

PR #27505 完整报告

sgl-project/sglang

[AMD] Fix test_deepseek_r1_mxfp4_8gpu.py : disable async-assert probes on AMD CI

合并时间: 2026-06-09 23:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27505>

执行摘要

- 一句话: AMD CI 暂时禁用 `async-assert` 以避免 MXFP4 测试崩溃
- 推荐动作: 这是一个临时修复, 建议尽快定位 MXFP4 EAGLE-MTP 路径中 `async-assert` 的真因, 并确保在 AMD 上也能安全启用。PR 本身值得快速合并以恢复 CI 稳定性。

功能与动机

PR #27461 在 AMD CI 中默认启用了 `SGLANG_ENABLE_ASYNC_ASSERT=1`, 但 MXFP4 EAGLE-MTP 解码路径中的 `torch._assert_async` 探针会触发 HSA 硬件异常, 导致调度器崩溃。为了恢复 AMD CI 的绿色状态, 需要在 AMD CI 脚本中临时禁用该环境变量。

实现拆解

1. 修改 `scripts/ci/amd/amd_ci_exec.sh` 中的环境变量映射: 将 `SGLANG_ENABLE_ASYNC_ASSERT` 从 1 改为 0。
2. 添加注释说明禁用原因: `torch._assert_async` 在 MXFP4 EAGLE-MTP 解码路径中导致 HSA 硬件异常。
3. 该变更仅影响 AMD CI 流程, NVIDIA CI 仍保持 `SGLANG_ENABLE_ASYNC_ASSERT=1`, 不涉及任何运行时或内核代码修改。

关键文件:

- `scripts/ci/amd/amd_ci_exec.sh` (模块 CI 脚本; 类别 `infra`; 类型 `infrastructure`): 这是唯一修改的文件, 通过将 `SGLANG_ENABLE_ASYNC_ASSERT` 设为 0 来禁用 AMD CI 上的异步断言, 从而避免 MXFP4 EAGLE-MTP 测试崩溃。

关键符号: 未识别

关键源码片段

`scripts/ci/amd/amd_ci_exec.sh`

这是唯一修改的文件, 通过将 `SGLANG_ENABLE_ASYNC_ASSERT` 设为 0 来禁用 AMD CI 上的异步断言, 从而避免 MXFP4 EAGLE-MTP 测试崩溃。

```
# scripts/ci/amd/amd_ci_exec.sh
# 环境变量映射表: 定义 AMD CI 运行时的环境变量
# 注意: SGLANG_ENABLE_ASYNC_ASSERT 被设置为 0, 因为 #27461 引入的异步断言探针
```

```
# 在 MXFP4 EAGLE-MTP 解码路径中触发 HSA 硬件异常，导致测试崩溃。
declare -A ENV_MAP=(
    [SGLANG_IS_IN_CI_AMD]=1
    [SGLANG_IS_IN_CI]=1
    # Disabled on AMD: the async-assert probes (#27461) fire torch._assert_async in
    # the MXFP4 EAGLE-MTP decode path and abort the queue with an HSA hardware
    # exception.
    [SGLANG_ENABLE_ASYNC_ASSERT]=0
    [SGLANG_USE_AITER]=1
)
```

评论区精华

维护者 HaiShaw 要求作者评估必要时在代码中设置 `SGLANG_ENABLE_ASYNC_ASSERT=1` 的必要性。但本 PR 仅是在 CI 级别暂时禁用，未深入追查根本原因。

- 是否需要永久修复而不是禁用 (question): 当前仅作临时禁用，未深入分析根本原因。

风险与影响

- 风险：风险极低：变更仅修改 AMD CI 环境变量，不涉及运行时逻辑或内核代码。但禁用 `async-assert` 后，AMD CI 上其他测试可能无法捕获潜在的异步错误，不过 NVIDIA CI 仍保留该功能。
- 影响：影响范围仅限于 AMD CI：使 `test_deepseek_r1_mxfp4_8gpu.py` 恢复绿色，避免 HSA 硬件异常导致的测试崩溃。但异步断言的覆盖率在 AMD 上降低，可能掩盖其他潜在问题。
- 风险标记：临时解决方案，降低断言覆盖率

关联脉络

- PR #27461 Enable `async-assert` probes in AMD CI: 本 PR 正是为了修复 #27461 引入的回归问题：在 AMD CI 中默认启用 `async-assert` 导致 MXFP4 测试崩溃。