Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
DIPARTIMENTO PER LA TRASFORMAZIONE DIGITALE

CINECA

Capitolato Tecnico

Tier1 Technical Specifications

The following specifications result from the work done
in collaboration with ACN

1 Project goals

1.1 Introduction

Italian prominence in the HPC field has taken a big step forward in recent years. CINECA is hosting Leonardo, that debuted 4th in the top500.org November 2022 list, and confirmed its position in the subsequent June 2023 list. While providing the most capable supecomputing system is still a priority for CINECA and its partners, it has become essential to populate the HPC ecosystem with more specialized infrastructures that will better address the needs of specific communities. Some examples of the latter are the system dedicated to the EUROfusion (Pitagora) with a clear focus in providing their HPC environment for the community in the next 6 years. Another example is the Tier-1 for the AIM (Agenzia Italia Meteo), INAF (Istituto Nazionale di Astrofisica), CNR (Consiglio Nazionale delle Ricerche) and UNIBO (University of Bologna). The implementation of this system addressed the needs of the stakeholders of leveraging a unified computing infrastructure, in order to exploit the best flexibility and elasticity of resources, while being characterized by different and dedicated storage infrastructures: high IOPS for AIM, high bandwidth for CNR and UNIBO, and high capacity for INAF (especially for the SKA - Square Kilometre Array - use cases).

Analogously CINECA, in strict collaboration with National Cybersecurity Authority agency (ACN), CNR and ICSC (Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing) , will define in this document the details of the Tier-1 infrastructure that will be deployed in the new data centre in Naples in the renovated area of San Giovanni a Teduccio. The geo-distribution of the HPC and cloud infrastructures improves the resiliency of the compute and storage services, providing the capability to host critical services or provide disaster recovery capability for critical workloads.

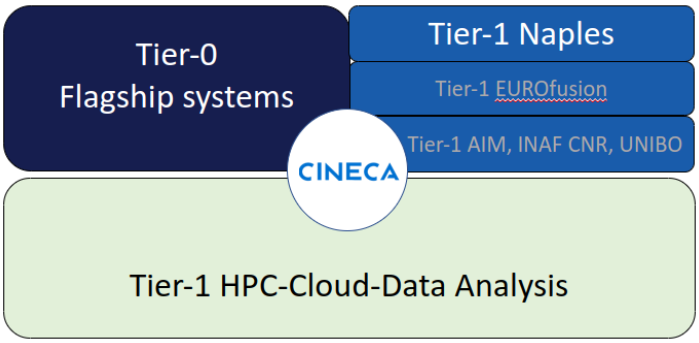

Figure 1. High level scheme of the HPC ecosystem

1.2 Goal of the procurement

This project aims to procure a hyper convergent HPC-Cloud infrastructure able to implement the following main characteristics:

- Provide a Tier-1 class computing HPC system, including a conventional (CPU) and accelerated (GPU) partition, with a clear focus in AI and traditional HPC workloads.
- Provide state-of-the-art parallel & multiprotocol storages, able to serve the Tier-1 partitions. It is paramount that workloads will run with adequate quality-of-service to comply with the service level agreements imposed by this critical national cybersecurity service.
- Provide an interconnection able to glue together the computing and storage partitions, with the flexibility to host Cloud and HPC workloads.

A primer in procuring and managing a HPC and Cloud convergent system was carried out by CINECA with the Galileo100 system[1] resulting from the Fenix-ICEI project (https://fenix-ri.eu/).

The procured resources, integrated in the newly built CINECA data centre, will represent a new step forward to tackle the new challenges in multiple scientific fields such as cybersecurity, material properties and design, and many others. In particular, this supercomputing infrastructure will play an essential role in addressing:

- National LLM model tuning and execution of inference workloads,
- Post-quantum Encryption,
- Advanced Attack and Threat simulations,
- Predictive risk analysis,
- High-performance detection,
- Confidential computing of sensitive data.
- Accelerated computing for material design,
- Earth simulations and environmental risk detection,
- Digital twins.

Moreover, this procurement targets the provisioning of border apparatus, firewall and network distribution switches to enable a capable external connectivity towards the systems hosted in the Naples data centre.

[1] https://ted.europa.eu/IT/notice/-/detail/218289-2019 and https://www.hpc.cineca.it/systems/hardware/galileo100/

2 Document definitions

2.1 Tendering procedure 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

2.2 Categories of requirements

The requirements and features within the documents are categorized 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 center 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 center integration of the offered solution into CINECA infrastructure.

Table 2: Categories of requirements

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 Project 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.

3.2.2 Time schedule

The procurer targets to start the production phase of the procured infrastructure according to the timeline defined in art. 4 of the tendering document "*Schema di Contratto*".

4 Data center facility

The new equipment will be hosted in CINECA data centre located in San Giovanni a Teduccio, c/o Polo Est dell'Università degli Studi Federico II, 80146 Napoli (NA), Italy. To set the logistic and data centre integration limits that the offers must comply, in the following Sections the data center specifications and MEP infrastructure design are reported.

4.1 Facility description

Errore. L'origine riferimento non è stata trovata. shows what is currently available in the Computing Data Hall in the Naples data centre.

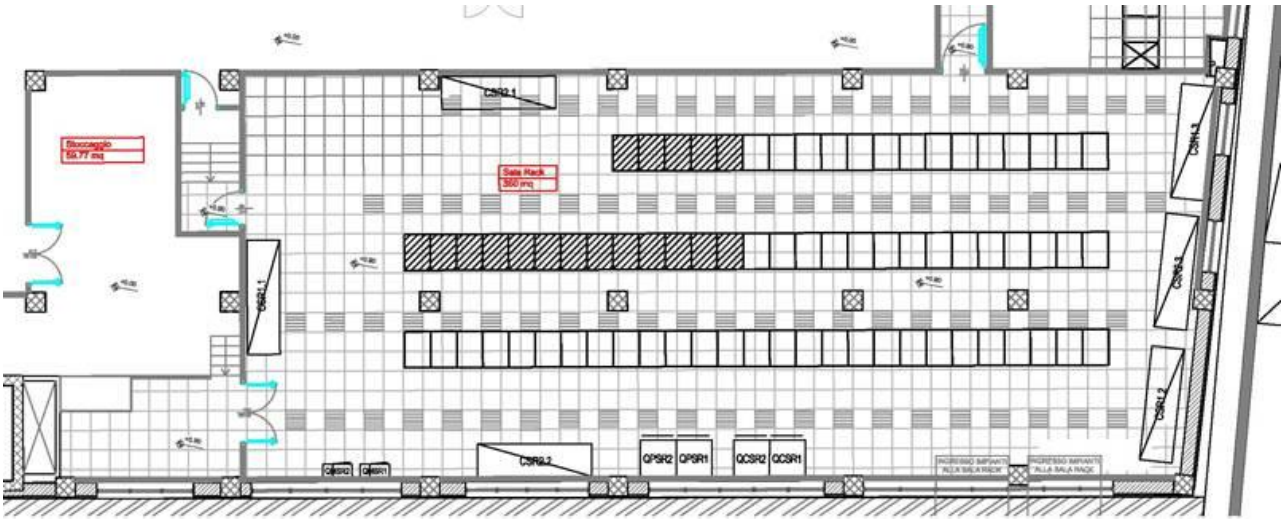

Figure 2. Data center floor plan. A tentative layout of the system, showing available rack's capacity and possible expansion in white boxes and grey boxes respectively. Square chess-like boxes denote pillars position. The Data Hall is empty. Perimeter chillers are also reported in the picture. Tile dimension is 600x600 mm.

4.2 Data center specifications

The Computing Data Hall is being adapted for the scope of hosting the Tier-1 system subject of this procurement procedure. The following specifications are part of the works execution plan of the data centre. Regarding the computing Data Hall the following specifications apply:

Req.	Description	Category
4.2-1	*Dedicated Data Hall*	DCS

	The offered solution infrastructure must be in installed the Data Hall shown in **Errore. L'origine riferimento non è stata trovata.**. In terms of whitespace available, the Data Hall is empty with a surface available in the order of 300 m^2.	
4.2-2	*Raised floor details* The data hall is equipped with a raised floor with height of 900 mm.	DCS
4.2-3	*Raised floor load* The raised floor is reinforced to host DLC racks. The maximum expected load is 30 kN/m^2 and 11 kN single point load.	DCS
4.2-4	*Rack maximum height* The cable tray is planned at 220 cm of height (see Figure 3). A maximum height per rack of 210 cm is envisioned.	DCS
4.2-6	*Room humidity* The offered infrastructure must comply with a relative humidity in the interval 20-60%.	DCS

A representation of the data center vertical section is represented in Figure 3.

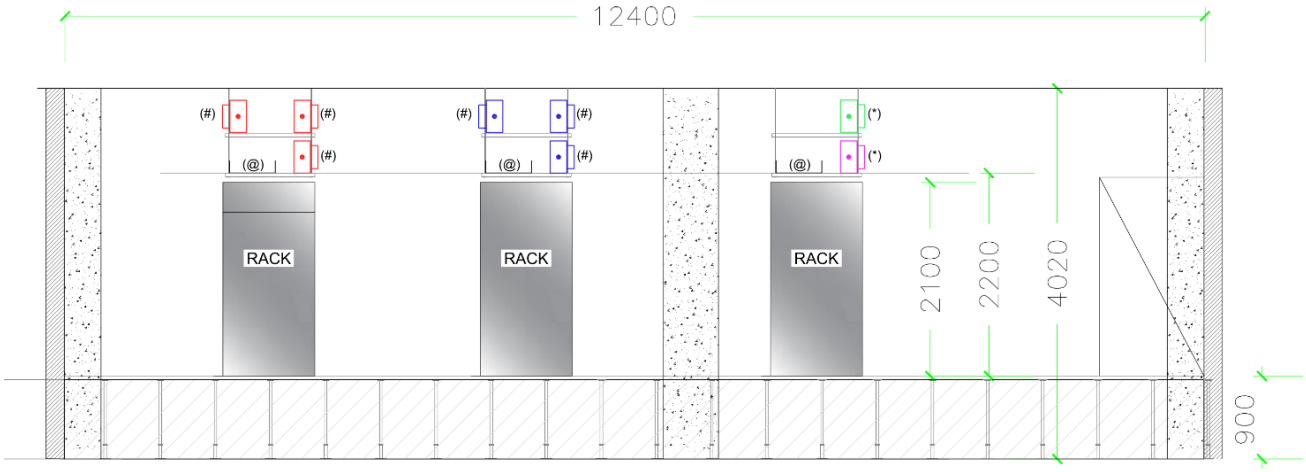

Figure 3. Section of the Data Hall with reference height of raised floor and cable trays.

