Revised RFP

Bid for GPU Compute for uses AI / ML at GSDC

1. **Eligibility Conditions:**

Sr. No.	Specific Requirement	Documents required
1.	The bidder should be a company registered in India under the Companies Act 1956, Act 2013 or a partnership registered under the India Partnership Act 1932, or a Partnership firm registered under the Limited Liability Partnership Act 2008 with their registered office in India in operation for the last three years	• Certificate of Incorporation • Memorandum and Article of association • Registered Partnership Deed • Copy of PAN card • Copies of relevant GST registration certificates.
2.	The bidder should have average Minimum Annual turnover of Rs. 50 crores in 3 years out of last 5 financial years from the last date of bid submission with positive net worth.	• Audited profit and loss statement and balance sheet • Auditor certificate clearly specifying the turnover and positive net worth.
3	The OEM should have average Annual Turnover of minimum Rs. 250 crores for the last five financial years from the last date of bid submission with positive net worth.	• Audited profit and loss statement and balance sheet • Auditor certificate clearly specifying the turnover and positive net worth.
3.	The Bidder Should have technical support center in Ahmedabad / Gandhinagar, Gujarat. If the bidder is not having any technical support center in Ahmedabad / Gandhinagar, Gujarat, then bidder should submit a letter of undertaking to open the office in Gujarat within 30 days from the date of issue of work order if (s) he is awarded the work	The Bidder should submit valid Proof (or) Bidder should submit Self-declaration duly Signed and stamped by the authorized Signatory in format described in RFP.
4.	The bidder should have experience of set up of similar GPU base solution with GPU or CPU with below criteria in last 5 Years in India as on last date of bid submission: • One project having similar works total value of INR 40 Cr (Should have min 45 GPU / 800 CPU) or • Two project having similar works total value of INR 25 Cr (Should have min 28 GPU / 500 CPU) or • Three project having similar works total value of INR 20 Cr (Should have min 22 GPU / 400 CPU) Note: Similar works means SITC OF GPU ACCELERATED with multiple GPU Node.	Copy of Work Order along with Completion / Go-Live certificate
4.1	The bidder should have experience of set up of similar GPU base solution with Cumulative 15 nos. of GPU based solution in last 5 Years in India.	Copy of Work Order along with Completion / Go-Live certificate.

5.	The OEM should have executed similar GPU setup for min 3 clients in last 5 Years in India as on date of bid submission. Out of which One client deployment should be One project having similar works total value of INR 125 Cr. Note: Similar works means SITC OF GPU ACCELERATED with multiple GPU Node.	Copy of Work Order along with Completion / Go-Live certificate
6.	The bidder should provide the authorization certificate from the OEM for a. Quoting the requirement and subsequent support for Hardware and Software (and) b. Proposed GPUs solution will not be End of Life (EOL) for 5 years from the date of installation	In Case of SI, should submit Manufacturer Authorization Form. In Case of OEM, Letter of Declaration on their letter head
7.	Neither OEM nor bidder should be blacklisted from supplying equipment to any Government/PSU/BFSI within India in the past.	Certificate of Undertaking for Non-blacklisting from supplying equipment to any Government/PSU/BFSI within India in the past.
8.	A Power of Attorney / Board Resolution in the name of the person signing the bid document.	Original Power of Attorney / Board Resolution Copy on a non-judicial stamp paper.

1. All details and the supportive documents for the above should be uploaded in the GeM bid.
2. Bidder's experience, bidder's turn over criteria will not be considered of GeM bid. However, bidder must match eligibility criteria, experience, bidder's turn over criteria, etc. as mentioned above (& in this document) and will be considered for evaluation. EMD and PBG should submitted by bidder as per GEM.

2. Clarification on Bidding Documents

- A prospective Bidder requiring any clarification of the bidding documents may seek clarifications by submitting queries on email Id: **mgrhninfra1-gil@gujarat.gov.in, dgmTech-gil@gujarat.gov.in** prior to the date of Pre-Bid Meeting.

- Tenderer will discuss the queries received from the interested bidders in the Pre-Bid Meeting and respond the clarifications by uploading on the website https://gil.gujarat.gov.in.

- No further or new clarification what so ever shall be entertained after the Pre-Bid Meeting.

- The interested bidder should send the queries as per the following format:

Bidder's Request For Clarification			
Name of Organization submitting Request		**Name & position of person submitting request:**	**Address of organization including phone, fax, email points of contact**
Sr. No.	**Bidding Document Reference (Clause /page)**	**Content of RFP requiring Clarification**	**Points of Clarification required**
1			
2			

3			

3. Scope of Work:

1. GPU Servers shall be supplied, installed, configured, tested and commissioned along with necessary software's, OS's and license's at GSDC located at Gandhinagar, Gujarat.
2. Bidder has to deploy propose solution for inference and AI training model.
3. All software and library licenses to be provided in the name of DST/ DIT, Government of Gujarat.
4. Gujarat GPU Compute for AI / ML solution must have rack mounted computing platform-based computer servers, either as rack or blade server design housed in its suitable chassis.
5. The bidder shall submit the detailed documentation on the implementation and deployment.
6. The solution should support remote console access to all the servers for cluster server's health monitoring at Fast Ethernet or better access speed.
7. The servers/chassis/enclosures should be populated fully with N+1 redundant power supplies of the suitable capacity rating available for the proposed model with the supplier. Failure of one of the Power supplies should not throttle the Compute nodes. In case the offered Power Supplies cannot take the HPL load of all the Compute Nodes in the chassis, lower number of Compute Nodes per chassis may be proposed.
8. The bidder will have to supply Server Rack along with provision of iPDU, TOR Switch, patch panel, cables, SFP modules, any other active/passive components etc. to host the HPC Cluster with GPUs at GSDC. Any other component required for the solution proposed by the supplier has to be incorporated for completion of the Solution.
9. Onsite comprehensive annual maintenance with warranty and OEM support for 5 Years from the date of completion of User Acceptance Test (Onsite warranty will include those sites to which the item supplied under the contract is moved, in case of migration of the equipment). Warranty should include but not limited to - On-going Firmware updates, Proactive bug fixes, Preventive Maintenance, Parts replacement, etc.
10. After completing the installation and integration, the bidder will demonstrate the compliance of the RFP and provide required training to the GSDC /TPA for executing UAT and further Operation.
11. All the items as required under this RFP should be delivered in a single lot.
12. If, during the warranty period, any system as a whole or any subsystem has any failure on two or more occasions in a period of 3 months, it shall be replaced by equivalent new equipment by the bidder at no cost. All defective items should be replaced / repaired within the SLA period.
13. The bidder shall be fully responsible for the manufacturer's warranty for all equipment, accessories, spare parts etc. against any defects arising from design, material, manufacturing, workmanship, or any act or omission of the manufacturer / bidder or any defect that may develop under normal use of supplied equipment during the warranty period.
14. The bidder shall replace the faulty hard disk at no cost, the department will not be returned the faulty disk after replacement of new disk.
15. The bidder should provide entire support of the required solution asked in the RFP and back-to-back support from the OEM.

