合同文本

以下为中标后签订本项目合同的通用条款，中标人不得提出实质性的修改。

江苏省政府采购合同（服务）（合同编号）

项目名称：新吴区智能算力服务项目　项目编号：JSZC-320214-JZCG-G2025-0140

甲方：（买方、需方、采购人）无锡市新吴区区域社会治理现代化指挥中心

乙方：（卖方、供方、供应商）无锡城市云计算中心有限公司

甲、乙双方根据无锡市公共资源交易中心新吴分中心新吴区智能算力服务项目公开招标的结果，签署本合同。

一、合同内容

1.1 标的名称：新吴区智能算力服务项目

1.2 标的质量：/

1.3 标的数量（规模）：具备提供总资源规模不低于 9Pflops（FB16/BF16）的智能算力服务能力，据实结算。

1.4 履行时间（期限）：自合同签订生效之日起 1 年

1.5 履行地点：无锡市新吴区

1.6 履行方式：根据招投标文件的相关要求，按需提供智能算力服务，据实结算。具体服务内容及要求见附件一

二、合同金额

2.1 本次招投标为优惠折扣报价，中标（成交）优惠折扣：百分之零点零壹（大写）；0.01%（小写），中标单价为 17.778222 万元/P/年，总金额不超过 160 万元，根据实际使用算力服务情况据实结算。

三、技术资料

3.1 乙方应按招标文件规定的时间向甲方提供服务（包含与服务相关的货物）的有关技术资料。

3.2 没有甲方事先书面同意，乙方不得将由甲方提供的有关合同或任何合同条文、规格、计划、图纸、样品或资料提供给与履行本合同无关的任何其他人。即使向履行本合同有关的人员提供，也应注意保密并限于履行合同的必需范围。

四、知识产权

4.1 乙方应保证甲方在使用、接受本合同服务（包含与服务相关的货物）或其任何一部分时不受第三方提出侵犯其专利权、版权、商标权和工业设计权等知识产权的起诉。一旦出现侵权，由乙方负全部责任。

五、产权担保

5.1 乙方保证所交付的服务（包含与服务相关的货物）的所有权完全属于乙方且

无任何抵押、查封等产权瑕疵。

六、履约保证金

6.1 履约保证金的缴纳和退付：本项目通过保函方式向甲方缴纳履约保证金。

6.2 履约保证金缴纳金额：壹佰陆拾万圆整（160万元）。

6.3 履约保证金缴纳形式：履约保函（保险）。

履约保证金的退付时间：/，退付办法：/。

履约保证金不予退还的情形：乙方未能履行合同规定的任何义务，甲方有权从履约保证金中取得补偿。

逾期退还履约保证金的违约责任：如甲方未按规定及时退还履约保证金，应当向乙方支付同期银行存款利息。

七、合同转包或分包

7.1 乙方不得将合同标的转包给他人履行。

7.2 乙方应按招标文件要求，如项目不允许分包的，乙方不得将合同标的分包给他人履行；如项目允许分包的，乙方应当按照投标文件中提供的分包意向协议履行合同。

7.3 乙方如有转包或未经甲方同意的分包行为，甲方有权解除合同。

八、合同款项支付

8.1 合同款项的支付方式及进度安排：每满半年度支付一次，根据考核结果和实际使用情况支付费用，具体考核要求见附件二。

8.2 结算与支付细则

8.2.1 甲乙双方同意，每半年度为一个结算周期。

8.2.2 每个结算周期届满后 10 个工作日内，乙方应向甲方提交该周期内经双方书面确认的算力服务使用量清单、费用计算明细及附件二规定的相关运维报告。

8.2.3 甲方应在收到乙方完整结算资料后 20 个工作日内，依据附件二《服务考核要求》对该周期服务进行考核评分，并书面通知乙方考核结果。

8.2.4 甲方支付金额计算公式为：当期应付费用=当期经确认的算力使用费×99.99%-依据附件二考核结果应扣除的款项。其中，依据附件二第 14 条产生的扣款，应在当期费用中直接扣除。

8.2.5 甲方在考核结果通知发出且收到乙方开具的等额合规发票后 15 个工作日内，支付经核算后的款项。

九、税费

9.1 本合同执行中相关的一切税费均由乙方负担。

十、项目验收

10.1 本项目无需验收。

十一、违约责任

11.1 甲方无正当理由拒绝接受乙方提供服务的，甲方向乙方偿付拒绝接受服务合同价款总值 10% 的违约金。

11.2 甲方无故逾期验收和办理合同款项支付手续的，甲方应按逾期付款总额每日万分之五向乙方支付违约金。但违约金总额不超过合同预估总价(160 万元)的 10%。

11.3 乙方逾期提供服务的，乙方应按逾期提供服务合同总额每日千分之六向甲方支付违约金，由甲方从待付合同款项中扣除。逾期超过约定日期 10 个工作日不能提供服务的，甲方可解除本合同。乙方因逾期提供服务或因其他违约行为导致甲方解除合同的，乙方应向甲方支付合同价款总额 10% 的违约金。如造成甲方损失超过违约金的，超出部分由乙方继续承担赔偿责任。

