(Revise RFP) Corrigendum-2 dated 15.03.2025

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

Please find the Pre-BID queries response and Revise RFP (Corrigendum-2) dated 15.03.2025 as below.

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

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
1	Eligibility Conditions	The bidder should have average Minimum Annual turnover of Rs. 50 crores in 3 years out of the last 5 financial years from the last date of bid submission with positive net worth.	Our request to provide a slight relaxation in Bidder turnover criteria. Our request to make it as below as other potential bidders can also participate.: "The bidder should have an average Minimum Annual turnover of Rs. 30 crores in 3 years out of the last 3 financial years from the last date of bid submission with positive net worth."	Please refer to corrigendum - 2
2	Eligibility Conditions	The bidder should have experience of setup of similar GPU base solution with GPU or CPU with below criteria in last 5 Years in India as on last date of bid submission:· One project having similar works total value of INR 40 Cr (Should have min 45 GPU / 800 CPU) or Two project having similar works total value of INR 25 Cr (Should have min 28 GPU / 500 CPU) or Three project having similar works total value of INR 20 Cr (Should have min 22 GPU / 400 CPU) Note: Similar works means SITC OF GPU ACCELERATED with multiple GPU Node	Our request to provide a slight relaxation in Bidder Experience criteria. Our request to make it as below as other potential bidders can also participate.: " The bidder should have experience of setup of similar GPU base solution with GPU or CPU with below criteria in last 5 Years in India as on last date of bid submission:· One project having similar works total value of INR 25 Cr (Should have min 20GPU / 500 CPU Cores) or Two project having similar works total value of INR 15 Cr (Should have min 10 GPU / 300 CPU Cores)	Please refer to corrigendum - 2
3	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	Our request to provide a slight relaxation in Eligibility Conditions. Our request to make it as below: "The OEM should have executed a similar GPU setup for 3 clients in the 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 65 Cr. Note: Similar works means SITC OF GPU ACCELERATED with multiple GPU Node."	Please refer to corrigendum - 2
4	Inference Node:	Infiniband Switch and FC Switches specifications are missing in RFP specs, while IB switches are mentioned in the connectivity diagram.	Please confirm, IB Switch.and FC switches Need to add in offer or GIL have enough port vacants in their own IB and FC Switches In their data centre.	Please refer to corrigendum - 2 for details on existing Switch at GSDC. However the bidder shall have to arrange for any additional required resources to complete the solution.
5	Inference Node	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux server Quoted OS should be under Enterprise support from OEM.Quoted model should be certified for RHEL, Ubuntu OS. The same shall be verifiable from OS OEMs website. Supply should include DC edition unlimited Guest OS licenses	Our Request, please consider Node OEM Compliance also a confirmation of OS support. And Compliance of OS support confirmation should be listed in the public domain. Our request to make it as below: "The system should support the latest version of Red Hat Enterprise Linux / Ubuntu Linux server Quoted OS should be under Enterprise support from OEM. Quoted model should be certified /Complied for RHEL, Ubuntu OS. The same shall be verifiable from OS /Node OEMs website. Supply should include DC edition unlimited Guest OS licenses "	No Change
6	Inference Node	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	This clause is complying offering thru only Foreign OEMs, so our request to please delete this clause or allow Node OEM Compliance for the same.	Please refer to corrigendum - 2
7	Inference Node	"Image segmentation (medical) 3D-Unet-99 Throughput for single NODE inference (99% offline) = 45 Samples/s or higher" "NLP Bert-99 Throughput for single NODE inference (99% offline) = 70000 Samples/s or higher" "Recommendation dlrm-v2-99 Throughput for single NODE inference (99% offline) = 610000 Samples/s or higher" "LLM Summarization gptj-99 Throughput for single NODE inference (99% offline) = 17000 Tokens/s or higher". "LLM Q&A LLama2-70B Throughput for single NODE inference (99% offline) = 27000 Tokens/s or higher" "LLM Text Generation (Math & Code) Mixtral-8x7B Throughput for single NODE inference (99% offline) = 53000 Tokens/s or higher" "Image Classification ResNet Throughput for single NODE inference (99% offline) = 716000 Samples/s or higher"	This clause is complying offering thru only Foreign OEMs, so our request to please delete this clause or allow Node OEM Compliance for the same.	Please refer to corrigendum - 2

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
		"Object Detection RetinaNet Throughput for single NODE inference (99% offline) = 11400 Samples/s or higher" "Image Generation Stable Diffusion-XL Throughput for single NODE inference (99% offline) = 14 Samples/s or higher"		
8	Inference Node	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.	This clause is complying offering thru only Foreign OEMs, so our request to please delete this clause or allow Node OEM Compliance for the same.	Please refer to corrigendum - 2
9	Storage Nodes	1. NVMe Storage offered must be supported /certified with the latest Nvidia DGX Super- POD and should be the same model with certification.	This clause is complying offering thru only Foreign OEMs, so our request to please delete this clause or allow Node OEM Compliance for the same.	Please refer to corrigendum - 2
10	42U Server rack	The rack shall be from OEM having ISO9001:2008, ISO14001:2004, ISO 45001:2018 & ISO 50001:2018 (Certificate to be submitted along with compliance)	Only One or Two OEM's are qualified on clause of ISO 50001:2018 certification, so please remove this ISO certification for a healthy participation.	Please refer to corrigendum - 2
11	Bid Submission End Time	25. Feb.2025,3.00 PM	Solution comprises multiple vendors, so our request to extend the bid submission end time at least 7 days more	Please refer to Corrigendum - 2
12	Revised RFP document, Point no.5 1. Eligibility Conditions under revised RFP document	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.	We request GIL to kindly generalize the OEM Eligibility Criteria to have a level playing field for all bidders. And kindly amend the clause as "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 50 Cr. Note: Similar works means SITC OF GPU ACCELERATED with multiple GPU Node. " It is important to highlight that even GeM terms and conditions (Referring to Clause no. 2 past performance) also suggest past experience in terms of value of deployments should be 80% of bid quantity, which translates to cost of the past deployment(s) should be less than or 80% of the desired current purchase. Whereas in your case cost of the past deployment is asked to be "125 Cr." which is multiple times the cost of desired purchase.	As per corrigendum dated 15.02.2025
13	Revised RFP document, under specs of Inference Node: Networking Switches Inference Node: Networking Switches	InfiniBand Switch specifications for Connectivity of AI Training nodes is missing. While Indicative Diagram provided by GIL, clearly specifies Two number of Infini Band Switches are required	Ambiguity in the tender technical specifications required to be clarified. We request GIL to kindly Confirm weather we need to refer to indicative diagram or specifications laid in tender document.	Please refer to corrigendum - 2
14	Revised RFP document, under specs of Inference Node: Certification Inference Node: 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.	Important to highlight that (i) As per tender specifications only H200 PCI Express variant with 2 GPU's populated can only qualify, Which requires OEM to offer model with 2 x H200 PCI Express GPU's and as per tender OEMs have to list offered product in GPU OEM website with 2 H200. (ii) Kindly refer Snapshot from GPU OEM website which clearly shows that there is no category for 2 x H200 PC express GPUs in their website or listing process. In view of above restrictions, it is not possible to offer a complied model meeting tender specifications, hence we request GIL to kindly DELETE this clause.	Please refer to corrigendum - 2
		"Image segmentation (medical) 3D-Unet-99 Throughput for single NODE inference (99% offline) = 45 Samples/s or higher" "NLP Bert-99 Throughput for single NODE inference (99% offline) = 70000 Samples/s or higher" "Recommendation dlrm-v2-99 Throughput for single NODE inference (99% offline) = 610000 Samples/s or higher"		