The bidder should provide Support/ Escalation Matrix & Portal details for logging tickets for any failure/performance incidents. Also there has to be mechanism wherein all licenses to be showcased on the portal.

Manpower for Hand Holding Support

- Successful bidder will have to depute **2 (two)** technical manpower as below to provide hand holding support for the contract period.

1. **System Administrator**

   - Total 5+ years of experience
   - Proficiency in Linux (e.g., Ubuntu, CentOS, RHEL)
   - Familiarity with cluster management tools like SLURM, Kubernetes
   - Understanding of high-speed interconnects - InfiniBand, Ethernet
   - Experienced in configuring network topologies for low-latency, high-throughput AI workloads.
   - Knowledge of GPUs (e.g., NVIDIA A100, H100), accelerators, and their deployment
   - Awareness of storage technologies and their AI workload implications (e.g., NVMe, SSDs, and parallel storage).
   - Experience with configuration management tools like Ansible

2. **AI / ML Deployment engineer**

   - Proficiency in Python and corresponding AI libraries – NumPy, SciKit, Pandas, CUDA Python, CuGraph, CuML etc
   - Experience with containerization (Docker) and orchestration (Kubernetes) tools
   - Hands-on experience with AI training frameworks like TensorFlow, PyTorch and deployment frameworks like Triton
   - Familiarity with deploying and scaling large language models (e.g., pre-training, fine-tuning, serving & inference pipelines).
   - Proficiency in data preprocessing, feature engineering, and handling large-scale datasets
   - Experience implementing MLOps pipelines for automating model lifecycle management
   - Experience with cloud services and APIs

- The deputed manpower will have to remain present during normal office hours of GSDC (9 AM to 7 PM) during working days and support GSDC for day-to-day maintenance and handling effective GPU infrastructure utilization.
- If require, the manpower will have to remain present on holyday(s) or after office hours based on the requirements of GSDC.

- The bidder shall have to provide backup resources in case of the deputed manpower is absent or on leave. The backup resource deputed shall be aware of the tasks and responsibility being carried out during that period at GSDC and should be able to execute the tasks with minimum on-call support.
- The manpower will have to report to GSDC authority. The bidder shall submit proof of attendance certified by the GSDC authority along with the Invoice for payment process.

4. <u>Warranty Support: As part of the warranty services bidder shall provide:</u>

I. Bidder shall provide a comprehensive on-site free warranty for 5 years from the date of acceptance of FAT (Final Acceptance Test) for proposed solution.

II. Bidder shall also obtain the 5 years OEM support (ATS/AMC) on all hardware and other equipment for providing OEM support during the warranty period.

III. Bidder shall provide the comprehensive manufacturer's warranty and support in respect of proper design, quality and workmanship of all hardware, equipment, Software, Licenses, accessories etc. covered by the bid. Bidder must warrant all hardware, equipment, accessories, spare parts, software etc. procured and implemented as per this bid against any manufacturing defects during the warranty period.

IV. Bidder shall provide the performance warranty in respect of performance of the installed hardware and software to meet the performance requirements and service levels in the bid.

V. Bidder is responsible for sizing and procuring the necessary hardware and software licenses as per the performance requirements provided in the bid. During the warranty period bidder, shall replace or augment or procure higher-level new equipment or additional licenses at no additional cost in case the procured hardware or software is not adequate to meet the service levels.

VI. Mean Time between Failures (MTBF): If during contract period, any equipment has a hardware failure on four or more occasions in a period of less than three months, it shall be replaced by equivalent or higher-level new equipment by the bidder at no cost. For any delay in making available the replacement and repaired equipment's for inspection, delivery of equipment's or for commissioning of the systems or for acceptance tests / checks on per site basis, DST/GIL/DIT reserves the right to charge a penalty.

VII. During the warranty period bidder, shall maintain the systems and repair / replace at the installed site, at no charge, all defective components that are brought to the bidder notice.

VIII. The bidder shall as far as possible repair/ replace the equipment at site.

IX. Warranty should not become void, if DST/GIL/DIT buys, any other supplemental hardware from a third party and installs it within these machines under intimation to the bidder. However, the warranty will not apply to such supplemental hardware items installed.

X. The bidder shall carry out quarterly Preventive Maintenance (PM), including cleaning of interior and exterior, of all hardware, if any, and should maintain proper records at each site for such PM. Failure to carry out such PM will be a breach of warranty and the warranty period will be extended by the period of delay in PM.

XI. Bidder shall monitor warranties to check adherence to preventive and repair maintenance terms and conditions.

XII. Bidder shall ensure that the warranty complies with the agreed Technical Standards, Security Requirements, Operating Procedures, and Recovery Procedures.

XIII. Bidder shall have to stock and provide adequate onsite and offsite spare parts and spare component to ensure that the uptime commitment as per SLA is met.