4.3 Electrical infrastructure

Req.	Description	Category
4.3-1	*Data hall power supply* The Data Hall has an electrical power supply characterized by 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 (also during the acceptance phase testing) of 1640 kW IT, of which 1.400 kW DLC with cooling water temperature at 36°C and 240 kW air cooling. Instead, for cooling water temperature of 32°C the electrical data are modified as follows: 1240 kW IT, of which 1.000 kW DLC and 240 kW air cooling.	DCS
4.3-2	*Electric load distribution* The IT load installed in the Data Hall can supply power for two rack rows with no UPS (the two rows at the top of Figure 2) and one rack row under UPS (the row at the bottom of Figure 2). 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. 4.3-1.	DCS
4.3-3	*Continual power system* Of the power IT supplied and detailed in Req. 4.3-1, 360 kW is available under UPS.	DCS

4.4 Cooling infrastructure

Req.	Description	Category
4.4-1	*Cooling infrastructure* The cooling infrastructure of the data center produces tempered water and chilled water. The tempered water is dedicated for DLC compute nodes and the	DCS

	chilled water is used for the air conditioning of the data halls. The offered solution must comply with the limits provided in Req. 4.4-2. and 4.4-3.	
4.4-2	*Liquid cooling* Tempered water will be used for direct cooling of the racks. The maximum cooling power is up to 1.400 kW DLC with cooling water temperature at 36°C, instead is 1.000 kW DLC with cooling water temperature at 32°C.	DCS
4.4-3	*Air cooling* The air-cooling infrastructure can provide 240 kW, considering redundancy, available for the offered solution.	DCS
4.4-4	*Flow rate* The maximum flow rate available for the Data Hall is 4400 l/min (also during the acceptance phase testing).	DCS

4.5 Data Hall MEP layout

Regarding the layout of the room, both in terms of air flows and water collectors, it is specified that the efficiency of the current plant is strongly linked to the current arrangement of the racks in the room,

radical layout changes could result in a strong reduction of the cooling efficiency despite having a considerable power.

The conveyance of cooling fluids is an essential element that has a significant impact on the overall effectiveness of the system.

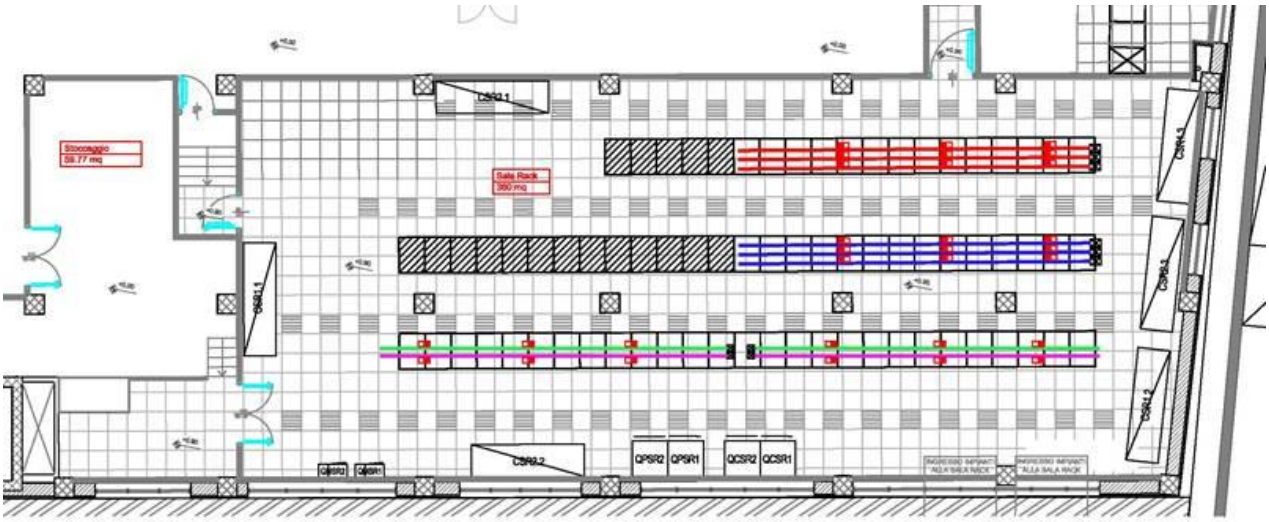

Figure 4. Power distribution layout of Data Hall. Red and blues represent electric branches with no UPS, green and purple represent lines under UPS.

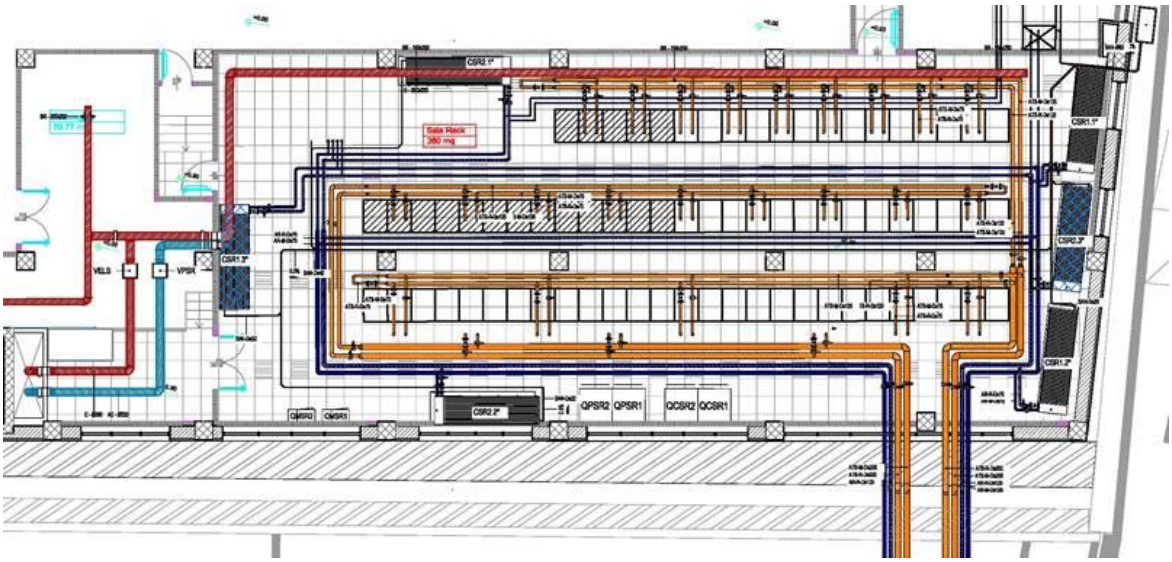

Figure 5. Cooling distribution layout of the Data Hall. Orange indicates the tempered water loops and blue the chilled water loops.

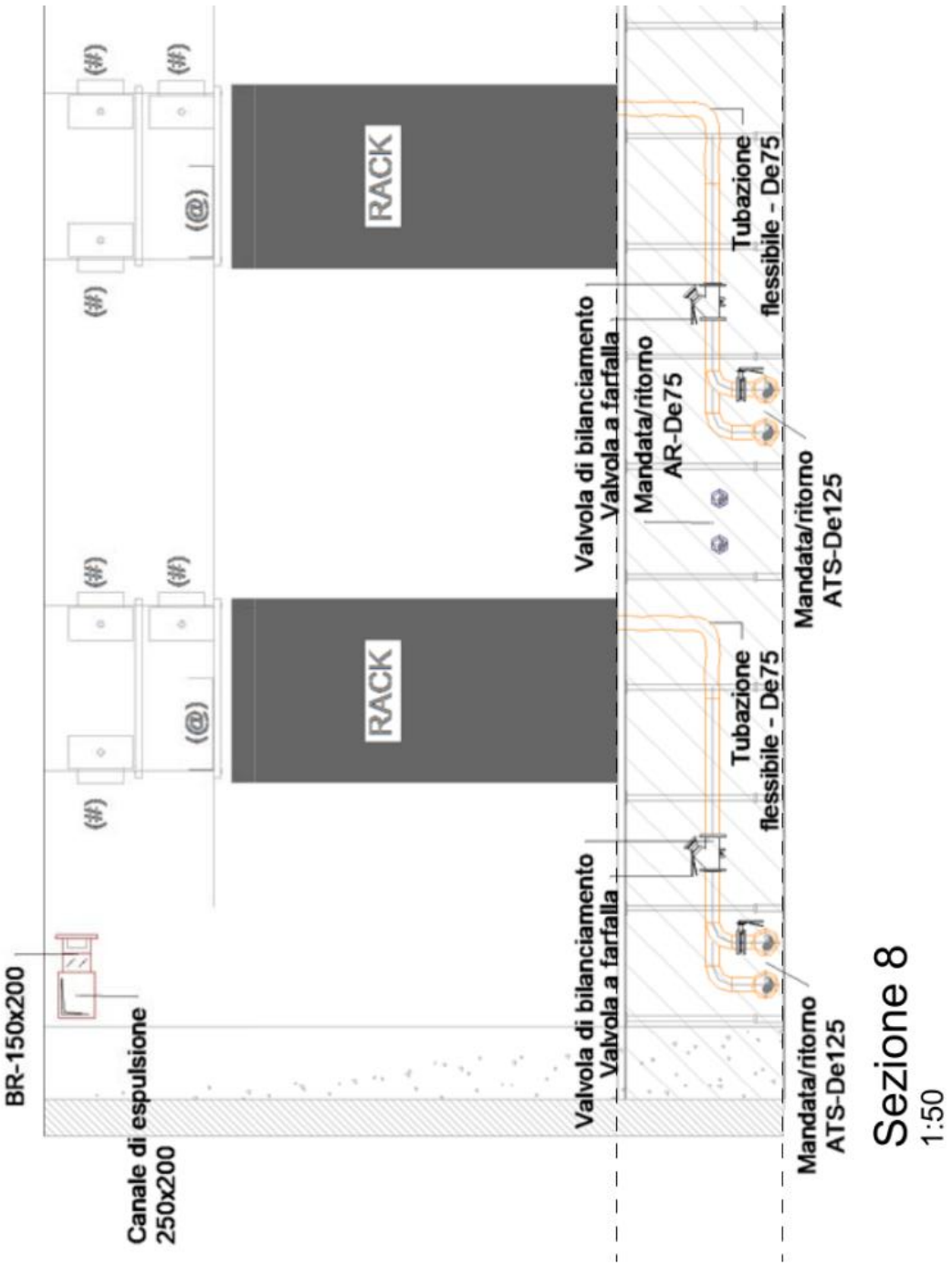

Figure 6. Section of the cooling distribution under the raised floor.

5 Data center network

5.1 Minimum technical requirements

The supplier must guarantee the supply of equipment and all the necessary components, guaranteeing the minimum general and specific requirements indicated in the following paragraphs.

