Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

CINECA

G00843 - Capitolato Tecnico

Tier1 Technical Specifications

FORNITURA HARDWARE TIER-1 PER IL TECNOPOLO DI BOLOGNA E RELATIVI SOFTWARE E SERVIZI CONNESSI

Cig B137697507

CUP CUP: D56G22000380006 (CINECA SPOKE0 CNHPC); J33C22001240001 (UNIBO ECOSISTER); J33C22002830006 (UNIBO FAIR); J33C22002920006 (UNIBO HEAL-ITALIA); J33C22002810001 (UNIBO SERICS); J33C22002880001 (UNIBO RESTART); B53C22006450001 (UNIBO PNC - DARE); C55F21002880001 (INAF)

The following specifications result from the work done
in collaboration with AIM, CNR, INAF, UNIBO and ICSC

Summary

Summary **3**
1. Project goals **5**
1.1 Introduction 5
1.2 Goal of the procurement 5
2. Document definitions and glossary **7**
2.1 Tendering procedure definitions 7
2.1.1 Definitions 7
2.1.2 Categories of requirements 7
2.2 Glossary 7
2.3 Unit of measure 9
3. Site context **11**
3.1 CINECA hosting entity 11
3.2 Implementation 11
3.2.1 Procedure 11
3.2.2 Time schedule 11
3.3 CINECA data centre 11
3.3.1 Facility description 11
3.3.2 Data centre specifications 12
3.3.3 Electrical infrastructure 13
3.3.4 Cooling infrastructure 13
3.3.5 Data Hall MEP layout 15
4. Technical specifications **18**
4.1 General requirements 18
4.1.1 Functional aspects 19
4.2 Interconnects 20
4.2.1 Compute Fabric 20
4.2.2 Core Ethernet Network (CEN) 21
4.2.3 Management network 23
4.3 Compute partition 25
4.3.1 CPU partition 25
4.3.2 GPU partition 26
4.4 Management partition 28
4.5 Front-end partition 31
4.5.1 Login partition 31
4.5.2 Visualization partition 32
4.6 Storage infrastructure 33
4.6.1 Data Movers 34

4.6.2 Tier-1 Storage – General requirements ... 35
4.6.3 Tier-1 AIM storage ... 37
4.6.4 Tier-1 CNR – INAF – UNIBO storage ... 39
4.6.5 Tier-3 SKA storage ... 40
4.7 Facility integration ... 42
4.8 System software and monitoring ... 44
5 Benchmarks ... 47
5.1 Introduction ... 47
5.2 Benchmark framework ... 47
5.2.1 Software metrics ... 47
5.2.2 Benchmark categories ... 47
5.2.3 Benchmark suite ... 47
5.3 Benchmark procedure ... 48
5.3.1 Benchmark rules ... 48
5.4 Benchmark execution ... 49
5.4.1 Retrieve and compilations of the codes ... 49
5.4.2 Input parameters ... 49
5.4.3 Execution ... 49
5.5 Benchmark analysis report ... 53
6 Maintenance and infrastructure availability ... 54
6.1 Maintenance and support requirements ... 54
6.2 Tier-1 Specialistic support ... 57
6.3 Licenses ... 57
6.4 Infrastructure availability ... 57
7 Installation and acceptance ... 59
7.1 Installation time schedule and project management ... 59
7.1.1 System Installation ... 59
7.1.2 Supply and installation project ... 60
7.2 Acceptance procedure ... 61
7.2.1 Documentation requirement ... 61
7.2.2 Execution of acceptance tests ... 61
7.2.3 Provisional acceptance tests ... 62
7.2.4 Pre-production qualification ... 64
7.2.5 Final acceptance ... 64

1. Project goals

1.1 Introduction

To provide the most competitive computing ecosystem to Italian and European researchers, CINECA is adopting a three-pillar strategy to extend the capabilities and performance of its computing and storage infrastructure.

Tier-1 systems are infrastructure that target the specificities of user communities and provide computing and storage services tailored for their needs.

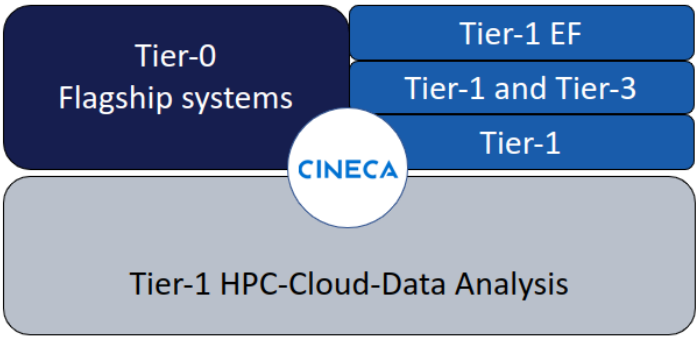

Figure 1. High level scheme of the HPC ecosystem

In doing so, CINECA can now leverage the availability of two new data centres: the new Big Data Technopole data centre currently hosting Leonardo, and a new data centre in Naples in the renovated area of San Giovanni a Teduccio. The geo-distribution of the infrastructure allows for improving the resiliency of the compute and storage services, providing the capability to host critical workloads. The evolution of the infrastructure will benefit from multiple projects and actions involving the likes of the recently founded national centre ICSC, established Italian research institutions such as CNR, INAF, OGS, INGV, and national agencies such as ACN (National Cybersecurity Authority) and AIM (Agenzia Italia Meteo). All of them are HPC stakeholders, relevant at national and international levels, bringing significant expertise in the field and providing new requirements for establishing a modern and competitive national computing and storage infrastructure. The infrastructure evolution investments amount to more than 125 million euros.

Moreover, the new data centre Big Data Technopole offers a strategic location for national infrastructures that will be dedicated to the Square Kilometre Array (SKA) and the weather forecast by the means of the Agenzia Italia Meteo (AIM).

1.2 Goal of the procurement

This project aims to procure an HPC infrastructure able to implement the following main characteristics:

- Provide a Tier-1 class computing HPC system, including a conventional (CPU) and accelerated (GPU) partition.
- Provide state-of-the-art parallel & multiprotocol storage infrastructures, able to serve all the Tier-1 partitions. It is paramount that weather forecast simulations will run with adequate quality-of-service to comply with the service level agreements imposed by this critical national service.
- Provide a Tier-3 class system, dedicated to process, handle, host and preserve hot and cold data produced within the SKA project. This system will include:

- An online multiprotocol data storage.
- A long-term data storage.

This procurement is a result of a strong collaboration between AIM, CNR, INAF, UNIBO and CINECA.

The procured resources, integrated in the existing CINECA data centre, will represent a new step forward to address the new challenges in multiple scientific fields such as weather forecast, material design, and astrophysics.

The infrastructure will play a significant role to store observational data produced via the SKA project[1]. To note that the SKA project aims to collect a huge amount of data - in the orders of hundreds of PByte - from the array of radio telescopes involved worldwide.

Moreover, paramount for the success of the procurement project is to have the procured infrastructure operational as soon as possible, in order to provide a replacement for the current resources used by AIM for the weather simulations.

[1] http://italy.skatelescope.org/

2. Document definitions and glossary

2.1 Tendering procedure definitions

2.1.1 Definitions

Term	Description
Candidate	The qualified economic operator eligible to contract
Supplier	The tenderer who is awarded the contract as part of this procurement.
Offer	The final bid submitted by the Tenderer

Table 1: Procurement procedure definitions

2.1.2 Categories of requirements

The requirements and features within the documents are categorised as follows.

Requirements priority	Requirements category	Description
Mandatory	**MRQ**	*Mandatory Requirements* These specifications are considered essential for the procured infrastructure and must be fulfilled by all best and final Offers. Mandatory Requirements will be assessed for each offers submitted.
Targeted	**TRQ**	*Highly targeted requirements* These are highly desired specifications for the procured system. In contrast to Mandatory Requirements, failure to provide targeted requirements will not lead to the rejection of the best and final Offer provided by the Candidate.
Mandatory	**DCS**	*Data Centre Specifications* The Offer must comply with the detailed data centre specifications. However, while complying with the requirements' framework, the Candidate is allowed to propose alternatives at its own cost, in order to provide the adequate data centre integration of the offered solution into CINECA infrastructure.

Table 2: Categories of requirements

2.2 Glossary

Term	Description
AIM	Agenzia Italia Meteo
CNR	Consiglio Nazionale delle Ricerche
INAF	Istituto Nazionale di Astrofisica
UNIBO	Università di Bologna
SKA	Square Kilometre Array
Backbone	Site-wide Ethernet network (40GE or 100GE)
CINECA	Interuniversity Consortium
DDR	Double Data Rate

DIMM	Dual In-line Memory Module
HPL	High Performance Linpack (see top500.org)
GPU or GPGPU	Graphic Processing Unit or General-Purpose Graphics Processing Units usable for computation
HA	High-Availability. Mechanism to ensure service availability in case one of a component failure
HPC	High-Performance Computing
LACP	Link Aggregation Control Protocol
NVM	Non-volatile memory
PCIe	Peripheral Component Interconnect Express
PDU	Power Distribution Unit
POSIX	Portable Operating System Interface for Unix
RAID	Redundant Array of Inexpensive Disks. Mechanism to prevent from disk failures by storing redundant information on additional disks (mirror, parity…)
SDRAM	Synchronous Dynamic Random Access Memory
SR-IOV	Single-root input/output virtualization
UPS	Uninterruptible Power Supply
VM	Virtual machine
RHEL	Red Hat Enterprise Linux
CDU	Cooling Distribution Unit
MN	Master nodes
SN	Service Nodes
CN	Compute Nodes, either being CPU- or GPU-based nodes
VN	Visualization Nodes
LN	Login Nodes
TOR	Top Of Rack
CEN	Core Ethernet Network
LTS	Long-Term data Storage
HDD	Hard Disk
CPU	Central Processing Unit

Table 3: List of acronyms and common terms

Concept	Definition
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.
Socket	Connector used to interface a Processor with a motherboard.
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.
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 intense workloads, other forms of co-processors, network interface cards and storage cards.
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.

Compute Node	A Node dedicated to compute workloads. Nodes of CPU and GPU partition are considered Compute Nodes. Those Nodes are typically managed by the Workload Manager.
CPU Node	A Compute Node that is part of the CPU partition based only on CPUs for data processing.
GPU Node	A Compute Node that is part of the GPU partition based with CPUs and GPUs for data processing.
Front-end Node	A Node dedicated for user access, software and data management. Login Nodes and Visualization Node are considered Front-end Nodes.
Login Node	A Node used by users to submit jobs in the System and to compile applications.
Visualization Node	A Node designed and used specifically for visualization workloads.
Management Node	A Node used for system management. Nodes of Service and Master partition are considered Management Nodes.
Service Node	A Node used for running specific system services (e.g., workload scheduler). A supercomputer may need many Service Nodes.
Master Node	A node used to System Administrator to manage the System. On these nodes are contained tools and applications used to manage the System.
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
Filesystem	Technology to manage non-volatile storage components by means of a file abstraction. The file-system technology may be compliant with (official) standards such as POSIX. Examples include XFS, Ext4, IBM Spectrum Scale, Lustre, NFS and pNFS.
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.
Swap	Space on disk (or comparable non-volatile storage components) used by the Operating System for memory paging.
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.
Batch System	Software component responsible for the management and the scheduling of resources (Nodes) and interactive or batch jobs.
Resource Management System	Software component responsible for the launch, execution and teardown of batch jobs on Nodes.
Workload Manager	Software component consisting of the combination of a Batch System and the Resource Management System

Table 4: List of technical definitions

2.3 Unit of measure

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 Proposal. For example:

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

The Proposal 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

3. Site context

3.1 CINECA hosting entity

CINECA – founded in 1969 – is a not-for-profit Consortium, made up of 117 members: the Italian Ministry of Education, the Italian Ministry of Universities and Research, 70 Italian universities and 45 Italian National Institutions. It is the largest Italian computing centre and one of the most important worldwide. With more than nine hundred employees, it operates in the high performance computing (HPC), technology transfer and information technology (IT) sectors. CINECA develops advanced IT applications and services with the main goal of supporting academia, public administration, and private companies.

At national and European level CINECA is expected to play a role as advanced research infrastructure provider, bringing its standout experience and expertise in HPC and the ability to support the actions of the center. In fact, CINECA offers state-of-the-art hardware resources and highly qualified personnel, and is committed to accelerate scientific discovery by continuously evolving its computing, data management and data analysis infrastructure and services. CINECA's HPC infrastructure and expertise support research across all domains, helping in tackle scientific and societal challenges in weather and climate forecasts, computational fluid dynamics, computational bioinformatics, genomics and so on.

CINECA has a proven track record of providing HPC systems at the top of the most powerful computing systems in the world - and three times in the top 10 - as ranked by the top500.org list. CINECA hosts and manages Leonardo, the fourth supercomputing system in the current top500 ranking. The HPC department works for the management, support, and exploitation of the HPC infrastructure, providing services to address computing research needs.

3.2 Implementation

3.2.1 Procedure

For this procurement, CINECA - as the procurer - and the involved partners elected to use a public open procedure. The goal of this document is to provide the requirements the supplier is requested to satisfy with the offered solution.

3.2.2 Time schedule

The implementation of the procured infrastructure will proceed according to the timeline defined in art. 4 of the "Schema di contratto" which is part of the tendering documents package.

3.3 CINECA data centre

The new equipment will be hosted in CINECA data centre located in DC Tecnopolo: Via Stalingrado, 86, 40128 Bologna (BO), Italy. To set the logistic and data centre integration limits that the offers must comply, in the following Sections the data centre specifications and architectural design are reported.

3.3.1 Facility description

Figure 2 shows what is currently deployed in the Big Data Technopole (Tecnopolo) data centre.

