INDIAN INSTITUTE OF TECHNOLOGY BOMBAY
MATERIALS MANAGEMENT DIVISION
Powai, Mumbai - 400076

PR No.1000017360 Rfx No. 6100001028

Technical Specification for
GPU Server - High Performance Servers (Qty 9 nos)

1. Processors
    - Dual ROME AMD processor with total of 128 CPU cores with minimum 2.25Ghz

2. Number and type of GPU
    - 8 x Nvidia A100 [or equivalent] GPU with 80GB GPU/V-RAM per GPU (total of 640 GB)

3. Performance
    - 160TF Double precision Performance,
    - 5 PetaFlops AI performance at single precision floating point
    - 10 PetaOPS INT8

4. Multi Instance GPU
    - Single GPU can be partitioned into as many as 7 GPU instances

5. Internal switches and GPU-GPU communication
    - Min 6 internal NV-Switches for GPU connectivity;
    - Minimum NVLink 3.0/ configured or NV Switch with minimum 600GB/s bidirectional communication bandwidth

6. System Memory
    - Minimum 2TB DDR4,

7. CUDA Cores
    - Minimum 5000 or above, per GPU

8. Tensor Cores
    - Minimum 600 or above per GPU

9. Network
    - Minimum 8 x Single port Mellanox IB HDR Ports (200Gbps)
    - Minimum 2 x Dual port Mellanox ConnectX (10/25/50/100/200Gb/sec Ethernet for storage connectivity
    - Should support GPU direct storage technology (Direct GPU to Storage access)

10. Internal Storage
    - OS - Minimum 2 X 1.92 TB NVMe RAID 1
    - Internal storage - Minimum 8 x 3.84 TB NVMe

11. Power requirements
    - 6.5 KW or less; hot plug & redundant power supply

12. Rack space
    - 10U or less

13. System Network (IPMI)
    - 1Gbps network

14. OS Support
    - Red Hat Enterprise Linux /CentOS/ Ubuntu Linux.
      Quoted OS should be under Enterprise support from OEM.

15. AI & HPC Software Containers Required DL SDKs
    - Nvidia NGC (Nvidia GPU Cloud) [or equivalent]containers with Nvidia NGC support for 3 years for each system.
    - Proposed system should be NGC certified system.
    - 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)

16. Scalability & Cluster software
    - System should be scalable with multi node cluster.
      Software support & cluster tools to be supplied along with product.

17. Warranty & Support
    - 3 Years warranty, next business day.
      Training should be provided at the site on system configuration, running benchmarks etc.

18. Qualifying Credential
    - The Quoted hardware must be listed under ML-Perf v1.0 or v1.1 for the below use cases for training on 8 identical nodes inparallel (8x8=64 GPUs):

1) Image Classification training on ImageNet using ResNet 50 v1.5
should yield 75.90% test accuracy within 5 minutes,

2) On COCO dataset 23% mAP using SSD should be achieved within 2 minutes

3) NLP, Wikipedia, BERT 4 minutes or less.

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
    - Bidders should submit authorization form from GPU OEM