5.1.1 Minimum general requirements

Req.	Description	Category
5.1.1-1	*Supply* The equipment supplied must be new (i.e. not used and/or remanufactured) and made with components produced by leading manufacturers worldwide. The offered data center network solution must be complete. All hardware and software components required to deliver the services and functionalities described in this chapter must be included in the Offer.	MRQ
5.1.1-2	*Manufacturing* All components should be produced by the equipment manufacturer (e.g. transceiver, etc.).	TRQ
5.1.1-3	*Software version* The software components integrated into the hardware resources must be supplied to the latest stable release.	MRQ
5.1.1-4	*Installation & configuration services* The Offer must include installation and configuration services.	MRQ
5.1.1-5	*Licenses* The Offer must include any software licenses required for a duration of at least 60 months.	MRQ
5.1.1-6	*Assistance and maintenance*	MRQ

	The Offer must include assistance and maintenance services for all hardware and software components supplied for a period of 60 months.	
5.1.1-7	*Power redundancy* The equipment must provide redundant and hot-swappable power supplies.	MRQ
5.1.1-8	*Rack mounted* All the network equipment and appliances (Router, Firewall, Management Platforms etc.) offered must be installable in standard 19" racks to be part of the supply and must also be powered by AC 230V/50Hz compatible with the specification in 4.3.	MRQ
5.1.1-9	*Cabling* The offer must include the realization and the supply of all the structured cabling, both copper and optical, useful for the connection of the network equipment: the data center border firewalls (5.2) the data center border routers (5.3), the data center management network (5.4) and the data center wireless network (5.5).	MRQ
5.1.1-10	*Cable certifications* All the required fiber optic and copper cabling must be certified and labeled, attested on patch panels provided by the Supplier and installed inside the racks involved; the requirement also applies to interfaces derived using breakout cables (e.g. to derive 100 Gb/s ports from 400 Gb/s ports). The certifications of the wiring must be delivered to CINECA for the purpose of final testing.	MRQ
5.1.1-11	*Integrated offered solution* The offer must be of a "turnkey" type, i.e. include everything necessary for the implementation of the proposed architecture, like: HW components, software components, various cables, fiber optic patches, any sleds and materials necessary for connection to both the data network and the	MRQ

	electricity network (i.e. fiber optic cables, breakout cables, UTP cables, DAC cables, etc.) and for installation inside racks, etc.	
5.1.1-12	*Airflow* Every equipment supplied and installed in the data center must include a heat dissipation system based on front-to-back airflow.	MRQ

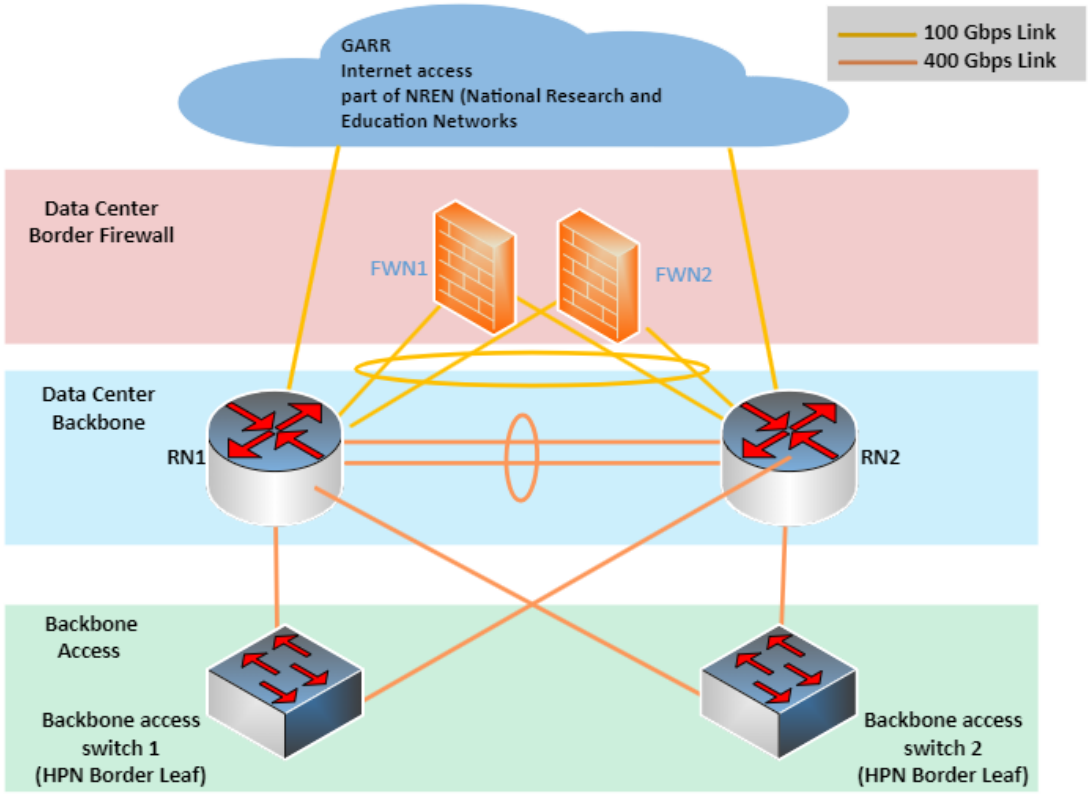

Figure 7: General Backbone Network Diagram

5.2 Data center Border Firewall

The Offer must include a Firewall instance. "Firewall instance" means perimeter security equipment complying at least with the minimum mandatory characteristics described in the following documentation. The firewall instance must be provided as a "black box" consisting of a certain number of firewall appliances (and therefore not necessarily two as shown in Figure 7) for traffic analysis and any

other relevant devices. The description of the architecture of the firewall instance must be provided in the Offer in order to allow the evaluation of the proposals.

Since in the Technopole Data Center in Bologna CINECA uses a firewall solution based on the Checkpoint Maestro architecture as a border firewall, it is requested that a solution based on the same technology be proposed for the firewall solution for the Tier 1 of the Naples data center in order to have completely compatible and homogeneous platforms from the point of view of configuration, as well as simplifying overall management and maintenance.

5.2.1 Border Firewall requirement

Req.	Description	Category
5.2.1-1	*Performance level* The firewall instance must provide the following performance in operation: • At least 100 Gb/s of traffic protected by the firewall instance with all devices working and with all the features provided among those listed from 5.2.1-4 to 5.2.1-11 simultaneously active, accepting a maximum degradation of no more than 50% of the protected traffic in the event of failure of no more than 50% of the equipment. The total minimum traffic protected by the firewall must be at least 60 Gb/s of IPsec VPN (IPv4) traffic or, alternatively, at least 50 Gb/s of TLS/SSL traffic to be inspected. • It must support at least 800 concurrent client-to-site IPsec and SSL VPN connections. • It must support at least 400 concurrent active IPsec site-to-site VPN connections. The inspection of TLS/SSL traffic must be able to be carried out both for inbound traffic (to protected company resources, through the installation of their X509 digital certificates and related private keys on the firewall) and for outbound traffic (from protected company resources, e.g. laptops to the Internet).	MRQ
5.2.1-2	*High availability* The firewall instance must consist of a cluster with High Availability (HA) characteristics; in general, in the event of failure of one of the components of the firewalls that constitute the High Reliability, this component must be able to be replaced without having to stop the provision of the firewall service. This	MRQ

	implies that the instance can be made up of a number of nodes, but in compliance with the following requirements: 1. The HA mode offered must be such that it does not require the use of multicast on the LAN between the instance and the network to which the firewall instance is attached. 2. The firewall instance must guarantee the minimum required performance (see 5.2.1-1) by accepting a maximum degradation of no more than 50% of the protected traffic in the event of failure of no more than 50% of the equipment.	
5.2.1-3	*General requirements* The firewall instance must have at least the following ports: • 16 x 100 Gb/s ports. • 16 x 10 Gb/s ports. • 4 x 1 Gb/s copper or 10 Gb/s ports (for management). This count excludes all the backend ports of the instance, i.e. those necessary for its proper internal functioning in relation to the proposed architecture. It is the responsibility of the Candidate to include the necessary network components (e.g. firewall modules, SFPs, switches, cables, etc.) for the implementation of the offered solution. It should also be noted that the requirement on the number and type of network ports does not apply to the management platform, but only to the components that are in charge of securing traffic. The management platform shall be configured by the Candidate in such a way that it can meet the requirements set out in Section 0. The ports indicated above, with the exception of the 1 Gb/s ones, must be supplied complete with SR transceivers. We also require the following additional transceivers to be used for future/backup purposes: • 20 transceiver 100 Gb/s SR. • 20 transceiver 10 Gb/s SR. The firewall instance must also support: • LACP (IEEE 802.3ad) to aggregate at least 4 interfaces for connection with the CINECA network • VLAN (IEEE 802.1Q) • MP-BGP • OSPFv2/v3	MRQ

	• GRE for GRE tunnel interface configuration The hardware virtualization feature must allow the implementation of at least 2 independent logical firewall instances. The offer should include any necessary licenses to activate this functionality.	
5.2.1-4	*Networking filtering rules* The firewall instance must have the ability to configure filtering rules based on IPv4/IPv6 address, protocol, and port. The rules must also allow to identify the source of traffic on a geographical basis and implement configurable rate limiting mechanisms.	MRQ
5.2.1-5	*Identity filtering rules* The firewall instance must have the ability to enable filtering rules that are conditioned by verifying the identity of users and/or devices. Integration with at least Active Directory and the availability of local authentication services to the Firewall (e.g. Captive Portal) are required. User recognition can be carried out even in the absence of specific and/or proprietary clients installed on each CINECA workstation, as long as this can also be done from Windows 10 or higher workstations.	MRQ
5.2.1-6	*Application filtering rules* The firewall instance must have the ability to configure filtering rules based on the recognition of major applications. This recognition must take place through the L7 analysis of network traffic and not through the mere IANA port-application association.	MRQ
5.2.1-7	*Encrypted traffic filtering rules* The firewall instance must have the ability to configure filtering rules to enable analysis of TLS/SSL encrypted traffic. The functionality must be present both for inbound traffic (to protected company resources, through the installation of their X509 digital certificates and related private keys on the firewall) and for outbound traffic (from protected company resources, e.g. laptops, to the Internet).	MRQ
5.2.1-8	*Web filtering rules*	MRQ

	The firewall instance must have the ability to configure filtering rules based on the classification of web resources (URLs). The classification of resources must take place dynamically, possibly also by means of a Cloud service of the vendor in continuous/periodic contact with the Firewall.	
5.2.1-9	*Intrusion prevention system* The Firewall must be equipped with an Intrusion Prevention service at least based on: IoC (including signatures related to known CVEs and IP reputations), identification of abnormal use of protocols (including DNS), identification of anomalies in the use of network resources or user behavior. The service must also have the ability to identify and block bot activity by controlling the active command and control channels within the analyzed traffic. The list of threats (IoCs, patterns, etc.) must be updated dynamically, possibly also by means of a Cloud service of the vendor in continuous/periodic contact with the Firewall.	MRQ
5.2.1-10	*Antivirus and antimalware support* The Firewall must be able to analyze traffic to identify and block, through Antivirus, the transmission of malicious executables or documents (viruses, malware, trojans, etc.). At a minimum, the antivirus must be signature-based. The list of signatures must be constantly updated. Protection can also be provided through integration with a Cloud service of the vendor in continuous/periodic contact with the Firewall.	MRQ
5.2.1-11	VPN support The firewall instance must have the ability to configure both IPsec and SSL VPNs. IPv6 must be supported for at least SSL VPN. The client-to-site VPN configuration must support multi-factor authentication, an SSL clientless mode (with an access portal integrated into the firewall), and a mode with or without a proprietary client as long as it works from at least Windows 10 or higher locations.	MRQ

5.2.2 Border Firewall Management Platform

Req.	Description	Category

5.2.2-1	*General requirements* The firewall instance must be managed by a single management platform that must be located on-premises. The management platform must allow: • The configuration, through a web interface and/or unified proprietary console, of the firewall instance covered by the specifications including security policies. • Receiving and indexing logs from firewall instances. • The possibility, through a web interface and/or proprietary console, to carry out customized searches among the logs received. The firewall instance management platform will need to be provided with all hardware, software, licenses, and maintenance services (for 60 months) necessary for full and proper operation.	MRQ
5.2.2-2	*Hardware appliances* The firewall instance management platform must each consist of 2 dedicated hardware appliances configured in active/stand-by high availability without requiring the use of multicast in the LAN between said platform and the network to which it is connected.	MRQ
5.2.2-3	*Management platform storage* Each appliance of the management platform must be equipped with at least 24 TB (raw) for log storage and have adequate computational and memory resources to receive, index and store in a sustained (non-peak) manner at least 20,000 log lines per second. The 24 TBytes mentioned above must be provided not necessarily within the appliances but also through additional components configured in high reliability and dedicated to log storage, as long as they are integrated with the appliances.	MRQ

