PR #22301 完整报告

sgl-project/sglang

Only upload CUDA coredumps on test failure

合并时间：2026-04-08 09:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22301

执行摘要

此 PR 优化了 CI 工作流，将 CUDA coredump 上传步骤从始终执行改为仅在测试失败时执行，以减少资源浪费。变更覆盖所有主要测试套件，包括 PR 测试、多模态生成测试、重跑测试和夜间测试，总计 32 处修改。风险较低，但对调试信息可用性需保持关注。

功能与动机

为什么做？根据 PR body 描述，coredumps 仅在测试崩溃时产生，因此在成功或取消的测试中上传是无效工作，浪费 CI 资源。作者旨在通过条件优化减少不必要的上传开销，提升 CI 效率。引用 PR body 关键表述："Coredumps are only produced on crashes, so uploading on success/cancel is wasted work"。

实现拆解

做了什么？统一修改了四个 GitHub Actions 工作流文件中的 upload-cuda-coredumps 步骤条件：

- 文件列表：
  - .github/workflows/nightly-test-nvidia.yml（16 处修改）
  - .github/workflows/pr-test.yml（12 处修改）
  - .github/workflows/pr-test-multimodal-gen.yml（3 处修改）
  - .github/workflows/rerun-test.yml（1 处修改）
- 关键变更：将所有 if: always() 替换为 if: failure()，确保 coredump 仅在上传步骤失败时触发。示例代码片段： ```yaml
- uses: ./.github/actions/upload-cuda-coredumps if: failure() # 原为 if: always() ```

评论区精华

讨论了什么？review 评论为空，无实质性技术讨论。仅有两条评论：

> gemini-code-assist[bot]: "You have reached your daily quota limit..." hnyls2002: "/tag-and-rerun-ci" 这表明变更直接通过，无争议或设计权衡，团队对 CI 优化达成共识。

风险与影响

风险分析：

- 调试信息丢失风险：如果测试失败但 coredump 未正确上传（如条件未捕获所有崩溃场景），可能影响问题排查。
- 边缘情况覆盖：需确保 failure() 条件能处理所有异常退出，包括超时或部分失败。 影响分析：
- 系统影响：减少 CI 存储和带宽开销，加快成功测试运行时间。
- 团队影响：提升 CI 资源利用率，但对调试流程需稍作调整以确保 coredump 可用。
- 用户影响：无直接影响，属于后台优化。

关联脉络

与历史 PR 的关系：

- PR 22284：为多模态生成 CI 添加快速失败机制，与本 PR 共同优化测试效率。
- PR 22288 和 22297：涉及夜间测试模型更新和撤销，显示团队对 CI 配置的持续调整，本 PR 是这一趋势的延续。
- 整体趋势：近期多个 PR（如 21931、22232）聚焦 CI 和性能优化，表明团队在提升系统稳定性和资源效率方面的持续投入。