

PR #31939 完整报告

sgl-project/sglang

Add Inkling to nightly test

合并时间: 2026-07-21 23:06

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31939>

执行摘要

- 一句话: 新增 Inkling-NVFP4 模型夜间测试
- 推荐动作: 值得精读: 该 PR 展示了如何为大型 MoE 模型 (975B) 添加端到端夜间测试, 包括启动参数配置、误差基线设定和硬件条件跳过。可作为类似测试的模板。

功能与动机

PR body 说明动机: Add Inkling-NVFP4 to the 8-GPU nightly, running both the accuracy and performance legs via run_combined_tests, matching the other test/registered/8-gpu-models/ entries.

实现拆解

1. 新建测试文件: 在 test/registered/8-gpu-models/ 下创建 test_inkling_nvfp4_nightly.py, 注册为 nightly-8-gpu-common suite。
2. 定义启动参数: 设置 NVFP4_ARGS 列表, 包含 --tp=8、--quantization=modelopt_fp4、--attention-backend=fa4 等 Blackwell 专用参数, 参考 cookbook 配置。
3. 编写测试类: TestInklingNVFP4Nightly 继承 unittest.TestCase, 通过 @unittest.skipIf(not is_blackwell_system(), ...) 跳过非 Blackwell 系统。
4. 执行精度与性能测试: 在 test_inkling_nvfp4 中调用 run_combined_tests, 分别传入 AccuracyTestParams (gsm8k, baseline=0.95) 和 PerformanceTestParams。
5. 启用统一基数树: 使用 envs.SGLANG_ENABLE_UNIFIED_RADIX_TREE.override(1) 启用缓存优化。

关键文件:

- test/registered/8-gpu-models/test_inkling_nvfp4_nightly.py (模块测试; 类别 test; 类型 test-coverage; 符号 TestInklingNVFP4Nightly, test_inkling_nvfp4): 新增 Inkling-NVFP4 模型的 8-GPU 夜间测试文件, 包含精度和性能测试的全部逻辑。

关键符号: TestInklingNVFP4Nightly.test_inkling_nvfp4

评论区精华

review 过程中仅包含 rerun-test 命令执行及结果确认, 无技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：本 PR 仅新增测试文件，不修改任何生产代码。可能风险：若 Blackwell 硬件环境不稳定，测试可能在 nightly CI 中产生 flaky 失败。另外，
baseline_accuracy=0.95 基于有限测量，若 FP4 内核偶发精度波动，可能导致误报。
- 影响：影响范围：仅限 Blackwell 8-GPU 夜间测试套件。新增测试约需 3600 秒运行时间，增加 CI 资源消耗但属于预期内。对用户无直接影响。
- 风险标记：测试覆盖有限，硬件依赖（Blackwell），基线精度可能波动

关联脉络

- 暂无明显关联 PR