PR #43186 完整报告

vllm-project/vllm

[CI] Lower granite-4.0-h-tiny gsm8k threshold for Hybrid SSM NixlConnector PD accuracy tests (4 GPUs)

合并时间：2026-05-21 01:00

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43186

执行摘要

- 一句话：降低 hybrid SSM PD 精度测试阈值以容忍内核变更导致的抖动
- 推荐动作：I 建议合入以解锁 CI，但同时需要创建 issue 跟踪 chunk_scan 与 selective_state_update 内核的数值差异修复，并在修复后及时回调精度阈值。

功能与动机

修复因 #42430 引入的 kernel divergence 导致 Hybrid SSM NIXL PD 精度测试持续失败，同时记录根本原因并解锁 CI。PR body 中详细说明了 chunk_scan 与 SSU 内核在 bf16 ULP 级别产生不同输出的事实。

实现拆解

1. 调整期望值: 在 tests/v1/kv_connector/nixl_integration/test_accuracy.py 中将 ibm-granite/granite-4.0-h-tiny 的 expected 从 0.80 改为 0.77。
2. 放宽容差: 将 RTOL 从 0.03 改为 0.05，并添加详细注释解释放宽原因，引用此 PR 编号及根本原因（SSU vs chunk_scan 数值差异）。
3. 记录 TODO: 在代码中添加 TODO 注释，指向此 PR 并建议在内核差异修复后收紧参数。

关键文件：

- tests/v1/kv_connector/nixl_integration/test_accuracy.py（模块 精度测试；类别 test；类型 test-coverage）：唯一修改的文件，调整 granite-4.0-h-tiny 的期望精度值和容差，并添加注释解释背景。

关键符号：未识别

关键源码片段

tests/v1/kv_connector/nixl_integration/test_accuracy.py

唯一修改的文件，调整 granite-4.0-h-tiny 的期望精度值和容差，并添加注释解释背景。

```
# 文件：tests/v1/kv_connector/nixl_integration/test_accuracy.py
# 关键常数变更：

# 之前：RTOL = 0.03 # 在 granite 模型上因 SSU vs chunk_scan 内核差异导致抖动过大
```

```
# 之后：
# TODO(#43186): 从 0.03 放宽以吸收 chunk_scan/SSU 数值抖动
# 待内核差异修复后可收紧此值
RTOL = 0.05

# 模型特定期望值字典中的变更项：
EXPECTED_VALUES = {
    # ... 其他模型保持不变 ...
    # 之前："ibm-granite/granite-4.0-h-tiny": 0.80,
    # 之后：调低至 0.77 以匹配 #42430 后 SSU 内核下的实际分布
    "ibm-granite/granite-4.0-h-tiny": 0.77,
    # ...
}
```

评论区精华

1. 数值差异合理性：评论者 tdoublep 质疑“why a large accuracy drop for algebraically equivalent kernels”，haosdent 解释了两者在 bf16 ULP 边界产生不同舍入路径，但 maintainers 同意先解锁 CI。
2. Eager vs CUDA Graph 一致性：NickLucche 最初担心 gating 会导致 eager 与 CG 模式不一致，但后续接受放宽阈值方案。
3. 精度波动范围：观察到 4 次运行结果为 0.7498, 0.7415, 0.7559, 0.78，波动达 ~0.04，确认需放宽。

- 精度下降的数值根源 (correctness)：接受暂时放宽阈值，但需后续单独调查内核数值差异。
- Eager vs CUDA Graph 一致性 (design)：放弃 gating 方案，统一放宽测试阈值。

风险与影响

- 风险：低风险。仅调整精度测试的期望值和容差，不影响任何生产逻辑。但需注意：宽松的阈值可能掩盖真正的精度回归，因此代码中已添加 TODO 提醒后续必须跟踪内核差异修复并收紧参数。
- 影响：仅影响 Hybrid SSM NIXL PD 精度测试（4 GPU 配置），解锁了因 #42430 持续失败的 CI 流水线。对用户无影响，对开发者是临时解决方案。
- 风险标记：测试阈值放宽可能掩盖回归

关联脉络

- PR #42430 [Bugfix] Use enable_sm120_family for per-tensor FP8 CUTLASS kernels on SM12.1: 根本原因：该 PR 将 recompute 从 chunk_scan 切换到 SSU，导致精度下降。