PR #46997 完整报告

vllm-project/vllm

[Bugfix][ROCm][MLA] Pass q/kv dtypes to get_mla_metadata_v1 in FP8 decode

合并时间：2026-06-30 20:31

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46997

执行摘要

- 一句话：修复 AITER MLA FP8 decode 元数据类型缺失导致精度退化
- 推荐动作：值得所有关注 ROCm 和 MLA 后端的工程师精读。本 PR 展示了如何通过参数传递的细微差异导致静默错误，并通过精巧的回归测试（spy + golden 对比）来防护。设计决策明确：优先根因修复而非 workaround。

功能与动机

PR body 指出：使用 AITER MLA 后端在 gfx950（MI355X）上以 FP8 KV cache 部署 MLA 模型时，gsm8k 准确率从 ~0.93 骤降至 ~0。根因是 AiterMLAMetadataBuilder 在调用 get_mla_metadata_v1 时未传递 dtype_q / dtype_kv（默认 None），导致 fold 路径产生错误的 split/reduce 元数据。

实现拆解

1. 保存 dtype 实例变量（rocm_aiter_mla.py__init__）：在计算出 q_dtype 和 kv_dtype 后，添加两行代码持久化到 self._mla_q_dtype 和 self._mla_kv_dtype。
2. 传递 dtype 给元数据函数（_build_decode）：在调用 get_mla_metadata_v1 时添加 dtype_q 和 dtype_kv 参数，与上方的 get_mla_metadata_info_v1 保持一致。
3. 新增回归测试（test_rocm_aiter_mla_decode_metadata.py）：构造 fold 路径配置，使用 mock.patch 拦截 get_mla_metadata_v1 验证 dtype 参数，并通过 golden 元数据逐字段比对确认内容正确。
4. 更新 CI 配置（test-amd.yaml）：在 Kernels (B200-MI355) 步骤中添加新测试命令，确保在 gfx950 硬件上自动运行。

关键文件：

- vllm/v1/attention/backends/mla/rocm_aiter_mla.py（模块 MLA 解码；类别 source；类型 core-logic；符号 init, _build_decode）：核心 bugfix 所在：在 __init__中保存并传递 dtype_q/dtype_kv 给 get_mla_metadata_v1。
- tests/kernels/attention/test_rocm_aiter_mla_decode_metadata.py（模块 测试覆盖；类别 test；类型 test-coverage；符号 _on_gfx950, _build_decode_metadata, spy, _compute_golden_metadata）：新增回归测试，覆盖 FP8 fold 路径，验证 dtype 正确传递并与 golden 元数据一致。

- .buildkite/test-amd.yaml（模块 CI 配置；类别 config；类型 configuration）：CI 配置，在 Kernels (B200-MI355) 步骤中集成新测试。

关键符号：AiterMLAMetadataBuilder.init, AiterMLAMetadataBuilder._build_decode, test_persistent_decode_metadata_matches_fp8_golden

关键源码片段

vllm/v1/attention/backends/mla/rocm_aiter_mla.py

核心 bugfix 所在：在 __init__中保存并传递 dtype_q/dtype_kv 给 get_mla_metadata_v1。

```
# 在 __init__ 中，计算 dtype 后持久化保存
kv_dtype = dtypes.d_dtypes.get(kv_cache_dtype_str, dtypes.bf16)
# Persist for get_mla_metadata_v1 (decode build): omitting these causes
# wrong split/reduce metadata for the gfx950 fp8 nhead=32 fold path.
self._mla_q_dtype = q_dtype
self._mla_kv_dtype = kv_dtype

# 在 _build_decode 中，向 get_mla_metadata_v1 传入 dtype
get_mla_metadata_v1(
    qo_indptr,
    paged_kv_indptr,
    paged_kv_last_page_len,
    self._num_attention_heads,
    1,
    True,
    self._mla_work_meta_data,
    self._mla_work_info_set,
    self._mla_work_indptr,
    self._mla_reduce_indptr,
    self._mla_reduce_final_map,
    self._mla_reduce_partial_map,
    page_size=1,
    kv_granularity=16,
    max_seqlen_qo=max_qo_len,
    uni_seqlen_qo=max_qo_len,
    fast_mode=True,
    dtype_q=self._mla_q_dtype, # 新增
    dtype_kv=self._mla_kv_dtype, # 新增
)
```

tests/kernels/attention/test_rocm_aiter_mla_decode_metadata.py

新增回归测试，覆盖 FP8 fold 路径，验证 dtype 正确传递并与 golden 元数据一致。

```
# 使用 spy 拦截 get_mla_metadata_v1 并记录参数
real_get_mla_metadata_v1 = aiter.get_mla_metadata_v1

def spy(*args, **kwargs):
    captured["args"] = args
```

```
    captured["kwargs"] = dict(kwargs)
    return real_get_mla_metadata_v1(*args, **kwargs)

with patch("aiter.get_mla_metadata_v1", spy):
    metadata = builder.build(...)

# 验证 dtype 参数正确
assert captured["kwargs"]["dtype_q"] == EXPECTED_Q_DTYPE # bfloat16
assert captured["kwargs"]["dtype_kv"] == EXPECTED_KV_DTYPE # float8_e4m3fn

# 与 golden 元数据逐字段对比（跳过 work_meta_data）
for field in _CONTENT_METADATA_FIELDS:
    golden = _compute_golden_metadata(
        captured["args"][:_NUM_INPUT_ARGS],
        captured["kwargs"],
        getattr(metadata, field),
    )
    assert torch.equal(getattr(metadata, field), golden), f"{field} mismatch"
```

评论区精华

AndreasKaratzas 在 review 中对测试文件提出三点建议：使用内置 is_aiter_found 而非 importlib.find_spec；移除不必要的 try/except（因已在非 ROCm 平台跳过测试）；简化文档字符串。作者在后续 commit 中全部采纳，最终 Andreas 给出 APPROVED。

- 使用内置函数替代 importlib.find_spec (style): 作者在后续 commit 中改用 is_aiter_found()。
- 移除不必要的 try/except (correctness): 作者在后续 commit 中移除了 try/except。
- 简化文档字符串 (style): 作者在后续 commit 中缩短了模块文档字符串。

风险与影响

- 风险：核心变更仅 6 行，逻辑简单，测试覆盖充分，风险极低。但需注意对 AITER 库接口的依赖；若未来 AITER 版本变更参数签名或行为，需同步更新。影响面仅限于 gfx950 + FP8 KV cache + fold 头数路径，不涉及其它后端或配置。
- 影响：仅影响使用 ROCM_AITER_MLA 后端、在 gfx950（MI355X）上使用 FP8 KV cache 且每 rank 头数触发 fold 路径的用户（如 TP2 下 32 头）。对该类用户是严重精度修复：gsm8k 从 0.01 恢复至 0.92。bf16 KV cache 和原生头数（16/64/128）用户无影响。
- 风险标记：核心元数据变更，AITER 依赖，gfx950 专用路径

关联脉络

- PR #46952 [Bugfix][ROCm][MLA] Pad FP8 MLA decode head count to a native value on gfx950 (fix AITER 0.1.16-post2): 同一个问题的 workaround（padding 头数从 32 到 64），本 PR 是根因修复。