

PR #33675 完整报告

sgl-project/sglang

Stop testing cu129 DeepGEMM wheels

合并时间: 2026-08-05 16:38

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33675>

执行摘要

- 一句话: 移除 cu129 DeepGEMM 回归测试, 统一 cu130 覆盖
- 推荐动作: 该 PR 适合关心 CI 成本与发布质量权衡的读者快速阅读, 不值得深入精读。值得关注的设计决策是“用单一 cu130 矩阵统一回归覆盖”的假设是否成立, 建议后续观察 cu129 wheel 是否有回归漏报。

功能与动机

PR 描述明确说明: 让 DeepGEMM 与其余 CUDA CI 对齐, 只测试 cu130, 因为 cu130 应当能捕获相同的回归 (Align DeepGEMM with the rest of CUDA CI by testing only cu130, which should catch the same regressions)。维护者希望通过移除重复测试矩阵来降低 CI 资源消耗, 同时保持回归覆盖可信度。

实现拆解

1. 变更入口: .github/workflows/release-whl-deepgemm.yml, 这是 DeepGEMM wheel 的构建、测试与发布流水线。
2. 删除 test-cu129 作业: 原作业通过矩阵覆盖 sm90 (8-gpu-h200) 与 sm100 (8-gpu-b200) 两台 8-GPU runner, 包含检出 DeepGEMM 仓库、下载 deepgemm-wheel-cuda12.9-* 产物、安装与 cu129 匹配的 torch、执行 sgl_deep_gemm/run_tests.sh 等步骤, 本次整体移除。
3. 简化发布依赖: release-cu129 的 needs 从 [build-cu129-matrix, test-cu129] 改为 build-cu129-matrix, 构建完成后直接进入发布流程, 不再等待 cu129 回归测试。
4. 配套变更: 无源码、测试或文档改动; 该 PR 本身是纯 CI 矩阵精简。

关键文件:

- .github/workflows/release-whl-deepgemm.yml (模块 发布流水线; 类别 infra; 类型 infrastructure): 唯一变更文件: 删除 cu129 回归测试作业并简化 release-cu129 发布依赖, 直接影响 DeepGEMM wheel 的发布质量门禁。

关键符号: 未识别

关键源码片段

[.github/workflows/release-whl-deepgemm.yml](https://github.com/sgl-project/sglang/blob/main/.github/workflows/release-whl-deepgemm.yml)

唯一变更文件：删除 cu129 回归测试作业并简化 release-cu129 发布依赖，直接影响 DeepGEMM wheel 的发布质量门控。

```
# .github/workflows/release-whl-deepgemm.yml (cu129 相关部分, 修改后)
jobs:
  build-cu129-matrix:
    # 保留构建任务：产出 deepgemm-wheel-cuda12.9-* 的 whl 产物
    ...

    # test-cu129 任务已整体删除：
    # 之前会在 8-gpu-h200 与 8-gpu-b200 上安装 cu129 wheel,
    # 并执行 DeepGEMM/sgl_deep_gemm/run_tests.sh 回归测试。
    # 现在 DeepGEMM 回归覆盖统一收敛到 cu130, 与 CUDA CI 其余部分对齐。

  release-cu129:
    needs: build-cu129-matrix # 原为 needs: [build-cu129-matrix, test-cu129]
    runs-on: ubuntu-latest
    steps:
      - uses: actions/checkout@v4
      # 下载构建产物并发布 wheel 的步骤保持不变
```

评论区精华

该 PR 没有产生任何 review 评论，审核者 Fridge003 直接批准。因此没有可提炼的设计交锋；关键决策依据是 PR 描述中“cu130 能捕获相同回归”的假设。

- 暂无高价值评论线程

风险与影响

- 风险：回归覆盖缺口：cu129 wheel 不再有受控测试，若 cu130 环境无法覆盖 cu129 特有的加载或运行问题（如 libcudart 预加载、torch 版本匹配），cu129 用户可能拿到未检出的回归。发布门控放宽：release-cu129 现在只等待构建完成，构建产物损坏或深层回归问题将不再被测试阶段拦截。资源收益：节省 8-gpu-h200 与 8-gpu-b200 两个 runner 各约 120 分钟的占用，发布流水线整体更快。
- 影响：影响范围限定在 CI/ 发布流程：减少约 2 个 8-GPU runner、每个 120 分钟的测试负载，缩短 DeepGEMM cu129 wheel 的发布周期。对于依赖 cu129 wheel 的用户，发布物的回归保障有所下降；对核心 SGLang 运行时与推理逻辑无影响。
- 风险标记：cu129 回归覆盖缺失，发布门控放宽，CI 资源节约

关联脉络

- PR #33637 [CI] Skip sglang-kernel and sgl-deep-gemm reinstall on version match: 同一 CI 优化线：同样针对 sgl-deep-gemm 依赖的安装 / 测试流程做精简。
- PR #33619 [CI] Speed up dependency install: dual-ABI Rust ext cache and prevalidation pruning: 同属 CI 提速系列，目标是减少依赖安装与重复验证开销。