TECHNICAL GUIDELINES FOR APPLICANTS TO PRACE 16th CALL (Tier-0)

The contributing sites and the corresponding computer systems for this call are:

System	Architecture	Site (Country)	Core Hours (node hours)	Minimum request
Curie and Irene-SKL	Bull Bullx cluster (Curie) and BULL Sequana (Irene-SKL, starting service in 2H 2018)	GENCI@CEA (FR)	128 million (3.9 million)	15 million core hours
Irene - KNL	BULL Sequana (starting service in 2H 2018)	GENCI@CEA (FR)	57 million (0.8 million)	15 million core hours
Hazel Hen	Cray XC40 System	GCS@HLRS (DE)	70 million (2.9 million)	35 million core hours
Juqueen successor	Multicore cluster	GCS@JSC (DE)	70 million (tbd)	35 million core hours
Marconi-Broadwell	Lenovo System	CINECA (IT)	36 million (1 million)	15 million core hours
Marconi-KNL	Lenovo System	CINECA (IT)	442 million (6.5 million)	30 million core hours
MareNostrum	Lenovo System	BSC (ES)	475 million (10 million)	15 million core hours
Piz Daint	Cray XC50 System	CSCS (CH)	510 million (7.5 million)	68 million core hours Use of GPUs
SuperMUC	IBM System X iDataplex/ Lenovo NextScale	GCS@LRZ (DE)	105 million (6.6 million)	35 million core hours

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, as long as proper benchmarks and resource request justification are provided on each of the requested systems.

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.

A - General Information on the Tier-0 systems available for PRACE 16th Call

		Curie TN	*Irene - SKL*	*Irene - KNL*	*Hazel Hen*	*Successor of Juqueen*	*Marconi Broadwell*	*Marconi KNL*	*MareNostrum 4*	*Piz Daint*	*SuperMUC Phase 1*	*SuperMUC Phase 2*
System Type		Bullx	Bull Sequana	Bull Sequana	Cray XC40	tbd	Lenovo System NeXtScale	Lenovo System Adam Pass	Lenovo	Hybrid Cray xC50	IBM System x iDataPlex	Lenovo NeXtScale
Compute	Processor type	Intel SandyBridge EP 2.7 GHz	Intel Skylake EP 2.7 GHz	Intel Knights Landing	Intel Xeon E5-2680v3 (Haswell)	Multicore	Intel Broadwell	Intel Knights Landing	Intel Xeon Platinum 8160 2.1 GHz	Intel® Xeon® E5-2690 v3 @ 2.60GHz (12 cores)	Intel Sandy Bridge EP	Haswell Xeon E5-2697 v3 (Haswell)
	Total nb of nodes	5 040	1656	684	7 712	tbd	504	3 600	3 456	5 320	9 216	3 072
	Total nb of cores	80 640	79 488	46 512	185 088	tbd	18 144	244 800	165 888	63 840	147 456	86 016
	Nb of accelerators/node	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	1 GPU per node	n.a.	n.a.
	Type of accelerator	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	NVIDIA® Tesla® P100 16GB	n.a.	n.a.
Memory	Memory / Node	64 GB	192 GB DDR4	96 GB DDR4 + 16 GB MCDRAM	128 GB	min. 96 GB	128 GB - DDR4	96 GB – DDR4 + 16 GB - MCDRAM	96 GB (200 nodes with 384GB)	64 GB	32 GB	64 GB
Network	Network Type	Infiniband QDR	TBA*	TBA*	Cray Aries	tbd	Intel Omni-Path Architecture 2:1	Intel Omni-Path Architecture 2:1	Intel Omni-Path Architecture	Cray Aries	Infiniband FDR10	Infiniband FDR14
	Connectivity	Fat tree	Fat Tree	Fat Tree	Dragonfly	tbd	Fat Tree	Fat Tree	Fat Tree	Dragonfly	Fat tree within island (512 nodes) pruned tree between islands	Fat tree within island (512 nodes) pruned tree between islands

