European High Performance Computing Joint Undertaking

Call for tenders EuroHPC/2023/OP/0002.

ACQUISITION, DELIVERY, INSTALLATION AND HARDWARE AND SOFTWARE MAINTENANCE OF THE UPGRADE OF THE EUROHPC SUPERCOMPUTER – DISCOVERER+

Open procedure

TENDER SPECIFICATIONS

[Part 2: Technical specifications]

TABLE OF CONTENTS

1. BACKGROUND AND OBJECTIVES

The goal of this call for tenders is the procurement, installation training and maintenance of an upgrade of a supercomputer system Dicoverer+ that is planned to be installed at SOFIA PETASCALE SUPERCOMPUTING CENTRE's supercomputing facility by March 2024.

The upgraded system will support EuroHPC JU and Sofia Petascale Supercomputing Centre - Bulgaria commitment to providing leading-edge innovative computing resources to European research.

2. DETAILED CHARACTERISTICS OF THE PURCHASE

2.1. Requirement Classification

While drafting the tender, the tenderer is invited to specifically describe how it accommodates the objectives that the EuroHPC JU has defined for the proposed supercomputer upgrade. The requirements and features are categorised as follows:

Requirements & Features	Priority	Description
Mandatory Requirements	**Mandatory**	These are considered essential for the procured system and must be fulfilled by all Final Tenders. Mandatory Requirements will be assessed for each Tender submitted. Final Tenders which will not be compliant with all Mandatory Requirements will be rejected.
Very High Target Capability	**Very High**	Target Capabilities are desirable features and desirable performance levels for the procured system. In contrast to Mandatory Requirements, failure to provide Target Capabilities will not lead to the rejection of the Final Tenders provided by the Tenderer. Tenders that provide the Target Capabilities will receive a higher score. Target Capabilities are prioritised. Very High priority Target Capabilities (VH-TC) are considered of higher importance than high Target Capabilities (H-TC).
High Target Capability	**High**	Target Capabilities are desirable features and desirable performance levels for the procured system. In contrast to Mandatory Requirements, failure to provide Target Capabilities will not lead to the rejection of the Final Tenders provided by the Tenderer. Tenders that provide the Target Capabilities will receive a higher score. Target Capabilities are prioritised.
Documentation	**Doc**	Documentation that must be included in the Tender. All documentation items are mandatory and must be provided by all Tenderers in their Tender.

Table 1 Requirements Categorisation

2.2. SOFIA PETASCALE SUPERCOMPUTING CENTRE Responsibilities

SOFIA PETASCALE SUPERCOMPUTING CENTRE takes responsibility for the following tasks:

1. Whenever possible and feasible, Discoverer Petascale Supercomputer Consortium will take responsibility for the procurement, installation, and operation of the management Ethernet networking equipment, except for NICs and associated optic fibers and transceivers.
2. Even so, the Tenderer should include in the offer (as an option) the designing of a management network and description of the necessary Ethernet equipment required to build it. The offer must also specify which elements in the proposed design are considered optional and may not be ordered/purchased.
3. Provision of the Ethernet backbone for the storage, UPS and GPU partition connectivity.
4. Preparation and operation of the facility.

2.3. SOFIA PETASCALE SUPERCOMPUTING CENTRE Specifications

The table enlists the technical characteristics of the site (hosting facility), whereupon the procured system needs to be delivered and installed. In there, the priority column indicates the degree of flexibility acceptable for fulfilling the requirements. "MANDATORY" means the actual values are considered fixed and non-changeable (for instance, the maximum height of the racks cannot be higher than 2.6 meters). "VERY HIGH" allows the requested numbers to be varied only within the specified ranges.

Computing Centre specifications	**Actual Values**	**Priority**
Datacentre	*Supercomputer centre building in zone 5 of Sofia Tech Park: 111, Tsarigradsko shose blvd., Sofia, Bulgaria*	*MANDATORY*
Maximum height of the rack	*2.6 m*	*MANDATORY*
Raised Floor	*· tiles of 0.6 m × 0.6 m;* *· 60 cm above the ground;* *· maximum weight withstanding: 20kN/m^2;* *· maximum single point load is 20kN.*	*VERY HIGH*
Concrete Floor	*capable of withstanding up to 20kN/m^2 in most places (except predetermined breaking points, etc.).*	*MANDATORY*
Room temperature	*air-conditioned to support 20-25 °C*	*VERY HIGH*
Room Humidity	*relative humidity of 40–60%.*	*VERY HIGH*
Maximum available space for the computer and related infrastructure	*500 m^2*	*MANDATORY*
Maximum available space (in the computer hall)	*Not more than 200 m^2*	*MANDATORY*
Electric power	*50 Hz with 3 x 400 Vac between phase lines and 230 Vac between phase and neutral lines.*	*VERY HIGH*
Maximal Power Consumption for the Discoverer +	*The maximum power consumption of the system at any point in time under any workload will be 100 KW, the Linpack consummation should be maximum 100 kW*	*VERY HIGH*

Maximum additional cooling capacity	*100kW*	*VERY HIGH*

Table 2 High Level Hosting Site Specifications

The specification of the sizes and location of the new system in the supercomputing centre is available at the end of this Annex. Sofia Petascale Supercomputing Centre reserves the right to change the position of the installation site and its sizes, without obsoleting the numbers given in Table 2.

2.4. Sofia Petascale Supercomputing Centre Architecture

Presently, the Sofia Petascale Supercomputing Centre site is fully operational. It is built as a frame structure of precast elements of reinforced concrete, on a foundation slab to withstand the necessary loads. Reception area, service area, loading ramps, fire suppression system, and server room are available in the buildings. The server room floor is elevated by 60 cm, resulting in a bearing capacity of 2200 kg/m^2. All applicable Eurocodes were promptly met during the designing and construction process.

2.5. Special Focus and Important Needs for GPU HPC Partition

The following performance targets and functional requirements are considered "important", and they need to become an entire part of the offer submitted for **GPU HPC Partition:**

1. The GPU HPC Partition must assume the delivery and installation of four homogeneous (based on the same hardware and software configuration) compute nodes (servers).
2. Each of the compute nodes hosts 8 GPU devices of the same type.
3. All GPU devices installed on each compute node must have the ability to become united "big" GPU device.
4. The maximum system power usage under full load cannot over exceed 50kW for all nodes.
5. The inlet operating temperature for each compute node can vary within 5–30°C.
6. Performance of each compute node must meet the following productivity goals:
7. GPU partition must be delivered including high speed interconnect network, that means including high speed interconnect switches and all cables required for connection between all hosts and switches. Required topology is non-blocking fat tree between all GPU nodes, all interfaces in GPU nodes and between GPU partition and connection to interfaces to storage system.
    a. Each node should be connected to the InfiniBand network with a network interface controller, at least one interface per each GPU card. Speed of interface must be at least 400 Gb/s
    b. It is required each high-speed interconnect switch to have a throughput at least 51 Tb/s
    c. The adopted communication protocols and infrastructure topology enhancements must be selected in a way that enables a direct data path between the local or remote storage, such as NVMe or NVMe over Fabric (NVMe-oF), and accessing directly the GPU memory over RDMA
    d. Infiniband network must have free ports with possibility to add more identical GPU nodes (at least 50 percent more than delivered) and connect them without need to change topology or add more switches
8. Each GPU node will have at least 2TB RAM with balanced configuration (all memory channels populated, all DIMMs same type and size)
9. Each GPU node will have internal NVMe storage for OS (at least 2 x 1,9TB) and for scratch (at least 8 x 1,9TB Brutto)

10. Each GPU node will have as well at least 2 x 400Gbit connection (two physical ports) for possibility to connect to high speed ethernet or Infiniband (NDR/HDR, possibility to connect to existing Infiniband infrastructure)
11. Each GPU node, all GPUs inside node must communicate with dedicated high speed interconnect bus. At least 900 GB/sec is required.
12. GPU partition will be delivered including racks and all equipment needed for installation of GPU nodes in the racks (like cabling, rackmount kits etc). All the racks will be identical and must be in dimensions to fit into existing DC space (maximum 60cm wide) and capable to carry all required HW (at least 120cm internal depth space). For cooling purposes, the rack has to be equipped with a hot water, direct liquid cooling door. The temperature of the water flow entering the rack should be 13 degrees Celsius. All racks must have active monitoring of temperatures and system for airflow and liquid flow regulation, to be neutral to DC environment and have possibility to set output temperature as required by tendered.
13. GPU partition will be connected to existing system, Discoverer cluster. For these purposes, it is required to use resources of existing Discoverer cluster (management nodes, login nodes, management and software tools, SLURM), so GPU partition will be managed from same environment as Discoverer as another partition to existing CPU partition. If it requires any SW or HW equipment addon to existing Discoverer system (HBAs, cabling, SW), must be delivered and installed as part of complex delivery of GPU partition.
14. Any SW or tool, related to GPU partition and required to meet criteria described in all other parts of RFP, related to SW part, storage part and other parts and needed for GPU partition to run, operate and work as required by tendered for their applications (like job orchestration, job scheduling, IO libraries etc) must be delivered as part of complex delivery of GPU partition.

