PR #29379 完整报告

sgl-project/sglang

[Fix] DSA: size cudagraph page_table to req_to_token width

合并时间：2026-06-27 03:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29379

执行摘要

- 一句话：修复 DSA CUDA Graph page_table 越界崩溃
- 推荐动作：建议审阅者重点关注：该修复对齐了 deepseek_v4_backend 的已有模式，但需确认 DSA 后端是否完全等价地处理 req_to_token 的额外列（尤其是 topk>1 或 page_size>1 场景）。

功能与动机

Long-generation runs with EAGLE speculative decoding on a DSA model (e.g. GLM-5.2-NVFP4) crash in the cudagraph verify path once a request's seq_len approaches the context length，报错为 RuntimeError: The size of tensor a (90006) must match the size of tensor b (90007) at non-singleton dimension 1。

实现拆解

1. 修改 init_cuda_graph_state 方法中 page_table 的宽度计算（python/sglang/srt/layers/attention/dsa_backend.py）：将原本的 self.max_context_len + (self.speculative_num_draft_tokens or 0) 改为 self.req_to_token.shape[1]。
2. 更新注释：新注释说明需要精确匹配 req_to_token 的宽度，因为推测解码会使 seq_len 短暂超过 context_len。
3. 对齐已有实现：deepseek_v4_backend 已使用 MAX_SEQ_LEN_FOR_CAPTURE = self.req_to_token.shape[1]，本改法遵循相同模式。

关键文件：

- python/sglang/srt/layers/attention/dsa_backend.py（模块 注意力；类别 source；类型 core-logic；符号 init_cuda_graph_state）：核心修复文件，修改了 CUDA Graph 的 page_table 缓冲区宽度定义，直接解决形状不匹配崩溃。

关键符号：init_cuda_graph_state

关键源码片段

python/sglang/srt/layers/attention/dsa_backend.py

核心修复文件，修改了 CUDA Graph 的 page_table 缓冲区宽度定义，直接解决形状不匹配崩溃。

```
# python/sglang/srt/layers/attention/dsa_backend.py
# 位于 init_cuda_graph_state 方法中
# 构建 CUDA Graph 的固定尺寸张量字典
self.decode_cuda_graph_metadata: Dict = {
    "cache_seqlens": torch.ones(max_num_tokens, dtype=torch.int32, device=self.device),
    "cu_seqlens_q": torch.arange(0, max_bs + 1, dtype=torch.int32, device=self.device),
    "cu_seqlens_k": torch.zeros(max_bs + 1, dtype=torch.int32, device=self.device),
    # 用于 sparse_prefill 的伪 page_table
    # 必须精确匹配 req_to_token 的宽度，因为推测解码会导致
    # seq_len 短暂超过 context_len（req_to_token 被额外分配了偏移）
    # 若不匹配，后续 copy_ 操作将因形状不同而崩溃
    "page_table": torch.zeros(
        max_num_tokens,
        self.req_to_token.shape[1], # 改为动态读取实际宽度
        dtype=torch.int32,
        device=self.device,
    ),
    "flashmla_metadata": (
        self._compute_flashmla_metadata(...) if self.dsa_decode_impl == "flashmla_kv" else None
    ),
}
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅变更了 page_table 缓冲区的宽度计算，从硬编码的 max_context_len + speculative_num_draft_tokens 改为动态读取 req_to_token.shape[1]。后者在训练阶段已确定且稳定，不会引入新的形状不匹配。
- 影响：影响范围限于 DSA 注意力后端的 CUDA Graph capture 路径。修复后，长序列（接近 context length）结合 EAGLE 推测解码的场景不再崩溃，提升了大上下文推断的可靠性。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #29142 [DeepSeek V3] Run routed experts on main stream in dual-stream MoE: 与 deepseek_v4_backend 共享类似模式，本 PR 参考了 deepseek_v4_backend 的 MAX_SEQ_LEN_FOR_CAPTURE 实现。