

PR #47376 完整报告

vllm-project/vllm

[XPU][CI] Split test_punica_ops into separate pytest invocations for stability

合并时间: 2026-07-02 15:50

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47376>

执行摘要

- 一句话: 拆分 XPU LoRA Punica CI 测试为独立进程
- 推荐动作: 值得合并。这是一个标准的最佳实践: 将大型测试文件拆分为独立进程执行, 消除相互影响。可作为 CI 稳定化的示例参考。

功能与动机

稳定 XPU LoRA Punica 内核 CI, 通过将 `test_punica_ops.py` 拆分为三个独立的 pytest 调用来避免所有测试一起运行时出现的失败, 并移除不断增长的 `--deselect` 排除列表。

实现拆解

1. 目标文件: `.buildkite/intel_jobs/lora_intel.yaml` 中的 LoRA Punica Kernels 步骤。
2. 变更前: 使用单条 `pytest -v -s lora/test_punica_ops.py` 命令附带 16 个 `--deselect` 参数逐个排除不稳定用例。
3. 变更后: 将该步骤拆分为三条独立的 `pytest -v -s lora/test_punica_ops.py::test_kernels`、`pytest -v -s lora/test_punica_ops.py::test_kernels_hidden_size` 和 `pytest -v -s lora/test_punica_ops.py::test_add_lora_fused_moe_early_exit` 命令, 每条命令只运行特定测试类。
4. 结果: 消除了所有 `--deselect` 排除项, 每个测试类独立运行避免相互干扰, CI 流程更简洁、更稳定。

关键文件:

- `.buildkite/intel_jobs/lora_intel.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 唯一变更文件, 将 `test_punica_ops.py` 的单进程测试拆分为三个独立 pytest 调用, 去除所有 `--deselect` 排除项, 提升 CI 稳定性。

关键符号: 未识别

评论区精华

无实质性 review 讨论。机器人 `claude[bot]` 因 PR 来自 fork 自动跳过; `jikunshang` 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅修改 CI 配置文件的命令部分，不涉及任何源码逻辑。拆分后三个测试类分别运行，若个别类仍存在资源竞争可能仍需进一步隔离，但整体方向正确。
- 影响：影响范围仅限于 Intel GPU CI 的 LoRA Punica Kernels 步骤。对用户无影响；对团队而言，CI 稳定性提升，维护成本降低（不再需要持续追加排除项）。
- 风险标记：仅配置变更，无代码风险

关联脉络

- 暂无明显关联 PR