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
15	Revised RFP document, under specs of Inference Node: GPUs Benchmark GPU's listing on Mlperf infrenceing Benchmark	"LLM Summarization gptj-99 Throughput for single NODE inference (99% offline) = 17000 Tokens/s or higher". "LLM Q&A LLama2-70B Throughput for single NODE inference (99% offline) = 27000 Tokens/s or higher" "LLM Text Generation (Math & Code) Mixtral-8x7B Throughput for single NODE inference (99% offline) = 53000 Tokens/s or higher" "Image Classification ResNet Throughput for single NODE inference (99% offline) = 716000 Samples/s or higher" "Object Detection RetinaNet Throughput for single NODE inference (99% offline) = 11400 Samples/s or higher" "Image Generation Stable Diffusion-XL Throughput for single NODE inference (99% offline) = 14 Samples/s or higher"	As per the revised GPU configuration specifications for the inference node, it has become quite restrictive to offer quite a specific GPU model and yet meeting strict criteria of Mlperf inferencing Benchmark, which requires at least 3-4 months from now, hence we request GIL to allow submission of "Benchmark report" instead of the "Benchmark listing ".	Please refer to corrigendum - 2
16	Revised RFP document, under specs of Inference Node: Performance Benchmark Inference Node: Performance Benchmark	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.	As per the revised GPU configuration specifications for the inference node, it has become quite restrictive to offer quite a specific model based on Genova or latest architecture (AMD) / (Intel) Emerald rapid or latest platform and yet meeting strict criteria of SPEC.org benchmark, which requires at least 5-6 months from now, hence we request GIL to allow submission of "Benchmark report" instead of the "Benchmark listing ".	Please refer to corrigendum - 2
17	Revised RFP document, under specs of Storage Nodes External Storage; Nvidia Integration	1. NVMe Storage offered must be supported /certified with the latest Nvidia DGX Super- POD and should be the same model with certification.	Since the DGX Super-POD certification for storage applies t**o Parallel file system and** not the storage appliance, hence we request GIL to kindly revise the clause as below. "Parallel File System for NVMe Storage must be supported /certified with the latest Nvidia DGX Super- POD and should be the same **Parallel File System w**ith certification."	Please refer to corrigendum - 2
18	Revised RFP document, under specs of42U Server rack 19. Certifications	The rack shall be from OEM having ISO9001:2008, ISO14001:2004, ISO 45001:2018 & ISO 50001:2018 (Certificate to be submitted along with compliance)	We request GIL to kindly generalize the OEM Eligibility Criteria to have a level playing field for all bidders. And kindly amend the clause as "The rack shall be from OEM having ISO9001:2008, ISO14001:2004, ISO 45001:2018"	Please refer to corrigendum - 2
19	Page No. 3, RFP clause no. 3	SOW	Kindly confirm the scope of Managed Services	As mentoned in RFP and time -to-time published corrigendum
20	Page No 3, RFP No. 3-6	The solution should support remote console access to all the servers for cluster server's health monitoring at Fast Ethernet or better access speed.	Please confirm the scope as looking the access over VPN access? If yes then confirm about the VPN connectivity.	Cluster Management should be accessable within Data Centre Premissis. However, VPN access shall be available as per GSDC policy.
21	Page no, 3, RFP no. 3-8	The bidder will have to supply Server Rack along with provision of iPDU, TOR Switch, patch panel, cables, SFP modules, any other active/passive components etc. to host the HPC Cluster with GPUs at GSDC. Any other component required for the solution proposed by thesupplier has to be incorporated for completion of the Solution.	Please confirm the scope of HPC Cluster in AI ML Solution Model	Please refer to Corrigendum - 2
22	Page no, 3, 3-9	Onsite comprehensive annual maintenance with warranty and OEM support for 5 Years from the date of completion of User Acceptance Test (Onsite warranty will include those sites to which the item supplied under the contract is moved, in case of migration of the equipment). Warranty should include but not limited to - On-going Firmware updates, Proactive bug fixes, Preventive Maintenance, Parts replacement, etc.	Please cofirm the criteria of UAT	As mentoned in RFP and time -to-time published corrigendum
23	Page no. 4, 1&2	Manpower for Hand Holding Support	1) Professiency for Hypervisor resource Reuirement are missing. 2) Assuming 2+2 Resoruce with including Backup Rerources for "System Admin" and "AI/ML Development Engineer".	As per corrigendum dated 15.02.2025

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
24	Page no. 7, 8.2	Supply of the Hardware including Licenses and OEM	45 Days is not enough for Hardware and Licensing	Please refer to corrigendum - 2
25	Page no. 10, 10	Following N/W are required: a) Infiniband / Ethernet (100Gb/s or higher) as required for quoted storage delivery to nodes b) Ethernet (100Gb/s or higher) for User delivery c) Ethernet (10GbE or higher) for cluster orchestration d) Ethernet (10GbE or higher) for perimeter connectivity e) Ethernet (1GbE or higher) for in-band management	Cluster and Perimeter Switch details are not cleared as per Architecture Diagram. Please provide the same for further participation.	The diagram in the published revised RFP is Indicative only, the bidder shall submit their own diagaram as per propossed solution to complete the solution as aked in this RFP and time to time corrigendum.
26	Page no. 10, 10	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux server Quoted OS should be under Enterprise support from OEM.	Ubuntu is an open-source OS. Do you need the support for it?	Yes , Premium / highest level OEM support is required for same
27	Page no. 10, 10	Hypervisor with Enterprise level highest license and should be provided from day one. Support	What is the scope of Hypervisor?	Hypervisor shall be used to cater requirement of virtual environment as asked in SOW and functionality in published RFP and time to time published corrigendum.
28	Page no. 10, 10	System should be scalable with multi node cluster. Software support & cluster tools and management hardware and software and licenses to be supplied along with product. Compatible with Full-stack reference designs of the Storage providers.	Can we offer open-source Cluster management tools?	Yes , you can offer open-source Cluster management tools. However, the complete responsibility shall be with the bidder and proposed open-source Cluster management tools Should cater requiremnt of RFP and should be supplied with premium or highest level support.
29	Page no. 10, 11	a) Minimum 8 nos of InfiniBand NDR ports or Ethernet (200Gb/s or higher) for compute communication for internode communication, b) 1 nos. of port for BMC (dedicated LAN port), c) Minimum 1 no. of 1 GbE port and 2 nos of 10 GbE or higher (Fiber/Copper) port. d) 1 nos. of InfiniBand / 100G or higher Ethernet 2 x twinport HCA as required for quoted storage delivery	Need clarification of Infiband Model	Please refer to corrigendum - 2
30	Page no. 11, 11	Following N/W are required: 1. NDR Infiniband / Ethernet (200Gb/s or higher) for compute communication 2. Infiniband/Ethernet (100Gb/s or higher) for storage delivery 3. Ethernet (Min 10 GbE or higher) for cluster orchestration 4. Ethernet (10 GbE or higher) for perimeter connectivity 5. Ethernet (1GbE) for in-band management	Need clarification of Infiband Model	Please refer to corrigendum - 2
31	Page no. 12, 12	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux server Quoted OS should be under Enterprise support from OEM	Ubuntu is an open-source OS. Do you need the support for it?	Yes , support is required for it and should be supplied with premium or highest level support.
32	Page no. 13, 12 Scalability, Cluster and Management Hardware and software System	System should be scalable with multi node cluster. Software support & cluster tools and management hardware and software and licenses to be supplied along with product. Full-stack reference designs with all of the leading Storage providers.	Can we offer open-source cluster management tools?	Yes , you can offer open-source Cluster management tools. However, the complete responsibility shall be with the bidder and proposed open-source Cluster management tools can cater requiremnt of RFP and should be supplied with premium or highest level support.
33	Page no. 14, 13 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.	2 x Accelerators Per Node is not enough for the requirement It should be atleast 4 X Accelrator	Please refer to corrigendum - 2
34	Page no. 15, 13 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	Need clarification on the Specification	Proposed GPU solution having capacity of partitioning of single GPU accellator into multiple GPU instances where both memory and compute of the GPU is divided into multiple instances.

