PR #21625 完整报告

sgl-project/sglang

[CI] [FlashInfer v0.6.7] Use offline quantized checkpoint for MXFP8 Gemm tests

合并时间：2026-03-30 13:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21625

执行摘要

本 PR 将 MXFP8 Gemm CI 测试从在线量化切换为使用离线量化检查点，以解决 FlashInfer v0.6.7 更新后测试不稳定的问题，同时临时禁用 Triton 测试以避免 PCG 捕获时间过长，从而提高测试可靠性和 CI 效率。

功能与动机

在 FlashInfer v0.6.7 更新后，MXFP8 Gemm 测试出现精度下降和不稳定（如 PR body 所示，在线量化路径导致精度从 0.84 降至 0.76）。通过基准测试对比，发现离线量化检查点能提供稳定结果（精度约 0.84），因此变更旨在修复 CI 测试的不可靠性。动机源于维护者 @humansand 的需求，确保量化功能验证的准确性。

实现拆解

主要修改了 test/registered/quant/test_fp8_blockwise_gemm.py 文件：

- 全局变量变更：将 BF16_MODEL_PATH 重命名为 MXFP8_MODEL_PATH，并更新为离线量化模型路径 'zianglih/Qwen3-4B-Instruct-2507-MXFP8'。
- 测试配置简化：在 MXFP8GemmBase.setUpClass 方法中，移除 --quantization mxfp8 参数，因为现在使用预量化检查点，无需在线量化步骤。
- 测试类调整：临时禁用 TestMXFP8GemmTriton 类，通过添加 @unittest.skip 装饰器，注释说明由于 PCG 捕获时间过长（5-7 分钟），待后续修复。

评论区精华

- 链接正确性检查：gemini-code-assist[bot] 在 review 评论中指出：'The link to the pull request seems to be broken as PR number 19835 does not exist. This appears to be a typo.' 这引发了对文档准确性的关注。后续 Issue 评论确认 #19835 已合并，但 typo 问题未进一步讨论。
- Triton 测试性能讨论：在 Issue 评论中，作者 @zianglih 说明：'TestMXFP8GemmTriton works after https://github.com/sgl-project/sglang/pull/19835 but currently compiling PCG takes 5-7mins, so I disable it again.' 讨论焦点从功能修复转向性能优化，结论是测试被暂时禁用以等待改进。

风险与影响

- 风险分析：变更风险较低。使用离线检查点可能增加模型版本管理复杂性，但路径明确指定，风险可控；Triton 测试被禁用可能暂时减少覆盖范围，但这是临时措施，作者计划优化后恢复。此外，需确保离线量化与在线量化行为一致，避免隐藏回归问题。
- 影响评估：直接影响 CI 测试套件，提升 MXFP8 Gemm 测试的稳定性和精度，减少误报失败，加速开发流程；间接增强团队对量化功能的信心；对最终用户无直接影响，因为是内部测试改进。

关联脉络

本 PR 与历史 PR #19835（'fix cuda graph capturing error in sm120 mxfp8 triton path'）直接相关，后者修复了 MXFP8 Triton 路径的 CUDA 图捕获错误。本 PR 在此基础上优化测试配置，通过使用离线量化检查点应对在线量化路径的不稳定性，体现了对量化功能测试的持续改进。此外，近期 PR 如 #21634 也涉及测试简化，显示团队在优化 CI 流程方面的趋势。