PR #40808 完整报告

vllm-project/vllm

[Bugfix] Disable FlashInfer CUTLASS MoE on SM110 (Jetson Thor AGX)

合并时间：2026-05-01 11:08

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40808

执行摘要

- 一句话：禁用 SM110 的 FlashInfer CUTLASS MoE 回退到 Triton
- 推荐动作：这是一个简单而正确的临时修复，值得精读。关注点：_supports_current_device() 的设计模式——通过白名单控制硬件特性选择；以及关联 PR #36286 引入的 oracle 流程如何自动暴露此前隐藏的兼容性问题。团队应跟踪上游 FlashInfer 是否发布 SM110 cubin 以移除此限制。

功能与动机

FlashInfer <= 0.6.8.post1 未提供 SM110 的 MoE cubin 文件，导致 Jetson Thor AGX 上运行未量化 BF16 MoE 模型（如 Nemotron）时，因为运行时复用 SM100 cubin 并选择了不支持 tile 256x256x128 而崩溃。PR body 明确说明："On SM110 the SM100 artifact is reused at runtime and the TMA-WS dispatcher picks tile 256x256x128 which has no SM100 registration, raising Unsupported tile shape config 256256128 for MoE gemm"。关联 Issue flashinfer-ai/flashinfer#3134 跟踪上游修复。

实现拆解

步骤 1：定位问题 分析发现 FlashInferExperts._supports_current_device() 方法中，SM110 被包含在设备能力家族 110 的检查中（p.is_device_capability_family(110)），该方法决定了 FlashInfer CUTLASS MoE 是否可用于当前 GPU。

步骤 2：排除 SM110 在文件 vllm/model_executor/layers/fused_moe/flashinfer_cutlass_moe.py 的第 132 行删除 or p.is_device_capability_family(110)，并添加注释 # SM110 excluded: flashinfer-ai/flashinfer#3134，将 SM120 直接保留。

步骤 3：效果 修改后，对于 SM110 设备，_supports_current_device() 返回 False，MoE oracle 不会选择 FlashInfer CUTLASS Experts，而是回退到 Triton 实现的 MoE kernel，从而避免崩溃。

配套说明 本次变更为一行代码，没有新增测试，但已在 Thor 上手动验证。该修复是临时性的，待 FlashInfer 上游提供 SM110 MoE cubin 后应移除该限制（见 PR body 中 "Remove once SM110 MoE cubins ship"）。

关键文件：

- vllm/model_executor/layers/fused_moe/flashinfer_cutlass_moe.py（模块 MoE 调度；类别 source；类型 core-logic）：核心变更文件，修改 _supports_current_device() 方法排除 SM110。

关键符号：未识别

关键源码片段

vllm/model_executor/layers/fused_moe/flashinfer_cutlass_moe.py

核心变更文件，修改 _supports_current_device() 方法排除 SM110。

```
@staticmethod
def _supports_current_device() -> bool:
    p = current_platform
    return (
        p.is_cuda()
        and (
            p.is_device_capability(90) # Hopper
            or p.is_device_capability_family(100) # Blackwell
            # SM110 excluded: flashinfer-ai/flashinfer#3134
            # FlashInfer <= 0.6.8.post1 lacks SM110 MoE cubins.
            # Reusing SM100 cubin leads to unsupported tile shape.
            or p.is_device_capability_family(120) # Grace Hopper / future
        )
        and has_flashinfer_cutlass_fused_moe()
    )
```

评论区精华

审阅者 NickLucche 直接批准（"Looks reasonable, thanks for the work @stecasta !"）。Gemini Code Assist 确认这是一个针对缺失 cubin 的规避措施。stecasta 在评论中指出 H100 测试失败与此 PR 无关。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅从白名单排除一个 GPU 架构，使其回退到已充分验证的 Triton MoE 实现。不会影响其他架构（如 SM90、SM100、SM120）。如果 SM110 用户本应使用 FlashInfer CUTLASS MoE 获得更好性能，此次回退可能带来性能下降，但这是必要的，因为原方案无法工作。上游修复后可安全恢复。
- 影响：影响范围：仅 Jetson Thor AGX（SM110）设备，且仅影响使用了未量化 BF16 MoE 模型的场景（如 Nemotron-3-Nano）。影响程度：从服务器崩溃变为正常服务（使用 Triton 回退），属于积极的 bugfix。对于其他设备无影响。
- 风险标记：仅临时修复，上游依赖

关联脉络

- PR #39825 [Bugfix] Disable FlashInfer CUTLASS MoE on SM121 (Jetson Thor AGX): 类似问题：为 SM121 做了相同的禁用，本 PR 是其延续。
- PR #36286 [MoE] Migrate unquantized MoE to full oracle flow: 该 PR 暴露了 SM110 的兼容性问题，使 FlashInfer CUTLASS 成为未量化 MoE 的默认路径。