TECHNICAL GUIDELINES FOR APPLICANTS TO PRACE 12th CALL (Tier-0)

Contributing sites and the corresponding computer systems for this call are:

BSC, Spain	**IBM System x iDataPlex “MareNostrum”**
CINECA, Italy	**IBM Blue Gene/Q “Fermi”**

The site selection is done together with the specification of the requested computing time by the two sections at the beginning of the online form. The applicant can choose one or several machines as execution system.

The parameters are listed in tables. The first column describes the field in the web online form to be filled in by the applicant. The remaining columns specify the range limits for each system.

The applicant should indicate the unit.

A - General Information on the Tier-0 systems available for PRACE 12th Call

		Fermi	***MareNostrum***	***MareNostrum Hybrid***
	System Type	Blue Gene/Q	IBM System x iDataPlex	IBM System x iDataPlex
Compute	Processor type	IBM PowerPC A2 (1.6 GHz) 16 cores/node	Intel Sandy Bridge EP	Intel Sandy Bridge EP
	Total nb of nodes	10 240	3 056	42
	Total nb of cores	163 840	48 896	672
	Nb of accelerators / node	n.a.	n.a.	2
	Type of accelerator	n.a.	n.a.	Intel Xeon Phi 5110P
Memory	Memory /Node	16 GB	32 GB	64 GB
Network	Network Type	IBM Custom	Infiniband FDR10	Infiniband FDR10
	Connectivity	5D Torus	Fat Tree	Fat Tree

		Fermi	***MareNostrum***	***MareNostrum Hybrid***
Home file system	Type	GPFS	GPFS	GPFS
	Capacity	114 TB	59 TB	59 TB
Work file system	Type	GPFS	GPFS	GPFS
	Capacity	1.7 PB	612 TB	612 TB
Scratch file system	Type	GPFS	GPFS	GPFS
	Capacity	854 TB	1.1 PB	1.1 PB
Archive	Capacity	On demand	3.7 PB	3.7 PB
Minimum required job size	Nb of cores	2 048	1 024	16

IMPORTANT REMARK:

Fermi:

Fermi will be replaced by a Marconi system mid-2016. General information is not yet available. It is recommended to applicants to prove that their application codes are able to scale on other architectures. This will be favourably considered for the technical evaluation of the project.

MareNostrum:

MareNostrum III will be replaced by a new system named MareNostrum IV. All applications running in MareNostrum III will also be supported on the newer system and for now the terms of access will be the same as for the existing system.

More details on the website of the centers:

Fermi:
http://www.hpc.cineca.it/content/fermi-reference-guide

MareNostrum:
http://www.bsc.es/marenostrum-support-services/mn3

Subsection for each system

Fermi, CINECA

Fermi system is 10 racks Blue Gene/Q with 1024 compute nodes per rack. Please be aware that 1 node consists of 16 cores with four-fold SMT and is equipped with 16 GB, i.e. each physical core has at most 1024 MB of main memory available. Pure MPI codes should use 16 tasks per node. In this case the amount of memory/task must be lower than 1 GB. Hybrid (multithread) applications too must cope with a maximum of 16 GB per node. In order to use the architecture efficiently, pure MPI and hybrid codes are highly recommended to use 32 (or even 64) tasks/threads per node.

MareNostrum, BSC

MareNostrum system consists of 36 IBM iDataPlex Compute Racks, and 84 IBM dx360 M4 compute nodes per rack. Each compute node has two 8-core SandyBridge-EP processors at 2.6 GHz, and 32 GB of main memory (2 GB/core), connected via Infiniband.

MareNostrum Hybrid is composed of 42 nodes with 16 cores, 64 GB of main memory, 2 Xeon Phi processors and one IB FDR10 link per node, for a peak performance of 100 teraFLOPS.

B – Guidelines for filling-in the on-line form

Resource Usage

Computing time

The amount of computing time has to be specified in core-hours (wall clock time [hours]*physical cores of the machine applied for). It is the total number of core-hours to be consumed within the twelve months period of the project.

Please justify the number of core-hours you request with **a detailed work plan**. Not doing so might result in decreasing the amount of core-hours or even in rejection of the proposal.

The project should be able to start immediately and is expected to use the resources continuously.

When planning for access, please take into consideration that the effective availability of the system is about 80 % of the total availability, due to queue times, possible system maintenance, upgrade, and data transfer time.

If less than 5 million core-hours in one of the Tier-0 system is required, the choice to use Tier-0 systems has to be justified as compared to the use of Tier-1 systems.