11.4 乙方所提供服务的标准不符合合同规定及招标文件规定标准的，甲方有权拒绝接受服务，并可单方面解除合同。

十二、不可抗力事件处理

12.1 在合同有效期内，任何一方因不可抗力事件导致不能履行合同，则合同履行期可延长，其延长期与不可抗力影响 期相同。

12.2 不可抗力事件发生后，应立即通知对方，并寄送有关权威机构出具的证明。

12.3 不可抗力事件延续 120 天以上，双方应通过友好协商，确定是否继续履行合同。

十三、解决争议的方法

13.1 双方在签订、履行合同中所发生的一切争议，应通过友好协商解决。如协商不成，由甲方住所地人民法院管辖。

十四、合同生效及其他

14.1 合同经双方法定代表人或授权委托代表人签字并加盖单位公章后生效。

14.2 本合同未尽事宜，遵照《中华人民共和国民法典》、《中华人民共和国政府采购法》有关条文执行。

14.3 本合同正本一式两份，具有同等法律效力，甲方、乙方各执一份。

甲方：

地址：

法定代表人或授权代表：

联系电话：

乙方：

地址：

法定代表人或授权代表：王其浩

联系电话：

签订日期：2026年 1月 7日

附件一、主要服务内容及要求

本项目涉及的智能计算资源池、智算资源管理云平台、人工智能大模型管理平台等服务的主要要求如下：

序号	类别	功能名称	功能参数要求
（一）	智能计算资源池服务	国产异构 GPU 物理服务器（国产异构 GPU 物理服务器，使用国产处理器，4U 机架式，国产品牌），需集群部署，支持弹性扩容。	
1		处理器	双路国产处理器；单颗 CPU 核心≥48 物理核心，主频≥2.6GHz
2		内存	内存≥1024GB
3		存储	1、标配 2 块 480GB 企业级 SSD 硬盘，2 块 3.84TB 企业级 SSD 硬盘； 2、标配独立 RAID 卡，2G 缓存，支持 Raid 0/1/5/10/50/60
4		GPU 卡	单机配置≥8 张国产单卡单芯片加速卡，单卡要求如下： FP32 非 tensor 算力≥24Tflops； TF32 算力≥98Tflops； GPU 显存≥64GB GDDR6； BF16 算力≥196Tflops； GPU 显存带宽≥896GB/s。
5		标准接口	1、双口 1G RJ45 网卡； 2、双口 10G 无模块光纤网卡*4； 3、10G/1G 自适应 SFP+多模光模块*8。
6		电源	220V 电压、≥2700W 电源模块*4
7		万兆业务交换机	1、固化 48 个 10G SFP+以太网光口，8 个 100G/40G QSFP28 光口，配置 2 个电源插槽； 2、100G QSFP28 接口光缆不少于 2 根。
8			
（二）	智算资源管理云平台服务	**智算资源管理云平台服务围绕国产信息技术应用创新要求及人工智能发展要求，提供对国产 X86 架构 CPU、人工智能深度计算单元（DCU）及 GPU 的支持，有效整合和利用基础设施资源，实现基础设施层的统一管理和调度，提供虚拟化、容器、GPU 独占、DCU 等多元算力，以满足多样化的应用需求和用户场景。**	
1		GPU 直通和虚拟化	管理平台应同时支持主流 GPU 设备的直通和虚拟化，支持云服务器挂载和卸载 GPU 设备，支持设置 GPU 配额管理，可基于租户和项目提供 GPU 配额管理，同时支持查看 GPU 监控，可查看监控指标功耗限制、内存总量、温度、功耗、芯片频率、算力使用率、内存已用量、内存使用率等指标。
2		裸金属管理	支持裸金属申请、创建、修改、监控、告警、删除生命周期管理，裸金属支持软装版和智能网卡版裸金属服

