

PR #46804 完整报告

vllm-project/vllm

[XPU][UT]Fix xpu pass_config.fuse_norm_quant assert issue

合并时间: 2026-06-30 12:13

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46804>

执行摘要

- 一句话: XPU pass_config.fuse_norm_quant 断言修复
- 推荐动作: 该 PR 是常规的平台适配 bug 修复, 代码量小, 逻辑清晰, 值得快速合并。对于关注 XPU 支持的开发者, 可以了解平台条件检查的模式。

功能与动机

修复 XPU 上运行测试时遇到的 `AssertionError: pass_config.fuse_norm_quant: expected True, got False`。PR body 明确指出了受影响的测试: `tests/test_config.py::test_vllm_config_defaults[RedHatAI/Qwen3-8B-speculator.eagle3-compilation_config5-2]`、`test_vllm_config_explicit_overrides`、`test_fusion_pass_op_priority`。

实现拆解

1. 修改条件检查: 在 `vllm/config/compilation.py` 的 `PassConfig.__post_init__` 方法中, 将原来的条件 `if self.enable_qk_norm_rope_fusion and not current_platform.is_cuda_alike()` 改为 `if self.enable_qk_norm_rope_fusion and not (current_platform.is_cuda_alike() or current_platform.is_xpu())`, 使得 XPU 平台也被认为是支持的, 从而不再禁用该 fusion。
2. 更新 warning 日志: 将对应的 warning 消息从 "CUDA or ROCm" 更新为 "CUDA, ROCm or XPU", 以准确反映支持的平台。
3. 调整测试断言: 在 `tests/test_config.py` 的 `test_vllm_config_explicit_overrides` 中, 将原先的 `assert config.compilation_config.pass_config.enable_qk_norm_rope_fusion is True` 改为根据平台条件动态判断: `current_platform.is_cuda_alike() or current_platform.is_xpu()`, 确保 XPU 上该字段为 `True`, 其他非 CUDA/XPU 平台为 `False`。

关键文件:

- `vllm/config/compilation.py` (模块 编译配置; 类别 source; 类型 core-logic): 核心修复文件, 修改了 `PassConfig.__post_init__` 中的平台条件检查, 将 XPU 纳入允许使用 QK Norm+RoPE fusion 的平台列表。
- `tests/test_config.py` (模块 配置测试; 类别 test; 类型 test-coverage): 测试文件, 更新了 `test_vllm_config_explicit_overrides` 中断言, 使其根据平台动态判断, 修复了在 XPU 上因硬编码 `True` 导致的断言失败。

关键符号：未识别

关键源码片段

vllm/config/compilation.py

核心修复文件，修改了 PassConfig.__post_init__ 中的平台条件检查，将 XPU 纳入允许使用 QK Norm+RoPE fusion 的平台列表。

```
# vllm/config/compilation.py
class PassConfig:
    ...
    def __post_init__(self) -> None:
        # ... other validations ...
        # 原先是：if self.enable_qk_norm_rope_fusion and not current_platform.is_cuda_alike():
        # 修复后：将 XPU 也作为支持的平台
        if self.enable_qk_norm_rope_fusion and not (
            current_platform.is_cuda_alike() or current_platform.is_xpu()
        ):
            logger.warning_once(
                "QK Norm + RoPE fusion enabled but the current platform is not "
                "CUDA, ROCm or XPU. The fusion will be disabled."
            )
            self.enable_qk_norm_rope_fusion = False
        # ... other validations ...
```

tests/test_config.py

测试文件，更新了 test_vllm_config_explicit_overrides 中断言，使其根据平台动态判断，修复了在 XPU 上因硬编码 True 导致的断言失败。

```
# tests/test_config.py
def test_vllm_config_explicit_overrides():
    # ... 之前的设置 ...
    pass_config = PassConfig(enable_qk_norm_rope_fusion=True)
    compilation_config = CompilationConfig(
        cudagraph_mode=CUDAGraphMode.NONE, pass_config=pass_config
    )
    config = VllmConfig(
        model_config=quantized_model,
        optimization_level=OptimizationLevel.O2,
        compilation_config=compilation_config,
    )
    assert config.compilation_config.cudagraph_mode == CUDAGraphMode.NONE
    # 原来硬编码为 True，现在根据平台动态判断
    # 在 CUDA 或 XPU 上应为 True，其他平台为 False
    assert config.compilation_config.pass_config.enable_qk_norm_rope_fusion is (
        current_platform.is_cuda_alike() or current_platform.is_xpu()
    )
    # 模式检查不变
    assert config.compilation_config.mode == CompilationMode.VLLM_COMPILE
```

评论区精华

PR 没有 review 评论讨论。审核由 jikunshang (批准) 和 yewentao256 (评论 LGTM) 完成, claude[bot] 自动评论提示从 fork 发起, 未进行自动审查。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低。变更仅在平台检查中添加了 XPU, 属于条件分支扩展, 不影响 CUDA/ROCm 原有行为。测试断言改为动态判断, 避免了硬编码 True 带来的平台兼容性问题。
- 影响: 影响范围限定在 XPU 平台。修复后的代码使得 XPU 用户可以正常使用 `enable_qk_norm_rope_fusion` 功能, 相关测试 (`test_vllm_config_explicit_overrides` 等) 在 XPU 上不再断言失败。对 CUDA 和 ROCm 无影响。
- 风险标记: 暂无

关联脉络

- PR #46987 [XPU] [RMSNorm] revert weightless change on xpu: 同为 XPU 平台修复, 涉及 `vllm/kernels/xpu_ops.py`, 表明近期 XPU 有多项适配工作。