高性能 GPU 服务器购置

任务需求说明

一、任务目的

为持续推进上海科学数据管理与共享服务平台数字化、智能化水平，落实国家算力基础设施自主可控、国内自主研发、商用密码应用安全相关政策要求，进一步完善平台国产化算力基础设施体系，支撑科学数据智能检索、智能审核、质量检验、大规模并行科学计算、科研数据开放服务等核心业务场景高效运转，现需采购一台全栈国产化高性能 GPU 服务器，用以构建标准化、高可靠、高密度、自主可控的高端国产 GPU 算力平台。通过夯实平台国产化算力底座，提升海量科学数据智能化处理效率与大规模并行计算能力，全面提升上海科学数据管理与共享服务平台的智能化水平与公共科研服务能级。

二、需求描述

1、 GPU 服务器配置要求

序号	配件类别	配置详情	数量	备注
1	CPU	▲国产化 X86 处理器，单颗物理核心≥48 核，单颗逻辑线程≥96 线程，基础主频≥2.6GHz；单颗三级缓存≥256MB；支持硬件虚拟化、高级向量扩展指令集；整机配置 2 颗同规格处理器配套使用。	2	适配大规模 AI 训练、并行科学计算场景，禁止采用同核心低线程、低缓存阉割款处理器
2	平台	高性能算力服务器平台 1．搭载国产化可扩展处理器，面向大规模人工智能与高性能计算场景设计； 2．PCIE 双 10Gb 以太网网卡； 3．▲存储扩展位≥12 个 NVMe U.2 高速盘位； 4．▲PCIe5.0 标准全速扩展插槽≥14 个，多设备同时工作不拆分带宽降速； 5．电源符合钛金能效标准，采用 N+M 冗余热插拔架构，最低配置 N≥4、M≥2，整机供电冗余预留量≥20%，主机供电与 GPU 供电物理隔离独立设计，保障 8 卡 GPU 满负载稳定供电； 6．▲整机支持 SM2、SM3、SM4 国密算法硬件加速，实现方式三选一：CPU 内置密码模块/板载国密芯片/PCIe 国密密码卡；国密硬件须具备国家商用密码产品认证证书； 7．整机满配并发压力下硬件 SM4 对称加密吞吐性能≥8GB/s； 8．25℃常温机房环境下，8 卡 GPU100%满载连续运行，GPU 核心稳定温度≤95℃，无过热降频、硬件告警。	1	专为大规模 AI 与高性能计算场景设计，支持多算力单元高带宽互联，整机交付前需完成 72 小时全负载烤机测试无故障

3	内存	▲整机配置国产 DDR5 ECC RDIMM 高速内存，单条内存基础频率≥4800MHz，整机总内存容量≥1536GB；整机内存物理插槽≥24 个，支持内存热插拔、单条内存硬件故障隔离；整机内存总带宽≥920GB/s，支持内存镜像数据保护功能。	1	允许采 64GB/128GB 等不同容量内存组合
4	系统盘	▲960GB NVME U.2 企业级固态盘，支持热插拔、断电保护，单盘持续读写性能不低于行业企业级标准。	2	用于操作系统、驱动、算力环境部署
5	存储盘	▲7.68TB NVME U.2 企业级固态盘，支持热插拔，单盘具备稳定高 IOPS、低延迟特性，适配海量科学数据高速读写。	8	承载科研数据集、大模型权重文件存储
6	网卡	▲2 张双口万兆光纤以太网网卡，配套原厂合规万兆光模块；支持标准数据中心网络协议，无丢包、低延迟数据传输。	1	适配数据中心内网高速数据交互
7	阵列卡	NVMe 兼容硬件阵列控制器，单卡缓存≥4GB，原生支持 RAID0/1/10/5 阵列模式，支持硬盘故障自动重构	2	保障高速 NVMe 存储数据冗余与故障恢复
8	IB 卡	▲单端口标称带宽≥200Gbps，同时支持 RoCE v2 协议、InfiniBand 协议；空载端到端传输延迟≤1.2μs；支持≥1000 路 RDMA 并发数据流；配套原厂高速线缆、合规光模块，多节点集群并行传输无丢包。	1	低延迟、高带宽节点互联，适配高性能计算集群协同运算场景
9	显卡	8 卡国产 GPU 算力模组，支持 OAM/PCIe5.0/高速卡间直连多类互联标准， ▲8 卡整体互联总带宽≥400GB/s； ▲单卡物理显存≥144GB， ▲显存带宽≥2.4TB/s； 单卡 72 小时满载持续稳定算力指标： ▲FP16/BF16≥385 TFLOPS ▲TF32≥192 TFLOPS ▲INT8≥771 TOPS， 禁止仅提供实验室短时峰值算力；单卡稳定功耗承载≥400W，配套独立散热风道，8 卡同时满载不降频；完整兼容国产深度学习框架、主流 AI4S 科学计算软件栈；（提供产品规格书、厂商技术白皮书或第三方检测报告）。	1	八路高协同算力模组，支撑大模型训练、大规模并行科学计算等高负载业务场景
10	电源	整机配套钛金能效电源，采用 N+M 热插拔冗余架构（最低 4+2 配置），整机供电冗余余量不低于 20%；单电源满载转换效率≥96%；GPU 供电回路与主板供电回路物理隔离，单电源故障可在线热插拔更换，整机业务不中断；满足 8 卡 GPU 长期满负载稳定供电需求。	1	高电能转换效率，多重供电冗余保障 7×24 小时不间断运行
11	机柜	配套标准 IDC 机房机柜，满足 8 卡 GPU 服务器上架要求：机柜整机承重≥1500kg，具备三级抗震结构；配套大功率 PDU 电源分配单元、独立接地防雷、防震固定支架、高速理线、整机通风散热全套配件；投标人若提供通用标准化机柜，须出具加盖投标人公章的适配改造方案，改造后整机散热、功耗、承重、供电均满足机房长期稳定运行标准，机柜与服务器一体化交付，同步完成布	1	配套机柜随服务器一体化交付，保障设备规范上架运维

