PR #38317 完整报告

vllm-project/vllm

[ROCm][CI] Enable hybrid chunked prefill test

合并时间：2026-03-30 10:30

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38317

执行摘要

该 PR 在 ROCm 平台上启用了混合分块预填充测试，通过移除 CUDA-only 跳过标记和添加特定模型跳过条件，同时扩展 CI 配置以支持 4xMI325 GPU 测试环境，提高了 vLLM 在 AMD 硬件上的测试覆盖率。

功能与动机

动机源于扩展 ROCm 平台测试覆盖的需求。PR 描述指出：' 移除测试文件中的全局 CUDA-only 跳过，让测试在 ROCm 上运行 '，并 ' 为 NVIDIA Nemotron 模型添加 ROCm 上的目标跳过，因为 modelopt 量化不在 ROCm 支持的量化列表中 '。这旨在确保混合分块预填充功能在 AMD 硬件上的正确性和兼容性。

实现拆解

- CI 配置扩展：在 .buildkite/test-amd.yaml 中新增一个测试步骤，使用 4xMI325 GPU 运行 test_hybrid_chunked_prefill.py 测试，指定硬件和依赖文件。
- 测试逻辑调整：在 tests/v1/e2e/test_hybrid_chunked_prefill.py 中：
  - 移除：@pytest.mark.skipif(not current_platform.is_cuda(), reason="CUDA not available")
  - 添加：为 NVIDIA Nemotron 模型参数添加 pytest.mark.skipif(not current_platform.is_cuda(), reason="modelopt quantization is supported only on CUDA")

评论区精华

review 中仅有一条讨论：tjtanaa 评论 'Since the original condition is not current_platform.is_cuda() let's retain the check as not current_platform.is_cuda()'，强调测试跳过条件的语法一致性。作者 AndreasKaratzas 迅速回应并修改代码，评论为 'Done :)'，确保条件正确设置。

风险与影响

风险：新增 CI 步骤可能增加资源消耗；特定模型跳过可能影响测试完整性；跨平台依赖外部量化支持列表。 影响：对用户无直接影响；对系统提升 ROCm 测试覆盖，有助于早期问题发现；对团队优化 CI 流程，支持多 GPU 测试。

关联脉络

与本 PR 相关的历史 PR 包括 38450（ROCm CI 修复）、38414（ROCm 变体更新）、38108（ROCm 测试修复），共同体现 vLLm 项目对 ROCm 平台的持续集成和测试扩展趋势。