GUJARAT INFORMATICS LIMITED
Block No. 2, 2ND Floor, C & D Wing, Karmayogi Bhavan, Sector -10A, GANDHINAGAR - 382010

Revised RFP (Corrigendum-03 dated 16.05.2025)

&

Responses to bid Queries dated 16.05.2025

Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Please find the below Corrigendum-3 dated 16.05.2025.

For more details visit www.gil.gujarat.gov.in

The bidder shall submit their queries, if any, within 07 days from the date of publication of Corrigendum-03. No further clarifications regarding Corrigendum-03 or any earlier published corrigendum shall be entertained by the Tenderer after this period. In addition, no new queries shall be accepted within this 7 day also.

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
1	**Corrigendum -2** , SLA & Penalties (Operational Penalties), Page-23	Timeline for resolution is within 4 hours from the time of call logged / reported to Bidder/OEM.	We request GIL to amend it as NBD (Next Business Day) resolution for Hardware related issues, during working days. As sparing locally for such High end GPU Hardware is not feasible for OEMS.	-	please refer corrigendum-03.
2	**Corrigendum 2 : 10. Minimum Technical Specification Master Node, Training Nodes, Inferecing Nodes Page-24, 27, 28**	**Certification :** Undertaking from Sever OEM for compatibility of the proposed sever with GPU under the quoted Inference node must be submitted (duly signed by authorized signatory , mentioning Bid reference)	We would like to draw GILs attention to these clauses that have been relaxed in Corrigendum 2 , will allow unproven systems being quoted, which are neither tested nor certified by GPU OEMS or benchmarked and submitted, audited results to reputed reference sites like MLcommons.org, that is referred, maintain by leading OEMs for GPU, Servers. This relaxation , will risk the unproven infrastructure being proposed, resulting into undesired, uncertain results. There are GIDs, who would just extrapolate and submit undertakings!! risking the complete bid.	-	As per RFP and time to time published corrigendum
		Benchmarks : If not listed on MLcommons, bidder shall be required to submit benchmark report for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details		-	As per RFP and time to time published corrigendum
		Benchmarks : If not listed on spec.org, bidder shall be required to submit benchmark report / logs for the Make/Model (same configuration) of the quoted server as part of bid submission		-	As per RFP and time to time published corrigendum
5	**Corrigendum 2** : 10. Minimum Technical Specification Master Node, Page-24 & Inferencing nodes, Page-27	32 Gbps Host Bus Adaptor with required SFP (at both end) for connecting with existing SAN switch and storage.	We request GIL to clarify, what does (at both end) means here !!!	-	Clarification: Both end in the clause refer to 32 Gbps SFP required for existing SAN Switch at GSDC and SFPs (if required) for the supplied servers maintaining a card level redundency Please refer to corrigendum for make & model of existing SAN and storage
6	**Corrigendum 2** : 10. Minimum Technical Specification Nodes for AI training, Page-25	Network : a) Infiniband / Ethernet (200GBps or higher) as required for quoted storage delivery to nodes	Network : a) Infiniband / Ethernet (200Gbps or higher) as required for quoted storage delivery to nodes **Justification** : The Network card be it infiniband or Ethernet the ratings are in Gbps and NOT GBps, the same can be noted in Training nodes, where Gbps is specified., we request GIL to amend appropraitly, refer to link : https://www.nvidia.com/content/dam/en-zz/Solutions/networking/ethernet-adapters/connectx-7-datasheet-Final.pdf	-	please refer Corrigendum-03.
7	**Corrigendum 2** : 10. Minimum Technical Specification AI Training Node, Page-25 & Inferecing Nodes, Page-27	Internal Storage : For Operating System: minimum 1.92 TB NVMe drives	**Modification Internal Storage** : For Operating System: minimum 1.92 TB NVMe drives OR M.2 960GB NVMe drives . **Justification** : Current changes in Corrigendum 2 are **restricting OEMs to participate,** earlier in Corrigendum 1, this was allowed. Hence request GIL to relax the clasue for wider OEMs participation. without impacting any perfomrance of the nodes. as Operating system are small in size and does not requried to be mised with performacne capacity drives and large dries .	-	please refer Corrigendum-03.
8	Corrigendum 2 : 10. Minimum Technical Specification AI Training Nodes, Page-25	**System Network** : Following N/W are required: 1. NDR Infiniband / Ethernet (400GBps or higher) for compute communication 2. Infiniband /Ethernet (200GBps or higher) for storage delivery	**Modification : Following N/W are required:** 1. NDR Infiniband / Ethernet (400Gbps or higher) for compute communication 2. Infiniband /Ethernet (200Gbps or higher) for storage delivery justification : The Network card are NOT rated in GBps, as suggested and sahred the link in earlier point, please refer again : https://www.nvidia.com/content/dam/en-zz/Solutions/networking/ethernet-adapters/connectx-7-datasheet-Final.pdf	-	please refer Corrigendum-03.
9	Corrigendum 2 : 10. Minimum Technical Specification Inferencing Nodes, Page-27	Power Requriement : Appropriate rated and energy efficient, redundant (N+N) hot swappable power supply and FAN	We reqeust GIL to defined the N+N, as N=1, would means N+1 redudnant hotswappable power supply to be offered	-	please refer corrigendum-03.
10	Corrigendum 2 : 10. Minimum Technical Specification, Inference Node, Page-27	**Number of GPUs and** **GPU Communication** : 2 x Accelerators per node, each with minimum 140 GB or higher GPU per Accelerator.	**We request GIL to open up this clause for wider OEMs particiation ,as current clause is restrictive in nature.** 2 x Accelarators per node, each with minimum 140 GB or higher GPU accelarators OR 4 x Accelarators per node , each with minimum 94GB or higher GPU OR 4x Accelarator per node, each with minimum 140GB or higher GPU (total 24 GPUs to be offerd acreoss 6 GPU Nodes for inferencing). OR 8x Accelarator per node, each with minimum 140GB or higher GPU (total 24 GPUs to be offerd acreoss 3 GPU Nodes for inferencing). **Justification : we request GIL to consider the suggested changes, as All leading OEM and GPU OEM does not have ratings, configurations available and benchmakred.** **Also this will help GIL to have better options with better Rack Space, Power Cooling for Data Center. without any compromise on performacne.**	-	As per RFP and time to time published corrigendum. However, Bidder may propose 2 or higher accelerator card per node (In inference Node) and numbers of inference node quantity may be change accordingly. However, total requirement of core and memory may be provided a s mentioned in RFP and time to time published corrigendum.
11	Nodes for AI training: Total Qty-4 sets. Benchmarks, Page-27	DLRM-dcnv2 – 3.6 minutes or less on single node GNN – 7.8 minutes or less on single node ResNet – 12.1 minutes or less on single node U-Net3D – 11.6 minutes or less on single node	**Modification** : We request GIL to revise the following benchmarks , as available on https://mlcommons.org/ DLRM-dcnv2 – 3.75 minutes or less on single node GNN – **Please clarify are we referring to RGAT benchmark for GNN**? ResNet – 13.25 minutes or less on single node U-Net3D – 12.42 minutes or less on single node **Justification : We requst to amend as request for Wider OEM participation in the bid.**	-	Please refer to last Corrigendum-2 dated 15.03.2025. Tolerance of 25% is allowed on the benchmark in AI Training model. if GNN is not available, RGAT is accepted instead of GNN.
12	Corrigendum 2 : 10. Minimum Technical Specification, Inference Node, Page-28	**OS Support** : The system should support latest version of Red Hat Enterprise Linux /Ubuntu Linux / RHEL AI / Red Hat OpenShift AI server, Quoted OS should be under Enterprise support from OEM with premium or highest level of support.	We requesg GIL to clarify, which of these to be quoted as part of solution for better clarity!!		please refer Corrigendum-03.
13	Corrigendum 2 : 10. Minimum Technical Specification, Inference Node, Page-28	**OS Support** : Supply should include DC edition unlimited Guest OS licenses	We request GIL to define the user base, as this ask of Unlimited Guest OS, will force bidder to quote highest level of subscritpion license, escalating the cost, which may not be utilized by GIL for years to come!! Hence defineing the user license will make it optimized.	-	please refer Corrigendum-03.

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
14	**Storage Nodes, External Storage, Page -30**	"The solution should be PFS (Parallel File System) based and delivered with 1PB (All NVMe) usable post RAID 6/equivalent or better protection, expandable up to 2PB in the same file system. "	**Modification** : **The solution should be PFS (Parallel File System) OR NFS (Network File System)** based and delivered with 1PB (All NVMe) usable post RAID 6/equivalent or better protection, expandable up to 2PB in the same file system. **Justification** :We request GIL to refer to corrigendum 1, NFS solution was allowed with required performcne, however in Corrigneudm 2 the NFS has be removed and PFS has been introduced with unresonably HIGH performacne number, which has no direct relationship with overall Training or Inferencing nodes throughput and network calcualtions!! Hence request GIL to relax this restrictive clasue for wider OEMs participation.	-	please refer Corrigendum-03.
15	**Storage Nodes, External Storage, Page -30**	Performance: Min 120 GBps Read and Min 60 GBps Write aggregated from day one and scalable up to 200% with a scale-out architecture and additional controllers/nodes in the future. IOPS: minimum 8,00,000 1. Storage must offer NVIDIA GPUDirect Storage connectivity to GPUs. 2. NVMe Storage offered must be certified with the proposed GPU OEM. Front-End Connectivity: 200GBE or higher Ethernet connectivity compatible with all nodes as per proposed solution.	**Modification** : **(As per Corrigendum 1)**. Performance: 28GBps Read and 14GBps Write from day one and scalable up to >100GBps read write combinations with a scale-out architecture and additional controllers/nodes in the future. **OR** Performance: Min 120 GBps Read and Min 60 GBps Write aggregated from day one and scalable up to 200% with a scale-out architecture and additional controllers/ nodes in the future. **Justification :: We request GIL to allow** NFS solution, as all current data is on NFS solution, converting, migrating NFS to PFS will take lot of time, that will further delay the productivity of the AI/ML setup. !! Relaxing **this restrictive clause** will help wider OEM participation with established results, deployments of AI/ML Solution , will Even NVIDIA does not have any bias towards PFS, as both perform well for AI/ML environment. **Current Specifications are OEM Specific and limiting participation by Leading OEMs.** Please refer the link shared again here with for detailed view, where it clearly shows NVIDIA SuperPod recommendation as well As per Nvidia documentation, for DGX superpod, available at Storage Architecture — NVIDIA DGX SuperPOD: Next Generation Scalable Infrastructure for AI Leadership Reference Architecture Featuring NVDIA DGX H200 , the storage performance is as under Single SU* (256 GPU) aggregate system read = 125GBps Single SU* (256GPU) aggregate system write = 62GBPs *A single Scalable Unit (SU) in an NVIDIA DGX SuperPOD consists of 32 DGX systems, each system with 8 GPUs of H200 **Hence for a 56GPU (4 training nodes with 8 GPU per node, and 12 inference server with 2 GPU per node) system, the maximum throughput we should need is as under:** **Aggregate system read = 28GBps** **Aggregate system write = 14GBps**	-	please refer Corrigendum-03.
16	**Master Node, Page-24, Network Section**	InfiniBand / Ethernet (200GBps or higher) as required for quoted storage delivery to nodes	**Modification**: Ethernet (200Gbps or higher) as required for quoted storage delivery to nodes **Justification**: **External Storage mentions**: "Front-End Connectivity: 200GBE or higher Ethernet connectivity" the mentioned GBps is not the correct rating ,as the HBA/NIC are available from leading OEMS.	-	please refer Corrigendum-03.
17	**Nodes for AI Training, Page-25, Network Section**	d) Required InfiniBand / 200G or higher Ethernet 2 x twin-port HCA as required for quoted storage delivery to node	**Modification**: Required 200G or higher Ethernet 2 x twin-port HCA as required for quoted storage delivery to node **Justification**: **External Storage mentions**: "Front-End Connectivity: 200GBE or higher Ethernet connectivity"	-	As per RFP and time to time published corrigendum
18	**Nodes for AI Training, Page-25, Network Section**	f) Required switch with 64 non-blocking ports with aggregate data throughput up to 51.2 Tb/s and required compatible cables of appropriate length to connect all 8 (compute communication) nos. of IB NDR ports / Ethernet of all nodes in non-blocking mode	**Modification**: Required **two** switches, **each** with 64 non-blocking ports and **1RU or 2RU form factor** with aggregate data throughput up to 51.2 Tb/s and required compatible cables of appropriate length to connect all 8 (compute communication) nos. of IB NDR ports / Ethernet of all nodes in non-blocking mode **Justification**: Chassis based switches can be very power hungry. 1 or 2RU form factor switches consume power around 2000W depending on the number of transceivers used. Also, with 2 switches, a redundancy can be formed in the back-end GPU connectivity so that failure of one switch can still allow GPU-to-GPU traffic to flow through at 400G	-	Clarification: The bidder shall have to deploy required quantity of switch with said functionality to complete the solution.
19	**Nodes for AI Training, Page-26, System Network Section**	2. InfiniBand /Ethernet (200GBps or higher) for storage delivery	**Modification**: Ethernet (200Gbps or higher) for storage delivery **Justification**: **External Storage mentions**: "Front-End Connectivity: 200GBE or higher Ethernet connectivity"	-	please refer Corrigendum-03.
20	**Inference Node, Page 28, System Network Section**	4. InfiniBand / Ethernet (200GBps or higher) as required for quoted storage delivery to nodes	**Modification**: Ethernet (200Gbps or higher) as required for quoted storage delivery to nodes **Justification**: **External Storage mentions**: "Front-End Connectivity: 200GBE or higher Ethernet connectivity"	-	please refer Corrigendum-03.
21	**Networking Switch, Page-28**	Min. Two or required 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	**Modification**: Min. **Four** or **higher** Nos. of Switch with 48 ***10/25** G SFP+ and 6 x 100G QSFP ports or higher to connect to Core for 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 and **EVPN-Multihoming** feature. **Justification**: Ideally, it is better to physically separate the perimeter traffic and cluster orchestration traffic. Across all servers, the total perimeter ports are 46 and orchestration node ports are 46. Additional layer of segregation can be done using VLAN, VxLAN and VRFs. Secondly, 25G standard is becoming highly common in DC use case. Having 25G as switch port option can allow adding servers with 25G ports, instead of buying new switches. Additionally, 6x 100G is adequate to connect to uplink towards Core / Spine switch. **Multihoming** allows the switch to form a setup similar to "MC-LAG" without needing to directly interconnect switches.	-	please refer corrigendum-03. Clarification: The bidder must deploy the required quantity of switches with the same or higher functionality to meet the solution requirements. The switch count shall be adjusted (increased/decreased) based on the actual port availability per device while maintaining the specified speed and functionality. EVPN -multihoming feature can be provided, in case it is required to complete the solution.

