**GPU服务器**

**1.1GPU服务器**

1）规格：机架式服务器，高度≤ 4U；

2）处理器：每台服务器配置≥2 颗EPYC 第三代处理器，每颗处理器核心数≥64核，基础主频≥2.2GHz，总浮点计算性能≥4.4TFLOPS；

3）高效数据系统：由处理器直接数据读取，工作主频≥3200MHz，可分配容量每个处理核心至少需要8GB的容量；

4）硬盘：每台服务器配置≥4块 15.36TB NVMe U.2企业级 SSD 硬盘；

5）硬盘扩展：前置支持≥8个热插拔3.5”硬盘位，其中可支持≥4个NVMe 硬盘位;

6）PCIE 扩展：每台服务器最大支持≥11个PCIe 4.0插槽，提供产品规格彩页证明

7）算力单元：额外增加算力单元，数量≥8个，采用主动散热，单个可提供≥82.5TFLOPS Peak FP32性能，缓存不低于24GB；

8）电源：≥2200W的2+2冗余供电模块。

9）网口：每台服务器配置≥ 1个4口万兆光纤网卡（含2个多模光模块和2个光转电模块）。

10）具备3C认证证书，提供复印件证明；

11）具备中国节能认证证书，提供复印件证明；

12）为保证技术服务专业性，设备生产厂商需为算力单元芯片厂商精英级合作伙伴，提供算力单元芯片厂商官网截图证明，且具备10名以上算力单元厂商认证的技术人员，提供证书复印件证明；

13）设备生产厂商具备ISO9001，ISO27001，ISO14001，ISO45001提供证书复印件；

14）服务器要求全新，产品为全配置装箱，在厂商整机配置清单上包含全部配件, 必须保证产品是同一批次出厂，通过官方的400服务热线可查询到配置详细信息，必须和采购要求匹配一致。

15）设备支持如下功能：

1. 资源虚拟化：采用轻量级容器虚拟化技术，实现对CPU、内存、磁盘等资源的虚拟化和统一管理。针对人工智能领域的特定需求，提供GPU等异构计算资源管理接口，实现对GPU等异构计算资源的虚拟化统一管理，支持为容器以直通方式挂载GPU等异构计算资源；

2．容器管理：支持快速创建多种深度学习开发调试环境的容器，支持web Terminal 访问容器（无需安装ssh服务），支持将创建的容器在线进行镜像打包，并支持将打包好的镜像上传镜像仓库，实现镜像版本的持续更新，提供截图证明；

3.算法开发：平台整合Jupyter、VSCode功能，用户访问增加权限控制，支持vnc功能，用户可以在平台上直接访问容器桌面环境，支持TensorBoard，支持RDMA，为保证安全性，其中远程桌面、Jupyter、TensorBoard、VSCode支持端口号和密码设置，提供功能截图证明；

4.底层使用非k8s或非基于k8s等开源调度系统二次开发，避免k8s中pod因内存或硬盘超过配额限制而重启的问题，提供厂商盖章承诺书或技术证明资料；

5. 资源限制：限制用户，用户组整体的资源；

6. 自动释放：可对运行中任务CPU、GPU使用率阀值限制，低于限制自动关闭任务，提供功能截图证明。；

7．排队功能：用户使用资源不足时，未超过单个资源池限制，启动排队功能，默认根据用户提交时间排队，管理员可修改用户排队顺序，提供功能截图证明。

8. 登录/操作日志：支持记录用户的登录时间、登录状态、登录IP、使用浏览器等；支持记录用户操作，包含功能名称、操作人、请求方式、操作状态、时间等，提供功能截图证明。

9. 文件共享与隔离：支持数据共享与数据隔离，即同一数据可供多用户同时访问，不同用户有自己的私有空间。也可以设定不同用户的访问权限，提供截图证明。

10．镜像管理：支持私有镜像仓库，集中化管理用户的镜像。能够提供新建项目、设置用户权限等功能，提供功能截图证明。

**1.2 GPU显卡**

配置数量不少于4个GPU显卡，采用主动散热，单个可提供≥82.5TFLOPS Peak FP32性能，缓存不低于24GB；配置GPU套件线缆、适配8卡服务器。

**1.3硬盘**

配置数量不少于2个容量高达15.36TB NVMe U.2 SSD企业级硬盘

**技术服务要求：**

1．设备安装调试: 在买方指定的地点完成安装调试，并配合买方进行测试验收

2．质保期自验收合格日起36个月

3．维修响应时间: 接到维修通知后，12小时内做出响应，24小时内到达现场排除故障

4．交货地点：用户指定地点