The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Digital Europe programme and the National Recovery and Resilience Plan "Greece 2.0" funded by the European Union – NextGenerationEU.

ΕΔΥΤΕ Α.Ε.
Εθνικό Δίκτυο Υποδομών Τεχνολογίας και Έρευνας

GRNET S.A.
National Infrastructures for Research and Technology

Introduction

GRNET S.A. – National Infrastructures for Research and Technology, operating since 1998, is one of the largest public sector technology companies in Greece. Since August 2019, it operates under the auspices of the Ministry of Digital Governance. GRNET provides networking, cloud computing, HPC, data management services and e-Infrastructures and services to academic and research institutions, to educational bodies at all levels, and to all agencies of the public sector. It offers:

- A nation-wide fiber optic network offering connectivity services amongst different institutions as well as inter-institutional departmental connectivity in Greece and abroad.
- High Performance Computing System (HPC-ARIS).
- Infrastructures (6), data centers.
- The National Competence Center for HPC.
- The National Academy of Digital Skills.
- The Greek Internet Exchange GR-IX facilitating the exchange of Internet traffic (IP).
- Internet, cloud computing, high-performance computing, authentication and authorization services, security services, as well as audio, voice and video services.

GRNET, holding the key role as the coordinator of all e-Infrastructures for Education and Research in Greece, has been assigned to design and implement the new Greek supercomputing installation, called DAEDALUS and carry out its operation. This facility is planned to be located at the Lavrion Technological and Cultural Park, in Attica - which is managed by the National Technical University of Athens (https://en.ltcp.ntua.gr).

The European High Performance Computing Joint Undertaking (EuroHPC JU) was established by Council Regulation (EU) 2021/1173 of 13 July September 2021. The mission of the EuroHPC JU is to develop, deploy, extend and maintain in the European Union an integrated world-class supercomputing and data infrastructure and to develop and support a highly competitive and innovative HPC ecosystem. The European Union's contribution from the Digital Europe Programme should cover up to 35 % of the acquisition costs plus up to 35 % of the operating costs of the mid-range supercomputers.

GRNET will co-own the 'mid-range' supercomputer DAEDALUS with EuroHPC JU and will operate DAEDALUS on behalf of the EuroHPC JU. GRNET will cover the share of the total cost of ownership of DAEDALUS that is not covered by the Union contribution, either until its ownership is transferred by the

EuroHPC JU to that hosting entity or until the supercomputer is sold or decommissioned in case there is no transfer of ownership.

This document describes the desired architecture of DAEDALUS, which shall meet the needs of both the Greek and European academic and research community as well as those of all potential users of the infrastructure. GRNET requests that stakeholders contribute to the implementation of the architectural design by studying the attached document and submitting their eventual comments or suggestions thereon. The comments and input submitted will be considered for the final specifications of the system in the forthcoming open international call for tenders.

Technical Requirements for DAEDALUS system.

The procured system will be a combination of high-performance compute, network and storage elements together with the software needed to be operational. The procurer expects the usage of the latest and most modern technologies for all used components at the time of delivery and a stable setup of all components including hardware, software and supporting equipment to enable the provision of advanced HPC, HPDA and AI oriented workloads. Part of this procurement will also be the delivery of implementation services, integration into the power and cooling infrastructure of the procurer, training of staff, warranty and support services provision. Finally, the system should be as "green" as possible in terms of the required power to achieve maximum performance.

The DAEDALUS system must allow effective execution of many concurrently running compute jobs in all phases of its life cycle (pre-processing/preparation, computation, post-processing/visualization) and of all different types of workloads (parallel, serial, batch, interactive) for a broad spectrum of applications and users. The system must allow a secure storage of the user's data with high speed and low latency access, an effective management of the whole system and its individual components, monitoring of available resources and the users. The system must provide a transparent, unified, shared user environment and unified access to all different compute and storage resources.

The procurer expects that the DAEDALUS system should have at least the following logical components:

- CPU-only compute partition
- Accelerated compute partition
- High-speed network
- HPC storage with parallel filesystems
- High IOPS Flash storage
- Service nodes
- Network infrastructure LAN and WAN
- Power and cooling equipment – integration into data center
- Software components

