

PR #35632 完整报告

sgl-project/sglang

[Fix] Keep deterministic GDN prefill on Triton

合并时间: 2026-08-20 19:38

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35632>

执行摘要

- 一句话: 修复确定性推理下 GDN prefill 后端选择, 强制 Triton
- 推荐动作: 该 PR 值得精读, 因为它展示了如何在性能优化与确定性保证之间做权衡, 并提供了清晰的错误处理方式。建议关注 `gdn_backend.py` 中的策略函数和新增的校验函数。

功能与动机

PR #34859 让 SM90 上的确定性推理默认使用 FlashInfer GDN prefill, 但 FlashInfer 的 prefill 是 batch 敏感的, 无法保证确定性, 这违背了确定性推理的语义。PR 的目标是恢复 Triton prefill, 确保确定性推理的稳定性。

实现拆解

1. 修改确定性模式下的默认策略: 在 `flashinfer_gdn_prefill_default` 中增加 `get_exec().deterministic.enable_deterministic_inference` 的条件, 当确定性推理启用时直接返回 `None`, 使默认后端回退到 Triton。
2. 新增显式校验函数: 新增 `_validate_gdn_linear_attn_backends`, 在 `GDNAttnBackend` 初始化时调用, 若确定性推理且 prefill 后端为 FlashInfer, 则抛出 `ValueError`, 拒绝启动。
3. 补充测试: 在 `test_gdn_prefill_backend_policy.py` 中新增两个测试, 分别验证默认策略在确定性模式下不选 FlashInfer, 以及显式指定 FlashInfer 时抛出异常。测试类改为 `CustomTestCase` 以支持更好的错误处理。

关键文件:

- `python/sglang/srt/layers/attention/linear/gdn_backend.py` (模块 核心逻辑; 类别 source; 类型 core-logic; 符号 `_validate_gdn_linear_attn_backends`): 核心修改文件, 控制 GDN prefill 后端选择策略, 新增确定性模式下的默认回退和显式校验。
- `test/registered/unit/layers/attention/test_gdn_prefill_backend_policy.py` (模块 测试; 类别 test; 类型 test-coverage; 符号 `TestFlashInferGDNPrefillBackendPolicy`, `test_declines_when_deterministic_inference_is_enabled`, `test_rejects_explicit_flashinfer_prefill_in_deterministic_mode`): 补充测试, 覆盖确定性模式下默认回退和显式拒绝的行为。

关键符号: `_validate_gdn_linear_attn_backends`, `flashinfer_gdn_prefill_default`

评论区精华

该 PR 有 5 条评论，均为 CI 重跑指令和结果，无实质讨论。Reviewer BBuf 批准了 PR，没有反对意见。

- 暂无高价值评论线程

风险与影响

- 风险：该修改主要影响确定性推理模式下的性能：在 SM90 GPU 上启用 `--enable-deterministic-inference` 时，prefill 默认不再使用 FlashInfer，可能带来性能回退，但这是为了保证确定性的必要取舍。校验函数在 GDNAttnBackend 初始化时抛出异常，可能影响显式指定 FlashInfer 的现有配置，但这是有意为之，需要用户调整配置。
- 影响：影响面集中在 GDN prefill 后端选择逻辑，主要影响使用 SM90/SM100 GPU 且启用确定性推理 (`--enable-deterministic-inference`) 的用户。该修改确保这些用户获得稳定、可重复的推理结果，代价是可能牺牲部分性能。
- 风险标记：潜在性能回退，配置变更

关联脉络

- PR #34859 *Make deterministic inference use FlashInfer GDN prefill on SM90*: 本 PR 正是修复该 PR 引入的性能优化导致的确定性缺失问题。