PR #49544 完整报告

vllm-project/vllm

[ROCm][Perf] gfx942: use FlyDSL fp8 MQA logits kernel (ROCm/aiter#3913)

合并时间：2026-08-17 00:50

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49544

执行摘要

- 一句话：gfx942 用 FlyDSL 替换 Triton fp8 MQA logits 内核
- 推荐动作：值得关注 ROCm 性能优化路线的读者精读。该 PR 设计极其克制：1 个文件、条件分支内 drop-in 替换，gfx950 等路径零改动，风险面小。值得学习的是其验证流程——AITER 版本确认（release notes 溯源）、双模型（GLM-5.2-FP8 与 DSv4 Flash）端到端基准、GSM8K/NIAH 精度测试以及合并前维护者触发 CI 复跑。对希望理解 vLLM 如何在 ROCm 上替换 Triton 内核的工程师，这是一个干净的参考案例。

功能与动机

在 gfx942 上，DeepSeek-V4 Flash 稀疏注意力的 fp8 MQA logits 索引器内核是长上下文 prefill 的显著瓶颈。关联的 ROCm/aiter#3913 提供了比 Triton 内核快 1.5-2.5 倍的 FlyDSL 内核，且不依赖操作数 dtype 配置。PR body 中的基准数据显示，ISL=128K 时 TTFT 中位数从 42,839ms 降至 21,871ms（-49%），输出吞吐从 52.44 提升到 85.34 tok/s（+63%），且短上下文无回归。

实现拆解

1. 入口与背景：变更位于 vllm/v1/attention/ops/rocm_aiter_mla_sparse.py 的 rocm_fp8_mqa_logits 函数中，该函数负责 DSV4 Flash 稀疏注意力的 FP8 MQA logits（索引器）内核分发。此前在 _ON_GFX942 且 AITER 启用时，走 vendored Triton 内核 fp8_mqa_logits_gfx942，并带有 TODO(ganyi) 的临时工作区注释。
2. 核心替换：在 _ON_GFX942 and rocm_aiter_ops.is_enabled() 分支中，将导入从 vllm.v1.attention.ops.triton_fp8_mqa_logits.fp8_mqa_logits_gfx942 改为 aiter.ops.flydsl.flydsl_fp8_mqa_logits，调用参数（q、k_fp8、scale、weights、cu_seqlen_ks、cu_seqlen_ke）与返回值语义完全一致，是纯 drop-in 替换，同时删除了临时 TODO 注释。之所以能直接替换，是因为 AITER 已将该内核合入主线（v0.1.19 起包含），无需再维护 vendored 版本。
3. 配套验证：未新增单元测试，依据 vllm/attention/ops/rocm_aiter_mla_sparse.py 的分支判断自动选择内核。作者在 PR 评论中补充了 DSv4 Flash 的端到端复测数据，并在合并前由维护者触发 /ci run 跑完整 CI。精度侧提供了 GSM8K（0.9416）与 NIAH 128K（8/10）验证。该改动依赖 AITER v0.1.19+，与当前 docker/Dockerfile.rocm_base 固定的版本一致。

关键文件：

- vllm/v1/attention/ops/rocm_aiter_mla_sparse.py（模块 注意力后端；类别 source；类型 core-logic；符号 rocm_fp8_mqa_logits）：唯一变更文件。在 rocm_fp8_mqa_logits 函数中，将 gfx942 + AITER 路径的 vendored Triton 内核替换为 aiter.ops.flydsl.flydsl_fp8_mqa_logits，删除临时 TODO，属于本次性能优化的核心改动。

关键符号：rocm_fp8_mqa_logits

关键源码片段

vllm/v1/attention/ops/rocm_aiter_mla_sparse.py

唯一变更文件。在 rocm_fp8_mqa_logits 函数中，将 gfx942 + AITER 路径的 vendored Triton 内核替换为 aiter.ops.flydsl.flydsl_fp8_mqa_logits，删除临时 TODO，属于本次性能优化的核心改动。

```
# vllm/v1/attention/ops/rocm_aiter_mla_sparse.py（按本 PR 变更整理）

# FP8 MQA logits 索引器内核的分发函数：
# 在 gfx942 且启用 AITER 时使用 FlyDSL 内核，其余平台继续走原有 Triton 实现。

def rocm_fp8_mqa_logits(
    q,
    kv,
    weights,
    cu_seqlen_ks,
    cu_seqlen_ke,
):
    """FP8 MQA logits 内核分发。

    kv 为 (k_fp8, scale) 元组；返回形状为 [M, N] 的 fp32 logits。
    """
    k_fp8, scale = kv

    # 仅 gfx942 且 AITER 启用时切换到 FlyDSL 内核。
    # flydsl_fp8_mqa_logits 与旧 Triton 内核的参数、返回值和
    # clean_logits 语义完全一致，是 drop-in replacement，调用方无需改动。
    if _ON_GFX942 and rocm_aiter_ops.is_enabled():
        from aiter.ops.flydsl import flydsl_fp8_mqa_logits
        return flydsl_fp8_mqa_logits(
            q,
            k_fp8,
            scale,
            weights,
            cu_seqlen_ks,
            cu_seqlen_ke,
        )

    # 其余路径（gfx950、非 AITER 等）继续走原有 Triton 实现……
```

评论区精华

> tjtanaa：which aiter version is this validated on??

