PR #27461 完整报告

sgl-project/sglang

Enable async-assert invariant probes by default in CI

合并时间：2026-06-07 07:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27461

执行摘要

- 一句话：在 CI 中默认启用异步断言探测
- 推荐动作：该 PR 值得阅读，尤其是对负责 CI 基础设施和测试策略的工程师。它展示了如何通过环境变量集中管控运行时检查，并在保持零同步开销的前提下扩大验证覆盖。设计上清晰分离了 CUDA/AMD 与 NPU/MUSA 的配置，未来如果 NPU/MUSA 支持 torch._assert_async，只需简单添加环境变量即可。

功能与动机

PR body 指出：异步断言探测可以在无 GPU-CPU 同步的前提下，在下一个同步点以干净断言暴露张量中的 NaN/Inf/OOB 等错误，而不是静默产生错误结果或非法地址崩溃。此前仅少数 spec/eagle 测试通过 envs.SGLANG_ENABLE_ASYNC_ASSERT.override(True) 选择启用，未发挥全面检测能力。本次将其提升为 CI 全局默认，使所有 CUDA 和 AMD 上的模型测试都能受益。

实现拆解

1. CUDA CI 工作流添加环境变量：在 .github/workflows/_pr-test-stage.yml、rerun-test.yml、nightly-test-nvidia.yml、weekly-test-nvidia.yml、nightly-72-gpu-gb200.yml 的 env: 块中添加 SGLANG_ENABLE_ASYNC_ASSERT: true，覆盖所有 CUDA PR 测试、夜间测试和每周测试。
2. AMD CI 脚本添加环境变量：在 scripts/ci/amd/amd_ci_exec.sh 的 ENV_MAP 中添加 [SGLANG_ENABLE_ASYNC_ASSERT]=1，通过 docker exec -e 注入到容器中，覆盖所有 AMD 测试套件。
3. 移除测试中的冗余 override：删除 10 个测试文件和 fixture 中的 with envs.SGLANG_ENABLE_ASYNC_ASSERT.override(True): 及相关 import，包括 eagle_fixture.py、test_eagle_constrained_decoding.py、test_deepseek_v3_fp4_mtp_small.py 等。这些测试现在直接依赖 CI 环境变量。

关键文件：

- .github/workflows/_pr-test-stage.yml（模块 CI 流程；类别 infra；类型 infrastructure）：核心 CI 工作流，是所有 PR 测试的入口，在此添加环境变量使后续所有 CUDA CI 任务默认启用异步断言。

- .github/workflows/nightly-72-gpu-gb200.yml（模块 CI 流程；类别 infra；类型 infrastructure）：72-GPU GB200 夜间测试工作流同样添加该环境变量，确保大规模测试也启用异步断言。
- python/sglang/test/server_fixtures/eagle_fixture.py（模块 测试夹具；类别 test；类型 test-coverage；符号 EagleServerBase.setUpClass）：最常用的 EAGLE 测试 fixture，移除了 async assert override，代表所有测试文件的清理模式。
- test/registered/spec/eagle/test_deepseek_v3_fp4_mtp_small.py（模块 推测测试；类别 test；类型 test-coverage；符号 TestDeepseekV3FP4MTP.setUpClass）：DeepSeek V3 FP4 MTP 测试文件，移除 override，验证在真实模型测试中的清理。

关键符号：EagleServerBase.setUpClass, TestDeepseekV3FP4MTP.setUpClass, TestEagleDPAttnServerSmall.setUpClass

关键源码片段

python/sglang/test/server_fixtures/eagle_fixture.py

最常用的 EAGLE 测试 fixture，移除了 async assert override，代表所有测试文件的清理模式。

```
# 从 sglang.srt.environ 导入 envs 已移除，因为不再需要 override
from sglang.srt.utils.common import kill_process_tree
from sglang.test.test_utils import (
    DEFAULT_DRAFT_MODEL_EAGLE,
    DEFAULT_TARGET_MODEL_EAGLE,
    DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    DEFAULT_URL_FOR_TEST,
    CustomTestCase,
    popen_launch_server,
)

class EagleServerBase(CustomTestCase):
    target_model = DEFAULT_TARGET_MODEL_EAGLE
    draft_model = DEFAULT_DRAFT_MODEL_EAGLE
    spec_algo = "EAGLE"
    spec_steps = 5
    spec_topk = 8
    spec_tokens = 64
    mem_fraction_static = 0.7
    extra_args = []

    @classmethod
    def setUpClass(cls):
        cls.base_url = DEFAULT_URL_FOR_TEST
        # 之前：with envs.SGLANG_ENABLE_ASYNC_ASSERT.override(True):
        # 现在：CI 环境已默认设置 SGLANG_ENABLE_ASYNC_ASSERT=true，
        # 因此直接启动服务器，无需上下文管理器。
        cls.process = popen_launch_server(
            cls.target_model,
```

```
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[
                f"--speculative-algorithm={cls.spec_algo}",
                f"--speculative-draft-model-path={cls.draft_model}",
                f"--speculative-num-steps={cls.spec_steps}",
                f"--speculative-eagle-topk={cls.spec_topk}",
                f"--speculative-num-draft-tokens={cls.spec_tokens}",
                f"--mem-fraction-static={cls.mem_fraction_static}",
            ]
            + cls.extra_args,
        )

    @classmethod
    def tearDownClass(cls):
        kill_process_tree(cls.process.pid, wait_timeout=60)
```

评论区精华

该 PR 由作者独立推动，review 过程中无实质讨论或争议。有一条 /rerun-test 命令用于重跑失败测试，已通过。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 新增 CI 失败风险：启用异步断言后，之前被掩蔽的 NaN/Inf 等数值错误将显式抛出，可能导致现有测试新增失败。但这是预期行为，有助于暴露潜在问题。
  - NPU/MUSA 未覆盖：CUDA 和 AMD 已全部覆盖，但 NPU 和 MUSA 平台未配置该变量，这些平台上的数值错误仍可能被忽视。
  - 依赖 torch._assert_async 支持：该 API 在较新 PyTorch 版本中可用，sglang 依赖版本已满足要求，风险极低。
  - 性能影响：异步断言无需 GPU-CPU 同步，计算开销接近于零。
- 影响：
  - 用户：无直接用户可见变化，仅 CI 内部机制增强。
  - 系统：所有 CUDA 和 AMD 上的模型测试现在都能在关键检查点执行张量合法性验证，及早捕获异常。
  - 团队：测试代码更简洁，不再需要每个测试单独开启异步断言；CI 配置集中化管理，维护成本降低。
  - 风险标记：NPU/MUSA 未覆盖，依赖 torch._assert_async 支持，可能导致现有测试新增失败

关联脉络

- PR #26972 Spec v2 tree drafting (topk>1) with page_size>1: 该 PR 引入了 EAGLE tree drafting 的 spec 测试，这些测试是本次移除 override 的主要对象之一。CI 默认启用异步断言使这些测试无需手动配置，与本次变更形成互补。