PR #25329 完整报告

sgl-project/sglang

Skip CI tests added in #24816 (broken on main)

合并时间：2026-05-15 09:12

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25329

执行摘要

- 一句话：临时跳过 CI 中断的测试
- 推荐动作：该 PR 是临时性 CI 维护变更，技术含量低。建议合并以保持 CI 绿色，但应尽快跟进根本原因修复。

功能与动机

PR #24816 添加的 FlashInfer SM90 cutlass MXFP4 单元测试和 DSv4-Flash FP4 FlashInferCutlass 端到端测试目前在 main 分支上运行失败，导致 CI 不能通过。作者希望暂时跳过这些测试，保持 CI 绿色，同时不影响已有的测试覆盖。

实现拆解

1. 禁用整个 MXFP4 单元测试文件：在 test/registered/unit/layers/quantization/test_mxfp4_sm90_cutlass.py 中，将 register_cuda_ci(est_time=120, stage="stage-b", runner_config="1-gpu-large") 改为 register_cuda_ci(est_time=120, stage="stage-b", runner_config="1-gpu-large", disabled="broken on main, see #24816")。该参数是 SGLang 标准约定，使该文件不再被 CI 调度。
2. 禁用单个测试类：在 test/registered/dsv4/test_deepseek_v4_flash_fp4_h200.py 中，在 TestDSV4FlashFP4H200FlashInferCutlass 类前添加 @unittest.skip("broken on main, see #24816")，仅跳过该损坏的类，而原有的 TestDSV4FlashFP4H200 (Marlin 路径) 继续运行。

关键文件：

- test/registered/unit/layers/quantization/test_mxfp4_sm90_cutlass.py（模块 量化测试；类别 test；类型 test-coverage）：通过 register_cuda_ci 的 disabled 参数禁用整个 MXFP4 单元测试文件，避免 CI 调度。
- test/registered/dsv4/test_deepseek_v4_flash_fp4_h200.py（模块 DeepSeek 测试；类别 test；类型 test-coverage）：为损坏的 TestDSV4FlashFP4H200FlashInferCutlass 类添加 @unittest.skip，仅跳过该类而不影响 Marlin 路径。

关键符号：未识别

关键源码片段

test/registered/unit/layers/quantization/test_mxfp4_sm90_cutlass.py

通过 register_cuda_ci 的 disabled 参数禁用整个 MXFP4 单元测试文件，避免 CI 调度。

```
"""Unit test for the SM90 cutlass MXFP4 path in :class:`Mxfp4MoEMethod`.
...
"""


from sglang.test.ci.ci_register import register_cuda_ci

# 变更前：register_cuda_ci(est_time=120, stage="stage-b", runner_config="1-gpu-large")
# 变更后：添加 disabled 参数，使该文件被 CI 排除
register_cuda_ci(
    est_time=120,
    stage="stage-b",
    runner_config="1-gpu-large",
    disabled="broken on main, see #24816", # 标准跳过机制
)
```

test/registered/dsv4/test_deepseek_v4_flash_fp4_h200.py

为损坏的 TestDSV4FlashFP4H200FlashInferCutlass 类添加 @unittest.skip，仅跳过该类而不影响 Marlin 路径。

```
@unittest.skip("broken on main, see #24816") # 新增行，跳过该损坏的测试类
@unittest.skipUnless(
    _flashinfer_has_sm90_cutlass_mxfp4(),
    "FlashInfer build lacks SM90 mixed-input MXFP4 helpers (PR #3084, >= 0.6.11)",
)
class TestDSV4FlashFP4H200FlashInferCutlass(ServerSanityMixin, CustomTestCase):
    """FlashInfer SM90 mixed-input cutlass MXFP4 backend..."""
    ...
```

评论区精华

该 PR 没有审核评论或讨论。作者在 body 中说明了动机：PR #24816 添加的测试在 main 上已损坏，需要暂时禁用以保持 CI 绿色。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅限于测试注册 / 跳过逻辑，不涉及任何生产代码。禁用的测试文件是 PR #24816 新添加的，原 main 分支没有这些测试，因此禁用不会降低现有测试覆盖率。需要确保后续修复后重新启用这些测试。
- 影响：影响范围：仅 CI 流程，不涉及用户或系统行为。 影响程度：低。临时禁用了两个与 FlashInfer SM90 MXFP4 相关的测试套件，不影响现有的 Marlin 路径测试。 后续行动：需要一个后续修复 PR 解决根本问题并重新启用这些测试。
- 风险标记：临时跳过，测试覆盖缺失，需要后续修复

关联脉络

- PR #24816 Add FlashInfer SM90 cutlass MXFP4 unit and server-sanity tests: 本 PR 禁用的测试正是 PR #24816 引入的，这些测试在 main 上已损坏。
- PR #25310 revert flashinfer 0.6.11 bumps: 回退 FlashInfer 0.6.11 至 0.6.8 的变更，可能与本 PR 涉及的 MXFP4 测试损坏相关——因为 FlashInfer 版本回退可能移除了 SM90 cutlass 支持。