PR #29461 完整报告

sgl-project/sglang

Fix FlashInfer A2A dispatcher during CUDA graph capture

合并时间：2026-06-28 17:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29461

执行摘要

- 一句话：修复 FlashInfer A2A 在 CUDA graph 捕获时的分发逻辑
- 推荐动作：值得精读，特别是对于理解 CUDA graph 捕获与运行时 token 分发的交互。修改简洁但设计决策正确，可推广到其他类似的分发器实现中。

功能与动机

CUDA graph 捕获时所有 EP rank 使用相同的捕获批次大小，因此 x.shape[0] 在所有 rank 上一致且安全。原有逻辑在 EP>1 且 require_mlp_tp_gather 为 False 时使用静态容量 self.max_num_tokens，但捕获期间这会导致图绑定错误的几何形状，影响正确性和性能。

实现拆解

1. 导入 get_is_capture_mode：在文件顶部从 sglang.srt.model_executor.runner_utils.capture_mode 新增导入，用于判断当前是否处于 CUDA graph 捕获模式。
2. 修改 dispatch 方法中的条件分支：原有的 EP>1 分支改为同时检查 not get_is_capture_mode()，即仅在非捕获状态下进入该分支。捕获状态下所有 EP rank 使用相同的批次大小，因此直接使用 x.shape[0]（Case 3）。
3. 添加详细的结构化注释：新增一段多行注释，清晰说明三种 runtime_max_tokens_per_rank 选择场景（DP attention、EP>1 非捕获、捕获及其他情况），提高代码可维护性。

关键文件：

- python/sglang/srt/layers/moe/token_dispatcher/flashinfer.py（模块 MoE 调度；类别 source；类型 core-logic）：核心变更文件，修改了 dispatch 方法中 runtime_max_tokens_per_rank 的选择逻辑，新增导入 get_is_capture_mode 并添加详细注释。

关键符号：未识别

关键源码片段

python/sglang/srt/layers/moe/token_dispatcher/flashinfer.py

核心变更文件，修改了 dispatch 方法中 runtime_max_tokens_per_rank 的选择逻辑，新增导入 get_is_capture_mode 并添加详细注释。

```
# 新增导入（文件顶部）
from sglang.srt.model_executor.runner_utils.capture_mode import get_is_capture_mode

# dispatch 方法中的关键改动（约第 237-249 行）
dp_global = get_dp_global_num_tokens()
if dp_global is not None and len(dp_global) > 1:
    # Case 1: DP attention 场景，使用所有 DP rank 的最大 token 数
    self.runtime_max_tokens_per_rank = max(dp_global)
elif (
    self.ep_size > 1
    and not get_is_capture_mode() # <-- 新增条件：仅在非 CUDA graph 捕获时进入此分支
    and not require_mlp_tp_gather(get_global_server_args())
):
    # Case 2: EP>1 且非捕获时，使用静态容量
    self.runtime_max_tokens_per_rank = self.max_num_tokens
else:
    # Case 3: 捕获模式、EP=1 或 SP 场景，使用实际输入张量大小
    # 捕获时所有 rank 批次大小一致，x.shape[0] 安全可用
    self.runtime_max_tokens_per_rank = x.shape[0]
if self.has_dummy_token:
    self.runtime_max_tokens_per_rank = max(self.runtime_max_tokens_per_rank, 1)
```

评论区精华

该 PR 无审核讨论，作者自行合并。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。修改仅添加了一个条件分支和导入，逻辑清晰。但需确保 get_is_capture_mode() 在所有推理场景中行为正确，特别是与 speculative decoding 等复合图捕获场景的交互。
- 影响：直接影响所有使用 FlashInfer MoE A2A 且启用 CUDA graph 的推理场景，特别是 EP>1 配置。修复后可避免捕获时错误使用静态容量，提升正确性和性能。影响范围局限在单个文件的单点逻辑。
- 风险标记：CUDA graph 捕获依赖

关联脉络

- PR #29232 [Spec] Replace shared-infra dflash special-cases with capabilities (WAR barrier + seq_lens_cpu): 涉及 CUDA graph 捕获模式相关的逻辑调整，可能与本次修改的 get_is_capture_mode 函数有关联。