* To be announced later

		Curie	*Irene*	*Hazel Hen*	*Successor of Juqueen***	*Marconi Broadwell*	*Marconi KNL*	*MareNostrum 4*	*Piz Daint*	*SuperMUC Phase 1*	*SuperMUC Phase 2*
Home file system	type	NFS	NFS	NFS	GPFS	GPFS	GPFS	GPFS	GPFS	NAS *	NAS *
	capacity	8 TB	TBA	60 TB	1.8 PB	200 TB	200 TB	32 TB	86 TB	10 PB	10 PB
Work file system	type	Lustre	Lustre	Lustre	n.a.	GPFS	GPFS	GPFS	GPFS	GPFS *	GPFS *
	capacity	600 TB	TBA	15 PB	n.a.	7.1 PB	7.1 PB	4.3 PB	5.7 PB	12 PB	12 PB
Scratch file system	type	Lustre	Lustre	n.a.	GPFS	GPFS	GPFS	GPFS	Lustre	GPFS *	GPFS *
	capacity	3.4 PB	4.6 PB	n.a.	5.3 PB	2.5 PB	2.5 PB	8.7 PB	6.2 PB	5.2 PB *	5.2 PB *
Archive	capacity	Unlimited	Unlimited	On demand	Unlimited	On demand	On demand	5.0 PB	NA	On demand	On demand
Minimum required job size	Nb of cores	1 024	1 024	4 096		540	2 040	1 024	6 nodes	512	512

* SuperMUC Phase 1 and Phase 2 share the same file systems

** Data refers to current storage. The system available for Call 16 will be at least of this size

IMPORTANT REMARK:

Applicants are strongly advised to apply for PRACE Preparatory Access to collect relevant benchmarks and technical data for the system the wish to use through Project Access (note: if requesting resources on Piz Daint, it is mandatory to show benchmarking on this system in your submission). Further information and support from HPC Technical teams can be requested during the preparation of the application through PRACE Peer-Review at peer-review@prace-ri.eu or directly at the centres.

More details on the website of the centers:

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

Irene:
TBA

Hazel Hen:
http://www.hlrs.de/systems/cray-xc40-hazel-hen/

Juqueen:
http://www.fz-juelich.de/ias/jsc/DE/Leistungen/Supercomputer/supercomputer_node.html

Marconi:
http://www.hpc.cineca.it/hardware/marconi
https://wiki.u-gov.it/confluence/display/SCAIUS/UG3.1%3A+MARCONI+UserGuide

MareNostrum:
https://www.bsc.es/marenostrum/marenostrum/technical-information
https://www.bsc.es/user-support/mn4.php

Piz Daint:
http://user.cscs.ch/index.html

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

Subsection for each system

Curie, GENCI@CEA

The Curie BULLx system is composed by 5040 compute 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 and accessing to a multi-layer Lustre parallel filesystem at 250 GB/s. The peak performance of this system is 1.7 petaflops.

Irene, with a first tranche available starting H2 2018, will be a BULL Sequana system based on 9 compute cells. The Irene SKL partition will be composed by 6 cells each containing 272 compute nodes with two 24-core Intel Skylake EP processors 2.7 GHz, 4 GB/core (192 GB/node). The other partition Irene KNL will be composed by 3 cells each containing 228 nodes with one Intel Knights Landing 68-core 1.4 GHz manycore processor with 16 GB of high-speed memory (MCDRAM) and 96 GB of main memory. All the compute nodes are interconnected through a high speed interconnect (to be announced later) and accessing to a multi-layer Lustre parallel file system at 500 GB/s. The peak performance of this system will be close to 9 petaflops.

A smooth transition will be organized between Curie and Irene in the first phase of the 16th call. Allocations on Curie and Irene SKL are linked: projects will begin their calculations on Curie during the first three months of the call (this period may vary and is subject to change) and will then be gradually ported on Irene SKL. Allocations on Irene KNL are independent and computing time on this partition will be available starting H2 2018.

A second tranche is planned to be installed during Q2 2019.

Hazel Hen, GCS@HLRS

Hazel Hen, a Cray XC40-system, is at the heart of the high performance computing (HPC) system infrastructure of the HLRS. With a peak performance of 7.42 Petaflops (quadrillion floating point operations per second), Hazel Hen is currently the most powerful HPC system in Germany. The HLRS supercomputer, which was taken into operation in October 2015, is based on the Intel® Haswell Processor and the Cray Aries network and is designed for sustained application performance and high scalability.

Successor of Juqueen, GCS@JSC

JSC is currently procuring the successor of its BlueGene/Q system Juqueen. This system will be the first module of JSC's future modular Supercomputer Complex. It will be a general purpose cluster module, based on multicore CPUs. Only a few nodes will be equipped with accelerators. The system will be connected to the next generation of JSC's central storage system JUST. The system is expected to be available not before June 1st, 2018. Therefore allocations can only be made for 10 months.

