AI and Data Intensive Applications Access call – Technical Guidelines

Peer-Review Office – Version 1 – 29/08/2024

Table of Contents

The contributing sites and the corresponding computer systems available are:

System	Architecture	Site (Country)	Fixed Allocation (node hours)
Vega GPU	BullSequana XH2001	IZUM Maribor (SI)	7,100
Karolina GPU	HPE Apollo 2000Gen10 Plus and HPE Apollo 6500	VSB-TUO, IT4Innovations, (CZ)	7,500
MeluXina GPU	BullSequana XH2005	LuxProvide (LU)	25,000
LUMI-G	HPE Cray EX	CSC (FI)	35,000
Leonardo Booster	BullSequana XH21355 "Da Vinci" blade	CINECA (IT)	50,000
MareNostrum5 ACC	Atos BullSequana EX3000	BSC (ES)	32,000

The site selection and the requested computing time are specified in the online application form. The applicant may choose only one system per application.

A. General information on the systems available for the AI and Data Intensive Applications call

	Compute						Memory	Network	
System	**System Type**	**Processor type**	**Total nb of nodes**	**Total nb of cores**	**Nb of acceler-ators /node**	**Type of accelera-tor**	**Memory / Node**	**Network Type**	**Connectivity**
Vega GPU	Bull Sequana	AMD Epyc 7H12 (64C, 2.6GHz)	60	7.68	4	NVIDIA A100	512 GB	Infiniband HDR100 Dragonfly+	5x100GbE
MeluXina GPU	Bull Sequana	AMD Epyc 7452 (32C, 2.35GHz)	200	12.8	44	Nvidia A100 40 GB HBM2	512 GB	Infiniband HDR 200 Dual-railInfiniband HDR 200	Dragonfly+
Karolina GPU	HPE	AMD Epyc 7763 (64C, 2.45GHz)	72	9.216	8	Nvidia A100-SXM4-40GB	1024 GB DDR4 + 8x40GB HBM	Infiniband HDR	4xHDR 800Gb/s
LUMI-G	HPE	AMD Epyc 7A53 (64C, 2.00GHz)	2928	187.392	4	AMD Instinct MI250X	512 GB + 4x128GB HBM	4x200Gbit Slingshot per node	Dragonfly
Leonardo Booster	Bull Sequana	Intel Xeon 8358 (32C, 2.60 GHz)	3456	110.592	4	NVIDIA®custom Ampere® A100, NVidia 3.0, 64GB	512 (8x64) GB DDR4 3200 MHz	NVIDIA Mellanox HDR DragonFly++ 200Gb/s	DragonFly++ 200Gb/s
MareNos-trum5 ACC	Bull Sequana	Intel Sapphire Rapids 8460Y+ (32C, 2.3 Ghz)	1120	71.680	4	Nvidia Hopper (64GB HBM)	512 GB	4x Infiniband NDR200 per node	Fat tree (with is-lands of 160 nodes full fat tree

	Home file system		Work file system		Scratch file system		Backup	Archive	Minimum re-quired job size
System	**Type**	**Capacity**	**Type**	**Capacity**	**Type**	**Capacity**	**Capacity**	**Capacity**	**Nb of cores**
Vega GPU	Ceph	100 GB	Ceph	On-demand	Lustre	20 GB	-	0	1 GPU
MeluXina GPU	Lustre	4.2 PB on the shared Tier2 Home-Project filesystem used as Work filesys-tem	Lustre	4.2 PB	Lustre	229 TB	2.2 PB	1.5 PB	64
Karolina GPU	NFS	31 TB	NFS	15 PB	Lustre	1361 TB	-	1500 TB	16
LUMI-G	Lustre	20 GB/User	Lustre	80 PB	Lustre	9 PB (flash)	N/A	N/A	1 GPU
Leonardo Booster	Lustre	50 GB/User	Lustre	10 PB	Lustre	41.4 PiB	-	40 PB	32
MareNos-trum5 ACC	IBM Storage Scale (GPFS)	272 TB (40 GB/user)	IBM Storage Scale (GPFS)	20 PB	IBM Storage Scale (GPFS)	160 PB	100 PB	40PB disk + 400PB tapes	64 (1 node)

IMPORTANT REMARKS

Applicants are advised to apply to EuroHPC Benchmark or Development calls to collect relevant benchmarks and technical data for the system they wish to use through this Access call.

