

PR #47094 完整报告

vllm-project/vllm

[ROCm][CI] Move LM Eval Large Models (8 GPUs) to mi300 pool

合并时间: 2026-06-30 09:59

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47094>

执行摘要

- 一句话: LM Eval 8 卡测试从 mi325 迁至 mi300 池
- 推荐动作: 可直接合并。此 PR 为纯基础设施配置调整, 变更简单且已获批准。推荐 CI 维护者关注迁移后测试是否在 mi300 池顺利调度。

功能与动机

将 8 GPU LM-eval 测试从 mi325 池迁移到 mi300 池, 以释放 mi325 资源并集中管理 mi300 上的评估任务。评审者 AndreasKaratzas 明确指出: "We moved this to empty the MI325 queue and reuse it elsewhere"。

实现拆解

1. 在 mi300 区域新增步骤: 在 .buildkite/test-amd.yaml 的 mi300 章节中, 紧接现有 "LM Eval Large Models (8xH200-8xMI300)" 步骤之后, 插入原属于 mi325 的 "ROCm LM Eval Large Models (8 GPUs)" 步骤, mirror_hardware 和 agent_pool 修改为 amdmi300 和 mi300_8。
2. 从 mi325 区域删除原步骤: 移除 mi325 章节中同名的旧步骤配置块 (mirror_hardware: [amdmi325], agent_pool: mi325_8)。
3. 保持所有其他参数不变: 包括 timeout_in_minutes: 180、num_gpus: 8、working_dir、source_file_dependencies、commands (仍执行 test_lm_eval_correctness.py --config-list-file=configs/models-large-rocm.txt --tp-size=8) 和 optional: true。

关键文件:

- .buildkite/test-amd.yaml (模块 CI 配置; 类别 config; 类型 configuration): 唯一变更文件, 包含 Buildkite pipeline 配置, 将 8 GPU LM-eval 步骤从 mi325 池迁移至 mi300 池。

关键符号: 未识别

评论区精华

无实质性讨论。Claude 自动评论因 PR 来自 fork 而禁用审核; AndreasKaratzas 直接批准并说明动机是清空 mi325 队列以复用资源。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 低风险：仅为 YAML 配置变更，未修改任何代码或测试逻辑。
 - 潜在回归：如果 mi300 8 卡池的 GPU 型号、驱动或环境配置与预期不符，测试可能失败，但已有其他 mi300 步骤运行正常，风险极低。
 - 容量影响：将 8 卡评测迁移至 mi300 后，该池负载增加，可能影响该池其他任务的调度。
- 影响：
 - 影响范围：仅限 ROCm CI 的 LM-eval 8 卡测试步骤。
 - 影响程度：低。测试行为完全不变，硬件池迁移不会改变测试结果。
 - 团队影响：CI 维护者通过集中管理 mi300 上的评估工作，可更灵活分配 mi325 资源。
 - 风险标记：低风险配置变更

关联脉络

- 暂无明显关联 PR