附件 9

“先进计算与新兴软件”重点专项
2025 年度项目申报指南

为落实“十四五”期间国家科技创新有关部署安排，国家重点研发计划启动实施“先进计算与新兴软件”重点专项。根据本重点专项实施方案的部署，现提出 2025 年度项目申报指南。

本专项总体目标是：针对新型计算系统结构、新型存储架构、新兴软件与新兴计算场景，构建神经元计算系统、图计算系统、存算一体系统、拟态计算系统等新型计算系统，系统效能相比传统计算技术提升至少一个数量级；针对大规模数据存储与新型计算需求，研制内存池化与分布式存储、近数据处理与智能存储、持久数据存储系统等新型存储系统与关键技术，存储性能提升一个量级；突破软硬协同关键技术，在晶圆级集成、数据流、机密计算、云边端协同、自然人机交互等领域取得支撑技术突破，构建新型架构上的系统软件、人机物融合系统、软件智能化开发等生态体系，支撑我国信息技术和产业平稳快速发展。

2025 年度项目申报指南部署坚持需求导向、问题导向，拟围绕领域专用软硬件协同计算系统和新兴软件与生态系统等 2 个技术方向，启动 8 项指南任务，拟安排国拨经费 2.02 亿元。其中，青年科学家项目拟安排国拨经费 800 万元，每个项目 200

万元。共性关键技术类项目配套经费与国拨经费比例不低于1:1。

项目统一按指南二级标题（如1.1）的研究方向申报。申报项目的研究内容必须涵盖二级标题下指南所列的全部研究内容和考核指标，实施周期不超过3年。共性关键技术类下设课题数不超过5个，项目参与单位总数不超过8家。项目设1名项目负责人，项目中每个课题设1名课题负责人。其中，指南1.2、2.1、2.2各推荐渠道均可推荐申报，但申报项目中至少有一个课题由之江实验室作为承担单位。项目鼓励企业参与。项目攻关形成的主要软件成果应适配国产开源操作系统，并在国家级开源软件平台上发布。

青年科学家项目不要求对指南内容全覆盖，不再下设课题，项目所含参与单位总数不超过3家。项目设1名项目负责人，项目负责人年龄要求不超过40岁（1985年1月1日后出生）。原则上团队其他参与人员年龄要求同上。

每个指南任务拟支持项目数为1项。

1. 领域专用软硬件协同计算系统

1.1 新一代国产混合万卡智能计算系统（共性关键技术类/部省联动项目）

研究内容：面向人工智能快速发展的大规模计算需求，构建新一代国产混合万卡智能计算系统，研究适配不同国产GPU的万卡规模智能计算系统架构，构建算力供应链安全的计算服务能力。研究统一算子与编译技术，实现不同加速卡的算子覆

盖增广、对齐计算精度要求。研究服务器内高扩展无阻塞卡间互连技术、服务器间大规模低延迟组网技术。研究统一集合通信技术，实现不同国产加速卡间在单一通信域高效集合通信。针对大模型训练过程中数据存取大并发、低时延、强一致性等需求，研究数据高效布局与迁移策略、海量小文件元数据管理机制、高效数据压缩技术等，提升数据的访问性能。研究大模型训练的非均匀并行策略框架，结合算网存协同优化技术。研究异构智算系统的模拟平台，实现快速准确模拟，降低调试与调优成本。集成算力调度、故障诊断与恢复、数据处理、模型训练、算法服务等工具套件，支持万亿参数大模型的全周期计算工序管理和高效能训练，拓展智算系统应用场景，面向三个以上典型领域进行应用示范，加强标准体系和生态建设。

考核指标： 智算系统支持智能算力不低于 3E FLOPS@FP16，支持加速卡总数量不低于 1 万张，支持训练万亿参数大模型；万卡规模训练平均每卡实测算力不低于 150TFLOPS，万卡集群训练卡利用率 MFU 不低于 50%；混合智算系统包含不少于 3 种不同品牌国产加速卡。平均每卡数据 I/O 独占时间占总训练时间低于 10%，与同等配置的业界主流存储系统相比性能提升 30%以上。检查点读写时间占总训练时间控制在 1%以内。针对不少于 3 款国产主流智能加速卡，完成训练框架、通信库、算子库等软件的移植和适配；可实现 3 种国产加速卡间高速直连通信，实现万卡规模在单一通信域高效集合通信，数据跨节点访问延时低于 10μs，集合通信性能相比现

