الجامعة الألمانية الأردنية
German Jordanian University

RFP for Designing, Delivering, and Commissioning of AI-Enabled Computing Solution

1. INTRODUCTION

GJU intends to acquire AI-Enabled Computing Solution at the main campus Almushaqar – Amman (Madaba Al Gharbi St.). Bidders are invited to submit their proposal for this solution. The required solution is divided into two integral parts:

- Part #1: AI-Enabled Hardware Solution.
- Part #2: On-Premises AI Software Solution.

2. GENERAL INSTRUCTIONS AND REQUIREMENTS

2.1 Bid Requirements

- The bidder must either present a reference project relevant to the tender's scope in Jordan or ensure that vendor-certified experts will carry out the implementation.
- Subcontracting or involving sub-bidders is not permitted; the selected bidder must execute the project independently.
- The chosen bidder must be a **Gold Certified Partner** of the vendor within the Hashemite Kingdom of Jordan and must provide only **new, vendor-certified equipment** (refurbished, renewed, or used equipment are not acceptable).
- The bidder must not propose any **End-of-Sale** or **End-of-Life** products or services. A detailed lifecycle for all proposed components must be included.
- All equipment and components must be from **reputable, well-known brands** with established service, support, and local partnership presence.
- The selected bidder will be held accountable for any damage caused to GJU's facilities or assets during the implementation phase.
- Both hardware and software must be registered under GJU's e-Account by the bidder/vendor.
- All proposed hardware must be **certified and/or qualified** for the specific Nvidia GPUs included in the proposal.

2.2 Format of Proposals

- A *Compliance Checklist Sheet* should be included in the Technical Proposal.
- The *Data Sheets* for Hardware, Software, Operating System (OS), and Licenses should be included in the offers.
- The Financial Proposal should quote the *Retail Price* for Hardware, Software, Operating System (OS), Licenses, etc.
- The proposed solution and design should be clearly illustrated and thoroughly detailed in the proposal.
- *Part Numbers* for each part of the solution hardware and software should be mentioned.

2.3 Scope of Work

The scope of work outlines the deliverables, responsibilities, and tasks required from the selected bidder to provide a complete AI-Enabled Computing Solution at GJU. This solution is divided into two integral parts: AI-Enabled Hardware Solution (Part #1) and On-Premises Software Solution (Part #2). The selected bidder is responsible for the following:

2.3.1 AI-Enabled Hardware Solution

a) Installation and Setup of Hardware:
- Transport, store, and assemble the hardware components.
- Install and configure the hardware (Servers, GPUs, Networking, etc.).
- Ensure all hardware is operational and integrated into GJU's existing infrastructure.
- Provide and install Fiber-Optics cables, SFPs, and cable spiral to secure redundant network connections to GJU core network.
- Provide a branded Power Distribution Unit (PDU) to secure redundant power distribution from two power sources via already existed industrial sockets.
- Install all equipment and any required accessories to produce a fully functioning system ready for handover to GJU.
- Any accessories that are needed for the installation and were not stated in the offer are the bidder's responsibility to provide at no extra cost.

b) Servers' Hardware:
- Supply and install two types of AI-enabled servers with specifications as detailed in the RFP (Processors, RAM, storage, GPUs, network interfaces, etc.).
- Ensure the servers are set up with GPU virtualization for eight concurrent users per server node.

c) Network Configuration:
- Ensure the servers are connected to the GJU network with appropriate and redundant network interfaces (10Gb and 1Gb Base-T ports).
- Provide any required networking accessories and cables for seamless integration.

d) System Management and Security:
- Configure Trusted Platform Module (TPM) or vTPM for VMs security.
- Implement and connect advanced or enterprise-level out-of-band management for system monitoring and control.

2.3.2 GPU On-Premises Software Solution

a) Operating System Installation and Configuration
- Provide and install an operating system capable of supporting CPU, RAM, IO, and disk virtualization, as well as Nvidia GPU virtualization.
- Configure the system to support at least eight concurrent guest (user/researcher) instances for both Windows and Linux environments.