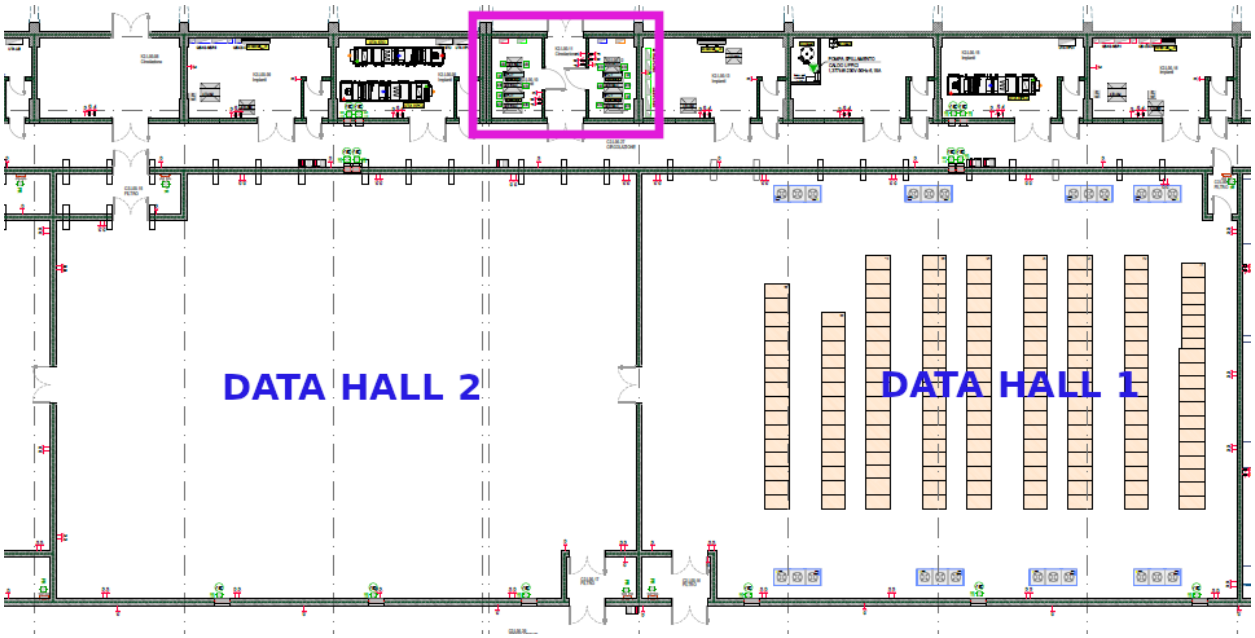

Figure 2. Technopole data centre floor plan. Leonardo layout is depicted in Data Hall 1 (on the right) and the procured System will be hosted in Data Hall 2 (on the left). The network room is indicated by a purple box at the top of the figure.

3.3.2 Data centre specifications

The floor space planimetry of the Data Hall 2 is reported in Figure 3. The Data Hall is empty. The tile size is 600x600 mm. In the orange rectangle it shown where is the expected location of the procured solution.

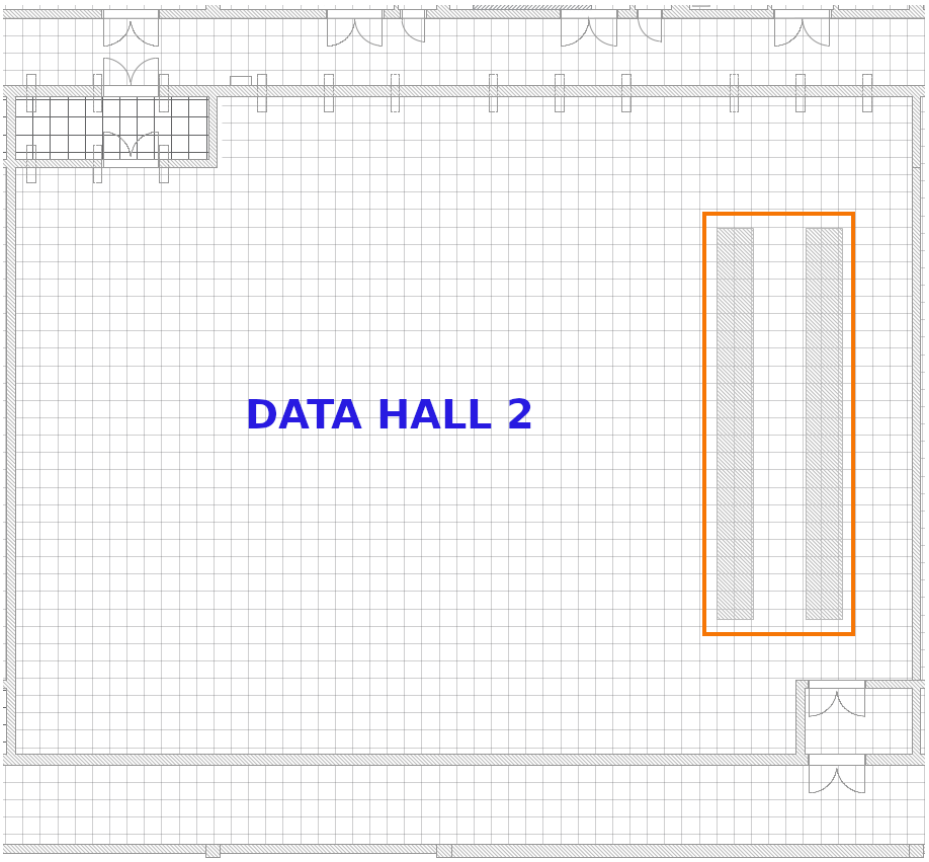

Figure 3. Floor plan of Data Hall 2 that will host the computing partitions of the infrastructure. The Data Hall is empty. The grey rectangles provide just an indication for the whitespace available. Busbars and cooling distribution allow - for the moment - to have racks just in the orange rectangle. Tile is 60x60 cm.

For the Data Hall 2 the following specifications apply:

Req.	Description	Category
3.3.2-1	*Dedicated data hall* The offered solution infrastructure must be installed the Data Hall 2. In terms of whitespace available, the Data Hall is empty with no pillars. Surface available for the offered solution is in the order of 120 m^2 considering the availability of the power and cooling distribution in the right partition of the data hall.	DCS
3.3.2-2	*Raised floor details* The data hall is equipped with a raised floor with height of 100 cm.	DCS
3.3.2-3	*Raised floor load* The raised floor will be reinforced to host DLC racks. The maximum expected load is 30 kN/m^2 and 11 kN single point load.	DCS
3.3.2-4	*Rack maximum height* The cable tray is at 240 cm of height.	DCS
3.3.2-5	*Room humidity* The offered infrastructure must comply with a relative humidity in the interval 20-60%.	DCS

3.3.3 Electrical infrastructure

Req.	Description	Category
3.3.3-1	*Data hall power supply* The Data Hall 2 has an electrical power supply with a frequency of 50 Hz, 3 x 400 Vac between the power lines, 230 Vac between the line and neutral. The hall can supply a power dedicated to the procured HPC infrastructure up to a maximum of 2.000 kW IT during the acceptance phase testing (HPL & HPCG), and 1.600 kW IT in operating conditions, to be protected, if necessary, by means of a power capping solution available on the procured infrastructure.	DCS
3.3.3-2	*Electric load layout* The IT load installed in the Data Hall 2 is available through 2 rack rows. Every single rack will be connected to 3 dedicated busbars installed above the racks. Upon request of the procurer (CINECA), the computing racks will be connected to the power supplies directly with cables, i.e., without the use of plugs and without compromising the guarantee of the offer's components. The total power distributed to the racks cannot in any case exceed the limits defined in Req. 3.3.3-1.	DCS

3.3.4 Cooling infrastructure

The cooling of the procured infrastructure will rely on the cooling infrastructure already in use for Leonardo. Any change in the cooling water temperature set points will impact Leonardo operations –

including its operating costs - and should be reduced to the minimum.

Req.	Description	Category
3.3.4-1	*Cooling infrastructure* The cooling infrastructure of the data centre produces tempered water and chilled water. The tempered water is dedicated to the DLC compute nodes and the chilled water is used for the air conditioning of the data halls. The procured infrastructure needs to comply with this cooling infrastructure requirement and the limits reported in reqq. 3.3.4-2 and 3.3.4-3.	DCS
3.3.4-2	*Liquid cooling* Tempered water will be used for direct cooling of the system. The tempered water circuit is at 36°C inlet - as it is for Leonardo system. It's possible to decrease the tempered water circuit down to 32°C. The tempered water circuit has a cooling capacity of 2.000 kWf dedicated for the procured infrastructure.	DCS
3.3.4-3	*Air cooling* The air-cooling system is based on 2 CRAH. Each device has a cooling power of roughly of 125 kWf (250kWf aggregated) for the procured System.	DCS
3.3.4-4	*Flow rate* The maximum flow rate available for Data Hall 2 is 2100 l/min per row during acceptance tests and 1700 l/min per row during operations.	DCS

3.3.5 Data Hall MEP layout

Power and mechanical distribution and their arrangement in the Data Hall 2 are reported in the figures below.

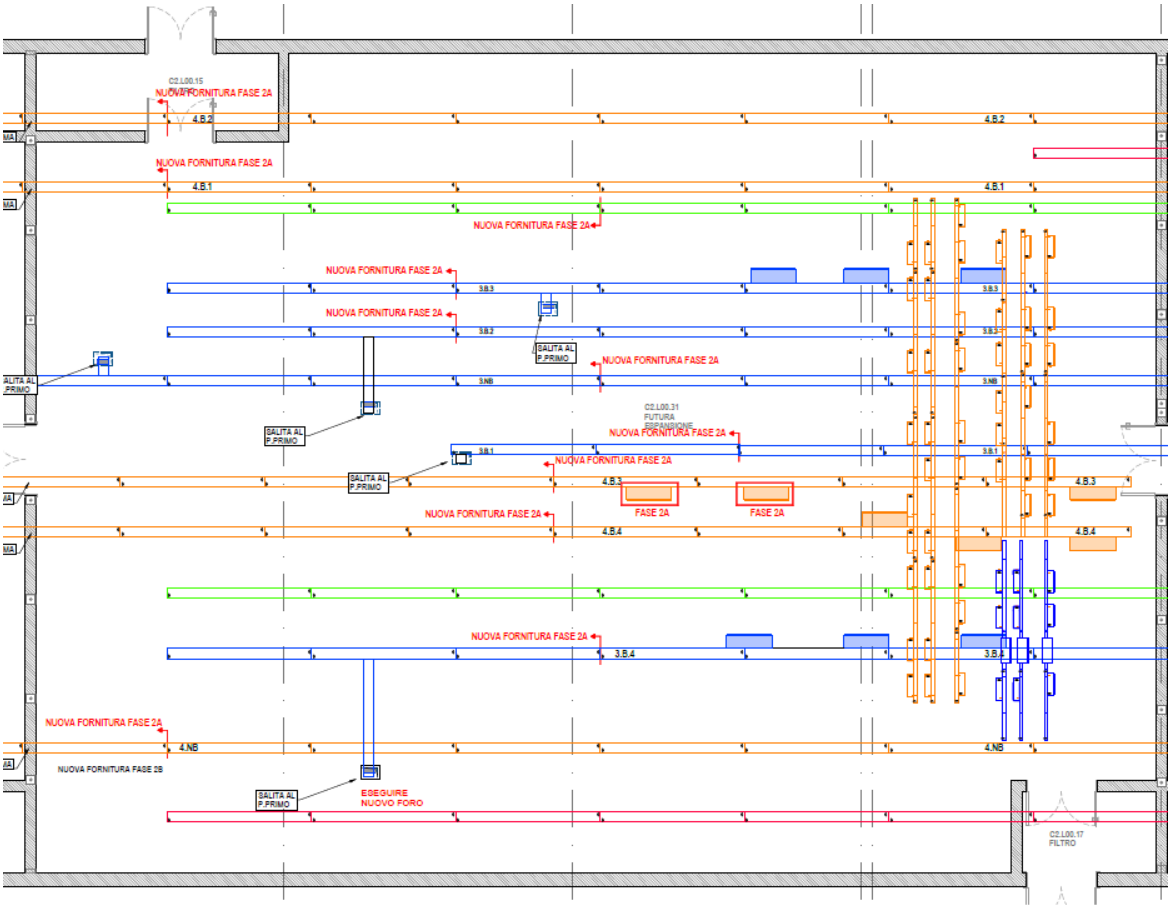

Figure 4. Power distribution layout of Data Hall 2. Each colour represents one of 4 electric branches.

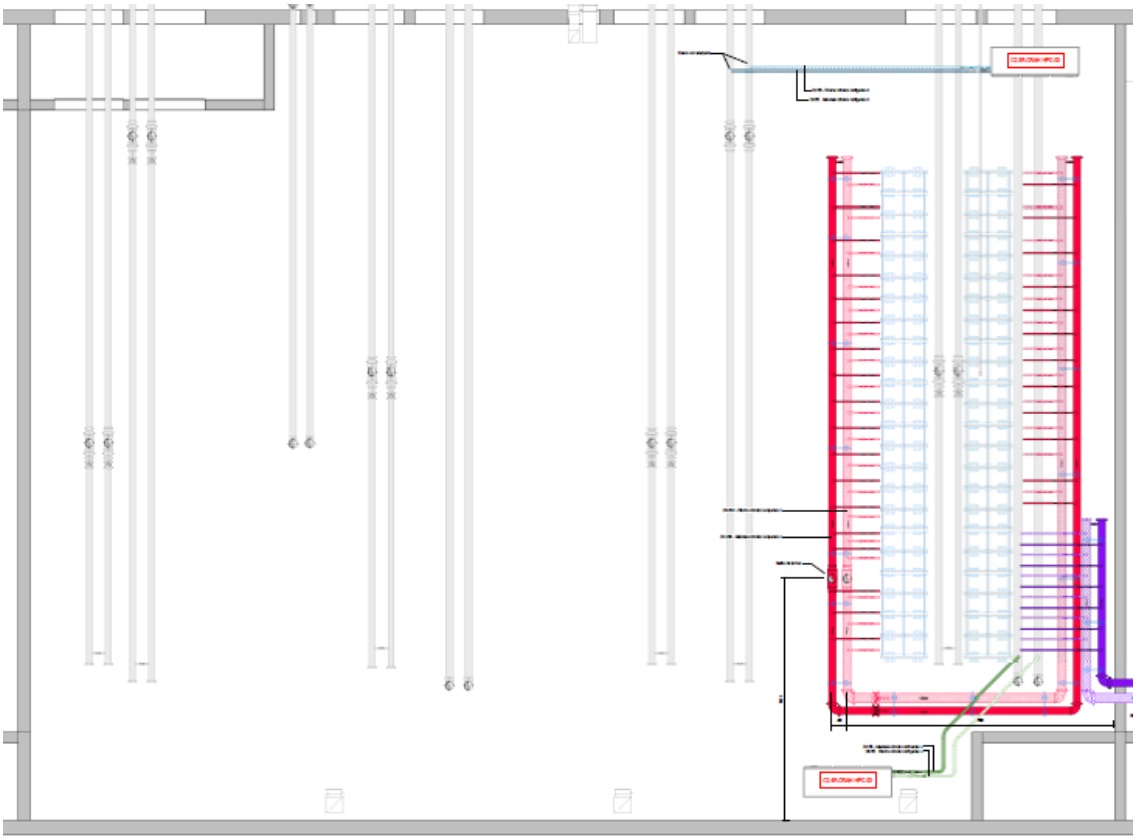

Figure 5. Cooling distribution layout of data Hall 2.

