The HPC Centers vision: CINECA

Marani Alessandro
HPC user support - CINECA
Italy
a.marani@cineca.it

September 19th 2014

- CINECA 2.0

- HPC computing infrastructure: FERMI, PICO, new Tier-1

- How to get acces to our systems

- User environment @cineca: job scheduling and modules

- Disk space, storage and data transfers

- Helpdesk, online documentation and training activities

CINECA
is a non profit Consortium, made up of **69 Italian universities**, **and 3 Institutions** (CNR, OGS and MIUR).
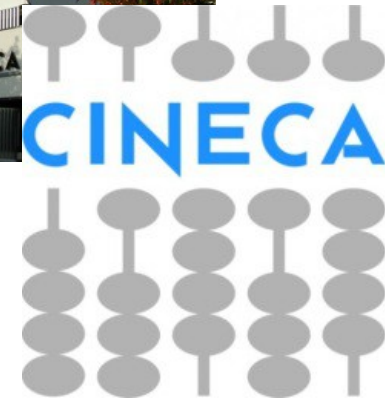
CINECA Today it is the largest Italian computing centre, one of the most important worldwide. . The **High Performance Systems department (SCAI: SuperComputing Applications and Innovation)** offers support to scientific and technological research through supercomputing and its applications.
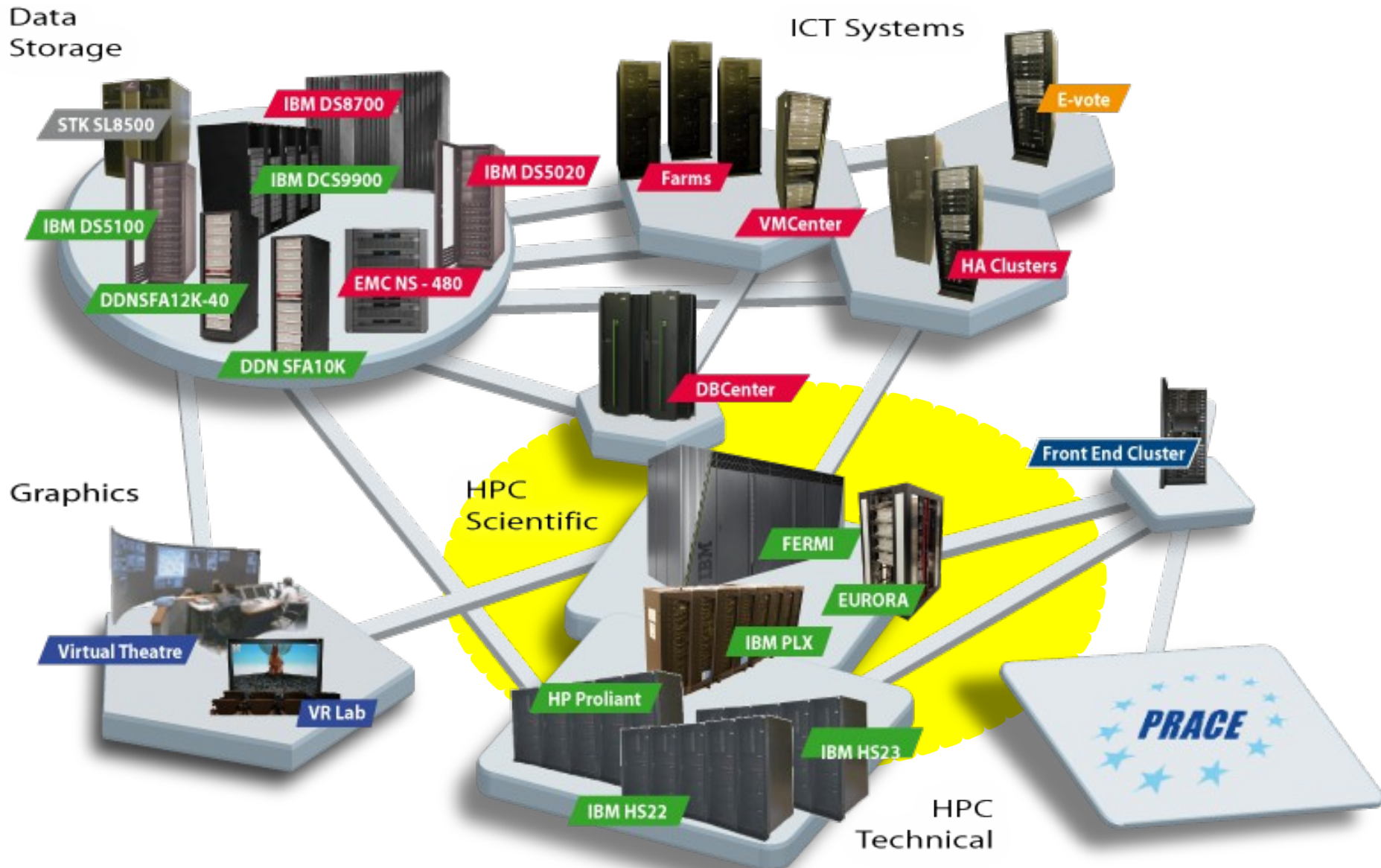
Architecture: BlueGene/Q
Processor: IBM PowerA2, 1.6 GHz
Number of processors (cores): 163840
Number of nodes: 10240 (16 cores per node)
RAM: 16 GB per node
Interconnection network: Internal (5D torus)
Disk space: 2 PB
Power consumption: ~1 MW
Operative system: Linux (on surface)
Peak performance: 2 PFlop/s
Compilers: Fortran, C, C++
Parallel libraries: MPI,OpenMP

