

PR #32555 完整报告

sgl-project/sglang

[Nemotron] Fix decode track-save reading the stale tail of the CUDA-graph track buffer

合并时间: 2026-07-30 10:03

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32555>

执行摘要

- 一句话: 修复 Mamba2 CUDA 图回放跟踪缓冲区尾部脏读
- 推荐动作: 此修复定位精准、改动最小、测试充分, 建议合并。开发者在合并前应确认所有相关 CI 通过。

功能与动机

Mamba2 混合模型 (如 NemotronH) 在启用 CUDA graphs、mamba_track_interval 边界跨越后, decoded 请求的跟踪状态被错误写入 radix 树, 导致后续 cache-hit 重新评分出现 KL ~1.55 的静默精度偏差 (禁用 CUDA graphs 时 KL ~0.002)。PR body 明确指出问题: No crash, no assert — silent corruption。

实现拆解

1. 修复核心逻辑: 在 `python/sglang/srt/layers/attention/hybrid_linear_attn_backend.py` 的 `_replay_metadata` 方法中, 将 `track_buf = self.mamba_track_indices_buf` 改为 `self.mamba_track_indices_buf[:bs]`, 确保回放时只传递已刷新的前 `bs` 个元素, 避免 Mamba2 后端切片 `[-num_decodes:]` 时读取未刷新的尾部。同时添加 `len(mamba_track_indices) >= bs` 断言, 防止传入的跟踪索引不足。
2. 新增回归测试: 在 `test/registered/radix_cache/test_mamba2_extra_buffer_kl.py` 中创建 `TestMamba2ExtraBufferKL` 类, 使用 Nemotron-Nano-9B-v2 模型和 `extra_buffer` 策略, 设置 `kl_div_thres=0.005` 检测 decode-seeded cache-hit 的 KL 偏差。测试注册为 `extra-a` 阶段 CI。
3. 测试参数调优: 设置 `--mamba-track-interval 128` (默认减半), 确保 512-token 的 decode 序列能跨越跟踪边界, 触发修复场景。

关键文件:

- `python/sglang/srt/layers/attention/hybrid_linear_attn_backend.py` (模块 注意力后端; 类别 source; 类型 core-logic; 符号 `_replay_metadata`): 核心修复文件, 仅修改一行关键逻辑 (`track_buf` 切片), 附带断言保护。
- `test/registered/radix_cache/test_mamba2_extra_buffer_kl.py` (模块 Mamba2 测试; 类别 test; 类型 test-coverage; 符号 `TestMamba2ExtraBufferKL`): 新增的回归测试文件, 使用 Nemotron-Nano-9B-v2 模型验证修复正确性, 注册为 CI。

关键符号: `_replay_metadata`

关键源码片段

python/sglang/srt/layers/attention/hybrid_linear_attn_backend.py

核心修复文件，仅修改一行关键逻辑（track_buf 切片），附带断言保护。

```
# python/sglang/srt/layers/attention/hybrid_linear_attn_backend.py
# _replay_metadata 方法中跟踪缓冲区处理部分（CUDA graph 回放路径）

# 刷新静态跟踪目标缓冲区（已翻译）；捕获的 track-save kernel
# 读取该缓冲区，保留传入的 InputBuffer 槽位为只读。
# 仅返回已刷新的 [:bs] 前缀 —— Mamba2 的 track-save 切片
# [-num_decodes:]，若传递完整 max_bs 缓冲区会绑定陈旧的尾部。
track_buf = None
if mamba_track_indices is not None:
    assert (
        len(mamba_track_indices) >= bs
    ), f"{len(mamba_track_indices)=} < {bs=}"
    track_buf = self.mamba_track_indices_buf[:bs]
    track_buf.copy_(self._translate_mamba_indices(mamba_track_indices[:bs]))
```

test/registered/radix_cache/test_mamba2_extra_buffer_kl.py

新增的回归测试文件，使用 Nemotron-Nano-9B-v2 模型验证修复正确性，注册为 CI。

```
# test/registered/radix_cache/test_mamba2_extra_buffer_kl.py
import unittest

from sglang.test.ci.ci_register import register_cuda_ci
from sglang.test.kits.kl_divergence_kit import KLDivergenceMixin
from sglang.test.server_fixtures.default_fixture import DefaultServerBase

register_cuda_ci(est_time=600, stage="extra-a", runner_config="1-gpu-large")

class TestMamba2ExtraBufferKL(KLDivergenceMixin, DefaultServerBase):
    """NemotronH (Mamba2) + extra_buffer: cache-hit logprobs match cold recompute."""

    # 使用 Mamba2 架构模型触发尾部分片错误
    model = "nvidia/NVIDIA-Nemotron-Nano-9B-v2"

    # decode-seeded 重用是回归触发源（图回放 track-save）；
    # 损坏路径 KL ~1.5，阈值 0.005 可明确区分
    kl_div_thres = 0.005
    kl_div_thres_prefill = 0.01
    kl_div_max_samples = 16

    # extra_buffer 策略 + 降低 track-interval 确保边界跨越
    other_args = [
        "--max-mamba-cache-size", "256",
        "--mem-fraction-static", "0.8",
```

```
    "--mamba-scheduler-strategy", "extra_buffer",  
    "--mamba-track-interval", "128", # 默认 256 减半, 确保 > mamba chunk size  
]
```

```
if __name__ == "__main__":  
    unittest.main()
```

评论区精华

无 review 评论。PR 被 alexnails 直接批准，CI rerun 请求均成功。

- 暂无高价值评论线程

风险与影响

- 风险：修复范围仅限 CUDA graph 回放路径中跟踪缓冲区的切片操作，对 eager 路径无影响（原有逻辑正确），对 GDN/KDA 等前端索引架构无影响。断言 `len(mamba_track_indices) >= bs` 可能捕获先前未暴露的编码错误，不会导致静默回归。主要风险是 `_replay_metadata` 的其他调用点未传递足够的 `mamba_track_indices`，但代码中该参数由调用方保证。
- 影响：直接影响 Mamba2 架构的混合模型（如 NemotronH、Nemotron-3-Super-120B）在 CUDA graphs 下的 decode 精度和 cache 命中率，修复前静默错误可能导致用户获得错误 logprobs 或生成质量下降。对其他模型无影响。
- 风险标记：CUDA graphs 路径变更，涉及 Mamba2 架构，新增断言可能暴露新问题

关联脉络

- PR #31833 fix(mamba): multi-request prefill offset defect in `_init_track_ssm_indices`: PR body 指出本修复与 #31833 不同但相关：#31833 修复多请求 prefill 偏移缺陷，本 PR 修复单请求图回放 decode 错误。