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
22	**Networking Switch, Page-28**	Min. One or required 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	**Modification**: Min. **Two** or **higher** Nos. of Switch with 48 *1G RJ45, 4* **10G** 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. **Justification**: There are a total of 23 servers (Master, Training and Inferencing). In addition, there will be other devices like switches, firewall, storage, load-balancers, etc. These devices will be spread across multiple racks considering the power constraint of the rack and the limit on the number of devices it can have. Having 1 OOB switch can become an issue with respect to cable laying across the racks including the distance. That's why to ensure that all bidders are able to follow appropriate DC design standard, minimum 2 OOB switches will be helpful / necessary.		please refer corrigendum-03. Clarification: The bidder shall deploy the required quantity of switches with equal or higher capability to meet the solution's needs. The switch count may be adjusted (increased/decreased) based on actual device port availability, provided the specified speed and functionality are maintained.
23	**Networking Switch, Page-28**	Min. One or required 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	**Modification**: Min. **Two** or **higher** 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 and **EVPN-Multihoming** feature. **Justification**: A Server NIC usually have 2x 100G NIC, instead of 1. Moreover, single connectivity to switch (instead in HA) can be a single point of failure and losing connectivity to user traffic is a switch or transceiver or fiber cable were to go bad. Therefore, specifying two switches which can be in HA will be good option to ensure all bidders comply to minimal baseline. **Multihoming** allows the switch to form a setup similar to "MC-LAG" without needing to directly interconnect switches.	-	please refer corrigendum-03. Clarification: The bidder shall have to deploy the required quantity of switches with said or higher functionality to complete the solution. EVPN-multihoming is not mandatory with the solution Increase/decrease the count of switches as per the ports available with the mentioned speed and functionality to complete the solution. EVPN -multihoming feature can be provided, in case it is required to complete the solution.
24	**Corrigendum -2** , SLA & Penalties (Operational Penalties), Page-23	Timeline for resolution is within 4 hours from the time of call logged / reported to Bidder/OEM.	We request GIL to amend it as NBD (Next Business Day) resolution for Hardware related issues, during working days. As sparing locally for such Hardware is not feasible for OEMS.	-	please refer corrigendum-03.
25	**Corrigendum 2 : 10. Minimum Technical Specification Master Node, Training Nodes, Inferecing Nodes Page-24, 27, 28**	**Certification :** Undertaking from Sever OEM for compatibility of the proposed sever with GPU under the quoted Inference node must be submitted (duly signed by authorized signatory , mentioning Bid reference)	We would like to draw GILs attention to these clasues that have been relaxed in **Corrigendum 2** , will allow unproven systems being quoted, which are neither tested nor certified by GPU OEMS or benchmarked and submiitted, auditted results to reputed refrence sites like MLcommons.org, that is reffrred, maintain by leading OEMs. This relaxation , will risk the unproven infrastrucure being proposed, resulting into undesired, uncertain results. There are OEMs, who would jsut extrapolate and and submitt understakings!! risking the compelet bid.	-	As per RFP and time to time published corrigendum
26		**Benchmarks :** If not listed on Mlcommons, bidder shall be required to submit benchmark report for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details		-	As per RFP and time to time published corrigendum
27		**Benchmarks :** If not listed on spec.org, bidder shall be required to submit benchmark report / logs for the Make/Model (same configuration) of the quoted server as part of bid submission		-	As per RFP and time to time published corrigendum
28	**Corrigendum 2** : 10. Minimum Technical Specification Master Node, Page-24 & Inferencing nodes, Page-27	32 Gbps Host Bus Adaptor with required SFP (at both end) for connecting with existing SAN switch and storage.	We request GIL to clarify, what does (at both end) means here !!! Which Existing SAN Switch we need to integrate. Kindly share Make/Model no of the SAN Switch.	-	Clarification: Both end in the clause refer to 32 Gbps SFP required for existing SAN Switch at GSDC and SFPs (if required) for the supplied servers maintaining a card level redundency. Please refer to corrigendum for make & model of existing SAN switches and storage.
29	**Corrigendum 2** : 10. Minimum Technical Specification Nodes for AI training, Page-25	Network : a) Infiniband / Ethernet (200GBps or higher) as required for quoted storage delivery to nodes	Network : a) Infiniband / Ethernet (200Gbps or higher) as required for quoted storage delivery to nodes **Justification** : The Network card be it infiniband or Ethernet the ratings are in Gbps and NOT GBps, the same can be noted in Training nodes, where Gbps is specified., we request GIL to amend appropraitly, refer to link : https://www.nvidia.com/content/dam/en-zz/Solutions/networking/ethernet-adapters/connectx-7-datasheet-Final.pdf	-	please refer Corrigendum-03.
30	**Corrigendum 2** : 10. Minimum Technical Specification AI Training Node, Page-25 & Inferecing Nodes, Page-27	Internal Storage : For Operating System: minimum 1.92 TB NVMe drives	**Modification Internal Storage** : For Operating System: minimum 1.92 TB NVMe drives OR M.2 960GB NVMe drives . **Justification** : Current changes in Corrigendum 2 are restricting OEMs to participate, earlier in Corrigendum 1, this was allowed. Hence request GIL to relax the clasue for wider OEMs participation. without impacting any perfomrance of the nodes. as Operating system are small in size and does not requried to be mised with performacne capacity drives and large dries .	-	please refer Corrigendum-03.
31	Corrigendum 2 : 10. Minimum Technical Specification AI Training Nodes, Page-25	**System Network** : Following N/W are required: 1. NDR Infiniband / Ethernet (400GBps or higher) for compute communication 2. Infiniband /Ethernet (200GBps or higher) for storage delivery	**Modification : Following N/W are required:** 1. NDR Infiniband / Ethernet (400Gbps or higher) for compute communication 2. Infiniband /Ethernet (200Gbps or higher) for storage delivery **Justification :** The Network card are NOT rated in GBps, as suggested and sahred the link in earlier point, please refer again : https://www.nvidia.com/content/dam/en-zz/Solutions/networking/ethernet-adapters/connectx-7-datasheet-Final.pdf	-	please refer Corrigendum-03.
32	Corrigendum 2 : 10. Minimum Technical Specification Inferencing Nodes, Page-27	Power Requriement : Appropriate rated and energy efficient, redundant (N+N) hot swappable power supply and FAN	We reqeust GIL to defined the N+N, as N=1, would means N+1 redudnant hotswappable power supply to be offered	-	please refer corrigendum-03.

