

PR #43164 完整报告

vllm-project/vllm

[ci] Move 3 entrypoints tests to h200_35gb queue

合并时间: 2026-07-18 23:43

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/43164>

执行摘要

- 一句话: 迁移三个 entrypoints 测试到 H200 MIG 队列
- 推荐动作: 建议合并。变更简单、经过验证, 属于常规 CI 基础设施优化。可以快速通过, 无需深入 code review。

功能与动机

为了优化 CI 资源利用, 将 entrypoints 相关测试从 L4 GPU 迁移到 H200 35GB MIG 队列, 以利用 H200 更强的算力和内存, 提升测试稳定性和速度。同时, Qwen3 模型在 H200 上对原始 prompt 'What is 123 * 456?' 会产生大量推理输出, 消耗几乎全部 6000 token 上下文, 导致测试失败 (status=incomplete), 因此需要精简 prompt 以避免该问题。

实现拆解

1. YAML 配置迁移: 在 .buildkite/test_areas/entrypoints.yaml 中为三个步骤添加 device: h200_35gb 属性: Entrypoints Unit Tests、Entrypoints Integration (LLM)、Entrypoints Integration (Responses API)。这些步骤原本未指定设备, 默认使用 1xL4 队列。
2. 测试 prompt 修复: 修改 tests/entrypoints/openai/responses/test_parsable_context.py 中 test_basic 函数的输入 prompt, 从 "What is 123 * 456?" 改为 "What is 123 * 456? Answer with only the number.", 以控制 Qwen3 模型的输出长度, 避免在 H200 上耗尽 6000 token 的 max_model_len。
3. 分支同步: 在完成两次提交后将分支合并到最新的 main, 确保基于当前主干。

关键文件:

- .buildkite/test_areas/entrypoints.yaml (模块 CI 配置; 类别 config; 类型 configuration) : CI 配置文件, 为三个 entrypoints 测试步骤显式指定 H200 35GB MIG 设备队列, 是本次迁移的核心变更。
- tests/entrypoints/openai/responses/test_parsable_context.py (模块 测试用例; 类别 test; 类型 test-coverage) : 测试文件, 修改了 Qwen3 测试的 prompt 以避免在 H200 上耗尽上下文窗口。

关键符号: 未识别

关键源码片段

tests/entrypoints/openai/responses/test_parsable_context.py

测试文件，修改了 Qwen3 测试的 prompt 以避免在 H200 上耗尽上下文窗口。

```
@pytest.mark.asyncio
@pytest.mark.parametrize("model_name", [MODEL_NAME])
async def test_basic(client: OpenAI, model_name: str):
    response = await client.responses.create(
        model=model_name,
        # 原 prompt 为 "What is 123 * 456?", Qwen3 在 H200 上会输出大量推理内容耗尽 6000
        token
        # 改为要求仅输出数字，限制输出长度，避免 status=incomplete
        input="What is 123 * 456? Answer with only the number.",
        temperature=0.0,
    )
    assert response is not None
    print("response: ", response)
    assert response.status == "completed"
    assert response.incomplete_details is None
```

评论区精华

无实质性 review 讨论。gemini-code-assist 机器人仅给出了无意见的评论。但 PR body 中作者详细描述了验证过程：在 H200 上运行测试时，原始 prompt 导致 `max_output_tokens=5980`，`status=incomplete`；修改 prompt 后测试通过。

- CI 迁移验证 (other): 确认迁移和 prompt 修改有效，所有目标步骤在 H200 上通过。

风险与影响

- 风险：风险极低。变更仅涉及 CI 配置和测试 prompt，不改变任何生产代码逻辑。唯一可能的风险是如果其他测试依赖于同一个 YAML 步骤的默认设备，但当前变更明确只修改了三个步骤，且通过 Buildkite 流水线验证 (pipeline #78536, #78551) 确认功能正确。测试 prompt 的修改只影响一个测试用例，且验证了在 H200 上通过。
- 影响：对用户无影响。对系统的影响仅限于 CI 基础设施：三个 entrypoints 测试套件将从 L4 GPU 迁移到 H200 MIG 队列，可能会改变测试执行时间和资源分配。对团队的影响是 CI 流程的调整，需要确保 H200 MIG 队列资源充足。这个 PR 是更广泛 CI 迁移 (Issue #43024) 的一部分，预计在全部分移完成后 L4 队列负载将降低。
- 风险标记：暂无

关联脉络

- PR #43024 [CI] Move remaining 1xL4 jobs to H200 35GB MIG: 关联的父级迁移任务，本 PR 是其中的一部分，处理了三个 entrypoints 测试步骤。
- PR #43030 PR #43030 (未提供详细信息，但 PR body 提及): 延续了 #43030 开始的迁移工作。