5.3 Data center Backbone Border Routers

Req.	Description	Category

5.3-1	*General requirements*
	The solution must provide 2 equal network devices with internet border functionality, i.e. RN1, RN2 as shown in Figure 7. Each of these network devices must:
	• Be based on non-blocking wirespeed architectures, i.e. with backplanes sized to guarantee transmission at the maximum allowed speed and without limitations (non-blocking) to at least the following interfaces (which must be included in this Offer, for each device):
	  ○ 8 x 400 Gb/s ethernet interfaces including 400GBASE-SR4 transceivers.
	  ○ 8 x 100 Gb/s ethernet interfaces including 100GBASE-SR4 transceivers.
	  ○ 8 x 10 Gb/s ethernet interfaces including 10GBASE-SR transceivers.
	•
	• Ensure that the router fits within a 2U (2 rack units) maximum form factor.
	• be equipped with a 1000BaseT ethernet management interface and a console interface.
	• have redundant power supply.
	• be equipped with all the licenses necessary to support both Layer 2 (L2) and Layer3 (L3) networking features and/or protocols with reference to the OSI model on Ethernet network and in particular at least the following:
	  ○ IPv4 and IPv6 Routing
	  ○ OSPF v2/v3
	  ○ MP-BGP
	  ○ GRE Tunnel
	  ○ Support for at least 6 VRFs with:
	  ▪ Supporting importing/exporting network prefixes between VRFs (VRF leaking) on the basis of user-definable criteria that do not necessarily require the user to insert static lists (e.g. possibility of import/export to/from a VRF of all prefixes received from a given routing protocol).
		MRQ

	▪ Ability to manage the full internet routing table with multiple peering for at least one VRF. o FHRP (es. VRRP, HSRP, etc.). o Support to implement traffic filtering rules with IP-based policies (e.g. ACLs, security policies, etc.). • Provide aggregation at least with multi-chassis protocols (e.g. MCLAG, VPC, E-VPN Multihoming, etc.) which will be used to aggregate the interconnection. Protocols that aggregate network equipment into a single system by unifying management and/or control plane plans (e.g. stacking protocols and/or stacked configuration) will not be allowed. The Candidates will have to provide everything necessary to achieve this functionality e.g.: transceivers, interfaces necessary to create the control channels (peer links), proprietary cables, etc. o LACP protocol. o L2 switching capabilities. o VLAN support (IEEE802.1Q). o L2 and L3 Multicast support. o Jumbo frame support. o Broadcast storm control features. o Mirroring the traffic of a router port and/or VLAN. o Sampling of traffic flows and forwarding to third-party equipment for analysis and monitoring activities (e.g. Netflow, Jflow, Sflow, etc.). • We also require the following additional transceivers to be used for future/backup purposes: o 16 transceiver 400 Gb/s SR o 16 transceiver 100 Gb/s SR. o 8 transceiver 100 Gb/s LR o 16 transceiver 10 Gb/s SR. • 100GBASE-SR4 and 10GBASE-SR ports can also be obtained by splitting the ports at higher speed with breakout cables that must be included in the offer. Finally, as required in 5.1.1-10, all interfaces must be attested on patch panels included in the Offer.	
5.3-2	*Geographical links* The Candidate will therefore have to provide for the certification of the geographical links directly on the RN1 and RN2 internet border routers provided. Therefore, these devices must have 100 Gb/s connections based on SR and LR transceivers for geolink attestation. If it is not possible to use directly	MRQ

	100 Gb/s transceivers within the 400 Gb/s port, it is required to include transceivers and breakout cables necessary to establish 100 Gb/s connections.	
5.3-3	*Management interfaces* The RN1 and RN2 internet border routers must be manageable via command line interface (CLI) to ensure independence from the HPEF Fabric described in Section **Errore. L'origine riferimento non è stata trovata.**. Specifically, it is required that RN1 and RN2 network devices are equipped with at least the following administrative functions at the operating system level: • Administrative shell for managing the system, configurations, local files, monitoring. • SSHv2 clients. • SNMPv2 protocol support. • Definition of multiple administrative users that can be configured to differentiate the privileges of access and management of the equipment. • Authentication of users and groups, both local and remote, using the Radius protocol. • Event and anomaly logging. • Support for log forwarding to remote server via Syslog protocol.	MRQ

5.4 Data Center Wireless Network

The Offer must include a Wireless Network that covers the Computing Data Hall, including all technical rooms. Since in the Technopole Data Center in Bologna CINECA uses a Wireless LAN solution based on Juniper Mist, it is requested that a solution based on the same technology be proposed in order to have completely compatible and homogeneous platforms from the point of view of configuration and management, as well as simplifying overall maintenance.

Req.	Description	Category
5.4-1	*WiFi access points* The Offer must include at least 10 access points and what is necessary for the implementation of a WiFi service for access to the Internet and to the systems installed at the Technopole datacenter with the following features: • support IEEE 802.11ax, 802.11ac, 802.11a/g/n protocols • support at least 4 SSIDs • ensure 2.4Ghz, 5Ghz and 6Ghz Radio support.	MRQ

	• support WPA3 and WPA2 in "Enterprise" modes with both RADIUS (802.1X) and PSK (Pre-Shared Key) authentication • Access to wireless networks for guests via a dedicated captive portal • support IEEE802.3af (POE) and IEEE.802.3at (POE+) protocols • support both wall and ceiling mounting.	
5.4-2	*WiFi management platform* The Offer must include a management platform for the WiFi service with the following features: • The traffic of users connected to the APs should not be sent to the cloud but should be managed exclusively locally. • The configurations of APs and data required for service level monitoring and troubleshooting must be appropriately encrypted and secured. The cloud service must be hosted at most on servers located in Europe and must comply with current regulations in terms of data storage, integrity and confidentiality and the provisions of the European GDPR regulation. • The management platform will have to collect and allow the visualization of administrators access logs (login/logout). Retention of wireless access logs of administrators must be kept in accordance with the law and available for at least 6 months. • The management platform must collect and allow the visualization of users' access logs (login/logout). Retention of wireless access logs of user must be kept in accordance with the law and available for at least 6 months. • The management platform must have the ability and include the licenses (1200 users) to interface directly with the Azure AD instance of CINECA for user authentication and must allow the creation of custom policies for WiFI access based on Azure AD security groups. • The management platform must include all the hardware and software components (including any licenses for 60 months) useful for its operation.	MRQ

5.5 Data center Network Management

The Offer must include a network management infrastructure dedicated to border routers and border firewalls as well for the connection of the wireless infrastructure.

Req.	Description	Category

5.5-1	*Management switch SWNM1 and SWNM2*
	We require the supply of 2 ethernet switches based on non-blocking wirespeed architectures, i.e. sized with backplanes capable of guaranteeing each transmission at the maximum allowed speed (wirespeed) and without limitations (non-blocking) to at least the following interfaces (which must be included in this supply, for each device):
	• 8 x 10Gb/s ethernet interfaces including 10GBASE-SR transceivers
	• 16 x 1Gb/s 1000BaseT ethernet interfaces supporting IEEE802.3af (POE) and IEEE.802.3at (POE+) standards
	• equipped with 1000BaseT ethernet management interface and console interface
	• with redundant power supply
	• with the following features:
	  ○ MLAG, LAG, LACP protocols
	  ○ Routing protocols for IPv4 (e.g., OSPF,OSPFv3).
	  ○ VLAN (IEEE802.1Q)
	  ○ L2 and L3 Multicast
	  ○ Jumbo frame
	  ○ Broadcast storm control
	  ○ Admin shell with commands for managing the system, local files, monitoring
	  ○ SSHv2 and Telnet Clients
	  ○ Definition of multiple administrative users that can be configured to differentiate access privileges and management of equipment.
	  ○ Authentication of both local and remote users and groups using the Radius protocol
	  ○ Event and anomaly logging
	  ○ Support for log forwarding to remote server via Syslog protocol
	  ○ Support for SNMPv2 and v3 protocols and "Management Information Base" (MIB)
		MRQ

	○ That they are fully compatible and provide the correct power delivery to the Access Points described in 5.5.	
5.5-2	*Management platform* The Offer must include an on-premise management platform for centralized monitoring of the devices described in Section: 5.2 (Border Firewall and Firewall Management Platform), 5.3 (Border router), 5.4 (Access Point), and 5.5 (Management Switch), with the following features: • high availability. • monitoring of the use of the main resources (HW, CPU, RAM, etc.) and the bandwidth usage of network interfaces. • receiving and consulting logs via syslog. • backup of network equipment configurations (RNx, FWNx and SWNMx) with the possibility of verifying differences between different versions and restoring previous configurations. • sending alarms and notifications via email.	MRQ
5.5-3	*Console Server* The Offer must include at least one console server for the attestation of the serial interfaces of the network equipment (including any licenses for 60 months) with at least 32 serial interfaces and allows to reach the consoles of the devices connected to them via IP by network.	MRQ

6 Tier-1 system infrastructure

6.1 General requirements

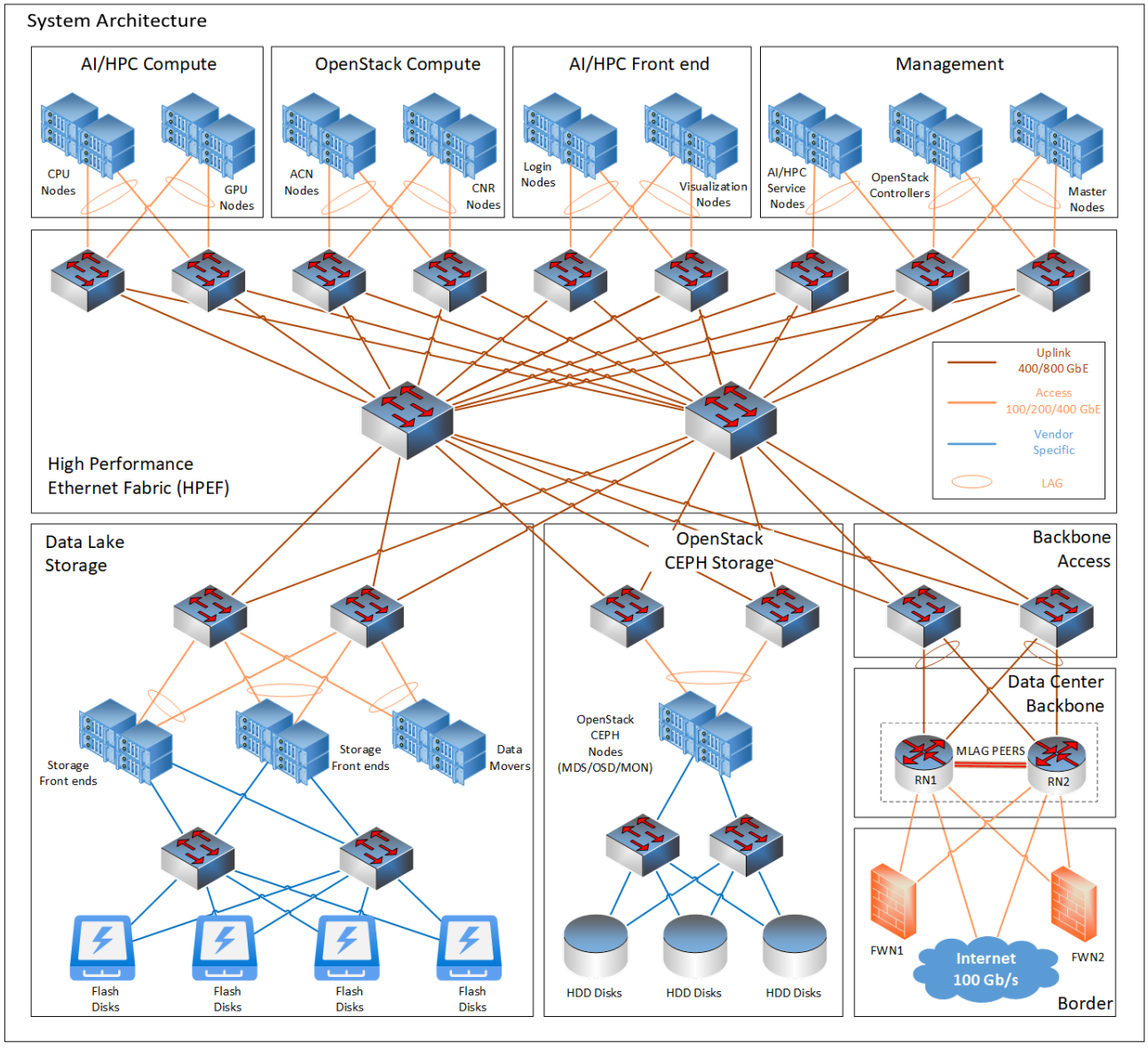

Figure 8: Reference design of the system architecture.

