# Technical requirements specification for IT4I complementary systems I

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.

## Components and diagram



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

The procurer requires that the delivered system must consist of following logical components:

* Compute partition 1 – based on ARM technology
* Compute partition 2 – based on Intel technologies
* Compute partition 3 – based on AMD technologies
* Compute partition 4 – reflecting Edge type of servers
* FPGA synthesis server
* High speed interconnect
* LAN infrastructure
* Software equipment
* Integration into data center

Following is a set of common requirements that must be fulfilled by **all servers of all compute partitions**:

* Nodes are equipped with baseboard management (if available on a given platform).
* A power monitoring system in-band or out-of-band is desired, however not required.

We also break the Complementary system 1 into following blocks used for partial delivery of the system as enabled by the contract proposal.

|  |  |
| --- | --- |
| Block # |  |
| 1 | Mandatory infrastructure **(must be part of the first installation)**   * FPGA synthesis server and software equipment * LAN infrastructure * Integration into data center |
| 2 | Compute partition 1 and software equipment, without   * IB Infiniband EDR or HDR100 interface |
| 3 | Compute partition 2 and software equipment, without   * IB Infiniband EDR or HDR100 interface * FPGA accelerators |
| 4 | FPGA accelerators for Compute partition 2 and software equipment |
| 5 | Compute partition 3 and software equipment, without   * IB Infiniband EDR or HDR100 interface * FPGA accelerators * GPU accelerators |
| 6 | FPGA accelerators for Compute partition 3 and software equipment |
| 7 | GPU accelerators for Compute partition 3 and software equipment |
| 8 | Compute partition 4 and software equipment |
| 9 | High speed interconnect including all IB Infiniband EDR or HDR100 interfaces for compute partitions 1,2 and 3 |

*Table 1. Definition of the delivery bocks of the Complementary system 1*

## Compute partitions and servers

### Compute partition 1

The **Compute partition 1** must be based on the processors with **Armv8.2-A** architecture with **SVE** extension of instruction set. We also require the processor to have an **HBM2** memory. An expected processor is **A64FX**[[1]](#footnote-2) as installed in the **Fugaku** supercomputer[[2]](#footnote-3).

**Hardware requirements:**

* min. 8 compute nodes
* node parameters:
  + 1x Fujitsu A64FX CPU
    - Arm v8.2-A ISA CPU with Scalable Vector Extension (SVE) extension
    - min. 48 cores at 2.0 GHz
    - min. 32 GB of HBM2 memory
  + min. 400 GB SSD (m.2 form factor) – mixed used type
  + 1x Infiniband EDR or HDR100 interface
    - connected via min. 16x PCI-e Gen3 slot to the CPU
* PSU: 2+1 redundancy

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

### Compute partition 2

The **Compute partition 2** must be based on Intel Ice Lake x86 architecture and must be delivered in form of a dual socket server. The key technologies installed in partition are Intel NVDIMM memories and Intel FPGA accelerators. This partition must contain two servers each with two FPGA accelerators installed in them.

**Hardware requirements:**

Both servers must fulfill following common parameters:

* must be 2U or larger server
* must contain:
  + 2x 3rd Gen Xeon Scalable Processors
    - min. 32-cores @ min. 2.00GHz
  + 16x 16GB RAM with ECC
    - min. DDR4-2933
  + 1x Infiniband EDR or HDR100 interface
    - connected to CPU via min. 16x PCI-e Gen3 or 8x PCI-e Gen4 interface
  + min. 3 TB NVMe local storage – mixed use type
  + 2x FPGA accelerators
    - Bitware 520N-MX: <https://www.bittware.com/fpga/520n-mx/>

In addition, each server must fulfill following parameters:

* Intel server 1 – low NVDIMM memory server with min. 2304 GB NVDIMM memory
  + 16x 128GB NVDIMM persistent memory modules
* Intel server 2 – high NVDIMM memory server with min. 8448 GB NVDIMM memory
  + 16x 512GB NVDIMM persistent memory modules

**Specific software requirements:**

* FPGA boards must support application development using following design flows:
  + OpenCL,
  + High-Level Synthesis (C/C++) including support for OneAPI
  + Verilog and VHDL

**Interconnection requirements for FPGAs in Compute Partition 2**

* Selected FPGAs cards contain min. 2x 100Gbps ports per card. We require to have these cards back-to-back connected using dedicated cables compatible with maximum speed of the FPGA card ports. Part of the delivery **must be 4 compatible cables** long enough to reach any position in a rack

### Compute partition 3

The **Compute partition 3** must be based on two servers equipped with AMD Milan x86 CPUs, AMD GPUs and Xilinx FPGAs architectures as this combination represents alternative to Intel based ecosystem in Compute partition 2[[3]](#footnote-4). The key requirement is a fast native interconnect between GPU accelerators.

