PRACE Project Access

Technical Guidelines – 24th Call for Proposals

Peer-Review Office – Version 1 – 25/10/2021

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*	HPE Apollo	GCS@HLRS (DE)	345.6 million (2.7 million)	100 million
Joliot-Curie KNL	BULL Sequana X1000	GENCI@CEA (FR)	37.5 million (0.6 million)	15 million
Joliot-Curie Rome	BULL Sequana XH2000	GENCI@CEA (FR)	195.3 million (1.5 million)	15 million
Joliot-Curie SKL	BULL Sequana X1000	GENCI@CEA (FR)	52.9 million (1.1 million)	15 million
JUWELS Booster**[1)]	BULL Sequana XH2000	GCS@JSC (DE)	85.2 million (1.78 million)	7 million Use of GPUs
JUWELS Cluster*	BULL Sequana X1000	GCS@JSC (DE)	35.04 million (0.73 million)	35 million
Marconi100**[2)]	IBM Power 9 AC922 Whiterspoon	CINECA (IT)	165 million (0.47 million)	35 million Use of GPUs
MareNostrum 4*	Lenovo System	BSC (ES)	TBA	30 million
Piz Daint**[3)]	Cray XC50 System	ETH Zurich/CSCS (CH)	510 million (7.5 million)	68 million Use of GPUs
SuperMUC-NG*	Lenovo ThinkSystem	GCS@LRZ (DE)	91 million	35 million

**At the time of opening the call, the volume of resources offered on the corresponding system cannot be definitively confirmed. The final volume is expected to be similar to previous calls and will be announced later.*

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 factor between the amount of core hours towards the available node hours on the JUWELS Booster is given by the number of host CPUs on each individual node (48 host CPUs per node).

[2)] On Marconi100 each node has 2 Power 9 processors each with 16 cores and 4 Nvidia V100 GPUs each with 80 streaming multiprocessors. Therefore, the Marconi100 cluster has 352 equivalent cores per node. This number of cores must be used in the budget estimation following the formula: Cumulative Core hours = 352*Node hours = 352 * (GPU hours / 4).

[3)] On Piz Daint each Cray XC50 node features a 12-core Intel Haswell processor and a Nvidia P100 GPU with 56 streaming multiprocessors. Therefore, Piz Daint has 68 equivalent cores per node and one node hour is equivalent to 68 core hours.

A - General Information on the Tier-0 systems available for PRACE 24th Call

		HAWK	Joliot-Curie KNL	Joliot-Curie Rome	Joliot-Curie SKL	JUWELS Booster	JUWELS Cluster	Marconi100	Mare Nostrum 4	Piz Daint	SuperMUC-NG
System Type		HPE	Bull Sequana	Bull Sequana	Bull Sequana	Bull Sequana	Bull Sequana	IBM Power 9 AC922 Whiterspoon	Lenovo	Hybrid Cray xC50	Lenovo ThinkSystem
Compute	Processor type	AMD Epyc Rome	Intel Knights Landing	AMD Epyc Rome	Intel Xeon Platinum 8168 2.7 GHz	AMD EPYC Rome	Intel Xeon Skylake Platinum 8168	2 * IBM POWER9 AC922 at 3.1 GHz per node	Intel Xeon Platinum 8160 2.1 GHz	Intel® Xeon® E5-2690 v3 @ 2.60GHz (12 cores)	Intel Skylake Xeon Platinum 8174
	Total nb of nodes	5 632	828	2 292	1 656	936	2 511	980	3 456	5 704	6 480
	Total nb of cores	720 896	52 992	293 376	79 488	44 928	120 528	31 360	165 888	68 448	311 040
	Nb of accelerators /node	n.a.	n.a.	n.a	n.a.	4	n.a.	4 GPU per node	n.a.	1 GPU per node	n.a.
	Type of accelerator	n.a.	n.a.	n.a	n.a.	NVIDIA® Ampere A100, 40 GB HBM2e	n.a.	NVIDIA® Volta® V100, Nvlink 2.0, 16GB	n.a.	NVIDIA® Tesla® P100 16GB	n.a.
Memory	Memory / Node	256 GB	96 GB DDR4 + 16 GB MCDRAM	256 GB	192 GB	512 GB	96 GB	256 GB DDR4 + up to 1.6 TB NVMe Memory per node	96 GB (200 nodes with 384GB)	64 GB	96 GB
Network	Network Type	Infiniband HDR	BULL BXI	Infiniband HDR 100	Infiniband EDR	InfiniBand HDR	InfiniBand EDR	Mellanox Infiniband EDR	Intel Omni-Path Architecture	Cray Aries	Intel Omni-Path Architecture
	Connectivity	9D enhanced Hypercube	Fat Tree	Dragonfly+	Fat Tree	Dragonfly+	Fat Tree	DragonFly+	Fat Tree	Dragonfly	Fat tree within island (786 nodes) pruned tree between islands

