

PR #29066 完整报告

sgl-project/sglang

[CI] Migrate JIT tests to runner config registration

合并时间: 2026-06-30 10:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29066>

执行摘要

- 一句话: 迁移 JIT 测试 CI 注册到 stage/runner_config 模型
- 推荐动作: 本 PR 值得 CI 维护者精读, 特别是 run_suite.py 中的命名约定和 skills 文档的示例。设计决策上, 将 suite 拆分为 stage 和 runner_config 提高了灵活性, 但命名中嵌入硬件规格 (如 1-gpu-large、4-gpu-b200) 可能造成耦合, 后续可考虑进一步抽象。

功能与动机

为了统一 CI 测试的注册方式, 让 /rerun-test 命令能够直接解析 runner 细节, 移除遗留的 suite-to-runner 映射。PR body 指出 'Remove the /rerun-test legacy suite-to-runner mapping so CUDA dispatch resolves runner details directly from runner_config.'

实现拆解

1. 修改所有 JIT 测试文件的注册调用: 将 register_cuda_ci 和 register_amd_ci 的 suite= 参数替换为 stage= 和 runner_config= 两个参数。涉及约 120 个测试文件, 如 test/registered/jit/benchmark/bench_clamp_position.py 等。
2. 更新 test/run_suite.py 中 CUDA 套件名称列表: 将 kernel 相关套件名称从 base-b-kernel-unit-1-gpu-large 等改为 base-b-kernel-unit-test-1-gpu-large 等, 自动添加 -test- 中缀以匹配生成的套件名称格式。
3. 更新 .claude/skills/add-jit-kernel/SKILL.md 和 .claude/skills/write-sglang-test/SKILL.md 中的示例, 展示新的注册语法, 并修正已变更的套件名称。
4. 保留 nightly 套件的遗留 suite= 注册形式, 因为它们不需要 /rerun-test 调度, 并在 _LEGACY_SUITE_TO_RUNNER_CONFIG 中恢复部分映射以支持 nightly/weekly 的 rerun 调度。
5. (配套) 无其他核心代码变更; 所有变更均限于测试注册和 CI 配置层面。

关键文件:

- test/run_suite.py (模块 测试调度; 类别 test; 类型 test-coverage): 集中定义了 CI 套件名称列表, 本 PR 更新了 CUDA kernel 相关套件名称以匹配新格式, 是 CI 调度的核心配置。
- test/registered/jit/benchmark/bench_clamp_position.py (模块 JIT 基准; 类别 test; 类型 test-coverage): 展示了 JIT 测试注册调用从 suite 到 stage/runner_config 的典型迁移。

移模式，是批改的样板文件之一。

- `.claude/skills/add-jit-kernel/SKILL.md` (模块 技能文档; 类别 docs; 类型 documentation)
: Review 中讨论了此文件的示例准确性, 需更新以反映新注册语法。

关键符号: 未识别

关键源码片段

`test/run_suite.py`

集中定义了 CI 套件名称列表, 本 PR 更新了 CUDA kernel 相关套件名称以匹配新格式, 是 CI 调度的核心配置。

```
# run_suite.py - 更新后的 CUDA 套件名称列表 (片段)
# 变更: kernel 套件名称增加了 -test- 中缀
CUDA_SUITES = [
    'base-a-test-1-gpu-small',
    'base-b-test-1-gpu-small',
    'base-b-test-1-gpu-large',
    'base-b-test-2-gpu-large',
    'base-b-test-4-gpu-b200',
    # 以下四个名称发生了变更:
    'base-b-kernel-unit-test-1-gpu-large', # 原 base-b-kernel-unit-1-gpu-large
    'base-b-kernel-unit-test-4-gpu-b200', # 原 base-b-kernel-unit-1-gpu-b200
    'base-b-kernel-unit-test-8-gpu-h200', # 原 base-b-kernel-unit-8-gpu-h200
    'base-b-kernel-benchmark-test-1-gpu-large', # 原 base-b-kernel-benchmark-1-gpu-large
    # ... 其余未变更
]
```

`test/registered/jit/benchmark/bench_clamp_position.py`

展示了 JIT 测试注册调用从 suite 到 stage/runner_config 的典型迁移模式, 是批改的样板文件之一。

```
# bench_clamp_position.py - 注册调用迁移
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

# 迁移前: register_cuda_ci(est_time=13, suite='base-b-kernel-benchmark-1-gpu-large')
# 迁移后: 使用 stage 和 runner_config 参数
register_cuda_ci(
    est_time=13, stage='base-b-kernel-benchmark', runner_config='1-gpu-large'
)
register_amd_ci(est_time=16, stage='jit-kernel-unit', runner_config='amd')
```

评论区精华

在 Review 中, hnyls2002 提出两个问题:

- 技能文档中的 `base-b-kernel-unit-test-4-gpu-b200` 可能过时, 因为历史上 `1-gpu-b200` 的 runner 配置实际上使用了 `4-gpu-b200` 的 runner。

- 建议将 stage 名称与 runner 配置解耦，不在注册语句中嵌入硬件规格（如 large、2-gpu）。kpha-sgl 回应称 nightly 迁移将在另一个 PR 进行，runner 配置问题需后续确认。
- b200 runner 配置一致性 (question): kpham-sgl 表示不确定，并指出 runner_configs.yml 中可能只有 4-gpu-b200，需后续确认。
- stage 名称与 runner 配置解耦 (design): kpham-sgl 同意该方向，但表示 nightly 迁移将在另一个 PR 进行。

风险与影响

- 风险：主要风险在于套件名称变更可能导致 CI 工作流无法匹配正确的 runner，造成测试被跳过或失败。具体地，b200 的 runner 配置映射可能不正确（历史上 1-gpu-b200 实际上使用 4-gpu-b200 的 runner），需要验证 runner_configs.yml 中的配置。此外，skills 文档中的示例若未完全同步可能误导开发者。需确保 `_LEGACY_SUITE_TO_RUNNER_CONFIG` 覆盖了所有必要的 nightly 套件。
- 影响：对用户无直接影响。对开发团队，JIT 测试的 CI 注册方式更新，需要在新测试文件中使用 stage 和 runner_config 参数；旧的 suite 参数不再被支持（pr-test 工作流）。`/rerun-test` 命令现在能直接根据新注册信息调度测试，提高了调试效率。套件名称的标准化有助于 CI 趋势监控。
- 风险标记：套件名称可能不匹配，runner 配置映射不一致，skills 文档示例可能过时

关联脉络

- 暂无明显关联 PR