PR #50578 完整报告

vllm-project/vllm

[ROCm][MLA] Use asm decode for non-divisor small head counts

合并时间：2026-08-07 02:56

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50578

执行摘要

- 一句话：ROCM MLA 小头数 decode 走 asm，长上下文提速 2.3x
- 推荐动作：值得精读。核心看点： 1) tile-and-slice 精确 padding 的设计——依赖 MLA per-head 独立性，简单且可证明正确，对比 append-only 方案的缺陷很有教学价值； 2) 架构门控 + 三态环境变量的双层内核路由设计，兼顾自动选择与用户可控； 3) review 中 maeehart 连续发现两个被遗漏的 Gluon 调用点，展示了 " 门控函数是否覆盖所有调用点 " 的审查思路； 4) 测试契约覆盖 1..15 全量头数与三种 env 模式，硬件端到端测试与 SDPA 参考对齐。

功能与动机

PR body 明确指出问题根因：ROCM_AITER_MLA 下少于 16 个 query head 的 decode 走 Gluon 内核，而 "Gluon parallelizes only over heads, so a single workgroup marches through the whole KV cache and per-token latency scales linearly with context length"。Kimi-K3 共 96 个 attention head（kv_lora_rank=512），TP8 时每 rank 12 个 head，是 16 的非约数：旧的 get_mla_padded_q 用 repeat_interleave(16 // num_heads) 补齐，对 12 头是 no-op，因此长期卡在慢速 Gluon 路径上。验证数据（MI355X/gfx950、TP8、mxfp4 单流 greedy）显示 32K prefill / 32K output 下 TPOT 从 54.1 ms 降到 23.6 ms（2.29x），且无短上下文回归。

实现拆解

1. padding 泛化（vllm/v1/attention/backends/mla/rocm_aiter_mla.py::AiterMLAHelper）：get_mla_padded_q 保留 16 的约数（1/2/4/8）走 repeat_interleave 的旧路径，新增非约数分支 q.repeat(1, reps, 1)[:, :16, :].contiguous()（reps = ceil(16 / num_heads)）；get_mla_unpadded_o 对应以 o[:, :num_heads, :] 切回真实 head。关键细节是平铺后切片会产生非连续 view，而 asm persistent decode 把 q 当作 packed [tokens, 16, head_dim] 缓冲读取，必须 .contiguous() 物化。该写法对任意 1..15 头数都精确补到 16，修复了早期 append-only 版本在 num_heads < 8 时补不足的缺陷（TP16=6 时旧写法只补到 12）。
2. 架构门控 _gluon_mla_decode_supported()：Gluon 小头内核的 tiling 需要约 160 KiB LDS，超过 CDNA3 的 64 KiB，因此 gfx942 上没有可回退的 build，选中即断言 mla_gluon requires gfx950。新增 @functools.lru_cache 的探测函数（仿照既有 _fp8_mla_prefill_supported()），use_gluon_decode 现在综合头数、max_qo_len、env

模式与架构四重条件决策。

3. 第二个 Gluon 调用点（DSpark 多 token verify）：forward_mqa 中 num_heads < 16 且 max_qo_len > 1 的 verify 展平分支直接调用 mla_gluon，不经过 use_gluon_decode，架构门控覆盖不到，gfx942 上一跑 speculative decoding 就断言。该分支现在同步受 _aiter_mla_small_head_mode() != "asm" 与 _gluon_mla_decode_supported() 门控；门控不过时落入 asm persistent decode，其本身支持 qlen > 1 verify（元数据非持久时对 MTP 未写入 lane 清零，且 qh16 ... mask1 内核自带因果掩码）。
4. 环境变量开关（vllm/envs.py）：按维护者 hongxiayang 要求新增 VLLM_ROCM_AITER_MLA_ASM_PADDING，三态 auto（默认，架构感知）/ gluon / asm，用 env_with_choices 注册限定取值，保证进入校验后的 env schema；gfx942 上 gluon 会回退 ASM 并打印一次性警告。_aiter_mla_small_head_mode() 读取该变量。
5. 测试配套：新增 tests/kernels/attention/test_rocm_aiter_mla_head_padding.py，覆盖 1..15 全量头数的 pad/unpad 往返、TP8=12 与 TP16=6 具体案例、约数路径行为不变、三种 env 模式下 use_gluon_decode 的选择，以及硬件门控的 12-head asm decode 与 SDPA 参考对比；修改 tests/kernels/attention/test_rocm_aiter_mla_causal_verify_mask.py，在测试内强制 _gluon_mla_decode_supported() 为 True，使 gfx942 也能跑到被 spy 替换的 Gluon verify 路径，保住因果窗口断言。