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
35	Page no. 16, 14 HBA Card	16 Gbps Host Bus Adaptor for connecting with existing storage	HBA Port is required ? Or we can connect it directly to Switch	Please refer to corrigendum - 2
36	Page no. 17, 14 Networking Switch	Two Nos. of Switch with 48 *10G SFP+ and 8 x 100G QSFP ports to connect all of “Master Node”, “Node for AI Training” and “Inference Node” to form Cluster Communication and Perimeter N/W. Switch must support of MLAG /MCLAG feature. Switch must support EVPN – VxLAN based network. Required cables of appropriate length, transceiversshould be supplied.Switches(s) should have redundant PowerSupply. 5 Years Comprehensive Onsite Warranty.2. One Nos. of Switch with 48 *1G RJ45,4* 25G SFP28 and 2 x 100GQSFP ports to connect all of “Master Node”, “Node for AI Training”and “Inference Node” to form In-band and BMC/Out-of-bandManagement N/W. Required cables of appropriate length,transceivers should be supplied. Switch should have redundant Power Supply. 5 Years Comprehensive Onsite Warranty.	Is they any Specific Switch Model or Vendor requried?	Bidder is free to quote any make and model for switch but proposed switch can cater asked functionality publihed RFP and time to time published corrigendum.
37	Page no. 18, 14 OS Support	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux server Quoted OS should be under Enterprise support from OEM. Quoted model should be certified for RHEL, Ubuntu OS. The same shall be verifiable from OS OEMs website. Supply should include DC edition unlimited Guest OS licenses	Ubuntu is an open-source OS. Do you need the support for it?	Yes , Premium / highest level OEM support is required for same
38	Page no. 19, 14 Hypervisor	Hypervisor with Enterprise level highest license and support available should be provided from day one.	What is the scope of Hypervisor?	Hypervisor shall be used to cater requirement of virtual environment as asked in SOW and functionality in published RFP and time to time published corrigendum.
39	Page no. 20, 14 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.	What is the scope of Hypervisor and VMS? Need to be clarify	Hypervisor and VMs shall be used to cater requirement of virtual environment as asked in SOW and functionality in published RFP and time to time published corrigendum.
40	Page no. 20, 15 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	Can we offer any open-source for cluster management tools?	Yes , you can offer open-source Cluster management tools. However, the complete responsibility shall be with the bidder and proposed open-source Cluster management tools can cater requiremnt of RFP and should be supplied with premium or highest level support.
41	Page no. 21, 16 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.	2 x Accelerators Per Node is not enough for the requirement It should be atleast 4 X Accelrator	Please refer to corrigendum - 2
42	Page no. 22, 17 External Storage	Disks: 1PB (NVMe) usable post RAID 6 or better configuration The storage should be distributed with namespace consistent across nodes.	Need clarfication if existing FC switch port can be utlized for New Storage Box.	Bidder to provide all items needed to run the AI cluster successfully. Also refer to corrigendum - 2 for details.
43	Page no. 23, 18	Existing Storage Net App NAS Storage Hitachi NVME SAN Storage	Please provide model details to understand exisitng storage box	Please refer to corrigendum - 2

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
44	Page no. 24, 18	The bidder has to integrate the existing storage with the GPU node. All necessary accessories, cabling, hardware, software, and licenses should be considered accordingly.	Please clarify the scope	To train on existing data of respective department of GOG which are availabled on existing storage.
45	Eligibility Condition s (Point No. 05)	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 clientdeployment should be One project having similar works total value of INR 125 Cr. Note: Similar works means SITC OF GPU. Copy of Work Order along with Completion / Go-Live certificate	We request you to Kindly remove / Relax this clause. As we are very keen to participate in this above-mentioned GeM bid.	As per corrigendum dated 15.02.2025
46	Eligibility Condition s (Point No. 3)	The OEM should have average Annual Turnover of minimum Rs. 250 crores for the last five financial years from the last date of bid submission with positive net worth. ● Audited profit and loss statement and balance sheet ● Auditor certificate clearly specifying the turnover and positive net worth	We request you to kindly relax this clause, as we are very keen to participate in this above mentioned GeM bid.	As per corrigendum dated 15.02.2025
47	Responses to Pre-BID queries dated 15.02.2025 and Revised RFP (Corrigendum).10.	Internal Storage : Minimum 4*3.84 TB capacity interface hot swap Enterprise NVMe SSDs.	Please provide clarity on the intended use case for this storage and how it will be integrated into the overall solution.	For Data purpose this storage is required.
48	Responses to Pre-BID queries dated 15.02.2025 and Revised RFP (Corrigendum).10. Master Node	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	The current specification limits us to older CPU series. Please revise the requirement to ensure that bidders must submit benchmarks based on the latest CPUs. These benchmarks should be provided before installation.	Please refer to corrigendum - 2
49	Responses to Pre-BID queries dated 15.02.2025 and Revised RFP (Corrigendum).12.Infer ence Node: Total Qty-	Internal Storage : Minimum 4*3.84 TB capacity interface hot swap Enterprise NVMe SSDs.	Please provide clarity on the intended use case for this storage and how it will be integrated into the overall solution.	For Data purpose this storage is required.
50	Responses to Pre-BID queries dated 15.02.2025 and Revised RFP (Corrigendum).12.Infer ence Node: Total Qty-	HBA Card : 16 Gbps Host Bus Adaptor for connecting with existing storage.	Will only the inference node access this storage, or will it be shared across all nodes? Is the storage still under warranty? Are the necessary connectivity ports available? Is an FC switch already in place, or will it need to be procured?	Both Inference and Master Nodes has to be connected with the existing storage through HBA Card. Please refer to corrigendum - 2 for detail.
51	Responses to Pre-BID queries dated 15.02.2025 and Revised RFP (Corrigendum).12.Infer	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.	We request you to consider GPU OEM Certifiy the RACK SERVER OEM. So request you to change this.	Please refer to corrigendum - 2
52	Responses to Pre-BID queries dated 15.02.2025 and Revised RFP	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	The current specification limits us to older CPU series. Please revise the requirement to ensure that bidders must submit benchmarks based on the latest CPUs. These benchmarks should be provided before installation.	Please refer to corrigendum - 2
53	Storage Node	External Storage: Performance: 20GBps Read/Write from day one and scalable up to 60GBps with a scale-out architecture and additional controllers/nodes in the future. NVIDIA Integration: 1. Storage must offer NVIDIA GPUDirect Storage connectivity to GPUs. 2. NVMe Storage offered must be certified with the latest Nvidia DGX SuperPOD and should be the same model with certification	This specification appears to be tailored to a specific OEM. We request its removal to ensure a more open and competitive evaluation process. We request you to remove this point of alter that with GPU OEM should Certify the Storage vendor	Please refer to corrigendum - 2
54	8. IMPLEMENTATION TIMELINES & PENALTIES	Supply of the Hardware including Licenses and OEM:T1=T+45 days from the date of issuance	We request an extension to a minimum of 90 days	Please refer to corrigendum - 2

