PR #24604 完整报告

sgl-project/sglang

[Bugfix] Fix a bug causing NVFP4 to be tested on all gpus like SM90 devices.

合并时间：2026-05-09 02:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24604

执行摘要

- 一句话：修复 NVFP4 测试在非 Blackwell GPU 上误跑的问题
- 推荐动作：本 PR 解决了测试隔离问题，值得合入。建议后续在 is_blackwell 函数中添加更多架构版本时同步更新此测试。

功能与动机

在 SM90 设备上运行 pytest test/registered/unit/layers -v 时，NVFP4 测试因只检查 CUDA 存在而错误执行，导致测试失败。PR 旨在将 NVFP4 测试限定在支持该特性的 Blackwell GPU 上。

实现拆解

1. 替换 skip 装饰器：在 test/registered/unit/layers/quantization/test_fp4_kv_cache_quant_method.py 中，将 skip_if_no_cuda 函数替换为 skip_if_no_blackwell_nvfp4，从 sglang.srt.utils 导入 is_blackwell 进行架构判断。
2. 动态获取 sm_version：在 test_quantize_dequantize_roundtrip 测试中，原本硬编码 sm_version=120，现改为通过 torch.cuda.get_device_capability() 动态获取当前 GPU 的计算能力，提高了测试的通用性。
3. 调整导入：新增对 is_blackwell 的导入，并删除了未被使用的 torch 相关导入（如 unittest 仍保留）。

关键文件：

- test/registered/unit/layers/quantization/test_fp4_kv_cache_quant_method.py（模块 量化测试；类别 test；类型 test-coverage；符号 skip_if_no_blackwell_nvfp4, test_quantize_dequantize_roundtrip）：唯一修改的文件，包含 skip 装饰器替换和 sm_version 动态获取的变更。

关键符号：skip_if_no_blackwell_nvfp4, test_quantize_dequantize_roundtrip

评论区精华

gemini-code-assist[bot] 在 review 中指出，NVFP4KVMethod 实现似乎支持 SM100 和 SM120，而初始 skip 条件 skip_if_no_sm120 仅限制在 SM120，建议放宽到 Blackwell 系列（SM100/SM120）或补充注释。作者 xz-keg 随后将函数改为 skip_if_no_blackwell_nvfp4 并

使用 is_blackwell()，以涵盖 SM100 和 SM120。

- Skip 条件应涵盖 SM100 和 SM120 (correctness): 作者将 skip 条件改为 skip_if_no_blackwell_nvfp4 并使用 is_blackwell() 以包含 SM100 和 SM120。

风险与影响

- 风险：低风险。变更仅涉及测试文件，不影响任何生产代码。主要风险是 is_blackwell() 或 is_blackwell 导入可能未正确匹配所有 Blackwell 架构，但已有测试验证其行为。
- 影响：影响范围局限于单元测试。SM90 等非 Blackwell GPU 将跳过 NVFP4 测试，避免误报失败。Blackwell GPU 上测试行为不变。
- 风险标记：测试隔离修复

关联脉络

- 暂无明显关联 PR