- vGPU provisioning and direct assignment into virtual instances.
- Configure containerized application deployment and support dynamic scheduling and management of virtualized workloads

b) Licensing for Nvidia Software:
   - Provide Nvidia licenses that support AI, Deep Learning, and high-performance simulation workloads in Virtual Machines.
   - Ensure the solution supports Windows and Linux instances (VMs) and enables GPU resource allocation for at least eight concurrent instances.
   - Ensure Nvidia licensing is in place for the duration of the warranty period, with academic pricing applied.
   - Enable GPU resource sharing and isolation across multiple containerized and virtualized workloads.

c) AI, ML, DL, and Inference Workloads Software Solution:
   - Provide a software solution capable of supporting AI, ML, DL, and Inference workloads on GJU's hardware.
   - The solution should integrate all hardware components into a single dashboard for management and monitoring.
   - The software solution must support multi-user tenancy, with secure file segregation for each user, and resource management for GPUs, CPUs, RAM, and disk.
   - Perform testing to confirm GPU virtualization is functional and capable of handling AI, DL, ML, and Inference workloads.
   - Support container orchestration, enabling scalable deployment, monitoring, and lifecycle management of user applications.
   - Support workload scheduling and automated resource distribution across available compute nodes based on user demand and predefined policies

d) Integration and Scalability:
   - Ensure the solution supports the integration of Microsoft Active Directory for Single Sign-On (SSO).
   - Support for container management and ability to distribute workloads across multiple physical GPUs and CPUs.
   - Enable scalability, allowing for the addition of server nodes and hardware upgrades in the future.
   - Support multi-node orchestration, resource pooling, and cross-node workload migration for high availability
   - Ensure that the solution includes a dedicated database and reporting capabilities.

e) User and Resource Management:
   - Provide the capability to allocate resources to users based on customizable profiles (e.g., GPU, CPU, RAM, and storage configurations).

- Support advanced user needs, including shell commands, reporting, and utilization dashboards.
- Support namespace-level isolation and granular access control to maintain secure user environments and resource quotas.

f) Support and Maintenance:
   - Offer support for installation, configuration, and troubleshooting of the software solution.
   - Provide necessary software patches and updates throughout the lifecycle of the project.
   - The involvement of administrative tools for orchestrating containerized environments, managing infrastructure state, and updating workloads with minimal downtime.

g) Security and User Authentication:
   - Implement user authentication and role-based access control to ensure secure usage of the system.
   - Provide secure handling of user data and prevent unauthorized access.
   - Supporting policy-based access controls, logging, and auditing of user and system activity for compliance and security.

2.3.3 General Responsibilities:

a) Coordination and Communication:
   - Maintain regular communication with GJU team during the implementation phase.
   - Ensure that all deliverables are provided on time and meet the specified requirements.

b) Testing and Validation:
   - Perform comprehensive testing on both hardware and software to ensure compatibility, performance, and compliance with the specifications.
   - Demonstrate and validate that the virtualized GPU environment supports the specified workloads and user concurrency.

c) Documentation:
   - Provide complete documentation, including technical specifications, installation guides, and system management instructions.
   - Deliver comprehensive documentation for configurations.
   - Ensure that all necessary training materials are provided for GJU staff to manage the system post-installation.
   - High-Level Design (HLD) and Low-Level Design (LLD) documentation.

d) Training and Knowledge Transfer:
    - Provide training sessions for GJU staff on how to use and manage the system, including both hardware and software components.

e) Consultation and Reporting
    - AI consultation for best practices implementations during support period.

f) Support and Maintenance
    - Signing the Service Level Agreement (SLA)
    - Continuous updates and security patches.

g) Handover and Final Acceptance:
    - Complete the installation, configuration, and testing to GJU's satisfaction before the final handover.
    - Ensure that all systems are functioning as expected and meet the specified requirements.

3 PART #1: AI-ENABLED HARDWARE SOLUTION

3.1 Introduction

This part encompasses exclusively the servers' hardware. GJU is seeking two types of AI-enabled servers (L40S and H100). The overall solution (tender) will be accepted only after installation, and only if GPU virtualization for eight users per server node is successfully achieved, enabling AI, DL, ML, and Inference workloads comparable to physical GPUs. Alternatively, the Software solution must be implemented, operated, tested, and approved according to the minimum requirements outlined in Part #2.

