

PR #36672 完整报告

sgl-project/sglang

[NPU] Chain PR test jobs and disable two DeepSeek-V4-Flash perf tests

合并时间: 2026-08-28 09:16

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36672>

执行摘要

- 一句话: NPU CI 串行化并禁用两个 DSV4-Flash 性能测试
- 推荐动作: 该 PR 是基础设施调整, 价值主要在于 CI 运维优化, 对功能开发参考价值不大。可以快速浏览以了解 NPU 测试流水线的组织方式, 但无需精读代码。

功能与动机

PR body 明确说明目的是“Chain the NPU PR test jobs into a sequential functional -> accuracy -> performance pipeline and disable two buggy DeepSeek-V4-Flash nightly performance tests”。这反映了维护者对 CI 稳定性和资源利用率的关注, 通过串行化避免性能测试在前置失败时仍被触发, 并通过禁用已知有问题的测试来降低夜间管线的失败率。

实现拆解

1. 修改 [.github/workflows/pr-test-npu.yml](#):
 - 为 base-c-test-acc-{2,4,8,16}-npu-a3 添加 needs 依赖, 包括所有功能作业 (base-a-test-1-npu-a2 和 base-b-test-{1,2,4,8,16}-npu-a3)。
 - 为 base-c-test-perf-{2,4,8,16}-npu-a3 添加 needs 依赖, 包括所有功能作业和所有准确率作业。
 - 这样形成 functional -> accuracy -> performance 的串行链。
2. 修改测试文件: 在两个 DSV4-Flash 性能测试文件中, 将 register_npu_ci 调用改为多行形式并添加 disabled="Testcase bug, temporarily disabled" 参数, 使其从夜间管线中排除。
3. 无测试 / 配置配套: 本次变更仅涉及 CI 配置和测试注册, 无源码逻辑改动。

关键文件:

- [.github/workflows/pr-test-npu.yml](#) (模块 CI 工作流; 类别 infra; 类型 infrastructure): 核心变更: 通过修改 needs 依赖实现 NPU PR 测试阶段串行化, 影响整个 CI 执行流程。
- [test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in32k_out1k_50ms.py](#) (模块 性能测试; 类别 test; 类型 test-coverage): 为 register_npu_ci 添加 disabled 参数, 从夜间管线中禁用有 bug 的测试。
- [test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in8k_out1k_50ms.py](#) (模块 性能测试; 类别 test; 类型 test-coverage): 类似地禁用另一个有 bug 的性能测试, 与上一个文件对称。

关键符号: register_npu_ci

关键源码片段

`test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in32k_out1k_50ms.py`

为 register_npu_ci 添加 disabled 参数，从夜间管线中禁用有 bug 的测试。

```
# 禁用有 bug 的性能测试，避免夜间管线失败
register_npu_ci(
    est_time=1800,
    suite="nightly-perf-16-npu-a3",
    nightly=True,
    disabled="Testcase bug, temporarily disabled", # 临时禁用，等待修复后移除
)
```

`test/registered/npu/performance/deepseek_v4_flash/test_npu_deepseek_v4_flash_w8a8_8p_in8k_out1k_50ms.py`

类似地禁用另一个有 bug 的性能测试，与上一个文件对称。

```
# 同样禁用有 bug 的性能测试
register_npu_ci(
    est_time=1800,
    suite="nightly-perf-16-npu-a3",
    nightly=True,
    disabled="Testcase bug, temporarily disabled", # 临时禁用，避免误报
)
```

评论区精华

该 PR 只有一个由 `sclang-npu-bot` 触发的 `/tag-and-rerun-ci` 命令评论，无实质性讨论。审核由 `sclang-npu-bot` 自动批准，无人工评论。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险在于串行化可能会增加整体 CI 运行时间，因为性能作业需要等待所有前置作业完成，可能导致流水线总时长变长。此外，禁用两个性能测试会暂时降低对 DSV4-Flash 性能回归的覆盖，若 bug 未修复，回归风险被隐藏。但这两个测试本身被标记为有 bug，禁用是合理的权衡。
- 影响：影响范围限于 NPU 相关的 CI 配置和测试注册，不涉及任何产品代码、模型推理逻辑或用户功能，对最终用户无影响。对团队而言，提高了 NPU CI 的可靠性和资源利用率，但可能增加单次 PR 验证时长。
- 风险标记：CI 串行化可能增加总时长，禁用测试可能隐藏回归风险

关联脉络

- PR #36736 [AMD][CI] Merge the four MI35x DeepSeek-V3.2 nightly jobs into two to save runtime: 同为调整 CI 作业以节省资源，本 PR 可能是 NPU 侧的类似优化。
- PR #36726 [Diffusion] Fix the five unit tests failing on main: 涉及禁用 / 修复导致 CI 失败的测试，与本 PR 的禁用测试目标相似。