CHRONOS Tier-0 Project

Technical Guidelines – Call for Proposals

The proper benchmarks and resource request justification should be provided on LUMI-G or Alps. The parameters are listed in the tables.

The system available will be LUMI-G system at CSC (the Swiss Share - available for benchmarking) and Alps research infrastructure at CSCS.

Via Trevano 131
6900 Lugano
Switzerland

Tel +41 (0)91 610 82 11
Fax +41 (0)91 610 82 82
www.cscs.ch

A - General Information on the LUMI-G system or Alps available for this Call

		LUMI-G
System Type		HPE Cray EX
Compute	Processor type	AMD Epyc 7A53 (64C, 2.00GHz)
	Total nb of nodes	2 928
	Total nb of cores	187 392
	Nb of accelerators /node	4 AMD MI250X GPUs per node
	Type of accelerator	AMD Instinct MI250X
Memory	Memory / Node	512 GB + 4x128GB HBM
Network	Network Type	4x200 Gbit Slingshot per node
	Connectivity	Dragonfly

		LUMI-G
Home file system	type	Lustre
	capacity	20 GB/User
Work file system	type	Lustre
	capacity	80 PB
Scratch file system	type	Lustre
	capacity	9 PB (flash)
Archive	capacity	n.a.
Minimum required job size	Nb of cores	1 GPU

		Alps
System Type		HPE Cray EX
Compute	Processor type	NVIDIA Grace Hopper 64 core ARM cores, H100 GPUs
	Total nb of nodes	1 200
	Total nb of sockets	4 800
	Nb of accelerators /node	4 NVIDIA GH per node
	Type of accelerator	H100 GPU
Memory	Memory / Node	128 LPDDR RAM + 96 GB HBM3 memory
Network	Network Type	HPC Cray Slingshot @ 200 Gbps injection per module / GPU
	Connectivity	Slingshot

		Alps
Scratch file system	type	Lustre
	capacity	100 + 10 PB on hard disk 5 + 1 PB on Solid State Disk (SSD)
Archive	capacity	n.a.
Minimum required job size	Nb of nodes	1 node (4 GPU)

More details on the website of the centre:

LUMI-G:

https://www.lumi-supercomputer.eu/

Alps

https://www.cscs.ch/computers/alps

Subsection for the CSCS system(s):

Swiss Share on LUMI-G

The LUMI-G partition provides the majority of the compute performance of LUMI. It consists of 2928 compute nodes. Each compute node has a single AMD 64 core CPU at 2.0 GHz and 512 GB of memory, the cores have support for 2-way simultaneous multithreading (SMT), however a number of cores are reserved for the operating system leaving 56 cores usable. Additionally, each node has 4 MI250X GPUs, Each GPU has a total of 128 GB of HBM2e memory, and is presented to the user as two logical devices. Each node also has 4 network adapters each providing 200 Gbit/s of connectivity.

CHRONOS on Alps Research Infrastructure

Alps is a general-purpose compute and data Research Infrastructure (RI) open to the broad community of researchers in Switzerland and the rest of the world. Alps will provide a high impact, challenging and innovative RI that will allow Switzerland to advance science and impact society.

Alps enables the creation of versatile clusters (vClusters) that can be tailored to the specific needs of users while maintaining confidentiality.

B – Guidelines

Resource Usage

Computing time

The amount of computing time has to be specified in GPU hours. It is the total number of GPU hours to be consumed within the twelve months period of the project.

Please justify the number of GPU 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>	LUMI-G	-
Able to write checkpoints <check button>	LUMI-G	Yes
Maximum time between two checkpoints (= maximum wall clock time for a job) (hours) <number>	LUMI-G	48 hours
Wall clock time of one typical simulation (hours) <number>	Alps	-
Able to write checkpoints <check button>	Alps	Yes
Maximum time between two checkpoints (= maximum wall clock time for a job) (hours) <number>	Alps	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>	LUMI-G, Alps	varies

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

Technical data needs to be provided on LUMI-G. Missing technical data (scaling, etc.) may result in rejection of the proposal.

Field in online form	*Machine*	*Min (cores)*
Expected job configuration (Minimum) <number>	LUMI-G	4 Accelerators
Expected number of cores (Average) <number>	LUMI-G	Multiple nodes
Expected number of cores (Maximum) <number>	LUMI-G	LUMI-G 1024 nodes (more by special arrangement
Expected job configuration (Minimum) <number>	Alps	4 Accelerators
Expected number of cores (Average) <number>	Alps	Multiple nodes
Expected number of cores (Maximum) <number>	Alps	Alps 1200 nodes

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	LUMI-G, Alps	500 TB	90 days retention
Total storage (Work) <number> Typical use: Result and large input files Lifetime: Duration of project	LUMI-G, Alps	20 GB/user	
Total storage (Home) <number> Typical use: Source code and scripts Lifetime: Duration of project	LUMI-G, Alps	N/A	
Total storage (Fast Scratch) <number>	LUMI-G, Alps	100 TB	30 days data retention

(*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>	LUMI-G	2000 k	

Number of files (Work) <number>	LUMI-G		
Number of files (Home) <number>	LUMI-G	100 k	
Number of files (Fast scratch) <number>	LUMI-G	1000 k	

For Alps these details will be defined later.

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>	LUMI-G, Alps	No enforced 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).