# PRACE Project Access

# Technical Guidelines - 19th Call for Proposals

Peer-Review Office – Version 5 – 06/03/2019

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

| System | Architecture | Site (Country) | Core Hours (node hours) | Minimum request (core hours) |
|---|---|---|---|---|
| ***HAWK*** (1) | HPE System | GCS@HLRS (DE) | 460 million | 35 millon |
| ***Joliot Curie - KNL*** | BULL Sequana X1000 | GENCI@CEA (FR) | 101 million (1.5 million) | 15 million |
| ***Joliot Curie - SKL*** | BULL Sequana X1000 | GENCI@CEA (FR | 142 million (3 million) | 15 million |
| ***JUWELS*** | BULL Sequana X1000 | GCS@JSC (DE) | 70 million (1.5 million) | 35 million |
| ***Marconi-Broadwell*** (2) | Lenovo System | CINECA (IT) | 36 million (1 million) | 15 million |
| ***Marconi-KNL*** (3) | Lenovo System | CINECA (IT) | 408 million (6 million) | 30 million |
| ***MareNostrum 4*** | Lenovo System | BSC (ES) | 240 million (5 million) | 30 million |
| ***Piz Daint*** | Cray XC50 System | ETH Zurich/CSCS (CH) | 544 million (8 million) | 68 million Use of GPUs |

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.

(1) The project runtime for this platform is from 01 December 2019 to 30 September 2020.
(2) At the date of publishing this document (06/03/2019), the Broadwell partition of Marconi is under maintenance. Upon completion, some aspects of the machine could be changed. In any case, CINECA can ensure that the architecture, in terms of CPUs (Broadwell), memory per node (128 GB), and interconnection (Intel Omnipath), will not change.
(3) By the end of 2019, Marconi KNL will be replaced by a new and more powerful partition. Currently, there is not sufficient information about this new partition. In any case, when it will be ready, the awarded core hours in this Call will be converted using as a reference the comparison between the theoretical performances of the two architectures. It cannot be guaranteed that the new architecture will be similar to the existing one. For this reason, it is recommended that applicants prove that their codes are able to scale on different architectures (including eventually accelerators). This will be favourably considered in the technical evaluation of the project.

# A - General Information on the Tier-0 systems available for PRACE 19th Call

| | | HAWK | JOLIOT CURIE - SKL | JOLIOT CURIE - KNL | JUWELS | Marconi Broadwell[(1)] | Marconi KNL[(2)] | MareNostrum 4 | Piz Daint |
|---|---|---|---|---|---|---|---|---|---|
| System Type | | HPE | Bull Sequana | Bull Sequana | Bull Sequana | Lenovo System NeXtScale | Lenovo System Adam Pass | Lenovo | Hybrid Cray xC50 |
| Compute | Processor type | AMD Epyc Rome | Intel Xeon Platinum 8168 2.7 GHz | Intel Knights Landing | Intel Xeon Skylake Platinum 8168 | Intel Broadwell | Intel Knights Landing | Intel Xeon Platinum 8160 2.1 GHz | Intel® Xeon® E5-2690 v3 @ 2.60GHz (12 cores) |
| Compute | Total nb of nodes | 5 000 | 1 656 | 828 | 2 511 | 720 | 3 600 | 3 456 | 5 704 |
| Compute | Total nb of cores | 640 000 | 79 488 | 52 992 | 120 528 | 25 920 | 244 800 | 165 888 | 68 448 |
| Compute | Nb of accelerators /node | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | 1 GPU per node |
| Compute | Type of accelerator | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | n.a. | NVIDIA® Tesla® P100 16GB |
| Memory | Memory / Node | 128 GB | 192 GB DDR4 | 96 GB DDR4 + 16 GB MCDRAM | 96 GB | 128 GB - DDR4 | 96 GB – DDR4 + 16 GB - MCDRAM | 96 GB (200 nodes with 384GB) | 64 GB |
| Network | Network Type | Infiniband HDR | Infiniband EDR | BULL BXI | InfiniBand EDR | Intel Omni-Path Architecture 2:1 | Intel Omni-Path Architecture 2:1 | Intel Omni-Path Architecture | Cray Aries |
| Network | Connectivity | 9 D enhanced Hypercube | Fat Tree | Fat Tree | Fat Tree | Fat Tree | Fat Tree | Fat Tree | Dragonfly |

