

PR #36242 完整报告

sgl-project/sglang

[CI] Register test_dflash_logits at its real cost

合并时间: 2026-08-25 09:22

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36242>

执行摘要

- 一句话: 修正 test_dflash_logits.py 的 CI 预估耗时从 1s 至 35s
- 推荐动作: 该 PR 为 CI 基础设施微调, 值得快速浏览以了解分片预估机制; 推荐作为 CI 配置维护的参考。

功能与动机

PR 描述指出 test_dflash_logits.py 注册的 est_time=1, 实际在 base-a-test-cpu 上耗时 32 秒, 是 CPU 测试套件中耗时与预估比最差的 (32 倍低估)。分片分配器在缺少实时统计时会使用该数字, 导致过度打包分片并引发分片超时。该问题是在调查 run 32786070189 的分片 7 超时时发现的。

实现拆解

1. 修改预估耗时: 在 test/registered/unit/spec/test_dflash_logits.py 中将 register_cpu_ci(est_time=1, suite="base-a-test-cpu") 改为 est_time=35, 这是唯一的代码变更。
2. 分析耗时根因: 耗时主要来自 CPU 上 cold Inductor 编译 (涉及 sglang/srt/models/dflash.py 中的 _grouped_conv、_score_edges、_follow_maps 三个 @torch.compile 函数)。量化测试显示 cold compile 约为 6.51s, 两次 cold compile 加上解释器导入开销可复现 32s。
3. 评估并否决优化方案: 尝试了延迟导入 triton kernel 和缩小 block-size 循环, 但均无收益; 保留编译路径的测试覆盖是合理的, 因此选择修正预估耗时而非改动测试。

关键文件:

- test/registered/unit/spec/test_dflash_logits.py (模块 测试注册; 类别 test; 类型 test-coverage): 唯一的变更文件, 修正 CI 预估耗时, 直接影响分片分配

关键符号: 未识别

关键源码片段

test/registered/unit/spec/test_dflash_logits.py

唯一的变更文件, 修正 CI 预估耗时, 直接影响分片分配

```
# test/registered/unit/spec/test_dflash_logits.py
```

```
import sys
from types import SimpleNamespace

import pytest
import torch

from sglang.srt.models.dflash import (
    CandidateSelector,
    DFlash2DraftModel,
    _grouped_conv,
)
from sglang.srt.speculative.dflash_utils import parse_dflash_draft_config
from sglang.test.ci.ci_register import register_cpu_ci

# 修正预估耗时：该测试在 CPU 上需约 32s（主要来自 cold Inductor 编译），
# 之前注册为 1s 导致分片分配器低估其开销，容易引发分片超时。
# 改动仅涉及 est_time，不改变测试逻辑。
register_cpu_ci(est_time=35, suite="base-a-test-cpu")
```

评论区精华

无 review 评论或讨论线程；PR 通过 CI 重跑请求验证测试仍通过。

- 暂无高价值评论线程

风险与影响

- 风险：无代码逻辑变更，风险极低；但预估耗时 35s 可能仍低估或高估其他硬件 / 环境下的实际耗时，可能影响分片平衡的精确性。
- 影响：该 PR 影响 CI 分片分配，使 test_dflash_logits.py 被分配足够时间，减少分片超时风险；对用户功能和运行时性能无影响。
- 风险标记：CI 基础设施调优，预估耗时可能随环境变化

关联脉络

- 暂无明显关联 PR