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
55	Generic		Please share the data centre layout for reference. With allotted space details for this project	Tentative layout diagram is enclosed with the corrigendum - 2. However the Bidder (at its own cost) can visit data centre in advance (before the bid submission date) during working days and hours to assess the rack positioning. Based on their assessment, they should quote their solution with the bid response.
56	1 Eligibility Criteria, Point no 4	The bidder should have experience of set up of similar GPU base solution with GPU or CPU with below criteria in last 5 Years in India as on last date of bid submission: One project having similar works total value of INR 40 Cr (Should have min 45 GPU / 800 CPU) or Two project having similar works total value of INR 25 Cr (Should have min 28 GPU / 500 CPU) or Three project having similar works total value of INR 20 Cr (Should have min 22 GPU / 400 CPU) Note: Similar works means SITC OF GPU ACCELERATED with multiple GPU Node	Kindly confirm if the count mentioned is for number of cores of the GPU / CPU or is it number of processors / GPU Cards also request to change clause to below: The bidder should have experience of set up of similar GPU base solution with GPU or CPU with below criteria in last 7 Years in India as on last date of bid submission:	Please refer to corrigendum - 2
57	1 Eligibility Criteria, Point no 4.1	The bidder should have experience of set up of similar GPU base solution with Cumulative 15 nos. of GPU based solution in last 5 Years in India.	Kindly change to : The bidder should have experience of set up of GPU based solution with some quantity of GPUs in last 7 Years in India.	As per corrigendum dated 15.02.2025
58	Page no. 1, Eligibility Condition, Sr. No 2	The bidder should have average Minimum Annual turnover of Rs. 50 crores in 3 years out of last 5 financial years from the last date of bid submission with positive net	**Amendment Request As Below:** The bidder should have average Minimum Annual turnover of Rs. ~~50~~ **30** crores in 3 years out of last 5 financial years from the last date of bid submission with positive net **Justification:-** Request you to please consider the above modification as it will enable a broader range of SIs to qualify and participate in the bid .	Please refer to corrigendum - 2
59	Page no. 1, Eligibility Condition, Sr. No 4	The bidder should have experience of set up of similar GPU base solution with GPU Copy of Work Order along with Completion / Go-Live certificate or CPU with below criteria in last 5 Years in India as on last date of bid submission: - One project having similar works total value of INR 40 Cr (Should have min 45 GPU / 800 CPU) or - Two project having similar works total value of INR 25 Cr (Should have min 28 GPU / 500 CPU) or - Three project having similar works total value of INR 20 Cr (Should have min 22 GPU / 400 CPU) Note: Similar works means SITC OF GPU ACCELERATED with multiple GPU Node.	**Amendment Request As Below:** The bidder should have experience of set up of similar GPU base solution with GPU Copy of Work Order along with Completion / Go-Live certificate or CPU with below criteria in last 5 Years in India as on last date of bid submission: - One project having similar works to- tal value of INR ~~40~~ **20** Cr (Should have min 45 GPU / 800 CPU) or - Two project having similar works total value of INR ~~25~~ **15**Cr (Should have min 28 GPU / 500 CPU) or - Three project having similar works total value of INR ~~20~~ **10**Cr (Should have min 22 GPU / 400 CPU) Note: Similar works means SITC OF GPU AC-CELERATED with multiple GPU Node **Justification:-** Request you to please consider the above modification as it will enable a broader range of SIs to qualify and participate in the bid .	Please refer to corrigendum - 2
60	Corrigendum -1 , SLA & Penalties (Operational Penalties), Page-22	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 Hardware resolution can be done, but Software related issues may take longer time to resolve.	Please refer to RFP and time-to-time published corrigendum

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
61	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.	**Modification** : Number of GPUs and GPU Communication : 4 x Accelerators per node, each with minimum 94 GB or higher GPU per Accelerator. **Justification** : We request GIL to consider the amendment, **as 140GB GPUs with NVL are not yet available from leading OEMs for 2 GPU Configurations**, as these are under design consideration due to High Power Consumption by GPUs with Higher Memory, Cores et al. Hence request GIL to consider the amended configuration. Also the Benchmarks as asked are available for 4 GPUs or 8 GPUs only, as High listed in the following point below.	Please refer to corrigendum - 2
62	10. Minimum Technical Specification, Inference Node, Page-28	"Image segmentation (medical) 3D-Unet-99 Throughput for single NODE inference (99% offline) = 45 Samples/s or higher" "NLP Bert-99 Throughput for single NODE inference (99% offline) = 70000 Samples/s or higher" "Recommendation dlrm-v2-99 Throughput for single NODE inference (99% offline) = 610000 Samples/s or higher" "LLM Summarization gptj-99 Throughput for single NODE inference (99% offline) = 17000 Tokens/s or higher" "LLM Q&A LLama2-70B Throughput for single NODE inference (99% offline) = 27000 Tokens/s or higher" "LLM Text Generation (Math & Code) Mixtral-8x7B Throughput for single NODE inference (99% offline) = 53000 Tokens/s or higher" "Image Classification ResNet Throughput for single NODE inference (99% offline) = 716000 Samples/s or higher" "Object Detection RetinaNet Throughput for single NODE inference (99% offline) = 11400 Samples/s or higher" "Image Generation Stable Diffusion-XL Throughput for single NODE inference (99% offline) = 14 Samples/s or higher"	We would like to bring to GIL's notice , the Benchmarks asked in Inference nodes areof 8 GPUs of NVIDIA H200 GPU. Whereas the ask for Inferensing GPU is of 2 GPUs. Plesae clarify , if we just need to submitt the perfomrance benchmarks of 8 GPU as listed ??? or for 2 GPUs?? **NOTE : The Mlcommons.org , DOES NOT have rating for 2 GPUs, it is either 4 GPU or 8 GPU ratings only.** **The Benchmarks should be as follows for 4 GPU Inferecing nodes ,as avaialbel on Mlcommons.org site as proof.** "Image segmentation (medical) 3D-Unet-99 Throughput for single NODE inference (99% offline) = 22.29 Samples/s or higher" "NLP Bert-99 Throughput for single NODE inference (99% offline) = 24850 Samples/s or higher" "Recommendation dlrm-v2-99 Throughput for single NODE inference (99% offline) = 208210 Samples/s or higher" "LLM Summarization gptj-99 Throughput for single NODE inference (99% offline) = 7190 Tokens/s or higher" ~~To remove : "LLM Q&A LLama2-70B Throughput for single NODE inference (99% offline) = 27000 Tokens/s or higher"~~ ~~"LLM Text Generation (Math & Code) Mixtral-8x7B Throughput for single NODE inference (99% offline) = 53000 Tokens/s or higher" (to be removed)~~ "Image Classification ResNet Throughput for single NODE inference (99% offline) = 250295 Samples/s or higher" "Object Detection RetinaNet Throughput for single NODE inference (99% offline) = 5435 Samples/s or higher" "Image Generation Stable Diffusion-XL Throughput for single NODE inference (99% offline) = 6.1 Samples/s or higher"	Please refer to corrigendum - 2
63	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	if GNN is not available, RGAT is accepted instead of GNN; Also, Please refer to corrigendum -2
64	Storage Nodes, Page-30	IOPS : 8,00,000	**Modification** : We request **GIL to REMOVE** this point, as NFS storage will not be able to such high IOPS, while the actual measure of performance is throughput ,.Even the GPU OEM NVIDIA recommendation of performance for GPU based servers is of throughput and not IOPS, please see https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-h100/latest/storage-architecture.html.	As per corrigendum dated 15.02.2025

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
65	Note, Page-32	In addition, Bidder has to connect proposed solution with existing storage at GSDC as following; **Existing Storage** Net App NAS Storage Hitachi NVME SAN Storage The bidder has to integrate the existing storage with the GPU node. All necessary accessories, cabling, hardware, software, and licenses should be considered accordingly	We reqeust GIL to claify these storage will have 10/25 or 100GBE conenctivity for requried data migration to GPU storage!! Or how GIL plan to do this integration!! As the current statement is not clear as how this needs to be accomplished, it has dependencies, compatibility to be clarified and pre-requisites to be provided by GIL.	Please refer to corrigendum - 2
66	Integratio nwith Existing Network	Networking details from existing Networking switches..	Existing Networking switch details requried along with Port Speed availabe to check on compatibility of propsoed Switches.	Please refer to corrigendum - 2
67	Integartion with GPU Cluster	The bidder has to integrate the existing storage with the GPU node. All necessary accessories, cabling, hardware, software, and licenses should be considered accordingly	**Clarification** : We reqeust GIL to clarify on the connectivity , as the GPU nodes are with 100GbE ethernet, the existing storage should have 100 GbE connecvity either on controllers or though Networking Switches.	The bidder has to integrate the existing storage with the Inference and Master Node. All necessary accessories, cabling, hardware, software, and licenses should be considered accordingly. Also, please refer to corrigendum - 2 for detail
68	Additional Security Point	System should support instant secure system erase, including NVMe drives	We request GIL to add these security features into the , Master Nodes, Training , Inferencing nodes at least, as this will ensure secure environment.	No Change
69		System should support dynamically enabled USB ports		No Change
70		System Should support Automatic , secure recovery of OS as well as BIOS and BMC		No Change
71	Networking Switch	Switch Type 1, 2 and 3 mention the similar information for all kinds of servers / storages, etc: "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"	Ideally, all categories of switches need to be different. For Inferencing, Training Nodes, Storage nodes and Master Nodes, as the port type mentioned are **100G**, applicable network switch need to be at least **32 or 48**x 100G to support 12x Inferencing nodes, 4x Traning Nodes (Front End) and 7x Master Nodes, along with the ability to form MC-LAG and BGP-EVPN capabilities for DC Network For OOB Network, 48x 1G, 4x 10G SFP+ and 2x 100G are adequate for connecting devices. 10G ports can be used to uplink to firewall, etc for access from outside. For backend training network, the applicable switch need to be at least 64x 200/400G. Since each training node has 8 NICs and the future scalability is 200 GPUs (200 NICs), going for 64-port switch (Ethernet / Infiniband) will be most suitable for such case For perimeter network access purpose, 48x 10/25G and 6x 100G is adequate and to connect them to Spines, while also forming MC-LAG between them. Therefore, suggestion is to add and mention 2x perimeter switch. 25G ports can be suitable for future scalability as number of GPUs will be scaled to 200+ requiring more ingestion of traffic through the perimeter. In addition, at least **2-3** Spines of 32/48x 100G ports will be needed to interconnect all the different switches for Inferencing, storage, master nodes and perimeter switches. 3 Spines can provide additional level of redundancy and adequate path for the traffic between different servers / storages	This is mimimum functional requirement. Bidder can quote higher specifications. Also please refer to corrigendum - 2
72	Delivery Time frame of GPU Nodes		**We request GIL to consider the Delivery Period of these GPU servers as minimum of 16-18 weeks from the date of confirm PO released.** **Justification :** AS GPU OEM takes longer lead times, we would reqeust GIL to consider the extended delivery time for the asked GPU solution.	Please refer to corrigendum - 2

