

采购需求

一、项目说明

采购标的需实现的目标：因项目需要拟采购一批符合技术指标要求的算力资源及配套存储和技术支持服务，针对上述所提供的计算资源，投标人应提供稳定的算力资源支持，并提供每周 7×24 小时，5 分钟内响应的技术服务支持保障，随时解答计算相关各类问题，确保项目顺利完成。

二、商务要求

1. 交付的时间和地点

★（1）交付时间：合同签订后 10 个工作日内交付符合技术指标要求的算力资源。

★（2）交付地点：线上交付。

★2. 付款条件（进度和方式）

合同签订后 10 个工作日内，采购人对中标人支付本包总预算的 50%，投标人应交付符合采购技术指标要求的算力资源，采购人应于标的物完成交付后对中标人所交付的算力配置及数量进行核验和验收，验收通过后支付剩余尾款。

★3. 服务期限

租赁服务期以合同签订日期为准，服务期内提供技术支持等服务。01 包合同服务期限至 2026 年 5 月 31 日届满；02 包合同服务期限至 2026 年 4 月 30 日届满。

4. 履约验收要求

组织专家或采购人自行验收。详见合同文本相关内容。

5. 保密/知识产权要求

详见合同文本相关内容。

6. 报价要求

6.1 投标人应在开标一览表中分别报出所投包的资源一、资源二、共享存储及互联网带宽的计费单价，分项报价超过单价最高限价的，**投标无效**。

6.2 本项目合同为固定单价合同，投标人须按照招标文件要求报投标单价，投标单价为包含完成本项目所有工作的全部费用的综合单价。合同价款以采购人实际

计算资源用量据实结算，每包合同价款的结算总金额最高不超过该包预算金额。

三、技术要求

1. 各包计算项目需求及包预算/最高限价

包号	项目名称	主要规格要求	预估数量	包预算/最高限价（万元）
1	GPU算力服务采购01	提供两种基于物理裸机的 GPU 算力资源服务，并在投标文件中提供资源按需使用单价（单位：元/卡时），采购人可根据项目需要动态调拨两种资源卡时可用额度，但合同结算总金额不超过本包最高限价。	资源一： 75,231卡时 资源二： 93,304卡时	163
2	GPU算力服务采购02	提供两种基于物理裸机的 GPU 算力资源服务，并在投标文件中提供资源按需使用单价（单位：元/卡时），采购人可根据项目需要动态调拨两种资源卡时可用额度，但合同结算总金额不超过本包最高限价。	资源一： 24,369卡时 资源二： 29,896卡时	52.8

01、02 包单价的最高限价如下：

报价项	单价最高限价
资源一（单价：元/卡时）	6.5
资源二（单价：元/卡时）	12
共享存储单价（单价：元/（1TB·天））	12
互联网带宽单价（单价：元/（100Mbps·天））	110

备注：投标报价（单价）超出以上单价最高限价，则投标无效。

01 包计算项目服务要求

1. 技术内容、形式和要求

提供基于物理裸机的 GPU 算力资源服务，投标人应同时提供两种算力资源，且资源应支持按需或整租模式使用，不得采用 vCPU、vGPU、云主机等虚拟化资源。同时提供其他配套服务：包括但不限于足够的数据存储空间、算力管理及访

问软件、资源监控报警、配备专业技术团队对所需配置环境、软件等提供每周 7×24 小时技术服务支撑等。

2. 技术要求

2.1 算力资源配置要求

备注：

（1）除非下表“配置要求”中明确说明投标人在投标文件中应提供的证明材料，否则均按下表“佐证材料类型”要求提供证明材料。

（2）因项目特殊性，要求投标人应同时提供以下两种算力资源配置。

资源一：

序号	类别	配置要求	重要性	佐证材料类型
1	处理器	配置≥2 颗通用 x86 CPU，支持全宽 AVX512 指令集（不低于），发布时间不得早于 2021 年 1 月 1 日（以官网规格说明为准），主频≥2.6GHz，单颗处理器物理核心数≥32 核。	★	提供符合配置要求的资源集群系统配置截图，及 CPU、GPU 的官网规格说明截图或技术彩页
2	内存	内存总量≥1024GB。	★	
3	GPU 卡	单卡显存≥80GB，单卡 FP32 性能≥19.5TFLOPS，单卡 TF32 性能（稀疏算法）≥300TFLOPS，FP16 性能（稀疏算法）≥600TFLOPS，单卡显存带宽≥2000 GB/s。	★	
4	卡间通信	节点内任意两卡间通信带宽≥400GB/s。	★	
5	节点互联	节点间应采用高速物理链路互联，聚合带宽≥800Gbps。	★	
6	共享存储	具有可提供不低于 20TB 共享存储，可用期限不低于 60 天的能力（按照采购人实际使用计费）。	★	提供资源配置承诺函
7	互联网带宽	具有可提供不低于 100Mbps 互联网带宽，可用期限不低于 60 天的	★	

		能力（按照采购人实际使用计费）。		
--	--	------------------	--	--

资源二：

序号	类别	配置要求	重要性	佐证材料类型
1	处理器	配置 ≥ 2 颗通用 x86 CPU，支持全宽 AVX512 指令集（不低于），发布时间不得早于 2023 年 1 月 1 日（以官网规格说明为准），主频 $\geq 2.1\text{GHz}$ ，单颗处理器物理核心数 ≥ 32 核。	★	提供符合指标描述的资源集群系统配置截图及 CPU、GPU 的官网规格说明截图或技术彩页
2	内存	内存总量 $\geq 2048\text{GB}$ 。	★	
3	GPU 卡	单卡 GPU 显存 $\geq 80\text{GB}$ ，单卡 FP32 性能 $\geq 60\text{TFLOPS}$ ，单卡 TF32 性能（稀疏算法） $\geq 980\text{TFLOPS}$ ，FP16 性能（稀疏算法） $\geq 1900\text{TFLOPS}$ ，单卡显存带宽 $\geq 3.3\text{TB/s}$ 。	★	
4	卡间通信	节点内任意两卡间通信带宽 $\geq 900\text{GB/s}$ 。	★	
5	节点互联	节点间应采用高速物理链路互联，聚合带宽 $\geq 3200\text{Gbps}$ 。	★	
6	共享存储	具有可提供不低于 20TB 共享存储，可用期限不低于 60 天的能力（按照采购人实际使用计费）。	★	提供资源配置承诺函
7	互联网带宽	具有可提供不低于 100Mbps 互联网带宽，可用期限不低于 60 天的能力（按照采购人实际使用计费）。	★	

2.2 性能测试要求

投标人应在本次所投资源平台上单机 8 卡配置环境下参照要求运行以下测试样例，并在投标文件中提供运行日志及符合要求的结果证明材料，否则视为未实际响应不予计分。

▲（1）资源一性能测试要求：

数据集	数据集：Pile 下载地址： https://www.kaggle.com/datasets/dschettler8845/the-pile-dataset-part-00-of-29 。投标人需自行从相应网站下载。
模型	模型：GLM-10B；模型代码： https://github.com/THUDM/GLM
算力	本次所投资源平台，单机 8 卡。
任务参数设置	模型参数与训练参数配置： <code>--hidden-size 4096 --num-attention-heads 64 --seq-length 1024</code> <code>--max-position-embeddings 1024</code>
结果要求	吞吐量 ≥ 10 samples/s。

（2）资源二性能测试要求：

▲（a）AIPerf-LLM 大模型训练基准性能测试

测试算例	投标人自行从官方下载： https://www.aiperf.org.cn/AIPerf-LLM-FastMoE.tar.gz
测试要求	本次所投资源平台，单机 8 卡，参数不限。
结果要求	输出性能 $\geq 3.5 \times 10^{15}$ FLOPS（请注意单位，以测试日志输出结果为准）。

▲（b）训练能力测试

测试算例	使用 GLM-10B 模型（ https://github.com/THUDM/GLM ）和 Pile 数据集（ https://www.kaggle.com/datasets/dschettler8845/the-pile-dataset-part-00-of-29 ），投标人需自行从相应网
------	--

	站上下载。
测试要求	本次所投资源平台，单机 8 卡，关键参数设置：“--hidden-size 4096 --num-attention-heads 64 --seq-length 1024 --max-position-embeddings 1024”。
结果要求	RunningAvgSamplesPerSec $\geq$ 20（以测试日志输出结果为准）。

▲（c）推理能力测试

测试算例	DeepSeek-V3.1-Terminus，下载地址： https://modelscope.cn/models/deepseek-ai/DeepSeek-V3.1-Terminus ）。 投标人需自行从相应网站上下载。
测试要求	所投资源平台，单机 8 卡，参数设置：FP8 精度，输入长度设置为 10 tokens，输出长度设置为 1000 tokens。
性能要求	单并发测试 TTFT $\leq$ 150ms，TPOT $\leq$ 7.5ms（以测试日志输出结果为准）。