General Requirements

1. The total cost of the system should be less than 33 million Euro, excluding VAT.

2. The target peak power consumption for the IT components should be less than 1.5 MW.
3. The Rmax (LINPACK Benchmark) performance of the accelerated compute partition should be at least 30 PFlops.
4. The Rmax performance of the CPU-only partition should be at least 5 PFlops.
5. The vast majority of the equipment (at least both the CPU-only and accelerated partitions) should be Direct Liquid Cooled.
6. The hardware equipment should be supported on a Next Business Day scheme.
7. The operating system of the machines should be supported for at least 5 years after installation.
8. Possible licenses for any part of the procured system should be included for 5 years.
9. All delivered systems and equipment must be accompanied by the EC declaration of conformity.
10. The Contracting Authority will give due consideration to the EU added value of the supercomputing system that will be proposed by the tenderer, assessing to what extent it contributes to achieving the objectives of the EuroHPC JU, as these are defined in the Regulation establishing the EuroHPC JU, and in particular Article 3 thereof[1]. In this regard, the inclusion of EU technologies and components in the supercomputing system of the tenderer is encouraged.
11. The candidate vendor should have at least 2 installed and supported systems in Europe with Rmax performance more than 12 PFlops.
12. The System (IT hardware) should be developed in the available area of around 150 square meters

Requirements -CPU-only compute partition

1. The CPU-only compute partition must consist of standard diskless nodes without accelerators and it should be based on either the x86 or the ARM CPU architecture.
2. The minimum supported instruction set should be AVX2 in the case of x86 and VSE2 in the case of ARM.
3. Each node should include an amount of RAM yielding at least 2 GB per core.
4. All memory channels/slots should be fully occupied to maximize memory bandwidth.
5. Memory should support ECC.
6. Each node should include a high bandwidth / low latency network interface with a bandwidth of at least 200 Gbps and a latency of less than 1 microsecond.

Requirements - Accelerated compute partition

1. The accelerated compute partition must consist of standard nodes with local flash type storage for local scratch. Its CPU should be based on either the x86 or the ARM CPU architecture.
2. The minimum supported instruction set should be AVX2 in the case of x86 and VSE2 in the case of ARM.
3. Each node should include an amount of RAM yielding at least 2 GB per CPU core.
4. All memory channels/slots should be fully occupied to maximize memory bandwidth.
5. Memory should support ECC.
6. Each node should have 4 accelerators.

[1] Official Journal of the European Union, 08.10.2018, L 252, pages 1-34.

7. Each node should include four high bandwidth / low latency network interfaces with a bandwidth of at least 200 Gbps and a latency of less than 1 microsecond each.

Requirements- High speed network

1. The high speed network should have an injection bandwidth of at least 200 Gbps for each port.
2. The preferred high speed network topology is either full non-blocking fat tree or Dragonfly+. Any other topology that guarantees non-blocking communications is acceptable.
3. The switches must have free ports to connect additional equipment (~10% of the occupied ports).
4. The communication between accelerated nodes should be non-blocking.
5. An optimized version of MPI for high speed networks should be provided and supported for 5 years after installation.
6. The high speed network should support the IP protocol in addition to RDMA and probably other protocols.
7. The necessary switches and cables should be provided.
8. The high speed network switches should support remote collection of statistics of switch as well as per port using snmp or other protocols.

Requirements - Typical HPC storage for parallel filesystems

1. The total usable capacity should be at least 10 PBytes.
2. The storage should support multiple volumes with different tuning. 5 volumes are planned for storage.
3. All filesystems should support user quota, group quota and snapshots.
4. The storage bandwidth from compute nodes should be at least 100 GB/s.
5. The metadata of the filesystems should reside on flash media.
6. The storage system should support collection of statistics and health status of all components (controllers, disks, etc.)

Requirements - High IOPS Flash storage

1. Total usable capacity should be at least 1 PB.
2. The file system should provide long-term sustainable random I/O performance for the block size of 4KiB and 80%/20% read/write mode of 3 million IOPs. The required performance must be achievable from compute nodes.

Requirements - Service Nodes

1. The system should include a number of service nodes. These nodes should have their own redundant local storage.
   1.1. Four login nodes if the CPU and accelerated partitions CPUs are exactly the same. If accelerated partition CPUs are different, two more login nodes with the same CPUs as those of accelerated partition CPUs.
   1.2. Cluster management, scheduler, monitor, accounting nodes. At least two nodes for each role.
   1.3. At least 10 general purpose nodes, with the same CPUs as those of the CPU partition nodes and local redundant storage for various services VMs.

Requirements - Network Infrastructure LAN and WAN

1. All the necessary ethernet switches and cables for the management and monitoring network should be provided.
2. Nodes exposed to both the internet and external private networks should have additional ethernet interfaces. The total bandwidth to/from internet of these nodes should be 200 Gbps. This connectivity should be integrated to WAN equipment.