(1) See note 2 on page 1.
(2) See note 3 on page 1.

| | | *HAWK* | *JOLIOT CURIE* | *JUWELS* | *Marconi Broadwell*[(1)] | *Marconi KNL*[(2)] | *MareNostrum 4* | *Piz Daint* |
|---|---|---|---|---|---|---|---|---|
| **Home file system** | type | NFS | NFS | GPFS | GPFS | GPFS | GPFS | GPFS |
| | capacity | 100 TB | TBA | 2.8 TB | 200 TB | 200 TB | 32 TB | 86 TB |
| **Work file system** | type | Lustre | Lustre | GPFS | GPFS | GPFS | GPFS | GPFS |
| | capacity | 25 PB | TBA | 2.3 PB | 7.1 PB | 7.1 PB | 4.3 PB | 4.7 PB |
| **Scratch file system** | type | n.a. | Lustre | GPFS | GPFS | GPFS | GPFS | Lustre |
| | capacity | n.a. | 4.6 PB | 9.1 PB | 2.5 PB | 2.5 PB | 8.7 PB | 8.8 PB |
| **Archive** | capacity | On demand | Unlimited | On demand | On demand | On demand | n.a. | n.a. |
| **Minimum required job size** | Nb of cores | 8 192[(*)] | 1 024 | | 540 | 2 040 | 1 024 | 6 nodes |

[(*)] Smaller jobs are NOT accepted in production run.
[(1)] See note 2 on page 1.
[(2)] See note 3 on page 1.

### *IMPORTANT REMARKS*

- Applicants are strongly advised to apply for PRACE Preparatory Access to collect relevant benchmarks and technical data for the system they 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 high performance computing (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.
- Please contact the peer review office of PRACE at peer-review@prace-ri.eu in order to request assistance from the high-level support team, at least 1 month before the submission deadline.

More details on the website of the centres:

**HAWK:**

https://www.hlrs.de/systems/

**JOLIOT CURIE:**

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

**JUWELS:**

http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUWELS/JUWELS_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:**

https://www.cscs.ch/computers/piz-daint (Brief description of the System)

https://user.cscs.ch (User Portal)

# Subsection for each system

**HAWK, GCS@HLRS**

HAWK is the new HPC system at HLRS and will become available in December 2019. HAWK provides 5000 nodes, each one equipped with next generation AMD processors, code name Rome, offering 128 cores and 128 GByte of main memory per node. The nodes are connected with a next generation Infiniband network.

The project runtime for this platform is from 01 December 2019 until 30 September 2020.

**JOLIOT CURIE, GENCI@CEA**

JOLIOT CURIE is a BULL Sequana system X1000 based on 9 compute cells integrated into 2 partitions:

- **The SKL partition** is composed of 6 cells, each containing 272 compute nodes with two 24-core Intel Skylake Platinium 8168 processors 2.7 GHz, 4 GB/core (192 GB/node). These 6 cells are interconnected by an Infiniband EDR 100 Gb/s high speed network;
- **The KNL partition** is composed of 3 cells, each containing 276 nodes with one Intel Knights Landing 68-core 7250 1.4 GHz manycore processor with 16 GB of high-speed memory (MCDRAM) and 96 GB of main memory. These 3 cells are interconnected by a BULL BXI 100 Gb/s high speed network. A KNL node provides 64 cores for user jobs and keeps 4 cores for the system. A node is configured in quadrant for the cluster node and in cache mode for the memory. The threadmultiple mode for the hybrid program is not yet supported.

