PR #27945 完整报告

sgl-project/sglang

fix(moe): make FlashInfer A2A robust to collapsed global_num_tokens (moe_dense_tp_size NaN)

合并时间：2026-06-13 07:11

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27945

执行摘要

- 一句话：修复 FlashInfer A2A 在 moe_dense_tp_size 设置时 NaN 问题
- 推荐动作：值得精读，尤其是 MoE 分布式推理中 global_num_tokens 跨 rank 一致性的设计权衡。require_mlp_tp_gather 作为通用检测函数值得在其他分发器中借鉴。建议后续增加自动化测试覆盖 moe_dense_tp_size != 1 + FlashInfer A2A 组合。

功能与动机

--moe-a2a-backend flashinfer 配合 --enable-dp-attention 和 --moe-dense-tp-size 1 时，模型推理产生 NaN/inf logits（gsm8k 准确率约 0%，或触发 probability tensor contains inf/nan 设备端断言）。根因是 moe_dense_tp_size 设置导致 batch.global_num_tokens 仅含本地 count 而非全局聚合值。

实现拆解

1. 新增 import：在 flashinfer.py 中添加 from sglang.srt.utils.common import require_mlp_tp_gather，用于检测 moe_dense_tp_size 是否导致 collapse。
2. 修改 dispatch 逻辑：在 dispatch() 方法的 runtime_max_tokens_per_rank 赋值处插入一个 elif 分支。当 dp_global 长度为 1 或为 None 时，如果 ep_size > 1 且 require_mlp_tp_gather 返回 False，说明 global_num_tokens 已被塌缩，此时使用 self.max_num_tokens（静态分配容量，所有 rank 相同）替代 x.shape[0]，确保 MoeAlltoAll 的固定几何结构跨 rank 一致。
3. 配套格式修复：在 protocol.py 中将 validate_function_tool 的返回类型注解从带引号的 "ResponseTool" 改为裸的 ResponseTool，符合 ruff UP037 规则（pre-commit 自动修复）。
4. 未包含测试：PR body 明确标记未添加单元测试，仅提供了手动 gsm8k 验证结果。

关键文件：

- python/sglang/srt/layers/moe/token_dispatcher/flashinfer.py（模块 分发器；类别 source；类型 dependency-wiring；符号 dispatch）：核心修复文件，新增 import 和 dispatch 逻辑的 elif 分支，确保跨 rank 的 runtime_max_tokens_per_rank 一致。
- python/sglang/srt/entrypoints/openai/protocol.py（模块 协议层；类别 source；类型 configuration；符号 validate_function_tool）：格式修复，仅去掉返回类型注解中的引号，属于 pre-commit 自动清理，与核心 Bug 无关。

关键符号：dispatch

关键源码片段

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

核心修复文件，新增 import 和 dispatch 逻辑的 elif 分支，确保跨 rank 的 runtime_max_tokens_per_rank 一致。

```
# flashinfer.py (dispatch 方法局部 )
dp_global = get_dp_global_num_tokens()
if dp_global is not None and len(dp_global) > 1:
    # DP attention: multiple DP ranks with different token counts.
    # Use the max across ranks so the A2A workspace fits the fattest.
    self.runtime_max_tokens_per_rank = max(dp_global)
elif self.ep_size > 1 and not require_mlp_tp_gather(get_global_server_args()):
    # require_mlp_tp_gather is False, so the scheduler collapsed
    # global_num_tokens to the local count; x.shape[0] then differs
    # across EP ranks and breaks the fixed-geometry MoeAlltoAll. Use
    # the static all-rank capacity instead.
    self.runtime_max_tokens_per_rank = self.max_num_tokens
else:
    # dp_size=1 or SP: use the actual input tensor size (post-scatter
    # in SP mode, full batch otherwise). Avoids the pre-scatter
    # scheduler count which can exceed the workspace cap.
    self.runtime_max_tokens_per_rank = x.shape[0]
```

评论区精华

ch-wan 在 review 中提出疑问：“should we check require_mlp_tp_gather?” 作者 YAMY1234 回应：“Switched to not require_mlp_tp_gather(get_global_server_args()) (with the import) instead of testing moe_dense_tp_size directly, full gsm8k (1319) is 0.977.” 最终采用 require_mlp_tp_gather 作为条件，而非直接检查 moe_dense_tp_size，更具鲁棒性。

- 使用 require_mlp_tp_gather 替代直接检查 moe_dense_tp_size (design): 使用 require_mlp_tp_gather 代替直接检查 moe_dense_tp_size，更具通用性。

风险与影响

- 风险：
  1. 回归风险：修复仅影响 ep_size > 1 且 require_mlp_tp_gather=False 的路径，正常路径无改动，回归风险低。但缺少自动化单元测试覆盖该特定条件。
  2. 性能风险：回退到 self.max_num_tokens 可能略微增大单 rank 的 A2A 传输量，因为使用静态容量而非实际 token 数。但对于出问题的配置，之前直接崩溃，因此可以接受。
  3. 兼容性：引入新的 import require_mlp_tp_gather，该函数需在 sglang.srt.utils.common 中可用，否则会报 ImportError。 - 影响：用户影响：修复了使用 BlackWell 架构（如 Qwen3.5-397B-A17B-NVFP4）搭配 --moe-dense-tp-size 1、

FlashInfer A2A 和 DP attention 的用户遇到的推理崩溃问题。系统影响：仅修改了 FlashInfer 分发器的条件逻辑，其他分发器（DeepEP、Mooncake 等）不受影响。团队影响：小范围修复，无需协调其他模块。 - 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR