PR #7268 完整报告

PaddlePaddle/FastDeploy

[CI] Set high-risk OOM tests for sequential execution

合并时间：2026-04-09 22:22

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7268

执行摘要

- 一句话：将 9 个高内存消耗测试标记为顺序执行，避免并发 OOM 导致的 CI 不稳定。
- 推荐动作：该 PR 值得快速浏览以了解 CI 测试执行优化策略，但无需深入代码细节。重点关注：1. 高风险测试的识别和分类逻辑；2. review 中关于硬编码和维护性的讨论，可作为未来 CI 脚本改进的参考。

功能与动机

PR body 中明确指出，在单 GPU 并行执行环境下，一些测试用例（如大批次、复杂算子或多阶段推理）内存消耗高，并发执行时可能触发 OOM 或被系统终止，导致 CI 不稳定和调试成本增加。

实现拆解

实现分为两个部分：1. 在 scripts/coverage_run.sh 的 classify_tests 函数中新增 Rule 5，硬编码 9 个高风险 OOM 测试文件路径，将其分类为 multi_gpu 类型，从而在 CI 中强制顺序执行；2. 在 scripts/unittest_requirement.txt 中添加 arctic_inference 依赖（版本 0.1.3），但此变更与 CI 优化主题关联性较弱。

关键文件：

- scripts/coverage_run.sh（模块 CI）：核心变更文件，新增 Rule 5 将 9 个高风险 OOM 测试标记为 multi_gpu 类型，实现顺序执行以避免并发内存争用。
- scripts/unittest_requirement.txt（模块 CI）：次要变更文件，添加 arctic_inference 依赖，但此变更与 CI 优化主题关联性弱，且引发版本不一致和维护性讨论。

关键符号：classify_tests

评论区精华

review 中主要讨论点：1. fastdeploy-bot 指出 arctic_inference 依赖与本次 CI 变更主题无关，建议确认是否应单独提交或移除；2. 依赖版本不一致问题，代码中提示安装 0.1.2 版本，但 PR 引入 0.1.3 版本；3. 测试文件硬编码方式维护性较差，建议通过配置文件等更灵活方式管理。ZhangYulongg 已批准 PR，但未直接回应这些建议。

- arctic_inference 依赖与 CI 变更主题无关 (design): 未在讨论中明确解决，PR 仍包含该依赖。

- 依赖版本不一致 (correctness): 未在讨论中明确解决，版本不一致可能持续存在。
- 测试文件硬编码维护性差 (design): 未在讨论中明确解决，硬编码方式被保留。

风险与影响

- 风险：技术风险包括：1. 维护性风险：scripts/coverage_run.sh 中硬编码 9 个测试文件路径，后续新增高风险测试需手动修改脚本，易遗漏或出错；2. 依赖管理风险：arctic_inference 依赖版本（0.1.3）与代码中错误提示版本（0.1.2）不一致，可能导致用户安装混淆或兼容性问题；3. 潜在回归风险：强制顺序执行可能延长 CI 总运行时间，但 PR 未评估时间影响。
- 影响：影响范围：1. 对系统：减少高内存测试并发执行时的 OOM 风险，提升 CI 稳定性和可靠性；2. 对团队：降低因 CI 失败导致的调试成本，但硬编码方式增加长期维护负担；3. 对用户：无直接影响，属于内部 CI 优化。影响程度中等，主要限于测试执行策略。
- 风险标记：硬编码维护性差，依赖版本不一致，潜在 CI 时间延长

关联脉络

- PR #7283 [CI] Add no_proxy configuration for docker execution: 同属 CI 优化类别，关注提升 CI 稳定性和网络访问可靠性。
- PR #7206 add deepe precision test: 涉及测试执行和 CI 稳定性，但聚焦于精度测试而非内存优化。
- PR #6730 [CI]【Hackathon 10th Spring No.33】config 单测补充：同属 CI 和测试改进，但侧重单元测试覆盖率提升。