

PR #33722 完整报告

sgl-project/sglang

[KDA] Fused-accept state advance for FlashInfer KDA MTP verify

合并时间: 2026-08-31 16:11

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33722>

执行摘要

- 一句话: KDA MTP verify 融合接受状态推进, SM100 端到端提速最高 62.5%
- 推荐动作:

功能与动机

PR body 指出 KDA MTP verify 在 FlashInfer 路径上是两阶段的: verify 内核把每个 draft token 的 post-state 写入 speculative intermediate_ssm scratch, 随后 commit_mamba_states_after_verify 将 accepted checkpoint 以全层 scatter 写回持久化池 temporal, 下一次 verify 再读回作为 seed, 这次 scatter 每请求每轮都发生, 数据量约 $\text{num_kda_layers} \times \text{HV} \times \text{V} \times \text{K} \times 4\text{B}$ (HV=16、K=V=128 fp32 时每层约 4 MB), 还伴随一次 kernel launch 与 host dispatch, 全部卡在 bs=1 MTP 的 inter-phase 接缝上。FlashInfer 的 recurrent_kda 已内置 num_accepted_tokens 参数 (fused speculative decode) ,

实现拆解

1. 核心源码路径 - `python/sglang/srt/layers/attention/linear/kda_backend.py` (core-logic) : 源码主路径; 涉及符号 `_can_fuse_accept_state`, `_fused_accept_indices`; 包含 导入关系调整、控制流调整、配置键调整; +108/-0; 关键符号: `_can_fuse_accept_state`, `_fused_accept_indices` - `python/sglang/srt/layers/attention/linear/kernels/kda_flashinfer.py` (core-logic) : 源码主路径; 涉及符号 `build_fused_accept_indices`; 包含 控制流调整、配置键调整、异常路径调整; +76/-28; 关键符号: `build_fused_accept_indices` - `python/sglang/srt/layers/attention/hybrid_linear_attn_backend.py` (core-logic) : 源码主路径; 包含 控制流调整、配置键调整; +31/-0
2. 测试与验证 - `test/registered/attention/unittests/hybrid_linear/test_kda_fused_accept_indices.py` (test-coverage) : 测试配套; 涉及符号 `TestBuildFusedAcceptIndices`, `test_real_slots_address_their_scratch_rows`, `test_padded_slot_rows_stay_fully_negative`, `test_nat_gather_reads_pool_values`; 包含 测试覆盖调整、导入关系调整、控制流调整; +141/-0; 关键符号: `TestBuildFusedAcceptIndices`, `test_real_slots_address_their_scratch_rows`, `test_padded_slot_rows_stay_fully_negative`, `test_nat_gather_reads_pool_values`

关键文件:

- python/sglang/srt/layers/attention/linear/kda_backend.py (模块 sglang/srt; 类别 source; 类型 core-logic; 符号 _can_fuse_accept_state, _fused_accept_indices) : 源码主路径; 涉及符号 _can_fuse_accept_state, _fused_accept_indices; 包含 导入关系调整、控制流调整、配置键调整; +108/-0
- test/registered/attention/unittests/hybrid_linear/test_kda_fused_accept_indices.py (模块 kda/fused/accept/indices; 类别 test; 类型 test-coverage; 符号 TestBuildFusedAcceptIndices, test_real_slots_address_their_scratch_rows, test_padded_slot_rows_stay_fully_negative, test_nat_gather_reads_pool_values) : 测试配套; 涉及符号 TestBuildFusedAcceptIndices, test_real_slots_address_their_scratch_rows, test_padded_slot_rows_stay_fully_negative, test_nat_gather_reads_pool_values; 包含 测试覆盖调整、导入关系调整、控制流调整; +141/-0
- python/sglang/srt/layers/attention/linear/kernels/kda_flashinfer.py (模块 sglang/srt; 类别 source; 类型 core-logic; 符号 build_fused_accept_indices) : 源码主路径; 涉及符号 build_fused_accept_indices; 包含 控制流调整、配置键调整、异常路径调整; +76/-28
- python/sglang/srt/layers/attention/hybrid_linear_attn_backend.py (模块 sglang/srt; 类别 source; 类型 core-logic) : 源码主路径; 包含 控制流调整、配置键调整; +31/-0
- python/sglang/srt/layers/attention/mamba/mamba2_metadata.py (模块 sglang/srt; 类别 source; 类型 core-logic) : 源码主路径; 包含 配置键调整; +7/-0

关键符号: _can_fuse_accept_state, _fused_accept_indices, TestBuildFusedAcceptIndices, test_real_slots_address_their_scratch_rows, test_padded_slot_rows_stay_fully_negative, test_nat_gather_reads_pool_values, TestFusedAcceptPerForwardCache, _build, test_shared_build_matches_a_per_layer_build, test_a_different_batch_builds_different_rows

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR