

题目编号：XH-202608

基于国产软件栈大模型推理前沿算子优化 比赛方案

一、发榜单位

沐曦集成电路（上海）股份有限公司

二、题目名称

基于国产软件栈大模型推理前沿算子优化

三、题目介绍

我国在 AI 算法创新与应用落地方面已取得显著成就，主流大模型在算法层面已与国际先进水平并跑乃至部分领跑。然而绝大多数可公开下载和部署的大模型，其底层推理代码与性能优化均深度适配于国际主流 AI 生态。这种深度绑定，在中美科技竞争长期化、复杂化的今天，已成为我国人工智能产业自主发展的潜在风险。在此背景下，需要构建基于全国产软件生态的核心算子库，为国产大模型和国产算力芯片的协同创新与规模化应用，浇筑自主可控的软件“地基”。

在此背景下，国产自主研发并于 2025 年开源的 TileLang，作为一种专为 AI 算子开发设计的领域专用语言（DSL），其核心是将高性能计算中的“分块技术”（Tile）作为首要优化对象，自动推导并实施适合目标硬件的优化策略，让开发者更专注于算法逻辑本身。

TileLang 具有高效率、自动化、强兼容和开放生态的特性，成为构建国产软件生态中重要的开源项目。沐曦股份 MXMACA 软件栈已与 TileLang 完成深度适配和验证，可作为后端接入 TileLang，构建贯穿上层框架到底层硬件的、全国产化的 AI 软件生态。

本次比赛聚焦于大模型推理中 3 个关键前沿算子，以 TileLang 作为核心开发语言，以沐曦股份 MXMACA 作为软件栈后端，以曦云系列 C500 处理器为硬件算力，开展基于全国产软件生态的深度性能优化与实践。选择的算子包括：

1. Fused Moe Gemm
2. DeepSeek V3/R1 所用 Multi-Head Latent Attention (QK dim576 + VO dim512)
3. Native Sparse Attention

其中 Fused moe gemm 是一种针对混合专家模型的关键核心算子。混合专家模型通过门控网络动态选择少数专家进行处理，通过稀疏激活特性带来参数规模的优势，但也带来计算细碎化，硬件算力资源利用不充分的问题。Fused moe gemm 通过计算融合优化技术解决这一技术挑战，该算子输入多个独立的专家权重矩阵、经过路由分配的输入激活张量，输出所有专家各自的计算结果，经后续处理（如聚合）后形成该 MoE 层的最终输出。该算子的技术核心在于通过将多个独立的专家矩阵

乘法（GEMM）分组并融合为一个统一的内核进行计算，提升硬件算力利用率，加速大模型推理。

DeepSeek V3 中的 MLA（Multi-head Latent Attention）算法通过将 KV cache 压缩为低秩潜在向量，在保持与标准多头注意力相当性能的同时，将 KV cache 显存占用大幅减小，显著降低了长序列推理的显存瓶颈；其创新的解耦旋转位置编码和优化的矩阵吸收机制，不仅提升了推理效率，还通过减少访存次数和增强计算强度，在大模型场景下实现了更高的吞吐量和更低的延迟，特别适合需要处理超长上下文的应用场景。有赖于 MLA 算法和 MOE 等关键技术，DeepSeek V3-671B 的性能在很长一段时间内领跑国际先进开、闭源模型的水平。

DeepSeek 提出的 Native Sparse Attention 算法通过硬件原生的稀疏计算模式，将超长文本的注意力计算复杂度从 $O(n^2)$ 降至 $O(n\sqrt{n})$ ，在保持模型性能几乎无损的同时，实现了数倍的推理速度提升和 50% 以上的显存节省；NSA 创新地通过分块稀疏模式和动态稀疏度调整机制，不仅避免了传统稀疏注意力中的负载不均衡问题，还通过优化的内存访问模式显著提高了 GPU 利用率，特别适合长上下文场景下的高效推理。实验表明，在 64k 长度序列上，NSA 在解码、前向传播和反向传播各个阶段都实现了显著加速，同时在通用基准测试、长

上下文任务和指令推理方面保持或超越了全注意力模型的性能，有效解决了长上下文建模中的计算瓶颈问题。

本次比赛聚焦于上述算子的性能优化，要求参赛选手理解算子的计算逻辑，TileLang 的基本语法和 GPU 算子的常见优化方法，在保证精度的同时提升算子性能。大赛将提供编程环境，功能和性能测试集，助力团队快速上手并开展实践。

四、参赛对象

学生赛道：参赛对象为 2026 年 6 月 1 日以前正式注册的国内全日制非成人教育的普通高等学校在校专科生、本科生、硕士和博士研究生（不含在职研究生），以及全日制职业教育本科、高职高专在校学生，可通过学生赛道申报作品参赛。

参赛对象可以团队或个人形式参赛，每个团队不超过 10 人，每件作品可由不超过 3 名指导教师进行指导。可以跨专业、跨学校、跨地域组队，但同一团队所有成员均应符合本赛道相关年龄、身份要求。每件作品只可由 1 所高等院校、科研院所等作为参赛主体提交申报。

五、答题要求

（一）初赛作品要求

根据提供的基础样例，使用 TileLang 在 MXMACA 平台上实现和优化 Fused Moe Gemm 算子，并需要在基础样例 Fused Moe Gemm 算法测试 API 端到端耗时性能

上有提升作为晋级决赛的基本要求。评判标准分两个层级：基本功能和性能加分层级。

1. 基本功能层级评判标准：

Fused Moe Gemm 功能 test 集通过率 99%以上，并有参考样例性能的提升。

2. 性能加分层级评判标准：

在满足基本功能层级评判标准的基础上，全部参赛选手实现的 **Fused Moe Gemm** 算法测试 API 端到端耗时（warmup 10 次，重复 100 次取平均耗时），按耗时长短排名，耗时越短排名越高，并且会在决赛有 10 分评分会根据初赛性能排序作为依据如下：

第 1 名 +10 分

第 2 名 +9 分

第 3 名 +8 分

第 4 名 +7 分

第 5 名 +6 分

第 6 名 +5 分

第 7 名 +4 分

第 8 名 +3 分

第 9 名 +2 分

第 10 名 +1 分

（二）决赛作品要求

1. 根据基础参考样例，使用 **TileLang** 提交 MLA 或

者 NSA 算子实现并优化性能，考虑到 2 个算子难度有差异，请选手从 2 个算子任选一个作为决赛参赛评比（注意：MLA 技术分最高 80 分，则决赛最终 $80 \times 0.6 = 48$ 分），2 个算子评价如下：使用 TileLang 在 MXMACA 平台上实现和优化 DeepSeekV3 MLA 算子【技术分满分：80 分】，技术分拆分如下：

基本功能层级评判标准（技术分：40 分）：

MLA 功能 test 集通过率 99%（按照通过 case 数目折算得分，例如：80%覆盖， $40 \times 80\% = 32$ 分）。性能加分层级评判标准（技术分：40 分）：

在满足基本功能层级评判标准的基础上，全部参赛选手实现的 MLA 算法测试 API 端到端耗时（warmup 10 次，重复 100 次取平均耗时），按耗时长短排名，耗时越短排名越高：

第 1 名 +40 分

第 2 名 +38 分

第 3 名 +36 分

第 4 名 +34 分

第 5 名 +32 分

第 6 名 +30 分

第 7 名 +28 分

第 8 名 +26 分

第 9 名 +24 分

第 10 名 +22 分

第 11 名 +20 分

第 12 名 +18 分

第 13 名 +16 分

第 14 名 +14 分

第 15 名 +12 分

第 16 名 +10 分

第 17 名 +8 分

第 18 名 +6 分

第 19 名 +4 分

第 20 名 +2 分

2. 使用 TileLang 在 MXMACA 平台上实现和优化 NSA【技术分满分：100 分】，技术分拆分如下：

基本功能层级评判标准（技术分：40 分）：

NSA forward API 功能 test 集通过率 99%（按照通过 case 数目折算得分，例如：80% 覆盖， $40 \times 80\% = 32$ 分）。

性能加分层级评判标准（技术分：60 分，例如：满分客观分为： $50 \times 0.3 = 15$ 分）：

在满足基本功能层级评判标准的基础上，全部参赛选手实现的 NSA 算法使用 64K seqLen 的输入测试 API 端到端耗时（warmup 10 次，重复 50 次取平均耗时），按耗时长短排名，耗时越短排名越高：

