UG2.3: Accounting

(updated: Jan 2018)

In this page:

- Usernames and Accounts
- The "saldo" command
- Billing policy
  - "Non-exclusive" mode usage: memory matters!
  - Accounting and accelerators
- Budget linearization

Our HPC resources can be used on a "pay for use" basis.

Currently, the cost is based on elapsed time and the effective number of cores (reserved, not used!) by the batch jobs. In general, most tools and applications from our Software Catalog can be used free of charge even if the program is burdened with a licence. Only in a few cases, you need to register and pay an additional fee in order to access special applications. All the information are reported in the description of the specific application: see application-software-science.

In order to run a batch job, a user must login to an HPC system using his/her username and password. The username must be associated with one or more active projects (Accounts) with available budgets.

Usernames and Accounts

In CINECA, the words "username" and "account" have different meanings.

**Username** identifies the individual connecting to the system. It is the string used (along with the password) for getting access to the system (through ssh, for example). It is 8 characters long and can be obtained, from your interactive unix session, by the command:

```bash
> echo $USER
```

Login credentials are to be considered as strictly personal, meaning that NO SHARING between members of the same working group is expected to happen. Each single user entitled with login credentials is to be considered personally responsible for any misuse that should take place.

**Account** indicates the grant or resource allocation which you can use for your batch jobs. Usually, a "budget" is associated to an Account and measures how many resources (computing hours) can be used within that Account. You can list all the Accounts attached to your username on the current cluster, together with the "budget" and the consumed resources, with the command "saldo" (see below).

One single username can use multiple accounts and one single account can be used by multiple usernames, all competing for the same budget.

The account that will be charged for the resource usage of a batch job is identified by a specific line in the script used for the job submission. Currently, all of our clusters have SLURM as their scheduler. For further documentation about how to write a jobscript for SLURM, please see the UserGuide about "Batch Scheduler SLURM".

Every account has a beginning date and an ending date, within which the budget must be used. On our HPC platforms, special constraints on how to use the budget are defined (see budget linearization below).

The elapsed time spent by the users' batch jobs, multiplied by the number of reserved cores (core-h), is used to decrease the account budget: once a day, at midnight, an automatic procedure will take into account all jobs completed in the past 24 hours and update your budget accordingly.

Whenever an account runs out of budget (in CPU hours), or when its expiring date is met, all the usernames referring to that account won't be able to submit batch jobs anymore using the corresponding project budget (but they will still be able to do so if they are authorized on other projects account).

Nevertheless, users with no active Account will still be able to access the HPC platforms in order to perform some lightweight post-processing (interactive runs) and/or to retrieve their data. Usernames will be kept alive for a whole year after their last (most recent) account has been shut down.

The mapping between users and Accounts is done by the CINECA staff, who is in charge of creating new projects and associating a PI to each of them. A PI, in turn, has the possibility of associating other users to a project as collaborators, via the UserDb page related to the project.

The "saldo" command

You can list all the Accounts attached to your username on the current cluster, together with the "budget" and the consumed resources, with the command:
One single username can use multiple Accounts and one single Account can be used by multiple usernames (possibly on multiple platforms), all competing for the same budget.

On systems like MARCONI, where different independent partitions are available (KNL and SKL) you should specify the host you are interested in:

```
> saldo -b (by default returns account_no active on SKL)
> saldo -b --knl (returns account_no active on KNL)
> saldo -b --skl (returns account_no active on SKL)
```

Another useful command is

```
> saldo -r
```

that prints daily resource usage report for selected usernames and/or Accounts on the local cluster.

For more information run the "saldo" command without any option.

**Billing policy**

The time spent in interactive work is not considered by the billing procedures, meaning it is free of charge.

For batch jobs, the billing is based on "elapsed time" (execution time, not necessarily wall clock time) and "effective number of cores" (reserved, not used!) by the batch job. When the node is used in a "non-exclusive" mode, the memory request is also taken into account (see below).

The basic idea is:

\[
\text{accounted hours} = \text{WallClockTime} \times \text{ReservedCores}
\]

Please note that every cluster has usually a "serial" queue, defined on front-end nodes, that allow for serial jobs for a short time limit (usually 4 hours). On these queues, accounting is not enabled, meaning that you can use them without being charged. As a consequence, serial queues are allowed to be used also when an account is expired or has exhausted all of its budget: it is useful for example for operations of post-processing or data transfer.

"Non-exclusive" mode usage: memory matters!

On some clusters (for example on GALILEO) you can also choose to allocate for your job part of the node, and are not forced to allocate all of it as it happens in clusters (like MARCONI) running in exclusive mode. In this case, the accounting procedure takes also into account the amount of memory you request for your job. If you ask for an amount of memory that is larger than the expected quantity in relation to the number of cores requested, the jobs will be billed for an "equivalent" (larger) number of cores.