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 _gluon_mla_decode_supported, _aiter_mla_small_head_mode, get_mla_padded_q, get_mla_unpadded_o）：核心变更文件：get_mla_padded_q / get_mla_unpadded_o 支持非约数头数精确补齐，use_gluon_decode 加入架构门控与三态 env 模式，forward_mqa 的 DSpark verify 分支同步门控，是全部路由逻辑的所在。
- tests/kernels/attention/test_rocm_aiter_mla_head_padding.py（模块 MLA 测试；类别 test；类型 test-coverage；符号 _rocm_aiter_available, _on_gfx950, _expected_tile_pad, _make_h12_decode_metadata）：新增测试文件，覆盖 1..15 全量头数的 pad/unpad 往返、TP8=12 与 TP16=6 具体案例、约数路径行为不变、三种 env 模式的 Gluon 选择，以及硬件端 12-head asm decode 与 SDPA 参考对比，是本 PR 契约的完整固化。
- vllm/envs.py（模块 环境变量；类别 source；类型 configuration；符号 VLLM_ROCM_AITER_MLA_ASM_PADDING）：注册 VLLM_ROCM_AITER_MLA_ASM_PADDING 三态环境变量并用 env_with_choices 限定取值，是维护者要求的 Gluon 路径保留开关的统一配置入口。
- tests/kernels/attention/test_rocm_aiter_mla_causal_verify_mask.py（模块 MLA 测试；类别 test；类型 test-coverage；符号 _run_verify_block, test_verify_flatten_rows_are_causal）：修复 gfx942 上因架构门控短路导致 Gluon spy 不触发的回归：测试内强制 patch _gluon_mla_decode_supported() 为 True，保住 verify 因果窗口断言在所有 ROCm AITER runner 上可运行。

关键符号：get_mla_padded_q, get_mla_unpadded_o, use_gluon_decode, _gluon_mla_decode_supported, _aiter_mla_small_head_mode, forward_mqa, check_num_heads_validity

关键源码片段

tests/kernels/attention/test_rocm_aiter_mla_head_padding.py

新增测试文件，覆盖 1..15 全量头数的 pad/unpad 往返、TP8=12 与 TP16=6 具体案例、约数路径行为不变、三种 env 模式的 Gluon 选择，以及硬件端 12-head asm decode 与 SDPA 参考对比，是本 PR 契约的完整固化。

```
def _expected_tile_pad(q: torch.Tensor, num_heads: int, m: int = 16) -> torch.Tensor:
    # get_mla_padded_q 平铺头部后切片到 m，即 padded 张量的第 i 个 head
    # 是输入的 head (i % num_heads)。
    idx = [i % num_heads for i in range(m)]
    return q[:, idx, :]


def test_h12_query_is_tile_padded_to_h16():
    # Kimi-K3 在 TP8 下每 rank 12 个 head，是 16 的非约数，
    # 必须走 tile-and-slice 补齐路径。
    q = torch.arange(2 * 12 * 4, dtype=torch.bfloat16).view(2, 12, 4)
    padded_q = AiterMLAHelper.get_mla_padded_q(12, q)

    assert padded_q.shape == (2, 16, 4)
    assert padded_q.is_contiguous() # asm kernel 需要 packed 连续缓冲
    # 真实 head 原样保留 ...
    torch.testing.assert_close(padded_q[:, :12], q)
    # ...4 个 padding head 是平铺回绕（head 0..3），不是零填充：
    # MLA 注意力按 head 独立，重复 query head 无害，输出端会被切掉。
    torch.testing.assert_close(padded_q[:, 12:], q[:, :4])


@pytest.mark.parametrize("num_heads", NON_DIVISOR_HEADS + DIVISOR_HEADS)
def test_all_small_head_counts_pad_to_16_and_round_trip(num_heads: int):
    # 全量 1..15 头数的 pad -> unpad 往返必须精确还原。
    q = torch.arange(2 * num_heads * 4, dtype=torch.float32).view(2, num_heads, 4)
    padded_q = AiterMLAHelper.get_mla_padded_q(num_heads, q)
    unpadded_o = AiterMLAHelper.get_mla_unpadded_o(num_heads, padded_q)

    assert padded_q.shape == (2, 16, 4)
    assert padded_q.is_contiguous()
    torch.testing.assert_close(unpadded_o, q)
```

