PR #34642 完整报告

sgl-project/sglang

Revert "[Kimi K3] Fuse MLA gate projection into QKV-A GEMM"

合并时间：2026-08-13 08:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34642

执行摘要

- 一句话：回滚 Kimi K3 MLA gate 融合优化，修复长序列回归
- 推荐动作：建议快速浏览。重点阅读 kimi_k3.py 中 _precompute_output_gate 与 _gated_o_proj_forward 的协作方式，以及原 PR #33623 的融合设计（权重合并、tile_m 特例、CUDA Graph 交互）。若团队计划重新实现融合，应先在 2048 token 全序列场景补充回归测试，并厘清与 alt-stream 重叠、breakable CUDA graph 的同步语义。

功能与动机

PR body 明确说明了回滚原因：Reverts sgl-project/sglang#33623 Breaks CI https://github.com/sgl-project/sglang/actions/runs/31650215286/job/94292793773#step:15:2133 and seems like a real regression given full 2048 tokens。原始 PR #33623 将 QKV-A 与 TP-local 输出 gate 融合为单次 GEMM 以降低投影开销，但融合路径与 CUDA Graph 捕获、alt-stream 重叠等机制的交互在完整序列下出现回归，因此需要撤回到未融合的稳健路径。

实现拆解

1. 模型层回滚（python/sglang/srt/models/kimi_k3.py）：删除 _merge_qkv_a_g_proj_weights 方法、_qkv_a_g_proj_weight / _qkv_a_g_proj_sizes 状态字段与 prepare_qkv_latent 重写，forward 中无条件恢复调用 _precompute_output_gate；同时移除 fused_a_gemm、UnquantizedLinearMethod 等已不再使用的 import。原因是融合路径在 2048 token 场景回归，需要回到 gate 独立 GEMM 的旧路径。
2. gate 消费逻辑微调：_gated_o_proj_forward 中的 wait_stream 条件由 precomputed is not None and precomputed[1] is not None 恢复为 precomputed is not None，因为删除融合路径后 _gate_precomputed 只由 _precompute_output_gate 产生，producer stream 必非空。
3. 内核特例回滚（python/sglang/kernels/jit/csrc/gemm/dsv3_fused_a_gemm.cuh）：删除 pick_tile_m 中为 K3 融合形状（hd_out=3648, hd_in=7168）返回 tile_m=32 的特例，恢复默认 tile 16，避免该特例影响其他模型形状。
4. 测试配套删除：整体删除 test/registered/unit/models/test_kimi_k3_mla_gate_fusion.py，包括 CPU 等价性测试 test_merged_projection_matches_separate_projections 与 SM90 CUDA Graph replay 测试 test_fused_a_cuda_graph_replay。

5. 回归验证：作者通过 issue 评论请求 rerun test/registered/models_e2e/test_kimi_k3_b300.py，8-gpu-b300 工作流通过。

关键文件：

- python/sglang/srt/models/kimi_k3.py（模块 模型层；类别 source；类型 core-logic；符号 _merge_qkv_a_g_proj_weights, prepare_qkv_latent, _precompute_output_gate, _gated_o_proj_forward）：回滚核心文件：删除 MLA gate 与 QKV-A 融合路径，恢复独立 gate GEMM 与 alt-stream 预计算，是本次变更的主战场。
- test/registered/unit/models/test_kimi_k3_mla_gate_fusion.py（模块 单元测试；类别 test；类型 test-coverage；符号 TestKimiK3MlaGateFusion, test_merged_projection_matches_separate_projections, test_fused_a_cuda_graph_replay）：随功能回滚整体删除的测试文件，包含融合逻辑的 CPU 等价性与 CUDA Graph replay 覆盖。
- python/sglang/kernels/jit/csrc/gemm/dsv3_fused_a_gemm.cuh（模块 GEMM 内核；类别 other；类型 core-logic；符号 pick_tile_m）：删除 K3 融合形状的 tile_m=32 特例，使 fused-A GEMM 的 tile 选择恢复通用默认值。

关键符号：_merge_qkv_a_g_proj_weights, prepare_qkv_latent, _precompute_output_gate, _gated_o_proj_forward, pick_tile_m