FERMI is currently ranked (as of June 2014) 17° in the top 500. It was ranked 7° at its debut in June 2012

It is our current top cluster (Tier-0) and it is mostly used by academic project research, both at a national (ISCRA) and at an european (PRACE) level

It is mainly designed for highly scalable applications with a massive parallel approach. It is not recommended for small or I/O-memory intensive softwares

Processor: Xeon E5-2670, 2.4 GHz

Number of processors (cores): 1360 (computing part)

Total number of nodes: 88

Number of computing nodes: 68 (20 CPUs per node)

RAM: 128 GB per node

Interconnection network: Infiniband

Disk space: 4 PB (+ tape)

Operative system: Linux

Non computing nodes include:
- FAT nodes (512 GB of RAM)
- "Big insight" nodes (3 TB of local disk space)
- I/O nodes
- Visualization nodes (2 NVIDIA K40 GPUs each)
- Nodes reserved for sperimentations (Hadoop, cloud computing, …)

Due to its characteristics, PICO isn't suited for the usual HPC applications, but it can be thought as our "BigData" cluster

The best applications for PICO are the ones with the following peculiarities:

- Low parallelism (less cores with respect to other systems, more cores/node)
- I/O intensive (SSD disk available)
- Memory intensive (more memory/core and /node)
- DB based (a lot of storage space)

Processor: Xeon E5-2630v3, 2.4 GHz

Number of processors (cores): 8256

Total number of nodes: 516 (16 CPUs per node)

RAM: 128 GB per node

Interconnection network: Infiniband, Cluster network

Disk space: yet unknown

Operative system: Linux

Compilers: Fortran, C, C++

Parallel libraries: MPI, OpenMP

The contract for the new Tier-1 system has been signed just yesterday. In a few months, the new cluster will replace the current Tier-1 system (PLX). Our goal is to have it ready for the next Top500 ranking in November 2014

Being a Tier-1 system, it will be mostly suited for small parallel applications not focused of big scalability but rather on high RAM necessity and/or I/O intensive requests

The users of this kind of clusters are mainly industrial and commercial researchers (like meteorologists or CFD experts), or small academic projects which don't suit for FERMI

How to get HPC resources

- Peer reviewed projects:
  you can submit a project that will be reviewed. If you win you will get the needed resources for free

  - Local (Lombardia region) → LISA
  - National → ISCRA
  - Europe → PRACE

- On demand/convention:
  sit is possible to buy a quantity of CPU hours to consume in a specified amount of time, or to sign a convention with CINECA in order to get access to our HPC clusters. Examples of conventioned users are SISSA, INFN, PoliMi and others

- The use of HPC resources is based on **projects** (o Account_no)
- Each project is defined on the basis of:
  - Budget (how many standard hours)
  - Validity (from – to date)
  - PI (Principal Investigator)
  - Collaborators
  - Host(s): one or more systems where the budget can be used
- The project is active as far as the budget is not exhausted and the validity period is OK
- The PI and Collaborators of a project can concurrently use its budget for batch processing.
- Each user gets a permanent **username** that enables him/her to access the HPC system/s
- Access is normally done via **ssh protocol** giving the username/pass information, making it possible for our users to connect remoltely to our clusters anytime, anywhere
- The username is **strictly personal**.

Once you are logged in a CINECA's HPC cluster, you can work in **interactive mode** directly on the front-end node (compiling, debugging, testing,…). In order to do real computation (and consume CPU hours) you have to write a **batch job** and submit it.