This configuration is completed with 5 fat nodes for pre/post processing (3 TB of memory each and a fast local storage based on NVMe) and 20 hybrid nodes used for remote visualisation.

These resources are federated across a multi-layer shared Lustre parallel filesystem with a first level (/scratch) of more than 5 PB at 300 GB/s.

The peak performance of this system is 9 petaflops.

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

**JUWELS, GCS@JSC**

JUWELS (Jülich Wizard for European Leadership Science) is designed as a modular system. The JUWELS Cluster module, supplied by Atos, based on its Sequana architecture, consists of about 2500 compute nodes, each with two Intel Xeon 24-core Skylake CPUs and 96 GiB of main memory. The compute nodes are interconnected with a Mellanox EDR InfiniBand interconnect. The peak performance of this CPU based partition is 10.4 petaflops. A booster module, optimized for massively parallel workloads, is currently scheduled for the beginning of 2020.

**Marconi, CINECA**

Currently, Marconi system consists of three partitions, from which two will be available for Call 19:

- **Marconi - Broadwell** consists of 10 Lenovo NeXtScale racks with 72 nodes per rack. Each node contains 2 Broadwell processors each with 18 cores and 128 GB of DDR4 RAM. This partition is currently in maintenance. When it will be completed, some aspect of the machine could be changed. In any case we can ensure that the architecture, both in terms of CPUs (Broadwell), memory per node (128 GB), and interconnection (Intel Omnipath), will not change.

- **Marconi - KNL** consists of 3600 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) are connected via Intel Omni-Path network. The aggregate peak performance of this system is more than 20 petaflops. By the end of 2019, Marconi KNL will be replaced by a new and more powerful partition. Currently, there is not sufficient information about this new partition. In any case, when it will be ready, the awarded core hours will be converted using as a reference the comparison between the theoretical performances of the two architectures. Currently, it cannot be guaranteed that the new architecture will be similar to the existing one. For this reason, it is recommended that applicants prove that their codes are able to scale on different architectures (including eventually accelerators). This will be favourably considered in the technical evaluation of the project.

### 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, ETH Zurich/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.

# B – Guidelines for filling-in the online form

## Resource Usage

### Computing time

To apply for PRACE Tier-0 resources there is a minimum amount of core hours for all systems. 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 and proportionally across the duration of the allocation.

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)**<br>**<number>** | Marconi<br>Piz Daint<br>HAWK<br>Other systems | 4 000 hours(*)<br>-<br>1 000 hours(*)<br>< 10 months |
| **Able to write checkpoints**<br>**<check button>** | All | Yes (apps checkpoints) |

| | | |
|---|---|---|
| **Maximum time between two checkpoints (= maximum wall clock time for a job) (<u>hours</u>) <number>** | All systems | 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).

### 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>** | Joliot-Curie | 25 (1 024 cores), 4 (8 192 cores)<br>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 |
| | JUWELS | 3 (more on demand) |
| | Marconi | 2-20 (depending on the job size) |
| | MareNostrum 4 | Dynamic(*) |
| | Piz Daint | No shared nodes: 1 job per node maximum |
| | HAWK | 15 (up to max. 320 000 cores)(**) |

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

(**) More cores/job available by special agreement

### 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. **<u>Typical production jobs should run at larger scale.</u>**

**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 behaviour 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 and mandatory on Piz Daint). Missing technical (scaling, etc.) data may result in rejection of the proposal.**