6.1.1 Functional aspects

Req.	Description	Category
6.1.1-1	*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
6.1.1-2	*Reboot time* Each partition must be fully rebooted in less than 60 minutes.	MRQ

6.1.1-3	*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:
		MRQ

	○ 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 components, when a pre-failure event is detected, the system must automatically send the alert.	
6.1.1-4	*Health monitoring* The procured infrastructure must 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
6.1.1-5	*Node power and energy measurement* The procured infrastructure must provide node power and energy measurements with a minimum of 95% accuracy and low impact on performance. The details of the implementation will be considered in evaluating the Offers that provide this capability.	MRQ
6.1.1-6	*Power capping*[2] The procured infrastructure must implement power capping mechanisms to meet the power constraint limits (see Section 4.3), the power consumption must be modified at unit level (e.g., rack) and at runtime with minimal impact on performances.	MRQ

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

6.1.1-7	*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
6.1.1-8	*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

6.2 Interconnects

6.2.1 High Performance Ethernet Fabric

Req.	Description	Category
6.2.1-1	*General requirements* The procured infrastructure must provide a High-Performance Ethernet Fabric (HPEF) used to interconnect all nodes. The fabric must respect the following characteristics: • Must be based on 800 Gb/s. • The minimum bandwidth for each access link must be at least 100 Gb/s with full bi-directional bandwidth per port. • The oversubscription of the topology must not exceed 2:1. • Provide full support to IPv4 and IPv6. • The topology must be based on layer 3 and BGP/EVPN/VXLAN. • Support RDMA over Converged Ethernet (RoCEv1 and v2). • Support spine-leaf architecture.	MRQ

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

	• Support MPI communications. • The HPEF must provide enough spine switches to accommodate future expansion of new leaf switches to increase the network up to 20% of the access ports. • The fabric topology must be redundant at both spine and leaf switches. Failures of single switch/link must not affect the network flow. The reference design of the HPEF is represented in Figure 8.	
6.2.1-2	*Fabric bandwidth* The HPEF should be full-fat tree with full bisection bandwidth (non-blocking[4]) is preferred to avoid any network congestion.	TRQ
6.2.1-3	*Switch characteristics* The switches of the fabric must support: • Several different logical networks (e.g. VLAN, EVPN-VXLAN). • Standard IEEE 802.X network protocols, in particular VXLAN (routing and bridging) and related protocols. • MLAG, LAG, LACP protocols. • EVPN multi-homing (ESI LAG). • Routing protocols for IPv4 and IPv6 (e.g., OSPF, BGP, MP-BGP, OSPFv3, etc.). • L2 and L3 Multicast support. • NVMe over TCP (NVMe/TCP). • Jumbo frames support.	MRQ
6.2.1-4	*Fabric participants* The nodes and the storage of the System must be connected to the HPEF, as shown in Figure 8, and at least to two different switches to provide high availability and redundancy in case of a switch failure.	MRQ
6.2.1-5	*Spine switches* Spine switches must have the following characteristics:	MRQ

[4] Oversubscription 1:1, on each 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).

	• High radix with at least 64 ports at 800 GbE. • All ports must support 100/200/400 GbE.	
6.2.1-6	*Leaf switches* • Must provide access ports at 100/200/400 GbE. • Must provide uplinks ports at 400 GbE and/or 800 GbE.	MRQ
6.2.1-7	*Monitoring and managing capabilities.* The following capabilities must be provided: • The HPEF must provide a SDN fabric manager to configure, control, and monitoring in near-real time the performance and the health of the fabric. • Each switch of the HPEF must provide an out-of-band management port based on 1Gb/s.	MRQ
6.2.1-8	*Advanced network features* The HPEF should support the following mechanisms: • *Lossless ethernet*: it should support lossless packet network. • *Out-of-order packet reordering*: it should support mechanism based of out-of-order packet reorder on the fabric endpoints. • *Smart NIC*: for the east-west traffic of AI/HPC partitions, the fabric should be based on end-to-end optimization based on smart NIC equipped on compute nodes in order to offload the network stack and to optimize the overall communications (e.g. packer reordering). • *RDMA adapting routing*: it should support adapting routing for RDMA communications. • *Telemetry-based congestion control*: it should implement congestion controls based on telemetry mechanism incorporated in the fabric. • *Noise isolation*: through adaptive routing mechanisms. • *Self-healing*: with re-routing based on auto-discovery changes of the topology. • *Security and performance isolation*: it should implement security mechanism and performance isolation to respect stringent SLAs imposed by storage infrastructures.	TRQ
6.2.1-9	*Power redundancy*	MRQ

	Each switch of the HPEF must be equipped with redundant & hot-swappable power supplies.	

6.2.2 Management network

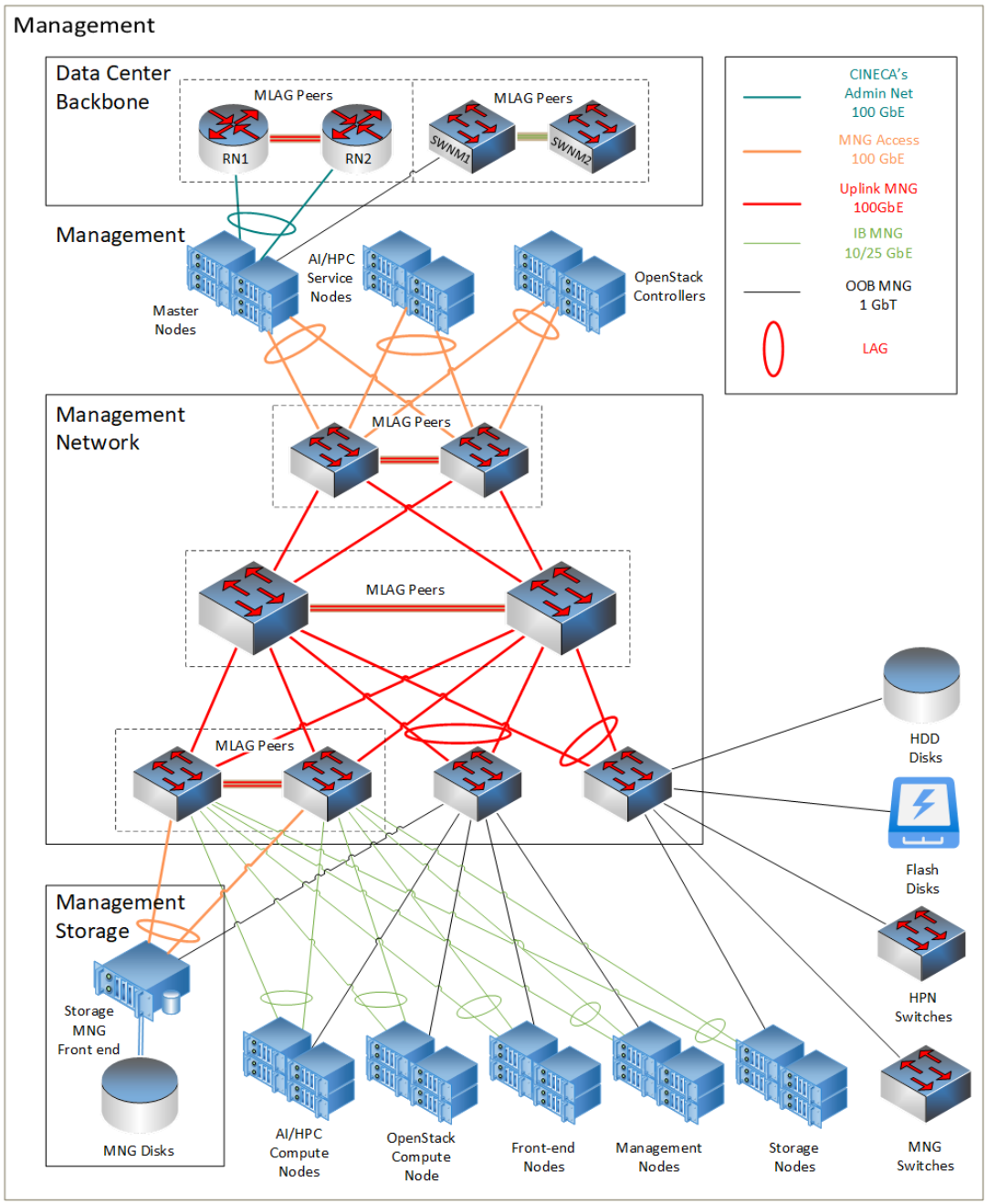

Figure 9: Reference design of the management network.

Req.	Description	Category
6.2.3-1	*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 reference design of the management network is represented in Figure 9.	MRQ
6.2.3-2	*Network Participants* • I*n-Band 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 management network*: used to manage the nodes through Board Management Controller (BMC) of all system's equipment (nodes, storage, networks, chassis, etc.).	MRQ
6.2.3-3	*Monitoring and managing capabilities* • 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 Gb/s connected to the OOB MNG network.	MRQ
6.2.3-4	*Spine switches* • Switches at spine level must provide ports at least 100 Gb/s to be connected to the leaf switches, management nodes, and to the management storage. The number of spine switches and the number	MRQ

	of ports will be defined by the Candidate in order to respect the req. 6.2.3-1. • Spine switches must support static and dynamic routing (Layer 3) and relative protocols (e.g., OSPF, BGP, MP-BGP, OSPFv3).	
6.2.3-5	*Leaf switches* • Must provide access ports at 1 Gb/s for OOB MNG equipment. • Must provide access ports at 10 Gb/s or 25 Gb/s for IB MNG nodes. • Must provide access ports at least 100 Gb/s for the Management Storage. • Must provide uplink ports at least 100 Gb/s.	MRQ
6.2.3-6	*Network capabilities* The switches of the management network must support: • Full support to IPv4 and IPv6. • MLAG, LAG, LACP, VLAN protocols and most common IEEE 802.X network protocols. • L2 and L3 Multicast support. • Jumbo frames support.	MRQ
6.2.3-7	*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
6.2.3-8	*Power redundancy* Each switch in the management network should be equipped with redundant & hot-swappable power supplies.	TRQ

6.3 AI/HPC Compute partition

The Compute partition includes two sub partitions:

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

6.3.1 AI/HPC CPU partition

Req.	Description	Category
6.3.1-1	*Partition size* The partition must feature at least 100 nodes.	MRQ
6.3.1-2	*Common node requirements* The nodes must implement the requirements provided in Section 6.1.1-3.	MRQ
6.3.1-3	*CPU Technology* The CPU must be based on x86_64 architecture and provide: • at least 64 cores. • a theoretical peak performance (FP64) of at least 3.5 TFlops. • high single-thread performance.	MRQ
6.4.1-4	*Node configuration* The node must be equipped with at least 2 CPUs.	MRQ
6.3.1-5	*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
6.3.1-6	*Network requirements* The nodes must be equipped with:	MRQ

	• 1 Smart NIC with 2 Ethernet ports connected to the HPEF with at least 200 Gb/s per port (400 Gb/s aggregated). The network ports must be connected to different switches. Those smart NIC 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.
	  ○ Single-root input/output virtualization (SR-IOV).
	  ○ OpenStack OVS hardware offloading.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10 Gb/s or 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.	