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
73	Corrigendum -1 , SLA & Penalties (Operational Penalties), Page-22	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 Hardware resolution can be done, but Software related issues may take longer time to resolve.	As mentoned in RFP and time -to-time published corrigendum
74	10. Minimum Technical Specification, Inference Node, Page-28	"Image segmentation (medical) 3D-Unet-99 Throughput for single NODE inference (99% offline) = 45 Samples/s or higher" "NLP Bert-99 Throughput for single NODE inference (99% offline) = 70000 Samples/s or higher" "Recommendation dlrm-v2-99 Throughput for single NODE inference (99% offline) = 610000 Samples/s or higher" "LLM Summarization gptj-99 Throughput for single NODE inference (99% offline) = 17000 Tokens/s or higher" "LLM Q&A LLama2-70B Throughput for single NODE inference (99% offline) = 27000 Tokens/s or higher" "LLM Text Generation (Math & Code) Mixtral-8x7B Throughput for single NODE inference (99% offline) = 53000 Tokens/s or higher" "Image Classification ResNet Throughput for single NODE inference (99% offline) = 716000 Samples/s or higher" "Object Detection RetinaNet Throughput for single NODE inference (99% offline) = 11400 Samples/s or higher" "Image Generation Stable Diffusion-XL Throughput for single NODE inference (99% offline) = 14 Samples/s or higher"	We would like to bring to GIL's notice , the Benchmarks asked in Inference nodes areof 8 GPUs of NVIDIA H200 GPU. Whereas the ask for Inferensing GPU is of 2 GPUs. Plesae clarify , if we just need to submitt the perfomrance benchmarks of 8 GPU as listed ??? or for 2 GPUs?? **NOTE : The Mlcommons.org , DOES NOT have rating for 2 GPUs, it is either 4 GPU or 8 GPU ratings only.** **The Benchmarks should be as follows for 4 GPU Inferecing nodes ,as avaialbel on Mlcommons.org site as proof.** "Image segmentation (medical) 3D-Unet-99 Throughput for single NODE inference (99% offline) = 22.29 Samples/s or higher" "NLP Bert-99 Throughput for single NODE inference (99% offline) = 24850 Samples/s or higher" "Recommendation dlrm-v2-99 Throughput for single NODE inference (99% offline) = 208210 Samples/s or higher" "LLM Summarization gptj-99 Throughput for single NODE inference (99% offline) = 7190 Tokens/s or higher" ~~To remove : "LLM Q&A LLama2-70B Throughput for single NODE inference (99% offline) = 27000 Tokens/s or higher"~~ ~~"LLM Text Generation (Math & Code) Mixtral-8x7B Throughput for single NODE inference (99% offline) = 53000 Tokens/s or higher" (to be removed)~~ "Image Classification ResNet Throughput for single NODE inference (99% offline) = 250295 Samples/s or higher" "Object Detection RetinaNet Throughput for single NODE inference (99% offline) = 5435 Samples/s or higher" "Image Generation Stable Diffusion-XL Throughput for single NODE inference (99% offline) = 6.1 Samples/s or higher"	Please refer to corrigendum - 2
75	Storage Nodes, Page-30	**IOPS : 8,00,000**	**Modification** : We request **GIL to REMOVE** this point, as NFS storage will not be able to such high IOPS, while the actual measure of performance is throughput ,.Even the GPU OEM NVIDIA recommendation of performance for GPU based servers is of throughput and not IOPS, please see https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-h100/latest/storage-architecture.html.	As per corrigendum dated 15.02.2025
76	Note, Page-32	In addition, Bidder has to connect proposed solution with existing storage at GSDC as following; **Existing Storage** Net App NAS Storage Hitachi NVME SAN Storage	We reqeust GIL to claify these storage will have 10/25 or 100GBE conenctivity for requried data migration to GPU storage!! Or how GIL plan to do this integration!! Asked for Dual - Port 16G FC Card for connecting existing Netapp & Hitachi based Storage with these server's How the connectivity is going to take place, need clarity that only Inference node are going to access that storage or its going to use with all the nodes. What is the leanth of the RACK to Rack, Help to share the layout of t he DC.	Please refer to corrigendum - 2
77	Integratio nwith Existing Network	Networking details from existing Networking switches..	Existing Networking switch details requried along with Port Speed availabe to check on compatibility of propsoed Switches.	Please refer to corrigendum - 2
78	Integartion with GPU Cluster	The bidder has to integrate the existing storage with the GPU node. All necessary accessories, cabling, hardware, software, and licenses should be considered accordingly	**Clarification** : We reqeust GIL to clarify on the connectivity , as the GPU nodes are with 100GbE ethernet, the existing storage should have 100 GbE connecvity either on controllers or though Networking Switches.	FC HBA are required for Inference & master node. Propossed Inference & Master Node should be compatible with existing Hitachi & Netapp Storage and SAN & FC Switch. Also, Please refer to corrigendum - 2 for detail.

