PR #20648 完整报告

sgl-project/sglang

[CI] Add Llama 3.1 8B Instruct FP4 CI test on SM120

合并时间：2026-04-02 09:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20648

执行摘要

本 PR 添加了针对 NVFP4 量化 Llama 3.1 8B Instruct 模型在 SM120 GPU 上的 CI 测试，以扩展量化测试覆盖并确保硬件兼容性。通过新增测试文件运行 GSM8K 评估，验证模型准确性，增强 CI 流水线的健壮性。

功能与动机

从 PR body 中：'Improves stage-b-test-small-1-gpu (SM120) coverage for quantized model tests. One of many addressing #20600.' 旨在填补单 GPU 量化测试空白，特别是针对 FP4 格式，确保量化模型在特定硬件的功能正确性。

实现拆解

实现集中于单个文件 test/registered/quant/test_nvfp4_gemm_sm120.py。关键改动如下：

- 定义基类 FP4GemmSM120Base，在 setUpClass 中设置服务器启动参数，包括模型路径、量化设置和禁用 piecewise CUDA graph。
- 添加测试方法 test_gsm8k，运行 GSM8K 评估并验证准确性阈值（>0.64）。
- 通过子类 TestFP4GemmSM120Auto 使用 'auto' 后端进行测试。
- 使用 register_cuda_ci 注册到 CI 套件，估计运行时间 90 秒。

评论区精华

Review 讨论中提炼出以下要点：

- 测试结构优化：gemini-code-assist[bot] 建议使用基类以提高可维护性，align with similar tests。
- 时间估计：b8zhong 询问时间准确性，DerekY2 回应实际约 50 秒，设置估计为 90 秒。
- 硬件要求：b8zhong 建议设置 SM ≥ 100，以允许在 SM100 上运行。
- 后端选择：b8zhong 建议测试默认后端而不是硬编码 cuDNN。
- 内核问题：DerekY2 报告 Piecewise CUDA Graph 导致 TorchDynamo 崩溃，通过添加 --disable-piecewise-cuda-graph 参数规避，但问题未根本解决。

风险与影响

风险：

- 测试时间估计可能不准确，影响 CI 流水线调度。
- TorchDynamo 崩溃问题被临时规避，可能掩盖内核或量化实现的潜在 bug。
- 准确性阈值设置需合理，避免误报。
- 测试仅覆盖 'auto' 后端，可能未全面验证所有后端选项。

影响：

- 对用户无直接影响，是内部测试改进。
- 对系统：增强 CI 测试覆盖，有助于早期发现问题，提高稳定性。
- 对团队：提供更全面测试数据，但可能增加 CI 运行时间和资源消耗。

关联脉络

与历史 PR 的关联：

- PR 20717：添加 FP8 测试在 SM120，类似扩展量化测试覆盖。
- PR 21576：集成 FlashInfer 量化 GEMM，涉及量化特性。
- PR 21888：修复 TorchDynamo 问题，与本 PR 中报告的内核崩溃相关。 这些 PR 共同反映了团队在量化测试和 CI 覆盖方面的持续演进。