PR #20997 完整报告

sgl-project/sglang

[NPU] [Diffusion] Update CI performance baseline for Wan2.2-T2V-A14B-Diffusers-w8a8

合并时间：2026-03-20 20:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20997

执行摘要

该 PR 将 NPU 扩散模型 CI 性能测试中 TextEncodingStage 的基准时间从 301.21 毫秒大幅提升至 1200.21 毫秒，旨在解决因 CI 服务器性能差异导致的偶发性测试超时失败。这是一个典型的 CI 基础设施调整，虽能立即提高测试通过率，但可能掩盖性能回归且未解决根本问题，建议结合 review 中的讨论关注长期测试策略。

功能与动机

- 问题背景：multimodal-gen-test-8-npu-a3 测试中 TextEncodingStage 阶段有时超时失败，但在本地无法复现，推测某些 CI 服务器性能较慢。
- 解决目标：通过放宽性能基线，减少环境差异导致的测试不稳定性，提升 CI 流水线可靠性。
- 关键引用：PR body 中说明“i just update time limit”，直接体现了以调整基线为快速解决方案的思路。

实现拆解

仅修改一个配置文件，具体变更如下：

文件路径	变更内容	影响
python/sglang/multimodal_gen/test/server/ascend/perf_baselines_npu.json	"TextEncodingStage": 301.21 → "TextEncodingStage": 1200.21	将 NPU 上 Wan2.2-T2V-A14B-Diffusers-w8a8 模型的文本编码阶段基准时间提升约 4 倍

评论区精华

review 中仅有的实质性讨论来自 gemini-code-assist[bot]，要点包括：

> “This increases the performance baseline by nearly 4x. While this will fix the immediate CI failures, such a large adjustment might mask underlying performance regressions or significant variability in CI runner performance.”

> “The new limit of 1200.21 ms is also very close to the failing time of 1198.02 ms shown in the PR description, which might lead to continued flakiness.”

这些评论指出了调整策略的潜在风险：一是可能掩盖真实性能问题，二是新基线过于接近失败阈值，不稳定性可能持续。但讨论未深入，变更最终被批准。

风险与影响

- 性能回归风险：4 倍的基线放宽可能使实际性能退化不被 CI 检测，影响 NPU 扩散模型的质量监控。
- 持续不稳定性：新基线仅比观察到的失败时间高约 2 毫秒，环境微小波动仍可能导致测试失败。
- 根本原因未解决：未探究 CI 服务器性能差异的根源，问题可能在其他测试或配置中重现。
- 影响范围：仅影响特定 NPU 扩散模型的 CI 测试通过标准，对用户功能和系统性能无直接影响。

关联脉络

- 近期相关 PR：PR #22031 同样调整多模态生成 CI 测试，但采用临时禁用策略而非修改基线，反映团队在应对 CI 不稳定性时的不同方法。
- NPU 性能演进：结合 PR #19246（NPU 优化 GLM4.7）可见，NPU 硬件上的性能优化和测试校准是持续主题，本 PR 属于测试基础设施的配套调整。
- CI 稳定性趋势：近期多个 PR（如 #22001、#22011）涉及 CI 工作流修复，表明团队正系统性提升 CI 可靠性，本 PR 是这一趋势中的具体实践。