6.3.1-7	*Node local storage* The nodes must be equipped with x2 SSD drives in RAID1 configuration with net space available of ≥ 0.8 TByte for the OS.	MRQ
6.3.1-8	*Virtualization support* The hardware and the operating system of the nodes must support virtualization and management of virtual machines through OpenStack infrastructure to have the possibility to convert them in OpenStack Compute Nodes.	MRQ
6.3.1-9	*Leaf switch for AI/HPC CPU Nodes* Leaf switches that connect AI/HPC CPU Nodes to the HPEF must not share ports with other partitions but must be fully dedicated only to AI/HPC CPU Nodes.	MRQ

6.3.2 AI/HPC GPU partition

Req.	Description	Category

6.3.2-1	*Partition performance* The partition must feature at least 100 nodes.	MRQ
6.3.2-2	*Common node requirements* The nodes must implement the requirements provided in Section 6.1.1-3.	MRQ
6.3.2-3	*CPU Technology* The CPU must be based on x86_64 architecture.	MRQ
6.3.2-4	*GPU technology* The GPUs must support the following characteristics: • Support IEEE-conforming single (FP32) and double precision (FP64) computations. • Provide at least a theoretical peak performance (FP64) of at least 60 TFlops. • Provide memory sharing with all the other GPU installed in the same node.	MRQ
6.3.2-5	*Node configuration* Each node must be equipped with 2 CPUs and 8 GPUs (8-way GPU configuration).	MRQ
6.3.2-7	*DRAM memory* • The node memory must be at least 1 TByte of DDR5 memory and 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
6.3.2-8	*Network requirements* The nodes must be equipped with: • 1 Smart NIC per GPU connected to the HPEF with at least 400 Gb/s (≥3.2 Tb/s aggregated) for the East-West (EW) traffic among GPUs as shown in Figure 10. Those network ports must support: ○ VLAN HW offloading.	MRQ

	○ 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. ○ Single-root input/output virtualization (SR-IOV). ○ OpenStack OVS hardware offloading. • 1 Smart NIC with 2 Ethernet ports connected to the HPEF with at least 200 Gb/s per port (400 Gb/s aggregated) for the storage and North-South (NS) traffic as shown in Figure 10. 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. ○ Single-root input/output virtualization (SR-IOV). ○ OpenStack OVS hardware offloading. • 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10 Gb/s or 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.	
6.3.2-9	*Node local storage* The nodes must be equipped with: • x2 SSD drives in RAID1 configuration with net space available of ≥ 0.8 TByte for the OS. • x8 SSD drives with a net space available for each disk of at least ≥ 3 TBytes used for a scratchpad storage. The drive array should support multiple RAID configurations such as RAID0/1/5/6 and global hot spare drives.	MRQ
6.3.2-10	*Fast distributed scratchpad*	TRQ

	The AI/HPC GPU partition should be supported with a dedicated fast distributed scratchpad built on top of the SSD drives equipped within the AI/HPC GPU Nodes and described in req. 6.3.2-9. This scratchpad should be shared among all the AI/HPC GPU Nodes, Data Movers, AI/HPC Front-end Nodes, and should provide the following characteristics: • At least 2 PBytes of net space available. • At least a total read throughput of 5 TB/s. • At least a total write throughput of 1 TB/s. • At least 150M IOPS (4k random reads) for NFS (or parallel file system) reads. • At least 45M IOPS (4k random writes) for NFS (or parallel file system) writes.	
6.3.2-11	*Virtualization support* The hardware and the operating system of the nodes must support virtualization and management of virtual machines through OpenStack infrastructure to have the possibility to convert them in OpenStack Compute Nodes.	MRQ
6.3.2-12	*Leaf switch for GPUs* As shown on Figure 10, NICs dedicated to east-west (EW) traffic among GPUs must be connected to leaf switches where all the ports must be dedicated only to the GPUs. Instead, storage and North-South (NS) traffic can be connected to switches with mixed ports with other partitions.	MRQ

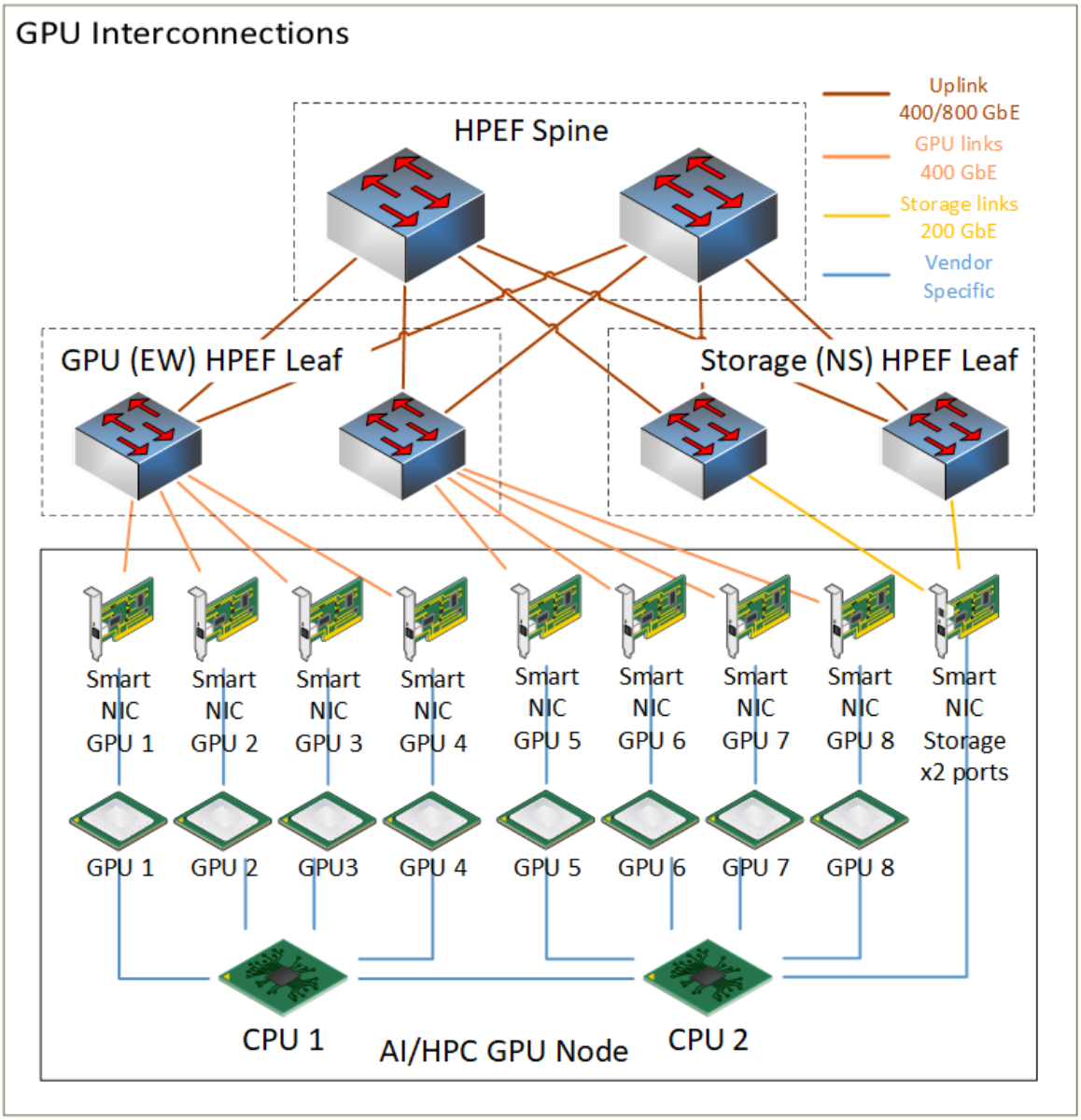

Figure 10: Reference design of the GPU and HPEF interconnection. The internal AI/HPC GPU node connection will be defined by the Candidate depending to the specific GPU technology.

6.4 OpenStack compute partition

The Compute partition includes two sub partitions:

- OpenStack *CNR partition*: it provides the conventional and accelerated compute nodes dedicated for CNR provided by the OpenStack infrastructure.
- OpenStack *ACN partition*: it provides the conventional and accelerated compute nodes dedicated for ACN provided by the OpenStack infrastructure.

6.4.1 OpenStack CNR partition

Considering the performance-demanding workloads planned on this partition, these OpenStack Compute Nodes will be operated providing the lowest possible virtualization impact on computing and storage performance, therefore providing bare-metal performance. No oversubscription will be applied to memory and CPU cores, and no hyperthreading will be activated.

Req.	Description	Category
6.4.1-1	*Partition size* The partition must feature 50 nodes. These 50 nodes must have bare-metal performance, and they must be accessible with system administrator permissions (root permissions).	MRQ
6.4.1-2	*Common node requirements* The nodes must implement the requirements provided in Section 6.1.1-3.	MRQ
6.4.1-3	*CPU technology* The CPUs must be based on x86_64 architecture and providing: • At least 64 cores per CPU. • A theoretical peak performance (FP64) of at least 3.5 TFlops. • high single-thread performance.	MRQ
6.4.1-4	*GPU technology* At least 40 of the 50 nodes of the OpenStack CNR partition must have the following set up: • State-of-the-art GPU technology with at least 1 GPU per node. • Each GPU must support IEEE-conforming single precision (FP32) and double precision (FP64) calculation.	MRQ

	• Each GPU must provide a theoretical peak performance (FP64) of at least 60 TFlops. • Each GPU must provide at least 80 GB of memory.	
6.4.1-5	*Node configurations* • Each one of the 50 nodes must be equipped with at least 2 CPUs. • At least 40 of the 50 nodes must feature at least 1 GPU.	MRQ
6.4.1-6	*DRAM memory* • The nodes must be equipped with at least 2 TBytes of DDR5 memory. • 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

6.4.1-7	*Network requirements*
	The nodes must be equipped with:
	• 1 NIC with 2 Ethernet ports connected to the HPEF 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.
	  ○ Single-root input/output virtualization (SR-IOV).
	  ○ OpenStack OVS hardware offloading.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10 Gb/s or 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

6.4.1-8	*Node local storage*	MRQ

	The nodes must be equipped with: • x2 SSD drives in RAID1 configuration with net space available of ≥ 0.8 TByte for the OS. • An array of NVMe SSD drives with net space available of ≥ 60 TB for the local data storage.	