			务，支持为裸金属打标签；支持对配置了算力加速卡的裸金属实例管理。
3		裸金属监控	支持查看 GPU 监控，支持配置了算力加速卡的裸金属实例管理，支持查看 GPU/DCU 监控信息，包括：指标功耗限制、内存总量、温度、功耗、芯片频率、算力使用率、内存已用量、内存使用率。
4		硬件设备直通	管理平台须支持 PCI（passthrough）透传方式使用服务器上的磁盘、GPU、USB、国产 GPU 物理设备；针对 GPU 设备支持 GPU 设备的启用、禁用操作，可查看 GPU 设备详细信息，包含：总线、类型、状态、容量、厂家、型号、所在物理主机、挂载实例、别名等信息。
5		多租户管理	支持多租户管理，提供统一的认证、授权、审计管理框架。支持设置内置用户、注册用户等，支持对每一种用户绑定角色类型，赋予相应的权限。
6		平台资源管理	支持用户对算力资源的各项服务进行重启、停止监控等操作。 支持用户对节点管理、网络管理、用户与配额管理等功能
7		运维与监控功能	支持提供一个针对平台系统组件、计算资源池、存储资源池、网络资源池、各种中间件、各种服务的全方位资源监控、统计、告警的功能。 1、支持通过监控面板查看监控数据的功能，将核心服务监控指标集中呈现在一张监控面板里。 支持自动监控和人工手动安装监控，支持自定义监控面板，支持自定义显示监控指标，支持按不同监控指标进行排序，支持导入导出监控数据。 2、支持用户从业务角度集中管理其业务涉及到裸金属服务器、带宽、数据库、弹性 IP 等资源。 3、支持不同维度指标配置告警规则，支持对多个服务器资源同时添加告警规则，支持随时对告警规则进行修改，支持对告警规则进行启用、停止、删除等灵活操作，支持告警规则的导入和导出。 4、支持所有智算服务的数量统计（包含正常和总数）。支持按主机集群维度、按物理节点维度、按项目维度统计各种智算服务资源。
（三）	人工智能大模型管理	**人工智能大模型管理平台服务是将 AI 算法模型及相关能力封装为服务，降低技术使用门槛；通过共享模型，提高资源利用效率；具备灵活的应用适配能力，能够快速融入各种业务场景，提供个性化的解决方案。**	
1		模型部署	1、模型部署功能支持将训练好的模型集成到生产环境中，使其能够在实际应用中处理数据和提供 AI 服务。在部署过程中，支持将模型的结构和权重、预处理/后处理逻辑一起封装，确保模型在部署后的正常运行。部署成功后支持提供可调用的服务，支持模型部署列表的上下线管理并支持根据模型状态、模型名称进行筛选和

	平台服务要求		搜索。 2、支持部署预置模型和用户训练后的自定义模型。
2		模型中心	模型中心用于展示和分享各类模型的公共空间，便于用户直观了解模型、发现新模型。 1、支持模型上架、模型删除、模型详情、模型仓库、模型广场等功能。 2、支持每个模型都有详细的介绍页面，包括模型的名称、所属领域、功能特点、API 示例、适用场景等信息。
3		模型体验	支持预置模型体验模块：为用户提供一个无需复杂设置，即可快速试用模型的环境，让用户在实际使用中感受模型的效果和性能，包括模型选择、文本对话、创建新对话等能力。
4		模型调优	模型调优模块为用户提供了一系列用于模型训练和调优的工具，帮助用户根据自己的需求对模型进行定制化开发。 1、支持模型调优列表功能，展示用户创建的所有调优任务及其相关信息，以及对每个调优任务，用户可以进行查看、删除、停止训练等操作。 2、支持训练新模型功能，支持用户使用自己的数据和平台提供的工具，基于基础模型训练一个全新的模型。在训练新模型时，支持用户进行一系列的参数设置。
5		数据管理	1、支持用户进行数据集的创建、管理和维护功能，确保用户可使用创建的数据进行模型训练和调优。 支持新建数据集,支持用户定义数据集的基本信息，如数据集名称、数据集类型、描述等信息。 2、支持数据集列表查看，支持列表展示用户创建的所有数据集，以及对数据集的相关操作，如数据预览、数据导入、新建版本、数据删除以及数据文件下载等功能。 3、支持数据向量化。支持把各种类型的数据转化为数值向量，以便进行后续的分析、建模和预测等任务。
6		智能体应用构建	支持快速构建智能体应用，创建过程中支持 AI 模型的温度、输出的 Token 数（0～1000000）等参数设置。支持通过引用知识库/非引用知识库的方式设置提示词模板，并可在线调试预览 Agent 智能体应用。
7		系统管理	支持 APIKEY 管理以及算力单元管理等功能。 1、支持提供 API KEY 的检索、生成、查看、删除等管理功能，并可动态配置 API KEY 列表的展示内容。 2、支持对 GPU 集群节点和加速卡的统一管理监控。包括节点数量、加速卡数量、加速卡显存总量、使用量、使用率占比、内存总量、使用量、使用占比的监测。对算力集群单个节点的监测包括节点名称、IP 地址、运行状态、加速卡型号、加速卡数量、加速卡显存、CPU