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
33	Corrigendum 2 : 10. Minimum Technical Specification, Inference Node, Page-27	**Number of GPUs and GPU Communication** : 2 x Accelerators per node, each with minimum 140 GB or higher GPU per Accelerator.	**We request GIL to open up this clause for wider OEMs particiation ,as current clause is restrictive in nature.** 2 x Accelarators per node, each with minimum 140 GB or higher GPU accelarators OR 4 x Accelarators per node , each with minimum 94GB or higher GPU OR 4x Accelarator per node, each with minimum 140GB or higher GPU (total 24 GPUs to be offerd acreoss 6 GPU Nodes for inferencing). OR 8x Accelarator per node, each with minimum 140GB or higher GPU (total 24 GPUs to be offerd acreoss 3 GPU Nodes for inferencing). **Justification :** we request GIL to consider the suggested changes, as All leading OEM and GPU OEM does not have ratings, configurations available and benchmakred. Also this will help GIL to have better options with better Rack Space, Power Cooling for Data Center. without any compromise on performacne.	-	As per RFP and time to time published corrigendum. However, Bidder may propose 2 or higher accelerator card per node (In inference Node) and numbers of inference node quantity may be change accordingly. However, total requirement of core and memory may be provided a s mentioned in RFP and time to time published corrigendum.
34	Nodes for AI training: Total Qty-4 sets. Benchmarks, Page-27	DLRM-dcnv2 – 3.6 minutes or less on single node GNN – 7.8 minutes or less on single node ResNet – 12.1 minutes or less on single node U-Net3D – 11.6 minutes or less on single node	**Modification** : We request GIL to revise the following benchmarks , as available on https://mlcommons.org/ DLRM-dcnv2 – 3.75 minutes or less on single node GNN – **Please clarify are we referring to RGAT benchmark for GNN**? ResNet – 13.25 minutes or less on single node U-Net3D – 12.42 minutes or less on single node **Justification : We requst to amend as request for Wider OEM participation in the bid.**	-	Please refer to last Corrigendum-2 dated 15.03.2025. if GNN is not available, RGAT is accepted instead of GNN.
35	Corrigendum 2 : 10. Minimum Technical Specification, Inference Node, Page-28	**OS Support** : The system should support latest version of Red Hat Enterprise Linux /Ubuntu Linux / RHEL AI / Red Hat OpenShift AI server, Quoted OS should be under Enterprise support from OEM with premium or highest level of support.	We requesg GIL to clarify, which of these to be quoted as part of solution for better clarity!!	-	please refer Corrigendum-03.
36	Corrigendum 2 : 10. Minimum Technical Specification, Inference Node, Page-28	**OS Support** : Supply should include DC edition unlimited Guest OS licenses	We request GIL to define the user base, as this ask of Unlimited Guest OS, will force bidder to quote highest level of subscritpion license, escalating the cost, which may not be utilized by GIL for years to come!! Hence defineing the user license will make it optimized.	-	please refer Corrigendum-03.
37	**Storage Nodes, External Storage, Page -30**	Performance: Min 120 GBps Read and Min 60 GBps Write aggregated from day one and scalable up to 200% with a scale-out architecture and additional controllers/nodes in the future. IOPS: minimum 8,00,000 1. Storage must offer NVIDIA GPUDirect Storage connectivity to GPUs. 2. NVMe Storage offered must be certified with the proposed GPU OEM. Front-End Connectivity: 200GBE or higher Ethernet connectivity compatible with all nodes as per proposed solution.	**Modification** : **(As per Corrigendum 1)**. Performance: 28GBps Read and 14GBps Write from day one and scalable up to >100GBps read write combinations with a scale-out architecture and additional controllers/nodes in the future. **OR** Performance: Min 120 GBps Read and Min 60 GBps Write aggregated from day one and scalable up to 200% with a scale-out architecture and additional controllers/nodes in the future. **Justification :: We request GIL to allow** NFS solution, as all current data is on NFS solution, converting, migrating NFS to PFS will take lot of time, that will further delay the productivity of the AI/ML setup. !! Relaxing **this restrictive claus**e will help wider OEM participation with established results, deployments of AI/ML Solution , will Even NVIDIA does not have any bias towards PFS, as both perform well for AI/ML environment. **Current Specifications are OEM Specific and limiting participation by Leading OEMs.** Please refer the link shared again here with for detailed view, where it clearly shows NVIDIA SuperPod recommendation as well. As per Nvidia documentation, for DGX superpod, available at Storage Architecture — NVIDIA DGX SuperPOD: Next Generation Scalable Infrastructure for AI Leadership Reference Architecture Featuring NVDIA DGX H200 , the storage performance is as under Single SU* (256 GPU) aggregate system read = 125GBps Single SU* (256GPU) aggregate system write = 62GBPs *A single Scalable Unit (SU) in an NVIDIA DGX SuperPOD consists of 32 DGX systems, each system with 8 GPUs of H200 **Hence for a 56GPU (4 training nodes with 8 GPU per node, and 12 inference server with 2 GPU per node) system, the maximum throughput we should need is as under:** **Aggregate system read = 28GBps** **Aggregate system write = 14GBps**	-	please refer Corrigendum-03.
38	**Master Node, Page-24, Network Section**	InfiniBand / Ethernet (200GBps or higher) as required for quoted storage delivery to nodes	**Modification**: Ethernet (200Gbps or higher) as required for quoted storage delivery to nodes **Justification**: **External Storage mentions**: "Front-End Connectivity: 200GBE or higher Ethernet connectivity" the mentioned GBps is not the correct rating ,as the HBA/NIC are available from leading OEMS.	-	please refer Corrigendum-03.
39	**Nodes for AI Training, Page-25, Network Section**	d) Required InfiniBand / 200G or higher Ethernet 2 x twin-port HCA as required for quoted storage delivery to node	**Modification**: Required 200G or higher Ethernet 2 x twin-port HCA as required for quoted storage delivery to node **Justification**: **External Storage mentions**: "Front-End Connectivity: 200GBE or higher Ethernet connectivity"	-	As per RFP and time to time published corrigendum
40	**Nodes for AI Training, Page-25, Network Section**	f) Required switch with 64 non-blocking ports with aggregate data throughput up to 51.2 Tb/s and required compatible cables of appropriate length to connect all 8 (compute communication) nos. of IB NDR ports / Ethernet of all nodes in non-blocking mode	**Modification**: Required **two** switches, **each** with 64 non-blocking ports and **1RU or 2RU form factor** with aggregate data throughput up to 51.2 Tb/s and required compatible cables of appropriate length to connect all 8 (compute communication) nos. of IB NDR ports / Ethernet of all nodes in non-blocking mode **Justification**: Chassis based switches can be very power hungry. 1 or 2RU form factor switches consume power around 2000W depending on the number of transceivers used. Also, with 2 switches, a redundancy can be formed in the back-end GPU connectivity so that failure of one switch can still allow GPU-to-GPU traffic to flow through at 400G	-	Clarification: The bidder must deploy the required quantity of switches with the same or higher functionality to meet the solution requirements. The switch count shall be adjusted (increased/decreased) based on the actual port availability per device while maintaining the specified speed and functionality.
41	**Nodes for AI Training, Page-26, System Network Section**	2. InfiniBand /Ethernet (200GBps or higher) for storage delivery	**Modification**: Ethernet (200GBps or higher) for storage delivery **Justification**: **External Storage mentions**: "Front-End Connectivity: 200GBE or higher Ethernet connectivity"	-	please refer Corrigendum-03.

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
42	**Inference Node, Page 28, System Network Section**	4. InfiniBand / Ethernet (200GBps or higher) as required for quoted storage delivery to nodes	**Modification**: Ethernet (200GBps or higher) as required for quoted storage delivery to nodes **Justification**: **External Storage mentions**: "Front-End Connectivity: 200GBE or higher Ethernet connectivity"	-	please refer Corrigendum-03.
43	**Networking Switch, Page-28**	Min. Two or required 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	**Modification**: Min. **Four** or **higher** Nos. of Switch with 48 ***10/25** G SFP+ and 6 x 100G QSFP ports or higher to connect to Core for 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 and **EVPN-Multihoming** feature. **Justification**: Ideally, it is better to physically separate the perimeter traffic and cluster orchestration traffic. Across all servers, the total perimeter ports are 46 and orchestration node ports are 46. Additional layer of segregation can be done using VLAN, VxLAN and VRFs. Secondly, 25G standard is becoming highly common in DC use case. Having 25G as switch port option can allow adding servers with 25G ports, instead of buying new switches. Additionally, 6x 100G is adequate to connect to uplink towards Core / Spine switch. **Multihoming** allows the switch to form a setup similar to "MC-LAG" without needing to directly interconnect switches.	-	please refer corrigendum-03. Clarification: The bidder shall have to deploy the required quantity of switches with said or higher functionality to complete the solution. EVPN-multihoming is not mandatory with the solution. The switch count may be adjusted (increased/decreased) based on actual device port availability, provided the specified speed and functionality are maintained. EVPN -multihoming feature can be provided, in case it is required to complete the solution.
44	**Networking Switch, Page-28**	Min. One or required 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	**Modification**: Min. **Two** or **higher** Nos. of Switch with 48 *1G RJ45, 4* **10G** 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. **Justification**: There are a total of 23 servers (Master, Training and Inferencing). In addition, there will be other devices like switches, firewall, storage, load-balancers, etc. These devices will be spread across multiple racks considering the power constraint of the rack and the limit on the number of devices it can have. Having 1 OOB switch can become an issue with respect to cable laying across the racks including the distance. That's why to ensure that all bidders are able to follow appropriate DC design standard, minimum 2 OOB switches will be helpful / necessary.	-	please refer corrigendum-03. Clarification: The bidder shall have to deploy the required quantity of switches with said or higher functionality to complete the solution. EVPN-multihoming is not mandatory with the solution. The switch count may be adjusted (increased/decreased) based on actual device port availability, provided the specified speed and functionality are maintained. EVPN -multihoming feature can be provided, in case it is required to complete the solution.
45	**Networking Switch, Page-28**	Min. One or required 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	**Modification**: Min. **Two** or **higher** 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 and **EVPN-Multihoming** feature. **Justification**: A Server NIC usually have 2x 100G NIC, instead of 1. Moreover, single connectivity to switch (instead in HA) can be a single point of failure and losing connectivity to user traffic is a switch or transceiver or fiber cable were to go bad. Therefore, specifying two switches which can be in HA will be good option to ensure all bidders comply to minimal baseline. **Multihoming** allows the switch to form a setup similar to "MC-LAG" without needing to directly interconnect switches.	-	please refer corrigendum-03. Clarification: The bidder shall have to deploy the required quantity of switches with said or higher functionality to complete the solution. EVPN-multihoming is not mandatory with the solution. The switch count may be adjusted (increased/decreased) based on actual device port availability, provided the specified speed and functionality are maintained. EVPN -multihoming feature can be provided, in case it is required to complete the solution.
46		Supply of the Hardware including Licenses and OEM Warranty Certificate. T1=T+60 days from the date of issuance of contract over GEM	Modification T1=T+90 days from the date of issuance of contract over GEM Justification: OEM Delivery timeline for Hardware taked minimum 3 months. Hence request authority to kindly look into this and amend the clause.	-	As per RFP and time to time published corrigendum.
47		POC to meet the benchmark as mentioned in this RFP document.	We request Authority to kindly provide POC Evaluation criteria	-	Please refer to RFP and time to time published Corrigendum
48		After Successful POC which is successful demonstration of the benchmark as mentioned in RFP,	We request Authorities to kindly elobarote the detail Benchmark criteria for successful DEMO and also provide scripts / tests to perform benchmark testing	-	Clarification: Will be share at the time of POC.
49		Submission Date Extension	Request authority to kindly give extension of BID subimission date of 20 working days from the date of publication of query responses.	-	please refer corrigendum-03.
50		Require Site Visit	We request Authority to kindly provide approval for site survey for study of - possibilities of integration - Understanding on integration / Networking with existing infrastructure - Overall Power requirement against needs	-	Please refer to RFP and time to time published Cor
51	24	32 Gbps Host Bus Adaptor with required SFP (at both end) for connecting with existing SAN switch and storage.	**Clarification-** No of 32G FC ports required per server and card level redundancy required or not	-	Clarification: Both end in the clause refer to 32 Gbps SFP required for existing SAN Switch at GSDC and SFPs (if required) for the supplied servers maintaining a card level redundency. Please refer to corrigendum for make & model of existing SAN switches and storage.
52	26	Appropriate rated and energy efficient, redundant (N+N) hot swappable power supply and FAN	**Change to**-Appropriate rated and energy efficient, redundant (N+1) or better hot swappable power supply	This in line with point 9 of page 18, where is is mentioned N+1 redundant Power supplies. Dense GPU servers typically come with N+1 redundancy. Dense GPU Servers do not come with Hot swappable fans, hence the request for change.	please refer corrigendum-03.
53	27	32 Gbps Host Bus Adaptor with required SFP (at both end) for connecting with existing SAN switch and storage.	**Clarification-** No of 32G FC ports required per server and card level redundacy required or not		Clarification: Both end in the clause refer to 32 Gbps SFP required for existing SAN Switch at GSDC and SFPs (if required) for the supplied servers maintaining a card level redundency. Please refer to corrigendum for make & model of existing SAN switches and storage.