The lots are fully independent it is not obligatory the bidders to offer all of them. The maximum budget for this lot 1 300 000 EUR.

2.6. Special Focus and Important Needs for two new storages

2.6.1. High performance HPC optimized 5 PB (or more) storage system with distributed file systems support

Currently, the Discoverer PetaSC supercomputer has only 2PB DDN of distributed file system storage. To increase the storage availability, and scalability, its storage capacity needs to be enhanced to support massive HPC calculations, which imposes a proper handling of the resulting large I/O loads that inevitably will occur on the storage servers. Therefore, a new, high performance, HPC class storage system with at least 5 PB of physical storage capacity, based on the Lustre File System (LustreFS), is demanded. The design of the new storage system must be able to handle at least 1 million IOPS and RRW > 24 GBps operating on NVMe flash pool, with minimum physical storage capacity of 500 TB. Its network connectivity needs to be based on at least 8 x InfiniBand HDR 200 Gbps ports, seamlessly connected to the existing 2PB DDN storage. The new storage system must be delivered along with all the necessary racks, InfiniBand HDR switches and cables, and then integrated into the existing IB HDR Dragon Fly+ network. The delivered, installed, and put into the operational storage system, must be covered with at least 3 years of technical support.

In details, the offered system must provide/include:

1. Integrated, high performance Lustre® parallel file system for HPC applications;
2. The delivered solution must provide global single-namespace, POSIX-compliant parallel file system;
3. Utilization of the parity of de-clustered RAID for data protection against drive failures;

4. Support for mixing LDISKFS and OpenZFS at the level of its back-end file system;
5. Support for horizontal data moving capabilities between multiple Lustre systems;
6. Integrated system management interface with GUI and CLI access system status, enhanced performance details, health status, configuration details etc.;
7. Management Service Unit that takes care of the storage systems management. The Management Service Unit should consist of a minimum of one pair of server nodes configured in a high availability configuration with storage failover. The two server nodes should include dedicated shared storage capacity of at least 5x 1.6TB mixed-use NVMe drives. Each server node should have at least one 1-port 200Gb/s InfiniBand HDR adapter. Cables for connecting all ports to the existing InfiniBand HDR infrastructure should be included;
8. Metadata Service Unit for persistently storing metadata information. The Metadata Service Unit should consist of a minimum of one pair of server nodes in a high availability configuration with storage failover enabled. The pair of server nodes should include a minimum of 24 x 3.2 TB mixed-use NVMe drives for dedicated shared storage capacity. Each server node should have at least two 1-port InfiniBand HDR adapters, 200Gb/s per port. Cables for connecting all ports to the existing InfiniBand HDR infrastructure should be included;
9. Object Storage Unit to store data on NVMe drives. A minimum of two pairs of server nodes should be used to store bulk storage of data on NVMe drives. They should be used in a high availability configuration with storage failovers. Each pair of server nodes should include a dedicated shared storage based on at least 24 NVMe drives. Each server node should have at least two 1-port InfiniBand HDR adapters, 200Gb/s per port. After data protection against minimum two simultaneous drive failures, the Object Services Unit should provide a total usable capacity of at least 500 TB for a bulk storage of data on NVMe drives. Additionally, additional spare capacity must be included based on the vendor's best practices. Cables for connecting all ports to the existing InfiniBand HDR infrastructure should be included;
10. A bulk storage of data on HDD drives included in the system. A minimum of two pairs of server nodes should be used for storage of bulk data on HDD drives. Each pair should be in a high availability configuration with storage failover. Each pair of server nodes should have a dedicated shared storage capacity based on HDD drives. There should be a minimum of 2 ports of 200Gb/s (per port) InfiniBand HDR on each server. The object services unit for storing bulk storage of data on HDD drives should provide a total usable capacity of minimum 5 PB after data protection against minimum two simultaneous drive failures. Additional spare capacity must be included based on the vendor's best practice. Cables for connecting all ports to the existing InfiniBand HDR infrastructure should be included;
11. At least two network switches for data management. Each switch should provide at least 48 x 1/10Gb Ethernet ports and 4 x 100 GB QSFP28 ports. The switch should provide support for advanced Layer 2/3 features, such as BGP, OSPF, VRF, ARP, and IPv6. There should be 2 100Gb cables for the inter-switch connection. The cables that are needed to manage the data on all the system units should be included;
12. All licenses required for all requested features, covering the entire support period;
13. Option for future use to enable tiering, combining NVMe and HDD into a single namespace;
14. At least 36 months of warranty on the vendors' hardware equipment. Additionally, the vendor must hand over to DISCOVERER a support service that includes software support with 8x5 coverage, incident logging, access to online self-serve and self-solve capabilities, remote support, and firmware updates. Up to 4-hour on-site attendance by a certified engineer is to be provided upon request as part of the support service;

15. Hardware and software installation and startup services.

2.6.2. High availability and performance multi-protocol storage system

High availability and performance storage are both essential for the productivity of the nodes in the new GPU HPC partitions. The Discoverer petscale system is hosting Nimbix infrastructure which is currently in use.

To provide reliable service, the GPU HPC compute nodes need fast accessible storage, based on NVMe disks, capable of handling dozens or hundreds of parallel reads and write operations simultaneously. Such levels of operation are necessary to support computational jobs run by scientists, engineers, government services, and business in the areas of (but are not limited to): artificial intelligence, training neural networks, machine learning, image recognition, open information architecture, digital twins, quantum chemistry, material sciences, geophysics, weather forecasting, simulation for preventing and forecasting natural disasters (floods, landslides, earthquakes, fires), transformation of very large-scale biological neural networks into digital equivalency, and their solution.

To meet the high demands for storage system, the provided offer needs to meet the following list of required features:

1. High-performance parallel file system with multiprotocol access to a global namespace;
2. Handling of distributed metadata without dedicated metadata servers;
3. Support for shared POSIX access to single global namespace;
4. NFS v3, v4 protocol support (over RDMA, whenever the protocol allows that);
5. SMB v3 protocol support;
6. S3 API support (create, delete, object/bucket list);
7. Container Storage Interface (CSI) support;
8. GPUDirect Storage (GDS) protocol support;
9. Access to the stored files via NFS, S3, SMB and GPUDirect Storage;
10. Support of at least 15000 snapshots;
11. Snapshotting should be instantaneous with no performance impact;
12. Support for configurable snapshots as read-only and writable;
13. Support for thin provisioning;
14. Support of file sizes up to 2PB;
15. Integrated Graphical User Interface for management;
16. Whenever a license is required for using and implementing any of the protocols and features requested, it must be included in the offer;
17. A tiering of inactive data to a lower-cost object storage tier should be available for activation in the system at any moment;
18. At least 430TB of usable capacity after data protection against at least two hardware failures at the same time ($N + 2$). Additional spare capacity must be added based on the best practices of the vendor;
19. Storage capacity stored on NVMe drives;
20. Scale-out architecture containing at least 8 clustered storage nodes;
21. The system must allow cluster expansion by addition of more storage nodes in future with support for minimum 128 nodes;
22. Adding NVMe drives on the storage nodes on demand/request;
23. At least 2 x InfiniBand 200Gb/s HDR ports per node, each one located on a separate InfiniBand adapter, and 4 x 1Gb/s Ethernet ports;

24. Full connectivity and compatibility with the existing InfiniBand HDR infrastructure at 200Gbps.
25. The solution must include all active cooling racks as well as all essential cables for connecting the devices to the infrastructure, and it must be properly installed and tested.
26. Least 36 months of warranty on the vendors' hardware equipment. Additionally, the vendor must hand over to DISCOVERER a support service that includes software support with 8x5 coverage, incident logging, access to online self-serve and self-solve capabilities, remote support, and firmware updates. Up to 4-hour on-site attendance by a certified engineer is to be provided upon request as part of the support service;
27. Hardware and software installation and startup services included;

Important requirement: The storage system components are required to be delivered and installed inside 42U racks (the racks are to be delivered too), together with all necessary 3-phase power distribution units, mounting and cooling accessories kits, cables and liquid-cooled solution for the rack.

