PR #44436 完整报告

vllm-project/vllm

[ROCm][CI] Add test for Aiter unified attn kernel

合并时间：2026-06-05 00:15

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44436

执行摘要

- 一句话：新增 ROCm AITER unified attention 核正确性测试
- 推荐动作：值得阅读，尤其是作为 ROCm 自定义 kernel 正确性测试的模板：展示了如何构造 block-sparse attention 输入、如何利用参考实现进行对比、如何参数化覆盖多种数据形状与数值精度。可借鉴到其他 kernel 测试中。

功能与动机

引入测试来比较 ROCm aiter_unified_attn kernel 输出与参考实现，确保 kernel 正确性，并集成到 AMD CI 中。

实现拆解

1. 跳过条件与模块常量：在文件顶部判断平台是否为 ROCm 且为 MI3xx 系列，否则跳过；定义 head_size、block_size、dtype、seq 长度组合等参数。
2. 辅助数据构造函数 _make_case：根据 seq_lens、head_size、block_size、dtype 等构建随机 query、key_cache、value_cache、block_tables、cu_seqlens 等输入，支持指定 kv_cache_dtype 和 q_dtype 以适应 FP8 场景。
3. FP8 专用构造 _make_fp8_case：调用 _make_case 并传入对应的 FP8 dtype 和 scale 参数，生成 FP8 kv-cache、FP8 query 或两者均为 FP8 的测试数据。
4. 参考输出计算 _ref_output：调用 ref_paged_attn（从 test_triton_unified_attention 导入）作为参考实现，得到标准 attention 输出。
5. AITER kernel 执行 _run_aiter_unified_attention：调用 aiter.ops.triton.unified_attention 计算待测输出。
6. 误差比较 _assert_matches_reference：比较待测输出与参考输出，允许指定 atol/rtol，默认 1.5e-2/1e-2，FP8 时放宽至 0.15/0.15。
7. 参数化测试用例：定义 test_aiter_unified_attn_decode、test_aiter_unified_attn_prefill、test_aiter_unified_attn_mixed_batch 三个测试函数，分别覆盖 decode、prefill、mixed 场景；使用 @pytest.mark.parametrize 组合 head_size、block_size、dtype 及 seq_lens 进行多轮验证。FP8 变体通过 test_aiter_unified_attn_fp8 单独覆盖。

关键文件：

- tests/kernels/attention/test_rocm_aiter_unified_attn.py（模块 ROCm 内核验证；类别 test；类型 test-coverage；符号 _require_aiter, _make_case, _make_fp8_case, _run_aiter_unified_attention）：唯一变更文件，新增 339 行测试代码，覆盖 AITER unified attention kernel 的多场景正确性验证。

关键符号：_require_aiter, _make_case, _make_fp8_case, _run_aiter_unified_attention, _ref_output, _assert_matches_reference, test_aiter_unified_attn_decode, test_aiter_unified_attn_prefill, test_aiter_unified_attn_mixed_batch, test_aiter_unified_attn_fp8

评论区精华

- 平台导入保护：tjtanaa 指出在非 ROCm 平台上直接导入 vllm.platforms.rocm 会失败，建议加 guard。divakar-amd 回复已添加 if current_platform.is_rocm(): 保护，并说明回退方案。
- Block size 支持范围：tjtanaa 询问 kernel 是否支持 block size 128/256。divakar-amd 确认支持，但生产环境仅使用 64（引用源码行），保留了更大 size 的测试以验证兼容性。
- 平台导入保护 (correctness): 作者已在相应位置添加 guard，确保非 ROCm 平台跳过导入，从而通过 pytest skip 安全退出。
- Block size 支持范围 (question): 作者确认 kernel 支持这些 size，但生产环境中 kv-cache block size 固定为 64（引用 v1 后端源码），测试保留更大 size 以验证兼容性。

风险与影响

- 风险：纯测试新增，无源码修改，不会对生产逻辑产生回归风险。测试仅在 AMD MI300/MI350 硬件上运行，在其他平台（包括 NVIDIA、CPU）自动跳过，不会影响 CI 流程。唯一的潜在风险是如果 Aiter 库的 API 在未来发生变化，该测试可能因未及时更新而失效，但这是测试维护的正常范围。
- 影响：对用户无直接影响；对 ROCm 平台开发和 CI 有正向作用，显著提高了 AITER unified attention kernel 的正确性覆盖度，便于在合入 kernel 改动前快速发现问题。对其他硬件平台无影响。
- 风险标记：仅在 MI3xx 上运行

关联脉络

- 暂无明显关联 PR