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
54	25	c) Minimum 1 no. of 1 GbE port and 2 nos of 10 GbE or higher (Fiber/Copper) port.	**Change to**-c) ~~Minimum 1 no. of 1 GbE port and 2~~ 4 nos of 10 GbE or higher (Fiber/Copper) port.	Wider Participant and get better perfomance and redundancy	Clarification: The bidder can quote the product on higher side meeting the requirements to complete the solution.
55	27 and 29	AI Training and Inference Nodes - If not listed on MIcommons, bidder shall be required to submit benchmark report for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details.	Kindly change it to ' If not listed on MIcommons, bidder shall be required to submit benchmark report OR give an undertaking to meet the Benchmark target timings during POC (with upto 25 % tolerance as mentioned in the RFP) for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details.	Kindly allow us to give an undertaking to meet the benchmark target timings during POC (with upto 25 % tolerance as mentioned in the RFP).	Please refer to RFP and time to time published Corrigendum.
56	30	The proposed storage array should be configured with no single point of failure, including required controllers, cache, power supply, cooling fans, etc. It should be scalable up to 12 additional controllers/nodes.	The proposed storage array should be configured with no single point of failure, including required controllers, cache (if applicable), power supply, cooling fans, etc. It should be scalable up to 12 additional controllers/nodes.	Some Storage solutions do not come with cache in the controllers, hence the request for change.	As per RFP and time to time published corrigendum
57	30	IOPS: minimum 8,00,000	IOPS: minimum 8,00,000 Read	AI Storage Systems require a high Read Performance, hence the suggestion for change.	please refer Corrigendum-03.
58	30	NVMe Storage offered must be certified with the proposed GPU OEM.	NVMe Storage offered must be certified/compatible with the proposed GPU Server OEM.	To ensure wider choice of Storage solutions. Kindly approve.	As per RFP and time to time published corrigendum
59	27	Page 26 - Scalability, Cluster and Management Hardware and software and Page 27 and 29 - Cluster Management & Scheduler and hardware	Page 26 already mentions Scalability, Cluster and Management Hardware and software. Kindly remove the clause of 'Cluster Management & Scheduler and hardware ', for the Training and Inference nodes since the features asked are proprietary.	Having only the requirement of 'Scalability, Cluster and Management Hardware and software', will help in providing a uniform cluster tool.	As per RFP and time to time published corrigendum
60	28	Networking Switch : (2). Min. One or required 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.	**Change to -** (2). Min. One or required Nos. of Switch with 48 *1G RJ45, 4* 25G SFP28 or 50G SFPS6 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.		please refer corrigendum-03. Clarification: The bidder shall have to deploy the required quantity of switches with said or higher functionality to complete the solution. The switch count may be adjusted (increased/decreased) based on actual device port availability, provided the specified speed and functionality are maintained.
61		The solution should be PFS (Parallel File System) 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 solution should be PFS (Parallel File System) or NFS over RDMA based and delivered with 1PB (All NVMe) usable post RAID 6/equivalent or better protection, expandable up to 2PB in the same file system.	NFS over RDMA is a superior choice to PFS for AI/ML workloads using GPU Direct due to its **lower latency, higher throughput, and improved scalability**. By leveraging RDMA's direct memory-to-memory transfer, NFS over RDMA reduces latency and overhead, making it ideal for massive data transfers required in AI/ML workloads. Additionally, NFS is a standardized protocol, simplifying management and integration with existing infrastructure. In contrast, PFS, while designed for high-performance computing, **can be complex to set up and manage, and may not scale as well as NFS over RDMA**. When combined with GPU Direct, NFS over RDMA enables faster data transfer, reduced latency, and improved overall performance by offloading data transfer tasks from the CPU and allowing it to focus on compute-intensive tasks.	please refer Corrigendum-03.
63		1PB (NVMe) usable post RAID 6 or better configuration	1PB (NVMe TLC drives) usable post RAID 6 or better configuration	TLC NVMe drives are well-suited for AI/ML workloads due to their high capacity, lower cost, improved performance, and increased endurance, offering a better balance of performance, capacity,while significantly outperforming traditional HDDs and SATA SSDs, making them a practical choice for AI/ML applications that require rapid data access, large datasets, and frequent data writes.	As per RFP and time to time published corrigendum
65		**Performance:** Min 120 GBps Read and Min 60 GBps Write aggregated from day one and scalable up to 200% with a scale-out architecture and additional controllers/nodes in the future.	**Performance:** Min ~~120~~ 40 GBps Read and Min ~~60~~ 15 GBps Write aggregated from day one and scalable up to 200% with a scale-out architecture and additional controllers/nodes in the future.	According to the validated architecture jointly developed by Nvidia and NetApp, a read performance of **45 GBps** is more than sufficient to support **128 GPUs**. Given that our current requirement is for **32 GPUs**, the proposed solution will provide ample headroom for future expansions, easily supporting up to 64 GPUs without compromising performance. This ensures a scalable and future-proof infrastructure that can grow with our evolving needs. **Please refer to the below reference** **https://docs.netapp.com/us-en/netapp-solutions/ai/aipod_nv_validation_sizing.html#solution-validation**	please refer Corrigendum-03.
66		IOPS: minimum 8,00,000	~~IOPS: minimum 8,00,000~~	Throughput is a more critical metric than IOPS for AI/ML workloads, as these workloads typically involve processing large datasets and sequential data access patterns, requiring high-throughput storage solutions to transfer data quickly, whereas IOPS measures small I/O operations, which is less relevant for AI/ML workloads that prioritize high-speed data transfer.	please refer Corrigendum-03.
67		1. Storage must offer NVIDIA GPUDirect Storage connectivity to GPUs.			Query not clear
68		2. NVMe Storage offered must be certified with the proposed GPU OEM.			Query not clear
69					Query not clear
70		**Front-End Connectivity:** 200GBE or higher Ethernet connectivity compatible with all nodes as per proposed solution.			Query not clear

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
71		**Additional points**	Security : 1. Offered Storage solution must offer tamperproof snapshots for the data with capability to automatically create snapshot and expire them by defining retention period. minimum 1000 snapshots must be supported. 2. Offered Storage solution must support native or addon solution to Identify Ransomware Attacks, Take autonomous actions to protect the data from ransomware attacks, report the attack to administrators and offer recovery capabilities to the administrators.	Including tamperproof snapshots and ransomware protection in the storage specifications for AI/ML workloads is crucial to ensure data integrity, version control, compliance, and rapid recovery. AI/ML workloads are data-intensive and require high-performance storage solutions, making it essential to protect data from unauthorized modifications, deletions, and ransomware attacks. Tamperproof snapshots provide a reliable way to track changes and maintain version control, while ransomware protection ensures real-time detection and prevention of attacks, enabling rapid recovery and minimizing downtime. Furthermore, these features are critical for regulated industries, such as healthcare and finance, where strict data protection and retention policies are mandatory.	As per RFP and time to time published corrigendum
72	24	32 Gbps Host Bus Adaptor with required SFP (at both end) for connecting with existing SAN switch and storage.	**Clarification-** No of 32G FC ports required per server and card level redundancy required or not		Clarification: Both end in the clause refer to 32 Gbps SFP required for existing SAN Switch at GSDC and SFPs (if required) for the supplied servers maintaining a card level redundency. Please refer to corrigendum for make & model of existing SAN switches and storage.
73	26	Appropriate rated and energy efficient, redundant (N+N) hot swappable power supply and FAN	**Change to**-Appropriate rated and energy efficient, redundant (N+1) or better hot swappable power supply	This in line with point 9 of page 18, where is is mentioned N+1 redundant Power supplies. Dense GPU servers typically come with N+1 redundancy. Dense GPU Servers do not come with Hot swappable fans, hence the request for change.	please refer corrigendum-03.
74	27	32 Gbps Host Bus Adaptor with required SFP (at both end) for connecting with existing SAN switch and storage.	**Clarification-** No of 32G FC ports requored per server and card level redundacy required or not		Clarification: Both end in the clause refer to 32 Gbps SFP required for existing SAN Switch at GSDC and SFPs (if required) for the supplied servers maintaining a card level redundency. Please refer to corrigendum for make & model of existing SAN switches and storage.
75	25	c) Minimum 1 no. of 1 GbE port and 2 nos of 10 GbE or higher (Fiber/Copper) port.	**Change to**-c) ~~Minimum 1 no. of 1 GbE port and 2~~ 4 nos of 10 GbE or higher (Fiber/Copper) port.	Wider Participant and get better perfomance and redundancy	Clarification: The bidder can quote the product on higher side meeting the requirements to complete the solution.
76	27 and 29	AI Training and Inference Nodes - If not listed on MIcommons, bidder shall be required to submit benchmark report for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details.	Kindly change it to ' If not listed on MIcommons, bidder shall be required to submit benchmark report OR give an undertaking to meet the Benchmark target timings during POC (with upto 25 % tolerance as mentioned in the RFP) for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details.	Kindly allow us to give an undertaking to meet the benchmark target timings during POC (with upto 25 % tolerance as mentioned in the RFP).	Please refer to RFP and time to time published Corrigendum.
77	30	The proposed storage array should be configured with no single point of failure, including required controllers, cache, power supply, cooling fans, etc. It should be scalable up to 12 additional controllers/nodes.	The proposed storage array should be configured with no single point of failure, including required controllers, cache (if applicable), power supply, cooling fans, etc. It should be scalable up to 12 additional controllers/nodes.	Some Storage solutions do not come with cache in the controllers, hence the request for change.	The proposed storage array should be configured with no single point of failure, including required controllers, cache, power supply, cooling fans, etc. It should be scalable up to 12 additional controllers/nodes. **If controller-based cache is unavailable, alternative acceleration mechanisms (e.g., NVMe, distributed caching, tiered memory) should be supported**
78	30	IOPS: minimum 8,00,000	IOPS: minimum 8,00,000 Read	AI Storage Systems require a high Read Performance, hence the suggestion for change.	please refer Corrigendum-03.
79	30	NVMe Storage offered must be certified with the proposed GPU OEM.	NVMe Storage offered must be certified/compatible with the proposed GPU Server OEM.	To ensure wider choice of Storage solutions. Kindly approve.	As per RFP and time to time published corrigendum
80	27	Page 26 - Scalability, Cluster and Management Hardware and software and Page 27 and 29 - Cluster Management & Scheduler and hardware	Page 26 already mentions Scalability, Cluster and Management Hardware and software. Kindly remove the clause of 'Cluster Management & Scheduler and hardware ', for the Training and Inference nodes since the features asked are proprietary.	Having only the requirement of 'Scalability, Cluster and Management Hardware and software', will help in providing a uniform cluster tool.	As per RFP and time to time published corrigendum
81	Clause No. 5 under "Eligibility Conditions,"	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.**		Kindly be noted that the above-men oned clause is a clear viola on of Office Memorandum P-45014/33/2021-BE-II (E-64737) dated 20th December 2022 & P-45021/121/2018-(B.E.-II) dated 20th June 2019 issued by DPIIT, which clearly cites "common examples of restric ve and discriminatory condi ons against the local suppliers" and Sub-clause 'e' of 'Clause 1' in the ANNEXURE-A expressly states that "Excessive past experience requirement, not commensurate with the proven experience expected from Bidder for successful execu on of contract" In additon to the above, we would like to highlight that the above highlighted pre-qualification conditon is deviating from GFR rules. General Financial Rules (GFR), 2017, Clause b i.e. Particular Construction Experience and Key Production Rates of subclause 2 (iii) i.e., Pre-qualification Criteria on Page No.33 and 34 under Chapter 3 of the Manual for Procurement of Works 2022 issued by DOE. The clause b of sub-clause (iii) states that: "The applicant should have:1. successfully completed or substan ally completed similar works during last seven years ending last day of month previous to the one in which applica ons are invited should be either of the following: -1.1 Three similar completed works cos ng not less than the amount equal to 40(forty) percent of the es mated cost; or1.2 Two similar completed works cos ng not less than the amount equal to 50 (fi y) percent of the es mated cost; or1.3 One similar completed work cos ng not less than the amount equal to 80 (eighty) percent of the es mated cost; In view of the above , it is per nent to men on here that clause no. 5 men oned in the Eligibility Condi ons is taking away the opportunity to par cipate from poten al OEMs who have strong experience in deploying GPU Clusters , which will limit the compe on . In view of above we request you to please modify the clause in order to avoid restric ve par cipa on and providing fair opportunity to all.	please refer Corrigendum-03.

