

PR #52139 完整报告

vllm-project/vllm

[Bugfix][ROCm][CI] Give the AITER MLA decode metadata stub its MLA dims

合并时间: 2026-08-14 02:35

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52139>

执行摘要

- 一句话: 修复 AITER MLA 测试 stub 维度缺失
- 推荐动作: 值得快速浏览而非精读: 它演示了『测试 stub 必须跟随生产代码读取属性集合』的维护模式, 以及 #51255 类大重构后容易漏改同类测试的教训。后续若再调整 `MLACommonMetadataBuilder` 的读取属性, 应全仓搜索 `SimpleNamespace` 形态的层 stub 一并更新。

功能与动机

PR body 指出 `tests/kernels/attention/test_rocm_aiter_mla_decode_metadata.py::test_persistent_decode_metadata_matches_fp8_golden` 在 main 上失败: `AttributeError: 'types.SimpleNamespace' object has no attribute 'q_lora_rank'`。原因是 #51255 将 `MLACommonMetadataBuilder.__init__` 改为从 `static_forward_context` 的层对象读取 MLA 潜在维度, 并基于 `kv_lora_rank + qk_rope_head_dim` 而非 `model_config.get_head_size()` 分配 chunked-prefill workspace; 该测试的层 stub 只包含 `prefill_backend`, builder 在读取第一个维度时就抛错。#51255 更新了兄弟测试与 MLA backend 测试, 唯独漏掉了本文件。

实现拆解

1. 定位根因: `MLACommonMetadataBuilder.__init__` 在 #51255 变更后从 `static_forward_context[layer_name]` 的层对象上读取 `q_lora_rank`、`kv_lora_rank`、`qk_rope_head_dim`、`qk_rope_head_dim`、`v_head_dim` 五个维度, 并据此分配 chunked-prefill workspace; 测试中的 stub 只有 `prefill_backend`, 属性访问直接抛 `AttributeError`。
2. 修改 stub: 在 `tests/kernels/attention/test_rocm_aiter_mla_decode_metadata.py` 的 `_build_decode_metadata` 中, 使用 `get_mla_dims(vllm_config.model_config)` 获取 DeepSeek-R1 的 MLA 维度, 并把五个维度全部注入 `SimpleNamespace`; 同时新增 `from vllm.model_executor.layers.attention.mla_attention import get_mla_dims` 导入。维度取自查点配置而非硬编码, 避免与测试围绕的 DeepSeek-R1 配置漂移。
3. 验证与排查: 作者在 gfx950 上同时跑 `decode metadata` 与 `causal verify mask` 两个测试, 均通过; 并搜索仓库确认只有这两个测试使用相同的层 stub 模式, 没有其他遗漏点。测试原有断言 (`persistent decode` 元数据与 `fp8 golden` 一致) 保持不变, 覆盖范围没有变化。

关键文件:

- tests/kernels/attention/test_rocm_aiter_mla_decode_metadata.py (模块 注意力测试; 类别 test; 类型 test-coverage; 符号 _build_decode_metadata, get_mla_dims) : 唯一变更文件。为 _build_decode_metadata 中的层 stub 补上 q_lora_rank、kv_lora_rank、qk_nope_head_dim、qk_rope_head_dim、v_head_dim 五个 MLA 维度, 修复 #51255 引入的 AttributeError, 恢复 ROCm CI 上的 AITER MLA decode metadata 测试。

