CHRONOS Tier-0 Project

Technical Guidelines – Call for Proposals

The proper benchmarks and resource request justification should be provided on Piz Daint[1)]. The parameters are listed in the tables.

The systems available will be Piz Daint at CSCS and LUMI-G system at CSC (the Swiss Share - not yet available for benchmarking).

[1)] 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 CSCS system available for this Call

		Piz Daint
System Type		Hybrid Cray xC50
Compute	Processor type	Intel® Xeon® E5-2690 v3 @ 2.60GHz (12 cores)
	Total nb of nodes	5 704
	Total nb of cores	68 448
	Nb of accelerators /node	1 GPU per node
	Type of accelerator	NVIDIA® Tesla® P100 16GB
Memory	Memory / Node	64 GB
Network	Network Type	Cray Aries
	Connectivity	Dragonfly

		Piz Daint
Home file system	type	GPFS
	capacity	160 TB
Work file system	type	GPFS
	capacity	6.3 PB
Scratch file system	type	Lustre
	capacity	8.8 PB
Archive	capacity	n.a.
Minimum required job size	Nb of cores	6 nodes

More details on the website of the centre:

Piz Daint:

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

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

Subsection for the CSCS system(s):

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 6 (section B) to provide correctly the required technical data for Piz Daint.

B – Guidelines

Resource Usage

Computing time

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

Please justify the number of node hours you request by providing a detailed work plan and the appropriate technical data on the systems of interest. Please applied for a preparatory project if needed.

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.

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>	Piz Daint	-
Able to write checkpoints <check button>	Piz Daint	Yes (apps checkpoints)
Maximum time between two checkpoints (= maximum wall clock time for a job) (hours) <number>	Piz Daint	24 hours

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>	Piz Daint	No shared nodes: 1 job per node maximum

Job Size

The next fields describe the job resource requirements, which are the number of 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 nodes in order to make efficient use of the architecture.

IMPORTANT REMARK

On Piz Daint, technical data needs to be provided on the Cray XC50 (see additional information on page 6). Missing technical data (scaling, etc.) may result in rejection of the proposal.

Field in online form	*Machine*	*Min (cores)*
Expected job configuration (Minimum) <number>	Piz Daint	6 nodes
Expected number of cores (Average) <number>	Piz Daint	6 to 2 400 nodes
Expected number of cores (Maximum) <number>	Piz Daint	4 400 nodes (prior agreement with CSCS staff is required to run jobs over 2400 nodes)

Additional information:

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 (1 node hour = 68 core hours).

Field in online form	*Machine*	*Max*
Memory (Minimum job) <number>	Piz Daint	5.3 GB per core or 64 GB per node (nodes are not shared)
Memory (Average job) <number>	Piz Daint	5.3 GB per core or 64 GB per node (nodes are not shared)
Memory (Maximum job) <number>	Piz Daint	5.3 GB per core or 64 GB per node (nodes are not shared)

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

IMPORTANT REMARK

All data must be removed from the execution system within 3 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. **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**	Piz Daint	8.8 PB	Without backup, clean-up procedure, Quota in nodes (max 1 million)

Total storage (Work) <number> Typical use: Result and large input files Lifetime: Duration of project	Piz Daint	250 TB (500 TB)(*1)	Read-only from compute nodes data kept only for duration of project
Total storage (Home) <number> Typical use: Source code and scripts Lifetime: Duration of project	Piz Daint	50 GB / user	With backup and snapshots
Total storage (Archive) <number>	Piz Daint	n.a.	

(*1) 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.

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>	Piz Daint	1 million	No limit while running, but job submission is blocked if the max number of files left on scratch is reached
Number of files (Work) <number>	Piz Daint	50 000 per TB	With backup and snapshots
Number of files (Home) <number>	Piz Daint	500 000	With backup and snapshots
Number of files (Archive) <number>	Piz Daint	n.a.	

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.

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>	Piz Daint	Currently no limit

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).