PR #43971 完整报告

vllm-project/vllm

[CI] Make Model Executor test hangs fail fast with a traceback

合并时间：2026-05-30 02:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43971

执行摘要

- 一句话：CI 模型执行器测试超时失败快速反馈
- 推荐动作：应立即合入，作为 CI 防御性措施。建议后续将类似超时机制推广到其他 GPU/CUDA 密集的 CI 步骤。此 PR 逻辑清晰，改动安全。

功能与动机

在构建 68772 中，Model Executor 测试步骤因单个测试 tests/model_executor/model_loader/fastsafetensors_loader/test_fastsafetensors_loader.py::test_model_loader_download_files 在 GPUModelRunner.__init__期间挂起，阻塞夜间构建约 10 小时。原有 timeout_in_minutes: 35 未生效，因为 CUDA 调用不可中断。

实现拆解

1. 在 .buildkite/test_areas/model_executor.yaml 的 commands 中添加 export PYTHONFAULTHANDLER=1，启用 faulthandler 在测试挂起时转储所有线程的栈跟踪。
2. 在 pytest 命令中添加 --timeout=900 --timeout-method=thread，每个测试单独拥有 900s 超时看门狗，thread 方法能中断 C/CUDA 内部的挂起并输出栈跟踪。
3. 对两个 pytest 调用（model_executor 和 tensorizer_entrypoint）均应用超时设置。
4. pytest-timeout 已存在于 requirements/test/cuda.txt，无需新增依赖。

关键文件：

- .buildkite/test_areas/model_executor.yaml（模块 CI 配置；类别 config；类型 configuration）：唯一修改文件，在 pytest 命令中添加 faulthandler 和线程模式超时，防止测试挂起阻塞 CI。

关键符号：未识别

评论区精华

PR 无 review 评论，仅有 njhill 的 APPROVED。无公开讨论。

- 暂无高价值评论线程

风险与影响

- 风险：本次变更为纯 CI 配置，不涉及源码修改。风险极低：pytest-timeout 在正常路径下是空操作（no-op），不会影响测试行为；只有挂起超过 900s 才触发。可能引入的新风险：超时时间（900s）是否足够覆盖正常慢测试？但 900s=15min，远大于常规测试时长，安全。
- 影响：仅影响 CI 中 Model Executor 测试步骤。影响程度低，但价值高：将诊断时间从 ~10h 缩短至 ~15min，快速定位 GPU 初始化挂起问题，避免阻塞夜间构建。对用户无影响。
- 风险标记：配置变更，无源码修改

关联脉络

- 暂无明显关联 PR