

PR #46018 完整报告

vllm-project/vllm

[Hardware][AMD][CI] Fix Spec Decode Eagle test group

合并时间: 2026-06-22 06:40

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46018>

执行摘要

- 一句话: 修复 AMD CI 中 Spec Decode Eagle 测试组
- 推荐动作: 可直接合入。对关注 AMD 平台测试的团队有一定参考价值。

功能与动机

PR body 明确指出 'fixes the Spec Decode Eagle test group', 目的是使该测试在 AMD CI 中正确通过。

实现拆解

1. 配置更新: 在 `.buildkite/test_areas/spec_decode.yaml` 中为 Spec Decode Eagle 步骤添加了 AMD 镜像配置, 指定设备 `mi325_1`、超时 45 分钟, 并扩展文件依赖以覆盖 ROCm 相关模块。
2. 测试修复: 在 `tests/v1/e2e/spec_decode/test_spec_decode.py` 中, 将 DeepSeek 模型的注意力后端从 `TRITON_MLA` 改为 `ROCM_AITER_MLA`, 以匹配 AITER 环境。
3. CI 调整: 在 `.buildkite/test-amd.yaml` 中, 将 Spec Decode Eagle 步骤的超时从 180 分钟缩短为 45 分钟, 并标记为 `optional: true`, 避免阻塞 CI 流水线。

关键文件:

- `.buildkite/test_areas/spec_decode.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 添加了 Eagle 测试的 AMD 镜像配置, 使该测试能在 AMD CI 中运行。
- `tests/v1/e2e/spec_decode/test_spec_decode.py` (模块 推测解码; 类别 test; 类型 test-coverage): 修复了 DeepSeek 模型在 AITER 环境下的注意力后端配置。
- `.buildkite/test-amd.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 调整了 Eagle 测试在 AMD CI 中的超时和可选性。

关键符号: 未识别

关键源码片段

`tests/v1/e2e/spec_decode/test_spec_decode.py`

修复了 DeepSeek 模型在 AITER 环境下的注意力后端配置。

```
# tests/v1/e2e/spec_decode/test_spec_decode.py
# 在 AITER 环境下, DeepSeek 模型应使用 ROCM_AITER_MLA 而非 TRITON_MLA
```

```
with monkeypatch.context() as m:
    m.setenv("VLLM_MLA_DISABLE", "1")

    if attn_backend == "ROCM_AITER_FA" and current_platform.is_rocm():
        if "deepseek" in model_setup[1].lower():
            m.setenv("VLLM_ROCM_USE_AITER", "1")
            m.delenv("VLLM_MLA_DISABLE", raising=False)
            # 修复：使用正确的 AITER MLA 后端
            attention_config = {"backend": "ROCM_AITER_MLA"}
        else:
            m.setenv("VLLM_ROCM_USE_AITER", "1")
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅限于 CI 配置和测试参数调整，不影响核心推理逻辑。注意力后端切换为 ROCM_AITER_MLA 可能影响 DeepSeek 模型在 ROCm 上的 Eagle 测试结果，但已通过测试验证。
- 影响：仅影响 AMD CI (ROCm) 中的 Spec Decode Eagle 测试组，使其能够正确运行并通过。不涉及用户侧或生产环境。
- 风险标记：暂无

关联脉络

- PR #43081 [SpecDecode] Support DFlash with FlashInfer: 同为推测解码相关功能，涉及注意力后端选择。
- PR #43132 [Spec Decode] Add Qwen3 architecture support for EAGLE3: 同为 Eagle 推测解码相关 PR。