# Specification of requirements for PROJECT data storage expansion Specification of requirements for PROJECT data storage integration into EURO\_IT4I system

## PROJECT Data Storage Description

### Purpose

IT4Innovations PROJECT data storage is the central user data storage of the IT4Innovations Supercomputer Center. The storage is also partially used for computations.

Short-term job computation and processing user data is primarily stored on SCRATCH storage of individual computing systems. Users with I/O-intensive computations are guided to use SCRATCH storage.

In some cases, PROJECT data storage will be used for computations directly. For example, in the case of low I/O intensive jobs and large data jobs, when transferring data to SCRATCH storage is ineffective (e.g. due to amount of data and capacity requirements).

### Computing Systems

The storage will be shared on all supercomputers of the center – on the existing systems Anselm, Salomon, Barbora, and DGX2; EURO\_IT4I system; other systems; and the LAN/WAN of the center.

### Data Structure

At the time of the analysis, data on a storage of similar purpose had the following properties:

* Large number of files – approx. 140 to 400 million files per 1PiB of used capacity
* Approx. 50% of files with the size of <= 4kiB
* Approx. 98% of files with the size of <= 32MiB
* More than 99% of the capacity is used by files with the size of >= 1MiB
* Approx. 95% of the capacity is used by files with the size of >= 32MiB

### PROJECT Data Storage Schema

The following figure provides a simplified illustration of PROJECT data storage solution in the contracting authority’s infrastructure.



Fig. 1 Data storage PROJECT schema

EURO\_IT4I system, Anselm, Salomon, Barbora, and DGX2 are the contracting authority’s computing systems.

„WAN/LAN“ shows the contracting authority’s centralized WAN/LAN infrastructure providing systems interconnection and Internet connectivity.

„GW“ – Data storage PROJECT Network Gateways

„DGW“ – Data storage PROJECT Data Gateways

### PROJECT Data Storage Components

PROJECT data storage contains *file storages* (hereinafter referred to as file storages). Purpose of file storages is to store and share file data. File storages are implemented as a complex solution of storages, servers (I/O servers, management servers, etc.), networks, and necessary software. File storages offer file services via NFS protocol. File storages are labeled as PROJECT1, PROJECT2, …

PROJECT data storage contains the *network infrastructure*,i.e. the network interconnection of components and systems to achieve the required functionality, provide access to individual services, and ensure performance, serviceability and security.

The network infrastructure consists mainly of PROJECT data storage *Access Network*, file storage *LANs,* and other networks according to the supplier's design (e.g. data networks, SAN).

*Access Network* is an Ethernet network providing storage services to IT4Innovations systems.

*Data networks* provide interconnection of storage devices and servers within file storages.

LAN networks are Ethernet networks providing the communication and device management within file storages.

Access Network consists of two L3 Ethernet switches called *Edge devices.*

PROJECT data storage services are provided to end systems exclusively through Access Network (with the exception of EURO\_IT4I system which also provides a direct connection to the jointly delivered file storage).

Selected computing systems are provided with PROJECT data storage services by connecting PROJECT Access Network with the compute network of computing system using *Network Gateways*.

Other systems in the contracting authority's WAN/LAN are provided with PROJECT data storage services by connecting PROJECT Access Network to the WAN/LAN.

*Network Gateways* provide connection of PROJECT Data Storage Access Network with the computing system network; they provide connection of networks of various physical layer technologies and their routing. Network Gateways are acquired as a part of computing systems in the scope of the computing systems procurement. Network Gateways are not intended for file storage services, they are intended only for network interconnection.

PROJECT data storage also contains additional servers designed to provide additional services over the data storage space known as *Data Gateways*. A specific Data Gateway is assigned to one specific file storage at a time. Data Gateways are intended solely for the provision of other services of the contracting authority, which are implemented and operated by the contracting authority. These are services for the provision and transmission of storage data by other protocols, e.g. GridFTP.

## PROJECT Data Storage Expansion

1. The delivery must expand PROJECT data storage by one file storage, two Data Gateways, including connection to PROJECT data storage Access Network.
2. PROJECT data storage expansion must be independent of EURO\_IT4I system, with the exception of providing access of EURO\_IT4I system to PROJECT file storages.
3. PROJECT data storage expansion solution must be hardware-independent of EURO\_IT4I system and must be placed in separate racks.
4. PROJECT data storage must be software-independent of EURO\_IT4I system and its functionality must not depend on accessibility of EURO\_IT4I system services.

### Requirements – File Storage

1. The file storage must be independent of other file storages of PROJECT data storage. The file storages (equipment and software solution) must be mutually independent. File storage systems must not share any infrastructure, with the exception of PROJECT Access Network (and services provided by the contracting authority). Outage of all equipment of one file storage must not have any negative impact on services of other file storages.
2. (I) In the offer, the supplier must include the file storage solution architecture, type and configuration of the offered equipment, and their connection.
3. (I) In the offer, the supplier must include the file storage software solution description.