In a batch job you specify how many resources do you need (nodes, CPUs, memory,…) and for how many time. You also specify your account so that the system knows who to account for the CPU time consumption

```
#!/bin/bash
#PBS -l walltime=1:00:00
#PBS -l select=1:ncpus=4:mpiprocs=1
#PBS -o job.out
#PBS -e job.err
#PBS -q parallel
#PBS -A <my_account>

cd $PBS_O_WORKDIR
export TMPDIR =
module load stata
stata-mp -b do "input"
```

Once submitted, your job will be processed by a scheduler. It will put your job in queue and execute it when the resources you requested will be available.

Queue waiting time depends on the amount of resources you asked, the number of jobs already scheduled, the internal priority settings and the queue you requested in your batch job

You can always check the status of your account with the "saldo" command

```
$ saldo –b
--------------------------------------------------------------------------------
account        start        end        total        localCluster      totConsumed      totConsumed
                                     (local h)    Consumed(local h)     (local h)           %
--------------------------------------------------------------------------------
try11_test    20110301    20111201    10000            2000              5000            50.0
cin_staff     20110323    20200323    200000000       479621            8933910          4.5

$ saldo -ba try11_test
erossi00 ffalcian mcestari fvitale0 sgiulian
--------------------------------------------------------------------------------
account        start        end        total        localCluster      totConsumed      totConsumed
                                     (local h)    Consumed(local h)     (local h)           %
--------------------------------------------------------------------------------
try11_test  20110301    20111201    10000            2000              5000            50.0
```

Module environment

The "module" environment provides the access to some of the most common HPC applications to our users in each of our clusters

By loading the proper module, a set of environment variables is defined so that the user can execute the application or link the library concerning that specific module. Module instructions can be set on the batch job script so that the compute nodes can take advantage of it as well.

New or updated modules can be installed from the User Support team at the request of the users

Module environment

```
$ module available

----------- /cineca/prod/modulefiles/profiles ---------------------
profile/advanced      profile/base(default)      profile/engineering
……
----------- /cineca/prod/modulefiles/base/libraries --------
PETSc/3.0.0--openmpi--1.3.3--gnu--4.1.2
PETSc/3.0.0--openmpi--1.3.3--intel--11.1--binary(default)

----------- /cineca/prod/modulefiles/base/compilers -----------
IntelMPI/4.0—binary          gnu/4.1.2(default)
gnu/4.5.2                    intel/11.1--binary(default)
openmpi/1.3.3--gnu--4.1.2    openmpi/1.3.3--intel--11.1—binary

----------- /cineca/prod/modulefiles/base/applications -----------
R/2.10.1                     gromacs/4.6.1
abinit/6.12.3                meep/1.1.1(default)
adf/2010.02b(default)        molcas/7.6
amber/11(default)            mopac/2009(default)
amber/12                     namd/2.8(default)
cp2k/2.3(default)            pyfrag/2007.02(default)
desmond/3.0.3                stata/10
….
```

Our cluster's disk space is usually divided into
3 main areas:

HOME: small, permanent area
used for small tasks as compilation or post-processing

SCRATCH: big, temporary area
Equipped with gpfs filesystem, suitable for actual production runs. It's user-defined, and will be replaced by the project-defined WORK

DATA: big, permanent, backupped area
Used for archiviation, not visible from the compute nodes

Data can be stored into **tapes** and archived into our data repository, upon request from our User Support team

Taped data is off-line and can't be seen by our HPC clusters. However, it can be accessed and managed by data management system **iRODS**
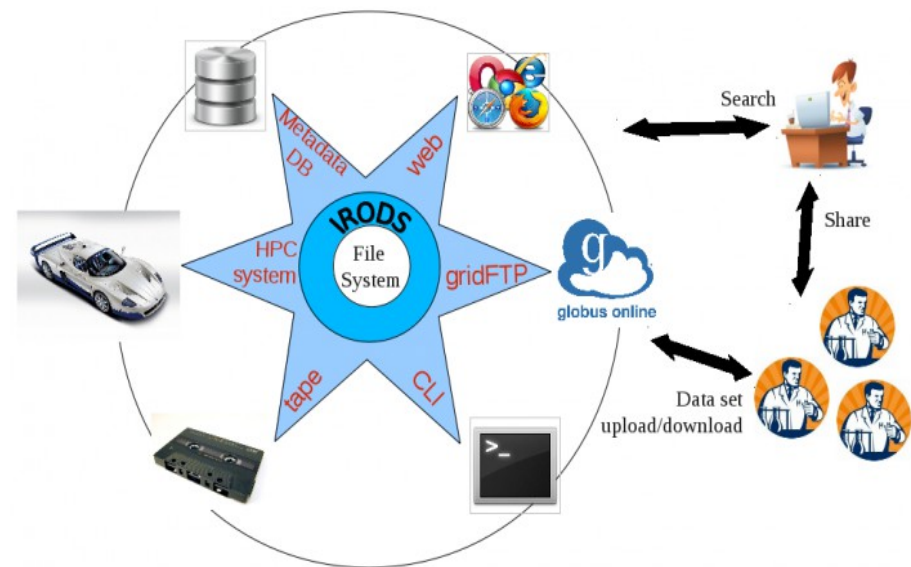
Data transfer

