# Technical requirements specification for Perun system

The scope of this public procurement is acquiring a modern HPC system called Perun integrated into a container datacenter (C-DC). The subject of the procurement is the whole HPC system together with the C-DC and all necessary supporting infrastructure. The procurer will provide a sufficiently prepared object for the C-DC installation with the base for the C-DC, power supply, water and sewage connections, data connectivity. The system will be procured as a part of a national project within the Resilience and Recovery Facility (RRF), therefore the timeline of the procurement as well as some specifications are tied to the RRF milestone definitions.

The procured HPC system integrated into C-DC is a complex of high performance compute, network and storage elements, as well as supporting technologies such as cooling and power backup 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 infrastructure of the procurer, training of staff, extended warranty and support services provision.

RRF milestone requirements are:

* acceptance of the system **no later than Q4/2025**
* ambition for Perun to be placed in **top 10 of the Green500** list in the first published list after the installation

The system size is defined by a user survey, which was combined with the above-mentioned conditions, arriving at the following approximate configuration:

|  |  |
| --- | --- |
| **Initial HPC and DC system requirements** | |
| CPU:FAT:CLOUD:GPU nodes ratio | ~ 4:1:1:6 |
| **Total number of cores (resulting from the survey)** | **> 57 000** |
| Computational performance | > 50 PFlop/s |
| Estimated power consumption | ~ 800 kW |
| PUE | < 1.1 |

The main priority parameters for the procurement will therefore be:

* CAPEX cap
* OPEX minimization
* number of cores at least 57k
* peak performance at least 50 PFlop/s (with respect to the RRF goals)

HPC system requirements:

The Perun system must allow effective execution of many concurrently running compute jobs in all phases of their 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 Perun system should at least consist of following logical components:

* Universal compute partition
* Accelerated compute partition
* Cloud infrastructure compute partition
* Large memory (FAT) 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 C-DC
* 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 approximately 4 GB of RAM per core (at least 256 GB per node). The partition should have an expected performance of 0.66 PFlop/s. This partition should be also very good for the inference part of AI workloads. The expected power consumption of the partition should be 100 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 at least 512 GB of RAM and 8 GPUs interconnected by a high speed memory coherent bus providing at least 384 GB of memory and 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 49 PFlop/s. 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 around 460 kW with a usual compute load.

Having in mind the ambition to be listed in the top 10 of the Green500 list, this partition should be considered a separate system to be benchmarked. That means it also has to be listed in the Top500 list and it has to show outstanding performance/power consumption rate.

The **Cloud infrastructure compute partition** should support the operation of partition nodes both ways – classical HPC and the *Infrastructure/HPC as a Service*. Primarily, 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 CPU(s) with a higher core count and smaller base frequency compared to the universal compute nodes (if applicable/reasonable), approximately 8 GB of RAM per core 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 partition should have an expected performance of 0.16 PFlop/s. This partition should be also very good for the inference part of AI workloads. The expected power consumption of the partition should be 25 kW with a usual compute load.

The **Large memory (FAT) 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 approximately 8 GB of RAM per core (at least 512 GB per node). The partition should have an expected performance of 0.16 PFlop/s. This partition should be also very good for the inference part of AI workloads. The expected power consumption of the partition should be 25 kW with a usual compute load.

The **High speed compute network** should be an RDMA based network supported by the GPU accelerators with link speeds of at least 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. 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 partition 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 at least 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 for traditional HPC workloads as a scratch file system. The expected power consumption should be around 20 kW.

The **PROJECT storage** is a centralized storage designed to store the data during the whole time of a user project (even multi-year projects). The total net capacity should be more than 10 PB with a throughput at least 250GB/s and IO rate exceeding 25 M IOPS. 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. 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.

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, cloud, 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 25 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 delivered with complete **power and cooling infrastructure** of the C-DC including 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, in-row cooling towers or similar.

The system should have warranty and should be serviced for 5 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 5 years.

C-DC and supporting infrastructure requirements:

C-DC should be equipped with a security monitoring system, fire protection system and controlled entry, e.g. using security cards for authorized personnel.

The delivery should include an energy efficient and dynamic cooling system, allowing efficient adaptation to workload and environmental conditions and resulting in lower OPEX for the procurer in the long-term. Additionally, redundancy in the cooling architecture is needed to mitigate any impact of single point failure and to allow for component replacement during normal operation. The cooling system should include a monitoring and alarming system, as well.

Interface with the waste heat recovery systems should be prepared as a part of the C-DC delivery. Waste heat recovery will not be implemented immediately after system installation, but rather during the system lifecycle. The C-DC design should enable the procurer to install such a system in the future without any additional major DC alterations.

The C-DC architecture should enable full power supply redundancy. It will be single-fed from one station. UPS and back-up power generators for core systems (management, service and login servers, networking, storage, cooling and the cloud partition (the procurer does not plan to back-up other compute servers).

The C-DC should have a warranty and should be serviced for 5 years.