PR #34856 完整报告

sgl-project/sglang

fix tpot by adjusting the sliding max-prefill-size window size

合并时间：2026-08-17 09:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34856

执行摘要

- 一句话：调整 NPU 性能用例 prefill delayer 窗口参数修复 TPOT
- 推荐动作：不建议精读，属于纯测试参数调优。值得留意的只有一点：prefill delayer 的滑动窗口大小会显著影响 NPU 上长序列场景的 decode pass 稳定性，后续若有平台调度调整，可参考本 PR 的窗口与 max-delay-passes 配套组合。

功能与动机

PR body 明确指出动机是 "prevent too fast max-prefill size decay to effect the decode pass"，即滑动窗口过小会导致 max-prefill size 衰减过快，进而影响 decode 阶段的性能表现（TPOT）。因此需要为NPU平台各性能用例放宽滑动窗口大小，并调整配套的延迟轮数参数。

实现拆解

1. 明确参数语义：SGLANG_PREFILL_DELAYER_MAX_PREFILL_BS_WINDOW_SIZE 控制 prefill batch size 滑动窗口大小，窗口越大，max-prefill size 衰减越平缓；SGLANG_PREFILL_DELAYER_MAX_DELAY_PASSES 和 --prefill-delayer-max-delay-passes 控制 prefill 延迟调度的最大轮数。本次改动围绕这两类参数展开。
2. 逐用例调整参数组合（全部位于 test/registered/npu/performance/）：
   - minimax_m2_5 用例：窗口 64 → 128，并在 --enable-prefill-delayer 后新增 --prefill-delayer-max-delay-passes 128；
   - qwen3_6_27b 1080p 多模态用例：窗口 64 → 128，同时在启动参数中新增 --prefill-delayer-max-delay-passes 300；
   - qwen3_6_27b 1024 多模态用例：环境变量 SGLANG_PREFILL_DELAYER_MAX_DELAY_PASSES 150 → 300，并新增窗口 128；
   - qwen3_235b_a22b 8 卡用例：SGLANG_PREFILL_DELAYER_MAX_DELAY_PASSES 100 → 50，并新增较小窗口 16（与其他用例方向相反）；
   - qwen3-8b、qwen3_30b_a3b、qwen3_32b 三个用例：仅新增窗口 128，其余参数不变。
3. 参数组合意图：除 235B 外，扩大窗口意味着 max-prefill size 的下降更平滑，避免 prefill batch 快速收缩导致 decode pass 的显存与调度抖动；235B 采用更小窗口 16 并降低 delay passes，应是针对 8 卡大规模 EP 场景单独验证过的配置。

4. 测试配套说明：所有用例均通过 register_npu_ci 注册到对应 CI 套件（如 base-c-test-perf-8-npu-a3、base-c-test-perf-2-npu-a3、full-2-npu-a3 等），无新增单元测试，也无配置、schema 或部署配套改动；CI 由 sglang-npu-bot 以 /tag-and-rerun-ci 触发验证。

关键文件：

- test/registered/npu/performance/minimax_m2_5/test_npu_minimax_m2_5_w8a8_4p_in64k_out1k_prefix90_50ms.py（模块 性能测试；类别 test；类型 test-coverage）：改动量最大的用例（+3/-1），将滑动窗口从 64 扩到 128 并新增 --prefill-delayer-max-delay-passes 128，是本 PR 的核心样例。
- test/registered/npu/performance/qwen3_6_27b/test_npu_qwen3_6_27b_1p_in1080p_30_out256_50ms.py（模块 性能测试；类别 test；类型 test-coverage）：改动量同样为 +3/-1，窗口 64 → 128 并在命令行显式新增 --prefill-delayer-max-delay-passes 300，覆盖多模态长文本场景。
- test/registered/npu/performance/qwen3_235b_a22b/test_npu_qwen3_235b_w8a8_8p_in3k5_out1k5_50ms.py（模块 性能测试；类别 test；类型 test-coverage）：参数组合与其它用例相反（delay passes 100 → 50、窗口仅 16），是 8 卡大规模场景下的特殊配置，值得关注其 CI 稳定性。
- test/registered/npu/performance/qwen3_6_27b/test_npu_qwen3_6_27b_1p_in1024x1024_30_out1024_50ms.py（模块 性能测试；类别 test；类型 test-coverage）：环境变量 SGLANG_PREFILL_DELAYER_MAX_DELAY_PASSES 从 150 抬到 300，并新增窗口 128，覆盖 1024x1024 图像输入场景。
- test/registered/npu/performance/qwen3-8b/test_npu_qwen3_8b_w8a8_1p_in3k5_out1k5_50ms.py（模块 性能测试；类别 test；类型 test-coverage）：仅新增窗口 128，是纯窗口扩大的最小改动样例。
- test/registered/npu/performance/qwen3_30b_a3b/test_npu_qwen3_30b_w8a8_1p_in3k5_out1k5_50ms.py（模块 性能测试；类别 test；类型 test-coverage）：仅新增窗口 128，与 8B/32B 用例保持一致。
- test/registered/npu/performance/qwen3_32b/test_npu_qwen3_32b_w8a8_2p_in3k5_out1k5_50ms.py（模块 性能测试；类别 test；类型 test-coverage）：仅新增窗口 128，与 8B/30B 用例保持一致。