The maximum value of computing time is limited by the total number of core-hours per system given in the terms of reference document for the 12th Call (see the "Call announcement" page at *www.prace-ri.eu/Call-Announcements*). **Any further limitation is specified in the terms of reference document of the corresponding Call for Proposals**.

Job Characteristics

This section describes technical specifications of simulation runs performed within the project.

Wall Clock Time

A simulation consists in general of several jobs. The wall clock time for a simulation is the total time needed to perform such a sequence of jobs. This time could be very large and could exceed the job wall clock time limits on the machine. **In that case the application has to be able to write checkpoints and the maximum time between two checkpoints has to be less than the wall clock time limit on the specified machine.**

Field in online form	*Machine*	*Max*
Wall clock time of one typical simulation (hours) **<number>**	All	< 10 months
Able to write checkpoints **<check button>**	All	
Maximum time between two checkpoints **(= maximum wall clock time for a job) (hours)** **<number>**	Fermi MareNostrum	24 hours 24 hours

Number of simultaneously running jobs

The next field specifies the number of independent runs which could run simultaneously on the system during normal production conditions. This information is needed for batch system usage planning and to verify if the proposed work plan is feasible during project run time.

Field in online form	*Machine*	*Max*
Number of jobs that can run simultaneously <number>	Fermi MareNostrum	2-10 (depending on the job size) dynamic*

* Depending on the amount of PRACE projects assigned to the machine, this value could be changed.

Job Size

The next fields describe the job resource requirements which are the number of cores and the amount of main memory. These numbers have to be defined for three different job classes (with minimum, average, or maximum number of cores).

Please note that the values stated in the table below are absolute minimum requirements, allowed for small jobs, which should only be requested for a small share of the requested computing time. Typical production jobs should run at larger scale.

Job sizes must be a multiple of the minimum number of cores in order to make efficient use of the architecture.

IMPORTANT REMARK

Please provide explicit scaling data of the codes you plan to work with in your project at least up to the minimum number of physical cores required by the specified site (see table below) using input parameters comparable to the ones you will use in your project (a link to external websites, just referencing other sources or "general knowledge" is not sufficient). ***Generic scaling plots provided by vendors or developers do not necessarily reflect the actual code behavior for the simulations planned. Missing scaling data may result in rejection of the proposal.***

Field in online form	*Machine*	*Min (cores)*
Expected job configuration (Minimum) <number>	Fermi MareNostrum	2 048 1 024
Expected number of cores (Average) <number>	Fermi MareNostrum	see above see above
Expected number of cores (Maximum) <number>	Fermi MareNostrum	32 768 -

Virtual cores (SMT is enabled) are not counted. Accelerator based systems (GPU, Xeo,, Phi, etc) need *special rules*.

Additional information:

FERMI

The minimum number of (physical) cores per job is 2 048.

However, this minimum requirement should only be requested for a small share of the requested computing time and it is expected that PRACE projects applying for FERMI can use at least 4096 physical cores per job on average.

Job sizes must use a multiple of 2 048 physical cores in order to fit into the architecture.

The maximum number of (physical) cores per job is 32 768. Larger jobs are possible in theory but the turnaround time is not guaranteed.

Please provide explicit scaling data of the codes you plan to work with in your project. A good scalability up to 4 096 physical cores must be demonstrated and the scaling behavior up to 8 192 physical cores must be shown using input parameters comparable to the ones you will use in your project.

For hybrid (multi-threaded) codes it is strongly recommended, that applicants show scaling data for different numbers of threads per task in order to exploit the machine most efficiently. Providing such kind of data will be favorably considered for the technical evaluation of the project.

Since FERMI will be replaced by MARCONI in mid 2016 (see also the terms of reference for this call), proving to be able to scale on other architectures will be favorably considered for the technical evaluation of the project.

Job Memory

The next fields are the total memory usage over all cores of jobs.

Field in online form	***Machine***	***Max***
Memory (Minimum job) **<number>**	Fermi MareNostrum	1 GB * #cores 2 GB * #cores
Memory (Average job) **<number>**	Fermi MareNostrum	see above see above
Memory (Maximum job) **<number>**	Fermi MareNostrum	see above -

The memory values include the resources needed for the operating system, i.e. the application has less memory available than specified in the table.

Storage

General remarks

The storage requirements have to be defined for four different storage classes (Scratch, Work, Home and Archive).

- **Scratch** acts as a temporary storage location (job input/output, scratch files during computation, checkpoint/restart files; no backup; automatic remove of old files).
- **Work** acts as project storage (large results files, no backup).
- **Home** acts as repository for source code, binaries, libraries and applications with small size and I/O demands (source code, scientific results, important restart files; has a backup).
- **Archive** acts as a long-term storage location, typically data reside on tapes. For PRACE projects also archive data have to be removed after project end. The storage can only be used to backup data (simulation results) during project's lifetime.

