INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

MATERIALS MANAGEMENT DIVISION

Powai, Mumbai - 400076

PR No.1000017360 Rfx No. 6100000922

Technical Specification for
GPU Server - High Performance Servers (Qty 6 nos)

SL. No.	Technical Specification
1	**Processors (Qty 1 no)**
	Dual ROME AMD processor with total of 128 CPU cores with minimum 2.25Ghz
2	**Number and type of GPU (Qty 8 nos)**
2.1	8 x Nvidia A100 GPU with 80GB GPU/V-RAM per GPU (total of 640 GB)
3	**Performance**
3.1	160TF Double precision Performance,
3.2	5 PetaFlops AI performance at single precision floating point
3.3	10 PetaOPS INT8
4	**Multi Instance GPU**
4.1	Single GPU can be partitioned into as many as 7 GPU instances
5	**Internal switches and GPU-GPU communication**
5.1	Min 6 internal NV-Switches for GPU connectivity;
5.2	Minimum NVLink 3.0/ configured or NV Switch with minimum 600GB/s bidirectional communication bandwidth
6	**System Memory (Qty 1 no)**
6.1	Minimum 2TB DDR4
7	**CUDA Cores**
7.1	Minimum 5000 or above, per GPU
8	**Tensor Cores**
8.1	Minimum 600 or above per GPU
9	**Network**
9.1	Minimum 8 x Single port Mellanox IB HDR Ports (200Gbps) **(Qty 8 nos)**
9.2	Minimum 2 x Dual port Mellanox ConnectX (10/25/50/100/200Gb/sec Ethernet for storage connectivity **(Qty 2 nos)**
9.3	Should support GPU direct storage technology (Direct GPU to Storage access)
10	**Internal Storage**
10.1	OS - Minimum 2 X 1.92 TB NVMe RAID 1 **(Qty 2 nos)**
10.2	Internal storage - Minimum 8 x 3.84 TB NVMe **(Qty 8 nos)**
11	**Power requirements**
11.1	6.5 KW or less; hot plug & redundant power supply

12	**Rack space**
12.1	10U or less
13	**System Network (IPMI)**
13.1	1Gbps network
14	**OS Support**
14.1	Red Hat Enterprise Linux /CentOS/ Ubuntu Linux.
14.2	Quoted OS should be under Enterprise support from OEM.
15	**AI & HPC Software Containers Required DL SDKs**
15.1	Nvidia NGC (Nvidia GPU Cloud) containers with Nvidia NGC support for 3 years for each system.
15.2	Proposed system should be NGC certified system.
15.3	CUDA toolkit,
15.4	CUDA tuned Neural Network (cuDNN) Primitives
15.5	TensorRT Inference Engine
15.6	Deep Stream SDK Video Analytics
15.7	CUDA tuned BLAS
15.8	CUDA tuned Sparse Matrix Operations (cuSPARSE)
15.9	Multi-GPU Communications (NCCL)
16	**Scalability & Cluster software**
16.1	System should be scalable with multi node cluster.
16.2	Software support & cluster tools to be supplied along with product.
17	**Warranty & Support**
17.1	3 Years warranty, next business day. Training should be provided at the site on system configuration, running benchmarks etc.
18	**Qualifying Credential**
	Following ML-Commons (ML-PERF v1.0) Training Benchmarks must be met for each server with 8x GPUs: 1) On ImageNet (ILSVRC2012) 75.90% classification test accuracy with ResNet-50 v1.5 should be achieved within 35 minutes while training on 8 GPUs, 2) on COCO dataset 0.377 Box min AP and 0.339 Mask min AP with Mask R-CNN should be achieved within 55 minutes while training on 8 GPUs, 3) on Wikipedia 2020/01/01 0.72 Mask-LM accuracy using BERT-large should be achieved within 28 minutes, and 4) on Go dataset 50% win rate vs. checkpoint using Mini Go model (based on Alpha Go paper) should be achieved within 300 minutes.
19	**Manufactures Authorization format**
19.1	Bidders should submit authorization form from GPU supplier