Questions and answers

Providing graphics processing units in high performance computing environment

1. Publication reference
EuropeAid/133025/MSUP/HR

2. Procedure
Local Open

3. Programme
IPA

4. Financing
IPA 2007 HR 16 I PO 001 Regional Competitiveness Operational Programme 2007-20011 for Community Assistance under the IPA Regional Development Component in Croatia

5. Contracting authority
Central Finance and Contracting Agency

QUESTIONS	ANSWERS
Date: 29.05.2012. 1. In Technical Specification (item 1.5.) it is requested that 1 memory module is installed per memory controller and that memory module size is 2GB and 1333 MHz (in total 288 GB on all nodes). In specified total rack unit height up to 14 U (item 1.1.) it is not possible to fulfill the maximum capacity of 288GB with this type of processor and request of installation of one 2 GB memory module per memory controller? Our recommendation for installation of 12 GB RAM per CPU would be the installation of 6 pieces of 2 GB module per memory controller (2 modules per memory channel).	Date 08.06.2012. The intention of the technical specification is to utilize all available CPU memory controllers, thus obtaining maximum CPU to memory bandwidth. In order to fulfill this requirement it is necessary to have at least one memory module installed per each CPU memory controller. If minimum size memory modules are used (2GB) and in order to fulfill total memory amount requirement - it is necessary to install more than one memory module per each CPU memory controller. Using 2 GB memory modules with 12 memory modules per node (2 memory modules per each memory controller of 2 CPUs with assumption that each CPU have 3 memory controllers) it is possible to reach required memory amount (288 GB) using 12 nodes (servers). With server height of 1U this configuration consumes 12U which is in compliance with the TS. Please bear in mind that requirements in these Technical Specifications are presented as a minimum standard which the offered goods must meet. Using larger memory modules (i.e. 4GB, which is also allowed because, as stated in TS, specifications are "presented as a minimum standard") total memory requirement can be implemented using fewer memory modules (using one 4GB memory module per CPU memory controller).

Date: 29.05.2012. 2. In Technical Specification (item 1.8) it is noted that Infiniband IO needs to have embedded support for GPU acceleration. Can you please clarify what exactly functionality is considered by "GPU acceleration"?	Date 08.06.2012. GPU acceleration for InfiniBand I/O is a technology used to accelerate communication between GPUs located in different servers that are interconnected with InfiniBand network. Currently, there is only one manufacturer of the InfiniBand HBA silicon on the market - Mellanox (with Mellanox InfiniBand HBAs being OEMd by different vendors – IBM, HP, Dell, Bull, etc.). The other manufacturer – qlogic, ceased InfiniBand silicon production since their InfiniBand division was acquired by Intel at the beginning of this year. The latest Mellanox InfiniBand silicon generations (ConnectX-3 & ConnectX-2) implement GPU acceleration technology branded as "GPUDirect".
Date: 29.05.2012. 3. In Technical Specification (items 1.14, 1.15 and 1.16) one Ethernet switch is requested. With only one switch the GPU cluster does not have network redundancy. It is necessary to include two Ethernet switches in this GPU cluster configuration and if yes, can you please confirm the technical requirements for the second Ethernet switch?	Date 08.06.2012. As correctly noted, in the Technical Specifications, per items 1.14, 1.15 and 1.1.6, one Ethernet switch is requested.
Date: 29.05.2012. 4.After the reading of the tender documents, we please answer to the following request of clarification: Bull complaints about unbalanced requirements for CPU (24) and GPU (36) counts. Explanation: Bull is leading European provider of HPC (high performance computing) solutions with the largest number of references. Based on its extensive experience with computing servers, Bull is convinced GPU clusters should be designed in a balanced way for uniform access to both CPUs and GPUs. By requesting unequal number of CPUs and GPUs this golden rule is broken. Hence, request is to change the number of CPUs and GPUs to be equal.	Date 08.06.2012. For planned HPC applications more than 1 GPU per CPU ratio is expected and therefore specification of different number of GPUs/CPUs within the same configuration is required. Besides configurations with more GPUs per single CPU, newer GPU generations (i.e. NVIDIA Tesla Kepler) allow for single GPU to be partitioned and used by different CPUs.