(*)Smaller jobs are NOT accepted in production runs. Jobs can only be requested with defined node numbers (64, 128, 256, 512, 1024, 2048 and 4096) in regular operation (1 node = 128 cores)
For details, please see https://kb.hlrs.de/platforms/index.php/Batch_System_PBSPro_(Hawk)#Topology_aware_scheduling

		HAWK	*Joliot-Curie*	*JUWELS*	*Marconi100*	*MareNostrum 4*	*Piz Daint*	*SuperMUC-NG*
Home file system	type	NFS	NFS	GPFS	GPFS	GPFS	GPFS	GPFS
	capacity	100 TB	0.5 TB	2.8 TB	200 TB	32 TB	160 TB	256 TB
Work file system	type	Lustre	Lustre	GPFS	GPFS	GPFS	GPFS	GPFS
	capacity	25 PB	9.2 PB	2.3 PB	3 PB	4.3 PB	6.3 PB	33 PB
Scratch file system	type	n.a.	Lustre	GPFS	GPFS	GPFS	Lustre	GPFS
	capacity	n.a.	5.2 PB	9.1 PB	2 PB	8.7 PB	8.8 PB	17 PB
Archive	capacity	On demand	On demand	On demand	On demand	n.a.	n.a.	On demand
Minimum required job size	Nb of cores	8 192(*)	1 024		2 nodes	1024	6 nodes	960

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/hpe-apollo-hawk/

Joliot-Curie:

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

JUWELS:

http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUWELS/JUWELS_node.html

Marconi100:

https://wiki.u-gov.it/confluence/display/SCAIUS/UG3.2%3A+MARCONI100+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)

SuperMUC-NG:

https://doku.lrz.de/display/PUBLIC/SuperMUC-NG

Subsection for each system

HAWK, GCS@HLRS

HAWK is the new HPC system at HLRS. HAWK provides 5632 nodes, each one equipped with 2nd generation AMD EPYC 7742 processors (Rome), offering 128 cores and 256 GByte of main memory per node. The nodes are connected with an Infiniband HDR network (200 Gb/s per node). The network topology as an enhanced 9D-Hypercube. Disk storage is a DDN Lustre system with 25 PB capacity. For pre- and post-processing there are several nodes with high memory capacity available.

Joliot-Curie, GENCI@CEA

The successor of Curie, Joliot-Curie is a BULL Sequana X1000/XH2000 system based on 14 compute cells integrated into 3 partitions:

- 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 Rome partition is composed of 5 cells containing 2292 nodes with two 64-core AMD Epyc 2nd gen (Rome) processors 2.5 GHz, 2 GB/core (256 GB/node). These cells are interconnected by an Infiniband HDR 100 Gb/s high speed network.
- The SKL partition is composed of 6 cells, each containing 272 compute nodes with two 24-core Intel Skylake 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.

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 and 32 nodes (each with 4 GPUs nVIDIA V100) for HPDA/AI workloads.

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 23 petaflops.

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, also based on the Sequana platform by Atos and optimized for massively parallel workloads, is added in 2020. It offers 936 nodes, AMD EPYC host CPUs and the latest generation on NVIDIA Ampere A100 GPUs. The Cluster and Booster are tightly integrated in the same InfiniBand fabric.

Marconi100, CINECA

Marconi100 is an IBM machine with 980 nodes (+ 8 login). Each node is equipped with 2 IBM POWER9 AC922 at 3.1 GHz (32 cores per node), 4 NVIDIA Volta V100 GPUs, Nvlink 2.0, 16GB, 256 GB of DDR4 RAM and 1.6 TB of NVMe Memory. The used network is a Mellanox Infiniband EDR with DragonFly+ topology.

IMPORTANT REMARK: Due to the concurrent presence of 32 cores on Power9 CPUs and 320 streaming multiprocessor on the GPUs, the Marconi100 cluster is considered to have 352 equivalent physical cores per node. This number of cores must be used in the budget estimation following the formula: Cumulative Core hours = 352*Node hours = 352 * (GPU hours / 4). Please do not use standard core hours, and report in the detailed document, when possible, also the estimation in NODE hours or in GPU hours.