#### Requirements – File Services

1. The file storage must provide shared network file system services.
2. From the user’s point of view, the file storage must behave as a single, continuous space with a single namespace. This means that the user uses a single namespace to access the files in the storage, and all the capacity and properties of the file storage are available within that single namespace.
3. File storage clients must provide standard file system functionalities.
4. On the client side, the file storage must be transparently integrated into the operating system, must allow for standard file operations and implement the usual semantics of native file systems, must support a native OS file system interface (API), and integrate operating system users as file system users.
5. The file storage must meet the following requirements:

* Unicode support in file names;
* Long file names support;
* Access control, standard Unix permissions (read, write, execute; user, group, others), and extended ACL;
* User quotas, limits for disk capacity and number of files configurable individually for each user;
* Group quotas, limits for disk capacity and number of files configurable individually for each group;
* Reporting of used capacity and number of files for individual users and groups;
* Support for files larger than 1TB;
* Symbolic links support;
* File locking support.

1. The file storage must be made available to Compute, Login, Visualization, and Data management nodes of EURO\_IT4I system using the native file storage protocol and using Compute Network of EURO\_IT4I system as interconnect network.
2. The file storage must provide NFS protocols version 3 a 4.

If the solutions requires implementation of NFS gateways (NFS export), these must be an integral part of the delivery. The supplier must not use Network Gateways and Data Gateways nodes for implementation of NFS Gateways.

1. The file storage must be made available to computing system nodes and to the WAN/LAN using the NFS protocol. The NFS protocols provided by file storage must be accessible on Access Network (at the edge devices), from where they will be provided to other contracting authority’s systems and networks.
2. The file storage must provide simultaneous connection of 2,000 active NFS clients, at a minimum.
3. The file storage must provide at least four NFS servers (servers providing NFS service) and aggregated network connection speed of these servers to the Access Network edge devices of 200Gb/s.
4. (I) In the offer, the supplier must include the file properties and provided protocols of the file storage solution.

#### Requirements – Properties and Parameters

1. The file storage must have a net available capacity on the provided filesystem-level of 2.5PB (2.5x10¹⁵byte).
2. The file storage must allow storing of 250 million files per petabyte of the file storage net available capacity.
3. The file storage must provide long-term sustainable sequential performance using the native protocol for the 1MiB (2²⁰B) block size of 18GB/s (18x10⁹ byte/s). The required performance must be achieved from clients connected to the Compute Network of EURO\_IT4I system (see SPEC\_13) using the native protocol.
4. The file storage must provide long-term sustainable sequential performance using the NFS protocol for the 1MiB (2²⁰B) block size of 12GB/s (12x10⁹ byte/s). The required performance must be achievable from clients connected to PROJECT Access Network edge devices (see SPEC\_15) using the NFS protocol.

The SPEC\_21 and SPEC\_22 performance requirements do not have to be met simultaneously.

1. The file storage must provide long-term sustainable random I/O performance using the native protocol for the block size of 4KiB (2¹²B) and 80%/20% read/write mode of 18,000 IOPS (I/O operations per second). The required performance must be achieved from clients connected to the Compute Network of EURO\_IT4I system (see SPEC\_13) using the native protocol.
2. The file storage must provide long-term sustainable random I/O performance using the NFS protocol for the block size of 4KiB (2¹²B) and 80%/20% read/write mode of 16,000 IOPS (I/O operations per second). The required performance must be achievable from clients connected to PROJECT Access Network edge devices using the NFS protocol.
3. File storage metadata must be stored on SSD or NVMe disks. The disks must be suitable for their designation and expected load.
4. The file storage must use technology ensuring effective storage and efficient provision of large number of small files. Effective storage is provided, for example, by variable block size technology, sub-block allocation, or storing data into metadata space. Effective data delivery is provided, for example, by specific space data storage technology – fast tier or metadata space.

Space for effective provision of small files must be implemented using SSD or NVMe disks. The disks must be suitable for their designation and expected load.

Space for effective provision of small files must allow storing of:

* 125 million files, 4KB per file, per 1PB of file storage net available capacity and
* 60 million files, 32KB per file, per 1PB of file storage net available capacity.

1. The file storage must effectively work with the provided SSD or NVMe disks and ensure uniform wearing of the flash memory cells.
2. (I) In the offer, the supplier must provide the net available capacity, maximum number of files, long-term sustainable sequential performance for the 1MiB block size and long-term sustainable random I/O performance for the block size of 4KiB and 80%/20% read/write mode as specified in SPEC\_19, SPEC\_20, SPEC\_21, SPEC\_22 SPEC\_23, and chapters 2.1.5 to 2.1.11.
3. (I) In the offer, the supplier must provide the method of saving file storage metadata and solution for provision of small files and their capacities as specified in SPEC\_25 and SPEC\_26.