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
79	Power		Training node required 15 KW each as power, how much power is there in a single RACK. Power is going to come from UPS ?	Please refer to corrigendum - 2
80	**Delivery of Goods**		**We request GIL to consider the Delivery Period of these GPU servers as minimum of 14-16 weeks from the date of confirm PO released.**	Please refer to corrigendum - 2
81	Tender submission		we request GIL , Currently RFP due date is 25 feb 2025 , kidnly extend for another 15 working days. After above clarification .	Please refer to corrigendum - 2
82	25 Nodes for AI Training	AI nodes for Training. For Operating System: minimum 1.92 TB M.2 NVMe drives.	Kindly change it to 'For Operating System: minimum 480GB M.2 NVMe drives'	Please refer to corrigendum - 2
83	25 Nodes for AI Training	AI nodes for Training. For Data: Minimum 8 * 3.84 TB U.2 NVMe drives	Kindly change it to 'Minimum 8 * 3.84 TB U.2 or EDSFF NVMe drives' since EDSFF drives are newer form factor drives.	Please refer to corrigendum - 2
84	26 Nodes for AI Training	AI Nodes for Training - Full-stack reference designs with all of the leading Storage providers.	Kindly relax the requirement for 'Full-stack reference design', since HPE will ensure that the proposed GPU nodes and Storage will be compatible in the solution.	Please refer to corrigendum - 2
85	26 Nodes for AI Training	AI Nodes for Training - 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.	Kindly relax this requirement and allow Server OEM to take responsibility to for end to end compatibility of the proposed solution. The reason is that some models take time for certification by Nvidia.	Please refer to corrigendum - 2
86	27 and 29 Nodes for AI Training	Cluster Management & Scheduler and hardware - for AI Training and Inference nodes	Kindly relax this requirement and allow Server OEM to offer cluster tools which can manage this cluster. The features asked are proprietary to some vendor.	No Change
87	27 Nodes for AI Training	Benchmarks - AI nodes for Training	Kindly change it to the following:	Please refer to corrigendum - 2
88	a Nodes for AI Training	DLRM-dcnv2 - 3.6 minutes or less on single node	Kindly change it to 4.2 mins	Please refer to corrigendum - 2
89	b Nodes for AI Training	GNN - 7.8 minutes or less on single node	Kindly remove since we have not yet run this benchmark.	Please refer to corrigendum - 2
90	c Nodes for AI Training	Llama 2 70B - 24.7 minutes or less on single node	Kindly change it to 27 mins	Please refer to corrigendum - 2
91	d Nodes for AI Training	ResNet - 12.1 minutes or less on single node	Kindly remove since we have not yet run this benchmark.	Please refer to corrigendum - 2
92	e Nodes for AI Training	RetinaNet - 34.3 minutes or less on single node	Kindly remove since we have not yet run this benchmark.	Please refer to corrigendum - 2
93	f Nodes for AI Training	U-Net3D - 11.6 minutes or less on single node	Kindly remove since we have not yet run this benchmark.	Please refer to corrigendum - 2
94	27 Nodes for AI Training	Benchmarks	Owing to limited time, kindly allow us to commit the benchmark numbers and demonstrate at the time of acceptance.	Please refer to corrigendum - 2
95	27 Nodes for AI Training	Benchmarks	Kindly allow 5% tolerance in the timings for both the Training and Inference nodes during acceptance, as a standard worldwide practice.	Please refer to corrigendum - 2
96	29 Nodes for AI Training	Benchmarks - for Inference	We request GIL to review the benchmark requirements. The target timings are pretty high for a node with 2 x Accelerators. These timings are similar to 8 x GPU numbers based on what we see in MLCommons website.	Please refer to corrigendum - 2
97	Master Node	Power Supply: Appropriate energy efficient redundant (N+N) hot swappable 80 Plus Titanium power supply and FAN	Appropriate energy efficient redundant (N+N) hot swappable 80 Plus ~~Titanium~~ **Platinum** power supply and FAN **Justification- OEM Specific**	Please refer to corrigendum - 2
98	6 Inference Node	Internal Storage: For Operating System: minimum 2*1.92 TB M.2 NVMe drives	For Operating System: minimum 2*1.92 TB ~~M.2~~ NVMe drives.	Please refer to corrigendum - 2
99		Internal Storage: Minimum 4 * 3.84 TB U.2 NVMe drives	Minimum 4 * 3.84 TB ~~U.2~~ NVMe drives	Please refer to corrigendum - 2
100	8	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.	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.	No Change

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
101	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).	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).	No Change
102	9 Inference Node	Power requirements: 2000W or more hot plug & redundant power supply 80 PLUS Titanium	~~2000~~ 1600W or more hot plug & redundant power supply 80 PLUS ~~Titanium~~Platinum	Please refer to corrigendum - 2
103	24	Generic Query External Storage	External Storage: Throughput is for large block sizes (1MB) in AI/HPC workload whereas IOPS is for small block size (4/8KB). The Nvidia RA documents refers only THROUGHPUT no's based on no of GPUs and is 100% READ or 100% WRITE. There are no references to IOPS in NVIDIA SUPERPOD documents. Where is the IOPs requirement coming from? What is the Block size for IO operations? Also clarify if the throughput requirement is for 100% Read or 100% Write.	Please refer to corrigendum - 2
104	PQ	The OEM should have average Annual Turn-over of minimum Rs. 250 crores for the last five financial years from the last date of bid submission with positive net worth.● Audited profit and loss statement and balance sheet ● Auditor certificate clearly specifying the turnover and positive net worth.	The Hardware OEM should have average Annual Turn-over of minimum Rs. 250 crores for the last five financial years from the last date of bid submission with positive net worth.● Audited profit and loss statement and balance sheet ● Auditor certificate clearly specifying the turnover and positive net worth.	As per corrigendum dated 15.02.2025
105	Hypervisor	Hypervisor with Enterprise level highest license and support available should be provided from day one.	Hypervisor and Container Platform with Enterprise level highest license and highest level of support available should be provided from day one including, Service Mesh, Local Image registry, Logging & Monitoring and tracebility / Obersevability.	No Change
106	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.		No Change
		All necessary and required software, SDK, libraries, tools to cater and run the AI/ML workload should be provided from day 1. Bidder to ensure that enterprise level OEM support & SLA is available for all OEM provided software and libraries. Licenses required must be included and shall be perpetual with no scaling restrictions.	Requesting GIL to modify as "Perpetual /Subscription "	Please refer to corrigendum - 2
		Some of the basic, SDK/library/containers to be used in the system are:		No Change
		a. CUDA toolkit,	Requesting GIL to Remove this as well as this pure nvidia tech.	
		b. CUDA tuned Neural Network (cuDNN) Primitives	Requesting GIL to Remove this as well as this pure nvidia tech.	
		c. TensorRT Inference Engine	OEM Specific, TensorRT inference engine is Nvidia specifc, there are far better inference engine like vLLM. Request to please modify as TensorRT/vLLM	
		d. CUDA tuned BLAS (cuBLAS)	Requesting GIL to Remove this as well as this pure nvidia tech.	
		e. CUDA tuned Sparse Matrix Operations (cuSPARSE)	Requesting GIL to Remove this as well as this pure nvidia tech.	
		f. Multi-GPU Communications (NCCL)		
		g. Industry SDKs – NVIDIA Merlin, DeepStream, ISAAC, Nemo, Morpheus	Industry SDKs- this points to nvidia NIM catalog. Again nvidia influenced. As this pure nvidia tech, please Modify this as "industry SDKs - NVIDIA Merlin, DeepStream, ISAAC, Nemo, Morpheus or other opensource OpenSource / OpenAI compliant with entreprise support"	
		h. Rapids, Tao, TensorRT, Triton Inference	Requesting GIL to Remove this as well as this pure nvidia tech, Request to mention generic requirement like. enterprise data management, Model Training framwork, Model Inference Engines	
		Bidder needs to submit proof of the quoted GPU meeting these MIcommons training benchmarks at the time of bidding :-		

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
107	Benchmarks	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	Offered Nodes should be listed under ML Commons Training (4.0 or higher) for the mentioned Benchmarks, supporting published link for any of the below 2must to be shared during bid submission	Please refer to corrigendum - 2
		Specifications:		
		a) BERT - 5.3 minutes or less on single node	https://www.redhat.com/en/blog/accelerating-generative-ai-adoption-red-hat-openshift-ai-achieves-impressive-results-mlperf-inference-benchmarks-vllm-runtime Supporting material :https://www.cerebrium.ai/blog/benchmarking-vllm-sglang-tensorrt-for-llama-3-1-api this is on the basis of the A100. We need to work with the bidder (including the Hardware OEM and GPU Eg H100 etc)	
		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	This is on the basis of the ResNet -50 https://www.redhat.com/en/blog/quantifying-performance-red-hat-openshift-machine-learning this is on the basis of the H100. We need to work with the bidder (including the Hardware OEM and GPU Eg H100 etc) e) ResNet 50 – 22.1 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		
108	**OS Support**	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux server Quoted OS should be under Enterprise support from OEM.	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.	please refer corrigendum-02.
109	**Hypervisor**	Hypervisor with Enterprise level highest license and support available should be provided from day one.	Hypervisor with Enterprise level highest license and support available should be provided from day one and must have a native open source LLM (large Language model) bundle in the proposed hypervisor. Since DIT/GSDC is planning to have inferencing / training on the proposed hardware as per Specs.	No Change
110	**AI Enterprise Software**	AI Enterprise software & subscription or equivalent for each and every GPUs to be included from day one of Installation. Software Support All necessary and required software, SDK, libraries, tools to cater and run the AI/ML workload should be provide from day 1.	AI Enterprise software & subscription or equivalent for each and every GPUs to be included from day one of Installation. Software Support All necessary and required software, SDK, libraries, tools to cater and run the AI/ML workload should be provide from day 1.bias and drift detection, better access to accelerators, and a centralized registry to share, deploy, and track models. the proposed solution can be open-source with enterprise support.	No Change
111	**MLOps**	New Specs to be Added.	the Proposed solution must have a MLOps, Pipeline for the project with native opensource workspace like Jyupter notebook, Pytorch Apache Spark etc. Tight integration with opensource supported CI/CD tools must be provided that will allows ML models to be quickly deployed iteratively, as needed by Govt entity / Dept.	Using appropriate tool- MLops practices and principles should be followed under training model without any additional cost to tenderer.
112	Open Source Model with indemnification	New Specs to be Added.	The proposed DSML platform should provide enterprise support and model IP indemnification for open source-licensed large language models (LLMs) from vendor providing enterprise open source support.	No Change