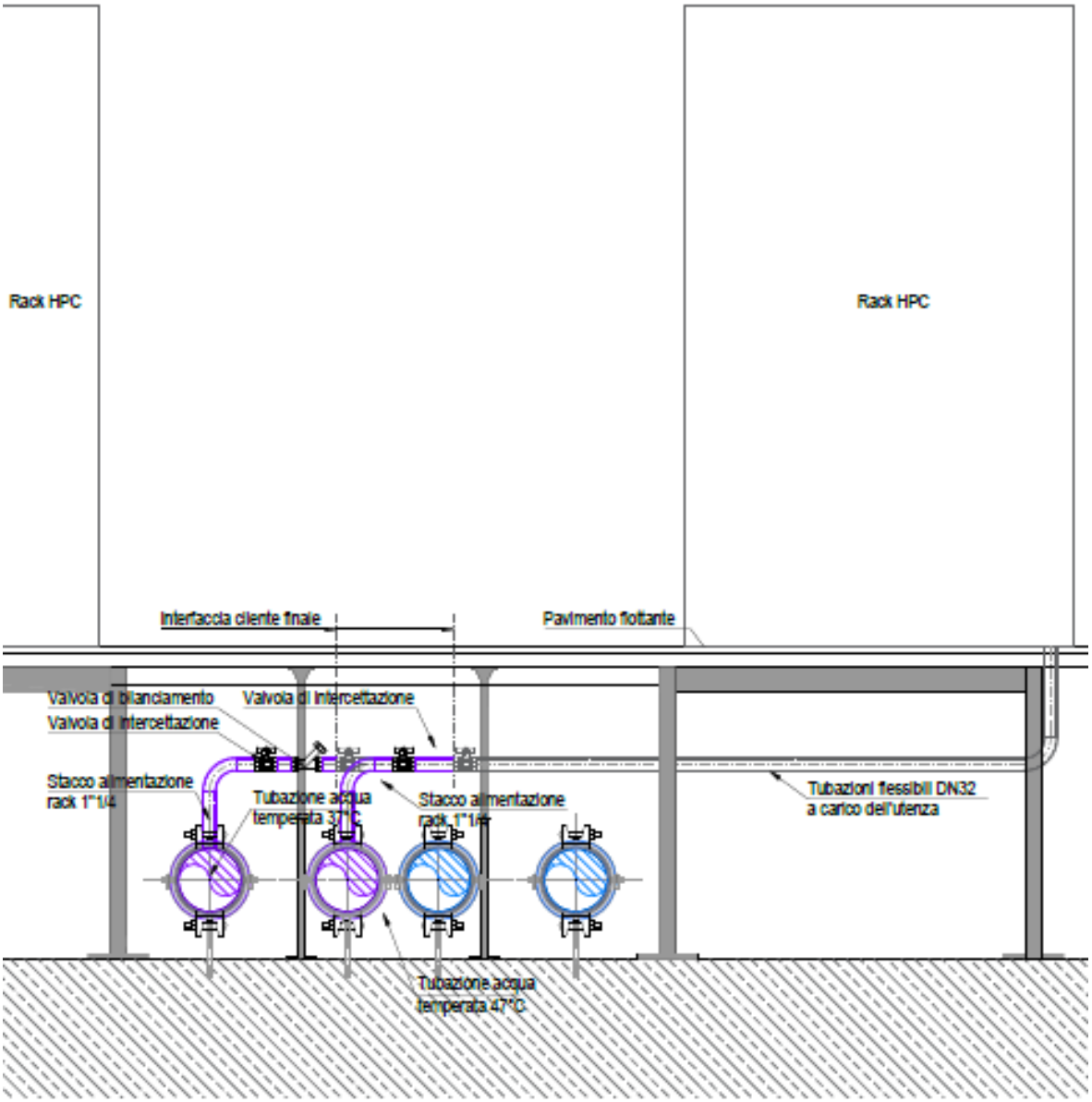

Figure 6. Section of the cooling distribution under the raised floor.

4. Technical specifications

4.1 General requirements

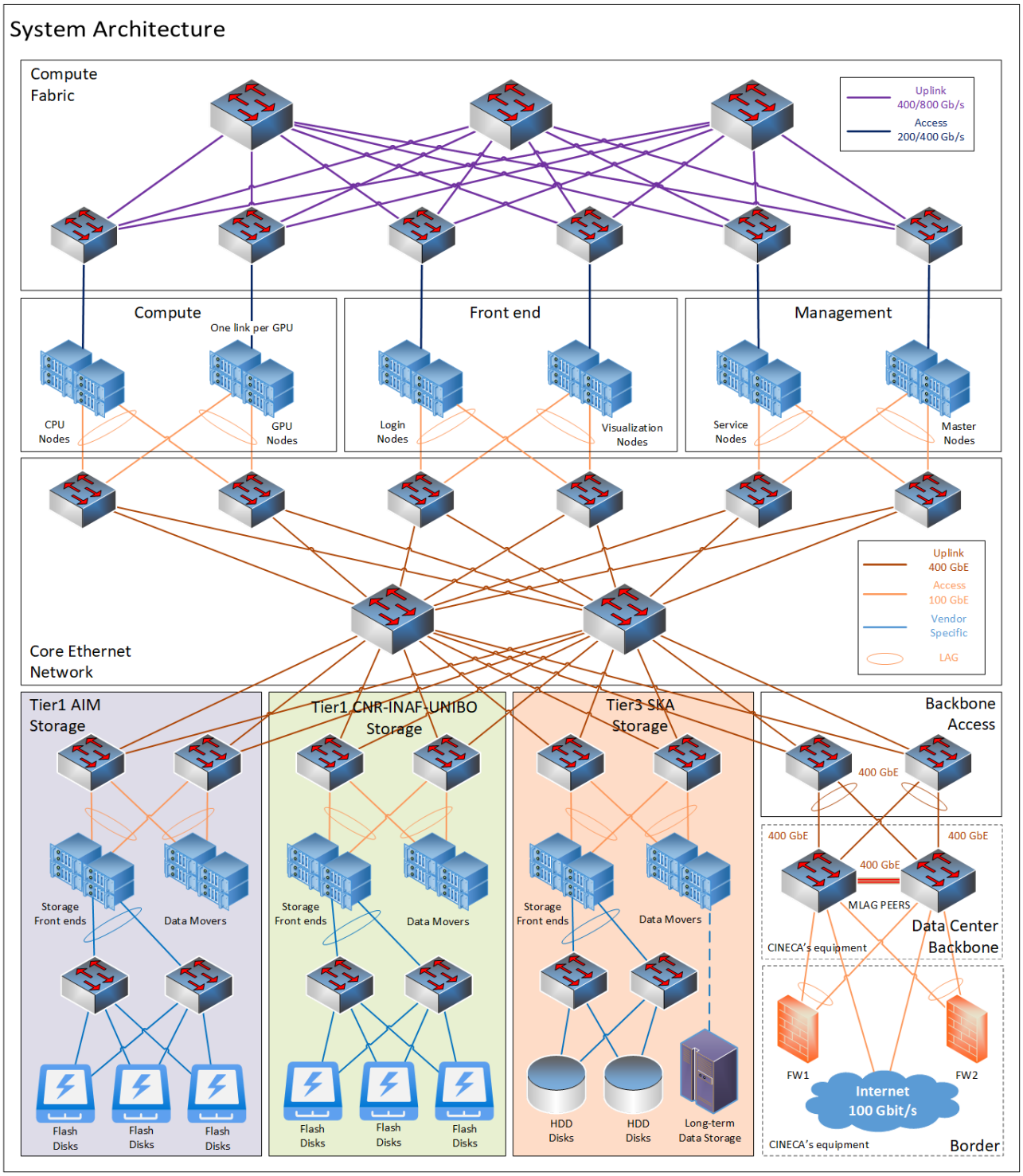

Figure 7: Reference design of the system architecture.

4.1.1 Functional aspects

Req.	Description	Category
4.1.1-1	*Integrated platform* The procured infrastructure is an integrated platform. All hardware and software components required to deliver service to users and manage the system must be included in the Offer.	MRQ
4.1.1-2	*Reboot time* Each partition must be fully rebooted in less than 60 minutes.	MRQ

4.1.1-3	*Common node features*
	
	The nodes must be equipped with:
	• Board Management Control (BMC) with the following features:
	  ○ Dedicated or shared Ethernet network port.
	  ○ Remote management protocols such as: VNC, Java & HTML5 GUI.
	  ○ Virtual console & VMedia functionalities.
	  ○ Scheduled automatic BIOS and internal component firmware updates.
	  ○ Server configuration and firmware lock-down functionalities.
	  ○ Digitally signed firmware updates.
	  ○ Firmware rollback capabilities.
	  ○ Protection features for firmware updates of internal components.
	  ○ Secure default password functionality.
	  ○ Secure erasure of all internal storage devices in the server (ISE).
	  ○ LDAP authentication support.
	  ○ IP blocking functionality.
	  ○ Agentless telemetry for system hardware components, power consumption, and temperatures.
	  ○ Air flow management functionality.
	• Diagnostic tool support:
	  ○ Support for detecting pre-failure events related to disk drives, RAM memory, power supplies, and fans. The diagnostic tools must be hardware and firmware-based and independent of the operating system.
	• Firmware upgrade support:
	  ○ Management system capable of automatically preparing a "service pack" with all the latest firmware for the machine (BIOS) and internal components (such as management card, LAN card, etc.) by directly connecting to the repositories provided by the vendor, without the need for specific Operating System agents. This allows the operator to select the relevant updates and independently proceed with the upgrade of the desired components.
	• Remote Monitoring and Alert Functionality:
	  ○ System capable of automatically sending an alert to the support, containing all the relevant information to diagnose the failure without any intervention from the System Administrators. Specifically, for RAM and disk drive
		MRQ

	components, when a pre-failure event is detected, the system must automatically send the alert.	
4.1.1-4	*Health monitoring* The System will provide the capability to monitor the health parameters of each component via adequate software/hardware infrastructure. The monitoring software infrastructure should expose open-source API/frameworks in order to be integrated with open-source tools. All hardware faults of the components that can affect the performance and stability of the nodes and devices of the system must be reported.	MRQ
4.1.1-5	*Node power and energy measurement* The procured infrastructure provides node power and energy measurements with a minimum of 95% accuracy and low impact on performance (<1% on HPL). The details of the implementation will be considered in evaluating the Offers that provide this capability.	MRQ
4.1.1-6	*Power capping*[2] Offered infrastructures relying on power capping to meet the power constraint limits (see Section 3.3), the power consumption must be modified at unit level (e.g., rack) and at runtime with minimal impact on performances.	MRQ
4.1.1-7	*Monitoring APIs* The monitoring and management systems of the procured infrastructure must provide APIs enabling the integration with third-party monitoring and management frameworks. APIs must provide information on life status of all the components of the infrastructure in a timely fashion to be alerted on incurring faults within 300 seconds from their occurrence.	MRQ
4.1.1-8	*Examon* The Offer should provide installation support and maintenance of Examon framework[3]. Examon framework will collect and monitor the activity of all the nodes and equipment of the system. Moreover, it should be integrated in the current installation deployed at CINECA on most of the HPC and facility systems.	TRQ

4.2 Interconnects

4.2.1 Compute Fabric

Req.	Description	Category
4.2.1-1	*General requirements* The procured infrastructure must provide a low-latency high bandwidth fabric used to interconnect compute nodes. The compute fabric must implement the following features: • Must be based on 400 Gb/s or 800 Gb/s.	MRQ

[2] If the Offer does not rely on power capping this requirement is not considered mandatory

[3] https://github.com/EEESlab/examon

	• The minimum bandwidth for each access link must be at least 200Gb with full bi-directional bandwidth per port. • Support RDMA communications. • The average latency of MPI Point-to-Point communication must be less than 3 microseconds. • Provide optimization for MPI communications. • Support Fat-tree and/or Dragonfly(+) topology. • The oversubscription of the topology must not exceed 2:1. The compute fabric is also represented in the Figure 7.	
4.2.1-2	*Bandwidth* A fabric with full bisection bandwidth is preferred.	TRQ
4.2.1-3	*Topology* A full-fat tree topology is preferred.	TRQ
4.2.1-4	*Participants* All Compute, Management, and Front-end nodes must be connected to the compute fabric as shown in Figure 7.	MRQ
4.2.1-5	*Monitoring and managing capabilities.* The following capabilities must be provided: • The compute fabric must provide managing mechanism and for near-real time collection of performance and health information. • Each switch of the compute fabric must provide an out-of-band management port based on 1GbT.	MRQ
4.2.1-6	*Advanced fabric features* The compute fabric should support the following mechanisms: • *In-network computing*: for collective offloads. • *Avoid congestion*: through adaptive routing mechanisms. • *Self-healing*: with re-routing based on auto-discovery changes of the topology.	MRQ
4.2.1-7	*Power redundancy* Each compute fabric switch must be equipped with redundant & hot-swappable power supplies.	MRQ

4.2.2 Core Ethernet Network (CEN)

Req.	Description	Category
4.2.2-1	*General requirements* The procured infrastructure must provide a Core Ethernet Network (CEN) to connect Compute, Management, and Front-end nodes with storage infrastructures and to propagate Internet connection. The CEN must implement the following features: • It must be based on 400G Ethernet. • The bandwidth of each access link will be defined by the Candidate to respect the requirement requested by each type of node.	MRQ

	• It must provide full support to IPv4 and IPv6. • It must support RDMA over Converged Ethernet (RoCEv1 and RoCEv2). The CEN is also represented in the Figure 7.	
4.2.2-2	*Network Bandwidth* The CEN must be non blocking[4] (oversubscription 1:1) to avoid any network congestion among different storages. This requirement addresses the stringent SLAs that CINECA must respect on the system.	MRQ
4.2.2-3	*Network Topology* • Spine-leaf topology based on layer 3 and BGP/EVPN/VXLAN. • The network topology must be redundant at both spine and leaf switches. Failures of single switch/link must not affect the network flow. • The CEN must provide enough spine switches to accommodate future expansion of new leaf switches to increase the network up to 20% of the access ports. This is particular important for future expansion of the storages. Due to the Software-Defined Networking (SDN) nature of CEN, each switch must have the capability to propagate several different logical networks (EVPN-VXLAN).	MRQ
4.2.2-4	*Spine switches* Switches at spine level must provide ports at 400 GbE to be connected with leaf switches. The number of spine switches and the type of port will be defined by the Candidate in order to respect the req. 4.2.2-2.	MRQ
4.2.2-5	*Leaf switches* • Must provide access ports at 100 GbE and uplinks ports at 400 GbE. • Each switch dedicated to a specific storage must not share ports with other storage and with any other nodes in the System as shown in Figure 7.	MRQ
4.2.2-6	*Network Participants* All nodes must have at least two links connected to two different leaf switches of the CEN to provide redundancy and high availability.	MRQ
4.2.2-7	*Monitoring and managing capabilities* • The CEN must provide methods for management and for near-real time collection of performance and health information (e.g., sFlow). • Each switch of the CEN must provide an OOB management port based on 1 GbT.	MRQ
4.2.2-8	*Network capabilities* The switches of CEN must support: • Standard IEEE 802.X network protocols, in particular VXLAN (routing and bridging) and related protocols.	MRQ

[4] On each CEN's switch the aggregated bandwidth of uplinks must be equal to the aggregated bandwidth of downlinks (except for the spine switches, which don't have uplinks).

	• MLAG, LAG, LACP, VLAN protocols to interconnect the data network with the CINECA's HPC Backbone Network. • Routing protocols for IPv4 and IPv6 (e.g., OSPF, BGP, MP-BGP, OSPFv3). • EVPN multi-homing (ESI LAG). • NVMe over TCP (NVMe/TCP). • Jumbo frames.	
4.2.2-9	*Power redundancy* Each switch of the CEN must be equipped with redundant & hot-swappable power supplies.	MRQ

4.2.3 Management network

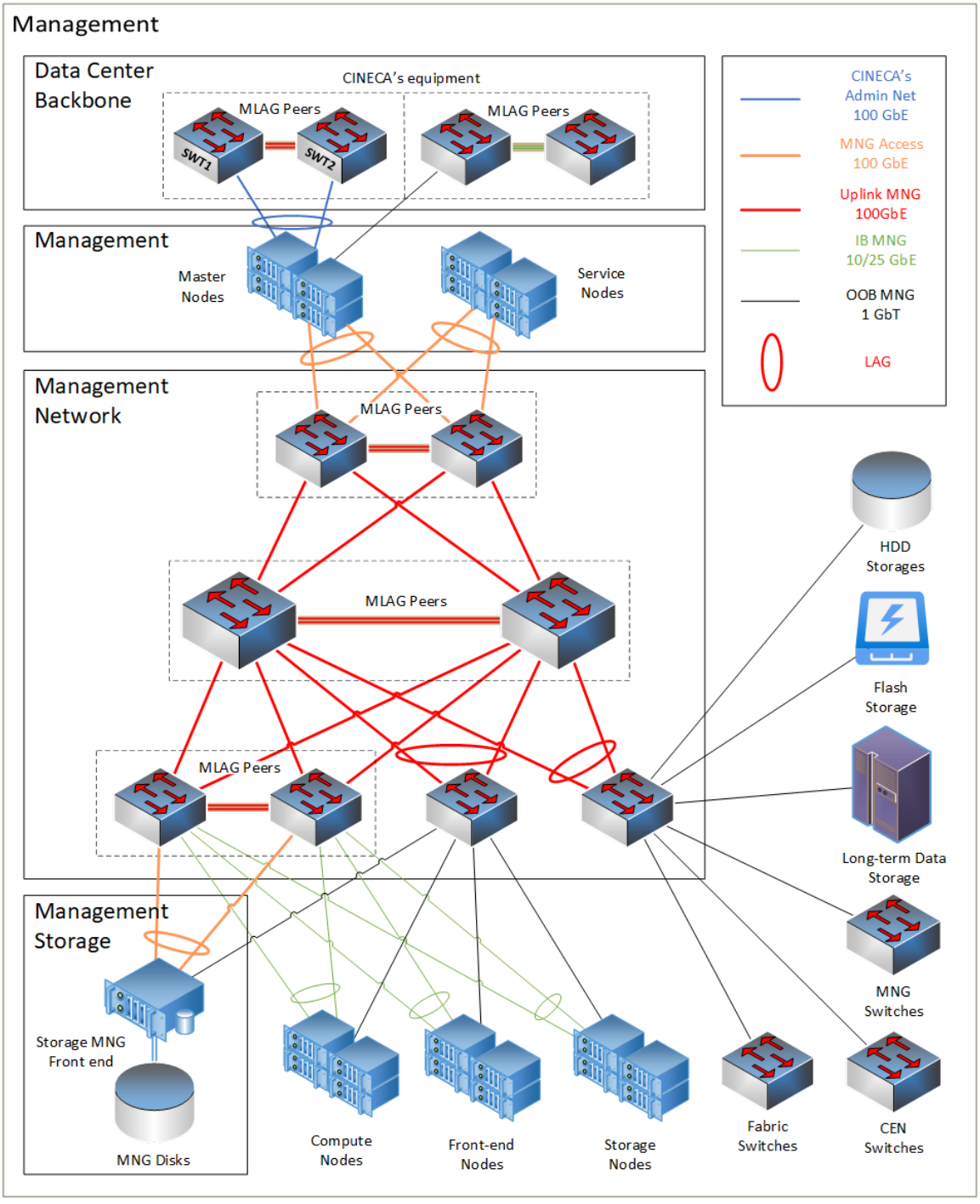

Figure 8: Reference design of the management network.

Req.	Description	Category
4.2.3-1	*Management network* The system must provide a physically dedicated Ethernet network for management purposes with the following characteristics: • Spine-leaf layer 2 based on MLAG topology is preferred. • The oversubscription of the topology must not exceed 4:1. • The network topology must be redundant both at inter-switch links and aggregation switches. Failures of single switch/link must not affect the network stability except for access switches. Redundancy for access switches is not required. • The management network must support two sub networks at logical link layer (VLAN): *In-Band (IB) and Out-Of-Band (OOB) management (MNG) network*. The management network is represented in Figure 8.	MRQ
4.2.3-2	*Network Participants* • I*n-Band (IB) management network*: used for managing and deploying compute nodes and data movers for operational services: bare metal/OS installation, OS monitoring and metering, etc. • *Out-Of-Band (OOB) management network*: used for managing Board Management Controller (BMC) of all system's equipment (storage, networks, chassis, etc.).	MRQ
4.2.3-3	*Monitoring and managing capabilities* The following capabilities must be provided: • The management network must provide methods for management and for near-real time collection of performance and health information (e.g., sFlow). • Each switch of the management network must provide an OOB MNG port based on 1 GbT connected to the OOB MNG network.	MRQ
4.2.3-4	*Spine switches* • Switches at spine level must provide ports at least 100 GbE to be connected to the leaf switches. The number of spine switches and the number of ports will be defined by the Candidate in order to respect the req. 4.2.3-1. • Spine switches must support static and dynamic routing (Layer 3) and relative protocols (e.g., OSPF, BGP, MP-BGP, OSPFv3, etc.).	MRQ
4.2.3-5	*Leaf switches* • Must provide access ports at 1 GbT for OOB MNG ports. • Must provide access ports at 10/25 GbE for IB MNG ports. • Must provide access ports at 100 GbE for the Management Nodes and storage. • Must provide uplink ports at 100 GbE.	MRQ
4.2.3-6	*Network capabilities* The switches of the management network must support:	MRQ

	• Full support to IPv4 and IPv6. • MLAG, LAG, LACP, VLAN protocols and most common IEEE 802.X network protocols.	
4.2.3-7	*Network performance* • Performance of these networks will allow for the full reconfiguration of the OS (without re-installation) of all nodes in less than 2 minutes. • Performance of these networks will allow for the (re-) installation of the OS of all CN nodes in less than 3 hours. • Performance of these networks will allow for cold reboot of all CN nodes in less than 60 minutes measured from the shut-down of the first node to the boot of the last non-faulted node. • Performance of these network will allow to collect all metrics and sensors of the management board of all compute nodes and network devices with open tools (i.e., IPMI tool, Redfish, Confluent, SNMP) in less than 20 seconds using as many parallel sessions as the monitoring infrastructure can use.	MRQ
4.2.3-8	*Power redundancy* Each switch in the management network should be equipped with redundant & hot-swappable power supplies.	TRQ

4.3 Compute partition

The Compute partition includes two sub partitions:

- *CPU partition*: it provides the conventional compute nodes based exclusively on CPU processing units.
- *GPU partition*: it provides the accelerated nodes based on GPUs devices.

4.3.1 CPU partition

Req.	Description	Category
4.3.1-1	*Partition Performance* The CPU partition must provide at least a HPL performance of 4 PFlops. Suitability of the partition for the scope of this procurement will be demonstrated by running efficiently the CPU benchmark suites described in Chapter 5 of this document.	MRQ
4.3.1-2	*Common node requirements* The nodes must implement the functional requirements provided in Section 4.1.1-3.	MRQ
4.3.1-3	*CPU Technology* The CPU must be based on x86_64 architecture, provide: • at least 56 cores. • a theoretical peak performance (FP64) of at least 3.5 TFlops. • high single-thread performance.	MRQ

4.3.1-4	*Node configuration* The node must be equipped with 2 CPUs.	MRQ
4.1.1-5	*DRAM memory* • The nodes must be equipped with at least 3 GBytes of DDR5 memory per core and not less than 512 GBytes. • The nodes must be configured to saturate all DDR memory channels of the CPUs (or in an optimal configuration to saturate the available memory bandwidth).	MRQ
4.3.1-6	*Memory Fat nodes configuration* At least 25 compute nodes of the CPU partition must feature 1 TByte of DDR5 memory or more.	MRQ
4.3.1-7	*Memory type* CPUs equipped with HBM memory are not allowed.	MRQ

4.3.1-8	*Network requirements*
	
	All CPU nodes must be equipped with:
	• 1 NIC with 1 port connected to the compute fabric with at least 200 Gb/s.
	• 1 NIC with 2 Ethernet ports connected to the CEN with at least 100 Gb/s per port (200 Gb/s aggregated). The network ports must be connected to different switches. Those network ports must support:
	  ○ VLAN HW offloading.
	  ○ Hardware offload of encapsulation and decapsulation of VXLAN.
	  ○ RDMA over Converged Ethernet (RoCE v1 and v2).
	  ○ RoCE over overlay networks.
	  ○ NVMe over Fabric target offloads.
	  ○ TCP/UDP/IP stateless offload.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10/25 Gb/s. The network ports must be connected to different switches and they must support:
	  ○ Pre-execution environment (PXE) boot.
	  ○ Remote boot over Ethernet.
	• 1 NIC with 1 Ethernet port dedicated to the BMC and connected to the OOB MNG network at 1Gb/s.
	
	OOB and IB MNG NICs can also share the same physical port.
		MRQ

4.3.1-9	*Local storage* The nodes must be equipped with x2 NVMe SSD drives in RAID1 configuration with net space available of ≥ 0.8 TBytes for the OS.	MRQ

4.3.2 GPU partition

Req.	Description	Category
4.3.2-1	*Partition performance* The GPU partition must provide at least a HPL performance of 15 PFlops.	MRQ

	Suitability of the partition for the scope of this procurement will be demonstrated by running efficiently the GPU benchmark suites described in Chapter 5 of this document.	
4.3.2-2	*Common node requirements* The nodes must implement the functional requirements provided in Section 4.1.1-3.	MRQ
4.3.2-3	*CPU technology* The CPU must be based on x86_64 architecture, provide: • at least 56 cores. • a theoretical peak performance (FP64) of at least 3.5 TFlops. • high single-thread performance.	MRQ
4.3.2-4	*GPU technology* • The GPUs must provide at least a theoretical peak performance (FP64) of 60 TFlops. • The GPUs must be able to share the memory with other GPU installed in the same node. • The GPUs must support IEEE-conforming single (FP32) and double precision (FP64) computations.	MRQ
4.3.2-5	*Node configuration* • The node must be equipped with 2 CPUs. • The node must be equipped with state-of-the-art GPU technology with at least 4 GPUs.	MRQ
4.3.2-6	*DRAM memory* • The nodes must be equipped with at least 512 GBytes of DDR5 memory and it must be greater equal to the sum of all GPU memories installed in the node. • The nodes must be configured to saturate all DDR memory channels of the CPUs (or in an optimal configuration to saturate the available memory bandwidth).	MRQ

4.3.2-7	*Network requirements*
	
	All GPU nodes must be equipped with:
	• 1 NIC with 1 port per GPU connected to the compute fabric with at least 200 Gb/s (≥800 Gb/s aggregated).
	• 1 NIC with 2 Ethernet ports connected to the CEN with at least 100 Gb/s per port (200 Gb/s aggregated). The network ports must be connected to different switches. Those network ports must support:
	  ○ VLAN HW offloading.
	  ○ Hardware offload of encapsulation and decapsulation of VXLAN.
	  ○ RDMA over Converged Ethernet (RoCE v1 and v2).
	  ○ RoCE over overlay networks.
	  ○ NVMe over Fabric target offloads.
	  ○ TCP/UDP/IP stateless offload.
		MRQ

	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10/25 Gb/s. The network ports must be connected to different switches and they must support:
	  ○ Pre-execution environment (PXE) boot.
	  ○ Remote boot over Ethernet.
	• 1 NIC with 1 Ethernet port dedicated to the BMC and connected to the OOB MNG network at 1 Gb/s.
	
	OOB and IB MNG NICs can also share the same physical port.	

4.3.2-8	*Local storage* The nodes must be equipped with x2 NVMe SSD drives in RAID1 configuration with net space available of ≥ 0.8 TBytes.	MRQ

4.4 Management partition

The Management partition includes two sub partitions and a shared storage:

- *Service partition*: A set of nodes dedicated to host all the general critical services (e.g., workload schedulers, system monitor, etc.).
- *Master partition*: A set of nodes dedicated for the whole cluster management (bare metal provisioning, internal networks management, etc.).
- *Management storage*: a shared storage infrastructure used from the Management Nodes to archive and collect information from the System.

Req.	Description	Category
4.4-1	*Common node requirements* The nodes must implement the functional requirements provided in Section 4.1.1-3.	MRQ
4.4-2	*General requirements* The management partition is used to host all system services and management tools of the System. The size of the management partition must be sufficient to support the operation of the system. Where possible, different services will be located on different (and possibly virtual) hosts. This partition will feature no less than 9 nodes. Candidates are invited to employ virtualization techniques to reduce the size of the service partition while complying with the above minimum.	MRQ
4.4-3	*Node configuration* The nodes must be equipped with two CPUs based on x86_64 architecture.	MRQ
4.4-4	*Memory configuration* All Management Nodes must feature a total of at least 256 GBytes of DDR5 memory.	MRQ
4.4-5	*Management storage* Management Nodes must have a shared storage with at least 200 TBytes to contain: 1. All the management software.	MRQ

	2. All the management databases and an historical daily (differential) backup of these databases for a year. 3. The aggregated system logs of all the node partitions for at least one year. 4. The aggregated audit logs of Compute, Front end, and Management Nodes for at least two years. 5. The performance and functional metrics collected from all nodes and equipment for at least two years. 6. The shared storage must be connected to the management network and must be possible to be mounted from all Management Nodes. 7. The management storage must be physically separated from the storage of the other partitions and must be shared between all the management nodes and seamlessly available to all the services. 8. The storage configuration must be resilient to the failure of at least two independent basic blocks (i.e., storage nodes, controllers, or disk chassis).	

4.4-6	*Networking configuration*
	
	Each Management Node must be equipped with:
	• 1 NIC with 1 port connected to the compute fabric with at least 200 Gb/s.
	• 1 NIC with 2 Ethernet ports connected to the CEN with at least 100 Gb/s per port (200 Gb/s aggregated). The network ports must be connected to different switches. Those network ports must support:
	  ○ VLAN HW offloading.
	  ○ Hardware offload of encapsulation and decapsulation of VXLAN.
	  ○ RDMA over Converged Ethernet (RoCE v1 and v2).
	  ○ RoCE over overlay networks.
	  ○ NVMe over Fabric target offloads.
	  ○ TCP/UDP/IP stateless offload.
	• 1 NIC with 2 Ethernet ports connected to the CINECA's Admin Network with 100 Gb/s.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network with 100 Gb/s. The network ports must be connected to different switches and they must support:
	  ○ Pre-execution environment (PXE) boot.
	  ○ Remote boot over Ethernet.
	• 1 NIC with 1 Ethernet port dedicated to the BMC and connected to the CINECA's OOB MNG network at 1Gb/s.
	
