

PR #48154 完整报告

vllm-project/vllm

[ROCm] Revert Part of `[ROCm] Fix pooling startup workspace lock` #47912

合并时间: 2026-07-10 04:34

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48154>

执行摘要

- 一句话: 回滚导致 CI 失败的 CUDA graph 预热变更
- 推荐动作: 该 PR 是紧急修复 CI 失败的临时措施, 值得关注其后续可能的重新实现或替代方案。对于 AMD CI 稳定性有直接改善。

功能与动机

PR#47912 修复了 pooling 启动 workspace 锁的问题, 但引入的 ROCm 特殊预热逻辑导致 AMD CI 中 `Model Executor` 测试失败 (如 `test_reload_weights`)。Issue 评论还指出该 revert 对 `test_draft_model_tensor_parallelism` 也是必要的。

实现拆解

1. 移除 ROCm MRV1 预热代码: 删除 `compile_or_warm_up_model` 中针对 ROCm、非 V2 model runner、非 eager mode 且为最后 ranks 的特殊预热逻辑 (约 15 行), 该逻辑原本将 `max_num_reqs` 加入预热大小列表以确保 CUDA graph 后 workspace 已锁定。
2. 调整 `cg_capture_sizes` 作用域: 将 `cg_capture_sizes` 的声明从函数开头移到 `VLLM_COMPILE` 条件分支内, 使其仅在编译模式下定义, 避免在非编译模式下使用未初始化的变量。

关键文件:

- `vllm/v1/worker/gpu_worker.py` (模块 工作器; 类别 source; 类型 core-logic; 符号 `compile_or_warm_up_model`): 核心变更文件, 移除了 ROCm MRV1 的 CUDA graph 预热特殊逻辑, 并调整了 `cg_capture_sizes` 作用域。

关键符号: `compile_or_warm_up_model`

关键源码片段

`vllm/v1/worker/gpu_worker.py`

核心变更文件, 移除了 ROCm MRV1 的 CUDA graph 预热特殊逻辑, 并调整了 `cg_capture_sizes` 作用域。

```
# vllm/v1/worker/gpu_worker.py (partial)
@instrument(span_name="Warmup (GPU)")
def compile_or_warm_up_model(self) -> CompilationTimes:
```

```

warmup_sizes: list[int] = []

if self.vllm_config.compilation_config.mode == CompilationMode.VLLM_COMPILE:
    # warm up sizes that are not in cudagraph capture sizes,
    # but users still want to compile for better performance
    compile_sizes = self.vllm_config.compilation_config.compile_sizes
    warmup_sizes = compile_sizes.copy() if compile_sizes is not None else []
    cg_capture_sizes: list[int] = [] # 作用域调整：移出外层，仅在此分支定义

    if self.vllm_config.compilation_config.cudagraph_mode != CUDAGraphMode.NONE:
        cg_sizes = self.vllm_config.compilation_config.cudagraph_capture_sizes
        cg_capture_sizes = [] if cg_sizes is None else cg_sizes
        warmup_sizes = [x for x in warmup_sizes if x not in cg_capture_sizes]

    # ... compile_ranges 处理 ...

# 移除了以下 ROCm MRV1 特殊预热代码块（约 15 行）：
# if current_platform.is_rocm() and not self.use_v2_model_runner ...
# warmup_sizes.append(max_num_reqs)

for size in sorted(warmup_sizes, reverse=True):
    logger.info("Compile and warming up model for size %d", size)
    self.model_runner._dummy_run(size, skip_eplb=True, remove_lora=False)
self.model_runner.maybe_remove_all_loras(self.model_runner.lora_config)

kernel_warmup(self)
# ... capture model ...

```

评论区精华

无实质性 review 讨论；AndreasKaratzas 直接批准了 PR。Issue 评论中 divakar-amd 补充说明该 revert 对 spec_decode 测试也是必要的。

- 暂无高价值评论线程

风险与影响

- 风险：回滚后，PR#47912 修复的 workspace 锁竞态问题可能重新暴露，但 PR#47912 的其他部分（如 warmup 顺序调整）仍保留，风险可控。回滚仅影响 ROCm 平台的 MRV1 模型 runner，不影响 MRV2 或 NVIDIA 平台。
- 影响：仅影响 ROCm 平台上的 MRV1 模型 runner，移除的预热逻辑原本用于避免 CUDA graph 捕获后 workspace 增长导致的锁问题。回滚后，使用 MRV1 的 AMD GPU 用户在特定场景（如 token-classification pooling 结合 GptOss/FusedMoE 模型）可能再次遇到 workspace 锁错误。
- 风险标记：竞态条件回归风险，仅影响 ROCm MRV1

关联脉络

- PR #47912 [ROCm] Fix pooling startup workspace lock: 本 PR 部分回滚了 PR#47912 的改动，该 PR 引入了导致 CI 失败的 ROCm 预热逻辑。