# Technical requirements specification for IT4I complementary systems II

**Important notice! This procurement is commenced repeatedly.**

**Alterations compared to the** [**previous tender procedure**](https://zakazky.vsb.cz/contract_display_1488.html) **reflected in technical requirements are highlighted in yellow and written in red color. Those alterations relate to the memory size of Compute partition 2 and the licensing model of Compute partition 4.**

Scope of this procurement is creation of development environment for user that need to port and optimize their codes and application for various hardware architectures and software technologies that are not currently present at IT4I production HPC systems or in Czech Republic. The focus is on modern HPC architectures that are either currently used by other HPC centers around the world or are planned for the near future for large scale pre-exascale or exascale systems. The contracting authority is interested in CPU architectures and accelerator architectures.

The procured system should contain small number of compute nodes of different architectures connected using a high performance interconnect together with software equipment needed for an effective operation.

No high-capacity storage is considered as user home directories are expected to reside on procurer’s NFS storage (called PROJECT). As scratch storage we envision to use only local NVMe’s, no global scratch is required.

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.



Fig. 1 Suggested topology of interconnection of the Complementary system II components

The system should at least consist of the following logical components:

* **Compute partition 1** (ARM + Nvidia GPU + DPU)
* **Compute partition 2** (Power10)
* **Compute partition 3** (high capacity L3 cache processor)
* **Compute partition 4** (virtual GPU accelerated workstations)
* PDUs with outlet level power monitoring capabilities
* High speed interconnect
* Network infrastructure - LAN
* Software equipment
* Integration into data center

A set of common requirements for **all compute partitions**:

* Nodes are equipped with local SSD NVMe disk.
* Nodes are equipped with baseboard management (if available on given platform).

## Compute partition 1

**Compute partition 1** should be based on the ARM processor technology and CUDA programmable GPGPU accelerators of Ampere architecture and DPU network processing units. The goal of this system is to enable development of GPU and DPU accelerated applications with ARM processor.

Node architecture specification:

* single socket with Arm v8.2+ 64-bit based processor designed for commercial/datacenter environment
  + min. 80 Cores at min. 3 GHz
  + e.g. Ampere Altra Q80-30 (80c, 3.0GHz), 210W
* two GPU accelerators that can be programmed using CUDA
  + each connected to the CPU via PCIe Gen4 x16
  + each min. FP64 peak performance 5TF, min. FP64 Tensor Core peak performance 10 TF
  + each min. 24 GB HBM2 memory
  + e.g. Nvidia A30
    - <https://www.nvidia.com/content/dam/en-zz/Solutions/data-center/products/a30-gpu/pdf/a30-datasheet.pdf>
* one dedicated network adapter for LAN connectivity
* one dedicated network adapter for High speed interconnect
* two network accelerators (DPU)
  + each accelerator must:
    - be connected to CPU via PCIe Gen4 x16
    - have min. 8 cores
    - have min. 16GB of On-board DDR4 Memory @ 3200MT/s
    - have min. 64GB eMMC memory
    - have min. 1x QSFP56 port
      * protocol support:
        + InfiniBand: min. EDR/HDR100/HDR
        + Ethernet: min. 10/100 Gb/s
  + e.g. NVIDIA BlueField-2 DPU - MBF2M345A
* fast local NVMe storage of size 3 TB
  + min. NVMe 4.0 x4
* min. 512 GB RAM
  + min. 16x 32 GB DIMMs DDR4, min. 3200MHz
  + with ECC support

We require **two nodes** for Compute partition 1.

A possible implementation of the Compute partition 1 is: <https://developer.nvidia.com/arm-hpc-devkit>

Required software installed: The installed software must seamlessly support the GPU and the DPU accelerated applications from the ARM architecture environment, we require Nvidia HPC SDK (available free of charge).

## Compute partition 2

**Compute partition 2** should be based on IBM Power10 architecture. The key technologies this platform should provide are: (i) Power10 architecture and (ii) PCIe gen 5 that can be potentially used for very fast local storage system. This node should have high memory capacity, with all memory slots occupied by 16 or 32GB modules, we expect DDR4 technology. The goal is to provide POWER 10 processor architecture, along with the ability to support heavy I/O capability to NVMe storage.

Node architecture specification:

* dual-socket server with two Power10 CPUs
* each CPU must meet the following parameters
  + min. 12 cores per CPU
* min. 512 GB of DDR4 memory
  + each CPU must have fully populated DDR4 DIMMs for maximum memory bandwidth
  + min. 350GB/s memory bandwidth per socket
* local NVMe storage
  + 2x min. 1,5TB - for system (min. PCIe Gen 4)
  + 2x min. 6TB - for data (min. PCIe Gen 4)
* one dedicated network adapter for LAN connectivity

Required software installed: We require the vendor-optimized scientific libraries and compilers installed, such as xlf, xlc and the ESSL library:

* IBM community editions: IBM XL C/C++&Fortran/ESSL CE

We require **one node** for Compute partition 2.