	OOB and IB MNG NICs can also share the same physical port.
		MRQ

4.4-7	*Connection with* CINECA's HPC Backbone The Management Nodes must be connected with the CINECA's HPC Backbone Network, which is based on two NVIDIA Spectrum SN4700[5] (represented in Figure 8 as *SWT1* and *SWT2*) with Cumulus Linux. The cabling length required for the connection is between 20 and 50 meters depending on the final racks' layout. All cables, optics, and 100/400 GbE splitters for both Management	MRQ

[5] https://www.nvidia.com/content/dam/en-zz/Solutions/networking/br-sn4000-series.pdf

	Nodes and CINECA's network equipment (NVIDIA Spectrum SN4700) must be provided in the Offer. See Figure 8 for the connection's details of the Management nodes.	
4.4-8	*Performance* The number of the Service Node must guarantee the requested performance levels for installation and reboot as well as effectiveness in collecting, storing, and processing all the metrics and logs. An excellent performance level must be guaranteed also during queries to collected data and all the management, troubleshooting, accounting, and security assessment activities.	MRQ
4.4-9	*High Availability* The service partition must include the required hardware components to configure all important system services in high availability. The service nodes must be configured in "cluster" mode, meaning that they should guarantee fault tolerance in terms of hardware and software services. All these functionalities must be available even in case of single or double node's fault with adequate levels of efficiency. Besides, a workload running on the Compute, Visualization and Login Nodes must be able to continue working without significant interruption. Any performance impact will be described in the Offer.	MRQ
4.4-10	*Health and Consistency Checks* The Candidate will provide tools to check the health and validate configuration of hardware and software components that can be integrated with the workload manager to ensure that only fully functional components are utilized for jobs. Where applicable, auto recovery actions will be performed and logged.	TRQ
4.4-11	*Rolling Updates* The system will provide "rolling update" mechanisms that allow reliable software updates and selected maintenance operations to be performed with minimal accumulated downtime. In full system maintenances, the idle time of Nodes incrementally grow prior to the start of the maintenance as running jobs finish and no new jobs start due to the pending maintenance reservation. The requested feature will significantly reduce this maintenance overhead.	MRQ
4.4-12	*Cluster Management Software* The Candidate will provide an integrated software solution for the management of all cluster resources, the provisioning of nodes and basic hardware and operating system monitoring. The software will offer support for the (out-of-band) management of all hardware components and node provisioning. The software will enable the automation of all the fundamental system management activities: • Installation of the OS on the nodes. • Reconfiguration of the OS of the nodes (and possibly of all apparatus). • Collection of nodes diagnostic information (and possibly of all apparatus). • Update of the firmware nodes.	MRQ

	This software is typically in execution on the Service nodes, must feature a redundant configuration mechanism, and preferably be open source and belonging to the OpenHPC initiative.	
4.4-13	*Basic Hardware Monitoring* The cluster management software will provide out-of-band and/or in-band monitoring of hardware events (e.g., system event log and machine check exceptions, if applicable). The events will be collected and stored at a central location.	MRQ

4.5 Front-end partition

The Front-end partition includes two sub partitions:

- *Login partition*: for external system access, compilation and data management activities, job submission as well interactive pre-/post-processing workloads.
- *Visualization partition*: to enables visualization of simulation results during and after the execution of jobs. Visualization Nodes may be operated in batch mode or as externally accessible interactive nodes.

4.5.1 Login partition

Req.	Description	Category
4.5.1-1	*Partition size* The login partition must feature at least 8 nodes.	MRQ
4.5.1-2	*Common node requirements* The nodes must implement the functional requirements provided in Section 4.1.1-3.	MRQ
4.5.1-3	*Node configuration* The login nodes must be organized in two different setups: • *CPU setup*: half of the Login Nodes must feature the same CPUs of the CPU partition's nodes. • *GPU setup*: half of the Login Nodes must feature the same CPUs of the GPU partition's nodes. They must also include at least one GPU with the same technology of the GPUs installed in the GPU partition.	MRQ
4.5.1-4	*Memory configuration* The Login Nodes must feature a total of at least 512 GBytes of DDR5 memory.	MRQ
4.5.1-5	*Local storage* All Login Nodes must feature at least two internal NVMe SSD drives that can be configured in RAID1 for a total of at least 7 TBytes of net storage for the OS.	MRQ
4.5.1-6	*Network requirements* All Login Nodes must be equipped with: • 1 NIC with 1 port connected to the compute fabric with at least 200 Gb/s.	MRQ

	• 1 NIC with 2 Ethernet ports connected to the CEN with at least 100 Gb/s per port (200 Gb/s aggregated). The network ports must be connected to different switches. Those network ports must support:
	  ○ VLAN HW offloading.
	  ○ Hardware offload of encapsulation and decapsulation of VXLAN.
	  ○ RDMA over Converged Ethernet (RoCE v1 and v2).
	  ○ RoCE over overlay networks.
	  ○ NVMe over Fabric target offloads.
	  ○ TCP/UDP/IP stateless offload.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10/25 Gb/s. The network ports must be connected to different switches and they must support:
	  ○ Pre-execution environment (PXE) boot.
	  ○ Remote boot over Ethernet.
	• 1 NIC with 1 Ethernet port dedicated to the BMC and connected to the OOB MNG network at 1Gb/s.
	
	OOB and IB MNG NICs can also share the same physical port.	

4.5.1-7	*Internet connectivity* The Login Nodes have direct access to Internet through the CEN, for this reason the CEN must propagate the Internet connection via a VXLAN tunnel.	MRQ
4.5.1-8	*Software installation* Login Nodes must allow the installation of all user software and applications that need to be run on the system.	MRQ

4.5.2 Visualization partition

Req.	Description	Category
4.5.2-1	*Partition size* The visualization partition must be composed to at least 4 nodes.	MRQ
4.5.2-2	*Common node requirements* The nodes must implement the functional requirements provided in Section 4.1.1-3.	MRQ
4.5.2-3	*Node configuration* The nodes must be equipped with: • Two state-of-the-art CPUs, binary compatible with the CPU equipping the nodes of CPU and GPU partition. • at least two high-end graphical cards supporting 3D acceleration through OpenGL graphics.	MRQ
4.5.2-4	*Memory configuration* The Visualization Nodes must feature a total of at least 512 GBytes of DDR5 memory.	MRQ
4.5.2-5	*Local storage*	MRQ

	The Visualization Nodes must feature at least two internal NVMe SSD drives that can be configured in RAID1 for a total of at least 7 TBytes of net storage for the OS.	

4.5.2-6	*Network requirements*
	
	All Login Nodes must be equipped with:
	• 1 NIC with 1 port connected to the compute fabric with at least 200 Gb/s.
	• 1 NIC with 2 Ethernet ports connected to the CEN with at least 100 Gb/s per port (200 Gb/s aggregated). The network ports must be connected to different switches. Those network ports must support:
	  ○ VLAN HW offloading.
	  ○ Hardware offload of encapsulation and decapsulation of VXLAN.
	  ○ RDMA over Converged Ethernet (RoCE v1 and v2).
	  ○ RoCE over overlay networks.
	  ○ NVMe over Fabric target offloads.
	  ○ TCP/UDP/IP stateless offload.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10/25 Gb/s. The network ports must be connected to different switches and they must support:
	  ○ Pre-execution environment (PXE) boot.
	  ○ Remote boot over Ethernet.
	• 1 NIC with 1 Ethernet port dedicated to the BMC and connected to the OOB MNG network at 1 Gb/s.
	
	OOB and IB MNG NICs can also share the same physical port.
		MRQ

4.5.2-7	*Internet connectivity* The Visualization Nodes have direct access to Internet through the CEN, for this reason the CEN must propagate the Internet connection via a VXLAN tunnel.	MRQ

4.6 Storage infrastructure

The storage infrastructure is organized on three sub storage:

- *Tier-1 AIM*: this storage will be dedicated to the "Agenzia Italia Meteo (AIM)" and it will contain Italian meteorological data.
- *Tier-1 CNR-INAF-UNIBO*: this storage will be of general use, serving communities coming from the Consiglio Nazionale delle Ricerche (CNR), Istituto Nazionale di Astrofisica (INAF), University of Bologna, ICSC national centre, and other communities, and it will mainly be devoted to process scientific data.
- *Tier-1 SKA*: this storage will contain astronomical data coming from the SKA project.

Each storage must be independent from the others, it will be interconnected on dedicated leaf swtiches, which it will be uplinked to spine switches of CEN as shown in Figure 7. All nodes of the procured system must be able to mount all the storages through a parallel file system (or equivalent). Each storage must also have the possibility to be connected on Internet to expose their S3 interface through a VXLAN tunnel. Each storage will also have a set of dedicated Data Movers, which are nodes utilized to move data in and out of the storage through Internet.

4.6.1 Data Movers

Req.	Description	Category
4.6.1-1	*Partition size* The Data Mover partition must be composed to at least 8 nodes organized as follow: • At least 2 Data Movers for Tier1 AIM storage. • At least 2 Data Movers for Tier1 INAF-CNR-UNIBO storage. • At least 4 Data Movers for Tier3 SKA storage.	MRQ
4.6.1-2	*Common node requirements* The nodes must implement the functional requirements provided in Section 4.1.1-3.	MRQ
4.6.1-3	*Node configuration* The nodes must be equipped with two CPUs based on x86_64 architecture.	MRQ
4.6.1-4	*Memory configuration* The nodes must feature a total of at least 256 GBytes of DDR5 memory.	MRQ
4.6.1-5	*Local storage* The nodes must feature at least two internal NVME SSD drives that can be configured in RAID1 for a total of at least 7 TBytes of net storage for the OS.	MRQ

4.6.1-6	*Network requirements*
	
	All nodes must be equipped with:
	• 1 NIC with 2 Ethernet ports connected to the CEN with at least 100 Gb/s per port (200 Gb/s aggregated). The network ports must be connected to different switches. Those network ports must support:
	  ○ VLAN HW offloading.
	  ○ Hardware offload of encapsulation and decapsulation of VXLAN.
	  ○ RDMA over Converged Ethernet (RoCE v1 and v2).
	  ○ RoCE over overlay networks.
	  ○ NVMe over Fabric target offloads.
	  ○ TCP/UDP/IP stateless offload.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10/25 Gb/s. The network ports must be connected to different switches and they must support:
	  ○ Pre-execution environment (PXE) boot.
	  ○ Remote boot over Ethernet.
	• 1 NIC with 1 Ethernet port dedicated to the BMC and connected to the OOB MNG network at 1 Gb/s.
	
	OOB and IB MNG ports can also share the same port.
		MRQ

4.6.1-7	*Internet connectivity* The nodes have direct access to Internet through the CEN, for this reason the CEN must propagate the Internet connection via a VXLAN tunnel.	MRQ

4.6.2 Tier-1 Storage – General requirements

In this Section, we provide a list of characteristics that the storage infrastructures of Tier-1 AIM, Tier-1 CNR-INAF-UNIBO, and Tier-1 SKA Online must respect. On each specific Section, we will provide incremental characteristics dedicated for each storage.

Req.	Description	Category
4.6.2-1	*Noisy neighbours* The architecture of the offered solution must guarantee a low impact of noisy neighbours to comply with the expect SLAs. In particular, the storage infrastructures must address this requirement.	MRQ
4.6.2-2	*Backup & Recovery* The procured solution needs to provide resources, hardware and software, to allow daily incremental backup of HOME namespaces. To size these resources dedicated for backup, it must be assumed - per each of the following storage infrastructures - to backup a filesystem with many small files, typically 10 billion of files with an average of 30 kB per file, and assuming an average of 0.2% of daily file changes. These resources need to provide the capability to traverse the filesystems and find all the daily file changes that need to be sent for backup.	MRQ
4.6.2-3	*Core Ethernet Network connection* • The storage infrastructures will be connected via CEN to all System's partitions. Each node of the System must be potentially able to mount and get access to the data contained in all the storage infrastructures. • At least a couple of ports on two different Front-end storage nodes must be dedicated to the Internet connection to provide public S3 interface. Each port dedicated to the Internet connectivity for S3 interface must support at least 100 Gb/s.	MRQ
4.6.2-4	*Management* • The storage infrastructures will be connected to the Management Networks for management operations and configuration tasks. • The System's management partition must be able to access the management interface of the storage infrastructures.	MRQ
4.6.2-5	*Accessibility* • All namespaces of all storage infrastructures must be accessible from all System's partitions. • The namespaces of storage infrastructures must be seen by the client nodes as a POSIX interface. • The storage infrastructures must support LDAP for authentication.	MRQ
4.6.2-6	*High availability & Resiliency* • The storage infrastructures must not have any single point of failure. • The storage infrastructures must provide transparent failover in the event of a node or network failure. • Data rebuilds must be fail-in-place and must begin automatically without device replacement or operator action. • All updates must be non-disruptive to all users.	MRQ

	• In case of a fault in any component of the storage infrastructures, the proposed System must provide an automatic procedure to recover from the fault. This process must not have impact on data integrity or normal accessibility. • In case of a replacement of a storage element or increase of capacity, the system must be able to redistribute and balance the data without service interruption.	
4.6.2-7	*Self-recovery time duration* The recovering process described in 4.6.2-2 should not take more than 4 hours and the impact on performance must be less than 10%.	MRQ
4.6.2-8	*Parallel filesystem* • The storage infrastructures must support a parallel filesystem (or equivalent) in order to read and write data across multiple storage front ends and to facilitate high-performance access through simultaneous, coordinated input/output operations between clients and storage front-end nodes. • The storage client must be able to support multipath across multiple links to get access to the storage infrastructures.	MRQ
4.6.2-9	*System Architecture* • The storage infrastructures must support client access via IPv4 and IPv6. • The storage infrastructures must be based on a scale-out architecture. • The storage infrastructures must allow for independent scaling of capacity (up to exascale level) and performance. • The storage infrastructures must support the integration of new generations of hardware. • The storage infrastructures shall provide transparent failover in the event of a storage front-end node or network failure. • Namespace must be scalable to support over 100 PByte of capacity.	MRQ
4.6.2-10	*Management and Security* • The storage infrastructures must be manageable via CLI, GUI (HTTPS), or RESTful API. • The vendor must provide centralized monitoring and analytics service. • Services must allow issues to be proactively discovered. • Service must securely collect log and analytics data. • The storage infrastructures must support Role Based Access (RBAC) control for data and administrative access. • The storage infrastructures must support LDAP authentication for administrative access. • The storage infrastructures must provide audit log of administrative access. • The storage infrastructures must provide audit log of client operations.	MRQ
4.6.2-11	*QoS*	TRQ

	• The storage infrastructures should support QoS throttles for shares/mounts in IOPS and/or MB/s. • The storage infrastructures should support QoS throttles for users and groups in IOPS and MB/s.	
4.6.2-12	*Software Licencing and Support* • Vendor must guarantee software licences and support, including replacement of all parts regardless of flash wear, will be offered on the storage infrastructures. • Software licenses shall be transferable to replacement hardware including hardware of later generations. • The software licenses for the backup & recovery services must be included in the Offer.	MRQ

4.6.3 Tier-1 AIM storage

Req.	Description	Category

4.6.3-1	*Specific requirements*
	
	The offered solution must provide:
	• A full-flash storage.
	• At least 5 PBytes of net space available.
	• At least a total read throughput of 60 GB/s.
	• At least a total write throughput of 20 GB/s.
	• At least 500K read IOPS (4k random reads) for NFS (or parallel file system).
	• At least 250K write IOPS (4k random writes) for NFS (or parallel file system).
	• The system will be configured with two main namespaces:
	  ○ Home: on this namespace will be contained the personal files of the users.
	  ○ Work: this namespace is used to contain project files and it used to process data.
	
