PR #26123 完整报告

sgl-project/sglang

Fix routed-experts device buffer overflow under DP attention

合并时间：2026-05-31 10:11

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26123

执行摘要

- 一句话：修复 DP attention 下 routed-experts 缓冲区溢出
- 推荐动作：建议尽快合并。这是一个明确正确的 bugfix，改动量小且风险低，解决了一个在 DP attention 路径下可导致静默数据损坏或崩溃的问题。值得关注的是作者将修复从大型 PR #23999 中独立拆分出来的做法，降低了集成风险。

功能与动机

在 DP attention 启用且 max_running_requests 大于 chunked_prefill_size * dp_size 的场景下，RoutedExpertsCapturer 的设备缓冲区会发生溢出，因为 _get_local_slice 使用 dp_rank * cuda_graph_batch 进行索引，当缓冲区仅按 max_running_requests 分配时，dp_rank > 0 的索引会超出边界。该问题在关联 Issue #23999 中被发现，作为独立修复从更大的 PR 中拆分出来。

实现拆解

仅修改 python/sglang/srt/state_capturer/routed_experts.py 中 RoutedExpertsCapturer.__init__ 方法，将 max_batch_size 计算从 max(chunked_prefill_size * dp_size, max_running_requests) 改为 max(chunked_prefill_size * dp_size, max_running_requests * dp_size)。同时更新注释，说明 dp_size 因子的必要性，并保留 spec decoding 尚未处理的 FIXME。

关键文件：

- python/sglang/srt/state_capturer/routed_experts.py（模块 状态捕获；类别 source；类型 core-logic；符号 RoutedExpertsCapturer.init）：唯一变更文件，修改了 RoutedExpertsCapturer.__init__中 max_batch_size 的计算，修复 DP attention 下缓冲区溢出。

关键符号：RoutedExpertsCapturer.init

关键源码片段

python/sglang/srt/state_capturer/routed_experts.py

唯一变更文件，修改了 RoutedExpertsCapturer.__init__中 max_batch_size 的计算，修复 DP attention 下缓冲区溢出。

```
# python/sglang/srt/state_capturer/routed_experts.py
# RoutedExpertsCapturer.__init__ 中关键计算部分
server_args = get_global_server_args()

# 按 dp_size 缩放，使缓冲区覆盖完整的 DP 聚合批次。
# _get_local_slice 会以 [attention_dp_rank * cuda_graph_batch, ...) 的方式索引，
# 当 max_running_requests > chunked_prefill_size 时，dp_rank > 0 的切片
# 会超出原来以 max_running_requests 为第一维的缓冲区。
# FIXME: spec decoding 的 num_verify_tokens 仍然未被考虑。
max_batch_size = max(
    server_args.chunked_prefill_size * server_args.dp_size,
    max_running_requests * server_args.dp_size, # 修复项：乘以 dp_size
)
```

评论区精华

审核者 alexnails 要求更新 FIXME 注释以反映变更，作者已照做。无其他技术争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：当 dp_size == 1 时表达式退化回原值；当 chunked_prefill_size * dp_size 已占主导时也不受影响。唯一可能的影响是轻微增加显存占用（增加量不超过 dp_size 倍），但这是正确性必须的。仍存在 spec decoding 的 num_verify_tokens 未考虑的问题（已在 FIXME 中记录）。
- 影响：影响范围小：仅修改一个文件的 2 行核心逻辑。影响用户：任何使用 MoE 模型且启用 DP attention、且 max_running_requests 超过 chunked_prefill_size dp_size 的用户将从隐式数据损坏 / 崩溃中得到修复。对 dp_size == 1 或 chunked_prefill_size 较大的用户无影响。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #23999 [moe] Capture routing softmax weights alongside routed experts and DP buffer-shape fix: 本修复最初是 #23999 的一部分，该 PR 还包含路由权重捕获功能，但因状态捕获器重构后需要重新调整而拆分。
- PR #24403 consolidate routed-experts capturer onto reusable base: 对该文件的 refactor，提取了 BaseTopkCapturer 基类，改变了文件结构。
- PR #24450 move topk capturers to srt/state_capturer/: 将路由专家捕获器移动到 state_capturer 目录，与此文件路径相关。