2.3 服务能力要求

（1）算力平台要求

投标人可提供支持以下功能的资源访问平台，按下表“佐证材料类型”的要求提供佐证材料；否则视为不满足。

序号	类别	技术指标	重要性	佐证材料类型
1	计算环境	应采用主流 64 位 Linux 操作系统；预装 GPU 驱动、编译器、数学库、科学计算工具包等并行计算环境；允许用户在线安装计算所需软件并提供安装协助。		提供集群系统计算环境配置截图
2	算力访问	针对按需资源，投标人应提供基于互联网的算力访问平台，支持文件传输、WEB 命令行、集成第三方 SSH 工具、作业管理、计费统计、余额查询等功能。		提供覆盖所有技术指标的软件系统功能截图
3		面向 Windows/MacOS 用户可提供具备上述功能的平台桌面客户端。面向 Linux/MacOS 可提供免密 SSH 访问算力和文件传输的命		

		令行客户端。		
4	消息通知	平台支持用户通过 WEB 自助设置预警通知类型及报警阈值，可将算力资源动态、用户作业异常、费用不足等报警信息及时推送到用户微信或邮箱。		
5	账号管理	平台支持主账号自助管理组织成员及超算账号的绑定关系、查询所有组织成员的超算作业状态和存储使用情况。平台支持用户自助共享目录给指定用户并设置只读或读写权限。	▲	
6	配额管理	投标人可为组织内的子账户设置机时定额（消耗限额），以限制子账户的计算消耗。	▲	
7	资源监控	投标人可提供实时监控算力节点系统指标（如处理器利用率、GPU 利用率、内存利用率、内存带宽读写、高速计算网络收发、文件系统读写等）的表现情况，并能将上述指标的平均值、瞬时值、峰值、总量等数值量化记录生成统计图及雷达图。	▲	
8	便捷查询	投标人可提供可实时查询作业状态和费用信息的微信小程序。	▲	提供符合要求的微信小程序名称和截图

(2) 技术服务要求

投标人在投标文件中应按下表要求提供佐证材料，否则视为不满足。

序号	类别	要求	重要性	佐证材料类型
1	服务标准	投标人应为本项目组建专项技术团队，建立专属微信服务群等即时沟通渠道并可承诺在机时未使用完期间均提供每周 7×24 小时技术支持服务，可以在 5 分钟内响应用户的使用诉求，随时解答用户提出的计算问题。	★	提供服务承诺函
2	服务内容	投标人可提供的技术服务应包括但不限于：资源及平台软件说明、用户培训、环境调试、异常排除、协助用户安装部		

		署 AI 计算相关软件等服务，并确保夜间与节假日均可以处理各类异常问题。		
3	数据转移	为保证采购人短时大量数据传输场景下网络传输效率受限的问题，投标人快递邮寄硬盘，全程协助采购人完成硬盘挂载、数据传输等转移操作。		
4	Slurm 集群支持能力	投标人应具备大规模 Slurm 集群的部署配置与故障排查能力，支持万核以上多机多核调度，确保 Slurm 调度系统与算力集群操作系统和其他软件环境的适配性。		大规模 Slurm 集群应用案例

3. 采购标需要满足的国家标准、行业标准及规范：无。

4. 其他要求

（1）投标人在投标文件中提供系统资源监控软件的界面截图，应能够实时采集并可视化展示核心系统资源性能指标。

（2）投标人在投标文件中提供 GPU 性能采集系统的界面截图，应能够实时采集并可视化展示 GPU 核心性能指标。

（3）投标人可提供 Deepseek、Llama、GLM、Bloom 等大模型应用在模型训练时各时刻 GPU 卡核心性能指标（如 Tensor Core 利用率、FP16 利用率、显存读写、卡间通信等）的表现情况图表。提供大模型训练性能表现图表并附分析解读。

02 包计算项目服务要求

1. 技术内容、形式和要求

提供基于物理裸机的 GPU 算力资源服务，投标人应同时提供两种算力资源，且资源应支持按需或整租模式使用，不得采用 vCPU、vGPU、云主机等虚拟化资源。同时提供其他配套服务：包括但不限于足够的数据存储空间、算力管理及访问软件、资源监控报警、配备专业技术团队对所需配置环境、软件等提供每周 7×24 小时技术服务支撑等。