有系统提升60%以上。计算系统资源按每卡每小时计量的周可用率达95%以上，支持千卡到万卡弹性调度；支持包括DeepSeek、GLM、FLM中不少于两种国产千亿参数以上开源模型，其中训练语料规模不低于1.5T tokens。支持故障诊断与恢复，故障定位速度达到分钟级。面向三个以上典型领域进行应用示范。

有关说明：由之江实验室牵头组织实施。

关键词：智能计算、国产智算加速卡、混合训练。

1.2 面向高质量大模型服务的跨域分布式边缘算力网络关键技术研究及示范应用（共性关键技术类/部省联动项目）

研究内容：面向大模型服务需求爆发式增长与服务质量极速提升的双重趋势，研究并突破跨域分布式边缘算力资源协同调度的关键技术，构建高可靠、高实时、高效能、高性价比、大规模的泛在算力基础设施，支撑大规模跨域的大模型业务服务和运营。感知边缘侧的算力资源信息，包括计算、存储、通信、服务等不同类型资源与服务的感知。对由多种算力单元组成的异构处理体系进行标准化的统一，针对网络、存储等多维资源，从多个维度进行算力资源的建模和评估。研究差异化用户计算意图模型，提出能够确保用户多样化需求的算力分配方案，并结合预测算法预判未来算力需求的变化。研究如何有效降低系统熵值，实现多类型、多层次异构算力资源的精准供需调度。锚定边缘算力网络的市场性，研究基于市场驱动的边缘算力资源流动模型，探索算力市场价格制定、算力网络多方动

态博弈问题，构建高效协同的算力交易机制，强化市场化闲散碎片算力的统筹汇聚能力。着重针对大模型的训练和推理任务，研究边缘算力网络中分布式模型训练和推理的高效调度与优化算法，重点解决大模型服务的并行化、资源调度与推理加速问题。针对网络敏感、服务质量敏感、时延敏感等三类不同场景的业务进行应用示范，构建针对大模型服务的广域分布式边缘算力网络的试验环境，并开展运营级示范。

考核指标: 构建不少于 1000 个分布式算力节点的分布式边缘算力网络，整合超 100T 带宽、10 万核 CPU、1000P（FP16 非稀疏）以上的分布式智算算力，广域范围覆盖 100 个城市智算节点，并兼容至少三种国产芯片。算力需求预测准确率达到 95%以上，用户差异化需求的满足率提升至 90%以上。用户意图感知的覆盖率到达 95%，误判率不超过 5%。支持万级并发用户的多意图处理。分布式算力网络系统整体熵降低 30%，用户服务响应时间减少 20%。在实时计算场景下，分布式用户的需求处理时间不超过 100ms，常规计算场景下不超过 500ms。提升分布式算力时空调度能力和市场化闲散碎片算力的汇聚能力，资源利用率提高 25%以上，交易效率提升 20%以上，算力价格波动稳定性提升 10%，交易成本效益比不低于 1000GFLOPS/RMB，算网交易供需比接近 1。纳管、调度并运营跨域多源异构的千卡级分布式边缘算力网络基础设施，为大模型业务提供高质量服务，分布式方式部署 10 种以上的各类大模型，实现模型训练和推理较原生版本加速 10 倍以上，开展不

少于5种典型大模型计算场景下的规模化运营级应用示范，日处理token数量超2000亿。

关键词：边缘算力网络、算力感知、算力度量、算力供需。

1.3 面向国产智算系统的网存算融合加速方法（青年科学家）

研究内容：针对大模型训练中计算加速需求，研究面向国产智算系统的网存算融合加速方法，研究智能网卡、可编程交换机、计算型存储盘与国产GPU等设备的高效协同技术。研究典型智算通信原语在网络设备上的卸载方法；研究典型智算算子在近数据计算硬件上的卸载方法；研究与RDMA、NVMe等标准兼容的网存算融合协议；研究集群环境下适配网存算融合协议的作业放置、路由和调度等策略；研发支持网存算融合加速的国产智算系统原型。

考核指标：研制一套网存算融合加速的国产智算系统原型，支持交换机、智能网卡和计算型存储盘等不少于4类硬件的融合加速；网存算融合协议可适配不少于3种国产GPU；支持AllReduce等不少于5种通信原语在网络设备上的卸载，支持GEMV、SoftMax等不少于5种算子在近数据计算硬件上的卸载；支持集合通信、数据预处理、数据存储等不少于5类计算过程加速；网络吞吐性能提升20%以上，存储吞吐性能提升20%以上，系统整体性能提升30%以上。