XIV. Any component that is reported to be down on a given date should be either fully repaired

or replaced by temporary substitute (of equivalent configuration) within the time frame indicated in the Service Level Agreement (SLA).

XV. Bidder shall develop and maintain an inventory database to include the registered hardware warranties.

XVI. To provide warranty support effectively, OEM should have spare depo in India and will be ask to deliver spare as per SLA requirement.

1. All supplied items must conform to the detailed technical specifications as mentioned in this document.
2. Install the equipment, obtain user acceptance and submit a copy of user acceptance to designated authority.
3. The agreement stipulates that the vendor shall maintain the system with uptime. It is required to maintain uptime of 99.741%. Further, bidder is responsible for providing comprehensive warranty and support (24x7) for the period of 5 years from the date of successful completion FAT.
4. The Bidder shall be responsible for providing all material, equipment and services specified or otherwise, which are required to fulfill the intent of ensuring operability, maintainability and the reliability of the complete work covered under this specification.
5. Manufacturer shall provide and support for installation, commissioning, spares, technical support in Gujarat.
6. All supporting equipment, tools shall be arranged by vendor himself.
7. Unpacking of goods shall be done in front of GIL/GSDC officer, Gandhinagar official and for any damage it is sole responsibility of vendor.
8. Delivery of goods: packing unpacking transportation loading unloading Octroi insurance and any other taxes and duties (excluding GST) shall be included in the bid price.
9. All the liabilities like human injury, incident, etc. pertain in the bidder scope. The bidder will be solely responsible to execute insurance for the said work as mentioned in this RFP.
10. All safety precaution should be taken as per Industrial practice by the bidder to take upmost care. In any case, the tenderer will not be liable to any obligation for any issue arise under this project.

5. <u>Lack of Information to Bidder:</u>

- The Bidder shall be deemed to have carefully examined all RFP documents to its entire satisfaction. Any lack of information shall not in any way relieve the Bidder of its responsibility to fulfil its obligation under the Contract.

6. <u>Payment Terms:</u>

1. 70% of the Capex cost shall be paid within 30 days after delivery duly certified by the GSDC and counter-signed/approved by the authority.
2. 20% of the Capex cost shall be paid against installation/configuration duly certified by the GSDC and counter-signed/approved by the authority.
3. Balance 10% of the Capex cost shall be paid after acceptance duly certified by the GSDC and counter-signed/approved by the authority.
4. Cost of **Comprehensive Annual Maintenance with warranty and OEM support** for 5 years will be equally distributed in 20 quarters and paid on Quarterly basis after FAT.

Note: Bidder has to submit invoices along with necessary legitimate supporting documents failing which invoices submitted are liable to be rejected/not accepted.

7. Final ACCEPTANCE TEST:

To be carried out based on followings but not limited to:

- GIL and GSDC reserves the right to inspect goods and services supplied as per the scope of this RFP document. The cost of all such tests shall be borne by the Vendor. Any inspected goods fail for confirm to the specification will be rejected, and Vendor shall have to replace the rejected goods as per the contract specification without any financial implication to the GIL/DIT.
- After successful installation of the System in accordance with the requirements as mentioned in Schedule of Requirement, Final Acceptance Test will be conducted. After successful testing, Acceptance Test Certificate will be issued by GIL/DIT and member of GSDC or its designated agency to the Bidder. The Bidder shall submit the certificate to GIL/DIT for further payment process.
- The date on which Final Acceptance certificate is issued shall be deemed to be the date of successful commissioning of the System.
- Any delay by the successful bidder in the Acceptance Testing shall render the successful bidder liable to the imposition of appropriate Penalties.
- Bidder is required to update the details of Hardware installed in the Assets Master or decided by GIL and member of GSDC Officer before completion of FAT.
- GIL/GSDC and/or an outside agency nominated by DST will conduct an acceptance test on the hardware after completion of installation and commissioning of hardware by the vendor. Acceptance test shall comprise of tests to verify conformity of technical requirements/specifications and performance. In case GIL/GSDC is not satisfied with the above then, the vendor will upgrade /replace them with equal or higher model after due approval of GSDC team without any extra cost. The exact details of acceptance test will be mutually decided after the installation of hardware.

8. IMPLEMENTATION TIMELINES & PENALTIES:

Successful bidder has to complete the Installation, Configure, Commissioning, Integration with Acceptance of the ordered work within the time period (s) specified in the below table. However, in case of any delay solely on the part of successful bidder TENDERER reserve the right to levy the appropriate penalties as per the below table:

IMPLEMENTATION TIMELINES & PENALTIES FOR PROPOSED HPC Cluster with GPUs AT GSDC					
S/n	**Work type**	**Time Limit for Execution**	**Penalty for Delay**	**Maximum Penalty**	**Overall Penalty Cap**
1	Submission of PBG	Within 15 Days from date of issuance of GEM contract	EMD may be forfeited and contract may be terminated or part thereof	-	
2	Supply of the Hardware including Licenses and OEM	T1=T+45 days from the date of issuance	0.5% of order value of delayed/pending work per week or	10% of GEM order value	Overall (Sr. no- 2 to 6) Penalty CAP not be

	Warranty Certificate.	of contract over GEM	part thereof		more than 10 % of the total GEM order value for IMPLEMENTATION TIMELINES & PENALTIES:
3	Installation, commissioning & integration of GPU servers at GSDC along with HLD , LLD documents	T2=T1+30	0.5% of order value of delayed/pending work per week or part thereof	10% of GEM order value	
4	Deployment of required Skilled Resource at GSDC	T2+7 Days	Rs. 10000/- day.	Rs. 250000/-	
5	Final Acceptance Testing (FAT) & User Acceptance Test (UAT)	T3=T2+15 days	0.5% of order value of delayed/pending work per week or part thereof	10% of GEM order value	
6	Training	10 Days from T3	Rs. 10000/- day.	Rs. 250000/-	

Note:

- Material supplied, installed and commission as per this Bid/contract should be covered under the warranty for a period of five years from the date of FAT acceptance.
- T= Date of issuance of contract over GEM.
- In case of any fault arises in the installed items during the warranty period of 5 years, bidder is requiring to either repair the faulty items or have to install the replacement (complying to the RFP specification) for faulty material without any additional cost to the Tenderer.
- Aforesaid penalty cap will not be applicable for any severe impact/incident/outage at GSDC, resulting in loss to Government of Gujarat.