#### Requirements – Availability, Redundancy

1. The PROJECT data storage solution must provide high availability. PROJECT data storage must not contain any component whose outage could cause a failure of the storage services (no “single point of failure” must exist).

Data storage solution components, specifically disks, power supply units, disk array controllers, switches, and servers must be redundant and hot-swappable without causing storage services outage.

A redundant disk array is not considered a “single point of failure” if the disk array only consists of dual-port disks and there are multiple independent data paths for all disks.

1. Failure or outage of any single file storage server or network device must not cause a failure of file storage services.

During a failure or outage of a file storage server or network device, the file storage performance can be lower than required.

1. PROJECT file storage solution must be designed for long-term heavy load.
2. (I) In the offer, the supplier must provide the details about the solution for high availability and redundancy of the supplied systems and equipment.

#### Requirements – Disk Redundancy

1. The file storage must provide sufficient redundancy so that a failure of any two disks will not cause data loss.
2. The file storage must provide recovery from disk failure, i.e. re-ensuring the required data redundancy (e.g. RAID group reconstruction using hot-spare disks). Recovery from disk failure must perform automatically, without operator’s intervention.
3. Recovery from disk failure, i.e. re-ensuring the required data redundancy, must be completed within 48 hours of disk failure. During recovery, the file storage performance can be temporarily lower than required.
4. The file storage configuration must ensure required data redundancy after a failure of any two disks of the storage, without operator intervention.
5. Each disk array (or similar equipment) of the file storage solution must provide a reserve capacity or spare disks in the number or capacity of

disks, at a minimum.

The max(a; b) function returns the larger value from a, b. The result is rounded up to the nearest integer. The total number of disks in disk array includes the disks with data and parity data, as well as spare disks and reserve capacity.

1. (I) In the delivery, the supplier must provide the details about the file storage redundancy, number and type of disks, RAID level, number of disks in RAID, number of spare disks, and file storage software solution description.

#### Definition – Storage Capacity

1. The data storage capacity (size) is required and must be stated as the net usable capacity, i.e. the storage capacity actually usable by the user at the highest service level provided.

The file storage capacity is the capacity of the file system provided by the storage.

1. The storage net usable capacity must be stated for offered/delivered configuration designed for normal operation.
2. All required/offered capacities (including the metadata capacity required for storing the requested number of files) must be achievable simultaneously.
3. Determination of the net usable capacity must not take into account the features of the system or its component, which could provide larger storing capacity under conditions that cannot be ensured (compression, deduplication, etc.).
4. Determination of the net usable capacity must not take into account the features of the system or its component, which could allow allocating more space than physically possible or practically feasible without further action (oversubscription).
5. The storage capacity is specified using prefixes of decimal multiples.

Gigabyte (GB) 10⁹ byte

Terabyte (TB) 10¹² byte

Petabyte (PB) 10¹⁵ byte

1. The net available storage capacity is determined on a suitable storage client system by writing to the storage until it is full, or reading the entire storage, or by a suitable system tool demonstrating the size/storage capacity.
2. The tools used to determine the capacity must provide reliable information and work with a known data block size or a known and accurate unit.

In Linux OS, the file system capacity in bytes can be displayed using the df -B1 command, value “Available“.

#### Definition – Storage Performance

1. The storage performance (sequential performance, I/O performance, metadata performance) is required and must be stated as the long-term performance realistically sustainable by the user from storage clients at the highest service level provided.
2. The file storage performance is the long-term, realistically sustainable performance of operations performed on file storage systems from file storage clients.
3. The storage performance must be stated for the offered/delivered configuration designed for normal operation.
4. The storage performance must be determined for highly occupied storage (see the benchmark specifications for details).
5. The storage performance must be determined for non-privileged user operations.
6. The performance must not be determined based on the assumption of specific, favorable conditions, or specific, advantageous measurement modes (e.g. cache operation), unless such conditions or modes are explicitly required.
7. The storage performance is specified using prefixes of decimal multiples.

#### Performance Measurement

1. The supplier must demonstrate compliance with the storage performance requirements by running performance tests (benchmarks) within the acceptance tests (performance measurements).
2. The performance tests must be run on test servers configured as file storage clients.

For native protocol performance metrics: EURO\_IT4I Compute servers configured as native clients system must be used for testing. The measurement will be performed exclusively on native protocol mountpoints on the test servers. All data traffic (native protocol communication) must take place via EURO\_IT4I system Compute Network.