In order to transfer local data to our HPC systems or viceversa, the most common data transfer protocols are available, like scp or rsync. Windows users can take advantage of tools like FileZilla.

Another efficient and very user-friendly system is globus on-line, a gridFTP client. It requires an x509 certificate.
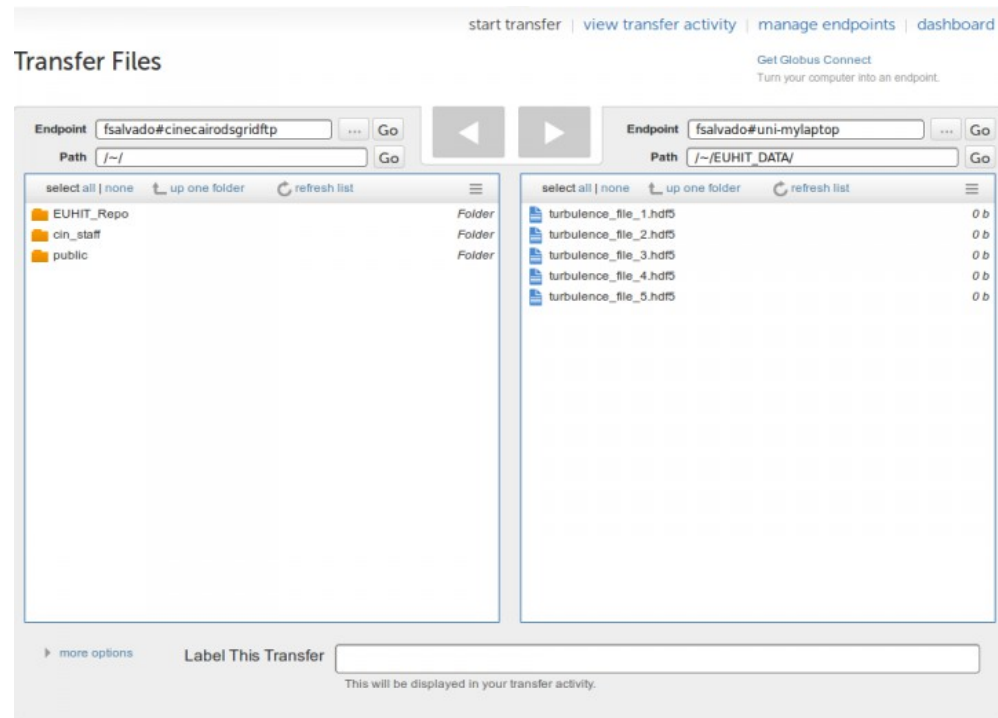
There are two different tools, that make use of accelerators (GPUs) for remote visualization
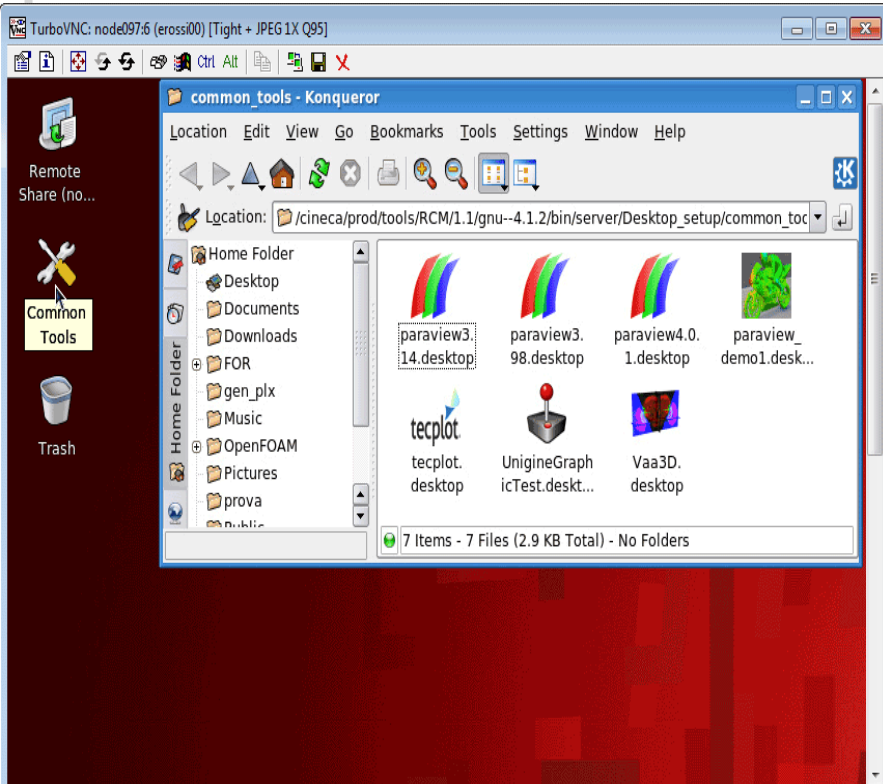
Remote Connection Manager (RCM) provides a graphical connection to the cluster in order to execute some graphical postprocess or visualization tools on the data stored in your disk space