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 is 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. Please read carefully the additional information of Piz Daint on page 11 (section B) to provide correctly the required technical data for Piz Daint.

SuperMUC-NG, GCS@LRZ

SuperMUC-NG provides 6480 Lenovo ThinkSystem dual-socket nodes equipped with 24 core Intel Skylake Xeon Platinum 8174 processors and 96 GB of main memory. A subset of 144 fat nodes yields 768 GB of main memory each. The nodes are connected via a fat-tree Omni-Path network.

The peak performance is at 26.7PF.

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) <number>	HAWK	1 000 hours(*)
	Marconi100	-
	Piz Daint	-
	SuperMUC-NG	2500 hours(*)
	Other systems	< 10 months
Able to write checkpoints <check button>	SuperMUC-NG	48 hours
	Other systems	Yes (apps checkpoints)
Maximum time between two checkpoints (= maximum wall clock time for a job) (hours) <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>	HAWK	15 (up to max. 524 288 cores)(*)
	Joliot-Curie	10 (1 024 cores), 2 (8 192 cores) for the KNL partition 25 (1 024 cores), 4 (8 192 cores) for the Rome partition 25 (1 024 cores), 4 (8 192 cores) for the SKL partition
	JUWELS Booster & JUWELS Cluster	3 (more on demand)
	Marconi100	2-20 (depending on the job size)
	MareNostrum 4	Dynamic(**)
	Piz Daint	No shared nodes: 1 job per node maximum
	SuperMUC-NG	5 (up to 768 nodes), 2 (up to 3072 nodes)

(*) More cores/job available by special agreement

(**) 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 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, since technical data needs to be provided on the Cray XC50 (see additional information on page 11). Missing technical data (scaling, etc.) may result in rejection of the proposal.**

Field in online form	*Machine*	*Min (cores)*
Expected job configuration (Minimum) <number>	HAWK	8 192 cores
	Joliot-Curie	1 024
	JUWELS	4 608 (Cluster), 24 nodes (Booster)
	Marconi100	2 nodes
	MareNostrum 4	1 024
	Piz Daint	6 nodes
	SuperMUC-NG	960
Expected number of cores (Average) <number>	HAWK	32 768 cores
	Joliot-Curie	4 096
	JUWELS	9 216 (Cluster), 96 nodes (Booster)
	Marconi100	>20 nodes (using 1 or more jobs at the same time)
	MareNostrum 4	4 096
	Piz Daint	6 to 2 400 nodes
	SuperMUC-NG	9 600
Expected number of cores (Maximum) <number>	HAWK	262 144 cores (524 288 on demand)
	Joliot-Curie	40 000 for the SKL & Rome partitions, 20 000 for the KNL partition (full machine on demand)
	JUWELS	24 576 on the Cluster (full machine on demand) 228 nodes on the Booster
	Marconi100	256 nodes (more could be possible, but only if approved by the hosting site and only for few jobs per project)
	MareNostrum 4	32 000 (for exceptional applications the usage of the full machine is possible)
	Piz Daint	4 400 nodes (prior agreement with CSCS staff is required to run jobs over 2400 nodes)
	SuperMUC-NG	147 456 (half machine; full machine on demand)

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

Additional information:

JUWELS Booster

To apply for the JUWELS Booster use of **GPUs is a must**. For the calculation of the necessary core hours, each node hour on the JUWELS Booster is calculated as 48 core hours which represents the number of host CPUs per node (even if these are not used for the calculation itself).

Marconi100

To apply for Marconi100 use of **GPUs is a must**. Scalability, performance and technical data have to be sufficient to justify the resource request. We will accept benchmarks performed only on very similar machines (Power9+V100). In any case the scalability at least up to the same number of GPUs to be used for production runs must be reported. A detailed description of the method used to estimate the requested budget must be reported.

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 provided in **node hours** therefore the breakdown of the resource request linked to the benchmark data of Piz Daint must be provided in node hours within the proposal. The equivalent number of core hours required in the PRACE submission form can be obtained multiplying the resource request expressed in node hours by the conversion factor 68 (1 node hour = 68 core hours).

SuperMUC-NG

The minimum number of (physical) cores per job is 960. However, it is expected that PRACE projects applying for this system can use more than 6 144 physical cores per job. When running several jobs simultaneously filling complete islands should be possible, but this is not mandatory.