			使用率、内存使用率、磁盘使用率等多个监测维度。对加速卡的监测维度包括：各型号加速卡数量占比、各型号加速卡显存总量、使用量、显存使用率、各型号加速卡算力使用率 TOP3、各型号加速卡显存平均使用率、最高算力使用率等维度。并支持对单个加速卡细粒度的使用监测，包括设备 ID、加速卡型号、所属算力节点、显存使用率、算力使用率、芯片频率等维度信息。
（四）	托管机房	机房要求	机房基础环境符合国家机房相关标准，提供完善的空调、供配电、防雷接地、消防等基础环境设施服务，提供不低于 42U 标准机柜，满足设备托管要求。机房至少需接入双路市电，机柜配置双 PDU 并分别接入一路市电，机房配备 UPS、柴油发电机等应急设施。
（五）	网络安全要求		支持多台集群定制部署，集群部署时包括但不限于万兆业务网交换机及辅料费用等配套设施。 需提供对物理服务器集群所需的网络服务。 需通过 2 条冗余裸光纤接入无锡市新吴区电子政务外网网络核心。

附件二、服务考核要求

采购人将对中标人实施半年度考核，考核结果未达到考核要求的情况下，根据考核分值扣除项目合同费用。考核内容（满分 100 分，扣完为止）：

序号	一级指标项	二级指标项	分值	定义及标准
1	资源储备（10 分）	算力储备规模	10	算力资源池的实施应保障一定的冗余储备，可应对业务对算力需求的变化。 算力冗余规模占总规模，低于 10%且在 7 个工作日内未补充算力资源的一次扣 2 分。
2	服务运营（15）	资源调度效率	10	算力资源持续、稳定、合规地被使用。 未快速分配、释放和复用 AI 加速资源的一次扣 2 分。
3		服务便捷性	5	预置主流大模型（如 DeepSeek、Qwen、智谱 AI 等），并根据实际情况补充、更新大模型。 大模型缺失或更新不及时的每个扣 1 分
4	服务支撑（15 分）	需求响应及时性	5	用户提出需求时需在 2 个工作日内做出响应给出具体解决方案。不及时或未响应的一次扣 1 分。
5		服务完成及时性	10	根据制定的方案需在约定时间内完成服务需求。未及时完成的一次扣 2 分。
6	服务保障（20 分）	服务安全性	6	被国家、省级或市级通报网络安全事件的进行扣分。国家级通报一次扣 6 分，省级通报一次扣 3 分，市级通报一次扣 2 分。
7		服务可靠性	14	提供的算力服务需稳定可靠支撑业务。 经研判是由算力服务影响业务开展的一次扣 3 分。
8	应急外置（16）	应急处置能力	16	具备快速、高效的应急处置能力，确保在突发事件发生时，能够第一时间响应并控制事态发展。。 1. 一般事件（未造成明显影响） 未在规定时间内启动应急响应或处置不及时：每发生一次，扣 2 分； 2. 较大事件（造成一定影响但未扩散） 未及时处置，导致影响扩大或处理延误：每发生一次，扣 4 分。 3. 重大事件（造成业务中断等不良影响） 未及时处置，造成系统中断、数据异常等不良后果的每发生一次，扣 8 分；并视情

				节严重性启动合同违约条款。
9	用户满意度（8 分）	用户投诉量	4	是否收到用户投诉。每接到一次投诉扣 1 分。
10		投诉解决情况	4	接到投诉后处理结果用户不满意的一次扣 2 分。
11	日常运维（12 分）	运维执行有效性	6	应有完备的运维制度，对运维管理的落实情况进行检查。每发生一次未按照运维制度执行的扣 1 分。
12		运行报告完备性	6	应按用户需要提供相应的报告，不限于算力资源使用报告、大模型使用报告等方面。 未提供报告或报告不真实的一次扣 2 分。
13	服务过程材料（4 分）	资料台账规范及完整度	4	出现一次不规范或不完整扣 1 分。
14	其他	突出成绩和重大失误		本次考核周期的取得突出成绩或发生严重失误、过错且造成重大影响的视情况进行酌情加减分。加分上限为考核总分不超过 100 分。

考核要求：服务项考评得分>=95 分。采购人将依据服务情况对中标人进行整体考评，每半年度实施一次。考核分数≥95 分，不进行扣款；得分小于 95 分后，每少一分扣款 5000 元（每半年度扣款上限为 10 万元）。