PR #42025 完整报告

vllm-project/vllm

[ROCm][CI] Stage B gating

合并时间：2026-05-15 16:49

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42025

执行摘要

- 一句话：AMD CI 第二阶段镜像门禁配置
- 推荐动作：值得关注该 PR 作为 AMD CI 基础设施分阶段扩展的一部分。建议确保所有 mirror 块包含必要的环境变量，并考虑逐步将 optional 测试提升为强制测试以提高覆盖质量。

功能与动机

PR 标题和描述明确指出这是“Stage B gating”，即第二阶段的门禁镜像。目的是将已有的测试组镜像到 AMD 硬件（mi300/mi250）上运行，以扩大 CI 覆盖并确保 AMD 平台的稳定性。同时移除不再需要的旧测试组，避免资源浪费。

实现拆解

1. 在 test_areas 中添加 AMD mirror 块：在 engine.yaml 中为 Engine (1 GPU) 和 e2e Scheduling (1 GPU) 添加 mirror 块，分别指定 mi300 和 mi250 设备，并设置超时和依赖。
2. 在 entrypoints.yaml 中添加 AMD mirror：为 API Server openai - Part 1/2/3 和 API Server 2 添加 mirror 块，复用顶层 commands，仅重写 device 和 timeout_in_minutes。
3. 在 kernels.yaml 中添加 AMD mirror：为 Kernels Quantization Test 添加 mirror 块，并额外包含 ROCm 特定源文件依赖（如 test_rocm_skinny_gemms.py 和 _aiter_ops.py）。
4. 在 models_language.yaml 中添加 AMD mirror：为 Language Models Test (Extended Pooling) 添加 mirror 块。
5. 调整 test-amd.yaml 主流水线：删除旧的 V1 Core + KV + Metrics 测试组（已由新镜像组替代），并为多个 entrypoints 测试组添加 optional: true，使其在 AMD 上不强制要求通过。
6. 根据 Issue 评论替换测试组：用 Entrypoints Integration (API Server openai - Part 1) 替换了原计划中的 V1 Core + KV + Metrics，用 Language Models Test (Extended Pooling) 替换了原计划中的 Kernels Attention Test %N。

关键文件：

- .buildkite/test-amd.yaml（模块 主流水线；类别 config；类型 configuration）：AMD CI 主流水线配置，删除旧测试组并添加 optional 标志

- .buildkite/test_areas/entrypoints.yaml（模块 测试分组；类别 config；类型 configuration）：测试分组配置，为 entrypoints 集成测试添加 AMD mirror
- .buildkite/test_areas/engine.yaml（模块 测试分组；类别 config；类型 configuration）：引擎和调度测试分组添加 AMD mirror
- .buildkite/test_areas/kernels.yaml（模块 测试分组；类别 config；类型 configuration）：内核测试分组添加 AMD mirror，包含 ROCm 特定依赖
- .buildkite/test_areas/models_language.yaml（模块 测试分组；类别 config；类型 configuration）：语言模型扩展池化测试分组添加 AMD mirror

关键符号：未识别

评论区精华

review 中 gemini-code-assist[bot] 指出在 misc.yaml 和 entrypoints.yaml 的 mirror 块中缺少 export VLLM_WORKER_MULTIPROC_METHOD=spawn，该环境变量在 ROCm 多进程场景下对防止死锁至关重要。该问题未被解决，但 PR 仍获得 khluu 的批准并合并。

- AMD mirror 缺少 VLLM_WORKER_MULTIPROC_METHOD=spawn (correctness): 未在 PR 中修正，但仍被批准合并。

风险与影响

- 风险：
  1. 环境变量遗漏：部分 mirror 块缺少 VLLM_WORKER_MULTIPROC_METHOD=spawn，可能导致 AMD CI 测试在多进程场景下死锁，降低 CI 可靠性。
  2. 可选测试退化风险：多个新镜像组被标记为 optional，若长期不维护可能变成常绿失败而不被关注，失去覆盖价值。
  3. 配置维护负担：部分 mirror 块完全重定义 commands，与顶层 commands 不一致，增加后续维护成本。 - 影响：对用户无直接影响。对开发和 CI 团队，AMD 测试覆盖范围扩大，有助于提前发现 AMD 特定问题；但 optional 标记降低了门禁的强制力，需要额外关注测试稳定性。长期看有助于提升 AMD 平台支持质量。 - 风险标记：环境变量遗漏，可选测试覆盖风险，配置维护负担

关联脉络

- 暂无明显关联 PR