PR #24725 完整报告

sgl-project/sglang

ci: tag-gated nightly migration — foundation + 40 whole-file moves

合并时间：2026-05-15 07:11

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24725

执行摘要

- 一句话：标签驱动 CI 夜间迁移，减少 per-commit 负担约 38.9%
- 推荐动作：本 PR 值得详细阅读：其展示了大型 CI 重构的策略——从基础设施抽象、文件级迁移到工作流重组，并包含了设计权衡（如放弃 tag-gated 改用 extra stages）。对于需要优化 CI 效能的团队具有较强的参考价值。关注点应放在标签映射的准确性以及 extra 工作流触发门槛对开发者体验的影响。

功能与动机

降低每次 commit 的 CI 负担，从 1232 分钟降至 753 分钟（该 PR 贡献 -305 分钟，结合 PR1 共 -38.9%）。通过标签路由让高频变更区域获得更多测试覆盖，而低频区域仅运行 baseline 测试。

实现拆解

1. 基础设施扩展：在 python/sglang/test/ci/ci_register.py 的 CIRegister 类中添加 tags 元组参数，AST 验证器接受元组 / 列表字面量。test/run_suite.py 新增 --include-tags 命令行参数、PER_COMMIT_TO_NIGHTLY 映射表，以及 nightly-1-gpu-5090 等 nightly suite 注册。
2. 工作流改造：.github/workflows/pr-test.yml 根据事件类型解析 INCLUDE_TAGS_FLAG（PR 用标签并集，schedule 用 *），并注入所有 run_suite.py 调用。实现 stage-c 到 stage-b 的层级折叠，删除 wait-for-stage-b，所有 stage-c 任务改为等待 stage-a 完成后并行执行。添加 baseline-only 标签支持，使大 PR 可以选择跳过额外测试。
3. 标签体系引入：.github/labeler.yml 新增 8 个标签（attention-backend、moe、rl、scoring、session、perf、scheduler、model-coverage），复用仓库已有标签如 Multi-modal、blackwell、deepseek、lora、quant、speculative-decoding、hicache。PR 标签与测试标签交集决定条件 CI 的拉入范围。
4. 测试文件迁移：将 53 个测试文件整文件从 per-commit 移至 nightly suite（如 stage-b-test-1-gpu-large → nightly-1-gpu），并更新 register_cuda_ci 调用。对 6 个测试文件进行类级别拆分，per-commit 保留核心变体，其他变体进入 nightly 并携带标签。
5. 公共代码抽离：创建 python/sglang/test/server_fixtures/ 下的多个 fixture 文件（streaming_session_fixture.py、standalone_fixture.py、pcg_spec_fixture.py、hybrid_attn_backend_fixture.py），将服务器启动和辅助函数从测试类中提取出来；同时

创建 python/sglang/test/kits/ 下的 kit 文件（streaming_session_kit.py）包含可复用的测试方法。最终测试文件仅通过继承这些 base 类并覆写类属性来定义变体，大幅减少重复代码。

6. 后续演化：合并前对方案进行了调整，最终放弃初始的 tag-gated 设计，改为在合并过程中引入 pr-test-extra.yml 独立工作流，由 run-ci-extra 标签触发。这简化了条件逻辑，但仍保留了基于标签的选择能力。

关键文件：

- python/sglang/test/ci/ci_register.py（模块 CI 注册表；类别 test；类型 core-logic；符号 CIRegistry, register_cuda_ci, register_amd_ci, register_cpu_ci）：条件 CI 的核心基础设施变更：为 CIRegistry 添加 tags 参数、AST 验证器支持元组 / 列表、运行时存根接收 tags。
- test/run_suite.py（模块 CI 运行器；类别 test；类型 core-logic；符号 PER_COMMIT_TO_NIGHTLY, NIGHTLY_SUITES, main）：实现 --include-tags 参数和 PER_COMMIT_TO_NIGHTLY 映射，决定条件 CI 拉入哪些 nightly 测试。
- .github/workflows/pr-test.yml（模块 CI 工作流；类别 infra；类型 core-logic；符号 INCLUDE_TAGS_FLAG, baseline-only, stage-c collapse）：主 CI 工作流：添加 INCLUDE_TAGS_FLAG 解析、stage-c 到 stage-b 的坍塌、baseline-only 标签支持、动态分区大小。
- python/sglang/test/server_fixtures/streaming_session_fixture.py（模块 测试 Fixture；类别 test；类型 test-coverage；符号 StreamingSessionServerBase, _abort_repro_generate, _concurrent_logprob_run, _stress_run_all）：新增的 streaming session 测试 fixture，集中管理服务器启动和公共辅助函数，被多个测试文件引用。
- test/registered/sessions/test_streaming_session.py（模块 流式会话测试；类别 test；类型 test-coverage；符号 TestStreamingSessionEagleV2RetractLargePage, TestStreamingSessionAbortLeakRepro）：展示 split 模式：per-commit 文件从 1092 行缩减至 82 行，仅保留 EagleV2RetractLargePage 和 AbortLeakRepro，其余移至 nightly 文件。
- test/registered/piecewise_cuda_graph/test_pcg_with_speculative_decoding.py（模块 PCG 推测解码测试；类别 test；类型 test-coverage；符号 TestPCGWithEAGLE3, PCGSpecBase）：展示类级别 split：原含 4 个变体（MTP/EAGLE3/STANDALONE/NGRAM），per-commit 仅保留 EAGLE3，其余移至 extra 文件。

