PR #47074 完整报告

vllm-project/vllm

[Bugfix] Use larger workspace size for Flashinfer MLA LSE

合并时间：2026-06-30 13:11

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47074

执行摘要

- 一句话：增大 FlashInfer MLA LSE 工作区大小
- 推荐动作：值得合并：修复明确，改动集中，测试通过（见 PR body 中 GSM8K 准确率 92.87%）。开发者可关注该模式：当未来引入更大模型时，可能需要进一步增大或动态调整缓冲区。

功能与动机

运行 Kimi K2.6 模型时，FlashInfer MLA LSE 因工作区不足抛出 Buffer overflow 错误：RuntimeError: Error in function 'aligned_alloc' ... Buffer overflow when allocating memory for trtllm_gen_softmax_workspace with size 135266304 and alignment 16, but only 125829120 bytes available in AlignedAllocator. Increase the workspace buffer size.（见 PR body 中的错误栈）。

实现拆解

1. 新增 FLASHINFER_MLA_LSE_WORKSPACE_BUFFER_SIZE = 256 * 1024 * 1024 常量，专用于 LSE 场景。
2. 引入模块级变量 _fi_workspace: torch.Tensor | None = None，替代原来的全局固定张量 g_fi_workspace。
3. 定义函数 _get_workspace_buffer(return_lse: bool) -> torch.Tensor，根据 return_lse 选择 256 MB 或 128 MB 大小，并在 _fi_workspace 为 None 或现有缓冲区不足时重新分配（惰性初始化）。
4. 在 FlashInferMLAImpl.__init__ 中移除 self._workspace_buffer = g_fi_workspace 赋值。
5. 在 forward_mqa 中调用 _get_workspace_buffer(return_lse) 获取工作区，并传入 trtllm_batch_decode_with_kv_cache_mla。

关键文件：

- vllm/v1/attention/backends/mla/flashinfer_mla.py（模块 注意力后端；类别 source；类型 core-logic；符号 _get_workspace_buffer）：唯一修改文件，包含所有核心变更：新增常量、惰性分配函数、按需选择 workspace 大小。

关键符号：_get_workspace_buffer

关键源码片段

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

唯一修改文件，包含所有核心变更：新增常量、惰性分配函数、按需选择 workspace 大小。

```
# vllm/v1/attention/backends/mla/flashinfer_mla.py

# 原有常量（128 MB），用于非 LSE 场景
FLASHINFER_MLA_WORKSPACE_BUFFER_SIZE = 128 * 1024 * 1024

# 新增常量（256 MB），用于需要返回 LSE 的场景，例如 Kimi K2.6
FLASHINFER_MLA_LSE_WORKSPACE_BUFFER_SIZE = 256 * 1024 * 1024

# 模块级变量，初始为 None，实现惰性分配
_fi_workspace: torch.Tensor | None = None


def _get_workspace_buffer(return_lse: bool) -> torch.Tensor:
    """根据是否需要 LSE 返回合适大小的 workspace buffer。"""
    global _fi_workspace

    # 选择缓冲区大小：使用 LSE 时需 256 MB，否则 128 MB
    buffer_size = (
        FLASHINFER_MLA_LSE_WORKSPACE_BUFFER_SIZE
        if return_lse
        else FLASHINFER_MLA_WORKSPACE_BUFFER_SIZE
    )
    # 仅在首次调用或现有缓冲区不足时重新分配
    if _fi_workspace is None or _fi_workspace.numel() < buffer_size:
        _fi_workspace = torch.zeros(
            buffer_size, dtype=torch.uint8, device="cuda"
        )
    return _fi_workspace

# forward_mqa 中的关键调用部分

# 根据当前解码是否需要 LSE 获取工作区
workspace_buffer = _get_workspace_buffer(return_lse)

# 调用 FlashInfer MLA 解码内核，传入工作区
kernel_out = trtllm_batch_decode_with_kv_cache_mla(
    query=q,
    kv_cache=kv_c_and_k_pe_cache.unsqueeze(1),
    workspace_buffer=workspace_buffer, # 之前是 self._workspace_buffer（固定 128 MB）
    qk_nope_head_dim=self.qk_nope_head_dim,
    kv_lora_rank=self.kv_lora_rank,
    qk_rope_head_dim=self.qk_rope_head_dim,
    ...
)
```

评论区精华

无 reviewer 讨论。作者 @wzhao18 在 issue 评论中说明该修复对运行 Kimi 2.6 DCP 是必需的。

- 暂无高价值评论线程

风险与影响

- 风险：低风险：变更仅影响 FlashInfer MLA 解码路径，通过判断 return_lse 决定缓冲区大小，非 LSE 场景行为不变。惰性分配确保仅在使用 LSE 时多占用额外 128 MB 显存。
- 影响：影响范围限于使用 FlashInfer MLA 后端且需要返回 LSE 的场景，主要是 Kimi K2.6 等大模型。修复后避免因工作区不足而崩溃。
- 风险标记：缺少测试覆盖

关联脉络

- PR #47079 [Bugfix][MLA] Fix LSE log-base mismatch in DCP + FlashInfer MLA *decode*: 同样涉及 FlashInfer MLA *decode* LSE 问题，与当前 PR 属于同一功能线（Kimi K2.6 DCP 支持）。