (using unix tar command) instead of many little ones
- Moving data outside CINECA NAS (Network Area Storage) heavly depends on the bandwidth of the network you are using

tar -cvf my_data.tar dir1/	creates data.tar archive
tar -rvf my_data.tar dir2/ dir3/ dir4/..	append data to archive
gzip my_data.tar	gzip archive

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 serial

. /cineca/prod/opt/environment/module/3.2.10/none/init/bash
cd $PBS_O_WORKDIR

sourc=/marconi_scratch/........          ## do not put the / here
dest=/marconi_work/MY_PROJECT/MY_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 managment.

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

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

Data Transfer: GridFTP clients

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

A X.509 certificate is issued by a Certificate Authority (CA) which checks the identity of the user and guarantees that the holder of this certificate is existing and his certificate is valid.

The certificate is used for authentication instead of the user's account to avoid the replication of the user's account. When authenticating to a site, the user's certificate is mapped to a local account under which all commands are executed.

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

Data Transfer: GridFTP clients

Basic procedure for globus-url-copy:

1. Get/create personal X.509 certificate
2. Create your proxy credential
   - Certificate from authority
   - Certificate from CA-CINECA

3. add your certificate on the UserDB under “personal data”
4. install the standard client.
5. transfer your data

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

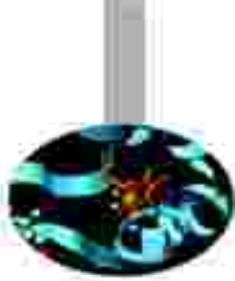
Data Transfer: GridFTP clients

Examples of data transfer using globus-url-copy:

1. for synching recursively a directory and its subdirectories to MARCONI (like with rsync)
globus-url-copy -cd -r -sync /path/to/your/dir/ gsiftp://gftp.marconi.cineca.it:2811/~/remote/dir/

2. for moving a big chunk of data from MARCONI to PICO, a stripped (-stripe option) configuration (using four parallel nodes) is available.
globus-url-copy -p 4 -stripe gsiftp://login.marconi.cineca.it:2811/~/path/to/file gsiftp://gftp.pico.cineca.it:2811/~/path/

3. for listing the file in your directory on iRODS repository.
globus-url-copy -list gsiftp://data.repo.cineca.it:2811/CINECA01/home/your-remote-dir/



Data transfer: use cases

Data to be transferred back at home. I'll need them for further processing on my local cluster	use rsync o gridFTP, or even GridFTP if your local host supports it. Depending on the throughput of your local network, consider several days to move 1TB of data.
Data to be transferred to another PRACE site I'll need them for further processing on another Tier-0 system	gridFTP is strongly recommended. Since a very efficient network is in place among PRACE sites, you can expect to spend few hours to move 1TB of data.