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
82	Clause No. 5 under "Eligibility Conditions,"	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.	In light of the above, we respectfully request that Clause No. 5 be reviewed and suitably amended in alignment with the GFR, 2017, to provide a fair and inclusive opportunity for all eligible participants. Furthermore, since the proposed solution involves multiple vendors, we kindly request an extension of the bid submission deadline by at least 7 additional days to accommodate the coordination and compliance requirements effectively.	It may please be noted that Clause No. 5 under "Eligibility Conditions," related to the pre-qualification criteria for participating OEMs, appears to be contrary to the General Financial Rules (GFR), 2017. The clause, under Particular Construction Experience and Key Production Rates, states that the applicant must have: 1. Successfully completed or substantially completed similar works during the last seven years ending on the last day of the month previous to the one in which applications are invited, in either of the following ways: o 1.1 Three similar completed works costing not less than 40% of the estimated cost; or o 1.2 Two similar completed works costing not less than 50% of the estimated cost; or o 1.3 One similar completed work costing not less than 80% of the estimated cost.	please refer Corrigendum-03.
83	1. Eligibility Conditions: (Point no 4.1) Page 16	The bidder should have experience of set up of GPU server base solution with Cumulative 15 nos. of GPUs in last 5 Years in India.	This eligibility point 4.1 contradicts to eligibility point 4.0 in revised RFP (corrigendum 2). Therefore, we request to consider bidders experience of supply & implementation of higher Core Servers as covered in point 4.0 OR remove this clause.	-	please refer Corrigendum-03.
84	8. IMPLEMENTATION TIMELINES & PENALTIES:	Supply of the Hard ware including Li censes and OEM Warranty Certificate. T1=T+60 days from the date of issuance of contract over GEM	Supply of the Hard ware including Li censes and OEM Warranty Certificate. **T1=T+90 days** from the date of issuance of contract over GEM	-	As per RFP and time to time published corrigendum
85		Installation, commissioning & integration of GPU serv ers at GSDC along with HLD , LLD doc uments T2=T1+30	Installation, com missioning & integration of GPU serv ers at GSDC along with HLD , LLD doc uments **T2=T1+60 days**	-	As per RFP and time to time published corrigendum
86		Deployment of required Skilled Resource at GSD T2+7 Days	Deployment of required Skilled Resource at GSD **T2+30 Days**	-	As per RFP and time to time published corrigendum
87		Overall (Sr. no- 2 to 6) Penalty CAP not be more than 10 % of the total GEM order value for IMPLEMENTATION TIMELINES & PENALTIES:	We request for **Maximum Penalty Capping @5% of the total GEM order value** for Implementation Timelines & Penalties.	-	As per RFP and time to time published corrigendum
88	b. SLA for Uptime (99.741%)	Uptime of solution <=99.741% In case of failure of proposed solution and non-maintaining targeted value, 0.5% of Billable Quarterly O&M and Man-power payment for every hourly delay or part thereof pro-portionately in resolution; with max cap of 10 % of GEM order value	We request you to revise this penalty clause as Uptime of solution <=99.741% In case of failure of proposed solution and non-maintaining targeted value, 0.05% of Billable Quarterly O&M and Man-power payment for every hourly delay or part thereof pro-portionately in resolution; with max cap of 10 % of O & M and Manpower cost / Value.	-	As per RFP and time to time published corrigendum
89	24	32 Gbps Host Bus Adaptor with required SFP (at both end) for connecting with existing SAN switch and storage.	**Clarification-** No of 32G FC ports required per server and card level redundancy required or not. Also, Regarding the existing FC port in DC, it is mentioned that 16G SFP are available, however as per RFP you have asked 32GB SFP at both end, kindly confirm if we can use 16GB or it is mandatory to use 32 GB SFP at both sides?		Clarification: Both end in the clause refer to 32 Gbps SFP required for existing SAN Switch at GSDC and SFPs (if required) for the supplied servers maintaining a card level redundency. Please refer to corrigendum for make & model of existing SAN switches and storage.
90	17	SOW, Bidder has to deploy propose solution for inference and AI training model	We are assuming that we need to set-up the required infra at GIL site, but all the AI/ML workloads and use cases will be responsibilities of GIL, bidder has no role to play in it once the underlying infra is ready.		Clarification: Please refer manpower clause of the RFP and time to time published corrigendum.
91	26	Appropriate rated and energy efficient, redundant (N+N) hot swappable power supply **and FAN**	**Change to**-Appropriate rated and energy efficient, redundant (N+1) or better hot swappable power supply	This in line with point 9 of page 18, where it is mentioned N+1 redundant Power supplies. Dense GPU servers typically come with N+1 redundancy. Dense GPU Servers do not come with Hot swappable fans, hence the request for change.	please refer corrigendum-03.
92	27	32 Gbps Host Bus Adaptor with required SFP (at both end) for connecting with existing SAN switch and storage.	**Clarification-** No of 32G FC ports required per server and card level redundancy required or not. Also, Regarding the existing FC port in DC, it is mentioned that 16G SFP are available, however as per RFP you have asked 32GB SFP at both end, kindly confirm if we can use 16GB or it is mandatory to use 32 GB SFP at both sides?		Clarification: Both end in the clause refer to 32 Gbps SFP required for existing SAN Switch at GSDC and SFPs (if required) for the supplied servers maintaining a card level redundency. Please refer to corrigendum for make & model of existing SAN switches and storage.
93	17	The proposed solution should support for sharing of GPU across multiple virtual environments and containers.	Kindly clarify the on below points 1. Which hypervisor will be used for VM-based workloads. 2. do we need to provison hypervisor licesnes or GIL will provide the same. 3. will container and VM co-exist on the same hardware.		Clarifications: The proposed solution should support GPU virtualization, enabling the efficient sharing of GPU resources across multiple virtual machines and containerized environments. It should be compatible with industry-standard hypervisors and container orchestration platforms, supporting vGPU or GPU passthrough mechanisms.The solution should support GPU virtualization, enabling shared GPU resources across multiple VMs and containerized applications instead of dedicating them to a single instance. It should allow simultaneous access for multiple workloads, support vGPU or GPU passthrough for fractional GPU allocation, and be compatible with industry-standard hypervisors and container platforms like VMware vSphere, Microsoft Hyper-V, KVM, Docker, and Kubernetes.
94		New Clause to be incorporated	Server, Storage & Switch OEM must have local service support depot in Gujarat preferably in Gandhinagar/Ahmedabad since last 5 years as on date of RFP release date.	Onsite replacement of faulty hardware and skills support directly from hardware OEM is utmost important in solution led bids wherein uptime and SLA are paramount and OEM skills and support on-site is mandatory and important.	As per RFP and time to time published corrigendum
95		New Clause to be incorporated	All OEMs (Hardware & Software) must be be a company registered in India under the Companies Act 1956, Act 2013	Both Bidder and OEM should be mandatorily registered under Indian Companies Act 1956, Act 2013 for Indian laws to be applicable on these entities and to make them accountable under Indian Judicial Laws.	As per RFP and time to time published corrigendum

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
96		New Clause to be incorporated	Security Features: 1. Immutable Silicon Root of Trust with component integration like network card, GPU etc. 2. Secure Recovery - Ability to rollback firmware 3. FIPS 140-2 validation 4. One-button Secure Erase 5. Common Criteria certification 6. Advanced Encryption Standard (AES) and Triple Data Encryption Standard (3DES) on browser 7. Support for Commercial National Security Algorithms (CNSA) 8. Secure Configuration Lock	Some basic Security Features are defined in the RFP, however, these addl. Parameters are more relevant and critical for system security.	As per RFP and time to time published corrigendum
97		New Clause to be incorporated	Security Features: 1. Trusted Platform Module 2.0 2. Secure Firmware 3. Detect and recover for BIOS tamper-free updates 4. Secure Recovery - Ability to rollback firmware 5. ACPI 6.3 Compliant 6. UEFI 2.8 7. SMBIOS 3.4 or later 8. Malicious Code Free design" (to be certified by OEM)	No Security Features are defined for AI Nodes in the RFP, hence these addl. parameters are must required and critical for system security.	As per RFP and time to time published corrigendum
98		New Clause to be incorporated	Security Features: 1. Immutable Silicon Root of Trust with component integration like network card, GPU etc. 2. Secure Recovery - Ability to rollback firmware 3. FIPS 140-2 validation 4. One-button Secure Erase 5. Common Criteria certification 6. Advanced Encryption Standard (AES) and Triple Data Encryption Standard (3DES) on browser 7. Support for Commercial National Security Algorithms (CNSA) 8. Secure Configuration Lock	Some basic Security Features are defined in the RFP, however, these addl. Parameters are more relevant and critical for system security.	As per RFP and time to time published corrigendum
99		Latest Generation Intel® Xeon® platinum or AMD Epyc scalable processors with Minimum Dual 32-Core, with Min 2 X GPU Accelerator.	Latest Generation Intel® Xeon® platinum or AMD Epyc scalable processors with Minimum Dual 32-Core, with Min 2 X GPU Accelerator. Total Qty of Inference server must contain atleast 24 x Accelerators.	Requirement mentions min. 2 x Accelerators per Node. Please Clarify, can we Offer more Accelerators per Node. Presently the RFP clause is contradictory as it mentions min. 2 x accelerators but the Inference Node Qty is Fixed. This will be disadvantage for vendor offering more than 2 x Accelerators per Inference Node. Request you to mention total accelerators required for the Inference Solution.	As per RFP and time to time published corrigendum. However, Bidder may propose 2 or higher accelerator card per node (In inference Node) and numbers of inference node quantity may be change accordingly. However, total requirement of core and memory may be provided a s mentioned in RFP and time to time published corrigendum.
100	Minimum Technical Specifications Storage Nodes	External Storage ***In case PFS (Parallel File System) is asked*** IOPS: minimum 8,00,000 NVMe Storage offered must be certified with the proposed GPU OEM	IOPS: minimum 8,00,000 NVMe Storage offered must be either certified with the proposed GPU OEM or must be self-certified by Storage OEM for proposed GPU.	Self-certification of PFS by Storage OEM for proposed GPU will suffice the interoperability and performance requirements of GIL.	please refer Corrigendum-03.
101	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** IOPS: minimum 8,00,000	IOPS: minimum 8,00,000 READ	There are no references to IOPS in NVIDIA SUPERPOD documents and AI NFS Storage will have very high number of READ IOPS, hence IOPS requirement needs to be READ IOPS.	please refer Corrigendum-03.
102	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** 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.	The proposed storage array should be configured with no single point of failure, including controllers **(at least 2 controllers per disk tier)**, cache, power supply, cooling fans, etc. It should be scalable up to 12 additional controllers / nodes.	2 Controllers are industry standard, please keep it to 2.	As per RFP and time to time published corrigendum
103	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** Performance: 20GBps Read/Write from day one and scalable up to 60GBps with a scale-out architecture and additional controllers/nodes in the future.	Performance: 40GBps 100% Read from day one and scalable up to 80GBps 100% Read with a scale-out architecture and additional controllers/nodes in the future. Data Availability: 99.9999% data availability guarantee on proposed Storage model duly certified by Storage OEM.	Performance in NFS Storage is better measured on Read throughput and also these numbers are available with all Storage OEMs. Data Availability Guarantee of Six-Nines (99.9999%) is practically must for this critical infrastructure.	please refer Corrigendum-03.
104	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** New Clause to be incorporated	Vendor shall ensure that concurrent failure of at-least 4 disks can be handled without any kind of downtime and vendor shall configure the erasure code accordingly	Required feature for giving different better resilance and performance	As per RFP and time to time published corrigendum
105	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** New Clause to be incorporated	1. Offered storage system shall be able to create native immutable snapshots for offered solution. 2. Offered storage system shall have capability for creating the immutable snapshot copies at both primary and DR location through replication engine shall provide the flexibility for having different retention period for each location. . 3. After defining the expiration period, it shall not be possible to reduce the expiration time. However, if the business need arise, then expiration period shall be shortened only through Dual authorization and through different set of authorized users only.	Required feature for giving better prtoction to storage	As per RFP and time to time published corrigendum
106	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** New Clause to be incorporated	1. Each offered file services front-end controller shall have minimum of 256GB memory and minimum 32 number of CPU cores. 2. Each front-end controller shall also be offered with 2 x 100Gbps Ethernet Front-end ports and shall also have 2 x 100Gbps backend ports for disk connectivity. 3. Every front-end controller shall have dual physical CPUs.	Minimum hardware to be proposed so that everyone proposes enough resources for performance.	As per RFP and time to time published corrigendum
107	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** New Clause to be incorporated	1. Offered Storage platform shall support NFS nconnect feature for increasing the NFS performance. It shall allow at-least 16 x TCP connections between each client and storage platform. 2. Offered storage platform shall support NFS over RDMA and multi-pathing feature for increasing the NFS performance while connecting the client to storage system. 3. Multi-pathing shall be able to work in conjunction with nconnect and NFS over RDMA. 4. Offered Storage platform shall also support byte range file locking for both NFS V 3.x and NFS 4.1.	Some advanced NFS features.	As per RFP and time to time published corrigendum
108	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** New Clause to be incorporated	1. Offered storage system shall provide the functionality of disaster recovery by replicating the required path or directory to DR or peer location. 2. Offered storage system shall ensure that data path between Primary and DR location is encrypted. Vendor shall offer required Software / License or hardware for achieving the required functionality. 3. Offered storage system shall support one to many and many to one replication so that one site can replicate to more than 1 DR site or replication peer and multiple Primary sites can replicate to single DR location.	Disaster Recovery capabilities	As per RFP and time to time published corrigendum