1. The test server disk caches must be cleared before each storage performance measurement.
2. The performance measurement must be carried out by procedures and under conditions that correspond to normal operation – normal provision of storage services. No action that would affect the measurement result must be taken prior to and during measurement of the storage performance.
3. (I) In the delivery, the supplier must state the proposal of the performance measurement implementation.

#### Using FIO Benchmark for Storage Performance Measurement

1. The file storage sequential performance and random I/O performance must be measured with fio version 3.13 in client-server mode. Other/newer version of the application can be used only with the approval of the contracting authority.

Fio is an open source tool (GPL license version 2) for I/O benchmarking and testing, available at https://github.com/axboe/fio.

Accessibility of the measured file storage is verified on each server dedicated to file storage measurement (test servers according to SPEC\_56) and the fio tool in server mode is run:

fio --server

The measurement is initiated from the selected server – the fio tool is run in client mode:

time fio \  
 --client=machinefile jobfile \  
 --output-format=normal,json |& \  
tee fio.out

In the *machinefile* file, the names of all test servers for storage measurement are specified; one name per line.

*Jobfile* – the fio tool configuration file – contains the test description according to the specified measurement. The *jobfile* file is stored on the server initiating the measurement.

Fio program must be executed by an unprivileged user.

The file storage directory used to store test measurement files (the directory setting in the *jobfile*) must be empty before the measurement.

The test files must be prepared in the directory before the sequential read performance measurement and before random I/O performance measurement. To create the test files, run the fio tool in the client mode and set the create\_only=1 parameter in the respective *jobfile*. By modifying the bs parameter, you can change the block size, as well.

Test directories and files must be owned by an unprivileged user.

The fio output on the server that initiated the measurement (i.e. the *fio.out* file) will be used to determine the result. The measurement summary information provided under “All clients” will be used – in JSON output format.

In JSON output, the last item of the client\_stats attribute is used, the value of the jobname attribute of this item is All clients, this item is hereinafter referred to as the summary.

The value of the bw\_bytes attribute of the read or write (depending on the operation type) attribute of the summary represents the respective sequential performance (throughput) in bytes per second.

To measure the random I/O performance, the sum of the values of the iops attribute of the read and write attributes of the summary represents the total number of I/O operations per second (IOPS).

#### Sequential Performance Measurement

1. The long-term sustainable sequential performance (SPEC\_21) must be measured with the fio tool according to SPEC\_60.

The Data storage sequential performance is the lower of the two values – file storage sequential read performance and file storage sequential write performance.

For file storage sequential write performance, use the following configuration file – fio *jobfile*:

[global]

rw=write

bs=1M

create\_on\_open=1

time\_based

runtime=2h

numjobs=NUMJOBS

[project]

directory=/projectN/test/fio

filesize=FILESIZE

For file storage sequential read performance, use the following configuration file – fio *jobfile*:

[global]

rw=read

bs=1M

time\_based

runtime=2h

numjobs=NUMJOBS

;create\_only=1

[project]

directory=/projectN/test/fio

filesize=FILESIZE

Replace the NUMJOBS and FILESIZE parameters with suitable values.

The NUMJOBS parameter indicates the number of parallel tasks performed on the test server; the value is an integer.

The FILESIZE parameter indicates the size of individual files that the job is working with; the value is the size according to the fio tool syntax (e.g. 30T).

Total size of the files with which the measurement works from one test server, i.e. the FILESIZE \* NUMJOBS value, must be greater than 20 times the memory size of each test server.

Total size of the files with which the measurement works from all used test servers, i.e. the FILESIZE \* NUMJOBS \* number\_of\_test\_servers value must be greater than 250TB.

#### Random I/O Performance Measurement

1. The long-term sustainable random I/O performance (SPEC\_23) must be measured with the fio tool according to SPEC\_60.

Use the following configuration file – fio *jobfile*.

[global]

rw=randrw

rwmixread=80

bs=4k

time\_based

runtime=2h

numjobs=NUMJOBS

;create\_only=1

[project]

directory=/projectN/test/fio

filesize=FILESIZE

Replace the NUMJOBS and FILESIZE parameters with suitable values.

The NUMJOBS parameter indicates the number of parallel tasks performed on the test server; the value is an integer.

The FILESIZE parameter indicates the size of individual files that the job is working with; the value is the size according to the fio tool syntax (e.g. 30T).

Total size of the files with which the measurement works from one test server, i.e. the FILESIZE \* NUMJOBS value, must be greater than 20 times the memory size of each test server.

Total size of the files with which the measurement works from all used test servers, i.e. the FILESIZE \* NUMJOBS \* number\_of\_test\_servers value must be greater than 250TB.