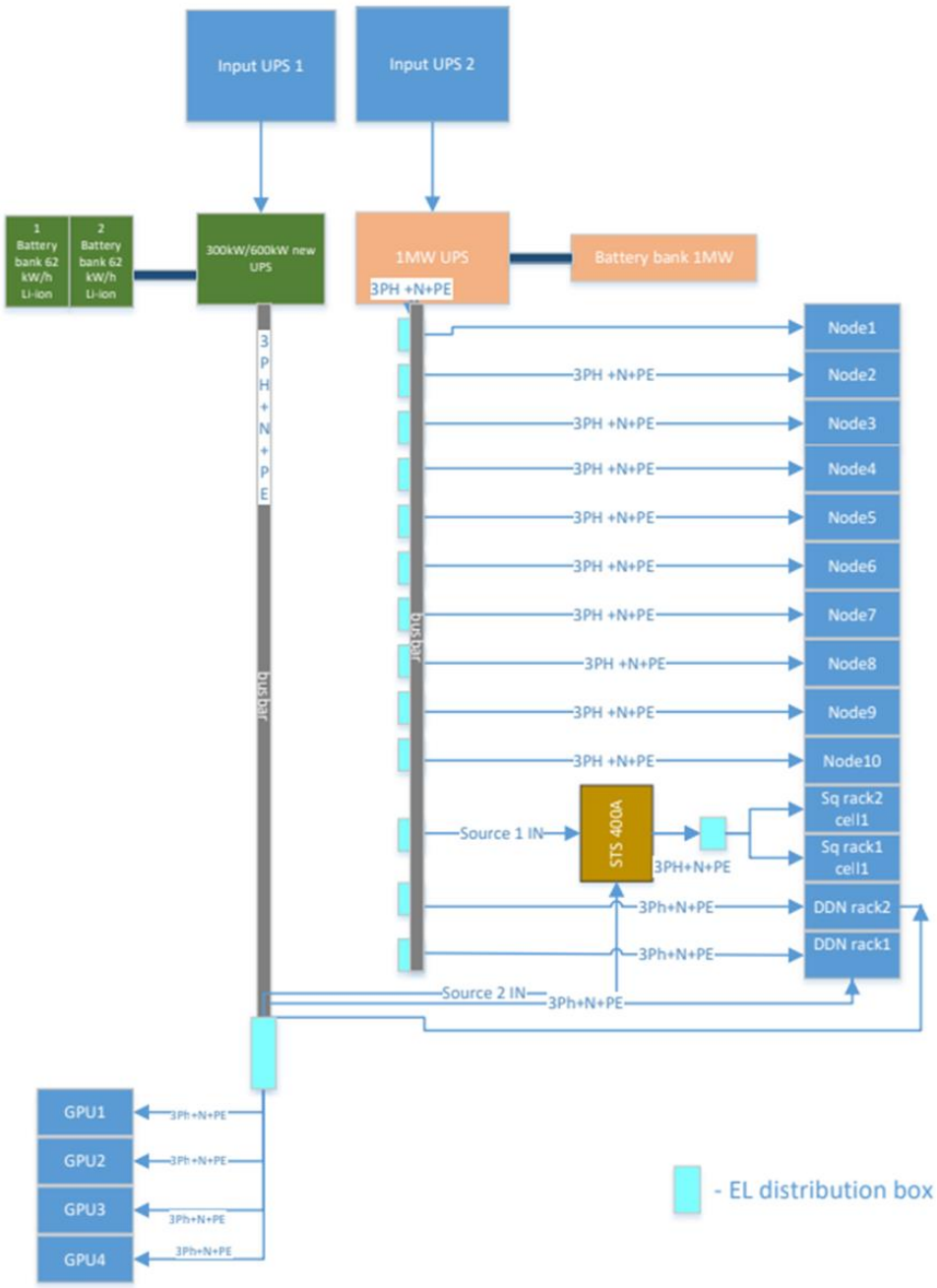
The lots are fully independent it is not obligatory the bidders to offer all of them. The maximum budget for this lot is 1 500 000 EUR.

2.7. Special Focus and Important Needs for additional UPS selection.

There are two independent UPS systems at the Sofia Petascale Supercomputing Centre. The first is dedicated to backing up the Discoverer Supercomputer. It is Galaxy VX 1000kVA, 400V. That UPS system is highly efficient, scalable, and has 3-phase power protection, flexible operating modes, and ECOnversion™ UPS. The second UPS is INGENIO PLUS 3-phase 160 kVA, responsible for supporting facilities for cooling (pumps, sensors, fens and others), security (cameras, DVRs, and others), emergency lighting and others. Currently, the Galaxy UPS is fully loaded without the possibility of offloading it by adding more capacity. This is the reason to request a new UPS system, which is responsible for power supplying the new equipment and the critical part of the old equipment.

Sofia Petascale Supercomputing Centre is supplied by three independent power supply lines, but it is not equipped with an on-site power generator, yet. To partially compensate for that lack of additional power source, a new UPS is requested. Its duty will be to power supply the nodes in the new GPU HPC partition, new storage system, new network equipment, two new air-cooled racks, and two existing liquid cooled Bull Sequana XH2000 DLC racks (those racks host critical components of communication, security, storage, login nodes, management nodes, performance node and compute nodes).

Principal scheme of new UPS, Racks and STS switch connection.



Technical requirements for the new modular 3-phase UPS, 300kW (250kW+50kW), N+1, hot-swap, extendable to 600kW.

	Characteristics (UPS)	Requirements	
		min	max
I	**Main characteristics (UPS)**		
	Maximum power	250kVA/230kW	250kVA/230kW
	Initial power (without redundant module)	250kVA/250kW	250kVA/250kW
	Initial power (redundant module included)	300kVA/300kW	300kVA/300kW
	Module power	50kW	50kW
	Input voltage range (full load)	3ph+N, 400VAC (310 - 475VAC)	3ph+N, 400VAC (310 - 475VAC)
	Input frequency	50 Hz +/- 10%	50 Hz +/- 20%
	Input PF at 50% load	>=0.99	>=0.99
	Input THDI at 100% load	<= 3%	<= 3%
	Output voltage	400 V +/- 1%	400 V +/- 1%
	Output frequency	50 Hz +/- 0.05Hz	50 Hz +/- 0.05Hz
	Output THDU at 100% non-linear load	<= 5 %	<= 5 %
	Cable entry	Top	Top
	Integrated by pass circuit breakers (input, output, manual and service)	Required	Required
	Max UPS+BYPASS size (HxWxD)	2000 x 1200 x 1100	2000 x 1200 x 1100
	Invertor overload at <= 125%	>= 5min	>= 10min
	Efficiency at load 30% (double conversion mode)	>=96%	>=96%
	Efficiency (ECO mode)	>= 98%	>= 98%
	Working temperature	5 to +35 C	0 to +40 C
	Relative humidity	20 - 80%	0 - 90%
II	**Main characteristics (batteries)**		
	Backup time at 200kW load	>=25min	>=25min
	Battery type	Li-Ion (NMC)	Li-Ion (NMC)
	Battery in racks	Required	Required
	Max number of racks	2	2
	Capacity per rack	>= 62kWh	>= 62kWh
	Max rack size (HxWxD)	2002 x 600 x 1090	2002 x 600 x 1090
	Cable entry	Top	Top
	Standard UN 38.3 (cell and module)	Required	Required
	Standard UL1642 (cell)	Required	Required
	IEC62619 (rack)	Required	Required
	Battery management system providing U, I, SoC and other data to the UPS	Required	Required
	Working temperature	5 to +35 C	5 to +35 C
	Relative humidity	20 - 80%	5 - 85%

III	**Other requirements**		
	Topology N+1 (5 x 50kW + 50kW)	Required	Required
	"ONLINE", double conversion	Required	Required
	Modules replacement (hot-swap) without going on by-pass	Required	Required
	SNMP remote monitoring	Required	Required
	Color touch screen	Required	Required
	Warranty period (UPS + batteries)	2 years	2 years
III	**Source Transfer Switch 400A**		
	Input voltage	3ph+N, 400VAC	3ph+N, 400VAC
	Input frequency	50 Hz +/- 10%	50 Hz +/- 10%
	Nominal surrent	400A	400A
	Switching poles	3 phases only, Neutral is hard connected	3 phases only, Neutral is hard connected
	Switching time (synchronized sources)	0ms	0ms
	Switching time (non synchronized sources)	<= 4ms	<= 4ms
	Switching method	Break Before Make	Break Before Make
	Efficiency	>= 99%	>= 99.3%
	Protection type	Circuit breakers	Circuit breakers
	Working temperature	5 to +35 C	0 to +40 C
	Relative humidity	20 - 80%	20 - 80%
	Cable entry	Bottom	Bottom
	Warranty period (STS/400A)	2 years	2 years

The lots are fully independent it is not obligatory the bidders to offer all of them. The maximum budget for this lot is 240 000 EUR.

2.8. Target Workload Description

2.8.1. Workload description of GPU HPC partition

The design of the GPU HPC partition aims to support high-capacity workloads for running massive computations in parallel. Under typical production conditions, the average GPU compute node utilization per job varies between 70-85% (on a single GPU device) to about 45-75% (32 or more GPU devices are taken on a single or multiple nodes).

The reported resource consumption profile is typical for a variety of applications and methods: AI, training neural network (NN), machine learning (ML), image recognition, open IA, digital twins, quantum chemistry and material sciences, geophysics (earthquake simulations and disaster estimation), weather forecasting, simulations for preventing and forecasting natural disasters (floods, landslides, earthquakes, fires), transformation of very large-scale biological neural networks in digital equivalency and their solution, and others.

2.8.2. Workload description of new UPS

The requested design of the new UPS can support the specified workloads while offering all required features for capability. Under typical production conditions, the load is expected to vary between 50-85% on the CPU HPC node.

Important: Backup time at 200kW load >=25min.

3. TECHNICAL REQUIREMENTS

There are seven different technical areas with requirements for this procurement:

1. Energy efficiency and power management;
2. Data management;
3. Programming environment and productivity;
4. Data centre integration;
5. Maintenance and support;
6. System and application monitoring;
7. Security.

These topics are defined in terms of common goals and objectives:

- **Goals:** Long term goals that may not necessarily be fully achieved within the expected lifetime of the system. They come from important ideas and needs that EuroHPC JU and the Hosting Entities want to push forward.
- **Objectives:** Measures that are attainable within the expected lifetime of the system. They might require collaborative efforts between the Procurers and the awarded Tenderer.

The Tender is asked to explain the extent their solution can meet these topics.

3.1. Energy efficiency and power management

Improvement in the energy efficiency of systems, by controlling the energy consumption through flexible and programmable resource optimization and accounting, is the major goal of this topic. Solutions that will reduce the Energy-to-Solution (ETS) of the typical workload with little or no impact on its performance are targeted. Please note that energy efficiency will play a role in the performance evaluation (see 0).

To make progress towards these goals, the following objectives are set within the system specifications:

- Enable correlation between power consumption and system workload;
- Enable dynamic power capping with graceful performance degradation of the system;
- Provide energy accounting mechanism;
- Allow energy profiling of applications to enable ETS optimization without causing performance degradation.

The procured system will feature advanced mechanisms for power and energy monitoring and control.

Req.No.	Priority	Description
EF1	High	Power Measurement Capabilities: The procured system will be able to measure GPU devices power consumption. The Tenderer will describe the power measurement capabilities, as well as the implemented data collection and storage mechanisms.
EF2	High	Energy Measurement Capabilities: The procured system will be able to measure energy at compute node level in time intervals of at least 60 seconds. Tools that can aggregate the measurements taken from multiple nodes will be provided. The error for measuring energy consumption over a period from 1 minute to 12 hours will not exceed 3%. The impact of the data collection on the application performance will be low. The Tenderer will describe the energy measurement capabilities, and the implemented data collection and storage mechanisms.
EF3	High	Job Energy Accounting: The Tenderer will provide tools and/or an API to enable energy accounting of Slurm batch jobs. The job energy data should preferably be stored along with other job accounting records in the database of the Slurm workload manager and be available via the same tools and the same API.

Table 3 Requirements on Energy efficiency and power management

3.2. Data management

The following objectives have been set to provide a solution:

- Reduce integrity or recovery times in case of failures.
- Improve monitoring of the healthiness of GPU HPC partition.
- Provide tools to get a global awareness of the usage of the GPU partition infrastructure.
- Optimize and reduce the time of data access of computing processes.

Req.No.	Priority	Description
DM1	Mandatory	An automatic recovery procedure must be executed and carried out if a component inside a GPU HPC partition infrastructure is reported/detected as failed. The proposed process must maintain data integrity and maintain normal accessibility. The evaluation will consider the performance of the recovery.
DM2	Very High	The duration of the recovery process of the procured system will not exceed 8 hours and the rebuild will cause less than 10% loss in the performance of the partition. The evaluation will consider the number of hours required and the impact on the performance.
DM3	Very High	Infrastructure, processes, firmware, drivers, and processes responsible for providing and controlling the I/O communication with the host and the hosted GPU devices (e.g. related to services like parallel file systems, network file systems, network communication protocols), must not create long-lasting high peaks in the I/O load of the procured system, if there are no running user applications that require intensive I/O communication

		with the GPU devices, the CPU and the memory on the host, or with the network adapters.

Table 4 Requirements on Data Management

3.3. CPU architecture and OS support

The proposed server system configuration must be equipped with at least two processors (CPUs) that support the 64-bit version of the x86 instruction set (x86-64). Each processor must have 32 CPU cores with two threads per core. AVX, AVX2, and AVX512 SIMD instructions must be supported in the processor. The AMX SIMD instructions support is considered optional, but preferable.

The server system must be capable of running Red Hat Enterprise Linux (RHEL), version 8.7 or higher, designed to operate on the x86-64 CPU architecture. The server system is expected to come with an installed Linux distribution, which can be utilized for initial testing. No RHEL installation with an active subscription must be shipped with the server system.

3.4. Programming environment and productivity

The goal is to provide application developers with up-to-date development software that satisfies their needs for programming language and parallel programming paradigm support and to foster a common modern environment that meets the needs of European users and is available for them on all EuroHPC systems.

To provide more flexibility and productivity to systems and to facilitate deployment of specific software stacks, e.g., platform software stacks of European research communities, the support of (lightweight) virtualization mechanisms are of interest.

To address these goals the following objectives are set:

- Provide support for recent implementations of common parallel programming standards;
- Provide uniform support for efficient execution of complicated multi-cores and multi-threading simulation workflow;
- Provide full support and system integration for customizable execution environments using containers based on common formats;
- Push level of virtualization towards support of virtual machines.

Req.No.	Priority	Description
PP1	Mandatory	The procured system must provide an adequate development environment to support at least the following programming languages: 1. Fortran: ISO/IEC 1539-1:2010 (aka Fortran 2008) or newer 2. C: ISO/IEC 9899:2011 or newer 3. C++: ISO/IEC 14882:2014 or newer 4. C, C++, and FORTRAN languages in versions and with syntax enhancements supported by LLVM compiler infrastructure model (version 14 and higher) with the following capabilities: MLIR Sparse compiler and GPU code generators, fixed GPU compute capability code generators, multi device GPU code generators,

		language-independent intermediate representation (IR) with dead-code elimination (DCE), code generators that support CPU–GPU unified memory programming model 5. C, C++. and FORTRAN languages with OpenACC syntax extensions that support GPU programming 6. Python 3.6 or newer To support direct programming and code profiling in real time, the system must be able to execute binary code compiled for the 64-bit version of the x86 instruction CPU set. It is essential to provide the ability to run LLVM Just-In-Time (JIT) compiler. The details of the implementation will be taken into account in evaluating the Tenders.
PP2	Medium	The procured system must be able to host and execute C, C++ and Fortran compilers that can link the generated binary code at compile-time to the libraries for parallel code generation and profiling tools OpenACC, ver. 2.7 or newer; OpenMP, ver. 4.5 or newer; TBB, ver. 2021.9 or newer; MPICH, ver. 4.1 or newer; OpenMPI, ver. 4.1 or newer; based on the 64-bit version of the x86 instruction CPU set and SIMD instructions AVX, AVX2, AVX512, and AMX. It must be able to handle the profiling of the produced parallel executable code.

