

PR #55919 完整报告

vllm-project/vllm

[CI][ROCm] Increase timeouts for AMD MI300 jobs

合并时间: 2026-09-09 05:06

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/55919>

执行摘要

本 PR 是一个纯 CI 配置变更，将四个在 AMD MI300 GPU 上运行的 Buildkite 测试作业的超时时间从 35-60 分钟增加到 50-75 分钟，以解决近期构建中因 CI 开销增加导致的非故障超时问题。变更范围小且风险低，旨在提升 AMD 平台 CI 的稳定性。

功能与动机

根据 PR body 描述，以下作业在 Buildkite 构建 #12703 和 #12711 中达到了当前超时限制：

- “Async Engine, Inputs, Utils, Worker” (35 → 50 分钟)
- “Entrypoints Integration (API Server Generate)” (50 → 65 分钟)
- “RayExecutorV2” (60 → 75 分钟)
- “Multimodal Models (Extended Pooling)” (50 → 75 分钟)

这些作业的测试在超时前仍在正常进度中，没有断言失败，表明是 CI 环境的额外开销（而非代码问题）导致了超时。因此，需要调整超时时间以提供足够的余量，同时为真正卡住的作业保留合理上限。

实现拆解

变更仅涉及两个 CI 配置文件，调整了四个作业的 `timeout_in_minutes` 字段：

1. 主配置文件 `.buildkite/test-amd.yaml`：

- 将 “Async Engine, Inputs, Utils, Worker” 作业的超时从 35 分钟增加到 50 分钟。
- 将 “Entrypoints Integration (API Server Generate)” 作业的超时从 50 分钟增加到 65 分钟。
- 将 “RayExecutorV2” 作业的超时从 60 分钟增加到 75 分钟。
- 将 “Multimodal Models (Extended Pooling)” 作业的超时从 50 分钟增加到 75 分钟。

2. 镜像配置文件 `.buildkite/test_areas/models_multimodal.yaml`：

- 将定义为 “:nvidia: (H200 MIG 18GB) Multimodal Models (Extended Pooling)” 作业的镜像（AMD 部分）超时时间从 60 分钟同步增加到 75 分钟，以保持与主配置文件中 MI300 版本的一致性。

评论区精华

评论很少，讨论不深入：

- AndreasKaratzas: 提供了批准 (Approval) 。
- claude[bot]: 指出这是一个 fork PR, 自动化审查被禁用。

这表明该变更被维护者视为直接、低风险的维护任务, 无需深入设计讨论。

风险与影响

技术风险:

- 超时值校准: 风险在于新的超时值是否足够精确。如果设置过高, 可能会延迟发现真正卡住的作业; 如果设置过低, 问题可能重现。PR 中提到新值基于观察到的运行时间并增加了余量, 表明进行了校准。
- 基础设施依赖: 如果未来 AMD MI300 CI 基础设施性能显著改善, 这些超时时间可能变得过于宽松。

影响范围:

- 用户: 无直接影响。
- 系统: 间接影响, 通过减少因 CI 超时导致的假失败, 提高了 AMD MI300 平台 CI 的可靠性。
- 团队: 正面影响, 减少了 CI 假失败对开发工作的干扰, 提升了开发体验。

关联脉络

从同仓库近期历史 PR 分析来看, 没有发现直接关联的 PR。该变更是一个独立的 CI 维护任务。然而, 它反映了对 AMD ROCm 平台持续投入和维护的趋势 (近期有多个 ROCm 相关 PR, 如 #54112、#54787、#55808) 。