PR #25907 完整报告

sgl-project/sglang

Fix FlashInfer A2A token cap sizing

合并时间：2026-05-21 14:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25907

执行摘要

- 一句话：修复 FlashInfer A2A token 容量双倍计数
- 推荐动作：此 PR 为针对 MoE 推理中 FlashInfer A2A dispatcher 的小型 bugfix，设计简洁，值得关注其默认值调整和注释中的空间计算解释。

功能与动机

PR body 明确指出旧代码在 env var 后乘以 ep_size，与 FlashInfer 内部的 moe_a2a_get_workspace_size_per_rank() 中的缩放重复，导致 workspace 尺寸被高估。commit 消息也详细解释了该问题。

实现拆解

1. 修正 max_num_tokens 计算：在 python/sglang/srt/layers/moe/token_dispatcher/flashinfer.py 的 FlashinferDispatcher.__init__ 中，移除 * self.ep_size 乘法。
2. 添加注释解释：新增注释说明 FlashInfer 的 workspace 分配逻辑，即 moe_a2a_get_workspace_size_per_rank() 会保留 ep_size * max_num_tokens * payload 的空间，因此无需在外部再次放大。
3. 调整默认值：第一次提交将默认值从 1024 提升至 16384 (per rank)，但后续提交发现该值与 ep_size 相乘后导致 CUDA graph capture 时 OOM，因此最终降为 4096 per rank。

关键文件：

- python/sglang/srt/layers/moe/token_dispatcher/flashinfer.py（模块 dispatcher；类别 source；类型 core-logic）：此文件是唯一的变更文件，包含了 FlashinferDispatcher 中 max_num_tokens 计算的核心 bugfix 和默认值调整。

关键符号：未识别

关键源码片段

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

此文件是唯一的变更文件，包含了 FlashinferDispatcher 中 max_num_tokens 计算的核心 bugfix 和默认值调整。

```
# python/sglang/srt/layers/moe/token_dispatcher/flashinfer.py (partial)
class FlashinferDispatcher(BaseDispatcher):
```

```
def __init__(self, ...):
    ...
    # TODO: Can this be a server arg and shared with deepep/mooncakeep?
    # FlashInfer sizes the workspace from the maximum dispatched tokens per
    # EP rank. See FlashInfer's moe_a2a_get_workspace_size_per_rank(),
    # which reserves ep_size * max_num_tokens * payload bytes, and the C++
    # dispatch op's epSize * runtimeMaxTokensPerRank payload buffer.
    self.max_num_tokens = get_int_env_var(
        "SGLANG_FLASHINFER_NUM_MAX_DISPATCH_TOKENS_PER_RANK", 4096
    )
    # Note: Previously this line multiplied by self.ep_size (double-counting)
    # because FlashInfer already accounts for ep_size internally.
```

评论区精华

无 review 评论，但作者通过 CI 重跑和 commit 消息中的调整说明了默认值 4096 的合理性：原 16384 per rank 在 4x B200 上被 ensure_cutedsl_wrapper 乘以 ep_size 后引起 OOM。

- 暂无高价值评论线程

风险与影响

- 风险：风险低：变更仅涉及一行 max_num_tokens 计算和默认值调整。若其他部分错误依赖旧的 ep_size 乘积值，可能引入 workspace 不足的 bug，但 FlashInfer 内部已正确缩放，因此风险很小。
- 影响：影响限于 FlashInfer A2A token dispatcher：修复了 workspace 尺寸双倍分配的问题，并提升了默认 token 容量（从 1024 到 4096 per rank），有助于减少因 workspace 不足导致的错误。对于使用自定义 env var 的用户，语义变化为 env var 现在表示 per-rank 的最大 token 数（而非总 token 数），需要相应调整。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR