

PR #47193 完整报告

vllm-project/vllm

[ROCm][CI] Enable LoRA TP Distributed Test Group In AMD CI

合并时间: 2026-07-01 14:24

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47193>

执行摘要

- 一句话: 修复 AMD CI LoRA TP 分布式测试环境变量和 Quark 兼容性
- 推荐动作: 推荐合并。此 PR 修复了 AMD CI 测试稳定性的两个实际问题, 修改范围小且经过验证 (已获得批准)。团队应在 Quark 0.12 发布后及时跟进, 移除跳过逻辑。

功能与动机

AMD CI 中的 LoRA TP 分布式测试组存在两个问题:

1. 从 CUDA CI 复制过来的 `PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True` 环境变量在 ROCm 上不必要且导致 `test_chatglm3_lora_tp4_fully_sharded_loras` 测试失败 (custom all reduce 出错);
2. Quark 量化库与 Torch 2.11 存在已知不兼容, 导致 `test_gptoss_tp.py` 测试失败, 需等到 Quark 0.12 修复。参见 PR body 和 Buildkite 日志。

实现拆解

1. 移除不兼容的环境变量 (`.buildkite/test-amd.yaml`): 删除 LoRA TP (Distributed) 步骤中的两行 `export` 命令——`VLLM_WORKER_MULTIPROC_METHOD=spawn` 和 `PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True`。前者是冗余设置 (已在别处保证), 后者在 ROCm 上会引发 custom all reduce 错误。
2. 添加 Quark 版本检测与测试跳过逻辑 (`tests/lora/test_gptoss_tp.py`): 在文件开头新增导入 `importlib.metadata`、`find_spec`、`torch`、`packaging.version`, 并定义变量 `QUARK_TORCH_COMPATIBLE`。该变量检查: 若 `torch >= 2.11`, 则要求 `amd-quark >= 0.12`; 若 `torch < 2.11`, 则直接认为兼容。然后在模块级别检查 `current_platform.is_rocm()` 且 `QUARK_TORCH_COMPATIBLE` 为假时, 调用 `pytest.skip` 跳过整个测试文件, 并附上跳转原因和上游 issue 链接。

关键文件:

- `tests/lora/test_gptoss_tp.py` (模块 测试文件; 类别 test; 类型 test-coverage; 符号 `QUARK_TORCH_COMPATIBLE`): 引入 Quark 版本检测和条件跳过逻辑, 解决 Torch 2.11 与 Quark 不兼容问题。
- `.buildkite/test-amd.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 移除导致 custom all reduce 失败的 CUDA 专属环境变量。

关键符号：未识别

关键源码片段

`tests/lora/test_gptoss_tp.py`

引入 Quark 版本检测和条件跳过逻辑，解决 Torch 2.11 与 Quark 不兼容问题。

```
# SPDX-License-Identifier: Apache-2.0
# SPDX-FileCopyrightText: Copyright contributors to the vLLM project

import importlib.metadata
from importlib.util import find_spec

import pytest
import torch
from packaging import version

import vllm
from vllm.lora.request import LoRARequest
from vllm.platforms import current_platform

from ..utils import multi_gpu_test

# Require amd-quark >= 0.12 on torch >= 2.11.
# Earlier torch releases work with older quark versions. See
# https://github.com/amd/Quark/issues/34
# TODO: Remove once amd-quark>=0.12.0
# 检查 Quark 是否安装并与当前 torch 版本兼容
QUARK_TORCH_COMPATIBLE = find_spec("quark") is not None and (
    version.parse(importlib.metadata.version("amd-quark")) >= version.parse("0.12.0")
    if version.parse(torch.__version__.split("+")[0]) >= version.parse("2.11")
    else True
)

# 仅在 ROCm 平台且 Quark 不兼容时跳过整个模块
if current_platform.is_rocm() and not QUARK_TORCH_COMPATIBLE:
    pytest.skip(
        "This test requires amd-quark >= 0.12 on torch >= 2.11.",
        allow_module_level=True,
    )

MODEL_PATH = "openai/gpt-oss-20b"
# ... 其余测试代码保持不变
```

评论区精华

无 review 讨论。机器人 Claude 评论此 PR 来自 fork，自动审查已禁用。AndreasKaratzas 批准了 PR。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。修改仅限于 CI 配置和测试文件：
 - `.buildkite/test-amd.yaml`：只删除两个环境变量，不影响其他步骤；
 - `tests/lora/test_gptoss_tp.py`：新增的版本检测逻辑只在 ROCm 平台且 Quark 版本不满足时跳过测试，对非 ROCm 平台无影响，测试逻辑本身未改变。但需注意：跳过 `test_gptoss_tp.py` 意味着该测试在 Quark 0.12 发布前在 ROCm CI 上不再运行，团队需跟踪恢复。
- 影响：影响范围：仅限 AMD/ROCm CI 的 LoRA TP 分布式测试组。用户影响：无；用户代码和模型推理不受影响。系统影响：AMD CI 的 LoRA TP 测试将更稳定，避免因不必要环境变量或 Quark 兼容性导致的误报失败。团队影响：需要开发者在 Quark 0.12 发布后回退跳过逻辑，重新启用测试。
- 风险标记：暂无

关联脉络

- PR #47242 [CI/Build] Fix LoRA testing: 同为 LoRA 测试修复，但针对多模态环境变量未恢复问题。
- PR #45368 [xpu][lora]: Align LoRA implementation with Punica GPU: fix `_apply_expand` rank mismatch, `add_inputs` hardcode, and MoE EP: 同为 LoRA 相关 bugfix，但针对 XPU 平台。