#### Metadata Performance Measurement

1. In order to verify the stability of the file storage metadata operations and to determine the file storage metadata performance, the long-term sustainable performance of the file storage metadata operations must be measured as acceptance tests.
2. The long-term sustainable metadata performance must be measured using the mdtest tool version 3.2.1 <https://github.com/hpc/ior>. Other/newer version of the application can be used only with the approval of the contracting authority.

For measurement, use the following command:

time \

mpirun -n $NPROC -machinefile \  
mdtest \  
-C -T -r \

-F \  
-d /projectN/test/mdtest \  
-I $FILES\_PER\_DIR \  
-i $ITERATIONS \  
-u \  
-z $TREE\_DEPTH -b $BRANCHING\_FACTOR \  
-L

where *machinefile* is the file containing the names of the test servers (as specified in SPEC\_55), one name per line,

the NPROC, FILES\_PER\_DIR ITERATIONS, TREE\_DEPTH, BRANCHING\_FACTOR variables are positive integers determined so that:

* The FILES\_PER\_DIR value is greater than or equal to 100;
* The number of files created in each iteration is 10⁸, at a minimum, i.e.

;

* The test run time is 2 hours, at a minimum.

The result is the values at the “File creation“ and “File stat“ lines, in the “Mean” column, of the “SUMMARY rate“ table.

### Requirements – Network

#### Requirements – Connection to EURO\_IT4I system Compute Network

1. The File Storage and Data Gateways of the procured PROJECT expansion must be connected to the Compute Network of EURO\_IT4I system.
2. The file storage connection to the Compute Network of EURO\_IT4I system must be redundant.
3. The file storage connection to the Compute Network of EURO\_IT4I system must provide connectivity sufficient for real use of long-term sustainable sequential performance for the 1MiB block size according to the SPEC\_21 requirement and for fulfillment of other requirements of the contracting authority.
4. (I) In the offer, the supplier must provide the description and parameters of the file storage and Data gateways connection to the Compute Network of EURO\_IT4I system.

#### Requirements – Connection to PROJECT Access Network

1. The File Storage and Data Gateways of the PROJECT expansion solution must be connected to PROJECT Access Network.
2. The file storage connection to PROJECT Access Network must be redundant.
3. The file storage connection to PROJECT Access Network must provide connectivity sufficient for real use of long-term sustainable sequential performance for a 1MiB block size according to the SPEC\_22 requirement and for fulfillment of other requirements of the contracting authority.
4. The PROJECT data storage expansion (PROJECT file storage and Data gateways) must be connected to each of the two edge devices of PROJECT Access Network using (up to two) 100Gb/s connections. For this purpose, the contracting authority reserves four 100Gb/s Ethernet QSFP28 type ports in edge devices of PROJECT data storage Access Network (two ports in each edge device). A part of the delivery are modules and optical cables required for connection to edge devices of PROJECT data storage Access Network. Network cabling in the Data center dropped ceiling is required.

PROJECT Access Network and cable routes are described in the Chapter 4.1.1 PROJECT Access Network.

1. The connection to edge devices of PROJECT Access Network must be implemented via Ethernet and must use private IPv4 networks in different routing instances.
2. The connection to PROJECT Access Network must provide especially the following types of communication:
   * Providing file storage data/file services to clients – through network gateways and to the contracting authority’s WAN/LAN;
   * Providing file storage Data gateway services to clients – through network gateways and to the contracting authority’s WAN/LAN;
   * Access to management interface of all storage equipment – from the contracting authority’s WAN/LAN to storage LAN;
   * Access to management services of all storage services – from the contracting authority’s WAN/LAN to storage LAN;
   * A communication with the contracting authority’s infrastructure services (see the requirements in Chapter 2.4 Requirements – Software) – to/from the contracting authority’s WAN/LAN from/to LAN storage.

The connection to PROJECT Access Network can provide other types of communication, as well. However, regardless of this fact, all contracting authority’s Network requirements must be met.

1. The connection to PROJECT Access Network must be divided in different L3 networks. For each L3 network, a different L2 network – VLAN – must be used. It is undesirable to use one L2 network for multiple L3 networks.
2. The connection to PROJECT Access Network must separate the following communication types using independent routing instances:
   * Providing file storage data services and Data Gateways services (numbers 1 and 2 of SPEC\_74) on private network IP addresses (Data Private);
   * Providing file storage data services and Data Gateways services (numbers 1 and 2 of SPEC\_74) on public network IP addresses (Data Public);
   * Access to device management interfaces and to management services of the procured file storage and communication with contracting authority’s infrastructure services (numbers 3, 4, and 5 of SPEC\_74) on private addresses (Management).
3. (I) In the offer, the supplier must provide the description and parameters of the file storage and Data gateways connection to PROJECT Access Network.