9. SLA & Penalties

a. Operational Penalty:

- The successful bidder shall repair/ replace all faulty material covered under the warranty within the shortest possible time thus ensuring minimum downtime, failing which applicable penalty will be imposed. In case of failure of appliance / solution for more than 3 consecutive time for the same issue within any of the single quarter during contract period, bidder would be bound to replace the product with no cost to DST / GIL/DIT.
- The successful bidder shall be responsible for maintaining the desired performance and availability of the system/services.
- Successful bidder should ensure the prompt service support during warranty period.
- Timeline for resolution is within 4 hours from the time of call logged / reported to Bidder/OEM. If the successful bidder fails to resolve the call as specified above, penalty will be imposed on each delayed day for Rs 5000 / hour or part thereof, which will be recovered against Performance bank guarantee submitted by the successful bidder on completion of warranty period.
- Down time will be calculated from the time complain is logged to service in charge of Successful Bidder (via email/call/written letter) till the GSDC's authorized / Nominated employee acknowledge the repair / service completion.

b. SLA for Uptime (99.741%)

SLA	Target	Penalties in case of breach in SLA

Uptime of solution	>=99.741%	No penalty
Uptime of solution	<=99.741%	In case of failure of proposed solution and non-maintaining targeted value, 0.5% of yearly CAMC payment for every hourly. delay in resolution; with max cap of 10 % of total 5 years CAMC value; it will be adjusted from Yearly CAMC payment .

- SLA will be calculating on monthly/quarterly basis, However, Final penalty deduction on the yearly payment of CAMC value i.e., (4* 3 quarter or 12*1 monthly SLA report penalty will be applied during CAMC yearly payment.)
- Bidder has to ensure support 365*24*7 for SLA calculation.

c. <u>Manpower related SLA and Penalties:</u>

1. Availability of the min required manpower should be 100%. The agency has to implement the attendance system and share the attendance report of each person deployed as part of team on monthly basis with the GSDC.

2. Replacement of a profile by the agency (only one replacement per technical profile – with equal or higher qualification and experience – would be permitted per year)

3. Prior Intimated Leave of absence will be allowed: If a resource proceeding on leave or becoming absent is replaced with a resource approved by authority, then such substitution will not be treated as absence.

For every SLA non-compliance reported and proved, there shall be a penalty as given below:

#	SLA	Timelines/ Event	Applicable Penalty
2	Replacement of resources by the agency on formal submission of resignation by the resource in the company.	There should be minimum 15 days overlap between the new deployed resource and the replaced resource.	No penalty- On timely replacement. Rs. 5000/- per resource per day for each day delay from stated timelines.
3	The deployed resources shall not be engaged in any activity other than that assigned by the TENDERER	-	Penalty of Rs. 50,000 per resource may be imposed on breach of SLA. On consecutive breach of 03 times may lead to termination of the contract.
4	Absence without prior approval from the TENDERER and No Backup resource arranged	-	Penalty of Rs. 5000/- per resource per day shall be imposed.

10. Minimum Technical Specification:

Master Node: (07 Nodes)

Components	Minimum Specifications
Processors	Latest Generation Intel® Xeon® platinum or AMD Epyc scalable processors with Minimum Dual 56-Core
Mother Board	OEM Supported Motherboard and chipset.
System Memory	1 TB DDR4 or higher SDRAM with ECC Advance.
Internal Storage	For Operating System: Minimum 2*3.84 TB capacity hot swap Enterprise NVMe SSDs Minimum 4*3.84 TB capacity interface hot swap Enterprise NVMe SSDs.
Storage Controller	Hardware RAID 0,1, 5, 6, 10, 50, 60 with 4GB cache Flash based cache protection module should be included, should support Gen 5.0 PCIe NVMe
Network	Following N/W are required: a) Infiniband / Ethernet (100Gb/s or higher) as required for quoted storage delivery to nodes b) Ethernet (100Gb/s or higher) for User delivery c) Ethernet (10GbE or higher) for cluster orchestration d) Ethernet (10GbE or higher) for perimeter connectivity e) Ethernet (1GbE or higher) for in-band management
External Port	One VGA port, 2 or more USB ports. Dedicated LAN port for Management Interface
Server Management	Dedicated IPMI 2.0 compliant management LAN port having support for system health monitoring, event log access, Virtual media over network, and Virtual KVM (KVM over IP). All required licenses to use IPMI features should be included. Licenses shall be perpetual to use.
Power Supply	Appropriate energy efficient redundant (N+N) hot swappable 80 Plus Titanium power supply and FAN
Failure Alerting Mechanism	Should be able to alert upcoming failures on maximum number of components such as Processor, memory, HDDs and expansion cards, etc.
OS Support	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux server Quoted OS should be under Enterprise support from OEM.
Hypervisor	Hypervisor with Enterprise level highest license and support should be provided from day one.
Software Support	All necessary and required software,SDK,libraries,tools to cater and run the AI/ML workload should be provide from day 1.
Scalability, Cluster and Management Hardware and software	System should be scalable with multi node cluster. Software support & cluster tools and management hardware and software and licenses to be supplied along with product. Compatible with Full-stack reference designs of the Storage providers.
Warranty & Support	5 Years comprehensive onsite warranty.

Security Features	System should support Secure Firmware Updates, Support for Trusted Platform Module enabled within the BIOS for secure cryptographic key generation, Secure storage space and Self encrypting drive. ACPI 6.4 Compliant, UEFI 2.8, Support for Trusted Platform Module enabled within the BIOS for secure cryptographic key generation, SMBIOS 3.5 or later, Malicious Code Free design" (to be certified by OEM).
Performance Benchmarks	1.Specrate2017_fp_base >690 2.Specrate2017_Int_base >530 The System OEM must have listed the SPEC benchmark score on www.spec.org for the same node model with the same CPU configuration and a memory configuration of at least 1TB.

