PR #39296 完整报告

vllm-project/vllm

[XPU][UT] update UTs in CI

合并时间：2026-04-09 09:38

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39296

执行摘要

本次 PR 更新了 XPU CI 的测试配置，通过忽略多个 hf3fs 相关的 KV 连接器单元测试来修复因 Torch 2.11 升级导致的测试失败。这是一个临时性基础设施调整，旨在确保 CI 流水线通过，为后续彻底解决依赖问题的 PR（#37947）创造条件。变更影响范围有限，主要涉及 CI 稳定性，但需注意测试覆盖减少的潜在风险。

功能与动机

为什么做？根据 Issue 评论，XPU CI 测试失败是由于 Torch 2.11 升级未同步升级 Triton 3.7 引发的兼容性问题。评论者 jikunshang 指出：

> failed case is due to torch 2.11 upgrade. it didn't upgrade triton 3.7 as well. https://github.com/vllm-project/vllm/pull/37947 will fix.

因此，本次 PR 作为临时修复，通过调整测试范围来确保 CI 通过，避免阻塞开发流程，同时等待 #37947 彻底解决依赖版本问题。

实现拆解

实现方案仅涉及两个 CI 配置文件的微小调整：

文件路径	变更内容	作用
.buildkite/intel_jobs/test-intel.yaml	在 pytest 命令的 --ignore 列表中新增 test_hf3fs_client.py、test_hf3fs_connector.py、test_hf3fs_metadata_server.py	定义 CI 流水线测试步骤，忽略特定测试文件
.buildkite/scripts/hardware_ci/run-xpu-test.sh	同步添加相同的忽略规则	确保本地和 CI 环境执行一致的测试命令

关键代码逻辑如下（以 yaml 文件为例）： pytest -v -s v1/kv_connector/unit \
--ignore=v1/kv_connector/unit/test_multi_connector.py \
--ignore=v1/kv_connector/unit/test_example_connector.py \
--ignore=v1/kv_connector/unit/test_lmcache_integration.py \
--ignore=v1/kv_connector/unit/test_hf3fs_client.py \

```
--ignore=v1/kv_connector/unit/test_hf3fs_connector.py \
--ignore=v1/kv_connector/unit/test_hf3fs_metadata_server.py
```

评论区精华

Review 讨论较为简单：

- gemini-code-assist[bot]的自动评论概括了变更："更新硬件 CI 测试脚本以忽略多个与 KV 连接器相关的单元测试，特别是排除 hf3fs 相关测试文件"。
- jikunshang直接批准，未提出技术争议。

核心讨论发生在 Issue 评论中，明确了测试失败的根因和修复路径，为本次 PR 提供了上下文。

风险与影响

风险分析：

1. 测试覆盖缺口：忽略 hf3fs 测试可能掩盖 KV 连接器模块中与 HuggingFace 文件系统集成的潜在缺陷。
2. 临时修复依赖：若 #37947 延迟或失败，测试忽略可能长期存在，增加回归风险。
3. 配置一致性：需确保其他测试环境（如开发者本地）同步调整，否则可能导致测试结果不一致。

影响分析：

- 对用户：无直接影响，属于内部 CI 调整。
- 对系统：提升 XPU CI 的稳定性，避免测试失败阻塞合并流程。
- 对团队：提供更可靠的 CI 反馈，但测试覆盖减少可能略微增加 KV 连接器模块的质量风险。

关联脉络

本次 PR 与历史 PR 的关联主要体现在：

- PR #37947：在 Issue 评论中被明确提及，将彻底解决 Torch/Triton 版本不匹配问题，本次 PR 为其铺平道路，属于同一问题修复链条的前置步骤。
- 其他 CI 相关 PR：如 #38580（ROCm CI 修复）、#37980（DeepGEMM 集成）等，反映了仓库在持续优化多平台 CI 基础设施，本次 PR 是 Intel GPU（XPU）方向的具体维护动作。

从更大视角看，这体现了 vLLM 项目在支持多样化硬件平台（如 Intel GPU、AMD ROCm）过程中，对 CI 稳定性的持续投入。