Data in the archive is stored on tapes. **Do not store thousands of small files in the archive, use container formats** (e.g. tar) to merge files (**ideal size of files: 500 – 1 000 GB**). Otherwise, **you will not be able to retrieve back the files from the archive within an acceptable period of time** (for retrieving one file about 2 minutes time (independent of the file size!) + transfer time (dependent of file size) are needed)!

IMPORTANT REMARK

*All data must be removed from the execution system within **2 months** after the end of the project.*

Total Storage

The value asked for is the maximum amount of data needed at a time. Typically this value varies over the project duration of 12 month (or yearly basis for multi-year projects). **The number in brackets in the ”Max per project” column is an extended limit, which is only valid if the project applicant contacted the center beforehand for approval**.

Field in online form	*Machine*	*Max per project*	*Remarks*
Total storage (Scratch) <number> **Typical use: Scratch files during simulation, log files, checkpoints** **Lifetime: Duration of jobs and between jobs**	Fermi MareNostrum	20 TB (100 TB) 100TB (200 TB)	without backup, cleanup procedure for files older than 30 days without backup, cleanup procedure
Total storage (Work) <number> **Typical use: Result and large input files** **Lifetime: Duration of project**	Fermi MareNostrum	20 TB (100TB)* 10 TB	without backup with backup
Total storage (Home) <number> **Typical use: Source code and scripts** **Lifetime: Duration of project**	Fermi MareNostrum	50 GB 40 GB	
Total storage (Archive) <number>	Fermi MareNostrum	20 TB (100 TB)** 100 TB	 typical file size should be > 5 GB

* The default value is 1 Tb. Please ask to CINECA User Support to increase your quota after the project will start
** Not active by default. Please ask to CINECA User Supportafter the project will start

When requesting more than the specified scratch disk space and/or larger than 1 TB a day and/or storage of more than 4 million files, please justify this amount and describe your strategy concerning the handling of data (pre/post processing, transfer of data to/from the production system, retrieving relevant data for long-term). If no justification is given the project will be proposed for rejection.

If you request more than 100 TB of disk space, please contact peer-review@prace-ri.eu before submitting your proposal in order to check whether this can be realized.

Number of Files

In addition to the specification of the amount of data, the number of files also has to be specified. If you need to store more files, **the project applicant must contact the center beforehand for approval**.

Field in online form	*Machine*	*Max*	*Remarks*
Number of files (Scratch) **<number>**	Fermi MareNostrum	2 Million 4 Million	without backup files older than 90 days will be removed automatically
Number of files (Work) **<number>**	Fermi MareNostrum	2 Million 2 Million	
Number of files (Home) **<number>**	Fermi MareNostrum	100.000 100.000	
Number of files (Archive) **<number>**	Fermi MareNostrum	10.000 1 Million	* *

* HSM has a better performance with a small amount of very big files

Data Transfer

For planning network capacities, applicants have to specify the amount of data which will be transferred from the machine to another location. Field values can be given in Tbyte or Gbyte.

Reference values are given in the following table. *A detailed specification would be desirable: e.g. distinguish between home location and other Prace Tier-0 sites.*

Please state clearly in your proposal the amount of data which needs to be transferred after the end of your project to your local system. Missing information may lead to rejection of the proposal.

Be aware that transfer of large amounts of data (e.g. tens of TB or more) may be challenging or even unfeasible due to limitations in bandwidth and time. Larger amounts of data have to be transferred continuously during project's lifetime.

Alternative strategies for transferring larger amounts of data at the end of projects have to be proposed by users (e.g. providing tapes or other solutions) and arranged with the technical staff.

Field in online form	*Machine*	*Max*
Amount of data transferred to/from production system **<number>**	Fermi MareNostrum	20 TB* 50 TB

* More is possible, but needs to be discussed with the site prior to proposal submission.

If one or more specifications above is larger than a reasonable size (e.g. more than tens of TB data or more than 1TB a day) the applicants must describe their strategy concerning the handling of data in a separate field (pre/post-processing, transfer of data to/from the production system, retrieving relevant data for long-term). In such a case, the application is *de facto* considered as I/O intensive.

I/O

Parallel I/O is mandatory for applications running on Tier-0 systems. Therefore the applicant must describe how parallel I/O is implemented (checkpoint handling, usage of I/O libraries, MPI I/O, netcdf, HDF5 or other approaches). Also the typical I/O load of a production job should be quantified (I/O data traffic/hour, number of files generated per hour).