**Hardware requirements:**

Both AMD servers must fulfill following parameters:

* must be 4U or larger server
* must contain:
  + 2x AMD Milan based CPU
    - min. 32 cores @ min 2.0 GHz
  + 16x 16GB RAM with ECC
    - min. DDR4-3200
  + 4x AMD GPU accelerators MI 100
    - Interconnected with AMD Infinity Fabric™ Link for fast GPU to GPU communication
  + 1x 100 GBps Infiniband EDR or HDR100
    - connected to CPU via min. 16x PCI-e Gen3 or 8x PCI-e Gen4 interface
  + 3 TB NVMe local storage – mixed use

In addition, each server must fulfill following parameters:

AMD server 1 must have:

* 2x FPGA accelerator
  + Xilinx Alveo U250: <https://www.xilinx.com/products/boards-and-kits/alveo/u250.html#solutions>

AMD server 2 must have:

* 2x FPGA accelerators,
  + Xilinx Alveo U280: <https://www.xilinx.com/products/boards-and-kits/alveo/u280.html#specifications>

**Interconnection requirements for FPGAs in Compute Partition 3**

* Selected FPGAs cards contain min. 2x 100Gbps ports per card. We require to have these cards back-to-back connected using dedicated cables compatible with maximum speed of the FPGA card ports. Part of the delivery **must be 4 compatible cables** long enough to reach any position in a rack

We expect that this and AMD partition will be updated with new PCIe cards by IT4I personnel. This cannot have an impact on the warranty.

**Specific software requirements:**

* both FPGA boards must support application development using following design flows:
  + OpenCL,
  + High-Level Synthesis (C/C++),
  + Verilog and VHDL
* developer tools and libraries for AMD GPUs.

### FPGA synthesis server

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 part of the systemmust be a dedicated server for this task. This server will be used by development tools needed for FPGA boards installed in both Compute partition 2 and 3. These tools must be available on this server.

**The FPGA synthesis server must fulfill following requirements:**

* single or dual socket machine
* CPU parameters:
  + min. 8 cores at min. 3.6 GHz nominal frequency
  + min. 8 memory channels with ECC
* min. 128 GB of DDR4-3200 or faster memory with ECC
  + memory must be fully populated to maximize memory subsystem performance
* 1x 10Gb Ethernet port used for connection to LAN
* NVMe local storage
  + 2x NVMe disks configured RAID 1, each with
    - capacity min. 3 TB
    - DWPD min. 3

**Software requirements for the FPGA synthesis server:**

* server must have installed developer tools and libraries for both Intel and AMD/Xilinx FPGA accelerators as defined in Compute partition 2 and 3

### Compute partition 4

The **Compute partition 4** must 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 cannot not require a data center to operate. The solution does not have to be rack-mountable, and it must be able to operate outside datacenter in industrial environment.

This must consists of one edge computing server with following parameters:

* 1x x86\_64 CPU
  + TDP max. 100 W,
  + min. 16 cores,
  + min. 210 GFlop/s theoretical max performance 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:
  + min. 2x 10 Gbps Ethernet,
  + min. WiFi 802.11 abgn,
  + LTE connectivity
* max. 500 W power consumption

## High speed interconnect

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) and latency lower than 10 microseconds. Compute partitions 4 and FPGA synthesis server are excluded from this requirement.

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

## 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. The compute partition 1 is exempt from this condition, at least 1Gb/s is required for nodes in this partition. 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 or CentOS 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 10 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 (excluding Compute partition 4 - edge computing server) 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 10.1 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.

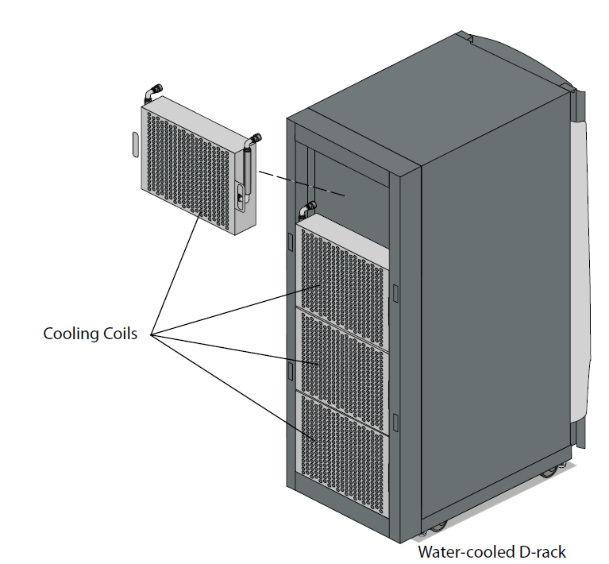


Fig. 1 D-Rack

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

Each rack contains 12 sets of 4x C13 and 1x C19 sockets (6 sets on the left side, 6 sets on the right side of the rack), i.e. 48x C13 and 12x C19 sockets in total. Each set of power sockets can provide current up to 16A.

