PR #38956 完整报告

vllm-project/vllm

[ci] Switch some CI jobs to H200 MIG slices

合并时间：2026-04-06 04:26

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38956

执行摘要

本 PR 将 25 个通过验证的 CI 测试步骤的设备目标切换为 H200 MIG 18GB 队列，基于构建 59734 的测试结果，旨在优化 CI 资源分配。变更仅涉及 Buildkite 配置文件更新，对代码逻辑无直接影响，需配套 ci-infra 队列支持。

功能与动机

为解决 CI 测试资源多样化需求，利用 H200 MIG 切片提供 18GB GPU 内存环境，提升测试覆盖和效率。PR body 明确提到“Adds device: h200_18gb to CI test steps that were validated as passing”，关联 Issue #325 添加了对应队列支持，以扩展 GPU 资源类型。

实现拆解

修改了 14 个 Buildkite 配置文件，在选定测试步骤中添加 device: h200_18gb 字段。关键文件包括：

- .buildkite/test_areas/basic_correctness.yaml：基础正确性测试
- .buildkite/test_areas/engine.yaml：引擎和调度测试
- .buildkite/test_areas/models_multimodal.yaml：多模态模型测试

提交历史显示初始标记 50 个步骤，后缩减到 25 个仅保留通过验证的步骤，避免包含失败或未运行作业。

评论区精华

Review 无实质技术讨论，仅 gemini-code-assist[bot] 自动评论。Issue 评论中，khluu 指出：“2 CPU tests failing are not related.. the rest of B200 jobs failing is because of infra issue”，确认变更范围有限，仅影响 25 个 L4 作业。

风险与影响

风险：

1. 依赖外部 ci-infra 变更，若队列支持不到位，可能导致测试排队失败。
2. 新设备类型可能引入未预见的兼容性问题，尽管已通过验证。
3. 多文件配置修改存在疏忽风险，但变更简单（仅添加字段）。

影响：

- CI 测试将部分使用 H200 MIG 资源，可能优化资源利用或测试多样性。
- 对用户无直接影响，仅内部测试环境调整。
- 团队需确保 ci-infra 同步，以维持测试正常执行。

关联脉络

与历史 PR 如 #38959、#38941、#38951 等同为 CI 配置调整，反映团队在持续优化测试基础设施。这些变更共同支持 vLLM 项目在多样化硬件环境下的测试稳健性。