6.4.2 OpenStack ACN partition

Req.	Description	Category
6.4.2-1	*Partition size* The partition must be featured with: • A set of CPU nodes featuring at least 100 nodes. • A set of GPU nodes featuring at least 96 Type-A GPUs (see req. 6.4.2-4). • A set of GPU nodes featuring at least 16 Type-B GPUs (see req. 6.4.2-5).	MRQ
6.4.2-2	*Common node requirements* The nodes must implement the requirements provided in Section 6.1.1-3.	MRQ
6.4.2-3	*CPU Technology* The CPUs of this partition must be based on x86_64 architecture and provide: • at least 128 cores for each CPU to the CPU nodes. • at least 56 cores for each CPU to the GPU nodes. N.B. see req. 6.4.2-1 for the node types.	MRQ
6.4.2-4	*Type-A GPU technology* The Type-A GPUs must comply with the following characteristics: • State-of-the-art GPUs with support to AI inference, 3D acceleration, and IEEE-conforming single precision (FP32) calculation. • Each GPU must provide at least 48 GB of memory.	MRQ
6.4.2-5	*Type-B GPU technology* The Type-B GPUs must comply with the following characteristics: • State-of-the-art GPUs with support to AI inference for LLMs. • Each GPU must provide at least 188 GB of memory.	MRQ
6.4.2-6	*Node configuration* • Each node must be equipped with at least 2 CPUs.	MRQ

	• The nodes equipped with GPUs must feature at least 4 Type-A GPUs or 2 Type-B GPUs.	
6.4.2-7	*DRAM memory* • The nodes must be equipped with at least 3.5 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

6.4.2-8	*Network requirements*
	The nodes must be equipped with:
	• 1 NIC with 2 Ethernet ports connected to the HPEF 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.
	  ○ Single-root input/output virtualization (SR-IOV).
	  ○ OpenStack OVS hardware offloading.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10 Gb/s or 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

6.4.2-9	*Node local storage* The nodes must be equipped with x2 SSD drives in RAID1 configuration with net space available of ≥ 0.8 TByte for the OS.	MRQ

6.5 Management partition

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

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

Req.	Description	Category
6.5-1	*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 15 nodes. The number of OpenStack Controller Nodes must be at least 5% of the OpenStack Compute Nodes. Candidates are invited to employ virtualization techniques to reduce the size of the service partition while complying with the above minimum.	MRQ
6.5-2	*Common node requirements* The nodes must implement the requirements provided in Section 6.1.1-3.	MRQ
6.5-3	*CPU technology* The partition must be equipped with CPU based on x86_64 architecture.	MRQ
6.5-4	*Node configuration* Each node must be equipped with at least 2 CPUs.	MRQ
6.5-5	*Memory configuration* All Management Nodes must feature a total of at least 256 GBytes of DDR5 memory.	MRQ

6.5-6	*Network requirements*
	
	The nodes must be equipped with:
	
	• 1 NIC with 2 Ethernet ports connected to the HPEF 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.
	  ○ Single-root input/output virtualization (SR-IOV).
	  ○ OpenStack OVS hardware offloading.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10 Gb/s or 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.
	• Only for Master Nodes: 1 NIC with 2 Ethernet ports connected to the CINECA's HPC Backbone Network with 100 Gb/s per port (200 Gb/s aggregated) as shown in Figure 9. The network ports must be connected to different CINECA's backbone switches.
	
	OOB and IB MNG NICs can also be the same NIC.
		MRQ

6.5-7	*Node local storage* The nodes must be equipped with x2 SSD NVMe drives in RAID1 configuration with net space available of ≥ 3TB.	MRQ
6.5-8	*Management storage* Management Nodes must have a shared storage with net available space of at least 400 TBytes to contain: 1. All the management software. 2. All the management databases and an historical daily (differential) backup of these databases for a year.	MRQ

	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. The storage must feature the following: 1. Connected to the management network and must be possible to be mounted to all Management Nodes. *2.* Be resilient to the failure of at least two independent basic blocks (i.e., storage nodes, controllers, or disk chassis) proving redundancy and high availability (e.g. RAID technology).	
6.5-9	*Connection with* CINECA's HPC Backbone The Master Nodes must be connected with the CINECA's HPC Backbone Network described in Section 5. See Figure 9 for the connection's details of the Management nodes.	MRQ
6.5-10	*AI/HPC Service Performance* The number of the AI/HPC 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
6.5-11	*High Availability* The Management partition will include the required hardware components to configure all important system services in high availability. The AI/HPC Service Nodes will 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 AI/HPC Compute and Front-end Nodes will be able to continue working without significant interruption. Any performance impact will be described in the Offer.	MRQ

6.5-12	*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
6.5-13	*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
6.5-14	*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. 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.	MRQ
6.5-15	*Basic Hardware Monitoring* The cluster management software will provide out-of-band or in-band monitoring of hardware events (e.g., system event log and machine check	MRQ

	exceptions, if applicable). The events will be collected and stored at a central location.	

6.6 AI/HPC Front-end partition

The Front-end partition includes two sub partitions:

- *AI/HPC Login partition*: for external system access, compilation and data management activities, job submission as well interactive pre-/post-processing workloads.
- *AI/HPC 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.

6.6.1 AI/HPC Login partition

Req.	Description	Category
6.6.1-1	*Partition size* The partition must feature at least 8 nodes.	MRQ
6.6.1-2	*Common node requirements* The nodes must implement the requirements provided in Section 6.1.1-3.	MRQ
6.6.1-3	*Node configuration* The nodes must be organized in two different setups: • *CPU setup*: half of the nodes must feature the same CPU of the AI/HPC CPU partition's nodes. • *GPU setup*: half of the nodes must feature the same CPU of the AI/HPC GPU partition's nodes and include at least one GPU with the same technology installed in the AI/HPC GPU partition's nodes.	MRQ
6.6.1-4	*Memory configuration* The nodes must feature a total of at least 512 GBytes of DDR5 memory.	MRQ
6.6.1-5	*Network requirements*	MRQ

	The nodes must be equipped with:
	• 1 NIC with 2 Ethernet ports connected to the HPEF 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.
	  ○ Single-root input/output virtualization (SR-IOV).
	  ○ OpenStack OVS hardware offloading.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10 Gb/s or 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.	

6.6.1-6	*Node local storage* The nodes must be equipped with x2 SSD drives in RAID1 configuration with net space available of ≥ 7 TBytes for the OS.	MRQ
6.6.1-7	*Internet connectivity* The Login Nodes must have direct access to Internet through the HPEF, for this reason the HPEF must propagate the Internet connection via VXLAN tunnel.	MRQ
6.6.1-8	*Software installation* Login Nodes must allow the installation of all user software and applications that need to be run on the system.	MRQ

6.6.2 AI/HPC Visualization partition

Req.	Description	Category

6.6.2-1	*Partition size* The partition must be composed to at least 2 nodes.	MRQ
6.6.2-2	*Common node requirements* The nodes must implement the requirements provided in Section 6.1.1-3.	MRQ
6.6.2-3	*CPU technology* A state-of-the-art CPU binary compatible with the CPUs installed in the AI/HPC Compute nodes.	MRQ
6.6.2-4	*GPU technology* A high-end GPU supporting 3D acceleration through OpenGL graphics.	
6.6.2-4	*Node configuration* The nodes must be equipped with at least 2 CPUs and 2 GPUs.	MRQ
6.6.2-5	*Memory configuration* The nodes must feature a total of at least 512 GBytes of DDR memory.	MRQ

6.6.2-7	*Network requirements*
	The nodes must be equipped with:
	• 1 NIC with 2 Ethernet ports connected to the HPEF 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.
	  ○ Single-root input/output virtualization (SR-IOV).
	  ○ OpenStack OVS hardware offloading.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10 Gb/s or 25 Gb/s. The network ports must be connected to different switches and they must support:
		MRQ

	○ 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.	
6.6.2-6	*Node local storage* The nodes must be equipped with x2 SSD drives in RAID1 configuration with net space available of ≥ 7 TBytes for the OS.	MRQ
6.6.2-8	*Internet connectivity* The nodes have direct access to Internet through the HPEF, for this reason the HPEF must propagate the Internet connection via VXLAN tunnel.	MRQ

6.7 Storage infrastructure

6.7.1 Data Movers

Req.	Description	Category
6.7.1-1	*Partition size* The Data Mover partition must be composed to at least 2 nodes.	MRQ
6.7.1-2	*Common node requirements* The nodes must implement the requirements provided in Section 6.1.1-3.	MRQ
6.7.1-3	*CPU technology* The nodes must be equipped with CPU based on x86_64 architecture.	MRQ
6.7.1-4	*Node configuration* Each node must be equipped with at least 2 CPUs.	MRQ
6.7.1-5	*Memory configuration*	MRQ

	The nodes must feature a total of at least 512 GBytes of DDR5 memory.	

6.7.1-6	*Network requirements*
	
	The nodes must be equipped with:
	
	• 1 NIC with 2 Ethernet ports connected to the HPEF 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.
	  ○ Single-root input/output virtualization (SR-IOV).
	  ○ OpenStack OVS hardware offloading.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10 Gb/s or 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

6.7.1-7	*Node local storage* The nodes must be equipped with x2 SSD drives in RAID1 configuration with net space available of ≥ 7 TByte for the OS.	MRQ
6.7.1-8	*Internet connectivity* The Data Movers have direct access to Internet through the HPEF, for this reason the HPEF must propagate the Internet connection via VXLAN tunnel.	MRQ

6.7.2 Data Lake

Req.	Description	Category
6.7.2-1	*Specific requirements*	MRQ

	The offered solution must provide at least: • A full-flash storage. • At least 46 PBytes (aggregated) of net space available. 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.	
6.7.2-2	*Geo-redundancy Backup* A certain amount of the storage (around 30%) will be installed at the Big Data Technopole data centre in Bologna in order to provide geo-redundancy backups for critical data. For this reason, the storage infrastructure should be splitable at least at 30% of the overall capacity.	MRQ
6.7.2-3	*Performance* The storage infrastructure installed in the Naples data centre must provide at least the following performance: • At least a total read throughput of 560 GB/s. • At least a total write throughput of 140 GB/s. • At least 3.5M IOPS (4k random reads) for NFS (or parallel file system) reads. • At least 800k IOPS (4k random writes) for NFS (or parallel file system) writes. Instead, the storage infrastructure installed in the Technopole data centre must provide at least the following performance: • At least a total read throughput of 160 GB/s. • At least a total write throughput of 40 GB/s. • At least 1M IOPS (4k random reads) for NFS (or parallel file system) reads. • At least 220k IOPS (4k random writes) for NFS (or parallel file system) writes.	

6.7.2-4	*HPEF connection* • The storage infrastructures will be connected through HPEF 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
6.7.2-5	*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
6.7.2-6	*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
6.7.2-7	*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. • 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.	MRQ
6.7.2-8	*Self-recovery time duration*	MRQ

	The recovering process described in 6.7.2-7 should not take more than 4 hours and the impact on performance must be less than 10%.	
6.7.2-9	*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
6.7.2-10	*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 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 PBytes of capacity.	MRQ
6.7.2-11	*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.	MRQ

	• The storage infrastructures must provide audit log of client operations.	
6.7.2-12	*Security features* • 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.	MRQ
6.7.2-13	*File system metadata* The offered solution must support the possibility to make the file system metadata available for query as a database table.	MRQ
6.7.2-14	*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
6.7.2-15	*S3 support* • 4 TBytes or larger maximum object size. • Support for >100 buckets/namespace. • S3 identity and bucket policies. • Multipart uploads. • Object Lock.	MRQ
6.7.2-16	*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.	MRQ

	• 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.	
6.7.2-17	*QoS* • 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.	MRQ
6.7.2-18	*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. • 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.	TRQ
6.7.2-19	*Advanced multi-site features* • The offered solution should support multi-site configurations. • Any cluster should present any folder/bucket from its namespace to one or more remote clusters where users can read or write to this global path as if it were just a folder in the local namespace. The offered solution should make data available where needed, whether that's sharing data generated in any of the data center sites.	TRQ
6.7.2-20	*Software Licensing and Support* • Vendor must guarantee software licenses and support, including replacement of all parts regardless of flash wear, will be offered on the storage infrastructures.	MRQ

	• 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.	

6.7.3 OpenStack CEPH

Req.	Description	Category
6.7.3-1	*Common node requirements* The nodes must implement the requirements provided in Section 6.1.1-3.	MRQ
6.7.3-2	*CPU configuration* The nodes must be equipped with CPU based on x86_64 architecture.	MRQ
6.7.3-3	*Node configuration* Each node must be equipped with at least 2 CPUs.	MRQ
6.7.3-3	*Memory configuration* The nodes must feature a total of at least 256 GB of DDR5 memory.	MRQ

6.7.3-4	*Network requirements*
	The nodes must be equipped with:
	• 1 NIC with 2 Ethernet ports connected to the HPEF 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.
	  ○ Single-root input/output virtualization (SR-IOV).
	  ○ OpenStack OVS hardware offloading.
	• 1 NIC with 2 Ethernet ports connected to the IB MNG network at 10 Gb/s or 25 Gb/s. The network ports must be connected to different switches and they must support:
		MRQ

	○ 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.	
6.7.3-5	*CEPH OSD Nodes* The solution must provide a number of nodes adequate to the offered CEPH storage. Those specific nodes must also provide a capacity optimized OSD nodes for a total capacity of 2 PBytes net space available. The offered CEPH solution will be used as block service for the OpenStack infrastructure. Hardware design and architecture of OSD nodes needs to take this into account. Moreover, these nodes must be equipped with: • x2 SSD drives ≥ 200 GBytes units for journaling. • Number of cores enough to support the IO frontend and OSDs as per CEPH best practices. • Planar and HBAs bandwidth must be adequate and balanced against the mix of SSD drives, SATA SSD drives, and NL-SATA disks hosted, to optimize the IO flows.	MRQ
6.7.3-6	*CEPH MON and MDS Nodes* The solution must provide a number of nodes adequate to the offered CEPH storage. Moreover, these nodes must be equipped with: • x2 SSD drives in RAID1 configuration with net space available ≥ 3 TBytes.	MRQ

6.8 Facility integration

Req.	Description	Category
6.8-1	*Integration with CINECA's HPC Backbone network*	DCS

	The system will be connected to the CINECA's *HPC* Backbone network through two dedicated HPEF switches as shown in Figure 8. See Section **Errore. L'origine riferimento non è stata trovata.** for the interconnection details.	
6.8-2	*Closed Innermost Cooling Loop* If applicable, the innermost cooling loop will be well separated from the data center 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
6.8 -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 center is prevented.	MRQ

6.9 System software and monitoring

Req.	Description	Category
6.9-1	*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
6.9-2	*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
6.9-3	*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	MRQ

	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.	
6.9-4	*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
6.9-5	*Lightweight Performance Profiling* The procured infrastructure will provide lightweight performance profiling capabilities that can be activated by the users on a job basis. Data at process, 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:	TRQ

	• 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).	
6.9-6	*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
6.9-7	*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
6.9-8	*Mechanisms for correlation* The Candidate will provide mechanisms that enable correlation of different metrics from different monitoring systems (including external ones). An open	TRQ

	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.	
6.9-9	*Optimized numerical libraries* 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.	MRQ
6.9-10	*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

7 Benchmarks

7.1 Introduction

This chapter describes the context and main goals of the benchmark procedure for assessing the performance of the offered solutions.

This Chapter is organized as follows:

- In Section 7.2 the benchmark suite is reported.
- In Section 7.3 the benchmark execution rules are described.
- In Section 7.4 the expected benchmark results are described.

7.2 Benchmark suite

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

#	Benchmark	Partition	Short description	website
1	**HPL**	AI/HPC GPU	HPL benchmark solves a linear system of equations of order n, measuring the sustained performance of the whole system.	see top500.org
3	**MLPerf Training**	AI/HPC GPU	MLPerf Training is a suite of ML benchmarks that represent both industrial and academic use cases	https://mlcommons.org/benchmarks/training/

Table 3: Benchmark application composing the benchmark suite.

7.3 Benchmark execution

Datasets, instructions, and rules on how to configure, set-up, and run the benchmarks are all available in the websites referred in Table 3.

7.4 Benchmark analysis report

To evaluate the benchmark results, a benchmark analysis report provided by the Candidate is required. The report will assess the performance, in terms of the selected benchmark applications, of the offered solution. HPL benchmark results projection should be provided for the whole AI/HPC GPU partition. Given that AI is one of the main drivers, results for MLPerf training benchmarks are expected to be reported. See also the tendering document "Disciplinare di gara" Article 8.5.

8 Maintenance and infrastructure availability

The Offer for the procured infrastructure must include a maintenance and support service that ensures high availability, described in Section 6.3, and stability of the procured infrastructure. In the following with the term Supplier, it is referred the Candidate awarded for this procurement.

8.1 Maintenance and support requirements

Req.	Description	Category
8.1-1	*Maintenance and support duration* The Candidate will offer maintenance and support of the offered infrastructure according to articles 1, 4 and 6. of the procurement document "Schema di Contratto".	MRQ
8.1-2	*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
8.1-3	*Special software support coverage*	TRQ

	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.	
8.1-4	*Reaction times* The Supplier guarantees an appropriate reaction time upon hardware and software issues.	MRQ
8.1-5	*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
8.1-6	*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 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.	TRQ
8.1-7	*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

8.1-8	*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
8.1-9	*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
8.1-10	*Escalation management process* The Supplier will provide an escalation management process to manage problems priority and critical/non-critical issues.	MRQ
8.1-11	*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
8.1-12	*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 9.	MRQ
8.1-13	*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
8.1-14	*Security patches and software updates* The Supplier will make security patches for all supported software components available for installation in an adequate period following the release of the	MRQ

	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.	
8.1-15	*Tier-1 maintenance and support service regulation* This service is considered "*a corpo*"[5] 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
8.1-16	*Maintenance periodic reporting* Periodic maintenance and maintenance activities must be reported, including: • Ticket lists issued by the call center, 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.	MRQ

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

	• Analysis of repeated failures. • Conformance ratios to SLAs. Candidate will describe this periodic reporting in the Offer.	
8.1-17	*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 center, etc.). The Supplier's representative will participate, if required, in regular meetings along with its representatives to 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.	MRQ
8.1-18	*Documentation requirements* The Candidate will describe the support workflow for software and hardware failures, including information about replacement part logistics and SLAs.	MRQ

8.2 Tier-1 specialistic support

Req.	Description	Category
8.2-1	*Professional services for high performance network* The Supplier must provide professional services for the HPEF infrastructure described in section 6.2. These services must target the set-up and tuning phase, and the production phase. Support must cover the lifetime of the System (60 months). These services must be described and included in the Offer.	MRQ

8.3 Data center network professional service

The following professional services must be part of the provision:

Req.	Description	Category
8.3-1	*Professional service requirements* • professional services for the design, installation and configuration of the devices and the equipment in the Offer; • professional post-implementation support services of at least 15 days that can also be provided remotely (e.g. for support in the implementation of network configurations to meet requirements not foreseen in the initial project, for analysis and troubleshooting of functional anomalies, etc.); • training service in Italian for CINECA staff on the technological architecture and operational management of the equipment supplied. The training must be provided through dedicated sessions of no more than 10 people. The training plan must be described by the provider (in terms of topics envisaged, number of days, proposed sessions, etc.); The Candidate must be a certified or concessionaire of the producers of the material to be supplied, which certifies its ability to successfully carry out the activities envisaged by these Specifications.	MRQ

8.4 Licenses

Req.	Description	Category
8.4-1	*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
8.4-2	*Licenses' list* Candidate must provide the complete list of all applicable licenses provided with their quantities.	MRQ

8.5 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:

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

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 or the cloud sashboard.
- 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.

For the AI/HPC partitions planned maintenance[6] interventions will be scheduled for up to 7 days per year and every month for a duration of 8 hours.

Req.	Description	Category

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

8.5-1	*Targeted monthly availability* The design of the procured infrastructure architecture and the maintenance service must aim for a monthly availability of 95%.	TRQ
8.5-2	*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 Installation and acceptance

9.1 Installation time schedule and project management

The Proposal for the procured infrastructure should 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.

9.1.1 System Installation

Req.	Description	Category
9.1.1-1	*Project Management* The Supplier will provide project management resources for the system installation.	MRQ
9.1.1-2	*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.	MRQ
9.1.1-3	*Installation Time* The timeline of delivery, installation, acceptance and start of operations are detailed in the tender document "Schema di Contratto", art. 4.	MRQ
9.1.1-4	*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.	MRQ

	• 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. • 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.	

9.1.2 Supply and installation project

Req.	Description	Category
9.1.2-1	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 staff is organized and foreseen. • Implementation plan for procured infrastructure acceptance (see Section 9.2). All interactions with CINECA staff, all training activities, as well as the documentation produced within the project, can be in Italian or English.	MRQ
9.1.2-2	*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
9.1.2-3	*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	MRQ

	expected impact and risk mitigation measures that will be implemented as part of the contract.	
9.1.2-4	*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

9.2 Acceptance procedure

9.2.1 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.

9.2.2 Execution of acceptance tests

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

For the acceptance procedure the following rules apply:

Req.	Description	Category
9.2.2-1	*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 combination for each benchmark. All tests will be performed using the production WLM.	MRQ

9.2.3 Provisional acceptance tests

9.2.3.1 Hardware checklist

Req.	Description	Category
9.2.3.1-1	*Hardware checklist* Completeness and consistency of the delivered and installed hardware will be checked against the Offer.	MRQ
9.2.3.1-2	*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 not functional. • Frontend nodes: all nodes must be functional. • Service nodes: all nodes must be functional. • HPEF links: less than 0.1% of the links may be not functional.	MRQ

9.2.3.2 Software Checklist

Req.	Description	Category
9.2.3.2-1	*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

9.2.3.3 Functional Tests

Req.	Description	Category

9.2.3.3-1	*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 9.2.3.4).
- 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 Sections 9.2.3.5-9.2.3.7),
- Verification of software specifications and offered features.
- Verification of other commitments made in the Offer.

9.2.3.4 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.

9.2.3.5 Application and Synthetic Benchmarks

The benchmarks included in the benchmark suite (see Section 7.2) 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 2% of relative tolerance.

9.2.3.6 Rules for HPL Benchmarks

The HPL benchmark will be executed on the CPU and GPU partition separately, according to the TOP500 list rules. During the LINPACK benchmark, the power consumption will be measured according to the GREEN500 run rules.

9.2.3.7 I/O Performance

The I/O performance of the filesystems will be measured using IOR or any relevant tool, in agreement between CINECA and the Supplier. The committed I/O performance (see Section 6.7.2) must be achieved within the relative tolerance of 2%.

9.2.4 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
9.2.4-1	*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 0 - that must result 90% or above.	MRQ

9.2.5 Final acceptance

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

Annex 1: Tier1 system glossary

Term	Description
Backbone	Site-wide Ethernet network (40GbE or 100GbE)
CINECA	Interuniversity Consortium
DDR	Double Data Rate
DIMM	Dual In-line Memory Module
HPL	High Performance Linpack (see top500.org)
CPU	Central Processing Unit
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
HPEF	High Performance Ethernet Network
LTS	Long-Term data Storage
LLM	Large-Language Model
NIC	Network Interface Card
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 acronyms and common terms for the Tier-1 infrastructure.