# PR #47209 完整报告

vllm-project/vllm

[ROCm][Bugfix] Fix Triton "out of resource: shared memory" Error In One-Shot LoRA MoE

合并时间：2026-07-01 14:24

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47209

## 执行摘要

- 一句话：修复 ROCm LoRA MoE Triton 核显存不足错误
- 推荐动作：建议合并。该 PR 修复了一个具体的运行时错误，改动简洁且经过平台隔离，逻辑正确。

## 功能与动机

PR body 指出，在 MI300 和 MI325 上运行 LoRA 测试 test_gpt_oss_lora 时，Triton 核因共享内存不足 (Required: 69632 bytes, Hardware limit: 65536 bytes) 而失败。错误提示需要减小 block 大小或 num_stages。由于共享内存限制是平台特定的，需要根据平台能力动态调整。

## 实现拆解

1. 新增导入：在 vllm/lora/ops/triton_ops/fused_moe_lora_op.py 中导入 get_max_shared_memory_bytes 函数，用于获取设备最大共享内存字节数。
2. 添加 fallback 逻辑：在 _run_fused_moe_lora_one_shot 函数中，当 num_stages = 3 且平台是 CUDA 类平台时，查询设备共享内存上限。若小于 68KB，则将 num_stages 降低至 2，从而减少共享内存需求。
3. 平台保护：使用 current_platform.is_cuda_alike() 检查，确保该 fallback 仅应用于 CUDA 类平台，避免在非 CUDA 平台上调用 get_max_shared_memory_bytes 可能引发错误。

关键文件：

- vllm/lora/ops/triton_ops/fused_moe_lora_op.py（模块 LoRA 算子；类别 source；类型 bugfix；符号 _run_fused_moe_lora_one_shot）：核心修复文件：新增共享内存检测和降级逻辑，解决 LoRA MoE 在 AMD GPU 上的共享内存溢出。

关键符号：_run_fused_moe_lora_one_shot

### 关键源码片段

#### vllm/lora/ops/triton_ops/fused_moe_lora_op.py

核心修复文件：新增共享内存检测和降级逻辑，解决 LoRA MoE 在 AMD GPU 上的共享内存溢出。

```
# vllm/lora/ops/triton_ops/fused_moe_lora_op.py

from vllm.platforms import current_platform
from vllm.utils.mem_utils import get_max_shared_memory_bytes # 新增导入

def _run_fused_moe_lora_one_shot(...):
    # ... 前面的逻辑确定 block_n, nw, ns
    if hidden_size >= 4096:
        block_n, nw, ns = 128, 8, 3
    else:
        block_n, nw, ns = 128, 4, 3

    # 修复共享内存溢出：若设备最大共享内存小于 68KB，则降级为 2 级流水线
    if current_platform.is_cuda_alike():
        max_shmem_bytes = 68 * 1024
        if get_max_shared_memory_bytes(device.index) < max_shmem_bytes:
            ns = min(ns, 2) # 将 num_stages 限制为 2
```

## 评论区精华

仅有一条来自 *jeejeelee* 的批准评论 "LGTM, thank you"，无实质性讨论。

- PR 审核 (other): PR 获得批准，无需进一步修改。

## 风险与影响

- 风险：风险极低。变更仅 9 行，加在一个已有 fallback 分支内，且仅在共享内存不足时才触发，不影响正常路径。对 non-CUDA 平台完全隔离。可能的风险是 get_max_shared_memory_bytes 在某些罕见平台上表现异常，但 is_cuda_alike() 已提供保护。
- 影响：影响范围小。仅修复了 AMD GPU 上特定 Triton 核的共享内存溢出问题，使 LoRA MoE 功能在受限 GPU 上可用。对其他平台无影响。
- 风险标记：平台特定修复，低风险小改动

## 关联脉络

- PR #47269 [ROCm][MiniMax-M3] Cross-layer lightning-indexer top-k sharing: 同为 ROCm 相关的性能与稳定性修复，涉及 GPU 资源优化。