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
109	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** New Clause to be incorporated	1. Offered storage system replication analytics engine shall have capability to showcase the overall RPO between sites and shall also showcase the RPO miss report so that required bandwidth, if required, planning can be done. 2. Offered storage system replication shall also show the overall data transfer over the replication connection so that, if required, bandwidth planning can be done for achieving the required RPO.	Disaster Recovery capabilities	As per RFP and time to time published corrigendum
110	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** New Clause to be incorporated	1. Offered storage system shall support Quality of service natively and shall be able to assign IOPS and required bandwidth for a given protected path / share. 2. It shall be possible to enable the quality of services within each provisioned tenant. 3. Offered quality of service functionality shall work in conjunction with quota management so that it can also be assigned for used capacity and provisioned capacity.	QOS features	As per RFP and time to time published corrigendum
111	Minimum Technical Specifications Storage Nodes	External Storage ***In case NFS based Storage is asked*** New Clause to be incorporated	1. Offered storage platform shall support and enabled with 256-bit-AES-XTS encryption and shall support both internal and external key management 2. Encryption object module shall be FIPS 140-2 validated. 3. Offered storage system shall also be supplied with FIPS enabled drives.	Encryption Features	As per RFP and time to time published corrigendum
112		Hypervisor	Hypervisor Integration with AI Solutions • The Hypervisor should support integration with AI solutions to enable seamless building, deployment, and management of AI workloads. • It should leverage both virtualization and containerization technologies and be officially supported and certified by the AI solution provider. Hypervisor Management Tools • The proposed solution should include robust hypervisor management tools that simplify the deployment, scaling, and operational management of AI workloads, thereby reducing operational complexity. GPU Acceleration Support • The hypervisor should support GPU acceleration to allow efficient utilization of GPU resources during AI training and inference. • This will ensure high performance and scalability of AI applications. Comprehensive GPU Reporting Capabilities • The solution should have the capability to generate comprehensive reports for GPU usage, performance metrics, compliance status, health monitoring, forecasting, and capacity planning across AI workloads. High Availability (HA) Support • The system should support High Availability for VM migration in case of a physical server failure. All virtual machines on the failed server should be capable of migrating automatically to another physical server running the same virtualization software. • Additionally, the solution should support HA for VMs utilizing passthrough PCIe devices or NVIDIA/other vendor vGPUs.		As per RFP and time to time published corrigendur
113		Hypervisor	1. The Hypervisor should support integrating with AI solutions to help build, deploy and manage AI workloads, leveraging the benefits of Virtualization and containerization. The Hypervisor should be supported and certified by AI Solution. 2. Solution should include hypervisor management tools, simplify the deployment, management and scaling of AI workloads, reducing operation complexity. 3. Hypervisor should support GPU acceleration, enabling efficient utilization of GPU resources for AI training and inference, ensuring high performance and scalability. 4. The solution should provide capability of generating reports for GPU usage, performance, compliance, health, forecasting, capacity, across AI workload. 5. Should support HA for migration of VMs in case one server fails all the Virtual machines running on that server shall be able to migrate to another physical server running same virtualization software. Should support HA for VMs with a passthrough PCIe device or a NVIDIA / other vGPUs.		As per RFP and time to time published corrigendum
114	Eligibility Conditions	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.	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 totalvalue of INR 20 Cr. or Total of 50 Cr from 2 Customers. Note: Similar works means SITC of Project which includes GPU ACCELERATED with multiple GPU Node.	This is required for wider participation and considering the budgeting of this project.	As per RFP and time to time published corrigendum
115	Master Node 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	Hardware RAID 0,1, 5, 6, 10, 50, 60 with 4GB cache Flash based cache protection module should be included, should support Gen 4/5.0 PCIe.	Every OEM has different architecture so please revise as requested for wider OEM participation as it is restricting us.	As per RFP and time to time published corrigendum.
116	Master Node Network	Following N/W are required: a) Infiniband / Ethernet (200GBps or higher) as required for quoted storage delivery to nodes b) Ethernet (100Gbps 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	Following N/W are required: a) Ethernet (200GBps or higher) as required for quoted storage delivery to nodes b) Ethernet (100Gbps 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/ oob management	Ethernet connectivity provides better throughput and performance , also ethernet is widely used and available with all leading OEMs while Infiniband is less used and available with some specific OEMs only so please revise accordingly. 1G connectivity is mostly required for oob connectivity and inband is used case for telemetry data so please revise for wider OEM participation	Clarification: Ethernet connectivity is already allowed.

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
117	Master Node Server Management	Dedicated IPMI 2.0 compliant management LAN port having support for system health monitoring, event log access, Virtual media over net- work, and Virtual KVM (KVM over IP). All required licenses to use IPMI features should be included. Licenses shall be perpetual/subscription base for entire contract period to use.	Dedicated IPMI 2.0 compliant management LAN port having support for system health monitoring, event log access, Virtual media over net- work, and Virtual KVM (KVM over IP). All required licenses to use IPMI features should be included. Licenses shall be perpetual/subscription base for entire contract period to use. Server management software should also provide below capabilities: •The management tool should be able to provide global resource pooling and policy management to enable policy-based automation and capacity planning with Zero-touch repository manager and self-updating firmware system, Automated hardware configuration and Operating System deployment to multiple servers * Virtual IO management / stateless computing and Server management software should provide capaibility to view health, inventory for third-party compute, network, storage, integrated systems, virtualization, and containers. * The management software should participate in server provisioning, device discovery, inventory, diagnostics, monitoring, fault detection, auditing, and statistics collection and should provide an alert in case the system is not part of OEM Hardware Compatibility list & should provide anti counterfeit. *The proposed management solution should provide proactive security & software advisory alerts and should outline the fixes required to address the issues and analyze current configurations & identify potential issues due to driver & firmware incompatibility * The proposed solution should have customizable dashboard to show overall faults / health / inventory for all managed infrastructure. With option to create unique dashboards for individual users. The user should have flexibility to select names for dashboards and widgets (ex:- health, utilization etc.)	Servers (specially like master nodes) play critical role in the Data centers as multiple applications dependo on it and external storage also connects to it due to this end to end server management sofwtare with mentioned features would be needed so please revise the clause acordingly	As per RFP and time to time published corrigendum.
118	Security Features	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).	min ACPI 6.2 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).	Every server architecture on different compliances get certified jointly by OEM and third parties due to that please revise with mentioned so that every OEM can participate	As per RFP and time to time published corrigendum
119	AI Training 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.	Min Dual 60 core latest Gen Intel® Xeon® platinum (5th Gen or higher) or AMD Epyc (Turin or higher) 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.	Defining processors's generation will standardize the type of processors offered by every OEM for equal participation also considering Training nodes require high compute performance so cores needs to be udpated as per processor's generation.So, please revise accordingly.	As per RFP and time to time published corrigendum. However, Bidder can quote higher side compute.
120	AI Training Network	a)Minimum 8 nos of InfiniBand NDR ports or Ethernet (400Gb/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)Required InfiniBand / 200G or higher Ethernet 2 x twin-port HCA as required for quoted storage delivery to node. e)Additionally, 1 nos of 100GbE or higher Ethernet (Fibre). f)Required switch with 64 non-blocking ports with aggregate data throughput up to 51.2 Tb/s and required compatible cables of appropriate length to connect all 8 (compute communication) nos. of IB NDR ports / Ethernet of all nodes in non-blocking mode.	a)Minimum 8 nos of SuperNIC's with minimum 8 arm cores capable of Supporting Infiband / Ethernet (400Gb/s or higher) for compute communication for internode communication b)1 nos. of port for BMC (dedicated LAN port), c)Minimum 2 nos of 10 GbE or higher (Fiber/Copper) port. d)Required 200G or higher DPU's with 2 x twin-port 200G as required for quoted storage delivery to node e)Required switch with 64 non-blocking ports with aggregate data throughput up to 51.2 Tb/s and required compatible cables of appropriate length to connect all 8 (compute communication) nos. of IB NDR ports / Ethernet of all nodes in non-blocking mode. The switch should provide RoCE v2 / equivalent, PFC, ECN, Telemetry capabilities to run the setup	GPU training nodes in AI architecture requires SuperNICs for east-west communication across the nodes and DPUs are reqired for North-South communication. Also RoCEv2 is very important & must needed critical protocol used for GPU to GPU communication across network switches which along with PFC, ECN make sures to provide lossless , low latency , high bandwidth communication, So please update this point accordingly and allow wider OEM participation	As per RFP and time to time published corrigendum
121	AI Training Internal Storage	• For Operating System: minimum 1.92 TB NVMe drives • For Data: Minimum 8 * 3.84 TB U.2 or EDSFF NVMe drives	• For Operating System: minimum 2x 960 GB M.2 SSD Boot drives • For Data: Minimum 8 * 3.84 TB U.2 Gen5 NVMe drives	Every OEM's architecture as per their testing and availability of drives is different so please update this clause for us to participate.	please refer Corrigendum-03.
122	AI Training Power requirements	Appropriate rated and energy efficient, redundant (N+N) hot swappable power supply and FAN	Appropriate rated and energy efficient,redundant (N+1) hot swappable power supply and FANs. In case of power failure, the system should be able to sustain 3 power supplies failure with GPU throttling no less than 60%	High availabilty and tolerance to powersupply is very important for crtical server having large datasets dependent on them for training so please update the clause for allowing us to participate	please refer corrigendum-03.
123	AI Training System Network	Following N/W are required: 1.NDR Infiniband / Ethernet (400GBps or higher) for compute communication 2.Infiniband /Ethernet (200GBps 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	Following N/W are required: 1. SuperNIC's with(400GBps or higher) for compute communication 2. DPU's (2x 200Gbps or higher) for storage delivery 3.DPU's (2x200 Gbps) for cluster orchestration 4.Ethernet (10 GbE or higher) for perimeter connectivity 5.Ethernet (1GbE) for in/oob-band management	GPU training nodes in AI architecture requires SuperNICs for east-west communication across the nodes and DPUs are reqired for North-South communication and also for storage communication. So , please revise the point to standardize the AI computing node connectivity and communication architecture and allowing every OEM to participate equally	As per RFP and time to time published corrigendum
124	AI Training 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.	Rack Servers should be certified by GPU Controller / Accelerator OEM, the Certificate or listing of offered Server model in GPU Controller / Accelerator OEM website or undertaking to be published by time of award must be submitted along with bid.	please generalise this clause for wider OEM participation as it is restrcting us to participate	As per RFP and time to time publish Corrigendum.
125	AI Training Benchmarks	Bidder needs to submit proof of the quoted GPU meeting these MIcommons training benchmarks at the time of bidding or If not listed on MIcommons, bidder shall be required to submit benchmark report for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details.	Bidder needs to submit proof of the quoted GPU meeting these MIcommons training benchmarks at the time of bidding or If not listed on MIcommons, bidder shall be required to submit GPU Accelerator's test results for the Make/Model (same GPU's) of the quoted server as part of bid submission.	Since AI training models are quite new to the market and their benchmarks are getting updated during course of time so please generalise this clause for wider OEM participation as it is restrcting us to participate	Please refer to RFP and time to time published Corrigendum.
126		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. Or If not listed on MIcommons, bidder shall be required to submit benchmark report for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details. 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 Up to 25% tolerance shall be accepted on aforementioned benchmarks during POC.	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. Or If not listed on MIcommons, bidder shall be required to submit undertaking to be provided that quoted systems Make/Model is certified by the GPU accelerator OEM The submission should be on OEM letterhead duly signed and referring the bidder and bid details. 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 Up to 25% tolerance shall be accepted on aforementioned benchmarks during POC	Since AI training models are quite new to the market and their benchmarks are getting updated during course of time so please generalise this clause for wider OEM participation as it is restrcting us to participate	Please refer to RFP and time to time published Corrigendum.

Responses to pre-bid queries and Revised RFP (Corrigendum-3 dated 16.05.2025)
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)

