# PR #46339 完整报告

vllm-project/vllm

[Bugfix] Re-enable FP8 MoE on NVIDIA Thor

合并时间：2026-06-24 22:35

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46339

## 执行摘要

- 一句话：修复 FP8 MoE 在 NVIDIA Thor 上的回归
- 推荐动作：该 PR 是必要的回归修复，变更简洁、风险低。建议阅读以了解 FP8 MoE kernel 的架构兼容性管理方式，特别是设备能力检查与 CMake 架构列表的同步机制。

## 功能与动机

PR #45277 中的更改导致 Qwen/Qwen3.5-35B-A3B-FP8 在 NVIDIA Thor (SM101 for CUDA 12, SM110 for CUDA 13) 上推理失败。需要部分回退以恢复 FP8 MoE 功能，并保持与现有架构的兼容。

## 实现拆解

1. 放宽设备能力检查：在 csrc/libtorch_stable/quantization/w8a8/cutlass/scaled_mm_entry.cu 中，将 cutlass_group_gemm_supported 函数内的 CUDA 设备能力上限从 110 修改为 120，使 SM101 和 SM110 设备能够通过检查并启用 FP8 MoE。
2. 更新 CMake 架构列表：在 CMakeLists.txt 中，针对 CUDA 13.0+ 编译器，将 SCALED_MM_ARCHS 添加 11.0f；针对更低版本编译器，添加 10.1a，确保 Thor 架构被正确编译。
3. 无测试配套修改：该 PR 未包含测试变更，仅做最小修复以快速合并和 cherry-pick。

关键文件：

- csrc/libtorch_stable/quantization/w8a8/cutlass/scaled_mm_entry.cu（模块 CUDA 内核；类别 source；类型 core-logic；符号 cutlass_group_gemm_supported）：核心修复文件，修改了控制 FP8 MoE 启用的设备能力检查条件，是 bug 的直接原因。
- CMakeLists.txt（模块 构建系统；类别 infra；类型 configuration）：更新了 SM 架构编译列表，确保 Thor 相关 CUDA 架构被正确包含。

关键符号：cutlass_group_gemm_supported

## 关键源码片段

### csrc/libtorch_stable/quantization/w8a8/cutlass/scaled_mm_entry.cu

核心修复文件，修改了控制 FP8 MoE 启用的设备能力检查条件，是 bug 的直接原因。

```
// 文件：csrc/libtorch_stable/quantization/w8a8/cutlass/scaled_mm_entry.cu
```

```
// 函数：cutlass_group_gemm_supported
// 判断是否支持 CUTLASS Grouped GEMM（用于 MoE）
bool cutlass_group_gemm_supported(int64_t cuda_device_capability) {
#if defined CUDA_VERSION
#if defined ENABLE_CUTLASS_MOE_SM100 && ENABLE_CUTLASS_MOE_SM100
  // 原本的上限是 110，排除 SM110 (Thor)
  // 改为 < 120，使 SM101 和 SM110 都能通过检查
  if (cuda_device_capability >= 100 && cuda_device_capability < 120) {
    return CUDA_VERSION >= 12080; // CUDA 12.8+ 才支持
  }
#endif
#endif
  return false;
}
```

### CMakeLists.txt

更新了 SM 架构编译列表，确保 Thor 相关 CUDA 架构被正确包含。

```
# 文件：CMakeLists.txt
# 构建配置：根据 CUDA 版本选择 SM 架构
if(${CMAKE_CUDA_COMPILER_VERSION} VERSION_GREATER_EQUAL 13.0)
  # CUDA 13: Thor 架构为 11.0f
  cuda_archs_loose_intersection(SCALED_MM_ARCHS "10.0f;11.0f" "${CUDA_ARCHS}")
else()
  # CUDA 12: Thor 架构为 10.1a
  cuda_archs_loose_intersection(SCALED_MM_ARCHS "10.0a;10.1a;10.3a" "${CUDA_ARCHS}")
endif()
```

## 评论区精华

审核中主要讨论点是架构命名的清晰性。Harry-Chen 建议添加如 ENABLE_CUTLASS_MOE_SM110 的宏，或重命名现有宏以明确支持 SM100 和 SM110。DarkLight1337 认为保持现状更简洁，避免代码重复，最终达成共识暂不重命名。讨论在友好的氛围中结束，审核者 Isotr0py 已批准合并。

- 架构命名和宏清晰性 (design)：暂不重命名，维持现状。

## 风险与影响

- 风险：风险较低。修改范围极小（2 个文件，各 2-3 行），且 SM100/110 设备能力上限放宽到 120 不会影响现有 SM100 设备的兼容性。但未对 SM110 进行独立性测试，可能存在未知的 kernel 兼容性问题。此外，CPUID 上限 120 可能覆盖未来架构，需谨慎。
- 影响：直接影响是恢复了 NVIDIA Thor（SM101/SM110）上 FP8 MoE 模型（如 Qwen3.5-35B-A3B-FP8）的推理能力。影响范围限于使用这些特定 GPU 和 FP8 MoE 的用户。由于是回归修复，对现有其他用户无影响。
- 风险标记：核心路径变更，缺少测试覆盖

## 关联脉络

- PR #45277 [Kernel][FP8] Add w8a8 cutlass 3x FP8 GEMM support for SM100 (Blackwell): 该 PR 的变更导致了当前 bug，当前 PR 部分回退了其影响。