The billing always follows the basic idea illustrated above, but a generalized parameter for the number of reserved cores, accounting for the memory request, is now used:

\[
\text{accounted hours} = \text{WallClockTime} \times \text{ReservedCoresEquiv}
\]

where

\[
\text{ReservedCoresEquiv} = \text{ReservedCores} \times \text{MemFactor}
\]

- **MemFactor = 1.**
  
  If the memory you ask for is smaller than or equal to the expected amount in relation to the number of reserved cores, in this case, you are not allocating more of the node capability than expected, and the amount of cpu-hours billed depends only on the number of requested cores.

- **MemFactor = \((\text{ReservedMemory} / \text{TotalMemory}) / (\text{ReservedCores} / \text{TotalCores})\)**

  If the memory you ask for is larger than the expected amount in relation to the number of reserved cores, in this case, you are allocating for more node capability than expected, and the amount of cpu-hours billed depends also on the amount of memory requested (i.e. the actual percentage of node allocated).

For example, on GALILEO the TotalMemory considered to calculate the MemFactor is 118000 MB of memory (around 115GB), and each compute node has 36 cores:

- TotalMemory = 115 GB
- TotalCore = 36

If you ask for only one core and 63 GB of memory (thus allocating for yourself half of the node even if you are using one core), the MemFactor is:

- ReservedMemory = 63 GB
ReservedCores = 1
MemFactor = (63 GB / 115 GB) / (1 / 36) = 20

Hence, with such a request for each hour of computation, your budget will be billed for 20 equivalent CPUs, i.e., for 20 hours.

This rule applies for each cluster based on its amount of total memory and cores.

Accounting and accelerators

Recently, accounting system has been extended to nodes equipped with accelerators. The principle is the same as the memory accounting: asking for a number of accelerators that will make you allocate a bigger portion of the node than what is suggested by the simple number of cpus requested will increase the consumption accordingly.

For GALILEO, every GPU will be treated as 18 cores in terms of accounting. That is because GPU nodes are Broadwell and have 36 CPUs each, so allocating 1 GPU out of the available 2 is equivalent to allocate half of the node (i.e. 18 CPUs).

Some examples based on GALILEO (1 node):

- cpus=24, gpus=1 ==> the number of GPUs requested is equal as having requested 18 CPUs, but since 24 of them have been requested in the standard way, they are not taken into account. Thus 24 CPUs will be billed;
- cpus=6, gpus=1 ==> the number of GPUs requested is equal as having requested 18 CPUs, which is higher than the number of CPUs requested. Thus 18 CPUs will be billed;
- cpus=24, gpus=2 ==> the number of GPUs requested is equal as having requested 36 CPUs, while 24 of them have been requested in the standard way, and they are not enough to cover for the GPU request. Therefore 36 CPUs will be billed;
- cpus=24, gpus=1, mem=115GB ==> the situation is similar to the first example (so 24 CPUs billed), but the memory request is higher than what is guaranteed by the simple allocation of the CPUs or GPUs, since it is equivalent of allocating the entire node. So, 36 CPUs will be billed.

Budget linearization

A linearization policy for the usage of project budgets is active on all clusters at Cineca. For each account, a monthly quota is defined as (total_budget / total_no_of_months). Starting from the first day of each month, the collaborators of any account are allowed to use the quota at full priority. As long as the budget is consumed, the jobs submitted from the account will gradually lose priority, until the monthly budget is fully consumed. At that moment, their jobs will still be considered for execution (so it is possible to consume more than the monthly quota), but with a lower priority than the jobs from accounts that still have some quota left.

The linearization effect on the priority is fine graduated, as the linearization parameter depends on the percentage of the monthly quota consumed. The job sorting formula depends also on other aspects, like walltime, resources requested or time spent waiting in queue, so a low priority job can still have some chance of being executed in a quick amount of time if well-tuned (but not as quick as jobs with the same tuning but advantaged in terms of linearization priority). You can check the usage of your monthly quota with the "saldo -b" command: the last two columns are about the quota defined for your account and the monthly consumption.

This policy is similar to those already applied by other important HPC centers in Europe and worldwide. The goal is to improve the response time, giving users the opportunity of using the cpu hours assigned to their project in relation to their actual size (total amount of core-hours). Please note that it is recommended to apply a sort of "linearization" of your project budget. Each month a given percentage of your budget is guaranteed, but non-linear usage is discouraged for the welfare of all the users that are simultaneously hosted by our HPC systems.