**RFP for GPU Server for DC with 2x NVIDIA L40S 48GB GPU card for 3 years warranty & 2 years AMC.**

**Deliver lead time:** 7 days to deliver to Brac Bank DC premises for Server & GPU card.

**Technical Specification:**

**Item Table:**

1. **GPU Card: 2 Qnty**

|  |  |  |  |
| --- | --- | --- | --- |
| **SL** | **Item** | **Unit** | **Total** |
| 1 | GPU Server for DC with 2x NVIDIA L40S 48GB GPU card with necessary kit. | Nos | 1 |

1. **Server: 1 Qnty**

| **SL** | **Item Name** | **Detail Required Specification** | **Bidder response** |
| --- | --- | --- | --- |
|  | Timeline | **Ready Stock** |  |
| 1 | Brand | HPE |  |
| 2 | Certification Quality | ISO 9001:2015 for the manufacturer, CE certification for quality assurance. |  |
| 3 | Model | HPE ProLiant DL380 Gen11 |  |
| 4 | Country of origin | To be mentioned by the bidder |  |
| 5 | Manufacturing Country | To be mentioned by the bidder |  |
| 6 | Form Factor | Bider must provide Rack Mountable Server with Rail Kit, Cable Management, and Bezel Kits. |  |
| 7 | Processors | 2x Intel Xeon-Gold 6526Y 2.8GHz 16-core 195W Processor |  |
| 8 | Cache L3 | 37.5MB |  |
| 9 | Chipset | To be mentioned by the bidder |  |
| 10 | Memory | Minimum 8x 64GB Dual-Rank 5600M/s modules |  |
| 11 | Storage array controller | Integrated Hardware RAID controller with 4GB cache supporting RAID 0, 1, 10, 5, 6 supporting SAS/SATA/NVMe devices. |  |
| 12 | Internal Storage | Shall be provided with a minimum of 3x 1.92TB NVMe 2.5-inch hot-pluggable SSD. |  |
| 13 | OS Boot Drive | Shall be provided with minimum 2x 480GB NVMe M.2 SSD in Hardware RAID1. |  |
| 14 | GPU | Minimum **2x NVIDIA L40S 48GB GPU accelerator card** from day1 and future scalable up to 3x NVIDIA L40S 48GB GPU. |  |
| 15 | Container & Orchestration Support | Must support Kubernetes, Docker, and NVIDIA GPU Operator out-of-the-box. |  |
| 16 | Implementation & Deployment | The deployment must include NVIDIA (NVDA) container services and any other GPU-related components necessary to fully enable GPU functionality in containerized environments (e.g., Docker, Kubernetes). Deployment must be performed by the OEM or bidder or an authorized partner of the OEM. The deployment method must follow the best practices recommended by NVIDIA |  |
| 17 | Resource Pool | The solution must support the deployment of multiple physical GPU cards grouped under a unified GPU resource pool. From this shared pool, GPU resources (different vGPU profile) must be dynamically assignable to virtual machines or containers based on workload demand. |  |
| 18 | GPU Framework Compatibility | Should support major ML/AI frameworks: TensorFlow, PyTorch, ONNX, RAPIDS, CUDA, and cuDNN libraries. |  |
| 19 | Power supply and Fan kit | Minimum 2x 1600W hot-pluggable power supply unit with hot-swappable redundant fan kit. |  |
| 20 | Network Card | Each server should be provided with-   * Minimum 1x 4-port 1G NIC. * Minium 2x 2-port10/25G SFP28 adapter with 25G with necessary SFP28 SR transceivers server end. * Minimum 1x 2-port 32Gbps FC HBA. |  |
| 21 | Expansion Slot (PCIe) | Minimum 8 PCIe slots with version 5.0. |  |
| 22 | Embedded Remote Management and firmware security | iLO advanced license |  |
| 23 | Server Management | HPE OneView software license |  |
| 24 | Hypervisor & Operating System Support | Should support the following operating systems and hypervisors   * Microsoft Windows Server * Red Hat Enterprise Linux * SUSE Linux Enterprise Server * VMware ESXi * Canonical Ubuntu * Citrix |  |
| 25 | Warranty & AMC | * Minimum 3Years OEM warranty & 2 years AMC * OEM must maintain an in-country local part depot center; authorization letters must be submitted. * The OEM manufacturer authorization letter must be submitted. |  |

**Annexure 1**

**High level Service Level Agreement (SLA):**

|  |  |  |  |
| --- | --- | --- | --- |
| Severity Level | General Evaluation Criteria | Response Time (24/7/365) | Resolution Time (24/7/365) |
| High | A problem that critically effects on ability to do regular business. System functionality is unavailable or down due to hardware or software failure. | By Email/Phone: 01 Hour  Onsite: 02 Hour | 04 Hours |
| Medium | Any Hardware or software component failure that leads to degradation of system performance. | 08 Hours |
| Low | System is having minor issues. Insignificant error without impacting usual business or system performance. | 24 Hours |

**Penalty Clause on Service Failure:**

0.5% of the work order value will be penalized for each failure to meet service level.

If resolution time is greater than agreed service level, then 0.025 % (of the work order value) will be penalized for each hour delay for severity level ‘High’ and 0.007 % (of the work order value) will be penalized for each hour delay for severity level ‘Medium’ and ‘Low’.

**Scope of Warranty Service:**

The scope of service is as below but not limited to:

* The product will be delivered as per RFP (All HW, GPU card, GPU necessary SW and other HW & SW accessories.
* Complete implementation & deployment all HW, SW, GPU Card (NVIDIA with necessary SW) & Container & Orchestration as per RFP in the Brac Bank premises will be carried out by the bidder along with OEM & local partner.
* Warranty for Server, GPU Card & all necessary software of NVIDIA mention in the RFP for 3 years and 2 years AMC.
* Install, configure in the Server systems as per the requirement of the Bank
* Maintain provided server, GPU card and software 24/7/365 days without any interruption.
* Provide patch, firmware, software upgradation, Security Enhancement, root cause analysis, VA mitigation of all HW & SW as per RFP.
* Pro-active support and notification from OEM or Local Partners for emergency Security fixes and bugs.
* Complete parts replacement of server, GPU card replacement for faulty will be carried out by the bidder along with OEM & local partner.
* Bridge between OEM and BRAC Bank for knowledge transfer, problem troubleshoot and best use of the product. OEM back-to-back support for bidding products will be applicable.
* Ensure new integration of hosts/servers with GPU card as per the requirement of the Bank
* To ensure a single point of contact, backup contact and support contact escalation matrix
* To provide a quarterly report on the health status of the Server system along with performance forecast and performance improvement plan to ensure engineer support for any kind of server movement.