Nodes for AI training: Total Qty-4 sets.

Components	Minimum Specifications
Processors & performance (per node, minimum)	Min Dual 56-core latest Gen Intel® Xeon® platinum or AMD Epyc scalable processors, with Min 8 X GPU Accelerators. providing 500TF or Higher Double Precision Tensor FP64 / TF64 Performance, 31 PetaFlops or Higher FP8 performance with sparsity.
Number of GPUs and GPU Communication	8 x Accelerators per node, each with minimum 140 GB or higher memory per Accelerator. Minimum 900GB/s bidirectional communication bandwidth per GPU. Should support Tensor core/Matrix core, CUDA / Stream Processors/ openCL /ROCm with Accelerators,
Multi Instance GPU	Capability to support partitioning of single GPU into multiple GPU instances where both memory and compute of the GPU is divided into multiple instances
System Memory	The system should be configured with Minimum 2TB DDR5 RAM with all slots populated in balanced configuration for maximum bandwidth.
Network	a) Minimum 8 nos of InfiniBand NDR ports or Ethernet (200Gb/s or higher) for compute communication for internode communication, b) 1 nos. of port for BMC (dedicated LAN port), c) Minimum 1 no. of 1 GbE port and 2 nos of 10 GbE or higher (Fiber/Copper) port. d) 1 nos. of InfiniBand / 100G or higher Ethernet 2 x twin-port HCA as required for quoted storage delivery to node. e) Additionally 1 nos of 100GbE Ethernet (Fibre).
Internal Storage	• For Operating System: minimum 1.92 TB M.2 NVMe drives • For Data: Minimum 8 * 3.84 TB U.2 NVMe drives
Security Features	System should support Secure Firmware Updates, Support for Trusted Platform Module enabled within the BIOS for secure cryptographic key generation, Secure storage space and Self encrypting drive.
Power requirements	hot plug & redundant power supply

System Network	Following N/W are required: 1. NDR Infiniband / Ethernet (200Gb/s or higher) for compute communication 2. Infiniband/Ethernet (100Gb/s or higher) for storage delivery 3. Ethernet (Min 10 GbE or higher) for cluster orchestration 4. Ethernet (10 GbE or higher) for perimeter connectivity 5. Ethernet (1GbE) for in-band management
OS Support	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux server Quoted OS should be under Enterprise support from OEM.
AI Enterprise Software	AI Enterprise software & subscription or equivalent for each and every GPUs to be included from day one of Installation. Software stack to be supported by GPU OEM for 5 years for each system. All necessary and required software, SDK, libraries, tools to cater and run the AI/ML workload should be provided from day 1. Bidder to ensure that enterprise level OEM support & SLA is available for all OEM provided software and libraries. Licenses required must be included and shall be perpetual with no scaling restrictions. Some of the basic, SDK/library/containers to be used in the system are: a. CUDA toolkit, b. CUDA tuned Neural Network (cuDNN) Primitives c. TensorRT Inference Engine d. CUDA tuned BLAS (cuBLAS) e. CUDA tuned Sparse Matrix Operations (cuSPARSE) f. Multi-GPU Communications (NCCL) g. Industry SDKs – NVIDIA Merlin, DeepStream, ISAAC, Nemo, Morpheus h. Rapids, Tao, TensorRT, Triton Inference
Software Support	All necessary and required software, SDK, libraries, tools to cater and run the AI/ML workload should be provided from day 1. Comprehensive software frameworks for the following should be provided: a) Accelerated ML and data processing b) LLM pre-training, fine-tuning & guard railing c) Microservices enabled framework for API based LLM model deployment & serving d) End to End flows for conversational AI - ASR, NMT, TTS e) Video, Audio and Image processing pipelines In addition customizable pre-built reference workflows for generative AI use cases shall also be covered as part of the software offerings
Scalability, Cluster and Management Hardware and software	System should be scalable with multi node cluster. Software support & cluster tools and management hardware and software and licenses to be supplied along with product. Full-stack reference designs with all of the leading Storage providers.
Certification	Rack Servers should be certified by GPU Controller / Accelerator OEM, the Certificate or listing of offered Server model in GPU Controller / Accelerator OEM website must be submitted along

	with bid.
Warranty & Support	5 Years comprehensive warranty with Enterprise level Highest/Premium Support. OEM Enterprise level Highest/Premium Support should reflect on OEM portal. Quoted all products including GPUs should not be End of support till 5 years from the date of issue of the bid. The product quoted should be manufactured in current year.
Cluster Management & Scheduler and hardware	Cluster management for system provisioning and monitoring needs to be included, The Cluster Manager must support multiple of hardware vendors. The Cluster Manager must allow for the easy deployment and management of servers across multiple data centers, the public cloud, and edge locations as a single shared infrastructure through a single interface. All necessary hardware, software and necessary licenses should be provided from day 1
Benchmarks	Bidder needs to submit proof of the quoted GPU meeting these Mlcommons training benchmarks at the time of bidding :-
	Offered Nodes should be listed under ML Commons Training (4.0 or higher) for the mentioned Benchmarks, supporting published link to be shared during bid submission Specifications: a) BERT - 5.3 minutes or less on single node b) DLRM-dcnv2 – 3.6 minutes or less on single node c) GNN – 7.8 minutes or less on single node d) Llama 2 70B – 24.7 minutes or less on single node e) ResNet – 12.1 minutes or less on single node f) RetinaNet – 34.3 minutes or less on single node g) Stable Diffusion – 41.4 minutes or less on single node U-Net3D – 11.6 minutes or less on single node

Inference Node: Total Qty-12 sets.

