PR #48483 完整报告

vllm-project/vllm

lower memory required for capturing cudagraphs for large cudagraph sizes

合并时间：2026-07-13 22:14

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/48483

执行摘要

- 一句话：修复 CUDA graph 捕获时 KV cache 过度分配
- 推荐动作：简单但重要的 bugfix，值得快速合入。建议阅读 _init_minimal_kv_cache_for_profiling 方法了解 profiling 流程。

功能与动机

PR body 指出：之前每个 token 分配 1 个 KV cache 块，对于大图尺寸会 OOM；实际只需要每个序列 1 个块。修复使大 CUDA graph 捕获可行。

实现拆解

在 vllm/v1/worker/gpu_model_runner.py 的 _init_minimal_kv_cache_for_profiling 方法中，将 min_blocks 从 self.compilation_config.max_cudagraph_capture_size or 1 改为 min(self.max_num_reqs, self.compilation_config.max_cudagraph_capture_size) or 1，确保最小 KV cache 分配量为序列数而非 token 数。

关键文件：

- vllm/v1/worker/gpu_model_runner.py（模块 模型运行器；类别 source；类型 data-contract；符号 _init_minimal_kv_cache_for_profiling）：修改了 CUDA graph profiling 时最小 KV cache 分配量的计算逻辑，从每个 token 1 块改为每个序列 1 块。

关键符号：_init_minimal_kv_cache_for_profiling

关键源码片段

vllm/v1/worker/gpu_model_runner.py

修改了 CUDA graph profiling 时最小 KV cache 分配量的计算逻辑，从每个 token 1 块改为每个序列 1 块。

```
# vllm/v1/worker/gpu_model_runner.py

def _init_minimal_kv_cache_for_profiling(self) -> None:
    from vllm.v1.core.kv_cache_utils import (
        get_kv_cache_config_from_groups,
        get_kv_cache_groups,
    )
```

```
kv_cache_spec = self.get_kv_cache_spec()
KVCacheSpecRegistry.check_kv_cache_spec_registry(kv_cache_spec)
kv_cache_groups = get_kv_cache_groups(self.vllm_config, kv_cache_spec)
# 修正：每序列仅需 1 个块，而非每 token 1 个块
min_blocks = (
    min(self.max_num_reqs, self.compilation_config.max_cudagraph_capture_size)
    or 1
)

# 临时覆盖 num_gpu_blocks_override 以分配最小 KV cache
saved_override = self.cache_config.num_gpu_blocks_override
self.cache_config.num_gpu_blocks_override = min_blocks
minimal_config = get_kv_cache_config_from_groups(
    self.vllm_config, kv_cache_groups, available_memory=0
)
self.cache_config.num_gpu_blocks_override = saved_override

self.initialize_kv_cache(minimal_config, is_profiling=True)
self.cache_config.num_gpu_blocks = minimal_config.num_blocks

logger.debug("Initialized minimal KV cache for CUDA graph profiling")
```

评论区精华

reviewer tomeras91 指出 max_cudagraph_capture_size 被误用，正确值应为 min(max_num_reqs, max_cudagraph_capture_size)，并追溯到原代码引入时的注释确认每序列只需 1 块。

- min_blocks 计算修正 (correctness): 一致同意修正，已合并。

风险与影响

- 风险：低风险。改动仅修改一行数值计算，逻辑上更正确；若 max_num_reqs 为 0，or 1 确保至少分配 1 块。未发现回归或安全风险。
- 影响：直接影响 CUDA graph 捕获阶段内存分配，使得大块序列（如 long context）也能成功完成 profiling，不会触发 OOM。用户无需任何配置更改即可受益。
- 风险标记：核心路径变更

关联脉络

- PR #30515 introduce cudagraph memory profiling: 原始引入 _init_minimal_kv_cache_for_profiling 的 PR，本次 PR 修复了其中的过度分配 bug。