

PR #28226 完整报告

sgl-project/sglang

[AMD] Relax allreduce-fusion residual accuracy tolerance to 1 bf16 ULP

合并时间: 2026-06-19 10:18

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28226>

执行摘要

- 一句话: 放宽 AMD allreduce 融合精度测试容差
- 推荐动作: 建议合并。该 PR 正确识别了 CI 失败的根因 (Docker 镜像 aiter 版本落后), 并通过合理的容差调整解决了问题。容差推导过程严谨, 值得参考。

功能与动机

CI 中 `test_fused_ar_rms_residual_accuracy` 测试持续失败, 因为预构建的 aiter 内核缺乏 ROCm/aiter#2586 (commit 43b7379b8) 中的精度修复。该修复在 f32 allreduce 累加后、残差加法前插入 bf16 往返, 使融合内核与非融合路径逐位匹配。原始测试要求 bit-exact, 因此在旧内核上失败。

实现拆解

1. 调整测试容差: 在 `test/registered/ops/test_aiter_allreduce_fusion_amd.py` 中, 将 `passed = max_diff == 0.0` 改为 `passed = max_diff <= ATOL`, `ATOL = 0.13`, 该值略大于 1 bf16 ULP 在典型输入量级下的最大值 0.125。
2. 移除冗余代码: 删除了 `tensor_model_parallel_all_reduce` 的导入和调用, 以及 `unfused_ar` 和 `ar_diff` 相关计算, 因为独立 allreduce 的正确性已由其他测试覆盖。
3. 更新文档字符串: 补充了容差推导过程和 aiter 修复引用的说明。

关键文件:

- `test/registered/ops/test_aiter_allreduce_fusion_amd.py` (模块 AMD 融合测试; 类别 test; 类型 test-coverage): 唯一被修改的文件。将 bit-exact 判断改为 1 bf16 ULP 容差, 移除冗余代码, 更新文档。

关键符号: `_run_residual_accuracy_check`

关键源码片段

`test/registered/ops/test_aiter_allreduce_fusion_amd.py`

唯一被修改的文件。将 bit-exact 判断改为 1 bf16 ULP 容差, 移除冗余代码, 更新文档。

```
def _run_residual_accuracy_check():  
    """Distributed entry point: residual accuracy across 1-stage/2-stage paths.  
  
    Regression test for the 1-stage kernel accuracy bug (ROCm/aiter#2586):
```

allreduce_fusion_kernel_1stage accumulated in f32 and added the residual before rounding to bf16, while the unfused path rounds allreduce to bf16 first. The fix (43b7379b8 in aiter) inserts a bf16 round-trip after accumulation so the fused kernel matches the unfused path bit-for-bit.

The tolerance here is 1 bf16 ULP (atol = bf16_eps * max_magnitude $\approx$ 0.125) rather than 0.0, because the prebuilt aiter kernel in the CI docker image may pre-date the fix. A diff of exactly 1 ULP indicates the unfixed kernel; a larger diff indicates a real regression and will fail the test.

"""

... (dist init omitted for brevity)

dtype = torch.bfloat16

eps = 1e-6

Allow at most 1 bf16 ULP of error in the residual output.

bf16 epsilon = 2^{-7} ; values in practice stay below ~ 16 , so 1 ULP $\leq$ 0.125.

A multi-ULP error (>0.125) indicates a real regression and fails the test.

Exactly 1 ULP indicates the prebuilt aiter kernel predates the fix in

ROCm/aiter#2586 (43b7379b8); the test still guards against regressions.

ATOL = 0.13 # Slightly above 1 ULP at magnitude ~ 16 ($2^{-3} = 0.125$)

all_pass = True

test_cases = [(m, n) for n in HIDDEN_DIMS for m in [1, 4, 8, 16, 32, 64, 128]]

... (loop omitted)

Reference: fused_ar (AR rounded to bf16, zero residual) + residual.

With the aiter fix (43b7379b8), this matches fused_res bit-for-bit.

Without the fix, fused_res may differ by exactly 1 bf16 ULP, which

is tolerated by ATOL but still guarded against larger regressions.

expected = fused_ar + residual

diff = (fused_res.float() - expected.float()).abs()

max_diff = diff.max().item()

frac_nonzero = (diff > 0).float().mean().item()

nbytes = m * n * dtype.itemsize

stage = "1-stage" if nbytes $\leq$ 128 * 1024 else "2-stage"

passed = max_diff $\leq$ ATOL # Changed from max_diff == 0.0

评论区精华

无实质性讨论。hubertlu-tw 直接批准，仅有一个 lint 串（移除未使用的导入、格式化长行、添加空行），已在后续提交中修复。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅修改测试容忍条件，不影响任何内核或运行时逻辑。如果将来 aiter 内核出现严重的数值回归，新容差仍能捕获（多 ULP 错误 > 0.13 ）。但存在细微的漏测风险：

如果未来融合内核的误差恰好控制在 1 ULP 内但实际精度退化，测试可能无法识别。

- 影响：仅影响 AMD CI 流水线中 stage-c-test-large-8-gpu-amd 的单一测试。修复后，使用旧 aiter 内核的 CI 镜像可通过测试，不再因为预构建内核版本差异而失效。对其他平台无影响。
- 风险标记：测试容差放宽

关联脉络

- PR #2586 Fix: Numerical Accuracy in allreduce_fusion_kernel_1stage: 该 PR 修复了 aiter 内核的精度问题，本 PR 调整测试容差以兼容修复前的旧内核。