PR #46880 完整报告

vllm-project/vllm

[Bugfix][NVFP4 MoE] Pad gated intermediate to 64 for FlashInfer TRT-LLM shuffle (M%128)

合并时间：2026-07-16 05:37

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46880

执行摘要

- 一句话：修复 NVFP4 MoE gated 中间维度对齐错误
- 推荐动作：该 PR 值得合并，修复逻辑清晰且数学上正确。建议精读理解 NVFP4 MoE 的权重对齐策略。

功能与动机

Issue #46879 报告 nvidia/Gemma-4-26B-A4B-NVFP4 在 B200 上以 --tensor-parallel-size 4 运行时，在 process_weights_after_loading 阶段因 FlashInfer 的 get_shuffle_matrix_sf_a_row_indices 中的 assert M % 128 == 0 而崩溃。问题根源是 gated NVFP4 MoE 的中间维度对齐不足。

实现拆解

该 PR 仅修改了一个文件，调整了 gated 路径的对齐参数。

1. 定位问题：在 prepare_nvfp4_moe_layer_for_fi_or_cutlass 函数中（flashinfer_fp4_moe.py:340），gated 路径的 min_alignment 被设为 16，这仅能保证 NVFP4 量化尺度块对齐，但 FlashInfer TRT-LLM shuffle 要求门控 / 上行融合维度 (up_mult * padded_intermediate，gated 时 up_mult=2) 是 128 的倍数。
2. 修复对齐：将 gated 路径的 min_alignment 从 16 改为 64，使得 padhed_intermediate 是 64 的倍数，从而 2 * padded_intermediate 是 128 的倍数，满足 shuffle 断言。
3. 数学验证：例如 Gemma-4-26B-A4B 在 TP=4 时 rank-local intermediate = 176 = 704/4，原逻辑 2 * round_up(176, 16) = 352，352 % 128 = 96 触发断言；改用 64 对齐后 2 * round_up(176, 64) = 384，384 % 128 = 0 通过断言。
4. 非 gated 路径不变：非 gated 路径（up_mult=1）继续使用 min_alignment=128，Marlin 路径（PR #45295）不受影响。

关键文件：

- vllm/model_executor/layers/quantization/utils/flashinfer_fp4_moe.py（模块 量化器；类别 source；类型 data-contract；符号 prepare_nvfp4_moe_layer_for_fi_or_cutlass）：修改了 gated 路径的对齐参数，是唯一变更的文件。

关键符号：prepare_nvfp4_moe_layer_for_fi_or_cutlass

关键源码片段

vllm/model_executor/layers/quantization/utils/flashinfer_fp4_moe.py

修改了 gated 路径的对齐参数，是唯一变更的文件。

```
# file: vllm/model_executor/layers/quantization/utils/flashinfer_fp4_moe.py
# 位于 prepare_nvfp4_moe_layer_for_fi_or_cutlass 函数中
# Align weights for FI NVFP4 MoE kernels.
# FlashInfer's TRT-LLM block-scale shuffle asserts the gate/up row dim
# (= up_mult * padded_intermediate, up_mult = 2 when gated) is a multiple of
# 128. So gated needs padded_intermediate % 64 (2*64=128); the old value 16
# left 2*intermediate a multiple of only 32, so an NVFP4 MoE whose rank-local
# intermediate is not 128-aligned at TP>1 (e.g. Gemma-4-26B-A4B at tp4) hit
# `assert M % 128 == 0`. Padded rows are zero -> outputs unchanged.
min_alignment = 64 if is_gated else 128
w13, w13_scale, w2, w2_scale, padded_intermediate = (
    align_fp4_moe_weights_for_fi(
        w13, w13_scale, w2, w2_scale, is_act_and_mul, min_alignment
    )
)
layer.moe_config.intermediate_size_per_partition = padded_intermediate
```

评论区精华

Review 中讨论较少，核心是修复的正确性和 FP8 路径的潜在问题。

- lucianommartins 确认了问题分析，并指出 FP8 TRT-LLM shuffle 可能也有相同的 128 行约束，建议跟进。
- mikekg 回应建议作为后续 PR，因为目前没有可重现的 FP8 模型。
- 最终 mgoin 批准了 PR。
- FP8 TRT-LLM shuffle 也可能有相同约束 (question): mikekg 建议作为后续 PR，因为暂无可重现的 FP8 模型。

风险与影响

- 风险：风险较低。改动仅改变对齐参数，填充行权重和尺度均为零，输出不变。非 gated 路径不受影响。但需注意：如果未来有其他模型或 TP 配置导致 intermediate 本身不是 128 的倍数，可能仍然需要更通用的方案。
- 影响：影响范围有限，主要影响使用 FlashInfer TRT-LLM 后端的 NVFP4 MoE 模型在 TP>1 时的 gated 中间维度对齐。修复了 Gemma-4-26B-A4B-NVFP4 在 TP=4 时的启动失败。其他模型和 TP 组合若 intermediate 已经 128 对齐则不受影响。
- 风险标记：单文件变更，核心路径变更，缺少测试覆盖

关联脉络

- PR #46879 [Bug] nvidia/Gemma-4-26B-A4B-NVFP4 fails assert M % 128 == 0 on FlashInfer TRT-LLM NVFP4 MoE at -tp 4 (Blackwell/B200): 直接关联的 bug 报告，PR

修复了该 issue。

- PR #45295 [Bugfix] NVFP4 MoE Marlin path alignment: PR body 中提及 Marlin 路径不受影响。