TECHNICAL GUIDELINES FOR APPLICANTS TO PRACE 11th CALL (Tier-0)

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

BSC, Spain	**IBM System X iDataplex "MareNostrum"**
CINECA, Italy	**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 systems (please check above which of these systems are available for this call)

		Curie FN	*Curie TN*	*Curie HN*	*Fermi*	*Hornet*	*MareNostrum*	*MareNostrum Hybrid*	*SuperMUC TN*	*SuperMUC FN**
	System Type	Bullx	Bullx	Bullx	Blue Gene/Q	Cray XC30	IBM System x iDataPlex	IBM System x iDataPlex	IBM System x iDataPlex	IBM BladeCenter HX5
Compute	Processor type	Intel Nehalem EX 2,27 Ghz	Intel SandyBridge EP 2,7 Ghz	Intel Westmere EP 2,67 Ghz	IBM PowerPC A2 (1,6 GHz) 16 cores/node	Intel Xeon E5-2680v3 (Haswell)	Intel Sandy Bridge EP	Intel Sandy Bridge EP	Intel Sandy Bridge EP	Intel Westmere EX
	Total nb of nodes	90	5040	144	10.240	3944	3028	42	9216	205
	Total nb of cores	11.520	80.640	1152	163.840	94656	48.448	672	147.456	8200
	Nb of accelerators / node	n.a.	n.a.	2	n.a.	n.a.	n.a.	2	n.a.	n.a.
	Type of accelerator	n.a.	n.a.	Nvidia M2090	n.a.	n.a.	n.a.	Intel Xeon Phi 5110P	n.a.	n.a.
Memory	Memory / Node	512 GB	64 GB	24 GB	16 GB	128GB	32	64 GB	32 GB	256 GB
Network	Network Type	Infiniband QDR 10	Infiniband QDR 10	Infiniband QDR 10	IBM Custom	Cray Aries	Infiniband FDR10	Infiniband FDR10	Infiniband FDR10	Infiniband QDR
	Connectivity	Fat tree	Fat tree	Fat tree	5D Torus	Dragonfly	Fat Tree	Fat Tree	Fat tree within island (8192 cores), pruned tree between islands	Fat Tree

* Only for pre-/post-processing purposes. Please verify with Gauss@LRZ for large computing resource requests on FN.

		Curie FN	***Curie TN***	***Curie HN***	***Fermi***	***Hornet***	***MareNostrum***	***MareNostrum Hybrid***	***MareNostrum TN***	***MareNostrum FN***
Home file system	type	NFS	NFS	NFS	GPFS	NFS	GPFS	GPFS	NAS	NAS
	capacity	8 TB	8 TB	8 TB	100 TB	60 TB	59 TB	59 TB	1,5 PB	1,5 PB
Work file system	type	Lustre	Lustre	Lustre	GPFS	Lustre	GPFS	GPFS	GPFS	GPFS
	capacity	600 TB	600 TB	600 TB	2PB	7 PB	612 TB	612 TB	7 PB	7 PB
Scratch file system	type	Lustre	Lustre	Lustre	GPFS	n.a.	GPFS	GPFS	GPFS	GPFS
	capacity	3,4 PB	3,4 PB	3,4 PB	1 PB	n.a.	1,1 PB	1,1 PB	3 PB	3 PB
Archive	capacity	Unlimited	Unlimited	Unlimited	On demand	On demand	2,4 PB	2,4 PB	30 PB On demand	30 PB On demand
Minimum required job size	Nb of cores	128	512	32	2048	2048	1024	16	512	512

More details on the website of the centers:

Curie:
http://www-hpc.cea.fr/en/complexe/tgcc-curie.htm

Fermi:
http://www.hpc.cineca.it/hardware/ibm-bgq-fermi

Hornet:
http://www.hlrs.de/systems/platforms/cray-xc30-hornet/

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

SuperMUC:
http://www.lrz.de/services/compute/supermuc/

Subsection for each system

Curie, GENCI@CEA

The Curie BULLx system encompasses three different partitions:

1. Curie Fat Nodes (FN): composed by 90 nodes, each node having 16 octo core Intel Nehalem EX processors 2,26 GHz, 4 GB/core (512 GB/node). These nodes are interconnected through an Infiniband QDR network.

