PR #41445 完整报告

vllm-project/vllm

[kv_offload+HMA][13/N]: Enable HMA support

合并时间：2026-05-01 19:30

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41445

执行摘要

- 一句话：启用 OffloadingConnector HMA 支持
- 推荐动作：值得精读，特别是测试框架如何通过 GPUBlock 和 kv_cache_groups 参数模拟多组场景，以及 SupportsHMA 接口的设计模式。

功能与动机

原 offloading connector 仅支持单一 KV 缓存组，无法兼容使用滑动窗口或混合注意力状态的 HMA 模型。用户 naveline67 在评论中反馈在 Qwen3.6 上使用 ——kv-offloading-size 16 工作正常，表明实际部署效果良好。

实现拆解

1. 在 OffloadingConnector 上实现 SupportsHMA 接口：继承 SupportsHMA，新增 request_finished_all_groups 方法，委托给 scheduler.request_finished（忽略 block_ids 参数）。
2. 调整 Scheduler 的 request_finished 签名：移除 block_ids 参数，简化接口。
3. 重构测试框架：引入 GPUBlock dataclass 表示带组索引的 GPU 块；重构 RequestRunner 支持可选的 kv_cache_groups 和 block_size_factor，模拟多组不同 block size 场景。
4. 增加多组测试用例：test_two_groups_full_and_sliding_window 和 test_two_groups_different_block_sizes 验证调度逻辑。
5. 增强端到端测试：用 MODEL_PARAMS 参数化替换硬编码列表，加入 google/gemma-3-1b-it 等 HMA 模型，测试 CPU offloading 的延迟与准确性。

关键文件：

- tests/v1/kv_connector/unit/offloading_connector/utils.py（模块 测试工具；类别 test；类型 test-coverage；符号 GPUBlock, _mock_get_layers, _update_gpu_block_idx, _update_gpu_blocks）：测试基础设施重构核心，引入 GPUBlock、多 group 和 block_size_factor 支持，使 RequestRunner 可模拟 HMA 场景。
- tests/v1/kv_connector/unit/offloading_connector/test_scheduler.py（模块 调度器测试；类别 test；类型 test-coverage；符号 test_two_groups_full_and_sliding_window, test_two_groups_different_block_sizes）：新增多缓存组和不同块大小的测试用例，验证 HMA 场景下 offloading 调度正确性。

- tests/v1/kv_connector/unit/test_offloading_connector.py（模块 端到端测试；类别 test；类型 test-coverage；符号 _latency_test, _accuracy_test, test_cpu_offloading）：引入 HMA 模型（Gemma-3, Mamba, Falcon-H1）的端到端 offloading 测试，验证多组场景的正确性。
- vllm/distributed/kv_transfer/kv_connector/v1/offloading_connector.py（模块 连接器；类别 source；类型 core-logic；符号 OffloadingConnector, request_finished_all_groups）：核心源码改动：OffloadingConnector 继承 SupportsHMA 接口，新增 request_finished_all_groups 方法。
- vllm/distributed/kv_transfer/kv_connector/v1/offloading/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 request_finished）：调整 request_finished 方法签名，移除 block_ids 参数，简化接口以适配 HMA。

关键符号：OffloadingConnector.request_finished_all_groups, OffloadingConnectorScheduler.request_finished, GPUBlock, _mock_get_layers, _update_gpu_block_idx, test_two_groups_full_and_sliding_window, test_cpu_offloading

评论区精华

gemini-code-assist[bot] 在 review 中指出测试工具中 offload_addresses_to_skip 的计算可能错误，应使用 logical_offset % block_size_factor 而非 -logical_offset % block_size_factor，但该问题未在代码中看到修复。markmc 审核并批准。

- offload_addresses_to_skip 计算可能存在逻辑错误 (correctness): 尚未在代码中看到对应的修复，但 PR 已合并。该问题可能影响测试准确性，但端到端测试结果正常。

风险与影响

- 风险：测试工具中潜在的 offload_addresses_to_skip 计算错误可能导致测试跳过错误数量的子块，影响某些场景的覆盖率可信度，但端到端测试通过。多 KV 组路径首次启用可能在极端配置下触发未预期行为，但已覆盖滑动窗口和不同块大小组合。
- 影响：对使用 offloading connector 的用户，现在可以支持 HMA 模型（如 Gemma-3、Mamba、Falcon-H1），提升兼容性。非 HMA 模型行为不变。测试框架重构可能影响其他基于此测试工具单元的测试，但概率低。
- 风险标记：测试工具潜在 Bug，多 KV 组路径首次启用

关联脉络

- PR #41228 [kv_offload+HMA][12/N]: Scheduler-side support for sliding window groups: 同一系列的前置 PR，为本 PR 提供调度器侧的滑动窗口组支持。
- PR #41361 [KV Offload] Use Collection instead of Sequence/Iterable for OffloadingManager key parameters: 对 KV offload 接口的类型统一，本 PR 依赖其重构的接口以支持 HMA。