PR #23056 完整报告

sgl-project/sglang

[Diffusion][NPU][CI] update perf numbers

合并时间：2026-04-21 00:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23056

执行摘要

此 PR 更新了 NPU 扩散模型的性能基线文件，通过手动调整解码阶段和去噪步骤的预期耗时，临时修复了因性能退化导致的 CI 测试失败。变更仅涉及一个测试文件，但 review 中指出了数据不一致的风险，建议后续使用脚本重新生成基线以确保准确性。

功能与动机

近期 Ascend NPU 上的扩散模型性能出现显著下降（如关联 Issue #23055 所示），导致 CI 测试失败。此 PR 作为临时解决方案，更新了性能基线文件中的数值，使测试能够通过，而性能退化的根本原因将在 Issue 中进一步调查。PR body 明确说明："update perf numbers to fix CI runs performance degradation will be investigated in https://github.com/sgl-project/sglang/issues/23055"。

实现拆解

变更仅涉及一个文件：python/sglang/multimodal_gen/test/server/ascend/perf_baselines_npu.json。该文件存储了 NPU 扩散模型性能测试的基准耗时数据，用于 CI 测试中的回归检查。

具体修改包括：

1. 更新解码阶段耗时：针对两个测试场景（wan2_1_t2v_1.3b_1_npu 和 wan2_2_t2v_14b_w8a8_8npu），将 DecodingStage 的数值从较低值调整为较高值，反映当前性能退化。
2. 调整去噪步骤耗时：在 wan2_1_t2v_1.3b_1_npu 场景中，将第一个去噪步骤（denoise_step_ms["0"]）从 103.56 毫秒更新为 153.0 毫秒。
3. 未更新聚合指标：如以下代码片段所示，聚合指标如 DenoisingStage 和 expected_e2e_ms 未同步更新，这可能导致基线数据内部不一致，影响测试准确性。

关键源码片段

python/sglang/multimodal_gen/test/server/ascend/perf_baselines_npu.json

这是唯一变更的文件，包含了 NPU 扩散模型性能测试的基准耗时数据，直接决定 CI 测试的通过与否。

```
{
  "wan2_1_t2v_1.3b_1_npu": {
    "stages_ms": {
```

```
            "InputValidationStage": 0.07,
            "TextEncodingStage": 876.11,
            "LatentPreparationStage": 0.25,
            "TimestepPreparationStage": 2.9,
            "DenoisingStage": 26188.0, // 注意：此聚合指标未随阶段耗时更新，可能导致数据不一致
            "DecodingStage": 650.1, // 从 320.03 更新为 650.1，反映解码阶段耗时增加
            "per_frame_generation": null
        },
        "denoise_step_ms": {
            "0": 153.0, // 从 103.56 更新为 153.0，反映第一个去噪步骤耗时增加
            "1": 329.59,
            // ... 其他步骤保持不变
        },
        "expected_e2e_ms": 38738.17, // 注意：此聚合指标未更新，可能不反映最新阶段耗时总和
        "expected_avg_denoise_ms": 523.62,
        "expected_median_denoise_ms": 536.23
    },
    "wan2_2_t2v_14b_w8a8_8npu": {
        "stages_ms": {
            "InputValidationStage": 0.07,
            "TextEncodingStage": 1200.21,
            "LatentPreparationStage": 0.2,
            "TimestepPreparationStage": 2.68,
            "DenoisingStage": 83661.46, // 注意：此聚合指标未更新
            "DecodingStage": 1080.05, // 从 232.94 更新为 1080.05，解码耗时显著增加
            "per_frame_generation": null
        },
        // ... 其他部分保持不变
    }
}
```

评论区精华

review 中，gemini-code-assist[bot] 指出了关键问题：

> “The update to DecodingStage and denoise_step_ms["0"] is not reflected in the aggregate metrics... It is recommended to regenerate the baseline using the gen_perf_baselines.py script to ensure all derived values are consistent.”

此评论强调了手动更新基线文件的风险——未同步更新聚合指标会导致数据不一致，可能使 CI 测试误判性能回归。但 PR 作者未回应此问题，PR 已合并，留下了潜在的技术债务。

风险与影响

- 测试准确性风险：由于聚合指标未更新，性能基线文件内部不一致，可能导致 CI 测试误报通过或失败，掩盖真实的性能退化程度。
- 维护风险：手动修改而非脚本生成基线文件，增加了未来更新时出错的风险，尤其是在多指标需同步调整的场景下。

- 影响范围：此变更仅影响 NPU 扩散模型的 CI 测试流程，对用户和系统运行时无直接影响，但确保了测试链的连续性，为后续性能调查（Issue #23055）争取了时间。

关联脉络

此 PR 与近期多个 diffusion 和 NPU 相关 PR 有间接关联：

- PR #23118 和 #23207 涉及 diffusion 模块的优化和重构，可能影响整体性能趋势。
- PR #22914 涉及 NPU 和上下文并行的代码去重，可能反映团队在性能优化上的持续努力。
- 关联 Issue #23055 提供了性能退化的背景，表明此 PR 是临时修复，后续需深入调查根本原因。 整体来看，此 PR 是 NPU 性能维护链条中的一环，突出了在快速迭代中平衡 CI 稳定性和数据准确性的挑战。