Annexure-A

TECHNICAL SPECIFICATIONS FOR GPU

(Node-3 & Node-8 in the tender document)

Components	**Description**	**Technical Compliance (Yes/No)**
Processor	Dual processors with at least 48 cores on each CPU or better	
System Memory	1024 GB DDR4 or higher	
GPU	8 x GPUs (Nvidia A100) or better	
Performance	Minimum 5 Peta-FLOPS AI	
GPU Memory	320 GB (8X40) total system	
CUDA Cores	Approx. 5000 per GPU	
Tensor Cores	Approx. 600 per GPU	
Power Requirements	7KW or less with hot plug & redundant power supply	
Rack space	6U or less	
Storage	OS: 2 X 1.92 TB NVMeSSDs in RAID-1 and additional 10TB (usable) storage using NVMe SSDs	
System Network	1. Two ports of IB HDR100 or better 2. Two ports of 10 GbEor better	
GPU communications protocol	NVLink 3.0 configured on NV Switch with minimum 600GB/s bidirectional communication bandwidthor equivalent	
OS Support	Red Hat Enterprise Linux /CentOS/ Ubuntu Linux	
USB Port	2	
VGA Port (or similar for connecting displays)	1	
Ethernet (RJ45) Ports	2	
Operating Temperature Range	Normal AC temperature	
Number of Simultaneous Users (Minimum)	16	
Software Support (Directly from OEM with updates & upgrades). Support portal should be enabled for min. 3 users.	CUDA toolkit CUDA tuned Neural Network (cuDNN) Primitives TensorRT Inference Engine DeepStream SDK Video Analytics CUDA tuned BLAS	

Vinod Kumar

Digitally signed by Vinod Kumar
Date: 2020.12.24 16:48:09 IST

Partner has to help build first model on-site with limited data-set	CUDA tuned Sparse Matrix Operations (cuSPARSE) Multi-GPU Communications (NCCL), Kubernetes TensorFlow , Caffe , PyTorch, Theano, Keras, caffe2, CNTK, NVidia HPC SDK	
OEM history	The OEM should provide a proof at-least 3 unique sites in India where the quoted model is being used for Development work in the areas of Artificial Intelligence (ML/DL)	
Scalability & Cluster software	System should be scalable with multi node cluster. Software support & cluster tools to be supplied along with product.	
Warranty	(3+2) Years comprehensive onsite warranty, details mentioned into the tender document	

Special Terms & Conditions and compliance to be submitted:

- The solution given for ML/DL workload should be certified by the respective OEM vendor to act as verified, tightly coupled architecture. Public document for the same should be available. All the supporting document for the same should be submitted along with bid.
- The solution should have ready to use container for different Big-data, ML, DL stack optimized for given architecture and configured to utilize GPUs fully.
- The solution should be supported for 5 years including all spare parts, software stack, DL frameworks and contract for the same should be with OEM directly.
- During the warranty all the updates and upgrades for software should be given for free.
- The solution provided should be highly scalable and should have reference architecture available for testing.
- Proposed architecture should be tested and verified by OEM jointly and proof for the same to be submitted on OEM letter head. The testing should also prove that architecture (combination of Server/storage/network) is designed jointly to get best optimized performance, deployment to be made quickly and have minimum overheads. RCB will integrating the same in existing Network & storage. Vendor to support in integration of the same.
- Proposed OEM should have min. 3 installation with similar system for Deep learning & Machine learning in different institutes (preferably in Education institutes, IITs, IISc, NIC, CSIR/DRDO/ISRO labs, large private players working in ML/DL etc.) with min. of 8 GPUs per node.
- SI should have have Engineer certified on Deep learning (Profile of Engr. to be attached). SI must support in initial project once annotated data is available with institute in choosing the right model and train the model using popular opensource frameworks for a period of

1 year on-site. The topics for training should include the usage of GPU libraries/applications such as CUDA toolkit, CUDA tuned Neural Network (cuDNN), Primitives TensorRT Inference Engine, DeepStream SDK Video Analytics CUDA tuned BLAS, CUDA tuned Sparse Matrix Operations (cuSPARSE) Multi-GPU Communications (NCCL), Kubernetes TensorFlow ,Caffe , PyTorch, Theano, Keras, caffe2, CNTK etc.

- SI must provide 5 days training on system administration, Deep learning & Machine learning, Frameworks, Practical's with few popular modules & Inferencing. This training mainly for the naive users of CUDA and should include one day for System Administrators.
- GPU nodes should also be connected to the proposed HPC facility with necessary Infiniband and/or network connectivity. The users should be able to fire jobs on GPUs through the HPC Job scheduler. GPU specific queues should be created during the commissioning of the HPC facility.