Responses to Pre-bid Queries				
Procurement of GPU Compute solution for Gujarat State Data Center, Gandhinagar (GeM No. GEM/2025/B/5824041 dated 16.01.2025)				
Sr. No.	**Page No. / Section No. / Clause No.**	**Tender Description**	**Query / Clarification / Suggestions from the Bidder**	**Responses of pre-bid Queries**
113	Eligibility Criteria: 4 | Page No: 1	The bidder should have experience of set up of similar GPU base solution with GPU or CPU with below criteria in last 5 Years in India as on last date of bid submission: • One project having similar works total value of INR 40 Cr (Should have min 45 GPU / 800 CPU) or • Two project having similar works total value of INR 25 Cr (Should have min 28 GPU / 500 CPU) or • Three project having similar works total value of INR 20 Cr (Should hav min 22 GPU / 400 CPU) Note: Similar works means SITC OF GPU ACCELERATED with multiple GPU Node.	The bidder should have experience of set up of similar GPU base solution with GPU or CPU with below criteria in last 5 Years in India as on last date of bid submission: • One project having similar works total value of INR 40 Cr (Should have min 45 GPU / 800 CPU / 2000 cores) or • Two project having similar works total value of INR 25 Cr (Should have min 28 GPU / 500 CPU / 1500 cores) or • Three project having similar works total value of INR 20 Cr (Should hav min 22 GPU / 400 CPU / 1000 cores) Note: Similar works means SITC OF GPU ACCELERATED with multiple GPU Node.	Please refer to corrigendum - 2
114	8. IMPLEMENTATION TIMELINES & PENALTIES | Page No: 8	Supply of the Hard-ware including Li-censes and OEM Warranty Certifi-cate. T1=T+45 days from the date of issuance of contract over GEM	The delivery timeline should be extended from 45 days to atleast 120 days due to the current market trend and high lead times for GPU systems delivery.	Please refer to corrigendum - 2
115	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.	As AI Deployment will be majorly on Containerzation and Resources should be assigned to those containerization, we would recommend that the Bid should ask for Sharing of GPU Across Multiple VMs and Containers via Dedicated GPU, Virtual GPU, MDI, Time-Slice and Fractional Allocation.	Please refer to corrigendum - 2
116	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	As AI Deployment will need central management and monitoring of all the Resources deployed in the AI Cluster, we would urge you to specify the Management and Monitoring for the Complete Deployment Architecture / AI Landscape.	As mentoned in RFP and time -to-time published corrigendum

Revised RFP (Corrigendum-02) dated 15.03.2025
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.	• 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 within the last five years in India, 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
4.1	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.	Copy of Work Order along with Completion / Go-Live certificate.
5.	The OEM should have executed similar GPU setup for min 3 clients in last 5 Years in India as on date of bid submission. Out of which One client deployment should be One project having similar works total value of INR 125 Cr. Note: Similar works means SITC OF GPU ACCELERATED with multiple GPU Node.	Copy of Work Order along with Completion / Go-Live certificate

6.	The bidder should provide the authorization certificate from the OEM for a. Quoting the requirement and subsequent support for Hardware and Software (and) b. Proposed GPUs solution will not be End of Life (EOL) for 5 years from the date of installation	In Case of SI, should submit Manufacturer Authorization Form. In Case of OEM, Letter of Declaration on their letter head
7.	Neither OEM nor bidder should be blacklisted from supplying equipment to any Government/PSU/BFSI within India in the past.	Certificate of Undertaking for Non-blacklisting from supplying equipment to any Government/PSU/BFSI within India in the past.
8.	A Power of Attorney / Board Resolution in the name of the person signing the bid document.	Original Power of Attorney / Board Resolution Copy on a non-judicial stamp paper.

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

2. <u>Clarification on Bidding Documents</u>

- A prospective Bidder requiring any clarification of the bidding documents may seek clarifications by submitting queries on email Id: **mgrhninfra1-gil@gujarat.gov.in, dgmTech-gil@gujarat.gov.in** prior to the date of Pre-Bid Meeting.
- Tenderer will discuss the queries received from the interested bidders in the Pre-Bid Meeting and respond the clarifications by uploading on the website https://gil.gujarat.gov.in.
- No further or new clarification what so ever shall be entertained after the Pre-Bid Meeting.
- The interested bidder should send the queries as per the following format:

Bidder's Request For Clarification			
Name of Organization submitting Request		**Name & position of person submitting request:**	**Address of organization including phone, fax, email points of contact**
Sr. No.	**Bidding Document Reference (Clause /page)**	**Content of RFP requiring Clarification**	**Points of Clarification required**
1			
2			
3			

3. <u>Scope of Work:</u>

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

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

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

1. No advance payment will be made to the bidder.

2. 70% of the Capex cost shall be paid within 30 days after POC (proof of concept) which is successful demonstration of the benchmark as 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. 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	EMD may be forfeited and contract	-	

		date of issuance of GEM 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 , LLD documents	T2=T1+30	0.5% of Capex value of delayed/pending work per week or part thereof	10% of GEM order value	
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)	T4=T3+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 4 hours from the time of call logged / reported to Bidder/OEM. If the successful bidder fails to resolve the call as specified above, penalty will be imposed on each delayed 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 Manage-ment Interface
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.
Power Supply	Appropriate rated and energy efficient, redundant (N+N) hot swappable power supply and FAN
Failure Alerting Mechanism	Should be able to alert upcoming failures on maximum number of components such as Processor, memory, HDDs and expansion cards, etc.
OS Support	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux server Quoted OS should be under Enterprise support from OEM.
Hypervisor	Hypervisor with Enterprise level highest license and support should be provided from day one.
Software Support	All necessary and required software, SDK, libraries, tools to cater and run the AI/ML workload should be provide from day 1.

Scalability, Cluster and Management Hardware and software	System should be scalable with multi node cluster. Software support & cluster tools and management hardware and software and licenses to be supplied along with product.
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 1.92 TB NVMe drives

	• For Data: Minimum 8 * 3.84 TB U.2 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+N) hot swappable power supply and FAN
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 server Quoted OS should be under Enterprise support from OEM.
AI Enterprise Software	AI Enterprise software & subscription or equivalent for each and every GPUs to be included from day one of Installation. Software stack to be supported by GPU OEM for 5 years for each system. All necessary and required software, SDK, libraries, tools to cater and run the AI/ML workload should be provided from day 1. Bidder to ensure that enterprise level OEM support & SLA is available for all OEM provided software and libraries. Licenses required must be included and shall be perpetual/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 2*1.92 TB ~~M.2~~ NVMe drives Minimum 4 * 3.84 TB U.2 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+N) hot swappable power supply and FAN
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 *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. Switch must support EVPN – VxLAN based network. Required cables of appropriate length, transceivers should be supplied. Switch should have redundant Power Supply. 5 Years Comprehensive onsite Warranty.
OS Support	The system should support latest version of Red Hat Enterprise Linux / Ubuntu Linux / RHEL AI / Red Hat OpenShift AI server, Quoted OS should be under Enterprise support from OEM with premium or highest level of support. 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, 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	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

