# Technical requirements specification for IT4I complementary systems I

All changes in the document compared to previous version are highlighted in red.

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, 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)
* Compute partition 2 (Intel)
* Compute partition 3 (AMD)
* Compute partition 4 (Edge)
* 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 it must contain processors with SVE extension of instruction set and HBM memory. An expected processor is A64FX as installed in the Fugaku machine.

Hardware requirements:

* min. 8 compute nodes (we expect one full 2U chassis)
* Node parameters:
  + 1x A64FX CPU with 48 cores at 2.0 GHz and 32 GB of HBM memory
  + min. 400 GB SSD (m.2 form factor) – mixed used type
  + 1x Infiniband EDR or HDR100 interface (connected to 16x PCI-e Gen3 slot)

Software requirements:

* compiler with commercial support for: C, C++, Fortran languages with support for OpenMP
* optimized vendor numerical libraries with commercial support: BLAS, LAPACK, ScaLAPACK
* MPI library with commercial support

The **Compute partition 2** should be based on x86 architecture (Intel Ice Lake) and in form of dual socket server. The key technologies installed in partition are NVDIMM and FPGA accelerators. We expect this partition to have two servers each with two FPGA accelerators installed. In terms of NVDIMM memory we expect it will be installed in both servers: higher capacity should be installed in one node and lower capacity in the second node.

Hardware requirements:

* We prefer 4U over 2U servers with high capacity for expansion cards so that it is easily upgradable in the future by new PCI-e Gen 4 accelerators. Ideal would be support for up to 8x with PCI-e Gen 4, but minimum should be 6 slots per server in total. We would like to have all slots with 16x PCI-e links.
  + Question – how many PCI-e links there will be per slot?
* both servers with FPGAs should have:
  + 2x 3rd Gen Xeon Scalable Processor
    - min. 32-cores @ min. 2.00GHz
  + 16x 16GB RAM with ECC - min. DDR4-2933
  + 1x Infiniband EDR or HDR100 interface (connected to 16x PCI-e Gen3 slot)
  + 4 TB NVMe local storage – mixed use
* Intel FPGA server 1 – low NVDIMM memory server (2304 GB RAM)
  + 16x Intel® Optane™ DC 128GB Persistent Memory Module
  + 2x FPGA accelerators, each with
    - min. 32 GB DDR4 memory
    - min. 2x 100Gbps links
    - i.e.:
      * Intel FPGA PAC D5005 (Stratix 10 SX): <https://www.intel.com/content/www/us/en/programmable/products/boards_and_kits/dev-kits/altera/intel-fpga-pac-d5005/overview.html>
      * or Bittware 520N: <https://www.bittware.com/fpga/520n/>
* Intel FPGA server 2 – high NVDIMM memory server (8448 GB RAM)
  + 16x Intel® Optane™ DC 512GB Persistent Memory Module
  + 2x FPGA accelerators, each with
    - min. 16 GB of HBM memory and min. 32 GB DDR4 memory
    - min. 2x 100Gbps links
    - 1x PCIe Gen 3.0 or 4.0 x16 interface
    - i.e: Bitware 520N-MX: <https://www.bittware.com/fpga/520n-mx/>

FPGAs design tools usually run for several hours to one day to generate a final bitstream (logic design) of large FPGA chips. These tools are usually sequential, therefore we would like to have a dedicated server for this task. We expect that it should contain one or two CPUs with very high per core performance and not necessarily high number of cores. Whether this CPU is Intel or AMD depends only on the performance in such tools, both options are open. In would be optimal if several (min. 4) such tasks can run on this server in parallel. This server should be used for both Intel and Xilinx FPGAs in this system.

Our initial research led us to these initial requirements:

* FPGA synthesis server - dedicated to FPGA synthesis tools
  + CPU with small number of cores running high frequency (Intel Xeon W or AMD Threadripper CPU)
  + min. 128 GB RAM
  + high IOPS local NVMe storage min. 4 TB

Software requirements:

* Developer tools and libraries for FPGA accelerators
  + Quartus, OpenCL FPGA SDK, board support package, …

Selected FPGAs cards contain min. 2x 100Gbps ports per card. We would like to have these card back-to-back connected using dedicated cables. We would like to have 4 cables long enough to reach any position in a rack.

**Question**: Can FPGAs be interconnected using dedicated or shared (with high performance network) 100 Gbps network switch?

**Question**: Please provide the availability of the Intel Agilex FPGA based accelerator cards. For instance the <https://www.bittware.com/fpga/ia-840f/>.

