

PR #47591 完整报告

vllm-project/vllm

[ROCm][CI] Increasing parallelism in Basic Models Tests (Extra Initialization)

合并时间: 2026-07-06 15:06

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47591>

执行摘要

该 PR 调整了 AMD ROCm CI 中 6 个测试任务的超时时间和并行度，旨在减少 Basic Models Tests 因串行执行导致的超时失败。变更仅涉及 `.buildkite/test-amd.yaml` 配置文件，无任何源代码修改，风险极低。

功能与动机

AMD CI 的 Basic Models Tests (Extra Initialization) 经常因串行执行耗时过长而超时，例如 [buildkite 构建 #10424](#)。为解决此问题，该 PR 增加部分任务的超时时间并提高其并行度。

实现拆解

变更集中在 `.buildkite/test-amd.yaml` 一个文件中，共 6 处修改：

1. vLLM IR Tests: 超时时间从 30 分钟 → 180 分钟。
2. Kernels KDA Test: 超时时间从 30 分钟 → 180 分钟。
3. Model Runner V2 Pipeline Parallelism (4 GPUs): 超时时间从 60 分钟 → 180 分钟。
4. Model Runner V2 Spec Decode: 超时时间从 45 分钟 → 180 分钟。
5. Basic Models Tests (Extra Initialization): 并行度从 2 → 6。
6. GGUF Plugin: 超时时间从 30 分钟 → 180 分钟。

所有修改均为 YAML 键值更新，不涉及逻辑或依赖变动。

评论区精华

无实质性讨论: claude[bot] 自动评论指出无法审查 fork 的 PR, tjtanaa 直接批准。

风险与影响

- 风险：几乎为零，仅涉及 CI 配置。但并行度提升至 6 可能会增加 AMD CI 集群的资源竞争，需观察后续排队情况。
- 影响：仅影响 AMD ROCm CI 流程，减少 Basic Models Tests 超时概率，对系统和用户无影响。

关联脉络

该 PR 与 #47730 (AMD CI 添加 TTY 支持) 同属 AMD CI 持续优化工作，体现了团队对 CI 稳定性的关注。