2. Curie Thin Nodes (TN): composed by 5040 blades, each node having 2 octo core Intel SandyBridge EP processors 2,7 GHz, 4 GB/core (64 GB/node) and around 64 GB of local SSD acting as local /tmp. These nodes are interconnected through an Infiniband QDR network.

3. Curie Hybrid Nodes (HN): composed by 144 nodes, each node having 2 GPU Nvidia M2090 coupled to 2 four cores CPU Westmere EP clocked at 2,67 GHz (8 cores and 2 GPU / node and1152 cores and 288 GPU for the full hybrid configuration). Each node has 24 Go of memory, let 3 Go / core by default, and each GPU has 6 Go. These nodes are interconnected through an Infiniband QDR network.

Purpose of Thin and Fat nodes on Curie: The Thin nodes and Fat nodes have the same ratio memory/core, therefore 4GB/core. The processor’s speed of Fat nodes is lower than the Thin nodes. The fat nodes are an assembly of 4 nodes with 32 cores each and the NUMA effect can be amplified. The main purpose of Fat nodes is an application who need more than 64GB per task, typically pre or post processing but it can be used also for SMP applications, with few memory requirement.

Fermi, CINECA

The 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.

Hornet, GCS@HLRS

The Cray XC30 “Hornet” consists of 21 racks with a total of 3944 compute nodes. One node consists of two sockets equipped with the new Intel Xeon 2680v3 processor providing 24 cores and a total of

128 GB main memory per node. To use the architecture efficiently, pure MPI code must use 24 processes per node. If the code can benefit from Hyperthreading, even 48 processes per node might be beneficial. Hybrid codes might use up to 24 threads (48 with Hyperthreading) per node.

If a project wants to have access to the archive, this has to be mentioned and justified in the project proposal.

MareNostrum, BSC

The 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.

SuperMUC, GCS@LRZ

The system consists of 18 thin node islands which are connected with Infiniband technology. Each node within the thin islands consists of 16 physical cores and typically 1.5 GB/core of main memory is available for use while the application is running. In some exceptional cases, when higher memory per task is required some cores of a node may be left idle to dedicate their memory to other cores but this scenario is not encouraged. The Fat node island has 205 nodes, each having 40 physical cores and a total of 240 GB main memory per node (6.0 GB/core) available for the application.

Most of the computing resources are available on the Thin Nodes. The fat nodes are mainly intended for pre-/post-processing jobs or for such parts of a project which really requires large memory. Fat nodes should not be considered as the only major integral part of the computing time request.

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 11th 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>	Curie Fat Nodes Thin Nodes Hybrid Nodes Fermi Hornet MareNostrum SuperMUC Fat nodes Thin nodes	24 hours 24 hours 24 hours 24 hours 24 hours (12 hours)* 24 hours 48 hours 48 hours

* This might be changed during project runtime, guaranteed minimum is the value in brackets.

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>	Curie Fat Nodes	10 (128 cores), 1 (4096 cores)
	Thin Nodes	50 (512 cores), 4 (8192 cores)
	Hybrid Nodes	10
	Fermi	2-10 (depending on the job size)
	Hornet	tbc
	MareNostrum	dynamic*
	SuperMUC Fat Nodes	4 (520 cores) or 1 (2080 cores)
	Thin Nodes	Up to max. 8 (512 cores) or 2 (8192 cores) or 1 (32.768 cores))

* 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>	Curie Fat Nodes	128
	Thin Nodes	512
	Hybrid Nodes	32
	Fermi	2048
	Hornet	2048
	MareNostrum	1024
	SuperMUC Fat Nodes	512
	Thin nodes	
Expected number of cores (Average) <number>	Other systems	see above
	Hornet	4096
		8192
	SuperMUC	
	Thin Nodes	

Expected number of cores (Maximum) <number>	Curie Fat Nodes Thin Nodes Hybrid Nodes Fermi Hornet SuperMUC Thin Nodes	4096 40.000 (80.000 on demand) 1152 32.768 tba 32.768

Virtual cores (SMT is enabled) are not counted. *GPU based systems need special rules*.

Additional information:

FERMI

The minimum number of (physical) cores per job is 2048.

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 2048 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 4096 physical cores must be demonstrated and the scaling behavior up to 8192 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.