Field in online form	*Machine*	*Max*
Memory (Minimum job) <number>	HAWK	No requirements
	Joliot-Curie	Jobs should use a substantial fraction of the available memory
	JUWELS	<92 GB per node on the Cluster; <500G on the Booster
	Marconi100	No requirements
	MareNostrum 4	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-NG	Jobs should use a substantial fraction of the available memory
Memory (Average job) <number>	HAWK	No requirement
	Joliot-Curie	Jobs should use a substantial fraction of the available memory
	JUWELS	<92 GB per node on the Cluster; <500GB on the Booster
	Marconi100	No requirements
	MareNostrum 4	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-NG	Jobs should use a substantial fraction of the available memory
Memory (Maximum job) <number>	HAWK	240 GB per node
	Joliot-Curie	4 GB per core SKL, 2GB per core Rome, 1.4 per core KNL
	JUWELS	<92 GB per node on the Cluster; <500GB on the Booster
	Marconi100	<240 GB per node
	MareNostrum 4	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-NG	2.0 GB* #cores or 96 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 – 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 Marconi100) 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> Typical use: Scratch files during simulation, log files, checkpoints Lifetime: Duration of jobs and between jobs	HAWK	n.a.	
	Joliot-Curie	100 TB	Without backup, automatic clean-up procedure
	JUWELS	90 TB	Without backup, clean-up procedure for files older than 90 days
	Marconi100	20 TB (100 TB)	Without backup, clean-up procedure for files older than 50 days
	MareNostrum 4	100 TB (more on demand)	Without backup, clean-up procedure

	Piz Daint	8.8 PB	Without backup, clean-up procedure, Quota in nodes (max 1 million)
	SuperMUC-NG	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	HAWK	100 TB(*1)	Without backups
	Joliot-Curie	5 TB	Without backup
	JUWELS	16 TB	With backup
	Marconi100	20TB (100 TB)(*2)	Without backup
	MareNostrum 4	10 TB (100 TB)	Without backup
	Piz Daint	250 TB (500 TB)(*3)	Read-only from compute nodes data kept only for duration of project
	SuperMUC-NG	100 TB (200 TB)	Without backups
Total storage (Home) <number> Typical use: Source code and scripts Lifetime: Duration of project	HAWK	10 GB / user	With snapshots
	Joliot-Curie	20 GB	With backup and snapshots
	JUWELS	10 GB	With backup
	Marconi100	50 GB	With backup
	MareNostrum 4	20 GB	With backup
	Piz Daint	50 GB / user	With backup and snapshots
	SuperMUC-NG	100 GB	With backup and snapshots
Total storage (Archive) <number>	HAWK	Upon agreement	
	Joliot-Curie	100 TB	File size > 1 GB
	JUWELS	(*4)	Ideal file size: 500 GB – 1000 GB
	Marconi 100	20 TB (100 TB)(*5)	
	MareNostrum 4	n.a.	
	Piz Daint	n.a.	
	SuperMUC-NG	100 TB(*6)	Typical file size should be > 5 GB

(*1) 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.
(*2) The default value is 1 TB. Please ask to CINECA User Support (superc@cineca.it) to increase your quota after the project will start.
(*3) 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.
(*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.
(*6) Long-term archiving or larger capacity has to be requested separately from 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 centre beforehand for approval.

Field in online form	*Machine*	*Max*	*Remarks*
Number of files (Scratch) <number>	HAWK	n.a.	
	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
	Marconi100	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
	SuperMUC-NG	1 million	Without backup, old files are removed automatically, Ideal file size: >100 GB
Number of files (Work) <number>	HAWK	100 000	
	Joliot-Curie	500 000	Extensible on demand, 10 000 files max per directory
	JUWELS	3 million	With backup
	Marconi100	2 million	Without backup
	MareNostrum 4	2 million	
	Piz Daint	50 000 per TB	With backup and snapshots
	SuperMUC-NG	1 million	Ideal file size: >100 GB
Number of files (Home) <number>	HAWK	100 000	
	Joliot-Curie	n.a.	
	JUWELS	40 000	With backup
	Marconi100	100 000	With backup
	MareNostrum 4	100 000	
	Piz Daint	500 000	With backup and snapshots
	SuperMUC-NG	100 000	With backup (snapshots)
Number of files (Archive) <number>	HAWK	Upon request	Typical file size should be > 5 GB
	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
	Marconi100	10 000(*)	Without backup
	MareNostrum 4	n.a.	
	Piz Daint	n.a.	
	SuperMUC-NG	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>	HAWK	Max. 5 Gb/s
	Joliot-Curie	100 TB
	JUWELS	100 TB
	Marconi100	20 TB(*)
	MareNostrum 4	50 TB
	Piz Daint	Currently no limit
	SuperMUC-NG	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).