

PR #50262 完整报告

vllm-project/vllm

[ROCm][CI] Fix Kimi K3 KDA on ROCm

合并时间: 2026-07-29 23:47

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/50262>

执行摘要

- 一句话: 修复 ROCm 平台 Kimi K3 KDA 测试路径与 FlashKDA 误选
- 推荐动作: 精读, 尤其是 `is_flashkda_supported` 的守卫添加方式, 可作为跨平台内核选型的参考模式。

功能与动机

AMD "Kernels KDA Test" CI 作业 (mi300_1) 持续报红。调查发现两个独立问题: 一是测试文件路径因 PR #50089 移动而失效, 二是 FlashKDA 作为 CUDA-only 扩展在 ROCm 上被错误选择, 导致导入崩溃。修复后 CI 从 1 failed 变为 0 failed。

实现拆解

1. 修复 CI 测试路径(`.buildkite/test-amd.yaml`): 将“Kernels KDA Test”任务的命令从 `pytest -v -s kernels/test_kda.py` 更新为 `pytest -v -s models/kimi_k3/test_kda.py models/kimi_k3/test_kda_metadata.py`, 并同步更新 `source_file_dependencies` 指向 Kimi K3 KDA 相关源文件。
2. 修复 FlashKDA 误选 ROCm(`vllm/models/kimi_k3/nvidia/kda.py`): 在 `is_flashkda_supported()` 函数开头添加 `if not current_platform.is_cuda(): return False` 守卫, 防止 ROCm 平台因计算能力主版本号匹配 SM90 而错误选择 CUDA-only 的 FlashKDA 预填充后端。

关键文件:

- `vllm/models/kimi_k3/nvidia/kda.py` (模块 模型内核; 类别 source; 类型 data-contract; 符号 `is_flashkda_supported`): 修复 FlashKDA 在 ROCm 上被错误选择的核心逻辑文件, 增加了平台守卫。
- `.buildkite/test-amd.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 修复因测试文件迁移导致的 CI 路径失效, 更新 `source_file_dependencies` 和测试命令。

关键符号: `is_flashkda_supported`

关键源码片段

`vllm/models/kimi_k3/nvidia/kda.py`

修复 FlashKDA 在 ROCm 上被错误选择的核心逻辑文件, 增加了平台守卫。

```

def is_flashkda_supported(
    head_dim: int,
    dtype: torch.dtype,
    lower_bound: float | None,
) -> bool:
    # ROCm 平台必须明确排除: CUDA 的 DeviceCapability major 9/10/12 匹配
    # 会与 ROCm gfx942 的 major=9 冲突。若不拦截, 后续 import vllm._flashkda_C 会崩溃。
    if not current_platform.is_cuda():
        return False

    capability = current_platform.get_device_capability()
    return (
        capability is not None
        and capability.major in (9, 10, 12)
        and head_dim == 128
        and dtype == torch.bfloat16
        and lower_bound is not None
    )

```

评论区精华

无 review 评论, PR 由 Andreaskaratzas 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低。变更仅涉及两处: CI 配置路径更新和增加平台守卫。is_cuda() 守卫只会影响 ROCm 平台, CUDA 行为不受影响, 且已验证 ROCm 上 test_flashkda_correctness 测试从失败变为跳过。
- 影响: 直接影响 AMD ROCm 的 Kimi K3 KDA 测试作业, 修复后 CI 可正常通过。对 CUDA 用户无影响。
- 风险标记: 平台兼容性修复, CI 配置变更

关联脉络

- PR #50089 [Model] Add Kimi K3 support: model files and kernels [1/N]: 此 PR 移动了 KDA 测试文件到 tests/models/kimi_k3/, 导致本 PR 需要修复 CI 路径。