PR #43030 完整报告

vllm-project/vllm

[ci] Route 28 gpu_1_queue tests to h200_35gb queue

合并时间：2026-05-19 12:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43030

执行摘要

- 一句话：28 个 CI 测试从 gpu_1_queue 迁移到 h200_35gb
- 推荐动作：建议跟进清理 misc.yaml 中的遗留 gpu: h100 字段，避免后续冲突。此外，可考虑对类似的配置进行统一审查，确保硬件分配清晰。

功能与动机

这些测试在 h200_18gb (1g.18gb MIG) 上出现 OOM 失败，但在 h200_35gb (1g.35gb MIG) 上全部通过（见 build #66777）。需要将测试路由到更大的 GPU 分片以维持 CI 稳定性。

实现拆解

1. 预验证：在 h200_35gb 队列的 build #66777 上运行 28 个候选测试，全部通过。
2. OOM 确认：同一批测试在 h200_18gb 队列的 build #66798 上运行，25/28 因内存不足失败。
3. 修改配置：在 .buildkite/test_areas/ 下的 12 个 YAML 文件中，为对应的测试步骤添加 device: h200_35gb 行。受影响文件包括 models_multimodal.yaml(5 处)、models_basic.yaml(4 处)、model_runner_v2.yaml(3 处)、models_language.yaml(3 处) 等。
4. 遗留问题：在 misc.yaml 中，对 "Acceptance Length Test (Large Models)" 步骤同时设置了 device: h200_35gb 和 gpu: h100，存在调度冲突隐患，尚未清理。

关键文件：

- .buildkite/test_areas/models_multimodal.yaml（模块 CI 配置；类别 config；类型 configuration）：修改了 5 处，是本次变更最多的文件，涉及多模态模型测试步骤
- .buildkite/test_areas/models_basic.yaml（模块 CI 配置；类别 config；类型 configuration）：修改了 4 处，涉及基础模型测试
- .buildkite/test_areas/model_runner_v2.yaml（模块 CI 配置；类别 config；类型 configuration）：修改了 3 处，涉及 V2 模型运行器测试
- .buildkite/test_areas/models_language.yaml（模块 CI 配置；类别 config；类型 configuration）：修改了 3 处，涉及语言模型测试

关键符号：未识别

评论区精华

唯一的 review 评论来自 gemini-code-assist[bot]，指出 .buildkite/test_areas/misc.yaml 中新增的 device 字段与已有的 gpu: h100 字段冲突，可能导致调度问题。该问题在 PR 中未解决，需要后续清理。

- misc.yaml 中 device 与 gpu 字段冲突 (other): 未在本次 PR 中解决；PR 合并后仍保留 gpu: h100，需后续清理。

风险与影响

- 风险：风险较低。主要风险是 misc.yaml 中同时指定 device 和 gpu 可能导致 BuildKite 调度器行为不确定，但实际运行中未报告问题。其他变更仅添加 device 字段，未修改现有字段，风险很小。
- 影响：对用户无直接影响。对 CI 系统，这些测试将始终在 h200_35gb 上运行，避免了在 h200_18gb 上的 OOM 失败，提高了 CI 可靠性。同时减轻了 h200_18gb 队列的压力，但增加了 h200_35gb 的负载。
- 风险标记：配置冲突未清理

关联脉络

- 暂无明显关联 PR