Table 5 Requirements on Programming environment and productivity

3.5. GPU support for scientific software

The support of the following software products is strongly preferred on the installed GPU devices as it is essential for the ongoing scientific community projects:

Amber PMEMD and AmberTools, ver 18, 19, and 20

GROMACS, ver. 2023 or newer

VMD, ver. 1.9.3 or newer

MagmaDNN, ver, 1.2 or newer

VASP, ver. 6 or newer

CP2K, ver. 2022 or newer, with ELPA support

GAUSSIAN, ver. 2016

Open Molcas, ver. 23 or newer

ABINIT, ver. 9.8.2 or newer

NAMD, ver. 2.14 or newer

GAMESS US, ver 2021 or newer

Quantum ESPRESSO, ver. 7.1 or newer

3.6. Data centre integration

Facility and infrastructure design and operation are crucial parts of each GPU partition deployment and day-to-day maintenance. This role is amplified by the increasing infrastructure complexity required to meet the higher density and power consumption of state-of-the-art supercomputers. At the same time, ease of operation, resilience, serviceability, and quick installation times remain important for supercomputing facilities. The goal of this topic is to foster developments of infrastructure technologies, including but not limited to liquid and air cooling, which strive to address these two, sometimes contradictory, targets.

The following objectives are set in the project:

- Speed up installation and simplify operation by minimizing the need for adapting the data centre infrastructure.
- Improve hardware and software installation process to shorten the overall system set up time;
- Improve energy efficiency of supercomputers and the facility (e.g. by decreasing the PUE);
- Set up communication links between the system and the facility to exchange status information and alarms.

Req.No.	Priority	Description
DC1	Very High	A re-installation of the procured GPU partition and software from scratch should take no longer than 1 working day when the hardware is already in place and in a stable condition. Site specific application software (additional user-accessible libraries) is not considered for estimating the duration of the process. The evaluation will consider the number of days required, for the Tenders that offer this capability.

Table 6 Requirements on Data centre integration

3.7. Maintenance and support

Each critical computer infrastructure should be designed and operated in a way that reduces the number and impact of planned or unplanned interruptions on the running jobs. To meet those high requirements, innovative solutions to reduce or even eliminate any maintenance downtime need to be implemented and available intentionally in the proposed new GPU HPC partition. Those requirements can be summarized as follows:

1. Reduction of the frequency and off time spent for the maintenance of the new GPU HPC partition;
1. Automatic hardware and software validation framework for examination of the health status of any hardware component in the partition;

2. Predictive failure detection systems for prevention and reduction of the maintenance, with a strong emphasis on the reduction of job failures due to problems with the hardware.

Req.No	Priority	Description
MS1	Very High	In order to reduce maintenance times, the "power on" or reboot time of the entire procured system, in the state "ready for production", will be minimized and take at most 20 minutes. In this case, the evaluation of the offer will consider the time (in minutes) proposed by the Tenderer.
MS2	Very High	The procured system needs to be capable of implementing tools to check and validate hardware and software health. To avoid the submission of new jobs onto the nodes with damaged components, those tools must first attempt to resolve the detected issue themselves (auto-recovery) and, if they fail, notify the Slurm workload manager at once. The implementation details will be considered when evaluating the offer submitted by the Tenderer.

Table 7 Requirements on Maintenance and support

3.8. System and application monitoring

Research and development on application profiling is an active and prolific field. However, often such tools are used solely during development, setup and optimization phases. The goal of this topic is to increase coverage of profiling tools by enabling continuous lightweight (application and job control upon the GPU device) sampling-based profiling of production jobs to identify optimization potential in production workflows and system software. This may encompass a variety of capabilities starting from the processing of system utilization information (which may be enough to identify improper thread count per node settings) up to fine-granular (e.g., function level) profiling information. Such a profiling technology should not be an "island solution" but rather must be well integrated in the overall system monitoring capabilities.

The following objectives are set in the project:

- Provide capabilities enabling collection, management and analysis of performance data for production jobs. Analysis of current as well as historical data should be possible;
- Provide tools for system operators to detect anomalies based on the collected data.

Req.No.	Priority	Description
SM1	High	The procured system is capable of providing lightweight, continuous performance profiling capabilities. The performance data collected at the level of process, job, and GPU device must be made available using scalable accumulation methods. Data retention times for the mentioned granularity levels may differ. The profiling and data collation technologies implemented should have minimal impact on application performance. The details of the implementation will be considered in evaluating the Tenderer that offers this capability. In an event of a work interruption (power failure, hardware failure) all intermittent data and jobs should be saved on NVMe discs located on each server, such that after the resuming the work of the machine, the job continues its work from the moment of interruption.

Table 8 Requirements on System and application monitoring

3.9. Security

Security is a strategic aspect of HPC facilities, to be mandatorily enforced at the appropriate level by legal regulations. The main goals for the operation of a secure HPC facility are:

- Maintaining the security level of the computer and data infrastructures in accordance with site policies and applicable laws;
- Guaranteeing user and group isolation at appropriate levels;
- Preventing downtime and service quality reductions due to malicious attacks, such as denying service attempts by means of system features and/or tools.

The following objectives are set in the project:

- Ensure the availability of periodic security patches and their quick application on the systems;
- Provide mechanisms to enforce isolation of different users/groups, including their data, on the systems.

Req.No.	Priority	Description
SE1	Very High	The procured system can provide features for job resources usage isolation (at user/group level) either on the batch scheduler, parallel file system (Lustre), block file system or all of them. The details of the Implementation will be considered in evaluating the Tenderer that offers this capability.

Table 9 Requirements on Security

The rules to be applied for the registration of private data mentioned above are based on Regulation (EC) 45/2001 of the European Parliament and of the Council of 18 December 2000[1]and Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016[2], and, if applicable, on the related national laws.

4. BENCHMARKS

4.1. Benchmark methodology

Measurement tools for architectural characteristics (synthetic benchmarks) and code tests (application benchmarks) will be provided to the Tenderer during the tender presentation meeting. These programs will be run by the Tenderer on a test system in accordance with the Final Tender.

All or part of the previous tests as well as workload tests (to assess the production system) can then be run by EuroHPC JU to check the results obtained by the Tenderer and to evaluate management tools of the environment.

4.2. Running a benchmark

Sources of codes to be used (including those of synthetic benchmarks) are the sources provided by the Contract Authority and not optimized versions possibly owned by the Tenderer.

Benchmarks must be performed by filling the processors and nodes in the best possible manner so that they consume a minimum number of compute nodes.

For pure performance measurement, if a dynamic GPU frequency management system is available, the latter must be deactivated to ensure reproducibility of the results. However, this mechanism can be activated for the measurement of energy.

Source code changes or settings provided will not be accepted except to prevent the code from failing and only after agreement of each Public Procurer. Such changes must be documented by the Tenderer to validate the committed performance.

Synthetic and application benchmarks must be run on the same computing environment: single operating system version, MPI and compilers (closest to the one that will be supported in production) with no reconfiguration or reboot between the two tests. The only authorized changes to the execution environment are those accessible to the user via environment variables or keywords of the batch management system.

The compiler options are freely editable only if they are also available in the compiler of the proposed solution.

The application benchmarks include several types of executions and input data on variable numbers of compute cores.

4.2.1. Benchmarks summary

A final summary of benchmarks for the test configuration (with an accurate description of this test configuration) should be provided by the Tenderer. For each result, the Tenderer must provide extrapolated value (if necessary) on which they will commit.

The method used for extrapolations must be described accurately.

4.2.2. Performance commitments

These extrapolations are part of the performance commitments and must be reproduced in the final configuration.

The Tenderer commits on all the specifications and the performance announced in their Tender and they will have to prove that all values are achieved in the final installed configuration.

5. SERVICES

5.1. Installation (including the project plan) of the Supercomputer

The deployment and installation of the entire system are split into the following phases:

1. Hardware and software installation;
2. Provisional acceptance;
3. Pre-production qualification;
4. Final acceptance.

5.1.1. Installation plan

The Tenderer must provide a "supply and installation project" for the GPU partition in the offered Tender. This project must detail:

- Planning and the delivery times of the various parts of the system;
- The offered system configuration and integration in the Sofia Petascale Supercomputing centre computing system architecture, complete with its setup and interconnection schemes;
- The Hardware and software installation plan, configuration and optimization of the components and partitions, and the interaction with the EuroHPC JU and Sofia Petascale Supercomputing centre personnel;
- Plan for testing and validation (see section 4.1.4 Pre-production qualification);
- Services for the installation, testing, and migration, in terms of activities and time schedule (GANTT chart);
- Characteristics of the offered training course (e.g., duration and course program);

In the project, the Tenderer must provide also:

- Constraints and assumptions, regarding minimum release software and matrix compatibility with existing Sofia Petascale Supercomputing centre's infrastructure architecture and technologies used;
- Accessory hardware and software components not supplied, that the Sofia Petascale Supercomputing centre should procure on its own, which are necessary for the full and correct use of the components provided in the project to implement the required functionality;
- Methodologies and activities needed to implement the required features.