#### Requirements – LAN Networks

1. The LAN networks of procured PROJECT expansion must be divided into different L3 networks. For each L3 network, a different L2 network must be used (VLAN or another active device). It is undesirable to use one L2 network for multiple L3 networks. LAN separation must ensure especially separation of the following operations:

* Communication between servers;
* LAN active devices management;
* Disk arrays and storage equipment management;
* Server (BMC, IPMI, etc.) management;
* Non-IT infrastructure (power supply, cooling, etc.) management.

1. The Ethernet network active devices must provide remote control and central management using encrypted management interface with strong encryption, strong keys, and strong hash algorithms.
2. (I) In the offer, the supplier must provide the LAN topology and list of used equipment.

### Requirements – Data Gateways

1. Each file storage of PROJECT data storage must include two dedicated servers reserved solely for the contracting authority’s purposes known as Data Gateways.
2. The data gateways are intended solely for the provision of other contracting authority’s services over the file storage data space (provision and transmission of storage data by other protocols, e.g. GridFTP). The contracting authority will implement its services on the servers. The supplier must not use the Data Gateways for any other purpose.
3. The supplier must provide a comprehensive solution for the native file storage client to the Data Gateway servers (operating system, file storage client, availability monitoring, etc.).
4. Each Data Gateway server must met the following requirements:

* Physical server;
* x86-64 architecture;
* Linux 64-bit, CENTOS 7, or RHEL 7 operating system;
* At least one processor, at least 16 CPU cores total;
* At least 128GiB RAM, operated in DDR4 mode, at least 2666MT/s with ECC;
* The RAM must consist of memory modules of the same parameters, evenly distributed across all server memory;
* Theoretical throughput of processor(s) to RAM at least 160GB/s (in the offered server configuration);
* At least 2 local SSD disks with the capacity of 120GB, at a minimum, in RAID1;
* Hot-swap disks;
* Redundant, hot-swap power supply units, redundant power supplies.

1. Each Data Gateway node must provide a connection to PROJECT Access Network with aggregated throughput of at least 100Gb/s reserved for contracting authority’s services. See also SPEC\_72. The connection must be implemented by links equally distributed to both edge devices. The connection to PROJECT Access Network must be resistant to link outage.
2. Each Data Gateway node must provide a connection to the respective file storage to ensure the functionality of the file storage client with an aggregate throughput of at least 100Gb/s.
3. (I) In the offer, the supplier must include the solution description and detailed configuration of the Data Gateway serves, including processors type and their connection to the network.

### Requirements – Software

1. A part of the delivery must be a comprehensive software solution of the file storage, i.e. all software equipment required to fulfill the contracting authority’s requirements (firmware, operating systems, drivers, file system, management and monitoring software, etc.).

All delivered software must be licensed for unlimited time, with no additional fees.

#### Software – OS and Applications

The contracting authority prefers a Linux-type operating system for servers, specifically Red Hat Enterprise Linux or CENTOS, major version 7 or newer.

The contracting authority further prefers the following software:

* For management, configuration, and automation - Puppet or Ansible
* For file versioning – GIT

1. Delivered operating systems and software must provide updates, in particular security, reliability, functionality, and performance updates.

The supplier must provide and install software updates according to conditions specified in the contract.

#### Software – Addressing, Name Services

1. Address prefixes/ranges, domain name suffixes, namespaces, user accounts and groups are managed and assigned solely by the contracting authority.
2. IP addressing must be in accordance with contracting authority’s address policy and plan. Unless already specified in another section of the documentation, used IP address ranges must be specified after consultation with the contracting authority.
3. IP addressing can be static or dynamic (using DHCP). The contracting authority can provide DHCP servers for assignment of IP addresses.

1. All used IP addresses must resolve to names. IP addresses must be resolved by DNS servers provided by the contracting authority; the supplier will provide a list of the devices network interfaces, their IP addresses, and names.
2. For the provided file services, the file storage must integrate and use user accounts and groups provided by the contracting authority.

The contracting authority provides centralized management of all IT4Innovations users and groups. The user and group information is provided by the contracting authority via LDAP, using the OpenLDAP software. LDAP services are provided by two LDAP servers (with a replicated database). For implementation of user accounts and groups LDAP schemata, posixAccount and posixGroup are used.

1. Access to the contracting authority’s LDAP services must use a secure LDAP protocol.

To provide high availability, access to contracting authority’s LDAP services must use both LDAP servers provided by the contracting authority.

For integration, the contracting authority prefers SSSD technology. In the case of SSSD technology, SSSD enumeration is preferred for better response.

