附件2

软硬件相关参数

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **序号** | **货物名称** | **技术参数要求** | **单位** | **数量** |
| **1** | GPU卡（GPU计算节点服务器） | 1.▲国产化要求：国产化通用GPU加速卡，保证核心IP技术自主可控； (投标时提供中国软件评测中心出具的评估报告扫描件)2.▲算力理论峰值：≥30TFLOPS@fp32，≥20TFLOPS@fp16/bf16，支持多精度数据类型，支持标准/混精。(投标时提供集成电路测试专业技术服务平台出具的第三方检测报告扫描件，以及官网截图和产品说明书证明材料扫描件)3.▲内存：≥32GB HBM2；4.接口：支持PCIe 4.0×16 lane；5.▲支持主流深度学习开发框架，兼容国内外主流软件生态。支持直接CUDA API 兼容 cuBlas，cuSparse，cuDNN，cuFFT， cuRAND， NCCL 等函数库，支持直接编译原生CUDA程序；（投标时提供产品说明书证明材料扫描件）6.单卡功耗≤250W；7.▲通过国产深度学习框架的PaddlePaddle的III级兼容性认证；（投标时提供互认证证书扫描件）8.大模型支持：支持智源Aquila-CodeGen、智源CPM-LM、智源GLM-Large、清华ChatGLM、 清华 CogView 、OpenAI GPT-2 、Meta LLaMA 、BlinkDL RWKV-LM 等大模型算法，支持DP、TP、PP等分布式并行训练部署，支持3D训练部署；（投标时提供以上模型中至少五项大模型算法测试结果截图证明）9.▲投标时提供有效期内（半年内）的与国产大模型ChatGLM2- 6B的厂商适配认证证书扫描件；10.▲支持包括但不限于CentOS/Ubuntu和国产操作系统等多类型操作系统；（投标时提供麒麟、统信、方德、龙蜥国产操作系统的兼容互认证书扫描件）11.▲异构算力服务：在同一大模型训练任务中，国产算力能够与英伟达训练产品混合部署，统一调度。投标时提供部署方案证明材料扫描件。 | 1,600 | 张 |
| **1** | **IaaS管理** | 1.支持IaaS/PaaS/SaaS一站式部署。**2.▲支持集群运行数据监控，包括但不限于CPU/GPU使用率、内存/显存使用率、磁盘读写速率、磁盘使用率、网络吞吐量，CPU负载、磁盘告警阈值，功耗，温度等相关监控数据设置阈值。(投标时提供至少五项相关功能界面截图)** 3.支持通过容器和虚拟机方式部署集群；投标时提供产品手册扫描件。4.支持数据监控对集群告警信息进行及时上报和处理；投标时提供产品手册扫描件。5.支持通过管理员身份登录管理平台进行运维，包括故障分析、日志查看和故障恢复等；投标时提供产品手册扫描件。6.支持分布式存储和超融合存储方式部署集群。**7.▲支持监测系统异常并自动恢复。投标时提供故障恢复统计截图。**8.支持将开发环境保存为私有镜像和转公有镜像功能。9.提供至少3种框架下的基础镜像。10.提供至少4种操作系统等多类型操作系统镜像。 | 1 | 项 |
| **2** | **PaaS&SaaS管理** | 1.**▲支持PyTorch/TensorFlow/PaddlePaddle/MindSpore等国内外主流算法框架；支持X86/ARM/LoogArch等CPU框架；支持国产GPU算力和国外主流GPU算力混合部署、统一调度，模型及环境可以无缝切换使用；投标时提供产品手册扫描件。**2.**▲支持应用市场，可根据不同算法框架和模型类型通过可视化界面快速上线推理任务；投标时提供产品手册扫描件。**3.**▲支持三维数字孪生模块，可基于平台进行通过可视化界面三维建模、在线预览，支持CPU类3D建模算法在GPU上并行加速；支持基于3D建模技术的模型差异检测；支持基于AI图像生成、大语言模型、语音文本互转、音唇同步算法的数字人方案；投标时提供产品手**册扫描件**。**4.支持基于分布式训练框架，在多机多卡环境进行大模型训练。投标时提供产品手册扫描件或视频截图。**5.▲支持动态挂载卸载算力，挂载卸载过程中对用户数据和环境无影响。投标时提供产品界面截图。**6.支持虚拟化，投标时提供产品手册扫描件。7.支持权限管理，包括用户账户增删改查、账户对于平台子功能权限限制和用户角色授权等；投标时提供产品手册扫描件。8.支持算力申请时长功能，到期后算力自动回归资源池；投标时提供用户使用手册扫描件。9.支持Jupyter开发工具，支持对当前开发环境的CPU使用率、内存使用率、GPU使用率等数据通过可视化图标方式展示；投标时提供用户使用手册扫描件。10.支持通过SSH远程登录云端开发环境，包括通过Terminal、VS Code、PyCharm等方式；投标时提供用户使用手册扫描件。11.可支持通过公有云方式上线运营算力能力：投标时提供使用手册扫描件。 | 1 | 项 |