**病理大模型GPU算力服务器需求**

**主要配置参数：**

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 设备名称 | 主要配置参数 | 数量 |
| 1 | GPU算力服务器 | **CPU：**2颗INTEL XEON五代CPU  不低于Intel 8558 (2.1GHz/48核/260MB) 或者  2颗AMD EPYC四代CPU不低于AMD Genoa 9654( 2.4GHz/96核/384MB)；  **显卡：**8张NVIDIA H20 141GB GPU卡；  **内存：**24根64GB DDR5 5600内存；  **硬盘：**2块1.92TB或以上SSD + 不少于24块3.84TB SSD；  **阵列卡：**1块8GB缓存及以上，支持RAID0/1/5/6，带掉电保护模块；  **网卡：**1\*4端口千兆网卡，2\*2端口25Gb光接口网卡（含模块），8\*单端口400G NDR IB网卡（含模块）；  **服务：**6年原厂质保与7×24四小时现场支持 | 3台 |
| 2 | 算力管理与前置处理服务器 | **CPU：**2颗Intel Xeon 5520或以上处理器；  **内存：**16\*32G-DDR5 5600内存;  **硬盘：**2\*960G SSD+不少于8块3.84TB SSD;  **阵列卡：**1块4GB缓存及以上，支持RAID0/1/5/6，带掉电保护模块；  **网卡：**1\*4端口千兆网卡，2\*2端口10G网卡（含模块）；  **其他：**冗余电源及风扇，导轨；  **服务：**6年原厂质保与7×24四小时现场支持 | 4台 |
| 3 | 大容量数据存储 | **存储架构：**一套分布式存储集群，包括不限容量的分布式存储软件授权及不少于3个分布式存储节点;  **容量：**集群配置处理器总核数不少于72核，集群采用纠删码或三副本冗余方式，提供可用容量不少于1PB；  **单节点配置要求：**  集群每节点配置2颗CPU，单颗处理器主频不低于2.1GHz，内存不小于256GB；  配置不少于2块480G SSD的系统盘，支持通过RAID卡做RAID1；  为提高集群读写性能，每节点应配置不少于3块3TB 以上 NVME高速固态硬盘；  集群每节点配置不少于2张双端口25G光口网卡作为存储业务冗余网络，配置不少于1张4口1G电口网卡作为管理网络；  配置冗余电源及风扇。  **服务：**6年原厂质保与7×24四小时现场支持(含硬盘不返还) | 1套 |
| 4 | 高速网络交换机 | 配套高速IB网络交换机要求：  **网口：**不少于40个NDR 400Gb IB端口，含24个NDR 400Gb模块及线缆；  **其他：**冗余电源及风扇，导轨，配套线缆等；  **服务：**6年原厂质保与7×24四小时现场支持 | 1台 |
| 5 | 业务网络交换机 | **主要配置**：25G光端口数量≥40个，40G/100G光接口≥8个，交换容量≥7.8Tbps，包转发率≥2300Mpps，≥1根100G 高速堆叠线缆，不少于24个 25G 模块；配置冗余电源及风扇。  **服务：**6年原厂质保与7×24四小时现场支持 | 2台 |
| 6 | 算力调度平台 | **用户管理：**支持web界面管理算力资源，支持划分系统管理员、租户管理员、普通用户权限。  **文件管理：**支持web界面管理文件夹、上传文件，FTP上传大文件，支持NFS和GPFS文件。支持镜像管理（私有、公共、外部）。  **资源计量：**支持资源使用量计算。  **异构管理：**支持多品牌GPU卡统一管理，包括但不限于英伟达、沐曦、天数、昆仑芯、英特尔、华为、RTX 4090等。支持多架构CPU同时纳管（X86、ARM）。支持按照MIG物理切分和百分比虚拟切分GPU卡。  **集群资源监控：**支持集群资源监控；支持CPU、内存、GPU、网络带宽等资源的实时使用曲线图；支持单节点的资源实时监控；支持多维度资源统计。支持任务报表  **数据标注：**支持数据标注功能，支持个人标注和团队标注，支持标注任务审核  **学习框架支持：**平台内置TensorFlow、Pytorch等常见深度学习框架和scikit-learn机器学习框架，支持多框分布式架建模训练。建模支持编码式和可视化建模。  **模型管理：**内置多种常用大模型，支持模型导入导出、支持模型格式转换，支持模型评估工具，支持模型推理。  **模型训练：**内置微调脚本和参数调优工具，支持故障自动检测，支持自动checkpoint断点续训，支持checkpoint存储加速 | 1套 |