# Technical requirements specification for EURO\_IT4I system

Scope of this public procurement is a modernization of the equipment of the procurer by acquiring a modern HPC system called EURO\_IT4I that should replace the current HPC system SALOMON. The procured system is a complex of high performance compute, network and storage elements together with software equipment needed for an effective operation. The procurer expects the usage of the latest and most modern technologies for all used components at the time of delivery and a complex setup of all components including the software part to enable provision of advanced HPC, HPDA and AI oriented workloads. Part of this procurement is also the delivery of implementation services, integration into the power and cooling infrastructure of the procurer, training of staff, warranty and support services provision.

The EURO\_IT4I 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 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 EURO\_IT4I system should have at least consist of following logical components:

* Universal compute partition
* Accelerated compute partition
* Data analytics compute partition
* Cloud infrastructure compute partition
* High speed compute network
* User’s data storage SCRATCH
* User’s data storage PROJECT
* User’s data storage HOME
* System’s data storage INFRA
* Login nodes
* Visualization nodes
* Data management nodes
* Infrastructure and management nodes
* Backup for infrastructure and management nodes
* Network infrastructure LAN
* Network infrastructure WAN – integration only
* Power and cooling equipment – integration into data center
* Software equipment

The **Universal compute partition** should consist of compute nodes without accelerators such as GPUs or FPGAs and should be based on x86 CPU architecture to provide quick accessibility for the users and their existing codes. It should have at least 2 GB of RAM per core (256 GB per node). Each node should have approximately 3.75 TFLOPS in Linpack giving the partition an expected performance of 2.23 PFLOPS. This partition should be also very good for the inference part of AI workloads. The expected power consumption of the partition should be 418 kW with a usual compute load.

The **Accelerated compute partition** should deliver most of the compute power usable for HPC but also excellent performance in HPDA and AI workloads, especially in the learning phase of Deep Neural Networks. It should consist of heavily accelerated nodes. Each node should have two x86 CPUs, 512 GB of RAM, 8 GPUs interconnected by a high speed memory coherent bus providing at least 384 GB of HBM memory and quad-rail high speed interconnects to the HPC network to enable the huge data ingest and output and parallelization between these nodes. The total Linpack performance of this partition should be around 6.3 PFLOPS. A small local flash storage for each node should be provided to supplement the support for bursts of data intensive workloads on these nodes. The expected power consumption of this partition should be less than 400 kW with a usual compute load.

The **Data analytics compute partition** should be oriented on supporting huge memory jobs by implementing a NUMA SMP system with 24 TB RAM, thirty two x86 CPUs and quad-rail high speed interconnects to the HPC network. The partition is preferred to be a single system allowing usage of all the resources by a single job (under a single instance of running OS) or on demand software partitioning using technologies like cgroups and cpusets to allow maximal utilization by multiple users with smaller requirements. New workload like quantum computing simulators can be run at such partition too still keeping the flexibility to run traditional HPC workloads. The total performance of the partition in Linpack should be at least 50 TFLOPS. Expected power consumption should be around 17.5 kW under usual compute load.

The **Cloud infrastructure compute partition** should support both the research (e.g. H2020 project LEXIS) and operation of the *Infrastructure/HPC as a Service*. This should further extend the virtualization and container based abstraction on the main compute partitions to allow the users   
to specifically design their own cloud based clusters using technologies like OpenStack and Kubernetes. Each node should have two x86 CPUs with the same architecture, core count and frequency as the universal nodes, 256 GB of RAM per node and extra Ethernet based network adapters to allow better utilization of the VMs running on the platform and the possibility to create user defined networks. A small local SSD storage should be provided for each node to even further enable the smooth setup of the cloud infrastructure and optimize the run of the VMs by providing a swap partition. The total Linpack performance should be around 131 TFLOPS. The expected power consumption of the partition should be less than 27 kW with a usual load.

The **High speed compute network** should be an RDMA based network supported by the GPU accelerators with link speeds of 100 Gb/s to the single port of an adapter, where the universal nodes should have a single port, while the accelerated ones 4 ports in 4 different cards. The data analytics node should also have 4 individual ports. An optimized topology such as dragonfly, multi-dimensional hypercube or similar should be used to reduce cost of the network while retaining the performance of the system and a balanced bandwidth to the centralized high speed storage across all of the nodes. A system wide Linpack should be possible using this configuration.

