# Technical requirements specification for IT4I complementary systems II

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 (in particular Japan and United States) 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.



Figure 1 Suggested topology of interconnection of the Complementary system 1 components

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

* **Compute partition 1** (ARM + Nvidia GPU + DPU)
* **Compute partition 2** (Power10)
* **Compute partition 3** (high L3 cache processor + DPU)
* **Compute partition 4** (virtual GPU accelerated workstations)
* **Compute partition 5** (Intel + Agilex FPGA + Intel Sapphire Rapids + Intel Xe GPU)
* PDUs with outlet level power monitoring capabilities
* High speed interconnections
* 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).
* A power monitoring system in-band or out-of-band is desired, however not required.

**Please keep in mind the following deadlines for this procurement:**

* **The contract must be signed by 12/2022.**
* **The system must be delivered and installed by 06/2023.**

The **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
  + i.e. 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
  + i.e. Nvidia A30
    - <https://www.nvidia.com/content/dam/en-zz/Solutions/data-center/products/a30-gpu/pdf/a30-datasheet.pdf>
* one dedicated network adaptor for LAN connectivity
* 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
  + i.e.
    - NVIDIA BlueField-2 DPU - MBF2M345A
  + **Question:** Please indicate the availability of NVIDIA BlueField-3 DPU and its applicability for this platform?
  + **Question**: Can DPUs be used as High speed interconnect NIC (100Gb/s Infiniband link)?
* 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 are considering **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 strongly consider Nvidia HPC SDK.**

The **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; ~~(iii) connection of multiple nodes to provide one cache coherent SMP 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. ~~and ability to support memory bound applications via high memory throughput and cache coherent SMP to provide memory volume~~.

Node architecture specification:

* dual-socket server with two Power10 CPUs
* each CPU must meet the following parameters
  + min. 12 cores per CPU
* min. 1 TB 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
  + 2x min. 6TB - for data
  + local NVMe storage connected through PCIe Gen 5 bus
* one dedicated network adaptor for High speed interconnect
* one dedicated network adaptor for LAN connectivity

Required software installed: We expect 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 are considering **one node** for Compute partition 2.

The **Compute partition 3** is aiming to provide our users modern CPU with a very large L3 cache, over 750 MB. ~~In addition, we want to equip it with identical network accelerator (DPU) as used in Compute partition one. In case Compute partition will have just single server this partition will be the counter part for development of the DPU accelerated applications.~~ 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.

Our initial idea of server/partition architecture:

* we expect a **single** server in this partition
* dual-socket server with two AMD Milan-X (or Trento) architecture CPUs, each with:
  + 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 adaptor for High speed interconnect
* one dedicated network adaptor for LAN connectivity

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

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: <https://www.nvidia.com/en-us/design-visualization/virtual-workstation/> which enables us to run multiple instances of OS on a single server.

Our initial idea of server/partition architecture:

* we expect two servers in this partition
* each server should be equipped with
  + two CPUs min. 24 cores
  + min. 512 GB of RAM
  + min. 3 TB of fast NVMe local storage
  + two GPU designed for professional 3D graphic, i.e. RTX A6000 or NVIDIA A40
    - two GPUs interconnected with NVLink bridge
    - <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>
  + min. one High speed interconnect link

Required software installed: We expect full integration/installation of both hardware, software and hypervisor solution in the extent of the <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 servers,
  + users should be able to launch the instances by themselves (using PBS, web interface, …)
* Windows OS (for guest OS) for all 8 instances,
* any other licenses required to run the NVIDIA RTX Virtual Workstation solution as a whole.

Finally, for **Compute partition 5** we consider the Intel-based solution, that will provide access to the latest architecture. The overall focus would be on delivering the environment for the development of accelerated HPC applications using OneAPI. The goal is to provide our users with an environment that enables the development of accelerated applications on modern heterogeneous architectures, in particular, the emerging Intel Xe GPU accelerators and the new generation Agilex FPGA accelerators using a single programming model.

Our initial idea of server/partition architecture:

* we expect two servers in this partition
* each server should have:
  + two Intel-based CPUs, we consider two possibilities, and we will decide based on availability
    - Sapphire Rapids with HBM memory (*preferred*) – **specify availability**
    - Sapphire Rapids without HBM memory (*less preferred*) – **specify availability**
  + Ponte Vecchio 2 Stacks (600W) – with two GPU accelerators connected with XeLink - **specify availability**
  + one Intel Agilex based FPGA accelerator - **specify availability** 
    - i.e. Bittware IA-840F: <https://www.bittware.com/fpga/ia-840f/>
  + min. 512 GB RAM
    - 16× 32GB DDR5 min. 4800 MHz – in the case of Saphire Rapids
  + one dedicated network adaptor for High speed interconnect
  + one dedicated network adaptor for LAN connectivity

Required software installed: With respect to the software environment, we expect **CPU, GPU and FPGA** of this server to be programmable by OneAPI. Adequate software should be installed.

**Infrastructure requirements**

**PDUs with outlet-level power monitoring capabilities.**

* sampling rate min 1 sample per second – higher is appreciated
* 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
* **Questions**: Please specify
  + which PDU you expect to use?
  + how the PDU will store samples for 24 hours?
  + how the samples can be transferred to external database?

**High speed interconnect**

Servers in partitions 1,2,3 and 4 should be connected using global high-speed interconnection intended for computations. We expect standard HPC (RDMA) network technology to be used, we consider Infiniband EDR or HDR technology. Interconnections should provide bandwidth at **least 100Gb/s** (port to port) and latency lower than 10 microseconds. Compute partitions 5 is excluded from this requirement.

High speed network connectivity must be configured and demonstrated as functional within partition 1, partition 2, partition 3, and partition 4.

**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 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.