3.2 AI-Enabled Solution Hardware – Server #1

Quantity	1
Processor	≥ 2x Intel Xeon Scalable Silver 4XXX 20C/40T, 35MB Cache
RAM	≥ 512 GB DDR5 RDIMM, 5600MT/s
Chipset	Intel Chipset
Storage	**Data Capacity** ≥ 8 TiB usable capacity after (RAID 6). **Type/Technology** ≥ NVMe Gen 4 Mix Use Drives (Disks). **RAID Controller** ≥ 6 GB Cache supporting all RAID levels (If Applicable).
GPU	**Type**: Nvidia L40S, PCIe, 48GB Passive, Double Wide, Full Height **Quantity**: Four (4) **Virtualization**: Required
Network Interfaces	≥ 2x Port 10Gb SFP+ including the SFPs ≥ 2x Port 1Gb Base-T ≥ 1x Out-of-Band NIC NIC Cooling: Passive
Security	Trusted Platform Module 2.0
Form Factor	2 U Rack mount with all needed accessories, media kits and cables to install the server in the cabinet and GJU datacenter network.
Power and Cooling	≥ Redundant Hot pluggable power supply ≥ Redundant Hot pluggable Air-Cooling Fans
OS Support	Windows Server 2016 ~ 2022 Ubuntu 14.04 ~ 24.04 RedHat Enterprise Linux VMware ESXi and Horizon
System Management	≥ (Advanced/Enterprise/Centralized) Out-of-Band Management
Brand	Well-Known Brand i.e., HPE, Dell, Lenovo, Fujitsu etc.
Warranty	**Option 1**: 3-Year Parts, Labor, Onsite support with next-day response. Mother Company warranty **Option 2**: 5-Year Parts, Labor, Onsite support with next-day response. Mother Company warranty

3.3 AI-Enabled Solution Hardware – Server #2

Quantity	1
Processor	≥ 2x Intel Xeon Scalable Silver 4XXX 20C/40T, 35MB Cache
RAM	≥ 512 GB DDR5 RDIMM, 5600MT/s
Chipset	Intel Chipset
Storage	**Data Capacity** ≥ 15.0 TB usable capacity after (RAID 6). **Maximum Disk Size**: 3.6 TB **Type/Technology** ≥ NVMe Gen 4 Mix Use Drives (Disks). **RAID Controller** ≥ 6 GB Cache supporting all RAID levels (If Applicable).
GPU	**Type**: NVIDIA H100 80GB PCIe **Quantity**: Four **Cooling**: Passive Cooling. **Interconnect:** NVIDIA NVLink Bridge **Virtualization**: Required
Network Interfaces	≥ 2x Port 10Gb SFP+ including the SFPs ≥ 2x Port 1Gb Base-T ≥ 1x Out-of-Band NIC NIC Cooling: Passive
Security	Trusted Platform Module 2.0
Form Factor	2 ~ 4 U Rack mount with all accessories needed, media kits and cables to install the server in the cabinet and GJU datacenter network.
Power and Cooling	≥ Redundant Hot pluggable power supply ≥ Redundant Hot Pluggable Air-Cooling Fans OR Server Embedded Liquid Cooling Module
OS Support	Ubuntu 14.04 ~ 24.04 RedHat Enterprise Linux VMware ESXi and Horizon
System Management	≥ (Advanced/Enterprise/Centralized) Out-of-Band Management
Brand	Well-Known Brand i.e., HPE, Dell, Lenovo, Fujitsu etc.
Warranty	**Option 1**: 3-Year Parts, Labor, Onsite support with next-day response. Mother Company warranty **Option 2**: 5-Year Parts, Labor, Onsite support with next-day response. Mother Company warranty

4 PART #2: ON-PREMISES SOFTWARE SOLUTION

GJU is looking for a software solution to activate and utilize the hardware solution in Part #1, this part is an integral part of Part #1, and the acceptance for this tender is conditional on achieving the whole tender cycle and minimum requirements, i.e. GPU, CPU, RAM, and Disk are available to eight concurrent users per server and securely segregate their files, activity, results, etc. The bidders are welcome to propose the software solution as mentioned in the following section (4.1).

4.1 GPU Virtualization Option

4.1.1 Operating System:

Bidders should offer operating system platform to be installed on the solution hardware (Host server) in their proposals as per the below minimum requirements.

- The OS should support CPU, RAM, IO, and Disk virtualization.
- The OS should support Nvidia GPU virtualization for AI, DL, ML, and Inference workloads.
- The OS should support eight concurrent guest (User/Researcher) instances.
- The OS should support Linux and Windows guest instances.
- The preferred option for the OS license is perpetual.
- The OS should support vGPU provisioning and direct assignment (Passthrough) into the virtual instance.
- Operating system licenses and options must be associated with a functional, non-personal email address.
- Operating system licenses and options are utilized at GJU, academic pricing should apply.

