PR #25239 完整报告

sgl-project/sglang

[FlashInfer v0.6.12] Support FlashInfer 4over6 NVFP4

合并时间：2026-06-05 05:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25239

执行摘要

- 一句话：支持 FlashInfer NVFP4 4over6 缩放模式
- 推荐动作：变更简洁清晰，适合快速精读。重点观察环境变量命名规范以及量化缩放因子的计算逻辑，后续维护时需注意与 FlashInfer 升级的同步。

功能与动机

PR body 指出父级依赖为 flashinfer PR#3264 和 sglang PR#22918，目的是支持 FlashInfer 4over6 NVFP4，以提升 per-token activation 量化路径的精度。作者在 issue 评论中说明需要等待 #26023 和 #26854 合并。

实现拆解

1. 注册环境变量：在 python/sglang/srt/environ.py 的 Envs 类中新增 FLASHINFER_NVFP4_4OVER6 和 FLASHINFER_NVFP4_4OVER6_E4M3_USE_256 两个布尔类型环境变量，默认均为 False。
2. 调整量化缩放因子：在 python/sglang/srt/layers/moe/moe_runner/flashinfer_trtllm.py 的 fused_experts_none_to_flashinfer_trtllm_fp4 函数中，当 SGLANG_FLASHINFER_NVFP4_PER_TOKEN_ACTIVATION 启用时，根据新增环境变量选择 e4m3_max（448.0 或 256.0），进而影响传给 nvfp4_quantize 的全局缩放因子 1.0 / (e4m3_max * 6.0)。
3. 更新文档：在 docs_new/docs/references/environment_variables.mdx 中为两个新环境变量添加表格行，说明用途和默认值。
4. 配套变更：无新增测试文件，但 CI 通过测试。

关键文件：

- python/sglang/srt/layers/moe/moe_runner/flashinfer_trtllm.py（模块 MoE；类别 source；类型 core-logic）：核心变更位置，修改了 NVFP4 量化路径中的缩放因子计算，根据环境变量动态选择 e4m3_max。
- python/sglang/srt/environ.py（模块 环境变量；类别 source；类型 configuration）：新增两个环境变量的声明，是配置的入口点。
- docs_new/docs/references/environment_variables.mdx（模块 文档；类别 other；类型 documentation）：文档中新增两个环境变量的说明，方便用户理解和配置。

关键符号：未识别

关键源码片段

python/sglang/srt/layers/moe/moe_runner/flashinfer_trtllm.py

核心变更位置，修改了 NVFP4 量化路径中的缩放因子计算，根据环境变量动态选择 *e*4m3_max。

```
# 在 fused_experts_none_to_flashinfer_trtllm_fp4 函数中，量化 hidden states 部分
if envs.SGLANG_FLASHINFER_NVFP4_PER_TOKEN_ACTIVATION.get():
    from flashinfer import SfLayout, nvfp4_quantize

    # 默认 E4M3 最大值为 448.0
    e4m3_max = 448.0
    # 如果开启了 4over6 并且使用 256 作为上限，则覆盖为 256.0
    if (
        envs.FLASHINFER_NVFP4_4OVER6.get()
        and envs.FLASHINFER_NVFP4_4OVER6_E4M3_USE_256.get()
    ):
        e4m3_max = 256.0

    hs_fp4_bytes, hs_sf_bytes, per_token_scale = nvfp4_quantize(
        hidden_states,
        1.0 / (e4m3_max * 6.0), # 全局缩放因子取决于 e4m3_max
        sfLayout=SfLayout.layout_linear,
        per_token_activation=True,
    )
```

评论区精华

在 review 中，b8zhong 建议将新环境变量的文档移至专门的 RL（Reinforcement Learning）章节，认为这些变量不适用于常规用户。zianglih 回应说该功能确实可用于常规 FP4 精度提升，因此维持在当前文档位置。

- 新环境变量的文档位置 (documentation): 维持现有文档位置（环境变量总表），不迁移。

风险与影响

- 风险：风险极低。核心风险在于环境变量与 FlashInfer 库的 4over6 配置必须严格对齐：若 FLASHINFER_NVFP4_4OVER6 未在 FlashInfer 侧同步开启，可能导致量化缩放因子计算错误，影响模型输出精度。此外，无新增测试覆盖，回归依赖现有 CI 用例。
- 影响：该 PR 仅影响启用了 SGLANG_FLASHINFER_NVFP4_PER_TOKEN_ACTIVATION 的 NVFP4 量化路径。对不使用的用户无任何影响。开启后，用户可通过环境变量调整缩放因子，可能提升低比特模型精度，性能不变。
- 风险标记：缺少测试覆盖，配置依赖外部行为

关联脉络

- PR #26023 unknown: PR 作者在 issue 评论中说明等待该 PR 合并。
- PR #26854 unknown: PR 作者在 issue 评论中说明等待该 PR 合并。
- PR #22918 unknown: PR body 中提及的父级依赖。