5.1.2. Hardware and software installation

The hardware and software installation phase are completed once the Supplier has delivered and installed all elements of the entire system in accordance with the Final Tender (hardware and software).

5.1.3. Provisional acceptance

After finalization of the hardware and software installation, and following the declaration of readiness by the Supplier, the system will be validated in this phase.

The Supplier will reproduce the committed performance values on the installed final system. They must also demonstrate that the various Target Capabilities proposed in the final technical solution are available and functional.

The following tests will be performed:

- Checking compliance with the contract of the hardware and software delivered;
- Tests carried out by representatives of the Supplier and the technical staff of the Sofia Petascale Supercomputing centre in the presence of the representatives of the Supplier to ensure proper functioning of the system and its environment;
- The benchmark results provided by the Supplier.

Provisional acceptance will be given at the end of this phase.

5.1.4. Pre-production qualification

The goal of pre-production qualification is to check that the GPU partition is performing as expected during an early operational stage. Key elements are the stability, reliability and performance and proper connection to the storages and CPU partition.

During this period, the proper functioning of the mechanisms necessary for the production environment is validated. During this period, the Supplier will make the adjustments necessary to achieve the availability during production requested by the EuroHPC/Sofia Petascale Supercomputing centre in the technical specifications.

The stability of the machine will be assessed using the methodology provided in the lot technical specification documents. The lot technical specification documents also define additional activities co-located with the pre-production qualification, such as the transfer of information and training for the technical staff of the Sofia Petascale Supercomputing centre.

5.1.5. Final acceptance

This final acceptance will validate the proper functioning of the entire system during the pre-production qualification period.

5.2. Maintenance and support of the GPU partition

In this section, we describe the SLA regarding maintenance and support services.

The Supplier must ensure the features of the maintenance and support service described below.

The Supplier must ensure the provision of the maintenance and support services, even if there is a dispute with EuroHPC JU. Such services will include the following activities:

- Direct access to database, software updates and fault opening procedures provided by equipment manufacturers;
- Support for malfunctioning of hardware or software components supplied;
- In case of failure, sending of replacement parts by the next business day (NBD);
- Release of software updates (firmware, drivers, micro-codes) for bug fixing or adding new features; note that the term “update” also refers to new versions ("releases") of the software;

- Assistance with configuration and tuning and best use of the products provided by the Sofia Petascale Supercomputing centre engineers responsible for the system management.

The maintenance and assistance service applies to all the products that are acquired by EUROHPC JU as part of the HPC systems. The maintenance service must include all activities required to ensure regulatory adjustments to software and equipment with reference to all European, national and regional regulations. All goods included in the service at its launch, even repaired or replacing parts, must comply with current regulations and their evolution. All maintenance and service interventions must be properly documented. The Supplier or their agent is required to provide the necessary technical assistance, strictly respecting the conditions and the intervention times defined in the specifications. The Supplier is responsible for the professionalism of the technicians in charge. All parts provided must bear the CE mark and comply with current technical and safety regulations or any regulations issued subsequently. The Supplier must specify the compliance of its systems with the applicable safety and emission regulations and electromagnetic compatibility at the time of their offer.

The Supplier or their authorized representatives may intervene in the following manner:

- Remote intervention for malfunctioning;
- On-site maintenance, if malfunction does not allow a remedy via remote intervention.

Maintenance activities requiring direct intervention must be agreed upon with the personnel appointed by EUROHPC JU. The intervention must ensure complete restoration of full operation, including technical specifications, the analysis and diagnosis of malfunctions, and may be carried out in collaboration with the staff of the Hosting Entities or other companies or personnel appointed by EUROHPC JU, if necessary.

Ordinary maintenance includes all of the hardware, with the obligation to replace any affected part without exception and at no charge for EUROHPC JU or the Sofia Petascale Supercomputing centre.

An intervention is considered concluded when all normal operating conditions have been restored and the operating conditions prior to the failure are fully restored. Only in suitable cases motivated by the Supplier or their authorized representatives, may alternative solutions be adopted.

5.2.1. SLA of maintenance and support services

The Supplier must ensure the following service levels, regardless of whether the service is provided directly by the Supplier or by the manufacturers of the various components or by other companies specializing in this type of service.

- Intervention time in case of blocking failure: less than 30 min from malfunction reporting (for 100% of the reported cases);
- Recovery time in case of blocking failure: less than 1 working day from malfunction reporting (for at least 100% of the reported cases);
- Recovery time in case of non-blocking failure: less than 2 working days from malfunction reporting (for at least 100% of the reported cases);
- Submission time of replacement parts: less than 2 working day from malfunction reporting (for 100% of the reported cases);
- Response time for service requests: no more than 4 business days.

Recovery time is the time interval between the malfunction reporting and the end of the troubleshooting procedure.

All of the above times are independent of the number of simultaneous incidents: in the event of multiple occurrences of simultaneous failures, the Supplier must provide technical and logistical support and timing for each individual intervention.

5.2.2. Planned maintenance

The Supplier must offer high availability design and network connectivity without planned maintenance windows. This is valid for the servers, network devices and links to the storages.
The Supplier and EuroHPC JU will agree on the content and duration of each maintenance session with the Petascale Supercomputer Bulgaria team. Maintenance beyond this period will lead to the system being considered unavailable.
Scheduled maintenance operations will take place as far as possible at the same time on the different systems running in the computing centre and they will be allowed only in case of extraordinary vendor requirements.
The Supplier must be present on site to ensure that the configurations are stopped and restarted correctly. They will be able to use these periods to carry out interventions (agreement of the Sofia Petascale Supercomputing centre is required).

5.2.3. Corrective maintenance

Malfunctions affecting the availability of the system are detected by an alarm system (set up by the Supplier), which systematically provides information on the GPU partition's health to the teams of the Supplier and of Sofia Petascale Supercomputing centre.

Outside of normal working hours, the monitoring system raises an alarm and alerts Sofia Petascale Supercomputing centre's on-call teams who are then responsible for notifying the Supplier's teams. These alert messages can also be sent directly to the Supplier. Remote point access to the Sofia Petascale Supercomputing centre site can also be provided to the Supplier (no guarantee of full-time availability) to carry out a diagnosis or even a remote intervention and thus reduce downtime.

5.2.4. Preventive maintenance

To respect its availability commitments, the Supplier will be assigned time to carry out preventive maintenance operations on hardware and software. These preventive actions carried out by the Supplier with no impact on the availability of the systems will also have to be documented using the configuration monitoring and intervention tools. A summary of these preventive actions will be provided in weekly meetings and site meetings.

5.2.5. Call centre and its SLA

During the period of providing maintenance and support services, a call centre must be reachable by phone via one or more numbers. The help desk functions of this call centre are related to the support following activities:

- Troubleshooting hardware or software components;
- Coordinate the delivery of replacements in case of failure;
- Coordinate the assignment of on-site technician(s) to replace failed hardware;
- Assist the configuration, tuning, and proper use of the products supplied by the Hosting Entity engineers, responsible for the system management;
- Assess requests for usage information, component functionality, and documentation related to the supported products (hardware, software).

The Supplier is required to ensure the minimum service levels given below, regardless of whether the service is delivered directly, or provided by the device manufacturer, or another company specializing in this type of service.

- *Call receiving period: weekdays from 09:00 to 18:00 EEST/EEDT;*
- *Response time: within 30 minutes, covering all submitted requests;*

The Supplier must ensure the commencement of the diagnosis of any malfunction reported by EuroHPC JU within 1 working hours from receiving the call, i.e., within 1 hours from receiving the call; EuroHPC JU must be contacted to start the problem determination phases.

5.2.6. Incident Management

Incident management is an important maintenance activity that allows for the investigation of reported incidents as quickly as possible. An incident is an event that significantly affects the availability of the GPU partition for the execution of a task at a given moment. Incidents should be categorized (grouped) based on the order of reporting. For example, problems with four (4) GPU devices, each reported at a different time, are counted as four separate incidents, while problems with two (2) GPU devices, reported at the same time, are counted as one incident.

This activity consists of:

- Administrative follow-up on availability incidents;
- First-level diagnosis of the reported incidents;
- In case of incidents requiring a second-level diagnosis, passing the incident on to the onsite support teams;
- Carrying out tests before putting elements into production.

Furthermore, the Supplier will have to set up and maintain an on-site tool to monitor interventions that accompany the incident solution, and specify the operations performed during each intervention (date and time of receiving the call, response time, duration of operations, operations carried out, replaced components, etc.).