**Question**: Intel provides support of OneAPI programing for Intel FPGAs. Please provide information which Intel FPGAs do support OpenAPI in the timeframe of Q4 2021.

The **Compute partition 3** architecture should be based on AMD Epyc CPUs in combination with AMD GPUs and Xilinx FPGAs. Key requirement is a fast native interconnect between GPU accelerators.

Hardware requirements:

* We prefer 4U over 2U servers with high capacity for expansion cards so that it is easily upgradable in the future by new PCI-e Gen 4 accelerators. Ideal would be support for up to 8x with PCI-e Gen 4, but minimum should be 6 slots per server in total. We would like to have all slots with 16x PCI-e links.
  + Question – how many PCI-e links there will be per slot?

Both AMD servers should have following parameters:

* 2x AMD Milan based CPU
  + min. 32 cores @ min 2.0 GHz
* 16x 16GB RAM with ECC - min. DDR4-3200
* 2x AMD GPU accelerators MI 100
  + with AMD Infinity Fabric™ Link for fast GPU to GPU communication
* 2x FPGA Xilinx Alveo U250 or preferably U280
* 4 TB NVMe local storage – mixed use
* 1x 100 GBps IB EDR or HDR100

Selected FPGA cards contain min. 2x 100Gbps ports per card. We would like to have these card back-to-back connected using dedicated cables. We would like to have 4 cables long enough to reach any position in a rack.

**Question**: Can FPGAs be interconnected using dedicated or shared (with high performance network) 100 Gbps network switch?

**Question**: We expect that this and AMD partition will be updated with new PCIe cards by IT4I personnel. What would be impact on the warranty?

Software requirements:

* Developer tools and libraries for FPGA accelerators for high level programing of FPGAs (equivalent to OpenCL) not only VHDL or Verilog.
* Developer tools and libraries for AMD GPUs.
* Environment should support to run code developed using ROCm on both GPU and FPGA accelerators
  + See <https://forums.xilinx.com/t5/Xilinx-Xclusive-Blog/AMD-and-Xilinx-Demonstrate-Converged-ROCm-Runtime-Technology/ba-p/1175091>

The **Compute partition 4** should provide overview of the so-called edge computing class of resources. While edge is a very wide class, this partition should cover solutions powerful enough to provide data analytic capabilities (both CPU and GPU) in a form factor which should not require a data center to operate. The solution does not have to be rack-mountable, but it must be able to operate outside datacenter in industrial environment.

This partition should have one node with following parameters:

* 1x x86\_64 CPU
  + TDP max. 100 W,
  + min. 16 cores,
  + min. 250 GFlop/s in double precision
* 1x CUDA programmable GPU
  + TDP max. 70W
  + theoretical performance: min. 8 TFlop/s in FP32
* min. 128 GB RAM
* min. 1.5 TB SSD storage
* required connectivity:
  + 10 Gbps Ethernet,
  + WiFi 802.11 abgn,
  + LoRa connectivity,
  + LTE connectivity
* max. 400 W maximum power consumption

**High speed interconnections**

Servers in partitions 1,2 and 3 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). Compute partitions 4 is excluded from this requirement.

**LAN infrastructure**

The system should 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 should 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 should provide bandwidth at least 10Gb/s, 100Gb/s is envisioned. For interconnection with the procurer’s WAN/LAN central devices two 100Gb/s links will be used.

**Software equipment**

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

**Integration into data center**

The system should contain necessary rack(s) and should be completely integrated into the power and cooling infrastructure of the procurer’s data center including connection to the central measurement-and-regulation system.

**Budget**

The contracting authority has dedicated funds in the amount of CZK 9 million excl. VAT.

**Required response**

**Please provide at least the following information:**

1. Draft of the technical solution and equipment; the contracting authority is interested in various options of the solution and is open to discussion and alternative supplier proposals.
2. Availability of the proposed equipment; expected date of acceptance of the delivery is Q4 2021.
3. Provide a draft of the standard warranty conditions; the contracting authority does not expect above-standard SLA conditions for this procurement. We prefer 3 years warranty time.
4. Would you consider the provided concept to be feasible, do you have any comments or reservations?
5. Estimated performance price, including price breakdown into cost of individual components and subcomponents.

The information on the technical solution and pricing referred to point No. 1 and No. 5 shall be structured according to logical components as mentioned above (at the turn of pages No. 1 and No. 2).