PR #39712 完整报告

vllm-project/vllm

[CI/Build] Enable FP8 on NVIDIA Thor

合并时间：2026-04-30 00:48

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39712

执行摘要

此 PR 在 CUTLASS 内核架构选择器中新增 enable_sm100_to_sm120，使 FP8 内核兼容 NVIDIA Thor (SM110)，同时删除死代码并优化错误消息。变更集中在 4 个 C++ 文件中，共 +13/-38 行，不影响推理逻辑，但需注意 Python 端 capability 检查尚未同步。

功能与动机

用户报告在 NVIDIA Thor 上加载 FP8 模型时因内核仅支持 SM100f/100a，启动阶段崩溃。此 PR 旨在通过放宽架构范围检查，让 FP8 推理在 Thor 上正常运行。

实现拆解

1. 重构架构选择器：在 csrc/cutlass_extensions/common.hpp 中新增 enable_sm100_to_sm120，覆盖 SM100-SM119，删除 enable_sm90_only 和 enable_sm100a_only 死代码，并更新 enable_sm120_only 和 enable_sm120_family 的错误提示。
2. 更新 GEMM 内核包装器：在 scaled_mm.cuh、scaled_mm_blockwise_sm100_fp8_dispatch.cuh、scaled_mm_sm100_fp8_dispatch.cuh 三处将 enable_sm100f_only 替换为 enable_sm100_to_sm120。
3. 错误信息清理：统一错误字符串命名，如 "sm120a"、"sm120f"、"sm[100, 120)"。

关键源码片段

csrc/cutlass_extensions/common.hpp

核心变更文件：新增 enable_sm100_to_sm120 架构选择器，删除 enable_sm90_only、enable_sm100a_only 死代码，更新错误消息。这是整个 PR 的基石。

```
template <typename Kernel>
struct enable_sm100_to_sm120 : Kernel {
  template <typename... Args>
  CUTLASS_DEVICE void operator()(Args&&... args) {
#if defined __CUDA_ARCH__
  #if (__CUDA_ARCH__ >= 1000 && __CUDA_ARCH__ < 1200)
    Kernel::operator()(std::forward<Args>(args)...);
  #else
    // Unsupported architecture: trigger trap
```

```
    asm("trap;");
  #endif
 #endif
  }
};
```

评论区精华

- gemini-code-assist[bot]指出 Python 端 cutlass_mla.py 的 capability 检查仍需更新，否则 MLA 后端在 Thor 上不会被选择。
- Isotr0py提醒架构选择器更新后，对应的错误消息也应同步修改，作者 DarkLight1337 随后调整了字符串并采用了范围命名。
- DarkLight1337讨论了用范围检查还是逐个架构添加，最终选择范围检查以降低维护成本，并清理了死代码。

风险与影响

- 风险：Python 端 MLA capability 检查未更新；范围检查可能覆盖未经充分测试的 SM 中间版本；缺少 CI 回归测试。
- 影响：解锁 Thor 上的 FP8 推理，代码量精简，无性能回退。

关联脉络

本次变更是对 NVIDIA Thor 硬件支持的独立改进，与近期其他 CI 或 kernel PR 无直接依赖。后续可追踪 Python 侧 capability 更新的 PR 以形成完整支持。