

PR #30835 完整报告

sgl-project/sglang

[Tiny] Enable Full Cuda Graph with Page size = 1

合并时间: 2026-07-14 01:07

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30835>

执行摘要

- 一句话: 修复 FA4 与 full prefill CUDA graph 的 page_size 冲突
- 推荐动作: 可合并, 变更简单且正确。建议后续增加测试覆盖该组合场景。

功能与动机

FlashAttention4 后端在非 MLA 模型上默认将 page_size 强制设为 128, 但 full prefill CUDA graph 路径 (#27988) 仅支持 page_size=1。当两者同时启用时会导致冲突, 本 PR 通过增加跳过条件来兼容。

实现拆解

在 `python/sglang/srt/arg_groups/overrides.py` 的 `_fa4_page_constraint` 函数中, 现有逻辑对 FA4 后端且非 MLA 且 SM100 支持等条件时, 会将 page_size 强制设为 128。现增加一条条件: `view.cuda_graph_config.prefill.backend != Backend.FULL`, 即当 prefill CUDA graph 后端为 FULL 时, 跳过 page_size 调整, 使得 FA4 可以与 full prefill CG 共存。

关键文件:

- `python/sglang/srt/arg_groups/overrides.py` (模块 配置重写; 类别 source; 类型 core-logic; 符号 `_fa4_page_constraint`, `Backend`): 核心修改文件, 在 `_fa4_page_constraint` 中新增条件跳过 page_size 强制调整。

关键符号: `_fa4_page_constraint`

关键源码片段

`python/sglang/srt/arg_groups/overrides.py`

核心修改文件, 在 `_fa4_page_constraint` 中新增条件跳过 page_size 强制调整。

```
# python/sglang/srt/arg_groups/overrides.py
@register_post_process
def _fa4_page_constraint(view: Any) -> dict:
    if (
        (
            view.attention_backend == "fa4"
            or view.decode_attention_backend == "fa4"
            or view.prefill_attention_backend == "fa4"
```

```

)
and not view.use_mla_backend()
and is_sm100_supported()
# EAGLE topk>1 spec runs the two-pass page-tree cascade, which the FA4
# CUTLASS kernel aborts on at page_size>1. That path only works at
# page_size==1, so skip the 128 auto-force for it and keep the default.
and (view.speculative_eagle_topk or 0) <= 1
# NEW: The full prefill CUDA graph runs the FA backend at page_size==1 only
# (#27988), so skip the 128 auto-force for it and keep the default.
and view.cuda_graph_config.prefill.backend != Backend.FULL
):
    logger.warning(
        f"FA4 backend only supports page size 128 for non-MLA model architectures, "
        f"changing page_size from {view.page_size} to 128."
    )
    return {"page_size": 128}
return {}

```

评论区精华

无 review 评论，仅有一名 reviewer (ispobock) 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：新增条件仅在 `_fa4_page_constraint` 中增加一个短路判断，不影响其他路径。但缺少测试覆盖，建议添加单元测试验证 FA4 + full prefill CG + `page_size=1` 的组合。
- 影响：仅影响同时启用 FA4 后端和 full prefill CUDA graph 的用户，这些用户将不再因 `page_size` 冲突而需要手动调整。对其他用户无影响。
- 风险标记：缺少测试覆盖

关联脉络

- PR #27988 Implement full prefill CUDA graph: 本 PR 依赖 #27988 的 full prefill CUDA graph 限制 (`page_size` 仅支持 1)，并为其添加兼容性跳转。