More details on the website of the centres:

Vega:

https://doc.vega.izum.si/

MeluXina:

https://docs.lxp.lu

Karolina:

https://www.it4i.cz/en/infrastructure/karolina

https://docs.it4i.cz/karolina/introduction/

LUMI:

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

Leonardo:

https://leonardo-supercomputer.cineca.eu/

https://wiki.u-gov.it/confluence/display/SCAIUS/UG3.4%3A+Leonardo+UserGuide

MareNostrum5:

https://www.bsc.es/innovation-and-services/marenostrum/marenostrum-5

A.1 EuroHPC JU systems specific information

Vega, IZUM (SI)

HPC Vega is an Atos BullSequana XH2000 system able to deliver more than 6.9 PFLOPS of aggregated sustained performance. It comprises computer partitions with different computing characteristics, and two high-performance storage systems, one based on Lustre and another on Ceph. It consists of 1020 compute nodes with at least 256 GB of RAM, all together 130560 CPU cores. Sustained performance on all CPUs is 3.8 PFLOPS. 240 GPU accelerators with all together 829440 FP64 CUDA cores and 103680 Tensor cores perform up to 3.1 PFLOPS.

The **CPU partition** consists of 10 BullSequana XH2000 DLC racks, with:

- 768 standard compute nodes (within 256 blades), each node with:
  - 2 CPUs AMD EPYC Rome 7H12 (64c, 2.6GHz, 280W), 256GB of RAM DDR4-3200, 1x HDR100 single port mezzanine, 1x local 1.92TB M.2 SSD
- 192 large memory compute nodes (within 64 blades), each node with:
  - 2 CPUs AMD EPYC Rome (64c, 2.6GHz, 280W), 1TB of RAM DDR4-3200, 1xHDR100 single port mezzanine 1x 1.92TB M.2 SSD

The **GPU partition** consists of 2 BullSequana XH2000 DLC racks, with:

- 60 GPU nodes (60 blades), each node with:
  - 2 CPUs AMD EPYC Rome (64c, 2.6GHz, 280W), 512 GB of RAM DDR4-3200, local 1.92 TB M.2 SSD
  - 4x NVIDIA Ampere A100 PCIe GPU (3456 FP64 CUDA cores, 432 Tensor cores, Peak FP64 9.7 TFLOPS, FP64 Tensor Core 19.5 TFLOPS), each with 40 GB HBM2

MeluXina, LuxProvide (LU)

MeluXina is an Atos BullSequana XH2000 system able to deliver more than 12.8 Petaflops of aggregated sustained performance. It comprises computer partitions with different computing characteristics, three high-performance storage systems based on Lustre, and a Tape archival system.

Hardware specifications – Compute environment:

- **Cluster Module (CPU)**: 573 CPU nodes, each with:
  - CPU: 2x AMD EPYC Rome 7H12 (64cores @ 2.6 GHz, 128 physical cores total)
  - RAM: 512 GB DDR4-3200
  - Interconnect: 1x HDR (200 Gbps InfiniBand)
  - No local storage

- **Accelerator Module (GPU):** 200 CPU-GPU hybrid nodes, each with:
  - CPU: 2x AMD EPYC Rome 7452 (2x 32cores @ 2.35 GHz, 64 physical cores total)
  - RAM: 512 GB DDR4-3200
  - Accelerator: 4x Nvidia Ampere A100-40 (40GB HBM, NVlink)
  - Interconnect: 2x HDR (200 Gbps InfiniBand, 400Gbps in dual-rail)
  - Local storage: 1.92TB SSD

- **Accelerator Module (FPGA):** 20 CPU-FPGA hybrid nodes, each with:
  - CPU: 2x AMD EPYC Rome 7452 (2x 32cores @ 2.35 GHz, 64 physical cores total)
  - RAM: 512 GB DDR4-3200
  - Accelerator: 2x BittWare 520N-MX (Intel Stratix 10MX,16GB HBM)
  - Interconnect: 2x HDR (200 Gbps InfiniBand, 400Gbps in dual-rail)
  - Local storage: 1.92TB SSD

- **Large Memory Module (CPU):** 20 CPU nodes with extended memory capacity, each with:
  - CPU: 2x AMD EPYC Rome 7H12 (64cores @ 2.6 GHz, 128 physical cores total)
  - RAM: 4 TB DDR4-3200
  - Interconnect: 2x HDR (200 Gbps InfiniBand, 400Gbps in dual-rail)
  - Local storage: 1.92TB NVMe

Karolina, IT4Innovations (CZ)

Karolina is an HPE Apollo system able to deliver more than 9.5 Petaflops of aggregated LINPACK performance. It comprises computer partitions with different computing characteristics, and high-performance storage system based on Lustre.

- **Universal partition (CPU)** consists of 720 nodes. Every node features 2x AMD EPYC 7H12 processors, 128 cores and 256GB of memory per node. The nodes are connected to the Infiniband HDR network at 100Gb/s rate. Via the network, nodes can access the SCRATCH, HOME and PROJECT storage. The partition provides 2.84PF of double precision performance;

- **The Accelerated partition (GPU)** consists of 72 nodes. Every node features 2x AMD EPYC 7763 processors, 128 cores and 1024GB of memory per node. Every node contains 8 Nvidia A100 GPUs with 40GB of HBM2 memory, attached via Gen4 PCIe bus. The 8 accelerators are interconnected by an NVLINK2 fabric featuring NVSwitch technology. This enables 320GB of HBM2 memory addressable across the accelerators in unified virtual address space. The nodes are connected to the Infiniband HDR network with 4x200Gb/s links to achieve very high throughput to the network and the SCRATCH storage. The partition provides in total 576 A100 GPUs and 6.75PF of LINPACK performance.

- **High performance SCRATCH storage** The all flash SCRATCH storage provides 1361 TB capacity, 730GB/s write performance and 1198BG/s read performance and over 5M IOPS performance It is accessible via the Infiniband network and is available from all login and computational nodes. The SCRATCH is based on Lustre parallel filesystem and is intended for temporary scratch data generated during the calculation as well as for high-performance access to input and output files. Extended ACLs are provided for sharing data with other users using fine-grained control.

Discoverer, Sofia Tech (BG)

Discoverer is an Atos BullSequana XH2000 system able to deliver more than 4.5 Petaflops of aggregated sustained performance. It comprises computer one standard memory CPU partition, one large memory CPU partition and one high performance storage based on Lustre.

Compute node design:

- **CPU model**: AMD EPYC 7H12, 64core, 2.6GHz, 280W; Next generation x86 “Zen2”
- CPU sockets per node: 2;
- CPU Cores per node: 128;
- Main memory per node: 256GB (Each of the 18x Fat nodes has 1024GB Memory)
- Memory type and frequency: 16GB DDR4 RDIMM 3200MT/s DR; (The fat nodes are equipped with 64GB DDR4 RDIMM 3200MT/s DR)
- Node DP TeraFlop/s peak: 5.325TFlops
- % DP TeraFlop/s peak vs Linpack: 74%;
- TFlop/s sustained Linpack: 3.940TFlops;
- Linpack node power consumption: 665.1 W per 256 GB compute node; 747.0 W per Fat compute node (Cooling subsystem power consumption excluded);
- Number and bandwidth of network interfaces: 1x 200Gbps HDR;

LUMI, CSC (FI)

LUMI is one of the three European pre-exascale supercomputers. It's an HPE Cray EX supercomputer consisting of several partitions targeted for different use cases. The largest partition of the system is the "LUMI-G" partition consisting of GPU accelerated nodes using a future-generation AMD Instinct GPUs. In addition to this, there is a smaller CPU-only partition, "LUMI-C" that features AMD EPYC "Milan" CPUs and an auxiliary partition for data analytics with large memory nodes and some GPUs for data visualization. Besides partitions dedicated to computation, LUMI also offer several storage partitions for a total of 119 PB of storage space.

- **LUMI-C: The CPU Partition**

The LUMI-C partition consists of 2048 compute nodes Each LUMI-C compute nodes are equipped with 2 AMD EPYC 7763 CPUs with 64 cores each running at 2.45 GHz for a total of 128 cores per node. The cores have support for 2-way simultaneous multithreading (SMT) allowing for up to 256 threads per node. The normal compute nodes in LUMI-C have 256 GB of memory, but there are also 128 nodes with 512 GB and 32 nodes with 1024 GB. Each compute node has one 200 Gbit/s network adapter.

- **LUMI-G: The GPU Partition**

The LUMI-C 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.

- **LUMI-D: The Data Analytics Partition**

LUMI-D is intended for interactive data analytics and visualization. It is also a good place run pre- and post-processing jobs that require a lot of memory. It consists of a 8 nodes with large memory capacity (4 nodes with 4 TB per node and 4 nodes with 8 TB per node) and 8 nodes with NVIDIA A40 GPUs. Each LUMI-D compute nodes are equipped with 2 AMD EPYC 7742 CPUs with 64 cores each running at 2.25 GHz for a total of 128 cores per node.

- **LUMI-P and F: Parallel Filesystems**

LUMI has two Lustre parallel file systems consisting of:

A main storage partition (LUMI-P) composed of 4 independent Lustre file systems with an aggregated performance of 240 GB/s and a 20 PB storage capacity each. Projects get assigned to one of these at project creation.

A flash storage partition (LUMI-F) optimized to support high IOPS rates with an aggregated performance in excess of 2000 GB/s and 9 PB of storage capacity

- **LUMI-O: The Object Storage**

Object storage is a data storage architecture that manages data as objects instead of a file hierarchy. Each object includes the data, the metadata and a globally unique identifier. This partition may be used for storing, sharing and staging your data. It's based on Ceph and has a storage capacity of 30 PB.

Leonardo, CINECA (IT)

Leonardo is the new pre-exascale Tier-0 EuroHPC supercomputer hosted by CINECA and currently built in the Bologna Technopole, Italy. It is supplied by ATOS, based on a BullSequana supercomputer nodes. The used network is a Mellanox Infiniband HDR with DragonFly+ topology.

Leonardo will provide to users **two main computing modules**:

- **Leonardo Booster**
  The Leonardo Booster module, supplied by ATOS, is based on a BullSequana XH2135 “Da Vinci” blade architecture. It was designed to satisfy the most computational-demanding requirements in terms of *time-to-solution*, while optimizing the *energy-to-solution*. The system consists of about 3456 computing nodes (+16 login) each equipped with 4 NVIDIA custom Ampere A 100 GPU, Nvlink 3.0, 64GB, 512 GB of DDR4 RAM driven by a single Intel Xeon 8358 CPU at 2.6 GHz (32 cores per node), This partition provides a peak performance over 238 Pflops.

- **Leonardo Data Centric General Purpose (Leonardo DCGP)**
  The Leonardo Data Centric module is based on the BullSequana X2140 compute blade architecture by Atos. The partition aim is to satisfy a broader range of applications. It offers 1536 compute nodes (the login nodes are shared with the Booster module) each equipped with 2 Intel Sapphire Rapids SPR03-LC (56 cores per CPU, 112 per node), 512 GB of DDR5 RAM and 8 TB of NVMe.

Marenostrum5, BSC (ES)

MareNostrum5 is a pre-exascale EuroHPC supercomputer located in the BSC-CNS. The system is supplied by Bull SAS combining Bull Sequana XH3000 and Lenovo ThinkSystem architectures. The machine will combine 2 main partitions, one dedicated to general purpose applications MareNostrum 5 GPP and another one based in accelerators MareNotrum5 ACC.

- **MareNostrum5 GPP**

MareNostrum5 GPP is a General-purpose partition using Intel Sapphire Rapids CPUs (2xIntel Shappire Rapids 8480+ 2Ghz 56C, per node), with a peak performance of more than 45 PFlops, it is one of the biggest machines in the world for general-purpose workflows. The machine contains 6408 nodes with 112 cores each one. The ratio of memory is 2GB/core except for 216 nodes high mem that will provide up to 1024 GB (8GB /core).

The connectivity for parallel jobs and for the storage is based on Infiniband NDR200, in this case sharing the 200GB/s by 2 nodes (providing 100Gb/s per node)

- **MareNostrum5 ACC**

Marenostrum5 ACC is an accelerated partition using Intel Shappire Rapids CPU and Nvidia Hopper GPU (2xIntel Shappire Rapids 8460Y+ 2.3Ghz 32C, per node and 4 GPUs Nvidia Hopper 64GB HBM). The machine has a peak performance of more than 224PFlops and have 512 GB or main memory per node.

The connectivity for parallel jobs and for the storage is based on Infiniband NDR200, the machine will provide 4xNDR200 per node, which sum up to 800GB/s per node.

- **MareNostrum5 Storage**

The Marenostrum5 machine will have a Spectrum Scale File system with up to 248PB distributed in different mount points. In addition, the system will have available an archive filesystem based in Spectrum Scale Archive with up to 40 PB of cache disk + 400PB using Tapes.

Deucalion, FCT (PT)

Deucalion is a petascale EuroHPC supercomputer located in Guimarães, Portugal. It is supplied by Fujitsu Technology Solutions, which combines a Fujitsu PRIMEHPC (ARM partition) and Atos Bull Sequana (x86 partitions). Deucalion is able to deliver more than 7.22 Petaflops of aggregated sustained performance and has a hybrid architecture with 2 computational clusters plus accelerated nodes with GPU.

Hardware specifications – Compute environment:

- **ARM CPU:** 1632 nodes, each with:
  - CPU: 1x Arm A64FX (2.0GHz, 48 Cores)
  - Memory: 32GiB (HBM2: 8GiB x4)
  - Interconnect: 1x HDR (100 Gbps InfiniBand)
  - Local storage: 1 x M.2 SSD 512GB NVMe

- **X86 CPU:** Bull Sequana X440 A5, 500 nodes, each with:
  - CPU: 2x AMD EPYC 7742 (2.25GHz, 64 Cores)
  - RAM: 256 GB DDR4
  - Interconnect: 1x HDR (100 Gbps InfiniBand)
  - Local storage: 1x 480GB SSD

- **X86 GPU:** Bull Sequana X410 A5, 33 nodes, each with:
  - CPU: 2x AMD EPYC 7742 (2.25GHz, 64 Cores)
  - RAM: 512GB DDR4
  - Accelerator: 4x Nvidia Ampere A100 NVlink (17 nodes with 40GB and 16 nodes with 80GB)
  - Interconnect: 2x HDR (200 Gbps InfiniBand)
  - Local storage: 1 x 480GB SSD
- **Storage:**
  - NetApp AFF A220 NAS subsystem with SSD 50 TB usable

- DDN EXAScaler Lustre PFS with HotPool NVME tier with 430 TB usable and 10 PB usable HDD Datapools with Aggregated Performance 340GB/s in reads, 260GB/s in writes

B. Guidelines for filling in the online form

B.1 Resource Usage

Computing time

The amount of computing time has to be specified in node hours (wall clock time [hours]*physical cores (nodes) of the machine applied for). It is the total number of node hours to be consumed within the 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. Applicants are strongly invited to apply to EuroHPC Benchmark and Development calls.

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 system availability depends on the system and it should be about 80% of the total availability, due to queue times, possible system maintenance, upgrade and data transfer time.

Proposals are required to respect the flat allocation request of resources as indicated on the EuroHPC website:

- AI and Data-intensive Applications Access call

B.2 Job Characteristics

This section describes the 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.**

<table>
<tr><th>Field in the online form</th><th>System</th><th colspan="3">Maximum</th></tr>
<tr><td rowspan="12">Wall clock time of one typical simulation (hours) <number></td><td rowspan="5">Vega</td><td>Partition</td><td>Nodes</td><td>Wall time</td></tr>
<tr><td>CPU</td><td>960</td><td>2 days</td></tr>
<tr><td>LARGEMEM</td><td>192</td><td>2 days</td></tr>
<tr><td>LONGCPU</td><td>6</td><td>4 days</td></tr>
<tr><td>GPU</td><td>60</td><td>4 days</td></tr>
<tr><td>MeluXina</td><td colspan="3">48 hours</td></tr>
<tr><td>Karolina</td><td colspan="3">48 hours</td></tr>
<tr><td>Discoverer</td><td colspan="3">48 hours</td></tr>
<tr><td>LUMI</td><td colspan="3">-</td></tr>
<tr><td>Leonardo</td><td colspan="3">24 hours</td></tr>
<tr><td>MareNostrum5</td><td colspan="3">72 hours</td></tr>
<tr><td>Deucalion</td><td colspan="3">48 hours</td></tr>
<tr><td rowspan="8">Able to write checkpoints <check button></td><td>Vega</td><td colspan="3">No</td></tr>
<tr><td>MeluXina</td><td colspan="3">Applications need to support an internal checkpoint-restart mechanism, and the data needs to fit within the project’s allocated storage</td></tr>
<tr><td>Karolina</td><td colspan="3">25TB (may be increased upon request, see below)</td></tr>
<tr><td>Discoverer</td><td colspan="3">Yes</td></tr>
<tr><td>LUMI</td><td colspan="3">Yes</td></tr>
<tr><td>Leonardo</td><td colspan="3">Yes</td></tr>
<tr><td>MareNostrum5</td><td colspan="3">Yes</td></tr>
<tr><td>Deucalion</td><td colspan="3">Applications need to support an internal checkpoint-restart mechanism, and the data needs to fit within the project’s allocated storage</td></tr>
<tr><td rowspan="8">Maximum time between two checkpoints (= maximum wall clock time for a job) (hours) <number></td><td>Vega</td><td colspan="3">N/A</td></tr>
<tr><td>MeluXina</td><td colspan="3">48 hours</td></tr>
<tr><td>Karolina</td><td colspan="3">48 hours</td></tr>
<tr><td>Discoverer</td><td colspan="3">48 hours</td></tr>
<tr><td>LUMI</td><td colspan="3">48 hours</td></tr>
<tr><td>Leonardo</td><td colspan="3">24 hours</td></tr>
<tr><td>MareNostrum5</td><td colspan="3">Recommended 24 hours</td></tr>
<tr><td>Deucalion</td><td colspan="3">N/A</td></tr>
</table>

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 the online form	System	Maximum
Number of jobs that can run simultaneously <number>	Vega	50
	MeluXina	Generally: maximum 100 jobs per user Some job types allow a maximum of 1 job per user for high priority testing, interactive development and non-scalable executions
	Karolina	Max. queued: no limit Max. running: 1296
	Discoverer	512
	LUMI	Varies
	Leonardo	Varies
	MareNostrum5	Depending on the system load
	Deucalion	Depending on the system load

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 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 REMARKS

If possible, 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. Generic scaling plots provided by vendors or developers do not necessarily reflect the actual code behaviour for the simulations planned. Scaling benchmarks should be representative of your study case and need

to support your resource request on every system of interest. Application to EuroHPC Benchmark and Development calls is recommended.

Field in the online form	System	Minimum (cores)
Expected job configuration (Minimum) <number>	Vega	1
	MeluXina	CPU and Large Memory nodes: 128 cores GPU and FPGA nodes: 64 cores + all accelerators on the same node
	Karolina	Karolina CPU : 128 cores Karolina GPU :16 cores + 1xA100
	Discoverer	64 cores
	LUMI	LUMI-C 128 cores (1 core on special partitions) LUMI-G 4 Accelerators (1 on special partitions)
	Leonardo	Leonardo Booster: 32 cores (1 nodes) Leonardo DCGP: 224 cores (2 nodes)
	MareNostrum5	1 full node per partition
	Deucalion	1 full node per partition
Expected number of cores (Average) <number>	Vega	512
	MeluXina	CPU nodes: 1024 cores (8 nodes) LargeMemory nodes: 128 cores (1 node) GPU nodes: 256 cores (4 nodes, 16 GPU accelerators) FPGA nodes: 64 cores (1 node, 2 FPGA accelerators)
	Karolina	Karolina CPU : 1024 cores Karolina GPU :128 cores + 8xA100
	Discoverer	512 cores
	LUMI	LUMI-C/G Multiple nodes
	Leonardo	Leonardo Booster: > 640 cores (20 nodes, using 1 or more jobs at the same time) Leonardo DCGP: > 64 nodes
	MareNostrum5	On demand
	Deucalion	Dependent on queue
Expected number of cores (Maximum) <number>	Vega	65536
	MeluXina	Maximums per job: For the default QOS, maximum 25% of each partition size. CPU nodes: 17920 cores (140 nodes) LargeMemory nodes: 640 cores (5 nodes) GPU nodes: 3200 cores (50 nodes) FPGA nodes: 320 cores (5 nodes)
	Karolina	Karolina CPU : 92160 cores Karolina GPU :9216 cores + 576xA100
	Discoverer	3072 cores
	LUMI	LUMI-C 512 nodes -> 65536 cores (more by special arrangement) LUMI-G 1024 nodes (more by special arrangement)
	Leonardo	Leonardo Booster: 8192 cores (256 nodes, 1024 GPUs, more could be possible, but only if approved by the hosting site and only for few jobs per projects) Leonardo DCGP: TBD (order of magnitude: 10000 cores, 90 nodes)
	MareNostrum5	On demand
	Deucalion	Dependent on queue

Additional information

Vega

Slurm partitions information: https://doc.vega.izum.si/slurm-partitions/

MeluXina

Job scheduling is done with node-level granularity, by default users have exclusive access to allocated nodes and all of their resources.
Compute nodes have hyperthreading enabled and HT cores are available by default to user jobs.
Time limits, job sizes and priorities are set through SLURM Quality of Service (QOS) configurations.
Specific SLURM reservations are available for rapid prototyping and interactive development, accessible through dedicated QOS.

Karolina

DATA Analytics: The Data analytics partition consists of a single HPE Superdome Flex node. The SMP node features 32 high end Intel Xeon 8268 processors with 24 cores each, amounting to 768 AVX-512 capable cores. The processors are equipped with 768GB of DDR4 memory and interconnected in an all-to-all topology by high speed internal network, thus providing over 24000 GB of shared memory. Further, the node is connected to the network and storage by 2x HDR network interfaces, with aggregated throughput of 400Gb/s. The Data analytics partition is intended to support huge memory jobs.

VISUALIZATON Nodes: Karolina includes two nodes for remote visualization via VirtualGL 2 and TurboVNC 2. Every node features 2x AMD EPYC 7452 processors, 64 cores and 256GB of memory and NVIDIA Quadro RTX6000 graphics card with OpenGL support. The nodes are connected to the Infiniband HDR network at 2x100Gb/s rate.

HOME Storage: The HOME storage is a small, 25TB storage to keep user home directories and configuration files. It is NFS based and accessible from all Karolina nodes.

PROJECT Storage: The PROJECT storage is an external, high capacity file storage available to the Karolina supercomputer. The storage is attached to the Karolina via dedicated gateways, providing up to 15PB capacity at an aggregated performance of 39GB/s and 57kIOPS. The PROJECT storage provides space for semi-permanent user data for the duration of user projects.