关键符号: _build_decode_metadata

关键源码片段

tests/kernels/attention/test_rocm_aiter_mla_decode_metadata.py

唯一变更文件。为 _build_decode_metadata 中的层 stub 补上 q_lora_rank、kv_lora_rank、qk_nope_head_dim、qk_rope_head_dim、v_head_dim 五个 MLA 维度, 修复 #51255 引入的 `AttributeError`, 恢复 ROCm CI 上的 AITER MLA decode metadata 测试。

```
def _build_decode_metadata():
    """构建 AITER MLA decode 元数据, 返回 (metadata, captured)。

    captured 记录 builder 传给 get_mla_metadata_v1 的位置参数与关键字参数,
    以使用完全相同的输入重算 golden 值。
    """
    from tests.v1.attention.utils import (
        BatchSpec,
        create_common_attn_metadata,
        create_vllm_config,
    )
    from vllm.config.vllm import set_current_vllm_config
    from vllm.model_executor.layers.attention.mla_attention import get_mla_dims
    from vllm.v1.attention.backends.registry import AttentionBackendEnum
    from vllm.v1.kv_cache_interface import MLAAttentionSpec
    from vllm.v1.worker.workspace import init_workspace_manager

    device = torch.device("cuda:0")

    # 测试围绕 DeepSeek-R1 配置构建, cache dtype 固定为 fp8。
    vllm_config = create_vllm_config(
        model_name="deepseek-ai/DeepSeek-R1",
        max_model_len=CONTEXT_LEN,
        # 每个 token 一个平坦页 (page_size=1), 并预留 null block 的缓冲。
        num_gpu_blocks=BATCH_SIZE * CONTEXT_LEN + 200,
        block_size=PAGE_SIZE,
        max_num_seqs=BATCH_SIZE,
        max_num_batched_tokens=8192,
        hf_config_override={"num_attention_heads": NUM_QUERY_HEADS},
    )
    vllm_config.cache_config.cache_dtype = "fp8"

    spec = MLAAttentionSpec(
        block_size=PAGE_SIZE,
```

```

num_kv_heads=1,
head_size=vllm_config.model_config.get_head_size(),
dtype=vllm_config.model_config.dtype,
cache_dtype_str="fp8",
)

builder_cls = AttentionBackendEnum.ROCM_AITER_MLA.get_class().get_builder_cls()

# builder 会从 static_forward_context 的层对象上读取 prefill_backend 和
# MLA 潜在维度，因此 stub 需要同时携带这两类属性。
# 维度值来自 get_mla_dims(model_config)，与 DeepSeek-R1 真实配置一致，
# 避免 stub 与模型配置漂移。
mla_dims = get_mla_dims(vllm_config.model_config)
layer_name = "placeholder"
vllm_config.compilation_config.static_forward_context[layer_name] = (
    types.SimpleNamespace(
        prefill_backend=torch.empty((1,)),
        q_lora_rank=mla_dims.q_lora_rank,
        kv_lora_rank=mla_dims.kv_lora_rank,
        qk_nope_head_dim=mla_dims.qk_nope_head_dim,
        qk_rope_head_dim=mla_dims.qk_rope_head_dim,
        v_head_dim=mla_dims.v_head_dim,
    )
)

init_workspace_manager(device)
# 后续构造 batch_spec，创建 builder，并用 spy 包装
# aiter.get_mla_metadata_v1 来捕获实际调用参数。

```

评论区精华

该 PR 没有产生实质性技术讨论：[claude\[bot\]](#) 自动回复注明来自 fork 的 PR 禁用自动 review；维护者 [AndreasKaratzas](#) 直接批准 (LGTM)，并触发 Buildkite CI (job 83766) 验证。PR body 中作者对根因的说明 (#51255 漏更新同模式测试) 即是主要决策依据。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 回归风险：stub 维度值来源于 `get_mla_dims(vllm_config.model_config)`，若 DeepSeek-R1 相关配置变更而测试未同步，仍可能漂移；但该测试本来就围绕 DeepSeek-R1 构建，风险低。
 - 测试逻辑耦合：`get_mla_dims` 现在是测试 stub 的数据源，而 builder 已不再走该函数；若未来 builder 读取属性集合再次变化，需要同步更新本文件与兄弟测试。
 - 无性能、安全、兼容性影响：纯测试文件变更，+13/-3。

- 影响：对用户与运行时零影响（仅改测试文件）；对 CI 有直接修复作用：恢复 dedicated AITER MLA job 与 sharded kernels/attention job 两个失败任务，消除 main 上的阻塞；对团队是低风险的小回归修复。
- 风险标记：上游 #51255 回归修复，测试 stub 与配置漂移风险，仅测试变更

关联脉络

- PR #51255 [Model] Add native Dots3 NOTE multimodal support: 本 PR 修复其引入的回归：#51255 将 MLACommonMetadataBuilder 改为从 static_forward_context 的层对象读取 MLA 维度，并更新了兄弟测试与 MLA backend 测试的 stub，却漏改了本测试文件。
- PR #51280 [ROCm][CI] Solidify entrypoint LLM lifecycle: 同为 ROCm CI 稳定性修复方向，关注测试生命周期与资源回收，与本 PR 共同服务于 ROCm 测试健康度。