PR #25685 完整报告

sgl-project/sglang

[SP] Fix runtime_max_tokens_per_rank for sequence parallelism

合并时间：2026-05-19 06:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25685

执行摘要

- 一句话：修复 SP 下 runtime_max_tokens_per_rank 过大问题
- 推荐动作：建议需要关注 MoE + 序列并行性能的团队成员精读该改动。设计思路清晰（区分 DP attention 与 SP 的 workspace 需求），可作为类似分配逻辑的优化参考。

功能与动机

当启用序列并行时，get_dp_global_num_tokens() 返回的是 pre-scatter 的 scheduler token 数，可能远超实际单个 worker 在 SP 下处理 token 数，导致 A2A workspace 分配过大，增加 TTFT。PR body 明确说明“the pre-scatter scheduler token count… can exceed the A2A workspace cap”。

实现拆解

1. 条件逻辑重写：在 flashinfer.py 的 dispatch 方法中，将原来单行三目运算拆分为更清晰的分支判断。
2. 多 rank DP attention 分支：当 dp_global 非空且列表长度大于 1 时，保持原有行为——取全局最大 token 数以确保所有 rank 的 workspace 能容纳最胖的输入。
3. 单 rank 或 SP 分支：否则使用 x.shape[0]（post-scatter 后的实际 token 数）。这个数值对于 SP 而言正好反映当前 rank 在 scatter 后真正的输入规模，不会大于实际传给 A2A 的 payload。
4. 配套影响：仅修改了一个文件（flashinfer.py），无其他模块、配置或测试变更。

关键文件：

- python/sglang/srt/layers/moe/token_dispatcher/flashinfer.py（模块 调度器；类别 source；类型 core-logic）：MoE token dispatcher 的核心分发逻辑，修复了序列并行下 workspace 分配过大的问题，直接影响 TTFT。

关键符号：未识别

关键源码片段

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

MoE token dispatcher 的核心分发逻辑，修复了序列并行下 workspace 分配过大的问题，直接影响 TTFT。

```
# python/sglang/srt/layers/moe/token_dispatcher/flashinfer.py
# dispatch 方法中关键计算逻辑（PR 改动部分）

dp_global = get_dp_global_num_tokens()
if dp_global is not None and len(dp_global) > 1:
    # DP attention: 多个 DP rank 有不同 token 数。
    # 取全局最大值以保证每个 rank 的 A2A workspace 都能容纳最胖的输入。
    self.runtime_max_tokens_per_rank = max(dp_global)
else:
    # dp_size==1 或 启用序列并行（SP）：
    # 使用实际的输入张量大小（SP 模式下为 post-scatter 后的 size，普通模式为完整 batch size）。
    # 避免使用 pre-scatter 的 scheduler 计数（该计数可能超过 workspace 上限）。
    self.runtime_max_tokens_per_rank = x.shape[0]
```

评论区精华

无 review 评论，PR 由作者 merrymercy 直接提交并合并。讨论仅在 CI 触发指令中出现过一次 /tag-and-rerun-ci。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。改动局限于 flashinfer.py 中 runtime_max_tokens_per_rank 的计算逻辑，且原有 x.shape[0] 回退路径已被保留。唯一可能的问题是当 get_dp_global_num_tokens() 返回非 None 但长度为 0（理论上不应发生），此时会进入 else 分支使用 x.shape[0]，行为与原代码后备一致，无额外风险。
- 影响：影响范围：仅影响启用了序列并行（SP）且使用 MoE 的模型（如 Llama4x、DeepSeek V4）。影响程度：降低了这些场景下的 TTFT，对非 SP 或 DP 注意力场景无行为变化。团队影响：开发者无需修改其他代码即可受益。
- 风险标记：核心路径变更

关联脉络

- PR #25688 Add no_combine support to cutlass_moe_fp4: 同为 MoE 模块的性能优化改动，关注不同的 MoE 内核路径。
- PR #25569 Add DeepSeekV4 fused MoE Triton autotune support: DeepSeekV4 的 fused MoE 优化，与 SP 下的 MoE 调度可能产生协同影响。
- PR #25509 [misc] Throw error when single batch overlap is enabled on Hopper: 与 MoE 运行配置相关，影响 Hopper GPU 上的 MoE 执行路径。