

PR #30125 完整报告

sgl-project/sglang

[MLX] Fix FakeOverlapScheduler test stub broken by forward_ct accounting

合并时间: 2026-07-07 04:43

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30125>

执行摘要

- 一句话: 修复 MLX 测试桩因 forward_ct 缺失而崩溃
- 推荐动作: 值得阅读以了解测试桩的维护模式, 特别是当 mixin 基类新增属性时, 需要同步更新所有子类桩。

功能与动机

PR #29217 引入的 `_finalize_mlx_pending_job` 会更新 `self.forward_ct` 并调用 `self.profiler_manager._profile_batch_predicate(...)`, 但测试文件 `test_attention_patching.py` 中的 `FakeOverlapScheduler` 桩未定义这两个属性, 导致 `test_finalize_pending_job_updates_scheduler_last_batch` 测试失败。该问题在 #29217 合并时未被发现 (当时没有 macOS MLX CI 运行该测试套件), 直到 #29691 合入后 CI 变为红色。

实现拆解

1. 在 `test/registered/unit/hardware_backend/mlx/test_attention_patching.py` 的 `FakeOverlapScheduler.__init__` 中新增两行:
 - `self.forward_ct = 0`: 初始化前向计数, 满足 mixin 的递增要求。
 - `self.profiler_manager = SimpleNamespace(_profile_batch_predicate=lambda batch: None)`: 使用文件中已使用的 `SimpleNamespace` 惯用法, 提供一个无操作的 profiler 谓词。
2. 该变更完全在测试代码内, 无生产代码修改。

关键文件:

- `test/registered/unit/hardware_backend/mlx/test_attention_patching.py` (模块 MLX 测试; 类别 test; 类型 test-coverage; 符号 `FakeOverlapScheduler.init`): 唯一变更文件: 在 `FakeOverlapScheduler.__init__` 中新增 `forward_ct` 和 `profiler_manager` 桩属性, 以匹配 mixin 新增的接口。

关键符号: `FakeOverlapScheduler.init`

关键源码片段

`test/registered/unit/hardware_backend/mlx/test_attention_patching.py`

唯一变更文件：在 FakeOverlapScheduler.__init__ 中新增 forward_ct 和 profiler_manager 桩属性，以匹配 mixin 新增的接口。

```
class FakeOverlapScheduler(SchedulerMlxOverlapMixin):
    def __init__(self, next_token_ids):
        self.tp_worker = FakeTpWorker(next_token_ids)
        self.last_batch = None
        self.processed_batch = None
        self.processed_result = None
        # 自从 #29217 后，_finalize_mlx_pending_job 会递增 forward_ct
        # 并调用 profiler 批次谓词（镜像 run_batch）；
        # 为这两个属性提供桩值，以便重叠记账逻辑能正常执行。
        self.forward_ct = 0
        self.profiler_manager = SimpleNamespace(
            _profile_batch_predicate=lambda batch: None
        )

    def process_batch_result(self, batch, result):
        self.processed_batch = batch
        self.processed_result = result
```

评论区精华

无实质讨论，仅 author 提及这是对 #29217 的跟进修复，并确认该修复与 CI 配置 PR #30121 无冲突。LarrySimingDeng 验证后表示 LGTM。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改测试桩，不涉及任何生产代码路径。不会引入回归或性能问题。
- 影响：影响范围仅限 MLX 硬件后端单元测试。修复后 test_attention_patching.py 的 39 个测试全部通过，MLX CI 通道恢复绿色。无用户或系统影响。
- 风险标记：暂无

关联脉络

- PR #29217 Add overlap scheduler forward_ct accounting and profiler predicate: 该 PR 引入了 forward_ct 和 profiler 调用来重叠调度，是本 PR 修复的直接原因。
- PR #29691 Enable mlx-unit-test CI lane: 该 PR 启用了 MLX 单元测试 CI 通道，使得本修复变得必要，因为测试桩的缺失在此之前未被发现。
- PR #30121 MLX CI plumbing pr-test-mlx.yml: 作者提及该 PR 与 #30121 独立不重叠，但 #30121 是 MLX CI 管道配置的一部分。