Marconi, CINECA

Marconi system consists of three partitions, from which two will be available for Call 16:

- **Marconi - Broadwell** consists of 7 Lenovo NeXtScale racks with 72 nodes per rack. Each node contains 2 Broadwell processors each with 18 cores and 128 GB of DDR4 RAM.
- **Marconi - KNL** consists of 3 600 Intel server nodes integrated by Lenovo. Each node contains 1 Intel Knights Landing processor with 68 cores, 16 GB of MCDRAM and 96 GB of DDR4 RAM.

All the nodes (for both systems) will be connected via Intel Omni-Path network.

MareNostrum 4, BSC

MareNostrum 4 consists of 48 Compute Racks with 72 compute nodes per rack. Each node has two Intel Xeon Platinum 8160 processors with 2.1 GHz, 24 cores per socket (48 cores/node) and 96 GB of main memory (2 GB/core), connected via Intel Omni-Path fabric at 100 Gbits/s.

There are a subset of 200 fat nodes available that have 384 GB of main memory (8 GB/core). Their use is restricted to a maximum of 50% of their hours for all projects combined during each PRACE call.

Piz Daint, CSCS

Named after Piz Daint, a prominent peak in Grisons that overlooks the Fuorn pass, this supercomputer is a hybrid Cray XC50 system and is the flagship system for national HPC Service. The compute nodes are equipped with Intel® Xeon® E5-2690 v3 @ 2.60GHz (12 cores) and NVIDIA® Tesla® P100 16GB, and 64 GB of host memory.

The nodes are connected by the "Aries" proprietary interconnect from Cray, with a dragonfly network topology.

SuperMUC, GCS@LRZ

SuperMUC Phase 1 consists of 18 Thin Node Islands with Intel Sandy Bridge processors and one Fat Node Island with Intel Westmere processors. Each Island contains (512 compute nodes, each node having 16 physcial cores) 8192 cores for the user applications. Each of these cores has approx. 1.6 GB/core available for running applications. Peak performance is 3.1 PF. All compute nodes within an individual Island are connected via a fully non-blocking Infiniband network (FDR10 for the Thin Nodes and QDR for the Fat Nodes). A pruned tree network connects the Islands.

SuperMUC Phase 2 consists of 6 Islands based on Intel Haswell-EP processor technology (512 nodes/island, 28 physical cores/node and available memory 2.0 GB/core for applications, 3 072 nodes, 3.6 PF). All compute nodes within an individual Island are connected via a fully non-blocking Infiniband network (FDR14). A pruned tree network connects the Islands.

Both system phases share the same Parallel and Home filesystems.

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

Resource Usage

Computing time

To apply for PRACE Tier-0 resources there is a minimum of 15 million core hours for all systems, but one where the threshold is higher. Proposals, which do not comply with this requirement, should apply in the Tier-1 national calls.

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

Please justify the number of core (node) hours you request by providing a detailed work plan and the appropriate technical data on the systems of interest. Applicants are strongly invited to apply for PRACE Preparatory Access.

Once allocated, the project has to 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.

Tier-0 proposals are required to respect the minimum and maximum request of resources as indicated in the Terms of Reference to be found here.

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>	Hazel Hen Other systems Piz Daint	2500 hours * < 10 months -
Able to write checkpoints <check button>	All	Yes (apps checkpoints)
Maximum time between two checkpoints (= maximum wall clock time for a job) (hours) <number>	Hazel Hen SuperMUC All other systems	24 hours (12 hours)** 48 hours 24 hours

* This is the time a job really can use the CPUs. This limited time is mainly due to waiting times in the queue especially for jobs with a limited scalability (using less than 10,000 cores)

** 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	25 (1 024 cores), 4 (8 192 cores)
	Irene	25 (1 024 cores), 4 (8 192 cores) for the SKL partition and 10 (1 024 cores), 2 (8 192 cores) for the KNL partition
	Hazel Hen	29 and at maximum 96.000 cores all jobs together
	Successor of Juqueen	tbd
	Marconi	2-10 (depending on the job size)
	MareNostrum	Dynamic*
	Piz Daint	No shared nodes: 1 job per node maximum
	SuperMUC	10 (512 cores), 3 (8 192 cores), 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 (nodes) 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/nodes).