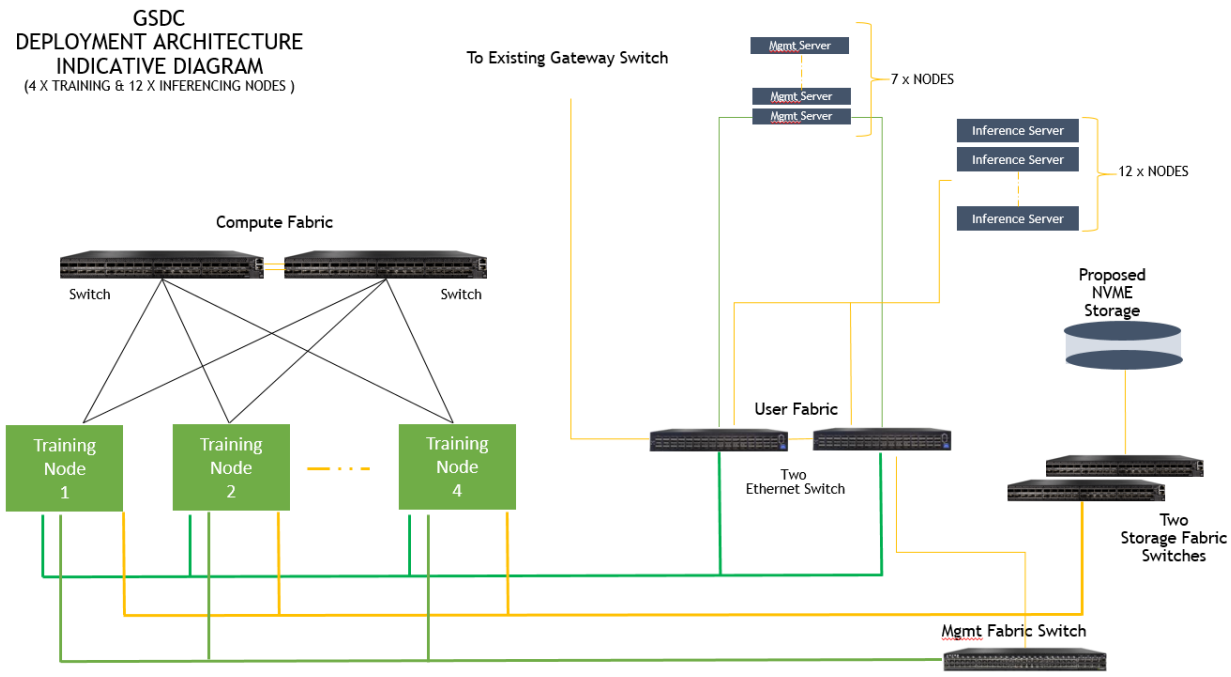
should be on OEM letterhead duly signed and referring the bidder and bid details.	Throughput for single NODE inference (99% offline) = 105000 Samples /s or higher Object Detection RetinaNet Throughput for single NODE inference (99% offline) = 1500 Samples /s or higher Image Generation Stable Diffusion-XL 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) 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 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.

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 & Construc-tion	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 An-gle	The 19” mounting angles should be provided 2 Nos. on front and rear side of the Rack. It should be adjustable full depth. 19” Mounting Angles made up of Steel 2mm Thickness with better mounting flexibility and maximizes usable mounting space.
5	"U" Identifica-tion	“U” numbering should be provided on the 19" mounting rails such that these unique 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 Man-ager Provision	Each rack should have 4 horizontal 1U Closed type cable manager.
8	Side Panels	Side Panel shall be covered with horizontally split steel panels The side panels should be easily detachable with locking provision.
9	Door	Front and Rear doors should be perforated and both front and rear doors should be at least 80
		% hexagonal perforated (Holes). Front & Rear Door should be with Minimum of
		138 degrees to allow easy access to the interior.
10	Door Perfora-tion	Hexagaonal Perforated Single Front Door will be Lockable and - handle Lock & Key should be provided.
11	Door Lock	Hexagonal Perforated Dual Rear Door will be Lockable and 3 Point Lock should be provided.
12	Castor	Rack should be with Plinth of 800 MMW, 100MM H and 1400 MM D. The rack shall be not having External height >2060 mm including Plinth.
13	Load Bearing	Minimum load bearing capacity supported by Base Frame should be static load of at least -1200 Kg.
14	Powder Coat-ing	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 Con-struction	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 reversi-ble so that it may open from either side.
18	Statutory Standard	100% assured compatibility with all equipment conforming to DIN 41494 / EIA 310-D standard(General industrial standard for equipment).
19	Certification	The rack shall be from OEM having ISO9001:2008,
		ISO14001:2004, ISO 45001:2018 ~~& ISO 50001:2018~~
		(Certificate to be submitted along with compliance)
21	Warranty	5 years onsite comprehensive warranty

Indicative Diagram



Note:

- Bidders should refer to the indicative diagram for reference and propose their 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

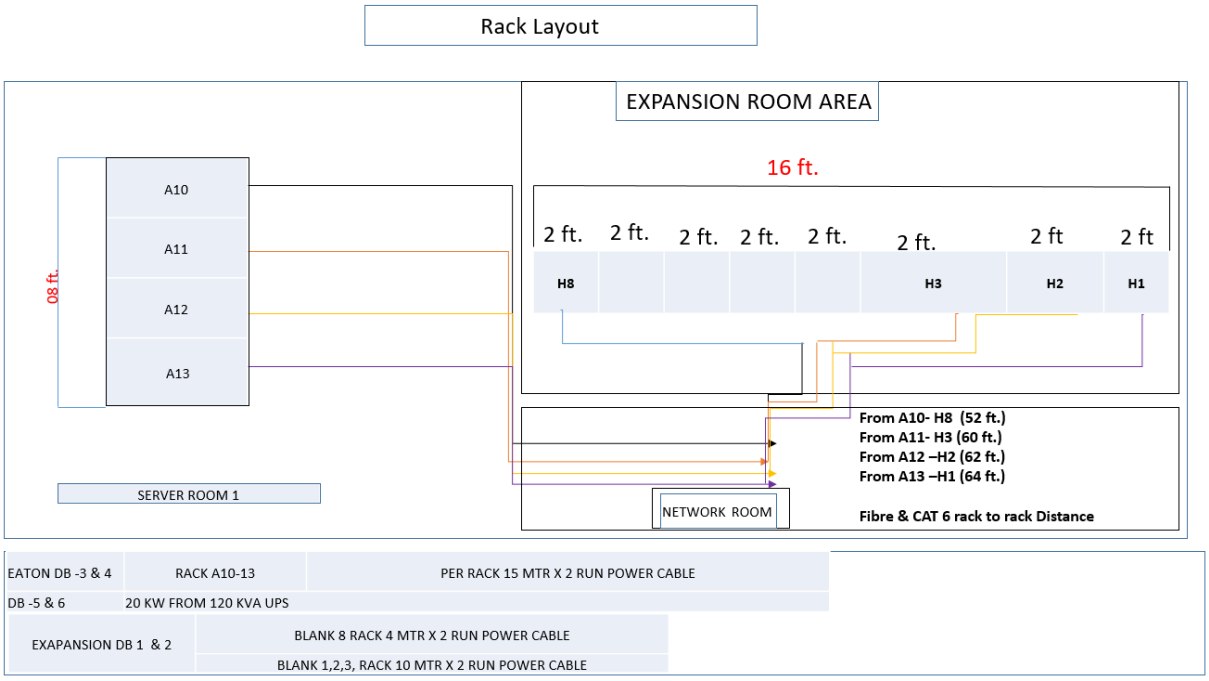
- **NetApp FAS8300 with Total ports 4 and Used ports 2**
- **Hitachi VSP 5600 with Total ports-64 ports and Used 64 ports**

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.



EATON DB -3 & 4	RACK A10-13	PER RACK 15 MTR X 2 RUN POWER CABLE
DB -5 & 6	20 KW FROM 120 KVA UPS	
EXAPANSION DB 1 & 2	BLANK 8 RACK 4 MTR X 2 RUN POWER CABLE	
	BLANK 1,2,3, RACK 10 MTR X 2 RUN POWER CABLE	

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>>**