PR #23972 完整报告

sgl-project/sglang

fix the compatibility between --moe-dense-tp-size 1 and piecewise cuda graph

合并时间：2026-04-30 17:12

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23972

执行摘要

- 一句话：修复 piecewise CUDA graph 与 MoE dense TP 的兼容性
- 推荐动作：值得精读，特别是关注 MoE 并行策略与 CUDA graph 捕获兼容性的开发者。设计简单且可维护。

功能与动机

PR body 明确指出：python -m sglang.launch_server --model-path deepseek-ai/DeepSeek-Coder-V2-Lite-Instruct --trust-remote-code --tp 8 --moe-dense-tp-size 1 之前会 crash。

实现拆解

1. 新增导入：在 piecewise_cuda_graph_runner.py 中导入 get_attention_cp_size 和 require_gathered_buffer 工具函数。
2. 过滤 capture sizes：在 __init__ 中，若 require_gathered_buffer 为真（即 --moe-dense-tp-size 1 且 TP group 与 attention TP group 不一致），计算 mul_base（attn_tp_size * attn_cp_size 的最小公倍数），仅保留能整除 mul_base 的 capture token 数。
3. 断言保护：若过滤后无合法大小，触发断言报错，避免静默非预期行为。
4. 无测试配套：本次变更未添加单元测试或集成测试。

关键文件：

- python/sglang/srt/model_executor/piecewise_cuda_graph_runner.py（模块 调度器；类别 source；类型 core-logic；符号 require_gathered_buffer, get_attention_cp_size, PiecewiseCudaGraphRunner.init）：核心变更文件，添加了 capture sizes 的整除性过滤逻辑，修复与 --moe-dense-tp-size 1 的兼容性。

关键符号：PiecewiseCudaGraphRunner.init

关键源码片段

python/sglang/srt/model_executor/piecewise_cuda_graph_runner.py

核心变更文件，添加了 capture sizes 的整除性过滤逻辑，修复与 --moe-dense-tp-size 1 的兼容性。

```
# 文件：python/sglang/srt/model_executor/piecewise_cuda_graph_runner.py
# 在 __init__ 方法中，设置 capture_num_tokens 后立即执行过滤

self.capture_num_tokens = self.compile_config.get_capture_sizes()

# 当 layer communicator 在 attention TP group 内做 scatter/gather 时
# （例如 --moe-dense-tp-size 1），模型的 reduce_scatter 需要 token 数
# 能被 attn_tp_size * attn_cp_size 整除。
# 丢弃不满足条件的 capture 大小（与常规 CUDA graph runner 中的过滤一致）。
if require_gathered_buffer(self.model_runner.server_args):
    mul_base = self.attn_tp_size
    attn_cp_size = get_attention_cp_size()
    if mul_base % attn_cp_size != 0:
        mul_base *= attn_cp_size
    filtered = [n for n in self.capture_num_tokens if n % mul_base == 0]
    # 确保至少有一个合法大小，否则断言失败
    assert (
        len(filtered) > 0
    ), f"No piecewise CUDA graph capture sizes are multiples of {mul_base}"
    self.capture_num_tokens = filtered
```

评论区精华

无 review 讨论，仅有 Oasis-Git 的 Approve 和一条 bot 配额耗尽消息。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，逻辑与常规 CUDA graph runner 中已有的过滤一致（注释明确提及 mirrors the filter used by the regular CUDA graph runner）。但过滤后可能减少可用的 capture 大小，对极端配置（如 attn_tp_size=1 且 attn_cp_size=1）无影响，当 mul_base 很大时可能导致可用 capture 数量不足，影响性能。断言保护可提前暴露问题。
- 影响：仅影响同时启用 --moe-dense-tp-size 1 和 piecewise CUDA graph 的用户，修复启动崩溃。对其他配置无影响。
- 风险标记：缺少测试覆盖，核心路径变更

关联脉络

- 暂无明显关联 PR