

PR #36092 完整报告

sgl-project/sglang

Npu single node test timeout config

合并时间: 2026-08-24 14:38

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36092>

执行摘要

- 一句话: NPU 单节点测试超时改为可配置, 流水线显式传 180 分钟
- 推荐动作: 这是一个简单且聚焦的 CI 配置改进, 值得快速浏览以了解其输入传递与覆写模式; 如果有 NPU 测试运行时间数据, 可以进一步验证 180 分钟是否合理, 并考虑将默认值也收敛到 180 以减少配置分歧。

功能与动机

PR 正文仅说明为 NPU 单节点测试增加超时配置。从提交演进看, 默认保持 300 分钟、夜间和 PR 流水线显式覆盖为 180 分钟, 目的是让 NPU 作业更快结束, 提高 CI 资源利用率, 同时保留默认安全值。

实现拆解

实现分为四步:

1. 在 `.github/workflows/_npu-single-node-test-stage.yml` 中新增 `test_timeout_minutes` 输入 (`required: false`, 默认 '300', 描述为 Run test 超时时间), 并将 “Run test” 步骤的 `timeout-minutes` 由硬编码 300 改为 `${{ fromJson(inputs.test_timeout_minutes) }}`, 使超时时间可被调用方覆写。
2. 在 `.github/workflows/nightly-test-npu.yml` 的 `nightly-acc-2-npu-a3` 作业中传入 `test_timeout_minutes: '180'`, 将夜间精度测试的超时从 300 缩短为 180 分钟。
3. 在 `.github/workflows/pr-test-npu.yml` 的 `base-c` 作业中同样传入 `test_timeout_minutes: '180'`, 保证 PR 测试也使用 180 分钟超时。
4. 没有新增源码测试或文档配套; 提交历史显示默认值经历了 300 → 180 → 300 (默认) + 180 (覆盖) 的迭代, 最终采用“默认 300 + 显式覆盖 180”的安全策略。

关键文件:

- `.github/workflows/_npu-single-node-test-stage.yml` (模块 CI 流水线; 类别 infra; 类型 infrastructure; 符号 `test_timeout_minutes`, Run test): 核心改动文件: 新增可配置的 `test_timeout_minutes` 输入, 并将 Run test 步骤的超时改为动态值。
- `.github/workflows/nightly-test-npu.yml` (模块 CI 流水线; 类别 infra; 类型 infrastructure; 符号 `test_timeout_minutes`): 夜间流水线显式传入 180 分钟超时, 缩短夜间精度测试的等待时间。

- [.github/workflows/pr-test-npu.yml](#) (模块 CI 流水线; 类别 infra; 类型 infrastructure; 符号 test_timeout_minutes) : PR 流水线 base-c 作业也改为 180 分钟超时, 保证 PR 测试与夜间测试行为一致。

关键符号: 未识别

关键源码片段

[.github/workflows/_npu-single-node-test-stage.yml](#)

核心改动文件: 新增可配置的 test_timeout_minutes 输入, 并将 Run test 步骤的超时改为动态值。

```
# 新增可覆写的超时输入: 默认 300 分钟, 调用方可按作业类型覆盖
test_timeout_minutes:
  required: false
  type: string
  default: '300'
  description: 'timeout-minutes for the Run test step'

# Run test 步骤不再硬编码 300, 改用输入值
- name: Run test
  # fromJson 将字符串参数转换为数值, 供 timeout-minutes 使用
  timeout-minutes: ${{ fromJson(inputs.test_timeout_minutes) }}
  env:
    SGLANG_USE_MODELSCOPE: true
    HF_ENDPOINT: https://hf-mirror.com
```

评论区精华

无人工 review 评论, sglang-npu-bot 自动 APPROVED。PR 内仅 CI 触发指令 ([/tag-and-rerun-ci](#)、[/rerun-failed-ci](#)), 没有关于设计权衡的实质讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 主要风险集中在 CI 行为变化: 若 NPU 测试实际运行时间超过 180 分钟, 会造成误报失败并增加重跑成本; fromJson 解析要求输入必须是合法整数, 否则 workflow 步骤会报错; 若调用方忘记传递新输入, 会静默回退到默认值 300, 导致不同作业间超时行为不一致。这些风险都限定在 NPU CI 工作流内部, 不影响产品代码。
- 影响: 影响范围限于 NPU CI 流水线, 包括夜间测试和 PR 测试作业; 单作业最长等待时间从 300 分钟降到 180 分钟, 加快失败反馈, 但可能增加重跑次数。对最终用户推理性能、模型输出无任何影响, 对团队而言主要是维护 CI 配置时需注意输入传递的完整性。
- 风险标记: CI 超时变更, 误报失败风险, 输入传递不一致

关联脉络

- 暂无明显关联 PR