Sr#	Reference (Clause /page)	Points of Clarification Requested	Suggestion/Clarification	Justification	Responses to pre-bid queries
127	Inference Node Number of GPUs and GPU Communication	2 x Accelerators per node, each with minimum 140GB or higher GPU per Accelerator.Should support Tensor core/Matrix core, CUDA / Stream Processors/ openCL /ROCm with Accelerators.	2 x Accelerators per node, each with minimum 90 GB or higher GPU per Accelerator. Should support Tensor core/Matrix core, CUDA / Stream Processors/ openCL /ROCm with Accelerators.	Type of GPUs are defined as per the different models/architecture of different Server OEMs so equivalent type of server having respective GPU needs to be updated to allow all OEMs to participate so request you to please as it is restrcting to participate	As per RFP and time to time published corrigendum. However, Bidder may propose 2 or higher accelerator card per node (In inference Node) and numbers of inference node quantity may be change accordingly. However, total requirement of core and memory may be provided a s mentioned in RFP and time to time published corrigendum.
128	Inference Node Internal Storage	For Operating System: minimum 2*1.92 TB ~~M.2~~ NVMe drives Minimum 4 * 3.84 TB U.2 or EDSFF NVMe drives	For Operating System: minimum 2*1.92 TB NVMe drives Minimum 4 * 3.84 TB U.2 /U.3 or EDSFF NVMe drives	Request you to please update the clause for wider OEM participation as different server OEMs has different types of drives certified and tested.	please refer Corrigendum-03.
129	Inference Node Security Features	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)	Min ACPI 6.2 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).	Every server architecture on different compliances get certified jointly by OEM and third parties due to that please revise with mentioned so that every OEM can participate	As per RFP and time to time published corrigendum
130	Inference Node 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	4 x PCIe Gen 5.0 x 16 FH FL Slots. All slots must operate at PCI Gen 4.0/5.0 speed when fully populated	Every OEM has different architecture so please revise as requested for wider OEM participation as it is restricting us.	As per RFP and time to time published corrigendum
131	Inference Node Mother Board	Appropriate Motherboard and chipset. Must support PCIe Gen 5.0 and compatible with selected processors and GPUs.	Appropriate Motherboard and chipset. Must support PCIe Gen 4.0/5.0 and compatible with selected processors and GPUs.	Every OEM has different architecture so please revise as requested for wider OEM participation as it is restricting us.	As per RFP and time to time published corrigendum
132	Inference Node Networking Switch	1.Min. Two or required 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.Min. One or required 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.Min. One or required 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.	= Please clarify this switch's usage as if this switch is required for master node connectivity also then it should have 400G switch ports also in it	need clarification	As per RFP and time to time published corrigendum
133	Bidder needs to submit proof of the quoted GPUs being listed in MLperf inferencing benchmarks at the time of bid submission. Or If not listed on MLperf, bidder shall be required to submit benchmark report for the Make/Model(same configuration) of the quoted should be on OEM letterhead duly signed and referring the bidder and bid details.	Image segmentation (medical) 3D-Unet-99 Throughput for single NODE inference (99% offline) = 09 Samples/s or higher NLP Bert-99 Throughput for single NODE inference (99% offline) = 10000 Samples/s or higher Recommendation dlrm-v2-99 Throughput for single NODE inference (99% offline) = 85000 Samples/s or higher LLM Summarization gptj-99 Throughput for single NODE inference (99% offline) = 2500 Tokens/s or higher Image Classification ResNet Throughput for single NODE inference (99% offline) = 105000 Samples /s or higher Object Detection RetinaNet Throughput for single NODE inference (99% offline) = 1500 Samples /s or higher Image Generation Stable Diffusion-XL Throughput for single NODE inference (99% offline) = 1 Samples/s or higher Up to 25% tolerance shall be accepted on aforementioned benchmarks during POC.	Bidder needs to submit proof of the quoted GPUs being listed in MLperf inferencing benchmarks at the time of bid submission. Or If not listed on MLperf, bidder shall be required to submit GPU Accelerator's test results for the Make/ Model (same GPU's) of the quoted server as part of bid submission.	Since AI models are quite new to the market and their benchmarks are getting updated during course of time so please generalise this clause for wider OEM participation as it is restrcting us to participate	Please refer to RFP and time to time published Corrigendum.
134	Nodes for AI training: Total Qty-4 sets. Page no.25	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	Kindly help to amend the clause as "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 FP32 / TF32 Performance, 31 PetaFlops or Higher FP8 performance with sparsity"	Suggested changes helps to participate in the bid process and this would give the department to make the bid more competitive and also help to participate respective other OEM's in the bid process.	As per RFP and time to time published corrigendum.

Revised RFP (Corrigendum-03)

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. 25 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. In case a Make in India OEM participates directly in this bid as the sole bidder, the OEM-specific eligibility criteria shall not be applicable. In such cases, only the bidder ' s turnover, relevant experience, and past performance shall be evaluated for qualification for Make in India OEM. The Make in India OEM must, however, comply with all technical and commercial requirements as specified in the RFP.	● Audited profit and loss statement and balance sheet ● Auditor certificate clearly specifying the turnover and positive net worth.
3.1	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 in setting up GPU or CPU core that meet the following criteria from any central or state Government / PSU/ Listed Company/BFSI sector in India or Global experience in Tier-04 Datacentre within the last five years as of the bid submission deadline. ☐ One project having total value of INR 40 Cr (Should have min 45 GPU / 800 core) or ☐ Two project having total value of INR 25 Cr (Should have min 28 GPU / 500 core) or ☐ Three project having total value of INR 20 Cr (Should have min 22 GPU / 400 core) Note: GPU Experience means GPU Installation in multiple server Nodes. CPU core experience means cores installed in Server Nodes.	Copy of Work Order along with Completion / Go-Live certificate. For Global experience bidder has to submit client's tier-04 datacenter certificate along with client official mail id, which should be reflect on client official website for due verification. Copy of Work Order along with Completion / Go-Live certificate.
~~**4.1**~~	~~The bidder should have experience of set up of GPU server base solution with Cumulative 15~~	~~Copy of Work Order along with Completion / Go-Live certificate.~~

	~~nos. of GPUs in last 5 Years in India.~~	
5.	The OEM should have executed similar GPU setup for min 3 clients from any central or state Government / PSU/ Listed Company/BFSI sector in India or Global experience in Tier-04 Datacentre in last 5 Years 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. In case a Make in India OEM participates directly in this bid as the sole bidder, the OEM-specific eligibility criteria shall not be applicable. In such cases, only the bidder ' s turnover, relevant experience, and past performance shall be evaluated for qualification for Make in India OEM. The Make in India OEM must, however, comply with all technical and commercial requirements as specified in the RFP.	Copy of Work Order along with Completion / Go-Live certificate. For Global experience bidder has to submit client's tier-04 datacenter certificate along with client official mail id, which should be reflect on client official website for due verification. 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. Bidder has to submit OEM MAF for Proposed hardware, Software and license part.
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.
3. Bid Evaluation Method – Lowest Price L1 based on Technical evaluation and / or PoC (Proof of Concept) Testing.
4. Bidder has to submit end-to-end OEM make and model details of proposed solution at the time of bid submission for further evaluation.

1.1 Criteria for Bid Evaluation

A three-stage procedure will be adopted for evaluation of proposals as follows:

1) Pre- Qualification or Eligibility Condition
2) PoC (Proof of Concept) Testing of benchmark mentioned in this RFP for who comply in Pre- Qualification or Eligibility Criteria
3) Financial bid opening for who comply in PoC (of Bench mark test mentioned in this RFP)

1.2 Technical Presentation cum Proof-of-Concept (PoC)

Technical Bids of only those Bidder / OEMs who meet the "Pre-Qualification or Eligibility Condition" criteria shall be considered for further Technical Evaluation.
The bidders who qualify for the pre-qualification shall be invited to demonstrate their proposed solution through POC to the Authorities of Tender Evaluation Panel (and potentially other representatives of the Authorities) and GSDC. The Demo will be conducted alphabetically of Bidder name. Timeline for POC as following;

Sr. NO.	POC Timeline	Work Type	Remarks
1	T0	Date of intimation to bidder for Delivery of minimum proposed solution for POC to pre-qualified/eligible bidders	-
2	T1=T0+20 Days	Delivery of minimum proposed solution to achieve mentioned Benchmark (in RFP) by each Bidder (In case of Complete end to end solution from the same OEM of multiple bidders, OEM wise POC will be conducted).	If the OEM / bidder fails to deliver proposed minimum solution within 15 days, they will be considered disqualified.
3	T2	Intimation for PoC demonstration schedule to bidders	-
3	T3=T2+20 Days	Installation, commissioning and completion of POC testing of minimum proposed solution to achieve mentioned benchmark (in RFP)	If the PoC is not completed within 20 days from the POC request date, the Bidders/OEM will be disqualified.

- If multiple bidders have proposed the same complete End to End OEM solution, the result of that PoC shall apply to all such bidders.
- In the event proposed solution of bidder fails in the PoC, **bidders quoting this solution** will be considered disqualified.
- Bidders must arrange and deploy all necessary infrastructure and accessories, including hardware, software, cables, racks, operating systems, and any other items required to conduct the PoC.
- GSDC shall provide only space, power, and cooling. All other logistical and PoC execution-related costs shall be borne entirely by the bidder/OEM.
- Upon completion of the PoC, the bidder must lift all deployed equipment from GSDC premises at their own cost.
- Bidders must demonstrate the required benchmarks as specified in the RFP during the PoC.
- Only those bidders who successfully qualify the PoC will be eligible for financial bid opening.
- The Tenderer reserves the right to accept or reject any or all bids in case it is not satisfied with the outcome of the PoC testing & benchmark required.
- The Tenderer reserves the right to accept or reject any or all bids or to re-tender at the Tenderer's sole discretion without assigning any reasons to anybody whatsoever.

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**	**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 proposed solution should support for sharing of GPU across multiple virtual environments and containers. Required license should be available from day one. Bidder to ensure premium level or highest level of OEM support to meet SLA for all OEM provided software and libraries.
6. MLops practices and principles should be followed under training model. If required, Bidder can use appropriate tool for the same without any additional cost to tenderer.
7. The bidder shall submit the detailed documentation on the implementation and deployment.
8. The solution should support remote console access as per GSDC policy to all the servers for cluster server's health monitoring at Fast Ethernet or better access speed.
9. 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.
10. 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 **GPU** 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.
11. Onsite comprehensive annual maintenance with warranty and OEM support for 5 Years from the date of completion of Functional 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.
12. 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 FAT and further Operation.
13. All the items as required under this RFP should be delivered in a single lot.
14. 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.

15. The bidder shall replace the faulty hard disk at no cost, the department will not returned the faulty disk after replacement of new disk.
16. The bidder should provide entire support of the required solution asked in the RFP and back-to-back support from the OEM.
17. 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 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. Lack of Information to Bidder:

- 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. Payment Terms:

1. No advance payment will be made to the bidder.
2. 70% of the Capex cost shall be paid within 30 days after Supply of the proposed product, software including Licenses mentioned in RFP of complete solution.
3. 20% of the Capex cost shall be paid within 30 days after Successful FAT (Final Acceptance Test) of complete solution duly certified by the GSDC and counter-signed/approved by the authority.
4. 10% of the Capex cost shall be paid after due acceptance by the GSDC and Go-live.
5. Cost of **O&M support and Manpower Cost (D mentioned in financial breakup)** 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, POC shall be executed.~~

- ~~Successful bidder has to complete the SITC of proposed complete solution and execute POC to meet the benchmark as mentioned in this RFP document. All cost with respect to execute the POC shall be borne by successful bidder.~~
- ~~If POC does not meet the benchmark, bidder shall lift deployed complete solution from GSDC without any cost to Tender. No payment will be made on failure of POC.~~
- ~~After Successful POC which is successful demonstration of the benchmark as mentioned in RFP, only then the bidder shall go for Final Acceptance Test~~.
- After successful installation of the System in accordance with the requirements as mentioned in Schedule of Requirement, Final Acceptance Test will be conducted. The GSDC or designated agency shall through review all aspects of the solution as per the ask of the RFP. 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 and Go-Live of the System.
- Any delay by the successful bidder in the POC or 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 as 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 **GPU** 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 Warranty Certificate.	T1=T+60 days from the date of issuance of contract over GEM	0.5% of Capex value of delayed/pending work per week or part thereof	10% of GEM order value	Overall (Sr. no- 2 to 6) Penalty CAP not be 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 ,	T2=T1+30	0.5% of Capex value of delayed/pending work per week or part thereof	10% of GEM order value	

	LLD documents				
4	~~POC to meet the benchmark as mentioned in this RFP document.~~	~~T3=T2+30 days~~	~~0.1 % of Capex value of delayed/pending work per week or part thereof.~~ ~~**In case of delay for more than 2(two) weeks after the defined milestone, the POC shall be treated as failed and the contract shall be terminated and PBG may be forfeited.**~~	~~10% of GEM order value.~~	
5	Final Acceptance Testing (FAT)	T3=T2+15 days	0.5% of Capex value of delayed/pending work per week or part thereof.	10% of GEM order value.	
6	Deployment of required Skilled Resource at GSDC	T3+7 Days	Rs. 10000/- day.	Rs. 250000/-	
7	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 NBD (Next Business Day) 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 hour for Rs. 5000 / hour or part thereof proportionately, which will be recovered against Performance bank guarantee or billable quarterly invoice amount submitted by the successful bidder.
- 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 Billable Quarterly O&M and Manpower payment for every hourly delay or part thereof proportionately in resolution; with max cap of 10 % of GEM order value.

- SLA will be calculated on quarterly basis, However, Final penalty deduction on the quarterly payment i.e., (4* 3 quarter SLA report penalty will be applied during O&M and Manpower quarterly payment.)
- Bidder has to ensure support 365*24*7 for SLA calculation.

c. Manpower related SLA and Penalties:

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 For Data: Minimum 4*3.84 TB capacity interface hot swap Enterprise NVMe SSDs.
HBA Card	32 Gbps Host Bus Adaptor with required SFP (at both end) for connecting with existing SAN switch and storage.
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 (200Gbps or higher) as required for quoted storage delivery to nodes b) Ethernet (100Gbps 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/subscription base for entire contract period to use.
Power Supply	Appropriate rated and energy efficient, redundant (N+1) hot swappable power supply (Mandatory) and Fan (Optional) .
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 / RHEL AI / Red Hat OpenShift AI server. However, the bidder shall deliver Ubuntu Linux version 22 or higher which should be delivered with Enterprise support from OEM with premium or highest level of support available. This must be provided with enterprise-level support from the OEM, specifically with premium or the highest level of support available.
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.
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. Or If not listed on spec.org, bidder shall be required to submit benchmark report / logs for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details.