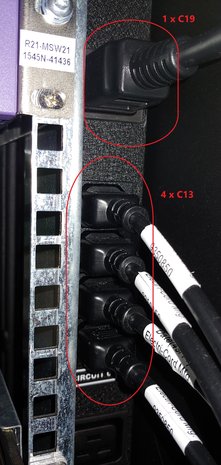


Fig. 2 PDU

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

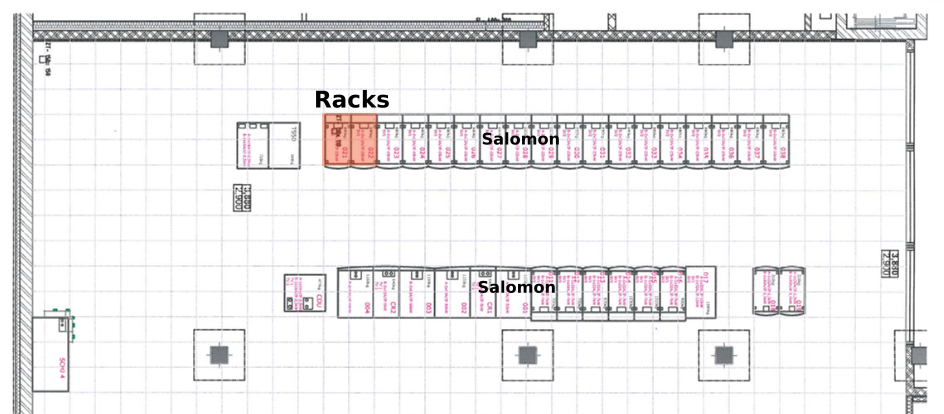


Fig. 3 Complementary systems racks placement

### 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.

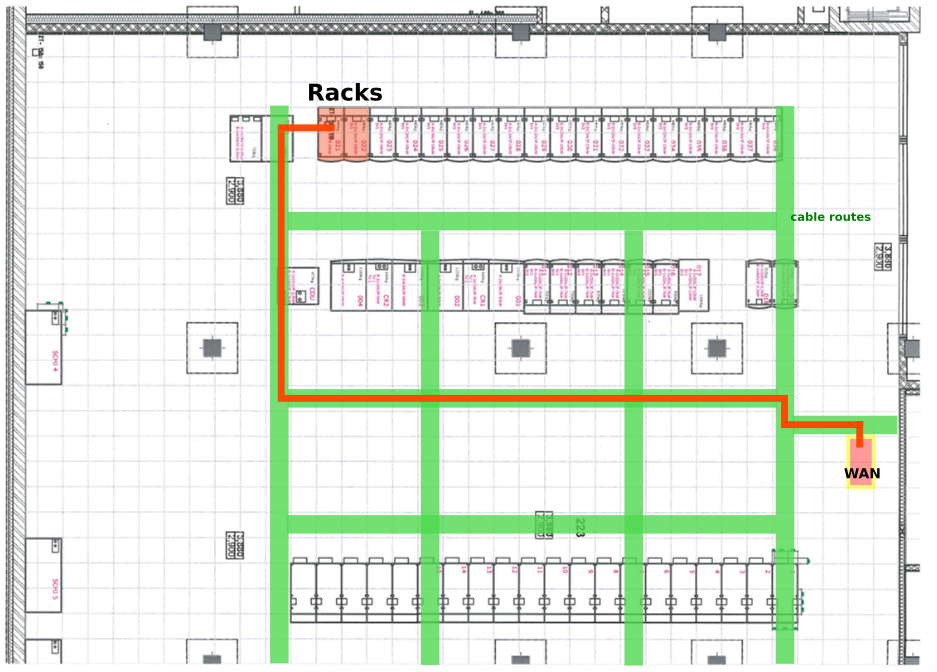


Fig. 4 WAN rack, cable routes

1. A64FX processor: <https://www.fujitsu.com/global/products/computing/servers/supercomputer/a64fx/> [↑](#footnote-ref-2)
2. Fugaku supercomputer: <https://www.fujitsu.com/global/about/innovation/fugaku/specifications/> [↑](#footnote-ref-3)
3. Unfortunately within the timeframe of this procurement Intel Xe GPU accelerators are not available. [↑](#footnote-ref-4)