DICE B2SAFE: Long term safe storage EUDAT B2SAFE service is locally provided and integrated with the Karolina supercomputer. B2SAFE is a way to distribute and store large volumes of data for

a long-term to those sites which are providing powerful data processing, analysis and data access facilities. The service is iRODS based and includes tools to set data management policies across different geographical and administrative domains in a trustworthy manner. Also, it allows to make data objects referenceable via globally unique persistent identifiers.

Leonardo Booster

To apply for Leonardo Booster 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. 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.

Field in the online form	System	Maximum
Memory (Minimum job) <number>	Vega	1 GB per core
	MeluXina	CPU, GPU and FPGA nodes: 512 GB (main memory per node) GPU nodes: 160 GB HBM2 (aggregated GPU accelerator memory per node) LargeMemory nodes: 4 TB (main memory per node)
	Karolina	Karolina CPU: 256 GB Karolina GPU: 1024 GB
	Discoverer	2 GB/core
	LUMI	LUMI-C 256 GB per node (less on special partitions) LUMI-G 512 GB per node (less on special partitions)
	Leonardo	No requirements
	MareNostrum5	Up to 2GB/core
	Deucalion	No requirements
Memory (Average job) <number>	Vega	2 GB per core
	MeluXina	Memory scaled to an expected average per job of: 8 CPU nodes, 4 GPU nodes, 1 LargeMemory and 1 FPGA node
	Karolina	Karolina CPU: 256 GB Karolina GPU: 1024 GB
	Discoverer	128 GB
	LUMI	-
	Leonardo	No requirements
	MareNostrum5	Up to 2GB/core
	Deucalion	No requirements
Memory (Maximum job) <number>	Vega	8 GB per core
	MeluXina	For the default job type, memory scaled to the maximum number of nodes per job depending on node type: 140 CPU nodes, 50 GPU nodes, 5 LargeMemory and 5 FPGA nodes
	Karolina	Karolina CPU: 256 GB Karolina GPU: 1024 GB
	Discoverer	1024 GB
	LUMI	LUMI-C 256 GB per node(4/8/32/64 GB/core on special partitions) LUMI-G 512 GB per node
	Leonardo	482 GB per node
	MareNostrum5	Up to 2GB per core (8GB/core on 216 high-mem nodes)
	Deucalion	CPU ARM: 32 GB per node CPU X86: 256 GB per node GPU: 512 GB per node

B.3 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 for most systems except for MeluXina (scratch storage is a faster data tier with no automatic file removal, projects can use it as their main storage area)).
- **Work** acts as project storage (large results files, no backup)
    - For MeluXina, the project storage data tier can be backed up for projects requesting it. MeluXina has a dedicated Backup data tier, in addition to Archival (off-site, tape-based) storage.
    - For Karolina, the storage can only be used to backup data (simulation results) during project's lifetime.
- **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 (not applicable for Discoverer)).
- **Archive** acts as a long-term storage location, typically data reside on tapes. For EuroHPC 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 on most systems. **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)! Additional data archive specifications of certain systems:

- **Vega** – Archive is not provided and data is not stored on tapes, some additional disc space can be provided on request.
- **MeluXina** – the Archive tier is not user-accessible. Data backup to the archival storage is defined based on the request of project, and is an automated process.
- **Karolina** – Third party archive storage should be used, such as national archive, EOSC or the EUDAT services. Long term safe storage EUDAT B2SAFE service is locally provided and integrated with the Karolina supercomputer.

IMPORTANT REMARKS

All data must be removed from the execution system within a period defined per centre after the end of the project. Data removal period per system is the following:

- **Vega** – 2 months
- **Karolina** – 12 months
- **MeluXina** – 1 month
- **Discoverer** – 2 months
- **LUMI** – 2 months
- **Leonardo** – 6 months
- **MareNostrum5** – 2 months
- **Deucalion** - 2 months

Total Storage

The value asked for is the maximum amount of data needed at a time. Typically, this value varies over the project duration. **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 the online form	System	Maximum per project	Remarks
Total storage (Scratch) <number> Typical use: Scratch files during simulation, log files, checkpoints Lifetime: Duration of jobs and between jobs	Vega	1 TB	
	MeluXina	No maximum capacity is set per project	Projects are allocated storage (max data & max inodes) based on their request and the available capacity relative to the access track.
	Karolina	700 TB	Default safety quota 20TB, increased by hosting entity upon request, short term burst limit up to 90% of free capacity (about 700TB)
	Discoverer	1-50 TB	Scratch and Work partitions are combined on Discoverer
	LUMI	500 TB	90 day retention
	Leonardo	No quota	Without backup, automatic clean-up procedure for files older than 40 days (time interval can be reduced in case of critical usage ratio of the area. In this case, users will be notified via HPC-News)
	MareNostrum5	On demand	90 days retention policy
	Deucalion	N/A	