The **High speed SCRATCH storage** should consist of flash based storage with the open source SW stack providing the parallel shared high performance file system. The total net capacity should be more than 1 PB with a throughput providing 350 GB/s – 1 TB/s and IO rate exceeding 5 M IOPS. A distributed system consisting of server nodes with SSD storage and a software layer like BeeGFS or a traditional Lustre based setup of storage and servers should be used. The storage is primarily designed for the IO intensive AI workload but should also serve extremely well also for traditional HPC workloads as a scratch file system. This storage has to be implemented as an independent solution from any others by the used HW components (disk drives, controllers, servers, network connections) and the SW layer as well (e.g. metadata server/DB cannot be shared with other storage provided). The expected power consumption should be around 50 kW with a usual load if implemented as the distributed file system or 22 kW if implemented as the traditional Lustre solution.

The **PROJECT storage** is a centralized storage operated at IT4Innovations for all clusters designed to store the data during the whole time of a user project (even multi-year projects). It’s designed on blocks which are independent and thus providing resiliency/redundancy for the data with snapshots allowing simplified backup of changing data. At the time of EuroHPC system installation the capacity of this storage will be 2 blocks with a total capacity between 5-8 PB and an aggregated throughput of 50 GB/s. As part of EURO\_IT4I another similar block of this storage should be acquired and integrated into the whole pool. The predicted additional capacity is approximately 2.5–3 PB with a 25 GB/s aggregated throughput, increasing the total to about 10 PB. The design is based on traditional rotating disks in storage enclosures with frontend servers allowing multiple access using protocols such as NFS, GridFTP and SSHFS/SCP. This flexibility of access allows mapping of the storage into traditional HPC environments as well as into cloud platforms. This storage has to be implemented as an independent solution from any others by the used HW components (disk drives, controllers, servers, network connections) and the SW layer as well (e.g. metadata server/DB cannot be shared with other storage provided). The expected power consumption should be around 20 kW with a usual load.

The **HOME storage** should allow users to store their data related to OS and application settings and should have a small capacity around 25 TB only.

The **INFRA storage** should be used to store the data of the managing, operating and monitoring infrastructure typically images of OSes of the compute nodes, all logs, user applications, scheduler/resource manager data, etc. It should be redundant/highly-available at least on the level of the underlying disk technologies (RAID, redundant controllers), preferably also on higher levels including the file servers and connectivity to the LAN network. Expected net capacity is around 25 TB.

The system should provide at least four **Login nodes** allowing the users interactive access from Internet doing all kind of pre-processing and post-processing activities related to their computations. Login nodes must be implemented as physical servers.

The system should provide at least two **Visualization nodes** that should be equipped with a hardware OpenGL accelerator (GPU) and SW to allow users remote access to the accelerated environment.

The system should provide at least four **Data management nodes** that should be dedicated to data movement of user’s data between the internal storages (HOME, SCRATCH, and PROJECT) and external ones (including other systems at the procurer premises or even completely remote locations). The foreseen protocols for such data transfers are SFTP, SCP, NFS and GridFTP, where the installation of such software environment is not part of this procurement.

The system should provide necessary amount of **Infrastructure and management nodes** that should serve to operate the whole system, especially fulfilling the tasks of job scheduling/resource management, license management and common operation services (DHCP, DNS, LDAP, monitoring, logging, etc.). The key services should be run in high-availability mode, preferably using the native mechanisms of the services.

**Backup for infrastructure and management nodes** should provide backup of data available on the individual infrastructure nodes, INFRA storage and HOME storage. The preferred technology is disk based with the effective use of de-duplication techniques.

The system should include equipment for complete implementation of **LAN infrastructure** and it’s integration with the existing WAN. The LAN should consist of individual L3 networks that should be based on individual L2 networks (represented either by a VLAN or by separate hardware equipment). The LAN part connecting login, visualization and data mover nodes to the WAN should be redundant at the hardware level. The connectivity of the aforementioned nodes should be redundant too. Bandwidth of the connection for the nodes should be at least 10 Gb/s. For login and data mover nodes, an aggregated bandwidth of 100 Gb/s should be provided. The minimal bandwidth for management and infrastructure part of the LAN should be 1 Gb/s. All the LAN equipment (switches and routers) should have remote central management and monitoring functionality included. The management interfaces of the LAN equipment should be integrated into the existing out-of-band network.

The system should be completely **integrated into the power and cooling infrastructure** of the data center including connection to the central measurement-and-regulation system where preferred source of cooling is a direct liquid cooling (DLC) approach with input temperature of the liquid at 32C minimum. For technologies that cannot be cooled using DLC, cooling using water of lower temperatures (16C) can be used by leveraging technologies like heat exchange rear doors on racks or in-row cooling towers.

The system should have warranty and should be serviced for 3 years.

The system should include implemented **software** stack covering at least operating systems for all equipment, firmware, drivers, libraries, job scheduler/resource manager, system services backup services, provisioning, shared parallel cluster file systems, monitoring, central logging, compilers and others needed for the overall effective operation. If needed appropriate licenses should be provided as well for 3 years. The operating system on all servers has to be RHEL or CentOS based.