Components	Minimum Specifications
Processors & performance (per node, minimum)	Latest Generation Intel® Xeon® platinum or AMD Epyc scalable processors with Minimum Dual 32-Core, with Min 2 X GPU Accelerator. The bidder should configured Head/ Master/ Management Node in (1+1) HA mode deliver the solution.
Number of GPUs and GPU Communication	2 x Accelerators per node, each with minimum 140 GB or higher GPU per Accelerator. Should support Tensor core/Matrix core, CUDA / Stream Processors/ openCL /ROCm with Accelerators.
Multi Instance GPU	Capability to support partitioning of single GPU into multiple GPU instances where both memory and compute of the GPU is divided into multiple instances.
System Memory	The system should be configured with Minimum 2TB DDR5 RAM with all slots populated in balanced configuration for maximum bandwidth.
Network	Minimum 2 x 100GbE Ethernet ports

	1 nos. of port for BMC (dedicated LAN port), minimum 1 no. of 1 GbE or 10 GbE (Fiber) port.
Internal Storage	For Operating System: minimum 2*1.92 TB M.2 NVMe drives Minimum 4 * 3.84 TB U.2 NVMe drives
HBA Card	16 Gbps Host Bus Adaptor for connecting with existing storage.
Security Features	System should support Secure Firmware Updates, Support for Trusted Platform Module enabled within the BIOS for secure cryptographic key generation, Secure storage space and Self encrypting drive. ACPI 6.4 Compliant, UEFI 2.8, Support for Trusted Platform Module enabled within the BIOS for secure cryptographic key generation, SMBIOS 3.5 or later, Malicious Code Free design" (to be certified by OEM).
Power requirements	2000W or more hot plug & redundant power supply 80 PLUS Titanium
PCI Express interface	4 x PCIe Gen 5.0 x 16 FH FL Slots. All slots must operate at PCI Gen 5.0 speed when fully populated
Mother Board	Appropriate Motherboard and chipset. Must support PCIe Gen 5.0 and compatible with selected processors and GPUs.
System Network	Following N/W are required: 1. Ethernet (10 GbE or higher) for cluster orchestration 2. Ethernet (10 GbE or higher) for perimeter connectivity 3. Ethernet (1GbE or higher) for in-band management 4. Infiniband / Ethernet (100Gb/s or higher) as required for quoted storage delivery to nodes
Networking Switch	1. Two Nos. of Switch with 48 *10G SFP+ and 8 x 100G QSFP ports to connect all of “Master Node”, “Node for AI Training” and “Inference Node” to form Cluster Communication and Perimeter N/W. Switch must support of MLAG /MCLAG feature. Switch must support EVPN – VxLAN based network. Required cables of appropriate length, transceivers should be supplied. Switches(s) should have redundant Power Supply. 5 Years Comprehensive Onsite Warranty. 2. One Nos. of Switch with 48 *1G RJ45, 4* 25G SFP28 and 2 x 100G QSFP ports to connect all of “Master Node”, “Node for AI Training” and “Inference Node” to form In-band and BMC/Out-of-band Management N/W. Required cables of appropriate length, transceivers should be supplied. Switch should have redundant Power Supply. 5 Years Comprehensive Onsite Warranty. 3. One Nos. of Switch with 32 * 100GbE QSFP ports or One Nos. of Switch with 64* 100GbE QSFP ports to connect all of “Master Node”, “Node for AI Training” and “Inference Node” to form User N/W. Switch must support of MLAG /MCLAG feature. Switch must support EVPN – VxLAN based network. Required cables of appropriate length, transceivers should be supplied. Switch should have redundant Power Supply. 5 Years Comprehensive onsite Warranty.
OS Support	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux server Quoted OS should be under Enterprise support from OEM. Quoted model should be certified for RHEL, Ubuntu OS. The same shall be verifiable from OS OEMs website. Supply should include DC edition unlimited Guest OS licenses

Hypervisor	Hypervisor with Enterprise level highest license and support available should be provided from day one.
Virtual GPU	Support for virtual GPU to share a physical GPU across multiple VMs. required license should be from day one. Bidder to ensure that enterprise level OEM support & SLA is available for all OEM provided software and libraries.
AI Enterprise Software	AI Enterprise software & subscription or equivalent for each and every GPUs to be included from day one. Bidder to ensure that enterprise level OEM support & SLA is available for all OEM provided software and libraries. All necessary and required software, SDK, libraries, tools to cater and run the AI/ML workload should be provide from day 1. Comprehensive software frameworks for the following should be provided: a) Accelerated ML and data processing b) Microservices enabled framework for API based LLM model deployment & serving d) End to End flows for conversational AI - ASR, NMT, TTS e) Video, Audio and Image processing pipelines In addition customizable pre-built reference workflows for generative AI use cases shall also be covered as part of the software offerings
Scalability, Cluster and Management Hardware and software	System should be scalable with multi node cluster. Software support & cluster tools and management hardware and software and licenses to be supplied along with product.
Certification	Rack Servers should be certified by GPU Controller / Accelerator OEM, the Certificate or listing of offered Server model in GPU Controller / Accelerator OEM website must be submitted along with bid.
Warranty & Support	5 Years comprehensive warranty with Enterprise level Highest/Premium Support. OEM Enterprise level Highest/Premium Support should reflect on OEM portal. Quoted all products including GPUs should not be End of support till 5 years from the date of issue of the bid. The product quoted should be manufactured in current year.
Cluster Management & Scheduler and hardware	Cluster management for system provisioning and monitoring needs to be included, The Cluster Manager must support multiple of hardware vendors. The Cluster Manager must allow for the easy deployment and management as a single shared infrastructure through a single interface. All necessary hardware, software and necessary licenses should be provided from day 1
Bidder needs to submit proof of the quoted GPUs being listed in MLperf inferencing benchmarks at the time of bidding	"Image segmentation (medical) 3D-Unet-99 Throughput for single NODE inference (99% offline) = 45 Samples/s or higher" "NLP Bert-99 Throughput for single NODE inference (99% offline) = 70000 Samples/s or higher" "Recommendation dlrm-v2-99 Throughput for single NODE inference (99% offline) = 610000 Samples/s or higher" "LLM Summarization gptj-99 Throughput for single NODE inference (99% offline) = 17000 Tokens/s or higher"