Please note that the values stated in the table below are absolute minimum requirements, allowed for small jobs, which should only be applicable to 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 (nodes) 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. Scaling benchmarks need to be representative of your study case and need to support your resource request on every system of interest (application to PRACE preparatory access project is strongly recommended). Missing scaling data may result in rejection of the proposal.***

Field in online form	*Machine*	*Min (cores)*
Expected job configuration (Minimum) <number>	Curie and Irene Hazel Hen Successor of Juqueen Marconi MareNostrum 4 Piz Daint SuperMUC	1 024 4 096 4 096 540 (Broadwell) 2 040 (KNL) 1 024 6 nodes 512 1 024
Expected number of cores (Average) <number>	Curie and Irene Hazel Hen Successor of Juqueen Marconi MareNostrum 4 Piz Daint SuperMUC	4 096 8 192 16 386 540 (Broadwell), 4 080 (KNL) 4 096 6 to 2 400 nodes 16 384 (Phase1), 7 168 (Phase2) 4 096
Expected number of cores (Maximum) <number>	Curie and Irene Hazel Hen Successor of Juqueen Marconi MareNostrum 4 Piz Daint SuperMUC	40 000 (full machine on demand) 40 000 for the SKL partition and 20 000 for the KNL partition (full machine on demand) 96 000 (using up to 180 000 cores is possible, but should be requested in the proposal) tbd 3 240 (Broadwell), 68 000 (KNL) 32 000 (for exceptional applications the usage of the full machine is possible) 4 400 nodes 65 536 (Phase1), 14 426 (Phase2)

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

Additional information:

Hazel Hen

The given number is the absolute minimum for a job. Job farming (starting several in principle independent jobs together) to reach this is NOT considered a valid measure by HLRS. We expect one job to make use of at least 4096 cores.

Marconi

The minimum number of (physical) cores per job is 2 040 (only for KNL partition. 540 on Broadwell).

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 Marconi can use at least a doubled value on average (some exceptions could be made for Broadwell partition, especially if the applicants will be able to launch several jobs at the same time: in this case a minimum number of 540 cores per job will be considered as sufficient).

The maximum number of (physical) cores per job is 68 000 (on Broadwell partition is 3240). Larger jobs are possible in theory (only if requested sending an email to superc@cineca.it), but the turnaround time is not guaranteed.

Please provide explicit scaling data of the codes you plan to work within your project. On KNL partition the scaling behavior at least up to 4080 physical cores must be shown and it must demonstrate a good

scalability at least up to 2040 physical cores. In any case, the applicant will have also to demonstrate that their codes can scale on a KNL system equipped with an Omnipath network at least up to the maximum number of required cores using a setup having size similar to the one proposed for the project. For hybrid (MPI+OpenMP) 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.

Proving to be able to scale on other architectures will be favorably considered for the technical evaluation of the project.

Piz Daint

Technical data needs to be provided on the Cray XC50, Piz Daint. To apply for Piz Daint use of **GPUs is a must**. Scalability, performance and technical data have to be sufficient to justify the resource request (≥1 million node hours).

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 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 and Irene	- Jobs should use a substantial fraction of the available memory
	Hazel Hen	- Jobs should use a substantial fraction of the available memory
	Sucessor of Juqueen	- Jobs should use a substantial fraction of the available memory
	Marconi	<118 GB per node (Broadwell), <90 GB per node (KNL)
	MareNostrum 4	- 2 GB * #cores or 8 GB * #cores (max 200 fat nodes)
	Piz Daint	- nodes are not shared
	SuperMUC	- Jobs should use a substantial fraction of the available memory
Memory (Average job) <number>	Curie and Irene	- Jobs should use a substantial fraction of the available memory
	Hazel Hen	- Jobs should use a substantial fraction of the available memory
	Sucessor of Juqueen	- Jobs should use a substantial fraction of the available memory
	Marconi	- <118 GB per node (Broadwell); <90 GB per node (KNL)
	MareNostrum	- 2 GB * #cores or 8 GB * #cores (max 200 fat nodes)
	Piz Daint	- 5.3 GB per core or 64 GB per node (nodes are not shared)
	SuperMUC	- Jobs should use a substantial fraction of the available memory

Memory (Maximum job) <number>	Curie and Irene Hazel Hen Sucessor of Juqueen Marconi MareNostrum4 Piz Daint SuperMUC	- 4 GB* #cores (or 64 GB * #nodes for CURIE and 192 GB * # nodes for Irene) - 5GB* #cores - 1 GB* #cores for the other cores - <88 GB per node - <118 GB per node (Broadwell), <90 GB per node (KNL) - 2 GB* #cores or 8 GB * #cores (max 200 fat nodes) - 5.3 GB per core or 64 GB per node (nodes are not shared) - 2 GB* #cores or 32 GB* #nodes (Phase1) - 2.2 GB* #cores or 64 GB* #nodes (Phase2)

