

PR #27526 完整报告

sgl-project/sglang

[XPU CI] Re-enable stage B with docker-pull flow and split tests

合并时间: 2026-06-09 09:26

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27526>

执行摘要

- 一句话: 重新启用 XPU CI 的 stage B 测试并拆分测试套件
- 推荐动作: 值得关注 CI 配置模式 (预构建镜像拉取 + 多阶段拆分) 的实践, 可作为其他硬件后端 CI 优化的参考。核心开发人员无需深入阅读代码细节。

功能与动机

之前 stage B 的测试被合并到 stage A, 导致 stage A 运行时间过长。PR #27282 将 stage A 切换为 docker-pull 流程后, 能够以相同方式重新启用 stage B, 分担重测试, 从而提升 CI 整体效率和稳定性。

实现拆解

1. 修改 CI 工作流 (.github/workflows/pr-test-xpu.yml): 取消注释原 stage B job 配置, 改为使用与 stage A 一致的 xpu_ci_start_container.sh 辅助脚本和预构建镜像拉取流程, 并添加对 wait-for-stage-a 的依赖。
2. 拆分测试套件: 将 6 个 XPU 测试文件的注册套件从 stage-a-test-1-gpu-xpu 改为 stage-b-test-1-gpu-xpu, 包括 test_deepseek_ocr、test_deepseek_ocr_triton、test_intel_xpu_backend、test_topk、test_chunk_gated_delta_rule。
3. 禁用 OOM 测试: test_gemma_4_e2b.py 因在 stage B 运行器上启动时内存不足 (OOM), 在 register_xpu_ci 调用中添加 disabled 参数跳过注册, 同时保留文件以通过 lint 钩子检查。

关键文件:

- .github/workflows/pr-test-xpu.yml (模块 CI 流程; 类别 infra; 类型 infrastructure): 核心 CI 配置文件, 重新启用 stage B 并采用预构建镜像拉取流程
- test/registered/xpu/test_gemma_4_e2b.py (模块 E2E 测试; 类别 test; 类型 test-coverage): 因 OOM 禁用该测试注册, 但保留文件通过 lint 检查
- test/registered/xpu/test_deepseek_ocr_triton.py (模块 OCR 测试; 类别 test; 类型 test-coverage): 将测试从 stage A 移至 stage B
- test/registered/attention/test_chunk_gated_delta_rule.py (模块 Attention 测试; 类别 test; 类型 test-coverage): 将 XPU 测试注册从 stage A 移至 stage B

- test/registered/xpu/test_deepseek_ocr.py (模块 OCR 测试; 类别 test; 类型 test-coverage) : 将测试从 stage A 移至 stage B
- test/registered/xpu/test_intel_xpu_backend.py (模块 XPU 后端; 类别 test; 类型 test-coverage) : 将测试从 stage A 移至 stage B
- test/registered/xpu/test_topk.py (模块 TopK 测试; 类别 test; 类型 test-coverage) : 将测试从 stage A 移至 stage B

关键符号: 未识别

关键源码片段

test/registered/xpu/test_gemma_4_e2b.py

因 OOM 禁用该测试注册, 但保留文件通过 lint 检查

```
from sglang.test.ci.ci_register import register_xpu_ci

# Single e2e test: boot + a short Q&A.
register_xpu_ci(
    est_time=240,
    suite="stage-b-test-1-gpu-xpu",
    # CI 注册时直接跳过, 但保留文件使 pre-commit lint 通过
    disabled="OOM on stage-b XPU runner (server launch fails with --mem-fraction-static)",
)
```

评论区精华

无实质性讨论。由项目维护者 [mingfeima](#) 直接批准, [gemini-code-assist\[bot\]](#) 自动评论但未提供具体反馈。

- 暂无高价值评论线程

风险与影响

- 风险: 低风险。变更仅涉及 CI 配置和测试注册, 不影响运行时代码或核心逻辑。主要风险包括:
 - test_gemma_4_e2b 被添加 disabled 参数, 可能遗漏回归直到内存问题解决。
 - stage B 测试依赖预构建镜像拉取, 若镜像不可用则全部失败。
 - 部分测试 (如 test_chunk_gated_delta_rule) 估计时间长达 900 秒, 可能导致 stage B 超时。
- 影响: 影响范围限于 XPU CI 测试。stage A 测试时间缩短 (约减少 600 秒), 但整个 CI 流水线增加 stage B, 总耗时可能略有增加。开发团队需监控两个 stage 的成功率和稳定性, 特别是 stage B 重测试的 flakiness。
- 风险标记: 测试覆盖缺失, CI 稳定性, 资源限制

关联脉络

- PR #27282 [XPU CI] Switch stage A to docker-pull flow: 本 PR 复用该 PR 引入的预构建镜像拉取流程和辅助脚本
- PR #27156 [XPU CI] Fold stage B tests into stage A: 之前将 stage B 测试折叠到 stage A, 本 PR 拆分回来