2. 技术要求

2.1 算力资源配置要求

注：

(1) 除非下表“配置要求”中明确说明投标人在投标文件中应提供的证明材料，否则均按下表“佐证材料类型”要求提供证明材料。

(2) 因项目特殊性，要求投标人应同时提供以下两种算力资源配置。

资源一：

序号	类别	配置要求	重要性	佐证材料类型
1	处理器	配置 ≥ 2 颗通用 x86 CPU，支持全宽 AVX512 指令集（不低于），发布时间不得早于 2021 年 1 月 1 日（以官网规格说明为准），主频 $\geq 2.6\text{GHz}$ ，单颗处理器物理核心数 ≥ 32 核。	★	提供符合配置要求的资源集群系统配置截图，及 CPU、GPU 的官网规格说明截图或技术彩页
2	内存	内存总量 $\geq 1024\text{GB}$ 。	★	
3	GPU 卡	单卡显存 $\geq 80\text{GB}$ ，单卡 FP32 性能 $\geq 19.5\text{TFLOPS}$ ，单卡 TF32 性能（稀疏算法） $\geq 300\text{TFLOPS}$ ，FP16 性能（稀疏算法） $\geq 600\text{TFLOPS}$ ，单卡显存带宽 $\geq 2000\text{ GB/s}$ 。	★	
4	卡间通信	节点内任意两卡间通信带宽 $\geq$	★	

		400GB/s。		提供资源配置承诺函
5	节点互联	节点间应采用高速物理链路互联，聚合带宽 $\geq 800\text{Gbps}$ 。	★	
6	共享存储	具有可提供不低于 20TB 共享存储，可用期限不低于 30 天的能力（按照采购人实际使用计费）。	★	
7	互联网带宽	具有可提供不低于 100Mbps 互联网带宽，可用期限不低于 30 天的能力（按照采购人实际使用计费）。	★	

资源二：

序号	类别	配置要求	重要性	佐证材料类型
1	处理器	配置 ≥ 2 颗通用 x86 CPU，支持全宽 AVX512 指令集（不低于），发布时间不得早于 2023 年 1 月 1 日（以官网规格说明为准），主频 $\geq 2.1\text{GHz}$ ，单颗处理器物理核心数 ≥ 32 核。	★	提供符合指标描述的资源集群系统配置截图，及 CPU、GPU 的官网规格说明截图或技术彩页
2	内存	内存总量 $\geq 2048\text{GB}$ 。	★	
3	GPU 卡	单卡 GPU 显存 $\geq 80\text{GB}$ ，单卡 FP32 性能 $\geq 60\text{TFLOPS}$ ，单卡 TF32 性能（稀疏算法） $\geq 980\text{TFLOPS}$ ，FP16 性能（稀疏算法） $\geq 1900\text{TFLOPS}$ ，单卡显存带宽 $\geq 3.3\text{TB/s}$ 。	★	
4	卡间通信	节点内任意两卡间通信带宽 $\geq 900\text{GB/s}$ 。	★	
5	节点互联	节点间应采用高速物理链路互联，聚合带宽 $\geq 3200\text{Gbps}$ 。	★	提供资源配置承诺函
6	共享存储	具有可提供不低于 20TB 共享存储，可用期限不低于 30 天的能	★	

		力（按照采购人实际使用计费）。		
7	互联网带宽	具有可提供不低于 100Mbps 互联网带宽，可用期限不低于 30 天的能力（按照采购人实际使用计费）。	★	

2.2 性能测试要求

投标人应在本次所投资源平台上单机 8 卡配置环境下参照要求运行以下测试样例，并在投标文件中提供运行日志及符合要求的结果证明材料，否则视为不满足。

▲（1）资源一性能测试要求：

数据集	数据集：Pile 下载地址： https://www.kaggle.com/datasets/dschettler8845/the-pile-dataset-part-00-of-29 。投标人需自行从相应网站上下载。
模型	模型：GLM-10B；模型代码： https://github.com/THUDM/GLM
算力	本次所投资源平台，单机 8 卡。
任务参数设置	模型参数与训练参数配置： <code>--hidden-size 4096 --num-attention-heads 64 --seq-length 1024 --max-position-embeddings 1024</code>
结果要求	吞吐量 ≥ 10 samples/s。

（2）资源二性能测试要求：

▲（a）AIPerf-LLM 大模型训练基准性能测试

测试算例	投标人自行从官方下载： https://www.aiperf.org.cn/AIPerf-LLM-FastMoE.tar.gz
测试要求	本次所投资源平台，单机 8 卡，参数不限。
结果要求	输出性能 $\geq 3.5 \times 10^{15}$ FLOPS（请注意单位，以测试日志输出结果为准）。

▲（b）训练能力测试

测试算例	使用 GLM-10B 模型 (https://github.com/THUDM/GLM) 和 Pile 数据集 (https://www.kaggle.com/datasets/dschettler8845/the-pile-dataset-part-00-of-29), 投标人需自行从相应网站上下载。
测试要求	本次所投资源平台, 单机 8 卡, 关键参数设置: “--hidden-size 4096 --num-attention-heads 64 --seq-length 1024 --max-position-embeddings 1024”。
结果要求	RunningAvgSamplesPerSec $\geq$ 20 (以测试日志输出结果为准)。

