PR #27973 完整报告

sgl-project/sglang

[DSV4] Use int64 for compressor out_loc tensors

合并时间：2026-06-12 08:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27973

执行摘要

- 一句话：DSV4 压缩器 out_loc 从 int32 升级为 int64
- 推荐动作：建议直接合并。这是一个纯粹的预防性修复，遵循了代码库中"索引使用 int64"的最佳实践。改动小、风险低、测试充分。

功能与动机

@merrymercy 在 #27091 review 中指出"prefer int64 for any indices"，原实现中 HiSparse C4 store 路径通过 .to(torch.int32) 将物理设备槽位截断为 int32，当 KV 池超过 2^31 槽位时存在潜在的截断风险。

实现拆解

1. metadata_kernel.py：将 _init_compressed_attn_metadata_triton 中 c4_out_loc 和 c128_out_loc 的创建 dtype 从 torch.int32 改为 torch.int64。
2. compressor_v2.py：在 HiSparse C4 store 路径中，使用 _translate_loc_to_hisparse_device(out_loc)（返回 int64）替代原来的 translate_loc_to_hisparse_device(out_loc).to(torch.int32)，去除显式截断。
3. fused_norm_rope_v2.cuh：将三个核函数变体（indexer、indexer-fp4、flashmla）中的 out_loc 参数类型从 const int32_t* 改为 const int64_t*，同时将内部 page/offset 运算的局部变量从 int32_t 升级为 int64_t。
4. test_fp4_indexer.py：测试 test_fp4_fused_norm_rope_store_layout 中 loc 张量的创建 dtype 从 torch.int32 改为 torch.int64。

其他 store 核函数（_set_k_and_s、store.cuh、fp4 index-cache 等）均已支持 int64，无需额外修改。

关键文件：

- python/sglang/srt/layers/attention/dsv4/metadata_kernel.py（模块 压缩器；类别 source；类型 core-logic；符号 _init_compressed_attn_metadata_triton）：核心变更文件：将 c4_out_loc 和 c128_out_loc 的 dtype 从 int32 改为 int64，这是整个变更的起点。
- python/sglang/srt/layers/attention/dsv4/compressor_v2.py（模块 压缩器；类别 source；类型 core-logic；符号 forward_unified）：移除 HiSparse 路径中显式的 int32 截断，改用直接返回 int64 的内部方法。

- python/sglang/jit_kernel/csrc/deepseek_v4/fused_norm_rope_v2.cuh（模块 JIT 核函数；类别 other；类型 core-logic；符号 fused_norm_rope_indexer, fused_norm_rope_indexer_fp4, fused_norm_rope_flashmla）：CUDA 核函数：out_loc 参数从 int32改为 int64，内部 page/offset 变量升级为 int64 以匹配。
- test/registered/jit/deepseek_v4/test_fp4_indexer.py（模块 测试；类别 test；类型 test-coverage；符号 test_fp4_fused_norm_rope_store_layout）：测试文件中 loc 张量的 dtype 从 int32 改为 int64，确保测试与生产代码一致。

关键符号：_init_compressed_attn_metadata_triton, forward_unified, fused_norm_rope_indexer, fused_norm_rope_indexer_fp4, fused_norm_rope_flashmla, test_fp4_fused_norm_rope_store_layout

关键源码片段

python/sglang/srt/layers/attention/dsv4/metadata_kernel.py

核心变更文件：将 c4_out_loc 和 c128_out_loc 的 dtype 从 int32 改为 int64，这是整个变更的起点。

```
# python/sglang/srt/layers/attention/dsv4/metadata_kernel.py

def _init_compressed_attn_metadata_triton(
    seq_lens: torch.Tensor,
    positions: torch.Tensor,
    raw_out_loc: torch.Tensor,
    page_table: Optional[torch.Tensor] = None,
    page_size: int = 0,
    compute_page_indices: bool = True,
) -> Tuple[...]:
    bs = seq_lens.shape[0]
    device = seq_lens.device

    # 变更：从 int32 改为 int64，避免当 KV 池槽位超过 2^31 时发生截断
    c4_out_loc = torch.empty(bs, dtype=torch.int64, device=device)
    c4_positions = torch.empty(bs, dtype=torch.int32, device=device) # 位置信息仍为 int32
    c4_seq_lens_raw = torch.empty(bs, dtype=torch.int32, device=device)
    c4_seq_lens_clamp1 = torch.empty(bs, dtype=torch.int32, device=device)

    c128_out_loc = torch.empty(bs, dtype=torch.int64, device=device)
    c128_positions = torch.empty(bs, dtype=torch.int32, device=device)
    c128_seq_lens_raw = torch.empty(bs, dtype=torch.int32, device=device)
    c128_seq_lens_clamp1 = torch.empty(bs, dtype=torch.int32, device=device)
    # ... 后续代码不变，所有消费 c4_out_loc / c128_out_loc 的 kernel 均已支持 int64
```

python/sglang/srt/layers/attention/dsv4/compressor_v2.py

移除 HiSparse 路径中显式的 int32 截断，改用直接返回 int64 的内部方法。

```
# python/sglang/srt/layers/attention/dsv4/compressor_v2.py
# 在 forward_unified 方法中，else 分支的 HiSparse 处理逻辑：
```

```
else:
    _, _, compress_kv_pool = token_to_kv_pool.layer_mapping[layer_id]
    assert compress_kv_pool is not None
    kv_cache = token_to_kv_pool.get_extra_key_buffer(layer_id)
    page_size = token_to_kv_pool.get_extra_key_page_size(layer_id)
    if hasattr(compress_kv_pool, "translate_loc_to_hisparse_device"):
        # 变更前：out_loc = compress_kv_pool.translate_loc_to_hisparse_device(out_loc).
        to(torch.int32)
        # 变更后：直接使用返回 int64 的内部方法，不进行 int32 截断
        out_loc = compress_kv_pool._translate_loc_to_hisparse_device(
            out_loc
        )
# 传递给压缩 kernel 的 out_loc 现在保持为 int64
self._forward_compress_all_in_one(
    ...
    out_loc=out_loc,
    ...
)
```

评论区精华

该 PR 无 review 评论，主要依据 #27091 中 @merrymercy 的关于“prefer int64 for any indices”的反馈。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：
  - 仅改变索引张量的 dtype，不改变算法或数据布局。
  - 其他所有涉及存储的核函数（_set_k_and_s、store.cuh 模版、fp4 index-cache、HIP 路径等）已确认都兼容 int64。
  - 测试覆盖了 int64 路径下的压缩注意力、fp4 indexer 等关键场景，结果全部通过。
  - 无性能影响：只有少量 per-token 索引的张量类型变宽，算术和存储布局不变。
  - 影响：影响范围限定在 DSV4 压缩器相关的 kernel 和测试文件，用户无感知。通过消除 int32 截断风险，使系统在更大 KV 池规模下更安全。
  - 风险标记：高风险消除（int32 截断），测试覆盖充分

关联脉络

- PR #27091 （推测）首次引入 HiSparse 或 out_loc 相关的 PR: 本 PR 直接提及 #27091，作为 @merrymercy 要求使用 int64 索引的来源。