PR #35162 完整报告

vllm-project/vllm

[Model Runner V2] Enable piecewise & full CUDA graphs for pipeline parallelism

合并时间：2026-03-23 04:48

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/35162

执行摘要

此 PR 为 vllm-project/vllm 仓库的 V2 模型 runner 添加了流水线并行下的 piecewise CUDA graph 支持，解决了此前 PP 场景只能使用 *eager* 模式导致的性能瓶颈。通过引入持久中间张量缓冲和 PP-aware 捕获逻辑，性能吞吐量提升约 66%，TTFT 和 TPOT 显著改善，使 V2 模型 runner 在 PP 下性能对齐 V1 基线。PR 包含关键设计权衡，如 num_reqs 调整 workaround，值得工程师精读以理解分布式 CUDA graph 优化。

功能与动机

为什么做：V2 模型 runner 在流水线并行启用时无法使用 CUDA graph 捕获，只能回退到 *eager* 模式，限制了推理性能。PR 作者在 body 中明确表示："V2 model runner did not support CUDA graph capture with PP, falling back to eager mode. This PR adds piecewise CUDA graph capture for PP." 关联 Issue #33960 可能提供了更多背景。目标是通过启用 CUDA graph 捕获来提升 PP 场景的性能，减少运行时开销。

实现拆解

关键改动模块：

1. model_runner.py：
   - 添加持久 self.intermediate_tensors 缓冲，为非首 PP rank 预分配内存，确保图形重放时地址稳定。
   - 更新 *capture_model* 函数，传入中间张量以支持捕获。
   - 在 *execute_model* 中，实现从接收张量到缓冲的复制逻辑，示例代码片段：python n = input_batch.num_tokens_after_padding for k, v in intermediate_tensors.tensors.items(): self.intermediate_tensors[k][:n].copy_(v[:n])
2. *cudagraph_utils.py*：
   - 扩展 *capture* 函数，添加 PP rank 状态判断和中间张量参数。
   - 引入 num_reqs 调整逻辑，确保 num_tokens 可被整除，以避免 TRTLLM *decode* 断言错误，代码片段：python if num_reqs > 0 and num_tokens > num_reqs and num_tokens % num_reqs != 0: tokens_per_req = cdiv(num_tokens, num_reqs) num_reqs = num_tokens // tokens_per_req if num_tokens % num_reqs != 0: num_reqs = 1

评论区精华

Review 讨论中聚焦于两个关键点：

1. 中间张量键一致性：gemini-code-assist[bot] 建议添加 assertion 以确保键匹配，但代码未实现，留下潜在风险。

"The code assumes that the keys in intermediate_tensors.tensors received from the previous pipeline stage are identical to the keys in the persistent self.intermediate_tensors.tensors... To improve robustness... it's safer to assert that the sets of keys are identical."

2. num_reqs 调整设计：yewentao256 询问调整原因，ZhanqiuHu 解释为应对 TRTLLM decode 断言，但不确定是否为正确方法。

"With CUDA graph capture, I ran into AssertionError: TRTLLM decode requires uniform query lengths per request... So I added this workaround... but not sure if this is the right approach." WoosukKwon 批准 PR 并提及将跟进 FULL graph 支持，表明此 PR 是阶段性改进。

风险与影响

技术风险：

- 中间张量键不匹配可能导致运行时错误或静默数据损坏。
- num_reqs 调整是 workaround，可能在某些场景下影响性能或引入非预期行为。
- PP 下 CUDA graph 捕获首次启用，需警惕内存对齐、跨 rank 同步等边缘情况。

影响分析：

- 性能提升：根据测试结果，吞吐量从 13.89 req/s 提升至 23.07 req/s，TTFT 从 231ms 降至 167ms，TPOT 从 17.5ms 降至 10.4ms，显著改善用户体验和系统效率。
- 系统演进：为 V2 模型 runner 添加核心功能，促进向新架构迁移，团队可借鉴中间张量管理设计。

关联脉络

与历史 PR 的关系：

- PR #34903 被 yewentao256 在 issue 评论中提及，同为 V2 模型 runner 的 CUDA graph 支持 PR，可能涉及 full graph 功能，可作为后续参考。
- 从近期历史 PR 看，仓库持续优化性能（如 FP8 kernel、ROCm 改进），此 PR 延续了性能提升趋势，聚焦于 CUDA graph 在分布式场景的应用。 整体上，此 PR 是 V2 模型 runner 演进中的关键一步，为流水线并行提供了高效的图形捕获方案，后续可能扩展至 full graph 支持。