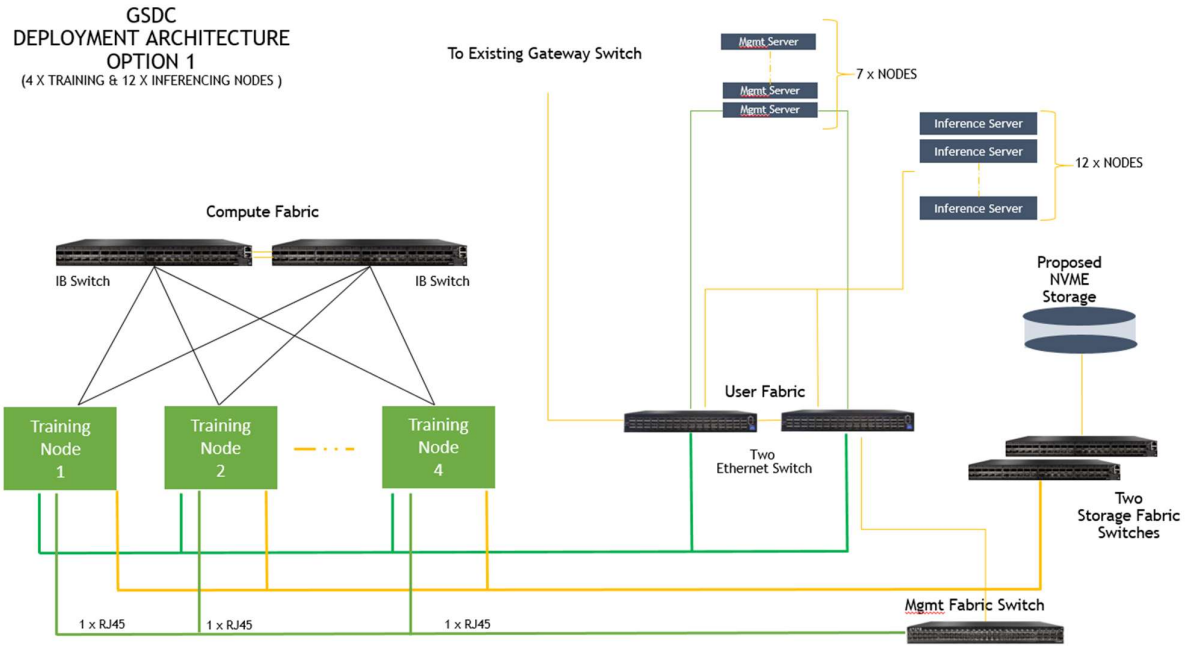
	"LLM Q&A LLama2-70B Throughput for single NODE inference (99% offline) = 27000 Tokens/s or higher" "LLM Text Generation (Math & Code) Mixtral-8x7B Throughput for single NODE inference (99% offline) = 53000 Tokens/s or higher" "Image Classification ResNet Throughput for single NODE inference (99% offline) = 716000 Samples/s or higher" "Object Detection RetinaNet Throughput for single NODE inference (99% offline) = 11400 Samples/s or higher" "Image Generation Stable Diffusion-XL Throughput for single NODE inference (99% offline) = 14 Samples/s or higher"
Performance Benchmarks	1.Specrate2017_fp_base >690 2.Specrate2017_Int_base >530 The System OEM must have listed the SPEC benchmark score on www.spec.org for the same node model with the same CPU configuration and a memory configuration of at least 1TB.

Storage Nodes	
External Storage	The solution should be NFS based and delivered with 1PB (All NVMe) usable post RAID 6/equivalent or better protection, expandable up to 2PB in the same file system.
	The proposed storage array should be configured with no single point of failure, including controllers (at least 3 controllers per disk tier), cache, power supply, cooling fans, etc. It should be scalable up to 12 additional controllers/nodes.
	Disks: 1PB (NVMe) usable post RAID 6 or better configuration The storage should be distributed with namespace consistent across nodes.
	Throughput: At least 20GBps performance for 8,00,000 IOPS, with linear scalability by adding capacity and controllers/nodes in the future. IOPS : 8,00,000 **Performance:** 20GBps Read/Write from day one and scalable up to 60GBps with a scale-out architecture and additional controllers/nodes in the future. NVIDIA Integration: 1. Storage must offer NVIDIA GPUDirect Storage connectivity to GPUs. 2. NVMe Storage offered must be certified with the latest Nvidia DGX Super-POD and should be the same model with certification. **Front-End Connectivity:** 100GBE or higher Ethernet connectivity compatible with proposed inferencing nodes.