Data Transfer: Globus Online

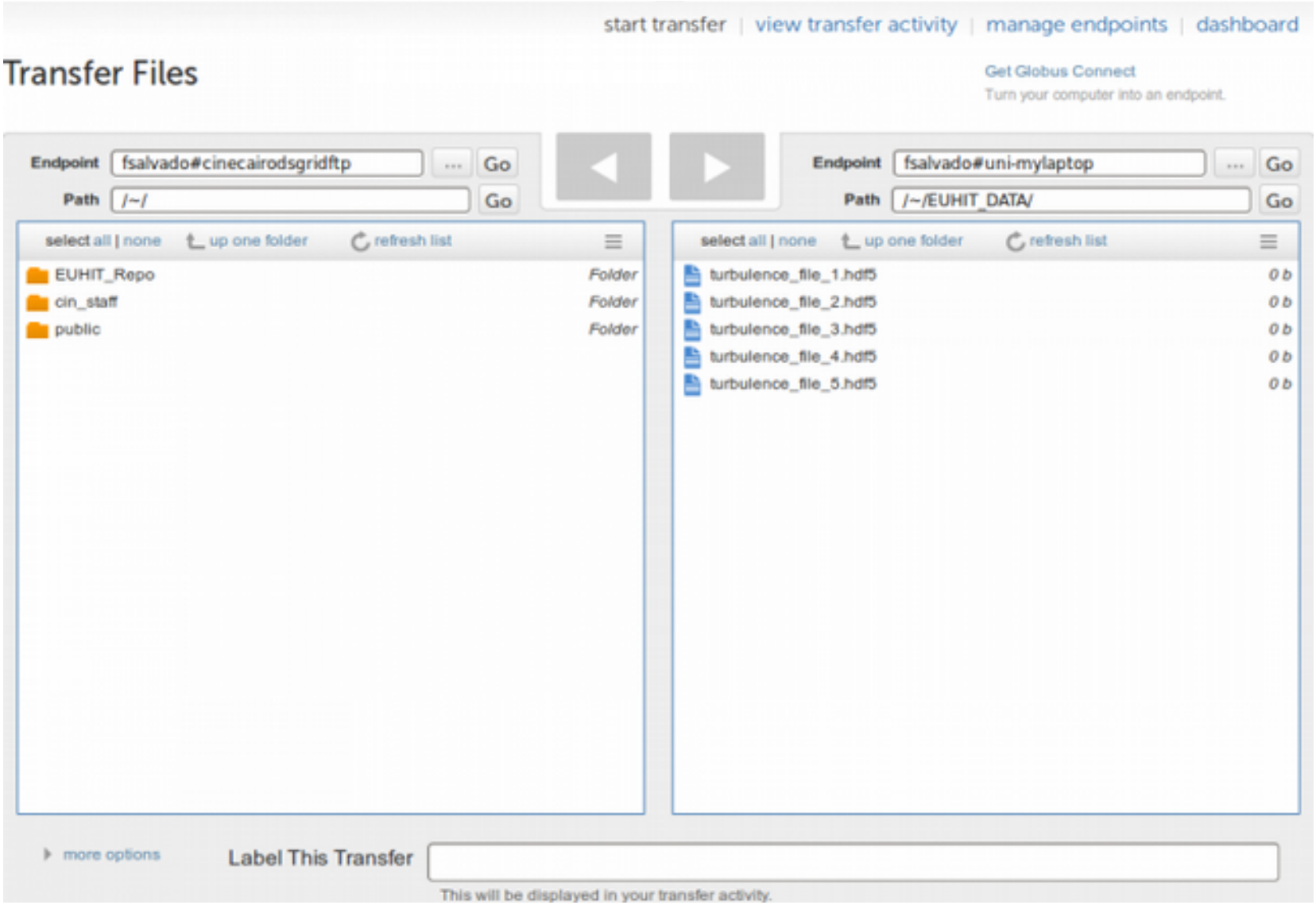
GlobusOnline allows data exchanging between defined and activated Endpoints.
The following steps help you to easily transfer data:

- between two different gridftp servers or
- between a gridftp server (e.g. CINECA gridftp server) and your local machine (e.g. your laptop), using Globus Online web interface.

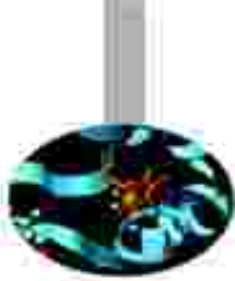
Basic usage to transfer files via Globus Online:

1. register on GlobusOnline on http://www.globusonline.org
2. Create your own proxy credentials
3. Login on GlobusOnline
4. Set your endpoints accordingly
5. Transfer your data

Data Transfer: Globus Online



iRODS-based REPO

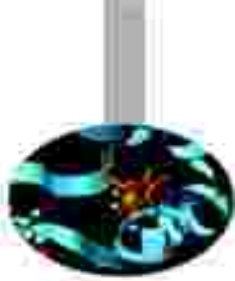


REPO is a CINECA service, implemented through iRODS (Integrated Rule-Oriented Data System), for the management of long lasting data.

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:

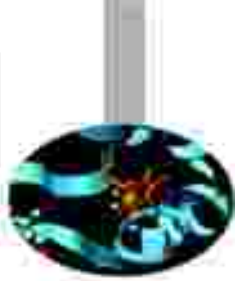
Upload/Download: the system supports high performance transfer protocols like GridFTP, or iRODS multi-threads transfer mechanism.

- The GridFTP protocol is supported as described in this page; 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/, but look down to know how to configure them.

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.

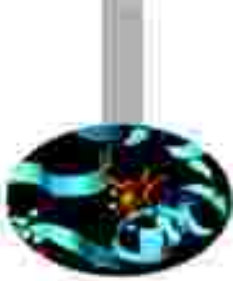
How to access the REPO space



There are two different ways to access data in the REPO:

- iRODS commands
- gridftp clients, such as globus-url-copy or Globus Online

How to access the REPO space



You can use the iCommands from CINECA HPC machines (MARCONI, PICO and GALILEO) or from your local linux machine.

- download iCommands
- download the file chain.pem
- create the .irods/.irods_environment.json config file in the home directory of the system where you use the icommand (MARCONI, PICO or GALILEO, your local linux machine)
- type the command iinit the first time you use irods . On default, the PAM authentication method is enabled ("irods_authentical_scheme parameter" in the json configuration file), so the password of your hpc username will be requested. Note the after some times (days...), you will need retype the iinit command to use the icommands.
- operate in REPO space, using the icommands. The documentation about the IRODS commands is available at this link.

https://wiki.u-gov.it/confluence/display/SCAIUS/iRODS-based+REPO