第 1 名 +60 分
第 2 名 +57 分
第 3 名 +54 分
第 4 名 +51 分
第 5 名 +48 分
第 6 名 +45 分
第 7 名 +42 分
第 8 名 +39 分
第 9 名 +36 分
第 10 名 +33 分
第 11 名 +30 分
第 12 名 +27 分
第 13 名 +24 分
第 14 名 +21 分
第 15 名 +18 分
第 16 名 +15 分
第 17 名 +12 分
第 18 名 +9 分
第 19 名 +6 分
第 20 名 +3 分

六、作品评选标准

提交作品最终评价及权重如下：

类别	评审维度	权重	说明
客观 评测	技术实现分	60%	根据决赛选择的题目的技术得分*0.6 给出技术实现分（备注：MLA 技术最高分 80 分，NSA 技术分最高 100 分）。
	初赛排名加分	10%	根据初赛名次取前 10 规则给与相应分数。
主观 评测	展示质量	20%	根据提交代码质量、Demo 视频与测试报告完整度给分。
	创新性与可扩展性	10%	是否体现独特机制、具备推广潜力和上游合并社区价值。

七、作品提交时间

2026 年 5 月至 9 月上旬，各参赛团队选择榜单中的题目开展研发攻关，各高校、科研机构等组织协调机构应组织学生参赛，安排专业人员给予指导，为参赛团队提供支持保障。

2026 年 9 月 5 日前，各参赛团队要向发榜单位完成作品提交，具体要求详见作品提交方式。

2026 年 9 月 20 日前，由发榜单位完成初审，确定入围终审擂台赛的晋级作品和团队。

2026 年 10 月，安排专门团队提供帮助和指导，各晋级团队完善作品。

2026 年 11 月，组织终审擂台赛，角逐“擂主”。

八、参赛报名及作品提交方式

（一）报名方式

（1）参赛选手登录“挑战杯”官网 www.tiaozhanbei.net，在“揭榜挂帅”擂台赛报名入口注册账号，登录大赛申报系统在线填写报名信息。报名信息提交后，下载打印系统生成的报名表。

（2）申报人在报名表对应位置加盖所在学校或所在单位公章。

（3）将盖章版报名表扫描件上传至报名系统，等待系统审核。请参赛选手注意查看审核状态，如审核不通过，需重新提交。

（4）系统开放报名时间为 2026 年 5 月 30 日—6 月 30 日，逾期后系统将自动关闭报名功能。

（二）作品提交方式

请已在官网报名成功的团队，于 9 月 5 日前将盖章的参赛申报表 pdf、作品所有相关材料发送至发榜单位邮箱 opensource@metax-tech.com。作品的提交除提到的客观评测外，参赛团队应将所有要求的材料，包括技术方案文档（PDF）、日志与结果文件、展示视频（10 分钟，总决赛）、算法介绍 PPT（总决赛）、源代码或 Notebook（可选，建议提交或开源）打包成.zip 格式压缩包。压缩包命名方式为：申报人所在单位－申报人姓名－作品名称－联系电话（例如：XX 大学－张 XX－XX 方案－手机号）。提交具

体作品时，务必一并提交 1 份报名系统中审核通过的参赛报名表（所有信息与系统中填报信息保持严格一致）。以上材料无需在“挑战杯”官网提交。

1. 算法设计报告：详细说明 **kernel** 设计的技术方案、创新点和实现步骤，给出算法伪代码

2. 功能测试报告：针对发榜单位提供的功能测试集合，提供完整的参数设置、计算结果的精度与 **golden** 校验对比报告

3. 性能测试报告：针对发榜单位提供的性能测试集合，提供完整的参数设置、性能测试结果分析报告

4. 源代码：用 **TileLang** 实现算子的完整源码代码（包括功能和性能测试源码），编译、运行 **test/benchmark** 方式的 **README.md** 文档

除参赛报名表外，各参赛组提交的文档、源代码和模型文件不得携带任何参赛学校、老师和学生的个人信息。同时，各参赛团队在提交作品时，同步报送 1 份经报名系统审核通过的参赛报名表，报名表所有信息须与系统内填报内容完全一致。