Specifications: 42U Server Rack		
Sr No	Parameter	Minimum Specifications
1	Form Factor with Width & Depth	42U Server Rack should be 800mm (Width) x — l400mm (Depth)
2	Cabinet Type & Construction	Rack Frame should be robust and made of welded steel frame that offers strong and sturdy support for installation of 19" equipment and accessories. Rack Frame made of Steel Pro-file and connected with Horizontal Profiles for Width and Depth. Depth support channel with adjustable mounting slots.
3	Cable Entry	Top and Bottom Panel with cable entry facility with Brush.
4	Mounting Angle	The 19” mounting angles should be provided 2 Nos. on front and rear side of the Rack. It should be adjustable full depth. 19” Mounting Angles made up of Steel 2mm Thickness with better mounting flexibility and maximizes usable mounting space.
5	"U" Identifica-tion	“U” numbering should be provided on the 19" mounting rails such that these unique num-bers are visible after mounting of the equipment also.
6	PDU Provision	Each rack should have provision for installation of two PDU with toolless mounting provi-sion to be connected to the two different sources individually.
7	Cable Manager Provision	Each rack should have 4 horizontal 1U Closed type cable manager.
8	Side Panels	Side Panel shall be covered with horizontally split steel panels The side panels should be easily detachable with locking provision.
9	Door	Front and Rear doors should be perforated and both front and rear doors should be at least 80
		% hexagonal perforated (Holes). Front & Rear Door should be with Minimum of
		138 degrees to allow easy access to the interior.
10	Door Perfora-tion	Hexagaonal Perforated Single Front Door will be Lockable and - handle Lock & Key should be provided.
11	Door Lock	Hexagonal Perforated Dual Rear Door will be Lockable and 3 Point Lock should be provided.
12	Castor	Rack should be with Plinth of 800 MMW, 100MM H and 1400 MM D. The rack shall be not having External height >2060 mm including Plinth.
13	Load Bearing	Minimum load bearing capacity supported by Base Frame should be static load of at least -1200 Kg.
14	Powder Coating	Rack shall be pre-treated and powder coated. The Powder coating process shall be ROHS compliant. Powder coating thickness shall be 80 to 100 microns. The color of the powder coat shall be Black.
15	PDU	Each rack shall be provided with 3 Nos. of 3 PHASE 63A PDU IEC C19 X 12 SKT (PER SOCKET IEC C19 X 4 SOCKET + 63A D Curve DP MCB) X 3 + 16 SQ MM 5 CORE 3.5 MTR FRLS CABLE WITH 5 PIN 63A IND PLUG (2No Vertical and 1 No Horizontal)
16	Shelf	1No Heavy Duty Shelf for keeping the Display & Keyboard
17	Door Construc-tion	All Racks & Doors are inherently grounded to Rack Frame. Both the front and rear doors should be designed with quick-release hinges allowing for quick and easy detachment with-out the use of tools. The front door of unit should be field reversible so that it may open from either side.
18	Statutory Stand-ard	100% assured compatibility with all equipment conforming to DIN 41494 / EIA 310-D standard(General industrial standard for equipment).
19	Certification	The rack shall be from OEM having ISO9001:2008,

		ISO14001:2004, ISO 45001:2018 & ISO 50001:2018
		(Certificate to be submitted along with compliance)
21	Warranty	5 years onsite comprehensive warranty

Indicative Diagram



Note:

- Bidders should refer to the indicative diagram for reference and propose their solutions accordingly.
- Bidder has to conduct site visits in advance (before the bid submission date) during working days and hours to assess the rack positioning. Based on this assessment, they should quote their solution in the bid submission.
- In addition, Bidder has to connect proposed solution with existing storage at GSDC as following;

  Existing Storage

  Net App NAS Storage
  Hitachi NVME SAN Storage
- The bidder has to integrate the existing storage with the GPU node. All necessary accessories, cabling, hardware, software, and licenses should be considered accordingly.

3. **PRICE BID SCHEDULE (On GEM):**

Sr. No.	Description	Cost including GST (Rs.)
1	**GPU Compute for uses AI / ML at GSDC:** a. Inclusive of all the required hardware, Software and necessary Licenses required to make the solution fully functional. b. As per the Scope of work, functional and technical requirement, including racks, cable & all other accessories (including active & passive components), Installation, testing, commissioning and training etc. c. Two Skilled Resource as per Scope d. Cost of **Comprehensive Annual Maintenance with warranty and OEM support** for 5 year	
	Total cost (Rs.)	

Note:

- L1 will be the lowest sum total of rates of all line items including GST as per GeM GTC.
- TENDERER/GIL may negotiate the prices with L1 Bidder, under each item/head offered by Bidder.
- The L1 Bidder shall share the Item Wise cost breakup with the tenderer for future reference for scalability and additional components within the solution.
- Enterprise level highest license and support for complete solution should be provided from day one.
- RA has been enabled in the GEM Bid.

Please submit the undertaking letter as per Ministry of Finance Memorandum No.: F. No.6/18/2019-PPD dated 23.07.2020 & Office Memorandum No.: F.18/37/2020-PPD dated 08.02.2021 as per Performa given below on OEM letterhead as well as on bidder's letterhead.

On Letterhead of Bidder

Sub: Undertaking as per Office Memorandum No.: F. No.6/18/2019-PPD dated 23.07.2020 & Office Memorandum No.: F.18/37/2020-PPD dated 08.02.2021 published by Ministry of Finance, Dept. of Expenditure, Public Procurement division

Ref: Bid Number: _____

I have read the clause regarding restriction on procurement from a bidder of a country that shares a land border with India. I certify that we as a bidder and quoted product from the following OEMs are not from such a country or if from such a country, these quoted products OEM has been registered with the competent authority. I hereby certify that these quoted product & its OEM fulfills all requirements in this regard and is eligible to be considered for procurement for Bid number_____.

No.	Item Category	Quoted Make & Model

In case I'm supplying material from a country which shares a land border with India, I will provide evidence for valid registration by the competent authority, otherwise GIL/End user Dept. reserves the right to take legal action on us.

(Signature)
Authorized Signatory of **M/s <<Name of Company>>**

On Letterhead of OEM

Sub: Undertaking as per Office Memorandum No.: F. No.6/18/2019-PPD dated 23.07.2020 & Office Memorandum No.: F.18/37/2020-PPD dated 08.02.2021 published by Ministry of Finance, Dept. of Expenditure, Public Procurement division

Ref: Bid Number: _____

Dear Sir,

I have read the clause regarding restriction on procurement from a bidder of a country that shares a land border with India. I certify that our quoted product and our company are not from such a country, or if from such a country, our quoted product and our company have been registered with the competent authority. I hereby certify that these quoted products and our company fulfills all requirements in this regard and is eligible to be considered for procurement for Bid number_____.

No.	Item Category	Quoted Make & Model

In case I'm supplying material from a country which shares a land border with India, I will provide evidence for valid registration by the competent authority; otherwise GIL/End user Dept. reserves the right to take legal action on us.

(Signature)
Authorized Signatory of **M/s <<Name of Company>>**