# 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.

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.

Our initial idea of the node architecture is:

* single or dual socket ARMv8 based processor designed for commercial/datacenter environment
  + i.e. ARM Ampere Altra Q80
* One or two GPU accelerators that can be programmed using CUDA
  + i.e. Nvidia A100 GPU we can also use lower cost solution like A30
    - <https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf>
    - <https://www.nvidia.com/content/dam/en-zz/Solutions/data-center/products/a30-gpu/pdf/a30-datasheet.pdf>
* one or two network accelerators (DPU)
  + i.e. NVIDIA BlueField-3 DPU or NVIDIA BlueField-2 DPU
* fast local NVMe storage of size 3 TB
* 256 or 512 GB RAM

We are considering **one or two nodes** for Compute partition 1, depending on pricing.

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

Required software installed power 10: 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) PCIe gen 5 that can be potentially used for very fast local storage system; (ii) connection of multiple nodes to provide one cache coherent SMP system. These nodes should have high memory capacity, with all memory slots occupied by 16 or 32GB modules, we expect DDR5 technology. The goal is to provide POWER 10 processor architecture, along with 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.

Our initial idea of the node and partition architecture is:

* two or four servers connected back-to-back with dedicated SMP interconnect
* each server with fully populated DDR5 DIMMs (16 or 32 GB per DIMM) for maximum memory bandwidth
  + we expect to reach about 1TB/s memory bandwidth per server
* local NVMe storage connected through PCIe Gen 5 bus – min. 5x size of DRAM

Required software installed: We expect the vendor optimized scientific libraries and compilers installed, such as xlf, xlc and the ESSL library as well as the IBM AI tools.

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

Our initial idea of server/partition architecture:

* we expect single server in this partition
* single or dual socket server with AMD Milan-X (or Trento) architecture
  + with min. 64 cores
  + min 750 MB L3 cache
* 256 GB RAM
* one or two network accelerators identical to ones installed in Compute partition 1

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 remote/virtual workstation running Microsoft Windows OS. The goal is to provide rich graphical environment with focus on 3D OpenGL or RayTracing based applications with smallest possible degradation of user experience.

In particular, we are interested in following solution: <https://www.nvidia.com/en-us/design-visualization/virtual-workstation/> which enables 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
    - <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>
  + single NIC for high performance network

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 6 to 8 instances are to be included.

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 environment for development of accelerated HPC application using OneAPI. The goal is to provide our user with environment that enables development of accelerated applications on modern heterogenous architectures, in particular the emerging Intel Xe GPU accelerators and the new generation Agilex FPGA accelerators using single programming model.

Our initial idea of server/partition architecture:

* we expect two servers in this partition
* each server should have:
  + two Intel Sapphire Rapids based CPUs – if possible, with HBM memory
  + one or two Intel Xe GPU accelerator – if possible, we are interested in dedicated GPU to GPU interconnect
  + one or two Intel Agilex based FPGA accelerators

Required software installed: With respect to software environment, we expect all parts 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 memory for recording the samples of min. 24 hours of monitoring of all outlets
* API to download samples from external computers