SuperMUC

The minimum number of (physical) cores per job is 512. However, it is expected that PRACE projects applying for this system can use more than 2048 physical cores per job. When running several jobs simultaneously, filling complete islands (approx. 8192 cores) is recommended as a best practice.

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>	Curie Fat Nodes Thin Nodes Hybrid Nodes Fermi Hornet MareNostrum SuperMUC Fat Nodes Thin Nodes	4 GB * #cores or up to 512 GB * #nodes 4 GB * #cores or 64 GB * #nodes 3 GB * #cores or 24 Gb * #nodes 1 GB * #cores Jobs should use a substantial fraction of the available memory 2 GB * #cores Jobs should use a substantial fraction of the available memory
Memory (Average job) <number>	Other systems SuperMUC Fat Nodes Thin Nodes	see above Jobs should use a substantial fraction of the available memory

Memory (Maximum job) <number>	Other systems	see above
	Curie Fat Nodes	4 GB * #cores or up to 512 GB * #nodes
	Thin Nodes	4 GB * #cores or 64 GB * #nodes
	Hybrid Nodes	3 GB * #cores or 24 Gb * #nodes
	Hornet	128GB*#nodes 1,5 GB* #cores or 24 GB* #nodes
		6,0 GB * #core or 240 GB* #nodes
	SuperMUC Thin Nodes	
	Fat Nodes	

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 – 1000 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 overthe project duration of 12 month. **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**	Curie Fat Nodes or Thin Nodes or Hybrid Nodes Fermi Hornet MareNostrum SuperMUC	20 TB (100 TB) 20 TB (100 TB) - 40 TB*[1] 100 TB (200 TB)	without backup, automatic cleanup procedure without backup, cleanup procedure for files older than 30 days HLRS provides a special mechanism for Work spaces, see next row without backup without backup, automatic cleanup procedure
Total storage (Work) <number> **Typical use: Result and large input files** **Lifetime: Duration of project**	Curie Fat Nodes or Thin Nodes or Hybrid Nodes Fermi Hornet MareNostrum SuperMUC	1 TB 20 TB (100TB) 250 TB 10 TB 100 TB (200 TB)	without backup *[2] with backup without backup
Total storage (Home) <number> **Typical use: Source code and scripts** **Lifetime: Duration of project**	Curie Fat Nodes or Thin Nodes or Hybrid Nodes Fermi Hornet MareNostrum SuperMUC	3 GB 50 GB 50 GB*[3] 100 GB 100 GB	with backup and snapshots (extensible on demand) no backup with backup and snapshots with backup and snapshots
Total storage (Archive) <number>	Curie Fat Nodes Thin Nodes Hybrid Nodes Fermi Hornet MareNostrum SuperMUC	100 TB *[4] *[5] 100 TB 100 TB *[6]	file size > 1 GB Typical file size should be > 5 GB

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

*[2] Numbers given are for a project requesting about 50 Million core hours on Hornet. Projects requiring less compute resources can only be granted analogical less storage space. More storage space is possible, but needs to be explicitly requested and justified in the proposal. In addition, this requirement needs to be discussed with the hosting site prior to proposal submission

*[3] The number given depends also on the number of users in the project.

*[4] To be arranged with CINECA staff, total storage lower than 10TB

*[5] Access to Hornet's archive needs a special agreement with HLRS and PRACE.

*[6] Long-term archiving or larger capacity must be negotiated separately with LRZ.

When requesting more than the specified scratch disk space and/or larger than 1TB 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 100TB 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>	Curie Fermi Hornet MareNostrum SuperMUC	2 Million 2 Million n.a. 4 Million 1 Million [*1]	days will be removed automatically
Number of files (Work) <number>	Curie Fermi Hornet MareNostrum SuperMUC	500.000 2 Million 2Million 2 Million 1 Million [*1]	Extensible on demand
Number of files (Home) <number>	Curie Fermi Hornet MareNostrum SuperMUC	n.a 100.000 100.000 10.000 100.000	with backup, includes the snapshots with backup, includes the snapshots
Number of files (Archive) <number>	Curie Fermi Hornet MareNostrum SuperMUC	100.000 10.000 10.000 1 Million 100.000	Extensible on demand * * *

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

[*1] Files must not be in a single directory but should be distributed in separate sub-directories.

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>**	Curie Fermi Hornet MareNostrum SuperMUC	100 TB 20 TB* 100 TB* 50 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).