PR #44370 完整报告

vllm-project/vllm

[ROCm][CI] Move Model Executor test step from MI250 to MI300 (gfx942)

合并时间：2026-06-04 01:23

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44370

执行摘要

- 一句话：将 Model Executor CI 步骤从 MI250 迁移至 MI300
- 推荐动作：值得关注：这是一个典型的“硬件代际迁移”操作，展示了在 CI 中如何因硬件能力差异（FP8 支持）而调整测试分配，对维护多硬件 CI 的团队有参考价值。

功能与动机

Model Executor 步骤在 MI250 上因 FP8 量化测试不兼容而失败，且执行时间约 38 分钟。将步骤重新指派到 MI300 可同时解决兼容性和性能问题（PR body 及 reviewer 评论）。

实现拆解

1. 删除 MI250 区段的 Model Executor 步骤：在 .buildkite/test-amd.yaml 中移除了位于 # mi250・model_executor 注释块下的整个步骤定义，包括 label、timeout_in_minutes、mirror_hardwares、agent_pool 等字段。
2. 在 MI300 区段新增 Model Executor 步骤：在 # mi300・lora 和 # mi300・models / language 之间插入新的步骤定义，配置完全一致，但 mirror_hardwares 改为 [amdexperimental, amdproduction, amdgfx942nightly, amdmi300]，agent_pool 改为 mi300_1。
3. 测试命令保持不变：commands 部分原封不动保留，包括安装依赖、设置环境变量以及执行 pytest -v -s model_executor -m '(not slow_test)' 和 pytest -v -s entrypoints/openai/completion/test_tensorizer_entrypoint.py。

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；将 Model Executor 测试步骤从 MI250 区段整体迁移到 MI300 区段，修改了 mirror_hardwares 和 agent_pool 配置。

关键符号：未识别

评论区精华

AndreasKaratzas 在 review 中指出 MI250 上的失败源于量化测试，建议直接将步骤整体迁移到 MI300（gfx942），而非 sharding 或跳过失败用例。JartX 随后按建议执行了纯净迁移。

- Model Executor 步骤迁移方案 (design): 采纳迁移方案，使用纯净的步骤移动（无并行 sharding）。

风险与影响

- 风险：风险较低：仅涉及 CI 配置 YAML 的步骤迁移，无任何源码或测试逻辑变更。潜在风险包括 MI300 资源竞争（需确保 agent_pool 容量足够）以及新步骤的 mirror_hardwares 列表是否与已有 CI agent 匹配。
- 影响：影响范围限于 AMD CI 流水线：Model Executor 测试将从 MI250 机器移动到 MI300 机器执行。FP8 相关测试将获得正常运行环境，整体测试稳定性预期改善。MI250 资源压力略有减轻，MI300 资源消耗增加。
- 风险标记：CI 配置变更，资源竞争风险

关联脉络

- 暂无明显关联 PR