PR #42062 完整报告

vllm-project/vllm

[ROCm] Enable gluon paged MQA logits on gfx950 (MI355X)

合并时间：2026-05-14 23:39

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42062

执行摘要

- 一句话：修复 MI355X (gfx950) 未使用 gluon 单核路径
- 推荐动作：值得快速合并。此 PR 修复了一个明显的性能回归问题，改动极小且逻辑清晰。对于关注 ROCm 性能和 MI355X 部署的团队值得精读，了解 GPU 架构分发条件的管理方式。

功能与动机

PR body 明确指出：deepgemm_fp8_paged_mqa_logits 在 gfx942 上使用了 gluon 单核路径，而 gfx950 回退到了两核慢路径（stage1 + sum(dim=0)）。AITER 内部已支持 gfx950 的 gluon 分发，但 vLLM 的 guard 未正确反映这一事实，导致 MI355X 性能受限。修复后 DeepSeek-V3.2 在 MI355X 上能正确调度 gluon kernel。

实现拆解

1. 修改导入语句：在 vllm/v1/attention/ops/rocm_aiter_mla_sparse.py 中，将原本只导入 _ON_GFX942 改为同时导入 _ON_GFX950，并在非 ROCm 平台添加 _ON_GFX950 = False 的 fallback。
2. 扩展条件判断：将 if _ON_GFX942: 改为 if _ON_GFX942 or _ON_GFX950:，使得 gfx950（MI355X）也能进入 gluon 单核路径。
3. 无需额外测试或配置：该变更仅涉及一行导入和一行条件判断，无其他配套文件改动。

关键文件：

- vllm/v1/attention/ops/rocm_aiter_mla_sparse.py（模块 注意力；类别 infra；类型 infrastructure；符号 rocm_fp8_paged_mqa_logits）：唯一修改的文件，包含导入语句和条件判断的修正，是 ROCm MLA 稀疏注意力运算的核心实现文件。

关键符号：rocm_fp8_paged_mqa_logits

关键源码片段

vllm/v1/attention/ops/rocm_aiter_mla_sparse.py

唯一修改的文件，包含导入语句和条件判断的修正，是 ROCm MLA 稀疏注意力运算的核心实现文件。

```
# vllm/v1/attention/ops/rocm_aiter_mla_sparse.py
```

```
# 导入扩展：新增 _ON_GFX950，使 gfx950 也能进入 gluon 路径
if current_platform.is_rocm():
    from vllm.platforms.rocm import _ON_GFX942, _ON_GFX950 # 原仅有 _ON_GFX942
else:
    _ON_GFX942 = False
    _ON_GFX950 = False # 新增 fallback，确保非 ROCm 平台不会报错

# ... 其他代码 ...

def rocm_fp8_paged_mqa_logits(
    # ... 参数省略 ...
):
    # ... 前置逻辑 ...
    aiter_paged_mqa_logits_module = paged_mqa_logits_module()
    if aiter_paged_mqa_logits_module is not None:
        # 关键修复：从 if _ON_GFX942 改为同时覆盖 gfx942 和 gfx950
        if _ON_GFX942 or _ON_GFX950:
            deepgemm_fp8_paged_mqa_logits = (
                aiter_paged_mqa_logits_module.deepgemm_fp8_paged_mqa_logits
            )
            # ... 后续 gluon 单核 kernel 调用 ...
        else:
            # 其他架构仍然使用两核慢路径（stage1 + sum(dim=0)）
            pass
```

评论区精华

gemini-code-assist[bot] 建议使用统一标志 _ON_MI3XX 替代分别判断不同架构，以提高可维护性。

> reviewer gemini-code-assist[bot]: "For better maintainability ... consider using the _ON_MI3XX flag ..."

maeehart 回应否决了该建议，指出需要为特定架构保留精细区分，因为 Mi308（可能指 MI308 或类似型号）可能存在例外情况，且代码中其他位置也需要分别引用这两个标志。

> reviewer maeehart: "In this case we have to be more specific as Mi308 might have some exceptions." reviewer maeehart: "We need both _on_gfx942 and _on_gfx950 elsewhere, so it does not make sense to apply above recommendation."

最终 PR 保持原有模式，仅扩展条件。

- 是否应使用统一 _ON_MI3XX 标志 (design): 保留 _ON_GFX942 和 _ON_GFX950 分别判断，不引入统一标志。

风险与影响

- 风险：风险极低。变更仅涉及两行代码：一行导入，一行条件扩展。未修改任何 kernel 逻辑或数据路径。gfx950 上的 gluon 路径在 AITER 内部已有支持，此 PR 只是修正了 vLLM 侧的分发条件。

- 影响：仅影响使用 ROCm 后端且 GPU 为 MI355X (gfx950) 并执行 MLA 稀疏注意力运算的场景。对于 gfx942 和其他架构无影响。用户将观察到 MI355X 上 DeepSeek 类模型推理性能提升（从双核路径变为单核 gluon 路径）。
- 风险标记：暂无

关联脉络

- PR #40871 Introduce gluon paged MQA logits path for gfx942: 此 PR 引入了 _ON_GFX942 guard，拆分 gluon 路径，本 PR 在此基础上扩展至 gfx950。