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 (400Gb/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) Required InfiniBand / 200G or higher Ethernet 2 x twin-port HCA as required for quoted storage delivery to node. e) Additionally, 1 nos of 100GbE or higher Ethernet (Fibre). f) Required switch with 64 non-blocking ports with aggregate data throughput up to 51.2 Tb/s and required compatible cables of appropriate length to connect all 8 (compute communication) nos. of IB NDR ports / Ethernet of all nodes in non-blocking mode.
Internal Storage	• For Operating System: Minimum capacity of 2*960 GB NVMe /M.2 NVMe drives • For Data: Minimum 8 * 3.84 TB U.2 / U.3 or EDSFF 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	Appropriate rated and energy efficient, redundant (N+1) hot swappable power supply (Mandatory) and Fan (Optional) .
System Network	Following N/W are required: 1. NDR Infiniband / Ethernet (400Gbps or higher) for compute communication

	2. Infiniband /Ethernet (200Gbps 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 / RHEL AI / Red Hat OpenShift AI server. However, the bidder shall deliver Ubuntu Linux version 22 or higher which should be delivered with Enterprise support from OEM with premium or highest level of support available.
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/subscription base for entire contract period 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, Tensor RT, 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) Micro services 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 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 or If not listed on Mlcommons, bidder shall be required to submit benchmark report for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details.
	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. Or If not listed on Mlcommons, bidder shall be required to submit benchmark report for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details. 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 Up to 25% tolerance shall be accepted on aforementioned benchmarks during POC.

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.
Internal Storage	For Operating System: Minimum capacity of 2*960 GB NVMe /M.2 NVMe drives Minimum 4 * 3.84 TB U.2 / U.3 or EDSFF NVMe drives

HBA Card	32 Gbps Host Bus Adaptor with required SFP (at both end) for connecting with existing SAN switch and 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	Appropriate rated and energy efficient, redundant (N+1) hot swappable power supply (Mandatory) and Fan (Optional) .
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 (200Gbps or higher) as required for quoted storage delivery to nodes 5. Minimum 1 x 100 GbE Ethernet ports for User Network 6. 1 nos. of port for BMC (dedicated LAN Port)
Networking Switch	1. Min. Two or required Nos. of Switch with **48 *10/25G or higher** SFP+ and **6 or higher 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. Min. One or required Nos. of Switch with 48 *1G RJ45, 4* 10/25G or higher SFP28 or higher 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. Min. two or required Nos. of Switch with 32 * 100GbE QSFP ports or One or required 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. Note: The bidder must deploy the required quantity of switches with the same or higher functionality to meet the solution requirements. The switch count shall be adjusted (increased/decreased) based on the actual port availability per device while maintaining the specified speed and functionality.
OS Support	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux / RHEL AI / Red Hat OpenShift AI server. However, the bidder shall deliver Ubuntu Linux version 22 or higher which should be delivered with Enterprise support from OEM with premium or highest level of support available.

	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	Undertaking from Sever OEM for compatibility of the proposed sever with GPU under the quoted Inference node must be submitted (duly signed by authorized signatory , mentioning Bid reference)
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 bid submission. Or If not listed on MLperf,	Image segmentation (medical) 3D-Unet-99 Throughput for single NODE inference (99% offline) = 09 Samples/s or higher NLP Bert-99 Throughput for single NODE inference (99% offline) = 10000 Samples/s or higher Recommendation dlrm-v2-99 Throughput for single NODE inference (99% offline) = 85000 Samples/s or higher

bidder shall be required to submit benchmark report for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details.	LLM Summarization gptj-99 Throughput for single NODE inference (99% offline) = 2500 Tokens/s or higher Image Classification ResNet Throughput for single NODE inference (99% offline) = 105000 Samples /s or higher Object Detection RetinaNet Throughput for single NODE inference (99% offline) = 1500 Samples /s or higher Image Generation Stable Diffusion-XL Throughput for single NODE inference (99% offline) = 1 Samples/s or higher Up to 25% tolerance shall be accepted on aforementioned benchmarks during POC.
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. If not listed on spec.org, bidder shall be required to submit benchmark report / logs for the Make/Model (same configuration) of the quoted server as part of bid submission. The submission should be on OEM letterhead duly signed and referring the bidder and bid details.

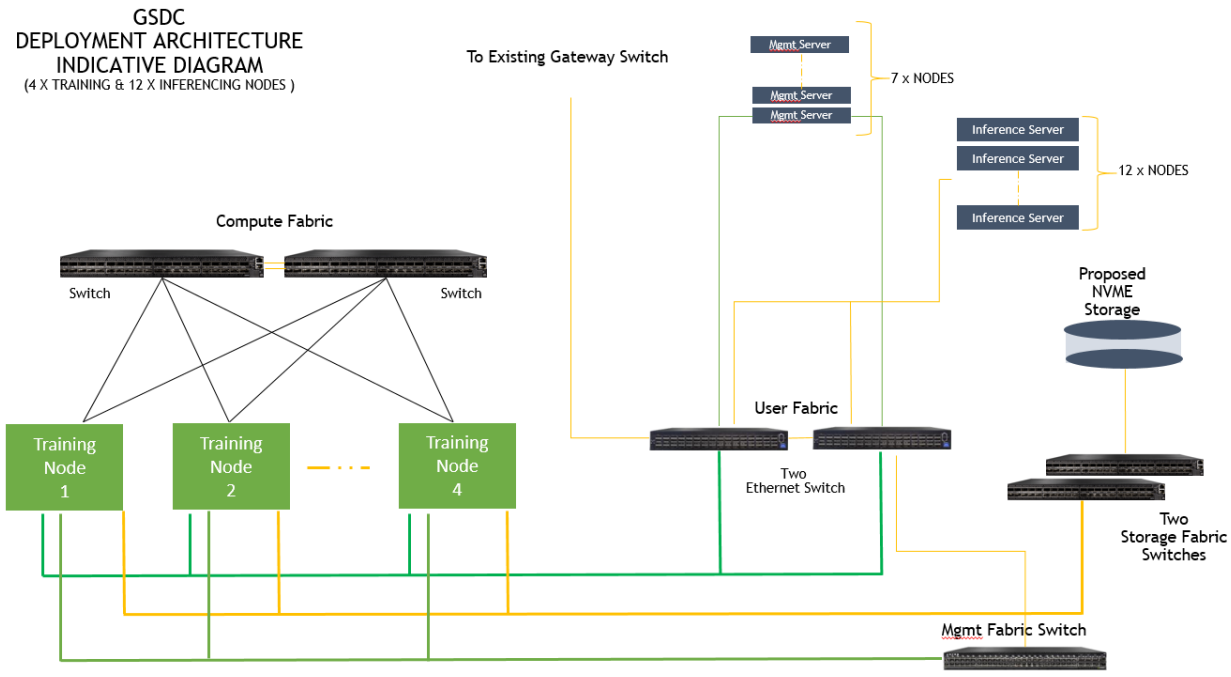
Storage Nodes	
External Storage	The solution should be PFS (Parallel File System) OR NFS (Network File System) based and delivered with 1PB (All NVMe) usable post RAID 6/equivalent or better protection, expandable up to 2PB in the same file system.
	The proposed storage array should be configured with no single point of failure, including required controllers, cache, power supply, cooling fans, etc. It should be scalable up to 12 additional controllers/nodes.
	1PB (NVMe) usable post RAID 6 or better configuration The storage should be distributed with namespace consistent across nodes.
	Performance: Min 32 GBps Read and Min 16 GBps Write aggregated from day one and scalable up to 200% with a scale-out architecture and additional controllers/nodes in the future. ~~IOPS: minimum 8,00,000~~ 1. Storage must offer NVIDIA GPUDirect Storage connectivity to GPUs. 2. NVMe Storage offered must be certified with the proposed GPU OEM. **Front-End Connectivity:** 200GBE or higher Ethernet connectivity compatible with all nodes as per proposed solution.

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 Profile 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" Identification	“U” numbering should be provided on the 19" mounting rails such that these unique numbers are visible after mounting of the equipment also.
6	PDU Provision	Each rack should have provision for installation of two PDU with toolless mounting provision 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 Perforation	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 Construction	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 without the use of tools. The front door of unit should be field reversible so that it may open from either side.
18	Statutory	100% assured compatibility with all equipment conforming to DIN 41494 /

	Standard	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 own solution to meet the requirement and ensuring minimum failure / no failure 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 Management and inference node with existing storage at GSDC as following;

Existing Storage

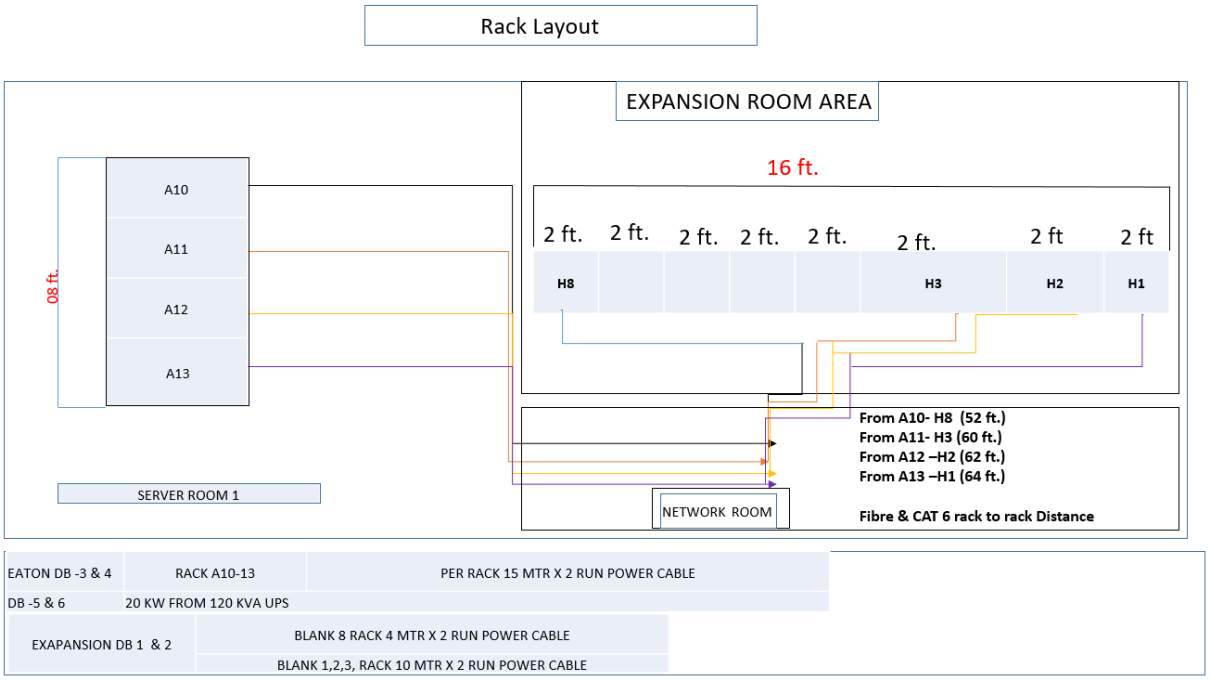
- **NetApp FAS8300 with Total ports 4 and Used ports 2**
- **Hitachi VSP 5600 with Total ports-64 ports and Used 64 ports**

CISCO MDS 9710 SAN Switch with 16 Gbps SFP (having port capacity of 32 Gbps) is used to connect with Storage. Port details are as below.

SAN Fabric Name	Total Ports	Used Ports	Available Ports
GSDC-Fabric-1	289	249	40
GSDC-Fabric-2	289	256	33

- For any additional requirement of ports over and above as aforementioned available ports, the bidder shall provide SAN switch with same or higher configuration compatible to connect inference node and management node with existing storage to complete the solution without any additional cost to the tenderer.
- The bidder has to ensure propose management and inference node solution should be compatible with aforementioned storage and switch. All necessary accessories, cabling, hardware, software, and licenses should be considered accordingly.

- Please find below tentative Layout diagram for Installation of RACKs.



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. Cost of **Comprehensive Annual Maintenance with warranty and OEM support** for 5 year D. Cost for O&M cost (including two skilled resources) for period of 5 Years	
	Total cost (Rs.)	

Note:

- CAPEX cost includes- **A, B, and C**. OPEX cost include **D.**
- 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>>**