vllm/envs.py

注册 VLLM_ROCM_AITER_MLA_ASM_PADDING 三态环境变量并用 env_with_choices 限定取值，是维护者要求的 Gluon 路径保留开关的统一配置入口。

```
    # 新增小头 MLA decode 路径选择开关，注册到 vllm/envs.py：
    # 在类型标注区声明（默认 "auto"），并在 envs 解析区用 env_with_choices
    # 限定合法取值，保证能出现在校验后的 env schema 中。
    VLLM_ROCM_AITER_MLA_ASM_PADDING: Literal["auto", "gluon", "asm"] = "auto"
    ...
```

```
    # Small-head (<16) AITER MLA decode kernel selection. Small head counts
    # (e.g. Kimi-K3: 12 heads/rank at TP8, 6 at TP16) can decode either through
    # the Gluon small-head kernel or through the padded persistent-scheduling
    # (PS) ASM kernel. "auto" (default) keeps Gluon for head counts that divide
    # 16 where a Gluon build exists (gfx950/CDNA4) and otherwise uses the
    # padded PS ASM decode; "gluon" forces the Gluon path wherever a build
    # exists; "asm" forces the padded PS ASM decode. On gfx942/CDNA3 there is
    # no Gluon build, so the ASM path is always used regardless of this
    # setting.
    "VLLM_ROCM_AITER_MLA_ASM_PADDING": env_with_choices(
        "VLLM_ROCM_AITER_MLA_ASM_PADDING",
        "auto",
        ["auto", "gluon", "asm"],
        case_sensitive=False,
    ),
```

评论区精华

核心交锋集中在四条线：① dllehr-amd 对 TP>8 时 padding 正确性的担忧直接引出 append-only 到 tile-and-slice 的重写；② maeehart 连续发现两个被遗漏的 Gluon 调用点——gfx942 约数头数仍选 Gluon 导致服务无法启动，以及 DSpark verify 分支直接调 mla_gluon 不受门控，并带来 CDNA3 LDS 容量限制（64 KiB vs 约 160 KiB）的架构洞察，两次修复均获 co-author credit；③ hongxiayang 与 tjtanaa 对环境变量策略的分歧——前者的用户可控性诉求落地为三态 env，后者的 flag 膨胀顾虑未被完全采纳但 auto 默认值缓解；④ maeehart 主动提交 gfx942 上因果 verify 测试的回归修复（架构门控使 spy 不再触发），被 cherry-pick 入分支。

- TP16（6 heads/rank）等小头数是否会被错误 padding (correctness): tile-and-slice 精确补齐并验证 1..15 头数均可还原真实 head；asm 内核只要求 padded count == 16。
- gfx942 上约数头数仍选 Gluon 导致 mla_gluon requires gfx950 断言 (correctness): 新增 _gluon_mla_decode_supported()（lru_cache 的 on_gfx950()），use_gluon_decode 同时按架构门控；gfx950 行为不变，gfx942 全部小头数走 asm。
- DSpark 多 token verify 分支是第二个 Gluon 调用点，未被架构门控覆盖 (correctness): 该分支同样加 _gluon_mla_decode_supported() 与 env 模式门控；gfx942 落入 asm decode（支持 qlen>1 verify，因果掩码），gfx950 行为不变。
- 需要环境变量允许用户 opt in/out asm padding 路径 (design): 新增 VLLM_ROCM_AITER_MLA_ASM_PADDING（auto/gluon/asm），auto 为默认，gluon 路径在 gfx950 上完整保留。
- 环境变量数量膨胀，倾向按架构自动路由 (design): PR 保留三态变量但 auto 默认即架构感知路由；该设计顾虑未完全采纳，属于遗留设计注记。
- 与 #50371 重叠，要求补 num_head=12 的单元与注意力类测试 (testing): 新增 tests/kernels/attention/test_rocm_aiter_mla_head_padding.py，覆盖 1..15 头数与三种 env 模式；测试结构借鉴 #50371。
- gfx942 上 causal verify 测试不再触发 Gluon spy (testing): 测试内强制 patch 该探测为 True，因果窗口断言与架构无关，可在任意 ROCm AITER runner 上运行；修复被