## Compute partition 3

**Compute partition 3** is aiming to provide our users modern CPU with a very large L3 cache, over 750 MB. The goal is to enable our users to develop algorithms and libraries that will efficiently utilize the new CPU with this very large new cache technology. The processor is expected to be very efficient, for example, for linear algebra on relatively small matrices.

Node architecture specification:

* dual-socket server with two AMD Milan-X (or Trento) architecture CPUs, each with:
  + min. 64 cores
  + min 750 MB L3 cache
* min. 256 GB RAM, min. DDR4 min. 3200 MHz
  + 16× 16 GB DIMMs, or
  + 16× 32 GB DIMMs
* one dedicated network adapter for High speed interconnect
* one dedicated network adapter for LAN connectivity
* local NVMe storage
  + 2x min. 3TB (min. PCIe Gen 4) in RAID 1 (mirror)

We require **one node** for Compute partition 3.

Required software installed: We require the vendor optimized scientific libraries and compilers are installed, such as the AOCC and related libraries.

## Compute partition 4

The **Compute partition 4** is designed to provide our infrastructure user with a remote/virtual workstation running Microsoft Windows OS. The goal is to provide a rich graphical environment with a focus on 3D OpenGL or RayTracing-based applications with the smallest possible degradation of user experience.

In particular, we are interested in the following solution: NVIDIA RTX Virtual Workstation (vWS) <https://www.nvidia.com/en-us/design-visualization/virtual-workstation/> which enables us to run multiple instances of OS on a single server.

Node architecture specification:

* two CPUs, each with min. 24 cores
* min. 512 GB of RAM per server
* 2x min. 3 TB of fast NVMe local disks in RAID 1 (mirror)
* two GPUs designed for professional 3D graphic, e.g. RTX A6000 or NVIDIA A40
  + each min. FP32 peak non-Tensor performance 37TF
  + each min. 48 GB GDDR6 memory with ECC, memory bandwidth min. 650 GB/s
  + <https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/quadro-product-literature/proviz-print-nvidia-rtx-a6000-datasheet-us-nvidia-1454980-r9-web%20(1).pdf>
  + <https://images.nvidia.com/content/Solutions/data-center/a40/nvidia-a40-datasheet.pdf>
* one dedicated network adapter for LAN connectivity

We require **two nodes** for Compute partition 4.

Required software installed: We require full integration/installation of both hardware, software and hypervisor solution in the extent of the NVIDIA RTX Virtual Workstation (vWS) <https://www.nvidia.com/en-us/design-visualization/virtual-workstation/>. Licenses for 8 instances are to be included for the following products:

* 8 instances for NVIDIA virtualization (NVIDIA RTX Virtual Workstation),
* hypervisor with support for all 8 instances and both nodes
* users should be able to launch the instances by themselves via web interface without elevated privileges.
* Windows OS (for guest OS) for 10 users with possibility to exchange users every 90 days – these licenses are part of the procurement process and can be assigned to any user of the IT4Innovations infrastructure,
* any other licenses required to run the Windows and Linux OS under the NVIDIA RTX Virtual Workstation solution as a whole.

## PDUs with outlet-level power monitoring capabilities.

* PDUs with outlet-level power monitoring
* Every ITC device will be connected, every power output must be monitored for electric power
* Sampling rate min 1 sample per second
* Internal or external memory for recording the samples of min. 24 hours of monitoring of all outlets
* API to download samples to external computers and databases

## High speed interconnect

Servers in partitions 1 and 3 should be connected using global high-speed interconnection intended for computations. We expect standard HPC (RDMA) network technology to be used, we coinsider Infiniband EDR or HDR technology. Interconnections should provide bandwidth at **least 100Gb/s** (port to port) and latency lower than 10 microseconds. Compute partitions 2 and 4 are excluded from this requirement.

High speed network connectivity must be configured and demonstrated as functional.

## LAN infrastructure

The system must include equipment for complete implementation of secure LAN infrastructure and its interconnection with the procurer’s WAN/LAN central devices. LAN is intended for internal communication as well for access to the system and external services; particularly for access to management of nodes, access to nodes, data access/transfers (file services), access to services. The LAN must consist of individual L3 networks that should be based on individual L2 networks (represented either by a VLAN or by separate hardware equipment). Connection to individual node must provide bandwidth at least 10Gb/s. For interconnection with the procurer’s WAN/LAN central devices two 100Gb/s links must be used.

LAN network devices must support user authentication using RADIUS or TACACS+ protocol, creating of different user roles for network management (operator, administrator, etc.), and issued commands logging.

LAN network devices must support import and export of its device configuration using TFTP, FTP, SCP, or SFTP protocol. Configuration must allow saving in the form, which allows its editing using text editor.

Active network devices must support SNMPv3 protocol enabling:

1. Reading of device state and ports utilization
2. Sending of SNMP traps for defined events

Management interfaces of active network devices must be connected to the contracting authority’s OOB network.

Serial management interfaces of edge devices must be connected to the contracting authority’s OOB router (Cisco 2901/K9). The supplier must provide required cables and installation.

For the purpose of interconnection with procurer network, single port in each WAN/LAN central devices WAN1 and WAN2 is reserved. Connection equipment (cables, modules) used to connect the LAN network to the central devices is to be compatible with Cisco Nexus 9336-FX2 model of WAN1 and WAN2 devices.