关键词：网存算融合、计算加速、可编程网络设备、计算型存储、国产GPU。

1.4 面向国产智算平台的推理模型持续后训练系统（青年科学家）

研究内容：针对新一代推理模型的国产硬件适配与高效训练的需求，研究面向国产智算平台的推理模型持续后训练框架及优化技术。研究面向持续后训练的多模态推理链数据的高效抽取、合成及知识关联技术；研究推理链数据实时制备流水线技术；研究推理模型持续知识扩充与后训练加速技术；研发适配异构国产智算平台的推理模型持续后训练框架和工具集，并在典型场景下开展应用验证。

考核指标：研发一套面向国产智算平台的推理模型持续后训练系统原型；后训练推理链数据实时制备流水线支持文本、图像、关系表等不少于 5 种模态数据的高效融合与动态知识关联；动态推理链数据生成端到端延迟 TPOT(Time Per Output Token）不超过 50ms，逻辑一致性验证准确率不低于 95%；后训练框架支持不少于 3 种国产 GPU，硬件资源平均 MFU 不低于 50%；推理模型持续后训练系统原型在 2 种以上典型场景进行应用验证。

关键词：国产智算集群、推理模型、实时推理链数据制备、持续后训练。

1.5 面向大模型推理加速的 GPU-PIM 异构算力协同方法（青年科学家）

研究内容：针对大模型推理的高吞吐低延迟需求，研究支持国产 GPU 和 PIM 异构算力协同的新型计算架构及方法。研究

多种典型访存密集型智算算子卸载至 PIM 的方法；研究面向 GPU-PIM 异构算力的计算任务调度、模型划分、数据放置等推理优化方法和技术；研究 GPU-PIM 架构下的统一共享内存管理方法。

考核指标：研发一套 GPU-PIM 异构算力系统原型，可适配不少于 3 种国产 GPU；支持 GEMV 等不少于 5 种访存密集型算子卸载至 PIM；实现硬件资源利用率提升 30%以上，数据访问开销降低 10%以上；典型智算算子计算效率提升 15%以上；相比同等算力的 GPU 集群，国产大模型（如 DeepSeek、Qwen）推理性能提升 20%以上。

关键词： GPU-PIM 协同加速、大语言模型推理、关键算子布局、数据一致性管理、调度策略。

1.6　面向国产智算系统的低精度混合训练框架研究（青年科学家）

研究内容：面向国产智算系统中大模型训练性能优化需求，研究低精度混合训练框架及其关键技术。研究支持低精度混合训练的统一训练框架，支持不同精度间的自动转换和误差补偿；研究低精度数据格式表示范围与模型精度损失评估方法；研究低精度训练对模型收敛性和泛化能力的影响模型；研究低精度训练的计算通信优化方法；研究低精度训练的算子优化方法。

考核指标： 研制一套低精度混合训练框架，适配包含不少于 3 种国产 GPU 的异构混合集群；提出一套适合国产 GPU 生态的低精度计算标准建议，支持 FP8、FP6、FP4、INT8 等不少

于 5 种低精度格式混合；完成对 Transformer、CNN 等不少于 3 类深度学习模型训练的应用验证；在相同 GPU 算力下，低精度混合训练框架相比主流训练框架，训练速度提升 50%，通信开销降低 50%，模型精度损失不超过 1%。

关键词：低精度计算、混合训练框架、国产 GPU。

2. 新兴软件与生态系统

2.1 新一代智能化软件的高效能资源管理方法与服务系统（共性关键技术类/部省联动项目）

研究内容：面向新一代智能化软件的算力需求，研究软件定义的算力基础设施体系结构模型，支撑大规模智能化软件的资源服务化方法；研究万卡级规模智算集群的资源管理方法，支持稠密模型、混合专家模型等多种模型架构的高效训练；研究支撑智能化软件的分离式推理架构，支持动态灵活的按需配置和并行策略优化；研究面向智能化软件不同领域的模型微调技术，支持基于基础模型之上多个微调模型的高效微调和推理；研究智能化软件的异构硬件互操作方法，支持智能化软件在不同异构硬件的互操作和迁移；关键技术在智能化软件典型场景下开展示范应用。

考核指标：单机本地支持 DeepSeek、Qwen、Llama 等国内外主流大模型；万卡级集群训练任务 MFU（Model FLOPs Utilization）提升不低于 20%；推理任务支撑序列长度不低于百万，在预填充阶段的 TTFT（Time to First Token)降低不少于 30%，在解码阶段的 TPOT（Time Per Output Token）降低不少于 20%；