cherry-pick 入分支。

- spec-decoding 与 asm MLA 组合的支持范围 (question): 作为已知限制记录在案；gfx942 的 spec decode 验证由 maeehart 完成（剩余失败定位为 AITER fused-MoE 问题 ROCm/aiter#4487，与 MLA 无关）。

风险与影响

- 风险：
  1. padding 正确性前提：非约数补齐依赖 MLA 每 query head 独立作用于共享 KV 的假设，padding head 的注意力结果会被直接丢弃。该假设对当前 MLA 结构成立，但若未来引入跨 head 交互（如 head 级 mask），需要重新审视。
  2. gfx950 与 AITER nightly 的耦合：auto 模式下 gfx950 的约数头数仍走 Gluon，而作者实测新 nightly 的 _mla_gluon 在 gfx950 上会崩溃（Triton layout 错误，mla_gluon.py:1029），这是 AITER 侧问题，vLLM 侧只能通过 asm 模式规避。
  3. spec-decoding 覆盖有限：gfx942 上 spec decode 走 asm 已验证（DSpark 草稿），但 gfx950 上 asm 强制模式与 spec decoding 的组合未验证，维护者明确标注为不支持范围。
  4. 性能风险低：非约数路径每次多一次 .contiguous() 拷贝，相对 kernel 时间可忽略；cudagraph 捕获后为固定路径，无运行时抖动。
  5. 配置面膨胀：新增 env 变量与既有大量 VLLM_ROCM_* 叠加，tjtanaa 明确反对，长期是维护成本。
  6. 影响范围：仅 ROCM_AITER_MLA（rocm_aiter_mla.py + envs.py），CUDA/FlashInfer/Triton 等其他 attention 后端不受影响；测试在非 ROCm + AITER 环境通过 skipif 跳过。 - 影响：用户侧：Kimi-K3（mxfp4）在 MI355X/MI325X 上的长上下文 decode 延迟大幅下降（最高 2.29x），gfx942 用户此前小头数 MLA 服务根本无法启动，现在可以正常 serve 并支持 250K 上下文。系统侧：变更局限于 ROCM_AITER_MLA 后端与小头数（<16）decode 路由，不影响 CUDA 等其他后端；VLLM_ROCM_AITER_MLA_ASM_PADDING 默认 auto 保持既有 gfx950 行为，无默认路径回归。团队侧：AMD ROCm 团队主导，多位内外部 reviewer（maeehart、dllehr-amd、tjtanaa、seungrokj）深度参与，最终由 hongxiayang 批准合并。 - 风险标记：ROCm/AITER 专属路径变更，padding 依赖 MLA 逐 head 独立假设，gfx950 Gluon 依赖 AITER nightly（外部崩溃风险），spec-decoding 与 asm MLA 组合未支持，新增环境变量（flag 膨胀）

关联脉络

- PR #50371 12-head MLA persistent decode（标题未在上下文中提供）：tjtanaa 指出与本 PR 功能重叠（已批准）；本 PR 的 head-padding 测试结构借鉴自它，作者以 co-author 形式给予 credit，两者在合并前协调了取舍。
- PR #51088 [ROCm][MLA] Add small-head PS ASM decode route (Kimi-K3 TP8): 本 PR 将其中提议的 env guard 整合并注册到 vllm/envs.py，形成统一配置入口，避免重复 flag。

- PR #51065 gfx942 相关集成栈（标题未在上下文中提供）：maeehart 用本 PR + #51065 + AITER 修复在 8×MI325X 上完成 250K 上下文集成验证，确认 gfx942 分支可启动、可捕获图、长上下文正确。
- PR #50613 [Attention][MLA] Per-request scheduling for MLA chunked context: 同属 MLA 后端长上下文性能改造方向：本 PR 解决 ROCm 侧小头数 decode 的算力利用问题，#50613 解决 MLA chunked context 调度，共同指向 MLA 长上下文端到端优化。