关键源码片段

python/sglang/srt/models/kimi_k3.py

回滚核心文件：删除 MLA gate 与 QKV-A 融合路径，恢复独立 gate GEMM 与 alt-stream 预计算，是本次变更的主战场。

```
# 回滚 #33623 后恢复的路径：gate 不再由融合的 QKV-A GEMM 产出，
# _gate_precomputed 唯一的生产者是 _precompute_output_gate，
# 因此 _gated_o_proj_forward 里的 wait_stream 判断可以还原为
# "precomputed is not None"。

# 下述 _gated_o_proj_forward 是 __init__ 中安装的实例级 wrap；
# 它拦截 o_proj 的 forward，在进入 o_proj GEMM 前乘上 sigmoid(gate)。

def _precompute_output_gate(self, hidden_states: torch.Tensor) -> None:
    """在 alt stream 上发起 gate GEMM，与 attention 核心重叠；
    不满足条件时回退到 o_proj wrap 中的 lazy path 计算。"""
    self._gate_precomputed = None
    if (
        self._gate_alt_stream is not None
        and get_is_capture_mode()
        # breakable CUDA graph 下跨 segment 的 wait 不可取，走 lazy path
        and not is_in_breakable_cuda_graph()
        and (0 < hidden_states.shape[0] <= self._gate_bs_limit)
    ):
        alt = self._gate_alt_stream
        alt.wait_stream(torch.cuda.current_stream())
```

```
            with torch.cuda.stream(alt):
                gate, _ = self.g_proj(hidden_states)
            # 记录 producer stream，供 _gated_o_proj_forward 中 wait_stream 对齐
            self._gate_precomputed = (gate, alt)

    def _gated_o_proj_forward(x, *args, **kwargs):
        gate_input = self._gate_hidden_states
        self._gate_hidden_states = None
        precomputed = self._gate_precomputed
        self._gate_precomputed = None
        if precomputed is not None:
            # 等待 alt stream 完成 gate 写入后再做 sigmoid 乘法，避免数据竞争
            torch.cuda.current_stream().wait_stream(precomputed[1])
        if gate_input is not None and not isinstance(x, tuple):
            gate = (
                precomputed[0]
                if precomputed is not None
                else self.g_proj(gate_input)[0]
            )
            from sglang.kernels.ops.kimi_k3 import mla_output_gate

            if mla_output_gate.covered(x, gate):
                # 单 kernel 完成 x * sigmoid(gate)，双舍入与未融合路径逐位一致
                x = mla_output_gate.kimi_k3_mla_output_gate(x, gate)
            else:
                x = x * torch.sigmoid(gate)
        return _orig_o_proj_forward(x, *args, **kwargs)
```

评论区精华

本 PR 没有 review 评论；唯一可见交互是作者在 issue 侧请求 /rerun-test test/registered/models_e2e/test_kimi_k3_b300.py，随后 GitHub Actions 报告 8-gpu-b300 测试通过。回滚决策本身由 CI 失败与 2048 token 回归观察驱动，未展开设计层面的争论。

- CI 回归验证（rerun B300 端到端测试） (testing): 8-gpu-b300 工作流通过，回滚后 K3 端到端测试恢复绿灯。

风险与影响

- 风险：
  1. 性能回退：回滚后 K3 的 MLA 需要恢复独立的 gate GEMM，原先 #33623 报告的 decode 吞吐提升（1.28%–4.47%）将丢失，这是预期代价。
  2. wait_stream 隐含假设：head 中 _gated_o_proj_forward 直接调用 wait_stream(precomputed[1])，若未来有代码路径设置 _gate_precomputed = (gate, None)，会触发空 stream 问题；当前仅 _precompute_output_gate 设置该字段，风险可控但属于脆弱契约。

3. 测试覆盖缺失：删除融合测试后，若后续重新启用融合，需要重新补充 CPU 等价性与 CUDA Graph replay 测试，否则回归难以被捕获。
4. CUDA Graph 交互：原回归可能来自融合路径与 breakable CUDA graph / alt stream 的同步问题，回滚后此类风险随之消失，但根因尚未定位。 - 影响：影响范围集中在 Kimi K3 模型的 MLA 输出 gate 路径（python/sglang/srt/models/kimi_k3.py）与相关单元测试；对仓库其他模型无影响。对用户而言，K3 推理正确性回归被修复，但性能回到融合前水平（约 1%–4.5% 的 decode 差距）；对团队而言，需要跟踪原 PR #33623 的回归根因，以便未来以更稳健的方式重新融合。 - 风险标记：核心模型路径回滚，性能回退预期，删除测试覆盖，CUDA Graph 交互敏感

关联脉络

- PR #33623 [Kimi K3] Fuse MLA gate projection into QKV-A GEMM: 本 PR 正是对其的整体回滚，动机源于它导致的 CI 失败与 2048 token 回归。
- PR #33521 Kimi K3 Fuse MLA gate projection into QKV-A GEMM（被自动关闭的前身）：33623 的前身，因 base 分支删除自动关闭，说明该融合方案此前已迭代多次。
- PR #33465 [Kimi-K3][NPU] Support Kimi-K3 on NPU: 同一模型文件 kimi_k3.py 的大规模改动，涉及 MLA 与 gate 路径，回滚需与其保持兼容。