关键符号：未识别

关键源码片段

test/registered/npu/performance/minimax_m2_5/test_npu_minimax_m2_5_w8a8_4p_in64k_out1k_prefix90_50ms.py

改动量最大的用例（+3/-1），将滑动窗口从 64 扩到 128 并新增 --prefill-delayer-max-delay-passes 128，是本 PR 的核心样例。

```
# MiniMax-M2.5 用例：滑动窗口从 64 上调到 128，并显式设置最大延迟轮数为 128。
# 目的是让 max-prefill size 衰减更平缓，避免 decode pass 被过快收缩的 prefill batch 拖累。
MINIMAX_M2_5_W8A8_4P_IN64K_OUT1K_PREFIX90_ENVS = {
```

```
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "STREAMS_PER_DEVICE": "32",
    "HCCL_SOCKET_IFNAME": "lo",
    "GLOO_SOCKET_IFNAME": "lo",
    # ... 其余环境变量省略 ...
    "SGLANG_PREFILL_DELAYER_MAX_PREFILL_BS_WINDOW_SIZE": "128", # 原值为 64
    "PYTHONPATH": f"{MINIMAX_M2_5_EAGLE3_MODEL_PATH}:{os.environ.get('PYTHONPATH',
    '')}",
}

MINIMAX_M2_5_W8A8_4P_IN64K_OUT1K_PREFIX90_OTHER_ARGS = [
    # ... 其余启动参数省略 ...
    "--enable-prefill-delayer",
    "--prefill-delayer-max-delay-passes", # 新增：限制 prefill 延迟调度的最大轮数，配合窗口扩大
    128,
    "--prefill-max-requests",
    10,
    # ... 其余启动参数省略 ...
]
```

test/registered/npu/performance/qwen3_6_27b/test_npu_qwen3_6_27b_1p_in1080p_30_out256_50ms.py

改动量同样为 +3/-1，窗口 64 → 128 并在命令行显式新增 --prefill-delayer-max-delay-passes 300，覆盖多模态长文本场景。

```
# Qwen3-6/27B 多模态用例（1080p）：窗口从 64 扩到 128，并显式传入最大延迟轮数 300。
# 该用例使用 NEXTN 推测解码，放大窗口有助于稳定长序列 prefill 到 decode 的过渡。
QWEN3_6_27B_1080P_ENVS = {
    "STREAMS_PER_DEVICE": "32",
    "HCCL_SOCKET_IFNAME": "lo",
    "GLOO_SOCKET_IFNAME": "lo",
    # ... 其余环境变量省略 ...
    "SGLANG_PREFILL_DELAYER_MAX_PREFILL_BS_WINDOW_SIZE": "128", # 原值为 64
    "ASCEND_USE_FIA": "1",
}

QWEN3_6_27B_1080P_OTHER_ARGS = [
    # ... 其余启动参数省略 ...
    "--enable-prefill-delayer",
    "--prefill-delayer-queue-min-ratio",
    0.45,
    "--prefill-delayer-max-delay-ms",
    5500,
    "--prefill-delayer-max-delay-passes", # 新增：显式配置最大延迟轮数 300
    300,
    "--enable-multimodal",
    # ... 其余启动参数省略 ...
]
```

评论区精华

本 PR 没有实质性的 review 技术讨论：唯一的 Issue 评论是 sglang-npu-bot 触发的 CI 指令 /tag-and-rerun-ci，审核状态为 sglang-npu-bot APPROVED，无人工 reviewer 评论。因此不存在未解决的争议点。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低，因为全部变更仅作用于测试用例的配置常量，不触碰产品代码路径。需要关注的具体点：
  - qwen3_235b_a22b 用例与其他 6 个用例的参数方向相反（delay passes 从 100 降至 50、窗口仅 16），若该组合是在小窗口下的激进策略，且 CI 环境波动时可能重新出现 TPOT 抖动，需要观察该 suite 的历史稳定性；
  - 窗口扩大后 prefill 延迟时间变长，可能影响同一 CI suite 中其他用例的排队时长，但不会影响产品行为；
  - 这些性能用例多为 nightly 或特定 NPU 套件，改动不覆盖 GPU 等其他硬件平台，无跨平台回归面。
  - 影响：影响范围限于 NPU 平台的性能 CI 套件（base-c-test-perf-8-npu-a3、base-c-test-perf-2-npu-a3、full-2-npu-a3 等），涉及 Minimax-M2.5、Qwen3-8B/30B/32B/235B、Qwen3-6/27B 多模态等模型场景的 TPOT 指标。对最终用户无影响，不改变任何服务端参数默认值或推理逻辑；对团队而言，该 PR 是 NPU 性能基准测试的配置维护，能提升 CI 对 prefill/decode 调度抖动的稳定性，但洞察价值有限。
  - 风险标记：仅测试参数调整，无产品代码影响，235B 用例参数方向相反

关联脉络

- PR #34996 Increase post-capture decode memory reserve: 同为通过调整调度 / 内存相关参数改善 decode 阶段性能的改动，与本文的 TPOT 修复目标一致。
- PR #34998 Add explicit EPLB balancedness reporting modes: 同为 NPU 平台下的调度参数与可观测性调整，涉及 server_args 与 NPU 测试，属于同一硬件平台的配置演进脉络。
- PR #34995 [VLM] Avoid synchronizing multimodal placeholder counts: 同为调度热路径性能优化，且涉及多模态调度逻辑，与 Qwen3-6/27B 多模态性能用例的稳定性目标相关。