PR #27758 完整报告

sgl-project/sglang

Revert "Share BCG output buffers across capture sizes"

合并时间：2026-06-10 11:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27758

执行摘要

- 一句话：回退 BCG 输出缓冲共享优化
- 推荐动作：建议合并。这是一个回退操作，目标明确、改动量小、风险低。后续如果仍需共享 buffer 优化，应在充分测试和 review 后重新提交。

功能与动机

PR body 明确说明 "Reverts sgl-project/sglang#27659"，且原 PR #27659 的 CI 状态显示为失败（:x:），因此回退是为了消除原优化引入的潜在正确性或稳定性问题。PR #27721 新增的 TP 服务进程 GPU 上下文回归测试也可能间接要求更保守的 BCG 行为。

实现拆解

1. 删除 _slice_output 方法：该方法负责递归地对输出中的 Tensor、PPProxyTensors、tuple/list 进行切片；回退后此逻辑不再需要。
2. 删除 _copy_output_to_buffer 方法：该方法处理不同 capture size 的输出拷贝到共享 buffer；回退后每个 capture size 拥有独立 buffer，无共享拷贝。
3. 简化 _capture_one 签名：移除 shared_output_buffer 参数，回退到原始单 buffer 分配。
4. 简化 _capture_all 中的循环：移除 shared_output_buffer 变量的初始化和追踪逻辑，每个 _capture_one 各自分配输出 buffer。
5. 调整 import：移除 typing.Any 的导入，因为 _slice_output 和 _copy_output_to_buffer 不再需要。

关键文件：

- python/sglang/srt/model_executor/breakable_cuda_graph_runner.py（模块 CUDA 图；类别 source；类型 core-logic；符号 _slice_output, _copy_output_to_buffer, _capture_one）：唯一修改的文件：删除了共享 buffer 逻辑（_slice_output、_copy_output_to_buffer）并简化了 _capture_one 和 _capture_all。

关键符号：_capture_one, _capture_all

关键源码片段

python/sglang/srt/model_executor/breakable_cuda_graph_runner.py

唯一修改的文件：删除了共享 buffer 逻辑（_slice_output、_copy_output_to_buffer）并简化了 _capture_one 和 _capture_all。

```
def _capture_all(self):
    """Capture breakable CUDA graphs for all token sizes."""
    with (
        freeze_gc(self.model_runner.server_args.enable_cudagraph_gc),
        graph_capture() as graph_capture_context,
        enable_breakable_cuda_graph(),
    ):
        stream = graph_capture_context.stream
        pool = get_global_graph_memory_pool()

        capture_range = (
            tqdm.tqdm(list(reversed(self.capture_num_tokens)))
            if get_tensor_model_parallel_rank() == 0
            else reversed(self.capture_num_tokens)
        )
        # 删除了 shared_output_buffer 变量
        for num_tokens in capture_range:
            # ...（省略进度条打印）
            # 回退前：graph, output = self._capture_one(num_tokens, pool, stream, shared_output_
            buffer)
            # 回退后：不再传递 shared_output_buffer
            graph, output = self._capture_one(num_tokens, pool, stream)
            # 删除了 shared_output_buffer = output 的赋值
            self.graphs[num_tokens] = graph
            self.output_buffers[num_tokens] = output
```

评论区精华

本 PR 无 review 评论。原 PR #27659 也未记录明显的讨论。回退的主要动机大概率来自 CI 失败或内部测试发现的问题。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：回退操作简单且目标明确，移除的代码是新增功能，回退后恢复为先前长期稳定运行的逻辑。无需担心新引入的回归。
- 影响：影响范围小：仅影响 breakable_cuda_graph_runner.py 一个文件，属于 BCG 模块。回退后 GPU 显存占用可能略微增加（因为不再共享 buffer），但消除了因共享 / 拷贝逻辑不完善可能引入的正确性风险。对用户无功能可见变化。
- 风险标记：暂无

关联脉络

- PR #27659 Share BCG output buffers across capture sizes: 本 PR 回退的对象：原优化因 CI 失败等原因被完全撤销。