| *Field in online form* | *Machine* | *Min (cores)* |
|---|---|---|
| **Expected job configuration (<u>Minimum</u>)**<br>**<number>** | Joliot-Curie<br>JUWELS<br>Marconi<br>MareNostrum 4<br>Piz Daint<br>HAWK | 1 024<br>4 608<br>540 (Broadwell), 2 040 (KNL)<br>1 024<br>6 nodes<br>8 192 cores |
| **Expected number of cores (<u>Average</u>)**<br>**<number>** | Joliot-Curie<br>JUWELS<br>Marconi<br>MareNostrum 4<br>Piz Daint<br>HAWK | 4 096<br>9 216<br>540 (Broadwell), 4 080 (KNL)<br>4 096<br>6 to 2 400 nodes<br>32 768 cores |
| **Expected number of cores (<u>Maximum</u>)**<br>**<number>** | Joliot-Curie<br><br><br>JUWELS<br>Marconi<br>MareNostrum 4<br><br>Piz Daint<br>HAWK | 40 000 (full machine on demand)<br>40 000 for the SKL partition and 20 000 for the KNL partition (full machine on demand)<br>24 576 (full machine on demand)<br>3 240 (Broadwell), 68 000 (KNL)<br>32 000 (for exceptional applications the usage of the full machine is possible)<br>4 400 nodes<br>327 680 cores |

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

## Additional information:

### Marconi

The minimum number of (physical) cores per job is 2 040 on KNL partition and 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 3 240). 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 behaviour at least up to 4 080 physical cores must be shown and it must demonstrate a good scalability at least up to 2 040 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 favourably considered for the technical evaluation of the project.

Considering the major upgrades of Marconi that will be completed by the end of 2019, proving to be able to scale on other architectures will be favourably 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). All technical data on Piz Daint must be in **node hours**.

| *Field in online form* | *Machine* | *Max* |
|---|---|---|
| **Memory (Minimum job) <number>** | Joliot-Curie | - Jobs should use a substantial fraction of the available memory |
| | JUWELS | - <92 GB per node |
| | Marconi | <118 GB per node (Broadwell), <86 GB per node (KNL) |
| | MareNostrum 4 | - 2 GB * #cores or 8 GB * #cores (max 200 fat nodes) |
| | Piz Daint | - nodes are not shared |
| | HAWK | - No requirements |

| | | |
|---|---|---|
| **Memory (Average job)<br><number>** | Joliot-Curie<br><br>JUWELS<br>Marconi<br>MareNostrum 4<br>Piz Daint<br><br>HAWK | - Jobs should use a substantial fraction of the available memory<br>- <92 GB per node<br>- <118 GB per node (Broadwell); <86 GB per node (KNL)<br>- 2 GB * #cores or 8 GB * #cores (max 200 fat nodes)<br>- 5.3 GB per core or 64 GB per node (nodes are not shared)<br>- No requirement |
| **Memory (Maximum job)<br><number>** | Joliot-Curie<br><br>JUWELS<br>Marconi<br>MareNostrum 4<br>Piz Daint<br><br>HAWK | - 4 GB* #cores (or 64 GB * #nodes for CURIE and 192 GB * # nodes for Joliot Curie)<br>- <92 GB per node<br>- <118 GB per node (Broadwell), <86 GB per node (KNL)<br>- 2 GB* #cores or 8 GB * #cores (max 200 fat nodes)<br>- 5.3 GB per core or 64 GB per node (nodes are not shared)<br>- 120 GB per node |

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 (6 on Marconi) 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 months (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 centre beforehand for approval**.

