PR #26600 完整报告

sgl-project/sglang

Revert "[CI] FA3: ascending cuda-graph capture to avoid varlen workspace IMA (#26532) (#26550)"

合并时间：2026-05-29 04:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26600

执行摘要

- 一句话：回退 FA3 升序 CUDA Graph 捕获顺序
- 推荐动作：该 PR 是紧急回退，值得精读以了解 CUDA Graph 捕获顺序与内存分配的关系。关注后续是否会有更好的修复方案（例如限制捕获批次大小或优化内存池）。

功能与动机

PR body 明确指出，升序捕获顺序 "appears to cause broader CI OOMs than the original issue it fixed"。作者在 Torch 升级 PR 和当前 main 分支上都观察到了相同的 CUDA Graph 捕获 OOM 模式。

实现拆解

1. 删除辅助函数：在 cuda_graph_runner.py 中删除了 _ci_use_ascending_capture_order 函数，该函数用于检测是否在 CI 中使用 FA3 后端并返回升序捕获标志。
2. 恢复捕获顺序逻辑：在 capture 方法的内部函数 _capture_one_stream 中，将条件分支（升序 / 降序）恢复为统一的降序（reversed(self.capture_bs)），并移除了相关的注释和 bs_seq 中间变量。

关键文件：

- python/sglang/srt/model_executor/cuda_graph_runner.py（模块 运行时；类别 source；类型 data-contract；符号 _ci_use_ascending_capture_order）：核心变更文件，删除了辅助函数并恢复了降序捕获逻辑。

关键符号：_ci_use_ascending_capture_order

关键源码片段

python/sglang/srt/model_executor/cuda_graph_runner.py

核心变更文件，删除了辅助函数并恢复了降序捕获逻辑。

```
# 删除了整个 _ci_use_ascending_capture_order 函数
# 原函数用于判断是否应在 CI + FA3 下使用升序捕获

class CudaGraphRunner:
    ...
```

```
def capture(self) -> None:
    ...
    def _capture_one_stream(stream_idx: Optional[int] = None):
        ...
        # 恢复为统一的降序捕获，以改善内存共享
        capture_range = (
            tqdm.tqdm(list(reversed(self.capture_bs)))
            if get_tensor_model_parallel_rank() == 0
            else reversed(self.capture_bs)
        )
        for i, bs in enumerate(capture_range):
            # 原本这里有条件分支：升序 / 降序，现全部走降序
            ...
```

评论区精华

本 PR 无 review 评论，但 PR body 中作者 @mmangkad 说明了回退原因：升序捕获导致更广泛的 CI OOM，并在 Torch 升级 PR 和 main 分支上都验证了该问题。cc @hnyls2002 作为原变更的作者。

- 暂无高价值评论线程

风险与影响

- 风险：回归风险：回退后，在 CI + FA3 场景下，FLASH_ATTENTION varlen workspace slot 的 IMA（非法内存访问）问题可能重新出现（原 #26532 修复的问题）。但 OOM 问题的影响面更大，因此回退是合理的权衡。无其他风险：变更仅涉及单文件、单函数，逻辑简单。
- 影响：影响范围：仅影响 CI 中使用 FA3 后端的 CUDA Graph 捕获流程。用户影响：无，回退仅针对 CI 环境。系统影响：CI 中 FA3 相关测试可能从 OOM 恢复，但可能面临原 IMA 问题。
- 风险标记：核心路径变更，回退可能引入原问题回归

关联脉络

- PR #26532 [CI] FA3: ascending cuda-graph capture to avoid varlen workspace IMA: 本 PR 直接回退了 #26532 引入的变更。
- PR #26550 follow-up for #26532: #26532 的后续 PR，本 PR 同时也回退了其影响。