CORRIGENDUM-2

Tender Name – Cloud Services for Deep Learning

Tender Reference Number - CS/MITES/054/2022/CLOUDSERVICE

Corrigendum details : Extension of Bid Submission Date and change in Annexure ‘A’ - Scope of Work.

EXTENSION OF BID SUBMISSION DATE:

The due date for the submission of bids has been extended to 12/12/2022 @ 3 PM. The bid opening is 13/12/2022 @ 3 PM.

Annexure ‘A’ - Scope of the Work to be considered only as per below mentioned requirements.

Corrigendum to Existing Technical Specification – Annexure 'A'

Cloud Services for Deep Learning

Scope of work

The bidder, is expected to meet the following compute and storage requirements and propose the right cloud services configuration to support this requirement.

Compute Requirements for Software Infrastructure

(Preferable to be located in India region for low-latency)

Compute Nodes

VM spec	Required instances
8 vCPUs, 32GB RAM, 160GB SSD	3 (optionally, 1 machine is expected to be a compute-optimized VM)
8 vCPUs, 16GB RAM, 256GB SSD	4
4 vCPUs, 16GB RAM, 100GB SSD	3
2 vCPUs, 8GB RAM, 64GB SSD	4
2 vCPUs, 4GB RAM, 64GB SSD	1

Managed DB Services

- 2 flexible (scalable) servers, for PostgreSQL
- 4 flexible (scalable) servers, for MongoDB

Compute Requirements for Training Infrastructure

- 5 DGX machines, with 8 A100 GPUs per box
    - With Slurm setup pre-configured
    - Configuration: Dual AMD EPYC 7742 (128 Phy Cores @2.25 GHz (base), 3.4 GHz (max boost), 1024 GB RAM, OS: 2x 1.92TB M.2 NVMe drives, Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives, 8x A100 GPUs, 1 IP Address
- A login node common to all the nodes

Compute Requirements for Deployment Infrastructure

(Preferable to be located in India region for low-latency)

- A managed service for Triton inference
    - Similar to Azure ML Studio, GCP Vertex AI
    - Current requirements:

- Scalable up to a maximum of 8 GPUs
- GPU Type: Nvidia 8 A30 (or V100 or equivalent)
    - Cloud GPU Server: Configuration:64 vCPUs, 360 GB RAM, 2560 GB SSD storage, 4 x NVIDIA A30-24 GB GPU RAM card, 1 IP Address

Bandwidth Requirements (across all environments)

- A maximum of 10TB ingress across all nodes, per month
- A maximum of 1TB egress across all nodes per month

Cloud Storage Requirements (for training environment)

- A maximum of 30TB data storage to begin with
- Expected to increase @ 2TB per month

Other Key Requirements

  - Data loss protection features are enabled as part of cloud infrastructure services to be provided by the bidder. This should be factored in as part of backup strategy for business continuity proposed by the bidder.
  - Provide necessary support / assistance to manage datasets available on existing cloud provider that AI4Bharat is currently using.
- The work approach would comprise of
  - Provisioning of applicable cloud services for processing deep learning workloads as per specifications outlined
  - Provide support services to ensure high reliability, availability and performance for users leveraging the cloud services to process their deep learning workloads.
  - Provide support / assistance required for data migration from existing cloud provider to the bidder's Cloud partner/ecosystem.

Key Criteria

- High availability and reliability and secure cloud services.
- Ensuring optimal performance for Deep learning (NLP) Processing workloads.
- Resolution of issues as per service level agreements.
- The unit cost per hour for the compute is optimal and is best in class among the bidders for the Cloud Infrastructure provisioning
- The unit cost per size (GB) for storage services is optimal and is best in class among the bidders.
- The bundled cost of services offered by the bidder is optimal and meets the specific scope requirements of this tender.

Delivery Schedule

- The provision of GPU services on the Cloud would be completed within a fortnight of issue of Purchase order to Bidder
- Learning support would be provided to users of these cloud services as part of onboarding to this cloud infrastructure in first month.

- The service level agreements would be defined and agreed with the selected Bidder in first month of engagement.
- This engagement is for a period of one year starting Dec 2022/Jan 2023 to Dec 2023 / Jan 2024 with a provision for extension by one more year subject to their performance during the first year meeting the above-mentioned key criteria.

Pricing

- The Saas based monthly Pricing will be provided by the Bidder separately for
    - Compute: CPU and GPUs
    - Storage
    - Bandwidth
      The pricing is expected to be inclusive of services to be provided by the Bidder.
    - To provide the unit cost (per hour) for compute for the infrastructure to be configured and provided by the bidder.
    - To provide the unit cost (Per GB) for specific storage services listed in scope section of this document.
- This pricing is expected to be valid for at least 2 years. The agreement with selected bidder could be extended at end of one year subject to their performance during the year.
- The Bidder would provide an option to extend this agreement for at least 1 year (2023-24) completion of agreement at the same price. This provided IIT Madras provides in writing request at least 2 months prior to the end date of agreement to be signed with the selected Bidder in Dec 2022/Jan 2023.
- The payment will be monthly and selected bidder would submit a monthly invoice detailing
    - Usage of Infrastructure as per scope defined
    - Usage of services.
- The payment terms will be t+30, where it is the date on which invoice submitted is accepted by IIT Madras.
- A simple governance mechanism would be in place to ensure to address and resolve the following effectively and with agility.
    - Issues in the Delivery of services and/or
    - Issues in meeting Service level agreements

Note: The Bidder would compensate for not adhering to any of key criteria in any month during which an active agreement exists with the Bidder. This would be at the minimum, applicable discounts in their pricing in the invoice that specific month.