

PR #37243 完整报告

vllm-project/vllm

[ROCm][CI] Refine gating tests

合并时间: 2026-05-06 13:05

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/37243>

执行摘要

- 一句话: 调整 ROCm CI 门控测试的设备映射与镜像配置
- 推荐动作: 对于关注 vLLM CI 基础设施和 ROCm 测试门控的团队成员, 本 PR 值得精读。特别是 run-amd-test.sh 中条件引用的修复, 体现了处理 CI 测试命令传递的细致考虑。其余配置更改较为直接, 留意评论中提及的潜在不一致即可。

功能与动机

PR body 明确指出 'This PR refines AMD mirror gating amidst new MI300 nodes', 作者在评论中补充关联 ci-infra PR#347 以协调基础设施变更。目的是配合新节点部署, 将 AMD 门控测试从旧设备平滑迁移到新设备。

实现拆解

1. 移除旧设备镜像: 在 engine.yaml 中删除 mi325_2 和 mi325_4 的 AMD 镜像 (V1 e2e 多卡测试); 在 misc.yaml、models_basic.yaml、models_language.yaml、models_multimodal.yaml 中删除对应的 mi325_1 镜像块。
2. 新增新设备镜像: 在 entrypoints.yaml 的 LLM 及 API Server 测试中加入 mi300_1 镜像; 在 misc.yaml 的 V1 Sample+Logits 和 V1 Core+KV+Metrics 中改为 mi300_1; 在 models_language.yaml 的 Hybrid 测试中新增 mi300_1 镜像; 在 models_multimodal.yaml 的所有标准测试中改为 mi300_1; 在 samplers.yaml 中改为 mi250_1。
3. 修复 run-amd-test.sh 引用问题: 引入 commands_source 变量区分命令来自环境变量还是参数列表, 仅在 argv 模式执行重新引用 (re_quoting) 以避免对已正确引用的环境变量命令造成二次破坏。
4. 更新顶层 pipeline 文件: 修改 .buildkite/test-amd.yaml 以包含调整后的测试组列表, 确保 CI pipeline 正确调度。

关键文件:

- .buildkite/test_areas/models_language.yaml (模块 测试配置; 类别 config; 类型 configuration): 核心演示了语言模型测试的 AMD 镜像从 MI325_1 切换到 MI300_1, 并新增了 Hybrid 测试的 AMD 镜像块。
- .buildkite/test_areas/models_multimodal.yaml (模块 测试配置; 类别 config; 类型 configuration): 多模态测试的四个 AMD 镜像全部从 mi325_1 更换为 mi300_1, 同时移

除了最后一个标准测试的镜像块。

- `.buildkite/scripts/hardware_ci/run-amd-test.sh` (模块 CI 脚本; 类别 infra; 类型 infrastructure) : 脚本改进避免了重复引用对 `VLLM_TEST_COMMANDS` 造成破坏, 体现了对 CI 健壮性的细致考虑。
- `.buildkite/test_areas/engine.yaml` (模块 测试配置; 类别 config; 类型 configuration) : 移除了两个多卡测试 (2GPUs、4GPUs) 的 AMD 镜像, 意味着这些测试不再在 AMD GPU 上运行。
- `.buildkite/test_areas/entrypoints.yaml` (模块 测试配置; 类别 config; 类型 configuration) : 在 LLM Integration 测试中新增 `mi300_1` 镜像, 同时移除了 API Server Part 1 和 Part 2 的 `mi325_1` 镜像, 改为不运行 AMD 镜像或仅在 LLM 测试中覆盖。
- `.buildkite/test_areas/misc.yaml` (模块 测试配置; 类别 config; 类型 configuration) : 调整了 Spec Decode 移除镜像, 同时 V1 Sample+Logits 和 V1 Core+KV+Metrics 改为 `mi300_1`。存在与 PR 描述不一致的潜在问题。
- `.buildkite/test-amd.yaml` (模块 顶层配置; 类别 config; 类型 configuration) : 顶层 CI pipeline 文件, 根据下层测试组的变化调整了包含的测试步骤列表。

关键符号: 未识别

关键源码片段

`.buildkite/scripts/hardware_ci/run-amd-test.sh`

脚本改进避免了重复引用对 `VLLM_TEST_COMMANDS` 造成破坏, 体现了对 CI 健壮性的细致考虑。

```
# 在 run-amd-test.sh 中, 引入 commands_source 变量以区分命令来源
if [[ -n "${VLLM_TEST_COMMANDS:-}" ]]; then
    commands="${VLLM_TEST_COMMANDS}"
    commands_source="env" # 标记命令来自环境变量, 引用已正确保留
    echo "Commands sourced from VLLM_TEST_COMMANDS (quoting preserved)"
else
    commands="$*"
    commands_source="argv"
    if [[ -z "$commands" ]]; then
        echo "Error: No test commands provided." >&2
        exit 1
    fi
fi

# 只有来自 argv 的命令需要重新引用处理, 避免破坏已经正确的引用
if [[ "$commands_source" == "argv" ]]; then
    commands=$(re_quote_pytest_markers "$commands")
    echo "After re-quoting: $commands"
else
    echo "Skipping re-quoting for VLLM_TEST_COMMANDS input"
fi
```

评论区精华

在 review 中，gemini-code-assist[bot] 提出了两个问题：

- misc.yaml 镜像标签与描述不符：PR 描述称要为 Examples 测试添加镜像，但实际改动在 Spec Decode 测试上，可能导致覆盖错误。该评论未收到人工回复。
- models_language.yaml 命令重复：新增的 AMD 镜像块中的 commands 与主测试 commands 几乎一致，机器人建议使用 YAML 锚点消除重复以降低维护成本。该建议未被采纳。
 - misc.yaml 镜像测试标签与 PR 描述不一致 (correctness): 未得到作者或维护者回应，问题未解决。
 - models_language.yaml 中 AMD 镜像命令重复 (design): PR 未采纳该建议，当前版本仍保留重复命令。

风险与影响

- 风险：
 - 配置不一致风险：misc.yaml 中镜像绑定可能不符合预期，导致门控测试无法正确覆盖目标硬件（Comment #1）。
 - 命令重复维护风险：models_language.yaml 中 AMD 镜像的命令与主测试命令重复，后续修改参数需同步两处，易致遗漏。
 - 脚本变更影响：run-amd-test.sh 的条件引用逻辑变更可能影响其他依赖该脚本的 CI 流程。
 - 设备资源依赖：新设备 mi300_1/mi250_1 资源不足可能导致测试排队或超时。
 - 影响：本 PR 仅影响 AMD ROCm CI 的测试调度和硬件映射，对 vLLM 产品的用户无直接功能影响。CI 维护者需关注新设备的稳定性和资源利用率。由于去掉了部分旧设备的镜像，门控覆盖率可能短暂下降，但长期将提升对新硬件的覆盖。
- 风险标记：配置不一致风险，命令重复维护，脚本变更影响，设备资源依赖

关联脉络

- 暂无明显关联 PR