

PR #36778 完整报告

sgl-project/sglang

fix: report the real backend for non-CUDA CI registrations in /rerun-test

合并时间: 2026-08-29 04:46

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36778>

执行摘要

- 一句话: 修复 /rerun-test 对非 CUDA 后端注册的误报
- 推荐动作: 值得略读, 属于低风险修复, 但展示了 CI 工具链中注册信息与调度逻辑的耦合。
关注点: 1) 如何通过共享数据结构 (REGISTER_MAPPING) 来避免硬编码后端列表; 2) 考虑将 AMD 等工作流纳入 /rerun-test 的调度范围, 以彻底消除此类误导。

功能与动机

PR body 指出, 对于通过 `register_amd_ci` 等注册的测试文件, /rerun-test 会错误地回复 “No register_cuda_ci(...) or register_cpu_ci() found”, 但这与事实不符。实测 `test/registered/` 目录下 354 个文件受到此误导, 容易让用户误入歧途, 以为需要修复注册, 而实际应使用 `/rerun-failed-ci`。

实现拆解

实现分为三个主要步骤:

1. 将 `_extract_legacy_suites` 重构为通用的 `_extract_suites(content, register_fn)`, 使其能提取任意注册函数的 suite 值, 同时保持对 `register_cuda_ci` 的兼容。
2. 新增 `_extract_other_backends(content)` 函数, 用于识别 `register_amd_ci`、`register_npu_ci`、`register_xpu_ci`、`register_musa_ci`、`register_mlx_ci` 等非 CUDA/CPU 注册, 并返回后端标签和对应的 suite 列表。
3. 在 `detect_suite` 中, 在原有 CUDA/CPU/legacy suite 检查之后, 新增对其他后端注册的检测分支。若检测到这类注册, 则返回一条新的错误信息, 明确指出该文件注册的后端、suite (若有), 并建议使用 `/rerun-failed-ci` 或手动调度对应工作流。该变更仅影响错误提示, 不改变任何调度逻辑。新增代码位于 `scripts/ci/utils/slash_command_handler.py` 中, 涉及符号 `_extract_suites`、`_extract_other_backends` 和 `detect_suite`。

关键文件:

- `scripts/ci/utils/slash_command_handler.py` (模块 CI 脚本; 类别 infra; 类型 infrastructure; 符号 `_extract_legacy_suites`, `_extract_suites`, `_extract_other_backends`, `detect_suite`): 唯一的变更文件, 包含核心逻辑修改: 新增 `_extract_other_backends` 函数, 重构 `_extract_suites`, 并在 `detect_suite` 中增加对非 CUDA/CPU 后端的检测分支。

关键符号: `_extract_suites`, `_extract_other_backends`, `detect_suite`

关键源码片段

`scripts/ci/utils/slash_command_handler.py`

唯一的变更文件，包含核心逻辑修改：新增 `_extract_other_backends` 函数，重构 `_extract_suites`，并在 `detect_suite` 中增加对非 CUDA/CPU 后端的检测分支。

```
# scripts/ci/utils/slash_command_handler.py
# 将原 _extract_legacy_suites 泛化为对任意注册函数提取 suite

def _extract_suites(content, register_fn):
    """Pull every single-string `suite=` from `
```

评论区精华

Codex 机器人提出了两点 P2 级别建议：

1. 避免承诺 `/rerun-failed-ci` 能重跑跳过 (skipped) 的文件，因为 `handle_rerun_failed_ci` 仅处理失败或跳过的 workflow，对于成功或从未运行的 workflow 无效，可能导致 `/rerun-group amd` 等命令无法重跑被跳过的文件。
 2. 过滤空的 suite 名称，因为部分 NPU 文件使用 `suite=""`，会导致响应中出现空的 suite 占位符，并推荐一个没有搜索模式的 `grep` 命令。这两点建议均未被作者明确回应，也未在最终合并版本中解决。
- 对 `/rerun-failed-ci` 的承诺是否过度 (design): 作者未回应，最终合并版本仍保留该建议，存在潜在误导。
 - 过滤空 suite 名称 (correctness): 作者未回应，最终合并版本未处理空 suite 值。

风险与影响

- 风险：主要风险在于新错误信息可能误导用户，尤其是当后端 workflow 从未运行或成功后，`/rerun-failed-ci` 无法重跑这些文件。此外，若 suite 为空字符串，新逻辑会生成不准确的提示信息。这些风险仅涉及 CI 提示文本，不会影响运行时行为，风险较低。
- 影响：影响范围限于 CI 斜杠命令 `/rerun-test` 的错误提示信息。受益方包括使用 AMD/NPU/XPU/MUSA/MLX 后端注册测试的开发者，他们现在能获得准确的重跑指引。团队维护成本增加，但逻辑保持简单。变更不会影响实际测试调度或任何运行时功能。
- 风险标记：对 `/rerun-failed-ci` 的承诺可能误导，空 suite 值未处理

关联脉络

- PR #36914 [Fix] Lazy-import aiter in DSv4 paged_decode to unbreak CPU CI: 同样涉及 CI 基础设施修复，聚焦非 CUDA 后端的 CI 稳定性。
- PR #36638 Fix KeyError on batch requests whose state is freed before it is read: 同为针对 CI/ 测试工具链的修复，但方向不同。