PR #48440 完整报告

vllm-project/vllm

Re-disable CUDA graph memory profiling on ROCm

合并时间：2026-07-13 11:59

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/48440

执行摘要

- 一句话：在 ROCm 上禁用 CUDA graph 内存分析
- 推荐动作：该 PR 内容简洁，值得快速 review 和合入。对于理解 vllm 多平台适配中的条件检查和性能权衡（特别是 CUDA graph 内存分析在不同硬件平台上的行为差异）有参考价值。无需特别精读，但可作为平台差异化处理的案例。

功能与动机

在 ROCm 平台上，PR #47366 引入的 CUDA graph 内存分析导致稳定态 decode 吞吐量下降，需要回退以恢复性能。PR 说明中明确提及 'On ROCm the profiling capture regresses steady-state decode throughput'。

实现拆解

1. 修改平台检查条件：在 vllm/v1/worker/gpu_worker.py 的 determine_available_memory 方法中，将 CUDA graph 内存分析的触发条件从 current_platform.is_cuda_alike() 改为 current_platform.is_cuda()。is_cuda_alike() 会覆盖 CUDA 和 ROCm（HIP），而 is_cuda() 仅针对 NVIDIA CUDA 平台。
2. 更新注释：将原本说明包含 ROCm 的注释替换为 'Skip on ROCm/HIP/XPU as graph pool handles and get_memory_info behave differently and can produce incorrect/negative estimates.'，明确了跳过 ROCm 的原因。
3. 更新后续注释：在 cudagraph_memory_estimate_applied 计算处添加注释说明 'On ROCm, cudagraph_memory_estimate is always 0 so this is a no-op.' 以澄清行为。
4. 保留测试侧修复：PR #47366 中为 ROCm CI 添加的 wait_for_rocm_memory_to_settle 测试辅助函数被保留，不涉及删除。

关键文件：

- vllm/v1/worker/gpu_worker.py（模块 工作节点；类别 source；类型 core-logic）：核心变更文件，修改了 CUDA graph 内存分析的条件判断和注释，直接影响 ROCm 平台内存初始化流程。

关键符号：未识别

关键源码片段

vllm/v1/worker/gpu_worker.py

核心变更文件，修改了 CUDA graph 内存分析的条件判断和注释，直接影响 ROCm 平台内存初始化流程。

```
# vllm/v1/worker/gpu_worker.py - determine_available_memory 方法中的关键片段 # Profile CUDA graph memory if graphs will be captured. # Skip on ROCm/HIP/XPU as graph pool handles and get_memory_info # behave differently and can produce incorrect/negative estimates. cudagraph_memory_estimate = 0 if ( current_platform.is_cuda() # 改为 is_cuda()，仅对 NVIDIA CUDA 生效 and self.vllm_config.compilation_config.cudagraph_mode != CUDAGraphMode.NONE ): cudagraph_memory_estimate = self.model_runner.profile_cudagraph_memory() # ... 后续代码 ... # On ROCm, cudagraph_memory_estimate is always 0 so this is a no-op. # On CUDA, respect the opt-in flag as originally designed. cudagraph_memory_estimate_applied = ( cudagraph_memory_estimate if envs.VLLM_MEMORY_PROFILER_ESTIMATE_CUDAGRAPHS else 0 )
```

注：is_cuda() 严格检查 NVIDIA CUDA 平台，而 is_cuda_alike() 还会匹配 AMD ROCm（HIP）。此修改确保 ROCm 上不执行 profile_cudagraph_memory()，避免性能退化。注释也相应更新。

评论区精华

该 PR 没有 review 评论线程。审核人 AndreasKaratzas 直接批准（LGTM），且 claude[bot] 的自动评论仅说明来自 fork 的 PR 审核受限。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险（低）：此 PR 实质上是回退 PR #47366 中的一部分改动，恢复到了更早的状态。如果 PR #47366 本意是为了统一平台行为，那么回退后 CUDA 和 ROCm 在 CUDA graph 内存分析上行为再次不一致。但鉴于 PR #47366 的改动在 ROCm 上被证实引起性能退化，此回退是合理的。
  2. 性能影响（低）：仅在 ROCm 平台上生效，CUDA 平台行为不变。ROCm 上跳过分析后，decode 吞吐量应恢复至 #47366 之前水平。
  3. 兼容性（低）：无 API 或配置变更。 - 影响：影响范围：仅影响 ROCm（AMD GPU）用户，在初始化时跳过 CUDA graph 内存分析步骤。CUDA 用户无影响。影响程度：中等。对 ROCm 用户而言，这一回退修复了显著的性能退化（decode 吞吐量下降），但代价是可能不再准确估计 CUDA graph 所需内存，可能导致在 GPU 内存紧张时出现 OOM 或 KV cache 分配不足。但根据注释，ROCm 上 profile_cudagraph_memory() 可能产生不准确 / 负值，因此估计不可靠，跳过是安全的选择。团队影响：无。
- 风险标记：平台行为不一致

关联脉络

- PR #47366 Enable CUDA graph memory profiling on ROCm: 本 PR 回滚了 #47366 对 gpu_worker.py 的改动，恢复了 ROCm 上 CUDA graph 内存分析的禁用状态。