

PR #28446 完整报告

sgl-project/sglang

[HiCache]Asymmetric pool support direct backend

合并时间: 2026-06-17 04:17

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28446>

执行摘要

- 一句话: Asymmetric 池新增 direct 后端支持, 移除 kernel 强制限制
- 推荐动作: 值得精读, 特别是 Asymmetric 池的布局 and IO 后端设计, 以及 K/V 独立传输的意图。审查者关注如何确保 stride 安全地由单独的 kernel 调用承载。

功能与动机

PR body 引用了一个 CI 失败链接和 #21631。此前 Asymmetric 池仅支持 kernel 后端和 page_first 布局, 使用 direct 后端时 K/V 传输会被拒绝。此 PR 旨在开放 direct 后端支持, 使 HiCache 在 asymmetric MHA 场景下也能利用 O_DIRECT 等高性能 IO 后端的优势。

实现拆解

1. AsymmetricMHATokenToKVPoolHost.init_kv_buffer (memory_pool_host.py) : 新增 page_first_direct 分支, 分配五维张量 (page_num, layer_num, page_size, head_num, head_dim) 作为 K/V 缓冲区, 与现有 page_first 的四维形状区分。 \
2. load_to_device_per_layer / backup_from_device_all_layer (memory_pool_host.py) : 为 io_backend='direct' 增加独立分支, 分别调用 transfer_kv_per_layer_direct_pf_lf 和 transfer_kv_all_layer_direct_lf_pf, 且 K 和 V 作为单独调用以确保 stride 计算正确 (因 head dim 不同)。 \
3. server_args._handle_model_specific_adjustments (server_args.py) : 移除 MiMoV2 启用 HiCache 时强制设置 hicache_io_backend='kernel' 和 hicache_mem_layout='page_first' 的代码, 改为仅注释说明两者均受支持。 \
4. 测试配套: 在 test_kvcacheio_asymmetric.py 中新增 test_asymmetric_mha_direct_page_first_direct_roundtrip; 在 test_asymmetric_mha_pool_host_unit.py 中将原有拒绝 direct 的测试改为验证分派正确性; test_mimo_v2.py 和 test_mimo_v2_flash.py 将启动参数切换为 page_first_direct 和 direct 以实现 e2e 覆盖。

关键文件:

- python/sglang/srt/mem_cache/memory_pool_host.py (模块 内存池; 类别 source; 类型 core-logic; 符号 init_kv_buffer, load_to_device_per_layer, backup_from_device_all_layer) : 核心类 AsymmetricMHATokenToKVPoolHost 新增 page_first_direct 布局 and direct IO 后端分支, 实现 K/V 独立传输

- test/registered/jit/test_kvcacheio_asymmetric.py (模块 KVCacheIO 测试; 类别 test; 类型 test-coverage; 符号 make_host_pool, _host_k_tokens, _host_v_tokens, test_asymmetric_mha_direct_page_first_direct_roundtrip) : 新增测试用例 test_asymmetric_mha_direct_page_first_direct_roundtrip, 覆盖 direct 后端的完整 roundtrip 流程
- test/registered/unit/mem_cache/test_asymmetric_mha_pool_host_unit.py (模块 AsymmetricMHA 单元测试; 类别 test; 类型 test-coverage; 符号 test_direct_load_splits_k_and_v_for_page_first_direct, test_direct_backup_splits_k_and_v_for_page_first_direct, test_direct_requires_page_first_direct_layout) : 新增单元测试验证 direct 后端的分派正确性 (K/V 分别调用)
- python/sglang/srt/server_args.py (模块 配置; 类别 source; 类型 core-logic; 符号 _handle_model_specific_adjustments) : 移除 MiMoV2 场景下强制使用 kernel/page_first 的限制, 允许用户选择 direct 后端
- test/registered/models_e2e/test_mimo_v2.py (模块 MiMoV2 e2e 测试; 类别 test; 类型 test-coverage) : 将 e2e 测试的布局 and IO 后端改为 page_first_direct/direct, 确保新路径被覆盖
- test/registered/models_e2e/test_mimo_v2_flash.py (模块 MiMoV2 Flash e2e 测试; 类别 test; 类型 test-coverage) : 同上, Flash 版本的 e2e 测试也切换到 direct 后端

关键符号: init_kv_buffer, load_to_device_per_layer, backup_from_device_all_layer, _host_k_tokens, _host_v_tokens, test_asymmetric_mha_direct_page_first_direct_roundtrip, test_direct_load_splits_k_and_v_for_page_first_direct, test_direct_backup_splits_k_and_v_for_page_first_direct, test_direct_requires_page_first_direct_layout

评论区精华

此 PR 无实质性 review 讨论, 仅 Gemini 自动检查 (无意见) 和 hzh0425 的快速批准。

- 暂无高价值评论线程

风险与影响

- 风险: 新增的 direct 分支仅在 page_first_direct 布局下生效, 若误用 page_first 布局会明确报错。风险较低, 但若未来新增其他布局 (如 page_head), 需确保该分支被正确处理。另外, 测试覆盖了 H200 单卡场景, 但未覆盖多卡或 AMD 平台, 可能存在兼容性风险。
- 影响: 使用 HiCache 且 K/V head dim 不同的模型 (如 MiMoV2) 用户现在可以选择 direct 后端, 可能获得更高的 IO 性能。server_args 移除强制限制后, 用户可以自由切换布局 and 后端。系统行为变化: 原先被拒绝的 direct 请求现在可正常工作。影响范围限于 HiCache 异步传输场景。
- 风险标记: 新增 direct 后端代码路径, 测试覆盖有限 (仅 H200 单卡), 布局兼容性风险

关联脉络

- PR #28375 [UnifiedTree]: Replace anonymous tuples with NamedTuples in UnifiedRadixCache: 同一模块 (mem_cache/unified_radix_cache) 的重构演进, 为 HiCache 提供更清晰的内部数据表示。
- PR #28389 [UnifiedTree]: Clean up some unused dead code.: 清除 UnifiedRadixCache 中的死代码, 间接影响 HiCache 的稳定性。