**采购需求**

标的名称：智能应用模型服务平台设备

|  |  |  |
| --- | --- | --- |
|  序号 |  参数性质 |  技术参数与性能指标 |
| 1 | ★ |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **货物****名称** | **技术参数** | **数量** | **单位** | **备注** |
| 1 | 智能应用模型服务平台设备 | **主要性能指标**：1、总体性能：平台节点支持横向扩展，本次可纳管节点≥3个（支持利旧节点）。2、单节点性能配置：2.1 处理器：≥2颗，单颗主频≥2.1Ghz、核数≥32；2.2 内存：容量≥512GB（8×64GB），DDR5；2.3 硬盘：≥2块2.5英寸960GB企业级SATA SSD盘（可扩容至8块），并支持后期扩容CPU直连NVMe SSD盘≥8块；2.4 RAID：≥1张RAID卡，缓存≥2GB，支持RAID 0/1/5/6；2.5 网卡：配置≥2个25G光端口（满配25G多模光模块）；2.6 GPU：配置≥8张GPU卡，单卡INT8|FPB算力≥280 TFLOPS，BF16|FP16算力≥140 TFLOPS，显存≥141GB HBM3，GPU内存带宽≥4.0 TB/s；2.7 管理要求：具备≥1个千兆独立的远程管理控制端口，可实现远程对服务器的完全控制，包括远程的开机关机、重启等操作，支持风扇、电源、温度、本地固件更新、错误日志、硬件告警等功能；2.8 质量认证：通过国家CCC检测认证；2.9 电源及其他：≥2+2冗余热插拔电源，机架安装导轨及电源线；**主要功能要求**：3、管理兼容性：支持兼容Nvidia Tesla系列、Quadro系列、GeForce系列等主流GPU卡；支持 CUDA 9.x至12.x版本、DTK-22.04至23.10、CANN；支持PyTorch 1.0-2.2、TensorFlow 1.8-2.15；兼容华为昇腾NPU、海光DCU、寒武纪MLU等主流国产化AI加速卡。提供平台与不少于五个国产厂商的AI加速卡相互兼容性测试认证证书（测试认证文件应由测试双方签章）；4、GPU资源池化能力；支持在同一GPU资源池中管理不同厂家、不同型号的GPU卡，支持卡级别的虚拟化，同一台GPU服务器内，支持虚拟卡和物理卡同时并存，并能任意切换任意卡的虚拟/物理状态。算力分配误差≤10%，GPU虚拟化功能算力损耗≤0.1 %；5、GPU资源切分能力：GPU算力自定义切分，GPU显存自定义切分，多种算力隔离模式，虚拟GPU资源动态调整。算力切分最小颗粒度≤1%，显存切分最小颗粒度≤ 1MB；6、 GPU调度能力：为容器调度指定GPU卡类型、型号和节点，多任务优先级设置及管理能力，队列调度能力。支持精细调度配额控制、调度策略设置；7、 GPU安全隔离能力：支持虚拟GPU卡强安全隔离，虚拟GPU卡中运行的服务异常，不影响同一张物理GPU卡下的其他虚拟GPU中的服务运行；8、环境部署：实现操作系统、容器环境等基础部署，提供DeepSeek大模型（DeepSeek-R1-671B）、开源前端页面、模型推理框架、开源向量数据库等组件安装配置；9、数据源管理：支持从S3/NFS/CIFS创建模数据源，并挂载到容器实例；支持对数据源设置共享范围，团队共同使用数据源；支持直接在平台上对数据源文件进行操作，包括文件上传、下载、删除、重命名等；（提供功能截图加盖公章）10、代码源管理：支持从GIT/S3/NFS/CIFS创建代码源，并挂载到容器实例，实现训练代码的共享和持久化；支持对代码源设置共享范围，团队共同使用代码源；支持直接在平台上对代码源文件进行操作，包括文件上传、下载、删除、重命名等；11、模型仓库管理：支持算法模型管理；支持同一个模型创建多个模型版本；支持对模型仓库设置共享范围，团队共同使用模型；支持直接在平台上对模型仓库文件进行操作，包括文件上传、下载、删除、重命名等；（提供功能截图加盖公章）12、AI推理任务管理: 需提供模型开发任务管理能力（支持创建和管理jupyterlab类型notebook开发任务、vscode类型notebook开发任务、在线WebIDE进入notebook开发机等）、模型训练任务管理能力（支持创建和管理pytorch、tensorflow、deepspeed单机和分布式训练任务等）、推理任务管理能力（支持查看推理任务的资源监控、日志和调度事件等）; （提供功能截图加盖公章）13、API要求：模型推理框架需支持北向管理API、业务API和监控API全面开放，可快速构建上层应用；14、提供平台设备三年原厂质保服务，提供原厂安装调试实施服务。 | 1 | 套 |  |

 |
| 2 | ★ | 其他要求：1.售后服务标准要求（1）3年 7×24设备维保服务（含原厂备品备件）。（2）负责培训校方维护人员，使维护工作人员能完全熟悉并掌握软硬件维护技能，及时排除一般的设备故障，培训内容包括：设备的操作和维护，硬件电路结构和原理，软件结构等。2.售后服务效率要求即时响应（包括电话响应）；电话响应无法解决时， 1 小时内到达现场。修复时间 4 小时内；如 4 小时内无法修复，应提供相应解决方案。 |
| 3 |  | 以上技术参数及商务要求必须满足，如有负偏离，则按废标处理。 |

**3.4商务要求**

**3.4.1交货时间**

采购包1：

自合同签订之日起 50 天内到货、安装调试并交付使用。

**3.4.2交货地点和方式**

采购包1：

西北大学数据中心机房。