5.2.6.1. Technical resolution engagement

The methods for monitoring the commitments to resolve incidents are presented below.

The different times of the incident are defined below:

- T0: the time at which the incident occurred;
- T1: the time at which the Supplier is notified;

- T2: the time at which the Supplier begins to work on the incident;
- Tq: the time at which the incident is qualified;
- T3: the time at which the incident is resolved or bypassed;
- T4: the time at which equipment affected by the incident is returned to production.

For incidents affecting the availability of the system, the Supplier will commit:

- To consider the incident in less than one hour after being informed (T2 - T1);
- To qualify the incident in less than 1 hours after the start of the intervention (Tq - T2);
- If the incident is the responsibility of the Supplier (hardware or software provided by the Supplier), they will resolve or bypass the incident in less than 2 hours (T3 - Tq);
- If the incident is not the Supplier's responsibility, they will apply any existing procedures provided by Sofia Petascale Supercomputing centre or they will escalate the problem to Sofia Petascale Supercomputing centre.

If the Supplier fails to comply with their commitments to resolve incidents, a global status will be presented at the site meeting and penalties might be imposed.

The Supplier will commit a maximum number of impacting incidents under their responsibility. During the first stage of system installation, this maximum number of incidents will be specified in the progress contract, beginning with a defined number of incidents and reaching a final number of incidents at the end of the progress contract.

For the entire system, this threshold will be set to 2 incidents per month, at the start of the progress contract, and might decrease up to 50 incidents per month, depending on the development of the situation.

Period	**Number of Incident**	**Measurement period**
T0 to T0+1 month	1	Monthly
T0+1 month to T0+2 months	2	Monthly
T0+2 month to T0+3 months	3	Monthly
After T0+3 months	4	Monthly

Table 10 Maximum number of incidents by period

5.2.7. Relations with the Supplier

The Supplier, in agreement with the EuroHPC JU, must nominate a representative to manage all relations with EuroHPC JU. The representative of the supplier must have the appropriate professional qualifications, must be named before signing the supply contract, and must carry out his duty during the execution of the supply contract. The primary responsibility of this individual is to coordinate the contacts with EuroHPC JU regarding issues that may not be resolved through communication with the Supplier, such as disputes involving the Supplier's sales manager, technician, or call centre. All necessary documentation for gaining access and using the maintenance service system (access credentials, etc.) will be provided by that representative. Another role of the Supplier's representative is

to attend regular meetings and provide updates on the progress of the supply project, and to handle any changes that need to be made to meet the terms set in the contract.

Upon the completion of his duties based on the supply contract, the Supplier's representative will attend especially arranged meetings to:

- Present new products and services, and/or provide information about feature updates to existing features available for the provided services;
- Analyse invoices and summary documents.

5.2.8. Operational Service Quality

For incidents that do not affect availability (for example redundant equipment or equipment with high availability (HA) capabilities provided by the Supplier), the Supplier must repair or bypass the incident within 2 working days. After this period, the equipment concerned will be considered unavailable. Afterwards, the incident will be considered for calculation of unavailability and for follow-up with a commitment to resolution.

Benchmarks (used for provisional acceptance) will be run regularly on systems configured as “in production” (systems configured with parameters and tools of the Sofia Petascale Supercomputing centre), particularly during maintenance.

The values obtained will serve as production reference values following a joint commitment between EuroHPC JU and the Supplier. These reference benchmarks will be reviewed regularly, in particular after each evolution of configurations of the system, and the results will be compared with the reference values preceding the evolution.

In the event of system regression (decrease in performance, malfunction of a code, or similar) or in the event of differences between the production and reference values, a precise analysis will be carried out jointly by the Sofia Petascale Supercomputing and Supplier.

If the identified problem is the responsibility of the Supplier, a hardware, firmware, or software problem will be opened and progress in solving the problem will be monitored during specific meetings.

5.2.9. Technical and administrative accountancy

Periodic maintenance and maintenance activities must be reported, including:

- Ticket lists issued by the call centre, including the relevant details;
- List of technical assistance interventions detailing the activities carried out and the total duration of the disruption;
- Reports of possible preventive maintenance interventions;
- Analysis of repeated failures;
- Conformance ratios to SLAs.

5.3. Training and knowledge transfer

Training and knowledge transfer from the Supplier to the Sofia Petascale Supercomputing centre is a key feature to fully understand the entire system at the beginning and its innovative capabilities. This knowledge leads to efficient administration and thus reduces the downtime of the system.

All Hosting Entities agree on a set of common requirements for “Training and knowledge transfer”:

- Documentation must be provided at the latest when deployment of the system starts, must
- cover in particular the innovative capabilities and must be updated on a regular basis over
- the system lifetime;
- Training targeting administrators must be provided;
- Training targeting users of the deployed solution must be provided.

5.3.1. Documentation

As part of the project installation, the Supplier must provide documentation describing the design and the full project installation log, explaining the different design decisions taken during the installation phase.

A physical layout must be provided for each of the racks, along with the cables that will be connected to each rack. Physical maps of all the networks of the solution must be provided, indicating clearly what is connected in each of the ports of the networks.

The documentation also must include all the operational procedures that need to be performed to keep the infrastructure working optimally.

5.4. Risk Management

The Tenderer shall include in the “supply and installation project”:

- A list of risks that could negatively affect the installation and early operation of the procured system. For each of the foreseen risk, the supplier shall give an indication of likeliness, and provide a description of the expected impact and a proposal for the measures to undertake to mitigate the risk;
- A description of the roles and responsibilities of all involved parties during the system installation in terms of a RACI (Responsible, Accountable, Consulted, Informed) – model.

5.5. Dismantling of the Supercomputer

The Supplier will provide the procedure for dismantling the delivered systems.

6. TRAINING AND KNOWLEDGE TRANSFER

The training and knowledge transfer requirements and TCs are defined in accordance with Section 4.3 Training and knowledge transfer of the Common Technical Specifications.

6.1. Documentation

Req.No.	Priority	Description
L1-DOC1	*Very High*	*System documentation: As part of the system delivery, the Supplier will provide full system documentation. In particular, the Supplier will provide the following documents: user manuals for all key software and hardware components, documentation of the configuration, the support workflow and a system-specific manual for the system administrators and operators. The Supplier commits to maintain and update the documents during the support timeframe.*

6.2. Training

Req.No.	Priority	Description
L1-TRA1	*Very High*	*Training for administrators: The Tender will include training for the system administrators as required to operate the system and to recover if software failure occurs.*
L1-TRA3	*Very High*	*Training for users: SOFIA PETASCALE SUPERCOMPUTING CENTRE organizes an introductory course for its user community every 3 months. Whenever possible, the Supplier commits to assist SOFIA PETASCALE SUPERCOMPUTING CENTRE with carrying out courses for users in different application areas (i.e., biochemistry, molecular-molecular interactions, computational fluid dynamics, others).*

7. COLLABORATION

SOFIA PETASCALE SUPERCOMPUTING CENTRE, potentially, together with additional suppliers of software components for the procured system, intends to enter an R&D cooperation with the selected Supplier for the development of the system over the lifetime of the system. Refer to Section 10 for additional information regarding the goals of the collaboration.

8. DEFINITIONS

Batch System	Software component responsible for the management and the scheduling of resources (Nodes) and interactive or batch jobs.
Tenderers	Economic Operators participating in the first stage of this Invitation to Tender by submitting a Request to Participate.

