TECHNICAL GUIDELINES FOR APPLICANTS TO PRACE 7th CALL (Tier-0)

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

- GCS@Jülich, Germany IBM Blue Gene/Q "JUQUEEN"
- GENCI@CEA, France Bull Bullx cluster "Curie"
- GCS@HLRS, Germany Cray XE6 "Hermit"
- GCS@LRZ, Germany IBM System X iDataplex "SuperMUC"
- 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

		Curie FN	*Curie TN*	*Curie HN*	*Fermi*	*Hermit*	*JUQUEEN*	*MareNostrum*	*SuperMUC TN*	*SuperMUC FN*
	System Type	Intel Bullx	Intel Bullx	Intel Bullx	Blue Gene/Q	Cray XE6	Blue Gene/Q	IBM System x iDataPlex	IBM System x iDataPlex	IBM BladeCenter HX5
Compute	Processor type	Nehalem EX 2,66 Ghz	SandyBridge EP 2,7 Ghz	Westmere EP 2,67 Ghz	IBM PowerPC A2 (1,6 GHz)	AMD Opteron 6276 (Interlagos)	IBM PowerPC® A2 1,6 GHz, 16 cores per node	Intel Sandy Bridge EP	Intel Sandy Bridge EP	Intel Westmere EX
	Total nb of nodes	90	5040	144	10.240	3552	28.672	3028	9216	205
	Total nb of cores	11.520	80.640	1152	163.840	113.664	458.752	48.448	147.456	8200
	Nb of accelerators / node	n.a.	n.a.	2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	Type of accelerator	n.a.	n.a.	Nvidia M2090	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Memory	Memory / Node	512	64	24	16 GB	32 GB/ 64 GB	16 GB	32	32	256
Network	Network Type	Infiniband QDR 10	Infiniband QDR 10	Infiniband QDR 10	IBM Custom	Cray Gemini	5D Torus network	Infiniband FDR10	Infiniband FDR10	Infiniband QDR
	Connectivity	Fat tree	Fat tree	Fat tree	5D Torus	3D Torus	5D Torus network	Fat Tree	Fat tree within island (8192 cores), pruned tree between islands	Fat Tree
Home file system	type	Nfs	Nfs	nfs	GPFS	NFS	GPFS	GPFS	NAS	NAS
	capacity	3 GB/user	3 GB/user	3 GB/user	100 TB	60 TB	1,1 PB	59 TB	1,5 PB	300 TB
Work file system	type	Lustre	Lustre	Lustre	NFS	Lustre	n.a.	GPFS	GPFS	NAS
	capacity	600 TB	600 TB	600 TB	230 TB	2,7 PB	n.a.	612 TB	7 PB	300 TB
Scratch file system	type	Lustre	Lustre	Lustre	GPFS	n.a.	GPFS	GPFS	GPFS	NAS
	capacity	3,4 PB	3,4 PB	3,4 PB	2 PB	n.a.	3,5 PB	1,1 PB	3 PB	300 TB
Archive	capacity	Unlimited	Unlimited	Unlimited	On demand	On demand	Unlimited	2,4 PB	30 PB	30 PB
Minimum required job size	Nb of cores	512	512	32	2048	2048	2048	1024	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

Hermit:
http://www.hlrs.de/systems/platforms/cray-xe6-hermit/

JUQUEEN:
http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN_node.html

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

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

Subsection for each system

Curie, GENCI@CEA
Curie encompasses three different partitions:

1. Curie Fat Nodes (FN): composed by 360 nodes, each node having 4 octo core Intel Nehalem EX processors 2,26 GHz, 4 GB/core (128 GB/node). These nodes are interconnected through an Infiniband QDR network.

   These 360 nodes are gathered into 90 super fat nodes by using a special hardware interface called BCS (Bull Coherent Switch) which allows to have per node a single system image and 16 octo core Intel Nehalem EX processors (for a total of 128 cores) and 4 GB/core (for a total of 512 GB of main memory).

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 which contains 2 GPU Nvidia M2090 coupled to 2 four cores CPU Westmere-EP clocked at 2,67 GHz, let 8 cores and 2 GPU / node and 1152 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.

Fermi, CINECA
The system is a 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 must use 16 tasks per node and hence the requested amount of memory/task cannot exceed 1 GB. Hybrid (multithread applications) cannot exceed 16 GB per task. However, in order to use the architecture efficiently, pure MPI codes (if requiring less than 512 MB per task) are highly recommended to use 32 (or even 64) tasks per node and hybrid codes the corresponding number of threads per task (from 32 to 64 threads per node).