维护者首先关心 AITER 版本兼容性，担心上游 v0.1.16.post5 不含 PR 3913，一度将 PR 置为 onhold。

> akii96：vLLM 当前 Dockerfile.rocm_base 已 pin AITER v0.1.19，FlyDSL FP8 MQA logits 内核已包含在内。

作者通过 release notes 链条澄清 3913 已包含在 post4/post5 中，并最终确认 v0.1.19 包含该内核，解除阻塞。

> akii96：在 DSv4 Flash 上重新验证了性能收益，该 PR 对 DSv4 Pro 也有益。

作者补充了 DeepSeek-V4-Flash TP8 的端到端基准（Triton vs FlyDSL），确认性能收益可复现，随后 tjtanaa 批准合并并触发 CI。

- AITER 版本是否包含 FlyDSL 内核 (question): 确认 AITER v0.1.19 包含所需内核，解除版本阻塞，PR 继续推进。
- DSv4 Flash 端到端性能复测 (performance): 性能收益在第二个模型上复现，维护者批准合并并触发 CI。

风险与影响

- 风险：
  1. AITER 版本兼容性：from aiter.ops.flydsl import flydsl_fp8_mqa_logits 在 AITER 低于 v0.1.19 时会导致 ImportError，gfx942 + AITER 用户启动即失败。当前 Dockerfile.rocm_base 已固定 v0.1.19，但使用旧镜像或自定义环境的用户可能受影响。
  2. 无单元测试覆盖：变更仅 1 个文件、11 行改动，没有针对内核选择分支的单元测试，回归检测完全依赖 CI 中的 ROCm 任务。
  3. fp8 精度依赖：内核计算结果依赖 AITER 的 fp8 实现，GSM8K 与 NIAH 测试通过，但 NIAH 128K 为 8/10，长上下文检索场景下存在一定的精度波动空间。
  4. 性能收益的普适性：基准数据集中在 H=64、D=128 及特定 KV 长度形状，其他模型或配置下收益可能不同；短上下文（8K）基本无提升，但已确认无回归。 - 影响：用户侧：ROCm gfx942（MI300X/MI325X）上运行 DeepSeek-V4 Flash/Pro 等稀疏注意力模型的用户，长上下文 prefill 延迟显著降低（GLM-5.2-FP8 实测 TTFT -49%、TPOT -37%），短上下文体验基本不变。系统侧：依赖 AITER v0.1.19+，与当前官方 ROCm 镜像版本一致，无额外部署成本；非 gfx942 平台完全不受影响。团队侧：为后续 ROCm 平台 Triton 内核到 FlyDSL 的迁移提供了最小改动的样板，验证了跨仓内核替换的流程（版本确认、双模型基准、精度测试、CI 复跑）。 - 风险标记：AITER 版本依赖，无单元测试覆盖，fp8 精度依赖

关联脉络

- PR #52356 [Bugfix][ROCm] Skip FP8 MLA prefill PS-metadata build for chunked-context batches: 同属 ROCm + FP8 MLA 注意力链路（rocm_aiter_mla.py），

与本 PR 的 fp8_mqa_logits 索引器内核共用同一执行路径。

- PR #51538 [Bugfix] Make DSV4 sparse MLA work end-to-end for plain decode, MTP, and DSpark: 修复 DSV4 稀疏 MLA 的端到端链路，包含 MLA indexer 相关改动，与本 PR 优化的 gfx942 索引器内核属于同一功能线。
- PR #52084 [Perf][DSV4] Optimize sparse top-k metadata kernels for higher prefill throughput: 同为 DSV4 稀疏注意力索引相关内核的性能优化，方向一致，可互相印证 ROCm 上稀疏 MLA 的优化脉络。
- PR #46730 [Bugfix] DSV4 Flash: make Q fnuz in the indexer: Issue 评论中明确提及的上游 dtype 修复（使 Q 与 fnuz KV cache 同 dtype），是 FlyDSL 单 dtype 路径真正生效的前置条件，体现跨仓库依赖关系。