▲（c）推理能力测试

测试算例	DeepSeek-V3.1-Terminus, 下载地址: (https://modelscope.cn/models/deepseek-ai/DeepSeek-V3.1-Terminus)。投标人需自行从相应网站上下载。
测试要求	所投资源平台, 单机 8 卡, 参数设置: FP8 精度, 输入长度设置为 10 tokens, 输出长度设置为 1000 tokens。
性能要求	单并发测试 TTFT $\leq$ 150ms, TPOT $\leq$ 7.5ms (以测试日志输出结果为准)。

2.3 服务能力要求

（1）算力平台要求

投标人可提供支持以下功能的资源访问平台, 要求按下表“佐证材料类型”要求提供佐证材料; 否则视为不满足。

序号	类别	技术指标	重要性	佐证材料类型
1	计算环境	应采用主流 64 位 Linux 操作系统; 预装 GPU 驱动、编译器、数学库、科学计算工具包等并行计算环境; 允许用户在线安装计算所需软件并提供安装协助。		提供集群系统计算环境配置截图
2	算力访问	针对按需资源, 投标人应提供基于互联网的算力访问平台, 支持文件传输、WEB		提供覆盖所有技术指标

		命令行、集成第三方 SSH 工具、作业管理、计费统计、余额查询等功能。		的软件系统功能截图
3		面向 Windows/MacOS 用户可提供具备上述功能的平台桌面客户端。面向 Linux/MacOS 可提供免密 SSH 访问算力和文件传输的命令行客户端。		
4	消息通知	平台支持用户通过 WEB 自助设置预警通知类型及报警阈值,可将算力资源动态、用户作业异常、费用不足等报警信息及时推送到用户微信或邮箱。		
5	账号管理	平台支持主账号自助管理组织成员及超算账号的绑定关系、查询所有组织成员的超算作业状态和存储使用情况。平台支持用户自助共享目录给指定用户并设置只读或读写权限。	▲	
6	配额管理	投标人可为组织内的子账户设置机时定额(消耗限额),以限制子账户的计算消耗。	▲	
7	资源监控	投标人可提供实时监控算力节点系统指标(如处理器利用率、GPU 利用率、内存利用率、内存带宽读写、高速计算网络收发、文件系统读写等)的表现情况,并能将上述指标的平均值、瞬时值、峰值、总量等数值量化记录生成统计图及雷达图。	▲	
8	便捷查询	投标人可提供可实时查询作业状态和费用信息的微信小程序。	▲	提供符合要求的微信小程序名称和截图

(2) 技术服务要求

投标人在投标文件中应按下表要求提供佐证材料,否则视为不满足。

序号	类别	要求	重要性	佐证材料类型
1	服务	投标人应为本项目组建专项技术团队,	★	提供服务

	标准	建立专属微信服务群等即时沟通渠道并可承诺在机时未使用完期间均提供每周7×24小时技术支持服务，可以在5分钟内响应用户的使用诉求，随时解答用户提出的计算问题。		承诺函
2	服务内容	投标人可提供的技术服务应包括但不限于：资源及平台软件说明、用户培训、环境调试、异常排除、协助用户安装部署AI计算相关软件等服务，并确保夜间与节假日均可以处理各类异常问题。		
3	数据转移	为保证采购人短时大量数据传输场景下网络传输效率受限的问题，投标人快递邮寄硬盘，全程协助采购人完成硬盘挂载、数据传输等转移操作。		
4	Slurm集群支持能力	投标人应具备大规模Slurm集群的部署配置与故障排查能力，支持万核以上多机多核调度，确保Slurm调度系统与算力集群操作系统和其他软件环境的适配性。		大规模Slurm集群应用案例

3. 采购标需要满足的国家标准、行业标准及规范：无。

4. 其他要求

（1）投标人在投标文件中提供系统资源监控软件的界面截图，应能够实时采集并可视化展示核心系统资源性能指标。

（2）投标人在投标文件中提供GPU性能采集系统的界面截图，应能够实时采集并可视化展示GPU核心性能指标。

（3）投标人可提供Deepseek、Llama、GLM、Bloom等大模型应用在模型训练时各时刻GPU卡核心性能指标（如Tensor Core利用率、FP16利用率、显存读写、卡间通信等）的表现情况图表。提供大模型训练性能表现图表并附分析解读。