1. The existence and status of user accounts and groups in the file storage must reflect the existence and status of user accounts and groups in the LDAP service provided by the contracting authority. Expressed in units of time, the maximum delay tolerated is 12 minutes.
2. Storage equipment and systems must use accurate time. Time synchronization must be done using the contracting authority’s NTP servers in the contracting authority network.

#### Software - Management

1. The storage solution must include tools for management of all systems and services, which are part of the system.
2. The storage solution must include tools for remote management of all hardware devices (servers, disk arrays, network switches, etc.) particularly allowing for configuration and management of the devices, and critical event detection and notification via SMTP (email).

The contracting authority can provide a virtual server in their infrastructure for installation of remote management tools.

#### Software – Monitoring

1. The storage solution must include monitoring of accessibility and status of components and services (hereinafter referred to as availability monitoring). Availability monitoring must provide information about accessibility of all delivered equipment available in network over IP and information about accessibility/status of all relevant components/services of supplied servers, storages, etc. Relevant components/services are those that affect the storage system services’ functionality, accessibility, and/or performance. Availability monitoring must classify accessibility/state based on severity (OK, Warning, Critical). Availability monitoring must allow configuring thresholds of monitored parameters for severity classification.

Availability monitoring must provide identification of the equipment, component, or service to which availability or status information relates. Availability monitoring must provide timestamp of objects’ accessibility/state change and record them in a log.

The contracting authority prefers the Icinga2 software in clustered, highly available configuration.

For implementation of the aforementioned preferred solution, the contracting authority can provide virtual servers in their infrastructure for the Icinga2 cluster. In other cases, the contracting authority does not provide any resources/servers for availability monitoring implementation.

1. All important states and inaccessibility of the file storage components and services must be reported to a single availability monitoring, according to SPEC\_100.
2. Availability monitoring as specified in SPEC\_100 must particularly provide the following information:
   1. Accessibility of device network interfaces in the IP networks (including accessibility of management interface, e.g. BMC interface) – for all equipment with network interface
   2. Equipment power supply status – for all servers, disk arrays, and network devices
   3. Equipment cooling status according to the temperature of the equipment and functionality of cooling fans – for all servers, disk arrays, and network devices
   4. Status of the equipment according to its indicated status and/or the record in the equipment’s log – for all servers, disk arrays, and network devices
   5. Servers RAM status (indicates memory error occurrence)
   6. Server CPU load status calculated for one CPU core of the server
   7. Server memory usage status
   8. Server status according to server file system capacity used
   9. Server status according to the total number of server processes
   10. Server status according to the number of server zombie processes
   11. Key server processes/services status (existence, basic response)
   12. Status of server network ports used by key server processes/services (accessibility)
   13. Network interfaces status (Ethernet, FC, InfiniBand, etc.) of the server (up/down)
   14. Server disks status according to disks’ state and properties (up/down, disk failures, interface error (SAS, SATA, etc.), multipath status, etc.)
   15. Status of disk array according to disk array controllers state and the status of communication/data interfaces and disk array ports (frontend and backend).
   16. Disk array status according to the state of logical objects of disk array (RAID/volume group, logical volume)
   17. Disk array status according to state of physical disks in disk array
   18. Network switches status according to the state and properties of network interfaces (up/down, line speed) with respect to the expected state and configuration
   19. Status of services provided in the high availability cluster and status of the high availability cluster itself
   20. Availability of NFS services of the file storage in Access network
3. The file storage must collect performance and capacity parameters (metrics) and provide their visualization (hereinafter referred to as performance monitoring). Performance monitoring must provide data collection and charts for :
4. current capacity usage and free capacity
5. current total number of files
6. current data throughput (read, write, total)
7. current number of I/O operations (read, write, total; metadata operations)
8. current capacity usage and throughput of physical storage (disk array, metadata/object storage target, etc.)
   1. current throughput through server network interface to Access network

at a minimum, for three years, in 5-minute intervals for the first month.

Performance monitoring must provide API for automated provision of data (metrics values).

The contracting authority prefers time metrics collection into time-series database (Graphite or InfluxDB) and chart visualization using the Grafana software.

For implementation of the aforementioned preferred solution, the contracting authority can provide a virtual server in their infrastructure for time-series database operation. Furthermore, the contracting authority can provide the supplier with access to their current Grafana system for visualization integration. In other cases, the contracting authority does not provide any resources/servers for performance monitoring implementation.

#### Software – Logging

1. The file storage must log and store information about activities, status changes, events, etc. Logs must include timestamp; system, service, and user identification; and event description.

Logs must be stored locally for a period of 7 days, at a minimum.

Logs must be forwarded to contracting authority’s remote syslog (for centralized processing).

#### Software – Mail

1. EURO\_IT4I system must send all email communication exclusively through the contracting authority’s SMTP servers.

The contracting authority will provide SMTP servers.

#### Software - Security

IT4Innovations National Supercomputing Center acquired the Information Security Management System Certification in accordance with ISO 27001 (ISO/IEC 27001:2013, ČSN ISO/IEC 27001:2014). EURO\_IT4I system must be implemented in accordance with the internal regulations of the contracting authority.

1. The system must provide access and services only to authorized users and systems. The system must not provide access or services to unauthorized users and systems. The system must be secured against data leaks, service misuse, and service and system breach.
2. Systems and services must use secure, strong passwords and secure keys, secure encryption, and secure protocols. Using default or weak passwords and/or keys is not permitted. Using identical authentication data for different accounts or services is not permitted.
3. Services not required for proper operation and functionality of the solution must not be operated on the nodes, preferably not even installed.
4. The storage system must not communicate with other systems without explicit contracting authority’s permission.
5. (I) In the offer, the supplier must describe the software solution and the names and number of licenses of the proposed software.

## PROJECT Data Storage Integration Equipment

1. The delivery must include equipment required for the integration of the contracting authority's PROJECT data storage (existing PROJECT file storages) into EURO\_IT4I system.

The contracting authority aims to commission PROJECT data storage services on the Compute nodes, Login nodes, Visualization nodes, and Data management nodes of EURO\_IT4I system. The integration will be a joint effort of the contracting authority, PROJECT data storage supplier, and EURO\_IT4I system supplier.

1. The delivery must include, at a minimum, two Network Gateways.

Network Gateways are intended for connecting PROJECT data Storage Access Network with EURO\_IT4I system Compute Network.

1. Each Network Gateway server must met the following requirements:

* Physical server;
* x86-64 architecture;
* Linux 64-bit, CENTOS 7 (or newer) or RHEL 7 (or newer) operating system;
* At least one processor, at least 16 CPU cores total;
* At least 32GiB RAM, operated in DDR4 mode, at least 2666MT/s with ECC;
* The RAM must consist of memory modules of the same parameters, evenly distributed across all server memory;
* Theoretical throughput of processor(s) to RAM at least 160GB/s (in the offered server configuration);
* At least 2 local SSD disks with the capacity of 120GB, at a minimum, in RAID1;
* Hot-swap disks;
* Redundant, hot-swap power supply units, redundant power supplies.

1. Each Network Gateway node must provide a connection to PROJECT Access Network with aggregated throughput of at least 100Gb/s. For this purpose, the contracting authority reserves two 100Gb/s Ethernet QSFP28 type ports in edge devices of PROJECT data storage Access Network (one port in each edge device). A part of the delivery are modules and optical cables required for connection to edge devices of PROJECT data storage Access Network. Network cabling in the Data center dropped ceiling is required.

PROJECT Access Network and cable routes are described in the Chapter 4.1.1 PROJECT Access Network.

1. Each Network Gateway node must provide a connection to EURO\_IT4I system Compute Network with aggregated throughput of at least 100Gb/s.
2. The supplier must cooperate with the contracting authority in:

* Provisioning of Compute network adapter software stack on Network Gateways;
* Provisioning of NFS mounpoints of PROJECT file storages on the Compute nodes, Login nodes, Visualization nodes, and Data management nodes.

1. (I) In the offer, the supplier must state count and detailed configuration of Network Gateways.

## Contracting Authority’s Infrastructure

PROJECT data storage is not yet implemented and the contracting authority does not know the exact equipment specifications.

#### PROJECT Access Network

PROJECT data storage Access Network consists of two Ethernet switches called Edge devices. Edge devices are L3 switch devices that provide L2 and L3 services at full speed on all device ports in a non-blocking manner. The edge devices are built as a multichassis with a common data plane and with multichassis etherchannel support or functionally identical technology. The throughput between the edge devices is at least 400Gb/s. Edge devices have redundant hot-swap power supplies. In the event of failure of half of the power supplies of each of the edge devices, network operation is not affected in any way.

The edge devices support IPv4 protocol – unicast and multicast – and IPv6 – unicast and multicast. The edge devices allow running a routing table with a total of 1,000 routing records. Each edge device supports at least 20 separate routing instances without using MPLS. The edge devices allow at least 250 VLANs, with a VLAN numbering from 1 to 4094. The edge devices allow 802.1Q encapsulation. The edge devices allow data traffic restrictions by access control lists (ACLs). The edge devices allow a separate ACL configuration for each device port. Each edge device allows the configuration of ACLs with a minimum of 2,000 output and 2,000 input rules.

The edge devices provide free ports for expanding PROJECT data storage and making it available to other contracting authority’s systems. Each edge device provides at least 12 free 100Gb/s Ethernet ports of the QSFP28 type.



Fig. 2 WAN rack, cable routes