	N.B. Storage infrastructure that provides net space using data reduction mechanisms (compression, deduplication, etc.) are allowed, but the Candidate must declare and commit on a certain level of available net space (with or without data reduction) and provide additional storage in case of violation in order to reach the commitment.
	Data reduction must not require administrators to manage, must be performed by the storage system inline before data is written to disk.
	Throughput and IOPS declared must be considered with data reduction overheads.
		MRQ

4.6.3-2	*General requirements* The storage infrastructure must comply with all the requirements provided in Section 4.6.2.	MRQ
4.6.3-3	*Security features*	MRQ

	• The storage infrastructure must have the possibility to encrypt all data before storing it in persistent media. • The storage infrastructure must support NFS 4.1 encryption in flight. • Encryption must support AES-256. • The storage infrastructure must provide immutable snapshots for Ransomware protection.	
4.6.3-4	*NFS support* • The storage infrastructures must support NFS v3 and v4.1. • The storage infrastructures must support POSIX and NFS v4.1 ACLs. • The storage infrastructures must support NFS over RDMA for NFS v3 and v4.1. • The storage infrastructures must support directory quotas that apply to NFS.	MRQ
4.6.3-5	*S3 support* • 40 TBytes or larger maximum object size. • Support for >100 buckets/namespace. • S3 identity and bucket policies. • Multipart uploads. • Object Lock.	MRQ
4.6.3-6	*Multi-protocol support* • The storage infrastructures must support object access via S3 APIs. • The storage infrastructures must support access via NFS v3. • The storage infrastructures shall support client access via IPv4 and IPv6. • The storage infrastructures must support directory quotas that apply to NFS. • The storage infrastructures must support quotas on both a per-user/group and per-folder basis. • Files created via NFS must be accessible as S3 objects. • Objects created via S3 must be accessible as NFS files. • The storage infrastructure must support asynchronous replication on the same storage technology but geographically distributed.	MRQ
4.6.3-7	*File system metadata* The offered solution should support the possibility to make the file system metadata available for query as database tables.	TRQ
4.6.3-8	*Advanced Analytics* • The offered solution should support the possibility to store in an integrated Database, metrics, data and metadata in a columnar table format natively stored within the data platform. • The Database should be integrated with modern Big Data frameworks like Spark, Trino, Dremio including query filter pushdowns into that Database to speed up significantly query performances. • The database should support ACID and must handle millions of transactions per second and scale to terabytes/second of query throughput.	TRQ

	• The system must provide a built-in tabular Database that should support schema evolutions to flexibly expand natural data types such as images, video, etc. along with data pertaining to files and objects.	

4.6.4 Tier-1 CNR – INAF – UNIBO storage

Req.	Description	Category

4.6.4-1	*Specific requirements*
	
	The offered solution must provide:
	• A full-flash storage.
	• At least 10 PBytes of net available space that will contain Work namespace used to host project files and to process data.
	  ○ This storage infrastructure must have at least a total read throughput of 120 GB/s.
	  ○ This storage infrastructure must have at least a total write throughput of 60 GB/s.
	
	N.B. Storage infrastructure that provides net space using data reduction mechanisms (compression, deduplication, etc.) are allowed, but the Candidate must declare and commit on a certain level of available net space (with or without data reduction) and provide additional storage in case of violation in order to reach the commitment.
	Data reduction must not require administrators to manage, must be performed by the storage system inline before data is written to disk.
	Throughput and IOPS declared must be considered with data reduction overheads.
		MRQ

4.6.4-2	*General requirements* The storage infrastructure must comply with all the requirements provided in Section 4.6.2.	MRQ
4.6.4-3	*NFS support* • The storage infrastructures must support NFS v3 and v4.1. • The storage infrastructures must support POSIX and NFS v4.1 ACLs. • The storage infrastructures must support NFS over RDMA for NFS v3 and v4.1. • The storage infrastructures must support directory quotas that apply to NFS.	MRQ
4.6.4-4	*S3 support* • 40 TBytes or larger maximum object size. • Support for >100 buckets/namespace. • S3 identity and bucket policies. • Multipart uploads. • Object Lock.	MRQ
4.6.4-5	*Multi-protocol support* • The storage infrastructures must support object access via S3 APIs. • The storage infrastructures must support access via NFS v3.	MRQ

	• The storage infrastructures shall support client access via IPv4 and IPv6. • The storage infrastructures must support directory quotas that apply to NFS. • The storage infrastructures must support quotas on both a per-user/group and per-folder basis. • Files created via NFS must be accessible as S3 objects. • Objects created via S3 must be accessible as NFS files. • The storage infrastructure must support asynchronous replication on the same storage technology but geographically distributed.	

4.6.5 Tier-3 SKA storage

4.6.5.1 Tier-3 SKA online storage

Req.	Description	Category
4.6.5.1-1	*Specific requirements* The offered solution must provide: • A capacity-oriented HDD-based storage. • At least 3 PBytes of net space available. • At least a total read throughput of 30 GB/s. • At least a total write throughput of 15 GB/s. N.B. No data reduction mechanisms are allowed on this storage infrastructure.	MRQ
4.6.5.1-2	*General requirements* The storage infrastructure must comply with all the requirements provided in Section 4.6.2.	MRQ
4.6.5.1-3	*NFS support* • The storage infrastructures must support NFS v3 and v4.1. • The storage infrastructures must support POSIX and NFS v4.1 ACLs. • The storage infrastructures must support NFS over RDMA for NFS v3 and v4.1. • The storage infrastructures must support directory quotas that apply to NFS.	MRQ
4.6.5.1-4	*S3 support* • 40 TBytes or larger maximum object size. • Support for >100 buckets/namespace. • S3 identity and bucket policies. • Multipart uploads. • Object Lock.	MRQ
4.6.5.1-5	*Multi-protocol support* • The storage infrastructures must support object access via S3 APIs. • The storage infrastructures must support access via NFS v3. • The storage infrastructures shall support client access via IPv4 and IPv6.	MRQ

	• The storage infrastructures must support directory quotas that apply to NFS. • The storage infrastructures must support quotas on both a per-user/group and per-folder basis. • Files created via NFS must be accessible as S3 objects. • Objects created via S3 must be accessible as NFS files. • The storage infrastructure must support asynchronous replication on the same storage technology but geographically distributed.	

4.6.5.2 Tier-3 SKA Long-term Data Storage (LTS)

Req.	Description	Category
4.6.5.2-1	*General requirements* The offered solution must provide a tape library with a net capacity of at least 5 PBytes.	MRQ
4.6.5.2-2	*Scalability* The LTS must be upgradable and scalable to reach at least 300 PBytes capacity in 10 years.	MRQ
4.6.5.2-3	*Performance* • The LTS must be able to host a consistent number of tape cartridges (e.g. LTO-9) with a nominal writing performance of at least 400 MB/s. • The LTS must be able to read/write at least 6 tape cartridges in parallel. • Each tape cartridge must provide at least 18 TB of uncompressed capacity. • The LTS must be able to deliver a performance with at least 64 Gbit/s of writing speed.	MRQ
4.6.5.2-4	*Data Retrieval* • The LTS must be easily accessible (read/write) directly (same file system as on-line storage) or via NFS by the on-line storage infrastructure. • The LTS must be seen as a unique volume or can be partitioned in different volumes at necessity. • Each tape cartridge must support the WORM capability. • Each file exceeding the total capacity of one tape cartridge must be automatically written on multiple tape cartridges but recognized as one (no manual split must be required). • The LTS must also foresee a backup technology (self-checking and self-recovery). • The LTS must be also become a logical partition of the Tier-3 SKA Online Storage. • Integrated native policy manager capable of managing the data movement between disks and tapes. • Availability of policies for automated data movement and archiving between different layers (e.g., flash, disk, tape, etc.).	MRQ

4.6.5.2-5	*Data retention* The tape cartridges must guarantee a lifespan at least of 10 years.	MRQ
4.6.5.2-6	*Interconnection* The LTS must be interconnected with one or multiple Data Movers of SKA storage in order to manage the data moving on Tier-3 SKA Online Storage.	MRQ
4.6.5.2-7	*Total Cost of Ownership* The offered long-term solution must be very efficient in terms of cost/TB and power consumption.	MRQ

4.7 Facility integration

Req.	Description	Category
4.7-1	*Integration with CINECA's HPC Backbone network* The system will be connected to the CINECA's *HPC* Backbone network through two CEN switches as shown in Figure 7. The Offer must include the CEN switches for the connection with the CINECA's network equipment and provide configuration and integration through professional services. The CINECA's HPC Backbone Network is based on two NVIDIA Spectrum SN4700[6] (represented in Figure 7 as *SWT1* and *SWT2*) with Cumulus Linux. The cabling length required for the connection is between 20 and 50 meters depending on the final racks' layout. The aggregated bandwidth between the CEN switches and the CINECA's HPC Backbone network equipment must be at least 800 Gb/s. All cables and optics for both CEN switches and CINECA's network equipment must be provided in the Offer.	DCS
4.7-2	*Closed Innermost Cooling Loop* If applicable, the innermost cooling loop will be well separated from the data centre infrastructure, e.g., by means of a heat exchanger or CDU. The regular control and maintenance of the water quality in this loop will be the responsibility of the Candidate.	MRQ
4.7-3	*Failure Detection and Reaction* The system will provide system-internal mechanism to detect, in real time, infrastructure related environment changes (e.g., leakage, pressure drops or temperature changes) and react, in (near-) real time, in such a way that damage of the system or data centre is prevented.	MRQ
4.7-4	*Communication with Facility* The system will provide mechanisms (e.g., APIs) to enable a link to be established between the system internal monitoring and the facility monitoring	MRQ

[6] https://www.nvidia.com/content/dam/en-zz/Solutions/networking/br-sn4000-series.pdf

	system. It will be possible to exchange infrastructure status information as well as events and alarms between the systems.	

4.8 System software and monitoring

Req.	Description	Category
4.8-1	*Operating System* The operating system for the System's nodes must be RHEL with 64-bit kernel version 5.14 or higher, and supporting remote management, network boot and system image delivery. It will be allowed to install security patches soon after their release independently from the constraints of the Compute and Front-end nodes (I.e., GPU or FS drivers/software).	MRQ
4.8-2	*User Management with LDAP* The design of the procured infrastructure must enable integration in CINECA's OpenLDAP-based directory service for user management in such a way that no single-point of failures exist.	MRQ
4.8-3	*Container support* The procured infrastructure enables the execution of containerized applications, i.e., applications utilizing a different system software stack than the one natively available on the Compute Nodes, with a small overhead relative to native execution. Container support is well integrated with the other system management components, in particular with the workload and resource management system. The system software will provide a mechanism to build and modify containers and ensure consistency between the container and the native execution environment (e.g., to ensure that the native kernel and containerized user-space components are compatible). At least one common container format will be supported, such as the Open Container Initiative (https://www.opencontainers.org/) v1.0 (or the current) or Image Specification used by Singularity's image format (https://sylabs.io/). Container construction will be possible based on portable recipes that enables user communities to utilize similar container images on different supercomputing sites. It will be possible to control, on a fine-grained basis, the permissions for container creation, modification, and execution to enable adoption to site security policies.	MRQ
4.8-4	*Support for recent programming environment standards.* All offered MPI implementations and compiler suites must support recent versions of the applicable standards: • MPI version 3.0 or newer. • OpenMP 4.5 or newer. • C ISO/IEC 9899:2011 or newer. • C++ ISO/IEC 14882:2014 or newer. • Fortran ISO/IEC 1539-1:2010 (aka Fortran 2008) or newer. • Python 3.8 or newer. Full stack software programming paradigm for accelerators, including at least C, C++, Fortran, and Python front-end.	MRQ
4.8-5	*Lightweight Performance Profiling* The procured infrastructure should provide lightweight performance profiling capabilities that can be activated by the users on a job basis. Data at process,	TRQ

	job and node level will be made available utilizing scalable accumulation methods. Data retention times for the mentioned granularity levels may differ. The technology will have minimal impact on application performance (less than 5% performance drop) and in particular not affect scalability of large jobs. A basic set of data must be gathered irrespectively of the user application, i.e., without requiring the users to link against specific libraries. At least the following system components will be covered: • CPU utilization (load avg.), IPC, Instruction mix information, memory footprint, cache utilization/hits/miss, TLB hits/miss, load/store ops, memory interface utilization: ○ For systems featuring multiple memory types and a deep(er) memory hierarchy, information will be gathered for all types and tiers. • I/O subsystem: number of reads/writes, read/write bandwidth. • Network: number of packets/reads/writes (RDMA), packet/segment length. • Accelerator: utilization. • MPI/communication libraries: Number of calls, time spent. It will be possible to flexibly extend the gathered observables (potentially with additional/higher overhead).	
4.8-6	*Performance report generation* The Candidate will provide tools and/or an API to create job performance reports for users based on the collected data. Ideally, the focus of the report will be controllable in terms of the level of detail as well as the considered system components. The level of detail and numbers of levels must be dynamically adjustable by administrators and users. An API for accessing the reports in a machine-readable format will be available (e.g., for the integration with the workload manager or external web portals). An integration with the workload manager, allowing appending reports to job output and email-based job notifications, is desirable. This integration requirement does not apply to offers that do not include the workload manager.	TRQ
4.8-7	*Anomaly detection* The Candidate will provide tools and algorithms to detect anomalies in the gathered performance data. This will provide operators with additional capabilities to detect problematic system components and assess the impact of System's changes (e.g., software updates) on application performance	TRQ
4.8-8	*Mechanisms for correlation* The Candidate will provide mechanisms that enable correlation of different metrics from different monitoring systems (including external ones). An open API for access to the data through the unification layer will be provided. This API must allow for the exporting of data in near-real time to other systems (e.g., external monitoring infrastructures). In addition, a graphical tool for operators is desirable.	TRQ
4.8-9	*Optimized numerical libraries*	MRQ

	The Candidate must provide highly optimized libraries providing API compatible replacements for BLAS, LAPACK and ScaLAPACK routines. The Proposal must include an optimized fast Fourier transform (FFT) library.	
4.8-10	*Parallel debugger* An adequate parallel debugger and profiler must be included in the software package to allow debugging of parallel application. It must be licensed for at least 10 CN.	MRQ
4.8-11	*Hardware counters* The hardware counters of the CPU/GPU must be mature and accessible (e.g., by a tool like LIKWID). In addition, a software tool must be provided which makes it possible to measure and automatically collect the performance of the users' applications running as batch job. The performance measurement should be based on the metrics of the performance counters.	MRQ

5 Benchmarks

5.1 Introduction

This chapter describes the context and main goals of the benchmark procedure for assessing the performance and the application portability on the Candidate proposed CPU and GPU partition. All instructions to run the benchmarks (source codes, instructions, dataset inputs, etc.) are available in the following public git repository:

https://gitlab.hpc.cineca.it/procurement/tier1-tecnopolo/benchmarks

This chapter is organized as follows:

- In Section 5.2 the benchmark framework, including criteria and metrics considered for the selection of the suite, are described.
- In Section 5.3 the benchmark procedure is described.
- In Section 5.4 it is described how the results of the benchmarks should be collected and provided to the Public Procurer.
- In Section 0 it is recalled to elaborate and provide the Benchmark analysis report.