There are two different tools, that make use of accelerators (GPUs) for remote visualization
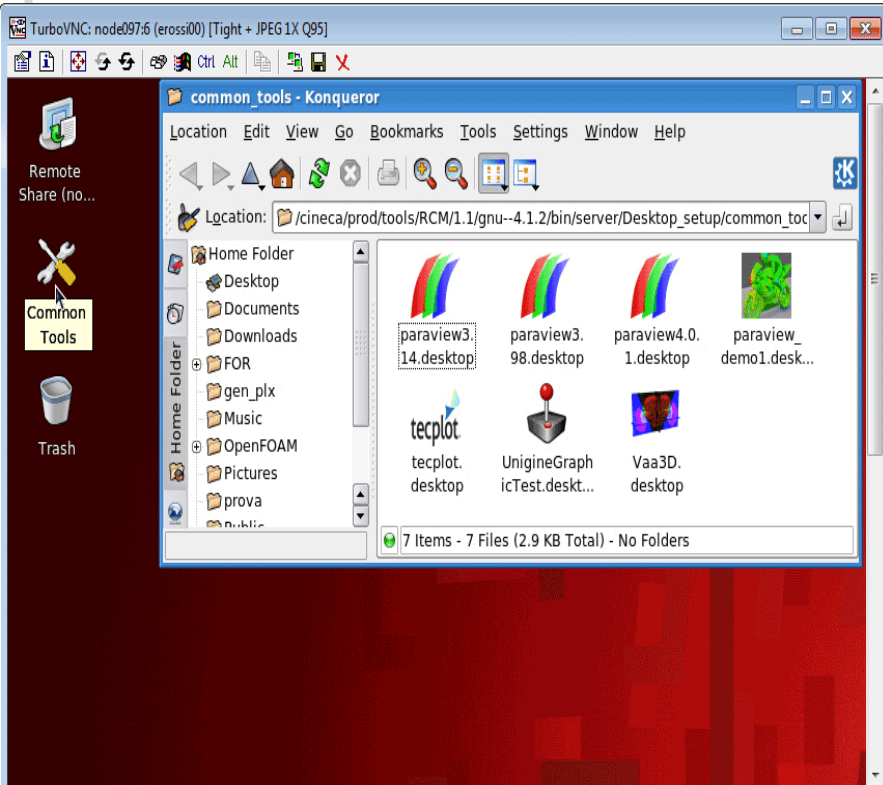
Webcompute provides different visualization services (like Xterm) directly from your browser

CINECA
Consorzio Interuniversitario

23rd Summer School on PARALLEL COMPUTING

19 - 30 May, 2014 - BOLOGNA
14 - 25 July, 2014 - ROME
8 - 19 September, 2014 - MILAN

CINECA offers various training activities spread among the year and the three sites. Topics covered range from basic scientific programming to parallel programming and tips on how to profile and debug in an HPC environment

You can check out our offer at the website:
http://www.hpc.cineca.it/content/training

Documentation and User Guide

http://www.hpc.cineca.it

http://www.hpc.cineca.it/content/hpc-user-guide

HelpDesk

- If you encounter any trouble, you can send a mail to superc@cineca.it
  - Always an on-call operator during working hours
  - You can also ask for domain-driven questions: the operator will find the right consultant for you