

PR #35245 完整报告

sgl-project/sglang

refactor(unified-memory): translate the KV write location once, at ForwardBatch construction

合并时间: 2026-08-31 14:52

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35245>

执行摘要

- 一句话:
- 推荐动作:

功能与动机

实现拆解

1. 核心源码路径 - `python/sglang/srt/mem_cache/kv_index_translator.py` (dependency-wiring) : 源码主路径; 涉及符号 `KVReadTables`, `KVIndexTable`, `sliding_window_read_ids`, `KVIndexTranslator`; 包含 导入关系调整、控制流调整、配置键调整; +358/-0; 关键符号: `KVReadTables`, `KVIndexTable`, `sliding_window_read_ids`, `KVIndexTranslator` - `python/sglang/srt/layers/attention/triton_backend.py` (core-logic) : 源码主路径; 涉及符号 `_translate_cuda_graph_shared_pool_locs`, `_fill_cuda_graph_write_locs`; 包含 导入关系调整、控制流调整、配置键调整; +91/-142; 关键符号: `_translate_cuda_graph_shared_pool_locs`, `_fill_cuda_graph_write_locs` - `python/sglang/srt/mem_cache/memory_pool.py` (dependency-wiring) : 源码主路径; 包含 导入关系调整、控制流调整、配置键调整; +35/-26
2. 测试与验证 - `test/registered/unit/model_executor/test_unified_out_cache_loc_rebind.py` (test-coverage) : 测试配套; 涉及符号 `_make_fb`, `_armed_source`, `_call_names`, `TestForwardBatchWiring`; 包含 测试覆盖调整、导入关系调整、控制流调整; +181/-0; 关键符号: `_make_fb`, `_armed_source`, `_call_names`, `TestForwardBatchWiring` - `test/registered/unit/mem_cache/test_kv_index_translator.py` (test-coverage) : 测试配套; 涉及符号 `_build_composite`, `_make_source`, `_reference_table`, `TestPassthrough`; 包含 测试覆盖调整、导入关系调整、控制流调整; +684/-0; 关键符号: `_build_composite`, `_make_source`, `_reference_table`, `TestPassthrough` - `test/registered/unit/server_args/test_unified_prefill_cuda_graph_gate.py` (test-coverage) : 测试配套; 涉及符号 `_run_handler`, `TestUnifiedPrefillCudaGraphGate`, `test_default_prefill_capture_is_auto_disabled`, `test_explicit_prefill_backend_is_refused`; 包含 测试覆盖调整、导入关系调整、控制流调整; +92/-0; 关键符号: `_run_handler`, `TestUnifiedPrefillCudaGraphGate`, `test_default_prefill_capture_is_auto_disabled`, `test_explicit_prefill_backend_is_refused`

3. 配置与运行环境 - [python/sglang/kernels/ops/kvcache/kv_read_table.py](#) (infrastructure) : 部署 / 基础设施; 涉及符号 `build_kv_read_table_kernel`, `build_kv_read_table`; 包含 导入关系调整、控制流调整、配置键调整; +140/-0; 关键符号 : `build_kv_read_table_kernel`, `build_kv_read_table` -
- [python/sglang/kernels/ops/kvcache/kv_indices.py](#) (infrastructure) : 部署 / 基础设施; 包含 控制流调整、配置键调整; +34/-10

关键文件:

- [python/sglang/srt/mem_cache/kv_index_translator.py](#) (模块 `sglang/srt`; 类别 `source`; 类型 `dependency-wiring`; 符号 `KVReadTables`, `KVIndexTable`, `sliding_window_read_ids`, `KVIndexTranslator`) : 源码主路径; 涉及符号 `KVReadTables`, `KVIndexTable`, `sliding_window_read_ids`, `KVIndexTranslator`; 包含 导入关系调整、控制流调整、配置键调整; +358/-0
- [python/sglang/srt/layers/attention/triton_backend.py](#) (模块 `sglang/srt`; 类别 `source`; 类型 `core-logic`; 符号 `_translate_cuda_graph_shared_pool_locs`, `_fill_cuda_graph_write_locs`) : 源码主路径; 涉及符号 `_translate_cuda_graph_shared_pool_locs`, `_fill_cuda_graph_write_locs`; 包含 导入关系调整、控制流调整、配置键调整; +91/-142
- [test/registered/unit/model_executor/test_unified_out_cache_loc_rebind.py](#) (模块 `unified/out/cache/loc/rebind`; 类别 `test`; 类型 `test-coverage`; 符号 `_make_fb`, `_armed_source`, `_call_names`, `TestForwardBatchWiring`) : 测试配套; 涉及符号 `_make_fb`, `_armed_source`, `_call_names`, `TestForwardBatchWiring`; 包含 测试覆盖调整、导入关系调整、控制流调整; +181/-0
- [test/registered/unit/mem_cache/test_kv_index_translator.py](#) (模块 `kv/index/translator`; 类别 `test`; 类型 `test-coverage`; 符号 `_build_composite`, `_make_source`, `_reference_table`, `TestPassthrough`) : 测试配套; 涉及符号 `_build_composite`, `_make_source`, `_reference_table`, `TestPassthrough`; 包含 测试覆盖调整、导入关系调整、控制流调整; +684/-0
- [test/registered/unit/server_args/test_unified_prefill_cuda_graph_gate.py](#) (模块 `unified/prefill/cuda/graph/gate`; 类别 `test`; 类型 `test-coverage`; 符号 `_run_handler`, `TestUnifiedPrefillCudaGraphGate`, `test_default_prefill_capture_is_auto_disabled`, `test_explicit_prefill_backend_is_refused`) : 测试配套; 涉及符号 `_run_handler`, `TestUnifiedPrefillCudaGraphGate`, `test_default_prefill_capture_is_auto_disabled`, `test_explicit_prefill_backend_is_refused`; 包含 测试覆盖调整、导入关系调整、控制流调整; +92/-0

关键符号: `KVReadTables`, `KVIndexTable`, `sliding_window_read_ids`, `KVIndexTranslator`, `init`, `make_capture_tables`, `_zeros`, `build_index_table`, `_translate_cuda_graph_shared_pool_locs`, `_fill_cuda_graph_write_locs`

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