5.2 Benchmark framework

5.2.1 Software metrics

The following metrics were considered in detailing the benchmark instructions.

- **Time-to-solution (TTS):** the wall-time spent to complete the execution of the single benchmark application.
- **System scale:** the mandatory minimum system partition to run the benchmark, expressed in terms of HPL performance.

5.2.2 Benchmark categories

The following main categories were considered in selecting the applications of the benchmark suite. They represent various scientific problems and computing patterns of particular interest for the AIM, INAF, CNR, and UNIBO communities.

- **Synthetic kernel applications**. These applications allow to evaluate system-wide performance, independently of specific workloads.
- **Scalable application**. Full applications that are expected to scale up to a large fraction of the system. For these codes, scalability to large number of processors allows for exploring new problems and defining the frontiers of knowledge and scientific insight.

5.2.3 Benchmark suite

The benchmark suite is composed by the following performance synthetic kernels and scalable applications.

#	Benchmark	Partition	type	Short description
1	**HPL**	CPU and GPU	Synthetic kernel	HPL benchmark solves a linear system of equations of order n, measuring the sustained performance of the whole system.
2	**HPCG**	CPU and GPU	Synthetic kernel	HPCG (high performance conjugate gradient) benchmark is intended to model the data-access patterns of real-world applications such as sparse matrix.
3	**Quantum Espresso**	CPU and GPU	Scalable Application	Quantum Espresso is an integrated suite of Open-Source computer codes for electronic-structure calculations and materials modelling at the nanoscale. It is based on density-functional theory, plane waves, and pseudopotentials.
3	**ICON**	CPU	Scalable Application	ICON is a flexible, scalable, high-performance modelling framework for weather, climate and environmental prediction that provides actionable information for society and advances our understanding of the Earth's climate system.
4	**PLUTO**	CPU	Scalable Application	PLUTO is a freely-distributed software for the numerical solution of mixed hyperbolic/parabolic systems of partial differential equations (conservation laws) targeting high Mach number flows in astrophysical fluid dynamics.

Table 5: Applications composing the benchmark suite.

Datasets and detailed instructions on how to configure, set-up, compile and run the applications (excluding HPL and HPCG) are available through the public git repository reported in Section 5.1. Regarding HPL and HPCG this benchmark suite relies on the source codes and the execution parameters complying with the rules for qualification for the Top500 List (top500.org).

5.3 Benchmark procedure

In this Section we provide the set of rules to be followed during the execution of the benchmarks, the information on how to execute the benchmarks and the reference values.

5.3.1 Benchmark rules

For all the scalable applications of the benchmark suite, (thus excluding HPL and HPCG) the following rules apply:

- No changes of the algorithms are allowed.
- Porting and optimization. Only minor modification of the GPU applications composing the benchmark suite are allowed. All source code modifications, including modifications of any of the provided third-party libraries, must be released along with the benchmark results. Modifications on CPU applications

are not allowed. In case changes are made, results must be provided with and without modification if this is possible.[7]

- When a code requires a library (beside BLAS, LAPACK and FFTW), the library name and version are shown in the README file. These libraries can be updated with: i) newer version of the libraries publicly available at the releasing date of this document, ii) other libraries only if publicly available at the releasing date of this document. The output of the run must be validated against the validation string reported in the Tables 6-11. In any of these cases, the version of the libraries, as well as their source code, and the releasing date to public availability, must be included in the provided results.
- There is no restriction on the use of the compile-line option. However, for each code, the compile-line option used must be reported along with the benchmark results.
- There is no restriction in terms of MPI Tasks/OpenMP threads/GPU threads, the candidate is invited to provide results for the best combination supported by the proposed system, but the nodes allocated by the benchmark run will be considered as completely dedicated to the application, regardless of the actual resources allocated within each node.
- For each code, the numerical results of each run must be checked according to the information provided in Tables 7-10 of this Chapter. If there's no agreement with the validation values, the performance result is considered not conforming.
- The Benchmark results must be projected for both CPU and GPU partitions.

For each application typically 2 different input decks are defined. Only one of them, the so-call *large input*, is expected to be used for the benchmark runs and the determination of time-to-solution (elapsed time) as clearly indicated in Tables 7-10. The *small input* is intended to be used for short test runs and is provided for the convenience of the Candidate. These *small inputs* are not relevant for evaluation and comparison of solutions provided by different Candidates. The Candidate should not assume CINECA be the owner of the provided benchmark codes. Copyright and licence conditions are defined in the code for each of the provided benchmarks.

Regarding HPL benchmarks the rules for qualification to TOP500/GREEN500 lists apply.

5.4 Benchmark execution

5.4.1 Retrieve and compilations of the codes

All the instructions to retrieve and compile the codes, excluding HPL and HPCG, will be available through the git repository.

5.4.2 Input parameters

Input datasets, excluding HPL and HPCG, will be available through the git repository.

5.4.3 Execution

For each code of the benchmark suite, we provide info about the input dataset, the reference output, the execution commands, the validation string, as well as the rules for the benchmark runs.

[7] If without modification the application cannot run on the benchmark system or offered system, then it does not make sense to provide results without modification.

HPL and HPCG	
Input dataset	see top500.org
Command line execution	see top500.org
Output	see top500.org
Extraction of the validation string	see top500.org
Extraction of timing	see top500.org

Table 6: Synthetic benchmark information

Quantum Espresso - CPU	
Input dataset name	https://gitlab.hpc.cineca.it/procurement/tier1-tecnopolo/benchmarks/-/blob/main/QuantumESPRESSO/LEONARDO-datacentric/inputfiles/csi.in?ref_type=heads
Minimum HPL performance of the proposed CPU partition to be used for executing the large input	150 TFlops
Command line execution	`mpirun -np $SLURM_NTASKS pw.x -nk $SLURM_NNODES -input csi.in > csi.out` https://gitlab.hpc.cineca.it/procurement/tier1-tecnopolo/benchmarks/-/blob/main/QuantumESPRESSO/LEONARDO-datacentric/outputfiles/submit.job?ref_type=heads
Output	https://gitlab.hpc.cineca.it/procurement/tier1-tecnopolo/benchmarks/-/blob/main/QuantumESPRESSO/LEONARDO-datacentric/outputfiles/csi.out?ref_type=heads
Extraction of the validation string	The reference value is the “total energy” printed at the last iteration in the standard output. It can be obtained with: cat <stdout> | grep "!" | tail -1 | awk '{print $5}' where <stdout> is the standard output of the simulation. The reference value is -6189.47(1) Ry using 9 scf iterations.

Extraction of timing	The reference value is the WALLTIME printed in the last lines of the standard output. It can be obtained with: cat <stdout> | grep "PWSCF " | tail -1 | awk '{print $5}' where <stdout> is the standard output of the simulation.

Table 7: Quantum Espresso – CPU Benchmark information

Quantum Espresso - GPU	
Input dataset name	https://gitlab.hpc.cineca.it/procurement/tier1-tecnopolo/benchmarks/-/tree/main/QuantumESPRESSO/LEONARDO-booster/inputfiles?ref_type=heads
Minimum HPL performance of the proposed GPU partition to be used for executing the large input	1 PFlops
Command line execution	`srun -n $SLURM_NTASKS --cpu-bind=cores --cpus-per-task=$SLURM_CPUS_PER_TASK pw.x -nk $(($SLURM_NTASKS/2)) -input csi.in > csi.out` ! The number of GPUs per node is the same of the number of MPI tasks per node https://gitlab.hpc.cineca.it/procurement/tier1-tecnopolo/benchmarks/-/blob/main/QuantumESPRESSO/LEONARDO-booster/outputfiles/submit.job?ref_type=heads
Output	https://gitlab.hpc.cineca.it/procurement/tier1-tecnopolo/benchmarks/-/blob/main/QuantumESPRESSO/LEONARDO-booster/outputfiles/csi.out?ref_type=heads
Extraction of the validation string	The reference value is the "total energy" printed at the last iteration in the standard output. It can be obtained with: cat <stdout> | grep "!" | tail -1 | awk '{print $5}' where <stdout> is the standard output of the simulation. The reference value is -6189.4(7) Ry using 51 scf iterations.

Extraction of timing	The reference value is the WALLTIME printed in the last lines of the standard output. It can be obtained with: cat <stdout> | grep "PWSCF " | tail -1 | awk '{print $5}' where <stdout> is the standard output of the simulation.

Table 8: Quantum Espresso – GPU Benchmark information

ICON	
Input dataset name	ICON–2I: for more information see gitlab repository
Minimum HPL performance of the proposed CPU partition to be used for executing the large input	75 TFlops
Command line execution	sbatch run_cineca.sh
Output	<MAIN_PATH>/RUN/projects/test_cineca/log/icon.2021102800.<SLURM_JOBID>.log HP <MAIN_PATH>/RUN/projects/test_cineca/log/timing.log
Extraction of the validation string	In >MAIN_PATH>/RUN/projects/test_cineca/run.2021102800: ~$ grib_get -p max,min,average test_cineca_2021102800_plevs_z+0060000.grb 18.9077 -21.4731 -3.23035 13.7041 -23.9922 -1.84731 10.5442 -23.7458 -2.38355 19.4833 -10.1109 4.10386 18.5083 -12.73 2.41878 19.2502 -15.0027 1.54569 261.598 250.093 256.868 278.945 268.957 273.654 287.297 274.36 281.609 0.00314712 2.22259e-05 0.000738327 0.00749114 6.98823e-05 0.00229181 0.0107859 0.000346031 0.0045783 56799.3 55320.5 56212 30991 30155.5 30555.5 15465.7 14498.7 15038.5 Tollerance <= 1%
Extraction of timing	In the file timing.log find the lines “start_ICON” and “end_ICON”, there is the starting and ending time of the model.

Table 9: ICON-CPU Benchmark information

PLUTO4.4	
Input dataset name	https://gitlab.hpc.cineca.it/procurement/tier1-tecnopolo/benchmarks/-/tree/main/PLUTO4.4/Test_Problems/MHD/Blast/Test_EnEff_Large
Minimum HPL performance of the proposed CPU partition to be used for executing the large input	300 TFlops
Command line execution	mpirun ./pluto -maxsteps 80 -no-write
Output	The log files are located in the working directory.
Extraction of the validation string	The reference value is the “dt” printed at the last iteration in the “pluto.0.log file” located in the working directory. It can be obtained with: cat pluto.0.log |grep step:80 | tail -1 | awk '{print $7}' | sed "s/;//" The reference value is 2.6068e-06 by using 80 integration steps
Extraction of timing	The elapsed time to initialize the system and perform 80 integration steps can be obtained with: cat pluto.0.log |grep Elapsed | tail -1 | awk '{print $4}'

Table 10: PLUTO4.4-CPU Benchmark information

5.5 Benchmark analysis report

To assess the benchmark results, a benchmark analysis report provided by the Candidate is required, as better detailed in the tendering document “Disciplinare di gara” art. 8.5.

6 Maintenance and infrastructure availability

The Offer for the procured infrastructure must include a maintenance and support service that ensures high availability, as described in Section 6.4, and stability of the procured infrastructure. In the following with the term Supplier, it is referred the Candidate awarded for this procurement.

6.1 Maintenance and support requirements

Req.	Description	Category
6.1-1	*Maintenance and support duration* The Supplier will offer maintenance and support of the offered infrastructure for five years.	MRQ
6.1-2	*Maintenance and support coverage* The maintenance and support will cover all key hardware and software (incl. firmware and all offered programming environment software) components of the procured infrastructure. This includes all infrastructure component (e.g., racks and power supplies) and network components except for those components provided by CINECA. The Candidate will describe all components not covered by the system maintenance and support. The customer maintenance and support times may be restricted to normal working hours. At least the standard working hours on all working days (excluding weekends and Italian public holidays), i.e., 5×8, will be covered. The Supplier must ensure the provision of the maintenance and support services, even if there is a dispute with CINECA.	MRQ
6.1-3	*Special software support coverage* Software, whose malfunction could harm the system stability and hardware health, will be supported by the Supplier. For example, if power capping techniques integrated in the workload manager are used for system operation, these workload manager capabilities must be fully supported by the Supplier.	TRQ
6.1-4	*Reaction times* The Supplier guarantees an appropriate reaction time upon hardware and software issues.	MRQ
6.1-5	*On-site stock* The Supplier will populate and maintain an on-site stock in CINECA's facility to ensure the availability of replacement parts, especially for components whose loss significantly affects system availability or utilization. However, the Supplier may rely on a facility other than CINECA's for stocking spare parts near CINECA's data centre (2-3 hours). The Candidate will provide a list of the intended spare parts included in the on-site stock.	MRQ
6.1-6	*On-site support* The Supplier will include one full time equivalent (FTE) position, based on one or multiple qualified persons, to ensure permanent on-site system support	TRQ

	during working hours. The on-site personnel will support CINECA primarily in failure analysis, hardware support (including spare and replacement part logistics if necessary) and software support for cluster management and system management software. If the FTE is based on multiple on-site persons, a reasonable team size must not be exceeded, and an appropriate coverage of the relevant support fields must always be ensured.	
6.1-7	*Preventive maintenance and early errors' detection actions:* The Supplier will perform preventive maintenance and early detection of errors actions to replace components that are likely to fail soon. Example of possible preventive maintenance actions are the replacement of components (disks, networking equipment) based on error counter information prior to the point where system operation is impacted and the replacement of memory components exhibiting high single-bit error rates prior to the occurrence of a (fatal) double-bit error. The Candidate will document these actions included in the Offer.	MRQ
6.1-8	*Data deletion* The Supplier will ensure that all client data stored on any, user accessible, non-volatile storage component (incl. HDD) are deleted when components are taken off-site as part of the system maintenance. Data deletion may occur off-site but must conform to common data protection guidelines. Alternative means that ensure the confidentiality of the data stored on non-volatile storage components (e.g., destruction by the customer) may be proposed.	TRQ
6.1-9	*Serviceability constraints* The Candidate will document all serviceability constraints affecting system availability. Examples of such serviceability constraints include sibling nodes that must be taken offline for a node replacement or rack components that need to be taken out of service for network servicing.	MRQ
6.1-10	*Escalation management process* The Supplier will provide an escalation management process to manage problems priority and critical/non-critical issues.	MRQ
6.1-11	*Regular maintenance and support meetings* CINECA intends to host regular (up to eight meetings per year) face-to-face meetings, to discuss the state of the installation and address any problems. The Supplier will ensure the availability of the necessary (travel) funds required for the attendance of the key support personnel in these meetings.	MRQ
6.1-12	*Pre-production qualification acceptance* The Supplier will declare whether the system design and maintenance concept enable the system to pass the acceptance tests proposed in Chapter 7.	MRQ
6.1-13	*Responsibilities and roles* The Candidate will describe the roles and responsibilities of all parties involved during system operation in the form of a RACI- (Responsible, Accountable, Consulted, Informed)-model.	MRQ
6.1-14	*Security patches and software updates*	MRQ