Change request	A request to change some aspect of the agreed baseline of a project (i.e., scope, requirements, deliverables, resources, costs, time frame or quality characteristics). It must be made in writing.
Commission Working Days EuroHPC Working Days	Monday to Friday, excluding European Union public holidays. The calendar of the European Union public holidays is published every year on Eurlex (e.g. for 2018 https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017D0217(01)&from=EN.
Compute Node	A Node dedicated to compute workloads. Compute Nodes are typically managed by the Workload Manager.
Contracting Authority	The European High Performance Computing Joint Undertaking- used interchangeably and referred to as the EuroHPC JU.
Contractor	Refers to natural or legal person with whom the CONTRACT has been concluded.
Core	Set of integer and floating calculation units managed by a control unit and capable of executing operations between internal registers and/or external memory. A single Processor may consist of several Cores.
Delivery note	Document listing the details of the delivery, acknowledging the fact that the delivery has taken place but in no way implying conformity with the specifications. It is compulsory for all deliverables.
Descriptive Document	This document, defining the Contracting Authority's needs, requirements and objectives and the exclusion, selection (prequalification) and award criteria for the procurement.
Device	Execution unit that performs specific computational or communication tasks to aid the processor in carrying out the execution of a process. Examples include graphics processor units, cards that offer acceleration for floating point intensive workloads, other forms of co-processors, network interface cards and storage cards.
Existing Services	The current services as they are being delivered under the responsibility of the EuroHPC JU Common Portal Unit and the service providers it may have.
Filesystem	Technology to manage non-volatile storage components by means of a file abstraction. The Filesystem technology must be Lustre parallel file system.
Final Tender	Submission by the Tenderers in response to the Invitation to Tender (ITT) during the Tendering stage.
Force majeure	Any unforeseeable, exceptional situation or event beyond the control of the parties that prevents either of them from fulfilling any of their obligations under the contract. The situation or event must not be attributable to error or negligence on the part of the parties or on the part of the subcontractors and must prove to be inevitable despite their exercising due diligence. Defaults, defects in equipment or material or delays in making them available, labour disputes, strikes and financial difficulties may not be invoked as force majeure, unless they stem directly from a relevant case of force majeure.
Hosting consortium	A group of Participating States that have agreed to contribute to the acquisition and operation of a EuroHPC supercomputer.
Sofia Petascale Supercomputing centre	Legal entity established in a Participating State to the Joint Undertaking that is a Member State which includes facilities to host and operate a EuroHPC supercomputer and which has been selected in accordance with the CEI.
Hosting site	is used to refer to the physical facilities at which Sofia Petascale Supercomputing centre shall host and operate the EuroHPC supercomputer and which is established in a Participating State that is a Member State.

Interconnect	Devices and apparatus that implement a network of Nodes featuring low communication latency and high bandwidth. Typically, all Compute Nodes, Login Nodes and potentially other Nodes are integrated in the Interconnect. The Interconnect hardware is accompanied by appropriate software components to enable message passing between processes on different Nodes. In addition, the Interconnect may integrate storage systems.
Linktest	A parallel ping-pong test between all possible MPI connections of a machine. Output of this program is a full communication matrix which shows the bandwidth and message latency between each processor pair and a report including the minimum bandwidth.
Liquidated Damages	The Euro amount by which the prices will be reduced based on Contractor's failure to achieve the SLRs. The mechanism to calculate all price reductions are specified in the CONTRACT (liquidated damages section).
Logic Node	A Node dedicated for user access, software and data management. The extent to which pre- and post-processing workloads are supported on Login Nodes is site specific.
Management Node	A Node used for system management. A system usually contains one or two Management Nodes.
Mandatory Requirement (MANDATORY)	Mandatory Requirements are considered essential for the procured system and must be fulfilled by all Final Tenders. Mandatory Requirements will be assessed for each Tender submitted. Final Tenders which will not be compliant with all Mandatory Requirements will be rejected.
Measurement interval	The period upon which performance will be calculated. This takes into consideration the impact of continuous outage. For example, a monthly measurement interval for a 99 percent Minimum Performance for a 24x7 system with eight hours of weekly planned downtime would allow 6.4 hours of a continuous outage, with no other outages during the calendar month. A weekly interval would only allow 1.6 hours of a continuous outage.
Measurement period	Any specified calendar period within which the metrics shall be measured and reported on for determining the Contractor’s performance to the SLR as specified with the Annexes 11.
Node	Set of Processors, memory areas and Devices. The Processors of a single Node access a shared memory address space through load/store instructions. Devices may feature a separate address space.
Parallel Filesystem	Filesystem accessible in a shared context through a network (potentially the Interconnect) that ensures global consistency (with specific implementation dependent semantics) of the address space.
Performance Target	The desired level of service the EuroHPC JU is seeking for that particular service level requirement
Processor	Execution unit constituted by one or more Cores and able to execute a portion of computation independently from the other Processors. Typically, a Processor is constituted by a single chip connected to the central memory and other hardware devices of the system via a single Socket.

Project	A temporary organizational structure which is setup to create a unique product or service (output) within certain constraints such as time, cost, and quality.
Regulation	Council Regulation (EU) No 2018/1488 of 28 September 2018 establishing the European High Performance Computing Joint Undertaking, and the statutes of the Euro HPC Joint Undertaking ('Statutes') annexed thereto.
Reporting period	The time span between two successive regular performance reporting.
Resource Management System	Software component responsible for the launch, execution and teardown of batch jobs on Nodes.
Service Level Agreement (SLA)	An agreement, between the Contractor and the EuroHPC JU, that describes the service, documents service level targets, and specifies the responsibilities of the Contractor and the EuroHPC JU.
Service Level Requirement (SLR)	The SLR documents the requirements for a service from the EuroHPC JU's viewpoint, defining the detailed service measures, performance targets, formula, measurement intervals and reporting periods.
Service Node	A Node used for running specific system services. A supercomputer may constitute many Service Nodes.
Services	The services that shall have to be delivered by the Service Provider as described in the Descriptive Document and its technical annexes.
Site visits	A visit to the key delivery centres of the Tenderers to better understand and further refine the solution proposed.
Socket	Connector used to interface a Processor with a motherboard.
Solution	The set of processes through which each Tenderer means to meet the needs and objectives of the EuroHPC JU.
Swap	Space on disk (or comparable non-volatile storage components) used by the Operating System for memory paging.
Target Capabilities	Target Capabilities are desirable features and desirable performance levels for the procured system. In contrast to Mandatory Requirements, failure to provide Target Capabilities will not lead to the rejection of the Final Tenders provided by the Tenderer. Tenders that provide the Target Capabilities will receive a higher score. Target Capabilities are prioritized. Level-one priority Target Capabilities (TC-1) are considered of higher importance than level-two Target Capabilities (TC-2).
Tender Specifications	Document defining the Contracting Authority's detailed needs and objectives (for which solutions can be provided by the tenders) and the award criteria for the award of the contract
Tiered Storage Solution	Storage solution based on different storage technologies, which are presented as a unique file namespace. The system provides an automatic procedure of data migration across different tiers (types) of storage devices and media.
Transition	Refers to the period during which the services are in the process of handover or takeover, whichever is applicable.
Weighting Factor	Numerically weighted values assigned for failure to meet specific SLR targets that are tracked and aggregated for the purpose of calculating and determining the Liquidated Damages. Weighting factors for SLA are included in the SLA tables within each relevant annex 11.

Working Days	Monday to Friday, excluding 1st January, Easter Monday, Ascension Thursday, Whit Monday, 1st November and 25 December.
Working languages	English
Workload manager	Software component consisting of the combination of a Batch System and the Resource Management System

8.1. Units of Measurement

Regarding units for memory and storage capacities, the following applies:

Unless stated otherwise, SI units (rather than ISO/IEC 80000 prefixes) are used in the technical specifications and should be used for the Tender. For example:

1 KB = 1000 bytes, 1 MB = 1000 KB, 1 GB = 1000 MB, 1 TB = 1000 GB, 1 PB = 1000 TB

The Tender should preferably exclusively use SI prefixes. Where this is not possible, the use of IEC (binary) prefixes must be made clearly visible.

The compute performance of a system may be assessed using the following unit:

1 KFlop/s = 1000 floating point operations per second

1 MFlop/s = 1000 KFlop/s

1 GFlop/s = 1000 MFlop/s

1 TFlop/s = 1000 GFlop/s

1 PFlop/s = 1000 TFlop/s

8.2. Glossary

Abbreviation	Description
24x7	24 hours a day, 7 days a week
AI	Artificial Intelligence
BS	Batch System
CAPEX	Capital Expenditure
CEI	Call for Expression of Interest
CN	Compute Node
DMZ	De-Militarized Zone (network)
ETS	Energy-to-Solution
FA	Final Acceptance

HDD	Hard-Disk Drive
HPC	High Performance Computing
HPDA	High Performance Data Analytics
IOPS	Input/Output Operations Per Second
ISA	Instruction Set Architecture
LN	Login Node
MN	Management Node
MPI	Message Passing Interface
NVM	Non-Volatile Memory
NWH	Non-Working Hours
OPEX	Operational Expenditure
OS	Operating System
PA	Provisional Acceptance
POSIX	Portable Operating System Interface
PPQ	Pre-Production Qualification
PRACE	Partnership for Advanced Computing in Europe
PUE	Power Usage Effectiveness
RMS	Resource Management System
RtP	Request to Participate
SN	Service Node
SSD	Solid-State Drive
TCO	Total Cost of Ownership
TEPS	Traversed Edges per Second
TTS	Time-to-Solution
VM	Virtual Machine
VN	Visualization Node
WH	Working Hours
WLM	Workload Manager

[1] Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)