

PR #29973 完整报告

sgl-project/sglang

[FIX] Prevent Lightning Attention extra-buffer mamba state corruption

合并时间: 2026-07-22 22:01

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29973>

执行摘要

- 一句话: 修复 Lightning Attention 与 mamba extra-buffer 的状态损坏
- 推荐动作: 此 PR 修复了一个微妙且难以发现的竞态问题, 涉及 radix cache、mamba 状态和 Triton 内核的交互。PR body 中的 ASCII 图示和分析非常透彻, 建议仔细阅读。

功能与动机

当 Lightning seg_la 与 mamba radix cache extra_buffer 一起使用时, 重复的服务运行会重用损坏的 mamba 状态快照, 导致 AIME 准确率崩溃。在 SGLang 中, mamba radix cache 必须保持 token/KV 前缀和循环 mamba 状态对齐。bug 在于 seg_la 没有为未对齐的扩展实现所需的中间循环状态, 导致调度器存储的 radix 条目获得了整个扩展的最终状态, 从而产生错配。

实现拆解

修复分解为三步:

1. 在 seg_la 内核中添加可选跟踪状态写入: 修改 seg_la_p_kernel, 当 TRACK_STATE 编译常量使能时, 在每个块迭代后检查 $n + b == \text{track_len}$ 并将当前 state 写入额外跟踪槽。
2. 在 Lightning 后端中准备跟踪信息: 新增 _prepare_seg_la_track_store 方法, 从 ForwardBatch 中提取跟踪相关字段, 计算边界长度和槽索引, 通过 forward_metadata 传递给 _linear_attention_entry 再传入 seg_la_fwd。
3. 添加内核级回归测试: 在 test_triton.py 中新增 test_seg_la_prefill_tracks_extra_buffer_state 方法, 验证单请求扩展场景下三个状态槽的正确性。

关键文件:

- python/sglang/srt/layers/attention/linear/lightning_backend.py (模块 注意力后端; 类别 source; 类型 dependency-wiring; 符号 _prepare_seg_la_track_store): 核心后端修改: 新增 _prepare_seg_la_track_store 方法, 修改 init_forward_metadata 和 init_forward_metadata_out_graph 以传播跟踪元数据。
- python/sglang/kernels/ops/attention/linear/seg_la.py (模块 Triton 内核; 类别 source; 类型 infrastructure): Triton 内核修改: 在预填充内核中添加可选跟踪状态写入, 通过 TRACK_STATE 编译常量控制。

- test/registered/attention/unittests/lightning/test_triton.py (模块 Triton 测试; 类别 test; 类型 test-coverage; 符号 test_seg_la_prefill_tracks_extra_buffer_state) : 测试文件 : 新增针对 unaligned extra-buffer 路径的回归测试。

关键符号: LightningAttentionBackend._prepare_seg_la_track_store,
LightningAttentionBackend.init_forward_metadata,
LightningAttentionBackend.init_forward_metadata_out_graph, seg_la_fwd,
seg_la_p_kernel

评论区精华

本 PR 无公开 review 讨论。关联的 Issue #29945 已先行处理了延迟 mamba COW/clear 顺序问题。合并者 BBuf 直接批准了 PR。

- PR 审查 (other): 无争议, 直接合并。

风险与影响

- 风险:

1. 回归风险: Triton 内核新增条件分支, 但仅在 extra-buffer 场景使能; 所有 seg_la_fwd 调用都会增加参数传递开销, 但极小。
2. 功能影响: 仅影响 linear_backend=seg_la 且 mamba-scheduler-strategy=extra_buffer 的配置。
3. 测试覆盖: 新增内核级单元测试, 但未覆盖 CUDA 图、多批次等复杂场景。
4. 兼容性: 依赖 get_server_args() 获取 mamba_cache_chunk_size, 第二个 commit 已适配新访问器。
 - 影响: 用户影响: 使用 Lightning+extra-buffer 的用户将修复准确率崩溃问题; 其他用户无感知。系统影响: 增加对 ForwardBatch.mamba_track_* 字段的依赖, 非跟踪场景快速短路。团队影响: 提供了清晰的跟踪元数据传播模式和测试范例。
 - 风险标记: Triton 内核修改, 特定配置依赖, 条件编译分支

关联脉络

- PR #29945 Move deferred mamba cow and clear: 本 PR 依赖 #29945 先行修复了延迟 mamba 操作顺序问题, 避免了在 Lightning 后端中重复处理 COW/clear。本 PR 聚焦于 seg_la 的跟踪状态损坏。