	The Supplier will make security patches for all supported software components available for installation in an adequate period following the release of the component by the vendor. Availability of Security Updates: All offered system components must receive security updates throughout the lifetime of the system. In case of (disclosed) major vulnerabilities, especially those allowing privilege escalation, the supplier will ensure the immediate collaboration to place mitigations and/or patches. Furthermore, the Supplier will ensure the release and application 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. Besides software updates the supplier will provide tested “recipes" for update processes, to ensure the whole system stability and reduce the possible services downtime or degradation.	
6.1-15	*Maintenance and support service regulation* This service is considered "*a corpo*"[8] and applies to all the products that are acquired by CINECA as part of this procurement. The maintenance and support 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 support 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 responds of 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, in particular those issued by the UNI and the CEI (Italian Electro technical Committee). 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. In particular, the Supplier must issue a Declaration of Conformity to Law no. 46 - "Safety Standards for Installations".	MRQ
6.1-16	*Maintenance periodic reporting* 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. Candidate will describe this periodic reporting in the Offer.	MRQ
6.1-17	*CINECA relation with the Supplier* Upon awarding the contract and for the conclusion of the contract, the Supplier, must nominate a representative to manage all relations with CINECA. The Supplier's representative is the point of contact for any issues that CINECA considers unresolved within the normal relationship with the Supplier (sales manager, technician, call centre, etc.). The Supplier's representative will participate, if required, in regular meetings along with its representatives to	MRQ

[8] Meaning the service included is to be considered as a complete package.

	update the status of the contract and to share any corrective action needed to comply with the contract. The representative will also be responsible for providing CINECA with all the documentation necessary for correct access and the use of maintenance and support service (access credentials, etc.). The Supplier's representative must have appropriate professional qualifications and must be available before the supply contract is signed.	
6.1-18	*Documentation requirements* The Candidate will describe the support workflow for software and hardware failures, including information about replacement part logistics and SLAs.	MRQ

6.2 Tier-1 Specialistic support

Req.	Description	Category
6.2-1	*Specialistic support for the storage infrastructure installation* The installation of the storage infrastructure described in Section 4.6 is critical for the success of the project. The Supplier must provide adequate support for its architecture design, deployment and tuning in terms of days of support. At least 15 days of specialistic support must be provided in the Offer.	MRQ
6.2-2	*Specialistic support for the storage infrastructure maintenance* The Supplier must provide adequate support for the storage infrastructure described in Section 4.6 during the entire lifetime of the System. At least 7 days per year of specialistic support must be provided in the Offer.	MRQ

6.3 Licenses

Req.	Description	Category
6.3-1	*Licenses* Where applicable, the Candidate must provide licenses for all offered software for the complete duration of the maintenance and support time frame. The software packages provided by CINECA are excluded from this.	MRQ
6.3-2	*Licenses' list* Candidate must provide the complete list of all applicable licenses provided with their quantities.	MRQ

6.4 Infrastructure availability

The procured infrastructure must seek the highest availability to the end users. CINECA will report on monthly availability for the server nodes (*Availability_c)* and for the storage (*Availability_s)*. They are calculated independently with the following formulas:

$$Availability\_c = \frac{\sum_i^N a_i}{\Delta T \cdot N - \sum_m \sum_i^{Nm} d_i}$$

Where:

- a_i is the availability of each single node "i" (front-end and compute nodes). Availability here means that the node is up and running, reachable by the users directly or through the WLM slurm.
- ΔT is the time interval considered (i.e., a month expressed in hours).
- N is the total number of compute nodes.
- m denotes a scheduled maintenance intervention.
- N_m is the number of nodes involved in the scheduled maintenance "m".
- d_i is the time that node "i" spent in scheduled maintenance "m". For each node involved in maintenance, the time d_i starts after action of system administrator that manually drains the node.

$$Availability\_s = \frac{\Delta T_a}{\Delta T - \Delta T_m}$$

Where:
- ΔT_a is the time when the storage infrastructure is available to the users.
- ΔT is the time interval considered (i.e., a month expressed in hours).
- ΔT_m is the time when the storage infrastructure is under scheduled maintenance and therefore unavailable to the users.
- m denotes a scheduled maintenance intervention.

It is planned to conduct planned maintenance[9] interventions for up to 7 days per year, and every month for a duration of 8 hours.

Req.	Description	Category
6.4-1	*Targeted monthly availability* The design of the procured infrastructure architecture and the maintenance service must aim for a monthly availability of 95%.	TRQ
6.4-2	*Minimum monthly availability* The design of the procured infrastructure architecture and the maintenance service must aim for a minimum monthly availability of 85%	MRQ

[9] Therefore, these interventions do not contribute for the calculation of the total Availability_s and Availability_c percentages as expressed in the respective formulas.

7 Installation and acceptance

7.1 Installation time schedule and project management

The Proposal for the procured infrastructure must include the necessary planning and project management resources for the installation of the system. In the following the term Supplier refers to the Candidate awarded for this procurement.

7.1.1 System Installation

Req.	Description	Category
7.1.1-1	*Project Management* The Supplier will provide project management resources for the system installation.	MRQ
7.1.1-2	*Benchmarking Support* The Supplier will provide expert support for the benchmarking of the system. The optimization of benchmark performance, rule conforming execution and the submission of the results to the official lists will be performed by the Supplier. CINECA plans to include the system in the TOP500 and GREEN500 lists.	MRQ
7.1.1-3	*Installation Time* The timeline of delivery, installation, acceptance and start of operations are detailed in the tendering document “Schema di Contratto”, art. 4.	MRQ
7.1.1-4	*Best practices and security* • During installation the system will be accessible only through CINECA VPNs or Bastion hosts. • All the administrative user’s passwords of all the installed equipment must be changed from their default values in the very early stages of deployment. • The used passwords must be adequately strong. • The passwords of admin users will be disclosed only to the essential staff. • Every person that doesn’t need to know a certain password to operate will be allowed to admin accounts by passwordless and/or ACL mechanisms. • Every person that doesn’t need admin access to a given equipment will be allowed with unprivileged user. • The secrets used for the management cluster and services must be different from the ones used for regular production nodes. • The secrets must not contain common substrings. The secret databases must be protected by adequate passwords.	MRQ

	• Directory services must not expose passwords (querable) and all the secrets must be stored in “hashed" form. • The Service node’s cluster networks must be segregated and only the essential services must be “published” to the regular nodes cluster networks. • The out-of-band management networks must be segregated and connected only to the management Service nodes. • Early access users, benchmarkers and all the people not involved in the installation process can access the system only via login nodes. • No shell enabling access to the Service or management nodes will be allowed through connections coming from regular cluster nodes. • A clear operational recipe must be made available to change the passwords and the secrets of all the services and nodes, as well as adequate automated helper procedures. • All the inactive and unused accounts and related secrets must be closed and/or deleted immediately. • All data in storage resources must be protected and disclosed only to the essential people. • In case of teams/subcontractors/main contractor handover during installation all the secrets and accounts must be contextually changed. • During acceptance handover all the secrets must be changed. • All the system logs during installation must be collected and aggregated using effective techniques to allow queries and forensic analysis.	

7.1.2 Supply and installation project

Req.	Description	Category
7.1.2-1	Installation project plan The Supplier is responsible for creating a supply and installation project for the procured components. This project must detail the delivery times of the various parts of the system, including any downtime that might affect the operation of CINECA infrastructure. The project must include: • Details of the offered configuration and integration in the CINECA computing system architecture, including setup and interconnection schemes. • Details of hardware and software installation plan, configuration and optimization of the components and partitions. • Details on how the interaction with CINECA personnel is foreseen. • Implementation plan for procured infrastructure acceptance (see Section 7.2). All interactions with CINECA staff, all training activities, as well as the documentation produced within the project, can be in Italian or English.	MRQ

7.1.2-2	*Time schedule* The Candidate will describe the time schedule for the system installation in detail and in terms of a GANTT chart. The time schedule will provide expected dates for the production and delivery of system components, installation, bring-up and acceptance of the procured infrastructure.	MRQ
7.1.2-3	*Project risks* The Candidate will provide a list of risks that could negatively affect the installation and early operation of the procured system. For each risk, the Candidate will give an indication of likeliness, provide a description of the expected impact and risk mitigation measures that will be implemented as part of the contract.	MRQ
7.1.2-4	*Responsibilities and roles* The Candidate will describe the roles and responsibilities of all parties involved during system installation and early operation in the form of a RACI-(Responsible, Accountable, Consulted, Informed)-model.	MRQ

7.2 Acceptance procedure

7.2.1 Documentation requirement

The Candidate will declare whether he agrees to the acceptance tests defined in this Section (with details specified in accordance with the Offer). Please note that all acceptance tests verifying committed functionality and performance values included in the Offer are non-negotiable.

7.2.2 Execution of acceptance tests

All acceptance tests will be performed by the Supplier together with, or directly informing, CINECA staff.

For the acceptance procedure and the verification of the committed benchmark results, the following rules apply:

Req.	Description	Category
7.2.2-1	*Acceptance rules* The Compute Nodes will run the full operating system stack. The latest security updates will be installed. The system may not be reconfigured for different benchmarks unless this process is fully integrated in the workload manager (WLM) and will be available at user-level during the production phase of the procured infrastructure. The benchmark runs will be performed using the offered compiler suite and MPI implementation. If the Offer includes multiple compiler suites or MPI implementations, the Supplier may choose a different	MRQ

	combination for each benchmark. All tests will be performed using the production WLM.	

7.2.3 Provisional acceptance tests

7.2.3.1 Hardware checklist

Req.	Description	Category
7.2.3.1-1	*Hardware checklist* Completeness and consistency of the delivered and installed hardware will be checked against the Offer.	MRQ
7.2.3.1-2	*Failure thresholds* The thresholds for defective components described in the following must not be exceeded. For equipment not listed below no fatal deficiencies may exist for the provisional acceptance test to be passed. • Compute nodes: less than 2% of nodes may be dysfunctional. • Frontend nodes: all nodes must be functional. • Service nodes: all nodes must be functional. • Ethernet links: less than 0.1% of the links may be dysfunctional. • High speed interconnect links: less than 0.1% of the links may be dysfunctional.	MRQ

7.2.3.2 Software Checklist

Req.	Description	Category
7.2.3.2-1	*Software checklist* Completeness and consistency of the delivered and installed software will be checked against the Offer. All components must be installed for the test to be passed.	MRQ

7.2.3.3 Functional Tests

Req.	Description	Category
7.2.3.3-1	*Acceptance plan* The Supplier - in agreement with CINECA - will provide an acceptance plan to verify, with a series of functional tests, the suitability of the components against the expected performance level reported in the Offer. All the components (hardware and software) must be checked against their performance level as described in the Offer. Components may be grouped together and verified with a single test in accordance with CINECA staff.	MRQ

The list of the functional tests included in the acceptance plan will ultimately depend on the system design and the Offer. For the benefit of the reader a non-exhaustive list of tests may include:

- Verification of the power and cooling infrastructure.
- Verification of power management system.
- Verification of health checks and monitoring in accordance with the Offer.
- Verification of cluster management including management network (collection of metrics, redundancy, node reinstallation and configuration).
- Verification of data network (reachability of compute nodes and service nodes, bandwidth, and latency performance).
- Verification of the stability of the system software, firmware and hardware (component stress tests, see Section 7.2.3.4).
- Verification of single functional component performance level (node, group of nodes, rack, group of racks).
- Verification of system level performance (Benchmark suite, IO partition, see Section 7.2.3.5-7.2.3.8),
- Verification of software specifications and offered features.
- Verification of other commitments made in the Offer.

7.2.3.4 Stress tests

The following synthetic tests are typically performed to stress the hardware components. In case of failure, the faulty components must be replaced, and the test will be restarted on the affected component. All these tests must be passed successfully.

- A local, optimized HPL will be run on each single node, in parallel for 30 minutes without failure.
- A memory stress test will be performed on the system. CINECA proposes a modified STREAM version which uses >95% of the system memory for this test. The Supplier may suggest an appropriate alternative tool.

7.2.3.5 Application and Synthetic Benchmarks

The synthetic and application benchmarks included in the benchmark suite (see Section 5.2.3) will be executed with the baseline values as provided by the Supplier in the Offer. For the test to be passed, all committed benchmark results must be achieved within a 3% of relative tolerance.

7.2.3.6 Rules for scalable application Benchmarks

The Supplier will dedicate a folder in the system to collect all the input files (application, libraries execution commands, and output files) required to verify and in case replicate the results.

7.2.3.7 Rules for HPL and HPCG Benchmarks

The HPL and HPCG benchmark will be executed on the CPU and GPU partition separately, according to the TOP500 list rules. During the HPL benchmark, the power consumption will be measured according to the GREEN500 run rules. The performance of HPL and HPCG benchmarks must confirm the values committed by the Supplier within the tolerance reported in Section 7.2.3.5.

7.2.3.8 I/O Performance

The I/O performance of the scratch file system will be measured using IOR or alternative benchmark application. The committed I/O performance (see Section 4.6) must be achieved within the relative tolerance reported in Section 7.2.3.5.

7.2.4 Pre-production qualification

The stability of the system will be tested over the course of one month under near-production conditions. For this purpose, the system will be filled with an arbitrary, well behaving, workload (i.e., a workload that does not trigger out-of-memory situations or other software exceptions). In this phase an early access can be provided to selected and experienced users.

Req.	Description	Category
7.2.4-1	*Pre-production availability* The Supplier will replace failed components and tune the infrastructure configuration during the pre-production phase to reach at least one week with an availability - as descripted in Section 6.4 - that must result 85% or above.	MRQ

7.2.5 Final acceptance

The final acceptance will validate the proper functioning of the entire system after the preproduction qualification period.

Il RUP
Sanzio Bassini