PR #25932 完整报告

sgl-project/sglang

[AMD] Fix AMD stage-a-test-small-1-gpu

合并时间：2026-05-21 11:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25932

执行摘要

- 一句话：校准 AMD EAGLE3 测试阈值并延长采样窗口
- 推荐动作：建议合入。该 PR 针对 AMD CI 回归问题提供了合理且最小侵入的校准方案，在保留测试覆盖率的同时适应平台差异。值得关注的决策：is_in_amd_ci() 条件分支适配策略，可作为跨平台 CI 测试参数化的参考模式。

功能与动机

AMD stage-a CI 在 TestBasicSanityEagle3.test_fwd_occupancy 上持续失败。观察到 AMD EAGLE3 样本的中位占用量约为 86.26，低于 CUDA 校准的 97.0 阈值，且重试时由于采样数量不足导致 NaN 样本过多。需要按 AMD 实际表现校准阈值并延长采样窗口，确保 CI 回归检测有效。

实现拆解

1. 在 test/registered/core/test_basic_sanity_eagle3.py 中新增 is_in_amd_ci() 导入附加工具函数。
2. 将静态阈值 97.0 改为条件赋值：AMD CI 环境使用 80.0，否则保持 97.0。
3. 新增 fwd_occupancy_max_new_tokens 属性，AMD CI 设为 4096（其他平台 2048），通过增加生成 token 数量收集更多有效样本减少 NaN 波动。

关键文件：

- test/registered/core/test_basic_sanity_eagle3.py（模块 测试；类别 test；类型 test-coverage；符号 TestBasicSanityEagle3.fwd_occupancy_threshold, TestBasicSanityEagle3.fwd_occupancy_max_new_tokens）：核心测试文件，修改了 EAGLE3 前向占用量测试的阈值和采样窗口大小，是此次变更的唯一文件。

关键符号：未识别

关键源码片段

test/registered/core/test_basic_sanity_eagle3.py

核心测试文件，修改了 EAGLE3 前向占用量测试的阈值和采样窗口大小，是此次变更的唯一文件。

```
# test/registered/core/test_basic_sanity_eagle3.py (head)
```

```
class TestBasicSanityEagle3(
    BasicAPIContractMixin,
    BasicDecodeCorrectnessMixin,
    BasicSchedulerStressMixin,
    FwdOccupancyMixin,
    HellaswagMixin,
    CustomTestCase,
):
    served_model_name = DEFAULT_TARGET_MODEL_EAGLE3
    # CUDA 5090 + Llama-3.1-8B measured ~99 median in CI. AMD EAGLE3
    # currently sustains lower single-batch occupancy and needs a longer
    # measurement window to avoid too few non-NaN samples.
    fwd_occupancy_threshold = 80.0 if is_in_amd_ci() else 97.0
    fwd_occupancy_max_new_tokens = 4096 if is_in_amd_ci() else 2048
```

评论区精华

无 review 讨论。审核者 bingxche 和 hnyls2002 均直接批准，未留下评论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅影响测试类 TestBasicSanityEagle3 的两个配置属性，不涉及任何运行时逻辑或其他模块。AMD 阈值 80.0 基于实际 CI runner 观察值设定，若后端性能提升可能需要后续重新校准。
- 影响：影响范围限定于 AMD 平台 EAGLE3 功能的 stage-a-test-1-gpu-small-amd CI 任务。修复后该测试不再跳过，保持对 EAGLE3 前向占用量的回归检测。CUDA 和其他平台行为完全不变。
- 风险标记：仅测试变更

关联脉络

- 暂无明显关联 PR