PR #27868 完整报告

sgl-project/sglang

fix(qwen3.5): keep CUDA dual-stream overlap (regressed by #25885)

合并时间：2026-06-15 21:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27868

执行摘要

- 一句话：修复 #25885 引入的 CUDA dual-stream overlap 性能回退
- 推荐动作：值得精读，尤其适合需要处理跨平台条件分支的开发者。PR 简洁地演示了如何修复因平台特定门控引入的回归，以及 Review 中建议的代码简化。

功能与动机

25885 为 AMD 平台增加 alt_stream 支持，但门控条件使用 _is_hip 且未保留 _is_cuda，导致原先对 CUDA 启用的 dual-stream overlap 被静默禁用。git bisect 定位 #25885 为性能回退首次坏提交，H200 上 Qwen3.5 decode 吞吐从 434 tok/s 降至 377 tok/s（-15%）。

实现拆解

该 PR 在 python/sglang/srt/models/qwen3_5.py 中做了两处核心修改：

1. 将模块级别的全局标志 _gdn_use_alt_stream 和 _qknorm_use_alt_stream 的定义从原来仅依赖环境变量与 _is_hip，改为 _is_cuda or (环境变量 and _hip_use_alt_stream)。这样在 CUDA 上这两个标志始终为 True，恢复原有行为；AMD 上仍通过环境变量控制。
2. 在 Qwen3_5AttentionDecoderLayer.__init__ 和 Qwen3_5DecoderLayer.__init__ 中，传递给 Qwen2MoeSparseMoeBlock 的 alt_stream 参数判断，从 alt_stream if _disable_shared_experts_fusion() else None 改为 alt_stream if (_is_cuda or _disable_shared_experts_fusion()) else None，保证 CUDA 上始终传入 alt_stream 参数，即使 shared experts fusion 未禁用。

关键文件：

- python/sglang/srt/models/qwen3_5.py（模块 模型层；类别 source；类型 core-logic；符号 _gdn_use_alt_stream, _qknorm_use_alt_stream, Qwen3_5AttentionDecoderLayer.init, Qwen3_5DecoderLayer.init）：唯一的修改文件，包含了全部三处逻辑变更：变量定义、MLP 参数传递、方法内条件（变量定义折叠后自动修复）。

关键符号：Qwen3_5GatedDeltaNet._forward_input_proj, Qwen3_5AttentionDecoderLayer._apply_qk_norm, Qwen3_5AttentionDecoderLayer.init, Qwen3_5DecoderLayer.init

关键源码片段

python/sglang/srt/models/qwen3_5.py

唯一的修改文件，包含了全部三处逻辑变更：变量定义、MLP 参数传递、方法内条件（变量定义折叠后自动修复）。

```
# 在模块顶部：CUDA 上始终启用 alt_stream，AMD 上需显式设置环境变量
_gdn_use_alt_stream = _is_cuda or (
    get_bool_env_var("SGLANG_GDN_QKVZ_BA_ALT_STREAM", "False") and _hip_use_alt_stream
)
_qknorm_use_alt_stream = _is_cuda or (
    get_bool_env_var("SGLANG_QK_NORM_ALT_STREAM", "False") and _hip_use_alt_stream
)

# 在 Qwen3_5AttentionDecoderLayer.__init__ 中传递 alt_stream 给共享专家模块
# 确保 CUDA 上始终使用 alt_stream，无论 shared experts fusion 是否禁用
self.mlp = Qwen2MoeSparseMoeBlock(
    ...
    alt_stream=(
        alt_stream
        if (_is_cuda or _disable_shared_experts_fusion())
        else None
    ),
    ...
)
```

评论区精华

Reviewer ispobock 建议将 _is_cuda 直接折叠到 flag 定义中，而不是在方法内分散使用 _is_cuda or，以提高代码简洁性。作者接受建议并更新了变量定义，同时移除了方法内显式的 _is_cuda or 检查（因为 flag 已包含该逻辑）。

- 将 _is_cuda 折叠到 flag 定义中简化代码 (design): 作者采纳建议，在最新提交中将 _is_cuda 折叠到 flag 定义中，并移除了方法内的显式检查。

风险与影响

- 风险：风险较低，因为变更仅恢复之前已存在的行为，未引入新逻辑。但需要注意： 1) 没有测试覆盖该修复，未来改动可能再次回退； 2) 如果 is_cuda() 函数判断逻辑发生变化，可能导致预期外行为； 3) 对 AMD 平台无影响（环境变量仍可控制）。
  - 影响：影响所有使用 Qwen3.5 模型的 CUDA 用户（如 H200、Blackwell、Ampere 等），恢复 decode 吞吐约 15%（434 vs 377 tok/s）。仅修改一个文件，影响范围明确，无 breaking change。

- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #25885 [AMD] Support alt stream for Qwen3.5 on AMD platform: 该 PR 是本 PR 修复的根源，错误地将 CUDA dual-stream overlap 限制为仅 AMD。