PR #38088 完整报告

vllm-project/vllm

[ROCm][CI] Increase OpenAPI schema test timeouts

合并时间：2026-03-25 18:06

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38088

执行摘要

此 PR 在 ROCm 平台上将 OpenAPI 模式测试的超时时间增加三倍，以应对 AMD CI nightly runs 中的超时失败问题，提升测试稳定性，影响范围仅限于 ROCm CI 测试。

功能与动机

动机源于 AMD CI nightly runs 中 test_openapi_stateless schemathesis test 频繁超时，如 PR body 所述："Multiple endpoints hit the 10s default timeout, and even endpoints with the 60s timeout fail on ROCm due to infra sluggishness。" 这导致 ROCm 测试不稳定，需要调整超时设置来适应基础设施缓慢。

实现拆解

实现仅修改 tests/entrypoints/openai/test_openai_schema.py 文件，关键变更如下：

- 引入平台判断变量：_ROCM_TIMEOUT_MULTIPLIER = 3 if current_platform.is_rocm() else 1
- 调整超时常数：
  - DEFAULT_TIMEOUT_SECONDS 从 10 秒改为 10 * _ROCM_TIMEOUT_MULTIPLIER
  - LONG_TIMEOUT_SECONDS 从 60 秒改为 60 * _ROCM_TIMEOUT_MULTIPLIER

这样，在 ROCm 平台上，超时分别提升至 30 秒和 180 秒，其他平台保持不变。

评论区精华

review 中无有价值讨论；gemini-code-assist[bot] 评论仅描述变更，DarkLight1337 直接批准，未引发任何技术交锋或争议。

风险与影响

- 风险：增加超时可能掩盖实际性能回归，需监控 ROCm 平台性能变化；变更局限于测试，但平台依赖超时设置可能在未来基础设施改善后需调整。
- 影响：对用户和系统功能无直接影响；可能减少 CI 失败率，提升团队开发效率，尤其针对 ROCm 测试流水线。

关联脉络

从历史 PR 看，ROCm 平台是活跃开发领域，多个 PR 如 #37483、#36702、#37640、#37787、#37924 均涉及 ROCm 测试、性能或 bugfix，显示持续优化跨模块的 ROCm 支持。本 PR 是这一脉络的一部分，专注于测试超时调整，与其他 ROCm 相关 PR 共同推动 ROCm 生态稳定性。