关键符号：register_cuda_ci, register_amd_ci, register_cpu_ci, CIRegistry, main (run_suite.py), StreamingSessionServerBase.setUpClass, StreamingSessionKitMixin.test_kv_cache_inheritance

评论区精华

该 PR 的 review 评论数为 0，但通过提交历史可以还原几项关键设计决策：

- Score API 利益相关方要求在 per-commit 保留 5 个 Score API 测试（commit 88cdbd18），体现了对生产 API 回归的重视。
- 初始 tag-gated 方案在合并阶段被替换为 explicit extra stages + run-ci-extra 标签的工作流（commit 503219ea），主要原因是条件 CI 复杂度较高且需要维护标签到 suite 的映射。
- 多次合并 main 分支以解决冲突，表明此 PR 跨度长、与同期 CI 变更交互频繁。
- Score API 测试从 nightly 移回 per-commit (design): 接受反馈，将 5 个 Score API 测试移回 per-commit（commit 88cdbd18）。
- 从 tag-gated 设计改为 extra workflow (design): 采用 extra workflow 方案，简化条件逻辑。extra 测试仅在 PR 同时带有 run-ci 和 run-ci-extra 标签时运行。

风险与影响

- 风险：
  1. 标签匹配风险：条件 CI 依赖 PR 标签与测试标签的交集，若标签遗漏或 glob 不准确，可能导致本该运行的测试被跳过，产生回归。.github/labeler.yml 中的路径模式需持续维护。
  2. extra 工作流门槛：pr-test-extra.yml 要求同时存在 run-ci 和 run-ci-extra 标签才能触发，可能因用户不熟悉或自动标注失败而漏跑。
  3. per-commit 测试缩减：部分测试（如 speculative decoding 的非 V2 版本）移至 nightly/extra，开发者若不主动带标签触发，可能直到合并前才发现失败。
  4. 新增 fixture/kit 的兼容性：大量测试类改为继承 fixture 基类，若基类有默认值错误或环境变量处理不当，会影响所有继承类。 - 影响：开发者体验：per-commit CI 提速明显（约 -4 小时），日常开发迭代更流畅。但需要理解标签机制，在涉及修改时添加对应标签以获得充分测试。测试覆盖：nightly/extra 阶段仍会运行全部测试，但仅在调度触发或带有 run-ci-extra 标签时。对于低变更频率模块，回归发现会延迟到合入前或 nightly。CI 维护团队：工作流配置更模块化，pr-test.yml 和 pr-test-extra.yml 分离，便于独立维护。标签定义和映射表需要持续更新。 - 风险标记：标签匹配遗漏风险，extra 工作流触发门槛，per-commit 测试覆盖缩减，频繁合并 main 导致冲突

关联脉络

- PR #24721 CI pruning effort PR1: 25 whole-file removals + 13 testcase archives: 同一 CI 精简计划的第一部分，为本 PR 的迁移奠定基础。
- PR #24731 Add nightly-test-general-1-gpu-5090 workflow job: 需要的配套 PR，提供 nightly-1-gpu-5090 套件，本 PR 中的 13 个 SM12.0 测试依赖它。合并顺序要求先合并此 PR。
- PR #25248 Support new-style register_cuda_ci(stage=, runner_config=) in slash handler + est-time updater: 在合并过程中，CI 脚本工具需要适配新式注册调用，该 PR 修复了相关问题。