PR #37906 完整报告

vllm-project/vllm

[ROCm][CI] Split Entrypoints Integration (API Server 1) into 3 jobs

合并时间：2026-03-24 09:48

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37906

执行摘要

此 PR 将 AMD CI 中的 Entrypoints 集成测试作业拆分为三个并行作业，以便在 ROCm 硬件上更有效地跟踪回归，优化测试执行时间，但引入了配置重复的维护风险。

功能与动机

动机源于需要在 AMD CI 中应用类似的测试拆分，以匹配其他 CI 配置（如 PR #37882）。PR body 明确表示："Applies the splitting in AMD CI external signal as well, so that we can easily track regressions on ROCm hardware too." 这有助于隔离故障并加速 CI 流程，特别是在多硬件平台上。

实现拆解

实现仅修改了 .buildkite/test-amd.yaml 文件：

- 原作业拆分：将"Entrypoints Integration (API Server 1)"作业拆分为三个新作业：
  - Part 1: 运行 entrypoints/openai/chat_completion 测试（排除特定文件如 test_chat_with_tool_reasoning.py）
  - Part 2: 运行 entrypoints/openai/completion 和 entrypoints/openai/speech_to_text 测试
  - Part 3: 运行剩余的 entrypoints/openai 测试（排除已覆盖的子目录）
- 代理池应用：为 mi325 和 mi355 两个代理池都实施相同拆分，确保测试执行一致性。

评论区精华

review 中，gemini-code-assist[bot] 提出了关键设计问题：

> "There is significant configuration duplication between the newly introduced Part 1, Part 2, and Part 3 jobs. This makes the CI configuration harder to maintain, as any change to common settings will need to be applied in three places." 建议使用 YAML anchors 创建可重用模板。然而，PR 被批准合并，未显示是否采纳此建议，留下了可维护性隐患。

风险与影响

风险：

- 配置重复：公共设置（如 timeout_in_minutes、source_file_dependencies）在多个作业中重复，增加维护错误风险。
- 测试覆盖：拆分可能意外遗漏某些测试，但基于现有 PR #37882 的模式，风险较低。

影响：

- 对用户：无直接影响。
- 对团队：改进 ROCm 测试回归跟踪，可能减少 CI 运行时间，提升开发效率。
- 对系统：CI 执行更细化，但配置复杂性增加，需额外维护。

关联脉络

此 PR 是仓库中 CI 优化趋势的一部分：

- PR #37882：直接在另一个 CI 配置中实施相同拆分，是本 PR 的灵感来源，显示跨平台测试拆分的一致性需求。
- PR #37016：早期拆分 V1 Others 测试的 PR，表明团队持续通过并行化测试作业来提升 CI 效率，特别是在多硬件支持场景。整体来看，vllm-project 正致力于优化测试基础设施以应对复杂硬件环境。