PR #51011 完整报告

vllm-project/vllm

[ROCm][MLA] [K3] Fix fp8 KV cache decode on the AITER MLA backend

合并时间：2026-08-10 23:42

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/51011

执行摘要

- 一句话：修复 ROCm AITER MLA 后端 fp8 KV cache 的静默错误解码
- 推荐动作：值得精读。该 PR 是典型的"路由正确性"修复，展示了如何用少量谓词把多层次架构约束（dtype、head 数、qlen、架构）收敛到一个决策点。三个可借鉴的设计：
  1. 用 reorder_batch_threshold 统一预测解码的元数据尺寸，避免方法名白名单漂移；
  2. 让 builder 与 impl 共用同一路由谓词，防止 gate 失效；
  3. 在无实测数据的架构上采用保守守卫并标注推断边界。建议读者关注后续 #50619 对 Gluon 能力门控的扩展。

功能与动机

PR body 和评论明确说明：Kimi-K3 在 TP8 下每 rank 有 12 个 MLA head，--kv-cache-dtype fp8 时无法正确服务。未修补的 main 上完整 GSM8K 得分为 74.00%，其中 285/1319 条答案退化；修复后 97.19%，与 bf16 完全一致。作者还区分了三种失败模式，特别强调第三种"静默错误输出"最危险——21.6% 的请求输出一两个合理 token 后重复单个 token 直到长度上限，这是 NaN 到达采样器的特征，短 smoke test 无法暴露，容易被误读为量化损失。

实现拆解

1. 修复推测解码查询长度推导：在 vllm/v1/attention/backends/mla/rocm_aiter_mla.py 的 AiterMLAMetadataBuilder.__init__ 中，将 self._mtp_decode_qlen 的来源从 ("mtp", "deepseek_mtp") 方法名白名单改为 self.reorder_batch_threshold or 1。reorder_batch_threshold 是 decode 能收到的最大查询长度，并且已经按 drafter 方案（含 parallel drafting 的 1 + 2 * num_spec）计算，因此 DSpark、Eagle 等未列入白名单的 drafter 也能获得正确元数据尺寸，避免 get_block_n_fp8[num_heads * qlen] 的 KeyError。
2. 新增 dtype 感知的路由谓词：use_gluon_decode 增加 kv_cache_dtype 参数，在 is_quantized_kv_cache(kv_cache_dtype) 为真时直接返回 False，且该检查先于 VLLM_ROCM_AITER_MLA_ASM_PADDING 模式开关；forward_mqa 中内联的 verify 分支被抽出为 use_gluon_verify 谓词，同样拒绝 fp8，使 builder 与 impl 对路由决策看到同一结果。

3. persistent metadata 门控跟随路由：_build_decode 中 use_persistent_metadata 从 num_heads >= 16 改为排除两个 Gluon 入口，并追加架构能力守卫 num_heads >= 16 or max_qo_len <= 4 or is_quantized_kv_cache(...)，防止 gfx942 上 bf16 小 head 且 spec step 大于 4 的形状请求不存在的 persistent kernel；同时 _kv_cache_dtype_str 被保留在 builder 实例上，供 _build_decode 决策使用。
4. 测试配套：新增 tests/v1/attention/test_rocm_aiter_mla_fp8_decode_routing.py（154 行），用 monkeypatch 固定架构门控与模式开关，参数化扫描 dtype、head 数、qlen，并钉住 use_gluon_decode / use_gluon_verify 的全部组合；补充 test_padded_query_is_contiguous 回归与 pad/unpad round-trip。tests/v1/attention/test_rocm_aiter_mla_mtp_split.py 扩展 drafter 方法（dspark、eagle）与 fp8、12-head 场景，新增 test_persistent_metadata_gate_without_gluon_build；tests/kernels/attention/test_rocm_aiter_mla_head_padding.py 适配新谓词签名（传入 "auto"）。

关键文件：

- vllm/v1/attention/backends/mla/rocm_aiter_mla.py（模块 MLA 后端；类别 source；类型 core-logic；符号 use_gluon_decode, use_gluon_verify, _build_decode, forward_mqa）：核心源码文件，所有路由与门控变更集中在此，包含 use_gluon_decode/use_gluon_verify 谓词、_mtp_decode_qlen 推导和 persistent metadata 门控的修改。
- tests/v1/attention/test_rocm_aiter_mla_fp8_decode_routing.py（模块 解码路由；类别 test；类型 test-coverage；符号 test_fp8_never_routes_to_gluon, test_fp8_never_routes_to_gluon_under_any_mode, test_large_head_counts_never_use_gluon, test_unquantized_divisor_heads_keep_gluon_decode）：新增回归测试，系统性地钉住所有路由谓词在 fp8/bf16 下的行为，是验证本 PR 正确性的核心测试文件。
- tests/v1/attention/test_rocm_aiter_mla_mtp_split.py（模块 MTP 拆分；类别 test；类型 test-coverage；符号 test_mtp_builder_init_sizes_native_fp8_metadata, test_persistent_metadata_gate, test_persistent_metadata_gate_without_gluon_build）：更新 persistent metadata 门控测试，覆盖 DSpark/Eagle drafter、fp8 缓存与 12-head 非约数形状，并新增无 Gluon 构建时的门控测试。
- tests/kernels/attention/test_rocm_aiter_mla_head_padding.py（模块 头部填充；类别 test；类型 test-coverage）：适配 use_gluon_decode 新增 kv_cache_dtype 参数，并修复 test_h12_aiter_mla_decode_matches_reference 的 impl 构造。