支持多 LoRA 模型的微调和推理，微调吞吐量提升不低于 20%，推理吞吐量提升不低于 30%；支持不少于 3 种国产异构硬件的互操作；研制原型系统并开源，云端验证规模不少于千卡，端侧验证规模不少于百万设备。

关键词：智能化软件、大模型、算力资源管理。

2.2 面向垂域场景的智能软件生态支撑技术与系统（共性关键技术类/部省联动项目）

研究内容：研究面向垂域场景的大模型智能软件应用开发框架与工具，实现基于国产硬件平台的智能体等软件应用的快速构建、调试和测试；研究面向垂域场景的模型-子任务智能适配及模型动态优化组合技术，实现面向多样化垂域场景的大/小模型混合编排及基于国产计算硬件的模型高效适配技术；通过模型及应用的协同迁移、领域知识自适应融合、模型轻量化等方法实现面向云边端环境的智能软件应用高效部署与运行支撑；研究垂域智能软件应用的运维技术，实现智能软件应用的可观测性分析、协同监控、根因分析，研究垂域智能软件应用的生态链构建机理、溯源追踪机制以及安全脆弱性分析机制；研究面向垂域场景的软件自动化测试技术；构建国内自主的智能应用开源社区，在政企服务、科学计算、金融等领域开展示范应用。

考核指标：智能软件应用的开发效率提升 30%以上，支持 3 种模型并行策略的自动构建;模型与子任务需求的匹配正确率达到 80%，支持 20 种以上垂域智算任务的大/小模型混合编排，

支持至少 3 种以上国产智能计算硬件协同计算；基座大模型支撑不少于 30 种领域知识的自适应融合，融合推理的准确率提升 50%，融合推理的效率提升 4 倍；智能软件应用的关键组件监控覆盖达到 95%以上，性能瓶颈定位与诊断准确率达到 80%以上，平均故障定位时间缩短至 5 分钟以内；智能软件应用的依赖与溯源分析的准确率达到 70%以上；垂域自动化测试软件缺陷检出率不低于 80%，软件缺陷检出正确率不低于 70%；形成开源社区，构建垂域智能应用不少于 1 万个，覆盖开发者超过 100 万；支撑政企服务、科学计算、金融等垂域生态的构建，形成至少 3 个典型领域智能应用生态建设的示范应用。

关键词：垂域大模型、开发运维、智能应用生态。

“先进计算与系统”重点专项
2025年度项目申报指南形式审查条件要求

本年度指南均采取一轮申报程序，申报项目须符合以下形式审查条件要求。

1.推荐程序和填写要求

（1）由指南规定的推荐单位在规定时间内出具推荐函。

（2）申报单位同一项目须通过单个推荐单位申报，不得多头申报和重复申报。

（3）项目申报书内容与申报的指南方向相符。

（4）项目申报书及附件按格式要求填写完整。

2.申报人应具备的资格条件

（1）项目（课题）负责人应为60周岁以下（1965年1月1日及以后出生），具有高级职称或博士学位，每年用于项目的工作时间不得少于6个月。

（2）青年科学家项目负责人应具有高级职称或博士学位，40周岁以下（1985年1月1日及以后出生），原则上团队其他参与人员年龄要求同上。

（3）港澳申报人员应爱国爱港、爱国爱澳。受聘于内地单位的外籍科学家及港、澳、台地区科学家可作为项目（课题）负责人，聘用期应覆盖所申报项目（课题）的执行期，并应提供相应聘用材料。其中，全职受聘人员应由内地聘用单位提供

全职聘用的有效材料，非全职受聘人员应由双方单位同时提供聘用的有效材料。

（4）参与重点专项实施方案或本年度项目指南编制的专家，原则上不能申报该重点专项项目（课题）。

（5）诚信状况良好，无在惩戒执行期内的科研严重失信行为记录和相关社会领域信用“黑名单”记录。

（6）中央和地方各级国家机关及港澳特别行政区的公务人员（包括行使科技计划管理职能的其他人员）不得申报项目（课题）。

（7）项目申报人员满足申报查重要求。

3.申报单位应具备的资格条件

（1）中国大陆境内注册的科研院所、高等学校和企业等独立法人单位，或由内地与香港、内地与澳门协商确定的港澳科研单位。

（2）中央和地方各级国家机关不得牵头或参与申报。

（3）注册时间在 2024年11月30日及以前。

（4）诚信状况良好，无在惩戒执行期内的科研严重失信行为记录和相关社会领域信用“黑名单”记录。

4.本重点专项指南规定的其他形式审查条件要求

青年科学家项目不再下设课题，项目参与单位总数不超过3家。

本专项形式审查责任人：丁莹