4.1.2 Nvidia Licensing:

Bidders should quote the below licenses in their proposals.

- Licenses solution should support AI, Deep Learning and Simulation (Compute-Intensive) workloads inside Virtual Machines, Host OS, Containers, etc.
- Licenses solution should support instance workloads running Windows and Linux.
- Licenses solution should be able to enable and allocate GPU resources for at least eight concurrent OS instances on each server node.
- Licenses should support High-end 3D visualization applications, AI training, and inference workloads.
- Licenses solution should include NVIDIA vWS (or equivalent) AND/OR NVIDIA AI Enterprise for Education (or equivalent).
- Licenses solution duration should at least cover the server solution warranty period and a perpetual option.
- Nvidia licensing must be associated with functional, non-personal emails.
- Nvidia licenses are utilized at GJU, academic pricing should apply.

4.1.3 AI, ML, DL, and Inference Workloads Support

An additional detail for running and serving ML, DL, High-end 3D visualization applications, AI training, and inference workloads as per the minimum requirements listed below:

- The solution should be installed on-premises, GJU main datacenter servers.
- The solution should support the common ML and DL environments.
- The solution should support OpenCV, Python, PyTorch, Keras, MXNet, TensorFlow, Nvidia Cuda, etc.
- The solution should connect the hardware solution in Part #1 into one dashboard.
- The solution should support a scalable server node for future addition.
- The solution should support multi-user tenancy.
- The solution must support integration with Microsoft Active Directory Domain Services (AD DS) to enable Single Sign-On (SSO).
- The solution should support time quotas and (GPU, CPU, RAM, Disk) resource quota and limitations.
- The solution should allow the users/tenants to upload their files to a segregated path to protect their files from other users/tenants.
- The solution should be accessed via a modern, easy Web UI.
- The solution should support containers to be managed centrally.
- The solution should manage the Host-OS, GPUs, RAM, Disk, and CPUs to be pooled to the admin user or any user/tenant.
- The solution should be able to distribute the workload to more than one physical GPU and/or CPU located on multiple servers.
- The solution should support shell commands for advanced users/tenants.
- The solution should support reporting, utilization dashboards, and lease statistics.
- The solution must include its own dedicated database.
- The solution should have the capability to create a resource profile, i.e., profile templates for user/tenant needs e.g., (1x L40s GPU OR virtual part of H100, 10 vCPU, 64 GB RAM, and 200 GB SSD).
- The solution should have a layer concept, i.e., Data layer, ML Layer, DL Layer, etc.
- The solution should support heterogeneous server platform technologies and generations, i.e., GJU can add different hardware vendor models and generations, in other words, the stack should support independent hardware in terms of generation, brand, and model.
- The solution must support container-based architecture capable of orchestrating workloads across multiple physical and virtual environments.
- The solution must support automated deployment, scaling, monitoring, and lifecycle management of containerized applications.
- The system must provide high availability and fault tolerance across multiple compute nodes.
- The solution should support declarative workload definitions and automated rollout/rollback capabilities for application updates.

- The platform should support integration with container registries and support secure container image management.
- The solution must support namespace-level isolation to ensure secure and segregated user or project environments.
- The system must provide granular role-based access control (RBAC) for users and administrators.
- The solution should support pluggable monitoring and logging frameworks for observability of containerized workloads and infrastructure components.

5. WARRANTY AND SUPPORT

The proposed solution (Hardware and Software) should be offered as per the options below. Bidders are required to provide pricing for all the following options:

- Option 1: 3-Year Parts, Labor, Onsite support with next-day response for Hardware, Software, Licenses, Mother Company warranty and support including the SLA.
- Option 2: 5-Year Parts, Labor, Onsite support with next-day response for Hardware, Software, Licenses, Mother Company warranty and support including the SLA.

Important Note: The overall evaluation and comparison of bids will be based on Option 1 pricing. Bidders must ensure that pricing for all options is fair, reasonable, and consistent. Any attempt to bias the pricing structure by disproportionately inflating the cost of Option 1 to influence selection toward Option 2 will be considered unfair and non-compliant, and such offers will be disqualified.

(Good Luck)