

PR #27955 完整报告

sgl-project/sglang

[CI] Fix GB300 TestDummyWithSBO crash: disable NaN assert and coredump for dummy weights

合并时间: 2026-06-12 05:17

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27955>

执行摘要

- 一句话: 禁用虚权重测试的 NaN 断言崩溃
- 推荐动作: 该 PR 是典型的测试基础设施修复, 改动简单直接, 适合快速合并。开发者在类似场景 (虚权重测试) 中可参考此模式来避免断言影响。

功能与动机

Dummy random weights legitimately produce NaN logits, which now trip the CI async NaN assert added in #27883 and kill the scheduler during warmup. 需要禁用该测试的崩溃机制, 让 NaN 以节流告警方式处理。

实现拆解

仅修改 `test/registered/4-gpu-models/test_deepseek_v3_cuteds1_4gpu.py` 中 `setUpClass` 方法的 `env` 字典, 添加 4 个环境变量覆盖:

1. `SGLANG_ENABLE_ASYNC_ASSERT=0`: 禁用异步 NaN 断言。
2. `SGLANG_CUDA_COREDUMP=0`: 禁用 GPU coredump。
3. `CUDA_ENABLE_COREDUMP_ON_EXCEPTION=0`: 显式覆盖, 因为当 `SGLANG_CUDA_COREDUMP=1` 时该变量已注入环境。
4. `SGLANG_CUDA_COREDUMP_BEFORE_CRASH=0`: 禁用崩溃前 coredump。

关键文件:

- `test/registered/4-gpu-models/test_deepseek_v3_cuteds1_4gpu.py` (模块 测试; 类别 `test`; 类型 `test-coverage`): 唯一变更文件, 通过环境变量禁用异步 NaN 断言和 GPU coredump, 修复虚权重导致的 CI 崩溃。

关键符号: 未识别

关键源码片段

`test/registered/4-gpu-models/test_deepseek_v3_cuteds1_4gpu.py`

唯一变更文件, 通过环境变量禁用异步 NaN 断言和 GPU coredump, 修复虚权重导致的 CI 崩溃。

```
# 在 setUpClass 方法中启动服务器时，通过环境变量传递
# Dummy 随机权重会正常产生 NaN logits，因此需要关闭 CI 崩溃机制
# 使用 NaN 告警而非崩溃调度器
cls.process = popen_launch_server(
    cls.model,
    cls.base_url,
    timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    other_args=other_args,
    env={
        **os.environ,
        "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK": "256",
        "SGLANG_MOE_NVFP4_DISPATCH": "0",
        # Dummy random weights legitimately produce NaN logits; turn
        # off the CI crash machinery (async assert, coredump on GPU
        # exception, crash-time coredump) so NaN is sanitized with a
        # warning instead of killing the scheduler.
        "SGLANG_ENABLE_ASYNC_ASSERT": "0", # 禁用异步 NaN 断言
        "SGLANG_CUDA_COREDUMP": "0", # 禁用 GPU coredump
        # 当 SGLANG_CUDA_COREDUMP=1 时该变量已被注入环境，需要显式覆盖
        "CUDA_ENABLE_COREDUMP_ON_EXCEPTION": "0",
        "SGLANG_CUDA_COREDUMP_BEFORE_CRASH": "0",
    },
)
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：该变更加入的条件检查仅在测试环境有效，不影响生产代码。风险极低，因为环境变量只在该测试的子进程中生效，且只是关闭了断言和 coredump，不会引入新问题。
- 影响：仅影响 DeepSeek V3 CuteDSL 4 GPU 测试，修复了因虚权重导致的 CI 崩溃，恢复该测试的稳定性。对其他测试和功能无影响。
- 风险标记：仅测试环境变更

关联脉络

- PR #27883 Add async NaN assert: 本 PR 修复了 #27883 引入的异步 NaN 断言导致虚权重测试崩溃的问题。