

序号	技术参数指标要求
1	深空目标识别多模态大模型并行计算模块
●1.1	外形：4U 机架式
★1.2	数据总线：采用 PCIe 5.0 x16 通道，提供 ≥ 8 个物理规格为 x16 的 PCIe 插槽，并支持同时扩展安装 ≥ 8 个计算单元。
★1.3	系统运行缓存：配置 12 条单条容量 $\geq 64\text{G}$ 高速随机存储条，运行参考频率 $\geq 4800\text{MHz}$ ，规格： $\geq \text{DDR5}$ ，扩展性：32 个插槽，最高支持 8T 总容量
★1.4	主控单元：数量 ≥ 2 颗。单枚通用逻辑处理单元 ≥ 32 个物理核心， ≥ 64 处理线程，基准工作频率 $\geq 2.1\text{GHz}$ ，系统级高速缓存 $\geq 160\text{MB}$
▲1.5	存储：配置 1 块 U.2 接口，NVME 协议，容量 $\geq 1.9\text{T}$ ；配置 6 块 Nvme4.0 接口，单块容量 $\geq 7.5\text{TB}$ ；均应为企业级
▲1.6	网卡：配置至少 2 个 10G 高速网口，1 个远程带外管理接口
●1.7	电源：数量 ≥ 3 块，单个电源模块 $\geq 2700\text{W}$ ，电源采用冗余设计
●1.8	IO 接口： ≥ 1 个视觉输出接口， ≥ 2 个 USB3.0 接口
▲1.9	框架支持：预装主流学习框架，例如：tensorflow、caffe、mxnet、pytorch、paddlepaddle 等主流的学习框架，提供软件功能截图证明
●1.10	资源虚拟化：采用轻量级容器虚拟化技术，实现对处理核心、运行存储、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，提供异构计算资源管理接口，实现对异构加速芯片等计算资源的虚拟化统一管理，支持为容器以直通方式挂载异构加速计算资源
●1.11	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
▲1.12	无卡模式：当加速计算资源不足时，允许用户可以使用纯逻辑运算模式启动任务，任务 ID 以及任务中的数据不丢失，提供软件功能截图证明
▲1.13	重置系统/更换镜像：对任务环境或者系统盘进行更换，更换完毕后，缓存盘与数据盘中的数据不丢失，提供软件功能截图证明
▲1.14	运算空间切分：支持对运算单元专用缓冲区进行切分，切分最小单位 1MB，提供软件功能截图证明
▲1.15	系统转换：支持虚拟化镜像与裸金属操作系统转换，提供软件功能截图证明

▲1.16	裸金属设备管理：在线添加裸金属服务器（支持本地系统盘与云系统盘两种方式），实现对裸金属服务器的开关机、重启、IP 配置、OS 更换、OS 重置、删除等操作，提供软件功能截图证明
▲1.17	裸金属操作系统下发：支持操作系统自动下发并自动部署，支持虚拟机镜像下发到裸金属服务器启动，提供软件功能截图证明
▲1.18	VPC 网络：支持创建多个 VPC 网络，每个网络可以创建多个子网能够提供安全、隔离的网络环境，为项目中的应用或虚拟机提供路由器 SNAT 能力，提供软件功能截图证明
●1.19	在线开发：支持在线模型开发功能，提供 JupyterLab、JupyterNoteBook、VS Code, Terminal 等在线编程环境
2	深空目标识别多模态大模型调度计算模块
★2.1	调度计算单元数：每套调度计算模块含 ≥ 3 个调度计算单元，该模块支持 ≥ 4 个计算单元槽位的升级
▲2.2	互联计算：支持 ≥ 2 个调度计算模块互联，异构通讯带宽 $\geq 600\text{GB/s}$
★2.3	单个调度计算单元能力：FP64 $\geq 25\text{TFLOPS}$; FP32 $\geq 50\text{TFLOPS}$
★2.4	专用存取空间 $\geq 80\text{GB}$ ，存取带宽 $\geq 2\text{TB/s}$
●2.5	I/O 带宽：PCIe 5.0 双向带宽 $\geq 128\text{GB/s}$
●2.6	系统接口：PCI Express 5.0 x16 (PCIe Gen5)