PR #23423 完整报告

sgl-project/sglang

[NPU] Fix mrope_position computation in Eagle Worker v2 with PlanStream

合并时间：2026-05-11 09:43

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23423

执行摘要

- 一句话：修复 NPU 推测解码中 mrope_position 竞态条件
- 推荐动作：值得精读。展示了异步流竞态修复的典型模式：在等待流同步后重新计算依赖值，并正确同步到 CUDA graph buffer。对理解 speculative decoding 中的流管理和 CUDA graph 缓冲有参考价值。

功能与动机

PR body 说明：'When SGLANG_ENABLE_SPEC_V2 and SGLANG_ENABLE_OVERLAP_PLAN_STREAM are enabled in multimodal scenarios, the Eagle Worker v2 computes mrope_position on the plan stream using draft outputs from the default stream. This race condition leads to incorrect mrope_position values.'

实现拆解

1. 方法公有化：在 forward_batch_info.py 中将私有方法 _compute_spec_mrope_positions 重命名为公有方法 compute_spec_mrope_positions，使其可以被外部模块调用。
2. 核心修复：在 eagle_worker_v2.py 的 verify 方法中，等待 plan stream 完成后，检测 NPU 后端、mrope 模型等条件，如果满足则调用 compute_spec_mrope_positions 重新计算正确的 mrope_position，覆盖 plan stream 中因竞态产生的错误值。
3. Graph buffer 同步：在 npu_graph_runner.py 的 replay 方法中，当启用 SGLANG_ENABLE_OVERLAP_PLAN_STREAM 且 forward_batch.mrope_positions 不为 None 时，将其拷贝到 self.buffers.mrope_positions，确保 CUDA graph 使用正确的 mrope_position。

关键文件：

- python/sglang/srt/model_executor/forward_batch_info.py（模块 前向批信息；类别 source；类型 data-contract；符号 _compute_spec_mrope_positions, compute_spec_mrope_positions）：方法重命名，将私有方法改为公有，是修复的前提条件。
- python/sglang/srt/hardware_backend/npu/graph_runner/npu_graph_runner.py（模块 NPU 图执行器；类别 source；类型 dependency-wiring；符号 replay）：新增在 replay 中将 mrope_positions 同步到 graph buffer，确保 CUDA graph 使用正确值。
- python/sglang/srt/speculative/eagle_worker_v2.py（模块 推测解码；类别 source；类型 core-logic；符号 verify）：核心修复逻辑：在 verify 中等待 plan stream 后重新计算

mrope_position，覆盖错误值。

关键符号：compute_spec_mrope_positions, verify, replay

关键源码片段

python/sglang/srt/speculative/eagle_worker_v2.py

核心修复逻辑：在 verify 中等待 plan stream 后重新计算 mrope_position，覆盖错误值。

```
# eagle_worker_v2.py verify 方法片段
# 等待 plan stream 完成同步
if self.plan_stream:
    torch.get_device_module(self.device).current_stream().wait_stream(
        self.plan_stream
    )
    # 修复：在 NPU + mrope + spec_info 非 idle 时，重新计算 mrope_position
    # 因为 plan stream 使用的 draft outputs 可能来自 default stream，
    # 导致竞态错误，此处确保在 default stream 上正确计算。
    if (
        _is_npu
        and self._target_worker.model_runner.model_is_mrope
        and batch.spec_info is not None
        and getattr(batch.spec_info, "positions", None) is not None
        and not batch.forward_mode.is_idle()
    ):
        verify_forward_batch.compute_spec_mrope_positions(
            self._target_worker.model_runner, batch
        )
```

评论区精华

- Todobe：建议将 mrope_position 重计算代码放在 wait_stream 之后以防止流冲突。（❓ 已修复）
- Todobe：建议先判断 _is_npu 以优化性能。（❓ 已调整顺序）
- Hexq0210：询问为什么将私有方法改为公有方法。silencejade 回复：因为需要作为公有函数在外部调用。
- 代码放置位置 (correctness): silencejade 已修复，将代码放在 wait_stream 后。
- 条件判断顺序 (performance): silencejade 已调整顺序。
- 方法重命名 (design): silencejade 解释因为需要作为公有函数在外部调用。

风险与影响

- 风险：变更仅影响 NPU 后端且必须同时启用 SPEC_V2、OVERLAP_PLAN_STREAM 和多模态输入，影响面窄。但该方法调用通过条件判断动态执行，未来若有新增调用点需确保正确性。缺少明确的单元测试覆盖该竞态场景。
- 影响：修复了特定配置下多模态推测解码的数值错误，用户无需修改配置即可获得正确结果。性能影响极小，仅增加一次同步后的额外计算和少量数据传输。

- 风险标记：特定配置触发，缺少测试覆盖，NPU 后端路径

关联脉络

- PR #23456 [SPEC V2] fix: skip stale state updates in spec-v2 overlap: 同属 speculative-decoding v2 重叠调度修复，共享 verify 流程中的重叠逻辑。
- PR #23819 [NPU] Fix warmup error with --disable-cuda-graph and mtp: 均为 NPU 后端 bugfix，共享硬件后端背景。