For such kind of applications, the use of SMT mode will be favorably considered for the project technical evaluation.

Hermit, GCS@HLRS

Please be aware that 1 node consists of 32 cores (AMD Interlagos) and is equipped with 32 GB, i.e. each core has at most 1024 MB of main memory available. In order to use the architecture efficiently, pure MPI codes must use 32 tasks per node, hybrid codes must use the corresponding number of threads per task (up to 32 threads per node). Therefore, pure MPI applications must use less than 1024 MB per MPI task, hybrid codes not more than 32 GB per node in order to be suited for the architecture. On 480 nodes, 64 GB of memory are available. When using these, the given numbers double. The requirement to make use of these nodes has to be explicitly stated in the proposal.

JUQUEEN, GCS@Jülich

The Blue Gene/Q system JUQUEEN consists of 28 Racks (28.672 compute nodes). One node consists of 16 cores with four-fold SMT and is equipped with 16 GB of main memory.
Typically about 15 GB of main memory per compute node are available for user applications. This means each physical core has a bit less than 1 GB of main memory available.
In order to use the Blue Gene/Q architecture efficiently, pure MPI codes **must use 32 tasks per node**, hybrid codes must use the corresponding number of threads per task (from 32 to 64 threads per node in total).
Therefore, pure MPI applications must use less than 512 MB per MPI task, hybrid codes less than 16 GB per node in order to be suited for the architecture.

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.

SuperMUC, GCS@LRZ

The system consists of 18 thin node islands plus one fat node island which are connected with Infiniband technology. Each node within the thin islands consists of 16 cores. Physical memory is 2 GB per core from which typically 1,5 GB are available for application use. If this memory is not sufficient, in exceptional cases some cores of a node can be left idle to dedicate their memory to other cores. Fat Nodes have 40 cores and a total of 256 GB. However, only 205 nodes or 8200 cores exist.
Purpose of Thin and Fat nodes on SuperMUC:
Much more compute resources are available as Thin Nodes. Therefore, most of the computations should be done on the Thin Nodes. The purpose of the fat nodes is mainly for pre-/postprocessing or for such parts of a simulation which really need large memory. *Only in the exceptional case when large memory is needed for the whole simulation should one apply for the Fat Nodes as a standalone resource.*

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 use of the Tier-0 system has to be justified (with no clear justification, you are advised to apply for a Tier-1 access).

The maximum value is limited by the total number of core hours given for the 7th Call for the corresponding machine (see Call announcement text).

Any further limitation may be specified in the Call.

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 Hermit JUQUEEN MareNostrum SuperMUC Thin nodes Fat nodes	24 hours 24 hours 24 hours 24 hours 24 hours (12 hours)* 24 hours (6 hours)* 24 hours 48 hours 48 hours **

* This might be changed during project runtime, guaranteed minimum is the value in brackets.
** For large jobs with more than 8192 cores, it is requested that jobs can write checkpoints more frequently and can use these checkpoint for restart in case of crashes

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 (512 cores), 2 (4096 cores)
	Thin Nodes	50 (512 cores), 4 (8192 cores)
	Hybrid Nodes	10
	Fermi	10
	Hermit	29 and at maximum 60.000 cores all jobs together
	JUQUEEN	15
	SuperMUC Fat Nodes	1(520 cores), 1(4096 cores)
	Thin Nodes	10 (512 cores), 2 (8192 cores), 1(32.768 cores)
	MareNostrum	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>	Curie Fat Nodes	512
	Thin Nodes	512
	Hybrid Nodes	32
	Fermi	2048
	Hermit	2048
	JUQUEEN	2048
	MareNostrum	1024
	SuperMUC Fat Nodes	520
	Thin nodes	512

Expected number of cores (Average) <number>	Other systems Hermit JUQUEEN SuperMUC Fat Nodes Thin Nodes	see above > 2048 4096 1040 8192
Expected number of cores (Maximum) <number>	Other systems Hermit JUQUEEN SuperMUC Fat Nodes Thin Nodes	see above 60.000 (using up to 113.664 cores is possible, but should be requested in the proposal) >4096 (up to the complete system) 4160 32.768 *

Virtual cores (SMT is enabled) are not counted. *GPU based systems need special rules*.
* Larger Jobs may be possible after the initial installation phase.

Additional information:

JUQUEEN

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 JUQUEEN can use more than 2048 physical cores per job on average (at least 4096 physical cores).