Field in the online form	System	Maximum per project	Remarks
Total storage (Work) <number> Typical use: Result and large input files Lifetime: Duration of project	Vega	100 TB	
	MeluXina	No maximum capacity is set per project	Projects are allocated storage (max data & max inodes) based on their request and the available capacity relative to the access track
	Karolina	2500 TB	Default quota 20TB, may be increased by beforehand approval by allocation entity and the hosting entity up to 2500TB
	Discoverer	1-50 TB	Scratch and Work partitions are combined on Discoverer
	LUMI		
	Leonardo	1 TB (default, 10-100TB expected)	- Permanent, project specific, local - No backup
	MareNostrum5	On demand	
	Deucalion	10 TB	
Total storage (Home) <number> Typical use: Source code and scripts Lifetime: Duration of project	Vega	10 TB per user	
	MeluXina	100 GB per user	
	Karolina	1000 GB	Default quota 20 GB, may be increased by beforehand approval by the hosting entity up to 1000 GB
	Discoverer	100 GB	
	LUMI	20 GB/user	
	Leonardo	50 GB	Permanent/backed up, user specific, local
	MareNostrum5	40 GB/user	
	Deucalion	10 GB/user	
Total storage (Archive) <number>	Vega	0 TB	
	MeluXina	Evaluated on demand	Projects are allocated archival-level backup storage (max data & max inodes) based on their request and the available capacity relative to the access track.
	Karolina	N/A	Limits set by external Archive service providers
	Discoverer	N/A	No archive partition installed or supported
	LUMI	N/A	
	Leonardo	0.5 PB per project (more is possible, but depending on the available resources)	No backup
	MareNostrum5	On demand	
	Deucalion	N/A	

Field in the online form	System	Maximum per project	Remarks
Total storage (Fast scratch) <number>	LUMI	100 TB	30 Days data retention
Total storage (Object storage) <number>	LUMI	Not decided yet	LUMIs object storage is expected to be available during the fall of 2022

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 (204 million files for Karolina), 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.** This is not applicable to MeluXina.

If you request more than 100 TB (20 TB for Karolina) of disk space, please contact the HPC centre 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 the online form	System	Maximum	Remarks
Number of files (Scratch) <number>	Vega	4 M	
	MeluXina	1 M	Default limit increased on demand and depending on available capacity
	Karolina	20 000 000	
	Discoverer	65 536	Scratch and Work partitions are combined on Discoverer
	LUMI	2000 k	
	Leonardo	No quota	
	MareNostrum5	8 M	
	Deucalion		
Number of files (Work) <number>	Vega	4 M	
	MeluXina	1 M	Default limit increased on demand and depending on available capacity
	Karolina	20 000 000	
	Discoverer	65 536	Scratch and Work partitions are combined on Discoverer
	LUMI		
	Leonardo	No quota	
	MareNostrum5	4 M	
	Deucalion	1M	
Number of files (Home) <number>	Vega	1 M	
	MeluXina	100 000	
	Karolina	500 000	
	Discoverer	86400	
	LUMI	100 k	
	Leonardo	No quota	
	MareNostrum5	100 k	
	Deucalion	100 k	
Number of files (Archive) <number>	Vega	0	
	MeluXina	Evaluated on demand	
	Karolina	N/A	Limits set by external Archive service providers
	Discoverer	N/A	No archive partition installed or supported
	LUMI		
	Leonardo	TBD	
	MareNostrum5	1 M	
	Deucalion		

Field in the online form	System	Maximum	Remarks
Number of files (Fast scratch) <number>	LUMI	1000 k	
Number of files (Object storage) <number>	LUMI	Not decided yet	

B.4 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 EuroHPC 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 the online form	System	Maximum
Amount of data transferred to/from production system <number>	Vega	500 GB per day
	MeluXina	1000 TB/day assuming full utilization of a single 100Gbps link through the GEANT network
	Karolina	500 TB/day
	Discoverer	1 TB/day
	LUMI	No enforced limit
	Leonardo	0.5 TB/day per project (more it is possible, e.g. up to 2 TB/day, but a detailed plan for moving the data needs to be reported)
	MareNostrum5	No enforced limit, but can be limited if the usage affects other users
	Deucalion	No enforced limit, but can be limited if the usage affects other users

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 in principle considered as I/O intensive.

I/O

Parallel I/O is advised but not mandatory for applications running on EuroHPC systems. Therefore, the applicant should 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). For the Karolina system, the users should also quantify in addition to the mentioned requirements, the total volume and throughput GB/s.