Power and Cooling Equipment

Power

Provide a general description in terms of electrical power distribution. The description must at least include the following:

1. Electrical Power Distribution type (Voltage/phases/frequency) per Rack. Distribution will be carried out through overhead busbars.
2. Quantities and type of power supply units per rack (plug type, PDU el. Diagrams and power monitor module), power cables routing, circuit size based on ELOT HD 384 or another equivalent standard. Primary and redundant supply units should be mentioned separately.
3. Input/output of network cables, type (i.e Copper, Fiber optics), quantity (estimated). Connection topology between racks but also with the Grnet Network should be provided together with a drawing /diagram.
4. Maximum absorbed power (kW and Amperes per Phase) per Rack configuration.
5. Power capacity (kW) per rack (referred to typical AC 400V/3ph/50Hz). Maximum capacity and designed capacity per Rack as of the proposed solution will be stated to define total redundant el. Power. IT and cooling (for the DLC system) power capacities should be mentioned separately.
6. Annual Total PUE of the combined system should not exceed 1,2.
7. Vendors must provide a detailed description and topology, regarding the interconnection of the proposed solution and future GRNET's BEMS (Building and Energy Management System). Supported network protocols and available capabilities of the proposed API (Application Programmable Interface) should also be stated and analyzed.

Cooling

The primary cooling system will be based on DLC (Direct Liquid Cooling) technology. DLC transfers heat away from processors and other heat-generating system components via liquid-cooled medium rather than using just air in the heat-exchange process.

Solution Scope is P.U.E. ratio minimization by mechanical cooling (compressors) usage deterioration.

Therefore, the crucial evaluation point will be the maximum acceptable primary circuit's inlet temperature.

The DLC system has two separate hydraulic circuits (primary and secondary) that are exchanging heat through CDUs (Cooling distribution unit). CDUs can be modules installed either in each rack or in separate row-based units.

Secondary cooling circuit, based on CDU, is the hydraulic circuit that transfers heat from IT equipment (CPUs, accelerated compute partitions etc.) while the primary is the one transferring heat from secondary circuit to the external environment through the building's cooling system.

The Solution Description (and the relative proposal) , provided in a technical description, should at least include the following:

1. General description of the proposed solution.
2. Redundancy method (water pumps, power supply, secondary circuit water pumps control). It is important to avoid any single point of failure.
3. Control points (cooling medium temperatures, flow etc) of the CDU PLC (or other equivalent control system).
4. Scenarios to avoid failure in primary or secondary circuit and device protection methods in case of unfavorable issues (i.e):
    a. Scenario 1: increased inlet primary or secondary cooling medium temperature
    b. Scenario 2: Decreased primary or secondary cooling medium flow
    c. Scenario 3: Major failure in the buildings cooling system (provided through a dry contact)
5. Cooling performance curves including upper and lower nominal operation limits:
    a. Pressure Drop of the primary circuit ΔP(bar) correlated to cooling medium flow Q (m 3/h)
    b. Cooling Medium flow of the primary circuit Q(m 3/h) correlated to cooling medium inlet temperature Tin(oC) for IT 50,75,100% Loads
    c. Temperature diff ΔT between inlet and outlet cooling medium of the primary circuit correlated to cooling medium inlet temperature Tin(oC) for 50,75,100% IT load and maximum cooling medium flow.
    d. Outlet cooling medium temperature of the primary circuit correlated to cooling medium inlet of the secondary circuit Tin(oC) for 50,75,100% IT load.
    e. Maximum IT load P(kW) correlated to cooling medium inlet temperature of the primary circuit Tin(oC) in case of raised water inlet above the nominal defined limits (provided the solution can respond to this situation).
6. Primary hydraulic connections (type, quantity, diameter etc.).
7. Option for 2-way or 3-way primary hydraulic valves.
8. The additional heat loads of the proposed solution to be rejected through air cooling. Also provides the DC required conditions (temperature and Relative Humidity) including upper and lower nominal operation limits.
9. Cooling medium quality requirements for both primary and secondary circuits.
10. Technical proposal for the heat removal of primary hydraulic circuit specifically customized to Athens environmental conditions and temperature profile (optional).
11. Drawing including the hydraulic connection to the primary circuit for the proposed solution (optional).

Software Components

1. All delivered software should have unlimited license at least for versions applicable at delivery time.
2. The system should include a resource and job schedule manager based on SLURM. Access to compute nodes for end users should be possible only via the resource manager.
3. The system should include all tools for management of all system parts: remote power on/off or reset, remote collection of health status, and environment parameters.
4. System should support network boot from a central boot images service. All compute nodes should be able to boot in 15 minutes. System should allow the creation / modification of boot images.
5. All nodes should run the same version of the Operating System that has to be RHEL based.
6. A vendor optimized version of Linpack should be delivered.
7. Vendor hardware optimized tools and libraries and tools should be delivered.