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**	Curie and Irene	20 TB (100 TB)	- without backup, automatic clean-up procedure;
	Hazel Hen		- HLRS provides a special mechanism for Work spaces, without backup.
	Successor of Juqueen	20 TB (100 TB)	- without backup, files older than 90 days will be removed automatically
	Marconi	20 TB (100 TB)*1	- without backup, clean-up procedure for files older than 30 days;
	MareNostrum 4	100 TB (more on demand)	- without backup, clean-up procedure.
	Piz Daint	6.2 PB	- without backup, clean-up procedure. Quota on inodes
	SuperMUC	100 TB (200 TB)	- without backup, automatic clean-up procedure.
Total storage (Work) <number> **Typical use: Result and large input files** **Lifetime: Duration of project**	Curie and Irene	1 TB	
	Hazel Hen	250 TB	- *2
	Successor of Juqueen	n.a.	
	Marconi	20 TB (100 TB)*1	- Without backup;
	MareNostrum 4	10 TB (100TB)	- With backup;
	Piz Daint	5.7 PB	- Input not readable from compute nodes; data kept only for duration of project
	SuperMUC	100 TB (200 TB)	- Without backup.
Total storage (Home) <number> **Typical use: Source code and scripts** **Lifetime: Duration of project**	Curie and Irene	3 GB	- with backup and snapshots;
	Hazel Hen	50 GB *3	- no backup;
	Successor of Juqueen	6 TB	- with backup;
	Marconi	50 GB	- with backup;
	MareNostrum 4	20 GB	- with backup;
	Piz Daint	86 TB	- with backup and snapshots
	SuperMUC	100 GB	- with backup and snapshots.
Total storage (Archive) <number>	Curie and Irene	100 TB	- File size > 1 GB
	Hazel Hen	*4	
	Successor of Juqueen	*5	- Ideal file size: 500 GB – 1000 GB
	Marconi	20 TB (100 TB)*6	
	MareNostrum 4	100 TB	
	Piz Daint	n.a.	
	SuperMUC	100 TB*7	- Typical file size should be > 5 GB

*1 The default value is 1 TB. Please ask to CINECA User Support (superc@cineca.it) to increase your quota after the project will start.

*2 The number given depends also on the number of users in the project. A larger project space is possible but needs an agreement with HLRS before proposal submission.

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

*4 Access to Hazel Hen's archive needs a special agreement with HLRS and PRACE.

*5 Due to limited file system cache for archive not more than 10 TB/week should be moved to this storage.

*6 Not active by default. Please ask to CINECA User Support after the project will start

*7 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 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>	Curie and Irene	2 Million	- 10 000 files max per directory, without backup, files older than 90 days will be removed automatically
	Hazel Hen	n.a.	
	Successor of Juqueen	4 Million	- Without backup, files older than 90 days will be removed automatically
	Marconi	2 Million	- Without backup, files older than 90 days will be removed automatically
	MareNostrum	2 Million	
	Piz Daint	1 Million	- No limit while running, but limits in number of files left on scratch.
	SuperMUC	1 Million	- Without backup, old files are removed automatically, Ideal file size: >100 GB.
Number of files (Work) <number>	Curie and Irene	500 000	- Extensible on demand, 10 000 files max per directory
	Hazel Hen	24 Million	
	Successor of Juqueen	n.a.	
	Marconi	2 Million	
	MareNostrum	2 Million	
	Piz Daint	50 000 per TB	- with backup and snapshots
	SuperMUC	1 Million	- Ideal file size: >100 GB
Number of files (Home) <number>	Curie and Irene	n.a.	
	Hazel Hen	100.000	
	Successor of Juqueen	12 Million	- With backup
	Marconi	100 000	- With backup
	MareNostrum	100 000	
	Piz Daint	100 000	- With backup and snapshots
	SuperMUC	100 000	- With backup (snapshots)
Number of files (Archive) <number>	Curie and Irene	100 000	- Extensible on demand, typical file size should be > 1 GB
	Hazel Hen	10 000	
	Successor of Juqueen	2 Million	- Ideal file size: 500 GB – 1000 GB
	Marconi	10 000*	
	MareNostrum	1 Million*	
	Piz Daint	n.a.	
	SuperMUC	100 000	- Typical file size should be > 5 GB

*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>	Curie and Irene Hazel Hen Successor of Juqueen Marconi MareNostrum Piz Daint SuperMUC	100 TB 100 TB* 100 TB 20 TB* 50 TB Currently no limit 100 TB

* More is possible, but this 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).