PR #49452 完整报告

vllm-project/vllm

[Bugfix][CI] Fix `topk_softplus_sqrt` no-op on non-XPU platforms

合并时间：2026-07-23 06:36

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49452

执行摘要

- 一句话：修复 topk_softplus_sqrt 在非 XPU 平台变成空操作
- 推荐动作：该 PR 为重要的回归 bugfix，推荐立即合并。变更逻辑简单明了，修复了明确的死代码问题。虽然只有一行改动，但其影响涵盖了 CUDA 和 ROCm 平台，是必要的修复。值得学习的是 review 中没有忽略这种看似微小的控制流错误，体现了对跨平台代码中条件分支正确性的重视。

功能与动机

49408 的 XPU workaround 意外地将 return 放在了函数体级别（非分支内），导致 torch.ops._moe_C.topk_softplus_sqrt(...) 在非 XPU 平台上变成不可达代码。Issue 评论中 @eugr 指出 "breaks models on CUDA too"。PR body 确认该 bug 导致所有 topk_softplus_sqrt 测试在 AMD GPU 上失败，输出全为零。

实现拆解

1. 定位问题：在 vllm/_custom_ops.py 的 topk_hash_softplus_sqrt 函数中，return 语句（第 2464 行）位于 if current_platform.is_xpu() 分支之外，导致所有非 XPU 路径在调用真正的 kernel 之前提前返回。
2. 修复方法：将第 2463 行的 return 移动到 if 分支内部（第 2462 行），使得只有 XPU 平台执行无 is_padding 参数的 kernel 调用后返回；非 XPU 平台继续执行带 is_padding 参数的 kernel 调用。
3. 测试验证：在 ROCm 开发容器中运行 pytest -q tests/kernels/moe/test_topk_softplus_sqrt.py，修复前全部失败，修复后 1657 passed, 3 skipped, 16 warnings。
4. 无额外改动：仅修改一行（加号 1，减号 1），未涉及测试、配置、文档的改动。

关键文件：

- vllm/_custom_ops.py（模块 自定义算子；类别 source；类型 core-logic；符号 topk_hash_softplus_sqrt）：修复的核心文件：将 return 移入 XPU 分支，恢复非 XPU 平台的 kernel 调用。单文件、单行逻辑更改，但影响整个 MoE 路由的正确性。

关键符号：topk_hash_softplus_sqrt

关键源码片段

vllm/_custom_ops.py

修复的核心文件：将 return 移入 XPU 分支，恢复非 XPU 平台的 kernel 调用。单文件、单行逻辑更改，但影响整个 MoE 路由的正确性。

```
def topk_hash_softplus_sqrt(
    topk_weights: torch.Tensor,
    topk_indices: torch.Tensor,
    token_expert_indices: torch.Tensor,
    gating_output: torch.Tensor,
    renormalize: bool = False,
    routed_scaling_factor: float = 1.0,
    e_score_correction_bias: torch.Tensor | None = None,
    input_tokens: torch.Tensor | None = None,
    hash_indices_table: torch.Tensor | None = None,
    is_padding: torch.Tensor | None = None,
) -> None:
    if current_platform.is_xpu():
        # TODO: Remove after vllm-xpu-kernels supports is_padding.
        # XPU 平台暂时不支持 is_padding，调用不带 is_padding 的旧版 kernel
        torch.ops._moe_C.topk_softplus_sqrt(
            topk_weights, topk_indices, token_expert_indices,
            gating_output, renormalize, routed_scaling_factor,
            e_score_correction_bias, input_tokens, hash_indices_table,
        )
        return # 将 return 移到分支内，确保非 XPU 不会提前返回

    # 非 XPU 平台：使用支持 is_padding 的 kernel 调用（修复前此处不可达）
    torch.ops._moe_C.topk_softplus_sqrt(
        topk_weights, topk_indices, token_expert_indices,
        gating_output, renormalize, routed_scaling_factor,
        e_score_correction_bias, input_tokens, hash_indices_table,
        is_padding, # 带 is_padding 参数
    )
```

评论区精华

代码审核中，yewentao256、jikunshang、mgoin 均批准（LGTM）。由于是 fork 的 PR，自动化 review 被禁用。没有实质性讨论线程。

- 暂无高价值评论线程

风险与影响

- 风险：该修复仅调整了 Python 控制流，将 return 移至正确的分支内，技术上风险极低。但需注意：XPU 平台的 workaround（调用无 is_padding 参数的 kernel）保持不变，如果未

来 is_padding 在非 XPU 平台上有预期行为，该修复正常无误。变更是纯 Python 控制流，不涉及 CUDA/ROCm 内核改动，回归风险很小。

- 影响：影响范围：修复了所有非 XPU 平台（ROCm、CUDA）上 topk_softplus_sqrt 内核的死代码 bug，直接恢复了这些平台上 MoE 相关功能的正确性。影响程度：关键 bug fix，因为受影响的是 MoE 路由核心调用，错误行为（输出全零）可能导致模型推理结果错误。修复后 AMD MI300、MI355 的 CI 测试能够通过。

用户影响：AMD 和 NVIDIA GPU 用户将不再遇到 MoE 路由输出为零的问题。 系统影响：CI 增加一次构建和测试的时间；社区用户无需手动修复。 团队影响：极低，仅一行改动，review 快速通过。

- 风险标记：核心路径变更，跨平台回归风险，向前兼容

关联脉络

- PR #49408 [XPU] WA of topk_softplus_sqrt arg mismatch on XPU: 本 PR 修复了 #49408 引入的回归：XPU workaround 将 return 放在了错误的作用域，导致非 XPU 平台 kernel 调用成为死代码。