# 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, accelerator architectures and as well as specialized appliance for AI and ML workloads.

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 (Power)
* Compute partition 4 (AMD)
* Compute partition 5 (Edge)
* Compute partition 6 (Graphcore)
* High speed interconnections
* Network infrastructure - LAN
* Software equipment
* Integration into data center

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

* Minimum number of nodes in each partition is two.
* Nodes are equipped with local SSD NVMe disk.
* Nodes are equipped with baseboard management (if available on given platform).
* If possible, the servers should use 2U form factor for further extension with new accelerator technologies in form of PCIe card.
* 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 set and HBM memory. An expected processor is A64FX as installed in the Fugaku machine and if possible, we are considering the TOFU interconnect.

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. If possible, it should contain also Intel Xe GPU(s), where multiple GPUs per node is important only if there is new interconnect technology (similar to NVLink) between them. We expect this partition to have two servers where FPGA accelerator(s) is installed in one and Xe GPU(s) in the second one. In terms of NVDIMM memory it should be proposed whether larger capacity should be installed in one node or lower in both nodes.

The **Compute partition 3** 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 32GB modules.

In Power9 generation there was a direct NVLink link between CPU and Nvidia GPUs. This support was removed from Power10 generation. In any case we would like to have an NVIDIA GPU accelerator(s) in this system, but we are not aiming for A100 as many of them will be in EURO\_IT4I installation. We aim for an T4 like card based on Ampere generation graphic cards.

The **Compute partition 4** architecture should be based on AMD Epyc CPUs in combination with AMD GPUs where the key requirements is fast interconnect between CPU and GPU accelerator. This partition should also contain a Xilinx FPGA accelerator.

The **Compute partition 5** 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.

* x86\_64 CPU, TDP max. 65 W, at least 8 cores
* CUDA programmable GPUs; GPU TDP max. 70W, at least 6 TFlop/s FP32
* The solution does not have to be rack-mountable.
* For example: Atos BullSequana Edge <https://atos.net/en/solutions/bullsequana-edge/atos-supports-nvidia-egx-ecosystem-as-an-edge-computing-driver>

The **Compute partition 6** should be based on single Graphcore AI/ML appliance which is delivered in form of 1U server and requires management node. For implementation of management node IT4I will provide a virtual machine in its server virtualization infrastructure. This partition is excluded from the common requirements for **all compute partitions.**

**In network computing functionality**

Contracting authority is also looking for in network computing functionality. This should be realized by inclusion of at least two programmable NIC cards into the Compute partition 2 or the Compute partition 4. Each NIC should also contain a tightly coupled GPU accelerator (on the same PCB) programmable with CUDA, see for example EGX 100 product by NVIDIA.

**High speed interconnections**

All nodes in each partition should be connected using high speed interconnections, these interconnections are intended for computations, Global interconnect connecting all partitions nodes is not mandatory but is allowed. Procurer is interested in “native” interconnect for a given architecture, “back-to-back” connection without switches is the preferred way (if possible). As an alternative, standard HPC (RDMA) network technology could be used, we consider Infiniband technology. Interconnections should provide bandwidth at least 100Gb/s (port to port). Compute partitions 5 and 6 are 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).