关键符号：use_gluon_decode, use_gluon_verify, _build_decode, forward_mqa, AiterMLAMetadataBuilder.init

关键源码片段

tests/v1/attention/test_rocm_aiter_mla_fp8_decode_routing.py

新增回归测试，系统性地钉住所有路由谓词在 fp8/bf16 下的行为，是验证本 PR 正确性的核心测试文件。

```
# tests/v1/attention/test_rocm_aiter_mla_fp8_decode_routing.py（新增）
# 固定架构门控与模式开关，让测试只专注 dtype 规则
@pytest.fixture
def gluon_available(monkeypatch):
    monkeypatch.setattr(rocm_aiter_mla, "_gluon_mla_decode_supported", lambda: True)
    monkeypatch.setattr(rocm_aiter_mla, "_aiter_mla_small_head_mode", lambda: "auto")

# fp8 下任意 head 数、任意 qlen 都不能进 Gluon；
# head 数特意扫过 16 的约数（bf16 留在 Gluon 的形状）验证 fp8 守卫覆盖它
@pytest.mark.parametrize("kv_cache_dtype", ["fp8", "fp8_e4m3", "fp8_e5m2"])
@pytest.mark.parametrize("num_heads", [1, 2, 4, 5, 6, 8, 12, 16, 32, 128])
@pytest.mark.parametrize("max_qo_len", [1, 2, 4, 5, 8, 15])
def test_fp8_never_routes_to_gluon(kv_cache_dtype, num_heads, max_qo_len):
    assert not AiterMLAHelper.use_gluon_decode(num_heads, max_qo_len, kv_cache_dtype)
    assert not AiterMLAHelper.use_gluon_verify(num_heads, max_qo_len, kv_cache_dtype)
```

评论区精华

Review 主要围绕三件事：

1. 门控中 use_gluon_decode 与 use_gluon_verify 是否重复（作者说明二者在 qlen 上互斥，但都必须排除）；
2. gfx942 影响评估，作者增加了持久化门控守卫并明确标注推断依据；
3. 反复要求精简注释。另外，作者与维护者讨论了与 #50578、#50619 的关系，确认本 PR 是 correctness-first 的 ASM 基线。

- 门控中两个 Gluon 谓词是否重复 (design): 作者保留两个检查，并在代码中加入一行注释说明互斥性。
- gfx942 影响评估与持久化门控守卫 (correctness): 守卫已合入，新增 test_persistent_metadata_gate_without_gluon_build 覆盖；作者标注该结论基于 kernel 表与 dispatch 推断，未在 gfx942 实测。
- 注释精简 (style): 作者将注释行从 37/72 缩减到 24/53，只保留代码无法直接体现的边界推导（reorder_batch_threshold 的由来、门控跟随路由的原因）。
- CI 七项失败与测试签名适配 (testing): 作者修复测试后，CI #83093 96/96 全绿。
- 与 #50578、#50619 的协作关系 (design): 本 PR 先合入，#50619 后续基于其上再接 Gluon 能力门控。

风险与影响

- 风险：变更集中于 vllm/v1/attention/backends/mla/rocm_aiter_mla.py 的 decode 路由，风险点包括：
  1. _mtp_decode_qlen 改为 reorder_batch_threshold 后，如果调度器阈值语义变化，元数据缓冲可能过大或过小（过小时 aiter 查表 KeyError）；
  2. persistent 门控新增的架构守卫在 gfx942 上基于 kernel 表与 dispatch 推断，未实测，存在错误的可能；

3. use_gluon_verify 让 fp8 走 asm fold 路径，若 AITER 包缺少对应 .co 内核会在运行时失败；
4. test 大量 monkeypatch，真实 kernel 组合（gfx942、不同 AITER 版本）覆盖不足，存在回归风险。 - 影响：从用户和系统角度看：影响范围是 ROCm 平台 + AITER MLA 后端 + fp8 KV cache 的用户，重点是 Kimi-K3（K3 标签）的 TP 部署。修复避免了静默错误输出（NaN 到达采样器导致 We!!!!!!... 式退化）和随机 KeyError 崩溃，GSM8K 从 74.00% 提升到 97.19%，并因不再跑满退化请求的 2048 token 上限而获得约 1.9 倍吞吐提升。对团队而言，该 PR 确立了 fp8 解码的 ASM 基线，为 #50619 的 DSpark Gluon 后端铺路，也补上了 #50578 未覆盖的 persistent 门控漏洞。 - 风险标记：核心路径变更，静默错误修复，跨架构验证不足，依赖较新 AITER 版本，测试依赖 monkeypatch

关联脉络

- PR #50578 [ROCm][MLA] Add VLLM_ROCM_AITER_MLA_ASM_PADDING knob: 作者在 PR body 与评论中多次对比两个 PR 的适用范围；#50578 的 knob 覆盖了本 PR 的 change 2 和 3 的一部分，但无法修复 change 4 的 persistent 门控问题。
- PR #50619 DSpark fp8 kv-cache-dtype on gluon-mla path: hongxiayang 在评论中建议查看该 PR；JohnQinAMD 说明本 PR 为 ASM 基线，后续 #50619 将在这上面叠加 capability-gated Gluon 后端。