The delivery must include the modules and fiber-optic cables required for connection of LAN edge devices to the WAN1 and WAN2 devices. It is required to install network cabling in the data center suspended ceiling.

Connection of LAN edge device and the contracting authority’s WAN devices must be realized over Ethernet and private IPv4 in separate routing instances.

LAN edge devices must be capable of filtering traffic with IP based access control list.

Edge devices must support IPFIX (Internet Protocol Flow Information eXport), or NetFlow v9, or sFlow v5. Data flow export must provide information about all IP packets (NetFlow standard, full NetFlow), or at least about one packet from 10 processed IP packets (sFlow, sampled NetFlow standards, sampling rate 1:10).

LAN network connectivity must be configured and demonstrated as functional within the LAN infrastructure, all servers and their management ports must be connected, configured and demonstrated as reachable.

## Software equipment

Delivery should provide **software equipment** (operating system, drivers, libraries, and development environment) required for efficient use of compute partitions. The operating system for nodes (except for partition 4) is Linux, RHEL based. If needed, appropriate licenses should be provided as well for 4 years.

Operating system on all servers must provide SSH server and client, bash shell, SSSD (including SSSD LDAP provider), idmapd, and NFSv4 client (using system security).

## Integration into data center

The supplier must implement and deploy Complementary systems in the contracting authority’s infrastructure – IT4Innovations data center (hereinafter referred to as integration into the data center).

Integration into the data center includes all deliveries and activities, the result of which must be deployment of Complementary systems in IT4Innovations data center premises and infrastructure.

Integration into the data center must be carried out in accordance with applicable legislation and regulations and in accordance with the requirements and recommendations of the manufacturers of individual systems. For installed systems, revisions required by legislation and regulations must be carried out.

The supplier must respect the contracting authority’s infrastructure. The infrastructure is described in chapter Contracting Authority Infrastructure.

Extension of the contracting authority’s infrastructure necessary for implementation of Complementary systems (if needed) must be a part of the delivery.

### Requirements – Placement

All ICT equipment of Complementary systems must be designed for installation into racks and must be installed into racks. Racks required for installation of the system are provided by the contracting authority, they are described in chapter Racks.

Cabling must provide reliable connection, handling of equipment (e.g. when servicing the equipment), and prevent damage to the cables during handling.

### Requirements – Transport

During equipment transport and installation, the load-bearing capacity of the floor on transport route must not be exceeded.

The load-bearing capacity of the floor in the data center and its access corridor (room 219 and 223) is 25kN/m2.

The load-bearing capacity of the floor in the access area (room 217 and 218) is 5kN/m2. For the transport of material, the supplier must temporarily install load distribution plates (e.g. plywood panels) in this area, so that the resulting load of the floor during the physical delivery does not exceed 5kN/m2 and thus prevent floor damage.

## Implementation and Further Actions

### Implementation

A part of the delivery must be a comprehensive implementation of Complementary systems so that all the contracting authority’s requirements are met.

The delivery must include design, delivery, installation, implementation, configuration, and performance of acceptance tests.

### Documentation

The supplier must provide a documentation of Complementary systems.

The documentation must include English documentation (datasheets/spec sheets, manuals, administrator, and user guides) of delivered hardware and software. Documentation must be provided in electronic form, allowing text copying.

### EC Declaration of Conformity

All delivered systems and equipment must be accompanied by EC declaration of conformity.

### Waste Disposal

A part of the delivery is the disposal of waste produced by implementation of the delivery.

The contracting authority is not obligated to store packaging and packaging material and will not do so.

## Warranty and Services

The system must have warranty and be serviced on-site for 4 years.

## Contracting Authority Infrastructure

Complementary systems will be installed and operated in the data room of the contracting authority’s data center in the IT4Innovations building. The IT4Innovations building is located on the campus of the Technical University of Ostrava, Studentská 6231/1B, 708 00 Ostrava-Poruba.

Existing systems of the contracting authority are located in the data room.

### Racks

For the installation of Complementary systems, the contracting authority provides up to two 42U 19-inch racks SGI D-Rack equipped with water cooled back doors (three water cooling coil assembly for 10U Space in each rack).

Rack width - the standard width of the 19-inch D-rack is 24 inches.  
Rack Depth - the standard depth of the 19-inch D-rack is 40 inches.



Fig. 2 D-Rack

Distance (gap) between left post (panel mount) and right post is 450mm (standard 19” rack), distance between front post and back post is 725mm.



Fig. 3 PDU

Expected placement of racks intended for installation of Complementary systems is shown in the figure below.



Fig. 4 Complementary systems racks placement

### WAN/LAN Network

Central devices of the contracting authority’s WAN/LAN are L3 switches Cisco Nexus 9336-FX2. The central devices are structured as multichassis with common data plane and multichassis EtherChannel support. The central devices provide HSRP technology.

The contracting authority’s WAN/LAN switches and OOB devices are located in the WAN rack.

The location of the WAN rack and cable routes is illustrated below.



*Fig. 5 WAN rack, cable routes*