

PR #32842 完整报告

sgl-project/sglang

[Kernel] Remove unreachable AOT headers

合并时间: 2026-07-30 22:08

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32842>

执行摘要: 此 PR 清理了 7 个在 AOT 编译中不可达的 CUDA 头文件, 净删 2852 行, 降低代码库维护负担。通过交叉检查编译单元和 manifest 确保安全, 并验证了 kernel inventory 测试和 H200 AOT 编译无回归。

功能与动机: PR body 明确指出要删除在 AOT 编译单元中无法到达的头文件。这些头文件已成孤立代码, 维护它们不仅浪费资源, 还可能误导开发者认为这些代码仍在活跃使用。删除后可以缩小内核代码搜索空间, 加快编译。

实现拆解: 作者通过以下步骤完成删除:

1. 交叉检查: 针对所有 AOT 构建根目录 (如 common_ops_sm90_build、common_ops_sm100_build) 和 manifest, 逐一验证每个头文件的包含关系, 确认 7 个头文件没有被任何编译单元引用。
2. 删除头文件: 从 python/sglang/kernels/aot/csrc/ 中删除以下文件:
 - gemm/marlin/marlin_template.h (主干 Marlin kernel 模板, 1629 行)
 - gemm/marlin/dequant.h (反量化逻辑, 504 行)
 - elementwise/pos_enc.cuh (位置编码 kernel, 467 行)
 - gemm/marlin/marlin.cuh (Marlin 辅助, 96 行)
 - gemm/marlin/marlin_dtypes.cuh (Marlin 数据类型, 82 行)
 - cutlass_extensions/gemm/dispatch_policy.hpp (FP8 调度策略, 38 行)
 - gemm/marlin/kernel.h (Marlin kernel 声明, 36 行)
3. 保留 JIT 副本: 确保 python/sglang/kernels/jit/ 下的对应文件不受影响, JIT 编译功能继续工作。
4. 验证: 运行 test/registered/kernels/test_kernel_inventory.py 全部通过 (5 passed)。在 H200 GPU 上构建 SM90 和 SM100 的 AOT 共享库并加载测试, 通过 rotary 和 GPTQ-Marlin 的 JIT smoke 测试。

关键源码片段: 作为示例, 下面展示被删除的 `dispatch_policy.hpp` 的完整内容。该文件定义了 FP8 Block Scaling 专属的主循环调度策略, 但由于 AOT 构建路径未包含此头, 其功能已通过 JIT 路径实现。

```
// Adapted from https://github.com/vllm-project/vllm/blob/main/csrc/cutlass_extensions/gemm/dispatch_policy.hpp
#include "cutlass/gemm/dispatch_policy.hpp"

namespace cutlass::gemm {
```

```

// FP8 related policies (including Blocked Scaled Accumulation)
// `ScaleGranularityM` specifies scaling granularity along M
template <int ScaleGranularityM = 0>
struct KernelTmaWarpSpecializedCooperativeFP8BlockScaledSubGroupMAccum
    : KernelTmaWarpSpecializedCooperative {};

// n-buffer in smem (Hopper TMA), pipelined with Hopper GMMA and TMA, Warp
// specialized dynamic schedule for FP8 kernels with Block Scaling
template <int Stages_, class ClusterShape_ = Shape<_1, _1, _1>,
        class KernelSchedule = KernelTmaWarpSpecialized,
        int ScaleGranularityM = 0>
struct MainloopSm90TmaGmmaWarpSpecializedBlockScalingSubGroupMFP8
    : MainloopSm90TmaGmmaWarpSpecialized<Stages_, ClusterShape_,
        KernelSchedule> {

    static_assert(
        cute::is_same_v<KernelSchedule,
            KernelTmaWarpSpecializedCooperativeFP8BlockScaledSubGroupMAccum<
                ScaleGranularityM>>,
        "KernelSchedule must be one of the warp specialized policies");
};

} // namespace cutlass::gemm

```

该文件被删除是因为它在 AOT manifest 中不可达。其中定义的 `ClusterShape_` 和 `KernelSchedule` 符号在当前代码库中不再有 AOT 引用。

评论区精华：无实质性的审查评论。

风险与影响：

- 风险：极低。删除的头文件均不可达，且已通过测试验证。
- 影响：对用户零影响。对开发者：减少了代码库体积，但需要确保不会重新引入这些头文件的依赖。当前 JIT 版本完整保留。

关联脉络：此 PR 是内核代码库持续清理的一部分。虽然没有直接关联的 PR，但与近期内核重构（如 PR#32813、#32886、#32887）共同提升了内核模块的维护性。