PR #30472 完整报告

sgl-project/sglang

Revert "Increase the KV cache pool when using indexShare by 15% (#30310)"

合并时间：2026-07-08 14:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/30472

执行摘要

- 一句话：回退导致 IMA 崩溃的 KV cache 池改动
- 推荐动作：建议审慎合入，后续应在修复 IMA 的根因后以正确方式重新实现 indexer 容量优化。同时建议补充针对 DSA 模型 + indexShare 场景的回归测试。

功能与动机

PR #30310 引入的变更触发了非法内存访问（IMA）崩溃，导致服务不稳定。PR body 明确说明："This causes an IMA, and reverting it fixes the crash"。回退是解决该问题的直接手段。

实现拆解

1. 回退 memory_pool.py 中 DSATokenToKVPool 类的变更：移除 __init__ 方法中的 skip_topk_layers 参数以及相关逻辑。所有层（包括原本跳过的 topk 层）现在都分配同样大小的 indexer KV cache buffer（index_k_with_scale_buffer）。move_kv_cache、get_cpu_copy、load_cpu_copy、get_state_buf_infos 方法中的层级跳过逻辑也随之移除。
2. 回退 pool_configurator.py 中 cell_size 计算：移除 _compute_cell_size 方法中根据 enable_hisparse 或 is_draft_worker 以及 dsa_layer_skips_topk 计算 num_indexer_layers 的复杂逻辑，改为对所有层使用 num_layers。
3. 回退 model_runner_kv_cache_mixin.py 中参数传递：移除在构造 DSATokenToKVPool 时传递 skip_topk_layers 参数的代码，并移除对 dsa_layer_skips_topk 的 import。

关键文件：

- python/sglang/srt/mem_cache/memory_pool.py（模块 内存池；类别 source；类型 core-logic；符号 DSATokenToKVPool.init, DSATokenToKVPool.move_kv_cache, DSATokenToKVPool.get_cpu_copy, DSATokenToKVPool.load_cpu_copy）：核心修改：移除 DSATokenToKVPool 中对 skip_topk_layers 的依赖，所有层分配相同大小的 indexer buffer。
- python/sglang/srt/model_executor/pool_configurator.py（模块 配置器；类别 source；类型 data-contract）：修改 _compute_cell_size 方法，移除按层计算 indexer 大小的逻辑，改为简单使用 num_layers。
- python/sglang/srt/model_executor/model_runner_kv_cache_mixin.py（模块 运行器；类别 source；类型 data-contract；符号 _init_pools）：移除向 DSATokenToKVPool 构造器传递 skip_topk_layers 的代码。

关键符号：DSATokenToKVPool.init, DSATokenToKVPool.move_kv_cache, DSATokenToKVPool.get_cpu_copy, DSATokenToKVPool.load_cpu_copy, DSATokenToKVPool.get_state_buf_infos

关键源码片段

python/sglang/srt/mem_cache/memory_pool.py

核心修改：移除 DSATokenToKVPool 中对 skip_topk_layers 的依赖，所有层分配相同大小的 indexer buffer。

```
# 在 DSATokenToKVPool.__init__ 中，移除 skip_topk_layers 参数后，
# 所有层都分配相同大小的 indexer KV cache buffer，不再为 topk 层分配零大小。

class DSATokenToKVPool(MLATokenToKVPool):
    def __init__(
        self,
        size: int,
        page_size: int,
        kv_lora_rank: int,
        dtype: torch.dtype,
        qk_rope_head_dim: int,
        layer_num: int,
        device: str,
        index_head_dim: int,
        enable_memory_saver: bool,
        kv_cache_dim: int,
        start_layer: Optional[int] = None,
        end_layer: Optional[int] = None,
        index_buf_size: Optional[int] = None,
        # skip_topk_layers 参数已被移除
    ):
        # ... 初始化代码 ...
        self.index_k_with_scale_buffer = [
            torch.zeros(
                (
                    (index_buf_size + page_size + 1) // self.page_size,
                    self.page_size
                    * (
                        index_head_dim + index_head_dim // self.quant_block_size * 4
                    ),
                ),
                dtype=self.index_k_with_scale_buffer_dtype,
                device=device,
            )
            for _ in range(layer_num) # 所有层都分配
        ]
```

评论区精华

由于 review 评论为空，且仅有 Fridge003 批准，本次变更为直接回退，无实质讨论。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 性能回退：回退后，原本被跳过的 topk 层的 indexer KV cache 将重新分配内存，导致 indexer 池大小缩减，可能影响高并发场景下的 token 容量。
  2. 无功能性损失：变更仅涉及回退，不引入新逻辑；但需验证其他依赖 skip_topk_layers 的代码（如 HiSparse）是否正常工作。
  3. 测试覆盖不足：PR 未包含针对回退后正确性的回归测试。 - 影响：直接影响使用 DeepSeek DSA 模型（indexShare）的用户，indexer KV cache 池大小恢复为原始值，可能降低可处理的 token 数量；但解决了 IMA 崩溃，提升了稳定性。对非 DSA 模型无影响。 - 风险标记：性能回退，缺少测试覆盖

关联脉络

- PR #30310 Increase the KV cache pool when using indexShare by 15%: 本 PR 回退的原因：PR #30310 引入的 IMA 崩溃导致回退。