| *Field in online form* | *Machine* | *Max per project* | *Remarks* |
|---|---|---|---|
| **Total storage (Scratch) <number>**<br>**Typical use: Scratch files during simulation, log files, checkpoints**<br>**Lifetime: Duration of jobs and between jobs** | Joliot-Curie<br>JUWELS<br>Marconi<br>MareNostrum 4<br>Piz Daint<br>HAWK | 20 TB (100 TB)<br>90 TB<br>20 TB (100 TB)(*1)<br>100 TB (more on demand)<br>8.8 PB<br>n.a. | - without backup, automatic clean-up procedure<br>- without backup, files older than 90 days will be removed automatically<br>- without backup, clean-up procedure for files older than 50 days<br>- without backup, clean-up procedure<br>- without backup, clean-up procedure, Quota on inodes (max 1 Million) |
| **Total storage (Work) <number>**<br>**Typical use: Result and large input files**<br>**Lifetime: Duration of project** | Joliot-Curie<br>JUWELS<br>Marconi<br>MareNostrum 4<br>Piz Daint<br>HAWK | 1 TB<br>16 TB<br>20 TB (100 TB)(*1)<br>10 TB (100TB)<br>250 TB (500 TB)(*2)<br>100 TB(*3) | - With backup<br>- Without backup<br>- With backup<br>- Read-only from compute nodes data kept only for duration of project<br>- Without backups |
| **Total storage (Home) <number>**<br>**Typical use: Source code and scripts**<br>**Lifetime: Duration of project** | Joliot-Curie<br>JUWELS<br>Marconi<br>MareNostrum 4<br>Piz Daint<br>HAWK | 3 GB<br>10 GB<br>50 GB<br>20 GB<br>10 GB x User<br>10 GB / user | - with backup and snapshots<br>- with backup<br>- with backup<br>- with backup<br>- with backup and snapshots<br>- with snapshots |
| **Total storage (Archive) <number>** | Joliot-Curie<br>JUWELS<br>Marconi<br>MareNostrum 4<br>Piz Daint<br>HAWK | 100 TB<br>(*4)<br>20 TB (100 TB)(*5)<br>n.a.<br>n.a.<br>Upon agreement | - File size > 1 GB<br>- Ideal file size: 500 GB – 1000 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) From 250 to maximum of 500 TB will be granted if the request is fully justified and a plan for moving the data is provided.
(*3) More workspace storage upon special request/agreement. On HAWK there is a disk accounting system in place. If the relative amount of allocated disk space is larger than the relative amount of used compute cycles per month, the allocated extra disk space will be accounted and deducted from the allocation.
(*4) Access to JUWELS' archive needs a special agreement with JSC and PRACE.
(*5) Not active by default. Please ask to CINECA User Support after 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 centre beforehand for approval.

| ***Field in online form*** | ***Machine*** | ***Max*** | ***Remarks*** |
|---|---|---|---|
| **Number of files (Scratch)** **<number>** | Joliot-Curie | 2 Million | - 10 000 files max per directory, without backup, files older than 90 days will be removed automatically |
| | JUWELS | 4 Million | - Without backup, files older than 90 days will be removed automatically |
| | Marconi | 2 Million | - Without backup, files older than 50 days will be removed automatically |
| | MareNostrum 4 | 2 Million | |
| | Piz Daint | 1 Million | - No limit while running, but job submission is blocked if the max number of files left on scratch is reached |
| | HAWK | n.a. | |
| **Number of files (Work)** **<number>** | Joliot-Curie | 500 000 | - Extensible on demand, 10 000 files max per directory |
| | JUWELS | 3 Million | - With backup |
| | Marconi | 2 Million | |
| | MareNostrum 4 | 2 Million | |
| | Piz Daint | 50 000 per TB | - with backup and snapshots |
| | HAWK | 100 000 | |
| **Number of files (Home)** **<number>** | Joliot-Curie | n.a. | |
| | JUWELS | 40 000 | - With backup |
| | Marconi | 100 000 | - With backup |
| | MareNostrum 4 | 100 000 | |
| | Piz Daint | 100 000 | - With backup and snapshots |
| | HAWK | 100 000 | |

| | | | |
|---|---|---|---|
| **Number of files (Archive)** **<number>** | Joliot-Curie | 100 000 | - Extensible on demand, typical file size should be > 1 GB |
| | JUWELS | 100 000 | - Ideal file size: 500 GB – 1 000 GB |
| | Marconi | 10 000(*) | |
| | MareNostrum 4 | n.a. | |
| | Piz Daint | n.a. | |
| | HAWK | Upon request | - 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>** | Joliot-Curie | 100 TB |
| | JUWELS | 100 TB |
| | Marconi | 20 TB(*) |
| | MareNostrum 4 | 50 TB |
| | Piz Daint | Currently no limit |
| | HAWK | Max 10 Gb/s |

(*) 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).