		线接地调试。		

2、整机交付与测试要求

操作系统预装：中标方按需预装国产服务器操作系统或 Ubuntu 系统，完整适配全套硬件驱动、算力环境、科学计算依赖库，交付时可直接上线使用；

整机压力测试：设备出厂、现场交付两个阶段均完成 72 小时 8 卡 GPU 全负载连续烤机测试，全程 CPU、内存、NVMe 硬盘、GPU 无降频、无硬件报错、无重启、无断流，保存完整测试日志作为验收材料；

散热验收标准：机房常温 25℃环境下，8 卡 100%满载运行 48 小时，GPU 核心温度持续≤95℃，无高温告警、自动降频现象；

PCIe 通道验收：所有 PCIe5.0 扩展插槽、GPU、IB 网卡、万兆光纤网卡同时满负载工作，各设备均跑满 PCIe5.0 全速带宽，无通道拆分限速；

性能核验规则：所有算力、带宽、网络延迟等核心性能指标，以第三方检测机构满载实测数据为最终验收依据（检测费包含于投标报价中），厂商白皮书、宣传参数仅作辅助参考。

3、证明材料响应要求

性能佐证材料：国产 GPU 需提供近 12 个月第三方检测机构测试报告；

架构合规材料：多卡互联、服务器平台冗余供电相关技术说明文件，佐证满足本需求最低性能门槛。

4、GPU 卡的先进性要求

要求 GPU 卡具备超大显存与先进多卡互联能力，兼容主流生态的 MXMACA 软件栈，迁移便捷。配合科学计算平台，可高效支撑材料模拟、生命科学、流体力学、分子模拟等 AI4S 前沿科研场景。

5、实施服务内容

乙方负责全程全包服务，包含原厂包装运输、现场卸货、机柜就位、服务器上架固定、强弱电布线、智能温控风机通风、系统安装、环境调试、整机联调、现场交付培训、设备清洁验收等全部人工与实施服务，实现设备到场即可上线投用。

6、人员要求

需提供一名项目经理负责本次 GPU 服务器的统筹工作，包含设备上架、系

统部署、设备布线等要求，要求有大型项目管理经验，项目经理持有信息系统项目管理师证书，或计算机专业高级工程师证书。

需提供至少一名项目实施工程师，负责本次项目 GPU 服务器的实施、交付工作，要求具备类似项目经验。

7、交付地点

上海市徐汇区钦州路 100 号 2 号楼 4 楼机房。

三、服务时间

自合同签订生效之日起一个月内完成实施上架并通过交付验收，服务器验收合格后免费质保期为五年。

四、质保与售后服务要求

整机及全部配件（CPU、内存、NVMe 固态、国产 GPU 模组、电源、机柜、网卡、阵列卡、配套机柜及所有辅材等）统一提供 5 年原厂整机质保，质保期自设备验收合格当日起计算；

质保期内提供 7×24 小时全天候技术电话支持；用户报修 1 小时内技术人员响应，2 小时抵达现场，2 个工作日内完成故障配件更换；

投标人提供加盖公章的书面质保服务承诺，若未按约定时限完成售后处置，按合同约定承担相应违约责任。

五、需求结果

1、通过 GPU 服务器标准化上架、安全散热、规范布线、稳定供电，满足运行规范与安全运维要求，设备整体安全性、稳定性、散热性、承重性达到专业算力机房标准。

2、高性能 GPU 服务器通过配置大内存、高速 NVMe 存储、Gen5 高速扩展架构、冗余电源架构，可稳定承载科学数据智能审核、智能检索、质量检验、大规模科学计算等多任务并发算力业务，满足 7×24 小时不间断运行。