九、赛事保障

1. 算力资源支持：参赛选手在完成报名后，提供对应线上的曦云 **C500** 在线算力资源。

2. 技术培训：赛前赛中至少组织 2 场线上培训，提供培训回放及答疑文档；另外届时会根据实际情况，决定是否组

织线下的专场培训。

3. 专家指导：建立线上答疑社群，由沐曦股份技术团队指导，定期回复技术问题。

4. 交流平台：在沐曦股份开发者社区开设赛事专属社区板块，支持参赛团队分享经验、交流问题，促进技术共创。

5. 技术文档和课程

(1) 提供 TileLang/MXMACA 技术文档

(2) 提供算子 baseline 示例

(3) 提供测试集、评测脚本

(4) 配备技术答疑团队

(5) 组织线上技术讲解/Q&A

十、设奖情况及奖励措施

1. 设奖情况

按参赛作品数量比例设奖，原则上评出“擂主”1个、特等奖5个，一等奖5个、二等奖6个、三等奖8个，获奖比例不超过参赛作品总数的30%，从特等奖中角逐出擂主团队。

2. 奖励措施

奖金奖励：擂主奖励10万元/个（叠加特等奖后奖金），特等奖奖励2万/个，一等奖奖励1万元/个，二等奖奖励0.5万元/个，三等奖奖励0.2万元/个。

高端 GPU 奖励：擂主团队2张 GPU 加速卡（不叠加特等奖 GPU 奖励），特等奖团队1张 GPU 加速卡。

实习奖励：优秀获奖者有机会参与之江&沐曦股份“南湖之新”联合培养计划；所有获奖团队可获得赛事荣誉证书。

曝光支持：获奖作品可在沐曦股份开发者社区展示，作品可提供成果孵化与应用推广支持。

备注：从特等奖中角逐出擂主团队，奖金 10 万元已叠加特等奖奖金，擂主不叠加特等奖 GPU 奖励。

3. 奖金发放方式

比赛结束后，单位比赛专班工作人员与获奖团队取得联系，填写奖金申请表，赛事终审结果公示无异议后（公示期约 1 个月），在 30 个工作日内通过银行转账一次性发放至团队负责人指定账户，上述所列奖金均为税后奖金。

十一、比赛专班联系方式

1. 专家指导团队

顾问专家：马老师，联系电话：13811784391

顾问专家：芦老师，联系电话：15395820130

顾问专家：刘老师，联系电话：18210506627

负责比赛期间技术指导保障。

2. 赛事服务团队

联络专员：杨老师，联系电话：15201842467

联络专员：章老师，联系电话：13501701786

负责比赛期间组织服务及后期相关赛务协调联络。

3. 联系时间

比赛期间工作日（9:00 – 17:00）

附：发榜单位简介

沐曦集成电路（上海）股份有限公司（股票代码：688802.SH）成立于2020年9月，于2025年12月成功登陆科创板。总部位于上海，并在北京、南京、成都、杭州、深圳、武汉、长沙等地设立全资子公司及研发中心。作为一家专注于全栈 GPU 芯片及解决方案的集成电路设计企业，致力于打造世界一流的 GPU 芯片及计算平台，成为数字经济的算力基石。

公司拥有技术完备、设计和产业化经验丰富的团队，核心成员平均拥有近20年高性能 GPU 产品端到端研发经验，曾主导过十多款世界主流高性能 GPU 产品研发及量产。目前，公司已推出全面覆盖人工智能训练和推理、通用计算、图形渲染和科学智能等场景的四大序列产品，并配套自研 MXMACA 软件栈，真正实现了“软硬协同”，满足“高能效”和“高通用性”的算力需求。2025年起，公司坚持“开放协同、自主可控”的方向，全面推进计算生态以及产业生态的建设。在计算生态层面，公司以自主研发的 MXMACA 全栈软件栈为核心，积极拥抱开源，打造自主、开放、兼容的通用计算开源生态；在产业生态层面，公司依托“1+6+X”战略布局，以数字算力底座为基，持续深耕金融、医疗健康、能源、教科研、交通、大文娱六大重点行业，同步积极探索具身智能、低空经济等新兴领域，为数字经济与新质生产力发展提供坚实算力支撑。