PR #41215 完整报告

vllm-project/vllm

[Bugfix] Use enable_sm120_family for per-tensor FP8 CUTLASS kernels on SM12.1

合并时间：2026-05-20 22:10

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41215

执行摘要

修复 NVIDIA SM12.1 设备（如 GB10）上 per-tensor FP8 CUTLASS 内核的崩溃。通过将架构守卫从 enable_sm120_only 替换为 enable_sm120_family，使内核在 SM12.x 全系列上可用。

功能与动机

Issue #40758 报告：Qwen3.6-35B-A3B-FP8 在 NVIDIA GB10（CUDA 13.0）上启动时崩溃，错误 cutlass_gemm_caller Error Internal。原因是 FP8 CUTLASS 内核使用 enable_sm120_only 守卫，仅匹配 __CUDA_ARCH__== 1200，而 SM12.1 的 __CUDA_ARCH__== 1210 被排除，导致内核陷阱。

实现拆解

1. 替换架构守卫：在 scaled_mm.cuh 和 scaled_mm_sm120_fp8_dispatch.cuh 中，将 enable_sm120_only 替换为 enable_sm120_family，后者检查 >= 1200 && < 1300，覆盖 SM12.0 和 SM12.1。
2. 添加安全陷阱：在 enable_sm120_family 实现中增加 kernel_trap，确保在不支持的架构上触发明确错误。
3. 拼写修正：修正错误消息中的拼写错误。

csrc/libtorch_stable/quantization/w8a8/cutlass/c3x/scaled_mm.cuh

核心变更文件，替换架构守卫以支持 SM12.1

```
// 文件：csrc/libtorch_stable/quantization/w8a8/cutlass/c3x/scaled_mm.cuh
// 将架构守卫从 enable_sm120_only 改为 enable_sm120_family
// enable_sm120_only 仅匹配 __CUDA_ARCH__ == 1200 (SM12.0)
// enable_sm120_family 匹配 __CUDA_ARCH__ >= 1200 && < 1300 (SM12.x)
// 这使得内核在 SM12.1 ( 如 NVIDIA GB10) 上也能使用
using GemmKernel = enable_sm120_family<cutlass::gemm::kernel::GemmUniversal<
    ...>>;
```

csrc/cutlass_extensions/common.hpp

enable_sm120_family 定义文件，添加 kernel_trap 增强错误安全性

```
// 文件：csrc/cutlass_extensions/common.hpp
// enable_sm120_family 定义：匹配 SM12.x 架构
// 添加 kernel_trap 确保在不支持的架构上触发错误
```

```
struct enable_sm120_family {
    template <typename T>
    using type = std::conditional_t<
        (__CUDA_ARCH__ >= 1200 && __CUDA_ARCH__ < 1300),
        T,
        T>; // 实际上通过 kernel_trap 终止
};
```

评论区精华

- gemini-code-assist[bot]指出 enable_sm120_family 缺少针对非 SM12x 硬件的安全陷阱，可能导致静默失败。作者随后添加了 kernel_trap 解决此问题。

风险与影响

- 风险：低。变更仅替换架构守卫，现有测试已覆盖 SM12.x 路径。添加的 kernel_trap 增强了错误安全性。
- 影响：仅限于 SM12.1 设备，修复 FP8 内核崩溃，不影响其他架构或功能。

关联脉络

本 PR 直接修复 Issue #40758 报告的 CI 失败。