PR #34542 完整报告

sgl-project/sglang

[MiniMax-M3] Overlap shared and routed experts

合并时间：2026-08-14 22:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34542

执行摘要

- 一句话：MiniMax-M3 共享与路由专家在 CUDA 图上双流重叠
- 推荐动作：值得精读。该 PR 展示了利用 CUDA 备用流在 MoE 内部重叠独立计算分支的通用手段，是 CUDA Graph 捕获模式下性能优化的典型范例。建议关注 get_is_capture_mode() 与 alt_stream 的组合使用方式，并为该路径补充单元测试或 CI 验证。

功能与动机

PR body 指出：MiniMax-M3 未融合的共享专家与路由专家分支在 CUDA Graph 执行期间串行运行，导致 7,680 token 的 FlashInfer TRT-LLM MXFP8 routed-MoE 调用在启用 PDL 时可能停顿。现有的 8,192 token PDL 限制未覆盖该形状，且未转发到 FP8 wrapper 路径。

实现拆解

本 PR 的最终版本只包含双流重叠改动，具体实现分为以下几步：

1. 引入备用流：从 sglang.srt.runtime_context 导入 get_stream，在 MiniMaxM3 的 __init__ 中通过 get_stream("alt") 获取备用 CUDA 流（仅 _is_cuda 时启用），并将该流传给 MiniMaxM3DecoderLayer 与 MiniMaxM3MoE，保存在 self.alt_stream 中。
2. 提取路由专家前向：把原来在 forward_normal 内联的 router logits 计算、topk 选择与 self.experts 调用封装为独立的 _forward_router_experts 方法，便于在备用流中整体调度。
3. 双流重叠执行：在 forward_normal 中，当满足 alt_stream is not None、shared_experts is not None 且处于 CUDA 图捕获模式（get_is_capture_mode()）时，主流先执行 _forward_shared_experts，备用流执行 _forward_router_experts，并通过 wait_stream 完成双向同步；否则走原串行路径。
4. PDL 限制调整（已回退）：原提交将 TRT-LLM MoE 的 PDL 上限从 8,192 降到 4,096 并应用到 FP8 wrapper，同时添加了单卡 MXFP8 CUDA Graph 复现器。由于 FlashInfer 0.6.15 已修复底层 2CTA 挂起（flashinfer-ai/flashinfer#3973），该变通方案被整体移除，现有 FP4 8,192 token 限制保持不变。

最终改动集中在一个文件，未新增测试或文档。

关键文件：

- python/sglang/srt/models/minimax_m3.py（模块 模型实现；类别 source；类型 core-logic；符号 _forward_router_experts, forward_normal, MiniMaxM3MoE.init, MiniMaxM3DecoderLayer.init）：核心实现文件，通过 alt_stream 在 CUDA 图捕获模式中重叠共享专家与路由专家，是性能提升的唯一来源。

关键符号：_forward_router_experts, forward_normal, MiniMaxM3MoE.init, MiniMaxM3DecoderLayer.init

关键源码片段

python/sglang/srt/models/minimax_m3.py

核心实现文件，通过 alt_stream 在 CUDA 图捕获模式中重叠共享专家与路由专家，是性能提升的唯一来源。

```
# 关键片段：forward_normal 中基于 alt_stream 的双流重叠
# 仅当存在备用流、共享专家可用且处于 CUDA 图捕获模式时启用，
# 避免影响普通 eager 推理路径的行为。
def forward_normal(
    self,
    hidden_states: torch.Tensor,
    should_allreduce_fusion: bool = False,
    use_reduce_scatter: bool = False,
) -> torch.Tensor:
    if hidden_states.shape[0] > 0:
        if (
            self.alt_stream is not None
            and self.shared_experts is not None
            and get_is_capture_mode()
        ):
            current_stream = torch.cuda.current_stream()
            # 保证当前流上的输入准备在备流开始前完成
            self.alt_stream.wait_stream(current_stream)
            # 主流执行共享专家（轻量）
            shared_output = self._forward_shared_experts(hidden_states)
            # 备用流执行路由专家（重计算，含 topk 和 experts）
            with torch.cuda.stream(self.alt_stream):
                final_hidden_states = self._forward_router_experts(hidden_states)
            # 主流等待备流完成，避免后续 all-reduce 或加法产生竞争
            current_stream.wait_stream(self.alt_stream)
        else:
            # 非 CUDA 图捕获或其他后端，保留原串行逻辑
            shared_output = self._forward_shared_experts(hidden_states)
            final_hidden_states = self._forward_router_experts(hidden_states)
    else:
        shared_output = None
        topk_output = self.topk.empty_topk_output(hidden_states.device)
        final_hidden_states = self.experts(hidden_states, topk_output)

    if shared_output is not None:
```

```
            final_hidden_states = final_hidden_states + shared_output
        if self.tp_size > 1 and not should_allreduce_fusion and not use_reduce_scatter:
            final_hidden_states = tensor_model_parallel_all_reduce(final_hidden_states)
        return final_hidden_states
```

评论区精华

唯一的实质性讨论来自作者在 issue 评论中的说明：作者在最新 main 上重新测试，发现 PDL 挂起不再复现，因为 FlashInfer 的 PR 3973 修复了底层 2CTA 挂起，于是移除了该 workaround 及其复现器。评审者 BBuf 直接批准，无其他评论。

- PDL 限制变通方案回退 (other): 保留现有 FP4 8192 token PDL 限制，删除 FP8 的 4096 token 改动，最终合并版本只含双流重叠。

风险与影响

- 风险：该改动引入多流同步逻辑，存在以下风险：
  - 在 CUDA 图捕获模式下，wait_stream 的时序若处理不当可能导致图捕获失败或数据竞争。当前实现中主流先 wait_stream(alt)，备流执行完后再 wait_stream(alt) 等待回主流，顺序正确，但未覆盖备流内部出错的情况。
  - 守卫条件依赖 self.shared_experts is not None，而该属性是否始终存在取决于模型结构；若未来共享专家融合逻辑变化，可能静默回退到串行路径。
  - 仅对 CUDA 生效，NPU/ROCM 等硬件不受影响，但缺少针对双流路径的单元测试，回归风险主要靠手工 benchmark 覆盖。
  - 影响：直接影响使用 MiniMax-M3 模型且启用 CUDA 图的用户，TPOT 降低约 10-13%，峰值交互性提升约 4-5%，LongBench v2 准确率略有提升（0.5484→0.6236，可能属噪声）。改动只涉及一个文件，对其他模型和硬件无影响。由于没有新增测试，团队后续需依赖现有 CI 和人工 benchmark 保证该路径的稳定性。
  - 风险标记：缺少测试覆盖，CUDA 图流同步风险，仅 CUDA 路径生效

关联脉络

- PR #33957 Unknown - split source PR: 本 PR 由作者从 #33957 拆分而来，仅保留 MiniMax-M3 共享 / 路由重叠和 TRT-LLM MoE PDL 限制更改。