Job sizes must use a multiple of 2048 physical cores in order to make efficient use of the architecture.

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. During the initial installation phase jobs should be limited to two islands (16.384 cores).
When running several jobs simultaneously, filling complete islands (approx. 8192 cores) should be possible but this is not mandatory.

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 Fermi Hermit JUQUEEN 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 1 GB * #cores 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 Hermit SuperMUC Thin Nodes Fat Nodes	see above 2GB* #cores (up to 15.360 cores) 1GB* #cores for the other cores 2 GB (1,5 GB) * #cores or 32 GB (24 GB)* #nodes 6,4 GB (5,8 GB) * #core or 256 GB (232 GB)* #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 over the 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>	Curie Fat Nodes or Thin Nodes or Hybrid Nodes	20 TB (100 TB)	without backup, automatic cleanup procedure
Typical use: Scratch files during simulation, log files, checkpoints	Fermi	20 TB (100 TB)	without backup, cleanup procedure for files older than 30 days
	Hermit	-	HLRS provides a special mechanism for Work spaces, see next row
Lifetime: Duration of jobs and between jobs	JUQUEEN	20 TB (100 TB)	without backup, files older than 90 days will be removed automatically
	MareNostrum	40 TB*[1]	without backup
	SuperMUC Thin nodes	100 TB (200 TB)	without backup, automatic cleanup procedure
	Fat nodes	5 TB (10 TB) *[2]	
Total storage (Work) <number>	Curie Fat Nodes Thin Nodes Hybrid Nodes	10 TB	
Typical use: Result and large input files	Fermi	20 TB	without backup
	Hermit	250 TB	*[3]
	JUQUEEN	n.a.	not available on JUQUEEN
Lifetime: Duration of project	MareNostrum	10 TB	with backup
	SuperMUC Thin nodes	100 TB (200 TB)	without backup
	Fat nodes	5 TB (10 TB) *[2]	
Total storage (Home) <number>	Curie Fat Nodes Thin Nodes Hybrid Nodes	3 GB (50 GB)	with backup and snapshots
Typical use: Source code and scripts	Fermi	50 GB	
	Hermit	50 GB*[4]	no backup
	JUQUEEN	6 TB	with backup
Lifetime: Duration of project	MareNostrum	100 GB	
	SuperMUC Thin nodes	100 GB	with backup and snapshots
	Fat nodes	100 GB	with backup and snapshots
Total storage (Archive) <number>	Curie Fat Nodes Thin Nodes Hybrid Nodes	100 TB	file size > 1 GB
	Fermi	*[5]	
	Hermit	*[6]	
	JUQUEEN	*[7]	Ideal file size: 500 GB – 1000 GB
	MareNostrum	100 TB	
	SuperMUC	100 TB *[8]	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] After integration of the fat nodes into the final system, this value can be increased up to the limit of the thin nodes.

*[3] Numbers given are for a project requesting about 60 Million core hours on Hermit. 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

*[4] The number given depends also on the number of users in the project.
*[5] To be arranged with CINECA staff
*[6] Access to Hermit's archive needs a special agreement with HLRS and PRACE.
*[7] Due to limited file system cache for archive not more than 10 TB/week should be moved to this storage.
*[8] 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	2 Million	
	Fermi	5 Million	
	Hermit	n.a.	
	JUQUEEN	4 Million	without backup, files older than 90 days
	MareNostrum	4 Million	will be removed automatically
	SuperMUC Thin nodes	1 Million	
	Fat nodes	100.000	
Number of files (Work) <number>	Curie	500.000 (4 Million)	
	Fermi	100.000	
	Hermit	4 Million	
	JUQUEEN	n.a.	
	MareNostrum	2 Million	
	SuperMUC Thin nodes	1 Million	
	Fat nodes	100.000	
Number of files (Home) <number>	Curie	n.a	
	Fermi	100.000	
	Hermit	100.000	
	JUQUEEN	2 Million	With backup
	MareNostrum	10.000	
	SuperMUC Thin nodes	100.000	This includes the snapshots
	Fat nodes	100.000	
Number of files (Archive) <number>	Curie	100.000 (1 Million)	
	Fermi	10.000	*
	Hermit	10.000	
	JUQUEEN	2 Million	Ideal file size: 500 GB – 1000 GB
	MareNostrum	1 Million	*
	SuperMUC	100.000	

* 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-0/1 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	100 TB
	Fermi	50 TB
	Hermit	100 TB*
	JUQUEEN	100 TB
	MareNostrum	50 TB
	SuperMUC	<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).